Advancing European Aquaculture by Genome Functional Annotation

Project no:	817923
Call:	H2020-SFS-2018-2
Start date:	1st May 2019
Duration:	48 months
Coordinator:	NMBU

D6.1. Genome-wide homology relationships established for AQUA-FAANG genomes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817923

Deliverable Name	Genome-wide homology relationships established for AQUA-FAANG genomes		
Deliverable No	D6.1		
Work package number(s)	WP6		
Document type (nature)	Report		
Due Date	31.03.2022.		
Responsible Partner	UEDIN		
Author(s) ***Name and Organisation***	Dan Macqueen, UEDIN		
Reviewer(s)	Peter Harrison		
Dissemination level	PU	Public	X
	CO	Confidential, only for members of the consortium (including the Commission Services)	
Short description	Description of resources generated for comparative genomics and homology prediction using the six AQUA-FAANG species		

Change Records			
Version	Date	Changes	Author

Disclaimer

The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Content

1 Executive summary

The grant agreement describes D6.1 as "Genome-wide homology relationships of functional and regulatory features established for AQUA-FAANG genomes; available for further analyses by partners, with long-term curation and visualization achieved via the Genomicus and Ensembl browsers, with an associated report on the main outcome. Can be used to measure achievement of objective 6.1 (and project specific Objective 4)." The work done to achieve D6.1 was addressed in Task 6.1, led and coordinated by EMBL, INRAe (working in collaboration with IBENS-ENS) and UEDIN. The work achieved has been delayed compared to the original project schedule (original D6.1 deadline was Month 16) owing to impacts of the COVID pandemic that have been well described in other reports. Importantly, the results achieved meet the standards described in the grant agreement.

This report provides a summary of the work achieved, which has generated valuable comparative tools available for use in AQUA-FAANG, that are publically available and will be of great value to the fish research community.

2 Final Ensembl Genomes

A key component of D6.1 is the availability of final high-quality reference genomes for all six AQUA-FAANG species, which were uptaken and annotated by Ensembl across 2021. For information on the genomes captured under AQUA-FAANG including accession numbers, see https://projects.ensembl.org/aqua-faang/ and Figure 1. Within Ensembl, several important comparative resources are available (Herrero et al. 2016; Yates el al. 2022), including gene trees and associated predictions for gene homology relationships (i.e. orthology relationships across species; paralogy relationships within genome), which can easily be extracted for global data analysis. Additionally, a whole genome alignment has been generated using Cactus (Armstrong et al. 2020) inclusive of the six AQUA-FAANG species and zebrafish. The gene trees from Ensembl were used as the start point for a more refined homology prediction done using the SCORPiOS approach (Section 4). Both the gene trees and Cactus alignments are being used in Tasks 6.2 and 6.3. The gene trees, associated sequence alignments, and Cactus whole genome alignment have been made available to project participants in advance of the Ensembl

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817923

106 release via an FTP site for the project hosted by EMBL (http://ftp.ensembl.org/pub/misc/aqua-faang/).

3 Genomicus browser

As an alternative comparative genomic framework for analysis of AQUA-FAANG species, the final Ensembl genomes (release 106) have been uptaken to the Genomicus database and webserver (Nguyen et al. 2022). This work was done by project collaborator IBENS-ENS in close liaison with WP6 (EMBL, UEDIN). Genomicus allows researchers to quickly visualize the conservation of genomic regions across species and duplicated regions within species at different physical scales, from specific genes of interest through to genome-wide chromosomal comparisons (e.g. Figure 2). A permanent link to the Genomicus database inclusive of Ensembl 106 AQUA-FAANG genomes can be found at: https://www.genomicus.bio.ens.psl.eu/genomicus-106.01/cgi-bin/search.pl. This database is publically available with full functionality, as of the release of Ensembl 106, but was shared in advance with AQUA-FAANG participants from January 2022.

4 Final AQUA-FAANG homology prediction

The Ensembl gene trees described in Section 2 were used to make a refined prediction of gene homology relationships using a strategy developed within the French National Research Agency project GenoFish (Grant No. ANR-16-CE12-0035) called SCORPiOs (Synteny-guided CORrection of Paralogies and Orthologies), which was recently published by Parey et al. (2021). This work was led by project collaborator IBENS-ENS in close liaison with WP6 (EMBL, UEDIN). The purpose of SCORPiOs is to correct gene trees for branching artefacts near to whole genome duplication nodes, which for the AQUA-FAANG species (and teleosts generally) include the ancestral teleost WGD in addition to lineage-specific WGDs in salmonid and common carp evolutionary history. SCORPiOs accounts for integrated synteny and phylogenetic information, and has been shown to correct ~20% of all trees with incorrect branching associated with the teleost WGD (Parey et al. 2021). The input to SCORPiOs was 53,978 Ensembl gene trees, which each captured homologous genes from up to 200 species, inclusive of a large number of fishes including the AQUA-FAANG species. The corrected trees were shared back with users in the consortium via the project FTP site hosted by EMBL (http://ftp.ensembl.org/pub/misc/aqua-faang/). These trees will form the basis of global comparative analyses performed in Task 6.2 and 6.3 and will support multiple WP6 deliverables.

5 References

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J, Marinescu VD, Alföldi J, Harris RS, Lindblad-Toh K, Haussler D, Karlsson E, Jarvis ED, Zhang G, Paten B. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020 Nov;587(7833):246-251. doi: 10.1038/s41586-020-2871-y.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817923

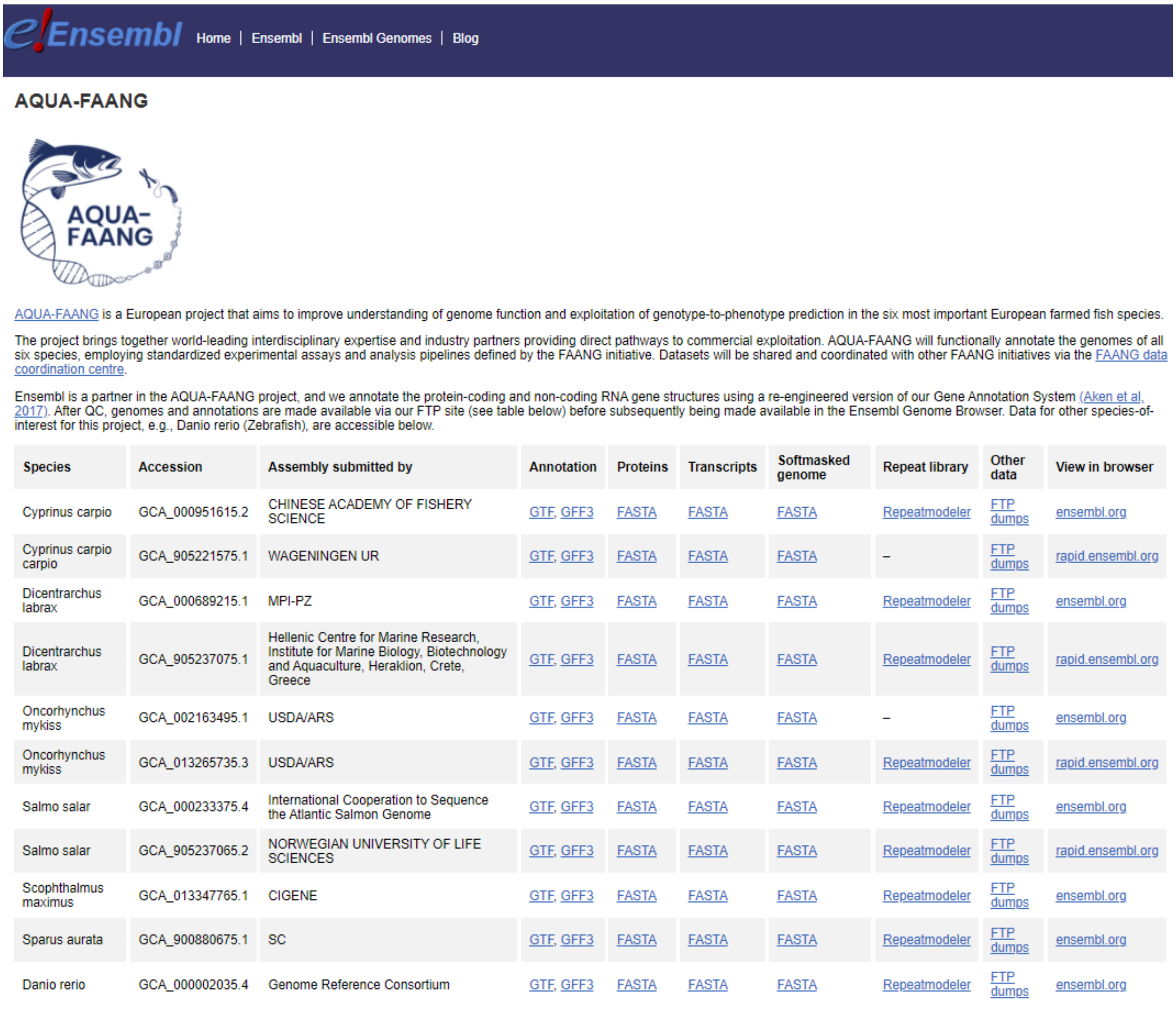
Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S, Spooner W, Kulesha E, Yates A, Flicek P. Ensembl comparative genomics resources. Database (Oxford). 2016 Feb 20;2016:bav096. doi: 10.1093/database/bav096.

Nguyen NTT, Vincens P, Dufayard JF, Roest Crollius H, Louis A. Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors. Nucleic Acids Res. 2022 Jan 7;50(D1):D1025-D1031. doi: 10.1093/nar/gkab1091.

Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications. Mol Biol Evol. 2020 Nov 1;37(11):3324-3337.

Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Carbajo Martinez M, Chakiachvili M, Chougule K, Christensen M, Contreras-Moreira B, Cuzick A, Da Rin Fioretto L, Davis P, De Silva NH, Diamantakis S, Dyer S, Elser J, Filippi CV, Gall A, Grigoriadis D, Guijarro-Clarke C, Gupta P, Hammond-Kosack KE, Howe KL, Jaiswal P, Kaikala V, Kumar V, Kumari S, Langridge N, Le T, Luypaert M, Maslen GL, Maurel T, Moore B, Muffato M, Mushtaq A, Naamati G, Naithani S, Olson A, Parker A, Paulini M, Pedro H, Perry E, Preece J, Quinton-Tulloch M, Rodgers F, Rosello M, Ruffier M, Seager J, Sitnik V, Szpak M, Tate J, Tello-Ruiz MK, Trevanion SJ, Urban M, Ware D, Wei S, Williams G, Winterbottom A, Zarowiecki M, Finn RD, Flicek P. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022 Jan 7;50(D1):D996-D1003. doi: 10.1093/nar/gkab1007.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817923



AQUA-FAANG

AQUA-FAANG is a European project that aims to improve understanding of genome function and exploitation of genotype-to-phenotype prediction in the six most important European farmed fish species.

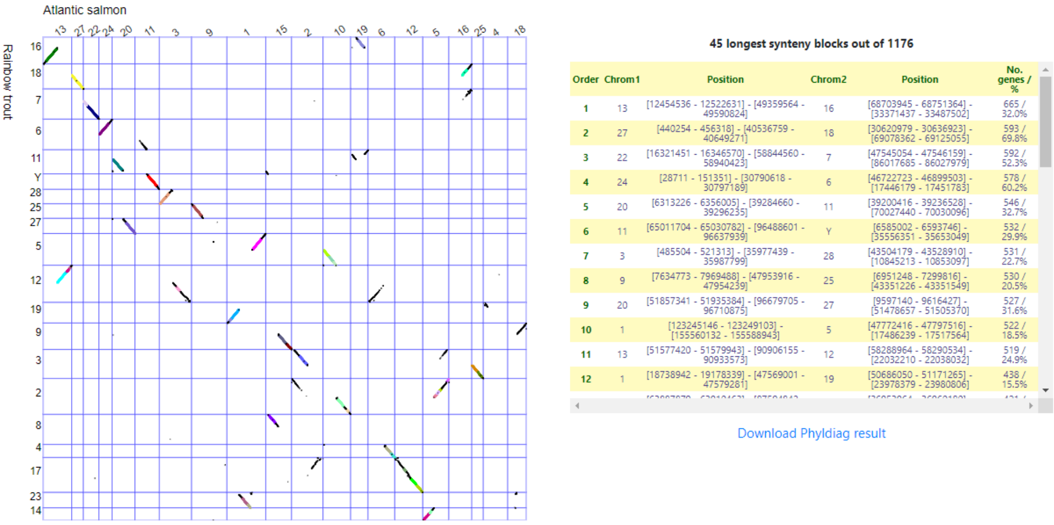
The project brings together world-leading interdisciplinary expertise and industry partners providing direct pathways to commercial exploitation. AQUA-FAANG will functionally annotate the genomes of all six species, employing standardized experimental assays and analysis pipelines defined by the FAANG initiative. Datasets will be shared and coordinated with other FAANG initiatives via the FAANG data coordination centre.

Ensembl is a partner in the AQUA-FAANG project, and we annotate the protein-coding and non-coding RNA gene structures using a re-engineered version of our Gene Annotation System (Aken et al, 2017). After QC, genomes and annotations are made available via our FTP site (see table below) before subsequently being made available in the Ensembl Genome Browser. Data for other species-of-interest for this project, e.g., Danio rerio (Zebrafish), are accessible below.

Species	Accession	Assembly submitted by	Annotation	Proteins	Transcripts	Softmasked genome	Repeat library	Other data	View in browser
Cyprinus carpio	GCA_000951615.2	CHINESE ACADEMY OF FISHERY SCIENCE	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org
Cyprinus carpio carpio	GCA_905221575.1	WAGENINGEN UR	GTF, GFF3	FASTA	FASTA	FASTA	–	FTP dumps	rapid.ensembl.org
Dicentrarchus labrax	GCA_000689215.1	MPI-PZ	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org
Dicentrarchus labrax	GCA_905237075.1	Hellenic Centre for Marine Research, Institute for Marine Biology, Biotechnology and Aquaculture, Heraklion, Crete, Greece	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	rapid.ensembl.org
Oncorhynchus mykiss	GCA_002163495.1	USDA/ARS	GTF, GFF3	FASTA	FASTA	FASTA	–	FTP dumps	ensembl.org
Oncorhynchus mykiss	GCA_013265735.3	USDA/ARS	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	rapid.ensembl.org
Salmo salar	GCA_000233375.4	International Cooperation to Sequence the Atlantic Salmon Genome	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org
Salmo salar	GCA_905237065.2	NORWEGIAN UNIVERSITY OF LIFE SCIENCES	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	rapid.ensembl.org
Scophthalmus maximus	GCA_013347765.1	CIGENE	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org
Sparus aurata	GCA_900880675.1	SC	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org
Danio rerio	GCA_000002035.4	Genome Reference Consortium	GTF, GFF3	FASTA	FASTA	FASTA	Repeatmodeler	FTP dumps	ensembl.org

Figure 1. Screenshot of final reference genomes and associated annotations captured under AQUA-FAANG including assemblies generated for common carp, European seabass, turbot, Atlantic salmon by consortium members. The final reference genomes are available on the main Ensembl browser (release 106) - https://www.ensembl.org/index.html, along with a number of comparative tools that will be widely exploited by the fish genomics research community.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817923



45 longest synteny blocks out of 1176

Order	Chrom1	Position	Chrom2	Position	No. genes / %
1	13	[12454536 - 12522631] - [49359564 - 49590824]	16	[68703945 - 68751364] - [33371437 - 33487502]	665 / 32.0%
2	27	[440254 - 456318] - [40536759 - 40649271]	18	[30620979 - 30636923] - [69078362 - 69125055]	593 / 69.8%
3	22	[16321451 - 16346570] - [58844560 - 58940423]	7	[47545054 - 47546159] - [86017685 - 86027979]	592 / 52.3%
4	24	[28711 - 151351] - [30790618 - 30797189]	6	[46722723 - 46899503] - [17446179 - 17451783]	578 / 60.2%
5	20	[6313226 - 6356005] - [39284660 - 39296235]	11	[39200416 - 39236528] - [70027440 - 70030096]	546 / 32.7%
6	11	[65011704 - 65030782] - [96488601 - 96637939]	Y	[6585002 - 6593746] - [35556351 - 35653049]	532 / 29.9%
7	3	[485504 - 521313] - [35977439 - 35987799]	28	[43504179 - 43528910] - [10845213 - 10853097]	531 / 22.7%
8	9	[7634773 - 7969488] - [47953916 - 47954239]	25	[6951248 - 7299816] - [43351226 - 43351549]	530 / 20.5%
9	20	[51857341 - 51935384] - [96679705 - 96710875]	27	[9597140 - 9616427] - [51478657 - 51505370]	527 / 31.6%
10	1	[123245146 - 123249103] - [155560132 - 155588943]	5	[47772416 - 47797516] - [17486239 - 17517564]	522 / 18.5%
11	13	[51577420 - 51579943] - [90906155 - 90933573]	12	[58288964 - 58290534] - [22032210 - 22038032]	519 / 24.9%
12	1	[18738942 - 19178339] - [47569001 - 47579281]	19	[50686050 - 51171265] - [23978379 - 23980806]	438 / 15.5%

Download Phyldiag result

Figure 2. Example of Genomicus functionality for comparative genomics using AQUA-FAANG species. The data shown highlights genome wide synteny (conservation of gene order) comparing the Atlantic salmon and rainbow trout Ensembl 106 release genome assemblies.