Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes

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

[a]Biosystematics Group, Wageningen University & Research, 6708PB Wageningen, The Netherlands

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and approved December 19, 2018 (received for review February 1, 2018)

A comprehensive analysis of relative gene order, or microsynteny, can provide valuable information for understanding the evolutionary history of genes and genomes, and ultimately traits and species, across broad phylogenetic groups and divergence times. We have used our network-based phylogenomic synteny analysis pipeline to first analyze the overall patterns and major differences between 87 mammalian and 107 angiosperm genomes. These two important groups have both evolved and radiated over the last ∼170 MYR. Secondly, we identified the genomic outliers or "rebel genes" within each clade. We theorize that rebel genes potentially have influenced trait and lineage evolution. Microsynteny networks use genes as nodes and syntenic relationships between genes as edges. Networks were decomposed into clusters using the Infomap algorithm, followed by phylogenomic copy-number profiling of each cluster. The differences in syntenic properties of all annotated gene families, including BUSCO genes, between the two clades are striking: most genes are single copy and syntenic across mammalian genomes, whereas most genes are multicopy and/or have lineage-specific distributions for angiosperms. We propose microsynteny scores as an alternative and complementary metric to BUSCO for assessing genome assemblies. We further found that the rebel genes are different between the two groups: lineage-specific gene transpositions are unusual in mammals, whereas single-copy highly syntenic genes are rare for flowering plants. We illustrate several examples of mammalian transpositions, such as brain-development genes in primates, and syntenic conservation across angiosperms, such as single-copy genes related to photosynthesis. Future experimental work can test if these are indeed rebels with a cause.

synteny networks | genome evolution | phylogenomic synteny profiling | mammals | angiosperms

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. The wealth of fully sequenced genomes of species across the phylogeny of mammals and angiosperms provides an excellent opportunity for comparative studies of evolutionary innovations underlying phenotypic adaptations (1). Phylogenetic profiling studies typically analyze the presence or absence of particular genes or gene families during the evolution of a lineage. For example, recent studies have investigated when particular gene families first evolved (2, 3) or have identified the loss of specific genes associated with a particular function (4–6). Less attention has been devoted to understanding changes in local gene position (genomic microcollinearity or microsynteny) in a phylogenetic context.

Synteny can be defined as evolutionarily conserved relationships between genomic regions. Synteny information provides a valuable framework for the inference of shared ancestry of genes, such as for assigning gene orthology relationships, particularly for large multigene families where phylogenetic methods may be nonconclusive (7–9). Finally, synteny data can speed the transfer of knowledge from model to nonmodel organisms.

While the basic characteristics of gene and genome organization and evolution are similar across eukaryote lineages, there are also significant differences that are not fully characterized or understood. 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 differ significantly between plants, animals, and other eukaryotes (10–13). It is known that genome organization and gene collinearity is substantially more conserved in mammals than plants (11), and thus identifying syntenic orthologs across mammals is more feasible and straightforward than in angiosperms. However, a comprehensive, comparative, and analytical analysis of microsynteny of all coding genes across these two groups has not yet been established. It is an opportune moment to do so due to the rapid increase in available completed genomes for these two groups.

One major characteristic of flowering plant genomes is the prevalent signature of shared and/or lineage-specific whole

Significance

Studying the organization of genes within genomes across broad evolutionary timescales can advance our understanding of the evolution of traits and clades. We have used a network approach to investigate genome dynamics of mammals and angiosperms. In general, genome organization and gene microcollinearity is much more conserved in mammals than in flowering plants. We then identified the genomic outliers or "rebel genes," within each clade. Genes that have moved are unusual for mammals, whereas highly conserved single-copy genes are exceptional for plants. How conservation and changes in synteny or fundamental differences in genome organization have contributed to the evolution of lineages could be a new scientific frontier.

Author contributions: M.E.S. designed research; T.Z. performed research; T.Z. and M.E.S. analyzed data; and T.Z. and M.E.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Datasets used in this study are available at DataVerse: (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BDMA7A). This includes protein sequences of all annotated coding genes in FASTA format and corresponding BED/GFF files indicating gene positions of all mammal and angiosperm reference genomes, as well as the entire microsynteny network database, subset of all BUSCO genes, and corresponding clustering results. The scripts for synteny network preparation (pairwise comparison, synteny block detection, and data integration), network clustering and statistics, and phylogenomic profiling are available at Github (https://github.com/zhaotao1987/SynNet-Pipeline).

[1]To whom correspondence should be addressed. Email: eric.schranz@wur.nl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1801757116/-/DCSupplemental.

Published online January 23, 2019.

genome duplications (WGDs) (14–19). In contrast, the genomes of mammals show evidence of only two shared and very old rounds of WGD, often referred to as "2R" (20–22). 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 reference genomes is growing exponentially, a major challenge is how to detect, represent, and visualize synteny relations across broad phylogenetic context. To remedy this, we have developed a network-based approach based on the *k*-clique percolation method to organize and display local synteny (23) and 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 (24). Such a network method is well suited for analyzing large complex datasets (25, 26) and is complementary to phylogenetic reconstruction methods that assume hierarchical bifurcating branching processes (27). Thus, independent and/or reciprocal changes in local gene synteny can be detected and assessed by analyzing network clusters in a phylogenetic context (i.e., phylogenetic profiling of synteny clusters, what we call "phylogenomic synteny profiling").

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. The goal then is to identify patterns of genome evolution that could provide insights into how genome dynamics have potentially contributed to trait evolution. To do so, we performed n^2 times (n stands for the number of species used) comparisons of all annotated genes, followed by $n(n+1)/2$ times synteny block detection using MCScanX (28) (Fig. 1*A*). All synteny blocks were integrated into one database. Syntenic genes derived from all inter- and intra species comparisons are interconnected into network clusters (Fig. 1*B*). The entire network database contains phylogenomic synteny trajectories of all of the annotated genes, which can be further utilized for specific purposes such as evaluating genome quality, characterizing relative syntenic strength, querying

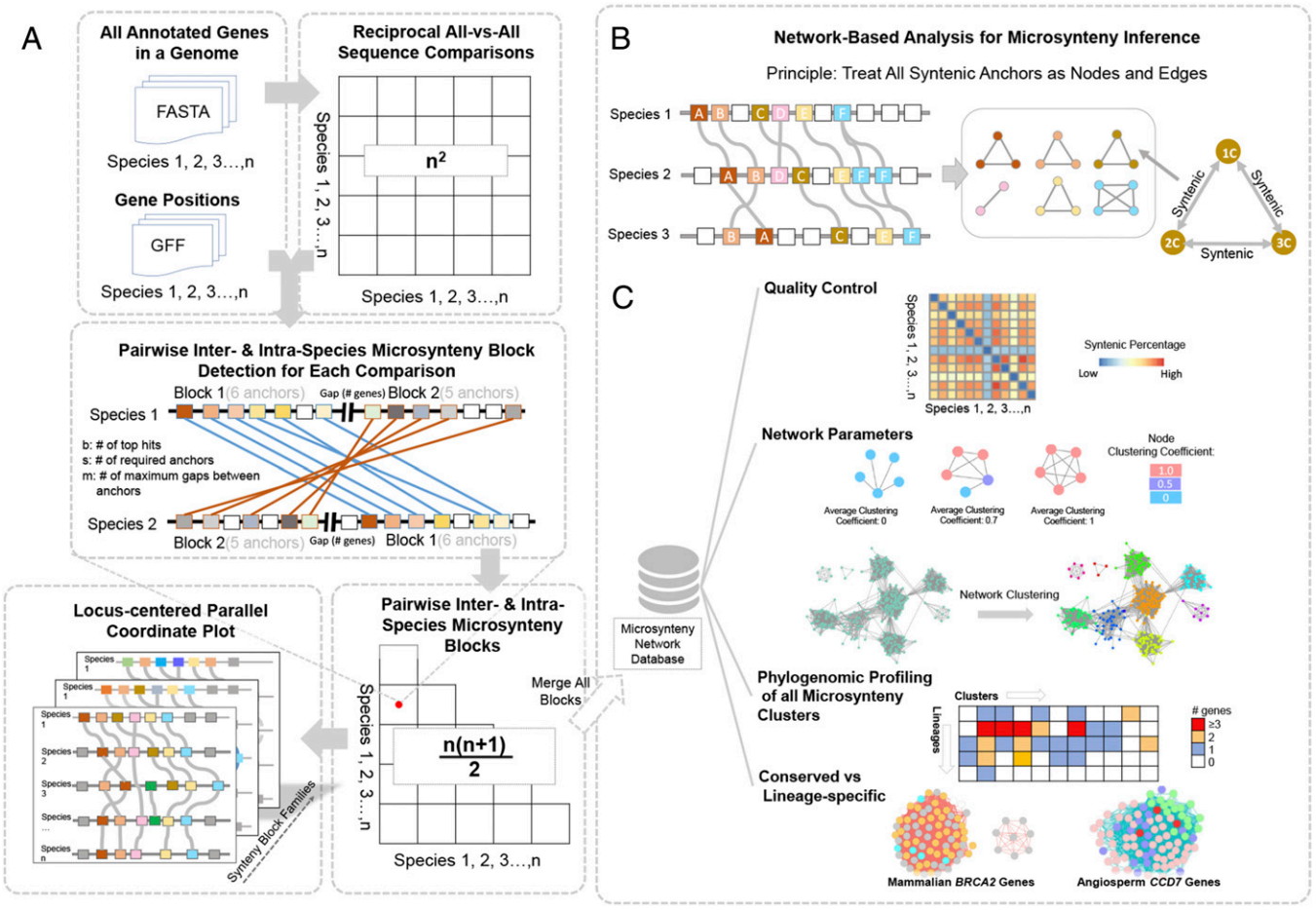

Fig. 1. Principles and applications of network-based microsynteny analysis. (*A*) For the genomes of *n* species, n^2 pairwise reciprocal all-vs.-all comparisons of all annotated genes are performed. Gene similarity relationships and relative gene positions are then used for collinearity/microsynteny block detection for each comparison (i.e., at least five syntenic anchor genes in a window of 20 genes). Syntenic anchor pairs were illustrated as colored boxes, black empty boxes represent nonsyntenic genes. All inter- and intraspecies blocks are extracted. Related blocks centered on a target locus (microsynteny block families) are traditionally organized into parallel coordinate plots. (*B*) Alternatively, we connect syntenic genes into clusters where nodes are genes and edges between the nodes means "syntenic"; cluster sizes depend on the number of related microsynteny blocks. (*C*) Network metrics and tools can then be utilized for a number of novel applications. For example, assessing overall genome quality that can be complementary to BUSCO. Principles of genome and gene family evolutionary dynamics across species can be inferred from network parameters such as clustering coefficients. Microsynteny network of multigene families can be decomposed using clustering algorithms. The clusters can then be analyzed by phylogenetic context (phylogenomic synteny profiling) to analyze gene copy number, long-term synteny conservation, and detection of lineage-specific changes in a syntenic context (i.e., gene transpositions).

particular gene families of interest, and phylogenomic synteny profiling (Fig. 1*C*). Synteny scores could be used as an alternative and complementary metric to other typical genome quality checks such as N50 values or the Benchmarking Universal Single-Copy Orthologs (BUSCO). Here we used Infomap as the network clustering method. Testing with various bench-marking input networks has shown Infomap to have excellent overall performance (29, 30). Infomap also scales better than the *k*-clique percolation method that we used in our previous study (24). The clusters produced are nonoverlapping. Furthermore, the number of clusters and cluster membership are determined by the algorithm, thus making results more comparable between different networks and independent from subjective user bias.

We analyzed the syntenic properties of 87 mammalian and 107 plant genomes (Fig. 2) which represent the main phylogenetic clades for both groups (17, 31–33). 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 in a phylogenetic context using synteny analysis. For mammals, the species used covered the three main clades of Afrotheria, Euarchontoglires, and Laurasiatheria, as well as basal 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* (Fig. 2). Some clades are more heavily represented than others such as primates (human relatives) and crucifers (*Arabidopsis*

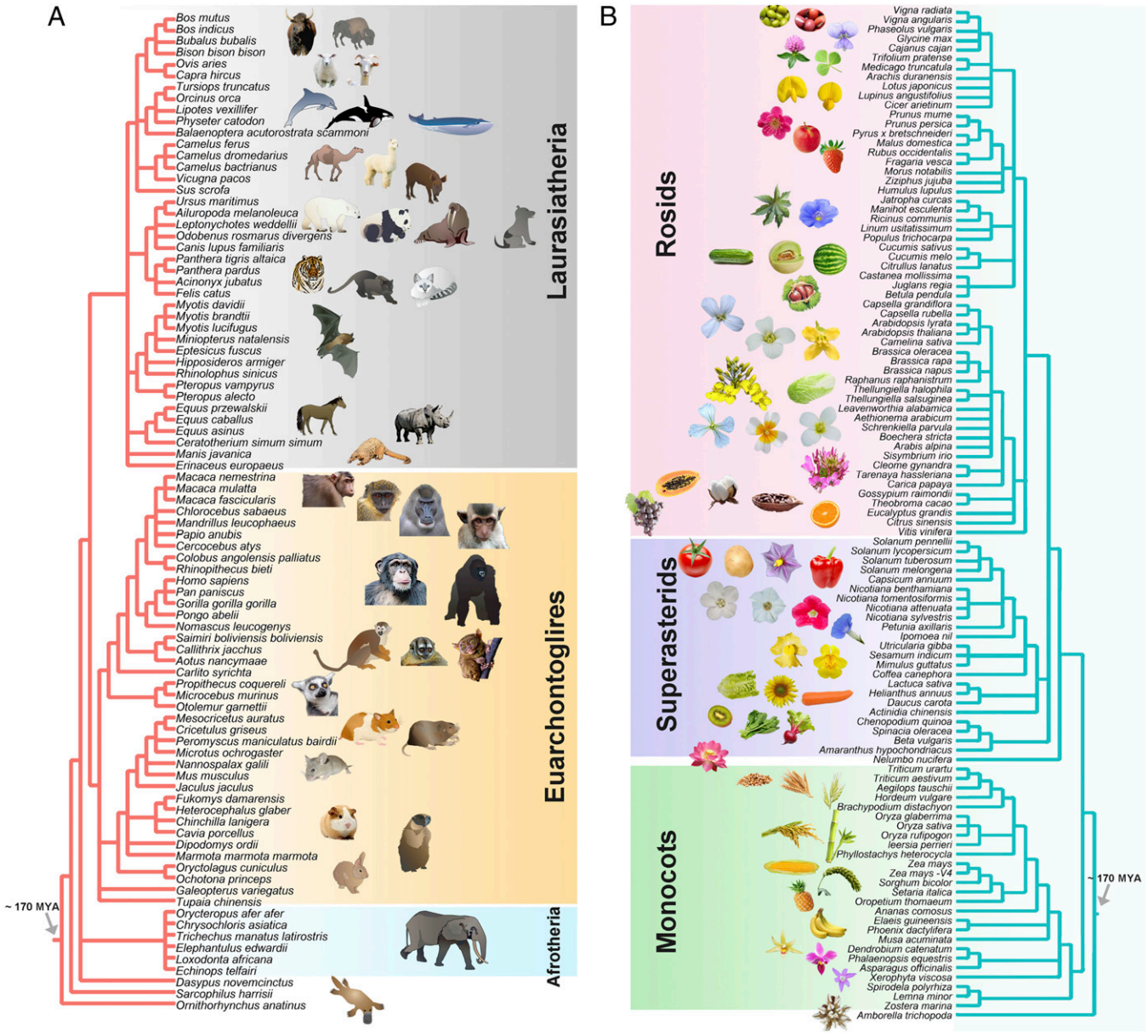

Fig. 2. Phylogenetic relationships of mammalian and angiosperm genomes analyzed. (*A*) Mammal genomes used (tree in red), highlighting the three main placental clades of Laurasiatherias (light-gray shading), Euarchontoglires (light-orange shading), and Afrotheria (light-blue shading). (*B*) Angiosperm genomes used (tree in blue), highlighting the three main clades of rosids (light-red shading), superasterids (light-purple shading), and monocots (light-green shading). The tree and clade shading is maintained in the latter figures. Mammal images courtesy of Tracey Saxby, Diana Kleine, Kim Kraeer, Lucy Van Essen-Fishman, Kate Moore, and Dieter Tracey, Integration and Application Network, University of Maryland Center for Environmental Science (ian.umces.edu/imagelibrary/).

relatives) due to research sampling biases. Mammalian and angiosperm linages have both evolved and radiated over the last ~170 MYR (17, 31–33) and have extremely rich research communities and a wealth of genomic resources, thus making such a comparative study of synteny of broad interest. Furthermore, we specifically identify unique sets of outliers between the two clades. In mammals, lineage-specific transpositions of genes are uncommon, whereas highly conserved syntenic single-copy genes are unusual in angiosperms. Being a "rebel gene" may be a signature of important or unique biological influence. The testing of this hypothesis could shed light on how genome dynamics may drive trait and lineage evolution.

Results and Discussion

Major Differences in Genomic Architecture Between Mammalian and Angiosperm Genomes Revealed by Pairwise Phylogenomic Microsynteny Analysis. Sequenced mammalian and angiosperm genomes were published at various qualities, as indicated by number of scaffolds, N50, and BUSCO (Dataset S1). Many are neither perfectly assembled nor annotated, with some poorly assembled genomes containing thousands of relatively small scaffolds. Since synteny detection based on genome annotations are subject to possible confounding factors, we tested 20 different settings, combining number of top hits for each gene (-b), and parameters of MCScanX (-m: MAX_GAPS, -s: MACH_SIZE) (*SI Appendix*, Fig. S1). Compared with angiosperms, we found mammals to be less sensitive to -m and -b, which indicates greater genome continuity and less impact of gene duplicates. The results show that under the same settings of -s and -m, increasing -b generally increases the pairwise syntenic percentages (except for mammals, under b15s3m25 and b20s3m25, compared with b5s3m25 and b10s3m25) (*SI Appendix*, Fig. S1 *A* and *B*). But this also leads to a decrease in the overall quality of detected syntenic blocks as reflected by the lower average clustering coefficients (*SI Appendix*, Fig. S1 *C* and *D*). Compared with angiosperm genomes, a lower -b for mammals generally increases the number of nodes while at the same time increasing the clustering coefficients. Mammalian genomes are also less sensitive to -m under the same -s (*SI Appendix*, Fig. S1*C*). Considering block quality and overall coverage, we used the setting of b5s5m15 for mammal genomes and b5m25s5 for angiosperm genomes for all subsequent synteny network analysis.

To assess the overall impact of phylogenetic distance, genome assembly quality and genome complexity, we summarized syntenic percentage (syntenic gene pairs plus collinear tandem genes relative to total number of annotated genes) for all pairwise comparisons of all annotated genes (3,828 times for mammals and 5,778 times for angiosperms) into color-scaled matrixes (Fig. 3) organized using the same species phylogenetic order as in Fig. 2.

The diagonals of the matrixes represent self- vs. self-comparisons and indicate the number of paralog/ohnolog pairs, that are indicative of recent and/or ancient WGDs (Fig. 3). 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 that we used. Identified poor-quality mammalian genomes include *O. anatinus* (platypus), *Galeopterus variegatus* (Sunda flying lemur), *Carlito syrichta* (Philippine tarsier), *Manis javanica* (Sunda pangolin), and *Tursiops truncates* (bottlenose dolphin) (Fig. 3*A*), and poor-quality

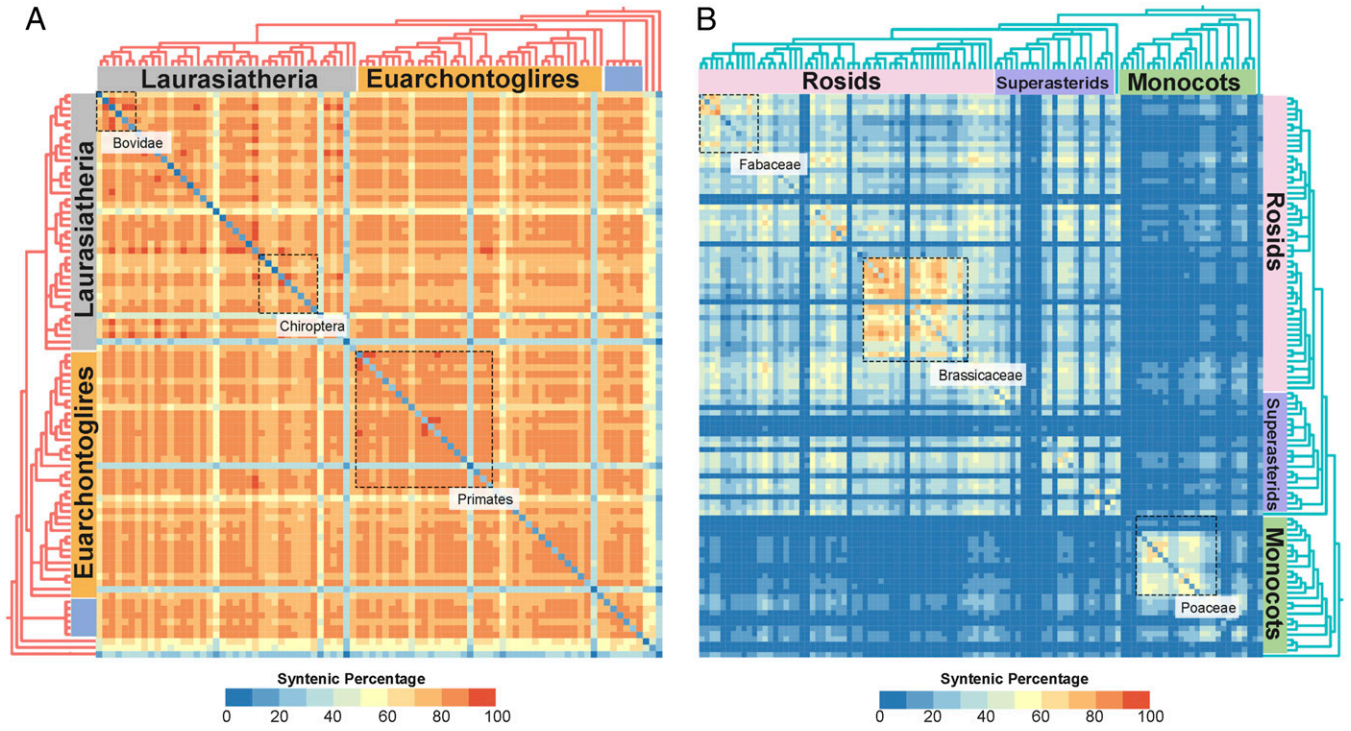

Fig. 3. Pairwise collinearity/microsynteny comparisons of mammalian and angiosperm genomes. (*A*) Pairwise microsynteny comparisons across mammal genomes. (*B*) Pairwise microsynteny comparisons across angiosperm genomes. The color scale indicates the syntenic percentage. Species are arranged according to the consensus phylogeny (Fig. 2). Overall, average microsynteny is much higher across mammals than plants. Also, the detected syntenic percentage does not show a strong phylogenetic signal. For example, contrasts are not higher for intra-Chiroptera (bats) or intra-Bovidae (cattle) than for distant pairwise contrasts. However, it is slightly higher for intraprimate contrasts, whereas, there is a much stronger phylogenetic signal seen for plant genomes such as intra-Brassicaceae or intra-Poaceae (grasses) contrasts than for interfamilial contrasts. The method also allows for easy detection of low-quality genomes. The diagonal for both plots represents intragenome comparisons which can detect potential recent and ancient WGDs. Note, that almost all plant genomes have higher intragenome microsyntenic pair scores than all mammal intragenome comparisons.

angiosperm genomes include *Humulus lupulus* (hop), *Raphanus raphanistrum* (wild radish), *Triticum urartu* (red wild einkorn wheat), *Aegilops tauschii* (Tausch's goatgrass), and *Lemna minor* (common duckweed) (Fig. 3*B*). For such species, better-quality genome assemblies/annotations than the ones we used (Dataset S1) hopefully will soon be available and thus improve the levels of synteny detected.

The matrices are based on all possible pairwise comparisons between genomes without correcting for phylogenetic distance. This was done to assess the effect of phylogenetic relationships on our results and to visualize overall differences in genome dynamics of mammals vs. angiosperms. As shown in the matrices, mammalian genomes overall are highly syntenic regardless of phylogenetic distance (Fig. 3*A* and Dataset S1) and groups with many completed genomes (such as bovines or bats) are not more obviously interconnected to one another. However, there is a slight increase in signal for primates (Fig. 3*A*). Whereas plant genomes show a stronger phylogenetic signal (e.g., grasses vs. grasses and crucifers vs. crucifers), the impact of recent WGD (e.g., *Brassica napus*) and more variability overall (due to assemblies/annotations from different research groups, different qualities, and multiple independent WGDs) (Fig. 3*B*). Almost all plant genomes have higher intragenome syntenic pair scores than all mammal intragenome comparisons due to the impact of ancient polyploidy.

To further illustrate the utility of our computed synteny scores for assessing genome quality, we compared it to more commonly used genome metrics and characteristics. Specifically, we plotted the average syntenic percentage against N50, genome size, number of scaffolds, and BUSCO (*SI Appendix*, Fig. S2). We found syntenic percentage was positively correlated to N50 and BUSCO and negatively correlated with genome size and the number of scaffolds (*SI Appendix*, Fig. S2). Mammalian genomes have significantly higher R-squared values (0.68) between BUSCO and syntenic percentage than that of the angiosperm genomes (0.35) (*SI Appendix*, Fig. S2). Synteny scores can thus provide alternative and complementary data for measuring and assessing genome quality, particularly for angiopsperms.

Distinct Network Properties of Phylogenomic Mammalian and Angiosperm Microsynteny Networks. The entire microsynteny networks are composed of all syntenic genes identified within all of the syntenic blocks. Specifically, there are 1,473,389 nodes (genes) and 50,396,484 edges (syntenic connections between genes) for mammals and 2,221,461 nodes and 47,737,321 edges for angiosperms, respectively (Fig. 4*A*). The average degree and clustering coefficient of the networks are significantly higher for mammals than that for angiosperms (mean node degree 68.4 for mammals compared with 43.0 for angiosperms; $P < 2.2e\text{-}16$ Mann–Whitney *U* test; mean clustering coefficient 0.88 for mammals compared with 0.65 for angiosperms; $P < 2.2e\text{-}16$ Mann–Whitney *U* test).

Fig. 4*B* shows the proportional degree distribution for the entire networks for mammals and angiosperms. The metrics for the two kingdoms are significantly different, but both distributions are clearly nonlinear (the scales of the axes are logarithmic), which would be the shape of scale-free networks if the distributions were governed by a power law (34). Specifically, for mammals a prominent peak occurs around node degree 50–100, where the corresponding cumulative fraction of nodes peaks rapidly from less than 0.2 to nearly 1 (especially around node degree 70–80 which represents the number of high-quality mammalian genomes). Such a curve indicates that most nodes have the same number of links and thus are very well interconnected (e.g., single-copy genes that are syntenic across all mammalian genomes). Comparatively, for angiosperms there are more nodes of lower node degree (over 25% for nodes with node degree less than 10). There are no major peaks observed; however,

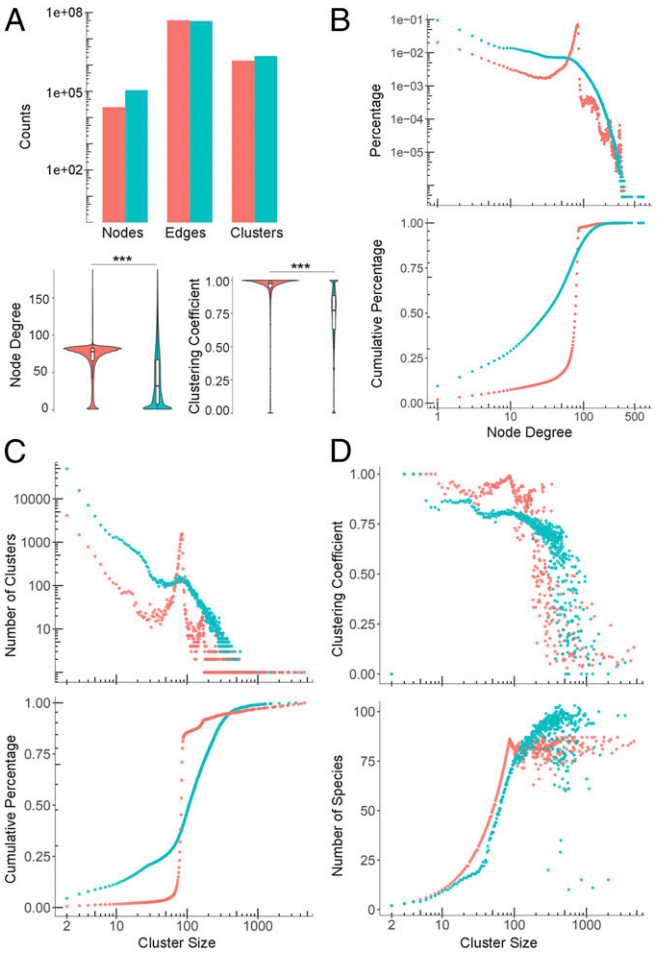

Fig. 4. Network statistics for mammal (red) and angiosperm (blue) microsynteny networks. (*A*) Number of total nodes, edges, and clusters. Note, compared with mammals, flowering plants have ~1.5 times total nodes, fewer (0.94) total edges, and ~4.5 times total number of clusters. Mammal mean node degree and clustering coefficient are significantly higher than that for flowering plants (***$P < 2.2e\text{-}16$). (*B*) Node degree distribution and corresponding cumulative percentage. The majority of mammal nodes peak around the degree 70–80. The scales of the axes are logarithmic. (*C*) Cluster size distribution by Infomap algorithm. Microsynteny cluster sizes vary from two to several thousand. (*D*) Corresponding clustering coefficient (median) and number of species (median) under certain sizes.

the distribution slightly bends from degree 10–30. Thus, there are many smaller nodes involving fewer taxa (e.g., extensive synteny is detected only across genomes from the same plant family).

The entire synteny networks of mammals and angiosperms were clustered into over 25,000 and 111,000 nonoverlapping clusters, respectively (Fig. 4*A*). We further summarized and compared the clustering results for mammals and angiosperms in terms of cluster-size distributions, corresponding clustering coefficients, and number of species included per cluster (Fig. 4 *C* and *D*). Overall, sizes of synteny clusters from mammal and angiosperm networks vary greatly from a minimum size of two up to thousands of nodes (Fig. 4*C*). This reflects the differences and dynamics of synteny conservation patterns among different genes and gene families. For example, clusters with bigger sizes could be genes maintained from several rounds of whole genome duplication events and/or tandem-duplicated arrays such as Hox genes, zinc finger proteins, and olfactory receptor genes in mammals and lectin receptor kinase genes and cytochrome P450 genes in angiosperms (Dataset S2). In contrast, small clusters could be

lineage-specific transpositions, for which synteny is shared only across a few closely related species such as transmembrane genes and keratin genes in mammals and F-box genes and NB-LRR genes in plants (Dataset S2).

Specifically, for mammals the cluster size distribution implies a strong correlation with its degree distribution, with the highest concentration of single-copy gene clusters around node size 70–100 (Fig. 4*C*). To the right, there is a second modest peak of duplicated (ohnolog) genes due to the ancient 2R WGD events (Fig. 4*C*). These peaks can be further understood by analyzing the corresponding average clustering coefficient and number of species relative to cluster size (Fig. 4*D*). We observe that the first peak is accompanied by a steady increasing trend of the clustering coefficient and the number of species involved (Fig. 4*D*). On the far left there is the rather modest proportion of lineage-specific genes, involving fewer species. Larger multigene families are found to the right where the number of species involved stays fairly constant but a general decrease in clustering coefficient is observed (Fig. 4*D*).

In contrast, angiosperm genomes show a very large proportion of lineage-specific clusters on the far left (Fig. 4*C*). For example, there are around 49,000 two-node clusters, accounting for ∼4.4% of the total nodes. Clusters with sizes ∼10–30, are mostly lineage specific as indicated by increased clustering coefficient (Fig. 4*D*). The size range reflects 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 is observed that are conserved across many lineages (Fig. 4*C*) 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.

Phylogenomic Synteny Profiling of All Gene Families Based on Microsynteny Networks Identifies Different Patterns of Conservation and Divergence. To classify conserved vs. specific genomic contexts, we profiled the patterns of gene copy number (0, 1, 2, and ≥3) across lineages and species of all of the clusters of mammals and angiosperms (Fig. 5 *A* and *B*). 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 (Fig. 5 *A* and *B*). Nearly empty rows of the less-syntenic species are consistent with the pairwise matrix in Fig. 3, very likely due to poor genome quality (*SI Appendix*, Fig. S2 and Dataset S1).

For mammals, a very large proportion (∼66%) of all clusters are largely syntenic and single copy (Fig. 5*A*) across all species with high-quality genomes. A smaller proportion of clusters (∼3.2%) are conserved and syntenic for duplicates derived from the 2R events or larger conserved multigene families (colored in red), for example gene clusters like the well-known Hox-gene clusters. We also detected lineage-specific clusters (∼23%) for mammalian clades with multiple species represented such as primates (including human, chimpanzee, macaque, and monkey), Rodentia (including hamster, mouse, and rat), Chiroptera (including bats and flying foxes), Felidae (including tiger, cheetah, and the house cat), Camelidae (including camels and alpaca), and Bovidae (including yak, cow, sheep, and goat) (Fig. 5*A*). These lineage-specific transpositions in mammals are the genomic outliers.

In contrast, for angiosperms only ∼8.7% of clusters are syntenically conserved between eudicot and monocot species (Fig. 5*B*). Strikingly, the remaining clusters are mostly lineage-specific clusters that appear as discrete columns (Fig. 5*B*). 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 (including quinoa, spinach, beet, and grain amaranth), Brassicaceae (including *Arabidopsis*, cabbage, and radish), Poaceae (including wheat, barley, rice, and maize), Fabaceae (including soybean, mung bean, red clover, and medicago), Rosaceae (including apple, peach, pear, and strawberry), and Solanaceae (including tomato, potato, pepper, petunia, and tobacco) (Fig. 5*B*). Such specific clusters were caused by transpositions and/or fractionation after WGD, which leads to changes/movements of genomic context. Results also highlight species with more gene copies per cluster (e.g., orange/red rows), likely due to recent WGD events

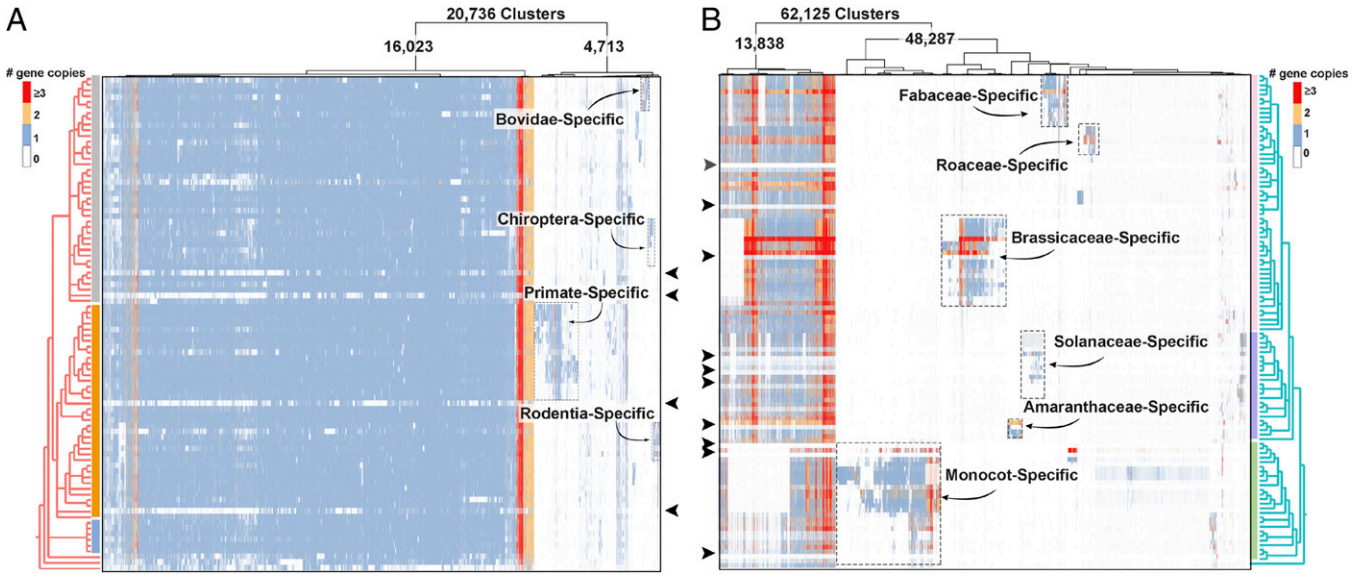

Fig. 5. Phylogenomic synteny profiling of mammal and angiosperm genomes. (*A*) Phylogenomic synteny profiling (copy-number profiling of microsynteny clusters across a phylogeny) of all mammalian clusters (size ≥ 3). Groups of lineage-specific clusters are boxed and labeled. (*B*) Phylogenomic synteny profiling of all angiosperm clusters (size ≥ 3). Groups of lineage-specific clusters are boxed and labeled. Black arrows mark nearly empty rows which indicate a poor genome quality. Overall, mammals have mostly syntenic (conserved) and single-copy genes, whereas angiosperms have many multicopy and/or lineage-specific microsynteny clusters.

such as for *Glycine max*, *B. napus*, and *Populus trichocarpa* (Fig. 5*B*). Thus, we observe a dramatically different pattern of genomic outliers in angiosperms than in mammals. It is the single-copy highly syntenic genes that represent the gene rebels in flowering plants.

In our earlier proof-of-principle publication, we analyzed the plant MADS-box gene family for angiosperms (24). The homeodomain family is a large multigene family in both plants and animals, playing critical roles in development, including the well-known Hox-gene clusters in animals. As a comparative example of an entire gene family for both mammals and plants, we give the complete homeodomain (35, 36) gene families for both lineages (*SI Appendix*, Fig. S3). We clearly show and verify that the mammalian Hox genes appear as interconnected synteny superclusters and also find synteny connections to the ParaHox genes, consistent with the numerous previous reports (37–39) (*SI Appendix*, Fig. S3). 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 (*SI Appendix*, Fig. S3).

Syntenic Properties of Mammal and Angiosperm BUSCO Genes. BUSCO genes are near-universal single-copy orthologs from OrthoDB (https://www.orthodb.org/), which are used to assess genome qualities and also as candidates for large-scale phylogenetic studies (40). Thus, investigating their positional history can provide complementary data for evolutionary studies. We identified candidate orthologous genes of 4,104 (mammalia_odb9) and 1,440 (embryophyta_odb9) benchmarking BUSCOs (gene families) for mammals and angiosperms from our dataset (Dataset S3). Although many BUSCO families are conserved as single-copy number and syntenic across species, we find many examples of both copy-number variation and of changes in genomic context across the phylogenies (i.e., multiple synteny clusters).

We use the number of synteny clusters to overall characterize synteny properties of BUSCOs. For example, if a BUSCO family is syntenic across all species, it would belong to only one synteny cluster. Overall 87.5% of mammal BUSCO families belong to only one synteny cluster and 11.8% of mammal BUSCO families have two clusters (Fig. 6*A*, *Left* and Dataset S4). Comparatively, only 11.9% of the plant BUSCO families have only a single synteny cluster. A total of 20.6% plant BUSCOs have two clusters, 19.5% plant BUSCOs have three clusters, and 21% plant BUSCOs have over five clusters (here no restriction to cluster sizes, minimum two nodes) (Fig. 6*B*, *Left* and Dataset S4). Changes in genomic context of benchmarking genes can provide important new complementary information for researchers using BUSCO genes to assess genome quality and for evolutionary studies. In particular, rebel genes could be particularly informative. Namely, two or more synteny clusters in mammals are less common whereas single-copy synteny clusters are unusual for angiosperms.

We highlight several examples of rebel genes that potentially have contributed to trait and lineage evolution. For mammals, we show lineage-specific gene transpositions (two synteny cluster BUSCO) with important functions. A gene that is transposed to a new genomic context could easily lead to new mechanisms of gene molecular evolution and/or regulation. For example, we found *BREAST CANCER 2* (*BRCA2*) and *Transformation/Transcription Domain Associated Protein* (*TRRAP*) genes form lineage-specific clusters for Chiroptera species (including all nine bat genomes used in this study: *Myotis brandtii*, *M. davidii*, *M. lucifugus*, *Miniopterus natalensis*, *Eptesicus fuscus*, *Hipposideros armiger*, *Rhinolophus sinicus*, *Pteropus alecto*, and *P. vampyrus*) [Fig. 6*A*, *Right*, cluster 4171(*BRCA2*) and cluster 4120 (*TRRAP*) in Dataset S4]. Both of the genes are known oncogenes also with roles in normal development (41–45). Zhang et al. (46) hypothesized that the evolution of flight in bats is linked to changes in metabolic capacity which would also require changes to DNA repair and DNA checkpoint genes, such as *BRCA2* which they reported to be positive selection. Interestingly, TRRAP also links to the DNA repair pathway (45, 47). Such lineage-specific transpositions of key genes like *BRCA2* and *TRRAP* potentially have facilitated the adaptive evolution of flight in bats.

We also found primate-specific clusters, including *MPHOSPH10* (*CT90*) and *CENPJ* (*CPAP*) [Fig. 6*A*, *Right*, cluster 4097 (*MPHOSPH10*) and cluster 4104 (*CENPJ*) in Dataset S4]. *MPHOSPH10* is an M-phase phosphoprotein 1 that localizes to the nucleolus and has been associated with the progression of some cancers (48). *CENPJ* (centromere protein J) is needed for normal spindle morphology

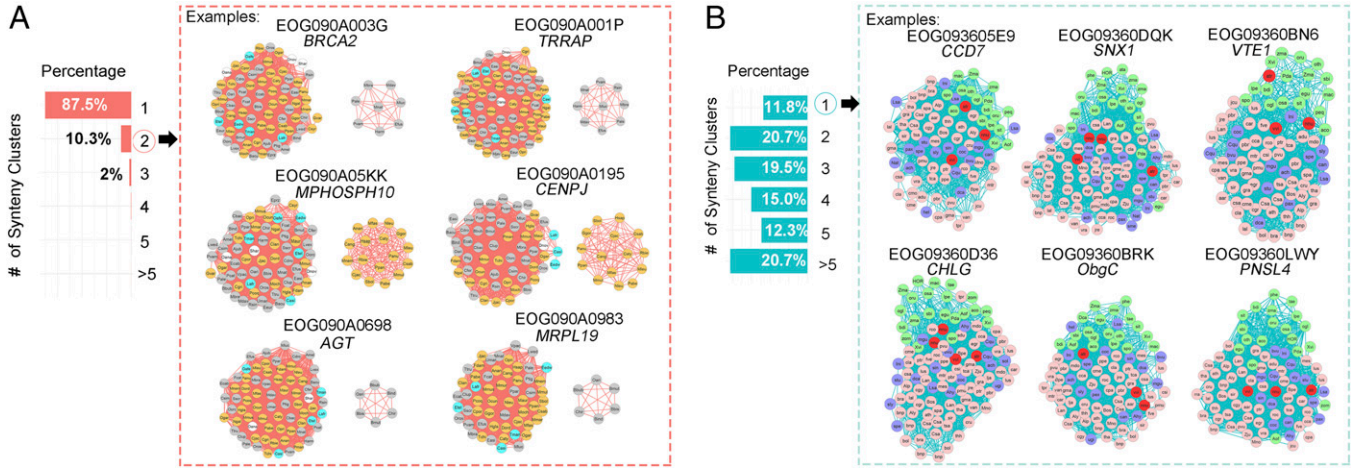

Fig. 6. Overall microsynteny conservation and examples of mammal and plant BUSCO genes. (*A*) Bar plot shows overall percentage of mammal BUSCOs that belong to certain number of synteny clusters. Most mammal BUSCO genes belong to only one synteny cluster. Examples of mammal BUSCO families which have two clusters are highlighted, including the oncogenes *BRCA2* and *TRRAPP* (Chiroptera specific), *MPHOSPH10* and *CENPJ* are associated with cell-divisions and possibly brain development (primate specific), and the peptide hormone angiotensin *AGT* and *MRPL19* (Bovidae specific). (*B*) Bar plot shows overall percentage of plant BUSCOs that have certain number of synteny clusters. Examples are highlighted of BUSCO gene families that belong to one synteny cluster, which are involved in hormone signaling (*CCD7* and *SNX1*) and photosynthesis (*VTE1*, *CHLG*, *ObgC*, and *PNSL4*). Node colors indicate lineages which are consistent with Fig. 3. Nodes for *Vitis vinifera* (basal rosids), *Nelumbo* (basal eudicots), and *Amborella* (basal angiosperm) are labeled red. Node labels are letter-coded species names which can be found in Dataset S1.

and it is involved in microtubule disassembly at the centrosome. Interestingly, changes in brain organization and brain size have been linked to changes in cell numbers and divisions, including specifically linked to *CENPJ* (49, 50). Primates have relatively larger brains compared with other mammals (51, 52). Note, two genes flanking *CENPJ* (namely *RNF17* and *ATP12A*) are cotransposed in primates [cluster 16942 (*RNF17*) and cluster 14351 (*ATP12A*) in Dataset S2]. The unique genomic context of primate genes potentially facilitated new and/or altered regulatory patterns and gene functions.

As a third set of mammalian rebel genes, we show Bovidae-specific clusters for *AGT* (also known as *ANHU*; *SERPINA8*; *hFLT1*) and *MRPL19* genes [Fig. 6*A*, cluster 4162 (*AGT*) and cluster 4159 (*MRPL19*) in Dataset S4]. *AGT* encodes the peptide hormone angiotensin that helps maintain blood pressure, body fluids, and electrolyte homeostasis (53, 54). It has been linked to both the control of thirst and to ovulation in cattle and sheep. *MRPL19* encodes a component of the mitochondrial large ribosomal subunit (mt-LSU) and is tightly linked to another gene that is also transposed (Dataset S4), GCFC2/C2orf3, that has recently been reported to be involved in intron splicing (55). The *MRPL19-C2orf3* gene pair is associated with dyslexia in humans (56, 57). How and if the transposition of angiotensin and dyslexia-related genes have affected bovines is unknown, but hopefully our results will generate hypotheses to be tested.

While changes in synteny patterns such as lineage-specific transpositions are the exception in animals, conserved synteny of single-copy genes are the rebel genes in flowering plants. For plant BUSCO gene families, we observed only 11.8% of angiosperm-wide conserved synteny clusters, for example, clusters for *CCD7* (cluster 280) and*SNX1* (cluster 27) genes (Fig. 6*B* and Dataset S4). *CCD7* (or *MAX3*) is required for the biosynthesis of strigolactones which are phytohormones synthesized from carotenoids and stimulate branching in plants and the growth of symbiotic arbuscular mycorrhizal fungi in the soil (58, 59). *SNX1* (SORTING NEXIN 1) plays a role in vesicular protein sorting and acts at the crossroads between the secretory and endocytic pathways. *Arabidopsis thaliana* SNX1 is involved in the auxin pathway by transporting PIN2 (60, 61).

GO enrichment analysis of the single-copy conserved syntenic BUSCO genes identified chloroplast-related genes as the most significantly enriched GO term (*SI Appendix*, Fig. S4), for example, *VTE1* (cluster 329), *CHLG* (cluster 23), *ObgC* (cluster 233), and *PNSL4* (cluster 256) genes (Fig. 6*B* and Dataset S4). VTE1 involved in the synthesis of both tocopherols and tocotrienols (vitamin E), which protect photosynthetic complexes from oxidative stress (62, 63). *CHLG* encodes a protein involved in one of the final steps in the biosynthesis of chlorophyll a (64). *ObgC* is the plant homolog of the bacterial *Obg* gene and encodes a GTP-binding protein involved in membrane biogenesis and protein synthesis in the chloroplast. ObgC is localized in chloroplast and is essential for early embryo development. Disruption in this locus results in embryonic lethality (65, 66). *PNSL4* encodes a subunit of the chloroplast NAD(P)H dehydrogenase (NDH) complex which mediates photosystem I (PSI) cyclic and chlororespiratory electron transport in higher plants (67). That chloroplast and photosynthesis-related genes are highly conserved across angiosperms highlights just how important this plant-specific organelle is for the success of plants, and suggests new possibilities to study links between plastid function and photosynthesis to conserved patterns of gene regulation (such as circadian regulation).

Previous work has shown how both gene positional conservation and dynamism can directly affect the evolution and development of individuals, species, and/or lineages. Phylogenetic profiling of genomic data has identified patterns of loss that correlate with phenotypic changes. For example, gene losses in bats were associated with shifts in diet (4) and gene losses in plants were associated with the loss of interactions with beneficial fungi (mycorrhizae) and/or bacteria (such as *Rhizobia*) (5, 6, 68). Nearly everyone appreciates and understands how rapid changes in genomic context of particular genes (such as by chromosomal breaks) can directly lead to many cancers (69–71). At the other extreme, the relative gene order and function of Hox genes is highly conserved across most animals and embryo development. There is also an increased appreciation of genomic changes that are unique to a species, such as for humans, that have affected our evolutionary trajectory (72).

Our analysis detected long-term conservation and lineage-specific changes in relative genomic context of genes across broad phylogenetic groups. How conservation and changes in synteny or fundamental differences in genome organization have contributed to the evolution of lineages could be a scientific frontier. For example, our results could be integrated with approaches examining evolutionary changes in the three-dimensional genomic environment, patterns of histone modifications throughout the nucleus, and transcriptional regulation (73, 74). We specifically highlighted rebel lineage-specific gene transpositions in mammals and conserved syntenic single-copy genes in angiosperms. Examples in this study are just the tip of the iceberg. Much remains to be explored. This study provides a foundation for future investigations of, for example, other phylogenetic groups, deeper evolutionary timescales, and to test if rebel genes do in fact have a cause.

Methods

Genome Resources. All reference genomes were downloaded from public repositories, including NCBI, Ensembl, CoGe, and Phytozome (Dataset S1). For each genome, we downloaded FASTA format files containing protein sequences of all predicted gene models and the genome annotation files (GFF/BED) containing the positions of all of the genes. We modified all peptide sequence files and genome annotation GFF/BED files with corresponding species abbreviation identifiers. An in-house script was used for batch downloading genomes and modifying gene names. We analyzed 87 mammalian genomes, presented according to the consensus species tree adopted from NCBI taxonomy (Fig. 2 and Dataset S1) which included 1 Prototheria (*O. 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 (*A. trichopoda*), 26 monocots (including 14 Poaceae), 80 eudicots [including 1 Proteales (*Nelumbo nucifera*), 23 superasterids (asterids and caryophyllales), and 56 rosids] (Fig. 2 and Dataset S1).

BUSCO completeness of each genome was performed using BUSCO v3.0 (40). Each genome containing all protein sequences was searched against plant (embryophyta_odb9, 1440 BUSCOs) or mammalian (mammalia_odb9, 4404 BUSCOs) reference databases.

Pairwise Comparison, Synteny Block Detection, and Network Construction. DIAMOND (75) was used to perform all inter- and intrapairwise all-vs.-all protein similarity searches (default parameters). In total, 7,569 and 11,449 whole genome comparisons (focused on protein-coding regions) were performed for 87 mammal genomes and 107 plant genomes. MCScanX (28) was used for pairwise synteny block detection, which is 3,828 times for mammals and 5,778 times for plants. We changed three main parameters: number of top homologous pairs for the input (-b: 5, 10, 15, and 20), number of max gene gaps (-m: 15, 25, and 35), and number of minimum matched syntenic anchors (-s: 3, 5, and 7), and performed microsynteny block detection under 20 different parameter settings, to check the impact to outputting synteny blocks. For each parameter, we also supplemented tandem duplicated genes that have been originally collapsed for the sake of microsynteny detection (28). This was performed by the script of "detect_collinear_tandem_arrays" of the MCScanX toolkit.

Each pairwise syntenic percentage was calculated using the number of syntenic pairs plus the number of collinear tandem genes relative to the number of all annotated genes. We merged syntenic gene pairs from all inter- and intraspecies synteny blocks into one two-columned tabular-format file, which can serve as an undirected synteny network/graph and be further analyzed or visualized in various tools [such as "igraph" (R package), Cytoscape, and Gephi, etc.]. In this synteny network, nodes are genes, edges stand for syntenic relationships between nodes, and edge lengths in this

study have no meaning (unweighted). Further details can be referred to in Github tutorial (https://github.com/zhaotao1987/SynNet-Pipeline).

We summarized pairwise syntenic percentages under different settings for mammalian genomes and angiosperm genomes (*SI Appendix*, Fig. S1 *A* and *B*, respectively). Also, we compared the total number of nodes against average clustering coefficient of the microsynteny network under each of the parameter settings (*SI Appendix*, Fig. S1 *C* and *D*).

Network Statistics. Network statistical analysis was carried out in the R environment (www.r-project.org), using the R package "igraph" (76). 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.

Network Clustering and Copy-Number Profiling of All Clusters. We used the Infomap method integrated in igraph to split the entire network, consisting of millions of nodes, into clusters (77). Clustering results were determined by topological edge connections; edges were unweighted and undirected. All microsynteny 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 (78), then hierarchically clustered by "ward.D," and visualized by "pheatmap." We illustrate all of the clusters of mammals and angiosperm, respectively, with cluster size over 2.

GO Functional Enrichment. GO analysis was performed for highly syntenic plant BUSCO genes. We regarded microsynteny clusters containing genes from 70+ of the 107 plant genomes in a single cluster as highly syntenic microsynteny clusters. Representative *Arabidopsis* genes from these clusters were used to identify enriched GO terms using agriGO (bioinfo.cau.edu.cn/agriGO/) (79).

ACKNOWLEDGMENTS. We thank Kitty Vijverberg, Klaas Bouwmeester, three anonymous reviewers, and the editor for constructive suggestions and ideas for improving the manuscript. T.Z. was supported by the China Scholarship Council (201306300016).

1. O'Brien SJ, et al. (1999) The promise of comparative genomics in mammals. *Science* 286:458–462, 479–481.
2. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285–4288.
3. Domazet-Lošo T, Tautz D (2010) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol* 8:66.
4. Sharma V, et al. (2018) A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat Commun* 9:1215.
5. Delaux P-M, et al. (2014) Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet* 10:e1004487.
6. van Velzen R, et al. (2018) Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proc Natl Acad Sci USA* 115:E4700–E4709.
7. Dewey CN (2011) Positional orthology: Putting genomic evolutionary relationships into context. *Brief Bioinform* 12:401–412.
8. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338.
9. Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14:360–366.
10. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11:204–220.
11. Murat F, Van de Peer Y, Salse J (2012) Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol Evol* 4: 917–928.
12. Gladyshev EA, Arkhipova IR (2007) Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci USA* 104: 9352–9357.
13. Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* 330:622–627.
14. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8:135–141.
15. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* 35:119–125.
16. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA (2016) On the relative abundance of autopolyploids and allopolyploids. *New Phytol* 210:391–398.
17. Jiao Y, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
18. Cui L, et al. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16:738–749.
19. Jiao Y, et al. (2012) A genome triplication associated with early diversification of the core eudicots. *Genome Biol* 13:R3.
20. Hokamp K, McLysaght A, Wolfe KH (2003) The 2R hypothesis and the human genome sequence. *Genome Evolution* (Springer, New York), pp 95–110.
21. Panopoulou G, Poustka AJ (2005) Timing and mechanism of ancient vertebrate genome duplications–The adventure of a hypothesis. *Trends Genet* 21:559–567.
22. Steinke D, Hoegg S, Brinkmann H, Meyer A (2006) Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol* 4: 16.
23. Zhao T, Schranz ME (2017) Network approaches for plant phylogenomic synteny analysis. *Curr Opin Plant Biol* 36:129–134.
24. Zhao T, et al. (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.
25. Barabási A-L, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.
26. Chasman D, Fotuhi Siahpirani A, Roy S (2016) Network-based approaches for analysis of complex biological systems. *Curr Opin Biotechnol* 39:157–166.
27. Carvalho DS, Schnable JC, Almeida AMR (2018) Integrating phylogenetic and network approaches to study gene family evolution: The case of the *AGAMOUS* family of floral genes. *Evol Bioinform Online* 14:1176934318764683.
28. Wang Y, et al. (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40:e49.
29. Lancichinetti A, Fortunato S (2009) Community detection algorithms: A comparative analysis. *Phys Rev E Stat Nonlin Soft Matter Phys* 80:056117.
30. Puxeddu MG, et al. (2017) Community detection: Comparison among clustering algorithms and application to EEG-based brain networks. *Conf Proc IEEE Eng Med Biol Soc* 2017:3965–3968.
31. Cifelli RL, Davis BM (2003) Paleontology. Marsupial origins. *Science* 302:1899–1900.
32. Bininda-Emonds OR, et al. (2007) The delayed rise of present-day mammals. *Nature* 446:507–512.
33. Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* 207:437–453.
34. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286: 509–512.
35. Krumlauf R (1994) Hox genes in vertebrate development. *Cell* 78:191–201.
36. Schena M, Davis RW (1992) HD-Zip proteins: Members of an Arabidopsis homeodomain protein superfamily. *Proc Natl Acad Sci USA* 89:3894–3898.
37. Lemons D, McGinnis W (2006) Genomic evolution of Hox gene clusters. *Science* 313: 1918–1922.
38. Ferrier DE, Holland PW (2001) Ancient origin of the Hox gene cluster. *Nat Rev Genet* 2:33–38.
39. Brooke NM, Garcia-Fernàndez J, Holland PW (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392:920–922.
40. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.
41. Wooster R, et al. (1995) Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* 378:789–792.
42. King M-C, Marks JH, Mandell JB; New York Breast Cancer Study Group (2003) Breast and ovarian cancer risks due to inherited mutations in *BRCA1* and *BRCA2*. *Science* 302: 643–646.
43. Howlett NG, et al. (2002) Biallelic inactivation of *BRCA2* in Fanconi anemia. *Science* 297:606–609.
44. McMahon SB, Van Buskirk HA, Dugan KA, Copeland TD, Cole MD (1998) The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell* 94:363–374.
45. Murr R, et al. (2006) Histone acetylation by Trrap-Tip60 modulates loading of repair proteins and repair of DNA double-strand breaks. *Nat Cell Biol* 8:91–99.
46. Zhang G, et al. (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* 339:456–460.
47. Lakin ND, Jackson SP (1999) Regulation of p53 in response to DNA damage. *Oncogene* 18:7644–7655.
48. Kanehira M, et al. (2007) Oncogenic role of MPHOSPH1, a cancer-testis antigen specific to human bladder cancer. *Cancer Res* 67:3276–3285.
49. Bond J, et al. (2005) A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nat Genet* 37:353–355, and erratum (2005) 37:555.
50. Gul A, et al. (2006) A novel deletion mutation in CENPJ gene in a Pakistani family with autosomal recessive primary microcephaly. *J Hum Genet* 51:760–764.
51. Kudo H, Dunbar R (2001) Neocortex size and social network size in primates. *Anim Behav* 62:711–722.
52. Byrne RW, Corp N (2004) Neocortex size predicts deception rate in primates. *Proc Biol Sci* 271:1693–1699.
53. Jeunemaitre X, et al. (1992) Molecular basis of human hypertension: Role of angiotensinogen. *Cell* 71:169–180.
54. Zhou A, et al. (2010) A redox switch in angiotensinogen modulates angiotensin release. *Nature* 468:108–111.
55. Yoshimoto R, Okawa K, Yoshida M, Ohno M, Kataoka N (2014) Identification of a novel component C2ORF3 in the lariat-intron complex: Lack of C2ORF3 interferes with pre-mRNA splicing via intron turnover pathway. *Genes Cells* 19:78–87.

56. Eicher JD, Gruen JR (2015) Language impairment and dyslexia genes influence language skills in children with autism spectrum disorders. *Autism Res* 8:229–234.
57. Anthoni H, et al. (2007) A locus on 2p12 containing the co-regulated *MRPL19* and *C2ORF3* genes is associated to dyslexia. *Hum Mol Genet* 16:667–677.
58. Booker J, et al. (2004) MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. *Curr Biol* 14:1232–1238.
59. Alder A, et al. (2012) The path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science* 335:1348–1351.
60. Jaillais Y, Fobis-Loisy I, Miège C, Rollin C, Gaude T (2006) AtSNX1 defines an endosome for auxin-Carrier trafficking in Arabidopsis. *Nature* 443:106–109.
61. Kleine-Vehn J, et al. (2008) Differential degradation of PIN2 auxin efflux Carrier by retromer-dependent vacuolar targeting. *Proc Natl Acad Sci USA* 105: 17812–17817.
62. Porfirova S, Bergmüller E, Tropf S, Lemke R, Dörmann P (2002) Isolation of an Arabidopsis mutant lacking vitamin E and identification of a cyclase essential for all tocopherol biosynthesis. *Proc Natl Acad Sci USA* 99:12495–12500.
63. Sattler SE, Cahoon EB, Coughlan SJ, DellaPenna D (2003) Characterization of tocopherol cyclases from higher plants and cyanobacteria. Evolutionary implications for tocopherol synthesis and function. *Plant Physiol* 132:2184–2195.
64. Gaubier P, Wu H-J, Laudié M, Delseny M, Grellet F (1995) A chlorophyll synthetase gene from *Arabidopsis thaliana*. *Mol Gen Genet* 249:58–64.
65. Bang WY, et al. (2012) Functional characterization of ObgC in ribosome biogenesis during chloroplast development. *Plant J* 71:122–134.
66. Garcia C, Khan NZ, Nannmark U, Aronsson H (2010) The chloroplast protein CPSAR1, dually localized in the stroma and the inner envelope membrane, is involved in thylakoid biogenesis. *Plant J* 63:73–85.
67. Peng L, Fukao Y, Fujiwara M, Takami T, Shikanai T (2009) Efficient operation of NAD(P)H dehydrogenase requires supercomplex formation with photosystem I via minor LHCI in Arabidopsis. *Plant Cell* 21:3623–3640.
68. Griesmann M, et al. (2018) Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361:eaat1743.
69. Stephens PJ, et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144:27–40.
70. Frampton GM, et al. (2013) Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* 31:1023–1031.
71. Loo LW, et al. (2004) Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res* 64:8541–8549.
72. O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM (2012) Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* 13:853–866.
73. Sotelo-Silveira M, Chávez Montes RA, Sotelo-Silveira JR, Marsch-Martínez N, de Folter S (2018) Entering the next dimension: Plant genomes in 3D. *Trends Plant Sci* 23:598–612.
74. Yu M, Ren B (2017) The three-dimensional organization of mammalian genomes. *Annu Rev Cell Dev Biol* 33:265–289.
75. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.
76. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *Int J Complex Syst* 1695:1–9.
77. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA* 105:1118–1123.
78. Dixon P (2003) VEGAN, a package of R functions for community ecology. *J Veg Sci* 14: 927–930.
79. Tian T, et al. (2017) agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res* 45:W122–W129.