MOLECULAR ECOLOGY
RESOURCES

De novo assembly transcriptome for the rostrum dace (Leuciscus burdigalensis, Cyprinidae: fish) naturally infected by a copepod ectoparasite.

Journal:	*Molecular Ecology Resources*
Manuscript ID:	MER-14-0526.R1
Manuscript Type:	Genomic Resources Note
Date Submitted by the Author:	n/a
Complete List of Authors:	Rey, Olivier; Station d'Ecologie Experimentale du CNRS, UMR 2936, CNRS Loot, Géraldine; Université de Toulouse, UPS, UMR-5174 (EDB), Bouchez, Olivier; GeT-PlaGe, Genotoul, INRA Auzeville F31326, ; INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F31326 , Blanchet, Simon; 1 Station d'Ecologie Expérimentale du CNRS à Moulis, USR 2936, ; Université de Toulouse, UPS, UMR-5174 (EDB),
Keywords:	Ecological Genetics, Fish, Host Parasite Interactions, Invasive Species, Parasitology, Transcriptomics

Title

***De novo* assembly transcriptome for the rostrum dace (*Leuciscus burdigalensis*, Cyprinidae: fish) naturally infected by a copepod ectoparasite.**

Authors

Olivier Rey[1*], Géraldine Loot[1,2], Olivier Bouchez[3, 4], Simon Blanchet[1,2*]

Affiliations

[1] Station d'Ecologie Expérimentale du CNRS à Moulis, USR 2936, 09 200 Moulis, France.
[2] Université de Toulouse, UPS, UMR-5174 (EDB), 118 route de Narbonne, 31062 Toulouse, Cedex 9, France
[3] GeT-PlaGe, Genotoul, INRA Auzeville F31326, Castanet-tolosan, FRANCE
[4] INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F31326 Castanet-Tolosan, France
[5] CNRS, UPS, ENFA, Evolution & Diversité Biologique (EDB) UMR 5174, 118 Route de Narbonne, 31062 Toulouse, Cedex 9, France

*Corresponding authors (olivier.rey.1@gmail.com / simon.blanchet@EcoEx-Moulis.cnrs.fr)

ABSTRACT

The emergence of pathogens represents substantial threats to public health, livestock, domesticated animals, and biodiversity. How wild populations respond to emerging pathogens has generated a lot of interest in the last two decades. With the recent advent of high-throughput sequencing technologies it is now possible to develop large transcriptomic resources for non-model organisms, hence allowing new research avenues on the immune responses of hosts from a large taxonomic spectra. We here focused on a wild population of the rostrum dace (*Leuciscus burdigalensis*) that is infected by *Tracheliastes polycolpus,* an emerging freshwater ectoparasite copepod. We used next generation Illumina sequencing technology to sequence the transcriptome of eight *L. burdigalensis* adult individuals collected *in natura* from the same sampling site. Four individuals were non-infected and four individuals were infected by *T. polycolpus*. We specifically focused on the spleen, the head kidney and epithelial cells and mucus from the fins, three tissues known to be involved in the immune response of fish. We used the *Trinity* methodology to reconstruct a *de novo* full-length transcriptome for *L. burdigalensis*. The resulting transcriptome will serve as an important broad-scale genomic resource for further studying the response of local population of *L. burdigalensis* to *T. polycolpus* pressures.

Introduction

The emergence of pathogens represents substantial threats to public health, livestock, domesticated animals, and biodiversity (Daszak *et al.*, 2000; Woolhouse, 2008). How wild populations respond to emerging pathogens has generated a lot of interest in the last two decades (Acevedo-Whitehouse & Cunningham, 2006; Bonneaud *et al.*, 2012; Gilbert *et al.*, 2013). So far, most of the molecular mechanisms and pathways underlying the host immune response have been identified in model species for which large genomic resources are available (Jenner & Young, 2005). In fish species, most of knowledge on molecular pathways underlying immunity response to pathogens stems from studies conducted on model organisms (e.g. *Danio rerio*) and/or economically relevant species in aquaculture and/or fisheries (e.g. salmonids, carps; Uribe *et al.*, 2011; Zhu *et al.*, 2013). With the recent advent of high-throughput sequencing technologies it is now possible to develop large transcriptomic resources even for non-model organisms (Metzker, 2010; Ekblom & Galindo, 2011), hence allowing new research avenues on the immune responses of hosts from a large taxonomic spectra (Dheilly *et al.*, 2014).

We here focused on the rostrum dace (Leuciscus *burdigalensis*) – *Tracheliastes polycolpus* host-parasite system (Loot *et al.*, 2004). The rostrum dace (also called beaked dace) belongs to the *Leuciscus* species complex (Cyprinidae: fish) and is endemic to streams and rivers from South-Western France and constitutes the principal host for *Tracheliastes polycolpus* (Loot *et al.*, 2004). *Tracheliastes polycolpus* is a copepod ectoparasite that has been introduced in Western Europe (e.g. France, Spain, United Kingdom) in the 1920's (Aubrook & Fryer, 1965; Tuffery, 1967; Fryer, 1982). Only females are parasitic and they attach to the fins of the individual host and feed on the mucus and epithelial cells, hence causing severe infections and lesions, both contributing to a reduction of host fitness (Loot *et al.*, 2004; Blanchet *et al.*, 2009a; Blanchet *et al.*, 2009b). How do local *L. burdigalensis*

populations respond to this new ectoparasite species remains an intriguing and challenging question, which has implications for predicting the evolutionary potential of the host populations.

We used next generation Illumina sequencing technology to sequence the transcriptome of eight *L. burdigalensis* adult individuals directly collected *in natura*. Four individuals were non-infected and four individuals were infected by *T. polycolpus*. We focused specifically on two tissues known to be involved in the immune response of fish -i.e. the spleen and the head kidney (Zapata *et al.*, 2006; Salinas *et al.*, 2011; Uribe *et al.*, 2011)-, as well as on epithelial cells and mucus from the fin on which parasites were anchored. We used the *Trinity* methodology to reconstruct a *de novo* full-length transcriptome for *L. burdigalensis*. The resulting transcriptome will serve as an important broad-scale genomic resource for further studying the response of local population of *L. burdigalensis* to *T. polycolpus* pressures. This transcriptome may also provide a huge genomic repertoire for further studies dealing with related cyprinid fish species (notably closely related species from the *Leuciscus* complex), the broadest fish family in terms of species, with special emphasis on immune responses.

Data access

NGS raw sequence files – NCBI BioProject PRJNA264971 (Individual SRA numbers are provided in Table 2).

Assembled contigs – The file Lburdigalensis_transcriptome_assemby.fasta is accessible on Dryad using the following url: http://datadryad.org/review?doi=doi:10.5061/dryad.6365v

Blast hits (with *D. rerio* cDNA database)- – The file Blast_Lburdigalensis_Drerio_cDNA.xls is accessible on Dryad using the following url: http://datadryad.org/review?doi=doi:10.5061/dryad.6365v

Meta information

Sequencing center - Plateforme Génomique Génopole Toulouse Midi-Pyrénées (Toulouse, France, https://genomique.genotoul.fr/).

Platform and model - HiSeq 2000 (Illumina)

Design description - Eight adult fish (*L. burdigalensis*) were collected from a single sampling site (X = 586643; Y = 1962631) on the Célé River in South-Western France using electric-fishing (DEKA 7000; 100-300 V; 1-3 A) in the early fall 2012. Fish were directly transferred and stored at the Station d'Ecologie Expérimentale du CNRS at Moulis (SEEM; France). Fish were maintained in a well-oxygenated 200 L tank containing water from the sampling site during 12 hours (i.e. overnight). This time lag before sampling tissues was used to minimize possible stress induced by the fishing and transportation to the laboratory. Tissues from the cephalic kidney, the spleen and fins of each fish were sampled with RNAse-killer-treated tools in a RNAse-free chirurgical room. All samples were directly stored in liquid nitrogen before storage at -80°C.

Analysis type - RNA / cDNA

Run date - 18 february 2013 (first lane) and 07 march 2013 (two subsequent lanes)

Library

Strategy - next-generation automated DNA sequencing (Illumina) of normalized cDNA

Taxon - *Leuciscus burdigalensis*

Sex - Unknown

Location - Célé River in South-Western France (X = 586643; Y = 1962631)

Tissue - Cephalic kidney, spleen and fin

Additional sample information - Information about the fish sampled is provided in Table 1.

Layout – Paired-end reads (2 X 101 bp)

Library construction protocol - Total individual RNA was extracted from each sampled tissue using the RNeasy Plus Mini Kit (Qiagen reference: 7413). The final product was eluted in 40 µL RNAse-free water. Each RNA extraction was dosed on a nanodrop ND-8000 (Thermo Scientific), which allows estimating the concentration as well as possible contamination by salts (260/230 ratio) and proteins (260/280 ratio). The quality of each RNA extraction was measured using a BioAnalyser (Agilent Technologies) based on the RIN estimation (i.e. RNA integrity number) and the 28S/18S ratio.

RNA-seq libraries have been prepared according to Illumina's protocols on a Tecan EVO200 liquid handler using the Illumina TruSeq RNA sample prep kit v2 to analyze RNA. Briefly, mRNA were selected using poly-T magnetic beads. Then, total RNAs were fragmented to generate double stranded cDNAs to be sequenced. 10 cycles of PCR were applied to amplify libraries. Libraries were tested qualitatively on an Agilent BioAnalyser and then quantified by QPCR using the KAPA Library Quantification Kit to obtain an accurate quantification. RNA-seq experiments have been performed on an Illumina HiSeq2500 (High Throughput mode) using a paired-end read length of 2x100 pb with the Illumina kits TruSeq SBS sequencing kits v3. The 24 libraries were multiplexed and sequenced three times on three independent lanes.

Processing

Pipeline – Sequencing files were first cleaned to remove adaptors from the methodological procedure (i.e. TruSeq adaptors) using Cutadapt software (Martin, 2011). Sequencing files were then filtered based on their quality using sickle software (Joshi & Fass, 2011) using the default settings (i.e. bases quality value with Phreds>30; minimum read length allowed after

trimming = 15 bases). At the end of this processing, we ended with a set of high quality read files for each library.

Reads from all libraries were pooled together and normalized according to depth of sequencing coverage as recommended by Haas *et al.* (Haas *et al.*, 2013). The *in silico* read normalisation was processed using the normalised_by_kmer_coverage.pl scripts included in the Trinity software package setting the maximum targeted coverage to 30 as recommended.

Based on the normalized reads obtained, we assembled *de novo* the transcriptome using the Trinity platform and a K- mer method following Haas *et al.* (Haas *et al.*, 2013) with Trinity parameters set as defaults (K-mer size = 25; minimum contig length = 200;,minimum k-mer coverage = 1).

To assess the quality of the obtained transcriptome, we examined the number of raw input RNA-seq reads that were well represented by the transcriptome assembly. We also blasted the transcriptome assembled *de novo* for *L. burdigalensis* to the transcriptome of *Danio rerio* available at: ftp://ftp.ensembl.org/pub/release-77/fasta/danio_rerio/cdna/ using an e-value threshold of $1e^{-10}$.

Results

Number of reads: Overall 1 464 928 470 reads were obtained from the three Illumina runs. After processing filtering and normalisation, we obtained 1 355 846 866 high quality reads with a mean length of 94.06 bases (Table 2).

De novo assembly: The final assembly obtained consisted of 659364 transcripts (i.e. contigs) composed of a total of 847630261 assembled bases. The transcript median sequence length was 567 bases. Only a small fraction of the overall raw reads (4.24%; Table 3) did not properly map to the *de novo* assemblage of *L. burdigalensis*. Moreover, a total of 218510 contigs aligned with 25625 sequences (i.e. transcripts) from *D. rerio* cdna database.

Acknowledgments

This work is part of the project INCLIMPAR (ANR-11-JSV7-0010) supported by a grant from the Agence National de la Recherche (ANR) awarded to GL. O.R. is also grateful to the "Region Midi-Pyrénées" for financial support. We also thank Charlotte Veyssière and Elise Mazé-Guilmo for their help with the sampling of *L. burdigalensis* samples on the field as well as Daphné Clet for his help with tissue collection and conditioning.

174 **Tables**

175 **Table 1. Details on collected fish used for the *de novo* transcriptome assembly**

Individual code	Individual status	Body length (mm)	Weight (g)	Total nb of parasites on fish individual	Nb of parasites on the collected fin	Collected fin
Leu01	parasitised	168	45.12	14	3	Dorsal
Leu02	healthy	175	55.43	0	0	Dorsal
Leu03	parasitised	165	43.05	5	3	Dorsal
Leu04	healthy	155	39.3	0	0	Dorsal
Leu05	parasitised	217	106.07	18	2	Dorsal
Leu08	healthy	183	63.12	0	0	Anal
Leu09	parasitised	205	72.14	7	1	Dorsal
Leu10	healthy	140	26.9	0	0	Dorsal

176

177 **Table 2. Number of reads obtained from Illumina sequencing**

Sample	Nreads before filtering	Nreads after filtering	Nreads filtered	% of reads deleted	SRA number
Leu01-fin	91991354	85450864	6540490	7.11	SRS733950
Leu01-spleen	65923200	60893730	5029470	7.63	SRS734012
Leu01-kidney	67552490	62922934	4629556	6.85	SRS734005
Leu02-fin	53296552	49465088	3831464	7.19	SRS734012
Leu02-spleen	56809568	52333622	4475946	7.88	SRS734013
Leu02-kidney	57621678	53468986	4152692	7.21	SRS734015
Leu03-fin	49565770	45857486	3708284	7.48	SRS734018
Leu03-spleen	72050396	66604748	5445648	7.56	SRS734019
Leu03-kidney	50569852	46448306	4121546	8.15	SRS734032
Leu04-fin	61349962	57078596	4271366	6.96	SRS734039
Leu04-spleen	55712798	51296084	4416714	7.93	SRS734040
Leu04-kidney	55793586	51695306	4098280	7.35	SRS734042
Leu05-fin	63023934	57790854	5233080	8.30	SRS734048
Leu05-spleen	54826002	50686718	4139284	7.55	SRS734049
Leu05-kidney	46233120	42961000	3272120	7.08	SRS734050
Leu08-fin	57508194	53328620	4179574	7.27	SRS734055
Leu08-spleen	61097996	56532080	4565916	7.47	SRS734057
Leu08-kidney	54784724	50804690	3980034	7.26	SRS734058
Leu09-fin	66687296	61506218	5181078	7.77	SRS734059
Leu09-spleen	58249430	53546634	4702796	8.07	SRS734060
Leu09-kidney	62414984	57621810	4793174	7.68	SRS734061
Leu10-fin	53392686	49502216	3890470	7.29	SRS734062
Leu10-spleen	57920062	53869736	4050326	6.99	SRS734063
Leu10-kidney	90552836	84180540	6372296	7.04	SRS734064
TOTAL	1464928470	1355846866	109081604	7.45	

178

179 **Table 3. Counts of raw reads aligned on the *de novo* assembled transcriptome**

Read classification	Count	Percentage
Proper pairing	45868764	89.38
Improper pairing	2173650	4.24
Right only	1673552	3.26
Left only	1604451	3.13

180

References

Acevedo-Whitehouse K, Cunningham AA (2006) Is MHC enough for understanding wildlife immunogenetics? *Trends in Ecology & Evolution* **21**, 433-438.

Aubrook EW, Fryer G (1965) The parasitic copepod *Tracheliastes polycolpus* Nordmann in some Yorkshire rivers: the first British records. *Naturalist London* **893**, 51-56.

Blanchet S, Méjean L, Bourque J-F*, et al.* (2009a) Why do parasitized hosts look different? Resolving the "chicken-egg" dilemma. *Oecologia* **160**, 37-47.

Blanchet S, Rey O, Berthier P, Lek S, Loot G (2009b) Evidence of parasite-mediated disruptive selection on genetic diversity in a wild fish population. *Molecular ecology* **18**, 1112-1123.

Bonneaud C, Balenger SL, Zhang J, Edwards SV, Hill GE (2012) Innate immunity and the evolution of resistance to an emerging infectious disease in a wild bird. *Molecular Ecology* **21**, 2628-2639.

Daszak P, Cunningham AA, Hyatt AD (2000) Emerging infectious diseases of wildlife - threats to biodiversity and human health. *Science* **287**, 443-449.

Dheilly NM, Adema C, Raftos DA*, et al.* (2014) No more non-model species: The promise of next generation sequencing for comparative immunology. *Developmental & Comparative Immunology* **45**, 56-66.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15.

Fryer G (1982) The parasitic copepoda and branchiura of British freshwater fishes., Cumbria.

Gilbert AT, Fooks AR, Hayman DTS*, et al.* (2013) Deciphering serology to understand the ecology of infectious diseases in wildlife. *Ecohealth* **10**, 298-313.

Haas BJ, Papanicolaou A, Yassour M*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494-1512.

Jenner RG, Young RA (2005) Insights into host responses against pathogens from transcriptional profiling. *Nature Reviews Microbiology* **3**, 281-294.

Joshi N, Fass J (2011) Sickle: A sliding window, adaptive, quality-based trimming tool for FastQ files (Version 1. 33). Available at https://github.com/najoshi/sickle.

Loot G, Poulet N, Reyjol Y, Blanchet S, Lek S (2004) The effects of the ectoparasite *Tracheliastes polycolpus* (Copepoda : Lernaeopodidae) on the fins of rostrum dace (*Leuciscus leuciscus burdigalensis*). *Parasitology Research* **94**, 16-23.

Martin M (2011) Cutadapt removes adapter sequences from high throughput sequencing reads. EMBnet.journal, North America, 17, May, 2011. Available at https://journal.embnet.org/index.php/embnetjournal/article/view/200.

Metzker ML (2010) Applications of next-generation sequencing technologies - the next generation. *Nature Reviews Genetics* **11**, 31-46.

Salinas I, Zhang Y-A, Sunyer JO (2011) Mucosal immunoglobulins and B cells of teleost fish. *Developmental & Comparative Immunology* **35**, 1346-1365.

Tuffery G (1967) Importance des considérations topographiques, biologiques, écologiques, lors de l'aménagement ou du classement d'un bassin hydrographique. *Bulletin français du pisciculture* **226**.

Uribe C, Folch H, Enriquez R, Moran G (2011) Innate and adaptive immunity in teleost fish: a review. *Veterinarni Medicina* **56**, 486-503.

Woolhouse MEJ (2008) Epidemiology: Emerging diseases go global. *Nature* **451**, 898-899.

Zapata A, Diez B, Cejalvo T, Gutiérrez-de Frias C, Cortés A (2006) Ontogeny of the immune system of fish. *Fish & Shellfish Immunology* **20**, 126-136.

Zhu L-y, Nie L, Zhu G, Xiang L-x, Shao J-z (2013) Advances in research of fish immune-relevant genes: A comparative overview of innate and adaptive immunity in teleosts. *Developmental & Comparative Immunology* **39**, 39-62.