Long-read sequencing and genome assembly of natural history collection samples and challenging specimens

Bernhard Bein [1,2,3#], Ioannis Chrysostomakis [4#], Larissa S. Arantes [5,6#], Tom Brown [5,6#], Charlotte Gerheim [1,2], Tilman Schell [1,2], Clement Schneider [7], Evgeny Leushkin [1,2], Zeyuan Chen [2], Julia Sigwart [1,2], Vanessa Gonzalez [8], N. Leena W.S. Wong [9], Fabricio R. Santos [10], Mozes P. K. Blom [11], Frieder Mayer [11], Camila J. Mazzoni [5,6], Astrid Böhne [4], Sylke Winkler [12,13], Carola Greve [1,2], Michael Hiller [1,2,3]*

[1] LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany
[2] Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt, Germany
[3] Institute of Cell Biology and Neuroscience, Faculty of Biosciences, Goethe University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany
[4] Center for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn, Adenauerallee 127, 53113 Bonn, Germany
[5] Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Königin-Luise-Straße 2-4, 14195 Berlin, Germany
[6] Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315 Berlin, Germany
[7] Senckenberg Research Institute, Am Museum 1, 02826 Görlitz, Germany
[8] Global Genome Initiative, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, USA
[9] Department of Aquaculture, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[10] Laboratório de Biodiversidade e Evolução Molecular, Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[11] Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstraße 43, 10115, Berlin, Germany
[12] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
[13] DRESDEN concept Genome Center, Technische Universität Dresden, 01062 Dresden, Germany

shared first authors
* michael.hiller@senckenberg.de

Abstract

Advances in long-read sequencing technology have shifted a key challenge in biodiversity genomics from DNA sequencing and genome assembly to obtaining suitable samples. Although flash-frozen samples remain the gold standard, they are often unattainable. In contrast, natural history collections worldwide harbor millions of ethanol-preserved samples, but such samples remain largely unutilized for long-read sequencing. Here, we use ethanol-preserved samples containing kilobase-sized DNA to show that amplification-free long-read sequencing can yield contiguous genome assemblies, which we exemplify using two catfish species. For other samples, where amplification-free protocols failed, we show that a modified protocol employing an alternative polymerase can successfully overcome issues related to long-read sequencing and PCR bias. For the maned sloth *Bradypus torquatus*, we demonstrate that this modified protocol together with long-range HiC data achieves a high-quality chromosome-level assembly of its 3.1 Gb genome, surpassing the previous 500 Mb genome size limit of the previous protocol. Moreover, our modified protocol also improves assembly quality and completeness for various other species that require amplification because of low sequencing performance or very small DNA amounts in tiny millimeter-sized organisms. Together, by highlighting collections as valuable but underexplored sample resources for long-read sequencing and by facilitating genome assemblies of small and challenging organisms, our study advances biodiversity genomics and global efforts to obtain reference genomes of all eukaryotes.

Introduction

High-quality genomes provide a powerful basis for understanding phylogenetic relationships, discovering fundamental principles of evolutionary processes, applying genomic methods to characterize, monitor and preserve biodiversity, and ultimately revealing the genetic blueprint underlying the fascinating diversity of life on our planet. Therefore, generating high-quality genomes of eukaryotic species has become a central goal in biological sciences (Lewin et al. 2018). Advances in short-read sequencing technology (with Illumina as the most prominent platform) enabled sequencing the genomes of a few thousand eukaryotes to date (Zoonomia Consortium 2020; Feng et al. 2020; Ronco et al. 2021; Kuderna et al. 2023). However, because eukaryotic genomes are often large and rich in repetitive DNA sequences, genome assembly from short reads ranging from 100 to 300 bp in size results in fragmented and incomplete assemblies (Zoonomia Consortium 2020; Feng et al. 2020; Kuderna et al. 2023), posing many limitations to downstream analyses. To generate highly contiguous genomes, the field has shifted to adopting long-read sequencing platforms from PacBio or Oxford Nanopore Technologies that can sequence DNA fragments with sizes of many kilobase pairs (kb) at once. Such long reads span most genomic repeats and do not suffer from the sequencing biases of short-read platforms in regions with very high or low GC content. Thus, long reads result in highly contiguous and complete genome assemblies, culminating in telomere-to-telomere assemblies (Nurk et al. 2022; Chen et al. 2023), and consequently enable complete genome annotations and comprehensive analyses (Warren et al. 2020; Kautt et al. 2020; Jebb et al. 2020; Rhie et al. 2021; Blumer et al. 2022; Vollger et al. 2022; Osipova et al. 2023; Shao et al. 2023).

A key limitation of long-read sequencing is the availability of high molecular weight DNA, ideally with fragment sizes of 50 kb or more. To obtain samples delivering such DNA, the best practice is to acquire fresh samples (which may require sacrificing an individual), flash-freeze in liquid nitrogen, and preserve samples permanently at -80°C until DNA is extracted (https://www.earthbiogenome.org/sample-collection-processing-standards). Such protocols are not practical or not possible for (i) rare or endangered species, where sacrificing even a single living individual is not permitted, (ii) species which are difficult to sample in the field (e.g. cetaceans), or (iii) situations where liquid nitrogen and freezer capacity is not practicable (e.g. in remote areas). Therefore, sample availability is a key challenge for biodiversity genomics (Blom 2021).

An alternative to get access to valuable or rare species that comprise Earth's biodiversity are samples that are available in museums and other research collections that house millions of specimens worldwide, including samples from extinct species (Johnson et al. 2023). As one example demonstrating the value of such collections for biodiversity genomics, several hundred bird genomes have been generated from dry samples stored in museum collections (Feng et al. 2020). However, since DNA of dry samples often exhibits various degrees of degradation, short-read sequencing was the only feasible technology, resulting in fragmented bird assemblies with an average contiguity of 43 kb (measured as contig N50 values, which state that 50% of the assembly consists of contiguous DNA segments – called contigs – of at least that size). Nevertheless, this and other studies using dry museum

samples and short-read sequencing approaches, including marker-based sequencing and genome skimming, provided valuable insights into taxonomy, phylogenomics and conservation genomics (Espeland et al. 2018; Heinicke et al. 2023; Tan et al. 2023).

In addition to dry material, collections worldwide also contain many millions of samples preserved in ethanol. In comparison to the logistical challenges associated with bringing liquid nitrogen to field trips and transporting flash-frozen samples without breaking the cold chain, preserving and transporting collected samples in ethanol is a notably simpler task. Since kilobase-sized DNA can be preserved in such samples (Mulcahy et al. 2016; Dahn et al. 2022), we explored whether ethanol-preserved samples are also suitable for long-read sequencing. We reasoned that even if DNA fragment sizes are substantially shorter than 50 kb, successfully sequencing reads of a few kilobases in size increases read length by at least an order of magnitude compared to short-read sequencing approaches, which in turn will improve assembly contiguity. In particular, we focused on the PacBio high-fidelity (HiFi) read protocol that instead of generating error-prone reads from “as long as possible” DNA fragments, sequences medium-sized fragments (10-15 kb) but with a high base accuracy of 99.8% (Wenger et al. 2019). HiFi sequencing enables assemblies that are both more contiguous and have a higher base accuracy than assemblies obtained with longer but more error-prone reads (Wenger et al. 2019; Sharma et al. 2021; Hiller et al. 2023; Chen et al. 2023), making it a promising technology to apply to ethanol-preserved samples.

In this study, we explored the utility of ethanol-preserved samples from collections for HiFi sequencing. Although we encountered DNA degradation and sample contamination as expected problems in some samples, we also successfully demonstrate that HiFi reads can be obtained from ethanol-preserved samples containing kilobase-sized DNA, either using amplification-free protocols or by using a modified amplification-based protocol that effectively addresses issues associated with HiFi sequencing and PCR bias. Using this modified protocol, we generate a high-quality assembly of the 3.1 Gb genome of the maned sloth *Bradypus torquatus*, demonstrating that the previous genome size limit of 500 Mb can be substantially extended. Beyond collection samples, we further show that our modified protocol improves the contiguity of assemblies of species belonging to other phyla such as Mollusca (Gastropoda, Bivalvia) and Arthropoda (Collembola), where amplification is typically required for long-read sequencing. The efficacy of this protocol facilitates genome assembly of challenging taxa and suggests that collections can serve as valuable sample sources for long-read sequencing.

Results

HiFi sequencing of ethanol-preserved samples with an amplification-free protocol

To investigate the effectiveness of PacBio HiFi sequencing from ethanol-preserved collection samples, we focused on vertebrates and used samples of four mammals (three-toed jerboa

149 *Dipus sagitta*, pen-tailed treeshrew *Ptilocercus lowii*, long-eared flying mouse *Idiurus macrotis*,
150 maned sloth *Bradypus torquatus*), two squamates (European blind snake *Xerotyphlops*
151 *vermicularis*, slow worm *Anguis fragilis)* and two fishes (the catfish species *Cathorops nuchalis*
152 and *Cathorops wayuu*), all lacking a genome assembly (Table 1, Supplementary Table 1). All
153 samples were collected in the field and preserved in technical or 96% ethanol. Apart from the
154 maned sloth and the catfishes, all samples were kept at room temperature. The samples of
155 the maned sloth and catfish were kept most of the time in a freezer at -20ºC; however, in
156 contrast to flash-frozen samples, freezing did not occur immediately after sampling and they
157 were kept at room temperature for extended periods of time, including during transportation.
158
159 We used a modified Circulomics Nanobind disk and a phenol/chloroform based protocol for
160 the extraction of genomic DNA (Methods). For *Dipus sagitta*, *Ptilocercus lowii*, and
161 *Xerotyphlops vermicularis*, we did not obtain a sufficient amount of DNA (< 400 ng) and/or
162 DNA fragments were shorter than 0.18 kb (Supplementary Table 1), showing that DNA is too
163 degraded to proceed with library preparation. For four species (*Anguis fragilis*, *Idiurus*
164 *macrotis*, *Cathorops nuchalis*, and *Cathorops wayuu*), the amount of DNA and the DNA
165 fragment sizes were sufficient to prepare an amplification-free PacBio low input library
166 (Supplementary Table 1). We sequenced all libraries on a PacBio Sequel IIe system, disabling
167 on-board calling of HiFi reads and instead applying the computationally expensive
168 DeepConsensus method (Baid et al. 2023) to maximize HiFi read yield and length. For
169 *Bradypus torquatus*, we did not obtain enough DNA and therefore proceeded with a PacBio
170 ultra-low input library (see below).
171
172 For the two catfish species, *Cathorops nuchalis* and *Cathorops wayuu*, we sequenced two
173 SMRT cells each and obtained HiFi reads with an average length of 8,832 and 8,783 bp,
174 respectively, providing a total of 43.8 and 41.2 Gb, which corresponds to a coverages of ~17X
175 and ~16.5X (Supplementary Table 1). Using HiFiasm with different parameters (Cheng et al.
176 2021), we were able to obtain a contig assembly for both species with a total length of 2.6 and
177 2.59 Gb and a contig N50 value of 3.2 and 2.1 Mb (Supplementary Table 2). To assess gene
178 completeness, we used compleasm (Huang and Li 2023) with the set of 3,640 ray-finned fish
179 (Actinopterygii) near-universally conserved genes (ODB10) (Manni et al. 2021), which showed
180 that 96.65% of these genes are fully present in the primary assembly of *C. nuchalis* and 95.6%
181 in that of *C. wayuu*. Although additional HiFi data would be needed to improve contiguity and
182 HiC data would be required to scaffold the contigs into chromosome-level scaffolds, our catfish
183 samples exemplify that an adequate genome assembly can be obtained from 10-year-old,
184 ethanol-preserved tissues.
185
186 In contrast to the catfish, we obtained very low sequencing yields for *Idiurus macrotis* and
187 *Anguis fragilis*, with only 0.3 Gb and 0.04 Gb of HiFi data (Supplementary Table 1). Quality
188 metrics showed that the polymerase N50 raw read lengths were very short and the local base
189 rates were low. For example, while the library from *Anguis* met the requirements for PacBio
190 sequencing with a mean fragment length of 12.2 kb, both the local base rate of 1.64 (expected
191 ~2.8) and the polymerase N50 raw read length of 32.3 kb (expected at least 200 kb) are very
192 low and insufficient to produce HiFi reads of most DNA fragments in the library. This indicates

193 that factors such as DNA damage, metabolites bound to the DNA, or contaminants
194 precipitated with the DNA inhibit the polymerase, highlighting sequencing challenges for
195 ethanol-preserved samples stored at room temperature.
196

species	year sampled	preservation	type of sample
northern three-toed jerboa (*Dipus sagitta*)	2006 & 1961	technical ethanol, room temperature	muscle, skin
pen-tailed treeshrew (*Ptilocercus lowii*)	1967		muscle
long-eared flying mouse (*Idiurus macrotis*)	2000		skin with hair
maned sloth (*Bradypus torquatus*)	2003	likely pure ethanol, mostly at -20°C (otherwise room temperature)	clogged blood
European blind snake (*Xerotyphlops vermicularis*)	2004 & 2011	technical ethanol, room temperature	skin and muscle
slow worm (*Anguis fragilis*)	2021		muscle from tail cross-section
catfish (*Cathorops nuchalis*)	2014	pure ethanol, transported multiple times at room temperature until final storage at -20°C	fin
catfish (*Cathorops wayuu*)	2014		fin

197 **Table 1:** Overview of the species and samples.
198
199

200 HiFi sequencing with the amplification-based ultra-low input protocol

201 We reasoned that a PCR-based amplification step prior to library preparation could render the
202 *Idiurus macrotis* and *Anguis fragilis* samples amenable to sequencing, as this procedure
203 should yield intact DNA devoid of potential polymerase-inhibiting metabolites. To this end, we
204 applied the PacBio ultra-low input library protocol (PacBio Biosciences 2020) to the samples
205 of *Idiurus macrotis* and *Anguis fragilis*. Although this protocol was originally designed for small
206 specimens providing very limited DNA amounts (Schneider et al. 2021) and is recommended
207 only for genome sizes of up to 500 Mb, the protocol includes a PCR amplification step using
208 two different undisclosed polymerases targeting DNA with average and high GC contents,
209 respectively. For simplicity, we refer to these polymerases as “A” and “B” in the following. We
210 also generated an ultra-low input library for the *Bradypus torquatus* sample that did not contain
211 enough DNA for the low input protocol.
212
213 Indeed, for *Idiurus macrotis* and *Anguis fragilis*, sequencing another SMRT cell each produced
214 10 and 19.6 Gb in HiFi reads with an average HiFi read length of 4,854 bp and 7,552 bp. The

first SMRT cell for *Bradypus torquatus* yielded 29.9 Gb in HiFi reads with an average HiFi read length of 10,850 bp (Supplementary Table 1).

For *Idiurus macrotis* and *Anguis fragilis*, we investigated whether a DNA repair step applied to the DNA extract before preparing the ultra-low library would increase HiFi read length and yield (Methods). In contrast to the previous sequencing results, adding the DNA repair step produced shorter HiFi reads (average read length 4,270 vs. 4,854 bp for *Idiurus macrotis* and 5,609 vs. 7,552 bp for *Anguis fragilis*) and a lower yield (6.4 vs. 10 Gb for *Idiurus macrotis* and 12.6 vs. 19.6 Gb for *Anguis fragilis*), suggesting that the DNA repair process is not advantageous for these samples (Supplementary Table 1).

Next, we investigated whether the sequenced DNA was contaminated with bacteria, fungi or other microorganisms. While little contamination was found in the *Bradypus torquatus* sample, the *Anguis fragilis* data had higher levels of contamination, and the vast majority of the sequencing data obtained from the *Idiurus macrotis* sample were contamination (Supplementary Figure 1). High levels of contamination were also detected for three additional ethanol-preserved samples, where we directly applied the ultra-low input protocol: Russian desman (*Desmana moschata*) sampled in 1947, Hazel dormouse (*Muscardinus avellanarius*) sampled in 2016, and a *Anguis fragilis* sample from 1878 (Supplementary Tables 1, 3). Together, while sample contamination is another challenge related to samples obtained from collections, our tests also show that amplifying DNA in the ultra-low input protocol prior to library preparation can enable PacBio HiFi sequencing of samples where the amplification-free low input library protocol failed.

PCR bias in the current protocol prevents high-quality assemblies of larger genomes

To investigate the feasibility of using the ultra-low input protocol to obtain a high-quality assembly of a genome that substantially exceeds the recommended size limit of 500 Mb, we focused on the maned sloth that has an estimated genome size exceeding 3 Gb and showed a low level of contamination. To obtain sufficient read coverage for genome assembly, we generated two additional libraries using the PacBio ultra-low protocol and sequenced four additional SMRT cells. In total, all five SMRT cells provided 140.2 Gb of HiFi reads, a total coverage of ~45X, with an average read length of 10.6 kb. However, using this data, we only obtained an assembly with a contig N50 of 405 kb (Figure 1, brown dashed line), which is unexpectedly low as similar read coverages typically yield mammalian assemblies with contig N50 values exceeding several megabases. Using compleasm (Huang and Li 2023) with the set of 11,366 near-universally conserved eutheria genes (ODB10) showed that only 85.3% of these genes are fully present in our assembly. Similarly, using TOGA (Kirilenko et al. 2023) to determine how many of the 18,430 ancestral placental mammal coding genes have an intact reading frame, revealed that only 68% of the ancestral genes are intact. Together, this indicates not only a low assembly contiguity but also a high level of incompleteness.

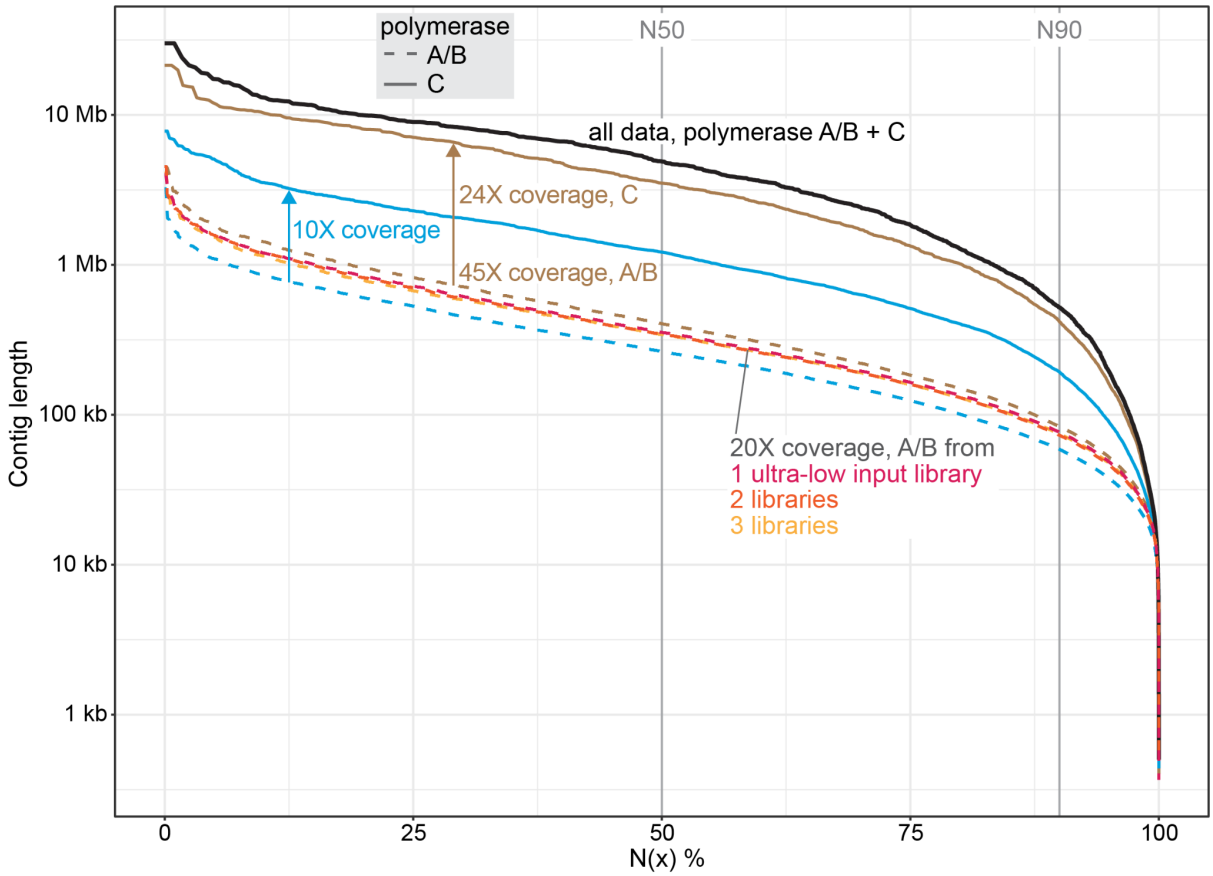

Figure 1: Contiguity of *B. torquatus* assemblies generated with data from ultra-low input libraries prepared with polymerases A/B and/or C at different coverages.
Assembly contiguity visualized as N(x) graphs that show contig sizes on the Y-axis, for which x percent of the assembly consists of contigs of at least that size (N50 and N90 values are indicated). Assemblies involving polymerase C read data are shown as solid lines, assemblies generated from polymerase A/B data are shown as dashed lines. Colors refer to different comparisons.

To further investigate the reasons for the poor quality of this assembly, we aligned our *Bradypus torquatus* HiFi reads against the high-quality genome of a related sloth species, *Choloepus didactylus* (Rhie et al. 2021). Despite both species being separated for 30 My (Gibb et al. 2016), we observed that 84.3% of the *Choloepus didactylus genome* was covered with *B. torquatus* HiFi reads at an average coverage of 38X. Inspecting the mapped reads in a genome browser revealed larger genomic regions, often spanning many kilobases, that completely lack any mapped reads (Figure 2). Since several of these regions contain highly-conserved genes, we reasoned that these read dropouts are probably not caused by high divergence between the sloth species. Instead, it is likely that despite relying on two polymerases, the PacBio ultra-low protocol has PCR bias on larger genomes, resulting in genomic regions that lack any reads.

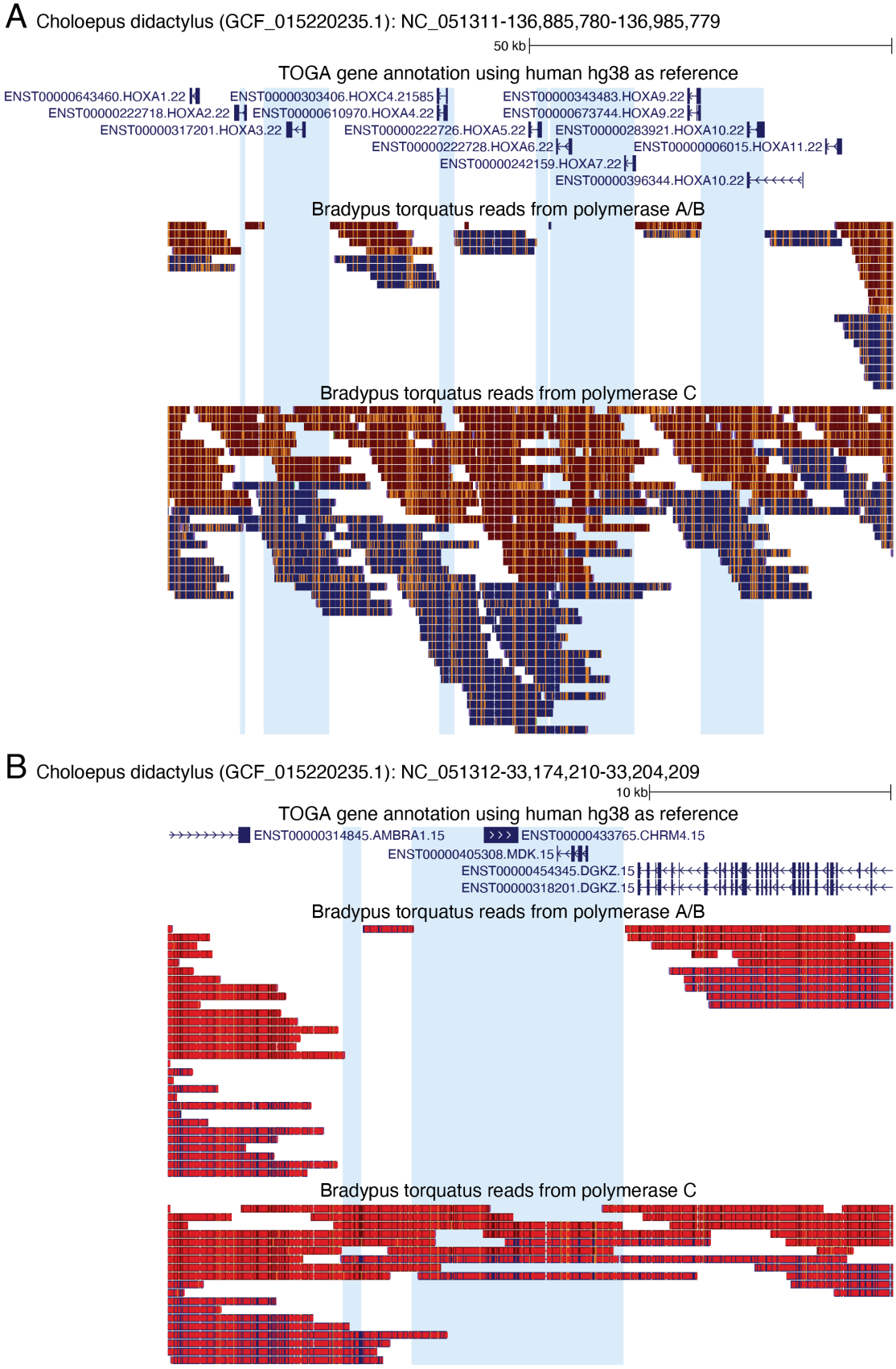

Figure 2: PCR bias in reads produced with polymerase A/B.
UCSC genome browser screenshots of the *Choloepus didactylus* assembly, together with the TOGA gene annotation and mapped HiFi reads of *B. torquatus* produced either with polymerase A/B or polymerase C.

(A) In the *HOXA* gene cluster, several regions, often covering parts or entire *HOX* genes, lack any reads produced with polymerase A/B (highlighted in blue). In contrast, these regions have a coverage of HiFi reads produced with polymerase C, which is sufficient for assembly.
(B) While reads produced with polymerase A/B do not cover the *CHRM4* and *MDK* genes, polymerase C reads cover the entire locus.

A different polymerase alleviates PCR bias and enables highly-complete assemblies of larger genomes

To alleviate PCR bias, we adapted the ultra-low input protocol and used a different polymerase, KOD Xtreme™ Hot Start DNA Polymerase. According to the specification sheet, this polymerase amplifies DNA fragments up to 24 kb at high fidelity, including templates with up to 90% GC content, which could help to overcome the underrepresentation of very low or high GC regions of the PacBio ultra-low input protocol. For simplicity, we refer to this polymerase as “C” in the following. Using a single library, we sequenced another three SMRT cells for the maned sloth, providing 91.7 Gb (corresponding to an additional 24X coverage) of reads with an average length of 10.2 kb.

Performing genome assembly using all HiFi reads obtained with polymerase A/B and C, produced a 3.13 Gb assembly with a contig N50 of 4.88 Mb (Figure 1, black line, Supplementary Table 4), which is 12 times higher than the previous assembly generated from reads obtained with polymerase A/B. Consistently, gene completeness estimated with compleasm improved from 85.3% to 96.4% and the percentage of intact ancestral placental mammal genes inferred with TOGA increased from 68 to 88.6%. Furthermore, mapping the polymerase C HiFi reads to the *C. didactylus* assembly covered the regions that completely lacked any read before (Figure 2). This confirms that previous read dropouts were not caused by sequence divergence between both sloth species or selective degradation of certain genomic regions in our sample, but by PCR bias associated with the polymerases in the PacBio ultra-low input protocol.

To directly compare the effect of polymerase A/B vs. C, taking differences in read coverages from the individual SMRT cells out of the equation, we downsampled our data to equal coverage and performed a number of tests (Supplementary Table 4). Since DNA fragments generated by polymerase A and B are pooled during the library preparation, we cannot investigate the effect of those two polymerases individually. Using an equal, downsampled coverage of ~10X, we found that the assembly produced from only polymerase C reads outperformed the assembly produced from only polymerase A/B reads by exhibiting a substantially higher contiguity (contig N50 1.22 Mb vs. 264 kb) and gene completeness (89.3% vs. 77.0% completely detected genes) (Figure 1, light blue lines). Remarkably, an assembly obtained from the complete polymerase C read data is substantially better than an assembly obtained from the complete polymerase A/B read data (contig N50 3.5 Mb vs. 405 kb, 96.4% vs. 85.3% completely detected genes) (Figure 1, brown lines), despite the polymerase C data having a substantially lower coverage (24X vs. 45X for polymerase A/B).

We next investigated how the number of libraries produced with polymerase A/B influences assembly, as additional libraries may increase complexity and reduce bias. However, sampling an equal coverage of 20X from either one, two or three libraries results in very similar assemblies in terms of contiguity and gene completeness (Figure 1, red/orange/yellow lines; Supplementary Table 4), indicating that inherent bias of polymerase A/B hampers assembly quality that cannot be overcome by producing several libraries.

Together, these tests show that *B. torquatus* assemblies generated with polymerase C reads are substantially better. To our knowledge, we provide the first high-quality contig assembly of a 3.1 GB genome that was produced using an adapted ultra-low input protocol combining polymerase A/B and C.

Chromosome-level assembly of the maned sloth

To obtain a final scaffolded assembly of *Bradypus torquatus*, we used the Arima HiC protocol to generate 97.5 Gb in long-range read pair data. Using the automated scaffolding software yahs (Zhou et al. 2023) and manual curation, our contig assembly could be scaffolded into chromosome-level scaffolds (Figure 3A). This final assembly consists of 2,915 scaffolds and 5,022 contigs. The scaffold N50 and N90 values are 157 Mb and 61.3 Mb, respectively (Figure 3B). The contig N50 and N90 values are 4.75 Mb and 519 kb, respectively. Using Merqury (Rhie et al. 2020) with the HiFi reads, we estimate a high base accuracy (QV=46.7), which represents an upper bound as the HiFi reads were also used for assembly. The assembly has a compleasm gene completeness score of 97.3% based on the eutheria ODB10 database with n=11,366 genes (Supplementary Table 4) and contains 90.72% of ancestral placental mammal genes.

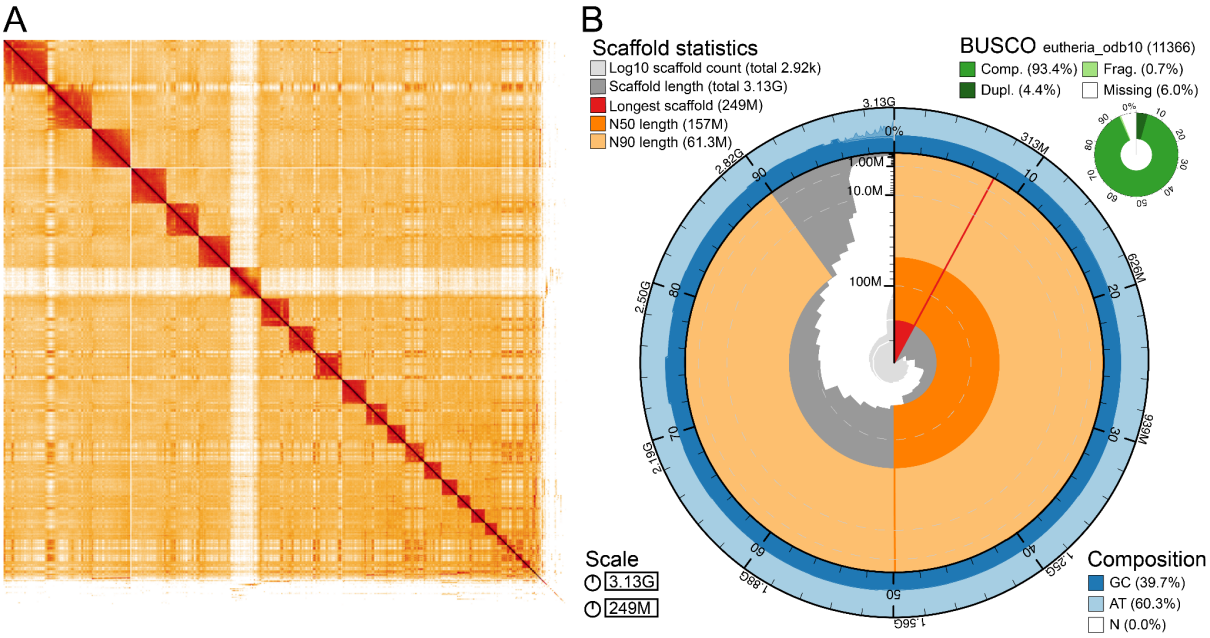

Figure 3: Chromosome-scale assembly of *Bradypus torquatus*.
(A) HiC interaction map after automated scaffolding by yahs and manual curation. The HiC map shows interactions in 3-dimensional space between two regions of the genome. Darker colors indicate a higher number of interactions. The region of low interaction between scaffold 7 and all other scaffolds indicates

this scaffold is the X chromosome, which was confirmed as this scaffold aligns to the human X chromosome.
(B) Snail plot showing lengths of all scaffolds, together with the longest scaffold (red), and the N50 (dark orange) and N90 length (light orange). The outer ring shows the GC content of the genome.

In comparison to existing genome assemblies of xenarthran species, our final assembly clearly outperforms the short-read based assembly of the sloth *Choloepus hoffmanni* in terms of contiguity and the number of intact ancestral placental mammal genes (Figure 4). Although other long-read based xenarthran assemblies, which were most likely generated from flash-frozen samples obtained from zoos and captive colonies, have even higher contiguities, our *Bradypus torquatus* assembly is a valuable addition for xenarthran and, more generally, mammalian comparative genomics.

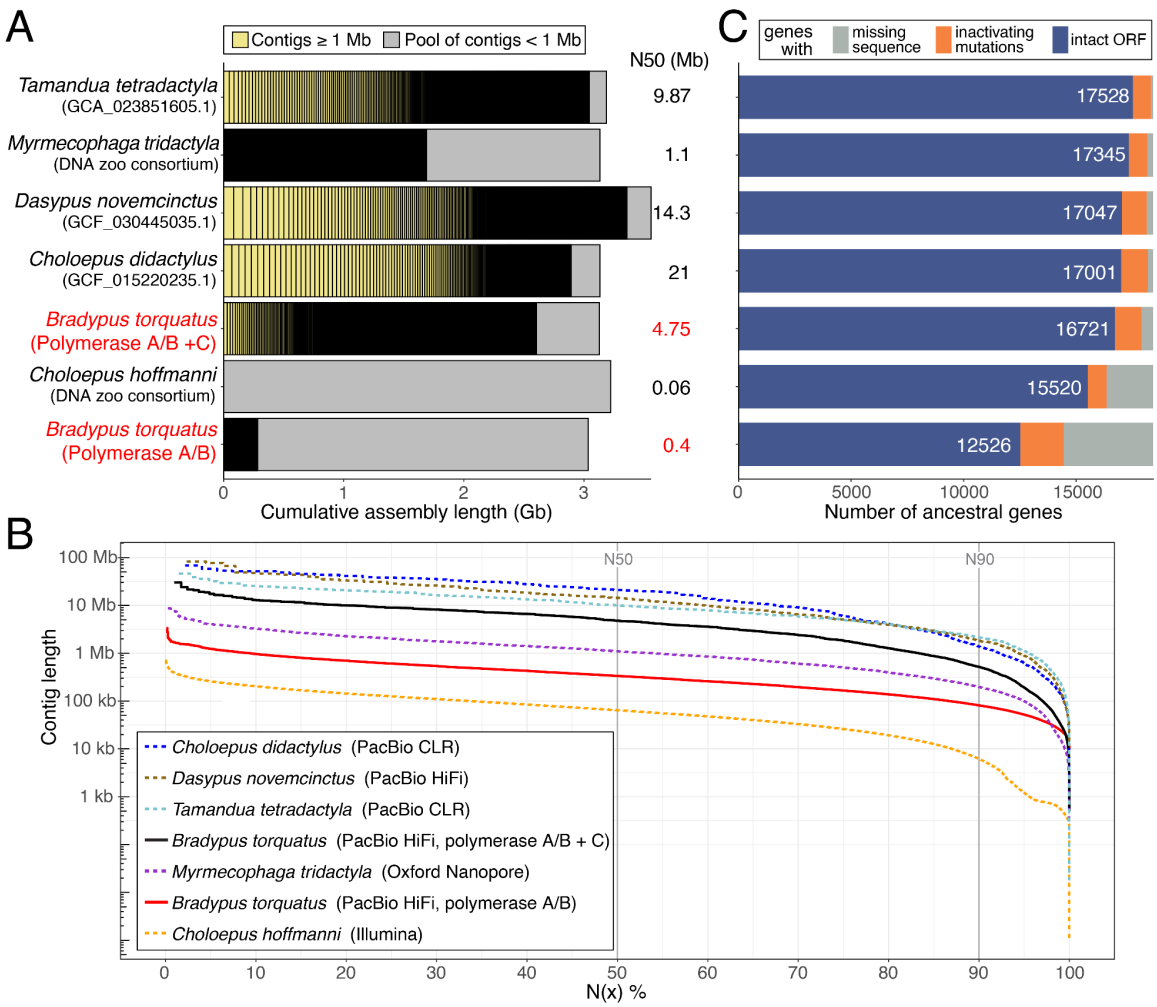

Figure 4: Comparison of xenarthran genome assemblies.
(A) Visualization of contig sizes of available xenarthran genome assemblies. Each bar represents the total assembly size. Contigs shorter than 1 Mb are not visualized individually, but shown as the grey portion of each bar. The final *B. torquatus* assembly and its preliminary assembly generated only from polymerase A/B reads are in red font. Assembly source or accession is listed in this panel, the sequencing technology used is listed in the inset in panel B.

380 (B) Visualization of assembly contiguity as an N(x) graph, showing contig sizes on the Y-axis, for which
381 x percent of the assembly consists of contigs of at least that size. Assembly order in the legend (inset)
382 is sorted by contig N50 value.
383 (C) TOGA classification of 18,430 ancestral placental mammal genes showing the number of genes
384 that have an intact reading frame (blue bar, number is given in white font), inactivating mutations (e.g.
385 frameshifts, stop codon, splice site mutations or exon deletions; orange bar), or missing coding
386 sequence parts often caused by assembly gaps or fragmentation (gray bar). Assemblies are sorted by
387 the number of intact genes.
388
389

390 Polymerase C improves assemblies for various species

391 We next explored whether polymerase C can also help to improve assemblies of other
392 species, using samples not obtained from collections. To provide a fair comparison, we
393 randomly downsampled the larger data set to obtain an equal coverage of HiFi reads
394 generated with polymerases A/B and C. To compare these polymerases for another mammal,
395 we used the human HG002 sample and generated assemblies for both human haplotypes.
396 Using an equal coverage of 23.5X, the polymerase A/B read data produced a 2.96 Gb
397 assembly for haplotype 1 with a contig N50 value of 642 kb, whereas the polymerase C data
398 generated a 3.03 Gb assembly with a substantially higher contig N50 value of 2.8 Mb, a 4.4
399 fold increase in contiguity. Consistently, gene completeness assessed with compleasm
400 (mammalia_odb10) increased substantially from 81.2 to 98.6%. Similar results were obtained
401 for the haplotype 2 assembly, where the polymerase A/B read data produced a 2.9 Gb
402 assembly with a contig N50 value of 558.8 kb and a gene completeness of 77.3%, whereas
403 the polymerase C read data produced a 3 Gb assembly with a contig N50 value of 2 Mb and
404 a gene completeness of 97.8%.
405
406 Next, we explored the application of polymerase C to three non-vertebrate taxa covering two
407 additional phyla, Mollusca (two taxonomic classes: Gastropoda and Bivalvia) and Arthropoda
408 (Collembola), using taxa where genome sequencing efforts typically rely on the amplification-
409 based protocols because of low sequencing performance with low input protocol or very small
410 DNA amounts.
411
412 For the sacoglossan gastropod *Elysia timida* (Mollusca), previous sequencing libraries created
413 with the low input protocol resulted in very poor sequencing performance. Therefore, we
414 applied the ultra-low input protocols, and compared two SMRT cells produced with polymerase
415 A/B, providing 16.6 and 20.8 Gb yield in reads with an N50 length of 6.5 and 5.8 kb, to one
416 SMRT cell produced with polymerase C, providing 23 Gb yield in reads with an N50 length of
417 7 kb (Supplementary Table 5). After subsampling to equal read coverage of 26.4X,
418 polymerase A/B and C read data generated assemblies with similar contig N50 values of 347.1
419 kb for polymerase A/B and 331 kb for polymerase C (Figure 5, Supplementary Table 5). Using
420 all polymerase A/B read data with a coverage of 42.5X increased the contig N50 value to
421 472.6 kb. Importantly, adding the 23 Gb of polymerase C reads, increased the contig N50
422 value 1.4 fold to 675.8 kb (Figure 5). While the gene completeness (metazoa_odb10) of 97.7

423 and 97.8% is similar between these assemblies, polymerase C data helped to improve
424 assembly contiguity for this mollusc.
425
426 To understand why polymerase C alone does not result in a more contiguous assembly, we
427 mapped both polymerase A/B and C reads to the *Elysia timida* assembly with highest
428 contiguity, generated from all read data. This showed that both polymerases A/B and C exhibit
429 bias; however, bias of one polymerase can be compensated by reads of the other
430 (Supplementary Figure 2), indicating that these polymerases may have taxon-specific
431 differences.
432
433 For the marine bivalve *Scintilla philippinensis* with an estimated genome size of 1.3 Gb, we
434 compared assemblies produced from 22.3 Gb of reads obtained from ultra-low input libraries
435 using polymerase A/B or C, which corresponds to a coverage of 17.1X. While the polymerase
436 A/B reads produced a 1.77 Gb assembly with a contig N50 value of 43.3 kb, the polymerase
437 C read data produced a 1.88 Gb assembly with a 1.2 fold increased contig N50 value of 52.1
438 kb. Gene completeness (metazoa_odb10) improved slightly from 89.1% (polymerase A/B) to
439 89.9% (polymerase C). Combining all polymerase A/B and C read data (coverage of 36.1X)
440 produced a 1.86 Gb assembly with an even higher contig N50 value of 75.1 kb (Figure 5,
441 Supplementary Table 5) and a higher gene completeness of 93.5%. Mapping polymerase A/B
442 and C reads to the assembly generated with all data also revealed regions that were covered
443 only by reads from one polymerase (Supplementary Figure 3A,B). While polymerase C reads
444 improved assembly contiguity of both mollusc species, the resulting assemblies have a
445 comparatively low contiguity, highlighting the challenges of sequencing molluscan DNA.
446
447 We next tested our adjusted protocol on a species having a very small body size, where
448 amplification of the limited amount of genomic DNA is required for long-read sequencing and
449 genome assembly (Schneider et al. 2021). We used an ethanol-preserved, whole single
450 specimen of the springtail *Podura aquatica* (Arthropoda: Collembola), which has a body size
451 of only 1.5 mm and an expected genome size of 200-300 Mb. The polymerase A/B run yielded
452 13.9 Gb of HiFi reads with an N50 length of 9.6 kb. The polymerase C run yielded 21.7 Gb of
453 reads but with a lower N50 read length of 5.7 kb, which is likely explained by sequencing DNA
454 one year after the initial extraction (the entire specimen was used for the initial DNA
455 extraction). Strikingly, at an estimated coverage of ~50X, the polymerase A/B read data
456 produced a 278.5 Mb assembly with a contig N50 value of only 919 kb, whereas the
457 polymerase C data generated a 269.3 Mb assembly with a contig N50 value of 2.77 Mb (Figure
458 5, Supplementary Table 5). This represents a 3 fold increase in contiguity, despite the
459 polymerase C reads being substantially shorter. Gene completeness (arthropoda_odb10)
460 increased slightly from 92.8% for polymerase A/B assembly to 93.4% for polymerase C
461 assembly. Combining all polymerase A/B and C read data resulted in a 284.7 Mb assembly
462 with an even higher contig N50 value of 5.74 Mb and the same gene completeness of 93.4%.
463 Similar to *Elysia* and *Scintilla*, aligning reads to the most contiguous assembly showed
464 complementary coverage dropouts (Supplementary Figure 3C,D).
465

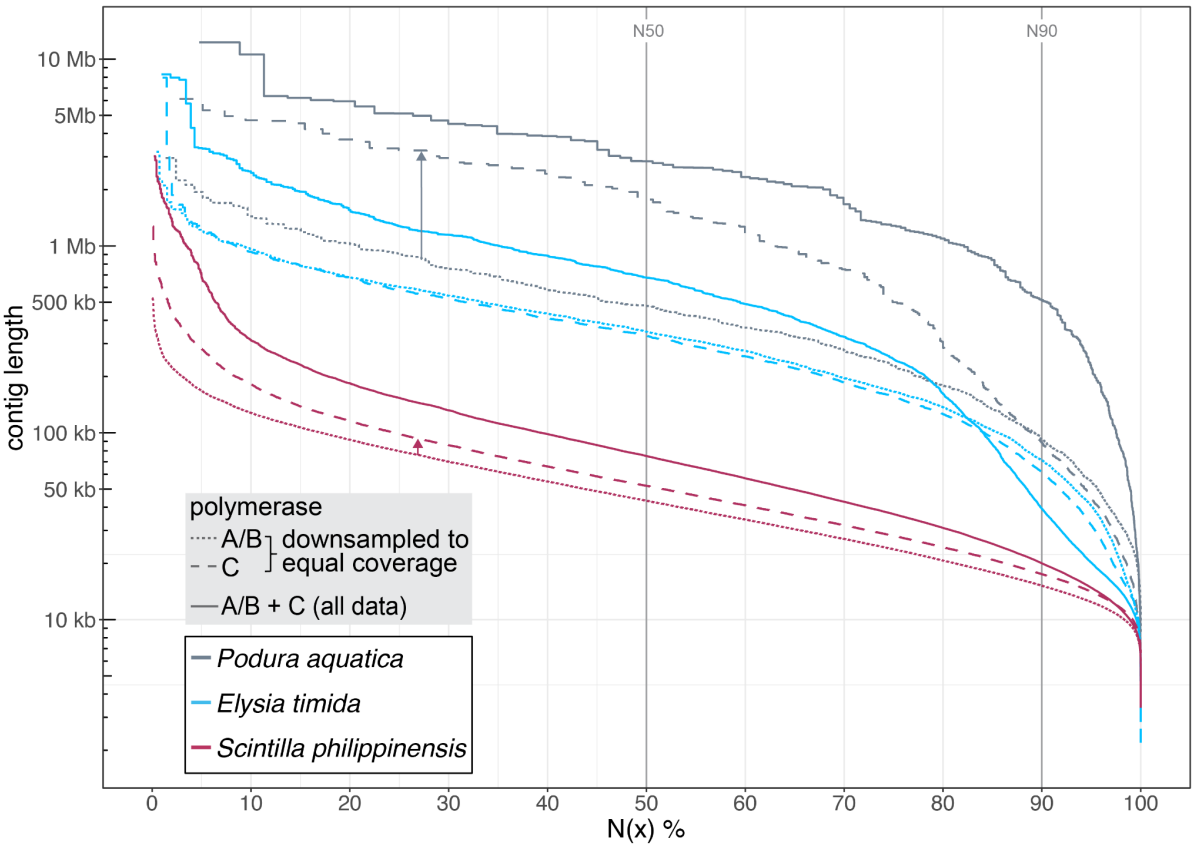

Figure 5: Impact of polymerase C on assemblies of mollusc and collembola species. Assembly contiguity visualized as N(x) graphs that show contig sizes on the Y-axis, for which x percent of the assembly consists of contigs of at least that size (N50 and N90 values are indicated). Assemblies are generated with an equal (downsampled) coverage of reads from polymerase A/B (dotted lines) and C (dashed lines). Assemblies generated with all data are shown as solid lines. Colors refer to different species.

Together, these tests confirm that polymerase C improves the assembly contiguity and sometimes gene completeness for a broad range of species, including species that rely on amplification-based library preparation protocols because their small size does not provide enough DNA from a single individual or because naturally-occurring metabolites presumably inhibit the polymerase during sequencing.

Discussion

Our investigation into utilizing collection samples for long-read sequencing confirms that ethanol-preserved samples can contain kilobase-sized DNA, long enough for long-read sequencing (Mulcahy et al. 2016; Dahn et al. 2022). For the two catfish species, we found that amplification-free protocols generated sequencing data sufficient to generate assemblies with contig N50 values surpassing 2 Mb. Application of amplification-free protocols is recommended whenever feasible, as they will not suffer from PCR bias. Our other tests indicate that mammal or reptile samples may necessitate amplification-based protocols. It

490 remains to be investigated for which taxonomic groups amplification-free protocols are
491 generally successful, and how this is influenced by sample age and storage temperatures. We
492 demonstrate that PCR bias associated with the amplification-based PacBio ultra-low input
493 protocol can be overcome or at least mitigated by employing an alternative polymerase. As a
494 proof of concept, the highly-contiguous 3.1 Gb genome assembly of *B. torquatus* shows that
495 a modified amplification-based protocol can produce high-quality assemblies of gigabase-
496 sized genomes.
497
498 Our study has a number of implications. First, the modified ultra-low input protocol improves
499 genome assembly of small specimens, where amplification is a requirement to obtain enough
500 DNA for sequencing. For example, the contiguity of the *Podura aquatica* genome increased
501 to an N50 of 5.7 Mb, and thus substantially exceeds the minimum standards of 100 kb set by
502 the Earth Biogenome Project for small species with limited DNA amounts (Lewin et al. 2018).
503 The modified protocol will likely not only be beneficial for species with diminutive body sizes
504 that represent a very large but mostly uncharacterized part of Earth's biodiversity, but also in
505 cases where only very limited amounts of material from non-lethal samplings (biopsies from
506 human patients or bat wing punches) are available. Second, long-read sequencing remains a
507 challenge for molluscs and other taxonomic groups, where satisfactory sequencing outputs
508 often require amplification-based protocols. Although achieving highly contiguous assemblies
509 with megabase contig N50 values remains challenging for these species, our investigations
510 suggest that employing a combination of different polymerases can at least help to improve
511 assembly contiguity. Third, while the PacBio ultra-low input protocol was previously limited to
512 genome sizes of up to 500 Mb, the successful application of the modified protocol to *B.*
513 *torquatus* with its 3.1 Gb genome extends its applicability to a broad range of species with
514 larger genome sizes. Together, the improved efficiency of the modified ultra-low input protocol
515 opens avenues for generating contiguous genomes across various species.
516
517 Our study raises the question of finding polymerases with minimal bias. While our tests with
518 *B. torquatus* and human indicate that polymerase C shows satisfactory performance for
519 mammals, we found that polymerase C also appears to exhibit bias for samples of molluscs
520 and collembola, albeit a different bias compared to polymerase A/B (Supplementary Figures
521 2, 3). Anticipating that DNA amplification will constitute a key step in the genome sequencing
522 procedure for numerous collection samples, challenging species, and species with diminutive
523 body sizes, future investigations could focus on identifying the most appropriate polymerase
524 or combination of polymerases that exhibit minimal bias for specific taxonomic groups.
525
526 More generally, our work suggests that collections can complement flash-frozen material as a
527 sample source for biodiversity genomics, especially for species that are hard to sample
528 because of rarity, protection status or other reasons. Thus, natural history collections as
529 extensive archives of biodiversity can help to achieve the ambitious goals of generating
530 reference genomes for all life on Earth.

Material and Methods

Sample sources

For *Bradypus torquatus*, we used a sample of ~50 mg of clogged blood, preserved in ethanol. This sample was collected under Brazilian license SISGEN number AF86294 and CITES number 138261. For *Idiurus macrotis*, we used ~12 mg of skin with hair preserved in technical ethanol at room temperature. For the *Anguis fragilis*, we used ~51 mg (for the one collected in 2021) and ~3 mg (for the one collected in 1878) of muscle tissue from a tail cross-section. Both samples were preserved in technical ethanol at room temperature. For both *Cathorops* species, we used fin samples stored in ethanol in frozen collections at the LIB Bonn. Originally, fin clips were taken in 2014, immediately placed into ethanol, but subsequently transported multiple times at room temperature until final storage at -20°C. For *Desmana moschata*, we used ~9 mg of muscle and skin tissue that was preserved in ethanol at room temperature. For *Muscardinus avellanarius*, we used ~19 mg of foot tissue that was preserved in technical ethanol at room temperature. For *Dipus sagitta*, we used ~12 mg (individual 95545) and ~16 mg (individual 95541) of muscle tissue and ~30 mg (individual 56492) of skin. All three samples were preserved in technical ethanol at room temperature. For *Ptilocercus lowii*, we used ~8 mg of muscle tissue that was preserved in technical ethanol at room temperature. For *Xerotyphlops vermicularis*, we used ~5 mg (individual collected in 2004) and ~3 mg (individual collected in 2011) of skin and muscle tissue preserved in technical ethanol at room temperature. For *Elysia timida*, we used a whole specimen (~1 cm body length) from our living culture. For *Scintilla philippinensis*, we used ~20 mg of muscle tissue preserved in ethanol, collected in Johor Malaysia under a collaboration agreement between Senckenberg and Universiti Putra Malaysia. For *Podura aquatica*, we used a single whole specimen (~1.5 mm body length) killed and preserved in 96% ethanol. Two libraries were produced either with polymerase A/B or polymerase C (below), and while the polymerase A/B experiment was done within the month following DNA extraction, the polymerase C experiment was conducted one year after DNA extraction, using DNA preserved at -20°C in TE buffer. Supplementary Table 1 lists sample sources, accessions and additional details.

DNA extraction

High molecular weight (HMW) gDNA was extracted from ethanol-preserved tissues of *Bradypus torquatus*, using a modified protocol version of the Circulomics Nanobind Tissue Big DNA kit, including the ethanol removing step described in ‘Guide and overview – Nanobind tissue kit’. We retrieved gDNA bound to the Nanobind disk as well as unbound gDNA in the precipitation solution. The gDNA bound to the Nanobind disk was eluted after several washing steps. The unbound gDNA in the precipitation solution was precipitated by centrifugation (18.000 xg for 30 min at 4°C). The resulting pellet was washed twice with 75% ice-cold ethanol, air dried for 20 min at room temperature and resuspended in 1x elution buffer. For both gDNA extractions we performed standard quality control, which involved Qubit quantification, Nanodrop measurement, and pulse field gel electrophoresis making use of the Femto Pulse system (Agilent).

574 For *Idiurus macrotis*, *Desmana moschata*, *Muscardinus avellanarius*, *Cathorops nuchalis*,
575 *Cathorops wayuu*, *Xerotyphlops vermicularis, Dipus sagitta*, *Ptilocercus lowii* and the two
576 *Anguis fragilis* samples, gDNA was extracted according to the protocol of (Sambrook and
577 Russell 2001). DNA concentration and DNA fragment length were assessed using the Qubit
578 dsDNA BR Assay kit on the Qubit Fluorometer (Thermo Fisher Scientific) and the Genomic
579 DNA Screen Tape on the Agilent 4150 TapeStation system (Agilent Technologies). For *Elysia*
580 *timida* and *Scintilla philippinensis*, gDNA was extracted using a CTAB-based method (Murray
581 and Thompson, 1980) including a pre-wash with sorbitol. The MagAttract HMW DNA Kit from
582 Qiagen was used to extract gDNA from *Podura aquatica.*
583
584 All details on the DNA yield and DNA fragment sizes can be found in Supplementary Table 1.
585

586 Low input PacBio HiFi library preparation

587 We prepared low input PacBio HiFi libraries according to the instructions of the SMRTbell
588 Express Prep Kit v2.0, except for the libraries of *Cathorops nuchalis* and *Cathorops wayuu*
589 which were prepared with the SMRTbell prep kit v3.0.
590

591 Ultra-low input PacBio HiFi library preparation

592 PacBio ultra-low input HiFi libraries were prepared with the SMRTbell Express Template Prep
593 Kit 2.0 according to the 'Procedure & Checklist - Preparing HiFi SMRTbell® Libraries from
594 Ultra-Low DNA Input' (PN 101-987-800 Version 02). To reduce potential PCR bias of
595 polymerase A/B, we used in our modified protocol a third PCR reaction, making use of
596 Polymerase C (KOD Xtreme™ Hot Start DNA Polymerase, Merk PN 71975), which is
597 optimized for the amplification of long strands and GC-rich DNA templates.
598
599 The amplified DNA from two PCR reactions with polymerase A and B was pooled equimolarly.
600 PCR fragments from polymerase C amplification were kept separately and processed
601 independently from the pooled fragments produced with polymerase A and B. Purified and
602 pooled amplified DNA libraries were size selected to remove smaller fragments
603 (Supplementary Table 1).
604
605 For *Anguis fragilis* and *Idiurus macrotis*, we prepared two additional libraries with DNA extracts
606 to which a DNA repair step was applied using the Sequential Reaction Protocol for PreCR
607 Repair Mix (New England BioLabs) prior to the actual library preparation.
608

609 PacBio sequencing

610 A total of 27 SMRT 8M cells were sequenced in CCS mode using the PacBio Sequel II / IIe
611 instrument. For low input libraries, where possible, libraries were loaded at an on-plate
612 concentration of 80 pM using adaptive loading and the Sequel II Binding kit 2.2 or 3.2 (Pacific
613 Biosciences, Menlo Park, CA). Ultra-low input libraries were loaded with up to 80 pM on plate
614 where possible using the SEQUEL II binding kit 2.2 or 3.2, and the sequencing kit 2.0. Pre-
615 extension time was 2 hours, run time was 30 hours.
616

617 ## HiC for scaffolding the *B. torquatus* assembly

618 Chromatin conformation capture was done using the Arima HiC+ Kit (Material Nr. A410110),
619 following the user guide for animal tissues (ARIMA-HiC 2.0 kit Document Nr: A160162 v00)
620 and processing 28 mg of tissue with the standard input approach. The subsequent Illumina
621 library preparation followed the ARIMA user guide for Library preparation using the Kapa
622 Hyper Prep kit (ARIMA Document Part Number A160139 v00). The barcoded HiC libraries
623 were run on an S4 flow cell of a NovaSeq6000 with 200 cycles.
624

625 ## Comparing polymerase A/B and C read assemblies

626 Aiming to evaluate the impact of libraries generated with polymerase A/B vs. C on the genome
627 assembly quality, we combined different datasets with varying coverages, library complexities
628 (number of libraries) and polymerase combinations (only A/B, only C, and A/B+C). For tests
629 that did not involve all read data, we randomly subsampled reads. Subsequently, we
630 assembled the read data into a contig assembly, as described below, and compared the
631 summary metrics, including contig N50, number of contigs and gene completeness. All results
632 are listed in Supplementary Tables 4 and 5.
633

634 ## Contig assembly

635 HiFi reads were called using a pipeline consisting of PacBio's tools ccs 6.4.0
636 (https://github.com/PacificBiosciences/ccs) and actc 0.3.1
637 (https://github.com/PacificBiosciences/actc) as well as samtools 1.15 (Danecek et al. 2021)
638 and DeepConsensus 0.2.0 or 1.2.0 (Baid et al. 2023). All commands were executed as
639 recommended in the respective guide for DeepConsensus
640 (https://github.com/google/deepconsensus/blob/v0.2.0/docs/quick_start.md; e.g. ccs --all). To
641 remove PCR adapters and PCR duplicates, which might originate from the whole genome
642 amplification during the ultra-low library preparation, PacBio's tools lima 2.6.0
643 (https://github.com/PacificBiosciences/barcoding) with options "--num-threads 67 --split-bam-
644 named --same" and pbmarkdup 1.0.2-0 with options "--num-threads 67 --log-level INFO --log-
645 file pbmarkdup.log --cross-library --rmdup"
646 (https://github.com/PacificBiosciences/pbmarkdup) were applied to samples prepared with the
647 ultra-low library preparation protocol. For the Catfish samples *Cathorops nuchalis* and *C.*
648 *wayuu* that were sequenced using the low-input library preparation protocol, PacBio
649 sequencing adapters were removed with HiFiAdapterFilt (Sim et al. 2022). The resulting reads
650 were merged and then decontaminated with kraken2 v. 2.1.3 (Wood et al. 2019) using the
651 kraken2 PlusPFP database downloaded in March 2023, with a confidence score of 0.51.
652

653 After HiFi calling, we used hifiasm v0.19.5 (Cheng et al. 2021, 2022) to assemble HiFi reads
654 obtained from the *Cathorops nuchalis, C. wayuu, Idiurus, Anguis, Elysia, Scintilla* and *Podura*
655 samples. For the two catfish samples *Cathorops nuchalis* and *C. wayuu*, because of
656 suboptimal performance with default parameters, we tested several hifiasm options before
657 deciding which parameters produce the best assembly in terms of gene completeness and
658 contiguity (Supplementary Table 2). To this end, we estimated the genome profile of these
659 two species with FastK (https://github.com/thegenemyers/FASTK) and Genescope.FK

660 (https://github.com/thegenemyers/GENESCOPE.FK) with k=30 to find the homozygous peak
661 that was then passed to hifiasm (Supplementary Table 2). In all other cases, we applied default
662 parameters with strict haplotig purging (-l3 parameter), and for the *Elysia* sample, we
663 additionally used available Arima HiC data for assembly phasing.
664
665 Contiguity statistics were calculated with Quast 5.0.2 (Mikheenko et al. 2018), gfastats v. 1.3.6
666 (https://github.com/vgl-hub/gfastats) and Merqury.FK
667 (https://github.com/thegenemyers/MERQURY.FK). Gene completeness was evaluated with
668 BUSCO 5.5.0 (Manni et al. 2021) as well as compleasm 0.2.5 (Huang and Li 2023). We used
669 the eutherian_odb10 dataset for *Bradypus torquatus*, and actinopterygii_odb10 for *C. nuchalis*
670 and *C. wayuu*, the mammalia_odb10 dataset for human, the arthropoda_odb10 dataset for
671 *Podura aquatica,* and the metazoa_odb10 dataset for *Elysia timida* and *Scintilla philippinensis*.
672
673 For *B. torquatus*, we initially obtained hifiasm (v0.19.5) assemblies that were of a size
674 expected from four haplotypes of this genome, consisting of a large number of small contigs
675 (Supplementary Table 6). Similar results were obtained with HiCanu (v2.2) (Nurk et al. 2020),
676 which is designed to break contigs at all joins in the assembly graph, meaning any divergences
677 between the four theoretical haplotypes would result in a new contig (in our case over 200,000
678 assembled contigs totaling almost 12 Gb of sequence). This indicated that the tissue samples
679 we obtained for this species consist of DNA sequenced from two individuals. While the
680 accuracy of the PacBio HiFi reads should in principle allow to distinguish all four haplotypes,
681 *B. torquatus* is expected to have a very low heterozygosity and high in-breeding rate due to
682 small population size, which results in assembly graphs where many regions collapse all
683 haplotypes due to the lack of sequence variation.
684
685 To overcome this problem, we used the assembler Flye (v2.9.2) (Kolmogorov et al. 2019),
686 which allows users to set the read error rate as an argument. Flye has been previously
687 suggested by the developers as a method for collapsing sequences from highly diverged
688 haplotypes into a single "pseudo-haplotype" sequence
689 (https://github.com/fenderglass/Flye/issues/636). Here, we found that a read error-rate of 3%
690 produced the most contiguous assembly, when combined with a reduced read-overlap of 5 kb
691 (Supplementary Table 6). The latter deviates from the default value selected by Flye, which
692 Flye would determine by the N90 of the input reads (in our case the N90 was 9 kb for the HiFi
693 library, which had a modal read length of ~10 kb). We then removed retained haplotigs using
694 purge-dups (Guan et al. 2020).
695

696 Contamination detection

697 Specimens stored in liquid preservation media are prone to various levels of DNA
698 contamination from non-target organisms (Raxworthy and Smith 2021), caused by different
699 handling and storage conditions that are often hard to retrace (Ruiz-Gartzia et al. 2022). To
700 detect levels of contamination from exogenous DNA in our assemblies, we used NCBI's
701 Foreign Contamination Screen (FCS 0.5.0) (Astashyn et al. 2023), which flags both putative
702 adapter sequences (FCS-adaptor) and contigs assigned to non-target species (FCS-GX).
703 Both FCS tools were executed from the provided singularity container using singularity 1.2.4.

FCS-adaptor was executed through the provided bash script (run_fcsadaptor.sh) with the option for eukaryotes (--euk). FCS-GX was executed by the python wrapper (fcs.py screen genome) together with the corresponding NCBI taxonomy ID and the GX database (as of Dec 5th, 2023).

Additionally, to assess pre-assembly read quality, we mapped reads obtained from samples of *Bradypus torquatus*, *Anguis fragilis* and *Idiurus macrotis* to available reference genomes of closely related species *Choloepus didactylus* (GCA_015220235.1), *Elgaria multicarinata* (GCA_023053635.1) and *Pedetes capensis* (GCA_007922755.1). Similarly, to identify regions of PCR coverage dropouts, we aligned reads from polymerase A/B or C libraries to the best (defined as highest contig N50, Supplementary Table 5) assemblies obtained for *Podura*, *Scintilla* and *Elysia*, and visually inspected mapped reads (Supplementary Figures S2, S3). All read mapping was performed using minimap2 v2.26 (Li 2021) with HiFi read mapping parameters (--ax map-hifi).

Scaffolding the final *B. torquatus* genome

To scaffold these *B. torquatus* contigs, we mapped HiC reads to the contig assembly using bwa-mem (v.0.7.17) (Li 2013), before the resulting HiC alignment file was filtered, sorted and deduplicated with pairtools parse, pairtools sort and pairtools dedup (v0.3.0), respectively. The processed HiC alignments were then used as input for scaffolder yahs (v1.2a.1.patch) (Zhou et al. 2023). A full list of commands is given in Supplementary Note 1. After initial automated scaffolding with yahs, we ran multiple rounds of manual curation based on the HiC interaction maps. This involved re-ordering and re-orienting the scaffolded sequences based on sequences close to each other in the genome, which are expected to have a higher number of HiC interactions than those further apart. Using this method, we were able to obtain chromosome-level scaffolds of the 24 autosomes and the X chromosome. This assembly was then again screened for adapter and foreign sequence contaminants using NCBI's FCS-adaptor and FCS-GX tools (Astashyn et al. 2023). We subsequently removed contaminant sequences by applying the python wrapper (fcs.py clean genome) together with the action report from "screen genome" and setting the minimum sequence length to 1 bp (--min-seq-len 1).

Competing interests

The authors have no competing interests.

Acknowledgment

We thank Deniz Kaya from PacBio for advice and suggestions on adapting the ultra-low input protocol, and Sarah Kingan, Juniper Lake, Ian McLaughlin and Jonas Korlach from PacBio for the polymerase C runs for the human sample. We also thank the Genome Technology Center (RGTC) at Radboudumc for the use of the Sequencing Core Facility (Nijmegen, The Netherlands), which provided the PacBio SMRT sequencing service on the Sequel IIe

platform, the Long Read Team of the DRESDEN Concept Genome Center, part of the MPI-CBG and the technology platform of the CMCB at the TU Dresden, supported by DFG (INST 269/768-1), and the HPC Service of FUB-IT, Freie Universität Berlin, for computing time (doi:10.17169/refubium-26754). We acknowledge Irina Ruf (Senckenberg Frankfurt), Carles Galià Camps (University of Barcelona), Madlen Stange, Claudia Koch, Morris Flecks, Jan Decher and Christian Montermann (Leibniz Institute for the Analysis of Biodiversity Change) for providing samples and Sandra Kukowka for helping in subsampling specimens.
This work was supported by a grant from the Leibniz Association's Competition Procedure (K419/2021) and the LOEWE-Centre for Translational Biodiversity Genomics (TBG) funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK).

Data and Code Availability

Genome assemblies are available at https://genome.senckenberg.de/download/GenomesCollectionsPolC. The raw sequencing data and assemblies for *Bradypus torquatus* will be available under BioProject PRJEB73341 and BioSample SAMEA115348596. The TOGA annotation for the *B. torquatus* is available at https://genome.senckenberg.de/download/TOGA/. No new computer code was generated in this study.

References

Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, Mozes E, Strope PK, Sylla PM, Wagner L, et al. 2023. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *bioRxiv*. http://dx.doi.org/10.1101/2023.06.02.543519.

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, et al. 2023. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol* **41**: 232–238.

Blom MPK. 2021. Opportunities and challenges for high-quality biodiversity tissue archives in the age of long-read sequencing. *Mol Ecol* **30**: 5935–5948.

Blumer M, Brown T, Freitas MB, Destro AL, Oliveira JA, Morales AE, Schell T, Greve C, Pippel M, Jebb D, et al. 2022. Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding. *Sci Adv* **8**: eabm6494.

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

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. 2022. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* **40**: 1332–1335.

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al. 2023. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet* **55**: 1221–1231.

Dahn HA, Mountcastle J, Balacco J, Winkler S, Bista I, Schmitt AD, Pettersson OV, Formenti G, Oliver K, Smith M, et al. 2022. Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *Gigascience* **11**. http://dx.doi.org/10.1093/gigascience/giac068.

788 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
789 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**.
790 http://dx.doi.org/10.1093/gigascience/giab008.

791 Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, Maunsell SC, Aduse-
792 Poku K, Talavera G, Eastwood R, et al. 2018. A Comprehensive and Dated Phylogenomic Analysis
793 of Butterflies. *Curr Biol* **28**: 770–778.e5.

794 Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC,
795 et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature*
796 **587**: 252–257.

797 Gibb GC, Condamine FL, Kuch M, Enk J, Moraes-Barros N, Superina M, Poinar HN, Delsuc F.
798 2016. Shotgun Mitogenomics Provides a Reference Phylogenetic Framework and Timescale for
799 Living Xenarthrans. *Mol Biol Evol* **33**: 621–642.

800 Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing
801 haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**: 2896–2898.

802 Heinicke MP, Nielsen SV, Bauer AM, Kelly R, Geneva AJ, Daza JD, Keating SE, Gamble T. 2023.
803 Reappraising the evolutionary history of the largest known gecko, the presumably extinct
804 Hoplodactylus delcourti, via high-throughput sequencing of archival DNA. *Sci Rep* **13**: 1–12.

805 Hiller M, Morales A, Ahmed A, Hilgers L, Kirilenko B, Kontopoulos D, Dong Y, Li X, Irving A, Brown
806 T, et al. 2023. Reference-quality bat genomes illuminate adaptations to viral tolerance and disease
807 resistance. *Research Square*. https://www.researchsquare.com/article/rs-2557682/latest
808 (Accessed September 18, 2023).

809 Huang N, Li H. 2023. compleasm: a faster and more accurate reimplementation of BUSCO.
810 *Bioinformatics* **39**. http://dx.doi.org/10.1093/bioinformatics/btad595.

811 Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermiin LS,
812 Skirmuntt EC, Katzourakis A, et al. 2020. Six reference-quality genomes reveal evolution of bat
813 adaptations. *Nature* **583**: 578–584.

814 Johnson KR, Owens IFP, Global Collection Group. 2023. A global approach for natural history
815 museum collections. *Science* **379**: 1192–1194.

816 Kautt AF, Kratochwil CF, Nater A, Machado-Schiaffino G, Olave M, Henning F, Torres-Dowdall J,
817 Härer A, Hulsey CD, Franchini P, et al. 2020. Contrasting signatures of genomic divergence during
818 sympatric speciation. *Nature* **588**: 106–111.

819 Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, Morales AE, Ahmed A-W,
820 Kontopoulos D-G, Hilgers L, et al. 2023. Integrating gene annotation with orthology inference at
821 scale. *Science* **380**: eabn3107.

822 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat
823 graphs. *Nat Biotechnol* **37**: 540–546.

824 Kuderna LFK, Gao H, Janiak MC, Kuhlwilm M, Orkin JD, Bataillon T, Manu S, Valenzuela A,
825 Bergman J, Rousselle M, et al. 2023. A global catalog of whole-genome diversity from 233 primate
826 species. *Science* **380**: 906–913.

827 Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards
828 SV, Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: Sequencing life for the future of
829 life. *Proc Natl Acad Sci U S A* **115**: 4325–4333.

830 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

arXiv [q-bioGN]. http://arxiv.org/abs/1303.3997.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572–4574.

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.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142–i150.

Mulcahy DG, Macdonald KS 3rd, Brady SG, Meyer C, Barker KB, Coddington J. 2016. Greater than X kb: a quantitative assessment of preservation conditions on genomic DNA quality, and a proposed standard for genome-quality DNA. *PeerJ* **4**: e2528.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291–1305.

Osipova E, Barsacchi R, Brown T, Sadanandan K, Gaede AH, Monte A, Jarrells J, Moebius C, Pippel M, Altshuler DL, et al. 2023. Loss of a gluconeogenic muscle enzyme contributed to adaptive metabolic traits in hummingbirds. *Science* **379**: 185–190.

PacBio Biosciences. 2020. Now available: Ultra-low DNA input workflow for SMRT sequencing. https://www.pacb.com/blog/introducing-the-ultra-low-input-protocol-for-smrt-sequencing/ (Accessed 2023).

Raxworthy CJ, Smith BT. 2021. Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol Evol* **36**: 1049–1060.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**: 737–746.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**: 245.

Ronco F, Matschiner M, Böhne A, Boila A, Büscher HH, El Taher A, Indermaur A, Malinsky M, Ricci V, Kahmen A, et al. 2021. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature* **589**: 76–81.

Ruiz-Gartzia I, Lizano E, Marques-Bonet T, Kelley JL. 2022. Recovering the genomes hidden in museum wet collections. *Mol Ecol Resour* **22**: 2127–2129.

Sambrook J, Russell WD. 2001. Protocol: DNA isolation from mammalian tissue. *Molecular Cloning: A Laboratory Manual (Sambrook J*.

Schneider C, Woehle C, Greve C, D'Haese CA, Wolf M, Hiller M, Janke A, Bálint M, Huettel B. 2021. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *Gigascience* **10**. http://dx.doi.org/10.1093/gigascience/giab035.

Shao Y, Zhou L, Li F, Zhao L, Zhang B-L, Shao F, Chen J-W, Chen C-Y, Bi X, Zhuang X-L, et al. 2023. Phylogenomic analyses provide insights into primate evolution. *Science* **380**: 913–924.

Sharma P, Al-Dossary O, Alsubaie B, Al-Mssallem I, Nath O, Mitter N, Rodrigues Alves Margarido

874 G, Topp B, Murigneux V, Kharabian Masouleh A, et al. 2021. Improvements in the sequencing and
875 assembly of plant genomes. *GigaByte* **2021**: gigabyte24.

876 Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 2022. HiFiAdapterFilt, a memory efficient read
877 processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their
878 negative impacts on genome assembly. *BMC Genomics* **23**: 157.

879 Tan HZ, Jansen JJFJ, Allport GA, Garg KM, Chattopadhyay B, Irestedt M, Pang SEH, Chilton G,
880 Gwee CY, Rheindt FE. 2023. Megafaunal extinctions, not climate change, may explain Holocene
881 genetic diversity declines in Numenius shorebirds. *Elife* **12**. http://dx.doi.org/10.7554/eLife.85422.

882 Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A,
883 Munson KM, Lewis AP, et al. 2022. Segmental duplications and their variation in a complete human
884 genome. *Science* **376**: eabj6965.

885 Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM,
886 Raveendran M, Hillier LW, et al. 2020. Sequence diversity analyses of an improved rhesus
887 macaque genome enhance its biomedical utility. *Science* **370**.
888 http://dx.doi.org/10.1126/science.abc6617.

889 Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan
890 A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing
891 improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.

892 Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol*
893 **20**: 257.

894 Zhou C, McCarthy SA, Durbin R. 2023. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**.
895 http://dx.doi.org/10.1093/bioinformatics/btac808.

896 Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and
897 conservation. *Nature* **587**: 240–245.