

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 most harsh environments. Animals living in desert
 4 environments are forced to endure intense heat and drought, and as a result, species having
 5 evolved in these environments are likely to possess specialized mechanisms that may enhance
 6 fitness. While living in deserts likely involves a large number of adaptive traits, 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 pro-
 9 cesses for all organisms, whether they be desert inhabitants or not. Most animals are exquisitely
 10 sensitive to changes in osmolality, with slight derangement eliciting physiologic compromise.
 11 When the loss of water exceeds dietary intake, dehydration - and in extreme cases, death - can
 12 occur. This process suggests that there is strong selection for mechanisms supporting osmoreg-
 13 ulation. Understanding these mechanisms will significantly enhance our understanding of the
 14 physiologic processes underlying osmoregulation in extreme environments, having implications
 15 for studies of human 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 to deserts 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 15 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 default 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 blastN procedure (default settings, *evalue* set to 10^{-10}) to identify contigs in the *P. eremicus* dataset that are likely 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) of the raw assembly. 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.

91 Assembled Sequence Annotation

92 The filtered assemblies were annotated using default settings of the blastN algorithm [26]
 93 against the Ensembl cDNA and ncRNA datasets described above, downloaded on 1 August
 94 2014. Amongst other things, the Ensemble transcript identifiers were used in the analysis
 95 of gene ontology, conducted in the PANTHER package [27]. Next, because rapidly evolv-
 96 ing nucleotide sequences may evade detection by blast algorithms, we used HMMER3 [28] to
 97 search for conserved protein domains contained in the dataset using the Pfam database [29].
 98 Lastly, I extracted putative coding sequences using Transdecoder version 4Jul2014 (<http://transdecoder.sourceforge.net/>)
 99

100
 101 To identify patterns of gene expression unique to each tissue type, I mapped sequence reads
 102 from each tissue type to the reference assembly using bwa-mem [30]. We estimated expression
 103 individually for the four tissues using default settings of the software eXpress [31]. Interesting
 104 patterns of expression, including instances where expression was limited to a single tissue type
 105 were identified and visualized.
 106

107 Population Genomics

108 In addition to the reference individual sequenced at four different tissue types, we sequenced
 109 15 other conspecific individuals from the same population, located in Palm Desert, California.
 110 Sequence data were mapped to the reference transcriptome using bwa-mem. The alignments
 111 were sorted and converted to BAM format, then passed to the program ANGSD version 0.610,
 112 which was used for calculating the folded site frequency spectrum (SFS) and Tajima's D [32].
 113

114 Natural Selection

115 To characterize natural selection on several genes related to water and ion homeostasis, we
 116 identified several of the transcripts identified as experiencing positive selection in a recent work
 117 on desert-adapted *Dipodomys* rodents. The coding sequence corresponding to these genes, Solute
 118 Carrier family 2 member 9 (Slc2a9) and the Vitamin D3 receptor (Vdr), were extracted from
 119 the dataset, aligned using the software MACSE [33] to homologous sequences in *Mus musculus*,
 120 *Rattus norvegicus*, *Peromyscus maniculatus*, and *Homo sapiens* identified by the conditional
 121 reciprocal best blast procedure (CRBB, [34]). An unrooted gene tree was constructed using the
 122 online resource Clustal-Omega, and together the tree and alignment were analyzed using the
 123 branch-site model (model=2, nsSites=2, fix_omega=0 versus model=2, nsSites=2, fix_omega=1,

124 omega=1) implemented in PAML version 4.8 [35,36].

125 **Results**

126 **RNA extraction, Sequencing, Assembly, Mapping**

127 RNA was extracted from the hypothalamus, renal medulla, testes, or liver from each individual
128 using sterile technique. TRIzol extraction resulted in a large amount of high quality (RIN \geq
129 8) total RNA, which was used as input. Libraries were constructed as per the standard Illu-
130 mina protocol, and ere sequenced as described above. The number of reads per library varied
131 from 56 million strand-specific paired-end reads in Peer360 kidney, to 9 million single-end reads
132 in Peer321 (Table 1, available on the Short Read Archive accession XXX). Adapter sequence
133 contamination and low-quality nucleotides were eliminated, which resulted in a loss of <2% of
134 reads.

135 **Table 1**

136

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

137

Table 1. The number of sequencing reads per sample

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 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. All tissue-specific assemblies to be made available on Dryad.

Figure 1

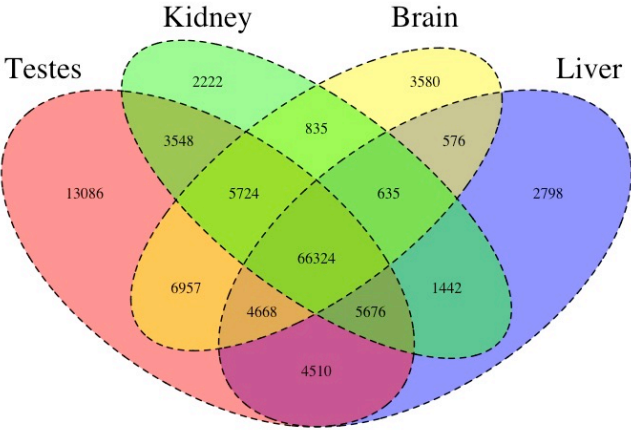


Figure 1. The Venn Diagram.

I then estimated gene expression on each of these tissue-specific datasets, which allowed me to understand expression patterns in the multiple tissues. Specifically, I constructed a Venn diagram (Figure 1), which allowed me to visualize the proportion genes whose expression was limited to a single tissue, and those where expression was ubiquitous. In addition to this, I estimated mean TMP (transcripts per million) for all transcripts. Table 2 consists of the 10 genes whose mean TMP was the highest.

After expression estimation, the filtered assemblies were concatenated together, and after removal of redundancy with cd-hit-est, 123,123 putative transcripts remained (To be available on Genbank). From this filtered concatenated dataset, I extracted 71626 putative coding sequences (72Mb, to be available on Dryad). 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). The results of a Pfam search conducted on the predicted amino acid sequences will be found on Dryad.

170
171
172

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).

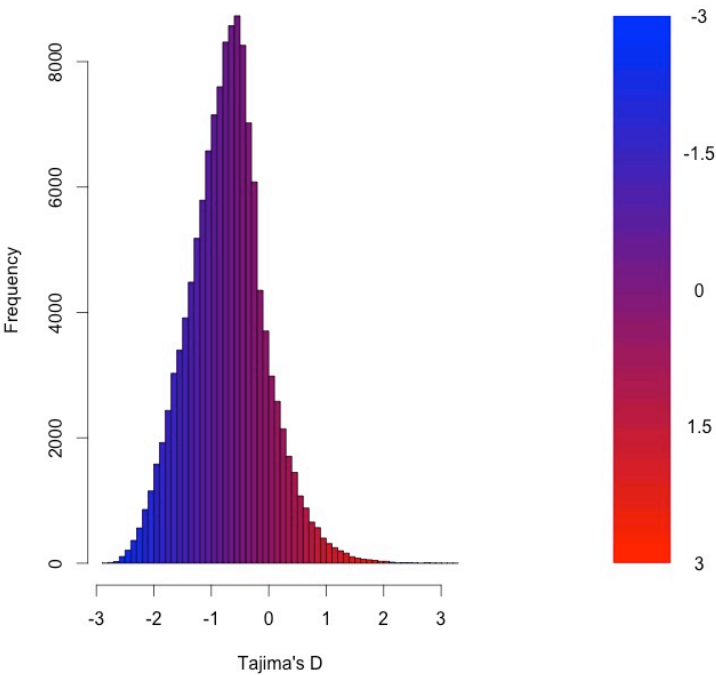
Population Genomics

As detailed above, the 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. The distribution of the results, shown in Figure 2, suggest that the vast majority of the transcriptome is under purifying selection ($D < 0$), with a much smaller fraction being subject to neutral or positive selection.

182

Figure 2

183



184

185 Figure 2. The distribution of Tajima’s D for all putative transcripts.

186 Table 3

187

Transcript ID	GenBank ID	Description	Tajima’s D
Transcript_49049	XM.006533884.1	heterogeneous nuclear ribonucleoprotein H1 (Hnrnph1)	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 is suggestive of positive selection.

190 **Table 4**

191

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 (HIV-1) 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 is suggestive of purifying selection.

194 **Natural Selection**

195 To test the hypothesis that selection on transcripts related to osmoregulation is enhanced in the
196 desert adapted *P. eremicus*, I implemented the branch-site test as described above, setting the
197 sequence corresponding to *P. eremicus* for both Slc2a9 and Vdr as the foreground lineages in
198 2 distinct program executions. The test for Slc2a9 was highly significant ($2\Delta\text{Ln}l=51.4$, $\text{df}=1$,
199 $p=0$), indicating enhanced selection in *P. eremicus* relative to the other lineages. The branch
200 site test for positive selection conducted on the Vdr gene was non-significant ($2\Delta\text{Ln}l=0.68$, $\text{df}=1$,
201 $p=1$).

202 **Discussion**

203 **Acknowledgments**

204 **References**

205 1. Walsberg G (2000) Small mammals in hot deserts: Some generalizations revisited. Bio-
206 science 50: 109–120.

207 2. Tatum R, Zhang Y, Salleng K, Lu Z, Lin JJ, et al. (2009) Renal salt wasting and chronic
208 dehydration in claudin-7-deficient mice. AJP: Renal Physiology 298: F24–F34.

- 209 3. Romero DG, Plonczynski MW, Welsh BL, Gomez-Sanchez CE, Zhou MY, et al. (2007)
 210 Gene expression profile in rat adrenal zona glomerulosa cells stimulated with aldosterone
 211 secretagogues. *Physiological Genomics* 32: 117–127.
- 212 4. Rojek A, Rojek A, Fuchtbauer E, Fuchtbauer E, Kwon T, et al. (2006) Severe urinary
 213 concentrating defect in renal collecting duct-selective AQP2 conditional-knockout mice.
 214 *Proceedings of The National Academy of Sciences of The United States of America* 103:
 215 6037–6042.
- 216 5. Nielsen S, Chou C, Marples D, Christensen E, Kishore B, et al. (1995) Vasopressin
 217 Increases Water Permeability of Kidney Collecting Duct by Inducing Translocation of
 218 Aquaporin-Cd Water Channels to Plasma-Membrane. *Proceedings of The National*
 219 *Academy of Sciences of The United States of America* 92: 1013–1017.
- 220 6. Mobasher A, Marples D, Young IS, Floyd RV, Moskaluk CA, et al. (2007) Distribution
 221 of the AQP4 Water Channel in Normal Human Tissues: Protein and Tissue Microarrays
 222 Reveal Expression in Several New Anatomical Locations, including the Prostate Gland
 223 Seminal Vesicles. *Channels* 1: 30–39.
- 224 7. Bedford JJ, Leader JP, Walker RJ (2003) Aquaporin expression in normal human kidney
 225 and in renal disease. *Journal of the American Society of Nephrology : JASN* 14: 2581–
 226 2587.
- 227 8. Nielsen S, Kwon T (1999) Physiology and Pathophysiology of Renal Aquaporins. *Journal*
 228 *of the ...*
- 229 9. Gallardo PA, Cortés A, Bozinovic F (2005) Phenotypic flexibility at the molecular and
 230 organismal level allows desert-dwelling rodents to cope with seasonal water availability.
 231 *Physiological and Biochemical Zoology* 78: 145–152.
- 232 10. Kaissling B, De Rouffignac C, Barrett JM, Kriz W (1975) The structural organization
 233 of the kidney of the desert rodent *Psammomys obesus*. *Anatomy and embryology* 148:
 234 121–143.
- 235 11. Altschuler EM, Nagle RB, Braun EJ, Lindstedt SL, Krutzsch PH (1979) Morphological
 236 study of the desert heteromyid kidney with emphasis on the genus *Perognathus*. *The*
 237 *Anatomical record* 194: 461–468.
- 238 12. Marra NJ, Romero a, DeWoody Ja (2014) Natural selection and the genetic basis of
 239 osmoregulation in heteromyid rodents as revealed by RNA-seq. *Molecular Ecology* 23:
 240 2699–2711.

- 241 13. Giorello FM, Feijoo M, a GD, Valdez L, Opazo JC, et al. (2014) Characterization of the
242 kidney transcriptome of the South American olive mouse *Abrothrix olivacea* 15: 1–10.
- 243 14. Veal R, Caire W (2001) *Peromyscus eremicus*. Mammalian Species 118: 1–6.
- 244 15. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, et al. (2007) Toward
245 a Molecular Phylogeny for *Peromyscus*: Evidence from Mitochondrial Cytochrome- b
246 Sequences. Journal of Mammalogy 88: 1146–1159.
- 247 16. Shorter KR, Owen A, anderson V, Hall-South AC, Hayford S, et al. (2014) Natural Ge-
248 netic Variation Underlying Differences in *Peromyscus* Repetitive and Social/Aggressive
249 Behaviors. Behavior genetics .
- 250 17. Panhuis TM, Broitman-Maduro G, Uhrig J, Maduro M, Reznick DN (2011) Analysis of
251 Expressed Sequence Tags from the Placenta of the Live-Bearing Fish Poeciliopsis (Poe-
252 ciliidae). Journal of Heredity 102: 352–361.
- 253 18. Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, et al. (2012) *Peromyscus* as a
254 Mammalian Epigenetic Model. Genetics Research International 2012: 1–11.
- 255 19. Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of
256 Mammalogists (2011) Guidelines of the American Society of Mammalogists for the use of
257 wild mammals in research. Journal of Mammalogy 92: 235–253.
- 258 20. Heo Y, Wu XL, Chen D, Ma J, Hwu WM (2014) BLESS: bloom filter-based error cor-
259 rection solution for high-throughput sequencing reads. Bioinformatics (Oxford, England)
260 30: 1354–1362.
- 261 21. MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence
262 data. Frontiers in Genetics 5.
- 263 22. Christoffels A (2014) A glance at quality score: implication for *de novo* transcriptome
264 reconstruction of Illumina reads : 1–5.
- 265 23. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly,
266 integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research
267 40: W622–7.
- 268 24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo*
269 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
270 generation and analysis. Nature protocols 8: 1494–1512.

- 271 25. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of
272 protein or nucleotide sequences. *Bioinformatics* (Oxford, England) 22: 1658–1659.
- 273 26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+:
274 architecture and applications. *BMC Bioinformatics* 10: 421.
- 275 27. Mi H (2004) The PANTHER database of protein families, subfamilies, functions and
276 pathways. *Nucleic Acids Research* 33: D284–D288.
- 277 28. Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. *Bioin-*
278 *formatics* (Oxford, England) 29: 2487–2489.
- 279 29. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein
280 families database. *Nucleic Acids Research* 40: D290–301.
- 281 30. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-
282 MEM .
- 283 31. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of
284 sequencing experiments. *Nature Methods* 10: 71–73.
- 285 32. Korneliussen TS, Moltke I, Albrechtsen A (2013) Calculation of Tajima’s D and other
286 neutrality test statistics from low depth next-generation sequencing data. *BMC*
- 287 33. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of
288 Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE* 6: e22594.
- 289 34. Aubry S, Kelly S, Kumpers BMC, Smith-Unna RD, Hibberd JM (2014) Deep Evolutionary
290 Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two
291 Independent Origins of C4 Photosynthesis. *PLOS Genetics* 10: e1004365.
- 292 35. Yang Z, dos Reis M (2011) Statistical Properties of the Branch-Site Test of Positive
293 Selection. *Molecular Biology and Evolution* 28: 1217–1228.
- 294 36. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular*
295 *Biology and Evolution* 24: 1586–1591.

296 **Figure Legends**

297 **Tables**