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

D1.2 Best-practices guidelines for sample collection and data curation

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

Deliverable Name	Best-practices guidelines for sample collection and data curation.		
Deliverable No	D1.2		
Work package number(s)	WP1		
Document type (nature)	Report		
Due Date	30 June 2021		
Responsible Partner	NMBU		
Author(s) ***Name and Organisation***	Matthew Kent, NMBU Peter Harrison, EMBL-EBI		
Reviewer(s)	Paul Flicek, EMBL-EBI		
Dissemination level	PU	Public	X
	CO	Confidential, only for members of the consortium (including the Commission Services)	
Short description	Written protocols describing biological sampling and standard requirements for metadata recording in the AQUA-FAANG project		

Disclaimer

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

Content

1 Executive summary

Mapping the functional landscape of genomes representing six different fish species is a keystone goal of AQUA-FAANG, and it is the purpose of WP1 to deliver sequencing data capturing functional element across species in a standardized and comparative way.

Deliverable 1.1 (submitted Feb 2020) reported on activities conducted within WP1 Task 1.1 ("Develop optimized functional annotation protocols in Atlantic Salmon") and describes the specific protocols (ATAC-Seq, ChIP-Seq and RNA-Seq) that have been optimized in the project to transform a variety of sample types (tissue and embryo) into sequence-ready libraries.

This deliverable (D1.2) relates to Task 1.2 ("Standardized sample collection and curation protocols") which seeks to define and standardized the activities researchers must perform before and after library production and sequencing, namely physical sampling and registration of metadata.

Physical Sampling: The fidelity of the data will be influenced by the manner in which sample are physically collected and stored. Non-standardized and/or inappropriate sampling protocols can undermine the integrity of the data and lessen its utility for downstream analysis.

Metadata registration: Data uploaded to publicly accessible databases must be accompanied by appropriate metadata that will allow any investigator to understand what biological context the data represents and how it was generated.

This Deliverable (D1.2) has experienced a significant delay (originally planned for Jan 2020, later revised to 30 June 2021) for several reasons. Within WP1, initial activity was directed at establishing working functional annotation protocols. This was more complex and took longer than was expected, but once established it was possible to interface these with appropriate physical sampling and preservation routines. Discussions related to data registration have been ongoing for some time, but as with everything else in the project, were severely disrupted due to the COVID pandemic. Over the last 18 months, we have prioritized supporting the WP1 partners in performing their experiments in an effort to minimize delays to the project. The protocols now being described in D1.2 have in reality been available to partners for many months and D1.2 is the formal reporting of this fact.

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

2 Physical Sampling

Within Task 1.2 lies Activity 1.2.1 which is defined as "Protocols developed for tissue sampling and preservation". For WP1 there are two primary sources of tissue; embryos (for DevMap) and organs (for BodyMap), each requires its own suite of methods.

For DevMap, protocols for collection and preservation of embryos collected at specific key developmental stages have already been described in D1.1, are available here, and will not be repeated here other than to say briefly:

- ATAC-seq, embryos are dissociated to single cells mechanically and/or chemically, cells are resuspended in cryopreservative and frozen.
- ChIP-seq, whole embryos are cross-linked with 1%PFA and frozen.
- RNAseq, whole embryos are mixed with Trizol reagent and homogenized with tissue-lyser before freezing the homogenate.

For BodyMap, a final comprehensive sampling and preservation protocol was made available to partners in June 2020 and is available here. Key elements are:

- Fish should be fasted 24hrs prior to sacrifice.
- Euthanasia must adhere to National regulations describing such procedures and prioritize animal welfare.
- At least 2 persons are required for efficient sampling, and they must record metadata according to the rules (discussed below in section 3).
- Detailed organ-by-organ sampling procedures (Brain, Distal Intestine, Gill, Gonad (male and female), Muscle, Liver and Head-kidney).
- Freezing procedure.
- Protocol for collection and preservation of gonad samples for histological assessment of developmental stage.

3 Metadata registration

A core principle of AQUA-FAANG is to make the data we generate publicly available and subject to the guidelines captured by "FAIR Guiding Principles for scientific data management and stewardship" (link). Essentially, data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. To do this, it is necessarily to collect metadata describing the samples, the experiment, and details about the lab assay and then connect this unambiguously to raw data files. Within WP1, Task 1.2 - Activity 1.2.2 is to "Define requirements for metadata associated with biological samples", that is decide on the specific novel information that must be collected to sufficiently describe the teleost fish data the project will generate.

Fortunately, the WP2 leader EMBL is managing the FAANG data coordination centre (DCC), which seeks to provide a full metadata solution for all FAANG projects. Following on from an initial video conference meeting on the 11th Feb 2020, a meeting held in Hinxton later in Feb 2020 (reported in Deliverable 7.4) allowed a detailed discussion about teleost metadata requirements. Over subsequent months, decisions were taken within WP1 about which sample information is to be collected, whether it is mandatory, recommended, or optional, and what

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

units or values are valid. This extended process has allowed EMBL to generate teleost-specific rulesets which have been included among the FAANG data portal (https://data.faang.org/home) sample Rule Groups and within data submission template files. As well as recording obvious general parameters such as species, tissue type, age etc, they also capture more nuanced and teleost specific parameters that are not relevant to livestock or existing model species such as average water oxygen and salinity. At this time, 13 new sample features have been defined related to "Teleostei embryos" (aka DevMap samples) and 25 sample features for "Teleostei post-hatching" (aka BodyMap samples). A complete description of these rules can be seen here (https://data.faang.org/ruleset/samples).

In October 2020, the FAANG DCC released a new validation and submission process for the submission of FAANG metadata. This service is now part of the main FAANG data portal https://data.faang.org/validation/samples), rather than the previous version that was hosted separately. This new version incorporates improved validation error handling and a more intuitive interface. The key change is the switch to automated brokering of submitted records to the underlying archives, so that the submission to the BioSamples and ENA archives is handled by the FAANG data portal on the users behalf, dramatically simplifying and quickening the submission process. To reflect the new validation and submission process the documentation has been reviewed and updated on the FAANG read-the-docs help pages, providing step by step instructions (https://dcc-documentation.readthedocs.io/en/latest/sample/biosamples_template/). To support this upgrade, online training/demonstration sessions with EMBL have been held in November 2020 and again in May 2021. Extensive discussion and trouble-shooting has also taken place using TEAMs within a dedicated Metadata channel which has seen intermittent but active usage as raw data has been returned to the project partners from the sequencing provider (Figure 1).

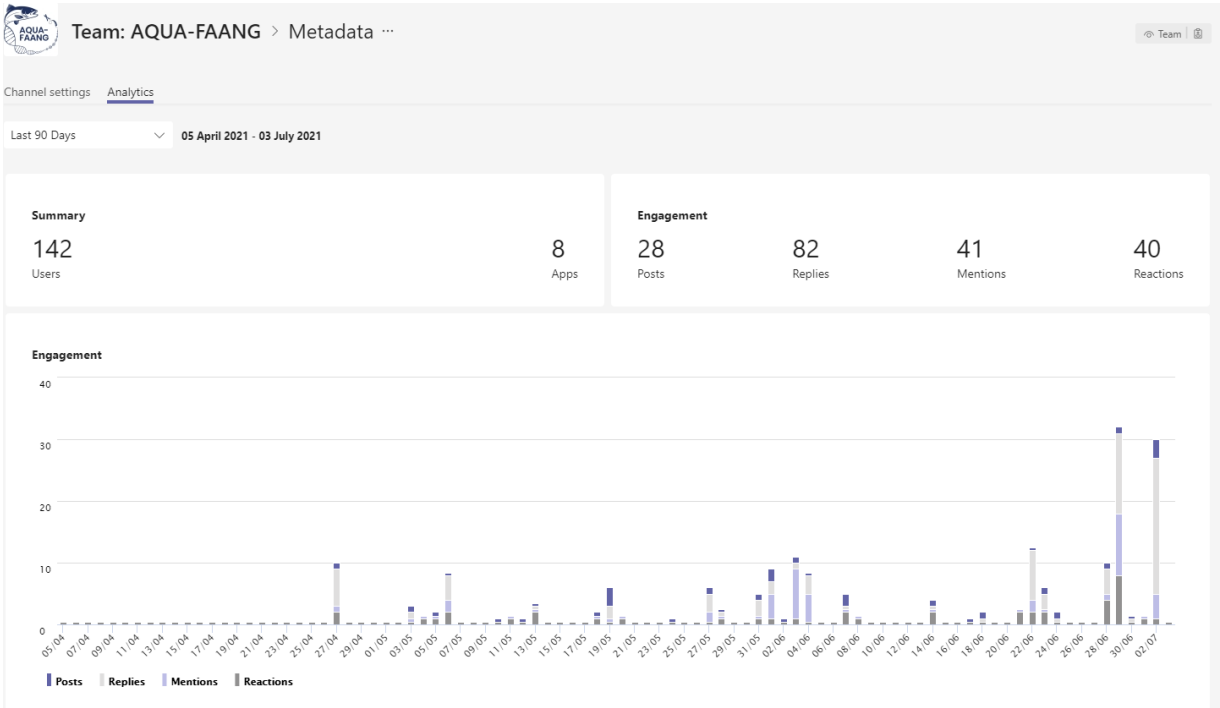

Figure 1. Histogram describing activity (posts, replies, mentions and reactions) related to the TEAMs "Metadata" channel over the last 90 days.

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

The current partner-level status is that all partners responsible for data generation have begun the process of data registration as illustrated in Figure 2 below.

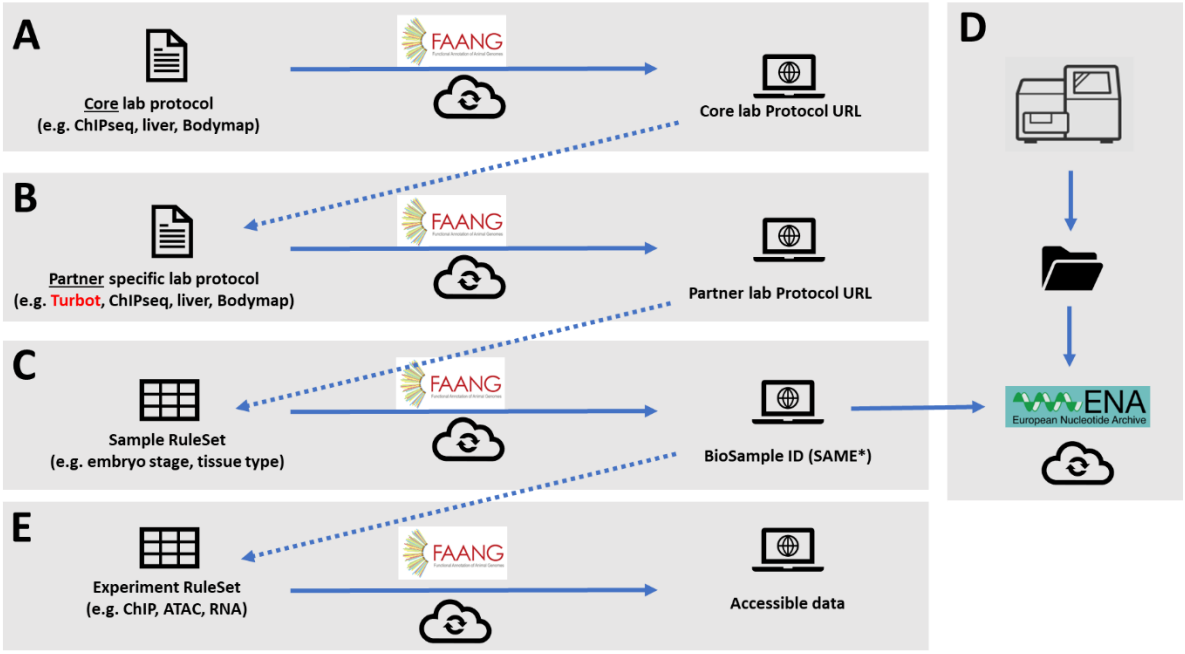

Figure 2: Schematic representation of metadata / data submission to EMBL. (A) Core lab protocols developed in WP1 (see D1.1) uploaded to FAANG data portal and assigned a unique, immutable URL, (B) partners upload accessory protocols referencing specific core protocols but including species or tissue specific modifications, URLs are assigned. (C) Sample information (including teleost specific sample features and protocols used for collection / preservation) are uploaded to generate a BioSample (SAME) ID which uniquely identifies the sample and sampling conditions. (D) Data (fastq files) generated for each library type (ChIP, ATAC, RNA) are associated with a unique BioSample ID and uploaded to the European Nucleotide Archive library, also generating an Accession number identifier. (E) The BioSample ID is integrated in the Experiment information sheet which includes lab related details about each sequencing library the sample has been used to produce.

Interdependencies exist between all stages of the multistep data submission process. At this time the core protocols and many partner specific protocols have been uploaded to FAANG-DCC and assigned URL's (stages A and B above). Furthermore, data from 3 batches of sequencing (representing almost all RNAseq and ATACseq data from WP1, 3 and 4) have been returned to partners from the sequencing provider and are ready for submission to ENA. Most partners are now engaged in competing the Sample Ruleset dataforms which will create BioSample ID and allow both data submission to ENA and continuation of Experimental Ruleset data submission. The completion of Sample and Experiment Rulesets is not trivial and there are several "first's" associated with AQUA-FAANG that have not been encountered with livestock samples, these have led to technical issues that have needed to be solved by the FAANG-DCC helpdesk. Our strategy has been to prioritize the successful submission for Atlantic salmon data and thereby generate example data sheets for Step C (sample metadata) that can be used to guide the process in other species.

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

All the partners have partially or completely finished step C thanks to the support provided by FAANG-DCC helpdesk and by Teams exchanges. Our strategy is now to complete a successful submission for a selected set of sequencing files to ENA (step D), in parallel to generate an example data sheet for step E (Experiment Metadata) that can be used to guide the process for the ensemble of the partners. We expect that NMBU (WP1 leader) will have completed all data submissions by the end of August 2021, and in so doing will provide partners with examples and experience to guide their submissions.

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