Multiple large-scale gene and genome duplications during the evolution of hexapods

Zheng Li*[1], George P. Tiley*[2,3], Sally R. Galuska[1], Chris R. Reardon[1], Thomas I. Kidder[1], Rebecca J. Rundell[1,4], Michael S. Barker[1]

Affiliations:

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721

[2]Department of Biology, University of Florida, Gainesville, FL 32611

[3]Department of Biology, Duke University, Durham, NC 27708

[4]Department of Environmental and Forest Biology, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210

Corresponding author: msbarker@email.arizona.edu

* Authors contributed equally

Keywords: polyploidy, whole genome duplication, large-scale genome duplication, genome evolution, hexapods

Abstract

Polyploidy or whole genome duplication (WGD) is a major contributor to genome evolution and diversity. Although polyploidy is recognized as an important component of plant evolution, it is generally considered to play a relatively minor role in animal evolution. Ancient polyploidy is found in the ancestry of some animals, especially fishes, but there is little evidence for ancient WGDs in other metazoan lineages. Here we use recently published transcriptomes and genomes from more than 150 species across the insect phylogeny to investigate whether ancient WGDs occurred during the evolution of Hexapoda, the most diverse clade of animals. Using gene age distributions and phylogenomics, we found evidence for 18 ancient WGDs and six other large-scale bursts of gene duplication during insect evolution. These bursts of gene duplication occurred in the history of lineages such as the Lepidoptera, Trichoptera, and Odonata. To further corroborate the nature of these duplications, we evaluated the pattern of gene retention from putative WGDs observed in the gene age distributions. We found a relatively strong signal of convergent gene retention across many of the putative insect WGDs. Considering the phylogenetic breadth and depth of the insect phylogeny, this observation is consistent with polyploidy as we expect dosage-balance to drive the parallel retention of genes. Together with recent research on plant evolution, our hexapod results suggest that genome duplications contributed to the evolution of two of the most diverse lineages of eukaryotes on Earth.

Introduction

Polyploidy has long been considered a major force of genome evolution and a generator of diversity. Evidence of paleopolyploidy is found in the genomes of many eukaryotes, such as yeasts, teleost fishes, and plants (Wolfe and Shields 1997; Berthelot et al. 2014; Barker, Husband, et al. 2016; Van de Peer et al. 2017). Speciation by genome duplication is perhaps most important among plants where nearly ⅓ of contemporary vascular plant species have recently duplicated genomes (Wood et al. 2009; Barker,

Arrigo, et al. 2016). All extant seed plants have also experienced at least one ancient WGD (Jiao et al. 2011; Li et al. 2015), and many flowering plants have undergone multiple rounds of paleopolyploidy (Wendel 2015; Barker, Husband, et al. 2016; Van de Peer et al. 2017). The creation of new genes (Edger et al. 2015; Gout and Lynch 2015), higher turnover of genome content (Arrigo and Barker 2012; Murat et al. 2017), and increased rates of adaptation (Selmecki et al. 2015) following polyploidy have likely contributed to the diversification of flowering plants (Arrigo and Barker 2012; Tank et al. 2015).

In contrast to plants, polyploid speciation among animals is generally regarded as exceptional (Orr 1990; Otto and Whitton 2000). The most well known polyploidization events in animals are two rounds of ancient WGD (the 2R hypothesis) that occurred in the ancestry of all vertebrates (Ohno and Susumu 1970; McLysaght et al. 2002). However, most known cases of polyploidy in animals are found among parthenogenetic and hermaphroditic groups (Gregory and Mable 2005). If paleopolyploidy is indeed fundamental to the evolution of animal life across deep time, as it is in plants, we would expect to find WGDs throughout the most species-rich animal lineages: molluscs and arthropods. Little is known about ancient WGD among invertebrates, but there is growing evidence for paleopolyploidy in molluscs (Hallinan and Lindberg 2011; Yoshida et al. 2011) and chelicerates (Nossa et al. 2014; Clarke et al. 2015; Schwager et al. 2017). There is no evidence of paleopolyploidy among Hexapoda, the most diverse lineage of animals on Earth. Only 0.01% of the more than 800,000 described hexapod species (De Wever A. eds. 2015) are known polyploids (Otto and Whitton 2000; Gregory and Mable 2005). However, until recently there were limited data available to search for evidence of paleopolyploidy among the hexapods and other animal clades. Thus, the contributions of polyploidy to animal evolution and the differences with plant evolution have remained unclear.

To search for evidence of WGDs among the hexapods, we leveraged recently released genomic data for the insects (Misof et al. 2014). Combined with additional data sets from public databases, we assembled 128 transcriptomes and 27 genomes with at

least one representative from each order of Hexapoda (SI Appendix, Dataset S1). We selected data from chelicerates, myriapods, and crustaceans as outgroups. Ancient WGDs were initially identified in the distributions of gene ages (Ks plots) produced by our DupPipe pipeline (Barker et al. 2008; Barker et al. 2010). We also used the MAPS algorithm (Li et al. 2015) to infer and locate WGDs or other large-scale genome duplications that are shared among descendant taxa. MAPS uses gene trees to infer the phylogenetic placement of ancient bursts of gene duplication, including those where evolutionary distances are too saturated to appear in Ks plots. We simulated gene family phylogenies with and without WGDs to identify significant bursts of large-scale genome duplication. Analyses of synteny within the *Bombyx mori* genome provided additional evidence that genome duplications inferred by our MAPS analyses result from large-scale genome duplication events. Together with MAPS, the orthologous divergence among lineages was used to place inferred genome duplications in phylogenetic context. Potential ancient WGDs detected in our gene age distributions were further corroborated by analyses of biased gene retention across 20 hexapod genomes.

Results

Inference of WGDs from Gene Age Distributions

Our phylogenomic analyses revealed that insects have a broad ancestry of polyploidy. Peaks of gene duplication consistent with WGDs were observed in the gene age distributions of 20 hexapod species (Fig. 1, SI Appendix, Fig. S1-S4, and Table S1). Each of the inferred WGDs was identified as a significant peak using SiZer and mixture model analyses (SI Appendix, Table S1, Dataset S2-S3). Fifteen of these appear as independent WGDs in our analyses because the sampled sister lineages lack evidence of the duplications (Fig. 2, SI Appendix, Fig. S1-S4). In two cases, multiple sister lineages contained evidence of paleopolyploidy in their Ks plots. All sampled species of Thysanoptera contained evidence of at least one peak consistent with paleopolyploidy in their Ks plots (Fig. 1*B*, SI Appendix, Fig. S2*I*-*K*). Analyses of orthologous divergence

among these taxa indicated that the putative WGD peaks are older than the divergence of these lineages, and we currently infer a single, shared WGD in the ancestry of Thysanoptera. Similarly, multiple taxa in the Trichoptera had evidence of WGD(s) in their Ks plots (SI Appendix, Fig. S1*O*-*R*, and Fig. S2*O*-*R*). Analyses of orthologous divergence indicated that each of these putative WGDs occurred independently (SI Appendix, Fig. S5*O* and *P*, and Table S2). A MAPS analysis also supported the independence of these WGDs and found evidence for a deeper duplication event shared among all the sampled Trichoptera (SI Appendix, Fig. S6*Y* and Table S3).

Overall, our analyses of gene age distributions found evidence for 18 independent paleopolyploidizations in the ancestry of 14 orders of hexapods (Fig. 2, SI Appendix, Fig. S7). We observed evidence for ancient WGDs in diverse lineages of hexapods including springtails, beetles, ants, lice, flies, thrips, moths, termites, sawflies, caddisflies, stoneflies, and mayflies. Some of these WGDs were of relatively modest synonymous divergence and may be correlated with the origins of families or clades of genera, such as the inferred paleopolyploidization in *Trichocera saltator* (Fig. 1*D*). However, many of these putative WGDs appear to have occurred early in the evolution of different hexapod orders with relatively high synonymous divergence among paralogs. For example, applying the *Drosophila* synonymous substitution rate of 5.8×10^{-9} substitutions/synonymous site/year (Cutter 2008) to thrips, we estimated that the thrips duplication occurred approximately 155 MYA based on the median paralog divergence of the WGD. However, if thrips have a slower rate of evolution than *Drosophila*, then this WGD would be older.

Phylogenomic Inference and Simulation of Ancient Large-scale Genome Duplications

Given the depth of the phylogeny, there may be many WGDs or other large-scale genome duplications in the ancestry of hexapods that do not appear in Ks plots due to saturation of substitutions. We implemented 33 MAPS (Li et al. 2015) analyses of 111,933 nuclear gene family phylogenies to infer large bursts of gene duplication deep in the history of all major clades of hexapods (SI Appendix, Fig. S6 and Dataset S4). By

examining evidence from multiple species, MAPS increases the signal of deep duplications and provides more resolution than a single species analysis. Overall, we found 25 branches in 22 MAPS analyses that contained at least one branch with significantly more shared gene duplications than expected compared to the null simulations (SI Appendix, Fig. S8 and Dataset S4). To further characterize these significant bursts of gene duplication, we simulated an additional set of gene trees with a WGD at the phylogenetic location of the duplication bursts in these 22 analyses. We found that 14 of the 25 bursts of gene duplication were statistically consistent with our positive simulations of WGDs (SI Appendix, Fig. S9 and Dataset S5). Six of these large-scale genome duplications had robust and consistent evidence across all of our MAPS analyses (Fig. 2, SI Appendix, Fig. S6, Table S3 and Dataset S4-S5). These included large-scale genome duplications in the ancestry of Odonota, Lepidoptera, and Trichoptera. We also observed seven episodic bursts of gene duplication in the insect phylogeny that had varying levels of significance across different MAPS analyses (eg. Coleoptera, Hymenoptera (in part) and Hemiptera (in part); Fig. 2, SI Appendix, Fig. S6-7 and Dataset S5). Although there was conflict among our analyses as to whether these seven events were statistically consistent with WGDs or other large-scale genome duplications, they do reflect significant increases in gene duplication at these locations in the insect phylogeny.

Considering the putative ancient nature of the MAPS inferred duplications, most are likely too saturated to appear in any of our Ks plots. Saturation of substitutions diminishes the signature peaks of polyploidy in Ks plots, and WGDs with a peak of Ks > 2 may become difficult to detect (Cui et al. 2006; Barker et al. 2008; Vanneste et al. 2013). To confirm that we should not expect to see these six large-scale genome duplications in Ks plots, we tested if the ortholog divergence among these lineages was Ks > 2. As expected, the median orthologs divergence within these MAPS analyses was Ks > 2 (SI Appendix, Fig. S5 and Table S2). Thus, all of the six large-scale genome duplications are likely too saturated to appear in Ks plots.

Synteny provides perhaps the most compelling evidence to characterize ancient genome duplication events. Although there are a number of high quality hexapod genomes, there are few high quality genomes in the clades where we inferred ancient duplications. The genome of *Bombyx mori (International Silkworm Genome Consortium 2008)*, the silkworm moth, is one of the few genomes that is reasonably well assembled and has an ancient genome duplication based on our MAPS analyses (SI Appendix, Figs. S8*AA* and S9*AA*, Dataset S6). If the duplication inferred by MAPS is a large-scale genome duplication event, we expect to find a significant association (Nakazato et al. 2006) of syntenic chains with the MAPS paralogs in the genome of *B. mori*. We identified 728 syntenic chains that included 2210 genes (SI Appendix, Fig. S11). Significantly more syntenic chains—83 chains—were associated with our MAPS paralogs than expected by chance (chi-square test p-value = 0.0001). The mean Ks between all pairs of syntenic blocks that included MAPS paralogs was highly saturated (>45), indicating that these duplicated blocks are ancient (SI Appendix, Fig. S12 and Dataset S7). Thus, syntenic evidence is consistent with the large-scale genome duplication inferred in the ancestry of Lepidoptera with MAPS.

Biased Gene Retention and Loss Following Inferred WGDs

Given the rarity of polyploid speciation among insects and the fact that no ancient WGDs have been previously observed in the clade, we sought to further characterize the nature of these putative WGDs. A common signature of paleopolyploidy is the biased retention and loss of some categories of genes relative to the background pattern of gene turnover. Surviving paralogs from ancient WGDs are often enriched with idiosyncratic gene ontology (GO) categories in plants (Barker et al. 2008; De Smet et al. 2013; Li et al. 2016; Mandáková et al. 2017; Rody et al. 2017), yeasts (Papp et al. 2003; Conant 2014), and animals (Berthelot et al. 2014; Lien et al. 2016; Session et al. 2016). Among the many hypotheses to explain the retention of paralogs (Kondrashov and Kondrashov 2006; Freeling 2009; Hahn 2009), the dosage balance hypothesis (DBH) is the only hypothesis that predicts the parallel retention and loss of

functionally related genes following WGDs across species (Freeling and Thomas 2006; Freeling 2009; Conant et al. 2014). The DBH predicts that genes with many connections or in dosage sensitive regulatory networks will be retained in duplicate following polyploidy to maintain the relative abundance of protein products. Conversely, these same genes will be preferentially lost following small-scale gene duplications to prevent the disruption of dosage (Freeling 2009; Birchler and Veitia 2012). Thus, if the signatures of gene duplication observed in the insects are the result of WGDs, then we expect to find a biased pattern of retention and loss that may be shared among the putative insect WGDs.

To test for biased gene retention and loss among our inferred WGDs, we used a HMMR based approach to annotate the genomes or transcriptomes of 20 hexapod and one outgroup species with the *Drosophila melanogaster* Gene Ontology (GO) data (Gene Ontology Consortium 2015). Paralogs were partitioned from the putative WGDs inferred in Ks plots by fitting a normal mixture model to the distributions (SI Appendix, Table S5). Using the numbers of genes annotated to each GO category, we performed a principal component analysis to assess the overall differences in GO category composition between all genes in the genome/transcriptome and paralogs retained from each putative WGD (Fig. 3, SI Appendix, Dataset S8). Significant differences ($p < 0.001$) between the GO composition of these two groups were found using a goodness of fit test. They also formed distinct clusters in the PCA with most data points in their 95% confidence intervals. Notably, the GO composition of WGD paralogs from each species formed a narrower confidence interval than the entire transcriptomes/genomes. We further identified GO categories that were significantly over- or under-represented among the paralogs from each WGD using a chi-square test. Our analyses found that paralogs from WGDs across all 21 species demonstrated biased patterns of gene retention and loss (Fig. 3, SI Appendix, Fig. S13-S14). Consistent with our PCA, many of the same GO categories were significantly over- and under-represented among the genes maintained in duplicate from the putative insect WGDs. For example, over 65% of our sampled species had paralogs from the ancient WGDs that were significantly

enriched for RNA metabolic processes (SI Appendix, Fig. S13). More than 50% of the WGD paralogs in the sampled species also showed significant over-retention of genes associated with the nucleus component, DNA binding function, and nucleobase-containing compound metabolic process. Similarly, many GO categories were significantly under-retained among the paralogs surviving from ancient WGDs (SI Appendix, Fig. S13). Over half of our sampled species demonstrated significant under-retention of genes associated with oxidoreductase activity, hydrolase activity, electron carrier activity, peptidase activity, and proteolysis. Although there is some noise in the patterns of over- and under-retention, this may not be surprising given the great divergence of most of these species from the annotation source, *Drosophila*, and the phylogenetic scale of the hexapods. Considering the great diversity of these lineages, the convergent pattern of gene retention and loss is consistent with our expectations following polyploidy.

Discussion

Our analyses provide the first evidence of paleopolyploidy in the hexapods. Combining our gene age distribution and MAPS analyses, we found evidence for 24 significant, episodic bursts of gene duplication in the insects. Of these bursts, 18 were detected as peaks in gene age distributions that are characteristic of ancient WGDs in 14 hexapod orders (Fig. 2, SI Appendix, Table S3). An additional six large-scale genome duplication events that may be ancient WGDs were inferred using phylogenomic analyses. Syntenic evidence from the *B. mori* genome also supported our inference of one of these events, a large-scale genome duplication in the ancestry of the Lepidoptera. These 24 large-scale genome duplications occurred in the ancestry of 15 orders of hexapods. Based on the phylogenetic position of the inferred duplication events, we estimate that approximately 17% of described hexapod species (De Wever A. eds. 2015) may have experienced at least one ancient genome duplication. Given that hexapods are the most diverse clade of animals, our results suggest that at least 16% of described animal species have experienced a genome duplication event. Together with evidence of

paleopolyploidy throughout the evolution of vascular plants (Jiao et al. 2011; Li et al. 2015; Barker, Husband, et al. 2016) and vertebrates (Otto and Whitton 2000), these results suggest that at least 28% of all described species on Earth (De Wever A. eds. 2015) may have a polyploid ancestry.

Although our results are largely unanticipated, there is previous evidence for polyploidy among the insects. Analyses of genome sizes have found evidence for recent polyploidy among the thrips (Jacobson et al. 2013). Closely related species of lepidoptera sometimes have large differences in the number of chromosomes (Orr 1990; White) that may be explained by polyploidy (Lukhtanov and Puplesiene 1999). Polyploidy has also been described among parthenogenic and hermaphroditic insects (Gregory and Mable 2005). Thus, it is not impossible for insects to experience WGD. Our inference of ancient WGDs among insects suggests that the number of contemporary polyploid species may be higher than currently appreciated. Even among plants where polyploidy is relatively well studied, estimates indicate that nearly 50% of contemporary plant polyploids are unnamed and there may be more than 50,000 cryptic flowering plant species (Barker, Arrigo, et al. 2016). Given the number of insect species, there could be many more undescribed or unrecognized polyploid species. A better understanding of contemporary polyploidy in insects would provide an improved context for understanding the expected frequency and consequences of the ancient WGDs and other large-scale genome duplications we inferred.

Our analyses of GO categories revealed a largely convergent pattern of biased gene retention and loss following paleopolyploidy in multiple species of insects. This biased pattern was also found among WGD paralogs in an outgroup species, *Ixodes scapularis* (deer tick), and suggests that it may be consistent across arthropods. Previous studies in plants have found that similar functional categories of genes have been maintained following different genome duplication events (Li et al. 2016; Rody et al. 2017), although there are often idiosyncratic patterns observed across families (Maere et al. 2005; Barker et al. 2008; Li et al. 2016; Mandáková et al. 2017). The 20 insect transcriptomes included in our GO category analyses represent diverse hexapod (SI

Appendix, Fig. S13) orders whose divergence times far exceed most previously studied plant examples. At least among the arthropods, these analyses suggest that biases in duplicate gene retention can be maintained over hundreds of millions of years. One explanation for the long consistency of this pattern is that insects have experienced a limited number of ancient WGDs that may influence gene retention patterns or large-scale shifts in gene network relationships. Most hexapod species included in our analyses only had one round of paleopolyploidy, whereas nearly all flowering plants have likely experienced at least three rounds of ancient polyploidization (Barker, Husband, et al. 2016).

The analyses of GO category retention and loss are consistent with our expectations of genome evolution following polyploidy rather than other types of gene duplication events. Based on the DBH, we expected to observe biased retention and loss of similar GO categories following WGDs. Our observation across multiple species is consistent with post-polyploid genome evolution as dosage-balance may drive the convergent retention of genes. Previous empirical studies supported that gene dosage balance plays an important role in gene retention following gene and genome duplications (Freeling 2009; Session et al. 2016; Mandáková et al. 2017). Other hypotheses of duplicate gene retention such as neo- and subfunctionalization are also important in duplicate gene retention, but only the DBH predicts a convergent pattern of gene retention following polyploidy across multiple species (Kondrashov and Kondrashov 2006; Freeling 2009). Thus, the signal of convergent gene retention across many of the duplication events is consistent with our inference that many were likely WGDs.

Our discovery of ancient polyploidy and other large-scale genome duplications throughout the history of insects and their close relatives raises a number of questions about the evolution of their diversity. Our results raise the possibility that gene and genome duplications could be associated with major transitions in insect evolution. The route to such profound transitions might involve the introduction of novelty via duplication, as has been shown in the expanding *Hox* genes of some Lepidoptera

(Ferguson et al. 2014). Our results lend support to long-standing hypotheses that gene and genome duplications are important forces in animal evolution (Ohno and Susumu 1970; Van de Peer et al. 2009). As in other groups of organisms, further analyses are needed to test if hexapod duplication events are associated with increases in net diversification rates, and to determine how genome duplications contributed to insect diversity. Recent research suggests that gene and genome duplications drove the co-evolutionary radiation of pierid butterflies and the Brassicales (Edger et al. 2015). Our observation of numerous episodes of duplication in the hexapod phylogeny raises the intriguing possibility of additional duplication driven co-evolutionary interactions with plants. Further phylogenomic sampling of insects will likely reveal more paleopolyploidy and other large-scale genome duplications. Indeed, given our sample size relative to the total number of hexapod species, the incidence and evolutionary significance of paleopolyploidy is likely to be underestimated. Large sequencing projects such as 1KP (Wickett et al. 2014), 1KITE (Misof et al. 2014), and i5K (i5K Consortium 2013) will improve our ability to place these events in the plant and hexapod phylogenies, both of which are replete with rapid and prolific diversification. As it stands, our results raise the intriguing possibility that gene and genome duplications may explain both Darwin's "abominable mystery" and Haldane's "inordinate fondness."

Materials and Methods

Data sampling

We compiled a phylogenetically diverse genomic dataset that comprised every hexapod order and outgroups from related arthropods (SI Appendix, Dataset S1). These data included 119 transcriptomes and 25 genomes for hexapods, as well as nine transcriptomes and two genomes from the Chelicerates, myriapods, and crustaceans as outgroups. We downloaded 128 published transcriptome assemblies from the GenBank Transcriptome Shotgun Assembly database (TSA), and 27 published genomes from multiple genome databases (SI Appendix).

DupPipe: single species inference of WGDs from paralog age distributions

For each data set, we used our DupPipe pipeline to construct gene families and estimate the age of gene duplications (Barker et al. 2010). We translated DNA sequences and identified reading frames by comparing the Genewise alignment to the best hit protein from a collection of proteins from 24 metazoan genomes from Metazome v3.0. For each node in our gene family phylogenies, we estimated synonymous divergence (Ks) using PAML with the F3X4 model (Yang 1997). For each species, we identified ancient WGDs as significant peaks of gene duplication in histograms of the age distribution of gene duplications (Ks plots) using mixture models (McLachlan and Peel 1999) and SiZer (Chaudhuri and Marron 1999).

MAPS: phylogenomic inference and placement of large-scale genome duplications from nuclear gene trees

To infer large-scale genome duplications, we used the Multi-tAxon Paleopolyploidy Search (MAPS) tool (Li et al. 2015), a gene tree sorting and counting algorithm. We circumscribed and constructed nuclear gene family phylogenies from multiple species for each MAPS analysis. Details of all 33 MAPS analyses are provided in Dataset S4. We translated each transcriptome into amino acid sequences using the TransPipe pipeline (Barker et al. 2010). Using these translations, we performed reciprocal BLASTP searches for each MAPS data set with an E-value of 10e-5 as a cutoff. We clustered gene families from these BLAST results using OrthoMCL v2.0 with the default parameters (Li et al. 2003) and only retained gene families that contained at least one gene copy from each taxon. We used SATé for alignment and phylogeny reconstruction of gene families (Liu et al. 2009). For each gene family, we ran SATé until we reached five iterations without an improvement in the likelihood score using a centroid breaking strategy. We constructed alignments using MAFFT (Katoh et al. 2002), employed Opal for mergers (Wheeler and Kececioglu 2007), and RAxML for tree estimation (Stamatakis 2014). We used the best scoring SATé tree for each gene family to infer large-scale genome duplications with MAPS.

Syntenic analysis of **Bombyx mori**

To validate our MAPS inferences of large-scale genome duplications in the insects, we examined the the *Bombyx mori* genome (International Silkworm Genome Consortium 2008) for syntenic evidence of an ancient large scale gene duplication event inferred in Lepidoptera (SI Appendix, Figs. S6*AA*-S9*AA*). The SynMap tool on the CoGe platform (Lyons et al. 2008) was used to identify syntenic regions. We used blastp with E-value cutoff of 10e-5. Quota Align Merge was selected to merge syntenic blocks and other parameters were set as default. Synteny was detected with a minimum of three genes to seed a chain and a Manhattan distance of 40. The divergence of syntenic blocks was estimated using the mean Ks for paralogs on the syntenic blocks provided in SynMap.

Estimating orthologous divergence to place large-scale genome duplications in relation to lineage divergence

We estimated ortholog divergences among major hexapod clades to place large-scale genome duplications in relation to lineage divergence (SI Appendix, Table S2). We used the RBH Ortholog pipeline (Barker et al. 2010) to estimate the median ortholog divergence between 24 species pairs (SI Appendix, Fig. S5 and Table S2). The median ortholog divergence was used to estimate the lower bound of paralog divergence for shared ancient large-scale genome duplication events.

Gene Ontology (GO) annotations and paleolog retention and lost patterns

We used the best hit with length of at least 100 bp and an E-value of at least 0.01 in phmmer (HMMER 3.1b1) for Gene Ontology (GO) annotation (Finn et al. 2015). For each species, we assigned paralogs to ancient WGDs based on the Ks ranges identified in mixture model analyses (SI Appendix, Table S5) (McLachlan and Peel 1999; Barker et al. 2008). We evaluated the overall differences between the genome or whole transcriptomes and WGD paralogs by performing a principal component analysis (PCA) using the rda function in R package vegan (Dixon 2003). We also tested for differences

among GO annotations across the inferred WGD events using chi-square tests (Team 2014) (SI Appendix, Fig. S13-S14).

Acknowledgments

We thank A. Baltzell, A. Baniaga, J. Bronstein, K. Dlugosch, S. Jorgensen, E. Lyons, X. Qi, J. Czekanski-Moir and Y. Tomoyasu for suggestions and discussion. Z. Wang, L. Tang, and X. Song for assistance with insect images. We also thank two anonymous reviewers for their careful reading and suggestions that improved the manuscript. Hosting infrastructure and services provided by the Biotechnology Computing Facility (BCF) at the University of Arizona. M.S.B. was supported by NSF-IOS-1339156 and NSF-EF-1550838.

References

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. 15:140–146.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytol. 210:391–398.

Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH. 2010. EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. Evol. Bioinform. Online 6:143–149.

Barker MS, Husband BC, Pires JC. 2016. Spreading Winge and flying high: The evolutionary importance of polyploidy after a century of study. Am. J. Bot. 103:1139–1145.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25:2445–2455.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5:3657.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences 109:14746–14753.

Chaudhuri P, Marron JS. 1999. SiZer for Exploration of Structures in Curves. J. Am. Stat. Assoc. 94:807–823.

Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. 2015. Spider Transcriptomes Identify Ancient Large-Scale Gene Duplication Event Potentially Important in Silk Gland Evolution. Genome Biol. Evol. [Internet]. Available from: http://gbe.oxfordjournals.org/content/early/2015/06/08/gbe.evv110.abstract

Conant GC. 2014. Comparative genomics as a time machine: how relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. Mol. Biol. Evol. 31:3184–3193.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. 19:91–98.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Cutter AD. 2008. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol. Biol. Evol. 25:778–786.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. U. S. A. 110:2898–2903.

De Wever A. eds. RYALOTNDKTCABNKPBTDREDW. 2015. Species 2000 & ITIS Catalogue of Life, 2015 Annual Checklist. Catalogue of Life [Internet]. Available from: www.catalogueoflife.org/annual-checklist/2015

Dixon P. 2003. VEGAN, a package of R functions for community ecology. J. Veg. Sci. 14:927.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. U. S. A. 112:8362–8366.

Ferguson L, Marlétaz F, Carter J-M, Taylor WR, Gibbs M, Breuker CJ, Holland PWH. 2014. Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. PLoS Genet. 10:e1004698.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. Nucleic Acids Res. 43:W30–W38.

Freeling M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. Annu. Rev. Plant Biol. 60:433–453.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43:D1049–D1056.

Gout J-F, Lynch M. 2015. Maintenance and loss of duplicated genes by dosage subfunctionalization. Mol. Biol. Evol. [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msv095

Gregory TR, Mable BK. 2005. Polyploidy in animals. In: Vol. 171. Elsevier Academic. p. 427–517.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J. Hered. 100:605–617.

Hallinan NM, Lindberg DR. 2011. Comparative Analysis of Chromosome Counts Infers Three Paleopolyploidies in the Mollusca. Genome Biol. Evol. 3:1150–1163.

i5K Consortium. 2013. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. J. Hered. 104:595–600.

International Silkworm Genome Consortium. 2008. The genome of a lepidopteran model insect, the silkworm Bombyx mori. Insect Biochem. Mol. Biol. 38:1036–1045.

Jacobson AL, Johnston JS, Rotenberg D, Whitfield AE, Booth W, Vargo EL, Kennedy GG. 2013. Genome size and ploidy of Thysanoptera. Insect Mol. Biol. 22:12–17.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

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

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. J. Theor. Biol. 239:141–151.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. Science 324:1561–1564.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. Science Advances 1(10):e1501084

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. Plant Cell 28:326–344.

Lukhtanov V, Puplesiene J. 1999. Polyploidy in bisexual Lepidoptera species (Insecta: Lepidoptera): old hypotheses and new data. Bonn. Zool. Beitr. 48:313–328.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Trop. Plant Biol. 1:181–190.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 102:5454–5459.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. [Internet]. Available from: http://dx.doi.org/10.1111/tpj.13553

McLachlan GJ, Peel D. 1999. The EMMIX Algorithm for the Fitting of Normal andt-Components. J. Stat. Softw. [Internet] 4. Available from: http://dx.doi.org/10.18637/jss.v004.i02

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31:200–204.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.

Murat F, Armero A, Pont C, Klopp C, Salse J. 2017. Reconstructing the genome of the

most recent common ancestor of flowering plants. Nat. Genet. [Internet]. Available from: http://dx.doi.org/10.1038/ng.3813

Nakazato T, Jung M-K, Housworth EA, Rieseberg LH, Gastony GJ. 2006. Genetic map-based analysis of genome structure in the homosporous fern Ceratopteris richardii. Genetics 173:1585–1597.

Nossa CW, Havlak P, Yue J-X, Lv J, Vincent KY, Brockmann H, Putnam NH. 2014. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. Giga Sci 3:9.

Ohno S, Susumu O. 1970. Evolution by Gene Duplication.

Orr HA. 1990. " Why Polyploidy is Rarer in Animals Than in Plants" Revisited. Am. Nat. 136:759–770.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.

Rody HVS, Baute GJ, Rieseberg LH, Oliveira LO. 2017. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. BMC Genomics 18:46.

Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, et al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol. 15:62.

Selmecki AM, Maruvka YE, Richmond PA, Guillet M, Shoresh N, Sorenson AL, De S, Kishony R, Michor F, Dowell R, et al. 2015. Polyploidy can drive rapid adaptation in yeast. Nature [Internet]. Available from: http://dx.doi.org/10.1038/nature14187

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. 2016. Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538:336–343.

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

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol. 207:454–467.

Team RC. 2014. R: A language and environment for statistical computing. Vienna,

Austria: R Foundation for Statistical Computing; 2013.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. 10:725–732.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. [Internet]. Available from: http://dx.doi.org/10.1038/nrg.2017.26

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. Mol. Biol. Evol. 30:177–190.

Wendel JF. 2015. The wondrous cycles of polyploidy in plants. Am. J. Bot. 102:1753–1756.

Wheeler TJ, Kececioglu JD. 2007. Multiple alignment by aligning alignments. Bioinformatics 23:i559–i568.

White M. JD, 1973 Animal Cytology and Evolution. Cambridge University

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U. S. A. 111:E4859–E4868.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. U. S. A. 106:13875–13879.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555–556.

Yoshida M-A, Ishikura Y, Moritaki T, Shoguchi E, Shimizu KK, Sese J, Ogura A. 2011. Genome structure analysis of molluscs revealed whole genome duplication and lineage specific repeat variation. Gene 483:63–71.

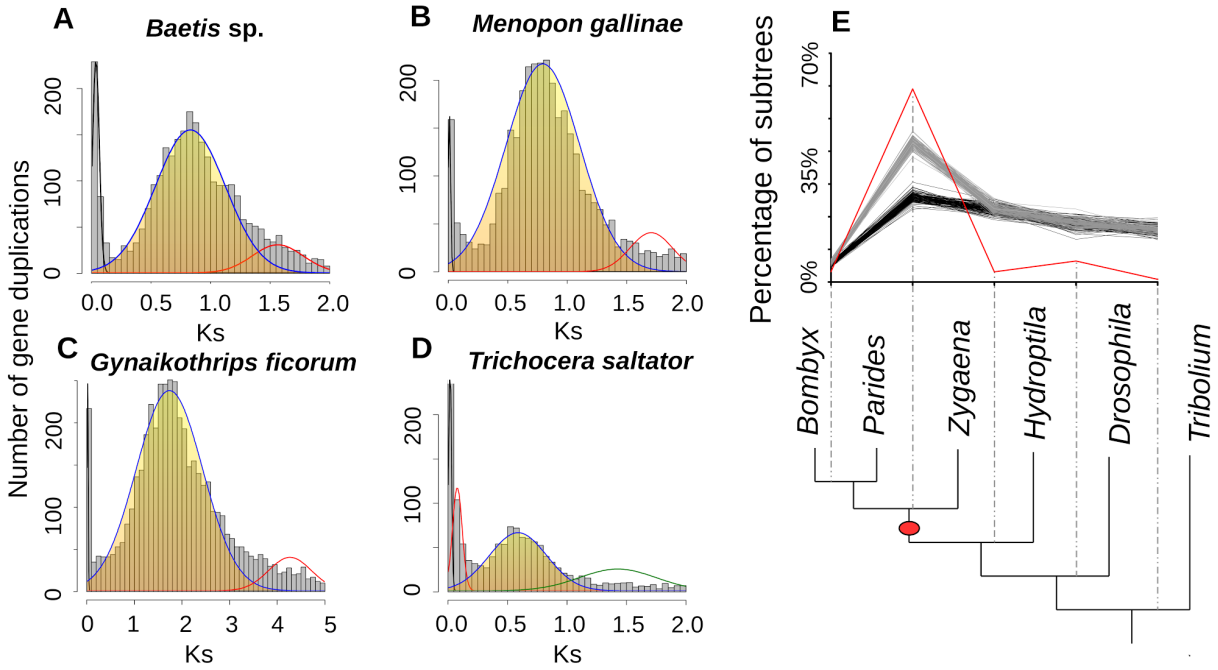

Fig. 1. Inferring ancient WGDs and large-scale genome duplications. Histograms of the age distribution of gene duplications (Ks plots) with mixture models of inferred WGDs for (**A**) *Baetis* sp. (Ephemeroptera); inferred WGD peak median Ks = 0.82. (**B**) *Gynaikothrips ficorum* (Thysanoptera); inferred WGD peak median Ks = 1.73. (**C**) *Menopon gallinae* (Psocodea); inferred WGD peak median Ks = 0.79. (**D**) *Trichocera saltator* (Diptera); inferred WGD peak median Ks = 0.59. The mixture model distributions consistent with inferred ancient WGDs are highlighted in yellow. (**E**) MAPS results from observed data and null simulations on the associated phylogeny. Percentage of subtrees that contain a gene duplication shared by descendant species at each node, results from observed data (red line) and 100 resampled sets of null simulations (multiple black lines). Red oval corresponds to the location of an inferred large-scale genome duplication event in Lepidoptera.

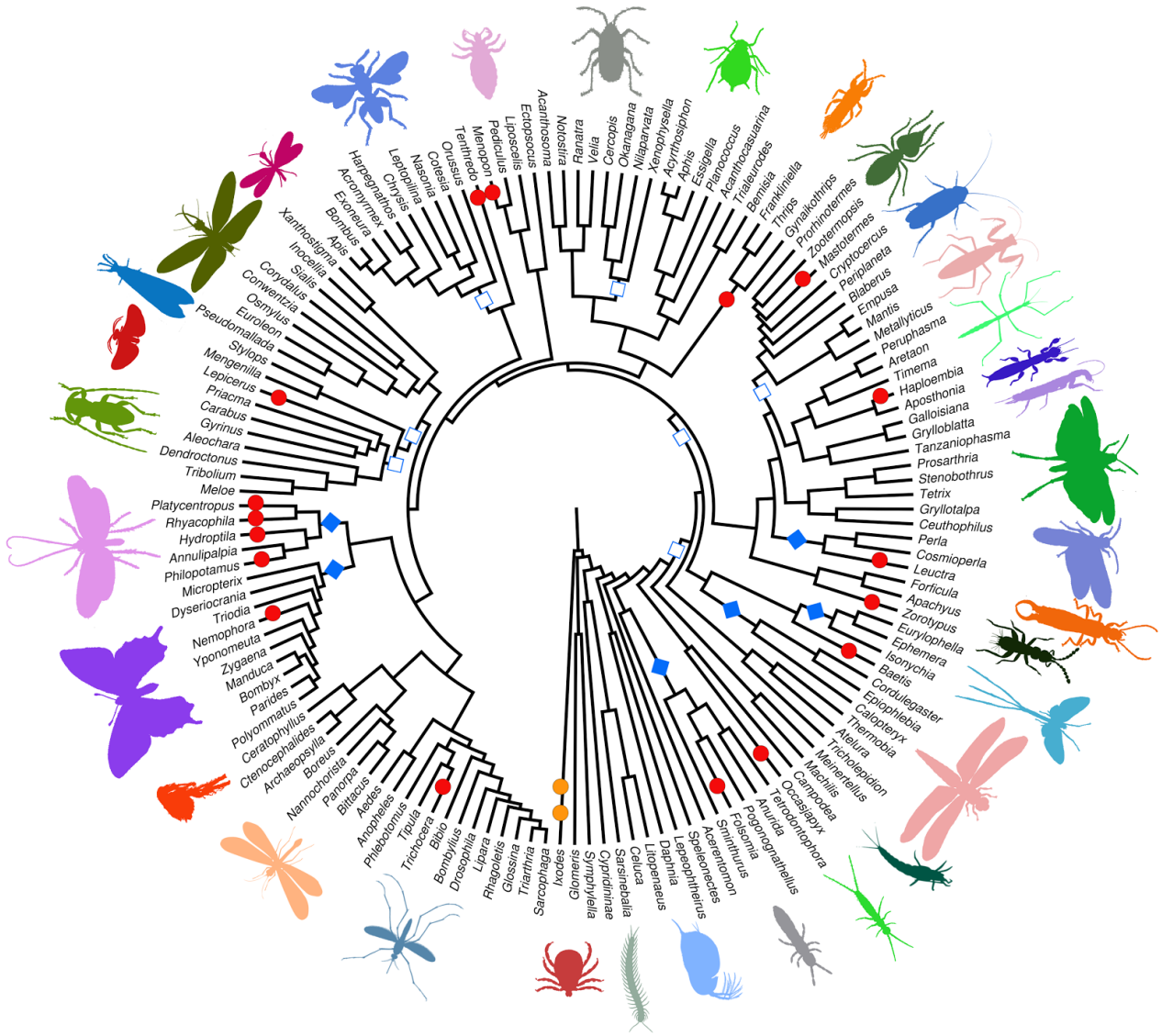

Fig. 2. Placement of inferred ancient genome duplications on the phylogeny of Hexapoda. Red circles = WGDs in hexapods inferred from Ks plots; Orange circles = WGDs in outgroups inferred from Ks plots; Blue diamonds = large-scale genome duplications inferred by MAPS analyses; Empty squares = episodic bursts of gene duplication with varying levels of significance across different MAPS analyses. Hexapod phylogeny adapted from Misof et al. 2014. The *Solenopsis invicta* (Hymenoptera) WGD inferred by Ks plot is not included on this phylogeny. Images of Raphidioptera, Coleoptera, and Neuroptera credit to Tang Liang, Zichen Wang and Zheng Li. Other images are in the public domain (Table S6).

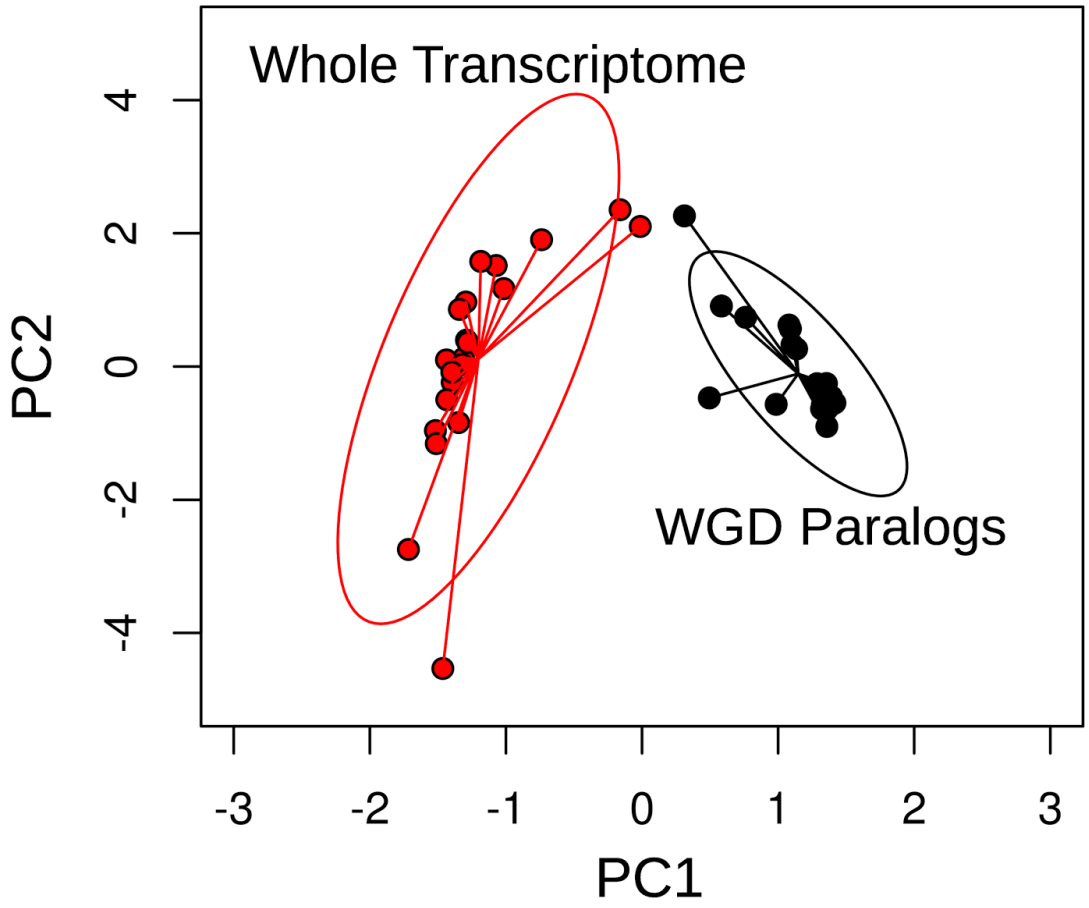

Fig. 3. Principal component analysis of the GO category composition of all genes in each genome/transcriptome and WGD paralogs. Red circles = number of genes annotated to each GO category in the whole genome or transcriptomes. Black circles = number of WGD paralogs annotated to each GO category. Ellipses represent the 95% confidence interval of standard deviation of point scores.