Dataset for sequencing and de novo assembly of the European endangered white-clawed crayfish (Austropotamobius pallipes) abdominal muscle transcriptome.

Frederic Grandjean, Han Ming Gan, Bouziane Moumen, Isabelle Giraud, Skander Hatira, Richard Cordaux, Christopher Austin

▶ **To cite this version:**

Frederic Grandjean, Han Ming Gan, Bouziane Moumen, Isabelle Giraud, Skander Hatira, et al.. Dataset for sequencing and de novo assembly of the European endangered white-clawed crayfish (Austropotamobius pallipes) abdominal muscle transcriptome.. Data in Brief, 2020, 29, pp.105166. 10.1016/j.dib.2020.105166 . hal-03021751

HAL Id: hal-03021751
https://hal.science/hal-03021751

Submitted on 25 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License

Data in brief 29 (2020) 105166

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Dataset for sequencing and *de novo* assembly of the European endangered white-clawed crayfish (*Austropotamobius pallipes*) abdominal muscle transcriptome

Frederic Grandjean [a], Han Ming Gan [b, c], Bouziane Moumen [a], Isabelle Giraud [a], Skander Hatira [a], Richard Cordaux [a], Christopher M. Austin [b, c, *]

[a] *Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Unité Mixte de Recherche 7267 Centre National de la Recherche Scientifique, Université de Poitiers, Poitiers, France*
[b] *Deakin Genomics Centre, Deakin University, Geelong, 3220, Victoria, Australia*
[c] *Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, 3220, Victoria, Australia*

A R T I C L E I N F O

Article history:
Received 22 December 2019
Accepted 15 January 2020
Available online 27 January 2020

Keywords:
Austropotamobius pallipes
Transcriptome
Muscle
Illumina
Crayfish

A B S T R A C T

The white-clawed crayfish (*Austropotamobius pallipes*) is an endangered species in Europe with limited genomic information. Despite its conservation status there is no transcriptomic data available for *A. pallipes* in public databases. The data here represents the first transcriptome profile of the white-clawed crayfish generated using Illumina stranded RNA sequencing. Pair-end reads were assembled *de novo* with three separate transcriptome assemblers (Trinity, RNABloom, and RNASpades) followed by transcript assembly reduction and gene reconstruction using the EvidentialGene pipeline. The transcriptome was functionally annotated using InterProScan and genes coding for carbohydrate-active enzymes were identified through the dbCAN2 server. Raw fastq reads and the final version of the transcriptome assembly have been deposited in the NCBI-SRA (SRR10549898) and NCBI-TSA (GICG01) databases.

© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

* Corresponding author. Deakin Genomics Centre, Deakin University, Geelong, 3220, Victoria, Australia.
E-mail address: c.austin@deakin.edu.au (C.M. Austin).

https://doi.org/10.1016/j.dib.2020.105166
2352-3409/© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Specifications Table

Subject	Biology
Specific subject area	Transcriptomics
Type of data	Sequencing raw reads, assembly, Table, Figure,
How data were acquired	Illumina HiSeq 2500
Data format	Raw Reads (fastq), Assembly (fasta)
Parameters for data collection	rRNA-depleted RNA from the abdominal muscle tissue of an adult specimen was used for library preparation and sequencing.
Description of data collection	Total RNA extraction was performed using QIAgen RNeasy mini kit (Qiagen) followed by rRNA removal using the Ribo-Zero rRNA depletion kit (Illumina). The rRNA-depleted RNA sample was subsequently processed with the Illumina TruSeq Stranded Total RNA kit (Illumina) following the manufacturer's instructions. Paired-end sequencing of the constructed library was performed on a HiSeq 2500 (2 × 150 bp run configuration).
Data source location	0°23′35.7691″ E; 46°56′10.6771″ N
Data accessibility	Raw data and final assembled contigs were deposited in the NCBI database under the Bioproject PRJNA592270. Additional files such as BUSCO analysis output, Interproscan annotation, transcript abundance estimation and the initial transcriptome assemblies from three separate transcriptome assemblers are available in the Zenodo database (http://doi.org/10.5281/zenodo.3581499)

Value of the Data

- First transcriptome dataset for the endangered white-clawed crayfish.
- High transcriptome completeness will enable its use in phylotranscriptomic studies of decapod crustaceans.
- The data will facilitate genetic management for the conservation of remaining white-clawed crayfish populations.
- The data adds to the limited genomic available for this and related species [1,2].

1. Data description

Stranded RNA sequencing was performed to generate the first *de novo* transcriptome assembly from the endangered white-clawed crayfish (*Austropotamobius pallipes*). Sequencing on the HiSeq 2500 generated a total of 58.88 million raw paired-end 150-bp reads. After Illumina adapter and quality trimming, 54.77 million paired-end reads longer than 50 bp were retained and subsequently used for *de novo* assembly. The final non-redundant and contaminant-filtered assembly consists of 79,886 contigs and exhibits a BUSCO transcriptome completeness of 88.1% (Table 1). A total of 18,026 and 11,809 transcripts coded for proteins containing InterProScan protein signatures and Gene Ontology (GO) terms, respectively. A majority of the functionally annotated transcripts with high abundance were associated with muscle function (Table 2). A total of 68 transcripts code for proteins with putative glycoside hydrolase activity with GH18 family (chitinase) having the highest transcript representation (Fig. 1A). The longest translated coding sequence from this assembly consists of 8570 amino acid residues (transcript length of 26,931 bp) and exhibits protein signatures that are commonly found in the titin-like giant proteins family (Fig. 1B).

2. Experimental design, materials, and methods

2.1. Crayfish, tissue sampling and RNA extraction

An adult white-clawed crayfish was collected from a wild population. Approximately 30 mg of abdominal tissue was aseptically dissected and immediately lysed in Buffer RLT provided in the RNeasy mini kit (Qiagen). Total RNA extraction was carried out as per the manufacturer's instructions.

Table 1
Transcriptome assembly statistics.

Transcriptome Assembly	
Assembled length	54,163,745 bp
N_{50} length	963 bp
Number of contigs	79,886
GC %	43.72%
BUSCO Completeness (Arthropoda odb9)	
Complete BUSCO	88.1%
Complete and single-copy BUSCO	81.0%
Complete and duplicated BUSCO	7.1%
Fragmented BUSCO	8.5%
Missing BUSCO	3.4%
Total BUSCO groups searched	1066

2.2. rRNA depletion, stranded RNA library construction and sequencing

Approximately 1 μg of total RNA as measured by the Qubit 3.0 fluorometer (Invitrogen, USA) was used as the input for Ribo-Zero rRNA removal kit (Illumina, San Diego, CA). The rRNA-depleted RNA was subsequently processed using the TruSeq Stranded RNA library preparation kit (Illumina, San Diego, CA). Sequencing of the stranded RNA library was performed on an Illumina HiSeq 2500 sequencing platform using the run configuration of 2 × 150 bp.

2.3. de novo *assembly and transcriptome completeness assessment*

Sequencing quality and yield before and after trimming was assessed using FASTQC v0.10.1 [3]. Data were filtered using fastp v0.20.0 [4] that performed adapter sequence and quality trimming. Only reads longer than 50 bp after trimming were used for assembly and transcript abundance quantification. The trimmed paired-end reads were assembled separately with Trinity v2.8.5 [5], RNABloom v1.1.0 [6]and rnaSPAdes v3.13.0 [7] using the default setting. These primary assemblies were merged and used as the input for EvidentialGene v2013.03.11 (default setting) that performed transcript assembly reduction with coding-sequence classifier [8]. The reduced assembly was submitted to NCBI Transcriptome shotgun assembly portal to screen for any residual adapter sequence contamination as well as human-derived contigs. After the removal of the remaining contaminant sequences, the transcriptome completeness was assessed using BUSCO v3 to obtain the percentage of single-copy orthologs represented in the Arthropod odb9 dataset [9].

2.4. Transcript annotation and abundance quantification

The translated coding sequences produced from EvidentialGene were functionally annotated using InterProScan v5.35−74.0 [10]. These sequences were also uploaded to the dbCAN2 meta server for

Table 2
Abundance and classification of the top 10 most highly expressed transcripts with functional annotation.

Transcript ID	Transcripts Per Million	Assembled transcript length	Putative function
evgE0.L.342066	154560	761	Winged-helix DNA binding protein
evgE0.L.188382	13074.6	236	transposase
evgE0.L.252356	6811.84	439	Thioredoxin-like superfamily
evgE0.U.533026	5834.55	1255	Adenylate cyclase
evgE0.L.106899	4535.13	1612	EF-hand calcium binding protein
evgE0.L.262753	4453.76	965	Insect cuticle protein
evgE0.L.487767	4373.56	1415	Troponin domain superfamily
evgvelvLoc19648ct1	4228.39	858	Tensin phosphatase
evgE0.U.552054	3405.58	10112	Myosin
evgE1.L.17840	3147.58	1499	Glyceraldehyde-3-phosphate dehydrogenase (GADPH)

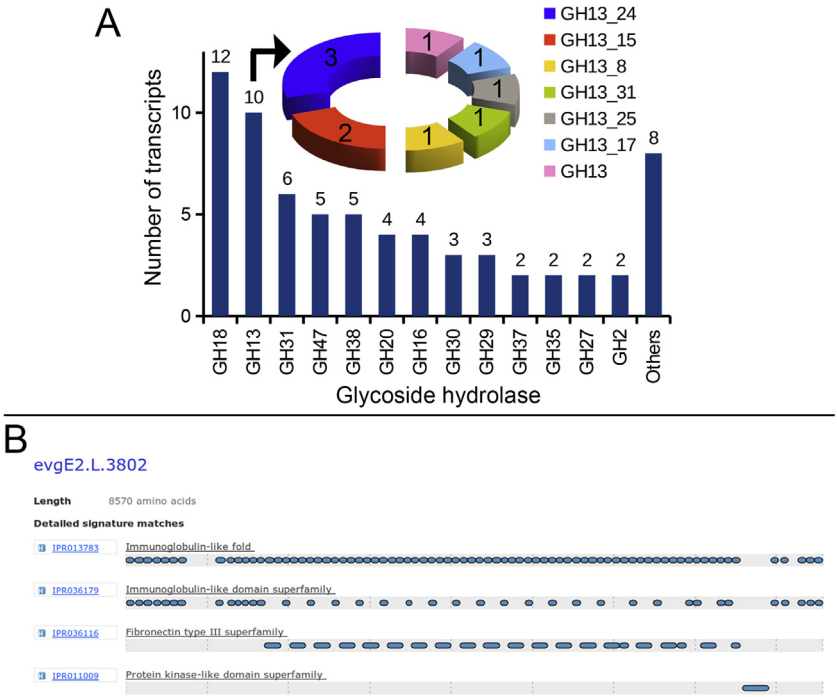

Fig. 1. (A) Distribution of transcripts with CAZy annotation (B) Visualization of protein domains identified on the putative *Austropotamobius pallipes* titin protein.

automated carbohydrate-active enzyme annotation based on the dbCAN HMMdb v8.0 database (E-Value < 1e-15, coverage > 0.35) [11]. For the quantification of transcript abundance, the final version of the transcriptome assembly was first indexed with Kallisto v0.46.0 [12] followed by pseudo-alignment of the trimmed paired-end reads with the "—bias" option activated to correct for sequence-specific, fragment-GC and positional biases.

Acknowledgments

This work was funded by Agence Nationale de la Recherche Grant ANR-15-CE32-0006 (CytoSexDet) to R.C. and T.R., the 2015—2020 State-Region Planning Contract and European Regional Development Fund, and intramural funds from the CNRS and the University of Poitiers.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2020.105166.

References

[1] F. Grandjean, M.H. Tan, H.M. Gan, Y.P. Lee, T. Kawai, R.J. Distefano, M. Blaha, A.J. Roles, C.M. Austin, Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish, Zool. Scripta 46 (2017) 718—728.

[2] F. Grandjean, M.H. Tan, H.Y. Gan, H.M. Gan, C.M. Austin, The complete mitogenome of the endangered white-clawed freshwater crayfish *Austropotamobius pallipes* (Lereboullet, 1858) (Crustacea: Decapoda: Astacidae), Mitochondrial DNA Part A 27 (2016) 3329–3330.
[3] S. Andrews, FastQC: a Quality Control Tool for High Throughput Sequence Data, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010.
[4] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.
[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, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (2011) 644.
[6] K.M. Nip, R. Chiu, C. Yang, J. Chu, H. Mohamadi, R.L. Warren, I. Birol, RNA-bloom Provides Lightweight Reference-free Transcriptome Assembly for Single Cells, bioRxiv, 2019, p. 701607.
[7] E. Bushmanova, D. Antipov, A. Lapidus, A.D. Prjibelski, rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data, GigaScience 8 (2019) giz100.
[8] D.G. Gilbert, Genes of the pig, Sus scrofa, reconstructed with EvidentialGene, PeerJ 7 (2019) e6374.
[9] R.M. Waterhouse, M. Seppey, F.A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E.V. Kriventseva, E.M. Zdobnov, BUSCO applications from quality assessments to gene prediction and phylogenomics, Mol. Biol. Evol. 35 (2017) 543–548.
[10] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (2014) 1236–1240.
[11] H. Zhang, T. Yohe, L. Huang, S. Entwistle, P. Wu, Z. Yang, P.K. Busk, Y. Xu, Y. Yin, dbCAN2: a meta server for automated carbohydrate-active enzyme annotation, Nucleic Acids Res. 46 (2018) W95–W101.
[12] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (2016) 525.