

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 Introduction

3 Deserts are widely considered one of Earth's harshest environments. Animals living in desert en-
 4 vironments are forced to endure intense heat and drought, and in turn, species having evolved in
 5 these environments are likely to have evolved specialised mechanisms that may enhance fitness.
 6 While living in deserts likely involves a large number of adaptive phenotypes, the ability to os-
 7 moregulate – to maintain the proper water and electrolyte balance – appears to be paramount [1].
 8 Indeed, the maintenance of water balance in animals is one of the most important physiologic
 9 processes for all animals, desert inhabitants or not. Most animals are exquisitely sensitive to
 10 changes in osmolality, with slight derangement eliciting physiologic compromise. When the loss
 11 of water exceeds dietary intake, dehydration - and in extreme cases, death - can occur, which
 12 suggests that there is strong selection for mechanisms supporting osmoregulation. Understand-
 13 ing these mechanisms will significantly enhance our understanding of the physiologic processes
 14 underlying osmoregulation in extreme environments, having implications for studies of human
 15 health, conservation, and climate change.

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

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 16 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>).

Assembled Sequence Annotation

The filtered assemblies were annotated using default settings of the blastN algorithm [26] against the Ensembl cDNA and ncRNA datasets described above, downloaded on 1 August 2014. Amongst other things, the Ensembl transcript identifiers were used in the analysis of gene ontology, conducted in the PANTHER package [?]. Next, because rapidly evolving nucleotide sequences may evade detection by blast algorithms, we used HMMER3 [27] to search for conserved protein domains contained in the dataset using the Pfam database [28]. Lastly, I extracted putative coding sequences using Transdecoder version 4Jul2014 (<http://transdecoder.sourceforge.net/>)

To identify patterns of gene expression 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 using default settings of the software eXpress [?]. 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, 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 protocol, and were 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. Adapter sequence contamination and low-quality nucleotides were eliminated, which resulted in a loss of $<2\%$ of reads.

Transcriptome assembly for each tissue type was accomplished using the program Trinity. 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 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 brain-derived reads to the brain transcriptome, 96.07% (88.13% properly paired) of liver-derived reads to the liver transcriptome, 96.81% (85.10% properly paired) of testes-derived reads to the testes transcriptome, and 91.87% (83.77% properly paired) of kidney-derived reads to the kidney transcriptome. Together, these statistics suggest that the tissue-specific transcriptomes are of extremely high quality.

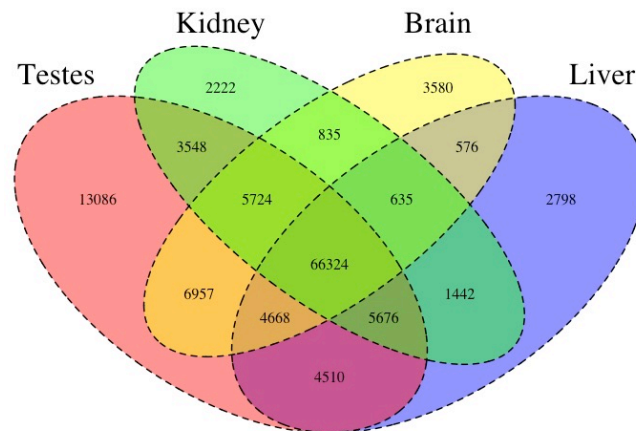


Figure 1. The Venn Diagram.

I then estimated gene expression on each of these tissue-specific datasets, which allowed me to understand expression patterns in the multiple tissues (Figure 1). 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 71626 putative coding sequences (72Mb). Of these 71626 sequences, 38221 were complete exons (containing both start and stop codons), while other were either truncated at the 5-prime end (20239 sequences), 3-prime end (6445 sequences), or were internal (6721 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

- 174 Reveal Expression in Several New Anatomical Locations, including the Prostate Gland
175 Seminal Vesicles. *Channels* 1: 30–39.
- 176 7. Bedford JJ, Bedford JJ, Leader JP, Leader JP, Walker RJ, et al. (2003) Aquaporin ex-
177 pression in normal human kidney and in renal disease. *Journal of the American Society*
178 *of Nephrology* : *JASN* 14: 2581–2587.
- 179 8. Nielsen S, KWON T (1999) Physiology and Pathophysiology of Renal Aquaporins. *Journal*
180 *of the*
- 181 9. Gallardo PA, Cortés A, Bozinovic F (2005) Phenotypic flexibility at the molecular and
182 organismal level allows desert-dwelling rodents to cope with seasonal water availability.
183 *Physiological and Biochemical Zoology* 78: 145–152.
- 184 10. Kaissling B, De Rouffignac C, Barrett JM, Kriz W (1975) The structural organization
185 of the kidney of the desert rodent *Psammomys obesus*. *Anatomy and embryology* 148:
186 121–143.
- 187 11. Altschuler EM, Altschuler EM, Nagle RB, Nagle RB, Braun EJ, et al. (1979) Morpholog-
188 ical study of the desert heteromyid kidney with emphasis on the genus *Perognathus*. *The*
189 *Anatomical record* 194: 461–468.
- 190 12. Marra NJ, Romero a, DeWoody Ja (2014) Natural selection and the genetic basis of
191 osmoregulation in heteromyid rodents as revealed by RNA-seq. *Molecular Ecology* 23:
192 2699–2711.
- 193 13. Giorello FM, Feijoo M, a GD, Valdez L, Opazo JC, et al. (2014) Characterization of the
194 kidney transcriptome of the South American olive mouse *Abrothrix olivacea* 15: 1–10.
- 195 14. Veal R, Caire W (2001) *Peromyscus eremicus*. *Mammalian Species* 118: 1–6.
- 196 15. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, et al. (2007) Toward
197 a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b
198 Sequences. *Journal of Mammalogy* 88: 1146–1159.
- 199 16. Shorter KR, Owen A, anderson V, Hall-South AC, Hayford S, et al. (2014) Natural Ge-
200 netic Variation Underlying Differences in *Peromyscus* Repetitive and Social/Aggressive
201 Behaviors. *Behavior genetics* .
- 202 17. Panhuis TM, Panhuis TM, Broitman-Maduro G, Broitman-Maduro G, Uhrig J, et al.
203 (2011) Analysis of Expressed Sequence Tags from the Placenta of the Live-Bearing Fish
204 *Poeciliopsis* (Poeciliidae). *Journal of Heredity* 102: 352–361.

- 205 18. Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, et al. (2012) *Peromyscus* as a
206 Mammalian Epigenetic Model. *Genetics Research International* 2012: 1–11.
- 207 19. Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of
208 Mammalogists (2011) Guidelines of the American Society of Mammalogists for the use of
209 wild mammals in research. *Journal of Mammalogy* 92: 235–253.
- 210 20. Heo Y, Wu XL, Chen D, Ma J, Hwu WM (2014) BLESS: bloom filter-based error cor-
211 rection solution for high-throughput sequencing reads. *Bioinformatics* (Oxford, England)
212 30: 1354–1362.
- 213 21. MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence
214 data. *Frontiers in Genetics* 5.
- 215 22. Christoffels A (2014) A glance at quality score: implication for *de novo* transcriptome
216 reconstruction of Illumina reads : 1–5.
- 217 23. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly,
218 integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*
219 40: W622–7.
- 220 24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo*
221 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
222 generation and analysis. *Nature protocols* 8: 1494–1512.
- 223 25. Li W, Li W, Godzik A, Godzik A (2006) Cd-hit: a fast program for clustering and
224 comparing large sets of protein or nucleotide sequences. *Bioinformatics* (Oxford, England)
225 22: 1658–1659.
- 226 26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+:
227 architecture and applications. *BMC Bioinformatics* 10: 421.
- 228 27. Wheeler TJ, Wheeler TJ, Eddy SR, Eddy SR (2013) nhmmer: DNA homology search
229 with profile HMMs. *Bioinformatics* (Oxford, England) 29: 2487–2489.
- 230 28. Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein
231 families database. *Nucleic Acids Research* 40: D290–301.
- 232 29. Korneliussen TS, Moltke I, Albrechtsen A (2013) Calculation of Tajima’s D and other
233 neutrality test statistics from low depth next-generation sequencing data. *BMC ...*
- 234 30. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of
235 Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE* 6: e22594.

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

239 32. Lopes JS, Arenas M, Posada D, Beaumont MA (2013) Coestimation of recombination,
240 substitution and molecular adaptation rates by approximate Bayesian computation 112:
241 255–264.

242 **Figure Legends**

243 **Tables**

244 **Table 1**

245

246

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

247 Table 1. The number of sequencing reads per sample