Comparative Phylogenomic Synteny Network Analysis of Mammalian and Angiosperm Genomes

Tao Zhao, M. Eric Schranz*
Biosystematics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
*Corresponding author: eric.schranz@wur.nl

Abstract

Background: 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 microsynteny studies. There is now a wealth of completed mammalian (animal) and angiosperm (plant) genomes, two very important lineages that have evolved and radiated over the last ~170 million years. Genomic organization and conservation differs greatly between these two groups; however, a systematic and comparative characterization of synteny between the two lineages using the same approaches and metrics has not been undertaken.

Results: We have built complete microsynteny networks for 87 mammalian and 107 angiosperm genomes, which contain 1,464,753 nodes (genes) and 49,426,268 edges (syntenic connections between genes) for mammals, and 2,234,461 nodes and 46,938,272 edges for angiosperms, respectively. Exploiting network statistics, we present the functional characteristics of extremely conserved and diversified gene families. We summarize the features of all syntenic gene clusters and present lineage-wide phylogenetic profiling, revealing intriguing sub-clade lineage-specific clusters. We depict several representative clusters of important developmental genes in humans, such as *CENPJ, p53* and *NFE2*. Finally, we present the complete homeobox gene family networks for both mammals (including Hox and ParaHox gene clusters) and angiosperms.

Conclusions: Our results illustrate and quantify overall synteny conservation and diversification properties of all annotated genes for mammals and angiosperms and show that plant genomes are in general more dynamic.

Keywords: synteny networks, genome evolution, gene family dynamics, phylogenetic profiling, mammals, angiosperms

Background

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 [1, 2]. 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 [1, 3, 4]. 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 [5]. 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 [6-9]. 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) [10-15]. While the genomes of mammalian vertebrates show evidence of only two shared and very old rounds of WGD; often referred to as “2R” [16-18]. 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 CoGe SynFind [19, 20]) describe collinear blocks surrounding a locus identifier and visualize the blocks at the local genomic scale. With the abundance of new genomic data, the changes for multispecies collinearity visualization are only exacerbated. We have developed a network-based approach to organize and display local synteny [21, 22] and have applied it to understand the evolution of the entire MADS-box transcription factor family across 51 plant genomes as a proof of

74 principle of the method [22]. We identified several evolutionary patterns including
75 extensive pan-angiosperm retention of certain gene clades, ancient retained tandem
76 duplications and lineage-specific transpositions such as the floral patterning genes in
77 Brassicaceae [22]. Our approach can be scaled to analyze not just one gene family, but
78 all gene families across a lineage.

79 The aim of this study is to investigate and compare the dynamics and properties of the
80 entire synteny networks of all annotated genes for mammals and angiosperms. To this
81 end, we analyzed the syntenic properties of 87 mammalian and 107 plant genomes
82 (Figure 1) which represent most major phylogenetic clades of both mammalian and
83 angiosperm groups across ~170 million years of evolution [13, 23-25]. For mammals, the
84 species used covered the three main clades of Afrotheria, Euarchontoglires, and
85 Laurasiatheria, as well as first-branching groups like *Ornithorhynchus anatinus*
86 (platypus). For angiosperms, the species also cover three main groups of Monocots,
87 Superasterids, and Rosids, as well as first-branching groups such as *Amborella*
88 *trichopoda* (Figure 1). Some clades are more heavily represented than others such as
89 primates (human relatives) and crucifers (Arabidopsis relatives) due to research sampling
90 biases. Regardless, most major lineages are represented. Also, there are differences in
91 the overall quality and completeness of the genome assemblies used, but this was a
92 factor we wanted to analyze and assess using synteny analysis.

93
94

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 1) 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 1).

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 [as described in 21, 22]. 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 (7,569 times for mammals and 11,449 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. For example, the mammalian genomes of *O. anatinus*, *Galeopterus variegatus*, *Carlito syrichta, Manis javanica*, and *Tursiops truncates* (Figure 2a) and for angiosperms *Humulus lupulus*, *Triticum urartu*, *Aegilops tauschii*, and *Lemna minor* (Figure 2b).

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 1). 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 1,464,753 nodes (genes) and 49,426,268 edges (syntenic connections between genes) for mammals, and 2,234,461 nodes and 46,938,272 edges 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 (Supplementary Figure 2) 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.

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-Matt-Whitney test) and retention percentage (mean 0.88 for mammals compared to 0.71 for angiosperm; $P < 0.001$, Wilcoxon-Matt-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, Supplementary Table 2). 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, Supplementary Table 2). 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, Supplementary Table 2). 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

176 genes shared only between a small number or related species (Figure 3b, Supplementary
177 Table 2).

178 ***Comparative synteny dynamics of gene families of mammals and angiosperms***

179 We investigated if gene families with similar C-P synteny dynamics (high-high, high-low,
180 low-high, and low-low), might also have similar functional annotations (e.g. GO terms)
181 [26, 27]. We tested for pathway and gene-function enrichment of gene families within
182 each of the four C-P profiles for both mammals and angiosperms (Figure 3c and 3d).
183 Over-representative terms are shown in a word-cloud with font sizes indicating the p-
184 value (Fisher's exact test with Bonferroni correction). For mammals, gene families with
185 "high-high" profiles are functionally enriched in DNA metabolic processes, such as "DNA
186 replication" and "DNA repair". Interestingly Alzheimer disease-amyloid secretase pathway
187 (P00003) genes are enriched in this category (Figure 3c). By contrast, "low-low" gene
188 families include functions in immune responses and pathways (e.g., "cellular response to
189 xenobiotic stimulus", "Collagen degradation", "Biological oxidations"), enriched protein
190 classes are "major histocompatibility complex antigen (PC00149)" and "cell adhesion
191 molecule (PC00069)" (Figure 3c). The mammalian "high-low" group is enriched for genes
192 that function in DNA-templated gene transcription and DNA binding, such as KRAB box
193 transcription factors (PC00029) [28] (Figure 3c). As transcription factors bind specific
194 promoters and thus regulate a variety of developmental and environmental processes.
195 Moreover, transcription factors commonly consist of multiple members. Thus, it can be
196 hypothesized that some gene family members are highly conserved and genomically
197 constrained, while other members are versatile and transposed into new genomic
198 positions. Finally the "low-high" group is enriched for genes involved in translation (e.g.
199 "peptide biosynthetic process", "peptide metabolic process") and ribosomal component
200 (e.g. "ribosomal subunit", "ribonucleoprotein complex"), most enriched Reactome
201 Pathways are closely related to translation processes (e.g. "eukaryotic translation", "Cap-
202 dependent translation initiation"), as well as infectious disease related pathways (e.g.
203 "Influenza infection", "Influenza life cycle", and "Influenza viral RNA transcription and
204 replication") (Figure 3c).

205 The functional enrichment analysis of angiosperms shows a different pattern than for
206 mammals (Figure 3d). Plant "high-high" gene families are enriched for organelle
207 components (e.g. "organelle part", "intracellular organelle", "chloroplast part", "organelle
208 organization", and "plastid part"), as well as acetyltransferase, transferase and
209 methyltransferase proteins for the processes such as "DNA repair", "ncRNA metabolic
210 process" and "methylation" (Figure 3d). Many of these categories are plant-specific
211 related to photosynthesis. By contrast, the plant "low-low" group is enriched by defense
212 response genes such as "peptidase inhibitor activity", "endopeptidase inhibitor", and "ADP
213 binding". "Low-high" gene families function in nuclear part components (e.g. "intracellular
214 organelle lumen", "organelle lumen"), biosynthetic process (e.g. "organonitrogen
215 compound biosynthetic process", "cellular aromatic compound metabolic process"), cell
216 surface proteins (e.g. "synthesis of glycosylphosphatidylinositol (GPI)) and gene

expression (e.g. “RNA polymerase complex”, “nucleic acid binding”, “RNA polymerase II transcription initiation”). Interestingly, “high-low” part of plant genes function in cell wall (e.g. “plant-type primary cell wall biogenesis”, “cellulose biosynthetic process”, “beta-glucan biosynthetic process”) (Figure 3d). 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 (Supplementary Table 2). 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 [29] and plant P450s and F-box genes) (Supplementary Table 2).

Comparative synteny network clustering

We next performed a clustering analysis for the entire mammal and angiosperm synteny networks. We used Infomap [30] as the clustering algorithm due to its efficiency and accuracy in handling large graphs with millions of nodes and because it has consistently out-performed other available methods [31]. The clustering results for mammals and angiosperms are summarized and compared in terms of cluster-size distributions (Figure 4a and 4b), corresponding clustering coefficients (Figure 4c and 4d), and number of species included per cluster (Figure 4e and 4f).

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 4a). These two peaks could be further explained by Figure 4c and Figure 4e that depict the corresponding average clustering coefficient and number of species, respectively. We observe that the peak in cyan in Figure 4a is accompanied by a steady increasing trend of the clustering coefficient and the number of species involved (Figure 4c). A similar trend was observed for the clusters forming the peak in yellow due to WGD (Fig 4a). 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 4a). 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 4a). Representative examples are labeled on the curve, and further depicted in Figure 4g and Figure 4h.

In contrast, angiosperm genomes show a very large proportion of lineage-specific clusters on the far left (shaded in purple, Figure 4b). 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 4d) and the cluster size for these lineage-specific clusters is mostly between 10 to 30 (shaded in cyan, Figure 4f), 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 4b) 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 4b). There is a variation for the number of species per cluster for these large multi-gene families in angiosperms (Figure 4f).

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 (Figure 4a, 4g and 4h). It has been reported that *CENPJ* regulates brain size [32, 33], and primates have relatively larger brains [34, 35]. It is interesting that we found primates formed a lineage-specific *CENPJ* synteny cluster (Figure 4g and 4h) 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 [36, 37]. *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 [38-41] and *NFE2-NFE2L1-NFE2L2-NFE2L3*, also with broad roles such as regulation of oxidative stress, aging and cancer cell proliferation [42-44].

Comparative phylogenetic profiling of synteny clusters

To further visualize and understand genomic diversity, we performed phylogenetic profiling of all synteny clusters of mammals and angiosperms (Figure 5a and 5b). 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) (Figure 5a and 5b). Nearly empty rows of the less-syntenic species are consistent with the pairwise matrix in Figure 2.

For mammals, a very large proportion of all genes are syntenic and single copy (Figure 5a) 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 5a).

In contrast, in angiosperms only ~10% of clusters are syntenically conserved between eudicot and monocot species (Figure 5b). The remaining clusters are mostly lineage-specific clusters that appear as discrete columns (Figure 5b). This indicates that angiosperm genomes are highly fractioned and reshuffled, with abundant examples of specific clusters for particular phylogenetic lineages/plant families, such as Amaranthaceae, Brassicaceae, Poaceae, Fabaceae, Rosaceae, and Solanaceae (Figure 5b). 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 5b).

Traditional phylogenetic profiling data typically show only the presence/absence of a gene family. Whereas, our synteny-based phylogenetic 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.

Synteny network for homeobox genes of mammals and angiosperms

To summarize and further illustrate synteny cluster properties between mammals and angiosperms species, we display synteny networks for the entire homeobox multi-gene family for both lineages (Figure 5c and 5d). For the mammals, the well-known Hox clusters, derived from WGD and tandem duplications [45, 46], were visualized as two huge clusters (*Hox1-8* and *Hox9-13*) connected by EVX gene cluster (*EVX1* and *EVX2*) (Figure 5c). ParaHox genes [47] *PDX1*, *GSX1*, and *GSX2* form one highly inter-connected cluster (Figure 5c), while the other three ParaHox genes *CDX1*, *CDX2*, and *CDX3* form respective independent clusters (Figure 5c). Moreover, we have found the synteny cluster of *DLX1-4*, and *DLX6* [48], cluster of *LHX2, 6*, and *9* [49], cluster of *NKX2-1* and *2-4* [50, 51], and cluster of *CERS5* and *6* [52] (Figure 5c).

Plant homeodomain proteins have been classified in the literature into various groups based on sequence similarity of their homeodomains [53-55]. Here the syntenic connections across the full set of homeobox genes provide novel insights to the origin and relationships of all homeobox subfamilies (Figure 5d). Some examples include conserved clusters (OCP3, RPL, and ATH1) [56-58]; WGD-derived clusters (*KNAT3-5, HAT1-3-HB2-HB4, HDG1-HDG7-ANL2-FWA,* and *HDG2-HDG3-PDF2-ATML1*) [59, 60]; eudicot-specific clusters (*STM, KNAT7, KNAT2-KNAT6, WOX1-PFS2* and *HB22-HB51*) [61-63], and monocot-specific clusters (i.e. *Os01g60270, Os06g04850, Os08g19590*) [64] (Figure 5d).

Synteny networks provide a complementary method to more traditional phylogenetic approaches for investigating the ancestry and homology relationships of (large) multi-gene families. For example, synteny information identified ancient tandem origins and lineage-specific transpositions of angiosperm MADS-box genes [22, 65, 66]. We have analyzed the mammalian homeobox genes. We clearly show and verify that the

mammalian Hox genes appear as inter-connected synteny super-clusters and also find synteny connections to the ParaHox genes, consistent with the numerous previous reports [45-47]. 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.

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 [7]. 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 phylogenetic 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 1). 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 1).)

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 searchs. 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” [67]. 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.

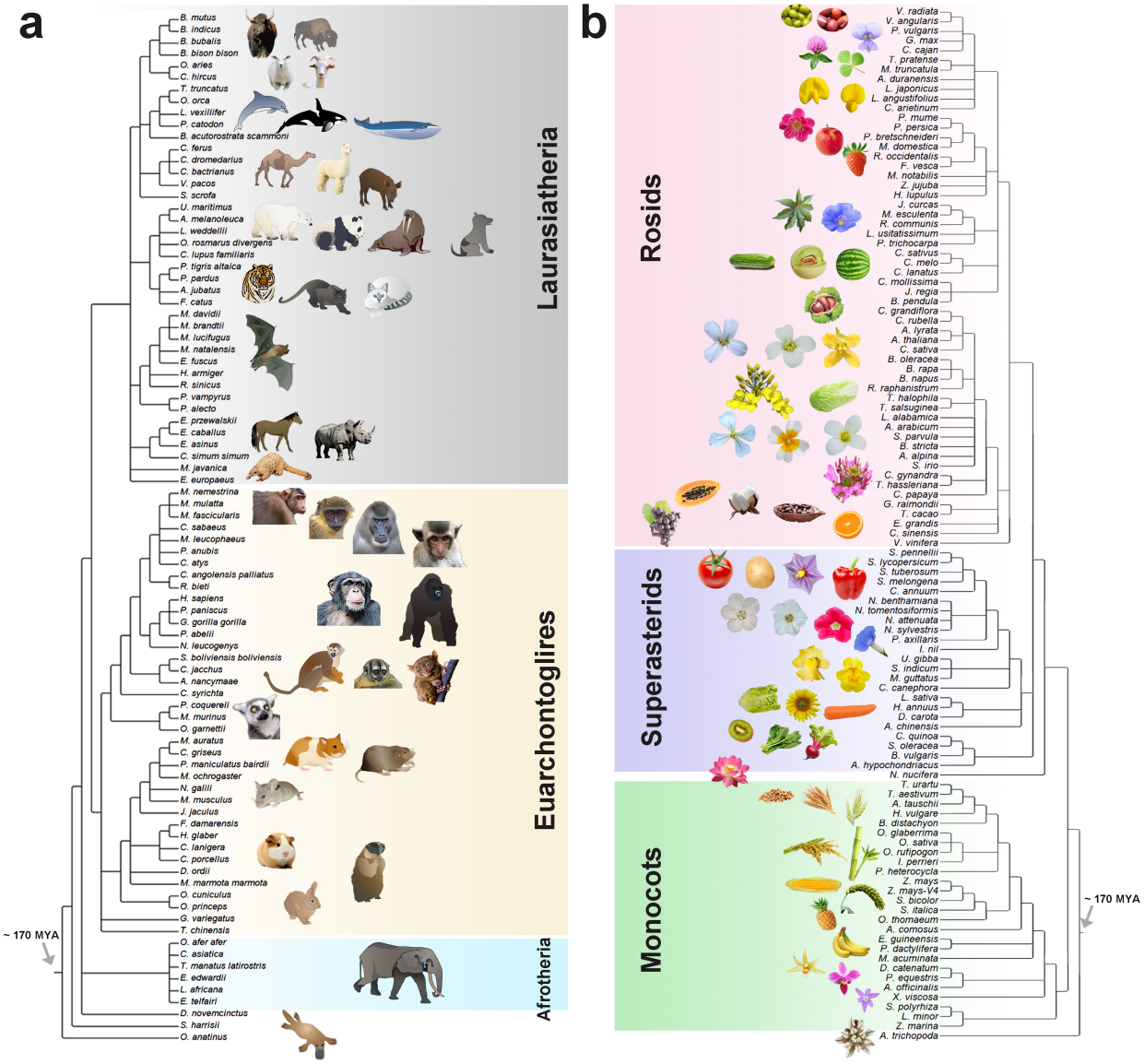
Gene annotation enrichment analysis

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 phylogenetic profiling

We used the infomap method to split the entire network, consisting of millions of nodes, into clusters [30]. 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 [68], then hierarchically clustered by “ward.D”, and visualized by “pheatmap”. We illustrate all the clusters of mammals (cluster size >= 2), and all angiosperm clusters with size >= 4.

436 **Figure Legends**



437

438 **Figure 1 Phylogenetic relationships of mammal and angiosperm genomes**
439 **analyzed.** (a) Mammal genomes used, highlighting the three main placental clades
440 Afrotheria, Euarchontoglires and Laurasiatherias. (b) Angiosperm genomes used,
441 highlighting the three main clades Monocots, Superasterids and Rosids.

442

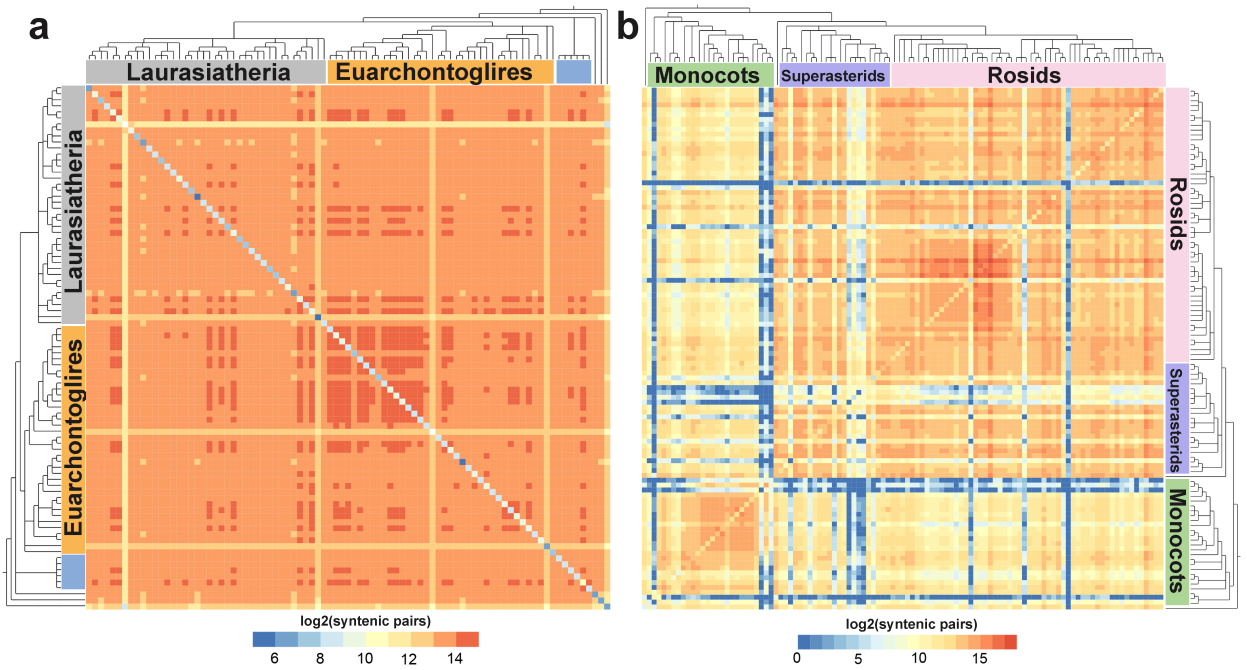


Figure 2 Pairwise synteny comparisons of mammal and angiosperm genomes. (a) Pairwise synteny comparison across Mammal genomes. (b) Pairwise synteny comparison across Angiosperm genomes. The logarithmic color-scale indicates the number of syntenic gene pairs. Species are ordered according to the consensus 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.

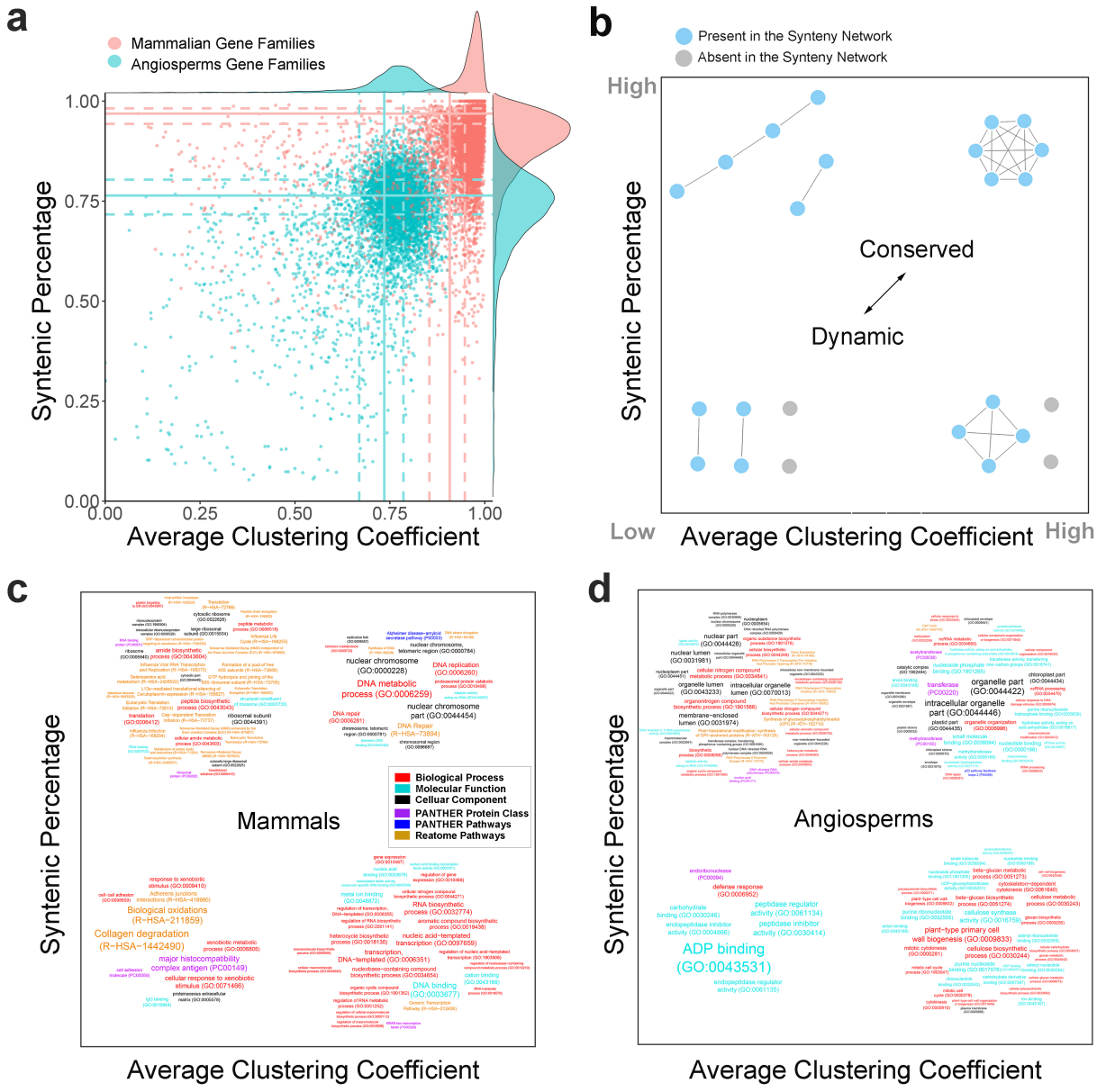


Figure 3 Network properties of gene families from mammal and angiosperm genomes. (a) Distributions of gene family dynamics of mammal (11,830 in red) and angiosperm (10,617 in blue) gene families plotted using percentage of syntenic genes and average clustering coefficients per family. Quartiles of average 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. (c, d) Comparative word clouds based on upper and lower quartiles for functional enrichment of significant terms with representative C-P profiles for mammals (c) and angiosperms (d). Font sizes are representative of adjusted p-values.

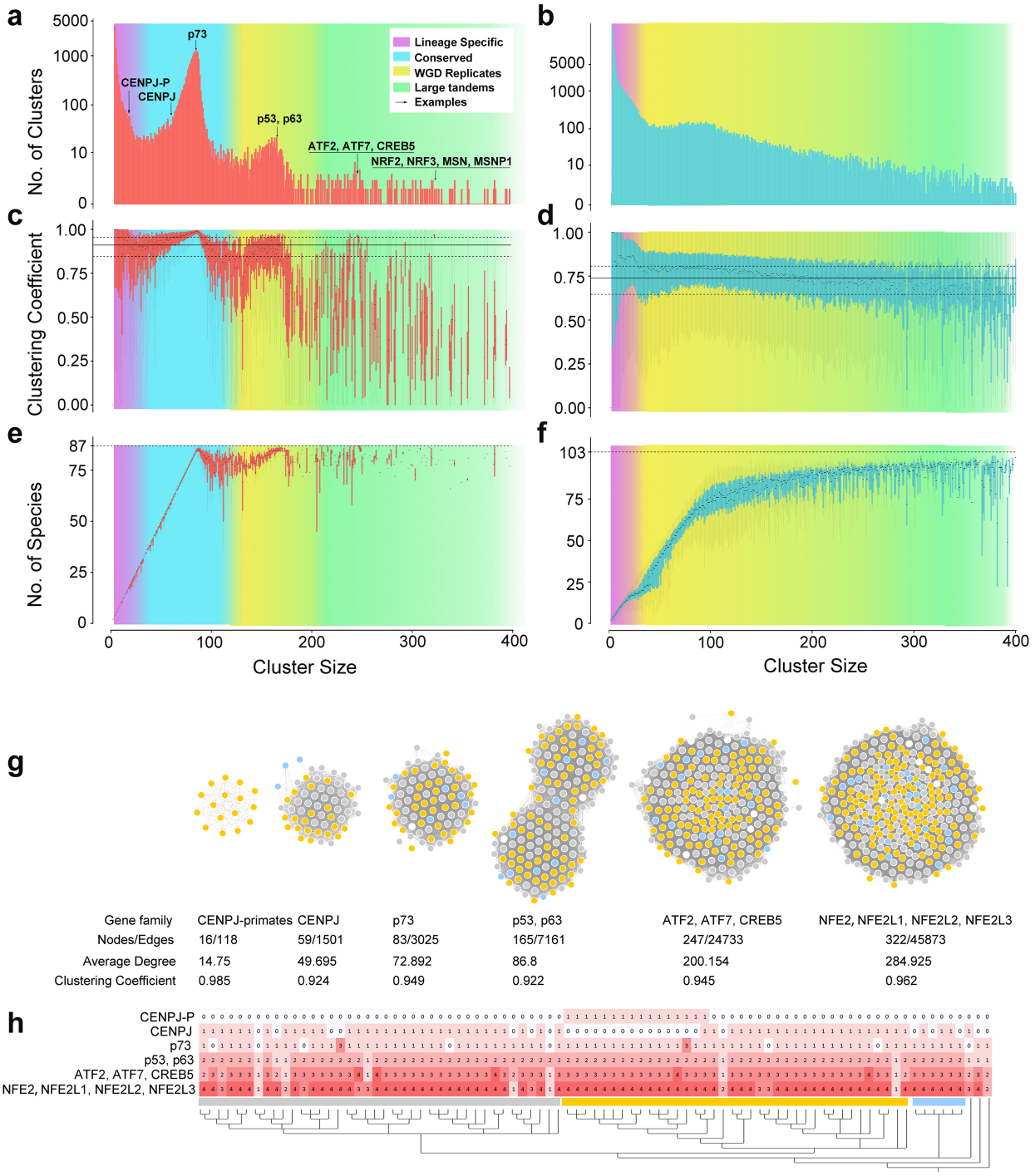


Figure 4 Synteny cluster statistics of mammal and angiosperm genomes and representative mammalian synteny clusters. Approximate size ranges for clusters of lineage-specific, conserved, WGD replicates, and large tandem genes are shaded in purple, cyan, yellow, and olive green, respectively. (a) Sizes 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

474 coefficient by mammalian cluster sizes. (d) Boxplot of clustering coefficient by angiosperm
475 cluster sizes. (e) Number of involving genomes for mammalian clusters by cluster sizes.
476 (f) Number of involving genomes for angiosperm clusters by cluster sizes. (g) Six
477 representative and diverse mammalian clusters of CENPJ (primate-specific one and the
478 others), p73, p53-p63, ATF2-ATF7-CREB5, and NFE2-NFE2L1-NFE2L2-NFE2L3. Total
479 number of nodes, edges, average degree, and clustering coefficient are indicated
480 accordingly below. (h) Phylogenetic profiling of the clusters from (g), a color gradient of
481 red indicates the number of syntelogs in each species.

482

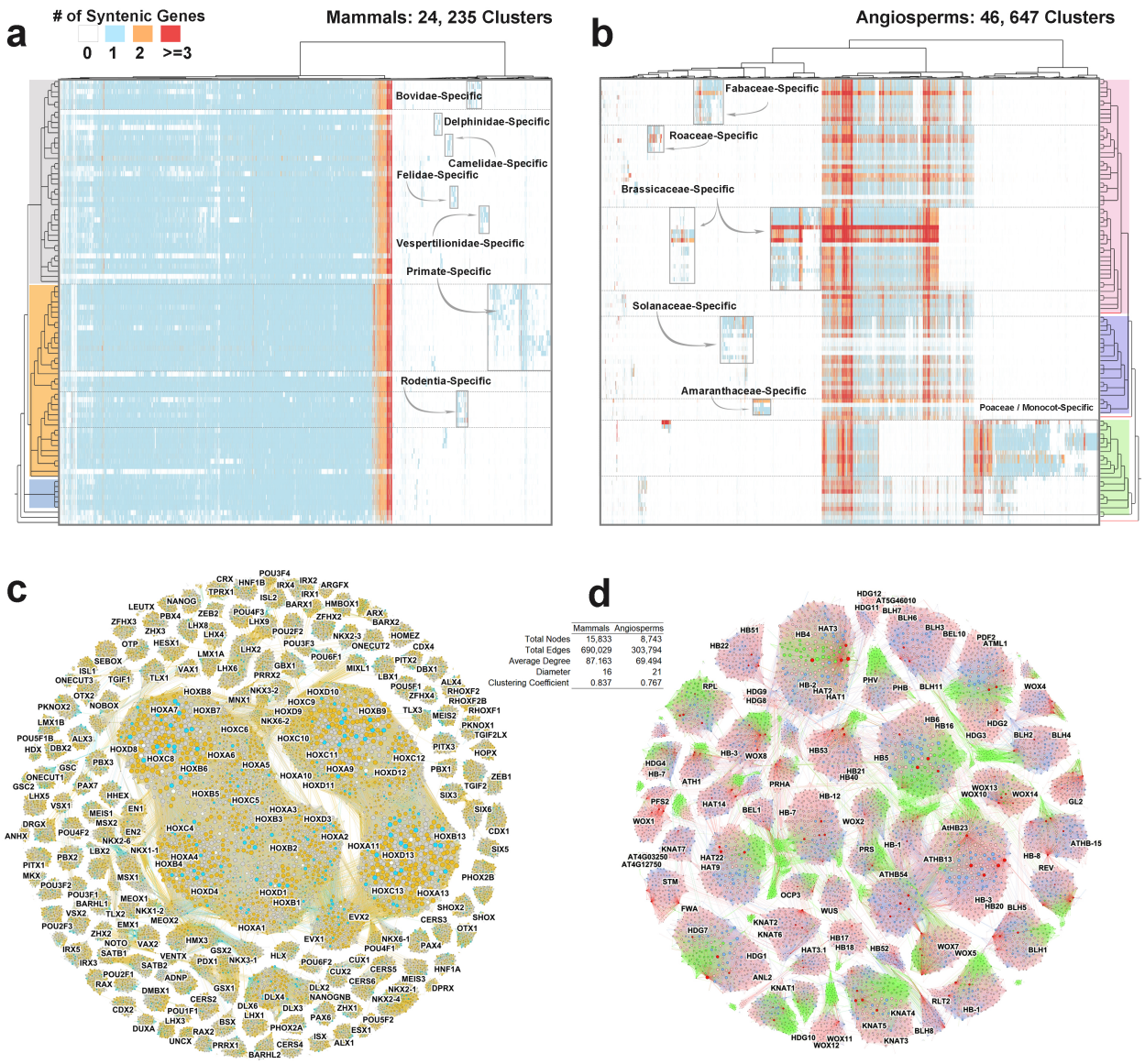
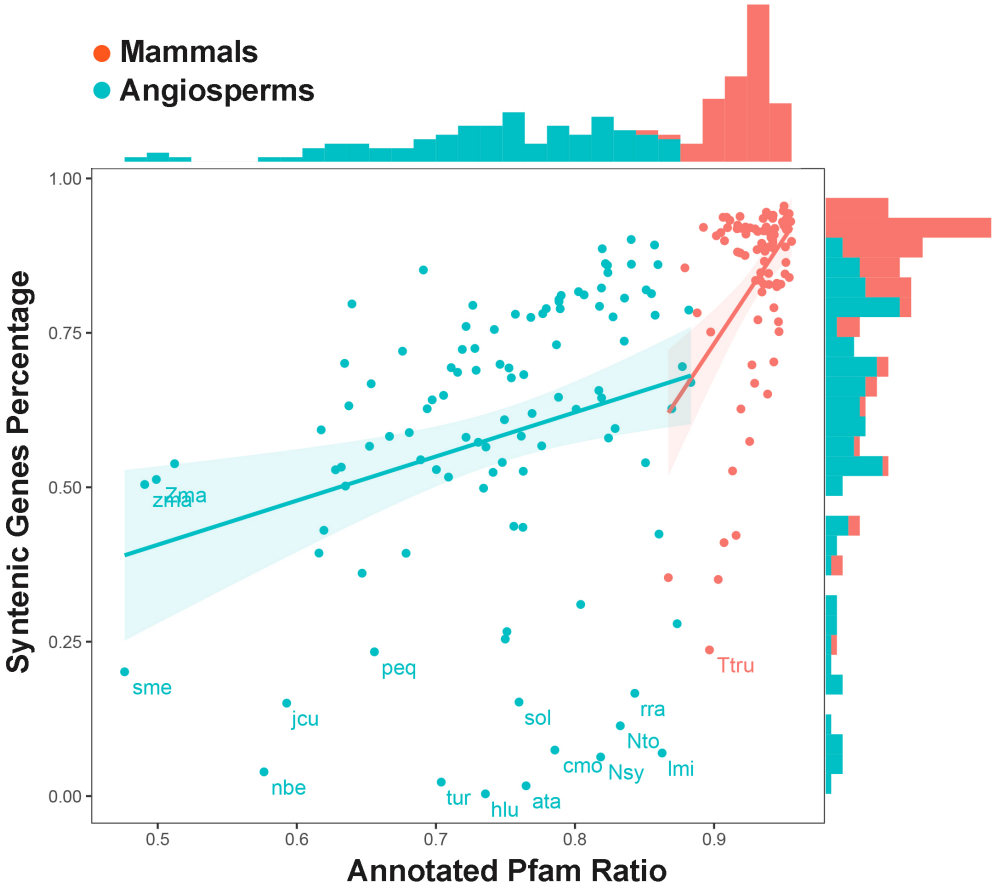


Figure 5 Phylogenetic profiling of all synteny clusters and complete Homeodomain multigene family synteny networks from mammal and angiosperm genomes. (a) Phylogenetic profiling of all mammalian clusters (size >= 2). Groups of lineage-specific clusters are boxed and labeled. (b) Phylogenetic profiling of all angiosperm clusters (size >= 3). Groups of lineage-specific clusters are boxed and labeled. (c, d) Synteny network of all homeo-domain proteins for mammals (c) and angiosperms (d), representative *H. sapiens* and *A. thaliana* genes are labeled, respectively.



492

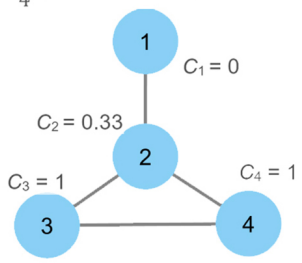
493 **Supplementary Figure 1** Plot of percentage syntenic genes again annotated (by Pfam)
494 percentage of all genomes. Species were highlighted with abbreviated names if syntenic
495 genes percentage lower than 0.25 or annotated proteins (by Pfam) lower than 0.5.

496

497

$$C_i = \frac{2L_i}{k_i\,(k_i - 1)}$$

$\bar{C}_i = \frac{1}{4}(1 + 1 + 0.33 + 0) = 0.58$



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*

498

499 **Supplementary Figure 2** Schematic diagram for the calculation of the average clustering
500 coefficient.

501 **Supplementary Table 1** Mammalian and angiosperm genomes used in this study

502 **Supplementary Table 2** Gene families with significant C-P features of mammals and
503 angiosperms.

504 **Declarations**

505 *Ethics approval and consent to participate*

506 Not applicable.

507 *Availability of data and material*

508 Data-sets and computer code used in this study are available at DataVerse:
509 (https://dataverse.harvard.edu/privateurl.xhtml?token=308d70cc-f489-435d-b7a5-
510 f4fc5acd4842). This includes the modified FASTA and BED files of all mammal and
511 angiosperm reference genomes. The scripts for network database preparation (pairwise
512 comparison, synteny block detection, and data integration), Pfam domain annotation,
513 network clustering and statistics, phylogenetic profiling, and for the figure preparation (if
514 applicable) are all included.

515 *Competing interests*

516 The authors declare that they have no competing interests.

517 *Authors' contributions*

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

521 *Acknowledgements*

522 TZ was supported by China Scholarship Council. Symbols for diagrams courtesy of the
523 Integration and Application Network (ian.umces.edu/symbols).

References

1. Dewey CN: Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform* 2011, 12:401-412.
2. Lv J, Havlak P, Putnam NH: Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. *BMC Bioinformatics* 2011, 12 Suppl 9:S11.
3. Koonin EV: Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005, 39:309-338.
4. Gabaldón T, Koonin EV: Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* 2013, 14:360-366.
5. Passarge E, Horsthemke B, Farber RA: Incorrect use of the term synteny. *Nature genetics* 1999, 23:387-387.
6. Law JA, Jacobsen SE: Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* 2010, 11:204-220.
7. Murat F, Peer YVd, Salse J: Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome biology and evolution* 2012, 4:917-928.
8. Gladyshev EA, Arkhipova IR: Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proceedings of the National Academy of Sciences* 2007, 104:9352-9357.
9. Feng S, Jacobsen SE, Reik W: Epigenetic reprogramming in plant and animal development. *Science* 2010, 330:622-627.
10. Adams KL, Wendel JF: Polyploidy and genome evolution in plants. *Current opinion in plant biology* 2005, 8:135-141.
11. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE: Polyploidy and genome evolution in plants. *Current opinion in genetics & development* 2015, 35:119-125.
12. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA: On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 2016, 210:391-398.
13. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al: Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011, 473:97-100.
14. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A: Widespread genome duplications throughout the history of flowering plants. *Genome research* 2006, 16:738-749.
15. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ: A genome triplication associated with early diversification of the core eudicots. *Genome biology* 2012, 13:R3.
16. Hokamp K, McLysaght A, Wolfe KH: The 2R hypothesis and the human genome sequence. In *Genome Evolution.* Springer; 2003: 95-110
17. Panopoulou G, Poustka AJ: Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. *TRENDS in Genetics* 2005, 21:559-567.
18. Steinke D, Hoegg S, Brinkmann H, Meyer A: Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC biology* 2006, 4:16.
19. Tang H, Bomhoff MD, Briones E, Zhang L, Schnable JC, Lyons E: SynFind: compiling syntenic regions across any set of genomes on demand. *Genome Biol Evol* 2015.
20. Lyons E, Freeling M: How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal* 2008, 53:661-673.
21. Zhao T, Schranz E: Network Approaches for Plant Phylogenomic Synteny Analysis. *Current Opinion in Plant Biology* 2017, 36:129-134.
22. Zhao T, Holmer R, de Bruijn S, Angenent GC, van den Burg HA, Schranz ME: Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific

Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation. *The Plant Cell* 2017, 29:1278-1292.

23. Cifelli RL, Davis BM: Marsupial origins. *Science* 2003, 302:1899-1900.
24. Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: The delayed rise of present-day mammals. *Nature* 2007, 446:507-512.
25. Magallón S, Gómez‐Acevedo S, Sánchez‐Reyes LL, Hernández‐Hernández T: A metacalibrated time‐tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* 2015, 207:437-453.
26. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R: Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell Online* 2016, 28:326-344.
27. Jiao Y, Paterson AH: Polyploidy-associated genome modifications during land plant evolution. *Philos Trans R Soc Lond B Biol Sci* 2014, 369.
28. Imbeault M, Helleboid P-Y, Trono D: KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 2017, 543:550-554.
29. Jones JD, Vance RE, Dangl JL: Intracellular innate immune surveillance devices in plants and animals. *Science* 2016, 354:aaf6395.
30. Rosvall M, Bergstrom CT: Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 2008, 105:1118-1123.
31. Lancichinetti A, Fortunato S: Community detection algorithms: a comparative analysis. *Physical review E* 2009, 80:056117.
32. Bond J, Roberts E, Springell K, Lizarraga S, Scott S, Higgins J, Hampshire DJ, Morrison EE, Leal GF, Silva EO: A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nature genetics* 2005, 37:353.
33. Gul A, Hassan MJ, Hussain S, Raza SI, Chishti MS, Ahmad W: A novel deletion mutation in CENPJ gene in a Pakistani family with autosomal recessive primary microcephaly. *Journal of human genetics* 2006, 51:760-764.
34. Kudo H, Dunbar R: Neocortex size and social network size in primates. *Animal Behaviour* 2001, 62:711-722.
35. Byrne RW, Corp N: Neocortex size predicts deception rate in primates. *Proceedings of the Royal Society B: Biological Sciences* 2004, 271:1693.
36. Levrero M, De Laurenzi V, Costanzo A, Gong J, Wang J, Melino G: The p53/p63/p73 family of transcription factors: overlapping and distinct functions. *J Cell Sci* 2000, 113:1661-1670.
37. Murray-Zmijewski F, Lane D, Bourdon J: p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death & Differentiation* 2006, 13:962-972.
38. Yoshida K, Maekawa T, Zhu Y, Renard-Guillet C, Chatton B, Inoue K, Uchiyama T, Ishibashi K-i, Yamada T, Ohno N: The transcription factor ATF7 mediates lipopolysaccharide-induced epigenetic changes in macrophages involved in innate immunological memory. *Nature immunology* 2015, 16:1034-1043.
39. Gupta S, Campbell D, Derijard B, Davis RJ: Transcription factor ATF2 regulation by the JNK signal transduction pathway. *SCIENCE-NEW YORK THEN WASHINGTON*- 1995:389-389.
40. Gozdecka M, Breitwieser W: The roles of ATF2 (activating transcription factor 2) in tumorigenesis. Portland Press Limited; 2012.
41. Bhoumik A, Fichtman B, DeRossi C, Breitwieser W, Kluger HM, Davis S, Subtil A, Meltzer P, Krajewski S, Jones N: Suppressor role of activating transcription factor 2 (ATF2) in skin cancer. *Proceedings of the National Academy of Sciences* 2008, 105:1674-1679.

42. **Chowdhury AMA, Katoh H, Hatanaka A, Iwanari H, Nakamura N, Hamakubo T, Natsume T, Waku T, Kobayashi A: Multiple regulatory mechanisms of the biological function of NRF3 (NFE2L3) control cancer cell proliferation. *Scientific reports* 2017, 7:12494.**
43. **Sykiotis GP, Bohmann D: Keap1/Nrf2 signaling regulates oxidative stress tolerance and lifespan in Drosophila. *Developmental cell* 2008, 14:76-85.**
44. **Kobayashi A, Ito E, Toki T, Kogame K, Takahashi S, Igarashi K, Hayashi N, Yamamoto M: Molecular cloning and functional characterization of a new Cap'n'collar family transcription factor Nrf3. *Journal of Biological Chemistry* 1999, 274:6443-6452.**
45. **Lemons D, McGinnis W: Genomic evolution of Hox gene clusters. *Science* 2006, 313:1918-1922.**
46. **Ferrier DE, Holland PW: Ancient origin of the Hox gene cluster. *Nat Rev Genet* 2001, 2:33-38.**
47. **Brooke NM, Garcia-Fernandez J, Holland PW: The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 1998, 392:920-922.**
48. **Panganiban G, Rubenstein JL: Developmental functions of the Distal-less/Dlx homeobox genes. *Development* 2002, 129:4371-4386.**
49. **Srivastava M, Larroux C, Lu DR, Mohanty K, Chapman J, Degnan BM, Rokhsar DS: Early evolution of the LIM homeobox gene family. *BMC biology* 2010, 8:4.**
50. **Sussel L, Marin O, Kimura S, Rubenstein J: Loss of Nkx2. 1 homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: evidence for a transformation of the pallidum into the striatum. *Development* 1999, 126:3359-3370.**
51. **Du T, Xu Q, Ocbina PJ, Anderson SA: NKX2. 1 specifies cortical interneuron fate by activating Lhx6. *Development* 2008, 135:1559-1567.**
52. **Pewzner-Jung Y, Ben-Dor S, Futerman AH: When do Lasses (longevity assurance genes) become CerS (ceramide synthases)? Insights into the regulation of ceramide synthesis. *Journal of Biological Chemistry* 2006, 281:25001-25005.**
53. **Schena M, Davis RW: HD-Zip proteins: members of an Arabidopsis homeodomain protein superfamily. *Proceedings of the National Academy of Sciences* 1992, 89:3894-3898.**
54. **Ariel FD, Manavella PA, Dezar CA, Chan RL: The true story of the HD-Zip family. *Trends in plant science* 2007, 12:419-426.**
55. **Mukherjee K, Brocchieri L, Bürglin TR: A comprehensive classification and evolutionary analysis of plant homeobox genes. *Molecular biology and evolution* 2009, 26:2775-2794.**
56. **Roeder AH, Ferrándiz C, Yanofsky MF: The role of the REPLUMLESS homeodomain protein in patterning the Arabidopsis fruit. *Current biology* 2003, 13:1630-1635.**
57. **Proveniers M, Rutjens B, Brand M, Smeekens S: The Arabidopsis TALE homeobox gene ATH1 controls floral competency through positive regulation of FLC. *The Plant Journal* 2007, 52:899-913.**
58. **Coego A, Ramirez V, Gil MJ, Flors V, Mauch-Mani B, Vera P: An Arabidopsis homeodomain transcription factor, OVEREXPRESSOR OF CATIONIC PEROXIDASE 3, mediates resistance to infection by necrotrophic pathogens. *The Plant Cell* 2005, 17:2123-2137.**
59. **Nakamura M, Katsumata H, Abe M, Yabe N, Komeda Y, Yamamoto KT, Takahashi T: Characterization of the class IV homeodomain-leucine zipper gene family in Arabidopsis. *Plant physiology* 2006, 141:1363-1375.**
60. **Carabelli M, Turchi L, Ruzza V, Morelli G, Ruberti I: Homeodomain-Leucine Zipper II family of transcription factors to the limelight: central regulators of plant development. *Plant signaling & behavior* 2013, 8:e25447.**
61. **Scofield S, Dewitte W, Nieuwland J, Murray JA: The Arabidopsis homeobox gene SHOOT MERISTEMLESS has cellular and meristem‐organisational roles with differential requirements for cytokinin and CYCD3 activity. *The Plant Journal* 2013, 75:53-66.**

665 **62.** **Li E, Wang S, Liu Y, Chen JG, Douglas CJ: OVATE FAMILY PROTEIN4 (OFP4) interaction with**
666 **KNAT7 regulates secondary cell wall formation in Arabidopsis thaliana. *The Plant Journal* 2011,**
667 **67:328-341.**
668 **63.** **Nakata M, Matsumoto N, Tsugeki R, Rikirsch E, Laux T, Okada K: Roles of the middle domain–**
669 **specific WUSCHEL-RELATED HOMEOBOX genes in early development of leaves in Arabidopsis.**
670 ***The plant cell* 2012, 24:519-535.**
671 **64.** **Jain M, Tyagi AK, Khurana JP: Genome - wide identification, classification, evolutionary**
672 **expansion and expression analyses of homeobox genes in rice. *The FEBS journal* 2008, 275:2845-**
673 **2861.**
674 **65.** **Ruelens P, de Maagd RA, Proost S, Theissen G, Geuten K, Kaufmann K: FLOWERING LOCUS C in**
675 **monocots and the tandem origin of angiosperm-specific MADS-box genes. *Nat Commun* 2013,**
676 **4:2280.**
677 **66.** **Cheng SF, van den Bergh E, Zeng P, Zhong X, Xu JJ, Liu X, Hofberger J, de Bruijn S, Bhide AS,**
678 **Kuelahoglu C, et al: The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait**
679 **and Genome Evolution of Crucifers. *Plant Cell* 2013, 25:2813-2830.**
680 **67.** **Csardi G, Nepusz T: The igraph software package for complex network research. *InterJournal,***
681 ***Complex Systems* 2006, 1695:1-9.**
682 **68.** **Dixon P: VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science***
683 **2003, 14:927-930.**
684
685