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

D8.4 Data Management Plan prepared

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

Deliverable Name	**Data Management Plan prepared**		
Deliverable No	8.4		
Work package number(s)	8		
Document type (nature)	R - Document, report (excluding the periodic and final reports)		
Due Date	31st October		
Responsible Partner	EMBL-EBI		
Author(s) ***Name and Organisation***	Peter Harrison and Paul Flicek, EMBL-EBI.		
Reviewer(s)	Sigbjørn Lien and Ross Houston		
Dissemination level	PU	Public	X
	CO	Confidential, only for members of the consortium (including the Commission Services)	
Short description	This deliverable report documents the first version of the AQUA-FAANG Data Management plan.		

Change Records			
Version	Date	Changes	Author
1.0	31.10.19	First version – approved by reviewers: Minor editorial corrections. Annex 2 was originally a separate file but is now included to the deliverable.	Paul Flicek and Peter Harrison
0.1	28.10.19	First draft	Paul Flicek and Peter Harrison

Disclaimer

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

Content

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

1 Executive summary

This deliverable report documents the first version of the AQUA-FAANG Data Management plan. The main outcome is the internal dissemination of the data management plan to the consortium so that all are aware and comply with the standards and requirements it sets. In accordance with the grant agreement the Data Management Plan will be maintained throughout the lifetime of the project. Revised versions will be distributed to all partners in the consortium so that they are aware of updated data management practices.

2 AQUA-FAANG Data Management Plan

The first version of the AQUA-FAANG data management plan has been prepared for the AQUA-FAANG consortium (See Annex 1). The plan was prepared in accordance to guidance set out by H2020 and utilised the H2020 data management plan template. The data management plan largely covers the activities of work package 2, led by EMBL-EBI, but is of importance to all work packages that are generating and analysing project data. The external requirements stipulated by AQUA-FAANG being a member of the FAANG coordinated action were considered when preparing this plan. Namely that the data management plan for AQUA-FAANG fully complies with the FAANG data sharing statement (See Annex 2) that in turn complies with the Fort Lauderdale principles and Toronto Workshop statement. The data management plan also complies with EU and H2020 requirements for open access, long term archival access to data, and strives for highest level of compliance with FAIR data principles.

The plan provides a summary of the projected 28.7 tera base pairs (Tbp) of sequence data that will be generated as part of the project. It covers the tasks and developments that the FAANG Data Coordination Centre (DCC) will conduct to ensure that AQUA-FAANG data is FAIR (findable, accessible, interoperable and reusable) compliant. It also documents the data security of the archives at EMBL-EBI that will act as the long-term storage of AQUA-FAANG data and the ethical aspects of data generated in the project.

The first version of the Data Management Plan document was distributed to all members of the consortium. The Data Management Plan is a living document, and future revisions to the plan will be similarly distributed to all partners in the consortium. The data management plan and any updated versions will be available to the consortium from the AQUA-FAANG Microsoft Teams files storage.

3 Annexes

Annex 1: AQUA-FAANG Data Management Plan
Annex 2. The FAANG Data sharing statement. Published version from 26th May 2015. The latest version can always be accessed from https://www.faang.org/data-share-principle

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

Annex 1. AQUA-FAANG Data Management Plan

Project Number: 817923

Project title: Advancing European Aquaculture by Genome Functional Annotation

Project Acronym: AQUA-FAANG

Version: 1.0

Date Published: 31st October 2019

Author: Peter Harrison (EMBL-EBI).

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.

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

Contents

1. Data Summary

The AQUA-FAANG project aims to generate functionally annotated genomes for six key commercial fish species of European aquaculture, exploiting commercially important trait variation with a particular focus on diseases resistance. The generation of 'omics data is therefore core to the overall objective of the project. The six selected fish species are Turbot, European seabass, Gilthead seabream, Atlantic salmon, Rainbow trout and Common carp. The generated data will create functional annotation maps for healthy and immune active (diseased) individuals at different developmental stages. These newly generated maps will be employed to predict disease resistance and commercially important traits, and to translate this into innovative practice for a sustainable and competitive aquaculture sector.

The project will produce extensive metadata documenting the new fish samples collected and the experiments performed on those samples in the generation of omics data. Data will be generated from 6,221 separate samples. In total the project will generate an estimated 28.7 tera base pairs (Tbp) of sequence data across the six species and a range of sequencing technologies (Table 1). The technologies employed are Whole Genome Sequencing, RNA-Sequencing, small RNA-sequencing (small and micro RNA species), single cell RNA-sequencing (just in Rainbow trout), ChIP-Seq to assess for histone modification, ChIP-Seq CTCF (to assess CCCTC-binding factor) and ATAC-Seq (Assay for Transposase-Accessible Chromatin). A single accredited provider will be subcontracted to perform all Illumina sequencing in the project, which will ensure an industry-standard for sequencing datasets with minimal batch effects due to different sequencing devices/chemistries. In addition, the project will generate SNP genotype data using a high density SNP array for samples of sea bass (approx n = 1500). A key objective for AQUA-FAANG is that all datasets generated will follow highly standardized

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

protocols designed to match the agreed standards of the FAANG initiative (WP1, Task 1.1; WP2, Task 2.3). All processed data generated in the project will be shared among the consortium using standard bioinformatic data file formats (i.e. FASTQ, FASTA, SAM/BAM, GFF/GTF, BED, and VCF).

Table 1. Data that is projected to be generated in AQUA-FAANG project. Shows the number of assays to be conducted with each technology and the total sequence data generated.

Species	WGS	RNA-Seq	sRNA-Seq	ATAC-Seq	ChIP-Seq (histone mod.)	ChIP-Seq (CTCF)	scRNA-Seq	*Sequence data generated*
Atlantic salmon	6	120	120	120	408	84	-	4.0 Tbp
Rainbow trout	6	146	146	146	472	84	21	4.4 Tbp
European seabass	56	720	160	160	408	84	-	7.7 Tbp
Gilthead seabream	26	120	120	120	408	84	-	4.0 Tbp
Common carp	26	140	120	120	408	84	-	4.0 Tbp
Turbot	26	120	120	120	408	84	-	3.6 Tbp
Total assays:	**146**	**1,366**	**786**	**786**	**2,512**	**504**	**21**	**28.7 Tbp**

The project will make extensive use of open access existing legacy datasets and datasets generated during the lifetime of the project identified and accessed from EMBL-EBI public archives. The project will also utilise existing data and the functional genome map of the Zebrafish as the only fish species where functional annotation is in an advanced state, through the extensive work of the DANIO-CODE initiative. This will be particularly important for homology in comparative analysis across the six species. The project includes two partners from the DANIO-CODE initiative.

The data generated by AQUA-FAANG will contribute six functionally annotated genome maps, complete with immune active and development tracks to the global FAANG and wider scientific community. The experimental, bioinformatic and comparative approaches developed will be translated to accelerate genome annotation in other species of importance for global aquaculture, FAASG and FAANG. The utilisation of the data and generated genome maps will have significant commercial and societal impact on the aquaculture sector. To achieve this aim the consortium includes involvement of SME producers, breeding and genetics service companies, in addition to an active dissemination, exploitation and knowledge exchange work package (WP7), that combined will ensure commercial relevance and rapid uptake of the technologies developed in AQUA-FAANG by the aquaculture industry.

2. FAIR data

2.1. Making data findable, including provisions for metadata

AQUA-FAANG data will be highly discoverable by fully conforming to the rich FAANG metadata standards (https://data.faang.org/ruleset/samples#standard). Every AQUA-FAANG sample, dataset deposited, and data file will receive a unique accession identifier from the EMBL-EBI archives, that is globally recognisable and supported by the comparable archives at the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and

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

DNA Databank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/index-e.html)/). These persistent and unique identifiers are easily searchable within the public archives, FAANG data portal and will be included in all publications. AQUA-FAANG will also include its own secondary identifier sample naming convention, the final structure is to be determined. It will be suggested to follow the FAANG recommended sample naming convention. This will contain a short species code, the project name, and a unique ID for each sample e.g. 'SSA_AQUAFAANG_R1'.

The FAANG data portal will provide the ability to find data associated with individual samples regardless of where the data is stored. The FAANG data portal will expose all of these metadata fields in its text based search (based on ElasticSearch technology with Google-style predictive text search; https://data.faang.org/search) and programmatic interfaces (https://data.faang.org/help/api) to allow for easy access to the AQUA-FAANG data. It will be possible to search for AQUA-FAANG data as part of all FAANG datasets or pre-limit the search specifically to only return AQUA-FAANG data results. The FAANG metadata also supports data browsing (https://data.faang.org/home) by providing powerful metadata filters that allow a user to explore the AQUA-FAANG data based on species, technology, breeds, sex, material, organism part, cell type, assay type, archive, and sequencing instrument. Versioning of experimental data is handled by the EMBL-EBI public archive accession versioning, software pipelines developed by AQUA-FAANG will be given major and minor version releases in accordance with best coding practices and Ensembl will version according to its established practices (https://www.ensembl.org/info/genome/stable_ids/index.html) .

2.2. Making data openly accessible

AQUA-FAANG sample, raw and analysis data will be made rapidly publicly available without embargo by deposition in the relevant EMBL-EBI public archives. This deposition route is well established with the FAANG Data Coordination Centre (DCC), that itself is based within the Molecular Archives cluster at EMBL-EBI. All data will include in its description an accompanying data reuse statement as recommended by the FAANG coordinated action.

"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 Consortium (faang@iastate.edu) to enquire. The full guidelines can be found at http://www.faang.org/data-share-principle."

This license statement will be available in the description field of all submitted datasets and thus accessible to web and programmatic users. Whilst the FAANG metadata is fully machine readable and the license is available to both web and programmatic users, further improvements will be investigated to further improve the machine readability. The FAANG DCC will investigate specific license API endpoints, html embedding of license links and license structure formatting to improve machine-based access.

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

Data will be available for direct download from the FAANG data portal (utilising underlying public archives) and from the public archives themselves. The archives make data available in a range of access and data transfer methods including web browser download, FTP, Aspera, Globus and API access. All of these download options are open source and the archives have extensive documentation on the various data access options. The FAANG data portal that collates all FAANG data from the various underlying archives includes supports web browser bulk data download, it will be easy to obtain all AQUA-FAANG data from the data portal. The FAANG API provides programmatic users with the access FTP addresses to make a secondary call to download the data files themselves.

Sample metadata will be deposited in the EMBL-EBI BioSamples archive (https://www.ebi.ac.uk/biosamples/) with protocols stored in the FAANG protocol FTP server (http://ftp.faang.ebi.ac.uk/ftp/protocols). Raw experimental data will be submitted to the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home), and European Variation Archive (https://www.ebi.ac.uk/eva/). Data from across the EMBL-EBI archives will be indexed within the FAANG data portal (https://data.faang.org/home) and eventually collated into a specific AQUA-FAANG project page containing all AQUA-FAANG data at the proposed address of (https://data.faang.org/projects/aqua-faang). AQUA-FAANG code will be deposited in the FAANG GitHub repository (https://github.com/FAANG) in publicly available repositories licensed under Apache 2, each repository will have the prefix 'project-aquafaang-'.

2.3. Making data interoperable

Ensuring that the data produced by AQUA-FAANG is interoperable is the responsibility of the FAANG DCC at EMBL-EBI. AQUA-FAANG will fully comply with the FAANG metadata standards for samples, experiments and analyses, and will as part of the project improve the standards for their application to fish species and sampling. The project will use the standard vocabulary and ontologies outlined by the FAANG Metadata and Data Sharing working group, as implemented by the FAANG DCC in the metadata standards. It will ensure its compliance with these standards by running all data through the FAANG validation software prior to submission to the public archives. Beyond a shared metadata standard, AQUA-FAANG through the FAANG DCC and FAANG working groups will liaise with other FAANG projects to ensure as far as possible that final datasets will be comparable. This includes agreements on the analyses being run, contribution to decisions on or shared analysis pipelines, file formats and protocol strategies. For pipeline development, AQUA-FAANG will strive wherever possible to use openly licensed software.

AQUA-FAANG will fully comply with the FAANG sample, experiment and analysis metadata standards (https://data.faang.org/ruleset/samples#standard). This involves utilising FAANG approved ontologies for metadata values and producing detailed protocols to accompany data submissions. The data submissions will also match the standards set by the public archives for mandatory fields and controlled values. An important component of the project will be to contribute improvements to the ontologies of importance to the project, particularly in identifying missing terms or inaccurate descriptions in terms of fish species. No project specific

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

ontologies will be used, the widely supported ontologies that will be used within AQUA-FAANG are:

OBI	https://www.ebi.ac.uk/ols/ontologies/obi
NCBI Taxonomy	https://www.ebi.ac.uk/ols/ontologies/ncbitaxon
EFO	https://www.ebi.ac.uk/ols/ontologies/efo
LBO	https://www.ebi.ac.uk/ols/ontologies/lbo
PATO	https://www.ebi.ac.uk/ols/ontologies/pato
VT	https://www.ebi.ac.uk/ols/ontologies/vt
ATOL	https://www.ebi.ac.uk/ols/ontologies/atol
EOL	https://www.ebi.ac.uk/ols/ontologies/eol
UBERON	https://www.ebi.ac.uk/ols/ontologies/uberon
CL	https://www.ebi.ac.uk/ols/ontologies/cl
BTO	https://www.ebi.ac.uk/ols/ontologies/bto
CLO	https://www.ebi.ac.uk/ols/ontologies/clo
SO	https://www.ebi.ac.uk/ols/ontologies/so
GO	https://www.ebi.ac.uk/ols/ontologies/go
NCIT	https://www.ebi.ac.uk/ols/ontologies/ncit
CHEBI	https://www.ebi.ac.uk/ols/ontologies/chebi

2.4. Increase data re-use (through clarifying licences)

AQUA-FAANG sample and raw data will be made rapidly publicly available without embargo to the EMBL-EBI public archives. All data will include in its description an accompanying data reuse statement as recommended by the FAANG coordinated action.

"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 Consortium (faang@iastate.edu) to enquire. The full guidelines can be found at http://www.faang.org/data-share-principle."

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

This statement outlines the principles of AQUA-FAANG data reuse, and the requirements of pre-publication use of the data. Analysis data generated by Ensembl (based at EMBL-EBI), that comprises new or revised functionally annotated genomes, complete with analysis tracks, will be scheduled for upcoming Ensembl data releases, which occur approximately every 3 months. This complies with FAANG's expectation for analysis data of multiple releases per year in the FAANG data sharing statement. Any stand-alone analysis files submitted to the public archives from Ensembl will also include the above FAANG data reuse statement. All data will be free to use by third parties without restriction after the end of the project or upon first publication in a global analysis of the data by AQUA-FAANG.

All stand-alone software packages and pipelines produced by AQUA-FAANG will be made publicly available through GitHub (https://github.com/) under a permissive Apache 2.0 licence. All repositories will display this license in the root folder of the repository in a file named 'LICENSE'.

All AQUA-FAANG data will be assessed with the latest guidelines on quality assurance, comply with directives of the public archives and with any quality guidance from the FAANG coordinated action if they are set within the lifetime of the project. Data will remain usable by the scientific community for the lifetime of the public archives at EMBL-EBI. It is expected that future projects will continue to generate new data that will feed into further improvements to the six species functional annotation maps created by AQUA-FAANG, but the data generated during this project will have long lasting impact on these genomes' construction, functional annotation and commercial and scientific impact.

3. Allocation of resources

Ensuring good data management practice is the responsibility of the FAANG Data Coordination Centre at EMBL-EBI, whose daily operations are led by Peter Harrison with overall responsibility by Paul Flicek (AQUA-FAANG WP2 lead). Other institution partners within the project will be responsible for the storage (at least temporarily) of raw, intermediate and analysed data until the raw and final analysed datasets are submitted to the public archives via the FAANG DCC. These partners will be responsible for their own infrastructure to ensure safe storage of these datasets, including appropriate backup systems. WP4 and WP5 will also generate datasets that go beyond the core FAANG raw and analysed assays typically processed by the FAANG DCC, the partners of these work packages will be responsible for the temporary storage of these datasets until they are submitted to appropriate permanent external repositories. AQUA-FAANG has specific tasks and deliverables for ensuring good data management, curation and governance of AQUA-FAANG data within work package 2. AQUA-FAANG will utilise and contribute to the improvement of the FAANG metadata rulesets, validation and submission software and accessibility of data through FAANG data portal and programmatic interfaces. This will be of benefit to the global FAANG community.

AQUA-FAANG has committed to make all generated data publicly available and ensure it meets FAIR data standards. The commitment to this policy is set by the AQUA-FAANG Executive Committee and will be implemented by the FAANG DCC at EMBL-EBI. The cost for ensuring

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

that the submitted datasets comply with FAIR policy has been factored into the FAANG Data Coordination Centre activities of work package 2. This covers the development of metadata standards and ontologies to support fish species in FAANG, the development of validation and submission procedures for highly discoverable data to be placed into the public archives, development of the indexing and public availability of data through the EMBL-EBI public archives and developments to the web search and programmatic interfaces of the FAANG data portal. The cost requirement is primarily personnel for development and curation that AQUA-FAANG has allocated to EMBL-EBI under WP2. AQUA-FAANG has also allocated funds for workshops that bring together key scientists and developers from the FAANG H2020 EU projects to harmonise practice on data recording, pipeline development and comparative analysis. The data generation and analysis work packages of AQUA-FAANG also have costs allocated for the accurate recording of metadata from sampling activities, sequencing and analysis, and the preparation of data files for submission to public archives.

The AQUA-FAANG functionally annotated genomes, analysis data files and raw omics data will have long term value to the community. As all data will be deposited in the EMBL-EBI public archives, there is no long-term cost consideration for the storage of data for AQUA-FAANG, as long-term storage is guaranteed by the EMBL-EBI public archives that are separately funded. The long-term and post-AQUA-FAANG curation of data will also be secured through the Ensembl genome browser, with costs covered by EMBL-EBI's involvement. The Ensembl browser was launched in 2000 and is currently one of the world's most widely used sources for accessing and viewing genome information for all species. The Ensembl visualization framework is currently in process of being redesigned using modern web frameworks and visualizations. This development will be completed at no cost to AQUA-FAANG, but the design team will engage other AQUA-FAANG partners to ensure that the visualizations created will benefit both the project itself and external users of the data.

4. Data security

Sample, experiment and analysis data will be submitted to the EMBL-EBI public archives at the earliest opportunity, with AQUA-FAANG committed to early pre-publication data sharing. The European Nucleotide Archive (ENA), BioSamples, European Variation Archive (EVA), BioStudies and BioImage archive are highly regarded and recognised repositories for the long-term preservation and curation of data. Each utilise persistent and unique identifiers for data objects and commit to long term preservation and access to datasets.

The EMBL-EBI public archives, that will host the AQUA-FAANG data, are located in state-of-the-art technical architecture distributed in three discrete Tier III plus data centres in different geographical locations to ensure long-term security. This provides AQUA-FAANG data with high protection through redundancy. The archives also form part of the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/) whereby raw reads, alignments and assemblies from AQUA-FAANG submitted to the ENA, along with their contextual metadata from the BioSamples archive, will be synced and made available from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and DNA Databank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/index-e.html) supplying even greater data redundancy and availability to users. The three institutions that form the INSDC

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

collaboration recognise each other's unique identifiers and make datasets submitted to any one centre discoverable from the web and programmatic interfaces of all three.

The EMBL-EBI public archives commit to store AQUA-FAANG data for the lifetime that the archives remain active, even if the FAANG data portal ceases to exist, the underlying data will remain available. Due to the commitment AQUA-FAANG places on extensive validated metadata, a standardised sample name format, a FAANG metadata project tag and persistent and unique identifiers, the AQUA-FAANG data will be discoverable in the EMBL-EBI public archives even if the indexing provided by the FAANG data portal ceases.

5. Ethical aspects

The AQUA-FAANG proposal meets the national legal and ethical requirements of the countries in which the tasks raising potential ethical issues are undertaken. The research will be undertaken fully within applicable international, EU and national laws. The AQUA-FAANG project does not have any known ethical or legal issues with data sharing, and informed consent was not required for data sharing or storing of personal data.

6. Other issues

The AQUA-FAANG project forms part of the Functional Annotation of Animal Genomes (FAANG) coordinated action. AQUA-FAANG within this Data Management Plan therefore fully complies with the requirements set out in the FAANG Data Sharing Statement, approved by the FAANG Steering Committee on 26th May 2015 (https://www.faang.org/data-share-principle). This in turn requires AQUA-FAANG to comply with the data sharing commitments of the Fort Lauderdale principles and Toronto Workshop statement. The requirements from FAANG for early access open data and FAIR data principles do not conflict with any EU H2020 data management requirements.

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

Annex 2. FAANG Data Sharing Statement

This document describes the principles of data sharing held by the FAANG consortium. This document is subject to approval by the FAANG steering committee. Any queries about this document should be sent to faang@iastate.edu.

FAANG recognizes that quickly sharing the data generated by the consortium with the wider community is a priority. Rapid data sharing before publication ensures that everyone can benefit from the data created by FAANG and can take advantage of improved understanding of the functional elements in these animal genomes to aid their own research.

All raw data produced for a FAANG associated project will be submitted to the archives without any hold until publication date, thus allowing the data to be publicly available immediately after successful archive submission and useful to the community as soon as possible.

The FAANG analysis group will turn the raw data into primary and integrated analysis results. Primary analysis results consistent of sample level analysis such as alignment to a reference genome or quantification of signal in the assay. Integrated analysis results represent analyses which drawn together data from multiple samples and/or experiments such as genome segmentation or differential analysis results.

Definitions

Archive means one of the archives hosted at the EBI, NCBI or DDBJ. These include the ENA, Genbank, ArrayExpress and Geo. A full list of the FAANG recommended archives is available as part of the FAANG metadata recommendations.

Submission means data and metadata submission to one of the FAANG recommended Archives.

FAANG member means an individual who has signed up to the FAANG consortium through the FAANG website and agreed to the FAANG core principles.

Data means any assay or metadata generated for or associated with FAANG experiments.

Analysis means any computational process where raw assay data is aligned, transformed or combined to produce a new product.

Internal means data that is only accessible via the FAANG private shared storage.

Private shared storage means a storage space hosted at EMBL-EBI that has password access via FTP, aspera and Globus Grid FTP technologies.

Public means all data available through the FAANG public FTP site, which has no password and is accessible to everyone.

The majority of these analysis results will not be archived before publication but FAANG recognizes the need to share them both within the consortium and with the community. Initially all files that are not archived will be shared between FAANG members in private shared storage hosted at the EMBL-EBI. Any individual who signs up to FAANG and agrees

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

to **the Toronto principles** [1] will be allowed access to this. There will be metadata files in the private data sharing area, which make credit for different datasets as clear as possible.

FAANG expects to make multiple releases each year. A data release will involve declaring a data freeze and copying all files associated with that data freeze from the private shared storage to the public FTP site. In the first instance these data freezes will contain the primary analysis results. As FAANG's analyses progress, the data freeze will be expanded to include integrative analysis too. The data freeze process will be coordinated by the FAANG Data Coordination Centre and will be based on consultation with FAANG members. FAANG will also aim to release all data associated with a paper before publication even if it lies outside this standard freeze cycle. The public data will be available to the whole community.

All FAANG public data is released under **Fort Lauderdale principles** [2]. The FAANG website, data portal and FTP site will all have clear data reuse statements on them.

When considering internal FAANG data, if one FAANG member wishes to publish using data generated by another FAANG member they should first contact the data generator and clarify the member's publication strategy. Collaboration is for everyone's benefit and is strongly encouraged. The FAANG Steering Committee commits to report to journal editors and the laboratories involved any event that disregards the rights of data creators (including biological measurements as well as analysis of such measurements).

All members of FAANG can and will continue to do experimental and analysis work outside of FAANG and the other data generated is not required to meet the same data sharing expectations.

Only FAANG data can be distributed through the private storage and public FTP site.

REFERENCES:

1. **Toronto International Data Release Workshop**: Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.
2. **Fort Lauderdale principles**: Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-scale Sequencing and Other Community Resource Projects.

(Approved by the FAANG Steering Committee on May 26, 2015)

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