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

***Authors for 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

Abstract:

The first insect genome (*Drosophila melanogaster*) was published two decades ago. Today, nuclear genome assemblies are available for a staggering 601 insect species representing 20 orders. In this study, we analyzed the most-contiguous assembly for each species and provide a

"state of the field" perspective, emphasizing taxonomic representation, assembly quality, gene completeness, and sequencing technologies. Relative to species richness, genomic efforts have been biased towards four orders (Diptera, Hymenoptera, Collembola, and Phasmatodea), Coleoptera are underrepresented, and 11 orders still lack a publicly available genome assembly. The average insect genome assembly is 439.2 megabases in length with 87.5% of single-copy benchmarking genes intact. Most notable has been the impact of long-read sequencing; assemblies that incorporate long-reads are ~48x more contiguous than those that do not. We offer four recommendations as we collectively continue building insect genome resources: (1) seek better integration between independent research groups and consortia, (2) balance future sampling between filling taxonomic gaps and generating data for targeted questions, (3) take advantage of long read sequencing technologies, and (4) expand and improve gene annotations.

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

Significance statement:

Since the first insect genome was sequenced ~20 years ago, sequencing technologies and the availability of insect genome assemblies have both advanced dramatically. In this study, we curated, analyzed, and summarized the field of insect genomics in terms of taxonomic representation, assembly quality, gene completeness, and sequencing technology. We show that 601 insect species have genome assemblies available, with some groups heavily overrepresented (e.g., Diptera) relative to others (e.g., Coleoptera). The major takeaway of our study is that genome assemblies produced with long reads are ~48x more contiguous than short read assemblies.

52 **Body:**

53 Since the publication of the *Drosophila melanogaster* genome (Adams, et al. 2000) sequencing
54 and analytical technologies have developed rapidly, bringing the power of genome sequencing to
55 an ever-expanding pool of researchers. More than 600 insects have now had their nuclear
56 genome sequenced and made publicly available in the GenBank repository (Sayers, et al. 2021).
57 While representing just 0.06% of the ~1 million described insects (Stork 2018), this breadth of
58 insect genome sequencing still spans ~480 million years of evolution (Misof, et al. 2014) and
59 roughly two orders of genome size from the tiny 99 megabase (Mb) genome of *Belgica antarctica*
60 (Kelley, et al. 2014) to the massive genome of *Locusta migratoria* at 6.5 gigabases (Gb; Wang,
61 et al. 2014).

62

63 Accumulating genomic resources have transformed biological research and precipitated major
64 advances in our understanding of the origins of biodiversity (Hug, et al. 2016; McGee, et al.
65 2020; McKenna, et al. 2019; Seehausen, et al. 2014). Considerable progress has been driven by
66 large-scale consortia [e.g., Human Genome Project (Collins, et al. 2003); Vertebrate Genome
67 Project (Rhie, et al. 2021)] and for insects, the most prominent consortium has been the i5K
68 initiative to sequence genomes for 5,000 different arthropods (i5K Consortium 2013; Robinson,
69 et al. 2011). The rise of long-read sequencing technologies—primarily Oxford Nanopore and
70 Pacific Biosciences (PacBio)—have also changed the landscape of genome sequencing by
71 providing an economical means for high-throughput generation of reads that are commonly 25
72 kilobases (Kb) or longer (Amarasinghe, et al. 2020), thereby greatly increasing the average size
73 of sequences used in assemblies. Genome sequencing efforts in insects, however, have not
74 been spread evenly. Aquatic insects, as a group, are underrepresented relative to their terrestrial
75 counterparts (Hotaling, et al. 2020). And, some orders (e.g., Diptera) are represented by far more
76 genome assemblies than their species diversity alone would warrant—likely reflecting the model
77 organisms within them—while many orders still have no genomic representation.

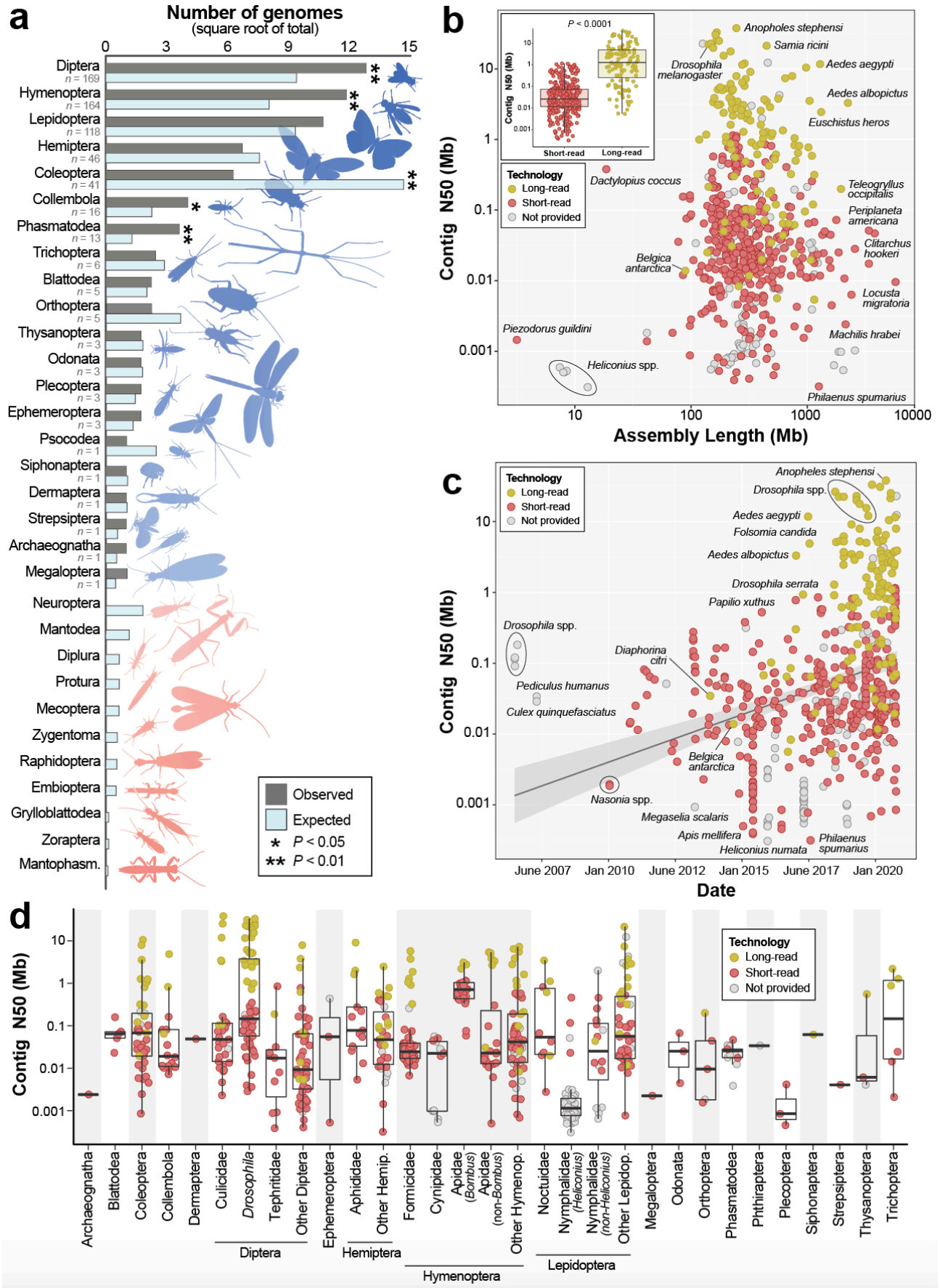

a
Number of genomes
(square root of total)
0
3
6
9
12
15
Diptera
n = 169
Hymenoptera
n = 164
Lepidoptera
n = 118
Hemiptera
n = 46
Coleoptera
n = 41
Collembola
n = 16
Phasmatodea
n = 13
Trichoptera
n = 6
Blattodea
n = 5
Orthoptera
n = 5
Thysanoptera
n = 3
Odonata
n = 3
Plecoptera
n = 3
Ephemeroptera
n = 3
Psocodea
n = 1
Siphonaptera
n = 1
Dermaptera
n = 1
Strepsiptera
n = 1
Archaeognatha
n = 1
Megaloptera
n = 1
Neuroptera
Mantodea
Diplura
Protura
Mecoptera
Zygentoma
Raphidoptera
Embioptera
Grylloblattodea
Zoraptera
Mantophasm.
Observed
Expected
* P < 0.05
** P < 0.01
b
P < 0.0001
Contig N50 (Mb)
Short-read
Long-read
Technology
Long-read
Short-read
Not provided
Anopholes stephensi
Samia ricini
Drosophila melanogaster
Aedes aegypti
Aedes albopictus
Euschistus heros
Dactylopius coccus
Teleogryllus occipitalis
Periplaneta americana
Clitarchus hookeri
Belgica antarctica
Locusta migratoria
Piezodorus guildini
Machilis hrabei
Heliconius spp.
Philaenus spumarius
Contig N50 (Mb)
Assembly Length (Mb)
c
Technology
Long-read
Short-read
Not provided
Anopheles stephensi
Drosophila spp.
Aedes aegypti
Folsomia candida
Aedes albopictus
Drosophila serrata
Papilio xuthus
Drosophila spp.
Diaphorina citri
Pediculus humanus
Culex quinquefasciatus
Belgica antarctica
Nasonia spp.
Megaselia scalaris
Apis mellifera
Heliconius numata
Philaenus spumarius
Contig N50 (Mb)
June 2007
Jan 2010
June 2012
Jan 2015
June 2017
Jan 2020
Date
d
Contig N50 (Mb)
Technology
Long-read
Short-read
Not provided
Archaeognatha
Blattodea
Coleoptera
Collembola
Dermaptera
Culicidae
Drosophila
Tephritidae
Other Diptera
Ephemeroptera
Aphididae
Other Hemip.
Formicidae
Cynipidae
Apidae (Bombus)
Apidae (non-Bombus)
Other Hymenop.
Noctuidae
Nymphalidae (Heliconius)
Nymphalidae (non-Heliconius)
Other Lepidop.
Megaloptera
Odonata
Orthoptera
Phasmatodea
Phthiraptera
Plecoptera
Siphonaptera
Strepsiptera
Thysanoptera
Trichoptera
Diptera
Hemiptera
Hymenoptera
Lepidoptera

Figure 1. Taxonomic representation, contiguity, and the timeline of availability for the most-contiguous nuclear genome assembly for 601 insect species in GenBank as of November 2020. Only one assembly per named species or subspecies is included. (a) *The taxonomic diversity of available insect genome assemblies.* Observed versus expected numbers of genome assemblies 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 orders (light red silhouettes) have no publicly available genome assembly. A breakdown of sequencing technology by order is shown in Fig. S1. (b) *Genome contiguity versus total assembly length.* Contiguity was assessed with 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 finer-scale breakdown by sequencing technology is shown in Fig. S2. (c) *The timeline of genome assembly 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): well-known or outlier genome assemblies in terms of either model status, 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 assemblies were available for a lower taxonomic level (family or genus). As in (b) and (c), each point represents a single insect genome assembly.

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 (2017)] to gain a more holistic overview of the field. Following similar studies (e.g., Hotaling, et al. 2020; Misof, et al. 2014; Petersen, et al. 2019), we defined insects to include all groups within the subphylum Hexapoda. We downloaded metadata from GenBank for all nuclear hexapod genome assemblies on an order-by-order basis (Sayers, et al. 2021; accessed 2

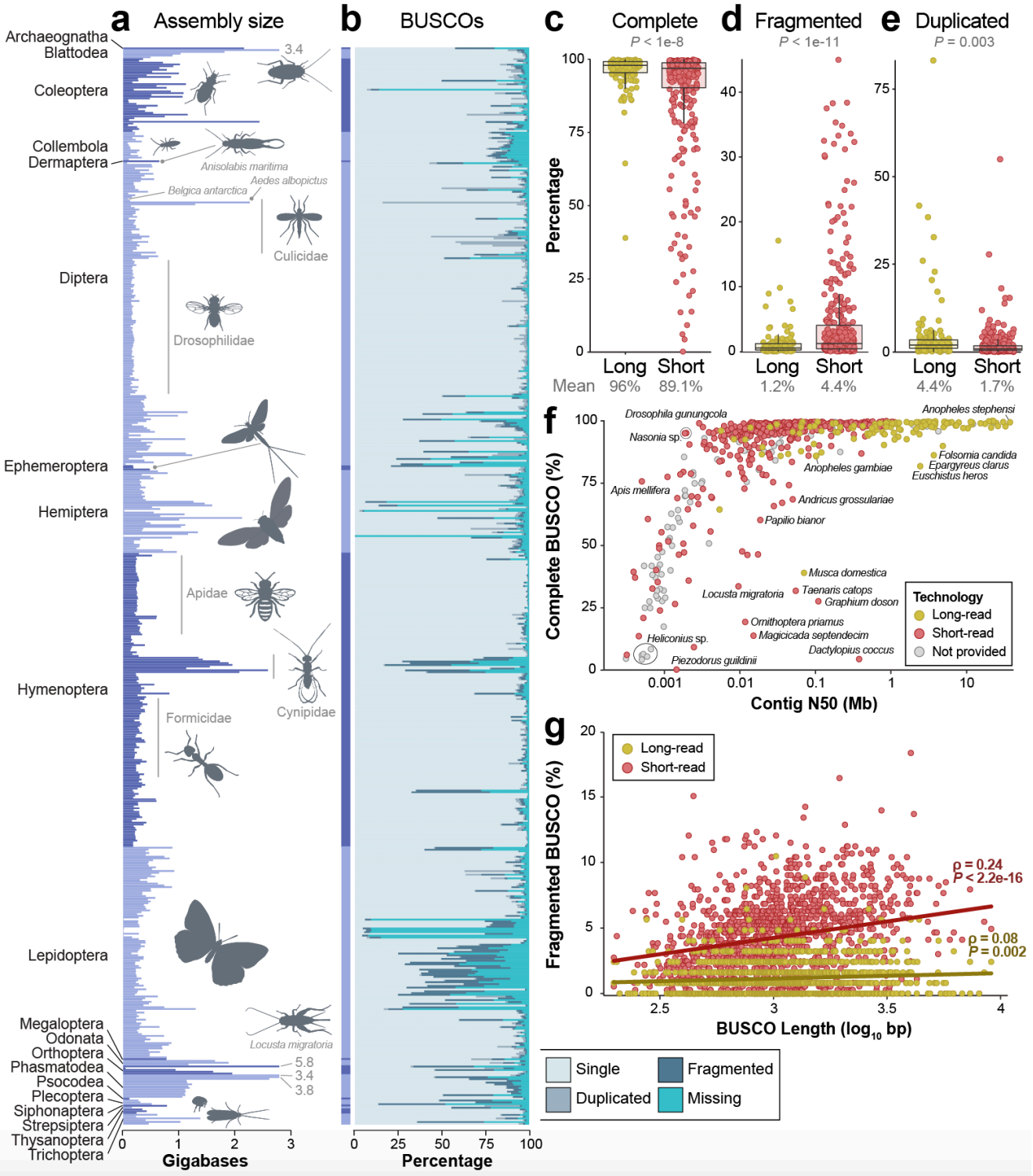
107 November 2020). We culled this data set to only include the assembly with the highest contig
108 N50 for each taxon and downloaded these assemblies for analysis. We acknowledge that this
109 filtering approach may introduce biases towards the present-day for assemblies that have been
110 improved over the years. Assemblies were classified as “short-read”, “long-read”, or “not
111 provided” based on whether only short-reads (e.g., Illumina) were used, any amount of long-read
112 sequences (e.g., PacBio) were used, or no information was provided. If an assembly used both
113 short- and long-reads (a “hybrid” assembly), it was classified as a long-read assembly in our
114 analysis.
115
116 To test if insect orders were under- or overrepresented in terms of genome assembly availability,
117 we compared the observed number of taxa with assemblies to the expected number given the
118 described diversity for a given order. We obtained totals for the number of insects described
119 overall and for each order from previous studies (Bellinger, et al. 2020; Zhang 2011). We
120 assessed significance between observed and expected representation with Fisher’s exact tests.
121 To assess gene completeness, we ran “Benchmarking Universal Single Copy Orthologs”
122 (BUSCO) v.4.1.4 (Seppey, et al. 2019) on each assembly using the 1,367 reference genes in the
123 OrthoDB v.10 Insecta gene set (Kriventseva, et al. 2019). It should be noted that Collembola
124 genome assemblies may have received slightly lower BUSCO scores in this analysis because
125 non-insect hexapod genomes were not used to generate the Insecta gene set. We tested for
126 differences in distributions of contig N50 or assembly size between short- and long-read
127 assemblies with Welch’s T-tests. Next, using the BUSCO gene set, we tested whether longer
128 genes were more likely to be missing or fragmented depending on sequencing technology (short-
129 or long-read) with Spearman’s correlations. We defined BUSCO gene length as the full
130 nucleotide sequence for the protein-coding portions of the consensus “ancestral” genes included
131 in the OrthoDB v.10 Insecta gene set. An extended version of the methods and the scripts used

132 for analysis are provided in the Supplementary Materials and GitHub repository
133 (https://github.com/pbfrandsen/insect_genome_assemblies).
134
135 As of November 2020, 601 different insect species representing 20 orders had nuclear genome
136 assemblies available in GenBank. These data were dominated by Diptera (*n* = 169 assemblies),
137 Hymenoptera (*n* = 164), and Lepidoptera (*n* = 118; Fig. 1a). Four orders were overrepresented
138 relative to their species diversity: Collembola, Diptera, Hymenoptera, and Phasmatodea (*P*,
139 Fisher's < 0.03; Fig. 1a). Coleoptera, with 387,100 described species (Zhang 2011), was
140 significantly underrepresented (41 assemblies versus ~228 expected; *P*, Fisher's < 0.01). Six
141 orders were represented by only one genome assembly and 11 orders had no publicly available
142 assembly. This lack of representation was particularly striking for Neuroptera (5,868 described
143 species, Zhang 2011).
144
145 On average, insect genome assemblies were 439.2 Mb in length (SD = 448.4 Mb; Fig. 2a) with a
146 mean contig N50 of 1.09 Mb (SD = 4.01 Mb) and 87.5% (SD = 21%) BUSCO completeness
147 (single and duplicated genes, combined). Substantial variation existed in all three metrics,
148 however, with assemblies ranging from the highly incomplete assembly of *Piezodorus guildini* at
149 just 3.2 Mb (contig N50 = 1.5 Kb, BUSCO completeness = 0.2%) to the exceptionally high-quality
150 140.7 Mb assembly of *D. melanogaster* (contig N50 = 22.4 Mb, BUSCO completeness = 99.9%;
151 Fig. 2, Table S1). For orders represented by >10 taxa, Hymenoptera assemblies were the most
152 complete (BUSCO completeness = 94%, SD = 14.3%) and Lepidoptera the least (74.6%, SD =
153 28.2%; Fig. 2b). At 15.3%, Lepidoptera had the lowest percentage of long-read assemblies (Fig.
154 S1) and *Heliconius* assemblies were particularly fragmented (Fig. 1d). For families represented
155 by >10 taxa, Drosophilidae assemblies were the most complete (BUSCO completeness = 98.4%,
156 SD = 2%) followed closely by Apidae assemblies (97.9%, SD = 3.7%; Fig. 1d, 2b). As expected,

157 assemblies with higher contig N50 lengths were also more complete (Fig. 2f) but assembly size

158 had little to no effect on gene completeness (Fig. S3).

159



160

161 **Figure 2. Variation in assembly size and Benchmarking Universal Single-Copy Ortholog (BUSCO) gene**
162 **completeness across Insecta.** (a) Assembly size for all insects, grouped by order then family. To improve
163 visualization, the upper display limit was set to 2.8 gigabases (Gb). Four genome assemblies exceeded this
164 value and are labeled with gray text (in Gb). Taxa silhouettes were either handmade or taken from PhyloPic
165 (http://phylopic.org). (b) BUSCO results for each insect genome assembly. Each horizontal bar represents one
166 assembly (n = 601 species) and corresponds to the same taxon in the assembly size plot to the left in (a). (c-e)
167 Long-read versus short-read genome assembly comparisons of (c) complete BUSCOs (single and duplicated
168 combined), (d) fragmented BUSCOs, and (e) duplicated BUSCOs only. Significance was assessed with Welch's
169 T-tests. (f) A comparison of BUSCO completeness versus contig N50. Each point represents the best available
170 assembly for one taxon and groups of taxa in the same genus are labeled with black circles. Unsurprisingly,
171 more contiguous genome assemblies also exhibit greater gene completeness. (g) Longer genes are more likely
172 to be fragmented in insect genome assemblies, regardless of the technology used. However, a much stronger
173 correlation exists between short-read assemblies and fragmentation of longer genes (Spearman's ρ: 0.24, P <
174 2.2e-16) than for long-read assemblies (Spearman's ρ: 0.08, P = 0.002). Unlike in (c-e), each circle in (g)
175 represents the percent of fragmentation for that BUSCO gene across all long- or short-read assemblies. Thus,
176 each gene is included twice (once for each technology). All 1,367 BUSCO genes in the OrthoDB v.10 Insecta
177 gene set (Kriventseva, et al. 2019) were used except one 2.02 Kb gene that was missing in >70% of assemblies
178 and subsequently removed from analysis and visualization. BUSCO gene lengths varied from 198 base pairs to
179 9.01 Kb.
180
181 The type(s) of sequence data used for genome assembly were obtained for ~82% of assemblies
182 (long-read = 126, short-read = 365; Table S1). Long-read assemblies were more contiguous than
183 short-read assemblies (Fig. 1b; *P,* Welch's T-test < 0.0001), averaging contig N50 values that
184 were ~4.4 Mb higher despite no difference in assembly size (*P,* Welch's T-test = 0.12; Fig. S4).
185 Gene regions were also far more complete in long-read assemblies (mean BUSCO
186 completeness = 96%, SD = 7%) versus those generated from short-reads (89.1%, SD = 19%; *P,*
187 Welch's T-test < 1e-8; Fig. 2c) with 70% fewer fragmented genes (*P,* Welch's T-test < 1e-11; Fig.
188 2d). Long-read assemblies, however, had ~2.6x more duplicated genes (4.4% vs. 1.7%; *P,*

189 Welch's T-test = 0.003; Fig. 2e). Longer BUSCO genes were also more likely to be fragmented
190 in both short-read (Spearman's ρ: 0.24, $P < 2.2e-16$) and long-read assemblies (Spearman's ρ:
191 0.08, $P = 0.002$; Fig. 2g) but they were less likely to be missing in both when compared to
192 shorter genes (short-read: Spearman's ρ: -0.08, $P = 0.002$; long-read: Spearman's ρ: -0.18, $P =$
193 9.7e-12; Fig. S5).
194
195 The rate at which new insect genome assemblies are becoming available is clearly accelerating
196 (Fig. 1c). Nearly 50% ($n = 292$) of the best-available insect assemblies were accessioned in
197 2019–2020 (Tables S1-S2). The same period also represented a high-water mark of contiguity
198 (mean contig N50, 2019–2020 = 1.77 Mb; Table S2). Much of the increase in contiguity was
199 driven by long-read assemblies which rose in frequency from 0% of all assemblies in 2011–2012
200 to 36.1% in 2019–2020. The contiguity of long-read assemblies also sharply increased in 2017
201 (Fig. S6, Table S2).
202
203 We have entered a new era of insect genome biology. Since 2019, a new species has had its
204 genome assembly deposited in GenBank every 2.3 days. These new assemblies are, on
205 average, markedly more contiguous than those of just a few years ago. As we continue
206 developing these resources, we offer four recommendations: first, we should recognize the
207 community-driven nature of these data and seek better integration between research groups and
208 consortia in terms of data sharing, best practices, and taxonomic focus. Progress towards these
209 goals is occurring (e.g., a proposed metric system for describing genome assembly quality with
210 associated benchmark standards from the Earth BioGenome Project, Lewin, et al. 2018) and will
211 accelerate as more researchers integrate these standards into their own workflows. Second, new
212 sequencing efforts should strive to balance sampling that fills taxonomic gaps and improves
213 existing resources with targeted sampling motivated by specific questions. Both approaches are
214 valuable and not mutually exclusive. The former—filling taxonomic gaps—is critical to broadly

215 understanding the evolution of insects, the most diverse animal group on Earth. While the
216 latter—targeted, question-driven sequencing—is critical to understanding specific aspects of
217 genome biology which are often best answered using dense sampling of specific groups.
218 Importantly, success for this recommendation will depend, in part, on our first recommendation.
219 Better integration and communication will limit redundancy of efforts where the same species'
220 genome is sequenced by multiple groups simultaneously. Third, we echo the findings of the
221 Vertebrate Genome Project (Rhie, et al. 2021)—long-read assemblies are vastly more
222 contiguous than short-read approaches—and recommend that these technologies be embraced
223 by insect genome scientists. And, fourth, as of 2019, only 40% of insect genome assemblies had
224 corresponding gene annotations in GenBank (Li, et al. 2019). Expanding and refining the
225 availability of gene annotations for insects will drive corresponding increases in the scale of
226 taxonomic comparisons that are possible for many analyses. Overcoming this challenge of
227 annotation quality and availability can be subdivided into two more specific calls: (1) whenever
228 possible, annotations should be made available alongside genome assemblies in GenBank or
229 similar public repositories and (2) researchers should consider using the NCBI Eukaryotic
230 Genome Annotation Pipeline (Thibaud-Nissen, et al. 2016) to limit variation introduced by
231 differing annotation approaches and maximize compatibility.
232
233 Beyond resource development, we must continue to leverage this data set to conduct new
234 studies of insect genome biology and evolution. These efforts are beginning to emerge and are
235 paying dividends. For instance, 76 arthropod genome assemblies were used to better
236 understand 500 million years of evolution by characterizing changes in gene and protein content
237 in a temporal and phylogenetic context, including the identification of novel gene families that
238 arose during diversification with links to key adaptations including flight (Thomas, et al. 2020).
239 Similarly, a study of 195 insect genomes revealed the high diversity of transposable elements
240 across insects with varying levels of conservation depending upon phylogenetic position (Gilbert,

241 et al. 2020). With genome assemblies representing 600+ taxa and ~480 million years of
242 evolution available in a public repository, the power and promise of insect genome research has
243 never been greater. While our focus was on insects, long-reads are likely revolutionizing genome
244 science in virtually all taxonomic groups with untapped genomic potential existing in public
245 repositories across the Tree of Life. The rise of long-read assemblies will, in particular, spur new
246 understanding of previously difficult to characterize aspects of the genome (e.g., genome
247 structure, highly repetitive regions). By continuing to build, curate, and make genomic resources
248 publicly available, we will gain tremendous insight into genome biology and evolution at broad
249 phylogenetic scales. We will also create a more inclusive and equitable discipline by expanding
250 access to resources for scientists whose participation has historically been limited by financial or
251 technological barriers.
252

253 **Acknowledgements:**
254 S.H. and J.L.K. were supported by NSF award OPP-1906015. J.H and S.U.P. were supported by
255 the LOEWE-Centre for Translational Biodiversity Genomics which is funded by the Hessen State
256 Ministry of Higher Education, Research and the Arts. J.S.S. was supported by an NSF
257 Postdoctoral Research Fellowship in Biology (DBI-1811930) and an NIH General Medical
258 Sciences award (R35GM119515) to A.M.L.
259

260 **Data availability:**
261 The data underlying this article are available in the Supplementary Materials, primarily in Table
262 S1, with associated scripts for analysis on GitHub:
263 https://github.com/pbfrandsen/insect_genome_assemblies.
264

265 **References:**

266 Adams MD, et al. 2000. The genome sequence of Drosophila melanogaster. Science 287: 2185-
267 2195.
268 Amarasinghe SL, et al. 2020. Opportunities and challenges in long-read sequencing data
269 analysis. Genome Biology 21: 1-16.
270 Bellinger PF, Christiansen KA, Janssens F 2020. Checklist of the Collembola of the World.
271 http://www.collembola.org.
272 Collins FS, Morgan M, Patrinos A 2003. The Human Genome Project: lessons from large-scale
273 biology. Science 300: 286-290.
274 Consortium AgG 2017. Genetic diversity of the African malaria vector Anopheles gambiae.
275 Nature 552: 96.
276 Gilbert C, Peccoud J, Cordaux R 2020. Transposable Elements and the Evolution of Insects.
277 Annual Review of Entomology 66.
278 Hotaling S, Kelley JL, Frandsen PB 2020. Aquatic insects are dramatically underrepresented in
279 genomic research. Insects 11: 601.
280 Hug LA, et al. 2016. A new view of the tree of life. Nature Microbiology 1: 1-6.
281 i5K Consortium 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human
282 health, agriculture, and the environment. Journal of Heredity 104: 595-600.
283 Kelley JL, et al. 2014. Compact genome of the Antarctic midge is likely an adaptation to an
284 extreme environment. Nat Commun 5: 4611. doi: 10.1038/ncomms5611
285 Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist,
286 bacterial and viral genomes for evolutionary and functional annotations of orthologs.
287 Nucleic Acids Res 47: D807-D811.
288 Lewin HA, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life.
289 Proceedings of the National Academy of Sciences 115: 4325-4333.
290 Li F, et al. 2019. Insect genomes: progress and challenges. Insect molecular biology 28: 739-
291 758.

292 McGee MD, et al. 2020. The ecological and genomic basis of explosive adaptive radiation.
293 Nature 586: 75-79.
294 McKenna DD, et al. 2019. The evolution and genomic basis of beetle diversity. Proceedings of
295 the National Academy of Sciences 116: 24729-24737.
296 Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science
297 346: 763-767. doi: 10.1126/science.1257570
298 Petersen M, et al. 2019. Diversity and evolution of the transposable element repertoire in
299 arthropods with particular reference to insects. BMC Evol Biol 19: 1-15.
300 Rhie A, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate
301 species. Nature 592: 737-746.
302 Robinson GE, et al. 2011. Creating a buzz about insect genomes. Science 331: 1386.
303 Sayers EW, et al. 2021. GenBank. Nucleic Acids Res 48: D84-D86.
304 Seehausen O, et al. 2014. Genomics and the origin of species. Nature Reviews Genetics 15:
305 176-192.
306 Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation
307 completeness. In. Gene prediction: Springer. p. 227-245.
308 Stork NE 2018. How many species of insects and other terrestrial arthropods are there on Earth?
309 Annual Review of Entomology 63: 31-45.
310 Thibaud-Nissen F, et al. 2016. The NCBI Eukaryotic Genome Annotation Pipeline. Journal of
311 Animal Science 94: 184-184.
312 Thomas GW, et al. 2020. Gene content evolution in the arthropods. Genome Biology 21: 1-14.
313 Wang X, et al. 2014. The locust genome provides insight into swarm formation and long-distance
314 flight. Nat Commun 5: 1-9.
315 Zhang Z-Q. 2011. Animal biodiversity: An outline of higher-level classification and survey of
316 taxonomic richness: Magnolia press.
317