GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Deliverable D3.2
Report on coordinated data production, recording and archiving within the European node of FAANG

Deliverable leader: EMBL
Authors: Peter Harrison (EMBL-EBI)

Version: 1.0

Due date of deliverable (as in DoA):	M12
Actual submission date:	18/12/2020, month M18

Dissemination level:

PU Public	**X**
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817998

Table of contents

1 Summary

Within this document we report on the establishment of the European node of the FAANG Data coordination Centre (DCC) at EMBL. The European node of the FAANG, now referred to as EURO-FAANG, has been formalised into an entity representing the three H2020 projects for the purposes of coordination, dissemination and cross project analysis. The DCC provides the coordination and technical developments to support the dissemination activities and the interoperability for data and methodology sharing for cross project comparative analyses to be performed.

The objective was to establish the European node of the FAANG DCC through agreement of scope, remit, objectives and priorities between the SFS-30 funded projects, which includes BovReg (Identification of functionally active genomic features relevant to phenotypic diversity and plasticity in cattle; Grant agreement ID: 815668) and AQUA-FAANG (Advancing European Aquaculture by Genome Functional Annotation; Grant agreement ID: 817923). The EURO-FAANG DCC has developed interoperability between the three projects as a basis for sharing of data and analysis methodologies. This framework will lead to cross project comparative analyses and publications in the coming years as part of an established cross-project comparative analysis working group.

EMBL led a series of meetings with the coordinators and co-coordinators of the three H2020 SFS-30 projects to explore and agree the scope, remit and early objectives of the European node of the FAANG Data coordination centre at EMBL-EBI. These early discussions made it clear that a more formal collaboration between the projects was desired, that evolved into the EURO-FAANG structure. This includes a governance committee made up of the coordinators and co-coordinators, and two working groups, one focussed on communication and dissemination and the other on cross-project comparative analyses.

The EURO-FAANG DCC was established and its scope and priorities outlined in a single roadmap document that covered shared and unique development across the projects. The roadmap outlines clearly the scope, remit, objectives and priorities of the DCC for AQUA-FAANG, BovReg and GENE-SWitCH. The DCC has already commenced working on laying the required interoperable foundations for coordinated data production, recording and archiving. The metadata standards, validation and brokered submission tools and protocol sharing services have all been updated for EURO-FAANG projects. Significant progress has also been made on coordination of common analysis pipelines to ensure that data produced by each consortium can be utilised for cross-project comparisons. The now established cross-project comparative analyses working group is meeting regularly to discuss the necessary data and analysis workflow harmonisations to lay the foundations for shared analyses and publications.

The work to establish a European node of FAANG was led by the FAANG DCC at EMBL. It has involved collaboration with the coordinator and co-coordinator at INRA to establish the scope, remit and objectives of EURO-FAANG with their counterparts in the other H2020 SFS-30 projects BovReg and AQUA-FAANG. EMBL has also collaborated with EAAP for the establishment of the shared dissemination working group of FAANG and writing of the EURO-FAANG communication and dissemination plan with counterparts in the other H2020 SFS-30 projects.

2 Introduction

The FAANG Data Coordination Centre was established at EMBL in 2015 to support the data coordination, archival and presentation of data generated by the global Functional Annotation of Animal Genomes (FAANG) project. Its role is to coordinate across the various worldline FAANG projects the functional annotation of animal genomes by Ensuring data is richly described, openly available, searchable, consistently reported and clearly presented. The DCC also coordinates the standardisation of protocols, analysis methods, all with the aim of facilitating data openness, reusability and cross-project analysis. AQUA-FAANG, BovReg and GENE-SWitCH have formed a closer relationship to coordinate these objectives within Europe. This includes coordinated training, development of potential collaborative analyses, infrastructure development within the Data Coordination Centre, communication and dissemination.

3 Results

3.1 Establishing the European node of the FAANG Data Coordination Centre

The FAANG DCC was initially funded by the UK Biotechnology and Biological Sciences Research Council, and is now collectively supported by the H2020 SFS-30 projects. The funding of three European FAANG related projects AQUA-FAANG, BovReg and GENE-SWitCH, brought the DCC into a new phase of development and established a need for improved coordination at the European level. This requirement was recognised by GENE-SWitCH during the planning of the project, which lead to the objective to establish the European node of the FAANG DCC. This node would conduct the necessary interoperability development work on behalf of the H2020 projects to enable the data sharing and shared analysis necessary for cross project comparative analyses to be performed.

3.1.1 Establishing the European node of the FAANG DCC: EURO-FAANG governance

In order to establish the European node of the FAANG DCC the first requirement was to collectively establish the scope, remit, objectives and priorities between the SFS-30 funded projects. EMBL arranged a series of virtual meetings between the coordinators and co-coordinators of each project, with the PIs of each project at EMBL to represent the DCC. This consisted of Sigbjørn Lien and Dan Macqueen (AQUA-FAANG), Christa Kuhn and Dominique Rocha (BovReg), Elisabetta Giuffra and Hervé Acloque (GENE-SWitCH) from each of the projects, and Paul Flicek, Guy Cochrane, Peter Harrison and Daniel Zerbino from EMBL. This group now forms the governance of EURO-FAANG and continue to meet regularly as required (and at least quarterly) to discuss DCC objectives, dissemination and cross project activity planning. Through this group we established the formalisation of EURO-FAANG as a collective entity to represent the interests of FAANG within Europe, under this groups governance. The group reviews and discusses the EURO-FAANG DCC shared roadmap, outlined below, high level cross project objectives and plans for comparative analyses. The group has also overseen through EMBL and EAAP leadership the formation of a cross project EURO-FAANG dissemination working group and as a key outcome of the February FAANG shared workshop a cross project comparative analysis working group to explore future analyses and publications. The DCC will provide the interoperable coordination and technical development necessary for these working groups to achieve their objectives.

3.1.2 EURO-FAANG DCC shared roadmap

As EMBL is a partner in all three projects, with the shared aim of delivering data coordination, analysis, archiving and presentation to each of the projects, it was important to establish the scope and shared developmental priorities across the funded projects. EMBL has adopted a transparent approach to the shared and unique technical and scientific developments required for the H2020 projects. To this end, we authored, sought approval and then published to the consortia the FAANG H2020 Shared Developmental Roadmap for the EURO-FAANG projects (https://bit.ly/H2020Roadmap). This document outlines the development strategy and work plan for the EMBL-EBI FAANG DCC and Ensembl in supporting the three successful H2020 SFS-30-2018-2019-2020 projects. Following a thorough review of the three grant proposals we outlined the shared and unique elements amongst the projects and the strategy for delivering the common and individual features. This was crucial for the shared development performed at EMBL-EBI and allow each of the projects fair input into the design of features and activity of the DCC and Ensembl. It was also key for the competing timelines for when different DCC features were required to be in production. The roadmap accounts for all FAANG activities and their associated charges to H2020 Agri-Aqua Labs projects.

It is a living document that will be updated throughout the projects life cycles and will evolve as the requirements and timelines of each of the projects change. The DCC and analysis tasks will be regularly reviewed, and the schedule of the work plan updated with the governance EURO-FAANG group of project coordinators and co-coordinators. The roadmap outlines clearly the scope, remit, objectives and priorities between the SFS-30 funded projects, the key objective of this deliverable 3.2.

3.1.3 FAANG Shared workshop: Foundations for the future

In February 2020, EMBL organised and hosted the FAANG Shared workshop: Foundations for the future event at their campus in Hinxton, UK. This brought together 20 representatives from each of the H2020 SFS30 projects and some invited observers from key US FAANG projects. This meeting aimed to promote interoperability of the projects to prepare for cross project coordination, dissemination and analyses. Key outcomes of the meeting included establishing the EURO-FAANG dissemination working group with a number of sessions on developing a shared communication and dissemination strategy across the EURO-FAANG projects. A cross project comparative analysis working group was also established to explore future research questions, what analysis pipelines will need to be developed and what data will be required from each project.

3.1.4 EURO-FAANG dissemination working group

EURO-FAANG has formed a dissemination working group that consists of communication and dissemination work package (WP) leaders and deputy leaders from each of the projects, supported by the communications team at EMBL (Table 1). This framework was established at the FAANG shared workshop that took place on 25-27 February 2020 at EMBL-EBI in Hinxton, UK.

Table 1. Members of the EURO-FAANG dissemination working group from each of the H2020 projects.

AQUA-FAANG	BovReg	GENE-SWitCH
Cagla Kaya	Riccardo Carelli	Cagla Kaya
Lise Marie Fjellsbø	Johanna Vilkki	Riccardo Carelli
Peter Harrison	Peter Harrison	Peter Harrison
Oana Stroe	Oana Stroe	Oana Stroe

The working groups first objective was to establish a EURO-FAANG communications strategy document that outlines communications procedures and best practices for the three projects that make up EuroFAANG. It provides an umbrella strategy that will be regularly updated and should be considered in addition to the communications objectives of each project. The communication strategy outlines eight key objectives:

1. Support AQUA-FAANG, BovReg and GENE-SWitCH in achieving their objectives by issuing effective, coordinated and timely communications
2. Raising awareness within the relevant scientific communities of these initiatives
3. Promoting the tools and data resources that EuroFAANG projects make available to the community
4. Providing clear information and call to actions to the scientific community about how they can get involved in the projects and dissemination
5. Promote the public acceptance of the tools/techniques developed by the three projects
6. Cross-promote the three projects whenever appropriate
7. Encourage synergy, maximise mutual benefits and map a collaboration path between the three projects' communication and dissemination objectives
8. Target different stakeholder groups jointly, in particular non-academic partners to increase the impact and effectiveness of the dissemination activities

3.1.5 Comparative analysis working group

Established during the FAANG Shared Workshop in February 2020, the EUROFAANG comparative working group is chaired by Dan Macqueen and Hendrik-Jan Megens. This group meets regularly to lay the foundations for future cross-project comparative analyses. This is currently focussing on cross-project understanding of the data being generated by each project, to allow for the scope of future analyses to be determined. Since the shared workshop the group has now established a number of sub working groups that met for also met for the first time in November 2020. These groups will each discuss interesting questions, sub-questions and potential comparative analyses, along with associated workflows and pipelines. For GENE-SWitCH the groups involved are from INRAE, WU and EMBL.

3.2 FAANG Data Coordination Centre interoperability development work

A key component of establishing the EURO-FAANG framework of the FAANG-DCC is the interoperability components that will enable cross project comparative analysis to be possible.

3.2.1 Metadata standards and protocols

The FAANG metadata standards are a key component of interoperability in the FAANG project, ensuring that data recording is standardised across FAANG records regardless of who and where data is generated[1]. The metadata standards have undergone an in-depth review and update in response to the new requirements of the EURO-FAANG projects, with updated rulesets published for FAANG (https://data.faang.org/ruleset/samples). The FAANG Data Co-ordination Centre has released a new validation and submission process for the submission of FAANG metadata, incorporating improved validation error handling and user interface (https://data.faang.org/validation/samples). This makes the submission process and provision of rich FAIR (Findable, Accessible, Interoperable and Reusable) metadata easier for the EURO-FAANG projects. The EURO-FAANG DCC also released a new protocol upload interface, to make it easier for the EURO-FAANG projects to upload their sampling, sequencing and analysis protocols to FAANG, in particular for sharing with the other EURO-FAANG partners (https://data.faang.org/upload_protocol). All of the protocols uploaded to FAANG are made available through the data portal (https://data.faang.org/protocol/samples).

3.2.2 Common analysis pipelines

Common analysis output across the EURO-FAANG projects takes three main forms, the standard annotation produced by the EMBL Ensembl team, the coordination of standardised primary and secondary analyses conducted individually within each project and those planned as part of cross-project comparative analysis working group. Ensembl (https://www.ensembl.org/) will produce a standardised annotation of the reference genome using its standardised regulatory build for all of the EURO-FAANG species. These annotations will be presented using the Ensembl genome browser and made available from the FAANG data portal.

In addition to this, each of the projects is conducting custom additional primary and secondary analysis to address research questions of interest. Each project will also contribute to analyses as part of the comparative analysis working group. For these analyses there is significant investment into developing and implementing common analysis pipelines across the projects. This was discussed in detail at the FAANG Shared Workshop in February 2020. The workshop produced and agreed upon a set of pipeline development guidelines, standards and coding principles. This standardises the guidelines and coding principles that must be adhered to for the development of bioinformatic pipelines within EURO-FAANG.

FAANG pipelines can easily be shared across EURO-FAANG through upload to the FAANG GitHub repository (https://github.com/FAANG). For example this GENE-SWitCH produced RNA-Seq analysis pipelines (https://github.com/FAANG/proj-gs-rna-seq). There is an agreement to utilise the NextFlow NF-Core platform for the standardisation and development for pipelines for FAANG (https://nf-co.re/). This provides a curated set of pipelines for FAANG that are containerised, to perform similarly and be easily useable on any of the EURO-FAANG partners chosen compute platforms. A key additional agreement, that was an outcome of the Shared Workshop in February, was to open additional slots in planned internal training events to members of the other EURO-FAANG projects. For example, a recent training event hosted by the BovReg project on “"Reproducible genomics workflows using Nextflow and nf-core”

that a number of GENE-SWitCH developers were able to attend. This event included a hackathon for attendees to contribute to development of shared analysis pipelines for use by EURO-FAANG.

3.2.3 Prepublication data sharing

All of the EURO-FAANG projects have as members of FAANG signed up to the FAANG prepublication data sharing principles (https://www.faang.org/data-share-principle). The FAANG Data Sharing Statement is based upon the Toronto[2] and Fort Lauderdale principles[3]. You can already see raw experimental data loaded by the projects, prior to any publications having been made (https://data.faang.org/projects), demonstrating each of the projects commitment to this process. This will allow the raw and primary analysis data from all of the EURO-FAANG projects to be available across all of the projects as a basis for comparative analyses to be performed. Each of the EURO-FAANG datasets include the following data sharing statement outlining their use:

"This study is part of the FAANG project, promoting rapid prepublication of data to support the research community. These data are released under Fort Lauderdale principles, as confirmed in the Toronto Statement (Toronto International Data Release Workshop. Birney et al. 2009. Pre-publication data sharing. Nature 461:168-170). Any use of this dataset must abide by the FAANG data sharing principles. Data producers reserve the right to make the first publication of a global analysis of this data. If you are unsure if you are allowed to publish on this dataset, please contact the FAANG Data Coordination Centre and FAANG Consortium (email faang-dcc@ebi.ac.uk and copy faang@iastate.edu) to enquire. The full guidelines can be found at http://www.faang.org/data-share-principle."

3.3 Current status of EURO-FAANG data production, recording and archiving

At the time of writing of this report all of the projects are submitting their first batches of generated 'omics sequencing data to the public archives through the FAANG validation and submission service (https://data.faang.org/validation/samples). Each of the projects data generation plans have been adversely affected to varying degrees by the COVID-19 pandemic in 2020. The EURO-FAANG governance, dissemination and comparative analysis working groups provide a key communication avenue for regular exchange of updated timelines of generated data in each of the projects. This assists with plans through the cross-project comparative analysis working group for when all of the data will be in place to start investigations. Shared spreadsheets that list the data that has already and will be generated is enabling planning of what future analyses and upon what tissues and data technologies can be performed. The EURO-FAANG DCC has developed project portal pages that collate all of the sample, raw and analysed datasets from the projects (https://data.faang.org/projects). These pages provide the latest overview of the coordinated data available from EURO-FAANG.

4 Conclusion

EMBL has led on the establishment of EURO-FAANG, the European node of the FAANG DCC. The scope, remit, objectives and priorities of the EURO-FAANG DCC was established through the formation of a governance committee consisting of EMBL principal investigators and the coordinators and co-coordinators of AQUA-FAANG, BovReg and GENE-SWitCH. The interoperability development work of the DCC in cross project coordination, data sharing and

analysis is transparently documented in the FAANG Shared Development Roadmap, that is regularly reviewed by the governance committee. The DCC has already commenced on the required interoperability development to support future cross project comparative analyses, including developments to metadata standards, validation systems, protocol sharing, prepublication data archival and sharing, and sharing of common analysis methodologies and pipelines. In addition to the interoperability development of the FAANG DCC, EMBL has led on the formation of a dissemination working group and cross project analysis working group. These were key outcomes of the FAANG Shared Workshop: Foundations for the future workshop held at EMBL Hinxton in February 2020. The dissemination working group that is constructed from the communications and dissemination work packages from each project, have produced a cross project communications strategy for EURO-FAANG. The cross-project comparative analyses working group is meeting regularly to discuss the necessary data and analysis harmonisations to lay the foundations for shared analyses and publications later in the project's life cycles.

5 Deviations or delays

This deliverable was delayed by six months due to the secondment of EMBL DCC staff from March to September to support the emergency European COVID-19 scientific response to develop the European COVID-19 data portal (https://www.covid19dataportal.org/).

6 Acknowledgements

With thanks to the EURO-FAANG project coordinators and co-coordinators that form the governance committee of H2020 for strategic input into the interoperability developments of the FAANG DCC to support the H2020 SFS30 projects. Also, thanks to the coordination and dissemination committee (Table 1) and cross project comparative analysis working group led by Dan Macqueen and Hendrik-Jan Megens, for furthering the cohesiveness of the three projects.

7 References

1 Harrison, P.W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., Archibald, A.L., Schmidt, C.J. and Flicek, P. (2018), FAANG, establishing metadata standards, validation and best practices for the farmed and companion animal community. Anim Genet, 49: 520-526. https://doi.org/10.1111/age.12736

2 Toronto International Data Release Workshop Authors. Prepublication data sharing. *Nature* **461,** 168–170 (2009). https://doi.org/10.1038/461168a

3 Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics--re-shaping scientific practice. *Nature reviews. Genetics*, *10*(5), 331–335. https://doi.org/10.1038/nrg2573

8 Glossary

DCC	Data Coordination Centre
EURO-FAANG	The European node of the Functional Annotation of Animal Genomes Data Coordination Centre
FAANG	Functional Annotation of Animal Genomes
FAIR	Findable, Accessible, Interoperable and Reusable