The Functional Annotation of Animal Genomes Data Portal

Current and future perspectives for data reuse

Peter Harrison

Genome Analysis Team Leader

EMBL-European Bioinformatics Institute

peter@ebi.ac.uk

@peterwharrison

EMBL-EBI

FAANG
Functional Annotation of Animal Genomes

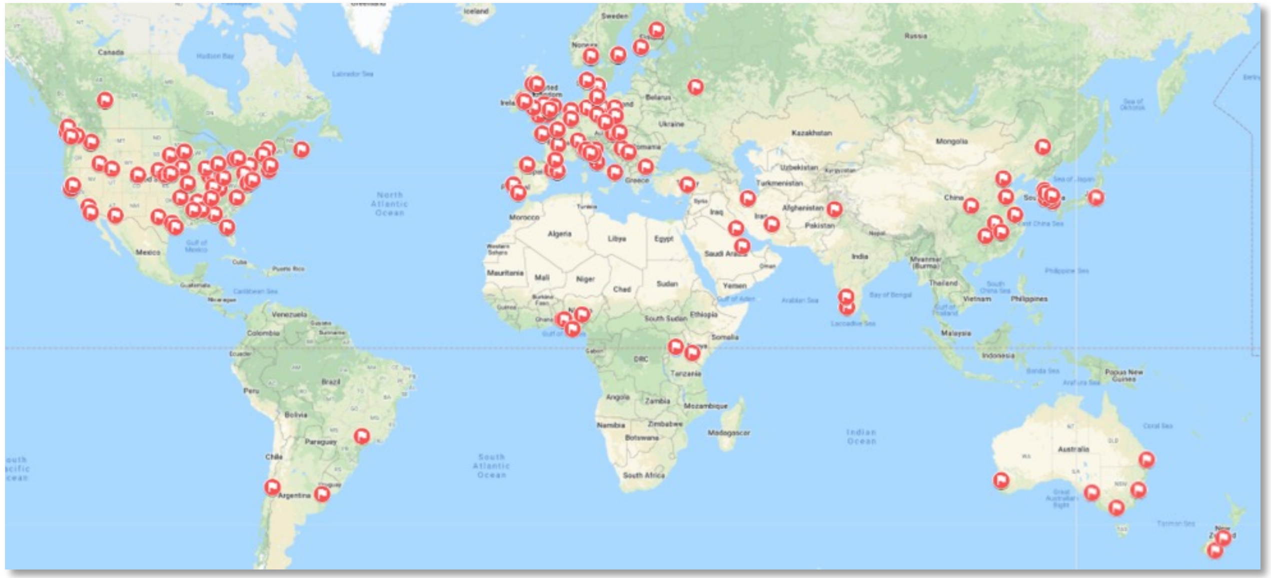
EuroFAANG

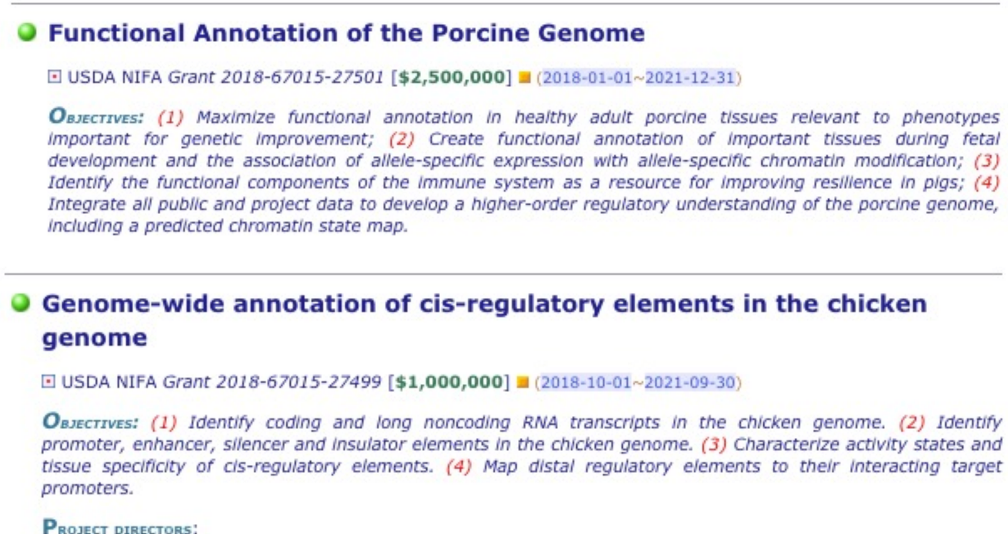
A coordinated international action to accelerate genome to phenome research

- Coordinated international effort to provide **high quality functional annotation** of animal genomes, with a focus on **livestock and aquaculture** communities, but extending to other animals.
- Core aims of data **openness, reusability, rich metadata and standardisation to** create a harmonised **rich genome to phenome resource**.
- EuroFAANG is a particularly coordinated effort within Europe to standardize our research processes to collectively improve animal production and welfare.

A global collaboration

- Steering committee, scientific advisory board, and working groups focus on range of key data reuse aspects.
- Comprised of multiple individual funded grants -> why coordination and standardisation is so important.
- FAANG projects collaborate and reuse existing genomic data to accelerate research and provide value to funders. Strong H2020 and USDA project data reuse.





Functional Annotation of the Porcine Genome

USDA NIFA *Grant 2018-67015-27501* [**$2,500,000**] (2018-01-01~2021-12-31)

Objectives: (1) Maximize functional annotation in healthy adult porcine tissues relevant to phenotypes important for genetic improvement; (2) Create functional annotation of important tissues during fetal development and the association of allele-specific expression with allele-specific chromatin modification; (3) Identify the functional components of the immune system as a resource for improving resilience in pigs; (4) Integrate all public and project data to develop a higher-order regulatory understanding of the porcine genome, including a predicted chromatin state map.

Genome-wide annotation of cis-regulatory elements in the chicken genome

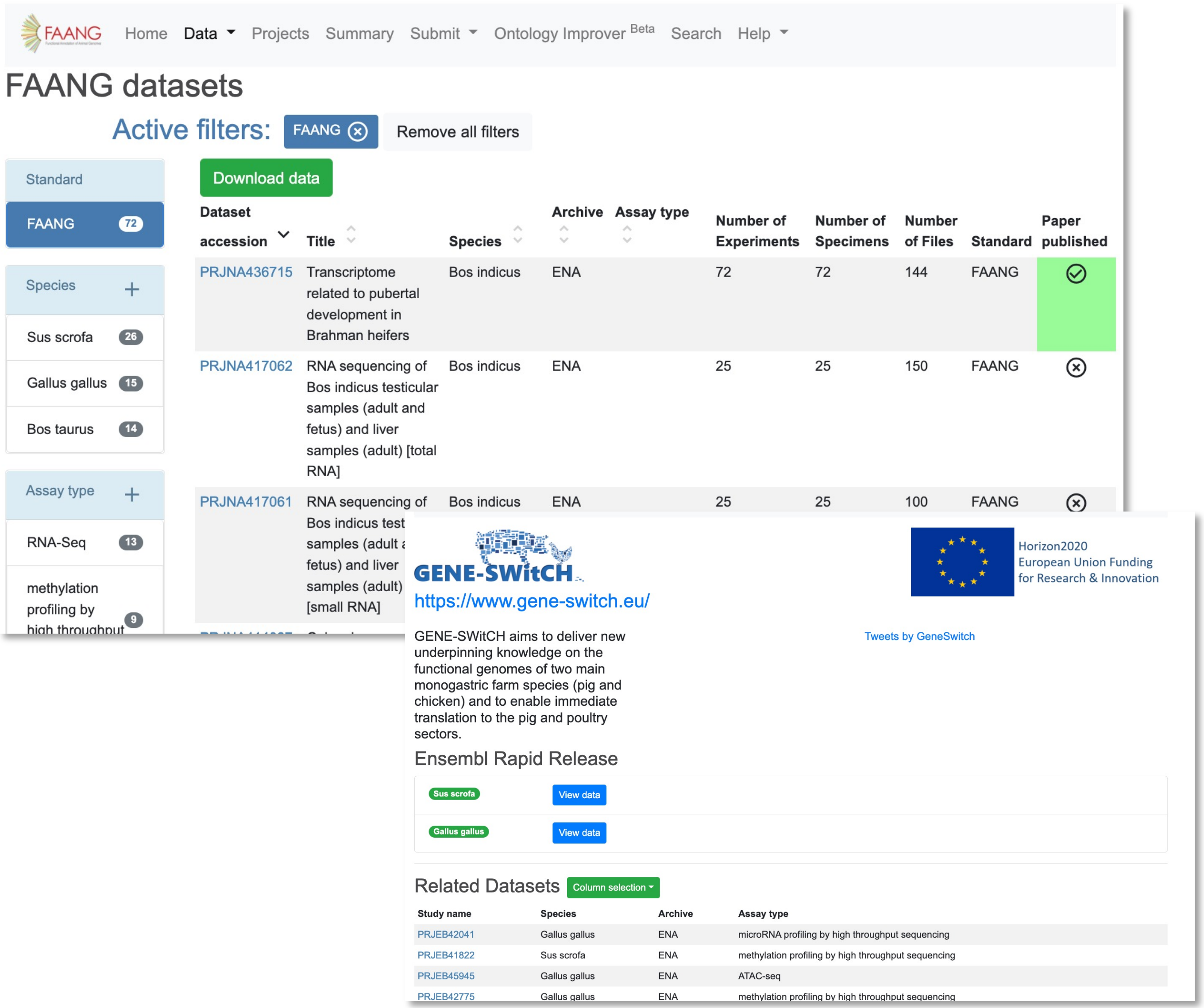
USDA NIFA *Grant 2018-67015-27499* [**$1,000,000**] (2018-10-01~2021-09-30)

Objectives: (1) Identify coding and long noncoding RNA transcripts in the chicken genome. (2) Identify promoter, enhancer, silencer and insulator elements in the chicken genome. (3) Characterize activity states and tissue specificity of cis-regulatory elements. (4) Map distal regulatory elements to their interacting target promoters.

Project directors:

The FAANG Data Portal

- A single access point to all FAANG metadata, data and publications.
- Providing direct access to download all data from various underlying public archives.
- Automatically identifies dataset (re)use in publications, and links these publications to datasets.
- Intuitive search and filtering.
- Has sub-project pages to access just that projects data.



FAANG datasets

Active filters: FAANG ⊗ Remove all filters

Standard

FAANG 72

Species +

Sus scrofa 26

Gallus gallus 15

Bos taurus 14

Assay type +

RNA-Seq 13

methylation profiling by high throughput 9

Dataset accession	Title	Species	Archive	Assay type	Number of Experiments	Number of Specimens	Number of Files	Standard	Paper published
PRJNA436715	Transcriptome related to pubertal development in Brahman heifers	Bos indicus	ENA		72	72	144	FAANG	✓
PRJNA417062	RNA sequencing of Bos indicus testicular samples (adult and fetus) and liver samples (adult) [total RNA]	Bos indicus	ENA		25	25	150	FAANG	⊗
PRJNA417061	RNA sequencing of Bos indicus test samples (adult a fetus) and liver samples (adult) [small RNA]	Bos indicus	ENA		25	25	100	FAANG	⊗

GENE-SWitCH

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

Horizon2020 European Union Funding for Research & Innovation

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.

Ensembl Rapid Release

Sus scrofa	View data
Gallus gallus	View data

Related Datasets

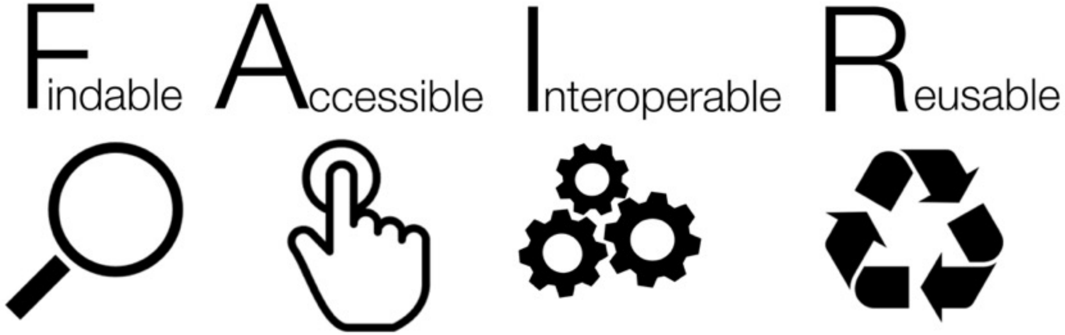
Study name	Species	Archive	Assay type
PRJEB42041	Gallus gallus	ENA	microRNA profiling by high throughput sequencing
PRJEB41822	Sus scrofa	ENA	methylation profiling by high throughput sequencing
PRJEB45945	Gallus gallus	ENA	ATAC-seq
PRJEB42775	Gallus gallus	ENA	methylation profiling by high throughput sequencing

https://data.faang.org/ https://data.faang.org/projects

EMBL-EBI

What makes the FAANG datasets special

- Rich, consistent and validated metadata descriptions.
- Standardised set of core assays from contributing projects.
- Mandatory sampling, experiment and analysis protocols connected to each dataset and available with the datasets for download.
- Many projects using standardised analysis pipelines attached to each dataset.
- A data platform and community drive that ensures data is open and FAIR.

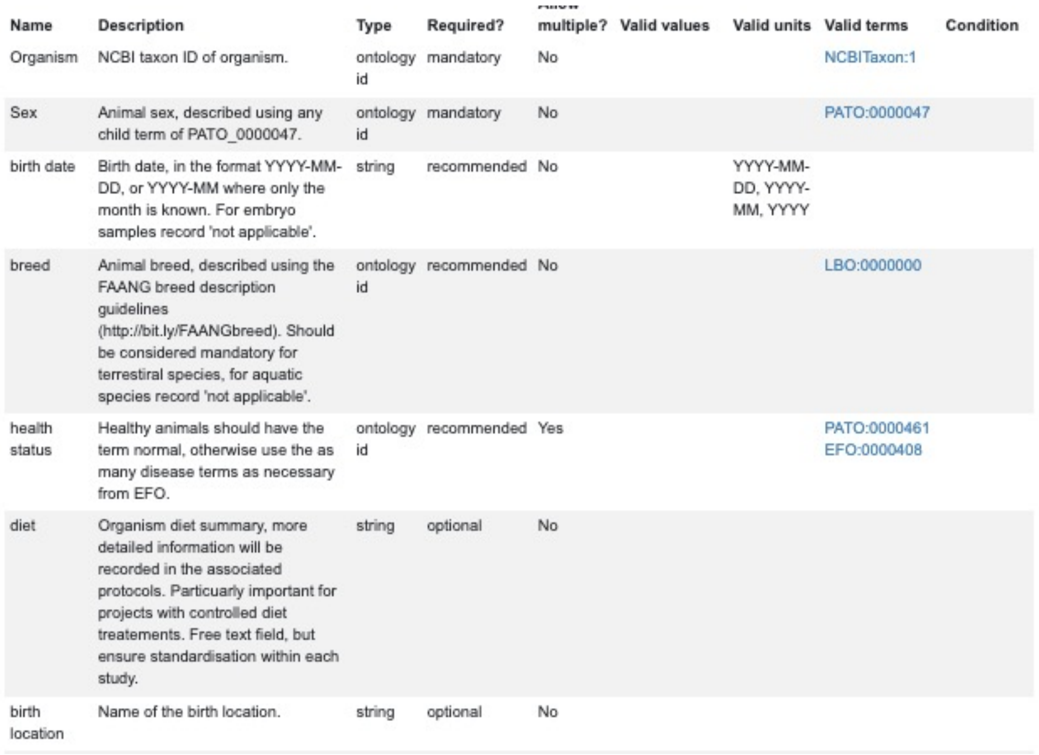


To accelerate genome to phenome research

Photo: CODATA

A full metadata solution for FAANG

- Requires **>200** different metadata questions for different studies.
- Constantly evolving with the community, recent changes include aquaculture, single cell sequencing and focus on developmental timepoints.
- Terminology controlled through standardised ontologies to make downstream search and analysis more powerful. Drives portal data filters.

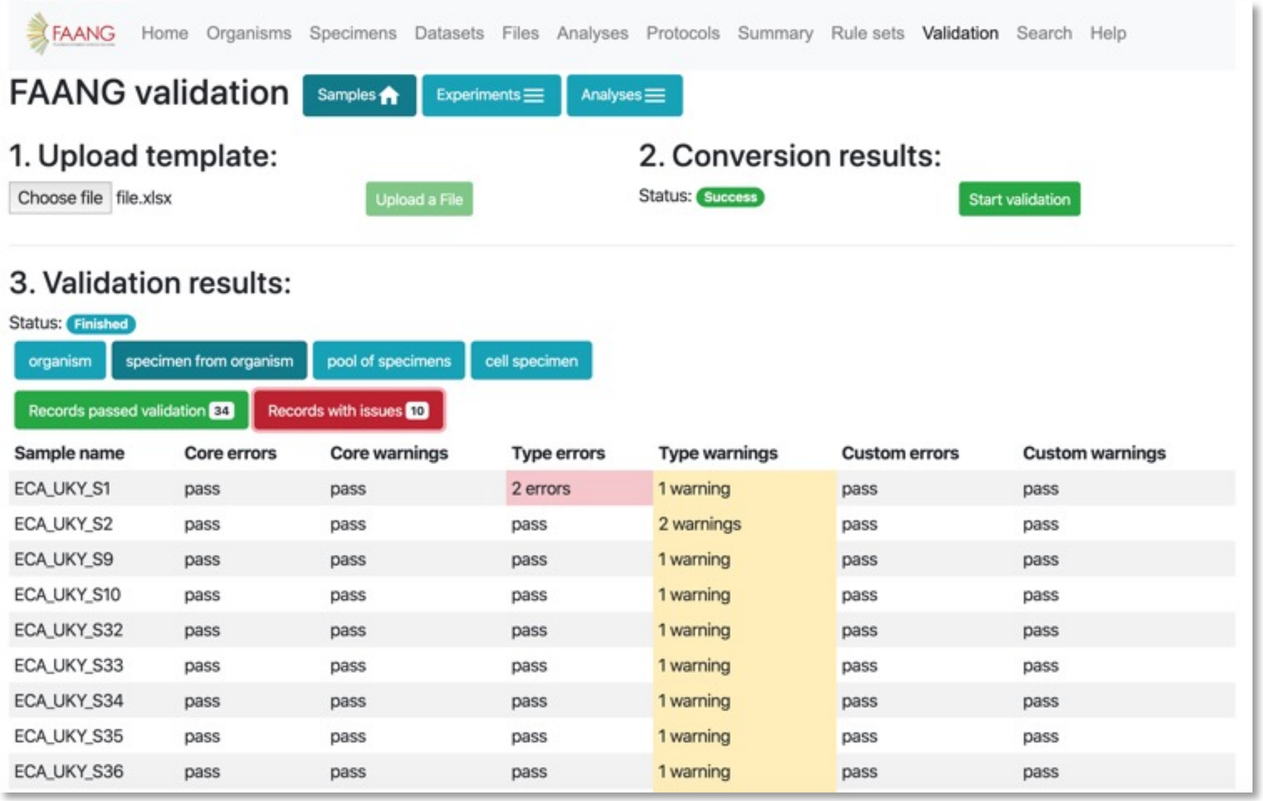


Name	Description	Type	Required?	Allow multiple?	Valid values	Valid units	Valid terms	Condition
Organism	NCBI taxon ID of organism.	ontology id	mandatory	No			NCBITaxon:1	
Sex	Animal sex, described using any child term of PATO_0000047.	ontology id	mandatory	No			PATO:0000047	
birth date	Birth date, in 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	
health status	Healthy animals should have the term normal, otherwise use the as many disease terms as necessary from EFO.	ontology id	recommended	Yes			PATO:0000461 EFO:0000408	
diet	Organism diet summary, more detailed information will be recorded in the associated protocols. Particuarly important for projects with controlled diet treatements. Free text field, but ensure standardisation within each study.	string	optional	No				
birth location	Name of the birth location.	string	optional	No				

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

Validation and brokered submission

- Rich metadata rulesets are only useful if they are met by all submissions.
- All FAANG data goes through pre-submission validation, that blocks submission till compliant.
- Validation service not only highlights errors, it warns on suggested improvements such as being more specific in ontologies.
- Brokered submission to underlying archives to ensure standard presentation.



FAANG validation

1. Upload template:

file.xlsx

2. Conversion results:

Status: Success

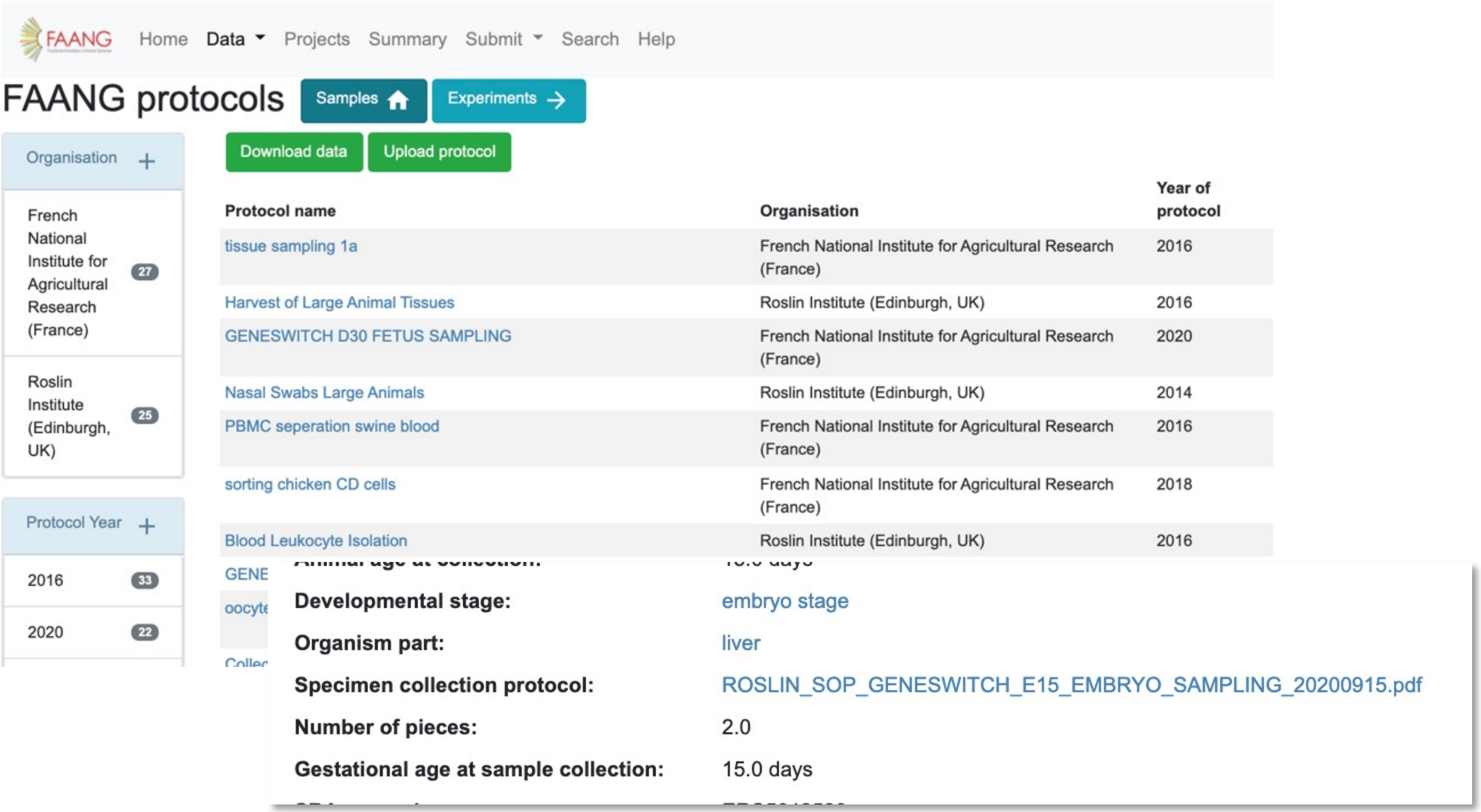
3. Validation results:

Status: Finished

Sample name	Core errors	Core warnings	Type errors	Type warnings	Custom errors	Custom warnings
ECA_UKY_S1	pass	pass	2 errors	1 warning	pass	pass
ECA_UKY_S2	pass	pass	pass	2 warnings	pass	pass
ECA_UKY_S9	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S10	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S32	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S33	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S34	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S35	pass	pass	pass	1 warning	pass	pass
ECA_UKY_S36	pass	pass	pass	1 warning	pass	pass

Detailed protocols mandatory with every submission

- Enhances reproducibility, reuse and comparative study possibilities.
- Our protocol browser shows all past protocols. A useful reference of methodologies for future studies.
- Encourages standardisation across future studies.



FAANG protocols

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

Protocol Year	
2016	33
2020	22

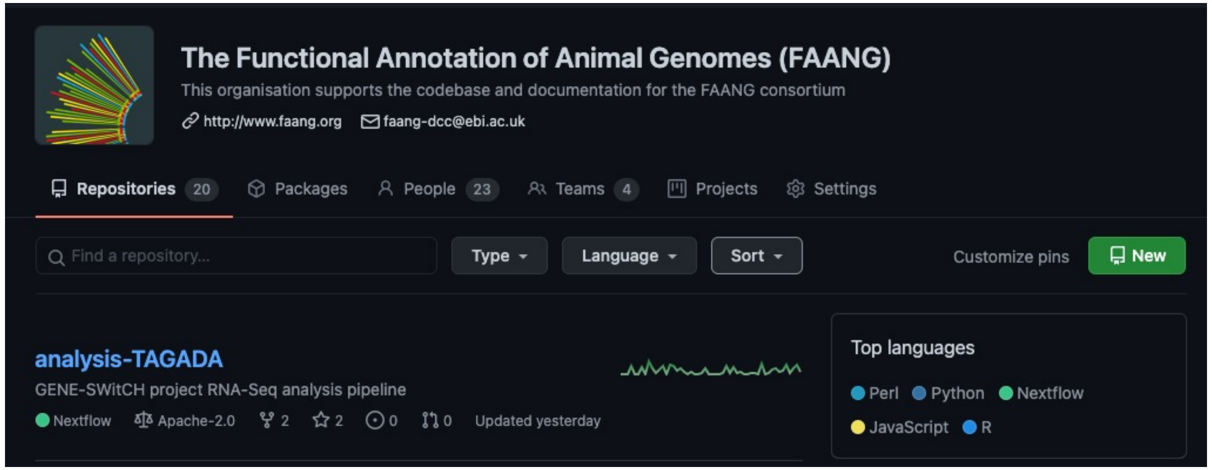
Protocol name	Organisation	Year of protocol
tissue sampling 1a	French National Institute for Agricultural Research (France)	2016
Harvest of Large Animal Tissues	Roslin Institute (Edinburgh, UK)	2016
GENESWITCH D30 FETUS SAMPLING	French National Institute for Agricultural Research (France)	2020
Nasal 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
GENE		
oocyte		
Collec		

Animal age at collection:	15.0 days
Developmental stage:	embryo stage
Organism part:	liver
Specimen collection protocol:	ROSLIN_SOP_GENESWITCH_E15_EMBRYO_SAMPLING_20200915.pdf
Number of pieces:	2.0
Gestational age at sample collection:	15.0 days

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

FAANG data analysis

- Shared development of a complete set of open pipelines across FAANG.
- Development based on the principles of open science, open source code and reproducible workflows.
- Researchers reuse and improve a common set of pipelines.
- EuroFAANG projects are exemplifying this shared development approach, focussing around nf-core.
- Will never be one pipeline that fits all, so we capture metadata on pipeline parameters for full reproducibility.



The Functional Annotation of Animal Genomes (FAANG)

This organisation supports the codebase and documentation for the FAANG consortium

http://www.faang.org faang-dcc@ebi.ac.uk

analysis-TAGADA

GENE-SWitCH project RNA-Seq analysis pipeline

● Nextflow   Apache-2.0   2   2   0   0   Updated yesterday

Top languages

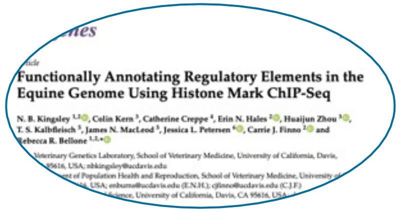
● Perl ● Python ● Nextflow ● JavaScript ● R

A community effort to collect a curated set of analysis pipelines built using Nextflow.

How FAANG promotes highly reusable standardised datasets

Rich, validated metadata

Automated Literature services



nes

icle

Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq

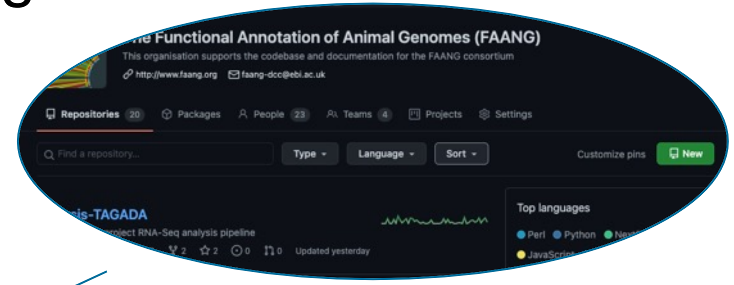
N. B. Kingsley [1,2], Colin Kern [3], Catherine Creppe [4], Erin N. Hales [2], Huaijun Zhou [3], T. S. Kalbfleisch [5], James N. MacLeod [5], Jessica L. Petersen [6], Carrie J. Finno [2] and Rebecca R. Bellone [1,2,*]

Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, 95616, USA; nbkingsley@ucdavis.edu

ent of Population Health and Reproduction, School of Veterinary Medicine, University of

USA; enburns@ucdavis.edu (E.N.H.); cjfinno@ucdavis.edu (C.J.F.)

Science, University of California, Davis, CA 95616, USA



he Functional Annotation of Animal Genomes (FAANG)

This organisation supports the codebase and documentation for the FAANG consortium

http://www.faang.org faang-dcc@ebi.ac.uk

is-TAGADA

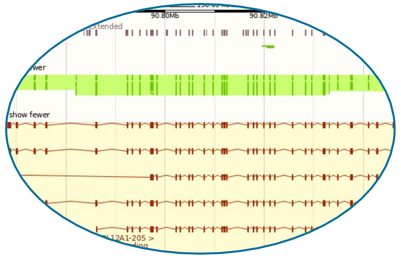
roject RNA-Seq analysis pipeline

2 2 0 0 Updated yesterday

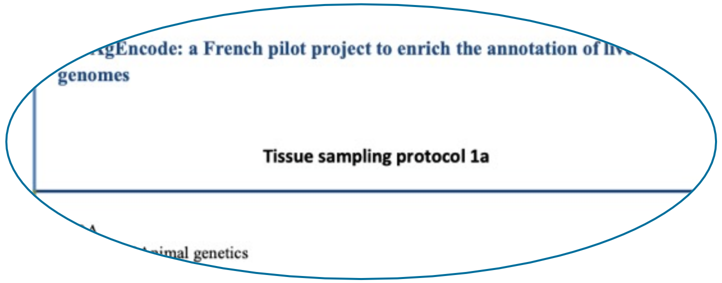
Top languages

● Perl ● Python ● Ne ● JavaScri

Accessible Analysis pipeline and parameters



FAANG dataset

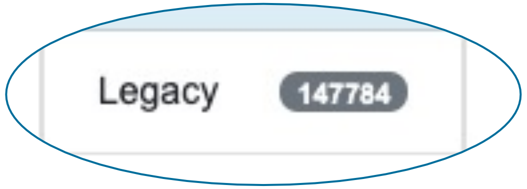


AgEncode: a French pilot project to enrich the annotation of live genomes

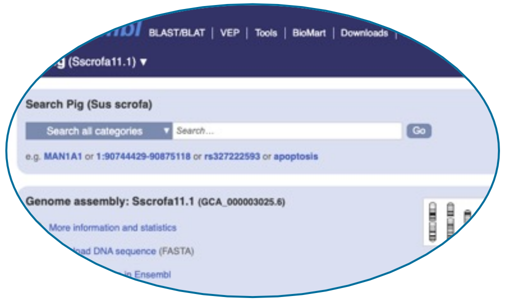
Tissue sampling protocol 1a

imal genetics

Detailed sampling and analysis protocols



Legacy contextual datasets



Search Pig (Sus scrofa)

e.g. MAN1A1 or 1:90744429-90875118 or rs327222593 or apoptosis

Genome assembly: Sscrofa11.1 (GCA_000003025.6)

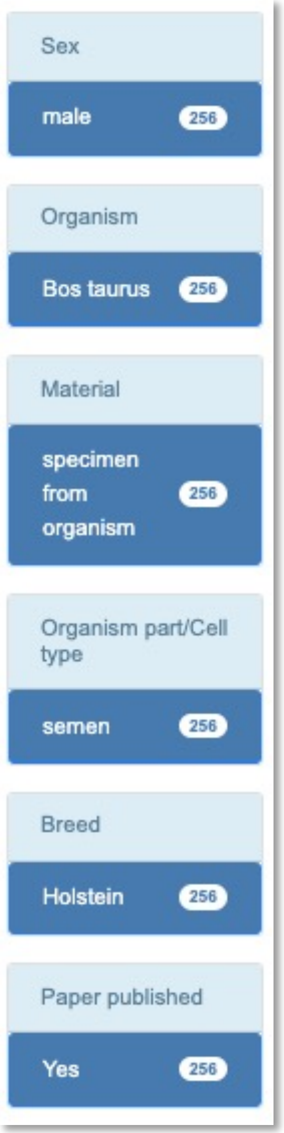
More information and statistics

load DNA sequence (FASTA)

in Ensembl

Secondary analyses and annotations

Intuitive identification of relevant reusable data



Sex

male 256

Organism

Bos taurus 256

Material

specimen from organism 256

Organism part/Cell type

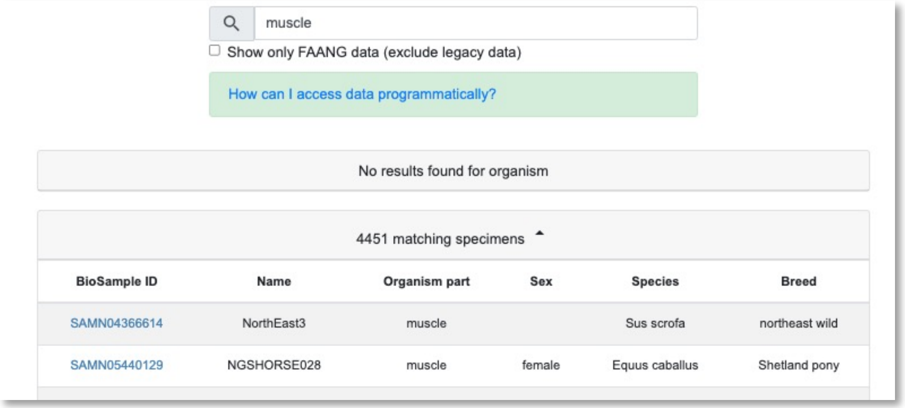
semen 256

Breed

Holstein 256

Paper published

Yes 256



muscle

□ Show only FAANG data (exclude legacy data)

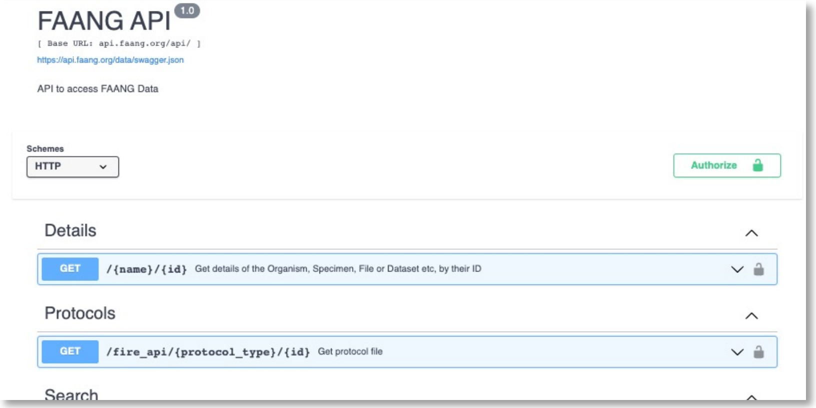
How can I access data programmatically?

No results found for organism

4451 matching specimens

BioSample ID	Name	Organism part	Sex	Species	Breed
SAMN04366614	NorthEast3	muscle		Sus scrofa	northeast wild
SAMN05440129	NGSHORSE028	muscle	female	Equus caballus	Shetland pony

Google style search



FAANG API 1.0

[Base URL: api.faang.org/api/]

https://api.faang.org/data/swagger.json

API to access FAANG Data

Schemes: HTTP

Authorize

Details

GET `/{name}/{id}` Get details of the Organism, Specimen, File or Dataset etc, by their ID

Protocols

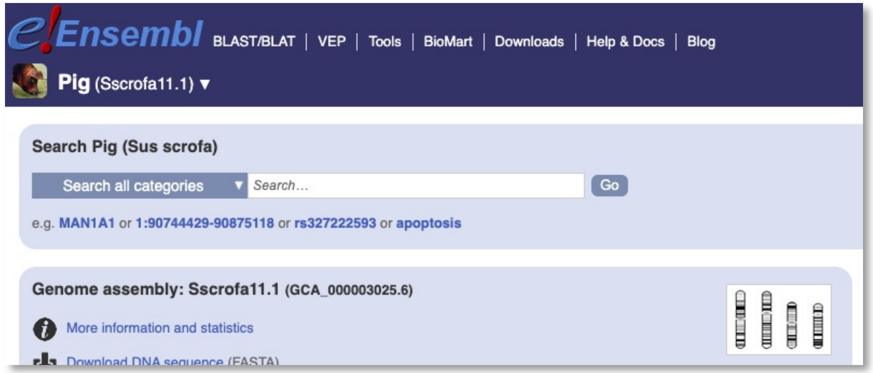
GET `/fire_api/{protocol_type}/{id}` Get protocol file

Search

APIs for external access -> FAANGmine.

Ontology and

metadata filters



e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Pig (Sscrofa11.1) ▾

Search Pig (Sus scrofa)

Search all categories ▾ Search… Go

e.g. MAN1A1 or 1:90744429-90875118 or rs327222593 or apoptosis

Genome assembly: Sscrofa11.1 (GCA_000003025.6)

More information and statistics

Download DNA sequence (FASTA)

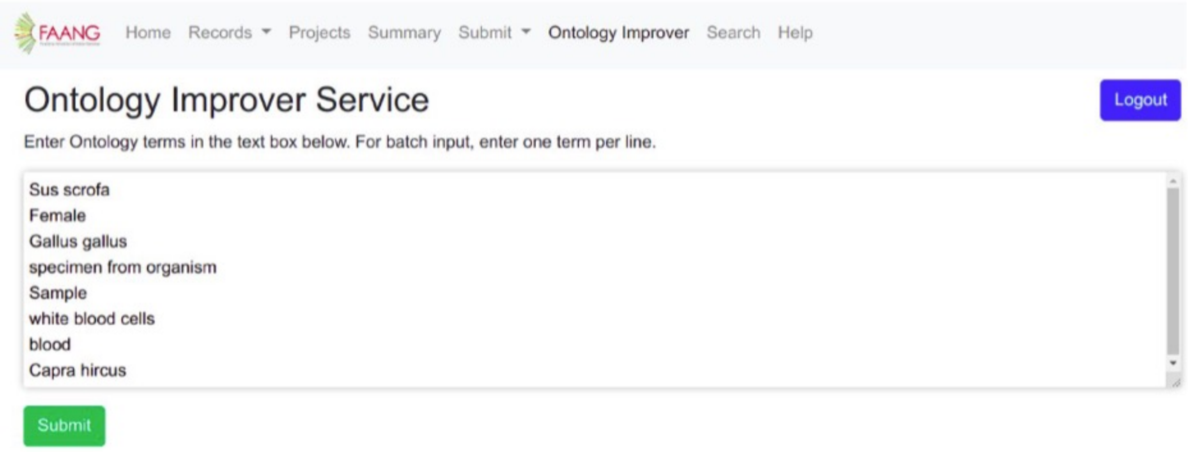
FAANG data drives improved Ensembl annotations

https://data.faang.org/

EMBL-EBI

FAANG Ontology improvement service Beta

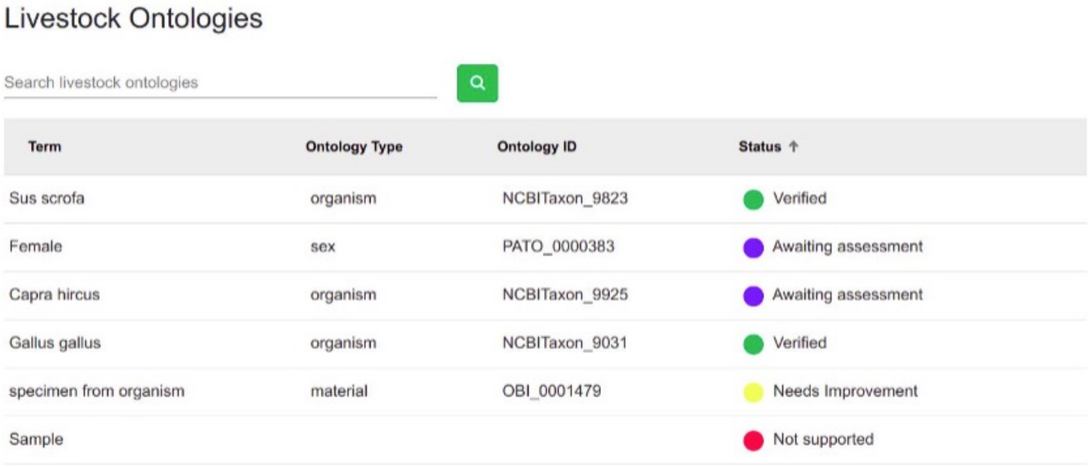
- Frequently ontologies are not appropriate for use in agriculture and G2P, as are medical or model organism based.
- We are developing a FAANG Ontology improvement service that will community source improvements to ontologies of importance in animal agriculture.
- Allows users to list ontologies of importance, flag ontologies that need improvement or are missing.
- Users submit improvements that are forwarded to underlying ontologies for update.



Ontology Improver Service

Enter Ontology terms in the text box below. For batch input, enter one term per line.

Sus scrofa
Female
Gallus gallus
specimen from organism
Sample
white blood cells
blood
Capra hircus



Livestock Ontologies

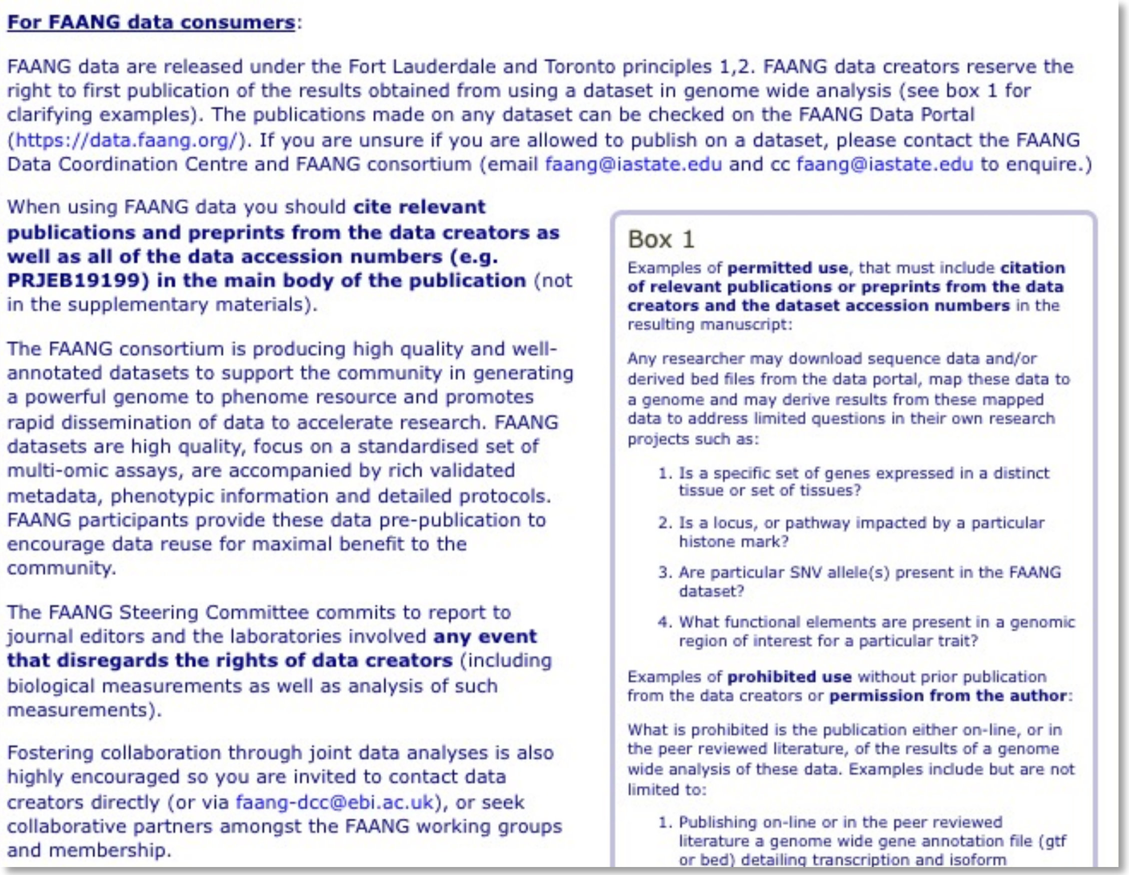
Search livestock ontologies

Term	Ontology Type	Ontology ID	Status ↑
Sus scrofa	organism	NCBITaxon_9823	● Verified
Female	sex	PATO_0000383	● Awaiting assessment
Capra hircus	organism	NCBITaxon_9925	● Awaiting assessment
Gallus gallus	organism	NCBITaxon_9031	● Verified
specimen from organism	material	OBI_0001479	● Needs Improvement
Sample			● Not supported

https://data.faang.org/ontology

Embargos and third party restrictions

- Necessary in many contexts, but must recognise the dampening of accelerated research if they limit open data reuse.
- FAANG encourages prepublication data archiving under Fort Lauderdale agreements to facilitate data reuse.
- Recently updated its data sharing statement to make data reuse expectations clearer.
- Still lots needed as a community to develop data openness and clarity of labelling.



For FAANG data consumers:

FAANG data are released under the Fort Lauderdale and Toronto principles 1,2. FAANG data creators reserve the right to first publication of the results obtained from using a dataset in genome wide analysis (see box 1 for clarifying examples). The publications made on any dataset can be checked on the FAANG Data Portal (https://data.faang.org/). If you are unsure if you are allowed to publish on a dataset, please contact the FAANG Data Coordination Centre and FAANG consortium (email faang@iastate.edu and cc faang@iastate.edu to enquire.)

When using FAANG data you should **cite relevant publications and preprints from the data creators as well as all of the data accession numbers (e.g. PRJEB19199) in the main body of the publication** (not in the supplementary materials).

The FAANG consortium is producing high quality and well-annotated datasets to support the community in generating a powerful genome to phenome resource and promotes rapid dissemination of data to accelerate research. FAANG datasets are high quality, focus on a standardised set of multi-omic assays, are accompanied by rich validated metadata, phenotypic information and detailed protocols. FAANG participants provide these data pre-publication to encourage data reuse for maximal benefit to the community.

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).

Fostering collaboration through joint data analyses is also highly encouraged so you are invited to contact data creators directly (or via faang-dcc@ebi.ac.uk), or seek collaborative partners amongst the FAANG working groups and membership.

Box 1

Examples of **permitted use**, that must include **citation of relevant publications or preprints from the data creators and the dataset accession numbers** in the resulting manuscript:

Any researcher may download sequence data and/or derived bed files from the data portal, map these data to a genome and may derive results from these mapped data to address limited questions in their own research projects such as:

1. Is a specific set of genes expressed in a distinct tissue or set of tissues?
2. Is a locus, or pathway impacted by a particular histone mark?
3. Are particular SNV allele(s) present in the FAANG dataset?
4. What functional elements are present in a genomic region of interest for a particular trait?

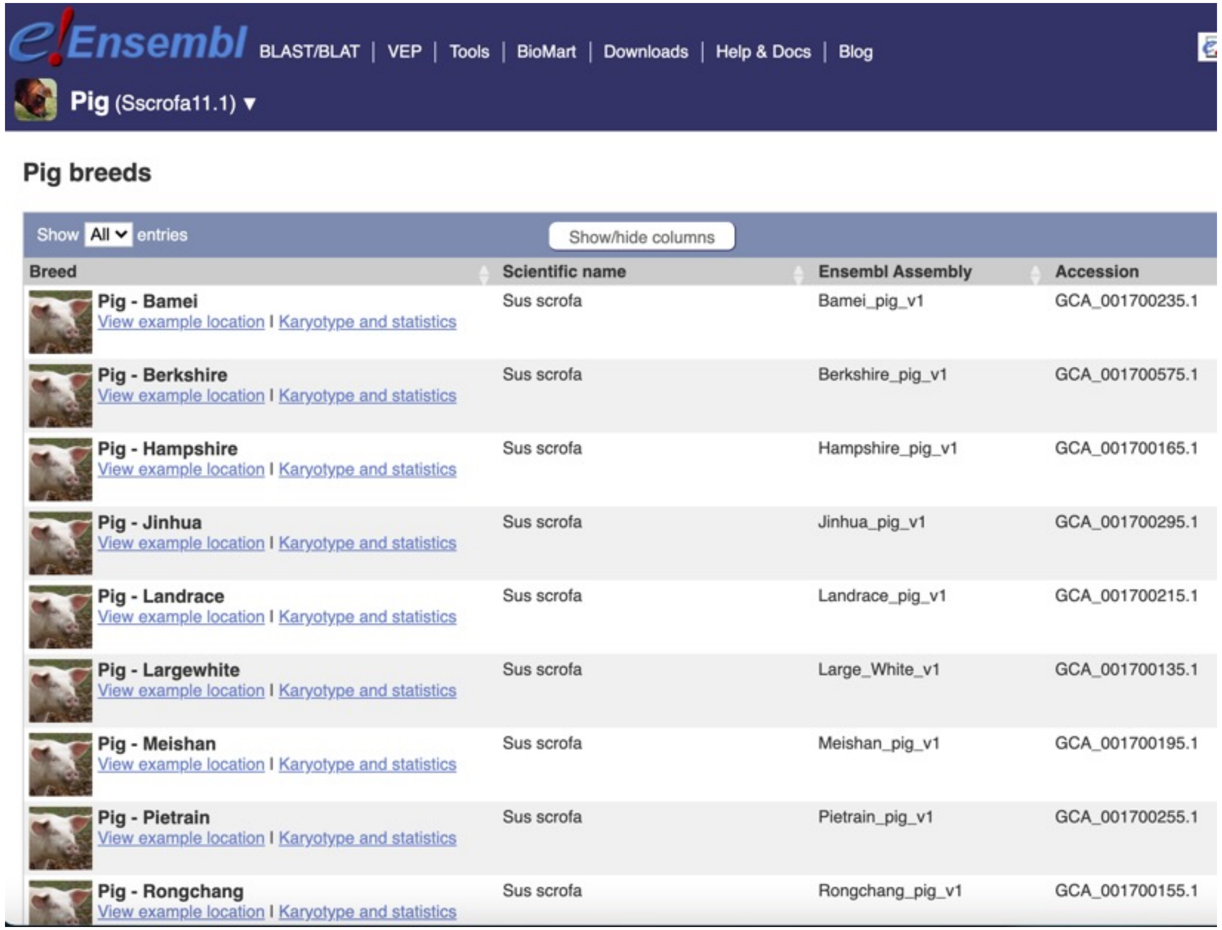
Examples of **prohibited use** without prior publication from the data creators or **permission from the author**:

What is prohibited is the publication either on-line, or in the peer reviewed literature, of the results of a genome wide analysis of these data. Examples include but are not limited to:

1. Publishing on-line or in the peer reviewed literature a genome wide gene annotation file (gtf or bed) detailing transcription and isoform

Potential challenge of Multiple references to data reuse

- Potential need to reanalyse past data and for researchers to agree and switch to new reference in coordination.
- Ensuring data is comparable and does not go out of date.
- Future advances of graph representations and scalability of genome browsers are needed.



Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Pig (Sscrofa11.1) ▾

Pig breeds

Show All entries Show/hide columns

Breed	Scientific name	Ensembl Assembly	Accession
Pig - Bamei View example location I Karyotype and statistics	Sus scrofa	Bamei_pig_v1	GCA_001700235.1
Pig - Berkshire View example location I Karyotype and statistics	Sus scrofa	Berkshire_pig_v1	GCA_001700575.1
Pig - Hampshire View example location I Karyotype and statistics	Sus scrofa	Hampshire_pig_v1	GCA_001700165.1
Pig - Jinhua View example location I Karyotype and statistics	Sus scrofa	Jinhua_pig_v1	GCA_001700295.1
Pig - Landrace View example location I Karyotype and statistics	Sus scrofa	Landrace_pig_v1	GCA_001700215.1
Pig - Largewhite View example location I Karyotype and statistics	Sus scrofa	Large_White_v1	GCA_001700135.1
Pig - Meishan View example location I Karyotype and statistics	Sus scrofa	Meishan_pig_v1	GCA_001700195.1
Pig - Pietrain View example location I Karyotype and statistics	Sus scrofa	Pietrain_pig_v1	GCA_001700255.1
Pig - Rongchang View example location I Karyotype and statistics	Sus scrofa	Rongchang_pig_v1	GCA_001700155.1

Some key Gaps for FAANG and wider communities to address

- Standardised ontologies across animals and crops, and improving existing ontologies from model organism and medical focus.
- Standardisation of minimal metadata standards between larger projects.
- Making data reuse conditions clearer, clearer labelling in molecular archives and data portals. Machine readable third party constraints.
- Need for cloud based preconfigured analyses to ease entry point and further standardise G2P analysis.
- Managing multiple references and graph genomics.

- These are some of the key points of focus for FAANG, AG2PI and Agbiodata.

Concluding Remarks

- FAANG promotes reuse through datasets having rich validated metadata, mandatory protocols, prepublication data sharing, standardised assays, standardised and documented analyses and an intuitive data portal.
- FAANG has produced functional maps of key animal species and continues to broaden its scope (FAANG to fork), whilst maintaining its core values of reusability and openness.
- Keen to address the data torrent grand challenges, in coordination with wider communities, to further promote data reuse, openness and standardisation.

Photo: Peter Harrison

Acknowledgments

peter@ebi.ac.uk
@peterwharrison

EMBL-EBI

- Alexey Sokolov (FAANG DCC Project Leader)
- Akshatha Nayak (FAANG DCC Bioinformatician)
- Koosum Roochun (FAANG DCC Bioinformatician)
- Raheela Aslam (DCC Web developer)

- H2020 EuroFAANG collaborators in particular form:
  - AQUA-FAANG
  - BovReg
  - GENE-SWitCH
- Members of FAANG steering committee and working groups
- Fantastic FAANG community for their continued data submission, feedback and support

EuroFAANG

Horizon 2020
European Union Funding
for Research & Innovation

AQUA-FAANG

BovReg
Understanding cattle genomes

GENE-SWitCH

Ensembl

wellcome