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

Matthew D. MacManes[1], Michael B. Eisen [2]

[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

## Abstract

As a direct result of intense heat and aridity, deserts are thought to be among the most harsh of environments, particularly for their 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. One such animal, *Peromyscus eremicus*, is native to the desert regions of the southwest United States and may live its entire life without oral fluid intake. As a first step toward understanding the genetics that underlie this phenotype, we present a characterization of the *P. eremicus* transcriptome. We assay four tissues (kidney, liver, brain, testes) from a single individual and supplement 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 estimates of Tajima's D. In addition, we used the branch-site test to identify a transcript – Slc2a9, likely related to desert osmoregulation – undergoing enhanced selection in *P. eremicus* relative to a set of related non-desert rodents.

## Introduction

Deserts are widely considered one of the harshest environments on Earth. Animals living in desert environments are forced to endure intense heat and drought, and as a result, species living in these environments are likely to possess specialized mechanisms to deal with them. While living in deserts likely involves a large number of adaptive traits, the 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*,

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

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

# Materials and Methods

## Animal Collection and Study Design

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

## RNA extraction and Sequencing

Total RNA was extracted from each tissue using a TRIzol extraction (Invitrogen) following 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 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 bless 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 (Bolger et al., 2014). 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 assembly 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 (Lowe et al., 2014; Pop, 2009), and assembly contiguity (Vijay et al., 2013). Individual tissues were assembled independently, as we hypothesize that tissue specific isoforms would be reconstructed with higher fidelity that 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/`)

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

## Assembled Sequence Annotation

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

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

## Population Genomics

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

## Natural Selection

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

# Results and Discussion

## RNA extraction, Sequencing, Assembly, Mapping

RNA was extracted from the hypothalamus, renal medulla, testes, and liver from each individual using sterile technique. TRIzol extraction resulted in a large amount of high 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 as part BioProject PRJNA242486). 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**

| 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-

tained 185425, 222096, 180233, and 514091 assembled sequences respectively. This assembly was filtered using a blastN procedure against the *Mus* cDNA and ncRNA and *P. maniculatus* cDNAs, which resulted in a final dataset containing 68331 brain-derived transcripts, 71041 liver-derived transcripts, 67340 testes-derived transcripts, and 113050 kidney-derived transcripts. Mapping the error-corrected adapter/quality trimmed reads to these datasets resulted in mapping 94.98% (87.01% properly paired) of the brain-derived reads to the brain transcriptome, 96.07% (88.13% properly paired) of the liver-derived reads to the liver transcriptome, 96.81% (85.10% properly paired) of the testes-derived reads to the testes transcriptome, and 91.87% (83.77% properly paired) of the kidney-derived reads to the kidney transcriptome. Together, these statistics suggest that the tissue-specific transcriptomes are of extremely high quality. All tissue-specific assemblies are to be made available on Dryad, and until then are stored on Dropbox (`https://www.dropbox.com/sh/2jwzd8p6n6eluco/AAAO3nSdXb_u4wtQZRTwqW9ia?dl=0`).

**Figure 1**

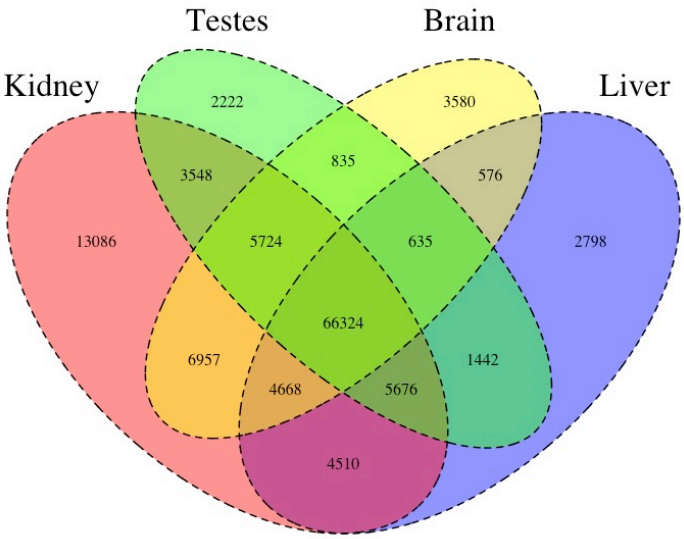


Figure 1. The Venn Diagram, which provides a visual representation of the overlap of expression of the four tissue types. The majority of transcripts (66,324) are expressed in all studied tissue types.

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/AAAO3nSdXb_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/AAAO3nSdXb_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/AAAO3nSdXb_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 |
| 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 |

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

## Population Genomics

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

The distribution of the estimates of Tajima's D for all of the assembled transcripts is shown in Figure 2. Although Tajima's D is known to be sensitive to demographic history, which is largely unknown for this population, the estimates may also be drive by patterns of selection. In general, the distribution is skewed toward negative values (mean=-0.89, variance=0.58), which may be the result of purifying selection, a model of evolution commonly invoked for coding DNA sequences (Chamary et al., 2006). Table 3 presents the 10 transcripts whose estimate of Tajima's D is the greatest, while Table 4 presents the 10 transcripts whose estimate of Tajima's D is the least. The former list of genes may contain transcripts experiencing balancing selection in the studied population. This list 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

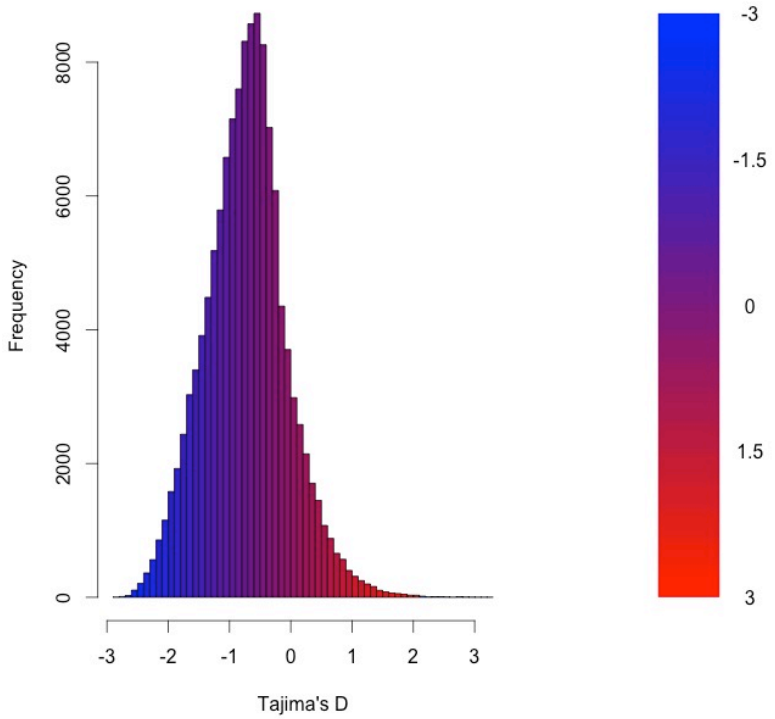


Figure 2. The distribution of Tajima's D for all putative transcripts.

## 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.

**Table 4**

| Transcript ID | GenBank ID | Description | Tajima's I |
|---|---|---|---|
| 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.

## Natural Selection

To begin to test the hypothesis that selection on transcripts related to osmoregulation is 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 six 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 ($2\Delta$Lnl=51.4, 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., Kümpers, 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.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):btu170–2120.

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-Duffin, K. P., Krueger, G. G., and Goldgar, D. E. (2007). Toward a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b Sequences. *Journal of Mammalogy*, 88(5):1146–1159.

Gallardo, P. A., Cortés, A., and Bozinovic, F. (2005). Phenotypic flexibility at the molecular and organismal level allows desert-dwelling rodents to cope with seasonal water availability. *Physiological and Biochemical Zoology*, 78(2):145–152.

Giorello, F. M., Feijoo, M., D'Elía, G., Valdez, L., Opazo, J. C., Varas, V., Naya, D. E., and Lessa, E. P. (2014). Characterization of the kidney transcriptome of the South American olive mouse *Abrothrix olivacea*. *BMC Genomics*, 15(1):446.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., Leduc, R. D., Friedman, N., and Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512.

Heo, Y., Wu, X.-L., Chen, D., Ma, J., and Hwu, W.-M. (2014). BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics*, 30(10):1354–1362.

Kaissling, B., De Rouffignac, C., Barrett, J. M., and Kriz, W. (1975). The structural organization of the kidney of the desert rodent *Psammomys obesus*. *Anatomy and Embryology*, 148(2):121–143.

Korneliussen, Thorfinn Sand Moltke, I., Albrechtsen, a., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14(1):289.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Lowe, E. K., Swalla, B. J., and Brown, C. T. (2014). Evaluating a lightweight transcriptome assembly pipeline on two closely related Ascidian species. *PeerJ Preprints*, pages 1–11.

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, 5:1–13.

MacManes, M. D. and Lacey, E. A. (2012). The Social Brain: Transcriptome Assembly and Characterization of the Hippocampus from a Social Subterranean Rodent, the Colonial Tuco-Tuco (*Ctenomys sociabilis*). *PLOS ONE*, 7(9):e45524.

Marra, N. J., Romero, A., and DeWoody, J. A. (2014). Natural selection and the genetic basis of osmoregulation in heteromyid rodents as revealed by RNA-seq. *Molecular Ecology*, 23(11):2699–2711.

Mbandi, S. K., Hesse, U., Rees, D. J. G., and Christoffels, A. (2014). A glance at quality score: Implication for *de novo* transcriptome reconstruction of Illumina reads. *Frontiers in Genetics*, 5:1–17.

Mi, H. (2004). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, 33(1):D284–D288.

Mobasheri, A., Marples, D., Young, I. S., Floyd, R. V., Moskaluk, C. A., and Frigeri, A. (2007). Distribution of the AQP4 Water Channel in Normal Human Tissues: Protein and Tissue Microarrays Reveal Expression in Several New Anatomical Locations, including the Prostate Gland Seminal Vesicles. *Channels*, 1(1):30–39.

Nielsen, S., Chou, C., Marples, D., Christensen, E., Kishore, B., and Knepper, M. (1995). Vasopressin increases water permeability of kidney collecting duct by inducing translocation of Aquaporin-CD water channels to plasma-membrane. *Proceedings of The National Academy of Sciences of The United States of America*, 92(4):1013–1017.

Nielsen, S., Kwon, T. H., Christensen, B. M., Promeneur, D., Frøkiaer, J., and Marples, D. (1999). Physiology and pathophysiology of renal aquaporins. *Journal of the American Society of Nephrology*, 10(3):647–663.

Panhuis, T. M., Broitman-Maduro, G., Uhrig, J., Maduro, M., and Reznick, D. N. (2011). Analysis of Expressed Sequence Tags from the Placenta of the Live-Bearing Fish *Poeciliopsis* (Poeciliidae). *Journal of Heredity*, 102(3):352–361.

Pop, M. (2009). Genome assembly reborn: Recent computational challenges. *Briefings In Bioinformatics*, 10(4):354–366.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40:D290–301.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, 6(9):e22594.

Roberts, A. and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73.

Rojek, A., Rojek, A., Fuchtbauer, E., Fuchtbauer, E., Kwon, T., Kwon, T., Frokiaer, J., and Nielsen, S. (2006). Severe urinary concentrating defect in renal collecting duct-selective AQP2 conditional-knockout mice. *Proceedings of The National Academy of Sciences of The United States of America*, 103(15):6037–6042.

Romero, D. G., Plonczynski, M. W., Welsh, B. L., Gomez-Sanchez, C. E., Zhou, M. Y., and Gomez-Sanchez, E. P. (2007). Gene expression profile in rat adrenal zona glomerulosa cells stimulated with aldosterone secretagogues. *Physiological Genomics*, 32(1):117–127.

Shorter, K. R., Crossland, J. P., Webb, D., Szalai, G., Felder, M. R., and Vrana, P. B. (2012). *Peromyscus* as a Mammalian Epigenetic Model. *Genetics Research International*, 2012:1–11.

Shorter, K. R., Owen, A., Anderson, V., Hall-South, A. C., Hayford, S., Cakora, P., Crossland, J. P., Georgi, V. R. M., Perkins, A., Kelly, S. J., Felder, M. R., and Vrana, P. B. (2014). Natural genetic variation underlying differences in *Peromyscus* repetitive and social/aggressive behaviors. *Behavior genetics*, 44(2):126–135.

Sikes, R. S., Gannon, W. L., and Animal Care and Use Committee of the American Society of Mammalogists (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *Journal of Mammalogy*, 92(1):235–253.

Tatum, R., Zhang, Y., Salleng, K., Lu, Z., Lin, J. J., Lu, Q., Jeansonne, B. G., Ding, L., and Chen, Y. H. (2009). Renal salt wasting and chronic dehydration in claudin-7-deficient mice. *AJP: Renal Physiology*, 298(1):F24–F34.

Veal, R. and Caire, W. (2001). *Peromyscus eremicus*. *Mammalian Species*, 118:1–6.

Vijay, N., Poelstra, J. W., Künstner, A., Wolf, J. B. W., and Wolf, J. B. W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Molecular Ecology*, 22(3):620–634.

Walsberg, G. (2000). Small mammals in hot deserts: Some generalizations revisited. *Bioscience*, 50(2):109–120.

Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yang, Z. and dos Reis, M. (2011). Statistical Properties of the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*, 28(3):1217–1228.