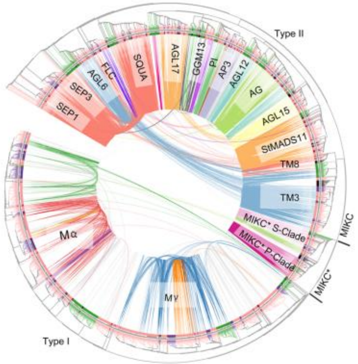
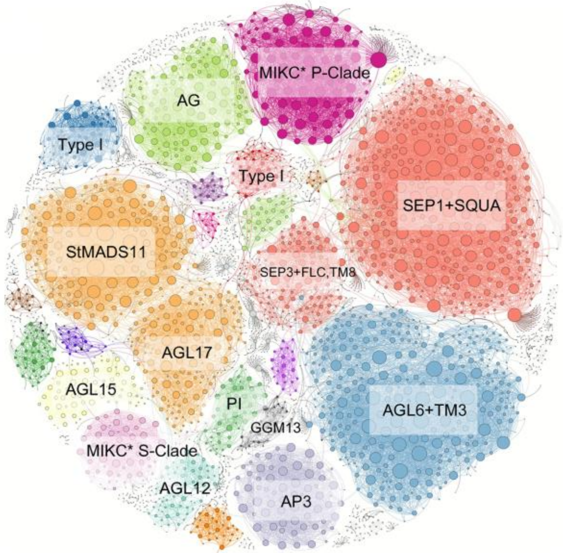


Synteny-Based Phylogenomic Networks for Comparative Genomics

T. Zhao

2018

Synteny-Based Phylogenomic Networks for Comparative Genomics



Tao Zhao

Propositions

1. Networks provide a powerful way to organize, present, and analyze comparative synteny data of plant genomes that have undergone recurrent polyploidy events.
   (this thesis)
2. Family-specific synteny clusters generated by gene transpositions may reveal genes contributing to phenotypic trait evolution.
   (this thesis)
3. A PhD program without a set research proposal may end up with better results.
4. Taking a step back and seeing the big picture can break tangled details, and help both in daily life and in research.
5. The moment you feel helpless might also be the opportunity you find your way and grow stronger.
6. Ones biggest contribution to society is to be happy in ones own life.

Propositions belonging to the thesis, entitled

Synteny-Based Phylogenomic Networks for Comparative Genomics

Tao Zhao

Wageningen, 17 September 2018

Synteny-Based Phylogenomic Networks for Comparative Genomics

Tao Zhao

Thesis committee

Promotor

Prof. Dr M. Eric Schranz

Professor of Biosystematics

Wageningen University & Research

Other members

Prof. Dr V. van Noort, Leiden University, the Netherlands

Dr B. Gravendeel, Leiden University, the Netherlands

Prof. Dr B. Snel, Utrecht University, the Netherlands

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

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

Synteny-Based Phylogenomic Networks for Comparative Genomics

Tao Zhao

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Monday 17 September 2018

at 11 a.m. in the Aula.

Tao Zhao

Synteny-Based Phylogenomic Networks for Comparative Genomics,

140 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)

With references, and summary in English

ISBN: 978-94-6343-320-4

DOI: https://doi.org/10.18174/456296

Contents

Chapter 1…………………………………………………………………………… **7**

General Introduction

Chapter 2…………………………………………………………………………… **19**

Network approaches for plant phylogenomic synteny analysis

Chapter 3…………………………………………………………………………… **29**

Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation

Chapter 4…………………………………………………………………………… **57**

Beyond the sequence: functional innovations in plant LATE EMBRYOGENESIS ABUNDANT proteins revealed by synteny network analysis

Chapter 5…………………………………………………………………………… **81**

Phylogenomic synteny networks reveal gene positional conservation and diversification across 170 MYR of mammal and angiosperm evolution

Chapter 6……………………………………………………………………...... **107**

General Discussion

References………………………………………………………………………… **113**

Summary…………………………………………………………………………… **133**

Acknowledgement...……………………………………………………………… **135**

Curriculum Vitae………………………………………………………………….. **137**

Education Statement…………………………………………………………….. **138**

Chapter 1

General Introduction

The code of life and genome sequencing

Life first appeared on earth ~ 4 billion years ago (Dodd et al., 2017) as vast colonies of single-celled bacteria that would ultimately gave rise to all other living creatures (Figure 1). Over the ages, our planet has witnessed explosions of life and massive extinctions. DNA (Deoxyribonucleic acid), the molecular blueprint (or recipe) for all known extant living organisms, except some viruses which use RNA (Johnston et al., 2001), acts as the raw material that has been molded and evolved into the many forms of life we observe today (Figure 1).

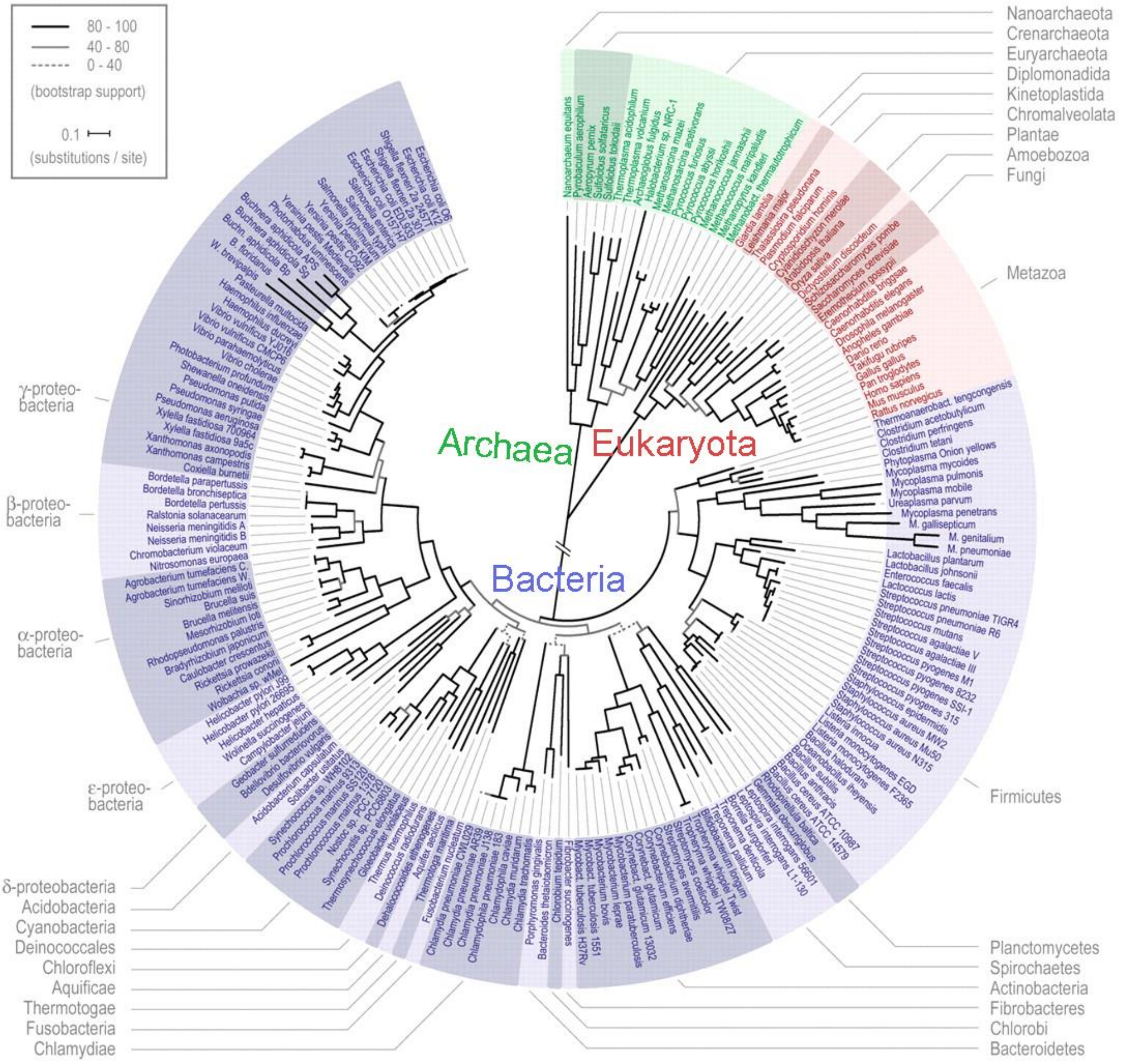


Figure 1 Phylogeny of the tree of life. Figure adapted and modified from Ciccarelli et al., 2006.

A genome contains complete set of DNA of an organism, including all its genes. A genome sequence is the complete list of the nucleotides (A, C, G, and T for DNA genomes) that make up all the chromosomes of an individual or a species. The human genome project, took an effort of 13 years, 3 billion dollars, and more than 200 scientists, to sequence and assemble the roughly 3 billion base pairs (3 Gb) of human DNA that comprise our genetic material (Consortium, 2001; Venter et al., 2001). Over the past four decades, DNA sequencing technology has developed tremendously (Shendure et al., 2017). High throughput “Next-Generation” sequencers can now read thousands of human genomes and a myriad of other genomes at high speeds with continuously improving accuracy and contiguity. DNA sequencing is now having crucial applications from molecular biology to metagenomics, from paleontology to clinical medicine and comparative evolutionary genomics.

Comparative genomics and synteny

The wealth of genomic data available nowadays is making comparative genomics an accessible focal point for the study of any form of life. Comparative genomics provides a powerful tool for studying evolutionary genomic features among organisms. Such genomic features may include changes/differences in DNA sequence, genes, gene order, regulatory sequences, and other genomic structural landmarks (Xia, 2013). These help to identify genes that are conserved or common among species, as well as genes that give each organism its unique characteristics.

Amid the aforementioned applications of comparative genomics, synteny reflects important relationships between the genomic context of genes both in terms of function and regulation and is often used as a proxy for the constraint and/or conservation of gene function (Dandekar et al., 1998; Dewey, 2011; Lv et al., 2011). Synteny was originally defined as pairs or sets of genes located on homologous chromosomes in two or more species, but not necessarily in the same order (Renwick, 1971; Passarge et al., 1999). However, the current widespread usage of the term synteny, which we adopt in this thesis, implies conserved collinearity and genomic context.

The analysis of synteny in the gene order sense has several applications in genomics. Shared synteny is one of the most reliable criteria for establishing the orthology of genomic regions in different species (Dewey, 2011; Altenhoff et al., 2016). Additionally, exceptional conservation of synteny can reflect important functional relationships between genes. For example, the order of genes in the "Hox cluster", a conserved cluster of homeobox domain transcription factors (Duboule, 2007), which are key determinants of the animal body plan and which interact with each other in critical ways, is generally preserved throughout the animal kingdom (Amores et al., 1998). However, in Octopus (*Octopus bimaculoides*) Hox genes are not organized into clusters as in most other bilaterian genomes, but are completely scattered (Albertin et al., 2015) (Figure 2).

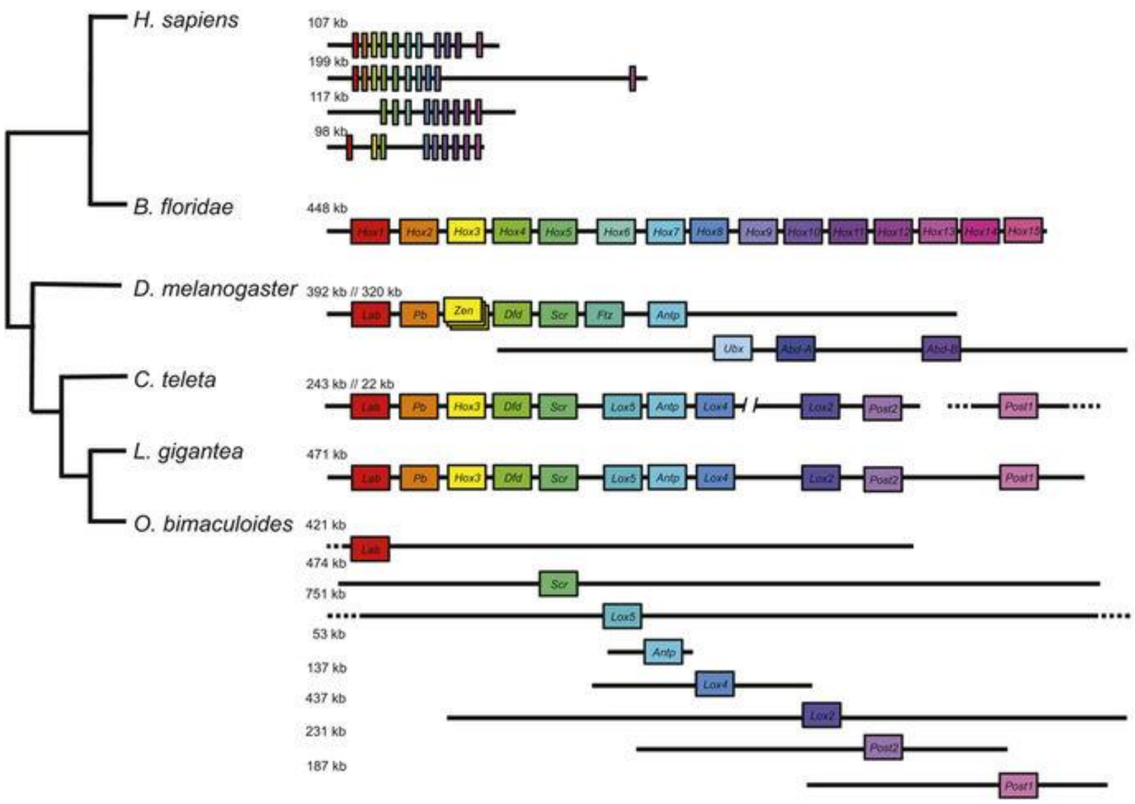


Figure 2. Synteny conservation of Hox clusters across five species (*Homo sapiens*, *Branchiostoma floridae*, *Drosophila melanogaster*, *Capitella teleta*, and *Lottia gigantea*), and its fragmentation in Octupus (*O. bimaculoide*s). Figure adapted from Albertin et al., 2015.

Synteny complements phylogeny

Phylogeny is the science of estimating evolutionary past, based on the comparisons of DNA or protein sequences (Baldauf, 2003). Phylogenetic analyses are important for understanding biodiversity, evolution, ecology, and genomes (Yang and Rannala, 2012). Phylogeny reconstruction has developed a set of tools and methods based on optimality criteria such as parsimony, likelihood and posterior probability in order to find best possible trees (Lewis, 2001). However character-based phylogeny could detect orthologous relations obscured by amino acid bias or DNA substitution rate variability among taxa and genes. In addition, different gene loss from the genome could also affect the reconstructed phylogeny (Figure 3). In such cases, additional evidence is then needed for poorly supported branches.

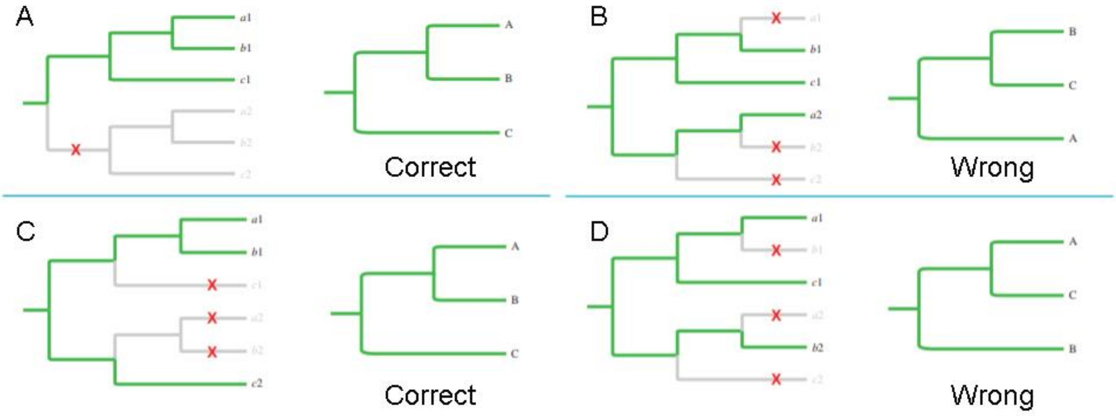


Figure 3. Different patterns of gene loss affect inferred phylogeny from single-copy genes. On the right panel are scenarios when constructed phylogeny will not truly reflect relationship among species A, B and C. Figure adapted and modified from Jiao and Paterson, 2014.

Synteny reveals local genomic gene arrangements, and presents a clearer picture of gene and gene family evolution, allowing one to decide which gene locations are ancestral. For example, the parallel coordinate synteny plot below depict relationships of multiple plant orthologous genes (taking the genes in red for example), deriving from different episodes of whole genome duplication (WGD) events (marked as α, β, and γ) (Figure 4A), thus the true relationship of these genes should be concluded as the phylogenetic tree 1 in Figure 4B. However sequence-based phylogenetic tree building without considering synteny information will most likely construct phylogenetic trees 2-4 (Figure 4B).

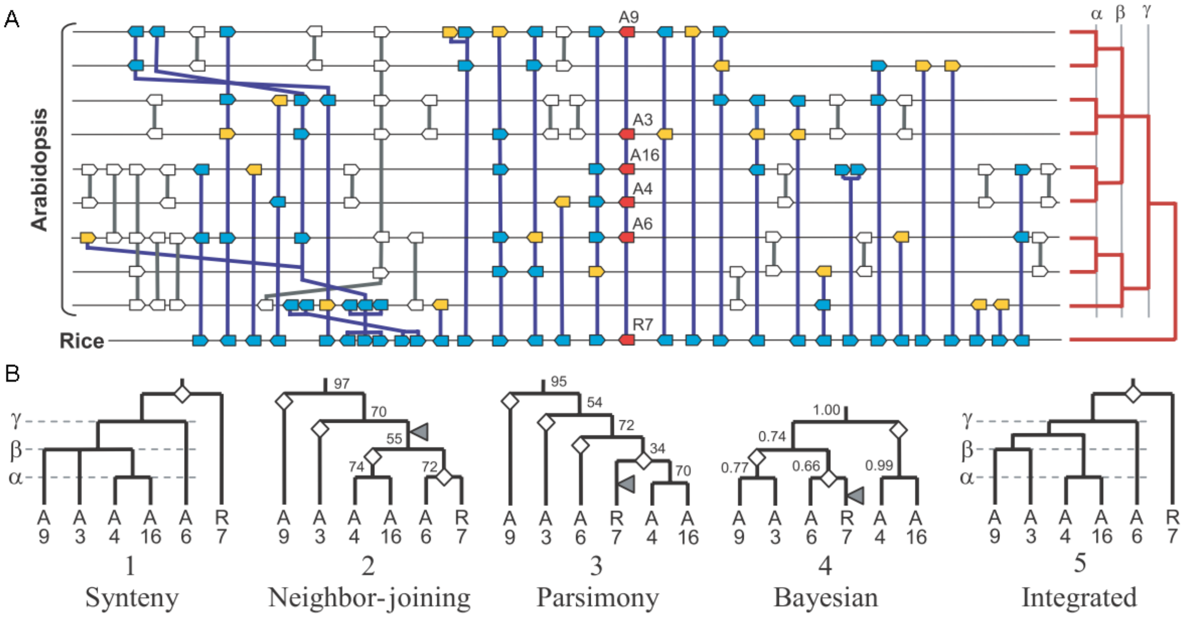


Figure 4. An example of synteny depicting six syntenic genes (highlighted red) from Arabidopsis and rice (A), and (B) sequence-based phylogenetic trees using Neighboring-joining (tree 2), Parsimony (tree 3), and Bayesian (tree 4). Figure adapted from Sampedro et al., 2005.

In the past few years, integration with synteny information has greatly facilitated studies in tracing gene evolution patterns (Causier et al., 2010; Ruelens et al., 2013; Zhao et al., 2017), and also for species delimitation (Magain et al., 2017).

Synteny detection tools and visualization

Since synteny holds key inference for gene orthology and evolutionary history, numerous tools and databases has been developed (Table 1). These tools differ greatly in input data format, multi-species support, processing scale, running time, as well as the resulting output. For comparative studies see Ghiurcuta and Moret, 2014 and Liu et al., 2018.

Table 1. Programs for detecting synteny

Program	Reference	#Citations (up to 1-5-2018)
Absynte (Archaeal and Bacterial Synteny) Explorer	Despalins et al., 2011	33
AutoGRAPH	Derrien et al., 2006	61
Cinteny	Sinha and Meller, 2007	126
ColinearScan	Wang et al., 2006	76
Cyntenator	Rödelsperger and Dieterich, 2010	41
DAGchainer	Haas et al., 2004	255
DiagHunter	Cannon et al., 2003	68
DRIMM-Synteny (Duplications and Rearrangements In Multiple Mammals-Synteny)	Pham and Pevzner, 2010	50
FISH (Fast Identification of Segmental Homology)	Calabrese et al., 2003	125
i-ADHoRe (iterative Automatic Detection of Homologous Regions)	Vandepoele et al., 2002	164
	Simillion et al., 2007	55
	Proost et al., 2012	71
MCMuSeC (Max-gap Clusters by Multiple Sequence Comparison)	Ling et al., 2009	35
MCScan (Multiple Collinearity Scan)	Tang et al., 2008b	329
MCScanX	Wang et al., 2012	373
MicroSyn	Cai et al., 2011	20
Mugsy	Angiuoli and Salzberg, 2010	265
MultiSyn	Baek et al., 2016	1
Murasaki	Popendorf et al., 2010	24
OrthoCluster	Zeng et al., 2008	36

Osfinder (Orthologous Segment finder)	Hachiya et al., 2009	37
PhylDiag	Lucas et al., 2014	6
Proteny	Gehrmann and Reinders, 2015	3
QUOTA-ALIGN	Tang et al., 2011	68
r2cat (Related Reference Contig Arrangement Tool)	Husemann and Stoye, 2009	94
RAGB (Reference-Anchored Gene Blocks	Benshahar et al., 2017	-
Satsuma)	Grabherr et al., 2010	65
Sibelia (Synteny Block Exploration tool)	Minkin et al., 2013	39
SyMAP (Synteny Mapping and Analysis Program)	Soderlund et al., 2011	130
SynChro	Drillon et al., 2014	26
SynFind	Tang et al., 2015	21
Synorth (Synteny and Ortholog)	Dong et al., 2009	25
Synteny Portal	Lee et al., 2016	4
SyntenyTracker	Donthu et al., 2009	29
SynteView	Lemoine et al., 2008	12
SyntTax	Oberto, 2013	52
SynMap	Lyons et al., 2008	121

Despite this amount of tools, comparing synteny of large eukaryotic genomes for multiple species still presents major challenges. First, the number of whole genome comparison is growing exponentially with the genomes to be compared, even for three genomes, nine times (i.e. all inter- and intra- genomes) whole genome comparison is required. Secondly, sequenced genomes are published at various qualities. Many are neither perfectly assembled nor annotated, with some poorly assembled genomes that have 10 to 100 times as many scaffolds as they have chromosomes. So synteny detection based on genome annotations are subject to many possible confounding factors. Last but not least, organizing and presenting syntenic regions of multiple genes across many species is probably the most burning question, which is especially true for flowering plant genomes due to recurrent whole genome duplications (WGDs) (see a recent review: Cheng et al., 2018a).

Pairwise dot plots and reference-based synteny comparisons are most widely adopted at the moment. Synteny analysis across related species has almost become a routine for genome sequencing papers (depending on the question, a careful selection of genomes to be compared is required). There are several search-based platforms such as PGDD (Lee et al., 2013), CoGe (Lyons and Freeling, 2008), Phytozome (Goodstein et al., 2012), Plaza (Proost et al., 2015), and Genomicus (Louis et al., 2012), providing an overview of homologous synteny blocks across multiple species (Figure 5). However it is difficult to automatically check and integrate synteny for, say an entire gene family, that may contains many members from various synteny block families, using these tools.

Considering the increasing number of available genomes, a systematic approach that is able to use as many genomes as possible, that organizes massive pieces of syntenic regions (synteny blocks), and also display/visualize synteny relations of any length of inquiry gene list, should be a target and new fashion for phylogenomic synteny studies.

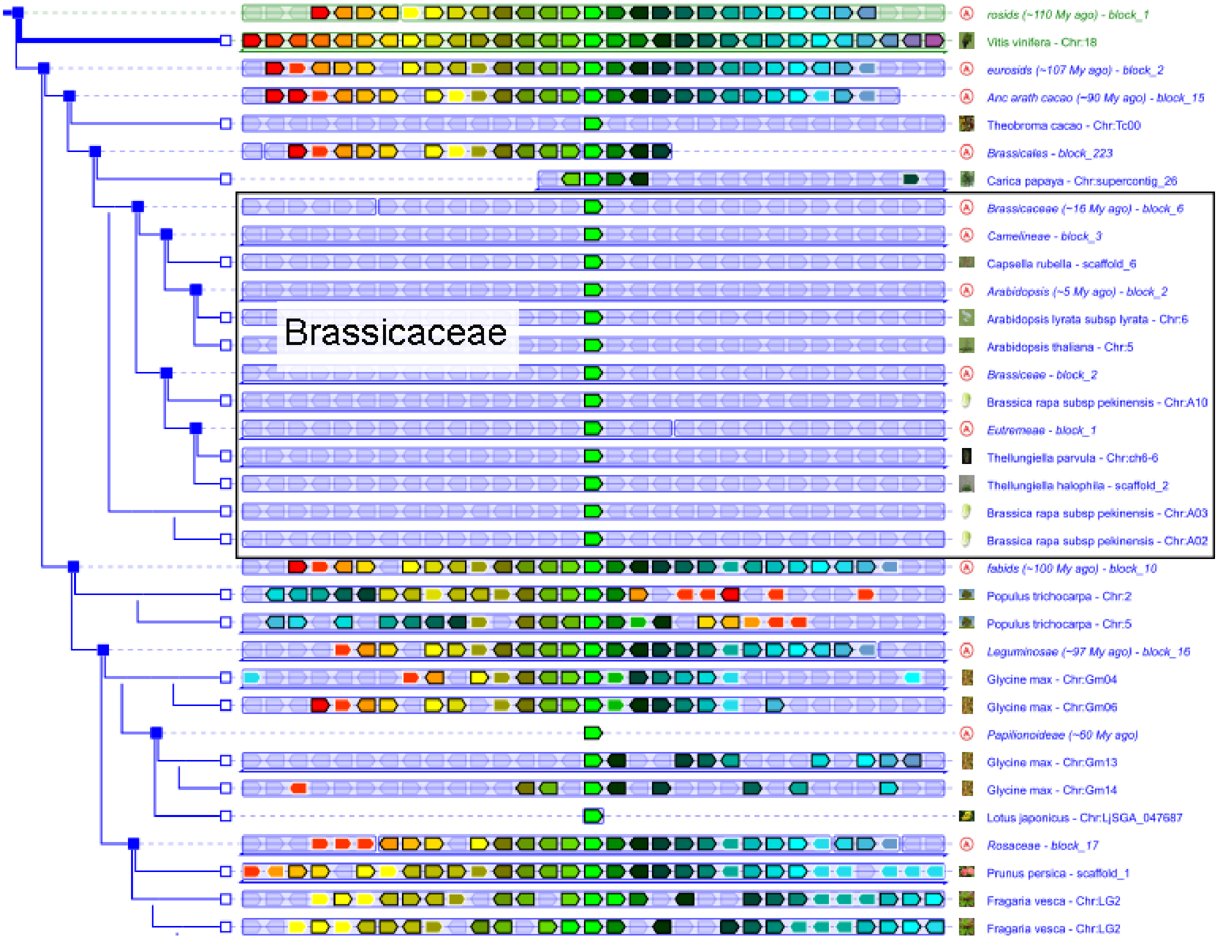

rosids (~110 My ago) - block_1
Vitis vinifera - Chr:18
eurosids (~107 My ago) - block_2
Anc arath cacao (~90 My ago) - block_15
Theobroma cacao - Chr:Tc00
Brassicales - block_223
Carica papaya - Chr:supercontig_26
Brassicaceae
Brassicaceae (~16 My ago) - block_6
Camelineae - block_3
Capsella rubella - scaffold_6
Arabidopsis (~5 My ago) - block_2
Arabidopsis lyrata subsp lyrata - Chr:6
Arabidopsis thaliana - Chr:5
Brassiceae - block_2
Brassica rapa subsp pekinensis - Chr:A10
Eutremeae - block_1
Thellungiella parvula - Chr:ch6-6
Thellungiella halophila - scaffold_2
Brassica rapa subsp pekinensis - Chr:A03
Brassica rapa subsp pekinensis - Chr:A02
fabids (~100 My ago) - block_10
Populus trichocarpa - Chr:2
Populus trichocarpa - Chr:5
Leguminosae (~97 My ago) - block_16
Glycine max - Chr:Gm04
Glycine max - Chr:Gm06
Papilionoideae (~60 My ago)
Glycine max - Chr:Gm13
Glycine max - Chr:Gm14
Lotus japonicus - Chr:LjSGA_047687
Rosaceae - block_17
Prunus persica - scaffold_1
Fragaria vesca - Chr:LG2
Fragaria vesca - Chr:LG2

Figure 5. Genomicus (Louis et al., 2012) phylogenomic synteny viewer of grape *PI* MADS-box gene (Vv18s0001g01760), the figure shows similar genomic contexts across rosid species, except for Brassicaceae species (no flanking genes, boxed).

Research scope: to grasp it as a whole

Synteny detection tools as listed in Table 1 usually start with pairwise all-vs-all sequence comparison for homologous relations before evaluating synteny. This step usually takes extensive computation time to complete, especially when analyzing many genomes (times of comparisons equals the square of the number of genomes used). In order to realize the aforementioned goal, new aligners developed these years has made large-scale eukaryotic genome comparisons possible with a previously unimaginable speed, such as RAPSearch2 (Zhao et al., 2012), DIAMOND (Buchfink et al., 2015), and MMseqs2 (Steinegger and Söding, 2017).

Networks organize rich data, and provide a wide range of mathematical tools for identifying and explaining the patterns they contain. Network science is nowadays a thriving interdisciplinary domain that focuses on the representation, analysis and modeling of complex social, biological and technological systems as networks or graphs (Barabási and Pósfai, 2016, Figure 6). In life science, protein interaction networks, co-expression networks, and ecological networks, present an overview of all pairwise relations from certain data resources, and provide ways of seeing, organizing, and guiding scientific thinking (Goh et al., 2007; Arabidopsis Interactome Mapping, 2011).

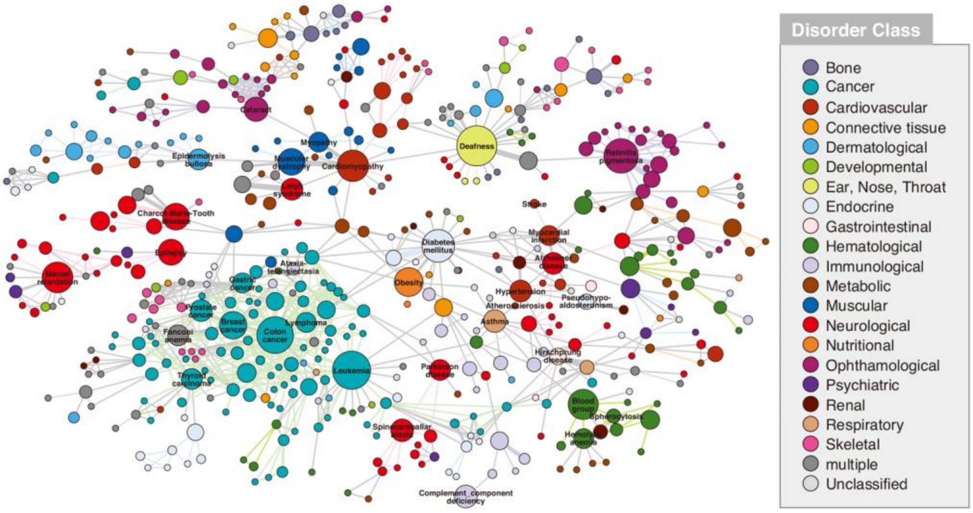


Figure 6. Human disease network. Nodes are diseases, two diseases are connected if the same genes are associated with them. Figure adapted from Goh et al., 2007.

In this thesis, I focus on the construction of synteny-based networks and its applications for the inference of gene evolutionary history. First in **Chapter 2**, we put forward the idea of using networks for phylogenomic synteny data. I present more background about current limitations of large-scale organization and curation of plant synteny connections, then propose an outline performing synteny network analysis, followed by a brief example of floral B-class genes (*AP3* and *PI*). The synteny network supports the previous reported B-gene synteny across species (Causier et al., 2010; Cheng et al., 2013), but now with 101 broadly distributed species grouped in the network graph, results are more clear, direct, robust, and systematic.

In **Chapter 3**, I analyzed synteny networks of the MADS-box transcription factor gene family from 51 complete plant genomes. Through this analysis, the relationships, approximate timing, gains and losses, and specific movements of these genes within the genome could be traced. Specifically, several novel evolutionary patterns were inferred and visualized from synteny network clusters. I found lineage-specific clusters that derive from transposition events for the regulators of floral development and flowering-time in the Brassicales and for the regulators of root-development in Poales. We also identified two large gene clusters that jointly support the idea that these genes are derived from an ancient tandem gene duplication that likely predates the radiation of the seed plants and then expanded by subsequent polyploidy events.

Besides regulative transcription factors such as MADS-box genes, in **Chapter 4** we applied our synteny network approach for an important abiotic stress-induced protective proteins, the LEAs (Late Embryogenesis Abundant proteins). This gene family include eight multi-gene families expressed in response to water loss during seed maturation and in vegetative tissues of desiccation tolerant species. The synteny network indicated that plant LEA families have distinct origins, and that most of them show synteny conservation across angiosperms. Recurrent tandem-duplications, and transpositions contributed to sequence diversification and functional innovations. For example, the dehydrin sub-family of LEAs has diversified ancestral synteny, which resulted in distinct evolution of amino acid sequences, biochemical properties, and gene expression patterns.

In **Chapter 5**, I scaled up the analysis to involve the genomes of 107 flowering plants and 87 mammals, covering major lineages that have evolved and radiated over the last ~170 million years. We built synteny networks for genomes within each kingdom and compare overall genomic architecture conservation and variation, with comparison to other genome metrics such as N50, BUSCO, and genome sizes. We characterized all synteny clusters with phylogenomic profiling, which illustrated all genomic innovations (i.e. duplications, gene transpositions, gene loss) in one graph. We also investigate synteny properties for BUSCO genes, which has been widely-used as conserved single copy genes for benchmarking genome qualities.

Finally in **Chapter 6**, I discuss future perspective of phylogenomic synteny network approach, and proposed future efforts to remedy existing problems.

Chapter 2

Network Approaches for Plant Phylogenomic Synteny Analysis

Tao Zhao[1], M. Eric Schranz[1]

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

Published in *Current Opinion in Plant Biology*

Volume 36, April 2017, Pages 129-134

DOI: https://doi.org/10.1016/j.pbi.2017.03.001

Abstract

Network analysis approaches have been widely applied across disciplines. In biology, network analysis is now frequently adopted to organize protein-protein interactions, organize pathways and/or to interpret gene co-expression patterns. However, comparative genomic analyses still largely rely on pairwise comparisons and linear visualizations between genomes. In this article, we discuss the challenges and prospects for establishing a generalized plant phylogenomic synteny network approach needed to interpret the wealth of new and emerging genomic data. We illustrate our approach with an example synteny network of B-class floral MADS-box genes. A broad synteny network approach holds great promise for understanding the evolutionary history of genes and genomes across broad phylogenetic groups and divergence times.

Introduction

Gene and genome duplication plays a fundamental role in evolution by providing the genetic material by which new traits can arise (Conant et al., 2014; Panchy et al., 2016). Thus, comparative analysis of synteny is a powerful approach to understand the evolutionary trajectory of genes and genomes and ultimately of traits and organisms (Lee et al., 2013; Jiao et al., 2014). However, in the plant kingdom recurrent whole genome duplication events, genomic rearrangements, and gene translocations complicate such analyses. Although several online tools now exist for analyzing synteny comparisons between species (Lyons et al., 2008; Proost et al., 2015; Veltri et al., 2016) , they represent syntenic blocks as either parallel coordinate plots or pairwise dot-plots. Such approaches to data organization and presentation restricts us from inferring global syntenic relationship patterns of entire multi-gene families or across multiple genomes simultaneously.

In this chapter, we discuss the prospects for a network-based approach for comparative synteny analysis. Genome syntenic blocks derived from intra- and inter-species comparisons can then be abstracted into vertices (nodes or points) and edges (lines between points). Sets of widely-available tools for characterizing networks can then be applied to such a synteny network to allow for visualization and clustering (Jarukasemratana and Murata, 2013; Domenico et al., 2015). We illustrate our concept of plant synteny network analysis for organizing, visualizing and finding patterns in the vast wealth of emerging plant genomic data by comparing such an approach with a previous published traditional parallel synteny coordinate plot example for the floral B-class MADS-box genes, *AP3* and *PI*. The new approach can provide new insights into the dynamics of gene and genome evolution such as the detection of gene transpositions and ancient gene tandem-duplications (Zhao et al., 2017).

Current limitations for visualizing global synteny relations across plants

Conserved synteny across species is an essential foundation for genomic research. Several existing computational programs and online tools have been developed to assess synteny for comparative genomic studies of two or more genomes, e.g. MCScanX (Wang et al., 2012), DAGchainer (Haas et al., 2004), i-ADHoRe (Proost et al., 2012), DRIMM (Pham and Pevzner, 2010), SynFind (Tang et al., 2015), and PGDD (Lee et al., 2013) among many others. To compare each of them is beyond the scope of this article, but for an overview please see (Ghiurcuta and Moret, 2014; Gehrmann and Reinders, 2015). As an output mummer-plots or parallel coordinate plots are widely adopted to display syntenic genomic regions (Cheng et al., 2013; Ruelens et al., 2013). Such visualization approaches can be quite helpful to trace the positional history of a particular gene of interest. However, the complexity increases dramatically when analyzing many species across broad phylogenetic groups (such as monocots vs. eudicots). This is particularly true because of the numerous independent polyploidy events (Li et al., 2015). For example, *A. thaliana* has undergone at least three polyploidy events (Vanneste et al., 2014) since its common ancestor with the entire angiosperm clade, and as a consequence ohnologs (paralogs derived from a polyploidy event) occupy multiple tracks using these tools. Furthermore, most tools display the information about flanking neighbor genes used to establish collinearity. However, when trying to understand the overall evolutionary pattern of large multi-gene families across multiple species it is nearly impossible to visualize.

Recent progress in pan-genomics has witnessed the application of network approaches for understanding multiple bacterial genomes and within species variation of large plant genomes (Marschall et al., 2016), such as Arabidopsis (Consortium, 2016), maize (Jiao et al., 2012b; Hirsch et al., 2014), and soybean (Li et al., 2014). New methods and tools are rapidly being developed for better visualization and improved data exploration (Baier et al., 2015; Sheikhizadeh et al., 2016). However, pan-genome approaches currently are more applicable to within species comparisons and not across broad phylogenetic scales of large genome species with shared and independent polyploidy events like plants.

Recently, evolutionary inferences for certain groups of genes has been made in several studies using networks to organize syntenic relationships based on PGDD-derived data across species (Li and Zhang, 2013; van Veen et al., 2014; Hammoudi et al., 2016). Networks focused just on the target loci can more easily display the relationships by visualizing syntenic connections into points (nodes/vertices) and lines (edges). In this chapter, a systematic methodology for large-scale synteny network construction and analysis will be outlined.

Sharpening the tools: outline of synteny network analysis

A network is comprised of all interconnected pairs of nodes. The data source for synteny networks is derived from all pair-wise computed syntenic blocks. Thus, synteny network construction consists of three primary steps: (1) pairwise whole-genome comparisons, (2) syntenic block detection and data fusion, and (3) network manipulation (Figure 1).

Pairwise whole genome comparisons

For a comparison of five plant genomes you would need to perform P (5, 2) + 5 = 25 whole genome all-against-all comparisons. For each pair of syntenic blocks, reciprocal comparisons are needed as well as all intra-species comparisons are needed to identify paralogs or “syntenic ohnolog pairs”. Currently, whole genome sequences are available for over 100 plant genomes (Goodstein et al., 2012; Jin et al., 2016). Accordingly, to build a synteny network for all these plant genomes over tens of thousand times of whole-genome protein comparisons are then needed. For such a task, standard tools like BLAST, or BLAST+ are too computationally demanding for most researchers (Altschul et al., 1990; Camacho et al., 2009). New algorithms like LAST or RAPSearch2 (>1000 times speedup over BLAST in the accelerating mode) are more capable for this type of analysis (Kiełbasa et al., 2011; Zhao et al., 2012).

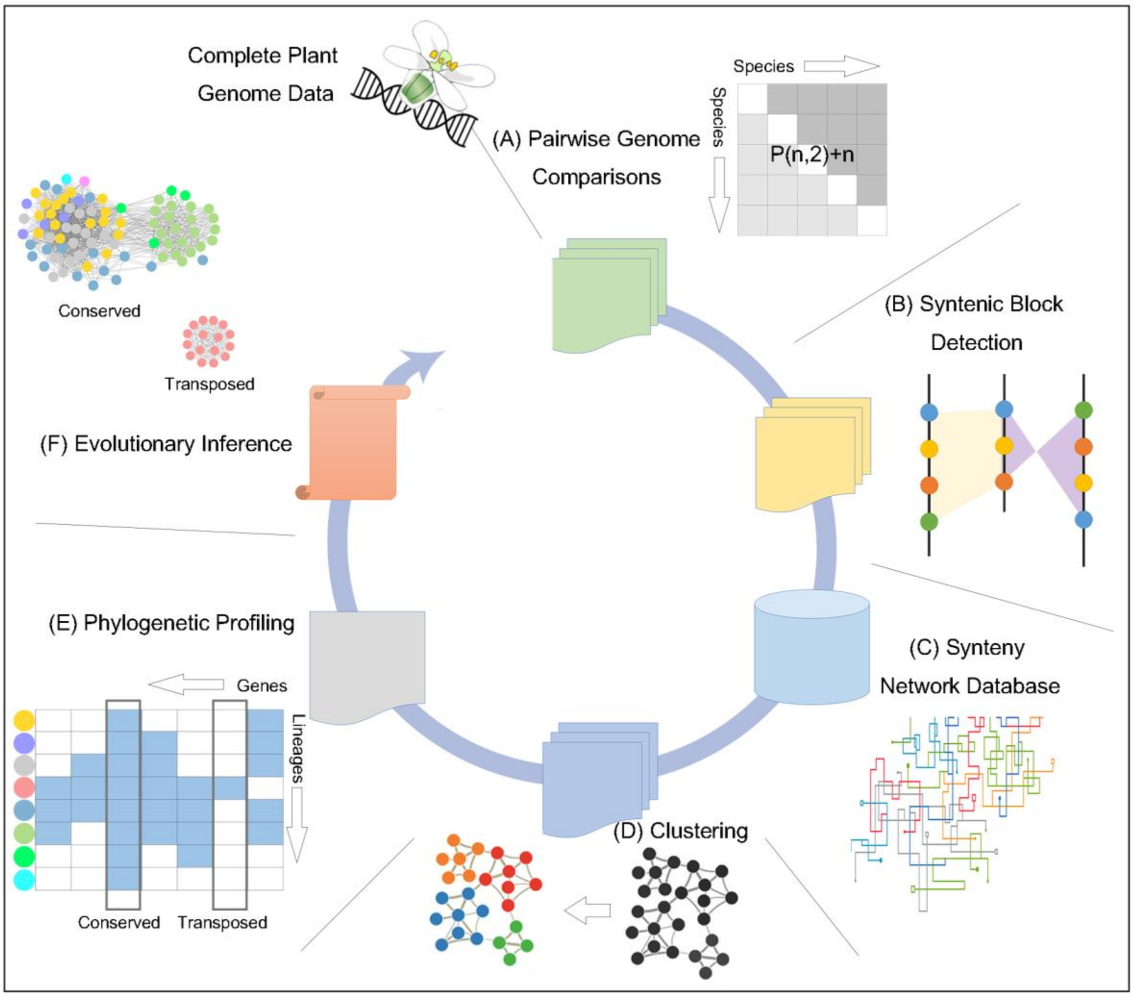


Figure 1. Synteny network construction and analysis consists of several steps: (A) Pairwise genome comparisons, followed by (B) syntenic block detection and data integration, in order to construct (C) a synteny network database that allows for (D) network data clustering and (E) phylogenetic profiling. The steps allow for (F) enhanced evolutionary inferences.

Syntenic block detection and data fusion

Syntenic blocks are found within a certain window size according to different scoring schemes (Proost et al., 2012; Wang et al., 2012). MCScanX for example takes a default window size of 50 neighboring genes for inferring significant collinearity. That is, one matched gene pair (one anchor) with adjacent potential anchors no further than 25 upstream and downstream genes respectively on genomic/chromosome/scaffold position (which can be inferred from the bed/GFF files). Syntenic blocks with a block-score over 250 (i.e. involving at least 5 collinear gene pairs) are reported (Wang et al., 2012a). Homologous collinear gene pairs within the block share the same block-score. Outputs can be merged in a tabular format keeping: (1) unique block ids, indicating species being compared and numeric index, (2) syntenic gene pairs, and (3) a block-score (for weighted network edges).

Network manipulation

The synteny network of plant genomes can contain many millions of nodes and edges. Such a complex "web" of syntenic connections is of too great complexity to interpret directly. Thus, a clustering step is needed to locate communities before further analysis. Identifying communities from a network is currently a popular topic, thus there has been much effort in the last decade for developing community detection algorithms (Fortunato, 2010). Considering synteny network characteristics, computation efficiency, and clustering quality, several methods for microsynteny network clustering are most appropriate: (1) clique percolation method (Palla et al., 2005) that can find overlapping dense groups of nodes in networks, providing insights of interconnected sub gene-families; (2) *k*-core decomposition (Alvarez-Hamelin et al., 2005; Khaouid et al., 2015) that efficiently partitions a large graph / network into layers from external to more central vertices; and (3) Infomap (Rosvall and Bergstrom, 2008; Rosvall et al., 2014) that is fast and accurate, providing remarkable performance for handling millions of nodes in short time.

Several factors that are important to consider during synteny network construction include: the quality of the genome assemblies and annotations, the synteny block detection settings (for example, block scores or number of syntenic genes used) and the settings and optimization of the clustering algorithms (for example, k value in k-cliques percolation). Altering these parameters can impact the robustness of the synteny networks (Zhao et al., 2017).

'Network Effect': deciphering phylogenomic evolutionary patterns

To illustrate our concept, we have built a synteny network for one hundred and one plant genomes (including 10 genomes in Fabaceae, 14 in Brassicaceae, 15 in Asterids, 14 in Poaceae, 4 in Pinaceae, 4 in Chlorophyta, etc.), which contains ~1.8 million nodes and 3.2 million edges (Figure 2). Such a network database thus containing a tremendous wealth of syntenic information. One application is to extract and visualize syntenic relationships for certain genes or gene families across and within species (depending on the question at hand). For example, *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) genes are important MADS-box transcription factors that specify petal and stamen identity (known as B-class floral regulatory genes) (Dodsworth, 2016; Theissen et al., 2016). A previous comparative study of floral MADS-box genes between the sister families Cleomaceae (*Tarenaya hassleriana*) and Brassicaceae (e.g. Arabidopsis and Brassica) found that *T. hassleriana AP3* genes are non-syntenic to Brassicaceae species, instead they are syntenic to other eudicot species (Figure 2a) and that is because Brassicaceae *AP3* genes were transposed probably due to the At-alpha whole genome duplication (Figure 2b). Interestingly in *T. hassleriana* one *PI* gene is syntenic to other eudicot species (Figure 2c) and the other one syntenic to Brassicaceae *PI* homologs (Figure 2d) (Cheng et al., 2013). Comparatively, for such a result, an updated synteny network approach clearly shows four distinct clusters for *AP3* and *PI* genes

(Figure 2, right panel) that can be more easily visualized than the four parallel synteny coordinate plots from Cheng *et al*. (Cheng et al., 2013) (Figure 2, left panel) . Each node represents an *AP3* (Figure 2a and 2b) or *PI* gene (Figure 2c and 2d) from species belonging to various lineages (indicated by different colors), two nodes are connected if they are syntenic to each other (unweighted, i.e., edge length not informative).

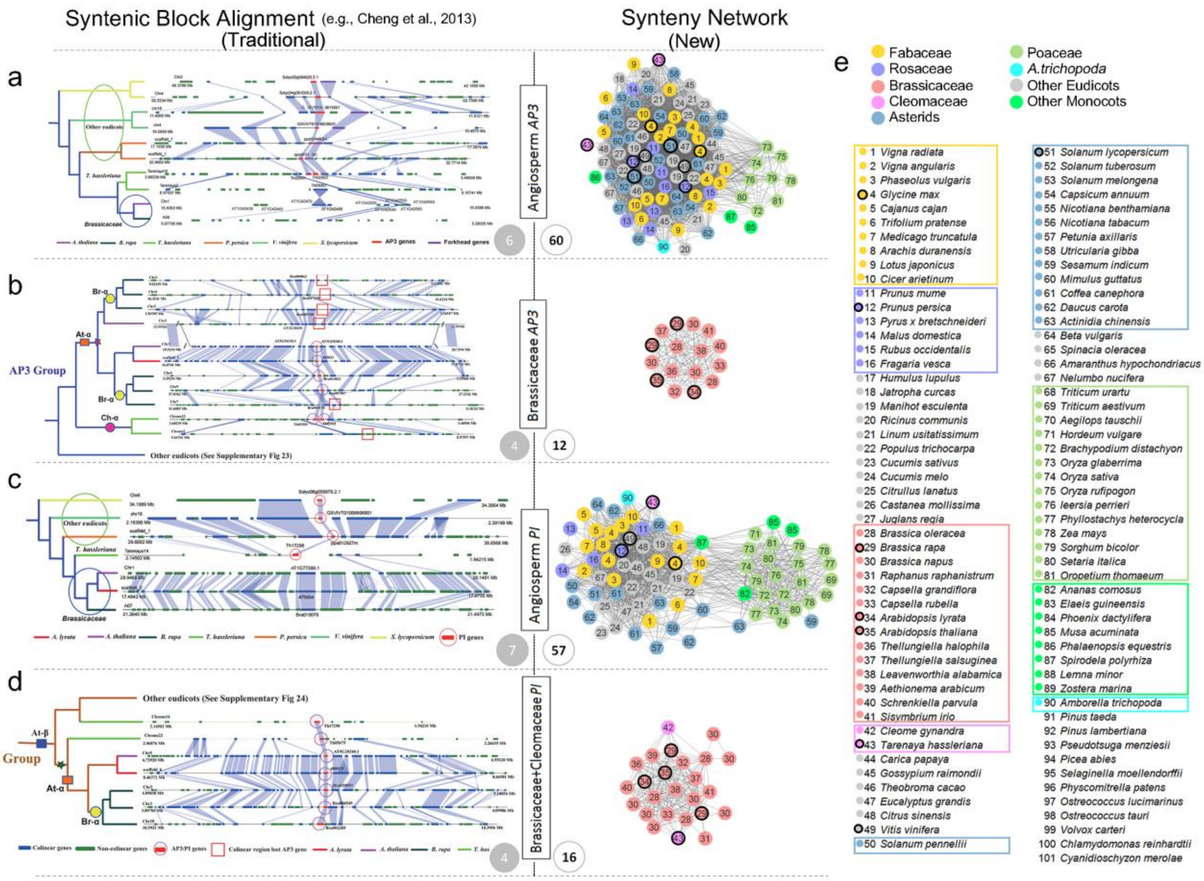


Figure 2. Comparison of traditional parallel synteny coordinate plots (adapted and modified from (Cheng et al., 2013) with permission from the authors) and new synteny network approach. (a) Angiosperm-wide synteny of AP3 genes, (b) synteny of Brassicaceae AP3 genes, (c) Angiosperm-wide synteny of PI genes, and (d) synteny of Brassicaceae and Cleomaceae PI genes. A number indicating involving species for synteny comparison is shown in the corner of each panel. Genes from (Cheng et al., 2013) at the left panel are highlighted as nodes with a thick black border at the right panel. (e) Species in different angiosperm lineages used are listed and indicated by different colored nodes.

In this way, only syntelogs derived from the pre-calculated synteny blocks are depicted in the network, from which we can infer the number of syntelogs and relationships among the remaining syntelogs after polyploidy events. For example, in the Brassicaceae AP3 synteny network (Figure 2b) most crucifers are represented by single nodes (i.e. Arabidopsis is node 35). Whereas, the mesopolyploid species *Brassica* *oleracea* and *B. rapa* (nodes 28 and 29) have two copies each. Thus, we can conclude that after the genome triplication of the Brassica lineage, two syntelogs have been retained and one has been lost. In the allopolyploid species, B. napus (nodes labelled 30), three AP3 syntelog nodes are detected (rather than the expected four: two from *B. rapa* and two from *B. oleracea*). Thus, one *B. napus* AP3 syntelog has either been lost since polyploidization or was not detected because it has transposed or was not assembled or annotated correctly. The synteny network of AP3 and PI genes support previous B-gene synteny across species (Causier et al., 2010; Cheng et al., 2013), but now with 101 broadly distributed species grouped in the network graph, results are more clear, direct and robust. The additional evidence from the network clusters give greater support for the B-gene transpositions and thus unique synteny of Brassicaceae and Cleomaceae homologs (Figure 2b and d). Whereas synteny of AP3 and PI homologs are highly conserved across most other angiosperms including eudicots, monocots and *A. trichopoda* (the sister lineage of other flowering plants (Albert et al., 2013)). Our network approach broadens the inference of the evolutionary history of the important MADS-box B-genes in a more comprehensive and systematic way.

Conclusion and future prospects

Synteny-based analysis is widely recognized as an effective and reliable approach for comparative genomics. However, effective large-scale organization and curation of microsynteny connections remains limited. With the rapid increase of completed plant whole genomes and a wealth of new algorithms and tools for network inference, we believe that our outline for building synteny networks from large syntenic connections of pairwise syntenic blocks across many species will garner new insights into genome and gene family evolutionary history. Furthermore, the network approach is a reliable method to identify and organize syntelogs derived from the frequent and independent ancient and recent WGDs in plants. Network statistical and mathematic parameters provide a framework for testing hypotheses of gene family expansion and contraction in a phylogenomic context. For example, node degree, clustering coefficient, betweenness centrality, and so on. can all be used to characterize and quantify various phylogenomic evolutionary features of specific genes and gene families (e.g. under certain cluster sizes, a higher clustering coefficient indicates a more stable genomic context). Gene families known to undergo rapid expansion/contractions and/or transposition events (such as NB-LRR, p450 and F-box genes) would have average low clustering coefficients whereas genes known to be preferentially retained in duplicate after whole-genome duplications (such as many transcription factor families) would have higher average clustering coefficients. A comprehensive analysis of pan-angiosperm synteny networks can serve as a novel starting point for understanding angiosperm genome organization, biology and evolutionary history.

Acknowledgements

TZ was supported by a funding from the China Scholarship Council (http://en.csc.edu.cn/). MES was supported by the Netherlands Organization for Scientific Research (NWO VIDI Grant 864.10.001).

Chapter 3

Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation

Tao Zhao[1], Rens Holmer[2], Suzanne de Bruijn[2], Gerco C. Angenent[2], Harrold A. van den Burg[3], M. Eric Schranz[1]*

[1]Biosystematics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[2]Laboratory for Molecular Biology, Wageningen University, 6708 PB Wageningen, The Netherlands
[3]Molecular Plant Pathology, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

Published in *The Plant Cell*

Volume 29, June 2017, Pages 1278-1292

DOI: https://doi.org/10.1105/tpc.17.00312

Abstract

Conserved genomic context provides critical information for comparative evolutionary analysis. With the increase in numbers of sequenced plant genomes, synteny analysis can provide new insight into gene family evolution. Here, we exploit a network analysis approach to organize and interpret massive pairwise syntenic relationships. Specifically, we analyzed synteny networks of the MADS-box transcription factor gene family using fifty-one completed plant genomes. In combination with phylogenetic profiling, several novel evolutionary patterns were inferred and visualized from synteny network clusters. We found lineage-specific clusters that derive from transposition events for the regulators of floral development (*APETALA3* and *PI*) and flowering-time (*FLC*) in the Brassicales and for the regulators of root-development (*AGL17*) in Poales. We also identified two large gene clusters that jointly encompass many key phenotypic regulatory Type II MADS-box gene clades (*SEP1, SQUA, TM8, SEP3, FLC, AGL6* and *TM3*). Gene clustering and gene trees support the idea that these genes are derived from an ancient tandem gene duplication that likely predates the radiation of the seed plants and then expanded by subsequent polyploidy events. We also identified angiosperm-wide conservation of synteny of several other less studied clades. Combined, these findings provide new hypotheses for the genomic origins, biological conservation and divergence of MADS-box gene family members.

Introduction

Conserved gene order can be retained for hundreds of millions of years and provides critical information about conserved genomic context and the evolution of genomes and genes. For example, the well-known “Hox gene cluster”, which regulates the animal body plan, is largely collinear across the animal kingdom (Lewis, 1978; Krumlauf, 1994; Ferrier and Holland, 2001). The term synteny was originally defined as a set of genes from two species located on the same chromosome, but not necessarily in the same order (Dewey, 2011, Passarge et al. 1999). However, the current widespread usage of the term synteny, that we adopt, now implies conserved collinearity and genomic context. Synteny data are widely used to establish the occurrence of ancient polyploidy events, to identify chromosomal rearrangements, to examine the expansion and contraction of gene families, and to establish gene orthology (Sampedro et al., 2005; Tang et al., 2008a; Dewey, 2011; Jiao and Paterson, 2014). Synteny likely reflects important relationships between the genomic context of genes both in terms of function and regulation and, thus, is often used as a ‘proxy for the conservation or constraint of gene function’ (Dewey, 2011; Lv et al., 2011). Syntenic relationships across a wide range of species thus provide crucial information to address fundamental questions on the evolution of gene families that regulate important developmental pathways. The origin of morphological novelty has been linked for example to the duplication of key regulatory transcription factors in the case of the Hox-genes in animals, but also the MADS-box genes in plants (Alvarez-Buylla et al., 2000b; Airoldi and Davies, 2012; Soshnikova et al., 2013). However, gene clusters are frequently dispersed or “broken-up” in certain lineages, like the Hox-cluster in the genomes of octopus (Lemons and McGinnis, 2006; Duboule, 2007; Albertin et al., 2015) and brachiopods (Schiemann et al., 2017), and this dispersion contributes to divergent gene expression and morphological novelties.

In plants, the MADS-box genes are critical transcription factors that regulate the developmental pattern of the floral organs, the reproductive organs, and other traits (Theissen, 2001; Becker and Theissen, 2003; Smaczniak et al., 2012). For instance, floral organ identity is controlled largely by MADS-box genes, as explained by the ABC(DE) model (Figure 1a) (Coen and Meyerowitz, 1991; Ditta et al., 2004) with, for example, the floral A-, B-, and E-function genes being required for petal identity (Figures 1a and 1b). Synteny data of the MADS-box genes have been used to infer the ancestral genetic composition of the B- and C-function (Causier et al., 2010), and the A- and E-function genes (Ruelens et al., 2013; Sun et al., 2014). However, these studies analyzed only a small number of species (fewer than 10) and the results were displayed as parallel coordinate plots (as in Figure 1c). A systematic comparison of the syntenic relationships for all the MADS-box genes across many plant species has not been done in a single study. That is because this gene family has undergone extensive duplications that have given rise to complicated relationships of orthology, paralogy, and functional homology (Jaramillo and Kramer, 2007). Hence, a systematic investigation in which all the possible syntenic relationships between the family members are sorted and visualized is challenging. With the increase of genomes that are simultaneously analyzed, it becomes increasingly more difficult to organize and display such syntenic relationships. This is due to the ubiquity of ancient and recent polyploidy events, as well as smaller scale events that derive from tandem and transposition duplications (Lynch and Conery, 2000; Bowers et al., 2003; Tang et al., 2008a; Schranz et al., 2012).

Here, we present a novel approach to cluster synteny networks and then analyze gene ancestry. Instead of presenting syntenic blocks as either parallel coordinate plots (Figure 1c) or pairwise dot-plots, we abstracted genome syntenic blocks (derived from intra- and inter-species comparisons) into vertices (nodes or points) and edges (lines between points). Syntelogs (syntenic homologous genes) of a target gene or gene family of interest can be highlighted in one graph without showing the flanking genes (Figure 1d). For example, the syntenic relationships of “Gene 2” across five species (Species A, D, E, F, and G) in Figure 1c can be represented as a cluster of five nodes, with edges representing their syntenic relationships (Figure 1d, Cluster 1). If one gene has undergone an additional duplication event, such as tandem and/or polyploid duplication, (for example “Gene 5” that is a tandem-duplicated in Species E (E5a and E5b) and ohnologs (syntelogs derived from polyploidy events) retained in Species F and G in Figure 1c), these duplicated syntelogs are included as nodes rather than adding additional linear panels to a parallel coordinate plot (Figure 1d, Cluster 2).

Potential ancient tandem-duplications can also be readily represented and detected by synteny network analyses. Cluster 3 illustrates an example where both “Gene 4” and “Gene 5” genes are found in one cluster (Figure 1d, Cluster 3). Such a result can occur when “Gene 4” and “Gene 5” belong to a same gene family and contain the same protein domain(s). Unlike the tandem duplication of the example of “Gene E5a” and “Gene E5b” (Figure 1d, Cluster 2), they may be derived from an ancient tandem duplication and thus evolved a certain degree of differences at the gene sequence level (and thus may even belong to different clades/subgroups of one gene family). “Gene 4” and “Gene 5” can be calculated as syntenic to each other by synteny detection programs when one of the loci was lost. For example, “Gene A4” is found to be syntenic to “Gene B5” because the best option “Gene B4” may has been lost in Species B. As a result, we obtain a twin-cluster layout with more “intra-links” than “inter-links” (Figure 1d, Cluster 3). It is worth mentioning that sometimes one specific node connects (radiates) to other unconnected nodes. For example, node “Gene E7” in Cluster 4 (Figure 1d, Cluster 4) radiates to seven other nodes of syntelogs of “Gene 8”, which belong to another gene family different from the one of “Gene 7”. This is because “Gene E7” contains both domains of “Gene 7” and “Gene 8”, either because of a potential genome mis-annotation or a real protein-domain fusion.

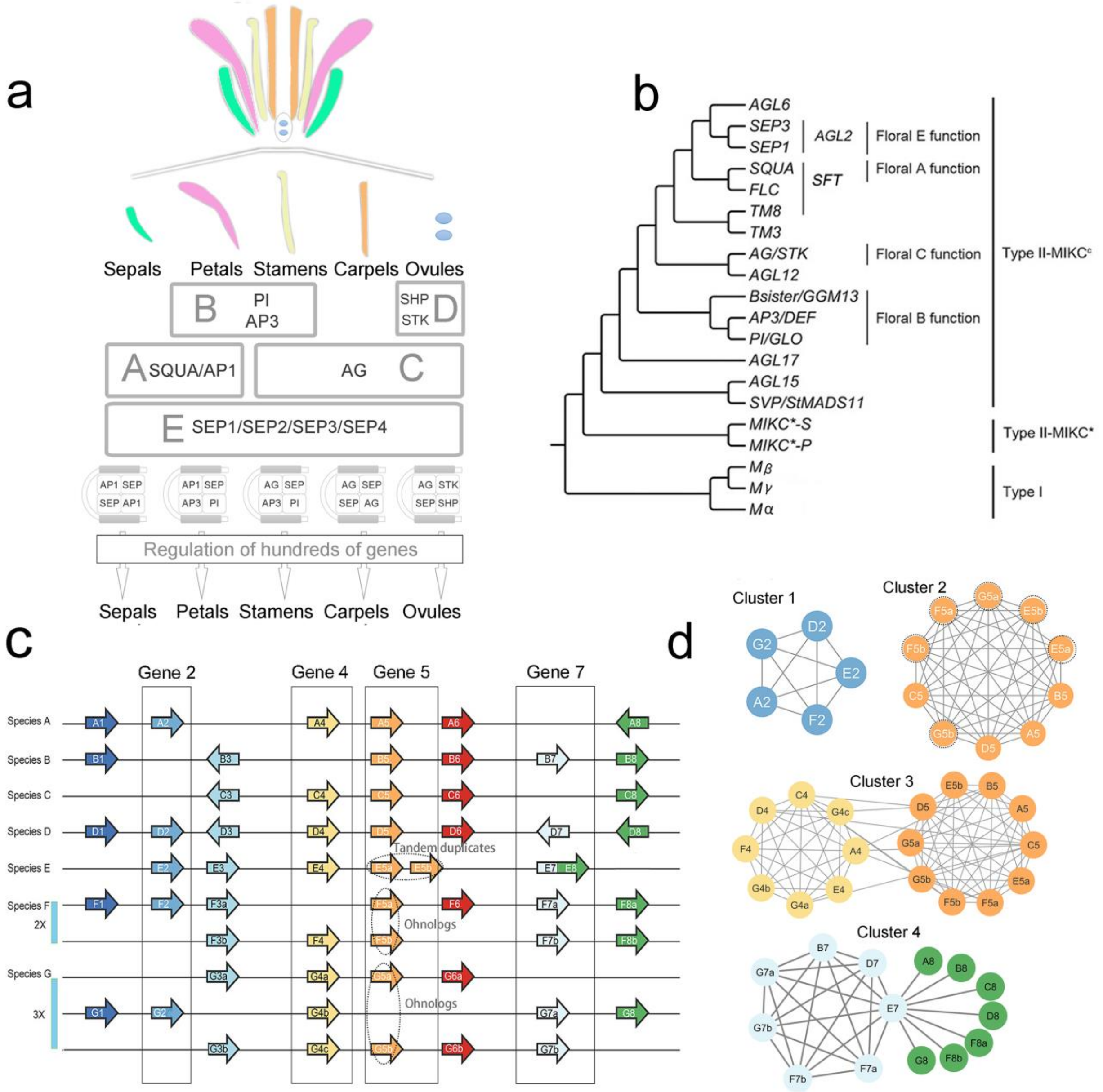


Figure 1 Summary of the MADS-box genes and principles of Phylogenomic Synteny Network Analysis. (**a**) Diagram illustrating the ABCDE floral development model. The A- and E-function genes are essential for sepal identity; the A-, B-, and E-function genes for petal identity; the B-, C-, and E-function genes for stamen identity; the C- and E-function genes for carpel identity, and the D-, C-, and E-function genes for ovule identity. (**b**) The consensus phylogenetic tree showing the relationships for the different functional gene clades of the MADS-box gene family. The combined clade containing the *SQUA*-, *FLC*-, and *TM8*-like genes is referred to as *SFT* in this study. (**c**) Hypothetical example of a parallel coordinate plot for synteny comparisons across seven species (A-G) for which species F has undergone a Whole Genome Duplication (WGD = 2x) and species G a Whole Genome Triplication (WGT = 3x). Examples of tandem duplicates and ohnologs/syntelogs of Gene 5 are indicated by the dotted ovals. Genes 2, 4, 5, and 7 are each boxed as examples of network view of synteny relationships. (**d**) Synteny network of Gene 2 (Cluster 1, less conserved), Gene 5 (Cluster 2, highly conserved and includes tandem duplicates and ohnologs), Gene 4 and Gene 5 (Cluster 3, where genes are members of larger gene family and thus are interconnected, in this case we suppose Gene 4 and Gene 5 belong

to a same gene family (which share similar domain(s)), and synteny is checked for this whole gene family), and Gene 7 (Cluster 4, where gene E7 represents an actual gene fusion of protein domain from neighboring gene or is an artifact due to gene mis-annotation, only synteny of Gene 7 homologs is being checked synteny). Nodes represent syntenic genes and edges (lines) represent a syntenic connection between two nodes. Edge-length in this example is non-informative.

With this background for visualizing synteny networks, we can proceed with their construction and use for understanding evolutionary patterns. We refer readers to a recently published outline of our generalized approach to construct synteny networks (Zhao and Schranz, 2017). The construction of synteny networks uses three main steps: (i) pairwise whole-genome comparisons, (ii) detection of syntenic blocks and data fusion, and (iii) finally network clustering. The first two steps provide a database of syntenic relationships between homologous genes for the genomes analyzed using standard programs, such as BLAST (Altschul et al., 1990) for genome comparisons and MCScan (Tang et al., 2008b) for synteny detection. The final step, the network clustering, can make use of a wide range of clustering algorithms and methods (for review see (Lancichinetti and Fortunato, 2009; Fortunato, 2010)) and are at the heart of our synteny network analysis. The resulting clusters can differ from each other according to the methods applied. Here we use CFinder to cluster our pairwise synteny data, which allows the detection of overlapping communities in network data by using the *k*-clique percolation method (Palla et al., 2005; Palla et al., 2007). *K*-clique corresponds to a fully connected sub-selection of *k* nodes (e.g., a *k*-clique of *k* = 3 is equivalent to a triangle). Two *k*-cliques are considered adjacent and thus form a *k*-clique-community if they share *k*-1 nodes (Derenyi et al., 2005; Palla et al., 2005).

To illustrate this approach, we analyzed the well-characterized MADS-box gene family. The relationships between the major clades of the plant MADS-box genes have already largely been inferred in various phylogenetic and evolutionary studies (Becker and Theissen, 2003; Martinez-Castilla and Alvarez-Buylla, 2003; Nam et al., 2003; Nam et al., 2004; Nam et al., 2005; Gramzow et al., 2012; Smaczniak et al., 2012; Kim et al., 2013; Ruelens et al., 2013; Gramzow et al., 2014; Sun et al., 2014; Yu et al., 2016b) (Figure 1b). However, these studies cannot fully resolve some of the deepest nodes of the MADS-box gene tree. The genome of the model plant *Arabidopsis thaliana* contains a total of 107 MADS-box genes, which derive from multiple gene duplication events (Martinez-Castilla and Alvarez-Buylla, 2003; Parenicova et al., 2003). The MADS-box genes can be divided into two major clades, termed Type I and Type II. The Type II lineage is further divided into the MIKC^C- and MIKC*-types (Henschel et al., 2002). The function and evolution of MADS-box genes have been extensively studied, especially the MIKC^C-types (for review see (Smaczniak et al., 2012)). For convenience, we hereafter refer to the hypothesized common ancestral genes of the ***S****QUA*-, ***F****LC*- and ***T****M8*-like genes as *SFT* genes.

Here we present and discuss the synteny network of all the detected MADS-box genes in 51 plant genomes. This network includes intra- and interspecies syntenic blocks that derive from both shared but also independent polyploidy events in these 51 species. In combination with phylogenetic analysis and phylogenetic profiling (Pellegrini et al., 1999), we could elucidate several previously undetected evolutionary patterns of gene transposition, gene duplication and shared deep ancestry for different MADS-box gene clades. Our approach sheds new light on the evolutionary trajectory of the MADS-box genes and thereby of the traits they control in different plant lineages. Our approach can be easily applied to other gene families and genomes following the step-by-step workflow given on GitHub (https://github.com/zhaotao1987/SynNet-Pipeline).

Results

Overview of the synteny network pipeline

In this study, we analyzed 51 plant genomes covering green algae, mosses, gymnosperms, and angiosperms (Supplemental Table 1 and Supplemental Figure 1). We analyzed all protein models from these genomes for all possible intra- and inter-species whole genome comparisons (Figure 2a). We then built a database that contains all the links between syntenic gene pairs present in syntenic genomic blocks identified by the tool MCScanX (Tang et al., 2008b; Wang et al., 2012). This database contains in total 921,074 nodes (i.e., genes that were connected by synteny with another gene) and 8,045,487 edges (i.e., pairwise syntenic connections); the data can be downloaded from GitHub (https://github.com/zhaotao1987/SynNet-Pipeline).

We used this database to investigate the syntenic relationships between the MADS-box genes. To this end we used HMMER (Finn et al., 2011) to screen the predicted protein sequences of the 51 genomes to identify all the MADS-box genes in these genomes (Supplemental Data Set 1, sheet 1). The resulting list with candidate MADS-box genes was subsequently used to extract the synteny sub-network for these MADS-box genes from the entire network database. This sub-network contained in total 3,458 nodes (MADS-box genes) that were linked by 25,500 syntenic edges (Supplemental Data Set 1, sheet 2). We visualized this sub-network using Gephi (Bastian et al., 2009) and color-coded the clusters using the *k*-clique percolation clustering method with *k* = 3 (Figure 2b). This network and its identified clusters give a first impression on how the MADS-box genes are positionally related to each other across all angiosperms lineages (Figure 2b). The network did not contain synteny information that linked to the non-angiosperm species, which is likely due to the extreme phylogenetic distance and the limited sampling of non-angiosperms species. The node size shown indicates the number-of-connections for each node (Figure 2b). To reveal syntenic relationships between distant gene clades, we then displayed pairwise syntenic relationships between the MADS-box genes in a gene tree that we constructed for the entire gene family (Figure 2c). The colors of the connecting lines indicate again the network communities defined at *k* = 3 from Figure 2b. Interestingly, we found genes from distal gene clades (shown in Figure 1b) that are syntenically connected, such as *SEP1*-like (floral E genes) with *SQUA*-like (floral A genes) genes, *AGL6*- with *TM3* (*SOC1*-like) genes, and *StMADS11* (*SVP*-like) with *AGL17*-like genes (Figures 2b and 2c).

Using CFinder, we detected all cliques of size *k* = 3 to *k* = 24 for the MADS-box gene synteny network and the number of *k*-clique-communities under each *k*-clique (Supplemental Figure 2a). Each of the community (cluster) sizes under a certain *k* is shown (Supplemental Figure 2a), which quantifies the strength of the syntenic connections across species. For example, the *AP3*-like genes of monocot species (green nodes) are only part of a community at relatively low *k* values (*k* < 8) (Supplemental Figure 2b). This could be due to several factors, including the larger genome sizes of monocots (making it more difficult to detect synteny), the limited number of monocot genomes included and/or the lack of phylogenetic sampling across the monocots (i.e. there are many Poales genomes included (7/11), but few other monocot lineages (4/11)).

A clique size of *k* = 3 to 6 was identified to best approximate the true number of communities (Derenyi et al., 2005; Palla et al., 2005; Porter et al., 2009; Xie et al., 2013). We obtained ninety-five clusters using *k* = 3 (Supplemental Data Set 2), and we used these clusters for phylogenetic profiling (Supplemental Figure 3). Each column depicts a syntenic occurrence for a certain MADS-box gene cluster in each plant species. Thereby the presence/absence of syntenic gene clusters across the 51 analyzed taxa are represented by their respective phylogenetic profiles to determine and infer evolutionary patterns (Supplemental Figure 3). We have highlighted twenty-six relevant (i.e. either broad conservation or lineage-specific) clusters in the phylogenetic profile (Figure 3a). For two monocot species, *Triticum urartu* (wheat) and *Hordeum vulgare* (barley), we did not find any syntenic regions for any of their MADS-box genes with other plant genomes. This is likely due to the fragmented early-version genome assemblies (partially due to their large genome sizes and transposon expansions) in these two grasses. Using the organic layout function in Cytoscape (Shannon et al., 2003), we further depicted an undirected and unweighted (e.g. edge-length of no meaning) network with related gene clade names (Figure 3b). From this, we can then infer the number of syntelogs and relationships among syntelogs generated via polyploidy and tandem-duplication events. Below, we highlight three novel insights into the evolution of the MADS-box gene family based on our synteny network cluster analysis: (I) lineage-specific transpositions; (II) ancient tandem gene arrangements and; (III) deep conservation of specific clades across angiosperms.

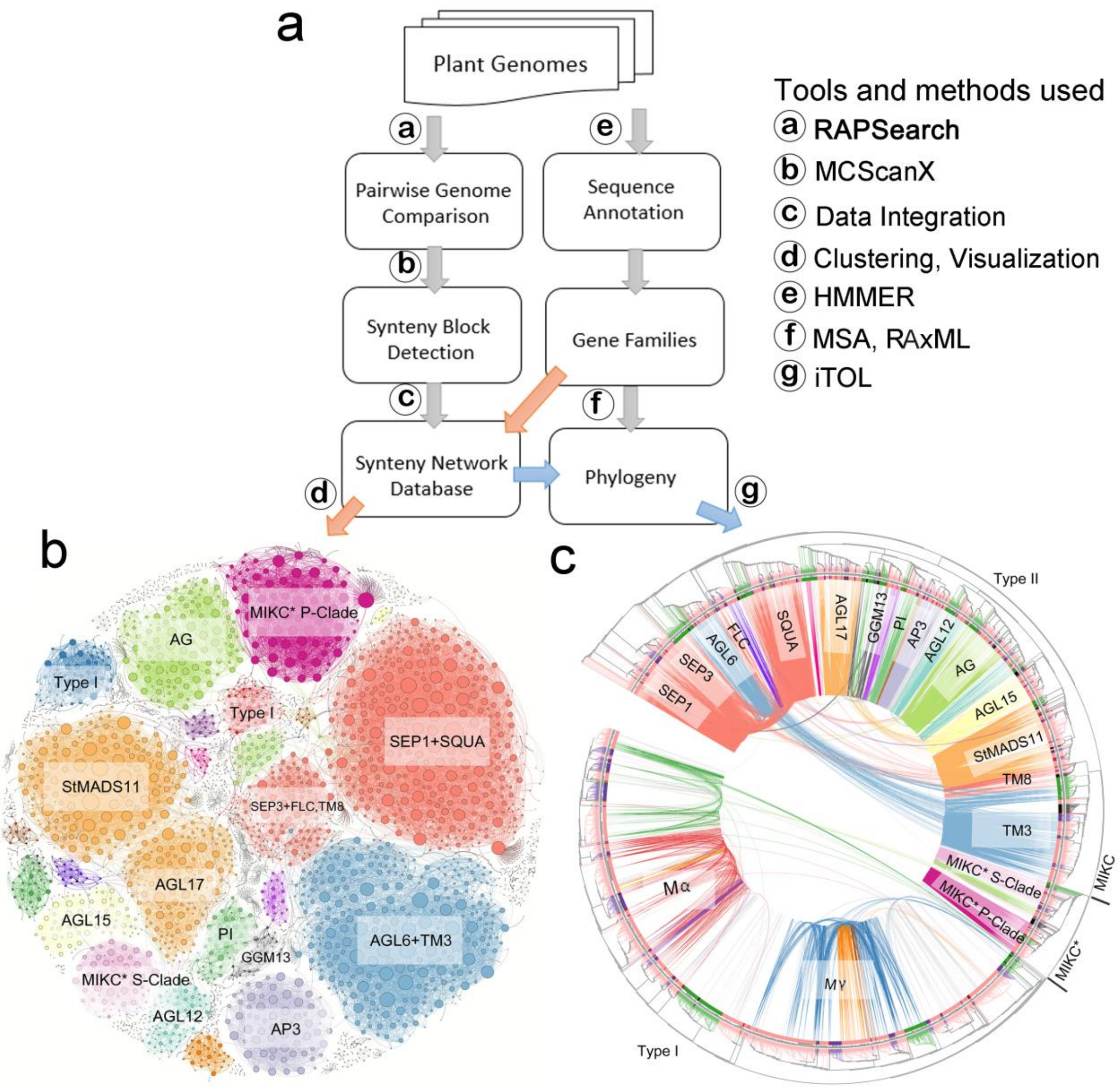


Figure 2 Workflow to create the Phylogenomic synteny network and example outputs for the global MADS-box gene family. (a) Workflow to create the phylogenomic synteny network. Annotated whole-genome sequences enter the pipeline and are used in two parallel modules. The left panel represents the analysis pipeline for pairwise genome comparisons and synteny calculations (synteny block detection), which creates the global syntenic network database. The right panel depicts the pipeline for a phylogenetic analysis including gene family identification and gene-tree construction. (b) Synteny network of the MADS-box gene family using all the detected syntenic relations in the synteny network database. Communities were rendered based on the clique percolation method at *k* = 3. The size of each node corresponds to the number of edges it has (node degree). Communities were labeled by the subfamilies/subfamily involved. (c) Maximum-likelihood gene tree for the MADS-box gene family and syntenic relationships between the genes. The subclades are indicated for the Type I, Type II, and MIKC- and MIKC*-Type II MADS-box genes on the tree. Terminal branch colors represent genes belonging to rosids (light pink), asterids (purple) and monocots (green). Genes belonging to

angiosperms in highly informative phylogenetic positions such as *Amborella trichopoda*, *Vitis vinifera*, *Beta vulgaris*, and *Nelumbo nucifera* are in red and genes of non-angiosperms belonging to *Chlamydomonas reinhardtii*, *Physcomitrella patens*, *Selaginella moellendorffii* and *Picea abies* are in black. Each connecting line located inside the inverted circular gene tree indicates a syntenic relationship between two MADS-box genes. The connecting lines are colored according to the discovered communities in panel b.

Section I: Lineage-specific synteny relationships

Important angiosperm families (such as Poaceae, Asteraceae, Fabaceae, Brassicaceae and Solanaceae) are readily identified by unique traits and floral characteristics. These major plant families are also characterized by having independent ancient polyploidy events at their origins (Soltis et al., 2009; Schranz et al., 2012; Tank et al., 2015). Morphological changes could thus be tied to these ancient polyploidy events or specific gene transposition events that place key-regulatory factors into new genomic contexts (Soltis et al., 2009; Freeling et al., 2012). Our synteny network approach can identify such lineage-specific transposition events for genes by clustering and phylogenetic profiling.

I.1 B-function (AP3 and PI) genes in the Brassicaceae and Cleomaceae families

The *AP3* and *PI* genes are important for petal and stamen specification (Jack et al., 1992; Goto and Meyerowitz, 1994; Jack et al., 1994; Zhang et al., 2013). In this study, we found that most *AP3* genes reside in a single cluster comprising homologs of both eudicot and monocot species, the basal angiosperm *A. trichopoda* and the basal eudicot *N. nucifera* (Figure 3, Cluster 9). However, the cluster lacks *AP3* homologs from the Brassicaceae family (Figure 3, Cluster 9). Instead, the *AP3* genes from the Brassicaceae form a separate cluster (Figure 3, Cluster 26) (except for *Aethionema arabicum*, where the *Aethionema AP3* gene was annotated on a scaffold lacking other genes (Gene ID: AA1026G00001, highlighted in Supplemental Data Set 1, sheet 1)).

A very similar picture emerges for the *PI* genes: the *PI* homologs from the analyzed six Brassicaceae species group together with a *PI* gene from *Tarenaya hassleriana* (a closely-related Cleomaceae species), while the *PI* homologs from most other species group with a second *PI* gene from *Tarenaya hassleriana* in another cluster (Figure 3, Cluster 24). To verify this pattern, we investigated the synteny relationships of the *PI* genes from grapevine (Vv18s0001g01760) and *A. thaliana* (AT5G20240) using the Genomicus parallel coordinate plot (Louis et al., 2012). Synteny was not detected with any Brassicaceae species when using the grape homolog of *PI* (Vv18s0001g01760) (Supplemental Figure 4a), while a unique synteny pattern is shared between the *A. thaliana* gene AT5G20240 and the Brassicaceae *PI* genes (Supplemental Figure 4b).

These two divergent synteny patterns suggest that in either cases (PI and AP3), a gene transposition, or genomic rearrangement event led to the unique genomic context seen for both genes in the Brassicaceae. Since one Cleomaceae *PI* gene belongs to the Brassicaceae *PI* cluster (Figure 3, Cluster 24) but the Brassicaceae *AP3* cluster does not contain any Cleomaceae *AP3* gene (Figure 3, Cluster 26), it is clear that *PI* transposed first and only later, and independently, did *AP3* transpose.

I.2 FLC-Like genes cluster in Brassicaceae

In *A. thaliana*, the *FLC* gene and its closely related MAF genes are floral repressors and major regulators of flowering time (Michaels and Amasino, 1999; Sheldon et al., 2000). We found a cluster comprising 21 syntelogs of *FLC* and the *MAF* genes across the six examined Brassicaceae species and one Cleomaceae species (*Tarenaya)* (Figure 3, Cluster 23).

This synteny cluster also contains one *FLC-like* gene from sugar beet. This sugar beet *FLC* homolog also shares synteny with a cluster comprising *StMADS11* (*SVP*-like) genes, which are found in an array of eudicot species (Figure 3b, Cluster 3; Supplemental Data Set 3). This sugar beet *FLC* gene thus connects the FLC/MAF genes of the Brassicales-lineage with the *StMADS11* genes of other eudicots. This highlights that likely a gene transposition or massive genome fractionation process has acted on the ancestral *FLC* gene in the Brassicales lineage after the split of the early branching Papaya, potentially near the time of the At-β WGD (Edger et al., 2015).

I.3 AGL17-Like genes cluster in Monocots

Also, the *AGL17*-like genes from six monocots specie (*Brachypodium distachyon, Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Setaria italica*, and *Elaeis guineensis*) form a distinct synteny cluster (Figure 3, Cluster 14, size 17). This may be to a specific transposition event and/or due to the ancient τ WGD shared by all monocot species (Jiao et al., 2014).

Section II: Inference of ancient tandem gene arrangements

Besides the distinctive lineage-specific clusters described above, larger clusters that comprise interconnected sub-clusters (with a force-directed or organic layout) can also be obtained when using the appropriate clustering methods (such as the *k*-clique percolation method that allows for community overlapping). As shown in Cluster 3 (Figure 1d), such clusters indicate long-conserved close genomic proximity of the genes involved (representing respective sub-clusters) and thus are of help to establish the trajectory of gene evolution.

II.1 Angiosperm-wide conserved SEP1-SQUA and SEP3-SFT tandems

The largest cluster (475 nodes) we identified comprises both the *AGL2* (*SEP*)-like and the *SFT*-like genes (Figure 3b, Cluster 1; Figure 4a). This cluster can be divided into

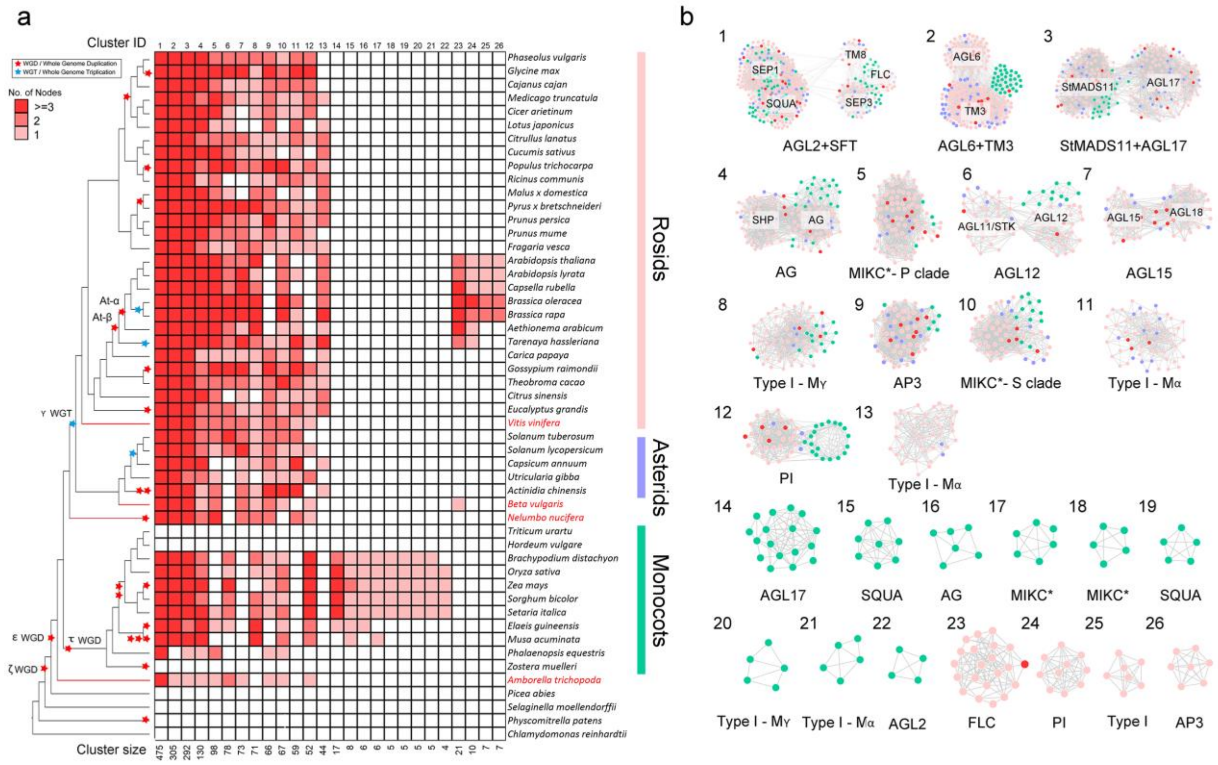


Figure 3 Phylogenetic profiling and detailed network representations for a set of selected important MADS-box synteny network clusters. (a) Phylogenetic profiling of twenty-six example clusters from all the ninety-five clusters at *k* = 3 (see Supplemental Figure 3). Red-colored cells depict the presence of at least one syntelog in the different species. The phylogenetic profiling approach identified large clusters containing genes from various subfamilies, such as "*AGL2+SQUA*" (Cluster 1) and "*AGL6+TM3*" (Cluster 2), or lineage-specific clusters such as of *PI* (Cluster 24) and *AP3* (Cluster 26) in the Brassicaceae, and "*AGL17*" (Cluster 14) and "*SQUA*" (Cluster 15) in monocots. Species names are shown on the right side. The basal rosid *Vitis vinifera*, the basal eudicots *Beta vulgaris* and *Nelumbo nucifera*, and the basal angiosperm *Amborella trichopoda* are highlighted in red. Red and blue stars on the tree indicate known WGD and WGT events, respectively, on the phylogenetic tree on the left side. The cluster ID and size are indicated at the top and bottom, respectively. The red color scale of the cell indicates the number of nodes (genes grouping in that cluster) found in a single species. (b) Clusters from (a), which can be divided into large-conserved clusters (Clusters 1-13) and lineage-specific clusters (Monocots (Cluster 14-22) and Brassicales (Cluster 23-26)). The node colors represent rosids (light pink), asterids (blue), monocots (green); the nodes belonging to *Amborella trichopoda*, *Vitis vinifera*, *Beta vulgaris*, and *Nelumbo nucifera* are shown in red.

two sub-groups: on the left are the *SEP1*-, and *SQUA*-like genes, while on the right are the *SEP3*-, *FLC*-, and *TM8*-like genes (Figure 4a). The *SEP1*-, and *SQUA*-like genes are highly interconnected between and within genomes (Figure 4a) with syntenic orthologs being present for both genes in a wide-range of angiosperm species including *A. trichopoda*, monocots and eudicots. As exemplified by Cluster 3 in the introduction (Figure 1d), *SEP1*- and *SQUA*-like genes are predominantly found in a tandem gene arrangement in most angiosperm species (Figures 4a and 4c, Supplemental Data Set 3) suggesting that this duplication occurred prior to or at the origin of the angiosperms.

For example, there is one *SEP1-SQUA* tandem gene arrangement in the *A. trichopoda* and three such tandem gene arrangements in the basal eudicot *V. vinifera* (Figures 4a and 4c, Supplemental Data Set 3), as a result of the Gamma hexaploidization (referred to as γ triplication) in eudicots.

On the right side of the network in Figure 4a, most eudicot and monocot *SEP3*-like genes group as a distinct cluster, which is relatively loosely connected to the nodes that represent the *FLC*-like genes and *TM8*-like genes (Figure 4a). Similar to the discovery of a *SEP1-SQUA* tandem, we also identified a *SEP3-FLC* tandem gene arrangement in 12 eudicots species (Figure 4c, Supplemental Data Set 3). This tandem arrangement was also found twice in monocots, namely *O. sativa* and *S. bicolor* (Figure 4c, Supplemental Data Set 3). However, the *SEP3-FLC* tandem gene arrangement is found less often than the *SEP1-SQUA* tandem gene arrangement. Besides this, we found that in *A. trichopoda* the *SEP3* and *TM8* homologs are also arranged in tandem (*SEP3-TM8*) (Figure 4c, Supplemental Data Set 3). None of the *FLC* homologs from Brassicaceae and Cleomaceae species are present in the angiosperm-*FLC* cluster in Figure 4a. As described in Section I, the Brassicales *FLC* syntelogs form an independent cluster (Figure 3b, Cluster 23).

II.2 Angiosperm-wide conserved AGL6-TM3 tandem

The second largest cluster identified in this study (*k* = 3, community size: 305) contains the *AGL6*- and *TM3* (*SOC1*)- like genes (Figure 3b Cluster 2, Figure 4b). Like the *SEP1-SQUA* and *SEP3-SFT* tandems in Figure 4a, we found that the *AGL6-TM3* tandem gene arrangement is widespread across angiosperms (Figures 4b and 4c, Supplemental Data Set 3). For example, there is one *AGL6-TM3* tandem gene pair in *A. trichopoda*, and two such tandems in *N. nucifera* likely due to the most recent WGD this species experienced (Ming et al., 2013; Wang et al., 2013) (Figure 4c, Supplemental Data Set 3). In *V. vinifera* we also found two *AGL6-TM3* tandems (Figure 4c, Supplemental Data Set 3) that likely originated from the γ triplication after which one tandem lost its *AGL6* locus. Like *V. vinifera*, *T. cacao* has not undergone any additional WGD after the γ triplication and in this genome two *AGL6-TM3* tandems also remain (Figure 4c, Supplemental Data Set 3).

Besides the prevalent *AGL6-TM3* tandem gene arrangement found in Figure 4b, we also found the tandem type of *TM3-TM3* in 10 species (seven eudicot species and three monocot species) (Figure 4c, Supplemental Data Set 3). Hence, the network has overall more *TM3*-like genes than *AGL6*-like genes (Figure 4b).

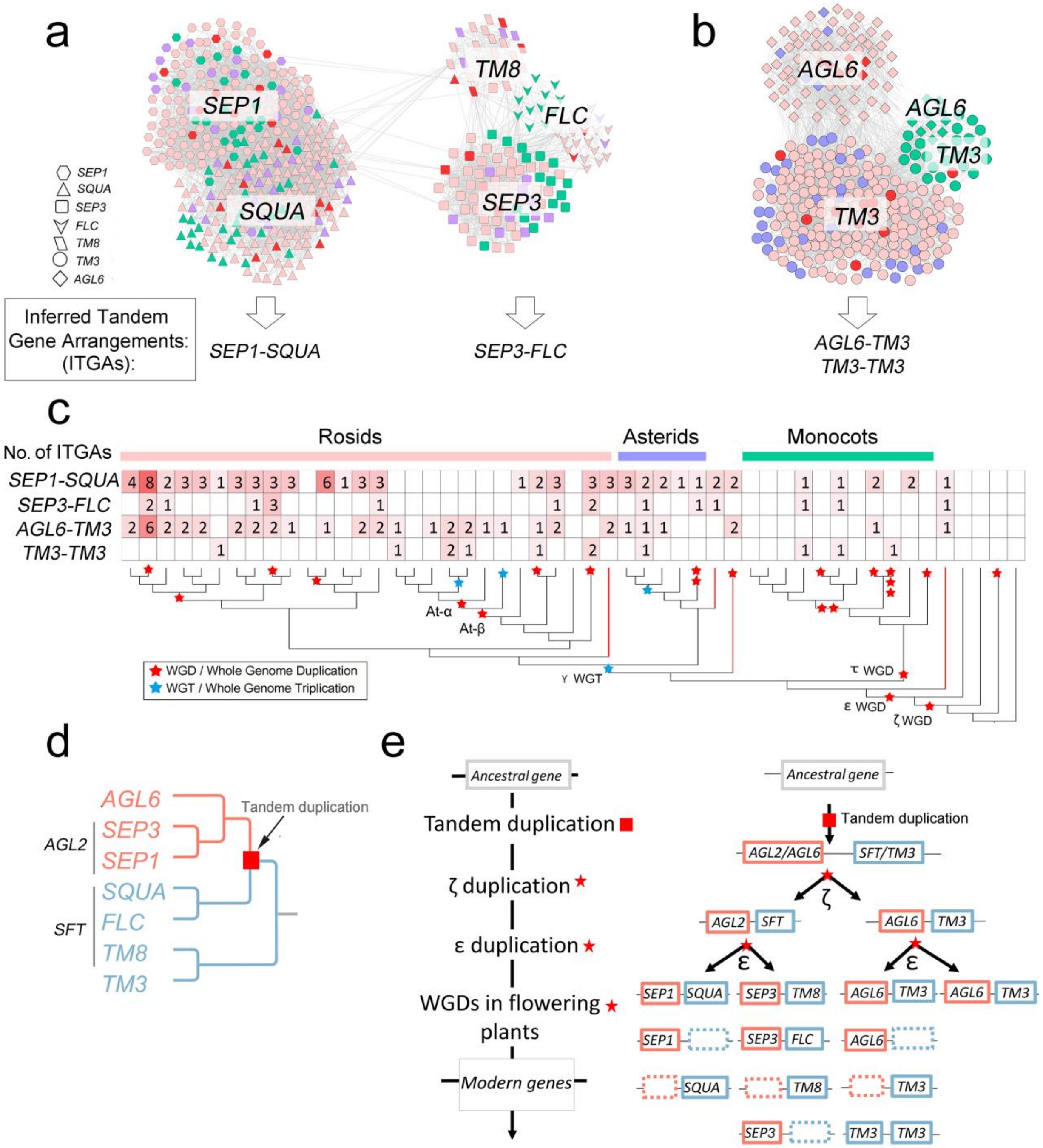


Figure 4 Evolutionary history of a major MIKC^c-type MADS-box gene super-clade is derived from Inferred Tandem Gene Arrangements (ITGAs). (a, b) Close-up of the networks for Cluster 1 (*AGL2*+*SFT*) and Cluster 2 (*AGL6*+*TM3*) from Figure 3, respectively. Node shapes represent different subfamilies. Node colors represent rosids (light pink), asterids (blue), monocots (green), and the nodes belonging to *Amborella trichopoda*, *Vitis vinifera*, *Beta vulgaris*, and *Nelumbo nucifera* are in red. The arrows below the clusters point to the Inferred Tandem Gene Arrangements (ITGAs) for the respective three clusters. (c) Summary of the occurrence of four ITGAs detected in the species analyzed. The two top rows give the number of *SEP1-SQUA* and *SEP3-FLC* tandems detected in Cluster 1 (panel a). While rows three and four give the number of *AGL6-TM3* and *TM3-TM3* tandems detected in Cluster 2 (panel b). The species tree is a

simplified version (without species names) of the tree used in Figure 3a and Supplemental Figure 1. (d) Mapping of the tandem duplication event onto the inferred consensus gene tree according to the three ITGAs in panel c. (e) Proposed evolutionary scenario for the origin of the MADS-box gene family clade in panel d. One single MIKC^c-type MADS-box gene initially got tandem duplicated, then this tandem arrangement underwent multiple rounds of polyploidy events, and eventually evolved to the superclade depicted in panel d.

Section III: Synteny relations across angiosperms for overlooked MADS-box gene clades

In addition to the many functionally characterized MADS-box genes, a large portion of the gene family members are poorly or not functionally characterized, such as MIKC*-type and Type I genes of the MADS-box gene family. However, the synteny network can provide evidence of synteny conservation for these genes over evolutionary time and thus suggest important conserved gene functions.

III.1 MIKC-Type genes*

The MIKC*-type genes form a monophyletic clade within the MADS-box genes (Alvarez-Buylla et al., 2000a; Henschel et al., 2002) (Figure 1b), with several of them being reported to play a major role in pollen development (Verelst et al., 2007a; Verelst et al., 2007b; Adamczyk and Fernandez, 2009). Using our synteny network analysis, we found two networks that are highly connected and contain (i) the angiosperm *AGL30*-, *AGL65*-, and *AGL94*-like genes (MIKC*-P clade) (Figure 3, Cluster 5) and (ii) the *AGL66*-, *AGL67*-, and *AGL104*-like genes (MIKC*-S clade) (Figure 3, Cluster 10), respectively.

Both clusters encompass eudicots and monocot species, as well as *A. trichopoda*. However, the MIKC*-S cluster appears to have expanded in monocots, while homologs of *N. nucifera* are absent in this cluster (Figure 3, Cluster 10). This means that both two MIKC* clades are broadly conserved across angiosperms.

Interestingly, MIKC* protein complexes play an essential role in late pollen development in *A. thaliana* and the formation of this protein complex requires MIKC* proteins from both clades. For example, the AGL30 and/or AGL65 proteins from the P clade form heterodimers with AGL104 or AGL66, which both group with the S clade (Verelst et al., 2007a; Verelst et al., 2007b). This suggests that these two clades (gene clusters) have been functionally retained during angiosperm evolution.

III.2 Type I MADS-box genes

Type I MADS-box genes show a higher rate of gene birth-and-death, often due to gene duplication-transposition, than Type II genes (Nam et al., 2004) (Freeling et al., 2008; Wang et al., 2016b). Also, the function of the different Type I genes are generally poorly characterized. However, several Type I genes have been reported to play a role in female gametogenesis, embryogenesis, and seed development (Portereiko et al., 2006; Bemer et al., 2010; Masiero et al., 2011).

With our approach, we found two distinct clusters that contain Type I MADS-box genes (Figure 3, Clusters 8 and 11). For example, the *PHERES1* (*PHE1*/*AGL37*) genes, which are regulated by genomic imprinting (Kohler et al., 2003), are in the same synteny network as *PHERES2* (*PHE2*/*AGL38*), *AGL35*- and *AGL36*-like genes, which all belong to the Mγ clade of the Type I MADS-box genes (Figure 3, Cluster 8). Likewise, we found one cluster that contains genes from the Mα clade (Figure 3, Cluster 11).

III.3 StMADS11 (SVP-Like) genes

In *A. thaliana*, the *StMADS11* gene clade is composed of two genes called *SVP* (*AGL22*) and *AGL24*. These two genes regulate the transition to flowering in *A. thaliana* (Hartmann et al., 2000; Michaels et al., 2003).

We found that the *SVP*- and *AGL24*-like genes are contained in one cluster for many of the angiosperms analyzed, which indicates that synteny has been retained for *SVP/AFL24* since the last common ancestor of angiosperms (Figure 3b, Cluster 3). It is worth noting that the *AGL17*-like genes from *A. trichopoda, N. nucifera* and most eudicot species form a cluster that is moderately connected to the cluster of *StMADS11*-like genes (Figure 3b, Cluster 3).

Discussion

Our phylogenomic synteny network analysis provides a novel approach to identify and visualize the relationships of genes of a targeted gene family across a broad range of species (Zhao and Schranz, 2017), which can be used to address fundamental questions on the origin of novel gene functions leading to morphological changes and adaptations. We have provided several new insights into the evolution of the MADS-box gene family from our synteny-based network analyses. These insights, in turn, generate new testable hypotheses on how the genomic context of a gene may (or may not) effect changes in its expression pattern, co-expression with other genes, epigenetic regulation and ultimately the evolution of plant phenotypes. Some possible hypotheses are discussed below. But, first we make a few comments regarding our methodology.

Factors affecting synteny network analysis

We have presented a methodological roadmap to construct synteny networks and an analysis pipeline, which can now be applied to any gene family across any set of genomes. The power of network analysis is the ability to organize large datasets and provide extrapolation and visualization beyond pairwise contrasts. As more plant species genomes are completed, particularly from under-represented lineages (such as non-angiosperm species), more robust network inferences can be made. Our network approach depends, however, on the quality of genome assemblies and their gene annotations. Genome collinearity is *de facto* more disrupted and difficult to detect in highly fragmented assemblies. Advances in genome sequencing and assembly means that chromosome level assemblies will be standard in the near future. With these advances, our network approach for synteny comparison will greatly benefit and improve.

The clustering methods used are pivotal for the interpretation of complex synteny networks, as it determines the size and structure of identified clusters. For example, when instead of our *k*-clique percolation method (at $k = 3$), other methods are used like *k*-core decomposition (Alvarez-Hamelin et al., 2006), MCL (Enright et al., 2002), infomap (Rosvall and Bergstrom, 2008), or CNM (Clauset et al., 2004), we would likely have obtained slightly different cluster topologies. Depending on the goals and objectives of a study, the appropriate clustering method should be established.

Lineage-specific genomic context of MADS-box genes: potential significant biological implications

In the model plant *A. thaliana*, the B-class AP3 and PI proteins form heterodimers and bind to the CArG-box *cis*-regulatory elements in promoters (Riechmann et al., 1996; Yang et al., 2003). Heterodimerization and/or homodimerization have contributed to the evolution of the highly diverse flower morphologies in angiosperms (Lee and Irish, 2011; Melzer et al., 2014; Bartlett et al., 2016). Brassicaceae species have rather uniform, or canalized, flowers (typical cross arrangement of the four petals). However, in its closest sister family Cleomaceae, which diverged from each other ~ 38 million years ago (MYA) (Schranz and Mitchell-Olds, 2006; Couvreur et al., 2010), more diverse floral morphologies are observed (Patchell et al., 2011). In this study, we found unique synteny patterns for the *T. hassleriana* B genes, which is consistent with previous findings (Cheng et al., 2013). One *T. hassleriana PI* gene resides in the cluster shared with most other eudicots and monocot species, while the other *T. hassleriana PI* gene sits in a cluster mostly composed of Brassicaceae species (Figure 3, Cluster 24). In Brassicaceae, we find the *PI* genes only in the new derived syntenic position. Furthermore, only in Brassicaceae we also find a unique syntenic position for the *AP3* genes (Figure 3, Cluster 26). *SEP*- and *SQUA*-like genes are also involved in petal formation according to the ABC(DE) flowering model (Figure 1a). Moreover, Brassicaceae (and Cleomaceae) species are absent from these *AGL2*-*SFT* type of tandems in comparison to other lineages (Figure 4c, Supplemental Data Set 3). It is unclear why the *PI*, *AP3* and *SEP3* genes are transposed in the Brassicaceae in comparison to other angiosperms. Potentially higher level inter- and intra-chromosomal chromatin interactions between loci, or new *cis*-regulatory elements, are required for crucifer B-specific gene expression patterns. It will be important to test such hypotheses and if potentially the derived genomic contexts of these genes have contributed to the canalization of the crucifer floral form.

FLC-like genes in the Brassicaceae and Cleomaceae are also in a derived genomic context compared to other angiosperms (Figure 3, Cluster 23). The vernalization process (prolonged cold exposure) is essential for many plants to initiate flowering. In *A. thaliana* and other crucifers, this process is mediated by cold-induced epigenetic repression of *FLC* genes, namely histone methylation (Bastow et al., 2004), chromatin structure modification with chromatin remodeling protein complexes (Kim and Sung, 2013), and the expression of long non-coding RNAs (Csorba et al., 2014). Genes flanking *FLC* are epigenetically coordinately regulated (Finnegan et al., 2004). Potentially the evolution of cold-specific epigenetic regulation was facilitated by the new genomic context of *FLC-like* genes in the Brassicales. It will be important to establish the patterns of epigenetic regulation of *FLC*-like genes outside of the Brassicales and which aspects are ancestral and which are derived.

A gene transposition event, likely after the split of monocot and eudicot species, has given rise to the specific synteny of the monocot *AGL17*-like genes found in this study (Figure 3, Cluster 14). In rice, the *AGL17/ ANR1*-like genes are preferentially expressed in root and responsive to various hormone treatments (Puig et al., 2013), and nutrient supply (Yu et al., 2014). Moreover, in rice the *AGL17* clade genes are specific targets of the miR444 miRNA family, and this miRNA family is specific to monocots (Sunkar et al., 2005; Wu et al., 2009; Li et al., 2010). MiR444 regulates nutrition signaling and root architecture in a monocot specific way (Yan et al., 2014), and together with its *AGL17* targets, they also play direct control in the rice antiviral pathway (Wang et al., 2016a). The synteny disruption of monocot *AGL17*-like genes, compared to eudicot species observed in this study, may be correlated with the origin of the miRNA-dependent regulation. Understanding this could be important for understanding the evolution of root architecture and responses to nutrient supplies, such as nitrogen.

Ancient tandems of MADS-box genes

The ancient *SEP1-SQUA* tandem gene arrangement, as revealed by our angiosperm-wide synteny network analysis (Figure 4a), is in agreement with other studies where the *SEP1-SQUA* tandem gene arrangement was found in eudicots (Ruelens et al., 2013). Another study also noted that most *AP1*-like genes (a subclade of the *SQUA*-like genes) and *SEP1*-like genes were tightly linked as genomic neighbors since the split of the basal eudicots (Sun et al., 2014). Another example is the ancient tandem arrangement of *SEP3-TM8*. *TM8* was first identified from *S. lycopersicum* (Pnueli et al., 1991), and this clade of genes has been reported to have undergone independent gene loss in different lineages based on phylogenic analyses (Becker and Theissen, 2003; Gramzow and Theissen, 2013). According to the consensus phylogeny based on studies by others (Figure 1b), the *TM8*-like genes are closely related to *TM3*-like genes and they both appear to share a common origin with the *AGL6*-, *AGL2*-, *SQUA*-, and *FLC*-like genes.

Based on synteny analyses, it was previously suggested that these *SEP3*- and *FLC*-like genes originated from an ancient tandem gene duplication (Ruelens et al., 2013). Our synteny analysis reveals a more broadly conserved, and thus potentially more ancient, tandem gene duplication that involves the last common ancestor of all *SEP3*- and *TM8*-like genes. Considering that *TM8*-like genes were already present in the last common ancestor of extant seed plants (Gramzow et al., 2014), it is likely that the *SEP3-TM8* tandem is more ancestral than the *SEP3-FLC* tandem (e.g. as defined by functions). Hence, the *FLC*-like genes could be derived from a *TM8* homolog in an ancestral plant species. According to the network structure and gene copy number of the *SEP3*-, *FLC*- and *TM8*-like gene clusters, we find that after the split of *A. trichopoda* from other angiosperms the *SEP3*- and *TM8*-like genes generally do not appear as a tandem gene pair within one species and *TM8*-like homologs tend to be lost from the tandem. This means that the *SEP3-TM8*/*FLC* tandem gene pair is more variable than the *SEP1-SQUA* tandem gene pair. In this study, both the *SEP1-SQUA* and *SEP3-TM8* tandem gene pair were found in *A. trichopoda* (Figure 4c, Supplemental Data Set 3). Hence, the duplication that led to these two tandems may be the ε WGD event, derived from one ancestral tandem gene pair of *AGL2-SFT* (Figures 4d and 4e) in a common ancestor of the angiosperms (Jiao et al., 2011; Li et al., 2015).

It is generally thought that the *AGL6*- and *AGL2* (*SEP*)-like genes are closely related sub-families (Figure 1b). It has been hypothesized that the combined ancestral gene of the *AGL6*- and *AGL2*-like genes was duplicated in a common ancestor of the seed plants (Spermatophytes) (Zahn et al., 2005; Kim et al., 2013), probably as a result of the ζ WGD (Jiao et al., 2011; Li et al., 2015). By interpreting our synteny networks, we found strong evidence of *SEP1*-*SQUA*, *SEP3*-*SFT*, and *AGL6*-*TM3* tandems (Figures 4a, 4b, and 4c), and evidence of monocot *TM3*-like genes connected to *SEP3*-, *SQUA*-, and *TM8*-like genes (Figure 4a). This enabled us to deduce the deep genealogy and to propose an evolutionary diagram that depicts how one ancestral locus that predates the last common ancestor of all seed plants has given rise to a large MADS-box gene clade with many subfamilies in angiosperms, which includes the *AGL2*-, *AGL6*-, *SQUA*-, *TM3*-, *TM8*-, and *FLC*-like gene clades (Figures 4c-4e). It can be inferred that in the last common ancestor of seed plants a gene tandem was already present that corresponds with the current *AGL2*/*AGL6*-*SFT*/*TM3* tandem gene arrangement (Figures 4c-4e). The ζ WGD that occurred shortly before the radiation of the extant seed plants (Jiao et al., 2011) is likely causal to the duplication of this original tandem gene pair, after which the *AGL2*- and *AGL6*-like genes diverged, as well did the *SFT*- and *TM3*-like genes (Figure 4e). As described above, a subsequent more recent WGD (the ε event), which occurred prior to the diversification of the extant angiosperms (Jiao et al., 2011; Li et al., 2015), allowed then the emergence of the *SEP1*- and *SEP3*-like genes from the ancestral *AGL2* locus, as well as the *SQUA*-, *TM8*-, and *FLC*-like genes from the ancestral *SFT* gene. During that same period only one copy of the *AGL6*-*TM3* tandem was retained from the ε WGD (Figure 4e). Altogether, this model hypothesized how one single MIKC^c-type MADS-box gene gives birth to a whole super-clade of genes composed of *AGL2* (*SEP*)-, *AGL6*-, *SFT*- (i.e., *SQUA*-, *FLC*-, and *TM8*-), and *TM3* (*SOC1*-like) genes/subfamilies due to a tandem duplication and subsequent WGDs.

Plant regulatory genes, such as MADS-box transcription factors, are generally not thought to be organized in co-expressed gene clusters like animal Hox or Para-Hox genes that do show coordinated gene expression (Lewis, 1978; Krumlauf, 1994; Ferrier and Holland, 2001). This could be due to the analysis techniques of plants employed to date, namely phylogenetic analyses and pair-wise synteny analyses, where ancient WGDs can dramatically complicate analyses. More recently, it has become apparent that many plant biosynthetic genes are organized into physical clusters that are co-regulated and co-expressed (Boutanaev et al., 2015; Nützmann et al., 2016; Yu et al., 2016a). Often, these biosynthetic clusters are lineage-specific and are not just due to tandem-duplication of a single ancestral gene.

With our approach, we have found several examples of highly conserved MADS-box collinearity and of lineage-specific transpositions. MADS protein-protein interactions or gene co-expression data are not obviously consistent with the parallel co-regulation model like for animal Hox-genes or plant biosynthetic gene clusters. However, potentially high-level chromatin-interacting domains within and between clusters that dictate their relative positions within the nucleus need to be tested for potential co-regulatory interactions. Although we describe several interesting patterns of evolution of the MADS-box genes, this is just an example of one gene family across fifty-one plant species. Thus, we are providing just a proof of concept and a view on the tip of a new genomic iceberg. Our approach is suited for analyzing the positional context of all genes across all completed genomes to examine patterns of genomic conservation and divergence.

Methods

Plant genomes analyzed

In total, fifty-one plant genomes were included in our analysis (Supplemental Figure 1, Supplemental Table 1 for detailed information), including thirty rosids, five asterids, *Beta vulgaris* (non-rosid non-asterid), eleven monocots, the early diverging angiosperm (*Amborella trichopoda*), and a single genome for gymnosperms (*Picea abies*), club moss (*Selaginella moellendorffii*), moss (*Physcomitrella patens*), and green alga (*Chlamydomonas reinhardtii*). For each genome, all annotated protein sequences (primary transcript only) in a FASTA file and a BED/GFF file indicating gene positions are needed.

Pairwise whole genome comparisons

Reciprocal all-against-all comparisons between pairwise genomes as well as intra-species comparisons are needed for synteny block detections. Thus for fifty-one species in this study, we need P (51, 2) + 51 = 2, 601 times whole-genome protein comparisons. RAPSearch2 (BLAST-like program, but much more efficient) was used for this task (Zhao et al., 2012).

Syntenic block calculation

MCScanX (Tang et al., 2008b; Wang et al., 2012) was used to compute genomic collinearity between all pairwise genome combinations using default parameters (minimum match size for a collinear block = 5 genes, max gaps allowed = 25 genes). The output files from all the intra- and inter- species comparisons were integrated into a single file named “Total_Synteny_Blocks”, including the headers “Block_Index”, “Locus_1”, “Locus_2”, and “Block_Score”, which served as the database file.

Synteny network for the MADS-box gene family

Candidate MADS-box genes were initially identified using HMMER3.0 with default settings (domain signature PF00319) (Finn et al., 2011) for each of the 51 genomes (Supplemental Data Set 1, sheet 1). Then this gene list containing all candidate MADS-box genes was queried against the “Total_Synteny_Blocks” file. Rows containing at least one MADS-box gene were retrieved into a new file termed “Syntenic_Blocks_MADS-box genes” (Supplemental Data Set 1, sheet 2). This file was then the final synteny network for the MADS-box genes, the network was imported and visualized in Cytoscape version 3.3.0 (Shannon et al., 2003) and Gephi 0.9.1 (Bastian et al., 2009).

Sequences were labeled based on the *A. thaliana* MADS-box genes plus three representative MADS-box genes that are not represented in *A. thaliana* (*TM8*-gene (GenBank Accession No. NP_001234105) from *Solanum lycopersicum*, *OsMADS32* gene (GenBank Accession No. XP_015642650) from rice, and *TM6* (GenBank Accession No. AAS46017) from *Petunia hybrida*) (Lee et al., 2003; Blanc and Wolfe, 2004; Daminato et al., 2014), using BLASTP (Altschul et al., 1990).

Network clustering

Clique percolation as implemented in CFinder (Derenyi et al., 2005; Palla et al., 2005; Fortunato, 2010) was used to locate all possible *k*-clique-communities for the MADS-box gene synteny network to identify communities (clusters of gene nodes). Increasing *k* values make the communities smaller and more disintegrated but also at the same time more connected.

Phylogenetic profiling of clustered communities

Communities (synteny clusters) derived from a certain *k* value were extracted and the node (i.e. gene) composition of each community was then mapped to the phylogenetic tree with 51 species (Smith et al., 2011). Presence (red) or absence (white) of homologs in a cluster was depicted for the different species in the phylogenetic tree, thus creating a phylogenetic profile of a synteny cluster (Supplemental Figure 3). Each column in the illustration represents one community (one synteny cluster), which is labeled at top of the x-axis based on its MADS-box name/annotation. Through such clustering and phylogenetic profiling steps, representative communities for the Type II (MIKC^{C}- and MIKC*- type) and Type I MADS box clades were found and then further analyzed.

Phylogenetic distance and tree construction

Amino acid sequences for the candidate MADS-box genes, both the genes represented in the synteny networks and the genes missing from the networks, were aligned using HmmerAlign (Kristensen et al., 2011). The alignment was then transferred into codon alignment using Pal2nal (Suyama et al., 2006). A phylogenetic tree was computed using RAxML (Stamatakis, 2014) with the GTRCAT (bootstrap = 100). The phylogenetic tree was annotated and depicted using iTOL v3 (Letunic and Bork, 2016).

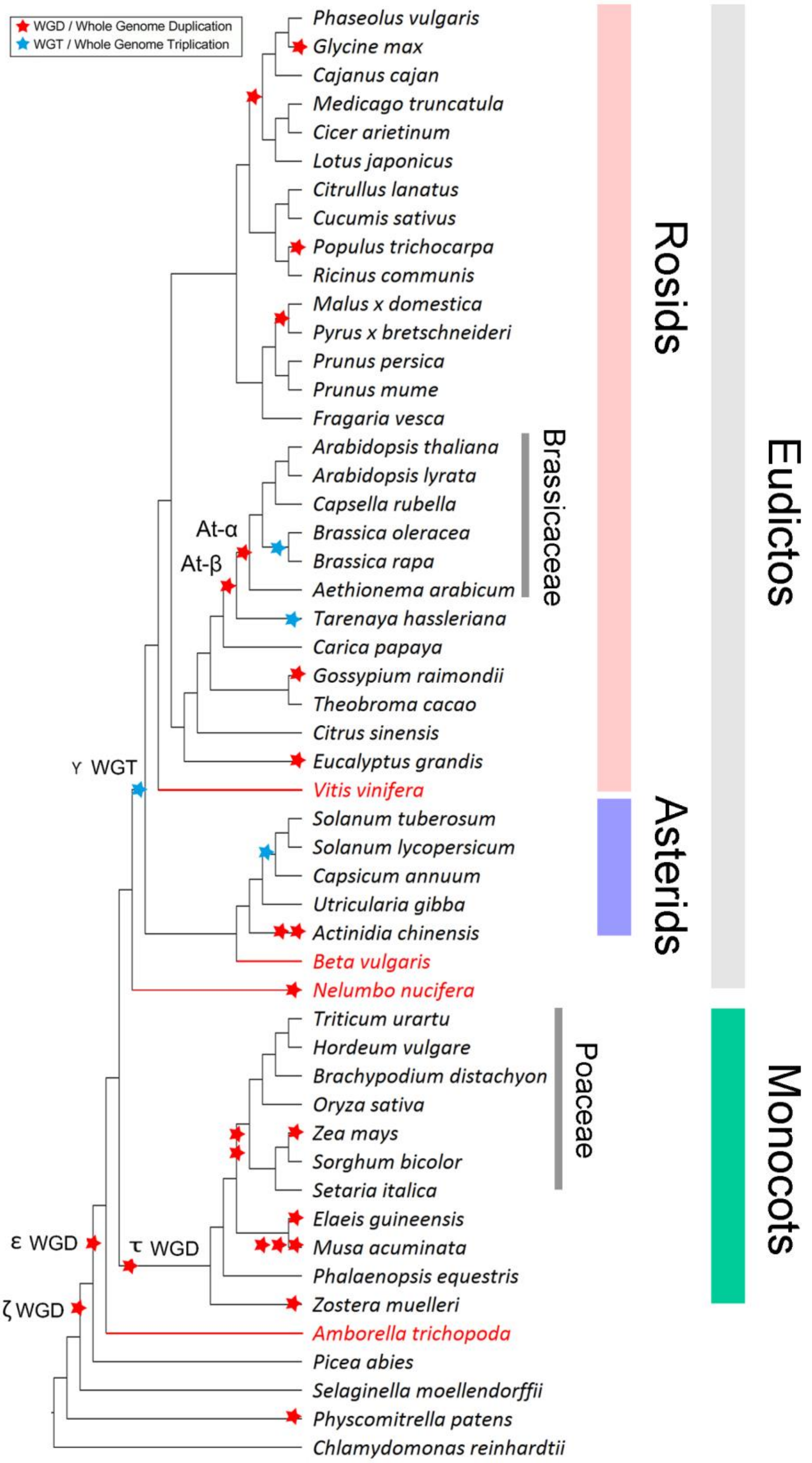
A script performing the above “Pairwise Whole Genome Comparisons” and “Syntenic Block Calculation” steps, and additional information about the method used in this work can be found at GitHub (https://github.com/zhaotao1987/SynNet-Pipeline).

Accession Numbers

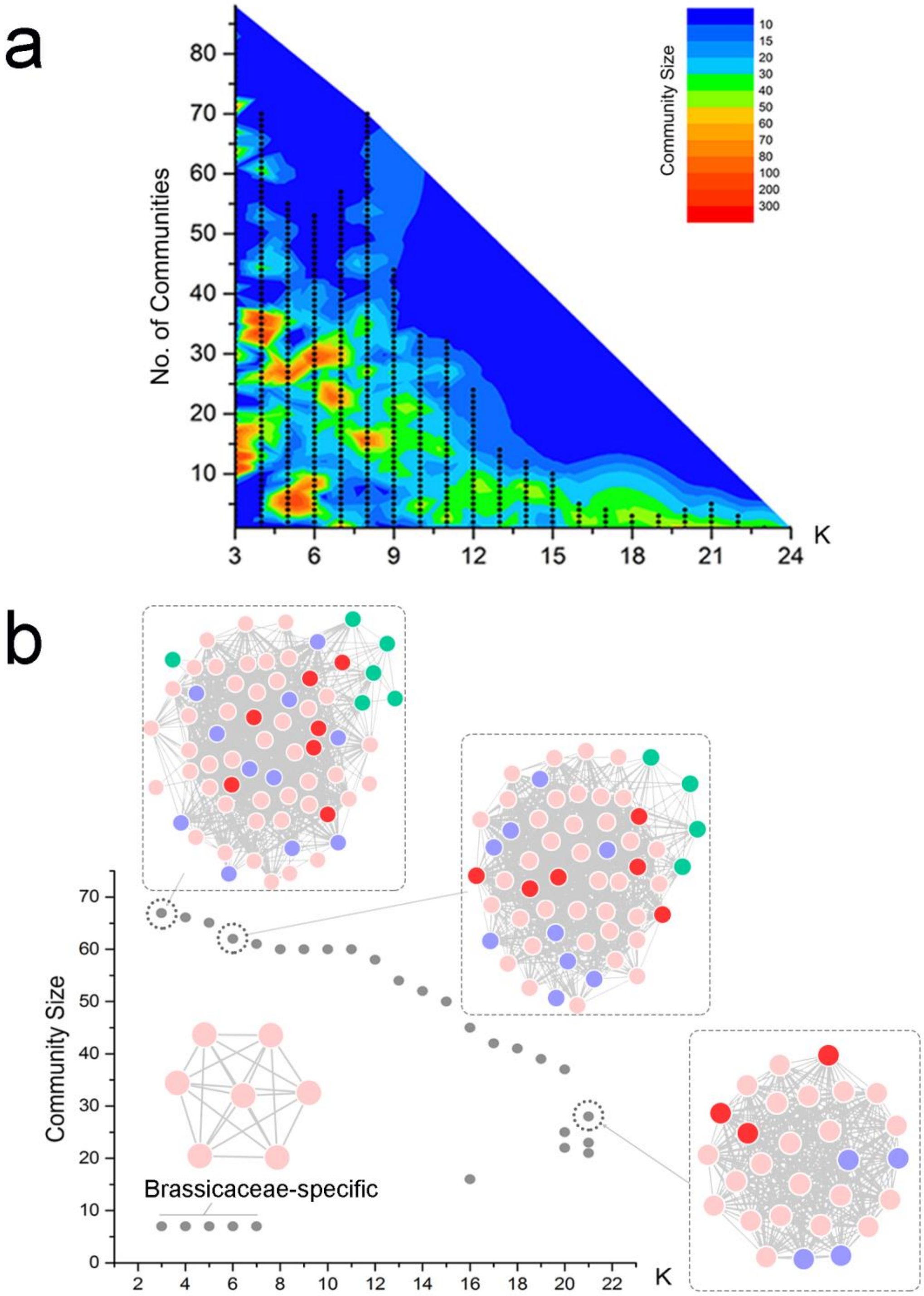
Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the following accession numbers:

Plant genomes used in this analysis are provide in Supplemental Table 1.

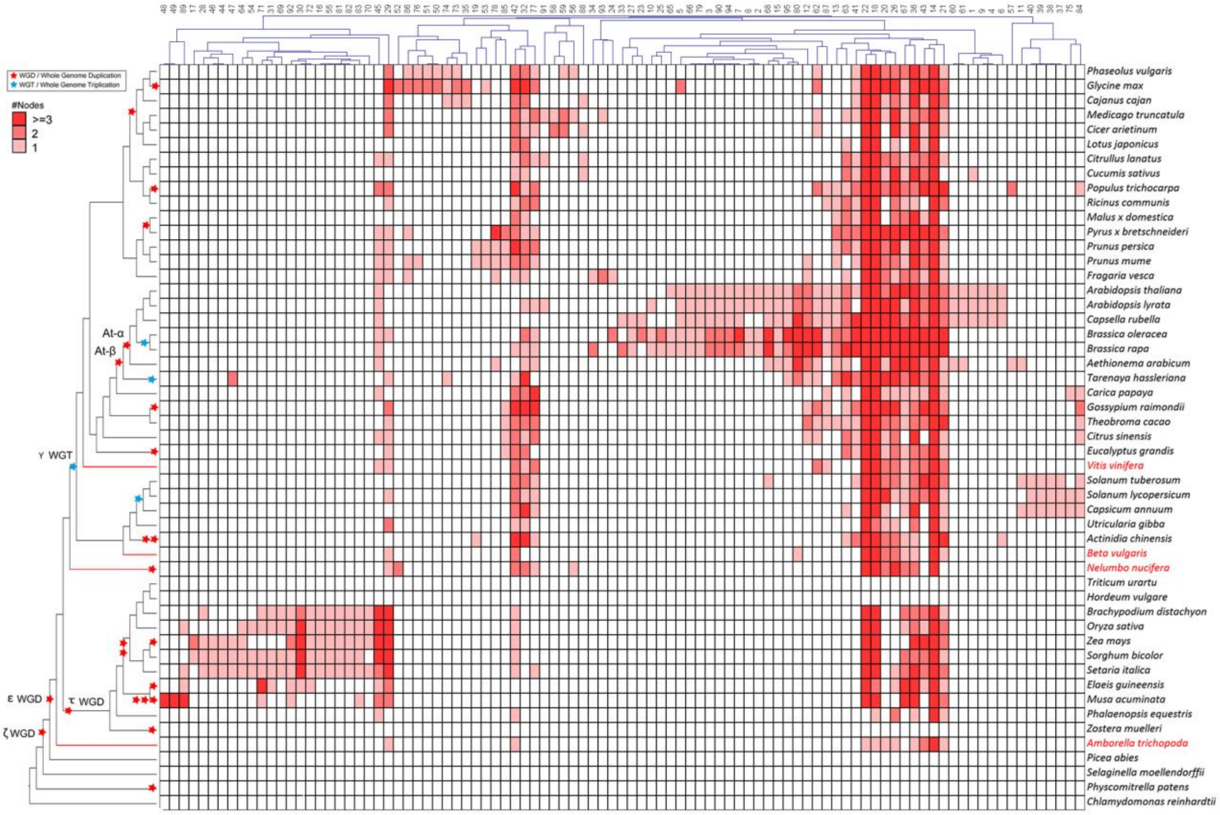
Supplemental Data



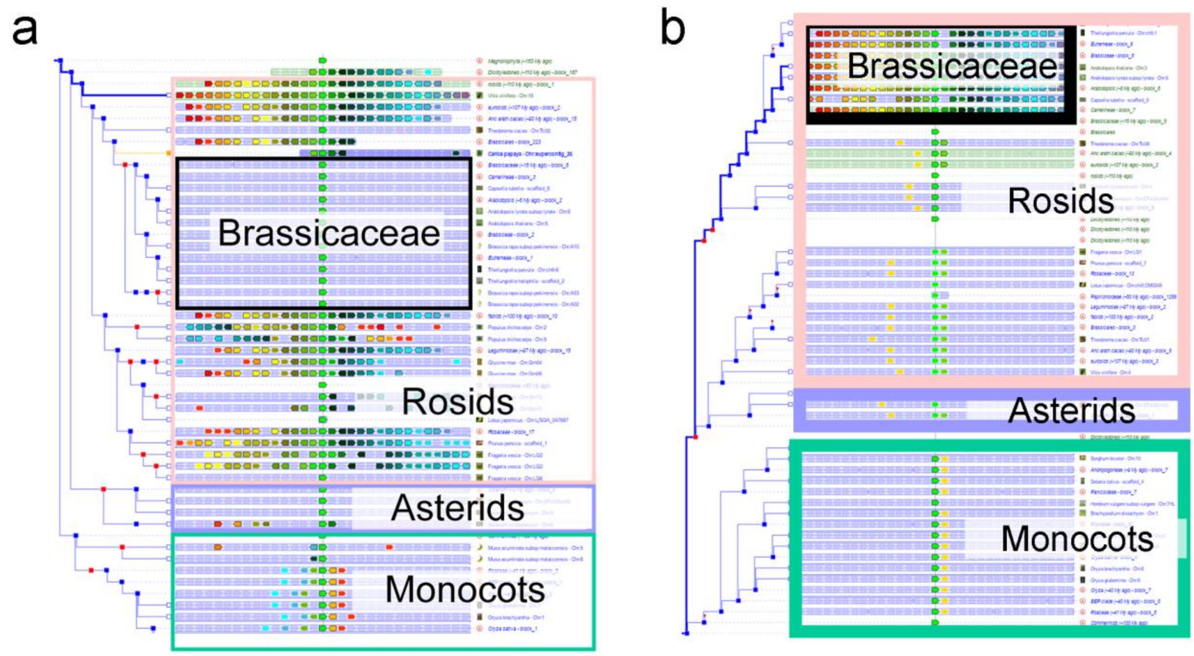
Supplemental Figure 1 Species used in this study. Known Whole Genome Duplications (WGD) and Whole Genome Triplications (WGT) events are indicated on the phylogenetic tree of the 51 analyzed genomes as red and blue stars, respectively. Clades of the species belong to “Brassicaceae”, “Poaceae”, “Rosids”, “Asterids”, “Eudicots”, and “Monocots” are indicated on the right side. The basal rosid *Vitis vinifera*, the basal eudicots *Beta vulgaris* and *Nelumbo nucifera*, and the basal angiosperm *Amborella trichopoda* are highlighted red.



Supplemental Figure 2 *k*-clique percolation of the synteny network for MADS-box genes. (a) Ternary contour plot of *k* (x-axis), the number of *k*-clique communities detected (y-axis, indicated by the number of black dots), and the community sizes detected (color range). At lower values of *k* (less connectivity required between nodes) there many more communities detected than at higher values of *k* (more connectivity between nodes). At low k values, synteny between just a few species can be found, however, at higher k values only highly-interconnected networks with many nodes (for example across angiosperms) are identified. (b) An example of how the community size of the cluster declines for the *AP3* MADS-box gene cluster as the *k*-cliques value increases. At lower values of *k* the number of inter-links can be lower, thus some less connected nodes are included in the network (such as the green labeled monocots). There are only six Brassicaceae genes included in our analysis, thus the network disappears at *k*=7. At the higher values of *k* represent more stringent communities of densely connected nodes with a high density of inter-links, for example eudicot AP3 genes, still form a community at *k*=21.



Supplemental Figure 3 Phylogenetic profiling for all the communities for *k*-clique = 3. Species names are shown on the right side. The basal rosid *Vitis vinifera*, the basal eudicots *Beta vulgaris* and *Nelumbo nucifera*, and the basal angiosperm *Amborella trichopoda* are highlighted in red. The red and blue stars on the species tree depict known WGD and WGT events, respectively. The number of nodes in a cell is reflected by the color scale. The columns are hierarchically clustered by the "Spearman Rank Correlation" method.



Supplemental Figure 4 Parallel coordinate synteny plots of PI derived from Genomicus (a) Synteny relationship of the grape homolog of PI (Vv18s0001g01760) across multiple lineages. (b) Synteny relationship of the A. thaliana PI gene (AT5G20240) across multiple lineages. Both images are derived from Genomicus: http://www.genomicus.biologie.ens.fr (Louis et al. 2012).

The following supplemental data can be accessed online: http://www.plantcell.org/content/early/2017/06/05/tpc.17.00312/tab-figures-data

Supplemental Table 1. Plant Genomes Used in this Analysis

Supplemental Data Set 1. Candidate MADS-box Genes (sheet1) and Synteny Network for MADS-box Genes (sheet2)

Supplemental Data Set 2. Node List and Edge List of the Communities at *k* = 3

Supplemental Data Set 3. Detail Information for the Inferred Tandem Gene Arrangements (ITGAs)

Acknowledgements

T.Z. was supported by the China Scholarship Council (CSC), H.v.d.B and M.E.S by a Netherlands Scientific Organization (NWO) Vernieuwingsimpuls Vidi grant (numbers 864.10.004 and 864.10.001, respectively). S.d.B received a NWO Experimental Plant Science (EPS) Graduate School “Master Talent” fellowship. We thank the three anonymous reviewers for their helpful comments and suggestions for improving the manuscript.

Author contributions

M.E.S and H.v.d.B designed the research. T.Z. performed the analysis. R.H., and S.d.B analyzed data. T.Z. and M.E.S wrote the article. G.C.A gave suggestions for the draft article. All authors discussed the result and commented on the article and approved of its final version for submission.

Chapter 4

Beyond the Sequence: Functional Innovations in Plant LATE EMBRYOGENESIS ABUNDANT Proteins Revealed by Synteny Network Analysis

Mariana Aline Silva Artur[1†,] Tao Zhao[2†], Wilco Ligterink[1], M. Eric Schranz[2], Henk W. M. Hilhorst[1*]

[1]Laboratory of Plant Physiology, Wageningen University, Droevendaalsesteeg 1, 6708PB, Wageningen, The Netherlands.
[2]Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB, Wageningen, The Netherlands.

[†]These authors contributed equally to this work.
*Correspondence and requests for materials should be addressed to H.W. M. H. (henk.hilhorst@wur.nl).

Genome Biology and Evolution, under review

Abstract

Late Embryogenesis Abundant proteins (LEAs) include eight multi-gene families expressed in response to water loss during seed maturation and in vegetative tissues of desiccation tolerant species. Despite their importance for water stress adaptation, a comprehensive understanding of LEAs evolution in plants is still elusive. We performed a phylogenomic synteny network analysis of the eight LEA gene families across 60 complete plant genomes, and investigated their distribution and diversification. Our results indicated that plant LEA families have distinct origins, and most of them show synteny conservation in angiosperms. Independent evolutionary patterns, such as ancestral diversification, recurrent tandem-duplications, and dynamic synteny evolution, contributed to sequence diversification and functional innovations. For example, ancestral synteny diversification in the Dehydrin family resulted in distinct evolution of amino acid sequences, biochemical properties, and gene expression patterns. We also identified the origin of a novel Dehydrin motif that may be specifically involved with abiotic stress tolerance. Together, our results show that distinct evolutionary patterns led to the independent synteny diversification between plant LEA families. The consequent structural and functional plasticity of LEA proteins may have contributed to plant adaptation to water-limiting environments.

Introduction

When plants colonized land 450Ma ago, they developed a wide range of adaptations including physiological, structural and regulatory mechanisms to cope with variable environments. Land plants (embryophytes) evolved from streptophyte algae, a paraphyletic group of green algae believed to be physiologically pre-adapted to terrestrial environments due their fresh water origin (Kenrick and Crane, 1997; Becker and Marin, 2009; Wodniok et al., 2011).

Plants also developed responses of desiccation tolerance (DT) as they colonized the land. DT is the ability to survive the removal of almost all cellular water without irreparable damage, and it is recurrent in reproductive structures of most vascular plants (e.g. during embryogenesis), in the vegetative body of non-vascular plants and in a few angiosperms species commonly known as 'resurrection plants' (Oliver et al., 2000; Illing et al., 2005; Leprince and Buitink, 2010; Farrant and Moore, 2011; Gaff and Oliver, 2013). Several genes that are thought to be important for DT are common amongst non-vascular and vascular plants, and are also present in their ancestral streptophyte algae (Rensing et al., 2008; Wodniok et al., 2011).

Within the conserved mechanisms of cellular protection involved in DT, a common group named Late Embryogenesis Abundant (LEA) proteins, has received considerable attention. LEAs were originally associated with the acquisition of DT in plant embryos due to the high gene expression and protein accumulation in the later stages of seed maturation (Galau et al., 1986; Dure et al., 1989; Espelund et al., 1992; Manfre et al., 2006; Delahaie et al., 2013). In vegetative tissues, LEAs were found to accumulate upon abiotic stresses such as drought, salinity, heat and freezing, and under desiccation in resurrection plants (Hoekstra et al., 2001; Cuming et al., 2007; Amara, 2014; Stevenson et al., 2016). Interestingly, LEA genes are also found outside the plant kingdom, suggesting a common mechanism of DT across distinct life forms (Browne et al., 2002; Tunnacliffe et al., 2005; Kikawada et al., 2006; Gusev et al., 2014).

LEA proteins exhibit peculiar biochemical properties such as high composition of polar amino acids, high hydrophilicity, and presence of intrinsically disordered regions (IDRs) (Dure et al., 1989; Garay-Arroyo et al., 2000; Goyal et al., 2005; Battaglia et al., 2008). Intrinsically disordered proteins (IDPs), have been proposed as critical for plant adaptation in new environments because of their ability to perform more than one function, the so called 'moonlighting' activity (Covarrubias et al., 2017). This property allow LEAs to perform anti-aggregation, protein stabilization, as well as molecular chaperone-like activities (Chakrabortee et al., 2007; Battaglia et al., 2008; Kovacs et al., 2008; Chakrabortee et al., 2012; Hincha and Thalhammer, 2012; Cuevas-Velazquez et al., 2017).

Several studies have attempted to identify, classify, and access LEAs function in plants (Battaglia et al., 2008; Hundertmark and Hincha, 2008; Shih et al., 2008; Amara, 2014), however, a comprehensive understanding of the evolutionary history and its relationship with the high diversification of protein sequence and function of LEAs in plants is still elusive.

With the increasing number of plant genomes available, a comprehensive analysis of the evolution and functional diversification of LEA gene families is now possible. The reconstruction of the evolutionary history of a protein family in an entire lineage involves homology identification by genome comparative analysis between different taxa, to provide a deeper understanding of the evolution of genomic complexity and lineage-specific adaptations (Koonin, 2005). Phylogenomic analysis (i.e. phylogenetic analysis at the genome scale) has often been employed in order to identify cross-species homologs and predict gene function by reconstructing the evolutionary history (Eisen, 1998).

In this study, we performed a large-scale phylogenomic analysis across 60 complete genomes, combining synteny network and phylogenetic analysis, in order to identify LEAs and investigate their origin and evolution in plants. Our synteny analysis reveals independent evolutionary patterns that shaped synteny diversification of LEAs in plants, and illustrates consequent functional novelties related to water-stress adaptation. Our work provides exciting opportunities for further classification and discovery of new LEA functions in plants.

Results

Global identification of LEAs across 60 genomes

We performed a genome-wide sequence homology search to identify the complete repertoires of LEAs across 60 genomes of diverse plant species (Figure 1). For that we used the most widely employed classification of LEAs that defines eight multi-gene protein families (Pfam): Dehydrin (DHN), LEA_1, LEA_2, LEA_3, LEA_4, LEA_5, LEA_6 and Seed Maturation Protein (SMP) (Hundertmark and Hincha, 2008). Based on the conservation of Hidden Markov Model (HMM) profiles of the eight LEA protein families we identified a total of 4,836 genes, with considerable copy number variation among the LEA families and the genomes investigated (Figure 1, Supplemental Table 1).

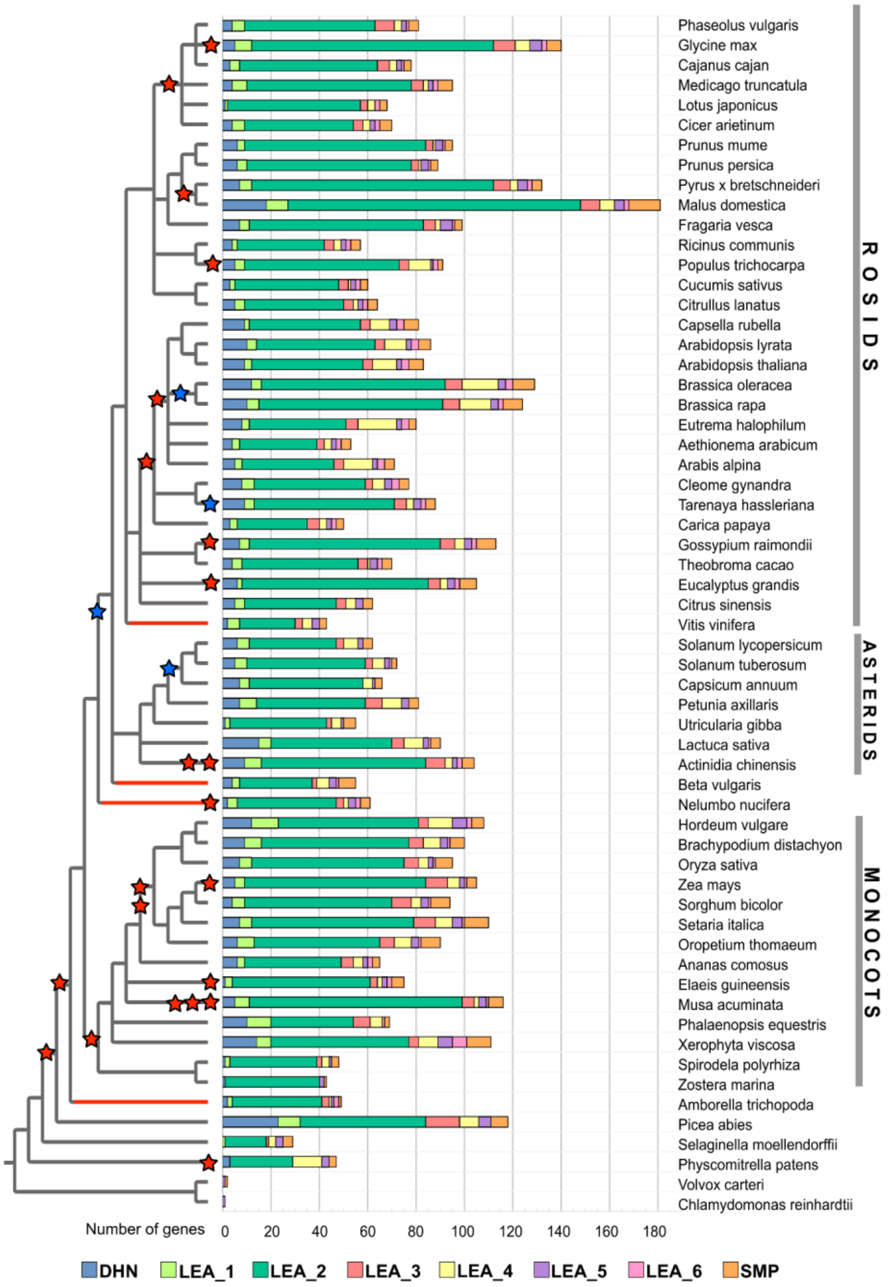


Figure 1 Species phylogeny and number of LEA genes identified in plant genomes. The species tree was inferred using NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/taxonomy). Each LEA family is represented by a specific colour. The red branches in the phylogenetic tree indicate the basal rosid *Vitis vinifera*, the basal eudicots *Beta vulgaris* and *Nelumbo nucifera*, and the basal angiosperm *Amborella trichopoda*. The red and blue stars on the phylogenetic tree indicate whole-genome duplication (WGD), and whole-genome triplication (WGT), respectively.

Single genes belonging to SMP and LEA_5 were found in algal genomes, suggesting an ancestral origin of these families. The Dehydrin, LEA_2 and LEA_4 families were identified in the bryophyte clade (*Physcomitrella patens*) and LEA_1 and LEA_3 families appeared in the lycophyte lineage (*Selaginella moelendorffii*). The LEA_6 family only emerged in early angiosperms (*Amborella trichopoda*), likely representing the most recent LEA family in plants. Overall, LEA_2 family was more abundant with 3,126 genes, being multi-copy in genomes of both angiosperms and lower plants. LEA_6, on the other hand, represents the smallest family with a total of 89 genes identified, with copy-number varying from 0 to 3, with the exception of *Xerophyta viscosa* containing six genes. The variable copy-number between different taxa suggests independent losses or duplication of genes in individual genomes. The under-representation of LEA genes in *Zostera marina* and *Spirodela polyrhiza* (Olsen et al., 2016), and the over-representation in *X. viscosa* (Costa et al., 2017) have already been reported and correlated with the respective desiccation sensitive and tolerant lifestyle of these species, suggesting that the evolution of LEAs contributed to water stress adaptation in plants.

Differential synteny conservation suggests independent evolution of LEA families in angiosperms

We used a synteny-based method to identify homology between the proteins and to explore the evolutionary history of LEAs in plants. Homologous genes comprise orthologs and paralogs, which are corresponding genes in different species that evolved from the same ancestral gene, and to genes duplicated within the same genome, respectively (Koonin, 2005; Gabaldon and Koonin, 2013). Generally, orthologs have equivalent functions in different taxa, while paralogs are prone to perform biologically distinct functions (Koonin, 2005; Gabaldon and Koonin, 2013). Synteny homologs (syntelogs) have similar genomic context and likely evolved from a common ancestor gene (Zhao et al., 2017; Zhao and Schranz, 2017). Syntelogs were inferred with the Synteny Network (Synets) method (Zhao et al., 2017; Zhao and Schranz, 2017) which enables detection of homologs in corresponding chromosomes in different species, as well as paralogs within a species. The output is a network in which the nodes represent anchor genes in a syntenic block and the edges indicate synteny similarity (Figure S1). Synteny communities can be detected in synteny networks using community detection methods (Zhao et al., 2017). Table 1 summarizes the percentage of syntelogs identified per LEA family as well as the number of synteny communities detected in each network (detailed information in Supplemental Table 2).

The variable percentage of syntenic genes and number of synteny communities suggest independent evolution between and within the LEA protein families. Genes not incorporated in synteny communities by our clustering method are likely to be species-specific singletons. No syntelogs were identified between angiosperm and non-angiosperm species, only a few 'in-paralogs' (paralogs from the same species) were detected in the basal species *Sellaginela moellendorffii* and *Physcomitrella patens* (Supplemental Table 2). Considering the long evolutionary distance between the species analyzed, we hypothesized that ancient and independent synteny diversification between LEA families may have functional implications.

Pfam	Total genes	Syntelogs (%)	Synteny communities
DHN	365	62.2	12
LEA_1	251	63.3	10
LEA_2	3126	76.0	130
LEA_3	274	79.9	16
LEA_4	298	77.2	18
LEA_5	153	67.3	4
LEA_6	89	76.4	8
SMP	280	59.3	11
Total	4836		209

Table 1 Summary of syntenic genes and synteny communities identified per LEA protein family.

Phylogenetic profile reveals angiosperm-wide and lineage-specific conservation of LEAs

We further analyzed the origin of the synteny communities detected with Synets in order to obtain information on the evolutionary conservation and diversification of LEAs in angiosperms. Presence or absence of a species syntelog in a community of the synteny network can be visualized as a phylogenetic profile, enabling inference of the origin, expansions, and contractions of the gene family in each clade of the phylogenetic tree (Figure 2a).

We subdivided the synteny communities into four main evolutionary categories: angiosperm-wide (AW), monocot-specific (MS), eudicot-specific (ES), and species-specific (SS) (Figure 2b, Supplemental Table 3). Angiosperm-wide are synteny communities that contain genes of at least one monocot and one eudicot species. Monocot-specific includes synteny communities containing only monocot genes, and eudicot-specific includes communities comprising eudicot genes only. Species-specific correspond to paralogs duplicated in an individual genome, also named ohnologs.

AW communities were found in all LEA families and encompasses the largest number of the syntelogs identified (Figure 2b), indicating that the majority of LEA genes have a common origin in angiosperms and are likely located in a more ancestral genomic context. The angiosperm-wide conservation of LEAs is specially observed in the families DHN, LEA_5 and SMP, where more than 80% of the syntelogs identified are shared amongst angiosperm species. Lineage-specific duplications (MS and ES) have also significantly contributed to the repertoire of LEAs in plants, especially in LEA_3 and LEA_6 families, were than 40% of the syntelogs are distributed over these two categories. SS paralogs were overall underrepresented or absent in the genomes investigated, likely due to low frequency of local gene duplications, or the duplicated copies were more likely to be lost in individual genomes. The finding of lineage-specific and species-specific synteny suggest that duplication events other than whole genome duplications (WGD) has significantly contributed to the expansion of LEA families in plant genomes.

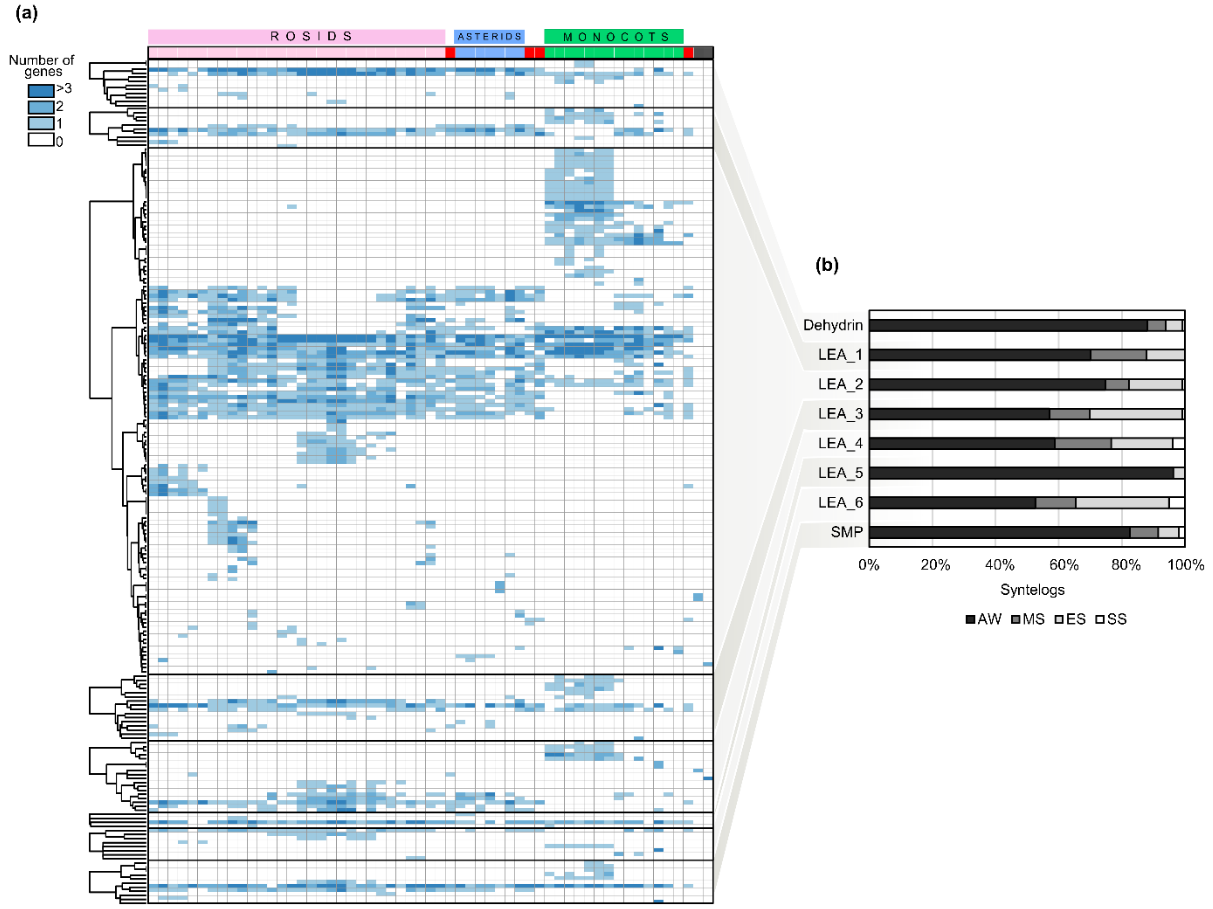


Figure 2 Phylogenetic profile and evolutionary categorization of syntenic LEAs in the genomes analyzed. (a) Phylogenetic profile showing the number and distribution of syntenic LEA genes in plants. Rows represent synteny communities and columns indicate species. The colours on top of the profile indicate rosids (pink), asterids (blue), monocots (green), basal angiosperm species (red) and *Physcomitrella patens* and *Selaginella moelendorffii* (dark grey). The species were ordered from the most recent to the most ancient, from the left to the right. (b) Distribution of syntenic genes in each evolutionary category. AW – angiosperm-wide, MS – monocot-specific, ES – eudicot-specific, SS – species-specific.

The fact that LEA_5 has the smallest number of synteny communities and that the majority of the genes belong to AW conserved genomic context indicate that this is the most conserved LEA family in plants. On the other hand, the large number of LEA_2 syntelogs in AW communities indicates that this is the most diverse LEA family in the plant lineage.

Synteny network analysis provides new insights into functional diversification of LEAs in plants

Duplication events can introduce a gene copy into a new regulatory context, leading to differential evolutionary and regulatory constraints, which is one of the main sources driving functional innovation within a gene family (Conant and Wolfe, 2008; Flagel and Wendel, 2009). Therefore, in the next sections we provide a few examples revealed by our synteny analysis of remarkable structural and functional innovations within LEA families resulted from differential evolution of the genomic context.

Dehydrin: Functional diversification via biochemical, structural, and regulatory innovations

Dehydrin (DHN) is classified as a LEA family due to the gene expression during late seed embryogenesis and ability to perform ‘classical’ chaperone activity, preventing heat-induced protein aggregation and inactivation in vitro (Kovacs et al., 2008; Liu et al., 2017). In our dataset, we found that DHN genes are distributed in two main angiosperm-wide synteny communities and a maximum likelihood tree supports the phylogenetic separation of these communities in angiosperms (Figure 3a).

A set of the DHNs are called hydrophylins because of their specific response to osmotic stress (Garay-Arroyo et al., 2000; Jaspard and Hunault, 2014). Hydrophylins play important role in protecting cell components from the adverse effects caused by low water availability due to biochemical properties such as high Glycine (Gly) content (> 6%) and low grand average hydropathy (GRAVY) (< -1) (Garay-Arroyo et al., 2000; Battaglia et al., 2008; Reyes et al., 2008). In order to investigate the distribution of hydrophylins in angiosperms, we analyzed the Gly content and GRAVY index of each protein within the two largest angiosperm-wide DHN communities (Figure 3b). Although both communities contain proteins with hydrophylin properties, community 1 contains proteins with more variable Gly/GRAVY composition, while community 2 proteins have a more homogeneous Gly/GRAVY distribution. These findings indicate that, even though hydrophylin-type proteins do not form an isolated synteny community, there is a clear biochemical divergence between proteins that evolved in distinct genomic contexts.

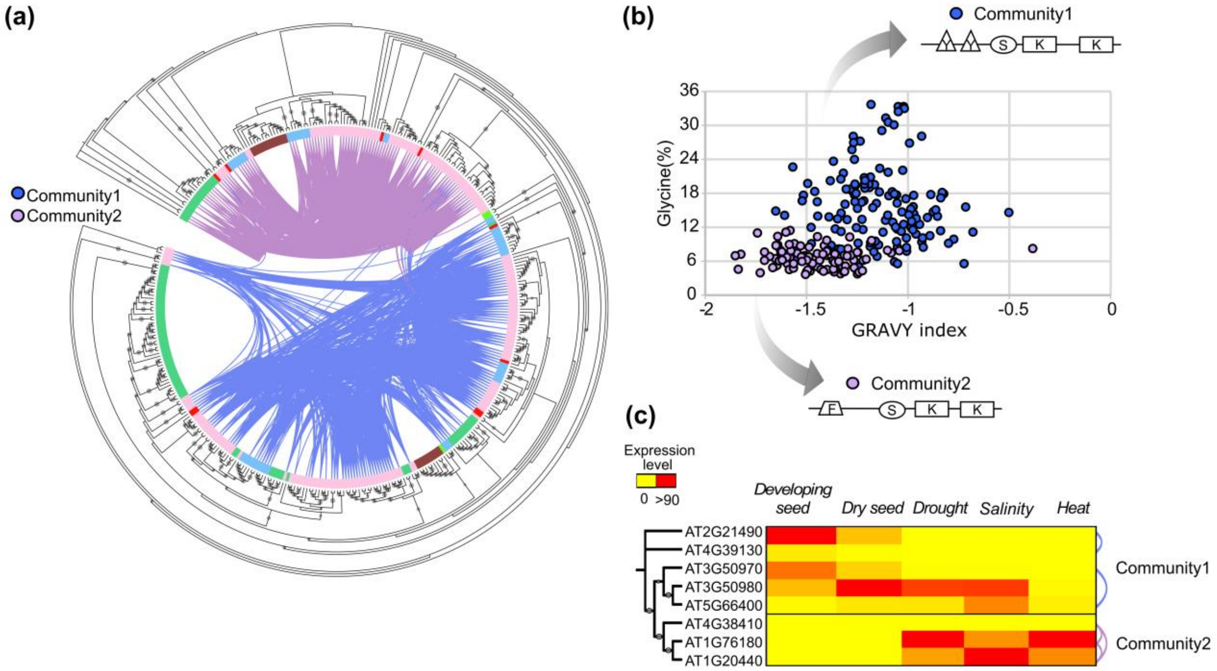


Figure 3 Characteristics of Dehydrin synteny communities. (a) Maximum likelihood tree of all DHN genes found in the genome of 60 species. The inner circle indicates species belonging to monocots (green), rosids (pink), asterids (blue), basal species (red), the gymnosperm *Picea abies* (brown), and the bryophyte *Physcomitrella patens* (light green). The connections between the branches indicate synteny between the gene pairs. Synteny communities 1 and 2 are indicated (blue and pink connections, respectively), dots on the branches represent bootstrap support values (>85). The larger the dots the higher the bootstrap values. (b) Glycine (Gly) content and GRAVY index plot (Gly/GRAVY plot) showing the distribution of hydrophylins between community 1 and 2. The arrows indicate a schematic representation of the consensus sequence of proteins of community 1 and 2, respectively. The F-, Y-, S-, and K- protein segments are indicated according to their position in the protein sequences. (c) Absolute expression values of DHN genes in *Arabidopsis thaliana*. The expression data was retrieved from the Bio-Array Resource for Arabidopsis Functional Genomics (http://bar.utoronto.ca/) and from Hundertmark and Hincha (2008). The dots on the branches of the phylogenetic tree indicate bootstrap support values (>75). Connections between the rows represent synteny relationships.

DHN proteins have also been functionally subdivided into four to five main architectures based on the presence and organization of specific motifs called Y-, S- or K- segments (Close, 1996; Hunault and Jaspard, 2010; Banerjee and Roychoudhury, 2016; Malik et al., 2017). We performed multiple sequence alignments of proteins from the DHN synteny communities 1 and 2 in order to investigate the diversification of the different functional motifs (Figure S2a, b). Our data indicate that the majority of proteins of community 1 comprises Y(n)SK(n) types (Figure 3b indicated with an arrow, Figure S2a), while community 2 contains mainly SK(n)-type proteins, lacking the Y-segment at the N-terminus (Figure S2b). Despite lacking the Y-segment, we found that proteins from community 2 possess a new conserved segment at the N-terminus (DRGLFDFLGKK).

This motif is named F-segment, and it was recently characterized as an overlooked motif in angiosperms and gymnosperms, with potential functional roles in membrane and protein binding (Strimbeck, 2017). Interestingly, genes encoding proteins belonging to community 1 express mainly during seed development in A. thaliana, and some of the genes can be induced by abiotic stress (Figure 3c). On the other hand, genes encoding the F-type DHN proteins of community 2 seem to be specifically induced by abiotic stresses such as drought, heat and salinity. These results combined indicate that the ancient synteny diversification DHN in angiosperms resulted in protein biochemical and sequence innovations, as well as changes in expression patterns that may be related to functional specificity within this protein family. To date, this is the first documentation of the evolution and diversification of the F-segment in angiosperms, and its association with abiotic stress in A. thaliana.

LEA_1: Ancient diversification of Intrinsically Disordered Proteins (IDPs) in angiosperms

LEA_1 proteins, also known as Group 4, also accumulates in the plant cell in response to water stress and has been proposed as model to study intrinsically disorder proteins (IDPs) in plants (Olvera-Carrillo et al., 2010; Cuevas-Velazquez et al., 2017). This family has also been subdivided into two main subclasses based on protein sequence features (Battaglia et al., 2008). One of the subgroups, named group 4A, comprises smaller proteins (80-124 residues) and the second group, 4B, has longer representatives (108-180 residues). Both subclasses possess a conserved portion at the N-terminal region, and a disordered C-terminal region predicted to form alpha-helices under water limiting conditions (Cuevas-Velazquez et al., 2017).

Our data indicates that LEA_1 members are distributed in 10 synteny communities, and 70% of the homologs identified with Synets belong to two angiosperm-wide (AW) communities (Figure 2b, Figure 4a). The absence of clear synteny and phylogenetic separation observed in the phylogenetic tree suggests that some of the ES and MS communities have originated by duplication or transposition of genes from AW communities (Figure 4b). We found differences between the consensus sizes of the multiple sequence alignments of protein from the two AW communities (Figure 4c, Figure S3a, b), what indicates that AW community 1 represents the subclass 4B of longer protein sequence, whereas community 2 contains members of subclass 4A of smaller proteins. It seems that the diversification of intrinsically disordered regions (IDRs) in LEA_1 occurred before the origin of monocots and eudicots, and that these protein types has been conserved in angiosperm genomes during evolution.

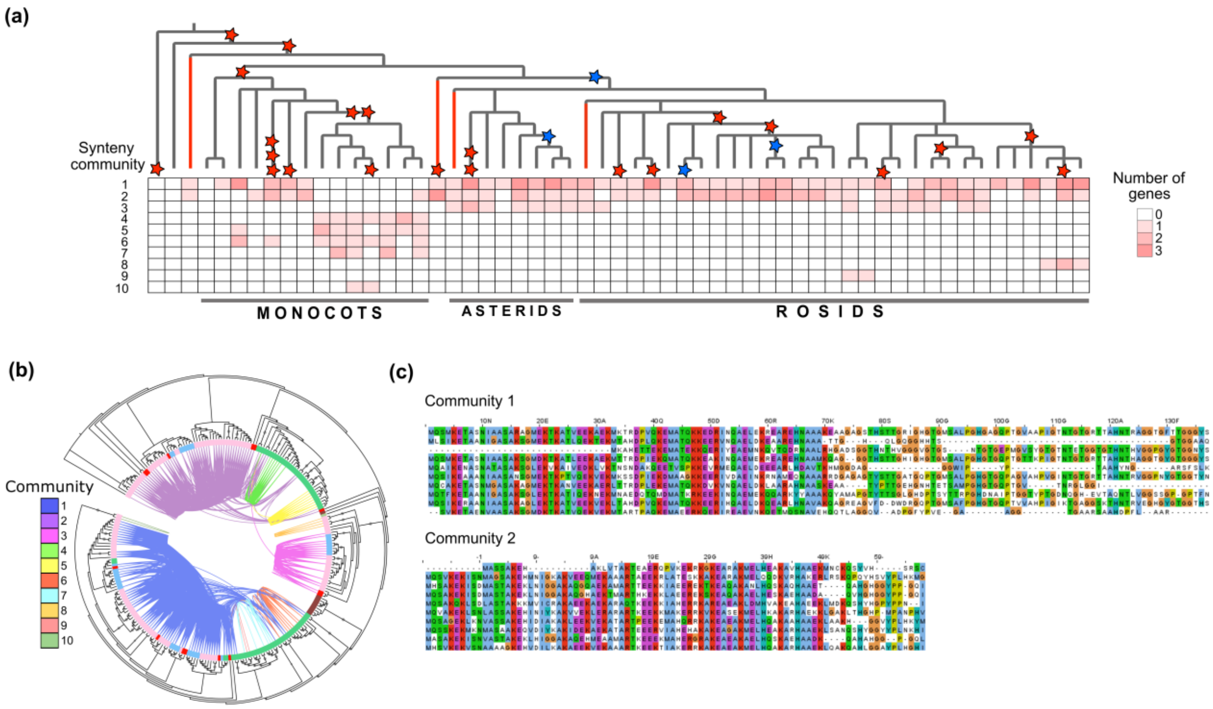


Figure 4. Phylogenetic ad synteny characteristics of LEA_1. (a) Phylogenetic profile of LEA_1 indicating the distribution of the synteny communities detected in the species phylogenetic tree. The red and blue stars indicate whole-genome duplication (WGD) and whole-genome triplication (WGT), respectively. (b) Maximum likelihood tree of the LEA_1 family. The circle inside the tree indicates species belonging to monocots (green), rosids (pink), asterids (blue), basal species (red), the gymnosperm *Picea abies* (brown), the bryophyte *Physcomitrella patens* (light green), and the lycophyte *Selaginella moellendorffii* (olive green). The connections between the branches indicate synteny between the gene pairs, and dots on the branches represent bootstrap support values (>85).The larger the dots the higher the bootstrap values. (c) Partial representation of the multiple sequence alignments of amino acid sequences of the communities 1 and 2 (top 10 sequences).

LEA_2: Expansion and diversification through recurrent tandem duplications

LEA_2 is the largest LEA family, and has been considered atypical because it contains proteins with more hydrophobic amino acids and more defined secondary structure in solution (Singh et al., 2005; Hundertmark and Hincha, 2008). Members of this family have been associated with hypersensitive response (HR) after microbial and parasitic nematode infection, which also differs from the other LEA families (VanderEycken et al., 1996; Escobar et al., 1999; Ciccarelli and Bork, 2005). However, functions associated with salinity, freezing, heat, UV radiation, osmotic, and oxidative stress in vitro have also been documented for LEA_2 proteins (He et al., 2012; Jia et al., 2014; Jiang et al., 2017).

Despite the large number of members, in general, synteny and phylogeny of the LEA_2 are in agreement, with highly supported branches in the phylogenetic tree connecting genes that belong to the same synteny community (Figure 5a). Interestingly, there is an evident interconnection between two of the largest LEA_2 synteny communities (Figure 5a, b). We found that these communities contain several tandem duplicates widespread in monocots and eudicots (Figure 5b, c). In fact, we also found several other tandem duplicates across other LEA_2 communities containing monocots and eudicots genes (Supplemental Table 4). These results indicate that tandem duplications have significantly contributed to the expansion and diversification of the large LEA_2 family in angiosperms, and may be one of the causes of the diversified functionality of this atypical LEA family.

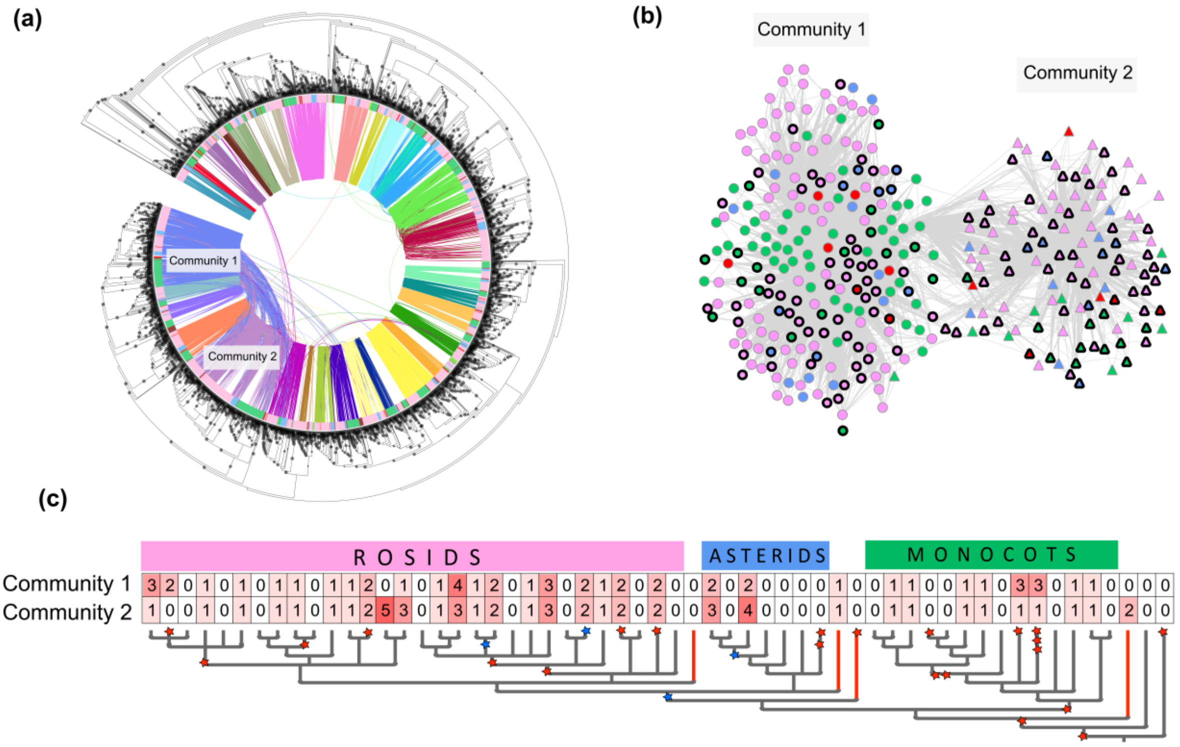

(a)
Community 1
Community 2
(b)
Community 1
Community 2
(c)
ROSIDS
ASTERIDS
MONOCOTS
Community 1 3 2 0 1 0 1 0 1 1 0 1 1 2 0 1 0 1 4 1 2 0 1 3 0 2 1 2 0 2 0 0 2 0 2 0 0 0 0 1 0 0 1 1 0 0 1 1 0 3 3 0 1 1 0 0 0 0
Community 2 1 0 0 1 0 1 0 1 1 0 1 1 2 5 3 0 1 3 1 2 0 1 3 0 2 1 2 0 2 0 0 3 0 4 0 0 0 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 1 0 2 0 0

Figure 5 Tandem duplications of the LEA_2 family. (a) Maximum likelihood tree containing all LEA_2 genes identified. The colours displayed in the inner circle indicate genes belonging to monocots (green), rosids (pink), asterids (blue), basal species (red), the gymnosperm *Picea abies* (brown), the bryophyte *Physcomitrella patens* (light green), and the lycophyte *Selaginella moellendorffii* (olive green). The connections between the branches indicate synteny between the gene pairs, and all the communities with at least 100 syntenic genes are displayed in different colours. Synteny communities 1 and 2 are indicated. The dots on the branches indicate bootstrap support values (>85). The larger the dots the higher the bootstrap values. (b) Synteny network of genes belonging to community 1 (circles) and community 2 (triangles). The colours displayed in the nodes represent the clades as indicated in (a). Tandem genes are indicated by a thicker black border. (c) Summary of the number of tandem duplicates in the synteny communities 1 and 2. The tree is a simplified version of the species tree presented in figure 1. Red stairs indicate WGD and blue stars indicate WGT.

LEA_4: Dynamic synteny in plant desiccation tolerance

LEA_4 genes, also known as group 3, are also found in non-plant organisms that display DT such as rotifers, arthropods, nematodes, and tardigrades (Browne et al., 2002; Tunnacliffe et al., 2005; Kikawada et al., 2006; Gusev et al., 2014) suggesting an association with the evolution of DT. In plants, LEA_4 is strongly associated with DT in basal and angiosperm resurrection species via an anciently conserved ABA signalling pathway (Cuming et al., 2007; Hundertmark and Hincha, 2008; Shinde et al., 2012; Delahaie et al., 2013; Stevenson et al., 2016). Our species set contained two desiccation tolerant species, the bryophyte Physcomitrella patens and the monocot *Xerophyta viscosa*, however, synteny cannot be detected between these species due to the large evolutionary distance. Nevertheless, we found that LEA_4 genes are distributed across several AW, MS, and ES synteny communities that are phylogenetically separated, suggesting a dynamic evolutionary history of this gene family in angiosperms (Figure 2b, Figure S4). Interestingly, only one of the eight LEA_4 genes identified in *X. viscosa* shares synteny with other angiosperm species, all the other duplicates are singletons or in-paralogs (Supplemental Table 5). In *X. viscosa*, LEA_4 family has expanded compared with other monocot species (Costa et al., 2017), which was correlated with higher desiccation response. Altogether, it seems that LEA_4 has evolved dynamically in angiosperm, and loss of synteny may result in fixation of these genes in the genome, resulting in improved contribution for DT in resurrection plants.

Discussion

How does the plant genome adapt to environmental stress? This question has been addressed frequently in recent years. It has been proposed that adaptation to novel or stressful environments is correlated with the retention of duplicated genes (Flagel and Wendel, 2009; Jiao et al., 2011; Kondrashov, 2012; Panchy et al., 2016). Among the many models for duplicated gene evolution (Conant and Wolfe, 2008; Innan and Kondrashov, 2010), it is suggested that genes that should be rapidly or constantly produced in response to environmental stress might be more prone to selection after duplication (Kondrashov, 2012).

In plants, the group of Late Embryogenesis Abundant (LEA) proteins, composed of eight multi-gene families (Dehydrin, LEA_1, LEA_2, LEA_3, LEA_4, LEA_5, LEA_6 and SMP), have been shown to play roles in water stress tolerance, and may represent a conserved and indispensable component of regulatory networks involved in environmental stress adaptation that allowed plants to endure the constraints associated with land adaptation (Shih et al., 2008; Hincha and Thalhammer, 2012; Amara, 2014). Evidences suggest that there is functional variability between and within each of the eight families (Hundertmark and Hincha, 2008), what raises questions such as what are the sources of functional variations, what are the precise biological functions of each family, if and how LEA families work as one entity, and which LEA genes are involved in plant development and stress tolerance.

To answer some of these questions we interrogated 60 whole genomes, ranging from green algae to angiosperms and analyzed the ancestry, conservation, and diversification of LEAs in plants. We found that LEA proteins belonging to LEA_5 and SMP families have arisen early in the plant lineage, while the other families appeared at later instants during plant evolution (Figure 1). Previous studies have already shown the presence and expression of ancient LEAs in algal genomes (Joh et al., 1995; Wodniok et al., 2011), corroborating the hypothesis that the ancestral fresh water lineages were pre-adapted to terrestrial environments, and the evolution of pre-existing and new gene families, including LEAs, may have facilitated the colonization of land (Rensing et al., 2008; Becker and Marin, 2009). It seems possible that later LEA families expanded and diversified in embryophytes as a result of the evolution of more specialized cells, tissues and organs such as spores and seeds, that required a better control of water retention and protection against desiccation and other stresses.

Synteny homology analysis indicated a clear genomic diversification of LEA genes during angiosperm evolution (Figure 2). The majority of LEA genes are located in angiosperm-wide conserved genomic regions, while the finding of clade-specific as well as species-specific gene copies indicates that the continuing expansion and diversification of angiosperm genome contributed to LEA gene families evolution. Stress-regulated genes retained after duplication events are more likely to neofunctionalize instead of inheriting the ancestral function, which might be in part related to changes in biochemical function and in cis-regulatory regions (Conant and Wolfe, 2008; Zou et al., 2009; Arsovski et al., 2015). As a result of these changes, complete or partial diversification of the interaction and regulatory networks in which the duplicated genes are involved might also occur. It is likely that the genes belonging to the same synteny community (positional homologs) display similar functions, and genes in different communities are likely to display functional innovations (Dewey, 2011).

We identified highly conserved synteny between LEA_5 genes in most genomes investigated, suggesting evolutionary constraints on maintaining the stability of their genomic context. These constraints may include the correct functioning of the maturation-induced desiccation program, where LEA_5 genes of A. thaliana were shown to play important roles (Manfre et al., 2006), and may be conserved across all orthodox angiosperm species.

We also found several examples of correlation between synteny diversification and functional innovations. Genes from the largely studied Dehydrin (DHN) family are localized in two distinct synteny communities across the angiosperm lineage (Figure 3a). Presumably, new regulatory elements were acquired in the duplicated copies, and differential evolutionary forces may have driven protein diversification, resulting in distinct biochemical properties (Figure 3b). The consequent differential gene expression (developmental or stress induced) may have allowed the preservation of duplicated copies in the different genomes, and amplified the stress tolerance response. The finding of functionally diverse Dehydrin types in *Physcomitrella patens* suggests that the colonization of land was one of the forces driving Dehydrin evolution (Ruibal et al., 2012; Agarwal et al., 2017). Similarly, LEA_1 have evolved into two angiosperm-wide synteny communities composed by two protein types containing distinct intrinsically disordered regions IDRs (Figure 4c). Our findings point toward an ancient functional divergence between LEA_1 members, what would explain their structural plasticity and 'moonlighting' properties associated with multiple abiotic stresses (Covarrubias et al., 2017; Cuevas-Velazquez et al., 2017).

Another source of evolutionary adaptations to environmental stress is gene family expansion via recurrent tandem duplications (Cannon et al., 2004; Hanada et al., 2008). Tandem duplications offers a pool of targets for evolutionary selection contributing to the maintenance of large gene families. These large gene families are enriched with genes important for rapid environmental adaptation such as biotic stress-responsive genes (Cannon et al., 2004; Hanada et al., 2008). We found several tandem duplicates in the synteny network of LEA_2 distributed across all angiosperm lineage (Figure5). This finding supports the atypical structured and hydrophobic nature of LEA_2 proteins and its broader spectra of gene expression in response to biotic and abiotic stresses (Ciccarelli and Bork, 2005; Singh et al., 2005; Hundertmark and Hincha, 2008).

Most of the LEA gene expression during seed development and environmental stresses is regulated via abscisic acid (ABA)-signalling pathways (Galau et al., 1986; Espelund et al., 1992; Shinde et al., 2012; Delahaie et al., 2013; Stevenson et al., 2016). The desiccation-induced LEA gene expression via ABA-responsive pathways is conserved across basal and angiosperm resurrection species (Cuming et al., 2007; Shinde et al., 2012; Stevenson et al., 2016). It seems that the acquisition of new genomic contexts by desiccation-related LEAs of the resurrection monocot *Xerophyta viscosa* is one important footprint of DT, and suggests that other regulatory mechanisms, likely independent on ABA, may also work to assure protection against desiccation.

Resurrection plants are species adapted to live in environments with low water availability, displaying specific molecular and genomic adaptations of DT (Oliver et al., 2000; Mundree, 2002; Illing et al., 2005; Farrant and Moore, 2011; Gaff and Oliver, 2013). The concept of DT is different from drought tolerance because drought tolerance refers to the tolerance to moderate water removal without removal of the bulk of cytoplasmic water (Shih et al., 2008), while DT refers to the tolerance to a further dehydration with an increased removal of the water shell and the capacity to survive long periods in the dry state (Hoekstra et al., 2001). Understanding the mechanisms underlying DT can help to improve drought tolerance in crops (Mundree, 2002; Leprince and Buitink, 2010; Costa et al., 2017). Several crops from the grass family (Poaceae) constitutes major contributors of global food security that have become targets of genomic programs aiming at improved drought tolerance. In grasses, overexpression of LEAs has already been shown to enhance tolerance to drought and other stresses (Babu, 2004; Fu et al., 2007; Xiao et al., 2007; Chen et al., 2015). We believe that comprehending the impact of synteny diversification in functional innovations in the LEA families may offer an extra powerful tool to select candidates for engineering drought and desiccation tolerant crops.

This data also opens several opportunities for hypothesis-driven fundamental and experimental characterization of the myriad of functions of LEA proteins, and the role of the diversification of the genomic context in plant evolution and adaptation to environmental stresses. Deciphering the evolution of eight gene families, with variable protein structure and diversified expression patterns over billions of years, is a challenging task. Despite the general association of LEAs water stress response, our work provides strong examples of a clear evolutionary divergence resulting in differential protein evolution. The diversity of LEA families in angiosperms is a result of extensive and dynamic synteny evolution, which indicates that the complexity of these gene families goes beyond the protein sequences.

Methods

Identification of LEA proteins

We used 60 fully sequenced genomes available in Phytozome (Goodstein et al., 2012) (https://phytozome.jgi.doe.gov/), and the recently published genome of *Xerophyta viscosa* (Costa et al., 2017). Our species list includes representative species belonging to green algae, mosses, lycophytes, gymnosperms, early angiosperms, monocots, early eudicots, asterids and rosids (Figure 1, Supplemental Table 1).

Several classifications have been proposed for LEA proteins (for a review, see (Battaglia et al., 2008). Here we used the Pfam annotation for protein families (Bateman et al., 2002) (http://pfam.xfam.org/) based on conserved protein domains (Hundertmark and Hincha, 2008). This annotation classifies LEAs into eight Pfams: Dehydrin (DHN) (PF00257), LEA_1 (PF03760), LEA_2 (PF03168), LEA_3 (PF03242), LEA_4 (PF02987), LEA_5 (PF00477), LEA_6 (PF10714), and Seed Maturation Protein (SMP) (PF04927). Hidden Markov Models (HMM) retrieved from the Pfam 3.0 database (http://Pfam.xfam.org) were queried against the 60 plant genomes to identify LEA proteins for each family using the program ‘hmmscan’ of the HMMER3.0 package (Finn et al., 2011). All proteins with significant hits (e-value < 0.001) were used in this analysis.

Synteny network construction and community detection

We used the Synets method (Zhao et al., 2017; Zhao and Schranz, 2017) for syntenic block calculations, network construction and community detection (https://github.com/zhaotao1987/SynNet-Pipeline). In summary, pairwise all-against-all comparisons were performed using RAPSearch (Zhao et al., 2012). Synteny block detection was performed with MCScanX software (Wang et al., 2012) with default parameters (minimum collinear block size = 5 genes, maximum gaps = 25 genes). The syntenic blocks containing the identified LEA sequences were used to build synteny networks (Synets) that were visualized and edited with Cytoscape 3.3.0 (Shannon et al., 2003) and Gephi 0.9.1 (Bastian, 2009) (https://gephi.org/). Infomap (Rosvall and Bergstrom, 2008) was used to find communities within the synteny networks, which is implemented under “igraph” package in R (http://igraph.org/r/doc/cluster_infomap.html). All synteny communities were numbered according to the largest to the smallest number of genes, and later renamed per LEA family accordingly (Supplemental Table 2). The synteny communities were further analyzed with a phylogenetic profiling. Phylogenetic profiling allows the visualization of the synteny communities that are lineage-specific or shared amongst different species. All synteny communities were decomposed into numbers of involved syntenic gene copies in each genome. Dissimilarity index of all clusters was calculated using the “Jaccard” method of the vegan package (Dixon, 2003), hierarchically clustered by “ward.D”, and visualized by “pheatmap”.

Multiple sequence alignments (MSAs) were built for each of the eight LEA families using MAFFT v.7 (Katoh et al., 2002). We used the automated method for the Pfam LEA_2 due to the large number of sequences, and the method G-INS-l for all other LEA Pfams. Phyutility 2.2.6 (Smith and Dunn, 2008) was used to trim gaps and maintain 75% the consensus alignment. The final MSAs were edited and displayed with Jalview 2.10.3 (Waterhouse et al., 2009). IQ-TREE v.1.5.1 (Nguyen et al., 2015) was used to infer Maximum Likelihood (ML) trees with 1000 bootstraps for each alignment. All phylogenetic trees were edited and displayed with the online tool iTOL (Letunic and Bork, 2016).

Physicochemical properties and expression data of Dehydrin proteins

The hydrophilicity index of Dehydrin proteins was calculated with the online GRAVY calculator (http://www.gravy-calculator.de/). More hydrophilic proteins have a more negative GRAVY score, and more hydrophobic proteins have a more positive GRAVY score. In order to reveal hydrophylin-type proteins (GRAVY < -1 and Gly > 6%), individual GRAVY scores were plotted against the percentage of Glycine (Gly) per protein sequence (Garay-Arroyo et al., 2000; Battaglia et al., 2008). Absolute gene expression values were retrieved from the e-Northern tool provided by the Bio-Array Resource for Arabidopsis Functional

Genomics (http://bar.utoronto.ca/) as well as from the datasets of seed and silique development, dry seed, drought and heat shock of Hundertmark and Hincha (2008).

Acknowledgements

M.A.S.A. received financial support from CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education (BEX 0857/14-9). T.Z. was supported by China Scholarship Council.

Author contributions

M.A.S.A, T.Z., W.L., M.E.S. and H.W.M.H. planned and designed the research, M.A.S.A and T.Z. performed the research and analyzed the data. M.A.S.A interpreted the data and wrote the manuscript with contributions of T.Z., W.L., M.E.S. and H.W.M.H. All authors edited and commented on the manuscript.

Data availability: The authors declare that all relevant data supporting the findings of this study are available within the paper and in its supplementary information files.

Supplementary information

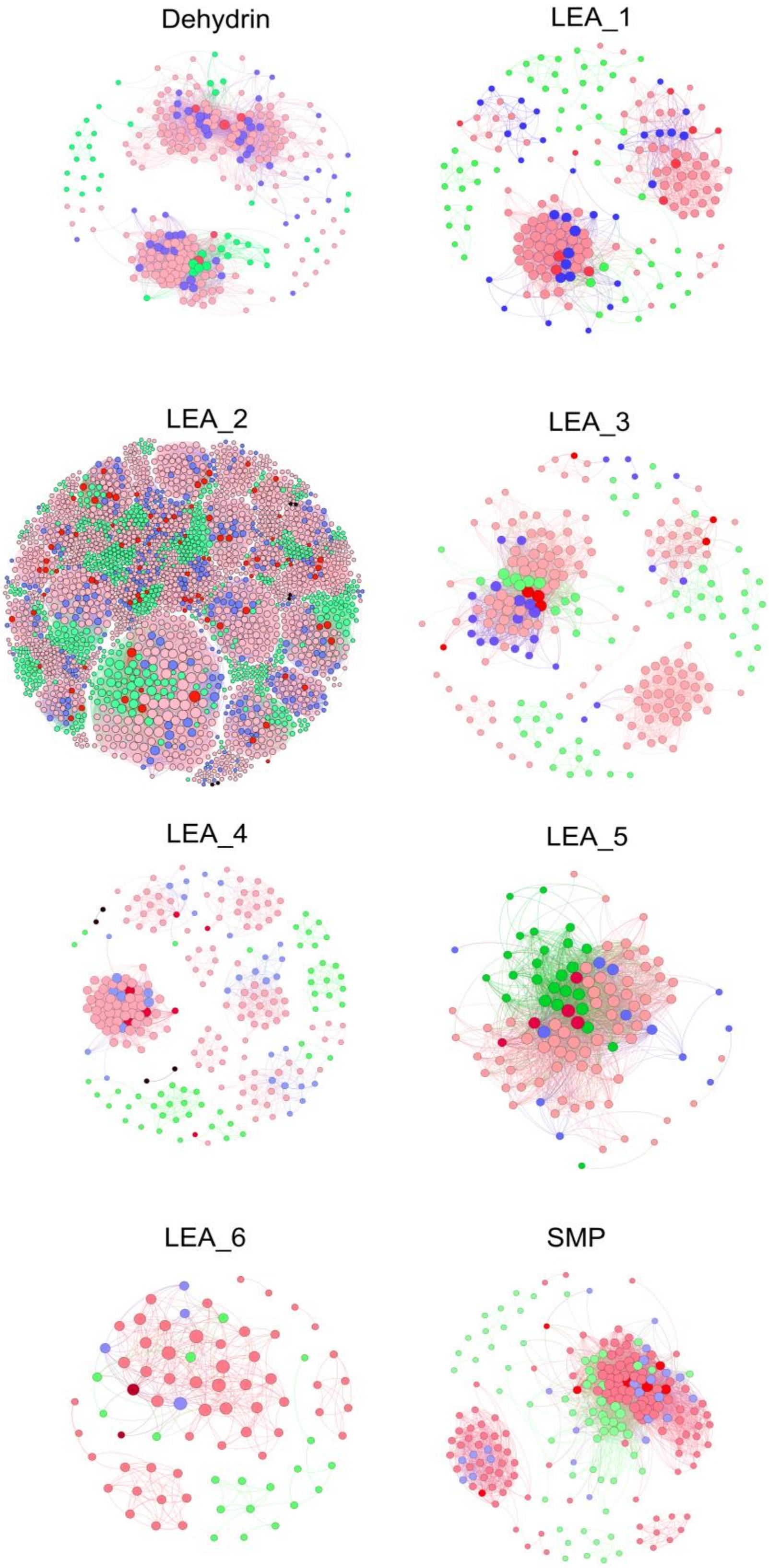


Figure S1. Synteny networks of LEA genes. Synteny networks were built with the synteny information of syntelogs identified with Synets. Node colours indicate species belonging to monocots (green), rosids (pink), asterids (blue), the basal rosid *Vitis vinifera*, the basal eudicots *Beta vulgaris* and *Nelumbo nucifera*, and the basal angiosperm *Amborella trichopoda* (red).

Physcomitrella patens and *Selaginella moelendorffii* syntelogs are indicated as black nodes. Node size indicates the number of connections: bigger nodes have more connections (stronger synteny relationships).

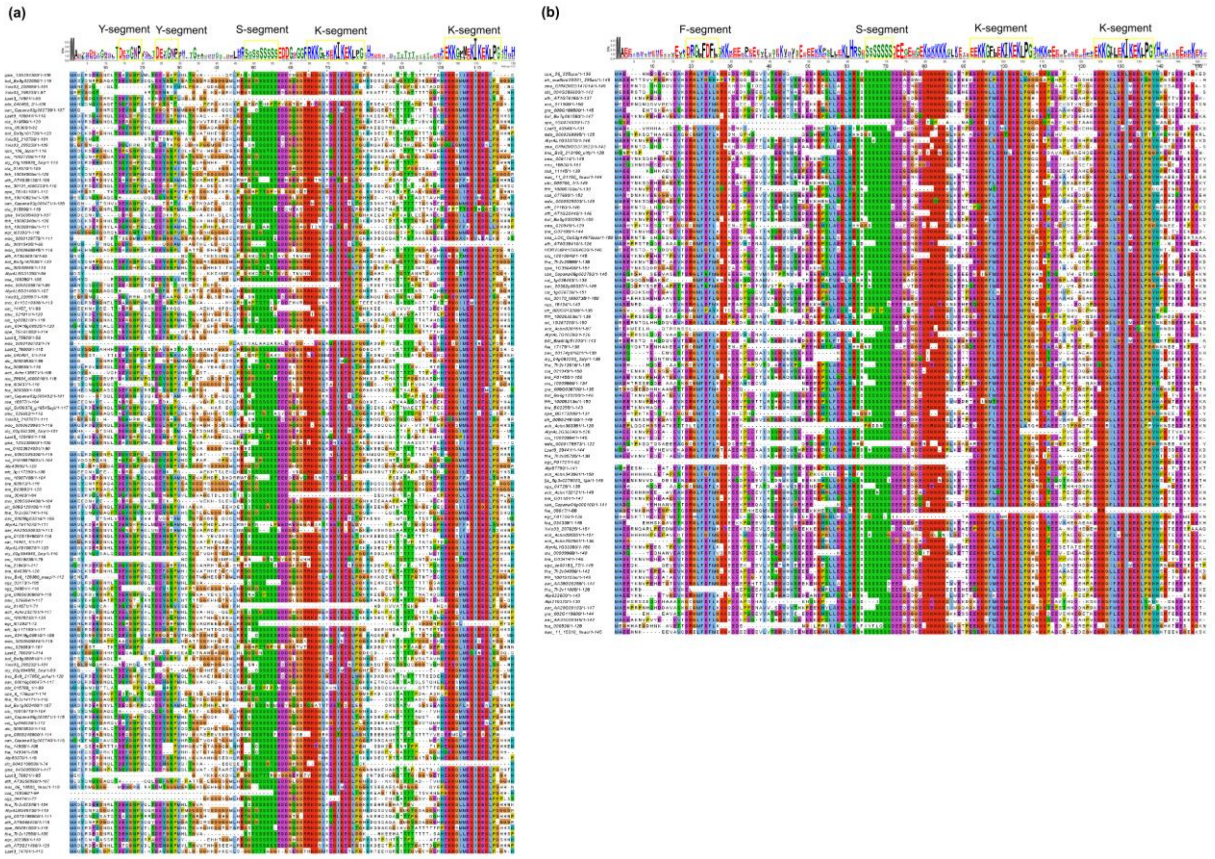


Figure S2. Multiple Sequence Alignments (MSA) of proteins belonging to DHN. MSA of proteins belonging to DHN synteny community 1 (a) and 2 (b). The consensus sequence is shown at the top of each alignment. The F-, Y-, S-, and K- protein segments are indicated.

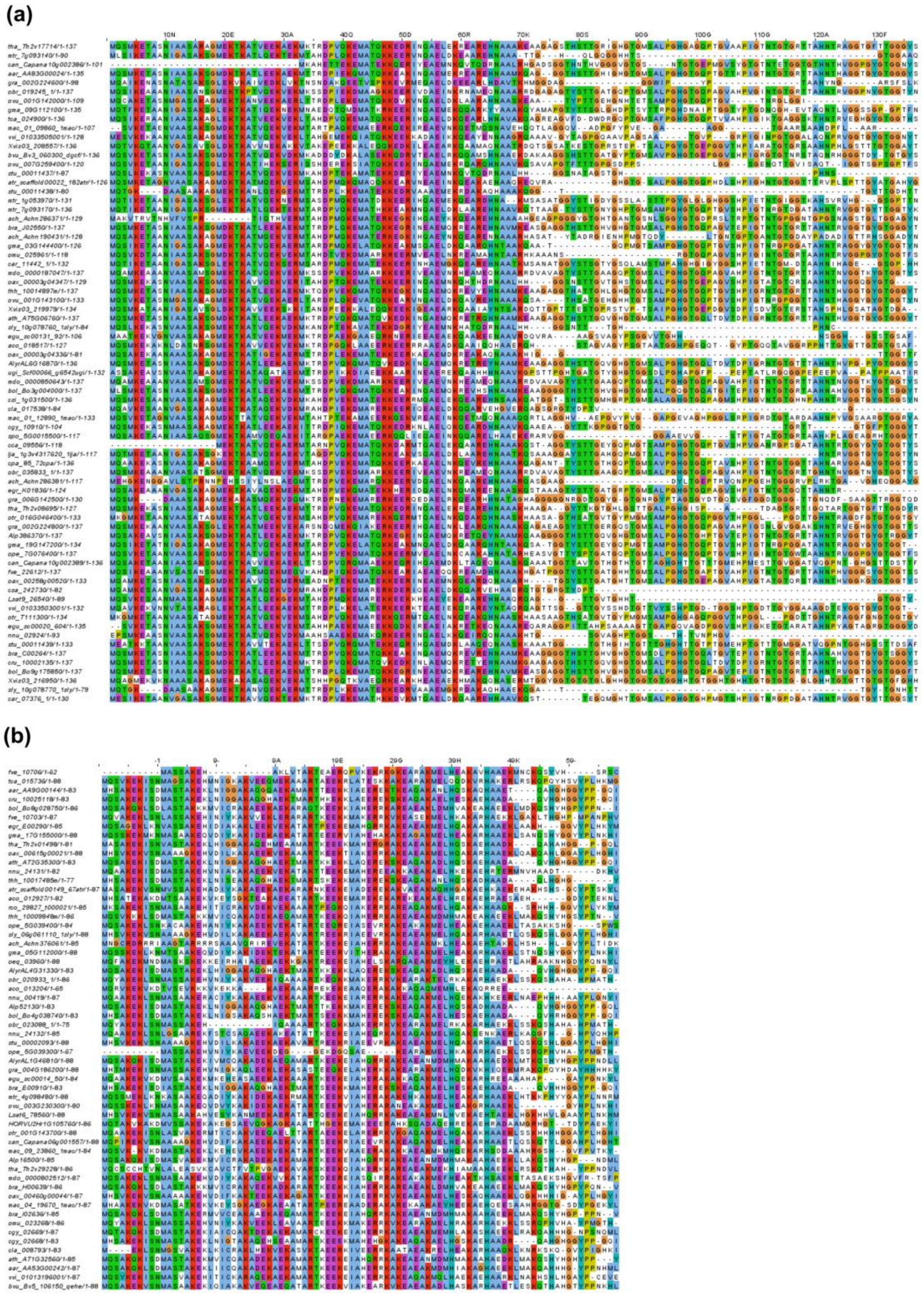


Figure S3. Multiple Sequence Alignments (MSA) of proteins belonging to LEA_1. MSA of proteins from synteny community 1 (a) and 2 (b).

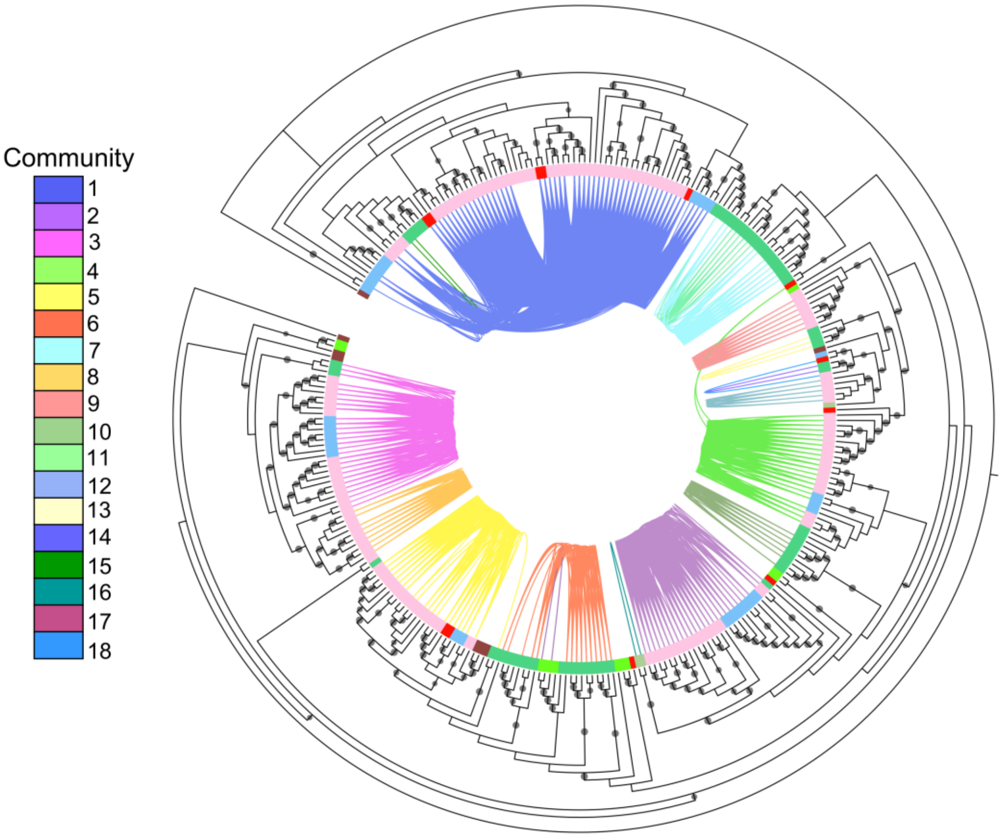


Figure S4. Maximum likelihood trees and synteny relationships of LEA_4. The circle inside the tree represents species belonging to monocots (green), rosids (pink), asterids (blue), basal species (red), the gymnosperm *Picea abies* (brown), the bryophyte *Physcomitrella patens* (light green), and the lycophyte *Selaginella moellendorffii* (olive green). The connection between the branches indicates synteny between the gene pairs and the dots on the branches indicates bootstrap support values (>85). The larger the dot the higher the bootstrap value. The colour scale for the different communities is indicated in the left.

Supplementary tables

Supplemental Table 1. Number of LEA proteins found in 60 species. The highlight colour represents the major clades.

Supplemental Table 2. Number of syntenic genes and syntenic communities.

Supplemental Table 3. Evolutionary categories of syntenic LEAs in angiosperms.

Supplemental Table 4. Tandem duplicates in the LEA_2 family.

Supplemental Table 5. Synteny information of *Xerophyta viscosa* LEA_4 genes.

Available at
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LHK9WL

Chapter 5

Phylogenomic Synteny Networks Reveal Gene Positional Conservation and Diversification Across 170 MYR of Mammal and Angiosperm Evolution

Tao Zhao[1], M. Eric Schranz[1]

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

To whom correspondence should be addressed:

Email: eric.schranz@wur.nl

PNAS, under revision

Abstract

Synteny analysis is a valuable approach for understanding eukaryotic gene and genome evolution, but still relies largely on pairwise or reference-based comparisons. Network approaches can be utilized to expand large-scale phylogenomic local synteny studies. Here we have built complete local synteny networks for 87 mammalian and 107 angiosperm genomes, major lineages that have evolved and radiated over the last ~170 million years. The networks contain ~1.5 million nodes (genes) and ~49.4 million edges (syntenic connections between genes) for mammals, and ~2.2 million nodes and ~47 million edges for angiosperms. By characterizing the entire networks with network parameters, such as clustering, size, average clustering coefficient, and node degree, we illustrate and quantify overall synteny conservation and diversification properties of all annotated genes for mammals and angiosperms. These provide new metrics for assessing genome evolution of mammal and angiosperm genomes. Further, we compare the functional characteristics of extremely conserved and diversified gene families, and perform phylogenomic profiling to identify lineage-specific clusters. We depict several representative clusters of important developmental genes in humans, such as *CENPJ, p53* and *NFE2*. Taken together, we illustrate how network approaches can enhance comparative genomic analysis.

Introduction

The patterns and differences of gene and genome duplication, gene loss, gene transpositions and chromosomal rearrangements can inform how genes and gene families have evolved to regulate and generate (and potentially constrain) the amazing biological diversity on Earth today. For comparative genomics, synteny reflects important relationships between the genomic context of genes both in terms of function and regulation and is often used as a proxy for the constraint and/or conservation of gene function (Dewey, 2011; Lv et al., 2011). Thus, syntenic relationships across a wide range of species provide crucial information to address fundamental questions on the evolution of gene families that regulate important traits. Synteny data can also be very valuable for assessing and assigning gene orthology relationships, particularly for large multigene families where phylogenetic methods maybe non-conclusive (Koonin, 2005; Dewey, 2011; Gabaldon and Koonin, 2013). Synteny was originally defined as pairs or sets of genes located on homologous chromosomes in two or more species, but not necessarily in the same order (Passarge et al., 1999). However, the current widespread usage of the term synteny, which we adopt, implies conserved collinearity and genomic context.

While the basic tenants of gene and genome organization and evolution are similar across major eukaryote lineages, there are also significant differences that are not fully characterized nor understood. For example, the length and complexity of genes and promoters, the types of gene families (shared or lineage-specific), transposon density, higher-order chromatin domains and the organization of chromosomes can differ significantly between plants, animals and other eukaryotes (Gladyshev and Arkhipova, 2007; Feng et al., 2010; Law and Jacobsen, 2010; Murat et al., 2012). In general, genome organization and gene collinearity is substantially less conserved in plants than in mammals. One major characteristic of flowering plant genomes is the prevalent signature of shared and/or lineage-specific whole genome duplications (WGDs) (Adams and Wendel, 2005; Cui et al., 2006; Jiao et al., 2011; Jiao et al., 2012a; Soltis et al., 2015; Barker et al., 2016). While the genomes of mammalian vertebrates show evidence of only two shared and very old rounds of WGD; often referred to as “2R” (Hokamp et al., 2003; Panopoulou and Poustka, 2005; Steinke et al., 2006). The variation in genomic organization between lineages is partially due to differences in fundamental molecular processes such as DNA-repair and recombination, but also likely reflect the historical biology of groups (such as mode of reproduction, generation times and relative population sizes). Differences in gene family and genome dynamics have significant effects on our ability to detect and analyze synteny.

While the number of quality reference genomes is growing exponentially, a major challenge is how to detect, represent, and visualize synteny relations of all members from a gene family across many genomes simultaneously. Conventional dot plots display macroscale collinear blocks between/within only two genomes in two-dimensional images. Parallel coordinate plots (like SynFind (Lyons and Freeling, 2008;

Tang et al., 2015)) describe collinear blocks surrounding a locus identifier and visualize the blocks at the local genomic scale. With the abundance of new genomic data, the challenges for multispecies collinearity visualization are only exacerbated. We have developed a network-based approach to organize and display local synteny (Zhao et al., 2017; Zhao and Schranz, 2017) and have applied it to understand the evolution of the entire MADS-box transcription factor family across 51 plant genomes as a proof of principle of the method (Zhao et al., 2017). We identified several evolutionary patterns including extensive pan-angiosperm retention of certain gene clades, ancient retained tandem duplications and lineage-specific transpositions such as the floral patterning genes in Brassicaceae (Zhao et al., 2017). Our approach can be scaled to analyze not just one gene family, but all gene families across a lineage.

The aim of this study is to investigate and compare the dynamics and properties of the entire synteny networks of all annotated genes for mammals and angiosperms. To this end, we analyzed the syntenic properties of 87 mammalian and 107 plant genomes (Figure 1) which represent most major phylogenetic clades of both mammalian and angiosperm groups across ~170 million years of evolution (Cifelli and Davis, 2003; Bininda-Emonds et al., 2007; Jiao et al., 2011; Magallón et al., 2015). For mammals, the species used covered the three main clades of Afrotheria, Euarchontoglires, and Laurasiatheria, as well as first-branching groups like *Ornithorhynchus anatinus* (platypus). For angiosperms, the species also cover three main groups of Monocots, Superasterids, and Rosids, as well as basal groups such as *Amborella trichopoda* (Figure 1). Some clades are more heavily represented than others such as primates (human relatives) and crucifers (Arabidopsis relatives) due to research sampling biases. Regardless, most major lineages are represented. Also, there are differences in the overall quality and completeness of the genome assemblies used, but this was a factor we wanted to analyze and assess using synteny analysis. We calculate average clustering coefficient for every gene family, and characterize gene functions of highly syntenically conserved versus dynamic. We decomposed the whole network into clusters, analysis of cluster composition, size distribution of all clusters indicate the landscape of specific genomic architecuture rearragements that may related to evolutioanry adaptions.

Results and Discussion

Genome collection, pairwise synteny comparisons

We used fully-sequenced genomes to investigate all syntenic blocks within and across genomes. Initially we searched public databases maintaining mammalian and angiosperms genome resources such as NCBI, Ensembl, CoGe and Phytozome. Candidate genomes had to contain downloadable complete predicted gene models and gene position annotations. Ultimately, we analyzed 87 mammalian genomes, presented according to the consensus species tree adopted from NCBI taxonomy (Figure 1, Supplemental Table S1) which included 1 Prototheia (*Ornithorhynchus anatinus*), 1 Metatheria (*Sarcophilus harrisii*), 1 Xenarthra (*Dasypus novemcinctus*), 6 Afrotheria, 38 Euarchontoglires and 40 Laurasiatheria species. For angiosperms, we analyzed 107 genomes including 1 Amborellaceae (*Amborella trichopoda*), 26 Monocots (including 14 Poaceae) and 80 eudicots (including 1 Proteales (*Nelumbo nucifera*), 23 Superasterids (Asterids and Caryophyllales), and 56 Rosids) (Figure 1, Supplemental Table S1).

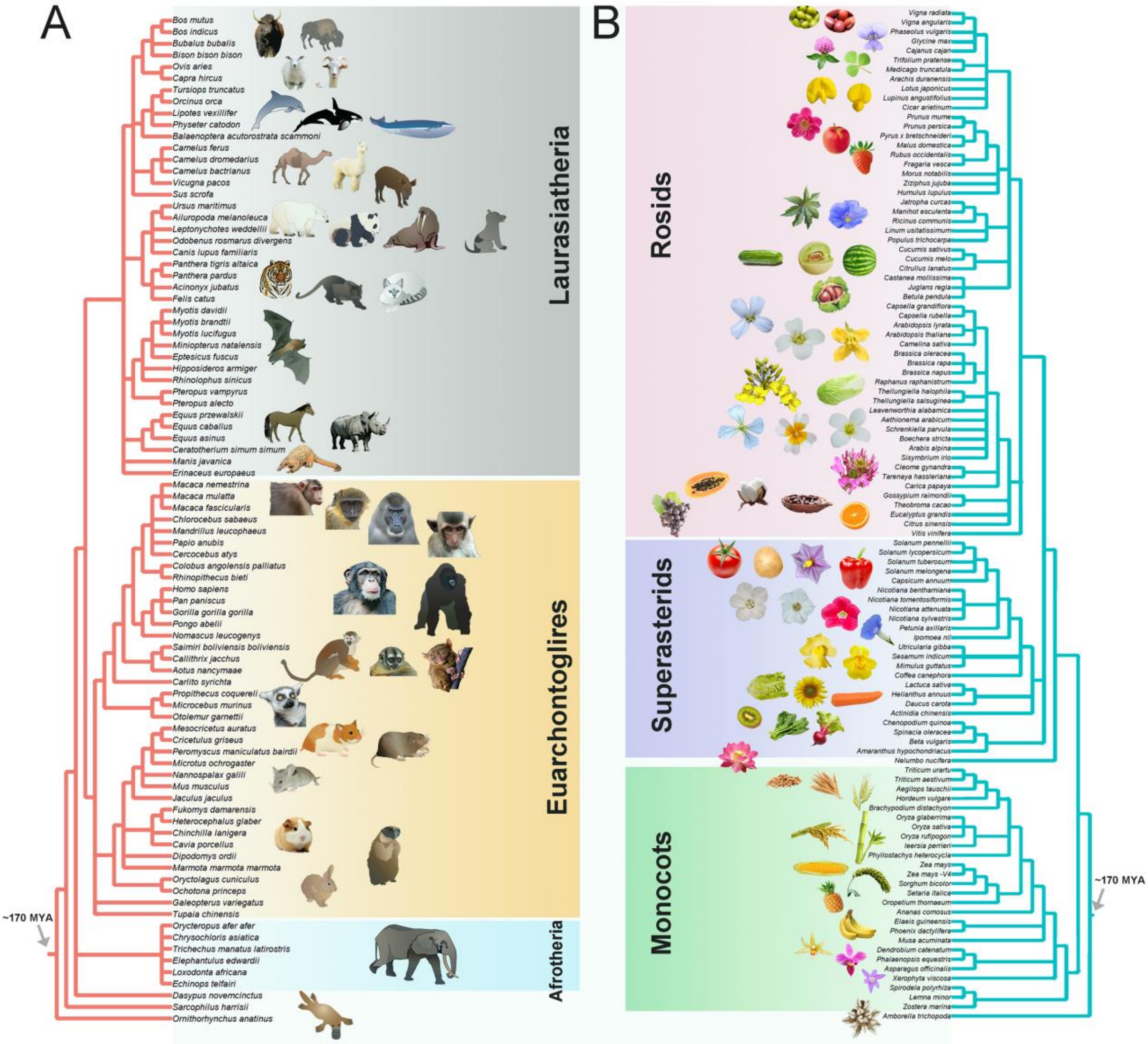


Figure 1 Phylogenetic relationships of mammal and angiosperm genomes analyzed. (A) Mammal genomes used, highlighting the three main placental clades Afrotheria, Euarchontoglires and Laurasiatherias. (B) Angiosperm genomes used, highlighting the three main clades Monocots, Superasterids and Rosids.

We modified all peptide sequence files and genome annotation GFF/BED files with corresponding species abbreviation identifiers, followed by pairwise all-vs-all genome comparisons for synteny block detection (Zhao et al., 2017; Zhao and Schranz, 2017).

To assess the overall impact of phylogenetic distance, genome assembly quality and/or genome complexity, we summarized the number of syntenic gene pairs for all pairwise genome comparisons (3,828 times for mammals and 5,778 times for angiosperms) into color-scaled matrixes (Figure 2) organized using the same species phylogenetic order as in Figure 1.

The diagonal of the matrix represents self- vs. self-contrasts and indicates the number of retained duplicate genes, which is indicative of recent and/or ancient WGDs. The lighter orange and blue rows with fewer syntenic links could reflect key biological or genomic differences, but is much more likely to be due to poor quality genome assemblies. Such as the mammalian genomes of *O. anatinus*, *Galeopterus variegatus*, *Carlito syrichta, Manis javanica*, and *Tursiops truncates* (Figure 2A) and genomes of *Humulus lupulus*, *Triticum urartu*, *Aegilops tauschii*, and *Lemna minor* in angiosperms (Figure 2B).

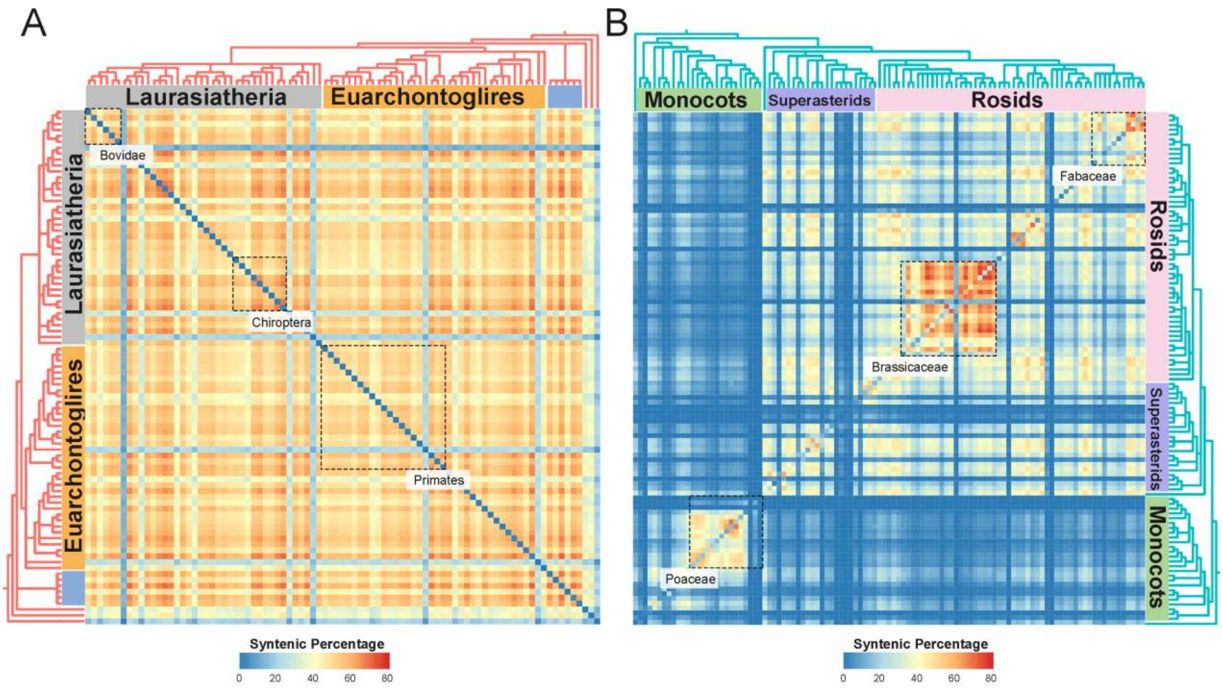


Figure 2 Pairwise synteny comparisons of mammal and angiosperm genomes. (A) Pairwise synteny comparison across Mammal genomes. (B) Pairwise synteny comparison across Angiosperm genomes. The color-scale indicates the syntenic percentage of each comparison. Species are arranged according to the consensus phylogeny (Figure 1). Overall, average synteny is much higher across mammals than plants. Also, there is a stronger phylogenetic signal seen for plant genomes. The method also allows for easy detection of potentially low-quality genomes (overall lower syntenic pair scores). The diagonal for both plots represents intra-genome comparisons which can detect potential recent and ancient WGDs. Note, that almost all plant genomes have higher intra-genome syntenic pair scores than all mammal intra-genome comparisons.

As shown in the matrixes, mammalian genomes overall are in general highly syntenic regardless of phylogenetic distance (Figure 2A) with primate vs primate comparisons showing marginally higher scores. Whereas plant genomes show more phylogenetic signal (e.g. monocots vs monocots and crucifers vs. crucifers), the impact of recent WGD (e.g. *Brassica napus*) and more variability overall (due to assemblies from different groups of researchers, different qualities, multiple independent WGDs) (Figure 2B). Note, that almost all plant genomes have higher intra-genome syntenic pair scores than all mammal intra-genome comparisons. We further checked genome characters by plotting syntenic gene percentage against Pfam annotation percentage for each genome (Supplemental Figure S1). We also compared pairwise synteny coverages to genome quality metrics such as N50s and BUSCO (Supplemental Figures S2, S3)

Based on these results, we removed four poor-quality plant genomes (*H. lupulus*, *T. urartu*, *A. tauschii*, and *L. minor*) before proceeding to the next step of our analyses.

Characterization of synteny networks

The entire synteny networks are composed of all syntenic genes identified within all the syntenic blocks. Specifically, there are 1,453,712 nodes (genes) and 49,035,861edges (syntenic connections between genes) for mammals, and 2,214,712 nodes and 49,035,861edges for angiosperms, respectively. To evaluate genomic conservation of gene families (for gene family assignments see Methods) over evolutionary time scales from the synteny network data, we introduce two estimators: average clustering coefficient (Supplemental Figure S4) and the percentage of genes in the family that are syntenic (syntenic percentage) for every gene-family (Figure 3A). A clustering coefficient is calculated for all nodes in the synteny network, as a measure of the degree to which nodes in a graph tend to cluster together. Genes can be mobilized (e.g. transposed) to other genomic contexts (e.g. unique or lineage-specific contexts) and thus will no longer be collinear or syntenic to other species or lineages. Thus, we use percentage (gene family members in the network/ total gene family members in the genomes) to quantify the proportion of the genes retaining synteny.

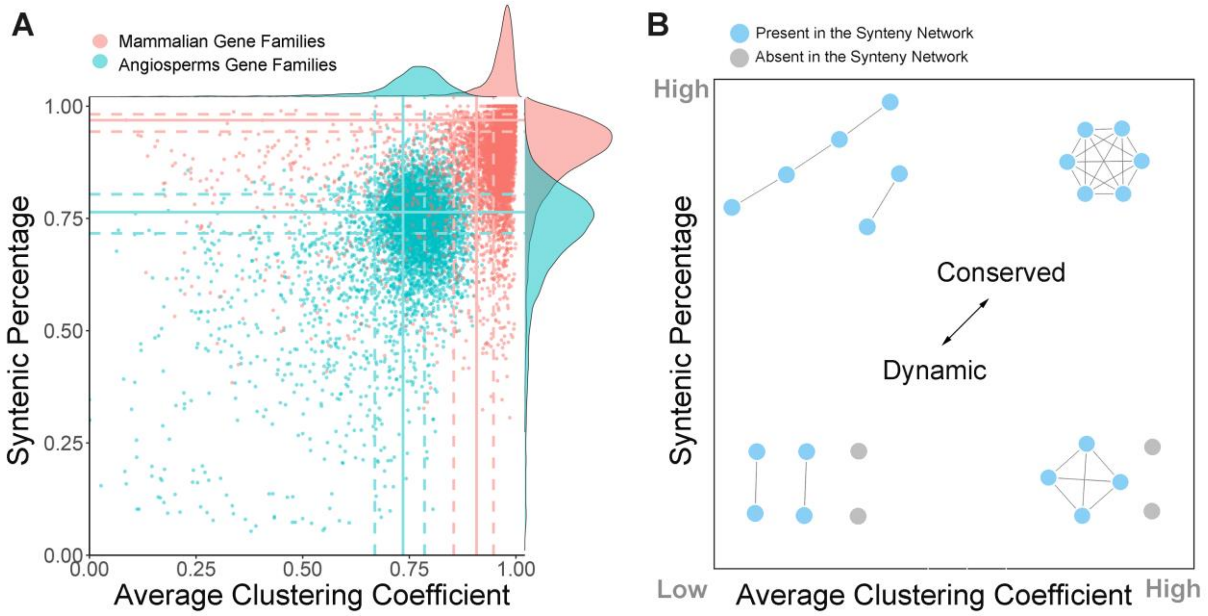


Figure 3 Network properties of gene families from mammal and angiosperm genomes. (A) Distributions of gene family dynamics of mammal (11,830 in red) and angiosperm (10,617 in blue) gene families plotted using percentage of syntenic genes and average clustering coefficients per family. Quartiles of average clustering coefficient and syntenic percentage for both mammals and angiosperms are indicated by dashed (25%/75%) and solid (median) lines. (B) Conceptual model depicting different patterns of synteny network connectivity, according to data distribution, with further analysis based on 25% quartiles.

We then plotted the average clustering coefficient and retention percentage of all the gene families for the mammalian (11,830 gene families) and angiosperm (10,617 gene families) synteny networks (Figure 3A). Mammalian gene families overall have significantly higher clustering coefficients (mean 0.92 for mammals compared to 0.72 for angiosperms; *P* < 0.001, Wilcoxon-Mann-Whitney test) and retention percentage (mean 0.88 for mammals compared to 0.71 for angiosperm; *P* < 0.001, Wilcoxon-Mann-Whitney test) than that of angiosperms (Figure 3A). This confirms that over large evolutionary time scales, genomic context is generally more conserved and constrained in mammals than for angiosperms.

Syntenic dynamics of all gene families could be classified and compared to other gene families by our C-P (Clustering coefficient vs Percentage) quartile analysis method, as conceptually depicted in Figure 3B. We defined values of the top 25% quartile as “high”, and the bottom 25% quartile as “low” for both mammals and angiosperms. The resulting four categories are highlighted (Figure 3B). The high clustering coefficient plus high retention percentage in the synteny network (“high-high” C-P values), indicates the both most syntenically conserved and most completely syntenic gene families, and thus the most inter-connected networks (Figure 3B, Supplemental Table S2). Genes in the category of “high-low” C-P detect gene families where certain gene sub-families and/or phylogenetic clades are highly syntenic, but overall many gene members are absent from the clusters (thus a low percentage). Non-syntenically connected gene family members may be prone to transposition (Figure 3B, Supplemental Table S2). In contrast, the category “low-high” C-P means that a high proportion of the gene family members are in the network, but not always well connected, for example due to tandem gene cluster expansions (Figure 3B, Supplemental Table S2). Lastly, the category “low-low” C-P represent gene families that are distributed dispersedly (such as across pericentromeric regions) and thus non-syntenic, or represent young transpositions or lineage-specific genes shared only between a small number or related species (Figure 3B, Supplemental Table S2).

Comparative synteny dynamics of gene families of mammals and angiosperms

We investigated if gene families with similar C-P synteny dynamics (high-high, high-low, low-high, and low-low), might also have similar functional annotations (e.g. GO terms) (Jiao and Paterson, 2014; Li et al., 2016). We tested for pathway and gene-function enrichment of gene families within each of the four C-P profiles for both mammals and angiosperms (Supplemental Table S3, Figures S5 and S6). The results shows that for mammals, gene families with “high-high” profiles are functionally enriched in DNA metabolic processes (such as “DNA replication” and “DNA repair”), intracellular organelle part, nucleoplasm, and telomere maintenance. (Supplemental Table S3, Supplemental Figure S5). By contrast, “low-low” gene families include functions in immune responses and pathways (e.g. “Biological oxidations”, “detection of chemical stimulus”, and “epoxygenase P450 pathway”), transmembrane receptor activities (e.g. “signaling receptor activity”, “G-protein coupled receptor activity”, and “olfactory receptor activity”), enriched protein classes are “antibacterial response protein”, “oxygenase”, “cytokine receptor”, and “defense/immunity protein” (Supplemental Table S3, Supplemental Figure S5). The mammalian “high-low” group is enriched for genes that function in DNA-templated gene transcription and DNA binding, such as KRAB box zinc finger transcription factors (Imbeault et al., 2017) (Supplemental Table S3, Supplemental Figure S5). Transcription factors bind specific promoters and thus regulate a variety of developmental and environmental processes. Moreover, transcription factors commonly consist of multiple members. Thus, it can be hypothesized that some gene family members are highly conserved and genomically constrained, while other members are versatile and transposed into new genomic positions. Finally the “low-high” group is enriched for genes involved in translation (e.g. “peptide biosynthetic process”, “peptide metabolic process”) and ribosomal component (e.g. “ribosomal subunit”, “ribonucleoprotein complex”), most enriched Reactome Pathways are closely related to translation processes (e.g. “eukaryotic translation”, “Cap-dependent translation initiation”), as well as infectious disease related pathways (e.g. “Influenza infection”, “Influenza life cycle”, and “Influenza viral RNA transcription and replication”) (Supplemental Table S3, Supplemental Figure S5).

The functional enrichment analysis of angiosperms shows a different pattern than for mammals (Supplemental Table S3, Supplemental Figure S6). Plant “high-high” gene families are enriched for organelle components (e.g. “organelle part”, “intracellular organelle”, “chloroplast part”, “organelle organization”, and “plastid part”), as well as acetyltransferase, transferase and methyltransferase proteins for the processes such as “DNA repair”, “ncRNA metabolic process” and “methylation” (Supplemental Table S3, Supplemental Figure S6). Many of these categories are plant-specific related to photosynthesis. By contrast, the plant “low-low” group is enriched for genes functions in biological oxidation and defense responses such as “secondary metabolic process”, “monooxygenase activity”, “UDP-glucosyltransferase activity”, and “Cytochrome P450s” (Supplemental Table S3, Supplemental Figure S6). “Low-high” gene families function in nuclear part components (e.g. “intracellular organelle lumen”, “organelle lumen”), biosynthetic process (e.g. “organonitrogen compound biosynthetic process”, “cellular aromatic compound metabolic process”), and gene expression (e.g. “RNA polymerase complex”, “nucleic acid binding”, “RNA polymerase II transcription initiation”) (Supplemental Table S3, Supplemental Figure S6). Similar to mammalian “high-low” C-P families, angiosperms “high-low” genes function in positive regulation of transcription (e.g. “RNA polymerase II regulatory region DNA binding”, “transcription factors”), interestingly MADS-box transcription factors controlling floral development also overrepresented. (Supplemental Table S3, Supplemental Figure S6).

Classifying and characterizing gene families according to their “synteny network C-P” scores allows for the relative comparisons of any gene family to all others across a lineage. The degree of conservation likely reflects functional constraints of the family. For example, gene families with a “high-high” C-P are responsible for fundamental functions (i.e. DNA repair and photosynthesis.) and “low-low” C-P gene families are highly mobile and functionally flexible (such as both animal and plant NLR family defense-related receptors (Jones et al., 2016) and plant P450s and F-box genes) (Supplemental Table S3).

Comparative phylogenomic profiling of synteny clusters

We next performed a clustering analysis for the entire mammalian and angiosperm synteny networks. We used Infomap (Rosvall and Bergstrom, 2008; Lancichinetti and Fortunato, 2009) as the clustering algorithm due to its efficiency and accuracy in handling large graphs with millions of nodes. To visualize and understand genomic diversity, we performed phylogenomic profiling of all synteny clusters of mammals and angiosperms (Figures 4A and 4B). Blue columns indicate conserved single copy syntenic clusters, orange columns indicate retained duplicate copy clusters (i.e. conserved ohnologs from WGD), and the red columns signify conserved clusters with more than two copies (e.g. conserved tandem clusters) (Figures 4A and 4B). Nearly empty rows of the less-syntenic species are consistent with the pairwise matrix in Figure 2.

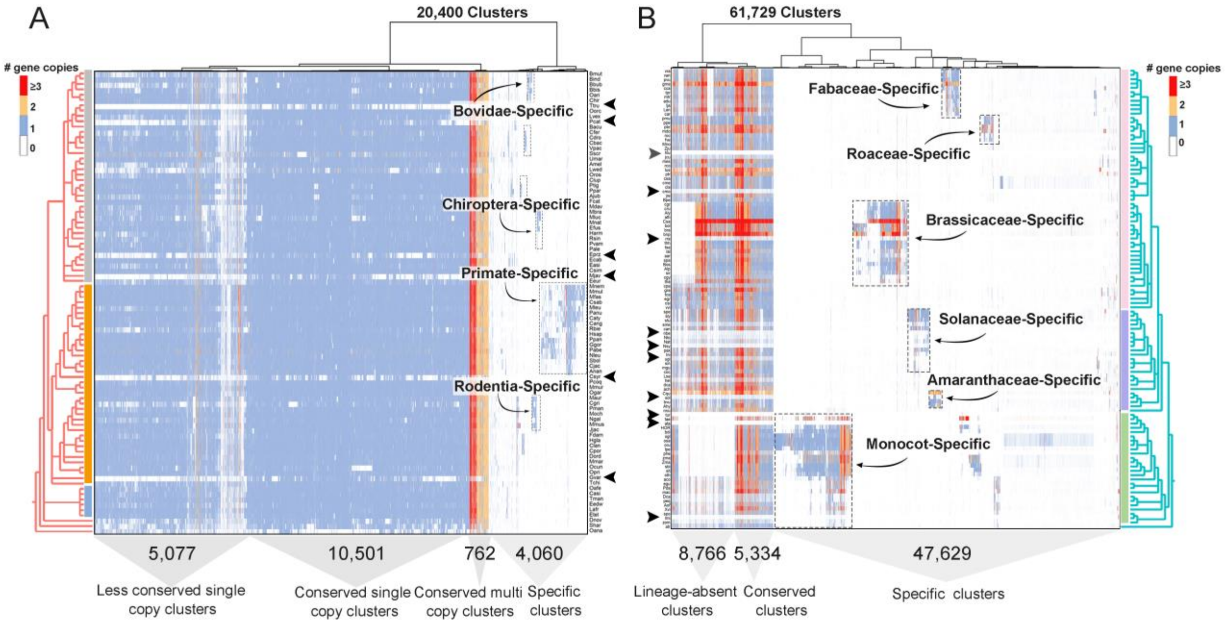


Figure 4 Phylogenomic profiling of all synteny clusters for mammal and angiosperm genomes. (A) Phylogenomic profiling of all mammalian clusters (size > 2). Groups of lineage-specific clusters are boxed and labeled. (B) Phylogenomic profiling of all angiosperm clusters (size > 2). Groups of lineage-specific clusters are boxed and labeled.

For mammals, a very large proportion of all genes are syntenic and single copy (Figure 4A) as mentioned above. Smaller proportions of mammalian genomes are conserved and syntenic for duplicates or larger conserved multi-gene families. Interestingly, lineage-specific clusters were observed for most of the included mammalian clades. For example, we found lineage-specific clusters for Primates (such as the CENPJ example discussed above), Rodentia, Vespertilionidae, Felidae, Camelidae, and Bovidae (Figure 4A).

In contrast, in angiosperms less than 10% of clusters are syntenically conserved between eudicot and monocot species (Figure 4B). The remaining clusters are mostly lineage-specific clusters that appear as discrete columns (Figure 4B). This indicates that angiosperm genomes are highly fractionated and reshuffled, with abundant examples of specific clusters for particular phylogenetic lineages/plant families, such as Amaranthaceae, Brassicaceae, Poaceae, Fabaceae, Rosaceae, and Solanaceae (Figure 4B). Results also highlight species with more gene copies per cluster (e.g. orange/red rows), likely due to recent WGD events such as for *G. max*, *B. napus* and *P. trichocarpa* (Figure 4B).

Traditional phylogenomic profiling data typically show only the presence/absence of a gene family. In contrast, our synteny-based phylogenomic profiling is based on conserved genomic collinearity of gene families across lineages which provides potential novel information about changes of genomic context (transpositions and/or expansions) or the origin of “novel genes” of specific gene families. Such changes in genomic context provide intriguing candidate gene sets for investigating trait evolution.

Comparative synteny network clustering

We further summarized and compared the clustering results for mammals and angiosperms in terms of cluster-size distributions (Figures 5A and 5B), corresponding clustering coefficients (Figures 5C and 5D), and number of species included per cluster (Figures 5E and 5F).

Mammalian genomes have a prevalent peak of syntenic gene families that are present only once per taxa (single copy orthologous gene cluster peak shaded in cyan, Figure 4A). To the right, there is a second modest peak of duplicated (ohnolog) genes due to the ancient 2R WGD events (shaded in bright yellow, Figure 5A). These two peaks could be further explained by Figures 5C and 5E that depict the corresponding average clustering coefficient and number of species, respectively. We observe that the peak in cyan in Figure 5A is accompanied by a steady increasing trend of the clustering coefficient and the number of species involved (Figure 5C). A similar trend was observed for the clusters forming the peak in yellow due to WGD (Figure 5A). On the far left there is the rather modest proportion of lineage specific genes (clusters of syntenic genes between only a subset of mammalian species or clade(s) (shaded in purple, Figure 5A). On the far right are large multigene clusters usually with multiple syntenic gene copies conserved across multiple species due to tandem duplications such the well-known Hox-genes (shaded in olive green, Figure 5A). Representative examples are labeled on the curve, and further depicted in Figures 5G and 5H.

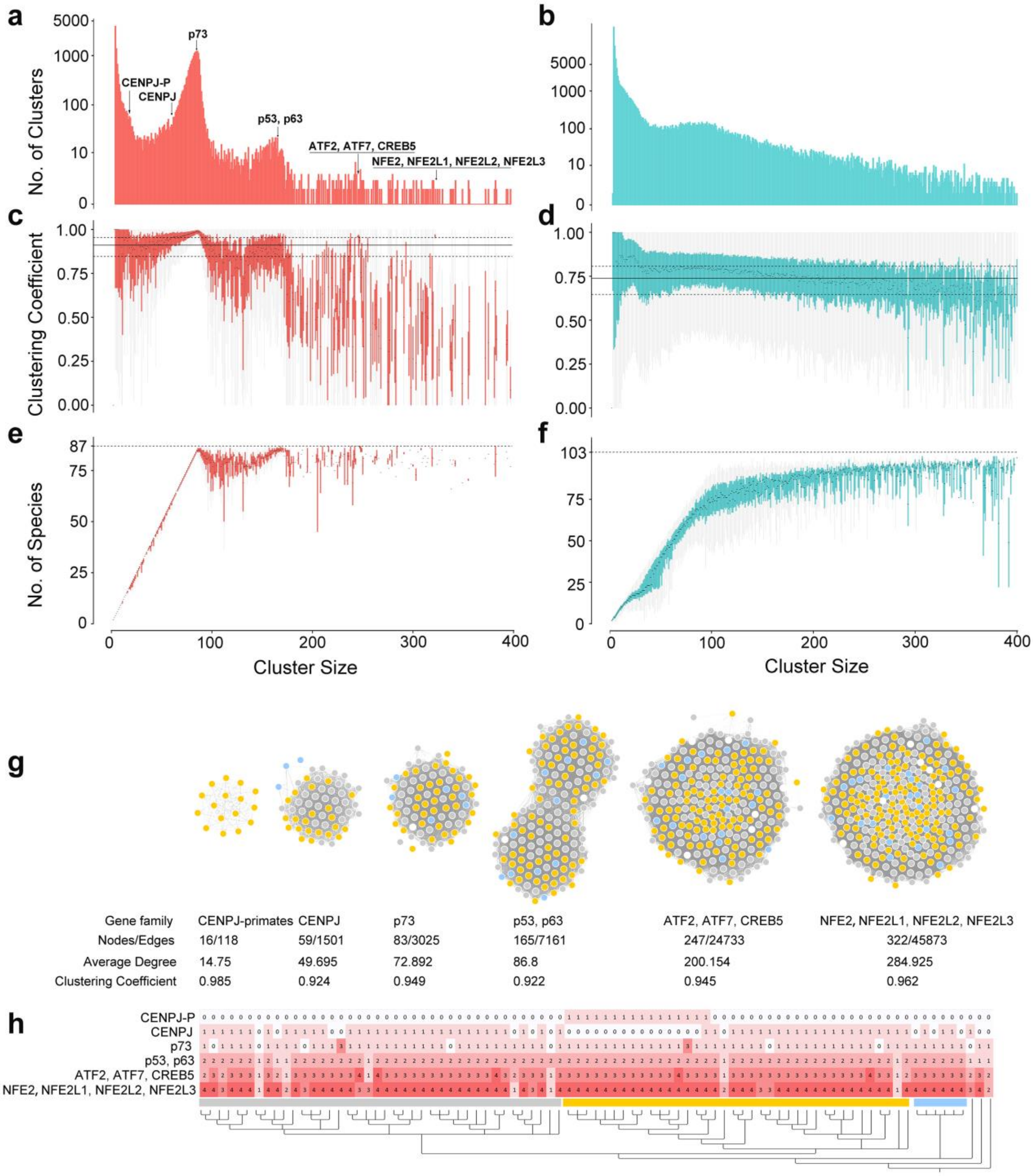


Figure 5 Synteny cluster statistics of mammal and angiosperm genomes and representative mammalian synteny clusters. Approximate size ranges for clusters of lineage-specific, conserved, WGD replicates, and large tandem genes are shaded in purple, cyan, yellow, and olive green, respectively. (A) Sizes distribution of all mammalian gene syntenic clusters. Representative examples are pointed and labeled on the curve. (B) Sizes distribution of all angiosperms gene syntenic clusters (C) Boxplot of clustering coefficient by mammalian cluster sizes. (D) Boxplot of clustering coefficient by angiosperm cluster sizes. (E) Number of involving genomes for mammalian clusters by cluster sizes. (F) Number of involving genomes for angiosperm clusters by cluster sizes. (G) Six representative and diverse mammalian clusters of *CENPJ* (primate-specific one and the others), *p73*, *p53-p63*, *ATF2-ATF7-CREB5*, and *NFE2-NFE2L1-NFE2L2-NFE2L3*. Total number of nodes, edges, average degree, and clustering

coefficient are indicated accordingly below. (H) Phylogenomic profiling of the clusters from (G), a color gradient of red indicates the number of syntelogs in each species.

In contrast, angiosperm genomes show a very large proportion of lineage-specific clusters on the far left (shaded in purple, Figure 5B). The clustering coefficients for these clusters is often above the threshold of "high" (top 25%, which was defined earlier for the C-P classification) (Figure 5D) and the cluster size for these lineage-specific clusters is mostly between 10 to 30 (shaded in cyan, Figure 5F), reflecting the number of species and gene copies within particular phylogenetic groups such as Fabaceae, Brassicaceae, and Poaceae. Next, a rather broad peak of gene clusters are observed that are conserved across many lineages (Figure 5B) of genes that are single-copy in some lineages and in two/more copies in other lineages due to WGD. Also, there is a larger proportion of large multigene families seen to the far right (shaded in olive green, Figure 5B). There is a variation for the number of species per cluster for these large multi-gene families in angiosperms (Figure 5F).

The combination of cluster size, corresponding clustering coefficient, and number of involved species were used to select representative synteny clusters for mammals. As an example of a lineage-specific cluster we show *CENPJ* (as an example an of a primate lineage-specific cluster), *p73* as an example of a single copy conserved cluster, *p53-p63* as an example of 2-ohnologs-retained WGD cluster, *ATF2-ATF7-CREB5* as an example of 3-ohnolog-retained WGD cluster, and *NFE2-NFE2L1-NFE2L2-NFE2L3* as example of 4-ohnolog-retained WGD cluster (Figures 5A, 5G and 5H). It has been reported that *CENPJ* regulates brain size (Bond et al., 2005; Gul et al., 2006), and primates have relatively larger brains (Kudo and Dunbar, 2001; Byrne and Corp, 2004). It is interesting that we found primates formed a lineage-specific *CENPJ* synteny cluster (Figures 5G and 5H) compared to other mammals. This indicates that *CENPJ* underwent a gene transposition event at or near the divergence of the primate ancestor from other mammals. Thus, the primate gene copy is in a unique genomic context facilitating potential new/altered regulatory patterns and gene functions. The *p53, p63* and *p73* genes compose a family of transcription factors involved in cell response to stress and development (Levrero et al., 2000; Murray-Zmijewski et al., 2006). *p63* is previously perceived close related to *p73* because of the similar protein domain compositions, however our result shows *p63* and *p53* are ohnolog duplicates retained after WGD. Other ohnolog clusters with strong support from our analyses include *ATF2-ATF7-CREB5*, transcription factors with broad roles such as activating CRE-dependent transcription, cancer progression and immunological memory (Gupta et al., 1995; Bhoumik et al., 2008; Gozdecka and Breitwieser, 2012; Yoshida et al., 2015) and *NFE2-NFE2L1-NFE2L2-NFE2L3*, also with broad roles such as regulation of oxidative stress, aging and cancer cell proliferation (Kobayashi et al., 1999; Sykiotis and Bohmann, 2008; Chowdhury et al., 2017), notably *NFE2L1* was recently reported to secure cellular protein quality control under cold adaptions by regulating brown adipose tissue (BAT) thermogenic function (Bartelt et al., 2018). As a comparative example of an entire gene family for both mammals and plants, we give the complete homeodomain (Schena and Davis, 1992; Krumlauf, 1994) gene families for both lineages (Figures S7 and S8). We clearly show and verify that the mammalian Hox genes appear as interconnected synteny super-clusters and also find synteny connections to the ParaHox genes, consistent with the numerous previous reports (Brooke et al., 1998; Ferrier and Holland, 2001; Lemons and McGinnis, 2006) (Supplemental Figure S7). In contrast, for plants we did not find any prominent tandem origin of homeobox clades, but did identify several examples of WGD-derived gene expansions and family-specific transpositions (Supplemental Figure S8).

Conclusions

Synteny analysis of multi-species genomics datasets has led to major advances in our understanding of evolutionary patterns and processes. However, few studies have systematically assessed and compared genomic properties across kingdoms (Murat et al., 2012). Synteny network statistical parameters provide new possibilities for systematically evaluating gene (syntenic) diversification and/or conservation patterns over long evolutionary time scales. In this study, we have presented an analytic framework for large-scale synteny comparisons using network analysis of all suitable mammalian and angiosperm genomes. Assessment metrics based on synteny intuitively illustrate genome contiguity and copy number depth due to (paleo)polyploidy. The C-P method provides a means to characterize gene family dynamics in a comparative evolutionary context. We have displayed and compared features of all synteny clusters from these two important lineages and performed their clade-wide phylogenomic profiling. The results illustrate the dramatic differences in genomic dynamics within and between the two groups, exemplified by synteny networks of primate-specific gene transpositions (i.e. *CENPJ*), extant ohnologs surviving 2R of mammals, and for all mammal and angiosperm homeobox genes.

Dissection of the properties of all synteny clusters provides intriguing insights into the differing genomic architectures and dynamics of mammal and flowering plants. Examples in this study are just the tip of the iceberg. Much remains to be explored, but this study provides an intriguing foundation for future investigations to better understand genome evolution and elucidate regulatory mechanisms underlying diverse evolutionary biological processes. Such approach can further be extended to other phylogenetic groups and deeper evolutionary time scales.

Methods

Genome resources

All reference genomes were downloaded from public repositories (Supplemental Table S1). For each genome, we needed a FASTA format file containing peptide sequences of all predicted gene models, as well as a genome annotation file (GFF/BED) showing the positions of all the genes. Original gene names in the FASTA file have been modified into a prefix (unique identifier indicating species) and numeric GenBank gene ID. An in-house script was used for batch downloading genomes and modifying gene names.

All mammalian genomes were downloaded from NCBI. Initially we utilized the total list of available mammal genomes on NCBI (https://www.ncbi.nlm.nih.gov/genome/browse/). Using the list with our script, some records did not contain the complete required information for our analysis (i.e. no genome annotation files, or no FASTA file of total peptide sequences). In the end, we retrieved 87 mammalian genomes suitable for our analysis. Angiosperm genomes were collected from various public databases such as Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html (Supplemental Table S1).)

Peptide sequence annotation

For gene family annotation, we used HMMER (hmmscan) to perform domain annotations against the Pfam database (version downloaded: Pfam 30.0, Pfam-A with 16,306 entries) for all the peptides of the utilized genomes. Domains identified from one sequence were combined, and used for gene family annotation. Multiple occurrences of the identical domain within one protein were counted only once.

Pairwise comparison, synteny blocks detection, and network construction

RAPSearch2 was used to perform all inter- and intra- pairwise all-vs-all protein similarity searches. MCScanX was used for synteny block detection with default settings (window size: 50, number of match genes: >= 5). All outputting collinear files were integrated and curated into one tabular-format file, each row contains information about “Block_ID”, “Block_Score”, and syntenic gene pairs. This file creates a database which contains the entire syntenic nodes and syntenic connections derived from the input genomes. Detail procedures can be referred to a Github tutorial (https://github.com/zhaotao1987/SynNet-Pipeline).

Network statistics

Network statistical analysis was carried out in the R environment (http://www.r-project.org), using the R package “igraph” (Csardi and Nepusz, 2006). We performed the analysis of the networks of mammal genomes and angiosperm genomes separately. The entire network must first be simplified to reduce duplicated edges (same syntenic pair may be derived from multiple detections), followed by the calculation of clustering coefficient, and node degree of each node.

We mapped gene family annotations to all the nodes, and computed the percentage for each gene family using its total occurrence in the synteny network against its total occurrence from the step “Peptide sequence Annotation“. We filtered gene families with at least 50 nodes and plot percentage against average clustering coefficient for all these gene families. Quartiles of percentage and average clustering coefficient was estimated according to their distributions. We describe values over Q3 (highest 25%) as high, and values below Q1 (lowest 25%) as low.

Gene annotation enrichment analysis

Gene families of special interest (“high-high”, “high-low”, “low-high”, and “low-low”) were extracted from the total analysis. We then mapped gene(s) from the model species *H. sapiens* (for mammals) or *A. thaliana* (for angiosperms) to each of the gene families. We then performed online PANTHER overrepresentation test (http://pantherdb.org/) for each of the gene lists, with Bonferroni correction for multiple testing. In addition to the annotation of GO enrichment (biological process, molecular function, and celluar component), we also included analysis of “Reactome pathways”, “PANTHER pathways”, and “PANTHER protein class”. Results containing significant enriched terms was downloaded and illustrated as word clouds, by the R package “tagcloud”. Font sizes determined by “-log10(p-value)”. We depicted a maximum of the top 40 most significant terms.

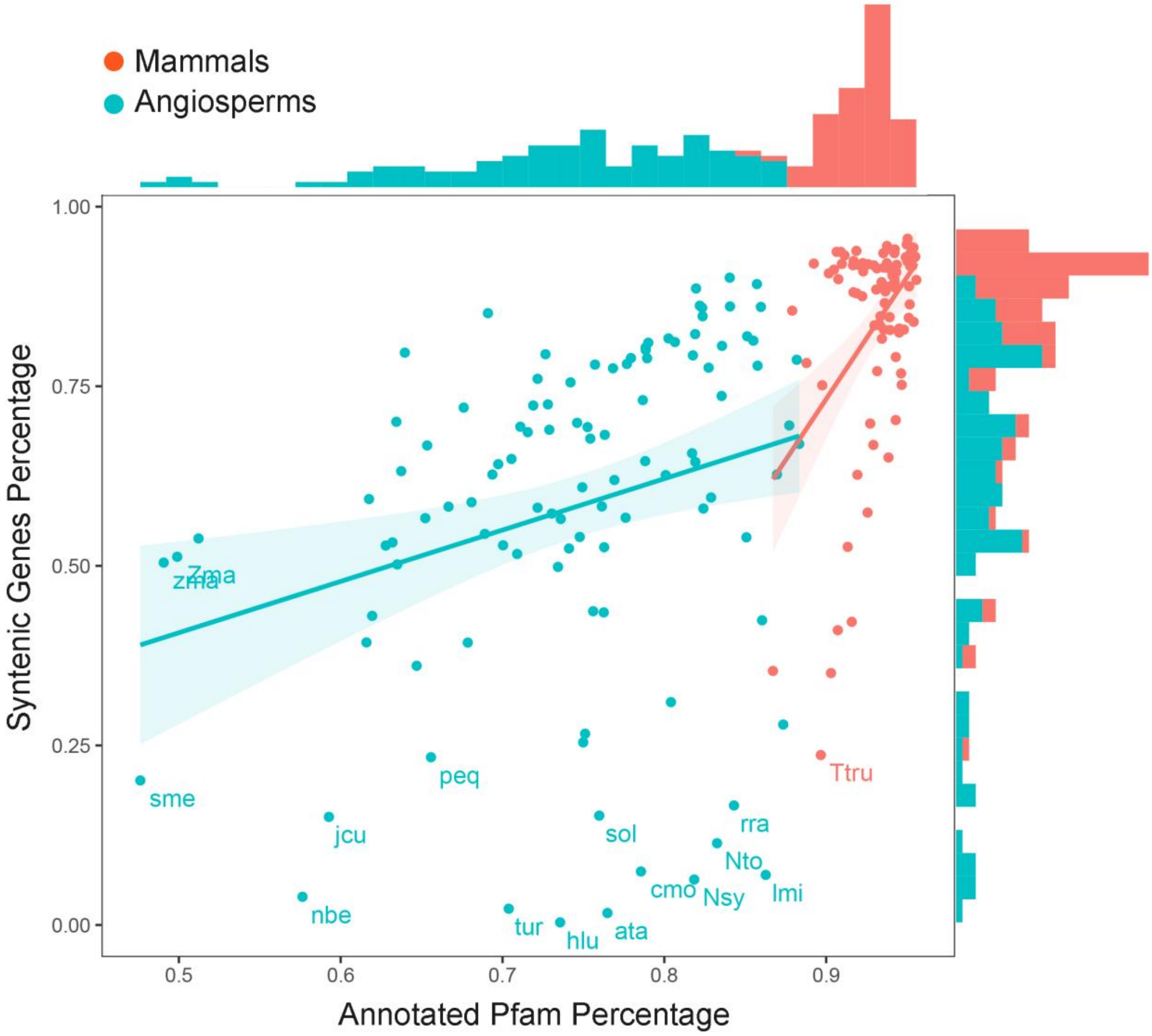
Network clustering and phylogenomic profiling

We used the infomap method to split the entire network, consisting of millions of nodes, into clusters (Rosvall and Bergstrom, 2008). Clustering results were determined by topological edge connections, edges were unweighted and undirected. All synteny clusters were decomposed into numbers of involved syntenic gene copies in each genome. Dissimilarity index of all clusters was calculated using the “Jaccard” method of the vegan package (Dixon, 2003), then hierarchically clustered by “ward.D”, and visualized by “pheatmap”. We illustrate all the clusters of mammals and angiosperm respectively with cluster size >2.

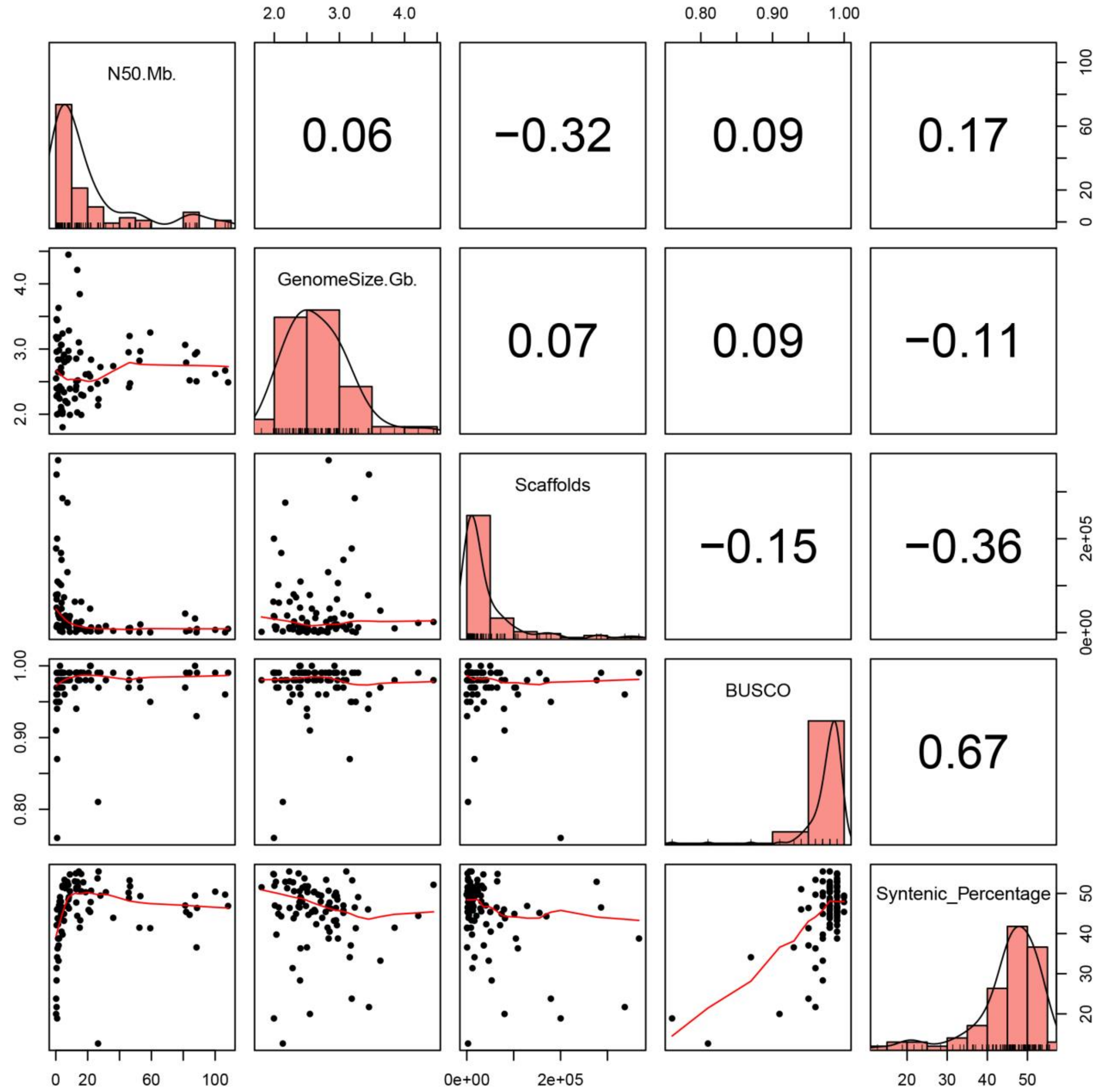
Acknowledgements

TZ was supported by China Scholarship Council. Mammal symbols in Figure 1 courtesy of the Integration and Application Network (ian.umces.edu/symbols).

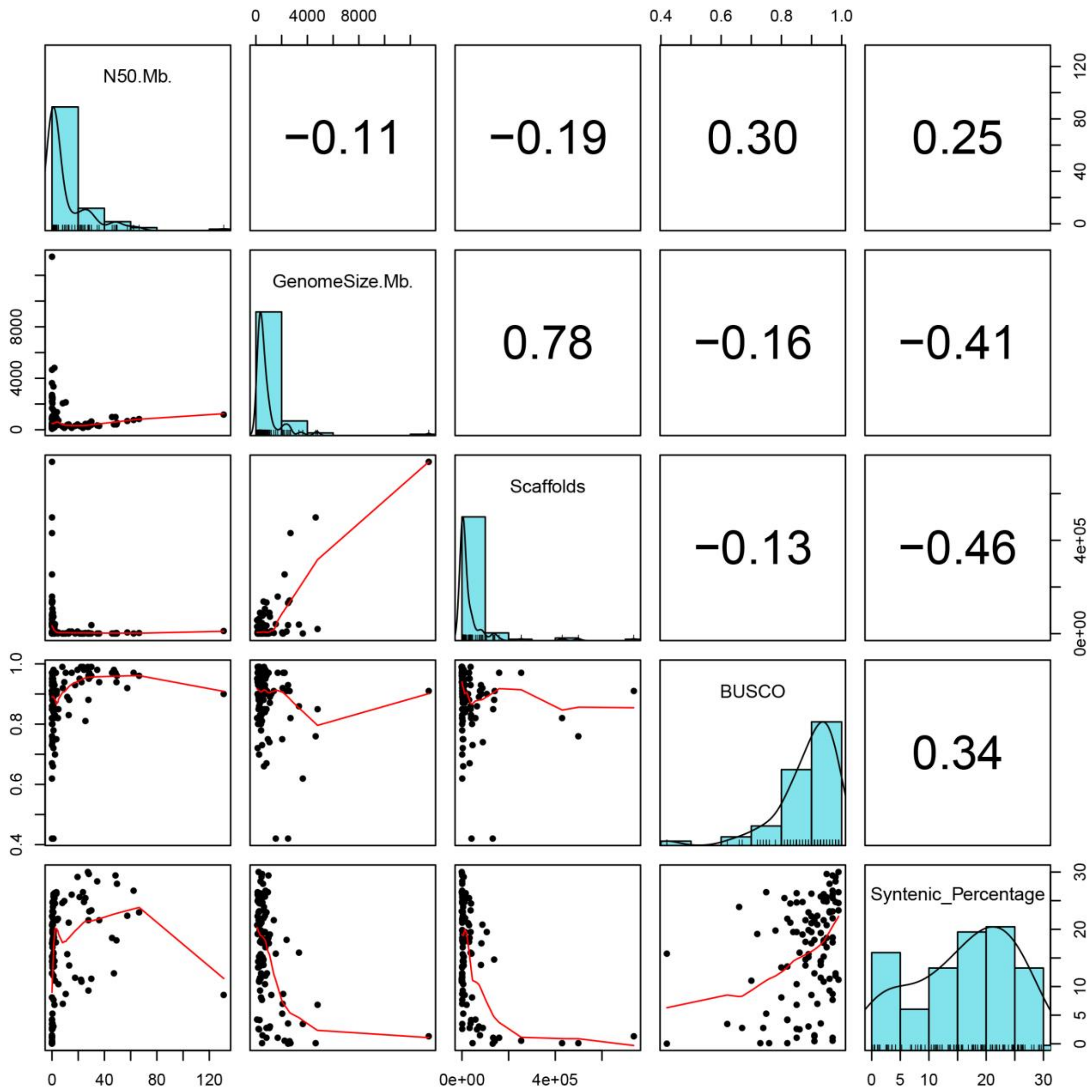
Supplemental Information



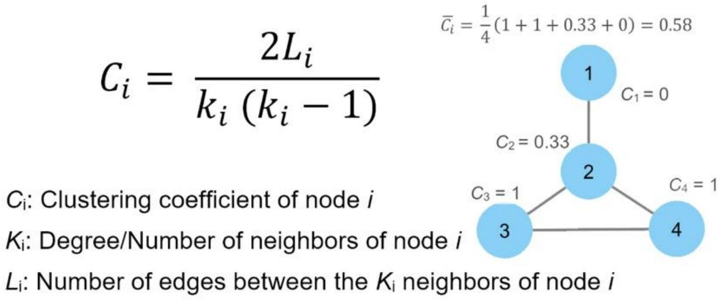
Supplemental Figure S1 Plot of percentage syntenic genes again annotated (by Pfam) percentage of all genomes. Species were highlighted with abbreviated names if syntenic genes percentage lower than 0.25 or annotated proteins (by Pfam) lower than 0.5.



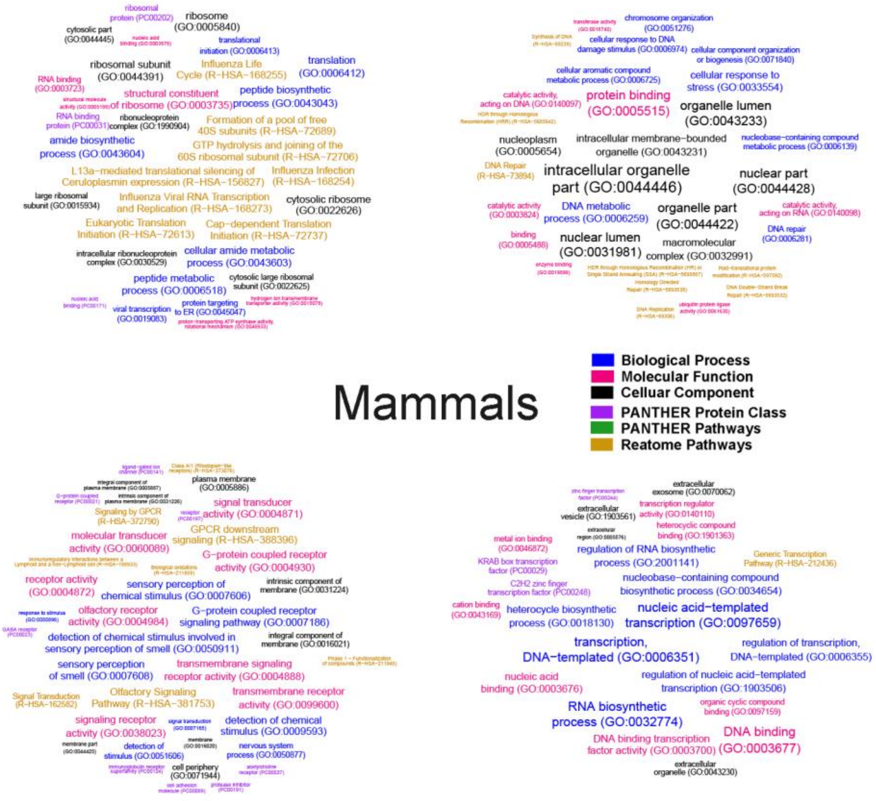
Supplemental Figure S2 Pairwise comparisons of genome metrics of mammal genomes, including N50, genome size, scaffolds, BUSCO, and average syntenic percentage.



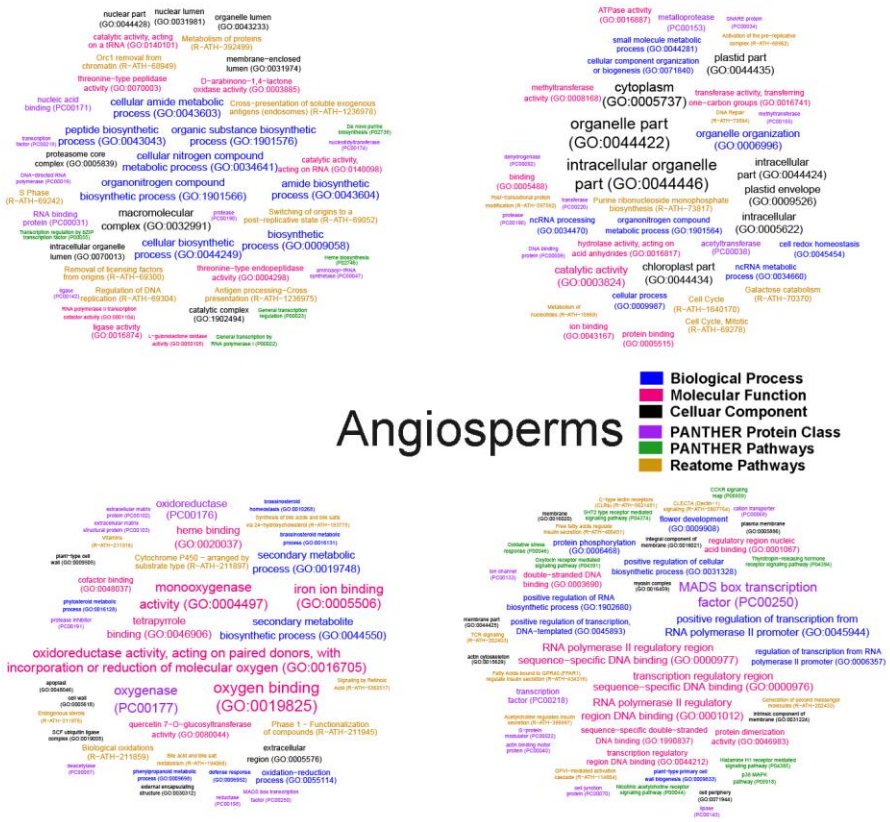
Supplemental Figure S3 Pairwise comparisons of genome metrics of angiosperm genomes, including N50, genome size, scaffolds, BUSCO, and average syntenic percentage.


$C_i = \frac{2L_i}{k_i\,(k_i - 1)}$
$\overline{C_i} = \frac{1}{4}(1 + 1 + 0.33 + 0) = 0.58$
1
$C_1 = 0$
$C_2 = 0.33$
2
$C_3 = 1$
$C_4 = 1$
3
4
C_i: Clustering coefficient of node i
K_i: Degree/Number of neighbors of node i
L_i: Number of edges between the K_i neighbors of node i

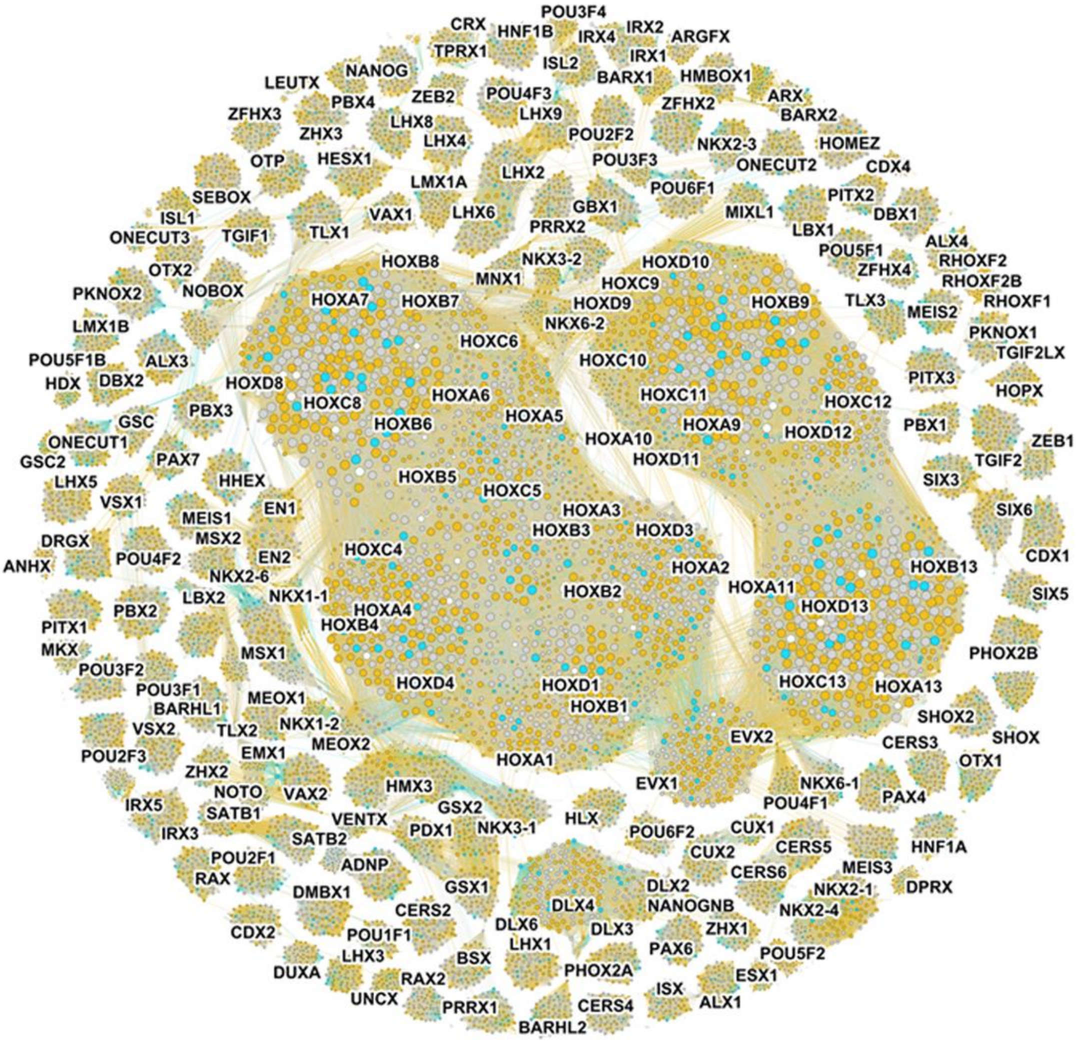
Supplemental Figure S4 Schematic diagram for the calculation of the average clustering coefficient.



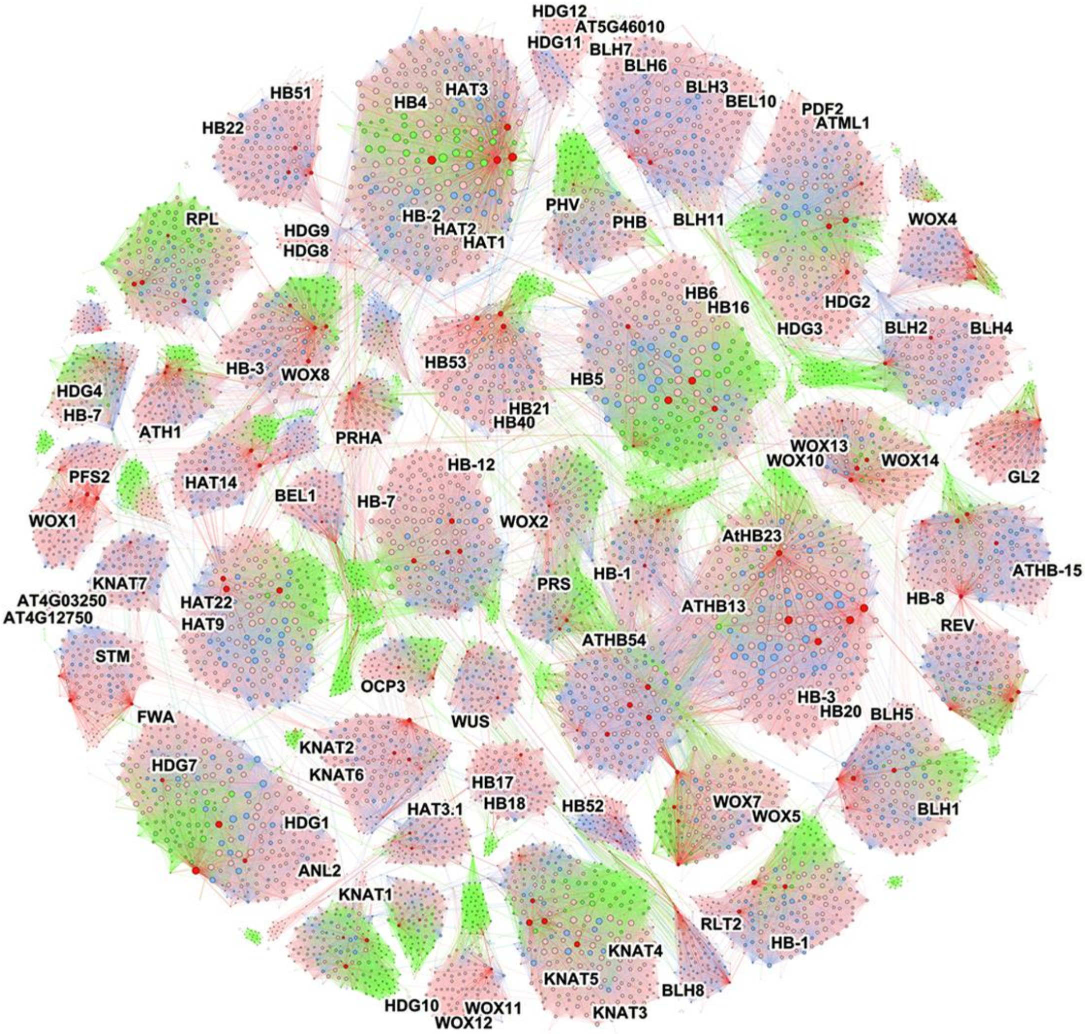
Supplemental Figure S5 Comparative word clouds based on upper and lower quartiles for functional enrichment of significant terms with representative C-P profiles for mammals. Font sizes are representative of adjusted p-values.



Supplemental Figure S6 Comparative word clouds based on upper and lower quartiles for functional enrichment of significant terms with representative C-P profiles for angiosperms. Font sizes are representative of adjusted p-values.



Supplemental Figure S7 Synteny network of all homeodomain proteins for mammals, representative *H. sapiens* are labeled. The well-known Hox clusters, derived from WGD and tandem duplications, were visualized as two huge clusters (*Hox1-8* and *Hox9-13*) connected by EVX gene cluster (*EVX1* and *EVX2*). ParaHox genes *PDX1*, *GSX1*, and *GSX2* form one highly inter-connected cluster, while the other three ParaHox genes *CDX1*, *CDX2*, and *CDX3* form respective independent clusters. Moreover, we have found the synteny cluster of *DLX1-4*, and *DLX6*, cluster of *LHX2*, *6*, and *9*, cluster of *NKX2-1* and *2-4*, and cluster of *CERS5* and *6*.



Supplemental Figure S8 Synteny network of all homeodomain proteins for angiosperms, representative *A. thaliana* genes are labeled. Some examples include conserved clusters (*OCP3*, *RPL*, and *ATH1*); WGD-derived clusters (*KNAT3-5, HAT1-3-HB2-HB4, HDG1-HDG7-ANL2-FWA,* and *HDG2-HDG3-PDF2-ATML1*); eudicot-specific clusters (*STM, KNAT7, KNAT2-KNAT6, WOX1-PFS2* and *HB22-HB51*), and monocot-specific clusters (i.e. *Os01g60270, Os06g04850, Os08g19590*).

Supplemental Table S1 Mammalian and angiosperm genomes used in this study.

Supplemental Table S2 Gene families with significant C-P features of mammals and angiosperms.

Supplemental Table S3 Gene function enrichment for gene families with distinguished C-P profiles of Mammals and Angiosperms.

Data deposition

Data-sets and computer code used in this study are available at DataVerse: (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BDMA7A). This includes the modified FASTA and BED/GFF files of all mammal and angiosperm reference genomes. The scripts for network database preparation (pairwise comparison, synteny block detection, and data integration), Pfam domain annotation, network clustering and statistics, phylogenomic profiling, and for the figure preparation (if applicable) are all included.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TZ and MES designed the study, TZ assembled the genomic data and performed the analysis. TZ and MES wrote the paper. All authors read and approved the final manuscript.

Chapter 6

General Discussion

A lot more genomes are coming…

Genome sequencing has advanced rapidly in the last decades, and thus the number of reference genomes is mushrooming. With richer information, comparative genomics may hold the key to clarify mechanisms that generated the great diversity of present-day life forms, and to answer some of the most fundamental questions. Toward this end, global joint efforts are being made, for example the 10KP (10,000 Plants) Genome Sequencing Project, as part of the EBP (Earth BioGenome Project, Lewin et al., 2018), will sequence and characterize representative genomes from every major clade of embryophytes, green algae, and protists (excluding fungi) within the next five years (Cheng et al., 2018b).

Are we ready for the forthcoming enormous volume of genomes? In comparative genomics, shared synteny provides important inference of genomic architecture variations. However, to compare synteny efficiently across many large eukaryotic genomes is a burning question. In this thesis, I discuss and propose an easy-to-follow systematic approach to organize phylogenomic synteny data by using networks. As shown in Chapters 2-5, synteny networks can nicely reflect and summarize the syntenic conservation and diversification status of any genes of interest across many genomes. Such a way of data organization frees us from integrating and aligning multiple synteny blocks that comes along with the newly added genomes.

Synteny networks: not only a representation

Gene duplications made large gene families (Ohno et al., 1968; Zhang, 2003). A combination of different duplications at different times has shaped the present-day gene family members in the genome. Phylogenetic tree reconstruction using the coding sequences is usually an entry point for the evolutionary inference of gene relationships. Now with the extra layer of synteny information, we are able to sketch the trajectory of how subfamilies diversified and are related by genomic context (such as the MADS-box genes AGL6 and AGL2 clades described in Chapter 3). A cursory glance over the cluster topology and gene IDs could help us to quickly inspect the genomic context conservation and diversification patterns. Together with phylogenetic profiling (copy-number profile of all clusters across phylogeny), we can distinguish from certain gene clusters (Chapter 3, Chapter 4) what kind of gene duplications/movements have been taken place.

Besides visualization, network science involves a set of mathematical parameters for characterizing networks. These includes measures of size, density, average degree, diameter, clustering, robustness, and centrality, as well as concepts such as scale-free, preferential attachment, small-world, percolation, link analysis, and associative mixing (Barabási and Pósfai, 2016). All these network parameters provide tools and opportunities to infer phylogenomic synteny in a new way. We could then use these as a framework to quantify and qualify synteny properties of all genes and gene families (Chapter 5). A clearer evolutionary history picture will guide our experimental design to explore gene function diversification and mechanisms.

Although the overall quality of newly published genomes keeps improving with new technologies, many available genomes are still of poor quality. N50 and BUSCO is not always a good indicator of how well the genome has been assembled (Chapter 5). But, one bonus of building synteny networks with many genomes is that, the cumulative effect make this network approach “error-tolerant”, which reduces the impact of several specific bad genomes. Besides we can use this as a metrics to eliminate these genomes from further analysis (Figure 3 Chapter 5).

Which clustering method best fits synteny data?

The entire synteny network could contains millions of nodes, and dozens of millions of edges. To interpret such complex networks, a clustering algorithm is needed to identify clusters or communities based on network topologies. In this thesis, we showed the application of two clustering methods, infomap (Rosvall and Bergstrom, 2008) and k-cliques (Derenyi et al., 2005). Next to that, I compared multiple other clustering methods for synteny network data, including Louvain (Blondel et al., 2008), MCL (Enright et al., 2002), spinglass (Reichardt and Bornholdt, 2006), walktrap (Pons and Latapy, 2005), fastgreedy (Clauset-Newman-Moore) (Clauset et al., 2004), Girvan-Newman (based on edgebetweeness) (Girvan and Newman, 2002), BigClam (Yang and Leskovec, 2013), kcores (Batagelj and Zaversnik, 2003), GCE (Lee et al., 2010), and cliquesmain (Leskovec and Sosič, 2016). These algorithms vary greatly at speed and quality, and were originally designed for community inference for different types of network data. So what is the best method for clustering synteny network data? To answer this question with a precise benchmarking is hard, because there is no ground-truth synteny data across multiple species for benchmarking. This is also an outstanding issue when comparing synteny detection programs. However I tested and compared the clustering results against each other by empirical evidence, using the same synteny network data. I found for example the fastgreedy (CNM) method (Clauset et al., 2004) would more likely link several loosely connected clusters as one community, while MCL (Enright et al., 2002) is prone to output too many clusters. For phylogenomic synteny data, we need some tailor-made criteria for clustering. For example, the clustering method should be tolerant for monocot species to have relatively fewer connections to the cluster formed by most dicots species in the network, rather than separating them into different clusters, due to the large evolutionary distance and extensive genomic rearrangements.

This thesis uses clique percolation method (CPM) based on k-cliques, and infomap in separate chapters. The k-cliques method (Derenyi et al., 2005) provides a good balance of presenting the whole picture while discarding noises. The k-cliques community brings together the gene clusters of SEP1, SQUA, SEP3, FLC, and TM8, which allows us to think about the relations and evolution of them as a whole (Figure 4, Chapter 3). The CPM algorithm (Derenyi et al., 2005) detects all complete sub graphs and then builds a clique-clique overlap matrix, and returns all connected components by a given k. On this account, the matrix could be huge and take substantial amounts of memory for very large networks. Algorithms such as fastgreedy (Clauset-Newman-Moore, Clauset et al., 2004), BigClam (Yang and Leskovec, 2013), Louvain (Blondel et al., 2008), and infomap (Rosvall and Bergstrom, 2007) would be more capable to efficiently identify communities inside large networks. Infomap is recognized as one of the most reliable algorithms according to several researches. To correctly evaluate the performance of the different clustering methods is still an issue, especially in the case of overlapping communities. The only way we can appreciate the quality of an algorithm is to test it on a graph with a built-in community structure, or where we know the ground-truth communities. Therefore, much could be further explored for synteny network clustering algorithms. New methods that are aware of phylogenetic distance between the nodes (weighted edges), and have high efficiency and accuracy would be most desirable.

Weakness and future perspective

Besides the main chapters within this thesis, I have also used this synteny network approach to investigate gene family or genome evolution with other collaborators. For example the analysis of fish genomes and birds genomes, and diverse gene families, such as ARFs, AP2s, NACs, HIPPs, LTPs, and TPS. Overall our synteny network approach provides exciting insights, which are consistent with wet-lab evidences. Nevertheless, there are several drawbacks with this approach, which I discuss below.

Firstly, for synteny detection we performed pairwise genome comparisons based on whole-genome annotations. Therefore genomes without annotations, represent in only raw scaffolds cannot be used directly. Also, only coding sequences are included as nodes, and as such intergenic non-coding regions that may contain important regulative conserved motifs across species cannot be inferred from the network.

Secondly, young tandem replicates could be underestimated from the network. This problem is derived from synteny detection softwares. We use MCScanX (the successor of MCScan) (Tang et al., 2008b; Wang et al., 2012), which is by far the most cited, and is stricter in its default settings (Liu et al., 2018). Tandem arrays could contains two to twenty genes continuously (e.g. P450s, TPS genes, LRR genes, etc.). In order to avoid blocks that consist of purely tandem duplicates, MCScanX by default collapses multiple tandem matches into one representative pair. Similar treatments for tandem repeats were also found for example in i-ADHoRe and SynFind. As a consequence, conserved tandem duplicates within the syntenic blocks across species will be greatly pruned (Figure 1). For example, the three important cold-resistance CBF genes in *Arabidopsis thaliana* (CBF1: AT4G25490, CBF2: AT4G25470, CBF3: AT4G25480), are in fact syntenic to three *A. lyrata* CBF genes (AL7G27600, AL7G27610, AL7G27630), but this cannot be directly inferred from standard synteny detection outputs (as exemplified in Figure 1B). A comparison of syntenic conserved tandem arrays across species could provide important insights of when and how such duplications occurred. Thus further complements to current synteny detection tools should be made to remedy such situation in the future.

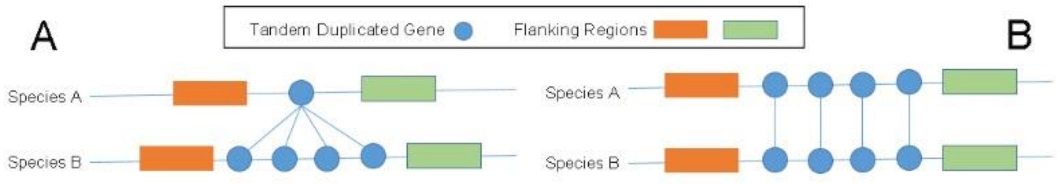


Figure 1. Two scenarios of synteny blocks containing tandem duplicated syntenic pairs, which can be oversimplified by a synteny detection software. (A) The gene was tandem duplicated in one species. (B) The gene was tandem duplicated in both species.

Thirdly, the synteny network approach provides visualization for particular genes of interest, without showing the flanking genes from the syntenic blocks. Thus it is not straightforward to compare flanking neighbouring genes surrounding the query locus. In fact for this need, we could always restore all involved synteny blocks using synteny block IDs and then align the blocks separately. However, a better way to organize, summarize, and visualize flanking neighbouring genes from all these blocks is the next challenge.

In this thesis, I showed that by using our novel synteny network approach, phylogenomic synteny conservation or diversification could be qualified and quantified in an easy way. Synteny properties could be used as another character besides chromatin data, methylation data, expression data, protein interaction data, etc. for machine-learning based computational biology, to confer more discoveries underlying trait evolution. Ever-evolving gene-sequencing technology, fast-increasing genomics data, fast-developing big-data-based infrastructure and analytics, as well as the application of targeted genome editing (CRISPR-Cas9) and synthetic biology, have put a spring in the step of understanding how genomes work through evolutionary comparative genomic analysis, and clarifying mechanisms that generate the great diversity of present-day life forms.

References

Adamczyk, B.J., and Fernandez, D.E. (2009). MIKC* MADS domain heterodimers are required for pollen maturation and tube growth in Arabidopsis. *Plant Physiol* 149, 1713-1723.

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8, 135-141.

Agarwal, T., Upadhyaya, G., Halder, T., Mukherjee, A., Majumder, A.L., and Ray, S. (2017). Different dehydrins perform separate functions in *Physcomitrella patens*. *Planta* 245, 101-118.

Airoldi, C.A., and Davies, B. (2012). Gene duplication and the evolution of plant MADS-box transcription factors. *J Genet Genomics* 39, 157-165.

Albert, V.A., et al. (2013). The Amborella genome and the evolution of flowering plants. *Science* 342, 1241089.

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., and Rokhsar, D.S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524, 220-224.

Altenhoff, A.M., et al. (2016). Standardized benchmarking in the quest for orthologs. *Nat Methods* 13, 425.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.

Alvarez-Buylla, E.R., Liljegren, S.J., Pelaz, S., Gold, S.E., Burgeff, C., Ditta, G.S., Vergara-Silva, F., and Yanofsky, M.F. (2000a). MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *Plant J* 24, 457-466.

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. (2000b). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci U S A* 97, 5328-5333.

Alvarez-Hamelin, J.I., Dall'Asta, L., Barrat, A., and Vespignani, A. (2005). K-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. arXiv preprint cs/0511007.

Alvarez-Hamelin, J.I., Dall'Asta, L., Barrat, A., and Vespignani, A. (2006). Large scale networks fingerprinting and visualization using the k-core decomposition. *Adv Neur Inform Proc Sys* 18, 41.

Amara, I., Zaidi, I., Masmoudi, K., Ludevid, M.D., Pagès, M., Goday, A. and Brini, F. (2014). Insights into late embryogenesis abundant (LEA) proteins in plants: from structure to the functions. *Am J Plant Sci* 5, 3440-3445.

Amores, A., et al. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science* 282, 1711-1714.

Angiuoli, S.V., and Salzberg, S.L. (2010). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334-342.

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. *Science* 333, 601-607.

Arsovski, A.A., Pradinuk, J., Guo, X.Q., Wang, S., and Adams, K.L. (2015). Evolution of cis-regulatory elements and regulatory networks in duplicated genes of Arabidopsis. *Plant Physiol* 169, 2982-2991.

Babu, R.C., Zhang, J., Blum, A., Ho, T.H.D., Wu, R., Nguyen, H.T. (2004). HVA1, a LEA gene from barley confers dehydration tolerance in transgenic rice (*Oryza sativa* L.) via cell membrane protection. *Plant Sci* 166, 855-862.

Baek, J.-H., Kim, J., Kim, C.-K., Sohn, S.-H., Choi, D., Ratnaparkhe, M.B., Kim, D.-W., and Lee, T.-H. (2016). MultiSyn: A webtool for multiple synteny detection and visualization of user's sequence of interest compared to public plant species. *Evol Bioinform* 12, EBO. S40009.

Baier, U., Beller, T., and Ohlebusch, E. (2015). Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform. *Bioinformatics*, btv603.

Baldauf, S.L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends Genet* 19, 345-351.

Banerjee, A., and Roychoudhury, A. (2016). Group II late embryogenesis abundant (LEA) proteins: structural and functional aspects in plant abiotic stress. *Plant Growth Regu* 79, 1-17.

Barabási, A.-L., and Pósfai, M. (2016). Network science. (Cambridge university press).

Barker, M.S., Arrigo, N., Baniaga, A.E., Li, Z., and Levin, D.A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol* 210, 391-398.

Bartelt, A., et al. (2018). Brown adipose tissue thermogenic adaptation requires Nrf1-mediated proteasomal activity. *Nat Med* 24, 292.

Bartlett, M., Thompson, B., Brabazon, H., Del Gizzi, R., Zhang, T., and Whipple, C. (2016). Evolutionary dynamics of floral homeotic transcription factor protein-protein interactions. *Mol Biol Evol* 33 1486-1501

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8, 361-362.

Bastian, M., Heymann, S., Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Proc 3rd Int AAAI Conf Web Soc Med* 8, 361-362.

Bastow, R., Mylne, J.S., Lister, C., Lippman, Z., Martienssen, R.A., and Dean, C. (2004). Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* 427, 164-167.

Batagelj, V., and Zaversnik, M. (2003). An O (m) algorithm for cores decomposition of networks. arXiv preprint cs/0310049.

Bateman, A., et al. (2002). The Pfam Protein Families Database. *Nucleic Acids Res* 30, 276-280.

Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F., and Covarrubias, A.A. (2008). The enigmatic LEA proteins and other hydrophilins. *Plant Physiol* 148, 6-24.

Becker, A., and Theissen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogenet Evol 29, 464-489.

Becker, B., and Marin, B. (2009). Streptophyte algae and the origin of embryophytes*. Ann Bot* 103, 999-1004.

Bemer, M., Heijmans, K., Airoldi, C., Davies, B., and Angenent, G.C. (2010). An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis. *Plant Physiol* 154, 287-300.

Benshahar, A., Chalifa-Caspi, V., Hermelin, D., and Ziv-Ukelson, M. (2017). A Biclique Approach to Reference-Anchored Gene Blocks and Its Applications to Genomic Islands. *J Comput Biol* 25, 214-235.

Bhoumik, A., et al. (2008). Suppressor role of activating transcription factor 2 (ATF2) in skin cancer. *Proc Natl Acad Sci U S A* 105, 1674-1679.

Bininda-Emonds, O.R., et al. (2007). The delayed rise of present-day mammals. *Nature* 446, 507-512.

Blanc, G., and Wolfe, K.H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679-1691.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008, P10008.

Bond, J., et al. (2005). A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nat Genet* 37, 353.

Boutanaev, A.M., Moses, T., Zi, J., Nelson, D.R., Mugford, S.T., Peters, R.J., and Osbourn, A. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proc Natl Acad Sci U S A* 112, E81-E88.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433-438.

Brooke, N.M., Garcia-Fernandez, J., and Holland, P.W. (1998). The *ParaHox* gene cluster is an evolutionary sister of the *Hox* gene cluster. *Nature* 392, 920-922.

Browne, J., Tunnacliffe, A., and Burnell, A. (2002). Anhydrobiosis - Plant desiccation gene found in a nematode. *Nature* 416, 38-38.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12, 59.

Byrne, R.W., and Corp, N. (2004). Neocortex size predicts deception rate in primates. *Philos Trans R Soc Lond B Biol Sci* 271, 1693.

Cai, B., Yang, X., Tuskan, G.A., and Cheng, Z.-M. (2011). MicroSyn: a user friendly tool for detection of microsynteny in a gene family. *BMC Bioinfomatics* 12, 79.

Calabrese, P.P., Chakravarty, S., and Vision, T.J. (2003). Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* 19, i74-i80.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.

Cannon, S.B., Kozik, A., Chan, B., Michelmore, R., and Young, N.D. (2003). DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* 4, R68.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol* 4, 10.

Causier, B., Castillo, R., Xue, Y., Schwarz-Sommer, Z., and Davies, B. (2010). Tracing the evolution of the floral homeotic B- and C-function genes through genome synteny. *Mol Biol Evol* 27, 2651-2664.

Chakrabortee, S., Boschetti, C., Walton, L.J., Sarkar, S., Rubinsztein, D.C., and Tunnacliffe, A. (2007). Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proc Natl Acad Sci U S A* 104, 18073-18078.

Chakrabortee, S., Tripathi, R., Watson, M., Schierle, G.S.K., Kurniawan, D.P., Kaminski, C.F., Wise, M.J., and Tunnacliffe, A. (2012). Intrinsically disordered proteins as molecular shields. *Mol Biosyst* 8, 210-219.

Chen, Y.S., Lo, S.F., Sun, P.K., Lu, C.A., Ho, T.H., and Yu, S.M. (2015). A late embryogenesis abundant protein HVA1 regulated by an inducible promoter enhances root growth and abiotic stress tolerance in rice without yield penalty. *Plant Biotechnol J* 13, 105-116.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X. (2018a). Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* 4, 258-268.

Cheng, S., et al. (2018b). 10KP: A phylodiverse genome sequencing plan. *GigaScience* 7, giy013.

Cheng, S., et al. (2013). The Tarenaya hassleriana genome provides insight into reproductive trait and genome evolution of crucifers. *Plant Cell* 25, 2813-2830.

Chowdhury, A.M.A., Katoh, H., Hatanaka, A., Iwanari, H., Nakamura, N., Hamakubo, T., Natsume, T., Waku, T., and Kobayashi, A. (2017). Multiple regulatory mechanisms of the biological function of *NRF3* (*NFE2L3*) control cancer cell proliferation. *Sci Rep* 7, 12494.

Ciccarelli, F.D., and Bork, P. (2005). The WHy domain mediates the response to desiccation in plants and bacteria. *Bioinformatics* 21, 1304-1307.

Ciccarelli, F.D., Doerks, T., Von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283-1287.

Cifelli, R.L., and Davis, B.M. (2003). Marsupial origins. *Science* 302, 1899-1900.

Clauset, A., Newman, M.E., and Moore, C. (2004). Finding community structure in very large networks. *Phys Rev E* 70, 066111.

Close, T.J. (1996). Dehydrins: Emergence of a biochemical role of a family of plant dehydration proteins. *Physiologia Plantarum* 97, 795-803.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature* 353, 31.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* 9, 938-950.

Conant, G.C., Birchler, J.A., and Pires, J.C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol* 19, 91-98.

Consortium, G. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*. 166, 481-491

Consortium, I.H.G.S. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860.

Costa, M.-C.D., et al. (2017). A footprint of desiccation tolerance in the genome of *Xerophyta viscosa* 3, 17038.

Couvreur, T.L., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol Biol Evol* 27, 55-71.

Covarrubias, A.A., Cuevas-Velazquez, C.L., Romero-Perez, P.S., Rendon-Luna, D.F., and Chater, C.C.C. (2017). Structural disorder in plant proteins: where plasticity meets sessility. *Cell Mol Life Sci* 74, 3119-3147.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Complex Syst* 1695, 1-9.

Csorba, T., Questa, J.I., Sun, Q., and Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at *FLC* during vernalization. *Proc Natl Acad Sci U S A* 111, 16160-16165.

Cuevas-Velazquez, C.L., Reyes, J.L., and Covarrubias, A.A. (2017). Group 4 late embryogenesis abundant proteins as a model to study intrinsically disordered proteins in plants. *Plant Signal Behav* 12, e1343777.

Cui, L., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16, 738-749.

Cuming, A.C., Cho, S.H., Kamisugi, Y., Graham, H., and Quatrano, R.S. (2007). Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, *Physcomitrella patens*. *New Phytol* 176, 275-287.

Daminato, M., Masiero, S., Resentini, F., Lovisetto, A., and Casadoro, G. (2014). Characterization of *TM8*, a MADS-box gene expressed in tomato flowers. *BMC Plant Biol* 14, 319.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23, 324-328.

Delahaie, J., Hundertmark, M., Bove, J., Leprince, O., Rogniaux, H., and Buitink, J. (2013). LEA polypeptide profiling of recalcitrant and orthodox legume seeds reveals ABI3-regulated LEA protein abundance linked to desiccation tolerance. *J Exp Bot* 64, 4559-4573.

Derenyi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Phys Rev Lett* 94, 160202.

Derrien, T., André, C., Galibert, F., and Hitte, C. (2006). AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics* 23, 498-499.

Despalins, A., Marsit, S., and Oberto, J. (2011). Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. *Bioinformatics* 27, 2905-2906.

Dewey, C.N. (2011). Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform* 12, 401-412.

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., and Yanofsky, M.F. (2004). The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. *Current Biol* 14, 1935-1940.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J Vegetation Science* 14, 927-930.

Dodd, M.S., Papineau, D., Grenne, T., Slack, J.F., Rittner, M., Pirajno, F., O'Neil, J., and Little, C.T. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* 543, 60.

Dodsworth, S. (2016). Petal, Sepal, or Tepal? B-Genes and Monocot Flowers. *Trends Plant Sci*. 22, 8-10

Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems. *Phys. Rev* 5, 011027.

Dong, X., Fredman, D., and Lenhard, B. (2009). Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol* 10, R86.

Donthu, R., Lewin, H.A., and Larkin, D.M. (2009). SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence. *BMC Res Notes* 2, 148.

Drillon, G., Carbone, A., and Fischer, G. (2014). SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* 9, e92621.

Duboule, D. (2007). The rise and fall of Hox gene clusters. *Development* 134, 2549-2560.

Dure, L., Crouch, M., Harada, J., Ho, T.H.D., Mundy, J., Quatrano, R., Thomas, T., and Sung, Z.R. (1989). Common amino-acid sequence domains among the lea proteins of higher-plants. *Plant Mol Biol* 12, 475-486.

Edger, P.P., et al. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proc Natl Acad Sci U S A* 112, 8362-8366.

Eisen, J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 8, 163-167.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575-1584.

Escobar, C., et al. (1999). Isolation of the *LEMMI9* gene and promoter analysis during a compatible plant-nematode interaction. *Mol Plant-Microbe I* 12, 440-449.

Espelund, M., Saeboelarssen, S., Hughes, D.W., Galau, G.A., Larsen, F., and Jakobsen, K.S. (1992). Late embryogenesis-abundant genes encoding proteins with different numbers of hydrophilic repeats are regulated differentially by abscisic-acid and osmotic-stress. *Plant J* 2, 241-252.

Farrant, J.M., and Moore, J.P. (2011). Programming desiccation-tolerance: from plants to seeds to resurrection plants. *Curr Opin Plant Biol* 14, 340-345.

Feng, S., Jacobsen, S.E., and Reik, W. (2010). Epigenetic reprogramming in plant and animal development. *Science* 330, 622-627.

Ferrier, D.E., and Holland, P.W. (2001). Ancient origin of the Hox gene cluster. *Nat Rev Genet* 2, 33-38.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39, W29-37.

Finnegan, E.J., Sheldon, C.C., Jardinaud, F., Peacock, W.J., and Dennis, E.S. (2004). A cluster of Arabidopsis genes with a coordinate response to an environmental stimulus. *Current Biol* 14, 911-916.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytol* 183, 557-564.

Fortunato, S. (2010). Community detection in graphs. *Phys Rep* 486, 75-174.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., and Lisch, D. (2008). Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res* 18, 1924-1937.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr Opin Plant Biol* 15, 131-139.

Fu, D., Huang, B., Xiao, Y., Muthukrishnan, S., and Liang, G.H. (2007). Overexpression of barley hva1 gene in creeping bentgrass for improving drought tolerance. *Plant Cell Rep* 26, 467-477.

Gabaldon, T., and Koonin, E.V. (2013). Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14, 360-366.

Gaff, D.F., and Oliver, M. (2013). The evolution of desiccation tolerance in angiosperm plants: a rare yet common phenomenon. *Func Plant Biol* 40, 315-328.

Galau, G.A., Hughes, D.W., and Dure, L. (1986). Abscisic-acid induction of cloned cotton late embryogenesis-abundant (LEA) messenger-rnas. *Plant Mol Biol* 7, 155-170.

Garay-Arroyo, A., Colmenero-Flores, J.M., Garciarrubio, A., and Covarrubias, A.A. (2000). Highly hydrophilic proteins in prokaryotes and eukaryotes are common during conditions of water deficit. *J Biol Chem* 275, 5668-5674.

Gehrmann, T., and Reinders, M.J. (2015). Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. *Bioinformatics* 31, 3437-3444.

Ghiurcuta, C.G., and Moret, B.M. (2014). Evaluating synteny for improved comparative studies. *Bioinformatics* 30, i9-i18.

Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 99, 7821-7826.

Gladyshev, E.A., and Arkhipova, I.R. (2007). Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A* 104, 9352-9357.

Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc Natl Acad Sci U S A* 104, 8685-8690.

Goodstein, D.M., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, D1178-D1186.

Goto, K., and Meyerowitz, E.M. (1994). Function and regulation of the Arabidopsis floral homeotic gene *PISTILLATA*. *Genes Dev* 8, 1548-1560.

Goyal, K., Walton, L.J., and Tunnacliffe, A. (2005). LEA proteins prevent protein aggregation due to water stress. *Biochem J* 388, 151-157.

Gozdecka, M., and Breitwieser, W. (2012). The roles of *ATF2* (activating transcription factor 2) in tumorigenesis (Portland Press Limited). *Biochem Soc Trans* 40, 230-234.

Grabherr, M.G., Russell, P., Meyer, M., Mauceli, E., Alfoldi, J., Di Palma, F., and Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* 26, 1145-1151.

Gramzow, L., and Theissen, G. (2013). Phylogenomics of MADS-Box genes in plants - two opposing life styles in one gene family. *Biology* (Basel) 2, 1150-1164.

Gramzow, L., Weilandt, L., and Theissen, G. (2014). MADS goes genomic in conifers: towards determining the ancestral set of MADS-box genes in seed plants. *Ann Bot* 114, 1407-1429.

Gramzow, L., Barker, E., Schulz, C., Ambrose, B., Ashton, N., Theissen, G., and Litt, A. (2012). Selaginella genome analysis - entering the "Homoplasy Heaven" of the MADS world. *Front Plant Sci* 3, 214.

Gul, A., Hassan, M.J., Hussain, S., Raza, S.I., Chishti, M.S., and Ahmad, W. (2006). A novel deletion mutation in *CENPJ* gene in a Pakistani family with autosomal recessive primary microcephaly. *J Hum Genet* 51, 760-764.

Gupta, S., Campbell, D., Derijard, B., and Davis, R.J. (1995). Transcription factor *ATF2* regulation by the JNK signal transduction pathway. *Science* 5196, 389-393.

Gusev, O., et al. (2014). Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. *Nat Commun* 5.4784

Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643-3646.

Hachiya, T., Osana, Y., Popendorf, K., and Sakakibara, Y. (2009). Accurate identification of orthologous segments among multiple genomes. *Bioinformatics* 25, 853-860.

Hammoudi, V., Vlachakis, G., Schranz, M.E., and van den Burg, H.A. (2016). Whole-genome duplications followed by tandem duplications drive diversification of the protein modifier SUMO in Angiosperms. *New Phytol* 211, 172-185.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* 148, 993-1003.

Hartmann, U., Hohmann, S., Nettesheim, K., Wisman, E., Saedler, H., and Huijser, P. (2000). Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. *Plant J* 21, 351-360.

He, S., Tan, L.L., Hu, Z.L., Chen, G.P., Wang, G.X., and Hu, T.Z. (2012). Molecular characterization and functional analysis by heterologous expression in *E. coli* under diverse abiotic stresses for *OsLEA5*, the atypical hydrophobic LEA protein from *Oryza sativa* L. *Mol Genet Genom* 287, 39-54.

Henschel, K., Kofuji, R., Hasebe, M., Saedler, H., Munster, T., and Theissen, G. (2002). Two ancient classes of MIKC-type MADS-box genes are present in the moss Physcomitrella patens. *Mol Biol Evol* 19, 801-814.

Hincha, D.K., and Thalhammer, A. (2012). LEA proteins: IDPs with versatile functions in cellular dehydration tolerance. *Biochem Soc Trans* 40, 1000-1003.

Hirsch, C.N., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26, 121-135.

Hoekstra, F.A., Golovina, E.A., and Buitink, J. (2001). Mechanisms of plant desiccation tolerance. *Trends in Plant Sci* 6, 431-438.

Hokamp, K., McLysaght, A., and Wolfe, K.H. (2003). The 2R hypothesis and the human genome sequence. In Genome Evolution (Springer), pp. 95-110.

Hunault, G., and Jaspard, E. (2010). LEAPdb: a database for the late embryogenesis abundant proteins. *BMC Genomics* 11,221.

Hundertmark, M., and Hincha, D.K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 9.

Husemann, P., and Stoye, J. (2009). r2cat: synteny plots and comparative assembly. *Bioinformatics* 26, 570-571.

Illing, N., Denby, K.J., Collett, H., Shen, A., and Farrant, J.M. (2005). The signature of seeds in resurrection plants: A molecular and physiological comparison of desiccation tolerance in seeds and vegetative tissues. *Integ Comp Biol* 45, 771-787.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543, 550-554.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models*. Nat Rev Genet* 11, 97-108.

Jack, T., Brockman, L.L., and Meyerowitz, E.M. (1992). The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens. *Cell* 68, 683-697.

Jack, T., Fox, G.L., and Meyerowitz, E.M. (1994). Arabidopsis homeotic gene *APETALA3* ectopic expression: transcriptional and posttranscriptional regulation determine floral organ identity. *Cell* 76, 703-716.

Jaramillo, M.A., and Kramer, E.M. (2007). Molecular evolution of the petal and stamen identity genes, *APETALA3* and *PISTILLATA*, after petal loss in the Piperales. *Mol Phylogenet Evol* 44, 598-609.

Jarukasemratana, S., and Murata, T. (2013). Recent Large Graph Visualization Tools: A Review. *Inform Media Tech* 8, 944-960.

Jaspard, E., and Hunault, G. (2014). Comparison of amino acids physico-chemical properties and usage of Late embryogenesis abundant proteins, hydrophilins and WHy domain. *Plos One* 9, e109570.

Jia, F.J., Qi, S.D., Li, H., Liu, P., Li, P.C., Wu, C.G., Zheng, C.C., and Huang, J.G. (2014). Overexpression of late embryogenesis abundant 14 enhances Arabidopsis salt stress tolerance. *Biochem Biophys Res Commun* 454, 505-511.

Jiang, S.J., et al. (2017). DrwH, a novel WHy domain-containing hydrophobic LEA5C protein from Deinococcus radiodurans, protects enzymatic activity under oxidative stress. *Sci Rep* 7, 9281.

Jiao, Y., and Paterson, A.H. (2014). Polyploidy-associated genome modifications during land plant evolution. *Philos Trans R Soc Lond B Biol Sci* 369.

Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26, 2792-2802.

Jiao, Y., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97-100.

Jiao, Y., et al. (2012a). A genome triplication associated with early diversification of the core eudicots. *Genome Biol* 13, R3.

Jiao, Y., et al. (2012b). Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 44, 812-815.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45, D1040-D1045.

Joh, T., Honjoh, K., Yoshimoto, M., Funabashi, J., Miyamoto, T., and Hatano, S. (1995). Molecular-cloning and expression of hardening-induced genes in chlorella-vulgaris C-27 - the most abundant clone encodes a late embryogenesis abundant protein. *Plant Cell Physiol* 36, 85-93.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasner, M.E., and Bartel, D.P. (2001). RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* 292, 1319-1325.

Jones, J.D., Vance, R.E., and Dangl, J.L. (2016). Intracellular innate immune surveillance devices in plants and animals. *Science* 354, 6316.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059-3066.

Kenrick, P., and Crane, P.R. (1997). The origin and early evolution of plants on land. *Nature* 389, 33-39.

Khaouid, W., Barsky, M., Srinivasan, V., and Thomo, A. (2015). K-core decomposition of large networks on a single PC. *Proc VLDB Endowment* 9, 13-23.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res* 21, 487-493.

Kikawada, T., Nakahara, Y., Kanamori, Y., Iwata, K.I., Watanabe, M., Mcgee, B., Tunnacliffe, A., and Okuda, T. (2006). Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid. *Biochem Biophys Res Commun* 348, 56-61.

Kim, D.-H., and Sung, S. (2013). Coordination of the vernalization response through a *VIN3* and *FLC* gene family regulatory network in Arabidopsis. *Plant Cell* 25, 454-469.

Kim, S., Soltis, P.S., and Soltis, D.E. (2013). *AGL6*-like MADS-box genes are sister to *AGL2*-like MADS-box genes. *J Plant Biol* 56, 315-325.

Kobayashi, A., Ito, E., Toki, T., Kogame, K., Takahashi, S., Igarashi, K., Hayashi, N., and Yamamoto, M. (1999). Molecular cloning and functional characterization of a new Cap'n'collar family transcription factor Nrf3. *J Biol Chem* 274, 6443-6452.

Kohler, C., Hennig, L., Spillane, C., Pien, S., Gruissem, W., and Grossniklaus, U. (2003). The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene *PHERES1*. *Genes Dev* 17, 1540-1553.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *P Roy Soc B-Biol Sci* 279, 5048-5057.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309-338.

Kovacs, D., Kalmar, E., Torok, Z., and Tompa, P. (2008). Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiol* 147, 381-390.

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., and Koonin, E.V. (2011). Computational methods for Gene Orthology inference. *Brief Bioinform* 12, 379-391.

Krumlauf, R. (1994). Hox genes in vertebrate development. *Cell* 78, 191-201.

Kudo, H., and Dunbar, R. (2001). Neocortex size and social network size in primates. *Anim Behav* 62, 711-722.

Lancichinetti, A., and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Phys Rev E* 80, 056117.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11, 204-220.

Lee, C., Reid, F., McDaid, A., and Hurley, N. (2010). Detecting highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827.

Lee, H.L., and Irish, V.F. (2011). Gene duplication and loss in a MADS box gene transcription factor circuit. *Mol Biol Evol* 28, 3367-3380.

Lee, J., Hong, W.-y., Cho, M., Sim, M., Lee, D., Ko, Y., and Kim, J. (2016). Synteny Portal: a web-based application portal for synteny block analysis. *Nucleic Acids Res* 44, W35-W40.

Lee, S., et al. (2003). Systematic reverse genetic screening of T-DNA tagged genes in rice for functional genomic analyses: MADS-box genes as a test case. *Plant Cell Physiol* 44, 1403-1411.

Lee, T.H., Tang, H.B., Wang, X.Y., and Paterson, A.H. (2013). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res* 41, D1152-D1158.

Lemoine, F., Labedan, B., and Lespinet, O. (2008). SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes. *BMC Bioinformatics* 9, 536.

Lemons, D., and McGinnis, W. (2006). Genomic evolution of Hox gene clusters. *Science* 313, 1918-1922.

Leprince, O., and Buitink, J. (2010). Desiccation tolerance: From genomics to the field. *Plant Sci* 179, 554-564.

Leskovec, J., and Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. *ACM T Intel Syst Tec* 8, 1.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees*.* *Nucleic Acids Res* 44, W242-W245.

Levrero, M., De Laurenzi, V., Costanzo, A., Gong, J., Wang, J., and Melino, G. (2000). The p53/p63/p73 family of transcription factors: overlapping and distinct functions. *J Cell Sci* 113, 1661-1670.

Lewin, H.A., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* 115, 4325-4333.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. *Nature* 276, 565-570.

Lewis, P.O. (2001). Phylogenetic systematics turns over a new leaf. *Trends Ecol Evol* 16, 30-37.

Li, Q., and Zhang, Y. (2013). The origin and functional transition of P34. *Heredity* 110, 259-266.

Li, Y.-h., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* 32, 1045-1052.

Li, Y.F., Zheng, Y., Addo.Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M.J., Zhang, W., and Sunkar, R. (2010). Transcriptome-wide identification of microRNA targets in rice. *Plant J* 62, 742-759.

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., and De Smet, R. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28, 326-344.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., and Barker, M.S. (2015). Early genome duplications in conifers and other seed plants. *Sci Adv* 1, e1501084.

Ling, X., He, X., and Xin, D. (2009). Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics* 25, 571-577.

Liu, D., Hunt, M., and Tsai, I.J. (2018). Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics* 19, 26.

Liu, Y., Song, Q.P., Li, D.X., Yang, X.H., and Li, D.Q. (2017). Multifunctional roles of plant dehydrins in response to environmental stresses. *Front Plant Sci* 8, 1018.

Louis, A., Muffato, M., and Crollius, H.R. (2012). Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res* 41, D700-705.

Lucas, J.M., Muffato, M., and Crollius, H.R. (2014). PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 15, 268.

Lv, J., Havlak, P., and Putnam, N.H. (2011). Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. *BMC Bioinformatics* 12 Suppl 9, S11.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151-1155.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* 53, 661-673.

Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008). The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop Plant Biol* 1, 181-190.

Magain, N., et al. (2017). Conserved genomic collinearity as a source of broadly applicable, fast evolving, markers to resolve species complexes: A case study using the lichen-forming genus Peltigera section Polydactylon. *Mol Phylogenetics Evol* 117, 10-29.

Magallon, S., Gomez-Acevedo, S., Sanchez-Reyes, L.L., and Hernandez-Hernandez, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* 207, 437-453.

Malik, A.A., Veltri, M., Boddington, K.F., Singh, K.K., and Graether, S.P. (2017). Genome Analysis of Conserved Dehydrin Motifs in Vascular Plants. *Front. Plant Sci* 8.

Manfre, A.J., Lanni, L.M., and Marcotte, W.R. (2006). The Arabidopsis group 1 LATE EMBRYOGENESIS ABUNDANT protein ATEM6 is required for normal seed development. *Plant Physiol* 140, 140-149.

Marschall, T., et al. (2016). Computational pan-genomics: Status, promises and challenges. bioRxiv, 043430.

Martinez-Castilla, L.P., and Alvarez-Buylla, E.R. (2003). Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. *Proc Natl Acad Sci U S A* 100, 13407-13412.

Masiero, S., Colombo, L., Grini, P.E., Schnittger, A., and Kater, M.M. (2011). The emerging importance of type I MADS box transcription factors for plant reproduction. *Plant Cell* 23, 865-872.

Melzer, R., Harter, A., Rumpler, F., Kim, S., Soltis, P.S., Soltis, D.E., and Theissen, G. (2014). DEF- and GLO-like proteins may have lost most of their interaction partners during angiosperm evolution. *Ann Bot* 114, 1431-1443.

Michaels, S.D., and Amasino, R.M. (1999). *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* 11, 949-956.

Michaels, S.D., Ditta, G., Gustafson-Brown, C., Pelaz, S., Yanofsky, M., and Amasino, R.M. (2003). *AGL24* acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. *Plant J* 33, 867-874.

Ming, R., et al. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol* 14, R41.

Minkin, I., Patel, A., Kolmogorov, M., Vyahhi, N., and Pham, S. (2013). Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In International Workshop on Algorithms in Bioinformatics (Springer), pp. 215-229.

Mundree, S.G., Baker, B., Mowla, S., Peters, S., Marais, S., Vander Willigen, C., Govender, K., Maredza, A., Muyanga, S., Farrant, J.M., Thomson, J.A. (2002). Physiological and molecular insights into drought tolerance. *Afr J Biotechnol* 1, 28-38.

Murat, F., Peer, Y.V.d., and Salse, J. (2012). Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol Evol* 4, 917-928.

Murray-Zmijewski, F., Lane, D., and Bourdon, J. (2006). p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death Diff* 13, 962-972.

Nam, J., dePamphilis, C.W., Ma, H., and Nei, M. (2003). Antiquity and evolution of the MADS-box gene family controlling flower development in plants. *Mol Biol Evol* 20, 1435-1447.

Nam, J., Kaufmann, K., Theissen, G., and Nei, M. (2005). A simple method for predicting the functional differentiation of duplicate genes and its application to MIKC-type MADS-box genes. *Nucleic Acids Res* 33, e12.

Nam, J., Kim, J., Lee, S., An, G.H., Ma, H., and Nei, M.S. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proc Natl Acad Sci U S A* 101, 1910-1915.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Bio Evol* 32, 268-274.

Nützmann, H.W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters–from genetics to genomics. *New Phytol* 211, 771-789.

Oberto, J. (2013). SyntTax: a web server linking synteny to prokaryotic taxonomy. *BMC Bioinformatics* 14, 4.

Ohno, S., Wolf, U., and Atkin, N.B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas* 59, 169-187.

Oliver, M.J., Tuba, Z., and Mishler, B.D. (2000). The evolution of vegetative desiccation tolerance in land plants. *Plant Ecol* 151, 85-100.

Olsen, J.L., et al. (2016). The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* 530, 331-335.

Olvera-Carrillo, Y., Campos, F., Reyes, J.L., Garciarrubio, A., and Covarrubias, A.A. (2010). Functional analysis of the group 4 late embryogenesis abundant proteins reveals their relevance in the adaptive response during water deficit in Arabidopsis. *Plant Physiol* 154, 373-390.

Palla, G., Barabasi, A.L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature* 446, 664-667.

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814-818.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. (2016). Evolution of gene duplication in plants. *Plant Physiol* 171, 2294-2316.

Panopoulou, G., and Poustka, A.J. (2005). Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. *Trends Genet* 21, 559-567.

Parenicova, L., et al. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. *Plant Cell* 15, 1538-1551.

Passarge, E., Horsthemke, B., and Farber, R.A. (1999). Incorrect use of the term synteny. *Nat Genet* 23, 387-387.

Patchell, M.J., Bolton, M.C., Mankowski, P., and Hall, J.C. (2011). Comparative floral development in Cleomaceae reveals two distinct pathways leading to monosymmetry. *Int J Plant Sci* 172, 352-365.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96, 4285-4288.

Pham, S.K., and Pevzner, P.A. (2010). DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics* 26, 2509-2516.

Pnueli, L., Abu-Abeid, M., Zamir, D., Nacken, W., Schwarz-Sommer, Z., and Lifschitz, E. (1991). The MADS box gene family in tomato: temporal expression during floral development, conserved secondary structures and homology with homeotic genes from Antirrhinum and Arabidopsis. *Plant J* 1, 255-266.

Pons, P., and Latapy, M. (2005). Computing communities in large networks using random walks. In International symposium on computer and information sciences (Springer), pp. 284-293.

Popendorf, K., Tsuyoshi, H., Osana, Y., and Sakakibara, Y. (2010). Murasaki: a fast, parallelizable algorithm to find anchors from multiple genomes. *PLoS One* 5, e12651.

Porter, M.A., Onnela, J.-P., and Mucha, P.J. (2009). Communities in networks. *Notices of the AMS* 56, 1082-1097.

Portereiko, M.F., Lloyd, A., Steffen, J.G., Punwani, J.A., Otsuga, D., and Drews, G.N. (2006). *AGL80* is required for central cell and endosperm development in Arabidopsis. *Plant Cell* 18, 1862-1872.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* 40, e11.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res* 43, D974-981.

Puig, J., Meynard, D., Khong, G.N., Pauluzzi, G., Guiderdoni, E., and Gantet, P. (2013). Analysis of the expression of the *AGL17*-like clade of MADS-box transcription factors in rice. *Gene Expr Pattern* 13, 160-170.

Reichardt, J., and Bornholdt, S. (2006). Statistical mechanics of community detection. *Phys Rev E* 74, 016110.

Rensing, S.A., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64-69.

Renwick, J.H. (1971). The mapping of human chromosomes. *Annu Rev Genet* 5, 81-120.

Reyes, J.L., Campos, F., Wei, H., Arora, R., Yang, Y.I., Karlson, D.T., and Covarrubias, A.A. (2008). Functional dissection of Hydrophilins during in vitro freeze protection. *Plant Cell Environ* 31, 1781-1790.

Riechmann, J.L., Krizek, B.A., and Meyerowitz, E.M. (1996). Dimerization specificity of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. *Proc Natl Acad Sci U S A* 93, 4793-4798.

Rödelsperger, C., and Dieterich, C. (2010). CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PloS one* 5, e8861.

Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* 105, 1118-1123.

Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., and Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nat Commun* 5, 4630.

Ruelens, P., de Maagd, R.A., Proost, S., Theissen, G., Geuten, K., and Kaufmann, K. (2013). *FLOWERING LOCUS C* in monocots and the tandem origin of angiosperm-specific MADS-box genes. *Nat Commun* 4, 2280.

Ruibal, C., Salamo, I.P., Carballo, V., Castro, A., Bentancor, M., Borsani, O., Szabados, L., and Vidal, S. (2012). Differential contribution of individual dehydrin genes from *Physcomitrella patens* to salt and osmotic stress tolerance. *Plant Sci* 190, 89-102.

Sampedro, J., Lee, Y., Carey, R.E., dePamphilis, C., and Cosgrove, D.J. (2005). Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant J* 44, 409-419.

Schena, M., and Davis, R.W. (1992). HD-Zip proteins: members of an Arabidopsis homeodomain protein superfamily. *Proc Natl Acad Sci U S A* 89, 3894-3898.

Schiemann, S.M., Martin-Duran, J.M., Borve, A., Vellutini, B.C., Passamaneck, Y.J., and Hejnol, A. (2017). Clustered brachiopod Hox genes are not expressed collinearly and are associated with lophotrochozoan novelties. *Proc Natl Acad Sci U S A* 114, E1913-E1922.

Schranz, M.E., and Mitchell-Olds, T. (2006). Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18, 1152-1165.

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Curr Opin Plant Biol* 15, 147-153.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504.

Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D., and Smit, S. (2016). PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* 32, i487-i493.

Sheldon, C.C., Rouse, D.T., Finnegan, E.J., Peacock, W.J., and Dennis, E.S. (2000). The molecular basis of vernalization: the central role of *FLOWERING LOCUS C* (FLC). *Proc Natl Acad Sci U S A* 97, 3753-3758.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345.

Shih, M.D., Hoekstra, F.A., and Hsing, Y.I.C. (2008). Late embryogenesis abundant proteins. *Adv Botanical Res*, Vol 48 48, 211-255.

Shinde, S., Nurul Islam, M., and Ng, C.K. (2012). Dehydration stress-induced oscillations in LEA protein transcripts involves abscisic acid in the moss, *Physcomitrella patens*. *New Phytol* 195, 321-328.

Simillion, C., Janssens, K., Sterck, L., and Van de Peer, Y. (2007). i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24, 127-128.

Singh, S., Cornilescu, C.C., Tyler, R.C., Cornilescu, G., Tonelli, M., Lee, M.S., and Markley, J.L. (2005). Solution structure of a late embryogenesis abundant protein (LEA14) from *Arabidopsis thaliana*, a cellular stress-related protein. *Protein Sci* 14, 2601-2609.

Sinha, A.U., and Meller, J. (2007). Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* 8, 82.

Smaczniak, C., Immink, R.G.H., Angenent, G.C., and Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* 139, 3081-3098.

Smith, S.A., and Dunn, C.W. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24, 715-716.

Smith, S.A., Beaulieu, J.M., Stamatakis, A., and Donoghue, M.J. (2011). Understanding angiosperm diversification using small and large phylogenetic trees. *Am J Bot* 98, 404-414.

Soderlund, C., Bomhoff, M., and Nelson, W.M. (2011). SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* 39, e68-e68.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. *Am J Bot* 96, 336-348.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* 35, 119-125.

Soshnikova, N., Dewaele, R., Janvier, P., Krumlauf, R., and Duboule, D. (2013). Duplications of hox gene clusters and the emergence of vertebrates. *Dev Biol* 378, 194-199.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312-1313.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35, 1026.

Steinke, D., Hoegg, S., Brinkmann, H., and Meyer, A. (2006). Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol* 4, 16.

Stevenson, S.R., et al. (2016). Genetic analysis of *Physcomitrella patens* identifies ABSCISIC ACID NON-RESPONSIVE, a regulator of ABA responses unique to basal land plants and required for desiccation tolerance. *Plant Cell* 28, 1310-1327.

Strimbeck, G.R. (2017). Hiding in plain sight: the F segment and other conserved features of seed plant SKn dehydrins. *Planta* 245, 1061-1066.

Sun, W., et al. (2014). Functional and evolutionary analysis of the AP1/SEP/AGL6 superclade of MADS-box genes in the basal eudicot Epimedium sagittatum. *Ann Bot* 113, 653-668.

Sunkar, R., Girke, T., Jain, P.K., and Zhu, J.-K. (2005). Cloning and characterization of microRNAs from rice. *Plant Cell* 17, 1397-1411.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34, W609-612.

Sykiotis, G.P., and Bohmann, D. (2008). Keap1/Nrf2 signaling regulates oxidative stress tolerance and lifespan in Drosophila. *Dev Cell* 14, 76-85.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008a). Synteny and collinearity in plant genomes. *Science* 320, 486-488.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18, 1944-1954.

Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12, 102.

Tang, H., Bomhoff, M.D., Briones, E., Zhang, L., Schnable, J.C., and Lyons, E. (2015). SynFind: compiling syntenic regions across any set of genomes on demand. *Genome Biol Evol* 12, 3286-3298.

Tank, D.C., Eastman, J.M., Pennell, M.W., Soltis, P.S., Soltis, D.E., Hinchliff, C.E., Brown, J.W., Sessa, E.B., and Harmon, L.J. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol* 207, 454-467.

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. *Curr Opin Plant Biol* 4, 75-85.

Theissen, G., Melzer, R., and Rumpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development* 143, 3259-3271.

Tunnacliffe, A., Lapinski, J., and McGee, B. (2005). A putative LEA protein, but no trehalose, is present in anhydrobiotic bdelloid rotifers. *Hydrobiologia* 546, 315-321.

van Veen, H., Akman, M., Jamar, D.C., Vreugdenhil, D., Kooiker, M., van Tienderen, P., Voesenek, L.A., Schranz, M.E., and Sasidharan, R. (2014). Group VII ethylene response factor diversification and regulation in four species from flood-prone environments. *Plant Cell Environ* 37, 2421-2432.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res* 12, 1792-1801.

VanderEycken, W., Engler, J.D., Inze, D., VanMontagu, M., and Gheysen, G. (1996). A molecular study of root-knot nematode-induced feeding sites. *Plant J* 9, 45-54.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res* 24, 1334-1347.

Veltri, D., Wight, M.M., and Crouch, J.A. (2016). SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Res*, W41-45.

Venter, J.C., et al. (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Verelst, W., Saedler, H., and Munster, T. (2007a). MIKC* MADS-protein complexes bind motifs enriched in the proximal region of late pollen-specific Arabidopsis promoters. *Plant Physiol* 143, 447-460.

Verelst, W., Twell, D., de Folter, S., Immink, R., Saedler, H., and Munster, T. (2007b). MADS-complexes regulate transcriptome dynamics during pollen maturation. *Genome Biol* 8, R249.

Wang, H., Jiao, X., Kong, X., Humaira, S., Wu, Y., Chen, X., Fang, R., and Yan, Y. (2016a). A signaling cascade from miR444 to *RDR1* in rice antiviral RNA silencing pathway. *Plant Physiol* 4, 2365-2377.

Wang, X., Shi, X., Li, Z., Zhu, Q., Kong, L., Tang, W., Ge, S., and Luo, J. (2006). Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics* 7, 447.

Wang, Y., Ficklin, S.P., Wang, X., Feltus, F.A., and Paterson, A.H. (2016b). Large-scale gene relocations following an ancient genome triplication associated with the diversification of core Eudicots. *PLoS One* 11, e0155637.

Wang, Y., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40, e49.

Wang, Y., et al. (2013). The sacred lotus genome provides insights into the evolution of flowering plants. *Plant J* 76, 557-567.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189-1191.

Wodniok, S., Brinkmann, H., Glockner, G., Heidel, A.J., Philippe, H., Melkonian, M., and Becker, B. (2011). Origin of land plants: Do conjugating green algae hold the key? *BMC Evol Biol* 11.

Wu, L., Zhang, Q., Zhou, H., Ni, F., Wu, X., and Qi, Y. (2009). Rice microRNA effector complexes and targets. *Plant Cell* 21, 3421-3435.

Xia, X. (2013). What is comparative Genomics? Comparative Genomics (Springer), pp. 1-20.

Xiao, B., Huang, Y., Tang, N., and Xiong, L. (2007). Over-expression of a LEA gene in rice improves drought resistance under the field conditions. *Theor Appl Genet* 115, 35-46.

Xie, J., Kelley, S., and Szymanski, B.K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comput Surv* 45, 43.

Yan, Y., Wang, H., Hamera, S., Chen, X., and Fang, R. (2014). miR444a has multiple functions in the rice nitrate-signaling pathway. *Plant J* 78, 44-55.

Yang, J., and Leskovec, J. (2013). Overlapping community detection at scale: a nonnegative matrix factorization approach. Proceedings of the sixth ACM international conference on Web search and data mining (ACM), pp. 587-596.

Yang, Y., Fanning, L., and Jack, T. (2003). The K domain mediates heterodimerization of the Arabidopsis floral organ identity proteins, *APETALA3* and *PISTILLATA*. *Plant J* 33, 47-59.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet* 13, 303.

Yoshida, K., et al. (2015). The transcription factor ATF7 mediates lipopolysaccharide-induced epigenetic changes in macrophages involved in innate immunological memory. *Nat Immunol* 16, 1034-1043.

Yu, C., Su, S., Xu, Y., Zhao, Y., Yan, A., Huang, L., Ali, I., and Gan, Y. (2014). The effects of fluctuations in the nutrient supply on the expression of five members of the *AGL17* clade of MADS-box genes in rice. *PloS one* 9, e105597.

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

Yu, X., Duan, X., Zhang, R., Fu, X., Ye, L., Kong, H., Xu, G., and Shan, H. (2016b). Prevalent exon-intron structural changes in the *APETALA1/FRUITFULL*, *SEPALLATA*, *AGAMOUS-LIKE6*, and *FLOWERING LOCUS C* MADS-box gene subfamilies provide new insights into their evolution. *Front Plant Sci* 7, 598.

Zahn, L.M., Kong, H., Leebens-Mack, J.H., Kim, S., Soltis, P.S., Landherr, L.L., Soltis, D.E., Depamphilis, C.W., and Ma, H. (2005). The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169, 2209-2223.

Zeng, X., Nesbitt, M.J., Pei, J., Wang, K., Vergara, I.A., and Chen, N. (2008). OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. Proceedings of the 11th international conference on Extending database technology: Advances in Database Technology (ACM), pp. 656-667.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* 18, 292-298.

Zhang, R., et al. (2013). Disruption of the petal identity gene APETALA3-3 is highly correlated with loss of petals within the buttercup family (Ranunculaceae). *Proc Natl Acad Sci U S A* 110, 5074-5079.

Zhao, T., and Schranz, M.E. (2017). Network approaches for plant phylogenomic synteny analysis. *Curr Opin Plant Biol* 36, 129-134.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A., and Schranz, M.E. (2017). Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell* 29, 1278-1292.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28, 125-126.

Zou, C., Lehti-Shiu, M.D., Thomashow, M., and Shiu, S.H. (2009). Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *Plos Genet* 5, e1000581.

Summary

For comparative genomics, relative gene orders or synteny holds key information to assess genomic innovations such as gene duplications, gene loss, or transpositions. While the number of reference genomes is growing exponentially, a major challenge is how to detect, represent, and visualize synteny relations of any genes of interest effectively across a large number of genomes.

In this thesis, I present six chapters centering on a network approach for large-scale phylogenomic synteny analysis, and discuss how such a network approach can enhance our understanding of the evolutionary history of genes and genomes across broad phylogenetic groups and divergence times.

In Chapter 1, I stress that synteny information is becoming more important at this genomics age with rapidly developing DNA sequencing technologies. It provides us another layer of data besides merely sequences, and could potentially be better used to improve phylogeny. I also summarized current available tools and gave an example of popular websites for synteny detection.

In Chapter 2, I propose an outline performing synteny network analysis, which is based on three primary steps: pairwise whole genome comparisons, syntenic block detection and data fusion, and network visualization. Then with comparison to a previous synteny comparison result which use traditional parallel coordinate plots, I show that the network approach could present us a much clear, strong, and systematic graph, with integrated synteny information from 101 broadly distributed species.

In Chapter 3, we analyzed synteny networks of the entire MADS-box transcription factor gene family from fifty-one completed plant genomes. We applied a k-cliques percolation method to cluster the synteny network. We found lineage-specific clusters that derive from transposition events for the regulators of floral development (*APETALA3* and *PI*) and flowering-time (*FLC*) in the Brassicales and for the regulators of root-development (*AGL17*) in Poales. We also visualized big difference of synteny properties between Type I MADS-box genes and Type II MADS-box genes. We identified two large gene clusters that jointly encompass many key phenotypic regulatory Type II MADS-box gene clades (*SEP1*, *SQUA*, *TM8*, *SEP3*, *FLC*, *AGL6* and *TM3*). This allows for a better understanding of how evolution has acted on a key regulatory gene family in the plant kingdom.

In Chapter 4, we performed synteny network analysis of LEA gene families, which includes eight different subfamilies (LEA_1 to LEA_6, SMP, and DHN) and has a relatively chaotic classification. Synteny clusters provide us better pictures of genomic innovations and function diversification. For example recurrent tandem duplications contributed to LEA_2 family expansion, whereas synteny and protein sequence were highly conserved during the evolution of LEA_5.

In Chapter 5, instead of the analysis of a particular gene family, I scale up the analysis to all the genes from all available genomes across kingdoms over significant evolutionary timescales. We used available genomes of 87 mammals and 107 flowering plants. We first compare synteny percentage with popular genome metrics such as BUSCO and N50, which reveal genomic architecture conservation and variation across kingdoms. We characterized and compare the properties of the whole network, using degree distribution and clustering results. Through phylogenomic profiling of size, degree and compositions of all clusters, we identified many phylogenomic genomic innovations (i.e. duplications, gene transpositions, gene loss), at the individual gene level, from tested mammal and angiosperm genomes.

In Chapter 6, I summarize the merits of taking a network-based approach for synteny comparisons, and discuss current clustering methods for synteny data. I also mentioned several weakness, which could be further complemented in the future.

Acknowledgements

Time flies, especially when you enjoyed it. Five years of PhD study at Wageningen University in the Netherlands was quite an experience for me. I would like to express my sincerest gratitude and appreciation to the people who helped me in finishing my PhD in all these years.

First of all, I would like to give the most grateful regards to my promoter and supervisor Prof. M. Eric Schranz. Eric, thanks for your patience and encouragement. Slowly, you have guided me into this fascinating research field. You gave me a better idea of how to think and be innovative in science, instead of mainly following the others. It took a long time for me to find the joyment of analysing the synteny network. I almost gave up, at that difficult time, what I heard most from you was "Don't worry Tao, we will finally get there.". I could always get courage and confidence from you when I felt down, and big praises when I made any progress. I used to miss you a lot during your holiday times, because I could not share the excitement easily with anyone about my new results. You are unique, brilliant, and not limited to formality. It is really great to have worked with you these years.

Special thanks to the people who gave me inspiration and good starts to real bioinformatic tools/techniques, it prompted me to learn more. Johannes A Hofberger taught me how to make use of windows with batch scripts. Erik van den Bergh showed me how to use bash/awk for text processing. The first synteny network script was wrote by Rens Holmer in R. Saulo Aflitos gave me a live demonstration of writing a python script with multiple modules for my specific problem. Siavash Sheikhizadeh Anari, Jiao Long and Xuenan Pi from the bioinformatics group also helped me with particular problems. Thank you all for the kind help, it means a lot to me.

Big gratitude to my collaborators. The discussions, meetings and emails I had with you broadened my minds. Mariana Silva Artur, thank you for your big efforts to the LEAs chapter, which helped me graduate earlier! Thank Henk and Wilco for the constructive discussions. Thank Sumanth Mutte and Dolf Weijers for the very enlightening meetings, trully exciting when we brainstormed and talked about plant deep evolution. Thank you Baocheng Guo for our fish project, the impact of fish-specific whole genome duplications are as fascinating as plant polyploidy events. Thank Guodong Wang for sharing your cool results and collaborating in metabolic synthetic clusters. Thank Alejandro Pereira Santana, it was so happy to discuss and exchange ideas with you, thanks for the home-made Mexican food, best wishes. Thank Sander and Arman for the intriguing discussions about LTP and ABC transporters. Thank Annemiek for sharing your passion about bird evolution. Thank you Steven Arisz, I am so happy the everlasting DGAT1 story has finally been accepted on Plant Physiology, you have made great efforts, I admire your humor and preciseness, unfortunately no more pizza delivery ;) Special thanks to Harrold van den Burg, it was so exciting and joyful to hear your ideas at UvA. I very much appreciate your insights, and your contribution in improving the text and figures of that MADS-box manuscript.

I have some friends who have already graduated and left Netherlands. They gave me much help when I was a junior PhD student, I should always remember them. Thank you Heo Jaeyun, you are always so kind and gentle to the others and hard to yourself. Thank you Yanli Wang, Tao Li, Zhen Wei, Xiaomei Yang, thank you so much for your help and suggestions to my PhD study and the life here in Netherlands. Thank you Hieng Ming Ting and Bo Wang. I wish you all big success to your new careers in China.

Special thanks to Michel Arts, Michele Bandecchi, and Jose Francisco Martinez Carcamo. Guys, I really miss you and our spartacus workout, that was a glorious memory. That persistence and spirit will always stimulates me. Thank you Pingping Arts, you always give me the best advices. Thank you Marek Fumic, man, you understand me, do come to visit China some day.

Thanks my colleagues from Biosystematics! Thank you all for the help. Thank you Phuong, Deedi, Andreas, Floris, Setareh, Sara, Thijmen, Niccolo, Lotte, Eddie, Wei, Xiuji, Lidija and Laurie. Thank you Freek and Lars for the interesting conversations! Thank Sabrina, Nina, Nynke, Roel, Gabriella, and Marco. Thank you Klaas Bouwmeester for the discussions and your nice suggestions to my paper and thesis chapters. Thank you Wilma for so much help, and advices beyond work. Thank you Frank, you are always a very happy man, thank you for helping me with those plates and seeds-sowing in the lab, I will miss your laughters! Many thanks to Kitty Vijverberg, I have expanded my expertise and learned new knowledge from participating your dandelion project. I appreciate your sincerity and suggestions.

Thank my Chinese friends for your accompany. I guess I could just skip thanking you guys here one by one. Fubiao, Wenbiao, Jinling, Haidong, Xiao Lin, Mandy, Huicui Lu, Defeng, Tian Zeng, Fengjiao, Ying Zheng, Huayi, Jinbin, Peirong, as well as Bing Bai, Li Xie, Yin Song, Yongfu Tang, Jiao Long, Yajun, Huchen, Feng Zhu, Junwei, Wei Song, thank you all! Of course, also thank Jianhua Zhang, Guiling Ren, YuanYuan, YanXia, Xiaoxue, Kaile and Yiqian.

Also, I would like to thank Chinese government and my master school for providing opportunity and financial support for my oversea study. Also, thank my master supervisor Prof. Fengwang Ma, who gives me enormous support and encouragement all the time. With the completion of my PhD, I'm also finishing my twenty-four years school life as a student. Here I would also like to thank all my mentors at different stages who has given me edification and encouragement.

Special thanks to my thesis committee: Prof Vera van Noort, and Barbara Gravendeel from Leiden University, Prof Berend Snel from Utrecht University, and Prof Dick de Ridder from Wageningen University & Research.

I've saved the best until last. My little mermaid, I am blessed to have you in my life. Without you, I'd be a withered husk. Thank you for your never-ending supply of love, patience, and understanding. I never forget how lucky I am!

Finally, thank my family - my parents, my elder brother and my sister-in-law.

Tao Zhao

Date of Birth: 06-25-1987
Contact: zhaotaoshine@hotmail.com

List of Publications

Zhao, T. and Schranz, M.E., (2018). Comparative Phylogenomic Synteny Network Analysis of Mammalian and Angiosperm Genomes. bioRxiv, p.246736.

Arisz, S.A., Heo, J.Y., Koevoets, I.T., **Zhao, T.**, van Egmond, P., Meyer, J., Zeng, W., Niu, X., Wang, B., Mitchell-Olds, T. and Schranz, M.E., (2018). Diacylglycerol acyltransferase 1 contributes to freezing tolerance. *Plant Physiology*, pp-00503.

Gamboa-Tuz, S.D., Pereira-Santana, A., **Zhao, T.**, Schranz, M.E., Castano, E. and Rodriguez-Zapata, L.C., (2018). New insights into the phylogeny of the TMBIM superfamily across the three of life: Comparative genomics and synteny networks reveal independent evolution of the BI and LFG families in plants. *Molecular Phylogenetics and Evolution*, 126, 266-278.

Xie, Y., Chen, P., Yan, Y., Bao, C., Li, X., Wang, L., Shen, X., Li, H., Liu, X., Niu, C., Zhu, C., Fang, N., Shao, Y., **Zhao, T**., Yu, J., Zhu, J., Xu, L., van Nocker, S., Ma, F., Guan, Q. (2018). An atypical R2R3 MYB transcription factor increases cold hardiness by CBF-dependent and CBF-independent pathways in apple. *New Phytologist*, 218, 201-218.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A., and Schranz, M.E. (2017). Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *The Plant Cell* 29, 1278-1292.
(Recommended by Zahn, L.M., 2017. Genomics trace plant gene evolution. *Science*, 356, 1347.)

Zhao, T., and Schranz, M.E. (2017). Network approaches for plant phylogenomic synteny analysis. *Current Opinion in Plant Biology* 36, 129-134.

Li, W., Zhang, F., Chang, Y., **Zhao, T**., Schranz, M.E., and Wang, G. (2015). Nicotinate O-glucosylation is an evolutionarily metabolic trait important for seed germination under stress conditions in *Arabidopsis thaliana*. *The Plant Cell* 27, 1907-1924.

Zhao, T., Xia, H., Liu, J., and Ma, F. (2014). The gene family of dehydration responsive element-binding transcription factors in grape (*Vitis vinifera*): genome-wide identification and analysis, expression profiles, and involvement in abiotic stress resistance. *Molecular Biology Reports* 41, 1577-1590.

Zhao, T., Liang, D., Wang, P., Liu, J., and Ma, F. (2012). Genome-wide analysis and expression profiling of the DREB transcription factor gene family in Malus under abiotic stress. *Molecular Genetics and Genomics* 287, 423-436.

Education Experiences

Wageningen University
PhD student, Biosystematics of Plant Science Group, Sep. 2013 – Sep. 2018.
Chinese Scholarship Council (CSC) Fellowship: Sep. 2013 – Aug. 2017.
Supervisor: Prof. Dr M. Eric Schranz

Northwest A & F University
Master's degree, State Key Laboratory of Crop Stress Biology for Arid Areas, Sep. 2010 – Jul. 2013.
Supervisor: Prof. Dr Fengwang Ma

ShanDong Agricultural University
Bachelor's degree, College of Horticulture, Sep. 2006 – Jul. 2010.
Supervisor: Prof. Dr Xuesen Chen

Expertise & Research Interest

Genome Evolution
Phylogenomic Comparative Genomics
Network Analysis
Data Visualization

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to:	**Tao Zhao**
Date:	**17 September 2018**
Group:	**Biosystematics**
University:	**Wageningen University & Research**

1) Start-up phase	*date*
► **First presentation of your project**	
Comparative analysis of cold tolerance in Brassicaceae and Cleomaceae	9 Apr 2014
► **Writing or rewriting a project proposal**	
Comparative analysis of cold tolerance in Brassicaceae and Cleomaceae	Oct 2013 - Apr 2014
► **Writing a review or book chapter**	
Network approaches for plant phylogenomic synteny analysis. Current Opinion in Plant Biology (2017), 36:129-134, DOI:10.1016/j.pbi.2017.03.001	Apr 2017
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*13.5 **

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student day, Leiden, the Netherlands	29 Nov 2013
6th European Plant Science Retreat, Amsterdam, the Netherlands	01-04 Jul 2014
EPS PhD student days 'Get2Gether', Soest, the Netherlands	29-30 Jan 2015
► **EPS theme symposia**	
EPS theme 4 symposium 'Genome Biology', Wageningen, the Netherlands	13 Dec 2013
EPS theme 3 symposium 'Metabolism and Adaptation', Wageningen, the Netherlands	11 Mar 2014
EPS theme 4 symposium 'Genome Biology', Amsterdam, the Netherlands	15 Dec 2015
EPS theme 4 symposium 'Genome Biology', Wageningen, the Netherlands	16 Dec 2016
► **National meetings (e.g. Lunteren days) and other National Platforms**	
Annual meeting 'Experimental Plant Sciences', Lunteren, the Netherlands	14-15 Apr 2014
Annual meeting 'Experimental Plant Sciences', Lunteren, the Netherlands	13-14 Apr 2015
Annual meeting 'Experimental Plant Sciences', Lunteren, the Netherlands	11-12 Apr 2016
Annual meeting 'Experimental Plant Sciences', Lunteren, the Netherlands	10-11 Apr 2017
► **Seminars (series), workshops and symposia**	
Mini-workshop: Bioinformatics PAML	25 Nov 2013
PhD masterclass: Dr. Laxmi Parida	10 Mar 2016
Seminar: Joel Salatin - 'It's the food, my friend!'	9 May 2015
Seminar: Prof. Alain Tissier - 'Insights into the inner workings of a metabolic cell factory: the tomato glandular trichome'	18 Mar 2016
Seminar: Dr. Sanjay Kapoor - 'Regulators of Reproductive Development in Rice'	29 Aug 2017
Seminar: Prof. Peter Linder - 'Why is the Cape flora so species rich? Insights from the Cape reeds (Restionaceae)'	06 Oct 2017
Seminar: Prof. John (Jack) H. Werren - 'Evolution of new gene functions: lateral gene transfers and expression evolution'	29 Jan 2018
Seminar: Dr. Teemu Teeri - 'Pelargonidin in flowers - why not? Gerbera and petunia flowers block pelargonidin biosynthesis in a different way'	14 Mar 2018
Seminar: Dr. Ronald Snijder - 'Modern domestication of Pelargonium in a commercial environment'	09 May 2018
Seminar: Dr. Bob Schmitz - 'Epigenomic Studies of Natural and Induced Epialleles in Plants'	06 Jun 2018
Symposium: Food for Future - Technological Food Innovations, Wageningen, the Netherlands	22 Jun 2018
► **Seminar plus**	
► **International symposia and congresses**	
Evolutionary Genetics & Genomics, Cold Spring Harbor Asia Conference, Suzhou, China	08-12 Oct 2014
15th European Conference on Computational Biology (ECCB 2016), The Hague, the Netherlands	03-07 Sep 2016
International PSE Symposium 'Plant Omics and Biotechnology for Human Health', Gent, Belgium	21-24 Nov 2016
XIX International Botanical Congress (IBC 2017), Shenzhen, China	23-29 Jul 2017
The Plant & Animal Genome XXVI Conference (PAG XXVI), San Diego, CA, USA	13-17 Jan 2018
► **Presentations**	
Talk: EPS theme 4 symposium, Amsterdam, the Netherlands	15 Dec 2015
Talk: B-Wise seminar, Wageningen, the Netherlands	03 May 2016
Talk: PAG XXVI, Plant and Animal Paleogenomics, San Diego, CA, USA	14 Jan 2018
Talk: PAG XXVI, Digital Tools and Resources Session 3, San Diego, CA, USA	16 Jan 2018
Poster: European Plant Science Retreat 2014, Amsterdam, the Netherlands	01-04 Jul 2014

	Poster: Evolutionary Genetics & Genomics, Cold Spring Harbor Asia, Suzhou, China	08-12 Oct 2014
	Poster: Annual Meeting 'Experimental Plant Sciences', Lunteren, the Netherlands	13-14 Apr 2015
	Poster: XIX International Botanical Congress (IBC 2017), Shenzhen, China	23-29 Jul 2017
	Poster: PAG XXVI, Methods: Bioinformatics, San Diego, CA, USA	13-17 Jan 2018
►	**IAB interview**	
►	**Excursions**	
	Subtotal Scientific Exposure	*23.2 **

3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	Systems Biology Course 'Statistical analysis of ~omics data', Wageningen, the Netherlands	15-19 Dec 2014
	Graduate Course 'Phylogenetics: principles & methods', Wageningen, the Netherlands	17-19 May 2016
	Postgraduate course 'Basic Statistics', Wageningen, the Netherlands	Jun - Jul 2016
	BioSB course 'Algorithms for Biological Networks' (5th Edition), Wageningen, the Netherlands	25-29 Jun 2018
►	**Journal club**	
	Journal Club Biosystematics group	2014-2017
►	**Individual research training**	
	Subtotal In-Depth Studies	*8.5 **

4) Personal development		*date*
►	**Skill training courses**	
	Practical English Plus, Wageningen, the Netherlands	Mar - Jun 2014
	WGS PhD Workshop Carousel, Wageningen, the Netherlands	2 Jun 2014
	SURF Boot Camp with workshops Introduction to UNIX, HPC Cloud & Introduction to cluster computing, Eindhoven, the Netherlands	15 Jun 2017
	2nd Silk Road International Symposium for Distinguished Young Scholars, Yangling, China	15-20 Nov 2017
	Co-writing grant proposal for the China Exchange Programme (CEP)	Jul - Aug 2016
	Co-writing Vernieuwingsimpuls grant proposal (NWO-VICI)	Jun - Aug 2017
►	**Organisation of PhD students day, course or conference**	
►	**Membership of Board, Committee or PhD council**	
	Subtotal Personal Development	*4.2 **

TOTAL NUMBER OF CREDIT POINTS	***49.4 ****

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

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

The support provided by China Scholarship Council for Tao Zhao is kindly acknowledged. This work was supported by the grant from NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design: Tao Zhao

Printed by: DigiForce || ProefschriftMaken