BMC Research Notes

DATA NOTE

Open Access

De novo assembly of the Brown trout (*Salmo trutta* m. *fario*) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery

J. Fibla[1,5*], N. Oromi[1†], M. Pascual-Pons[1†], J. L. Royo[1,2], A. Palau[3] and M. Fibla[4]

Abstract

Objectives: The Brown trout is a salmonid species with a high commercial value in Europe. Life history and spawning behaviour include resident (*Salmo trutta* m. *fario*) and migratory (*Salmo trutta* m. *trutta*) ecotypes. The main objective is to apply RNA-seq technology in order to obtain a reference transcriptome of two key tissues, brain and muscle, of the riverine trout *Salmo trutta* m. *fario*. Having a reference transcriptome of the resident form will complement genomic resources of salmonid species.

Data description: We generate two cDNA libraries from pooled RNA samples, isolated from muscle and brain tissues of adult individuals of *Salmo trutta* m. *fario*, which were sequenced by Illumina technology. Raw reads were subjected to de-novo transcriptome assembly using Trinity, and coding regions were predicted by TransDecoder. A final set of 35,049 non-redundant ORF unigenes were annotated. Tissue differential expression analysis was evaluated by Cuffdiff. A False Discovery Rate (FDR) ≤ 0.01 was considered for significant differential expression, allowing to identify key differentially expressed unigenes. Finally, we have identified SNP variants that will be useful tools for population genomic studies.

Keywords: Rnaseq, De novo transcriptome, Brain & muscle transcriptome, SNP discovery, *Salmo trutta* m. *fario*

Objective

Brown trout (*Salmo trutta*) has been extensively studied by its commercial and biological importance. From the sixty-six species in this family, *S. trutta* is a species native to Europe with a wide distribution area that includes Atlantic and Mediterranean European basins, as well as northern Africa and western Asia basins [1, 2].The specie has been introduced in North and South America and Australia by its commercial exploitation for sport fishing, as well as farmed for food and game fish, extending their actual geographical distribution as discontinuous populations on all continents except Antarctica [3].

Life history traits of Brown trout populations include resident forms such as riverine (*S. trutta* m. *fario*) and migratory forms such as anadromous (*S. trutta* m. *trutta*) ecotype [4, 5]. Anadromous and non-anadromous forms coexist in the same river being apparently genetically indistinguishable [6, 7]. An extended literature on Brown trout research has been produced that includes physiological, ecological and genetic aspects [8–10]. As a contribution to this global effort, here we provide a comprehensive transcriptome data set derived from brain and muscle tissues of *Salmo trutta*

*Correspondence: joan.fibla@cmb.udl.cat
†N. Oromi and M. Pascual-Pons contributed equal to this work
[5] Complex Diseases Genetics, Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida-IRBLLEIDA, Campus de Ciències de la Salut, Edifici Biomedicina I despatx b2.17, Av. Rovira Roure, 80, 25198 Lleida, Spain
Full list of author information is available at the end of the article

m. *fario* ecotype by using RNA-seq technology. We also evaluated differential transcript expression among these two tissues identifying key differentially expressed unigenes. Finally, we applied an in-silico pipeline that allow us to discover SNP variants useful for population genomic studies. The generated data could provide new valuable genomic resources for population genetic and genomic studies that can help to answer opened questions about the live history traits of riverine *S. trutta* m. *fario* as well as differences among *S. trutta* ecotypes.

Data description

Salmo trutta m. *fario*. brain and muscle tissues were collected from 25 wild type individuals (15 females) captured at the Falmisell river (Lleida, Catalonia). RNA pools from brain (10.2 µg) and muscle (11.4 µg) tissues were obtained with equimolar concentration from each subject. The TruSeq™ RNA sample Prep Kit (Illumina, Madrid, Spain) was used to build cDNA libraries according to manufacturer instructions (Table 1, Data file 1). FASTQ sequence reads were assembled using Trinity [11] run on the paired end sequences with the

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Methodology description	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.12902474.v1
Data file 2	Descriptive statistics of assembly-sequencing	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.12902474.v1
Data file 3	FigS1 Size_distribution	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.12902405.v2
Data file 4	FigS2 GeneOntology	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.12902405.v2
Data file 5	FigS3 Differential_expression	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.12902405.v2
Data file 6	Raw RNA-seq. Reads Brain tissue	Fastq files (.fastq)	*NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP151838
Data file 7	Raw RNA-seq. Reads Muscle tissue	Fastq files (.fastq)	*NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP151838
Data file 8	Trinity144	Fasta file (.fasta)	*Figshare* https://doi.org/10.6084/m9.figshare.7326464
Data file 9	Predicted non-redundant Open Reading Frames (ORFs)	Fasta file (.fasta)	*NCBI GenBank* https://identifiers.org/ncbi/insdc:GHGR00000000.1
Data file 10	Megablast hit aligment of non-redundant ORF unigenes to reference nucleotide databases	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.7712708.v4
Data file 11	Blastx homology search of non-redundant ORF unigenes to reference protein databases	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.7712708.v4
Data file 12	Krona_pie_chart_on_Non_redundant_ORF_to_NCBI_nt_and_rnaREF_seq_2018__HTML_html	HTML file (.html)	*Figshare* https://doi.org/10.6084/m9.figshare.7712708.v4
Data file 13	Protein family (Pfam) assignation to non-redundant ORF unigenes	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.12905777.v2
Data file 14	GOslim annotation of non-redundant ORF unigene sequences	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.12905777.v2
Data file 15	KEGG pathway annotation of non-redundant ORF unigene sequences	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.12905777.v2
Data file 16	Raw_Cufflinks_Brain_transcript_expression	Cufflinks output file (.txt)	*Figshare* https://doi.org/10.6084/m9.figshare.12905747.v1
Data file 17	Raw_Cufflinks_Muscle_transcript_expression	Cufflinks output file (.txt)	*Figshare* https://doi.org/10.6084/m9.figshare.12905747.v1
Data file 18	Raw_Cuffdiff_Brain_Muscle_transcript_differential_expression_testing	Cuffdiff output file (.txt)	*Figshare* https://doi.org/10.6084/m9.figshare.12905747.v1
Data file 19	Differentialy expressed non-redundant ORF unigenes at FDR_0.01	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.12905747.v1
Data file 20	Salmo trutta m. Fario—mapped SNP_to_ORF	Varian Call Format file (.vcf)	*Figshare* https://doi.org/10.6084/m9.figshare.12905831.v1
Data file 21	SNP context sequence	Spreadsheet (.xlsx)	*Figshare* https://doi.org/10.6084/m9.figshare.12905831.v1

fixed default k-mer size of 25 and minimum contig length of 200. Descriptive statistics of assembly and sequencing is found at Table 1 (Data file 2 and Data file 3). Among the 144,984 contigs predicted by Trinity (Table 1, Data file 4 and Data file 8), we identify protein coding regions using TransDecoder package [11]. We retained the longest ORF predicted for each contig sequence with a minimum of 100 amino acids long. Transcript redundancy was further reduced by CD-hit [12], obtaining a final set of 35,189 non-redundant ORF unigenes as best cluster representatives (Table 1, Data file 5). Size distribution for clustered ORF unigenes is presented in Table 1 (Data file 3). This final set was characterized by homology search to nucleotide and protein databases (Table 1, Data file 10 and Data file 11). Taxonomic representation showed the top hits for a large fraction of unigenes (≈88%) to *Neopterigii* taxon, with 66% of unigenes assigned to family *Salmonidae* (*Salvelius sp.* (1%), *Onchorrinchus sp.* (14%) and *Salmo sp.* (51%) (Table 1, Data file 12). A total of 4337 protein motif were assigned to 23,616 ORF unigenes, being the RNA recognition motif (6.4%), Immunoglobulin domain (4.8%), Tetratricopeptide repeat (4.8%) and Protein kinase domain (3.4%) the most prevalent (Table 1, Data file 13).

Similarity search by Blast2GO renders a total of 28,132 (80%) unigenes with GO annotation. GO term were then simplified using a generic GOSlim vocabulary [13] (Table 1, Data file 14). The ten top GO terms among the Cellular Component (18,071, 64%), Molecular Function (20,691, 74%) and Biological Process (23,954, 85%) ontology at level 2 are shown in Table 1 (Data file 4). Mapping unigenes to the reference canonical pathways in the KEGG database, yields a total of 13,957 (39.8%) ORF unigenes assigned to 3421 KEGG terms (KO) defining a total of 386 pathways (Table 1, Data file 15).

Tissue specific transcriptome expression analysis was performed by normalization of raw reads (FPKM, fragments per kilobase of exon per million fragments) obtained from both tissues (Table 1, Data file 16 and Data file 17). Analysis reveals 1172 ORF unigenes expressed only in muscle, 8595 expressed only in brain and 12,072 expressed in both tissues (Table 1, Data file 5, FigS3). Differentially expressed unigenes at FDR < 0.01 and best homologous sequences are shown at Table 1 (Data file 18 and Data file 19).

Finally, we have identified 73,237 putative SNPs (Table 1, Data file 20) and extracted 150 bp sequence context to each SNP as a source for the design of PCR primers useful for genotyping protocols (Table 1, Data file 21).

Limitations

The use of pooled RNA samples does not allow us to detect sex or individual specific transcript expression profiles as well as limit our capability to detect transcripts expressed at low level in a specific individual. In addition, pooled samples avoid us to resolve SNP frequency distribution, being this parameter indirectly estimated according to the observed SNP sequence coverage in the pooled sample.

Abbreviations

BAM: Binary Sequence Alignment/Map; BLAST: Basic local alignment search tool, bp: base pair; CDS: Coding sequence; FPKM: Fragments Per Kilobase of exon model per Million mapped reads; GO: Gene Ontology; KEGG: Kyoto Encyclopaedia of Genes and Genomes; ORF: Open reading frame; Pfam: Protein families database; FDR: False Discovery Rate; SAMtools: Sequence Alignment/Map tools; SNP: Single nucleotide polymorphism.

Acknowledgements

We are grateful to all participants of Gesna Estudis Ambientals S.L. (R. Rocaspana and E. Aparicio) and Eccus Proyectos Técnicos, Medioambientales y Obras SL (MA MarínVitalla), who have participated in the sampling procedure. We would like to dedicate this paper in memory of M.A. Marín Vitalla (Nines), who passed away last year. She is sorely missed.

Authors' contributions

JF, NO and AP designed the study, NO, MP-P and MF captured animals and processed samples. NO, MP-P and JLR, carried out lab work and assisted with data analysis, JF obtained the founding, perform data analysis and drafted the manuscript. All authors read and approved the final manuscript.

Funding

This study has been supported and financed by the Biodiversity Conservation Plan of ENDESA, S.A. (ENEL Group) to JF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The data described in this Data note can be freely and openly available on *Figshare* (https://doi.org/10.6084/m9.figshare.12902474.v1; https://doi.org/10.6084/m9.figshare.12902405.v2; https://doi.org/10.6084/m9.figshare.7326464.v1; https://doi.org/10.6084/m9.figshare.7712708.v4; https://doi.org/10.6084/m9.figshare.12905777.v2; https://doi.org/10.6084/m9.figshare.12905747.v1; https://doi.org/10.6084/m9.figshare.12905831.v1). Assembly of non-redundant ORF unigene sequences are available from NCBI transcriptome shotgun assembly (TSA) database (https://identifiers.org/ncbi/insdc:GHGR00000000.1). Raw sequence reads are available from the NCBI sequence read archive (SRA) database (https://identifiers.org/insdc.sra:SRP151838). Please see Table 1 and references list [14–22] for details and links to the data.

Ethics approval and consent to participate

Permissions for electrofishing and capture of ***S. trutta*** m. ***fario*** individuals, was approved by the competent authorities: Departament de Medi Ambient i Habitatge de la Generalitat de Catalunya (current Departament d'Agricultura, Ramaderia, Pesca, Alimentacio i Medi Natural) (SF/602) of the regional authorities of Catalonia.

Consent for publication

Not applicable.

Competing interest

The authors did not report any competing interests.

Author details

[1] Institute of Biomedical Research of Lleida (IRBLleida), University of Lleida, Lleida, Spain. [2] Area of Biochemistry and Molecular Biology, School

of Medicine, University of Malaga, Malaga, Spain. [3] Environment and Soil Sciences Department, ETSEA, University of Lleida, Lleida, Spain. [4] Animal Science Department, ETSEA, University of Lleida, Lleida, Catalonia, Spain. [5] Complex Diseases Genetics, Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida-IRBLLEIDA, Campus de Ciències de la Salut, Edifici Biomedicina I despatx b2.17, Av. Rovira Roure, 80, 25198 Lleida, Spain.

Received: 4 September 2020 Accepted: 21 October 2020
Published online: 02 November 2020

References

1. Bagliniere JL. Introduction: the brown trout (*Salmo trutta* L)—its origin, distribution and economic and scientific significance. Biology and ecology of the Brown and Sea Trout. 3rd ed. London: Springer London; 2000. pp. 1–12. https://doi.org/https://doi.org/10.1007/978-1-4471-0775-0_1
2. Klemetsen A, Amundsen PA, Dempson JB, Jonsson B, Jonsson N, O'Connell MF, et al. Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories. Ecol Freshwater Fish. 2nd ed. 2003;12:1–59. https://doi.org/https://doi.org/10.1034/j.1600-0633.2003.00010.x
3. MacCrimmon HR, Marshall TL. World distribution of Brown Trout, Salmo trutta. J Fish Res Board Can. 2011;25:2527–48. https://doi.org/10.1139/f68-225.
4. Elliott JM. Quantitative ecology and the Brown Trout. Oxford University Press, USA; 1994. https://doi.org/https://doi.org/10.1577/1548-8659-123.6.1006
5. Poćwierz-Kotus A, Bernaś R, Dębowski P, Kent MP, Lien S, Kesler M, et al. Genetic differentiation of southeast Baltic populations of sea trout inferred from single nucleotide polymorphisms. Anim Genet. 2014;45:96–104. https://doi.org/10.1111/age.12095.
6. Charles K, Guyomard R, Hoyheim B, Ombredane D, Bagliniere JL. Lack of genetic differentiation between anadromous and resident sympatric brown trout (*Salmo trutta*) in a Normandy population. Aquat Living Resour. 2005;18:65–9. https://doi.org/10.1051/alr:2005006.
7. Charles K, Roussel JM, Lebel JM, Bagliniere JL, Ombredane D. Genetic differentiation between anadromous and freshwater resident brown trout (*Salmo trutta* L.): insights obtained from stable isotope analysis. Ecol Freshwater Fish. 2006;15:255–63. https://doi.org/10.1111/j.1600-0633.2006.00149.x.
8. Harvey J. Ecology of Atlantic Salmon and Brown Trout: habitat as a template for life histories. Freshw Biol. 2012;57:1531–41. https://doi.org/10.1007/978-94-007-1189-1.
9. Boel M, Aarestrup K, Baktoft H, Larsen T, Søndergaard Madsen S, Malte H, et al. The physiological basis of the migration continuum in brown trout (*Salmo trutta*). Physiol Biochem Zool. 2014;87:334–45. https://doi.org/10.1086/674869.
10. Oromi N, Jové M, Pascual-Pons M, Royo JL, Rocaspana R, Aparicio E, et al. Differential metabolic profiles associated to movement behaviour of stream-resident brown trout (Salmo trutta). PLoS ONE. 2017;12:e0181697. https://doi.org/10.1371/journal.pone.0181697.
11. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512. https://doi.org/10.1038/nprot.2013.084.
12. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9. https://doi.org/10.1093/bioinformatics/btl158.
13. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS, et al. AgBase: a unified resource for functional analysis in agriculture. Nucleic Acids Res. 2007;35:D599-603. https://doi.org/10.1093/nar/gkl936.
14. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. *Salmo trutta* m. *fario* Raw sequence reads. NCBI Sequence Read Archive; 2020. https://identifiers.org/insdc.sra:SRP151838
15. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. TSA: *Salmo trutta fario*, transcriptome shotgun assembly. GenBank; 2020. https://identifiers.org/ncbi/insdc:GHGR00000000.1
16. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. SNP discovery of Non-redundant ORF unigenes to Gene Ontology, KEGG and Protein Family databases of "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. https://doi.org/10.6084/m9.figshare.12905831.v1
17. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. Functional annotation of Non-redundant ORF unigenes to Gene Ontology, KEGG and Protein Family databases of "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. https://doi.org/10.6084/m9.figshare.12905777.v2
18. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. Tissue differential expression profile of "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. https://doi.org/10.6084/m9.figshare.12905747.v1
19. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. Supplementary Files for "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. https://doi.org/10.6084/m9.figshare.12902474.v1
20. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. Supplementary Figures for "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. Figure. https://doi.org/10.6084/m9.figshare.12902405.v2
21. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A and Fibla M. Annotation of Non-redundant ORF unigenes to nucleotide and protein databases of "De novo assembly of the Brown trout (Salmo trutta m. fario) brain and muscle transcriptome: transcript annotation, tissue differential expression profile and SNP discovery". Figshare; 2020. https://doi.org/10.6084/m9.figshare.7712708.v4
22. Fibla J, Oromi N, Pascual-Pons M, Royo JL, Palau A, Fibla M. De novo assembled contigs. Figshare. 2020. https://doi.org/10.6084/m9.figshare.7326464.v1.

Publisher's Note