Data in brief 27 (2019) 104577

Contents lists available at ScienceDirect

Data in brief

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

Data Article

The mixed liver and heart transcriptome dataset of the New Zealand brushtail possum, *Trichosurus vulpecula*

Daniel J. White [a, b, *], Katherine Trought [a], Brian Hopkins [a]

[a] *Manaaki Whenua Landcare Research, Lincoln, 7608, New Zealand*
[b] *School of Biological Sciences, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*

ARTICLE INFO

Article history:
Received 6 May 2019
Received in revised form 8 September 2019
Accepted 20 September 2019
Available online 15 October 2019

Keywords:
Vertebrate
Invasive pest
Brushtail possum
New Zealand
Transcriptome
Annotated

ABSTRACT

New Zealand suffers greatly from invasive mammal predators including rats, stoats, feral cats and possums all of which not only damage or prey on New Zealand's unique terrestrial biodiversity, but also have huge impact on NZ's economy as many of these pests act as vectors of disease to farm and game animals. As such, the NZ government has invested nearly $90 m to support an ambitious plan to make the country predator free by 2050. Although there are adequate means to control invasive predator populations, it is widely agreed that current technologies are not sufficient for total eradication and that improved technologies are required. The Achilles Heel approach is one such developmental technology that attempts to exploit variation in the genes of target species that are vital to key physiological or cellular pathways within the body, such that interference with these genes will cause a species-specific death without the harmful effects on the environment and non-targets species that the current suite of control agents engender. Interference could either be through species-specific gene knock-down using such agents as siRNA and/or the use of species-selective chemical toxicants specifically developed against these targets. To assist with identifying species-specific gene targets in the New Zealand brushtail possum (*Trichosurus vulpecula*)

[*] Corresponding author. School of Biological Sciences, University of Western Australia, 35 Stirling Highway, Crawley, WA, 6009, Australia.
E-mail address: daniel.white@uwa.edu.au (D.J. White).

https://doi.org/10.1016/j.dib.2019.104577

we have assembled and annotated a possum mixed heart and liver transcriptome.

Specifications Table

Subject area	Biology
More specific subject area	Vertebrate transcriptomics
Type of data	Transcriptome sequences and associated annotations
How data was acquired	Total RNA was isolated from liver tissue and heart tissue of the NZ brushtail possum. 2 × 100 bp paired-end RNAseq reads of RNA were sequenced on a HiSeq 2500 system.
Data format	Raw data in FASTQ and transcriptome sequences in FASTA
Experimental factors	5 mm cubed of heart and liver tissue was sampled from one brushtail possum immediately after euthanasia and placed in DNA/RNA shield, stored at −80 °C.
Experimental features	*De novo* assembly, quality checking and functional annotation of one adult male NZ brushtail possum combined transcriptome using liver and heart tissue
Data source location	Lincoln, NZ
Data accessibility	This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GHKB00000000. The version described in this paper is the first version, GHKB01000000. The overall BioProject ID is PRJNA525264 and the BioSample accessions are SAMN11044341 and SAMN11044342

Value of Data

- First *de novo* transcriptome of the NZ brushtail possum to be published, giving public access to the coding regions of an invasive vertebrate that has substantial economic and biodiversity impact
- Data presented here will directly benefit researchers and managers interested in developing species-specific genetic tools for NZ brushtail possum management such as gene knock-down and knock-out, as they provide a valuable gene-mining resource [1]
- Having full sequence data for the genes of the NZ brushtail possum will allow reliable prediction of the proteome and accurate development of protein-targeted and gene-targeted toxins
- These data will help studies aimed at increasing the understanding of brushtail possum traits that may make it a successful pest species such as development of tolerance to toxins and diverse climates
- The NZ brushtail possum transcriptome could be useful reference sequence for the assembly of brushtail possum whole genomes
- More broadly, the annotated NZ brushtail possum transcriptome could contribute to studies focused on the discovery of genes involved in marsupial-specific traits, as well as identifying functional variation in other possum species with more critical conservation status

1. Data

Data presented in this article include raw reads for two paired-end libraries generated from mRNA which was isolated from heart and liver tissue of a male New Zealand brushtail possum, *Trichosurus vulpecula*. The BioSample accession numbers for the heart and liver paired-end libraries are SAMN11044341 and SAMN11044342, respectively. These raw reads were quality checked and filtered and then *de novo* assembled into a raw transcriptome. This raw transcriptome was improved by merging open reading frames and predicting protein domains, referred to as the refined transcriptome. This refined transcriptome has been deposited at Manaaki Whenua Landcare Research's data repository, DataStore, with doi https://doi.org/10.7931/nzcx-1176. Any non-mammalian transcripts were screened out before this final clean transcriptome was functionally annotated. The clean transcriptome represents the New Zealand brushtail possum combined heart and liver transcriptome and has been deposited as a Transcriptome Shotgun Assembly project at DDBJ/EMBL/GenBank under the accession GHKB00000000. The overall BioProject ID is PRJNA525264.

2. Experimental design, materials and methods

2.1. Tissue samples

Tissue from the heart and liver were obtained from an adult male brushtail possum (*Trichosurus vulpecula*) immediately following euthanasia. This animal was euthanised for a different project at Manaaki Whenua Landcare Research, following all appropriate animal ethics guidelines. Small segments of tissue (5 mm cubed) were placed into 1.7 ml microcentrifuge tubes containing 1 ml DNA/RNA shield ™ (Zymo Research, CA, USA) and stored at −80 °C until required for RNA extraction.

2.2. RNA isolation

RNA was extracted from the tissue samples using the Easy-spin Total RNA Extraction Kit (iNtRON Biotechnology Inc, Gyeonggi-do, South Korea) following a modified version of the manufacturer's instructions. Briefly, 50 mg of tissue was added to 700 μl lysis buffer. The sample was homogenised in a bead mill (TissueLyser II, Qiagen Inc, CA USA) with a 3-mm stainless steel ball at a frequency of 30/second for 60 seconds. Samples were processed from this point onward following the manufacturer's instructions until the elution step, where an additional wash step of 700 μl 80% ethanol was placed in the spin column to remove any additional salts. The spin column was dried for 2 minutes at 16,000×*g* and the RNA eluted in 50 μl ultra-pure water. The eluted RNA was treated with the RNase-free Dnase set (Qiagen Inc, CA USA) using the protocol outlined in Appendix E of the RNeasy mini handbook (DNase Digestion of RNA before RNA Cleanup (Qiagen Inc, CA USA)). The purified RNA was eluted in a final volume of 50 μl ultrapure water. RNA quantity and quality were assessed using the LabChip GX Touch HT (PerkinElmer Inc, MA, USA) and 2.5–7.5 μg total RNA was loaded into RNAstable tubes (Biomatrica Inc, San Diego, USA). The RNA was stabilised and dried as per the manufacturer's instructions and shipped at ambient temperature for library preparation and transcriptome sequencing to Macrogen, South Korea. 100 base-pair (bp) strand specific paired-end mRNA libraries were made for each tissue type and sequenced on one lane of an Illumina HiSeq 2500 platform.

2.3. Bioinformatic analysis

Raw paired-end reads were quality checked using FASTQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

Adapters were trimmed from the 3' end of reads using CUTADAPT v1.16 [2], then quality filtered in SOLEXAQA++ v3.1.7 based on base-pair quality (p set to 0.01) and minimum read length of 25 bp [3]. After filtering any reads lacking their pair were removed, so only properly paired reads were included

Table 1
NZ brushtail possum raw transcriptome assembly statistics.

No. libraries	2 – heart tissue, liver tissue
Total no. bases assembled	201,171,876
No. transcripts	203,458
N50	2451 nt
Ex90N50	3055 nt
Ex83N50[a]	3381 nt
Shortest contig	176 nt
Longest contig	18,175 nt
No. transcripts in transcriptome after merging predicted ORFs, and pfam domain and peptide sequence homology	66,610

[a] Ex83N50 is the optimal ExN50 value.

Table 2
NZ brushtail possum clean transcriptome assembly statistics.

No. transcripts	50,484
Total length of transcripts	62,738,637 nt
Mean length	1242 nt
N50	1737 nt
Shortest contig	258 nt
Longest contig	15,819 nt

in the assembly. As no reference genome existed for the brushtail possum at the time of assembly, a *de novo* assembly was done, using reads from both the heart and liver tissue paired-end libraries in TRINITY v2.8.4 with the SS_lib_type flag set to RF [4]. Assessment of the assembly was made by calculating N50, read content of the assembly and ExN50, which is a measurement of N50 at desired percentages of total expressed transcripts, using TRINITY [5]. N50 values can be driven downwards by overabundance of small incomplete transcripts beyond the optimal ExN50 value, and E90N50 is a commonly chosen value for quality assessment. Assembly statistics are provided in Table 1.

TRANSDECODER v5.5.0 [5] was run to select the best transcripts from the raw transcriptome. First, open reading frames (ORFs) from the same gene were merged into the longest representative open reading frame. This set of transcripts were then processed by gene prediction software and screened against the SwissProt database using the blastp option in BLAST v2.6.0 (evalue threshold set to 1e-5) and the Pfam database using HMMER v3.1 with default parameters to detect protein domains [6,7]. This generated a refined transcriptome which was purged of possible non-mammal contaminants by screening for homology to the mammal subset of the non-redundant nr database at Genbank using blastx-fast (which uses a word length of 6) in Blast2Go Pro, with the evalue threshold set to 1e-3 [8,9]. Any transcripts that did not find a match were excluded from further analysis, and the retained transcripts were considered the final, clean NZ brushtail possum transcriptome (see Table 2). Due to less information available in databases for marsupial mammals, search criteria were left at default values to prevent excluding potential possum proteins.

For annotation and characterisation of the clean transcriptome, a new project was set up in Blast2Go Pro and the final sub-set of transcripts were screened once more against the mammalian subset of the nr database at Genbank. This time regular blastx (word length of 3) and an evalue threshold of 1e-3, as well as running a functional analysis using InterProScan, were run in Blast2Go Pro [8,9]. InterProScan predicts domains and important sites based on numerous databases and prediction algorithms and classifies proteins into families. CDD and Pfam resources were used to search for domains [10,11],

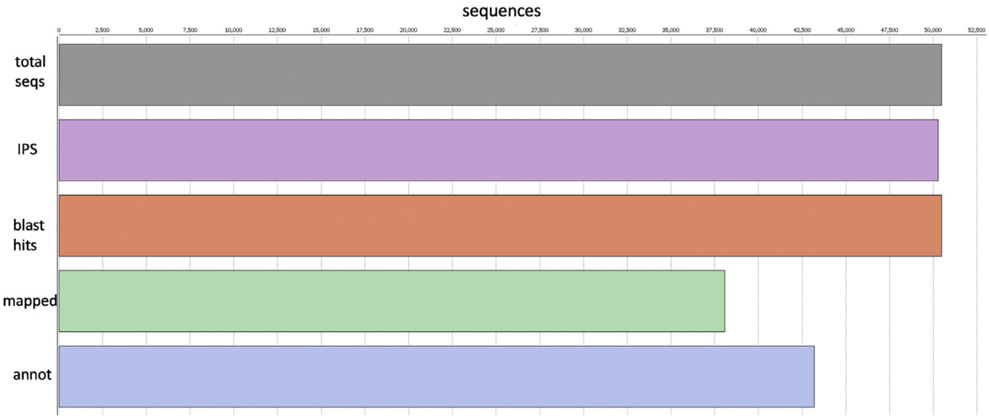

Fig. 1. A summary of blastx, InterProScan, mapping and annotation results for 50,484 contigs in the final clean NZ brushtail possum transcriptome.

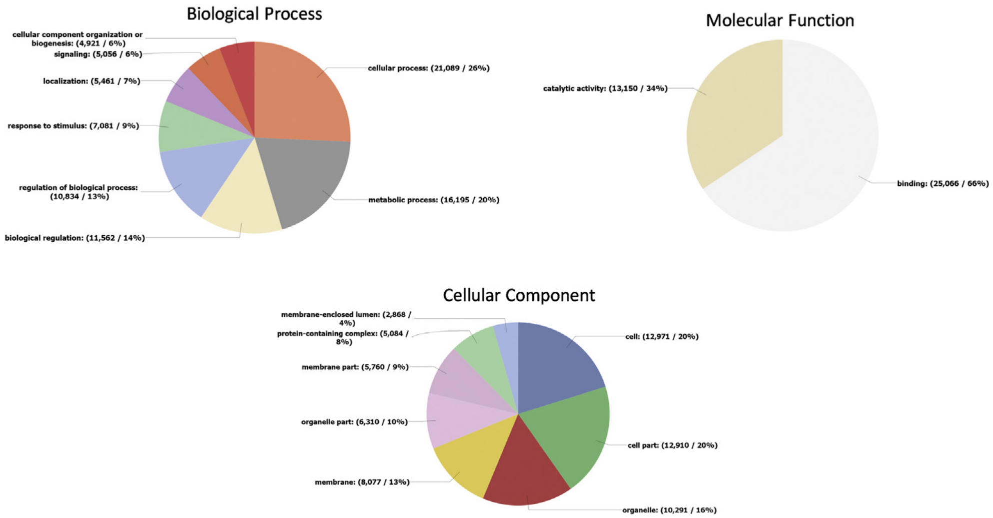

Fig. 2. Percentage of annotated transcripts with level 2 gene ontology terms for biological process, cellular component and molecular function ontologies.

Panther was used to identify families and associate proteins with function [12], Prosite was used to identify families based on sites and patterns [13] and Superfamily was used to predict functional domains [14]. Of the 50,484 transcripts in the clean transcriptome 50,477 (100%) had blastx hits and 47,150 (93%) had IPS hits (Fig. 1). A summary of transcript function based on gene ontology level 2 terms for biological process, cellular component and molecular function is provided in Fig. 2.

Acknowledgements

This work was funded by MBIE Endeavour Smart Idea grant C09X1615 and supported by the BioHeritage National Science Challenge Project 2.3 High-tech solutions to invasive mammal pests. We are grateful to the staff at MWLR's Animal Facility, Lincoln for access to samples. We acknowledge the contribution of the NeSI high-performance computing facilities and the staff at the Centre for eResearch at the University of Auckland. We thank Peter Maxwell and Dinindu Senanayake (NeSI), Carlos Menor (BioBam) and Jihye Ham (Macrogen) for technical assistance.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P.K. Dearden, N.J. Gemmell, O.R. Mercier, P.J. Lester, M.J. Scott, R.D. Newcomb, T.R. Buckley, J.M.E. Jacobs, S.G. Goldson, D.R. Penman, The potential for the use of gene drives for pest control in New Zealand: a perspective, J. R. Soc. N. Z. 48 (2018) 225–244.

[2] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal 17 (2011) 10–12.

[3] M.P. Cox, D.A. Peterson, P.J. Biggs, SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data, BMC Bioinf. 11 (2010).

[4] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q.D. Zeng, Z.H. 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 (2011) 644. U130.

[5] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, 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 using the Trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512.

[6] S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[7] HMMER, Available from: http://hmmer.org. (Accessed 10 April 2019).

[8] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.

[9] S. Gotz, J.M. Garcia-Gomez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talon, J. Dopazo, A. Conesa, High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res. 36 (2008) 3420–3435.

[10] A. Marchler-Bauer, Y. Bo, L.Y. Han, J.E. He, C.J. Lanczycki, S.N. Lu, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z.X. Wang, R.A. Yamashita, D.C. Zhang, C.J. Zheng, L.Y. Geer, S. H. Bryant, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures, Nucleic Acids Res. 45 (2017) D200–D203.

[11] 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 (2012) D290–D301.

[12] H.Y. Mi, A. Muruganujan, D. Ebert, X.S. Huang, P.D. Thomas, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, Nucleic Acids Res. 47 (2019) D419–D426.

[13] C.J.A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, P. Bucher, PROSITE: a documented database using patterns and profiles as motif descriptors, Briefings Bioinf. 3 (2002) 265–275.

[14] J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, J. Mol. Biol. 313 (2001) 903–919.