Received: 7 August 2019 | Revised: 3 March 2020 | Accepted: 6 March 2020

DOI: 10.1111/1755-0998.13155

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues

Jocelyn P. Colella[1,2] | Anna Tigano[1,2] | Matthew D. MacManes[1,2]

[1]Molecular, Cellular, and Biomedical Sciences Department, University of New Hampshire, Durham, NH, USA

[2]Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, USA

Correspondence
Jocelyn P. Colella, Molecular, Cellular, and Biomedical Sciences Department, University of New Hampshire, Durham, NH, USA.
Email: Jocelyn.Colella@unh.edu

Funding information
National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R35GM128843

Abstract

High-throughput sequencing technologies are a proposed solution for accessing the molecular data in historical specimens. However, degraded DNA combined with the computational demands of short-read assemblies has posed significant laboratory and bioinformatics challenges for de novo genome assembly. Linked-read or "synthetic long-read" sequencing technologies, such as 10× Genomics, may provide a cost-effective alternative solution to assemble higher quality de novo genomes from degraded tissue samples. Here, we compare assembly quality (e.g., genome contiguity and completeness, presence of orthogroups) between four new deer mouse (*Peromyscus* spp.) genomes assembled using linked-read technology and four published genomes assembled from a single shotgun library. At a similar price-point, these approaches produce vastly different assemblies, with linked-read assemblies having overall higher contiguity and completeness, measured by larger N50 values and greater number of genes assembled, respectively. As a proof-of-concept, we used annotated genes from the four *Peromyscus* linked-read assemblies and eight additional rodent taxa to generate a phylogeny, which reconstructed the expected relationships among species with 100% support. Although not without caveats, our results suggest that linked-read sequencing approaches are a viable option to build de novo genomes from degraded tissues, which may prove particularly valuable for taxa that are extinct, rare or difficult to collect.

KEYWORDS

10× genomics, assembly quality, natural history collections, *Peromyscus*

1 | INTRODUCTION

A disconnect between the capabilities of high-throughput sequencing technologies and the quality, or lack thereof, of historical museum specimens has proven a major barrier to accessing molecular data from degraded samples. Natural history collections (NHCs) store a wide variety of species from across the globe, including those that are now difficult to collect or are extinct in the wild. Voucher specimens housed in NHCs are an invaluable source of morphological material as they provide a reference for measuring change across both space and time. Specimens have also been recognized as important repositories of genetic sequence data (Holmes et al., 2016; Payne & Sorenson, 2002; Wandeler, Hoeck, & Keller, 2007) and have provided insight into the phylogenetic relationships and origins of species (McLean et al., 2016; Suarez & Tsutsiu, 2004). Quick progress in genomics methods has significantly expanded our ability to interrogate museum archives molecularly (Holmes et al., 2016) and enabled the generation of genomic-level data sets despite the DNA degradation that characterizes many of these historical samples. "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), is increasingly resolving the phylogenetic Tree of Life (Lessa, Cook, D'Elia, & Opazo, 2014; Teeling & Hedges, 2013; Wood, González, Lloyd, Coddington, & Scharff, 2018) and has myriad expanded applications, including identifying functional variants implicated in ecological adaptations (Opazo, Palma, Melo, & Lessa, 2005), genome sequencing of extinct species (Feigin et al., 2018; Green et al., 2010), and estimating mutation rates and the timing of evolutionary events (Pélissié, Crossley, Cohen, & Schoville, 2018).

Over time, DNA degrades into short fragments through exposure to UV light, temperature, pH, salt, etc. (Dean & Ballard, 2001; Dessauer, Cole, & Hafner, 1990; Lindahl, 1993; Willerslev & Cooper, 2005), which can complicate the application of genomic methods to museum specimens. Since the 1970s, when museums began archiving tissues, collection and preservation methods have varied widely. Preservation methods generally evolved to accommodate changing analytical technologies, resulting in the variety of preservation media (formalin, ethanol, frozen, etc.) and tissue qualities 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.

To retrieve genetic information from degraded samples, exome capture (Bi et al., 2013) and other reduced-representation approaches (Targeted capture, Jones & Good, 2016) have been the recommended approaches. However, in addition to complex laboratory work, these approaches retrieve only a subset of the genome, leaving out potentially useful information for understanding the different genomic targets (e.g., coding and noncoding, transposable elements, structural variation) of adaptive evolution and speciation (Andolfatto, 2005; Brooks, Turkarslan, Beer, Lo, & Baliga, 2011; Mack & Nachman, 2017). Additionally, targeted capture approaches such as exome-capture can introduce biases due to the potentially confounding effect of purifying selection on exonic coding regions (Jackson, Campos, & Zeng, 2015) and have limited ability to parse paralogous gene families and variation in gene copy numbers among species (Fromer et al., 2012; Mandelker et al., 2016). The generation of a de novo genome assembly is valuable to maximize the retrieval of genetic information from a single specimen, to avoid some of the biases or limitations inherent in reduced-representation approaches, and to build a reference sequence for whole genome resequencing, and is increasingly seen as a viable option for even moderately to highly degraded samples (Zhang, Lehmann, Shyr, & Guo, 2017).

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 repetitive regions that extend beyond the sequenced read length (Alkan, Sajjadian, & Eichler, 2011; Nagarajan & Pop, 2013). 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. The ability to resolve sequencing artefacts in assembly improves with increasing read length, by leveraging long-range information present in intact DNA and/or long-reads that span areas that are difficult to assemble. One of these methods, long-read sequencing (>10–50 kb), such as Pacific Biosciences and Oxford Nanopore Technologies, has addressed some computational complexities of de novo genome assembly, but require large quantities of HMW DNA for library preparation, making these methods inaccessible for degraded tissue samples (Rowe et al., 2011). Before the development of long-read sequencing, the most common approach to de novo genome assembly involved a combination of shotgun short-insert (<500-bp) and mate-pair long-insert (>2,000-bp) libraries, where the first would be used for assembly and the second for scaffolding. 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)—based on single short-insert shotgun libraries sequenced to ~60× 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 (with genome sizes <400 Mb, Edelman et al., 2019), but is significantly more expensive than other approaches for organisms with larger genomes 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). With this approach, 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 as a scaffolding guide only (Gnerre, Lander, Lindblad-Toh, & Jaffe, 2009). While reference-guided assembly offers a partial solution, high-quality, closely related references—a prerequisite for this approach—are not yet available for a large number of ecologically relevant taxa.

In the grey area between short- (up to 250-bp paired-end) and long-read (>10s of kb) 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 tissue samples. These methods allow the assembly of pseudo-long reads up to tens of kb from short-read data, and with higher accuracy compared to true long-read sequencing (Jiao & Schneeberger, 2017). Initially introduced by Illumina (Kuleshov et al., 2014; McCoy et al., 2014), there are few but increasing numbers of publications using linked-read methods, particularly in the field of museomics (Etherington et al., 2019; Latorre et al., 2020; Lutgen et al., 2020). 10× 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. 10× 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 derive from the same original DNA fragment (van Dijk, Jaszczyszyn, Naquin, & Thermes, 2018; Goodwin, McPherson, & McCombie, 2016). This method therefore increases assembly confidence and contiguity by geographically linking smallreads in genome space. 10× Genomics requires low input material (<2 ng) and although HMW DNA is optimal for any method, the barcodes attached to unique DNA fragments maximize the limited long-range information contained in degraded DNA. 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). Thanks to the low input requirements of 10× Genomics, DNA purification techniques that eliminate lowmolecularweight (LMW) fragments can be useful for selecting only the longest fragments for assembly, as these contain the long-range information necessary for high-quality genome assembly. While eliminating LMW DNA fragments further reduces the amount of DNA available, newer innovations such as the Short Read Eliminator Kit (Circulomics) offer promising solutions to minimize DNA loss.

While 10× Genomics protocols are currently optimized for human genomes and are most often applied to cancer and biomedical research (Zheng et al., 2016), linked-read methods are easily extended to other mammals (sea otters and beluga whales, Jones, et al., 2017a, 2017b), which are expected to have similar genome sizes and structure (proportion of repeats, GC content, etc.). Linked-read methods have also been successfully used for genome assembly in other taxa, including plants, insects and birds (e.g., orchids, Zhang et al., 2017; ladybird beetle, Ando et al., 2018; Gouldian finch, Toomey et al., 2018). As a proof-of-concept, we compare contiguity and completeness of four deer mouse (*Peromyscus*) genome assemblies derived from low-quality historical tissue samples collected from 1982 to 2006, produced using the 10× Genomics linked-read approach (hereafter 10×). Second, we contrast these 10× assemblies against four publicly available shotgun Illumina mammalian genome assemblies generated at an equivalent cost and using similar read volumes. Third, we use the 10× assemblies to reconstruct phylogenetic relationships between the newly assembled *Peromyscus* genomes and eight additional mammal species. We demonstrate the utility of this economical approach for de novo genome assembly from degraded tissue samples for researchers interested in diverse questions related to diversification and adaptive evolution, but limited by sample quality for their species of interest.

2 | MATERIALS AND METHODS

Twenty-five micrograms of frozen liver tissue from each of four field-collected *Peromyscus* museum specimens (*P. attwateri*, *P. aztecus*, *P. melanophrys*, *P. nudipes*) were loaned from the Museum of Southwestern Biology (MSB; Table 1). Three of the specimens were collected internationally and collection dates ranged from 1982 to 2006 (Table 1). For comparison, a high-quality tissue subsample from *Peromyscus boylii*, collected in 2017, and a moderate-quality tissue subsample from *Peromyscus pectoralis* loaned from the MSB and collected in 1997 were processed in parallel. Genomic DNA was extracted using a standard Qiagen Genomic Tip protocol. DNA was quantified with a Qubit fluorometer (Thermofisher Scientific) and quality was assessed using an Agilent TapeStation. As fragment size distribution greatly influences the contiguity of the genome assembly, the samples were further processed using

TABLE 1 Natural history data for specimens sequenced using 10× Genomics (*Peromyscus* spp.) and for publicly available, shotgun assemblies used for comparison

Common name	Genus	Species	Collection. year	Collection locality	Voucher	Publication
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, Upham, Golding, Ojeda, and Ojeda (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. (2018), Unpubl.

the Circulomics short read eliminator kit, which removes DNA molecules shorter than 10 kb, and progressively up to 25 kb. The size-selected DNA from each of these samples was sent to the Genomics Core Facility at Icahn School of Medicine at Mount Sinai (New York), where samples were first run on a Femto Pulse (Agilent) to assess fragment size distribution post-Circulomics and then used for library preparation. The resulting 10× libraries were sequenced at Novogene using 150-bp paired reads generated in one lane of Illumina HiSeq X for each species. Raw 10× data were assembled with SUPERNOVA version 2.1.1 (Weisenfeld et al., 2017) and the final fasta file was generated using the "pseudohap style" option in SUPERNOVA MKOUTPUT using default settings. All commands used for this work are available at https://github.com/macmanes-lab/museum_genomics.

To compare 10× and shotgun approaches, four publicly available genome assemblies generated from a single shotgun library and sequenced on an Illumina platform were selected. To minimize differences in genome size and structure that could confound the comparison between these two approaches, the four species were mammals with similar genome size: *Tympanoctomys barrerae*, *Octomys mimax* (Evans, Upham, Golding, Ojeda, & Ojeda, 2017), *Phodopus sungorus* (Bao, Hazelerigg, Prendergast, & Stevenson, 2016) and *Tolypeutes matacus* (Gibb et al., 2016; Table 1). Comparative shotgun assemblies were also selected based on the number of total reads sequenced (~200 million paired-end reads) for a comparison against 10× assemblies based on equivalent sequencing costs. Read counts and assembly details for each externally sourced genome are available in Table 2.

Genome quality was assessed through comparison of N50 values for contiguity and presence of BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) genes for completeness. The N50 statistic was calculated using the *assemblathon_stats.pl* script available at: https://github.com/KorfLab/Assemblathon. Because the mammalian genomes considered here are generally similar in size, N50 values are comparable and normalization by genome size is not necessary. BUSCO version 3 (Simão et al., 2015) statistics were used as metrics of genome completeness based on presence of genes conserved across Mammalia (mammalia_odb9).

All genomes were annotated using MAKER version 2.3.1 (Cantarel et al., 2008) using the *Mus musculus* (GCF_000001635.26) reference proteome. ORTHOFINDER version 2.3.3 (Emms & Kelly, 2015) was used to identify orthogrups (e.g., groups of genes with a shared evolutionary history) from the annotated genes in each genome. To demonstrate the phylogenetic utility of genome assemblies produced with 10×, a consensus species tree was built from shared orthogroups among the four *Peromyscus* sequenced here, plus three publicly available *Peromyscus* genomes (*P. maniculatus* [GCA_003704035], *P. leucopus* [GCF_004664715] and *P. polionotus* [GCA_003704135]), and five additional outgroup genomes (*Rattus norvegicus* [GCA_000001895.4], *Mus musculus* [GCF_000001635.26_GRCm38], *Onychomys torridus* [GCA_004026725], *Neotoma lepida* [GCA_001675575.1] and *Sigmodon hispidus* [GCA_004025045]). For consistency, all genomes

TABLE 2 Qubit concentrations of original DNA extractions, DNA quantifications before and after size selection with Circulomics (short-read eliminator kit), and the percentage (%) of DNA lost Peak sizes before and after size selection

Species	Collection year	Qubit concentration (ng/µl)	DNA quantity pre-Circulomics	DNA quantity (µg) post-Circulomics	Percentage lost	Peak size (kb) pre-Circulomics[a]	Peak size (kb) post-Circulomics[a]	Peak size (kb) post-Circulomics[b]
P. attwateri	1995	390	10	0.391	96	24	30	16
P. aztecus	1982	370	10	0.755	92	12	18	11
P. melanophrys	2006	191	10	4.2	58	—	—	9
P. nudipes	1995	486	10	0.115	99	27	27	10
P. pectoralis	1997	282	10	2.4	76	26	26	—
P. boylii	2017	208	10	7.1	29	>60	>60	—

[a] TapeStation measurement.

[b] Femto Pulse measurement.

were re-annotated using the approach described above and the orthogroups were identified de novo excluding the low-quality shotgun genome assemblies used for methodological comparison. The phylogeny, based on sets of orthogroups shared across the 12 Rodentia taxa, was generated using default settings in IQTREE, with the best-fitting substitution model autodetected (Nguyen, Schmidt, von Haeseler, & Minh, 2015). As genome assembly quality varied greatly across the 12 species included in the phylogeny, the number of shared orthogroups between pairs of taxa was visualized using the R package PLOTLY (Plotly Technologies Inc, 2015). Finally, it should be noted that a variety of programs were used for genome assembly (Table 3), with each being tuned to properly assemble a specific type of sequence data. Specifically, DISCOVAR DENOVO (Broad Institute, 2015; Weisenfeld et al., 2017) was developed to assemble 250-bp paired-end PCR-free reads, while SOAPDENOVO (Luo et al., 2012) was developed to assemble and scaffold short-insert and mate-pair reads. SUPERNOVA (Weisenfeld et al., 2017) is currently the only fully benchmarked assembler developed for use with sequencing libraries produced on the 10× platform. Therefore, the relative performance of these assembly algorithms cannot be tested here due to their data type specificity.

3 | RESULTS

Concentrations of DNA extracted from the six *Peromyscus* tissue samples ranged from 191 to 486 ng/µl (Table 2). DNA was moderately degraded for five of the *Peromyscus* samples examined, albeit with differences among species. Distributions of DNA fragments for all except the high-quality *P. boylii* sample peaked at molecular weights well below the recommended 50 kb, and as low as 12 kb. Circulomics size selection removed the vast majority of DNA from degraded samples. More than 90% of the DNA was lost in three of the four low-quality samples, while only 29% was lost from the highest quality sample (Table 2). Peak fragment size increased in two of three cases, by 25% and 50% in *P. attwateri* and *P. aztecus*, respectively. Across low-quality tissue samples, the least DNA was lost from *P. melanophrys*, which was also the most recently collected. Although estimation of fragment size distribution after size selection differed between TapeStation and Femto Pulse, they concordantly indicated that fragment size distribution peaked at the largest size in *P. attwateri*. Femto Pulse data for the samples before size selection or TapeStation data for *P. melanophrys* before or after size selection were not collected. The number of input reads for each assembled genome is available in Table 3. Additional assembly statistic (n:500, L50, N80, N20, E-size, etc.) are available online at https://github.com/macmanes-lab/museum_genomics/blob/master/assembly_stats.md.

N50 values for the four 10× assemblies ranged from 32,396 to 40,046 bp (36,376 bp on average), compared to lower N50 values for shotgun assemblies (range: 2,392–10,217 bp; average: 5,545 bp; Table 3). Among 10× and comparative shotgun assemblies, the number of genes annotated ranged from 3,233 (*P. sungorus*) to 19,008 (*P. attwateri*), with 10× assemblies containing 18,068 genes on average, 66% more than in shotgun assemblies on average (10,900). BUSCO measures of genome completeness ranged from 12.7% (*P. sungorus*) to 66.4% (*P. attwateri*) and were again highest for 10× assemblies (average: 61.6%) and lowest for shotgun assemblies (average: 23.7%; Table 3). Among *Peromyscus*, *P. attwateri* peaked at the largest fragment size and its assembly had the highest contig N50, completeness, number of orthogroups and genes, demonstrating the importance of fragment size for genome assembly quality. Annotations and predicted transcripts and proteins are available at http://doi.org/10.5281/zenodo.3351485. ORTHOFINDER identified significantly fewer orthogroups in shotgun assemblies than in 10× assemblies (5,305 vs. 9,112 on average, respectively; Table 2; Figure 1).

When analysed with other higher-quality genome assemblies ("Phylogenetic outgroup taxa," Table 3), the number of orthogroups retrieved in the four 10× *Peromyscus* assemblies was 48% higher on average than when analysed with the shotgun assemblies only (Table 3; Figure 1). Among species included in the phylogenomic analysis, *P. polionotus* had the highest total number of orthogroups identified (15,667) and *P. melanophrys* had the lowest (13,080, Table 3; Figure 1). In total, 9,254–14,265 orthogroups were shared between pairs of taxa (Table 3; Figure 1). The number of orthogroups shared between taxa was highest between higher-quality assemblies and closer phylogenetic relationships (Figure 1). However, the number of orthogroups shared between *Peromyscus* was higher when one high-quality assembly was included in the comparison and highest when both assemblies were of high quality, suggesting that assembly quality has a greater impact than phylogenetic distance in the identification of shared orthogroups, at least at the genus level. The maximum-likelihood species tree resolved relationships with 100% bootstrap support (Figure 1) and is consistent with previous phylogenetic investigations of *Peromyscus* (Bradley et al., 2007). Raw reads and assemblies are available through The European Nucleotide Archive (ENA) under project number PRJEB33530.

4 | DISCUSSION

Linked-read sequencing facilitates the production of higher quality de novo genomes from historical tissue samples, in less time and with less effort than traditional shotgun-based methods. As such, linked-read sequencing shows great potential for maximizing the retrieval of genomic information from NHCs and to enable the investigation of a broad range of evolutionary and ecological questions at greater resolution. Despite the loss of more than 90% of the DNA to size selection (Table 2), 10× assemblies had greater contiguity and completeness relative to de novo assemblies based on shotgun libraries and comparable read volumes (Table 3). Although the number of raw reads varies within both groups—shotgun-based and 10× assemblies—it is not correlated with genome quality, indicating that differences in assembly quality are not driven by differences in sequencing depth. With the same sequencing effort (e.g., 200 million, 150-bp paired-end reads), linked-read sequencing resulted in a

TABLE 3 Sequencing and assembly quality statistics for each examined genome, including: sequencing platform, assembler, number of reads PE [paired-end], SE [single-end]), scaffold N50, longest contig, percentage complete BUSCOs, and the number of genes annotated

Species	Library preparation	Sequencing platform	Assembler(s)	Number of reads (million)	Scaffold N50 (bp)	Longest scaffold (bp)	BUSCO (%)	Orthogroups	Number of genes
10× Assemblies									
P. attwateri	10× Genomics	Illumina Hiseq X	SUPERNOVA	409 PE	40,046	284,598	66.4	14,035 [9,560]	19,008
P. aztecus	10× Genomics	Illumina Hiseq X	SUPERNOVA	423 PE	34,920	176,898	61.7	13,525 [9,122]	18,061
P. melanophrys	10× Genomics	Illumina Hiseq X	SUPERNOVA	405 PE	32,296	386,405	55.1	13,080 [8,748]	17,244
P. nudipes	10× Genomics	Illumina Hiseq X	SUPERNOVA	377 PE	38,243	240,034	63.1	13,533 [9,188]	17,960
Comparative shotgun Assemblies									
T. barrerae	Shotgun	SL Illumina HiSeq	ABYSS	342 PE (7 SE)	4,698	113,146	23	[5,305]	11,177
O. mimax	Shotgun	SL Illumina HiSeq	ABYSS	168 PE (3 SE)	4,874	75,086	16.9	[4,154]	9,631
P. sungorus	Shotgun	SL Illumina HiSeq	SOAP DENOVO	na	2,392	34,960	12.7	[2,337]	3,233
T. matacus	Shotgun	SL Illumina HiSeq	DISCOVAR DENOVO	na	10,217	251,906	42.1	[7,171]	19,557
Phylogenetic outgroup taxa									
P. maniculatus	na	Illumina HiSeq	na	na	115,033,041	193,310,054	95.3	15,593	20,060
P. polionotus	na	Illumina	na	na	117,603,569	213,001,178	95.4	15,667	19,768
P. leucopus	Shotgun + SMRTbell	PacBio, Illumina HiSeq, Hi-C	Platanus[a], Falcon[b], Quiver[c], Pilon[d]	794M PE	114,273,790	193,658,164	94.5	15,147	19,740
N. lepida	Shotgun	Illumina HiSeq	ALLPATHS	674M PE	151,693	4,085,094	78	13,322	17,282
O. torridus	Shotgun	Illumina HiSeq	DISCOVAR DENOVO	na	20,878	372,325	63.4	13,905	18,310
S. hispidus	Shotgun	Illumina HiSeq	DISCOVAR DENOVO	na	101,373	1,032,205	86.6	14,351	24,474
R. norvegicus	Shotgun + BAC	Sanger, SOLiD, PacBio	ATLAS	na	145,729,302	282,763,074	91.6	14,257	27,491
M. musculus	Shotgun + BAC	na	ATLAS	>44M	130,694,993	195,471,971	95.2	13,883	21,907

Notes: SL, single lane. Number of orthogroups identified among the 12 species included in phylogenetic analysis and, in square brackets, the number of orthogroups identified between 10× and shotgun assemblies.

[a] Kajitani et al. (2014).

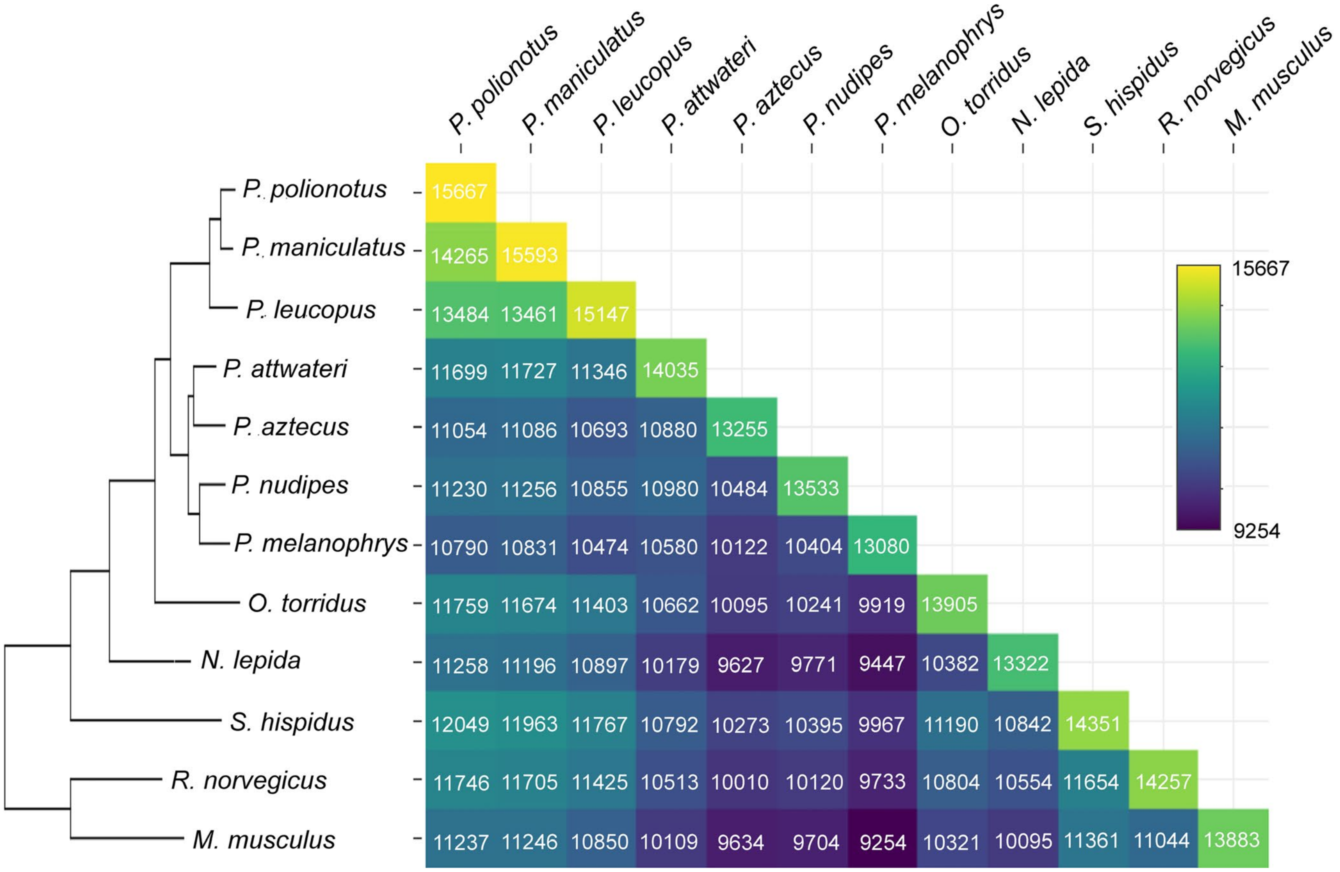
[b] Chin et al. (2016).

[c] Chin et al. (2013).

[d] Walker et al. (2014).

[e] Butler et al. (2008).

[f] Atlas Assembly Suite, Havlak et al. (2004).



	P. polionotus	P. maniculatus	P. leucopus	P. attwateri	P. aztecus	P. nudipes	P. melanophrys	O. torridus	N. lepida	S. hispidus	R. norvegicus	M. musculus
P. polionotus	15667											
P. maniculatus	14265	15593										
P. leucopus	13484	13461	15147									
P. attwateri	11699	11727	11346	14035								
P. aztecus	11054	11086	10693	10880	13255							
P. nudipes	11230	11256	10855	10980	10484	13533						
P. melanophrys	10790	10831	10474	10580	10122	10404	13080					
O. torridus	11759	11674	11403	10662	10095	10241	9919	13905				
N. lepida	11258	11196	10897	10179	9627	9771	9447	10382	13322			
S. hispidus	12049	11963	11767	10792	10273	10395	9967	11190	10842	14351		
R. norvegicus	11746	11705	11425	10513	10010	10120	9733	10804	10554	11654	14257	
M. musculus	11237	11246	10850	10109	9634	9704	9254	10321	10095	11361	11044	13883

FIGURE 1 Maximum-likelihood phylogeny of the four new linked-read *Peromyscus* genome assemblies, three publicly available *Peromyscus* assemblies and five outgroup assemblies within Rodentia. The phylogeny, generated from consensus orthogroups, demonstrates complete resolution (100% bootstrap support for all nodes). The heat map details the number of shared orthogroups across taxa, with the diagonal indicating the total number of orthogroups identified for each species [Colour figure can be viewed at wileyonlinelibrary.com]

six-fold increase in N50 values and doubled the number of assembled genes, without requiring a reference sequence from a closely related species. 10× assemblies allowed us to retrieve more than 18,000 genes on average and up to 66% of complete BUSCOs, proving an affordable and fast method to retrieve genomic sequence data from degraded samples for use in phylogenetic analyses and other genome-scale investigations. Additionally, these and other de novo assemblies can be used as a reference genome for mapping population-level whole genome resequencing data, which can be obtained even from moderately to highly degraded samples. For population-level investigations based on temporal sampling or for taxa that are at-risk, elusive or otherwise cannot be resampled for a variety of reasons, an improved reference genome will increase the amount of variation, both sequence and structural, available for genotyping within and among species.

Often limited by technology, molecular investigations of museum specimens were traditionally centred around systematic inquiry and phylogenetics (Holmes et al., 2016). Following broad application of single- and multilocus investigations (Hickerson et al., 2010), high-throughput sequencing methods have further increased our ability to interrogate historical archives molecularly and to address a variety of evolutionary questions at increased resolution. The growing field of museomics (Bi et al., 2013; Schmitt, Cook, Zamudio, & Edwards, 2018) continues to explore the tree of life (Guschanski et al., 2013; Teeling & Hedges, 2013), elucidate early human histories (Green et al., 2010; Meyer et al., 2012) and identify evolutionary adaptations of nonmodel species (Cheviron & Brumfeld, 2012; Jones, Mills, Jensen, & Good, 2019; Tigano, Colella, & MacManes, 2020). However, to date, the majority of genome-scale specimen investigations have relied on reduced-representation sequencing approaches or standard Illumina shotgun deep sequencing. As increasing sequencing depth does not necessarily increase genome contiguity, a linked-read approach can facilitate de novo genome assembly or enhance the quality of existing lower-quality assemblies with low to moderate coverage by adding long-range information (depending on starting DNA quality; Yeo, Coombe, Warren, Chu, & Birol, 2018).

Although the contiguity and completeness of linked-read assemblies depend on DNA integrity, linked-read methods may be especially useful to build genome assemblies 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 geographical 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 may be the best currently available option for accessing data from preserved tissues of these species, even if the generation of highly contiguous genomes for these taxa may not be attainable. For example, the genome of the now-extinct thylacine or Tasmanian tiger (*Thylacinus cynocephalus*) was recently assembled from shotgun sequencing data generated from a 108-year-old, alcohol-preserved specimen stored at the Museums Victoria in Australia (Feigin et al., 2018). The resulting 3.2-kb N50 is equivalent to the shotgun-based mammalian genome assemblies examined here (Table 3), suggesting that linked-read methods could help to improve the contiguity of the thylacine assembly and expand the available sequence for inquiry into this extinct species and related taxa. De novo genome assemblies may provide more information than reference-based assemblies, especially when the target species is highly divergent from available reference sequences, as is often the case for extinct species or otherwise exceptionally divergent taxa (e.g., monotypic genera [*Ailurus*, *Eira*] or families [Dugongidae, Orycteropodidae]).

Linked-reads present some caveats. For example, these methods are not optimized for genomes larger than ~3 Gb (e.g., human-sized). Although linked-read methods are generally appropriate for mammalian species and have been tested across a wide range of genome sizes (140 Mb to 3.2 Gb), their application to other species with much larger genome sizes (>4 Gb) is not supported and remains to be explored in greater detail. Linked-reads alone cannot produce chromosome-level assemblies, even with the best samples as starting material. However, when DNA quality and quantity allow, linked-read assemblies can be improved with other types of data, including long-reads, Hi-C data, and linkage and optical mapping. The higher quality assemblies obtainable with these hybrid sequencing and assembly approaches are necessary for analyses of structural variation or genotype–phenotype associations, which require high genome completeness. Finally, low-quality frozen tissue samples, similar to those used here, are often unavailable for many premolecular era specimens, and linked-reads have not been tested, and may or may not be appropriate, for extracting quality sequence data from more degraded samples such as skin, bone or hair. However, the ability to generate quality de novo genomes from relatively recently collected species is invaluable if these are the only samples available for a specific taxon, and can still provide important information on the time these samples were collected and evolutionary change since, especially for organisms with short generation times. Targeted capture approaches or enriched sequencing methods (Bi et al., 2013; Jones & Good, 2016; McCormack et al., 2012; Staats et al., 2013) may remain the most effective means of extracting genetic data from severely degraded specimens or when many individuals are required to address a research question. However, as with all reduced-representation methods, there are several biases that should be considered during experimental planning (Graham et al., 2015; Jackson et al., 2015).

Ultimately, our results underline the importance of continued scientific collecting and archival of legacy collections into NHCs, as new technologies will continue to improve our ability to extract molecular information from degraded and aged samples in the future. The centralization of biological resources and associated information in NHCs ensures the broad utility of these specimens to the scientific community and facilitates tests, such as these, to determine the best available means of extracting meaningful sequence data from lower-quality samples. In particular, we endorse maximizing the utility of a specimen through the archival of multiple tissue types, through multiple storage media (liquid nitrogen, ethanol, RNAlater [Sigma-Aldrich], etc.) to maximize future applications of these archives as technology evolves (Lessa et al., 2014; McLean et al., 2016). The ability to generate quality de novo assemblies from field-preserved tissues encourages the expansion of remote field expeditions and resurvey projects that incorporate tissue collection and preservation to expand our ability to extract genomic data from extinct and difficult-to-sample taxa. In an era of unprecedented ecological and environmental change (Ceballos, Ehrlich, & Dirzo, 2017), maximizing the extraction of genomic information from historical samples holds promise for helping us to understand the evolutionary responses of natural populations to environmental perturbation and lay the foundation for predicting future responses and undertaking proactive management (Malaney & Cook, 2013; Wandeler et al., 2007).

ACKNOWLEDGEMENTS

We thank the Museum of Southwestern Biology for the *Peromyscus* tissue loans, Christopher Tracy for access to the Boyd Deep Canyon Reserve, Anthony Westbrook for computational support, and the UNH Premise computing cluster. Thank you also to the anonymous reviewers for their constructive feedback. This work was funded by the National Institute of Health National Institute of General Medical Sciences to M.D.M. (1R35GM128843).

AUTHOR CONTRIBUTIONS

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

DATA AVAILABILITY STATEMENT

All commands used for this work are available at https://github.com/macmanes-lab/museum_genomics. Raw 10× 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.

ORCID

Jocelyn P. Colella https://orcid.org/0000-0003-2463-1029
Anna Tigano https://orcid.org/0000-0001-9240-3058
Matthew D. MacManes https://orcid.org/0000-0002-2368-6960

REFERENCES

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

Ando, T., Matsuda, T., Goto, K., Hara, K., Ito, A., Hirata, J., … Kobayashi, M. (2018). Repeated inversions within a pannier intro drive diversification of intraspecific colour patterns of ladybird beetles. *Nature Communications*, *9*, 1–13.

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*). GenBank. https://www.ncbi.nlm.nih.gov/nuccore/1056038647. Deposited 18 August 2016.

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.

Bradley, R. D., Durish, N. D., Rogers, D. S., Miller, J. R., Engstron, M. D., & Kilpatrick, C. W. (2007). Toward a molecular phylogeny for *Peromyscus:* Evidence from mitochondrial cytochrome-*b* sequences. *Journal of Mammalogy*, *88*(5), 1146–1159.

Broad Institute. (2015). DISCOVAR: Assemble genomes, find variants. Retrieved from 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.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., … Jaffe, D. B. (2008). ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Research*, *18*(5), 810–820.

Cantarel, B. L., Korf, I., Robb, S. M., 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. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences USA*, *114*(30), E6089–E6096.

Cheviron, Z., & Brumfeld, R. (2012). Genomic insights into adaptation to high-altitude environments. *Heredity*, *108*(4), 354–361.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–571.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., … Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), 1050.

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 D. M. Hillis, & C. Moritz (Eds.), *Molecular systematics* (pp. 29–47). Sunderland, MA: Sinauer Associates.

Edelman, N., Frandsen, P., Miyagi, M., Clavijo, B., Davey, J., Dikow, R., … Mallet, J. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science*, *366*(6465), 594–599.

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

Etherington, G. J., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G., Di Palma, F. (2019). Sequencing smart: *De novo* sequencing and assembly approaches for non-model mammals. *bioRxiv*, 1-39. https://doi.org/10.1101/723890

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.

Feigin, C. Y., Newton, A. H., Doronina, L., Schmitz, J., Hipsley, C. A., Mitchell, K., … Pask, A. J. (2018). Genome of Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. *Nature Ecology and Evolution*, *2*, 182–192.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., & Kirov, G. (2012). Discover and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, *91*(4), 597–607.

Gibb, G. C., Condamine, F. L., Kuch, M., Enk, J., Moraes-Barros, N., Superina, M., … & Delsuc, F. (2016). Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living xenarthrans. *Molecular Biology and evolution*, *33*(3), 621–642.

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.

Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., … Wilson, J. Y. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequences (RADSeq). *Molecular Ecology Resources*, *15*, 1304–1315.

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

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., … Savolainen, V. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology*, *62*(4), 539–554.

Havlak, P., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X.-Z., … & Gibbs, R. A. (2004). The Atlas genome assembly system. *Genome research*, *14*(4), 721–732.

Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., … Yoder, A. D. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, *54*(1), 291–301.

Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., LaBarbera, K., Wommack, E. A., … Nachman, M. W. (2016). Natural history collections as windows on evolutionary processes. *Molecular Ecology*, *25*, 864–881.

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.

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. Retrieved from 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, M. R., Mills, L. S., Jensen, J. D., & Good, J. M. (2019). Convergent evolution of seasonal camouflage in response to reduced snow cover across snowshoe hare range. *BioRxiv*, 1-34. https://doi.org/10.1101/851766

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

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., … Itoh, T. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*(8), 1384–1395.

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.

Latorre, S. M., Herrmann, M., Paulsen, M. J., Rödelsperger, C., Dréau, A., Röseler, W., Burbano, H. A. (2020). Museum phylogenomics of extinct Oryctes beetles from the Mascarene Islands. *bioRxiv*, 1-33. https://doi.org/10.1101/2020.02.19.954339

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. https://doi.org/10.1101/048603

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.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., & … Tang, J. (2012). SOALDENOVO2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, *1*(1), 18.

Lutgen, D., Ritter, R., Olsen, R. A., Schielzeth, H., Gruselius, J., Ewels, P. A., Burri, R. (2020). Linked-read sequencing enables haplotype-resolved resequencing at population scale. *bioRxiv*, 1-19. https://doi.org/10.1101/2020.01.15.907261

Mack, K. L., & Nachman, M. W. (2017). Gene regulation and speciation. *Trends in Genetics*, *33*(1), 68–80.

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.

Mandelker, D., Schmidt, R. J., Ankala, A., McDonald Gibson, K., Bowser, M., Sharma, H., … Funke, B. (2016). Navigating highly homologous genes in a molecular diagnostic setting: A resource for clinical next-generation sequencing. *Genetics in Medicine*, *18*(12), 1282.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfeld, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Research*, *22*, 746–754.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., … Fiston-Lavier, A.-S. (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.

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

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.

Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC: Plotly Technologies Inc. https://plot.ly

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.

Schmitt, C. J., Cook, J. A., Zamudio, K. R., & Edwards, S. V. (2018). Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philosophical Transactions of the Royal Society B*, *374*, 20170387.

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.

Staats, M., Erkens, R. H. J., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., … Bakker, F. T. (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.

Tigano, A., Colella, J. P., & MacManes, M. D. (2020). Comparative and population genomics approaches reveal the basis of adaptation to deserts in a small rodent. *Molecular Ecology*, in press. https://doi.org/10.1111/mec.15401

Toomey, M. B., Marques, C. I., Andrade, P., Araújo, P. M., Sabatino, S., Gazda, M. A., … Carneiro, M. (2018). A non-coding region near Follistatin controls head color polymorphism in the Gouldian finch. *Proceedings of the Royal Society B*, *285*, 20181788.

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.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). PILON: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11), e112963.

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.

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.

Yeo, S., Coombe, L., Warren, R. L., Chu, J., & Birol, I. (2018). ARCS: Scaffolding genome drafts with linked reads. *Bioinformatics*, *34*(5), 725–731.

Zhang, P., Lehmann, B. D., Shyr, Y., & Guo, Y. (2017). The utilization of formalin fixed-paraffinembedded specimens in high throughput genomic studies. *International journal of genomics*, *2017*.

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., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., … Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, *34*(3), 303–311.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Colella JP, Tigano A, MacManes MD. A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues. *Mol Ecol Resour*. 2020;20:871–881. https://doi.org/10.1111/1755-0998.13155