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

Introduction

For biologists interested in understanding the relationship between fitness, genotype, and phenotype, modern sequencing technologies provide for an unprecedented opportunity to gain a deep understanding of genome level processes that together, underlie adaptation. Unlike more traditional approaches (e.g. QTL mapping), second generation sequencing allow researchers to access genome wide patterns of gene expression, nucleotide variation and ulitimately pattens of natural selection in non-model organisms lacking an extensive set of genomic tools. This approach has been used widely in recent years in characterizing X, Y and Z.

One interesting example of adaptation lies in animal's ability to survive desert conditions. Here, heat *and* drought provide for powerful selective forces, testing animals' ability to osmoregulate and thus to survive [1]. Indeed, the maintenance of water balance in animals is one of the most important physiologic processes, and is critical to desert survival. Mammals 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. Unlike most mammals, animals living in desert habitats are subjected to long periods of extreme heat and intense drought. As a result, desert animals have evolved mechanisms through which physiologic homeostasis is maintained despite severe and prolonged dehydration. Understanding these mechanismsm will significantly enhance our understanding of the physiologic processes underlying osmoregulation in extreme environments, having implications for studies of human health, conservation, and climate change.

Genes responsible for the maintinance of water balance are well characterized in model organisms such as mice [?], rats [?, ?, ?], and humans [?, ?, ?]. In addition to these studies, a ong standing interest in deserts adaptation has resulted in a number of studies that looked at the morphology or expression of single genes in the renal tissues of desert adapted rodents *Phyllotis darwini* [?], *Psammomys obesus* [?], and *Perognathus penicillatus* [?]. More recently, full renal transcriptomes have been generated for *Dipodomys spectabilis* and *Chaetodipus baileyi* [?] as well as *Abrothrix olivacea* [?].

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 the desert adapted a cricetid rodent endemic to the Southwest United States [?]. 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*) [2] with growing genetic and genomic resources [3–5] 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 [6].

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

Following recommendations from MacManes [7] and Mbandi [8], adapter sequence contamination was removed, and low quality nucleotides (defined as PHRED <2) were removed from the dataset using the program Trimmomatic version 0.32 [9]. We concatenated sequence data from each reference tissue type and assembled them jointly using the Trinity beta version released 16 March 2014 [10]. 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 the XSEDE computer resource Blacklight. To filter the raw sequence assembly, I estimated TPM for each assembled sequence using bwa-mem version 0.75 [11] and eXpress version 1.51 [12], removing all contigs whose expression was less than TPM=1 [10].

Assembled Sequence Annotation

From the filtered assembly, I extracted putative coding sequences using Transdecoder version 16Jan2014 (`http://transdecoder.sourceforge.net/`). These putative protein coding sequences were annotated using default settings of the blastx algorithm [13] against the SwissProt database 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 [14] to search for conserved protein domains contained in the Pfam database [15].

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 [16].

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 ≥ 8) total RNA, which was used as input. Libraries were constructed as per the standard Illumina protocol, and ere 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 was accomplished using the program Trinity. The raw assembly contained 743314 assembled sequenced measuring 418Mb. This assembly was filtered using TMP >1 as a threshold. The filtered assembly contained 130764 sequences measuring 149Mb. from this filteres 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. Bioscience 50: 109–120.

2. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, et al. (2007) Toward a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b Sequences. Journal of Mammalogy 88: 1146–1159.

3. Shorter KR, Owen A, anderson V, Hall-South AC, Hayford S, et al. (2014) Natural Genetic Variation Underlying Differences in *Peromyscus* Repetitive and Social/Aggressive Behaviors. Behavior genetics .

4. Panhuis TM, Panhuis TM, Broitman-Maduro G, Broitman-Maduro G, Uhrig J, et al. (2011) Analysis of Expressed Sequence Tags from the Placenta of the Live-Bearing Fish Poeciliopsis (Poeciliidae). Journal of Heredity 102: 352–361.

5. Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, et al. (2012) *Peromyscus* as a Mammalian Epigenetic Model. Genetics Research International 2012: 1–11.

6. Sikes RS, Gannon WL, 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: 235–253.

7. MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics 5.

8. Christoffels A (2014) A glance at quality score: implication for *de novo* transcriptome reconstruction of Illumina reads : 1–5.

9. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research 40: W622–7.

10. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8: 1494–1512.

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

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

13. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

14. Wheeler TJ, Wheeler TJ, Eddy SR, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics (Oxford, England) 29: 2487–2489.

15. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Research 40: D290–301.

16. Korneliussen TS, Moltke I, Albrechtsen A (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC . . . .

Figure Legends

Tables

Table 1

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

Table 1. The number of sequencing reads per sample