



AQUAculture infrastructures for EXCELlence
in European fish research towards 2020 —
AQUAEXCEL2020

D3.5: Bioinformatic tool performing meta-analysis of fish genomic data and implementation of public access

*Patrick Prunet, Ambre-Aurore Josselin, Thomas
Chaussepieds, Isabelle Leguen, Jérôme Montfort*



Executive Summary

Objectives:

Fish and Chip is a bioinformatic tool which enables meta-analysis of fish microarray data in order to highlight conserved gene signatures between multiple independent studies performed in fish.

Rationale: To improve performance of Fish and Chip tool created in the previous AQUAEXCEL project, we have developed several additional modules.

Main Results:

The following tasks have been developed within Fish and Chip tool:

- Extension of Fish and Chip to RNAseq data: The capability of Fish and Chip developed in the previous AQUAEXCEL project was limited to microarray. The development of numerous studies in fish using RNAseq lead us to extend the Fish and Chip database to RNAseq data, to allow comparisons with microarray data.
- Extension of Fish and Chip to meta-analysis of gene regulation through information on regulatory motifs of gene promoters: to carry out this analysis, information on regulatory motifs of gene promoters (extracted from available fish genome sequences) was incorporated in the Fish and Chip tool.
- The updated Fish and chips tool is publicly available at the following address: <http://fishandchips.toulouse.inra.fr>

Authors/Teams involved:

INRA/LPGP:

Patrick Prunet, Ambre-Aurore Josselin, Thomas Chaussepieds, Isabelle Leguen, Jérôme Montfort

Table of contents

Executive Summary.....	2
Table of contents	3
1. Introduction and rationale.....	4
2. Update of the genomic Fish and Chip tool.....	5
2.1. Extension of Fish and Chip to RNAseq data.....	5
2.2. Extension of Fish and Chip to meta-analysis of gene regulation through information on regulatory motifs of genepromoters.....	6
3. Conclusion.....	6
Document information.....	7
Annex 1: Check list.....	8

1. Introduction and rationale

During the last 10 years, a large number of studies in fish have developed transcriptomic approaches using microarray technologies and more recently RNAseq analysis. These methodologies have been used to answer a large diversity of biological questions in fish, including development, toxicology, physiology, behavior, pathology. These transcriptomic data have been accumulated in public generic databases and they are all associated with specific experimental protocols developed in various fish species (Qian et al., 2014). Collecting the different transcriptomic data sets for meta-analysis adds a new dimension to gene expression analysis by combining a large set of experimental conditions. Despite this abundance of transcriptomic data, no such meta-analysis has been pursued using fish data whereas such approach is used, for example, in humans to better understand diseases (ex. Kuhn et al., 2007; Calura et al., 2008). In this context, a tool that makes fish transcriptome data and their meta-analysis easily accessible to any user has been developed in the previous AQUAEXCEL project (2011-2014). The objective of this tool was to enable meta-analysis **of fish microarray data** in order to highlight conserved gene signatures between multiple independent studies performed in fish.

The general strategy to develop this meta-analysis tool was as a system tool dedicated to a muscle transcriptome database and presented in the MADMuscle database (Baron et al., 2011). In the previous AQUAEXCELproject (2009-2014) a similar tool was developed for fish transcriptome data using the same approach and the following tasks have been carried out:

- Pooling of fish data sets: This included development of a database containing all available information EST from several fish species (seabream, salmon, trout, seabass....) and also genome information of 4 model species (fugu, zebrafish, stickleback, medaka). These data were extracted from public repositories (UniGene, NCBI).
- Gene annotation for each fish species: This was carried out on the 4 species, seabream, seabass, salmon and trout. For each identified cluster, functional annotation was performed using Gene Ontology (GO). This step on the quality of gene annotation appeared to have a strong impact on the quality of the results of the meta-analysis.
- Implementation of microarray data: Collection and implementation of fish microarray data coming from about 400 data sets: These data sets were extracted from public repositories (GEO and ArrayExpress) using the keyword “fish”.
- Development of program for data re-normalization and clustering: This step aimed at normalizing the pre-normalized microarray data to remove artefacts and ensure that each entry of the database follows the same procedure. Then, hierarchical classification was used to allow investigation of relationships between samples and genes.
- Development of a program which enables setting up orthology relationships between distant fish species.

Completion of these different tasks lead to the implementation of a functional tool dedicated to meta-analysis of fish microarray data. This tool is implemented on the Genotoul bioinformatics platform at the following address:

<http://fishandchips.toulouse.inra.fr>

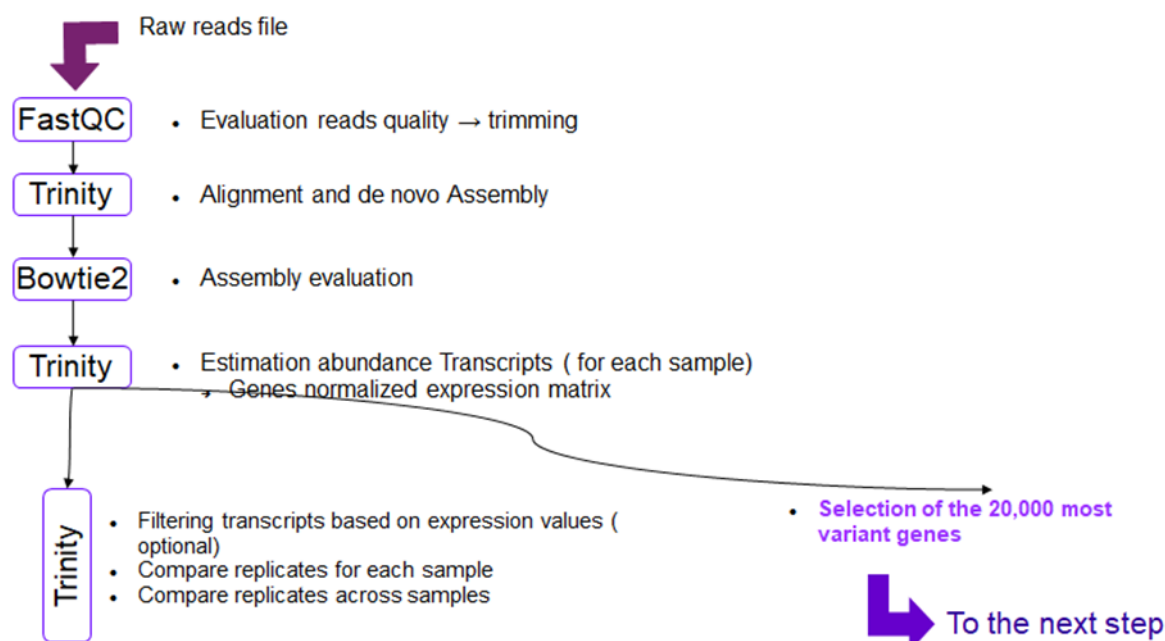
To login to Fish and Chip tool, users can request a password to the webmaster at the following address: Jerome.montfort@inra.fr. These information are indicated on the home page of the web interface.

2. Update of the genomic Fish and Chip tool

The capability of Fish and Chip developed in the previous AQUAEXCEL project was limited to microarray. The development of numerous studies in fish with RNAseq data lead us to extend the Fish and Chip database to RNAseq data, to allow comparisons with microarray data. Finally, as a way to enable the extension of Fish and Chip to meta-analysis of gene regulation, information on regulatory motifs of gene promoters (extracted from available fish genome sequences available) was developed in Fish and Chip tool.

2.1. Extension of Fish and Chip to RNAseq data.

Completion of this task required to retrieve RNAseq experiments from the NCBI Sequence Read Archive (SRA) and to use their raw reads files. This information was further processed using the Trinity pipeline which allowed evaluation the quality of the reads, to align them, to perform a *de novo* assembly which was further evaluated (Grabherr et al., 2011). Then, an estimation of the abundance of transcripts was obtained for each sample which gave us a gene normalized expression matrix. Finally, from this information, we extracted the 20,000 most variant genes for further processing in the Fish and Chip tool, i.e. annotation and integration of these data in Fish and Chip databases where they could be used in addition to microarray data. The overall procedure is summarized in fig. 1 (below).



assemble transcripts and identify the most variable genes from RNAseq data

2.2. Extension of Fish and Chip to meta-analysis of gene regulation through information on regulatory motifs of gene promoters.

In order to extend Fish and Chip to meta-analysis of gene regulation, we have developed a tool for comparison of regulatory motifs of the genes contained in the clusters of interest. Gene promoters included in a cluster of interest were extracted from Ensembl. Then we used the MEME suite to analyse motifs conservation among genes. The MEME algorithm allows us to discover motifs and the AME algorithm identifies known motifs for a transcription factor in promoters. After identification, enrichment for each motif is calculated. If this score is relevant we can hypothesise that genes that shared this motif in their promoters could be regulated by the same transcription factor.

2.3. Implementation of public access

Both above mentioned extensions have been implemented in the Fish and Chips tool which is available at <http://fishandchips.toulouse.inra.fr> with the same procedure as explained before.

3- Conclusion

The work performed on Fish and Chip tool during AQUAEXCEL²⁰²⁰ allowed us to extend this tool to the analysis of RNAseq data, to develop meta-analysis of gene regulation. This tool is now available on the web and accessible to users. However, to be fully competitive, the Fish and Chip data base containing gene expression data and gene annotation should be updated regularly and such process needs to be done automatically. This issue is discussed in the report of deliverable D3.7.

References

Baron D., Dubois E., Buhouée A., Teusan R., Steenman M., Jourdon P., Magot A., Péréon Y., Veitia R., Savagner F., Ramstein G., Houlgatte R. BMC Genomics 2011, 12: 113.

Calura E., Cagnin S., Raffaello A., Laveder P., Lanfranchi G., Romualdi C. Meta-analysis of expression signatures of muscle atrophy : gene interaction networks in early and late stages. BMC Genomics 2008, 9: 630.

Calduch-Giner J.A., Echasserieau Y., Crespo D., Baron D., Planas J.V., Prunet P., Perez-Sanchez J. Transcriptional assessment by microarray analysis and large-scale meta-analysis of the metabolic capacity of cardiac and skeletal muscle tissues to cope with reduced nutrient availability in Gilthead Sea Bream (*Sparus auratus* L.). Mar. Biotechnol. 2014, 16: 423-435.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol. 29(7):644-52.

Kuhn A., Goldstien D.R., Hodges A., Starnd A.D., Sengstag T., Kooperberg C., et al. Mutant huntingtin's effects on striatal gene expression in mice recapitulates changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage. Hum. Mol. Genet. 2007, 16: 1845-1861.

Document information

EU Project N°	652831	Acronym	AQUAEXCEL ²⁰²⁰
Full Title	AQUAculture Infrastructures for EXCELlence in European Fish Research towards 2020		
Project website	www.aquaexcel.eu		

Deliverable	N°	D3.5	Title	Bioinformatic tool performing meta-analysis of fish genomic data and implementation for public access
Work Package	N°	3	Title	Common standards and tools

Date of delivery	Contractual	15/010/2019 (Month 48)	Actual	15 29/10/2019 (Month 48 49)
Dissemination level	X	PU Public, fully open, e.g. web		
		CO Confidential, restricted under conditions set out in Model Grant Agreement		
		CI Classified, information as referred to in Commission Decision 2001/844/EC.		

Authors (Partner)	INRA LPGP			
Responsible Author	Name	Patrick Prunet	Email	Patrick.prunet@inra.fr

Version log			
Issue Date	Revision N°	Author	Change
23/10/2019			Accepted version

Annex 1: Check list

Deliverable Check list (to be checked by the “Deliverable leader”)

	Check list		Comments
BEFORE	I have checked the due date and have planned completion in due time	X	<i>Please inform Management Team of any foreseen delays</i>
	The title corresponds to the title in the DOW	X	<i>If not please inform the Management Team with justification</i>
	The dissemination level corresponds to that indicated in the DOW	X	
	The contributors (authors) correspond to those indicated in the DOW	X	
	The Table of Contents has been validated with the Activity Leader	X	<i>Please validate the Table of Content with your Activity Leader before drafting the deliverable</i>
	I am using the AQUAEXCEL ²⁰²⁰ deliverable template (title page, styles etc)	X	<i>Available in “Useful Documents” on the collaborative workspace</i>
The draft is ready			
AFTER	I have written a good summary at the beginning of the Deliverable	X	<i>A 1-2 pages maximum summary is mandatory (not formal but really informative on the content of the Deliverable)</i>
	The deliverable has been reviewed by all contributors (authors)	X	<i>Make sure all contributors have reviewed and approved the final version of the deliverable. You should leave sufficient time for this validation.</i>
	I have done a spell check and had the English verified	X	
	I have sent the final version to the WP Leader, to the 2 nd Reviewer and to the Project coordinator (cc to the project manager) for approval	X	<i>Send the final draft to your WP Leader, the 2nd Reviewer and the coordinator with cc to the project manager on the 1st day of the due month and leave 2 weeks for feedback. Inform the reviewers of the changes (if any) you have made to address their comments. Once validated by the 2 reviewers and the coordinator, send the final version to the Project Manager who will then submit it to the EC.</i>