Characterization of the transcriptome, nucleotide sequence polymorphism and 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 Institution Name, 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. One interesting example of adaptation lies in animals ability to survive desert conditions. Here, heat *and* drought provide for powerful selective forces, testing animals' ability to osmoregulate and thus to survive.

Specifically, the maintenance of water balance in animals is one of the most important physiologic processes, and is critical to desert survival. Indeed, 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.

One such desert-adapted rodent, a cricetid rodent endemic to the Southwest United states is a novel model for the study of adaptation to desert environments. They 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 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*) (**?**). There are growing genetic and genomic resources available (**???**).

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

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 (**?**) and Mbandi (**?**), 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 (**?**). We concatenated sequence data from each reference tissue type and assembled them using the Trinity beta version released 16 March 2014 (**?**). 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 0.77 (**?**) and eXpress version 1.51 (**?**), removing all contigs whose expression was less than TPM=1 (**?**).

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 (**?**) 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 (**?**) to search for conserved protein domains contained in the Pfam database (**?**).

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 (**?**).

Results

Subsection 1

Subsection 2

Discussion

Acknowledgments

Figure Legends

Tables