GENE-SWitCH

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

Deliverable D3.3
Fully functional data coordination platform for production use

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

Version: 1.1

Due date of deliverable (as in DoA):	M18
Actual submission date:	24/06/2021, month M24

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 website (DEC) report we detail the fully functional FAANG Data Coordination platform that has been developed for the production, submission and presentation of GENE-SWitCH datasets. This report will provide the website links, screenshots and descriptions of the different components that comprise the Data Coordination platform. The platform is in production use with, at the time of this report, eleven datasets in the GENE-SWitCH project page of the coordination platform (https://data.faang.org/projects/GENE-SWitCH) that comprise an array of different assay types for both chickens and pigs.

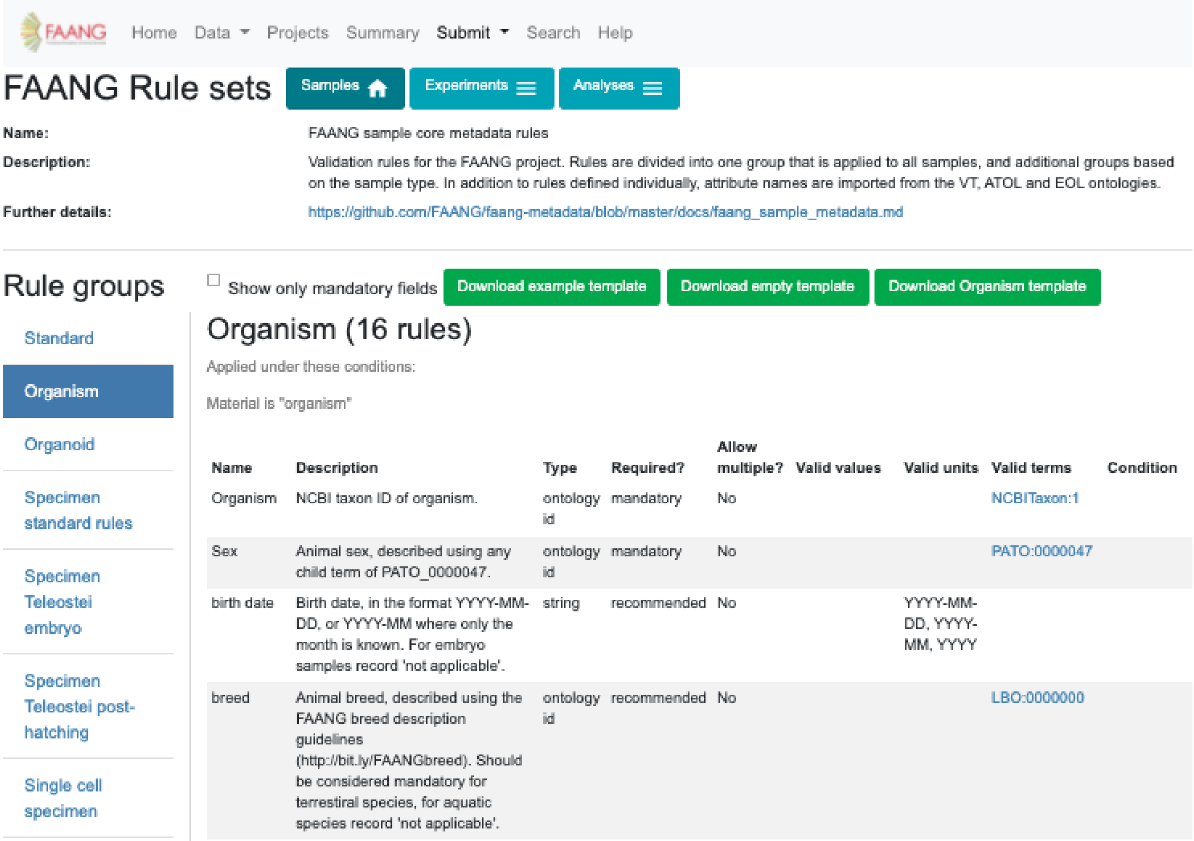
2 The FAANG Data Coordination Platform

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. The Data Coordination Platform has now been completely overhauled to support the data submission and presentation requirements of GENE-SWitCH. This report documents the different components that comprise the FAANG Data Coordination Platform (https://data.faang.org/) that is in production use by the project.

2.1 Metadata rulesets

https://data.faang.org/ruleset/samples
https://data.faang.org/ruleset/experiments
https://data.faang.org/ruleset/analyses

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 which consortium member and in what location data is generated. The metadata standards have undergone an in-depth review and update in response to the new requirements of the EuroFAANG projects (see https://www.gene-switch.eu/eurofaang.html), with updated rulesets published for FAANG in the underlying GitHub and portal presentation pages (https://data.faang.org/ruleset/samples). The metadata pages provide a complete overview of the FAANG requirements for sample, experiment, and analysis submissions. Importantly this details whether the information is mandatory and whether controlled terminology or ontological limits are applied (Figure 1). This resource is crucial for consortium members to prepare data submissions to FAANG and ensuring that they collect and submit the required metadata information.



FAANG Rule sets

Name:	FAANG sample core metadata rules
Description:	Validation rules for the FAANG project. Rules are divided into one group that is applied to all samples, and additional groups based on the sample type. In addition to rules defined individually, attribute names are imported from the VT, ATOL and EOL ontologies.
Further details:	https://github.com/FAANG/faang-metadata/blob/master/docs/faang_sample_metadata.md

Rule groups

□ Show only mandatory fields

Download example template
Download empty template
Download Organism template
Standard
Organism
Organoid
Specimen standard rules
Specimen Teleostei embryo
Specimen Teleostei post-hatching
Single cell specimen

Organism (16 rules)

Applied under these conditions:

Material is "organism"

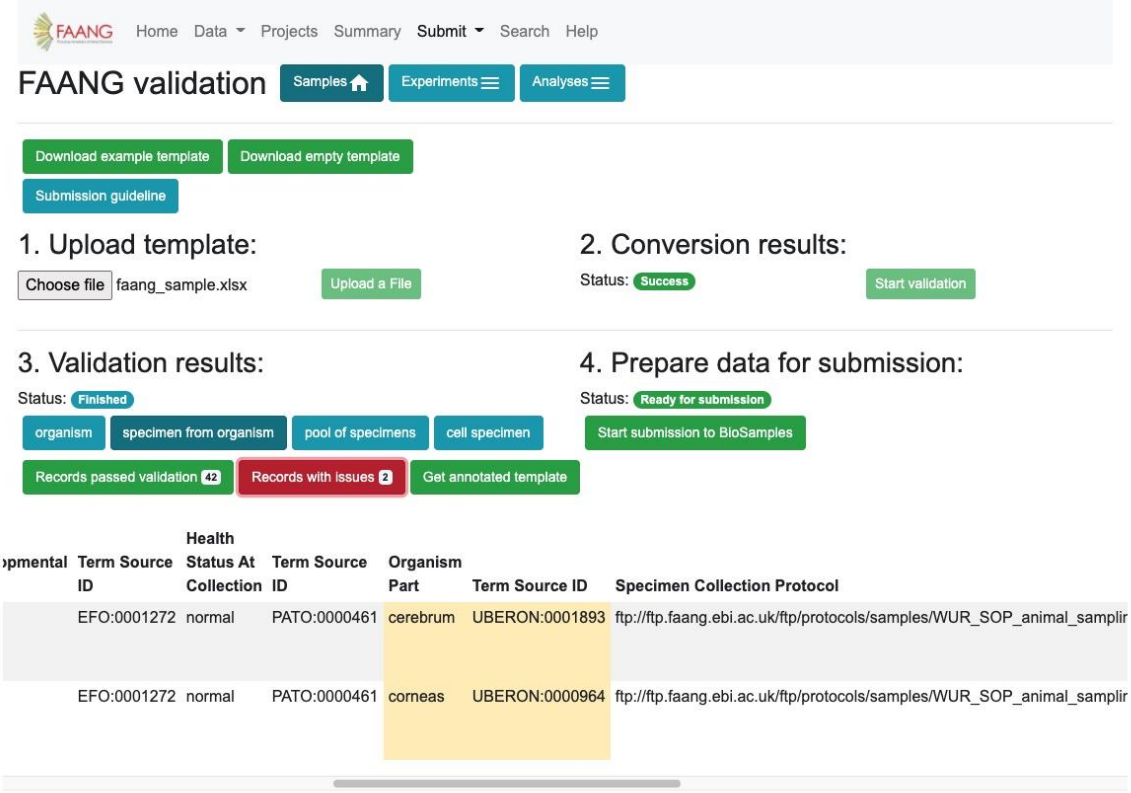
Name	Description	Type	Required?	Allow multiple?	Valid values	Valid units	Valid terms	Condition
Organism	NCBI taxon ID of organism.	ontology id	mandatory	No			NCBITaxon:1	
Sex	Animal sex, described using any child term of PATO_0000047.	ontology id	mandatory	No			PATO:0000047	
birth date	Birth date, in the format YYYY-MM-DD, or YYYY-MM where only the month is known. For embryo samples record 'not applicable'.	string	recommended	No		YYYY-MM-DD, YYYY-MM, YYYY		
breed	Animal breed, described using the FAANG breed description guidelines (http://bit.ly/FAANGbreed). Should be considered mandatory for terrestiral species, for aquatic species record 'not applicable'.	ontology id	recommended	No			LBO:0000000	

Figure 1. The metadata ruleset guide pages of the FAANG Data Coordination Platform, including different metadata rule groups, level of requirement and controlled terminology.

2.2 Validation and brokered submission

https://data.faang.org/validation/samples
https://data.faang.org/validation/experiments
https://data.faang.org/validation/analyses

Metadata rulesets are only useful if submitters are enabled and required to meet them. The FAANG Data Coordination Centre has released a new validation and submission process for the submission of GENE-SWitCH metadata, incorporating improved validation error handling and an updated user interface (https://data.faang.org/validation/samples; Figure 2). This makes the submission process and provision of rich FAIR (Findable, Accessible, Interoperable and Reusable) metadata easier for the EuroFAANG projects. The service takes a completed metadata spreadsheet from a GENE-SWitCH submitter, converts it to a JSON file format, and conducts a range of detailed validation checks. The service not only provides errors for metadata inaccuracies, but also provides warnings and suggestions for improvements that the users could make to their metadata submissions such as being more specific with ontologies. Once validated, the service takes the users archive submission credentials and submits the metadata to the underlying archive on the users behalf.



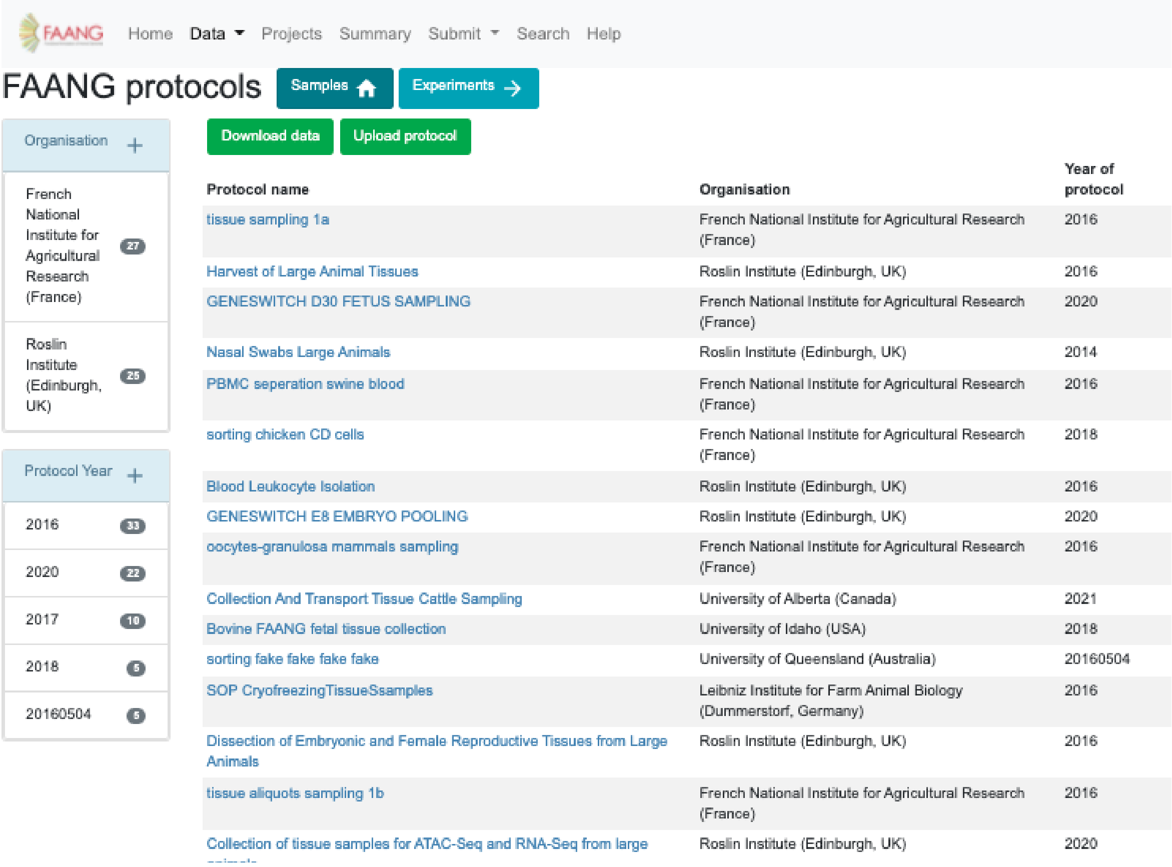
›pmental	Term Source ID	Health Status At Collection	Term Source ID	Organism Part	Term Source ID	Specimen Collection Protocol
	EFO:0001272	normal	PATO:0000461	cerebrum	UBERON:0001893	ftp://ftp.faang.ebi.ac.uk/ftp/protocols/samples/WUR_SOP_animal_samplir
	EFO:0001272	normal	PATO:0000461	corneas	UBERON:0000964	ftp://ftp.faang.ebi.ac.uk/ftp/protocols/samples/WUR_SOP_animal_samplir

Figure 2. Validation and brokered submission component of the FAANG Data Coordination Platform.

2.3 Protocol upload and display

https://data.faang.org/protocol/samples

The EuroFAANG DCC also released a new protocol upload interface, to make it easier for the EuroFAANG projects to upload their sampling, sequencing, and analysis protocols to FAANG, in particular for sharing with the other EuroFAANG 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). Protocols used within the GENE-SWitCH project are attached to the datasets so that the protocol can be obtained alongside the dataset.



FAANG protocols

Organisation	
French National Institute for Agricultural Research (France)	27
Roslin Institute (Edinburgh, UK)	25

Protocol Year	
2016	33
2020	22
2017	10
2018	5
20160504	5

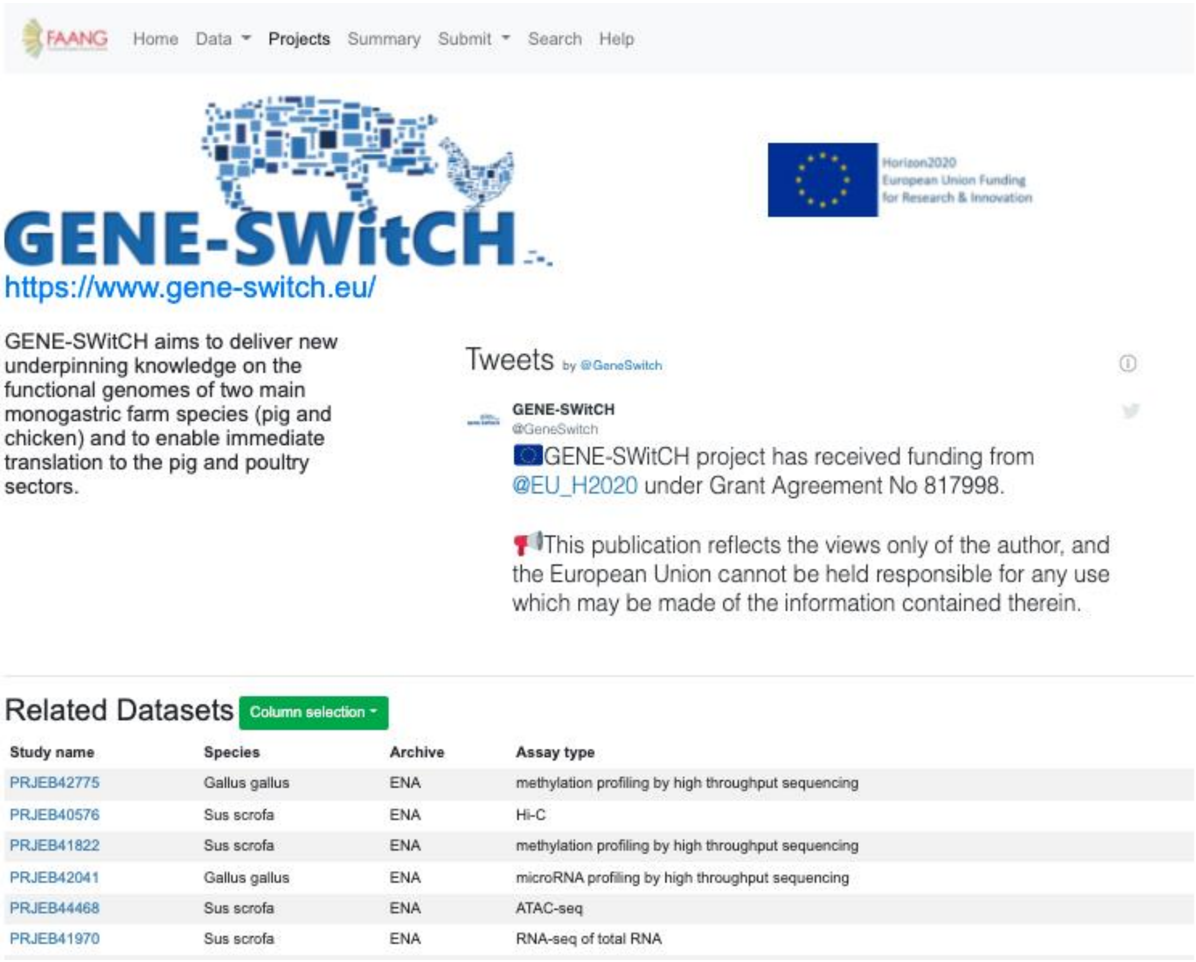
Protocol name	Organisation	Year of protocol
tissue sampling 1a	French National Institute for Agricultural Research (France)	2016
Harvest of Large Animal Tissues	Roslin Institute (Edinburgh, UK)	2016
GENESWITCH D30 FETUS SAMPLING	French National Institute for Agricultural Research (France)	2020
Nasal Swabs Large Animals	Roslin Institute (Edinburgh, UK)	2014
PBMC seperation swine blood	French National Institute for Agricultural Research (France)	2016
sorting chicken CD cells	French National Institute for Agricultural Research (France)	2018
Blood Leukocyte Isolation	Roslin Institute (Edinburgh, UK)	2016
GENESWITCH E8 EMBRYO POOLING	Roslin Institute (Edinburgh, UK)	2020
oocytes-granulosa mammals sampling	French National Institute for Agricultural Research (France)	2016
Collection And Transport Tissue Cattle Sampling	University of Alberta (Canada)	2021
Bovine FAANG fetal tissue collection	University of Idaho (USA)	2018
sorting fake fake fake fake	University of Queensland (Australia)	20160504
SOP CryofreezingTissueSsamples	Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany)	2016
Dissection of Embryonic and Female Reproductive Tissues from Large Animals	Roslin Institute (Edinburgh, UK)	2016
tissue aliquots sampling 1b	French National Institute for Agricultural Research (France)	2016
Collection of tissue samples for ATAC-Seq and RNA-Seq from large	Roslin Institute (Edinburgh, UK)	2020

Figure 3. Protocol browser within the FAANG Data Coordination Platform.

2.4 GENE-SWitCH presentation platform

https://data.faang.org/projects/GENE-SWitCH

The GENE-SWitCH project presentation platform is a fully customisable view of the FAANG data specific to the GENE-SWitCH project. It contains a project description, logo, funding logo and automated twitter stream. Data that is known to be part of GENE-SWitCH through the setting of the secondary project tag on data submission is automatically synced to this portal view. The page has interactive tables of GENE-SWitCH publications, datasets, files, organisms, and specimens.



https://www.gene-switch.eu/

GENE-SWitCH aims to deliver new underpinning knowledge on the functional genomes of two main monogastric farm species (pig and chicken) and to enable immediate translation to the pig and poultry sectors.

Tweets by @GeneSwitch

GENE-SWitCH
@GeneSwitch

GENE-SWitCH project has received funding from @EU_H2020 under Grant Agreement No 817998.

This publication reflects the views only of the author, and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Related Datasets

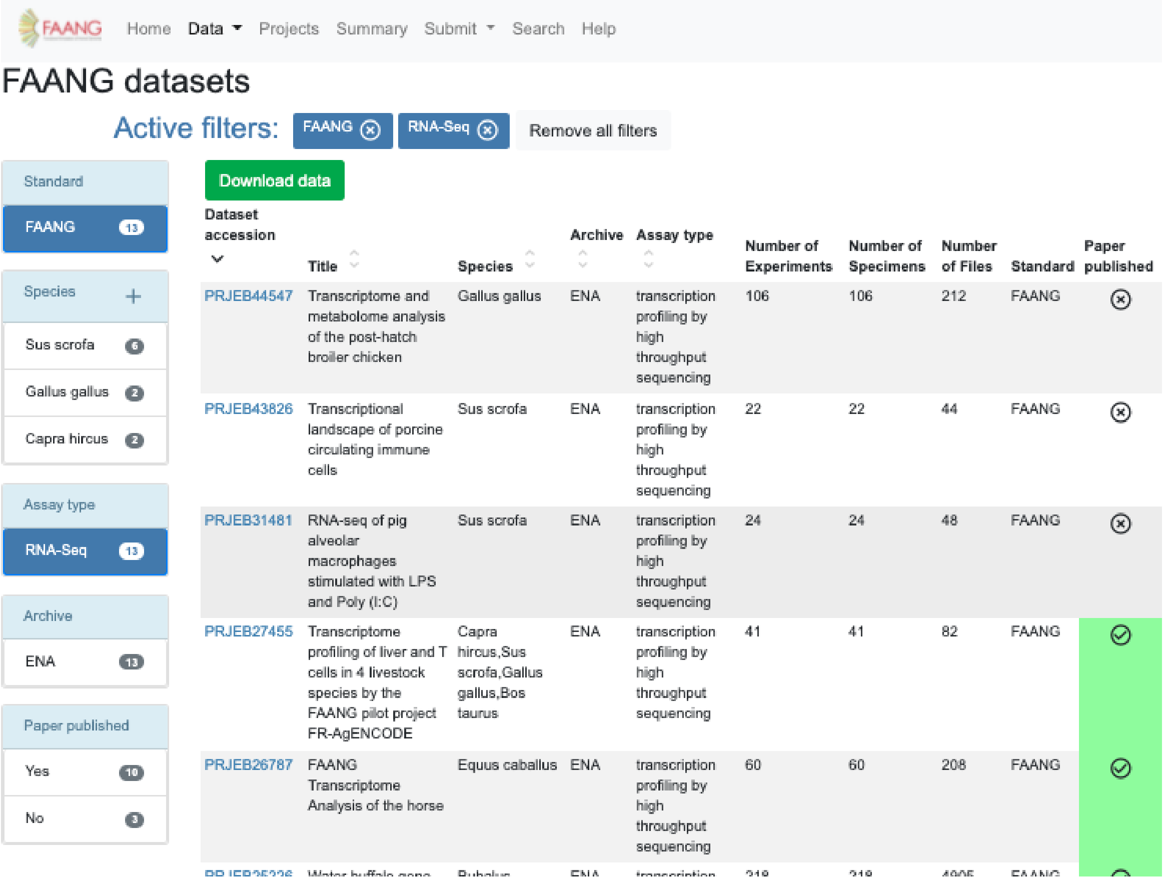
Study name	Species	Archive	Assay type
PRJEB42775	Gallus gallus	ENA	methylation profiling by high throughput sequencing
PRJEB40576	Sus scrofa	ENA	Hi-C
PRJEB41822	Sus scrofa	ENA	methylation profiling by high throughput sequencing
PRJEB42041	Gallus gallus	ENA	microRNA profiling by high throughput sequencing
PRJEB44468	Sus scrofa	ENA	ATAC-seq
PRJEB41970	Sus scrofa	ENA	RNA-seq of total RNA

Figure 4. GENE-SWitCH project view of the FAANG Data Coordination Platform.

2.5 FAANG dataset portal pages

https://data.faang.org/dataset

The data portal pages provide interactive tables of the organisms, specimens, datasets, and files of FAANG data. Preconfigured filters are provided to identify data of interest by filtering by fields such as species, assay type, breed, and instrument. Each record has links to a full page with additional metadata detail and links to datasets in the underlying archive. The pages provide summary files of the tabulated metadata that can be downloaded. The datasets can be sorted by different metadata columns.



FAANG datasets

Active filters: FAANG ⊗ RNA-Seq ⊗ Remove all filters

Standard
FAANG 13

Species +
Sus scrofa 6
Gallus gallus 2
Capra hircus 2

Assay type
RNA-Seq 13

Archive
ENA 13

Paper published
Yes 10
No 3

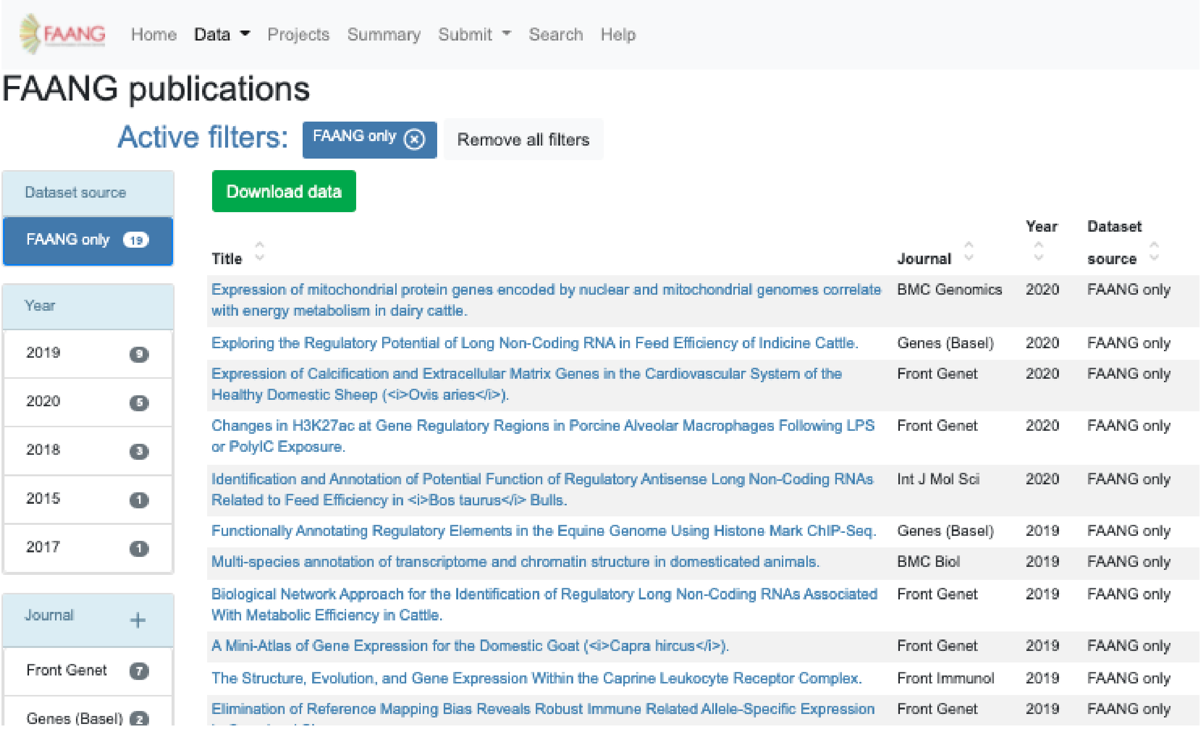
Dataset accession	Title	Species	Archive	Assay type	Number of Experiments	Number of Specimens	Number of Files	Standard	Paper published
PRJEB44547	Transcriptome and metabolome analysis of the post-hatch broiler chicken	Gallus gallus	ENA	transcription profiling by high throughput sequencing	106	106	212	FAANG	⊗
PRJEB43826	Transcriptional landscape of porcine circulating immune cells	Sus scrofa	ENA	transcription profiling by high throughput sequencing	22	22	44	FAANG	⊗
PRJEB31481	RNA-seq of pig alveolar macrophages stimulated with LPS and Poly (I:C)	Sus scrofa	ENA	transcription profiling by high throughput sequencing	24	24	48	FAANG	⊗
PRJEB27455	Transcriptome profiling of liver and T cells in 4 livestock species by the FAANG pilot project FR-AgENCODE	Capra hircus,Sus scrofa,Gallus gallus,Bos taurus	ENA	transcription profiling by high throughput sequencing	41	41	82	FAANG	✓
PRJEB26787	FAANG Transcriptome Analysis of the horse	Equus caballus	ENA	transcription profiling by high throughput sequencing	60	60	208	FAANG	✓
PRJEB25226	Water buffalo gene	Bubalus	ENA	transcription	218	218	4905	FAANG	✓

Figure 5. Data portal tables of the FAANG Data Coordination Platform.

2.6 Publication browser

https://data.faang.org/article

Publications on any FAANG dataset are automatically identified using the EMBL Literature services EuropePMC literature scraping technology (https://europepmc.org/). This identifies INSDC FAANG identifiers used within the main text of any publication, these publications are then imported into the FAANG data portal and linked to the dataset quoted. Links to the full text of the publication are included, and in the main data portal tables datasets that have publications associated with them are clearly labelled.



FAANG publications

Active filters: FAANG only ⊗ Remove all filters

Dataset source	
FAANG only	19

Year	
2019	9
2020	5
2018	3
2015	1
2017	1

Journal +	
Front Genet	7
Genes (Basel)	2

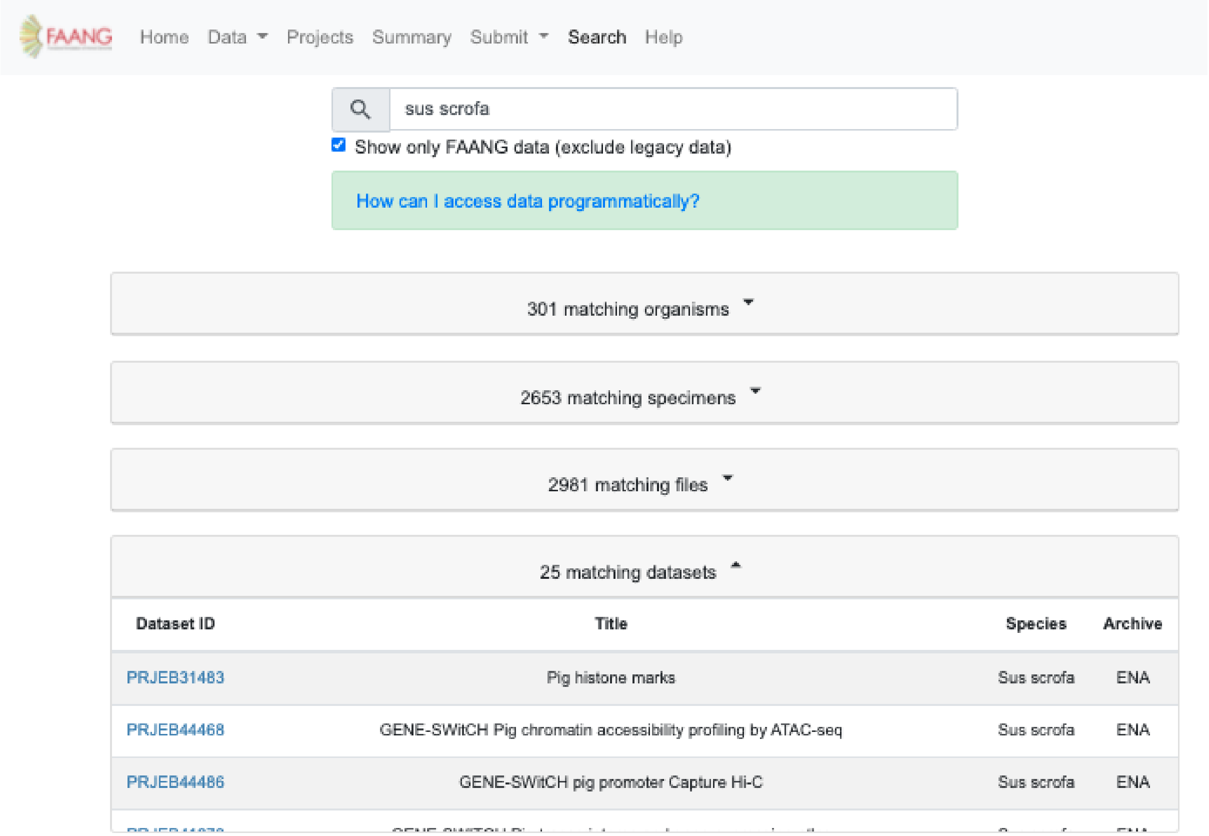
Title	Journal	Year	Dataset source
Expression of mitochondrial protein genes encoded by nuclear and mitochondrial genomes correlate with energy metabolism in dairy cattle.	BMC Genomics	2020	FAANG only
Exploring the Regulatory Potential of Long Non-Coding RNA in Feed Efficiency of Indicine Cattle.	Genes (Basel)	2020	FAANG only
Expression of Calcification and Extracellular Matrix Genes in the Cardiovascular System of the Healthy Domestic Sheep (<i>Ovis aries</i>).	Front Genet	2020	FAANG only
Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages Following LPS or PolyIC Exposure.	Front Genet	2020	FAANG only
Identification and Annotation of Potential Function of Regulatory Antisense Long Non-Coding RNAs Related to Feed Efficiency in <i>Bos taurus</i> Bulls.	Int J Mol Sci	2020	FAANG only
Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq.	Genes (Basel)	2019	FAANG only
Multi-species annotation of transcriptome and chromatin structure in domesticated animals.	BMC Biol	2019	FAANG only
Biological Network Approach for the Identification of Regulatory Long Non-Coding RNAs Associated With Metabolic Efficiency in Cattle.	Front Genet	2019	FAANG only
A Mini-Atlas of Gene Expression for the Domestic Goat (<i>Capra hircus</i>).	Front Genet	2019	FAANG only
The Structure, Evolution, and Gene Expression Within the Caprine Leukocyte Receptor Complex.	Front Immunol	2019	FAANG only
Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression	Front Genet	2019	FAANG only

Figure 6. Publication browser of the FAANG Data Coordination Platform.

2.7 FAANG search interface

https://data.faang.org/search

The search interface provides a keyword search that utilises ElasticSearch tokenisation. It simultaneously searches across metadata from organisms, specimens, files and datasets. It has the option to exclude legacy data that does not met the full FAANG standards. Search results provide the number of hits, an interactive table with a limited summary view of the data and links to view the full information within the data portal pages.



sus scrofa

☑ Show only FAANG data (exclude legacy data)

How can I access data programmatically?

301 matching organisms

2653 matching specimens

2981 matching files

25 matching datasets

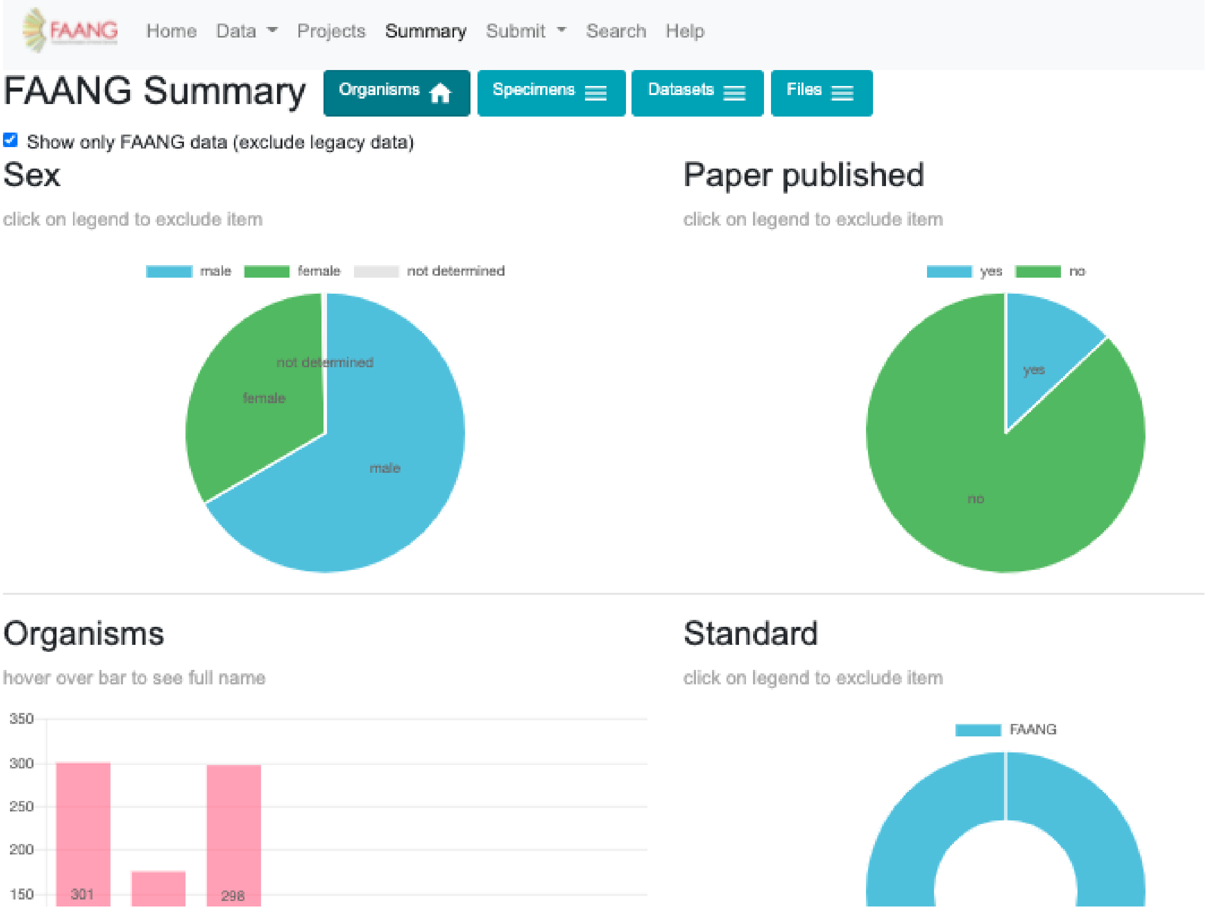
Dataset ID	Title	Species	Archive
PRJEB31483	Pig histone marks	Sus scrofa	ENA
PRJEB44468	GENE-SWitCH Pig chromatin accessibility profiling by ATAC-seq	Sus scrofa	ENA
PRJEB44486	GENE-SWitCH pig promoter Capture Hi-C	Sus scrofa	ENA

Figure 7. Search interface of the FAANG Data Coordination Platform.

2.8 Summary statistics

https://data.faang.org/summary/organisms

The summary statistics page of the FAANG Data Coordination Platform provides overall summary statistics on FAANG data for organisms, specimens, datasets and files (Figure 8). It provides statistic plots on sex, whether data is associated with a publication, organisms, legacy status, breeds, cell type and assay type.



☑ Show only FAANG data (exclude legacy data)

Figure 8. Summary statistics pages of the FAANG Data Coordination Platform.

3 Conclusion

The FAANG Data Coordination Platform has been developed and deployed to meet the data coordination, archiving and presentation requirements of the GENE-SWitCH project. The platform is in full production used with eleven datasets already coordinated, archived and on display in the data portal.

4 Deviations or delays

This deliverable was delayed by six months due to the secondment of EMBL DCC staff from March to September 2020 to support the emergency European COVID-19 scientific response to develop the European COVID-19 data portal (https://www.covid19dataportal.org/).

5 Glossary

DCC	Data Coordination Centre
EuroFAANG	The European node of the Functional Annotation of Animal Genomes Data Coordination Centre
EMBL	European Molecular Biology Laboratory
FAANG	Functional Annotation of Animal Genomes
FAIR	Findable, Accessible, Interoperbale and Reusable