Supplemental Material for:

Long Identical Multispecies Elements in Plant and Animal Genomes

Jeff Reneker[a], Eric Lyons[b,1] , Gavin C. Conant[c,d], J. Chris Pires[d,e], Michael Freeling[b], Chi-Ren Shyu[a,d], Dmitry Korkin[a,d*]

S1. Syntenic analysis of LIMEs in human, mouse and rat genomes

The CoGe comparative genomics platform was used to identify syntenic regions of each LIME in Table S1, and the results are summarized in Table S2. Synteny was determined by the presence of a similar pattern of homologous genes within the region (±100,000 bp) surrounding the element. In the absence of any nearby genes, homology was determined by the presence or absence of multiple high-scoring segment pairs HSPs. Therefore, these are subjective classifications and, in the absence of an obvious choice, the classification is '?'. If a LIME was within a gene, then the gene name was given, otherwise the element was intergenic. LIME 2 is a subsequence of the heterogeneous nuclear ribonucleoprotein hnRNP-A3 (39), which binds Pol transcripts and is involved in many RNA-related activities. In their paper, Makeyev et al. provided details for over 34 hnRNP-A genes and pseudogenes in human, mouse and other species, while demonstrating that in human, mouse and rat, only one hnRNP-A3 gene per genome had introns and showed the features of an active gene locus. Our analysis confirmed that LIME 2 was within a large region (250,000+ bp) of homology between human chromosome 2, mouse chromosome 2, and rat chromosome 3 (http://tinyurl.com/ycyfv9b). Makeyev *et al*. also demonstrated that the other copies in human and mouse are pseudogenes caused by retrotranspositions that had occurred over a long period of time as evidenced by the differing numbers of accumulated mutations in each copy. Our data demonstrated that LIME 2 was present in 8 copies of the gene/pseudogene in human, mouse and rat. Building on Makeyev *et al*. findings, it would also suggest that the copies on mouse chromosomes 4, 14, 16, and X as well as rat chromosome 13 were relatively recent retrotransposition events.

* Correspondence should be addressed to D.K. (korkin@korkinlab.org)
[a]Department of Computer Science, University of Missouri, Columbia, MO 65211, USA.
[b]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94704, USA.
[c]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA.
[d]Informatics Institute, University of Missouri, Columbia, MO 65211, USA.
[e]Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211, USA
[1]Present address: Bio5 Institute, iPlant Collaborative, University of Arizona, Tucson, AZ 85721, USA.

S2. Analysis of LIME 4 in human, mouse and rat

We next used CoGe to identify syntenic regions of each genome containing LIME 4, which is a subsequence of LIME 5. As shown in Table S1, LIME 4 is present once in human and mouse and twice in rat. The results are shown in Figure S11 and can also be viewed interactively at http://tinyurl.com/ybbzuub. From top to bottom, the four regions shown are from human chromosome 6, mouse chromosome 13, rat chromosome 4 and rat chromosome 17. The region containing LIME 4 has been circled in black on each chromosome and the lines drawn connect high-scoring segment pairs HSPs (rectangles) from the blastz alignments in the immediate vicinity of the element. There are multiple green HSPs as well as a similar distribution of homologous genes (thin, grey arrows parallel to dotted lines) between mouse chromosome 13 and rat chromosome 17, indicating that these regions share a large block (110,000+ bp) of synteny. A lower level of homology exists between human chromosome 6 and both mouse chromosome 13 and rat chromosome 17, as indicated by the fewer numbers of orange and brown HSPs, respectively, but the genes shown are still syntenic in terms of annotation. There is no significant homology between rat chromosome 4 and any other sequence, except for the 3,067 bp region containing the element. However, every chromosome (except rat chromosome 4) contains homologous copies of a gene, indicated by blue arrows, which are flanked by HSPs having 89+% cross-species identity. This gene is named C6orf62, BC005537 and LOC498750 in human, mouse and rat, respectively. One possible explanation for the presence of the 3,067 bp region of rat chromosome 4 is that a retrotransposition has occurred. In other words, an RNA splice variant of LOC498750 on chromosome 17 could have been reverse transcribed and inserted back into chromosome 4 at this site. The AceView (Thierry-Mieg and Thierry-Mieg, 2006) description of LOC498750 lists an mRNA variant with a premessenger length of 11,957 bases. LIME 4 maps to 5' junction between the non-coding region and the first exon with 97 bases in the non-coding region and 104 bases in the first exon that code for the first 34 amino acids of the protein: MGDPNSRKKQALNRLRAQLRKKKESLADQFDFKM. The 3,067 bp region of rat chromosome 4 contains two adjacent HSPs aligning to rat chromosome 17. The first HSP is fully contained over the 5' non-coding region and the first exon of the premessenger while the second HSP is fully contained over the last few bases of the last intron, the last exon and 2/3 of the 3' untranslated region. Therefore, the entire 3,067 bp region is transcribed into the premessenger

RNA which gives further evidence for retrotransposition. The AceView annotation for C6orf62 in human indicates that the gene is expressed in brain, lung, liver, testis, eye, uterus, placenta and 109 other tissues. There is no known phenotype and the gene's in vivo function is yet unknown. However, there are currently 54 proteins expressed in 18 different species that have protein-to-protein Blast scores >1,000 to C6orf62, which implies perfect or near-perfect matches.

S3. Human complex LIME analysis to mitochondrial and rDNA sequence

Every human complex LIME was used in BLAST searches of the human mitochondrial genome and only two matches were found, both from human-macaque. CoGe was employed to compare the location of LIME ID 1089409 in human chromosome 9 and macaque chromosome 15 as well as the two mitochondrial genomes (http://genomevolution.org/r/42s8). The LIME was found to be in a large region of sequence identity between the two genomic sequences, suggesting mitochondrial insertion. Similarily, LIME ID 514066 was also found in a large region of sequence identity between human chromosome 2, macaque chromosome 12 and the two mitochondrial genomes (http://genomevolution.org/r/42s9). In addition, every human complex LIME was used in BLAST searches of the human ribosomal DNA complete repeating unit (GenBank: U13369.1), and 250 distinct LIMEs were found to contain matching sequences. All but 9 of these mapped to human-macaque exclusively, while six mapped to human-mouse exclusively and one mapped only to human-chicken. There were also two mapping to human-mouse-macaque while dog and rat LIMEs did not contain any matches.

S4. Annotation of LIMEs associated with the supersequences in *Arabidopsis*

The “supersequences” of complex LIMEs shared by at least three species and joined together, if their locations on a chromosome overlapped, were used as the query sequences for the Nucleotide BLAST sequence similarity search. The annotation of the LIMEs associated with the supersequences in *Arabidopsis* included non-coding as well as coding elements that were not associated with ribosomal genes. The only exception was a 140 bp sequence (LIME ID 834 and 1st 140 bp of LIME ID 835) of a single LIME presented in *Arabidopsis*, rice and sorghum, which was found to have mitochondrial origin in other species. The genomic region containing the latter element appeared to evolve differently across different species and, similarly to the ribosomal LIMEs, might be important in translation. For instance, in *Arabidopsis* and rice genomic DNA

(as well as rice mitochondria), the element was located 500 – 600 bp downstream of a tRNA-Tyr gene, while in sorghum mitochondria it was 274 bp downstream of tRNA-Met gene.

S5. BLAST search and TAIR annotation of complex LIMEs that are common to at least three plant species

For some plant LIMEs, the absence of synteny could be the result of their recurrent origins, which were essentially 'created in place' rather than being universally inherited from the common ancestor. For other LIMEs, the non-syntenic nature could be explained by their origin through the transfer of the genetic material from an organellar to the nuclear genome or by their being parts of mobile elements that have not been annotated as such. To address the question of evolutionary origins of the remaining non-syntenic, complex, non-organellar plant LIMEs, we performed a nucleotide BLAST search against the non-redundant database at NCBI (40) using supersequences of complex LIMEs as queries. These sequences derived from LIMEs shared by at least three species were merged if their locations on a chromosome overlapped. The annotation of the returned hits for each element was consistently one of several ribosomal subunits (16S, 18S, 23S, 26S or 45S) across multiple organisms, except for one 140 bp element (LIME ID 834/5) present in *Arabidopsis*, rice and sorghum, which was mitochondrial in origin. Thus, the lack of synteny in the case of LIMEs associated with the ribosomal subunits belies a true functional conservation, most likely as a result of the increased possibility for rearrangements in the highly duplicated ribosomal genes. We further suggest that one reason these ribosomal genes possess LIMEs is that the process of gene conversion among the multiple copies has resulted in long-term sequence homogenisation among these genes throughout the history of the angiosperms.

CoGe analysis of the last element (LIME ID 834/5) in *Arabidopsis* showed that it was within a large region of identity (99.66% over 52,089 bp) between chromosome 2 and the mitochondrial genome (Fig. S8A). The element was 498 bp downstream of AT2G07792 (tRNA-Tyr anticodon: GTA) on chromosome 2 and within the fourth intron of ArthMp044 (NADH dehydrogenase subunit 1) in the mitochondrial genome. It was also present twice in rice mitochondria. One copy was 585 bp downstream of OrsajM_t19 (tRNA-Tyr) and 53 bp downstream of OrsajM_p52 (NADH dehydrogenase subunit 2) while the other copy was 405,541 bp upstream of OrsajM_p52 (its nearest gene). On rice chromosome 1, the element was 56 bp downstream of Os01g0790900

(H+-transporting two-sector ATPase, delta/epsilon subunit family protein) and 585 bp downstream of Os01g0790800 (tRNA-Tyr). The largest High-scoring Segment Pair (HSP) between rice chromosome 1 and rice mitochondria (which contains the LIME) was 6,256 bp long and has 100% identity. In sorghum chromosome 9, the element was within the unannotated Sb09g014016 gene. In sorghum mitochondria, it was 274 bp downstream of SobioMt09 (tRNA-Met) on one strand while, on the other strand, 88 bp of it are in the 5th and last exon of SobioMp21 (NADH dehydrogenase subunit 2) and 52 bp are downstream of the gene. Sequence analysis suggested that the regions containing this element had evolved differently in *Arabidopsis*, rice and sorghum. For instance, in *Arabidopsis*, the element was within a large identical region between genomic and mitochondrial DNA, while in sorghum there was little similarity of the corresponding genomic and mitochondrial regions; the similarity of rice genomic and mitochondrial regions was in between those two extremes.

S6. Mitochondrial LIMEs shared by *Arabidopsis* and other plant genomes.

In addition to their nuclear genomes, most plants also pass two other genomes to their offspring, the mitochondrial and plastid genomes. Occasionally, the latter genomes can become fully or partially incorporated into the nuclear genome; a copy of almost the entire *Arabidopsis* mitochondrial genome (a ~280-kbp segment) is found on chromosome 2, albeit with a few rearrangements. This region contains 18 LIMEs (LIME IDs 830-847), 13 of which are exonic or partly exonic, with TAIR annotations that include cytochrome *c*, tRNA-Met, NADH dehydrogenase and ribosomal protein S4. The elements were identified through exact matches to grape (7), soybean (4), rice (6) and sorghum (1). Notably, all nine complex, non-artefactual exonic LIMEs found in our analysis that were shared by *Arabidopsis* and at least one other species were identified as mitochondrial insertions (Table S4). However, the corresponding genomic cross-species regions showed very limited homology at best and sometimes only matched at the elements themselves. The mitochondrial genomes of each species except soybean (for which no sequence has been published) were searched, and exact matches to all but three elements were found. The latter three elements had nearly exact matches, differing by only one or two nucleotides in *Arabidopsis*. Using the mitochondrial matches as anchoring points in a cross-species comparison frequently resulted in relatively short (50–1,000 bp) high-scoring segment pairs that had been extensively rearranged between mitochondrial genomes. Therefore, the

surrounding mitochondrial and nuclear sequences seem to be rearranged and/or diverged, while still retaining these few elements throughout evolution.

S7. Functional repeats in plants.

Each genome contains a fraction of 1,699 possible distinct tandem repeat motifs of 2–7 bp; the sizes of these sets varied from 228 motifs in *Arabidopsis* to 680 motifs in soybean (Table 2). Only a small subset of each motif set contributed to the repetitive LIMEs; however, these subsets were remarkably similar among the genomes of all six species. A simple statistical model was considered (see Tables 2, 3 and *Material and Methods* section) to determine that the number of distinct repetitive element motifs shared by six genomes was unlikely to be obtained by a chance: the probability of sharing a common repertoire of 12 motifs by six randomly selected sets of motifs with pairwise overlaps of the corresponding sizes was estimated to be $\sim 3 \times 10^{-69}$ (see Tables 2, 3 and *Material and Methods* section). The fact that repeats containing certain motifs shared extreme conservation across multiple genomes can be attributed to the common functions carried by the motifs. Indeed, the TTTAGGG motif is the telomeric repeat sequence in *Arabidopsis* (41). Telomere sequences containing this motif are found in a wide variety of plant species, although the pattern is not universal among angiosperms, as Asparagales species appear to lack the repeat (42-44). Our analysis found telomeric repetitive elements containing motif TTTAGGG in all six genomes (Figs. 1b, S1-S5). In addition, centromeric $(TTTAGGG)_n$ repeat sequences were found in *Arabidopsis* (chromosome 3) and possibly in grape (chromosome 9). Another example of functional repeats are the GAGA repeats in soybean (45), which have a known binding protein, GAGA-binding protein (GBP), and are thought to regulate gene function(s). Similarly, the GAGA repeats in *Drosophila melanogaster* have been shown to be the binding sites for protein complexes and have a regulatory role (46). The large number of potential target sequences for GAGA-binding proteins in plant genomes suggests that those proteins may affect the expression of a variety of genes involved in different plant processes. It is possible that other repetitive LIMEs, such as ATACAT and perhaps ATTAT, are manufactured in a similar fashion by a yet-to-be-discovered mechanism. For instance, ATTAT repeats are fully conserved in the noncoding *trnL* intron of the chloroplast genome in the orchid species of the *Desa* genus (47).

S8. Determining co-localized plant LIMEs

The spatial distribution of complex LIMEs was analyzed by an agglomerative clustering algorithm, in which the maximum distance between any two adjacent/overlapping elements on a chromosome was not allowed to exceed a certain threshold. The threshold was varied in 100 bp increments from 100 bp (1,769 clusters) to 1,000,000 bp (380 clusters) and all 85 chromosomes from the 6 plant genomes were considered (Fig. S10). A region of LIMEs was subsequently defined by setting the threshold at 60,000 base pairs, resulting in 627 clusters); further increase of the threshold resulted in a significantly slower decrease of the number of clusters, compared to the initial increase of the threshold from 100 to 60,000 bp.

S9. Network of clustered plant LIMEs within a single genome and across multiple genomes

We constructed two networks of complex and repetitive LIMEs that define the relationship between LIME clusters through identical copies of LIMEs shared between clusters within the same and across different species (Fig. S9). The nodes in the network of complex LIME locations corresponded to individual LIMEs, each node was colored by the corresponding species. There were two types of intra-species edges: intra-cluster (colored the same as the nodes of that species) and inter-cluster (colored dark green). Two clusters could be connected with at most one edge through an arbitrary selected pairs of representative LIMEs, one from each cluster. Edges colored red corresponded to the presence of at least one shared LIME between a pair of clusters across species. The network of complex LIMEs consisted of 8,788 nodes.

An analysis of the network topology of complex sequences revealed that the network appeared to be scale-free, as the node degrees followed a power law distribution. Defining network hubs as nodes having 100 or more edges leaves 11 hubs in the network with an average of 619.6 edges per hub. A hub of the highest degree is located in soybean chromosome 13 and contains 5,087 edges. Overall, there were 8,157 intra-cluster edges, 2,147 inter-cluster edges and 1,070 inter-species edges making an average of ~1.3 edges per node. The nodes in the network of repetitive elements corresponded to individual elements that had been merged when overlapping on a chromosome since 99.05% of them overlapped. This drops the total number of nodes from 3,628,645 to just 4,343. The network was not scale-free as the node degrees followed a

polynomial distribution. Overall, there were 747 intra-cluster edges, 1,877,343 inter-cluster edges and 2,620,855 inter-species edges making an average of ~1,036 edges per node. Both networks were constructed using the Large Graph Layout LGL software (48).

When analysing connectivity within superclusters, we found that LIMEs that belonged to the same cluster in one species were dispersed into multiple clusters in another species. For instance, in a supercluster that included a single complex LIME from *Arabidopsis* (LIME ID 1516), the average number of inter-species connections for one cluster was ~3.4 (red edges, Fig. 4). Similarly, the intra-species copies of a multi-copy LIME often did not co-localise in the same cluster (dark green edges in Figs. 4 and S10).

S10. Examples of multiple-copy plant LIMEs

We next performed the CoGe analysis of a 121 bp LIME (LIME ID 23928) that occurred three times in rice and twice in sorghum. In rice, one copy was located within gene Os05g0293600 and another was less than 500 bp downstream of Os10g0355000; both of these genes were labeled as "DNA-directed RNA polymerase beta chain". The third copy was located in an intergenic region. In sorghum, one copy of the LIME was located within gene Sb04g009441 and the other was within gene Sb03g017630 (the data can be reproduced in CoGe: http://tinyurl.com/lgcb82); both of the genes have been annotated "similar to DNA-directed RNA polymerase subunit beta".

S11. A model to test paleopolyploidy hypothesis for multi-copy plant LIMEs

To find whether the appearance of multiple-copy LIMEs could be explained exclusively by the paleopolyploidy events, we compared multiple-copy LIMEs in rice and sorghum genomes. These genomes were selected, since they were found to share more multiple-copy elements (240) than other pairs of genomes. In addition, the genomes were thought to undergo the same ancient duplication events (49, 50). Rice has 965 copies of multi-copy LIMEs shared between the two genomes, while sorghum has 554, giving a ratio of 1.74 to 1. Since LIMEs often have a close spatial relationship and would subsequently be expected to evolve as a unit, an alternative scenario was considered where LIMEs were grouped into clusters based on the distance between them ($\leq$ 60,000 bp, see section S8 for more details), and the link between a LIME cluster in the rice genome and a cluster in the sorghum genome was established if both cluster had at least one LIME in common. We found that the rice genome has 85 LIME clusters that link to 19 clusters in

sorghum, resulting in a 4.47:1 cluster ratio. If multiple copies of LIMEs in both genomes were due to one or several paleopolyploidy event, one would expect to see similar numbers of copies. Based on the results of both scenarios, one can conclude that paleopolyploidy seems not to be the exclusive mechanism behind the multiple-copy LIME phenomenon.

S12. Copy numbers for identified plant LIMEs.

With the ability of our new method to match a fragment of one genome sequence to multiple positions of another genome, we discovered that both repetitive and complex LIMEs often occurred in multiple locations across each genome, although, curiously, we did not find complex LIMEs occurring in multiple copies in cottonwood and grape. The four LIMEs shared by all six genomes were no exception: the longest LIME occurred in 89 different locations of soybean chromosome 13, while the second shortest element was in two locations of sorghum chromosome 5, and the shortest element occurred twice on sorghum chromosome 1 (Table 1). On average, a plant genome from our set contained 4,938 LIMEs occurring in multiple copies and 6,047 single-copy LIMEs (Table 4). The genomes varied greatly in the number of multiple-copy LIMEs, varying from a single instance in cottonwood to 212 cases in rice among the complex elements and from 692 (in *Arabidopsis*) to 11,053 (in soybean) for repetitive elements. Unsurprisingly, the number of single-copy LIMEs in a genome always exceeded the number of multiple-copy elements. The number of copies of a multiple-copy LIME also varied significantly: 2 to 55 copies of each complex and 18 to 160 of each repetitive LIME, with an unusually high number of multiple copies located on chromosome 13 of the soybean genome (see Section S10 for specific examples). These observations resemble the patterns of gene family size seen in complete genomes (51) and may have similar origins. Some elements were present multiple times in multiple plant genomes (see Section S10 for specific examples). One possible source of multiple-copy LIMEs is whole-genome duplication events, a recurrent feature of plant genome evolution. For example, in the lineage leading to Arabidopsis, there are a paleopolyploidy shared among all eurosids (52, 53) and two subsequent sequential tetraploidies (54). Likewise, the sequenced monocot genomes, all grasses, are reported to share three tetraploidy events (55, 56). By (potentially) duplicating all of the LIMEs in a genome, such events could be responsible for the extant multiple-copy LIMEs. However, although the rice and sorghum genomes appear to have experienced the same genome-duplication events (49, 50), sorghum had twice as many LIME

copies as did rice (see Section S11), suggesting that paleopolyploidy is unlikely to explain different occurrences of multiple-copy UCE in extant genomes.

Supplementary Tables

Supplementary Table 1| Previously unreported human, mouse and rat elements of extreme conservation found in 2004 builds.

LIME Number	H_sapiens Build 34	M_musculus Build 30	R_norvegicus Build 3.1
1	Chromosome 1 38,013,280 - 38,013,484	Chromosome 4 119,243,379 - 119,243,583	Chromosome 5 143,910,554 - 143,910,758
2	Chromosome 2 178,286,889 - 178,287,109	Chromosome 2 73,547,266 - 73,547,486 Chromosome 4 8,772,976 – 8,773,196 Chromosome 14 113,443,776 – 113,443,996 Chromosome 16 95,324,396 – 95,324,616 Chromosome X 61,403,393 – 61,403,613	Chromosome 3 58,253,488 - 58,253,708 Chromosome 13 68,887,235 – 68,887,455
3	Chromosome 2 208,142,426 - 208,142,626	Chromosome 1 61,514,263 - 61,514,463	Chromosome 9 62,833,729 - 62,833,929
4	Chromosome 6 24,826,771 – 24,826,972	Chromosome 13 20,999,097 – 20,999,298	Chromosome 4 102,649,865 – 102,650,066 Chromosome 17 47,330,345 – 47,330,546
5	Chromosome 6 24,826,771 - 24,826,973	Chromosome 13 20,999,096 - 20,999,298	Chromosome 17 47,330,345 - 47,330,547
6	Chromosome 8 66,201,165 - 66,201,409	Chromosome 3 15,377,956 - 15,378,200	Chromosome 2 103,702,515 - 103,702,759
7	Chromosome 11 16,171,435 - 16,171,635	Chromosome 7 98,432,093 - 98,432,293	Chromosome 1 173,877,668 - 173,877,868
8	Chromosome 14 48,043,388 - 48,043,638	Chromosome 12 60,391,309 - 60,391,559	Chromosome 6 91,120,792 - 91,121,042 91,326,559 - 91,326,809
9	Chromosome 14 48,043,398 – 48,043,627	Chromosome 12 60,391,320 – 60,391,549	Chromosome 6 91,120,802 – 91,121,031 91,326,570 – 91,326,799 Chromosome 12 22,723,039 – 22,723,268
10	Chromosome 14 48,043,404 – 48,043,609 48,319,331 – 48,319,536	Chromosome 12 60,391,338 – 60,391,543	Chromosome 6 91,120,808 – 91,121,013 91,326,588 – 91,326,793 Chromosome 12 22,723,057 – 22,723,262
11	Chromosome X 24,370,511 - 24,370,711	Chromosome X 74,525,227 - 74,525,427	Chromosome X 80,729,034 - 80,729,234
12	Chromosome X 121,297,012 - 121,297,216	Chromosome X 24,702,346 - 24,702,550	Chromosome X 3,482,862 - 3,483,066

Supplementary Table 2| Syntenic analysis of previously unreported human, mouse and rat elements

LIME Number	Synteny / Homology	Gene Name / Intragenic
1	homology	intergenic
2	?	HNRPA3
3	homology	intergenic
4,5	synteny (Fig. S11)	C6orf62
6	homology	intergenic
7	synteny	SOX6
8, 9, 10	?	intergenic
11	synteny	ARX
12	synteny	GRIA3

Supplementary Table 3| LIME repeated motifs in animals and plants. The twelve plant motifs are colored red.

Species/Motif	Cf	Gg	Hs	Mouse (Mm)	Macaque (Mm)	Rn	Total	At	Gm	Os	Pt	Sb	Vv	Total
GGGGGCTGCAAGGGAGGCTGTGGCTCCTGT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGCTTGCAGCAGCTGGACTGACAGCAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGCTTACAGCAGCTGGACTGACAGCAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGATGCTGGGCAGGGATGCTGGACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGGACAGAGACCCAGAGAGAGA	0	0	1	0	1	0	**2**	0	0	0	0	0	0	**0**
GGGGTGGCAGGGCTCAGGAGCCTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGTTGGAGGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGTTGGAGTG	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGTTGGAGTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGTCTTGTTGGAGTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGTTAGAGGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGTTGAAGGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGTGGCCTTGCTGGAGTG	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGGAGGAGGAAGAGAA	0	0	1	1	0	0	**2**	0	0	0	0	0	0	**0**
GAAAGAAAGAAAGAAAG	0	1	1	0	0	0	**2**	0	0	0	0	0	0	**0**
GATAGATAGATAGAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGGCAGAGGCAGA	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAATATATATATAT	0	0	1	0	1	0	**2**	0	0	0	0	0	0	**0**
GGATGGATGGATGAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCAGAGGCAGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAAACAGAGAGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCACACACACACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAGGAAAGAAA	0	1	0	0	1	0	**2**	0	0	0	0	0	0	**0**
GCTGTGTGTGCTGT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAGAAAGAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GAAAGAAAGAAAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**
GGGAAGGGAAAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGCGTGCGCGTGC	0	0	1	0	1	0	**2**	0	0	0	0	0	0	**0**
GAAAAAAAGAAAA	0	1	0	0	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAGGAAAAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**
GGGAAAAGGGAA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GGGAGAAGGGAA	1	0	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGAAAGGAGAAA	1	0	0	1	1	0	**3**	0	0	0	0	0	0	**0**
GATAGATACATA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GGGGAAAGGAAA	1	0	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGGGGAAGGGAA	1	0	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGGAAGGGAAGA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GGAAAGAGGAAA	0	0	0	1	1	0	**2**	0	0	0	0	0	0	**0**
GGGAAAGGAAA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GGGAAAGGGAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGAAGGAGGAA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAGGAGAA	0	0	0	1	1	0	**2**	0	0	0	0	0	0	**0**
GAAAGAGAGA	0	1	0	1	1	0	**3**	0	0	0	0	0	0	**0**
GGAGAAGAGA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GAGAGAGACA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GGGAAAGGAA	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GGAGGAGGCA	0	0	0	1	1	0	**2**	0	0	0	0	0	0	**0**
GGAGAGAAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGGGCAGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAAGGCAA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGCAGAGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGCGAGGC	0	0	0	1	1	0	**2**	0	0	0	0	0	0	**0**

GGGCCTTGTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGGTGTGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAGAAGA	0	1	1	1	1	1	**5**	0	0	0	0	0	0	**0**
GGAGGAGAA	1	0	1	1	1	1	**5**	0	0	0	0	0	0	**0**
GGAAGGAAA	1	1	0	1	1	1	**5**	0	0	0	0	0	0	**0**
GAAAGAAAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**
GGGAAGGAA	0	1	0	1	1	0	**3**	0	0	0	0	0	0	**0**
GGGAGGAGA	1	0	0	1	1	0	**3**	0	0	0	0	0	0	**0**
GATGTATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAAGAA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGGAGAGGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAAAGAAAG	1	1	0	1	1	1	**5**	0	0	0	0	0	0	**0**
GGAGAGGCA	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAAGGAGA	0	1	1	1	0	0	**3**	0	0	0	0	0	0	**0**
GATGATGTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAAAGA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGAAGAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGCAGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GTGTATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGAGAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GTATATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAGGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GATATATA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGCAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAGAGA	1	1	1	1	0	1	**5**	0	0	0	0	0	0	**0**
GGGAAGGA	1	1	0	1	1	1	**5**	0	0	0	0	0	0	**0**
GGAAGAGA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GGGAGGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAGGCA	1	1	0	1	1	1	**5**	0	0	0	0	0	0	**0**
GGAAAAGA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GAGAGACA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GGGAAGAA	1	0	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGGAGAAA	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GTATCTAT	0	0	1	1	1	1	**4**	0	0	0	0	0	0	**0**
GATAGACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGAAGA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGGACCCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GCAAGAAA	1	0	1	1	1	1	**5**	0	0	0	0	0	0	**0**
GGATTGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTTGGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGTGGAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GTATTTAT	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGGATGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCGGAGT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GCTGTGCT	0	1	0	0	0	1	**2**	0	0	0	0	0	0	**0**
GTAAATAA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTCTCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGCTGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGATGAAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGACATA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGAGA	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GAAAGAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**
GGAGGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
ATTATAT	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GAAAAAA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAAGAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**

GGCAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GATATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGCAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGAGAA	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GGAGAAA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GGGAGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAGAG	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GAGAAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GTATATA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTTGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGAGA	0	0	1	1	0	0	**2**	0	0	0	0	0	0	**0**
GAATATA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAGACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAATACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGCCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAGAA	0	0	1	1	1	1	**4**	0	0	0	0	0	0	**0**
GTGTGCT	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GGGTCCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGAGCTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGCCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGCCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTGAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGAAA	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGGCTCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
TTTAGGG	0	0	0	0	0	0	**0**	1	1	1	1	1	1	**6**
GGCAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGGAGA	1	1	1	1	0	1	**5**	0	0	0	0	0	0	**0**
ATACAT	0	1	1	1	0	1	**4**	0	0	1	0	1	0	**2**
GAGAAA	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GATATA	1	0	1	1	0	1	**4**	0	0	0	0	0	0	**0**
GGAAGA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GAAAAA	1	1	0	1	1	0	**4**	0	0	0	0	0	0	**0**
GGGTTA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGGGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GTGTCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAAAA	1	1	0	1	0	0	**3**	0	0	0	0	0	0	**0**
GTGTAT	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAGCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAAGCA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
TTTATA	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGGGAA	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GGGAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGCACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GATACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GTCTAT	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAGGC	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAACAA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGATA	1	1	1	1	0	1	**5**	0	0	0	0	0	0	**0**

GGGTCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTTTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTGT	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGTATA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GATGTA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGTGCA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GGGCTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTCC	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGTCA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GTGGAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGATA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGCTTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GCTGGA	1	0	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GTTGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GAGTTA	0	0	1	1	0	0	**2**	0	0	0	0	0	0	**0**
GGTAGA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGATAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGAACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGAGA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GAGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GAAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
TTTTA	1	1	1	1	0	1	**5**	0	0	0	0	0	0	**0**
GGCAA	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GGATA	0	1	0	1	1	1	**4**	0	0	0	0	0	0	**0**
GGGGA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GAATA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GTTTT	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
GGCTC	0	1	1	1	0	1	**4**	0	0	0	0	0	0	**0**
GGGCA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GTGCT	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GCTCT	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GCTTT	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GTTCT	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
GTCTT	0	1	0	1	1	1	**4**	0	0	0	0	0	0	**0**
GTATT	0	0	1	1	1	1	**4**	0	0	0	0	0	0	**0**
GTATA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GTAGA	1	1	1	1	0	1	**5**	0	0	0	0	0	0	**0**
GTACA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGTA	0	1	0	0	0	1	**2**	0	0	0	0	0	0	**0**
GGGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGAT	0	1	0	1	1	1	**4**	0	0	0	0	0	0	**0**
GGACA	0	1	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GGAAT	1	1	0	1	0	1	**4**	0	0	0	0	0	0	**0**
GCTAT	0	0	1	1	1	1	**4**	0	0	0	0	0	0	**0**
GAGAC	0	1	0	1	0	0	**2**	0	0	0	0	0	0	**0**
ATTAT	0	0	0	0	0	0	**0**	0	0	1	0	1	0	**2**
GAAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGAA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGAT	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GGCA	0	1	1	1	1	1	**5**	0	0	0	0	0	0	**0**
GAAT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
ATGT	0	0	1	1	1	1	**4**	1	0	1	0	1	0	**3**

ATCT	1	0	1	1	1	1	**5**	0	0	1	0	1	0	**2**
GTGC	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGGA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
GTCT	0	0	1	1	1	1	**4**	0	0	0	0	0	0	**0**
GCAA	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
GTAA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GGTA	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GGCT	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
AAAT	1	0	0	1	0	1	**3**	0	0	0	0	0	0	**0**
CTT	1	1	1	1	1	1	**6**	1	1	1	0	1	0	**4**
GGA	1	1	1	1	1	1	**6**	0	0	0	0	0	0	**0**
CAT	0	1	1	1	1	1	**5**	1	1	0	0	1	0	**3**
GTT	0	0	0	1	0	1	**2**	0	1	1	1	1	1	**5**
GTA	0	0	0	1	0	1	**2**	0	0	0	0	0	0	**0**
GCT	0	0	0	1	1	1	**3**	0	0	0	0	0	0	**0**
GAC	0	0	1	1	0	1	**3**	0	0	0	0	0	0	**0**
ATT	0	0	0	0	0	0	**0**	1	1	1	0	1	1	**5**
GA	1	1	1	1	1	1	**6**	1	1	1	1	1	1	**6**
GT	0	1	1	1	1	1	**5**	0	1	1	1	1	1	**5**
AT	1	0	1	1	1	1	**5**	0	1	1	1	1	1	**5**
Sum Within Species	**69**	**91**	**59**	**233**	**69**	**194**		**6**	**8**	**11**	**5**	**12**	**6**	

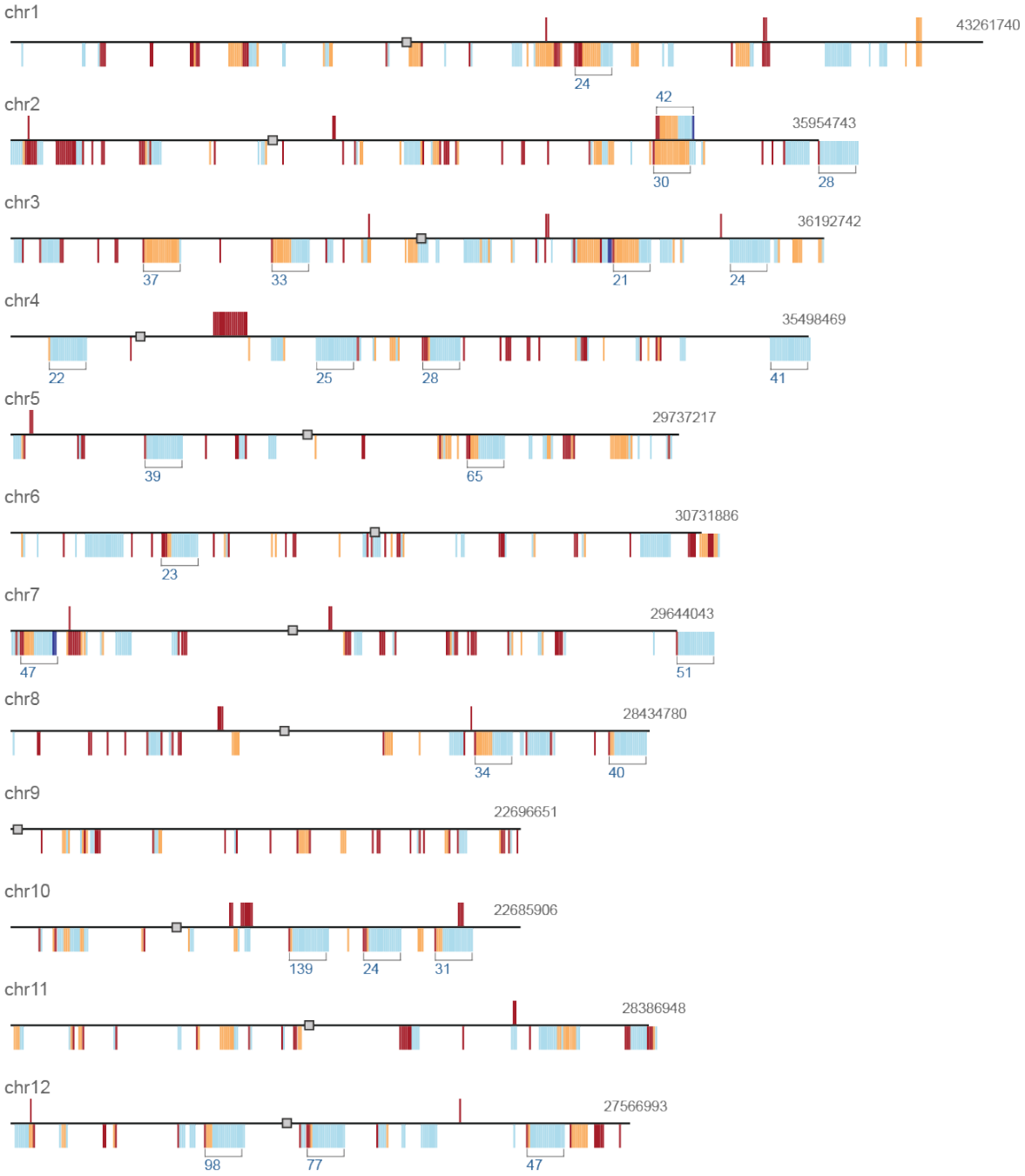
Supplementary Table 4| TAIR functional annotation of complex, non-artifactual exonic LIMEs in *Arabidopsis* that are neither ribosomal nor transposon-associated.

ID	Gene ID	Strand	LIME position		Functional feature		Functional annotation
			Begin	End	Begin	End	
830	AT2G07771.2	R	3239856	3239961	3239690	3240460	cytochrome c biogenesis protein-related
837	AT2G07785.1	F	3348235	3348387	3348111	3350015	NADH-ubiquinone oxidoreductase, putative
838	AT2G07698.1	F	3363285	3363397	3362519	3364124	Identical to ATP synthase subunit alpha, mitochondrial
839	AT2G07698.1	F	3363437	3363539	3362519	3364124	Identical to ATP synthase subunit alpha, mitochondrial
840	AT2G07709.1	F	3390562	3390666	3388451	3394513	similar to NADH dehydrogenase
841	AT2G07709.1	F	3391038	3391210	3388451	3394513	similar to NADH dehydrogenase
842	AT2G07709.1	F	3392211	3392343	3388451	3394513	similar to NADH dehydrogenase
845	AT2G07711.1	R	3398893	3399014	3397155	3399431	similar to NADH dehydrogenase subunit 5
846	AT2G07786.1	R	3399690	3399794	3399654	3399909	similar to NADH dehydrogenase (ubiquinone)

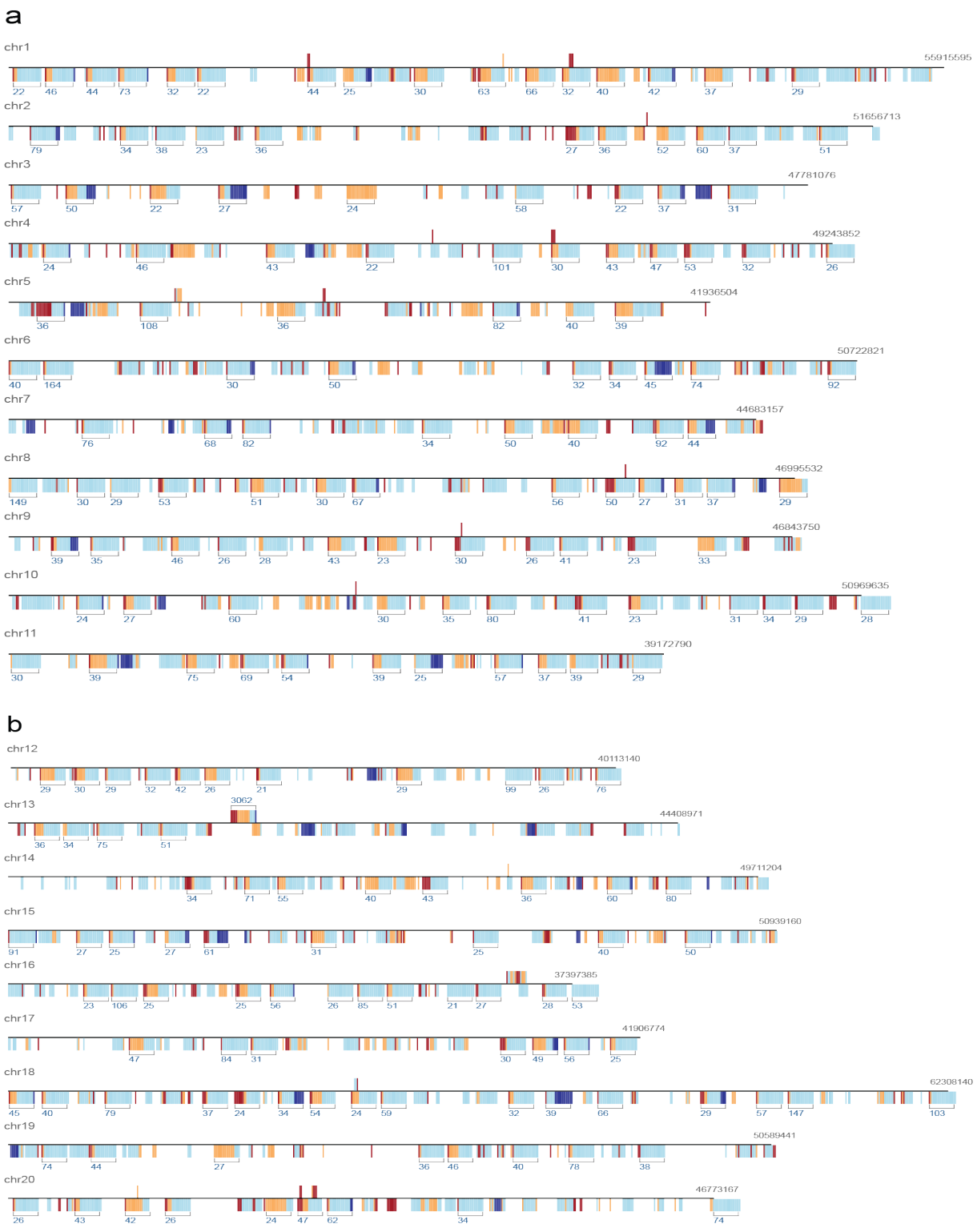
Supplementary Table 5| Structural taxonomy of annotated plant and animals LIMEs. Shown are the numbers for *Arabidopsis* LIMEs in plants, and for the human, mouse, rat and chicken LIMEs in animals. The following abbreviations are used: Rep for repetitive, Com for complex, Te for telomeric, H for heterochromatic, Tr for transposon, R for rRna, M for mitochondrial, and O for other.

Plant (*Arabidopsis*)							Animal (human, mouse, rat, chicken)					
Syntenic	Non-syntenic						Syntenic		Non-syntenic			
Rep	Rep		Com				Rep	Com	Rep	Com		
Te	H	Tr	R	Tr	M	O	Te	O	H	R	Tr	M
1,597	1,180	1	96	67	2	49	752	1,553,351	268,955	15,612	768	2,849

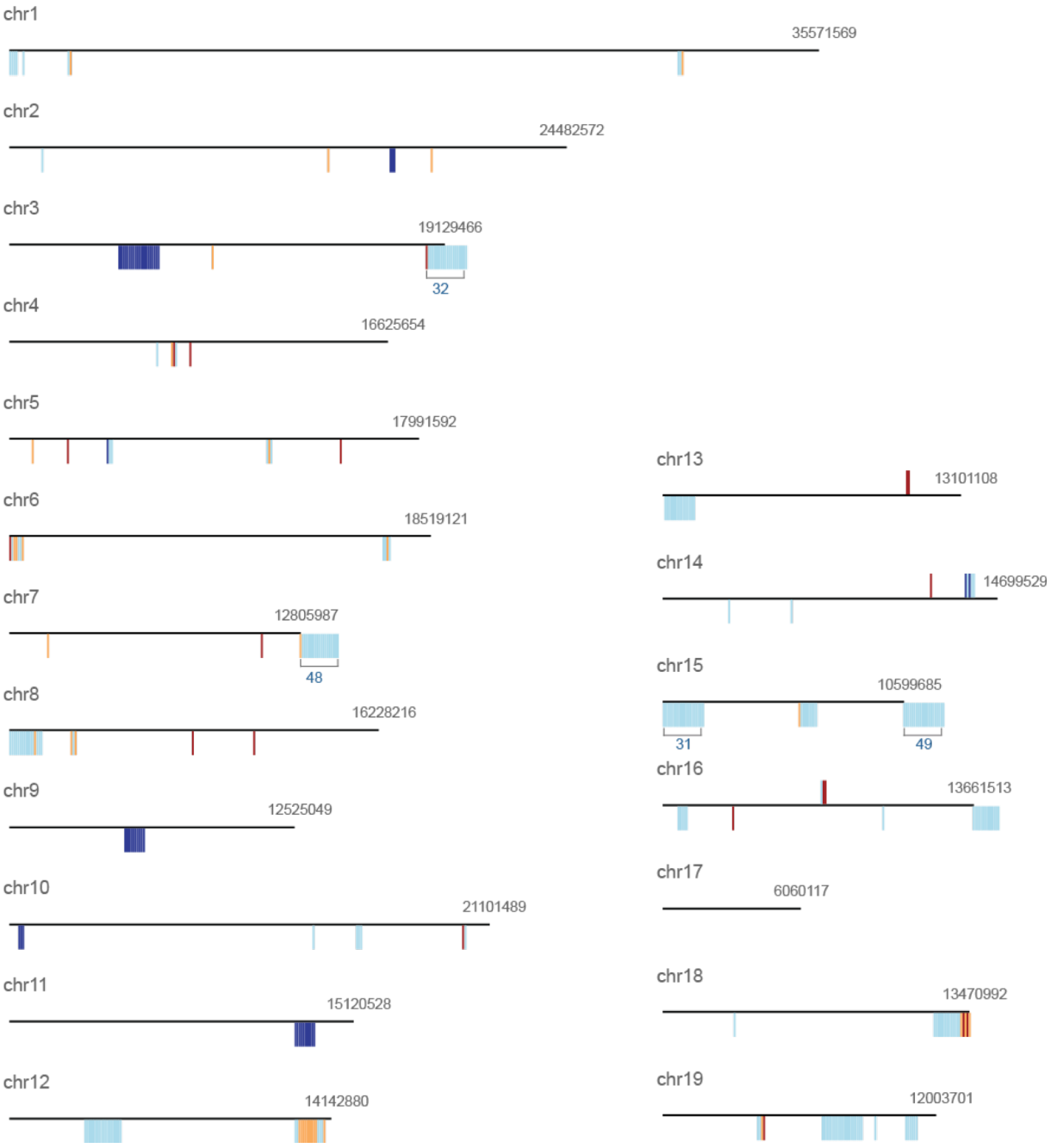
Supplementary Figures



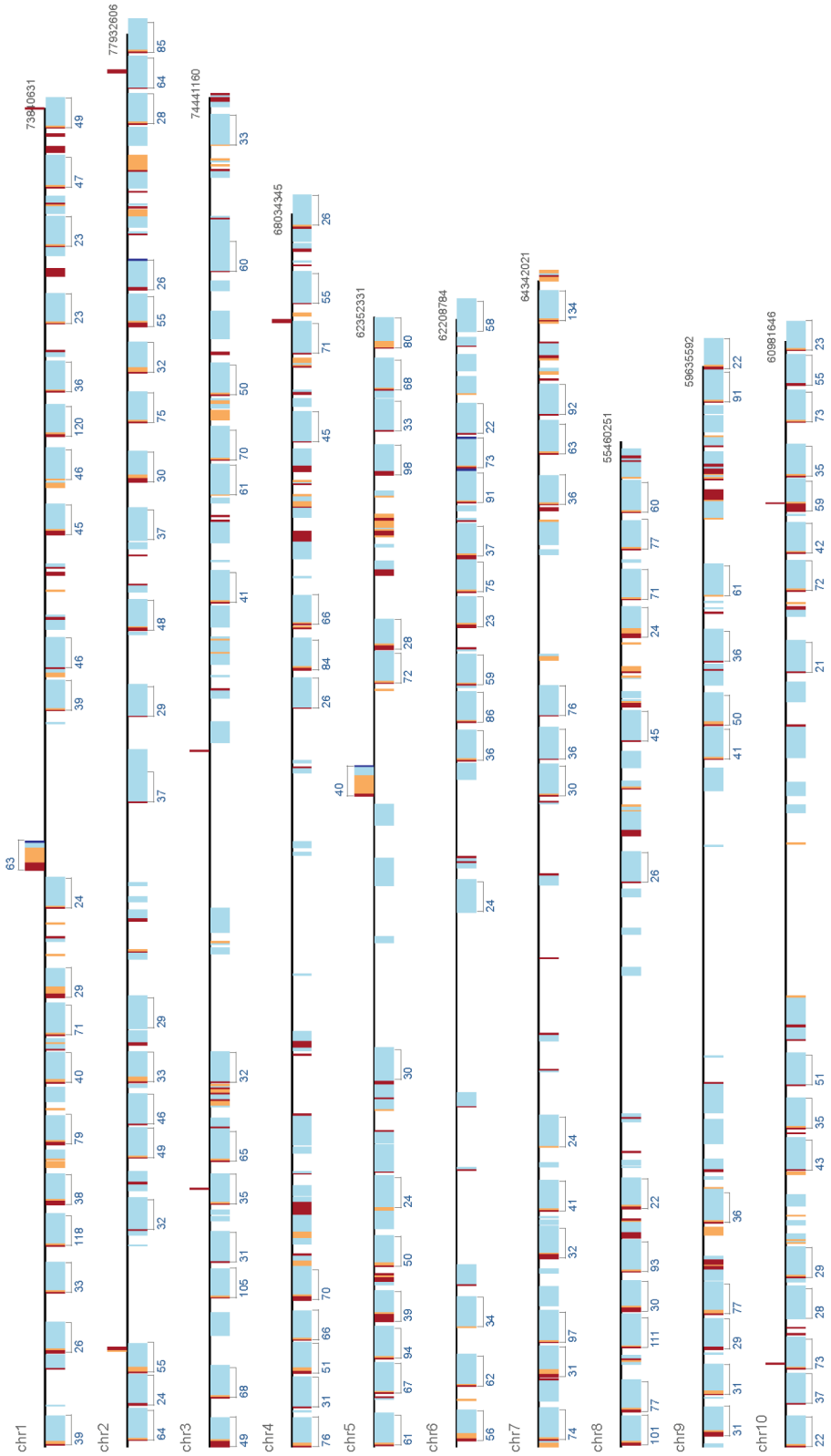
Supplementary Figure 1| Plant LIMEs in rice genome. The LIMEs are depicted as colored ticks with complex LIMEs above and repetitive LIMEs below each chromosome sequence. Tick color corresponds to the number of genomes sharing an LIME: red for 3 genomes, orange for 4, light blue for 5, and dark blue for 6. When two LIMEs are 45kbp or less apart, they are grouped in the same box. Once there are more than 20 LIMEs in such box, the box size is unchanged, but correct proportions of LIMEs shared by 3, 4, 5, and 6 genomes are depicted by the relative thickness of the colored parts. Orange numbers specify the total number of LIMEs per each box, blue correspond to motif ID for one or multiple repetitive LIMEs. Identified centromere positions are shown as grey boxes.



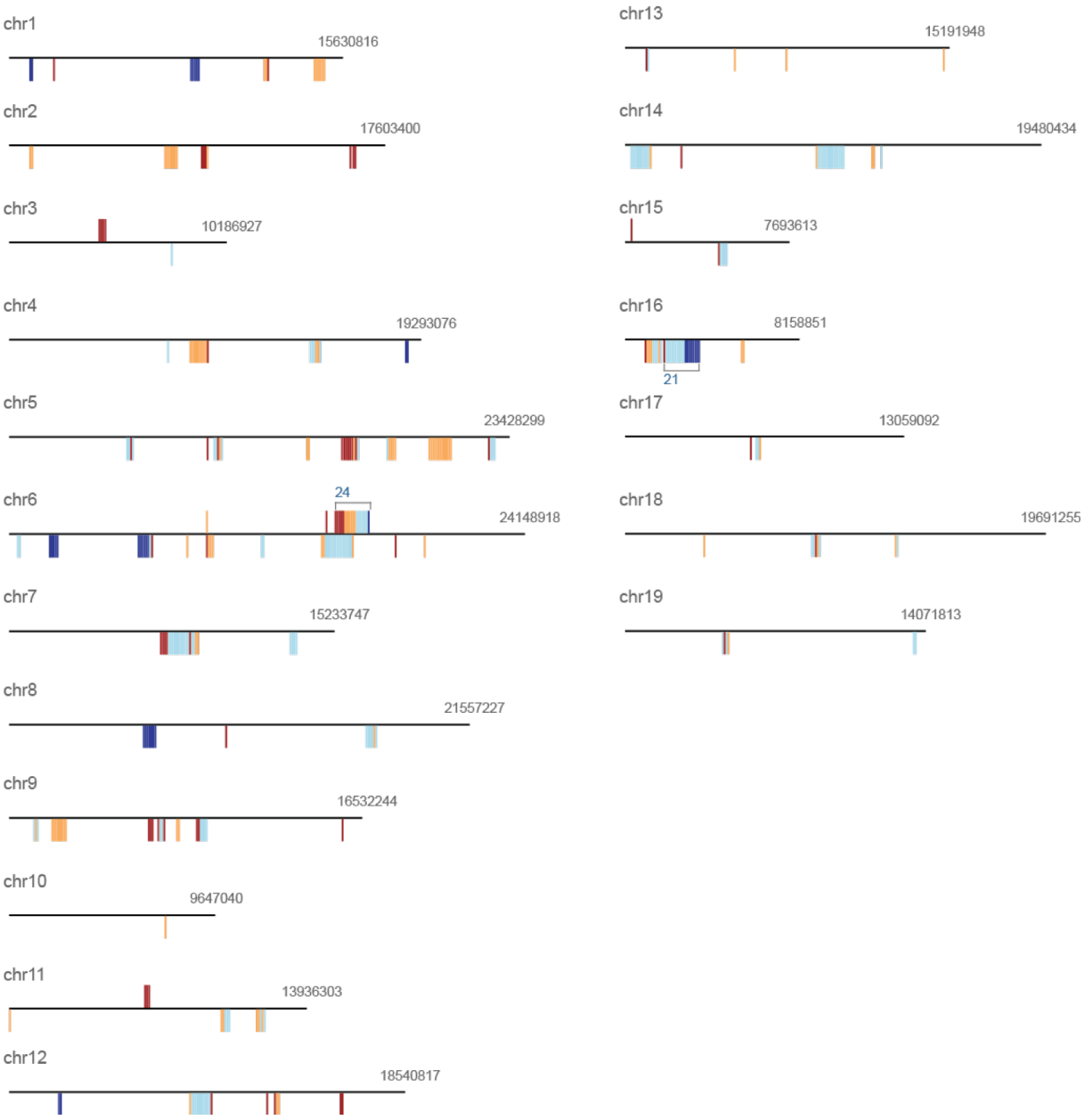
Supplementary Figure 2| Plant LIMEs in soybean genome. a, Chromosomes 1–11. **b**, Chromosomes 12–20. Figure description is the same as in Fig. S1.



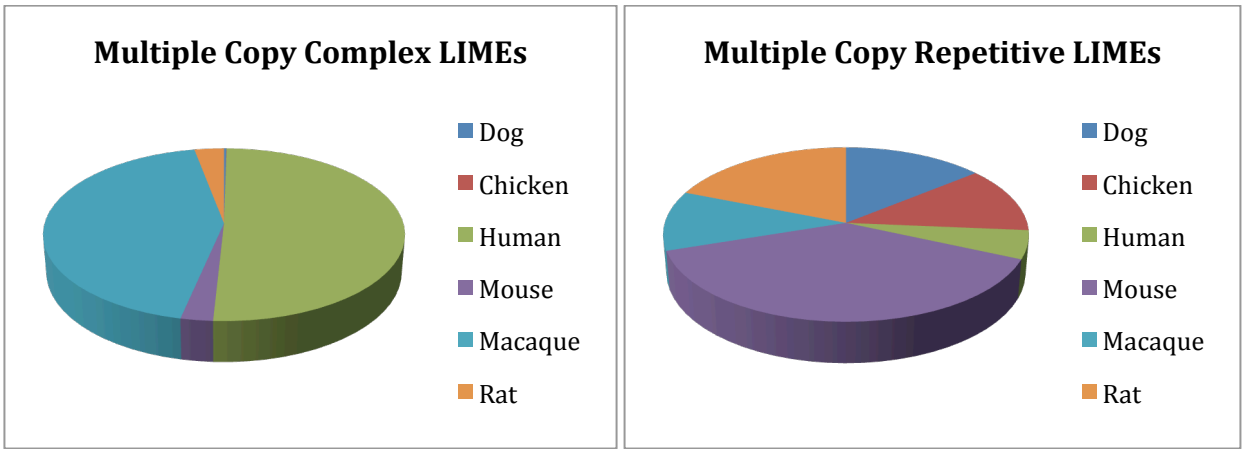
Supplementary Figure 3| Plant LIMEs in cottonwood genome. Figure description is the same as in Fig. S1.



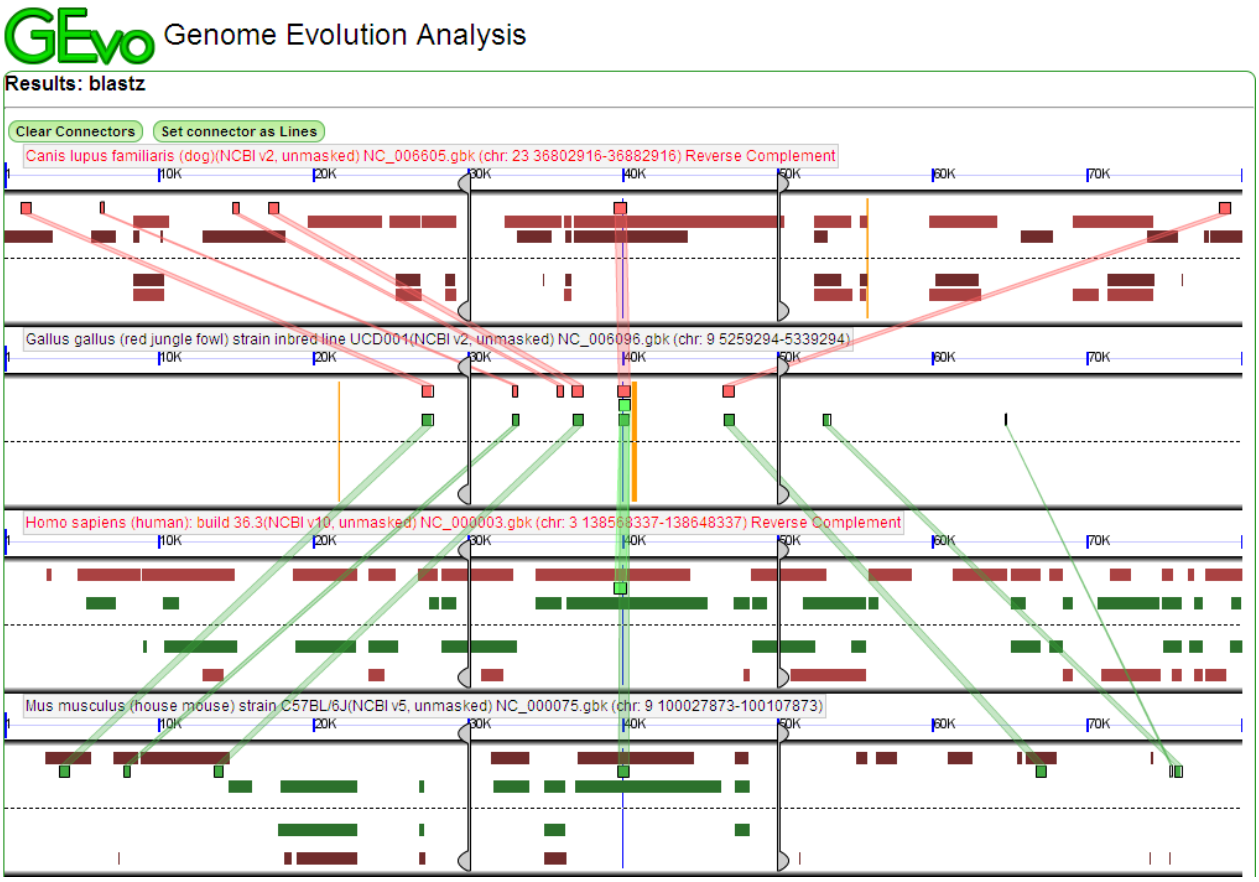
Supplementary Figure 4| Plant LIMEs in sorghum genome. Figure description is the same as in Fig. S1.



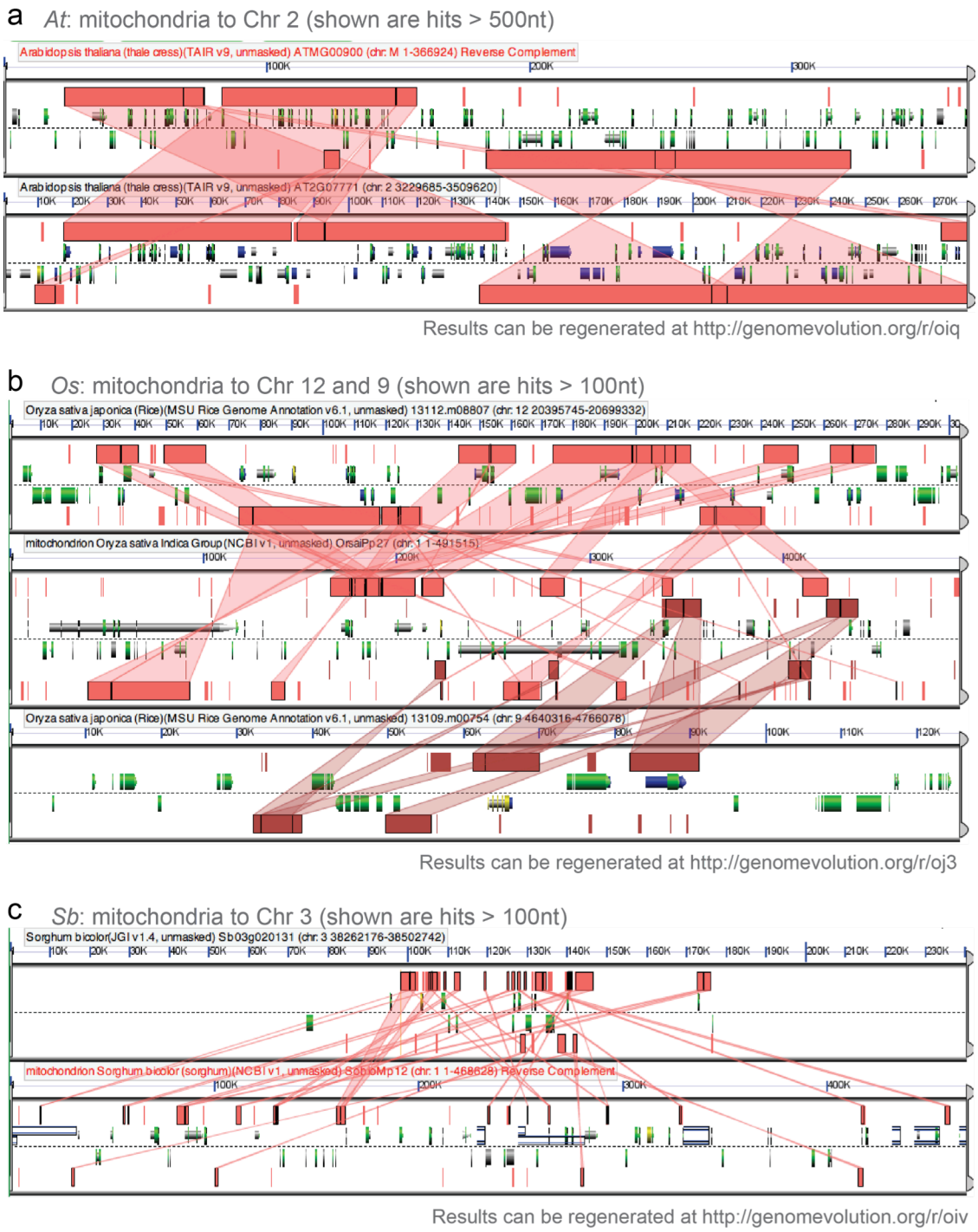
Supplementary Figure 5| Plant LIMEs in grape genome. Figure description is the same as in Fig. S1.



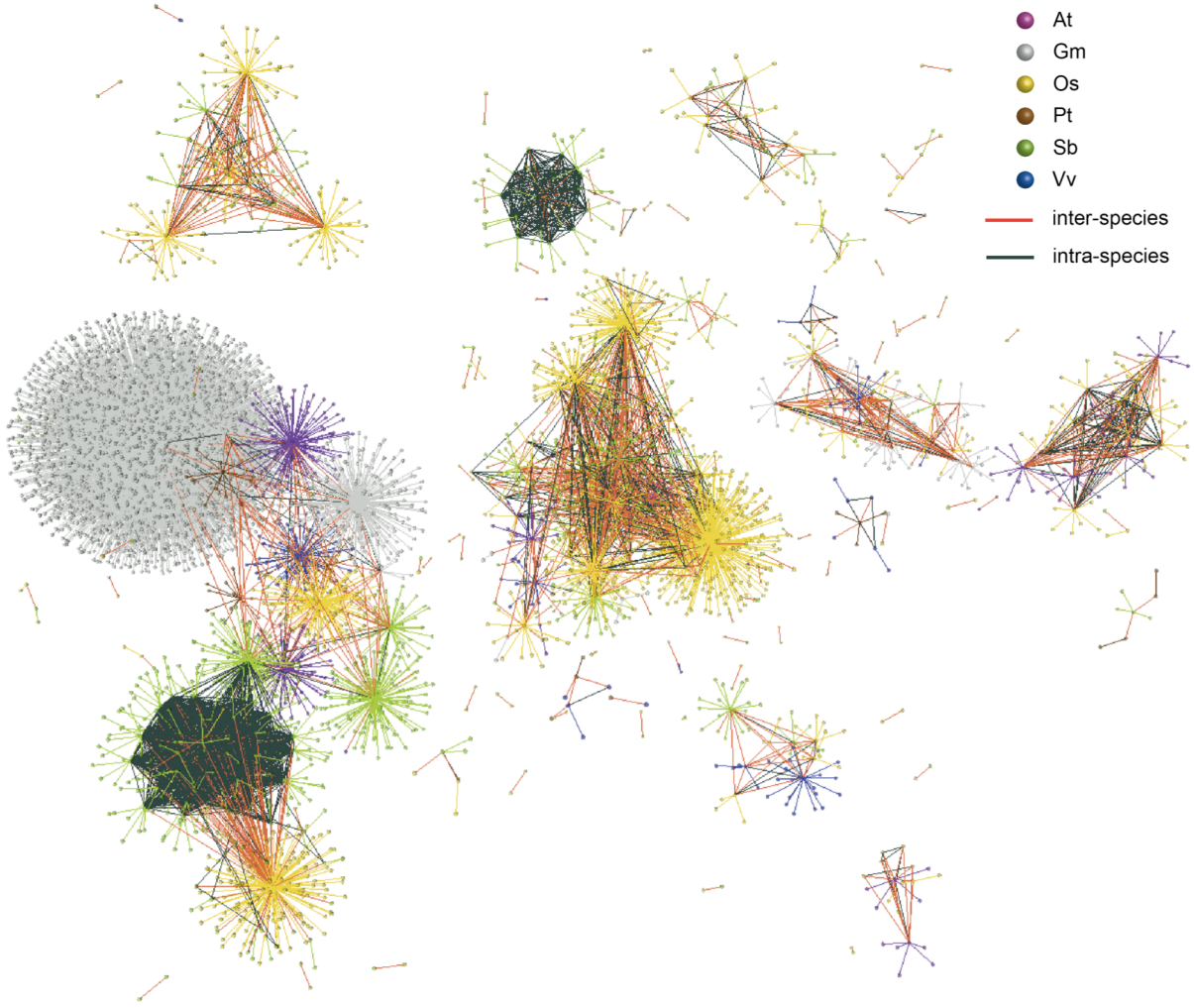
Supplementary Figure 6| Distribution of multiple copy animal LIMEs. Pie charts of multiple copy complex and repetitive LIMEs for six animal genomes.



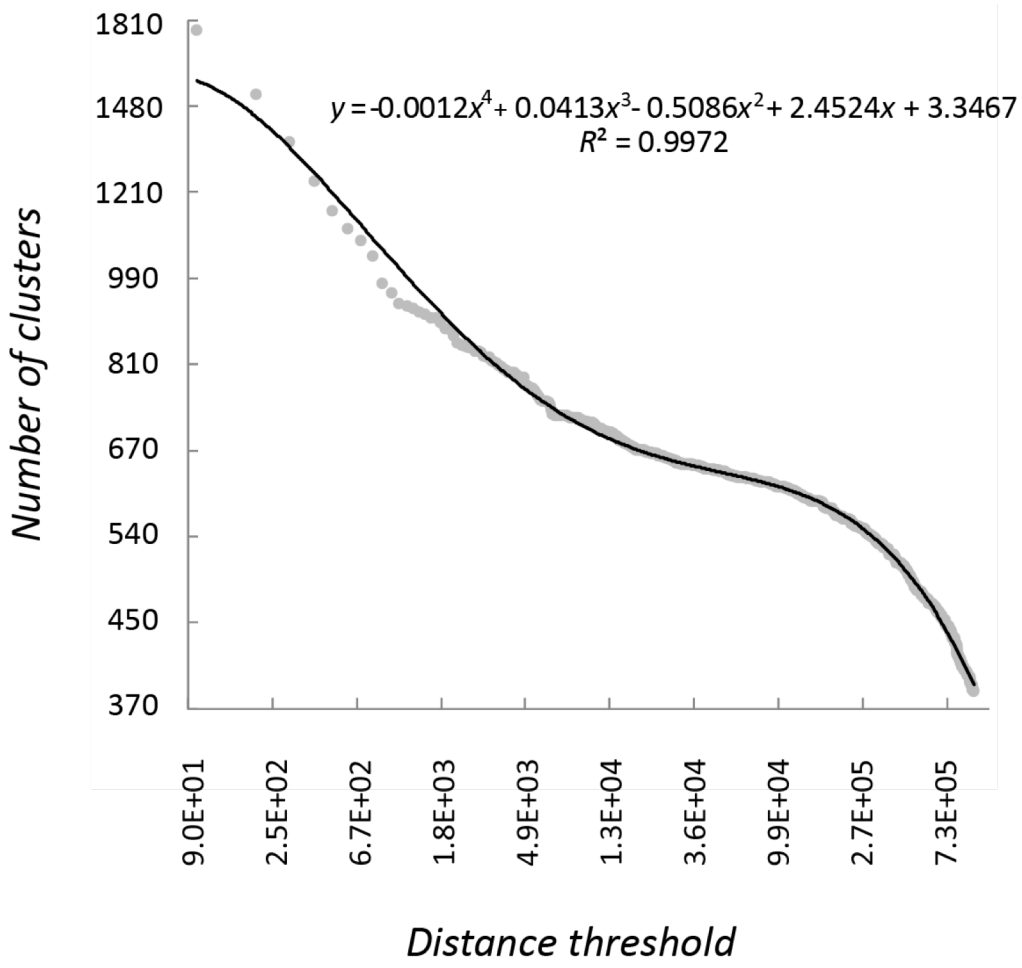
Supplementary Figure 7| Mapping of animal complex element ID 126872 (271 bp) using the comparative genomics platform CoGe. From top to bottom: dog, chicken, human and mouse regions are shown. Each panel represents a genomic region with the dashed line separating the top and bottom strands of DNA. Transparent wedges connect tandem regions of sequence similarity between chicken and the other genomes, as identified by BlastZ; HSPs in the (++) and (+-) orientations are drawn above and below the gene models respectively. Orange blocks represent Ns. There are no genes or other genomic markers present. Rat and macaque also contain element 126872.



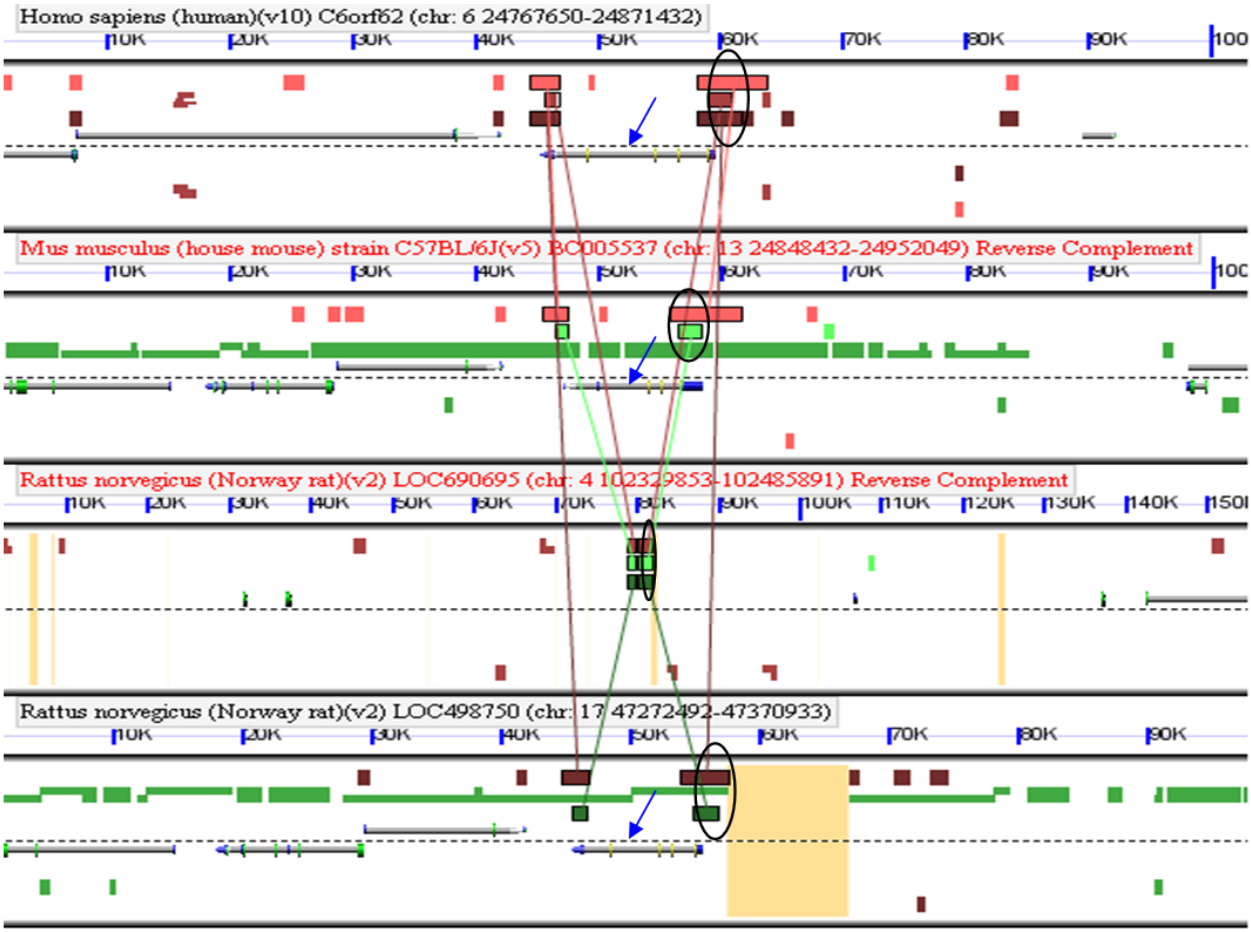
Supplementary Figure 8| Mapping of the mitochondrial to nuclear genomes using the comparative genomics platform CoGe. a, *Arabidopsis* (At). **b**, rice (Os). **c**, sorghum (Sb). Each panel represents a genomic region with the dashed line separating the top and bottom strands of DNA. Protein coding regions are colored green or yellow (gene used to anchor the genomic region in the analysis), mRNAs are colored blue, and genes are colored gray. Gene models drawn above and below the dashed line are transcribed from the top and bottom strands of DNA respectively. Red transparent wedges connect regions of sequence similarity, as identified by BlastZ; HSPs in the (++) and (+-) orientations are drawn above and below the gene models respectively.



Supplementary Figure 9| Clusters of plant LIMEs are organized into networks. The network of complex LIMEs from *Arabidopsis* (At, maroon node), soybean (Gm, grey nodes), rice (Os, gold nodes), cottonwood (Pt, brown nodes), sorghum (Sb, green nodes), and grape (Vv, blue nodes). All elements in one cluster are connected to a selected representative with the edges of the same colour as nodes. Clusters of LIMEs within one species are connected through the representative nodes with dark green edges if they share one or more multi-copy complex LIMEs. Clusters sharing LIMEs across multiple species are connected through their representatives with red edges.



Supplementary Figure 10| Number of LIME clusters. Log-log plot of a maximal distance allowed between two LIMEs in the same cluster against the total number of clusters obtained as a result of agglomerative clustering using the maximal distance as a threshold (grey dots). Solid black line is a fitted polynomial function of degree 4.



Supplementary Figure 11| CoGe screen shot of the regions in human, mouse and rat containing element #4 from Table S1. Above and below each dotted line indicates forward and reverse directions on the DNA strand, respectively. Orange blocks are regions of sequence containing Ns. From top to bottom, the four regions shown are from human chr. 6, mouse chr. 13, rat chr. 4 and rat chr. 17. Other symbols are described in the section "Exact match subsequence in human, mouse and rat".

References

39. Makeyev AV*, et al.* (2005) HnRNP A3 genes and pseudogenes in the vertebrate genomes. (Translated from eng) *J Exp Zool A Comp Exp Biol* 303(4):259-271 .
40. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403-410.
41. Richards EJ & Ausubel FM (1988) Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. *Cell* 53(1):127-136.
42. Cox AV*, et al.* (1993) Comparison of plant telomere locations using a PCR-generated synthetic probe. *Annals of Botany* 72(3):239.
43. Fuchs J, Brandes A, & Schubert I (1995) Telomere sequence localization and karyotype evolution in higher plants. *Plant Systematics and Evolution* 196(3):227-241.
44. Adams SP*, et al.* (2001) Loss and recovery of Arabidopsis-type telomere repeat sequences 5'-(TTTAGGG)(n)-3' in the evolution of a major radiation of flowering plants. *Proc Biol Sci* 268(1476):1541-1546.
45. Sangwan I & O'Brian MR (2002) Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. *Plant Physiol* 129(4):1788-1794.
46. Lehmann M (2004) Anything else but GAGA: a nonhistone protein complex reshapes chromatin structure. (Translated from eng) *Trends Genet* 20(1):15-22 .
47. Bellstedt DU, Linder HP, & Harley EH (2001) Phylogenetic relationships in Disa based on non-coding trnL-trnF chloroplast sequences: evidence of numerous repeat regions. *Am. J. Bot.* 88(11):2088-2100.
48. Adai A, Date S, Wieland S, & Marcotte E (2004) LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *Journal of Molecular Biology* 340(1):179-190.
49. Tang H*, et al.* (2008) Synteny and collinearity in plant genomes. *Science* 320(5875):486-488.
50. Adams KL & Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8(2):135-141.
51. Koonin EV, Wolf YI, & Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420(6912):218-223.
52. Jaillon O*, et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161):463-467.
53. Lyons E*, et al.* (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* 148(4):1772-1781.
54. Bowers JE, Chapman BA, Rong J, & Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433-438.
55. Yu J*, et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296(5565):79-92.
56. Tang H, Bowers JE, Wang X, & Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A* 107(1):472-477.