Long-reads are revolutionizing 20 years of insect genome sequencing

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

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

Correspondence:
Scott Hotaling, School of Biological Sciences, Washington State University, Pullman, WA, 99164, USA; Email: scott.hotaling@wsu.edu: Phone: (828) 507-9950

Paul B. Frandsen, Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, 84602, USA; Email: paul_frandsen@byu.edu; Phone: (804) 422-2283

Keywords: genome biology, Insecta, Arthropoda, arthropod genomics, long-read sequencing, Pacific Biosciences, Oxford Nanopore

Running head: The state of insect genome biology

The first insect genome (*Drosophila melanogaster*) was published two decades ago. Today, nuclear genome assemblies are available for a staggering 601 different insects representing 20 orders. Here, we analyzed the best assembly for each insect and provide a "state of the field" perspective, emphasizing taxonomic representation, assembly quality, gene completeness, and sequencing technology. We show that while genomic efforts have been biased towards specific groups (e.g., Diptera), assemblies are generally contiguous with gene regions intact. Most notable, however, has been the impact of long-read sequencing; assemblies that incorporate long-reads are ~48x more contiguous than those that do not.

Since the publication of the *Drosophila melanogaster* genome[1], sequencing and analytical technologies have dramatically expanded, bringing the power of genome sequencing to an ever-expanding pool of researchers. More than 600 insects have now had their nuclear genome sequenced and made publicly available on the most popular repository, GenBank[2]. While representing just 0.06% of the ~1 million described insects[3], this breadth of insect genome sequencing still spans ~480 million years of evolution[4] and roughly two orders of genome size from the tiny 99 megabase (Mb) genome of *Belgica antarctica*[5] to the massive genome of *Locusta migratoria* at 6.5 gigabases (Gb)[6].

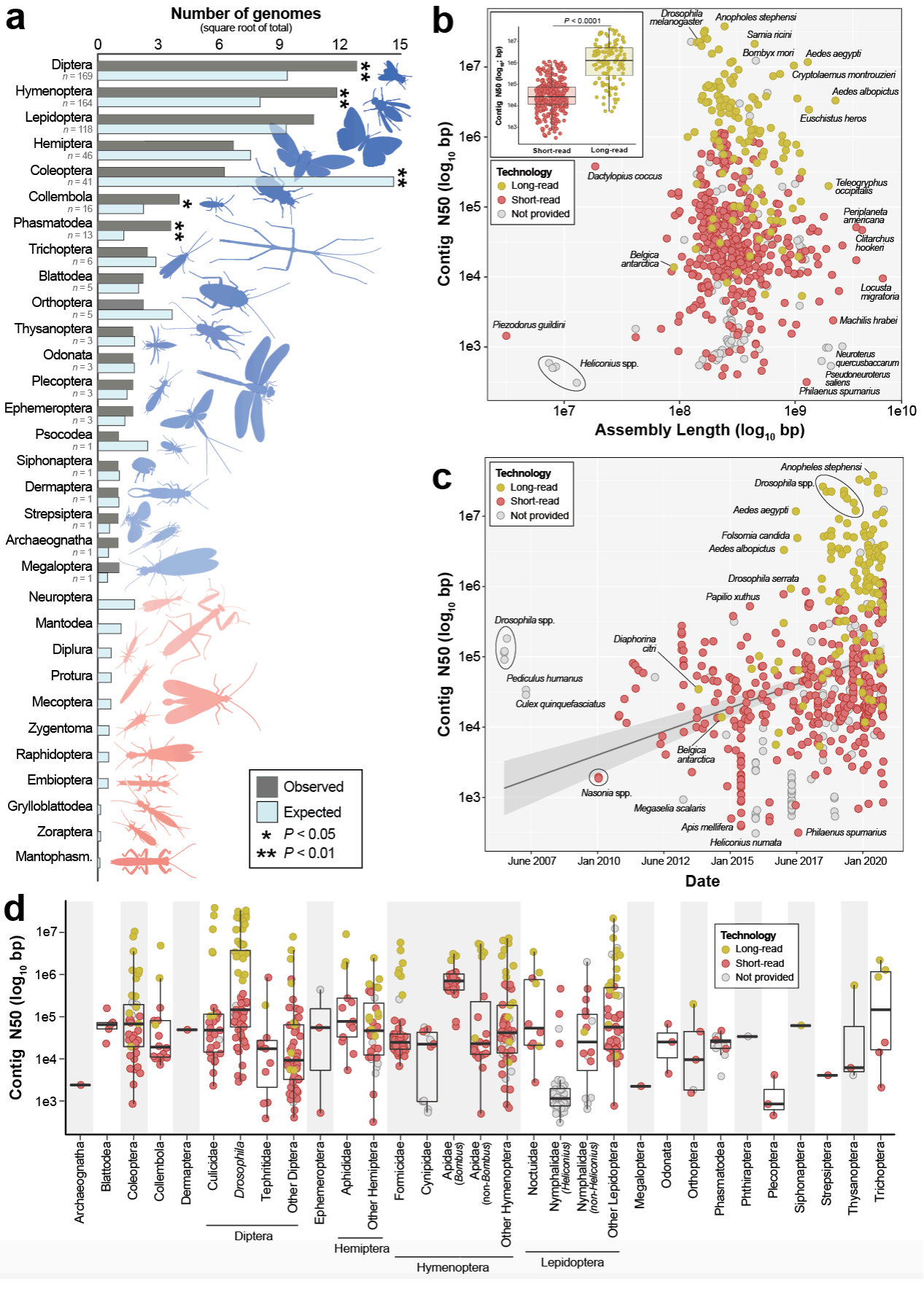

Figure 1. Taxonomic representation, contiguity, and the timeline of availability for all nuclear insect genome assemblies (*n* = 601) as of November 2020. Only one assembly per named species or subspecies is included. (a) The taxonomic diversity of insect genomes on GenBank. Observed versus expected numbers of genomes represent the total number of available assemblies versus those that would be expected given the proportion that each order comprises of all described insect diversity. Significance was assessed with Fisher's exact tests. One order is underrepresented (Coleoptera) while four orders are overrepresented (Diptera, Hymenoptera, Collembola, Phasmatodea). Eleven insect orders (light red silhouettes) have no publicly available genome sequence. A breakdown of sequencing technology by order is shown in Fig. S1. (b) Genome contiguity versus total assembly length. Contiguity was assessed according to contig N50, the mid-point of the contig distribution where 50% of the genome is assembled into contigs of a given length or longer. The inset plot shows a comparison of contig N50 distributions for short-read (*n* = 365) versus long-read (*n* = 126) assemblies. Significance was assessed with a Welch's T-test. A more expansive version of the inset comparison that includes a fine-scale breakdown of the sequence technology used for each assembly is shown in Fig. S2. (c) The timeline of genome availability for insects according to the GenBank publication date. A steady increase in contiguity is largely precipitated by the rise of long-read sequencing. Labeled in (b) and (c): The smallest (*Belgica antarctica*, 99 Mb[5]), largest (*Locusta migratoria*, 6.5 Gb[6]), and many other outlier genome assemblies in terms of either assembly size or contiguity. Groups of species in the same genus are labeled with black circles. (d) Contig N50 by taxonomic group. Generally, taxa were grouped into orders except when 10 or more genomes were available for a lower taxonomic group (family or genus). As in (b) and (c), each point represents a single insect genome.

Expanding genomic resources have transformed biological research and precipitated major advances in our understanding of the origins of biodiversity[7-10]. Considerable progress has been driven by large-scale consortia (e.g., Human Genome Project[11]; Vertebrate Genome Project[12]) and for insects, the most prominent consortium has been the i5K initiative to sequence genomes for 5,000 different insects[13]. The rise of long-read sequencing technologies—primarily Oxford Nanopore and Pacific Biosciences (PacBio)—have also changed the landscape of genome sequencing by providing an economical means for high-throughput generation of reads that are commonly 25 kilobases (Kb) or longer[14], thereby greatly increasing the mean size of sequence fragments used in assemblies. Genomic effort in insects, however, has not been spread evenly. Aquatic insects, as a group, are woefully underrepresented relative to their terrestrial counterparts[15]. And, some orders (e.g., Diptera) are represented by far more genome assemblies than their species diversity alone would warrant, while others have no genomic representation (Fig. 1a).

Here, we curated and analyzed the best available genome assembly for 601 insects (species or subspecies). We provide a "state of the field" perspective emphasizing taxonomic representation, assembly quality, gene completeness, and sequencing technology. We focused on taxonomic breadth rather than within-group efforts (e.g., The *Anopheles gambiae* 1000 Genomes Consortium[16]) to gain more holistic insight on the field. Following similar studies[4,15,17], we defined insects to include all groups within the subphylum Hexapoda. We then downloaded metadata for all nuclear Insecta genomes on GenBank[2] (accessed 2 November 2020). We culled this data set to only include the genome with the highest contig N50 for each taxon and downloaded assemblies for analysis. 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.

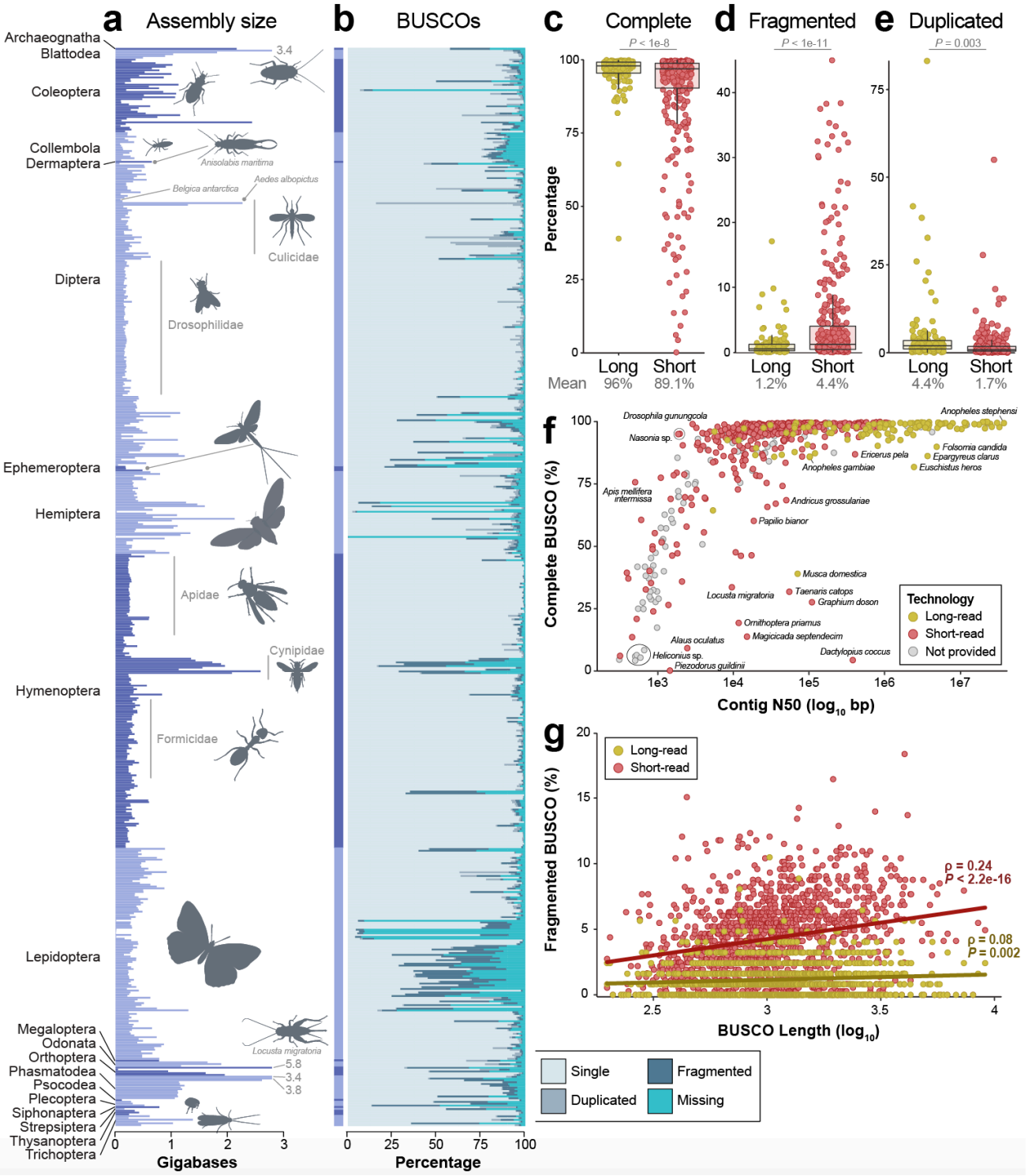

Figure 2. Variation in assembly size and Benchmarking Universal Single-Copy Ortholog (BUSCO) gene completeness across Insecta. (a) Assembly size for all insects, ordered by family. To improve visualization, the upper display limit was set to 2.8 gigabases (Gb). Four genome assemblies exceeded this value and their values are labeled with gray text (in Gb). (b) BUSCO results for each insect genome. Each horizontal bar represents one genome (n = 601 genomes) and corresponds to the same taxon in the assembly size plot to the left in (a). (c-e) Long-read versus short-read genome assembly comparisons of (c) complete BUSCOs (single and duplicated combined), (d) fragmented BUSCOs, and (e) duplicated BUSCOs only. Significance was

assessed with Welch's T-tests. (f) A comparison of BUSCO completeness versus contig N50. Each point represents the best available assembly for one taxon and groups of taxa in the same genus are labeled with black circles. Unsurprisingly, more contiguous genomes also exhibit greater gene completeness. (g) Longer genes are more likely to be fragmented in insect genome assemblies, regardless of the technology used. However, a much stronger correlation exists between short-read assemblies and fragmentation of longer genes (Spearman's ρ: 0.24, $P < 2.2e-16$) than for long-read assemblies (Spearman's ρ: 0.08, $P = 0.002$). Unlike in (c-e), each circle in (g) represents the percent of fragmentation for that BUSCO gene across all genomes in a given group (long- or short-read assemblies). Thus, each gene is included twice (once for each technology). All BUSCO genes in the OrthoDB v.10 Insecta gene set[18] (n = 1,367) were used except one 2.02 Kb gene that was missing in >70% of assemblies and subsequently removed from analysis and visualization. BUSCO gene lengths varied from 198 base pairs to 9.01 Kb.

To test if insect orders were under- or overrepresented in terms of genome availability, we compared the observed number of taxa with assemblies to the expected number given the described diversity for a given order. We obtained totals for the number of insects described overall and for each order from previous studies[19,20]. We assessed significance between observed and expected representation with Fisher's exact tests. To assess gene completeness, we ran BUSCO v.4.1.4[21] on each assembly using the OrthoDB v.10 Insecta gene set[18] (n = 1,367 reference genes). We tested for differences in distributions of contig N50 or assembly size between short- and long-read genomes with Welch's T-tests. Next, using the BUSCO gene set, we tested whether longer genes were more likely to be missing or fragmented depending on sequencing technology (short- or long-read) with Spearman's correlations. An extended version of the methods with the scripts used for analysis are provided in the Supplementary Materials and GitHub repository (https://github.com/pbfrandsen/insect_genome_assemblies).

As of November 2020, 601 nuclear genome assemblies representing taxa from 20 insect orders were available on GenBank. These data were dominated by Diptera (n = 169 genomes), Hymenoptera (n = 164), and Lepidoptera (n = 118; Fig. 1a). Four orders were overrepresented relative to their species diversity: Collembola, Diptera, Hymenoptera, and Phasmatodea (P, Fisher's < 0.03; Fig. 1). Coleoptera, with 387,100 described species[19], was significantly underrepresented (41 genomes versus ~228 expected; P, Fisher's < 0.01). Six orders were represented by only one genome and 11 orders had no publicly available genome. This lack of representation was particularly striking for Neuroptera (5,868 described species[19]).

On average, insect genome assemblies were 439.2 Mb in length (SD = 448.4 Mb; Fig. 2a) with a mean contig N50 of 1.09 Mb (SD = 4.01 Mb) and 87.5% (SD = 21%) BUSCO completeness (single and duplicated genes, combined). Substantial variation existed in all three metrics, however, with assemblies ranging from the highly incomplete genome of *Piezodorus guildini* at just 3.2 Mb (contig N50 = 1.5 Kb, BUSCO completeness = 0.2%) to the exceptionally high-quality 140.7 Mb assembly of *D. melanogaster* (contig N50 = 22.4 Mb, BUSCO completeness = 99.9%; Fig. 2, Table S1). For orders represented by >10 taxa, Hymenoptera assemblies were the most complete (94%, SD = 14.3%) and Lepidoptera the least (74.6%, SD = 28.2%; Fig. 2b). At 15.3%, Lepidoptera had the lowest percentage of long-read genomes (Fig. S1) and *Heliconius* genomes were particularly fragmented (Fig. 1d). For families represented by >10 taxa, Drosophilidae assemblies were the most complete (98.4%, SD = 2%) followed closely by Apidae (97.9%, SD = 3.7%; Fig. 1d, 2b). As expected, assemblies with higher contig N50 scores were also more complete (Fig. 2f) but assembly size had little to no effect on gene completeness (Fig. S3).

The type(s) of sequence data used for genome assembly were obtained for ~82% of genomes (long-read = 126, short-read = 365; Table S1). Long-read assemblies were more contiguous than short-read assemblies (Fig. 1b; P, Welch's T-test < 0.0001), averaging contig N50 values that were ~4.4 Mb higher despite no difference in assembly size (P, Welch's T-test = 0.12; Fig. S4).

Gene regions were also far more complete in long-read genomes (mean BUSCO = 96%, SD = 7%) versus short-read genomes (89.1%, SD = 19%; *P,* Welch's T-test < 1e-8; Fig. 2c) with 70% fewer fragmented genes (*P,* Welch's T-test < 1e-11; Fig. 2d). Long-read genomes, however, had ~2.6x more duplicated genes (4.4% vs. 1.7%; *P,* Welch's T-test = 0.003; Fig. 2e). Longer BUSCO genes were also more likely to be fragmented in both short-read (Spearman's ρ: 0.24, *P* < 2.2e-16) and long-read assemblies (Spearman's ρ: 0.08, *P* = 0.002; Fig. 2g) but they were less likely to be missing in both when compared to shorter genes (short-read: Spearman's ρ: -0.08, *P* = 0.002; long-read: Spearman's ρ: -0.18, *P* = 9.7e-12; Fig. S5).

The rate at which new insect genomes are becoming available is accelerating (Fig. 1c). Nearly 50% (*n* = 292) of all insect assemblies were accessioned in 2019–2020 (Tables S1-S2). Despite such an influx of new genomes, the same period also represented a high-water mark of contiguity (mean contig N50, 2019–2020 = 1.77 Mb; Table S2). Much of this increase was driven by long-read assemblies which rose in frequency from 0% in 2011–2012 to 36.1% of all assemblies in 2019–2020. Long-read assemblies also underwent a major shift in contiguity in 2017 (Fig. S6, Table S2).

We have entered a new era of insect genome biology. Since 2019, a new species has been sequenced every 2.3 days and these new assemblies are, on average, vastly superior to those of just a few years ago. As we continue building this data set, we offer four recommendations: first, we should recognize the community-driven nature of these data and seek better integration between research groups and consortia in terms of data sharing, best practices, and taxonomic focus. Second, new sequencing efforts should strive to balance sampling that fills taxonomic gaps and improves existing resources with targeted sampling motivated by specific questions. Both approaches are valuable and not mutually exclusive. The former—filling taxonomic gaps—is critical to broadly understanding the evolution of insects, the most diverse animal group on Earth. While the latter—targeted, question-driven sequencing—is critical to understanding specific aspects of genome biology which are often best answered using dense sampling of model groups. Third, we echo the initial findings of the Vertebrate Genome Project[12]—long-read assemblies are vastly more contiguous than short-read approaches—and recommend that these technologies be embraced by insect genome scientists. And, fourth, as of 2019, only 40% of insect genome assemblies had corresponding gene annotations on GenBank[22]. Expanding and refining the availability of gene annotations for insects will drive corresponding increases in the scale of taxonomic comparisons that are possible for many analyses.

Beyond resource development, we must continue to leverage this data set to conduct new studies of insect genome biology and evolution. These efforts are beginning to emerge and are paying dividends. For instance, 76 arthropod genomes were used to identify novel gene families that arose during insect diversification with links to key adaptations including flight[23]. Similarly, a study of 195 insect genomes revealed the high diversity of transposable elements across insects with varying levels of conservation depending upon taxonomy[24].

With genomes representing 600+ taxa and ~480 million years of evolution available in a public repository, the power and promise of insect genome research has never been greater. The rise of long-read assemblies, in particular, will spur new understanding of previously difficult to characterize aspects of the genome (e.g., repeat dynamics). While our focus was on insects, long-reads are likely revolutionizing genome science in virtually all taxonomic groups with untapped genomic potential existing in public repositories across the Tree of Life. By continuing to build, curate, and expand access to these resources, we will gain tremendous insight into genome biology and evolution at broad phylogenetic scales while also creating a more inclusive, equitable discipline.

Acknowledgements:

S.H. and J.L.K. were supported by NSF award OPP-1906015. J.H and S.U.P. were supported by the LOEWE-Centre for Translational Biodiversity Genomics which is funded by the Hessen State Ministry of Higher Education, Research and the Arts. J.S.S. was supported by an NSF Postdoctoral Research Fellowship in Biology (DBI-1811930) and an NIH General Medical Sciences award (R35GM119515) to A.M.L.

References:

1 Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).

2 Sayers, E. W. *et al.* GenBank. *Nucleic Acids Research* **48**, D84-D86 (2020).

3 Stork, N. E. How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology* **63**, 31-45 (2018).

4 Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767, doi:10.1126/science.1257570 (2014).

5 Kelley, J. L. *et al.* Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nature Communications* **5**, 4611, doi:10.1038/ncomms5611 (2014).

6 Wang, X. *et al.* The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications* **5**, 1-9 (2014).

7 Seehausen, O. *et al.* Genomics and the origin of species. *Nature Reviews Genetics* **15**, 176-192 (2014).

8 Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 1-6 (2016).

9 McKenna, D. D. *et al.* The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences* **116**, 24729-24737 (2019).

10 McGee, M. D. *et al.* The ecological and genomic basis of explosive adaptive radiation. *Nature* **586**, 75-79 (2020).

11 Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286-290 (2003).

12 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* (2020).

13 Robinson, G. E. *et al.* Creating a buzz about insect genomes. *Science* **331**, 1386 (2011).

14 Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 1-16 (2020).

15 Hotaling, S., Kelley, J. L. & Frandsen, P. B. Aquatic insects are dramatically underrepresented in genomic research. *Insects* **11**, 601 (2020).

16 Consortium, A. g. G. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature* **552**, 96 (2017).

17 Petersen, M. *et al.* Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology* **19**, 1-15 (2019).

18 Kriventseva, E. V. *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 Research* **47**, D807-D811 (2019).

19 Zhang, Z.-Q. *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness*. (Magnolia press, 2011).

20 Bellinger, P. F., Christiansen, K. A. & Janssens, F. Checklist of the Collembola of the World. http://www.collembola.org. (2020).

21 Seppey, M., Manni, M. & Zdobnov, E. M. in *Gene prediction*. p. 227-245 (Springer, 2019).

22 Li, F. *et al.* Insect genomes: progress and challenges. *Insect molecular biology* **28**, 739-758 (2019).

23 Thomas, G. W. *et al.* Gene content evolution in the arthropods. *Genome Biology* **21**, 1-14 (2020).

24 Gilbert, C., Peccoud, J. & Cordaux, R. Transposable Elements and the Evolution of Insects. *Annual Review of Entomology* **66** (2020).