

# 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 Introduction

3 Deserts are widely considered one of Earth's harshest environments. Animals living in desert  
4 environments are forced to endure intense heat and drought, and in turn, species having evolved  
5 in these environments are likely to have evolved specialised mechanisms that may enhance  
6 fitness. While living in deserts likely involves a large number of adaptive phenotypes, the abil-  
7 ity to osmoregulate – to maintain the proper water and electrolyte balance – appears to be  
8 paramount [1]. Indeed, the maintenance of water balance in animals is one of the most impor-  
9 tant physiologic processes, and is critical to desert survival. Mammals are exquisitely sensitive  
10 to changes in osmolality, with slight derangement eliciting physiologic compromise. When the  
11 loss of water exceeds dietary intake, dehydration - and in extreme cases, death - can occur. Un-  
12 derstanding these mechanisms will significantly enhance our understanding of the physiologic  
13 processes underlying osmoregulation in extreme environments, having implications for studies  
14 of human health, conservation, and climate change.

15

16 The genes and structures responsible for the maintenance of water and electrolyte balance  
17 are well characterized in model organisms such as mice [2], rats [3–5], and humans [6–8]. These  
18 studies, many of which have been enabled by newer sequencing technologies, serve as a founda-  
19 tion for studies of renal genomics in non-model organisms. In particular, because researchers  
20 have long been interested in desert adaptation, a number of studies have looked at the mor-  
21 phology or expression of single genes in the renal tissues of desert adapted rodents *Phyllotis*  
22 *darwini* [9], *Psammomys obesus* [10], and *Perognathus penicillatus* [11]. More recently, full re-  
23 nal transcriptomes have been generated for *Dipodomys spectabilis* and *Chaetodipus baileyi* [12]  
24 as well as *Abrothrix olivacea* [13].

25

26 These studies provide a rich context for the current and future work, aimed at developing

a synthetic understanding of the 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, 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 it's 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] which together suggest 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 here to lay the groundwork by characterizing the transcriptome from four distinct tissues (brain, liver, kidney, testes). These data will be included in 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 bases of adaptation in *P. eremicus*.

## Materials and Methods

### Animal Collection and Study Design

To begin to understand how genes may underlie desert adaptation, I collected 15 individuals from a single population *P. eremicus* over a two year time period (2012-2013). These individuals were captured in live traps, then euthanized using isoflurane overdose and decapitation. Immediately post-mortem, the abdominal and pelvic organs were removed, cut in half (in the case of kidney), placed in RNAlater and flash frozen in liquid Nitrogen. Removal of the brain, with similar preservation techniques, followed that. 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 follow 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 manufacturers 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 using the rapid-mode of the HiSeq 2500 sequencer at The Hubbard Center for Genome Sciences (University of New Hampshire). The remaining 14 libraries were similarly multiplexed and sequenced in a mixture of 100nt paired and single end across two lanes of an Illumina HiSeq 2000 at the Vincent G. Coates Gnome Center (University of California, Berkeley).

## Sequence Data Preprocessing and Assembly

The raw sequence reads were error corrected using the software `bleed` [20], using `kmer=25`, based on the developers recommendations. The error corrected adapter and quality trimmed following recommendations from MacManes [21] and Mbandi [22]. Specifically, adapter sequence contamination was removed, 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 `Trinity` version released 17 July 2014 [24]. We used flags indicating the stranded nature of sequencing reads and set 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, I 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/`). I used a relaxed `blastN` procedure (evalue set to  $10^{-10}$ ) to identify contigs in the *P. eremicus* dataset that are likely biological in origin. I then concatenated the filtered assemblies from each tissue into a single file, reducing redundancy using the software `cd-hit-est` [25] using default setting except that sequences were clustered based on 95% sequence similarity. The resulting assembly file was characterized using the software package `transrate` (`https://github.com/Blahah/transrate`).

89

## Assembled Sequence Annotation

From the filtered assembly, I extracted putative coding sequences using `Transdecoder` version 4ul2014 (`http://transdecoder.sourceforge.net/`). These putative protein coding sequences were annotated using default settings of the `blastx` algorithm [26] against the SwissProt database

93

downloaded on 1 March 2014. Because transcriptome assemblies typically contain non-coding elements (e.g. ncRNA) in addition to protein coding sequence, we annotated the entire filtered dataset using the NCBI nt dataset, downloaded on 1 March 2014. Lastly, because rapidly evolving nucleotide sequences may evade detection by blast algorithms, we used HMMER3 [27] to search for conserved protein domains contained in the Pfam database [28].

To identify sequences unique to each tissue type, I mapped sequence reads from each tissue type to the reference assembly using bwa-mem. We estimated expression individually for the four tissues. Interesting patterns of expression, including instances where expression was limited to a single tissue type were identified.

## Population Genomics

In addition to the reference individual sequenced at four different tissue types, we sequenced 15 other conspecific individuals from the same population, located 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, which was used for calculating the folded site frequency spectrum (SFS) and Tajima's D [29].

## 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 *Dipodomys* rodents. The coding sequence corresponding to these genes, Solute Carrier family 2 member 9 (Slc2a9) and the Vitamin D3 receptor (Vdr), were extracted from the dataset, aligned using the software MACSE [30] to homologous sequences identified by the conditional reciprocal best blast procedure (CRBB, [31]) implemented in transrate. These alignments were inputted into the software codeABC version 1.6.0 [32].

## Results

### RNA extraction, Sequencing, Assembly, Mapping

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

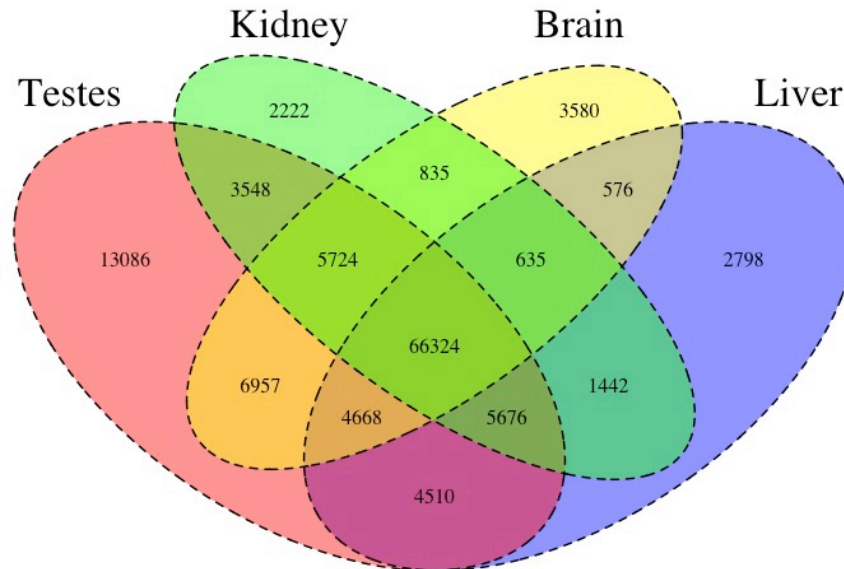
125 protocol, and ere sequenced as described above. The number of reads per library varied from  
 126 56 million strand-specific paired-end reads in Peer360 kidney, to 9 million single-end reads in  
 127 Peer321. Adapter sequence contamination and low-quality nucleotides were eliminated, which  
 128 resulted in a loss of <2% of reads.

129

130 Transcriptome assembly for each tissue type was accomplished using the program Trinity.  
 131 The raw assembly for brain, liver, testes and kidney contained 185425, 222096, 180233, and  
 132 514091 assembled sequences respectively. This assembly was filtered using a blast procedure  
 133 with resulted in a final dataset containing 68331 brain-specific transcripts, 71041 liver-specific  
 134 transcripts, 67340 testes-specific transcripts, and 113050 kidney-specific transcripts. Mapping  
 135 the error-corrected adapter/quality trimmed reads to these datasets resulted in mapping 94.98%  
 136 (87.01% properly paired) of brain-derived reads to the brain transcriptome, 96.07% (88.13%  
 137 properly paired) of liver-derived reads to the liver transcriptome, 96.81% (85.10% properly  
 138 paired) of testes-derived reads to the testes transcriptome, and 91.87% (83.77% properly paired)  
 139 of kidney-derived reads to the kidney transcriptome. Together, these statistics suggest that the  
 140 tissue-specific transcriptomes are of extremely high quality.

141

142 I then estimated gene expression on each of these tissue-specific datasets, which allowed me  
 143 to understand expression patterns in the multiple tissues Figure 1.



144

Figure 1. The Venn Diagram.

After expression estimation, the filtered assemblies were concatenated together, and after removal of redundancy with cd-hit-est, 123,123 putative transcripts remained. From this filtered concatenated dataset, I extracted 64355 putative coding sequences (60Mb). Of these 64355 sequences, 37960 were complete exons (containing both start and stop codons), while other were either truncated at the 5-prime end (16880 sequences), 3-prime end (4203 sequences), or were internal (5312 sequencing having neither stop nor start codon).

## Subsection 2

## Discussion

## 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. *Proceedings of The National Academy of Sciences of The United States of America* 103: 6037–6042.
5. Nielsen S, CHOU C, MARPLES D, CHRISTENSEN E, KISHORE B, et al. (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: 1013–1017.
6. Mobasheri A, Marples D, Young IS, Floyd RV, Moskaluk CA, et al. (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: 30–39.

- 174 7. Bedford JJ, Bedford JJ, Leader JP, Leader JP, Walker RJ, et al. (2003) Aquaporin ex-  
 175 pression in normal human kidney and in renal disease. Journal of the American Society  
 176 of Nephrology : JASN 14: 2581–2587.
- 177 8. Nielsen S, KWON T (1999) Physiology and Pathophysiology of Renal Aquaporins. Journal  
 178 of the . . . .
- 179 9. Gallardo PA, Cortés A, Bozinovic F (2005) Phenotypic flexibility at the molecular and  
 180 organismal level allows desert-dwelling rodents to cope with seasonal water availability.  
 181 Physiological and Biochemical Zoology 78: 145–152.
- 182 10. Kaissling B, De Rouffignac C, Barrett JM, Kriz W (1975) The structural organization  
 183 of the kidney of the desert rodent *Psammomys obesus*. Anatomy and embryology 148:  
 184 121–143.
- 185 11. Altschuler EM, Altschuler EM, Nagle RB, Nagle RB, Braun EJ, et al. (1979) Morpholog-  
 186 ical study of the desert heteromyid kidney with emphasis on the genus *Perognathus*. The  
 187 Anatomical record 194: 461–468.
- 188 12. Marra NJ, Romero a, DeWoody Ja (2014) Natural selection and the genetic basis of  
 189 osmoregulation in heteromyid rodents as revealed by RNA-seq. Molecular Ecology 23:  
 190 2699–2711.
- 191 13. Giorello FM, Feijoo M, a GD, Valdez L, Opazo JC, et al. (2014) Characterization of the  
 192 kidney transcriptome of the South American olive mouse *Abrothrix olivacea* 15: 1–10.
- 193 14. Veal R, Caire W (2001) *Peromyscus eremicus*. Mammalian Species 118: 1–6.
- 194 15. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, et al. (2007) Toward  
 195 a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b  
 196 Sequences. Journal of Mammalogy 88: 1146–1159.
- 197 16. Shorter KR, Owen A, anderson V, Hall-South AC, Hayford S, et al. (2014) Natural Ge-  
 198 netic Variation Underlying Differences in *Peromyscus* Repetitive and Social/Aggressive  
 199 Behaviors. Behavior genetics .
- 200 17. Panhuis TM, Panhuis TM, Broitman-Maduro G, Broitman-Maduro G, Uhrig J, et al.  
 201 (2011) Analysis of Expressed Sequence Tags from the Placenta of the Live-Bearing Fish  
 202 *Poeciliopsis* (Poeciliidae). Journal of Heredity 102: 352–361.
- 203 18. Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, et al. (2012) *Peromyscus* as a  
 204 Mammalian Epigenetic Model. Genetics Research International 2012: 1–11.

- 205 19. Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of  
206 Mammalogists (2011) Guidelines of the American Society of Mammalogists for the use of  
207 wild mammals in research. *Journal of Mammalogy* 92: 235–253.
- 208 20. Heo Y, Wu XL, Chen D, Ma J, Hwu WM (2014) BLESS: bloom filter-based error cor-  
209 rection solution for high-throughput sequencing reads. *Bioinformatics* (Oxford, England)  
210 30: 1354–1362.
- 211 21. MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence  
212 data. *Frontiers in Genetics* 5.
- 213 22. Christoffels A (2014) A glance at quality score: implication for *de novo* transcriptome  
214 reconstruction of Illumina reads : 1–5.
- 215 23. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly,  
216 integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*  
217 40: W622–7.
- 218 24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo*  
219 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference  
220 generation and analysis. *Nature protocols* 8: 1494–1512.
- 221 25. Li W, Li W, Godzik A, Godzik A (2006) Cd-hit: a fast program for clustering and  
222 comparing large sets of protein or nucleotide sequences. *Bioinformatics* (Oxford, England)  
223 22: 1658–1659.
- 224 26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+:  
225 architecture and applications. *BMC Bioinformatics* 10: 421.
- 226 27. Wheeler TJ, Wheeler TJ, Eddy SR, Eddy SR (2013) nhmmer: DNA homology search  
227 with profile HMMs. *Bioinformatics* (Oxford, England) 29: 2487–2489.
- 228 28. Punta M, Cogill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein  
229 families database. *Nucleic Acids Research* 40: D290–301.
- 230 29. Korneliussen TS, Moltke I, Albrechtsen A (2013) Calculation of Tajima’s D and other  
231 neutrality test statistics from low depth next-generation sequencing data. *BMC . . . .*
- 232 30. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of  
233 Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE* 6: e22594.



234 31. Aubry S, Kelly S, Kämpers BMC, Smith-Unna RD, Hibberd JM (2014) Deep Evolutionary  
235 Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two  
236 Independent Origins of C4 Photosynthesis. PLOS Genetics 10: e1004365.

237 32. Lopes JS, Arenas M, Posada D, Beaumont MA (2013) Coestimation of recombination,  
238 substitution and molecular adaptation rates by approximate Bayesian computation 112:  
239 255–264.

240 **Figure Legends**

241 **Tables**

242 **Table 1**

243

244

DATASET	NUM. RAW READS
PEER360 TESTES	32M PE
PEER360 LIVER	53M PE
PEER360 KIDNEY	56M PE
PEER360 BRAIN	23M PE
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

245      Table 1. The number of sequencing reads per sample