Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data on the first functionally-annotated *de novo* transcriptome assembly for North American flying squirrels (genus *Glaucomys*)

Michael G.C. Brown[a,*], Jeff Bowman[b], Paul J. Wilson[c]

[a] Environmental and Life Sciences Graduate Program, Trent University, Peterborough, Canada
[b] Wildlife Research and Monitoring Section, Ontario Ministry of Natural Resources and Forestry, Peterborough, Canada
[c] Biology Department, Trent University, Peterborough, Canada

ARTICLE INFO

Article history:
Received 11 March 2021
Revised 16 June 2021
Accepted 22 July 2021
Available online 25 July 2021

Keywords:
RNA-seq
Transcriptome
De novo assembly
Functional annotation
High-throughput sequencing
Hybridization
Flying squirrels

ABSTRACT

We report the first functionally-annotated *de novo* transcriptome assembly for North American flying squirrels (genus *Glaucomys*). RNA was extracted from tissue samples obtained from two northern flying squirrels and two southern flying squirrels sampled from Ontario, Canada, and sequenced on an Illumina paired-end sequencing platform. We reconstructed 702,228 *Glaucomys* transcripts using 193,323,120 sequence read pairs and captured sequence homologies, protein domains, and gene function classifications. Introgressive hybridization between northern (*Glaucomys sabrinus*) and southern flying squirrels (*G. volans*) has been observed in some areas of North America. However, existing molecular markers lack the resolution to discriminate late-generation introgressants and describe the extent to which hybridization influences the *Glaucomys* gene pool. These genomic resources can increase the resolution of molecular techniques used to examine the dynamics of the *Glaucomys* hybrid zone.

* Corresponding author.
E-mail address: michaelgcbrown@outlook.com (M.G.C. Brown).
Social media: (J. Bowman), (P.J. Wilson)

https://doi.org/10.1016/j.dib.2021.107267

Specifications Table

Subject	Biochemistry, Genetics and Molecular Biology (General)
Specific subject area	Transcriptomics
Type of data	Table Figure XLS Worksheet RNA raw sequencing data Assembled contigs
How data were acquired	Illumina HiSeq 2500 platform. The sequence read-pairs were assembled *de novo* using Trinity, and the resulting transcriptome was annotated using the Trinotate pipeline.
Data format	Illumina Hiseq 2500 raw sequence reads in FASTQ format, *de novo* assembled trasncriptome in FASTA format, Trinotate annotation report in Microsoft Excel 97–2003 Worksheet (.xls) format.
Parameters for data collection	Brain tissue was collected from two Southern flying squirrels (*Glaucomys volans*) and two Northern flying squirrels (*G. sabrinus*).
Description of data collection	Total RNAs isolated from each *Glaucomys* individual were sequenced on an Illumina HiSeq 2500 platform.
Data source location	Institution: Trent University City/Town/Region: Peterborough, Ontario Country: Canada Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: *G. sabrinus* (sample NFS6525): 44.633563, −78.726709 *G. sabrinus* (sample NFS50254): 45.674719, −78.322426 *G. volans* (sample SFSCC1): 45.172767, −78.837508 *G. volans* (sample SFS25428): 42.633097, −80.612048
Data accessibility	Trimmed RNA paired-end sequence reads are deposited in the NCBI database under SRA accession number PRJNA705604. https://www.ncbi.nlm.nih.gov/sra/PRJNA705604 The associated annotation data generated with the Trinotate program are available as Supplementary Material.

Value of the Data

- This functionally annotated *de novo* transcriptome assembly of the North American flying squirrel genus (*Glaucomys*) represents ecologically-important forest obligate species that exhibit climate-induced interspecific hybridization.
- The functionally annotated *Glaucomys* RNA sequence data present a useful source of transcribed genomic variation.
- This dataset will permit downstream queries of differential expression and genomic variation that will allow researchers to investigate the dynamics of introgressive hybridization in response to a warming climate.
- The data provided can be used to investigate the dynamics of climate-induced introgressive hybridization in *Glaucomys*, by informing the development of markers to investigate expression divergence in zones of hybridization, detect late-generation introgressants with greater accuracy, and survey for evidence of adaptive introgression by characterizing interspecific functional genomic polymorphisms responsible for modulating gene expression and protein function.

1. Data Description

Illumina RNA-*seq* produced a total of 193,323,120 raw read-pairs among four samples. The mean number of paired-end sequence reads in each of the cDNA libraries was 48,330,780, and one-third of the sequence reads contained adapters. Approximately 6 million bp were processed in each of the cDNA libraries using trimming software, and ~4% of the total number of base pairs were quality trimmed of adapters, low quality sequences and empty reads. Of the original raw sequence reads, ~96% were retained as cleaned reads after the quality trimming process. The Trinity program assembled a transcriptome consisting of 702,228 transcripts contained within 584,007 Trinity 'genes'. The quality of the transcriptome assembly was validated by assessing the RNA-*seq* read representation, which is accomplished by aligning the original reads to the assembly. The read composition of the assembly was high; alignment rates of read pairs used to generate the *Glaucomys* assembly were ~95%. Summary statistics for the assembled transcriptome are shown in Table 1. The raw RNA sequence reads for each individual and the assembled transcripts are available in the Appendix.

Queries of Trinity transcripts and TransDecoder-predicted proteins in the *Glaucomys* transcriptome using BLAST and HMMER (Table 2) yielded 803,744 total annotations containing 2705,275 hits. Of the total annotations and hits, 329,538 and 116,325 were unique, respectively. A TransDecoder search of protein sequences identified 157,682 candidate protein-coding genes from the *Glaucomys* assembly. A BLASTp search found that over half of the candidate protein-coding genes had sequence homologies with the Swiss-Prot database. BLASTx queries of Trinity transcripts accounted for the highest amount of total sequence homology hits and number of annotated transcripts. Of the total Trinity transcripts, approximately 25% mapped to homologous protein sequences in the Swiss-Prot database. The distribution of top BLAST species hits (Tables 3, 4) were similar among both methods, with human (*Homo sapiens*) and house mouse (*Mus musculus*) accounting for most hits in each. BLASTp hits were dominated by human which included almost half of the total hits, while human hits in BLASTx were much lower. Norway rat (*Rattus norvegicus*) and bovine (*Bos taurus*) were also common among BLAST hits, however accumulating much less. The top-hit species distribution from BLASTx and BLASTp searches of

Table 1
Summary statistics of a *de novo Glaucom*ys transcriptome assembled with the program Trinity, using RNA-*seq* reads sampled from brain tissue of two *G. volans* and two *G. sabrinus.*

Parameter	Value
Total number of base pairs assembled (bp)	23,309,493,176
Total Trinity 'genes'	584,007
Total number of assembled contigs (Trinity transcripts)	702,228
Candidate protein sequences	157,682
Average contig length (bp)	721
Minimum contig length (bp)	201
Maximum contig length (bp)	18,691
Total length of all contigs in assembly (bp)	506,799,925
GC content of contigs (%)	46.46
N50 (bp)	1261
N30 (bp)	2523
N90 (bp)	279

Table 2
Summary of BLAST and HMMER queries of Swiss-Prot and Pfam databases, respectively, from the Trinotate functional annotation of a *de novo Glaucomys* transcriptome assembly.

Query type	No. of unique annotations	No. of unique top hits	No. of annotated sequences	No. of total hits
BLASTx/Swiss-Prot	140,644	31,045	173,419	314,401
BLASTp/Swiss-Prot	71,639	23,539	93,781	124,623
HMMER/Pfam	53,406	5888	71,223	176,592

Table 3
Top hit species distribution of Swiss-Prot BLASTx queries for a *de novo Glaucomys* transcriptome assembly.

Rank	ID	Scientific name	Common name	No. of hits	% of total hits
1	HUMAN	*H. sapiens*	human	147,554	27.1
2	MOUSE	*M. musculus*	house mouse	61,502	11.3
3	ARATH	*Arabidopsis thaliana*	mouse-ear cress	21,841	4.0
4	RAT	*R. norvegicus*	Norway rat	21,184	3.9
5	BOVIN	*B. taurus*	bovine	13,725	2.5

Table 4
Top hit species distribution of Swiss-Prot BLASTp queries for a *de novo Glaucomys* transcriptome assembly.

Rank	ID	Scientific name	Common name	No. of hits	% of total hits
1	HUMAN	*H. sapiens*	human	63,333	42.9
2	MOUSE	*M. musculus*	house mouse	20,941	14.2
3	RAT	*R. norvegicus*	Noway rat	7684	5.2
4	BOVIN	*B. taurus*	bovine	5922	4.0
5	PONAB	*P. abelii*	orangutan	3325	2.3

Table 5
Top 10 Pfam domains identified from candidate proteins in a *de novo Glaucomys* transcriptome assembly.

Rank	Pfam domain	No. of hits
1	Zinc finger, C2H2 type	7791
2	Ankyrin repeat	3136
3	WD domain, G-beta repeat	2972
4	C2H2-type zinc finger	2914
5	Ankyrin repeats	2546
6	Ankyrin repeats	2465
7	Protein kinase domain	2260
8	Protein tyrosine kinase	2200
9	C2H2-type zinc finger	1899
10	Ankyrin repeat	1795

homologous protein sequences in Swiss-Prot had similar representations and were both dominated by human, house mouse, and Norway rat. The BLASTp top hit species distribution results differed in that Norway rat, bovine and Sumatran orangutan (*Pongo abelii*) were the most common top hits after human and house mouse.

HMMER searches found that almost half of the candidate protein sequences had homologies with the Pfam database. Of the top 10 protein domain hits (Table 5) in the *Glaucomys* transcriptome, 'Zinc finger, C_2H_2 type' (4.4%) had the most hits. The zinc finger protein family (C_2H_2 type) was highly represented among the top Pfam domains identified in the *Glaucomys* transcriptome, accounting for approximately one-third of the top domains. Ankyrin repeat protein families were also highly represented, accounting for almost half of the top Pfam domains identified in the *Glaucomys* transcriptome and 5.6% of total domain hits.

Protein sequence homologies captured from Swiss-Prot and Pfam databases using BLAST and HMMER, respectively, generated almost 2 million GO terms from the GO database (Table 6). Over 25% of transcripts within the *Glaucomys* assembly were annotated with GO terms, exceeding that of COG, eggNOG and KEGG annotations. COG, KEGG, and eggNOG annotations only return one term when a match is detected in Swiss-Prot or Pfam. The number of unique annotations retrieved, and number of annotated transcripts and candidate proteins from Swiss-Prot-captured GO and KEGG terms were very similar, while the number of transcripts annotated with Pfam protein domains and COGs were nearly exact.

Almost half of all Swiss-Prot captured GO terms for transcripts and candidate protein sequences belonged to the 'biological processes' domain (Table 7). The least amount of Swiss-Prot-captured GO terms were associated with 'cellular components', however, with a similar

Table 6
Summary totals for Gene Ontology (GO) and functional ortholog annotations for a *de novo Glaucomys* transcriptome assembly.

Database	No. of unique annotations	No. of unique top terms	No. of annotated sequences	No. of total terms
COG	1307	–	44,413	–
eggNOG	7113	–	97,301	–
KEGG	28,817	–	129,005	–
GO BLAST	24,244	17,124	149,686	1723,200
GO Pfam	2368	1492	44,913	95,737

Table 7
Functional representation of GO term domains queried from Swiss-Prot and Pfam databases for a *de novo Glaucomys* transcriptome assembly.

Ontology domain	No. of terms (Swiss-Prot)	% of total terms (Swiss-Prot)	No. of terms (Pfam)	% of total terms (Pfam)
Biological process	811,624	47.1	28,416	29.7
Cellular component	503,982	29.2	14,350	15.0
Molecular function	407,594	23.7	52,971	55.3

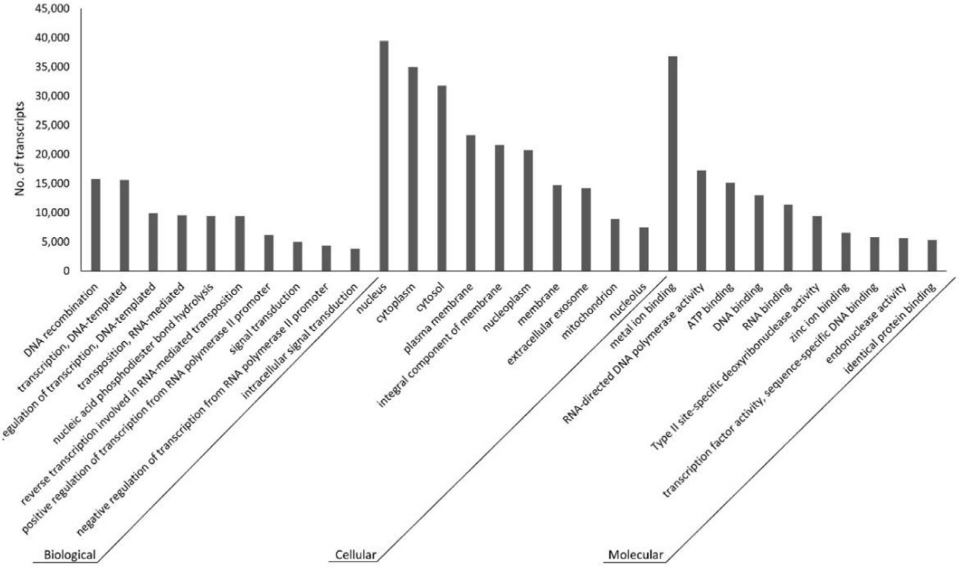

Fig. 1. Distribution of the top-10 GO terms for each ontology domain retrieved from Swiss-Prot sequence homology hits for the *de novo Glaucomys* transcriptome assembly.

representation to that of ‘molecular function’. This distribution of GO terms was not found with those captured by Pfam; ontology domains were highly represented by molecular function, followed by biological process and cellular component. The most frequent GO term queried from Swiss-Prot was ‘nucleus’ within the cellular component domain, followed by ‘metal ion binding’ within the molecular component domain (Fig. 1). Within the cellular component domain, a large proportion of the total GO terms were distributed among the top terms, with six term categories (‘nucleus’, ‘cytoplasm’, ‘cytosol’, ‘plasma membrane’, ‘integral component of membrane’, and ‘nucleoplasm’) exceeding the highest term frequencies found within the biological process domain. The distribution of the number of GO terms in the biological process and molecular function domains were more evenly distributed among each term.

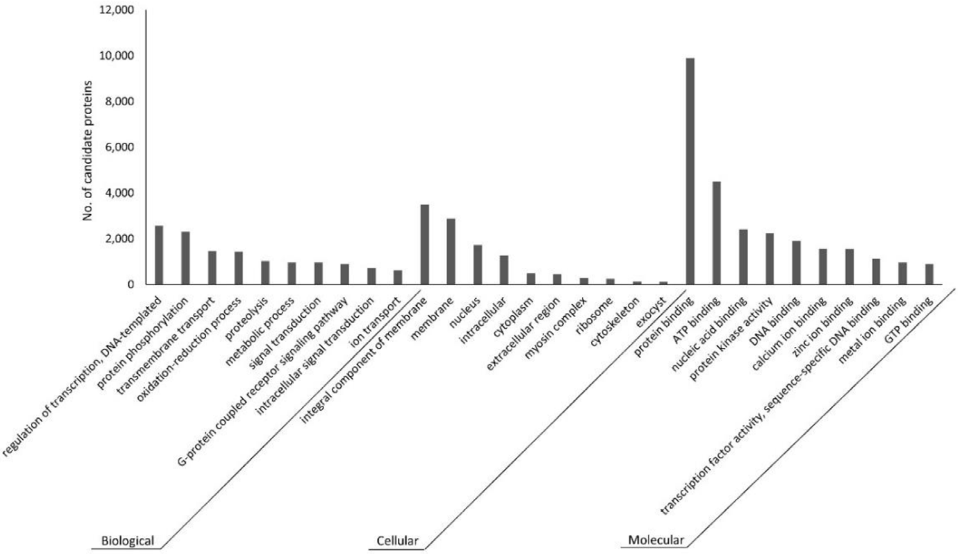

Fig. 2. Distribution of the top-10 Pfam-derived protein sequence homology GO term hits within each ontology domain for the *de novo Glaucomys* transcriptome assembly.

Among the GO terms retrieved from Pfam, the frequency of 'protein binding' and 'ATP binding' annotations within the 'molecular function' domain was highly represented compared to other top terms (Fig. 2). The frequency of each GO term within biological process domain displayed a more even distribution than the other domains, with the top 10 terms capturing most of the total GO term frequency for that domain. The top four GO terms within the cellular component domain captured most annotations within that domain.

We retrieved pathway annotations for 18.7% of our Trinity transcripts from the KEGG database, and of those transcripts, ~10% were found to have orthologs in the KO database (Fig. 3). Most KO gene annotations were categorized as having 'organismal system function' (30.6%), followed by 'metabolism' (24.8%) and 'environmental information processing' (22.1%). 'Genetic information processing' accounted for the least number of KO annotations (8.4%). Within the top represented category of 'organismal system function', orthologs related to 'immune system function' were highly enriched (26.8%), as was 'endocrine system' (25.3%). The genes grouped within 'signal transduction', and 'global and overview maps', represented within the 'categories of environmental information processing' and 'metabolism', respectively, displayed the highest enrichment of orthologs. The gene that was most enriched among KO annotations was 'metabolic pathways', grouped under 'global and overview maps' within the 'metabolism' classification. There were 881 genes with 'metabolic pathways' orthologs, outnumbering the next most enriched ortholog – biosynthesis of secondary metabolites (233) – found in the same category.

Annotations from Orthologous Groups (OGs) databases were compiled for 20% of Trinity transcripts, with eggNOG annotations more than doubling those of COGs. Seven of the top 10 annotations were shared between each database; six of which were found in the same distributional order. The frequency of top eggNOG annotations was highly distributed among three categories, two of which were an 'endo/exonuclease/phosphotase' functional category (Fig. 4). The frequency of COGs annotations were highly represented among the 'zinc finger protein' category, which is the top represented protein domain retrieved from Pfam (Fig. 5) and the second-most frequent eggNOG annotation. Ankyrin repeat motifs which were one of the top Pfam hits, are also represented as one of the top hits in both eggNOG and COGs databases. Mammalian genomes are highly enriched with these protein families.

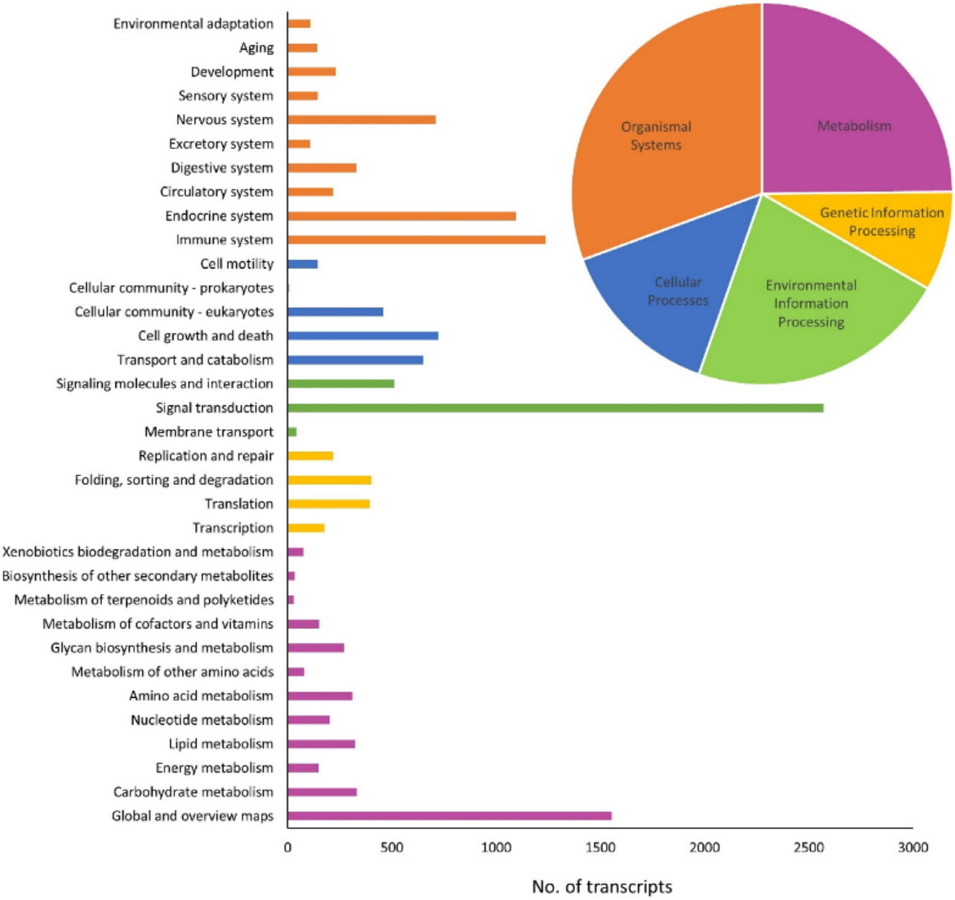

Fig. 3. Distribution of KEGG Orthology (KO) categories of high-order functions and utilities of the biological system for a *de novo Glaucomys* transcriptome assembly. Each color represents a high-order functional category.

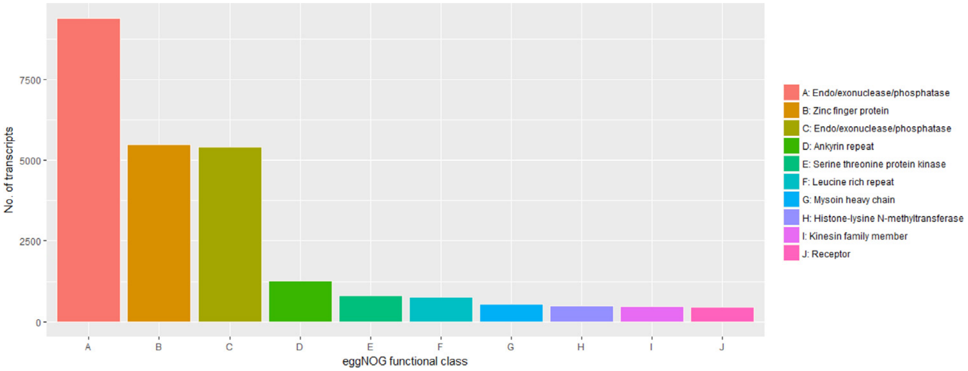

Fig. 4. Distribution of the top-10 eggNOG functional categories annotated for the *de novo Glaucomys* transcriptome assembly.

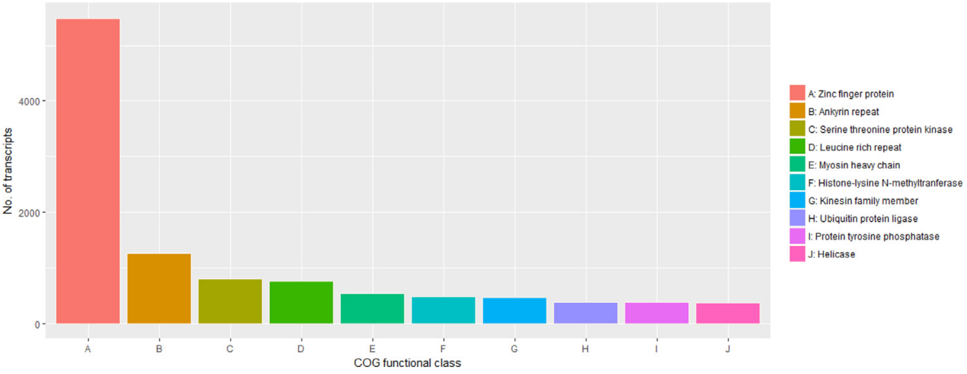

Fig. 5. Distribution of the top-10 COG functional categories annotated for the *de novo Glaucomys* transcriptome assembly.

2. Experimental Design, Materials and Methods

2.1. De novo *transcriptome assembly*

2.1.1. Sample preparation

We isolated total RNA from brain tissue of two adult *G. volans* and two adult *G. sabrinus* for RNA sequencing (RNA-*seq*), with both sexes represented for each species. *Glaucomys sabrinus* individuals were collected from near Kawartha Highlands Signature Site Park and in Algonquin Provincial Park, Ontario, Canada, and *G. volans* individuals were from Sherborne Lake and Clear Creek, Ontario, Canada. All individuals were subjects of previous incidental live-trap mortality and were obtained from a wildlife specimen repository. Live-trapping was conducted as part of a long-term *Glaucomys* population monitoring project and under an approved Animal Care Protocol Application for Wildlife and Field Work Research issued by the Trent University Animal Care Committee. Approximately 1.0 g of frozen brain tissue was removed from the hindbrain of each individual and immediately stored in RNA*later*-ICE (Invitrogen, Carlsbad, CA, USA) to prevent RNA degradation. We followed the protocol for purification of total RNA from animal tissues found in the RNeasy Mini Kit (Qiagen Inc, Hilden, Germany).

2.1.2. RNA isolation and detection

For RNA-*seq,* we extracted and submitted 404–655 ng of total RNA input with high RNA integrity number (RIN) scores (criterion RIN > 8, range 9.1–9.8) from each sample. The quantity of input RNA was measured using a Bioanalyzer 2100 RNA Nano chip (Agilent Technologies, Santa Clara, CA, USA), and the concentration was measured with a Qubit RNA HS Assay on a Qubit fluorometer (Thermo Fisher).

2.1.3. cDNA library construction and high-throughput sequencing

Library preparation was performed following the New England Biolab's NEBNext Ultra Directional RNA Library Preparation protocol. Between 404–655 ng of RNA input was enriched for poly-A mRNA and incubated for 4 min at 94 °C to fragment into a 200–300-base range. RNA fragments were converted to double-stranded cDNA and end-repaired and adenylated at 3′ with overhang A, to ligate with overhang T Illumina adapters. The cDNA library fragments were amplified under the following PCR reaction conditions: (1) denaturation at 98 °C for 10 s, (2) 10 × 98 °C for 10 s, (3) 60 °C for 30 s, (4) 72 °C for 30 s, and (5) an extension step of 72 °C for 5 min. During the amplification step, each of the samples were amplified with different barcode adapters for multiplex sequencing. After amplification, cDNA library size was measured by a Bioanalyzer 2100 DNA High Sensitivity chip (Agilent Technologies) using 1 µl of the final cDNA library. The cDNA libraries were quantified by qPCR using the Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems, Massachusetts, USA).

2.1.4. Assembly and read mapping

We cleaned raw sequence reads by removing Illumina TruSeq Universal adapters using the software Cutadapt (version 1.12) [1]. To improve the quality of SNP detection without loss of coverage [2], ambiguous nucleotides and low-quality reads (Phred quality score ≤30) were trimmed from sequence reads using TrimGalore (version 0.4.2) [3]. We used FastQC (version 0.11.5) [4], to perform simple quality control checks on raw sequence data to confirm the quality of the trimmed sequence reads.

We used Trinity (version 2.3.2) [5] to generate a *de novo* transcriptome assembled from cDNA libraries, as this program has outperformed other *de novo* assembling programs by recovering more full-length transcripts across a wide range of expression levels [5, 6]. We pooled ~200 million cDNA sequence read-pairs from all individuals to generate a consensus *Glaucomys* transcriptome. The following settings were used:

–seqType fq –max*_memory 300 G –bflyHeapSpaceMax 10 G –bflyCPU 2 –SS_lib_type RF –CPU 2 –bflyGCThreads 2*

We quality checked the *Glaucomys* transcriptome assembly by measuring the representation of RNA-*seq* reads that are aligned as proper pairs. Fragmented or short transcripts may cause only one fragment of a read pair to align to a contig, reducing the RNA-*seq* read representation of the assembly. We used bowtie2 (v2.3.0) [7] and SAMtools (v1.5) [8] to measure the proper-pair read composition of the *Glaucomys* assembly by aligning paired-end sequence reads to the transcriptome assembly, using the following settings:

–local –no-unal -q; -Sb -@ 8 -m 4 G -n -o

Read-pairs that successfully map to the assembly will be found as a proper pair, and 70–80% of paired-end reads in a high-quality Trinity assembly will be found as proper pairs [9].

2.2. Functional annotation

We functionally annotated the *Glaucomys* assembly using the Trinotate (version 3.0.2) [10] pipeline [11,12,13]. Trinotate is a comprehensive annotation program suite that incorporates a variety of well-referenced sequence, protein domain, and annotation databases to generate a robust annotation report for *de novo* transcriptome assemblies. Trinotate operates by searching for homologous protein sequences in a reference protein database, then retrieving relevant biological information from ontological databases. The Trinotate annotation report contains sequence homology hits, and gene and protein annotations.

We followed the Trinotate guidelines to leverage the latest Swiss-Prot [14] and Pfam (version 31.0) [15] protein sequence databases for sequence homologies and protein domains using the Basic Local Alignment Search Tool (BLAST+, version 2.6.0) [16] and HMMER (version 3.1b2) [17], respectively. The Pfam database contains a large collection of protein families represented by multiple sequence alignments and hidden Markov Models, derived from the UniProt knowledgebase (KB). The Swiss-Prot protein sequence database is also a component of the UniProtKB. BLASTx searches protein databases using a translated nucleotide query to identify potential protein products, and BLASTp searches protein databases using a protein query. Querying transcript and protein sequences, and protein domains, is important because UniProtKB-sourced databases produce Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] pathways, Gene Ontology (GO) [19], evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG version 3.0) [20], and Clusters of Orthologous Groups (COGs) [21] annotations. Before executing sequence analyses, we partitioned the *Glaucomys* assembly into 20 equal components using BBMap (version 37.50) [22] to avoid segmentation fault errors that arise from processing large files. We used BLASTx to query *Glaucomys* assembly transcripts in all six open reading frames (ORFs) against Swiss-Prot. Because BLASTp queries protein sequences to capture sequence homologies with protein databases, we first had to generate the most probable longest ORF peptide candidates from the *Glaucomys* assembly using TransDecoder (version 3.0.1) [23], which identifies candidate coding regions, or candidate proteins, among transcript sequences. Once predicted proteins sequences were generated, BLASTp captured protein sequence homologies with Swiss-Prot. HMMER was used to query the probable longest ORF peptide candidates against the Pfam database of protein families. The annotation output generated from searching Swiss-Prot and Pfam databases were uploaded to an SQLite database to produce a full Trinotate annotation report. Each BLAST query was conducted using the following settings:

-db uniport_sprot.pep, -num_threads 32, -max*_target_seqs 1, -outfmt 6*

The results from searches of protein sequence homologies in Swiss-Prot and Pfam databases were used to capture GO term and KEGG orthology annotations. The Gene Ontology Consortium produces common terminology to describe gene function in any organism. The GO database is linked to gene and protein databases such as Swiss-Prot and Pfam [19]. KEGG is a reference knowledgebase used to query genomes using series of molecular networks that represent cell, organism and ecosystem function. The KEGG Orthology (KO) system links genes to these molecular networks by translating genomic information into ortholog groups to define genes within a hierarchal context of metabolism, genetic information processing, environmental information processing, cellular processes and organismal systems [18]. The eggNOG database incorporates various taxonomic levels of functionally annotated Orthologous Groups (OGs) of proteins, using an algorithm that leverages previous Clusters of Orthologous Groups (COGs) methodologies [24].

To summarize the Trinotate annotation report output, we used the R package, trinotateR [25] which splits multiple hit BLAST homologies, Pfam protein domains, and GO annotations, and calculates simple annotation statistics.

Ethics Statement

The *Glaucomys* individuals sampled in this study were subjects of previous incidental live-trap mortality and were obtained from a wildlife specimen repository. Live-trapping was conducted as part of a long-term *Glaucomys* population monitoring project with the Ontario Ministry of Natural Resources and Forestry and under an approved Animal Care Protocol Application (Protocol # 24337) for Wildlife and Field Work Research issued by the Trent University Animal Care Committee in 2015.

CRediT Author Statement

Michael Brown: Data Curation (lead); Formal Analysis (lead); Investigation (lead); Methodology (equal); project administration (equal); software (lead); validation (equal), visualization (lead); writing – original draft preparation (lead); writing – review and editing (equal); **Jeff Bowman:** Conceptualization (lead); funding acquisition (lead); methodology (equal); project administration (equal); resources (lead); supervision (lead); validation (equal); writing – original draft preparation (supporting); writing – review and editing (equal); **Paul Wilson:** Conceptualization (lead); funding acquisition (lead); methodology (equal); project administration (equal); resources (lead); supervision (lead); validation (equal); writing – original draft preparation (supporting); writing – review and editing (equal).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgements

We thank Jillian Lalor and Bridget Redquest for assistance in creating the data used in this project; NSERC (Discovery Grants to JB and PJW), NSERC CREATE-Enviro program, and the Ontario Ministry of Natural Resources and Forestry for funding; and Drs. Christina Davy, Joanna Freeland, Rebekah Horn, and Rebecca Taylor for their constructive feedback and insight during original draft preparation.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107267.

References

[1] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal 17 (1) (2011) 10–12, doi:10.14806/ej.17.1.200.

[2] C. Del Fabbro, S. Scalabrin, M. Morgante, F.M. Giorgi, An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis, PLoS ONE 8 (12) (2013) DOI 10.1371/journal.pone.0085024, doi:10.1371/journal.pone.0085024.

[3] F. Krueger, A Wrapper Around Cutadapt and FastQC to Consistently Apply Adapter and Quality Trimming to FastQ files, With Extra Functionality For RRBS Data, 2016. https://github.com/FelixKrueger/TrimGalore. Accessed May 2, 2016.

[4] S. Andrews, FastQC: a Quality Control Tool For High Throughput Sequence Data, 2010 http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed March 15, 2016.

[5] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (7) (2011) 644–652, doi:10.1038/nbt.1883.

[6] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, D. Philip, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, *De novo* transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity, Nat. Protoc. 8 (8) (2014) 1–43, doi:10.1038/nprot.2013.084.

[7] B. Langmead, S. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359, doi:10.1038/nmeth.1923.

[8] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079, doi:10.1093/bioinformatics/btp352.

[9] B. Haas, RA Seq Read Representation by Trinity Assembly, 2018. https://github.com/trinityrnaseq/trinityrnaseq/wiki/RNA-Seq-Read-Representation-by-Trinity-Assembly. Accessed March 22, 2017.

[10] B. Haas, Trinotate: Transcriptome Functional Annotation and Analysis, 2015 https://trinotate.github.io/. Accessed February 9, 2016.

[11] A. Sayadi, E. Immonen, H. Bayram, G. Arnqvist, The *de novo* transcriptome and its functional annotation in the seed beetle *callosobruchus maculatus*, PLoS ONE 11 (7) (2016), doi:10.1371/journal.pone.0158565.

[12] S.L. Fernandez-Valverde, A.D. Calcino, B.M. Degnan, Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*, BMC Genomics 16 (2015) 387, doi:10.1186/s12864-015-1588-z.

[13] N. Ghaffari, A. Sanchez-Flores, R. Doan, K.D. Garcia-Orozco, P.L. Chen, A. Ochoa-Leyva, A.A. Lopez-Zavala, J.S. Carrasco, C. Hong, L.G. Brieba, E. Rudiño-Piñera, P.D. Blood, J.E. Sawyer, C.D. Johnson, S.V. Dindot, R.R. Sotelo-Mundo, M.F. Criscitiello, Novel transcriptome assembly and improved annotation of the whiteleg shrimp (*Litopenaeus vannamei*), a dominant crustacean in global seafood mariculture, Sci. Rep. 4 (2014) 7081, doi:10.1038/srep07081.

[14] The Uniprot Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169, doi:10.1093/nar/gkw1099.

[15] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The pfam protein families database, Nucleic Acids Res. 40 (1) (2012) D290–D301, doi:10.1093/nar/gkr1065.

[16] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, doi:10.1016/S0022-2836(05)80360-2.

[17] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37, doi:10.1093/nar/gkr367.

[18] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular datasets, Nucleic Acids Res. 40 (2012) D109–D114, doi:10.1093/nar/gkr988.

[19] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matesem, J.E. Richardson, M. Ringwalk, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, doi:10.1038/75556.

[20] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L.J. Jensen, C. von Mering, P. Bork, eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges, Nucleic Acids Res. 40 (2012) D284–D289, doi:10.1093/nar/gkr1060.

[21] R.L. Tatusov, M.Y. Galperin, D.A. Natale, E.V. Koonin, The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Res. 28 (1) (2000) 33–36, doi:10.1093/nar/28.1.33.

[22] B. Bushnell, BBMap Short Read aligner, and Other Bioinformatic Tools, 2015 https://sourceforge.net/projects/bbmap/. Accessed June 28, 2016.

[23] B. Haas, TransDecoder (Find Coding Regions Within Transcripts), 2018. https://github.com/TransDecoder/TransDecoder/wiki. Accessed July 17, 2016.

[24] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M.C. Walter, T. Rattei, D.R. Mende, S. Sunagawa, M. Kuhn, L.J. Jensen, C. von Mering, P. Bork, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, Nucleic Acids Res. 44 (2017) D286–D293, doi:10.1093/nar/gkv1248.

[25] D.M. Bryant, K. Johnson, T. DiTommaso, T. Tickle, M.B. Couger, D. Payzin-Dogru, T.J. Lee, N.D. Leigh, T.H. Kuo, F.G. Davis, J. Bateman, S. Bryant, A.R. Guzikowski, S.L. Tsai, S. Coyne, W.W. Ye, R.M. Freeman Jr., L. Peshkin, C.J. Tabin, A. Regev, B.J. Haas, J.L. Whited, A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors, Cell Rep. 18 (3) (2017) 762–776, doi:10.1016/j.celrep.2016.12.063.