EuroFAANG Data Infrastructure

Standardizing and presenting BovReg and global FAANG data

Peter Harrison

Genome Analysis Team Leader

EMBL-European Bioinformatics Institute

peter@ebi.ac.uk

Horizon 2020
European Union Funding
for Research & Innovation

EMBL-European Bioinformatics Institute

50 YEARS | 1974–2024

- EMBL-EBIs mission is to help scientists realise the potential of big data in biology.
  - freely available data and services,
  - cutting edge research and technology,
  - advanced bioinformatics training.

- We provide services and expertise to support coordination, archiving, annotation and presentation of data at scale.

BovReg Understanding cattle genomes EuroFAANG EMBL-EBI

FAANG
Functional Annotation of Animal Genomes
10 Years

EuroFAANG

Data Coordination Centre at EMBL-EBI

Supports core FAANG aims of data **openness, reusability, and standardisation to** create a harmonised rich **genome to phenome resource**.

- Ensuring data is richly described, available and searchable.
- Ensuring data is consistently reported and presented.
- Facilitating standardisation of formats, protocols and analysis methods.
- Facilitating data openness, reusability and cross-project analysis.

Funded by EU EuroFAANG projects, but serves the global FAANG community.

BovReg
Understanding cattle genomes
EuroFAANG
EMBL-EBI

EuroFAANG

- Coordination across six large EU projects to standardize research processes, analysis and methods.
- Increases effectiveness at collectively improving animal production, welfare and supporting industry applications.
- Leverages global FAANG expertise and predictive power through combining standardized data.
- Data Infrastructure forms key component of future Research Infrastructure for Europe, that is in its concept development phase.

EuroFAANG

Leveraging functional genomic information to make animal production more robust and sustainable in Europe and beyond

Research aims

- ☑ Increase effiency through precision breeding
- ☑ Increase disease resistance
- ☑ Minimise environmental impact

Joint strategies

- ☑ Communication & Dissemination
- ☑ Training
- ☑ Research Methodology

BovReg Understanding cattle genomes EuroFAANG EMBL-EBI

Research Infrastructure and DCC

- The FAANG DCC and Data Portal have been funded by EuroFAANG projects and is now a key component of the future EuroFAANG Research Infrastructure.
- Expanding DCC to support new breeding, phenotyping, biobanking, genome editing and genomic technologies.
- Will retain a global focus and connection to FAANG

EuroFAANG
RESEARCH INFRASTRUCTURE

Horizon2020
European Union Funding
for Research & Innovation

EMBL-EBI ‘s role in BovReg consortium

A focus on **scientific reproducibility** through development of standardised pipelines, integration and consistency of data, comparative analysis and the improvement of data descriptions through ontologies (Lead: CRG).

CRG Centre for Genomic Regulation

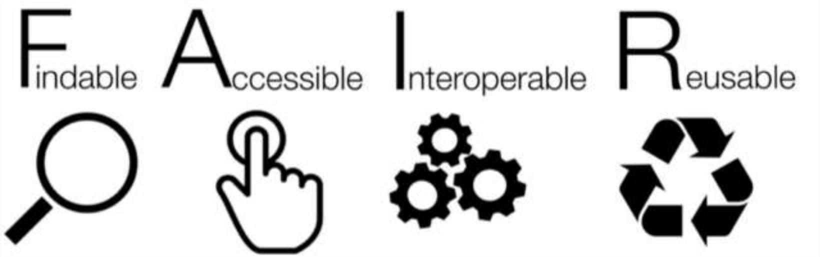
EMBL-EBI

A focus on the **open distribution** of BovReg data to the community and the application of **FAIR** data principles (Lead: EMBL).

Findable Accessible Interoperable Reusable

What makes FAANG datasets special

- Rich, consistent and validated metadata descriptions.
- Standardised set of core assays from each contributing project.
- Mandatory sampling, experiment and analysis protocols connected to each dataset.
- Links to standardised analysis workflows with evert analysis file.
- A data platform and community drive that ensures data is open and FAIR.
- Innovative data presentation and contextual genome visualisation.



To accelerate genome to phenome research

Photo: CODATA

Fully reproducible analyses

BioSamples

All samples and datasets linked together with BioSamples

Prepublication open data sharing of ‘omics datasets

nf-core

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

Containerised and parameterised workflows for every analysis

Rich validated metadata and ontologies

Mandatory detailed sampling and analysis protocols

Europe PMC

Automated literature tracking

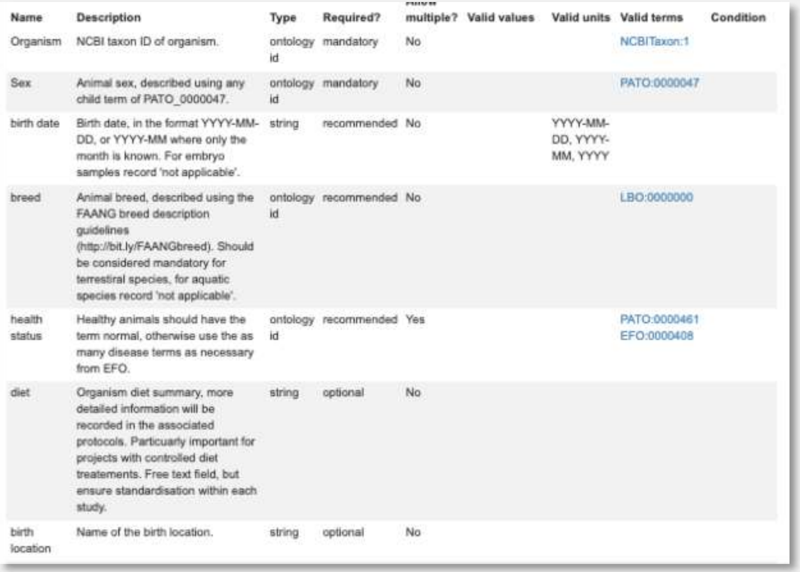
Ensembl

Standardised Ensembl annotations and mapped tracks

BovReg Understanding cattle genomes EuroFAANG EMBL-EBI

A full metadata solution for FAANG

- FAANG metadata aims to:
  - be easily understandable, highly repeatable and reusable.
  - use consistent terminology
  - be well structured, rich and specific.
- Enhancing reproducibility, accelerating research and enabling cross-depositor analyses to be performed.

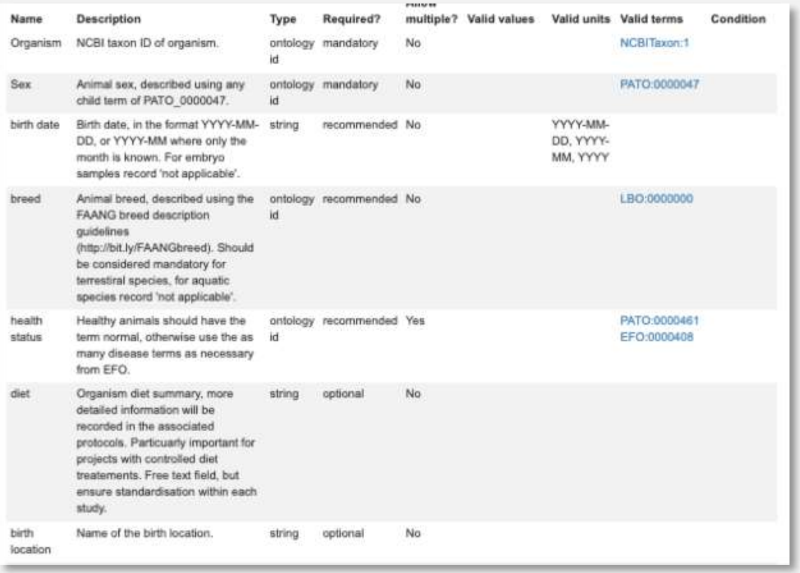


Name	Description	Type	Required?	Allow multiple?	Valid values	Valid units	Valid terms	Condition
Organism	NCBI taxon ID of organism.	ontology id	mandatory	No			NCBITaxon:1	
Sex	Animal sex, described using any child term of PATO_0000047.	ontology id	mandatory	No			PATO:0000047	
birth date	Birth date, in 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

A full metadata solution for FAANG

- FAANG metadata aims to:
  - be easily understandable, highly repeatable and reusable.
  - use consistent terminology
  - be well structured, rich and specific.
- Enhancing reproducibility, accelerating research and enabling cross-depositor analyses to be performed.
- Key part of standardisation is controlling terminology through standardised ontologies to make downstream search and analysis more powerful.
- Why it's important ontologies are accurate.

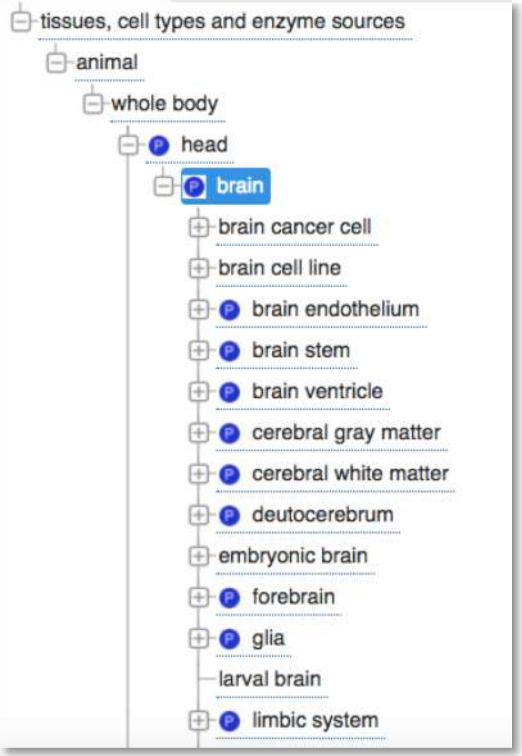


Name	Description	Type	Required?	multiple?	Valid values	Valid units	Valid terms	Condition
Organism	NCBI taxon ID of organism.	ontology id	mandatory	No			NCBITaxon:1	
Sex	Animal sex, described using any child term of PATO_0000047.	ontology id	mandatory	No			PATO:0000047	
birth date	Birth date, in the format 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

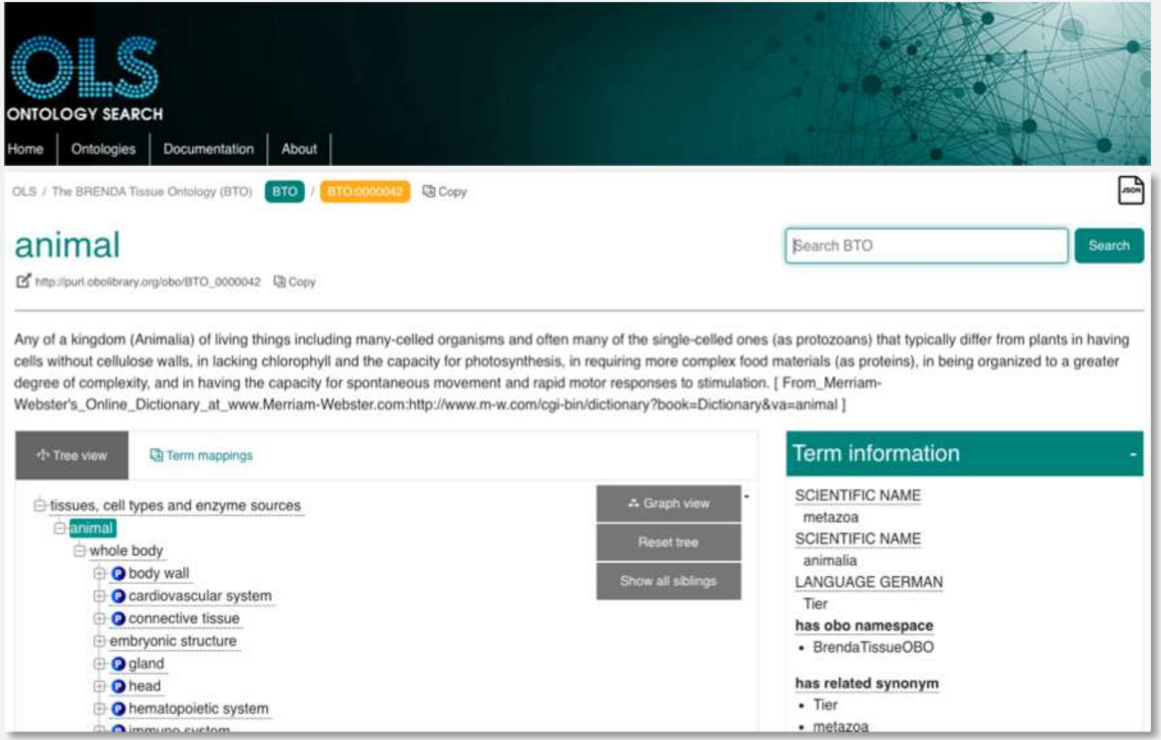
Ontologies

- In biology we use ontologies to unambiguously describes a biological entity.
- Each ontology entity has a unique identifier and groups a number of synonyms under it.
- Ontologies have hierarchy allowing for intuitive search at different levels of specificity.
- For example, in this tissue ontology ‘cerebral grey matter’ is a child of the ‘brain’ ontology term.



- tissues, cell types and enzyme sources
  - animal
    - whole body
      - head
        - brain
          - brain cancer cell
          - brain cell line
          - brain endothelium
          - brain stem
          - brain ventricle
          - cerebral gray matter
          - cerebral white matter
          - deutocerebrum
          - embryonic brain
          - forebrain
          - glia
          - larval brain
          - limbic system

EMBL-EBI Ontology Lookup Service



animal

http://purl.obolibrary.org/obo/BTO_0000042

Any of a kingdom (Animalia) of living things including many-celled organisms and often many of the single-celled ones (as protozoans) that typically differ from plants in having cells without cellulose walls, in lacking chlorophyll and the capacity for photosynthesis, in requiring more complex food materials (as proteins), in being organized to a greater degree of complexity, and in having the capacity for spontaneous movement and rapid motor responses to stimulation. [From_Merriam-Webster's_Online_Dictionary_at_www.Merriam-Webster.com:http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=animal]

- tissues, cell types and enzyme sources
  - animal
    - whole body
      - body wall
      - cardiovascular system
      - connective tissue
      - embryonic structure
      - gland
      - head
      - hematopoietic system

Term information

SCIENTIFIC NAME
metazoa
SCIENTIFIC NAME
animalia
LANGUAGE GERMAN
Tier

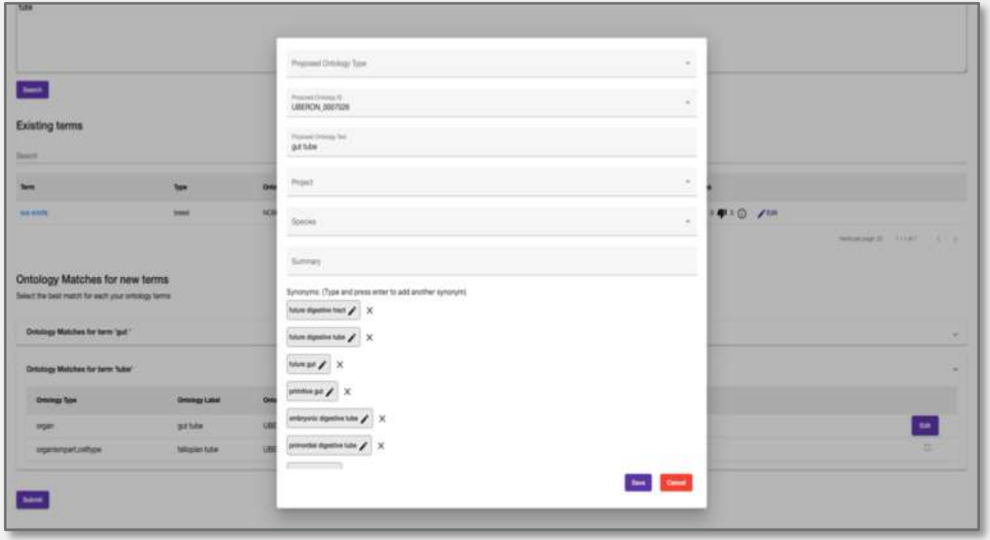
has obo namespace
- BrendaTissueOBO

has related synonym
- Tier
- metazoa

https://www.ebi.ac.uk/ols/

Ontology Improvement tool

Collective call from H2020 EuroFAANG projects and wider communities to improve ontology descriptions for animal agriculture and genotype to phenotype research.



Proposed Ontology ID: UBERON_0007026

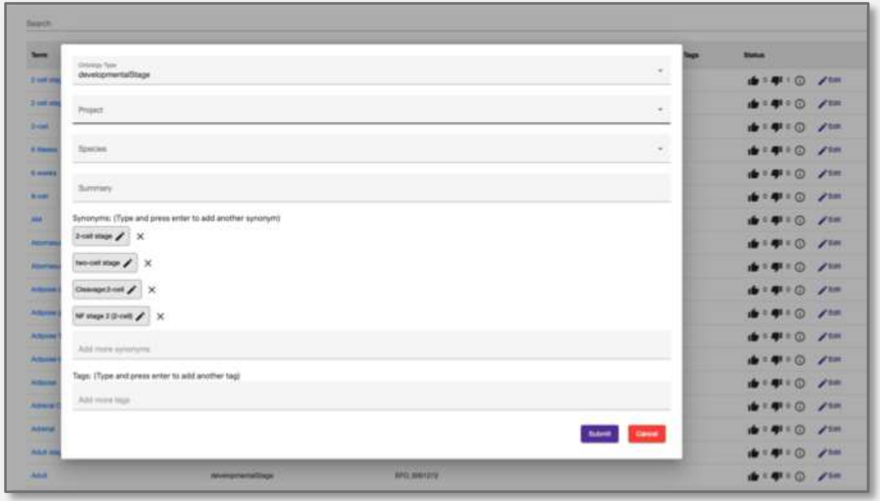
Proposed Ontology Text: gut tube

Synonyms: (Type and press enter to add another synonym)

- future digestive tract
- future digestive tube
- future gut
- primitive gut
- embryonic digestive tube
- primordial digestive tube

Enhancing Ontologies for BovReg

- Users assess ontology suitability, can be linked to projects such as BovReg.
- Authenticated login that tracks provenance of proposed changes.
- Users can propose ontology improvements, work on them collaboratively and vote on changes.
- Changes will be forwarded to underlying ontologies.



Ontology Type: developmentalStage

Synonyms: (Type and press enter to add another synonym)

2-cell stage

two-cell stage

Cleavage:2-cell

NF stage 2 (2-cell)

Tags: (Type and press enter to add another tag)

https://data.faang.org/ontology

Join the FAANG metaFAIR task force

FAANG

Functional Annotation of Animal Genomes

Functional Annotation of ANimal Genomes (FAANG) Project

— A coordinated international action to accelerate Genome to Phenome

Task Force: metaFAIR

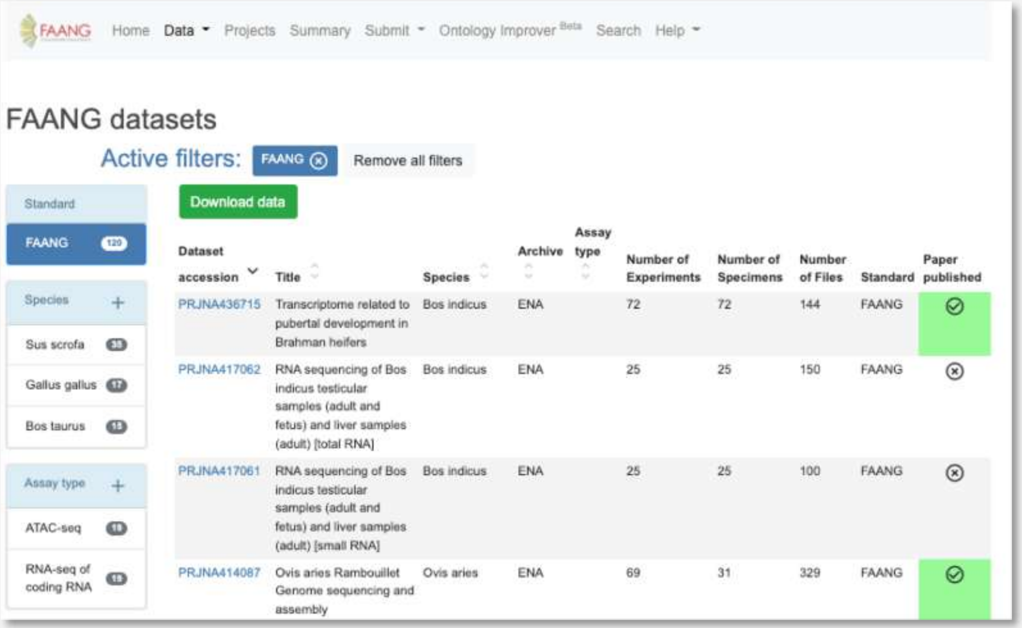
— FAANG Metadata, Ontologies and FAIR Data Sharing Task Force

Sign up at
www.faang.org

- Task force dedicated to managing the FAIR FAANG metadata standards.
- We need your expertise.
- Help update key standards of relevance to your current and future work

The FAANG Data Portal

- A single access point to all FAANG metadata, data, pipelines and publications.
- Direct access to download data from underlying public archives.
- Automatically identifies dataset (re)use in publications, and links publications to datasets.
- Intuitive search and filtering.
- Data visualisation and cross referencing.



FAANG datasets

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 testicular samples (adult and fetus) and liver samples (adult) [small RNA]	Bos indicus	ENA		25	25	100	FAANG	⊗
PRJNA414087	Ovis aries Rambouillet Genome sequencing and assembly	Ovis aries	ENA		69	31	329	FAANG	✓

https://data.faang.org/

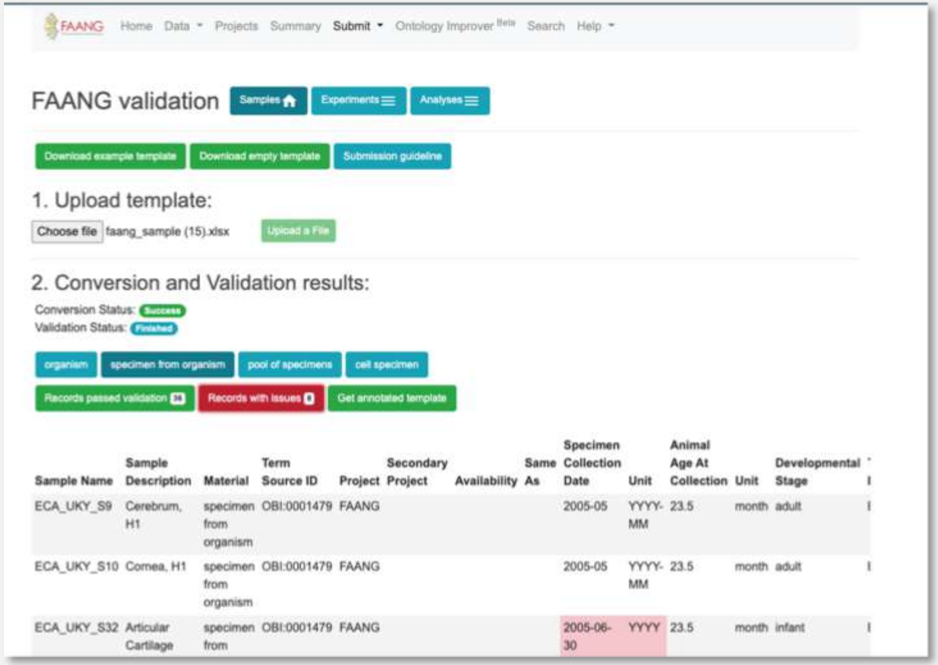
FAANG Functional Annotation of Animal Genomes 10 Years

BovReg Understanding cattle genomes EuroFAANG EMBL-EBI

Validation and brokered submission

FAANG
Functional Annotation of Animal Genomes
10 Years

- All FAANG data has pre-submission validation, ensuring rich metadata.
- Not only highlights errors, suggests improvements such as more specific ontologies.
- Brokered submission to underlying archives simplifies process for users.
- Also submit protocols, track hubs and workflows with configuration files.



FAANG validation

1. Upload template:

faang_sample (15).xlsx

2. Conversion and Validation results:

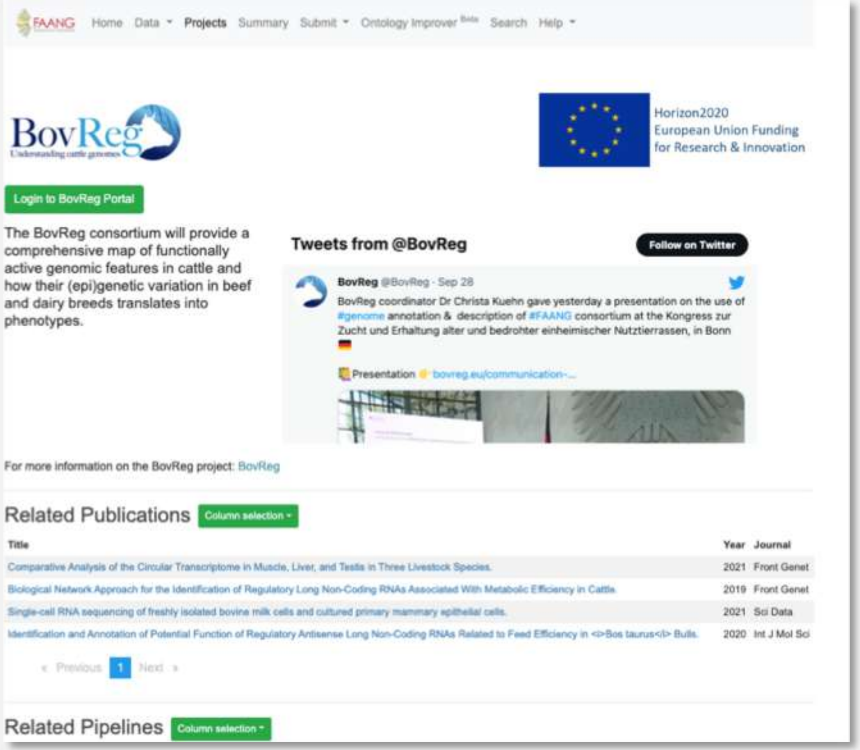
Conversion Status: Success
Validation Status: Finished

Sample Name	Sample Description	Material	Term Source ID	Project	Secondary Project	Availability	Same As	Specimen Collection Date	Unit	Animal Age At Collection	Unit	Developmental Stage
ECA_UKY_S9	Cerebrum, H1	specimen from organism	OBI:0001479	FAANG				2005-05	YYYY-MM	23.5	month	adult
ECA_UKY_S10	Cornea, H1	specimen from organism	OBI:0001479	FAANG				2005-05	YYYY-MM	23.5	month	adult
ECA_UKY_S32	Articular Cartilage	specimen from	OBI:0001479	FAANG				2005-06-30	YYYY	23.5	month	infant

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

BovReg Understanding cattle genomes
EuroFAANG
EMBL-EBI

BovReg Public Data Portal



The BovReg consortium will provide a comprehensive map of functionally active genomic features in cattle and how their (epi)genetic variation in beef and dairy breeds translates into phenotypes.

Tweets from @BovReg

BovReg @BovReg · Sep 28

BovReg coordinator Dr Christa Kuehn gave yesterday a presentation on the use of #genome annotation & description of #FAANG consortium at the Kongress zur Zucht und Erhaltung alter und bedrohter einheimischer Nutztierrassen, in Bonn

Presentation bovreg.eu/communication-...

For more information on the BovReg project: BovReg

Related Publications

Title	Year	Journal
Comparative Analysis of the Circular Transcriptome in Muscle, Liver, and Testis in Three Livestock Species.	2021	Front Genet
Biological Network Approach for the Identification of Regulatory Long Non-Coding RNAs Associated With Metabolic Efficiency in Cattle.	2019	Front Genet
Single-cell RNA sequencing of freshly isolated bovine milk cells and cultured primary mammary epithelial cells.	2021	Sci Data
Identification and Annotation of Potential Function of Regulatory Antisense Long Non-Coding RNAs Related to Feed Efficiency in <i>Bos taurus</i> Bulls.	2020	Int J Mol Sci

Related Pipelines

- A dedicated slice of the FAANG Data Portal presenting all data generated by BovReg.
- Contains all public samples and datasets produced by the project.
- Automatically identifies any publications that mention BovReg datasets in their text.
- Includes tracking of workflows/pipelines utilised by BovReg and in nf-core.

Related Pipelines

Pipeline name	Assay type	Pipeline link	Pipeline documentation	Platform
BovReg/nf-core-rnaseq	RNA-seq	https://github.com/BovReg/nf-core-rnaseq	https://github.com/BovReg/nf-core-rnaseq#readme	nf-core

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

Track Hub Brokering Platform

Track Hubs

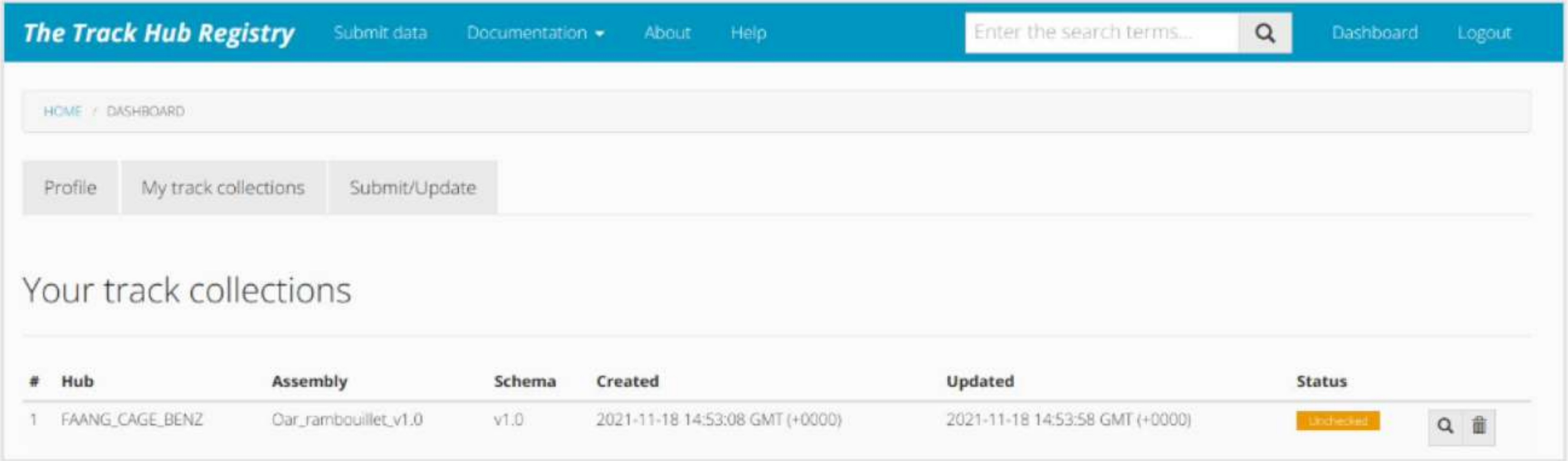
- web-accessible directories of genomic data
- store genome annotations within specifically formatted files
- visualise data on genome browsers
- convenient way to view and share large datasets

FAANG Track Hubs brokering tool

A solution for brokered upload, storage, registration and presentation of FAANG tracks data

Track Hub Brokering Platform

- Track Hubs are registered on behalf of the user thorough **EMBL-EBI Track Hub Registry**
- Track Hub Registry - A global centralised collection of publicly accessible track hubs
- Makes FAANG Data searchable and discoverable for viewing in Genome Browsers

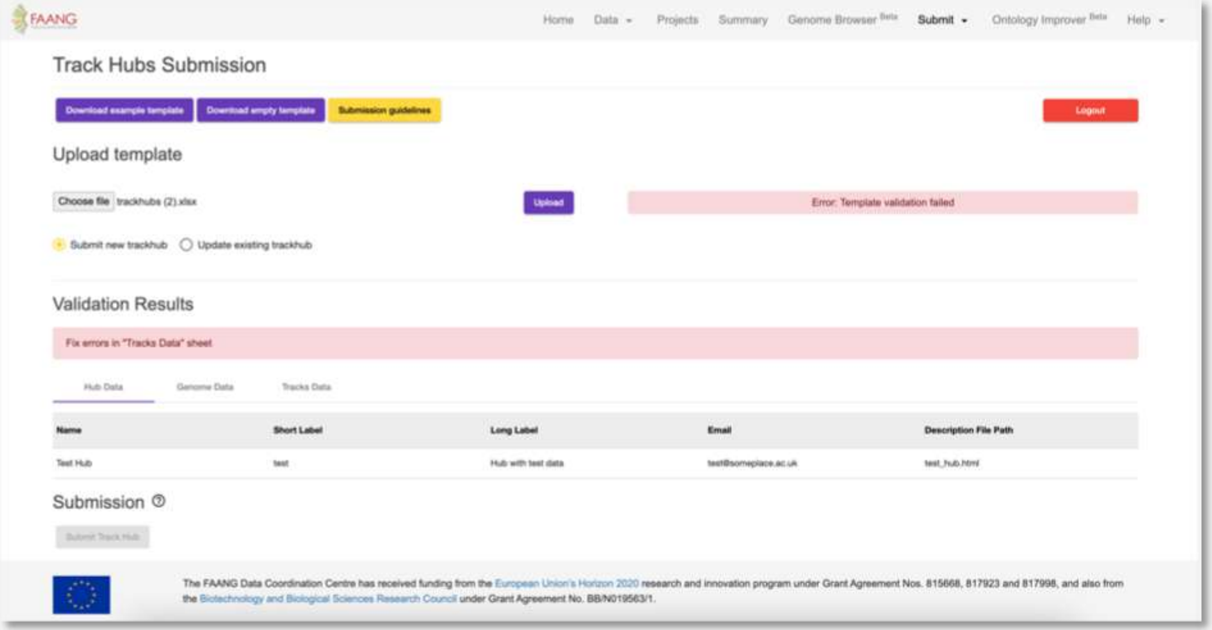


Your track collections

#	Hub	Assembly	Schema	Created	Updated	Status
1	FAANG_CAGE_BENZ	Oar_rambouillet_v1.0	v1.0	2021-11-18 14:53:08 GMT (+0000)	2021-11-18 14:53:58 GMT (+0000)	Unchecked

https://data.faang.org/trackhubs

Track Hub submission



Track Hubs Submission

Upload template

trackhubs (2).xlsx

Error: Template validation failed

● Submit new trackhub ○ Update existing trackhub

Validation Results

Fix errors in "Tracks Data" sheet

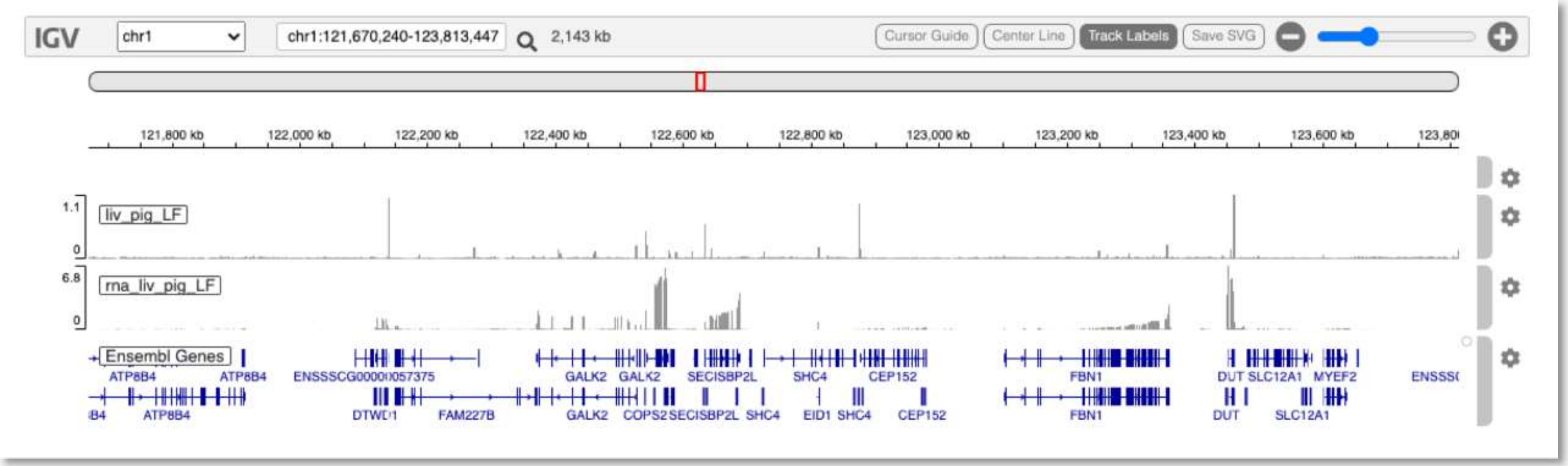
Name	Short Label	Long Label	Email	Description File Path
Test Hub	test	Hub with test data	test@someplace.ac.uk	test_hub.html

Submission ⓘ

The FAANG Data Coordination Centre has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement Nos. 815668, 817923 and 817998, and also from the Biotechnology and Biological Sciences Research Council under Grant Agreement No. BB/N019563/1.

- FAANG brokered submission, validation, storage and presentation of community genomic track hubs.
- Tracks are automatically linked to relevant data and connections shown in the FAANG Data portal to view in Genome browsers.

FAANG genome browser



- Track hub genome browser embedded in FAANG Data Portal.
- Based on IGV browser, providing powerful visual interaction.
- Highly customisable viewer to present your project track hubs.
- Select your reference of choice, setup multi track views, and incorporate Ensembl annotations.

Ensembl Genome Browser

GLOBAL CORE BIODATA RESOURCE

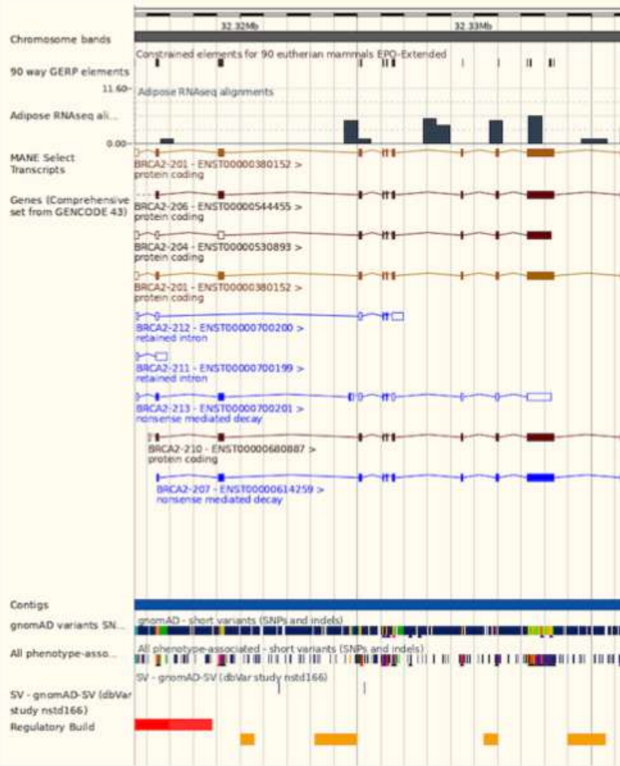
- A genomics platform and browser for enabling and accelerating downstream science.
- Standardized, high quality and richly described FAANG data is driving new annotations in Ensembl.
- Range of tools such as Variant Effect Predictor, EMBL industry promotion and ongoing training.

Comparative analysis

Gene and transcript annotation

Variation

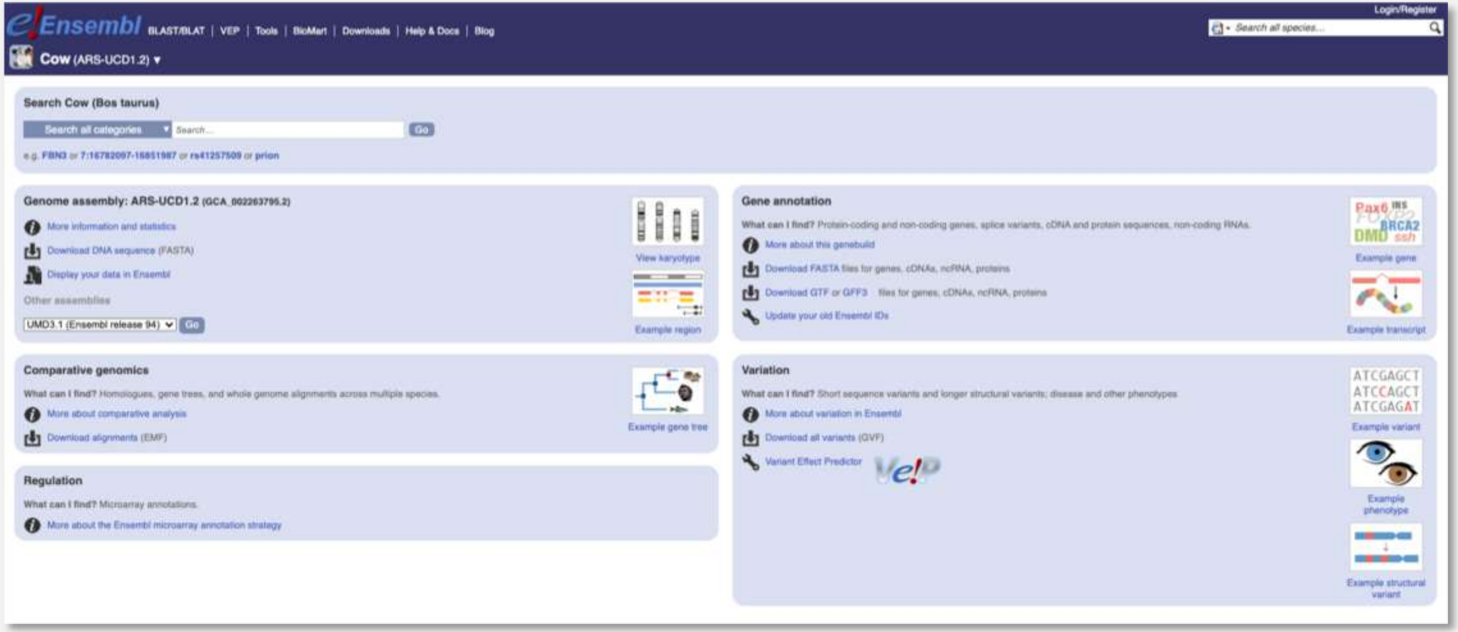
Regulatory elements



www.ensembl.org

High-quality annotations will be improved with wealth of BovReg data.

- Annotations will be rapidly and freely available through Ensembl once BovReg releases the data publicly.



Search Cow (Bos taurus)

e.g. FBN3 or 7:16782097-16851987 or rs41257509 or prion

Genome assembly: ARS-UCD1.2 (GCA_002263795.2)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Other assemblies

UMD3.1 (Ensembl release 94) Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download FASTA files for genes, cDNAs, ncRNA, proteins

Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

Update your old Ensembl IDs

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Variant Effect Predictor

Regulation

What can I find? Microarray annotations.

More about the Ensembl microarray annotation strategy

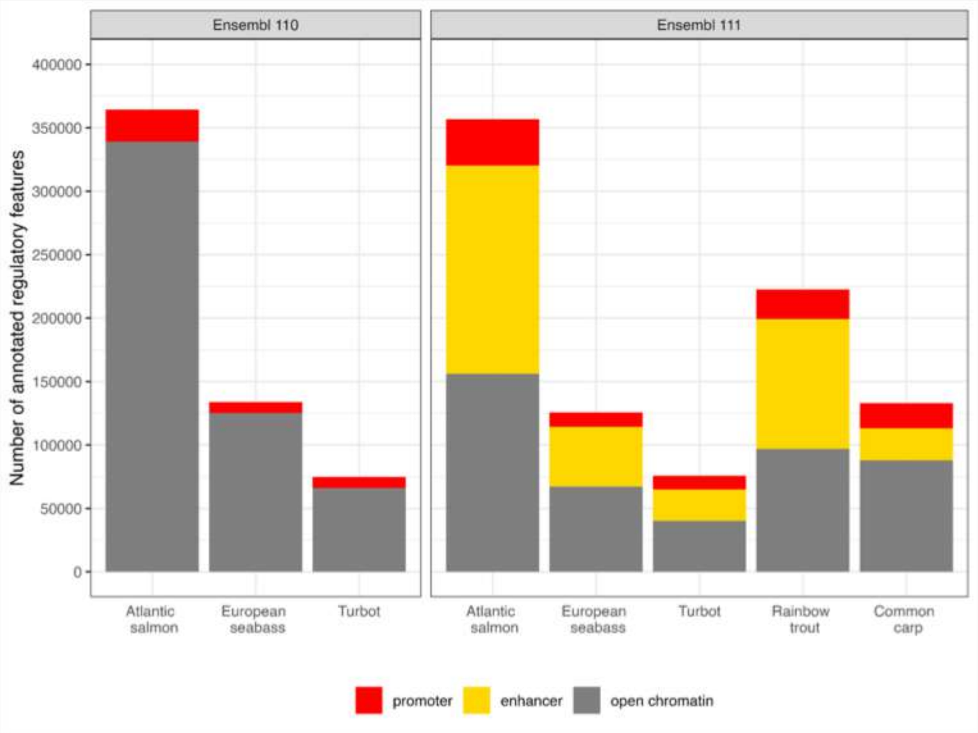
e!Ensembl First Regulatory builds

EMBL 50 YEARS | 1974–2024

- In collaboration with EuroFAANG projects released Ensembl regulatory builds for chicken, pig and 5 fish species.
- Driven new Ensembl annotation methods and visualisation.
- Viewable in the Ensembl browser.

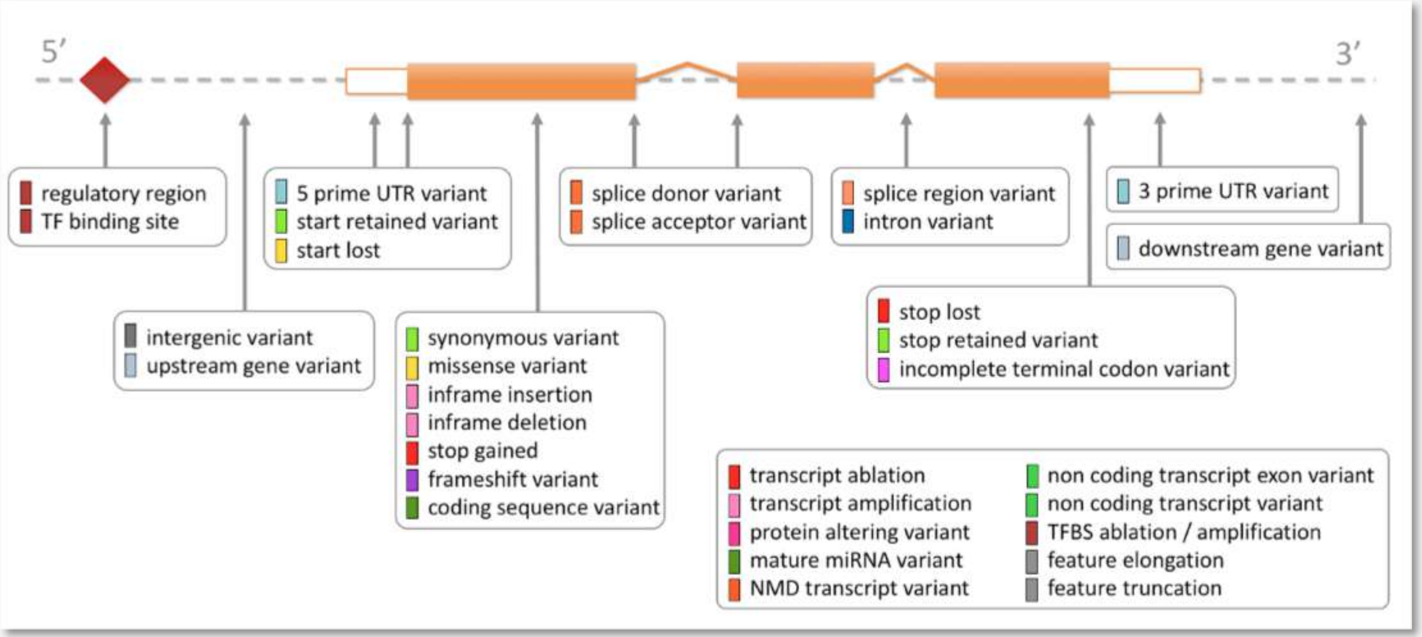
AQUA-FAANG

GENE-SWitCH



Ensembl Variant Effect Predictor Tool

Ensembl VEP maps variants to transcripts and regulatory features and predicts their molecular consequence. Results are reported using Sequence Ontology terms.



- 🟥 regulatory region
- 🟥 TF binding site

- ⬛ intergenic variant
- ⬜ upstream gene variant

- 🟦 5 prime UTR variant
- 🟩 start retained variant
- 🟨 start lost

- 🟩 synonymous variant
- 🟨 missense variant
- 🟪 inframe insertion
- 🟪 inframe deletion
- 🟥 stop gained
- 🟪 frameshift variant
- 🟩 coding sequence variant

- 🟧 splice donor variant
- 🟧 splice acceptor variant

- 🟧 splice region variant
- 🟦 intron variant

- 🟥 stop lost
- 🟩 stop retained variant
- 🟪 incomplete terminal codon variant

- 🟦 3 prime UTR variant

- ⬜ downstream gene variant

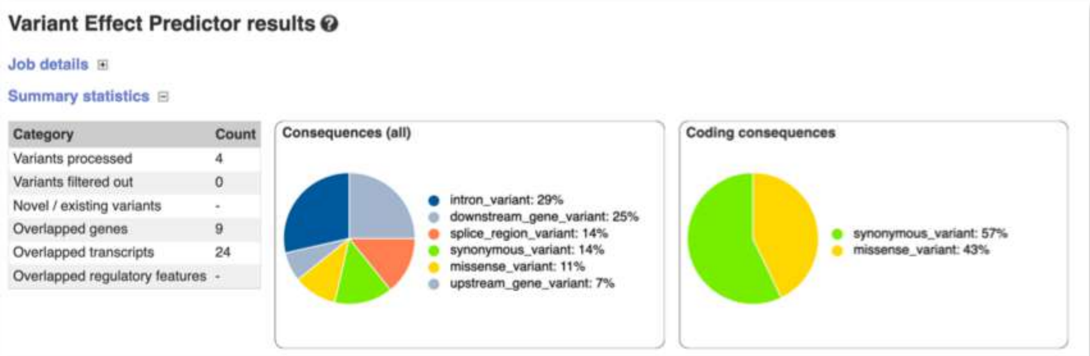
- 🟥 transcript ablation
- 🟪 transcript amplification
- 🟪 protein altering variant
- 🟩 mature miRNA variant
- 🟧 NMD transcript variant
- 🟩 non coding transcript exon variant
- 🟩 non coding transcript variant
- 🟥 TFBS ablation / amplification
- ⬛ feature elongation
- ⬛ feature truncation

The Ensembl Variant Effect Predictor. McLaren W. et al. *Genome Biology* 17:122(2016)

Ensembl VEP web tool output

Can quickly analyse a set of GWAS top hits

The filtering tool enables custom variant prioritisation



Variant Effect Predictor results

Job details

Summary statistics

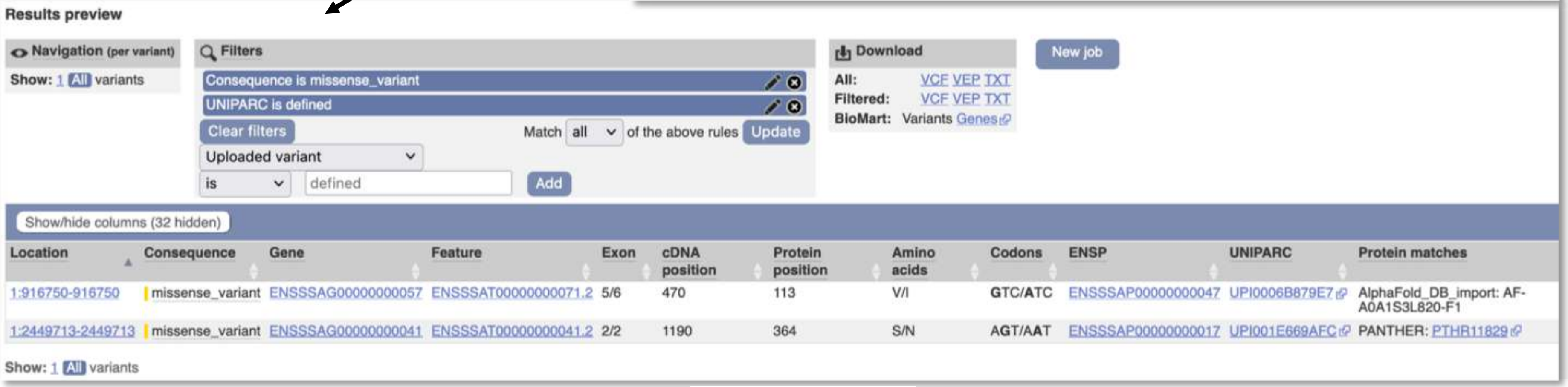
Category	Count
Variants processed	4
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	9
Overlapped transcripts	24
Overlapped regulatory features	-

Consequences (all)

- intron_variant: 29%
- downstream_gene_variant: 25%
- splice_region_variant: 14%
- synonymous_variant: 14%
- missense_variant: 11%
- upstream_gene_variant: 7%

Coding consequences

- synonymous_variant: 57%
- missense_variant: 43%



Results preview

Navigation (per variant)

Show: 1 All variants

Filters

Consequence is missense_variant

UNIPARC is defined

Download

All: VCF VEP TXT

Filtered: VCF VEP TXT

BioMart: Variants Genes

Location	Consequence	Gene	Feature	Exon	cDNA position	Protein position	Amino acids	Codons	ENSP	UNIPARC	Protein matches
1:916750-916750	missense_variant	ENSSSAG00000000057	ENSSSAT00000000071.2	5/6	470	113	V/I	GTC/ATC	ENSSSAP00000000047	UPI0006B879E7	AlphaFold_DB_import: AF-A0A1S3L820-F1
1:2449713-2449713	missense_variant	ENSSSAG00000000041	ENSSSAT00000000041.2	2/2	1190	364	S/N	AGT/AAT	ENSSSAP00000000017	UPI001E669AFC	PANTHER: PTHR11829

Show: 1 All variants

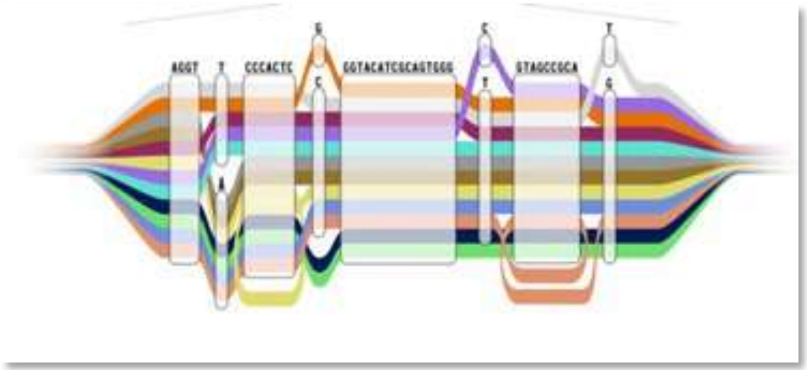
Ensembl Cattle (*Bos taurus*)

- Based on all publicly available data (will be updated again when further BovReg data publicly released)
- ARS-UCD1.3 available via ensembl.org now
- ARS-UCD2.0 available via rapid.ensembl.org in next few weeks
- Also gathering more breed annotations

Assembly	ARS-UCD1.3 (GCA_002263795.3)
# protein-coding genes	23,739
BUSCO completeness	C:96.9%[S:95.7%,D:1.2%],F:1.2%,M:1.9%,n:12234

Future directions

- Ensembl will continue to update and improve reference annotations and looking forward to adding T2T references.
- Always looking for further species and breeds/strains of economic and research interest.
- This variation builds towards a pangenomic view for each species, following work being done in human and crops (Fergal Martin's talk later today).
- Through Earth BioGenome Projects Ensembl is looking to annotate all eukaryotic species.



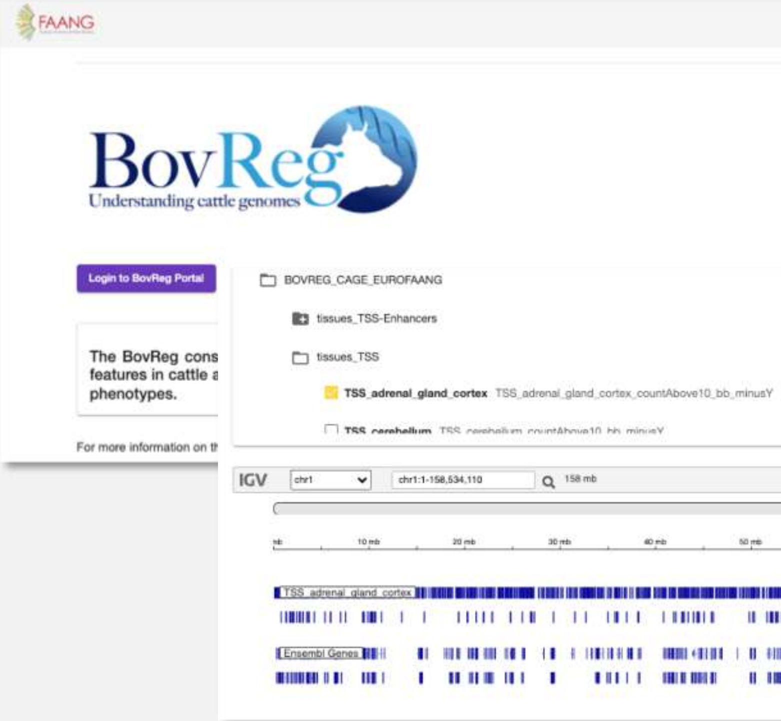
Elixir Domestic Animal Genomes and Phenomes

- Launching new task force driven by EuroFAANG and emerging Research Infrastructures on Phenotyping.
- Elixir brings together scientists from 23 countries and over 250 research institutes:
  - Coordinate data and technology discussions
  - Develop standards
  - Standardise analysis and workflows
  - Organise training, workshops and hackathons
  - Connect to global efforts
- Contact peter@ebi.ac.uk if interested.

Summary

FAANG
Functional Annotation of Animal Genomes
EuroFAANG

- BovReg FAANG data portal presents wealth of standardised BovReg data to the community.
- Metadata and ontology standards customised for BovReg needs.
- BovReg data will drive improved Ensembl annotations, and enable first Ensembl cattle regulatory build.
- Data Infrastructure developed during BovReg project will be a key component of future EuroFAANG and FAANG data coordination.



The BovReg cons
features in cattle a
phenotypes.

For more information on th

BovReg
Understanding cattle genomes
EuroFAANG
EMBL-EBI

EMBL-EBI DCC Acknowledgements

Alexey Sokolov

Agricultural and Biodiversity Coordination Project Leader
Lead developer of FAANG data portal

Koosum Roochun

Agricultural and Biodiversity Coordination
Bioinformatician

Milena Mansurova

Agricultural and Biodiversity Coordination
Bioinformatician

Raheela Aslam

Agricultural and Biodiversity Coordination
Full Stack Developer

Contact:

faang-dcc@ebi.ac.uk

peter@ebi.ac.uk

*This project has received funding from the European Union's **Horizon** Programme under the grant agreement **n° 815668.***

BovReg *PARTNERS*

AARHUS UNIVERSITY

CRG Centre for Genomic Regulation

diagenode Innovating Epigenetics Solutions

edinethics

EMBL-EBI

ETH zürich

FBN

FACULDADE DE MEDICINA VETERINÁRIA UNIVERSIDADE DE LISBOA

INSTYTUT GENETYKI I HODOWLI ZWIERZĄT PAN

INRAE

LIÈGE université GIGA Medical Genomics

Université de Limoges

Luke LUONNONVARAKESKUS

MARTIN-LUTHER UNIVERSITÄT HALLE-WITTENBERG

ROSLIN

UNIVERSITY OF ALBERTA

THE UNIVERSITY OF EDINBURGH

Universiteit Utrecht

WAGENINGEN UNIVERSITY & RESEARCH

Thank you for your attention

www.bovregproject.eu

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 815668