Repetitive elements in the era of biodiversity genomics: insights from 600+ insect genomes

John S. Sproul[1,2,*], Scott Hotaling[3,4,*], Jacqueline Heckenhauer[5,6,*], Ashlyn Powell[7], Amanda M. Larracuente[2], Joanna L. Kelley[3], Steffen U. Pauls[5,6,8], and Paul B. Frandsen[5,7,9]

Affiliations:
[1] Department of Biology, University of Nebraska Omaha, Omaha, NE, USA
[2] Department of Biology, University of Rochester, Rochester, NY, USA
[3] School of Biological Sciences, Washington State University, Pullman, WA, USA
[4] Department of Watershed Sciences, Utah State University, Logan, UT, USA
[5] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany
[6] Department of Terrestrial Zoology, Entomology III, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Germany
[7] Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, USA
[8] Department of Insect Biotechnology, Justus-Liebig-University Gießen, Germany
[9] Data Science Lab, Smithsonian Institution, Washington, DC, USA
* Contributed equally

Correspondence:
John S. Sproul, Department of Biology, University of Nebraska Omaha, Omaha, NE, 68182, USA; Email: johnssproul@gmail.com; ORCID: 0000-0002-6747-3537

Keywords: genome biology, transposable elements, Insecta, arthropod genomics, long-read sequencing, repeats

Abstract:
Repetitive elements (REs) are integral to the composition, structure, and function of eukaryotic genomes. Yet, RE dynamics remain understudied in many taxonomic groups, preventing holistic understanding of how genomes and species evolve. Here, we investigated REs across 601 insect species (20 orders) to better understand the RE landscape of insects and to evaluate automated RE annotation methods in the era of biodiversity genomics. We identified wide variation in the types and frequency of REs across insect groups. We quantified associations between REs and protein-coding genes and found an elevated frequency of associations in insects with abundant long interspersed nuclear elements (LINEs). Sequencing technology impacts RE detection; ~36% more REs could be identified in long-read versus short-read assemblies. Long terminal repeats (LTRs) showed markedly improved detection in long-read assemblies (162% more), while DNA transposons and LINEs showed less respective technology-related bias. We illustrate fundamental challenges to efficient study of REs in diverse groups, showing that in most insect lineages, 25–85% of repetitive sequences were "unclassified" compared to only ~13% of unclassified repeats in *Drosophila* species. Our findings suggest this RE-annotation bottleneck, driven largely by uneven taxonomic representation in RE reference databases, is worsening. Although the diversity of available insect genomes has rapidly expanded, the rate of community contributions to RE databases (essential for RE annotation) has not kept pace, preventing high resolution study of REs in most groups. We highlight the tremendous opportunity and need for the field of biodiversity genomics to embrace REs and suggest collective steps for making progress towards this goal.

Introduction:
Repetitive elements (REs) comprise large proportions of eukaryotic genomes and are fundamental to the evolutionary process (1, 2). Broadly, REs can be classified as *interspersed* or *tandem repeats*. Interspersed repeats include transposable elements (e.g., retrotransposons) which encode for proteins that facilitate their movement and proliferation in genomes. Tandem repeats (e.g., satellite DNAs) can form large blocks (e.g., megabases) of relatively short non-coding sequences in repeated arrays (reviewed in 3). Together, interspersed and tandem repeats comprise ~67% of the human genome (4). Despite their major genomic footprint, REs are understudied in genome science due to a history of technical challenges associated with their sequencing and assembly (5, 6); however, long-read sequencing is ameliorating this challenge through dramatic improvements in genome assembly contiguity (7).

Although understudied, REs can play critical roles in the organization, stability, regulation, and evolution of genomes (1, 8). At broad scales, REs shape chromatin domains across chromosomes and drive the three-dimensional organization of DNA (9, 10). Rapidly evolving blocks of REs are common sites of recombination and chromosome rearrangements (e.g., 11, 12). At finer scales, shifts in RE location and abundance (e.g., through transposition of retrotransposons) can alter gene expression and phenotype evolution (13–15). Across evolutionary scales, rapid RE evolution (e.g., tandem repeats) is associated with hybrid incompatibilities between species (16–18). In short, REs exhibit an array of structural and evolutionary effects on genome evolution across species.

We have entered the era of biodiversity genomics with availability and quality of genome assemblies rapidly increasing in plants and animals (7, 19, 20). As a critical mass of assemblies accumulates within a group, phylogenetically informed meta-analyses of REs can illuminate their impact on genome dynamics and evolution (2). With more than 1 million described species, insects account for the bulk of Earth's animal biodiversity (21). While some insects have become model genetic organisms (e.g., *Drosophila melanogaster*) and thus considerable attention is devoted to many aspects of their genome biology, including REs (e.g., 22, 23), for the vast majority of insects, repetitive genomic components remain largely unexplored.

In this study, we analyzed the RE landscape in genome assemblies of more than 600 insect species spanning ~400 million years of evolution (24). We used this dataset to gain a broad view of RE dynamics in insects and to assess how sequencing technology and taxonomic representation in reference databases (e.g., Repbase, 25) shapes our ability to identify and classify REs with widely used automated RE annotation tools in a comparative framework. Given the potential for REs to impact protein-coding genes (e.g., through epigenetic silencing of adjacent DNA sequences), we also investigated the frequency of associations between REs and protein-coding genes. Our findings yield new insight into the RE landscape of insect genomes from a much wider taxonomic perspective than previous analyses. Beyond insects, we highlight the opportunities and obstacles for investigating RE dynamics in biodiversity studies with emphasis on RE annotation bottlenecks. We conclude by describing the ways in which the biodiversity genomics community can alleviate challenges of RE annotation (e.g., RE database curation and taxonomy) to build towards a more holistic understanding of genomic natural history and evolution.

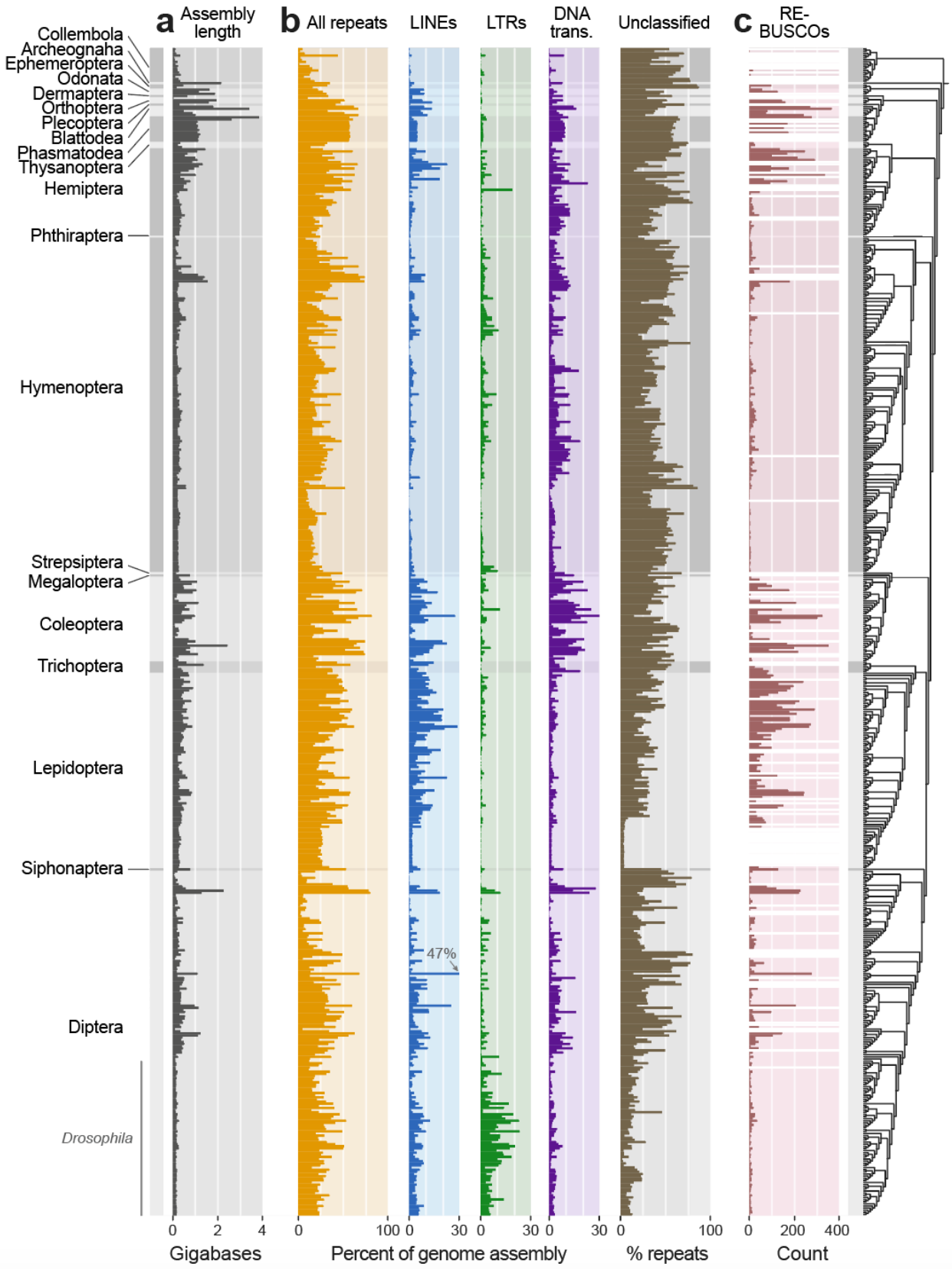
a
Assembly length
b
All repeats
LINEs
LTRs
DNA trans.
Unclassified
c
RE-BUSCOs
Collembola
Archeognaha
Ephemeroptera
Odonata
Dermaptera
Orthoptera
Plecoptera
Blattodea
Phasmatodea
Thysanoptera
Hemiptera
Phthiraptera
Hymenoptera
Strepsiptera
Megaloptera
Coleoptera
Trichoptera
Lepidoptera
Siphonaptera
Diptera
Drosophila
47%
0
2
4
Gigabases
0
100
0
30
0
30
0
30
Percent of genome assembly
0
100
% repeats
0
200
400
Count

Figure 1. The repetitive element landscape of insects. Left bars with alternating shades of gray indicate taxonomic boundaries and track across plots. (a) Total genome assembly length. (b) Overall RE abundance followed by LINEs, LTRs, and DNA transposons all as a percent of the overall genome assembly. One species (*Hermetia illucens*) exceeded the scale for LINEs (indicated at 47%). Any REs that could not be classified ("unclassified") are shown as a percentage of all repeats identified for a given species. (c) Abundance of RE-associated BUSCOs. For (a) and (b), all assemblies with BUSCO completeness ≥ 50% were included (n = 548). For (c), because we were concerned that BUSCO completeness would alter our capacity to detect RE-associated BUSCOs, we only included assemblies with BUSCO completeness ≥ 90% (n = 493). Assemblies that were excluded in (c) are indicated with white bars. To the right of the plots, the phylogeny inferred in this study that was used to place species in their phylogenetic context is shown.

Results:

We assessed RE content for genome assemblies of 601 insect species across a total of 20 orders. Of the 601 assemblies, 548 and 441 assemblies had BUSCO completeness ≥ 50% and ≥ 90%, respectively. We report results for three data sets: all assemblies, ≥ 50% BUSCO, and ≥ 90% BUSCO. For clarity, we used the "all assemblies" data set for analyzing taxonomic representation in Repbase, the "≥ 50%" data set for assessing overall RE trends in insects, and the "≥ 90%" for all other analyses.

The proportion of REs in insect genomes ranged widely from 1.6–81.5% (mean = 30.8%; Figs. 1 & 2). Based on mean genomic proportion of specific RE categories, DNA transposons were the most abundant overall and particularly so in Coleoptera (Figs. 1, 2a–d), yet conspicuously uncommon in Lepidoptera. LINEs were the next most abundant RE type and exhibited wide variance across and within orders (Figs. 1, 2a–d). For example, within Diptera, LINE abundances ranged from ~1% in 29 species to ~47% in *Hermetia illucens* (Fig. 1b). However, LINEs were notably uncommon in Hymenoptera (1.8% ± 1.7% genomic proportion, n = 157 species). LTRs were generally uncommon but were abundant within *Drosophila* (order Diptera; Fig. 1b). Since LTRs are particularly difficult to identify due to their size and complexity (26), and because *Drosophila* LTRs are better characterized in RE databases than other insect lineages (see below), this trend may reflect methodology more than biology. Consistent with previous studies (27, 28), we showed that RE abundance correlates with genome size (Fig. 2j). SINEs showed greatest abundance in lepidopterans and some dipterans, tandem repeats were most common in Hymenoptera and Diptera, while "Other" repeats were abundant in Lepidoptera reflecting the high number of Helitrons in some lineages (Fig. S1).

Sequencing technology influences the study of REs. In insects, long-read assemblies are on average ~48 times more contiguous than short-read technologies (7). For REs, we identified 36.1% more REs in long-read assemblies versus short-reads (Welch two-sample T-test, P = 0.04; Fig. 2f). Furthermore, this difference in total REs identified was not due to assembly length, which did not differ between technologies (Welch two-sample T-test, P = 0.42; Fig. 2e). Long-read assemblies had the greatest influence on LTR detection (162% increase, Welch two-sample T-test, P < 0.001; Fig. 2h), followed by DNA transposons (47% increase; Welch two-sample T-test, P = 0.03; Fig. 2i). Although LINEs showed increased average detection in long-read assemblies the difference was not significant (Welch two-sample T-test, P = 0.42; Fig. 2g). These trends set a general expectation for sequencing technology-related bias, with LTRs being under-detected in short-read assemblies, whereas DNA transposons, LINEs, and other RE classes, show moderate/low sensitivity to sequencing technology (Fig. 2f-i).

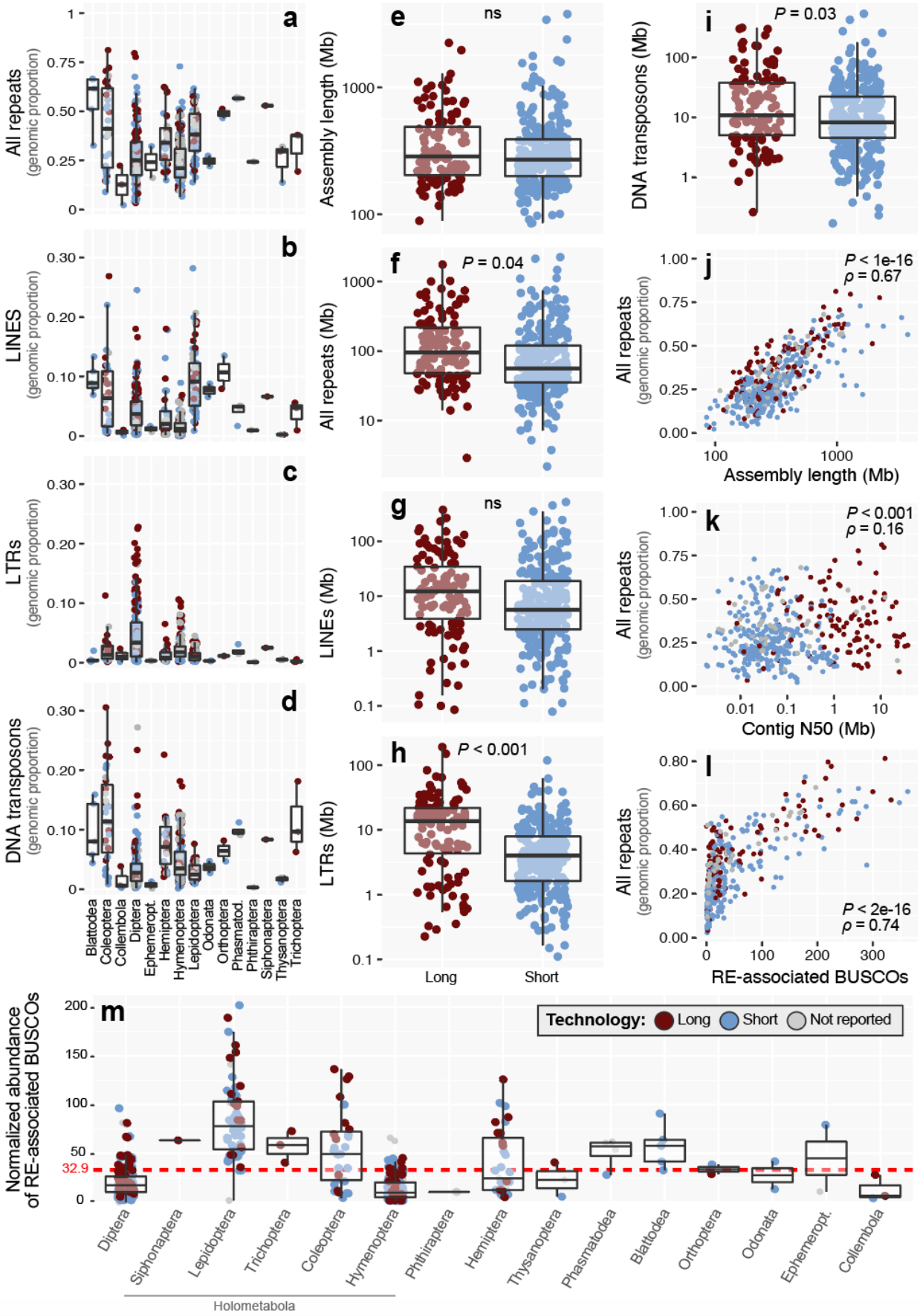
a
All repeats (genomic proportion)
b
LINES (genomic proportion)
c
LTRs (genomic proportion)
d
DNA transposons (genomic proportion)
Blattodea
Coleoptera
Collembola
Diptera
Ephemeropt.
Hemiptera
Hymenoptera
Lepidoptera
Odonata
Orthoptera
Phasmatod.
Phthiraptera
Siphonaptera
Thysanoptera
Trichoptera
e
ns
Assembly length (Mb)
f
P = 0.04
All repeats (Mb)
g
ns
LINEs (Mb)
h
P < 0.001
LTRs (Mb)
Long
Short
i
P = 0.03
DNA transposons (Mb)
j
P < 1e-16
ρ = 0.67
All repeats (genomic proportion)
Assembly length (Mb)
k
P < 0.001
ρ = 0.16
Contig N50 (Mb)
l
P < 2e-16
ρ = 0.74
RE-associated BUSCOs
m
Technology: Long Short Not reported
Normalized abundance of RE-associated BUSCOs
32.9
Holometabola

Figure 2. (a-d) Genomic proportion of all repeats (a), LINEs (b), LTRs (c), and DNA transposons (d) across insect orders in the data set. Note: To improve visualization, y-axis scales differ between (a) and (c-d). (e-i) Sequencing technology comparisons for (e) assembly length, (f) all repeats, (g) LINEs, (h) LTRs, and (i) DNA transposons. Significance was assessed with Welch two-sample T-tests. ns: not significant at $P < 0.05$. (j-l) Spearman correlations between genomic proportion of repeats versus (j) assembly length, (k) contig N50, (l) and number of RE-associated BUSCO genes. (m) Normalized abundances of RE-associated BUSCOs across orders and organized by the phylogeny shown in Fig. 1. For all plots, log-transformed data were used for visualization while statistics were performed on the untransformed data.

As a surrogate measure for associations between REs and protein-coding genes, we quantified RE presence in BUSCO genes (termed hereafter RE-associated BUSCOs) following Heckenhauer et al. (27). RE-associated BUSCOs increased with assembly length (Spearman's correlation = 0.74, $P < 2e\text{-}16$; Fig. 2i). However, assembly length alone did not explain increased RE-associated BUSCO abundance. For example, Lepidoptera and Coleoptera species had 5.8- and 4.4-fold respective average increases in RE-associated BUSCOs compared to Hymenoptera after correcting for assembly length (Fig. 2m). Overall, RE-associated BUSCOs were most abundant in species with high proportions of LINEs (e.g., Hemiptera, Blattodea, Coleoptera, Trichoptera, Lepidoptera) (Figs. 1, 2m, and S2). In some lineages (e.g., some Blattodea, Coleoptera, and Hemiptera), RE sequences were detected in upwards of 25% of all BUSCO genes, while RE-associated BUSCOs averaged ~1-2% of all BUSCOs in Hymenoptera and Diptera. To address whether general trends in RE-associated BUSCOs could be driven by an artifact of assembly errors (which might simply be more numerous in larger assemblies), rather than true associations between REs and BUSCO genes, we predicted that less contiguous, short-read assemblies would show inflated RE-BUSCO associations compared to more contiguous, long-read assemblies. However, this comparison revealed the opposite pattern; RE-associated BUSCOs are ~60% more common in long-read assemblies (Welch two-sample T-test, $P = 0.007$; Fig. S3).

Since most RE annotation relies on reference databases (i.e., Repbase, Dfam, 29, 30), we expected bias in database representation to impact our RE annotation. The proportion of unclassified REs in a given assembly increased with its genetic distance from *D. melanogaster* (Spearman's correlation = 0.4, $P < 2e\text{-}16$; Fig. 3a). For reference, unclassified repeats comprised only 13.1% of all repeats in the 71 *Drosophila* species but accounted for 40.5% total repeats on average in all other taxa. High fractions of unclassified repeats were especially evident in poorly sampled, early diverging insect orders. For example, in Thysanoptera and Ephemeroptera, 72.0% and 85.1% of respective REs are unclassified despite having similar genomic proportions of REs as *Drosophila* (~25% in all three groups; Fig. 3b).

To clarify the impact of uneven taxonomic representation in reference databases on RE annotation, we quantified representation of insect orders and families in Repbase (25, 30). Repbase is the most widely cited repository of RE sequences and is integral to the standard RE identification and annotation programs RepeatModeler2 (26) and RepeatMasker (31). Of the 20 insect orders in our data set, 14 are represented in Repbase, however, of those six are represented by a single family (Fig. 3d). Of 154 insect families in our data set, just over one-third ($n = 57$) had any representation in Repbase. Taxonomic bias is more extreme when the number of reference sequences is considered. Just two families, Drosophilidae and Culicidae (order Diptera), account for ~60% ($n = 8{,}453$) of all insect sequences in Repbase (Fig. 3d) and ~70% of all LTR sequences ($n = 5{,}908$). Nearly 75% of all insect families in Repbase are represented by fewer than 100 sequences and only four families [Culicidae (Diptera), Drosophilidae (Diptera), Formicidae (Hymenoptera), and Acrididae (Orthoptera)] have > 1000

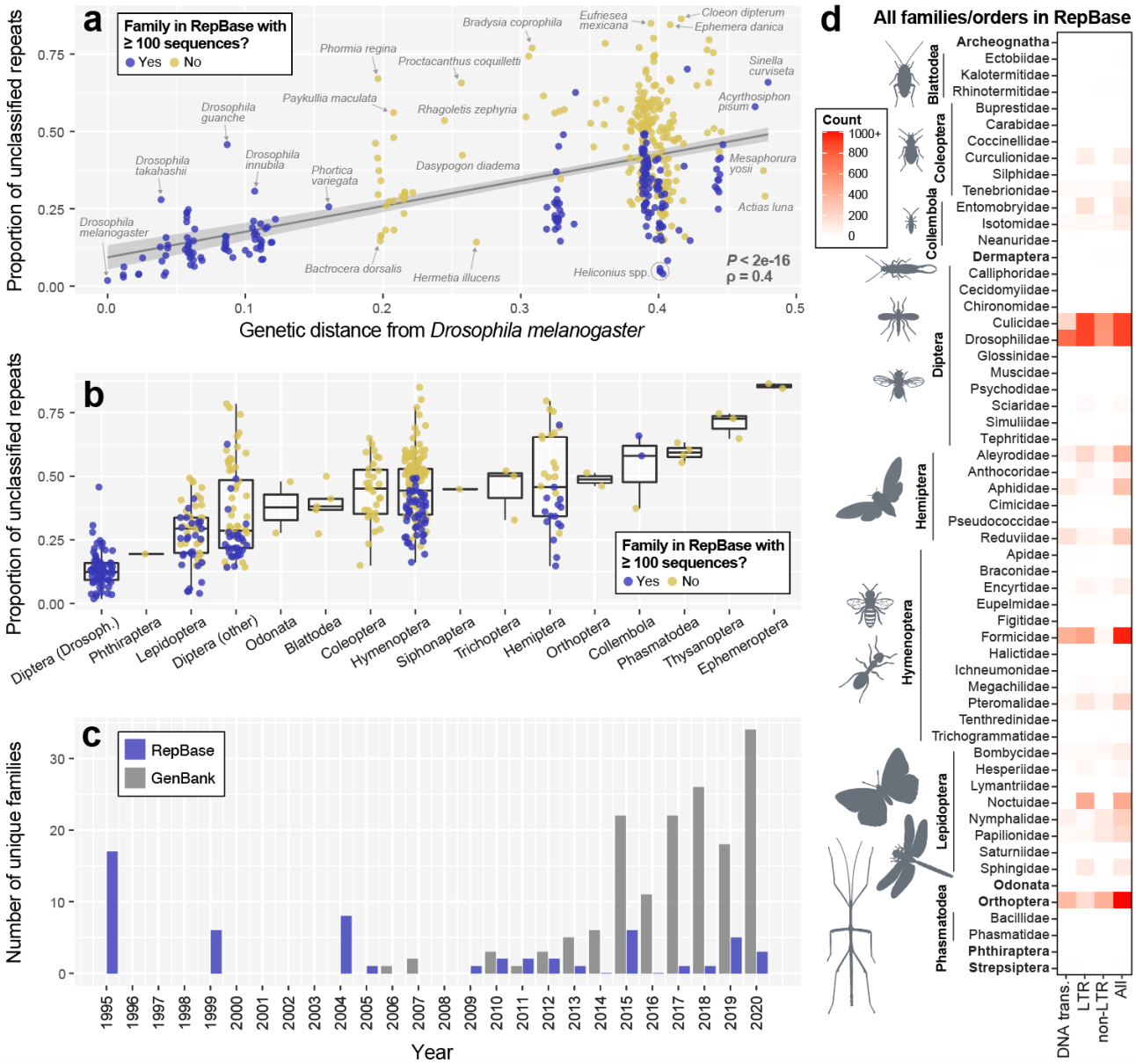

Figure 3. (a) A comparison of the proportion of total repeats that are unclassified in each insect's genome assembly versus its genetic distance from *Drosophila melanogaster*. (b) The same data presented in (a) but grouped by order except for Diptera which are divided into family Drosophilidae and all other Diptera. In both (a) and (b), a "yes" reflects family-level representation of 100 or more sequences in RepBase. (c) Unique insect entries at the family-level submitted to RepBase or GenBank from 1995–2020. Data for GenBank submissions were taken from Hotaling et al. (2021). Of note, for 2020, only GenBank submissions through October 2020 were included. (d) Heatmap showing the abundance (count) of RE sequence entries in RepBase by order (bold) or family. Of the 154 insect families in our dataset, roughly one third, those listed here, have any representation in Repbase. Of those, many are represented by few RE sequences, e.g., essentially white boxes indicate only 1–10 sequences are present. If a single family was present, it is labeled with the broader order name, if two or more families from the same order were present, they are listed with a line encompassing them to the left.

sequences (Fig. 3d). Species belonging to a family represented by ≥ 100 sequences in Repbase had, on average, 24.5% unclassified REs, whereas insects belonging to families

represented by ≤ 99 sequences had nearly double the proportion of unclassified repeats (45.8%). The gap between available genome assemblies for insects and Repbase representation appears to be increasing. Since insect genome assemblies began to proliferate on GenBank in ~2010, submissions to Repbase have not exhibited similar growth (Fig. 3c).

Discussion:

In the present study, we extended previous efforts (e.g., 2, 28) by describing RE dynamics for 600+ insect species. In the process, we evaluated the efficacy of automated RE annotation pipelines in a large, taxonomically diverse data set and clarified expectations for RE annotation success in diverse clades.

REs in insects: new insight from a broad taxonomic comparison

Insects account for over half of all described animal species (21). To understand the genomic basis of this diversity, we must understand repeat evolution, as repeats comprise major fractions of nearly all insect genomes. We identified wide variation in RE abundance both within and among major clades. For example, DNA transposons were generally abundant in most insect orders yet conspicuously uncommon in Lepidoptera. Similarly, LINEs are abundant in many orders (e.g., Coleoptera, Trichoptera, Hemiptera) but largely absent in Hymenoptera (Figs. 1–2). These order-level patterns indicate both deep phylogenetic constraints in RE architecture (e.g., within orders), as well as major shifts between lineages. For example, in Holometabola, LINEs shift from low abundance in Hymenoptera to higher abundance in the next-branching lineages (i.e., Lepidoptera, Trichoptera, Coleoptera), then back to lower abundance in Diptera (Fig. 2m). suggesting shifts in strategies for maintaining genome stability and TE regulation across groups.

Abundant LINEs in several lineages may have broad consequences for phenotype evolution given that RE associations with protein-coding genes (i.e., BUSCOs) are disproportionately abundant in LINE-rich lineages (e.g., LINEs and RE-associated BUSCOs in Lepidoptera versus Hymenoptera; Figs. 1, 2m, S3). Genomes suppress RE activity through epigenetic silencing of repetitive sequences (e.g., heterochromatin formation, 32). Because silent marks may occur near regulatory gene regions and spread to adjacent sequences (33, 34), movement of REs to new genomic loci can have immediate impact on expression of nearby genes. Over longer timescales, RE sequences can be co-opted to form genome-wide regulatory networks of gene expression (13). Thus, our finding of abundant and dynamically evolving RE-gene associations in insects identifies new potential for studying RE impacts on coding regions and phenotype evolution.

Our broad taxonomic sampling illustrates that non-model insects tend to have larger, more repeat-rich genomes than the model species (e.g., *D. melanogaster*) that seeded much of our present knowledge of RE dynamics (Fig. 1a–b). While REs can have both deleterious and adaptive impacts on host genomes (35, 36), their dynamics are understudied in complex, repeat-rich genomes. The few large-genome groups that have been comprehensively studied for REs (e.g., maize) suggest an ecosystem-like environment where REs adopt diverse strategies and impacts within their various niches in the genome (37). Investigating the diversity of insect models with varying genome sizes and complexity can expand our perspectives on genome evolution. For example, in caddisflies (Trichoptera), clades containing large genomes show higher species diversity and ecological breadth than small-genome lineages, raising the potential for adaptive advantages of maintaining high repeat loads (27, 38). With high species diversity and broad distributions in nearly all habitat types, insects may be particularly useful for understanding factors driving temporal dynamics of TE activity including population demographics (39) and environmental stress (40, 41). In addition, a “many-model” phylogenetic

framework offered by insects may be key to connecting patterns of genome size, REs, and developmental constraints with ecological factors (42).

Sequencing technology and RE analysis

Our analyses showed that sequencing technology influences RE detection. Specifically, long-read assemblies contain 36% more REs than short-read assemblies. LINEs and DNA transposons showed low or modest impact from technology differences (e.g., differences in LINEs detection were non-significant, Fig. 2g). This, combined with their overall genomic abundance, even in lineages with poor representation in RE databases (e.g., Coleoptera, Blattodea; Figs. 1, 3d), suggests robustness to both technology differences and limited database representation for LINEs and DNA transposons. By contrast, LTRs showed 162% increase in long-read assemblies. LTRs are difficult to identify with standard approaches due to their length and sequence complexity (26) and this finding suggests technology advances are closing the assembly and annotation gaps for historically problematic elements. Recent studies that report telomere-to-telomere assemblies (e.g., 43) further illustrate the impact of sequencing technology advances in resolving assemblies at repetitive regions. Tandem repeats may now be the last RE type for which assembly remains largely intractable. Although long-read assemblies showed modest gains in tandem repeat detection (~25% increase), large blocks (e.g., megabases) of tandem repeats such as satellite DNAs are common in insects and other groups (3) and will remain unresolved in assemblies for the near future.

Challenges and opportunities for RE biology in biodiversity genomics

We provide an empirical illustration of fundamental challenges that limit thorough RE annotation in all but a few model species and their close relatives. Given the scale of repeats that could not be annotated in many lineages (i.e., unclassified repeats, Figs. 1, 3a–b) we show how deep insights into RE dynamics across phylogenetic scales remain impractical until we can map the finer details of RE landscapes in any species.

To realize the potential that biodiversity genomics offers for the study of REs in insects and beyond, we must be able to efficiently study homologous REs across clades. Two main challenges have slowed progress toward this goal: assembly fragmentation in repetitive regions and comprehensive RE annotation. The rise of long-read sequencing technology has dramatically improved assembly of repetitive regions (e.g., 7) and largely ameliorated this first challenge. This advance has been driven primarily by industry research and incentives paired with buy-in from the genomics community, including consortia (e.g., Earth BioGenome Project, (44). However, advances in RE annotation rely largely on the academic community with fewer financial or related incentives. Although many tools for automated identification and annotation of REs exist (5), annotation tools are limited by the quality of reference databases and specifically the breadth of known REs that can be used to annotate unknown REs in focal assemblies.

As such, community-led RE database curation is not trivial. Two specific obstacles to effective annotation exist: (1) RE taxonomies are in early stages of curation such that redundantly described, or undescribed REs are both common; and (2) taxonomic representation in existing RE databases is woefully incomplete (Fig. 3d). Although these issues have been raised in the RE community for more than a decade (5, 45–47), our results add quantification to an abstract challenge and highlight that despite major progress in biodiversity genomics overall, the RE "database issue" is growing worse rather than improving (Fig. 3c). As it stands, an average of 40.5% of total repeats are unclassified in all non-*Drosophila* taxa, while just 13.1% are unclassified on average in the 71 *Drosophila* species sampled (Fig. 3b). The numbers are much worse in early diverging insect orders such as Thysanoptera and Ephemeroptera (72.0% and

85.1% unclassified, respectively) despite their having similar genomic proportions of REs as *Drosophila* (~25% in all three groups; Fig. 3b). Without a concerted effort to improve RE curation and annotation, we expect unclassified percentages of REs to increase as additional assemblies are sequenced. These problems are not likely specific to insects and present a fundamental challenge that impedes deep understanding of genomes that genomicists seek.

To be clear, we applaud the efforts of many groups that develop, maintain, and curate RE repositories such as Repbase and Dfam (25, 29, 30, 48, 49). We also acknowledge the valuable efforts from research groups studying classical model species (e.g., *Drosophila melanogaster*) whose contributions form a basis of broad understanding about RE biology. As biodiversity genomics continues to grow and diversify, concerted efforts should be made to support RE research and make the importance of their annotation central to broader goals of the field (i.e., similar to generating new genome assemblies or gene annotation tools).

We view biodiversity science as a large-scale solution to many challenges facing RE biology. With a long history of deep expertise in phylogenetics, taxonomy, and specimen acquisition, the infrastructure, experience, and human resources within biodiversity science could be a boon for improving RE taxonomy, curation, and taxonomic representation. However, we emphasize the need for care when embracing this challenge. A primary lesson learned from decades of taxonomy and phylogenetic inference is that thorough taxon sampling is critical to avoid mistakes in both endeavors. Thus, a stable RE taxonomy hinges upon the mapping of REs in taxonomically diverse groups, establishing homology through robust phylogenetic analysis of specific elements within and across groups, *and* submitting curated RE sequences to existing databases (25, 29, 30, 48, 49). In turn, studying REs in diverse clades can offer reciprocal benefits to biodiversity science in that repetitive elements are an underutilized source of signal that can add resolution to evolutionary studies (50–52).

As we move forward in this new era of biodiversity genomics, we need to simultaneously meet the challenge of studying RE dynamics across broad taxonomic scales. To bridge this gap, we offer three key ways for the genomics community to contribute:

> **(1) Embrace RE biology.** Rather than viewing REs as nuisance sequences to be masked (53), seek to understand their interesting and diverse roles in genome biology. Many excellent, accessible reviews exist (e.g., Bourque et al., 2018; Wells & Feschotte, 2020) and more RE literacy and interest will no doubt improve RE science.
>
> **(2) Document REs in new (and existing) genome assemblies.** Whether generating a new genomic resource or evaluating one as a reviewer, editor, or peer, encourage reporting and documentation of REs. This will add to the RE knowledge base and accelerate literacy of both REs and the software tools available for their study.
>
> **(3) Invest in RE library curation and database submission within your area of taxonomic expertise.** To meet the challenge of RE annotation with accelerating availability of genome assemblies, RE library curation and database submission needs to become a mainstream step in data archiving. There are many resources designed to streamline contribution and data sharing, including RE curation guidelines (Goubert et al. 2022), descriptions of Repbase and Dfam databases and submission (48, 54), TE library curation tools (e.g., Ou et al., 2019), and group-specific RE resources (45).

From single, difficult-to-obtain genome assemblies ~20 years ago to dozens of new, highly contiguous assemblies being published every day, an exciting, new discipline of biodiversity

genomics has emerged. By investing in solutions to address bottlenecks for studying REs and any similar challenges, we can build the foundation for unprecedented new understanding of genome biology in insects and across the tree of life.

Materials and Methods:
An extended version of Materials and Methods with additional details of phylogenetic inference, and RE-gene associations, and statistical analysis is provided in Supplemental Materials.

Data acquisition
Following Hotaling et al. (56), we used the assembly-descriptors function in the NCBI datasets command line tool to download metadata for all nuclear genomes available for insects on GenBank (accessed 2 November 2020; (57). We then culled our data set to include only one representative genome per taxon (species or subspecies) by selecting the assembly with the highest contig N50 (the mid-point of the contig distribution where 50% of the genome is assembled into contigs of a given length or longer). Using provided NCBI metadata on the sequencing read technology used for assembly, assemblies were classified as “short-read”, “long-read”, or “not provided” based on whether only short-reads (e.g., Illumina) were used, any amount of long-read sequences (e.g., PacBio) were used, or no information was provided. After identifying our focal genome set, we downloaded the relevant genomes for downstream analysis. Analysis scripts used in this study, including those that were used for data collection, are included in this study’s GitHub repository. A full list of the genome assemblies used in this study are provided in Table S1 (Supplemental Material).

Quantifying assembly completeness and phylogenetic inference
To assess gene completeness, we ran “Benchmarking Universal Single Copy Orthologs” (BUSCO) v.4.1.4 (58) on each assembly using the 1,367 reference genes in the OrthoDB v.10 Insecta gene set (59) and the “--long” analysis mode. We divided our data set into three subsets: (1) the full data set with no filtering, (2) only assemblies with BUSCO gene content ≥ 50%, or (3) only assemblies with BUSCO gene content ≥ 90%. To organize our results in a phylogenetic framework, we then estimated a species tree for our full data set using single-copy orthologs resulting from the BUSCO analyses.

Repeat element identification and annotation
We identified REs in genome assemblies using RepeatModeler2.0 (26) with search engine “ncbi”, which also generates a library of repeat consensus sequences. We annotated repeats in assemblies through two rounds of annotation with RepeatMasker4.1.0 (31), the first round used custom repeat libraries generated by RepeatModeler2 for each respective assembly and with search engine “ncbi” and option -xsmall. We then converted the softmasked assembly resulting from the first RepeatMasker round to a hardmasked assembly using the lc2n.py script (https://github.com/PdomGenomeProject/repeat-masking), and re-ran RepeatMasker on the hard-masked assembly with RepeatMasker’s internal arthropod repeat library and species “Arthropoda”. We then merged RepeatMasker output tables from both runs to summarize the abundance of RE categories. We studied patterns of repeat dynamics within and across taxonomic groups by parsing RepeatMasker output tables and visualizing the distribution and abundance of major RE categories using custom python and R scripts.

Correlation analyses
We tested for correlations between RE abundance and a range of aspects for each genome assembly, including sequencing technology, using R version 3.6.3 (60). These included a comparison of total REs identified as well as specific classes (e.g., LINEs) versus the primary sequencing technology used (short- or long-reads). For all correlation analyses, we tested for

normality in our data sets with a Shapiro-Wilk test and since the null hypothesis was rejected for all data sets ($P < 0.05$), we used Spearman's rank correlation tests.

Repetitive element and protein coding gene associations
For all assemblies with ≥ 90% BUSCO gene content, we measured RE-gene associations (i.e., RE sequences inserted within or adjacent to protein-coding genes) following Heckenhauer et al. (27). Their study validated a new approach to quantifying RE sequences associated with BUSCOs. In some cases, RE fragments are embedded within BUSCOs, and in others, REs with open reading frames that are immediately adjacent to BUSCOs are inadvertently classified by the BUSCO algorithm as being part of the BUSCO. They showed that quantifying such instances of RE sequences in BUSCOs can serve as a proxy for genome-wide RE-gene associations. Our approach adapted theirs to suit a higher throughput analysis, as described in more detail in Supplemental Materials.

Investigating the effects of taxonomic sampling bias
We investigated effects of taxonomic sampling bias on our understanding of REs in insects by analyzing the composition of the Repbase repository for RE sequences and the resulting impact on repeat annotation in our assemblies. We used custom scripts to parse the Repbase database and quantify the taxonomic representation of insect orders and families included in our data set, as well as the rate of insect repeat submissions over time.

Acknowledgements:
J.S.S. was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology (Division of Biological Infrastructure DBI-1811930) and a National Institutes of Health General Medical Sciences grant (R35GM119515) awarded to A.L. S.H. and J.L.K. were supported by NSF award #OPP-1906015. J.H and S.U.P. were supported by LOEWE-Centre for Translational Biodiversity Genomics funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK).

Data availability:
All genomes are publicly available on GenBank. Accession numbers are provided in Table S1. Species-specific repeat libraries generated by RepeatModeler2 and summary tables are available on FigShare (https://doi.org/10.6084/m9.figshare.c.6024905.v1). All scripts used in analyses are available on GitHub (https://github.com/johnssproul/Insect_REs).

References:

1. G. Bourque, *et al.*, Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).

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

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

4. A. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, D. D. Pollock, Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).

5. C. M. Bergman, H. Quesneville, Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **8**, 382–392 (2007).

6. C. G. Sotero-Caio, R. N. Platt, A. Suh, D. A. Ray, Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* **9**, 161–177 (2017).

7. S. Hotaling, *et al.*, Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. *Genome Biol. Evol.* **13**, evab138 (2021).

8. J. N. Wells, C. Feschotte, A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).

9. L. Sun, *et al.*, Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nat. Commun.* **11**, 1886 (2020).

10. D. J. Winter, *et al.*, Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae. *PLoS Genet.* **14**, e1007467 (2018).

11. J. L. Bennetzen, Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269 (2000).

12. M. Cáceres, J. M. Ranz, A. Barbadilla, M. Long, A. Ruiz, Generation of a Widespread Drosophila Inversion by a Transposable Element. *Science* **285**, 415–418 (1999).

13. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

14. L. Schrader, J. Schmitz, The impact of transposable elements in adaptive evolution. *Mol. Ecol.* **28**, 1537–1549 (2019).

15. T. Stuart, *et al.*, Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**, e20777 (2016).

16. C. L. Brand, M. T. Levine, Cross-species incompatibility between a DNA satellite and the Drosophila Spartan homolog poisons germline genome integrity. *Curr. Biol.* (2022) https:/doi.org/10.1016/j.cub.2022.05.009 (May 31, 2022).

17. P. M. Ferree, D. A. Barbash, Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLOS Biol.* **7**, e1000234 (2009).

18. M. Jagannathan, Y. M. Yamashita, Defective satellite DNA clustering into chromocenters underlies hybrid incompatibility in Drosophila. *Mol. Biol. Evol.* **38**, 4977–4986 (2021).

19. S. Hotaling, J. L. Kelley, P. B. Frandsen, Toward a genome sequence for every animal: Where are we now? *Proc. Natl. Acad. Sci.* **118**, e2109019118 (2021).

20. R. A. Marks, S. Hotaling, P. B. Frandsen, R. VanBuren, Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **7**, 1571–1578 (2021).

21. N. E. Stork, How many species of insects and other terrestrial arthropods are there on Earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).

22. Y. B. Kim, *et al.*, Divergence of Drosophila melanogaster repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. *Proc. Natl. Acad. Sci.* **111**, 10630–10635 (2014).

23. C. Vargas-Chavez, *et al.*, Transposable element variants and their potential adaptive impact in urban populations of the malaria vector Anopheles coluzzii. *Genome Res.* **32**, 189–202 (2022).

24. B. Misof, *et al.*, Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).

25. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

26. J. M. Flynn, *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

27. J. Heckenhauer, *et al.*, Genome size evolution in the diverse insect order Trichoptera. *GigaScience* **11**, giac011 (2022).

28. M. Petersen, *et al.*, Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol. Evol.* **19**, 11 (2019).

29. R. Hubley, *et al.*, The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).

30. J. Jurka, *et al.*, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).

31. A. Smit, R. Hubley, *RepeatMasker Open-4.1. 2019*.

32. R. K. Slotkin, R. Martienssen, Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).

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

34. K. H.-C. Wei, D. Mai, K. Chatla, D. Bachtrog, Dynamics and Impacts of Transposable Element Proliferation in the Drosophila nasuta Species Group Radiation. *Mol. Biol. Evol.* **39**, msac080 (2022).

35. 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).

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

37. M. C. Stitzer, S. N. Anderson, N. M. Springer, J. Ross-Ibarra, The genomic ecosystem of transposable elements in maize. *PLOS Genet.* **17**, e1009768 (2021).

38. L. K. Olsen, *et al.*, Draft Genome Assemblies and Annotations of Agrypnia vestita Walker, and Hesperophylax magnus Banks Reveal Substantial Repetitive Element Expansion in Tube Case-Making Caddisflies (Insecta: Trichoptera). *Genome Biol. Evol.* **13**, evab013 (2021).

39. L. Schrader, *et al.*, Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.* **5**, 5495 (2014).

40. V. Horváth, M. Merenciano, J. González, Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends Genet.* **33**, 832–841 (2017).

41. S. Signor, G. Yocum, J. Bowsher, Life stage and the environment as effectors of transposable element activity in two bee species. *J. Insect Physiol.* **137**, 104361 (2022).

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

43. K. H. Miga, *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).

44. H. A. Lewin, *et al.*, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* **115**, 4325–4333 (2018).

45. T. A. Elliott, *et al.*, TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mob. DNA* **12**, 16 (2021).

46. D. R. Hoen, *et al.*, A call for benchmarking transposable element annotation methods. *Mob. DNA* **6**, 13 (2015).

47. B. Piégu, S. Bire, P. Arensburger, Y. Bigot, A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.* **86**, 90–109 (2015).

48. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).

49. T. J. Wheeler, *et al.*, Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2012).

50. S. Dodsworth, *et al.*, Genomic Repeat Abundances Contain Phylogenetic Signal. *Syst. Biol.* **64**, 112–126 (2015).

51. S. Negm, A. Greenberg, A. M. Larracuente, J. S. Sproul, RepeatProfiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles. *Mol. Ecol. Resour.* **21**, 969–981 (2021).

52. J. S. Sproul, L. M. Barton, D. R. Maddison, Repetitive DNA profiles Reveal Evidence of Rapid Genome Evolution and Reflect Species Boundaries in Ground Beetles. *Syst. Biol.* (2020) https:/doi.org/10.1093/sysbio/syaa030 (May 15, 2020).

53. R. K. Slotkin, The case for not masking away repetitive DNA. *Mob. DNA* **9**, 15 (2018).

54. O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474 (2006).

55. S. Ou, *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

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

57. E. W. Sayers, *et al.*, GenBank. *Nucleic Acids Res.* **49**, D92–D96 (2021).

58. M. Seppey, M. Manni, E. M. Zdobnov, "BUSCO: Assessing Genome Assembly and Annotation Completeness" in *Gene Prediction: Methods and Protocols*, Methods in Molecular Biology., M. Kollmar, Ed. (Springer, 2019), pp. 227–245.

59. E. V. Kriventseva, *et al.*, OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).

60. R Core Team, R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna Austria URL HttpwwwR-Proj.* (2013).