Greenhalgh Robert (Orcid ID: 0000-0003-2816-3154)
Holding Matthew L. (Orcid ID: 0000-0003-3477-3012)
Parchman Thomas (Orcid ID: 0000-0003-1771-1514)

Trio-binned genomes of the woodrats *Neotoma bryanti* and *Neotoma lepida* reveal novel gene islands and rapid copy number evolution of xenobiotic metabolizing genes

Running title: Xenobiotic metabolism shapes *Neotoma* genomes

Author list: Robert Greenhalgh[1*], Matthew L. Holding[2a], Teri J. Orr[1b], James B. Henderson[3], Thomas L. Parchman[4], Marjorie D. Matocq[2], Michael D. Shapiro[1] and M. Denise Dearing[1]

[1]School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, Utah 84112, USA

[2]Department of Natural Resources & Environmental Science, University of Nevada, Reno, 1664 North Virginia Street, Reno, Nevada 89775, USA

[3]Center for Comparative Genomics, California Academy of Sciences, 55 Music Concourse Drive, San Francisco, California 94118, USA

[4]Department of Biology, University of Nevada, Reno, 1664 North Virginia Street, Reno, Nevada 89775, USA

[*]Corresponding author: robert.greenhalgh@utah.edu

[a]Present address: Life Sciences Institute, University of Michigan, 210 Washtenaw Avenue, Ann Arbor, Michigan 48109, USA

[b]Present address: Department of Biology, New Mexico State University, 1780 East University Avenue, Las Cruces, New Mexico 88003, USA

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13650

Conflict of interest statement: The authors have no conflicts of interest to disclose.

Abstract

The genomic architecture underlying the origins and maintenance of biodiversity is an increasingly accessible feature of species, due in large part to third-generation sequencing and novel analytical toolsets. Applying these techniques to woodrats (*Neotoma* spp.) provides a unique opportunity to study how herbivores respond to environmental change. *Neotoma bryanti* and *N. lepida* independently achieved a major dietary feat in the aftermath of a natural climate change event: switching to the novel, toxic food source creosote bush (*Larrea tridentata*). To better understand the genetic mechanisms underlying this ability, we employed a trio binning sequencing approach with a *N. bryanti* × *N. lepida* F_1 hybrid, allowing the simultaneous assembly of genomes representing each parental species. The resulting phased, chromosome-level, highly complete haploid references enabled us to explore the genomic architecture of several gene families — cytochromes P450, UDP-glucuronosyltransferases (UGTs), and ATP-binding cassette (ABC) transporters — known to play key roles in the metabolism of naturally occurring toxic dietary compounds. In addition to duplication events in the ABCG and UGT2B subfamilies, we found expansions in three P450 gene families (2A, 2B, 3A), including the evolution of multiple novel gene islands within the 2B and 3A subfamilies, which may have provided the crucial substrate for dietary adaptation. Our assemblies demonstrate that trio binning from an F_1 hybrid rodent effectively recovers parental genomes from species that diverged more than a million years ago.

Key words

Adaptive evolution, dietary adaptation, functional genomics, genome sequencing, metabolic adaptation, xenobiotic metabolism.

Introduction

At every meal, mammalian herbivores confront the possibility of being poisoned by their food (Dearing et al., 2005). In their seminal work on dietary strategies, Freeland and Janzen (Freeland & Janzen, 1974) predicted that liver biotransformation enzymes dictated detoxification ability and, therefore, diet breadth in herbivores. Despite decades of pharmacological research on laboratory species and humans, we know relatively little about the mechanisms used by mammalian herbivores to metabolize plant toxins or adapt to new diets (Dearing et al., 2005; Foley et al., 1999; Marsh et al., 2006; Moore et al., 2015; Sorensen et al., 2006; Sorensen & Dearing, 2006). Nevertheless, this fundamental information on how mammalian herbivores metabolize dietary toxins is critical in understanding the evolution of dietary shifts that entail exploiting new food sources, or shifts that are imposed by flora changing in response to natural and anthropogenic climate change. By studying woodrats (genus *Neotoma*), which include many species that have weathered environmental change (F. A. Smith et al., 1995) and adapted their diets to include novel toxic plants (Nielsen & Matocq, 2021), we can gain insight into both of these areas.

Woodrats are the sister genus to *Peromyscus* (Figure 1A), and are endemic to North and Central America (Hall, 1982). Though most *Neotoma* species have distinct ranges, in some areas, closely related species come into contact and produce fertile hybrids (Coyner et al., 2015; Patton et al., 2007; Shurtliff et al., 2014). Around 17,000 years ago, the region that is now the southwestern United States experienced a temperature increase as the Late Glacial Cold Stage ended. This caused substantial disruption to the vegetation throughout the region and resulted in the establishment of the arid deserts of the American Southwest (Van Devender, 1977; Van Devender & Spaulding, 1979; Wells & Woodcock, 1985). In the wake of this climate change event, juniper (*Juniperus* spp.) (Figure 1B) populations declined, and large portions of their

former range were replaced with what ultimately became the dominant shrub of the region: creosote bush (*Larrea tridentata*) (Figure 1C) (Van Devender, 1977; Van Devender & Spaulding, 1979; Wells & Hunziker, 1976). This vegetation shift forced resident woodrat species to switch from a diet of juniper and cactus to one containing significant quantities of creosote. These food sources have radically different plant secondary compounds and presumably exerted disparate selective forces on the hepatic detoxification systems of *Neotoma bryanti* (Figure 1D) and *N. lepida* (Figure 1E) (Dearing et al., 2005; Holchek et al., 1990; Mabry et al., 1977). If selection for this novel diet was facilitated by adaptation, then evidence for this is likely to be found in the overall structure of the xenobiotic metabolizing genes in both species.

Identifying and characterizing the putative genetic basis of dietary adaptations demands high-quality and annotated reference genomes; however, these resources were lacking for *Neotoma* woodrats. Prior to this study, a draft genome for *N. lepida* (Campbell et al., 2016) was the sole genomic reference available for the genus, and, while nearly complete, it is highly fragmented (scaffold N50 of 151 kb), thereby limiting genetic analyses. To address this deficiency, we generated new genomic resources for both *N. lepida* and *N. bryanti.* These two woodrat species are sister taxa with an estimated 1.5-million-year divergence time (Patton et al., 2007). The desert woodrat (*N. lepida*) is found across many habitats in the western United States, from as far north as Idaho and Oregon down through Utah and eastern California. In contrast, Bryant's woodrat (*N. bryanti*) has a smaller range primarily in the coastal woodlands and chaparrals of southern California, with some populations extending into desert habitats (Patton et al., 2007). Both species have populations that feed on creosote bush (Weinstein et al., 2021). The ranges of these two species are largely distinct, but in the few locations where they overlap (e.g., Morongo Valley, California), hybrids occur (Patton et al., 2007). We leveraged the ability of these two species to hybridize and applied the recently developed trio binning approach implemented in Canu (Koren et al., 2018) to an F_1 cross between *N. bryanti* and *N.*

lepida in order to simultaneously generate haploid reference genomes for both species. This approach combines PacBio long read sequencing of the hybrid individual with Illumina short read sequencing of the parental individuals. *k*-mers identified in the Illumina read data are used to bin the hybrid reads by parent-of-origin, and each parental haplotype is then assembled separately from the binned PacBio reads. By using this method, we avoided some of the issues that plague the reconstruction of diploid vertebrate genomes, while also keeping sequencing and computational costs low. Although this technique has previously been used to generate genome assemblies for hybrid species within the Bovidae (Rice et al., 2020) and Felidae (Bredemeyer et al., 2020) families, this marks the first time this approach has been used to generate genomes for two rodent species and for two species of naturally hybridizing mammals.

The resulting reference genomes are assembled to the chromosome level, and their sequence content and annotations are highly complete by multiple metrics. To better understand the genomic basic of adaptation to different toxic diets, we interrogated the genetic architecture and possible expansion dynamics of several xenobiotic metabolizing gene families, with a particular focus on the cytochromes P450 (henceforth P450) 2A, 2B, and 3A genes. The P450 enzymes produced by these genes are critical in the biotransformation of toxins like those present in creosote bush and juniper (Huo et al., 2017; Magnanou et al., 2009; Shah et al., 2016; Skopec et al., 2013; Wilderman et al., 2014). Novel dietary niches involving toxic food are hypothesized to exert strong selection pressure for herbivore defense against potential toxic effects (Dearing et al., 2005), and previous work in these woodrats and other herbivorous mammals suggests that the evolution of gene copies, specifically those in P450 subfamily 2B, play an important role in dietary adaptation (Kitanovic et al., 2018; Magnanou et al., 2009; Malenke et al., 2012). Therefore, we expected to document copy number expansion of the 2B and potentially other P450 gene subfamilies in comparison to other rodents.

Materials and Methods

Animal collection and handling

The female *N. bryanti* used in the F_1 cross was captured in March 2018 near Pioneertown, California, USA (lat. 34.151, long. −116.479). The male *N. lepida* was captured in October 2018 near Kelso Depot, Mojave Preserve in California, USA (lat. 35.009, long. −115.645). *Neotoma lepida* individuals (N = 4) used for the transcriptomics portion of the work were captured in March 2018 in Lytle Ranch Preserve, Utah, USA (lat. 37.070, long. −114.000). Woodrats were live-trapped at their respective sites using 7.6 × 8.9 × 22.9 cm Sherman traps (H. B. Sherman Traps, Inc., Tallahassee, Florida, USA). The *N. lepida* captured near Kelso Depot was obtained under the California Department of Fish and Wildlife permit number SC-2105, MOJA-00038 issued to James Patton, while all other animals were trapped under the California Fish and Game permit SCP-8123 and Utah Game and Fish permit 1COLL5194-2 issued to M. Denise Dearing. Animal work was approved under the University of Utah's Institutional Animal Care and Use Committee (Protocol 16-02011).

After transport to the University of Utah's School of Biological Sciences animal facility, animals were housed individually (except for mating) in 48 × 27 × 20 cm shoebox cages. The facility was maintained at 24°C, 15–20% relative humidity and with a 12-hour light/dark cycle. Animals were provided rabbit chow (Teklad formula 2031; Envigo, Indianapolis, Indiana, USA) and water *ad libitum*. The *N. bryanti* × *N. lepida* cross yielded three hybrid individuals (two females, one male) in March 2019, and one female from this litter was selected for sequencing to ensure the X chromosome could be reconstructed for each haplotype.

Animals were dispatched prior to dissection. For DNA sequencing, liver tissue was extracted and stored at −80°C. For RNA sequencing, liver tissue was placed in RNAlater (Thermo Fisher Scientific, Waltham, Massachusetts, USA) prior to storage at −80°C.

Genomic library preparation and sequencing

For the F_1 parents, DNA was extracted from liver tissue using the DNeasy Blood & Tissue Kit (QIAGEN, Germantown, Maryland, USA); the concentration (~70 ng/μL) was validated using a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA). Following shearing with a Covaris S2 Focused-ultrasonicator (Covaris, Woburn, Massachusetts, USA), libraries with an average insert size of 450 bp were prepared by the University of Utah's Huntsman Cancer Institute High-Throughput Genomics Core Facility using the Illumina TruSeq DNA PCR-Free Library Prep Kit (Illumina, San Diego, California, USA). 2 × 150 bp reads for each parent were generated on an Illumina NovaSeq 6000 instrument using the NovaSeq S2 reagent kit (Illumina, San Diego, California, USA). 67.44× and 81.10× coverage for the *N. bryanti* mother and *N. lepida* father, respectively, were obtained based on a genome size estimate of 2.4 Gb derived from the previous *N. lepida* ASM167557v1 assembly (Campbell et al., 2016).

For the hybrid offspring, DNA extraction, library preparation and sequencing were carried out by the DNA Sequencing Center at Brigham Young University. Following DNA extraction from liver tissue with the QIAGEN MagAttract HMW DNA Kit (QIAGEN, Germantown, Maryland, USA), libraries were generated using the SMRT Bell Template Preparation Kit (Pacific Biosciences, Menlo Park, California, USA), and sequencing was performed using eight SMRT Cells on a PacBio Sequel II instrument (Pacific Biosciences, Menlo Park, California, USA). This yielded 787.76 Gb of sequence data, corresponding to approximately 328× coverage.

RNA library preparation and sequencing

RNA was extracted from liver tissue using the QIAGEN RNeasy Mini Kit (QIAGEN, Germantown, Maryland, USA). The concentration (~75 ng/μL) and RNA quality (RIN ≥ 9.0) for each sample was validated with a Qubit RNA BR Assay Kit (Thermo Fisher Scientific, Waltham,

Massachusetts, USA) and Agilent Technologies RNA ScreenTape Assay (Agilent Technologies, Santa Clara, California, USA), respectively, before library preparation and sequencing was carried out by the University of Utah's Huntsman Cancer Institute High-Throughput Genomics Core Facility. Libraries were generated using the Illumina TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, California, USA), and 2 × 150 bp reads were produced on an Illumina NovaSeq 6000 instrument using the NovaSeq S2 reagent kit (Illumina, San Diego, California, USA). For sample counts, see Table S1.

Mitochondrial genome assembly

Illumina reads for each parent were imported into CLC Genomics Workbench v.12 (https://digitalinsights.qiagen.com) and trimmed and assembled using the default settings of the "Trim Reads" and "De novo assembly" tools, respectively. Sequences were aligned with BLAT v.36 (Kent, 2002) against the *Rattus norvegicus* Rnor 6.0 mitochondrial genome (Rat Genome Sequencing Project Consortium, 2004) to identify scaffolds corresponding to the mitochondrion. For each parent, the mitochondrion appeared to be contained in a single sequence, and subsequent BLASTN (Camacho et al., 2009) searches against the NCBI Nucleotide database (NCBI Resource Coordinators, 2016) on December 3, 2020 confirmed these findings.

Read partitioning and haplotype assembly

PacBio reads were converted to the FASTA format using bam2fasta v.1.3.0 (https://github.com/PacificBiosciences/bam2fastx). Employing *k*-mers identified in the parental Illumina read data, Canu v.1.9 (Koren et al., 2018) was used to partition the PacBio reads of the hybrid by haplotype (and therefore species). The resulting species-specific reads were assembled separately using the default settings of Canu with an estimated genome size of 2.4 Gb.

Genome polishing

BAM files containing species-specific reads identified by Canu were generated from the original PacBio subread BAMs and indexed using pbindex v.1.0.7 (https://github.com/PacificBiosciences/pbbam) to create input files for the pbmm2 aligner. Using the default settings of pbmm2 v.1.2.1 (https://github.com/PacificBiosciences/pbmm2; H. Li, 2018), the species-specific reads were aligned against the appropriate Canu reference. Arrow v.2.3.3 (https://github.com/PacificBiosciences/GenomicConsensus) was run using default settings to polish each assembly; due to the high memory requirements of this program, the pbmm2 BAMs and Canu reference FASTAs were split into individual files by contig, Arrow was run on each contig individually, and the resulting FASTAs were combined into a single file once the polishing was complete. These polished assemblies were subjected to a second round of pbmm2 and Arrow to generate the final contigs.

Identification and masking of repetitive elements

RepeatModeler v.2.0.1 (Flynn et al., 2020) was run with the LTR structural search pipeline on each polished assembly to generate species-specific repeat libraries. Utilizing these libraries as well as the RepeatMasker v.20181026 libraries from Repbase (Jurka, 1998), RepeatMasker v.4.0.9 (https://www.repeatmasker.org) was used to mask repetitive elements in each assembly. As the resulting FASTAs were to be used for automated gene annotation, RepeatMasker was run with the "-nolow" argument to avoid masking low complexity regions that could potentially comprise parts of genes (Tables S2 and S3).

Processing and alignment of transcriptomic reads

Transcriptomic reads were supplemented with the RNA-seq reads used to annotate the previous ASM167557v1 draft release of the *N. lepida* genome (Campbell et al., 2016). Quality control and trimming of all RNA-seq reads was performed with Trim Galore! v.0.65 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which utilized FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Cutadapt v.2.10 (Martin, 2011). 99.55% of the reads were retained following Trim Galore! processing. Processed reads were aligned to each genome using the two-pass mode of STAR v.2.5.7a (Dobin et al., 2013), and merged into a single BAM file with SAMtools v.1.9 (H. Li et al., 2009). See Table S1 for mapping statistics.

Automated gene prediction

Gene models on the repeat-masked references were predicted with BRAKER v.2.1.5 (Hoff et al., 2019). The STAR RNA-seq alignments, as well as all protein sequences from the OrthoDB v10 database (Kriventseva et al., 2019), were supplied to the BRAKER command line arguments "--bam" and "--prot_seq", respectively, and the application was run using the "--etpmode" and "--softmasking" arguments. Gene hints from the protein sequences and RNA-seq alignments were identified by the GeneMark-ES v.4.59 suite (Brůna et al., 2020), and a subset of these predictions determined to be high quality based on the supporting OrthoDB and transcriptome evidence were selected by BRAKER for training. After aligning these predictions against each other with DIAMOND v.2.0.0 (Buchfink et al., 2015) to remove redundant models, these genes were used to train AUGUSTUS v.3.3.3 (Stanke et al., 2004) parameters for each species prior to gene prediction. Custom Python scripts were used to reformat the resulting BRAKER GFF3 for sorting and validation with GenomeTools v.1.6.1 (Gremme et al., 2013).

Gene model filtering and functional annotation

Protein sequences were obtained from each species' BRAKER GFF3 and polished reference FASTA with scripts utilizing the Biopython v.1.76 libraries (Chapman & Chang, 2000). InterProScan v.5.51-85.0 (Jones et al., 2014), was run on each of these sequences, and all gene predictions lacking an InterPro group ID were removed from the BRAKER annotations. The filtered GFF3s were analyzed and processed by the GFF3toolkit v.2.0.3 (Chen et al., 2019) to resolve incomplete and spuriously duplicated models. All remaining protein sequences were aligned with BLASTP v.2.11.0+ (Camacho et al., 2009) against a database containing protein sequences from *Homo sapiens* release GRCh38 (Schneider et al., 2017), *Mus musculus* GRCm38.p6 (Mouse Genome Sequencing Consortium, 2002), *Peromyscus leucopus* PerLeu 2.1 (Long et al., 2019), *Peromyscus maniculatus* Pman 2.1 and *Rattus norvegicus* Rnor 6.0 (Rat Genome Sequencing Project Consortium, 2004); with the exception of *P. leucopus*, all protein sequences were retrieved from Ensembl (Yates et al., 2019). When a protein prediction had a hit with an E-value under 10^{-5}, a description and gene name were retrieved from the corresponding annotation and added to the BRAKER GFF3 alongside all relevant InterPro group IDs and Gene Ontology categories.

Annotation of P450 genes

Several sources of information were leveraged to identify and curate the P450 genes in each genome. We produced a set of gene models in BITACORA v.1.3 (Vizueta et al., 2020), which combines Hidden Markov Model (HMM) protein profiles and contig aware linear searches to annotate tandemly arrayed genes, using the default BITCORA settings and the Pfam PF00067 HMM profile. Additionally, we used the P450 amino acid sequence dataset from Thomas (Thomas, 2007) to query our genomes using TBLASTN searches (v.2.11, E-value = 10^{-5}). Finally, we used HISAT2 v.2.2.1 (Kim et al., 2019) to conduct splice-aware RNA-seq alignments to each genome using the "--no-unal --no-mixed --no-discordant" options, and filtered the

resulting alignment map files in SAMtools v.1.9 (H. Li et al., 2009) to retain only the top alignment for each read pair. An additional GFF3 was also generated by aligning CDS sequences from the five animal genomes employed for functional annotation to each of the woodrat references using GMAP v.2020-06-01 (Wu & Watanabe, 2005) with "--min-trimmed-coverage" set to 0.8. The GMAP GFF3 was processed with the GFF3toolkit (Chen et al., 2019) and sorted and validated with GenomeTools (Gremme et al., 2013).

Manual curation of cytochromes P450

The contigs, BLAST hits, RNA-seq alignments, and various gene model annotations were loaded into Geneious v.10.2.6 (https://www.geneious.com) for manual curation. The P450 models from the combined BRAKER, GMAP, and BITCORA annotations were used for producing a final gene model with the following characteristics: adjacent exons were bounded by AG/GT splice sites, exons included appropriate reading frames, and internal stop codons were absent. When RNA-seq data covered the entire model, the splice sites supported by the RNA-seq alignments were given priority. Putative P450s with early stops, or those with largely incomplete coding sequences based on TBLASTN searches, were marked as pseudogenes. Pseudogene designations included any BLAST hit that was near a functional gene or P450 island, and ranged from essentially complete sequences to hits matching single exons.

Inference of syntenic P450 gene islands and copy number differences

Using the Geneious v.10.2.6 viewer, we visualized the genomic organization of P450 2ABFGST and 3A genes for both *Neotoma* as well as an Old World rat (*R. norvegicus*) and New World *Peromyscus* mice, which are the sister genus to *Neotoma* (*P. maniculatus* for the CYP2ABFGST genes and *P. leucopus* for the CYP3A genes were chosen based on the completeness of the clusters in each species). Inference of expansions in *Neotoma* represents the most conservative

comparison that can be made to its sister genus based on available resources. We also visualized nearby non-P450 genes to help establish regional synteny. To derive gene trees, the amino acid sequences of all functional genes within each subfamily were aligned in MAFFT v.7.45 (Katoh & Standley, 2013) using the G-INS-I algorithm and BLOSUM62 scoring matrix. We then used the "backtrans" function in TreeBeST v.1.9.2 (http://treesoft.sourceforge.net) to back translate the amino acid alignment to a protein-guided codon alignment of nucleotide sequences, and finally used the TreeBeST "best" function guided by the species tree (downloaded from http://www.timetree.org) with nodal support derived from 100 bootstraps. The models of each gene island and gene trees were then edited in Adobe Illustrator (Adobe Inc., San Jose, California, USA), using both gene order and clustering in the gene tree to call putatively syntenic P450 clusters.

Resolution of the genomic region containing the CYP2B array in *N. lepida*

While the CYP2B array was assembled as a single contig in the *N. bryanti* Canu assembly, the *N. lepida* array was significantly more fragmented, with CYP2B sequences identified across seven contigs. To resolve the region in *N. lepida*, haplotype-specific PacBio reads were aligned to the *N. lepida* assembly with minimap2 v.2.17-r941 (H. Li, 2018), and only those mapping to the seven identified contigs were retained. Two Canu v.1.9 (Koren et al., 2018) assemblies were constructed with these reads: the first was generated using all reads (hereafter "AR"; genome size was left at 2.4 Gb and stopOnLowCoverage was set to 0.7), while the second was generated using Canu's default coverage settings (hereafter "40×"; genome size was set to 19 Mb, the approximate size of the seven CYP2B-containing contigs). Each of these assemblies was polished with two rounds of pbmm2 v.1.2.1 (https://github.com/PacificBiosciences/pbmm2; H. Li, 2018) and Arrow v.2.3.3 (https://github.com/PacificBiosciences/GenomicConsensus).

The 40× assembly appeared to resolve the entire region into two primary contigs: one 6.58 Mb and another 11.98 Mb in length. Though more fragmented overall, the AR assembly contained a single contig 13.63 Mb in length that, based on MUMmer v.3.23 (Kurtz et al., 2004) alignments (Figure S1A, B), contained portions of both contigs from the 40× assembly. Using quickmerge v.0.3 (Chakraborty et al., 2016) with "--length_cutoff" set to 1,000,000, the two 40× contigs were aligned against the 13.63-Mb AR contig to generate a single sequence 18.55 Mb in length. To polish this sequence, reads were first aligned with minimap2 to the AR and 40× assemblies and only those mapping to the three contigs of interest were retained; these reads were then used for two rounds of polishing with pbmm2 and Arrow. Based on MUMmer alignments, the resulting sequence resolved the order and orientation of the largest four contigs in the initial Canu assembly (Figure S1C). The remaining three contigs could not be definitively placed and appeared to be repetitive sequences possibly arising from unresolved sequencing errors (Figure S1D). RepeatMasker v.4.0.9 (https://www.repeatmasker.org) and the *N. lepida* RepeatModeler v.2.0.1 (Flynn et al., 2020) library were used to mask repetitive elements on this sequence as previously described, and gene models from the initial assembly were transferred to this sequence using Liftoff v.1.6.1 (Shumate & Salzberg, 2021), with the manually curated P450s transferred using Geneious v.10.2.6 (https://www.geneious.com).

Annotation and curation of additional xenobiotic metabolizing families and transcription factors

For the annotation of the ATP-binding cassette transporter, flavin-containing monooxygenase, glutathione-*S*-transferase, sulfotransferase and UDP-glucuronosyltransferase genes, along with the aryl hydrocarbon, pregnane X and constitutive androstane receptors (respectively encoded by the *AHR* and nuclear hormone receptor 1I2 and 1I3 genes — *NR1I2* and *NR1I3*), complete rodent protein sequences belonging to each family or transcription factor were obtained from the NCBI Protein database on September 23, 2021; sequences flagged as "partial" or "low quality"

were omitted. Using the default settings of NCBI TBLASTN v.2.9.0 (Camacho et al., 2009) and GenomeThreader v.1.7.1 (Gremme et al., 2005), rodent sequences were aligned to each *Neotoma* genome to provide additional sources of evidence. Gene models from the *Mus musculus* and *Rattus norvegicus* annotations were also mapped to each *Neotoma* genome using the default settings of Liftoff v.1.6.1 (Shumate & Salzberg, 2021) to assist with annotation. The STAR RNA-seq alignments — which were converted to strand-specific bigWig files with deepTools 3.5.1 (Ramírez et al., 2014) — along with the protein alignments, Liftoff models, BRAKER predictions and GMAP GFF3s were loaded into Apollo 2.6.5 (Dunn et al., 2019) for manual curation. Genes were required to have intact coding frames, priority was given to splice sites supported by the RNA-seq data (provided the entire model had RNA-seq coverage), and models with partial or disrupted coding sequences were marked as pseudogenes. Homology data were additionally sourced from the Rat Genome Database (J. R. Smith et al., 2019) on November 11, 2021 to compile species counts for each gene family.

Assessing genome completeness

Completeness for both genomes was assessed using the "genome" mode of BUSCO against the Glires_odb10 dataset of 13,798 genes. We ran BUSCO v.4.1.4 (Seppey et al., 2019), which uses the gene predictions of both AUGUSTUS and TBLASTN, and the newer BUSCO v.5.0.0, which employs MetaEuk (Levy Karin et al., 2020). The results of these searches were largely overlapping, but some genes were found by only one version. We combined the gene lists from each run to produce our final BUSCO metrics (Table 1).

Jellyfish v.2.3.0 (Marçais & Kingsford, 2011) was used to tabulate *k*-mer counts of length 25 in the trimmed and corrected reads generated for each of the *Neotoma* assemblies. After discarding low count *k*-mers (those that appeared less than 8 times for *N. bryanti* and 11 times for *N. lepida*), the genome size was estimated by multiplying each count value by the total

number of *k*-mers present at that value, then dividing that number by the *k*-mer count value that appeared most frequently (24 for *N. bryanti* and 30 for *N. lepida*). Plots of the *k*-mer distributions for *N. bryanti* and *N. lepida* are shown in Figure S2.

Genome scaffolding and syntenic analysis

Contigs for both woodrats were aligned to the chromosome scaffolds of *Peromyscus leucopus* (Long et al., 2019), *P. maniculatus* and *P. nasutus* (https://www.dnazoo.org/assemblies/Peromyscus_nasutus), the closest relatives for which chromosome-level assemblies were available. Chromosome numbering and orientation are based on the *P. maniculatus* reference, with the corresponding chromosome sequences in *P. leucopus* and *P. nasutus* determined by MUMmer v.3.23 (Kurtz et al., 2004) alignments of those assemblies against the *P. maniculatus* reference. Due to the potential for misalignment due to repetitive elements, only *Neotoma* contigs exceeding 1 Mb in length with at least 100 kb of sequence uniquely aligned to a single chromosome were used in the scaffolding process. Contigs were aligned with the default settings of MUMmer v.4.0.0 beta 2 (Marçais et al., 2018), and assigned to the chromosome to which they had the greatest amount of sequence uniquely aligned. The orders and orientations of contigs along each chromosome were determined by the median position of their alignments, and were manually adjusted if discontinuities were evident when the MUMmer alignments were visualized (these discontinuities were almost universally large-scale inversions relative to the *Peromyscus* references). Once preliminary chromosome scaffolds were generated, the *N. bryanti* and *N. lepida* sequences were aligned against each other for validation as well as to correct any outstanding contig placement issues. For both species, a total of 82 joins between adjacent contigs were made for the final chromosome scaffolds. These joins, each denoted by a stretch of 500 N characters, account for the 41,000 bp of gap sequence present in each assembly (Table 1). Gene models were transferred to the new

coordinates using custom scripts employing the Biopython v.1.76 libraries (Chapman & Chang, 2000).

After scaffolding, syntenic regions between each *Neotoma* genome and *P. leucopus* and *P. maniculatus* were identified using BLASTALL v.2.2.26 (Altschul et al., 1990) protein alignments with an E-value of 10^{-20} and the default settings of Synima (GitHub release a0bc445) (Farrer, 2017). MUMmer plots against the *Peromyscus* genomes are available for all *N. bryanti* chromosomes as Figures S3–S5, and all *N. lepida* chromosomes as S6–S8. MUMmer plots between the *Neotoma* chromosome scaffolds are available in Figures S9 and S10. Information on the relative order and orientation of contigs for each chromosome is contained in Table S4.

Estimating nuclear and mitochondrial divergence

The *N. bryanti* and *N. lepida* chromosome scaffolds were aligned to each other using the NUCmer program of MUMmer v.3.23 (Kurtz et al., 2004), and delta-filter was run on the alignments to retain only regions that had one-to-one mappings between both species. Variants in these regions were then called using the MUMmer show-snps program, and metrics were calculated using a custom Python script. The same process was performed for the mitochondrial scaffolds. Complete statistics for both the nuclear and mitochondrial genomes are contained in Table S5.

Results and Discussion

Read binning produces highly contiguous and complete assemblies

The read lengths provided by recent advances in sequencing technology, combined with the trio binning approach, yielded an almost perfect separation of parent-of-origin sequencing data from the F_1 hybrid. After removing all PacBio reads considered too short (<1 kb, which accounted for

1.97 Gb, or 0.25% of the total), Canu assigned 49.46% (389.60 Gb) of reads to the maternal *N. bryanti* haplotype and 50.16% (395.14 Gb) to the paternal *N. lepida* haplotype. Only 0.13% (1.05 Gb) was unassignable. This marks a slight advancement over that reported for trio binning within *Bos taurus* (49.3% and 49.6% assigned to each haplotype, respectively) (Koren et al., 2018), and is in line with the results reported for the recent cross-species *Bos* (Rice et al., 2020) and feline (Bredemeyer et al., 2020) assemblies. The highly accurate partitioning of trio binning yielded minimal loss of usable sequencing data, further demonstrating the promise this technique holds for generating genomes from hybridized taxa.

Though well over 2,000 contigs were produced for each woodrat reference, the N50 values indicate that the vast majority of each assembly lies within large, contiguous sequences (Table 1). Furthermore, although less than 5% of contigs exceeded 1 Mb (110 for *N. bryanti* and 107 for *N. lepida*), these sequences contained over 91% of each 2.6-Gb assembly. Of particular note were the extremely long contigs generated for each species; with lengths exceeding 100 Mb, these sequences likely represent most (or all) of a chromosome.

BUSCO analysis of the assemblies indicated that they were remarkably complete, with *N. bryanti* containing 97.62% and *N. lepida* possessing 97.63% of the 13,798 single-copy orthologs present in the Glires 10 dataset (Table 1). Additionally, both assemblies contained few partial or multi-copy BUSCOs, indicating that fragmentation and sequence duplication were minimal. We investigated this small number of multi-copy BUSCOs to see if they contained an enrichment of xenobiotic metabolizing genes, but found relatively few in this category. This is not surprising, however; BUSCO is designed to locate highly-conserved, single-copy orthologs (Seppey et al., 2019), and as many xenobiotic metabolizing subfamilies have members with a high degree of sequence similarity — a feature which complicates accurate ortholog identification — their underrepresentation in the BUSCO databases is not unexpected. Further bolstering the completeness of the references, the total assembly lengths were similar to size

estimates based on *k*-mer counts from the PacBio sequencing data (2.72 Gb and 2.62 Gb for *N. bryanti* and *N. lepida*, respectively), with the assembled and estimated values differing by less than 5%. Together, these results indicate that the vast majority of the nuclear genome for both species is contained in the assemblies.

Similar results were found for the mitochondrial genomes reconstructed from the parental short read data. Each mitochondrial scaffold — 16.7 kb for *N. bryanti* and 16.3 kb for *N. lepida* — was very close in size to the 16.3-kb mitochondrial assemblies for *R. norvegicus* (Rat Genome Sequencing Project Consortium, 2004) and *M. musculus* (Mouse Genome Sequencing Consortium, 2002). These completeness findings were further strengthened with BLASTN (Camacho et al., 2009) searches against the NCBI Nucleotide database (NCBI Resource Coordinators, 2016), which found query coverages of 97% and 99% for *N. bryanti* and *N. lepida*, respectively, for the closely related *N. magister* and *N. mexicana* mitochondrial genomes.

Scaffolding indicates most chromosomes were assembled from few contigs

As no chromosome-level *Neotoma* assemblies exist, we used alignments against the genomes of *Peromyscus leucopus*, *P. maniculatus* and *P nasutus* to scaffold our contigs into chromosomal sequences. Using the 24 chromosome scaffolds from the *Peromyscus* assemblies as references, we scaffolded 96.36% of contigs exceeding 1 Mb for *N. bryanti* (106/110), and 99.07% (106/107) for *N. lepida* (Figure 2A, B). The scaffolded chromosomes resulting from this approach contained the vast majority of each species' genome, comprising 91.86% and 91.85% of the *N. bryanti* and *N. lepida* assemblies, respectively (Table 1). Though the number of chromosomes in *Neotoma* (1N = 26 chromosomes in *N. lepida*) may be slightly higher than in *Peromyscus* (Mascarello & Hsu, 1976), given the close phylogenetic proximity of these genera, our scaffolding approach based on the 24 *Peromyscus* chromosomes provides a reasonable approximation of each *Neotoma* species' chromosomal architecture.

Among the assembled chromosomes, five were composed of single contigs; these were chromosomes 4 and 19 for *N. bryanti* and 12, 13 and 19 for *N. lepida* (Figure 2A, B). Further, of the chromosomes scaffolded for both species, the vast majority (23/24 for *N. bryanti* and 22/24 for *N. lepida*) were assembled from eight or fewer contigs — only the X chromosome for *N. bryanti* (23 contigs) and chromosomes 1 and X for *N. lepida* (12 and 17 contigs, respectively) were composed of more (Table S4). Given that repetitive elements comprise large portions of the X chromosome in mammals (Ross et al., 2005) and can be notoriously difficult to assemble, the fragmentation of this chromosome in both assemblies is not surprising. Nevertheless, the selection of a female hybrid offspring for sequencing enabled us to reconstruct the X chromosome for both species.

Large-scale chromosome structure and synteny are conserved

Alignments of the scaffolded *Neotoma* chromosome sequences to the three available *Peromyscus* genomes reveal that the majority of the chromosomal sequence is present in the same order and orientation in both genera (Figures S3–S8). While a number of structural rearrangements (primarily inversions) are present, we see no evidence to suggest translocation events between chromosomes. Furthermore, the number of inversions detected might be inflated. Very few of these structural features are present in all three *Peromyscus* references (Figures S3–S8), leading to the possibility that some inferred inversions might be assembly artifacts resulting from the different scaffolding approaches taken for each *Peromyscus* genome. Synteny findings, too, are consistent with the alignment results, with gene positions strongly conserved between the *Neotoma, P. leucopus,* and *P. maniculatus* genomes (Figure 2C, D).

Though our alignment-based approach cannot rule out large-scale structural differences, individual contig alignments between *N. bryanti* and *N. lepida* lend no support to their existence (Figures S9, S10). This is intriguing given the chromosome variation previously observed within

Neotoma. Chromosome counts range from 1N = 19 in *N. phenax* to 1N = 28 in *N. fuscipes*, with 1N = 26 being the most common for this genus (Baker & Mascarello, 1969); research on *N. lepida* has also found evidence for chromosomal polymorphisms within this species that vary by population (Mascarello & Hsu, 1976). The lack of such structural changes, however, may be instrumental in maintaining the viability and fertility of hybrid offspring.

Nuclear and mitochondrial sequence divergence between *N. bryanti* and *N. lepida*

Sequence divergence was 1.27% between the *N. bryanti* and *N. lepida* nuclear genomes (Table S5), which agrees with an estimated 1.5-million-year divergence (Galewski et al., 2006; Patton et al., 2007). For the mitochondrial genome, divergence was 6.96%; this is not substantially different than the 9% divergence previously found for cytochrome *b* alone (Patton et al., 2007).

Gene counts and repetitive element annotations are similar to model rodents

22,444 protein coding genes were identified for *N. bryanti*, and 21,453 were identified for *N. lepida*. These numbers align with the 22,549 and 21,587 protein coding genes in the current *M. musculus* (Mouse Genome Sequencing Consortium, 2002) and *R. norvegicus* (Rat Genome Sequencing Project Consortium, 2004) annotations, respectively. RepeatMasker results are also in agreement with *M. musculus* (http://www.repeatmasker.org/species/mm.html) and *R. norvegicus* (http://www.repeatmasker.org/species/rn.html), with 40.62% and 41.44% of the *N. bryanti* and *N. lepida* assemblies, respectively, comprised of repetitive elements (Tables S2 and S3).

Dynamic evolution of cytochrome P450 gene islands and the genomic basis of toxin tolerance

After confirming the completeness of the *Neotoma* assemblies and annotations, we examined several P450 subfamilies (CYP2A, CYP2B, CYP3A) because of their role in xenobiotic

metabolism (Thomas, 2007). All three subfamilies showed clear signs of dynamic evolution and lineage-specific expansion, in contrast to the relative stability of *CYP2F*, *CYP2G*, *CYP2S*, and *CYP2T* genes, which are in the same P450 tandem array as the CYP2A and CYP2B subfamilies (Figure 3A). *R. norvegicus* and *P. leucopus* have three CYP2A genes (*CYP2A1*, *CYP2A2*, and *CYP2A3*), which have expanded to seven and six genes in *N. bryanti* and *N. lepida*, respectively (Table 2, Figure 3A). Based on the gene tree (Figure 3C), separate duplication events of *CYP2A2*-like and *CYP2A3*-like genes are responsible for the increase in CYP2A counts in *Neotoma*. Following the relationships among species, the most parsimonious interpretation of structural changes in CYP2A appears to be expansion in *Neotoma* rather than copy loss in other genera.

The pattern of expansion is even more remarkable in the CYP2B subfamily (Figure 3B). *CYP2B* copy number variation among *Neotoma* species was previously documented using quantitative PCR and cDNA cloning approaches (Kitanovic et al., 2018; Malenke et al., 2012), and we can now start to understand the spatial organization of this variation. We infer the evolution of a novel gene island containing only tandemly arrayed *CYP2B* gene copies in New World rats and mice (*Neotoma* and *Peromyscus*). Typically, *CYP2B* genes are found only between the single-copy *CYP2G* and *CYP2S* genes in the CYP2ABFGST gene array (Figure 3A, B). This is the arrangement for *Mus musculus*, humans and other primates (Hoffman & Hu, 2007; Thomas, 2007). However, examination of this gene array in other systems has shown dynamic rearrangement via duplication, deletion, and inversion within a genomic region bounded by the *Egln2* and *Axl* genes (Hu et al., 2008). Similarly, New World rodents also show variation in the arrangement of *CYP2B* genes within the ancestral location of this array, evidenced by two extra functional copies and several pseudogenes in *P. maniculatus* compared to woodrats, and the inverted orientation of the *CYP2B* gene closest to *CYP2S* in woodrats (Figure 3A). Within *Neotoma*, we find a novel gene island outside the bounds of the ancestral array, comprised

solely of variable numbers of *CYP2B* genes (Figure 3B). Based on the relationships among these copies, the novel island likely resulted from a duplication and translocation of the *CYP2B* gene closest to *CYP2S* to a location more than 1 Mb away on the same chromosome, bounded by genes encoding zinc-finger proteins, *Tescl* and *Lypd5*.

The novel *CYP2B* gene island consists of duplicates of sequences most similar to *CYP2B37* (Figure S11), while *CYP2B35* and *CYP2B36* appear to be contained within the bounds of the ancestral island (Wilderman et al., 2014). We identified and named *CYP2B35*, *36*, and *37* in concordance with previous work on *Neotoma lepida* (Malenke et al., 2012; Wilderman et al., 2014). In addition to the unique position of the *CYP2B37* genes (Figure 3B), their proteins show unique enzymatic activity and substrate binding relative to CYP2B35 and CYP2B36 (Wilderman et al., 2014). It is possible that translocation to this new position facilitated the process of neofunctionalization of these genes, as has been proposed in other systems (Deng et al., 2010; Wen et al., 2006). Additionally, there is copy variation between *N. bryanti* and *N. lepida* (Figure 3B), with two functional copies in *N. bryanti* and nine functional copies plus seven pseudogenes in *N. lepida*. These findings, in combination with the known functional uniqueness of CYP2B37 proteins, are consistent with the hypothesis that evolution within unstable gene islands is a source of adaptive novelty in the arms-race between herbivores and the plants on which they feed (Thomas, 2007; Wen et al., 2006; Wilderman et al., 2014).

The pattern observed in the *CYP3A* subfamily was similar to the dynamic evolution apparent in *CYP2B* (Figure 4), and contrasted with the relative conservativism of the CYP genes involved in biosynthesis. First, all *CYP3A* genes are found on a single chromosome (chromosome 12 in *R.* norvegicus, chromosome 23 in *Peromyscus* and our *Neotoma* assemblies). In *R. norvegicus*, the two *CYP3A* gene islands are approximately 7 Mb apart (Figure 4A). The first island is small and contains the *CYP3A9*-like genes, which play a role in steroid hormone synthesis (Xue et al., 2003). This island showed general conservation among

rodent species, with two functional copies in *R. norvegicus*, two in *P. leucopus*, three in *N. bryanti*, and one plus two pseudogenes in *N. lepida*. Moreover, these genes clustered together in the gene phylogeny, and this conservation may reflect the homeostatic roles *CYP3A9* genes play.

In contrast, the remaining *CYP3A* gene islands all have evolved dynamically among the *Peromyscus* and *Neotoma* species included here. The second of two gene islands in *R. norvegicus* contains four functional *CYP3A* genes and is bounded by several non-P450 genes that appear in the other taxa. However, between *CYP3A* islands syntenic with the *R. norvegicus* island lies another island that appears to be unique to the New World rodents (Figure 4B). This array is bounded by *Vmn2r* and *Anhx*, and contains nine functional copies of *CYP3A* in *P. leucopus* (*CYP3A13* in RefSeq) and five in each woodrat. Using *CYP5A1* genes as an outgroup, we found that the island containing the *CYP3A13*-like sequences is the sister group of the other *CYP3A* genes (Figure 4D). Given this relationship, the *CYP3A13*-like genes can be inferred to be duplicated from the common ancestor of the *CYP3A11* and *CYP3A9*-like genes, both of which are present in New and Old Word rodents. This raises the intriguing possibility that the *CYPA13*-like genes were present in the ancestor of all rodents studied herein, and lost in the lineages leading to *R. norvegicus.* Alternatively, a dynamic history of duplication with gene conversion may be obscuring these deeper relationships within the gene tree, as suggested by the low branch support for the ancestral node leading to the *CYP3A11*-like genes. Broader comparative analyses with other highly contiguous rodent genomes will be needed to address the order of these large duplication events involving the *CYP3A* genes.

On chromosome 23, we encountered *CYP3A11*-like genes similar in number and sequence to their putative orthologs in *R. norvegicus* (Figure 4C). In *Mus*, CYP3A11 is highly expressed in the liver and neurons, where it protects against a wide variety of bioactive compounds and cytotoxic chemicals (Hagemeyer et al., 2003). However, the arrangement of

CYP3A11-like gene copies changes drastically with the inclusion of the two *Neotoma* species; the inferred ancestral single gene island is instead two distinct islands separated by the non-P450 gene *Usp12*. The novel fourth *CYP3A* gene island in *Neotoma* appears to have arisen, in part, due to duplication and inversion of a large segment of DNA. The *Zscan25* gene, which bounds one side of this gene island in both *R. norvegicus* and *P. leucopus*, appears on each end of the novel island of *CYP3A11*-like genes in *N. bryanti* (Figure 4C), and a large *CYP3A* pseudogene now separates *Zkscan5* and *Zscan25* on the 5' side of the array. Together, these duplicated genes and novel islands account for the increase in *CYP3A* genes in New World rodents compared to the six functional copies of *CYP3A* in *R. norvegicus* (Table 2), with woodrats having the largest number of functional copies among the species examined.

Expansions in other xenobiotic metabolizing families, but not transcription factors, may play a role in novel toxin degradation

We also explored the dynamics of several phase I, II and III biotransformation families, as well as key transcription factors, to further our understanding of xenobiotic metabolism in *Neotoma*. As many of these families are not well-studied within Rodentia outside of *M. musculus* and *R. norvegicus*, we restricted all comparisons of *N. bryanti* and *N. lepida* to these two rodent species. Therefore, it is possible that a number of the expansions discovered may be present not only in the genus *Neotoma*, but in other closely related cricetid species as well.

Flavin-containing monooxygenases (FMOs), another phase I family, exhibit a different pattern compared to the P450s. Exceptionally efficient at oxygenating substrates into more polar metabolites (Katzung, 2018), FMOs are highly conserved across the tree of life (Huang et al., 2021) and have similar copy numbers across mammalian lineages (Koukouritaki et al., 2002). Consistent with this conservation, we found no evidence for expansion in this family, with all putatively functional FMOs in mouse and rat (*FMO1–6*) present as single copies in both

Neotoma species (Tables 3, S6). Additionally, intact models for *FMO9, 12* and *13*, which appear to have complete coding frames but nevertheless may be pseudogenes in *M. musculus* (Hernandez et al., 2004), were located in each woodrat genome.

Among the phase II enzymes, which conjugate various chemical groups to xenobiotics (Katzung, 2018), we explored the three enzyme classes responsible for the vast majority of phase II reactions in animals: glutathione-*S*-transferases (GSTs), sulfotransferases (SULTs) and UDP-glucuronosyltransferases (UGTs) (Daniel, 1993; Dutton, 1980; Runge-Morris & Kocarek, 2009). For the GSTs, which neutralize substrates through the addition of a glutathione group (Katzung, 2018), we observed modest expansions compared to mouse and rat. Out of the six subfamilies (Nebert & Vasiliou, 2004), these duplications were confined to just two: GST alpha (GSTA) and mu (GSTM). Among the remaining subfamilies — omega (GSTO), pi (GSTP), theta (GSTT) and zeta (GSTZ) — no expansions or contractions were evident (Tables 3, S7). Though not officially part of the GST family, we found similar results for the more distantly related kappa (GSTK), microsomal (MGST) and prostaglandin E synthase subfamilies (Nebert & Vasiliou, 2004). GSTK and MGST numbers matched those of mouse and rat, while the prostaglandin E synthase subfamily appeared to have one duplicated gene (*PTGES*) in *N. lepida*.

This same pattern extended to the SULT family, a group of enzymes that metabolize compounds through sulfation (Katzung, 2018), where only a few modest expansions were discovered (Tables 3, S8). Three subfamilies, SULT1, 2 and 3, appeared to have a few duplications in *N. bryanti* and *N. lepida*, while the SULT4, 5 and 6 numbers agreed with those of *M. musculus* and *R. norvegicus* (Alnouti & Klaassen, 2006; Gamage et al., 2006; J. R. Smith et al., 2019). Additionally, the two *PAPS* genes, which serve as the sulfate group donor for all SULT enzymes (Katzung, 2018), were present as single copies in both *Neotoma* genomes, indicating they had not undergone expansion. As the reactions catalyzed by the GST and SULT families can be metabolically more expensive than those of the phase I enzymes (Klaassen,

2001) and result in the loss of conjugate bases that may be limiting (e.g., sulfur, amino acids), our findings of modest duplications in these two families are not surprising from an efficiency standpoint.

In stark contrast to the other phase II enzymes, we found profound expansions in the UGT genes, a class of enzymes that catalyze the conjugation of a substrate with glucuronic acid (Katzung, 2018) (Tables 3, S9). Though the additions were limited to a single subfamily, UGT2B, the numbers of apparently intact genes, 20 for *N. bryanti* and 27 for *N. lepida,* dwarf the seven in mouse (Buckley & Klaassen, 2007) and rat (J. R. Smith et al., 2019). While we were not able to assemble these genes on a single contig for either *Neotoma* species, MUMmer alignments indicate that this is a single genomic region on chromosome 10 that appears to have undergone recent and substantial tandem duplication. Interspersed among the intact UGT2B genes were 18 and 26 pseudogenes for *N. bryanti* and *N. lepida*, respectively, demonstrating an unusually high degree of turnover. Outside of UGT2B, numbers for the remaining UGT subfamilies (UGT1A, 2A and 3A) were all in line with *M. musculus* and *R. norvegicus*, though an extra UGT3A gene was found in *N. bryanti*. As a family, UGTs are exceptionally versatile in substrate recognition (Klaassen, 2001), and the UGT2B subfamily in particular is involved in the metabolism of a wide range of both xenobiotic and endogenous metabolites (Meech et al., 2019). Notably, increased glucuronidation levels have been previously reported for *N. lepida* individuals fed high amounts of *Larrea tridentata* (Haley et al., 2008; Mangione et al., 2001), so it is a distinct possibility that the UGT2B expansions observed in both genomes play a role in enabling these woodrats to subsist on a diet rich in creosote bush. In addition, biotransformation of xenobiotics by P450s is often followed with glucuronidation, and a concomitant expansion of UGTs and P450s makes sense biologically.

Lastly, as part of our efforts to better understand woodrat phase III detoxification — the efflux of xenobiotic compounds across cellular membranes — we also explored the ATP-binding

cassette (ABC) transporter family (Ishikawa, 1992; S. Li et al., 2015). Similar to the UGTs, we observed substantial expansion confined to a single subfamily, in this case ABCG (Tables 3, S10). This subfamily has broad substrate recognition, but is primarily involved in the transport of hydrophobic compounds (both endogenous and xenobiotic) including lipids, bile salts and sterols (Kerr et al., 2011). The expansion was driven by the proliferation of a single member of the ABCG subfamily on chromosome 10: *ABCG3*. *ABCG3* genes appear to be unique to rodents and have a distinct ATP-binding domain (Mickley et al., 2001), but are otherwise poorly studied. The region encompassing these genes was assembled in its entirety for both *Neotoma* species, and much like the CYP2B and UGT2B subfamilies, tandem duplication with a high degree of gene turnover is evident, suggesting relatively recent expansion. Modestly increased gene counts were also observed in the ABCB family, which plays a role in xenobiotic metabolism in mammals (Sarkadi et al., 2006); *ABCB3*/*TAP2* was duplicated in both species, and extra copies of *ABCB1* and *ABCB2*/*TAP1* were present in *N. lepida.* All ABCA genes were present in both *Neotoma* genomes, though unlike *M. musculus* and *R. norvegicus*, only a single copy of *ABCA8* was present. The remaining ABC families (C, D, E and F) each had numbers matching those of mouse and rat, and no expansions were evident.

Finally, we also investigated whether or not duplication among transcription factors may play a role in regulating the response to xenobiotic metabolism, but found this was unlikely to be the case. The aryl hydrocarbon receptor (*AHR*) and nuclear hormone receptors 1I2 and 1I3 (*NR1I2* and *NR1I3*; these encode the pregnane X and constitutive androstane receptors, PXR and CAR, respectively), which are the primary transcription factors involved in coordinating the expression of multiple biotransformation genes in mammals (Tolson & Wang, 2010; Wang & LeCluyse, 2003; Zhang et al., 2008), were intact in each sequenced genome and present only as single copies.

Conclusions

Our study contributes the first nearly complete, thoroughly annotated genomes of the genus *Neotoma*. Given the complexity, repetitiveness and high degree of turnover in several biotransformation families, and the P450 islands in particular, long-read DNA sequencing technology was crucial to our ability to accurately reconstruct the evolutionary history of the genes studied herein. These genomes now provide the basis for understanding aspects of adaptive evolution in these and related species, and in particular, the dynamics of three specific biotransformation families involved in diet-related adaptations: P450 2B and 3A, UGT2B and ABCG. As anthropogenic climate change over the coming decades is expected to lead to vast alterations in vegetation (Rosenzweig et al., 2007) — and therefore drastic changes in mammalian herbivore diets — woodrats can provide insight into the physiological and genetic factors that are crucial for survival and species conservation. These new genomic resources lay the groundwork for future studies into the adaptability of this unique genus (and vertebrate herbivores in general), and the success and efficiency of the trio binning method should encourage other researchers studying hybridizing species to strongly consider utilizing this approach.

Acknowledgments

We thank Dr. Richard Clark for generously sharing his computational resources with us for this work and Sebastian Smith for assistance with UNR's Pronghorn. We are also indebted to Dr. James Patton for providing the male *Neotoma lepida* used for the F_1 cross, Dr. James R. Halpert for his input on the xenobiotic metabolizing families, Madeline Nelson for performing the animal husbandry, Jennifer Dixon and Kika Kitanovic for carrying out RNA preparation for the transcriptome work, Whitney Kenner for preparing the DNA samples, and Danny Nielson for initial genotyping to confirm the identity of genome animals. We are grateful to the editor, the two

reviewers, and Dylan Klure for their comments and insightful suggestions on improving the manuscript. The photographs in Figure 1 were generously provided by Dr. Sara Weinstein (juniper and creosote images), Maggie Doolin (*N. bryanti* image) and Dr. Kevin Kohl (*N. lepida* image). This research was supported by grants from the National Science Foundation to M.D.D. and M.D.S. (IOS-1656497), and to M.D.M. (IOS-1457209 and OIA-1826801).

References

Alnouti, Y., & Klaassen, C. D. (2006). Tissue Distribution and Ontogeny of Sulfotransferase Enzymes in Mice. *Toxicological Sciences*, *93*(2), 242–255. https://doi.org/10.1093/toxsci/kfl050

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Baker, R. J., & Mascarello, J. T. (1969). Karyotypic analyses of the genus *Neotoma* (Cricetidae, Rodentia). *Cytogenetic and Genome Research*, *8*(3), 187–198. https://doi.org/10.1159/000130061

Bredemeyer, K. R., Harris, A. J., Li, G., Zhao, L., Foley, N. M., Roelke-Parker, M., O'Brien, S. J., Lyons, L. A., Warren, W. C., & Murphy, W. J. (2020). Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (*Felis catus*) and Asian Leopard Cat (*Prionailurus bengalensis*). *Journal of Heredity*, esaa057. https://doi.org/10.1093/jhered/esaa057

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Buckley, D. B., & Klaassen, C. D. (2007). Tissue- and Gender-Specific mRNA Expression of UDP-Glucuronosyltransferases (UGTs) in Mice. *Drug Metabolism and Disposition*, *35*(1), 121–127. https://doi.org/10.1124/dmd.106.012070

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Campbell, M., Oakeson, K. F., Yandell, M., Halpert, J. R., & Dearing, D. (2016). The draft genome sequence and annotation of the desert woodrat *Neotoma lepida*. *Genomics Data*, *9*, 58–59. https://doi.org/10.1016/j.gdata.2016.06.008

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, gkw654. https://doi.org/10.1093/nar/gkw654

Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, *20*(2), 15–19.

Chen, M.-J. M., Lin, H., Chiang, L.-M., Childers, C. P., & Poelchau, M. F. (2019). The GFF3toolkit: QC and Merge Pipeline for Genome Annotation. In S. J. Brown & M. E. Pfrender (Eds.), *Insect Genomics* (Vol. 1858, pp. 75–87). Springer New York. https://doi.org/10.1007/978-1-4939-8775-7_7

Coyner, B. S., Murphy, P. J., & Matocq, M. D. (2015). Hybridization and asymmetric introgression across a narrow zone of contact between *Neotoma fuscipes* and *N. macrotis* (Rodentia: Cricetidae): Hybridization in *Neotoma*. *Biological Journal of the Linnean Society*, *115*(1), 162–172. https://doi.org/10.1111/bij.12487

Daniel, V. (1993). Glutathione S-Transferases: Gene Structure and Regulation of Expression. *Critical Reviews in Biochemistry and Molecular Biology*, *28*(3), 173–207. https://doi.org/10.3109/10409239309086794

Dearing, M. D., Foley, W. J., & McLean, S. (2005). The influence of plant secondary metabolites on the nutritional ecology of herbivorous terrestrial vertebrates. *Annual Review of Ecology Evolution and Systematics*, *36*, 169–189.

Deng, C., Cheng, C.-H. C., Ye, H., He, X., & Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences*, *107*(50), 21593–21598. https://doi.org/10.1073/pnas.1007883107

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2019). Apollo: Democratizing genome annotation. *PLOS Computational Biology*, *15*(2), e1006790. https://doi.org/10.1371/journal.pcbi.1006790

Dutton, G. F. (1980). *Glucuronidation of drugs and other compounds*. CRC Press.

Farrer, R. A. (2017). Synima: A Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics*, *18*(1), 507. https://doi.org/10.1186/s12859-017-1939-7

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Foley, J. W., Iason, G. R., & McArthur, C. (1999). Role of plant secondary metabolites in the nutritional ecology of mammalian herbivores: How far have we come in 25 years?,. In H.

G. J. and G. C. Fahey (Ed.), *Nutritional Ecology of Herbivore* (pp. 130–209). American Society of Animal Science.

Freeland, W. J., & Janzen, D. H. (1974). Strategies in herbivory by mammals the role of plant secondary compounds. *The American Naturalist*, *108*(961), 269–289. https://doi.org/10.1086/282907

Galewski, T., Tilak, M., Sanchez, S., Chevret, P., Paradis, E., & Douzery, E. J. (2006). The evolutionary radiation of Arvicolinae rodents (voles and lemmings): Relative contribution of nuclear and mitochondrial DNA phylogenies. *BMC Evolutionary Biology*, *6*(1), 80. https://doi.org/10.1186/1471-2148-6-80

Gamage, N., Barnett, A., Hempel, N., Duggleby, R. G., Windmill, K. F., Martin, J. L., & McManus, M. E. (2006). Human Sulfotransferases and Their Role in Chemical Metabolism. *Toxicological Sciences*, *90*(1), 5–22. https://doi.org/10.1093/toxsci/kfj061

Gremme, G., Brendel, V., Sparks, M. E., & Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, *47*(15), 965–978. https://doi.org/10.1016/j.infsof.2005.09.005

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(3), 645–656. https://doi.org/10.1109/TCBB.2013.68

Hagemeyer, C. E., Rosenbrock, H., Ditter, M., Knoth, R., & Volk, B. (2003). Predominantly neuronal expression of cytochrome P450 isoforms cyp3a11 and cyp3a13 in mouse brain. *Neuroscience*, *117*(3), 521–529. https://doi.org/10.1016/S0306-4522(02)00955-7

Haley, S. L., Lamb, J. G., Franklin, M. R., Constance, J. E., & Dearing, M. D. (2008). “Pharm-Ecology” of Diet Shifting: Biotransformation of Plant Secondary Compounds in

Creosote (*Larrea tridentata*) by a Woodrat Herbivore, *Neotoma lepida*. *Physiological and Biochemical Zoology*, *81*(5), 584–593. https://doi.org/10.1086/589951

Hall, E. R. (1982). *The mammals of North America*. Blackburn Press.

Hernandez, D., Janmohamed, A., Chandan, P., Phillips, I. R., & Shephard, E. A. (2004). Organization and evolution of the flavin-containing monooxygenase genes of human and mouse: Identification of novel gene and pseudogene clusters. *Pharmacogenetics*, *14*(2), 117–130. https://doi.org/10.1097/00008571-200402000-00006

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 65–95). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_5

Hoffman, S. M. G., & Hu, S. (2007). Dynamic evolution of the CYP2ABFGST gene cluster in primates. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *616*(1–2), 133–138. https://doi.org/10.1016/j.mrfmmm.2006.11.004

Holchek, J. L., Munshikpu, A. V., Saiwana, L., Nunez Hernandez, G., Valdez, R., Wallace, J. D., & Cardenas, M. (1990). Influences of six shrub diets varying in phenol content on intake and nitrogen retention by goats. *Tropical Grasslands*, *24*(93–98), Article 93–98.

Hu, S., Wang, H., Knisely, A. A., Reddy, S., Kovacevic, D., Liu, Z., & Hoffman, S. M. G. (2008). Evolution of the CYP2ABFGST gene cluster in rat, and a fine-scale comparison among rodent and primate species. *Genetica*, *133*(2), 215–226. https://doi.org/10.1007/s10709-007-9206-x

Huang, S., Howington, M. B., Dobry, C. J., Evans, C. R., & Leiser, S. F. (2021). Flavin-Containing Monooxygenases Are Conserved Regulators of Stress Resistance and Metabolism. *Frontiers in Cell and Developmental Biology*, *9*, 630188. https://doi.org/10.3389/fcell.2021.630188

Huo, L., Liu, J., Dearing, M. D., Szklarz, G. D., Halpert, J. R., & Wilderman, P. R. (2017). Rational Re-Engineering of the O-Dealkylation of 7-Alkoxycoumarin Derivatives by Cytochromes P450 2B from the Desert Woodrat *Neotoma lepida*. *Biochemistry*, *56*(16), 2238–2246. https://doi.org/10.1021/acs.biochem.7b00097

Ishikawa, T. (1992). The ATP-dependent glutathione S-conjugate export pump. *Trends in Biochemical Sciences*, *17*(11), 463–468. https://doi.org/10.1016/0968-0004(92)90489-V

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jurka, J. (1998). Repeats in genomic DNA: Mining and meaning. *Current Opinion in Structural Biology*, *8*(3), 333–337. https://doi.org/10.1016/S0959-440X(98)80067-5

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Katzung, B. G. (Ed.). (2018). *Basic & Clinical Pharmacology* (Fourteenth Edition). McGraw Hill Education.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kerr, I. D., Haider, A. J., & Gelissen, I. C. (2011). The ABCG family of membrane-associated transporters: You don't have to be big to be mighty: The biochemistry and pharmacology of ABCG proteins. *British Journal of Pharmacology*, *164*(7), 1767–1779. https://doi.org/10.1111/j.1476-5381.2010.01177.x

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kitanovic, S., Orr, T. J., Spalink, D., Cocke, G. B., Schramm, K., Wilderman, P. R., Halpert, J. R., & Dearing, M. D. (2018). Role of cytochrome P450 2B sequence variation and gene copy number in facilitating dietary specialization in mammalian herbivores. *Molecular Ecology*, *27*(3), 723–736. https://doi.org/10.1111/mec.14480

Klaassen, C. D. (2001). *Cararett and Doull's Toxicology: The Basic Science of Poisons.* McGraw Hill.

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L., & Phillippy, A. M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, *36*(12), 1174–1182. https://doi.org/10.1038/nbt.4277

Koukouritaki, S. B., Simpson, P., Yeung, C. K., Rettie, A. E., & Hines, R. N. (2002). Human Hepatic Flavin-Containing Monooxygenases 1 (FMO1) and 3 (FMO3) Developmental Expression. *Pediatric Research*, *51*(2), 236–243. https://doi.org/10.1203/00006450-200202000-00018

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *2004*(5), R12. https://doi.org/10.1186/gb-2004-5-2-r12

Levy Karin, E., Mirdita, M., & Söding, J. (2020). MetaEuk—Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48. https://doi.org/10.1186/s40168-020-00808-x

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Zhang, W., Yin, X., Xing, S., Xie, H. Q., Cao, Z., & Zhao, B. (2015). Mouse ATP-Binding Cassette (ABC) Transporters Conferring Multi-Drug Resistance. *Anti-Cancer Agents in Medicinal Chemistry*, *15*(4), 423–432.

Long, A. D., Baldwin-Brown, J., Tao, Y., Cook, V. J., Balderrama-Gutierrez, G., Corbett-Detig, R., Mortazavi, A., & Barbour, A. G. (2019). The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. *Science Advances*, *5*(7), eaaw6441. https://doi.org/10.1126/sciadv.aaw6441

Mabry, T. J., Difeo, D. R. J., Sakakibara, M., Bohnstedt, C. F. J., & Seigler, D. (1977). The natural products chemistry of *Larrea*. In T. J. H. H. Mabry & J. R. D. R. Difeo (Eds.), *Us/Ibp (International Biological Program) Synthesis Series, Vol. 6. Creosote Bush. Biology and Chemistry of Larrea in New World Deserts. Xvi+284p. Illus. Maps. Dowden, Hutchinson and Ross, Inc.: Stroudsburg, Pa., USA (Dist. By Academic Press: New York, N.Y., USA; London, Encland). Isbn 0-87933-282-4. 1977. 115-134.* (PREV197815053986 Copyright BIOSIS 2003.).

Magnanou, E., Malenke, J. R., & Dearing, M. D. (2009). Expression of biotransformation genes in woodrat (*Neotoma*) herbivores on novel and ancestral diets: Identification of candidate

genes responsible for dietary shifts. *Molecular Ecology*, *18*(11), 2401–2414. https://doi.org/10.1111/j.1365-294X.2009.04171.x

Malenke, J. R., Magnanou, E., Thomas, K., & Dearing, M. D. (2012). Cytochrome P450 2B Diversity and Dietary Novelty in the Herbivorous, Desert Woodrat (*Neotoma lepida*). *PLoS ONE*, *7*(8), e41510. https://doi.org/10.1371/journal.pone.0041510

Mangione, A. M., Dearing, D., & Karasov, W. (2001). Detoxification in Relation to Toxin Tolerance in Desert Woodrats Eating Creosote Bush. *Journal of Chemical Ecology*, *27*(12), 2559–2578. https://doi.org/10.1023/A:1013639817958

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Marsh, K. J., Wallis, I. R., Andrew, R. L., & Foley, W. J. (2006). The detoxification limitation hypothesis: Where did it come from and where is it going? *J Chem Ecol*, *32*(6), 1247–1266. https://doi.org/10.1007/s10886-006-9082-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Mascarello, J. T., & Hsu, T. C. (1976). Chromosome evolution in woodrats, genus *Neotoma* (Rodentia: Cricetidae). *Evolution*, *30*(1), 152–169. https://doi.org/10.1111/j.1558-5646.1976.tb00892.x

Meech, R., Hu, D. G., McKinnon, R. A., Mubarokah, S. N., Haines, A. Z., Nair, P. C., Rowland, A., & Mackenzie, P. I. (2019). The UDP-Glycosyltransferase (UGT) Superfamily: New

Members, New Functions, and Novel Paradigms. *Physiological Reviews*, *99*(2), 1153–1222. https://doi.org/10.1152/physrev.00058.2017

Mickley, L., Jain, P., Miyake, K., Schriml, L. M., Rao, K., Fojo, T., Bates, S., & Dean, M. (2001). An ATP-binding cassette gene (ABCG3) closely related to the multidrug transporter ABCG2 (MXR/ABCP) has an unusual ATP-binding domain. *Mammalian Genome*, *12*(1), 86–88. https://doi.org/10.1007/s003350010237

Moore, B. D., Wiggins, N. L., Marsh, K. J., Dearing, M. D., & Foley, W. J. (2015). Translating physiological signals to changes in feeding behaviour in mammals and the future effects of global climate change. *Animal Production Science*, *55*(3), 272. https://doi.org/10.1071/AN14487

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19. https://doi.org/10.1093/nar/gkv1290

Nebert, D. W., & Vasiliou, V. (2004). Analysis of the glutathione S-transferase (GST) gene family. *Human Genomics*, *1*(6), 460. https://doi.org/10.1186/1479-7364-1-6-460

Nielsen, D. P., & Matocq, M. D. (2021). Differences in dietary composition and preference maintained despite gene flow across a woodrat hybrid zone. *Ecology and Evolution*, *11*(9), 4909–4919. https://doi.org/10.1002/ece3.7399

Patton, J. L., Huckaby, D. G., & Álvarez-Castañeda, S. T. (2007). *The evolutionary history and a systematic revision of woodrats of the* Neotoma lepida *group*. University of California Press.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(W1), W187–W191. https://doi.org/10.1093/nar/gku365

Rat Genome Sequencing Project Consortium. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, *428*(6982), 493–521. https://doi.org/10.1038/nature02426

Rice, E. S., Koren, S., Rhie, A., Heaton, M. P., Kalbfleisch, T. S., Hardy, T., Hackett, P. H., Bickhart, D. M., Rosen, B. D., Ley, B. V., Maurer, N. W., Green, R. E., Phillippy, A. M., Petersen, J. L., & Smith, T. P. L. (2020). Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *GigaScience*, *9*(4), giaa029. https://doi.org/10.1093/gigascience/giaa029

Rosenzweig, C., Casassa, G., Karoly, D. J., Imeson, A., Liu, C., Menzel, A., Rawlins, S., Root, T. L., Seguin, B., Tryjanowski, P., & Hanson, C. E. (2007). Assessment of observed changes and responses in natural and managed systems. In M. L. Parry, O. F. Canziani, J. P. Palutikof, & P. J. van der Linden (Eds.), *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 79–131). Cambridge University Press.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Frankish, A., Lovell, F. L., Howe, K. L., Ashurst, J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., … Bentley, D. R. (2005). The DNA sequence of the human X chromosome. *Nature*, *434*(7031), 325–337. https://doi.org/10.1038/nature03440

Runge-Morris, M., & Kocarek, T. A. (2009). Regulation of Sulfotransferase and UDP-Glucuronosyltransferase Gene Expression by the PPARs. *PPAR Research*, *2009*, 1–14. https://doi.org/10.1155/2009/728941

Sarkadi, B., Homolya, L., Szakács, G., & Váradi, A. (2006). Human Multidrug Resistance ABCB and ABCG Transporters: Participation in a Chemoimmunity Defense System. *Physiological Reviews*, *86*(4), 1179–1236. https://doi.org/10.1152/physrev.00037.2005

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 227–245). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14

Shah, M. B., Liu, J., Huo, L., Zhang, Q., Dearing, M. D., Wilderman, P. R., Szklarz, G. D., Stout, C. D., & Halpert, J. R. (2016). Structure-Function Analysis of Mammalian CYP2B Enzymes Using 7-Substituted Coumarin Derivatives as Probes: Utility of Crystal Structures and Molecular Modeling in Understanding Xenobiotic Metabolism. *Mol Pharmacol*, *89*(4), 435–445. https://doi.org/10.1124/mol.115.102111

Shumate, A., & Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics*, btaa1016. https://doi.org/10.1093/bioinformatics/btaa1016

Shurtliff, Q. R., Murphy, P. J., & Matocq, M. D. (2014). Ecological segregation in a small mammal hybrid zone: Habitat-specific mating opportunities and selection against hybrids

restrict gene flow on a fine spatial scale. *Evolution*, *68*(3), 729–742. https://doi.org/10.1111/evo.12299

Skopec, M. M., Malenke, J. R., Halpert, J. R., & Denise Dearing, M. (2013). An In Vivo Assay for Elucidating the Importance of Cytochromes P450 for the Ability of a Wild Mammalian Herbivore (*Neotoma lepida*) to Consume Toxic Plants. *Physiological and Biochemical Zoology*, *86*(5), 593–601. https://doi.org/10.1086/672212

Smith, F. A., Betancourt, J. L., & Brown, J. H. (1995). Evolution of Body Size in the Woodrat over the Past 25,000 Years of Climate Change. *Science*, *270*(5244), 2012–2014. https://doi.org/10.1126/science.270.5244.2012

Smith, J. R., Hayman, G. T., Wang, S.-J., Laulederkind, S. J. F., Hoffman, M. J., Kaldunski, M. L., Tutaj, M., Thota, J., Nalabolu, H. S., Ellanki, S. L. R., Tutaj, M. A., De Pons, J. L., Kwitek, A. E., Dwinell, M. R., & Shimoyama, M. E. (2019). The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Research*, gkz1041. https://doi.org/10.1093/nar/gkz1041

Sorensen, J. S., & Dearing, M. D. (2006). Efflux transporters as a novel herbivore countermechanism to plant chemical defenses. *Journal of Chemical Ecology*, *32*(6), 1181–1196.

Sorensen, J. S., Skopec, M. M., & Dearing, M. D. (2006). Application of pharmacological approaches to plant–mammal interactions. *Journal of Chemical Ecology*, *32*(6), 1229–1246. https://doi.org/10.1007/s10886-006-9086-z

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(Web Server), W309–W312. https://doi.org/10.1093/nar/gkh379

Stecher, G., Tamura, K., & Kumar, S. (2020). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution*, *37*(4), 1237–1239. https://doi.org/10.1093/molbev/msz312

Thomas, J. H. (2007). Rapid Birth–Death Evolution Specific to Xenobiotic Cytochrome P450 Genes in Vertebrates. *PLoS Genetics*, *3*(5), e67. https://doi.org/10.1371/journal.pgen.0030067

Tolson, A. H., & Wang, H. (2010). Regulation of drug-metabolizing enzymes by xenobiotic receptors: PXR and CAR. *Advanced Drug Delivery Reviews*, *62*(13), 1238–1249. https://doi.org/10.1016/j.addr.2010.08.006

Van Devender, T. R. (1977). Holocene Woodlands in the Southwestern Deserts. *Science*, *198*(4313), 189–192. https://doi.org/10.1126/science.198.4313.189

Van Devender, T. R., & Spaulding, W. G. (1979). Development of Vegetation and Climate in the Southwestern United States. *Science*, *204*(4394), 701–710. https://doi.org/10.1126/science.204.4394.701

Vizueta, J., Sánchez-Gracia, A., & Rozas, J. (2020). BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources*, *20*(5), 1445–1452. https://doi.org/10.1111/1755-0998.13202

Wang, H., & LeCluyse, E. L. (2003). Role of Orphan Nuclear Receptors in the Regulation of Drug-Metabolising Enzymes. *Clinical Pharmacokinetics*, *42*(15), 1331–1357. https://doi.org/10.2165/00003088-200342150-00003

Weinstein, S. B., Martínez-Mota, R., Stapleton, T. E., Klure, D. M., Greenhalgh, R., Orr, T. J., Dale, C., Kohl, K., & Dearing, M. D. (2021). Microbiome stability and structure is governed by host phylogeny over diet and geography in woodrats (*Neotoma* spp.). *Proc Natl Acad Sci U S A*, *118*(47), e2108787118. https://doi.org/10.1073/pnas.2108787118

Wells, P. V., & Hunziker, J. H. (1976). Origin of the Creosote Bush (*Larrea*) Deserts of Southwestern North America. *Annals of the Missouri Botanical Garden*, *63*(4), 843. https://doi.org/10.2307/2395251

Wells, P. V., & Woodcock, D. (1985). Full-glacial vegetation of Death Valley, California: Juniper woodland opening to *Yucca* semidesert. *Madroño*, *32*(1), 11–23.

Wen, Z., Rupasinghe, S., Niu, G., Berenbaum, M. R., & Schuler, M. A. (2006). CYP6B1 and CYP6B3 of the Black Swallowtail (*Papilio polyxenes*): Adaptive Evolution through Subfunctionalization. *Molecular Biology and Evolution*, *23*(12), 2434–2443. https://doi.org/10.1093/molbev/msl118

Wilderman, P. R., Jang, H.-H., Malenke, J. R., Salib, M., Angermeier, E., Lamime, S., Dearing, M. D., & Halpert, J. R. (2014). Functional characterization of cytochromes P450 2B from the desert woodrat *Neotoma lepida*. *Toxicology and Applied Pharmacology*, *274*(3), 393–401. https://doi.org/10.1016/j.taap.2013.12.005

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Xue, L., Zgoda, V. G., Arison, B., & Almira Correia, M. (2003). Structure–function relationships of rat liver CYP3A9 to its human liver orthologs: Site-directed active site mutagenesis to a progesterone dihydroxylase. *Archives of Biochemistry and Biophysics*, *409*(1), 113–126. https://doi.org/10.1016/S0003-9861(02)00582-9

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2019). Ensembl 2020. *Nucleic Acids Research*, gkz966. https://doi.org/10.1093/nar/gkz966

Zhang, B., Xie, W., & Krasowski, M. D. (2008). PXR: A xenobiotic receptor of diverse function implicated in pharmacogenetics. *Pharmacogenomics*, *9*(11), 1695–1709. https://doi.org/10.2217/14622416.9.11.1695

Data Accessibility

The CLC assemblies, Canu assemblies, chromosome scaffolds and accompanying annotations for *Neotoma bryanti* and *Neotoma lepida* have been deposited at the Center for Open Science (https://doi.org/10.17605/osf.io/xck3n). The genomic and RNA-seq reads generated for this work have been deposited at the SRA under BioProject PRJNA818341. The custom scripts used throughout this project are available at https://github.com/robertgreenhalgh/trio-scripts.

Author Contributions

M.D.D. and M.D.S. designed the study, with input from M.D.M. and T.L.P. T.J.O. performed fieldwork, animal husbandry and tissue dissection. R.G. performed genome assembly and generated automated annotations. R.G. and M.L.H. performed genomic analyses. R.G. and M.L.H. curated gene models with input from J.B.H. R.G. and M.L.H. wrote the manuscript with input from all authors.

Tables

Table 1: Assembly metrics and BUSCO assessments for *Neotoma bryanti* and *Neotoma lepida*.

Metric	*Neotoma bryanti*	*Neotoma lepida*
Assembly size (bp)	2,600,174,201	2,612,755,627
Contig count	2,397	2,421
Contig N50 length (bp)	39,917,500	52,364,606
Longest contig (bp)	151,130,156	139,341,509
Shortest contig (bp)	1,077	1,059
Chromosome scaffold size (bp)	2,388,581,836	2,399,855,743
Chromosome scaffold count	24	24
Chromosome scaffold N50 length (bp)	111,123,070	111,355,628
Longest chromosome scaffold (bp)	187,705,395	189,504,864
Shortest chromosome scaffold (bp)	45,124,132	44,361,528
Chromosome scaffold gaps (bp)	41,000	41,000
BUSCO completeness (Glires – 13,798 BUSCOs)	97.62% (13,469)	97.63% (13,471)
Complete, single-copy	96.46% (13,309)	96.42% (13,304)
Complete, duplicated	1.16% (160)	1.21% (167)
Fragmented	0.34% (47)	0.26% (36)
Missing	2.04% (282)	2.11% (291)

Table 2: Cytochrome P450 subfamily gene counts.

P450 subfamily	*Rattus norvegicus*	*Peromyscus sp.*	*Neotoma bryanti*	*Neotoma lepida*
CYP2A	3	3†,‡	7	6
CYP2B	8	9†	7	14
CYP3A	6	13‡	18	16

† *Peromyscus maniculatus* used for gene counts. ‡ *Peromyscus leucopus* used for gene counts. Subfamily completeness varies between the two available *Peromyscus* genomes, with neither having all three clusters intact.

Table 3: Intact gene and subfamily counts for additional phase I, II and III biotransformation families.

Gene family/ subfamily	*Mus musculus*	*Rattus norvegicus*	*Neotoma bryanti*	*Neotoma lepida*
FMO (phase I)	**9**	**8**	**9**	**9**
GST (phase II)	**22**	**22**	**29**	**28**
GSTA	5	6	10	10
GSTM	7	8	10	9
GSTO	2	2	2	2
GSTP	3	1	2	2
GSTT	4	4	4	4
GSTZ	1	1	1	1
SULT (phase II)	**21**	**14**	**28**	**24**
SULT1	6	7	9	7
SULT2	9	4	10	9
SULT3	2	0	5	4
SULT4	1	1	1	1
SULT5	1	1	1	1
SULT6	2	1	2	2
UGT (phase II)	**12**	**11**	**26**	**32**
UGT1A	1	1	1	1
UGT2A	2	2	2	2
UGT2B	7	7	20	27
UGT3A	2	1	3	2
ABC (phase III)	**53**	**57**	**66**	**65**
ABCA	16	16	15	15
ABCB	12	12	13	16
ABCC	11	11	11	11
ABCD	4	4	4	4
ABCE	1	1	1	1
ABCF	3	3	3	3
ABCG	6	10	19	15

Figures

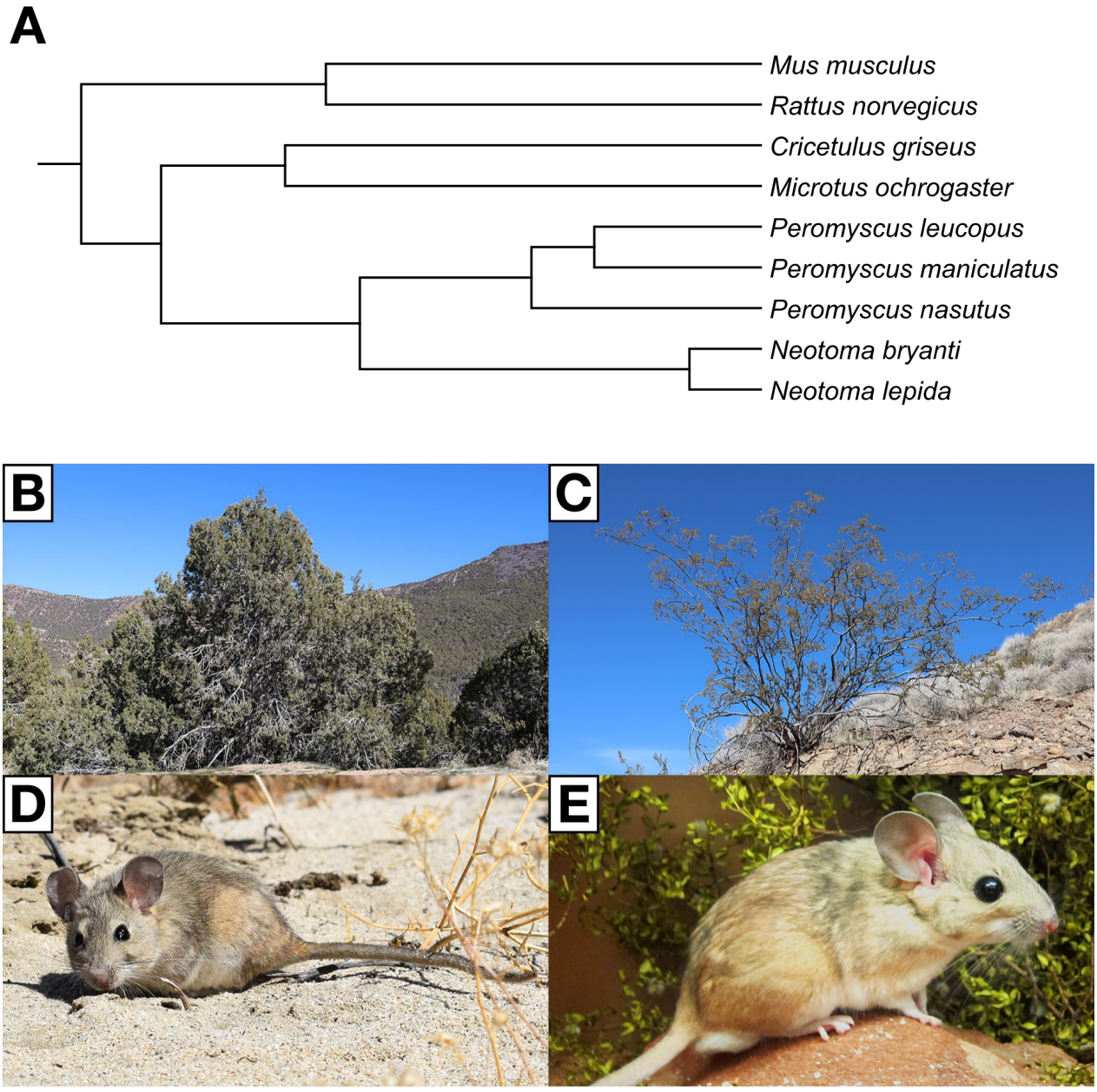

Figure 1: Phylogeny of Cricetidae rodents, common dietary items, and photographs of *Neotoma bryanti* and *Neotoma lepida*. (**A**) *Neotoma* woodrats are closely related to deer mice of the genus *Peromyscus* and more distantly related to other members of the family Muridae. The phylogeny for this panel was downloaded from http://www.timetree.org and visualized in MEGA X (Stecher et al., 2020). (**B**, **C**) Images of juniper (*Juniperus* spp.) (**B**), an ancestral *Neotoma* food source, and creosote bush (*Larrea tridentata*) (**C**), the shrub that replaced it in

many low elevation localities in the Mojave Desert. (**D**, **E**) Photographs of *N. bryanti* (**D**) and *N. lepida* (**E**)**,** the two woodrat species selected for the trio binning approach.

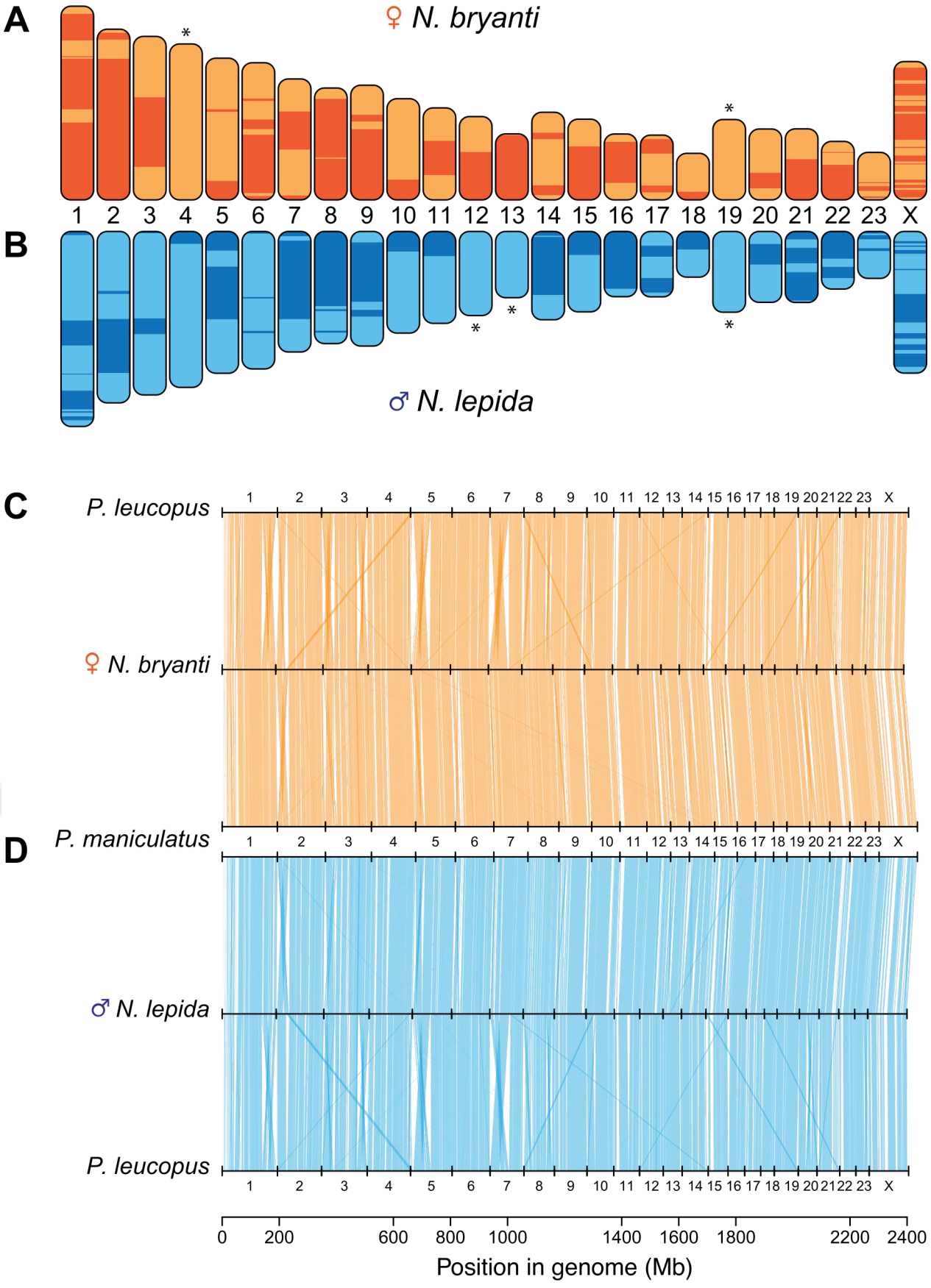

Figure 2: Contig scaffolding and synteny results for both *Neotoma* genomes. (**A**, **B**) Visualization of contigs scaffolded into chromosomes for *Neotoma bryanti* (**A**) and *N. lepida* (**B**). Alternating colors denote individual contigs, and chromosomes indicated with an asterisk were assembled in their entirety as a single contig. (**C**, **D**) Identification of syntenic gene alignments by Synima for the *N. bryanti* (**C**) and *N. lepida* (**D**) autosome scaffolds 1–23 and X chromosome scaffold against the *Peromyscus leucopus* and *P. maniculatus* genomes. Note that while there are a number of instances of cross-alignments among the chromosomes (likely due to cross-mapping between genes in duplicated families), the positions of the majority of genes are strongly conserved between *Neotoma* and *Peromyscus*.

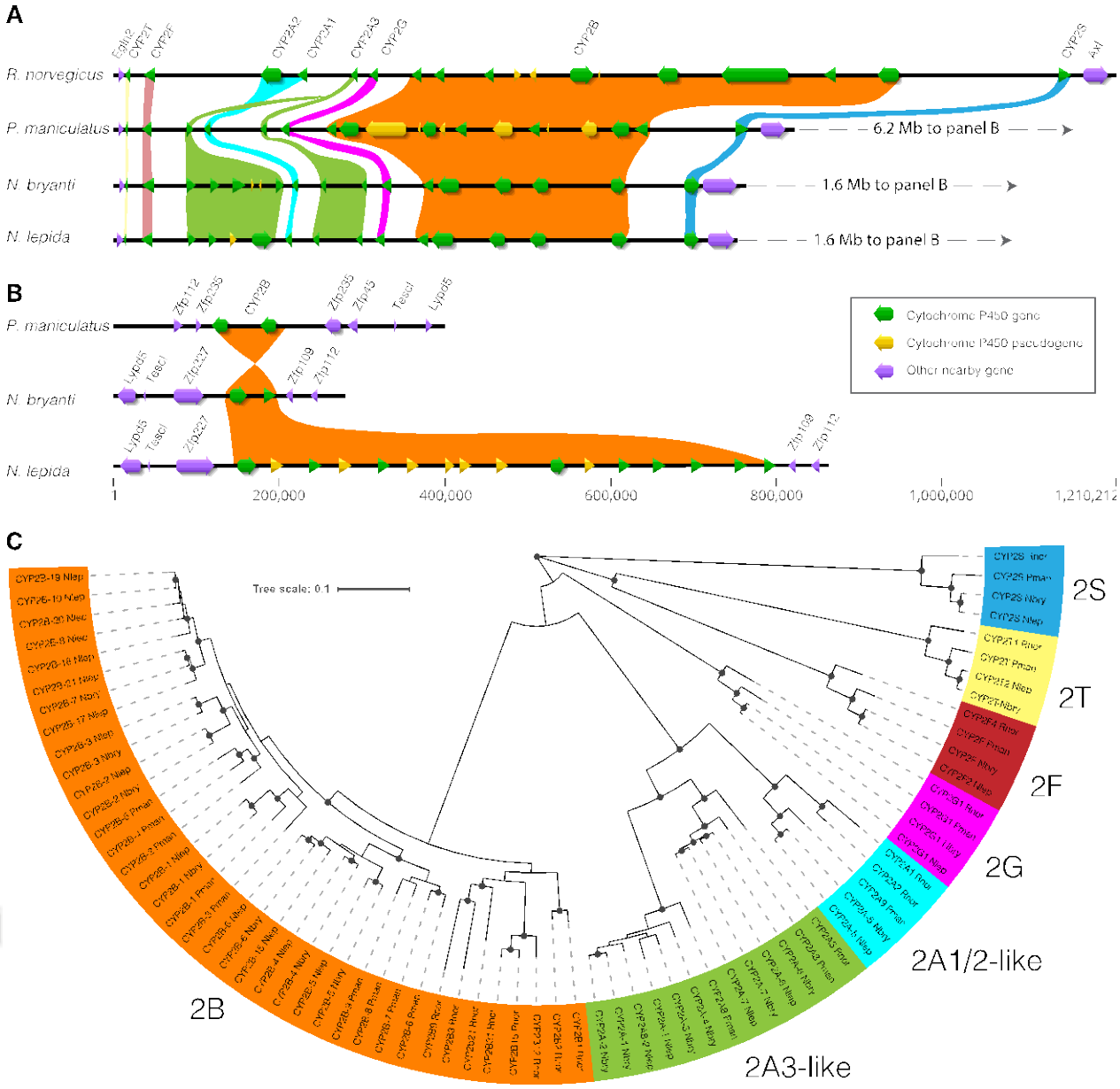

Figure 3: Genomic architecture and evolution of the CYP2ABFGST gene islands in four rodents. Visualization of the contigs containing these genes was based on annotations for the Norway rat (*Rattus norvegicus*; mRatBN7.2) and deer mouse (*Peromyscus maniculatus*; HuPman2.1), and manual annotation of these regions for *Neotoma bryanti* and *N. lepida*. CYP2 families are color coded similarly across all figure panels. (**A**) A highly conserved gene island is bounded by *Egln2* and *Axl* and retains subfamily order in *Neotoma* and *Peromyscus*; gene order is also similar to that of other vertebrates. (**B**) A second gene island composed of variable numbers of *CYP2B* genes arose in the New World rodents only, and underwent further expansion in *N. lepida*. (**C**) A tree of the inferred functional genes from each species was used to determine gene cluster identity/orthology. Circles at nodes indicate bootstrap values of 70% or higher.

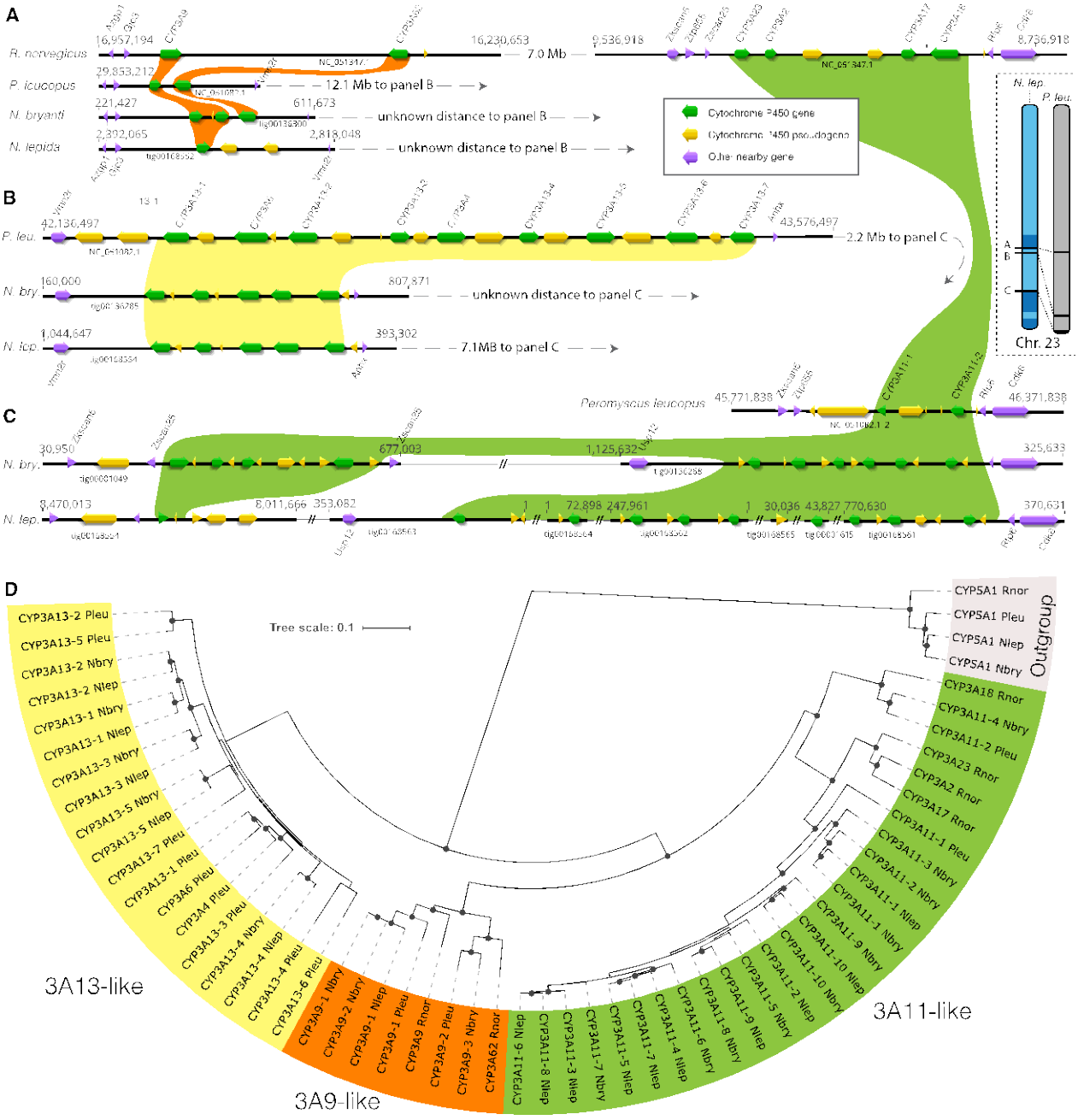

Figure 4: Genomic architecture and evolution of the *CYP3A* gene islands in four rodents. Visualization of the contigs containing these genes was based on existing annotations for the Norway rat (*Rattus norvegicus*; mRatBN7.2) and white-footed mouse (*Peromyscus leucopus*; UCI_PerLeu_2.1), and manual annotation of these regions for *Neotoma bryanti* and *N. lepida*. (**A**) Organization of *R. norvegicus CYP3A* genes on chromosome 12 (top contig), as well as the largely conserved *CYP3A9*-like genes found in *P. leucopus* and both *Neotoma*. (**B**) A second gene island has no clear orthology with a region in *R. novegicus*. (**C)** The downstream *R. norvegicus* genes in (A) are related to a third downstream gene island in *P. leucopus* that is further divided into a third and fourth gene island in *Neotoma*. (**D**) A gene tree of the functional genes from each species was used to determine gene cluster identity/orthology, and is color coded similarly across all figure panels. Circles at nodes indicate bootstrap values of 70% or higher. The inset (top right) indicates the approximate locations of panels (A), (B) and (C) on chromosome scaffold 23 for *N. lepida* (blue; 45.6 Mb in length) and *P. leucopus* (gray; 46.5 Mb in length). Alternating colors denote the individual contigs comprising chromosome scaffold 23 in *N. lepida*.

Supplemental Figure Legends

Figure S1: Alignments of assemblies used to resolve the CYP2 region in *Neotoma lepida*. Alignments of the (**A**) 6.58 Mb and 11.98 Mb contigs from the 40× assembly and (**B**) 13.63 Mb contig from the all reads (AR) assembly against the 18.55 Mb sequence produced by merging the three contigs with quickmerge 0.3. Note that the 13.63 Mb contig is composed of sequence contained in both contigs from the 40× assembly. (**C**) Alignment of the four CYP2 contigs from the *N. lepida* Canu assembly that appear to compose the entirety of this sequence (Nlep_tig00001124, Nlep_tig00001166, Nlep_tig00001922 and Nlep_tig00418827). (**D**) Alignment of the remaining three CYP2 contigs from the Canu assembly that appear to be potential repetitive sequences or assembly errors (Nlep_tig00004740, Nlep_tig00010068 and Nlep_tig00418826). These contigs align to the same region as that occupied by Nlep_tig00418827 (top).

Figure S2: 25-mer frequency spectra for the partitioned *Neotoma bryanti* and *Neotoma lepida* reads. 25-mer spectra counts generated by Jellyfish 2.3.0 are shown for **(A)** *N. bryanti* and **(B)** *N. lepida*.

Figure S3: *Neotoma bryanti* contigs aligned against *Peromyscus leucopus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S4: *Neotoma bryanti* contigs aligned against *Peromyscus maniculatus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S5: *Neotoma bryanti* contigs aligned against *Peromyscus nasutus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S6: *Neotoma lepida* contigs aligned against *Peromyscus leucopus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S7: *Neotoma lepida* contigs aligned against *Peromyscus maniculatus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S8: *Neotoma lepida* contigs aligned against *Peromyscus nasutus* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 250.

Figure S9: *Neotoma bryanti* contigs aligned against *Neotoma lepida* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 2,500.

Figure S10: *Neotoma lepida* contigs aligned against *Neotoma bryanti* chromosome scaffolds. These MUMmer plots were generated using a minimum cluster size of 2,500.

Figure S11: Phylogeny of CYP2ABFGST genes. Unrooted phylogeny of the CYP2ABFGST genes from the four rodent species studied, as well as protein sequences of the genes named *CYP2B35*, *CYP2B36*, and *CYP2B37* from Wilderman et al. (2014).