Realizing the promise of biodiversity genomics with highly accurate long reads

Scott Hotaling[1]*, Edward R. Wilcox[2], Jacqueline Heckenhauer[3,4], Russell J. Stewart[5], and Paul B. Frandsen[3,6,7]*

Affiliations:

[1] Department of Watershed Sciences, Utah State University, Logan, UT, USA

[2] DNA Sequencing Center, Department of Biology, Brigham Young University, Provo, UT, USA

[3] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany

[4] Department of Terrestrial Zoology, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt 60325, Germany

[5] Department of Biomedical Engineering, University of Utah, Salt Lake City, UT, USA

[6] Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, USA

[7] Data Science Lab, Smithsonian Institution, Washington, DC, USA

***Authors for Correspondence:**

Scott Hotaling, Department of Watershed Sciences, Utah State University, Logan, UT, USA; Email: scott.hotaling1@gmail.com; Phone: (828) 507-9950

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

Abstract:

Generating the most contiguous, accurate genome assemblies given available sequencing technologies is a long-standing challenge in genome science. With the rise of long-read sequencing, assembly challenges have shifted from merely increasing contiguity to correctly assembling complex, repetitive regions of interest, ideally in a phased manner. At present, researchers largely choose between two types of long read data: longer, but less accurate

sequences, often generated with Oxford Nanopore (ONT) technology, or highly accurate, but shorter, reads typically generated with Pacific Biosciences HiFi. To understand how both technologies influence genome assembly and to clarify how scale of data (i.e., mean length and sequencing depth) influence outcomes, we compared genome assemblies for a caddisfly, *Hesperophylax magnus*, generated with ONT and HiFi data. Despite shorter reads and less coverage, HiFi reads outperformed ONT reads in all assembly metrics tested and allowed for accurate assembly of the repetitive ~20-Kb *H-fibroin* gene. Next, we quantified the influence of data type on genome assemblies across 6,750 plant and animal genomes. We show that HiFi reads consistently outperform all other data types for both plants and animals and may represent a particularly valuable tool for assembling complex plant genomes. To realize the promise of biodiversity genomics, we call for greater uptake of highly accurate long-reads in future studies.

Keywords: Insecta, Oxford Nanopore, PacBio, HiFi, caddisfly, genome biology

Significance statement:

Understanding how types of sequence data influence genome assembly is an important aspect of genome science. In general, more data–i.e., longer reads, greater depth of coverage–often yields better genome assemblies. However, it is unclear how highly accurate long-read sequence data (e.g., PacBio HiFi) compare to noisier long-read data. We showed that HiFi outperformed noisier long-read data for a caddisfly species in terms of assembly contiguity and resolution of the highly repetitive ~20-Kb *H-fibroin* gene. We also showed that this outperformance likely extends to all animals and plants via a field-wide meta-analysis. Thus, long-read accuracy should be emphasized in future genome studies.

Body:

As genome sequencing has been revolutionized by high-throughput sequencing, a general rule has emerged: more data—e.g., longer reads or greater depth of coverage—yields more

57 contiguous, accurate genome assemblies. This is particularly evident when read length is
58 considered; third-generation long reads, which are often tens or even hundreds of thousands of
59 base pairs in length, have dramatically improved genome assemblies across the Tree of Life
60 (Hotaling, et al. 2021a; Hotaling, et al. 2021b; Marks, et al. 2021; Rhie, et al. 2021). For
61 coverage, an increase can limit the impacts of erroneous read calls through more replication of
62 potential variants (Sims, et al. 2014). However, a shortcoming of second-generation, short-read
63 platforms (e.g., Illumina) for genome assembly is that no amount of data will allow for resolution
64 of repeat-driven gaps that exceed read lengths (Sims, et al. 2014). The power of long reads to
65 mitigate this issue is well-documented (e.g., Hotaling, et al. 2022) but not all long reads are
66 created equal; indeed, different platforms yield different length versus error profiles
67 (Amarasinghe, et al. 2020; De Coster, et al. 2021). This difference is particularly important since
68 some long, repeat-rich genomic regions–including many genes of phenotypic relevance–pose
69 assembly challenges even when long read data are used. Generally speaking, past evidence
70 would suggest that to resolve difficult genomic regions with long read data, longer reads at
71 greater depth of coverage will outperform shorter reads and/or less dense coverage. However, it
72 is unclear how this expectation jibes with read accuracy, particularly for two common types of
73 long-read data that are currently in use: longer but noisier long reads [e.g., Oxford Nanopore
74 (ONT)] and more accurate, but shorter, long reads (e.g., PacBio HiFi).
75
76 We inadvertently tested this prediction—that longer, higher coverage, and noisier reads would
77 outperform shorter, but more accurate, reads at lower sequencing depth—in a silk-producing
78 caddisfly, *Hesperophylax magnus*, and the highly repetitive *heavy fibroin chain* gene *(H-fibroin*).
79 Caddisflies share an evolutionary origin of silk with butterflies and moths, including the primary
80 protein component of silk, *H-fibroin* (Frandsen, et al. 2019). *H-fibroin* commonly spans ~20 Kb
81 and its amino acid sequence consists of conserved termini and a highly repetitive internal
82 region. Early efforts to assemble the complete *H-fibroin* gene with short reads were

unsuccessful due to its long repetitive region (Ashton, et al. 2013; Yonemura, et al. 2009). However, long-read assemblies have since yielded full-length sequences (Frandsen, et al. 2019; Kawahara, et al. 2022; Luo, et al. 2018).

In this study, we compared two long-read genome assemblies for *H. magnus* produced with two technologies: ONT and HiFi. For comparison, we considered genome-wide metrics as well as accurate assembly of the *H-fibroin* gene. We chose to focus on *H-fibroin* as a surrogate for complex but phenotypically important genes where we expected the benefits of highly accurate long reads to be most obvious. To estimate whether assembly outperformance with HiFi data was unique to our focal caddisfly or reflects a broader trend in genome biology, we performed a meta-analysis of contig N50, assembly length, and sequencing technology for all publicly available plant and animal genomes on GenBank. For our caddisfly case study and meta-analysis, highly accurate HiFi sequence data dramatically outperform all other types of sequence data.

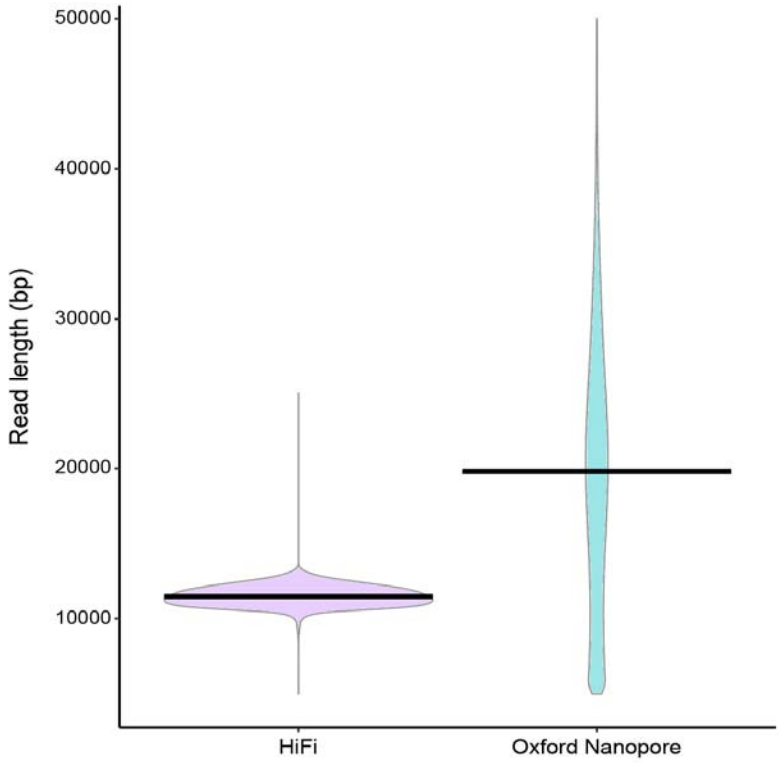
For our caddisfly genome comparison, we sequenced two individuals of *H. magnus* from the same population. For the first individual, we generated a combination of noisy ONT sequencing (R.9.4.1, LSK-109 ligation library prep kit) and Illumina sequencing (NovaSeq). For the second individual, we only generated HiFi reads via CCS sequencing on the PacBio Sequel II platform. We used Guppy v.5.0.11 to base-call the ONT reads and removed all reads under 5 Kb for further analysis. We used SMRTlink v.10 to generate HiFi reads (reads with quality >Q20). To assemble genomes for the two individuals, we tested a range of assemblers. For ONT, we tested Canu v.1.8 (Koren, et al. 2017), wtdbg2 v.2.4 (Ruan and Li 2020), and a hybrid approach with MaSuRCA (Zimin, et al. 2013). For HiFi, we tested Hifiasm (Cheng, et al. 2021) and HiCanu (Nurk, et al. 2020). We selected the best assembly based on contiguity and "Benchmarking Universal Single-Copy Orthologs" (BUSCO) scores. We ran BUSCO v.5.2.2

(Manni, et al. 2021) using the 1,367 reference genes in the OrthoDB v.10 Insecta gene set (Kriventseva, et al. 2019). To evaluate recovery and assembly of *H-fibroin*, we used tblastn to identify conserved terminal sequences from existing transcriptomes (Ashton, et al. 2013). We then extracted *H-fibroin* from both assemblies with 1,000 additional bps from each terminus and annotated it using Augustus v.3.3.2 (Stanke, et al. 2006). To visualize mismatches between reads and the assembly, we mapped raw reads to the assembled *H-fibroin* gene using Minimap2 (Li 2018) and visualized the results in Geneious 2022.0.2 (Kearse, et al. 2012).

To assess the influence of sequencing technology on genome assembly across the Tree of Life, we extracted all available genome assemblies for plants (class Embryophyta) and animals (class Metazoa) from GenBank using the “summary genome” function in v.10.9.0 of the NCBI Datasets command-line tool on 13 November 2021. We then used the “lineage” function of TaxonKit (Shen and Xiong 2019) to retrieve taxonomic information for all entries in our genome assembly list. Next, we gathered additional metadata (e.g., sequencing technology) for each entry using a custom web scraper script (modified from Hotaling, et al. 2021b). Next, we removed duplicate assemblies and alternative haplotypes for a given assembly that were identified through keyword searching in either the BioProject information or assembly title.

We binned assemblies into four sequencing technology categories: short-reads (e.g., Illumina), long-read ONT (ONT long-reads with or without short-reads), long-read PacBio (non-HiFi PacBio long-reads with or without short reads), or HiFi (any assembly where HiFi long-reads were used). To assist in categorization, we considered any assembly that was generated before 2017—when long-read assemblies began to emerge (Hotaling, et al. 2021b)—to be a short-read assembly. We also used self-reported information on the genome assembly algorithms used to classify assemblies. For instance, if an assembly only reported PacBio sequence data but a HiFi-specific assembler (e.g., Hifiasm) was used, we classified it as a HiFi assembly. After

135 binning, we removed any assembly for which the sequence data type used could not be
136 established. We tested for differences in the distributions of assembly size and contig N50
137 among plant and animals or our sequence type categories within the overarching plant or
138 animal grouping using Welch Two Sample T-tests or one-way ANOVAs followed by Tukey HSD
139 tests in R v3.6.3 (R Core Team 2021). While all statistical tests were performed on the
140 untransformed data, we visualized log-transformed comparisons using ggplot2 (Wickham 2011).
141



142

143 **Figure 1.** Violin plots of read lengths for the HiFi and Oxford Nanopore data sets used to assemble
144 *Hesperophylax magnus* genomes in this study. Width of the colored areas indicate numbers of reads according
145 to lengths on the y-axis. Dark lines represent the medians of each distribution.
146

147 The ONT library had a wider distribution of read lengths with a median of 19.9 Kb for 33.5 Gb of
148 raw data. The HiFi dataset had a median read length of 11.3 kb bp for 28 Gb of raw data (Fig.
149 1). The best ONT assembly (Genbank #GCA_016648045.1) was generated with MaSuRCA and
150 spanned 1.23 Gb. For HiFi, the best assembly was produced with Hifiasm (Genbank

#JAIUSX000000000) and was nearly identical in length at 1.22 Gb. However, we observed a dramatic difference in contiguity; the ONT assembly had a contig N50 of 0.7 Mb versus 11.2 Mb for the HiFi assembly. The ONT assembly also contained fewer complete BUSCOs (93%) versus the HiFi assembly (95.6%). In both assemblies, the full-length *H-fibroin* locus was present but the quality of the annotations differed greatly (Fig. 2). For ONT, Augustus annotated a dozen genes in the ~30Kb region, most of which did not include the characteristic repeats known from previous data. For HiFi, the annotation included a single gene with a single intron in the n-terminus region. The second exon was large (25.3 Kb) and fully in-frame including a well-resolved repetitive structure, giving high confidence in the accuracy of the assembly.

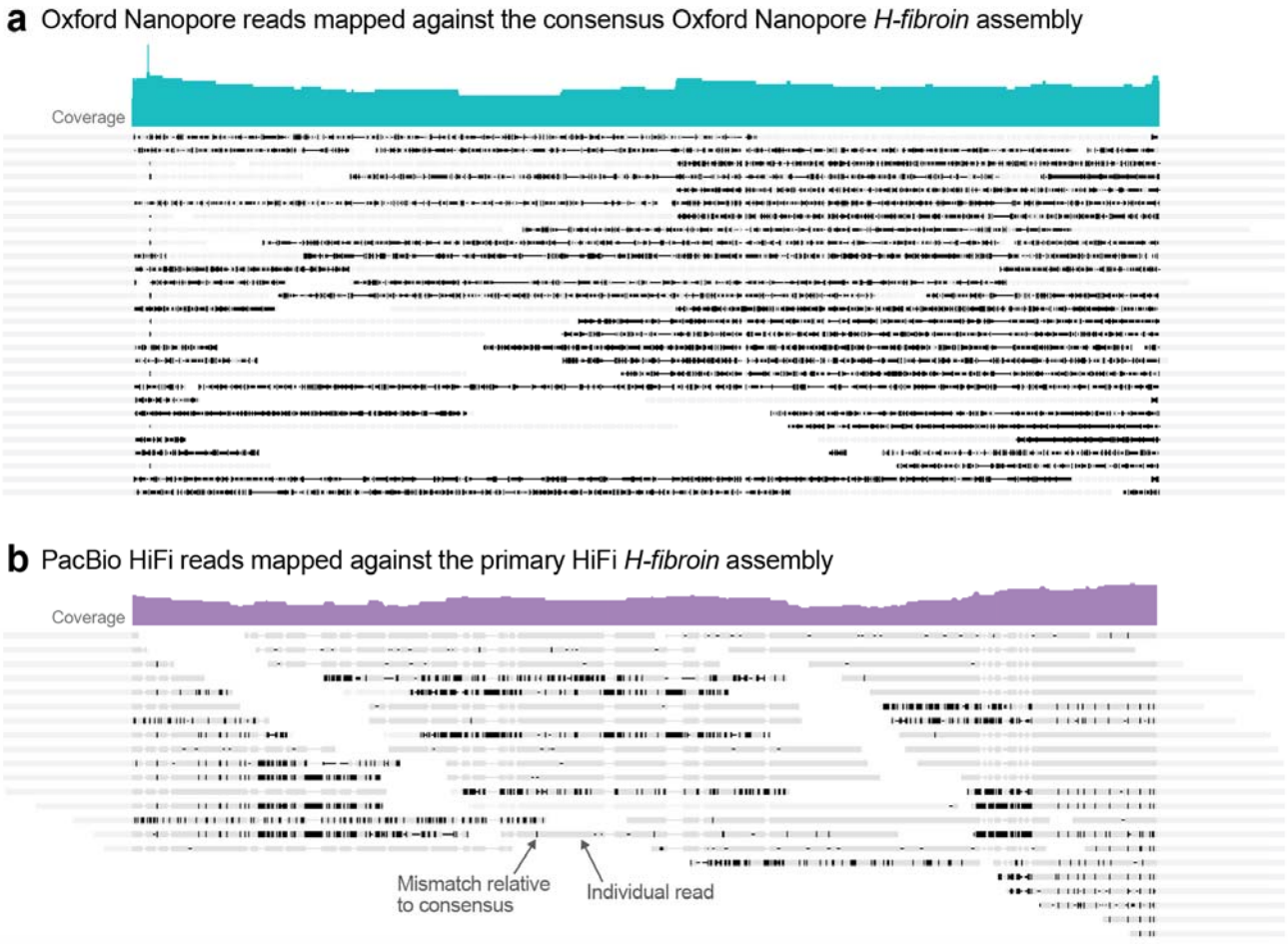


Figure 2. A case study comparing the capacity for two long-read sequencing technologies to assemble the complex gene underlying silk production in caddisflies, *H-fibroin.* (a) Raw Oxford Nanopore (ONT) reads

mapped to the consensus *H-fibroin* sequence from the ONT assembly. (b) Raw HiFi reads mapped to the primary *H-fibroin* sequence from the phased HiFi assembly. Dark lines indicate mismatches relative to consensus. In (b), mismatches reflect an *H-fibroin* length polymorphism that can be resolved by subsampling reads based on their allele-specificity.

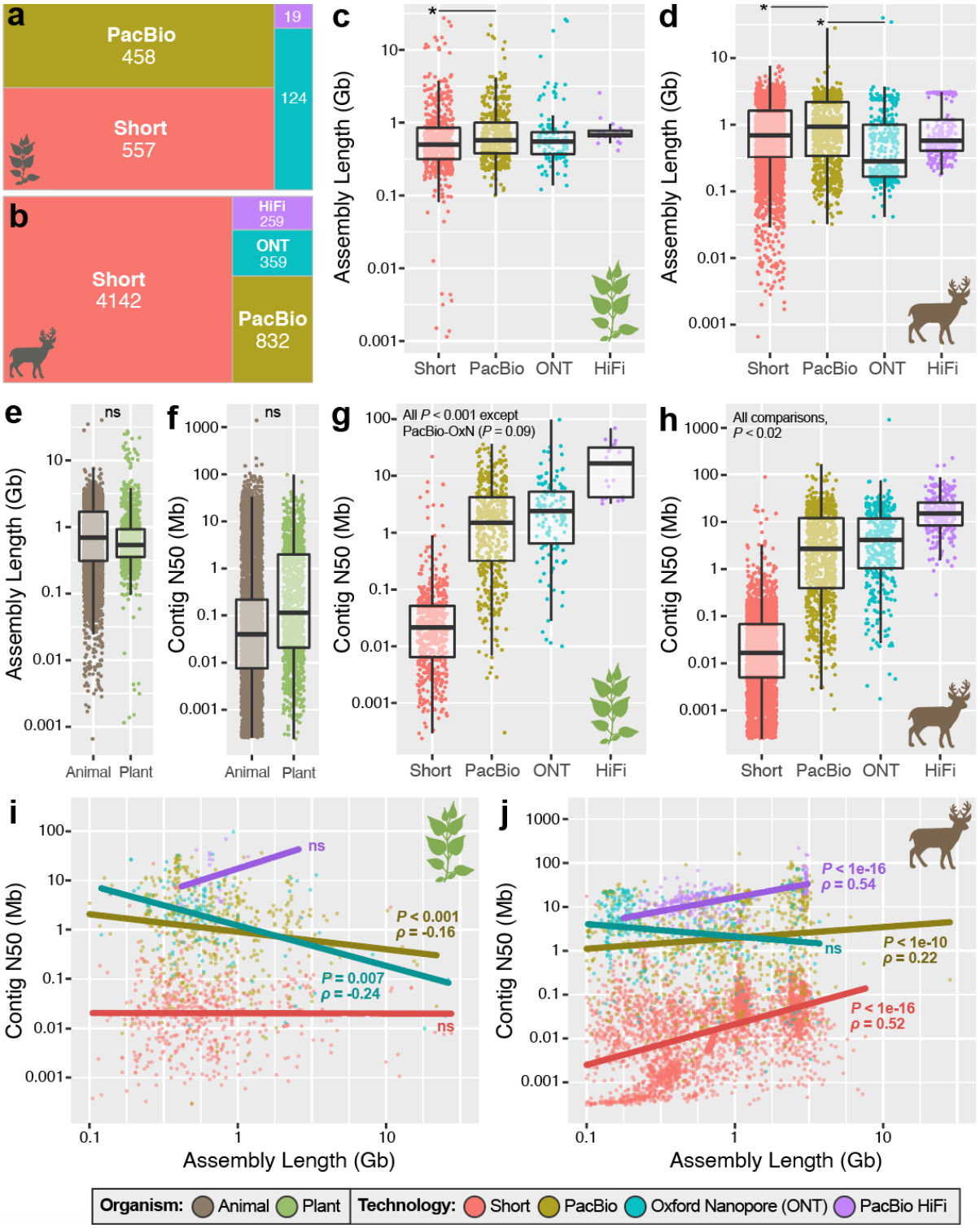


Figure 3. Sequencing technology representation and genome assembly quality across all animal and plant assemblies deposited in GenBank as of November 2021. A breakdown of the sequencing technology used for genome assemblies in (a) plants and (b) animals. Total assembly length broken down by sequencing technology for (c) plants and (d) animals. (e) Assembly length across all plant and animal genomes, regardless of technology. (f) Contig N50 across all plant and animal genomes and broken down by technology for (g) plants and (h) animals. Spearman's correlations between contig N50 and assembly length for (i) plants and (j) animals. For (i) and (j), correlation statistics were generated for the full data sets but for visualization, assemblies less than 0.1 Gb in length or with contig N50 > 1 Gb have been excluded. For (c-h), asterisks and thin dark lines indicate significant differences at $P < 0.05$.

After filtering, our data set contained 6,750 genome assemblies (animals = 5,592; plants = 1,158; Table S1). For plants, short-read assemblies (48.1%; *N* = 557) and long-read assemblies generated with non-HiFi PacBio data were similarly common (39.6%; *N* = 458; Fig. 3a). ONT assemblies, however, were much less common (10.7%) and HiFi assemblies were exceptionally rare, comprising just 1.6% of all assemblies (*N* = 19; Fig. 3a). For animals, the majority of assemblies were generated with short-read data (74.1%; *N* = 4,142). Non-HiFi long-read assemblies, generated with either PacBio or ONT reads, were also common, comprising 21.4% of the data set. HiFi assemblies were again the least common with just 259 assemblies (4.6%; Fig. 3b).

On average, animal genome assemblies were neither longer (*P,* Welch T-test = 0.80, Fig. 3e) nor more contiguous than those of plants (*P,* Welch T-test = 0.10, Fig. 3f). When broken down by technology, assembly lengths did differ for plants (*P*, one-way ANOVA = 0.006) and animals (*P*, one-way ANOVA < 0.001). For plants, only one length comparison was significantly different–non-HiFi PacBio assemblies were longer than those generated with short-reads (*P*, Tukey HSD = 0.005; Fig. 3c). For animals, non-HiFi PacBio assemblies were longer than both short-read (*P*, Tukey HSD = 0.002) and ONT assemblies (*P*, Tukey HSD < 0.001; Fig. 3d). In terms of assembly contiguity, contig N50 was significantly different for all comparisons in plants (*P*, Tukey HSD < 0.001) and animals (*P*, Tukey HSD < 0.02) with one exception: in plants, non-

HiFi PacBio assemblies were not different from ONT assemblies (*P*, Tukey HSD = 0.09; Fig. 3g-h). For both groups, HiFi reads dramatically outperformed all other long-read technologies tested. In plants, the average HiFi assembly was 501% more contiguous (mean contig N50 = 20.5 Mb) than assemblies generated with other long-reads (mean contig N50 = 4.1 Mb; Fig. 3g). For animals, HiFi assemblies were 226% more contiguous (mean contig N50 = 20.9 Mb) versus other long-read assemblies (mean contig N50 = 9.3 Mb; Fig. 3h).

When assembly size was compared to contiguity, striking patterns emerged. For plants, when non-HiFi long reads are used, contiguity declines with increasing assembly length ($P < 0.008$; Spearman's ρ, PacBio = -0.16, Spearman's ρ, ONT = -0.24; Fig. 3i). The same trend isn't present for short reads (*P*, Spearman's $\rho = 0.18$) nor HiFi (*P*, Spearman's $\rho = 0.37$). However, for HiFi, this lack of significance is likely a product of small sample size ($N = 19$; Fig. 3i). For animals, however, three of four read types (short, PacBio, HiFi) exhibit positive correlations between contig N50 and assembly length (*P*, Spearman's $\rho < 1e\text{-}10$) with the steepest trends for the two most accurate sequencing technologies: HiFi (Spearman's $\rho = 0.54$) and short-reads (Spearman's $\rho = 0.52$; Fig. 3j).

As high-throughput sequencing technologies have matured, so has a common strategy: to generate better assemblies, more data is better. While simplistic, this "more is better" approach has been supported by empirical data and echoed by genome sequencing overviews (e.g., Ekblom and Wolf 2014). The practicality of this approach has also been empirically observed in the long-read era. For instance, benchmarking of long-read assemblies in maize found that lower sequencing depths (< 30x) with mean read lengths less than 11 Kb yielded highly fragmented assemblies (Ou, et al. 2020). Stepping back, the outperformance of long-reads relative to short-reads in terms of basic assembly metrics is dramatic and independent of taxonomy (Hotaling, et al. 2021b; Marks, et al. 2021). Thus, support exists for the premise that

to generate the most high-quality assemblies, researchers should maximize depth of coverage and mean read length regardless of the technology being used.

However, since we cannot maximize perfectly accurate deep sequencing with the longest possible reads, achieving the best possible genome assemblies under the current landscape of sequencing technologies appears to require more nuance. Our results highlight that in at least one instance, a smaller amount of shorter, but more accurate, HiFi reads outperformed ONT data for genome assembly in the same taxon. Since this outperformance likely stems from base pair accuracy and the potential for sequence errors to confound assembly, we expected the benefit of highly accurate reads to scale with genome and/or gene-region complexity. The dramatic outperformance of HiFi reads when assembling the *H-fibroin* locus supports this expectation. Similar results have also been obtained for other taxonomic groups. For the leaf rust fungus, *Puccinia triticina*, HiFi reads outperformed noisier long-reads, particularly in areas of the genome that were difficult to assemble (Duan, et al. 2022). Among rice genomes, however, the results were less clear: ONT ultra-long reads yielded higher overall contiguity but more errors than HiFi (Lang, et al. 2020).

Moving beyond our case study to a large-scale comparison of animal and plant genome assemblies, key broader themes emerged. First, HiFi assemblies were significantly more contiguous than all other types of sequence data tested. And, given the benefits it affords, highly accurate long-read sequencing–e.g., HiFi–remains underrepresented, particularly in plant genetics. Second, all significant relationships between genome size and contiguity for sequencing technologies in animals were positive (Fig. 3j) whereas no correlation was positive in plants (Fig. 3i). This suggests that as genome size increases in plants, so too does complexity, likely at a rate that outpaces the capacity for modern assembly algorithms to assemble it. Notable, however, was the sharply positive trend for HiFi sequencing in plants

where contig N50 appears to rapidly increase with assembly length (Fig. 3i). While not statistically significant at $P < 0.05$–likely due to a low sample size–this pattern suggests that HiFi and similarly accurate long-read technologies may represent a valuable means to overcome challenges of genome complexity in plants. Finally, we acknowledge that while we focused on specific technologies in this study and found strong evidence for HiFi efficacy, our point is more about long-read accuracy than specific technologies. Indeed, other technologies–including ONT–will likely approach and may even surpass HiFi in terms of accuracy in the months and years to come.

With the rise of highly accurate long-read sequencing, little room remains for another revolution in genome assembly quality. It is now possible to generate reference-quality assemblies for virtually any species with modest resources. The only exceptions, for now, are species that are very small, difficult to obtain, and/or with exceptionally large or complex genomes. What is lacking now are appropriate metrics for contemporary genome assembly benchmarking. For instance, contig N50–the most common metric for assessing contiguity–scales with assembly length in highly accurate long-read assemblies (Fig. 3i-j) and thus, its upper limit is tied to chromosome length, making comparisons among groups difficult. For gene content assessment–i.e., BUSCO scores–one challenge lies in how accurately BUSCO scores reflect true gene content, particularly when more repeat-rich genes are considered. For instance, the median gene length in the 1,467-gene “Insecta” gene set is ~1 Kb (longest = 9.1 Kb; Hotaling, et al. 2021b) yet phenotypically relevant genes like *H-fibroin* can be *much* longer (i.e., >20 Kb). Thus, while BUSCO scores for the ONT and HiFi assemblies of *H. magnus* reflect marginal differences in their gene completeness (93% vs. 95.6%), the true gap is likely much greater. We expect differences in assembly quality to scale with genomic region complexity–a result that is not captured by BUSCO scores.

Acknowledgements:

J.H was supported by the LOEWE-Centre for Translational Biodiversity Genomics, which was funded by the Hessen State Ministry of Higher Education, Research and the Arts.

Data availability:

Both genome assemblies generated for this study are available on Genbank (#GCA_016648045.1, #JAIUSX000000000) as well as the raw reads used (ONT: SRX9290148; Illumina: SRX9290147; HiFi: SRR15840267). The full data set used for the animal and plant genome analysis is provided in Table S1.

References:

Amarasinghe SL, et al. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biology 21: 1-16.

Ashton NN, Roe DR, Weiss RB, Cheatham III TE, Stewart RJ 2013. Self-tensioning aquatic caddisfly silk: Ca2+-dependent structure, strength, and load cycle hysteresis. Biomacromolecules 14: 3668-3681.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18: 170-175.

De Coster W, Weissensteiner MH, Sedlazeck FJ 2021. Towards population-scale long-read sequencing. Nature Reviews Genetics 22: 572-587.

Duan H, et al. 2022. Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. Genome Biology 23: 1-27.

299 Ekblom R, Wolf JB 2014. A field guide to whole-genome sequencing, assembly and annotation.
300 Evol Appl 7: 1026-1042. doi: 10.1111/eva.12178

301 Frandsen PB, et al. 2019. Exploring the underwater silken architectures of caddisworms:
302 comparative silkomics across two caddisfly suborders. Philosophical Transactions of the Royal
303 Society B 374: 20190206.

304 Hotaling S, Desvignes T, Sproul JS, Lins LS, Kelley JL 2022. Pathways to polar adaptation in
305 fishes revealed by long-read sequencing. Mol Ecol.

306 Hotaling S, Kelley JL, Frandsen PB 2021a. Toward a genome sequence for every animal:
307 Where are we now? Proceedings of the National Academy of Sciences 118: e2109019118.

308 Hotaling S, et al. 2021b. Long-reads are revolutionizing 20 years of insect genome sequencing.
309 Genome Biology and Evolution evab138.

310 Kawahara AY, et al. 2022. Long-read HiFi Sequencing Correctly Assembles Repetitive heavy
311 fibroin Silk Genes in New Moth and Caddisfly Genomes. bioRxiv.

312 Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform
313 for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649.

314 Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer
315 weighting and repeat separation. Genome research 27: 722-736.

316 Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist,
317 bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic
318 Acids Res 47: D807-D811.

Lang D, et al. 2020. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. GigaScience 9: giaa123.

Li H 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.

Luo S, Tang M, Frandsen PB, Stewart RJ, Zhou X 2018. The genome of an underwater architect, the caddisfly Stenopsyche tienmushanensis Hwang (Insecta: Trichoptera). GigaScience 7: giy143.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38: 4647-4654.

Marks RA, Hotaling S, Frandsen PB, VanBuren R 2021. Representation and participation across 20 years of plant genome sequencing. Nature Plants 7: 1571-1578.

Nurk S, et al. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome research 30: 1291-1305.

Ou S, et al. 2020. Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nat Commun 11: 1-10.

R Core Team 2021. R: A language and environment for statistical computing.

Rhie A, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592: 737-746.

339 Ruan J, Li H 2020. Fast and accurate long-read assembly with wtdbg2. Nature Methods 17:
340 155-158.

341 Shen W, Xiong J 2019. TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit. bioRxiv:
342 513523.

343 Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP 2014. Sequencing depth and coverage: key
344 considerations in genomic analyses. Nature Reviews Genetics 15: 121-132.

345 Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids
346 Res 34: W435-W439.

347 Wickham H 2011. ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3: 180-185.

348 Yonemura N, Mita K, Tamura T, Sehnal F 2009. Conservation of silk genes in Trichoptera and
349 Lepidoptera. Journal of Molecular Evolution 68: 641-653.

350 Zimin AV, et al. 2013. The MaSuRCA genome assembler. Bioinformatics 29: 2669-2677.
351