

Characterization of the transcriptome, nucleotide sequence polymorphism, and natural selection in the desert adapted mouse *Peromyscus eremicus*

Matthew D. MacManes¹, Michael B. Eisen²,

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

² HHMI and University of California, Berkeley, Berkeley, CA, USA

* E-mail: macmanes@gmail.com, @PeroMHC

1 Abstract

2 As a direct result of intense heat and aridity, deserts are thought to be among the most
 3 harsh of environments, particularly for their mammalian inhabitants. Given that os-
 4 moregulation can be challenging for these animals, with failure resulting in death, strong
 5 selection should be observed on genes related to the maintenance of water and solute
 6 balance. One such animal, *Peromyscus eremicus*, is native to the desert regions of the
 7 southwest United States and may live its entire life without oral fluid intake. As a first
 8 step toward understanding the genetics that underlie this phenotype, we present a char-
 9 acterization of the *P. eremicus* transcriptome. We assay four tissues (kidney, liver, brain,
 10 testes) from a single individual and supplement this with population level renal transcrip-
 11 tome sequencing from 15 additional animals. We identified a set of transcripts undergoing
 12 both purifying and balancing selection based on estimates of Tajima's D. In addition, we
 13 used the branch-site test to identify a transcript – Slc2a9, likely related to desert os-
 14 moregulation – undergoing enhanced selection in *P. eremicus* relative to a set of related
 15 non-desert rodents.

16

17 Introduction

18 Deserts are widely considered one of the harshest environments on Earth. Animals living
 19 in desert environments are forced to endure intense heat and drought, and as a result,
 20 species living in these environments are likely to possess specialized mechanisms to deal
 21 with them. While living in deserts likely involves a large number of adaptive traits, the
 22 ability to osmoregulate – to maintain the proper water and electrolyte balance – appears

to be paramount (Walsberg, 2000). Indeed, the maintenance of water balance is one of the most important physiologic processes for all organisms, whether they be desert inhabitants or not. Most animals are exquisitely sensitive to changes in osmolality, with slight derangement eliciting physiologic compromise. When the loss of water exceeds dietary intake, dehydration - and in extreme cases, death - can occur. Thus there has likely been strong selection for mechanisms supporting optimal osmoregulation in species that live where water is limited. Understanding these mechanisms will significantly enhance our understanding of the physiologic processes underlying osmoregulation in extreme environments, which will have implications for studies of human health, conservation, and climate change.

The genes and structures responsible for the maintenance of water and electrolyte balance are well characterized in model organisms such as mice (Tatum et al., 2009), rats (Romero et al., 2007; Rojek et al., 2006; Nielsen et al., 1995), and humans (Mobasheri et al., 2007; Bedford et al., 2003; Nielsen et al., 1999). These studies, many of which have been enabled by newer sequencing technologies, provide a foundation for studies of renal genomics in non-model organisms. Because researchers have long been interested in desert adaptation, a number of studies have looked at the morphology or expression of single genes in the renal tissues of desert adapted rodents *Phyllotis darwini* (Gallardo et al., 2005), *Psammomys obesus* (Kaissling et al., 1975), and *Perognathus penicillatus* (Altschuler et al., 1979). More recently, full renal transcriptomes have been generated for *Dipodomys spectabilis* and *Chaetodipus baileyi*, (Marra et al., 2014) as well as *Abrothrix olivacea* (Giorello et al., 2014).

These studies provide a rich context for current and future work aimed at developing a synthetic understanding of the genetic and genomic underpinnings of desert adaptation in rodents. As a first step, we have sequenced, assembled, and characterized the transcriptome (using four tissue types - liver, kidney, testes and brain) of a desert adapted cricetid rodent endemic to the southwest United States, *Peromyscus eremicus*. These animals have a lifespan typical of small mammals (Veal and Caire, 2001), and therefore an individual may live its entire life without ever drinking water. Additionally, they have a distinct advantage over other desert animals (e.g. *Dipodomys*) in that they breed readily in captivity, which enables future laboratory studies of the phenotype of interest. In addition, the focal species is positioned in a clade of well known animals (e.g. *P. californicus*,

57 *P. maniculatus*, and *P. polionotus*) (Feng et al., 2007) with growing genetic and genomic
 58 resources (Shorter et al., 2014; Panhuis et al., 2011; Shorter et al., 2012). Together, this
 59 suggests that future comparative studies are possible.

60
 61 While the elucidation of the mechanisms underlying adaptation to desert survival is
 62 beyond the scope of this manuscript, we aim to lay the groundwork by characterizing the
 63 transcriptome from four distinct tissues (brain, liver, kidney, testes). These data will be
 64 included in the current larger effort aimed at sequencing the entire genome. Further, via
 65 sequencing the renal tissue of a total of 15 additional animals, we characterize nucleotide
 66 polymorphism and genome-wide patterns of natural selection. Together, these investiga-
 67 tions will aid in our overarching goal to understand the genetic basis of adaptation to
 68 deserts in *P. eremicus*.

69 **Materials and Methods**

70 **Animal Collection and Study Design**

71 To begin to understand how genes may underlie desert adaptation, we collected 16 adult
 72 individuals (9 male, 7 female) from a single population of *P. eremicus* over a two-year time
 73 period (2012-2013). These individuals were captured in live traps and then euthanized
 74 using isoflurane overdose and decapitation. Immediately post-mortem, the abdominal and
 75 pelvic organs were removed, cut in half (in the case of the kidneys), placed in RNAlater and
 76 flash frozen in liquid nitrogen. Removal of the brain, with similar preservation techniques,
 77 followed. Time from euthanasia to removal of all organs never exceeded five minutes.
 78 Samples were transferred to a -80C freezer at a later date. These procedures were approved
 79 by the Animal Care and Use Committee located at the University of California Berkeley
 80 (protocol number R224) and University of New Hampshire (protocol number 130902) as
 81 well as the California Department of Fish and Game (protocol SC-008135) and followed
 82 guidelines established by the American Society of Mammalogy for the use of wild animals
 83 in research (Sikes et al., 2011).

84 **RNA extraction and Sequencing**

85 Total RNA was extracted from each tissue using a TRIzol extraction (Invitrogen) following
 86 the manufacturer's instructions. Because preparation of an RNA library suitable for

sequencing is dependent on having high quality, intact RNA, a small aliquot of each total RNA extract was analyzed on a Bioanalyzer 2100 (Agilent, Palo Alto, CA, USA). Following confirmation of sample quality, the reference sequencing libraries were made using the TruSeq stranded RNA prep kit (Illumina), while an unstranded TruSeq kit was used to construct the other sequencing libraries. A unique index was ligated to each sample to allow for multiplexed sequencing. Reference libraries (n=4 tissue types from Peer360, a male mouse used for generating a genome sequence - not part of the current study) were then pooled to contain equimolar quantities of each individual library and submitted for Illumina sequencing using two lanes of 150nt paired end sequencing employing the rapid-mode of the HiSeq 2500 sequencer at The Hubbard Center for Genome Sciences (University of New Hampshire). The remaining 15 libraries were similarly multiplexed and sequenced in a mixture of 100nt paired and single end sequencing runs across several lanes of an Illumina HiSeq 2000 at the Vincent G. Coates Genome Center (University of California, Berkeley).

Sequence Data Preprocessing and Assembly

The raw sequence reads corresponding to the four tissue types were error corrected using the software *bleed* version 0.17 (Heo et al., 2014) using *kmer*=25, based on the developer's default recommendations. The error-corrected sequence reads were adapter and quality trimmed following recommendations from MacManes (MacManes, 2014) and Mbandi (Mbandi et al., 2014). Specifically, adapter sequence contamination and low quality nucleotides (defined as *PHRED* <2) were removed using the program *Trimmomatic* version 0.32 (Lohse et al., 2012). Reads from each tissue were assembled using the *Trinity* version released 17 July 2014 (Haas et al., 2013). We used flags to indicate the stranded nature of sequencing reads and set the maximum allowable physical distance between read pairs to 999nt. We elected to assemble reads derived from a single deeply sequenced individual (Peer360, a male) to reduce polymorphism and thus the complexity of the de Bruijn graph, which has important implications for runtime, hardware requirements (Pop, 2009), and assembly contiguity (?). Individual tissues were assembled independently, as we hypothesize that tissue specific isoforms would be reconstructed with higher fidelity than if all tissues were assembled together.

The assembly was conducted on a linux workstation with 64 cores and 512Gb RAM. To filter the raw sequence assembly, we downloaded *Mus musculus* cDNA and ncRNA datasets from Ensembl (ftp://ftp.ensembl.org/pub/release-75/fasta/mus_musculus/)

120 and the *Peromyscus maniculatus* reference transcriptome from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Peromyscus_maniculatus_bairdii/RNA/). We used a blastN
 121 (version 2.2.29+) procedure (default settings, evaluate set to 10^{-10}) to identify contigs
 122 in the *P. eremicus* dataset likely to be biological in origin. This procedure, when a ref-
 123 erence dataset is available, retains more putative transcripts than a strategy employing
 124 expression-based filtering (remove if transcripts per million (TPM) < 1 (**MacManes** and
 125 Lacey, 2012)) of the raw assembly. We then concatenated the filtered assemblies from
 126 each tissue into a single file and reduced redundancy using the software cd-hit-est version
 127 4.6 (Li and Godzik, 2006) using default settings, except that sequences were clustered
 128 based on 95% sequence similarity. This multi-fasta file was used for all subsequent anal-
 129 yses, including annotation and mapping.

131

132 Assembled Sequence Annotation

133 The filtered assemblies were annotated using the default settings of the blastN algorithm
 134 (Camacho et al., 2009) against the Ensembl cDNA and ncRNA datasets described above,
 135 downloaded on 1 August 2014. Among other things, the Ensemble transcript identi-
 136 fiers were used in the analysis of gene ontology conducted in the PANTHER package
 137 (Mi, 2004). Next, because rapidly evolving nucleotide sequences may evade detection by
 138 blast algorithms, we used HMMER3 version 3.1b1 (Wheeler and Eddy, 2013) to search
 139 for conserved protein domains contained in the dataset using the Pfam database (Punta
 140 et al., 2012). Lastly, we extracted putative coding sequences using Transdecoder version
 141 4Jul2014 (<http://transdecoder.sourceforge.net/>)

142

143 To identify patterns of gene expression unique to each tissue type, we mapped sequence
 144 reads from each tissue type to the reference assembly using bwa-mem (version cloned from
 145 Github 7/1/2014) (Li, 2013). We estimated expression for the four tissues individually
 146 using default settings of the software eXpress version 1.51 (Roberts and Pachter, 2013).
 147 Interesting patterns of expression, including instances where expression was limited to a
 148 single tissue type, were identified and visualized.

149

150 Population Genomics

151 In addition to the reference individual sequenced at four different tissue types, we se-
 152 quenced 15 other conspecific individuals from the same population in Palm Desert, Cali-
 153 fornia. Sequence data were mapped to the reference transcriptome using bwa-mem. The
 154 alignments were sorted and converted to BAM format, then passed to the program ANGSD
 155 version 0.610, which was used for calculating the folded site frequency spectrum (SFS)
 156 and Tajima's D (Korneliussen et al., 2013).

158 Natural Selection

159 To characterize natural selection on several genes related to water and ion homeostasis, we
 160 identified several of the transcripts identified as experiencing positive selection in a recent
 161 work on desert-adapted Heteromyid rodents (Marra et al., 2014). The coding sequences
 162 corresponding to these genes, Solute Carrier family 2 member 9 (Slc2a9), the Vitamin
 163 D3 receptor (Vdr) and several of the Aquaporin genes (Aqp1,2,4,9), were extracted from
 164 the dataset, aligned using the software MACSE version 1.01b (Ranwez et al., 2011) to
 165 homologous sequences in *Mus musculus*, *Rattus norvegicus*, *Peromyscus maniculatus*, and
 166 *Homo sapiens* as identified by the conditional reciprocal best blast procedure (CRBB,
 167 (Aubry et al., 2014)). An unrooted gene tree with branch lengths was constructed using
 168 the online resource ClustalW2-Phylogeny ([http://www.ebi.ac.uk/Tools/phylogeny/
 169 clustalw2_phylogeny/](http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/)), and the tree and alignment were analyzed using the branch-
 170 site model (model=2, nsSites=2, fix_omega=0 versus model=2, nsSites=2, fix_omega=1,
 171 omega=1) implemented in PAML version 4.8 (Yang and dos Reis, 2011; Yang, 2007).
 172 Significance was evaluated via the use of the likelihood ratio test.

174 Results and Discussion

175 RNA extraction, Sequencing, Assembly, Mapping

176 RNA was extracted from the hypothalamus, renal medulla, testes, and liver from each
 177 individual using sterile technique. TRIzol extraction resulted in a large amount of high
 178 quality (RIN \geq 8) total RNA, which was then used as input. Libraries were constructed

as per the standard Illumina protocol and sequenced as described above. The number of reads per library varied from 56 million strand-specific paired-end reads in Peer360 kidney, to 9 million single-end reads in Peer321 (Table 1, available on the Short Read Archive accession XXX). Adapter sequence contamination and low-quality nucleotides were eliminated, which resulted in a loss of <2% of the total number of reads. These trimmed reads served as input for all downstream analyses.

Table 1

186

DATASET	NUM. RAW READS	SRA ACCESSION
PEER360 TESTES	32M PE/SS	SRR1575398
PEER360 LIVER	53M PE /SS	SRR1575397
PEER360 KIDNEY	56M PE/SS	SRR1575396
PEER360 BRAIN	23M PE/SS	SRR1575395
PEER305	19M PE	SRR1575434
PEER308	15M PE	SRR1575437
PEER319	14M PE	SRR1575439
PEER321	9M SE	SRR1575441
PEER340	16M PE	SRR1575443
PEER352	14M PE	SRR1575464
PEER354	9M SE	SRR1575466
PEER359	14M PE	SRR1575492
PEER365	16M PE	SRR1575493
PEER366	16M PE	SRR1575494
PEER368	14M PE	SRR1575624
PEER369	14M PE	SRR1575625
PEER372	17M SE	SRR1576070
PEER373	23M SE	SRR1576071
PEER380	16M SE	SRR1576072
PEER382	14M SE	SRR1576073

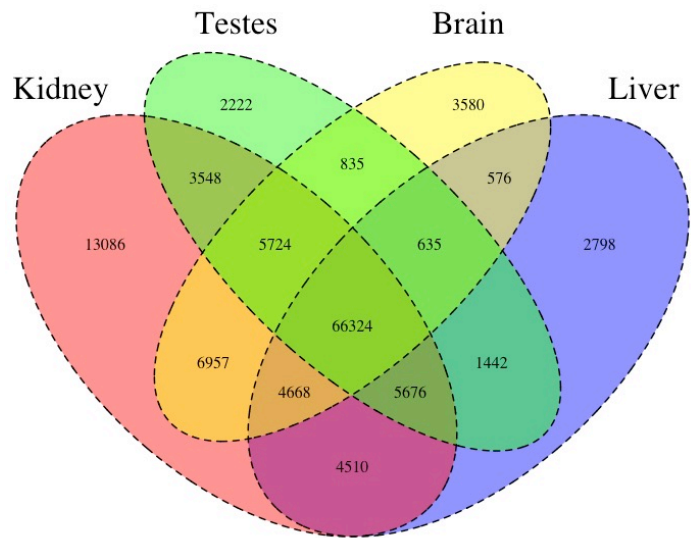
Table 1. The number of sequencing reads per sample, whose identity is indicated by Peer[number]. PE=paired end, SS=strand specific, SE=single end sequencing.

Transcriptome assemblies for each tissue type was accomplished using the program Trinity (Haas et al., 2013). The raw assemblies for brain, liver, testes, and kidney con-

192 tained 185425, 222096, 180233, and 514091 assembled sequences respectively. This as-
193 sembly was filtered using a blastN procedure against the *Mus* cDNA and ncRNA and
194 *P. maniculatus* cDNAs, which resulted in a final dataset containing 68331 brain-derived
195 transcripts, 71041 liver-derived transcripts, 67340 testes-derived transcripts, and 113050
196 kidney-derived transcripts. Mapping the error-corrected adapter/quality trimmed reads
197 to these datasets resulted in mapping 94.98% (87.01% properly paired) of the brain-
198 derived reads to the brain transcriptome, 96.07% (88.13% properly paired) of the liver-
199 derived reads to the liver transcriptome, 96.81% (85.10% properly paired) of the testes-
200 derived reads to the testes transcriptome, and 91.87% (83.77% properly paired) of the
201 kidney-derived reads to the kidney transcriptome. Together, these statistics suggest that
202 the tissue-specific transcriptomes are of extremely high quality. All tissue-specific assem-
203 blies are to be made available on Dryad, and until then are stored on Dropbox (https://www.dropbox.com/sh/2jwzd8p6n6eluco/AAA03nSdXb_u4wtQZRTwqW9ia?dl=0).
204

205
206

Figure 1



207
208
209
210

Figure 1. The Venn Diagram, which provides a visual representation of the overlap of expression of the four tissue types. The majority of transcripts (66,324) are expressed in all studied tissue types.

We then estimated gene expression on each of these tissue-specific datasets, which allowed us to understand expression patterns in the multiple tissues (Pero.tissue.xprs, will be made available on Dryad, until then on Dropbox (https://www.dropbox.com/sh/2jwzd8p6n6eluco/AAA03nSdXb_u4wtQZRTwqW9ia?dl=0)). Specifically, we constructed a Venn diagram (Figure 1) that allowed us to visualize the proportion of genes whose expression was limited to a single tissue and those whose expression was ubiquitous. 66324 transcripts are expressed on all tissue types, while 13086 are uniquely expressed in the kidney, 2222 in the testes, 3580 in the brain, and 2798 in the liver. The kidney appears to an outlier in the number of unique sequences, though this could be the result of the recovery of more lowly expressed transcripts or isoforms.

In addition to this, we estimated mean TPM (number of transcripts per million) for all transcripts. Table 2 consists of the 10 genes whose mean TPM was the highest. Several genes in this list are predominately present in a single tissue type. For instance Transcript_126459, Albumin is very highly expressed in the liver, but less so in the other tissues. It should be noted, however, that making inference based on uncorrected values for TPM is not warranted. Statistical testing for differential expression was not implemented due to the fact that no replicates are available.

After expression estimation, the filtered assemblies were concatenated together, and after the removal of redundancy with cd-hit-est, 123,123 putative transcripts remained (to be made available on Genbank, and until then are stored on Dropbox https://www.dropbox.com/sh/2jwzd8p6n6eluco/AAA03nSdXb_u4wtQZRTwqW9ia?dl=0). From this filtered concatenated dataset, we extracted 71626 putative coding sequences (72Mb, to be made available on Dryad). Of these 71626 sequences, 38221 contained complete open reading frames (containing both start and stop codons), while the others were either truncated at the 5-prime end (20239 sequences), the 3-prime end (6445 sequences), or were internal (6721 sequencing with neither stop nor start codon). The results of a Pfam search conducted on the predicted amino acid sequences will be found on Dryad, and until then are stored on Dropbox https://www.dropbox.com/sh/2jwzd8p6n6eluco/AAA03nSdXb_u4wtQZRTwqW9ia?dl=0.

Table 2

	Transcript ID	Testes	Liver	Kidney	Brain	Genbank ID	Gene ID
	Transcript_83842	2.05E+03	6.40E+03	1.03E+04	5.47E+03	DQ073446.1	COX2
	Transcript_126459	1.43E+01	2.22E+04	2.77E+01	6.73E+00	XM_006991665.1	Alb
	Transcript_128937	4.39E+00	1.91E+04	4.74E+02	2.23E+00	XM_007627625.1	Apoa2
	Transcript_81233	1.71E+03	5.23E+03	6.11E+03	3.08E+03	XM_006993867.1	Fth1
245	Transcript_94125	3.67E+01	1.08E+04	2.09E+03	2.75E+00	XM_006977178.1	CytP450
	Transcript_119945	5.03E+03	1.15E+03	1.33E+03	3.71E+03	XM_008686011.1	Ubb
	Transcript_5977	4.95E+00	1.01E+04	3.05E+02	3.58E+02	XM_006978668.1	Tf
	Transcript_4057	2.62E+01	9.32E+03	1.34E+02	8.38E+01	XM_006994871.1	Apoc1
	Transcript_112523	4.07E+02	7.36E+03	7.78E+02	9.54E+02	XM_006994872.1	Apoe
	Transcript_98376	1.98E+00	8.66E+03	1.02E+00	2.68E+00	XM_006970208.1	Ttr

246 Table 2. The 10 transcripts with the highest mean TPM (transcripts per million).

247 Population Genomics

248 As detailed above, RNAseq data from 15 individuals were mapped to the reference tran-
 249 scriptome with the resulting BAM files being used as input to the software package
 250 ANGSD. The Tajima's D statistic was calculated for all transcripts covered by at least
 251 14 of the 15 individuals. In brief, a negative Tajima's D - a result of lower than expected
 252 average heterozygosity - is often associated with purifying or directional selection, recent
 253 selective sweep or population bottleneck. In contrast, a positive value for Tajima's D
 254 represents higher than expected average heterozygosity, often associated with balancing
 255 selection.

256

257 The distribution of the estimates of Tajima's D for all of the assembled transcripts is
 258 shown in Figure 2. Although Tajima's D is known to be sensitive to demographic history,
 259 which is largely unknown for this population, the estimates may also be drive by patterns
 260 of selection. In general, the distribution is skewed toward negative values (mean=-0.89,
 261 variance=0.58), which may be the result of purifying selection, a model of evolution com-
 262 monly invoked for coding DNA sequences (Chamary et al., 2006). Table 3 presents the
 263 10 transcripts whose estimate of Tajima's D is the greatest, while Table 4 presents the
 264 10 transcripts whose estimate of Tajima's D is the least. The former list of genes may
 265 contain transcripts experiencing balancing selection in the studied population. This list
 266 includes, interestingly, genes obviously related to solute and water balance (e.g. Clcnkb

and a transmembrane protein gene) and immune function (a interferon-inducible GTPase and a Class 1 MHC gene). The latter group, containing transcripts whose estimates of Tajima's D are the smallest are likely experiencing purifying selection. Many of these transcripts are involved in core regulatory functions where mutation may have strongly negative fitness consequences.

Figure 2

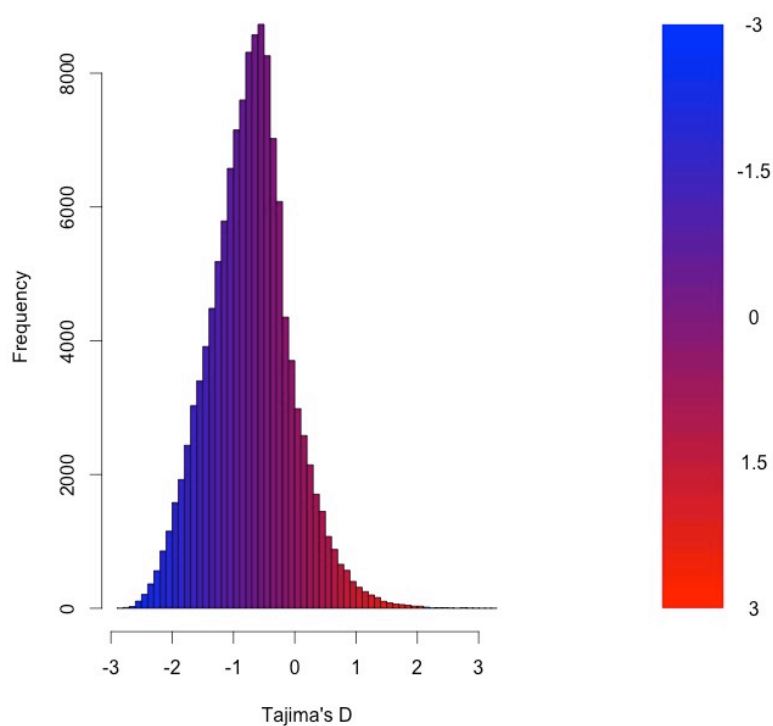


Figure 2. The distribution of Tajima's D for all putative transcripts.

Table 3

Transcript ID	GenBank ID	Description	Tajima's D
Transcript_49049	XM_006533884.1	heterogeneous nuclear ribonucleoprotein H1 (Hnrnph1)	3.26
Transcript_38378	XM_006522973.1	Son DNA binding protein (Son)	3.19
Transcript_126187	NM_133739.2	transmembrane protein 123 (Tmem123)	3.02
Transcript_70953	XM_006539066.1	chloride channel Kb (Clcnkb)	2.96
Transcript_37736	XM_006997718.1	h-2 class I histocompatibility antigen	2.92
Transcript_21448	XM_006986148.1	zinc finger protein 624-like	2.84
Transcript_47450	NM_009560.2	zinc finger protein 60 (Zfp60)	2.82
Transcript_122250	XM_006539068.1	chloride channel Kb (Clcnkb)	2.81
Transcript_78367	XM_006496814.1	CDC42 binding protein kinase alpha (Cdc42bpa)	2.78
Transcript_96470	XM_006987129.1	interferon-inducible GTPase 1-like	2.77

Table 3. The 10 transcripts with the highest values for Tajima's D, which suggests balancing selection.

280 **Table 4**

281

Transcript ID	GenBank ID	Description	Tajima's D
Transcript_84359	XM_006991127.1	nuclear receptor coactivator 3 (Ncoa3)	-2.82
Transcript_87121	XM_006970128.1	methyl-CpG binding domain protein 2 (Mbd2)	-2.82
Transcript_125755	EU053203.1	alpha globin gene cluster	-2.78
Transcript_87128	XM_006976644.1	membrane-associated ring finger (March5)	-2.76
Transcript_55468	XM_006978377.1	Vpr binding protein (Vprbp)	-2.75
Transcript_116042	XM_006980811.1	membrane associated guanylate kinase (Magi3)	-2.75
Transcript_18966	XM_006982814.1	ubiquitin protein ligase E3 component n-recognin 5 (Ubr5)	-2.75
Transcript_122204	XM_008772511.1	zinc finger protein 612 (Zfp612)	-2.75
Transcript_100550	XM_006971297.1	bromodomain adjacent to zinc finger domain, 1B (Baz1b)	-2.74
Transcript_33267	XM_006975561.1	pumilio RNA-binding family member 1 (Pum1)	-2.75

Table 4. The 10 transcripts with the lowest values for Tajima's D, which suggests purifying or directional selection.

285 Natural Selection

286 To begin to test the hypothesis that selection on transcripts related to osmoregulation is
 287 enhanced in the desert adapted *P. eremicus*, we calculated Tajima's D as described above,

and implemented the branch-site test using alignments produced in MACSE. These alignments were manually inspected, and were relatively free from indels and internal stop codons. We set the sequence corresponding to *P. eremicus* for both Slc2a9, Vdr, and several of the Aquaporin genes (Aqp1,2,4,9) as the foreground lineages in 2 distinct program executions. These transcripts Slc2a9 and Vdr were chosen specifically because they - the former significantly - were recently linked to osmoregulation in a desert rodent (Marra et al., 2014). The test for Slc2a9 was highly significant ($2\Delta\text{LnL}=51.4$, $\text{df}=1$, $p=0$, Table 5), indicating enhanced selection in *P. eremicus* relative to the other lineages. The branch site test for positive selection conducted on the Vdr and Aquaporin genes were non-significant. While the branch site test of positive selection is largely non-significant, estimating Tajima's D for these few candidate loci demonstrates that either a selective or demographic process may be influencing the genome at these functionally relevant sites.

Table 5

Transcript ID	Description	Tajima's D	Branch Site Test p.value
Transcript_106085	Slc2a9	2.15	p=0
Transcript_114624	Vdr	1.97	p=1
Transcript_128972	Aqp1	1.39	p=1
Transcript_33960	Aqp2	1.78	p=1
Transcript_22154	Aqp4	2.10	p=1
Transcript_107677	Aqp9	2.06	p=1

Table 5. Several candidate genes were evaluated using Tajima's D and the branch site test implemented in PAML.

Conclusions

As a direct result of intense heat and aridity, deserts are thought to be amongst the harshest environments, particularly for mammalian inhabitants. Given that osmoregulation can be challenging for these animals - with failure resulting in death - strong selection should be observed on genes related to the maintenance of water and solute balance. This study aimed to characterize the transcriptome of a desert-adapted rodent species, *P. eremicus*. Specifically, we characterized the transcriptome of four tissue types (liver, kidney, brain, and testes) from a single individual and supplemented this with population-level renal

transcriptome sequencing from 15 additional animals. We identified a set of transcripts undergoing both purifying and balancing selection based on Tajima's D. In addition, we used a branch site test to identify a transcript, likely related to desert osmoregulation, undergoing enhanced selection in *P. eremicus* relative to a set of non-desert rodents.

Acknowledgments

This manuscript was greatly improved by careful review from C. Titus Brown, Elijah Lowe, and an anonymous reviewer, as well as by Matthew Hahn, who provided feedback on an earlier version of the manuscript posted on bioRxiv.

References

- Altschuler, E. M., Nagle, R. B., Braun, E. J., Lindstedt, S. L., and Krutzsch, P. H. (1979). Morphological study of the desert heteromyid kidney with emphasis on the genus *Perognathus*. *The Anatomical Record*, 194(3):461–468.
- Aubry, S., Kelly, S., Kumpers, B. M. C., Smith-Unna, R. D., and Hibberd, J. M. (2014). Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLOS Genetics*, 10(6):e1004365.
- Bedford, J. J., Leader, J. P., and Walker, R. J. (2003). Aquaporin expression in normal human kidney and in renal disease. *Journal of the American Society of Nephrology*, 14(10):2581–2587.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.
- Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2):98–108.
- Feng, B.-J., Sun, L.-D., Soltani-Arabshahi, R., Bowcock, A. M., Nair, R. P., Stuart, P., Elder, J. T., Schrodi, S. J., Begovich, A. B., Abecasis, G. R., Zhang, X.-J., Callis-

- 341 Duffin, K. P., Krueger, G. G., and Goldgar, D. E. (2007). Toward a Molecular Phylogeny
342 for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b Sequences. *Journal of*
343 *Mammalogy*, 88(5):1146–1159.
- 344 Gallardo, P. A., Cortés, A., and Bozinovic, F. (2005). Phenotypic flexibility at the molec-
345 ular and organismal level allows desert-dwelling rodents to cope with seasonal water
346 availability. *Physiological and Biochemical Zoology*, 78(2):145–152.
- 347 Giorello, F. M., Feijoo, M., D’Elía, G., Valdez, L., Opazo, J. C., Varas, V., Naya, D. E.,
348 and Lessa, E. P. (2014). Characterization of the kidney transcriptome of the South
349 American olive mouse *Abrothrix olivacea*. *BMC Genomics*, 15(1):446.
- 350 Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J.,
351 Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J.,
352 Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel,
353 R., Leduc, R. D., Friedman, N., and Regev, A. (2013). *De novo* transcript sequence
354 reconstruction from RNA-seq using the Trinity platform for reference generation and
355 analysis. *Nature Protocols*, 8(8):1494–1512.
- 356 Heo, Y., Wu, X.-L., Chen, D., Ma, J., and Hwu, W.-M. (2014). BLESS: Bloom filter-
357 based error correction solution for high-throughput sequencing reads. *Bioinformatics*,
358 30(10):1354–1362.
- 359 Kaissling, B., De Rouffignac, C., Barrett, J. M., and Kriz, W. (1975). The structural
360 organization of the kidney of the desert rodent *Psammomys obesus*. *Anatomy and*
361 *Embryology*, 148(2):121–143.
- 362 Korneliussen, Thorfinn Sand Moltke, I., Albrechtsen, a., and Nielsen, R. (2013). Calcula-
363 tion of Tajima’s D and other neutrality test statistics from low depth next-generation
364 sequencing data. *BMC Bioinformatics*, 14(1):289.
- 365 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-
366 MEM. *arXiv.org*.
- 367 Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large
368 sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

- 369 Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., and Usadel,
370 B. (2012). RobiNA: A user-friendly, integrated software solution for RNA-Seq-based
371 transcriptomics. *Nucleic Acids Research*, 40(Web Server issue):W622–7.
- 372 MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence
373 data. *Frontiers in Genetics*, 5.
- 374 Marra, N. J., Romero, a., and DeWoody, J. a. (2014). Natural selection and the genetic
375 basis of osmoregulation in heteromyid rodents as revealed by RNA-seq. *Molecular*
376 *Ecology*, 23(11):2699–2711.
- 377 Mbandi, S. K., Hesse, U., Rees, D. J. G., and Christoffels, A. (2014). A glance at quality
378 score: Implication for *de novo* transcriptome reconstruction of Illumina reads. *Frontiers*
379 *in Genetics*, 5:17.
- 380 Mi, H. (2004). The PANTHER database of protein families, subfamilies, functions and
381 pathways. *Nucleic Acids Research*, 33(Database issue):D284–D288.
- 382 Mobasheri, A., Marples, D., Young, I. S., Floyd, R. V., Moskaluk, C. A., and Frigeri,
383 A. (2007). Distribution of the AQP4 Water Channel in Normal Human Tissues: Pro-
384 tein and Tissue Microarrays Reveal Expression in Several New Anatomical Locations,
385 including the Prostate Gland Seminal Vesicles. *Channels*, 1(1):30–39.
- 386 Nielsen, S., Chou, C., Marples, D., Christensen, E., Kishore, B., and Knepper, M. (1995).
387 Vasopressin increases water permeability of kidney collecting duct by inducing translo-
388 cation of aquaporin-CD water channels to plasma-membrane. *Proceedings of The Na-*
389 *tional Academy of Sciences of The United States of America*, 92(4):1013–1017.
- 390 Nielsen, S., Kwon, T. H., Christensen, B. M., Promeneur, D., Frøkiaer, J., and Marples, D.
391 (1999). Physiology and pathophysiology of renal aquaporins. *Journal of the American*
392 *Society of Nephrology*, 10(3):647–663.
- 393 Panhuis, T. M., Broitman-Maduro, G., Uhrig, J., Maduro, M., and Reznick, D. N. (2011).
394 Analysis of Expressed Sequence Tags from the Placenta of the Live-Bearing Fish Poe-
395 ciliopsis (Poeciliidae). *Journal of Heredity*, 102(3):352–361.
- 396 Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings In*
397 *Bioinformatics*, 10(4):354–366.

- 398 Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang,
399 N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L.,
400 Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database.
401 *Nucleic Acids Research*, 40:D290–301.
- 402 Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: Multiple
403 Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS*
404 *ONE*, 6(9):e22594.
- 405 Roberts, A. and Pachter, L. (2013). Streaming fragment assignment for real-time analysis
406 of sequencing experiments. *Nature Methods*, 10(1):71–73.
- 407 Rojek, A., Rojek, A., Fuchtbauer, E., Fuchtbauer, E., Kwon, T., Kwon, T., Frokiaer, J.,
408 and Nielsen, S. (2006). Severe urinary concentrating defect in renal collecting duct-
409 selective AQP2 conditional-knockout mice. *Proceedings of The National Academy of*
410 *Sciences of The United States of America*, 103(15):6037–6042.
- 411 Romero, D. G., Plonczynski, M. W., Welsh, B. L., Gomez-Sanchez, C. E., Zhou, M. Y.,
412 and Gomez-Sanchez, E. P. (2007). Gene expression profile in rat adrenal zona glomeru-
413 losa cells stimulated with aldosterone secretagogues. *Physiological Genomics*, 32(1):117–
414 127.
- 415 Shorter, K. R., Crossland, J. P., Webb, D., Szalai, G., Felder, M. R., and Vrana, P. B.
416 (2012). *Peromyscus* as a Mammalian Epigenetic Model. *Genetics Research Interna-*
417 *tional*, 2012:1–11.
- 418 Shorter, K. R., Owen, A., anderson, V., Hall-South, A. C., Hayford, S., Cakora, P.,
419 Crossland, J. P., Georgi, V. R. M., Perkins, A., Kelly, S. J., Felder, M. R., and Vrana,
420 P. B. (2014). Natural genetic variation underlying differences in *Peromyscus* repetitive
421 and social/aggressive behaviors. *Behavior genetics*, 44(2):126–135.
- 422 Sikes, R. S., Gannon, W. L., and Animal Care and Use Committee of the American
423 Society of Mammalogists (2011). Guidelines of the American Society of Mammalogists
424 for the use of wild mammals in research. *Journal of Mammalogy*, 92(1):235–253.
- 425 Tatum, R., Zhang, Y., Salleng, K., Lu, Z., Lin, J. J., Lu, Q., Jeansonne, B. G., Ding,
426 L., and Chen, Y. H. (2009). Renal salt wasting and chronic dehydration in claudin-7-
427 deficient mice. *AJP: Renal Physiology*, 298(1):F24–F34.

- 428 **MacManes**, M. D. and Lacey, E. A. (2012). The Social Brain: Transcriptome Assembly
429 and Characterization of the Hippocampus from a Social Subterranean Rodent, the
430 Colonial Tuco-Tuco (*Ctenomys sociabilis*). *PLOS ONE*, 7(9):e45524.
- 431 Veal, R. and Caire, W. (2001). *Peromyscus eremicus*. *Mammalian Species*, 118:1–6.
- 432 Walsberg, G. (2000). Small mammals in hot deserts: Some generalizations revisited.
433 *Bioscience*, 50(2):109–120.
- 434 Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: DNA homology search with profile
435 HMMs. *Bioinformatics*, 29(19):2487–2489.
- 436 Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular*
437 *Biology and Evolution*, 24(8):1586–1591.
- 438 Yang, Z. and dos Reis, M. (2011). Statistical Properties of the Branch-Site Test of Positive
439 Selection. *Molecular Biology and Evolution*, 28(3):1217–1228.