Evolution of the Largest Mammalian Genome

Ben J. Evans[1,*], Nathan S. Upham[1,2,3,*], Goeffrey B. Golding[1], Ricardo A. Ojeda[4], and Agustina A. Ojeda[4]

[1]Biology Department, McMaster University, Hamilton, Ontario, Canada

[2]Field Museum of Natural History, Chicago, IL

[3]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT

[4]Grupo de Investigaciones de la Biodiversidad (GIB), Instituto Argentino de Investigaciones de Zonas Áridas (IADIZA), Mendoza, Argentina

*Corresponding authors: E-mails: evansb@mcmaster.ca; nathan.upham@yale.edu.

Accepted: June 21, 2017

Data deposition: Sequence data and assemblies from this study are deposited in the Sequence Read Archive of the National Center for Biotechnology Information (NCBI), study SRP102508, bioproject PRJNA380259, including RNA sequencing (RNAseq) reads and transcriptome assemblies for *O. mimax*, *T. barrerae*, *X. tropicalis*, and *X. laevis*, and whole genome sequencing (WGS) reads, for *O. mimax* and *T. barrerae*. Contigs >200 bp in length from the draft whole genome assemblies for *O. mimax* and *T. barrerae* have been deposited at DDBJ/ENA/GenBank in subproject ID SUB2517200 under the accession NDGM00000000 and NDGN00000000. The versions described in this paper are version NDGM01000000 and NDGN01000000. High abundance k-mer contig assemblies are provided in the Supplementary Material online.

Abstract

The genome of the red vizcacha rat (Rodentia, Octodontidae, *Tympanoctomys barrerae*) is the largest of all mammals, and about double the size of their close relative, the mountain vizcacha rat *Octomys mimax*, even though the lineages that gave rise to these species diverged from each other only about 5 Ma. The mechanism for this rapid genome expansion is controversial, and hypothesized to be a consequence of whole genome duplication or accumulation of repetitive elements. To test these alternative but nonexclusive hypotheses, we gathered and evaluated evidence from whole transcriptome and whole genome sequences of *T. barrerae* and *O. mimax*. We recovered support for genome expansion due to accumulation of a diverse assemblage of repetitive elements, which represent about one half and one fifth of the genomes of *T. barrerae* and *O. mimax*, respectively, but we found no strong signal of whole genome duplication. In both species, repetitive sequences were rare in transcribed regions as compared with the rest of the genome, and mostly had no close match to annotated repetitive sequences from other rodents. These findings raise new questions about the genomic dynamics of these repetitive elements, their connection to widespread chromosomal fissions that occurred in the *T. barrerae* ancestor, and their fitness effects—including during the evolution of hypersaline dietary tolerance in *T. barrerae*.

Key words: whole genome duplication, repetitive DNA, mammals, Rodentia, Caviomorpha, Octodontidae.

Introduction

Mammalian gametes have an average genome size of ~3.5 picograms (pg) (Gregory 2005), which corresponds to ~3.5 billion base pairs of DNA (Dolezel et al. 2003), placing the gamete genome size (the *C*-value) of humans (3.1 pg) slightly below average. However, a common ancestor of the red and golden vizcacha rats (Rodentia, Octodontidae, genera *Tympanoctomys*, and *Pipanacoctomys*), underwent a striking genome expansion, resulting in *C*-values of ~8.4 and 7.7 pg, respectively (Gallardo et al. 1999, 2004, 2003, 2006) (fig. 1), and making them the largest of any mammal by a considerable margin (the next highest is 6.3 pg in the golden mole *Chrysochloris*, Gregory 2005). These massive genomes are packaged into about twice as many chromosomes as their close relative, the mountain vizcacha rat *Octomys mimax*, whose genome is about half as large, even though the lineages that gave rise to these two species diverged from each other only ~5–6 Ma (Upham and Patterson 2012, 2015; Suárez-Villota et al. 2016, fig. 1). These observations raise the question of how the genomes of some vizcacha rats became so large so quickly.

Two mechanisms have been proposed to explain these enormous genomes: (1) whole genome duplication (WGD)

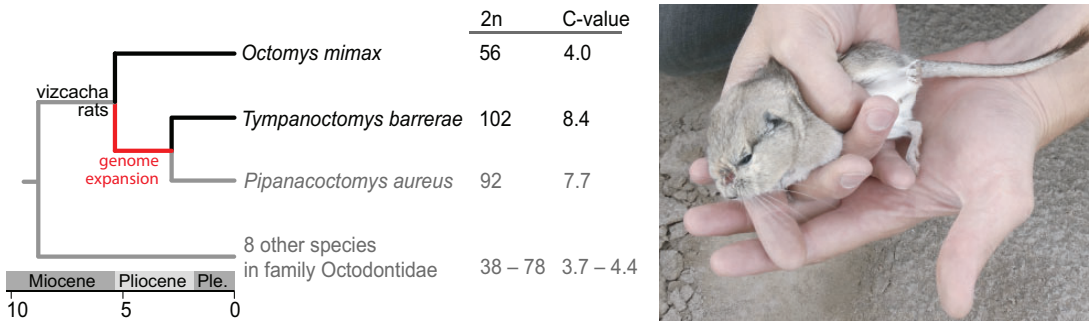

Fig. 1.—Evolutionary relationships, somatic chromosome number (2*n*), and genome size in picograms (*C*-value) of vizcacha rats and other members of the family Octodontidae (left) based on the analysis of Upham and Patterson (2015) and data from Gallardo et al. (2003). Scale bar is in millions of years; dark branches on phylogeny subtend focal species in this study, with the presumed timing of genome expansion indicated with a red branch. According to Upham and Patterson (2015), the divergence of vizcacha rats from other members of the family Octodontidae has a 95% confidence interval (CI) of 7.4–10.5 Ma, divergences of the *Octomys* from other vizcacha rats has a 95% CI of 3.9–7.3 Ma, and divergence of *Pipanacoctomys* from *Tympanoctomys* has a 95% CI of 1.8–4.3 Ma. Not depicted in this phylogeny are the *Tympanoctomys* species *T. loschalchalerosorum* and *T. kirchnerorum*. Some researchers have included *P. aureus* as a member of *Tympanoctomys* (Diaz et al. 2015), but we opt to maintain this separate genus to reflect their molecular divergence. Depicted on the right is the red vizcacha rat *T. barrerae* in El Nihuil, Mendoza, Argentina (photo credit: Fernanda Cuevas).

and (2) expansion of repetitive genomic elements (RGE). The WGD and RGE hypotheses both involve genomic redundancy and are nonexclusive in that either one or both could have contributed to genome expansion. A key factor that distinguishes these hypotheses is that genomic redundancy generated by WGD is genome-wide (or perhaps partially so, see proposed hybrid/allotriploid origin below), whereas genomic redundancy generated by RGE affects only repetitive elements that 1) constitute a portion rather than the entirety of the genome and 2) should be rare in genes due to purifying selection. In order for the RGE hypothesis to fully account for observed genomic characteristics in *T. barrerae*, it would also have to be preceded, accompanied, or followed by chromosomal fissions, which resulted in their high diploid chromosome number ($2n = 102$). As such, the RGE hypothesis requires more steps than WGD to accomplish the transformation in both genome size and karyotype.

The WGD hypothesis is supported by (1) entirely metacentric chromosomes, except the acrocentric Y (Gallardo et al. 2004), (2) a haploid number (n) of 51 and 46 for *T. barrerae* and *P. aureus*, respectively, and DNA content that is about twice that of *O. mimax* ($n = 28$, *C*-value = 4.0 pg) (Gallardo et al. 2003), and (3) the detection of a few duplicated genes (Bacquet et al. 2008; Gallardo et al. 2006, 2004). In plants, WGD can increase the sizes of structures such as leaves, flowers, and cells (including gametes); this is collectively known as the gigas effect (Acquaah 2007). Related to this, another line of evidence supporting the WGD hypothesis is (4) the large head size of spermatozoa (Gallardo et al. 2002). Gallardo et al. (2007) also proposed a hybrid/allotriploid origin of *T. barrerae* wherein an unreduced ($2n = 56$) diploid gamete from *O. mimax* and a haploid ($n = 46$) gamete from *P. aureus* fused to generate the 102 chromosomes in somatic cells of *T. barrerae*. Using silver staining of chromosomal preparations, one active pair of chromosomes containing a nucleolar organizer region was detected in *T. barrerae* and in *O. mimax* but four and two hybridization spots, respectively, were identified using fluorescence *in situ* hybridization (FISH) with ribosomal DNA probes to metaphase spreads (Gallardo et al. 2006). This information was interpreted to suggest nucleolar dominance in an allopolyploid genome of *T. barrerae*, in which one of two the sets of duplicated ribosomal RNA (i.e., the ribosomal DNA from one of the two putative ancestral species of an allopolyploid descendant species) were transcriptionally silenced by epigenetic modifications (Gallardo et al. 2006, 2007).

The RGE hypothesis is supported by FISH with 14 chromosome-specific probes from another caviomorph rodent, *Octodon degus*, that painted only one orthologous chromosome pair, or portion of a pair, of *T. barrerae* (Svartman et al. 2005). This study also found four chromosome-specific probes from *O. degus* that painted several chromosomes of *T. barrerae* and, based on reciprocal hybridizations with whole genome probes, identified a greater proportion of the genome as species-specific in *T. barrerae* than *O. degus* (Svartman et al. 2005). Extensive heterochromatin in the karyotype of *T. barrerae* was also considered to support RGE-based genome expansion in association with tandem fissions (Contreras et al. 1990).

Cytogenetic evidence for unique genomic regions in the expanded genomes of vizcacha rats was also recovered by Suárez-Villota et al (2012). When meiotic chromosomes of *T. barrerae* were blocked with genomic DNA from *P. aureus* and then probed with genomic DNA from *O. mimax*, 30 chromosomes were painted and 26 were not, but when the blocking and probe DNA were reversed, 26 chromosome pairs of *T. barrerae* were painted (Suárez-Villota et al. 2012).

This result was interpreted as being consistent with the hybrid/triploid origin proposed by Gallardo et al. (2007) (Suárez-Villota et al. 2012). On balance, another explanation is that expansion of RGE disproportionately affected a subset of the chromosomes in an ancestor of (*T. barrerae* + *P. aureus*), a possibility that was proposed by Svartman et al. (2005).

Different Types of Genetic Similarity Are Predicted by the WGD and RGE Hypotheses

In general, genetic similarity within or among genomes can either be due to descent from a common ancestor (homology) or convergence (homoplasy). Here we focus on homologous similarity, which can be further divided into similarity due to speciation from a common ancestor (resulting in orthologous sequences in different species), versus similarity within or between species due to small scale gene duplication (resulting in paralogous sequences) or WGD (resulting in homeologous—also known as ohnologous—sequences). If species hybridize, gene flow can homogenize orthologous sequences. Introgressed genes in different species, including new variation that arises after gene flow, can still be considered orthologous because they are not convergently evolved (homoplasious) and because they are not derived from gene or genome duplication (paralogous and ohnologous, respectively).

Considering nucleotide sequences from transcriptomes and genomes, the WGD hypothesis predicts that redundancy in the expanded genomes will generally be ohnologous, although paralogous variation is also expected because duplication of genes and expansion of RGEs also occur in polyploids (i.e., independently from WGD). With the exception of ohnologous redundancy stemming from very ancient genome duplications (Dehal and Boore 2005), the RGE hypothesis predicts that *all* genomic redundancy is paralogous.

This is the first study to leverage sequence data from complete transcriptomes and whole genomes of vizcacha rats to explore the origin of their huge genomes, the largest known among mammals. We studied both types of data because they are expected to be influenced in distinct ways by RGE and WGD. For comparison, we additionally analyzed transcriptome and whole genome data from a diploid and tetraploid species pair—the African clawed frogs *Xenopus tropicalis* and *X. laevis*—whose ploidy levels have been confirmed by whole genome sequences. Our goals were thus to search for diverged ohnologs and repetitive DNA in the genomes and transcriptomes of *T. barrerae* and *O. mimax*, and to evaluate our findings in the context of the WGD and RGE hypotheses.

Materials and Methods

Samples and Data

We collected RNAseq and WGS data from a female *T. barrerae* individual (field identification number AO245) and a male *O. mimax* individual (field identification number AO248). Vouchers for each individual are deposited in the mammal collection at the Instituto Argentino de Investigaciones de Zonas Áridas (IADIZA), Mendoza, Argentina. The *T. barrerae* individual was collected in El Nihuil, Mendoza Province, Argentina, and the *O. mimax* individual was collected in Ischigualasto Provincial Park, San Juan Province, Argentina. RNAlater (Thermo Fisher Scientific) and ethanol preserved genetic samples were then exported from Argentina to McMaster University, Canada, for analysis.

RNA was extracted from six tissues (liver, heart, muscle, lung, kidney, and gonad—testis or ovary) for each individual using the RNeasy Mini kit (Qiagen). Libraries for RNAseq were prepared with New England Biolab's NEBNext Ultra RNA Library Prep kit for Illumina. RNAseq was performed using 100 bp paired-end sequencing on two lanes of an Illumina HiSeq 2500 machine, with the six transcriptomes from each species multiplexed on one lane per species. Total cellular DNA (including mitochondrial and nuclear DNA) was extracted from muscle tissue from each individual using the DNeasy extraction kit (Qiagen). Libraries for WGS were prepared using the Illumina TruSeq Nano DNA library Preparation kit. WGS was performed for both individuals and multiplexed on one HiSeqX lane, using 150 bp paired-end sequencing, and with 2/3rds of the lane loaded with the *T. barrerae* library and 1/3rd with the *O. mimax* library. In this way, we ended up with a similar level of coverage for each individual (See Results). All sequence data from this study are deposited in the Sequence Read Archive of the NCBI (study SRP102508, bioproject PRJNA380259).

Sequence Trimming and Assembly; Transcriptomes

RNAseq data were trimmed with TRIMMOMATIC version 0.32, retaining sequences with a minimum length of 36 base pairs and an average Phred-scaled quality score of at least 15 in a sliding window of four base pairs. After trimming, a total of 132,894,650 and 144,063,591 paired-end RNAseq reads were used to assemble transcriptomes of *T. barrerae* and *O. mimax*, respectively. Analysis with FASTQC (Andrews et al. 2010) indicated that, in general, these RNAseq data were of very high quality, with almost all reads having very high per base sequence qualities their entire length, very high per tile and per read sequence quality, and a long trimmed sequence length.

TRINITY version 2.1.1 was used to assemble, for each individual, concatenated RNAseq data from the six transcriptomes. Only read pairs where both partners passed the trimming step were assembled. The "cd-hit-est" function of CDHIT version v4.6.1-2012-08-27 was then used to remove identical sequences from the transcriptome assemblies.

Sequence Trimming and Assembly; Draft WGS

WGS data were trimmed with TRIMMOMATIC version 0.36, respectively, using the same settings as for the RNAseq data detailed above. We then used the k-mer based approach of

QUAKE (Kelley et al. 2010) to identify and filter reads in the WGS data. K-mers are sequence motifs of size *k*, and their analysis can provide insights into genome size and redundancy, and also sequencing error. In high coverage WGS data, most low frequency k-mers are due to sequencing errors. JELLYFISH (Marçais and Kingsford 2011) was used to count 19-mers, and the cov-mod.py script of QUAKE was used to assess the cutoff value for trimming, which was 1 for both WGS data sets. Then the "correct" program of QUAKE was used to filter the WGS data. Similar to the RNAseq data, analysis with the program FASTQC (Andrews et al. 2010) indicated that the trimmed and filtered WGS data were of very high quality.

A total of 342,002,329 and 168,247,437 paired-end and 7,202,820 and 3,375,490 single-end WGS reads were used to assemble the genomes of *T. barrerae* and *O. mimax*, respectively, using ABYSS version 1.9.0 (Simpson et al. 2009) with a k-mer value of 64. Each genome assembly included paired-end and single-end reads.

Read Mapping and Genotyping

Under the WGD hypothesis, diverged ohnologous genes are expected in *T. barrerae* but not *O. mimax*. To test this, we mapped RNAseq data from *T. barrerae* to the transcriptome assembly of *O. mimax*, and vice versa. Under the WGD hypothesis, we expected heterozygosity to be substantially higher for the heterospecific mapping where RNAseq data from *T. barrerae* is mapped to the transcriptome assembly from *O. mimax* than for the heterospecific mapping where RNAseq data from *O. mimax* is mapped to the transcriptome assembly from *T. barrerae*. This is because in the former mapping, divergent ohnologous sequences from *T. barrerae* are expected to co-map to one ortholog of *O. mimax*, with their divergent sites generating heterozygous genotypes.

The MEM function of BWA version 0.7.8-r455 (Li and Durbin 2010) and SAMTOOLS version 1.3.1 (Li and Durbin 2010) were used to map sequence reads to conspecific or heterospecific transcriptome or genome assemblies. Prior to mapping, PCR duplicates were removed with PICARDTOOLS version 1.131 (http://broadinstitute.github.io/picard/). Genotype calling and filtering was performed with SAMTOOLS and BCFTOOLS (Li and Durbin 2010). Genotype filtering removed insertion/deletions, and genotypes that were missing, had a depth of coverage below 10X, or a phred-scaled genotype quality below 20.

Analysis of Assembled Transcriptome Sequences: Disomy, Triads, and Diads

The genome of *T. barrerae* is disomic (Gallardo et al. 2006), meaning that chromosomes form bivalents during meiosis rather than multivalents. One possibility is that this genome has always been diploid and disomic. Alternatively disomy can evolve after WGD. In disomic polyploids, ohnologous genomic regions diverge from each other due to a lack of recombination (Wolfe 2001). If the *T. barrerae* genome was formed by WGD, two ohnologous genes would initially be present in the *T. barrerae* genome for every one orthologous gene in *O. mimax*. Eventually one or both of the ohnologous genes may become pseudogenized, but the expectation that both copies of some ohnologous pairs continue to be expressed holds for many millions of generations (e.g., Lynch and Conery 2000). The disomic genome of the tetraploid frog *Xenopus laevis*, for example, still expresses both copies of >60% of its ohnologous pairs even though allopolyploidization occurred >15 Ma (Session et al. 2016).

Thus, homologous genes from a tetraploid species and a closely related diploid species include "triads" comprising two expressed ohnologs from the tetraploid and one ortholog from the diploid or, if one ohnolog is lost or lowly expressed in the polyploid, "diads" comprising one ortholog in the tetraploid and one in the diploid. We identified triads and diads in *T. barrerae* and *O. mimax* using an approach (Chain et al. 2011) based on reciprocal best megablast matches ("hits") identified with the basic local alignment tool for nucleotides (BLASTN) version 2.2.25+ (Altschul et al. 1990). In triads, two sequences from *T. barrerae* both had the same top BLASTN hit to the *O. mimax* transcriptome which, when queried against the *T. barrerae* transcriptome, also had both *T. barrerae* sequences as its top two hits. BLASTN searches used an expect value of 1^{-20}, with the task set to megablast (to match highly similar sequences). A minimum match between orthologous comparisons of 200 base pairs (bp) was required in order for a triad to be recognized. Expression of a gene also might only be detected in one species, and these were not considered in the diad/triad analysis. Perl scripts were used to compile triads and diads based on BLASTN results, extract homologous sequences and generate fasta files which were piped to MAFFT version v7.205 (Katoh and Standley 2013) for alignment using the "adjustdirectionaccurately" option. We then used another perl script to quantify uncorrected pairwise genetic distances between the aligned sequences.

For comparative purposes, we performed the same analysis using the *O. mimax* transcriptome as the putative polyploid and the *T. barrerae* transcriptome as the diploid. We also performed this exercise on assembled transcriptomes from another closely related species pair, described next, whose ploidy levels are established to be diploid and tetraploid by whole genome sequencing.

Transcriptome Analysis of a Confirmed Tetraploid/Diploid Pair

As a proof of concept, we performed a parallel analysis of an RNAseq data set from an uncontroversial polyploid species, the African clawed frog *X. laevis* (Session et al. 2016), and a closely related diploid species *X. tropicalis* (Hellsten et al. 2010). Similar to the disomic genome of *T. barrerae* (Gallardo et al. 2006), the genome of *X. laevis* is disomic

(Tymowska et al. 1991). RNA was extracted from liver tissue from one individual from each of these species using the same protocol as above, and RNAseq was performed as described above. The sex of the *X. laevis* individual (identification number BJE4168) is unknown and the *X. tropicalis* individual (identification number BJE3909) was female. After trimming using the same TRIMMOMATIC parameters as for the rodent RNAseq data detailed above, a total of 47,952,989 and 54,091,238 reads were retained for *X. tropicalis* and *X. laevis*, respectively. Similar to the rodent data, these frog data were also of very high quality in terms of per base sequence quality, per tile sequence quality, and per read sequence quality.

An important caveat to this comparison (discussed below) is that the timing of WGD in the ancestor of *X. laevis* was ~17 Ma (Session et al. 2016), considerably older than the hypothesized timing of WGD in *T. barrerae* after divergence from *O. mimax* ~5–6 Ma (Upham and Patterson 2012, 2015; Suárez-Villota et al. 2016).

Analysis of Draft WGS Assembly

As a complement to the transcriptome analyses, we generated and analyzed data from the entire genomes of *T. barrerae* and *O. mimax*. We used these data to explore the possibility that there was widespread pseudogenization in *T. barrerae* that could have occurred following WGD. This could have been favored by natural selection, for example, to restore pre-WGD dosage of autosomal and sex-linked genes.

If WGD occurred very recently in *T. barrerae*, some ohnologous exons may not be substantially diverged from one another. However, divergence in linked, noncoding regions may nonetheless be present, and for this reason even nondiverged ohnologous exons might be assembled into different contigs in the draft genome assembly. To test this possibility, we used each species' conspecific transcriptome assemblies to query their draft genome assemblies. We then quantified how often each section of each unique assembled transcript (which potentially correspond to separate exons) matched more than one genomic region in the genome assembly. We considered an assembled transcript to match multiple genomic regions if one or more portions of the query sequence had more than one megablast hit with an expect value of $<1^{-20}$, and where portion of the query sequence with multiple hits was at least 20 bp long.

One concern with our analyses of the draft genome assemblies is that repetitive sequences may be under represented because their repetitive nature makes them difficult to assemble. As a complement to this analysis, the frequency distribution of k-mers can provide insights into the degree of redundancy in a genome (Liu et al. 2013), by showing how many times a k-mer is observed (the occurrence; *x*-axis) relative to the number of unique k-mers that are observed that many times (the count; *y*-axis). K-mer analysis can be performed directly on raw sequence reads, and thus is not influenced by the degree to which sequences can be assembled. With no redundancy and low heterozygosity, the frequency distribution of k-mers follows Poisson expectations with the mean ($\bar{x}$) equal to the average depth of coverage. $\bar{x}$ is thus the number of times that the highest number of unique k-mers are observed in the data (i.e., the highest peak in a frequency distribution of k-mers). Each heterozygous sites increases the frequency that unique k-mers are observed 0.5 $\bar{x}$ times. In a genome that has experienced WGD, divergent ohnologous genomic regions are expected to generate a k-mer peak at 1$\bar{x}$, with the height of this peak proportional to the time since recombination between ohnologous regions stopped. Nondiverged ohnologous regions generate a k-mer peak at 2$\bar{x}$ because their k-mers are observed twice as often as diverged ohnologous regions. Heterozygous sites in polyploid genomes generate additional k-mer peaks at 0.5$\bar{x}$ and 1.5$\bar{x}$, respectively, due to k-mers associated with single nucleotide polymorphisms that are present on one or three of the two pairs of ohnologous alleles, respectively. Depending on the age of a polyploid, multiple k-mer peaks may or may not be observed. For example, a k-mer distribution for *X. laevis* is unimodal because most ohnologs are diverged (Session et al. 2016), but a k-mer distribution for another younger polyploid *Xenopus* (*X. mellotropicalis*) is bimodal (unpublished data not shown).

After trimming and filtering as described above, we used JELLYFISH (Marçais and Kingsford 2011) to count k-mers of size 19, 25, and 35 bp in the WGS data of *T. barrerae* and *O. mimax*. We ignored k-mers containing bases with a quality below 20. The occurrence distribution was used to estimate genome size and the proportion of the genome containing highly repetitive sequences.

With an aim of identifying and classifying the nature of high abundance repetitive elements in the genome of *T. barrerae*, we used REPARK (Koch et al. 2014) and VELVET (Zerbino and Birney 2008) to assemble high abundance k-mers into contigs. The threshold was selected by REPARK and was 46 for AO248 and 50 for AO245 and a k-mer size of 31 was used in this analysis. We then used BLASTN to attempt to match the high abundance k-mer contigs to conspecific whole genome and transcriptome assemblies. A Perl script was used to determine, in each sample, the proportion of bp of the assembled transcripts that matched any of the conspecific high abundance k-mer contigs. This proportion was calculated in putative noncoding and coding regions in the unique sequences from each assembled transcriptome based on coding regions that were identified by TRANSDECODER version 2.0 (Haas et al. 2013). A few of the most abundant or longest of the high abundance k-mer contigs were also used as queries to search GenBank to determine whether any were similar to annotated repeat elements in other species, and all of them were used as a query against a rodent database of repetitive sequences (Repbase version 21.12; Bao et al. 2015).

Results

A Similar Number of Transcripts Are Assembled for *T. barrerae* and *O. mimax*

Under the WGD hypothesis, we expected more unique transcripts to be assembled from the RNAseq data of *T. barrerae* than *O. mimax*. However, we instead found more transcripts for *O. mimax* ($n = 308{,}854$) than *T. barrerae* ($n = 280{,}712$), of which 306,324 and 279,172 were unique, respectively. The mean length of all transcripts was somewhat higher for *T. barrerae* than *O. mimax*, but less so when only the longest isoform of each gene was considered, and the median lengths were similar either way (supplementary Table S1, Supplementary Material online). As detailed in the Methods, the number of reads used in each assembly was somewhat lower for *T. barrerae* (133 million pairs) compared with *O. mimax* (144 million pairs). Sex-related differences in expression between the female *T. barrerae* and male *O. mimax* individuals—especially in the gonads—probably contribute to some degree to differences in the number and composition of transcripts in these assemblies.

For comparison, we assembled and analyzed transcriptomes of a species pair—the African clawed frogs *Xenopus tropicalis* and *X. laevis*—whose ploidy levels are established to be diploid and tetraploid, respectively, by WGS (Session et al. 2016; Hellsten et al. 2010). As expected for a tetraploid genome, considerably more assembled transcripts were generated for the tetraploid *X. laevis* ($n = 123{,}88$) than for the diploid *X. tropicalis* ($n = 71{,}067$), of which 123,277 and 70,747 were unique. As expected for a comparison between a tetraploid and diploid transcriptome, the total length of assembled transcripts was higher for *X. laevis* than *X. tropicalis* (supplementary Table S1, Supplementary Material online). This disparity in total length was less pronounced for the longest isoforms, which is probably because some ohnologous transcripts of *X. laevis* were incorrectly classified as isoforms in *X. laevis* by TRINITY. The number of reads used in each assembly was somewhat higher for *X. laevis* compared with *X. tropicalis*. The considerably lower number of assembled transcripts in the *Xenopus* transcriptomes compared with the rodent transcriptomes is presumably largely a consequence of them being from only one tissue type instead of six, and the lower number of reads used in the *Xenopus* assemblies. Overall, these results are consistent with the expectation that more transcripts would be assembled from a tetraploid than a diploid transcriptome, although the higher number of reads from the tetraploid *X. laevis* (Materials and Methods) may have contributed to the magnitude of this difference.

A Similar Change in Heterozygosity in Heterospecific Mappings of RNAseq Data from *T. barrerae* and *O. mimax*

Under the WGD hypothesis, heterozygosity is expected to be substantially higher for the heterospecific mapping where RNAseq data from *T. barrerae* is mapped to the transcriptome assembly from *O. mimax* than for the heterospecific mapping where RNAseq data from *O. mimax* is mapped to the transcriptome assembly from *T. barrerae*. Indeed, levels of heterozygosity were mildly elevated in both heterospecific mappings by factors of 1.35 and 1.27 for RNAseq data of *T. barrerae* and *O. mimax*, respectively, compared with conspecific mapping. The relative increase in heterozygosity of heterospecific mapping is thus 1.11 (=1.35/1.27) fold higher for the heterospecific mapping of RNAseq data from *T. barrerae* compared with that for *O. mimax*. Conspecific mapping of each RNAseq data set indicates that the heterozygosity within the *T. barrerae* individual was higher than that within the *O. mimax* individual (0.00083 and 0.00048, respectively). Together these findings are consistent with expectations of the WGD hypothesis, but the magnitude of this increase (1.11-fold) is modest.

When the same analysis was performed for the *Xenopus* assemblies, a much more pronounced signature of WGD in the tetraploid *X. laevis* was recovered. Heterozygosity in the heterospecific mapping with tetraploid *X. laevis* RNAseq data to the diploid *X. tropicalis* transcriptome assembly increased almost 6-fold (0.02011) and compared with the conspecific mapping for *X. laevis* (0.00346). In *X. tropicalis*, the heterospecific mapping to the *X. laevis* transcriptome assembly had slightly lower heterozygosity (0.00134) compared with the *X. tropicalis* conspecific mapping (0.00193).

No Sign of Diverged Ohnologs in the *T. barrerae* Transcriptome

We used a modified reciprocal best BLASTN approach to identify triads and diads in the transcriptome assemblies of *T. barrerae* and *O. mimax* (Methods). When *T. barrerae* was the putative tetraploid, a total of 81,733 diads and triads were identified, but only 3,044 (3.7% of the total) of them were triads (i.e., with ohnologous divergence >0%), and 849 of these triads (1.0% of the total) had ohnologous divergence >1%. For comparison, when the *O. mimax* assembly was used as the putative polyploid transcriptome, a total of 61,742 diads and triads were identified, and 1,219 (2.0% of the total) of them were triads, and 1,057 of these (1.7% of the total) had ohnologous divergence >1%. Thus, contrary to the WGD expectation, similar proportions of diverged expressed duplicates were detected in the transcriptomes of *T. barrerae* and *O. mimax*, and in both species these proportions were low.

Conducting the same analysis on the *Xenopus* transcriptomes found an unambiguous indication of divergent ohnologous variation in the tetraploid *X. laevis*. Using *X. laevis* as the putative tetraploid identified 17,018 triads and diads, 5,572 (32.7% of the total) were triads, and 4,530 (26.6% of the total) had divergence >1%. For comparison, when the diploid species *X. tropicalis* was used as the putative tetraploid, 17,173 triads and diads were identified, 698 (4.1% of the

total) were triads, and only 42 (0.2% of the total) had ohnologous divergence >1%. These results verify that the modified reciprocal best BLASTN approach can identify triads in a polyploid species with diverged ohnologs, including in transcriptomes assembled from much less data than we collected for the vizcacha rats. Of course, this approach has less power when the ohnologous divergence is of similar magnitude to within locus heterozygosity, which might be the case if a polyploid originated very recently. This analysis also carries caveats of not identifying all triads that have divergent patterns of expression between ohnologs, or triads that were sequenced incompletely or in nonoverlapping (nonhomologous) regions of their transcripts.

Overall, we failed to identify a substantial number of triads representing diverged ohnologs in the transcriptome of *T. barrerae* and an ortholog in *O. mimax*, suggesting either that WGD did not occur, or that if it did, it occurred so recently that the level of divergence between ohnologous transcripts in *T. barrerae* is extremely low.

No Genome-Wide Evidence of a Partner Pseudogene for Expressed Transcripts

The draft genome assemblies of *T. barrerae* and *O. mimax* that we assembled were very similar in size, with the total lengths of these assemblies being 2.43 Gigabase pairs (Gbp) and 2.45 Gbp, respectively. The contig N50 and scaffold N50 values (i.e., the shortest length *N* for which 50% of the bp are in contigs or scaffolds whose lengths are smaller than *N*) were 4,794 and 5,293 for the draft *T. barrerae* assembly and 4,872, and 5,223 for the draft *O. mimax* assembly. The contig L50 and scaffold L50 values (i.e., the number of contigs or scaffolds whose length is ≥ the N50 value) were 135,809 and 120,691 for the draft *T. barrerae* assembly and 139,636 and 129,153 for the draft *O. mimax* assembly. The number of scaffolds in the draft *T. barrerae* and *O. mimax* assemblies was 18,311,798 and 5,510,567, respectively. Thus, even after accounting for the ~2-fold larger genome size, the draft assembly of *T. barrerae* was considerably more fragmented than that of *O. mimax*. After removing PCR duplicates from reads that were mapped to each assembly, the average depth of coverage per site of conspecific reads mapped to each conspecific genome assembly was 22.3× and 15.2×, respectively, for *T. barrerae* and *O. mimax*. These estimates are higher and less accurate than the k-mer based estimates of coverage discussed below, because they are inflated by reads from repetitive regions in different genomic regions that were assembled into one contig.

Overall, the fragmented nature and small size of these draft genome assemblies indicates that both are far from complete; this effort thus is best considered as a random sequencing survey of genomic variation in each one. The similar sizes of the draft genome assemblies of these species, which are known to differ dramatically in their actual genome size, could arise if 1) WGD occurred so recently that the subgenomes within *T. barrerae* are essentially identical and assembled together, or 2) the genome of *T. barrerae* consists of a high proportion of repetitive elements which were either assembled together or that generated complexities that obfuscated assembly (e.g., bubbles in the de Bruijn graph; Simpson et al. 2009).

Under the WGD hypothesis, we expected exons from individual transcripts from a putative polyploid to match two (ohnologous) genomic regions in a conspecific genome assembly. However, BLAST analysis indicated that a similar proportion of assembled transcriptome sequences in each species had at least one portion (putative exons) that matched multiple genomic regions in the conspecific genome assembly. Specifically, 45% (138,020 out of 306,324 unique queries) of the assembled *O. mimax* transcriptome sequences had portions with multiple hits to the conspecific draft genome assembly and 47% (131,598 out of 279,172 queries) of the assembled *T. barrerae* transcriptome sequences had portions with multiple hits to the conspecific draft genome assembly. That these proportions are similar in each species is inconsistent with the expectations of the WGD hypothesis. Instead, these multiple hits in transcriptomes of both species could be due to paralogous (e.g., due to small scale duplication of individual genes or individual exons) rather than ohnologous variation (i.e., due to WGD).

When the same analysis was performed using our assembled transcriptomes and the published whole genome sequences from the *Xenopus* species (versions 7.1 and 9.1 for *X. tropicalis* and *X. laevis*, respectively), results were consistent with the expectations based on the ploidy level of each species. We found 36% of the assembled transcripts (25,523 out of 70,747 queries) from the diploid *X. tropicalis* had portions of the query sequence that matched multiple genomic regions of the conspecific genome assembly, whereas more than double of the assembled transcripts from the tetraploid *X. laevis* (85% of sections, 104,887 out of 123,277 queries) had portions with multiple matches. That the proportion of matches is higher than the estimate of the proportion of retained duplicates in *X. laevis* (~60%; Session et al. 2016) is consistent with the conclusion from the triad analysis that some of the unique assembled transcriptome sequences are nonoverlapping portions of the same transcript. Overall, this comparison between these frog species confirms our expectation that assembled transcripts from a tetraploid species match multiple conspecific genomic regions more frequently than do assembled transcripts from a diploid species.

The *T. barrerae* Genome Contains More Highly Redundant Sequences Than *O. mimax*

Highly repetitive sequences, such as those from transposable elements, are associated with much higher occurrence than the genome-wide coverage ($\bar{x}$, see Methods). To test for this,

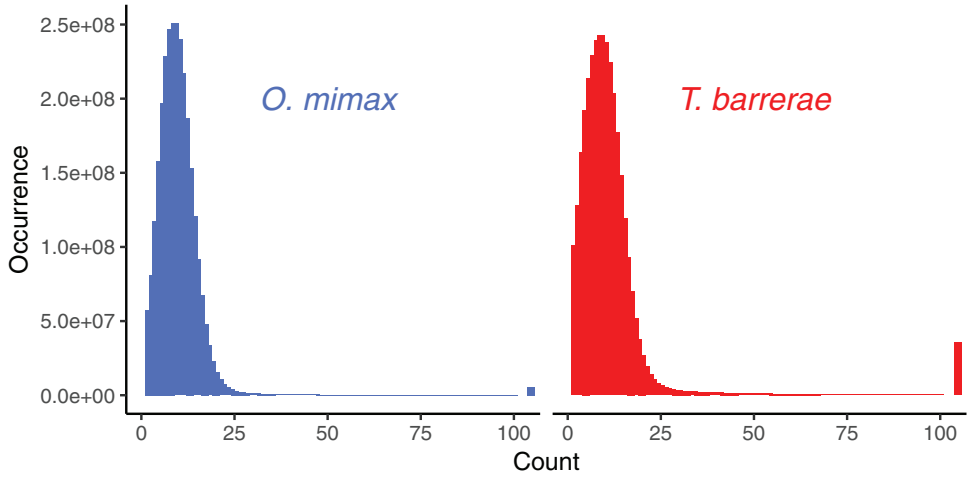

Fig. 2.—Occurrence of 35-mers in trimmed and filtered WGS data from *O. mimax* (left) and *T. barrerae* (right). The occurrence of 35-mers that were observed only once is not shown; the sum of occurrences of 35-mers that were observed >100 is represented by a bar on the right side of each graph (which were 5.7 million and 35.7 million occurrences for *O. mimax* and *T. barrerae*, respectively).

we generated k-mer distributions from the trimmed WGS data from *T. barrerae* and *O. mimax* with an aim to further characterize the extent and nature of redundancy in these genomes; an example of the 35-mer distribution is shown in figure 2. It is clear from this analysis that the genome of *T. barrerae* contains considerably more highly redundant sequences than does that of *O. mimax*. For example, even though the depth of coverage in each species was similar (see below), in *T. barrerae*, 35,697,009 different 35-mers were detected >100 times whereas in *O. mimax*, only 5,683,317 different 35-mers were detected >100 times (fig. 2). Results were consistent using other k-mer sizes (data not shown).

The peak occurrence of k-mer distributions suggests coverage was 9× for 35-mers (11× and 12×, respectively, for 25-mers and 19-mers; hereafter results are itemized in this order) for *T. barrerae* and 9× (10×, 11×) for *O. mimax*. A rough estimate of genome size was obtained by summing the product of the k-mer counts and occurrence and then dividing this sum by the coverage. We excluded from this estimate k-mers with frequency equal to 1, the bulk of which are due to sequencing error. The (rough) genome size estimate for *T. barrerae* was 6.19 (5.47, 5.11) Gbp and for *O. mimax* was 3.16 (3.14, 2.99) Gbp. The magnitudes of the genome size estimates based on k-mers are consistent with previous inferences of a large disparity in genome size between these species, but are lower and less accurate than the other estimates based on flow cytometry (Gallardo et al. 2003) due to variation in coverage, genomic regions that are difficult to sequence, and the omission of rare k-mers in the k-mer estimates.

The portion of the k-mer distribution that corresponds to the single copy regions can be approximated by overlaying a Poisson distribution with mean equal to the depth of coverage on the k-mer distribution. Applying a conservative threshold of 40× (i.e., much higher than the Poisson expectation, fig. 2) for both species yielded similar estimated sizes of the single copy portion of each genome: 3.35 (2.79, 2.44) Gbp for the *T. barrerae* genome and 2.79 (2.64, 2.34) Gbp for the *O. mimax* genome. Expressing this as a proportion of the total genome size and subtracting from 1 yields estimated genomic proportions of highly redundant sequences, which are 45.8% (48.9%, 52.3%) and 11.7% (15.7%, 21.6%) for *T. barrerae* and *O. mimax*, respectively. Thus, k-mer analysis suggests highly redundant sequences comprise about half of *T. barrerae* genome but only one fifth or less of the *O. mimax* genome.

Assembly of Highly Redundant Sequences

Using the programs RepARK and velvet, highly abundant 31-mers were assembled into 309,756 and 41,854 high abundance k-mer contigs for *T. barrerae* and *O. mimax*, respectively, with a median coverage of 3.0 and 3.0 and median length of 57 and 88 bp, respectively. Thus there are many more types of highly repetitive elements in the genome of *T. barrerae*, although they are typically shorter than the highly repetitive elements in *O. mimax* (fig. 3).

In *O. mimax*, two high abundance k-mer contigs had coverages of 244× 236×, and were 59 and 61 bp long, respectively (fig. 3). Neither one had a close match in NCBI GenBank. All of the other high abundance k-mer contigs in *O. mimax* had coverage <50×. The longest high abundance k-mer contig in *O. mimax* was 16,330 bp and matched with 89% identity the complete mitochondrial genome sequence of *T. barrerae* (the mitochondrial genome sequence of *O. mimax* is not currently in GenBank). The coverage of this longest high abundance k-mer contig was 3×.

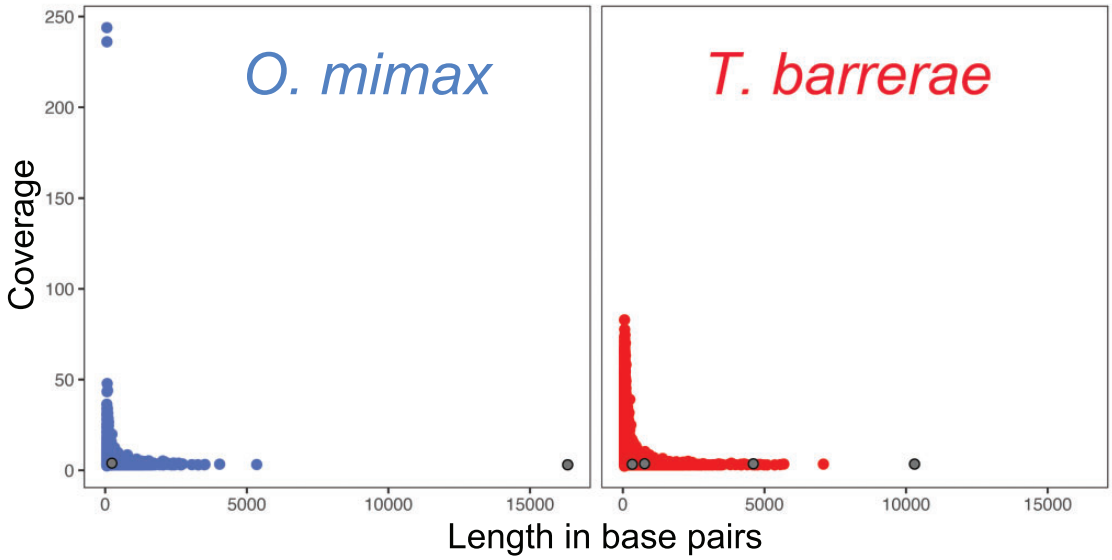

Fig. 3.—Length and coverage of high abundance k-mer contigs in *O. mimax* (left) and *T. barrerae* (right). In both plots, gray dots indicate contigs with length >200 that match mitochondrial DNA of *T. barrerae* GenBank accession number HM544132.1.

In *T. barrerae*, 217 of the high abundance k-mer contigs had coverage ≥50×. Thus, many more high abundance contigs had very high coverage in *T. barrerae* than in *O. mimax*. The longest of these was 10,301 bp, had coverage of 3.5X, and matched with 98% identity the complete mitochondrial genome sequence of *T. barrerae* (fig. 3).

We also searched for matches to the complete mitochondrial genome sequence of *T. barrerae* (GenBank accession number HM544132.1, length 16,863 bp) in the high abundance k-mer contigs from *T. barrerae* and *O. mimax*. A total of 15 and 2 matches, respectively, were recovered and the sum of the lengths of matches in each species was 16,501 and 16,654 bp. These results indicate that the vast majority of the high abundance k-mer contigs from each species are not of mitochondrial origin.

To explore the degree to which the high abundance k-mer contigs were different from each other, we used BLASTN to query each high abundance k-mer contig against all conspecific high abundance k-mer contigs, and then quantified how many of the queries had more than one match (a unique contig should only match itself). Out of 309,756 high abundance k-mer contigs from the *T. barrerae* WGS data, 90,953 (30%) matched at least one other conspecific high abundance k-mer contig. Out of 41,854 high abundance k-mer contigs from the *O. mimax* data, 9,080 (22%) matched at least one other conspecific high abundance k-mer contig. Thus, there is a much higher number of distinct sequence motifs that are highly repetitive in *T. barrerae* compared with *O. mimax*. These results also suggest that genome expansion of *T. barrerae* was not due to just one or a small number of repetitive sequences, but instead due to expansion of many distinctive and highly repetitive small sequence motifs.

What Are the Highly Redundant Sequences?

With a goal of better understanding the nature of the highly redundant sequences we identified, we blasted the high abundance k-mer contigs against the rodent database of repetitive DNA in Repbase version 21.12 (Bao et al. 2015). When the search was performed using discontinuous megablast (for somewhat dissimilar sequences), 6% of the high abundance k-mers from *T. barrerae* had a match and 9% of those from *O. mimax*) had a match. In both cases, they were mostly to LINE-1 sequences (80% for *T. barrerae* and 75% for *O. mimax*). Overall then, almost all of the high abundance k-mer contigs did not have a close match to known rodent repetitive sequences; the small proportion of these sequences that did match annotated repetitive sequences generally were to LINE-1 sequences. Summaries of high abundance k-mer contigs of *T. barrerae* and *O. mimax* discontinuous blast hits to annotated Repbase sequences are presented in supplementary Tables S2 and S3, Supplementary Material online, respectively.

Highly Redundant Sequences in the Transcriptome Assemblies

When the high abundance k-mer contigs from each species were blasted against their conspecific transcriptome assembly for *T. barrerae* there were 226,206 hits to 42,872 transcripts out of a total of 279,172 assembled transcripts. For *O. mimax*, there were 83,844 hits to 30,928 transcripts out of 306,324 assembled transcripts. The proportions of bp in each assembly that matched high abundance k-mer contigs was 1.9% and 3.3% in putative coding and noncoding transcribed regions of *T. barrerae*, and 0.7% and 1.6% in putative coding and noncoding transcribed regions of *O. mimax*. These results

indicate that high abundance k-mer contigs are more prevalent in 1) noncoding than coding regions of transcribed DNA in both species, 2) nontranscribed than transcribed DNA in both species, and 3) nontranscribed and transcribed DNA of *T. barrerae* than the corresponding regions of *O. mimax*.

Highly Redundant Sequences in the Genome Assemblies

We also evaluated how high abundance k-mer contigs from each species matched their conspecific genome assemblies using BLASTN. When those from *O. mimax* were blasted against the draft genome assembly from this species, there were 4,913,004 hits to 2,139,053 scaffolds out of a total of 5,510,567 scaffolds in the draft assembly (39%). The maximum number of unique hits on one scaffold was 189, and the scaffold with the maximum number of hits was 10,334 bp in length. When this scaffold was used as a query to NCBI, it was found to match a predicted vomeronasal type-2 receptor 116-like (LOC101592925) mRNA from the caviomorph rodent *O. degus*. Similarly, the four *O. mimax* scaffolds containing the next highest numbers of unique high abundance contigs (which ranged in length from 10,611 to 29,674 bp) matched vomeronasal type-2 receptor 116-like from *O. degus* or uncharacterized LOC105742023 from *O. degus*. Overall, this indicates that some genomic regions of *O. mimax* that carry a high diversity of repetitive elements (that are abundant elsewhere in the genome) are associated with protein coding regions, including (in the case of the vomeronasal receptor) detection of conspecific pheromones (Francia et al. 2014).

When this analysis was performed for the high abundance k-mer contigs for *T. barrerae*, there were 48,340,064 hits to 12,512,520 scaffolds out of a total of 18,311,798 scaffolds in the draft assembly (68%). Thus, the number of scaffolds with highly repetitive elements in *T. barrerae* was higher than in *O. mimax*. The maximum number of unique hits on one scaffold was 346, and the scaffold with the maximum number of hits was 10,004 bp in length. When this scaffold was used as a query to NCBI, it was found to match a clone of 1_c RPCS satellite sequence from the caviomorph rodent *Ctenomys haigi*. Similarly, the four *T. barrerae* scaffolds (which ranged in size from 4,679 to 10,398 bp) containing the next highest numbers of unique high abundance contigs also all matched sequences that were annotated as satellite or microsatellite DNA from species in the caviomorph rodent genus *Ctenomys*. However, in all of these matches, $<$10% of the *T. barrerae* query sequence matched the *Ctenomys* sequences, indicating that most sequence in each of these scaffolds had no close match in NCBI.

Overall, a diversity of redundant elements was detected in several genomic regions of *O. mimax* with known biological function, whereas the highest diversity of highly redundant elements in *T. barrerae* were detected in genomic regions with unknown function (satellite or microsatellite DNA). We note that these characterizations are anecdotal in nature, and do not take into account differences in scaffold sizes in the genome assemblies, or differences in the coverage of high abundance contigs that they contain.

Discussion

There is considerable debate about why the red vizcacha rat (*T. barrerae*) has such a huge genome (Gallardo et al. 1999, 2004, 2003, 2006; Svartman et al. 2005; Suárez-Villota et al. 2012). Our analyses of RNAseq and WGS data demonstrate that about half of the *T. barrerae* genome is diverse but highly repetitive sequences, but found no strong evidence for WGD. The only support we recovered for the WGD hypothesis was slightly higher heterozygosity (1.1×) of the heterospecific mapping of *T. barrerae* RNAseq data to the *O. mimax* transcriptome assembly. However, we speculate that this modest increase could have been associated with the higher within-individual heterozygosity of *T. barrerae*, as opposed to stemming from WGD. Furthermore, the WGD hypothesis was not strongly supported by other findings such as that, in both species, a similar number of transcripts were assembled, and a similar proportion of diverged expressed duplicates were detected, and putative exons in assembled transcripts generally had similar levels of redundancy in the conspecific genome sequence. In contrast, when we performed the same analyses on a pair of tetraploid and diploid frog species whose ploidy levels have been confirmed by high quality WGS, we found strong signs of tetraploidy as expected in each of these analyses (the tetraploid had much higher heterozygosity when mapped to a heterospecific diploid transcriptome, and compared with the diploid, the tetraploid had many more assembled transcripts, many diverged expressed duplicates, and putative exons from the tetraploid had a higher degree of redundancy when compared with the conspecific genome sequence).

Using k-mer-based approaches and draft assembly of WGS data, we identified thousands of distinctive and highly abundant sequence motifs in the genome of *T. barrerae* that are either not found or not common in the genome of *O. mimax*, a closely related species whose genome is about half as large (fig. 1). Most of these elements had no close match in the rodent repeat element database (RepBase). One reason for this could be that most of the rodent data in RepBase are from mice or rats, and the ancestor of these two species diverged from the ancestor of (*T. barerae* + *O. mimax*) early during the diversification of the Order Rodentia, some 64–74 Ma (Meredith et al. 2011). Thus, that most of the high abundance kmer contigs did not closely match annotated repetitive elements does not demonstrate an absence of homology to known repetitive elements. Some of these elements were identified in transcribed DNA of *T. barrerae*—more frequently

in putatively noncoding than in coding portions of the transcriptome—but in general they were far less common in transcribed genomic regions than nontranscribed regions.

Taken together, these results provide insights into the origin of the largest mammalian genome and support that 1) it evolved by expansion of a diverse mosaic of repetitive sequences, and 2) the genomic distribution of these elements is not uniform. The nonuniform distribution of repeats could either be because they expanded throughout the genome and then were removed from most transcribed regions by natural selection, or because they mostly originated in genomic regions with a dearth of transcribed regions (e.g., telomeres or centromeres), or some combination of these possibilities. That some chromosomes appear to contain more repetitive elements than others (Svartman et al. 2005; Suárez-Villota et al. 2012) suggests that their expansion involves mechanisms acting in *cis* rather than *trans*. The major type of satellite DNA in the caviomorph rodent of the genus *Ctenomys*, for example, has been proposed to expand via a *cis*-acting rolling mechanism (Rossi et al. 1995). We were not able to test this hypothesis due to the fragmented nature of our draft genome assemblies. Additional data from long read sequences will be important to permit more comprehensive characterization of repetitive elements and their genomic distributions.

The Y chromosome of placental and marsupial (therian) mammals is relevant to the WGD hypothesis because it is "degenerate" in that during divergence from the X it lost almost all of the genes it once carried (Charlesworth and Charlesworth 2000). Caveats to our analyses discussed below notwithstanding, there are at least two reasons that WGD would be surprising in a species that has a degenerate Y chromosome. The first, articulated by Orr (1990), is that a newly generated dioecious polyploid individual would have to backcross with a diploid from the ancestral population; this would generate a triploid individual whose dosage of autosomal and X-linked genes would be different from the ancestral dosage. This difference and the associated disruption of the ancestral dosage compensation system presumably would be highly deleterious. The second reason relates to the genotypic consequences of having a genome with two pairs of sex chromosomes that were generated by WGD after a breeding population of polyploids became established. Offspring could inherit from zero Y chromosomes (an XXXX individual) to three (an XYYY individual with one Y being sex-linked and the other two being paleo-Ys with autosomal inheritance). Because the Y chromosome is missing many genes that are carried by the X, variation in gene dosage associated with these different genotypes could pose substantial challenges to a newly formed polyploid population, at least until the nonsex linked paleo-Y were removed by natural selection (Evans et al. 2012). Hence, our data support the long-standing view that polyploidization is an unlikely mode of speciation in mammals that have a degenerate (Y) sex chromosome (Coyne and Orr 2004).

Caveats and Future Directions

Overall, this study supports the conclusions of Svartman et al. (2005) based on cytogenetic evidence that genome expansion in *T. barrerae* was due to the expansion of highly repetitive elements. However, caveats exist to our interpretations and many questions remain. Data analyzed in this study are derived from a female *T. barrerae* individual and a male *O.mimax* individual. It would have been ideal if the sexes of our samples were the same, and this difference between samples generates caveats to our interpretations. For example, when mapped to a heterospecific assembly, sequence data from the *O. mimax* Y chromosome could map to gametologous regions of the *T. barrerae* X chromosome. This could increase the genome-wide level of heterozygosity compared with the reciprocal mapping (where reads from a female *T. barrerae* were mapped to an assembly from a male *O. mimax*), which would make our inferences based on heterozygosity of heterospecific mapping less conservative with respect to the WGD hypothesis. However, we expect this factor to have a negligible effect on our conclusions because the X and Y chromosomes of therian mammals diverged from each other >160 Ma, and differ dramatically in size, structure, gene content, and repetitive elements (Graves 2015), and thus are unlikely to extensively comap to one another. A recently sequenced complete Y-chromosome from a mouse provides support for the assertion that rodent sex chromosomes are diverged from each other in gene content: only 2% of the genes on the male-specific portion of the mouse Y chromosome are derived from the ancestral autosomes and >95% of genes on the mouse Y chromosomes are within three ampliconic (extensively duplicated and homogenized by gene conversion) gene families (Soh et al. 2014). The number of ampliconic genes on the mouse Y chromosome is unusual compared with other mammalian Y chromosomes but in general, gene content on the X and Y chromosomes of placental mammals tends to be very different (e.g., Bellott et al. 2014).

The timing of WGD and the associated statistical power to detect it is also a potential issue with our searches for the divergent ohnologs. Using data from a confirmed diploid and tetraploid species pair of African clawed frogs (*Xenopus*), we demonstrated that our analyses can detect genomic signatures of WGD in a tetraploid (*X. laevis*) with fairly diverent ohnologs (~7%; Evans and Kwon 2015). However, these analyses may fail to detect signs of WGD that occurred so recently that there exists almost no divergence between ohnologs. We view this possibility as exceedingly unlikely because the estimated divergence time between *T. barrerae* and *P. aureus*—whose large genomes suggest a common mechanism—is ~2.9 Ma (Upham and Patterson 2015; Suárez-Villota et al. 2016; fig. 1). Thus genome

expansion in the ancestor of (*T. barrerae* + *P. aureus*) may have predated the Pleistocene, which provides many rodent generations for mutations to distinguish duplicated genes, if they were present.

Another concern with the transcriptome analyses is that gene silencing may potentially occur very soon after WGD and/or affect a large number of ohnologous gene pairs. For example, transcriptional silencing after polyploidization can occur within a few generations in several plants (Adams et al. 2004; Kashkush et al. 2002; Feldman and Levy 2009). Transcriptional shock, in which expression patterns of ohnologous genes are rapidly partitioned soon after WGD has been reported in allotetraploid *Tragopogon miscellus* (Asteraceae) plants (Buggs et al. 2011). However, whatever the causes, patterns, or extent of gene silencing after WGD, if diverged ohnologous sequences were present in the *T. barrerae* we would expect to identify some in the draft WGS assembly. This expectation holds even if DNA from one half of the duplicated genome (the subgenome) were lost at a greater rate than the other, such as in *Nicotiana tabacum* (tobacco) (Renny-Byfield et al. 2011, 2012), *Arabidopsis* (Thomas et al. 2006), and *Xenopus laevis* (Session et al. 2016).

Even though they appear to not be derived from WGD, the rapid evolution of the large genomes of some vizcacha rats is nonetheless striking, and the support for the RGE hypothesis presented here raises new questions. For example, it remains unclear why a dramatic increase in chromosomal complement occurred alongside the expansion in genome size in the (*Tympanoctomys* + *Pipanacoctomys*) lineage. Chromosome numbers in Octodontidae are among the most diverse for mammalian families, and in 13 species range from 38 to 102 (mode of $2n = 58$), including the $2n = 78$ *Octodon lunatus*, in addition to the vizcacha rats (Gallardo et al. 2003). Chromosomal rearrangements are common in the Ctenomyidae sister family (tuco-tucos), where chromosome numbers are also highly variable ($2n$ ranges from 10 to 70), and where changes in satellite DNA abundance have been linked to cladogenesis (Slamovits et al. 2001). It is therefore possible that chromosomal number variation in these rodents, and in the broader South American radiation of caviomorph rodents ($2n$ ranges from 10 to 118; Dunnum et al. 2001), may promote or be driven by RGE accumulation. Alternatively, it is possible that RGE evolution is decoupled from chromosomal changes entirely. The diversity in chromosomal number in Caviomorpha does not appear to be mirrored by correlated genome size variation, although only 13% of species have known genome sizes (34/264 extant species; Gregory 2005).

Species of *Tympanoctomys* and *Pipanacoctomys* also have a highly specialized ecophysiology that allows them to feed on hypersaline vegetation from the saltbush plant *Atriplex* (Diaz et al. 2000; Berman 2003; Ojeda et al. 1999; Torres-Mura et al. 1989; Giannoni et al. 2000; Ojeda et al. 1996; Mares et al. 1997). This is an exceedingly rare adaptation in mammals that allows them to survive in arid habitats of Argentina's Monte and Patagonian Deserts. Their close relative *O. mimax*, which is also a desert specialist, lacks this adaptation and instead feeds to a greater extent on watery plants such as cactus (Ojeda et al. 1996; Sobrero et al. 2010). The consequences of genome expansion on this adaptation are unknown, although the few other species that independently evolved this adaptation have genome sizes that are fairly typical for mammals (e.g., kangaroo rats; Hatch et al. 1976).

Also unclear is why the genomic distribution of repeat elements appears to be clustered on a subset of the chromosomes (Svartman et al. 2005), what accounts for the high diversity of repetitive elements in *T. barrerae*, as well as why these chromosomes (except the Y) are all metacentric (Contreras et al. 1990). Additionally, the degree to which these expanded genomes are deleterious, whether mechanisms to suppress further expansion have already evolved, and whether expansion is still underway are open questions. Thus there is still much to learn about the largest mammalian genome.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Ben Furman and Fernanda Cuevas for assistance with fieldwork, and Ben Furman, Caroline Cauret, and Jared Simpson for helpful discussions about analysis. We also thank three anonymous reviewers for many helpful comments. This work was supported by grants from the Natural Science and Engineering Research Council of Canada to B.J.E. (RGPIN/283102-2012) and to G.B.G. (RGPIN-2015-04477), and from grants from Consejo Nacional de Investigaciones Científicas y Técnicas to A.A.O. (PICT 2253) and R.A.O. (PIP CONICET 1122015 0100258; PICT Agencia 2015-1636; and PICT-E 0193). N.S.U. was supported by a United States National Science Foundation grant (DEB 1441737). We also thank supporters of our 2015 Instrumentl crowdfunding campaign, including Steadman Upham, Jon Upham, Gyan Sandhu, Travis Knowles, Molly McDonough, Joan Stewart, Mark Stewart, and Bob Syren.

Literature Cited

Acquaah G. 2007. Principles of plant genetics and breeding. Malden, MA: Blackwell Publishing.

Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics 168:2217–2226.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Andrews S, et al. 2010. FastQC: A quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bacquet C, et al. 2008. Epigenetic processes in a tetraploid mammal. Mamm Genome 19:439–447.
Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6:11.
Bellott DW, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508:494–499.
Berman SL. 2003. A desert octodontid rodent, *Tympanoctomys barrerae*, uses modified hairs for stripping epidermal tissue from leaves of halophytic plants. J Morphol. 257:53–61.
Buggs RJ, et al. 2011. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol. 21:551–556.
Chain FJ, Dushoff J, Evans BJ. 2011. The odds of duplicate gene persistence after polyploidization. BMC Genomics 12:599.
Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. Philos Trans Roy Soc Lond B Biol Sci. 355:1563–1572.
Contreras L, Torres-Mura J, Spotorno A. 1990. The largest known chromosome number for a mammal, in a South American desert rodent. Experientia 46:506–508.
Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.
Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.
Diaz GB, Ojeda RA, Gallardo MH, Giannoni SM. 2000. *Tympanoctomys barrerae*. Mamm Species 1–4.
Diaz M, Barquez R, Verzi D. 2015. Genus Tympanoctomys. In: Patton JL, Pardiñas UF, and D'Elía G, editors. Mammals of South America, Vol. 2: Rodents. Chicago (IL): University of Chicago Press, p. 1043–1048.
Dolezel J, Bartos J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. Cytometry 51:127.
Dunnum J, Salazar-Bravo J, Yates T. 2001. The Bolivian bamboo rat, *Dactylomys boliviensis* (Rodentia: Echimyidae), a new record for chromosome number in a mammal. Mamm Biol. 66:121–126.
Evans BJ, Kwon T. 2015. Molecular polymorphism and divergence of duplicated genes in tetraploid African clawed frogs (*Xenopus*). Cytogenet Genome Res. 145:243–252.
Evans BJ, Pyron RA, Wiens JJ. 2012. Polyploidization and sex chromosome evolution in amphibians. In: Soltis PS and Soltis DE, editors. Polyploidy and genome evolution. Berlin: Springer, p. 385–410.
Feldman M, Levy AA. 2009. Genome evolution in allopolyploid wheat: a revolutionary reprogramming followed by gradual changes. J Genet Genomics 36:511–518.
Francia S, Pifferi S, Menini A, Tirindelli R. 2014. Vomeronasal receptors and signal transduction in the vomeronasal organ of mammals. In: Neurobiology of Chemical Communication, p. 297. (Mucignat-Caretta, Ed), Boca Raton, FL: CRC Press Taylor and Francis Group.
Gallardo M, Bickham J, Kausel G, Köhler N, Honeycutt R. 2003. Gradual and quantum genome size shifts in the hystricognath rodents. J Evol Biol. 16:163–169.
Gallardo M, Ojeda R, González C, Ríos C. 2007. The Octodontidae revisited. In: Kelt DA, Lessa EP, Salazar-Bravo JA, and Patton JL, editors. The quintessential naturalist: honoring the life and legacy of Oliver P. Pearson, Vol. 134. California: University of California Publications in Zoology, p. 1–981.
Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Kohler N. 1999. Discovery of tetraploidy in a mammal. Nature 401:341.
Gallardo MH, Conzález CA, Cebrián I. 2006. Molecular cytogenetics and allotetraploidy in the red vizcacha rat, *Tympanoctomys barrerae* (Rodentia, Octodontidae). Genomics 88:214–221.
Gallardo MH, et al. 2004. Whole-genome duplications in South American desert rodents (Octodontidae). Biol J Linn Soc. 82:443–451.
Gallardo MH, Mondaca F, Ojeda R, Köhler N, Garrido O. 2002. Morphological diversity in the sperms of caviomorph rodents. Mastozool Neotrop. 9:159–170.
Giannoni SM, Borghi CE, Ojeda RA. 2000. Feeding behaviour of *Tympanoctomys barrerae*, a rodent specialized in consuming Atriplex leaves. J Arid Environ. 46:117–121.
Graves JAM. 2015. Evolution of vertebrate sex chromosomes and dosage compensation. Nat Rev Genet. 17:33–46.
Gregory TR. 2005. Genome size evolution in animals. Evol Genome 1:4–87.
Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8:1494–1512.
Hatch FT, Bodner AJ, Mazrimas JA, Moore DH. 1976. Satellite DNA and cytogenetic evolution. DNA quantity, satellite DNA and karyotypic variations in kangaroo rats (Genus *Dipodomys*). Chromosoma 58:155–168.
Hellsten U, et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328:633–636.
Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160:1651–1659.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11:1.
Koch P, Platzer M, Downie BR. 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 42:e80.
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595.
Liu B, et al. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint arXiv:1308.2012.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770.
Mares MA, et al. 1997. How desert rodents overcome halophytic plant defenses. BioScience 47:699–704.
Meredith RW, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.
Ojeda RA, et al. 1999. Evolutionary convergence of the highly adapted desert rodent *Tympanoctomys barrerae* (Octodontidae). J Arid Environ. 41:443–452.
Ojeda RA, et al. 1996. Ecological observations of the red vizcacha rat *Tympanoctomys barrerae* in desert habitats of Argentina. Mastozool Neotrop. 3:183–191.
Orr HA. 1990. Why polyploidy is rarer in animals than in plants" revisited. Amer Nat. 136:759–770.
Renny-Byfield S, et al. 2011. Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. Mol Biol Evol. 28:2843–2854.
Renny-Byfield S, et al. 2012. Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of Nicotiana tabacum. PLoS One 7:e36963.
Rossi MS, Pesce C, Kornblihtt AR, Zorzópulos J. 1995. Origin and evolution of a major satellite DNA from South American rodents of the genus *Ctenomys*. Rev Chil Hist Nat. 68:171–183.
Session AM, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 538:336–343.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
Slamovits CH, Cook JA, Lessa EP, Rossi MS. 2001. Recurrent amplifications and deletions of satellite DNA accompanied chromosomal diversification in South American tuco-tucos (genus *Ctenomys*, Rodentia: Octodontidae): a phylogenetic approach. Mol Biol Evol. 18:1708–1719.

Sobrero R, Campos VE, Giannoni SM, Ebensperger LA. 2010. *Octomys mimax* (Rodentia: Octodontidae). Mamm Species 42:49–57.
Soh YS, et al. 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell 159:800–813.
Suárez-Villota EY, González-Wevar CA, Gallardo MH, Vásquez RA, Poulin E. 2016. Filling phylogenetic gaps and the biogeographic relationships of the Octodontidae (Mammalia: Hystricognathi). Mol Phylogenet Evol. 105:96–101.
Suárez-Villota EY, et al. 2012. Distribution of repetitive DNAs and the hybrid origin of the red vizcacha rat (Octodontidae). Genome 55:105–117.
Svartman M, Stone G, Stanyon R. 2005. Molecular cytogenetics discards polyploidy in mammals. Genomics 85:425–430.
Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.
Torres-Mura JC, Lemus ML, Contreras LC. 1989. Herbivorous specialization of the South American desert rodent *Tympanoctomys barrerae*. J Mammal. 70:646–648.
Tymowska J, Green D, Sessions S. 1991. Polyploidy and cytogenetic variation in frogs of the genus Xenopus. In: Green DM, Sessions SK, editors. Amphibian cytogenetics and evolution. San Diego (CA): Academic Press, p. 259–297.
Upham NS, Patterson BD. 2012. Diversification and biogeography of the Neotropical caviomorph lineage Octodontoidea (Rodentia: Hystricognathi). Mol Phylogenet Evol. 63:417–429.
Upham NS, Patterson BD. 2015. Evolution of Caviomorph rodents: a complete phylogeny and timetree for living genera. Mastozool Neotrop. 23:63–120.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2:333–341.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Associate editor: Jay Storz