Genome size evolution in the diverse insect order Trichoptera

Authors

Jacqueline Heckenhauer[1,2*], Paul B. Frandsen[1,3,4], John S. Sproul[5], Zheng Li[6], Juraj Paule[7], Amanda M. Larracuente[5], Peter J. Maughan[3], Michael S. Barker[6], Julio V. Schneider[2], Steffen U. Pauls[1,2,8]

Affiliations

[1]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany, jacqueline.heckenhauer@senckenberg.de

[2]Department of Terrestrial Zoology, Entomology III, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Germany

[3]Department of Plant & Wildlife Sciences, Brigham Young University, Provo, UT

[4]Data Science Lab, Smithsonian Institution, Washington, DC

[5]Department of Biology, University of Rochester, Rochester, NY, USA

[6]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA

[7]Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Germany

[8]Institute for Insect Biotechnology, Justus-Liebig-University, Gießen, Germany

Abstract

Genome size can vary widely over relatively short evolutionary time scales and is implicated in form, function and ecological success of a species. Here, we generated 17 new de novo whole genome assemblies and present a holistic view on genome size diversity of the highly diversified, non-model insect order, Trichoptera (caddisflies). We detect large variation in genome size and find strong evidence that transposable element (TE) expansions are the primary driver of genome size evolution: TE expansions contribute to larger genomes in clades with higher ecological diversity and have a major impact on protein-coding gene

regions. These TE-gene associations show a linear relationship with increasing genome size. Our findings suggest new hypotheses for future testing, especially the effects of TE activity and TE-gene associations on genome stability, gene expression, phenotypes, and their potential adaptive advantages in groups with high species, ecological, and functional diversities.

Introduction

Genome size can vary widely over relatively short evolutionary time scales, although our knowledge is still scarce for many non-model groups. This is especially true for insects, which account for much of the earth's species diversity. To date 1,345 insect genome size estimates have been published, representing less than 0.15% of all described insect species (Gregory, 2005: Animal Genome Size Database: http://www.genomesize.com, last accessed 2021-04-30). Genome size in insects ranges 240-fold from 69 Mb in chironomid midges (1) to 16.5 Gb in the mountain grasshopper *Podisma pedestris* (2).

Genome size variation relates poorly to the number of coding genes or the complexity of the organism (C-value enigma, (3–6) and evolutionary drivers of genome size variation remain a topic of ongoing debate (e.g. 7–10). Two principally different mechanisms are proposed as primary drivers of eukaryotic genome size evolution: Whole genome duplication (WGD, i.e. polyploidy) or smaller duplication events and expansion of repetitive elements (REs, (5). While WGD is ubiquitous in plant evolution, it has been regarded as the exception in animals (11,12). However, ancient WGD has been hypothesized to be an important driver of evolution of mollusks (e.g. (13), amphibians (e.g. (14–15), fish (e.g. (16 –18) and arthropods (e.g. (19–21), including multiple putative ancient large-scale gene duplications within Trichoptera (22). REs are the primary driver of genome size variation in many eukaryotic genomes (23–24). The two major categories of REs are tandem repeats (satellite DNA) and mobile transposable elements (TEs). TEs are classified into class I [retrotransposons: endogenous retroviruses (ERVs), related long terminal repeat (LTR) and non-LTR retrotransposons: SINEs (Short Interspersed Nuclear Elements), LINEs (Long Interspersed Nuclear Elements)] and class II elements (DNA transposons, (25). In insects, the genomic proportion of TEs ranges from 1% in the antarctic midge *Belgica antarctica* (26) to 65% in the migratory locust *Locusta migratoria* (27). Broad-scale analysis of TE abundance in insects suggests that some order-specific signatures are present, however, major shifts in TE abundance are also common at shallow taxonomic levels (28–29) including in Trichoptera (*30*). The movement and proliferation of REs can have deleterious consequences on gene function and genome stability

(31–35). Moreover, repeat content and abundance can turn over rapidly even over short evolutionary time scales (reviewed in (36). This rapid evolution has consequences for genome evolution and speciation, e.g. repeat divergence causes genetic incompatibilities between even closely related species (37). However, TEs can also be sources of genomic innovation with selective advantages for the host (38–43) and they contribute to global changes in gene regulatory networks (44–46). Investigating RE dynamics in diverse clades provides a powerful lens for understanding their capacity for contributing to genome function and evolution. In addition, studying REs in species-rich groups may be critical for understanding how environmental factors are linked to genome evolution in natural populations (see (47)). With more than 16,500 species, caddisflies (Trichoptera) are among the most diverse of all aquatic insects (48). Their species richness is reflective of their ecological diversity, including, e.g. microhabitat specialization, full array of feeding modes, and diverse use of underwater silk secretions (49-50)An initial comparison of six caddisfly species found wide genome size variation in Trichoptera (raniging from 230 Mbp to 1.4 Gbp) and hypothesized that the observed variation was correlated with caddisfly phylogeny and that TEs contributed to a suborder-specific increase of genome size (30). Here, we present a multi-faceted analysis to investigate genome size evolution in the order Trichoptera, as an example for highly diversified non-model organisms. Specifically, we (1.) estimate genome size for 40 Trichoptera species across the order to address the hypothesis that there are phylogenetic patterns in the distribution of genome size variation of Trichoptera and (2.) use 26 (including 17 new) Trichoptera genomes to investigate which processes (WGDs, expansion of REs) play a role in genome size evolution of caddisflies. Studying the genomic diversity of this highly diversified insect order adds new insights into drivers of genome size evolution with potential to shed light on how genome size is linked to form, function, and ecology.

Results

New genomic resources

We combined long- and short-read sequencing technologies to generate 17 new *de novo* genome assemblies across a wide taxonomic range, covering all major lineages of Trichoptera. These new assemblies increase available assemblies in the order from nine to 26, nearly tripling the number of publicly available caddisfly reference genomes. The new genomes are of comparable or better quality than other Trichoptera genomes previously reported in terms of BUSCO completeness and contiguity (Table 1, Data S1) and expand the

availability of genomic resources of aquatic insects which are underrepresented in genomic research (51).

Genome size evolution in Trichoptera

Based on genomes of six trichopteran species, Olsen et al. (30) found a 3-fold suborder-specific increase of genome size and hypothesized that genome size variation is correlated with their phylogeny. To test this hypothesis, we first reconstructed phylogenetic relationships by analyzing ~2,000 single-copy BUSCO genes from the 26 study species (Fig. 1 & Fig. 2, Fig. S1). Phylogenetic relationships in this study agree with recent phylogenetic hypotheses (52), see supplementary note 1) which showed that Trichoptera is divided into two suborders: Annulipalpia (Figs 1 & 2: Clade A, blue) and Integripalpia [consisting of basal Integripalpia (Fig. 1: Clade B1-3, light green) and infraorder Phryganides (Fig. 1: clade B4, dark green)]. Trichopterans use silk to build diverse underwater structures (see illustrations Fig.1; supplementary note 1, Fig. S1). Thus, we refer to Annulipalpia as 'fixed retreat- and net-spinners', to Phryganides (Integripalpia) as 'tube case-builders', and to basal Integripalpia as 'cocoon-builders'.

We used three common approaches for estimating genome size across Trichoptera: *k-mer* distribution-estimates, backmapping of sequence data to available draft genomes, and flow cytometry (FCM, results are reported in Data S1). FCM estimates can be affected by chromatin condensation, the proportion of cells in G0 to G1 phases (53–54) and endoreplication in insect cells and tissues (55) while *kmer*-based estimates can be affected by repetitive elements in the genome resulting in smaller genome size estimates (e.g. (56–58)). Bland-Altman plots (Fig. S2) revealed general agreement of all three methods in our study (Fig. S2). However, the FCM estimates were generally higher compared to the sequence-based estimates (Fig. 1, Data S1) and, among all three approaches, this measure is expected to be the most accurate (8). We observe that variation among the methods increased with genome size, indicating issues potentially caused by repeat content (see Results *Repeat dynamics*).

We observed large variations in genome size across the order. Genome size tends to be lower in 'fixed retreat- and net-spinners' and 'cocoon-builders' compared to 'tube case-builders' (Fig. 1). Specifically, we observe that genome size varies ~14-fold ranging from 1C = 153.52 Mb in 'cocoon-builders' (Fig. 1, B1: Hydroptilidae) to 1C = 2129.05 Mb in 'tube case-builders' (Fig. 1, clade B4: Limnephilidae). Of the 29 species analyzed by FCM, *Halesus digitatus* (Fig. 1, clade B4: Limnephilidae, Intergripalpia) possessed the largest genome (1C =

2129.05 Mb), while the genome of *Hydropsyche saxonica* (Fig. 1, clade A: Hydropsychidae, 'fixed retreat- and net-spinners') was the smallest (1C = 241.54 Mb). Genome size estimates based on sequence-based methods (*kmer*-based and back-mapping) range from 1C = 153.52 - 160.00 Mb in *Agraylea sexmaculata* (Fig. 1, clade B1: Hydroptilidae, 'cocoon-builders') to 1C = 1238.19 - 1400.00 Mb in *Sericostoma* sp. (Fig. 1, clade B4: Sericostomatidae, 'tube case-builders').

Repeat Dynamics

Repetitive element abundance and classification

To understand the structural basis of genome size variation across the order Trichoptera we explored repetitive element (RE) content. We found that major expansions of transposable elements (TEs) contribute to larger genomes in 'tube case-' and some 'cocoon-builders', but particularly in 'tube case-builders' with an average of ~600 Mb of REs compared to ~138 Mb in 'fixed retreat- and net-spinners' (Fig. 2 A, B). LINEs are the most abundant classified TEs in 'cocoon-' and 'tube case-builders' and comprise >154 Mb on average in 'tube case-builders', or an average genome proportion of 16.9% (range = 5.6–34.7%). This represents a 1.8- and 2.8-fold increase in genome proportion relative to 'cocoon-builders' and 'fixed retreat- and net-spinners' respectively. The LINE abundance of >312 Mb in *Odontocerum albicorne* exceeds the entire assembly lengths (152–282 Mb) of the three smallest genome assemblies (*Hydropsyche tenuis*, *Parapsyche elsis*, and *Agraylea sexmaculata*, respectively) (Fig. 2 A, B). DNA transposons also comprise large genomic fractions in both 'cocoon-' and 'tube case-builders' (averages of 54.4 Mb and 32.8 Mb, respectively). However, despite containing a large number of bps, they make up a smaller fraction of total bps in the genomes of 'cocoon-' and 'tube case-builders' than in 'fixed retreat- and net-spinners' (average genome proportion = 5.9%, 4.5%, and 11.1% in 'tube case-builders', 'cocoon-builders', and 'fixed retreat- and net-spinners', respectively) (Fig. 2 B), and thus cannot explain the larger genome sizes. SINEs, LTRs, *Penelope* (grouped with "other" repeats in Fig. 2), and satDNAs show a disproportionate increase in 'cocoon-' and 'tube case-builders', however, all categories combined make up a relatively small proportion of their genomes (all less than 3% on average in Integripalpia) (Fig. 2, B). Unclassified repeats are the most abundant repeat category across all Trichoptera but also show disproportionate expansions in both 'cocoon-' and 'case-builders' relative to 'fixed retreat- and net-spinners' (Fig. 2 A, B). The general trends noted in our assembly-based analysis of REs were corroborated by our reference-free analysis of repeat abundance (Fig. S3, supplementary note 2).

TE age distribution analysis

To test whether the observed abundance patterns of specific TEs are driven by shared ancient proliferation events or more recent/ongoing activity of the respective TEs, we analyzed TE age distributions. TE age distributions showed high abundance of recently diverged TE sequences in 'cocoon-' and 'tube case-builders', particularly in LINEs, DNA transposons, and LTRs in which the majority of TE for a given class show 0–10% sequence divergence within copies of a given repeat (Fig. 3). This trend was particularly pronounced among 'tube case-builders' with several species showing high abundance of LINEs and DNA transposons with 0–5% sequence divergence (Fig. 3). This pattern suggests that the observed TE expansion is due primarily to ongoing TE activity within lineages rather than a few shared bursts of activity in ancestral lineages. This is further supported by our analysis of repeat sub-classes with landscape plots (Fig. S4). For example, in our study, LINE abundance is often due to expansion of different LINE subclasses even between species in the same sub-clade (e.g., compare *Lepidostoma* with *Micrasema*, *Himalopsyche* with *Glossosoma*; Fig. S4). We also find evidence of shared ancient bursts of SINE activity in 'cocoon-' and 'tube case-builders', although SINEs are not an abundant repeat class in any species (avg. genomic proportion=1.9% stdev=1.7%) (Fig. S4).

Associations between TE sequences and protein-coding genes

During early exploration of our sequence data, it became apparent that some lineages showed an amplification in the copy number of fragments of BUSCO genes. We refer to these as 'repetitive BUSCOs'. We found major expansions of repetitive BUSCO gene fragments in 'cocoon-' and 'tube case-builders' (Fig. 4A) that are significantly correlated with total repeat abundance, as well as the genomic proportion of LINEs and DNA transposons (supplementary Fig. S6). To test whether repetitive fragments are due to TE insertions or to proliferation of "true" BUSCO gene fragments, we examined patterns of BUSCO gene structure in pairwise alignments of three species pairs in which one species has a repetitive fragment (i.e., the "inflated" species) and its counterpart does not (i.e., the "reference" species). 73 of 75 randomly sampled alignments were consistently non-contiguous in inflated regions in the reference sequence (Fig. 4B). The number of BLAST hits in the genome assembly of an "inflated species" when using the BUSCO sequence from the "reference species" as a query was significantly fewer than when using the BUSCO sequence from the "inflated species" as a query (Fig. 4C, Data S2). Our intersect analysis showed that large

fractions of BLAST hits for repetitive BUSCOs (when BLASTed to their own assembly) are in fact annotated transposable element sequences, especially LINEs and DNA transposons (Fig. 4D). Taken together these findings provide strong evidence that the repetitive fragments we observe in BUSCO genes are derived from TE sequences. These sequences may represent remnants left behind following historical insertion events, or TE sequences with open reading frames that are immediately adjacent to protein-coding genes and inadvertently classified as BUSCO fragments. Given the large fraction of BUSCO genes shown to have TE fragments (averages of 11.2% and 21.4% in 'cocoon-' and 'tube case-builders', respectively, compared to 0.062% in annulipalipans), this finding highlights the impact of TEs on the composition of protein-coding genes in species with repeat-rich genomes.

Gene and genome duplications

Recently, a transcriptome-based study found evidence for putative ancient gene and genome duplications in hexapods, including potential WGD events in caddisflies (22), suggesting that duplication events could be responsible for some genome size variation in Trichoptera. We investigated whether this pattern persists with whole genome data and found that the age distribution of duplications in 18 genomes were significantly different compared to the background rate of gene duplication (Figs. S7 & S8). To identify if any significant peak is consistent with a potential WGD, we used mixture modeling to identify peaks in these gene age distributions. We found no obvious peak consistent with an ancient WGD suggesting that WGD did not play a major role in genome size variation in Trichoptera.

Discussion

The drivers and evolutionary consequences of genome size evolution are a topic of ongoing debate. Several models have been proposed (8). Some hypothesize genome size to be a (mal)adaptive trait by impacting phenotypic traits such as developmental/life history, body size and other cell-size related effects (59–62) reviewed in (8). On the other hand, neutral theories suggest that DNA accumulation occurs only by genetic drift without selective pressures playing a major role in the accumulation or loss of DNA [the mutational hazard hypothesis (MHH, (23) and the mutational equilibrium hypothesis (MEH, (24)]. The MHH includes only small deleterious effects for the accumulation of extra DNA which is accompanied by higher mutation rates in larger genomes (23), while the MEH focuses on the balance between insertions and deletions. It suggests that genome expansions arise by 'bursts' of duplication events or TE activity and genome shrinkage may be caused by a more constant

rate of small deletions (24).

In this study, we observe that genome size varies ~14-fold across the order Trichoptera, with lower genome size estimates in ‘fixed retreat- and net-spinners’ and ‘cocoon-builders’ compared to ‘tube case-builders’ and explore potential drivers of genome size evolution. Although, recent genomic studies have shown evidence of bursts of gene duplication and gene family expansion during the evolution of hexapods (22, 63), the presence of ancient genome duplication events are still in debate (64–66). We found no evidence of ancient WGD in the gene age distribution in our Trichoptera genomes, although we recognize that our current genome assemblies might be too fragmented to infer synteny. Future studies using chromosome-level genome assemblies are needed to allow a deeper exploration of putative ancient duplications observed by Li *et al.* (22).

We find evidence that TE expansions (especially LINEs) are the primary driver of genome size evolution in Trichoptera (Fig. 2, Fig. S3), which is consistent with the MEH explained above. Our TE age distribution analysis suggests that the high abundance of LINEs is due to ongoing/recent activity occurring independently across ‘cocoon-’ and particularly ‘tube case-builders’ (Fig. 3, Fig. S4). Thus, the shift to large genomes in these lineages does not appear to be due to a single (or few) shared ancient events, rather the they maintain dynamic turnover in composition in their large genomes. Mutational bias affecting pathways tied to TE-regulation may affect insertion/deletion ratios and subsequently lead to lineage-specific shifts in genome size equilibrium (67). Such changes may be stochastic (e.g., due to drift), or linked to traits that evolve on independent trajectories as lineages diverge, and are thereby constrained by phylogeny. Ecological factors, demographic history and effective population size can further impact mutation rates. For example, environmental stress can trigger bursts of TE activity and elevated mutation rates (68–70) driving lineages that occupy niche space with frequent exposure to environmental stress toward increased TE loads and larger genomes. Similarly, lineages with small effective population sizes or which are prone to population bottlenecks may have higher mutation rates and/or reduced efficacy of natural selection which would otherwise purge mildly deleterious TE load.

Although our study is not designed to pinpoint specific forces maintaining large genomes in some lineages, the pattern we observe in the distribution of genome size (i.e. lower genome size estimates in ‘fixed retreat- and net-spinners’ and ‘cocoon-builders’ compared to ‘tube case-builders’) leads us to hypothesize that ecological factors may play a role in genome size

evolution in the order. The three focal groups discussed here exhibit markedly different ecological strategies. Larvae of ‘fixed retreat- and net-spinners’ generally occupy relatively narrow niche space in oxygen-rich flowing-water (mostly stream/river) environments where they rely on water currents to bring food materials to their filter nets. The evolutionary innovation of tube-case making is thought to have enabled ‘tube case-builders’ to occupy a much greater diversity of ecological niche space by allowing them to obtain oxygen in stillwater (e.g., pond, lake, marsh) environments which are much more variable in temperature and oxygen availability than stream environments (71–72). Given the species and ecological diversification in habitats that are more variable at both short (daily, seasonally) and long temporal scales, it is plausible these lineages experience greater environmental stress, and less stable population demographics that could lead to both more frequent TE bursts, and reduced efficacy of natural selection in purging deleterious effects of TE expansions as described above (23–24).

We show that TE expansions (especially LINEs and DNA transposons) in ‘cocoon-’ and ‘tube case-builders’ have a major impact on protein-coding gene regions (Fig. 4). These TE-gene associations show a linear relationship with increasing genome size. This trend is particularly pronounced among ‘tube case-builders’ in which an average of 21.4% of BUSCO genes have internal or adjacent repetitive sequences (compared to 6.2% in annulipalpians). This finding corroborates other studies highlighting the role of TEs as drivers of rapid genome evolution (73–76) and highlights their impact on genomic regions that have potential effects on phenotypes. Questions remain as to what evolutionary roles such changes in genic regions may play. In general, TE insertions are considered to have deleterious effects on their host’s fitness activity (77–78). They are particularly prone to interrupt genes (33), pose a risk of ectopic recombination that can lead to genome rearrangements (34, 31, 79), and have epigenetic effects on neighboring sequences (80–81). Therefore, purifying selection keeps TEs at low frequencies (33). However, there is growing evidence that TE activity can also be a critical source of new genetic variation driving diversification via chromosomal rearrangements and transposition events which can result in mutations (82), including examples, of co-option (83), e.g. recent research in mammals has shown that DNA transposon fragments can be co-opted to form regulatory networks with genome-wide effects on gene expression (44).

Ecological correlates with genome size are widely discussed in other taxa (84–88). Caddisflies and other diverse insect lineages that feature various microhabitat specializations,

feeding modes, and/or the use of silk represent evolutionary replicates with contrasting traits and dynamic genome size evolution. They thus have high potential as models for understanding links between ecology and the evolution of REs, genomes, and phenotypes. Our study lays a foundation for future work in caddisflies that investigates the potential impact of TE expansions on phenotypes and tests for evidence of co-option/adaptive impacts of TE-rich genomes against a null of neutral or slightly deleterious effects.

Materials and Methods

DNA extraction, library preparation, sequencing, and sequence read processing

We extracted high molecular weight genomic DNA (gDNA) from 17 individuals (15 species) of caddisfly larvae (for sampling information, see Data S1) after removing the intestinal tracts using a salting-out protocol adapted from (*89*) as described in supplementary note 3. We generated gDNA libraries for a low-cost high-contiguity sequencing strategy, i.e. employing a combination of short (Illumina) and long read (Nanopore or PacBio) technologies as described in supplementary note 4, Supplementary Material online. For details on sequencing coverage for each specimen see Data S1.

De novo genome assemblies and quality assessment

We applied different assembly strategies for different datasets. First, we applied a long-read assembly method with subsequent short-read polishing as this method revealed good results in previous *de novo* assemblies in caddisflies (57). In cases where this pipeline did not meet the expected quality regarding contiguity and BUSCO completeness, we applied de novo hybrid assembly approaches supplementary note 5. Illumina-only data was assembled with SPAdes (90) explained in supplementary note 5).

We calculated assembly statistics with QUAST v5.0.2 (91) and examined completeness with BUSCO v3.0.2 (92 –93) using the Endopterygota odb9 dataset with the options *--long, –m = genome* and *–sp= fly*. A summary of the assembly statistics and BUSCO completeness is given in Data S1. The final genome assemblies were screened and filtered for potential contaminations with taxon-annotated GC-coverage (TAGC) plots using BlobTools v1.0 (94). Details and blobplots are given in supplementary note 6 & Figs. S9-25.

Species tree reconstruction

We used the single-copy orthologs resulting from the BUSCO analyses to generate a species tree. We first combined single-copy ortholog amino acid files from each species into a single

FASTA for each ortholog. We then aligned them with the MAFFT L-INS-i algorithm (95). We selected amino acid substitution models for each ortholog using ModelFinder (option -m mfp,(96) in IQtree v.2.0.6 (97) and estimated a maximum likelihood tree with 1000 ultrafast bootstrap replicates (98) with the BNNI correction (option -bb 1000 -bnni). We combined the best maximum likelihood tree from each gene for species tree analysis in ASTRAL-III (99). We visualized the trees using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Genome size estimations and genome profiling

Genome size estimates of 29 species were conducted using flow cytometry (FCM) according to Otto (100) using Lycopersicon esculentum cv. Stupické polnítyčkové rané (2C = 1.96 pg;(101)) as internal standard and propidium iodine as stain. Additionally, we used the short-read data sequenced in this study, to conduct genome profiling (estimation of major genome characteristics such as size, heterozygosity, and repetitiveness) using a k-mer distribution-based method (GenomeScope 2.0, (102). Genome scope profiles are available online (see links Data S1 & Fig. S26-28). In addition, we applied a second sequencing-based method for genome size estimates, which uses the backmapping rate of sequenced reads to the assembly and coverage distribution (backmap.pl v0.1, (103). Details of all three methods are described in supplementary note 7. Coverage distribution per position and genome size estimate from backmap.pl are shown in Figs. S29-52). We assessed the congruence among the three quantitative methods of measurement (Genomescope2, Backmap.pl and FCM) with Bland-Altman-Plots using the function BlandAltmanLeh::bland.altman.plot in ggplot2 (104) in RStudio (RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/; supplementary note 8, Fig. S2).

Repeat dynamics

Repeat abundance and classification

We identified and classified repetitive elements in genome assemblies of each species using RepeatModeler2.0 (105). We annotated repeats in assemblies with RepeatMasker 4.1.0 (http://www.repeatmasker.org) using the custom repeat libraries generated from RepeatModeler2 for each respective assembly with the search engine set to “ncbi” and using the -xsmall option. We converted the softmasked assembly resulting from the first RepeatMasker round into a hardmasked assembly using the lc2n.py script (https://github.com/PdomGenomeProject/repeat-masking). Finally, we re-ran RepeatMasker on the hard-masked genome with RepeatMasker’s internal arthropod repeat library using -

species “Arthropoda”. We then merged RepeatMasker output tables from both runs to summarize the abundance of repeat categories.

We also estimated repetitive element abundance and composition using RepeatExplorer2 (106–107) This reference-free approach quantifies repeats directly from unassembled short-read data using clustering-based methods. This analysis allowed us to test for general consistency of patterns with our assembly-based approach described above, and to test for the presence of abundant repeat categories such as satellite DNAs which can comprise large fractions of genomes yet can be prone to poor representation in the genome assembly. Prior to analysis in RepeatExplorer2 we normalized input data sets to 0.5x coverage using RepeatProfiler (108) and seqtk (https://github.com/lh3/seqtk), and then ran RepeatExplorer2 clustering with the Metazoa 3.0 database specified for annotation.

TE age distribution analysis

We further characterized repetitive element dynamics in Trichoptera by analyzing TE landscapes, which show relative age differences among TE sequences and their genomic abundance. We used these analyses to test whether abundance patterns of specific TEs are driven by shared ancient proliferation events or more recent/ongoing activity of the respective TEs. For example, if shared ancient proliferation is driving abundance patterns of a given TE, the majority of its copies would show moderate to high sequence divergence (e.g., >10% pairwise divergence). In contrast, if abundance patterns are driven by recent/ongoing activity of a given TE, we would expect the majority of its sequences to show low sequence divergence (e.g., 0–10%). We generated TE landscape plots using dnaPipeTE v1.3.1 (*109*) with genomic coverage for each species sampled to 0.5X prior to analysis and the -RM_lib flag set to the Repbase v20170127 repeat library.

TE sequence associations with protein-coding genes

During the early exploration of our sequence data we noted that copy number profiles of some BUSCO genes showed regions of unexpected high copy number with the frequency of such genes varying widely across species. Following this observation, we analyzed all BUSCO genes for all species to quantify the abundance of repetitive BUSCOs across samples and test two hypotheses that could account for the observed patterns: (1.) inflated copy-number of BUSCO gene fragments occur when sequences of repetitive elements (e.g. TEs) are inserted into (or immediately adjacent to genes; (2.) inflated regions are BUSCO gene fragments that

have proliferated throughout the genome (e.g., by hitch-hiking with TEs). We investigated associations between TEs and genic sequences in Trichoptera lineages by quantifying the abundance of repetitive sequences found within or adjacent to BUSCO genes (for presence and absence of repetitive BUSCOs see Fig. S5, Data S2). This analysis also allowed us to quantify shifts in associations between TEs and genic regions across Trichoptera lineages with varying repeat abundance. We used copy number profiles to identify BUSCO genes with repetitive sequences based on coverage profiles and quantified their genomic abundance by using each repetitive BUSCO as a query in a BLAST search against their respective genome assembly. We then conducted intersect analysis for all unique BUSCO hits from repetitive sequences to determine if repetitive fragments were annotated as TEs. We calculated the total number of bases in filtered BLAST after subtracting the number of bases at the locus belonging to all 'complete' BUSCO genes and categorize BUSCO repeats based on their annotation status and repeat classification using custom scripts (available at LINK). We plotted the number of BUSCO repeats belonging to repetitive element categories (i.e., classes and subclasses) alongside plots of the relative genomic abundance of each respective category. For details of this method readers are directed to supplementary note 9. In addition, we investigated BUSCO genes with repetitive sequences by pairwise alignments. Specifically, we visualized alignments of BUSCOs with repetitive fragments alongside orthologous BUSCOS that lack repetitive fragments taken from closely related species. To test whether repetitive fragments are due to TE insertions or to proliferation of true BUSCO gene fragments we examined patterns of BUSCO gene structure in pairwise alignments of species in which one species has a repetitive fragment (i.e., the "inflated" species) and its counterpart does not (i.e., the "reference" species). To reduce the likelihood that indels due to sequence divergence (i.e., not due to REs) would complicate the analysis we conducted this test with three pairs of closely related species: *Micrasema longulum* and *Micrasema minimum*; *Rhyacophila brunnea* and *Rhyacophila evoluta*; and *Hesperophylax magnus* and *Halesus radiatus*. We predicted that if repetitive regions were due to RE sequence insertion (i.e., Hyp1 above) that inflated species would show indels at coordinates of repetitive fragments relative to reference species. If repetitive regions were due to proliferation of true BUSCO sequence (i.g., Hyp2 above) we predicted that alignments would be contiguous over the repetitive region. We further tested this prediction by taking the set of BUSCOs that are only repetitive in the inflated species and contrasted results of two BLAST searches followed by an intersect analysis. A detailed description of this method is provided in supplementary note 9.

Gene and Genome duplications

Genome Annotation

Prior to annotating the individual genomes with MAKER2 v2.31.10 (110–111) we used RepeatModeler v1.0.11 and RepeatMasker v4.0.7, to identify species-specific repetitive elements in each of the assemblies, relative to RepBase libraries v20181026; www.girinst.org). Transcriptome evidence for the annotation of the individual genomes included their species-specific or closely related *de novo* transcriptome provided by 1KITE (http://www.1kite.org/; last accessed November 11, 2019, Data S1) or downloaded from Genbank as well as the cDNA and protein models from *Stenopsyche tienmushanensis* (112) and *Bombyx mori* (AR102, GenBank accession ID# GCF_000151625.1). Additional protein evidence included the uniprot-sprot database (downloaded September 25, 2018). We masked repeats based on species-specific files produced by RepeatModeler. For *ab initio* gene prediction, species specific AUGUSTUS gene prediction models as well as *Bombyx mori* SNAP gene models were provided to MAKER. The EvidenceModeler (113) and tRNAscan (114) options in MAKER were employed to produce a weighted consensus gene structure and to identify tRNAs genes. MAKER default options were utilized for BLASTN, BLASTX, TBLASTX searches. Two assemblies (*Agapetus fuscipens* GL3 and *Micrasema longulum* ML1) were not annotated because of their low contiguity. All protein sequences were assigned putative names by BlastP Protein–Protein BLAST 2.2.30+ searches (115) and were functionally annotated using command line Blast2Go v1.3.3 (116), see supplementary note 10, Figs. S53-82).

Inference of WGDs from gene age distributions

To recover signal from potential WGDs, for each genome, we used the DupPipe pipeline to construct gene families and estimate the age distribution of gene duplications (117), https://bitbucket.org/barkerlab/evopipes/src/master/). We translated DNA sequences and identified ORFs by comparing the Genewise (118) alignment to the best-hit protein from a collection of proteins from 24 metazoan genomes from Metazome v3.0. For all DupPipe runs, we used protein-guided DNA alignments to align our nucleic acid sequences while maintaining the ORFs. We estimated synonymous divergence (K_s) using PAML with the F3X4 model (119) for each node in the gene family phylogenies. We first identified taxa with potential WGDs by comparing their paralog ages to a simulated null distribution without ancient WGDs using a K-S goodness-of-fit test (120). We then used mixture modeling to identify if any significant peaks consistent with a potential WGD and to estimate their median paralog K_s values. Significant peaks were identified using a likelihood ratio test in the boot.comp function of the package mixtools in R (121).

References

1. R. Cornette, O. Gusev, Y. Nakahara, S. Shimura, T. Kikawada, T. Okuda, Chironomid midges (Diptera, chironomidae) show extremely small genome sizes. *Zoolog. Sci.* **32**, 248–254 (2015).

2. M. Westerman, N. H. Barton, G. M. Hewitt, Differences in DNA content between two chromosomal races of the grasshopper Podisma pedestris. *Heredity*. **58**, 221–228 (1987).

3. C. A. Thomas, The genetic organization of chromosomes. *Annu. Rev. Genet.* **5**, 237–256 (1971).

4. J. Bernard, *The Eukaryote Genome in Development and Evolution* (Springer Netherlands, 1988; https://www.springer.com/gp/book/9780045750337).

5. T. R. GREGORY, The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Ann. Bot.* **95**, 133–146 (2005).

6. T. A. Elliott, T. R. Gregory, What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140331 (2015).

7. H. Abdel-Haleem, The Origins of Genome Architecture. *J. Hered.* **98**, 633–634 (2007).

8. J. Blommaert, Genome size evolution: towards new model systems for old questions. *Proc. R. Soc. B Biol. Sci.* **287**, 20201441 (2020).

9. A. Canapa, M. Barucca, M. A. Biscotti, M. Forconi, E. Olmo, Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenet. Genome Res.* **147**, 217–239 (2015).

10. D. A. Petrov, Evolution of genome size: new approaches to an old problem. *Trends Genet. TIG*. **17**, 23–28 (2001).

11. H. A. Orr, "Why Polyploidy is Rarer in Animals Than in Plants" Revisited. *Am. Nat.* **136**, 759–770 (1990).

12. S. P. Otto, J. Whitton, Polyploid Incidence and Evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).

13. N. M. Hallinan, D. R. Lindberg, Comparative analysis of chromosome counts infers three paleopolyploidies in the mollusca. *Genome Biol. Evol.* **3**, 1150–1163 (2011).

14. M. A. Alexandrou, B. A. Swartz, N. J. Matzke, T. H. Oakley, Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* **69**, 514–523 (2013).

15. M. A. Campbell, M. C. Hale, G. J. McKinney, K. M. Nichols, D. E. Pearse, Long-Term Conservation of Ohnologs Through Partial Tetrasomy Following Whole-Genome Duplication in Salmonidae. *G3 GenesGenomesGenetics*. **9**, 2017–2028 (2019).

16. I. G. Woods, C. Wilson, B. Friedlander, P. Chang, D. K. Reyes, R. Nix, P. D. Kelly, F. Chu, J. H. Postlethwait, W. S. Talbot, The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15**, 1307–1314 (2005).

17. C. Berthelot, F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J.-M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J.-N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. R. Crollius, Y. Guiguen, The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* **5**, 3657 (2014).

18. S. M. K. Glasauer, S. C. F. Neuhauss, Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics MGG*. **289**, 1045–1060 (2014).

19. T. H. Clarke, J. E. Garb, C. Y. Hayashi, P. Arensburger, N. A. Ayoub, Spider Transcriptomes Identify Ancient Large-Scale Gene Duplication Event Potentially Important in Silk Gland Evolution. *Genome Biol. Evol.* **7**, 1856–1870 (2015).

20. N. J. Kenny, K. W. Chan, W. Nong, Z. Qu, I. Maeso, H. Y. Yip, T. F. Chan, H. S. Kwan, P. W. H. Holland, K. H. Chu, J. H. L. Hui, Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*. **116**, 190–199 (2016).

21. E. E. Schwager, P. P. Sharma, T. Clarke, D. J. Leite, T. Wierschin, M. Pechmann, Y. Akiyama-Oda, L. Esposito, J. Bechsgaard, T. Bilde, A. D. Buffry, H. Chao, H. Dinh, H. Doddapaneni, S. Dugan, C. Eibner, C. G. Extavour, P. Funch, J. Garb, L. B. Gonzalez, V. L. Gonzalez, S. Griffiths-Jones, Y. Han, C. Hayashi, M. Hilbrant, D. S. T. Hughes, R. Janssen, S. L. Lee, I. Maeso, S. C. Murali, D. M. Muzny, R. Nunes da Fonseca, C. L. B. Paese, J. Qu,

M. Ronshaugen, C. Schomburg, A. Schönauer, A. Stollewerk, M. Torres-Oliva, N. Turetzek, B. Vanthournout, J. H. Werren, C. Wolff, K. C. Worley, G. Bucher, R. A. Gibbs, J. Coddington, H. Oda, M. Stanke, N. A. Ayoub, N.-M. Prpic, J.-F. Flot, N. Posnien, S. Richards, A. P. McGregor, The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol.* **15**, 62 (2017).

22. Z. Li, G. P. Tiley, S. R. Galuska, C. R. Reardon, T. I. Kidder, R. J. Rundell, M. S. Barker, Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci.* **115**, 4713–4718 (2018).

23. M. Lynch, J. S. Conery, The Origins of Genome Complexity. *Science*. **302**, 1401–1404 (2003).

24. D. A. Petrov, Mutational Equilibrium Model of Genome Size Evolution. *Theor. Popul. Biol.* **61**, 531–544 (2002).

25. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).

26. J. L. Kelley, J. T. Peyton, A.-S. Fiston-Lavier, N. M. Teets, M.-C. Yee, J. S. Johnston, C. D. Bustamante, R. E. Lee, D. L. Denlinger, Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat. Commun.* **5**, 4611 (2014).

27. X. Wang, X. Fang, P. Yang, X. Jiang, F. Jiang, D. Zhao, B. Li, F. Cui, J. Wei, C. Ma, Y. Wang, J. He, Y. Luo, Z. Wang, X. Guo, W. Guo, X. Wang, Y. Zhang, M. Yang, S. Hao, B. Chen, Z. Ma, D. Yu, Z. Xiong, Y. Zhu, D. Fan, L. Han, B. Wang, Y. Chen, J. Wang, L. Yang, W. Zhao, Y. Feng, G. Chen, J. Lian, Q. Li, Z. Huang, X. Yao, N. Lv, G. Zhang, Y. Li, J. Wang, J. Wang, B. Zhu, L. Kang, The locust genome provides insight into swarm formation and long-distance flight. *Nat. Commun.* **5**, 2957 (2014).

28. M. Petersen, D. Armisén, R. A. Gibbs, L. Hering, A. Khila, G. Mayer, S. Richards, O. Niehuis, B. Misof, Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol. Evol.* **19**, 11 (2019).

29. C. Gilbert, J. Peccoud, R. Cordaux, Transposable Elements and the Evolution of Insects. *Annu. Rev. Entomol.* **66**, 355–372 (2021).

30. L. K. Olsen, J. Heckenhauer, J. S. Sproul, R. B. Dikow, V. L. Gonzalesz, M. P. Kweskin, A. M. Taylor, S. B. Wilson, R. J. Stewart, X. Zhou, R. Holzenthal, S. U. Pauls, P. B. Frandsen, De novo whole genome assemblies of Agrypnia vestita Walker, and Hesperophlax magnus Banks reveal substantial repetitive element expansion in tube case-making caddisflies (Insecta: Trichoptera). *Genome Biol. Evol.* (2021), doi:10.1093/gbe/evab013.

31. E. Montgomery, B. Charlesworth, C. H. Langley, A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. *Genet. Res.* **49**, 31–41 (1987).

32. B. Charlesworth, C. H. Langley, Transposition of copia elements in Drosophila. *Nature*. **332**, 21–21 (1988).

33. B. Charlesworth, C. H. Langley, The Population Genetics of Drosophila Transposable Elements. *Annu. Rev. Genet.* **23**, 251–287 (1989).

34. E. A. Montgomery, S. M. Huang, C. H. Langley, B. H. Judd, Chromosome rearrangement by ectopic recombination in Drosophila melanogaster: genome structure and evolution. *Genetics*. **129**, 1085–1098 (1991).

35. P. A. Mieczkowski, F. J. Lemoine, T. D. Petes, Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. *DNA Repair*. **5**, 1010–1020 (2006).

36. Đ. Ugarković, M. Plohl, Variation in satellite DNA profiles—causes and effects. *EMBO J.* **21**, 5955–5959 (2002).

37. P. M. Ferree, D. A. Barbash, Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila. *PLOS Biol.* **7**, e1000234 (2009).

38. L. J. Gahan, F. Gould, D. G. Heckel, Identification of a Gene Associated with Bt Resistance in Heliothis virescens. *Science*. **293**, 857–860 (2001).

39. S. Chen, X. Li, Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. *BMC Evol. Biol.* **7**, 46 (2007).

40. J. González, K. Lenkov, M. Lipatov, J. M. Macpherson, D. A. Petrov, High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLOS Biol.* **6**, e251 (2008).

41. J. González, T. L. Karasov, P. W. Messer, D. A. Petrov, Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. *PLOS Genet.* **6**, e1000905 (2010).

42. K. Itokawa, O. Komagata, S. Kasai, Y. Okamura, M. Masada, T. Tomita, Genomic structures of Cyp9m10 in pyrethroid resistant and susceptible strains of Culex quinquefasciatus. *Insect Biochem. Mol. Biol.* **40**, 631–640 (2010).

43. A. E. van't Hof, P. Campagne, D. J. Rigden, C. J. Yung, J. Lingley, M. A. Quail, N. Hall, A. C. Darby, I. J. Saccheri, The industrial melanism mutation in British peppered moths is a transposable element. *Nature*. **534**, 102–105 (2016).

44. C. Feschotte, Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).

45. C. E. Ellison, D. Bachtrog, Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. *Science*. **342**, 846–850 (2013).

46. M. E. Santos, I. Braasch, N. Boileau, B. S. Meyer, L. Sauteur, A. Böhne, H.-G. Belting, M. Affolter, W. Salzburger, The evolution of cichlid fish egg-spots is linked with a cis -regulatory change. *Nat. Commun.* **5**, 5149 (2014).

47. K. Alfsnes, H. P. Leinaas, D. O. Hessen, Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecol. Evol.* **7**, 5939–5947 (2017).

48. Trichoptera World Checklist, (available at https://entweb.sites.clemson.edu/database/trichopt/).

49. G. B. Wiggins, R. J. Mackay, Some Relationships between Systematics and Trophic Ecology in Nearctic Aquatic Insects, with Special Reference to Trichoptera. *Ecology*. **59**, 1211–1220 (1978).

50. R. J. Mackay, G. B. Wiggins, Ecological Diversity in Trichoptera. *Annu. Rev. Entomol.* **24**, 185–208 (1979).

51. S. Hotaling, J. L. Kelley, P. B. Frandsen, Aquatic Insects Are Dramatically Underrepresented in Genomic Research. *Insects*. **11**, 601 (2020).

52. J. A. Thomas, P. B. Frandsen, E. Prendini, X. Zhou, R. W. Holzenthal, A multigene phylogeny and timeline for Trichoptera (Insecta). *Syst. Entomol.* **45**, 670–686 (2020).

53. S. J. Hanrahan, J. S. Johnston, New genome size estimates of 134 species of arthropods. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **19**, 809–823 (2011).

54. Y.-S. Yu, S. Jin, N. Cho, J. Lim, C.-H. Kim, S.-G. Lee, S. Kim, J.-S. Park, K. Kim, C. Park, S.-J. Cho, Genome Size Estimation of Callipogon relictus Semenov (Coleoptera: Cerambycidae), an Endangered Species and a Korea Natural Monument. *Insects*. **12**, 111 (2021).

55. E. E. Hare, J. S. Johnston, Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol. Biol. Clifton NJ*. **772**, 3–12 (2011).

56. C. M. Austin, M. H. Tan, K. A. Harrisson, Y. P. Lee, L. J. Croft, P. Sunnucks, A. Pavlova, H. M. Gan, De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read. *GigaScience*. **6** (2017), doi:10.1093/gigascience/gix063.

57. J. Heckenhauer, P. B. Frandsen, D. K. Gupta, J. Paule, S. Prost, T. Schell, J. V. Schneider, R. J. Stewart, S. U. Pauls, Annotated Draft Genomes of Two Caddisfly Species Plectrocnemia conspersa CURTIS and Hydropsyche tenuis NAVAS (Insecta: Trichoptera). *Genome Biol. Evol.* **11**, 3445–3451 (2019).

58. J. M. Pflug, V. R. Holmes, C. Burrus, J. S. Johnston, D. R. Maddison, Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3 GenesGenomesGenetics*. **10**, 3047–3060 (2020).

59. M. D. Bennett, R. Riley, The duration of meiosis. *Proc. R. Soc. Lond. B Biol. Sci.* **178**, 277–299 (1971).

60. T. Cavalier-Smith, Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* **34**, 247–278 (1978).

61. T. R. Gregory, P. D. N. Hebert, The Modulation of DNA Content: Proximate Causes and Ultimate Consequences. *Genome Res.* **9**, 317–324 (1999).

62. T. CAVALIER-SMITH, Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. *Ann. Bot.* **95**, 147–175 (2005).

63. G. W. C. Thomas, E. Dohmen, D. S. T. Hughes, S. C. Murali, M. Poelchau, K. Glastad, C. A. Anstead, N. A. Ayoub, P. Batterham, M. Bellair, G. J. Binford, H. Chao, Y. H. Chen, C. Childers, H. Dinh, H. V. Doddapaneni, J. J. Duan, S. Dugan, L. A. Esposito, M. Friedrich, J. Garb, R. B. Gasser, M. A. D. Goodisman, D. E. Gundersen-Rindal, Y. Han, A. M. Handler, M. Hatakeyama, L. Hering, W. B. Hunter, P. Ioannidis, J. C. Jayaseelan, D. Kalra, A. Khila, P. K. Korhonen, C. E. Lee, S. L. Lee, Y. Li, A. R. I. Lindsey, G. Mayer, A. P. McGregor, D. D. McKenna, B. Misof, M. Munidasa, M. Munoz-Torres, D. M. Muzny, O. Niehuis, N. Osuji-Lacy, S. R. Palli, K. A. Panfilio, M. Pechmann, T. Perry, R. S. Peters, H. C. Poynton, N.-M. Prpic, J. Qu, D. Rotenberg, C. Schal, S. D. Schoville, E. D. Scully, E. Skinner, D. B. Sloan, R. Stouthamer, M. R. Strand, N. U. Szucsich, A. Wijeratne, N. D. Young, E. E. Zattara, J. B. Benoit, E. M. Zdobnov, M. E. Pfrender, K. J. Hackett, J. H. Werren, K. C. Worley, R. A. Gibbs, A. D. Chipman, R. M. Waterhouse, E. Bornberg-Bauer, M. W. Hahn, S. Richards, Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).

64. Y. Nakatani, A. McLysaght, Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1816–1818 (2019).

65. Z. Li, G. P. Tiley, R. J. Rundell, M. S. Barker, Reply to Nakatani and McLysaght: Analyzing deep duplication events. *Proc. Natl. Acad. Sci.* **116**, 1819–1820 (2019).

66. D. Roelofs, A. Zwaenepoel, T. Sistermans, J. Nap, A. A. Kampfraath, Y. Van de Peer, J. Ellers, K. Kraaijeveld, Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biol.* **18**, 57 (2020).

67. S. Lamichhaney, R. Catullo, J. S. Keogh, S. Clulow, S. V. Edwards, T. Ezaz, A bird-like genome from a frog: Mechanisms of genome size reduction in the ornate burrowing frog, Platyplectrum ornatum. *Proc. Natl. Acad. Sci.* **118** (2021), doi:10.1073/pnas.2011649118.

68. P. Capy, G. Gasperi, C. Biémont, C. Bazin, Stress and transposable elements: co-evolution or useful parasites? *Heredity*. **85 (Pt 2)**, 101–106 (2000).

69. A. Pecinka, H. Q. Dinh, T. Baubec, M. Rosa, N. Lettner, O. M. Scheid, Epigenetic Regulation of Repetitive Elements Is Attenuated by Prolonged Heat Stress in Arabidopsis. *Plant Cell*. **22**, 3118–3129 (2010).

70. M. Tittel-Elmer, E. Bucher, L. Broger, O. Mathieu, J. Paszkowski, I. Vaillant, Stress-Induced Activation of Heterochromatic Transcription. *PLOS Genet.* **6**, e1001175 (2010).

71. G. Wiggins, *Larvae of the North American Caddisfly Genera (Trichoptera)* (University of Toronto Press, 2018; https://www.degruyter.com/document/doi/10.3138/9781442623606/html).

72. G. B. Wiggins, C. I. for S. and T. Information, R. O. Museum, *Caddisflies: The Underwater Architects* (University of Toronto Press, 2004).

73. D. Kordiš, N. Lovšin, F. Gubenšek, Phylogenomic Analysis of the L1 Retrotransposons in Deuterostomia. *Syst. Biol.* **55**, 886–901 (2006).

74. I. A. Warren, M. Naville, D. Chalopin, P. Levin, C. S. Berger, D. Galiana, J.-N. Volff, Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* **23**, 505–531 (2015).

75. A. Suh, G. Churakov, M. P. Ramakodi, R. N. Platt II, J. Jurka, K. K. Kojima, J. Caballero, A. F. Smit, K. A. Vliet, F. G. Hoffmann, J. Brosius, R. E. Green, E. L. Braun, D. A. Ray, J. Schmitz, Multiple Lineages of Ancient CR1 Retroposons Shaped the Early Genome Evolution of Amniotes. *Genome Biol. Evol.* **7**, 205–217 (2015).

76. F. C. Grandi, W. An, Non-LTR retrotransposons and microsatellites. *Mob. Genet. Elem.* **3**, e25674 (2013).

77. T. F. C. Mackay, Transposable elements and fitness in Drosophila melanogaster. *Genome* (2011), doi:10.1139/g89-046.

78. E. G. Pasyukova, S. V. Nuzhdin, T. V. Morozova, T. F. C. Mackay, Accumulation of Transposable Elements in the Genome of Drosophila melanogaster is Associated with a Decrease in Fitness. *J. Hered.* **95**, 284–290 (2004).

79. C. H. Langley, E. Montgomery, R. Hudson, N. Kaplan, B. Charlesworth, On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52**, 223–235 (1988).

80. J. D. Hollister, B. S. Gaut, Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428 (2009).

81. Y. C. G. Lee, G. H. Karpen, Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *eLife*. **6**, e25762 (2017).

82. K. Kraaijeveld, Genome Size and Species Diversification. *Evol. Biol.* **37**, 227–233 (2010).

83. R. L. Cosby, N.-C. Chang, C. Feschotte, Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* **33**, 1098–1116 (2019).

84. D. C. Hardie, P. D. Hebert, Genome-size evolution in fishes. *Can. J. Fish. Aquat. Sci.* (2011), doi:10.1139/f04-106.

85. D. J. R. J. Rees, F. D. Dufresne, H. G. Glémet, C. B. Belzile, Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. *Genome* (2007), doi:10.1139/G06-155.

86. F. Dufresne, N. Jeffery, A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Res.* **19**, 925–938 (2011).

87. G. Lertzman-Lepofsky, A. Ø. Mooers, D. A. Greenberg, Ecological constraints associated with genome size across salamander lineages. *Proc. R. Soc. B Biol. Sci.* **286**, 20191780 (2019).

88. A. E. Vinogradov, Larger genomes for molluskan land pioneers. *Genome*. **43**, 211–212 (2000).

89. S. A. Miller, D. D. Dykes, H. F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16**, 1215 (1988).

90. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

91. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: Quality assessment tool for genome assemblies. *Bioinforma. Oxf. Engl.* **29** (2013), doi:10.1093/bioinformatics/btt086.

92. F. Simão, R. Waterhouse, P. Ioannidis, E. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma. Oxf. Engl.* **31** (2015), doi:10.1093/bioinformatics/btv351.

93. R. M. Waterhouse, M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, E. M. Zdobnov, BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

94. D. R. Laetsch, M. L. Blaxter, BlobTools: Interrogation of genome assemblies. *F1000Research*. **6**, 1287 (2017).

95. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

96. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*. **14**, 587–589 (2017).

97. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

98. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

99. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. **19**, 153 (2018).

100. F. Otto, DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. *Methods Cell Biol.* **33**, 105–110 (1990).

101. J. Dolezel, P. Binarova, S. Lucretti, Analysis of Nuclear DNA content in plant cells by Flow cytometry. *Biol Plant*. **31**, 113–120 (1989).

102. T. R. Ranallo-Benavidez, K. S. Jaron, M. C. Schatz, GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11** (2020), doi:10.1038/s41467-020-14998-3.

103. T. Schell, B. Feldmeyer, H. Schmidt, B. Greshake, O. Tills, M. Truebano, S. D. Rundle, J. Paule, I. Ebersberger, M. Pfenninger, An Annotated Draft Genome for Radix auricularia (Gastropoda, Mollusca). *Genome Biol. Evol.* **9**, 585–592 (2017).

104. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer International Publishing, ed. 2, 2016; https://www.springer.com/de/book/9783319242750), *Use R!*

105. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

106. P. Novák, P. Neumann, J. Pech, J. Steinhaisl, J. Macas, RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. **29**, 792–793 (2013).

107. P. Novák, L. Ávila Robledillo, A. Koblížková, I. Vrbová, P. Neumann, J. Macas, TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111–e111 (2017).

108. S. Negm, A. Greenberg, A. M. Larracuente, J. S. Sproul, *bioRxiv*, in press, doi:10.1101/2020.05.22.111252.

109. C. Goubert, L. Modolo, C. Vieira, C. ValienteMoro, P. Mavingui, M. Boulesteix, De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biol. Evol.* **7**, 1192–1205 (2015).

110. B. L. Cantarel, I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, M. Yandell, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

111. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. **12**, 491 (2011).

112. S. Luo, M. Tang, P. B. Frandsen, R. J. Stewart, X. Zhou, The genome of an underwater architect, the caddisfly Stenopsyche tienmushanensis Hwang (Insecta: Trichoptera). *GigaScience*. **7** (2018), doi:10.1093/gigascience/giy143.

113. B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, J. R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

114. T. M. Lowe, P. P. Chan, tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).

115. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).

116. S. Götz, J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talón, J. Dopazo, A. Conesa, High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).

117. M. S. Barker, K. M. Dlugosch, L. Dinh, R. S. Challa, N. C. Kane, M. G. King, L. H. Rieseberg, *Evol. Bioinforma.*, in press, doi:10.4137/EBO.S5861.

118. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).

119. Z. Yang, PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

120. L. Cui, P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, C. W. dePamphilis, Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).

121. T. Benaglia, D. Chauveau, D. R. Hunter, D. S. Young, mixtools: An R Package for Analyzing Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2009).

122. G. B. Wiggins, W. Wichard, Phylogeny of Pupation in Trichoptera, with Proposals on the Origin and Higher Classification of the Order. *J. North Am. Benthol. Soc.* **8**, 260–276 (1989).

123. evolution and classification of Trichoptera. I. The groundplan of Trichoptera, (available at https://agris.fao.org/agris-search/search.do?recordID=US201301471380).

124. H. H. Ross, Evolution and classification of the mountain Caddis-flies, 21.

125. A.-M. Waldvogel, A. Wieser, T. Schell, S. Patel, H. Schmidt, T. Hankeln, B. Feldmeyer, M. Pfenninger, The genomic footprint of climate adaptation in Chironomus riparius. *Mol. Ecol.* **27**, 1439–1456 (2018).

126. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

127. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12 (2011).

128. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

129. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

130. J. Ruan, H. Li, Fast and accurate long-read assembly with wtdbg2. *Nat. Methods*. **17**, 155–158 (2020).

131. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

132. R. Vaser, I. Sovic, N. Nagarajan, M. Sikic, *Genome Res.*, in press, doi:10.1101/gr.214270.116.

133. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*. **12**, 733–735 (2015).

134. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

135. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*. **9**, e112963 (2014).

136. M. J. Roach, S. A. Schmidt, A. R. Borneman, Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*. **19**, 460 (2018).

137. A. V. Zimin, G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, J. A. Yorke, The MaSuRCA genome assembler. *Bioinformatics*. **29**, 2669–2677 (2013).

138. A. V. Zimin, D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J. Dvořák, S. L. Salzberg, Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).

139. D. Antipov, A. Korobeynikov, J. S. McLean, P. A. Pevzner, hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*. **32**, 1009–1015 (2016).

140. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013) (available at http://arxiv.org/abs/1303.3997).

141. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. **27**, 764–770 (2011).

142. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

143. J. Greilhuber, J. Dolezel, M. A. Lysák, M. D. Bennet, The Origin, Evolution and Proposed Stabilization of the Terms ‘Genome Size’ and ‘C-Value’ to Describe Nuclear DNA Contents. *Ann. Bot.* **95**, 255–260 (2005).

144. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

145. H. Weigand, M. Weiss, H. Cai, Y. Li, L. Yu, C. Zhang, F. Leese, Fishing in troubled waters: Revealing genomic signatures of local adaptation in response to freshwater pollutants in two macroinvertebrates. *Sci Total Environ*. **633**, 875–891 (2018).

146. L. Ferguson, F. Marlétaz, J.-M. Carter, W. R. Taylor, M. Gibbs, C. J. Breuker, P. W. H. Holland, Ancient Expansion of the Hox Cluster in Lepidoptera Generated Four Homeobox Genes Implicated in Extra-Embryonic Tissue Formation. *PLOS Genetics*. **10**, e1004698 (2014).

147.H. Weigand, M. Weiss, H. Cai, Y. Li, L. Yu, C. Zhang, F. Leese, Deciphering the origin of mito-nuclear discordance in two sibling caddisfly species. *Molecular Ecology*. **26**, 5705–5715 (2017).

Acknowledgments

The authors thank Ralph Holzenthal for providing illustrations of larval Trichoptera and

The structures they build.

Funding

This work is a result of the LOEWE-Centre for Translational Biodiversity Genomics funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK) that supported JH and SUP, as well as internal funds of Senckenberg Research Institute provided to JP. JSS was supported by an NSF Postdoctoral Research Fellowship in Biology (DBI-1811930) and an NIH General Medical Sciences award (R35GM119515) to AML. Sequencing was, in part, supported by BYU start-up funds to PBF.

Author contributions:

Conceptualization –JH, JSS, PBF, SUP

Data curation – JH

Formal Analysis - JH, JM, JP, JSS, PBF, ZL

Funding acquisition - SUP, PBF

Investigation – JH, JP, JSS, PBF, ZL

Methodology – AML, JSS, JP, JVS, PBF

Project administration – SUP

Resources – JP, MB, PBF, SUP

Visualization - JH, JSS

Writing – original draft – JH, JSS, PBF, ZL

Writing – review & editing – AML, JH, JM, JP, JSS, JVS, MB, PBF, SUP, ZL

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: This project has been deposited at NCBI under BioProject ID: PRJNA558902.

Figures

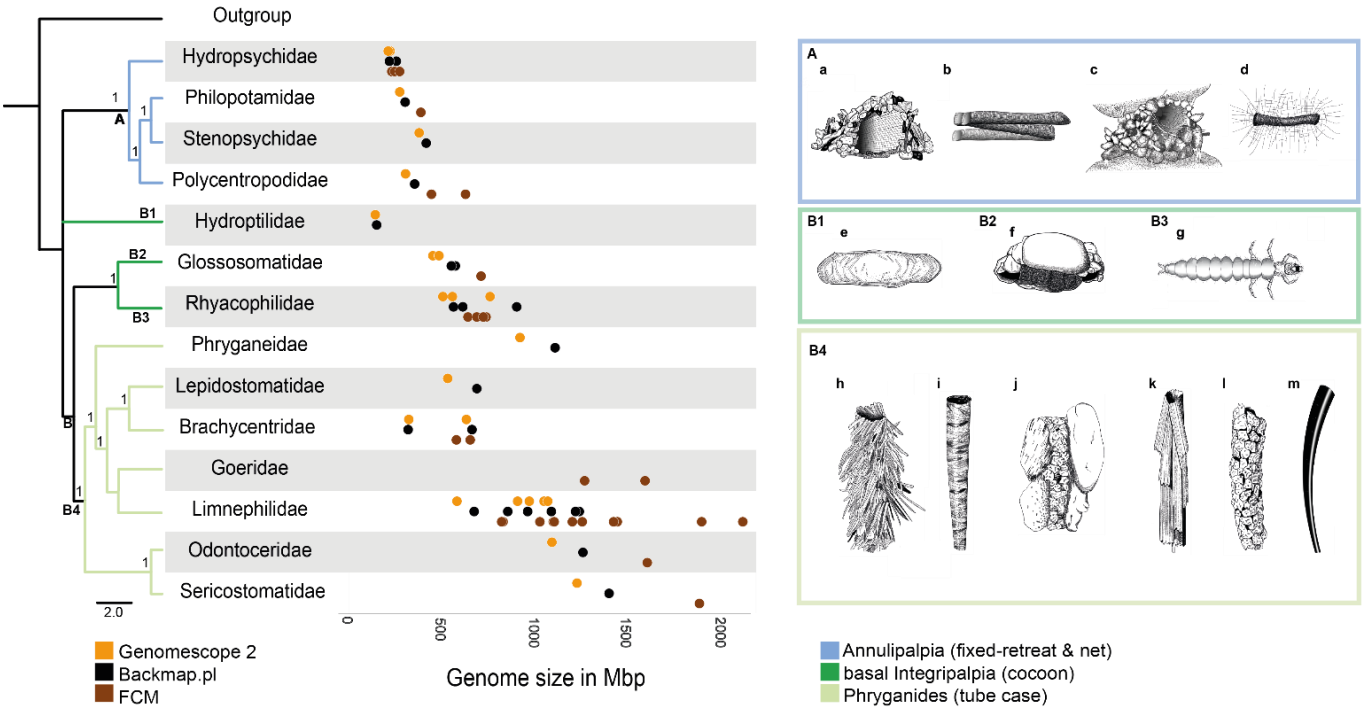

Fig. 1: Ecological diversity (right) and genome size (left) in caddisflies. Phylogenetic relationships derived from ASTRAL-III analyses using single BUSCO genes. Goeridae, which was not included in the BUSCO gene set, was placed according to (52). ASTRAL support values (local posterior probabilities) higher than 0.9 are given for each node. Nodes with less than 0.9 support were collapsed. Taxa were collapsed to family level. Trichoptera are divided into two suborders: Annulipalpia ('fixed retreat- and net-spinners', clade A: blue) and Intergripalpia (clade B: green) which includes basal Integripalpia ('cocoon-builders', clades B1-B3, dark green) and Phryganides or 'tube case-builders' (clade B4: light green). 'Cocoon-builders' are divided into 'purse case'- (clade B1), 'tortoise case-building' (clade B2) and 'free-living' (clade B3) families. Genome size estimates based on different methods (Genomescope2: orange, Backmap.pl: black, Flow Cytometry (FCM): brown) are given for various caddisfly families. Each dot corresponds to a mean estimate of a species. For detailed information on the species and number of individuals used in each method see Data S1. Colors and clade numbers in the phylogenetic tree refer to colored boxes with illustrations. The following species are illustrated by Ralph Holzenthal: a: *Hydropsyche sp.* (Hydropsychidae); b: *Chimarra sp.* (Philopotamidae); c: *Stenopsyche sp.* (Stenopsychidae); d: *Polycentropus* sp. (Polycentropodidae); e: *Agraylea sp.* (Hydroptilidae); f: *Glossosoma sp.* (Glossosomatidae); g: *Rhyacophila sp.* (Rhyacophilidae); h: *Fabria inornata* (Phryganeidae); i: *Micrasema* sp. (Brachycentridae); j:*Goera fuscula* (Goeridae); k: *Sphagnophylax meiops* (Limnephilidae); l: *Psilotreta sp.* (Odontoceridae), m: *Grumicha grumicha* (Sericostomatidae).

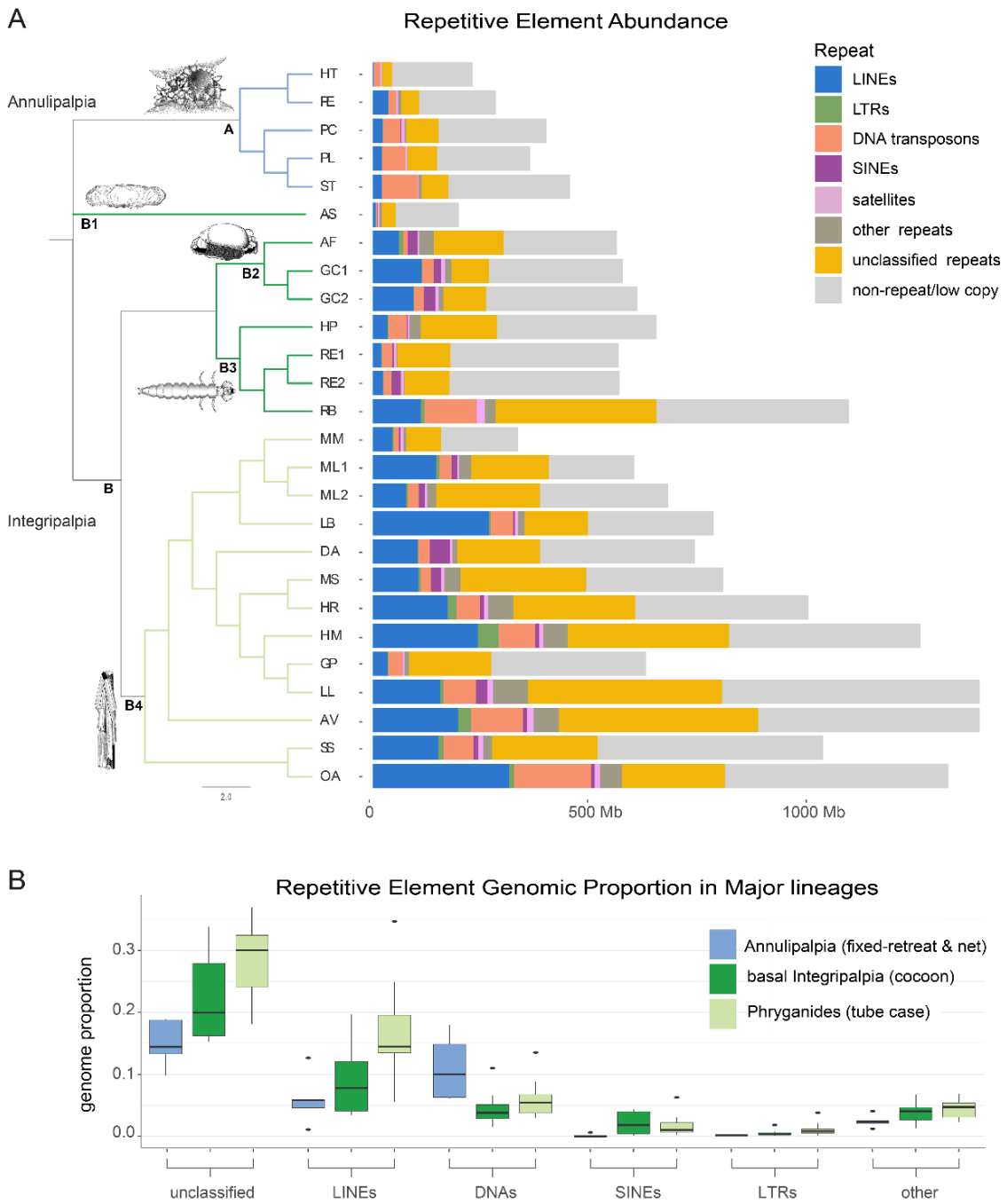

Fig. 2: Repeat abundance and classification in 26 caddisfly genomes. Number of bp for each repeat type is given for each caddisfly genome. A: Repeat abundance and classification. Phylogenetic tree was reconstructed with ASTRAL-III using single BUSCO genes from the genome assemblies. Nodes with ASTRAL support below 0.9 were collapsed. Species names corresponding to the abbreviations in the tree can be found in Table 1. Trichoptera are divided into two suborders: Annulipalpia ('fixed retreat- and net-spinners', clade A: blue) and Intergripalpia (clade B: green) which includes basal Integripalpia ('cocoon-builders', clades B1-B3, dark green) and Phryganides or 'tube case-builders' (clade B4: light green). 'Cocoon-builders' are divided into 'purse case'- (clade B1), 'tortoise case-building' (clade B2) and 'free-living' (clade B3) families. An illustration of a representative of each clade is given. The "other_repeats" category includes: rolling-circles, *Penelope*, low-complexity, simple repeats, and small RNAs. B: Boxplots summarizing shifts in the genomic proportion of RE categories in major Trichoptera lineages.

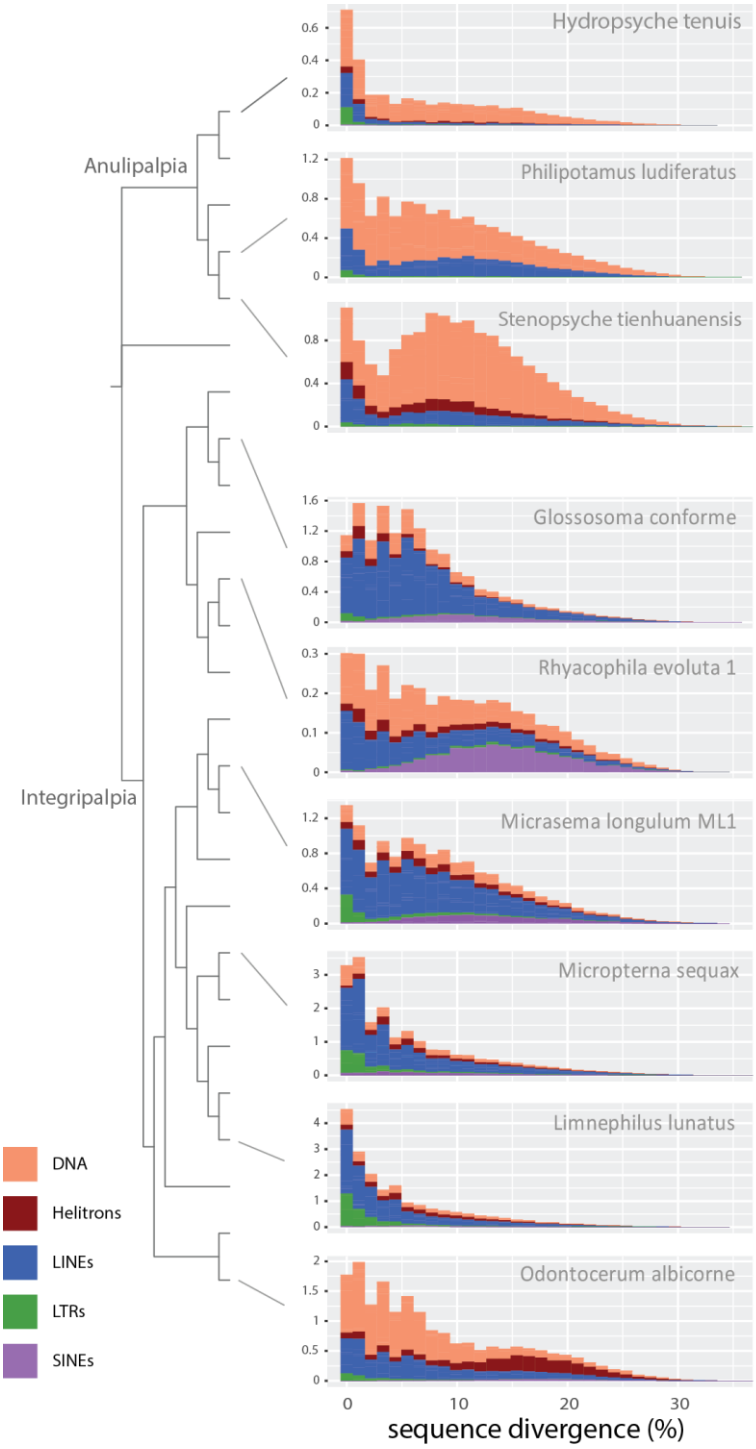

Fig. 3: Transposable element age distribution landscapes. Representative examples are chosen from major Trichoptera lineages. The y-axis shows TE abundance as a proportion of the genome (e.g., 1.0 = 1% of the genome). The x-axis shows sequence divergence relative to TE consensus sequences for major TE classes. TE classes with abundance skewed toward the left (i.e., low sequence divergence) are inferred to have a recent history of diversification relative to TE classes with right-skewed abundance. Plots were generated in dnaPipeTE. Plots for all species are shown in supplementary Fig. 4. For tip labels of the phylogenetic tree see Fig. 2.

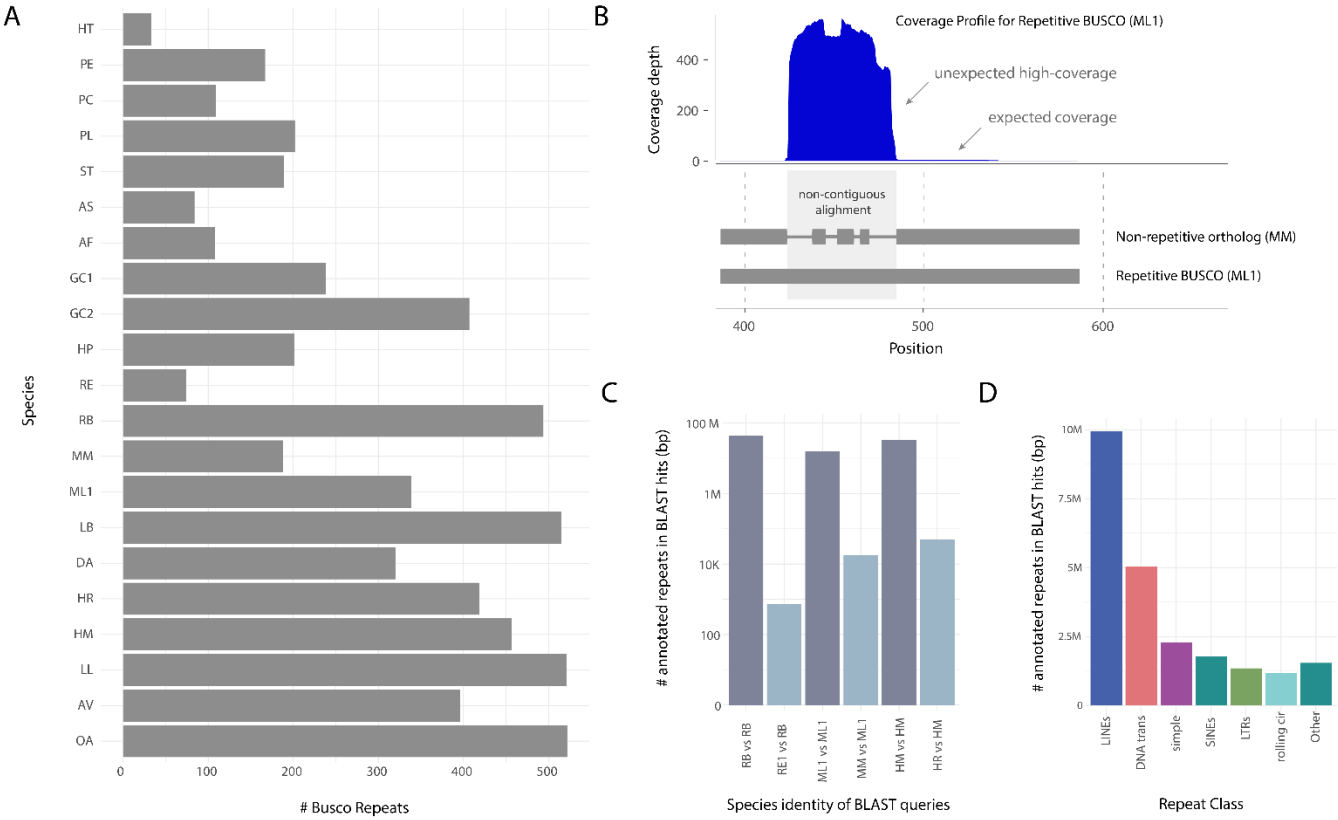

Fig. 4: TE-BUSCO-gene associations in Trichoptera species. (A) Raw abundance of BUSCO with repetitive fragments present in the assembly out of a total of 2442BUSCOs studied in the OrthoDB 9 Endopterygota dataset. (B) Upper plot: An example of a copy number profile of a repetitive BUSCO gene (BUSCO EOG090R02Q9 from ML1 ('inflated species')) which shows unexpected high coverage in the second exon putatively due to the presence of a TE-derived sequence fragment. Lower plot: alignments of repetitive BUSCOs to non-repetitive orthologs taken from a closely related species ('reference species') are consistently non-contiguous in regions of inflated coverage in the repetitive BUSCO. (C) Summary of total bases annotated as REs obtained from each of two BLAST searches. First, when repetitive BUSCOs are BLASTed against an assembly for the same species BLAST hits included megabases of annotated repeats (dark plots). Second, when non-repetitive orthologs of the repetitive BUSCOs in the first search are taken from a close relative and BLASTed against the inflated species, there is a dramatic drop in BLAST hits annotated as REs. Note log scale on the y-axis. (D) Summary of annotations for BLAST hits for classified REs when repetitive BUSCOs are BLASTed against an assembly of the same species.

Table 1. Comparison of assembly and annotation statistics of all available Trichoptera Genomes. *Assemblies produced in this study. **$N_{Arthropoda}$=2442, % of present BUSCOS = complete + fragmented.

Species	Abbre-viation	Accession number	Length (bp)	N50 (kbp)	No. of contigs	BUSCOS %**
Agapetus fuscipes*	AF	JAGTXP000000000	552,637,417	2.8	291,536	75.5
Agraylea sexmaculata*	AS	JAGTTH000000000	196,044,125	86	7,050	96.5
Agrypnia vestita (30)	AV	JADDOH000000000	1,352,945,503	111.8	25,541	92.8
Drusus anulatus*	DA	JAGWCC000000000	727,941,535	1,043.7	2,401	96.8
Glossosoma conforme*	GC1	JAGTXR000000000	568,249,599	2,212.1	653	91.8
Glossosoma conforme (145)	GC2	GCA_003347265.1	604,293,666	17.1	119,821	92.1
Glyphotaelius pellucidula (146)	GP	Glyphotaelius_pellucidus_k51_scaffolds	623,431,006	1.6	461,749	68.6
Halesus radiatus*	HR	SUB9540325	973,361,729	125.2	12,483	90.6
Himalopsyche phryganea*	HP	JAGVSL000000000	633,785,554	4,634	710	97.8
Hesperophylax magnus (30)	HM	JADDOG000000000	1,275,967,528	768.2	6,877	95.2
Hydropsyche tenuis (57)	HT	GCA_009617725.1	229,663,394	2,190.1	403	97.6
Lepidostoma basale*	LB	JAGTTH000000000	769,209,094	1,052	1,620	97.0
Limnephilus lunatus (63)	LL	GCA_000648945.2	1,369,180,260	69.1	58,718	91
Micrasema longulum*	ML2	JAGXCS000000000	668,638,047	2.5	368,343	69.8
Micrasema longulum*	ML1	JAGVSM000000000	585,245,295	170.5	5,451	84.5
Micrasema minimum*	MM	JAGVSQ000000000	329,257,313	69.5	7,561	67.1
Micropterna sequax*	MS	JAGUCF000000000	778,696,759	7.9	144,286	68.9
Odontocerum albicorne*	OA	JAGTXQ000000000	1,305,984,461	266.4	9,303	95.9
Parapysche elsis*	PE	JAGVSN000000000	282,185,525	5,591.7	159	97.4
Philopotamus ludificatus*	PL	JAGXCT000000000	360,302,175	67.4	37,275	92.6
Plectrocnemia conspersa (57)	PC	GCA_009617715.1	396,695,105	869	1,614	97.8
Rhyacophila brunnea*	RB	JAGYXB000000000	1,086,880,777	1,043.7	2,105	97.3
Rhyacophila evoluta*	RE2	JAGVSQ000000000	565,830,460	9.9	114,057	91.7
Rhyacophila evoluta*	RE1	JAGVSO000000000	562,550,625	9.7	111,706	91.7
Sericostoma sp. (147)	SS	GCA_003003475.1	1,015,727,762	3.2	561,698	60.3
Stenopsyche tienhuanesis (112)	ST	GCA_008973525.1	451,494,475	1,296.7	552	97.6