

1 Characterization of a Male Reproductive Transcriptome for *Peromyscus eremicus* (Cactus
2 mouse)

3

4 Lauren Kordonowy^{1*} and Matthew MacManes¹⁺

5 1. University of New Hampshire Department of Molecular Cellular and Biomedical Sciences
6 Durham, NH, USA

7 * lauren.kordonowy@unh.edu

8 + matthew.macmanes@unh.edu

9

10 **Corresponding Author:**

11 Lauren Kordonowy

12 University of New Hampshire

13 Rudman Hall (MCBS)

14 46 College Road

15 Durham, NH, 03824

16 lauren.kordonowy@unh.edu

17

18

19

20

21

22

23

24

25

26

27 **Abstract:**

28 Rodents of the genus *Peromyscus* have become increasingly utilized models for investigations
29 into adaptive biology. This genus is particularly powerful for research linking genetics with
30 adaptive physiology and behaviors, and recent research has capitalized on the unique
31 opportunities afforded by the ecological diversity of these rodents. However, well characterized
32 genomic and transcriptomic data is intrinsic to explorations of the genetic architecture
33 responsible for ecological adaptations. This study characterizes a reproductive transcriptome of
34 male *Peromyscus eremicus* (Cactus mouse), a desert specialist with extreme physiological
35 adaptations to water limitation. We describe a reproductive transcriptome comprising three
36 tissues in order to expand upon existing research in this species and to facilitate further studies
37 elucidating the genetic basis of potential desert adaptations in male reproductive physiology.

38

39 **Introduction:**

40 The rapid infusion of novel bioinformatics approaches in the fields of genomics and
41 transcriptomics has enabled the coalescence of the fields of genetics, physiology and ecology
42 into innovative studies for adaptation in evolutionary biology. Indeed, studies on the biology of
43 adaptation had previously been dominated by research painstakingly documenting morphological
44 shifts associated with ecological gradients. However, the discipline of bioinformatics has
45 breathed new life into the field of adaptation biology. Specifically, while the morphological
46 basis as well as the physiological mechanisms of adaptation have been explored for a variety of
47 species in extreme environments, the genetic underpinnings of these adaptations has only
48 recently become a larger area of research (Cheviron & Brumfield, 2011). High-throughput
49 sequencing technology of model and non-model organisms (Ellegren, 2014) enables
50 evolutionary biologists to conduct genome and transcriptome wide analyses and link patterns of
51 gene selection with functional adaptations.

52 Studies on the genetic basis of adaptation have included a wide variety of taxa. For
53 example, butterflies in the *Heliconius* genus have been a particularly effective study systems for
54 determining the genetic basis of pigmentation patterns, and there is evidence of interspecific

55 introgression for genes enabling adaptive mimicry patterns (Hines et al., 2012; The Heliconius
56 Genome Consortium, 2012). In addition, a population genomic study in three-spine sticklebacks
57 has elucidated many loci responsible for divergent adaptations from marine to freshwater
58 environments (Jones et al., 2012). Another active area of adaptation genetics research focuses on
59 species residing in extreme environments. High altitude adaptations to hemoglobin variants have
60 been identified in multiple organisms, including humans (Lorenzo et al., 2014), several species
61 of Andean ducks (McCracken et al., 2009a; 2009b), and deer mice, *Peromyscus maniculatus*
62 (Storz et al., 2010; Natarajan et al., 2015). The genetic pathways responsible for physiological
63 adaptations to desert habitats remain enigmatic; however, considerable progress has been made
64 developing candidate gene sets for future analyses (e.g. Guillen et al., 2015; MacManes & Eisen,
65 2104; Marra et al., 2012; Marra, Romero, & deWoody, 2014). Functional studies will stem from
66 this foundational research aimed at identifying the genomic underpinnings of adaptations to
67 extreme environments; yet, it is inherently challenging and critically important to demonstrate
68 that specific loci are functionally responsible for adaptations (Storz & Wheat 2010).

69 Rodents of the genus *Peromyscus* have been at the forefront of research elucidating the
70 genetic basis for adaptation (reviewed in Bedford & Hoekstra, 2015). This diverse genus has
71 served as an ideal platform for adaptation research spanning from the genetic basis of behavioral
72 adaptations – such as complex burrowing in *Peromyscus polionotus* (Weber & Hoekstra; 2009;
73 Weber, Peterson & Hoekstra, 2013) – to the loci responsible for adaptive morphology – such as
74 coat coloration in *Peromyscus polionotus leucocephalus* (Hoekstra et al., 2006) and including
75 kidney desert adaptations in *Peromyscus eremicus* (MacManes & Eisen, 2014). We are currently
76 using *Peromyscus eremicus* as a model species for investigating the genetic bases of desert
77 adaptations, and this paper will describe efforts to meet this research aim.

78 Initial steps toward understanding the genetics of adaptation must include the genomic
79 and transcriptomic characterization of target study species (MacManes & Eisen, 2014). Toward
80 this end, we assembled and characterized a composite transcriptome for three male reproductive
81 tissues in the desert specialist, *P. eremicus*. This species is an exceptional example of desert
82 adaptation, as individuals may live exclusively without water access (Veal & Clare, 2001).
83 MacManes and Eisen (2014) assembled transcriptomes from kidney, hypothalamus, lung, and
84 testes of this species, and they identified several candidate genes potentially underlying adaptive

85 renal physiology. However, to our knowledge, potential physiological reproductive adaptations
86 to water limitation have not been studied in this species or in other desert rodents. In order to
87 pursue this novel line of adaptation research, we developed a transcriptome comprising multiple
88 reproductive tissues in this species. Specifically, we assembled a composite reproductive
89 transcriptome for three male reproductive tissues – the epididymis, testes, and vas deferens.
90 Here, we describe and compare tissue specific transcriptomic data in the context of transcript
91 abundance and relevant database searches.

92

93 **Methods:**

94 *Tissue Samples, RNA extraction, cDNA Library Preparation and Sequencing*

95 A single reproductively mature *P. eremicus* male was sacrificed via isoflurane overdose
96 and decapitation. This was done in accordance with University of New Hampshire Animal Care
97 and Use Committee guidelines (protocol number 130902) and guidelines established by the
98 American Society of Mammalogists (Sikes et al., 2011). Testes, epididymis, and vas deferens
99 were immediately harvested (within ten minutes of euthanasia), placed in RNAlater (Ambion
100 Life Technologies) and stored at -80 degree Celsius until RNA extraction. We used a standard
101 TRIzol, chloroform protocol for total RNA extraction (Ambion Life Technologies). We
102 evaluated the quantity and quality of the RNA product with a Qubit 2.0 Fluorometer (Invitrogen)
103 and a Tapestation 2200 (Agilent Technologies, Palo Alto, USA).

104 We used a TURBO DNase kit (Ambion) to eliminate DNA from the samples prior to the
105 library preparation. Libraries were made with a TruSeq Stranded mRNA Sample Prep LS Kit
106 (Illumina). Each of the three samples was labeled with a unique hexamer adapter for
107 sequencing for identification in multiplex single lane sequencing. Following library completion,
108 we confirmed the quality and quantity of the DNA product with the Qubit and Tapestation. We
109 submitted the multiplexed sample of the libraries for running on a single lane at the New York
110 Genome Center Sequencing Facility (NY, New York). Paired end sequencing reads of length
111 125bp were generated on an Illumina 2500 platform. Reads were parsed by tissue type
112 according to their unique hexamer IDs in preparation for transcriptome assembly.

113 *Reproductive Transcriptome assembly*

114 The composite reproductive transcriptome was assembled with reads from the testes,
115 epididymis and vas deferens using the previously developed Oyster River Protocol for *de novo*
116 transcriptome assembly pipeline (MacManes, 2016). Briefly, the reads were error corrected with
117 Rcorrector v1.0.1 (Song & Florea, 2015). We used the *de novo* transcriptome assembler Trinity
118 v2.1.1 (Haas et al., 2013; Grabherr et al., 2011). Within the Trinity platform, we ran
119 Trimmomatic (Bolger, Lohse and Usadel, 2014) to remove the adapters, and we also trimmed at
120 PHRED < 2, as recommended by MacManes (2014).

121 Next we evaluated transcriptome assembly quality and completeness using BUSCO
122 v1.1b1 and Transrate v1.0.1. BUSCO (Simão et al., 2015) reports the number of complete,
123 fragmented, and missing orthologs in assembled genomes, transcriptomes, or gene sets relative
124 to compiled ortholog databases. We ran BUSCO on the assembly using the ortholog database
125 for vertebrates, which includes 3,023 genes. The assembly was also analyzed by Transrate using
126 the *Mus musculus* peptide database from Ensembl (downloaded 2/24/16) as a reference. The
127 Transrate score provided a metric of *de novo* transcriptome assembly quality, and the software
128 also generated an improved assembly comprised of highly supported contigs (Smith-Unna et al.,
129 2015). Finally, we re-ran BUSCO on the improved assembly generated by Transrate to
130 determine if this assembly had similar metric scores for completeness as the original assembly
131 produced by Trinity. As alternatives to the original Trinity assembly and the optimized
132 Transrate assembly, we proceeded with our optimization determinations by filtering out low
133 abundance contigs from the original Trinity assembly. First we calculated the relative abundance
134 of the transcripts with Kallisto v0.42.4 and Salmon v0.5.1. Kallisto utilizes a pseudo-alignment
135 algorithm to map RNA-seq data reads to targets for transcript abundance quantification (Bray et
136 al., 2015). In contrast, Salmon employs a lightweight quasi-alignment method and a high speed
137 streaming algorithm to quantify transcripts (Patro, Duggal & Kingsford, 2015). After
138 determining transcript abundance in both Kallisto and Salmon, we removed contigs with
139 transcripts per million (TPM) estimates of less than 0.5 and of less than 1.0 in two separate
140 optimization trials (as per MacManes, 2016). Finally, we evaluated these two filtered assemblies
141 with Transrate and BUSCO to determine the relative quality and completeness of both
142 assemblies. We chose the optimal assembly version by comparing Transrate and BUSCO metrics
143 and also through careful consideration of total contig numbers across all filtering and optimizing

144 versions. The chosen assembly was the Transrate optimized TPM > 0.5 filtered assembly, and
145 this assembly was used for all subsequent analyses.

146 *Annotation, Transcript Abundance, and Database Searches*

147 We used dammit v0.2.7.1 (Scott 2016) to annotate the optimized transcriptome assembly
148 (as per MacManes, 2016). Within the dammit platform, we predicted protein coding regions for
149 each tissue with TransDecoder v2.0.1 (Haas et al., 2013), which was used to find open reading
150 frames (ORFs). Furthermore, dammit utilizes multiple database searches for annotating
151 transcriptomes. These database searches include searches in Rfam v12.0 to find non-coding
152 RNAs (Nawrocki et al., 2014), searches for protein domains in Pfam-A v29.0 (Sonnhammer,
153 Eddy & Durbin, 1997; Finn et al., 2016), the execution of a LAST search for known proteins in
154 the UniRef90 database (Suzek et al., 2007; Suzek et al., 2015), ortholog matches in the BUSCO
155 databases, and orthology searches in OrthoDB (Kriventseva et al., 2015).

156 Next we used the assembly annotated by dammit to re-run Kallisto to determine
157 transcript abundance within each of the three tissue types. Highly abundant transcripts were
158 found by sorting and selecting the transcripts with the 10 highest TPM counts for each tissue.
159 Ensembl accession numbers generated by dammit were searched within the web browser
160 (ensembl.org) to determine the protein and gene matches corresponding to these transcripts. In
161 addition, we used TPM counts of expression for all three tissues to generate counts of transcripts
162 specific to and shared across tissue types.

163 We also downloaded the ncRNA database for *Mus musculus* from Ensembl (v 2/25/16),
164 and we did a BLASTn (Altschul et al., 1990; Madden, 2002) search for these ncRNAs in our
165 assembly. This database has 16,274 sequences, and we determined the number of transcript ID
166 matches and the number of unique ncRNA sequence matches for our assembly. We also counted
167 how many transcript matches were present in each of the tissues, and we referenced the
168 corresponding Kallisto derived TPM values to determine the number of unique and ubiquitous
169 transcript matches for each tissue.

170 We searched the annotated assembly for transporter protein matches within the
171 Transporter Classification Database (tcdb.org). This database has 13,846 sequences representing

172 proteins in transmembrane molecular transport systems (Saier et al., 2014). We executed a
173 BLASTx (Altschul et al., 1990; Madden, 2002) search to find the number of transcript ID
174 matches and the number of unique transporter protein matches within the assembly. Next we
175 determined how many transcript ID matches were found in each of the three tissues. As
176 previously described above, we also cross-referenced these matches with the Kallisto derived
177 TPM values to find the number of ubiquitous and unique transcript matches by tissue type.

178 Of note, the code for performing all of the above analyses can be found at
179 https://github.com/macmanes-lab/peer_reproductive-transcriptome/blob/master/code.md. The
180 data files that are in Dropbox, which will later be submitted to Dryad, can be found at
181 <https://www.dropbox.com/home/PEReproductiveTranscriptome/To%20Submit%20to%20Dryad>.

182

183 **Results and Discussion:**

184 *Reproductive Transcriptome assembly*

185 There were 45-94 million paired reads produced for each of the three transcriptome
186 datasets, yielding a total of 415,960,428 reads. The raw reads are available at the European
187 Nucleotide Archive under study accession number PRJEB13364.

188 We assembled a *de novo* composite reproductive transcriptome with reads from testes,
189 epididymis and vas deferens. The evaluation of alternative optimized assemblies allowed us to
190 generate a substantially complete transcriptome of high quality. The alternative assemblies had
191 raw Transrate scores ranging from 0.194-0.208 (**Table 1**). However, the scores for the improved
192 assemblies generated by Transrate, consisting of only highly supported contigs, ranged between
193 0.295-0.349, which is well above the threshold Transrate score of 0.22 for an acceptable
194 assembly. The BUSCO results indicated that the assemblies were highly complete, with
195 complete matches ranging from 73-90% of vertebrate orthologs (**Table 2**). These BUSCO
196 benchmark values are consistent with the most complete reported assessments for transcriptomes
197 from other vertebrate taxa (busco.ezlab.org). Furthermore, our BUSCO values exceed that of the
198 only available reported male reproductive tissue (from a coelacanth: *Latimeria menadoensis*
199 testes), which was 71% complete (Simão et al., 2015). The assembly version which was of

200 highest quality in relation to the Transrate metrics was the Transrate optimized Trinity assembly;
201 specifically, the optimized Transrate score was 0.3495, and the percent coverage of the reference
202 assembly was also highest, with 45% of the mouse database represented. This assembly was
203 highly competitive for completeness, as indicated by the BUSCO metric of 85% orthologs found.
204 However, this assembly had an exorbitantly high number of contigs (657,952 contigs), which is
205 nearly an order of magnitude more contigs than the next best performing assembly: the Transrate
206 optimized TPM > 0.5 filtered assembly (78,424 contigs). In consideration of the dramatically
207 more realistic contig number for the Transrate optimized TPM > 0.5 filtered assembly, and in
208 light of its second best performance for Transrate score (0.3013), reasonable Transrate mouse
209 reference assembly coverage (37%), and sufficiently high BUSCO completeness (73% orthologs
210 found), we chose this assembly as our optimized transcriptome. Therefore, we proceeded with
211 this optimized assembly version as our finalized transcriptome assembly for our analyses, which
212 is available in Dropbox (to be posted on Dryad after this manuscript's acceptance).

213 *Annotation, Transcript Abundance and Database Searches*

214 The reproductive transcriptome assembly annotations were produced by dammit, and
215 they are available through Dropbox (this file, and all other data files will be posted on Dryad
216 after acceptance) in a gff3 file format. Furthermore, TransDecoder was used to predict coding
217 regions in the assembly. TransDecoder predicted that 49.5% (38,342) of the transcripts (78,424
218 total) contained ORFs, of which 63.9% (24,808) had complete ORFs containing a start and stop
219 codon. The predicted protein coding regions generated by TransDecoder are reported in five file
220 types, and they are available on Dropbox. Furthermore, the Pfam results yielded 30.7% of
221 transcripts (24,107) matching to the protein family database. In contrast, the LAST search found
222 that 75.9% of transcripts (59,503) matched to the UniRef90 database. We have uploaded the
223 homology search results generated by Pfam and UniRef90 matches onto Dropbox. In addition,
224 1.04% (816) of transcripts matched to the Rfam database for ncRNAs, and these results are
225 posted in Dropbox. Of note, 80.1% (62,835) of the transcripts were annotated using one of more
226 of the above described methods (the dammit.gff3 file is posted in Dropbox), and it is this final
227 annotated assembly that was used for all subsequent analyses.

228 The transcripts with the 10 highest TPM counts generated by Kallisto for each tissue type
229 corresponded with protein matches in Ensembl (**Tables 3-5**). All three tissues had highly
230 abundant transcripts for mitochondrially encoded cytochrome c oxidase subunits, which are
231 involved in cellular respiration. Highly abundant testes proteins included protamine 2 – which is
232 involved in spermatogenesis – and sperm autoantigenic protein 17 – a zona pellucida binding
233 protein. Similarly, the highly abundant epididymis transcripts consisted of a protein involved in
234 spermatozoa maturation, cysteine-rich secretory protein 1, as well as Cd52 (also known as
235 epididymal secretory protein E5). In contrast, the vas deferens had abundant transcripts for
236 proteins involved in muscle contraction, specifically several actin subunits. The highly abundant
237 transcripts in all three tissues were consistent with our expectations of their physiology and in
238 keeping with findings in humans. Specifically, protamine-2 was the most highly expressed gene
239 in the human testes, and zona pellucida binding proteins were also highly expressed in human
240 testes (Djureinovic et al., 2014).

241 The Kallisto generated TPM counts of expression (available on Dropbox) were also
242 utilized to determine which transcripts were ubiquitous and specific to the three tissue types,
243 which we have depicted in a Venn diagram format (**Figure 1**). The assembly consisted of
244 78,424 different transcript IDs, of which 64,553 were shared across all three tissues. The
245 number of unique transcripts were as follows: 3,563 in testes, 342 in epididymis, and 502 in vas
246 deferens. The relatively large number of unique transcripts in the testes is consistent with
247 previous findings which describe the testes as the tissue with the highest number of tissue-
248 enriched genes in the human body (Uhlen et al., 2015).

249 In addition, we searched for *Mus musculus* ncRNA sequence matches within our
250 assembly. There were 15,964 transcript matches, which correspond to 2,320 unique ncRNA
251 matches, and they are posted on Dropbox. The transcript matches by tissue type were found
252 using the Kallisto TPM determinations, and they were as follows, testes: 15,260, epididymis:
253 15,552, and vas deferens: 15,558. A Venn Diagram depicts unique and shared transcript matches
254 by tissue type (**Figure 2**). The majority of transcript matches were ubiquitous to all three tissues
255 (14,724), and there were far fewer tissue specific matches. The testes had more unique transcript
256 matches (185) than the epididymis (26) or the vas deferens (45). These findings are consistent

257 with our results above regarding the relative numbers of total unique transcripts in the assembly
258 by tissue type.

259 Our search for transporter protein matches within the Transporter Classification Database
260 yielded 7,521 different transcript matches, corresponding to 1,373 unique transporter protein
261 matches, and they are posted on Dropbox. The number of transcript matches was highly similar
262 between the tissue types (testes 7,025; epididymis 7,115; vas deferens: 7,071). We generated a
263 Venn Diagram to display the numbers of shared and unique transcript matches to the transporter
264 protein sequences (**Figure 3**). Most transcript matches were present in all three tissues (6,472),
265 and there were relatively few unique matches in the three tissue types. However, the testes had
266 the highest number of unique transcript matches (215) relative to the epididymis (19) and the vas
267 deferens (37). These results are in keeping with those reported for the ncRNA sequence matches
268 and the complete assembly dataset. Furthermore, our BLASTx search of this transporter protein
269 database yielded transcript matches for multiple solute carrier proteins. We are particularly
270 interested in solute carrier proteins because previous research has found candidate genes in this
271 protein family for desert adaptations in kidneys of the kangaroo rat (Marra et al. 2012; Marra,
272 Romero & deWoody 2014) and the Cactus mouse (MacManes and Eisen 2014). In addition, we
273 had multiple matches to aquaporins, which are water channels allowing transport across cellular
274 membranes. One transcript matched specifically to Aquaporin 3, a sperm water channel found in
275 mice and humans, which is essential to maintaining sperm cellular integrity in response to the
276 hypotonic environment within the female reproductive tract (Chen et al., 2011).

277

278 **Conclusions:**

279 This study describes a composite transcriptome from three male reproductive tissues in
280 the desert specialist *Peromyscus eremicus*. Our analyses include quality and completeness
281 assessments of this reproductive assembly, which was generated using reads from testes,
282 epididymis and vas deferens of a male Cactus mouse. We also describe transcript expression
283 levels, generate annotations, and search relevant databases for ncRNAs and transporter protein
284 sequences. Finally, we describe the degree of ubiquity between transcripts among the three
285 tissues as well as identify transcripts unique to those tissues. Our future research will investigate

286 potential male reproductive physiology adaptations to water limitation in Cactus mouse, and the
287 characterization of this reproductive transcriptome will form the foundation of studies along this
288 vein. Moreover, this research contributes transcriptomic materials to a larger body of work in
289 the expanding field of adaptation genetics, which benefits tremendously from enhanced
290 opportunities for comparative analyses.

291

292 **Tables:**

293 **Table 1: Transrate results for the reproductive transcriptome assembly produced by**
 294 **different optimization methods. * This is the score of the Transrate optimized assembly in**

295 **Table 2**

296

Assembly	Transrate Score	Optimized Score*	# Read Pairs (fragments)	Contigs (n_seqs)	# Good Contigs	%Good Contigs
Trinity Original	0.1944	0.3492	207,980,214	856,711	657,952	0.77
Filter TPM<0.5	0.1672	0.3013	207,980,214	147,966	78,424	0.53
Filter TPM<1.0	0.156	0.2854	207,980,214	80,165	54,140	0.68

297

298 **Table 2: BUSCO metrics for the reproductive transcriptome assembly produced by**
 299 **different optimization methods.**

Assembly	% Complete	%Duplicated	%Fragmented	%Missing
Trinity original	90	49	3.4	5.5
Transrate Optimized	85	44	4.3	9.7
Filter TPM<0.5	85	38	3.0	11
Transrate TPM<0.5	73	31	3.9	22
Filter TPM<1.0	80	28	2.8	16
Transrate TPM<1.0	74	25	3.4	21

300

301 **Table 3: Testes transcript annotations for top 10 TPM results generated by Kallisto.**

302

Transcript ID	Ensembl Accession Number	TPM	Protein	Gene
Transcript_72932	ENSRNOP00000014866	23136.8	Sperm autoantigenic protein 17	Spa17
Transcript_56114	ENSMUSP00000047925	19154.7	Protamine 2	Prm2
Transcript_37923	ENSMODP00000000682	18016.4	Uncharacterized Protein	N/A
Transcript_197	ENSLACP00000009213	4841.98	ubiquitin A-52 residue ribosomal protein fusion product 1	Uba52
Transcript_56089	ENSMUSP00000080993	4236.73	mitochondrially encoded cytochrome c oxidase I	Mt-co1
Transcript_21368	ENSRNOP00000065673	4186.17	N/A	N/A
Transcript_70782	N/A	4085.53	N/A	N/A
Transcript_67718	ENSMUSP00000080997	3840.31	mitochondrially encoded cytochrome c oxidase III	Mt-co3
Transcript_51810	N/A	3013.74	N/A	N/A
Transcript_67719	ENSRNOP00000046414	2950.84	mitochondrially encoded cytochrome c oxidase II	Mt-co2

303

304 **Table 4: Epididymis transcript annotations for top 10 TPM results generated by Kallisto.**

305

Transcript ID	Ensembl Accession Number	TPM	Protein	Gene
Transcript_48473	N/A	99593	N/A	N/A
Transcript_48471	ENSMUSP00000026498	59217.6	cysteine-rich secretory protein 1	Crisp1
Transcript_43156	N/A	26929.8	N/A	N/A
Transcript_56089	ENSMUSP00000080993	24654.4	mitochondrially encoded cytochrome c oxidase I	Mt-co1
Transcript_67718	ENSMUSP00000080997	19450.7	mitochondrially encoded cytochrome c oxidase III	Mt-co3
Transcript_67719	ENSRNOP00000046414	16564.4	mitochondrially encoded cytochrome c oxidase II	Mt-co2
Transcript_65120	ENSRNOP00000020688	14990.1	CD52 molecule	Cd52
Transcript_48472	ENSMUSP00000026498	12385.4	cysteine-rich secretory protein 1	Crisp1
Transcript_49280	ENSMUSP00000026498	10974.7	cysteine-rich secretory protein 1	Crisp1
Transcript_73235	ENSSTOP00000007662	9807.83	Fatty acid-binding protein, adipocyte	Fabp4

306

307 **Table 5: Vas Deferens transcript annotations for top 10 TPM results generated by Kallisto.**

308

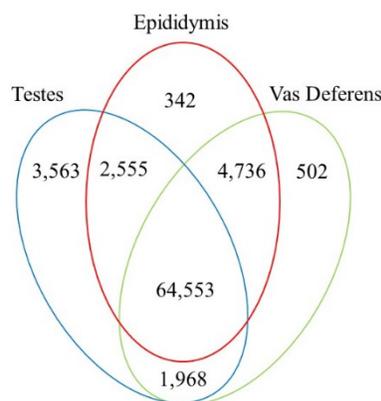
Transcript ID	Ensembl Accession Number	TPM	Protein	Gene
Transcript_67718	ENSMUSP00000080997	29027.1	mitochondrially encoded cytochrome c oxidase III	Mt-co3
Transcript_56089	ENSMUSP00000080993	19198.9	mitochondrially encoded cytochrome c oxidase I	Mt-co1
Transcript_48473	N/A	18225.6	N/A	N/A
Transcript_67719	ENSRNOP00000046414	13942.7	mitochondrially encoded cytochrome c oxidase II	Mt-co2
Transcript_43156	N/A	12965.1	N/A	N/A
Transcript_34960	ENSPPYP00000002849	11641.8	actin, alpha 2, smooth muscle, aorta	Acta2
Transcript_67721	ENSDORP00000013349	11397.6	actin, alpha 1, skeletal muscle	Acta1
Transcript_34961	ENSDNOP00000025628	10435.2	actin, gamma 2, smooth muscle, enteric	Actg2
Transcript_67765	ENSRNOP00000043141	8521.31	mitochondrially encoded NADH dehydrogenase 4	Mt-nd4
Transcript_37923	ENSMODP00000000682	7813.27	Uncharacterized protein	N/A

309

310 **Figures:**

311 **Figure 1: Venn Diagram of transcript expression differences and similarities between the**
 312 **three reproductive tissues. The total number of transcripts is 78,424.**

313

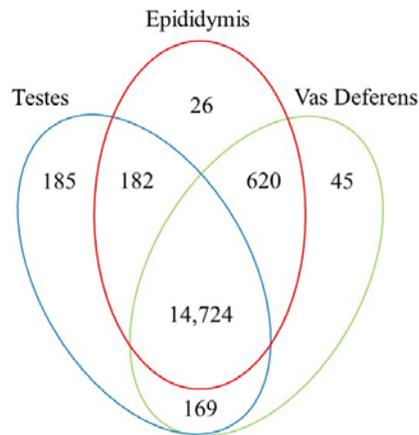


314

315

316 **Figure 2: Venn Diagram of transcript matches between the three reproductive tissues to**
317 **ncRNA sequences in *Mus musculus*. The total number of transcript matches across the**
318 **tissue types is 15,964.**

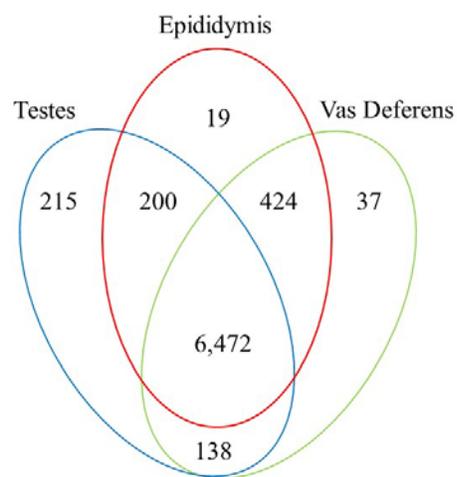
319



320

321

322 **Figure 3: Venn Diagram of transcript matches between the three reproductive tissues to**
323 **protein sequences in the Transporter Classification Database. The total number of**
324 **transcript matches across the tissue types is 7,521.**



325

326

327

328 **Dropbox Data File List:**

329 Final Annotated Reproductive Tissue Transcriptome: *reproductive.annotated.fasta* (127 MB)

330 Transdecoder (Five Files):

331 *transdecoder.gff3* (49 MB)

332 *transdecoder.pep* (27 MB)

333 *transdecoder.cds* (62 MB)

334 *transdecoder.mRNA* (172 MB)

335 *transdecoder.bed* (10 MB)

336 Pfam Annotation: *reproductive.pfam.gff3* (32 MB)

337 Rfam Annotation: *reproductive.rfam.gff3* (191 KB)

338 Dammit Annotation: *reproductive.dammit.gff3* (98 MB)

339 UniRef90 Annotation: *reproductive.uniref.gff3* (9.5 MB)

340 Kallisto Results for Annotated Transcriptome (Three Files):

341 *kallisto.testes.tsv* (3.4 MB)

342 *kallisto.epi.tsv* (3.4 MB)

343 *kallisto.vas.tsv* (3.4 MB)

344 ncRNA Database Matches (three files):

345 *epi.tpm.plus.ncRNA.txt* (4.3 MB)

346 *vas.tpm.plus.ncRNA.txt* (4.3 MB)

347 *testes.tpm.plus.ncRNA.txt* (4.3 MB)

348 tcdb Database Matches (three files):

349 *epi.tpm.plus.tcdb.txt* (946 KB)

350 *vas.tpm.plus.tcdb.txt* (946 KB)

351 *testes.tpm.plus.tcdb.txt* (946 KB)

352

353

354

355 **References:**

- 356 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search
357 tool. *J Mol Biol* 5:215(3):403-10.
- 358 Bedford NL, Hoekstra HE. 2014. *Peromyscus* mice as a model for studying natural variation.
359 *eLife*. doi: 10.7554/eLife.06813.
- 360 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence
361 data. *Bioinformatics* 30(15): 2114-2120. doi:10.1093/bioinformatics/btu170
- 362 Bray N, Pimentel H, Melsted P, Pachter L. (2015) Near-optimal RNA-seq quantification.
363 arXiv:1505.02710 [q-bio.QM]
- 364 Chen Q, Peng H, Lei L, Zhang Y, Kuang H, Cao Y, Shi Q, Ma T, Duan T. 2011. Aquaporin 3 is
365 a sperm water channel essential for postcopulatory sperm osmoadaptation and migration. *Cell*
366 *Research* 21: 922-933.
- 367 Cheviron ZA, Brumfield RT. 2011. Genomic insights into adaptation to high-altitude
368 environments. *Heredity*, 108(4), 354-361
- 369 Djureinovic D, Fagerburg L, et al. 2014. The human testes-specific proteome defined by
370 transcriptomics and antibody-based profiling. *Molecular Human Reproduction*, 20 (6):476-488.
371 Doi:10.1093/molehr/gau018.
- 372 Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms.
373 *Trends in Ecology and Evolution*, 29(1), 1-13.
- 374 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M,
375 Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein
376 families database: towards a more sustainable future. *Nucleic Acids Research: Database Issue*
377 44:D279-D285
- 378 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
379 Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F,
380 Birren BW, Nusbaum C, Lindbald-Toh K, Friedman N, Regev A. 2011. Full-length
381 transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*
382 29: 644-652.
- 383 Guillen Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, et al. 2015. Genomics of
384 Ecological Adaptation in Cactophilic *Drosophila*. *Genome Biology and Evolution*, 7(1), 349-
385 366.
- 386 Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-Seq: reference
387 generation and analysis with Trinity. *Nature Protocols* 8 (8): 1494-1512.
388 doi:10.1038/nprot.2013.084

- 389 *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of
390 mimicry adaptations among species. *Nature* 94:487. doi:10.1038/nature11041.
- 391 Hines HM, Papa R, Ruiz M, Papanicolaou A, Wang C, Nijhout HF, McMillan WO, Reed RD.
392 2012. Transcriptome analysis reveals novel patterning and pigmentation genes underlying
393 *Heliconius* butterfly wing pattern variation. *BMC Genomics* 13:288.
- 394 Hoekstra HE, Hirschmann RJ, Bunday RA, Insel PA, Crossland JP. 2006. A single amino acid
395 mutation contributes to adaptive beach mouse color patterns. *Science* 313:101-104. doi:
396 10.1126/science.1126121.
- 397 Jones FC, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*
398 484. 55-61. doi:10.1038/nature10944
- 399 Kriventseva EV, Tegenfeldt R, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, Ioannidis
400 P, Zdobnov EM. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the
401 underlying free software. *Nucleic Acids Research*:43(Database issue):D250-6.
402 doi: 10.1093/nar/gku1220. Epub 2014 Nov 26.
- 403 Lorenzo FR, Huff C, Myllymäki M, Olenchok B, Swierczek S, Tashi T, et al. 2014. A genetic
404 mechanism for Tibetan high-altitude adaptation. *Nature Genetics*, 46(9), 951-956.
- 405 MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data.
406 *Frontiers in Genetics* 5:13. doi.org/10.3389/fgene.2014.00013
- 407 MacManes MD, Eisen MB. 2014. Characterization of the transcriptome, nucleotide sequence
408 polymorphism, and natural selection in the desert adapted mouse. *PeerJ*, 2, e642-14
- 409 MacManes MD. 2016. Establishing evidenced-based best practice for the de novo assembly and
410 evaluation of transcriptomes from non-model organisms.
411 bioRxivdoi: <http://dx.doi.org/10.1101/035642>
- 412 Madden T. 2002 Oct 9 [Updated 2003 Aug 13]. The BLAST Sequence Analysis Tool. In:
413 McEntyre J, Ostell J, editors. *The NCBI Handbook* [Internet]. Bethesda (MD): National Center
414 for Biotechnology Information (US); 2002-. Chapter 16. Available from:
415 <http://www.ncbi.nlm.nih.gov/books/NBK21097/>
- 416 Marra NJ, Eo SH, Hale MC, Waser PM, DeWoody JA. 2012. *A priori* and *a posteriori*
417 approaches for finding genes of evolutionary interest in non-model species: Osmoregulatory
418 genes in the kidney transcriptome of the desert rodent *Dipodomys spectabilis* (banner-tailed
419 kangaroo rat). *Comparative Biochemistry and Physiology, Part D* 7: 328-339.
- 420 Marra NJ, Romero A, DeWoody A. 2014. Natural selection and the genetic basis of
421 osmoregulation in heteromyid rodents as revealed by RNA-seq. *Molecular Ecology* 23. doi:
422 10.1111/mec.12764

- 423 McCracken K, Barger C, Bulgarella M, Johnson K, Kuhner M, Moore A, et al. 2009a. Signatures
424 of High-Altitude Adaptation in the Major Hemoglobin of Five Species of Andean Dabbling
425 Ducks. *The American Naturalist*, 174(5), 631-650.
- 426 McCracken KG, Barger CP, Bulgarella M, Johnson KP, Sonsthagen SA, Trucco J, et al. 2009b.
427 Parallel evolution in the major haemoglobin genes of eight species of Andean waterfowl.
428 *Molecular Ecology*, 18(19), 3992-4005
- 429 Natarajan C, Hoffman FG, Lanier HC, Wolf CJ, Cheviron ZA, Spangler ML, Weber RE, Fago
430 A, Storz JF. 2015. Intraspecific polymorphism, interspecific divergence, and origins of function-
431 altering mutation in deer mouse hemoglobin. *Mol. Biol. Evol.* doi:10.1093/molbev/msu403.
- 432 Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner
433 PP, Jones TA, Tate J, Finn RD. 2014. Rfam 12.0: updates to the RNA families database. *Nucleic
434 Acids Research*: 10.1093/nar/gku1063
- 435 Patro R, Duggal G, Kingsford C. 2001 Accurate, fast, and model-aware transcript expression
436 quantification with Salmon. *Birxiv.org*, pages 1-35.
- 437 Saier MH, Reddy VS, Tamang DG, Vastermark A. 2014. The transporter classification database.
438 *Nucleic Acids Research*, 42(1):D251-8 [24225317].
- 439 Scott C. dammit: an open and accessible de novo transcriptome annotator. 2016.
440 www.camillescott.org/dammit
- 441 Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of
442 Mammalogists. 2011. Guidelines of the American society of mammalogists for the use of wild
443 mammals in research. *Journal of Mammalogy* 92(1):235-253
- 444 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
445 assessing genome assembly and annotation completeness with single-copy orthologs.
446 *Bioinformatics*. pii: btv351
- 447 Smith-Unna RD, Bournnell C, Patro R, Hibberd JM, Kelley S. 2015. TransRate: reference free
448 quality assessment of de-novo transcriptome assemblies. *BioRxiv*.
449 doi: <http://dx.doi.org/10.1101/021626>
- 450 Song L, Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq
451 reads. *Gigascience*. 4:48. doi: 10.1186/s13742-015-0089-y. eCollection 2015.
- 452 Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein
453 families based on seed alignments. *Proteins*, 28:405-420.
- 454 Storz JF, Runck AM, Moriyama H, Weber RE, Fago A. 2010. Genetic differences in
455 hemoglobin function between highland and lowland deer mice. *Journal of Experimental
456 Biology*, 213, 2565-2574.

- 457 Storz JF, Wheat CW. 2010. Integrating evolutionary and functional approaches to infer
458 adaptation at specific loci. *Evolution*, 64(9), 2489-2509.
- 459 Suzek BE, Huang H, McGarvey PB, Mazumder R, Wu CH. 2007. UniRef: comprehensive and
460 non-redundant UniProt reference clusters. *Bioinformatics*, 23 (10):1282-1288.
461 doi:10.1093/bioinformatics/btm098First published online: March 22, 2007
- 462 Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. 2015. UniRef clusters: a comprehensive
463 and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31 (6): 926-
464 932.doi:10.1093/bioinformatics/btu739First published online: November 13, 2014
- 465 Uhlen M, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science*,
466 347(6220):1260419. doi: 10.1126/science.1260419.
- 467 Veal R, Caire W. 2001. *Peromyscus eremicus*. *Mammalian Species* (118):1-6. Available at
468 <http://www.science.smith.edu/msi/pdf/i0076-3519-118-01-0001.pdf>.
- 469 Weber JN, Hoekstra HE. 2009. The evolution of burrowing behavior in deer mice (genus
470 *Peromyscus*). *Animal Behavior* 77: 603-609. doi: 10.1016/j.anbehav.2008.10.031
- 471 Weber JN, Peterson BK, Hoekstra HE. 2013. Discrete genetic modules are responsible for
472 complex burrow evolution in *Peromyscus* mice. *Nature* 493: 402-405. doi: 10.1038/nature11816.