iScience, Volume *25*

Supplemental information

Transposons and non-coding regions drive the intrafamily differences of genome size in insects

Yuyang Cong, Xinhai Ye, Yang Mei, Kang He, and Fei Li

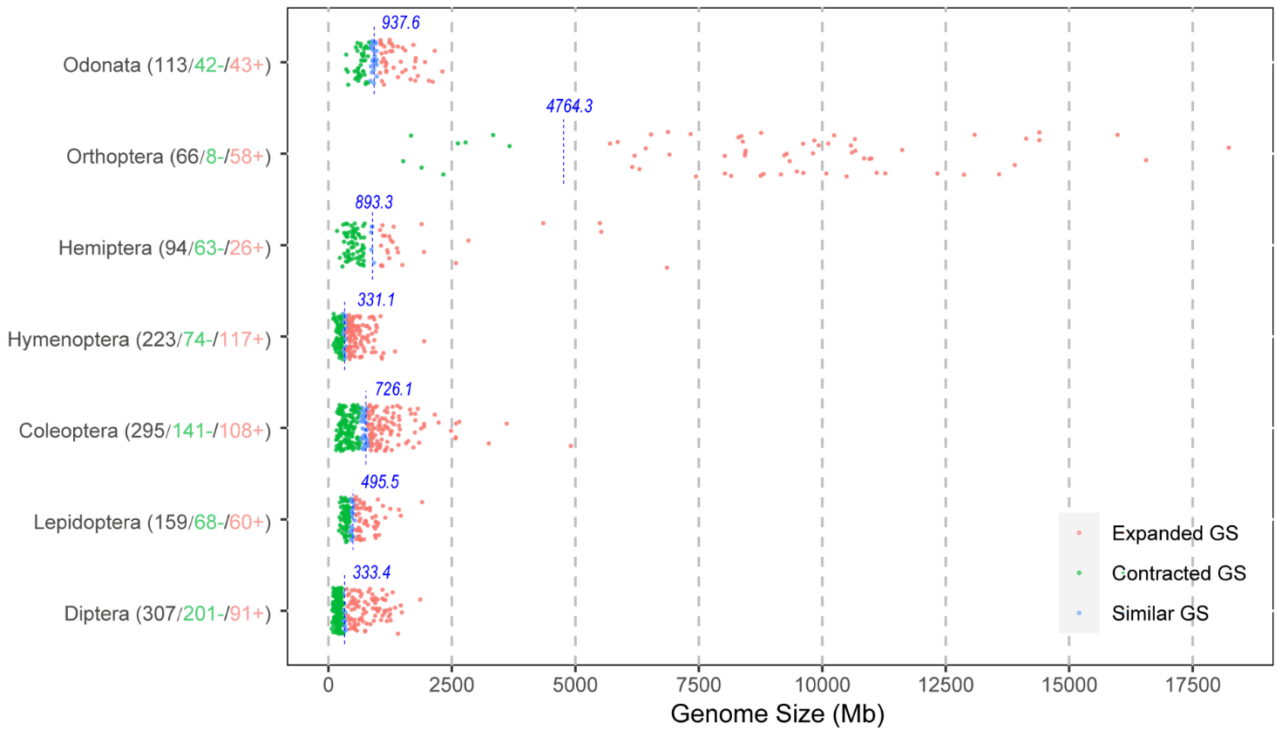

Figure S1. Distribution pattern of GS expansion and contraction in seven insect orders. Blue dashed lines labeled with specific numbers represent estimated values of ancestral GS in respective clades. GS dots which are 10 % greater than the ancestral values are defined as contracted GS, 10 % less than ancestral value as expanded GS and the rest as similar GS. Corresponding species numbers (total /contracted /expanded) are marked in the left. Related to Figure 1b.

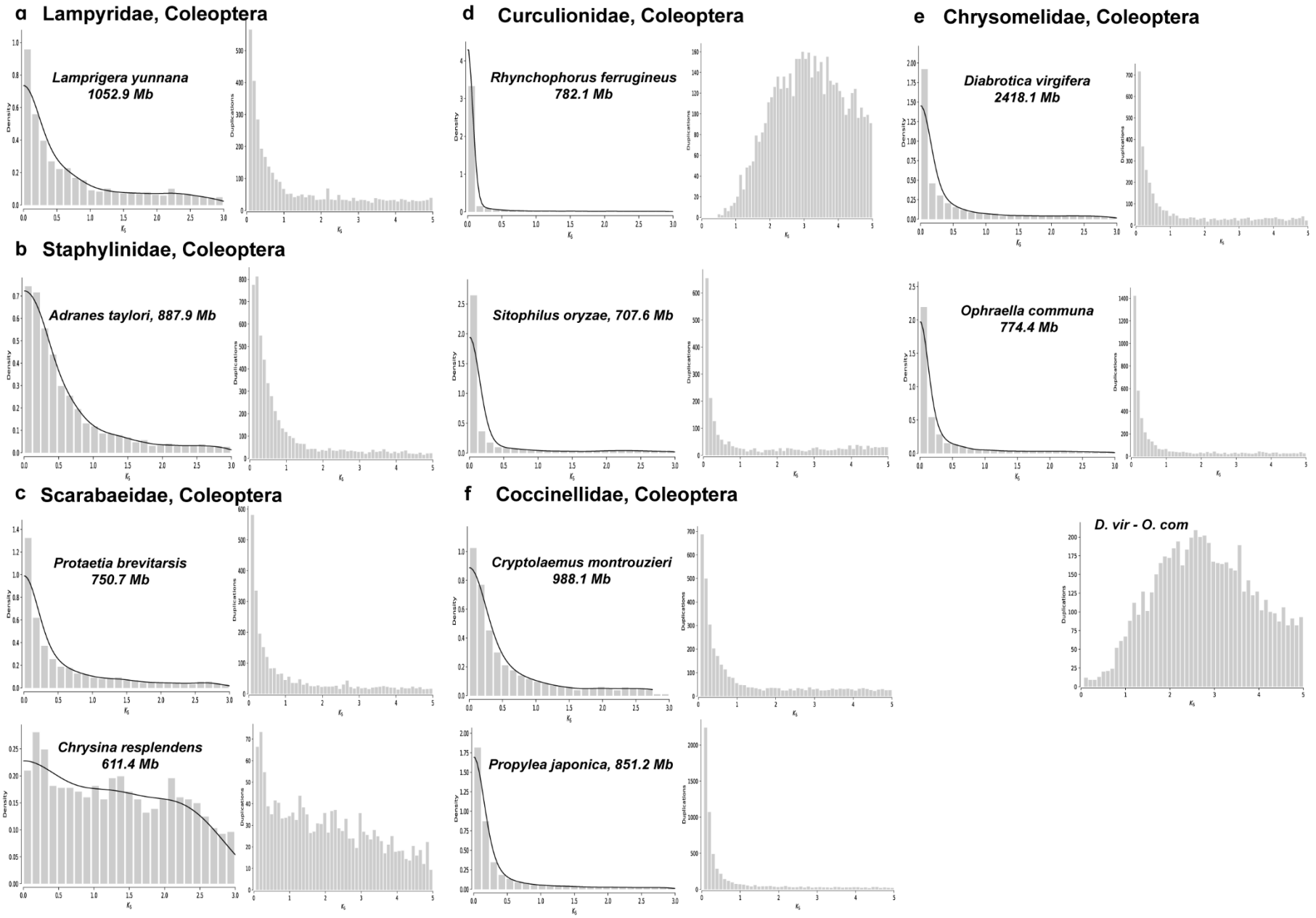

Figure S2. Inferring ancient WGD events based on the Ks age distribution of gene duplicates in six families of Coleoptera. Plots on the right are histograms of Ks (y-axis: numbers of duplicated paralog genes) and the left are the corresponding density fitting plot by Kernel density estimates (KDE) using the wgd tools (Zwaenepoel and Van de Peer 2019). Related to Figure 3.

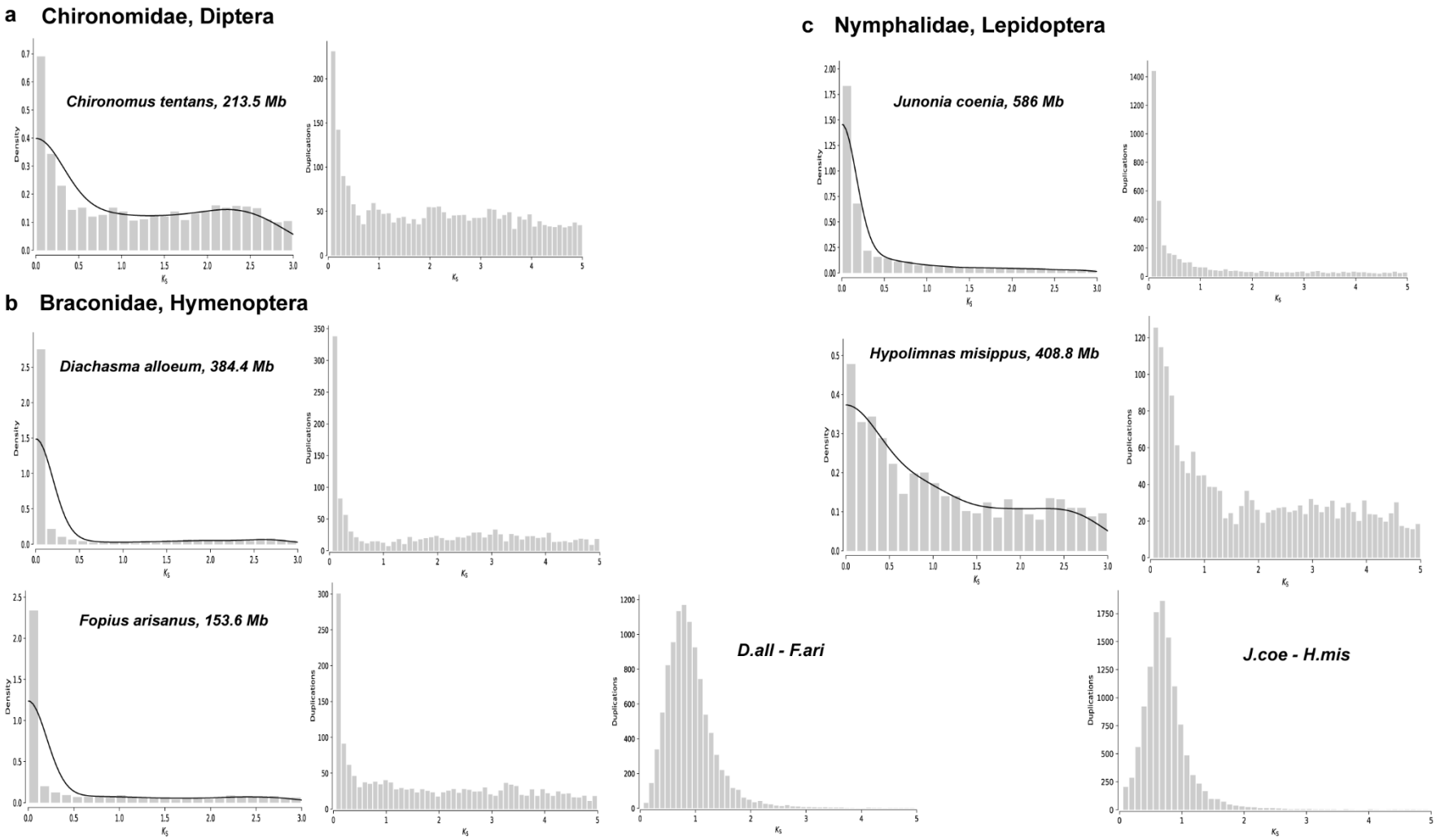

Figure S3. Inferring ancient WGD events based on the Ks age distribution of gene duplicates in (a) Family Chironomidae of Dipetra, (b) Family Braconidae of Hymenoptera, and (c) Family Nymphalidae of Lepidoptera. Plots on the right are histograms of Ks (y-axis: numbers of duplicated paralog genes) and the left are the corresponding density fitting plot by Kernel density estimates (KDE) using the wgd tools (Zwaenepoel and Van de Peer 2019). Related to Figure S4, S7, S9.

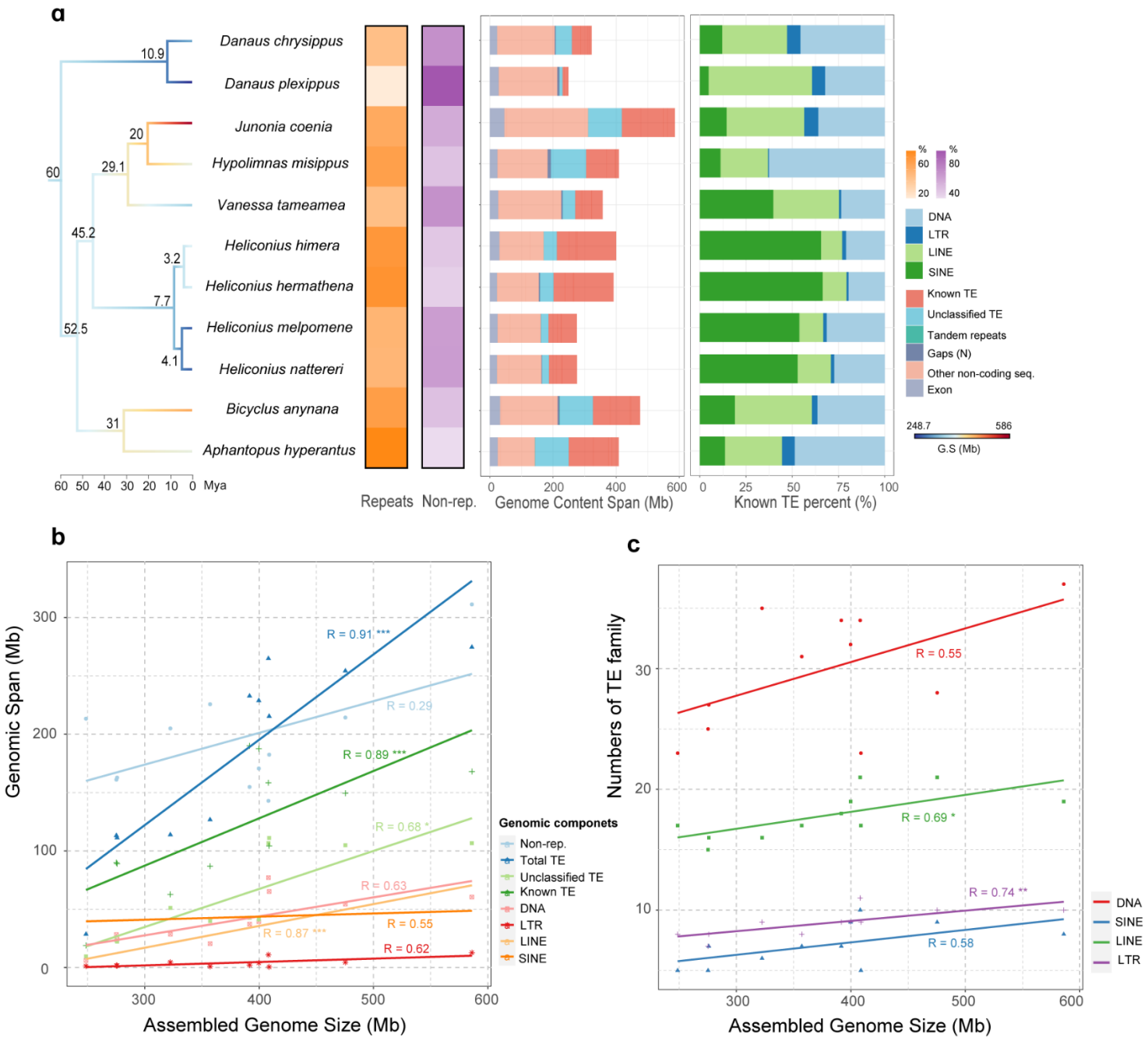

Figure S4. Genomic landscapes, correlation between GS and genomic components and TE diversity of 11 butterflies in the family of Nymphalidae, Order Lepidoptera. a) Genomic landscapes of 11 butterflies. Time-calibrated phylogenetic relationship inferred from the BUSCO core gene sets using ML (see Method details) is shown on the left, along with the heatmaps of the coverage of repetitive (orange) and non-repetitive (purple) content (%) in each lineage. Genome content divided into six major genomic components and the relative abundance of annotated known TEs are shown by bar charts on the right. **b)** Correlation between assembled GS and contents of the genomic components in 11 butterflies. Numbers following the regression lines correspond to Pearson's correlation coefficient under the phylogenetic independent contrast (PIC) with *p*-values (***＜0.001; 0.001≤**＜0.01; 0.01≤*＜0.05). **c)** Correlation between TE diversity (the numbers of the annotated TE families) and the assembled GS of butterflies in Nymphalidae. Pearson correlation coefficients are noted near the regression lines with *p*-values (***＜0.001; 0.001≤**＜0.01; 0.01≤*＜0.05). Related to Figure 3, Table S6, Table S7d, Table S8.

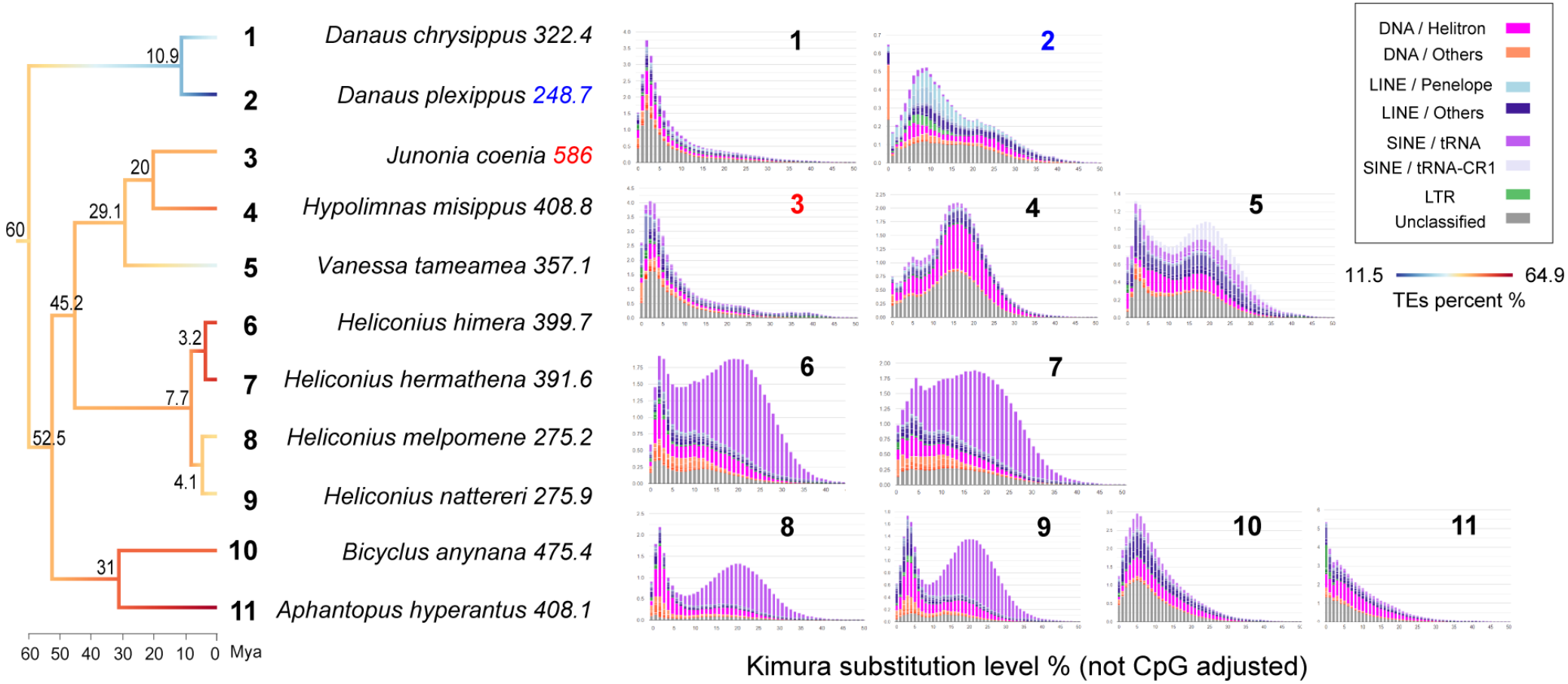

Figure S5. TE divergency landscapes of 11 butterflies in the family of Nymphalidae, Order Lepidoptera. Branches on the time-calibrated phylogeny are colored according to the total abundance of TEs. Species' Latin names are followed by their GS number, the biggest GS are in red and the smallest in blue. Each species is numbered, and each number corresponds to a small graph on the right showing the TE divergency of species. The y-axis represents genome coverage of various types of TE, and the x-axis represents the Kimura substitution level % (TE divergency based on the Kimura distance of TE copies to the corresponding consensus sequences) from 0 to 50. Related to Figure 4.

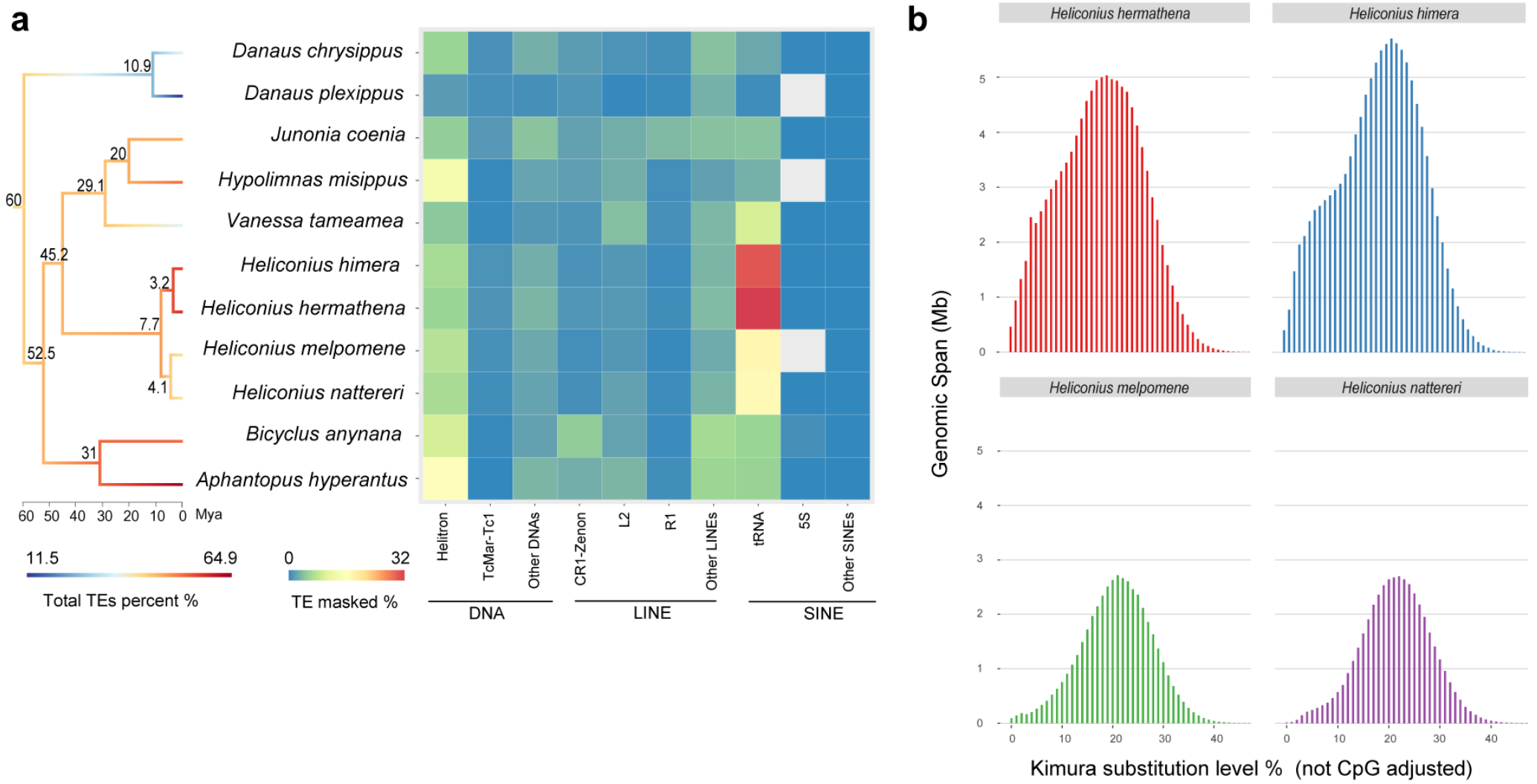

Figure S6. Genomic abundance and divergency of several primary TEs in 11 butterflies. a) Time-calibrated phylogeny of 11 butterflies in the family Nymphalidae of Lepidoptera with total TEs coverage mapped on the multi-colored branches. The heatmap on the right represents genome coverage of the top three TEs among all TE types. **b)** Kimura substitution level % (TE divergency based on the Kimura distance of TE copies to the corresponding consensus sequences) of SINE/tRNA elements in four butterflies of the *Heliconius* genus. Related to Figure 5, Table S9b.

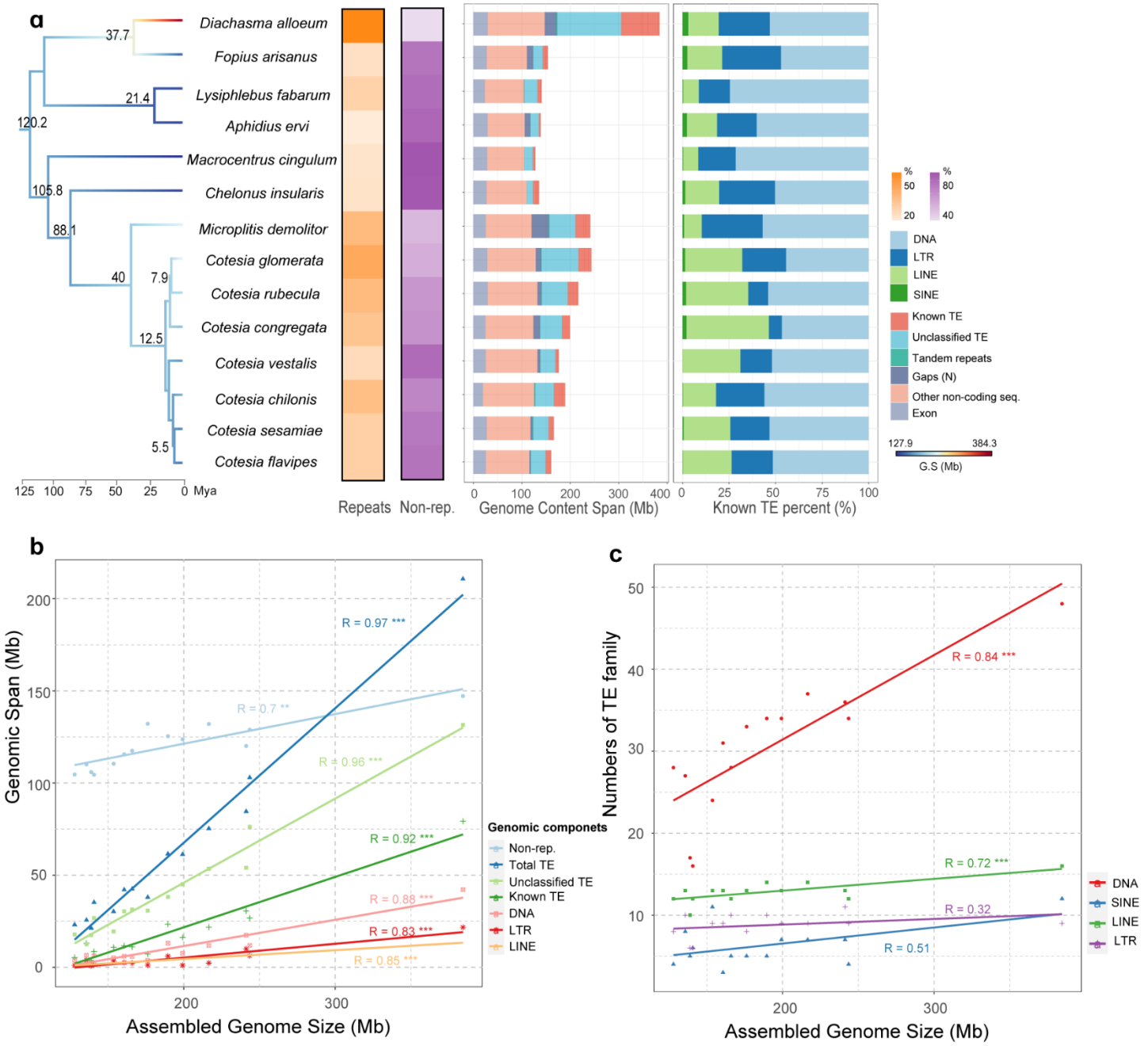

Figure S7. Genomic landscapes, correlation between GS and genomic components and TE diversity of 14 bees in the family of Braconidae, Order Hymenoptera. a) Genomic landscapes of 14 bees. Time-calibrated phylogenetic relationship inferred from the BUSCO core gene sets using ML (see Method details) is shown on the left, along with the heatmaps of the coverage of repetitive (orange) and non-repetitive (purple) content (%) in each lineage. Genome content divided into six major genomic components is shown by bar charts on the right, along with the relative abundance of annotated known TEs. **b)** Correlation between assembled GS and contents of the genomic components in 14 bees. Numbers following the regression lines correspond to Pearson's correlation coefficient under the phylogenetic independent contrast (PIC) with *p*-values (***$<$0.001; 0.001$\leq$**$<$0.01; 0.01$\leq$*$<$0.05). **c)** Correlation between TE diversity (the numbers of the annotated TE families) and the assembled GS of bees in Braconidae. Pearson correlation coefficients are noted near the regression lines with *p*-values (***$<$0.001; 0.001$\leq$**$<$0.01; 0.01$\leq$*$<$0.05). Related to Figure 3, Table S6, Table S7c, Table S7d, Table S8.

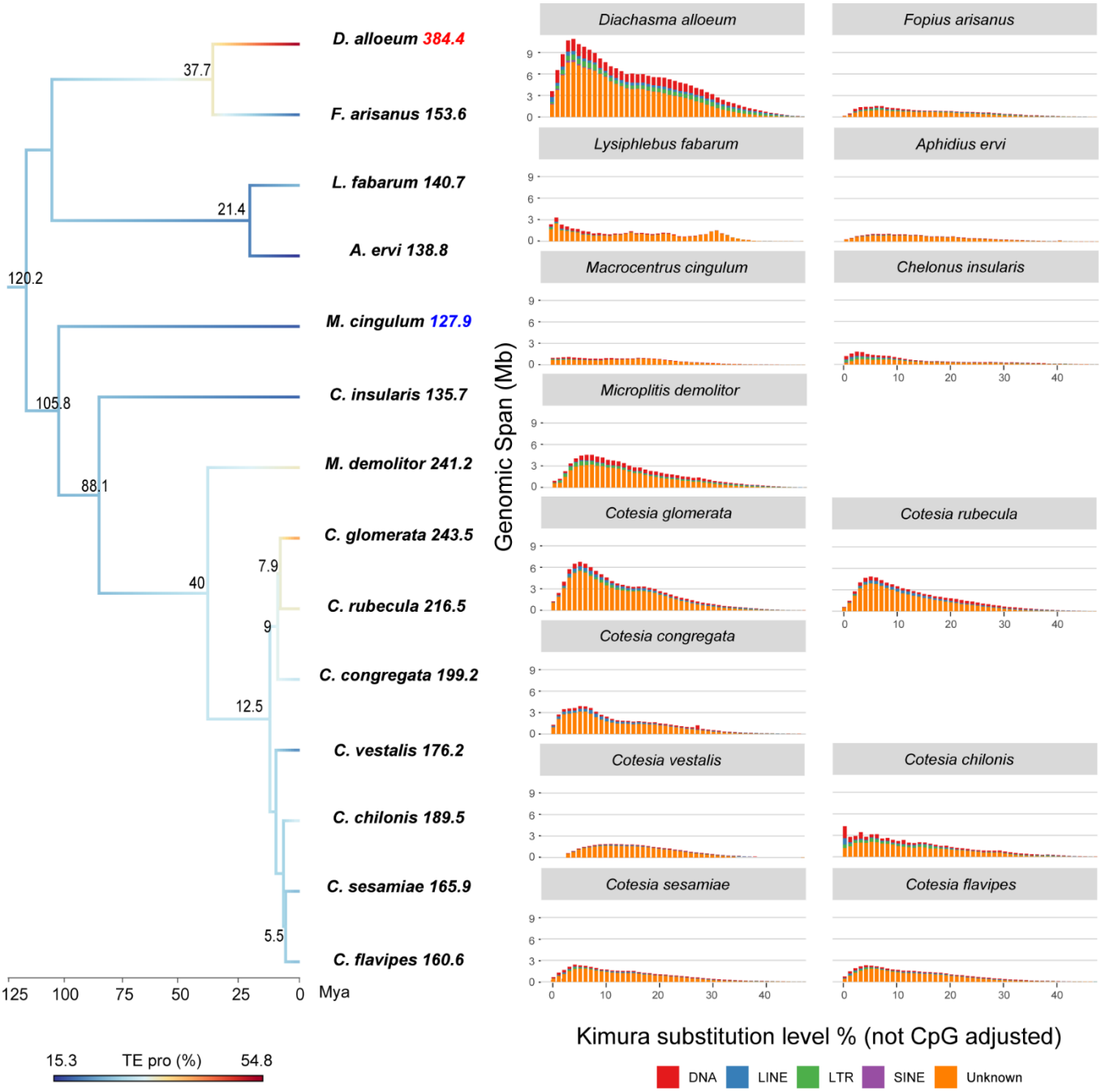

Figure S8. TE divergency landscapes of 14 bees in the family of Braconidae, Hymenoptera. Branches on the time-calibrated phylogeny are colored according to the total abundance of TEs. Species' Latin names are followed by their GS number, the biggest GS are in red and the smallest in blue. Small graphs on the right show the TE divergency of the corresponding species. The y-axis represents genome coverage of various types of TE, and the x-axis represents Kimura substitution level % (TE divergency based on the Kimura distance of TE copies to the corresponding consensus sequences) from 0 to 50. Related to Figure 4.

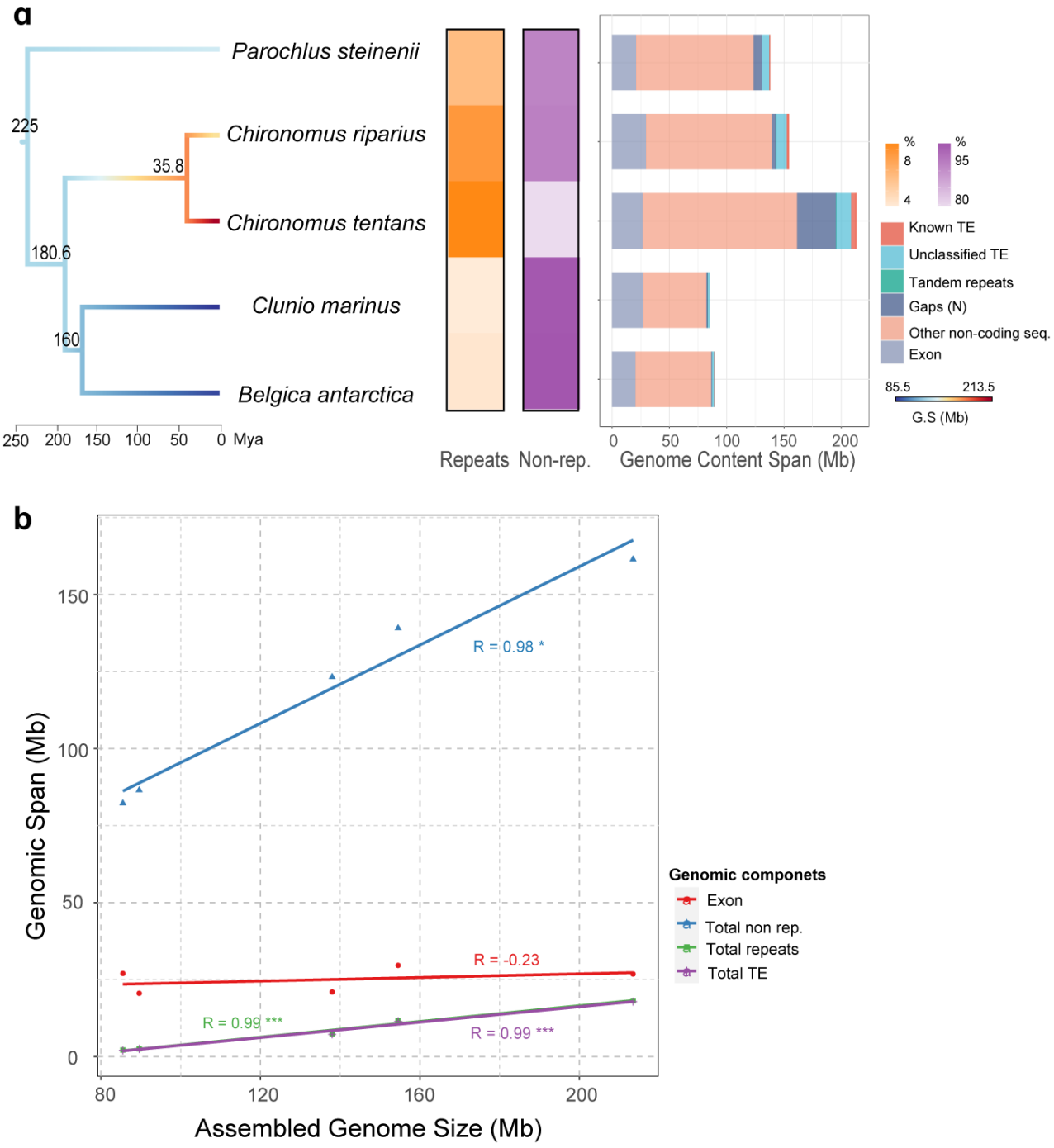

Figure S9. Genomic landscapes and correlation between GS and genomic components of 11 butterflies in the family of Chironomidae, Order Diptera. a) Genomic landscapes of five midges in the family. Time-calibrated phylogenetic relationship inferred from the BUSCO core gene sets using ML (see Method details) is shown on the left, along with the heatmaps of the coverage of repetitive (orange) and non-repetitive (purple) content (%) in each lineage. Genome content divided into six major genomic components is shown by bar charts on the right, along with the relative abundance of annotated known TEs. **b)** Correlation between assembled GS and contents of the genomic components in five midges. Numbers following the regression lines correspond to Pearson's correlation coefficient under the phylogenetic independent contrast (PIC) with *p*-values (***＜0.001; 0.001≤**＜0.01; 0.01≤*＜0.05. Related to Figure 3, Table S6, Table S7b, Table S8a.

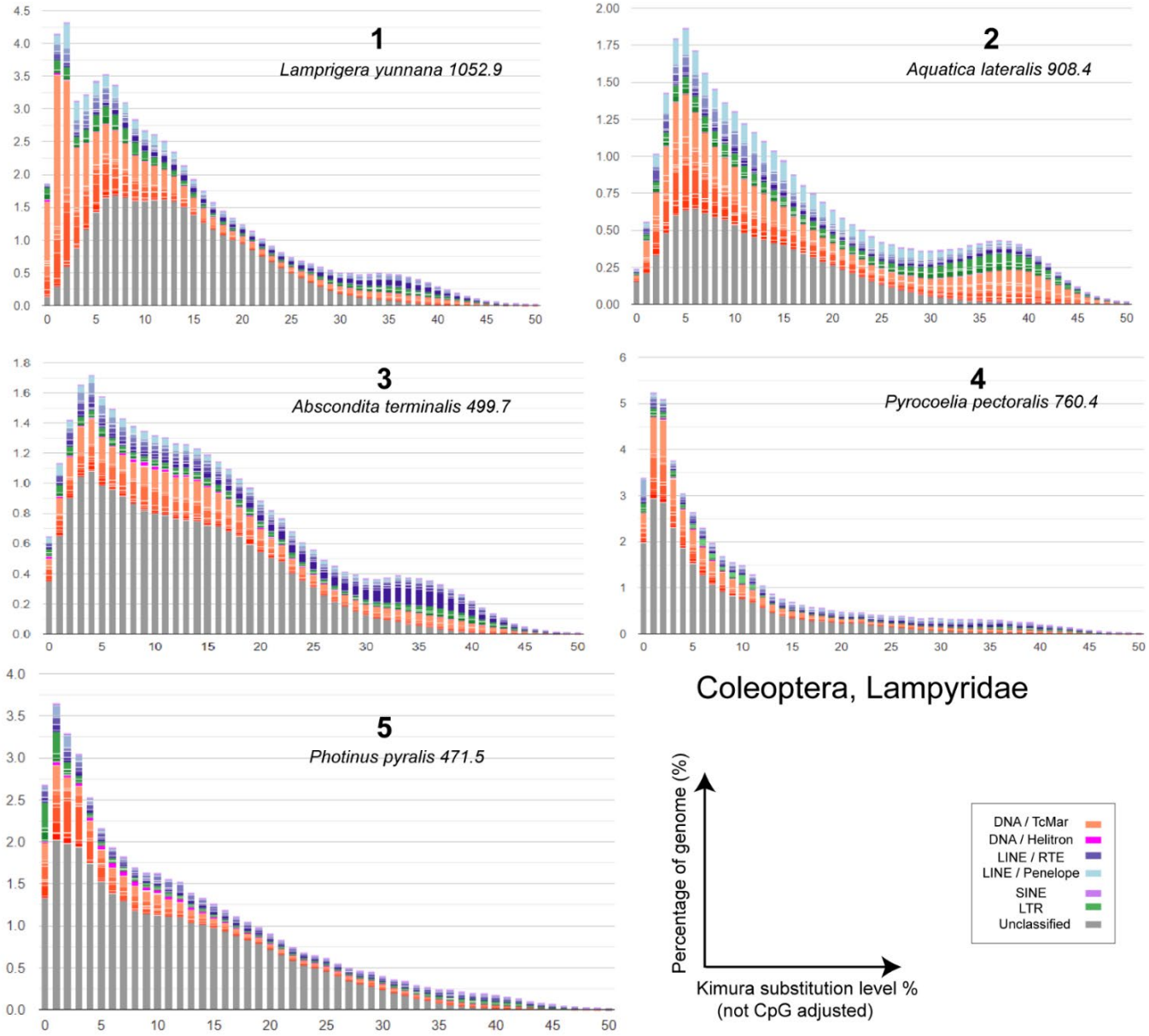

Figure S10. TE divergency landscapes of beetles in Family Lampyridae, Order Coleoptera. Related to Figure 4.

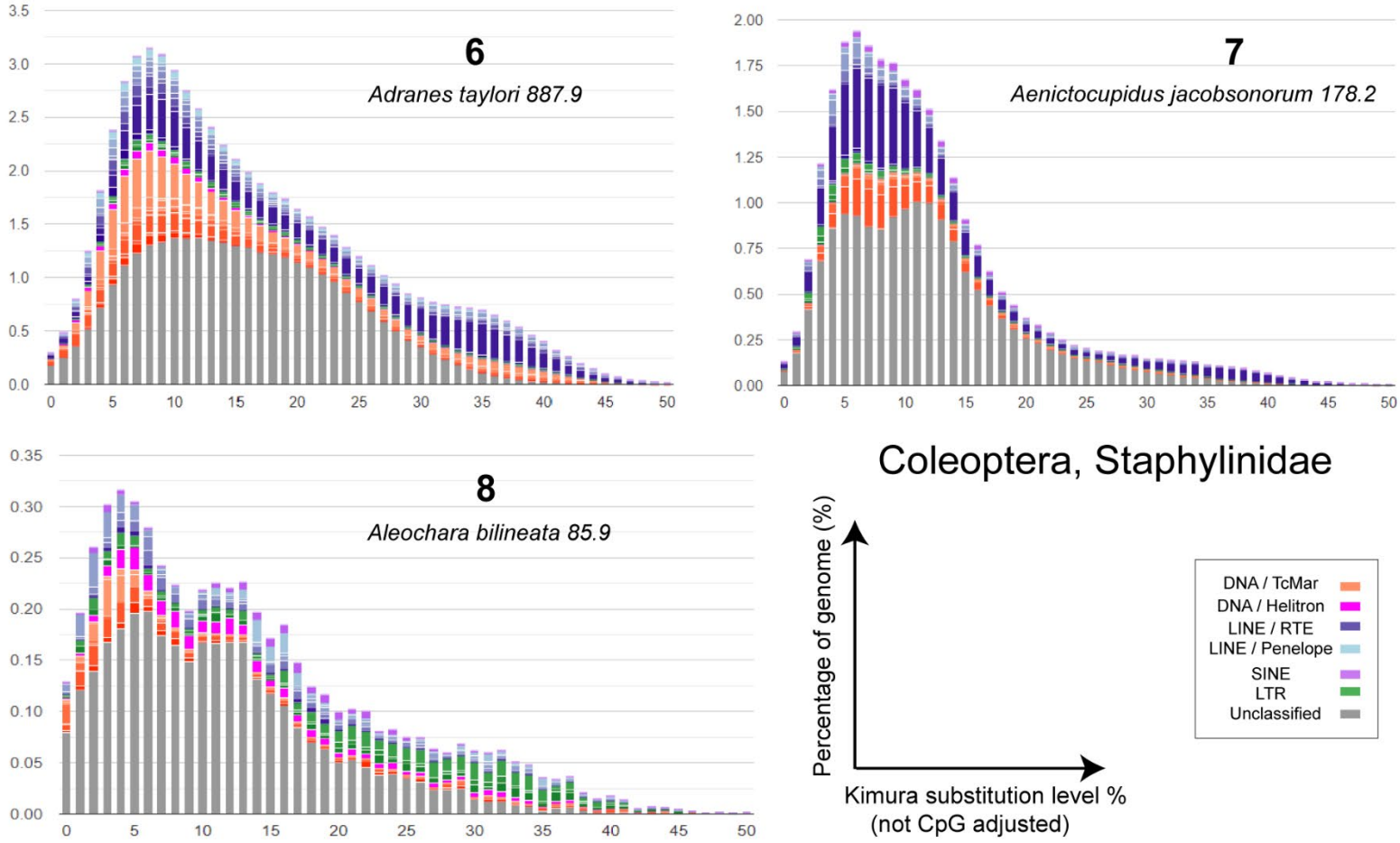

Figure S11. TE divergency landscapes of beetles in Family Staphylinidae, Order Coleoptera. Related to Figure 4.

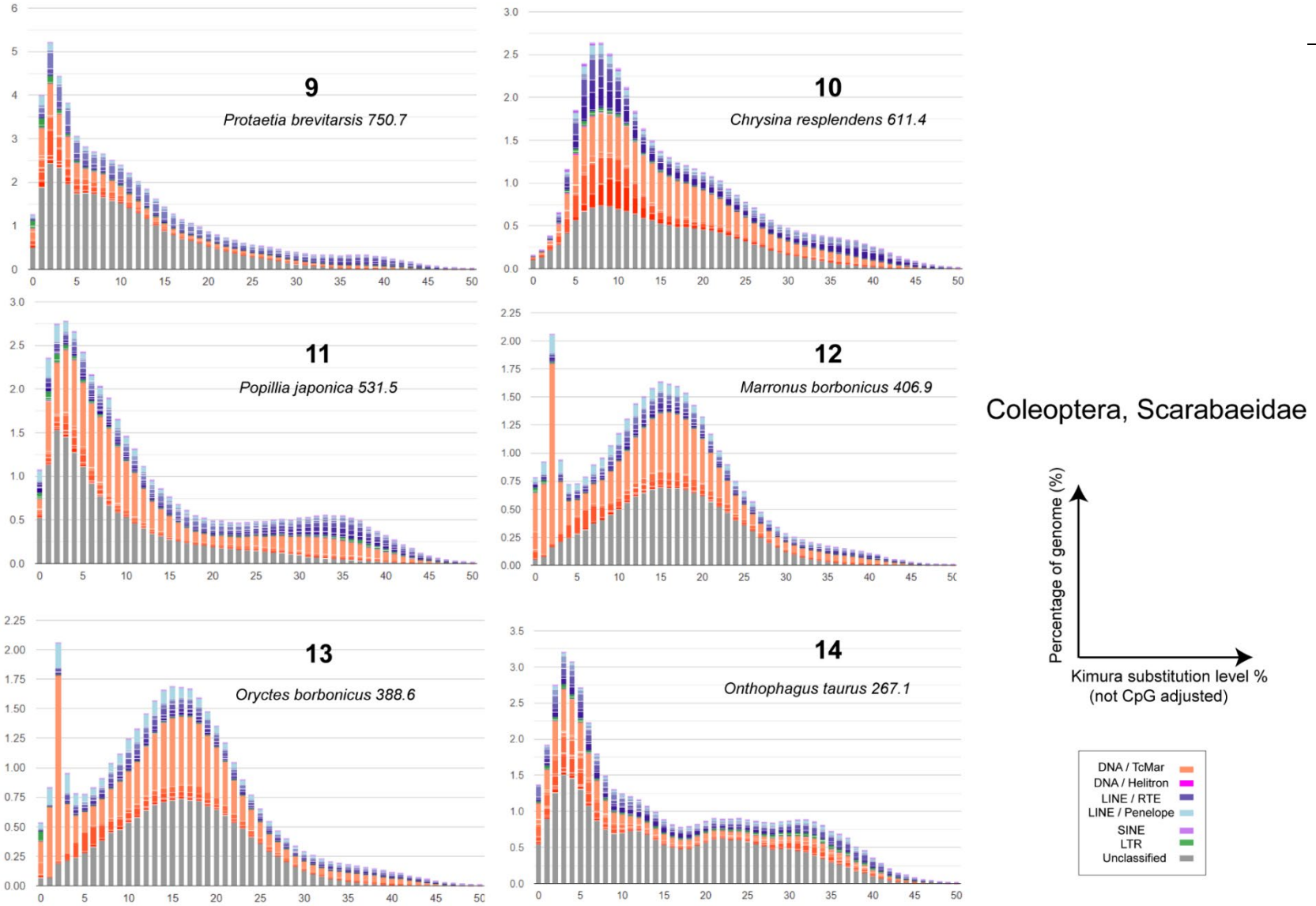

Figure S12. TE divergency landscapes of beetles in Family Scarabaeidae, Order Coleoptera. Related to Figure 4.

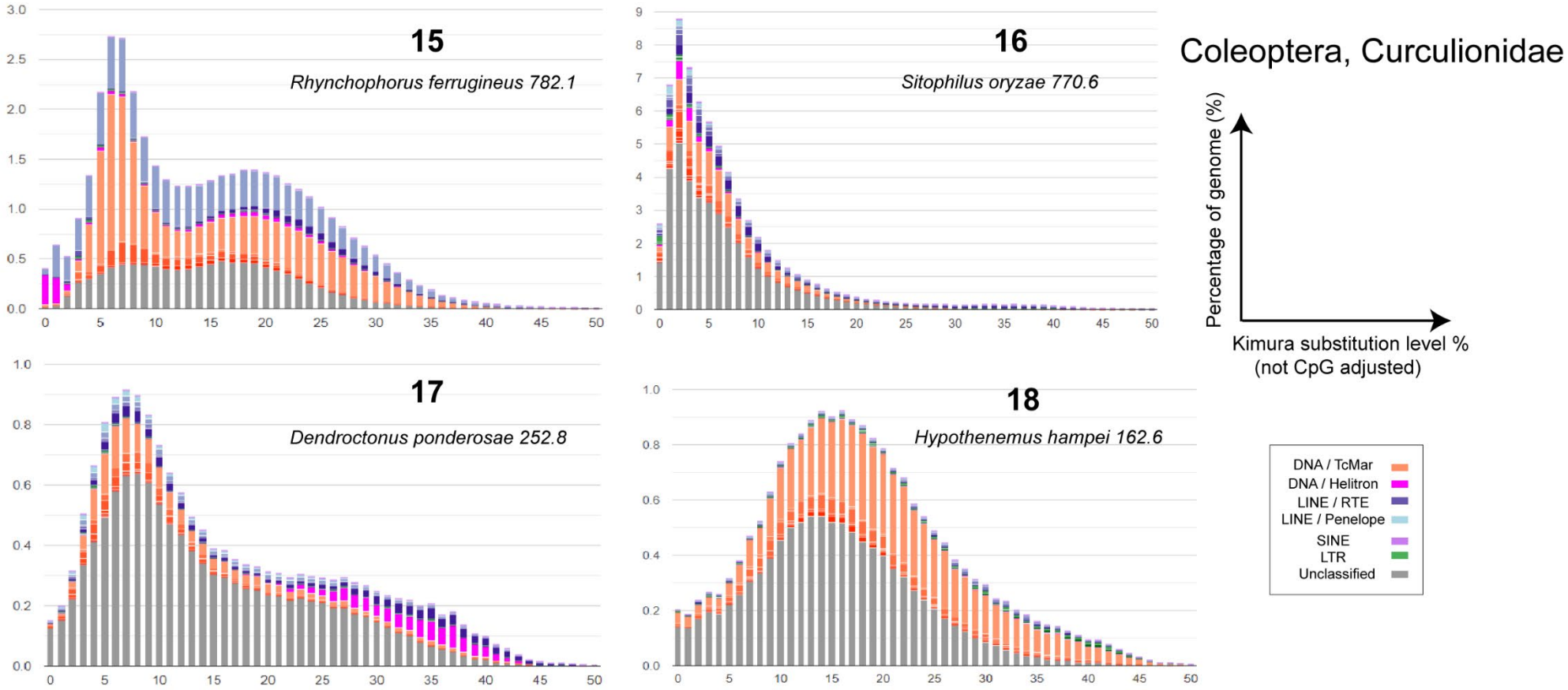

Figure S13. TE divergency landscapes of beetles in Family Curculionidae, Order Coleoptera. Related to Figure 4.

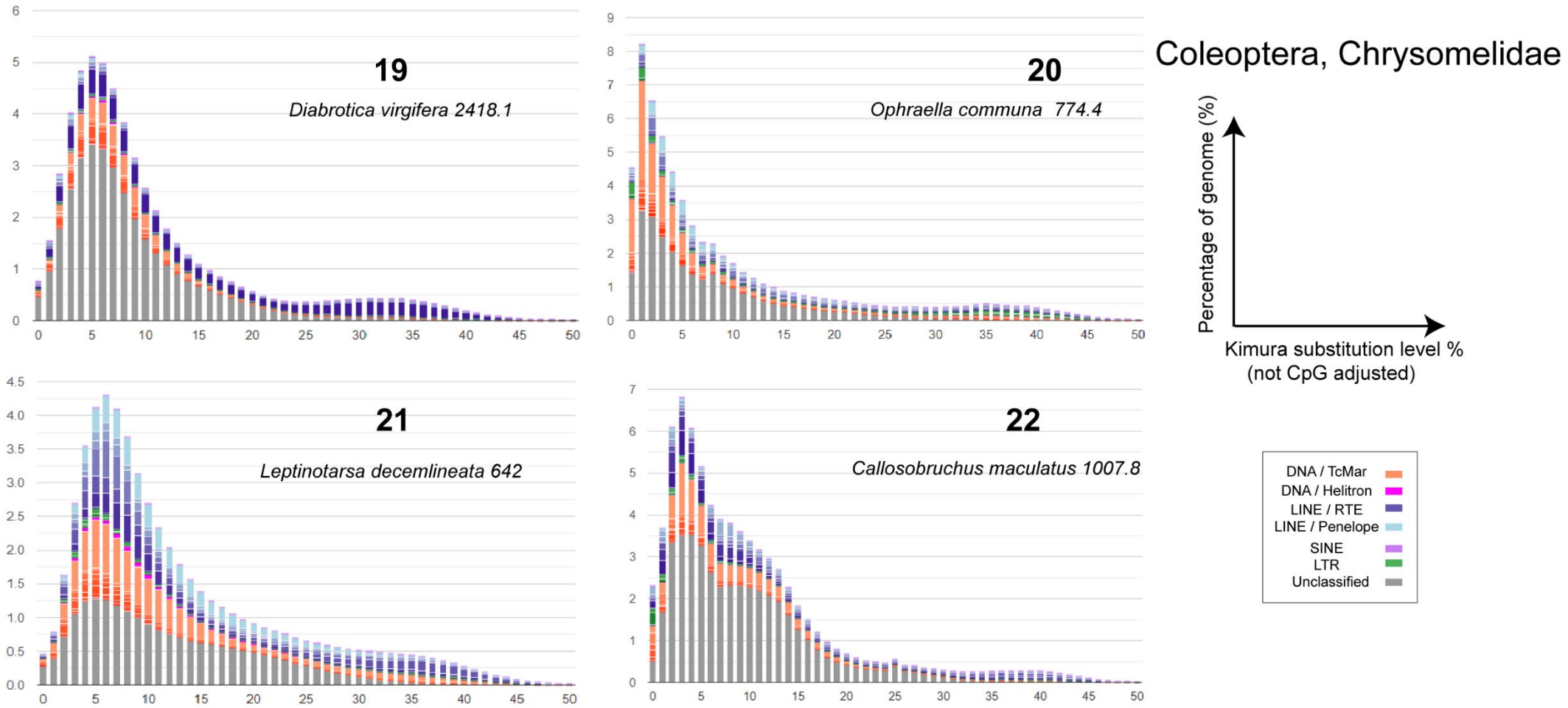

Figure S14. TE divergency landscapes of beetles in Family Chrysomelidae, Order Coleoptera. Related to Figure 4.

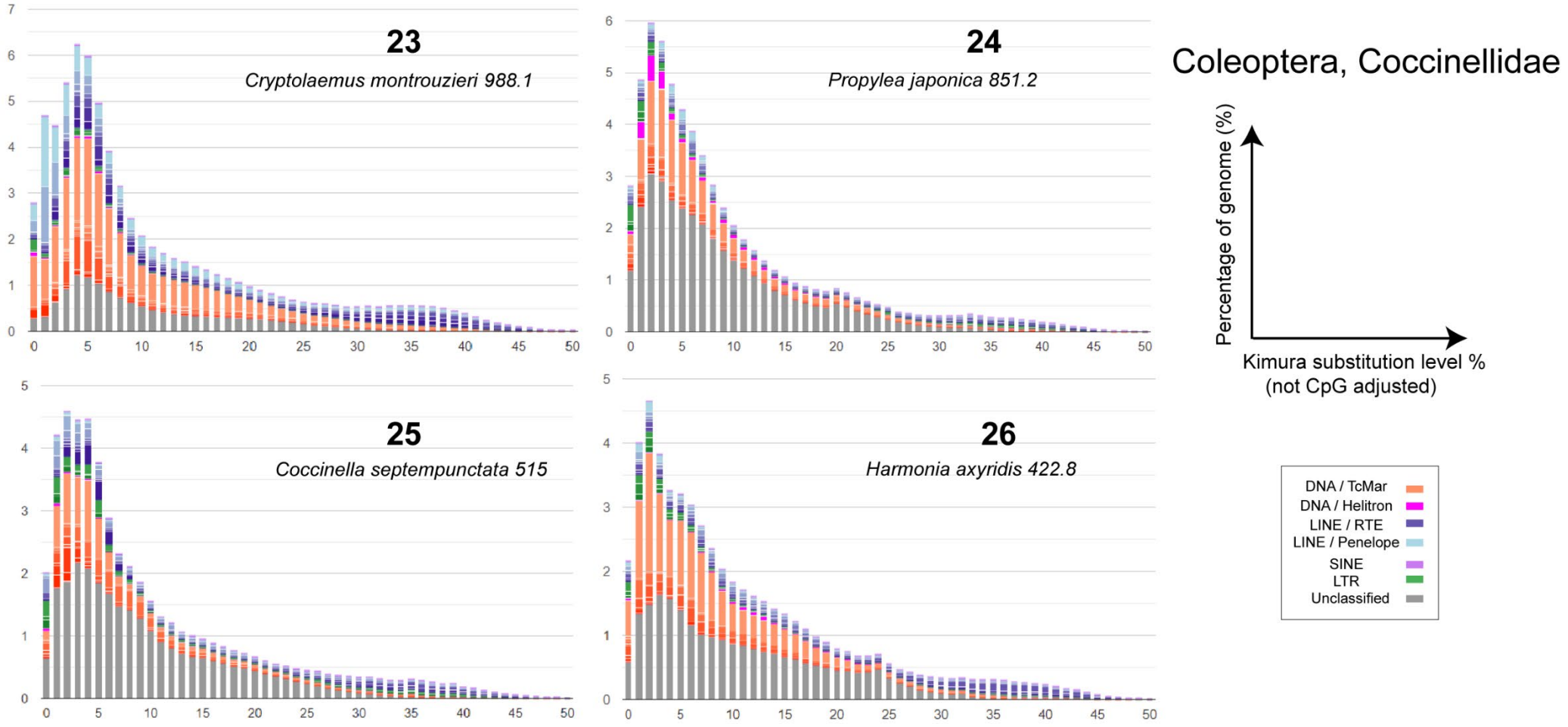

Figure S15. TE divergency landscapes of beetles in Family Coccinellidae, Order Coleoptera. Related to Figure 4.