

# SCIENTIFIC REPORTS



OPEN

## *De novo* assembly, annotation, and characterization of the whole brain transcriptome of male and female Syrian hamsters

Received: 28 July 2016  
Accepted: 06 December 2016  
Published: 10 January 2017

Katharine E. McCann, David M. Sinkiewicz, Alisa Norvelle & Kim L. Huhman

Hamsters are an ideal animal model for a variety of biomedical research areas such as cancer, virology, circadian rhythms, and behavioural neuroscience. The use of hamsters has declined, however, most likely due to the dearth of genetic tools available for these animals. Our laboratory uses hamsters to study acute social stress, and we are beginning to investigate the genetic mechanisms subserving defeat-induced behavioural change. We have been limited, however, by the lack of genetic resources available for hamsters. In this study, we sequenced the brain transcriptome of male and female Syrian hamsters to generate the necessary resources to continue our research. We completed a *de novo* assembly and after assembly optimization, there were 113,329 transcripts representing 14,530 unique genes. This study is the first to characterize transcript expression in both female and male hamster brains and offers invaluable information to promote understanding of a host of important biomedical research questions for which hamsters are an excellent model.

Syrian hamsters (*Mesocricetus auratus*) have been used in biomedical research for decades because they are uniquely suited for the study of a wide variety of behaviours and diseases. In recent years, however, the use of hamsters has declined<sup>1</sup>. A PubMed search of 'Syrian hamster' yields 2,280 publications before 1995, 856 publications from 1995–2004, and only 463 publications from 2005–2015. This decline is likely due to the advancement in genetic and molecular tools for other rodents, namely mice, and is not due to a reduction in the utility of hamsters in biomedical research. For example, hamsters provide an excellent model with which to study many types of cancer<sup>2,3</sup>, a variety of tumours<sup>4,5</sup>, and even pathogens such as Ebola virus<sup>6,7</sup>. Hormone release from the hypothalamic-pituitary-adrenal (HPA) axis, the so-called stress axis, in humans is more similar to that of hamsters than it is to that of other rodents, making hamsters a valuable model for studying behavioural and neurochemical responses to stress<sup>8–12</sup>. In addition, hamsters display robust circadian rhythms<sup>13,14</sup>, which make them an ideal subject for the study of the neurobiological basis of circadian rhythmicity. Finally, both male and female hamsters display a rich array of social and communicative behaviours, including intraspecific aggression and striking behavioural responses to social defeat stress<sup>15–19</sup>, allowing for the study of sex differences in a wide variety of endpoints using this species.

Historically, the vast majority of scientific research has used primarily male subjects, whether the study involved non-humans or humans. This has certainly been the case in the majority of neuroscience research using rodent models<sup>20</sup>. This bias towards males has historically been attributed to the complexity and variability introduced by working with females that have pronounced fluctuations in hormonal state, but it is also the case that, among mammals, some behaviours are not prominently produced by females (e.g., territorial aggression). Female rats and mice, for example, rarely produce any aggression outside of maternal defence of pups<sup>21</sup>. It is clearly the case, however, that female humans can be highly aggressive even outside of defence of offspring, thus rats and mice may not represent the best choice with which to model human agonistic behaviour. Female hamsters, on the other hand, readily display a range of social and agonistic behaviours toward male and female conspecifics<sup>18,22–25</sup>, presenting the opportunity to study social behaviour in both sexes rather than trying to generalize findings from males to females.

Neuroscience Institute Georgia State University, 161 Jesse Hill Jr. Drive Atlanta, Georgia 30303, USA. Correspondence and requests for materials should be addressed to K.E.M. (email: kmccann3@gsu.edu)

Sample Pool	RNA integrity number (RIN)	Concentration (ng/ $\mu$ l)
Female A	7.7	802
Female B	7.3	1286
Female C	7.3	848
Male A	7.4	1231
Male B	7.7	915
Male C	7.4	992

**Table 1.** Individual sample quality and concentration of each sample pool used for sequencing.

Our lab has established a model of social stress-induced behavioural change in Syrian hamsters that we have termed conditioned defeat. Conditioned defeat is the dramatic shift from territorial aggression to submission and social avoidance that is exhibited by both males and females after losing even a *single* agonistic encounter<sup>9,18,26</sup>. We have begun to examine the genetic and epigenetic markers of conditioned defeat but have been limited in this pursuit by a lack of specific probes and primers that are selective for hamster gene sequences. Thus, to advance the tools with which to investigate potential genetic mechanisms leading to conditioned defeat as well as to sexual dimorphisms in social behaviour, we sequenced the entire brain transcriptome of males and females. Here, we provide a detailed analysis of the brain transcriptome of male and female hamsters. This novel information about transcript expression in hamster brain will be of wide utility in a variety of fields that currently use hamsters as well as to fields that currently rely on mouse models of illnesses or behaviours for which hamsters would be ideal subjects.

## Results and Discussion

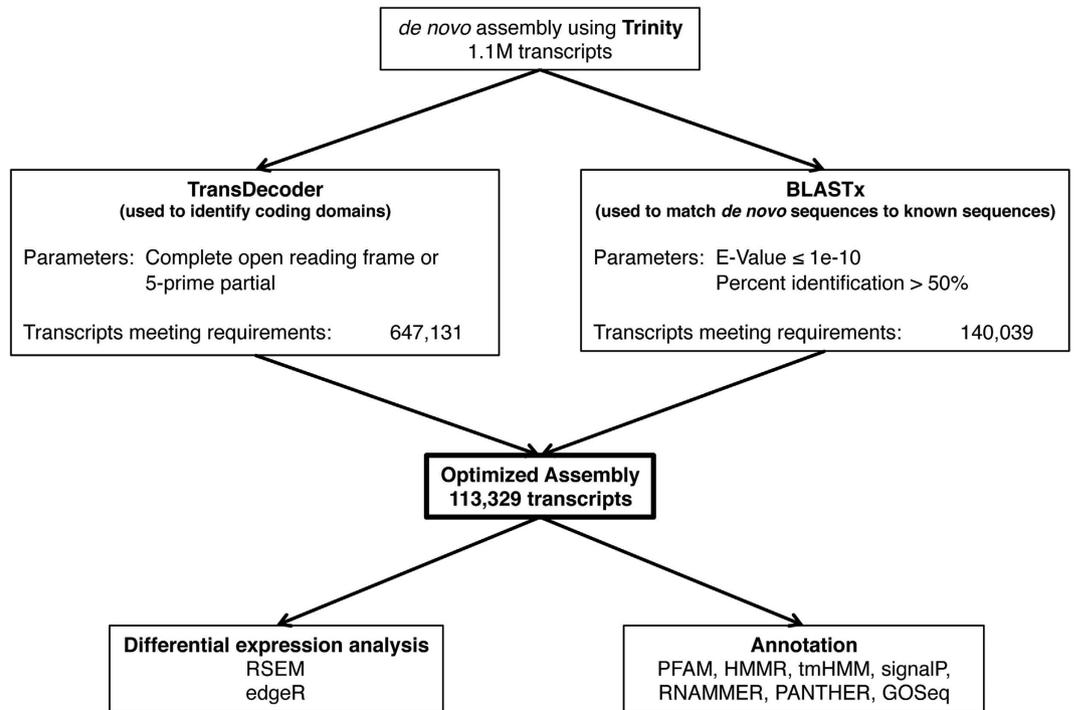
**Sample quality and description of raw reads.** All RNA samples were measured with the Agilent Bioanalyzer before sequencing. The RNA integrity numbers (a measure of sample quality) of all samples were good, falling between 7–8 (maximum value of 10), and all above the recommended cut-off of 6. Table 1 shows the RNA quality and concentration of each sample. Final raw sequence data was run through a quality assurance test (FastQC) to ensure minimal bias in sequencing and to confirm quality of starting library material. This test provides confidence in the quality of the sequence output before proceeding to assembly and annotation. Per base sequence quality scores all fell in the “very good” range (Phread score above 28) giving us the confidence to move forward with transcriptome assembly.

**Transcriptome assembly.** We assembled the Syrian hamster brain transcriptome using *de novo* techniques because, while there is a partially annotated Syrian hamster genome available (NCBI NW\_004801604.1, APMT 00000000.1), we were unable to reliably use this for a genome-guided assembly for several reasons. First, the genome currently available was sequenced from a single female hamster, thus eliminating the sequences of any Y-linked genes. One of the goals of this project was to develop tools to be able to directly compare males and females, so having Y-linked sequences would not only provide a positive control when looking at sex differences but would also lead to a more complete and representative transcriptome. In addition, the incomplete annotation of the current hamster genome leads to a number of problems when trying to build a transcriptome. The software currently available for building genome-guided assemblies assumes complete, or near-complete, annotation, and therefore returns error messages for any sequence that is not already annotated. Thus, we moved forward with a *de novo* assembly for more accurate and complete results.

The *de novo* assembly using Trinity revealed 1,002,166 total Trinity ‘genes’ and 1,147,108 transcripts from 973,648,406 total assembled bases. The average contig, or presumptive transcript, was 848.79 bases (median 440) with a percent GC content of 45.62. After completing the *de novo* assembly, raw reads were aligned back to the assembly. Proper pairs (both left and right reads aligned to same contig) accounted for 80.83% (539,735,450) of the 667,738,987 total aligned reads. Of the remaining pairs, left-only reads accounted for 9.68% (64,655,456) and right-only for 7.85% (52,410,243). Improper pairs, in which left and right reads align but to different contigs due to fragmentation, accounted for only 1.64% (10,937,838) of the total reads. These data provide an excellent starting point with which to build a usable transcriptomic database for Syrian hamster brains.

**Assembly optimization and annotation.** Trinity ‘genes’ are transcripts that may or may not code for a specific gene. Trinity *de novo* sequencing builds transcripts from sequence patterns that are *likely* to code for a gene. Without a genome to guide the assembly, some guesswork is involved in assembling the bases into known sequences. Thus, the approximation of the *de novo* assembly calls for several additional parameters to be put in place to build a more confident and usable transcriptome database. In order to gain confidence in our assembly and to minimize false positives as well as artificial sequences created by the *de novo* assembly, we ran a number of programs (see Methods) to optimize the assembly into an accurate representation of transcripts present in Syrian hamster brain, as done previously with other *de novo* assemblies in several fish and rodent species<sup>27–31</sup>. See Fig. 1 for a schematic of the assembly optimization process.

First, TransDecoder was run to determine the number of probable coding sequences within the assembly. Complete coding sequences accounted for 456,234 of the total number of open-reading frames (790,773). There were 108,213 3′-partial, 190,897 5′-partial, and 35,429 internal sequences. The sequencing protocol had a 3′ bias, thus we included all transcripts with 5′-partial and complete coding sequences for the initial assembly

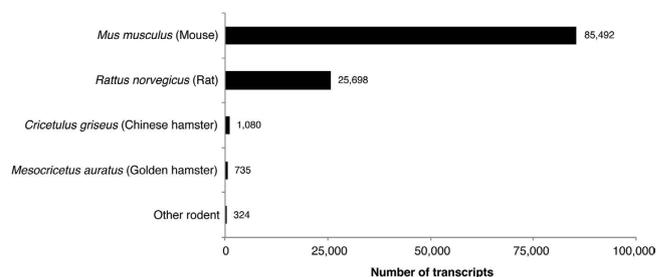


**Figure 1. Schematic of *de novo* assembly optimization and analysis.** After initial *de novo* assembly using Trinity, we optimized the assembly using several programs to omit falsely assembled sequences or sequences that were not likely to code for an actual gene. After optimization, we used RSEM to generate expected counts of each transcript from the raw reads and used those reads to calculate differential expression between males and females using edgeR. Annotation of the optimized assembly was completed using a series of annotation steps, PANTHER, and GOSeq.

optimization (647,131), as these transcripts were most likely to represent expressed genes<sup>32</sup>. We also filtered the assembly using data obtained from BLASTx using the Uniprot-rodent database (1/21/16) to ensure that all transcripts matched a known rodent sequence. BLASTx returned 1,219,140 matches, however many of these were at very low confidence parameters, thus only those with an E-value of  $\leq 1e-10$  and a percent identification match of  $\geq 50$  were included (140,039). These stringent parameters provide enhanced confidence in the quality of our optimized and annotated transcriptome<sup>28,29</sup>. Finally, we combined the output from TransDecoder and BLASTx, which left 113,329 transcripts meeting all the above stated criteria. An additional 27 transcripts were identified as containing a vector sequence during submission to the NCBI database and were removed from the final assembly. While this reduction process may have eliminated some sequences that represent true genes within hamster brain, these steps were necessary in order to eliminate a large number of false positives that can occur in *de novo* sequencing. Furthermore, BUSCO analysis revealed that 89% of the highly conserved sequences among vertebrates were present in the optimized assembly (2695 out of 3023), while 92% of the conserved genes across all eukaryotes were present (396 out of 429). These data also provide enhanced confidence in the quality and completeness of the optimized brain transcriptome.

We used the rodent database from Uniprot in order to maximize the number of transcripts in our assembly that matched a known sequence. Almost all of the transcripts matched *Mus musculus* (mouse) (85,492) and/or *Rattus norvegicus* (rat) (25,698), while 735 transcripts matched *Mesocricetus auratus* (golden hamster) as the top hit (Fig. 2). This is not surprising considering that the mouse genome is the most highly curated rodent genome available. Of the 113,329 individual transcripts in the optimized assembly, there were only 14,530 unique gene identifiers from BLAST, suggesting that there are multiple isoforms of some genes present in the assembly. This is consistent with data in mice and humans showing that there are approximately 17,000–25,000 genes in their respective genomes, with at least 10x the number of transcripts<sup>33–36</sup>. Of the 735 transcripts in the optimized transcriptome that matched *M. auratus*, there were 155 unique gene identifiers from BLAST. There are only 274 reviewed and annotated Syrian hamster genes in the Uniprot database, and more than half of those sequences match sequences from our *de novo* transcriptome assembly. Furthermore, many of the *de novo* sequences match multiple species in BLAST and only the top hit is recorded for this annotation. Therefore, it is likely that many more transcripts matched *M. auratus* but also matched another rodent (e.g., *M. musculus* or *R. norvegicus*) with an equal or higher score. Overall, the close alignment with the partially annotated hamster genome further validates our *de novo* assembly.

**Transcript expression analyses.** Using expected read counts from RSEM, we first compiled a matrix to determine which transcripts were most highly expressed in Syrian hamster brain. The genes represented by these transcripts are shown in Table 2 and, not surprisingly, represent genes that are highly expressed in brain tissue



**Figure 2. Number of transcripts matching specific rodent species.** The majority of transcripts in the optimized assembly matched *Mus musculus* as the top hit during annotation. Over 700 transcripts matched the partially annotated *Mesocricetus auratus* genome, suggesting a strong alignment of the *de novo* assembly with the available hamster genomic sequences.

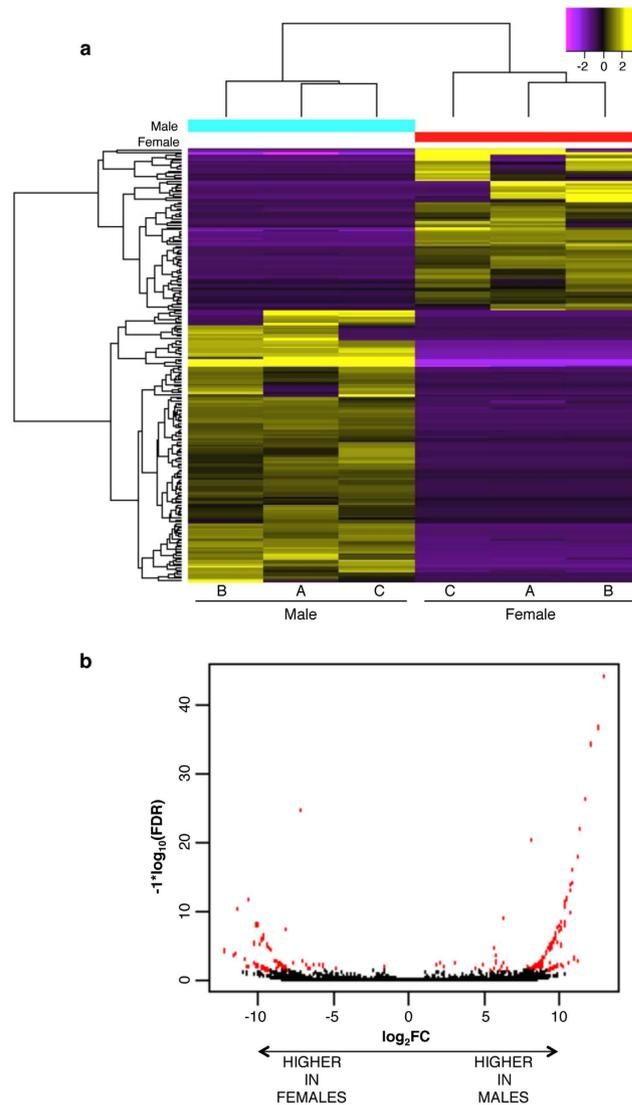
Gene ID	Gene	Uniprot ID
Nlrc3	Protein NLRC3	NLRC3_MOUSE
Plp1	Myelin proteolipid protein	MYPR_RAT
Scd2	Acyl-CoA desaturase 2	ACOD2_RAT
Hspa8	Heatshock cognate 71 kDa	HSP7C_RAT
Mbp	Myelin basic protein	MBP_MOUSE
Eef1a1	Elongation factor 1-alpha-1	EF1A1_RAT
Gapdh	Glyceraldehyde-3-phosphate dehydrogenase	G3P_CRIGR
Ywhag	14-3-3 protein gamma	1433G_RAT
Hsp90aa1	Heat shock protein HSP 90-alpha	HS90A_MOUSE
Sptbn1	Spectrin beta chain, non-erythrocytic 1	SPTB2_MOUSE
Atp5b	ATP synthase subunit beta, mitochondrial	ATPB_RAT
Glul	Glutamine synthase	GLNA_ACOCA
Aldoa	Fructose-bisphosphate aldolase A	ALDOA_RAT
Camk2n1	Calcium/calmodulin-dependent protein kinase II inhibitor 1	CK2N1_RAT
Atp2a2	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	AT2A2_MOUSE
Snrpn	Small nuclear ribonucleoprotein-associated protein N	RSMN_RAT
Psap	Prosaposin	SAP_RAT
Map1a	Microtubule-associated protein 1A	MAP1A_MOUSE
Serinc1	Serine incorporator 1	SERC1_RAT
Gpm6a	Neuronal membrane glycoprotein M6-a	GPM6A_RAT

**Table 2. Most highly expressed genes.** Top 20 genes that are the most highly expressed in Syrian hamster brain (both males and females).

of other species. For example, prosaposin is important for nervous system development and maintenance and microtubule-associated protein 1a is critical for neurogenesis and is found at its highest levels in brain tissue of rodents and humans<sup>37,38</sup>. Furthermore, several of the top expressed transcripts are nervous system-specific, including two of the top five expressed transcripts that are myelin-related (myelin proteolipid protein and myelin basic protein), as well as neuronal membrane glycoprotein M6-a.

We next completed differential expression analysis on the optimized transcriptome to determine what transcripts, if any, were differentially expressed in male and female brains. Excluding transcripts that did not meet the minimum expression cut off (see Methods), 207 transcripts were differentially expressed in the whole brain, the majority of which were higher in males compared with females (130 higher in males, 77 higher in females) (Fig. 3). Some of the differentially expressed transcripts (DETs) matched the same BLAST entry, suggesting that there may be differential regulation of multiple isoforms of these genes. The full list of genes represented by the DETs can be found in Supplemental Table S1.

There are several important considerations regarding DETs that should be addressed. First, they are presented here based on which sex had higher expression. It should be noted that the differential expression could, in fact, be the result of a decrease in expression of the opposite sex or a combination of an increase in one and a decrease in the other. Second, 207 is a reasonable number to expect for overall sex differences in whole brain based on data from both humans and *Drosophila*<sup>39,40</sup>, however this number can vary greatly depending on the statistical test and parameters used. Here, we use a stringent analysis previously used in other *de novo* assemblies and one recommended by the Trinity package<sup>28,41</sup>. Lastly, the differences reported here are representative of the entire brain, thus some sexually dimorphic genes may not be represented in our dataset due to differential regulation in different brain regions that may act to counterbalance or eliminate overall differences in expression. It is interesting to note



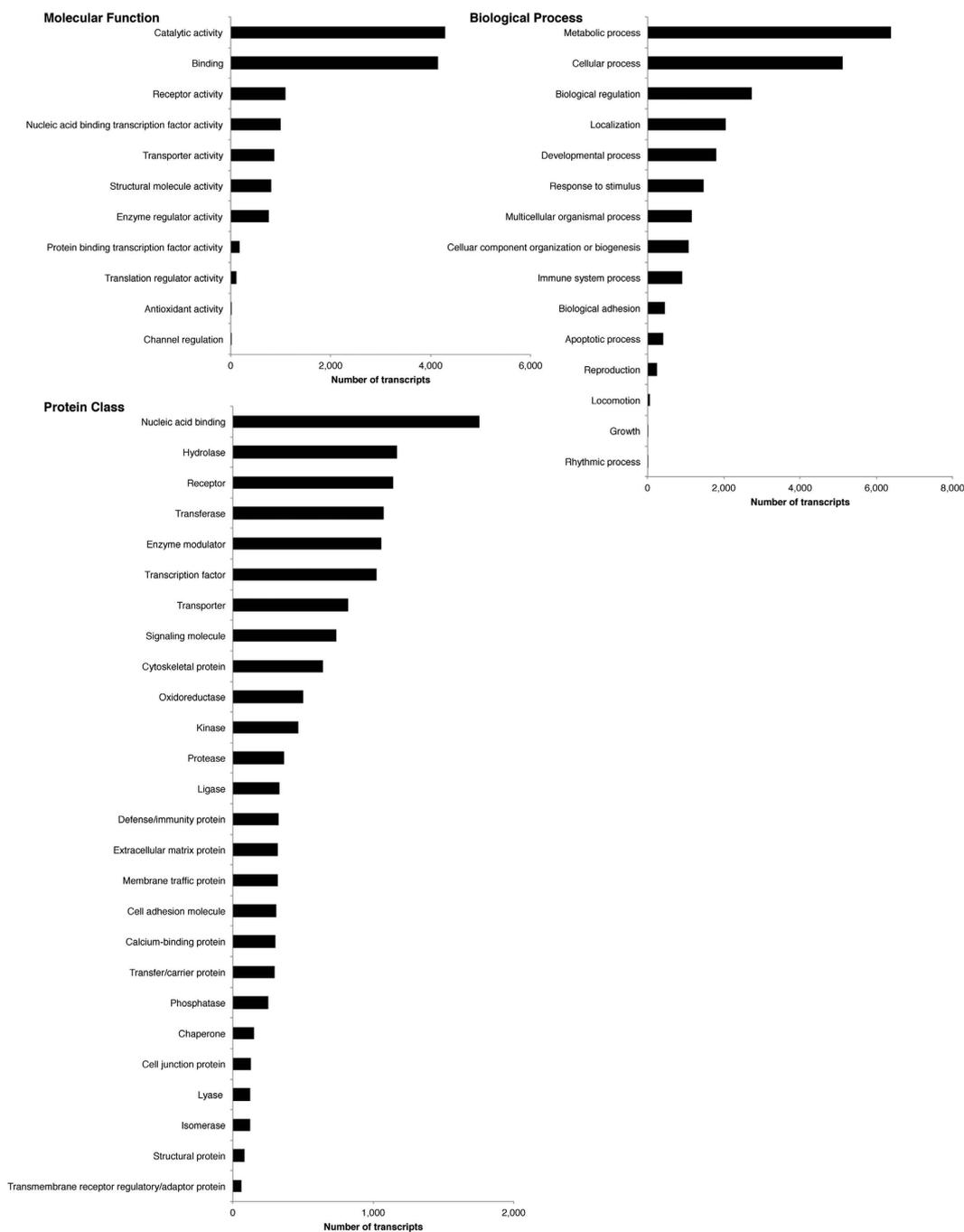
**Figure 3. Visualization of differential expression between male and female hamster brain.** (a) Heatmap showing the 207 differentially expressed transcripts between males and females. The left side of the heatmap indicates how the transcripts group together and fold change is shown by colour (yellow designating positive fold change, purple designating negative fold change). More transcripts (130) were higher in males than were higher in females (77). (b) Volcano plot of the transcripts expressed in hamster brain. Red indicates a significant difference in expression between males and females (FDR < 0.05).

that some of the DETs from the transcriptome represent genes that have been shown to be sexually dimorphic in other species. For example, one isoform of tolloid-like protein 1 (Tll1) was more highly expressed in females, while another isoform was higher in males. Tll1 has been linked to sex differences in behavioural response to stress in mice<sup>42</sup> and, based on the current data it may be of interest to further define the role of specific isoforms of this gene in both males and females. The consistency of sexual dimorphism in our hamster transcriptome compared with other species indicates that this *de novo* assembly will be a powerful toolkit for future use in hypothesis-driven investigation of gene expression in male and female hamsters.

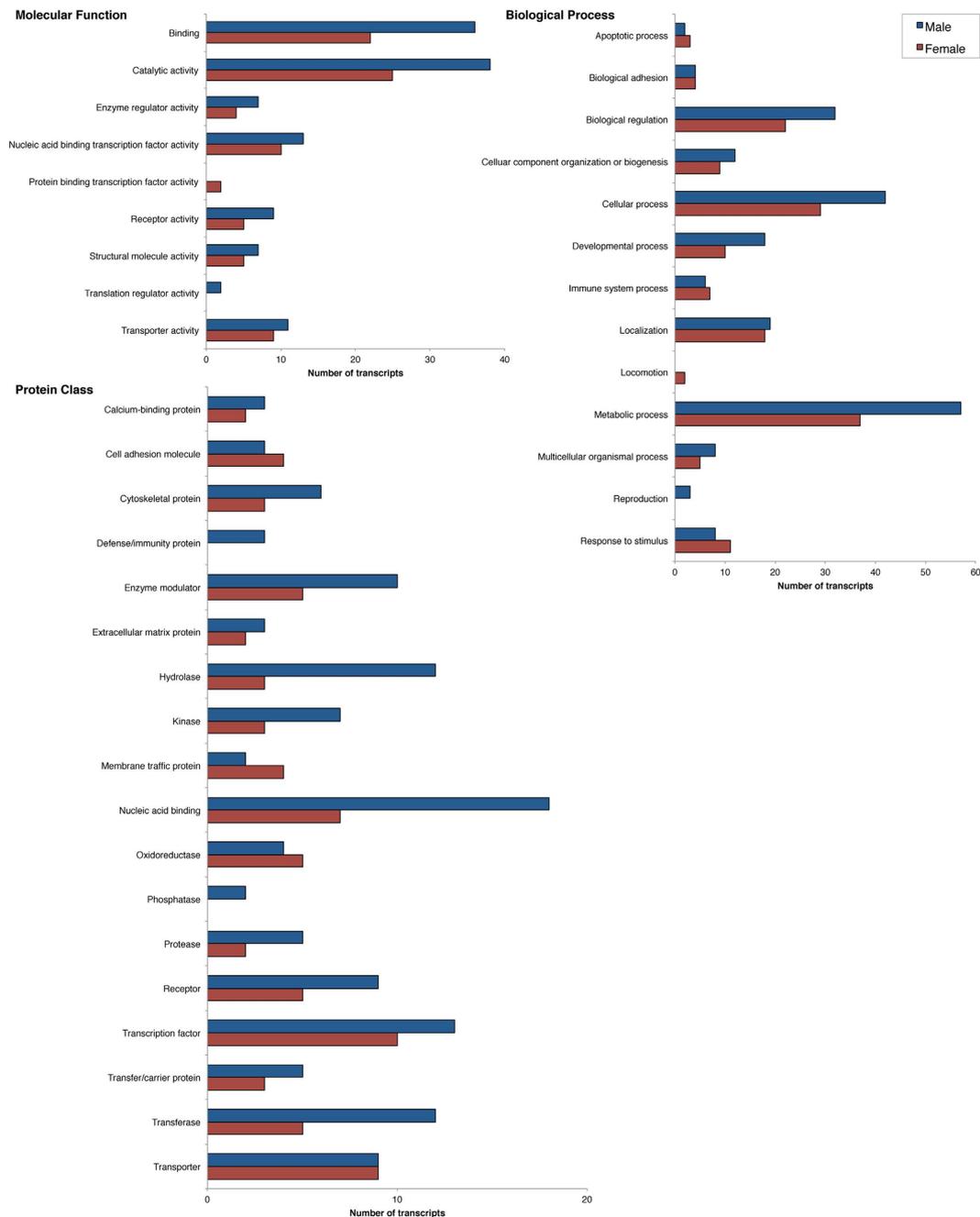
**Functional annotation and gene ontology (GO) analysis.** Annotation of the optimized assembly was completed using the steps outlined in Methods. The results from the assembly annotation are shown in Table 3. In order to complete functional annotation of the full brain transcriptome, we next filtered the annotated assembly through PANTHER analysis to determine which GO terms were highly represented in the optimized brain transcriptome. The top hits for each classification (molecular function, biological process, protein class) are presented in Fig. 4. Next, we examined the subsets of DETs to determine if any specific GO terms differed in their representation in these transcripts as compared with the complete transcriptome. The highest represented terms for each classification in males and females are presented in Fig. 5. Catalytic activity and binding were the most represented molecular functions in the full assembly as well as in the subsets of DETs. Likewise, the highest number of transcript matches for biological processes were cellular and metabolic processes.

Annotation Step	Purpose	Number of transcripts
RNAMMER	Identify rRNA transcripts	44
PFAM	Identify protein domains	103,916
SignalP	Predict signal peptides	29,004
tmHMM	Predict transmembrane regions	36,978

**Table 3. Annotation of assembly.** Number of transcripts represented in each step of the annotation process.



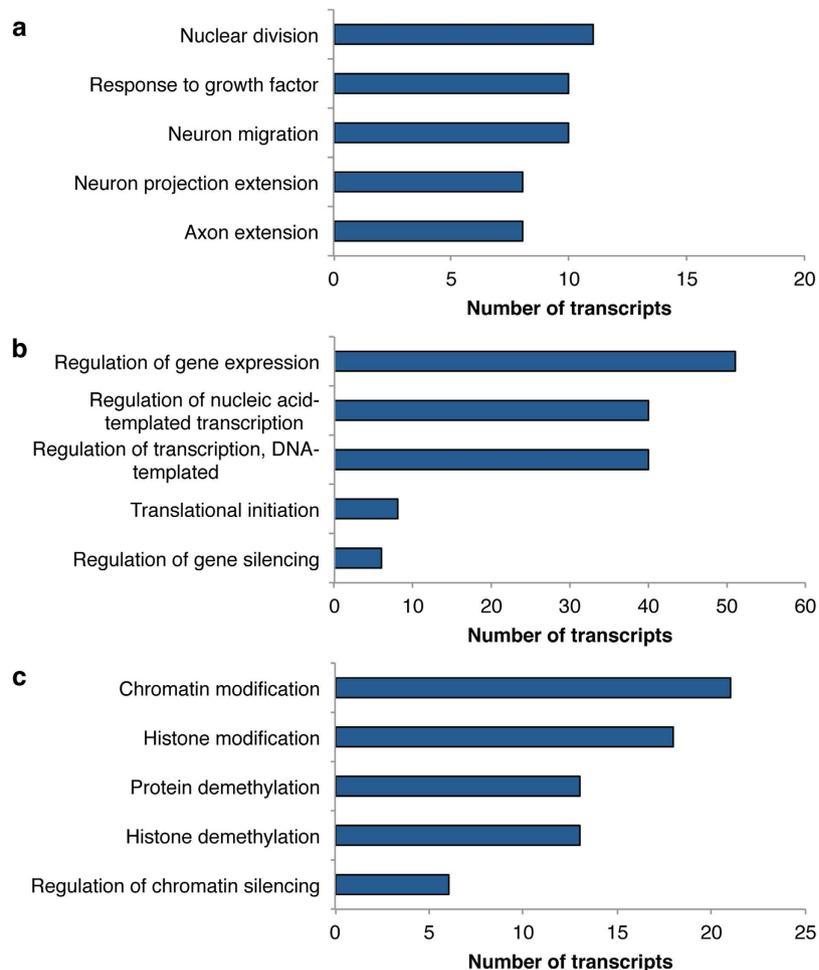
**Figure 4. Highest represented gene ontology terms from the optimized whole brain transcriptome.** We used PANTHER analysis to match the 14,530 unique genes in the optimized transcriptome to gene ontology terms for functional annotation of the assembly. These are the most represented functions in Syrian hamster brain.



**Figure 5. Highest represented gene ontology terms in the subsets of differentially expressed genes.** Highest represented gene ontology terms from PANTHER for the 130 genes more highly expressed in males (blue) and the 77 genes more highly expressed in females (red) in Syrian hamster brain.

Each category represented in Figs 4 and 5 has subcategories into which the transcripts can be further classified, and several interesting trends emerge when comparing the DETs. For example, the vast majority of transcripts associated with Localization in males (85.1%) and females (81.9%) matched the highest categories for the whole brain, including Vesicle, Protein, Ion, and Lipid Transport (81.8%). In addition, the majority of Receptors classified in the optimized brain transcriptome represented G-protein Coupled Receptor Activity (42.5%) but none of the transcripts that were differentially expressed between males and females were classified by this subcategory. In fact, Glutamate Receptor Activity was the only subcategory of Receptor represented in the transcripts with differential expression (higher in females). These functional classifications of the DETs may help to identify more precise targets for understanding sex differences in behaviour and future studies can explore these possibilities.

Finally, an enrichment analysis using GOSeq revealed 142 GO terms that were enriched in DETs that were more highly expressed in males. The majority of these terms (i.e., 100) were in the category of biological process and involved gene expression, epigenetic modification, and growth. A subset of these terms is highlighted in Fig. 6 and a full list can be viewed in Supplemental Table S2.



**Figure 6. Enriched gene ontology terms from GOSeq analysis.** GOSeq revealed 142 gene ontology terms that were enriched in males compared with females, the majority of which were categorized as Biological Processes. Subsets of these terms are shown here and involve processes related to (a) growth, (b) gene expression, and (c) epigenetic modification.

## Conclusions

These data represent the first comprehensive report of the Syrian hamster brain transcriptome and the first time that transcripts of both male and female hamsters have been sequenced and analysed. The differential expression analyses presented here between male and female baseline expression are not meant to provide a detailed analysis of sex differences in the brain but rather to provide a good starting point for analysing potential genetic and epigenetic mechanisms underlying sex differences in behaviour. Our lab is currently investigating site-specific sex differences in transcript expression in the brain using the tools developed here. Ultimately, the sequences obtained from this project will permit those conducting biomedical research to use Syrian hamsters when appropriate and to design custom primers and probes using hamster-specific sequences to answer important molecular and genetic questions.

## Methods

**Animals and tissue collection.** Six adult male and six adult female Syrian hamsters were obtained from Charles River Laboratories (Danvers, MA). Animals were approximately 10 weeks old upon arrival and weighed between 120–130 g. Subjects were singly housed for at least 2 weeks and handled daily. During handling, oestrous cycles of females were monitored for at least two cycles via vaginal swabs to confirm oestrous cycle stage and stability. For brain collection, animals were anesthetized via isoflurane exposure and then decapitated. All females were sacrificed on Dioestrus 2 to minimize variation in gene expression based on day of the oestrous cycle. This day of the cycle was chosen because we most often test female behaviour on Dioestrus 2 in our laboratory<sup>43</sup>. An equal number of males were sacrificed at the same time as the females. After decapitation, whole brains were rapidly extracted, frozen immediately in isopentane on dry ice, and stored at  $-80^{\circ}\text{C}$  until processing. All procedures and protocols were approved by the Georgia State University Institutional Animal Care and Use Committee and are in accordance with the standards outlined in the National Institutes of Health Guide for Care and Use of Laboratory Animals.

**RNA extraction.** Two brains from same-sex animals were pooled together for each RNA extraction in order to minimize the effect of individual variability. This sample size supplies sufficient power for downstream analyses while minimizing the total number of animals sacrificed<sup>44,45</sup>. We used Trizol (Life Technologies, Grand Island, NY) for extractions, following a modified version of the manufacturer's protocol. In brief, frozen brains were cut into large pieces and placed in 50 mL conical tubes on ice. Brains were homogenized on ice with 20 mL Trizol. After full homogenization, the sample was allowed to settle at room temperature for 5 min. The homogenate was then mixed with 4 mL of chloroform, allowed to stand at room temperature for 2–3 min and centrifuged at  $5,250 \times g$  for 45 min at 4 °C to separate the phases. The aqueous RNA phase was removed and dispensed into a new conical tube. 200  $\mu$ L/mL of chloroform was added to the aqueous phase, mixed well, allowed to stand 2–3 min, and then centrifuged at  $12,000 \times g$  for 10 min at 4 °C. For enhanced visualization of the pellet, 3  $\mu$ L/mL of GlycoBlue (Life Technologies, Grand Island, NY) was added and mixed gently. For RNA precipitation, 500  $\mu$ L/mL of 100% isopropanol was added, mixed gently and allowed to stand at room temperature for 10 min. To obtain an RNA pellet, the solution was centrifuged at  $12,000 \times g$  for 20 min at 4 °C. The remaining liquid was carefully removed and the pellet was washed twice in 75% RNase-free ethanol and centrifuged at  $7,500 \times g$  for 5 min at 4 °C. The pellet was allowed to air dry for approximately 5 min and was then re-suspended in 125  $\mu$ L of ultrapure water and immediately stored at –80 °C.

**RNA quality assurance and RNA sequencing.** RNA quality was assessed using the Agilent RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA) on the Agilent Bioanalyzer, following the manufacturer's instructions. RNA integrity numbers and concentration (ng/ $\mu$ L) were recorded and sent with the samples for sequencing. Samples (n = 6) were sent on dry ice to Beckman Coulter Genomics (Danvers, MA) for Illumina Automated RNA sequencing and were sequenced in paired-end 100 bp reads, averaging 110 M reads per sample. While it is true that 110 M reads may not allow for the identification of the entire transcriptome (e.g., microRNAs, non-coding RNAs), it should identify without any issue the mRNA landscape of male and female brain tissue<sup>46</sup>.

**Transcriptome assembly and optimization.** In order to produce a comprehensive brain transcriptome, we completed a *de novo* transcriptome assembly with Trinity (<https://github.com/trinityrnaseq/trinityrnaseq>)<sup>47,48</sup> using the jaccard clip parameter to minimize potential fusion transcripts. All data were acquired using the computing resources at Georgia State University<sup>49</sup>. After assembly, TransDecoder (<https://transdecoder.github.io/>)<sup>48</sup> was used to identify coding domain sequences with a minimum cut-off of 50 amino acids<sup>28</sup>. Assembled transcripts were also run through NCBI's BLASTx (National Center for Biotechnology Information's Basic Local Alignment Search Tool, <http://blast.ncbi.nlm.nih.gov/Blast.cgi>)<sup>50</sup> using the Uniprot-rodent database from January 21, 2016 (<http://www.uniprot.org/>)<sup>51</sup> to match *de novo* sequences to known genes.

Annotation of the assembly was accomplished using a series of annotation steps, including NCBI's BLAST to match sequences to known genes, Pfam<sup>52</sup> and HMMR<sup>53</sup> to identify protein domains, tmHMM<sup>54</sup> to predict transmembrane regions, signalP<sup>55</sup> to predict signal peptides, and RNAMMER<sup>56</sup> to identify rRNA transcripts. Finally, we compared our annotated assembly to a database of highly conserved orthologs using the BUSCO (Benchmarking Universal Single Copy Orthologs, <http://busco.ezlab.org/>) database to determine the completeness of our optimized assembly<sup>31,57</sup>.

We further identified gene ontology terms associated with our annotated transcripts using PANTHER (Protein Analysis Through Evolutionary Relationships, <http://pantherdb.org/>)<sup>58–61</sup>. We compared all genes using *Mus musculus* as the reference organism in PANTHER and identified the molecular functions, biological processes, and protein classes associated with the fully annotated transcriptome and the subsets of DETs, described below. GOSeq was used to perform the enrichment analysis on the differentially expressed transcripts<sup>62</sup>.

**Differential expression analysis.** Differential transcript expression in male and female hamster brains was calculated using an exact test in the Bioconductor R package (<https://www.r-project.org/>), edgeR (Empirical Analysis of Digital Gene Expression Data in R, <https://bioconductor.org/packages/release/bioc/html/edgeR.html>)<sup>63,64</sup>. We used RSEM (RNA-Seq by Expectation-Maximization, <http://deweylab.github.io/RSEM/>)<sup>65</sup> to generate read counts for the optimized assembled transcriptome to input into edgeR. EdgeR normalizes raw input data using a trimmed mean of M-values (TMM) and transcripts with artificially low counts (<1 count across all samples) after normalization were excluded before differential expression analysis was completed. Transcripts were considered to significantly differ in expression between males and females if the false discovery rate (FDR) was <0.05. All transcripts that met this criteria had a log<sub>2</sub> fold change of <–2 or >2, with the exception of two transcripts (–1.67, 1.86). These stringent parameters used as the cut-off for considering transcripts to be differentially expressed greatly minimizes the possibility of false positives.

## References

- Gao, M. *et al.* Generation of transgenic golden Syrian hamsters. *Cell research* **24**, 380–382; doi: 10.1038/cr.2014.2 (2014).
- Vairaktaris, E. *et al.* The hamster model of sequential oral oncogenesis. *Oral oncology* **44**, 315–324; doi: 10.1016/j.oraloncology.2007.08.015 (2008).
- LaRocca, C. J. *et al.* Oncolytic adenovirus expressing interferon alpha in a syngeneic Syrian hamster model for the treatment of pancreatic cancer. *Surgery* **157**, 888–898; doi: 10.1016/j.surg.2015.01.006 (2015).
- Li, J. J. & Li, S. A. Estrogen-induced tumorigenesis in hamsters: roles for hormonal and carcinogenic activities. *Archives of toxicology* **55**, 110–118 (1984).
- Gimenez-Conti, I. B. & Slaga, T. J. The hamster cheek pouch carcinogenesis model. *Journal of cellular biochemistry. Supplement* **17F**, 83–90 (1993).
- Wahl-Jensen, V. *et al.* Use of the Syrian hamster as a new model of ebola virus disease and other viral hemorrhagic fevers. *Viruses* **4**, 3754–3784; doi: 10.3390/v4123754 (2012).
- Prescott, J., Falzarano, D. & Feldmann, H. Natural Immunity to Ebola Virus in the Syrian Hamster Requires Antibody Responses. *The Journal of infectious diseases* **212** Suppl 2, S271–276; doi: 10.1093/infdis/jiv203 (2015).

8. Kollack-Walker, S., Watson, S. J. & Akil, H. Social stress in hamsters: defeat activates specific neurocircuits within the brain. *The Journal of neuroscience: the official journal of the Society for Neuroscience* **17**, 8842–8855 (1997).
9. Potegal, M., Huhman, K., Moore, T. & Meyerhoff, J. Conditioned defeat in the Syrian golden hamster (*Mesocricetus auratus*). *Behavioral and neural biology* **60**, 93–102 (1993).
10. Wommack, J. C. & Delville, Y. Repeated social stress and the development of agonistic behavior: individual differences in coping responses in male golden hamsters. *Physiology & behavior* **80**, 303–308 (2003).
11. Huhman, K. L., Moore, T. O., Ferris, C. F., Mougey, E. H. & Meyerhoff, J. L. Acute and repeated exposure to social conflict in male golden hamsters: increases in plasma POMC-peptides and cortisol and decreases in plasma testosterone. *Hormones and behavior* **25**, 206–216 (1991).
12. Huhman, K. L., Moore, T. O., Mougey, E. H. & Meyerhoff, J. L. Hormonal responses to fighting in hamsters: separation of physical and psychological causes. *Physiology & behavior* **51**, 1083–1086 (1992).
13. Albers, H. E. & Ferris, C. F. Neuropeptide Y: role in light-dark cycle entrainment of hamster circadian rhythms. *Neuroscience letters* **50**, 163–168 (1984).
14. Antle, M. C. & Mistlberger, R. E. Circadian clock resetting by sleep deprivation without exercise in the Syrian hamster. *The Journal of neuroscience: the official journal of the Society for Neuroscience* **20**, 9326–9332 (2000).
15. Albers, H. E., Huhman, K. L. & Meisel, R. L. Hormonal basis of social conflict and communication. *Hormones, brain and behavior* **1**, 393–433 (2002).
16. Bell, M. R. & Sisk, C. L. Dopamine mediates testosterone-induced social reward in male Syrian hamsters. *Endocrinology* **154**, 1225–1234; doi: 10.1210/en.2012-2042 (2013).
17. Huhman, K. L. Social conflict models: can they inform us about human psychopathology? *Hormones and behavior* **50**, 640–646; doi: 10.1016/j.yhbeh.2006.06.022 (2006).
18. Huhman, K. L. *et al.* Conditioned defeat in male and female Syrian hamsters. *Hormones and behavior* **44**, 293–299 (2003).
19. Kollack-Walker, S. & Newman, S. W. Mating and agonistic behavior produce different patterns of Fos immunolabeling in the male Syrian hamster brain. *Neuroscience* **66**, 721–736 (1995).
20. Beery, A. K. & Zucker, I. Sex bias in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews* **35**, 565–572 (2011).
21. St John, R. D. & Corning, P. A. Maternal aggression in mice. *Behavioral biology* **9**, 635–639 (1973).
22. Faruzzi, A. N., Solomon, M. B., Demas, G. E. & Huhman, K. L. Gonadal hormones modulate the display of submissive behavior in socially defeated female Syrian hamsters. *Hormones and behavior* **47**, 569–575; doi: 10.1016/j.yhbeh.2004.11.023 (2005).
23. Hennessey, A. C., Huhman, K. L. & Albers, H. E. Vasopressin and sex differences in hamster flank marking. *Physiology & behavior* **55**, 905–911 (1994).
24. Solomon, M. B., Karom, M. C. & Huhman, K. L. Sex and estrous cycle differences in the display of conditioned defeat in Syrian hamsters. *Hormones and behavior* **52**, 211–219; doi: 10.1016/j.yhbeh.2007.04.007 (2007).
25. Taravosh-Lahn, K. & Delville, Y. Aggressive behavior in female golden hamsters: development and the effect of repeated social stress. *Hormones and behavior* **46**, 428–435; doi: 10.1016/j.yhbeh.2004.03.007 (2004).
26. McCann, K. E. & Huhman, K. L. The effect of escapable versus inescapable social defeat on conditioned defeat and social recognition in Syrian hamsters. *Physiology & behavior* **105**, 493–497; doi: 10.1016/j.physbeh.2011.09.009 (2012).
27. Albertin, C. B. *et al.* The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**, 220–224; doi: 10.1038/nature14668 (2015).
28. Feng, N. Y., Fergus, D. J. & Bass, A. H. Neural transcriptome reveals molecular mechanisms for temporal control of vocalization across multiple timescales. *BMC Genomics* **16**, 408; doi: 10.1186/s12864-015-1577-2 (2015).
29. MacManes, M. D. & Lacey, E. A. The social brain: transcriptome assembly and characterization of the hippocampus from a social subterranean rodent, the colonial tuco-tuco (*Ctenomys sociabilis*). *PLoS one* **7**, e45524; doi: 10.1371/journal.pone.0045524 (2012).
30. Sharma, E. *et al.* Transcriptome assemblies for studying sex-biased gene expression in the guppy, *Poecilia reticulata*. *BMC Genomics* **15**, 400; doi: 10.1186/1471-2164-15-400 (2014).
31. Theissingner, K. *et al.* De Novo assembly and annotation of the freshwater crayfish *Astacus astacus* transcriptome. *Marine Genomics*, doi: 10.1016/j.margen.2016.02.006 (2016).
32. Senatore, A., Edirisinghe, N. & Katz, P. S. Deep mRNA sequencing of the *Tritonia diomedea* brain transcriptome provides access to gene homologues for neuronal excitability, synaptic transmission and peptidergic signalling. *PLoS one* **10**, e0118321; doi: 10.1371/journal.pone.0118321 (2015).
33. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
34. Hastie, N. D. & Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**, 761–774 (1976).
35. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6062–6067 (2004).
36. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
37. Carvelli, L., Libin, Y. & Morales, C. R. Prosaposin: a protein with differential sorting and multiple functions. *Histol Histopathol* **30**, 647–660; doi: 10.14670/HH-30.647 (2015).
38. Liu, Y., Lee, J. W. & Ackerman, S. L. Mutations in the microtubule-associated protein 1A (Map1a) gene cause Purkinje cell degeneration. *The Journal of neuroscience: the official journal of the Society for Neuroscience* **35**, 4587–4598; doi: 10.1523/JNEUROSCI.2757-14.2015 (2015).
39. Catalan, A., Hutter, S. & Parsch, J. Population and sex differences in *Drosophila melanogaster* brain gene expression. *BMC Genomics* **13**, 654; doi: 10.1186/1471-2164-13-654 (2012).
40. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun* **4**, 2771; doi: 10.1038/ncomms3771 (2013).
41. Fraser, B. A., Weadick, C. J., Janowitz, I., Rodd, F. H. & Hughes, K. A. Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. *BMC Genomics* **12**, 202; doi: 10.1186/1471-2164-12-202 (2011).
42. Tamura, G., Olson, D., Miron, J. & Clark, T. G. Tolloid-like 1 is negatively regulated by stress and glucocorticoids. *Brain Res Mol Brain Res* **142**, 81–90; doi: 10.1016/j.molbrainres.2005.09.016 (2005).
43. McCann, K. E. *et al.* Histone deacetylase and acetyltransferase inhibitors modulate behavioral responses to social stress. *Psychoneuroendocrinology* **75**, 100–109; doi: 10.1016/j.psyneuen.2016.10.022 (2017).
44. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13; doi: 10.1186/s13059-016-0881-8 (2016).
45. Sonesson, C., Gerster, S. & Delorenzi, M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS one* **9**, e100335; doi: 10.1371/journal.pone.0100335 (2014).
46. Francis, W. R. *et al.* A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* **14**, 167; doi: 10.1186/1471-2164-14-167 (2013).
47. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644–652; doi: 10.1038/nbt.1883 (2011).
48. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494–1512; doi: 10.1038/nprot.2013.084 (2013).

49. Sarajilic S. *et al.* Orion: Discovery Environment for HPC Research and Bridging XSEDE Resources. *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* (XSEDE16). ACM, New York, NY, USA, Article 54; doi: 10.1145/2949550.2952770 (2016).
50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410; doi: 10.1016/S0022-2836(05)80360-2 (1990).
51. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204–212; doi: 10.1093/nar/gku989 (2015).
52. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40**, D290–301; doi: 10.1093/nar/gkr1065 (2012).
53. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–37; doi: 10.1093/nar/gkr367 (2011).
54. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567–580; doi: 10.1006/jmbi.2000.4315 (2001).
55. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**, 785–786; doi: 10.1038/nmeth.1701 (2011).
56. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108; doi: 10.1093/nar/gkm160 (2007).
57. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212; doi: 10.1093/bioinformatics/btv351 (2015).
58. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29; doi: 10.1038/75556 (2000).
59. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols* **8**, 1551–1566; doi: 10.1038/nprot.2013.092 (2013).
60. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* **44**, D336–342; doi: 10.1093/nar/gkv1194 (2016).
61. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol* **563**, 123–140; doi: 10.1007/978-1-60761-175-2\_7 (2009).
62. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14; doi: 10.1186/gb-2010-11-2-r14 (2010).
63. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140; doi: 10.1093/bioinformatics/btp616 (2010).
64. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria (ISBN 3-900051-07-0) <http://www.R-project.org/> (2014).
65. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 1–16; doi: 10.1186/1471-2105-12-323 (2011).

## Acknowledgements

We gratefully acknowledge the use of Orion that is supported by Georgia State University's Research Solutions. The authors would also like to acknowledge N Edirisinghe, A Signore, AD Guzman Bambaren, BM Thompson, and GMF Jones for their assistance with this project. Research reported here was supported by Brains and Behavior Fellowships awarded to KEM and DMS, by the Honeycutt Fellowship awarded to KEM, by the National Institute of Mental Health of the National Institutes of Health under Award Number R01MH062044 awarded to KLH, and by a Seed Grant awarded to KLH from the Brains and Behavior Program at Georgia State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Georgia State University.

## Author Contributions

K.E.M., A.N., and K.L.H. designed experiments. K.E.M. conducted all animal work and sample preparation. K.E.M. and D.M.S. completed transcriptomic assembly and all analyses. K.E.M., D.M.S., A.N., and K.L.H. prepared the manuscript.

## Additional Information

**Accession codes:** This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GEMX00000000 (BioProject PRJNA320732). The version described in this paper is the first version, GEMX01000000.

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** McCann, K. E. *et al.* *De novo* assembly, annotation, and characterization of the whole brain transcriptome of male and female Syrian hamsters. *Sci. Rep.* **7**, 40472; doi: 10.1038/srep40472 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017