

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

Matthew D. MacManes<sup>1</sup>, Michael B. Eisen<sup>2</sup>,

<sup>1</sup> Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham NH, USA

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

## 16 Introduction

17 Deserts are widely considered one of the harshest environments on Earth. Animals living in  
 18 desert environments are forced to endure intense heat and drought, and as a result, species liv-  
 19 ing in these environments are likely to possess specialized mechanisms to deal with them. While  
 20 living in deserts likely involves a large number of adaptive traits, the ability to osmoregulate –  
 21 to maintain the proper water and electrolyte balance – appears to be paramount [1]. Indeed,  
 22 the maintenance of water balance is one of the most important physiologic processes for all  
 23 organisms, whether they be desert inhabitants or not. Most animals are exquisitely sensitive  
 24 to changes in osmolality, with slight derangement eliciting physiologic compromise. When the  
 25 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 [2], rats [3–5], and humans [6–8]. 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* [9], *Psammomys obesus* [10], and *Perognathus penicillatus* [11]. More recently, full renal transcriptomes have been generated for *Dipodomys spectabilis* and *Chaetodipus baileyi*, [12] as well as *Abrothrix olivacea* [13].

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 [14], *Peromyscus eremicus*. These animals have a lifespan typical of small mammals, and therefore an individual may live its entire life without ever drinking water. These rodents have 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*) [15] with growing genetic and genomic resources [16–18]. 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 individuals 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 University of California Berkeley Animal Care and Use Committee and followed guidelines established by the American Society of Mammalogy for the use of wild animals in research [19].

### 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) 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* [20] using *kmer*=25, based on the developer's default recommendations. The error-corrected sequence reads were adapter and quality trimmed following recommendations from MacManes [21] and Mbandi [22]. Specifically, adapter sequence contamination and low

quality nucleotides (defined as PHRED  $<2$ ) were removed using the program Trimmomatic version 0.32 [23]. Reads from each tissue were assembled using the Trinity version released 17 July 2014 [24]. We used flags to indicate the stranded nature of sequencing reads and set the maximum allowable physical distance between read pairs to 999nt. 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/](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/](ftp://ftp.ncbi.nlm.nih.gov/genomes/Peromyscus_maniculatus_bairdii/RNA/)). We used a blastN procedure (default settings, evaluate 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 than a strategy employing expression-based filtering (remove if TMP  $<1$  [25]) 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 [26] using default setting, 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 [27] 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 [28]. Next, because rapidly evolving nucleotide sequences may evade detection by blast algorithms, we used HMMER3 [29] to search for conserved protein domains contained in the dataset using the Pfam database [30]. 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 [31]. We estimated expression for the four tissues individually using default settings of the software eXpress [32]. Interesting patterns of expression, including instances where expression was limited to a single tissue type, were identified and visualized.

## 126 Population Genomics

127 In addition to the reference individual sequenced at four different tissue types, we sequenced  
 128 15 other conspecific individuals from the same population in Palm Desert, California. Sequence  
 129 data were mapped to the reference transcriptome using bwa-mem. The alignments were sorted  
 130 and converted to BAM format, then passed to the program ANGSD version 0.610, which was  
 131 used for calculating the folded site frequency spectrum (SFS) and Tajima's D [33].

## 133 Natural Selection

134 To characterize natural selection on several genes related to water and ion homeostasis, we iden-  
 135 tified several of the transcripts identified as experiencing positive selection in a recent work on  
 136 desert-adaptor *Dipodomys* rodents. The coding sequences corresponding to these genes, Solute  
 137 Carrier family 2 member 9 (Slc2a9) and the Vitamin D3 receptor (Vdr), were extracted from  
 138 the dataset, aligned using the software MACSE [34] to homologous sequences in *Mus musculus*,  
 139 *Rattus norvegicus*, *Peromyscus maniculatus*, and *Homo sapiens* as identified by the conditional  
 140 reciprocal best blast procedure (CRBB, [35]). An unrooted gene tree was constructed using the  
 141 online resource Clustal-Omega, and the tree and alignment were analyzed using the branch-site  
 142 model (model=2, nsSites=2, fix\_omega=0 versus model=2, nsSites=2, fix\_omega=1, omega=1)  
 143 implemented in PAML version 4.8 [36,37]. Significance was evaluated via the use of the likeli-  
 144 hood ratio test.

## 146 Results and Discussion

### 147 RNA extraction, Sequencing, Assembly, Mapping

148 RNA was extracted from the hypothalamus, renal medulla, testes, and liver from each individual  
 149 using sterile technique. TRIzol extraction resulted in a large amount of high quality (RIN  $\geq$   
 150 8) total RNA, which was then used as input. Libraries were constructed as per the standard  
 151 Illumina protocol and sequenced as described above. The number of reads per library varied  
 152 from 56 million strand-specific paired-end reads in Peer360 kidney, to 9 million single-end reads  
 153 in Peer321 (Table 1, available on the Short Read Archive accession XXX). Adapter sequence  
 154 contamination and low-quality nucleotides were eliminated, which resulted in a loss of  $<2\%$  of  
 155 the total number of reads. These trimmed reads served as input for all downstream analyses.

### 156 Table 1

DATASET	NUM. RAW READS
PEER360 TESTES	32M PE/SS
PEER360 LIVER	53M PE /SS
PEER360 KIDNEY	56M PE/SS
PEER360 BRAIN	23M PE/SS
PEER305	19M PE
PEER308	15M PE
PEER319	14M PE
PEER321	9M SE
PEER340	16M PE
PEER352	14M PE
PEER354	9M SE
PEER359	14M PE
PEER365	16M PE
PEER366	16M PE
PEER368	14M PE
PEER369	14M PE
PEER372	17M SE
PEER373	23M SE
PEER380	16M SE
PEER382	14M SE

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

Transcriptome assembly for each tissue type was accomplished using the program Trinity [24]. The raw assembly for brain, liver, testes, and kidney contained 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-specific transcripts, 71041 liver-specific transcripts, 67340 testes-specific transcripts, and 113050 kidney-specific 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.

**Figure 1**

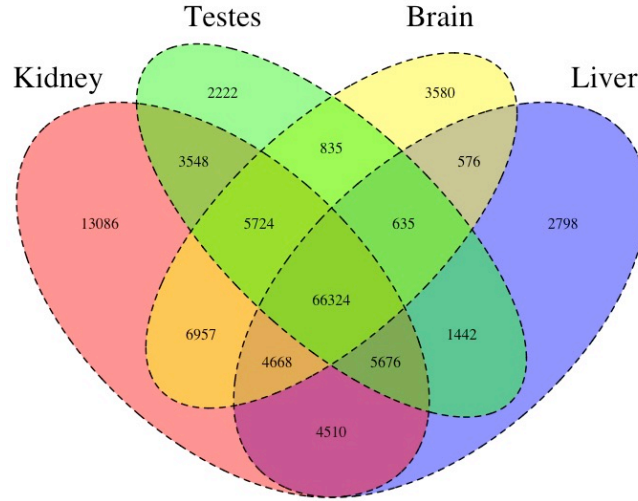


Figure 1. The Venn Diagram.

We then estimated gene expression on each of these tissue-specific datasets, which allowed us to understand expression patterns in the multiple tissues. 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 that may be the result of deeper sequencing.

In addition to this, we estimated mean TMP (number of transcripts per million) for all transcripts. Table 2 consists of the 10 genes whose mean TMP 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). From this filtered concatenated dataset, we extracted 71626 putative coding sequences (72Mb, to be made available on Dryad). Of these 71626 sequences, 38221 were complete exons (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.

**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

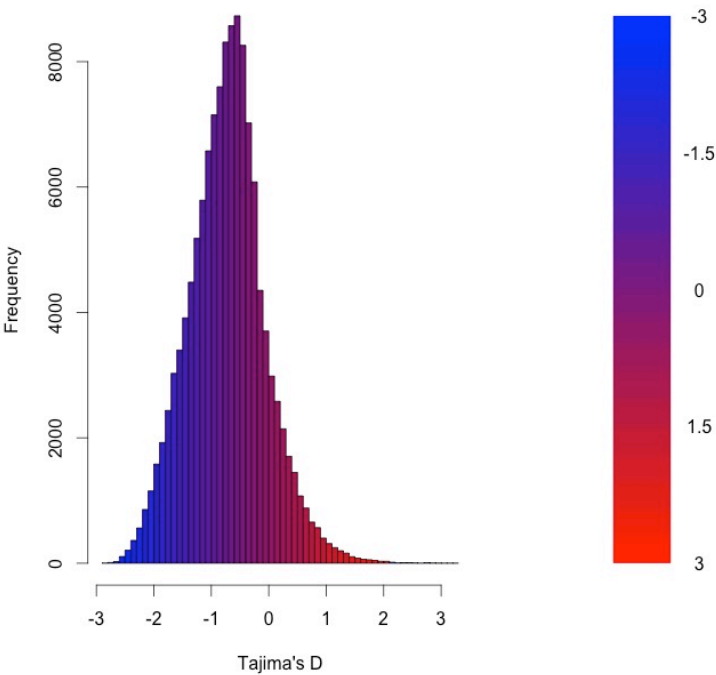
As detailed above, RNAseq data from 15 individuals were mapped to the reference transcriptome with the resulting BAM files being used as input to the software package ANGSD. The Tajima's D statistic was calculated for all transcripts covered by at least 14 of the 15 individuals. In brief, a negative 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. The distribution is skewed toward negative values (mean=-0.89, variance=0.58), which is likely the result of purifying selection, a model of evolution commonly invoked for coding DNA sequences [38]. 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



221 least. The former list of genes is likely to contain transcripts experiencing balancing selection  
222 in the studied population. This list includes, interestingly, genes obviously related to solute  
223 and water balance (e.g. *Clcnkb* and a transmembrane protein gene) and immune function (a  
224 interferon-inducible GTPase and a Class 1 MHC gene). The latter group, containing transcripts  
225 whose estimates of Tajima's D are the smallest are likely experiencing purifying selection. Many  
226 of these transcripts are involved in core regulatory functions where mutation may have strongly  
227 negative fitness consequences.  
228

229 **Figure 2**



230  
231 Figure 2. The distribution of Tajima's D for all putative transcripts.

232 **Table 3**

233

Transcript ID	GenBank ID	Description	Tajima's D
Transcript_49049	XM_006533884.1	heterogeneous nuclear ribonucleoprotein H1 (Hnrnp1)	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.

236 **Table 4**

237

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.

## 240 Natural Selection

241 To begin to test the hypothesis that selection on transcripts related to osmoregulation is en-  
 242 hanced in the desert adapted *P. eremicus*, we implemented the branch-site test as described  
 243 above by setting the sequence corresponding to *P. eremicus* for both Slc2a9 and Vdr as the  
 244 foreground lineages in 2 distinct program executions. These two transcripts were chosen specifi-  
 245 cally because they - the former significantly - were recently linked to osmoregulation in a desert

rodent [12]. The test for Slc2a9 was highly significant ( $2\Delta\text{Lnl}=51.4$ ,  $\text{df}=1$ ,  $p=0$ ), indicating enhanced selection in *P. eremicus* relative to the other lineages. The branch site test for positive selection conducted on the Vdr gene was non-significant ( $2\Delta\text{Lnl}=0.68$ ,  $\text{df}=1$ ,  $p=1$ ). This limited analysis of selection is to be followed up by an analysis of genome wide patterns of natural selection.

## 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

## References

1. Walsberg G (2000) Small mammals in hot deserts: Some generalizations revisited. *Bio-science* 50: 109–120.
2. Tatum R, Zhang Y, Salleng K, Lu Z, Lin JJ, et al. (2009) Renal salt wasting and chronic dehydration in claudin-7-deficient mice. *AJP: Renal Physiology* 298: F24–F34.
3. Romero DG, Plonczynski MW, Welsh BL, Gomez-Sanchez CE, Zhou MY, et al. (2007) Gene expression profile in rat adrenal zona glomerulosa cells stimulated with aldosterone secretagogues. *Physiological Genomics* 32: 117–127.
4. Rojek A, Rojek A, Fuchtbauer E, Fuchtbauer E, Kwon T, et al. (2006) Severe urinary concentrating defect in renal collecting duct-selective AQP2 conditional-knockout mice.

- 276 Proceedings of The National Academy of Sciences of The United States of America 103:  
277 6037–6042.
- 278 5. Nielsen S, Chou C, Marples D, Christensen E, Kishore B, et al. (1995) Vasopressin  
279 increases water permeability of kidney collecting duct by inducing translocation of  
280 aquaporin-CD water channels to plasma-membrane. Proceedings of The National  
281 Academy of Sciences of The United States of America 92: 1013–1017.
- 282 6. Mobasheri A, Marples D, Young IS, Floyd RV, Moskaluk CA, et al. (2007) Distribution  
283 of the AQP4 Water Channel in Normal Human Tissues: Protein and Tissue Microarrays  
284 Reveal Expression in Several New Anatomical Locations, including the Prostate Gland  
285 Seminal Vesicles. Channels 1: 30–39.
- 286 7. Bedford JJ, Leader JP, Walker RJ (2003) Aquaporin expression in normal human kidney  
287 and in renal disease. Journal of the American Society of Nephrology 14: 2581–2587.
- 288 8. Nielsen S, Kwon TH, Christensen BM, Promeneur D, Frøkiaer J, et al. (1999) Physiology  
289 and pathophysiology of renal aquaporins. Journal of the American Society of Nephrology  
290 10: 647–663.
- 291 9. Gallardo PA, Cortés A, Bozinovic F (2005) Phenotypic flexibility at the molecular and  
292 organismal level allows desert-dwelling rodents to cope with seasonal water availability.  
293 Physiological and Biochemical Zoology 78: 145–152.
- 294 10. Kaissling B, De Rouffignac C, Barrett JM, Kriz W (1975) The structural organization  
295 of the kidney of the desert rodent *Psammomys obesus*. Anatomy and Embryology 148:  
296 121–143.
- 297 11. Altschuler EM, Nagle RB, Braun EJ, Lindstedt SL, Krutzsch PH (1979) Morphological  
298 study of the desert heteromyid kidney with emphasis on the genus *Perognathus*. The  
299 Anatomical Record 194: 461–468.
- 300 12. Marra NJ, Romero a, DeWoody Ja (2014) Natural selection and the genetic basis of  
301 osmoregulation in heteromyid rodents as revealed by RNA-seq. Molecular Ecology 23:  
302 2699–2711.
- 303 13. Giorello FM, Feijoo M, D’Elía G, Valdez L, Opazo JC, et al. (2014) Characterization of  
304 the kidney transcriptome of the South American olive mouse *Abrothrix olivacea*. BMC  
305 Genomics 15: 446.
- 306 14. Veal R, Caire W (2001) *Peromyscus eremicus*. Mammalian Species 118: 1–6.

- 307 15. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, et al. (2007) Toward  
308 a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b  
309 Sequences. *Journal of Mammalogy* 88: 1146–1159.
- 310 16. Shorter KR, Owen A, anderson V, Hall-South AC, Hayford S, et al. (2014) Natural genetic  
311 variation underlying differences in *Peromyscus* repetitive and social/aggressive behaviors.  
312 *Behavior genetics* 44: 126–135.
- 313 17. Panhuis TM, Broitman-Maduro G, Uhrig J, Maduro M, Reznick DN (2011) Analysis of  
314 Expressed Sequence Tags from the Placenta of the Live-Bearing Fish *Poeciliopsis* (Poe-  
315 ciliidae). *Journal of Heredity* 102: 352–361.
- 316 18. Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, et al. (2012) *Peromyscus* as a  
317 Mammalian Epigenetic Model. *Genetics Research International* 2012: 1–11.
- 318 19. Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of  
319 Mammalogists (2011) Guidelines of the American Society of Mammalogists for the use of  
320 wild mammals in research. *Journal of Mammalogy* 92: 235–253.
- 321 20. Heo Y, Wu XL, Chen D, Ma J, Hwu WM (2014) BLESS: Bloom filter-based error correc-  
322 tion solution for high-throughput sequencing reads. *Bioinformatics* 30: 1354–1362.
- 323 21. MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence  
324 data. *Frontiers in Genetics* 5.
- 325 22. Mbandi SK, Hesse U, Rees DJG, Christoffels A (2014) A glance at quality score: Impli-  
326 cation for *de novo* transcriptome reconstruction of Illumina reads. *Frontiers in Genetics*  
327 5: 17.
- 328 23. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: A user-friendly,  
329 integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*  
330 40: W622–7.
- 331 24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo*  
332 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference  
333 generation and analysis. *Nature Protocols* 8: 1494–1512.
- 334 25. **MacManes** MD, Lacey EA (2012) The Social Brain: Transcriptome Assembly and Char-  
335 acterization of the Hippocampus from a Social Subterranean Rodent, the Colonial Tuco-  
336 Tuco (*Ctenomys sociabilis*). *PLOS ONE* 7: e45524.

- 337 26. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of  
338 protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
- 339 27. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+:  
340 architecture and applications. *BMC Bioinformatics* 10: 421.
- 341 28. Mi H (2004) The PANTHER database of protein families, subfamilies, functions and  
342 pathways. *Nucleic Acids Research* 33: D284–D288.
- 343 29. Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. *Bioin-*  
344 *formatics* 29: 2487–2489.
- 345 30. Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein  
346 families database. *Nucleic Acids Research* 40: D290–301.
- 347 31. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-  
348 MEM .
- 349 32. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of  
350 sequencing experiments. *Nature Methods* 10: 71–73.
- 351 33. Korneliussen I Thorfinn Sand Moltke, Albrechtsen a, Nielsen R (2013) Calculation of  
352 Tajima’s D and other neutrality test statistics from low depth next-generation sequencing  
353 data. *BMC Bioinformatics* 14: 289.
- 354 34. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of  
355 Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE* 6: e22594.
- 356 35. Aubry S, Kelly S, Kämpers BMC, Smith-Unna RD, Hibberd JM (2014) Deep Evolutionary  
357 Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two  
358 Independent Origins of C4 Photosynthesis. *PLOS Genetics* 10: e1004365.
- 359 36. Yang Z, dos Reis M (2011) Statistical Properties of the Branch-Site Test of Positive  
360 Selection. *Molecular Biology and Evolution* 28: 1217–1228.
- 361 37. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular*  
362 *Biology and Evolution* 24: 1586–1591.
- 363 38. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at  
364 synonymous sites in mammals. *Nature Reviews Genetics* 7: 98–108.