


METHOD

Open Access



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}, Clément Schneider⁷, Evgeny Leushkin^{1,2}, Zeyuan Chen², Julia Sigwart^{1,2}, Vanessa Gonzalez⁸, Nur Leena W. S. Wong⁹, Fabricio R. Santos¹⁰, Mozes P. K. Blom¹¹, Frieder Mayer¹¹, Camila J. Mazzoni^{5,6}, Astrid Böhne⁴, Sylke Winkler^{12,13}, Carola Greve^{1,2} and Michael Hiller^{1,2,3*} 

[†]Bernhard Bein, Ioannis Chrysostomakis, Larissa S. Arantes and Tom Brown are shared first authors.

*Correspondence: michael.hiller@senckenberg.de

¹ LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, Frankfurt 60325, Germany
Full list of author information is available at the end of the article

Abstract

Museum collections harbor millions of samples, largely unutilized for long-read sequencing. Here, we use ethanol-preserved samples containing kilobase-sized DNA to show that amplification-free protocols can yield contiguous genome assemblies. Additionally, using a modified amplification-based protocol, employing an alternative polymerase to overcome PCR bias, we assemble the 3.1 Gb maned sloth genome, surpassing the previous 500 Mb protocol size limit. Our protocol also improves assemblies of other difficult-to-sequence molluscs and arthropods, including millimeter-sized organisms. By highlighting collections as valuable sample resources and facilitating genome assembly of tiny and challenging organisms, our study advances efforts to obtain reference genomes of all eukaryotes.

Keywords: Long-read sequencing, PCR amplification, Genome assembly, Museum collections

Background

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 [1]. Advances in short-read sequencing technology (with Illumina as the most prominent platform) enabled sequencing the genomes of a few thousand eukaryotes to date [2–5]. 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 [2, 3, 5], posing many



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

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 [6, 7], and consequently enable complete genome annotations and comprehensive analyses [8–16].

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 [17].

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 [18]. 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 [3]. 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 [19–21].

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 [22, 23], 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% [24]. 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 [7, 16, 24, 25], 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 often 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 *Dipus sagitta*, pen-tailed treeshrew *Ptilocercus lowii*, long-eared flying mouse *Idiurus macrotis*, maned sloth *Bradypus torquatus*), two squamates (European blind snake *Xerotyphlops vermicularis*, slow worm *Anguis fragilis*), and two fishes (the catfish species *Cathorops nuchalis* and *Cathorops wayuu*), all lacking a genome assembly (Table 1, Additional File 2: Table S1). All samples were collected in the field and preserved in technical or 96% ethanol. Apart from the maned sloth and the catfishes, all samples were kept at room temperature. The samples of the maned sloth and catfish were kept most of the time in a freezer at – 20 °C; however, in contrast to flash-frozen

Table 1 Overview of the species and samples

Species	Year sampled	Preservation	Type of sample
Northern three-toed jerboa (<i>Dipus sagitta</i>)	2006 and 1961	Technical ethanol, room temperature	Muscle, skin
Pen-tailed treeshrew (<i>Ptilocercus lowii</i>)	1967	Technical ethanol, room temperature	Muscle
Long-eared flying mouse (<i>Idiurus macrotis</i>)	2000	Technical ethanol, room temperature	Skin with hair
Maned sloth (<i>Bradypus torquatus</i>)	2003	Likely pure ethanol, mostly at – 20 °C (otherwise room temperature)	Clogged blood
European blind snake (<i>Xerotyphlops vermicularis</i>)	2004 and 2011	Technical ethanol, room temperature	Skin and muscle
slow worm (<i>Anguis fragilis</i>)	2021	Technical ethanol, room temperature	Muscle from tail cross-section
Catfish (<i>Cathorops nuchalis</i>)	2014	Pure ethanol, transported multiple times at room temperature until final storage at – 20 °C	Fin
Catfish (<i>Cathorops wayuu</i>)	2014		Fin

samples, freezing did not occur immediately after sampling and they were kept at room temperature for extended periods of time, including during transportation.

We used a modified Circulomics Nanobind disk and a phenol/chloroform-based protocol for the extraction of genomic DNA (Methods). For *Dipus sagitta*, *Ptilocercus lowii*, and *Xeratyphlops vermicularis*, we did not obtain a sufficient amount of DNA (<400 ng) and/or DNA fragments were shorter than 0.18 kb (Additional File 2: Table S1), showing that DNA is too degraded to proceed with library preparation. For four species (*Anguis fragilis*, *Idiurus macrotis*, *Cathorops nuchalis*, and *Cathorops wayuu*), the amount of DNA and the DNA fragment sizes were sufficient to prepare an amplification-free PacBio low input library [26] (Additional File 2: Table S1). We sequenced all libraries on a PacBio Sequel IIe system, disabling on-board calling of HiFi reads and instead applying the computationally expensive DeepConsensus method [27] to maximize HiFi read yield and length. For *Bradypus torquatus*, we did not obtain enough DNA and therefore proceeded with a PacBio ultra-low input library (see below).

For the two catfish species, *Cathorops nuchalis* and *Cathorops wayuu*, we sequenced two SMRT cells each and obtained HiFi reads with an average length of 8832 and 8783 bp, respectively, providing a total of 43.8 and 41.2 Gb, which corresponds to coverages of ~17X and ~16.5X (Additional File 2: Table S1). Using HiFiasm with different parameters [28], we obtained a contig assembly for both species with a total length of 2.6 and 2.59 Gb and a contig N50 value of 3.2 and 2.1 Mb (Additional File 2: Table S2). To assess gene completeness, we used compleasm [29] with the set of 3640 ray-finned fish (Actinopterygii) near-universally conserved genes (ODB10) [30], which showed that 96.65% of these genes are fully present in the primary assembly of *C. nuchalis* and 95.6% in that of *C. wayuu*. Although additional HiFi data would be needed to improve contiguity and HiC data would be required to scaffold the contigs into chromosome-level scaffolds, our catfish samples exemplify that an adequate genome assembly can be obtained from 10-year-old, ethanol-preserved tissues.

In contrast to the catfish, we obtained very low sequencing yields for *Idiurus macrotis* and *Anguis fragilis*, with only 0.3 Gb and 0.04 Gb of HiFi data (Additional File 2: Table S1). Quality metrics showed that the polymerase N50 raw read lengths were very short and the local base rates were low. For example, while the library from *Anguis* met the requirements for PacBio sequencing with a mean fragment length of 12.2 kb, both the local base rate of 1.64 (expected ~2.8) and the polymerase N50 raw read length of 32.3 kb (expected at least 200 kb) are very low and insufficient to produce HiFi reads of most DNA fragments in the library. This indicates that factors such as DNA damage, metabolites bound to the DNA, or contaminants precipitated with the DNA inhibit the polymerase, highlighting sequencing challenges for ethanol-preserved samples stored at room temperature.

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

We reasoned that a PCR-based amplification step prior to library preparation could render the *Idiurus macrotis* and *Anguis fragilis* samples amenable to sequencing, as this procedure should yield intact DNA devoid of potential polymerase-inhibiting metabolites. To this end, we applied the PacBio ultra-low input library protocol [31] to the samples of *Idiurus macrotis* and *Anguis fragilis*. Although this protocol was originally

designed for small specimens providing very limited DNA amounts [32] and is recommended only for genome sizes of up to 500 Mb, the protocol includes a PCR amplification step using two different undisclosed polymerases targeting DNA with average and high GC contents, respectively. For simplicity, we refer to these polymerases as “A” and “B” in the following to distinguish them from a third polymerase “C” that we also investigate as described below. We also generated an ultra-low input library for the *Bradypus torquatus* sample that did not contain enough DNA for the low input protocol.

Indeed, for *Idiurus macrotis* and *Anguis fragilis*, sequencing another SMRT cell each produced 10 and 19.6 Gb in HiFi reads with an average HiFi read length of 4854 bp and 7552 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 (Additional File 2: Table S1).

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 4270 vs. 4854 bp for *Idiurus macrotis* and 5609 vs. 7552 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 (Additional File 2: Table S1).

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 (~200 kb mostly assigned to plants), the *Anguis fragilis* data had higher levels of contamination (~200 Mb assigned to various bacterial groups), and the vast majority of the sequencing data obtained from the *Idiurus macrotis* sample were contamination (~75 Mb assigned to various groups of bacteria) (Additional file 1: Figs. S1, S2). High levels of contamination (71–90% of sequenced reads) 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 an *Anguis fragilis* sample from 1878 (Additional File 2: Tables S1, S3). Together, while sample contamination with bacteria, protists and bacterial viruses or cross-contamination with human DNA is another challenge related to samples obtained from collections [33–35], 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 (Fig. 1, brown dashed line), which is unexpectedly

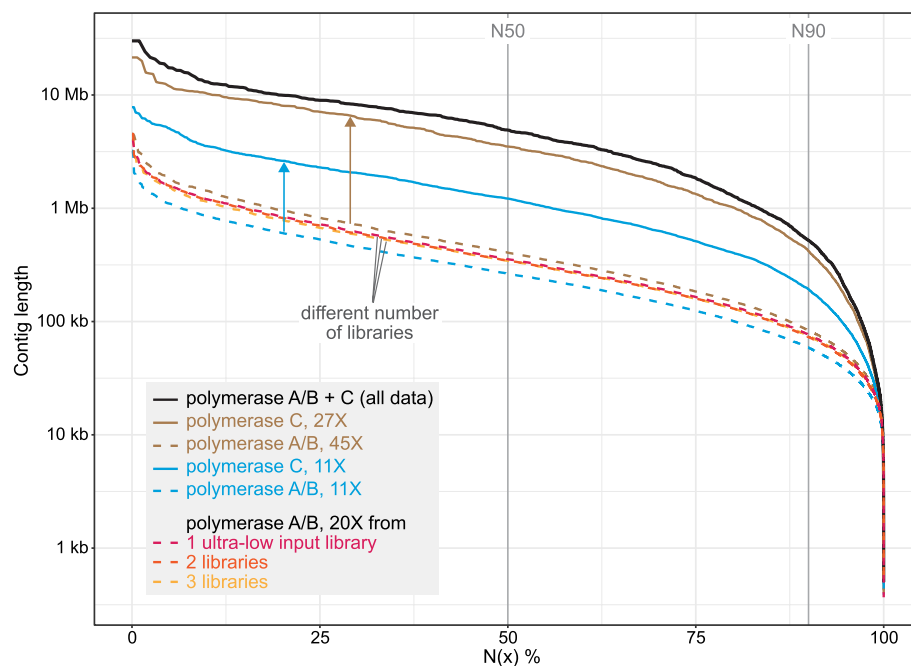


Fig. 1 Contiguity of *B. torquatus* assemblies generated with data from ultra-low input libraries prepared with polymerases A/B and/or C at different coverages. Assembly contiguity is visualized as N(x) graphs that show contig sizes on the Y-axis, for which x percent of the assembly consists of contigs of at least that size. The N50 and N90 values are shown as vertical gray lines and indicate contig sizes for which 50% and 90% of the assembly consists of contigs of at least that size, respectively. 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 discussed in the text and summarized in the inset

low as similar read coverages typically yield mammalian assemblies with contig N50 values exceeding several megabases. Using compleasm [29] 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 [36] 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.

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* [11]. Despite both species being separated for 30 My [37], 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 (Fig. 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.

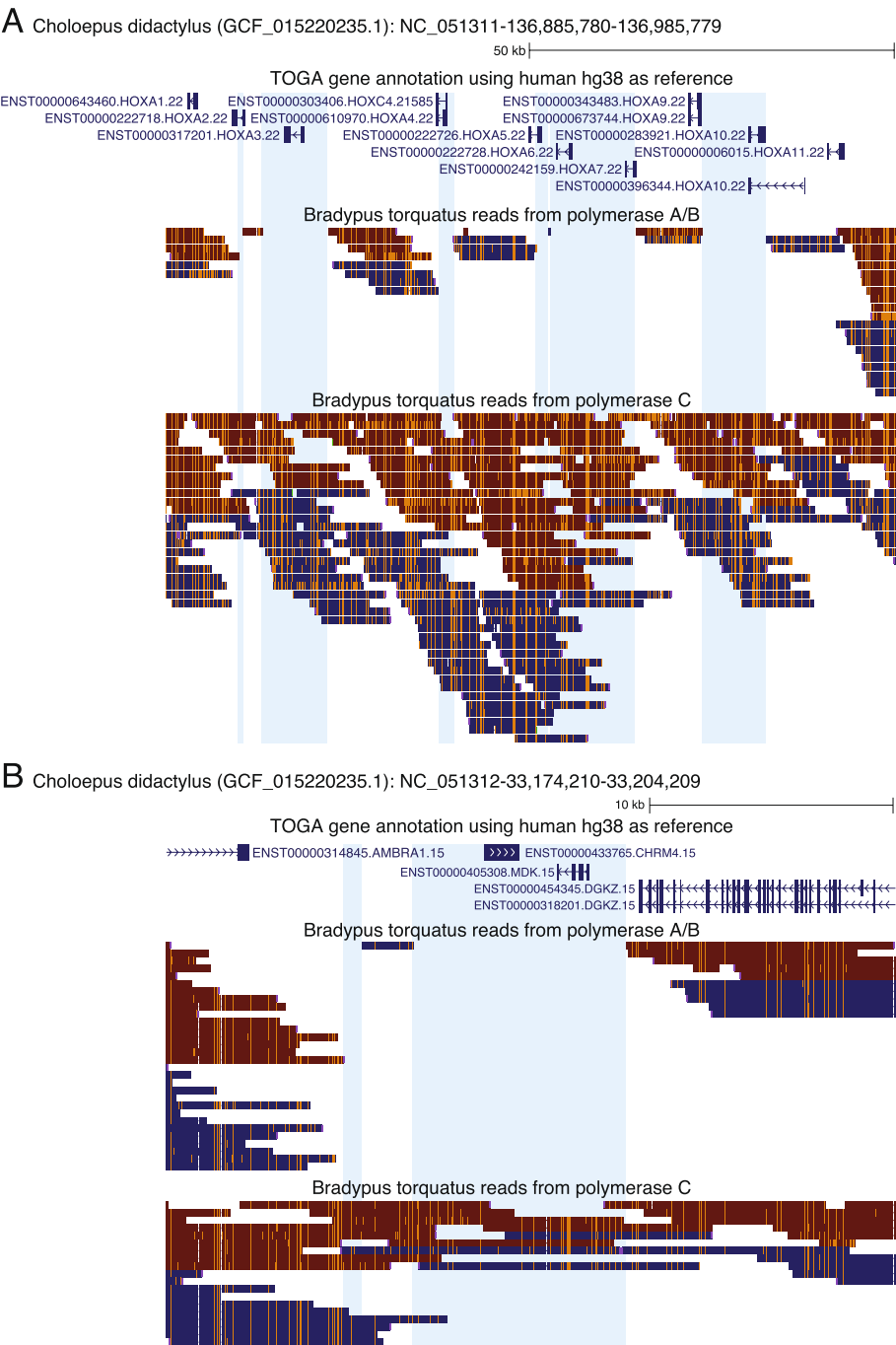


Fig. 2 PCR bias in reads produced with polymerase A/B. UCSC genome browser screenshots of the *Choloepus didactylus* assembly, together with the TOGA gene annotation and mapped HiFi reads of *B. torquatus* produced either with polymerase A/B or polymerase C. The TOGA gene annotation is shown in blue with boxes representing coding exons, connecting horizontal lines representing introns, and arrowheads indicating the direction of transcription (+ or —strand). Mapped HiFi reads are shown below as boxes with orange tickmarks representing insertions in the *B. torquatus* reads relative to the *C. didactylus* assembly. Reads in blue and red align to the + and —strand, respectively. **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 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 (Merck). 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 27X 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 (Fig. 1, black line, Additional File 2: Table S4), which is 12 times higher than the previous assembly generated from reads obtained with polymerase A/B. 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 (Fig. 2). Consistent with a higher PCR bias for polymerase A/B, we found that the normalized read coverage in exonic and repeat regions is biased towards a lower coverage for the polymerase A/B data compared to polymerase C data (Additional file 1: Fig. S3). 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 coverage from the individual SMRT cells out of the equation, we downsampled our data to equal coverage and performed a number of tests (Additional File 2: Table S4). 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 ~11X, 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) (Fig. 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) (Fig. 1, brown lines), despite the polymerase C data having a substantially lower coverage (27X 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 (Fig. 1, red/orange/yellow lines; Additional File 2: Table S4), 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, which is applicable to ethanol-preserved samples [23, 38], to generate 97.5 Gb in long-range read pair data. Using the automated scaffolding software yahs [39] and manual curation, our contig assembly could be scaffolded into chromosome-level scaffolds (Fig. 3A). This final assembly consists of 2915 scaffolds and 5022 contigs. The scaffold N50 and N90 values are 157 Mb and 61.3 Mb, respectively (Fig. 3B). The contig N50 and N90 values are 4.75 Mb and 519 kb, respectively. Using Merquary [40] 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 (Additional File 2: Table S4) and contains 90.72% of ancestral placental mammal genes.

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 (Fig. 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.

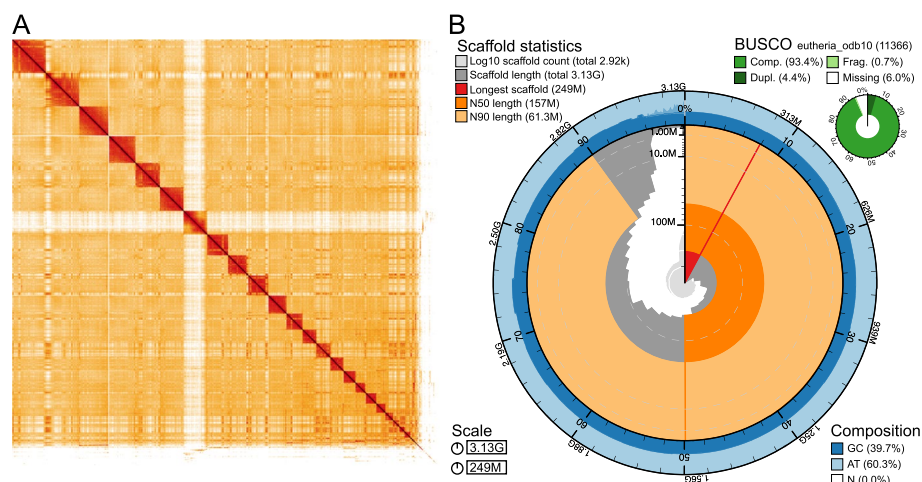


Fig. 3 Chromosome-scale assembly of *Bradypus torquatus*. **A** HiC interaction map after automated scaffolding by yahs and manual curation. The HiC map shows interactions in 3-dimensional space between two regions of the genome. Darker colors indicate a higher number of interactions. The region of low 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 lengths (light orange). The outer ring shows the GC content of the genome

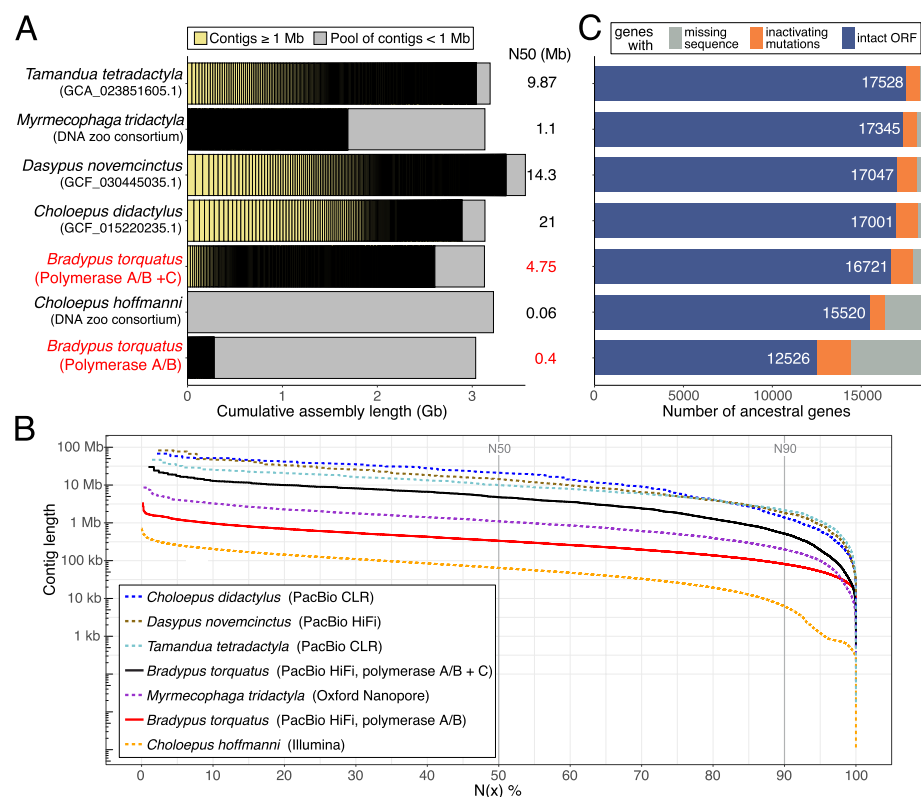


Fig. 4 Comparison of xenarthran genome assemblies. **A** Visualization of contig sizes of available xenarthran genome assemblies. Each bar represents the total assembly size. Contigs shorter than 1 Mb are not visualized individually but shown as the gray portion of each bar. The final *B. torquatus* 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**. **B** Visualization of assembly contiguity as an N(x) graph, showing contig sizes on the Y-axis, for which x percent of the assembly consists of contigs of at least that size. Assembly order in the legend (inset) is sorted by contig N50 value. **C** TOGA classification of 18,430 ancestral placental mammal genes showing the number of genes that have an intact reading frame (blue bar, number is given in white font), inactivating mutations (e.g., frameshifts, stop codon, splice site mutations or exon deletions; orange bar), or missing coding sequence parts often caused by assembly gaps or fragmentation (gray bar). Assemblies are sorted by the number of intact genes

Polymerase C improves assemblies for various species

We next explored whether polymerase C can also help to improve assemblies of other species, using samples not obtained from collections. To provide a fair comparison, we randomly downsampled the larger data set to obtain an equal coverage of HiFi reads generated with polymerases A/B and C. To compare these polymerases for another mammal, we used the human HG002 sample and generated assemblies for both human haplotypes. Using an equal coverage of 23.5X, the polymerase A/B read data produced a 2.96 Gb assembly for haplotype 1 with a contig N50 value of 642 kb, whereas the polymerase C data generated a 3.03 Gb assembly with a substantially higher contig N50 value of 2.8 Mb, a 4.4-fold increase in contiguity. Consistently, gene completeness assessed with compleasm (mammalia_odb10) increased substantially from 81.2 to 98.6%. Similar results were obtained for the haplotype 2 assembly, where the polymerase A/B read data produced a 2.9 Gb assembly with a contig N50 value of 558.8 kb and a gene completeness of 77.3%, whereas the polymerase C read data

produced a 3 Gb assembly with a contig N50 value of 2 Mb and a gene completeness of 97.8%.

Since PCR amplification may produce chimeric reads [41], we used available non-amplified human HiFi reads produced from the HG002 sample as a baseline to compare the amount of chimeric HiFi reads generated by polymerase A/B and C. We mapped reads to the HG002 assembly [6] and computed the number of reads with supplementary alignments, which indicate chimeras. We found that the fraction of chimeric alignments is very low ($\leq 0.81\%$) across all three libraries, with polymerase C reads having the lowest fraction (Additional file 1: Fig. S4A). We next included available HG002 read data obtained by Multiple Displacement Amplification (MDA) in this comparison. Consistent with previous observations [41, 42], the majority of MDA alignments (69.3%) are chimeras, which is further supported by the observation that the primary alignment lengths are much shorter than the MDA reads (Additional file 1: Fig. S4E). We therefore conclude that long range PCR amplification used in the original and modified ultra-low input protocol does not create more chimeric reads than non-amplified libraries and orders of magnitude fewer chimeric reads than MDA libraries.

Next, we explored the application of polymerase C to three non-vertebrate taxa covering two additional phyla, Mollusca (two taxonomic classes: Gastropoda and Bivalvia) and Arthropoda (Collembola), using taxa where genome sequencing efforts often rely on the amplification-based protocols because of low sequencing performance with low input protocol or very small DNA amounts.

For the sacoglossan gastropod *Elysia timida* (Mollusca), previous sequencing libraries created with the low input protocol resulted in very poor sequencing performance. Therefore, we applied the ultra-low input protocols and compared two SMRT cells produced with polymerase 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 SMRT cell produced with polymerase C, providing 23 Gb yield in reads with an N50 length of 7 kb (Additional File 2: Table S5). After subsampling to equal read coverage of 26.4X, polymerase A/B and C read data generated assemblies with similar contig N50 values of 347.1 kb for polymerase A/B and 331 kb for polymerase C (Fig. 5, Additional File 2: Table S5). Using all polymerase A/B read data with a coverage of 42.5X increased the contig N50 value to 472.6 kb. Importantly, adding the 23 Gb of polymerase C reads, increased the contig N50 value 1.4-fold to 675.8 kb (Fig. 5). While the gene completeness (metazoa_odb10) of 97.7 and 97.8% is similar between these assemblies, polymerase C data helped to improve assembly contiguity for this mollusc.

To understand why polymerase C alone does not result in a more contiguous assembly, we mapped both polymerase A/B and C reads to the *Elysia timida* assembly with the highest contiguity, generated from all read data. This showed that both polymerases A/B and C exhibit bias; however, bias of one polymerase can be compensated by reads of the other (Additional file 1: Fig. S5), indicating that these polymerases may have taxon-specific differences.

For the marine bivalve *Scintilla philippinensis* with an estimated genome size of 1.3 Gb, we compared assemblies produced from 22.3 Gb of reads obtained from ultra-low input libraries using polymerase A/B or C, which corresponds to a coverage of 17.1X. While the polymerase A/B reads produced a 1.77 Gb assembly with a contig N50 value of 43.3 kb, the polymerase C read data produced a 1.88 Gb assembly with a 1.2-fold

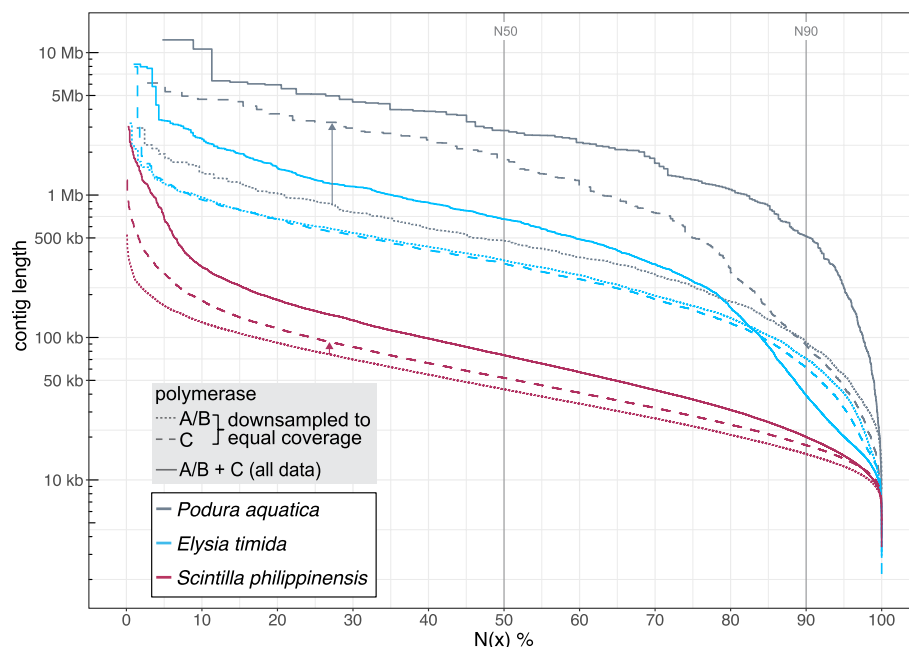


Fig. 5 Impact of polymerase C on assemblies of mollusc and collembola species. Assembly contiguity visualized as N(x) graphs that show contig sizes on the Y-axis, for which x percent of the assembly consists of contigs of at least that size (N50 and N90 values are indicated). Assemblies are generated 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

increased contig N50 value of 52.1 kb. Gene completeness (metazoa_odb10) improved slightly from 89.1% (polymerase A/B) to 89.9% (polymerase C). Combining all polymerase A/B and C read data (coverage of 36.1X) produced a 1.86 Gb assembly with an even higher contig N50 value of 75.1 kb (Fig. 5, Additional File 2: Table S5) and a higher gene completeness of 93.5%. Mapping polymerase A/B and C reads to the assembly generated with all data also revealed regions that were covered only by reads from one polymerase (Additional file 1: Fig. S6A,B). While polymerase C reads improved assembly contiguity of both mollusc species, the resulting assemblies have a comparatively low contiguity, highlighting the challenges of sequencing molluscan DNA.

We next tested our adjusted protocol on a species having a very small body size, where amplification of the limited amount of genomic DNA is required for long-read sequencing and genome assembly [32]. We used an ethanol-preserved, whole single specimen of the springtail *Podura aquatica* (Arthropoda: Collembola), which has a body size of only 1.5 mm and an expected genome size of 200–300 Mb. The polymerase A/B run yielded 13.9 Gb of HiFi reads with an N50 length of 9.6 kb. The polymerase C run yielded 21.7 Gb of reads but with a lower N50 read length of 5.7 kb, which is likely explained by sequencing DNA 1 year after the initial extraction (the entire specimen was used for the initial DNA extraction). Strikingly, at an estimated coverage of ~50X, the polymerase A/B read data produced a 278.5 Mb assembly with a contig N50 value of only 919 kb, whereas the polymerase C data generated a 269.3 Mb assembly with a contig N50 value of 2.77 Mb (Fig. 5, Additional File 2: Table S5). This represents a three-fold increase in contiguity, despite the polymerase C reads being substantially shorter.

Gene completeness (arthropoda_odb10) increased slightly from 92.8% for polymerase A/B assembly to 93.4% for polymerase C assembly. Combining all polymerase A/B and C read data resulted in a 284.7 Mb assembly with an even higher contig N50 value of 5.74 Mb and the same gene completeness of 93.4%. Similar to *Elysia* and *Scintilla*, aligning reads to the most contiguous assembly showed complementary coverage dropouts (Additional file 1: Fig. S6C,D).

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 [22, 23]. 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 remains to be investigated for which taxonomic groups amplification-free protocols are generally successful. We demonstrate that PCR bias associated with the amplification-based PacBio ultra-low input protocol can be overcome or at least mitigated by employing an alternative polymerase. As a proof of concept, the contiguous 3.1 Gb genome assembly of *B. torquatus* shows that a modified amplification-based protocol can produce high-quality assemblies of gigabase-sized genomes.

Contamination caused by sample decomposition, human handlers, or commensal bacteria is expected for collection samples that have been stored under non-sterile conditions [33]. It is difficult to assess contamination prior to sequencing, and we find different levels of contamination in our samples, ranging from most of the sequenced reads stemming from contaminants to almost no contamination. Analyzing a low coverage of sequencing reads for contamination before sequencing a sample to the coverage required for assembly could therefore be a cost-efficient strategy to select those samples that contain sufficiently low contamination levels. Furthermore, the resulting assemblies should be carefully screened for contamination using existing methods [43, 44].

Consistent with previous observations [34], we find that sample age alone is not an accurate predictor of input DNA quality and sample suitability for sequencing. For example, while the *B. torquatus* sample was collected in 2003, several younger samples exhibited high degrees of DNA degradation (Additional File 2: Table S1). Hence, in addition to sample age, other factors such as storage temperature and conditions, storage medium, or tissue type likely influence DNA quality. From our experience, samples consistently stored at -20°C and preserved in 96% ethanol perform well, but a systematic assessment of larger sample numbers is needed to substantiate this.

Our study has a number of implications. First, the modified ultra-low input protocol improves genome assembly of small specimens, where amplification is a requirement to

obtain enough DNA for sequencing. For example, the contiguity of the *Podura aquatica* genome increased to an N50 of 5.7 Mb and thus substantially exceeds the minimum standards of 100 kb set by the Earth Biogenome Project for small species with limited DNA amounts [1]. The modified protocol will likely not only be beneficial for species with diminutive body sizes that represent a very large but mostly uncharacterized part of Earth's biodiversity but also in cases where only very limited amounts of material from non-lethal samplings (biopsies from human patients or bat wing punches) are available. Second, long-read sequencing remains a challenge for molluscs and other taxonomic groups, where satisfactory sequencing outputs often require amplification-based protocols. Although achieving highly contiguous assemblies with megabase contig N50 values remains challenging for these species, our investigations suggest that employing a combination of different polymerases can at least help to improve assembly contiguity. Third, while the PacBio ultra-low input protocol was previously limited to genome sizes of up to 500 Mb, the successful application of the modified protocol to *B. torquatus* with its 3.1 Gb genome extends its applicability to a broad range of species with larger genome sizes. Together, the improved efficiency of the modified ultra-low input protocol opens avenues for generating contiguous genomes across various species.

Our study raises the question of finding polymerases with minimal bias. While our tests with *B. torquatus* and human indicate that polymerase C shows satisfactory performance for mammals, we found that polymerase C also appears to exhibit bias for samples of molluscs and collembola, albeit a different bias compared to polymerase A/B (Additional file 1: Figs. S5, S6). Anticipating that DNA amplification will constitute a key step in the genome sequencing procedure for numerous collection samples, challenging species, and species with diminutive body sizes, future investigations could focus on identifying the most appropriate polymerase or combination of polymerases that exhibit minimal bias for specific taxonomic groups.

Apart from the ultra-low input protocol, several new approaches have recently been developed to make small amounts of input DNA accessible for long read sequencing. This includes the above-mentioned MDA [41, 42, 45], adapter ligation via tagmentation [46, 47], and Picogram input multimodal sequencing (PiMmS) [48]. We show here that the ultra-low input protocol produces very few chimeric reads in contrast to MDA. Furthermore, the ultra-low input protocol can generate average read lengths of ~10 kb, which is similar to read lengths generated by PiMmS [48], but substantially longer than those generated with tagmentation based approaches (2.5–5 kb averages) [46, 47]. Nevertheless, different methods likely have ideal application ranges that depend on the input sample, its quality, and amount of DNA. Future research should therefore benchmark which library preparation method is optimal for which sample type.

Conclusions

Our work suggests that collections can complement flash-frozen material as a sample source for biodiversity genomics, especially for species that are hard to sample because of rarity, protection status or other reasons. Thus, natural history collections as extensive archives of biodiversity can help to achieve the ambitious goal of generating reference genomes for all life on Earth.

Methods

Sample sources

For *Bradypus torquatus*, we used a sample of ~50 mg of clogged blood, preserved in ethanol. This sample was collected in 2003 and provided by the Taxonomic Collection Center of the Federal University of Minas Gerais (CCT-UFMG). The sample was exported under CITES license number 138261, and the access to genetic resources of Brazil is registered at SISGEN number AF86294. We note that a recent taxonomic review [49] suggested that maned sloths should be split into two species: the northern (*Bradypus torquatus*) and southern (*Bradypus crinitus*) maned sloths. Under this classification, which is not yet officially adopted, our genome assembly, generated from a sample collected in the state of Espírito Santo in Brazil, represents the *Bradypus crinitus* lineage. For *Idiurus macrotis*, we used ~12 mg of skin with hair preserved in technical ethanol at room temperature. For *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 Leibniz Institute for the Analysis of Biodiversity Change (LIB) Bonn. Originally, fin clips of specimens acquired from local fishermen were taken in 2014, immediately placed into ethanol, but subsequently transported multiple times at room temperature until final storage at -20°C . The exact time between catch and sampling is unknown but was likely a few hours. 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 95,545) and ~16 mg (individual 95,541) of muscle tissue and ~30 mg (individual 56,492) 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, which we immediately homogenized for DNA extraction after euthanization. This sample was collected under license ESNC 205 issued by the Spanish “Dirección General de Biodiversidad, Bosques y Desertificación del Ministerio para la Transición Ecológica y el Reto Demográfico”. 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 immediately 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. Additional File 2: Table S1 lists sample sources, accessions and additional details.

DNA extraction

High molecular weight (HMW) gDNA was extracted from ethanol-preserved clogged blood 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 ($18,000 \times g$ 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 $1 \times$ 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 Technologies).

For *Idiurus macrotis*, *Desmana moschata*, *Muscardinus avellanarius*, *Cathorops nuchalis*, *Cathorops wayuu*, *Xerotyphlops vermicularis*, *Dipus sagitta*, *Ptilocercus lowii*, and the two *Anguis fragilis* samples, gDNA was extracted according to the protocol of [50]. DNA concentration and DNA fragment length were assessed using the Qubit dsDNA BR Assay kit on the Qubit Fluorometer (Thermo Fisher Scientific) and the Genomic DNA Screen Tape on the Agilent 4150 TapeStation system (Agilent Technologies). For *Elysia timida* and *Scintilla philippinensis*, gDNA was extracted using a CTAB-based method [51] and a bead-based protocol [52], respectively, including a pre-wash with sorbitol. The MagAttract HMW DNA Kit from Qiagen was used to extract gDNA from *Podura aquatica*. For these gDNA extractions, DNA concentration and DNA fragment length were assessed using Qubit quantification (Thermo Fisher Scientific), the Agilent 2200 TapeStation system (Agilent Technologies) and the Femto Pulse system (Agilent Technologies).

All details on the DNA yield and DNA fragment sizes can be found in Additional File 2: Table S1.

Low input PacBio HiFi library preparation

The low input protocol allows generating PacBio libraries for samples with limited DNA content without amplification [26]. We prepared low input PacBio HiFi libraries according to the instructions of the SMRTbell Express Prep Kit v2.0, except for the libraries of *Cathorops nuchalis* and *Cathorops wayuu* which were prepared with the SMRTbell prep kit v3.0.

Ultra-low input PacBio HiFi library preparation

PacBio ultra-low input HiFi libraries were prepared with the SMRTbell Express Template Prep Kit 2.0 according to the “Procedure & Checklist—Preparing HiFi SMRTbell® Libraries from Ultra-Low DNA Input” (PN 101–987-800 Version 02). To reduce potential PCR bias of polymerase A/B, we used in our modified protocol a third PCR reaction, making use of Polymerase C (KOD Xtreme™ Hot Start DNA Polymerase, Merck PN 71975), which is optimized for the amplification of long strands and GC-rich DNA

templates. A detailed report of the Ultra-Low Input protocol for Polymerase C can be found in Additional File1: Note 1.

The amplified DNA from two PCR reactions with polymerase A and B was pooled equimolarly. PCR fragments from polymerase C amplification were kept separately and processed independently from the pooled fragments produced with polymerase A and B. Purified and pooled amplified DNA libraries were size selected to remove smaller fragments (Additional File 2: Table S1).

For *Anguis fragilis* and *Idiurus macrotis*, we prepared two additional libraries with DNA extracts to which a DNA repair step was applied using the Sequential Reaction Protocol for PreCR Repair Mix (New England BioLabs) prior to the actual library preparation.

PacBio sequencing

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

HiC for scaffolding the *B. torquatus* assembly

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

Comparing polymerase A/B and C read assemblies

Aiming to evaluate the impact of libraries generated with polymerase A/B vs. C on the genome assembly quality, we combined different datasets with varying coverages, library complexities (number of libraries), and polymerase combinations (only A/B, only C, and A/B + C). For tests that did not involve all read data, we randomly subsampled reads. Subsequently, we assembled the read data into a contig assembly, as described below, and compared the summary metrics, including contig N50, number of contigs and gene completeness. All results are listed in Additional File 2: Tables S4 and S5.

Contig assembly

HiFi reads were called using a pipeline consisting of PacBio's tools ccs 6.4.0 (<https://github.com/PacificBiosciences/ccs>) and actc 0.3.1 (<https://github.com/PacificBiosciences/actc>) as well as samtools 1.15 [53] and DeepConsensus 0.2.0 or 1.2.0 [27]. All commands were executed as recommended in the respective guide for DeepConsensus (https://github.com/google/deepconsensus/blob/v0.2.0/docs/quick_start.md; e.g., ccs

–all). To remove PCR adapters and PCR duplicates, which might originate from the PCR amplification during the ultra-low library preparation, PacBio’s tools lima 2.6.0 (<https://github.com/PacificBiosciences/barcoding>) with options “–num-threads 67 –split-bam-named –same” and pbmarkdup 1.0.2–0 with options “–num-threads 67 –log-level INFO –log-file pbmarkdup.log –cross-library –rmdup” (<https://github.com/PacificBiosciences/pbmarkdup>) were applied to samples prepared with the ultra-low library preparation protocol. For the Catfish samples *Cathorops nuchalis* and *C. wayuu* that were sequenced using the low-input library preparation protocol, PacBio sequencing adapters were removed with HiFiAdapterFilt [54]. The resulting reads were merged and then decontaminated with kraken2 v. 2.1.3 [55] using the kraken2 PlusPFP database downloaded in March 2023, with a confidence score of 0.51.

After HiFi calling, we used hifiasm v0.19.5 [28, 56] to assemble HiFi reads obtained from the *Cathorops nuchalis*, *C. wayuu*, *Idiurus*, *Anguis*, *Elysia*, *Scintilla*, and *Podura* samples. For the two catfish samples *Cathorops nuchalis* and *C. wayuu*, because of sub-optimal performance with default parameters, we tested several hifiasm options before deciding which parameters produce the best assembly in terms of gene completeness and contiguity (Additional File 2: Table S2). To this end, we estimated the genome profile of these two species with FastK (<https://github.com/thegenemyers/FASTK>) and Genescope.FK (<https://github.com/thegenemyers/GENESCOPE.FK>) with $k=30$ to find the homozygous peak that was then passed to hifiasm (Additional File 2: Table S2). In all other cases, we applied default parameters with strict haplotig purging (-l3 parameter), and for the *Elysia* sample, we additionally used available Arima HiC data for assembly phasing.

Contiguity statistics were calculated with Quast 5.0.2 [57], gfastats v. 1.3.6 (<https://github.com/vgl-hub/gfastats>) and Merqury.FK (<https://github.com/thegenemyers/MERQURY.FK>). Gene completeness was evaluated with BUSCO 5.5.0 [30] as well as compleasm 0.2.5 [29]. We used the eutherian_odb10 dataset for *Bradypus torquatus*, the actinopterygii_odb10 dataset for *C. nuchalis* and *C. wayuu*, the mammalia_odb10 dataset for human, the arthropoda_odb10 dataset for *Podura aquatica*, and the metazoa_odb10 dataset for *Elysia timida* and *Scintilla philippinensis*.

For *B. torquatus*, we initially obtained hifiasm (v0.19.5) assemblies that were of a size expected from four haplotypes of this genome, consisting of a large number of small contigs (Additional File 2: Table S6). Similar results were obtained with HiCanu (v2.2) [58], which is designed to break contigs at all joins in the assembly graph, meaning any divergences between the four theoretical haplotypes would result in a new contig (in our case over 200,000 assembled contigs totaling almost 12 Gb of sequence). This indicated that the tissue samples we obtained for this species originated from two different individuals. While the accuracy of the PacBio HiFi reads should in principle allow distinguishing all four haplotypes, *B. torquatus* is expected to have a very low heterozygosity and high in-breeding rate due to small population size, which results in assembly graphs where many regions collapse all haplotypes due to the lack of sequence variation.

To overcome this problem, we used the assembler Flye (v2.9.2) [59], which allows users to set the read error rate as an argument. Flye has been previously suggested by the developers as a method for collapsing sequences from highly diverged haplotypes into a single “pseudo-haplotype” sequence (<https://github.com/fenderglass/Flye/issues/636>).

Here, we found that a read error rate of 3% produced the most contiguous assembly, when combined with a reduced read-overlap of 5 kb (Additional File 2: Table S6). The latter deviates from the default value selected by Flye, which Flye would determine by the N90 of the input reads (in our case the N90 was 9 kb for the HiFi library, which had a modal read length of ~10 kb). We then removed retained haplotigs using purge dups [60].

Contamination detection and read coverage analysis

Specimens stored in liquid preservation media are prone to various levels of DNA contamination from non-target organisms [33], caused by different handling and storage conditions that are often hard to retrace [61]. To detect levels of contamination from exogenous DNA in our assemblies, we used NCBI's Foreign Contamination Screen (FCS 0.5.0) [43], which flags both putative adapter sequences (FCS-adaptor) and contigs assigned to non-target species (FCS-GX). 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 December 5, 2023). Furthermore, to visualize contamination across the respective contig-level assemblies before FCS-filtering, we used blobtoolkit v4.1.4 [44], which assigns all contigs from a given assembly to a taxonomic group based on best blast hits (Additional file 1: Fig. S2).

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, Additional File 2: Table S5) assemblies obtained for *Podura*, *Scintilla* and *Elysia*, and visually inspected mapped reads (Additional file 1: Figs. S5, S6).

To further quantify PCR bias, we calculated the normalized coverage (coverage of each nucleotide divided by the average coverage) of each polymerase A/B and C *Bradypus torquatus* library, using either the *Choloepus didactylus* genome or the best assembly of *Bradypus torquatus*. We also calculated normalized coverage of a non-amplified human library (downloaded from <https://downloads.pacbc.loud.com/public/revio/2022Q4/HG002-rep1/>; last accessed 19 Sep 2024) as well as polymerase C (produced in this study) and A/B amplified libraries (NCBI, BioProject PRJNA657245, accessions SRR12454519 and SRR12454520) sequenced from the human cell line HG002, using the human HG002 assembly [6] (v.1.1, maternal haplotype). We then computed normalized coverage across nucleotides assigned to exonic and repeat sequences. For *C. didactylus* and *B. torquatus*, exons were annotated by TOGA v1.0.0 and repeats were annotated with RepeatModeler [62] and RepeatMasker v4.1.4 (<https://www.repeatmasker.org/>). For human, exons were annotated by RefSeq (v110 from CHM13, JHU v5.2), <https://ccb.jhu.edu/T2T.shtml>) and annotated repeats [63] were downloaded from the UCSC table browser [64] (last accessed

19 Sep 2024). Read mapping was performed using minimap2 v2.26 [65] with HiFi read mapping parameters (`-ax map-hifi`), and absolute coverage per base and across annotations was computed with samtools v1.17 [53], using the “samtools depth” and “samtools bedcov” commands, respectively. For the HG002 gene annotation, we filtered the annotation to only include coding exons to enable a fair comparison with the TOGA annotations that do not include non-coding transcripts or UTRs.

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) [66], 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) [39]. A full list of commands is given in Additional file 1: Note 2. 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 [43]. 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`).

Read chimer analysis

To investigate whether our modified amplification-based protocol creates more chimeric reads, we mapped reads (all obtained from the human HG002 sample) against the HG002 reference genome [6] (v.1.1, maternal haplotype, <https://github.com/marbl/hg002?tab=readme-ov-file>), using minimap2 v2.26 [65] with HiFi read mapping parameters (`-ax map-hifi`). We used reads amplified with polymerase C and polymerase A/B (NCBI BioProject PRJNA657245, accessions SRR12454519 and SRR12454520), as well as the non-amplified reads (<https://downloads.pacbcloud.com/public/revio/2022Q4/HG002-rep1/>; last accessed 19 Sep 2024), and reads amplified with MDA (NCBI BioProject PRJNA1005794, accession SRR25653511). To calculate the fraction of alignments classified as primary alignments, secondary alignments, supplementary alignments, and unmapped, we counted the flags assigned by minimap using samtools v1.17 [53] with the command “samtools view.” Raw read lengths and alignment lengths of primary and supplementary alignments were extracted from raw fastq-files and sam-files created by minimap2, respectively.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03487-9>.

Additional file 1: Fig. S1: Levels of exogenous DNA contamination. For the samples sequenced with the PacBio ultra-low library protocol, we assembled the HiFi reads and screened the contigs for exogenous DNA contamination.

The left plots show the total length of sequences not flagged or flagged as contamination by FCS. The right plots show the total length of sequences assigned to a certain contaminant taxa. Sequences tagged by the FCS status “exclude” and “trim” show clear non-endogenous taxonomic signals, while sequences tagged with “review” are borderline cases. For contig assembly of reads obtained from the *Idiurus macrotis* sample, most of the sequenced DNA is contamination from bacteria. Aligning the sequenced reads to the genome of a close relative revealed an extremely low mapping rate of < 0.1%, confirming that essentially no endogenous DNA is in the sample. For contig assembly of reads obtained from the *Anguis fragilis* sample from 2021, the majority of the assembled sequence is putatively endogenous DNA; however bacterial contamination is also present. Aligning the sequenced reads to the genome of a close relative showed a moderate mapping rate of 36.27%, confirming that a part of the sequenced DNA is endogenous. For the final chromosome-level assembly of the maned sloth, no contamination was found. FCS identifies very low levels of plant contamination. Consistent with these results for the final assembly, screening the HiFi reads of the first SMRT cell with kraken2 [54] indicated very low levels of non-mammalian DNA in the sample. The *Elysia timida* hifi-asm-phased assembly of haplotype 1, generated from HiFi data obtained with both polymerases A/B and C, exhibits low levels of contamination mostly coming from bacteria.

Fig. S2: Breakdown of taxa contributing to exogenous DNA contamination. Blobplots created with blobtoolkit [44] for assemblies created with the ultra-low input protocol. The central plot depicts assembly contigs as blobs, with the blob sizes corresponding to contig sizes and colors representing the assigned taxon as inferred from the best blast hits. The Y-axis shows the coverage of sequenced reads mapped back to the respective contig and the X-axis shows the average GC-content of each contig. Histograms on top and right hand side summarize average GC-content and coverage of assembly contigs, partitioned by the assigned taxa. The contig assembly of *Idiurus macrotis* reads shows that most of the assembled contigs are assigned to different bacteria, and coverage as well as GC content of assembled contigs varies widely, indicating that the assembly mostly consists of contaminant sequence. For the contig assembly of the *Anguis fragilis* sample from 2021, the majority of the assembly is assigned to either “Chordata” or “no-hit”, and “no-hit” contigs partly resemble “Chordata” contigs in terms of coverage and GC-content. On the other hand, the longest contigs in this assembly are assigned to “Pseudomonadota”, meaning that there is evident bacterial contamination present in the assembly. This sample was collected as a roadkill, and partial decomposition by bacteria provides an explanation for the observed contamination. In the final contig level assembly of *Bradypus torquatus*, only two small contigs are assigned to “Streptopytha”, highlighting the absence of systematic contamination. The majority of contigs of the *Elysia timida* hifi-asm-phased assembly of haplotype 1 is assigned to “Mollusca”, but the blobplot clearly shows several other clusters consisting of large contigs assigned to “Pseudomonadota” and other bacterial taxa or “no-hit”. As the individual sequenced here was immediately killed and homogenized for sequencing without prior fixation, sample decomposition is unlikely. Rather, *Elysia timida* is too small to dissect before sequencing, therefore the observed contamination is likely caused by the gut microbiome and other commensal microorganisms. It should be noted that mollusca are underrepresented in the blast database, which likely increases false-positive hits and “no-hit” assignments [66].

Fig. S3: Amplification with polymerase C creates more even read coverage across exons and repeats. Violin plots depict the distribution density of normalized read coverage, calculated as the read coverage at a given genomic position divided by the average coverage of the respective library. Horizontal lines show 25% quartile boundaries, median and 75% quartile boundaries, respectively. Dotted lines represent a normalized coverage of 0.25. Only genomic regions annotated as exons or repeats were considered. Reads from polymerase C cover exonic and repeat regions more evenly, as the mode of the distribution is more pronounced around ~ 1.0 normalized coverage. Number of unique exons having ≤ 0.25 normalized coverage. The overall number of unique exons is 256,011, 246,006, and 239,251. There is an up to ten-fold excess of exons having ≤ 0.25 coverage for polymerase A/B compared to polymerase C, highlighting potential drop-outs that can lead to a fragmented assembly and missing annotations. *Bradypus torquatus* sequencing reads obtained with polymerase A/B or C, aligned against the *Choloepus didactylus* genome or the *Bradypus torquatus* genome. Human sequencing data aligned against the HG002v1.1 maternal haplotype assembly, comparing reads obtained with polymerase A/B or C and reads obtained without amplification.

Fig. S4: Low fraction of chimeric reads in amplified PacBio HiFi-read libraries. The Y-axis shows different types of alignments as classified by minimap2 [64] when mapping human reads to the HG002v1.1 assembly. The X-axis depicts the fraction of alignments from a library. Supplementary alignments of a read indicate that the read is a chimera consisting of two different genomic regions. Alignments of non-amplified ultra-low input reads serve as a baseline and show a similarly low fraction of supplementary alignments as reads obtained with polymerase A/B and polymerase C. In stark contrast, the majority of MDA reads have supplementary alignments and are likely chimeric. Histograms of raw read and alignment lengths of different PacBio libraries, following the color scheme of A. The filled histogram areas indicate raw read length. The red and blue lines show primary and supplementary alignment lengths excluding soft- and hard-clipped bases. Whereas primary alignment lengths closely follow raw read length distributions for ultra-low and non-amplified reads, indicating full length mapping, for MDA reads, the mode of the primary alignment distribution is located at around ~ 5000 nt compared to ~ 11,000 nt length for raw reads, clearly showing that primary alignments tend to be severely truncated. In accordance with, the amount and length of supplementary alignments is very low in non-amplified and ultra-low reads, but exceptionally high in MDA reads.

Fig. S5: Both polymerase A/B and C exhibit PCR bias for *Elysia timida*. IGV screenshots show alignments of HiFi reads produced with the ultra-low input protocol using either polymerase A/B or polymerase C. Reads were aligned to an assembly of *Elysia timida* generated with all reads from all three polymerases. Two examples of genomic regions where HiFi read coverage is very low or drops to zero for polymerase A/B while polymerase C reads cover the region. This exemplifies regions difficult to sequence with polymerase A/B. Two examples of loci where read coverage drops to zero for polymerase C but not polymerase A/B. While inreads from polymerase A/B still cover the locus, it should be noted that there are fewer reads compared to the flanking regions. This indicates that it is also difficult for polymerase A/B to amplify these genomic regions.

Fig. S6: Polymerase A/B and C biases for *Podura aquatica* and *Scintilla philippinensis*. IGV screenshots show alignments of HiFi reads produced with the ultra-low input protocol using either polymerase

A/B or polymerase C. Reads were aligned to assemblies of *Scintilla philippinensis* and *Podura aquatica* generated with all reads from polymerases A/B and C. Two examples of genomic regions where HiFi read coverage is very low or drops to zero for polymerase A/B, while polymerase C reads cover these regions. This exemplifies regions difficult to sequence with polymerase A/B. Two examples of loci where read coverage drops to zero for polymerase C but not polymerase A/B. Of note, in panel D, polymerase A/B coverage is also comparatively low, and in panel B, there is a polymerase A/B coverage dropout upstream of the polymerase C dropout. Together, this illustrates that for certain genomic regions one or both polymerases have difficulty amplifying DNA. Note 1: PacBio ultra-low library preparation based on PCR amplification with KOD Xtreme™ Hot Start DNA Polymerase (Merck). Note 2: Commands used for scaffolding the *Bradypus torquatus* assembly.

Additional file 2: Table S1: Detailed list of species and samples, together with DNA extraction and long-read sequencing results of each SMRT cell. Table S2: Assembly statistics for *Cathorops nuchalis* and *Cathorops wayuu*. Table S3: High levels of contamination in samples of Russian desman, Hazel dormouse and slow worm. Table S4: *Bradypus torquatus* genome assemblies generated with data from ultra-low input libraries prepared with polymerases A/B and/or C at different coverages. Table S5: Genome assemblies generated with polymerases A/B and/or C data at different coverages for *Elysia timida*, *Scintilla philippinensis*, and *Podura aquatica*. Table S6: Assembly statistics for all tested contig assemblers using sequenced PacBio HiFi reads from four SMRTs produced with polymerase A/B and one SMRT produced with polymerase C for *Bradypus torquatus*.

Additional file 3: Review history.

Acknowledgements

We thank Deniz Kaya from PacBio for advice and suggestions on adapting the ultra-low input protocol and Sarah Kingan, Juniper Lake, Ian McLaughlin, Aaron Wenger, 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 (<https://doi.org/10.17169/refubium-26754>). We acknowledge Irina Ruf (Senckenberg Frankfurt), Carles Galà 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.

Peer review information

Kin Fai Au and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Review history

The review history is available as Additional File 3.

Funding

Open Access funding enabled and organized by Projekt DEAL. 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 (LOEWE/1/10/519/03/03.001(0014)/52).

Data availability

Assemblies and sequencing data created in this study are provided under the NCBI umbrella project "Collomic Genome Assemblies" (Accession: PRJEB80235). The raw sequencing data and assemblies for *Bradypus torquatus* are available at NCBI under BioProject PRJEB73341 and BioSample SAMEA115348596. An improved version of the *Elysia timida* assembly after incorporating additional polymerase C reads and Hi-C scaffolding [67] is available under Bioproject PRJNA1119176 and Biosample SAMN42332041. The *Scintilla philippinensis* assembly and raw sequencing data are available under Bioproject PRJNA1120792. Genome assemblies and raw sequencing data of both catfish genomes are available under Bioproject PRJNA1162287 (*Cathorops nuchalis*) and PRJNA1162286 (*Cathorops wayuu*). The *Podura aquatica* assembly and sequencing data are available under Bioproject PRJNA1163304. Raw reads of the contaminated samples *Idiurus macrotis* and *Anguis fragilis* are available under BioProject PRJNA1211858 and PRJNA1212865. Raw reads and assemblies obtained with polymerase C for HG002 are available on <https://downloads.pacbcloud.com/public/review/2023Q3/KODXtreme/>. The TOGA annotation for the *B. torquatus* is available at <https://genome.senckenberg.de/download/TOGA/>. Ultra-low input-based assemblies generated in this study are also available at <https://genome.senckenberg.de/download/GenomesCollectionsPolC>. No new computer code was generated in this study.

Declarations

Competing interests

The authors declare no competing interests.

Author details

¹LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, Frankfurt 60325, Germany. ²Senckenberg Research Institute, Senckenberganlage 25, Frankfurt 60325, Germany. ³Institute of Cell Biology and Neuroscience, Faculty of Biosciences, Goethe University, Max-Von-Laue-Str. 9, Frankfurt 60438, Germany. ⁴Center for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn, Adenauerallee 127,

Bonn 53113, Germany. ⁵Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Königin-Luise-Straße 2-4, Berlin 14195, Germany. ⁶Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, Berlin 10315, Germany. ⁷Senckenberg Research Institute, Am Museum 1, Görlitz 02826, Germany. ⁸Global Genome Initiative, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, USA. ⁹International Institute of Aquaculture and Aquatic Sciences, Universiti Putra Malaysia, Port Dickson, Negeri Sembilan 71050, Malaysia. ¹⁰Laboratório de Biodiversidade E Evolução Molecular, Departamento de Genética, Universidade Federal de Minas Gerais, Ecologia E Evolução, Belo Horizonte, Minas Gerais, Brazil. ¹¹Museum Für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstraße 43, Berlin 10115, Germany. ¹²Max Planck Institute of Molecular Cell Biology and Genetics, Pfötenhauerstr. 108, Dresden 01307, Germany. ¹³DRESDEN Concept Genome Center, Technische Universität Dresden, Fetscherstraße 105, Dresden 01307, Germany.

Received: 1 May 2024 Accepted: 27 January 2025

Published online: 10 February 2025

References

- Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth biogenome project: sequencing life for the future of life. *Proc Natl Acad Sci U S A*. 2018;115:4325–33.
- Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature*. 2020;587:240–5.
- Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020;587:252–7.
- Ronco F, Matschiner M, Böhne A, Boila A, Büscher HH, El Taher A, et al. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*. 2021;589:76–81.
- Kuderna LFK, Gao H, Janiak MC, Kuhlwiilm M, Orkin JD, Bataillon T, et al. A global catalog of whole-genome diversity from 233 primate species. *Science*. 2023;380:906–13.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376:44–53.
- Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet*. 2023;55:1221–31.
- Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*. 2020;370. <https://doi.org/10.1126/science.abc6617>.
- Kautt AF, Kratochwil CF, Nater A, Machado-Schiaffino G, Olave M, Henning F, et al. Contrasting signatures of genomic divergence during sympatric speciation. *Nature*. 2020;588:106–11.
- Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, et al. Six reference-quality genomes reveal evolution of bat adaptations. *Nature*. 2020;583:578–84.
- Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–46.
- Blumer M, Brown T, Freitas MB, Destro AL, Oliveira JA, Morales AE, et al. Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding. *Sci Adv*. 2022;8: eabm6494.
- Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. *Science*. 2022;376: eabj6965.
- Osipova E, Barsacchi R, Brown T, Sadanandan K, Gaede AH, Monte A, et al. Loss of a gluconeogenic muscle enzyme contributed to adaptive metabolic traits in hummingbirds. *Science*. 2023;379:185–90.
- Shao Y, Zhou L, Li F, Zhao L, Zhang B-L, Shao F, et al. Phylogenomic analyses provide insights into primate evolution. *Science*. 2023;380:913–24.
- Morales AE, Dong Y, Brown T, Baid K, Kontopoulos DG, Gonzalez V, et al. Bat genomes illuminate adaptations to viral tolerance and disease resistance. *Nature*. 2025. <https://doi.org/10.1038/s41586-024-08471-0>.
- Blom MPK. Opportunities and challenges for high-quality biodiversity tissue archives in the age of long-read sequencing. *Mol Ecol*. 2021;30:5935–48.
- Johnson KR, Owens IFP, Global Collection Group. A global approach for natural history museum collections. *Science*. 2023;379:1192–4.
- Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, et al. A comprehensive and dated phylogenomic analysis of butterflies. *Curr Biol*. 2018;28:770–8.e5.
- Heinicke MP, Nielsen SV, Bauer AM, Kelly R, Geneva AJ, Daza JD, et al. Reappraising the evolutionary history of the largest known gecko, the presumably extinct *Hoplodactylus delcourti*, via high-throughput sequencing of archival DNA. *Sci Rep*. 2023;13:1–12.
- Tan HZ, Jansen JFF, Allport GA, Garg KM, Chattopadhyay B, Irestedt M, et al. Megafaunal extinctions, not climate change, may explain Holocene genetic diversity declines in Numenius shorebirds. *Elife*. 2023;12. <https://doi.org/10.7554/eLife.85422>.
- Mulcahy DG, Macdonald KS 3rd, Brady SG, Meyer C, Barker KB, Coddington J. Greater than X kb: a quantitative assessment of preservation conditions on genomic DNA quality, and a proposed standard for genome-quality DNA. *PeerJ*. 2016;4:e2528.
- Dahn HA, Mountcastle J, Balacco J, Winkler S, Bista I, Schmitt AD, et al. Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *Gigascience*. 2022;11. <https://doi.org/10.1093/gigascience/giac068>.
- Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37:1155–62.

25. Sharma P, Al-Dossary O, Alsubaie B, Al-Mssallem I, Nath O, Mitter N, et al. Improvements in the sequencing and assembly of plant genomes. *GigaByte*. 2021;2021:gigabyte24.
26. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A high-quality DE novo genome assembly from a single mosquito using PacBio sequencing. *Genes*. 2019;10: 62.
27. Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol*. 2023;41:232–8.
28. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5.
29. Huang N, Li H. compleasm: a faster and more accurate reimplement of BUSCO. *Bioinformatics*. 2023;39. <https://doi.org/10.1093/bioinformatics/btad595>.
30. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38:4647–54.
31. PacBio Biosciences. Now available: ultra-low DNA input workflow for SMRT sequencing [Internet]. 2020. Available from: <https://www.pacb.com/blog/introducing-the-ultra-low-input-protocol-for-smrt-sequencing/>. Cited 2023.
32. Schneider C, Woehle C, Greve C, D'Haese CA, Wolf M, Hiller M, et al. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *Gigascience*. 2021;10. <https://doi.org/10.1093/gigascience/giab035>.
33. Raxworthy CJ, Smith BT. Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol Evol*. 2021;36:1049–60.
34. Irestedt M, Thörn F, Müller IA, Jönsson KA, Ericson PGP, Blom MPK. A guide to avian museomics: insights gained from resequencing hundreds of avian study skins. *Mol Ecol Resour*. 2022;22:2672–84.
35. Strunov A, Kirchner S, Schindelar J, Kruckenhauser L, Haring E, Kapun M. Historic museum samples provide evidence for a recent replacement of *Wolbachia* types in European *Drosophila melanogaster*. *Mol Biol Evol*. 2023;40. <https://doi.org/10.1093/molbev/msad258>.
36. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene annotation with orthology inference at scale. *Science*. 2023;380:eabn3107.
37. Gibb GC, Condamine FL, Kuch M, Enk J, Moraes-Barros N, Superina M, et al. Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living xenarthrans. *Mol Biol Evol*. 2016;33:621–42.
38. Osipova E, Ko MC, Petricek KM, Yung Wa Sin S, Brown T, Winkler S, et al. Convergent and lineage-specific genomic changes contribute to adaptations in sugar-consuming birds. *bioRxiv*. 2024. <https://doi.org/10.1101/2024.08.30.610474>.
39. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*. 2023;39. <https://doi.org/10.1093/bioinformatics/btac808>.
40. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245.
41. Hård J, Mold JE, Eisefeldt J, Tellgren-Roth C, Häggqvist S, Bunikis I, et al. Long-read whole-genome analysis of human single cells. *Nat Commun*. 2023;14:5164.
42. Lee Y-C, Ke H-M, Liu Y-C, Lee H-H, Wang M-C, Tseng Y-C, et al. Single-worm long-read sequencing reveals genome diversity in free-living nematodes. *Nucleic Acids Res*. 2023;51:8035–47.
43. Astashyn A, Tvedte ES, Sweeney D, Sapozhnikov V, Bouk N, Joukov V, Mozes E, Strobe PK, Sylla PM, Wagner L, Bidwell SL. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol*. 2024;25(1):60. <https://doi.org/10.1101/2023.06.02.543519>.
44. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*. 2020;10(4):1361–74.
45. Roberts NG, Gilmore MJ, Struck TH, Kocot KM. Multiple Displacement Amplification Facilitates SMRT Sequencing of Microscopic Animals and the Genome of the Gastrotrich *Lepidodermella squamata* (Dujardin 1841). *Genome Biol Evol*. 2024;16(12):evae254.
46. Jia H, Tan S, Cai Y, Guo Y, Shen J, Zhang Y, et al. Low-input PacBio sequencing generates high-quality individual fly genomes and characterizes mutational processes. *Nat Commun*. 2024;15:5644.
47. Nanda AS, Wu K, Irklyenko I, Woo B, Ostrowski MS, Clugston AS, et al. Direct transposition of native DNA for sensitive multimodal single-molecule sequencing. *Nat Genet*. 2024;56:1300–9.
48. Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, et al. Ancient diversity in host-parasite interaction genes in a model parasitic nematode. *Nat Commun*. 2023;14:7776.
49. Miranda FR, Garbino GS, Machado FA, Perini FA, Santos FR, Casali DM. Taxonomic revision of maned sloths, subgenus *Bradypus* (Scaelopis), Pilosa, Bradypodidae, with revalidation of *Bradypus crinitus* Gray, 1850. *J Mammal*. 2023;104:86–103.
50. Sambrook J, Russell WD. Protocol 1: DNA isolation from mammalian tissue. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: New York. 2001. p. 623–7.
51. Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res*. 1980;8:4321–5.
52. Mayjonade B, Gouzy J, Donnadiou C, Pouilly N, Marande W, Callot C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques*. 2016;61:203–5.
53. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10. <https://doi.org/10.1093/gigascience/giab008>.
54. Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics*. 2022;23:157.
55. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20:257.
56. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol*. 2022;40:1332–5.

57. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*. 2018;34:i142–50.
58. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020;30:1291–305.
59. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019;37:540–6.
60. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36:2896–8.
61. Ruiz-Gartzia I, Lizano E, Marques-Bonet T, Kelley JL. Recovering the genomes hidden in museum wet collections. *Mol Ecol Resour*. 2022;22:2127–9.
62. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117:9451–7.
63. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science*. 2022;376: eabk3112.
64. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32:D493–6.
65. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
66. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*. 2013. Available from: <http://arxiv.org/abs/1303.3997>
67. Männer L, Schell T, Spies J, Galià-Camps C, Baranski D, Ben Hamadou A, Gerheim C, Neveling K, Helfrich EJ, Greve C. Chromosome-level genome assembly of the sacoglossan sea slug *Elysia timida* (Risso, 1818). *BMC genomics*. 2024;25(1):941.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.