Genomic resources for the North American water vole (*Microtus richardsoni*) and the montane vole (*Microtus montanus*)

Drew J. Duckett[1*], Jack Sullivan[2], Stacy Pirro[3], Bryan C. Carstens[1]

[1]Department of Evolution, Ecology, and Organismal Biology. The Ohio State University. 1315 Kinnear Rd., Columbus OH, 43212

[2]Department of Biological Sciences, Box 443051, University of Idaho, Moscow ID, 83844-3051

[3]Iridian Genomes, Inc., 6213 Swords Way, Bethesda MD 20817

*email: duckettdj@gmail.com

Abstract

Background: Voles of the genus *Microtus* are important research organisms, yet genomic resources in the genus are lacking. Providing such resources would benefit future studies of immunology, phylogeography, cryptic diversity, and more. *Findings:* We sequenced and assembled nuclear genomes from two subspecies of water vole (*Microtus richardsoni*) and from the montane vole (*Microtus montanus*). The water vole genomes were sequenced with Illumina and 10X Chromium plus Illumina sequencing, resulting in assemblies with ~1,600,000 and ~30,000 scaffolds respectively. The montane vole was assembled into ~13,000 scaffolds using Illumina sequencing also. In addition to the nuclear assemblies, mitochondrial genome assemblies were also performed for both species. We conducted a structural and functional annotation for the best water vole nuclear genome, which resulted in ~24,500 annotated genes, with 83% of these receiving functional annotations. Finally, we find that assembly quality statistics for our nuclear assemblies fall within the range of genomes previously published in the genus *Microtus*, making the water vole and montane vole genomes useful additions to currently available genomic resources.

Keywords: genome assembly; genome annotation; mitochondrial genome; 10X Chromium; Illumina sequencing

Context

The genus *Microtus* consists of 62 species of voles distributed throughout North America, Europe, and Asia [1]. *Microtus* is believed to have experienced rapid speciation and diversification, with all speciation events occurring within the past four million years [2, 3], and it has been suggested that some nominal species, such as *M. pennsylvanicus*, contain cryptic diversity [4]. *Microtus* has been an important model system across multiple biological disciplines, including studies of adaptation (e.g., [5]), infectious disease (e.g., [6]), parental care (e.g., [7]), and population dynamics (reviewed in [8]). The rapid radiation of *Microtus* voles has hindered systematic classification, leading to multiple taxonomic revisions and conflicting phylogenetic analyses [1, 9, 10]. Consequently, both species boundaries and relationships among species are difficult to infer. Genomic resources within *Microtus* will help resolve these questions, and resources have steadily increased in recent years. Currently, four *Microtus* species have assembled genomes on GenBank, two European species (*M. agrestis* and *M. arvalis*) and two North American species (*M. ochrogaster;* [11], and *M. oeconomus*). The present study provides resources for two additional species: *M. richardsoni* and *M. montanus*.

The North American water vole (*M. richardsoni*) is adapted to a semiaquatic lifestyle, relying on alpine and sub-alpine streams for creating burrows and escaping predators [12]. Like other semiaquatic mammals (e.g., otters), it is likely that adaptations to this lifestyle have been driven by natural selection [13-15]. Water voles are among the largest species of *Microtus* and are known for making runways of stamped-down vegetation along streams through frequent movement [12, 16]. Unlike most other vole species, *M. richardsoni* does not appear to experience regular population boom and bust cycles, although population size in the species may be correlated to levels of precipitation [17]. Despite being listed as *Least Concern* by the IUCN

Redlist [18], the species is listed as *Critically Imperiled* by the Wyoming Natural Diversity Database due to its specific habitat requirements, which can be substantially degraded by livestock grazing [19]. *Microtus richardsoni* occupies a large, disjunct distribution in the Pacific Northwest of North America, with habitat in the Cascades Mountains and the Rocky Mountains, spanning from southern Canada into central Utah. Four subspecies are currently recognized: *M. r. arvicoloides* in the Cascades Mountains, *M. r. richardsoni* in the Canadian Rocky Mountains, *M. r. macropus* in the central Rocky Mountains and Wyoming, and *M. r. myllodontus* in Utah. Due to the subspecific classifications and the disjunct range of the species, *M. richardsoni* has been included in multiple studies of phylogeography in the Pacific Northwest [20-22]. These studies were based solely on mitochondrial DNA, and the results of analyses that investigated species limits and demographic history were limited to inferences that can be derived from a single gene tree. Genomic resources for *M. richardsoni* will provide a rich source of data to address these knowledge gaps.

The montane vole (*M. montanus*) is partially sympatric with *M. richardsoni* and can be found throughout most of the water vole's range with the exception of the Canadian Rockies. However, *M. montanus* can be found farther south and east including areas of California, Nevada, Colorado, Arizona and New Mexico [23]. The species has been divided into fifteen subspecies, including *M. m. canescens* in the Cascades Mountains., *M. m. nasus* in the central Rocky Mountains, and *M. m. amosus* in northern Utah. Notably, *M. montanus* does not exhibit a break in its range in the Columbia Basin, likely because it is not restricted to riparian areas like *M. richardsoni*. The species as a whole is listed as *Least Concern* by the IUCN Redlist, but *M. m. arizonicus* has been listed as endangered by the New Mexico State Game Commission Regulation [23], and *M. m. ricularis* has been noted as being of concern due to a small range and declining population size [24]. Genomic resources in *M. montanus* will provide a wealth of data to assess subspecies boundaries, quantify gene flow among subspecies, and aid in conservation efforts of threatened subspecies.

The present study provides two nuclear and one mitochondrial genome assembly for *M. richardsoni* along with single nuclear and mitochondrial genome assemblies for *M. montanus*. Furthermore, a structural and functional annotation are performed with one of the *M. richardsoni* genomes to aid in future studies of adaptation. Genome-level comparisons are made between the new genome assemblies and other *Microtus* genome assemblies to examine differences in assembly quality and repeat content.

Sequencing and Nuclear Genome Assembly

Frozen tissue from a single *M. r. arvicoloides* individual collected from the southern Cascades Mountain range (JMS_292; 44.016667N, -121.750000E; [20]) was sent to Hudson Alpha (Huntsville, AL) for high molecular weight DNA extraction and 10X Chromium library preparation [25]. In the 10X method, each extracted DNA fragment receives a different barcode before the fragment is sheared for library preparation. After sequencing, these barcodes are used to connect sequencing reads for a more contiguous assembly. After sequencing with a single run on an Illumina HiSeqX, the resulting 150 base pair (bp) paired-end reads were input into Supernova for *de novo* genome assembly with --maxreads=all [26].

Additional tissue was obtained from a single *M. r. macropus* individual collected from the northern Rocky Mountains (JMG_88; 46.333333N, -114.633333E; [20]). DNA was extracted using a Qiagen DNeasy Blood and Tissue Kit, and the DNA was sent for library preparation and sequencing by Iridian Genomes, Inc (Bethesda, MD). 150bp paired-end reads were sequenced on two runs of an Illumina HiSeqX. Genome assembly was performed using two different deBruijn graph-based programs, SOAPdenovo and Discovar de novo [27, 28]. For SOAPdenovo, quality trimming was performed using fastQC and Trimmomatic with settings ILLUMINACLIP: 2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:36 [29, 30]. SOAPdenovo assemblies were performed with settings max_rd_len=150, avg_ins=300, reverse_seq=0, asm_flags=3, rd_len_cutoff=150, rank=1, pair_num_cutoff=3, and map_len=32. SOAPdenovo was run with kmer values of 63, 89, 95, and 101 based on analysis of optimal kmer values in kmerGenie [31]. Raw reads were used as input for *de novo* genome assembly with Discovar as recommended in the program documentation.

In an attempt to provide the most contiguous assembly for *M. richardsoni*, a hybrid assembly was performed using the ARCS+LINKS pipeline [32, 33]. The ARCS+LINKS pipeline uses barcoding information from the 10X Chromium reads to scaffold the contigs from a separate genome assembly. Barcoded reads from *M. r. arvicoloides* were mapped to the *M. r. macropus* Discovar assembly with bwa mem [34] before converting the mapped reads to BAM format and sorting with SAMTools [35]. ARCS and LINKS were then run with settings –s 98 –c 5 –l 0 –z 500 –d 0 –r 0.05 –m 50-10000 –e 30000 and –d 4000 –k 20 –l 5 –t 2 –a 0.3 –o 0 –a 0.3 –z 500 respectively.

As part of a separate project, a single *M. montanus* individual from Utah (UMNH:Mamm:30891; 38.19381N, -111.5824E) was misidentified as *M. richardsoni*. DNA was extracted from the sample using a Qiagen DNeasy Blood and Tissue Kit before being sent to the University of California Davis Genome Center for library preparation and sequencing. Paired-end 150bp sequences were collected with a single shared run on an Illumina NovaSeq. Species identity was confirmed using the Barcode of Life Database (BOLD; [36]). Reads were checked and trimmed for quality with fastQC and Trimmomatic as above before mapping reads to the mitochondrial cytochrome oxidase I (COI) sequence of *M. r. macropus* [37] using bwa mem. The resulting mapped reads were converted to BAM format, sorted, and indexed with SAMTools. PCR duplicates were identified and removed with Picard [38], resulting reads were piled with SAMTools mpileup using base and mapping quality scores of 30, consensus sequences were generated with bcftools [39], and consensus sequences were converted to fastq format using vcfutils with a minimum depth filter of 5 and maximum depth filter of 10000 [35]. The resulting sequence was input into BOLD. Due to the low sequencing coverage, *de novo* genome assembly was not appropriate for *M. montanus*. To provide a preliminary genome sequence, a reference-guided genome assembly was performed with RaGOO [40]. Raw reads were input into Discovar to generate an initial genome assembly, misassembly correction was performed with RaGOO using reads trimmed with the same settings as the *M. r. macropus* reads, and RaGOO was then used to scaffold the Discovar contigs onto the *M. r. arvicoloides* assembly, which is more closely related to *M. montanus* than the other available *Microtus* genome assemblies [3]. Since *M. montanus* has less than half the chromosomes of *M. richardsoni* (2n = 22-24 in *montanus* versus 56 in *richardsoni* [41]), the possibility of structural errors in the *M. montanus* assembly was examined by calculating the percentage of reads that mapped back to the assembly using bwa mem and bamtools [42].

The final assemblies were submitted to GenBank [43], where screening was performed to identify any contamination, and contaminated scaffolds were removed. All assemblies were evaluated with QUAST [44], bbmap [45], custom Python scripts (https://github.com/djlduckett/Genome_Resources/), and BUSCO using the Euarchontoglires reference set [46]. After comparing assembly statistics from the different assemblies of *M. r.*

macropus, the Discovar assembly was selected as best because it had less fragmentation, higher N50 and L50, and a higher BUSCO score than the SOAPdenovo assemblies (Table 1). Genome sequencing of *M. r. arvicoloides* produced over 800 million (M) reads and 47x genome sequencing coverage. The final genome assembly consisted of ~32 thousand (K) scaffolds with an N50 of 2.3 megabase pairs (Mb), 1.3% missing data (N), and a BUSCO score of 85.8%. Supernova estimated the length of the genome assembled to be ~2.4Gb and the total genome size to be ~2.6 gigabase pairs (Gb). *Microtus richardsoni macropus* sequencing produced over 600M reads and 35x coverage. Genome assembly with Discovar resulted in ~1.6M scaffolds with an N50 of 16 kilobase pairs (Kb), 0.06% Ns, and a BUSCO score of 54.5%. Given that there are many programs to perform *de novo* genome assembly from short reads, it is possible that another program would have produced a more contiguous *M. r. macropus* assembly, but previous studies have shown Discovar performs well compared to other programs [47, 48]. The hybrid assembly produced with the ARCS+LINKS pipeline had ~1.6M scaffolds, an N50 of 38Kb, 0.09% Ns, and a BUSCO score of 59.8%. Because of the poor quality of the hybrid assembly, it was not used for further analyses, and the *M. richardsoni* subspecies assemblies were kept separate. It seems likely that the high fragmentation of the hybrid assembly is due to the fragmentation of the Discovar input assembly. Published results with this hybrid pipeline often include a much higher sequencing coverage of the input contigs to produce a better starting point for the pipeline. Therefore, additional Illumina sequencing with *M. r. macropus* in the future could substantially improve the hybrid assembly. 108M reads (13x coverage) were used to produce the preliminary *M. montanus* genome, resulting in ~13K scaffolds, an N50 of ~3.1Mb, 8.8% Ns, and a BUSCO score of 82.6%. Additionally, 89.3% of reads mapped back to the *M. montanus* assembly.

Table 1

	Discovar	SOAPdenovo	SOAPdenovo	SOAPdenovo	SOAPdenovo
Kmer	NA	63	89	95	101
Length	2.54Gb	2.72Gb	2.88Gb	2.89Gb	3.21Gb
Scaffolds	1.6M	4.1M	4.0M	4.1M	6.7M
Max Scaffold	264Kb	186Kb	146Kb	174Kb	139Kb
N50	16.1Kb	4.5Kb	3.4Kb	3.4Kb	1.5Kb
L50	35.7K	117K	156K	163K	371K
BUSCO	54.5%	38.1%	37.1%	35.9%	25.9%
% N	0.06	1.45	0.99	0.94	0.90
% GC	42.13	41.92	41.91	41.92	41.98

Comparison of genome assembly strategies for *M. r. macropus*. NA: not applicable.

Mitochondrial Genomes

The complete mitochondrial genomes of *M. r. arvicoloides* and *M. montanus* were assembled using the genomic sequencing reads. The mitochondrial genomes were assembled by both mapping reads to a reference mitochondrial genome and using the reference-guided assembly program Novoplasty [49]. For the mapping assembly, reads were mapped to the *M. r. macropus* mitochondrial genome, using the same steps as the *M. montanus* BOLD analysis. The mitochondrial assemblies were 16,285bp and 16,268bp in length with an average depth of coverage of 7886x and 6805x for *M. r. arvicoloides* and *M. montanus* respectively. Reference guided mitochondrial assemblies with Novoplasty used the *M. r. macropus* mitochondrial genome as the reference along with settings *Genome Range=12000-22000, K-mer=33, Read*

Length=150, and Insert size=400. Because the *M. r. arvicoloides* dataset contained many reads, 25% of reads were subsampled to use for assembly, as suggested in the program documentation. The assemblies for *M. r. arvicoloides* and *M. montanus* were 16,298bp and 16,319bp in length with average depths of coverage of 5131x and 14,713x respectively. To compare mitochondrial assemblies between methods, the assemblies were aligned using the MUSCLE plugin in Geneious v. R9 with eight iterations and an open gap score of -1 [50, 51]. This comparison showed the Novoplasty assemblies contained multiple insertions compared to the mapped assemblies and the reference mitochondrial genome. These insertions were up to 13bp long in multiple genes, including trnT, trnK, and ATP8. Comparison to other *Microtus* mitochondrial genomes (*M. ochrogaster*; NC_027945.1 and *M. fortis*; NC_015243.1) showed that the Novoplasty assemblies were the only mitochondrial assemblies to exhibit these insertions. Therefore, the mapping assemblies were used for further analyses. The mapping assemblies for both species included ambiguous bases, which were much more frequent for *M. montanus* than *M. r. arvicoloides*. These may be the result of using the mitochondrial genome of a different subspecies (for *M. r. arvicoloides*) or species (for *M. montanus*) for mapping the reads. Additionally, the presence of nuclear DNA of mitochondrial origin (NUMTs; [52, 53]) may have influenced these results. If mitochondrial segments have been incorporated into the nuclear genomes and subsequent mutations have occurred, both nuclear and mitochondrial sequences could be mapped to the same mitochondrial region during assembly and result in the ambiguous bases observed here. It is likely that NUMTs are present, as they have been documented in other species of *Microtus* [54-56]. Both mitochondrial genomes were annotated using MITOS [57]. The annotations each consisted of 22 tRNA genes, 2 rRNA genes, and 13 protein coding genes.

Microtus Genome Assembly Comparison

The available *Microtus* genome assemblies, *M. agrestis* (GCA_902806755.1), *M. arvalis* (GCA_007455615.1), *M. ochrogaster* (GCA_000317375.1), and *M. oeconomus* (GCA_007455595.1), were downloaded from GenBank. Assembly summary statistics were calculated using QUAST, bbmap, and custom Python scripts (https://github.com/djlduckett/Genome_Resources/). To compare repeat content among all genomes, including the three produced by the current study, repeats were first identified *de novo* using RepeatModeler [58]. RepeatMasker was then used to further identify repeats using a combined repeat library that included the repeats identified from RepeatModeler and those from the RepeatMasker *Rodentia* database [59]. The percentage of the genome consisting of each type of repeat element was extracted from the RepeatMasker log file for each genome assembly.

All genome assemblies used some form of Illumina sequencing (Table 2), although assembly continuity varied greatly among assemblies from 1366 scaffolds in *M. agrestis* to 1.6 M scaffolds in *M. r. macropus*. Genome coverage was similarly varied, from 13x in *M. montanus* to 35x in *M. r. macropus* to 77x in *M. arvalis* and *M. oeconomus*. The percent of repetitive regions ranged from 31.7% in *M. montanus* to 44.1% in *M. arvalis* (Figure 1), and repeat content did not appear to be associated with phylogenetic relatedness as repeats between the two subspecies of *M. richardsoni* were not more similar to each other than to other *Microtus* species. However, it is possible that the repeat content is affected by the continuity of the genome assemblies, and further research is needed to confirm this relationship.

Table 2

Species	***M. agrestis***	***M. arvalis***	***M. montanus****	***M. ochrogaster***	***M. oeconomus***	***M. r. arvicoloides****	***M. r. macropus****
Distribution	Europe	Europe	North America	North America	North America	North America	North America
Year	2020	2019	2020	2012	2019	2020	2020
Accession (GCA_)	902806775.1	7455615.1	xxxxxxxxxxx	317375.1	7455595.1	xxxxxxxxxxx	xxxxxxxxxxx
Sequencing	10X Chromium + Illumina	Illumina	Illumina	Illumina	Illumina	10X Chromium + Illumina	Illumina
Assembler	Supernova	Discovar	RaGOO	ALLPATHS	Discovar	Supernova	Discovar
Length	2.03Gb	2.62Gb	2.34Gb	2.29Gb	2.31Gb	2.36Gb	2.54Gb
Coverage	50	77	13	94	77	47	35
# Scaffolds	1,366	1,081,432	12,962	6,341	562,436	31,632	1,648,927
Longest Scaffold	56.96Mb	0.80Mb	748.72Mb	126.73Mb	0.93Mb	16.00Mb	0.26Mb
N50	13.35Mb	0.53Mb	3.08Mb	61.81Mb	0.11Mb	2.30Mb	0.02Mb
L50	45	11,870	91	14	5,556	278	35,660
%N	2.87	0.07	8.81	8	0.12	1.29	0.06
%GC	42.33	41.71	42.38	42.25	42.18	42.21	42.13

Genome assembly comparison among *Microtus* species. Assemblies with a * were produced by the present study. Note: in-depth methods for *M. agrestis* are not available, and it is possible that the assembly includes additional sequencing and/or methods.

Genome Annotation

The *M. r. arvicoloides* genome assembly was annotated with the MAKER pipeline [60], loosely following the tutorial provided by Daren Card (https://gist.github.com/darencard /bb1001ac1532dd4225b030cf0cd61ce2). Briefly, the pipeline consists of masking repeats followed by multiple rounds of annotation with both evidence-based and ab-initio gene models. Repeats were identified as described above. Complex repeats were then extracted from RepeatMasker results using grep with keywords “Satellite” and “rich”. Within Maker, the model_org argument was set to “simple” so Maker would soft mask simple repeats, and the RepeatMasker results were provided to hard mask complex repeats. Evidence-based gene discovery used protein and mRNA sequences from the previous genome annotation of *M. ochrogaster* (GCF_000317375.1) as well as an additional RNASeq assembly from *M. pennsylvanicus* (GSM3499528; [61]). Hidden Markov models (HMMs) for ab-initio gene prediction were trained using both SNAP and Augustus [62, 63]. With SNAP, gene models identified by MAKER were filtered using an Annotation Edit Distance (AED) of 0.5 and an amino acid length of 50. After validating these models with SNAP’s Fathom utility, removing likely errors, and including 1000bp surrounding each training sequence, the training sequences were passed to the hmm-assembler script. For Augustus, training sequences plus 1000bp on each side were obtained from the first round of MAKER mRNA annotations. Augustus was used to train the HMM using the --long option in BUSCO and the Euarchontoglires reference set. MAKER was then run again with the previously annotated gene models and the HMM models from SNAP and Augustus. After the initial MAKER run, two cycles of ab-initio gene prediction and annotation with MAKER were performed. To prevent overfitting, results were compared after each round of MAKER. Because the increase in AED score was minimal between the first and second rounds of ab-initio gene prediction, further analysis was conducted on the results after the first round only. This round annotated ~24K genes with a mean gene length of 7445bp (Table 3), which is within the range found in previous studies of *M. ochrogaster* (22,427 genes; GCF_000317375.1) and *Arvicola amphibious* (25,136 genes; GCF_903992535.1). Of these annotations all occurred on scaffolds greater than 1Kb in length and 97% occurred on scaffolds greater than 10Kb in length.

Functional annotation of the *M. r. arvicoloides* genome was performed using GOfeat, an online functional annotation tool that uses multiple protein databases including UniProt, InterPro, and Pfam [64-67]. An input file for GOfeat was generated by supplying the genome assembly FASTA file and the MAKER General Feature Format (GFF3) file to the Python package gffread [68]. GOfeat annotated 83.49% of genes. Biological Processes accounted for 42.46% of annotations, Cellular Components accounted for 30.29%, and Molecular Functions comprised 27.25%. The most frequent gene ontology (GO) terms were *positive regulation of transcription by RNA polymerase II, negative regulation of transcription by RNA polymerase II,* and *DNA-templated regulation of transcription* for Biological Processes, *cytoplasm* and *plasma membrane* for Cellular Components, and *metal ion binding* and *calcium ion binding* for Molecular Functions.

Table 3

	Before Gene Modeling	**Gene Modeling Round 1**	**Gene Modeling Round 2**
Genes	20,945	24,548	23,811
Exons	139,845	192,974	179,225
mRNA	20,945	24,548	23,811
tRNA	-	24,504	24,539
5’ UTR	-	1,229	1,180
3’ UTR	-	503	642
Mean Gene Length	-	7,445	7,132
AED < 0.50	0.993	0.881	0.888
AED < 0.25	0.672	0.543	0.520
BUSCO (Complete)	-	67.7%	70.5%

Structural annotation summary after each round of MAKER. UTR: untranslated region; AED: annotation edit distance. Values with dashes were not analyzed prior to gene modeling with SNAP and Augustus.

Conclusion

The current study details the assembly and annotation of three nuclear and two mitochondrial genomes. Compared to previously published nuclear genomes, the *M. r. arvicoloides* and *M. montanus* genomes are of high quality as evidenced by the low number of scaffolds, high N50/L50 values, and high BUSCO scores. While not as complete as the other *Microtus* genomes, the nuclear genome of *M. r. macropus* will still be useful for mapping low coverage reads or reduced representation sequencing data. Furthermore, the mitochondrial genomes contributed here add to a growing number for the genus *Microtus* and reinforce earlier suggestions that high-quality mitochondrial genomes can be obtained as byproducts of nuclear sequencing (e.g., [69, 70]). Overall, the data presented serve as an example that even though they do not include chromosomal information, high-quality draft genomes can be produced from widely available and very cost-effective methods like the 10X Chromium protocol. These references can aid a variety of studies including those examining genus and species adaptation [71, 72], phylogenetics [10], phylogeography [22, 73], and disease dynamics [6, 74]. However, some activities, like exploring changes to chromosome structure, will not be possible due to the fragmentation and lack of chromosomal mapping for these assemblies. Finally, the *M. r. macropus* and *M. montanus* sequencing data and preliminary assemblies will serve as the building blocks of more accurate reference genomes in the future.

Availability of Supporting Data and Materials

Raw sequences, nuclear assemblies, and mitochondrial assemblies are available from GenBank under BioProjects PRJNA673719, PRJNA509068, and PRJNA673873 for *M. r. arvicoloides*, *M. r. macropus*, and *M. montanus* respectively. The custom python script used to calculate genome assembly summary information is available on GitHub (https://github.com/djlduckett/Genome_Resources/). Full BUSCO tables, structural annotation gff files, functional annotation tables, and repeat libraries are available in the GigaScience data repository (http://gigadb.org/).

Figures:

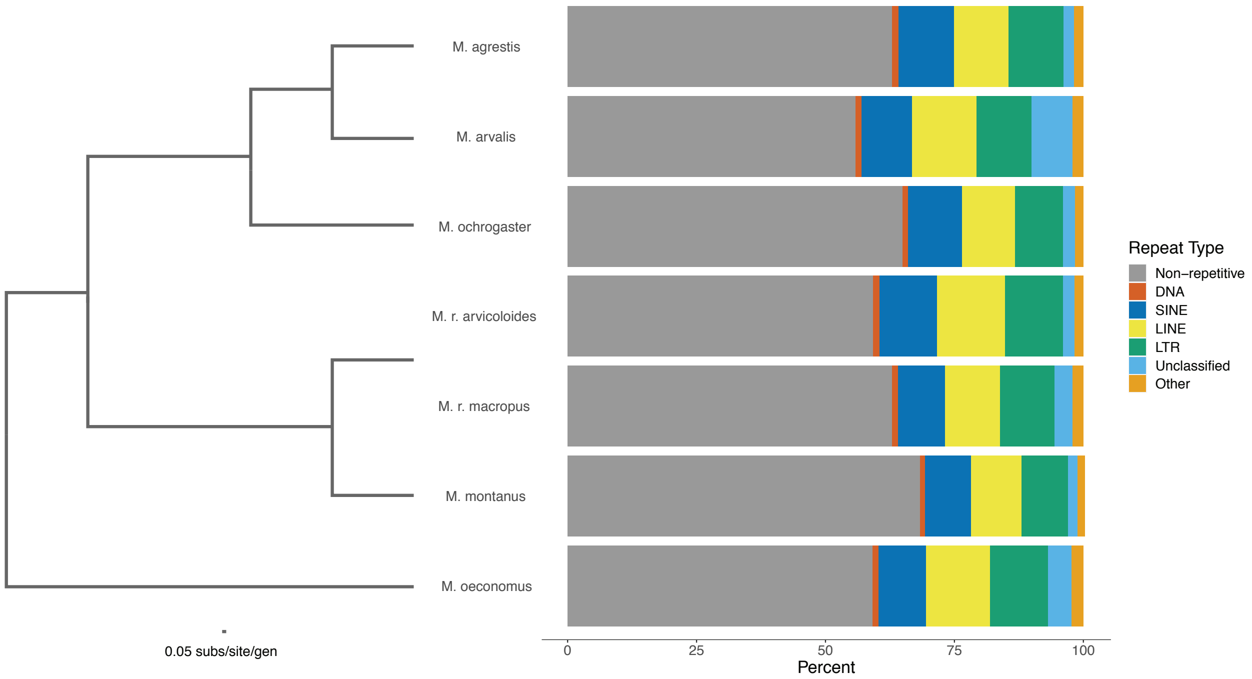

Figure 1: Repeat Content among *Microtus* genomes. SINE: short interspersed nuclear element; LINE: long interspersed nuclear element; LTR: long terminal repeat; Other consists of small RNA, satellite, simple, and low complexity repeats. The phylogeny displayed was recreated from [3] by pruning unincluded species from the data alignment and rerunning RAxML [75] with the same settings used in the original analysis.

Abbreviations

BOLD: Barcode of Life Database; bp: base pair; COI: cytochrome oxidase I; M: million; K: thousand; Gb: gigabase pairs; Mb: megabase pairs; Kb: kilobase pairs; SINE: short interspersed nuclear element; LINE: long interspersed nuclear element; LTR: long terminal repeat; AED: annotation edit distance; HMM: hidden Markov model; NUMT: nuclear DNA of mitochondrial origin

Competing Interests

SP is the director of Iridian Genomes, Inc.

Funding

Sequencing was funded by Iridian Genomes, Inc., as well as the National Science Foundation (DEB-1457519). Salary support for DD was provided by The Ohio State University and the National Science Foundation (DBI-1945347).

Author Contributions

DD, JS, and BC conceived the study. JS, SP, and BC provided funding for sequencing. DD performed DNA extractions, assembled genomes, and annotated genomes with input from SP. DD and BC wrote the manuscript with input from JS and SP. DD and SP submitted the resources to GenBank.

Acknowledgements

We thank Jeffrey Good and Eric Rickart/Utah Museum of Natural History for tissue samples, Michael Broe for advice with genome assembly and annotation, and the Ohio Supercomputer Center (OSC) for computational resources.

References

1. Wilson DE, Reeder DM, editors. Mammal species of the world: a taxonomic and geographic reference. JHU Press; 2005.
2. Reig OA. Karyotypic repatterning as one triggering factor in cases of explosive speciation. In: Fontdevila A, editor. Evolutionary biology of transient unstable populations. Springer, Berlin, Heidelberg; 1989 p. 246-289.
3. Steppan SJ, Schenk JJ. Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. PloS one. 2017;12:8.
4. Jackson DJ, Cook JA. A precarious future for distinctive peripheral populations of meadow voles (*Microtus pennsylvanicus*). Journal of Mammalogy. 2020;101:1:36-51.
5. Monarca RI, Speakman JR, da Luz Mathias M. Energetics and thermal adaptation in semifossorial pine-voles *Microtus lusitanicus* and *Microtus duodecimcostatus*. Journal of Comparative Physiology B. 2019;189:2:309-18.
6. Wanelik KM, Begon M, Birtles RJ, Bradley JE, Friberg IM, Jackson JA, Taylor CH, Thomason AG, Turner AK, Paterson S. A candidate tolerance gene identified in a natural population of field voles (*Microtus agrestis*). Molecular ecology. 2018;27:4:1044-52.
7. Seelke AM, Perkeybile AM, Grunewald R, Bales KL, Krubitzer LA. Individual differences in cortical connections of somatosensory cortex are associated with parental rearing style in prairie voles (*Microtus ochrogaster*). Journal of Comparative Neurology. 2016;524:3:564-77.
8. Oli MK. Population cycles in voles and lemmings: state of the science and future directions. Mammal Review. 2019;49:3:226-39.
9. Bailey V. Revision of American voles of the genus *Microtus*. North American Fauna. 1900;17:1-88.

10. Barbosa S, Paupério J, Pavlova SV, Alves PC, Searle JB. The *Microtus* voles: Resolving the phylogeny of one of the most speciose mammalian genera using genomics. Molecular phylogenetics and evolution. 2018;125:85-92.
11. McGraw LA, Davis JK, Lowman JJ, ten Hallers BF, Koriabine M, Young LJ, De Jong PJ, Rudd MK, Thomas JW. Development of genomic resources for the prairie vole (*Microtus ochrogaster*): construction of a BAC library and vole-mouse comparative cytogenetic map. BMC genomics. 2010;11:1:1-8.
12. Klaus M, Beauvais GP. Water Vole (Microtus richardsoni): A Technical Conservation Assessment. Prepared for USDA Forest Service, Rocky Mountain Region, Species Conservation Project. 2004.
13. Stein BR. Bone density and adaptation in semiaquatic mammals. Journal of mammalogy. 1989;70:3:467-76.
14. Dunstone N, Gorman ML, editors. Behaviour and ecology of riparian mammals. Cambridge University Press; 2007.
15. Beichman AC, Koepfli KP, Li G, Murphy W, Dobrynin P, et al. Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter. Molecular biology and evolution. 2019;36:12:2631-55.
16. Ludwig DR. Microtus richardsoni. Mammalian Species. 1984;223:1-6.
17. Klaus M, Moore RE, Vyse E. Impact of precipitation and grazing on the water vole in the Beartooth Mountains of Montana and Wyoming, USA. Arctic, Antarctic, and Alpine Research. 1999;31:3:278-82.
18. Cassola, F. *Microtus richardsoni*. The IUCN Red List of Threatened Species 2016.
19. Wyoming Natural Diversity Database. University of Wyoming. 2020. https://wyndd.org/species_list/.

20. Cartens BC, Brunsfeld SJ, Demboski JR, Good JM, Sullivan J. Investigating the evolutionary history of the Pacific Northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework. Evolution. 2005;59:8:1639-52.
21. Carstens BC, Richards CL. Integrating coalescent and ecological niche modeling in comparative phylogeography. Evolution: International Journal of Organic Evolution. 2007;61:6:1439-54.
22. Espíndola A, Ruffley M, Smith ML, Carstens BC, Tank DC, Sullivan J. Identifying cryptic diversity with predictive phylogeography. Proceedings of the Royal Society B: Biological Sciences. 2016;283:1841:20161529.
23. Sera WE, Early CN. *Microtus montanus*. Mammalian Species. 2003;716:1-10.
24. Cassola, F. *Microtus montanus*. The IUCN Red List of Threatened Species 2016.
25. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology. 2016;34:3:303-11.
26. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome research. 2017;27:5:757-67.
27. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:1:2047-17.
28. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, Nusbaum C. Comprehensive variation discovery in single human genomes. Nature genetics. 2014;46:12:1350.
29. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ .

30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:15:2114-20.
31. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30:1:31-7.
32. Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 2018;34:5:725-31.
33. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. GigaScience. 2015;4:1:s13742-015.
34. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:5:589-95.
35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:16:2078-9.
36. Ratnasingham S, Hebert PD. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Molecular ecology notes. 2007;7:3:355-64.
37. Alqahtani F, Duckett D, Pirro S, Mandoiu II. Complete mitochondrial genome of the water vole, *Microtus richardsoni* (Cricetidae, Rodentia). Mitochondrial DNA Part B. 2020;5:3:2498-9.Broad Institute. Picard tools. 2016. http://github.com/broadinstitute/picard.
38. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:21:2987-93.
39. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome biology. 2019;20:1:1-7.

40. Modi WS. Phylogenetic analyses of chromosomal banding patterns among the Nearctic Arvicolidae (Mammalia: Rodentia). Systematic Zoology. 1987;36:2:109-36.
41. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27:12:1691-2.
42. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic acids research. 2012;41:D1:D36-42.
43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:8:1072-5.
44. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014. https://jgi.doe.gov/data-and-tools/bbtools/.
45. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:19:3210-2.
46. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:13:i142-50.
47. Fernandez-Silva I, Henderson JB, Rocha LA, Simison WB. Whole-genome assembly of the coral reef Pearlscale Pygmy Angelfish (Centropyge vrolikii). Scientific reports. 2018;8:1:1-1.
48. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research. 2017;45:4:e18-.
49. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:12:1647-9.

50. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32:5:1792-7.

51. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of molecular evolution. 1994;39:2:174-90.

52. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (nuMountains) in sequenced nuclear genomes. PLoS Genet. 2010;6:2:e1000834.

53. Triant DA, DeWoody JA. Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications. Gene. 2007;401:1-2:61-70.

54. Triant DA, DeWoody JA. Molecular analyses of mitochondrial pseudogenes within the nuclear genome of arvicoline rodents. Genetica. 2008;132:1:21-33.

55. Triant DA, DeWoody JA. Demography and phylogenetic utility of numt pseudogenes in the Southern Red-Backed Vole (*Myodes gapperi*). Journal of mammalogy. 2009;90:3:561-70.

56. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution. 2013;69:2:313-9.

57. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008. http://www.repeatmasker.org .

58. Smit AF, Hubley R, Green P. RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org .

59. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research. 2008;18:1:188-96.

60. Young RL, Ferkin MH, Ockendon-Powell NF, Orr VN, Phelps SM, Pogány Á, Richards-Zawacki CL, Summers K, Székely T, Trainor BC, Urrutia AO. Conserved transcriptomic profiles underpin

monogamy across vertebrates. Proceedings of the National Academy of Sciences. 2019;116:4:1331-6.

61. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004;5:1:59.
62. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19:ii215-25.
63. Araujo FA, Barh D, Silva A, Guimarães L, Ramos RT. GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data. Scientific reports. 2018;8:1:1-4.
64. UniProt Consortium. UniProt: a hub for protein information. Nucleic acids research. 2015;43:D1:D204-12.
65. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD. InterPro: the integrative protein signature database. Nucleic acids research. 2009;37:D211-5.
66. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ. The Pfam protein families database. Nucleic acids research. 2004;32:D138-41.
67. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research. 2020;9.
68. Voigt O, Erpenbeck D, Wörheide G. A fragmented metazoan organellar genome: the two mitochondrial chromosomes of Hydra magnipapillata. BMC genomics. 2008;9:1:350.
69. Smith DR. Not seeing the genomes for the DNA. Briefings in functional genomics. 2012;11:4:289-90.
70. Fink S, Excoffier L, Heckel G. Mitochondrial gene diversity in the common vole Microtus arvalis shaped by historical divergence and local adaptations. Molecular Ecology. 2004;13:11:3501-14.

71. Fischer MC, Foll M, Excoffier L, Heckel G. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (Microtus arvalis). Molecular Ecology. 2011;20:7:1450-62.
72. Frey JK. Genetics of allopatric populations of the montane vole (Microtus montanus) and Mogollon vole (Microtus mogollonensis) in the American Southwest. Western North American Naturalist. 2009;69:2:215-22.
73. Tołkacz K, Alsarraf M, Kowalec M, Dwużnik D, Grzybek M, Behnke JM, Bajer A. Bartonella infections in three species of *Microtus*: prevalence and genetic diversity, vertical transmission and the effect of concurrent *Babesia microti* infection on its success. Parasites & vectors. 2018 Dec 1;11(1):491.
74. Silvestro D, Michalak I. raxmlGUI: a graphical front-end for RAxML. Organisms Diversity & Evolution. 2012;12:4:335-7.