Higher quality *de novo* genome assemblies from degraded museum specimens: a linked-read approach to museomics

Jocelyn P. Colella, Anna Tigano, Matthew D. MacManes

Molecular, Cellular, and Biomedical Sciences Department, University of New Hampshire, Durham, NH 03857; jocelyn.colella@unh.edu (JPC), anna.tigano@unh.edu (AT), matthew.macmanes@unh.edu (MDM)

* Corresponding author: Jocelyn.Colella@unh.edu

Key words: 10X Genomics, assembly quality, natural history collections, *Peromyscus*

ABSTRACT

High-throughput sequencing technologies are a proposed solution for accessing the molecular data in historic specimens. However, degraded DNA combined with the computational demands of short-read assemblies has posed significant laboratory and bioinformatics challenges. Linked-read or 'synthetic long-read' sequencing technologies, such as 10X Genomics, may provide a cost-effective alternative solution to assemble higher quality *de novo* genomes from degraded specimens. Here, we compare assembly quality (e.g., genome contiguity and completeness, presence of orthogroups) between four published genomes assembled from a single shotgun library and four deer mouse (*Peromyscus* spp.) genomes assembled using 10X Genomics technology. At a similar price-point, these approaches produce vastly different assemblies, with linked-read assemblies having overall higher quality, measured by larger N50 values and greater gene content. Although not without caveats, our results suggest that linked-read sequencing technologies may represent a viable option to build *de novo* genomes from historic museum specimens, which may prove particularly valuable for extinct, rare, or difficult to collect taxa.

INTRODUCTION

A disconnect between the capabilities of high-throughput sequencing technologies and the quality, or lack thereof, of historic museum specimens has largely neutered the ability of genomic methods to access molecular data from degraded specimens. Natural history collections (NHCs) store a wide variety of species from across the globe, including those that are difficult to collect or extinct in the wild. Voucher specimens housed in NHCs have been an invaluable source of morphological material as they provide a reference for measuring change across both space and time. More recently, specimens contained in NHCs have been recognized as important repositories of genetic data (Payne & Sorenson, 2002; Wandeler, Hoeck & Keller, 2007) and have provided insight into the phylogenetic relationships and origins of species (Suarez & Tsutsui, 2004; McLean et al., 2015). Quick progress in genomics methods are now enabling the use of museum specimens in ways that were not imaginable until only a few years ago. “Museomics”, or the application of genomic techniques to museum specimens, has already uncovered reticulate evolutionary histories across hominids (Green et al. 2010; Meyer et al., 2012, 2014) and is increasingly resolving the phylogenetic Tree of Life (Teeling & Hedges, 2013; Lessa, Cook, D’Elia, & Opazo, 2014; Wood, González, Lloyd Coddington, & Scharff, 2018), with expanded applications including, but not limited to, identifying functional variants implicated in ecological adaptations (Opazo, Palma, Melo, & Lessa,. 2005) and estimating mutation rates and the timing of evolutionary events (Pélissié, Crossley, Cohen, & Schoville, 2018).

Over time, however, and through exposure to agents known to degrade nucleic acids (UV, temperature, pH, salt, chemical modification, etc.; Dessauer, Cole, & Hafner,

1990; Lindahl, 1993; Dean & Ballard, 2001; Willerslev & Cooper, 2005), DNA degrades into short fragments, which can complicate the application of genomic methods to museum specimens. Since the 1970s, when museums widely began archiving tissues, collection and preservation methods have varied widely, but generally evolved to accommodate changing analytical technologies, resulting in the variety of preservation methods (e.g., formalin, ethanol, ground, frozen, etc.) and quality of tissue collections available to researchers today. In addition to the challenges of tissue preservation, field conditions including weather, processing speed, and available cold storage options are inherently unpredictable, resulting in further inconsistencies in field-collected tissue quality.

De novo genome assembly is the computational process of optimally fitting short-read fragments output from sequencers into a larger contiguous whole-genome sequence, recovering critical information about the locations of genes and variants that are lost in the sequencing process. Assembly methods are based on the often-incorrect assumption that similar DNA fragments originate from the same position within the genome; therefore, assembly can be complicated by the presence of extended repeats or regions of high divergence that extend beyond the sequenced read length (Alkan, Sajjadian, & Eichler, 2011; Nagarajan & Pop, 2013). Unfortunately, methods that yield the highest quality *de novo* genome assemblies often require large quantities of high molecular weight (HMW) DNA as starting material for library preparation, as the ability to resolve sequencing artefacts in assembly improves with increasing read length. This prerequisite often makes these methods inaccessible to degraded specimens. For example, although the recent emergence of long-read sequencing technologies (>10-50

kb) has significantly improved the computational complexities of *de novo* genome assembly, long-read sequencing requires large quantities of HMW DNA as a starting material, making these methods impractical for most museum samples (Rowe et al., 2011). Prior to the development of long-read sequencing, the most common approach to *de novo* genome assembly has involved a combination of shotgun short-insert (<500 bp) and mate-pair long-insert (>2000 bp) libraries of varying insert size, where the first would be used for assembly and the second for scaffolding. Once again, scaffolding would be limited by fragmented DNA, as input molecules must be longer than the selected insert size. More recently, the protocol accompanying the assembler DISCOVAR denovo (Broad Institute, 2015; Weisenfeld, Kumar, Shah, Church, & Jaffe, 2017), which is based on single short-insert shotgun libraries sequenced to ~60X using 250 bp paired-end reads, appears to be a viable option for genome assembly from degraded samples. This approach proved cost-effective for the genome assembly of 20 *Heliconius* species (Edelman et al., 2018), but for organisms with larger genomes this option is significantly more expensive than other approaches due to the high coverage and longer read lengths required. An appealing alternative is reference-guided assembly (Rowe et al., 2011; Staats et al., 2013), where either raw reads are mapped to an existing high-quality reference genome from a closely related species to build a consensus sequence (Pop, 2009) or a related reference genome is used only as a scaffolding guide (Gnerre, Lander, Lindblad-Toh, & Jaffe, 2009). While this approach may offer a partial solution, high quality, closely related references, a prerequisite for this approach, are not available for a large number of ecologically relevant taxa yet. To overcome this obstacle, other studies have recommended avoiding whole-genome

sequencing (WGS) of museum specimens altogether, suggesting exome capture (Bi et al., 2013) or other reduced-representation approaches (Jones & Good, 2018) as an alternative proxy for accessing molecular data from museum specimens. However, in addition to complex laboratory work, these approaches retrieve only a restricted subset of sequence data relative to WGS. In addition to the potentially confounding effect of pervasive purifying selection on exonic coding regions (Jackson, Campos, & Zeng, 2014), exome sequencing further fails to represent significant regulatory or non-coding regions essential to phylogenetic reconstruction (Nei & Tateno, 1975; Lynch, 1989), and, with increasing awareness, for understanding the targets of adaptive evolution (Andolfatto, 2005; Brooks, Turkarslan, Beer, Lo, & Baliga, 2011).

In the grey area between second and third generation sequencing, linked-read or 'synthetic long-read' (SLR, Voskoboynik et al., 2013) sequencing may provide a cost-effective solution for *de novo* genome sequencing from degraded specimens. These methods allow the assembly of pseudo-long reads up to 18kb from short-read data with higher accuracy compared to true long-read sequencing techniques (Jiao & Schneeberger, 2017). Initially introduced by Illumina (Kuleshov et al., 2014; McCoy et al., 2014), SLR methods have not been widely adopted by evolutionary biologists and museum scientists. 10X Genomics (Zheng et al., 2016), a newer technology loosely based on innovations developed by the Illumina SLR technique, offers several advantages for museum science applications. Specifically, this method requires as little as 1 nanogram of input material and it is robust to the effects of input DNA quality. 10X Genomics uses microfluidics to split extracted DNA fragments across >100,000 partitions or 'GEM's (gel-coated beads). Each 'GEM' then contains a fraction (< 0.5%) of

the genome, which is further sheared and barcoded. Reads from the same partition or 'GEM' are sequenced via conventional Illumina short-read sequencing and assembled locally, by barcode, as they must be derived from the same original DNA fragment (Goodwin, McPherson, & McCombie, 2016; van Dijk, Jaszczyszyn, Naquin, & Thermes, 2018). Although HMW DNA is optimal for any method, the physical separation of DNA fragments in a 'GEM' largely eliminates the issue of degraded DNA and increases assembly confidence by geographically linking small-reads in genome-space, thereby reducing misassembly. This library preparation method can also facilitate allele phasing and the detection of structural variants, although its power will depend on the quality of starting DNA (Lee et al., 2016; Zheng et al., 2016).

While linked-reads are currently optimized for human genomes (kb.10xgenomics.com) and most often applied to cancer and biomedical related questions (Zheng et al., 2016), these methods are beginning to be explored in other taxa (orchids, Zhang et al., 2017; sea otters and beluga whales, Jones et al., 2017a, b). As a consequence of phylogeny, 10X Genomics methods are most easily extended to other mammalian taxa, expected to have similar genome size and structure (*e.g.,* repeat content, heterozygosity, etc.). Thus, as a proof-of-concept, we compare assembly quality and content of four deer mouse (*Peromyscus*) genomes sequenced and assembled using the 10X Genomics linked-read approach with — at a comparable cost — four publically-available shotgun Illumina mammalian genome assemblies generated from comparable read volumes. We demonstrate the utility of this economical approach to whole genome reconstruction for researchers interested in questions related to systematics and functional genomics.

MATERIALS & METHODS

Twenty-five micrograms of frozen liver tissue from each of four field-collected museum specimens (*Peromyscus attwateri* [MSB:Mamm:84733], *Peromyscus aztecus* [MSB:Mamm:48205], *Peromyscus melanophrys* [MSB:Mamm:273915], and *Peromyscus nudipes* [MSB:Mamm:70743]) were loaned from the Museum of Southwestern Biology (MSB). Three of the specimens were collected internationally and collections dates ranged from 1982 to 2006 (Table 1). Genomic DNA was extracted using a standard QIAGEN Genomic Tip (Valencia, CA, USA) protocol. DNA was quantified with Qubit and its quality was assessed using an Agilent TapeStation (Santa Clara, CA, USA). DNA from each of the four species contained a distribution of DNA fragments heavily skewed towards smaller molecular weights. As fragment size distribution greatly influences the contiguity of the genome assembly, we further processed the samples using the Circulomics short read eliminator kit (Baltimore, MD, USA), which removes DNA molecules shorter than 10kb, and progressively up to 25 kb, thereby removing the vast majority of our DNA sample. The remaining DNA from each of these samples was sent to the Genomics Core Facility at Icahn School of Medicine at Mount Sinai for library preparation, where samples were run on a Femto Pulse (Agilent, Santa Clara, CA, USA) to assess fragment size distribution post-Circulomics (Supplemental Information). Resulting 10X libraries were sequenced at Novogene (Novogene, Sacramento, CA, USA) using 150 bp paired reads generated in one lane of Illumina HiSeq X for each species. Raw data were assembled with SUPERNOVA V. 2.1.1 (Weisenfeld *et al.* 2017) and the final fasta file was generated using the 'pseudohap

185 style' option in *supernova mkoutput* using default settings. All commands used for this
186 work are available at https://github.com/macmanes-lab/museum_genomics.
187 We downloaded four publically available genome assemblies generated from a
188 single shotgun library sequenced on an Illumina platform. To minimize differences in
189 genome structure that could affect the performance of different methods, we selected
190 four mammalian species of similar genome size including *Tympanoctomys barrerae,*
191 *Octomys mimax* (Evans, Upham, Golding, Ojeda, & Ojeda, 2017), *Phodopus sungorus*
192 (Bao, Hazelerigg, Prendergast, & Stevenson, 2016), and *Tolypeutes matacus* (Johnson
193 et al., 2017).
194 To compare 10X versus shotgun-based assemblies, we assessed genome
195 quality through comparison of relative N50 values, genome completeness using
196 presence of BUSCO genes and of orthologous groups. Because the mammalian
197 genomes considered here are generally similar in size, N50 values are comparable and
198 normalization by genome size is not necessary. Comparative shotgun assemblies were
199 selected based on the number of total reads sequenced (~200M), as an equivalent
200 sequencing cost comparison against 10X Genomics. Read counts and assembly details
201 for each externally sourced genome are available in Table 2.
202 All genomes were annotated using MAKER V. 2.3.1 (Cantarel et al., 2008) using
203 the *Mus musculus* (GCF_000001635.26) reference proteome. The N50 statistic was
204 calculated via the *abyss-fa*c tool included in the ABYSS package (Simspon et al., 2009).
205 Benchmarking Universal Single-Copy Orthologs (BUSCO V. 3; Simão, Waterhouse,
206 Ioannidis, Kriventseva, & Zdobnov, 2015) statistics were used as metrics of genome

completeness based on gene content for genes conserved across Mammalia (mammalia_odb9).

We grouped genes from each species into orthogroups using ORTHOFINDER v. 2.3.3 (Emms & Kelly, 2015) and determined the number of orthogroups we could retrieve from each assembly. As proof of concept, a species tree was built for the *Peromyscus* taxa sequenced here, three publically available *Peromyscus* genomes (*P. maniculatus* [GCA_003704035], *P. leucopus* [GCF_004664715], and *P. polionotus* [GCA_003704135]), and four outgroup sequences (*Rattus norvegicus* [GCA_000001895.4], *Mus musculus* [GCF_000001635.26_GRCm38], *Onychomys torridus* [GCA_004026725], and *Sigmodon hispidus* [GCA_004025045] based on the ORTHOFINDER sets of orthologous genes, using IQTREE and default settings (Nguyen, Schmidt, von Haeseler, & Minh, 2015).

RESULTS

Read counts for each analyzed genome are available in Table 2. Additional assembly statistic (n:500, L50, N80, N50, N20, E-size, etc.) are available online at https://github.com/macmanes-lab/museum_genomics/blob/master/assembly_stats.md.

N50 values ranged from 2,626 to 39,982 bp with the highest values for 10X Genomics assemblies (36,160 bp on average) and lowest for single-lane Illumina assemblies (6,457 bp on average; Table 2). The number of genes annotated ranged from 3,233 (*P. sungorus*) to 19,008 (*P. attwateri*), with 10X Genomics assemblies 18,068 genes on average, compared to 10,900 genes in the average shotgun-based assembly. BUSCO measures of genome completeness ranged from 16.9% (*O. mimax)*

to 66.4% (*P. attwateri*) and were again highest for 10X Genomics assemblies (average: 61.6%) and lowest for shotgun assemblies (average: 27.3%; Table 2).

Annotations and predicted transcripts and proteins are available at http://doi.org/10.5281/zenodo.3351485. ORTHOFINDER identified 5,305 orthologous groups on average in shotgun-based assemblies and 9,112 on average in 10X assemblies ($p < 0.05$; Table 2). Our basic maximum-likelihood species tree resolved relationships with 100% bootstrap support (Fig. 1). Raw reads and assemblies are available through The European Nucleotide Archive (ENA) under project number PRJEB33530.

DISCUSSION

Linked-read sequencing facilitates the production of higher quality *de novo* genomes from historic samples, in less time, and with less effort than traditional shotgun based methods, providing a new option for accessing the genomes of aged samples. As such, linked-read sequencing may be the long-awaited key to unlocking the molecular secrets of NHCs and have applications across a broad range of evolutionary and ecological questions. *De novo* assemblies from linked-reads have greater contiguity and completeness relative to *de novo* assemblies based on shotgun libraries for comparable read volumes (Table 2). With the same sequencing effort (*e.g.,* 200 million 150 bp paired-end reads) linked-read sequencing results in a six-fold increase in N50 values and increases the number of represented genes by three times, without requiring a reference sequence from a closely related species. Using 10X assemblies as reference genomes to map population-level whole genome resequencing data, which are only

minimally affected by DNA degradation, will also increase the amount of variation, both sequence and structural, available for genotyping within and among species. Note that, once a *de novo* reference genome is available, shotgun libraries sequenced with short reads will still be a viable and cost-cost effective methodological choice for whole-genome resequencing. Linked-read methods facilitate the detection of rare alleles and enable haplotype phasing, both of which may be key to identifying emerging model taxa for biomedical research, investigations of rare genetic disorders in humans, and analyses of introgression, by enabling estimates of local ancestry (Tennessen et al., 2012; Janzen, Wang, & Hufford, 2019). Previously limited by technology, molecular investigations of museum specimens traditionally centered around systematic inquiry and phylogenetics. Now, the ability economically generate quality *de novo* assemblies for lower-quality tissue resources increases the power of these historic archives to address new questions.

Although the quality and completeness of linked-read assemblies are still dependent on DNA integrity, the application of linked-read methods may be especially impactful for rare or extinct species or when the collection of new material is difficult or impossible (Payne & Sorenson, 2002) due to the conservation status or geographic location (*e.g.,* international) of the target species. As new or higher-quality tissue samples will never again be available for extinct species, linked-reads offer an improved method for accessing data from preserved tissues of these species, even if the generation of perfectly contiguous genomes for these taxa is not attainable. In cases where the target species is highly divergent from available reference sequences, such as the case for extinct species or otherwise exceptionally divergent taxa (*e.g.,*

monotypic genera [*Ailurus, Eira*] or families [*Dugongidae, Orycteropodidae*]), *de novo* genome assemblies, rather than reference-based assemblies may provide more information.

As a caveat, 10X Genomics methods have not yet been tested for genomes larger than ~3Gb (e.g., human-sized), so although they are appropriate for many mammalian species, they may be less applicable to species with larger genomes. Detailed analysis of structural variation, as is often implicated in ecological adaptation (Wellenreuther, Mérot, Berdan, & Bernatchez, 2019), remains under the purview of long-read or hybrid (short and long reads) *de novo* sequencing methods and N50 statistics for linked-read assemblies are still limited relative to true long-read methods. Although the number of raw reads are variable within both groups — the shotgun-based and 10X Genomic assemblies — the number of reads is not correlated with genome quality. This leads us to conclude that differences in assembly quality are not driven by differences in sequencing depth. Finally, although our results are derived from lower quality frozen tissue samples, tissues remain unavailable for many pre-molecular era specimens. While linked-reads may be a solution to produce a *de novo* genome from poor quality tissues, this method has not been applied to and may not be appropriate for highly degraded museum study skins or destructively-sampled bone remains. Reduced-representation genetic approaches (Sanger sequencing, RADseq) or enriched sequencing methods (Bi et al., 2013; Staats et al., 2013; Jones & Good, 2016) may remain the most effective means of extracting data from more historic specimens in the absence of a closely related reference genome.

298 Ultimately, our results underscore the importance of continued scientific
299 collecting and the archival of personal legacy collections into NHCs into the future, as
300 new technologies will continue to improve our ability to extract molecular information
301 from degraded and aged samples. The centralization of biological resources and
302 associated information ensures the broad utility of these specimens to the scientific
303 community and facilitates tests, such as these, to determine the best available means of
304 extracting meaningful sequence data from lower quality DNA. In particular, we endorse
305 maximizing the utility of a specimen through the archival of multiple tissue types,
306 through multiple storage media (liquid nitrogen, ethanol, RNA later® [Sigma-Aldrich, St.
307 Louis, Missouri, USA], etc.) to maximize future uses of these archives as technologies
308 continue to evolve (Lessa et al., 2014; McLean et al., 2015).The ability to generate
309 WGS data from field-preserved tissues, further encourages the expansion of resurvey
310 projects (such as the NSF funded Grinnell Resurvey Project from the Museum of
311 Vertebrate Zoology, UC Berkeley) as a means for measuring change through time
312 (Moritz, Patton, Conroy, Parra, White, & Beissinger, 2008) and opens the possibility of
313 sequencing *de novo* genomes from now extinct species with preserved tissues
314 available. In an era of unprecedented ecological and environmental change (Ceballos,
315 Ehrlich, & Dirzo, 2017), genomic analyses of historic samples will help us understand
316 the evolutionary responses of natural populations to environmental perturbation and
317 hence lay the foundation for proactive management initiatives and predicting future
318 responses (Wandeler et al., 2007; Malaney & Cook, 2013).
319
320

ACKNOWLEDGEMENTS

We thank the Museum of Southwestern Biology for the *Peromyscus* tissue loans that made this work possible and the National Institute of Health (Grant Number: NIH R35GM128843) for funding.

DATA ACCESSIBILITY

All commands used for this work are available at https://github.com/macmanes-lab/museum_genomics. Raw reads and assemblies are available through The European Nucleotide Archive (ENA) under project number PRJEB33530, with assembly IDs: *P. attwateri* (ERZ1029326), *P. nudipes* (ERZ1029275)*, P. melanophrys* (ERZ1029325)*, and P. aztecus* (ERZ1029324).Assembly statistics are available online at https://github.com/macmanes-lab/museum_genomics/blob/master/assembly_stats.md. Annotations and predicted transcripts and proteins are available at http://doi.org/10.5281/zenodo.3351485.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this manuscript. MDM assembled the *Peromyscus* genomes.

TABLES & FIGURES

Figure 1. Maximum-likelihood phylogeny of examined *Peromyscus* species generated from consensus orthogroups, demonstrates complete resolution (100 bootstrap support for all nodes).

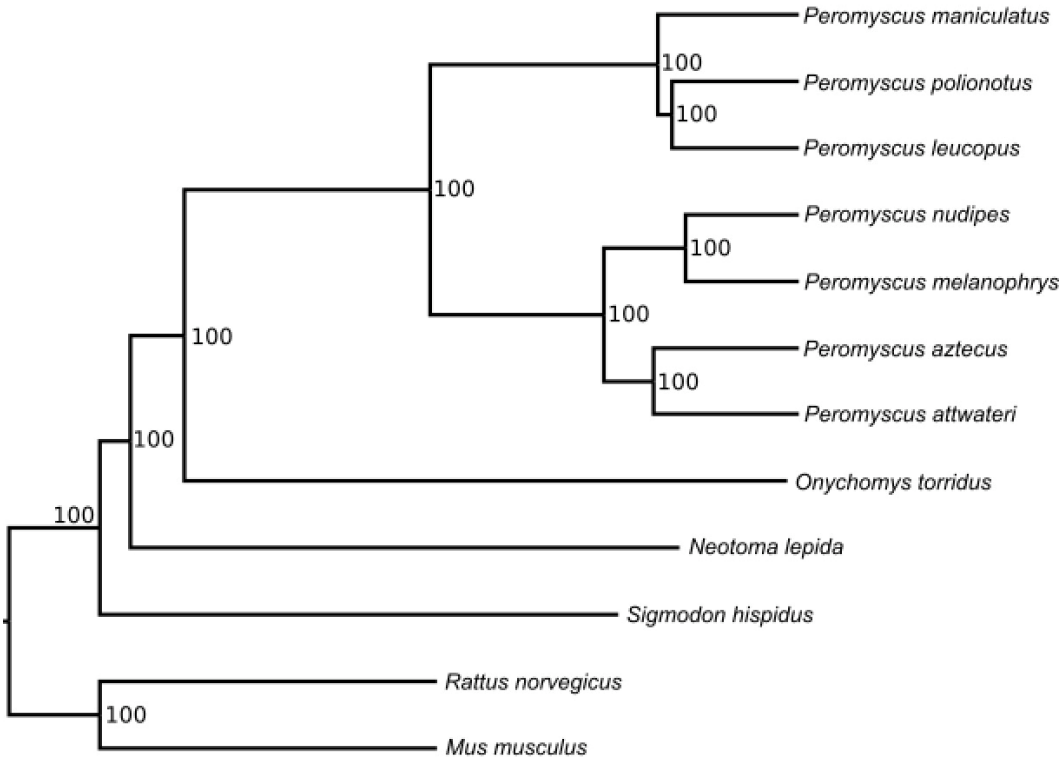

Table 1. Natural history data for specimens sequenced using 10X Genomics (*Peromyscus* spp.) and publically available single-land Illumina assemblies. Coll. = collection. Pub. = publication.

Common Name	Genus	Species	Coll. Year	Coll. Locality	Voucher	Pub.
Texas deer mouse	*Peromyscus*	*attwateri*	1995	Texas, USA	MSB:Mamm:84733	This Study
Aztec deer mouse	*Peromyscus*	*aztecus*	1982	Michoacan, Mexico	MSB:Mamm:48205	This Study
Plateau deer mouse	*Peromyscus*	*melanophrys*	2006	Coahuila, Mexico	MSB:Mamm:273915	This Study
La carpintera deer mouse	*Peromyscus*	*nudipes*	1995	Guanacaste, Costa Rica	MSB:Mamm:70743	This Study
Red vizcacha rat	*Tympanoctomys*	*barrerae*	na	Mendoza, Argentina	AO245	Evans et al. 2017
Mountain vizcacha rat	*Octomys*	*mimax*	na	San Juan, Argentina	AO248	Evans et al. 2017
Siberian hamster	*Phodopus*	*sungorus*	na	Laboratory	Unvouchered	Bao et al. 2016, Unpubl.
Three-banded Armadillo	*Tolypeutes*	*matacus*	na	na	Voucher not reported	Johnson et al. 2017, Unpubl.

Table 2. Sequencing and assembly quality statistics for each examined genome, including: Sequencing Platform (Seq. Platform), Assembler, Number of Reads (# of Reads [M = million], PE [paired-end], SE [single-end]), Contig N50 in base pairs (bp), longest contig (bp), percent (%) complete (C) BUSCOs, and the number of genes (# Genes) annotated.

Species	Seq. Platform	Assembler	# Reads (M)	Contig N50 (bp)	Longest contig (bp)	BUSCO (C:%)	Ortho-groups	# Genes
P. attwateri	10X Genomics	SuperNova	409M PE	39,982	283,668	66.4	9,560	19,008
P. aztecus	10X Genomics	SuperNova	423M PE	34,606	172,978	61.7	9,122	18,061
P. melanophrys	10X Genomics	SuperNova	405M PE	32,063	382,165	55.1	8,748	17,244
P. nudipes	10X Genomics	SuperNova	377M PE	37,990	239,464	63.1	9,188	17,960
T. barrerae	SL Illumina HiSeq	Abyss	342M PE (7M SE)	5,293	74,984	23	5,305	11,177
O. mimax	SL Illumina HiSeq	Abyss	168M PE (3M SE)	5,223	113,044	16.9	4,154	9,631
P. sungorus	SL Illumina HiSeq	SOAPdenovo	na	2,626	34,960	12.7	2,337	3,233
T. matacus	SL Illumina HiSeq	DISCOVAR denovo	na	12,685	251,506	42.1	7,171	19,557

References

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*, 61–65.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437, 1149–1152.

Bao, R., Hazelerigg, D., Prendergast, B., & Stevenson, T. J. (2016). The sequence and *de novo* assembly of the Siberian hamster genome (*Phodopus sungorus*). Unpublished, available online: https://www.ncbi.nlm.nih.gov/nuccore/MCBN000000000.1/.

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., & Moritz, C. (2013). Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, 22(23), 6018–6032.

Broad Institute. (2015). DISCOVAR: Assemble genomes, find variants. https://Www.broadinstitute.org/Software/Discovar/Blog.

Brooks, A., Turkarslan, S., Beer, K., Lo, F., & Baliga, N. (2011). Adaptation of cells to new environments. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(5), 544–561.

Cantarel, B., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1), 188–196.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2917). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, 114(30), E6089–E6096.

Dean, M. D., & Ballard, W. O. (2001). Factors affecting mitochondrial DNA quality from museum preserved *Drosophila simulans*. *Entomologia Experimentalis*, 98, 279–283.

Dessauer, H. C., Cole, C. J., & Hafner, M. S. (1990). Collection and storage of tissues. In Molecular Systematics (D.M. Hillis and C. Moritz (eds), pp. 29–47). Sunderland, MA: Sinauer Associates.

Edelman, N., Frandsen, P., Miyagi, M., Clavijo, B., Davey, J., Dikow, R., … Mallet, J. (2018). Genomic architecture and introgression shape a butterfly radiation. *BioRxiv*, 466292.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.

Evans, B., Upham, N., Golding, G., Ojeda, R., & Ojeda, A. (2017). Evolution of the largest mammalian genome. Genome Biology and Evolution, 9(6), 1711–1724.

Gnerre, S., Lander, E. S., Lindblad-Toh, K., & Jaffe, D. B. (2009). Assisted assembly: how to improve a *de novo* genome assembly by using related species. *Genome Biology*, *10*(8), R88.

Goodwin, S., McPherson, J. D., & McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews: Genetics*, 17(6), 333.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328, 710–722.

Jackson, B. C., Campos, J. L., & Zeng, K. (2015). The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. *Heredity*, 114(2), 163–174.

Janzen, G. M., Wang, L., & Hufford, M. B. (2019). The extent of adaptive wild introgression in crops. *New Phytologist*, 221, 1279–1288.

Jiao, W.-B., & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36, 64–70.

Johnson, J., Muren, E., Swofford, R., Turner-Maier, J., Marinescue, V. D., Genereux, D. P., … Lindblad-Toh, K. (2018). The 200 mammals project: sequencing genomes by a novel cost-effective method, yielding a high-resolution annotation of the human genome. Unpublished, available at: https://www.ncbi.nlm.nih.gov/nuccore/PVIB000000000.1/.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1), 185–202.

Jones, S. J., Haulena, M., Taylor, G. A., Chan, S., Bilobram, S., Warren, R. L., … Jones, S. (2017). The genome of the northern sea otter (*Enhydra lutris kenyoni*). *Genes*, *8*(12), E379.

Jones, S., Taylor, G., Chan, S., Warren, R., Hammond, S., Bilobram, S., … Haulena, M. (2017). The genome of the beluga whale (*Delphinapterus leucas*). *Genes*, *8*(12), E378.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., … Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nature Biotechnology*, 32(3), 261.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., … Schatz, M. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*, 48603.

Lessa, E. P., Cook, J. A., D'Elia, G., & Opazo, J. C. (2014). Rodent diversity in South America: transitioning into the genomics era. Frontiers in Ecology and Evolution, 2, 1–7.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362, 709–715.

Lynch, M. (1989). Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. *Evolution*, 43(1), 1–17.

Malaney, J. L., & Cook, J. A. (2013). Using biogeographical history to inform conservation: the case of Preble's meadow jumping mouse. *Molecular Ecology*, 22(24), 6000–6017.

McCoy, R., Taylor, R., Blauwkamp, T., Kelley, J., Kertesz, M., Pushkarev, D., … Fiston-Lavier, A. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One*, 9, e106689.

McLean, B. S., Bell, K. C., Dunnum, J. L., Abrahamson, B., Colella, J. P., Deardorff, E. R., … Cook, J. A. (2016). Natural history collections-based research: progress, promise, and best practices. *Journal of Mammalogy*, 97(1), 287–297.

Meyer, M., Fu, Q., Aximu-Petri, A., Glocke, I., Nickel, B., Arsuaga, J.-L., … Pääbo, S. (2014). A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, 505(7483), 403.

Meyer, M., Kircher, M., Gansauge, M., Li, H., Racimo, F., Mallick, S., … Sudmant, P. H. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104), 222–226.

Moritz, C., Patton, J., Conroy, C., Parra, J., White, G., & Beissinger, S. (2008). Impact of century of climate change on small-mammal communities in Yosemite National Park. USA *Science*, 322, 258–261.

Nachman, M. W. (2013). Genomics and museum specimens. *Molecular Ecology*, 22(24), 5966–5968.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews: Genetics*, *14*, 157–167.

Nei, M., & Tateno, Y. (1975). Interlocus variation of genetic distance and the neutral mutation theory. *Proceedings of the National Academy of Sciences*, 72(7), 2758–2760.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.

Opazo, J. C., Palma, R. E., Melo, F., & Lessa, E. (2005). Adaptive evolution of the insulin gene in Caviomorph rodents. *Molecular Biology and Evolution*, *22*(5), 1290–1298.

Payne, R. B., & Sorenson, M. D. (2002). Museum collections as sources of genetic data. *Bonn Zoological Bulletin*, 51, 97–104.

Pélissié, B., Crossley, M., Cohen, Z., & Schoville, S. (2018). Rapid evolution in insect pests: the importance of space and time in population genomics studies. *Current Opinion in Insect Science*, *26*, 8–16.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354–366.

Rowe, K. C., Singhal, S., MacManes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., … Moritz, C. C. (2011). Museum genomics: low-cost and hihg-accuracy genetic data from historical specimens. *Molecular Ecology Resources*, 11, 1082–1092.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 32-10–3212.

Simpson, J., Wong, K., Jackman, S., Schein, J., Jones, S., & Birol, I. (2009). ABySS: a parallel assembled for short read sequence data. *Genome Research*, 19(6), 1117–1123.

Staats, M., Erkens, R., van de Vossenberg, B., Wieringa, J., Kraaijeveld, K., Stielow, B., … Bakker, F. (2013). Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLOS ONE*, 8(7), e69189.

Suarez, A., & Tsutsiu, N. (2004). The value of museum collections for research and society. *BioScience*, 54(1), 66–74.

Teeling, E. C., & Hedges, S. B. (2013). Making the impossible possible: rooting the tree of placental mammals. *Molecular Biology and Evolution*, 30, 1999–2000.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E., Gravel, S., … Akey, J. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090), 64–69.

van Dijk, E., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, 34(9), 666–681.

Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., … Quake, S. (2013). The genome sequence of the colonial chordate, *Botrylius schlosseri*. *eLife*, 2, e00569.

Wandeler, P., Hoeck, P. E., & Keller, L. F. (2007). Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution*, 22(12), 634–642.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, 27, 757–767.

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203–1209.

Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, 272, 3–16.

Wood, H. M., González, V., Lloyd, M., Coddington, J., & Scharff, N. (2018). Next-generation museum genomics: Phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palpimanoidea). *Molecular Phylogenetics and Evolution*, 127, 907–918.

Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., … Liu, Z.-J. (2017). The *Apostasia* genome and the evolution of orchids. *Nature*, *549*, 379–383.

Zheng, G. X. Y., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C., Kyriazopoulou-Panagiotopoulou, S., … Ji, H. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing*. Nature Biotechnology*, 34(3), 303–311.