Genomics 4.0 - Syntenic Gene and Genome Duplication Drives Diversification of Plant Secondary Metabolism and Innate Immunity in Flowering Plants

- Advanced Pattern Analytics in Duplicate Genomes -

Johannes A. Hofberger

Thesis committee

Promotor
Prof. Dr M. Eric Schranz
Professor of Experimental Biosystematics
Wageningen University

Other members
Prof. Dr Bart P.H.J. Thomma, Wageningen University
Prof. Dr Berend Snel, Utrecht University
Dr Klaas Vrieling, Leiden University
Dr Gabino F. Sanchez, Wageningen University

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

Genomics 4.0 - Syntenic Gene and Genome Duplication Drives Diversification of Plant Secondary Metabolism and Innate Immunity in Flowering Plants

- Advanced Pattern Analytics in Duplicate Genomes -

Johannes A. Hofberger

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 18 May 2015
at 4 p.m. in the Aula.

Johannes A. Hofberger
Genomics 4.0 - Syntenic Gene and Genome Duplication Drives Diversification of Plant Secondary Metabolism and Innate Immunity in Flowering Plants
83 pages.

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summaries in Dutch and English

ISBN: 978-94-6257-314-7

PROPOSITIONS

1. Ohnolog over-retention following ancient polyploidy facilitated diversification of the glucosinolate biosynthetic inventory in the mustard family.
   (this thesis)

2. Resistance protein conserved in structurally stable parts of plant genomes confer pleiotropic effects and expanded functions in plant innate immunity.
   (this thesis)

3. Integrating the science of management with the science of life leads to more and better results.

4. In every personality, obvious attributes are linked to hidden attributes like genes are linked when sharing one chromosome.

5. If you recognize what is in your sight, that which is hidden from you will become plain to you for there is nothing hidden which cannot become manifest

6. It doesn't matter if it is genes, thoughts, people or stocks: the whole is more than the sum of its parts.

Propositions belonging to the thesis, entitled

"Genomics 4.0 - Syntenic Gene and Genome Duplication Drives Diversification of Plant Secondary Metabolism and Innate Immunity in Flowering Plants"

Johannes A. Hofberger,
Mai 18th, 2015

TABLE OF CONTENTS

Genomics 4.0 - Syntenic Gene and Genome Duplication Drives Diversification of Plant Secondary Metabolism and Innate Immunity in Flowering Plants

Johannes A. Hofberger[1, 2, 3]

[1] Biosystematics Group, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands (*August 2012 – December 2013*)

[2] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands (*December 2010 – July 2012*)

[3] Chinese Academy of Sciences/Max Planck Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai 200031, PR China (*January 2014 – December 2014*)

TWO-SENTENCE SUMMARY

Large-scale comparative analysis of Big Data from next generation sequencing provides powerful means to exploit the potential of nature in context of plant breeding and biotechnology. In this thesis, we combine various computational methods for genome-wide identification of gene families involved in (a) plant innate immunity and (a) biosynthesis of defense-related plant secondary metabolites across 21 species, assess dynamics that affected evolution of underlying traits during 250 Million Years of flowering plant radiation and provide data on more than 4500 loci that can underpin crop improvement for future food and live quality.

GENERAL ABSTRACT

As sessile organisms, plants are permanently exposed to a plethora of potentially harmful microbes and other pests. The surprising resilience to infections observed in successful lineages is due to a complex defense network fighting off invading pathogens. Within this network, a sophisticated plant innate immune system is accompanied by a multitude of specialized biosynthetic pathways that generate more than 200,000 secondary metabolites with ecological, agricultural, energy and medicinal importance. The rapid diversification of associated genes was accompanied by a series of duplication events in virtually all plant species, including local duplication of short sequences as well as multiplication of all chromosomes due to meiotic errors (plant polyploidy). In a comparative genomics approach, we combined several bioinformatics techniques for large-scale identification of multi-domain and multi-gene families that are involved in plant innate immunity or defense-related secondary metabolite pathways across 21 representative flowering plant genomes. We introduced a framework to trace back duplicate gene copies to distinct ancient duplication events, thereby unravelling a differential impact of gene and genome duplication to molecular evolution of target genes. Comparing the genomic context among homologs within and between species in a phylogenomics perspective, we discovered orthologs conserved within genomic regions that remained structurally immobile during flowering plant radiation. In summary, we described a complex interplay of gene and genome duplication that increased genetic versatility of disease resistance and secondary metabolite pathways, thereby expanding the playground for functional diversification and thus plant trait innovation and success. Our findings give fascinating insights to evolution across lineages and can underpin crop improvement for food, fiber and biofuels production.

GENERAL INTRODUCTION

From the Neolithic to the Genomics revolution: 12,000 years of plant biology

The understanding of plant biology has facilitated the origin of modern civilization and greatly contributed to human life quality ever since. The history of modern civilization began with a first agricultural ("Neolithic") evolution, starting around 12,000 years ago in the Holocene with domestication of various types of plants [1]. In this context, specialized food-crop cultivation led to increased yields and triggered a gradual transformation of small and mobile groups of hunter-gatherers to larger non-nomadic societies based in build-up settlements [2].

A second agricultural revolution rationalized crop production and coincided with the onset of industrialization 200 years ago [3]. Yields rose beyond subsistence and facilitated a near-exponential growth of the human population [4]. However, the increasing population led to land conversion and a decrease of the limited area of arable land [5]. Likewise, the potential of many available crop varieties reached its limits and yield overages were shrinking gradually. At the peak of this development, many countries came to the brink of food shortage and related famine in the mid of the 20th century [6].

In a third agricultural ("Green") revolution, a series of research, development and technology transfer initiatives were initiated between the late 1940s and the late 1960s with the aim to create higher-yield crops [7]. Due to these efforts, the total production of cereals doubled in developing nations between the years 1961–1985 [8]. Once again, a detailed understanding of plant biology ensured the availability of food and hence better life quality for an estimated billon of people [9].

Uncovering the double helix structure of the DNA macromolecule and subsequent development of polymerase chain reaction (PCR)-technology in the 1980s facilitated high-throughput genotyping of plants [10]. Before that, crop improvement was still limited to classical forward genetics approaches that rely on time-consuming cross-breeding of individuals that carry unusual traits [11]. Research on varieties bred during the Green revolution and many others resulted in functional characterization of genes associated with increased yields and better food quality [12, 13]. However, information on the genetic basis of underlying traits was limited to few model species and domestic crop lineages. Furthermore, the lack of whole genome assemblies made it difficult to identify putative orthologous loci and encoded functions in distant species. Therefore, insights into evolutionary processes contributing to functional gene family dynamics were difficult to obtain in most cases and thus incompletely understood [14].

The release of the first fully sequenced flowering plant genome *Arabidopsis thaliana* marked the onset of a fourth agricultural (or Genomics-) revolution with the begin of the 21st century (Genomics 4.0) [15, 16]. Within the last 15 years, progress in next generation sequencing technology has accelerated with a rate beyond Moore's law, stating that computer chips of a given price double their performance every two years (**fig. 1**) [17, 18]. This boom resulted in an overwhelming abundance of genomics data that to date comprise more than 50 well-assembled flowering plant genomes ("Big Data") [19]. Hand in hand with the increased availability of large biological datasets, information technology has become an essential part for better understanding of plant systems [20, 21]. More and more algorithms are now emerging in the rapidly growing fields of bioinformatics and systems biology, capable of inferring non-apparent relationships from the analysis and comparison of genes, genomes and other datasets [22]. For example, comparing novel draft genomes to genomes with multitudes of characterized loci provides important means for plant breeding [23-26]. In this context, various applications facilitate the localization of target genes for genetic modification

or migration between lineages with little functional data at hand [27]. Furthermore, it is now possible to monitor dynamics that affect functional gene families during flowering plant evolution in a phylogenomics framework due to the a better coverage of sequenced genomes in many lineage representatives [23]. Together with better understanding the evolution of gene families associated with nutritionally and economically important traits, the genomics revolution facilitates genomics-based plant breeding crop improvement for faster production of better food, fiber and biofuels. In this thesis, we designed and employed a novel combination of several bioinformatics tools to identify multi-gene families associated with four independent key traits across 21 representative flowering plant genomes (**fig. 5**). We perform large-scale evolutionary analysis of both plant secondary metabolism and plant innate immunity and provide data that will underpin genomics-based crop breeding for better food and life quality.

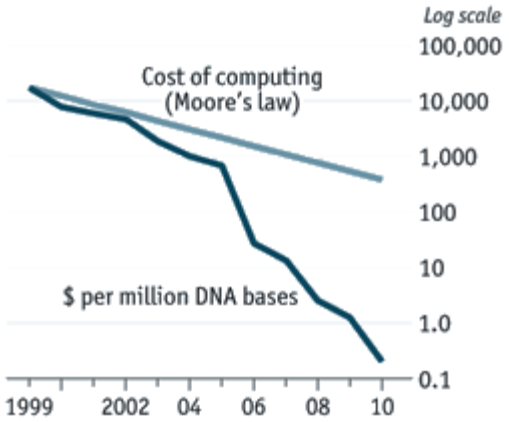


FIG. 1.— Cost of genome sequencing compared to Moore's law for computers (adapted from [28]).

Gene and genome duplication and multi-gene family evolution

In the 1970s, Ohno postulated a key role of gene duplication for molecular evolution and trait innovation [29]. With the genomics revolution, next generation sequencing and bioinformatics tools led to a steady increase of functional and structural genomics data [30]. Following Ohno's hypothesis, this now facilitates a large-scale comparison of different duplication modes in many lineages and integration of available data on function of gene duplicates. A better understanding of duplication history and gene family evolution can underpin efficient curation of functional genes associated with specific traits for plant breeding.

Comparing the genomic distribution of homologous genes within and between plant species revealed the important contribution of short sequence duplication events to gene family evolution. Such events include tandem duplication of one or few neighboring genes resulting in supergene clusters or tandem arrays of neighboring homologs. Tandem duplication can be due to unequal crossing-over of chromosomes or errors during DNA repair and affected 15% of all protein-coding genes in *A. thaliana* (**fig. 2A**) [31]. Likewise, gene transposition duplication leads to homologs embedded in distant genomic positions and are due to transposon activity [32]. In *A. thaliana*, 14% of the protein-coding genes transposed at least once during lineage evolution (**fig. 2B**) [33, 34].

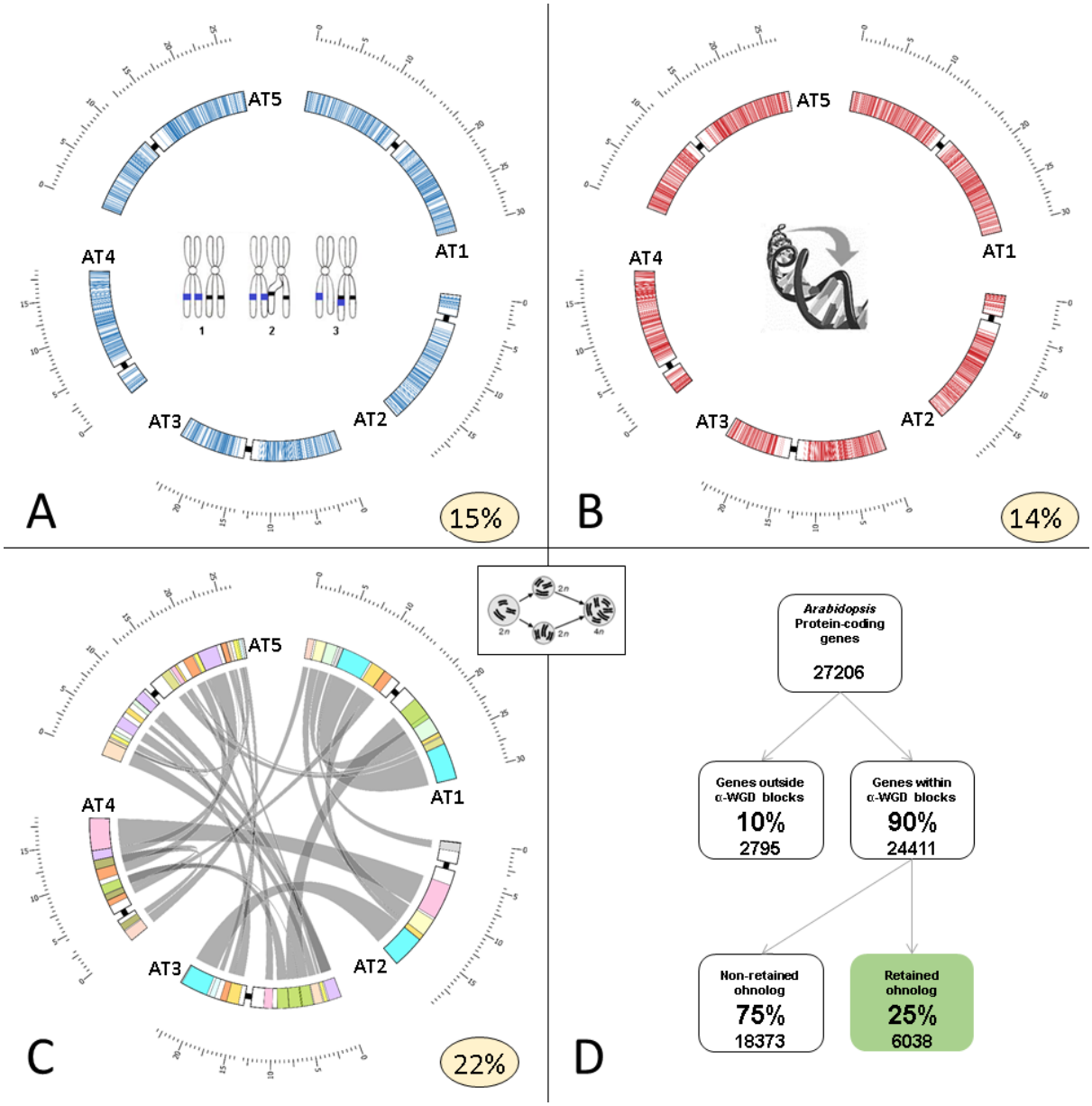


FIG. 2.— Genome-wide distribution of duplicates among all protein-coding genes in *A. thaliana* according to origin of duplication. **A**. 15% of all protein-coding genes are organized in supergene clusters due to tandem duplication. **B**. 14% of all protein-coding genes comprise copies due to gene transposition duplication. **C**. 22% of all protein-coding genes comprise duplicate gene pairs due to whole genome duplication (ohnologs) and are organized in 26 syntenic blocks (marked in color-coding as introduced by). **D**. 90% of all protein-coding genes locate within syntenic blocks but only 25% of genes covered by blocks retain the ohnolog copy. AT1-AT5 indicates *A. thaliana* chromosomes 1-5. Scale indicates chromosome length in million base pairs (MB). For methods, see Chapter 1.

Within- and between-species comparison of large genomic regions has established a common history of recurring ancient duplication events that affect all genes/chromosomes at once due to meiosis errors and are shared by all flowering plant lineages (polyploidy events) [35-38]. Following these polyploidy events, a genome-wide fractionation process retained groups of duplicate genes that reveal block-like patterns, also known as “syntenic blocks” of genes (ohnologs or paralogs created from WGDs) scattered throughout the replicated genome (**fig. 3A**) [26, 39, 40]. In *Arabidopsis*, 22% of all protein-coding genes comprise ohnologs organized in 26 syntenic blocks

retained following the most recent At-α WGD event (**fig. 3B, 2C**). While At-α blocks cover around 90% of the genome, not every gene within syntenic block boundaries is retained after genome multiplication. For example, about 25% of genes duplicated during the last WGD event stably retained pairwise in *Arabidopsis*, whereas 75% got lost (fractionated) after duplication during lineage evolution (**fig. 2D**) [35].

The model plant *A. thaliana* underwent at least five polyploidy events during lineage evolution, two preceding and three following the origin of flowering plants (**fig. 5**) [37, 41, 42]. The At-α whole genome duplication (WGD) event is shared by all other mustard family members, including the extant sister clade Aethionemeae [35, 43, 44]. This was predated by the At-β WGD event that occurred after the split of the *Carica* lineage, but is shared by most other lineages in the Brassicales [45, 46]. The more ancient polyploidy event detectable in the *Arabidopsis* genome is a genome triplication (WGT) termed At-γ, shared by all Asterids and Rosids, grape (Vitales) as well as more basal eudicot clades such as *Pachysandra terminalis* (Buxales) and *Gunnera manicata* (Gunnerales) [47, 48]. Similarly, the Poaceae lineage (grasses) underwent at least three independent polyploidy events after flowering plant radiation. The most recent rho (ϱ) WGD event is predated by the sigma (σ) WGD event. An even more ancient WGD event was made evident predating monocot radiation (monocot tetraploidy) [36, 49, 50]. All aforementioned polyploidy events are shared by larger clades with multiple completed genome sequences. In addition more recent, clade-specific genome multiplications have been identified in various lineages (**fig. 5**). The mesopolyploid genome of the crop *Brassica rapa* underwent a triplication (Br-α WGT) [51, 52]. Another clade-specific WGT event was discovered in *Tarenaya hasslerania* (formerly *Cleome spinosa*, Cleomaceae), family sister to the Brassicaceae [45, 53, 54]. Likewise, soybean (*Glycine max*) and corn (*Zea mays*) genomes have both been subject of recent genome doublings [55-61].

Gene identification – homologs, orthologs, paralogs and ohnologs

Both whole genome- and short sequence duplication create novel genes with sequence homology to at least one duplicate copy (homologs). Orthologs refer to homolog genes from different species that are due to one ancestral locus and diverged due to speciation [62]. Paralogs refer to homologs within one species that are due to short sequence duplication events [63]. In contrast, ohnologs refer to paralogs that are due to whole genome multiplication events. By definition, ohnologous genes comprise syntenic orthologs (**fig. 4**). Identification of orthologs involved in desired traits is of paramount importance for plant breeding. In a first step, it is necessary to identify all homologs present in a genome that belong to a distinct gene family. In a second step, the distinction of orthologs, paralogs and ohnologs provides insights into dynamics that affected the evolution of the specific trait. In this context, scoring of synteny can provide efficient means to distinguish orthologs from paralogs, thereby illustrating both genome structural and functional evolution.

Increasing evolutionary distance between species presents limitations for homolog (ortholog and/or paralog) identification based on DNA sequence homology (such as blastn) due to the degeneration of the genetic code on the 3rd codon position as well variable gaps of non-coding sequences [64, 65]. Hence, protein sequence identity and profile searches are used to infer homologs of protein-coding loci across distant clades with better sensitivity [66, 67]. Notably, proteins are organized in functional units termed domains [68]. For example, 37% of the *A. thaliana* Col-0 TAIR10 representative proteins contain more than one characterized protein domain [69]. Therefore, robust ortholog gene identification based on the encoded protein sequence involves identification of more than one pool of homolog genes, each specific for one domain only. This is followed by overlapping

all identified pools to detect multi-domain encoding genes sharing a combination of all desired protein motifs.

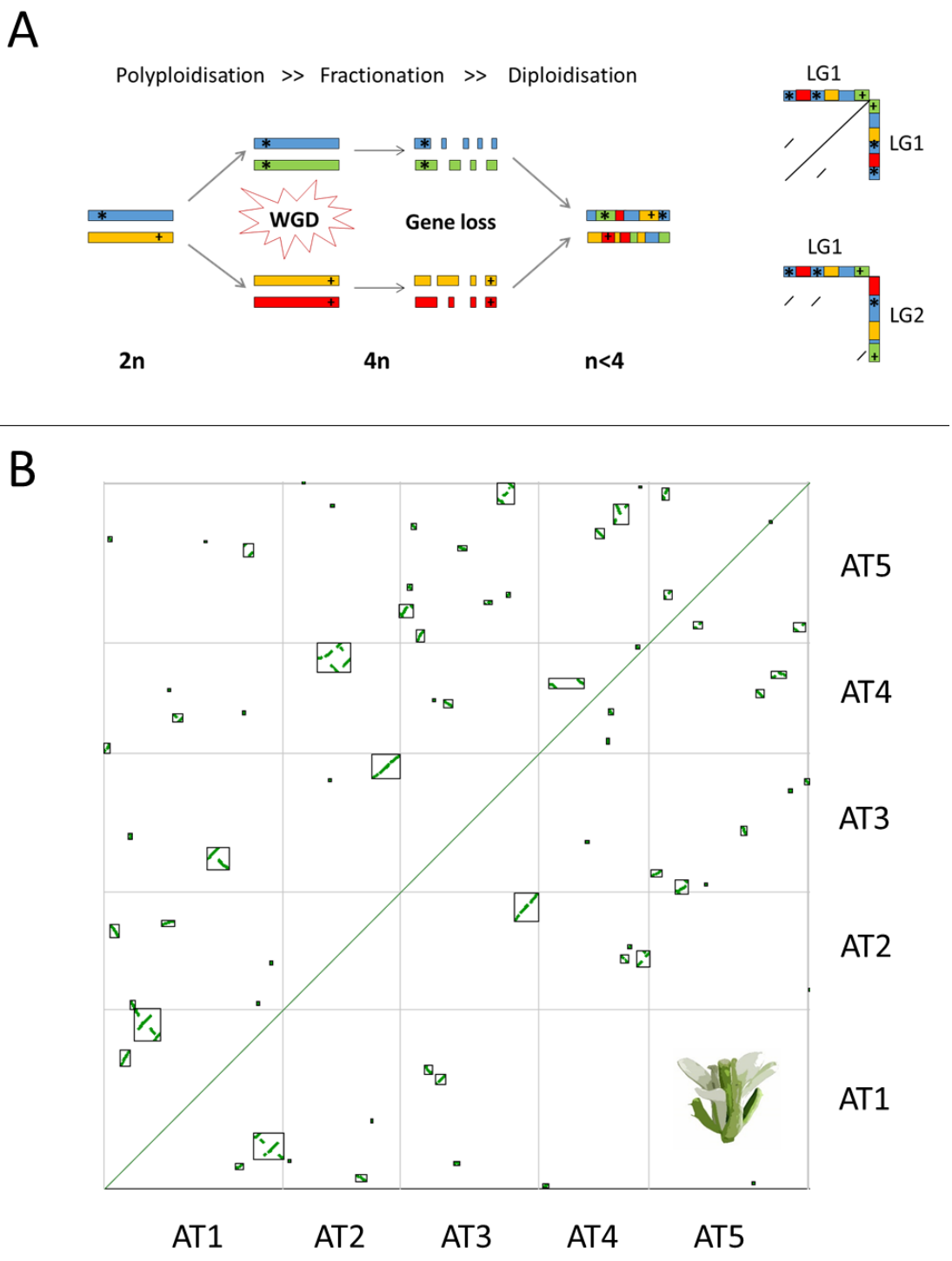


FIG. 3.— Ancient whole genome duplication (polyploidy or WGD) events are followed by gene fractionation that retains syntenic patterns of genomic blocks scattered throughout the replicated genome. **A**. Proposed order of events following duplication of the initial diploid set of chromosomes. Shown left are two different chromosomes without their diploid copy before and after a WGD event. Shown right are cartoon representations of dot-plots that visualize duplicate regions in novel linkage groups retained after fractionation and re-diploidisation, revealing syntenic block patterns. **B**. Syntenic dot-plot of the *Arabidopsis thaliana* genome, showing 26 collinear blocks due to the At-α WGD event distributed across all five chromosomes. This experiment can be reproduced online

following the CoGe link https://genomevolution.org/r/elkt (last accessed on December 13th, 2014). For methods, see Chapter 3.

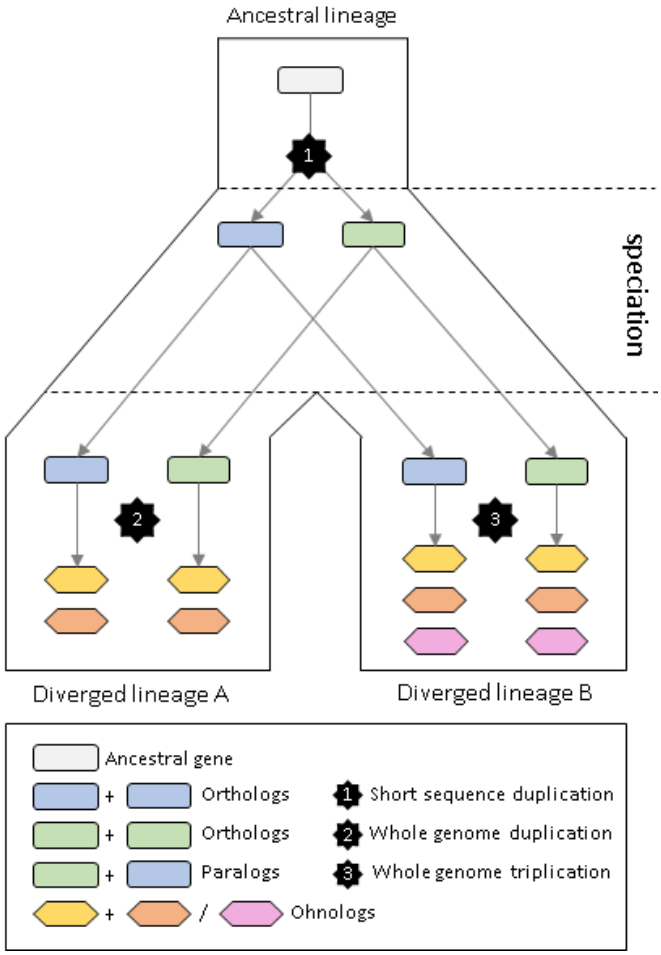
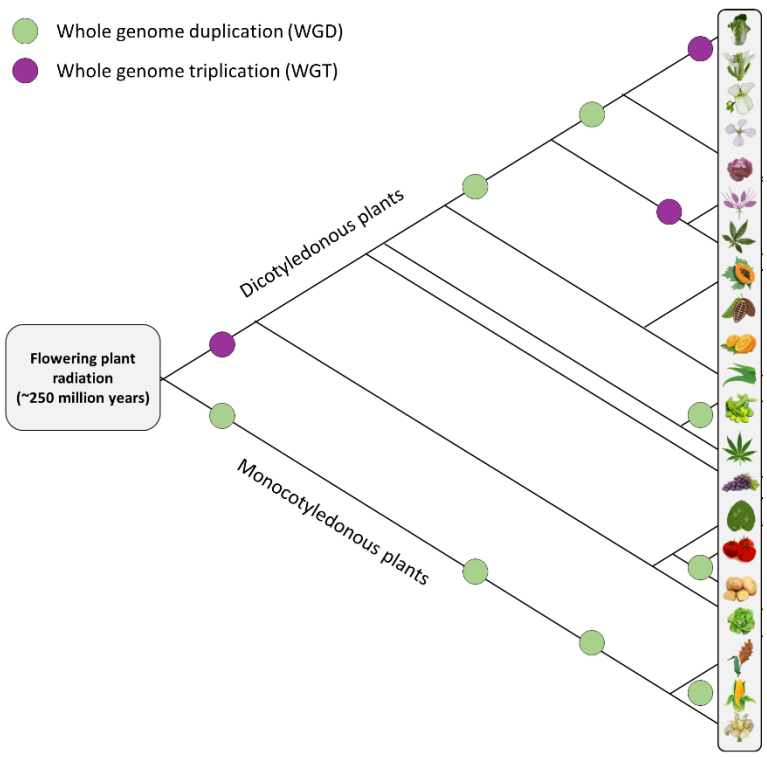


FIG. 4.— The copy number variation of homolog genes is influenced by both short sequence- and whole genome duplication. Depending on the harboring linage and the origin of duplication, homologs are referred to as orthologs, paralogs and ohnologs.

Currently, there are several methods available for homolog detection. Among them, the determination of reciprocal best blast hits (RBH) at key phylogenetic nodes is the easiest and hence the most common method to infer families of orthologous genes that represent the modern descendants of ancestral gene sets (used in, for example, [70]). RBH comprise pairs of genes in two different genomes that are more similar to each other than either is to any other gene in the other genome. It is now evident that RBH-only approaches miss up to 60% of true orthologs in duplicate-rich species, which are particularly prevalent among angiosperm plants [71]. Blasting Hidden Markov Model (HMM)-generated protein domain consensus sequences against a translated genome assembly is a more accurate way (used in, for example, [72]). However, the quality of results depends on the accuracy of the input sequences used to generate the consensus as well as the exonic structure of the proteins (i.e. some domains can span multiple small exons which can reduce the sensitivity of the translated blast). This creates challenges when analyzing highly-diverged gene

families in more distant clades. In contrast, OrthoMCL [73] employs a Markov Cluster algorithm to group putative orthologs and paralogs in all-vs.-all blast screen within and between given genome assemblies. Gene families defined by OrthoMCL therefore represent relatively compact and coherent clusters of similar proteins. However, OrthoMCL is unaware of the domain structure found in the family members, is computationally intensive and impracticable for application to dozens of genome annotations given a limited timeframe.

None of the similarity-based methods utilize information provided by the genomic context of the putative ortholog / paralog gene pairs. This is relevant because two members of the same gene family can evolve differently in two lineages to a degree that the ortholog in lineage A produces a RBH to the paralogs in lineage B and vice versa [44]. This produces false ortholog assignments but can be clarified by scoring and weighting gene synteny evidence [52]. Likewise, synteny evidence can lead to novel assignment of highly diverged genes to functional families in cases where sequence homology or domain composition is ambiguous [74]. In this thesis, we designed and employed a novel, iterative approach by combining blast, HMM modeling and genomic contextual information provided by synteny to identify robust multi-domain and multi-gene families. Likewise, we investigated specific duplication modes that affected gene copy number, thereby providing in-depth information on the evolutionary history of identified gene families. In the first part of this thesis, we focused on plant secondary metabolite pathways (**fig. 5**). In Chapter 1, we identified and analyzed the extended biosynthetic inventory of the glucosinolate (GS) pathway within two sister lineages of the mustard family. GS comprise a group of sulfurous plant secondary metabolites with important roles in ecology, human health and nutrition [75]. In Chapter 2, we identified and analyzed more than 1,900 genes involved in modular terpenoid biosynthesis in an Angiosperm-wide perspective. Terpenoids comprise plant secondary metabolites with various roles in plant-environmental interaction, such as pollinator attraction or pathogen defense [76]. Likewise, terpenoid biosynthetic genes are important targets for plant breeding due to their connection to crop smell and scent. In the second part of this thesis, we investigated multi-gene families with key roles in plant innate immunity (**fig. 5**). In Chapter 3, we identified and analyzed more than 2,000 Angiosperm resistance proteins of the NB-LRR type that convey important roles in pathogen effector recognition and the second layer of plant innate immunity [77]. In Chapter 4, we identified and analyzed more than 300 pattern recognition receptors of the LecRK type that are capable of pathogen detection following perception of pathogen-associated molecular patterns (PAMPs) in the first layer of plant innate immunity [78]. For all analyzed traits and genomes, we provide data on all identified loci in order to contribute to future experiments for generation of more and better crops, ultimately necessary to meet the global food production and to maintain life quality for an ever-increasing and demanding population.



PLANT SPECIES	CHAPTER 1 Glucosinolates	CHAPTER 2 Terpenoids	CHAPTER 3 *NB-LRR* genes	CHAPTER 4 *LecRK* genes	
Brassica rapa		page 37 ff.	page 72 ff.	page 98 ff.	Brassicaceae
Arabidopsis thaliana	page16 ff.	page 37 ff.	page 72 ff.	page 98 ff.	
Arabidopsis lyrata			page 72 ff.	page 98 ff.	
Eutrema parvulum			page 72 ff.	page 98 ff.	
Aethionema arabicum	page16 ff.		page 72 ff.	page 98 ff.	
Tarenaya hasslerania		page 37 ff.	page 72 ff.	page 98 ff.	Brassicales
Cleome gynandra		page 37 ff.			
Caraica papaya		page 37 ff.	page 72 ff.	page 98 ff.	
Theobroma cacao				page 98 ff.	
Citrus sinensis		page 37 ff.	page 72 ff.		Fabidae
Eucalyptus grandis		page 37 ff.			
Glycine max		page 37 ff.			
Cannabis sativa		page 37 ff.			
Vitis vinifera		page 37 ff.	page 72 ff.	page 98 ff.	
Nicotiana benthamiana		page 37 ff.	page 72 ff.		Solanaceae
Solanum lycopersicum		page 37 ff.	page 72 ff.		
Solanum tuberosum		page 37 ff.	page 72 ff.		
Lactuca sativa		page 37 ff.			Poaceae
Sorghum bicolor		page 37 ff.			
Zea mays		page 37 ff.			
Amborella trichopoda		page 37 ff.			
	Analysis of genes involved in **plant secondary metabolism**		Analysis of genes involved in **plant innate immunity**		

FIG. 5.—Thesis overview: Shown left is a phylogeny that comprises 21 representative flowering plant genomes subjected to our analysis, including ancient polyploidy events that affected lineage evolution. Shown right is a matrix illustrating the species-wise investigation of four groups of genes distributed across the four independent research chapters of this Thesis.

Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family

Johannes A. Hofberger[1], Eric Lyons[2], Patrick P. Edger[3], J. Chris Pires[3] and M. Eric Schranz*[1]

[1] Biosystematics Group, Wageningen University, 6708 PB, Wageningen, The Netherlands
[2] School of Plant Sciences; iPlant Collaborative, BIO5 Institute, 1657 East Helen Street, University of Arizona, AZ 85721 Tuscon, USA
[3] Division of Biological Sciences, 311 Bond Life Sciences Center, University of Missouri, MO 65211 Columbia, USA

ABSTRACT

Plants share a common history of successive whole genome duplication (WGD) events retaining genomic patterns of duplicate gene copies (ohnologs) organized in conserved syntenic blocks. Duplication was often proposed to affect the origin of novel traits during evolution. However, genetic evidence linking WGD to pathway diversification is scarce. We show that WGD and Tandem Duplication (TD) accelerated genetic versatility of plant secondary metabolism, exemplified with the glucosinolate (GS) pathway in the Mustard Family. GS biosynthesis is a well-studied trait, employing at least 52 biosynthetic and regulatory genes in the model plant *Arabidopsis.* In a phylogenomics approach, we identified 67 GS loci in *Aethionema arabicum* of the tribe Aethionemeae, sister group to all Mustard Family members. All but one of the *Arabidopsis* GS gene families evolved orthologs in *Aethionema* and all but one of the orthologous sequence pairs exhibit synteny. The 45% fraction of duplicates among all protein-coding genes in *Arabidopsis* was increased to 95 and 97% for *Arabidopsis* and *Aethionema* GS pathway inventory, respectively. Compared to the 22% average for all protein-coding genes in *Arabidopsis,* 52 and 56% of *Aethionema* and *Arabidopsis* GS loci align to ohnolog copies dating back to the last common WGD event. While 15% of all *Arabidopsis* genes are organized in tandem arrays, 45% and 48% of GS loci in *Arabidopsis* and *Aethionema* descend from TD, respectively. We describe a sequential combination of tandem- and whole genome duplication events driving gene family extension, thereby expanding the evolutionary playground for functional diversification and thus potential novelty and success.

KEYWORDS: comparative genomics, glucosinolates, whole genome duplication, functional diversification, *Arabidopsis, Aethionema,* Brassicaceae

*****Author for Correspondence:

M. Eric Schranz | Biosystematics Group | Wageningen University & Research Center | Wageningen, The Netherlands | Tel. +31(0)317-483160 | email: eric.schranz@wur.nl

INTRODUCTION

Gene duplication has played an important evolutionary role in angiosperm adaptation and success, for example by contributing to regulatory and enzymatic pathways involved in generating the >200,000 diverse biochemical plant secondary metabolites in the Angiosperm lineage [79]. Functional diversification refers to processes of gene duplication followed by sub- or neo-functionalization of the enzymes encoded by duplicate copies [29, 80, 81], mediating specificities to extended classes of substrates or catalysis of novel reactions [82]. Fast expansion of gene copy number occurs in various ways. In this study we focus on whole genome duplication (WGD), Tandem Duplication (TD) and gene transposition duplication (GTD). For example, nearly 45% of the *Arabidopsis* nuclear protein-coding genes have been affected by such processes [31, 32, 35]. In this study, we investigated the impact of gene duplication to the diversification of plant secondary metabolites exemplified with glucosinolate (GS) biosynthesis. Glucosinolate biosynthesis is a well-studied key trait shared by all Brassicales including the Mustard family (Brassicaceae) crown-group [83] and its sister lineage Aethionemeae. Comparative genomics analysis unraveled a history of successive paleo-polyploidy events commonly shared by almost all Angiosperms [35]. The *Arabidopsis* lineage underwent at least five polyploidy events in the history of life, two preceding and three following Angiosperms radiation [35, 37]. The most recent WGD is commonly referred to as At-α and occurred approximately 30 million years (MA) ago in the ancestor of all Brassicaceae, including the sister group Aethionemeae [84]. As a result, pairwise syntenic regions are scattered throughout the genome (genomic blocks), defined as copies of consecutive ohnologs derived from At-α [35]. It is known that polyploidy is succeeded by a genome-wide process of biased fractionation, preferentially targeting one sub-genome to retain clusters of dosage-sensitive genes organized in functional modules [85]. Furthermore, several studies have established a potential link of polyploidy to natural variation due to differential expression of ohnolog copies [81], seed and flower origin and diversification [37, 86, 87], morphological complexity [88], and survival of plant lineages at the Cretaceous-Tertiary extinction event [89]. In this study, we provide solid evidence for the link of WGD to pathway expansion of a distinct key trait relevant for herbivore defense and hence highly connected to fitness. Interestingly, polyploidy also affects other kinds of duplication, creating network of factors with mutual influence. Recent studies have shown an interaction between polyploidy and the fractionation rate of tandem duplicate (TD) copies in both *Arabidopsis* and *B. rapa* (having undergone an additional genome triplication). Hence, we analyzed short-sequence duplications to utilize the evolutionary significance of different duplication classes.

Tandem Duplication (TD) of short sequences can be caused by unequal crossing-over or template slippage during DNA repair, producing tandem arrays (TARs) of homologous genes in close genomic vicinity [90]. Depending on the number of allowed gene spacers, TAR genes include about 10 – 15 % of the *A. thaliana* genome (0 and 10 spacers, respectively) [31]. Comparison of tandem arrays in *Arabidopsis* and rice revealed enrichment of genes encoding membrane proteins and function in biotic and abiotic stress [31]. Notably, the impact of TD to trait evolution has been elucidated in multiple taxa, including disease resistance in Solanaceae [91] and Brassicaceae [92]. Likewise, TD played a role in the evolution of signal transduction, for example the expansion and functional diversification of the F-box type transcriptional activator gene family in Fabaceae [93]. Moreover, TD is an important factor for increasing versatility of defense response in Brassicaceae. In GS biosynthesis, sub-functionalization of TAR genes is evident for 2-oxoglutarate-dependent

dioxygenases (AOP) [94], flavin-monooxygenases (FMO_{GSOX}) [95] and methyl-thioalkylmalate synthases (MAM) [96-98]. In this study, we integrate previous findings to dissect the influence of polyploidy with tandem- and gene transposition duplication (GTD) in the last ~30 MA of GS pathway expansion since *Aethionema* and *Arabidopsis* lineage divergence.

Duplicate gene copies can move to a new genomic location. The observed frequency of gene movements explains the observed erosion of synteny between plant genomes during evolution [99], defining the limits of synteny-based approaches for ortholog detection. Gene movements are often caused by transposition. Gene transposition duplication (GTD) events occur when a single non-transposon gene relocates to a new position, and segregants contain duplicates [100]. Whereas transposable elements (TEs) account for approximately 10% of the *Arabidopsis* genome [32] and show non-random association to syntenic blocks [101], 14% of all protein-coding genes transposed at least once during Rosid evolution [33, 34]. Importantly, a novel genomic context of the transposed copy potentially influences rates of gene expression [102] and might thereby contribute to the phenotypic consequences of the duplication event [103]. Accordingly, TE activity was shown to foster variation of NBS-resistance proteins in grape [104] as well as natural growth variation and expansion of ERF family transcriptional regulators in *Arabidopsis* [105, 106]. In contrast, evolutionary dynamics of GTD events affecting genetic versatility of plant secondary metabolism has not yet been investigated.

Glucosinolates (GS) comprise a class of secondary plant metabolites derived from amino acids and sugars, part of a 2-component chemical defense against herbivory in Brassicales [107-109]. Myrosinase enzymes are the other component of the defense system and confer GS hydrolysis activity. They are released from the vacuole upon tissue damage, producing a plethora of GS degradation products such as nitriles, isothiocyanates, thiocyanates and ephithioalkanes with various bioactivities [110, 111]. Glucosinolates are of particular interest for human health because they can inhibit carcinogen activation [112, 113] and carcinogenesis by triggering cell cycle arrest and stimulating apoptosis [114, 115]. The observed variation in GS biochemistry across Brassicales is due to the differences in biochemistry among their amino acid precursors [108, 116] and allows GS grouping to 4 distinct classes. Oxidative deamination of phe and tyr initiates biosynthesis of indolic GS (I); trp is the substrate for indolic GS production (II); ala, val, leu and Ile are precursors for biosynthesis of aliphatic GS (III) [75]. While aromatic and aliphatic GS have been detected in other eudicot families including Phytolaccaceae, Euphorbiaceae and Pittosporaceae [116, 117], indolic GS are Brassicales-specific. Met-derived GS form a fourth class of GS (IV), referring to a subset of aliphatic GS specific to the Brassicales crown group, including the sister group Aethionemeae. The utilization of trp- and met-derived amino acids for GS production may be tied to pathway expansion caused by ancient WGD events [83].

The genus *Aethionema* of the tribe Aethionemeae is an ideal group for comparative genomics of polyploidy and GS pathway evolution. First, it shares the composite GS chemotype observed in the larger and more diverse Brassicaceae crown group [84]. Second, phylogenetic analysis highly support the tribe Aethionemeae as the earliest diverged clade and extant sister to the crown group Brassicaceae [118] with an estimated split of the two lineages approximately 30 MA ago. However, a high degree of inter-species synteny is maintained (see Results section). Third, the most recent WGD event identified in the lineage of *Arabidopsis* (referred to as At-α) predated the divergence of *Arabidopsis* and *Aethionema*. Furthermore, it was not succeeded by an additional species-specific genome polyploidization, preventing additional fractionation of synteny [35, 43]. In contrast, *B. rapa*

underwent an additional genome triplication event [51], complicating efforts to analyze the potential impact of At-α on the evolution of the GS pathway inventory.

MATERIALS & METHODS

Ae. arabicum genome assembly and set of annotated genes

Sequence assembly and annotation of the *Ae. arabicum* genome was obtained from [43].

RNA isolation and sequencing

Aethionema arabicum RNA was isolated from fresh apical meristematic tissue or very young leaves using an RNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Samples were kept on liquid nitrogen prior to RNA isolation. The optional step of heating the lysis solution to 65°C was used to maximize RNA yield. RNA was eluted into a final volume of 100 µL RNase-free water. Total mass of RNA and quality was estimated using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). Samples were deemed acceptable if RIN scores were greater than 8.0. A minimum of 20 µg of total RNA was required for library building and sequencing. RNA-seq [119] paired-end libraries with average fragment lengths of 250 base pairs (bp) were constructed, and each library was sequenced on a single lane of an Illumina GAIIX sequencer flow cell (Illumina, San Diego, California, USA) to generate a minimum of 3 gigabases of 75 bp, paired-end sequences.

Database of *A. thaliana* genes retaining an ohnologous copy dating back to the At-α WGD event

First, we generated a spreadsheet with information on all 33,323 *A. thaliana* nuclear genes annotated in the TAIR database v10, including (a) *Arabidopsis* gene identifiers (AGIs), (b) locus type, (c) locus name and (d) short description of encoded function. Second, we integrated optional affiliation to (e) syntenic block and (f) ohnolog copy dating back to At-α WGD event as previously described [88]. The corresponding authors did not account for every gene in their analysis, and inferred the genomic location of ohnolog blocks dating back to At-α using the TIGR *Arabidopsis* genome annotation v5 from 2005. Fourth, we added an additional column (i), to indicate coverage of the gene in Feeling's study (yes / not considered / not present in TIGR5).

Database for glucosinolate biosynthetic gene identification in *A. thaliana* and *Ae. arabicum*

Files containing the coding sequences representing the complete set of glucosinolate biosynthetic and regulatory (AtGS) genes in *A. thaliana* [120] were acquired from the TAIR database v10 (www.arabidopsis.org, last accessed on December 13th, 2014). We highlighted the AGIs in the spreadsheet covering all nuclear genes in *Arabidopsis*.

Interpolation of novel putative glucosinolate biosynthetic genes and retained At-α ohnolog pairs in *A. thaliana*

We utilized similarity among ohnologous gene copies. We employed the spreadsheet containing information on genomic location of AtGS genes as well as α-blocks with optional retained duplicates therein. We visually screened for all ohnolog copies of AtGS genes not sharing annotation as AtGS genes themselves. Differential expression of ohnolog pairs was tested using the botany array resource (http://bar.utoronto.ca/welcome.htm, last accessed on December 13th, 2014). We included all ohnolog copies for our analysis to create an extended AtGS gene set. For the spreadsheet, see (**supplementary table S1,** Supplementary Material online**)**. Previous analysis of ohnologous gene pair identification did not consider every protein-coding gene [35, 85]. To minimize resulting errors

for analysis of AtGS loci, we performed a BLASTP screen without a cut-off e-value, querying all AtGS genes with non-retained At-α ohnologs against all other *Arabidopsis* genes with non-retained At-α ohnologs. Highest-scoring sequence pairs (HSPs) sharing genomic location within converse copies of the same α-block were tested for synteny and positives defined as additional pairs of retained At-α ohnolog copies (marked as "our addition / _oa" in **table 8** and **supplementary table S1,** Supplementary Material online).

Database of *A. thaliana* tandem arrayed (TAR) genes

A database of *A. thaliana* coding sequences organized in tandem arrays was generated for the TAIR annotation v10 as previously described [31], using a low-stringency approach with a number of N=10 allowed gene pacers. Information was updated to TAIR10 and included to the spreadsheet covering all *Arabidopsis* nuclear genes (**supplementary table S1,** Supplementary Material online).

Database of *A. thaliana* GS genes affected by gene transposition duplication (GTD)

A database of the epoch-independent positional history of all *Arabidopsis* genes was generated as previously described for TAIR9 [34]. We updated all putative gene transposition duplication (GTD) copies to TAIR10. Woodhouse et al. scored gene duplicates as transposed based on a function of synteny across taxa in the direction *A. thaliana -> A .lyrata -> C. papaya -> P. trichocarpa - > V. vinifera*. For analysis of Brassicaceae genome evolution, methodical restrictions apply due to the low resolution within that clade, covered by only two tribes. Thus, we screened the genomic context of AtGS genes within a narrow window of 3kb for flanking TE-like sequences, using the GEvo function from the CoGe comparative genomics package (http://genomevolution.org/CoGe/GEvo.pl, last accessed on December 13th, 2014) [26]. Graphical highlights of TE-like sequences have been customized by choosing "show other features" in the "results visualization" tab. By that means, we confirmed AtGS genes that transposed at least once during lineage evolution as defined by Woodhouse et al. and identified further GTDs missed by that approach due to lack of synteny data (i.e. GTDs pre-dating *Vitis* speciation as well as recent GTDs of Brassicaceae-specific genes; marked by asterisks in **table 3**). Information on GTD events was added to an additional column in (**supplementary table S1**, Supplementary Material online).

Analysis of putative GS genes not affected by TD, GTD or At-α ohnolog retention in *Arabidopsis*

We performed additional analysis of AtGS loci beyond the above-mentioned types of duplication by considering more ancient WGD events. Information on *Arabidopsis* genome-wide distribution of ohnolog duplicate pairs dating back to the At-ß and At-γ WGD events [35] were added to an additional column in (**supplementary table S1,** Supplementary Material online). GS genes were referenced accordingly. Remnants did not show significant similarities to any other locus in *Arabidopsis* by definition and evolutionary stability was confirmed using the *Arabidopsis* transpositional history database (http://geco.iplantcollaborative.org/athaliana/, last accessed on December 13th, 2014) as well as data on AtGS syntelogs in *B. rapa* [70].

Orthologous gene identification of *Arabidopsis* glucosinolate biosynthetic genes in *Ae. arabicum*

We considered multiple lines of evidence for identification of orthologs between *A. thaliana* GS loci and *Ae. arabicum.* We defined orthologous pairs of *A. thaliana* and *Ae. arabicum* GS loci as reciprocal best hits (RBH) within a given region of gene collinearity (synteny). First, we screened for regions in the *Ae. arabicum* genome displaying synteny to genomic regions in *A. thaliana* harboring

GS loci, using the "Synfind" function with standard parameters from the CoGe comparative genomics package (www.genomeevolution.org, last accessed on December 13th, 2014) [52]. Second, we determined reciprocal best hits (RBH) between *A. thaliana* GS genes and *Ae. arabicum* genes within the syntenic regions from (i), using BLASTP with a minimum query coverage of N=0.5 and a cut-off e-value of 1E-10. Third, we queried all putative *Ae. arabicum* GS loci against the *Ae. arabicum* genome in a BLASTP-screen with a cut-off e-value of 1E-30. We screened for subject sequences not sharing the query sequence scaffold and identified syntenic regions in *A. thaliana*. If GS biosynthetic gene was present in syntenic *A. thaliana* region (BLASTP with a cut-off e-value of 1E-30), we defined the aligned *Ae. arabicum* subject sequence as ortholog to the *A. thaliana* query sequence.

Tandem arrayed (TAR) gene copy identification of putative glucosinolate biosynthetic genes in *Ae. arabicum*

We queried all putative *Ae. arabicum* GS loci against the *Ae. arabicum* genome in a BLASTP-screen with a cut-off e-value of 1E-30. For identification of Tandem Duplications, GS query sequences were grouped with the subset of respective subject sequences located within a window of N=10 allowed gene spacers to form *Ae. arabicum* super-families of putative TAR genes. Tandem Duplications were visualized using the MAFFT package (http://mafft.cbrc.jp/alignment/software/, last accessed on December 13th, 2014) [121]. We further confirmed *Ae. arabicum* GS genes expression by querying the RNA-seq data. Transcriptome data were mined for expression of GS genes using TBLASTX with a cut-off E-value=1E−10 (data not shown).

Identification of lineage-specific GS gene transposition duplications comparing putative glucosinolate biosynthetic genes in *Ae. arabicum and A. thaliana*

We queried all putative *Ae. arabicum* GS loci against the *Ae. arabicum* genome in a BLASTP-screen with a cut-off e-value of 1E-30. For identification of gene transposition duplications following divergence of these lineages, we screened for subject sequences not sharing the query sequence scaffold and identified syntenic regions in *A. thaliana*. If GS biosynthetic gene is absent in syntenic *A. thaliana* region (BLASTP with a cut-off e-value of 1E-30), we defined the aligned *Ae. arabicum* subject sequence as lineage-specific GTD copy.

Phylogenetic and Similarity Analysis.

A number of *Arabidopsis* flavin-monooxygenases involved in GS biosynthesis (*FMO GS-OX*) are encoded in clusters consisting of retained ohnolog copies as well as both tandem- and gene transposition duplicates. To visualize the evolution of *FMO*-like sequences in Brassicales, *Carica papaya* and *Tarenaya hasslerania*, *FMO* orthologs from these species were obtained using the CoGe comparative genomics package. A phylogenetic tree was constructed using the maximum-likelihood method with PhyML 3.1 software [122], employing the LG model for amino acid substitution. Protein sequence similarity analysis were performed using the Needle program from the EMBOSS software package (http://emboss.sourceforge.net/, last accessed on December 13th, 2014) [123].

Genome Data Visualization and Statistics.

Fisher exact test for count data was performed using the R package for statistical computing (www.r-project.org, last accessed on December 13th, 2014). Circular visualization of genome data was performed using the circos package (www.circos.ca, last accessed on December 13th, 2014) [124]

and graphically edited with the GIMP-package (www.gimp.org, last accessed on December 13th, 2014).

RESULTS

The Influence of the At-α WGD Event to GS Pathway Evolution in *Arabidopsis*

We first updated the genomic location of all ohnolog blocks dating back to the At-α WGD event (α-blocks thereafter) in *A. thaliana* from the TIGR5 to the TAIR10 annotation, leading to minor changes in the list published by [85] (**supplementary table S1**, Supplementary Material online). As a first step to understanding the dynamics of GS pathway evolution, we divided the 52 to-date known AtGS genes into three groups: first, genes with a retained At-α ohnolog copy (**table 1**). Second, genes with lost At-α ohnolog copy, but a genomic location covered by α-blocks (**table 3**). Third, genes located outside the genomic borders of α-blocks (**table 4**). For the original set of AtGS genes published by [120], we found an increased retention rate of 49% (24/49) for retained ohnolog copies dating back to the At-α WGD event (**fig. 1**), compared with a 22% average observed for all *Arabidopsis* protein-coding genes (**fig. 2A, B**). These 24 canonical AtGS genes group to six ohnolog pairs with annotation to GS metabolism and 12 loci lacking annotation of one ohnolog copy to GS biosynthesis (**figs. 1, 3**). Notably, the 12 ohnolog pairs sharing GS annotation and the six ohnolog pairs lacking GS annotation of one member (forming 18 AtGS ohnolog copy pairs in total) either display high degrees of pairwise similarity and/or show similar tendencies in gene expression following treatment with methyljasmonic acid, an organic volatile important for plant defense signaling [125] (**tables 5, 6**). Therefore, we inferred functional redundancy of ohnolog copies due to structural homology. We propose a significant contribution to GS metabolism and consistently include all 12 ohnolog copies lacking GS annotation to our analysis, forming 12 pairs of two ohnolog copies each. We thereby created an extended set of 64 putative AtGS genes (**figs. 1, 3**). Among genes located within α-block boundaries, we found an At-α ohnolog retention rate of 59% (36/61) for the extended AtGS set (**fig. 1**), which is more than double of the observed 22% average rate for ohnolog retention among all *Arabidopsis* protein-coding loci harbored within the boundaries of α-blocks (**fig. 2B**).

Quantification of TD and GTD influence to GS pathway evolution in *Arabidopsis*

In the next step, we quantified the impact of TD to GS pathway versatility in *Arabidopsis*. Minor changes were made in the list of *Arabidopsis* TAR genes by [31] due to the gene updates to TAIR10 (**supplementary table S1**, Supplementary Material online). We mined the 1,497 *Arabidopsis* TARs comprising 4,034 duplicate gene copies for AtGS genes. Forty-five percent (29/64) of AtGS genes are members of TARs, compared with a genome-wide average of 15% (**figs. 2B, D**). Next, we quantified the influence of GTD to GS pathway evolution in *Arabidopsis*. Initially, a list of 4,575 genes with putative origin due to a GTD was proposed for TAIR9 [34]. Our update to TAIR10 retained 4,539 loci clearly referenced to transposition events (**supplementary table S1**, Supplementary Material online), illustrating a 14% average for GTD genes among protein-coding loci in *Arabidopsis* (**fig. 2B**). Among those, we confirmed all 13 references to AtGS loci (**table 2**), using the GEvo function from the CoGe package for comparative genomics (see Materials & Methods section). We thereby discovered that four additional AtGS genes (marked by asterisks in **table 2**) lost all syntenic anchor genes in genomic proximity (±1,000 kb) but are surrounded by TE-like sequences. Thus they may have transposed following the At-α WGD event. These additional genes may be Brassicaceae specific but lost secondarily in *A. lyrata*. This might explain their absence in the *Arabidopsis* gene transpositional history database (that mainly scores pre-α GTD events due to the lack of further Brassicaceae

synteny data necessary for scoring of post-α GTDs). Hence, the total fraction of GTD copies among AtGS genes sums up to 27% (18/ 67) (**fig. 2D**).

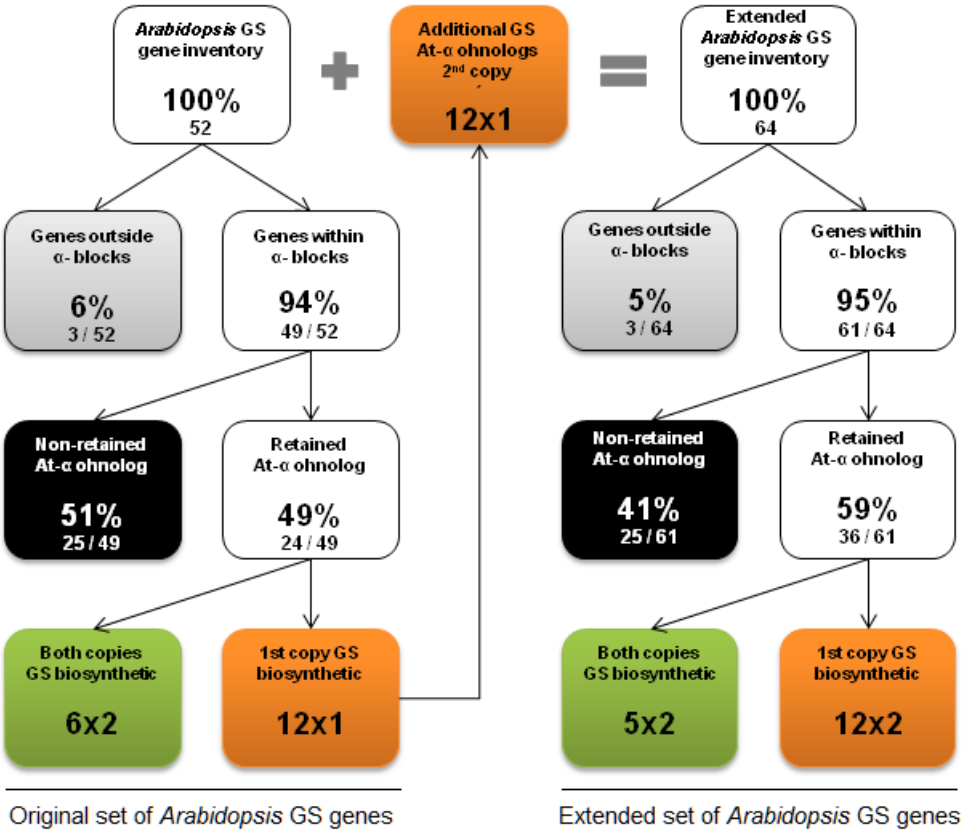


FIG. 1.— Distribution of GS pathway inventory relative to At-α WGD event. AtGS genes are shown before (left) and after (right) interpolation of ohnolog duplicate copies. We hypothesize functional redundancy of 12 additional ohnologs to canonical GS biosynthetic genes.

Analysis of GS genes not affected by TD, ohnolog retention or transposition

In addition, we performed a more in-depth analysis of the three AtGS genes lacking TD, retained At-α ohnolog copy or evidence for transposition during evolution of the *Arabidopsis* lineage (**table 7**). Among those, *MYB34* is the only locus retaining an ohnologous copy dating back to the At-ß WGD event, leaving two putative non-duplicate genes in AtGS pathway inventory: *GSH1* and *IIL1*, functioning in GS co-substrate pathways and side-chain elongation, respectively (**table 3**). To confirm the observed evolutionary stability of these genes, we identified syntelogs in *V. vinifera* (*IIL1*) and *B. rapa* (*GSH1*), respectively. Syntelogs in *Vitis* proof that *IIL1* did not transpose since the birth of the Rosids. Therefore, this gene represents a very ancient unigene. In case of duplication before *Vitis* lineage evolution, all copies were lost subsequently before radiation of the Rosid clade. In contrast, *GSH1* may be Brassicaceae-specific unigene that likewise lost all duplicates with above-threshold similarity.

TABLE 1.— Retained At-α ohnolog duplicate gene pairs in *Arabidopsis* and *Aethionema* GS pathway inventory

Protein Name[A]	AGI	α-Block	Evident SSD[B]	AabID[C]	Syntelog	% identity	Col-0 -> Aab[D]
Core-structure formation							
UGT74C1	AT2G31790	A02N051	TD	Aab37175	Yes	79.44	6->10
[UGT-like]	AT1G05670	A02N051	TD	Aab31930	Yes	81.11	6->10
FMO-GSOX-2	AT1G62540	A03N117	TD	Aab10869	Yes	76.6	11->8
[FMO-like]	AT1G12130	A03N117	TD	Aab13543	Yes	65.09	11->8
CYP79F2	AT1G16400	A05N062	TD	-	-	-	8->9
CYP79C1	AT1G79370	A05N062	GTD	Aab34143	Yes	76.8	8->9
SOT16	AT1G74100	A05N186	TD	Aab14278	Yes	91.07	3->3
SOT17	AT1G18590	A05N186		Aab19675	Yes	86.05	3->3
SUR1	AT2G20610	A10N194		Aab31155	Yes	89.15	3->2
[SUR-like]	AT4G28420	A10N194	TD	Aab30136	Yes	57.93	3->2
CYP79B2	AT4G39950	A10N257	GTD	Aab17805	Yes	81.38	8->9
CYP79B3	AT2G22330	A10N257	GTD	Aab19477	Yes	81.4	8->9
GGP1	AT4G30530	A10N314	TD	Aab24374	Yes	87.6	5->5
[GGP-like]	AT2G23960	A10N314	TD	Aab11021	Yes	61.38	5->5
GSTF11	AT3G03190	A12N102		Aab14996	Yes	77.1	4->4
[GSTF12]	AT5G17220	A12N102		Aab14791	Yes	77.84	4->4
Co-substrate pathways							
[AAO3]	AT2G27150	A02NOA1	GTD	Aab27016		77.67	2->2
AAO4	AT1G04580	A02NOA1		Aab24896	Yes	79.58	2->2
APK1	AT2G14750	A10NOA2		Aab32150	Yes	83.75	2->2
APK2	AT4G39940	A10NOA2	GTD	Aab17804	Yes	86.05	2->2
Side-chain elongation							
BCAT4	AT3G19710	A08N074		Aab21007	Yes	75	6->6
[BCAT7]	AT1G50090	A08N074	TD	Aab22548	Yes	76.9	6->6
IPMI1	AT3G58990	A11N226		Aab13092	Yes	83	3->3
[IPMI-like]	AT2G43090	A11N226	TD	Aab19619	Yes	85.99	3->3
BCAT3	AT3G49680	A19N002		Aab33782	Yes	76.02	6->6
[BCAT5]	AT5G65780	A19N002	TD	Aab23605	Yes	75.3	6->6
BAT5	AT4G12030	A20N095		Aab32285	Yes	76.21	2->2
[BAT-like]	AT4G22840	A20N095		Aab23321	Yes	91.82	2->2
TF - regulation							
OBP2	AT1G07640	A02N142		Aab18330	Yes	80.28	2->2
[OBP-like]	AT2G28810	A02N142		Aab24559	Yes	70.03	2->2
MYB122	AT1G74080	A05N185		Aab14276	Yes	57.56	6->4
MYB51	AT1G18570	A05N185		Aab19683	Yes	59.89	6->4
IQD1	AT3G09710	A14N046		Aab18852	Yes	65.59	2->2
[IQD2]	AT5G03040	A14N046		Aab18368	Yes	77.39	2->2
MYB28	AT5G61420	A26N034		Aab12163	Yes	67.39	6->4
MYB29	AT5G07690	A26N034	TD	Aab33585	Yes	65.13	6->4
	36% TD (13/36) 14% GTD (5/36)			31% TD (11/35) 14% GTD (5/35)		Ø 76.58	

[A] Squared brackets indicate ohnolog copies of GS biosynthetic genes without GO!-annotation to GS biosynthetic process

[B] SSD refers to short sequence duplication, TD refers to members of tandem arrays (TARs) and GTD refers to the history of transposition in *Arabidopsis*

[C] Predicted *Aethionema* CDS

[D] Change of gene family size in Col-0 -> Aab order

TABLE 2.— Gene transposition duplicates (GTD) in *Arabidopsis* and *Aethionema* GS pathway inventory.

Protein Name	AGI[A]	α-Block	AabID[B]	Syntelog	% identity	Lineage specific	Col-0 -> Aab[C]
GS genes with retained α-ohnolog							
[AAO3]	AT2G27150	A02NOA1	Aab27016	Yes	77.67	both	2->2
CYP79C1	AT1G79370	A05N062	Aab34143	Yes	76.8	both	8->9
APK2	AT4G39940	A10NOA2	Aab17804	Yes	86.05	both	2->2
CYP79B2	AT4G39950	A10N257	Aab17805	Yes	81.38	both	8->9
CYP79B3	AT2G22330	A10N257	Aab19477	Yes	81.4	both	8->9
GS genes with tandem duplicate copy							
AOP1	AT4G03070	A01	Aab37231	Yes	70.03	both	2->1
AOP3	AT4G03050	A01	-	-	-	*Arabidopsis*	2->1
CYP79C2	AT1G58260	A03	Aab17711	Yes	71.85	*Aethionema*	8->9
		A11	Aab22600	No	61.8	*Aethionema*	8->9
CYP83A1	AT4G13770	A15	Aab32506	Yes	69.67	both	2->3
		-	Aab30975	No	82.8	*Aethionema*	2->3
CYP81F2	AT5G57220	A22	-	-	-	*Arabidopsis*	2->1
GS genes without α-ohnolog or tandem duplicate copy							
UGT74B1	AT1G24100*	A05	Aab07827	Yes	80.65	both	6->10
		A05	Aab07826	Yes	70.35	none	6->10
FMO-GSOX-1	AT1G65860*	A25	Aab30109	Yes (minimum)	58.45	both	11->8
CYP79A2	AT5G05260*	A14	Aab36760	Yes	73.37	both	8->9
CHY1	AT5G65940*	A19	Aab05851	Yes	80.75	both	2->2
IMD1	AT5G14200	A12	Aab14760	Yes	89.5	both	2->1
IMD3	AT1G31180	A06	-	-	-	*Arabidopsis*	2->1
CYP83B1	AT4G31500	A10	Aab12019	No	92.73	both	2->3
GSL-OH	AT2G25450	A10	-	-	-	*Arabidopsis*	1->0
	24% TD (4 / 17) 29% retained α-ohnolog (5/17)		13% TD (2 / 16) 31% retained α-ohnolog (5/16)		Ø 76.78%		

[A]Asteriks mark GTDs infered by flanking TE-like sequences using GEvo

[B]Change of gene family size in Col-0 -> Aab order

[C]Predicted *Aethionema* CDS

Glucosinolate biosynthetic gene identification from draft *Ae. arabicum* genome

On the basis of the *Ae. arabicum* genome v1.0 and 37,839 annotated genes [43], we identified homologs of *A. thaliana* GS biosynthetic and regulatory genes. Combining reciprocal best BLASTP hits with LAST screens for large scale gene collinearity/synteny (100 kb–1.2 Mb) (employed by the Synfind algorithm, see Materials & Methods section), we found putative *Ae. arabicum* orthologs covering 57 of the 64 proposed AtGS genes with an observed nucleotide sequence identity of 45–94% (**tables 1, 3-4**). Among those, seven loci gave rise to 10 further paralogs due to TD and GTD in *Aethionema*, thereby extending the copy number of six multigene families to a total of 67 putative AabGS genes. The mRNA sequencing data for *Ae. arabicum* supported the evidence that all 67 putative AabGS genes were expressed (data not shown).

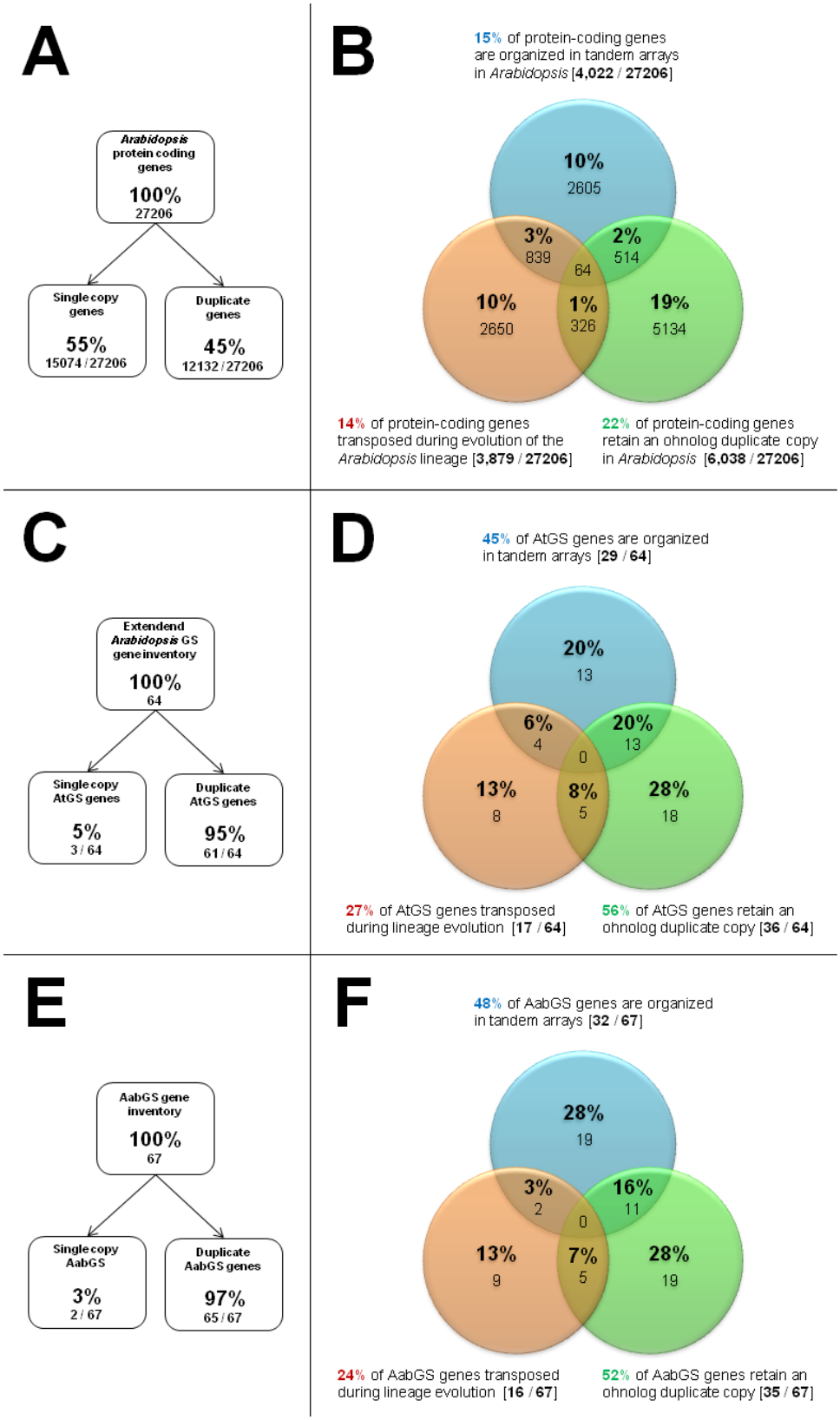


FIG. 2.— Duplicate distribution among (A, B) *Arabidopsis* protein-coding genes compared with (C, D) AtGS and (E, F) Aethionema GS loci. Shown are retained ohnologs (green), tandem duplicates (blue), and gene transposition duplicates (orange). GS metabolic versatility resulted from a combination of increased ohnolog retention and TD rates.

TABLE 3.- Genes with non-retained At-α ohnolog duplicate gene copy in *Arabidopsis* and *Aethionema* GS pathway inventory

Protein Name	AGI	α-Block	Evident SSD[A]	AabID[B]	Syntelog	%identity	Col-0 -> Aab[C]
Core-structure formation							
AOP1	AT4G03070	A01	TD / GTD	Aab37231	Yes	70.03	2->1
AOP3	AT4G03050	A01	TD / GTD	-	-	-	2->1
UGT74-like	Aab specific	A02	TD	Aab37178	Yes	82.05	6->10
UGT74-like	Aab specific	A02	TD	Aab37179	Yes	77.63	6->10
UGT74-like	Aab specific	A02	TD	Aab37180	Yes	78.33	6->10
GSTF10	AT2G30870	A02	TD	Aab28612	Yes	91.59	4->4
FMO-GSOX-3	AT1G62560	A03	TD	Aab10867	Yes	71.9	11->8
FMO-GSOX-4	AT1G62570	A03	TD	Aab10866	Yes	55.2	11->8
FMO-GSOX-5	AT1G12140	A03	TD	Aab13546	Yes	71.9	11->8
CYP79C2	AT1G58260	A03	TD	Aab17711	Yes	71.85	8->9
	Aab specific	A11		Aab22600	No	61.8	8->9
CYP79F1	AT1G16410	A05	TD	Aab27579	Yes	72.79	8->9
SOT18	AT1G74090	A05	TD	Aab14277	Yes	83.9	3->3
GSTU20	AT1G78370	A05	TD	Aab06999	Yes	67.29	5->6
	Aab specific			Aab6995	Yes	48.86	5->6
UGT74B1	AT1G24100	A05	GTD	Aab07826	Yes	70.35	6->10
	Aab specific			Aab07827	Yes	80.65	6->10
CYP83B1	AT4G31500	A10	GTD	Aab12019	No	92.73	2->3
GSL-OH	AT2G25450	A10	GTD	-	-	-	1->0
CYP79A2	AT5G05260	A14	GTD	Aab36760	Yes	73.37	8->9
CYP83A1	AT4G13770	A15	GTD	Aab32506	Yes	69.67	2->3
	Aab specific			Aab30975	No	82.8	2->3
CYP81F2	AT5G57220	A22	TD / GTD	-	-	-	2->1
FMO-GSOX-1	AT1G65860	A25	GTD	Aab30109	Yes	58.45	11->8
Co-substrate pathways							
CHY1	AT5G65940	A19	GTD	Aab05851	Yes	80.75	2->2
GSH1	AT4G23100	A20	N/A	Aab22781	Yes	91.81	2->2
BZO1	AT1G65880	A25	TD	Aab31601	Yes	70.04	2->4
			TD	Aab31602	Yes	69.4	2->4
Side-chain elongation							
IMD1	AT5G14200	A12	GTD	Aab14760	Yes	89.5	2->1
IMD3	AT1G31180	A06	GTD	-	-	-	2->1
IPMI2	AT2G43100	A11	TD	Aab19630	Yes	78.71	3->3
IIL1	AT4G13430	A15	N/A	Aab18132	Yes	93.9	1->1
TF-regulation							
MYB76	AT5G07700	A26	TD	-	-	-	6->4
	60% TD (15 / 25)			57% TD (16 / 28)		Ø 76.46%	
	48% GTD (12 / 25)			39% GTD (11/ 28)			

Note.-NA, not applicable; SSD, short sequence duplication

[A] TD (tandem duplicates) refers to members of TARs and GTD (gene transposition duplicates) refers to the history of transposition in *Arabidopsis*

[B] Predicted *Aethionema* CDS

[C] Change of gene family size in Col-0 -> Aab order

TABLE 4.— Genes not covered by α-blocks in *Arabidopsis* and *Aethionema* GS pathway inventory.

Protein Name	AGI	α-Block	Evident SSD[A]	AabID[B]	Syntelog	% identity	Col-0 -> Aab[C]
Side-chain elongation							
MAM1	AT5G23010	-	TD	Aab12229	Yes	72.31	2 -> 4
				Aab12230	Yes	71.5	2 -> 4
MAM-L	AT5G23020	-	TD	Aab12225	Yes	70.67	2 -> 4
				Aab12226	Yes	68.36	2 -> 4
TF - regulation							
MYB34	AT5G60890	-	N/A	-	-	-	6 -> 4
	66% TD (2/3)			100% TD (4/4)		Ø 70.71%	
	0% GTD			0% GTD			

Note.-NA, not applicable

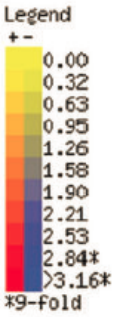
[A] SSD refers to short sequence duplication and TD (tandem duplicates) refers to members of TARs (tandem arrays)

[B] Predicted *Aethionema* CDS

[C] Change of gene family size in Col-0 -> Aab order

TABLE 5 .- Intraspecies protein similarities for At-α ohnolog pairs sharing GS annotation, shown with differential expression in *Arabidopsis* following MeJA treatment.

α-Block	Protein Name	GO[A]	AGI	Similarity in Col-0	AabID[B]	Similarity in Aab (%)	Expression change[C]		
Both ohnologs with annotated to GS biosynthesis							Time ->	0.5	1	3
A05N062	CYP79C1	Yes	AT1G79370	60.90%	Aab34143	N/A		0	0.3	0.5
	CYP79F2	Yes	AT1G16400[2]		-			0.8	0.3	1.1
A05N185	MYB122	Yes	AT1G74080	68.80%	Aab14276	61.50%		0.6	3.1	0.7
	MYB51	Yes	AT1G18570		Aab19683			-0.8	-0.6	0
A05N186	SOT17	Yes	AT1G18590	83.50%	Aab19675	83.20%		1	1.1	0.8
	SOT16	Yes	AT1G74100		Aab14278			0.8	1.2	1.4
A10NOA1	APK1	Yes	AT2G14750	67.50%	Aab32150	62.10%		0.7	1.4	1.6
	APK2	Yes	AT4G39940		Aab17804			1.3	1.9	1.6
A10N257	CYP79B3	Yes	AT2G22330	92.10%	Aab19477	29.50%		0.9	2	2.6
	CYP79B2	Yes	AT4G39950		Aab17805			0.5	1.4	2.2
A26N034	MYB29	Yes	AT5G07690	72.30%	Aab33585	66.00%		0.1	0.2	-0.6
	MYB28	Yes	AT5G61420		Aab12163			-0.2	-0.3	-0.7
				Ø 74.18%		Ø 60.46%				



Note.-NA, not applicable; MeJa, Methyl-jasmonic acid

[A] GO!-column indicates if genes is annotated to canonical GS pathway inventory as defined by (Sonderby, et al. 2010)

[B] Predicted *Aethionema* CDS

[C] Whole wild-type plant averages of log-transferred expression change according to ATH1microarray data

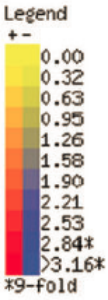
Glucosinolate gene families with expanded copy number in *Aethionema*

Among the 10 novel paralogs, eight were identified as descendants from TD events. Intriguingly, the *Arabidopsis* methyl-thiomalate synthase array *MAM1/MAM-L* underwent a further duplication in *Aethionema*, retaining four *MAM*-like loci (**supplementary fig. S1**, Supplementary Material online). Likewise, we observed a further TD of the *Arabidopsis* benzoate-CoA ligase array *BZO1/BZO*-like in *Aethionema*, adding two paralogs to the set of putative AabGS genes (**table 8**). Notably, both clusters encode functions in GS side-chain elongation (MAM) or co-substrate pathways (BZO), indicating the connection of TD to metabolic versatility in both *Arabidopsis* and *Aethionema*. Furthermore, TD extended the gene inventory of GS core-structure modification in two cases. First, we detected one additional duplicate of the *Arabidopsis* tau-type gluthation-s-transferase array *GSTU19-GSTU23* in *Aethionema* (table 8). Second, we identified extension of the *UGT*-like superfamily that is present with five members in *Arabidopsis* and organized in two TARs of distant genomic location (**fig. 3B**). Intriguingly, both regions represent the sister copies of α-block a02 (**fig. 3F**). Furthermore, both *UGT*-like TARs comprise neighboring pairs of the At-α ohnologs duplicates A02N051 (*UGT74C1* or AT2G31790/ *UGT*-like or AT1G05670) and A02N053 (*UGT74D1* or AT2G31750/*UGT74E2* or AT1G05680) (**fig. 3B, table 8**), indicating a pre-At-α TD event generating both precursors of the above-mentioned *UGT*-like ohnolog pairs. In *Aethionema*, we found a further TD-driven extension of this superfamily, adding three more copies to reach a total number of 8 *UGT*-like sequences (**table 8**). Therefore, the diversity of *UGT*-like sequences in Brassicaceae is expanded by the combination of WGD with pre- and post-At-α TD events. In contrast, GTD accounts for the copy number expansions of two putative AabGS loci. Both cases involve *CYP*-like genes that play a role in GS core-structure formation [120]. In *Aethionema*, the TAR formed by (1) *CYP79C2* (At1G85260) and (2) the *CYP*-like locus AT1G58265 transposed an additional copy of the TAR to a different genomic location (**supplementary fig. S2**, Supplementary Material online). Likewise, we identified an additional GTD of *CYP83A1* in *Aethionema* (**table 3**). CYP83A1 metabolizes oximes in GS

biosynthesis, is not redundant to CYP83B1, and interestingly also possesses a history of GTD events in *Arabidopsis* [126].

TABLE 6.- Intras-pecies protein similarities for At-α ohnolog pairs not sharing GS annotation, shown with differential expression in *Arabidopsis* following MeJA treatment.

α-Block	Protein Name[A]	GO![B]	AGI	Similarity in Co	AabID[C]	Similarity in Aab	Expression change[D]		
One hnologs with out annottation to GS biosynthesis							Time ->	0.5	1	3
A02NOA2	AAO4	Yes	AT1G04580	84.10%	Aab24896	79.90%		17	-0.2	-0.4
	[AAO3]	No	AT2G27150		Aab27016			0.8	-0.2	-0.1
A02N051	UGT74C1	Yes	AT2G31790	22.30%	Aab37175	19.50%		0.4	0.4	0.3
	[UGT-like]	No	AT1G05670		Aab31930			0.3	0.1	0
A02N142	OBP2	Yes	AT1G07640	68.50%	Aab18330	34.30%		-0.1	0.1	-0.2
	[OBP-like]	No	AT2G28810		Aab24559			-0.1	-0.4	-0.2
A03N117	FMO-GSOX-2	Yes	AT1G62540	75.90%	Aab10869	70.60%		-0.3	0.3	0.4
	[FMO-like]	No	AT1G12130		Aab13543			-0.8	-0.6	2.1
A08N074	BCAT4	Yes	AT3G19710	72.20%	Aab21007	76.80%		0.1	0.2	0.5
	[BCAT7]	No	AT1G50090		Aab22550			0.4	0	0.7
A10N194	SUR1	Yes	AT2G20610	84.00%	Aab31155	69.90%		0.6	1	0.8
	[SUR-like]	No	AT4G28420		Aab30136			0.9	0.6	-2.5
A10N314	GGP1	Yes	AT4G30530	84.90%	Aab24374	83.40%		0.8	1	1.4
	[GGP-like]	No	AT2G23960		Aab11021			-0.2	-0.7	-0.7
A11N226	IPMI1	Yes	AT3G58990	80.60%	Aab13092	73.20%		0.2	0.6	0.9
	[IPMI-like]	No	AT2G43090		Aab19619			0	0.1	0.2
A12N102	GSTF11	Yes	AT3G03190	83.60%	Aab14996	41.80%		1.4	1.7	1.4
	[GSTF12]	No	AT5G17220		Aab14791			-0.5	0.2	1.1
A14N046	IQD1	Yes	AT3G09710	67.10%	Aab18852	52.60%		-0.2	-0.3	-0.1
	[IQD2]	No	AT5G03040		Aab18368			-0.2	-0.2	-0.4
A19N002	BCAT3	Yes	AT3G49680	8.20%	Aab33782	79.90%		0	-0.1	0
	[BCAT5]	No	AT5G65780		Aab23605			-0.1	-0.2	0
A20N095	BAT5	Yes	AT4G12030	78.90%	Aab32285	63.30%		0.4	-0.1	0.4
	[BAT-like]	No	AT4G22840		Aab23321			0	0	0.6
				Ø 67.52%		Ø 62.1%				



Note.-NA, not applicable; MeJa, Methyl-jasmonic acid

[A] Squared brackets indicate ohnolog copies of GS biosynthetic genes without GO!-annotation to GS biosynthetic process

[B] GO!-column indicates if genes is annotated to canonical GS pathway inventory as defined by (Sonderby, et al. 2010)

[C] Predicted *Aethionema* CDS

[D] Whole wild-type plant averages of log-transferred expression change according to ATH1 microarray data

TABLE 7. - Putative single-copy gene in *Aethionema* and *Arabidopsis* GS biosynthetic inventory

AGI	Name	α-Block	Retained At-ß / -γ ohnolog	Most ancient syntelog	Closest paralog	BLASTP E-value	Name	α-Block	Retained At-ß / -γ ohnolog
AT4G23100	GSH1	A20	No	*B. rapa*	AT1G19220[B]	0.19	ARF11	A05	No
AT4G13430	IIL1	A15	No	*V. vinifera*	AT4G26970	1.00E-16	ACO2	A22N121	No
AT5G60890[A]	MYB34	-	B20N001	*V. vinifera*	AT1G74080	4.00E-62	MYB122	A05N185	B20N004

[A] Absent in *Aethionema*

[B] gene transposition duplicate copy

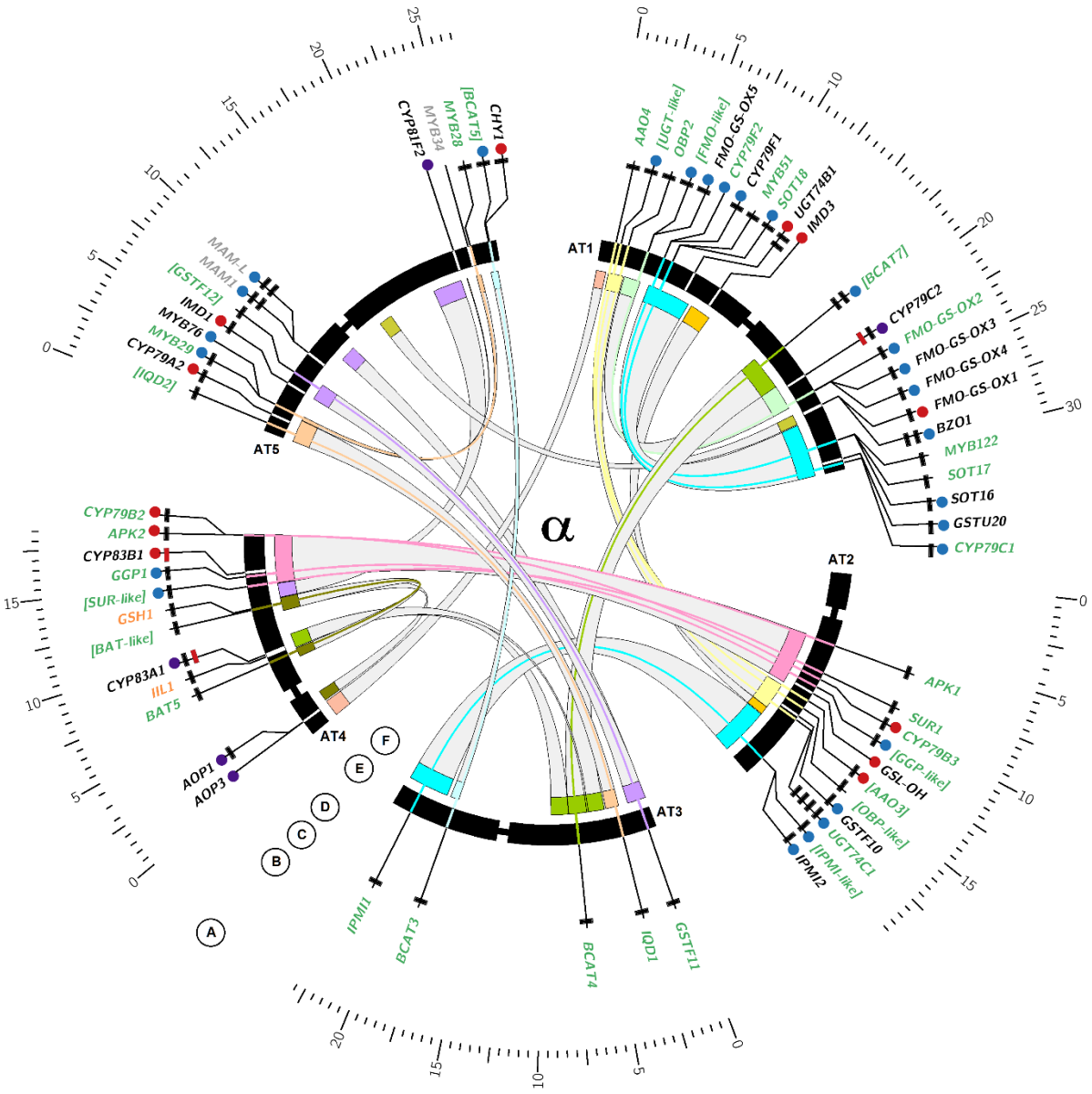


FIG. 3.— Ideogram of *Arabidopsis thaliana* chromosomes with GS biosynthetic genes. Circos plot visualizing the evolutionary contribution of different duplication types to GS pathway inventory in *Arabidopsis* and *Aethionema*. **A**. Inner chromosome scale (Mb). **B**. *Arabidopsis thaliana* GS biosynthetic genes. Gray text indicates genomic location outside ohnolog blocks. Black text indicates genomic location within ohnolog blocks but non-retained ohnolog copy. Green text indicates retained pairs of ohnolog copies with missing GO annotation to GS biosynthetic process shown in edged brackets. Orange text indicates single copy genes without clear paralogs in both species. **C**. Blue circles indicate genes organized in TARs (i). Red circles indicate genes with transpositional history (ii). Purple circles indicate loci sharing (i) and (ii). **D**. Number of rectangles indicates number of homologs present in the *Aethionema arabicum* draft genome (0–4). Color of rectangles indicates presence (black) or absence (red) of synteny between *A. thaliana* and *Ae. arabicum* in the genomic context of the target gene. **E**. *Arabidopsis thaliana* chromosomes with labels showing GS biosynthetic genes. Bands for genes retained in ohnolog pairs are connected with colors of corresponding ohnolog blocks, as defined by [35]. **F**. Genomic location of ohnolog block copies harboring GS biosynthetic genes in *A. thaliana*, connected by gray bands. All ranges are in scale.

Arabidopsis* Glucosinolate loci without orthologs in *Aethionema

In seven cases, RBH- and synteny-based evidence was not sufficient to clearly assign orthologs to AtGS loci in the *Ae. arabicum* draft genome (**tables 1, 3-4**), leading to the contraction of six multigene families and loss of one single-copy gene. Four of those loci, namely *AOP3*, *CYP79F2*, *IMD3*, and *MYB76*, are likewise absent in the *B. rapa* genome and may therefore be specific to *A. thaliana* and more closely related species [70]. *AOP1/AOP3* and *CYP79F1/2* represent two neighboring TARs with evident sub-functionalization in *Arabidopsis* [94, 127]. While *AOP3* functions in GS side-chain elongation in *Arabidopsis* [94], *CYP79F2* encodes an enzyme involved in core structure formation of long-chain aliphatic GS. Furthermore, over-expression of the *MYB76* transcription factor correlates with increased levels of both long-chained and short-chained aliphatic GS in *Arabidopsis* [128]. However, experiments with *Arabidopsis* myb76 T-DNA insertion lines to date did not show any significant change in GS chemotype, making a strict requirement of *MYB76* for GS biosynthesis unlikely [128]. Moreover, *IMD3* encodes a predicted enzyme with proposed functional redundancy to (as well as strong co-expression with) *IMD1*, encoding a protein that was shown to be involved in GS accumulation in *Arabidopsis* [129-132]. Therefore, absence of *IMD1* (*IMD3*) in *B. rapa* (*Ae. arabicum*) supports the hypothesized capability of mutual phenotype rescue among *IMD1/3* double knock-outs in Brassicaceae, eventually preventing significant alterations of GS chemotype due to fractionation of *IMD*-like genes in *Aethionema*.

The other three of the seven AtGS loci that lack a clear ortholog in Aethionema are not found in the *B. rapa* genome: *MYB34*, *CYP81F2*, and *GSL-OH* (**tables 3, 4**). Therefore, they represent *Aethionema* lineage-specific gene losses. *MYB34* was shown to control indolic GS biosynthesis in *Arabidopsis* [133]. Interestingly, over-expression of *MYB34* in *Arabidopsis* partially rescued the altered GS chemotype caused by *MYB51* knockout [134]. Because of functional redundancy of *MYB51/34*, the loss of *MYB34* likely does not cause significant changes in GS chemotype in *Aethionema*. In contrast, *CYP81F2* and *GSL-OH* encode functions associated with secondary modification of the GS core structures [120]. *CYP81F2* has been shown to control a quantitative trait loci for indole GS modification in *Arabidopsis*, catalyzing the conversion of indole-3-yl-methyl GS to 4-hydroxy-indole-3-yl-methyl GS [135]. Notably, these metabolites play a significant ant role in MAMP-triggered immunity in *Arabidopsis* [136]. Among various known cytochrome p450s active in indolic GS biosynthesis, *CYP81F2* is the only locus impairing callose deposition after detection of the non-self infection in *Arabidopsis* [137]. On the basis of these findings, we concluded a *CYP81F2*-specific onset of sub/neofunctionalization from GS biosynthesis toward plant innate immunity after divergence of the *Arabidopsis* and *Aethionema* lineages, thereby mitigating fatal consequences of the absent ortholog in *Aethionema.* Moreover, we determined the absence of the 2-oxoacid-dependend oxygenase activity GSL-OH in *Aethionema* (**table 3**). GSL-OH is necessary for biosynthesis of 2-hydroxy-but-3-enyl GS in *Arabidopsis* [130, 138] and present in the *B. rapa* genome [70]. Noteworthy, *GSL-OH* was the only multigene family member within the extended AtGS set whose loss in *Aethionema* was not accompanied by copy number expansion of one or more paralogs (**table 3**). Accordingly, we concluded an *Aethionema*-specific loss of this locus after divergence from the *Arabidopsis* lineage, creating measurable differences in GS chemotype among *Arabidopsis* and *Aethionema*. Consistently, we could not detect traces of 2-hydroxybut-3-enyl GSs in *Ae. arabicum* root-, leaf-, and seed extract using uHPLC (data not shown).

TABLE 8.-Tandem duplicate genes in *Arabidopsis* and *Aethionema* GS pathway inventory. Underlined Duplicates are pre-α.

Protein Name[A,B]	AGI	α-Block[C]	AabID[D]	Syntelog	% identity[E]	Lineage specific	Col-0 -> Aab[F]
GS genes with retained α-ohnolog							
UGT74C1	AT2G31790	A02N051	Aab37175	Yes	79.44	No	6->10
UGT74D1_oa	AT2G31750	A02N053	Aab37181	Yes	78.95	not considered	6->10
			Aab37178	Yes	82.05	*Aethionema*	6->10
			Aab37179	Yes	77.63	*Aethionema*	6->10
			Aab37180	Yes	78.33	*Aethionema*	6->10
[UGT-like]	AT1G05670	A02N051	Aab31930	Yes	81.11	No	6->10
UGT-like_oa	AT1G05675	A02	Aab31932	Yes	71.4	not considered	6->10
UGT74E2_oa	AT1G05680	A02N053	Aab31933	Yes	59.8	not considered	6->10
FMO-GSOX-2	AT1G62540	A03N117	Aab10869	Yes	76.6	No	11->8
FMO-GSOX-3	AT1G62560	A03	Aab10867	Yes	71.9	No	11->8
FMO-GSOX-4	AT1G62570	A03	Aab10866	Yes	55.2	No	11->8
FMO-like_oa	AT1G62580	A03	-	-	-	not considered	10->7
FMO-like_oa	AT1G62600	A03	-	-	-	not considered	10->7
FMO-like_oa	AT1G62620	A03	-	-	-	not considered	10->7
[FMO-like]	AT1G12130	A03N117	Aab13543	Yes	65.09	No	11->8
FMO-GSOX-5	AT1G12140	A03	Aab13546	Yes	71.9	No	11->8
FMO-like_oa	AT1G12200	A03	Aab13549	Yes	66.66	not considered	10->7
CYP79F2	AT1G16400	A05N062	-	-	-	*Arabidopsis*	8->9
CYP79F1	AT1G16410	A05	Aab27579	Yes	72.79	*Arabidopsis*	8->9
SOT16	AT1G74100	A05N186	Aab14278	Yes	91.07	No	3->3
SOT18	AT1G74090	A05	Aab14277	Yes	83.9	No	3->3
GSTU20	AT1G78370	A05	Aab06999	Yes	67.29	No	5->6
GSTU23_oa	AT1G78320	A05	Aab06994	Yes	81.74	not considered	5->6
GSTU22_oa	AT1G78340	A05	Aab06997	Yes	71.1	not considered	5->6
GSTU21_oa	AT1G78360	A05	Aab06998	Yes	76.71	not considered	5->6
			Aab06995	Yes	48.86	*Aethionema*	5->6
GSTU19_oa	AT1G78380	A05N104	Aab07000	Yes	83.41	not considered	5->6
[BCAT7]	AT1G50090	A08N074	Aab22550	Yes	69	No	6->6
BCAT-like_oa	AT1G50110	A08	Aab22548	Yes	78.12	not considered	6->6
[SUR-like]	AT4G28420	A10N194	Aab31154	Yes	47.42	No	3->2
SUR-like_oa	AT4G28410	A10	Aab31155	Yes	63.96	not considered	3->2
GGP1	AT4G30530	A10N314	Aab24374	Yes	87.6	No	5->5
GGP-like_oa	AT4G30540	A10	Aab24373	Yes	75	not considered	5->5
GGP3_oa	AT4G30550	A10	Aab24372	Yes	82	not considered	5->5
[GGP-like]	AT2G23960	A10N314	Aab11018	Yes	69.67	No	5->5
GGP-like _oa	AT2G23970	A10	Aab11021	Yes	83.6	not considered	5->5
[IPMI-like]	AT2G43090	A11N226	Aab19619	Yes	85.99	No	3->3
IPMI2	AT2G43100	A11	Aab19630	Yes	78.71	No	3->3
[BCAT5]	AT5G65780	A19N002	Aab23605	Yes	75.3	No	6->6
LINC4_oa	AT5G65770	A19	Aab23607	Yes	70.08	not considered	6->6
MYB29	AT5G07690	A26N034	Aab33585	Yes	65.13	*Arabidopsis*	6->4
MYB76	AT5G07700	A26	-	-	-	*Arabidopsis*	6->4
GS genes with non-retained α-ohnolog							
AOP1	AT4G03070	A01	Aab37231	Yes	70.03	*Arabidopsis*	2->1
AOP3	AT4G03050	A01	-	-	-	*Arabidopsis*	2->1
GSTF10	AT2G30870	A02	Aab28612	Yes	91.59	No	4->4
GSTF9_oa	AT2G30860	A02	Aab28613	Yes	89.76	not considered	4->4
UGT74B1	AT1G24100	A05	Aab07827	Yes	80.65	*Aethionema*	6->10
		A05	Aab07826	Yes	70.35	*Aethionema*	6->10
CYP79C2	AT1G58260	A03	Aab17711	Yes	71.85	No	8->9
CYP-like_oa	AT1G58265	A03	Aab17712	Yes	60.71	not considered	8->9
CYP81F2	AT5G57220	A22	-	-	-	*Arabidopsis*	2->1
CYP71B10 _oa	AT5G57260	A22	Aab25774	Yes	73.21	not considered	2->1
BZO1	AT1G65880	A25	Aab31601	Yes	70.04	No	2->4
			Aab31602	Yes	69.4	*Aethionema*	2->4
BZO-like_oa	AT1G65890	A25	Aab31603	Yes	68.85	not considered	2->4
			Aab31604	Yes	67.83	not considered	2->4

(continued)

TABLE 8.— Continued

Protein Name[A,B]	AGI	α-Block[C]	AabID[D]	Syntelog	% identity[E]	Lineage specific	Col-0 -> Aab[F]
GS genes outside α-blocks							
MAM1	AT5G23010	-	Aab12229	Yes	72.31	No *Aethionema*	2->4
			Aab12230	Yes	71.5		2->4
MAM-L	AT5G23020	-	Aab12225	Yes	70.67	No *Aethionema*	2->4
			Aab12226	Yes	68.36		2->4
	AtGS genes: 45% TD (29/64)		AabGS genes: 46% TD (31/67)		Ø 73.43%		

[A] Square brackets indicate ohnolog copies of GS biosynthetic genes without GO-annotation to GS biosynthetic process.

[B] The "_oa" suffix indicates tandem duplicate copies of GS biosynthetic genes without GO!-annotation to GS biosynthetic process.

[C] Underlined line-items refer to offspring of one pre- α tandem duplication event.

[D] Predicted *Aethionema* CDS

[E] In case of Aethionema-specific TAR expansion, the corresponding *Arabidopsis* sequence for identity comparison was determined based on both genomic location and homology criteria
homology criteria.

[F] Change of gene family size in Col-0 -> Aab order

Glucosinolate gene families with lower copy number in *Aethionema*

Considering the *Aethionema*-specific loss of *GSL-OH*, six putative GS-annotated multigene families display a lower copy number in *Aethionema* (*AOP*x, *CYP79*x, *CYP81*x, *GSL*x, *IMD*x, and *MYB*x) (**tables 1-4, 8**). In sum, their total gene count increased from 20 genes in *Aethionema* to 27 observed in *Arabidopsis*, thereby mediating a 35% increase. Although *IMD3* and *GSL-OH* possess GTD copies in *Arabidopsis*, these loci are absent in *Aethionema*. In contrast, *AOP1/2*, *IMD1/3*, *MYB29/76*, and *CYP81F2* with its *CYP*-like neighbor AT1G58265 comprise TARs in *Arabidopsis* but are likewise absent in *Aethionema* (**table 8**) and *B. rapa* [70]. Therefore, the underlying TD events may be *Arabidopsis*-specific. Thus, TD facilitated GS pathway expansion in the *Arabidopsis* lineage after split from the tribe Aethionemeae. Furthermore, we found evidence for *Arabidopsis*-specific TD of three neighboring *FMO*-like loci (*FMO GS-OX*$_{2\text{-}4}$) (**figs. 3-4**), leading to lower copy number in *Aethionema*. These genes were lost in *B. rapa*[70], illustrating a degree of gene and genome plasticity across Brassicaceae. *FMO*-like loci comprise a multigene family with five members annotated to GS biosynthesis in *Arabidopsis* mapping to three distant genomic locations (**fig. 3B**). Among those, two regions are embedded in ohnolog copies of α-block A02 (**fig. 3E**) and contain the retained α-pair of AT1G62540 (*FMO GS-OX2*) and AT1G12130 (*FMO*-like) (fig. 3B). The latter is not annotated to GS biosynthesis in *Arabidopsis*. However, AT1G12130 is member of a *FMO*-like four-gene TAR with its 3'-neighbor *FMO GS-OX5* involved in aliphatic GS biosynthesis [95]. The third genomic region in *Arabidopsis* harboring a *FMO*-like sequence with encoded function in GS metabolism is defined by AT1G65860 (*FMO GS-OX1*), representing a transposed duplicate gene copy (**fig. 4**). Interestingly, *FMO GS-OX*$_{1\text{-}4}$ share broad substrate specificity and catalyze the conversion from methyl-thioalkyl GS to the related methyl-sulfinyl GS independent of chain length. In contrast, *FMO GS-OX5* shows substrate specificity for 8-methyl-thiooctyl GS [95, 139]. This example is similar to the case of *UGT*-like loci (see above) and again illustrates the combination of ohnolog retention with tandem- and GTD leading to increased GS pathway versatility in Brassicaceae (**fig. 4**).

Deduction of total duplicate frequencies in *Aethionema* and *Arabidopsis* GS pathway inventory and comparison to *Arabidopsis* genome-wide average

We found the fraction of retained ohnolog duplicate gene pairs among *Arabidopsis* (56%) and *Aethionema* (52%) GS biosynthetic and regulatory genes significantly increased compared with the genome-wide average in *Arabidopsis* (22%) (**fig. 2, table 9**). Moreover, 46% (31/67) of AabGS genes

are organized in TARs (**fig. 2F**), compared with a 22% average for all protein-coding genes in *Arabidopsis*, thereby significantly surpassing the TAR coverage rate of 45% (29/64) observed for AtGS loci (**fig. 2, table 9**). For duplication by gene transposition, we detected 27% (17/67) of affected AabGS loci (**fig. 2**). In summary, we found no significant enrichment of GTD events among GS pathway inventory in both species (**table 9**).

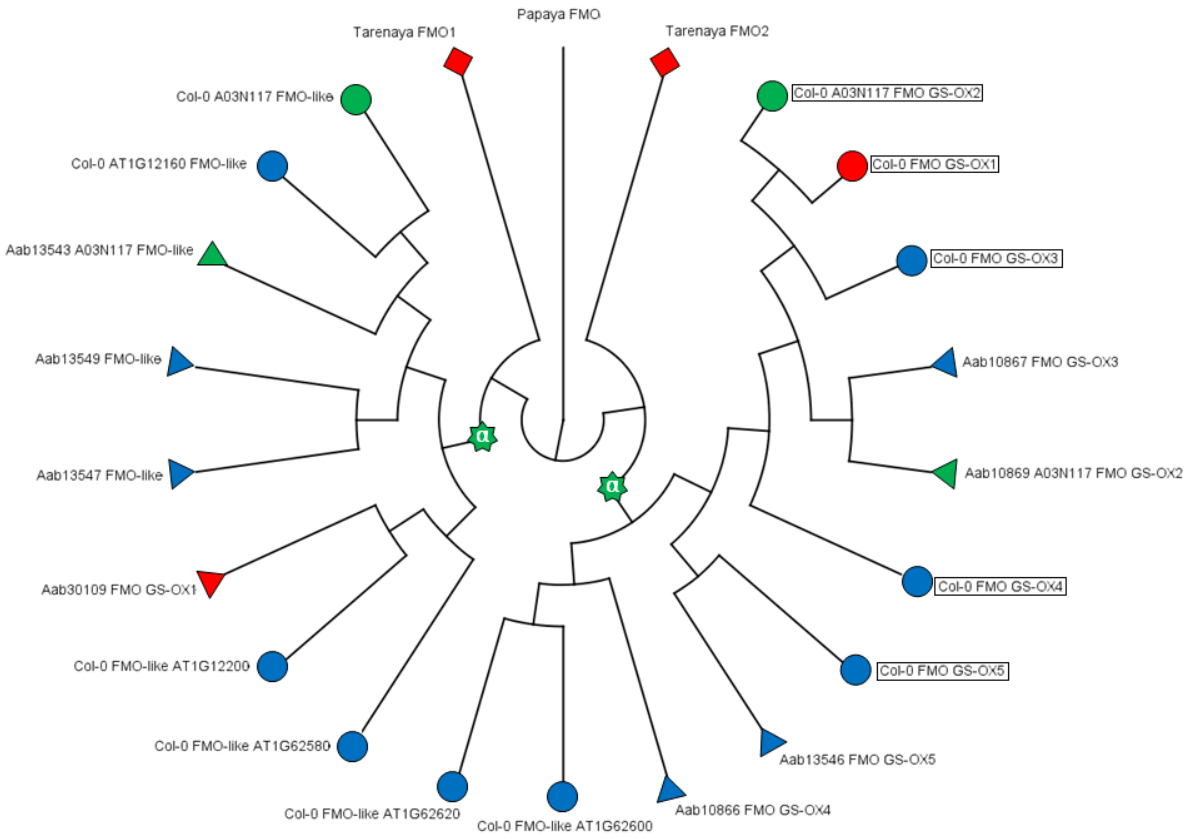


FIG. 4.— Phylogenetic relationships among FMO proteins. Col-0, Aab, and papaya refer to *Arabidopsis thaliana* (circles), *Aethionema arabicum* (triangles), and *Carica papaya* (colorless), respectively. Boxes indicate annotation to GS metabolic activity in *Arabidopsis*. *Tarenaya hasslerania* (diamonds) represents the closest-related outgroup of Brassicaceae, including the sister clade Aethionemeae. Stars: At-α WGD event. Blue: proteins encoded by members of TARs. Red: protein encoded by GTD locus. The At-α WGD leads to duplication of a *FMO* locus, resulting in two clades comprising all *FMO*-like sequences of both *Aethionema* and *Arabidopsis*. Thus, *FMO* versatility has been promoted by a combination of an increased degree of ohnolog retention and TD events.

TABLE 9.— Statistical test[A] on duplicate fractions in *Arabidopsis* and *Aethionema* GS pathway inventory compared to genome-wide average in *Arbidopsis*.

	Arabidopsis genome	*Arabidopsis* GS genes[B]	*Aethionema* GS genes
Protein-coding genes	27206	64	67
Tandem Duplicates	4022 / 15%	29 / 45%	32/ 48%
P-value	->	5.71E-09	1.77E-10
Gene Transposition Duplicates	3879 / 14%	17 / 27%	16 / 24%
P-value	->	0.04384	0.03407
Retained At-α ohnologs	6038 / 22%	36 / 56%	35 / 52%
P-value	->	3.87E-09	1.26E-07
Sum duplicates	12132 / 45%	61 / 95%	65 / 97%
P-value	->	2.20E-16	2.20E-16

[A] Fisher's exact test on count data
[B] extended set (see fig. 1)

DISCUSSION

The Aethionemeae/Brassicaceae crown-group/sister-group lineages split about 30–60 MA shortly after the last common WGD event and independently evolved ever since [75, 83, 118]. Notably, the radiation process evident for the Brassicaceae lineage created about 3,700 species [118]. In contrast, the species-poor *Aethionema* lineage [84] is well established as most ancient Brassicaceae extant sister and may therefore possess a more "ancient" genome organization when compared with *Arabidopsis*. This facilitates the recognition and quantification of common factors underlying rapid innovation of complex traits shared by both species. We exploit novel genomics resources for evolutionary analysis of the complete GS pathway inventory in both *A. thaliana* and *Ae. arabicum* to utilize the impact of different kinds of duplication classes to diversification of plant secondary metabolites. In a comparative genomics approach, we employ the phylogenetic relationship of *Ae. arabicum* and *A. thaliana* to identify key factors driving GS pathway divergence. In this context, we establish GS genetics/genomics as a scaffold to incorporate further phenotypic data for better understanding the impact of duplication to rapid evolution of novel key traits. In *Arabidopsis*, several GS genes retained duplicate gene copies dating back to the last WGD event but lacking annotation to GS metabolic processes (**fig. 1**). Illustrating high degrees of protein similarities among these ohnolog copy pairs and/or similar responses in gene regulation following GS pathway induction (**tables 5, 6**), we identified 12 novel putative *Arabidopsis* genes associated to GS biosynthesis (**figs. 1, 3**). Given the fact that these loci remained unknown despite their putative relevance for an experimentally very well-studied trait-like GS biosynthesis, we highlight the importance of considering ohnolog copies when analyzing a plethora of other highly diverged multigene pathways (i.e., terpenoid biosynthesis). We thereby provided an easy-to-follow framework on how to use existing data on WGD in *Arabidopsis* to better understand the networks of functional redundancy, especially involving genes that are targeted for knock-out experiments in functional studies. Evolutionary analysis of homologous GS loci in *Arabidopsis* and *Aethionema* found a majority (all but two) comprising duplicate groups organized in multigene families (**figs. 2, 3**). This underlined the dominant role of duplication for creation and expansion of biochemical diversity in plant (secondary) metabolism.
Clear orthologs of seven *Arabidopsis* GS genes are absent in the *Aethionema* draft genome (due to three *Aethionema*-specific GS gene losses and four *Arabidopsis*-specific TDs). Evolution of 10 additional *Aethionema* paralogs (two due to gene transposition and eight due to TD events, **fig. 3**) lead to an almost 100% conserved GS pathway inventory across the crown group/sister group system.. This sheds light upon the relevance of genome plasticity for key trait maintenance despite of scattered gene losses. To test this hypothesis, we indicate the requirement of further research on additional multigene pathways in a deeper phylogenetic resolution. Identification of *Aethionema* GS gene homologs allowed confirming the increased frequency of duplicates in lineages that diverged more than 30–60 MA. The absence of lineage-specific polyploidy events in either species facilitated the comparative analysis of genes duplicated due to the common ancient WGD events (particularly At-α) as well as lineage-specific gene tandem and transposition duplications. Partitioning the duplicate genes set in GS pathway inventory revealed significant enrichments of retained At-α ohnologs and tandem duplicates (but not GTD events) in both species compared to the average observed for protein-coding genes in *Arabidopsis* (**table 2**). We therefore conclude that WGD and TD facilitated the early and continued evolution of GS biosynthesis in the mustard family. To our knowledge, this is the first study providing distinct indications on a genetics level for the connection of WGD to the emergence of key traits in planta.

Various duplicates of different GS gene families code for proteins encoding functions in consecutive steps of GS biosynthesis [94, 139]. Among GS biosynthetic and regulatory genes, pairs of retained At-α ohnolog duplicates in distant genomic location further expand to TARs (**fig. 3**). In *Arabidopsis*, the S-oxygenase activity FMO is provided by a pair of retained ohnologs on distant arms on chromosome 1 (**figs. 3, 4**). Both copies evolved further tandem duplicates with different substrate specificities [95]. Different groups of substrates are products of SOT-type sulfotransferases provided by another retained ohnolog pair on At1 with additional TD copies sharing annotation to GS production [35, 140] (**fig. 3**). The reaction delivering substrates for GS SOT-type sulfotransferases is catalyzed by UGT-type proteins, likewise encoded by a pair of retained ohnologs that evolved multiple tandem and gene transposition duplicates in both *Aethionema* and *Arabidopsis* (**fig. 3**). It is thus inferred that subfunctionalization of both TD and retained At-α ohnolog pairs caused functional diversification of GS biosynthetic and regulatory elements. Showing mutual influence of ohnolog retention and TD rate across a crown group–sister group system, we describe a complex network of gene duplication fostering the expansion of a composite trait, thereby contributing to the means of mutation and selection to create evolutionary innovation in a limited time-frame. Evidence for the model of evolution by gene duplication can be found in comparative GS pathway analysis. Thus, GS may provide a framework for investigating the expansion of complex traits.

ACKNOWLEDGEMENTS AND FUNDING INFORMATION

This work was funded by a Netherlands Organization for Scientific Research (NWO) VIDI and Ecogenomics grant (M.E.S.). I am grateful for the input and support of Erik van den Bergh, Nicole van Dam, Mike Freeling, Tom Mitchell-Olds, Benjamin Schweßinger, Cyril Zipfel and Martin Parniske. Likewise, I want to acknowledge the contributions of two anonymous reviewers.

SUPPLEMENTARY MATERIAL

Supplementary figures S1, S2 and supplementary table S1 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/, last accessed on December 13th, 2014).

Supplementary Fig. S1.— MAFFT-Dot plot visualizing synteny of *Arabidopsis* and *Aethionema MAM*-regions (marked in red / blue) including Tandem Duplication of the *Arabidopsis MAM1/MAM-L* array (green arrows) in *Aethionema* (black arrows).

Supplementary Fig. S2.— GEvo graphic of a BLASTZ five-way multiple alignement including a gene transposition duplicate copy of *CYP79C2* and the neighboring CYP-like element AT1G58265 (marked in red). At and Aab refer to *A. thaliana* and *Ae. arabicum*. F and R indicate forward and reverse strand. Depending on sequence pairs, synteny is highlighted in blue, black and dark green. 3rd lane: *A. thaliana* sequences on chromosome 1 marked in red. Orange regions indicate transposon-like sequences. 2nd lane: *Ae. arabicum* scaffold 4412, harboring one of 2 RBH (marked in bright green) to GS *CYP* query sequence (marked in red). Top lane: Genomic region on *A. thaliana* chromosome 2 displying synteny to *A. arabuicum* scaffold 4412. Lanes 4 and 5: neighboring *Aethionema* scaffolds displaying synteny to genomic context of *AtCYP79C2.* Two copies are present in *Aethionema* and tranposed in *Arabidopsis* after divergence of the lineages. This experiment can be reproduced following the GEvo link http://genomevolution.org/r/8rnv, last accessed on December 13th, 2014.

Supplementary Table S1.— Overview on doifferent uplication modes affecting all protein-coding genes in *A.thaliana*

Large-scale evolutionary analysis of genes and supergene clusters from terpenoid modular pathways provides insights into metabolic diversification in flowering plants

Johannes A. Hofberger[1,2], Aldana M. Ramirez[3], Xinguang Zhu[2], Harro J. Bouwmeester[1], Robert C. Schuurink[3] and M. Eric Schranz[1]*

[1] Biosystematics Group, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
[2] Chinese Academy of Sciences/Max Planck Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai 200031, PR China
[3] University of Amsterdam, Swammerdam Institute for Life Sciences, Science Park 904, 1098 XH Amsterdam, The Netherlands

ABSTRACT

An important component of plant evolution is the plethora of pathways producing more than 200,000 biochemically diverse specialized metabolites with pharmacological, nutritional and ecological significance. To unravel dynamics underlying metabolic diversification, it is of paramount importance to determine lineage-specific gene family expansion in a phylogenomics framework. However, robust functional annotation is often only available for core enzymes catalyzing committed reaction steps within few model systems. In a genome informatics approach, we extracted information from early-draft gene-space assemblies and non-redundant transcriptomes to identify protein families involved in isoprenoid biosynthesis. Isoprenoids comprise terpenoids with various roles in plant-environment interaction, such as pollinator attraction or pathogen defense. Combining lines of evidence provided by synteny, sequence homology and Hidden-Markov-Modelling, we screened 17 genomes including 12 major crops and found evidence for 1,904 proteins associated with terpenoid biosynthesis. Our terpenoid genes set contains evidence for 840 core terpene-synthases and 338 triterpene-specific synthases. We further identified 190 prenyltransferases, 39 isopentenyl-diphosphate isomerases as well as 278 and 219 proteins involved in mevalonate and methylerithrol pathways, respectively. Assessing the impact of gene and genome duplication to lineage-specific terpenoid pathway expansion, we illustrated key events underlying terpenoid metabolic diversification within 250 million years of flowering plant radiation. By quantifying Angiosperm-wide versatility and phylogenetic relationships of pleiotropic gene families in terpenoid modular pathways, our analysis offers significant insight into evolutionary dynamics underlying diversification of plant secondary metabolism. Furthermore, our data provide a blueprint for future efforts to identify and more rapidly clone terpenoid biosynthetic genes from any plant species.

KEYWORDS: systems biology, genome informatics, big data, comparative genomics, terpenoids, trichomes, functional diversification

***Author for Correspondence:**
M. Eric Schranz | Biosystematics Group | Wageningen University & Research Center | Wageningen, The Netherlands | Tel. +31(0)317-483160 | email: eric.schranz@wur.nl

INTRODUCTION

To elucidate the dynamics underlying metabolic diversification across multiple lineages, it is critical to identify and distinguish the complete set of orthologous and paralogous loci present within multiple genome annotations in a phylogenetic framework [1]. Two homologous genes are referred to as orthologs if they descend from one locus present in the common ancestor lineage and split due to speciation [2,3]. By definition, orthologous genes are embedded in chromosomal segments derived from the same ancestral genomic locus, thus sharing high inter-species synteny between closely related lineages [4]. In contrast, paralogous loci refer to homologs within one lineage and are due to, for example, tandem-, transposition- or whole genome duplications (WGDs) [5,6]. Large-scale synteny is not observed for paralogs derived from small-scale events like tandem- and transposition duplication. In contrast, paralogs derived from WGDs are located within intra-species syntenic genomic blocks, and can be referred to as ohnologs or syntelogs [7,8]. Supergene loci refer to clusters of genes in close genomic proximity, often causing linkage disequilibrium [9,10]. Tandem duplicates comprise arrays of paralog supergenes that are due to, for example, errors in meiosis like unequal crossing over and have been connected to metabolic diversification in plants [11-13].

Together with the continuous progress and use of next generation sequencing techniques, genome-wide analysis of syntelog distribution provided evidence for a history of ancient (shared and/or lineage-specific), successive polyploidy events for all flowering plant lineages. [4]. For example, the lineage of the model plant Arabidopsis underwent at least five polyploidy events during evolution, two preceding and three following angiosperm evolution [14]. Among those, the most recent WGD event is commonly referred to as “At-α” and is shared by all other mustard family members, including the extant sister clade of the Aethionemeae [15,16]. The more ancient At-β WGD event is in turn shared by core species in the order Brassicales and excepting early-branching lineages such as papaya [17,18] and therefore occurred after split of the Carica lineage. At-γ refers to an older whole genome triplication (WGT) event with evidence in all Asterids (including tomato) and Rosids, grape (Vitales) and basal clades such as Pachysandra terminalis (Buxales) and Gunnera manicata (Gunnerales) [19,20]. Crops like Brassica rapa (Br-α WGT), Solanum lycopersicum and Solanum tuberosum (Sol-α WGD/WGT) also show evidence of ancient genome multiplications [21,22]. As a consequence, the level of “genome multiplicity” expected from successive WGDs/WGTs in B. rapa (defined as “syntenic depth”) is 36x when compared to the 1x eudicot ancestor (3x due to At- γ, 2x due to At-β, another 2x due to At-α as well as 3x due to Br- α, see above).

Evidence is now accumulating for significant impact of ancient and recent gene and genome duplication events to birth and diversification of key biological traits. Duplication was proposed to be a key factor in expansion of regulatory and enzymatic pathways involved in generation of >200,000 diverse biochemical secondary metabolites in the flowering plant lineage [23-25]. For example, a differential impact of various duplication modes has been revealed for plant resistance proteins [26]. Likewise, the last three polyploidy events of the Arabidopsis-lineage (see above) likely contributed to shaping the genetic versatility of the glucosinolate pathway, a class of plant secondary metabolites with beneficial effects to human health and nutrition [25]. Similarly, polyploidy has been brought in connection to the origin of C4-photosynthesis in Cleomaceae [27].

Little is known about the impact of genome duplication to diversification of isoprenoid pathways. Isoprenoids form a highly diverse class of metabolites commonly found in all Angiosperm lineages [28]. For example, phytol side-chain substitutes of chlorophyll and carotenoid pigments as well as

phytohormones like gibberellin or brassinosteroids are well-characterized isoprenoids involved in basic metabolic processes that are essential for plant growth and development [29]. The most abundant group of plant isoprenoid derivatives comprises compounds of the terpenoid class [30]. Similar to glucosinolates, terpenoids are defined as specialized or secondary metabolites that play major roles in plant-insect interactions like, for example, attraction of beneficial organisms or defense against herbivores [30,31]. Boutanaev et al. investigated core terpene synthase (TPS) genes (which generate terpene scaffold diversity) and identified micro-syntenic clusters that have arisen within recent evolutionary history by gene duplication, acquisition of new function and genome reorganization [32]. Note that in concert with TPS genes, terpenoid biosynthesis depends on various independent pathways (referred to as modules hereafter). Here, we performed further extended comparative analysis of various independent terpenoid biosynthetic modules in context of gene- and genome duplication.

Briefly, a sequential combination of six distinct reaction modules acts in concert to convert primary metabolites to longer-chain compounds mediating designated biological function. Therefore, plant terpenoid biosynthesis displays “modular” organization, including (1) TPS genes, (2) IPP isomerases (IDI), (3) prenyltransferases (PTF), (4) genes from MVA and (5) MEP pathways as well as (6) triterpene-specific synthases (see fig. 1 for a comprehensive overview). Notably, genes involved in the latter three modules share a common evolutionary origin (i.e. genes are homologous) as previously described based on analysis of Solanaceae [33]. All terpenoids are synthesized from two universal C5-isoprenoid building blocks (a) isopentenyl diphosphate (IPP) and (b) its isomer dimethylallyl diphosphate (DMAPP). In plants, IPP is synthesized independently by the mevalonate (MVA, shown in black in fig. 1) and methylerythritol phosphate (MEP, Shown in purple in fig. 1) pathways. In contrast, DMAPP is synthesized by enzymes of the MEP pathway only [34]. Both DMAPP and IPP compounds can be isomerized by enzymes of the IPP isomerase type (IDI, shown in turquois in fig. 1) [35]. Due to the economic relevance of enzymes involved in MEP and MVA pathways as well as IPP isomerases, the underlying biochemistry has been thoroughly investigated. Note that both MVA and MEP pathways comprise sequential arrangements of consecutive reaction steps leading to formation of intermediate products [36]. Analysis of stoichiometry indicated dosage-dependent effects regarding both pathways in yeast [37]. Going beyond yeast, comparative network analysis of MVA and MEP pathways in prokaryotes and the model plant A. thaliana characterized dosage-dependent effects of enzymes in both pathways and elevations of corresponding metabolite concentrations in plants and humans. This indicates that enzymes involved in MVA and MEP pathways operate concentration-dependent across all kingdoms of life [36,38]. Similarly, genetic engineering of Escherichia coli in context of industrial terpenoid production revealed that enzymes of the IDI group function in a dosage-dependent manner [39]. This was confirmed by mechanistic investigations of IDI enzymes in Thermus thermophiles due to their relevance for a wide range of biotechnological applications [40]. Likewise, dosage-dependent effects have been revealed for plant-derived IDI enzymes. For example, the economic potential of in vitro production of caoutchouc led to cloning, heterologous expression and functional characterization (i.e. determination of biochemical function) of IDI loci from the rubber tree Hevea brasiliensis [41].

Enzymes of the prenyltransferase class (PTF, shown in green in fig. 1) subsequently catalyze formation of C10-prenyl diphosphate molecules. Moreover, they can mediate the (optional) elongation of the C10-backbone by the addition of further C5-isopentenyl diphosphate units necessary for formation of di- and sesquiterpenes including longer-chain (C25-C55) tetra- and

polyterpenes [42-44]. Terpene synthases (encoded by TPS genes, shown in red in fig. 1) catalyze conversion of specific C10-, C15 or C20 isoprenoid precursors to specialized monoterpenes (C10), sesquiterpenes (C15) and diterpenes (C20), building a module further downstream within terpenoid biosynthesis, respectively [29,45]. Specialized triterpene synthases catalyze formation of pentacyclic triterpenes (such as lupane and squalene) (C30, shown in blue in fig. 1) [46-48]. Note that those compounds can be further modified in distant branches of plant secondary metabolism, for example to triterpene alcohols (such as lanosterol and cycloarthenol) with various bioactivities [49,50]. Entry to the aforementioned MEP pathway was previously proposed to be catalyzed by two divergent 1-deoxy-D-xylulose 5-phosphate synthase isoforms in S. lycopersicum (SlDXS1 / SlDXS2) and A. thaliana (AtDXS1 / AtDXS2) [51-54]. In tomato, DXS1 is ubiquitously expressed whereas DXS2 transcripts are abundant in a few tissue types including glandular trichomes. Trichomes are hair-like structures present in the aerial parts of many plant species. They exhibit tremendous diversity but are of general interest to plant breeders since they are often responsible for the production of plant secondary metabolites with various bioactivities, including terpenoids [55-57]. Interestingly, knock-down of DXS2 led to a differential distribution of mono- and sesquiterpenes within tomato glandular trichomes as well as to a significant increase of trichome density, giving rise to economic and ecological potential to this small gene family [51].

The core-TPS gene family has been most intensively studied in the model plant Arabidopsis thaliana. The ecotype Col-0 contains 32 full-length functional and 8 pseudogenes of the terpene synthase type, of which about a third have been annotated to a designated biochemical function by functional characterization [29,58-64]. Most of the Col-0 core-TPS genes are constitutively expressed in roots, flowers or leaves for production of mono-, di- or sesquiterpenes whereas some are up-regulated under presence of specific stress-related stimuli [58,64,65]. Notably, 27 of the 32 Col-0 core-TPS genes comprise supergene clusters organized in 16 tandem arrays [11], whereas two of them constitute an ohnolog duplicate gene pair due to the most recent At-α ancient whole genome duplication event (see above) [8]. Beyond A. thaliana, efforts to identify core-TPS genes have been published for tomato (S. lycopersicum) [66], orange (C. sinensis) [67], eucalyptus (E. grandis) [68], grape (V. vinifera) [69], millet (S. bicolor), apple (M. domesticus) [70] and the basal Angiosperm Amborella (A. trichopoda) [71]. However, functional characterization (i.e. distinct biochemical function) of all TPS genes present within a species are currently available for tomato and the model plant Arabidopsis only. The complete set of biosynthetic elements involved in both MEP and MVA pathways as well as other terpenoid-associated prenyltransferases including triterpene-specific synthases has to-date only been described in Arabidopsis with a total of 34 genes [34]. Among those, 15 possess prenyltransferase activity, whereas nine and eight belong to the MVA and MEP pathway, respectively. Furthermore, the A. thaliana genome contains two genes encoding proteins of the IPP isomerase (IDI) type with similar bioactivities [34,35]. In total, the gene count of all modules within the complete terpenoid biosynthetic pathway therefore rises to 66 including 32 functional core-TPS genes in Arabidopsis, with a 64% (42 / 66) fraction of tandem duplicate supergenes and 21% (14 / 66) comprising ohnolog duplicate gene pairs dating back to the At-α, At-γ or At-β ancient whole genome multiplication events (see above) [8,11,72].

In this study, we employed a meta-method by combining evidence provided by sequence homology (BLAST), HMM Modelling (interpro scan) and genomic context (SynMap) for robust annotation of genes involved in all modules of terpenoid biosynthesis on a uniquely broad phylogenomics framework. First, we infer novel annotation for loci previously not brought in connection to terpenoid biosynthesis within 17 genome assembly including twelve major crops, thereby providing

insights to diversification of plant secondary metabolism during 250 MA of flowering plant evolution. Second, we assessed and compared key factors contributing to copy number variation across all terpenoid biosynthetic modules, thereby providing evidence for the impact of gene- and genome duplication to metabolic diversification in plants. Third, we established a novel clade of duplicate genes with pleiotropic effects in control of trichome density and terpenoid biosynthesis, thereby providing data that support the concept of functional divergence following gene and genome duplication. In summary, our data offer significant insight into evolutionary dynamics underlying diversification of plant secondary metabolism. Furthermore, we provide a blueprint for future efforts to identify and more rapidly modify terpenoid biosynthetic genes across all modules in any flowering plant species.

MATERIALS & METHODS

Software prerequisites

All employed Perl and Python scripts required perl (strawberry v5.18) and Python (v2.7) libraries including Bioperl (v1.6.910) and Biopython (v1.63) modules, respectively. The iprscan_urllib.py-script for HMM-based domain annotation (see below) required SOAPy, NumPy and urllib Python modules. For BLAST screens, we employed the stand-alone command line version of NCBI BLAST 2.2.27+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/, last accessed on December 13th, 2014) [73]. Fisher's exact test for count data was performed using the R package for statistical computing (www.r-project.org, last accessed on December 13th, 2014).

Genome annotations

In total, we analyzed 15 draft genomes as well as two gene-space assemblies represented by non-redundant transcriptomes. The Complete sets of representative genes and proteins for 12 of these 17 datasets were downloaded using www.phytozome.net, last accessed on December 13th, 2014) [74] and the CoGe package for comparative genomics [4]. We included *Amborella trichopoda* EVM27 [71], *Arabidopsis thaliana* TAIR10 [75], *Brassica rapa* v1.1 [76], *Carica papaya* v0.5 [18], *Citrus sinensis* v1 [77], *Eucalyptus grandis* v1.1 [78], *Glycine max* Wm82.a2.v1 [79], *Sorghum bicolor* v1.4 [80], *Solanum tuberosum* v3.2.10 [81], *Solanum lycopersicum* v2.40 (Potato Genome Consortium 2012), *Vitis vinifera* Genoscope.12X [20] and *Zea mays* 5a.59 [82]. *Tarenaya hasslerania* v5 [83] (Weber, Schranz et al, unpublished data), *Cleome gynandra* v2 (Weber, Schranz et al. 2014, unpublished data) and *Nicotiana benthamiana* v0.42 [84] genome annotations were made available by the authors. Non-redundant transcriptomes of *Cannabis sativa* ChemDawg (marijuana) [85] and *Lactuca sativa* (Mitchelmoore et al, unpublished data) early-draft gene-space assembly of were extracted from Genbank [86].

De novo protein annotation of early-draft gene-space assemblies

Non-redundant transcriptome data of *Cannabis sativa* ChemDawg (marijuana) and *Lactuca sativa* derive from unpublished early-draft gene-space assembly (see above) and therefore contain significant parts of non-coding sequences as well as putative sequencing errors. We therefore subjected both datasets to the translatedna.py script v1.75 (https://github.com/jenhantao/HiSeq/blob/master/translatedna.py, last accessed on December 13th, 2014). First, single mRNA sequences were translated in all six frames. Second, the peptide fragment encoded by the largest open reading frame was printed. All other parts were discarded. The output comprises an approximation on the non-redundant set of proteins for both species.

Confirmation and expansion of multi-gene families associated with the terpenoid biosynthetic module in *Arabidopsis thaliana* ("run 1")

Functional annotation of target genes across all organisms was an interlaced approach consisting of 3 independent BLAST screens (run 1 – 3). For run 1 in *A. thaliana*, we obtained 32 core-*TPS* genes from [29] as well as 34 genes acting in modules further up- and downstream in terpenoid

biosynthesis [34]. We queried all 66 sequences against the TAIR10 *A. thaliana* genome annotation in a BLAST screen without e-value threshold (forward run). We extracted all target sequences and queried them back against the *A. thaliana* TAIR10 genome annotation with an applied target sequence maximum threshold of 2 (under consideration of self-hits produced by Col-0 genes within this pool) (reverse run). After removal of self-hits, we scored loci as associated with the *A. thaliana* terpenoid specialized metabolism if they were part of the target sequence pool in the forward run, and aligned to a terpenoid biosynthetic gene as defined by [29,34] in the reverse run. We thereby created an extended set of *A. thaliana* TP-associated loci (85 genes).

Species-wise determination of putative homologous gene anchors ("run 2")

In the next step for large-scale specialized terpenoid biosynthetic gene identification, run 2 determined unidirectional best BLAST hits for both (a) protein and (b) coding DNA sequences between *A. thaliana* Col-0 and all other 14 genome annotations in a screen without e-value thresholds (for early-draft gene-space assemblies, only protein data were used). Since terpenoid biosynthetic loci can comprise multiple domain types connected by partially conserved linkers, the BLAST approach can result in false positives due to short but highly conserved highest-scoring sequence pairs (HSPs) in functionally non-relevant (i.e. structural) parts of the protein. Therefore, we developed a python script to discard target sequences with a query/target sequence length ratio below 0.5 and above 2.0 as previously described to avoid false positive BLAST results due to short but highly conserved highest-scoring sequence pairs (HSPs) in functionally non-relevant (i.e. structural) parts of the protein [26]. We determined (c) additional, length-filtered HSP pairs (based on both CDS and proteins) for these loci within the aforementioned length ratio scope to form a 2nd line of evidence for homolog gene detection as previously described [26].

Syntelog / ohnolog determination

Calculation of pairwise syntenic blocks within and between genomes is based on integer programming [87] but implemented to an easy-to-use web interface termed CoGe package for comparative genomics (www.genomevolution.org, last accessed on December 13th, 2014) [4,88]. Within all genome assemblies, we determined genes sharing the same genomic context to counterparts in the *A. thaliana* Col-0 genome annotation (defined as syntelogs) using the DAGchainer [89] and Quota-Align [87] algorithms implemented to the "SynMap" function within CoGe. To mask noise generated by successive duplication(s) of ohnolog blocks including segmental duplications, we applied Quota-Align ratios for the "coverage depth"-parameter that are consistent with the syntenic depth (defined as the level of "genome multiplicity" expected from the multiplication of successive WGDs/WGTs) calculated for each genome annotation. For merging of adjacent syntenic blocks, we applied a threshold of n=350 gene spacers. For within-species ohnolog counterparts of target genes, we applied the "Synfind" function within the CoGe package (https://genomeevolution.org/CoGe/SynFind.pl, last accessed on December 13th, 2014). To decrease false-positive scoring of recent segmental duplications, we set maximum threshold values of 1.5 for the Ks-value averages between duplicate gene copies. This facilitates selective scoring of ohnolog duplicate pairs within genomic blocks that are due to polyploidy as previously described [4]. Please note that we appended URLs to regenerate genome-wide ohnolog identification for 13 out of 17 genomes subjected to this analysis (see Results section).

Determination of tandem duplicate gene copies

Following a widely-used method for tandem duplicate identification, we queried the complete set of proteins encoded in the whole genome assembly against itself in a BLAST screen without any e-value threshold (this ensures the identification of most homologs including highly diverged ones) and filtered our final set of target sequences from above outside a window of n=10 allowed gene spacers in both directions from the query sequences (this ensures the identification of adjacent duplicates organized in arrays among all homologs scored above) as previously described for the identification of tandem duplicates [11]. We acknowledged that determination of genome-wide tandem duplicate frequencies following this approach decreases in accuracy with increased degrees of assembly fragmentation (i.e. total number of scaffolds/contigs). This means that false-negatives singletons are more likely scored in genomes with many short scaffolds ("gene-space assemblies") compared to annotations with few scaffolds in the size-range of chromosome pseudo-molecules which is due to the lack of information on the relative order of scaffolds. Similarly, it is not possible to score tandem duplicates based on non-redundant transcriptomes because those represent collections of single transcripts without information of the genomic context. As a result, our analysis of tandem duplicate fractions was restricted to 13 genome assemblies.

Scoring of putative gene transposition duplicate pairs among *Arabidopsis DXS*-like genes

Scoring of gene transposition duplicate pairs among *DXS* genes involved three steps. First, we obtained all tandem- and ohnolog duplicates present within the gene family as described above. Second, we queried CDS sequences of non-tandem/non-ohnolog duplicate target genes against the *Arabidopsis* genome in a BLAST screen without e-value threshold. Third, we generated (B)LastZ two-way alignments of the genomic regions that harbor (a) query as well as (b) highest-scoring non-self target sequence within a 40 kb window (20 kb on each side). This was accomplished using the GEvo function from the CoGe comparative genomics package (http://genomevolution.org/CoGe/GEvo.pl, last accessed on December 13th, 2014) [4]. Graphical highlights of transposon-like sequences have been customized by choosing "show other features" in the "results visualization" tab. We scored *DXS*-like gene pairs as gene transposition duplicates if they comprise highest-scoring sequence pairs embedded in otherwise non-syntenic regions, while both loci showing evidence for adjacent fragments of transposable elements as previously described [25].

Determination of anchor paralogs and generation of extended multi-gene family pools across all analyzed species ("run 3")

Since ortholog detection based on unidirectional or reciprocal best BLAST hits can miss many "real" orthologs in duplicate-rich species like animals or plants [90], a separate run was necessary to increase accuracy. For run 3, we defined the initial homologous genes set as the merged set consisting of five HSP partner groups (first group: based on length-filtered protein pairs; second group: based on non-length-filtered protein pairs; third group: based on non-length-filtered CDS pairs; fourth group: based on length-filtered CDS pairs; fifth group: based on syntelogs, see above for length filter criteria). We thereby created a set of putative homologous loci anchoring all *A. thaliana* gene families in all other analyzed genome annotations ("anchor pool"). In a next step, we performed a BLAST search without e-value thresholds to query all homologous anchor genes against all 17 genomes in a species-wise manner to determine putative paralogs of the anchor gene set ("run 3 forward"). We extracted all target sequences and queried them against the *A. thaliana* Col-0 TAIR10 genome annotation with a target sequence maximum threshold of 2 ("run 3 reverse"). After removal of self-hits, we scored loci as associated with terpenoid biosynthesis within their species if

they align to any member of the extended terpenoid biosynthetic loci in *A. thaliana* (see above). We defined all members of this pool as homologous to the anchor pool if they were not present within the set of homologous anchor genes (see above).

Hidden Markov Modeling and prediction of protein domains

Since we included highest-scoring sequence partners based on BLAST as well as syntelogs, the above-mentioned extended multi-gene family pool of terpenoid biosynthetic genes is based on both sequence homology and genomic location of its members. However, we observed an erosion of synteny across lineages relative to their phylogenetic distance. Furthermore, DNA sequence homology decreases with phylogenetic distance due to wobble rules for the 3rd codon position. Likewise, the protein sequence homology between distant multi-gene family members can decrease due to synonymous substitutions of amino acids belonging to the same chemical class (i.e. aliphatic, aromatic, basic, cyclic). Therefore, we applied a final filtering step to remove false-positive loci from the extended terpenoid biosynthetic genes pool across all genomes (including the extended terpenoid biosynthetic genes pool in *Arabidopsis*, see above). Using the iprscan_urllib.py script provided by the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) (https://www.ebi.ac.uk/Tools/webservices/download_clients/python/urllib/iprscan_urllib2.py, last accessed on December 13th, 2014), we queried every member of the terpenoid biosynthetic genes pool (including the extended set determined for *A. thaliana*, see above) to 14 algorithms that apply Hidden Markov Models for (protein domain) signature recognition (BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther and Gene3D) [91]. We overcame the one-sequence-at-a-time limitation of the EMBL server by writing batch wrappers for 25x-fold parallelization. As a result, we mapped all protein domains present in the putative multi-gene family pool onto their genes in less than a day, and discarded all false positive genes from the whole set (i.e. genes not encoding at least one domain common to at least one reaction module). Referencing of all identified genes to distinct terpenoid biosynthetic modules was based on presence of module-specific protein domains.

Multiple protein alignments

To generate multiple alignments of protein sequences, the stand-alone 64-bit version of MAFFT v7 was employed (http://mafft.cbrc.jp/alignment/software/, last accessed on December 13th, 2014) [92]. First, all terpenoid biosynthetic proteins were aligned species-wise using the command line mafft.bat --anysymbol --thread 4 --threadit 0 --reorder --auto input > output. Mesquite v2.75 (http://mesquiteproject.org, last accessed on December 13th, 2014) was used with multi-core preferences to trim MAFFT multiple alignments down to gap-free sites. Trimmed blocks were re-aligned using MAFFT with the command line mafft.bat --anysymbol --thread 4 --threadit 0 --reorder --maxiterate 1000 --retree 1 –localpair input > output.

Microarray-based gene expression analysis in *Arabidopsis*

To test differential and trichome-specific expression of *DXS*-like genes in *Arabidopsis*, we have used a Col-0 wild type trichome-specific transcriptome dataset (available at the TrichOME database, http://www.planttrichome.org/, last accessed on December 13th, 2014) [93]. Normalized values of three independent experiments performed with the ATH1 microarray were generated and averaged

as described [94]. For calculation of relative gene expression, we referenced the bHLH-motif containing house-keeping gene AT4G34720 [95].

Quantitative PCR-based gene expression analysis in tomato (*S. lycopersicum*)

Leaves, stems and roots were collected in triplicate from 4-week-old *Solanum lycopersicum* cultivar Moneymaker plants. Part of the stems were left intact and part were used for trichome isolation by shaking the stems in liquid nitrogen. Frozen isolated trichomes, stems that remained after trichome removal, intact stems, leaves and roots were ground to a fine powder and subjected to RNA isolation with Tri Reagent (Sigma) and DNase treatment (TURBO DNase, Ambion) according to the manufacturer's instructions. cDNA was synthesized from 1 µg of total RNA using the RevertAid kit (Fermentas). RT-qPCR was used to study the expression of 1-deoxy-d-xylulose 5-phosphate synthase isoforms 1, 2 and 3 (*DXS1*, *DXS2*, and *DXS3*) in cDNA derived from different tissues. Gene specific primers were designed using Primer3Plus (*DXS1*-F: 5'-ATTGGGATATGGCTCAGCAG-3'; *DXS1*-R: 5'-CAGTGGTTTGCAGAAACGTG-3'; *DXS2*-F: 5'-TTTACCGACCGCAACCTTAG-3'; *DXS2*-R: 5'-GTGCTTGAGGTCCAATTTGC-3'; *DXS3*-F: 5'- AATGGAGCCTTCACTTCACC-3'; *DXS3*-R: 5'-ACCCAGCTGCAAATGTTACC-3'). Tomato RUB1 conjugating enzyme-encoding (*RCE1*) gene (Gen-Bank accession no. AY004247) (*RCE*-F: 5'- GATTCTCTCTCATCAATCAATTCG-3' and *RCE*-R '5-GAACGTAAATGTGCCACCCATA-3') was used for normalization. PCR reactions were prepared in duplicate by mixing cDNA equivalents of 10 ng RNA with the SYBR Green Real-Time PCR master mix (Invitrogen) and 300 nM of each primer. Quantification of the transcript level was performed in an ABI 7500 Real-Time PCR System (Applied Biosystems) with the following cycling program: 2 min, 50 °C, 15 min 95 °C, 45 cycles of 15 sec at 95 °C and 1 min at 60 °C followed by a melting curve analysis. At the end of each run, amplified products were sequenced to verify their identity. Relative expression values were calculated using the efficiency δCt method as previously described [96]. All wet-lab expression analysis were performed in four independent biological and three technical replicates.

Phylogenetic and similarity/identity analysis

We performed Bayesian Markov chain Monte Carlo (MCMC) analysis using MrBayes version 3.2.2 (http://mrbayes.sourceforge.net/, last accessed on December 13th, 2014) [97] with the following parameters: Dirichlet model; uniform gamma shape parameter variation 0.00-200.00; 50 million generations; 2 independent runs, 4 chains each; temperature heating 0.2; sample taking every 5000 generations; burn-in time at 12500000 samples. Bayesian inference trees were constructed using the CIPRES package (http://www.phylo.org/sub_sections/portal/, last accessed on December 13th, 2014) [98]. Model convergence was checked in Tracer version 1.5 (http://tree.bio.ed.ac.uk/software/tracer/, last accessed on December 13th, 2014) [99]. FigTree v1.3.1 was used to generate and edit phylogenetic trees (http://tree.bio.ed.ac.uk/software/figtree/, last accessed on December 13th, 2014) [100]. Results were scored reliable once the effective sampling size of all parameters was above 100. Tree branches supported with posterior probabilities (PP) below 0.7 were considered weak and above 0.9 as strong. Protein sequence similarity analysis were performed using the Needle program from the EMBOSS software package (http://emboss.sourceforge.net/, last accessed on December 13th, 2014) [101].

Ethics Statement

The authors hereby state that no specific permissions were required for any activities and/or locations that are connected to this research. Likewise, the authors hereby confirm that the research summarized in this article did not involve endangered or protected species. In addition, the authors hereby state clearly that all sampling procedures and/or experimental manipulations were reviewed/specifically approved and no field permit was required. Wageningen University & Research Center and all other institutes affiliated with this work comprise legal entities that do not act on any basis that is prohibited by local, state or federal law.

RESULTS

(Re)annotation of the *Arabidopsis* terpenoid biosynthetic inventory and expansion of genes associated with all reaction modules

As an initial step for identification of genes involved in all terpenoid biosynthetic modules, we reviewed current literature to pool all published *Arabidopsis* core-TPS genes with biosynthetic elements acting further up- and downstream in the pathway (**fig. 1**) [29,34,61]. As a result, we generated a list of 66 biosynthetic elements previously identified in the model plant (**table 1A**). This compilation represents a patchwork of information containing both genes found by functional studies as well as genes with computationally inferred association to terpenoid metabolism. Hence, uniform standards of gene identification have not been applied for curation of this initial list.

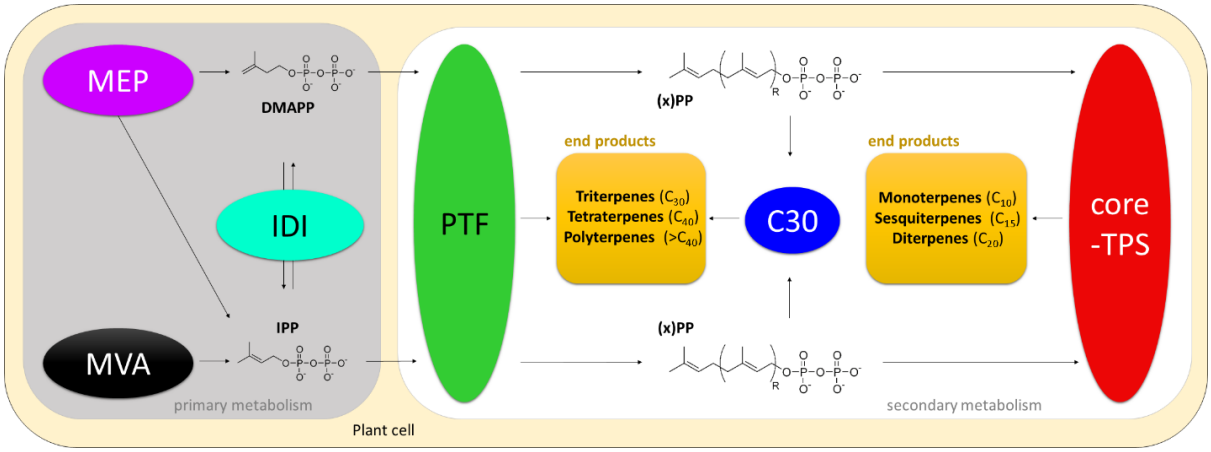


FIG. 1.—Overview of all plant specialized terpenoid biosynthetic modules. Proteins involved in the mevalonic (MVA, shown in black) and methylerythritol phosphate (MEP, shown in purple) pathways synthesize the universal C_5-isoprenoid building blocks isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). Those compounds can be isomerized using enzymes of the IPP isomerase class (IDI, shown in turquois). Subsequently, the C_5- blocks are transferred by enzymes of the prenyltransferase (PTF, shown in green) group to the isoprenoid intermediates with variable carbon backbone chain lengths (i.e. C_{10} for geranyl pyrophosphate; C_{15} for farnesyl pyrophosphate; C_{20} for geranylgeranyl phosphate and nerolidol diphosphate). Terpene synthase (core-*TPS*) gene products (shown in red) further catalyze biosynthesis of C_{10} (mono-), C_{15} (sesqui-) or C_{20} (di-) terpenes (end products, shown in yellow). C30 (shown in blue) refers to enzymes catalyzing biosynthesis of specific triterpenes (end products, shown in yellow). Likewise, prenyltransferases are involved in biosynthesis of longer-branched tetra- and polyterpenes (shown in green).

In a next step, we therefore screened for additional members of all involved gene families within the Col-0 genome that may have been missed in previous studies in an interlaced bioinformatics approach (see Materials & Methods section). Briefly, we combined layers of information based on

sequence similarity, gene synteny and Hidden Markov Modelling. First, we queried all 66 genes against the Col-0 genome in a very sensitive BLAST screen (no e-value cutoff). After removing self hits, we included every target sequence for further analysis if it formed a highest-scoring sequence pair with a distant member of the initial list. This included identification of ohnologs as well as annotation and mapping of protein domains. Interestingly, this led to evidence for 19 additional genes that produce highest-scoring sequence pairs (HSPs) with genes mentioned in **table 1A**. In essence, these genes comprise tandem-, transposition- as well as ohnolog duplicate copies of genes previously known to be associated with terpenoid metabolism (**table 1B**). Note that segmental duplications are excluded from this analysis due to technical reasons (see below). Interestingly, 18 of those 19 genes are annotated as triterpene-specific synthases. Thirteen of those have been functionally characterized in previous efforts [46-48,102]. Five further ones lack functional data but have been assigned to three-letter codes in the *Arabidopsis* information resource based on computational inferences (i.e. sequence homology). Interestingly, all those loci encode an oxysqualene synthase domain (**supplementary table 1,** Supplementary Material online). In contrast, one is a *DXS*-like gene with its closest homolog involved in the MEP pathway (**table 1B**). Note that the initial set of query genes applied to our first, above-mentioned BLAST analysis did not contain triterpene-specific synthases (**table 1A**). Therefore, our results indicate sequence homology of triterpene-specific synthases to genes of the core-*TPS* as well as the prenyltransferase class. Based on these findings, we hypothesize that all three groups go back to one common ancestral gene family with subsequent rounds of duplication going hand in hand with functional diversification as previously described in many cases including proline-rich proteins [103].

TABLE 1A.- The published terpenoid biosynthetic module in *Arabidopsis* . Gene abbreviations are adapted from the *Arabidopsis* Information Resource[A].

Gene ID	Annotation	Tandem duplicate	Bowers pair[B]	Reference[C]
Isopentenyl diphosphate (IPP) isomerases				
AT3G02780	*IDI1*	-	A12N076	Campbell et al., 1998
AT5G16440	*IDI2*	-	A12N076	Campbell et al., 1998
Mevalonic acid (MVA) pathway				
AT1G31910	*HMG1*	Yes	-	Benveniste et al., 2002
AT1G76490	*PMK*	Yes	C2N120	Caelles et al., 1989
AT2G17370	*HMG2*	-	C2N120	Caelles et al., 1989
AT2G38700	*MVD1*	-	A11N067	Cordier et al., 1999
AT3G54250	*MVD2*	-	A11N067	Benveniste et al., 2002
AT4G11820	*HMGS*	-	-	Montamat et al., 1995
AT5G27450	*MK*	-	-	Riou et al., 1994
AT5G47720	*ACT1*	Yes	-	Ahumada et al., 2008
AT5G48230	*ACT2*	Yes	-	Ahumada et al., 2008
Methylerythritol phosphate (MEP) pathway				
AT1G63970	*MDS*	-	-	Hsieh and Goodman, 2006
AT2G02500	*MCT*	-	-	Rohdich et al., 2000
AT2G26930	*CMK*	-	-	Hsieh et al., 2008
AT4G15560	*DXS1*	-	A15N013	Lange et al., 2003
AT4G34350	*HDR*	-	-	Hsieh and Goodman, 2005
AT5G11380[D]	*DXS3*	-	-	Lange et al., 2003
AT5G60600	*HDS*	-	-	Rodríguez-Concepción et al., 2002
AT5G62790	*DXR*	-	-	Schwender et al., 1999
Prenyltransferases (PTF)				
AT1G49530	*GGPS1*	Yes	-	Zhu et al., 1997a
AT2G18620	*GGPS2*	Yes	A10N118	Wang and Dixon, 2009
AT2G18640	*GGPS3*	Yes	-	Okada et al., 2000
AT2G23800	*GGPS4*	Yes	A10N309	Zhu et al., 1997b
AT2G34630	*GPS1*	-	-	Bouvier et al., 2000
AT3G14510	*GGPS5*	Yes	-	Finkelstein et al., 2002
AT3G14530	*GGPS6*	Yes	-	Wang and Dixon, 2009
AT3G14550	*GGPS7*	Yes	-	Okada et al., 2000
AT3G20160	*GGPS8*	-	-	Zhu et al., 1997a
AT3G29430	*GGPS9*	Yes	-	Finkelstein et al., 2002
AT3G32040	*GGPS10*	Yes	-	Finkelstein et al., 2002
AT4G17190	*FPS1*	Yes	A21N001	Cunillera et al., 2000
AT4G36810	*GGPS11*	-	A10N118	Okada et al., 2000
AT4G38460	*GGR*	-	-	Oh et al., 2002
AT5G47770	*FPS2*	Yes	A21N001	Delourme et al., 1994

(continued)

TABLE 1A (continued) .- The published terpenoid biosynthetic module in *Arabidopsis* . Gene abbreviations are adapted from the *Arabidopsis* Information Resource[A].

Gene ID	Annotation	Tandem dup	Bowers pair[B]	Reference[C]
Core terpene synthases				
AT1G31950	*TPS29*	Yes	-	Lange et al., 2003
AT1G33750	*TPS22*	-	-	Lange et al., 2003
AT1G48800	*TPS28*	Yes	-	Lange et al., 2003
AT1G61120	*TPS4*	Yes	-	Herde et al., 2008
AT1G61680	*TPS14*	Yes	-	Chen et al., 2003
AT1G66020	*TPS26*	Yes	-	Lange et al., 2003
AT1G70080	*TPS6*	Yes	-	Lange et al., 2003
AT1G79460	*TPS32*	Yes	-	Yamaguchi et al., 1998
AT2G23230	*TPS05*	Yes	-	Dal Bosco et al., 2003
AT2G24210	*TPS10*	-	-	Bohlmann et al., 2000
AT3G14490	*TPS17*	Yes	-	Dal Bosco et al., 2003
AT3G14520	*TPS18*	Yes	-	Lange et al., 2003
AT3G14540	*TPS19*	Yes	-	Lange et al., 2003
AT3G25810	*TPS24*	Yes	-	Chen et al., 2003
AT3G25820	*TPS27*	Yes	-	Chen et al., 2004
AT3G25830	*TPS23*	Yes	-	Chen et al., 2004
AT3G29110	*TPS16*	Yes	-	Lange et al., 2003
AT3G29190	*TPS15*	Yes	-	Lange et al., 2003
AT3G29410	*TPS25*	Yes	-	Dal Bosco et al., 2003
AT3G32030	*TPS30*	Yes	-	Lange et al., 2003
AT4G02780	*TPS31*	-	-	Mann et al., 2010
AT4G13280	*TPS12*	Yes	-	Ro et al., 2006
AT4G13300	*TPS13*	Yes	-	Ro et al., 2006
AT4G15870	*TPS1*	Yes	-	Aubourg et al., 1997
AT4G16730	*TPS2*	Yes	-	Huang et al., 2010
AT4G16740	*TPS3*	Yes	-	Fäldt et al., 2003
AT4G20200	*TPS7*	Yes	-	Lange et al., 2003
AT4G20210	*TPS8*	Yes	A21N124	Tholl and Lee, 2011
AT4G20230	*TPS9*	Yes	-	Dal Bosco et al., 2003
AT5G23960	*TPS21*	-	-	Chen et al., 2003
AT5G44630	*TPS11*	-	A21N124	Tholl et al., 2005
AT5G48110	*TPS20*	Yes	-	Dal Bosco et al., 2003

[A] TAIR10, www.arabidopsis.org, last accessed on December 13th, 2014

[B] Ohnolog pair according to Bowers et al., 2003 [8]

[C] for a comprehensive review, see Tholl and Lee, 2011 and Phillips et al., 2008)

[D]assosciation of *AtDXS3* to MEP pathway is subject of scientific debate (see Phillips et al., 2008)

To account for the whole range of sequence diversity found among *Arabidopsis* terpenoid biosynthetic genes, we merged tables 1A and B and obtained pool of 85 genes putatively involved in all Col-0 terpenoid biosynthetic modules ("extended set") (sum of all Col-0 gene entries in **supplementary table 2,** Supplementary Material online). Initially, we dissected all members into three groups based on putative affiliation with a certain module (**fig. 1**): (a) prenyltransferases and triterpene-specific synthases, (b) core terpene synthases and (c) genes involved in MEP and MVA pathways including IPP isomerases, respectively. Visual comparison of duplicate fractions revealed striking differences between the subsets but also when comparing to the fraction of duplicates among all protein-coding genes (**fig. 2**). For the whole set of 85 genes, we found a 68%-fraction of tandem duplicate supergene clusters and a 15%-fraction of ohnolog duplicate pairs (**fig. 2A**, **table 2 and table 3**). For subgroup (a), we report a 70%-fraction of tandem duplicate supergenes and a 13%-fraction of ohnolog duplicate copies (**fig. 2B**). In contrast, 94% of subgroup (b) comprise members of tandem arrays, while the ohnologs fraction drops to 6% (**fig. 2C, table 2 and table 3**). Interestingly, subgroup (c) contains only 16% of tandem duplicate genes but a 27% fraction of genes retained after ancient polyploidy events (**fig. 2D**).

In summary, we reported a connection to biosynthesis of mono-, di- and sesquiterpenes for 19 additional genes in *Arabidopsis* that are homologous to but absent from the published set of terpenoid biosynthetic genes. Likewise, we showed an asymmetric distribution of duplicates among genes involved in different modules of terpenoid biosynthesis in *Arabidopsis*.

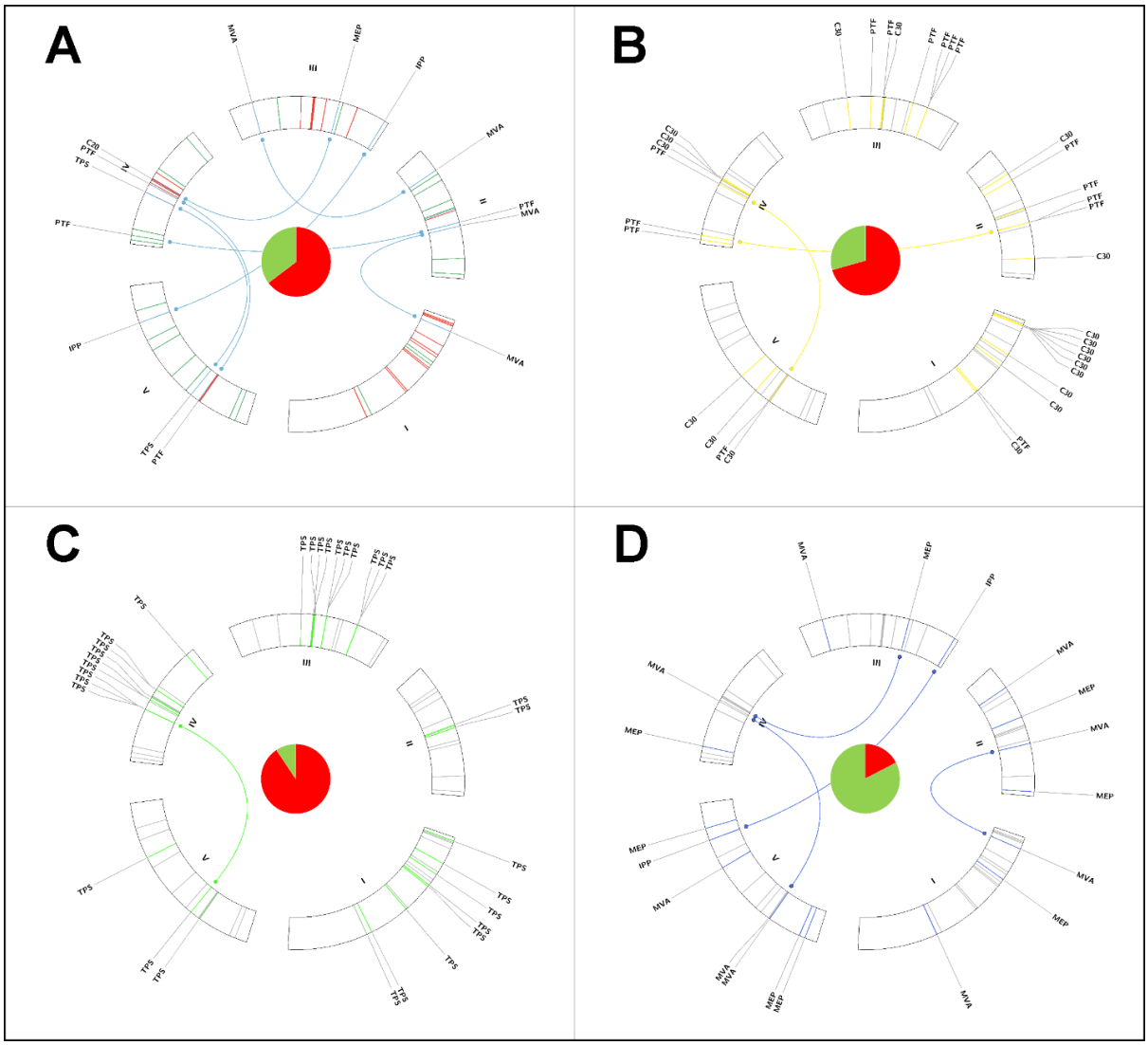


FIG. 2.—Circos ideogram showing 5 *Arabidopsis* chromosomes with the extended set of genes associated with major terpenoid biosynthetic modules. **A**. Gene inventory of the complete terpenoid biosynthetic pathway after initial expansion of published modules. Tandem duplicate supergenes are marked in red. Singletons are marked in green. Ohnolog duplicate gene pairs are marked in blue. Central pie chart shows a 68% tandem duplicate supergenes fraction. **B**. Subset of prenyltransferases and specific triterpene synthases marked in yellow. Central pie chart shows a 70% tandem duplicate supergenes fraction. **C**. Subset of core terpene synthase (*TPS*) genes marked in bright green. Central pie chart shows an 84% tandem duplicate supergenes fraction. **D**. Subset of genes associated with MEP and MVA pathways, including IPP isomerases, marked in blue. Central pie chart shows a 16% tandem duplicate supergenes fraction.

TABLE 1B.- The extended terpenoid phenotypic module in *Arabidopsis,* including triterpene- specific (C_{30}) synthases. Three letter gene abbreviations are adapted from the *Arabidopsis* Information Resource[a].

Gene ID	Annotation	Description[A]	Tandem duplicate	Bowers pair[B]	Reference[C]
Methylerythritol phosphate (MEP) pathway					
AT4G15560	*DXPS2*	Desoxy-xylulosephosphate synthase 2	Yes	A15N013	Lange et al., 2003
Triterpene-specific synthases					
AT1G62730	-	N/A; Squalene/phytoene synthase	No	-	Wang et al., 2008
AT1G66960	*LUP5*	Lupeol synthase 5	Yes	-	Herrera et al., 1998
AT1G78480	-	N/A; Prenyltransferase/squalene oxidase	Yes	-	Hanada et al., 2010
AT1G78500	*PEN6*	Pentacyclic triterpene synthase 6	Yes	-	Husselstein-Muller et al., 2001
AT1G78950	*LUP4*	Lupeol synthase 4	Yes	-	Benveniste et al., 2002
AT1G78955	*CAMS1*	Camelliol synthase 1	Yes	-	Kushiro et al., 1998
AT1G78960	*LUP2*	Lupeol synthase 2	Yes	-	Herrera et al., 1998
AT1G78970	*LUP1*	Lupeol synthase 1	Yes	-	Herrera et al., 1998
AT3G29255	-	N/A; Squalene cyclase (InterPro:IPR01833	Yes	-	this manuscript
AT2G07050	*CAS1*	Cycloartenol synthase 1	-	-	Lange et al., 2003
AT3G45130	*LAS1*	Lanosterol synthase 1	-	-	Benveniste et al., 2002
AT4G15340	*PEN1*	Pentacyclic triterpene synthase 1	Yes	-	Husselstein-Muller et al., 2001
AT4G15370	*PEN2*	Pentacyclic triterpene synthase 2	Yes	-	Husselstein-Muller et al., 2001
AT5G36150	*PEN3*	Pentacyclic triterpene synthase 3	-	-	Husselstein-Muller et al., 2001
AT5G42600	*MRN1*	Marernal Synthase 1	-	-	Benveniste et al., 2002
AT5G48010	*THAS1*	Thalianol Synthase 1	Yes	-	Benveniste et al., 2002
Function not clear					
AT1G48820	-	N/A; tandem duplicate of *TPS28*	Yes	-	Lange et al., 2003
AT2G37140	-	N/A; best BLAST hit is *TPS1*	-	-	Lange et al., 2003

[A] TAIR10, www.arabidopsis.org, last accessed on December 13th, 2014

[B] Ohnolog pair according to Bowers et al., 2003 [8]

[C] for a comprehensive review, see Tholl and Lee, 2011 [30]

TABLE 2.- Tandem Duplicates fractions among terpenoid specialized biosynthetic module in 13[A] genomes. Red indicates absence of tandem duplicates. Green indicates significant enrichment compared to genome-wide tandem duplicate fraction based on fisher's exact test on count data (p-value threshold: 0.01). For absolute gene numbers and p-values, see Supplementary Table 5.

Species	Genome-wide	core-TPS genes	MEP-pathway	MVA-pathway	IPP-isomerases	Prenyltransferases	Triterpene synthases	Average[B]
A. thaliana	**15%**	94%	0%	33%	0%	73%	68%	**68%**
B. rapa	**20%**	51%	0%	7%	0%	42%	40%	**33%**
T. hasslerania	**17%**	62%	18%	7%	0%	53%	20%	**37%**
C. papaya	**17%**	52%	0%	0%	0%	0%	54%	**32%**
C. sinensis	**44%**	37%	50%	56%	100%	25%	37%	**39%**
E. grandis	**33%**	73%	18%	19%	0%	60%	75%	**63%**
G. max	**73%**	42%	12%	0%	50%	7%	33%	**20%**
V. vinifera	**32%**	91%	21%	35%	0%	0%	96%	**78%**
S. lycopersicum	**28%**	80%	0%	13%	50%	55%	55%	**56%**
S. tuberosum	**51%**	51%	0%	21%	0%	0%	25%	**36%**
S. bicolor	**35%**	68%	20%	0%	0%	14%	76%	**52%**
Z. mays	**44%**	42%	31%	40%	33%	10%	58%	**40%**
A. trichopoda	**24%**	57%	18%	0%	0%	0%	67%	**34%**
Average[B]	**32%**	**59%**	**13%**	**17%**	**17%**	**25%**	**53%**	**46%**

[A] *C. Sativa, L. sativa* and *N. benthamiana* and *C. gynandra* are excluded from this analysis due to technical reasons (see Materials & Methods section)

[B] Averages based on numbers of tandem and singleton genes, not on pentages values since gene counts in subsets are not equal

Table 3.- Ohnolog duplicates fractions among the terpenoid specialized biosynthetic module in 13[A] genomes. Red indicates absence of ohnolog duplicates. Green indicates above-average fraction of ohnolog duplicates compared to the genome-wide background. For absolute values, see Supplementary Table 5.

Species	Genome-wide	CoGe-link[C]	core-TPS genes	MEP-pathway	MVA-pathway	IPP-isomerases	Prenyltransferases	Triterpene synthases	Average[B]
A. thaliana	22%	bit.ly/1t7DH7A	6%	13%	22%	100%	33%	5%	**15%**
B. rapa	53%	bit.ly/1uVTlHt	26%	54%	73%	100%	74%	33%	**49%**
T. hasslerania	48%	bit.ly/1r0khkj	26%	27%	60%	100%	47%	45%	**40%**
C. papaya	7%	bit.ly/1yt11Ap	0%	0%	0%	0%	17%	0%	**2%**
C. sinensis	6%	bit.ly/1xRKTJh	3%	0%	22%	0%	0%	0%	**3%**
E. grandis	18%	bit.ly/1p2oGrm	10%	0%	50%	0%	0%	0%	**11%**
G. max	62%	bit.ly/1yt2QNw	48%	56%	85%	100%	53%	27%	**57%**
V. vinifera	22%	bit.ly/1uefADr	2%	0%	15%	0%	40%	0%	**4%**
S. lycopersicum	19%	bit.ly/1xsTXoT	14%	0%	33%	0%	0%	0%	**12%**
S. tuberosum	11%	bit.ly/1yWDKGZ	4%	0%	17%	0%	0%	0%	**6%**
S. bicolor	23%	bit.ly/1xRLLxE	0%	0%	27%	0%	29%	20%	**11%**
Z. mays	27%	bit.ly/11xv3rs	8%	23%	55%	66%	60%	16%	**28%**
A. trichopoda	7%	bit.ly/1x3SpxM	0%	1%	17%	0%	33%	0%	**8%**
Average[B]	**28%**		**10%**	**19%**	**42%**	**48%**	**33%**	**10%**	**18%**

[A] *C. sativa* and *L. sativa* , *N. benthamiana* and *C. gynandra* are excluded from this analysis due to technical restrictions

[B] Averages based on numbers of tandem and singleton genes, not on percentages values since gene count in subsets is not equal

[C] link to the CoGe platform for comparative genomics for online-regeneration of the analysis for ohnolog identification

Protein domain annotation of extended genes set associated with all terpenoid biosynthetic modules in *Arabidopsis*

Increasing phylogenetic distance of plant species can lead to increased sequence diversity in homologs while the broad class of biological function remains unchanged [1,104]. For example, amino acid substitutions within the same chemical group (i.e. aliphatic, aromatic) may have little or no effects on protein function, but may result in decreased accuracy in orthologous and paralogous gene detection by sequence homology (such as BLAST) [90,105,106]. We therefore performed Hidden Markov Modelling (HMM)-driven protein motif searches and annotation among all subsets of genes involved in the extended set of Col-0 terpenoid biosynthetic genes in order to screen for additional homologs (see Materials & Methods section). Briefly, we submitted all 85 target sequences to the "Interpro5" algorithm that performs parallelized prediction of protein domains (see Materials & Methods section) [91,107-109]. This is based on machine learning for pattern recognition rather than direct sequence comparisons. As the "training" dataset for domain modelling for the submitted protein sequences, Interpro5 uses the HMM-generated profiles of all protein motif entries and associated sequences present within the pfam and various other databases [110]. Notably, benchmarking of profile HMMs and the BLAST algorithm previously revealed a higher sensitivity of HMM-based methods that is mirrored by an increased alignment quality [111].

As a result of HMM-driven protein domain annotation, we obtained a collection of all motifs encoded by all genes present in the initial set (**table 4**). First, we pooled all terpenoid pathway-associated genes from the extended set into (a) core-TPS proteins, (b) IPP isomerases, (c) genes involved in the MEP pathway, (d) MVA pathway-associated proteins, (e) prenyltransferases as well as (f) triterpene-specific synthases and subjected all six sets to the Interpro5 algorithm [91], thereby querying a total of 14 protein motif databases. Five among those recognized motifs shared by every single member of at least 2 pools and were selected for further analysis: Interpro, Pfam, Panther, Gene3D as well as Superfamily [107,110,112-114] (**table 4**). In a next step, we screened for protein motif entries within these 5 databases that are specific for any of the 6 aforementioned subsets of genes associated with all Col-0 terpenoid biosynthetic modules (**supplementary table 1,** Supplementary Material online). Interestingly, our approach identified 38 domains associated with more than one subgroup due to accurate modelling of protein domain signatures (**table 4**). Those were found either for both core-TPS proteins and prenyltransferases, or for both core-TPS genes and triterpene-specific synthases. Together with the sequence homology determined in the initial BLAST screen that formed the extended set of Col-0 target genes, this illustrates that those gene families are similar in terms of both sequence and domain structure as described above and hence might share a common evolutionary origin and function (**supplementary table 1,** Supplementary Material online). As a result, it is not possible to affiliate one distinct homologous genes to one of these three subsets based on domain composition in every case without functional data at hand, also when utilizing specific combinations of two or more domains. In summary, we performed in-depth investigation of protein domains among all enzymes involved in every *Arabidopsis* terpenoid biosynthetic module, thereby curating a set of all detectable domains involved in the terpenoid biosynthetic pathway within the *Arabidopsis* model plant.

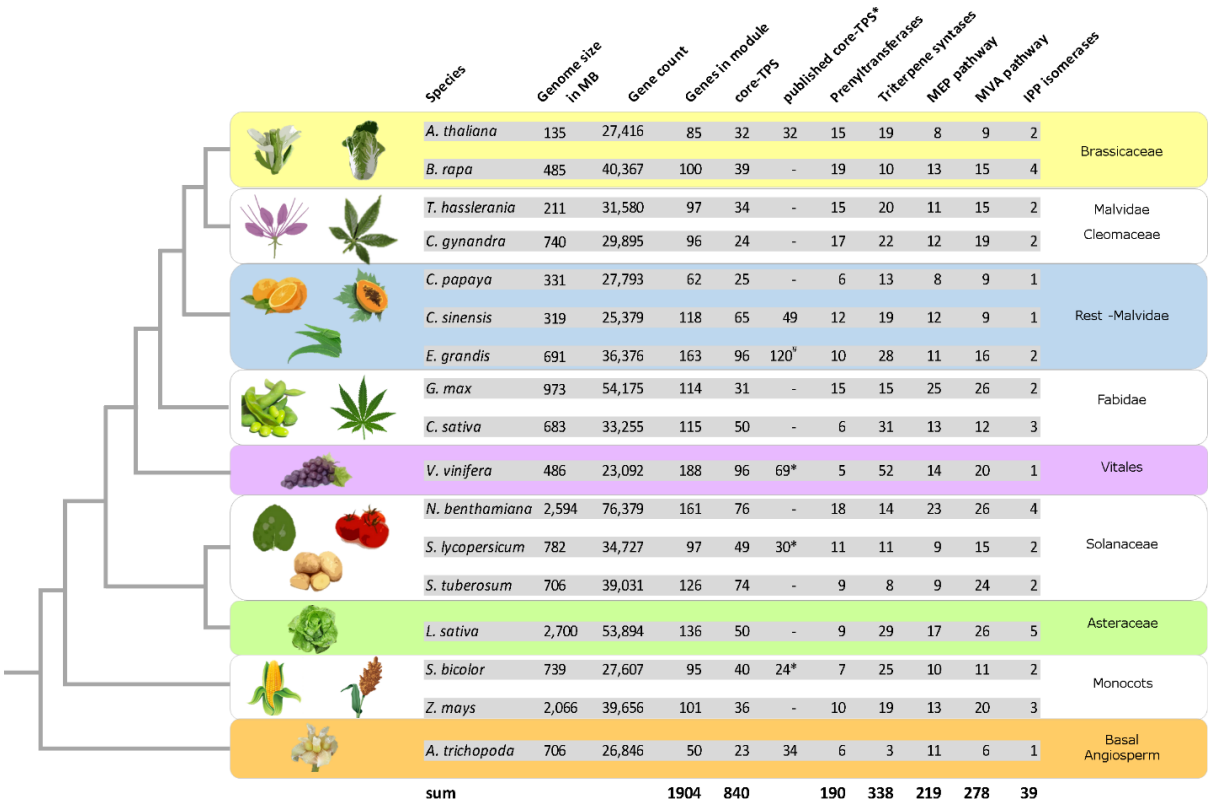
TABLE 4.- Overview of protein domain annotation for the extended set of *Arabidopsis* terpenoid biosynthetic genes[A]

Database	Predicted domains	Predicted domains specific for functional module	Genes with predicted domains	Genes with module-specific domains
Interpro	64	49	85	48
Panther	20	18	85	59
Pfam	25	17	85	43
Superfamily	16	10	84	11
Gene3D	16	9	83	10
Total	**141** (100%)	**103** of 141 (73%)	**85** of 85 (100%)	**59** of 85 (69%)

[A] 85 target genes in the extended set of *Arabidopsis* terpenoid biosynthetic genes

Annotation of genes in all terpenoid biosynthetic modules across 17 target species based on both sequence homology and protein domain composition

We obtained a list of 1,904 protein-coding genes with putative annotation to a terpenoid biosynthetic module. To cross-reference every member to one of the six designated functional modules (**fig. 1**), we mapped all aforementioned protein motifs onto all target genes. For genes with ambiguous domain composition (i.e. presence of 38 domains without clear referencing to one functional module within the terpenoid biosynthesis, see above), we used the annotation of its highest-scoring target sequence alignment in *Arabidopsis*. Depending on sequence homology as well as on presence/absence of the aforementioned module-specific protein domains (**supplementary table 1,** Supplementary Material online), we describe a total of 840 core terpene synthase genes (shown red in **fig. 1**), 190 prenyltransferases (shown in green in **fig. 1**), 338 triterpene-specific synthases (shown in blue in **fig. 1**) as well as 219 and 278 genes associated with the MEP (shown in purple in **fig. 1**) and MVA pathways (shown in black in **fig. 1**), respectively. Likewise, we found a total of 39 IPP isomerases (shown in turquois in **fig. 1**), summing up to 1,904 target genes in total (**fig. 3**)**.** Please note that all sequence identifiers are appended in (**supplementary table 2,** Supplementary Material online).



Species	Genome size in MB	Gene count	Genes in module	core-TPS	published core-TPS*	Prenyltransferases	Triterpene synthases	MEP pathway	MVA pathway	IPP isomerases	
A. thaliana	135	27,416	85	32	32	15	19	8	9	2	Brassicaceae
B. rapa	485	40,367	100	39	-	19	10	13	15	4	
T. hasslerania	211	31,580	97	34	-	15	20	11	15	2	Malvidae
C. gynandra	740	29,895	96	24	-	17	22	12	19	2	Cleomaceae
C. papaya	331	27,793	62	25	-	6	13	8	9	1	Rest -Malvidae
C. sinensis	319	25,379	118	65	49	12	19	12	9	1	
E. grandis	691	36,376	163	96	120[N]	10	28	11	16	2	
G. max	973	54,175	114	31	-	15	15	25	26	2	Fabidae
C. sativa	683	33,255	115	50	-	6	31	13	12	3	
V. vinifera	486	23,092	188	96	69*	5	52	14	20	1	Vitales
N. benthamiana	2,594	76,379	161	76	-	18	14	23	26	4	Solanaceae
S. lycopersicum	782	34,727	97	49	30*	11	11	9	15	2	
S. tuberosum	706	39,031	126	74	-	9	8	9	24	2	
L. sativa	2,700	53,894	136	50	-	9	29	17	26	5	Asteraceae
S. bicolor	739	27,607	95	40	24*	7	25	10	11	2	Monocots
Z. mays	2,066	39,656	101	36	-	10	19	13	20	3	
A. trichopoda	706	26,846	50	23	34	6	3	11	6	1	Basal Angiosperm
sum			**1904**	**840**		**190**	**338**	**219**	**278**	**39**	

FIG. 3.—Illustration showing the complete set of genes associated with all terpenoid biosynthetic modules identified in this study across 17 genome assemblies, based on the HMM-generated profiles of table 4. For core-*TPS* genes, numbers of previously published full-length target genes is included if available. Asterisks indicate number of previously identified full-length *TPS* open reading frames and hence putative number of functional terpene synthase enzymes. Incomplete protein fragments are not included.

Compared to the total number of protein-coding genes present in the genome, *V. vinifera* (grapevine) possesses the most expanded inventory of terpenoid biosynthetic genes including all modules, but also for individual modules like core terpene synthases, triterpene-specific synthases and both MEP and MVA pathways. In contrast, the highest number of prenyltransferases relative to the total number of protein-coding genes is encoded by the C4-species *C. gynandra*. The small gene family of IPP isomerases is most abundant (i.e. target gene count compared to number of all genes per genome) in *B. rapa*. Note that the *B. rapa* genome possesses the highest syntenic depth level among all species analyzed in this study (**supplementary table 3,** Supplementary Material online).

In contrast, the basal Angiosperm *A. trichopoda* possesses the leanest inventory relative to the number of all protein-coding genes when looking at all terpenoid biosynthetic modules. Same counts for triterpene-specific synthases and for genes associated with the MVA pathway. For core-*TPS* genes, the *G. max* (soybean) genome encodes the smaller relative number of proteins. For prenyltransferases, we found that the *L. sativa* genome encodes the smallest relative number. In contrast, the MEP pathway in *S. tuberosum* (potato) recruits the lowest number of genes compared to all of its protein-coding genes. Finally, we found the lowest relative number of IPP isomerases within the *C. papaya* genome (**supplementary table 3,** Supplementary Material online). In summary, we provide evidence for annotation of 1,904 genes to every major module of terpenoid biosynthesis within 17 target genomes, many of which have not been connected to this trait so far. Similar to functional annotation of the *Arabidopsis* genome, computational inferences of gene function comprise an important step for the future collection of functional data in wet-lab experiments [75].

General and subset-specific cross-referencing of supergene clusters and ohnolog duplicates to terpenoid biosynthetic elements among all species

After curating a set of 1,904 target genes across 17 species, we first scored supergenes organized in tandem arrays as well as ohnolog duplicates due to polyploidy events. Second, we compared the obtained duplicate fractions between all six modules of terpenoid biosynthesis. For detection of potential enrichment or depletion of duplicate frequencies within these subsets, a species-wise comparison to the genome-wide average of tandem/ohnolog duplicates fraction was necessary. Due to technical reasons, these genome-wide fractions can't be accurately determined for *C. gynandra*, *N. benthamiana, C. sativa and L. sativa* (for *Cannabis* and *Lactuca*, non-redundant RNAseq data are available only whereas the *C. gynandra* and *N. benthamiana* assemblies are highly fragmented, leading to a highly error-prone determination of genome-wide duplicates fractions, see Materials & Methods section). Therefore, our genome-wide analysis of duplicates fractions was restricted to 13 genome assemblies.

On average, 46% of all curated genes associated with terpenoid biosynthesis comprise supergenes with duplicates organized in tandem arrays. Compared to the 32% average observed for the genome-wide tandem duplicate fraction determined across the 13 genome assemblies subjected to this part of our analysis, our results highlight a significant enrichment of supergene clusters for terpenoid biosynthetic genes according to statistical analysis based on Fisher's exact test on count data (**table 2, supplementary table 4,** Supplementary Material online). Next, we investigated the species-wise fractions of tandem duplicates among all identified terpenoid biosynthetic genes for comparison to the respective genome-wide background. Similar to our findings for genome-wide tandem duplicate fractions across all analyzed genomes, the significant enrichment for supergene clusters holds up for all organisms except *C. sinensis*, *G. max*, *Z. mays* and *A. trichopoda* (**table 2**). However, comparison of duplicate frequencies within different functional modules of terpenoid biosynthesis across the 13 genomes subjected to tandem duplicate analysis did reveal certain subsets that are enriched for duplicates (five genomes were not applicable to this analysis due to technical reasons, see above). For example, triterpene-specific synthases are significantly enriched for tandem arrayed supergenes compared to the genome-wide background in *G. max*. Similarly, core-*TPS* genes are enriched for tandem duplicates compared to genome-wide average in the basal angiosperm *A. trichopoda* (**table 2**). We have found that only *Citrus* and maize lack significant enrichment for tandem duplicates among all subsets of genes involved in terpenoid biosynthesis. Based on the enriched fraction of tandem duplicates specific for certain terpenoid biosynthetic modules, we deducted a general pattern. Both core-*TPS* genes and triterpene-specific synthases were found to be significantly more enriched for tandem duplicates across most of the analyzed species, whereas MEP and MVA pathways as well as IPP isomerase functions retained few or no supergene clusters within most analyzed species (**table 2**, **supplementary table 4,** Supplementary Material online).

In a next step, we determined the cumulative fraction of duplicate genes retained after ancient polyploidy events (ohnologs). Similar to the analysis of tandem duplicates, ohnolog identification relies on gene contextual information and is hence not applicable to highly fragmented gene-space assemblies or translated transcriptome datasets (see above) [4,89]. Please note that we appended URLs for online-regeneration of ohnolog identification in 13 genomes out of 17 genomes (**table 3**). We again measured genome-wide averages wherever possible and compared them to the fractions among all subsets as described above for tandem duplicate supergenes (**table 3**). On average, 18% of all genes associated with all modules of terpenoid biosynthesis comprise ohnolog duplicate gene

copies. Compared to the 28% fraction of genome-wide ohnolog merged across all analyzed species, Fisher's exact test on count data indicates absence of significant ohnolog enrichment for this set (**supplementary table 5,** Supplementary Material online). In contrast, species-wise analysis revealed a significant enrichment of ohnologs among all terpenoid biosynthetic genes identified in *Z. mays* (**table 3**). Moreover, analysis of species-specific ohnolog distributions among different terpenoid biosynthetic modules highlighted differential trends. In essence, we revealed patterns of above-average ohnolog retention opposite to those described for tandem duplicates. For example, dosage-independent modules like core-TPS synthases and triterpene-specific synthases contain below-average ohnolog fractions in all analyzed species (**table 3**), while recruiting highest fractions of supergene clusters as shown above (**table 2**).Strikingly, genes associated with dosage-dependent modules like the MVA pathway and IPP isomerases show the highest fractions of ohnolog duplicates merged across all genomes (**table 3**). In contrast, both subsets include low fractions of tandem duplicates compared to other subsets (**table 2**).

However, ohnolog fractions of dosage-dependent modules vary greatly between different species in many cases. The small gene family of IPP isomerases, for example, consists of 100% ohnolog duplicates within *Arabidopsis*, *Brassica*, *Tarenaya* as well as *Glycine*. In contrast, we did not detect retained ohnologs within this gene families within *Carica*, *Citrus*, *Eucalyptus*, *Vitis*, all analyzed Solanaceae as well as *Sorghum* based on the applied preferences. This is likely due to technical reasons (see Materials & Methods section). Briefly, the scoring method of SynMap depends on presence of long colinear regions and hence the N50 value indicating the "fragmentation" of the assembly. This means that false-negatives are more likely scored in genomes with many short scaffolds compared to few in the size-range of chromosome pseudo-molecules, due to the lack of information on the relative order of scaffolds.

In summary, we showed above-average fractions of ohnologs combined with below-average fractions of supergene clusters recruited by two dosage-dependent terpenoid biosynthetic modules (IPP isomerases and genes involved in the MVA pathway) (**table 2, table 3**). In addition, we revealed a below-average rate of ohnolog retention combined with a significantly increased rate of tandem duplicates for stoichiometrically insensitive genes (i.e. genes that are not acting in a dosage-dependent way) like core-terpene synthases as well as triterpene-specific synthases.

Identification and phylogenetic analysis of key genes controlling isoprenoid profiles and trichome density

The aforementioned biosynthetic inventory of all plant terpenoid biosynthetic modules is necessary and sufficient for production of related compounds with designated biochemical function. However, some terpenoids are autotoxic and can only be produced in high amounts in specialized hair-like aerial structures termed glandular trichomes [56,115] where they are stored or secreted to the surface in order to facilitate ecological interactions (i.e. repelling herbivores or attracting beneficial organisms). Biogenesis and distribution of trichomes is controlled by various biosynthetic and regulatory processes, often mediated by pleiotropic genes [116,117]. In this context, it has recently become evident that trichome density on the leaf surface is amongst other factors influenced by a class of pleiotropic genes that also catalyzes the entry step to the MEP pathway [34]. In tomato, two deoxy-xylulosephosphate synthase genes (*DXS*) have previously been identified. Interestingly, differential and tissue-specific expression was observed: While *DXS1* is ubiquitously expressed, *DXS2* was found to be abundant in only a few tissues including trichomes. Reduction of *DXS2* expression in cultivated tomato led to an increase in glandular trichome density [51]. To identify additional *DXS*-

like homologs, we screened our curated genes set and found evidence for 79 encoded proteins within all genomes subjected to our analysis (**fig. 4, supplementary table 5,** Supplementary Material online). In addition, we included four *DXS*-like genes that were previously identified in the moss *Physcomitrella patens* to reconstruct the evolutionary history of all 83 target genes during Angiosperm radiation [52,118].

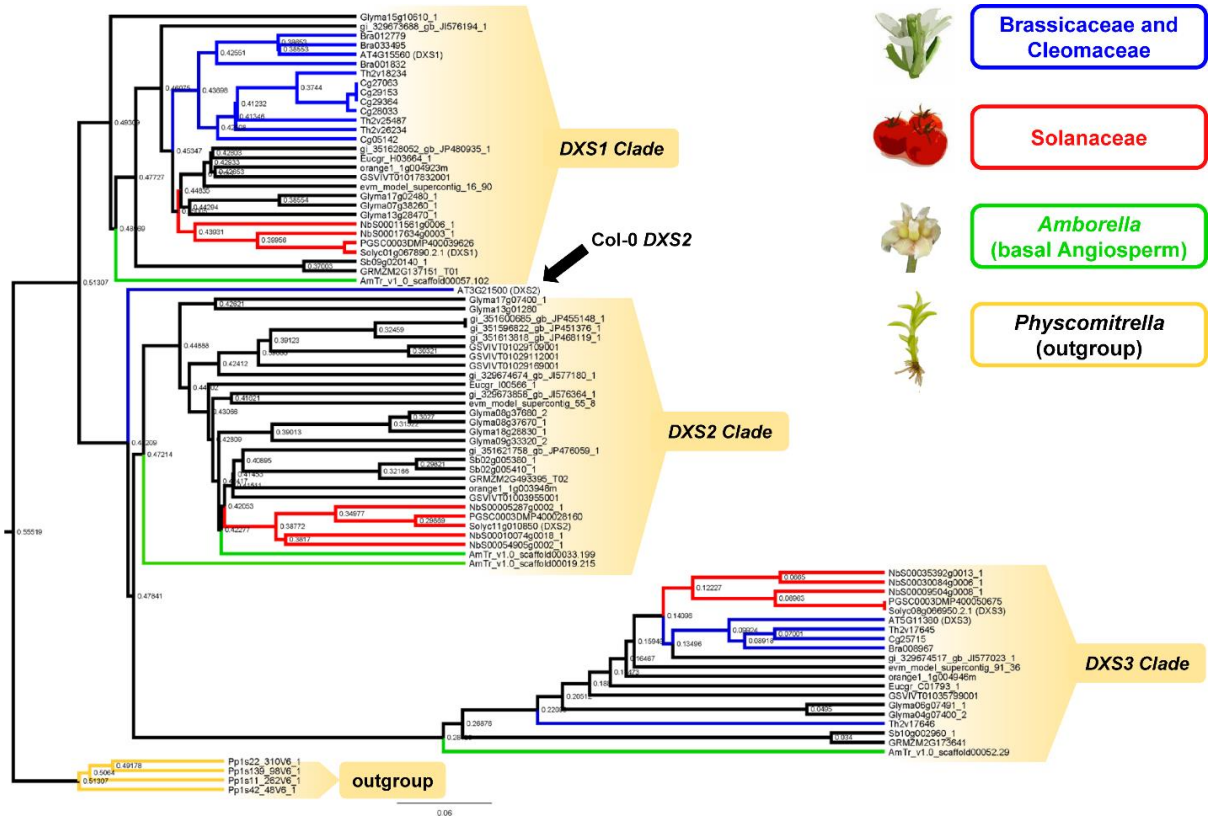


FIG. 4.—Phylogenetic relationships among 83 DXS-like proteins. Brassicaceae and Cleomaceae are marked in blue. Solanaceae are marked in red. Proteins encoded in the basal Angiosperm Amborella are marked in green. The moss *Physcomitrella* comprises the outgroup and is marked in yellow. *DXS*-like genes group in three distinct clades since the origin of Angiosperms. Notably, all analyzed Brassicaceae have lost *DXS2*-like genes. However, the model plant *Arabidopsis* contains one highly diverged member of clade two that groups closer to clade one than to any other clade two homologs (marked by black arrow).

As previously reported, the *Physcomitrella DXS*-like genes form a monophyletic clade that groups distant to all other analyzed Angiosperm target genes. Strikingly, we have identified multiple gene family members that remained un-characterized within all other analyzed annotations except maize. Within the Angiosperm clades, we found that *DXS*-like genes always group in three distinct clades that form monophyletic groups rooted by basal members present in the *Amborella* genome (**fig. 4**), which is commonly placed at or near the base of the flowering plant lineage [71,119]. Within those clades, we observed grouping of closely related species consistent with the evident phylogenetic relationships of these species as a whole (**fig. 4**). However, our analysis also revealed clade-specific differences. First, we did not detect any proteins grouping to clade three within the non-redundant *Cannabis* transcriptome (**supplementary table 5,** Supplementary Material online). Second, whereas most organisms encode at least one member of every clade, Brassicaceae and Cleomaceae have lost *DXS*-like genes belonging to clade two (**fig. 4**, **supplementary table 5,** Supplementary Material online). Interestingly, the model plant *Arabidopsis* forms the only exception, because it possesses

one *DXS2* locus (At*DXS2*/At*DXL1* or AT3G21500) that is highly diverged from any other members present in that clade (marked by black arrow in **fig. 4**). Our analysis revealed that *AtDXS2* forms a basal sister to all other clade two members and groups closer to its clade one ortholog present in the basal Angiosperm *Amborella* compared to any other clade two members. Note that first evidence supports functional specialization at both the expression and biochemical level within the plant DXS family in *Arabidopsis (*see introduction section) [52]. In this context, the authors reveal the occurrence and putative relevance of lineage-specific gene duplications. Therefore, the plant *DXS* family emerges as an interesting model to examine the molecular evolutionary basis of plant secondary metabolism diversification, giving rise to further investigation of this gene family in a broader phylogenomics framework, as we presented in this study.

Next, we assessed the contribution of gene and genome duplication to *DXS*-like gene family composition among four further genome annotations with most accurate determination of ohnolog blocks (**table 5**). To our knowledge, the contribution of genome duplication to *DXS*-like family evolution has previously not been assessed to that extend. For *A. thaliana*, *B. rapa*, *T. hasslerania* and *G.* max, we found 24 *DXS*-like genes in total, organized in eight duplicate groups (defined as set of genes comprising descendants from one distinct ancestral singleton due to one or more rounds of duplication) and distributed across all three *DXS*-like clades (**fig. 4**, **supplementary table 5**, Supplementary Material online). Strikingly, 100% of those are due to ancient polyploidy events, either directly when forming pairs (WGD) or triplets (WGT) of ohnolog copies or indirectly when forming tandem- or transposition duplicates (GTD) of ohnolog group members (**table 5**). In *A. thaliana*, for example, *DXS1* (AT4G15560) and *DXS2* (AT3G21500) form the ohnolog duplicate gene pair A15N013, dating back to the At-α WGD event [8,72] (**tables 1A, B**). The encoded proteins share 78.8% of protein sequence similarity (**table 5**). Likewise, the corresponding genes are differentially expressed and pleiotropic (see introduction section; i.e. involved in terpenoid biosynthesis, plastid development and trichome formation [34,52,53,120,121]). Further analysis of *DXS3* (AT5G11380) indicated its putative origin due to gene transposition duplication of *DXS1*. First, both genes form a highest-scoring sequence pair based on our BLAST analysis after removal of self-hits in *Arabidopsis* (see Materials & Methods section). Second, both genes are embedded in a non-syntenic genomic regions that contain remnants of transposon-like sequences (**fig. 5**). Considering the increased phylogenetic distance between this pair of genes and its reduced degree of protein sequence similarity (**table 5**) compared to the pair of DXS1/DXS2 (**fig. 4**), this illustrates that genetic versatility within the *Arabidopsis DXS* family was further leveraged by a gene transposition duplication (GTD). Taken together, these results give rise to the onset of functional diversification of the A15N013 ohnolog pair following the At-α WGD event in Brassicaceae (see Discussion section). Similarly, short sequence duplication may have contributed to functional diversification of *DXS*-like genes. Based on those results, we further assessed the impact of various duplication modes to all other identified *DXS*-like genes in all analyzed genome assemblies including analysis of expression and sequence diversity.

TABLE 5.- Overview of gene and genome duplication responsible for *DXS* -like cluster extension; shown are all target genes for four genomes[A]

Species	Gene Identifier	Clade	Origin of Duplication	Duplicate Group	Similarity[B] to Duplicate copy	Identity[B] to Duplicate copy
A. thaliana	AT3G21500	1	At-α WGD (A15N013)	1	78.8%	72.5%
A. thaliana	AT4G15560	1	At-α WGD (A15N013)	1		
A. thaliana	AT5G11380	3	GTD (AT4G15560)[C]	1	68.6%	53.3%
B. rapa	Bra001832	1	Br-α WGT[D]	2	77.9% - 81.8%	73.4% - 77.3%
B. rapa	Bra012779	1	Br-α WGT	2	92.0%	93.7%
B. rapa	Bra033495	1	Br-α WGT	2		
B. rapa	Bra008967	3	GTD (Bra033495)[C]	2	67.0%	52.6%
T. hasslerania	Th2v17645	3	Tandem (Th2v17646)[E]	3	4.3%	6.5%
T. hasslerania	Th2v17646	3	Tandem (Th2v17645)[E]	3		
T. hasslerania	Th2v18234	1	Th-α WGT[D]	4	92.4% - 93.3%	88.8% - 89.5%
T. hasslerania	Th2v26234	1	Th-α WGT	4	87.6%	83.9%
T. hasslerania	Th2v25487	1	Th-α WGT	4		
G. max	Glyma07g38260	1	*Glycine* WGD (I)	5	94.4%	91.7%
G. max	Glyma17g02480	1	*Glycine* WGD (I)	5		
G. max	Glyma15g10610	1	*Glycine* WGD (I)	5	51.7%	48.9%
G. max	Glyma13g28470	1	*Glycine* WGD (I)	5		
G. max	Glyma04g07400	3	*Glycine* WGD (II)	6	97.0%	94.3%
G. max	Glyma06g07491	3	*Glycine* WGD (II)	6		
G. max	Glyma17g07400	2	*Glycine* WGD (III)	7	45.4%	44.8%
G. max	Glyma13g01280	2	*Glycine* WGD (III)	7		
G. max	Glyma18g28830	2	*Glycine* WGD (IV)	8	96.8%	94.1%
G. max	Glyma08g37670	2	*Glycine* WGD (IV)	8		
G. max	Glyma08g37680	2	Tandem (Glyma08g37670)	8	97.6%	96.1%
G. max	Glyma09g33320	2	Segmental (Glyma08g37670)[F]	8	92.2%	86.5%
-	-	-	-	**average**	**77.6%**	**73.4%**

[A] Analysis restricted to Genomes with most accurate identification of ohnologs due to technical limitation

[B] Based on encoded protein sequence

[C] Origin of GTD Duplicate based on lowest blastp e-value for alignment to other family members

[D] Embedded in most fractionated subgenome; similarity and identity scores shown relative to ohnologs in both other subgenomes

[E] Note significant length difference of both genes in this array; low similarity and identity scores indicate annotation error dividing one ORF into two neighbouring genes. Both values are exluded for calculation of average.

[F] Gene scored as Segmental Duplicate due to high synteny score of harbouring region while other members of duplicate group are sufficient to cover the synthenic depth of this genome (i.e. no WGT evident).

Initially, we assessed divergence levels among both pairs of DXS-like protein sequences and compared those following various modes of duplication by testing for differential and tissue-specific expression of all three *DXS*-like genes in *Arabidopsis*. Please note that glandular trichomes are absent in the model plant [122]. Notably, *DXS1* is the only member of its gene family that is annotated to "trichome specific up-regulation" in the plant ontology database (PO:0000282) [123-125]. However, we confirmed expression of all three loci in *Arabidopsis* non-glandular trichomes (and various other tissue types) based on publically available microarray data [94]. Furthermore, we uncovered consistent patterns of differential expression across several tissue types. Compared to housekeeping gene expression, *DXS1* transcript are most abundant in all analyzed tissues. The ohnolog duplicate *DXS2* shows lowest expression levels, whereas the transposed duplicate *DXS3* forms an intermediate across all analyzed tissues (**fig. 6**).

To assess and compare *DXS*-like gene family divergence in further species, we have performed two separate approaches. First, we performed *DXS*-like gene expression analysis. Second, we assessed

and compared the protein sequence identities of *DXS*-like duplicate groups due to different duplication events.

Gene duplication can result in transposition of the novel duplicate copy to a distant genomic location, leading to the presence of other cis-acting elements including promotors or enhancers that influence gene expression [126,127]. This results in sub-functionalization of segregants on the expression level. To extend the aforementioned findings concerning sub-functionalization of *DXS* genes in *A. thaliana*, we have tested expression of *S. lycopersicum* target genes in every clade. In addition to increased expression of *DXS2* in trichomes and global expression of *DXS1* that was previously made evident [51] (**fig. 7**), we have uncovered that transcript levels of *DXS3* are almost 2-fold higher in trichomes compared to any other analyzed tissue type (**fig. 8**).

In addition to frequent changes in gene expression, recent analysis revealed an accelerated rate of amino acid changes when comparing ohnolog duplicates to their paralogs [128]. High rates of amino acid substitutions lead to decreased levels of protein sequence identities when comparing gene copies due to different duplication modes. For example, polyploidy facilitated rapid diversification of protein sequences and sub-functionalization on a biochemical level in several cases, including glucosinolate biosynthesis, resistance proteins of the NB-LRR type as well as L-type lectin receptor-like kinases [25,26,129]. In all three cases, functional diversification among certain duplicate pairs correlates with differentially decreased protein sequence identities when comparing "novel" gene copies due to certain duplication events. Therefore, we assessed protein sequence similarity/identity among all other seven *DXS*-like duplicate groups (i.e. sets of genes due to duplication of one distinct ancestral singleton), thereby screening for indications of putative sub- or neo-functionalization (**table 5**). Values for protein sequence similarity (identity) range from 45.4% (44.8%) (*G. max*, duplicate group 7) to 96.8% (94.1%) (*G. max*, duplicate group 8). In summary, DXS-like proteins share an average of 77.6% (73.4%) for sequence similarity (identity) among all groups, thereby reaching a cumulative divergence level similar to that observed in *A. thaliana*, for which data on differential target gene expression following gene and genome duplication are available (see above).

In summary, we have analyzed three clades of *DXS*-like genes present in every analyzed genome annotation. We have assessed differential and tissue-specific expression for two distant lineages, thereby collecting indications for putative sub-functionalization following gene and genome duplication within this group of target genes. To further support this hypothesis, more sequence and expression data are necessary from basal angiosperms in order to facilitate comparison of the observed profiles in a more ancestral state.

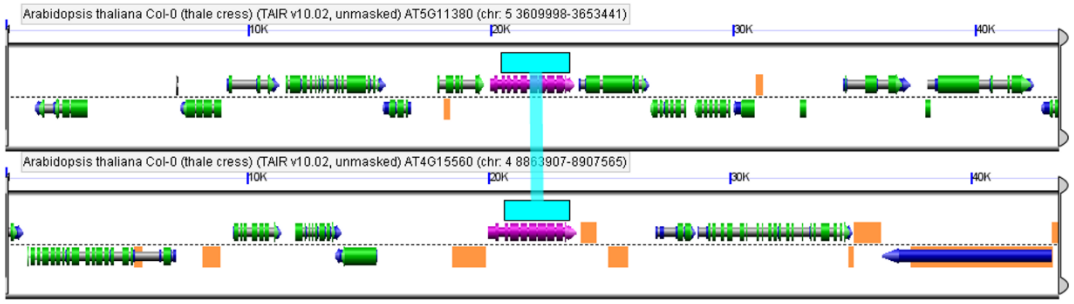


FIG. 5.—(B)LastZ two-way multiple alignment of 40kb-windows harboring the putative *Arabidopsis* gene transposition duplicate gene pair *DXS3* (AT5G11380) (upper lane, marked in purple) and *DXS1* (AT4G15560) (lower lane, marked in purple). Non-syntenic coding sequences are marked in green. Both duplicate copies form a highest-scoring sequence pair (marked in turquoise). Transposon-like sequences are marked in orange. Pseudogenes are marked in blue. Analysis can be regenerated online following the CoGe link https://genomevolution.org/r/eooq (last accessed on December 13th, 2014).

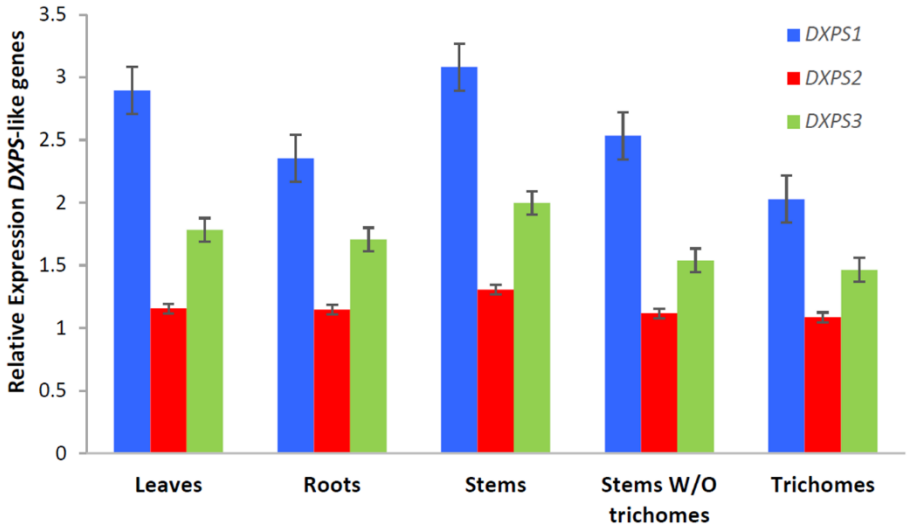


FIG. 6.—Comparative tissue-specific expression of all *Arabidopsis DXS*-like genes relative to the *bHLH* housekeeping gene. Values comprise averages of four independent ATH1 microarray experiments (Experiment ID: E-MEXP-2008, see Materials & Methods section). Notably, *DXS1* is the only member with annotation to "trichome" plant ontology (PO:0000282). The error bars represent the standard error.

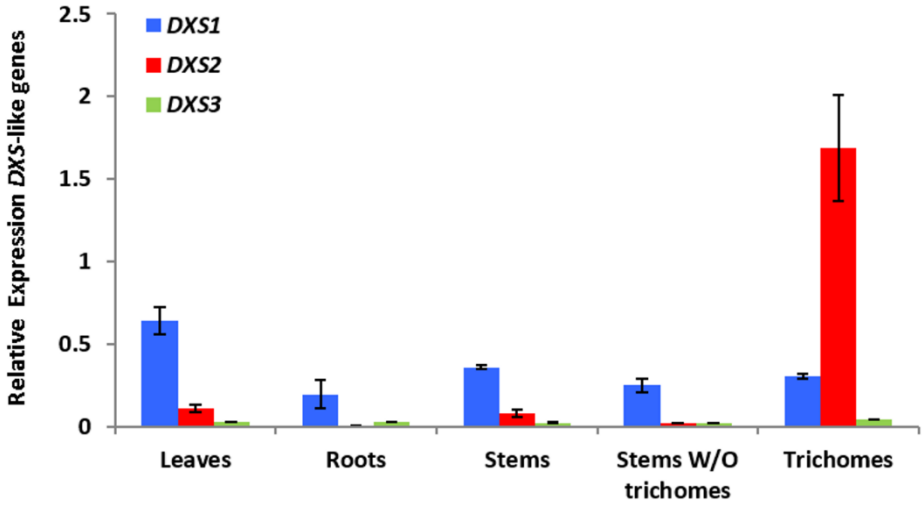


FIG. 7.—Transcript levels of *S. lycopersicum DXS*-like genes in different parts of the plant (leaves, roots, stems, stems without (W/O) trichomes, and isolated stem trichomes) relative to those of the reference gene *RCE1* (Solyc10g039370.1.1). Transcript levels were determined by real-time qPCR with four biological and three technical replicates for each biological sample. The error bars represent the standard error.

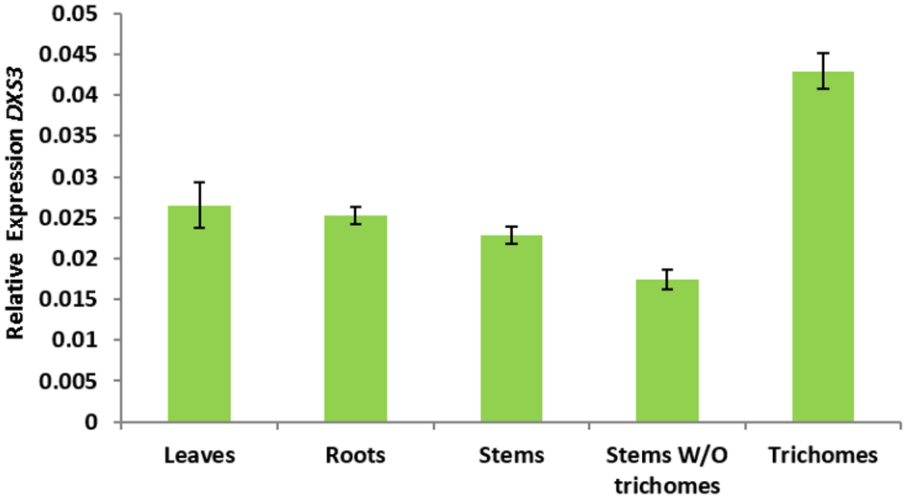


FIG. 8.—Transcript levels of the *S. lycopersicum DXS*3 gene in different parts of the plant (leaves, roots, stems, stems without (W/O) trichomes, and isolated stem trichomes) relative to those of the reference gene *RCE1* (Solyc10g039370.1.1). Transcript levels were determined by real-time qPCR with four biological replicates and three technical replicates for each biological sample. The error bars represent the standard error.

DISCUSSION

A combination of synteny, sequence similarity and protein domain modelling facilitates large-scale gene identification and novel annotations in all modules of terpenoid biosynthesis

In a genome informatics approach, we combined a novel and easy-to-follow meta-method for gene and supergene cluster identification with a custom pipeline for *de novo* protein annotation for large-scale identification of biosynthetic elements associated with plant secondary metabolism. The method provided in this article is novel because it integrates information provided by the genomic location of target genes to information on sequence homology and to the information on encoded protein domain composition. Our method can be applied to a collection of final stage genome annotations, early-stage gene-space assemblies as well as non-redundant transcriptomes, thereby facilitating uniform standards for gene identification. In this context, we analyzed various datasets of different quality for a flowering-plant wide comparative survey of genes building up a major pathway of plant specialized metabolism. In summary, we curated a set of genes associated with all modules of terpenoid biosynthesis and determined key factors shaping metabolic diversification in an Angiosperm-wide scale.

First, we investigated 17 species including twelve major crops. During this initial part of our analysis, we discovered previously uncharacterized genes of the (a) *TERPENE SYNTHASE*- as well as the (b) *DXS*-like types in all species except *Arabidopsis* and *Eucalyptus*. These gene families have often been subjected to species-specific analysis in the past because they are involved in (a) generating a diverse set of terpenoid compounds and (b) in control of trichome density on the leaf surface, thereby providing significant economic and ecologic potential. The provided data on novel annotations of target genes in most species elucidated the power of our approach in a proof-of-concept and may act as a blueprint for future efforts to more rapidly find and clone functional core-*TPS* and *DXS*-like genes from any flowering plant in context of plant breeding and biotechnology.

Second, we identified various genes that have previously not been associated with a distinct function and established computational inferences to encoded prenyltransferases and triterpene-specific synthases across all lineages. These enzymes are commonly associated with the biosynthesis of di-, sesqui-, tri- , tetra- and polyterpenes. Assessing similarities to core-*TPS* genes in both coding sequence on the DNA level and protein domain composition, we provided indications for the common evolutionary origin shared among all three gene families. Furthermore, we monitored the underlying variation of gene copy number in a phylogenomics framework and thereby described a framework that increased genetic versatility to create the necessary basis for metabolic diversification within a timeframe of 250 MA corresponding to flowering plant radiation.

Third, our approach identified homologs of all genes currently annotated to MVA and MEP pathways including *DXS*-like genes in *Arabidopsis* across all analyzed genomes. Large-scale annotation of genes employed by those pathways has to date not yet been made available for every analyzed species except *Arabidopsis*, tomato and potato. In this context, our study provides and important prerequisite for future efforts aiming at metabolic engineering within any of the analyzed crop lineages.

Both gene- and genome duplication mediated a dramatic increase of genetic versatility underlying modular terpenoid biosynthesis in all species

In the next part of our analysis, we screened for gene and genome duplication events that affected copy number of all loci involved in distant modules of terpenoid biosynthesis across all investigated species. In this context, genetic versatility is defined as the number of homologs within one gene family. Including novel annotations of previously un-identified genes to all six modules (see above), we described a 376%-increase of terpenoid biosynthetic gene copy number ("genetic versatility") ranging from the leanest state found in the basal Angiosperm *Amborella* (50 genes) up to most versatile genotypes found for *Vitis* that has been subjected to extensive human domestication (188 genes). Merging the genetic inventory associated with all six modules, we revealed that this increase is driven by a combination of gene and genome duplication across 250 MA corresponding with the radiation time of flowering plants. However, individual differences apply when considering single terpenoid biosynthetic modules separately. To our knowledge, this is the to-date most intensive and systematic study of plant gene family expansion that influenced metabolic diversification in a phylogenomics framework.

Please note that segmental duplications are excluded from our analysis. In this context, we acknowledged an error rate due to false-positive scoring as ohnolog duplicates affecting ancient segmental duplications of large genomic regions. Briefly: It is currently not possible to accurately distinguish large segmental duplications from fractionated blocks due to genome multiplications in all cases. Likewise, very short segmental duplications with high degree of fractionation may be accidentally scored as a series of distant gene transposition duplication. This is mainly due to technical reasons, because the SynMap algorithm controls scoring of synteny merely based on a function of collinear genes in a certain density as previously described [4,87,88]. However, most segmental duplications that did not emerge roughly at the same time than any of the investigated genome duplications will display significantly different averages for ka/ks values, and are therefore excluded from synteny analysis due to the cut-offs thresholds applied in the SynMap preferences (see Materials & Methods section).

Enzymes catalyzing the committed step of end product biosynthesis are more often encoded by supergene clusters due to tandem duplication

We highlighted a consistently asymmetric distribution of supergene clusters across all terpenoid biosynthetic modules. Generally, core terpene synthases as well as triterpene-specific synthases comprise enzymes catalyzing the committed step for biosynthesis of designated end products (mono-, di-, sesqui-, and triterpenes). We revealed that those are most enriched for tandem duplicate copies across all analyzed genomes. Please note that in alignment with this these findings, the role of syntenic core-*TPS* supergene clusters that include adjacent loci involved in different modules was recently made evident for diversification of terpenoid pathway assembly during radiation of various Angiosperm clades (see below) [32]. Moreover, it has become evident that single-featured polymorphisms affecting those genes are sufficient to alter, amongst others, herbivore behavior in otherwise isogenic lines [130-132]. In the opinion of the authors, such processes may have correlated with human efforts of plant domestication and crop breeding in multiple cases. It seems possible that sub-functionalization following tandem duplication of target genes influenced key traits (i.e. scent, taste), making the plant more suitable for further selection. This hypothesis is supported by the high target gene count for highly domesticated species with high content of terpenoids (like *Vitis*, *Cannabis* and *Lactuca*). Although *Eucalyptus* possesses the highest

terpenoid biosynthetic gene count among all species analyzed in this study, it did not undergo major processes of human domestication [78]. However, several herbivores are known to respond differently to *Eucalyptus* inter- and intraspecific variation of secondary metabolite profiles with potential effects on target gene evolution [133]. Please also note that intensive domestication may also lead to a low *TPS* gene count in some cases, for example as a result of selection towards different key traits negatively influenced by genes in linkage disequilibrium to *TPS* genes [134].

Dosage-dependent enzymes in modules mediating intermediate reaction steps are more often encoded by ohnolog duplicates – Introducing a two-step model for rapid plant pathway diversification

Compared to the above-mentioned asymmetric distribution of tandem duplicate copies across all subsets of genes involved in terpenoid biosynthesis, we reported opposite tendencies for retained ohnologs. We made evident that multi-gene family members involved in the MVA pathway as well as IPP isomerases more often tend to originate from whole genome multiplication events. For the MVA pathway, ohnolog fractions greatly outreach genome-wide averages for all genome annotations except papaya. IPP isomerases comprise 100% of retained ohnologs in Brassicaceae, Cleomaceae as well as *Glycine*. These groups of gene copies are due to duplication of a distinct ancestral singleton ("duplicate groups") but encode enzymes involved in different terpenoid functional modules, working together by catalyzing neighboring reactions and isomerization of intermediate products (IPP or MVA/MEP modules). According to the gene balance hypothesis, duplicate loci are preferentially retained when functioning together in a dosage-dependent way [6,135]. In this context, we showed an asymmetric ohnologs distribution among the modules acting up- and downstream of core terpene scaffold synthesis.

Based on those findings, we hypothesize a two-step mechanism for the rapid plant pathway- and trait diversification observed in nature. This proposed mechanism depends on both gene- and genome duplication and affects different groups of genes at different times. In a first step, ancient polyploidy plays a paramount role by mediating the described expansion of certain genetic networks involved in plant primary metabolism (like MEP/MVA and *IDI* loci, see **fig. 1**), thereby creating a certain degree of "pathway redundancy". Due to stoichiometric effects, the following post-polyploidy rate of plant survival depends on parallel retention of most (if not all) duplicated genes present in affected metabolic modules. Both functional diversification of ohnolog duplicates and/or incomplete module retention may lead to detrimental effects due to altered fractions of primary metabolite concentrations, as previously hypothesized and backed up by gene network analysis in context of mustard family evolution [135,136]. In a second step, more recent, short sequence duplications (including tandem and gene transposition duplication) creates an extended pool of trans-acting elements (like, for example, additional core-*TPS* or *DXS* genes). Since increased copy number of those genes does not lead to detrimental effects due to stoichiometry as described above, functional diversification may create extended capabilities to catalyze biosynthesis of extended product ranges (novel functions). The aforementioned polyploidy-induced primary module duplication created a superabundance of primary metabolites, thereby providing a "playground" for the evolution of novel functions catalyzed by novel gene copies due to short sequence duplicates.

In a nutshell, our results provided evidence for a partial polyploidy-driven expansion of plant secondary metabolism and strongly supported the gene-balance hypothesis for the dosage-dependent subset of involved key genes. Such trends have often been suggested for plants

[14,23,137], but solid evidence on a genetic level was to-date only available for glucosinolates and plant resistance proteins of the NB-LRR type [25,26].

Duplicate gene copies of ancestral singletons diversified in metabolic function following gene and genome duplication: the case of *DXS*-like genes

Recent analysis strongly support the concept of functional specialization following gene duplication as the evolutionary fate explaining retention of the duplicated gene pair DXS1/DXS2 in *Arabidopsis* [52]. Based on this approach, we performed follow-up analysis of *DXS*-like gene family evolution on a broader phylogenomics scale. In summary, we showed that certain sets of duplicate gene copies that descend from duplication of one ancestral singleton (i.e. duplicate groups) contain genes encoding different enzymes for the same pathway in *Arabidopsis* and tomato. Some of those convey pleiotropic effects due to published annotation to different traits (i.e. control of trichome density and terpenoid biosynthesis). Additionally, we identified common protein motifs present (a) within and (b) across different modules of terpenoid biosynthesis. We conclude an expansion of isoprenoid pathways by gene family diversification following gene and genome duplication, thereby resulting in the complex, modular architecture of terpenoid biosynthesis and the plethora of produced compounds observed across the Angiosperm clade. Because supergene clusters tend to be younger than genes preferentially retained after ancient polyploidy events [11,138], ohnologs are likely prone to acquire additional roles over time as previously described (sub- and neo-functionalization) [135,139,140].

Moreover, we have found evidence for the preferential (i.e. above-average) retention of *IPP* genes following various independent, successive polyploidy events for the Brassicaceae-Cleomaceae sister group system [83]. Similar to DXS-like proteins, IPP isomerases convey pleiotropic functions because they are relevant for the biosynthesis of other isoprenoid compounds beyond plant terpenoid biosynthesis. They also have been brought in connection with plant development in *Arabidopsis*, thereby mediating a check-point for primary metabolism (e.g. hormones) and different branches of specialized metabolism [141-143]. The observed trend of *IDI* over-retention is consistent for species-specific WGT events (Th-α for *Tarenaya* and Br-α for *Brassica*) as well as for the more ancient At-α WGD event shared by all Brassicaceae [8,17,72,76,144]. Similarly, we observed a rising *IDI* gene counts following soybean polyploidy. We concluded a preferential retention of this gene family following polyploidy that might be due to reported dosage-sensitivity (see Introduction section) and is likely visible especially in the aforementioned genomes due to their high levels of syntenic depth (i.e. high levels of genome multiplicity due to more successive WGDs/WGTs compared to other genomes). However, the case of *Arabidopsis* provides an exception which might be due to its reductive genome state that has been previously reported for the genus of the model plant [145].

Furthermore, our results further support the concept of sub-functionalization among *DXS*-like genes on a broader phylogenomics scale than previously reported [52]. In addition, we assessed and compared the differential impact of various duplication modes (i.e. WGD and short sequence duplication) to functional diversification of *DXS*-like genes, thereby uncovering novel aspects shaping target gene family evolution. Similar to *IDI* loci, *DXS*-like genes have been associated with more than one trait. Two among three *DXS*-like genes in *Arabidopsis* comprise the retained ohnolog pair A15N013, dating back to the At-α that is shared by all Brassicaceae. While both *DXS1* (AT4G15560) and *DXS2* (AT3G21500) are annotated to the MEP pathway, *DXS1* is also involved in plastid development [8,51,72,120,121]. In addition to the reported control of isoprenoid profiles, functional evidence for control of trichome density on the leaf surface has been made evident [51]. Initially, we

discovered a whole new clade of *DXS*-like genes with members in Solanaceae and Brassicaceae including *Arabidopsis*. Next, we scored the contribution of ohnolog retention to the set of target loci identified the Brassicaceae-Cleomaceae sister group system as well as the legume *G. max*. We showed that all target genes within the aforementioned four genome annotations date back to ancient polyploidy events, either directly by comprising ohnolog duplicate groups or indirectly by comprising tandem- or transposition copies of ohnologs. Furthermore, we unraveled phylogenetic relationships within the target gene family that groups to three clades of encoded DXS-like proteins. We brought those clades in connection with a expression polymorphisms following gene- and genome duplication in tomato and the model plant *Arabidopsis*, thereby elucidating another case of putative sub-functionalization following duplication.

Modified terpenes: Future work or going beyond the plant terpenoid biosynthetic module

Recently, Boutanaev et al. published a very conclusive investigation of core-*TPS* gene diversification across an evolutionary timeframe similar to the scope of our study (see introduction) [32]. The authors defined an (incomplete) "terpenome" that merely consists of (some, but not all present) core-*TPS* genes and supergene clusters that consist of both core-*TPS* and *CYP*-like genes. *CYP*-like genes encode cytochrome P450 enzymes that catalyze downstream modifications of various secondary metabolite core structures including alkaloids, glucosinolates and terpene post-modification reactions [146-148]. The authors infer an important role of (micro)synteny and *TPS*/*CYP*-locus linkage disequilibrium for terpenoid pathway assembly in plants, and suggest a differential mechanism of trait diversification in monocots and dicots [32]. However, terpenoid biosynthesis in plants is modular because it consists of more than just the core-*TPS* gene family (**fig. 1**). Likewise, *CYP*-like genes are not the only family mediating terpene post-modification reactions [34]. Due to our large-scale annotation of terpenoid biosynthetic genes across all pathway modules within 17 representative genomes, our results provide a valuable basis for future efforts to further investigate the role of synteny and genetic linkage disequilibrium in context of a more complete "terpenome". This includes the possibility to better elucidate the effects of genetic co-segregation with many other gene families that convey terpene downstream modifications, similar to the aforementioned case study published by Boutanaev et al. [32]. Such gene families may include, for example, UDP glucuronosyltransferases and many other pleiotropic genes involved in biosynthesis of terpenoids and, beyond that, various other plant secondary metabolites [149]. Ultimately, the data provided in our study will facilitate a better understanding of plant secondary metabolite pathway assembly in Angiosperms with various implications for plant breeding and metabolic engineering in context of medicine, flavor, fragrance and pigment production.

ACKNOWLEDGEMENTS AND FUNDING INFORMATION

We would like to thank three anonymous reviewers for their helpful comments and Mariam Neckzei for her help with all illustrations. This work was funded by a Netherlands Organization for Scientific Research (NWO) VIDI and Ecogenomics grant (M.E.S.).

SUPPLEMENTARY MATERIAL

Supplementary files 1-5 are available at PLoS ONE online (http://www.plantphysiology.org/, last accessed on December 13th, 2014).

Supplementary Table 1.— HMM-driven protein domain prediction among the extended set of *Arabidopsis* terpenoid biosynthetic genes

Supplementary Table 2.— Cross-referencing of 1, 904 target genes among 17 genomes to a specific subset of genes acting in the terpenoid biosynthetic module

Supplementary Table 3.— Species-specific relative size of terpenoid biosynthetic modules. Numbers are quotients of the module-wise gene count of terpenoid biosynthetic pathways and the number of all protein-coding genes within the whole genome. Species with highest and lowest relative pathway size among all analyzed species are color-coded as indicated in the legend.

Supplementary Table 4.— Species-wise distribution of *DXPS*-like genes among three subgroups

Supplementary Table 5.— Comparison of genome-wide numbers of tandem/ohnolog duplicates to numbers among subsets of the terpenoid biosynthetic module, including p-values from Fisher's exact test on count data. Red indicates absence of tandem/ohnolog duplicates. Green indicates significant enrichment among terpenoid biosynthetic genes compared to background with threshold of 0.01.

A novel approach for multi-domain and multi-gene family identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants

Johannes A. Hofberger[1,2], Beifei Zhou[2,3], Haibao Tang[4,5], Jonathan D.G. Jones[6] and M. Eric Schranz[1]*

[1] Biosystematics Group, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708 PB Wageningen, Gelderland, The Netherlands
[2] Chinese Academy of Sciences/Max Planck Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai 200031, PR China
[3] Heidelberg Institute for Theoretical Studies-HITS, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Baden-Württemberg, Germany.
[4] Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, PR China
[5] J. Craig Venter Institute, 4120 Torrey Pines Road, La Jolla, CA 92037, USA
[6] The Sainsbury Laboratory, Norwich Research Park, Colney Lane, Norwich, Norfolk NR4 7UH, UK

BACKGROUND: Recent advances in DNA sequencing techniques resulted in more than forty sequenced plant genomes representing a diverse set of taxa of agricultural, energy, medicinal and ecological importance. However, gene family curation is often only inferred from DNA sequence homology and lacks insights into evolutionary processes contributing to gene family dynamics. In a comparative genomics framework, we integrated multiple lines of evidence provided by gene synteny, sequence homology and protein-based Hidden Markov Modelling to extract homologous super-clusters composed of multi-domain resistance (R)-proteins of the NB-LRR type (for NUCLEOTIDE BINDING/LEUCINE-RICH REPEATS), that are involved in plant innate immunity.

RESULTS: To assess the diversity of R-proteins within and between species, we screened twelve eudicot plant genomes including six major crops and found a total of 2,363 *NB-LRR* genes. Our curated R-proteins set shows a 50% average for tandem duplicates and a 22% fraction of gene copies retained from ancient polyploidy events (ohnologs). We provide evidence for strong positive selection and show significant differences in molecular evolution rates (Ka/Ks-ratio) among tandem- (mean = 1.59), ohnolog (mean = 1.36) and singleton (mean = 1.22) R-gene duplicates. To foster the process of gene-edited plant breeding, we report species-specific presence/absence of all 140 *NB-LRR* genes present in the model plant *Arabidopsis* and describe four distinct clusters of *NB-LRR* "gatekeeper" loci sharing syntenic orthologs across all analyzed genomes.

CONCLUSION: By curating a near-complete set of multi-domain R-protein clusters in an eudicot-wide scale, our analysis offers significant insight into evolutionary dynamics underlying diversification of the plant innate immune system. Furthermore, our methods provide a blueprint for future efforts to identify and more rapidly clone functional *NB-LRR* genes from any plant species.

KEYWORDS: Systems biology, big data, comparative genomics, molecular evolution, innate immunity

***Author for Correspondence:**
M. Eric Schranz | Biosystematics Group | Wageningen University & Research Center | Wageningen, The Netherlands | Tel. +31(0)317-483160 | email: eric.schranz@wur.nl

INTRODUCTION

Plants have evolved a two-layered innate immune system against microbial and other pathogens [271]. In a first layer of defense, transmembrane pattern recognition receptors (PRRs), usually with extracellular LRR-type domains, recognize pathogen associated molecular patterns (PAMPs) and initiate downstream signaling events including defense gene induction [272], and lead also to cell wall reinforcement by callose deposition and SNARE-mediated secretion of anti-microbial compounds [273, 274]. This is referred to as PAMP- or pattern-triggered immunity (PTI).

Successful pathogens have evolved virulence factors (effectors) that act in the apoplast or inside the host cell to overcome PTI [275]. As a second layer of the innate immune response, many host plant lineages evolved intracellular R-proteins of the NB-LRR type that respond to virulence factors, either directly or through their effects on host targets [276]. Plants producing a specific R-gene product are resistant towards a pathogen that produces the corresponding effector gene product (avirulence factors encoded by *Avr* genes), leading to gene-for-gene resistance [277]. This is referred to as effector-triggered immunity (ETI). Rounds of ETI and effector-triggered susceptibility (ETS) due to novel *Avr* genes on the pathogen side can result in an evolutionary arms-race, generating a "zigzagzig" amplitude of host resistance and susceptibility [271].

R-genes play a major role in defending crops against microbial infection and thus are of great interest in plant breeding programs and efforts to meet increased global food production. In potato, for example, R-proteins of the NB-LRR type confer resistance to the oomycete *Phytophthora infestans*, a hemibiotrophic pathogen that causes late blight [278, 279]. In *Arabidopsis*, R-proteins of the NB-LRR type have been studied extensively in terms of molecular function, structural organization, sequence evolution and chromosomal distribution [72, 280-282]. This superfamily is encoded by scores of diverse genes per genome and subdivides into TIR-domain-containing (for TOLL/INTERLEUKIN-LIKE RECEPTOR/RESISTANCE PROTEIN; TIR-NB-LRR or TNL) and non-TIR-domain-containing (NB-LRR or NL), including coiled-coil domain-containing (CC-NB-LRR or CNL) R-protein subfamilies [77, 283]. For example, the TNL type R-protein RPP1 confers resistance to *Hyaloperonospora arabidopsidis* (downy mildew) in *Arabidopsis* [284]. Similarly, the RPS5 CNL type R-protein interacts in a gene-for-gene relationship with the avrPphB effector from Pseudomonas syringae to activate innate immune responses [285]. The TNL type R-protein RRS1, in concert with the TNL protein RPS4, confers resistance to the soil microbe *Ralstonia solanacearum* in *Arabidopsis* [286, 287]. The latter also contains a C-terminal WRKY transcription factor-like domain for DNA binding [288], increasing the number of domains common to the NB-LRR super-family to five. This number is further extended by cases with presence of additional, C-terminal domains mediating extended gene function. For example, the *Arabidopsis NB-LRR* locus *CHILLING-SENSITIVE3* (*CHS3* or *DAR4*) encodes a mutated allele of a C-terminal LIM-type domain-containing TNL protein, leading to constitutive activation of defense responses and increased chilling susceptibility [289]. The *NB-LRR* gene *ADR1-L1* encodes an N-terminal RPW8-domain whose functional importance has previously been reported [290]. However, many *RPW8*-like genes encode transmembrane proteins without NB-ARC-domain but impact on resistance to powdery mildew in *Arabidopsis* [291-293].

TIR- and non-TIR NB-LRR protein clusters share a conserved central NB-ARC-domain including three sub-domains (NB, ARC1, and ARC2). Together, these confer ATPase function [294]. The C-terminal part of NB-LRR proteins harbors a leucine-rich repeat (LRR)-domain for recognition of intracellular effector molecules upon infection, leading to a conformational shift within the NB-

ARC-domain [295] upon recognition of the corresponding effector or a change in the surveyed plant protein. In the case of the soybean (*Glycine max*) CNL-class R-protein RPSk-1, defense genes are induced upon *Phytophthora sojae* effector recognition. This includes differential regulation of transcription factor activity as previously proposed [296-298].

A genome-wide comparison of multi-gene families in *A. thaliana* Col-0 revealed a high frequency of gene duplication among the *NB-LRR* gene superfamily and impact on genomic distribution [299]. For example, 63% of all reported *NB-LRR* genes are members of tandem arrays in both *A. thaliana* (101/159) and *A. lyrata* (118/185) [72]. Notably, *NB-LRR* loci are subjected to positive selection [300]. In this context, [72] re-assessed rates of molecular evolution for both sets of tandem and non-tandem (singleton hereafter) genes and found significant differences in selection rates. In this study, we went a step further by distinguishing the frequency of tandem and ohnolog duplicates to *NB-LRR* gene family expansion and diversity within a wider phylogenomics perspective, thereby covering an evolutionary timeframe of approximately 100 million years (MA hereafter) that corresponds to the radiation of core eudicots [41, 301]. We compared the average rates of molecular evolution for singleton, tandem and ohnolog duplicate R-genes. We further provided evidence for strong positive, but significantly different, selection rates acting on all copy classes of *NB-LRR* duplicates, illustrating the impact of gene and genome duplication to the diversification of plant key traits across approximately 100 MA of genome evolution.

To elucidate the dynamics underlying pathway and trait evolution across multiple lineages, it is of paramount importance to identify and distinguish the complete set of orthologous and paralogous loci present within multiple genome annotations in a phylogenetic framework [23]. Two homologous genes are referred to as orthologs if they descend from one locus present in the common ancestor lineage and diverged due to speciation [62]. By definition, orthologous genes are embedded in chromosomal segments derived from the same ancestral genomic region, thus sharing high inter-species synteny between closely related lineages [26]. In contrast, paralogous loci refer to homologs within one lineage and are due to, for example, tandem, transposition- or whole genome duplications (WGDs) [29, 100]. Large-scale synteny is not observed for paralogs derived from small-scale events like tandem and transposition duplication. In contrast, paralogs derived from WGDs are located within intra-species syntenic genomic blocks, and can be referred to as ohnologs or syntelogs [35, 63].

Recent analysis of genome-wide ohnolog distribution have revealed a common history of ancient, successive polyploidy events that are a common feature of genome evolution shared by all flowering plant linages [26]. For example, the *Arabidopsis* lineage underwent at least five polyploidy events that we know of, two preceding and three following angiosperm radiation [37]. The most recent WGD event for the *Arabidopsis* lineage is termed At-α and shared by all Brassicaceae including the extant sister clade Aethionemeae [43, 84]. The older At-β WGD is shared by most species in the order Brassicales, but occurred after the papaya lineage split [45, 46]. The more ancient At-γ event is a whole genome triplication (WGT) that is shared by most eudicots including all Rosids, all Asterids (including tomato), Grape (Vitales) and more distant and basal clades such as Gunnera manicata (Gunnerales) and Pachysandra terminalis (Buxales) [47, 48]. In addition to ancient polyploidy events, more recent, species-specific WGDs/ WGTs occurred in various lineages, such as genome triplications in *B. rapa* [52] (Br-α WGT), *T. hasslerania* (Th-α WGT) [45, 53] and the tomato genome (*Solanum lycopersicum*) [144]. Hence, the "syntenic depth" (defined as the level of genome multiplicity expected from the multiplication of successive WGDs/WGTs) of the *Brassica rapa*

genome is 36x compared to the putative 1x eudicot ancestor (3x due to At-γ, 2x more due to At-β, 2x more due to At-α and finally 3x due to Br-α). Under consideration of two polyploidy rounds at or near the origin of angiosperms as well as 2x at or near the origin of seed plants [37], the syntenic depth of the *B. rapa* genome would be expected to be increased to 144x ("rho-mu-delta-ploidy" genome).

Polyploidy events also influence other kinds of duplication, thereby creating a network of factors with mutual influence. In *Brassica rapa* (that underwent an additional species-specific genome triplication event, see above), arrays of tandem duplicate (TD) genes (TAR genes) fractionated dramatically after the Br-α WGT event when compared either to non-tandem genes in the *B. rapa* or to tandem arrays in closely related species that have not experienced a recent polyploidy event [263]. Errors in DNA replication due to template slippage or unequal crossing-over can result in tandem duplication (TD), producing tandem arrays (TAR) of paralog genes in close genomic proximity [90]. It is known that TAR genes are enriched for genes functioning in biotic and abiotic stress [31]. For disease resistance, there are multiple taxa with an evident impact of TD to trait evolution, including members of Brassicaceae [92], Solanaceae [91] and Fabaceae [93].

Evidence is accumulating for the connection of ancient WGD events to birth and diversification of key biological traits. It was made evident that WGD is often followed by a genome-wide process of biased fractionation that preferentially targets one sub-genome to retain clusters of dosage-sensitive genes often organized in functional modules [85, 88, 302]. In Brassicaceae, WGD shaped the genetic versatility of the glucosinolate pathway [44], a key trait mediating herbivore resistance and thus highly connected to reproductive fitness of the population. Similarly, starch biosynthesis in grasses, origin and diversification of seed and flowering plants as well as increased species survival rates on the Cretaceous–Tertiary (KT)-boundary are hypothesized to be linked to ancient polyploidy events [41, 86, 87, 89, 303].

In this study, we utilized an iterative approach by combining BLAST, HMM modeling and genomic contextual information provided by synteny to determine the fraction of tandem- and whole genome duplicate copies among all (re)annotated full-length *NB-LRR* genes across twelve species in the context of a phylogenomics perspective, based on uniform standards facilitating all comparisons. After utilization of duplicate classes, we assessed and compared rates of molecular evolution to describe a complex interplay of TD and WGD events driving R-protein superfamily extension, both of which expanded the evolutionary playground for functional diversification and thus potential novelty and success.

MATERIALS & METHODS

Hardware resources and software prerequisites

All analysis were performed on a commercial Lenovo ultrabook, model Thinkpad X1 Carbon with 8GB RAM and Intel Core i7 3667U CPU (two physical / four virtual cores). The in-house developed perl and python scripts required perl (strawberry v5.18) and python (v2.7) libraries including bioperl (v1.6.910) and biopython (v1.63) modules. The iprscan_urllib.py-script for HMM-based domain annotation (see below) required SOAPy, NumPy and urllib python modules. For BLAST screens, we employed the stand-alone command line version of NCBI BLAST 2.2.27+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/, last accessed on December 13th, 2014) [66]. For platform-independent coupling and parallelization of all employed scripts and programs, we wrote batch wrappers using the notepad++ editor (www.notepad-plus-plus.org, last accessed on December 13th, 2014).

Genome annotations

The Complete sets of representative genes and proteins for twelve genome annotations were downloaded using phytozome (www.phytozome.net, last accessed on December 13th, 2014) [186]. We included *Arabidopsis thaliana* TAIR10.02 [69], *Arabidopsis lyrata* v107 [269], *Eutrema parvulum* v2 [304], *Brassica rapa* v1.1 [51], *Carica papaya* v0.5 [46], *Citrus sinensis* v1 [187], *Vitis vinifera* v2 [48], *Solanum tuberosum* v3.2.10 [190] and *Solanum lycopersicum* v2.40 [144]. *Aethionema arabicum* v0.2 [43], *Tarenaya hasslerania* v4 [53] and *Nicotiana benthamiana* v0.42 [191] genome annotations were made available by the authors.

Confirmation and extension of the *NB-LRR* multi-gene family in *Arabidopsis thaliana*

We obtained 138 *NB-LRR* genes from [72] and queried them against the TAIR10 *A. thaliana* genome annotation in a BLAST screen without e-value threshold (forward run). We extracted all target sequences and queried them back against the *A. thaliana* TAIR10 genome annotation with an applied target sequence maximum threshold of two (reverse run). After removal of self-hits, we scored loci as *NB-LRR* genes if they were part of the target sequence pool in the forward run, and aligned to a *NB-LRR* gene as defined by [72] in the reverse run. We thereby created an extended set of *A. thaliana NB-LRR* loci**.**

Determination of orthologous gene anchors

In a first step for large-scale *NB-LRR* gene identification, we determined reciprocal best BLAST hits (RBH) for both (a) protein and (b) coding DNA sequences between *A. thaliana* Col-0 and all other eleven genome annotations in a BLAST screen without e-value thresholds. Since *NB-LRR* loci can comprise up to seven different domain types connected by partially conserved linkers, the RBH approach can result in false positives due to short but highly conserved alignments of highest-scoring sequence pairs (HSPs) in functionally non-relevant (i.e. structural) parts of the protein. Therefore, we developed a python script to discard RBH pairs with a query/target sequence length ratio below 0.5 and above 2.0. We determined (c) additional, length-filtered RBH pairs for these loci within the aforementioned length ratio scope to form a third line of evidence for orthologous gene detection.

Syntelog/ohnolog determination

Calculation of pairwise syntenic blocks within and between genomes is based on integer programming [194] but implemented to an easy-to-use web interface termed CoGe package for comparative genomics (www.genomevolution.org, last accessed on December 13th, 2014) [26]. Within all genome assemblies, we determined genes sharing the same genomic context to counterparts in the *A. thaliana* Col-0 genome annotation (defined as ohnologs or syntelogs) using the DAGchainer [195] and Quota-Align [194] algorithms implemented to the "SynMap" function within CoGe. To mask noise generated by successive duplication(s) of ohnolog blocks, we applied Quota-Align ratios for the "coverage depth"-parameter that are consistent with the syntenic depth calculated for each genome annotation (defined as the level of "genome multiplicity" expected from the multiplication of successive WGDs/WGTs). For merging of adjacent syntenic blocks, we applied a threshold of n = 350 gene spacers. For ohnolog gene pairs, we calculated rates of synonymous substitutions (Ks-values) using CodeML of the PAML package [305] implemented to SynMap and applied Ks-value thresholds for ancient WGD events as previously described [35]. For determination of within-species ohnologs (comprising ohnolog blocks due to autopolyploidy events), we proceeded similar with the difference that we queried the target genomes against themselves instead of against *Arabidopsis*, using the "SynMap" function within the CoGe package for comparative genomics (parameters: gene order = relative/minimum cluster size = 5 genes/maximum chaining distance = 20 genes/scoring function = collinear). The latter parameter enforces, together with the maximum chaining distance, scoring dense arrangement of collinear gene pairs as previously described [26, 34] and provides a de facto density cutoff. Note that gene density cutoffs per Kb/ Mb would not be consistent between different synteny runs since values vary greatly across genomes, or even across different regions within the same genome as previously described [26, 34]. For the lineage-specific WGD events known for *B. rapa*, *T. hasslerania*, *S. tuberosum* and *S. lycopersicum*, we set maximum thresholds for Ks value averages of ohnolog blocks (1.5) to eliminate noise of recent duplication events. Due to minimum requirements on assembly quality that apply for usage of SynMap, it was not possible to determine the fraction of ohnolog duplicates for the current gene-space assemblies of *Aethionema*, *Carica*, *Citrus*, *Vitis* and *Nicotiana* with the available algorithms. Synteny of genes within and between lineages was visualized using the GEvo function implemented to the CoGe package for comparative genomics (see above).

Determination of anchor paralogs and generation of extended multi-gene family cluster pool

We defined the orthologous gene sets as sum of three groups of RBH pairs (first group: based on length-filtered protein pairs; second group: based on non-length-filtered protein pairs; third group: based on non-length-filtered CDS pairs; see above for length filter criteria). We merged the orthologous gene sets with the ohnolog genes set to create a set of putative homologous loci anchoring all *A. thaliana* gene families in all other analyzed genome annotations ("anchor pool"). In a next step, we performed a BLAST search without e-value threshold to query all homologous anchor genes against all twelve genomes to determine putative paralogs of the anchor genes set (forward run). We extracted all target sequences and queried them against the *A. thaliana* Col-0 TAIR10 genome annotation with a target sequence maximum threshold of two (reverse run). After removal of self-hits, we scored loci as *NB-LRR* if they aligned to any member of the extended *NB-LRR* locus family in *A. thaliana* (see above). We defined all members of this pool as anchor paralogs if they are not present within the set of homologous anchor genes (see above), thereby creating a highly accurate super-cluster of *NB-LRR* genes across twelve genomes.

Hidden Markov modeling and prediction of protein domains

The above-mentioned extended multi-gene family pool of *NB-LRR* genes is based on both sequence homology and genomic location of its members. However, we observed an erosion of synteny across lineages relative to their phylogenetic distance. Furthermore, DNA sequence homology decreases with phylogenetic distance due to wobble rules for the third codon position. Likewise, the protein sequence homology between distant multi-gene family members can decrease due to synonymous substitutions of amino acids belonging to the same chemical class (i.e. aliphatic, aromatic or indolic). Therefore, we applied a final filtering step to remove false-positives from the extended *NB-LRR* gene pool across all genomes. Using the iprscan_urllib.py script provided by the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) (https://www.ebi.ac.uk/Tools/webservices/services/archive/pfa/iprscan_rest, last accessed on December 13th, 2014), we queried every member of the extended *NB-LRR* pool to 14 algorithms that apply Hidden Markov Models for (protein domain) signature recognition (BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther and Gene3D) [196]. We overcame the one-sequence-at-a-time limitation of the EMBL server by writing batch wrappers for 25x-fold parallelization. To form a second layer of control we additionally tested all target genes for an encoded LRR-domain using the "LRRfinder"-algorithm version 2.0 available online (http://www.lrrfinder.com, last accessed on December 13th, 2014) [306]. As a result, we mapped all protein domains present in the putative multi-gene family pool onto their genes in less than a day, and discarded all false positive genes (i.e. genes not coding for at least one cluster-common domain). Final referencing of proteins with both NB-ARC- and LRR-domains was performed using a multi-vlookup array function in MS excel 2013.

Determination of tandem duplicate gene copies

To determine the fraction of tandem duplicate gene copies, we queried the complete protein annotation of every genome assembly against itself in a BLAST screen without any e-value threshold and filtered our final set of target sequences from above outside a window of n = 10 allowed gene spacers in both directions from the query sequence as previously described [31]. Likewise, we have filtered hits with genomic location on distant chromosomes/scaffolds/contigs to avoid false-positive scoring of transposition duplicates.

Multiple protein alignments

To generate multiple alignments of protein sequences, the stand-alone 64-bit version of MAFFT v7 was employed (http://mafft.cbrc.jp/alignment/software/, last accessed on December 13th, 2014) [121]. First, all NB-LRR proteins were aligned species-wise together with the HMM-generated consensus sequence of the NB-ARC-domain (available at http://niblrrs.ucdavis.edu/At_RGenes/, last accessed on December 13th, 2014) as well as the LRR-domain (available at http://smart.embl.de/smart/do_annotation.pl?DOMAIN=SM00370, last accessed on December 13th, 2014) using the command line mafft.bat –anysymbol –thread 4 –threadit 0 –reorder –auto input > output. Mesquite v2.75 (http://mesquiteproject.org, last accessed on December 13th, 2014) was used with multi-core preferences to trim MAFFT multiple alignments down to the NB-ARC- and LRR-domain blocks. Trimmed blocks were re-aligned using MAFFT with the command line mafft.bat –anysymbol –thread 4 –threadit 0 –reorder –maxiterate 1000 –retree 1 –localpair input > output.

Codon alignments and determination of substitution rates

Re-aligned NB-ARC- and LRR-domain blocks were transferred to codon alignments using the CDS sequence counterparts and the pal2nal.pl script v14 (http:// www.bork.embl.de/pal2nal/distribution/pal2nal.v14.tar.gz, last accessed on December 13th, 2014) [307]. Gaps were allowed but manually edited wherever necessary. We allowed unusual symbols (i.e. ambiguous base pair positions) and manually edited mismatches between CDS and protein sequences wherever necessary. Synonymous and non-synonymous substitution rates were determined using the "KaKs_Calculator" software (https:// code.google.com/p/kaks-calculator/wiki/KaKs_Calculator, last accessed on December 13th, 2014) [308] including ten substitution rate estimation methods (model averaging was applied). Divergence rates are generally determined between pairwise alignments of homologous sequences. For determination of average divergence rates among singletons (i.e. non-TD non-ohnolog loci), we aligned singleton *NB-LRR* loci with the best non-self BLAST hit among all target genes within one species. For determination of average divergence rates among retained ohnolog duplicates, we aligned all ohnolog *NB-LRR* loci with the best non-self BLAST hit among all ohnologs within one species. In case of ohnolog triplets, we only considered the highest-scoring sequence pair (HSP). For determination of average divergence rates among arrays of tandem duplicate *NB-LRR* genes, we aligned the first with the last member of every array, thereby covering the majority of all tandem arrays (see Results section). In a control step, we determined average divergence rates for all pairwise combinations within the largest tandem array in every species and did not find significant deviations (data not shown).

Generation and graphical editing of figures

Ideograms of plant chromosomes/scaffolds/contigs were generated using the circos package (http://circos.ca/, last accessed on December 13th, 2014) [124]. Histograms and Venn-diagrams were generated using the matplotlib package (http://matplotlib.org/, last accessed on December 13th, 2014). Other figures were generated with MS office and graphically edited using the GIMP package (http://www. gimp.org/, last accessed on December 13th, 2014).

RESULTS

Determination of protein domain-specific sub-clusters

Encoded architecture of *NB-LRR* loci in plants is variable and can comprise up to seven different domains in *Arabidopsis* (**fig. 1**). In contrast to previous studies [282], we defined functional NB-LRR proteins as composite units sharing both the NB-ARC-domain and a LRR-domain signal due to at least one repeat. Hence, TIR-NB-, LRR-only, NB-only or TIR-only proteins are not assigned as NB-LRR proteins by definition. To determine the number of *NB-LRR* loci within a given genome annotation, we combined layers of information provided by sequence homology, protein identity as well as genomic context of target genes in a custom, iterative approach using batch programming (**fig. 2**).

In the first step, we identified putative orthologous (defined as size-filtered reciprocal best BLAST hits for both protein and DNA sequences, see Materials & Methods section) and/or syntenic (based on conserved genomic context, see Materials & Methods sections) "anchor" genes (a) present in the most up-to-date genome annotations of (1) *A. lyrata*, (2) *B. rapa*, (3) *E. parvulum*, (4) *Ae. arabicum*, (5) *T. hasslerania*, (6) *C. papaya*, (7) *C. sinensis*, (8) *V. vinifera*, (9) *N. benthamiana*, (10), *S. lycopersicum* and (11) *S. tuberosum* as well as (b) aligning to any gene present in the (12) *A. thaliana* Col-0 TAIR10 genome annotation. This step resulted in an ortholog genes dataset anchoring every gene family present in *Arabidopsis* to all of the aforementioned lineages, hence providing valuable means for gene identification with any kind of target trait known in core-eudicot plants. Subsequently, we screened for genes encoding (i) an LRR-domain, (ii) a NB-ARC-domain or (iii) a TIR-domain (extended set of target genes defined in this study, see Materials & Methods section) (**supplementary file 1**, Supplementary Material online). In a second step, we screened for anchor gene paralogs present in every aforementioned genome annotation to form an extended cluster of homologous genes containing at least one of the aforementioned domains (**fig. 2**). In a third step, we applied multiple machine learning methods (see Materials & Methods section) to filter false-positives to obtain three highly accurate, functional domain cluster (NB-ARC/LRR/TIR) (**supplementary file 2**, Supplementary Material online). We performed the third (filtering) step three times (once for every aforementioned domain).

We identified 8,292 genes encoding a LRR-domain in total (**fig. 3**). Among those, the lowest number of genes containing an LRR-domain is 302 for the *C. papaya* genome annotation v0.5. In contrast, the highest number of genes encoding an LRR-domain is 1,344 for the *C. sinensis* genome annotation v1. Interestingly, both annotations share a syntenic depth of 1x representing the lowest-copy genomes subjected to our analysis (i.e. no major evidence for WGD since At-γ). We identified 2,571 genes encoding a NB-ARC-domain in total (**fig. 3**). Likewise, the lowest number was found within the *C. papaya* genome annotation v0.5 (48 loci). Again, the highest number of genes encoding a NB-ARC-domain was found in the *C. sinensis* genome annotation v1 (459 loci). We identified a pool of 1,075 genes encoding at least one TIR-domain (**fig. 3**). Similar to the aforementioned domains, the *C. papaya* genome annotation v0.5 encodes the lowest number of TIR-like loci (16 genes). In contrast to the aforementioned cases, the *A. lyrata* annotation v1.07 (but not *C. sinensis*) contains the highest number of encoded TIR-domains (170 loci). Notably, the syntenic depth of *A. lyrata* is double that of papaya or orange.

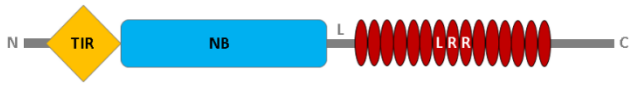
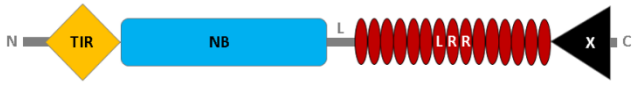
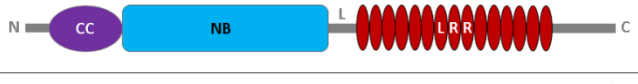
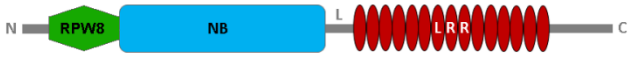
	AT1G56510 *ADR2* AT1G64070 *RLM1* AT3G44480 *RPP1*	85 / 140 (61%)
	AT5G45260 *RRS1* AT5G17890 *DAR4*	2 / 140 (<2%)
	AT1G12220 *RPS5* AT3G46530 *RPP13* AT5G43470 *RPP8*	48 / 140 (34%)
	AT4G33300 *ADR1-L1* AT5G04720 *ADR1-L2*	5 / 150 (4%)
Nucleotide binding - APAF-1 / R-gene / CED4 domain Toll / interleukin -1 receptor homology domain Powdery mildew resistance domain Coiled-coil domain Leucine-rich repeat domain Variable domain	*Arabidopsis NB-LRR* loci	

FIG. 1.— Domain composition overview for NB-LRR proteins (adapted from [77]). The NB-LRR multi-gene family comprises five common and one C-terminal variable domain. Four gene clusters in Arabidopsis are defined in this study based on domain compositions. Left: Frequent domain combinations. Middle: Well-characterized class representative in *Arabidopsis thaliana* Col-0. Right: relative abundance of target *NB-LRR* locus class in *Arabidopsis thaliana* Col-0. In case of *RRS1*, "X" refers to a WRKY TF-like DNA binding domain. In case of *DAR4* (*CHS3*), "X" refers to a LIM-domain.

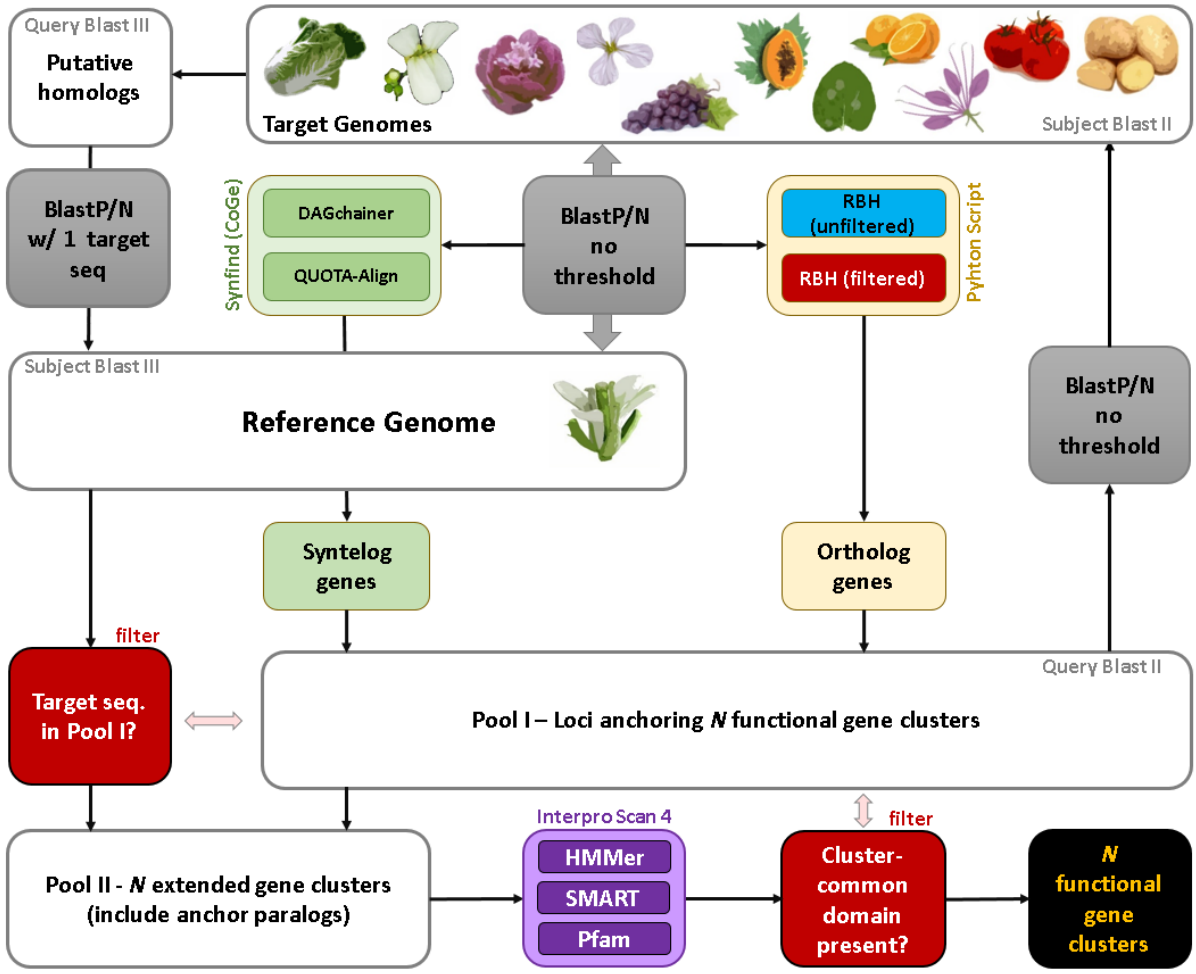


FIG. 2.— Custom flowchart showing the integrated approach for the identification of conserved multi-domain protein clusters. A bi-directional BLAST screen between a reference genome and n target genomes marks the entry point to the pipeline (grey box in the upper middle). Grey indicates BLAST screens. Red indicates filtering steps. White boxes indicate pools of FASTA-formatted sequences. Purple indicates Hidden Markov Modelling steps to predict and map protein domains. Green refers to the CoGe package for comparative genomics (see Materials & Methods section). Ochre indicates custom python scripts. Flowchart starts with two-per-genome bidirectional BLAST screens (middle) and ends with highly accurate functional protein clusters (black, bottom right).

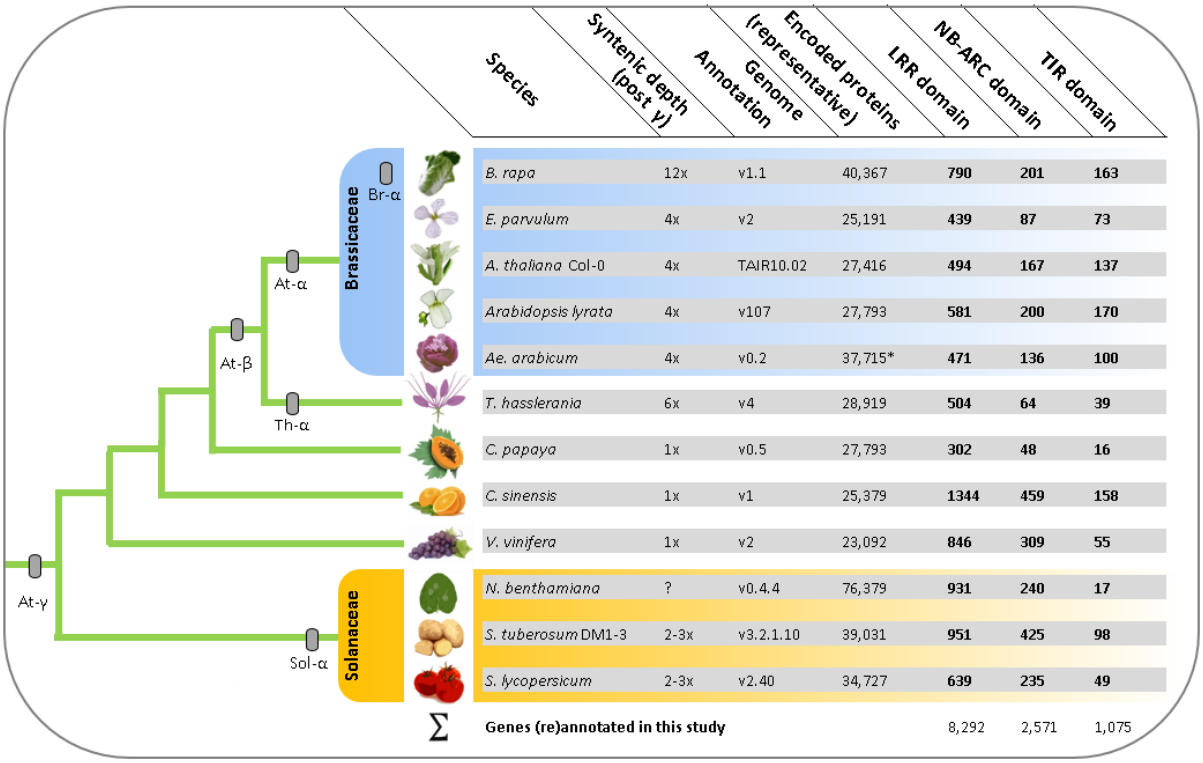


FIG. 3.— Results overview of plant Resistance (R)-gene domain (re)annotation. Shown left are phylogenetic relationships among all analyzed plant lineages and rough placement of whole-genome duplication events. Shown right are numbers of target genes per domain cluster and information on annotation build.

Determination of *NB-LRR* multi-gene family size by overlapping domain-specific sub-clusters

For every analyzed plant species, we determined the multi-gene family size of all annotated *NB-LRR* candidate genes by overlaying each filtered domain clusters. Note that statements about target loci missing or flawed within the gene annotations are beyond the scope of this section, but can likewise be considered *in silico* by applying sequence scaffolds/contigs instead of gene models to our customized pipeline (see Discussion section). For the A. thaliana Col-0 TAIR10 genome annotation, we have found 140 non-redundant *NB-LRR* loci (**fig. 4A**). Previous studies found 166 [309], 178 [282], 174 [310, 311] and 138 [72] *NB-LRR* loci present in the model plant. In contrast, TAIR10 domain annotation efforts reported 127 target loci [69]. The differences in our study resulted from usage of the updated TAIR10.02 annotation and more stringent criteria; namely the exclusive combination of machine learning with sequence identity and consideration of the genomic context (e.g. synteny). For example, we focus on protein-coding genes only and ignore non-functional (i.e. pseudogenized) loci due to the scope of this study to provide information relevant for breeding of gene-edited crops. Moreover, we defined NB-LRR proteins as sharing both NB-ARC- and LRR-domains, whereas many previous studies score anything as a *NB-LRR* gene that partially aligns to any one domain common to the cluster (i.e. TIR-only, NB-only, LRR-only genes).

For the *A. lyrata* genome annotation v0.2, we identified 166 non-redundant *NB-LRR* loci (**fig. 4B**). Previous studies reported evidence for 182 [311] and 138 [72] *NB-LRR* loci present in the *A. lyrata* genome assembly. Chen et al. score pseudogenes as well as loci that do not contain both NB- and LRR-domains, leading to the higher number of target genes than reported in this study [311]. The difference between our results and those of [72] is likely due to false-negative target genes with a divergence level that cannot be recognized by their applied HMM-generated NB-ARC consensus

sequence . We were able to score these more divergent loci using synteny data anchoring locus determination and subsequent *de novo* domain prediction using a combination of 14 HMM algorithms (see Materials & Methods section). For example, the *A. lyrata* locus fgenesh1_pg.C_scaffold_8000651 displays only moderate homology (e-value: 1e-34) to the closest related sequence in *A. thaliana*, a P-loop-containing nucleoside triphosphatase that is not defined as *NB-LRR* locus. However, we found both encoded NB-ARC- and LRR-domains within that gene in *A. lyrata*.

For the crop plant *B. rapa* (genome annotation v1.1), we found 167 non-redundant *NB-LRR* candidate genes (**fig. 4D**), while previous studies reported a sum of 92 [312] and 206 [313] target loci. The latter number includes proteins without an LRR-domain (for example TIR-NB or CC-NB). Removing those, [313] identified 139 genes encoding both NB-ARC- and LRR-domains, 28 less than we proposed. This differences may be due to our consideration of synteny and application of 14 different HMM algorithms, whereas [313] employed HMMER V3.0 only. Note that [312] did not have the whole genome assembly available, and hence identified R-proteins based on 1,199 partially redundant BAC clones mostly from a single chromosome. The authors acknowledge a significant degree of sequence redundancy within the available dataset that covers 19-28% of the *B. rapa* genome only. Likewise, [312] performed ab-initio gene annotation based on the fgenesh algorithm only [314], and solely use protein sequence homology (based on BLASTP) for R-protein homolog identification. In contrast, we used the whole gene-space assembly (including every to-date annotated protein-coding gene) as well as three layers of information for homolog identification (see Materials & Methods section).

To our knowledge, we performed the first analyses of R-proteins for *E. parvulum*, *Ae. arabicum*, *T. hasslerania* and *N. benthamiana*. For the extremophile saltwater cress *E. parvulum* (previously known as *Thellungiella parvula*, genome annotation v2), we found 72 non-redundant *NB-LRR* loci (**fig. 4C**). For *Ae. arabicum*, the extant sister lineage to all other mustard family members (genome annotation v0.2), we identified 112 non-redundant *NB-LRR* loci (**fig. 4E**). For the *T. hasslerania* genome annotation v4 (previously known as *Cleome spinosa*), we identified 59 non-redundant *NB-LRR* loci for this species (**fig. 4F**), that underwent a lineage-specific genome triplication event (**fig. 3**) and has been established as the mustard family outgroup [45, 53, 54]. For the Solanaceae and tobacco relative *N. benthamiana*, we identified 233 non-redundant NB-LRR proteins (**fig. 4J**). Notably, *N. benthamiana* is widely used as system for transient over-expression and silencing of various genes involved in plant innate immunity to elucidate downstream signaling events after PAMP-mediated priming. In this context, our results provide accurate mapping of all *NB*-LRR-like sequences encoding functions characterized in *A. thaliana* down to the *Nicotiana* gene-space assembly (**supplementary file 2**, Supplementary Material online), thereby facilitating adjusted planning of aforementioned experiments and better understanding of results in the Solanaceae.

For the crop plant *C. papaya* (genome annotation v0.5), we identified 44 non-redundant R-proteins of the NB-LRR type (**fig. 4G**). Among all species we have analyzed so far, the papaya gene-space assembly encodes the lowest number of R-gene candidates. We again acknowledge the possibility of incomplete gene annotations in this context (see Discussion section). However, the low gene count of the *NB-LRR* locus family was previously revealed for the available papaya genes set [315]. The authors found 54 target loci using a combination of TBLASTX [66] and the pfam HMM algorithm to search for the pfam NB (NB-ARC) family PF00931 domain [244]. The difference in gene-family size estimates is due to an updated genome annotation we have used, as well as more stringent criteria

for target gene scoring (i.e. NB-LRR proteins are defined as sharing both NB-ARC- and LRR-domains, see above).

Our analysis revealed 455 non-redundant loci of the *NB-LRR* type for the crop plant *C. sinensis* (orange) (**fig. 4H**). Evidence for the high R-gene count in orange has been noted previously. For example, the plant resistance gene database (prgdb) lists 3,230 R-genes (including LRR-domain-containing receptor-like kinases/proteins) for this crop plant [316], many of which are redundant. To our knowledge, our study comprises the first efforts to cross-reference both NB-ARC- and LRR-domains among R-genes in orange.

For grape (*V. vinifera*), we found 294 non-redundant R-proteins sharing both NB-ARC- and LRR-domains (**fig. 4I**). Previous efforts identified 300 target genes [310]. The differences are due to an updated genome assembly as well as more stringent criteria for *NB-LRR* locus definition.

In addition, we subjected the potato crop (*S. tuberosum* Group Phureja DM1-3) genome annotation v3.2.10 to our customized pipeline for identification of homologous gene clusters. We identified 402 encoded non-redundant NB-LRR proteins within the potato genome (**fig. 4K**). Previous efforts identified 438 target genes [317] from the annotated proteins set using the MEME and MAST algorithms [318] as well as 755 target genes for the *NB-LRR* gene repertoire based on reduced representation analysis of DNA enriched [319] (referred to as "Renseq" hereafter [320-322]). Referring to [319], we acknowledge the inability of our pipeline to identify genes present in the crop but flawed or missing from the annotation or the assembly. The difference between our value and [317] results from more stringent criteria in *NB-LRR* locus identification. For example, at least 34 of the 438 genes from [317] do not contain both NB-ARC- and LRR-domains, whereas at least two do not contain any of the required domains.

For tomato (*Solanum lycopersicum* Heinz 1706), we have found 219 non-redundant R-proteins of the NB-LRR type (**fig. 4L**). Previous studies identified 221 target genes sharing both NB-ARC- and LRR-domains in a very conclusive approach [323]. The minor difference in numbers is due to a different build of the annotation based on the genome version 2.4 (fusion of loci/locus fragments) and illustrates the thoroughness of the corresponding authors work. In contrast, application of Renseq to tomato genomic and cDNA recently identified 355 NB-LRR genes, thereby highlighting further potential of improvement for *de novo* genome assembly and annotation [324]. Again, we stress that the error rate of our pipeline depends on the quality of the input data (i.e. genes missing in the assembly or annotation can't be detected). In total, we identified 2,363 R-proteins of the NB-LRR type. CDS sequences are appended including translation to protein sequences (**supplementary files 3-4**, Supplementary Material online).

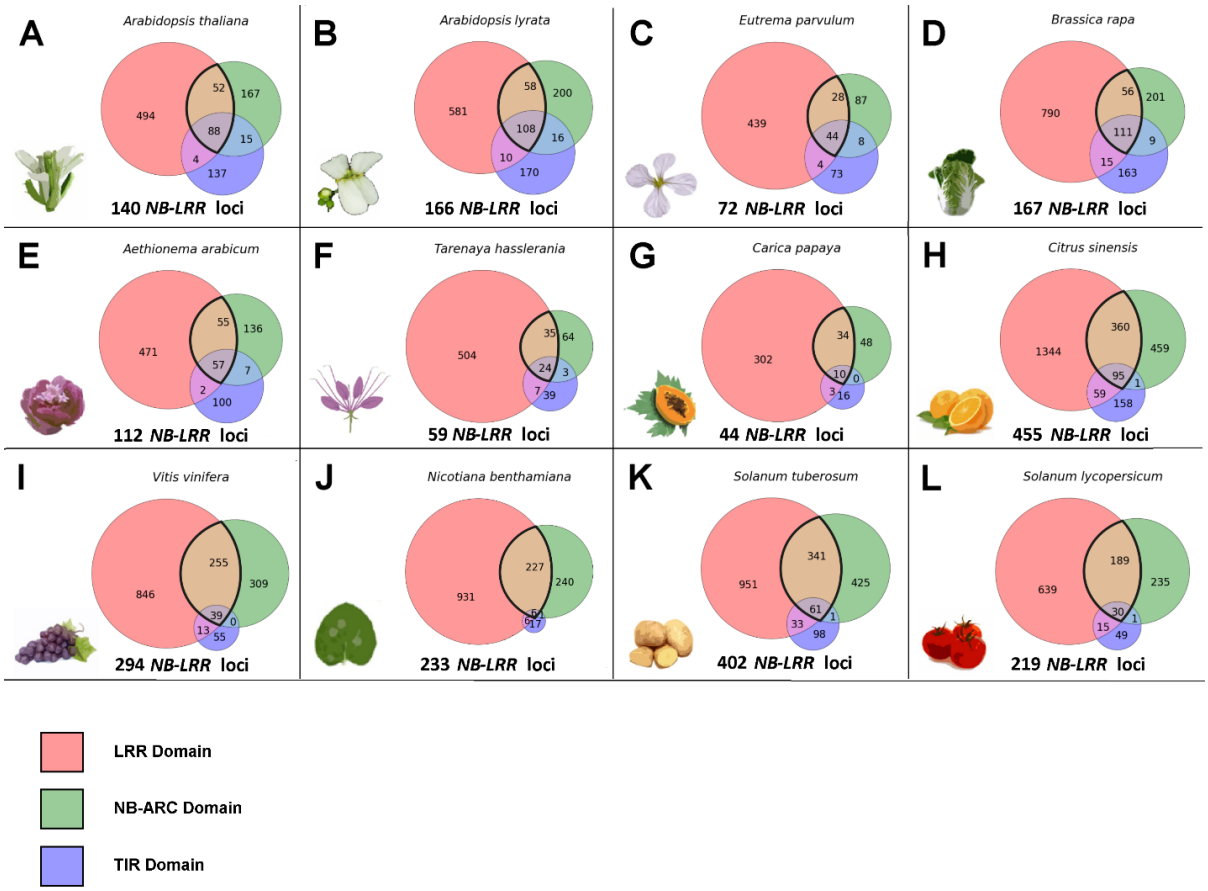


FIG. 4.— Area-weighted Venn diagrams illustrating the distribution of three main functional domains common to the *NB-LRR* gene family shown for twelve species. Domain-specific sub-clusters and overlaps are color-coded according to the legend (red: LRR-domain; green: NB-ARC- domain; blue: TIR-domain). Cartoons and italicized Latin names indicate target species: **A**. *Arabidopsis thaliana* **B**. *Arabidopsis lyrata* **C**. *Eutrema parvulum* **D**. *Brassica rapa* **E**. *Aethionema arabicum* **F**. *Tarenaya hasslerania* **G**. *Carica papaya* **H**. *Citrus sinensis* **I**. *Vitis vinifera* **J**. *Nicotiana benthamiana* **K**. *Solanum tuberosum* **L**. *Solanum lycopersicum*. Please note that we required a *NB-LRR* gene to harbor both NB-ARC- and LRR-domains.

Localization of genes with both NB-ARC- and LRR-domains and determination of tandem duplicate fractions

We localized all reported *NB-LRR* loci onto the corresponding chromosomes/scaffolds/contigs present in all analyzed genome assemblies except *N. benthamiana* (excluded from Circos plot due to insufficient assembly quality, see Materials & Methods section). Application of a number of n = 10 allowed gene spacers (see Materials & Methods section) allowed determination of a global rate of 53% tandem duplicates (**fig. 5**). Notably, we have found significant differences in tandem array fractions between the analyzed species (up to a factor of 2.8). For example, 70 *NB-LRR* genes present in the *V. vinifera* genome annotation v2 are members of tandem arrays (**table 1**). In contrast, the *N. benthamiana* genome annotation v0.4.4 contains only one fourth of tandem duplicates among all present *NB-LRR* loci. The latter represents a fragmented gene-space rather than a genome assembly, leading to a likely under-estimation of tandem duplicates fraction. Hence, the global tandem duplicates fraction drops after inclusion of *N. benthamiana* loci (**table 1**). For the mean gene count per *NB-LRR* tandem array, *Aethionema* scores highest. Likewise, the extant mustard family sister clade contains the largest tandem array we found so far. In contrast, the largest orange (*C. sinensis*) *NB-LRR* tandem array comprises less than half the number of target genes, leading to a very low genome-wide average of *NB-LRR* genes per tandem array for that species (**table 1**). Please note that we required presence of both encoded NB-ARC- and LRR-domains for *NB-LRR*-type R-gene curation. Therefore, some of the aforementioned tandem arrays may be further extended due to the presence of partial sequences in close proximity. We do not exclude a biological significance of such fragments *per se*, but set the scope to full-length candidate genes exclusively to obtain a uniform dataset to facilitate comparisons of molecular evolution rates (see below).

However, our data indicate that both aforementioned outlier situations with high (*Aethionema*) and low (*Citrus*) maximums for gene counts per *NB-LRR* tandem array are outliers beyond the average degree of *NB-LRR* gene tandem array extension. The majority of all 1,191 tandem duplicates (60%) are organized in arrays with two genes only. Three gene members per array occur in less than one fifth of all cases, whereas four, five and more than five genes per array occur with a cumulative frequency below 10% (**fig. 6**).

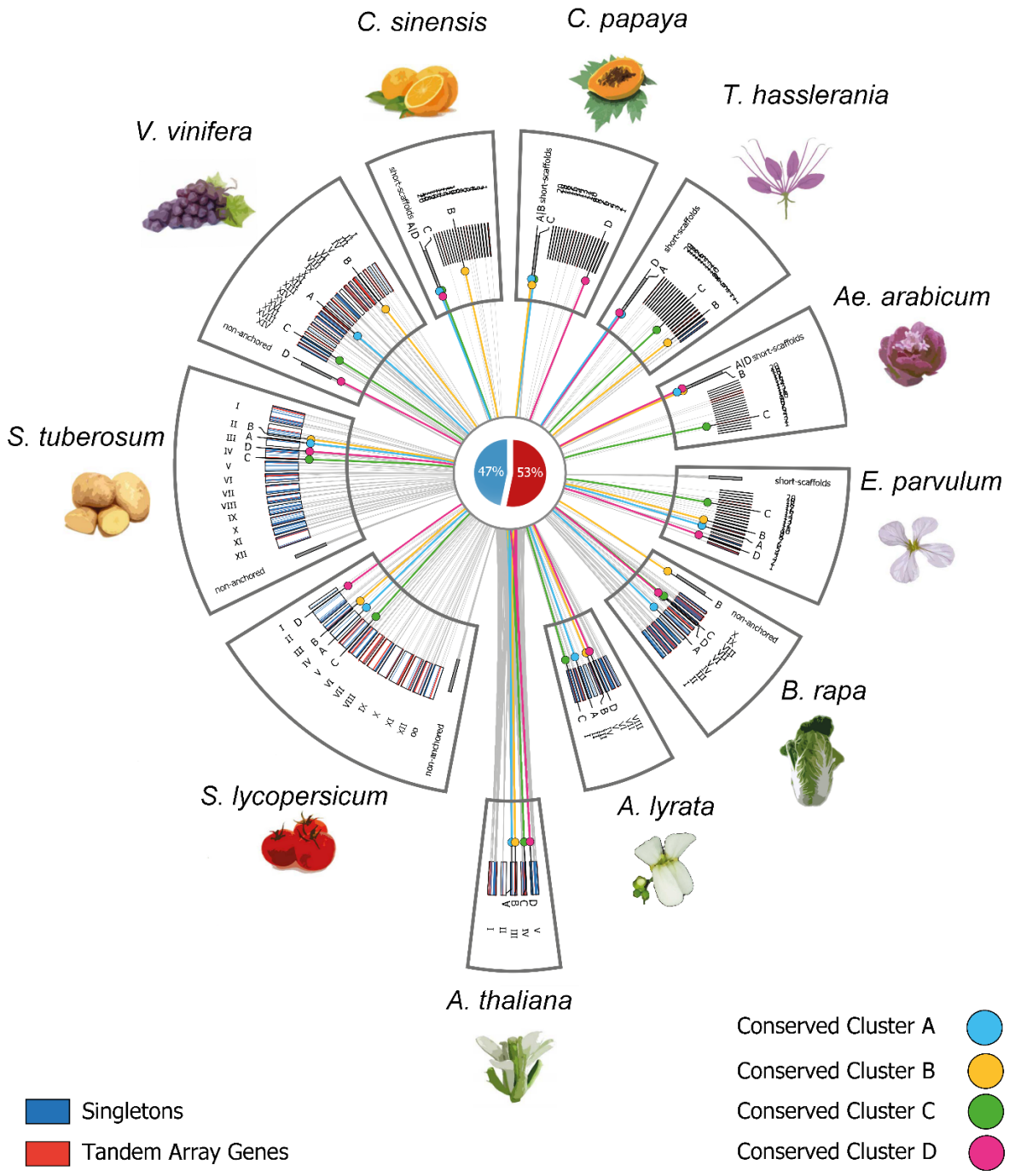


FIG. 5.— Circos ideogram with 2,363 *NB-LRR* loci localized on eleven genome annotations. Latin numbers refer to chromosome pseudo-molecules. Loose scaffolds and contigs not anchored to the genome assembly are shown shifted in radius but not in length scale. For genomes without assembly to the chromosome level, the 20 largest scaffolds are displayed and named in in ascending order with Arabic numbers. Beginning at the bottom block in counter-clockwise orientation, shown are (1) *Arabidopsis thaliana* Col-0, (2) *Arabidopsis lyrata*, (3) *Brassica rapa*, (4) *Eutrema parvulum*, (5) *Aethionema arabicum*, (6) *Tarenaya hasslerania*, (7) *Carica papaya*, (8) *Citrus sinensis*, (9) *Vitis vinifera*, (10) *Solanum lycopersicum* and (11) *Solanum tuberosum*. Tandem duplicate gene copies are highlighted in red. Singleton genes are highlighted in dark blue. "Conserved Cluster A-D" refers to four distinct *A. thaliana NB-LRR* loci that have been coded in distant colors for easy visual distinction (A: AT3G14470; B: AT3G50950; C: AT4G33300; D: AT5G17860) including ohnologs in all other ten genomes. For genome assembly versions used in this analysis, see fig.3. Please note that due to the fragmented assembly status of *Nicotiana benthamiana*, all scaffolds of this annotation are below visible length threshold.

Genome-wide determination of retained ohnolog fractions and cross-referencing of *NB-LRR* genes

We determined the genome-wide fraction of retained duplicate groups due to ancient polyploidy events (ohnologs), including all identified *NB-LRR* loci. Screening of pairwise synteny blocks within the analyzed genome assemblies was accomplished using an integer programming approach implemented by the CoGe package for comparative genomics (see Materials & Methods section) [194]. Due to technical restrictions, this was possible for seven genomes (i.e. minimum requirements in the N50 index, requiring a minimum of approximately 50 kb, see Materials & Methods section). The high degree of tandem duplicates among R-proteins in all species results in a low degree of retained ohnolog duplicates by definition, because ohnologs mainly comprise groups of two or three duplicates, whereas tandem arrays can have up to eleven members (**fig. 6**). Notably, the *B. rapa* genome possesses the highest syntenic depth value among all analyzed genome assemblies with 12x in total (**fig. 3**). Consistently we found the highest fraction of retained ohnolog duplicates both genome-wide and among *NB-LRR* genes present in this crop with in total (**table 2**). In contrast, the potato crop (*S. tuberosum*) contains the lowest fractions of retained ohnolog duplicates for both genome-wide average and the set of *NB-LRR* genes (**table 2**). On average, about one third of present in the seven analyzed genome assemblies comprise retained ohnolog duplicate groups. This fraction drops among all *NB-LRR* loci. This apparent under-representation of ohnologs among R-proteins highlights the high relative contribution of tandem duplication in R-protein cluster extension for the group of genome assemblies subjected to this analysis (**table 2**).

TABLE 1.- Arrays of Tandem Duplicate Copies among *NBS-LRR* loci

	NBS-LRR genes	Number of Tandem Duplicates	Number of Duplicates	Fraction of Tandem	Tandem Arrays	Number of Genes per Array	Average Number of	Number of Genes in largest Array
B. rapa	167	92	55%	31	2.9	8		
E. parvulum	72	37	51%	13	2.8	9		
A. thaliana Col-0	140	94	67%	32	2.9	8		
A. lyrata	166	71	43%	23	3.1	9		
Aet. arabicum	112	71	63%	21	3.4	11		
T. hasslerania	59	26	44%	10	2.6	6		
C. papaya	44	32	72%	10	3.2	5		
C. sinensis	455	136	30%	61	2.2	5		
V. vinifera	294	206	70%	62	3.3	10		
N. benthamiana	233	58	25%	26	2.2	5		
S. tuberosum	402	238	59%	77	3.1	8		
S. lycopersicum	219	125	57%	40	3.1	7		
Σ	2,363	1,186	50%*	406	2.9	7.6		

* Difference of value compared to fig. 5 is due to presence on *N. benthamiana*

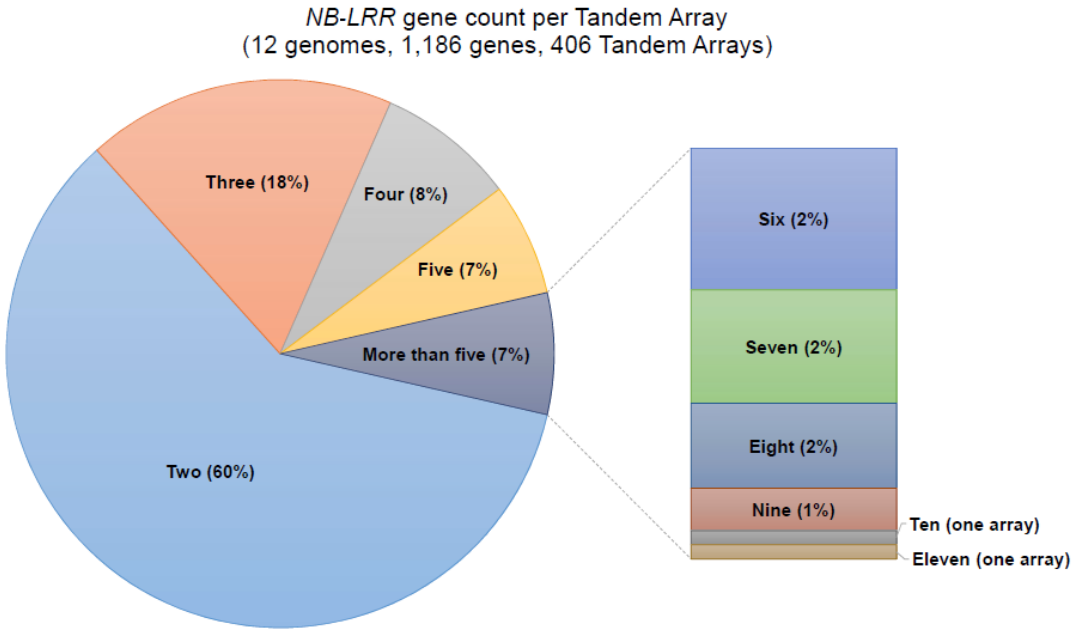


FIG. 6.— Gene count listing of full-length tandem duplicate *NB-LRR* genes observed within all twelve analyzed genomes. 60% of all tandem arrays comprise two duplicate gene copies.

TABLE 2.- Retained ohnolog duplicate copies among *NBS-LRR* loci

	Syntenic depth #	Genome-wide average	*NBS-LRR* loci Among	enrichement Ohnolog
B. rapa	12x	53%	42%	no
E. parvulum	4x	32%	29%	no
A. thaliana Col-0	4x	22%	17%	no
A. lyrata	4x	33%	23%	no
T. hasslerania	6x	44%	27%	no
S. tuberosum	2-3x	10%	5%	no
S. lycopersicum	2-3x	19%	16%	no
Σ		30.3%	22.7%	**no**

* Genomes with low assembly quality are excluded from this analysis due to technical reasons (see methods)
# Post-γ ploidy level

Uncovering differential patterns of selection acting on subsets of *NB-LRR* loci pooled according to duplicate origin

We performed a genome-wide analysis of molecular evolution acting on all encoded NB-LRR proteins based on both the NB-ARC- and LRR-domain. In a first step, we grouped (a) members of tandem arrays, (b) retained ohnolog duplicates as well as (c) singleton genes (defined as non-tandem array genes without retained ohnolog duplicate). By analyzing non-synonymous substitutions per non-synonymous sites, compared to synonymous substitutions per synonymous site (Ka/Ks ratio or ω, dN/dS), patterns of strong positive selection were uncovered among all three groups. Strikingly, we also found differences in molecular evolution rates among all three groups. Members of tandem arrays evolved fastest with a ω mean of 1.59. In contrast, all analyzed retained ohnolog duplicates evolved with an intermediate rate (ω mean = 1.36). We reported the slowest rate of molecular evolution for singleton NB-LRR genes with a ω mean of 1.22 (**fig. 7**). Values for ω above one indicate positive or Darwinian selection, less than one implies purifying (or stabilizing) selection whereas ratios of one are indicative for neutral (i.e. absence of) selection [325].

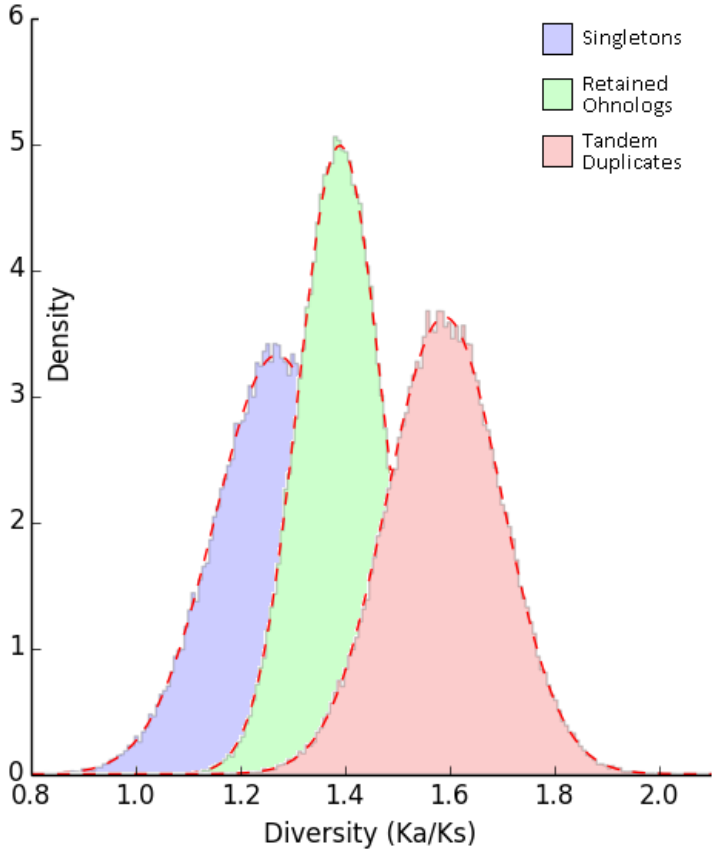


FIG. 7.— Selection in action between gene pairs of three major duplicates categories – singletons, tandem duplicates and WGD duplicates (ohnologs). Strong positive selection following gene and genome duplication of *NB-LRR loci*, as indicated by higher Ka/Ks values.

Assessing structural dynamics of genomic regions with conserved *NB-LRR* loci

Utilizing the wealth of *NB-LRR* functional and molecular data available in *Arabidopsis* as a reference, we composed a species-wide matrix of R-protein presence/absence based on sequence homology (i.e. filtered/non-filtered reciprocal best BLAST hits, referred to as "RBH" hereafter) and synteny (**supplementary files 1 and 2**, Supplementary Material online). Among the extended set of 140 distinct *NB-LRR* loci present in the model plant (see above), we found four conserved clusters of "gatekeeper" genes sharing syntenic orthologs across all twelve analyzed genomes (**fig. 5 and supplementary file 1**, Supplementary Material online). Please note that genomic regions displaying conserved synteny across lineages (**fig. 8**) define evolutionary immobile parts of plant genomes [26]. For two among those, functional data are available in *Arabidopsis*, whereas members of the other two gene clusters have not yet been characterized in any of the analyzed plant lineages. The non-TIR non-CC NB-LRR (NL) class R-protein AT3G14460 is a "gatekeeper" because it forms one of four conserved clusters together with all of its aforementioned ohnologs (**supplementary file 1**, Supplementary Material online and "Conserved Cluster A" in **fig. 5**). Interestingly, there are yet no functional data available concerning this gene, neither in *Arabidopsis* nor in any of the other eleven analyzed genome/gene-space assemblies. For example, this NL-class "gatekeeper" AT3G14460.1 [282, 326] forms syntenic RBH pairs with fgenesh2_kg.3__1571 (*A. lyrata*), Bra027333 (*B. rapa*), Tp3g12770 (*E. parvulum*), AA_scaffold578_71 (*Ae. arabicum*), Th16129 (*T. hasslerania*), supercontig_77.89 (*C. papaya*), GSVIVT01013307001 (*V. vinifera*), Solyc03g078300.1 (*S. lycopersicum*) as well as PGSC0003DMG400005046 (*S. tuberosum*). For *C. sinensis*, the RBH partner orange1.1g000782m is harbored by a very small scaffold (~12.6 kb) with three genes only, making the scoring of gene synteny impossible. However, the locus orange1.1g000782m in turn forms RBH pairs with the aforementioned genes supercontig_77.89 (*C. papaya*) as well as GSVIVT01013307001 (*V. vinifera*), thereby closing the gap in a phylogenetic framework (data not shown). Likewise, the *N. benthamiana* gene NB00009911g0001.1 forms RBH pairs with the aforementioned syntenic orthologs in tomato and grape-vine, overcoming the lack of synteny data for this early-stage draft genome assembly (data not shown). Notably, the underlying locus underwent tandem duplication after grape-vine lineage split, leading to presence of a tandem array in all Brassicales including orange, but an evident singleton gene in Solanaceae and *V. vinifera* (**fig. 8**).

The TIR-*NB-LRR* (TNL)-class "gatekeeper" locus AT5G17680 is anchoring another group of syntenic orthologs shared by all lineages (**supplementary file 1**, Supplementary Material online, "Conserved Cluster D" in **fig. 5**). Similarly, this locus lacks evidence on gene function in any of the analyzed plant lineages.

In contrast, conserved clusters B and C are anchored by *ZAR1* (*HOPZ-ACTIVATED RESISTANCE 1* or AT3G50950) and the *NB-LRR* gene *ADR1-L1* (*ACTIVATED DISEASE RESISTANCE 1* or AT4G33300), that confers pleiotropic effects in *Arabidopsis* innate immunity (**supplementary file 1**, Supplementary Material online, "Conserved Cluster B and C" in **fig. 5**). *ZAR1* encodes a CC-NB-LRR (CNL) class R-protein of the FLARE group (Flagellin Rapidly Elicited, due to rapid up-regulation following exposure to the PAMP flg22) [327]. *ZAR1* confers allele-specific recognition of the *Pseudomonas syringae* HopZ1a type III effector in *Arabidopsis* and acts independent of several gene products required by other R-protein signaling pathways [328]. In contrast, *ADR1-L1* over-expression results in a dwarf phenotype and activation of defense-related gene expression in *Arabidopsis* [290, 327]. Note that *ADR1-L1* encodes an R-protein conferring pleiotropic roles due to function as "helper" *NB-LRR* that can transduce signals subsequent to specific pattern recognition receptor activation during effector-triggered immunity [329]. Furthermore, *ADR1-L1* encodes the N-terminal RPW8-like domain, whose

functional importance in plant innate immunity has been previously reported [292, 330]. Interestingly, the *Arabidopsis RPW8*-like "gatekeeper" was found to be necessary and sufficient to confer induced resistance to powdery mildew in the distant lineage of Solanaceae (*Nicotiana tabacum*) [291]. This case excludes restricted taxonomic functionality and provides additional evidence for functional conservation of syntenic orthologs as defined by "gatekeepers" on a broad phylogenomics range.

In summary, we found four *NB-LRR* genes conserved in sequence as well as linked to structurally immobile parts of the core-eudicot pan-genome. At least one of those confers pleiotropic effects and extended functions in Arabidopsis as a "helper-*NB-LRR*" [331]. Although both synteny and sequence conservation across lineages during a timeframe of approximately 250 MA provides strong indications for conservation in function, this may not always be the case. However, we hypothesize that structural stability of the harboring genomic region sup- ports evolution of pleiotropic effects conferred by "gate- keeper" R-proteins (see below).

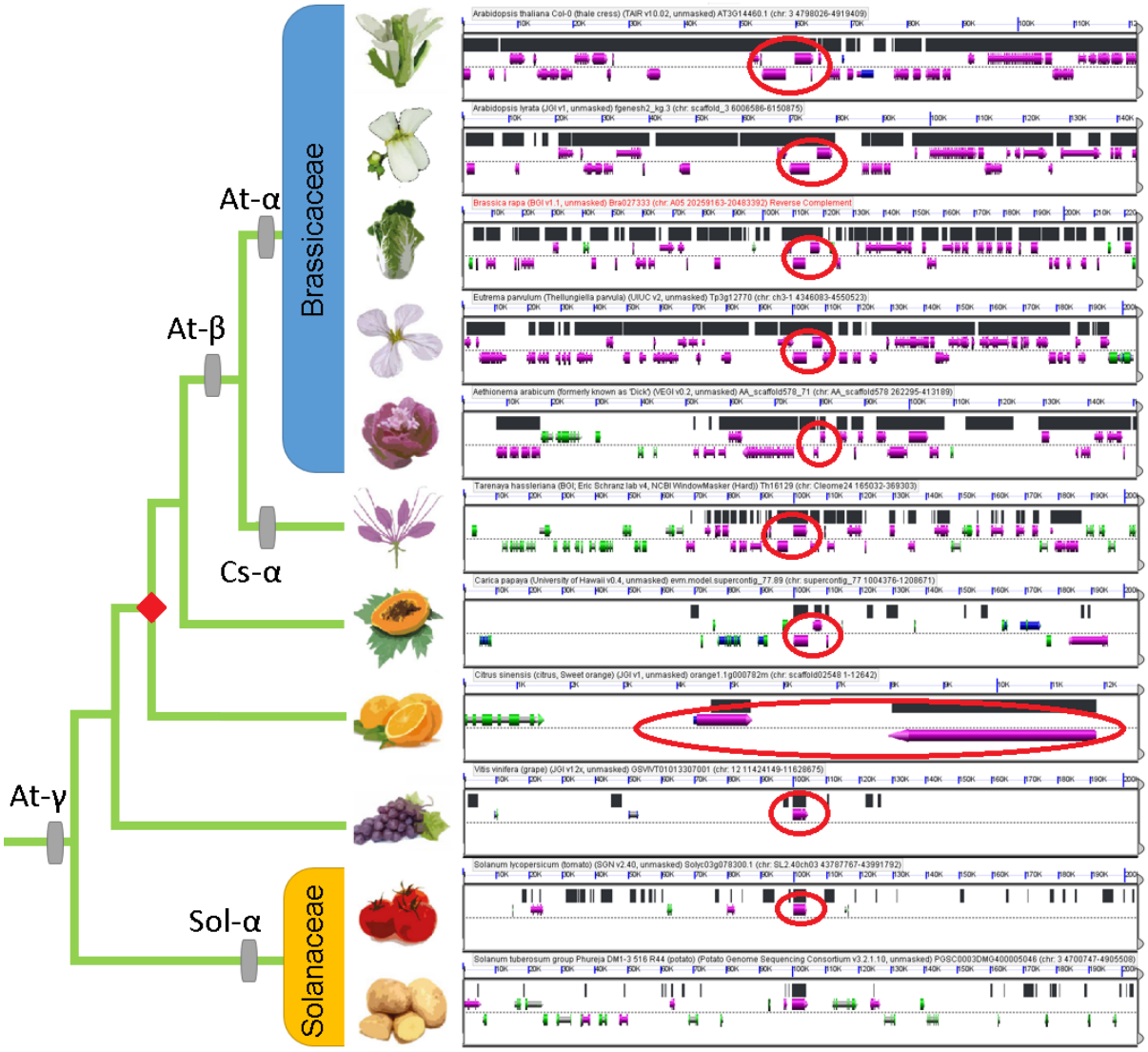


FIG. 8.— (B)LastZ eleven-way multiple alignment of conserved cluster A from fig. 5. Shown left is the phylogenetic relationship among all eleven species (*Nicotiana benthamiana* is excluded from this analysis due to technical reasons). Shown right is the genomic context of the syntenic regions (marked in black). The regions in focus include one *NB-LRR* gene that expanded to a tandem array in the *Arabidopsis* lineage after split of Solanaceae. Diamond indicates a tandem duplication event. Genes not overlapped by highest-scoring sequence pairs are shown in green. In case of *C. sinensis*, the orthologous genes are harbored by a very small scaffold (~12.6 kb), therefore scaled differently from other panels in GEvo.

DISCUSSION

The proliferation of high-throughput DNA sequencing and genome informatics approaches enables an accelerated production rate of draft genomes from a wide phylogenetic sampling of plant taxa, highlighting a need for robust methods and a comparative framework for gene and genomic comparisons. We therefore have developed a custom approach to identify functional groups of plant proteins applying a novel and highly complementary combination of available algorithms and data-sets. We have applied this to R-proteins and annotated 2,363 loci of the *NB-LRR* type in total. This set contains genes that previously remained un-identified for all species except tomato and potato. For Solanaceae, we stress that re-sequencing approaches based on complexity reduction such as target gene capture have been successfully applied for a similar purpose (referred to as Renseq) [319, 332]. However, it is not unreasonable to assume that the onset of next generation sequencing and genome informatics will continue with acceleration beyond Moore's law and hence lead to more and better algorithms for *de novo* generation of gene annotations. Therefore, the added value of the computational pipeline shown in this study will rise with the same rate. For future references, we are working on customization of our approach to make it suitable for application to whole sequence scaffolds/contigs rather than sets of annotated genes/proteins. We intend to generate a computational pipeline for *in silico* target gene capture based on scoring of combined hits outside the annotated gene-space within a size window common to protein-coding genes, thereby overcoming the evident limitations of currently available algorithms for *de novo* gene annotation (Jupe F, personal communication). The pipeline shown in this study represents the first step towards this goal.

Since tandem duplicates represent the majority of the R-gene duplicates that typically have a higher turnover rate, and additionally most of the R-genes have experienced high birth-and-death rate due to the persistent arms-race with the evolution of pathogen target effectors, most R-genes should have a fairly limited cross-taxonomic coverage [333, 334]. However, a limited set of R-gene clusters are more stable, such as the four gene clusters that we have shown here to be conserved over 100 MA in most (if not all) core eudicot genomes. Could these gene clusters represent shared immunity responses to common pathogens? In addition, the genes in these clusters could also act as "helper *NB-LRRs*", mediating signal transduction downstream of various different NB-LRR receptors for activation during effector-triggered immunity, thereby leveraging functional constraints as previously made evident for *ADR1* family in *A. thaliana* [331, 335]. Please note that members of the RPW8-domain-encoding *ADR1*- like family have been identified across all angiosperms, providing hints towards relevance of "gatekeepers" in a broad phylogenomics range across the whole angiosperm clade [336] (Zhao and Schranz, unpublished results). More studies need to be done in order to unravel gene function underlying the retention of these unusually "stable" R-gene loci. This is stressed by the fact that (some degree of) functional evidence accumulated for two of our four *NB-LRR* "gatekeeper" functions in *Arabidopsis*; in at least one case "gatekeeper" R-proteins confer pleiotropic effects as "helper" *NB-LRR*s. In contrast, such data lacks for the other two "gatekeepers", notably including one TNL class R-protein. We hypothesize significant potential for extension of gene functional data regarding all four "gatekeeper" loci, either by gene-for-gene resistance towards yet-undiscovered pathogen effectors or by facilitating pleiotropic effects and effector-triggered signaling downstream of other *NB-LRR* genes similar to "helper *NB-LRR*s". Notably, a combination of both scenarios is evident in *Arabidopsis* and hence not unreasonable to occur in other cases.

We highlight the need for "uniform" standards for comparative studies, such as the method we used in this study that is applicable but by no means limited to R-gene families. In contrast to most past computational pipelines of gene identification that only employ DNA sequence similarity, our approach consolidates multiple tiers of evidence, including the basic protein sequence identity, domain compositions, and genomic context (synteny). Uniform standards also ensure that our gene family member counts are directly comparable with one another, making in-depth studies of the expansion-contraction dynamics of gene families possible. Furthermore, our method allows efficient screening of genome assemblies for near-complete curation of multi-domain and multi-gene family clusters. In the case of *NB-LRR* type R-genes, the resulting raw data provide a detailed overview of nucleotide diversity among all target genes within and between twelve lineages covering the whole core-eudicot clade. Utilizing the wealth of genomics and gene functional data in *A. thaliana*, this leads to species-wise mapping (presence/absence) of every *NB-LRR* sequence present in the model plant. Notably, these data can be used by breeders to identify both target loci as well as small RNA sequence requirements for fast and efficient migration of resistance locus A to organism B using the emerging techniques of genome editing in case restricted taxonomic R-gene functionality doesn't apply. For example, the particular *NB-LRR* gene conferring the desired resistance can be selected from our curated dataset followed by calculation of the smallest nucleotide distance (or closest related) target gene in the desired organism. The sequence of the small RNA(s) necessary for engineering of nucleases in context of genome editing can be inferred accordingly in order to design a minimum set of experiments necessary and sufficient for gene-editing and thus generating an extended spectrum of resistance in any of the crop subjected to our analysis. However, note that taxonomic restrictions may apply for at least some encoded R-gene functions. Going beyond plant innate immunity, we provide data on a network of anchor genes present in all analyzed genome assemblies, thereby referencing orthologs and paralogs of every gene family present in the model plant *Arabidopsis*. We thereby excel future efforts to extract plant gene function, ultimately necessary for crop improvement and increased rates of global food production.

CONCLUSION

We highlight three major findings in this study: (a) higher frequency of tandem gene expansion in R-genes, (b) higher selection ratio in tandem duplicates compared to ohnologs and singletons and (c) evolutionary stable, orthologous R-gene clusters established within structurally immobile parts of plant genomes. Those are likely to indicate a common functional constraint ("gatekeepers"). R-genes typically show an unusually high turnover rate due to strong selection to keep up in a biological arms race with plant pathogens [300, 311]. We suggest such R-genes follow a different evolutionary trajectory than genes with regulatory roles [100]. In this context, the added value of our study lies within the wide phylogenomics scope of the underlying approach. Although similar findings are available in *Arabidopsis*, monitoring dynamics underlying target gene evolution for approximately 100 MA (corresponding to radiation time of the core eudicots) results in higher confidence in the validity of our inferences.

ACKNOWLEDGEMENTS AND FUNDING INFORMATION

We would especially like to thank Florian Jupe for his valuable input with proof-reading of the manuscript and support during all stages of the underlying research. Likewise, thanks go to Detlef Weigel and the whole BMAP team for their inspiration and discussions during the onset of this project. The authors are grateful to Xinguang Zhu for his support at CAS and to Mariam Neckzei for

her help with graphical editing of the figures. Finally, we would like to thank three anonymous reviewers for their helpful comments. This work was funded by a Netherlands Organization for Scientific Research (NWO) VIDI and Ecogenomics grant (M.E.S.).

SUPPLEMENTARY MATERIAL

Supplementary files 1-4 are available at BMC Genomics online (www.biomedcentral.com/bmcgenomics, last accessed on December 13th, 2014).

Supplementary file 1.— Syntelogs and orthologs of all Arabidopsis *NB-LRR* genes across all analyzed species.

Supplementary file 2.— *NB-LRR* gene IDs, duplicate classes and closest homolog in *Arabidopsis*.

Supplementary file 3.— CDS sequences of identified genes encoding both NB-ARC- and LRR-domains.

Supplementary file 4.— Translated protein sequences of identified genes encoding both NB-ARC- and LRR-domains.

A Complex Interplay of Tandem- and Whole Genome Duplication Drives Expansion of the L-type Lectin Receptor Kinase Gene Family in the Brassicaceae

Johannes A. Hofberger[1,2,#], David L. Nsibo[1,#], Francine Govers[3], Klaas Bouwmeester[3,4], and M. Eric Schranz[1]*

[#] These authors contributed equally to this work

[1] Biosystematics Group, Wageningen University, 6708 PB, Wageningen, The Netherlands

[2] Chinese Academy of Sciences/Max Planck Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai 200031, PR China

[3] Laboratory of Phytopathology, Wageningen University, 6708 PB, Wageningen, The Netherlands

[4] Plant-Microbe Interactions, Department of Biology, Faculty of Science, Utrecht University, 3584 CH, Utrecht, The Netherlands

ABSTRACT

The comparative analysis of plant gene families in a phylogenetic framework has greatly accelerated due to advances in next generation sequencing. In this study, we provide an evolutionary analysis of the L-type lectin receptor kinase and L-type lectin domain proteins (L-type LecRKs and LLPs) that are considered as components in plant immunity, in the plant family Brassicaceae and related outgroups. We combine several lines of evidence provided by sequence homology, HMM-driven protein domain annotation, phylogenetic analysis and gene synteny for large-scale identification of L-type *LecRK* and *LLP* genes within nine representative core-eudicot genomes. We show that both polyploidy and local duplication events (tandem duplication and gene transposition duplication) have played a major role in L-type *LecRK* and *LLP* gene family expansion in the Brassicaceae. We also find significant differences in rates of molecular evolution based on the mode of duplication. Additionally, we show that *LLP*s share a common evolutionary origin with L-type *LecRK*s and provide a consistent gene family nomenclature. Finally, we demonstrate that the largest and most diverse L-type LecRK clades are lineage-specific. Our evolutionary analyses of these plant immune components provide a framework to support future plant resistance breeding.

KEYWORDS: Comparative genomics, polyploidy, gene duplication, Brassicaceae, L-type lectin receptor kinases, plant innate immunity

***Author for Correspondence:**
M. Eric Schranz | Biosystematics Group | Wageningen University & Research Center | Wageningen, The Netherlands | Tel. +31(0)317-483160 | email: eric.schranz@wur.nl

INTRODUCTION

During plant evolution, individual genes and gene families have undergone selection for copy number via duplications, transpositions and/or deletions. Such events can be detected by screening for patterns of syntenic or collinear genes [34, 337]. Gene duplication and subsequent gene retention or loss (fractionation) is often attributed to recent and/or ancient whole genome polyploidy events, for example at the origin of seed plants and angiosperms [37]. Whole genome duplications (WGDs) can act as mechanism to buffer gene functions due to increased genetic redundancy and hence provide an important source of sub- or neo-functionalization driving genetic innovation [89, 338]. For example, ohnolog genes (paralogous genes derived specifically from a WGD) encoding structurally similar enzymes have been shown to evolve towards extended substrate specificities or catalysis of novel reactions, while its ancestral gene retains its designated function [80]. Similarly, distant genomic locations of ohnologs can lead to differential gene expression [339]. Hence, it has been hypothesized that WGDs contributed to species diversity by driving trait evolution [84]. In this context, several studies highlight the contribution of WGD to the observed diversity across lineages as well as to extended gene function in a variety of organisms, including mammals [340], amphibians [341, 342] and plants [343-345]. Large-scale synteny is not observed for paralogs derived from small-scale events like tandem- and gene transposition duplication (GTD).

The Brassicaceae, also known as the mustard family, has many advantages to study and understand the contributions of whole genome and gene duplications on plant genome evolution. It comprises several species for which well-assembled genomes are available, including *Arabidopsis thaliana, Arabidopsis lyrata*, *Brassica rapa*, *Thellungiella halophila* and *Aethionema arabicum* [43, 51, 69, 118, 269, 346]. Analysis of many of these genome assemblies has provided insights into patterns of gene evolution, retention and functionality [85]. Within the Brassicaceae family at least five polyploidy events can be detected that have occurred in the *A. thaliana* lineage, three of which have been studied extensively [35, 347]. This includes the “At-γ” event that occurred approximately 111 million years (MA hereafter) before the lineage split of *A. thaliana* and the common grape vine *Vitis vinifera* [348] and which is shared by all eudicots [47, 48]. The less ancient ”At-β” event occurred approximately 72 MA after the split of the *Carica papaya* and *A. thaliana* lineages [46], and is restricted to the order Brassicales [34]. The most recent polyploidy event termed “At-α” occurred after the split of Brassicaceae and Cleomaceae approximately 40 MA ago, and was followed by a lineage separation of *A. thaliana* and *A. lyrata* approximately 10 MA ago [35, 84, 269]. In addition to WGD events in plant genomes, local duplication events such as tandem- and gene transposition duplication (GTD) contributed to gene copy number variation and are currently the best understood drivers of gene retention and cluster expansion. Tandem duplication (TD) is a result of single unequal crossing over (UCO) events, and/or multiple repeats thereof during DNA repair. UCO produces tandem duplicate genes organized in tandem arrayed genes (TAR genes) that individually cluster with up to ten intervening genes [90]. This lead to copy number variation in many plant gene families including several involved in plant disease resistance and glucosinolate biosynthesis [44, 91, 92, 94]. Interestingly, UCO can result in gene copies positioned in a head-to-tail direct orientation [282]. Alternatively, TD can also result from intra-chromosomal rearrangements between direct and indirect repeats, producing gene copies with opposite head-to-head orientation within a tandem array. Note that depending on the orientation of adjacent tandem duplicates, common promoters can be shared. For example, intra-chromosomal rearrangements caused the formation of the *A. thaliana* gene array *RRS1* and *RPS4*, that function as a dual resistance gene system in defense against bacterial and fungal plant pathogens [287]. Similarly, TD has significantly influenced the divergence

of many disease resistance genes (i.e. *NB-LRRs*) that confer race-specific resistance in Brassicaceae and Solanaceae [92, 145, 349]. In contrast to TD, gene transposition duplication (GTD) results in gene relocation to distant genomic positions and hence induces gene family dispersion across the entire genome. GTD copies transpose from ancestral to novel positions with the ancestral loci having fewer insertions and deletions (InDels) with shorter maximum InDel lengths. In addition, ancestral GTD "seed" loci have longer coding-regions and exon lengths than the novel copies [350]. Overall, TD and GTD have been reported to frequently occur in diversified high-copy number gene families, such as those comprising *NB-LRR* disease resistance, Type I MADS-box transcription factor, F-Box, and B3 gene families [33, 100].

In plants, the perception of extracellular stimuli and subsequent signal transduction is often mediated by receptor-like kinases (RLKs), which can be divided in various subfamilies based on their extracellular domains [351, 352]. Plant RLKs underwent a dramatic expansion in comparison to those of other organisms − with at least 610 and 1100 members in *A. thaliana* and rice, respectively [353, 354] −, indicating their importance during plant adaptation. Several RLKs have been shown to play pivotal roles as pattern recognition receptors (PRRs) to mediate basal defense. Amongst these are the lectin receptor kinases (LecRKs), which are membrane-spanning receptors that contain an extracellular lectin domain and an intracellular Ser/Thr kinase domain [355, 356]. LecRKs can be further subdivided based on their lectin domain composition into three categories; i.e. the G-, C- and L-type LecRKs [355, 357]. The G-type LecRKs, also known as S-domain RLKs (SRKs), comprise functions in both plant self-incompatibility and defense [358, 359]. C-type LecRKs are named after their extracellular calcium-dependent lectin domain. This domain is commonly found in a plethora of innate immune receptors in mammalians [360], but is rare in plants. The function of the C-type LecRKs remains thus far enigmatic [355]. The third category consists of the L-type LecRKs which contain an extracellular legume-like lectin domain. L-type LecRKs are ubiquitous in plants, and have been identified in a variety of plant species, e.g. cotton, cucumber and rice [361-363]. *A. thaliana* was shown to contain 45 L-type LecRKs, which could be divided into nine distinct clades and seven additional so-called singletons that group distantly (termed "ambiguous" hereafter) [355].

Recently, evidence has accumulated pointing towards roles of L-type LecRKs in biotic stress responses [364-367]. LecRK-I.9, for example, was identified as a *Phytophthora* resistance component [364], while LecRK-V.5 is involved in susceptibility to bacterial pathogens [365]. In addition, Singh and co-workers [366] showed that LecRK-VI.2 is critical in defense against both hemibiotrophic and necrotrophic phytopathogenic bacteria in *A. thaliana*. L-type LecRKs also function in insect resistance, e.g. *A. thaliana* LecRK-I.8 was shown to play a crucial role in defense triggered by egg-derived elicitors of the cabbage butterfly *Pieris brassicae* [368]. In addition, few L-type LecRKs are thus far described to function in response to abiotic stimuli and plant development [367, 369-371].

Here, we employ several bioinformatics methods for the identification and comparison of L-type *LecRK* family members encoded in nine representative core-eudicot genomes. In a phylogenomics approach, we provide data to assess the differential impact of duplication modes driving L-type *LecRK* copy number expansion observed across the plant family Brassicaceae.

MATERIALS & METHODS

Plant genome annotations

Genome annotations for four Brassicaceae species: i.e. *Aethionema arabicum* v0.2 [43], *Arabidopsis thaliana* TAIR10 [69], *Arabidopsis lyrata* v1.07 [269], *Brassica rapa* [51] and *Thellungiella halophila* v1 [346]; one Cleomaceae species: *Tarenaya hasslerania* v4 [53]; one Caricaceae species: *Carica papaya* v0.5 [46]; one Malvaceae species: *Theobroma cacao* v1 [372] and one Vitaceae species: *Vitis vinifera* v2 [48] were obtained from Phytozome v9.1 (http://phytozome.org, last accessed on December 13th, 2014) [186].

Re-annotation of L-type *LecRK*s and *LLPs*

Protein and gene sequences of *A. thaliana* L-type LecRKs and LLPs were obtained using the Arabidopsis Information Resource website (TAIR10, http://www.arabidopsis.org, last accessed on December 13th, 2014). Possible pseudogenisation of *A. thaliana* L-type *LecRKs* and *LLPs* was analyzed using available ATH1 microarray datasets at TAIR (data not shown). To identify orthologous L-type LecRKs and LLPs across the nine plant genomes, the Reciprocal Best Blast Hits (RBH) were determined using both *A. thaliana* gene and protein sequences as queries against the remaining eight plant genomes using NCBI BLAST 2.2.28+ (http://www.ncbi.nlm.nih.gov/news/04-05-2013-blast-2-2-28, last accessed on December 13th, 2014) [66, 373] with an e-value threshold of 1e-10. A total of three RBH sets (i.e., a length filtered protein pair set; a non-length filtered protein pair set, and a non-length coding sequence pair set with a size-filter threshold of 0.5-to-2 gene lengths) were retrieved after BLAST as previously described [145].

Ohnolog identification and analysis

Ohnolog (collinear or syntenic copies of genes) of all putative L-type *LecRK* orthologs were identified via analysis of gene collinearity within and between all genomes using the "SynMap" algorithm within the CoGe package for comparative genomics (www.genomeevolution.org, last accessed on December 13th, 2014) [26]. Firstly, genes of each analyzed species that share syntenic orthologs to the *A. thaliana* L-type *LecRKs* and *LLPs* were determined by making use of DAGchainer [195] and quota align algorithms [194] within the CoGe package for comparative genomics (http://genomeevolution.org/CoGe/GEvo.pl, last accessed on December 13th, 2014). The following parameter settings were used: merging neighboring syntenic blocks, maximum distance between two blocks fixed at 350 genes; synonymous substitutions rates (Ks) with an average of 1.7 determined using CoDeML of the PAML package [305] implemented in SynMap; five collinear genes to seed a syntenic block; maximum of 20 non-syntenic genes between syntenic genes to interrupt genomic blocks as previously described [34, 194]. Secondly, within-species ohnologs (i.e. paralogs due to polyploidy) were determined by querying the target genomes against themselves. Microsynteny analysis within and between genomes was performed with GEvo (http://genomeevolution.org/CoGe/GEvo.pl, last accessed on December 13th, 2014). The obtained syntenic gene set output was thereafter cleaned using a retention maximum of three ohnologs for each of the analyzed species.

Anchor paralog identification and protein domain prediction

Ortholog and ohnolog gene sets were combined to create a pool of homologous "anchor" genes. These gene sets of the analyzed target genomes were queried against the *A. thaliana* genome with a

maximum target sequence threshold of 1. Each query sequence that aligned to an *A. thaliana* L-type *LecRK* or *LLP,* but not belonging to the "anchor" gene set, was defined as an anchor paralog. With the above-mentioned steps a complete set of L-type LecRK and LLP-encoding homologs present in every analyzed target species (orthologs, paralogs and ohnologs) was created. As this approach may lead to false positives due to alignment of highly conserved linker sequence pairs an additional filtering step was applied based on HMM-driven protein domain annotation using the iprscan_urlib.py script (https://www.ebi.ac.uk/Tools/webservices/download_clients/python/urllib/iprscan_urllib2.py, last accessed on December 13th, 2014) querying the EMBL server (http://smart.embl-heidelberg.de, last accessed on December 13th, 2014) [374]. Protein motifs were determined using InterProScan 4 (http://www.ebi.ac.uk/Tools/pfa/iprscan, last accessed on December 13th, 2014) [241, 242] and the bioinformatics tools SMART, Superfamily, ProDom, PRINTS, PROSITE, PIR, Pfam, TIGRFAMs, PANTHER, Profile, Gene3D, HAMAP, TMHMM and SignalP.

Identification of mode of gene duplication

Arabidopsis thaliana L-type *LecRK* ohnolog gene copies were obtained based on the blocks described by Bowers, et al. [35] and updated according to Thomas, et al.[85]. Determination of ohnolog duplicates in all other genomes was utilized using the "SynMap" algorithm integrated into the CoGe package for comparative genomics with above-described preferences. Tandem duplicate genes were obtained using BLASTP hits within a maximum of ten consecutive intervening gene spacers as previously described [31]. To identify gene transposition duplicate (GTD) partners among homolog genes, all non-tandem non-ohnolog duplicate target sequences were queried against the whole set of target genes using BLASTP with an e-value threshold of 1e-30. Closest homologs were scored as GTD partners. Putative transpositions were confirmed using the gene transpositional database [33]. Duplicated gene copies belonging to tandem-duplicated ohnologs (TD-α genes) by sharing similar evolutionary patterns with tandem duplicates were obtained and confirmed using the methods described by [350]. Statistical significance of retained ohnolog fractions among target genes compared to the background of genome-wide ohnolog fractions was determined using a Fisher's exact test on count data integrated to the R package for statistical computing (http://www.r-project.org, last accessed on December 13th, 2014).

Coding sequence alignment and determination of Ka/Ks-values to assess divergence

Coding sequence alignments of homologous genes were compiled in Mesquite [375] and manually cleaned to remove premature stop codons and gaps. Other alignments were generated using Prank (http://www.ebi.ac.uk/goldman-srv/webprank, last accessed on December 13th, 2014) [376] with default settings. Ka/Ks was calculated using the KaKs calculator (https://code.google.com/p/kaks-calculator/wiki/KaKs_Calculator, last accessed on December 13th, 2014) [308]. Average divergence rates between respective tandem, ohnologs, gene transposition duplicates and tandem-ohnolog homologous sequences were computed as previously described [145].

Sequence annotation and alignment

Alignments of full length protein sequences were compiled using Mesquite version 2.74 [375]. Removal of stop codons and sequence trimming was performed as previously described [145] ENREF 77. Sequence alignment was performed using Prank relying on default settings (http://www.ebi.ac.uk/goldman-srv/webprank, last accessed on December 13th, 2014) [377].

Phylogenetic analysis

Maximum Likelihood phylogenetic trees were constructed with full-length protein sequences using the RAxML web-server at the CIPRES portal (http://sco.h-its.org/exelixis/web/software/raxml/index.html; last accessed on December 13th, 2014) [378]. Maximum Likelihood searches and estimate proportion of invariable sites were selected as parameters. The robustness of the phylogenetic trees was assessed by performing bootstrap resampling using 100 replicates. All phylogenetic trees were rooted with protein sequences of WAK1 (AT1G21250), PERK1 (AT3G24550), the C-type LecRK AT1G52310, and the G-type LecRKs ARK1 (AT1G65790) and CES101 (AT3G16030). MrBayes version 3.2.2 (http://www.phylo.org/portal2/oldmrbayeshybrid_tg!input.action, last accessed on December 13th, 2014) [201] was used to generate Bayesian trees using the following parameters: rates allowed to vary among four gamma categories; nucleotide state frequencies mixed (Dirichlet model); a uniform gamma shape parameter allowed to vary between 0 to 200 analysis to run for 50 million generations; each generation consisting of two independent runs for four chains each, one of which was heated at a temperature of 0.2 to keep the heated chain in motion; samples were taken every 5000 generations; burn-in time was set at 12,500,000 samples. Bayesian inference trees were constructed by using CIPRES (http://www.phylo.org/sub_sections/portal, last accessed on December 13th, 2014) [202]. Convergence of the parameters and model likelihood between runs were checked in Tracer version 1.5 (http://beast.bio.ed.ac.uk/Tracer, last accessed on December 13th, 2014) after which .p- and .t-files were combined as previously described [379]. Con files (.con) were generated in MrBayes and contained the Bayesian 50%-majority rule consensus trees. FigTree software was used to generate and edit the phylogenetic trees (http://tree.bio.ed.ac.uk/software/figtree, last accessed on December 13th, 2014). Results were scored positive once the effective sampling size (ESSs) of all parameters was above 100. Tree branches supported by posterior probabilities (PP) below 0.7 were considered as weak and above 9.0 as strong.

RESULTS

Curation of *A. thaliana* L-type *LecRK*s and *LLP*s

In a first step, we compiled a list of the 45 L-type *LecRK*s that have previously been described in *A. thaliana* (**supplementary table 1,** Supplementary Material online) [78, 355, 364]. Phylogenetic analysis placed 37 L-type *LecRK*s into nine distinct clades and identified seven singleton genes (e.g. no clear relationship to one distinct clade or "ambiguous" genes). We confirmed these previous results with our phylogenetic analysis (**fig. 1A, B**). Furthermore, we included the ten *LLP* genes previously identified by Armijo, et al. [380] that encode so-called Legume-like lectin proteins. LLPs contain a legume-like lectin domain but lack a kinase domain. In addition, we found another *LLP*; bringing the total count of *A. thaliana LLPs* to eleven **(table 1)**. In summary, the *A. thaliana* genome encodes 56 proteins containing a putative legume-like lectin domain (IPR001220) (**fig. 2A**). For the eleven *LLP*s, we propose a uniform gene nomenclature based on their phylogenetic relationship (**fig. 1A**). LLPs form two strongly supported monophyletic clades, one consisting of six and the other of four members. The remaining one is an ambiguous LLP because it groups distantly. In line with the nomenclature proposed by Bouwmeester and Govers [355], the *LLP* clades were named using Roman numerals. The largest clade of six comprises LLPs that lack a transmembrane domain, for which we propose the term L-type lectin proteins (Clade I: LecPs). In contrast, the other LLP members share in addition to the legume-like lectin domain a transmembrane domain, and are herewith proposed to be named L-type lectin receptor proteins (Clade II: LecRPs) **(table 1, fig. 2A)**. Interestingly, our phylogenetic analysis shows that LecP-I.1 (At1g07460) groups with the LecRK-III clade (**fig. 1A**), whereas all other LecPs show a shared sequence similarity with L-type LecRKs belonging to clade VII (**fig. 1A**), and this indicates that LLPs share independent evolutionary histories with L-type LecRKs.

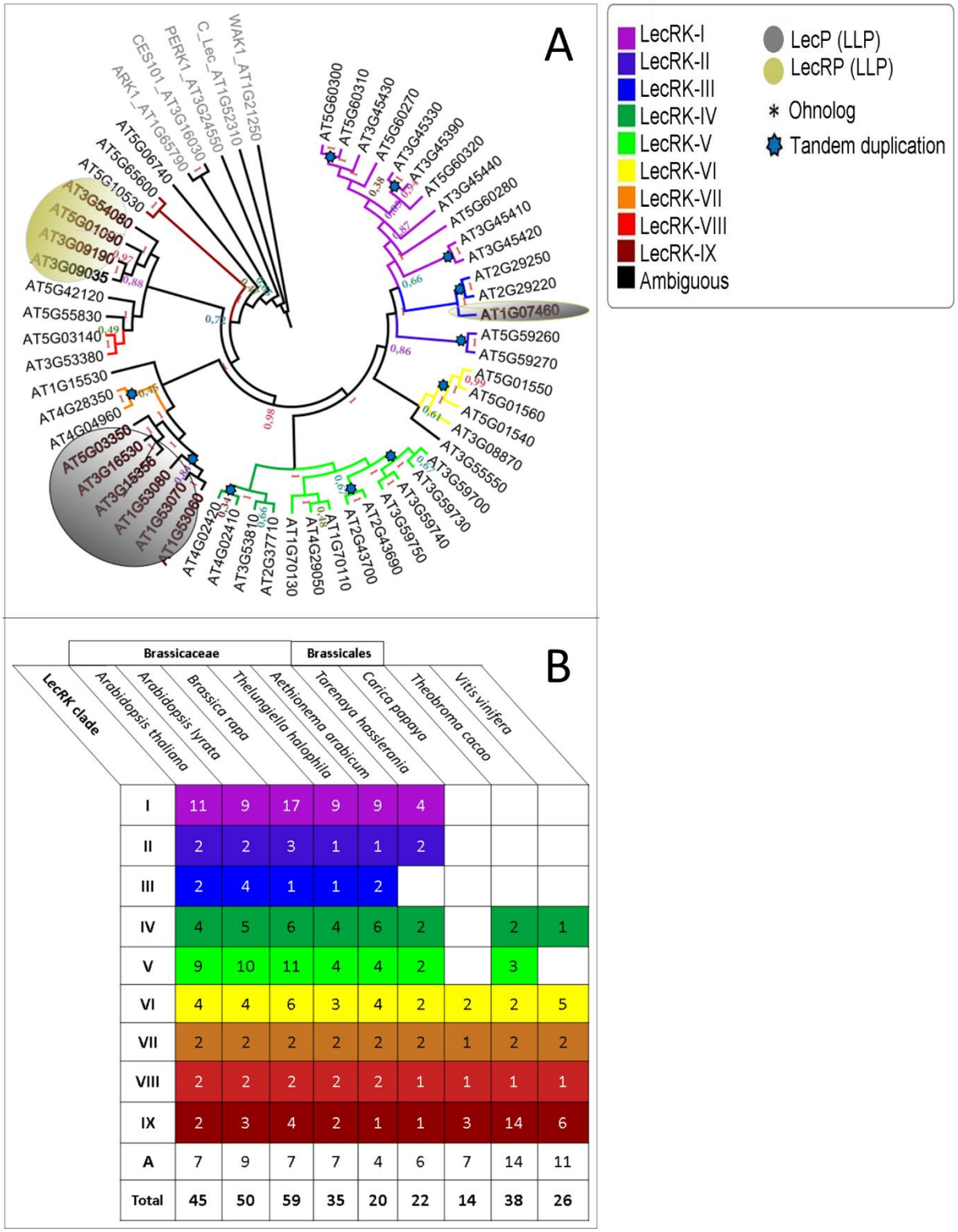


FIG. 1.— Phylogeny and classification of A. thaliana L-type LecRKs and LLPs.
A. Phylogeny of 43 full-length L-type LecRKs and eleven LLPs in *A. thaliana*. We identified two LLP clades; LecPs (lacking transmembrane domains) and LecRPs (with transmembrane domains) which are highlighted in dark grey and ochre, respectively. Color-coding was adapted according to Bouwmeester and Govers (2009) [44]. Tandem duplication events are indicated by light blue stars. The tree was rooted using the *A. thaliana* G-type *LecRK*s CES101 and ARK1, the C-type LecRK AT1G21250 and the Wall-associated kinases WAK1 and PERK1. Clade-support bootstrap values range from to 0.80 to 0.94. **B**. Clade assignment of 309 *LecRK*s identified across nine analyzed genome annotations. Colors represent the nine clades as previously described [44]. A refers to ambiguous genes (singletons).

TABLE 1- Classification of *LLP* loci in *A. thaliana* including information on encoded proteins

	Gene information					Protein information				
Proposed classification	Proposed gene name	Locus	Tandem duplcate	Length (bp)	Uniprot accession	Length (AA)	Signal peptide	No. of TM motifs	Domain configuration	Reference
LecRP	*LecRP-I.1*	AT3G09035	yes	1017	Q3EBA4	338	yes	1	Lectin_legB-(TM)	Armijo *et al* .(2013)
	LecRP-I.3	AT3G09190	yes	1038	Q9SS71	345	yes	1	Lectin_legB-(TM)	this manuscript
	LecRP-S.1	AT3G54080	no	1053	Q9M395	350	yes	1	Lectin_legB-(TM)	Armijo *et al.* (2013)
	LecRP-I.2	AT5G01090	no	1062	Q9LFC7	353	yes	1	Lectin_legB-(TM)	Armijo *et al.* (2013)
LecP	*LecP-I.1*	AT1G07460	no	777	Q4PT39	258	no	0	Lectin_legB	Armijo *et al.* (2013)
	LecP-I.2	AT1G53060	yes	729	Q9LNN1	242	no	0	Lectin_legB	Armijo *et al.* (2013)
	LecP-I.3	AT1G53070	yes	819	Q9LNN2	272	yes	0	Lectin_legB	Armijo *et al.* (2013)
	LecP-I.4	AT1G53080	yes	852	Q9LNN3	283	yes	0	Lectin_legB	Armijo *et al.* (2013)
	LecP-I.5	AT3G15356	no	816	Q9LJR2	271	yes	0	Lectin_legB	Armijo *et al.* (2013)
	LecP-I.6	AT3G16530	no	831	Q9LK72	276	yes	0	Lectin_legB	Armijo *et al.* (2013)
	*LecP-I.7**	AT5G03350	no	825	Q9LZF5	274	yes	0	Lectin_legB	Armijo *et al.* (2013)

* alias *SAI-LLP1* (Armijo *et al* ., 2013)

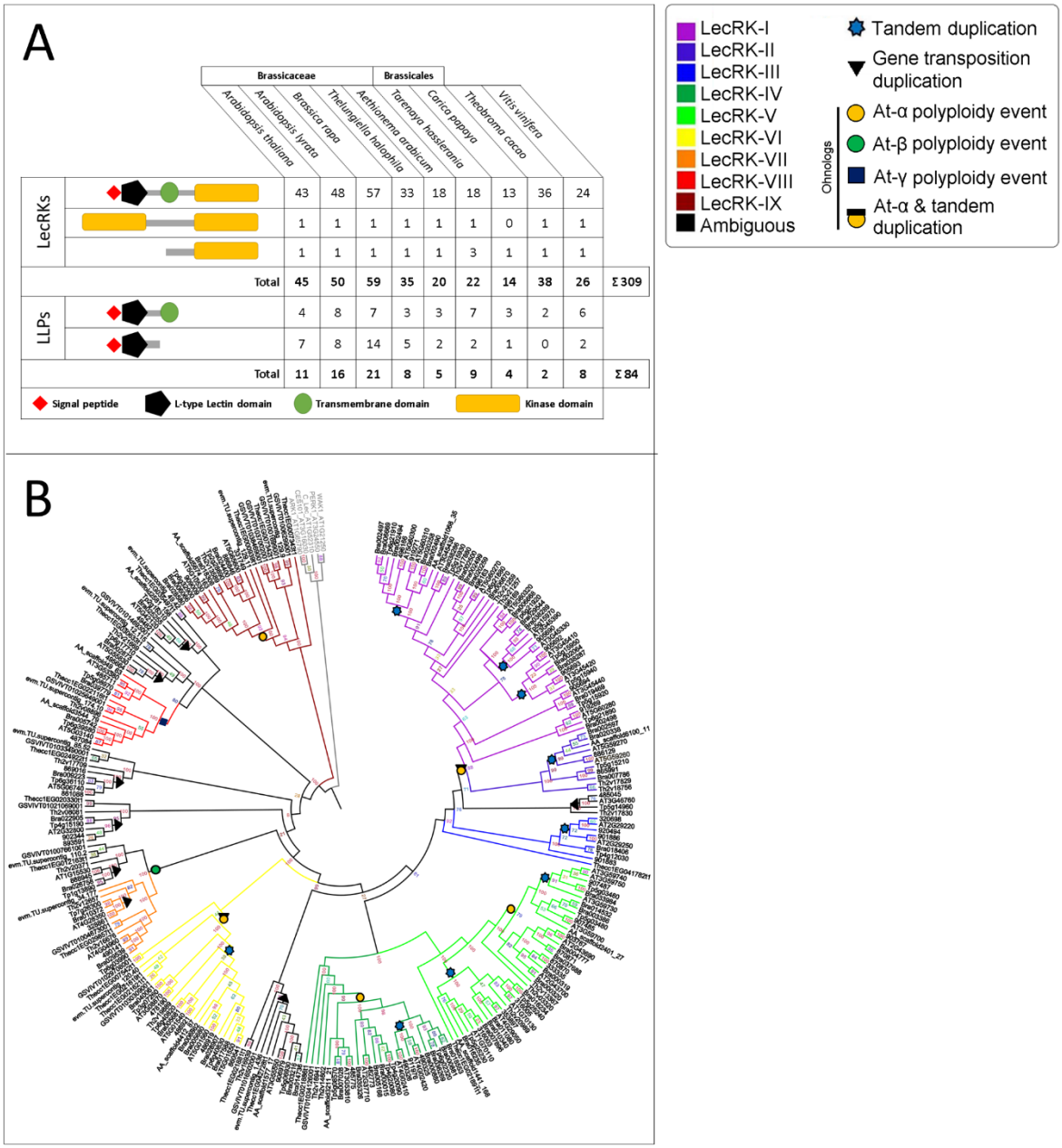


FIG. 2.— **Classification of L-type LecRKs and LLPs identified in nine plant species**. **A**. Domain composition of 309 L-type LecRKs and 84 LLPs across Brassicaceae, Brassicales, *T. cacao* and *V. vinifera*. L-type LecRKs containing two kinase domains are present in all analyzed species except *C. papaya.* Note that the *T. cacao* lacks LLPs. **B**. Cladogram based on the legume-like lectin domains of Brassicaceae L-type LecRKs from *A. thaliana*, *A. lyrata*, *B. rapa, T. halophila* and *A. arabicum.* Further included are 63 legume-like lectin domain sequences from three other families: *T. hasslerania* (Cleomaceae), *T. cacao (Malvaceae), C. papaya* (Caricaceae) and *V. vinifera* (Vitaceae) with support values indicated on key nodes. Number-only IDs refer to expressed genes present in the "Araly1"-annotation (*A. lyrata*). The phylogenetic tree was rooted with the extracellular domains of the G-type LecRKs CES101 and ARK1, the C-type LecRK AT1G52310 and the Wall-associated kinases WAK1 and PERK1 as outgroup sequences. Clade support bootstrap values range from 0.70 to 0.95. For all species, the L-type LecRKs cluster to nine distinct clades (colored) corresponding to the clade assignment of the *A. thaliana* L-type LecRKs including those without clear affiliation to a distinct clade (ambiguous). Symbols placed on nodes represent the different duplication modes: i.e. At-α whole genome duplication event (orange circles), At-α ohnologs subjected to tandem duplication (TD-α genes) (orange circle with black square), tandem duplication event (light blue stars), gene transposition duplicates (black triangle) and more ancient polyploidy events: At-β (blue square) and At-γ (green circles). Symbols mark last common duplication events. Six of nine clades are specific to Brassicaceae, Cleomaceae and Caricaceae while the rest of

the clades are shared between Brassicales and Vitales. Ambiguous LecRKs are spread across the tree and across families.

Duplication analysis of *A. thaliana* L-type *LecRK*s and *LLP*s

To establish the relationship between gene duplication and genetic divergence, the chromosomal locations of the L-type *LecRKs* were confirmed using the *A. thaliana* locus codes. Results show that the L-type *LecRK*s are organized in nine gene clusters distributed over the five *A. thaliana* chromosomes, with the highest density on chromosome V followed by chromosome III [355]. All L-type *LecRKs* are located to regions covered by ohnolog blocks due to the most recent ancient polyploidy event. Furthermore, we localized the two *A. thaliana* clusters possessing the highest target gene density at two independent tandem duplicate supergene clusters on chromosomes III and V. *Arabidopsis thaliana* clade V L-type *LecRK*s are located in proximity on chromosomes I, II and III **(supplementary table 1)**. Notably, one large tandem array containing *LecRK-V.5* (At3g59700), *LecRK-V.6 (*At3g59730), *LecRK-V7* (At3g59740) and *LecRK-V8* (At3g59750) was found to be specific for *A. thaliana* since orthologs in all other species were singletons (array 14 in **supplementary table 1,** Supplementary Material online). Likewise, we investigated the genomic locations of the *A. thaliana LLP*s. These were predominantly located on chromosomes III and V, of which several cluster together with L-type *LecRK*s. Among these are *LecRP-I.1* (At3g09035) and *LecRP-I.3* (At3g09190) that share chromosomal location with *LecRK-VI.1* (At3g08870). Moreover, *LecRP-I.2* (At5g01090) and *LecRK-VI.2* (At5g01540) are located in each other chromosomal proximity **(Supplemental table 1)**. Again, this shows that that *LLP*s share an evolutionary history with L-type *LecRK*s. In this context, the observed degree of sequence similarity and domain conservation may be due to ancient sub- and neo-functionalization following gene and genome duplication.

Domain conservation and ortholog retention across the Brassicaceae

As a next step, a combination of *A. thaliana* L-type *LecRK* orthologs was obtained for eight genome assemblies by RBH analysis (**supplementary table 2,** Supplementary Material online). Likewise, L-type *LecRK* ohnologs were curated for all analyzed genomes (**supplementary table 3,** Supplementary Material online). Both datasets were merged to create a pool of "anchor" genes for every analyzed genome annotation. This pool of putative "anchor" genes was used in an additional BLAST analysis against the various genomes to screen for target gene paralogs. This additional screen was necessary because it became evident that ortholog assignment based on RBH only misses many true orthologs in lineages with duplicate-rich genomes [71]. In this way, we identified a total of 393 genes encoding a legume-like lectin domain, of which 309 are L-type *LecRKs* (**fig. 2A**). In line with the phylogenetic relationship of the *A. thaliana* L-type *LecRK*s, all Brassicaceae contain the nine clades of L-type *LecRK*s and at least four ambiguous gene family members that encode proteins with the conserved L-type LecRK domain composition (**fig. 2**). However, species-specific differences apply with increased phylogenetic distance. For example, *T. hasslerania* of the Cleomaceae is a closely related sister lineage to all mustard family members but its genome annotation does not contain clade III orthologs. The more distant species *T. cacao* lacks L-type LecRKs aligning to clades I, II and III, and *C. papaya* lacks target genes from clade I-V, as well as the orthologs of the two "ambiguous" target genes At2g32800 (*LecRK-S.2*) and At3g46760 (*LecRK-S.3*). The common grape vine *V. vinifera*, the most distant Brassicaceae outgroup analyzed in this study, lacks orthologs grouping to L-type *LecRK* clades I, II, III and V.

L-type *LecRK* orthologs and ohnologs across the Brassicaceae and several outgroups

When investigating the genomic context of orthologous target gene pairs, we found that all analyzed genomes retained a fraction of the respective orthologs within a given syntenic region (i.e. are syntenic to *A. thaliana* L-type *LecRK* orthologs or ohnologs) **(supplementary tables 2-3**, Supplementary Material online**)**. Notably, the closest related sister lineage *A. lyrata* has ohnologs to 39 *A. thaliana* L-type *LecRK*s, corresponding to a retention score of 87% (**supplementary table 3,** Supplementary Material online). This score decreases with increased phylogenetic distance of Brassicaceae lineages, as indicated by the values for the crop *B. rapa* (78%), the saltwater cress *T. halophila* (62%), the early-diverged mustard *A. arabicum* (69%), and the closest mustard outgroup *T. hasslerania* (36%), as well as the more diverged crop species *C. papaya* (18%), *T. cacao* (29%) and *V. vinifera* (16%). These results are consistent with previous studies reporting an erosion of synteny across lineages relative to their phylogenetic distance [26, 145]. In addition, we investigated the retention of *LLP*s in the various Brassicaceae species. This revealed that between 91% (*A. lyrata*) and 45% (*A. arabicum*) of all *LLP*s identified in *A. thaliana* are retained within the analyzed Brassicaceae species. Interestingly, the Brassicaceae outgroup *T. hasslerania* retained a higher fraction of *LLP* orthologs (73%) than the basal Brassicaceae *A. arabicum*, and this is consistent with the species-specific genome triplication event evident for this Cleomaceae species [53]. In contrast, all other Brassicales as well as *V. vinifera* only contain one *LLP* gene which is orthologous to *A. thaliana LecRP-I.2* (AT5G01090), corresponding to a retention score of 10% (**supplementary table 4-5**, Supplementary Material online).

Different modes of duplication affect L-type *LecRK* and *LLP* copy number variation

In a next step, we identified both tandem- and whole genome duplication events that have influenced copy number variation and molecular evolution of the L-type *LecRK* and *LLP* gene families across all analyzed genomes. For *A. thaliana* L-type *LecRKs* and *LLPs*, we scored both tandem- and ohnolog duplicates based on previously published definitions (see Materials & Methods section) (**table 2, supplementary table 6,** Supplementary Material online). The obtained results revealed that a relatively large fraction of ohnologs (37%) was retained from ancient polyploidy events among all identified L-type *LecRK* and *LLP* genes within all genomes. Compared to the average of genome-wide ohnolog fractions across all genomes (i.e. 30%), this indicates a significant over-retention of whole genome duplicates among L-type *LecRK*s and *LLP*s (**table 3**). Note that species-specific differences apply. For example, *B. rapa* and *T. hasslerania*, which both underwent a lineage-specific genome triplication event, show higher fractions of genome-wide ohnologs compared to the other lineages (53% and 48%, respectively, compared to 22% for *A. thaliana*). In summary, statistical analysis based on a Fisher's exact test revealed a significant enrichment of ohnologs among genes encoding a legume-like lectin domain for five of the nine genomes that we investigated (**table 3**). Likewise, we identified a 55% fraction of genes in tandem arrays among all identified L-type *LecRK*s and *LLP*s (**table 4**). All identified tandem duplicate genes group to a sum of 54 distant tandem arrays distributed across all analyzed genomes (**supplementary table 1,** Supplementary Material online), with an average of 2.9 genes per tandem array and 5.9 genes in the largest identified tandem array (**table 4**). Again, differences were detected in the species-wise tandem duplicate fractions among target genes, varying from 29% in *A. lyrata* to 68% in *T. cacao* (**table 4**). We could, however, not detect any tandem duplicate genes in *A. arabicum*. This could be due to the fact that we have used a draft version of the *A. arabicum* genome annotation that is based on large-scale integration of RNAseq data. This draft version may include mis-annotations of small open reading frames with fusions of tandem duplicates due to similar transcripts. The number of independent clusters of

tandem duplicates was found to vary across species; i.e. from eleven distant tandem arrays in the Brassicaceae species *B. rapa* (that underwent a species-specific genome triplication) to one tandem array only in the Brassicales crop *C. papaya*. The more distant Brassicales crop *T. cacao* contains the highest average number of genes per array, whereas tandem arrays in the most distant analyzed outgroup *V. vinifera* are lowest in gene count across all analyzed species. Assessment of gene count within the largest array present in all analyzed species revealed maximums of ten for *B. rapa* and minimums of three in *C. papaya* and *V. vinifera* (**table 4**).

TABLE 2. - Duplicate *LLP* gene pairs in *A. thaliana* and mode of duplication

Duplicate one		**Duplicate two**		
AGI	Name	AGI	Name	Duplication mode
AT1G53060	*LecP-I.2*	AT1G53070	*LecP-I.3*	Tandem duplication
AT1G53070	*LecP-I.3*	AT1G53080	*LecP-I.4*	Tandem duplication
AT1G53080	*LecP-I.4*	AT1G53070	*LecP-I.3*	Tandem duplication
AT3G09035	*LecRP-I.1*	AT3G09190	*LecRP-I.3*	Gene transposition duplication
AT3G09190	*LecRP-I.3*	AT3G09035	*LecRP-I.1*	Gene transposition duplication
AT3G15356	*LecP-I.5*	AT3G16530	*LecP-I.6*	Gene transposition duplication
AT3G16530	*LecP-I.6*	AT3G15356	*LecP-I.5*	Gene transposition duplication
AT5G03350	*LecP-I.7*	AT3G15356	*LecP-I.5*	Gene transposition duplication
AT3G54080	*LecRP-S.1*	AT5G01090	*LecRP-I.2*	Ohnolog duplication
AT5G01090	*LecRP-I.2*	AT3G54080	*LecRP-S.1*	Ohnolog duplication
AT1G07460	*LecP-I.1*	AT2G29220	*LecRK-III.1*	Tandem & ohnolog duplication (TD-α genes)

TABLE 3.- Ohnolog duplicate fractions among genes encoding a legume-like lectin domain

	Genome-wide		Genes encoding an L-type lectin domain				
Species	**Number of genes**	**Ohnolog fraction**	***LecRKs***	***LLPs***	**Sum**	**Ohnolog fraction**	**Enrichment***
Arabidopsis thaliana	27,416	22%	45	11	56	29%	yes
Arabidopsis lyrata	32,670	28%	50	16	66	35%	yes
Brassica rapa	40,367	53%	59	21	80	40%	no
Thelungiella halophila	25,191	32%	35	8	43	40%	yes
Aethionema arabicum	22,230	29%	20	5	25	56%	yes
Tarenaya hasslerania	31,580	48%	22	9	31	48%	no
Carica papaya	27,793	7%	14	4	18	11%	no
Theobroma cacao	29,452	32%	38	2	40	33%	no
Vitis vinifera	23,092	22%	26	8	34	38%	yes
	Σ	**30%**	**309**	**84**	**393**	**37%**	**yes**

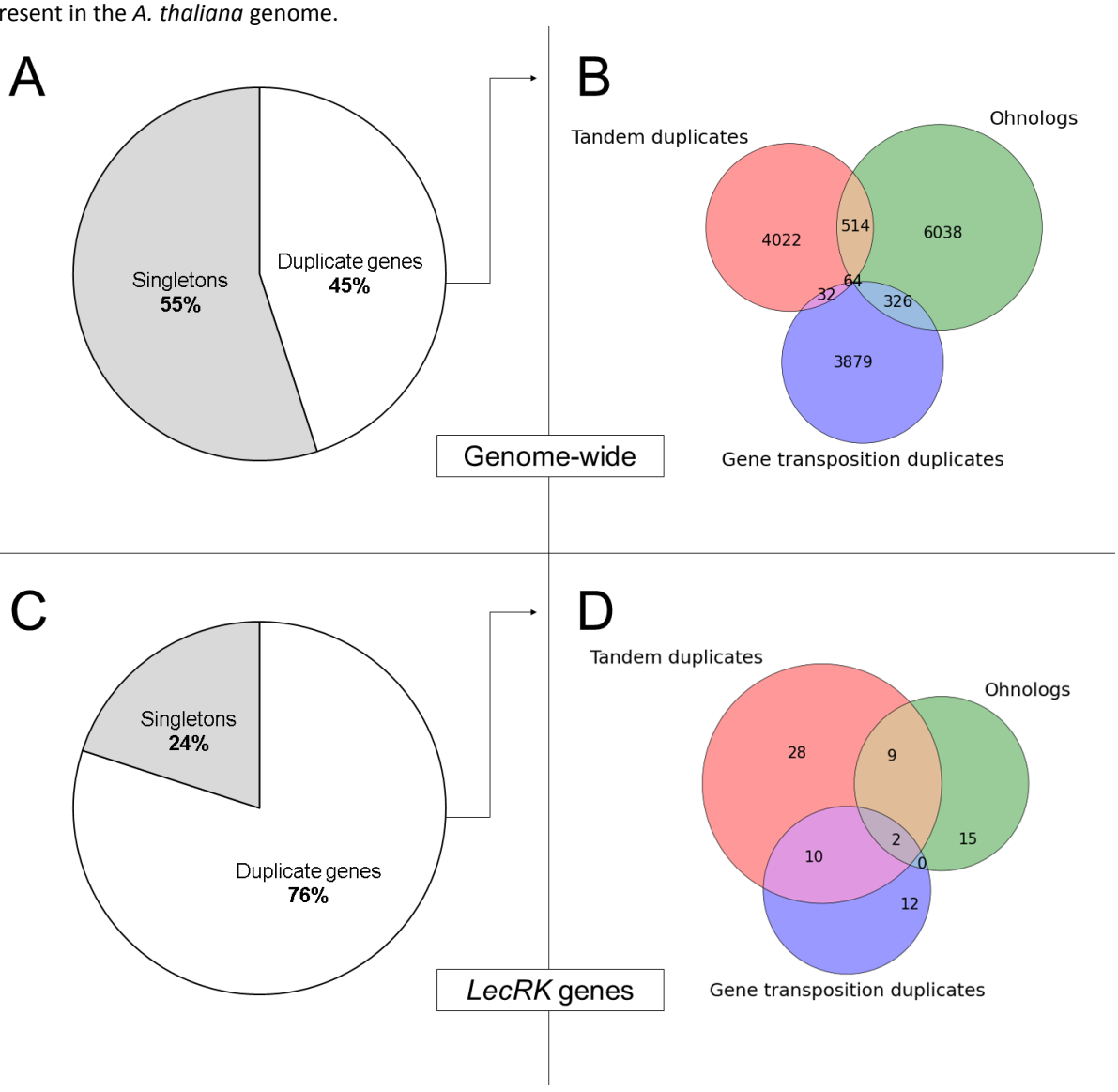
*Accoording to Fisher's exact test (p<0.01)

TABLE 4.- Tandem duplicate fractions among genes encoding a legume-like lectin domain

Species	Genes encoding an L-type lectin domain	Number of tandem duplicates	Fraction of tandem duplicates	Number of tandem arrays	Average size of arrays*	Number of genes in largest array*
Arabidopsis thaliana	56	31	55%	10	3.1	6
Arabidopsis lyrata	66	19	29%	8	2.4	4
Brassica rapa	80	34	43%	11	3.1	10
Thelungiella halophila	43	19	44%	7	2.7	5
Aethionema arabicum #	25	0	0%	0	0	0
Tarenaya hasslerania	31	10	32%	4	2.5	4
Carica papaya	14	3	16%	1	3.0	3
Theobroma cacao	40	27	68%	8	3.4	7
Vitis vinifera	34	11	32%	5	2.2	3
Σ	**389**	**154**	**40%**	**54**	**2.9**	**4.57**

*Tandem array refers to a locus containing one discinct cluster of tandemly arrayed genes

FIG. 3.— Venn-diagrams illustrating genome-wide average and L-type *LecRK* gene duplication fractions. Tandem duplicates (red), ohnolog duplicates (green) and gene transposition duplicates (blue). **A**. Duplicates among all protein-coding genes present in the *A. thaliana* genome. **B**. Duplicates among all L-type *LecRK*s present in the *A. thaliana* genome.



Previous reports indicated that tandem duplicate gene clusters are the birthplace of transposed duplicate copies [100, 350]. For the *A. thaliana* genome, a gene transposition database has previously been made available [33]. These data facilitated scoring of GTD L-type *LecRK* copies. As a result, we referenced all L-type *LecRK* duplicates to either tandem-, ohnolog- or gene transposition duplication modes and compared those to the observed genome-wide fractions of duplicate classes. Initially, we found that 45% of all protein-coding genes in the *A. thaliana* genome comprise duplicate genes (**fig. 3A**). As previously reported for all *A. thaliana* protein-coding genes [44], ohnolog copies comprise 22%, whereas copies due to TD or GTD comprise 15% and 14%, respectively (4022/27416 for TD and 3879/27416 for GTD) (**fig. 3B**). For the subset of L-type *LecRK* genes, we observed different trends in duplication. In *A. thaliana*, 34 of the 45 L-type *LecRK*s comprise duplicates, corresponding to a 76% fraction (**fig. 3C**). In total, we found that 26% of the L-type *LecRKs* in *A. thaliana* transposed at least once after the origin of the Brassicales (i.e. 12 out of 45). Note that all L-type *LecRK* GTD copies are members of tandem duplicate gene clusters (**fig. 3D)**. For the *LLP*s, the GTD fraction is 45%, i.e. 5 out of the 11 *A. thaliana LLPs*. (**supplementary table 1,** Supplementary Material online). Transposition times of most *A. thaliana* GTD copies have been estimated previously [33, 85]. Based on this, we estimated the transposition times for the transposed L-type *LecRKs* to the epoch of At-α (approximately 25-50 MA ago) and even earlier polyploidy events, for example At-β (approximately 50-72 MA ago) which are shared by Brassicales [34, 37]. Many other genes have been reported to have been expanded due to transposition duplication including *B3*, *LCR* and *TRAF* genes that duplicated after *A. thaliana* diverged from *C. papaya* [34]. In this context, we uncovered a connection of GTD and other types of duplications with consequences for molecular evolution (see below).

Hereafter, we assessed the fractions of *A. thaliana* L-type *LecRK* and *LLP* ohnologs that have been subjected to tandem duplication following polyploidy, hereafter termed TD-α duplicates (**table 2, supplementary table 6,** Supplementary Material online**)**. This revealed a 20% fraction of TD-α duplicates among *A. thaliana* L-type *LecRK* genes (nine out of 34 non-singleton genes) (**fig. 3D, supplementary table 1,** Supplementary Material online). This value is consistent with the 20% of TD-α duplicates found among the glucosinolate biosynthetic genes in *A. thaliana* [44] (see Discussion section). In contrast, none of the tandem duplicates among *LLP*s contain ohnologs that date back to the At-α WGD event (**supplementary table 1,** Supplementary Material online). Furthermore, our phylogenetic analysis revealed that TD-α genes are prone to clades I and V, and Brassicales-specific. These two clades are hence the most dynamic L-type *LecRK* clades amongst the analyzed plant species (**fig. 1A**). Here, we show that a 29% fraction of genes retained after ancient polyploidy events for the merged set of *A. thaliana LLP* and L-type *LecRKs* (**table 3**). Likewise, 55% of genes within this merged set comprise members of tandem arrays (**table 4**). Moreover, 30% of L-type *LecRK* and *LLP* genes transposed at least once after the origin of Brassicales, whereas the *A. thaliana* L-type *LecRK*s were found to belong to a GTD fraction of 26% (**supplementary table 1,** Supplementary Material online). In comparison to the genome-wide average, there is a significant difference in the proportions of tandem-, gene transposition- and ohnolog duplicate fractions in L-type *LecRKs* (**fig. 3**). In addition, a clear impact of both tandem- and whole genome duplication (TD-α genes) was detected among the L-type *LecRK* genes.

Molecular evolution of L-type *LecRK*s is impacted by different modes of duplication

Determination of synonymous substitution rates per synonymous sites (Ks) is a common procedure to determine the evolutionary age and divergence level of gene copies [339, 350]. In this context,

comparing divergence rates provides insights into the differential impact of gene duplication modes [81, 318, 339]. Hence, we calculated the Ka/Ks values of the L-type *LecRKs* that date back to different duplication modes in *A. thaliana*. We observed differential patterns of selection following all analyzed duplication modes **(table 5, fig. 4A)**. Tandem duplicate L-type *LecRKs* show the highest average rates of molecular evolution (Ka/Ks=1.23), indicating strong positive or Darwinian selection. Interestingly, lower rates of positive selection were determined for TD-α genes that comprise tandem duplicate ohnolog copies (Ka/Ks=1.13) as well as ohnolog duplicate gene pairs (Ka/Ks=1.11). Ka/Ks-values equal to one indicate neutral (or absence of) selection. L-type *LecRK* copies due to GTD showed the lowest rate of molecular evolution, i.e. a Ka/Ks value of 0.94, implying moderate purifying (or stabilizing) selection **(table 5, fig. 4A)**. The GTD duplicate class comprises mostly ambiguous L-type *LecRKs* and members of clades V and VII (**supplementary table 6,** Supplementary Material online).

Furthermore, we compared gene lengths of L-type *LecRK* copies due to different duplication events using gene-coding sequences (CDS). All CDS were compiled and clustered based on the duplication modes and the difference in coding-region lengths was estimated **(fig. 4B)**. In this analysis, tandem duplicate gene copies display the lowest observed average both for coding-region length and variation thereof, whereas GTD copies display the highest. In contrast, coding-region length of TD-α duplicates display the highest variation. These findings are consistent with previous studies, uncovering a connection between gene length and duplicate origin [34].

TABLE 5.— Molecular evolution rates following different modes of *LecRK* duplication

Duplication mode	Ka*	Ks#	Ka/Ks
Gene transposition duplicates	2.6	2.78	**0.94**
Ohnolog duplicates	2.98	2.68	**1.11**
Tandem & ohnolog duplicates (TD-α genes)	2.58	2.29	**1.13**
Tandem duplicates	2.72	2.42	**1.23**

* Ka = non-synonymous substitutions per non-synonymous site

Ks = synonymous substitutions per synonymous site

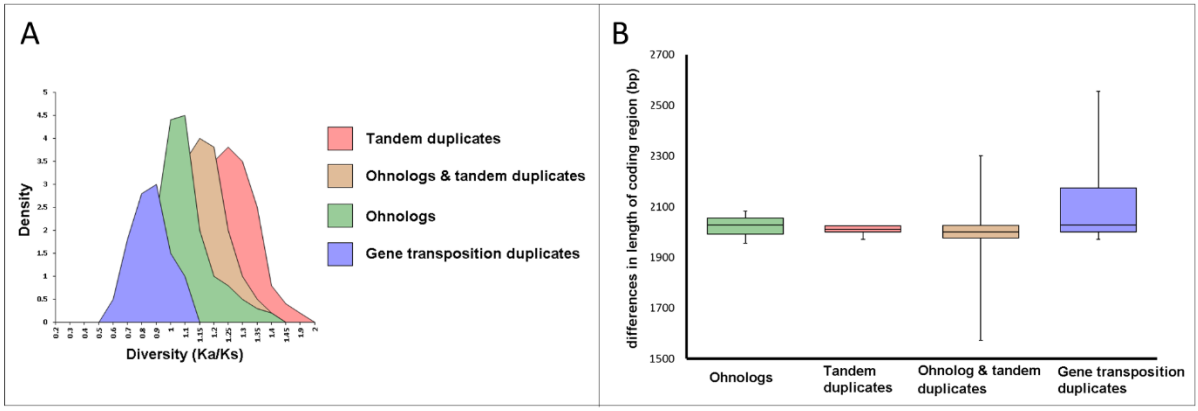


FIG. 4.— Analysis of divergence of L-type *LecRK*s based on mode of gene duplication in *A. thaliana*.

A. Molecular evolution rates of L-type *LecRK* gene pairs based on Ka/Ks values following tandem duplication (red), gene transposition duplication (blue), divergence of ohnologs due to whole genome duplication (green) and divergence of ohnologs that have been subjected to tandem duplication (TD- α genes) (ochre).

B. Divergence of duplicate gene coding sequence length following the aforementioned duplication modes with identical color-coding.

DISCUSSION

As sessile organisms, plants are permanently exposed to a plethora of microbes, including plant pathogens. Hence, the perception of biotic stimuli is crucial for plant survival. The initial detection of these stress factors and subsequent induction of defense signaling is largely governed by receptor-like kinases (RLKs). One class of RLKs considered to function as potential immune receptors are the L-type LecRKs, which comprise an extracellular legume-like lectin domain hypothesized to perceive non-self-associated molecules [364-367]. Also, the LLPs that similar to L-type LecRKs contain a legume-like lectin domain, but are lacking a kinase domain, have been suggested to play roles in plant defense [357, 380].

In this study we used bioinformatics techniques in a comparative genomics approach to elucidate the evolutionary history of the superfamily of legume-like lectin domain-encoding genes in Brassicaceae and related families. This methodology confirmed all previously identified L-type *LecRK*s [355] and identified eleven *LLP* genes in *A. thaliana*; ten of which were described before by Armijo and co-workers [380] (**table 4**). We revealed that 37% of all target genes identified across all species comprise ohnolog gene copies due to whole genome duplication (WGD) events. Compared to the genome-wide averages of duplicates due to polyploidy in all analyzed species, we uncovered a significant enrichment for ohnologs among genes encoding a legume-like lectin domain (**table 1**). Investigating local duplication events, we scored tandem duplicate gene copies among L-type *LecRK*s and *LLP*s in all analyzed species and revealed that the majority of target genes localize to arrays of tandem duplicate genes in *A. thaliana* and *T. cacao* (55% and 68%, respectively). Including all other genome assemblies, a global 40% fraction of all identified L-type *LecRK*s and *LLP*s are organized in tandem arrays (**table 2**). Based on rates for molecular evolution (i.e. Ka/Ks values), we find that tandem duplicate *LecRK*s potentially have been subjected to stronger positive selection in comparison to copies resulting from other duplication modes; a characteristic that also has been described for *NB-LRR* resistance genes in Brassicaceae and Solanaceae [145]. Overall this indicates that the tandem duplication events drive divergence of L-type *LecRK* paralogs and orthologs and thereby could influence functional specialization in plant immunity.

Tandem arrays exist as a result of unequal crossing over (UCO) [90], and result into duplicate genes positioned with a direct orientation. Nine L-type *LecRK* tandem arrays were detected to exhibit a head-to-tail orientation. The exception is the gene pair of *LecRK-V.7* and *LecRK-V.8*, which is positioned in a head-to-head orientation, indicating a potentially shared promoter region. This phenomenon is attributed to intra-chromosomal recombination between direct and indirect repeats. These two L-type *LecRKs* fall under the fraction of tandem duplicate genes that exist as a result of intra-chromosomal recombination in *A. thaliana* [381]. Head-to-head orientation of tandem duplicates has been shown to be relevant for gene function. This includes genes involved in plant innate immunity, as previously shown for the *RRS1*/*RPS4* gene pair that encodes a dual *NB-LRR*-mediated resistance system [287]. Further functional studies to elucidate the contribution of the spatial orientation of *LecRK-V.7* and *LecRK-V.8* in plant immunity is needed, especially since *LecRK-V7* seems to play a role in defense against *Phytophthora* pathogens and the bacterium *Pseudomonas syringae* [367]. For the *A. thaliana* L-type *LecRK*s, our results further demonstrate a significantly increased fraction of gene copies due to a combination of whole genome- and tandem duplication (TD-α genes) compared to the genome-wide average (**fig. 3**). Interestingly, TD-α gene pairs evolve

faster than ohnologs following duplication. The majority of TD-α L-type *LecRK*s groups to clades I and V (**supplementary table 5,** Supplementary Material online). Note that the largest L-type *LecRK* tandem array in *A. thaliana* contains ohnolog copies also while grouping to an under-fractionated homologous genomic region (**supplementary table 1,** Supplementary Material online). Hence, expansion of gene copy number within the L-type *LecRK* clades I and V is largely due to a combination of whole genome- and duplication (TD-α duplication), indicating that their evolution is more dynamic compared to other L-type *LecRK* clades. We hypothesize that the underlying increased copy number occurred at the time of the At-α polyploidy event after the Brassicaceae and Cleomaceae lineage split [54, 118]. This phenomenon was also reported amongst glucosinolate biosynthetic genes, which show a 20% fraction of genes due to TD-α duplication [44]. Also, recent WGD and TD seem to have greatly influenced the expansion and retention of L-type *LecRKs* in clades I, II, III and V amongst core Brassicales, which might be related to the increased degree of functional divergence observed for target genes in this family. More ancient WGD events also had an impact on L-type *LecRK* cluster expansion. We determined that several L-type *LecRKs* duplicated due to more ancient WGD events dating back to the time of divergence of the *A. thaliana* and *C. papaya* lineages approximately 72 MA ago [46] and the divergence of *A. thaliana* and *V. vinifera* 111 MA ago, respectively (**supplementary table 5,** Supplementary Material online). Our comparative analysis also showed evidence for the impact of gene transposition duplication to L-type *LecRK* gene copy number and divergence. All ambiguous L-type *LecRKs*, i.e. those that do not belong to a distinct clade (**fig. 2A**), showed evidence for GTD, which was confirmed in our phylogenetic analysis (**fig. 2B)**. Subjection of genes to GTD may also result into fractionation of gene collinearity, thereby introducing target genes to a novel genomic context and thus influencing functional divergence across L-type *LecRK* clades or even genomes.

Here, we demonstrate that L-type *LecRKs* have undergone all modes of duplication in their evolutionary history, with the highest fraction of duplicates due to WGD and TD. Recent whole genome and tandem duplication have by far most influenced the birth of L-type *LecRKs* and might be a factor for their functional divergence. L-type *LecRK*s form a family whose stability is manifested in the syntenic retention across Brassicaceae species and other closely related species. Earlier findings showed that different duplication events occurred at different times during evolution [34, 46-48, 53, 84, 348]. However, our results demonstrate an exceptional simultaneous occurrence of whole genome and tandem duplication for L-type *LecRKs* across species. This makes the L-type *LecRK* family a highly dynamic and interesting exception amongst several other studied gene families [44, 145]. We also established that *LLP*s cluster into two clades based on sequence homology. It is likely that their origin is due to domain loss from L-type LecRK proteins. Hence, LLPs likely acquired novel functions, however future functional analysis is important to confirm this hypothesis. In this study, we propose a uniform nomenclature for the *A. thaliana* LLPs based upon two criteria: (i) clustering in the phylogenetic tree with PP values >0.9, and (ii) the presence or absence of a transmembrane domain (i.e. the LecRPs versus LecPs) (**fig. 1**). This was inspired by the nomenclature given to the L-type LecRKs by Bouwmeester and Govers (2009) [355]. LLPs share an evolutionary history with L-type LecRKs based on synteny and the monophyletic grouping with specific L-type *LecRK* clades.

Overall, our findings reveal a dynamic evolutionary history of genes encoding a legume-like lectin domain. This divergence is attributed to a complex interplay of whole genome- and tandem duplication events, thus resulting into domain retention and/or loss with subsequent sub- or neofunctionalization. We believe that the highly dynamic birth-death and expansion of these genes have contributed to plant immunity.

ACKNOWLEDGEMENTS AND FUNDING INFORMATION

We would like to acknowledge Tao Zhao for technical support and discussion. This research was supported by The Netherlands Fellowship Program (DLN), a VENI grant from The Netherlands Organization for Scientific Research (KB), and a VIDI and Ecogenomics grant from The Netherlands Organization for Scientific Research (MES).

SUPPLEMENTARY MATERIAL

Supplementary tables 1-6 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/, last accessed on December 13th, 2014).

Supplementary Table 1.— All identified genes encoding a legume-like lectin domain and modes of duplication across nine analyzed genome assemblies

Supplementary Table 2.— Syntelogs to *A. thaliana LecRK* genes across eight target genomes

Supplementary Table 3.— Orthologs to *A. thaliana LecRK* genes across eight target genomes

Supplementary Table 4.— Orthologs to *A. thaliana LLP* genes across eight target genomes

Supplementary Table 5.— Syntelogs to *A. thaliana LLP* genes across eight target genomes

Supplementary Table 6.— Duplicates among *A. thaliana LecRK* genes and duplication mode

GENERAL CONCLUSION

For thousands of years, a detailed understanding of plant biology has accelerated the continued growth and prosperity of mankind. In contrast to the limited area of arable land available on earth, human population realized a near-exponential growth curve within the last 200 years [9]. As a result, crop improvement is now more important than ever in order to ensure human life quality by feeding our planet without destroying it [5].

Recently, a series of scientific and technological innovations facilitated the availability of Big Data covering many or most important crop species on a genomics level [19]. This now facilitates in-depth computational and comparative analysis of genes and genomes in a phylogenomics perspective, thereby unravelling the whole range of within- and between-species diversity present in biochemical pathways associated with specific traits. The work summarized in this thesis is a result of this this fourth or "genomics revolution" and includes a novel and efficient framework for large-scale identification of multi-domain and multi-gene families involved in any desired pathway present in the Angiosperm clade ("Genomics 4.0"). We exemplified this with the analysis of various key gene families involved in traits important for human health and nutrition and provide data that will underpin more rapid gene selection and cloning leading to faster production of more and better food.

Investigating the genomic context of homologous genes summarizes the core innovation provided by this thesis. We highlight five major added values mediated by synteny scoring. First, this allows the curation of duplicate groups due to distinct chromosomal duplication events (ohnolog pairs), thereby revealing the contribution of polyploidy to gene family extension. For example, the results provided in **Chapter 1** highlight a significant over-retention of ohnologs among glucosinolate biosynthetic and regulatory genes in *Arabidopsis*, thereby showing the impact of the most recent whole genome duplication event to innovation within the plant secondary metabolism across the mustard family. Furthermore, our analysis of terpenoid modular pathways summarized in **Chapter 2** confirmed the connection of genome doubling to diversification of secondary metabolite pathways on a broader phylogenetic scale, but also revealed a bias of ohnolog retention towards stoichiometry sensitive sub-modules within composite specialized pathways.

Second, since duplicate groups due to polyploidy often contain highly diverged copies whose affiliation with a distinct gene family is ambiguous based on sequence homology [350], synteny scoring facilitates more accurate identification of all multi-gene family members. This added value is illustrated by the results provided in **Chapter 2** (terpenoid biosynthetic genes) and **Chapter 3** (*NB-LRR* genes). For both gene families, we found previously un-identified loci within many species that have been subjected to extensive, genome-wide analysis in the past. Third, synteny scoring facilitates accurate determination of orthologs and distinguishing paralogs and ohnologs in a phylogenomics order [23]. As a consequence, the ancestral, genome-wide distribution of loci prior to gene family expansion can be determined, leading to clear view of multi-gene family evolution and genome plasticity coinciding with 250 MA of flowering plant radiation when appropriate lineages are selected. For example, the results provided in **Chapter 3** illustrate a dramatic increase of genes associated with all modules of terpenoid biosynthesis, ranging from 50 loci in the basal Angiosperm *Amborella* up to 188 genes in common grapevine *V. vinifera* and include and overview to responsible key factors within all analyzed lineages. Fourth, since structural rearrangements of genomic regions lead to an erosion of synteny across lineages [26], between-species comparison of syntenic regions in a phylogenomics order allows the determination of genomic regions that remained structurally

immobile during flowering plant evolution [35]. In **Chapter 2**, we identified four ortholog groups of *NB-LRR* genes that comprise loci displaying synteny across all twelve analyzed core-eudicot species. Hence, these “gatekeeper” genes are conserved in structurally immobile parts of plant genomes. Interestingly, two of those “gatekeepers” have been shown to convey pleiotropic effects and extended functions in plant innate immunity [335]. In this context, our analysis highlights a connection of genome structural and functional evolution. Fifth, the distinction of homologs due to different duplication events facilitates a comparative survey of gene molecular evolution rates following tandem-, gene transposition- and whole genome duplication, thereby providing insights to the genomics basis of plant variation, innovation and success. In this context, the results provided in **Chapter 4** show a differential impact to molecular evolution following polyploidy and short sequence duplication. In addition, the results of **Chapter 4** illustrate a complex interplay of tandem- and whole genome duplication modes that forms a previously un-investigated duplication class and highlights its far-ranging consequences for copy number and divergence rates of L-type *LecRK* genes within various representative species.

In summary, we took advantage of recent developments in science and technology and followed the mission statement of Wageningen University & Research Center to exploit the potential of nature to improve quality of life. First, we introduced an easy-to-use meta-method for gene identification with multitudes of possible applications all across the broad community of genetics and genomics researchers. Second, we curated data on more than 4500 loci that can underpin crop improvement for food production and made them available to the public domain by publication in several scientific high-impact open access journals. Third, we shed light upon the consequences of plant polyploidy to trait evolution and thereby contributed to the increased understanding of plant biology that has facilitated the prosperity of mankind ever since the origin of modern civilization.

1. Barker G: **The Agricultural Revolution in Prehistory: Why did Foragers become Farmers?: Why did Foragers become Farmers?**: Oxford University Press; 2006.
2. Bocquet-Appel J-P: **When the world's population took off: the springboard of the Neolithic Demographic Transition**. *Science* 2011, **333**(6042):560-561.
3. Thompson FM: **The second agricultural revolution, 1815–1880**. *The Economic History Review* 1968, **21**(1):62-77.
4. Strulik H, Weisdorf J: **Population, food, and knowledge: a simple unified growth theory**. *Journal of Economic Growth* 2008, **13**(3):195-216.
5. Döös BR: **Population growth and loss of arable land**. *Global Environmental Change* 2002, **12**(4):303-311.
6. Khush GS: **Green revolution: the way forward**. *Nature reviews Genetics* 2001, **2**(10):815-822.
7. Hazell PB: **The Asian green revolution**, vol. 911: Intl Food Policy Res Inst; 2009.
8. Conway G: **The doubly green revolution: food for all in the twenty-first century**: Cornell University Press; 1998.
9. Godfray HC, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C: **Food security: the challenge of feeding 9 billion people**. *Science* 2010, **327**(5967):812-818.
10. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H: **Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. 1986**. *Biotechnology* 1992, **24**:17-27.
11. Rasmusson D, Phillips R: **Plant breeding progress and genetic diversity from *de novo* variation and elevated epistasis**. *Crop Science* 1997, **37**(2):303-310.
12. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F: **'Green revolution'genes encode mutant gibberellin response modulators**. *Nature* 1999, **400**(6741):256-261.
13. Hedden P: **The genes of the Green Revolution**. *Trends in genetics : TIG* 2003, **19**(1):5-9.
14. Buckler IV ES, Thornsberry JM: **Plant molecular diversity and applications to genomics**. *Current opinion in plant biology* 2002, **5**(2):107-111.
15. Initiative AG: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**(6814):796.
16. Beck S, Sterk P: **Genome-scale DNA sequencing: where are we?** *Current opinion in biotechnology* 1998, **9**(1):116-121.
17. Moore GE: **Cramming more components onto integrated circuits**. In.: McGraw-Hill New York, NY, USA; 1965.
18. Metzker ML: **Sequencing technologies—the next generation**. *Nature Reviews Genetics* 2009, **11**(1):31-46.
19. Michael TP, Jackson S: **The first 50 plant genomes**. *The Plant Genome* 2013, **6**(2).
20. Ouzounis CA: **Rise and demise of bioinformatics? Promise and progress**. *PLoS computational biology* 2012, **8**(4):e1002487.
21. Weckwerth W: **Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology**. *J Proteomics* 2011, **75**(1):284-305.
22. Ma C, Zhang HH, Wang X: **Machine learning for Big Data analytics in plants**. *Trends in plant science* 2014, **19**(12):798-808.
23. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences**. *The Plant Journal* 2008, **53**(4):661-673.
24. Newell MA, Jannink J-L: **Genomic Selection in Plant Breeding**. In: *Crop Breeding.* Springer; 2014: 117-130.
25. Varshney RK, Terauchi R, McCouch SR: **Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding**. *PLoS biology* 2014, **12**(6):e1001883.

26. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D *et al*: **Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids**. *Plant physiology* 2008, **148**(4):1772-1781.
27. Bohra A: **Emerging paradigms in genomics-based crop improvement**. *The Scientific World Journal* 2013, **2013**.
28. Carr G: **Biology 2.0: A special report on the human genome**: Economist Newspaper; 2010.
29. Ohno S: **Evolution by gene duplication**, vol. 1970, 1st edition edn. New York: Springer Publishing Group; 1970.
30. Martinez M: **From plant genomes to protein families: computational tools**. *Computational and structural biotechnology journal* 2013, **8**:e201307001.
31. Rizzon C, Ponger L, Gaut BS: **Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice**. *PLoS computational biology* 2006, **2**(9):e115.
32. Huang CR, Burns KH, Boeke JD: **Active transposition in genomes**. *Annual review of genetics* 2012, **46**:651-675.
33. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D: **Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales**. *Genome research* 2008, **18**(12):1924-1937.
34. Woodhouse MR, Tang H, Freeling M: **Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids**. *The Plant cell* 2011, **23**(12):4241-4253.
35. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**(6930):433-438.
36. Tang H, Bowers JE, Wang X, Paterson AH: **Angiosperm genome comparisons reveal early polyploidy in the monocot lineage**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(1):472-477.
37. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al*: **Ancestral polyploidy in seed plants and angiosperms**. *Nature* 2011, **473**(7345):97-100.
38. Blanc G, Wolfe KH: **Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes**. *The Plant cell* 2004, **16**(7):1667-1678.
39. Vision TJ, Brown DG: **Genome archaeology: detecting ancient polyploidy in contemporary genomes**. In: *Comparative genomics.* Springer Netherlands; 2000: 479-491.
40. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in *Arabidopsis***. *Science* 2000, **290**(5499):2114-2117.
41. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ *et al*: **A genome triplication associated with early diversification of the core eudicots**. *Genome biology* 2012, **13**(1):R3.
42. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome**. *Genome research* 2003, **13**(2):137-144.
43. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM *et al*: **An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions**. *Nat Genet* 2013, **45**(8):891-898.
44. Hofberger JA, Lyons E, Edger PP, Chris Pires J, Eric Schranz M: **Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family**. *Genome biology and evolution* 2013, **5**(11):2155-2173.
45. Barker MS, Vogel H, Schranz ME: **Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales**. *Genome biology and evolution* 2009, **1**:391-399.

46. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL *et al*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.
47. Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, Maere S, Van de Peer Y, Geuten K: **Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for *MADS*-Box Gene and Species Diversification**. *Mol Biol Evol* 2012, **29**(12):3793-3806.
48. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**(7161):463-467.
49. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(26):9903-9908.
50. Jiao Y, Li J, Tang H, Paterson AH: **Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots**. *The Plant cell* 2014, **26**(7):2792-2802.
51. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F *et al*: **The genome of the mesopolyploid crop species *Brassica rapa***. *Nat Genet* 2011, **43**(10):1035-1039.
52. Tang H, Lyons E: **Unleashing the genome of *Brassica rapa***. *Frontiers in plant science* 2012, **3**:172.
53. Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, de Bruijn S, Bhide AS, Kuelahoglu C *et al*: **The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers**. *The Plant cell* 2013, **25**(8):2813-2830.
54. Schranz ME, Mitchell-Olds T: **Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae**. *The Plant cell* 2006, **18**(5):1152-1165.
55. Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP *et al*: **Genome duplication in soybean (*Glycine* subgenus soja)**. *Genetics* 1996, **144**(1):329-338.
56. Schlueter JA, Scheffler BE, Jackson S, Shoemaker RC: **Fractionation of synteny in a genomic region containing tandemly duplicated genes across *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana***. *The Journal of heredity* 2008, **99**(4):390-395.
57. Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ: **Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families**. *Syst Biol* 2005, **54**(3):441-454.
58. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al*: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**(7278):178-183.
59. Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J: **Close split of *Sorghum* and maize genome progenitors**. *Genome research* 2004, **14**(10A):1916-1923.
60. Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S *et al*: **Physical and genetic structure of the maize genome reflects its complex evolutionary history**. *PLoS genetics* 2007, **3**(7):e123.
61. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
62. Fitch WM: **Distinguishing homologous from analogous proteins**. *Systematic zoology* 1970, **19**(2):99-113.
63. Wolfe K: **Robustness—it's not where you think it is**. *Nat Genet* 2000, **25**(1):3-4.
64. Wall DP, Fraser HB, Hirsh AE: **Detecting putative orthologs**. *Bioinformatics* 2003, **19**(13):1710-1711.

65. Alkatib S, Scharff LB, Rogalski M, Fleischmann TT, Matthes A, Seeger S, Schöttler MA, Ruf S, Bock R: **The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code**. *PLoS genetics* 2012, **8**(11):e1003076.
66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
67. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling**. *J Mol Biol* 1994, **235**(5):1501-1531.
68. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA: **Structure, function and evolution of multidomain proteins**. *Current opinion in structural biology* 2004, **14**(2):208-216.
69. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L *et al*: **The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation**. *Nucleic acids research* 2008, **36**(suppl 1):D1009-D1014.
70. Wang H, Wu J, Sun S, Liu B, Cheng F, Sun R, Wang X: **Glucosinolate biosynthetic genes in *Brassica rapa***. *Gene* 2011, **487**(2):135-142.
71. Dalquen DA, Dessimoz C: **Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals**. *Genome biology and evolution* 2013, **5**(10):1800-1806.
72. Guo YL, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D: **Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis***. *Plant physiology* 2011, **157**(2):757-769.
73. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome research* 2003, **13**(9):2178-2189.
74. Mazourek M, Cirulli ET, Collier SM, Landry LG, Kang BC, Quirin EA, Bradeen JM, Moffett P, Jahn MM: **The fractionated orthology of *Bs2* and *Rx/Gpa2* supports shared synteny of disease resistance in the Solanaceae**. *Genetics* 2009, **182**(4):1351-1364.
75. Mithen R, Bennett R, Marquez J: **Glucosinolate biochemical diversity and innovation in the Brassicales**. *Phytochemistry* 2010, **71**(17-18):2074-2086.
76. Cheng A-X, Lou Y-G, Mao Y-B, Lu S, Wang L-J, Chen X-Y: **Plant Terpenoids: Biosynthesis and Ecological Functions**. *J Integr Plant Biol* 2007, **49**(2):179-186.
77. McHale L, Tan X, Koehl P, Michelmore RW: **Plant NBS-LRR proteins: adaptable guards**. *Genome biology* 2006, **7**(4):212.
78. Yan W, Bouwmeester K, Beseh P, Shan W, Govers F: **Phenotypic analyses of *Arabidopsis* T-DNA insertion lines and expression profiling reveal that multiple L-type lectin receptor kinases are involved in plant immunity**. *Molecular Plant-Microbe Interactions* 2014(ja).
79. Hartmann T: **From waste products to ecochemicals: fifty years research of plant secondary metabolism**. *Phytochemistry* 2007, **68**(22-24):2831-2846.
80. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA: **Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms**. *Journal of experimental zoology Part B, Molecular and developmental evolution* 2007, **308**(1):58-73.
81. Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Paterson AH: **Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms**. *PloS one* 2011, **6**(12):e28150.
82. Stehle F, Brandt W, Schmidt J, Milkowski C, Strack D: **Activities of *Arabidopsis* sinapoylglucose: malate sinapoyltransferase shed light on functional diversification of serine carboxypeptidase-like acyltransferases**. *Phytochemistry* 2008, **69**(9):1826-1831.
83. Schranz ME, Edger PP, Pires JC, van Dam NM, Wheat CW: **Comparative Genomics in the Brassicales: Ancient Genome Duplications, Glucosinolate Diversification and *Pierinae* Herbivore Radiation**. In: *Genetics, genomics and breeding of oilseed brassicas.* 2011: 206-218.

84. Schranz ME, Mohammadin S, Edger PP: **Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model**. *Current opinion in plant biology* 2012, **15**(2):147-153.
85. Thomas BC, Pedersen B, Freeling M: **Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes**. *Genome research* 2006, **16**(7):934-946.
86. De Bodt S, Maere S, Van de Peer Y: **Genome duplication and the origin of angiosperms**. *Trends in ecology & evolution* 2005, **20**(11):591-597.
87. Irish VF, Litt A: **Flower development and evolution: gene duplication, diversification and redeployment**. *Current opinion in genetics & development* 2005, **15**(4):454-460.
88. Freeling M, Thomas BC: **Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity**. *Genome research* 2006, **16**(7):805-814.
89. Fawcett JA, Maere S, Van de Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(14):5737-5742.
90. Kane J, Freeling M, Lyons E: **The evolution of a high copy gene array in *Arabidopsis***. *Journal of molecular evolution* 2010, **70**(6):531-544.
91. Parniske M, Wulff BB, Bonnema G, Thomas CM, Jones DA, Jones JD: **Homologues of the *Cf-9* disease resistance gene (*Hcr9s*) are present at multiple loci on the short arm of tomato chromosome 1**. *Molecular plant-microbe interactions : MPMI* 1999, **12**(2):93-102.
92. Leister D: **Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene**. *Trends in genetics : TIG* 2004, **20**(3):116-122.
93. Bellieny-Rabelo D, Oliveira AE, Venancio TM: **Impact of whole-genome and tandem duplications in the expansion and functional diversification of the F-box family in legumes (Fabaceae)**. *PloS one* 2013, **8**(2):e55127.
94. Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T: **Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis***. *The Plant cell* 2001, **13**(3):681-693.
95. Li J, Hansen BG, Ober JA, Kliebenstein DJ, Halkier BA: **Subclade of flavin-monooxygenases involved in aliphatic glucosinolate biosynthesis**. *Plant physiology* 2008, **148**(3):1721-1733.
96. Heidel AJ, Clauss MJ, Kroymann J, Savolainen O, Mitchell-Olds T: **Natural variation in MAM within and between populations of *Arabidopsis lyrata* determines glucosinolate phenotype**. *Genetics* 2006, **173**(3):1629-1636.
97. Textor S, de Kraker JW, Hause B, Gershenzon J, Tokuhisa JG: **MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in *Arabidopsis***. *Plant physiology* 2007, **144**(1):60-71.
98. Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T: **Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100 Suppl 2**:14587-14592.
99. Wicker T, Buchmann JP, Keller B: **Patching gaps in plant genomes results in gene movement and erosion of colinearity**. *Genome research* 2010, **20**(9):1229-1237.
100. Freeling M: **Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition**. *Annual review of plant biology* 2009, **60**:433-453.
101. Hughes AL, Friedman R, Ekollu V, Rose JR: **Non-random association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana***. *Molecular phylogenetics and evolution* 2003, **29**(3):410-416.
102. Wang X, Weigel D, Smith LM: **Transposon variants and their effects on gene expression in *Arabidopsis***. *PLoS genetics* 2013, **9**(2):e1003255.

103. Kliebenstein DJ: **A role for gene duplication and natural variation of gene expression in the evolution of metabolism**. *PloS one* 2008, **3**(3):e1838.
104. Malacarne G, Perazzolli M, Cestaro A, Sterck L, Fontana P, Van de Peer Y, Viola R, Velasco R, Salamini F: **Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving *NBS* resistance genes**. *PloS one* 2012, **7**(1):e29762.
105. Vlad D, Rappaport F, Simon M, Loudet O: **Gene transposition causing natural variation for growth in *Arabidopsis thaliana***. *PLoS genetics* 2010, **6**(5):e1000945.
106. Nakano T, Suzuki K, Fujimura T, Shinshi H: **Genome-wide analysis of the ERF gene family in Arabidopsis and rice**. *Plant physiology* 2006, **140**(2):411-432.
107. Rodman J: **Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies.** *American journal of botany* 1996, **85**(7):10.
108. Windsor AJ, Reichelt M, Figuth A, Svatos A, Kroymann J, Kliebenstein DJ, Gershenzon J, Mitchell-Olds T: **Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae)**. *Phytochemistry* 2005, **66**(11):1321-1333.
109. Beekwilder J, van Leeuwen W, van Dam NM, Bertossi M, Grandi V, Mizzi L, Soloviev M, Szabados L, Molthoff JW, Schipper B *et al*: **The impact of the absence of aliphatic glucosinolates on insect herbivory in *Arabidopsis***. *PloS one* 2008, **3**(4):e2068.
110. Bones AM, Rossiter JT: **The enzymic and chemically induced decomposition of glucosinolates**. *Phytochemistry* 2006, **67**(11):1053-1067.
111. Rask L, Andreasson E, Ekbom B, Eriksson S, Pontoppidan B, Meijer J: **Myrosinase: gene family evolution and herbivore defense in Brassicaceae**. *Plant molecular biology* 2000, **42**(1):93-113.
112. Hecht SS: **Inhibition of carcinogenesis by isothiocyanates**. *Drug metabolism reviews* 2000, **32**(3-4):395-411.
113. Nakajima M, Yoshida R, Shimada N, Yamazaki H, Yokoi T: **Inhibition and inactivation of human cytochrome P450 isoforms by phenethyl isothiocyanate**. *Drug metabolism and disposition: the biological fate of chemicals* 2001, **29**(8):1110-1113.
114. Wittstock U, Kliebenstein DJ, Lambrix V, Reichelt M, Gershenzon J: **Glucosinolate hydrolysis and its impact on generalist and specialist insect herbivores**. In: *Integrative Phytochemistry: from Ethnobotany to Molecular Ecology.* Edited by Romeo JT. Amsterdam: Elsevier; 2003: 101 - 126.
115. Hayes JD, Kelleher MO, Eggleston IM: **The cancer chemopreventive actions of phytochemicals derived from glucosinolates**. *European journal of nutrition* 2008, **47 Suppl 2**:73-88.
116. Fahey JW, Zalcmann AT, Talalay P: **The chemical diversity and distribution of glucosinolates and isothiocyanates among plants**. *Phytochemistry* 2001, **56**(1):5-51.
117. Rodman JE, Karol KG, Price RA, Sytsma KJ: **Molecules, Morphology, and Dahlgren's Expanded Order Capparales**. *Systematic Botany* 1996, **21**(3):289.
118. Couvreur TL, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K: **Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae)**. *Mol Biol Evol* 2010, **27**(1):55-71.
119. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.
120. Sonderby IE, Geu-Flores F, Halkier BA: **Biosynthesis of glucosinolates--gene discovery and beyond**. *Trends in plant science* 2010, **15**(5):283-290.
121. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic acids research* 2002, **30**(14):3059-3066.
122. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0**. *Syst Biol* 2010, **59**(3):307-321.

123. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends in genetics : TIG* 2000, **16**(6):276-277.
124. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome research* 2009, **19**(9):1639-1645.
125. Cheong JJ, Choi YD: **Methyl jasmonate as a vital substance in plants**. *Trends in genetics : TIG* 2003, **19**(7):409-413.
126. Naur P, Petersen BL, Mikkelsen MD, Bak S, Rasmussen H, Olsen CE, Halkier BA: **CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes metabolizing oximes in the biosynthesis of glucosinolates in *Arabidopsis***. *Plant physiology* 2003, **133**(1):63-72.
127. Prasad KV, Song BH, Olson-Manning C, Anderson JT, Lee CR, Schranz ME, Windsor AJ, Clauss MJ, Manzaneda AJ, Naqvi I *et al*: **A gain-of-function polymorphism controlling complex traits and fitness in nature**. *Science* 2012, **337**(6098):1081-1084.
128. Gigolashvili T, Engqvist M, Yatusevich R, Muller C, Flugge UI: **HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana***. *The New phytologist* 2008, **177**(3):627-642.
129. Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K *et al*: **Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(15):6478-6483.
130. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ: **Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways**. *PLoS genetics* 2007, **3**(9):1687-1701.
131. Sawada Y, Kuwahara A, Nagano M, Narisawa T, Sakata A, Saito K, Hirai MY: **Omics-based approaches to methionine side chain elongation in *Arabidopsis*: characterization of the genes encoding methylthioalkylmalate isomerase and methylthioalkylmalate dehydrogenase**. *Plant & cell physiology* 2009, **50**(7):1181-1190.
132. He Y, Chen B, Pang Q, Strul JM, Chen S: **Functional specification of *Arabidopsis* isopropylmalate isomerases in glucosinolate and leucine biosynthesis**. *Plant & cell physiology* 2010, **51**(9):1480-1487.
133. Celenza JL, Quiel JA, Smolen GA, Merrikh H, Silvestro AR, Normanly J, Bender J: **The *Arabidopsis* ATR1 Myb transcription factor controls indolic glucosinolate homeostasis**. *Plant physiology* 2005, **137**(1):253-262.
134. Gigolashvili T, Berger B, Mock HP, Muller C, Weisshaar B, Flugge UI: **The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in *Arabidopsis thaliana***. *The Plant journal : for cell and molecular biology* 2007, **50**(5):886-901.
135. Pfalz M, Vogel H, Kroymann J: **The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in *Arabidopsis***. *The Plant cell* 2009, **21**(3):985-999.
136. Bednarek P, Pislewska-Bednarek M, Svatos A, Schneider B, Doubsky J, Mansurova M, Humphry M, Consonni C, Panstruga R, Sanchez-Vallet A *et al*: **A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense**. *Science* 2009, **323**(5910):101-106.
137. Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM: **Glucosinolate metabolites required for an *Arabidopsis* innate immune response**. *Science* 2009, **323**(5910):95-101.
138. Hansen BG, Kerwin RE, Ober JA, Lambrix VM, Mitchell-Olds T, Gershenzon J, Halkier BA, Kliebenstein DJ: **A novel 2-oxoacid-dependent dioxygenase involved in the formation of the goiterogenic 2-hydroxybut-3-enyl glucosinolate and generalist insect resistance in *Arabidopsis***. *Plant physiology* 2008, **148**(4):2096-2108.

139. Hansen BG, Kliebenstein DJ, Halkier BA: **Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in *Arabidopsis***. *The Plant journal : for cell and molecular biology* 2007, **50**(5):902-910.
140. Piotrowski M, Schemenewitz A, Lopukhina A, Muller A, Janowitz T, Weiler EW, Oecking C: **Desulfoglucosinolate sulfotransferases from *Arabidopsis thaliana* catalyze the final step in the biosynthesis of the glucosinolate core structure**. *The Journal of biological chemistry* 2004, **279**(49):50717-50725.
141. Koonin EV: **Orthologs, paralogs, and evolutionary genomics 1**. *Annu Rev Genet* 2005, **39**:309-338.
142. Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, Baxter S, Abanto M, Bermingham E, Humphray SJ, Rogers J *et al*: **A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies**. *PLoS biology* 2006, **4**(10):e303.
143. Smith JM: **Evolutionary genetics**: Oxford University Press; 1989.
144. Consortium TTG: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**(7400):635-641.
145. Hofberger JA, Zhou B, Tang H, Jones JD, Schranz ME: **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* 2014, **15**(1):966.
146. Gershenzon J, Dudareva N: **The function of terpene natural products in the natural world**. *Nature chemical biology* 2007, **3**(7):408-414.
147. Matsuba Y, Nguyen TT, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schafer P, Kudrna D, Wing RA, Bolger AM *et al*: **Evolution of a complex locus for terpene biosynthesis in *Solanum***. *The Plant cell* 2013, **25**(6):2022-2036.
148. Tholl D, Lee S: **Terpene Specialized Metabolism in *Arabidopsis thaliana***. *The Arabidopsis book / American Society of Plant Biologists* 2011, **9**:e0143.
149. Phillips MA, D'Auria JC, Gershenzon J, Pichersky E: **The *Arabidopsis thaliana* type I isopentenyl diphosphate isomerases are targeted to multiple subcellular compartments and have overlapping functions in isoprenoid biosynthesis**. *The Plant Cell Online* 2008, **20**(3):677-696.
150. Vranová E, Coman D, Gruissem W: **Network analysis of the MVA and MEP pathways for isoprenoid synthesis**. *Annual review of plant biology* 2013, **64**:665-700.
151. Gruchattka E, Hädicke O, Klamt S, Schütz V, Kayser O: ***In silico* profiling of *Escherichia coli* and *Saccharomyces cerevisiae* as terpenoid factories**. *Microbial cell factories* 2013, **12**(1):84.
152. Shack S, Gorospe M, Fawcett TW, Hudgins WR, Holbrook NJ: **Activation of the cholesterol pathway and *Ras* maturation in response to stress**. *Oncogene* 1999, **18**(44):6021-6028.
153. Campos N, Rodriguez-Concepcion M, Sauret-Gueto S, Gallego F, Lois L, Boronat A: ***Escherichia coli* engineered to synthesize isopentenyl diphosphate and dimethylallyl diphosphate from mevalonate: a novel system for the genetic analysis of the 2-C-methyl-D-erythritol 4-phosphate pathway for isoprenoid biosynthesis**. *Biochem J* 2001, **353**:59-67.
154. Johnston JB: **Mechanistic Investigations of Types I and II Isopentenyl Diphosphate Isomerase**: ProQuest; 2007.
155. Han K-H, Kang H-S, Oh S-K, Shin D-h, Yang J-M: **Isopentenyl diphosphate isomerase from *Hevea brasiliensis* and rubber producing method using the same**. In.: Google Patents; 2001.
156. Kang JH, Gonzales-Vigil E, Matsuba Y, Pichersky E, Barry CS: **Determination of residues responsible for substrate and product specificity of *Solanum habrochaites* short-chain cis-prenyltransferases**. *Plant physiology* 2014, **164**(1):80-91.
157. Akhtar TA, Matsuba Y, Schauvinhold I, Yu G, Lees HA, Klein SE, Pichersky E: **The tomato cis-prenyltransferase gene family**. *The Plant journal : for cell and molecular biology* 2013, **73**(4):640-652.
158. Oldfield E, Lin FY: **Terpene biosynthesis: modularity rules**. *Angewandte Chemie* 2012, **51**(5):1124-1137.

159. Bohlmann J, Meyer-Gauen G, Croteau R: **Plant terpenoid synthases: molecular biology and phylogenetic analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(8):4126-4133.
160. Aubourg S, Lecharny A, Bohlmann J: **Genomic analysis of the terpenoid synthase (*AtTPS*) gene family of *Arabidopsis thaliana***. *Molecular genetics and genomics : MGG* 2002, **267**(6):730-745.
161. Kolesnikova MD, Wilson WK, Lynch DA, Obermeyer AC, Matsuda SPT: ***Arabidopsis* camelliol C synthase evolved from enzymes that make pentacycles**. *Org Lett* 2007, **9**(25):5223-5226.
162. Shibuya M, Xiang T, Katsube Y, Otsuka M, Zhang H, Ebizuka Y: **Origin of structural diversity in natural triterpenes: direct synthesis of seco-triterpene skeletons by oxidosqualene cyclase**. *Journal of the American Chemical Society* 2007, **129**(5):1450-1455.
163. Lodeiro S, Xiong Q, Wilson WK, Kolesnikova MD, Onak CS, Matsuda SP: **An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis**. *Journal of the American Chemical Society* 2007, **129**(36):11213-11222.
164. Moses T, Pollier J, Thevelein JM, Goossens A: **Bioengineering of plant (tri)terpenoids: from metabolic engineering of plants to synthetic biology *in vivo* and *in vitro***. *The New phytologist* 2013, **200**(1):27-43.
165. Laszczyk MN: **Pentacyclic triterpenes of the lupane, oleanane and ursane group as tools in cancer therapy**. *Planta medica* 2009, **75**(15):1549-1560.
166. Paetzold H, Garms S, Bartram S, Wieczorek J, Uros-Gracia EM, Rodriguez-Concepcion M, Boland W, Strack D, Hause B, Walter MH: **The isogene 1-deoxy-D-xylulose 5-phosphate synthase 2 controls isoprenoid profiles, precursor pathway allocation, and density of tomato trichomes**. *Molecular plant* 2010, **3**(5):904-916.
167. Carretero-Paulet L, Cairó A, Talavera D, Saura A, Imperial S, Rodríguez-Concepción M, Campos N, Boronat A: **Functional and evolutionary analysis of *DXL1*, a non-essential gene encoding a 1-deoxy-D-xylulose 5-phosphate synthase like protein in *Arabidopsis thaliana***. *Gene* 2013, **524**(1):40-53.
168. Heyndrickx KS, Vandepoele K: **Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources**. *Plant physiology* 2012, **159**(3):884-901.
169. Tissier A: **Glandular trichomes: what comes after expressed sequence tags?** *The Plant Journal* 2012, **70**(1):51-68.
170. Markus Lange B, Turner GW: **Terpenoid biosynthesis in trichomes—current status and future opportunities**. *Plant biotechnology journal* 2013, **11**(1):2-22.
171. Wagner GJ, Wang E, Shepherd RW: **New approaches for studying and exploiting an old protuberance, the plant trichome**. *Annals of botany* 2004, **93**(1):3-11.
172. Tholl D, Chen F, Petri J, Gershenzon J, Pichersky E: **Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from *Arabidopsis* flowers**. *The Plant journal : for cell and molecular biology* 2005, **42**(5):757-771.
173. Yamaguchi S, Sun T, Kawaide H, Kamiya Y: **The *GA2* locus of *Arabidopsis thaliana* encodes ent-kaurene synthase of gibberellin biosynthesis**. *Plant physiology* 1998, **116**(4):1271-1278.
174. Bohlmann J, Martin D, Oldham NJ, Gershenzon J: **Terpenoid Secondary Metabolism in *Arabidopsis thaliana*: cDNA Cloning, Characterization, and Functional Expression of a Myrcene-β-Ocimene Synthase**. *Archives of Biochemistry and Biophysics* 2000, **375**(2):261-269.
175. Chen F, Tholl D, D'Auria JC, Farooq A, Pichersky E, Gershenzon J: **Biosynthesis and emission of terpenoid volatiles from *Arabidopsis* flowers**. *The Plant cell* 2003, **15**(2):481-494.
176. Chen F, Ro DK, Petri J, Gershenzon J, Bohlmann J, Pichersky E, Tholl D: **Characterization of a root-specific *Arabidopsis* terpene synthase responsible for the formation of the volatile monoterpene 1,8-cineole**. *Plant physiology* 2004, **135**(4):1956-1966.

177. Fäldt J, Arimura G-i, Gershenzon J, Takabayashi J, Bohlmann J: **Functional identification of *AtTPS03* as (E)-β-ocimene synthase: a monoterpene synthase catalyzing jasmonate-and wound-induced volatile formation in *Arabidopsis thaliana***. *Planta* 2003, **216**(5):745-751.
178. Herde M, Gartner K, Kollner TG, Fode B, Boland W, Gershenzon J, Gatz C, Tholl D: **Identification and regulation of TPS04/GES, an *Arabidopsis* geranyllinalool synthase catalyzing the first step in the formation of the insect-induced volatile C16-homoterpene TMTT**. *The Plant cell* 2008, **20**(4):1152-1168.
179. Huang M, Abel C, Sohrabi R, Petri J, Haupt I, Cosimano J, Gershenzon J, Tholl D: **Variation of herbivore-induced volatile terpenes among *Arabidopsis* ecotypes depends on allelic differences and subcellular targeting of two terpene synthases, TPS02 and TPS03**. *Plant physiology* 2010, **153**(3):1293-1310.
180. Falara V, Akhtar TA, Nguyen TT, Spyropoulou EA, Bleeker PM, Schauvinhold I, Matsuba Y, Bonini ME, Schilmiller AL, Last RL *et al*: **The tomato terpene synthase gene family**. *Plant physiology* 2011, **157**(2):770-789.
181. Dornelas MC, Mazzafera P: **A genomic approach to characterization of the *Citrus* terpene synthase gene family**. *Genet Mol Biol* 2007, **30**(3):832-840.
182. Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF: **Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways**. *BMC genomics* 2009, **10**(1):452.
183. Martin DM, Aubourg S, Schouwey MB, Daviet L, Schalk M, Toub O, Lund ST, Bohlmann J: **Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays**. *BMC plant biology* 2010, **10**(1):226.
184. Nieuwenhuizen NJ, Green SA, Chen XY, Bailleul EJD, Matich AJ, Wang MY, Atkinson RG: **Functional Genomics Reveals That a Compact Terpene Synthase Gene Family Can Account for Terpene Volatile Production in Apple**. *Plant physiology* 2013, **161**(2):787-804.
185. Amborella Genome P: **The *Amborella* genome and the evolution of flowering plants**. *Science* 2013, **342**(6165):1241089.
186. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al*: **Phytozome: a comparative platform for green plant genomics**. *Nucleic acids research* 2012, **40**(D1):D1178-D1186.
187. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP *et al*: **The draft genome of sweet orange (*Citrus sinensis*)**. *Nat Genet* 2013, **45**(1):59-66.
188. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D *et al*: **The genome of *Eucalyptus grandis***. *Nature* 2014, **510**(7505):356-362.
189. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al*: **The *Sorghum bicolor* genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
190. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J *et al*: **Genome sequence and analysis of the tuber crop potato**. *Nature* 2011, **475**(7355):189-195.
191. Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB: **A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research**. *Molecular plant-microbe interactions : MPMI* 2012, **25**(12):1523-1530.
192. van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE: **The draft genome and transcriptome of *Cannabis sativa***. *Genome biology* 2011, **12**(10):R102.
193. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic acids research* 2008, **36**(suppl 1):D25-D30.
194. Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M: **Screening synteny blocks in pairwise genome comparisons through integer programming**. *BMC bioinformatics* 2011, **12**:102.

195. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny**. *Bioinformatics* 2004, **20**(18):3643-3646.
196. Zdobnov EM, Apweiler R: **InterProScan – an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**(9):847-848.
197. Dai X, Wang G, Yang DS, Tang Y, Broun P, Marks MD, Sumner LW, Dixon RA, Zhao PX: **TrichOME: a comparative omics database for plant trichomes**. *Plant physiology* 2010, **152**(1):44-54.
198. Marks MD, Wenger JP, Gilding E, Jilk R, Dixon RA: **Transcriptome analysis of *Arabidopsis* wild-type and *gl3-sst sim* trichomes identifies four additional genes required for trichome development**. *Molecular plant* 2009, **2**(4):803-822.
199. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR: **Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis***. *Plant physiology* 2005, **139**(1):5-17.
200. Livak KJ, Schmittgen TD: **Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method**. *methods* 2001, **25**(4):402-408.
201. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space**. *Syst Biol* 2012, **61**(3):539-542.
202. Miller M, Holder M, Vos R, Midford P, Liebowitz T, Chan L, Hoover P, Warnow T: **The CIPRES Portals. CIPRES**. *http://wwwphyloorg/sub_sections/portal* 2009.
203. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees**. *BMC Evol Biol* 2007, **7**(1):214.
204. Rambaut A: **FigTree version 1.3. 1**. *Computer program distributed by the author, website: http://treebioedacuk/software/figtree/* 2009.
205. Xiang T, Shibuya M, Katsube Y, Tsutsumi T, Otsuka M, Zhang H, Masuda K, Ebizuka Y: **A New Triterpene Synthase from *Arabidopsis thaliana* Produces a Tricyclic Triterpene with Two Hydroxyl Groups**. *Org Lett* 2006, **8**(13):2835-2838.
206. Dvorakova L, Cvrckova F, Fischer L: **Analysis of the hybrid proline-rich protein families from seven plant species suggests rapid diversification of their sequences and expression patterns**. *BMC genomics* 2007, **8**:412.
207. Campbell M, Hahn FM, Poulter CD, Leustek T: **Analysis of the isopentenyl diphosphate isomerase gene family from *Arabidopsis thaliana***. *Plant molecular biology* 1998, **36**(2):323-328.
208. Benveniste P: **Sterol metabolism**. *The Arabidopsis book / American Society of Plant Biologists* 2002, **1**:e0004.
209. Caelles C, Ferrer A, Balcells L, Hegardt FG, Boronat A: **Isolation and structural characterization of a cDNA encoding *Arabidopsis thaliana* 3-hydroxy-3-methylglutaryl coenzyme A reductase**. *Plant molecular biology* 1989, **13**(6):627-638.
210. Cordier H, Karst F, Berges T: **Heterologous expression in *Saccharomyces cerevisiae* of an *Arabidopsis thaliana* cDNA encoding mevalonate diphosphate decarboxylase**. *Plant molecular biology* 1999, **39**(5):953-967.
211. Montamat F, Guilloton M, Karst F, Delrot S: **Isolation and characterization of a cDNA encoding *Arabidopsis thaliana* 3-hydroxy-3-methylglutaryl-coenzyme A synthase**. *Gene* 1995, **167**(1):197-201.
212. Riou C, Tourte Y, Lacroute F, Karst F: **Isolation and characterization of a cDNA encoding Arabidopsis thaliana mevalonate kinase by genetic complementation in yeast**. *Gene* 1994, **148**(2):293-297.
213. Ahumada I, Cairó A, Hemmerlin A, González V, Pateraki I, Bach TJ, Rodríguez-Concepción M, Campos N, Boronat A: **Characterisation of the gene family encoding acetoacetyl-CoA thiolase in *Arabidopsis***. *Functional Plant Biology* 2008, **35**(11):1100-1111.

214. Hsieh MH, Goodman HM: **Functional evidence for the involvement of *Arabidopsis* IspF homolog in the nonmevalonate pathway of plastid isoprenoid biosynthesis**. *Planta* 2006, **223**(4):779-784.
215. Rohdich F, Wungsintaweekul J, Eisenreich W, Richter G, Schuhr CA, Hecht S, Zenk MH, Bacher A: **Biosynthesis of terpenoids: 4-diphosphocytidyl-2C-methyl-D-erythritol synthase of *Arabidopsis thaliana***. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(12):6451-6456.
216. Hsieh MH, Chang CY, Hsu SJ, Chen JJ: **Chloroplast localization of methylerythritol 4-phosphate pathway enzymes and regulation of mitochondrial genes in *ispD* and *ispE* albino mutants in *Arabidopsis***. *Plant molecular biology* 2008, **66**(6):663-673.
217. Lange BM, Ghassemian M: **Genome organization in *Arabidopsis thaliana*: a survey for genes involved in isoprenoid and chlorophyll metabolism**. *Plant molecular biology* 2003, **51**(6):925-948.
218. Hsieh M-H, Goodman HM: **The *Arabidopsis IspH* homolog is involved in the plastid nonmevalonate pathway of isoprenoid biosynthesis**. *Plant physiology* 2005, **138**(2):641-653.
219. Rodríguez-Concepción M, Boronat A: **Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics**. *Plant physiology* 2002, **130**(3):1079-1089.
220. Schwender J, Müller C, Zeidler J, Lichtenthaler HK: **Cloning and heterologous expression of a cDNA encoding 1-deoxy-D-xylulose-5-phosphate reductoisomerase of *Arabidopsis thaliana***. *FEBS letters* 1999, **455**(1):140-144.
221. Zhu X, Suzuki K, Saito T, Okada K, Tanaka K, Nakagawa T, Matsuda H, Kawamukai M: **Geranylgeranyl pyrophosphate synthase encoded by the newly isolated gene *GGPS6* from *Arabidopsis thaliana* is localized in mitochondria**. *Plant molecular biology* 1997, **35**(3):331-341.
222. Wang G, Dixon RA: **Heterodimeric geranyl(geranyl)diphosphate synthase from hop (*Humulus lupulus*) and the evolution of monoterpene biosynthesis**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(24):9914-9919.
223. Okada K, Saito T, Nakagawa T, Kawamukai M, Kamiya Y: **Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in *Arabidopsis***. *Plant physiology* 2000, **122**(4):1045-1056.
224. Zhu XF, Suzuki K, Okada K, Tanaka K, Nakagawa T, Kawamukai M, Matsuda K: **Cloning and functional expression of a novel geranylgeranyl pyrophosphate synthase gene from *Arabidopsis thaliana* in *Escherichia coli***. *Plant & cell physiology* 1997, **38**(3):357-361.
225. Bouvier F, Suire C, d'Harlingue A, Backhaus RA, Camara B: **Molecular cloning of geranyl diphosphate synthase and compartmentation of monoterpene synthesis in plant cells**. *The Plant journal : for cell and molecular biology* 2000, **24**(2):241-252.
226. Finkelstein RR, Gampala SS, Rock CD: **Abscisic acid signaling in seeds and seedlings**. *The Plant cell* 2002, **14 Suppl**(suppl 1):S15-45.
227. Cunillera N, Arró M, Forés O, Manzano D, Ferrer A: **Characterization of dehydrodolichyl diphosphate synthase of *Arabidopsis thaliana*, a key enzyme in dolichol biosynthesis**. *FEBS letters* 2000, **477**(3):170-174.
228. Oh K, Hardeman K, Ivanchenko MG, Ellard-Ivey M, Nebenführ A, White T, Lomax TL: **Fine mapping in tomato using microsynteny with the *Arabidopsis* genome: the *Diageotropica* (*Dgt*) locus**. *Genome biology* 2002, **3**(9):049.
229. Delourme D, Lacroute F, Karst F: **Cloning of an *Arabidopsis thaliana* cDNA coding for farnesyl diphosphate synthase by functional complementation in yeast**. *Plant molecular biology* 1994, **26**(6):1867-1873.
230. Dal Bosco C, Lezhneva L, Biehl A, Leister D, Strotmann H, Wanner G, Meurer J: **Inactivation of the chloroplast ATP synthase gamma subunit results in high non-photochemical**

fluorescence quenching and altered nuclear gene expression in *Arabidopsis thaliana*. *The Journal of biological chemistry* 2004, **279**(2):1060-1069.

231. Mann FM, Prisic S, Davenport EK, Determan MK, Coates RM, Peters RJ: **A single residue switch for Mg2+-dependent inhibition characterizes plant class II diterpene cyclases from primary and secondary metabolism**. *Journal of Biological Chemistry* 2010, **285**(27):20558-20563.
232. Ro D-K, Ehlting J, Keeling CI, Lin R, Mattheus N, Bohlmann J: **Microarray expression profiling and functional characterization of *AtTPS* genes: Duplicated *Arabidopsis thaliana* sesquiterpene synthase genes At4g13280 and At4g13300 encode root-specific and wound-inducible (Z)-γ-bisabolene synthases**. *Archives of biochemistry and biophysics* 2006, **448**(1):104-116.
233. Aubourg S, Takvorian A, Chéron A, Kreis M, Lecharny A: **Structure, organization and putative function of the genes identified within a 23.9-kb fragment from *Arabidopsis thaliana* chromosome IV**. *Gene* 1997, **199**(1-2):241-253.
234. Wang Y, Zhang W-Z, Song L-F, Zou J-J, Su Z, Wu W-H: **Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in *Arabidopsis***. *Plant physiology* 2008, **148**(3):1201-1211.
235. R Herrera JB, Bartel B, Wilson WK, Matsuda S: **Cloning and characterization of the *Arabidopsis thaliana* lupeol synthase gene**. *Phytochemistry* 1998, **49**(7):1905-1911.
236. Hanada K, Sawada Y, Kuromori T, Klausnitzer R, Saito K, Toyoda T, Shinozaki K, Li WH, Hirai MY: **Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana***. *Mol Biol Evol* 2011, **28**(1):377-382.
237. Husselstein-Muller T, Schaller H, Benveniste P: **Molecular cloning and expression in yeast of 2,3-oxidosqualene-triterpenoid cyclases from *Arabidopsis thaliana***. *Plant molecular biology* 2001, **45**(1):75-92.
238. Kushiro T, Shibuya M, Ebizuka Y: **Beta-amyrin synthase--cloning of oxidosqualene cyclase that catalyzes the formation of the most popular triterpene among higher plants**. *European journal of biochemistry / FEBS* 1998, **256**(1):238-244.
239. Abel S, Savchenko T, Levy M: **Genome-wide comparative analysis of the *IQD* gene families in *Arabidopsis thaliana* and *Oryza sativa***. *BMC Evol Biol* 2005, **5**:72.
240. Dessimoz C, Gabaldon T, Roos DS, Sonnhammer EL, Herrero J: **Toward community standards in the quest for orthologs**. *Bioinformatics* 2012, **28**(6):900-904.
241. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic acids research* 2001, **29**(1):37-40.
242. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database**. *Nucleic acids research* 2009, **37**:D211-215.
243. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification**. *Bioinformatics* 2014, **30**(9):1236-1240.
244. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic acids research* 2004, **32**:D138-141.
245. Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.
246. Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA: **Gene3D: modelling protein structure, function and evolution**. *Nucleic acids research* 2006, **34**:D281-284.
247. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome research* 2003, **13**(9):2129-2141.

248. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *Journal of Molecular Biology* 2001, **313**(4):903-919.
249. Lipka V, Panstruga R: **Dynamic cellular responses in plant-microbe interactions**. *Current opinion in plant biology* 2005, **8**(6):625-631.
250. Hauser MT: **Molecular basis of natural variation and environmental control of trichome patterning**. *Frontiers in plant science* 2014, **5**:320.
251. Pattanaik S, Patra B, Singh SK, Yuan L: **An overview of the gene regulatory network controlling trichome development in the model plant, *Arabidopsis***. *Frontiers in plant science* 2014, **5**:259.
252. Soltis PS, Soltis DE: **Angiosperm phylogeny: A framework for studies of genome evolution**: Springer; 2013.
253. Guimil S, Dunand C: **Cell growth and differentiation in *Arabidopsis* epidermal cells**. *Journal of experimental botany* 2007, **58**(14):3829-3840.
254. Consortium PO: **The Plant Ontology™ consortium and plant ontologies**. *International Journal of Genomics* 2002, **3**(2):137-142.
255. Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM *et al*: **The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations**. *Nucleic acids research* 2008, **36**:D449-454.
256. Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M *et al*: **Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages**. *Comparative and functional genomics* 2005, **6**(7-8):388-397.
257. Kamolsukyunyong W, Sukhaket W, Ruanjaichon V, Toojinda T, Vanavichit A: **Single-feature polymorphism mapping of isogenic rice lines identifies the influence of terpene synthase on brown planthopper feeding preferences**. *Rice (N Y)* 2013, **6**(1):18.
258. Sallaud C, Giacalone C, Topfer R, Goepfert S, Bakaher N, Rosti S, Tissier A: **Characterization of two genes for the biosynthesis of the labdane diterpene Z-abienol in tobacco (*Nicotiana tabacum*) glandular trichomes**. *The Plant journal : for cell and molecular biology* 2012, **72**(1):1-17.
259. Mendes MD, Barroso JG, Oliveira MM, Trindade H: **Identification and characterization of a second isogene encoding γ-terpinene synthase in *Thymus caespititius***. *J Plant Physiol* 2014.
260. Lawler I, Foley W, Eschler B, Pass D, Handasyde K: **Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials**. *Oecologia* 1998, **116**(1-2):160-169.
261. Trontin C, Tisné S, Bach L, Loudet O: **What does *Arabidopsis* natural variation teach us (and does not teach us) about adaptation in plants?** *Current opinion in plant biology* 2011, **14**(3):225-231.
262. Edger PP, Pires JC: **Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes**. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 2009, **17**(5):699-717.
263. Fang L, Cheng F, Wu J, Wang X: **The Impact of Genome Triplication on Tandem Gene Evolution in *Brassica rapa***. *Frontiers in plant science* 2012, **3**:261.
264. Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, Stacey G, Doerge RW, Jackson SA: **The fate of duplicated genes in a polyploid plant genome**. *The Plant Journal* 2013, **73**(1):143-153.
265. Mühlhausen S, Kollmar M: **Whole genome duplication events in plant evolution reconstructed and predicted using myosin motor proteins**. *BMC evolutionary biology* 2013, **13**(1):202.

266. Chen R, Harada Y, Bamba T, Nakazawa Y, Gyokusen K: **Overexpression of an isopentenyl diphosphate isomerase gene to enhance trans-polyisoprene production in *Eucommia ulmoides* Oliver**. *BMC biotechnology* 2012, **12**(1):78.
267. Berthelot K, Estevez Y, Deffieux A, Peruch F: **Isopentenyl diphosphate isomerase: A checkpoint to isoprenoid biosynthesis**. *Biochimie* 2012, **94**(8):1621-1634.
268. Okada K, Kasahara H, Yamaguchi S, Kawaide H, Kamiya Y, Nojiri H, Yamane H: **Genetic evidence for the role of isopentenyl diphosphate isomerases in the mevalonate pathway and plant development in Arabidopsis**. *Plant & cell physiology* 2008, **49**(4):604-616.
269. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H *et al*: **The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change**. *Nat Genet* 2011, **43**(5):476-481.
270. Estévez JM, Cantero A, Romero C, Kawaide H, Jiménez LF, Kuzuyama T, Seto H, Kamiya Y, León P: **Analysis of the expression of *CLA1*, a gene that encodes the 1-deoxyxylulose 5-phosphate synthase of the 2-C-methyl-D-erythritol-4-phosphate pathway in *Arabidopsis***. *Plant physiology* 2000, **124**(1):95-104.
271. Jones JD, Dangl JL: **The plant immune system**. *Nature* 2006, **444**(7117):323-329.
272. Chinchilla D, Bauer Z, Regenass M, Boller T, Felix G: **The *Arabidopsis* receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception**. *The Plant cell* 2006, **18**(2):465-476.
273. Zipfel C, Robatzek S: **Pathogen-associated molecular pattern-triggered immunity: veni, vidi…?** *Plant physiology* 2010, **154**(2):551-554.
274. Collins NC, Thordal-Christensen H, Lipka V, Bau S, Kombrink E, Qiu JL, Huckelhoven R, Stein M, Freialdenhoven A, Somerville SC *et al*: **SNARE-protein-mediated disease resistance at the plant cell wall**. *Nature* 2003, **425**(6961):973-977.
275. Schwessinger B, Zipfel C: **News from the frontline: recent insights into PAMP-triggered immunity in plants**. *Current opinion in plant biology* 2008, **11**(4):389-395.
276. Hann DR, Dominguez-Ferreras A, Motyka V, Dobrev PI, Schornack S, Jehle A, Felix G, Chinchilla D, Rathjen JP, Boller T: **The *Pseudomonas* type III effector HopQ1 activates cytokinin signaling and interferes with plant innate immunity**. *The New phytologist* 2013:n/a-n/a.
277. Van der Biezen EA, Jones JD: **Plant disease-resistance proteins and the gene-for-gene concept**. *Trends in biochemical sciences* 1998, **23**(12):454-456.
278. Vleeshouwers VG, Raffaele S, Vossen JH, Champouret N, Oliva R, Segretin ME, Rietman H, Cano LM, Lokossou A, Kessel G *et al*: **Understanding and exploiting late blight resistance in the age of effectors**. *Annual review of phytopathology* 2011, **49**:507-531.
279. Ballvora A, Ercolano MR, Weiss J, Meksem K, Bormann CA, Oberhagemann P, Salamini F, Gebhardt C: **The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/*NBS*/*LRR* class of plant resistance genes**. *The Plant journal : for cell and molecular biology* 2002, **30**(3):361-371.
280. Young ND: **The genetic architecture of resistance**. *Current opinion in plant biology* 2000, **3**(4):285-290.
281. Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB, Jones JD: **Three genes of the *Arabidopsis RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants**. *The Plant cell* 1998, **10**(11):1847-1860.
282. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW: **Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis***. *The Plant cell* 2003, **15**(4):809-834.
283. Maekawa T, Kufer TA, Schulze-Lefert P: **NLR functions in plant and animal immune systems: so far and yet so close**. *Nature immunology* 2011, **12**(9):817-826.
284. Boisson B, Giglione C, Meinnel T: **Unexpected protein families including cell defense components feature in the N-myristoylome of a higher eukaryote**. *The Journal of biological chemistry* 2003, **278**(44):43418-43429.

285. Warren RF, Henk A, Mowery P, Holub E, Innes RW: **A mutation within the leucine-rich repeat domain of the *Arabidopsis* disease resistance gene RPS5 partially suppresses multiple bacterial and downy mildew resistance genes**. *The Plant cell* 1998, **10**(9):1439-1452.
286. Deslandes L, Olivier J, Peeters N, Feng DX, Khounlotham M, Boucher C, Somssich I, Genin S, Marco Y: **Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(13):8024-8029.
287. Narusaka M, Shirasu K, Noutoshi Y, Kubo Y, Shiraishi T, Iwabuchi M, Narusaka Y: ***RRS1* and *RPS4* provide a dual Resistance-gene system against fungal and bacterial pathogens**. *The Plant journal : for cell and molecular biology* 2009, **60**(2):218-226.
288. Bernoux M, Timmers T, Jauneau A, Brière C, de Wit PJ, Marco Y, Deslandes L: **RD19, an *Arabidopsis* cysteine protease required for RRS1-R–mediated resistance, is relocalized to the nucleus by the *Ralstonia solanacearum* PopP2 effector**. *The Plant Cell Online* 2008, **20**(8):2252-2264.
289. Yang H, Shi Y, Liu J, Guo L, Zhang X, Yang S: **A mutant CHS3 protein with TIR-NB-LRR-LIM domains modulates growth, cell death and freezing tolerance in a temperature-dependent manner in *Arabidopsis***. *The Plant journal : for cell and molecular biology* 2010, **63**(2):283-296.
290. Kato H, Shida T, Komeda Y, Saito T, Kato A: **Overexpression of the *Activated Disease Resistance 1-like1* (*ADR1-L1*) Gene Results in a Dwarf Phenotype and Activation of Defense-Related Gene Expression in *Arabidopsis thaliana***. *J Plant Biol* 2011, **54**(3):172-179.
291. Xiao S, Charoenwattana P, Holcombe L, Turner JG: **The *Arabidopsis* genes *RPW8.1* and *RPW8.2* confer induced resistance to powdery mildew diseases in tobacco**. *Molecular plant-microbe interactions : MPMI* 2003, **16**(4):289-294.
292. Xiao S, Calis O, Patrick E, Zhang G, Charoenwattana P, Muskett P, Parker JE, Turner JG: **The atypical resistance gene, *RPW8*, recruits components of basal defence for powdery mildew resistance in *Arabidopsis***. *The Plant journal : for cell and molecular biology* 2005, **42**(1):95-110.
293. Dangl JL, Jones JD: **Plant pathogens and integrated defence responses to infection**. *Nature* 2001, **411**(6839):826-833.
294. van Ooijen G, Mayr G, Kasiem MM, Albrecht M, Cornelissen BJ, Takken FL: **Structure-function analysis of the NB-ARC domain of plant disease resistance proteins**. *Journal of experimental botany* 2008, **59**(6):1383-1397.
295. Boller T, Felix G: **A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors**. *Annual review of plant biology* 2009, **60**(1):379-406.
296. Tameling WI, Vossen JH, Albrecht M, Lengauer T, Berden JA, Haring MA, Cornelissen BJ, Takken FL: **Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation**. *Plant physiology* 2006, **140**(4):1233-1245.
297. Takken FLW, Goverse A: **How to build a pathogen detector: structural basis of NB-LRR function**. *Current opinion in plant biology* 2012, **15**(4):375-384.
298. Bhattacharyya MK: ***RPSk-1* gene family, nucleotide sequences and uses thereof**. In.: Google Patents; 2007.
299. Cannon SB, Mitra A, Baumgarten A, Young ND, May G: **The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana***. *BMC plant biology* 2004, **4**(1):10.
300. Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS: **Patterns of positive selection in the complete *NBS-LRR* gene family of *Arabidopsis thaliana***. *Genome research* 2002, **12**(9):1305-1315.

301. Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Moore MJ *et al*: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III**. *Bot J Linn Soc* 2009, **161**(2):105-121.
302. Schnable JC, Springer NM, Freeling M: **Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(10):4069-4074.
303. Paterson AH, Freeling M, Tang H, Wang X: **Insights from the comparison of plant genome sequences**. *Annual review of plant biology* 2010, **61**(1):349-372.
304. Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C *et al*: **The Reference Genome of the Halophytic Plant *Eutrema salsugineum***. *Frontiers in plant science* 2013, **4**:46.
305. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.
306. Offord V, Werling D: **LRRfinder2.0: a webserver for the prediction of leucine-rich repeats**. *Innate immunity* 2013, **19**(4):398-402.
307. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments**. *Nucleic acids research* 2006, **34**:W609-612.
308. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J: **KaKs_Calculator: calculating Ka and Ks through model selection and model averaging**. *Genomics, proteomics & bioinformatics* 2006, **4**(4):259-263.
309. Richly E, Kurth J, Leister D: **Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution**. *Mol Biol Evol* 2002, **19**(1):76-84.
310. Yang S, Zhang X, Yue JX, Tian D, Chen JQ: **Recent duplications dominate NBS-encoding gene expansion in two woody species**. *Molecular genetics and genomics : MGG* 2008, **280**(3):187-198.
311. Chen Q, Han Z, Jiang H, Tian D, Yang S: **Strong positive selection drives rapid diversification of R-genes in *Arabidopsis* relatives**. *Journal of molecular evolution* 2010, **70**(2):137-148.
312. Mun JH, Yu HJ, Park S, Park BS: **Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa***. *Molecular genetics and genomics : MGG* 2009, **282**(6):617-631.
313. Yu J, Tehrim S, Zhang F, Tong C, Huang J, Cheng X, Dong C, Zhou Y, Qin R, Hua W *et al*: **Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana***. *BMC genomics* 2014, **15**:3.
314. Salamov AA, Solovyev VV: ***Ab initio* gene finding in Drosophila genomic DNA**. *Genome research* 2000, **10**(4):516-522.
315. Porter BW, Paidi M, Ming R, Alam M, Nishijima WT, Zhu YJ: **Genome-wide analysis of *Carica papaya* reveals a small *NBS* resistance gene family**. *Molecular genetics and genomics : MGG* 2009, **281**(6):609-626.
316. Hermoso A, Vlasova A, Sanseverino W, D'Alessandro R, Andolfo G, Frusciante L, Roma G, Ercolano M, Lowy E: **The Plant Resistance Gene Database (PRGdb): a Wiki-based system for the annotation of R-genes**. *IWBBIO Proceedings* 2009.
317. Jupe F, Pritchard L, Etherington GJ, Mackenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD *et al*: **Identification and localisation of the *NB-LRR* gene family within the potato genome**. *BMC genomics* 2012, **13**:75.
318. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic acids research* 2009, **37**:W202-208.
319. Jupe F, Witek K, Verweij W, Sliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ *et al*: **Resistance gene enrichment sequencing (RenSeq) enables reannotation of the *NB-LRR* gene family from sequenced plant genomes and rapid mapping of resistance**

loci in segregating populations. *The Plant journal : for cell and molecular biology* 2013, **76**(3):530-544.
320. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ *et al*: **Direct selection of human genomic loci by microarray hybridization**. *Nature methods* 2007, **4**(11):903-905.
321. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ *et al*: **Genome-wide *in situ* exon capture for selective resequencing**. *Nat Genet* 2007, **39**(12):1522-1527.
322. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J: **Targeted enrichment strategies for next-generation plant biology**. *American journal of botany* 2012, **99**(2):291-311.
323. Andolfo G, Sanseverino W, Rombauts S, Van de Peer Y, Bradeen JM, Carputo D, Frusciante L, Ercolano MR: **Overview of tomato (*Solanum lycopersicum*) candidate pathogen recognition genes reveals important *Solanum* R locus dynamics**. *The New phytologist* 2013, **197**(1):223-237.
324. Andolfo G, Jupe F, Witek K, Etherington GJ, Ercolano MR, Jones JD: **Defining the full tomato *NB-LRR* resistance gene repertoire using genomic and cDNA RenSeq**. *BMC plant biology* 2014, **14**:120.
325. Hurst LD: **The Ka/Ks ratio: diagnosing the form of sequence evolution**. *Trends in genetics : TIG* 2002, **18**(9):486.
326. Tan X, Meyers BC, Kozik A, West MA, Morgante M, St Clair DA, Bent AF, Michelmore RW: **Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in *Arabidopsis***. *BMC plant biology* 2007, **7**:56.
327. Navarro L, Zipfel C, Rowland O, Keller I, Robatzek S, Boller T, Jones JD: **The transcriptional innate immune response to flg22. Interplay and overlap with *Avr* gene-dependent defense responses and bacterial pathogenesis**. *Plant physiology* 2004, **135**(2):1113-1128.
328. Lewis JD, Wu R, Guttman DS, Desveaux D: **Allele-specific virulence attenuation of the *Pseudomonas syringae* HopZ1a type III effector via the *Arabidopsis* ZAR1 resistance protein**. *PLoS genetics* 2010, **6**(4):e1000894.
329. Bonardi V, Cherkis K, Nishimura MT, Dangl JL: **A new eye on NLR proteins: focused on clarity or diffused by complexity?** *Current opinion in immunology* 2012, **24**(1):41-50.
330. Wang W, Zhang Y, Wen Y, Berkey R, Ma X, Pan Z, Bendigeri D, King H, Zhang Q, Xiao S: **A comprehensive mutational analysis of the *Arabidopsis* resistance protein RPW8.2 reveals key amino acids for defense activation and protein targeting**. *The Plant cell* 2013, **25**(10):4242-4261.
331. Bonardi V, Tang S, Stallmann A, Roberts M, Cherkis K, Dangl JL: **Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(39):16463-16468.
332. Uitdewilligen JG, Wolters AM, D'Hoop B B, Borm TJ, Visser RG, van Eck HJ: **A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato**. *PloS one* 2013, **8**(5):e62355.
333. Michelmore RW, Meyers BC: **Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process**. *Genome research* 1998, **8**(11):1113-1130.
334. Ratnaparkhe MB, Wang X, Li J, Compton RO, Rainville LK, Lemke C, Kim C, Tang H, Paterson AH: **Comparative analysis of peanut *NBS-LRR* gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny**. *The New phytologist* 2011, **192**(1):164-178.
335. Roberts M, Tang S, Stallmann A, Dangl JL, Bonardi V: **Genetic requirements for signaling from an autoactive plant NB-LRR intracellular innate immune receptor**. *PLoS genetics* 2013, **9**(4):e1003465.

336. Collier SM, Hamel LP, Moffett P: **Cell death mediated by the N-terminal domains of a unique and highly conserved class of NB-LRR protein**. *Molecular plant-microbe interactions : MPMI* 2011, **24**(8):918-931.
337. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L: **Chromosome evolution in eukaryotes: a multi-kingdom perspective**. *Trends in genetics : TIG* 2005, **21**(12):673-682.
338. Chapman BA, Bowers JE, Feltus FA, Paterson AH: **Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(8):2730-2735.
339. Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y: **Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana***. *Genome biology* 2006, **7**(2):R13.
340. Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Kohler N: **Discovery of tetraploidy in a mammal**. *Nature* 1999, **401**(6751):341.
341. Comber SC, Smith C: **Polyploidy in fishes: patterns and processes**. *Biological Journal of the Linnean Society* 2004, **82**(4):431-442.
342. Ptacek MB, Gerhardt HC, Sage RD: **Speciation by Polyploidy in Treefrogs: Multiple Origins of the Tetraploid, *Hyla versicolor***. *Evolution* 1994, **48**(3):898.
343. Wendel JF: **Genome evolution in polyploids**. In: *Plant Molecular Evolution.* Springer; 2000: 225-249.
344. Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V *et al*: **Understanding mechanisms of novel gene expression in polyploids**. *Trends in genetics : TIG* 2003, **19**(3):141-147.
345. Tate JA, Soltis DE, Soltis PS: **Polyploidy in plants**. *The evolution of the genome* 2005:371-426.
346. Taji T, Sakurai T, Mochida K, Ishiwata A, Kurotani A, Totoki Y, Toyoda A, Sakaki Y, Seki M, Ono H *et al*: **Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, *Thellungiella halophila***. *BMC plant biology* 2008, **8**(1):115.
347. Van de Peer Y, Maere S, Meyer A: **The evolutionary significance of ancient genome duplications**. *Nature reviews Genetics* 2009, **10**(10):725-732.
348. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE: **Rosid radiation and the rapid rise of angiosperm-dominated forests**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(10):3853-3858.
349. Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BB, Jones JD: **Novel Disease Resistance Specificities Result from Sequence Exchange between Tandemly Repeated Genes at the *Cf-4/9* Locus of Tomato**. *Cell* 1997, **91**(6):821-832.
350. Wang Y, Tan X, Paterson AH: **Different patterns of gene structure divergence following gene duplication in *Arabidopsis***. *BMC genomics* 2013, **14**(1):652.
351. Shiu SH, Bleecker AB: **Plant receptor-like kinase gene family: diversity, function, and signaling**. *Science's STKE : signal transduction knowledge environment* 2001, **2001**(113):re22.
352. Shiu SH, Bleecker AB: **Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10763-10768.
353. Shiu SH, Bleecker AB: **Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in *Arabidopsis***. *Plant physiology* 2003, **132**(2):530-543.
354. Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH: **Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice**. *The Plant cell* 2004, **16**(5):1220-1234.
355. Bouwmeester K, Govers F: ***Arabidopsis* L-type lectin receptor kinases: phylogeny, classification, and expression profiles**. *Journal of experimental botany* 2009, **60**(15):4383-4396.

356. Singh P, Zimmerli L: **Lectin receptor kinases in plant innate immunity**. *Frontiers in plant science* 2013, **4**:124.
357. Lannoo N, Van Damme EJ: **Lectin domains at the frontiers of plant defense**. *Frontiers in plant science* 2014, **5**.
358. Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME: **Self-incompatibility in the genus *Arabidopsis*: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana***. *The Plant cell* 2001, **13**(3):627-643.
359. Gilardoni PA, Hettenhausen C, Baldwin IT, Bonaventure G: ***Nicotiana attenuata* LECTIN RECEPTOR KINASE1 suppresses the insect-mediated inhibition of induced defense responses during *Manduca sexta* herbivory**. *The Plant cell* 2011, **23**(9):3512-3532.
360. Geijtenbeek TB, Gringhuis SI: **Signalling through C-type lectin receptors: shaping immune responses**. *Nature Reviews Immunology* 2009, **9**(7):465-479.
361. Vaid N, Macovei A, Tuteja N: **Knights in action: lectin receptor-like kinases in plant development and stress responses**. *Molecular plant* 2013, **6**(5):1405-1418.
362. Phillips SM, Dubery IA, van Heerden H: **Identification and molecular characterisation of a lectin receptor-like kinase (*GhLecRK-2*) from cotton**. *Plant Molecular Biology Reporter* 2013, **31**(1):9-20.
363. Liu Y, Wu H, Chen H, Liu Y, He J, Kang H, Sun Z, Pan G, Wang Q, Hu J *et al*: **A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice**. *Nat Biotechnol* 2014.
364. Bouwmeester K, de Sain M, Weide R, Gouget A, Klamer S, Canut H, Govers F: **The lectin receptor kinase LecRK-I.9 is a novel *Phytophthora* resistance component and a potential host target for a RXLR effector**. *PLoS pathogens* 2011, **7**(3):e1001327.
365. Desclos-Theveniau M, Arnaud D, Huang TY, Lin GJ, Chen WY, Lin YC, Zimmerli L: **The *Arabidopsis* lectin receptor kinase LecRK-V.5 represses stomatal immunity induced by *Pseudomonas syringae* pv. tomato DC3000**. *PLoS pathogens* 2012, **8**(2):e1002513.
366. Singh P, Kuo YC, Mishra S, Tsai CH, Chien CC, Chen CW, Desclos-Theveniau M, Chu PW, Schulze B, Chinchilla D *et al*: **The lectin receptor kinase-VI.2 is required for priming and positively regulates *Arabidopsis* pattern-triggered immunity**. *The Plant cell* 2012, **24**(3):1256-1270.
367. Wang Y, Govers F, Bouwmeester K: **Lectin receptor kinases: sentinels in defense against plant pathogens**. In: *Book of Abstracts COST 2014-2nd Annual Conference of the SUSTAIN Action: 2014*; 2014: 25.
368. Gouhier-Darimont C, Schmiesing A, Bonnet C, Lassueur S, Reymond P: **Signalling of *Arabidopsis thaliana* response to Pieris brassicae eggs shares similarities with PAMP-triggered immunity**. *Journal of experimental botany* 2013, **64**(2):665-674.
369. Wan J, Patel A, Mathieu M, Kim SY, Xu D, Stacey G: **A lectin receptor-like kinase is required for pollen development in Arabidopsis**. *Plant molecular biology* 2008, **67**(5):469-482.
370. Deng K, Wang Q, Zeng J, Guo X, Zhao X, Tang D, Liu X: **A lectin receptor kinase positively regulates ABA response during seed germination and is involved in salt and osmotic stress response**. *J Plant Biol* 2009, **52**(6):493-500.
371. Xin Z, Wang A, Yang G, Gao P, Zheng ZL: **The *Arabidopsis* A4 subfamily of lectin receptor kinases negatively regulates abscisic acid response in seed germination**. *Plant physiology* 2009, **149**(1):434-444.
372. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN *et al*: **The genome of *Theobroma cacao***. *Nat Genet* 2011, **43**(2):101-108.
373. Camacho C, Madden T, Ma N, Tao T, Agarwala R, Morgulis A: **BLAST Command Line Applications User Manual**. 2013.
374. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource**. *Nucleic acids research* 2012, **40**:D302-305.

375. Maddison W, Maddison D: **Mesquite: a modular system for evolutionary analysis. 2010. Version 2.74**. *Available at: mesquiteproject.org/mesquite/download/download html* 2010.
376. Loytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(30):10557-10562.
377. Loytynoja A, Goldman N: **webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser**. *BMC bioinformatics* 2010, **11**(1):579.
378. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.
379. Rambaut A, Drummond A: **Tracer v1. 5 Andrew Rambaut**. In.; 2009.
380. Armijo G, Salinas P, Monteoliva MI, Seguel A, Garcia C, Villarroel-Candia E, Song W, van der Krol AR, Alvarez ME, Holuigue L: **A salicylic acid-induced lectin-like protein plays a positive role in the effector-triggered immunity response of *Arabidopsis thaliana* to *Pseudomonas syringae* Avr-Rpm1**. *Molecular plant-microbe interactions : MPMI* 2013, **26**(12):1395-1406.
381. Schuermann D, Molinier J, Fritsch O, Hohn B: **The dual nature of homologous recombination in plants**. *Trends in genetics : TIG* 2005, **21**(3):172-181.

CURRICULUM VITAE

JOHANNES A. HOFBERGER *MSc, PhD*

Specialities:
Big Data
Bioinformatics
Life Sciences

Personal Data:
Date of birth: September 10th, 1984
Place of birth: Dachau (Germany)
Nationality: German

Current Address:
Freiligrathstr. 4
60385 Frankfurt am Main
Germany

Contact:
johannes.hofberger@gmail.com
www.linkedin.com/in/hofberger
+49 151 7546 1519

EDUCATION

12/2010 – 11/2014	**PhD**; Majors: Bioinformatics, Big Data analysis; Wageningen University & Research Centre / University of Amsterdam (*Amsterdam and Wageningen, The Netherlands*)
09/2004 – 09/2010	**MSc** (with honors); Major: Life Sciences; Ludwig-Maximilians-University Munich (*Munich, Germany*)
09/1995 – 07/2004	**A-level**; Majors: Biology, English, Chemistry, History; Ignaz-Taschner-Gymnasium Dachau (*Dachau, Germany*)

EXPERIENCE (MANAGEMENT)

02/2014 – 03/2014	**Project Management**; Main coordinator for international research collaboration; Beijing Genome Institute (华大基因) (*Shenzhen, PR China*)
11/2012 – 12/2014	**Management Consulting**; Intern with Life Sciences practice; Accenture Management Consulting (*Frankfurt am Main, Germany*)
02/2012 – 04/2012	**Head** of Dutch research delegation and management of a local project team; Shanghai Institute for Biological Sciences (上海生科院) (*Shanghai, PR China*)
12/2010 – 07/2012	**Supervision** of MSc students and planning of experiments; University of Amsterdam (*Amsterdam, The Netherlands*)
04/2006 – 06/2009	**Head** and independent organizer of tutorials for MSc students; Ludwig-Maximilians-University Munich (*Munich, Germany*)

EXPERIENCE (RESEARCH & DEVELOPMENT)

04/2014 – 09/2014	**Development** of informatics tools for Big Data analysis; Visiting scientist at Partner Institute of Max-Planck-Society and Chinese Academy of Sciences (中国科学院) (*Shanghai, PR China*)
08/2012 – 01/2014	**Research** and data analysis in Life Sciences; PhD student at Wageningen University & Research Centre (*Wageningen, The Netherlands*)
12/2009 – 09/2010	**Research** and data analysis in Life Sciences; MSc student at The Sainsbury Laboratory (*Norwich, United Kingdom*)
07/2009 – 08/2009	**Research** and data analysis in Life Sciences; Intern at Max-Planck-Institute for Biochemistry (*Munich, Germany*)
02/2007 – 07/2007	**Research** and data analysis in Life Sciences; Intern at Shanghai Normal University (上海师范大学) (*Shanghai, PR China*)
09/2006 – 02/2007	**Development** of diagnostics tools; Intern and laboratory assistant at biotech start-up Agrobiogen (*Laretshausen, Germany*)
04/2006 – 06/2009	**Research** and data analysis in Life Sciences; Laboratory assistant at Ludwig-Maximilians-University Munich (*Munich, Germany*)

CONFERENCES & WORKSHOPS (MANAGEMENT)

04/07/2013	**Marketing & sales**, workshop "Roche Round Tables", Roche Diagnostics (*Mannheim, Germany*)
02/07/2013	**Technology**, panel discussion "Technology Vision 2013", Accenture (*Kronberg im Taunus, Germany)*
14/06/2013	**Technology**, international conference "Big Techday 6 – How Big Data will Change the World", TNG Technology Consulting (*Munich, Germany*)
13/06/2013	**Management Consulting**, workshop "Chances@KPMG", KPMG (*Berlin, Germany*)
24/05/2013	**Operations**, workshop "Lean Six Sigma", Accenture (*Kronberg im Taunus, Germany*)
12/04/2013	**Management Consulting**, workshop "Campus for Strategy & Leadership 2013", WHU – Otto Beisheim School of Management (*Düsseldorf, Germany*)
25/01/2013	**Management Consulting**, workshop "Medical Biotechnology in Germany", The Boston Consulting Group (*Berlin, Germany*)

CONFERENCES & WORKSHOPS (RESEARCH & DEVELOPMENT)

26/05/2014	**Invited speaker**, workshop "Integrative Genomics", Institute for Genetics and Developmental Biology (中科院遗传发育所) (*Beijing, PR China*)
15/01/2014	**Invited speaker**, international conference "Plant Animal Genome XXIII" (*San Diego, USA*)
20/07/2013	**Invited speaker**, international conference "5th European EPS Retreat" (*Gent, Belgium*)
14/03/2012	**Invited speaker**, symposium "Genomics of Polyploidy" (*Shanghai, PR China*)
16/07/2011	**Invited speaker**, international conference "3rd European EPS Retreat" (*Paris, France*)
14/07/2010	**Invited speaker**, workshop "Plant Innate Immunity", Max-Planck-Institute for Plant Breeding Research (*Cologne, Germany*)

COMMITMENTS & SCHOLARSHIPS

11/2012 – to date	Careerloft and e-fellows.net[#] online-scholarships
12/2010 – 12/2012	Dutch Organization for Scientific Research (NWO) (**€ 160.000**)
12/2010 – to date	Membership in European Plant Science Organization (EPSO)
12/2009 – 09/2010	Biotechnology and Biological Sciences Research Council (BBSRC) (**£ 8.000**)
02/2007 – 07/2007	German Academic Exchange Service (DAAD) (**€ 6.000**)
04/2006 – 02/2007	Independent organization of seminars on bioethics for students at Ludwig-Maximilians-University Munich (*Munich, Germany*)

LANGUAGES

German (native)	Dutch (fluent)
English (fluent)	Mandarin (basics)

IT-SKILLS

Python; Perl; Batch-scripting
MS Excel / MS PowerPoint / MS Word
MS Project, MS Visio, Think-Cell

OTHER INTERESTS

#Interview in German language: http://bit.ly/hofberger

Social psychology (i.e. Tracom Social Styles, Power Mapping); *(Social) network analytics; Travelling through Southeast Asia (Tibet, Nepal, Kashmir, Ladakh);* Chess

[#]Interview in German language: http://bit.ly/hofberger

The research described in this thesis was financially supported by The Dutch Financer. Financial support from the University of Amsterdam, Wageningen University & Research Centre, the Chinese Academy of Sciences and the Netherlands Organization for Scientific Research is greatly appreciated.