Identification of functionally active genomic features relevant to phenotypic diversity and plasticity in cattle

Deliverable 3.3

Report on ontologies used across BovReg

Grant agreement no°: 815668

Due submission date

2020-08-31

Actual submission date

2021-02-26

Responsible author(s)

Peter Harrison peter@ebi.ac.uk

Daniel Zerbino zerbino@ebi.ac.uk

Confidential No

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668. The content of this report reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

DOCUMENT CONTROL SHEET

Deliverable name	Report on ontologies used across BovReg
Deliverable number	3.3
Partners providing input to this Deliverable	EMBL-EBI
Draft final version circulated by lead party to: On date	WP leader – Cedric Notredame 17/02/2021
Approved by (on date)	FBN as Coordinator (2021-02-26)
Work package no	3
Dissemination level	Public (PU)

REVISION HISTORY

Version number	Version date	Document name	Lead partner
Vs 1	17/02/2021	D3.3_Report_On_Ontologies	EMBL-EBI
Vs 2	18/02/2021	D3.3_Report_On_Ontologies_vs2	EMBL-EBI
Vs 3	18/02/2021	D3.3_Report_On_Ontologies_vs3	EMBL-EBI
Vs 3.1	21/02/2021	D3.3_Report_On_Ontologies_vs3.1	EMBL-EBI
Vs 3.2	24/02/2021	D3.3_Report_On_Ontologies_vs3.2	EMBL-EBI
Vs 4	25/02/2021	D3.3_Report_On_Ontologies_vs4	EMBL-EBI

Changes with respect to the DoA (Description of Action)

N/A

Dissemination and uptake

Public

Table of Content

1. Summary of results

The purpose of this deliverable is to catalogue and report on the ontology usage in metadata descriptions for the BovReg project. This is to ensure that the data produced by BovReg will be fully interoperable with other FAANG initiatives and in farm animal breeding and management and be aligned with best practice. Additionally, it is a first step towards development of high-quality ontology usage that will be a key component of ensuring accurate genotype to phenotype annotations. This ontology assessment shows the current ontology usage within the initial public sample and experimental submissions from the BovReg project.

From the first BovReg submissions, examples of ontologies used show the need for further development as existing descriptions do not fit use in cattle. For example, the BRENDA Tissue Ontology (BTO) term for milk gland is only suitable for use in fly projects yet is the only currently available term for this use.

This report demonstrates the known need within the project, and wider community, for improvements to ontologies for use in cattle agricultural genomics. The report also highlights the ontology terms added specifically to FAANG for the BovReg project to support single cell RNA-Seq metadata recording.

This report's findings flow naturally to the next stage of development within the Data Coordination Centre for the BovReg project, with development commencing on the FAANG Ontology Improvement Tool. The first version of this tool is due for release in summer 2021 (as a result of clustering activities across EuroFAANG, the three FAANG projects funded under the SFS-30- 2018 call). The catalogue of ontologies described in this deliverable, will be used to initially populate the ontology improvement tool with ontologies of key importance for the project and wider community. These key ontologies, as well as further terms proposed by users of the tool, will be collectively crowd-source reviewed for potential improvements. The tool will allow the community to propose new ontology terms, provide corrections to existing ontology terms to improve compatibility with farm animal / cattle usage and add new cattle specific synonyms.

It is worth noting that some BovReg submissions have been delayed by Covid-19 issues throughout 2020.

2. Introduction

This deliverable catalogues ontology usage with the first public International Nucleotide Sequence Database Collaboration (INSDC) FAANG data submissions from the BovReg and other cattle-related FAANG projects. It utilises the publicly available FAANG BovReg data as viewable on the BovReg data portal page https://data.faang.org/projects/BovReg. This assessment furthers our understanding of ontology usage within the project, with the aim of subsequently improving the ontologies to boost interoperability with the other FAANG initiatives (Harrison et al. 2018). This work will expand the suitability for cattle genomics of community ontologies such as the Experimental Factor Ontology (EFO; Malone et al.

2010), the NCBI Taxonomy (NCBITaxon; Schoch et al. 2020), the Ontology for Biomedical Investigations (OBI; Bandrowski et al. 2016), the Cell Line Ontology (CL; Diehl et al. 2016) and the Livestock Breed Ontology (LBO; https://www.animalgenome.org/bioinfo/projects/lbo/).

It is clear that there are missing terminologies and descriptions in the ontologies required for cattle genomics. These missing ontologies and improved ontology descriptions will be collected by the FAANG Ontology Improvement Tool for BovReg and the wider FAANG community.

3. Core report

3.1. Report on ontologies used across BovReg

Report on the ontologies used within BovReg samples and experiments submitted to the International Nucleotide Sequence Database Collaboration (INSDC) public archives as of 12th February 2021, and available from https://data.faang.org/projects/BovReg. It is worth noting that this represents a smaller assessment than originally planned as many datasets were delayed due to Covid-19 restrictions. A custom python script was utilised to extract the different ontologies used by the project from a metadata dump file produced by the FAANG data portal.

BovReg used 20 distinct ontology terms across their samples (Table 1) and experiments (Table 2), initial FAANG submissions, with a total of 1219 ontology usages within the metadata. This covers usage of the Ontology for Biomedical Investigations (OBI), the NCBI Taxonomy Ontology (NCBITaxon), the Phenotype and Trait Ontology (PATO; Gkoutos et al. 2005), the Livestock Breed Ontology (LBO), the Experimental Factor Ontology (EFO; Malone et al. 2010), the Uber-anatomy Ontology (UBERON; Mungall et al. 2012), the BRENDA Tissue Ontology (BTO; Gremse et al. 2011), the Cell Ontology (CL) and the Chemical Entities of Biological Interest Ontology (ChEBI; Degtyarenko et al. 2008).

A prolific problem with ontology usage in cattle genomics, and FAANG projects in general, is that ontology descriptions are heavily focussed on model organisms such as humans and flies, or on medical applications. Already in this initial submission from BovReg there are a number of ontologies applied whose descriptions are not accurate for use in cattle. An obvious example is from the BRENDA Tissue Ontology (BTO), with the use of 'milk gland' (BTO_0005704). The term itself is exactly what is required for the BovReg project, but the ontology description is so far only applicable to Drosophila projects:

"Larval nutrition is provided via a modified accessory gland, a milk gland, that empties into the uterus. The milk gland is connected to the dorsal side of the uterus and expands throughout the abdominal cavity of the fly as bifurcating tubules intertwining with fat body tissue. The lumen of the milk gland is surrounded by secretory and epithelial cells."

This is an example where BovReg either requires the ontology description to be updated to be more generic for usage in cattle, or the creation of a new term to cover FAANG requirements. There are a diverse range of ontologies that will require improvement for the BovReg project, as further and more complex datasets are generated. This includes ontologies covering developmental stage, feeding status, lactation, lactation stage, potential castration, weaned status, and organism parts from tissue collections. This deliverable's survey of ontology use in the early submissions of the BovReg project is just the first stage of the required improvements for the project. These requirements will be fed into the FAANG Ontology Improvement Tool as part of an ongoing collaboration with the other EuroFAANG projects (AQUA-FAANG and GENE-SWitCH).

Table 1. BovReg ontology usage for public INSDC sample submissions

Metadata question	Ontology text	Ontology term	Count
	Standard		
Material	specimen from organism	OBI_0001479	194
	organism	OBI_0100026	50
	cell specimen	OBI_0001468	2
	Organism		
Organism	Bos taurus	NCBITaxon_9913	50
Sex	male	PATO_0000384	24
	female	PATO_0000383	26
Breed	Cattle crossbreed	LBO_0001036	48
	Holstein	LBO_0000132	2
Health status	normal	PATO_0000461	50
	Specimen		
Developmental stage	adult	EFO_0001272	194
Health status at collection	normal	PATO_0000461	194
Organism part	jejunal mucosa	UBERON_0000399	48
	liver	UBERON_0002107	48
	rumen	UBERON_0007365	48
	skeletal muscle tissue	UBERON_0001134	48
	milk gland	BTO_0005704	1
	milk	UBERON_0001913	1
	Cell specimen		
Cell type	luminal epithelial cell of the lactiferous duct	CL_0002662	1
	cells isolated from milk	CL_0000548	1

Table 2. BovReg ontology usage for public INSDC experiment submissions

Metadata question	Ontology text	Ontology term	Count
	RNA-seq of total RNA		
experiment target	total RNA	CHEBI:33697	189

To assess ontology quality for future BovReg submissions we took the proposed tissues from the BovReg Description of Action (Table S1: BovReg tissues collected) as a test case. These tissue descriptions were run through Zooma an automated ontology annotation tool provided by EMBL-EBI (https://www.ebi.ac.uk/spot/zooma/). This application predicts what ontology terms are most appropriate for each given text description, and essentially for this report predicts what terms will be available for the project when these tissues are ready for submission to FAANG. The Zooma application also assigns a quality score for mapping confidence to each assignment. In every case the mapping quality was good.

The report highlights at least one term that is missing from the Uber-Anatomy Ontology (UBERON), the preferred ontology for FAANG for tissue descriptions. A term for "cerebrum cortex" will need to be added to the UBERON ontology before this tissue can be accurately recorded in the BovReg submission to FAANG.

Table 3. Automated ontology assignment to proposed tissues of the BovReg project

Tissue from Description of Action	Automatically assigned ontology label	Automatically assigned ontology term	Mapping Confidence
adrenal gland cortex	adrenal cortex	UBERON_0001235	Good
cerebellum	cerebellum	UBERON_0002037	Good
cerebrum cortex	cortex of cerebrum	FMA_83910	Good
colon	colon	UBERON_0001155	Good
duodenum	duodenum	UBERON_0002114	Good
heart	heart	UBERON_0000948	Good
hypothalamus	hypothalamus	UBERON_0001898	Good
ileum	ileum	UBERON_0002116	Good
jejunum	jejunum	UBERON_0002115	Good
kidney	kidney	UBERON_0002113	Good
liver	liver	UBERON_0002107	Good
lung	lung	UBERON_0002048	Good
lymph node	lymph node	UBERON_0000029	Good
mammary gland	mammary gland	UBERON_0001911	Good
ovary	ovary	UBERON_0000992	Good
pancreas	pancreas	UBERON_0001264	Good
pituitary gland	pituitary gland	UBERON_0000007	Good
rumen	rumen	UBERON_0007365	Good
skeletal muscle	skeletal muscle tissue	UBERON_0001134	Good

spleen	spleen	UBERON_0002106	Good
subcutaneous fat	subcutaneous adipose tissue	UBERON_0002190	Good
testis	testis	UBERON_0000473	Good
thyroid gland	thyroid gland	UBERON_0002046	Good
uterus	uterus	UBERON_0000995	Good

The diversity in required ontologies demonstrates the challenge faced for the project in coordinated ontology improvement for cattle genomics. This diversity results in improvements needing to be submitted to an array of different ontology providers, with different curation requirements, submission processes and timescales for responsiveness. This is why a centralised ontology tool for BovReg and the wider FAANG community is planned under the lead of EMBL-EBI to manage some of this administrative burden and track the status of required ontology changes.

Data from WP2 (BAM files on tissue assays, see D2.1 and D2.2) will be the first to necessitate further ontology terms used within the BovReg project in Spring 2021.

3.2. New ontologies added to FAANG for the BovReg project

The BovReg project includes single cell RNA-Seq experiments, that are novel to FAANG. This required the generation of new metadata rulesets and thus new ontology selections for FAANG. Table 4 lists the ontologies that were added to FAANG for use by the BovReg project.

Table 4. Ontologies added to FAANG for use in scRNA-Seq submissions

Metadata question	Ontology text	Ontology term
	Standard	
Material	Single cell specimen	OBI_0002127
	Organism	OBI_0100026
	Cell specimen	OBI_0001468
	Single Cell Specimen	
Cell type	Any cell type term from CL Ontology	CL_0000000
	scRNA-Seq	
Experiment target	Ribonucleic acid	CHEBI_33697

3.3. Next steps, the Ontology Improvement Tool

The catalogue of ontologies described in this deliverable, will be used to initially populate the proposed FAANG Ontology Improvement Tool with ontologies of key importance for the project and community. The first release of this tool is planned for the summer 2021, and new feature releases will be made through subsequent years of the project. The first version of the tool will include the ability to catalogue ontologies requiring improvement and for authenticated users to propose improvements to those terms. The tool will provide authenticated users with the opportunity to review ontology terms deemed important for the project by their usage in public submissions or to propose their own terms.

Terms will be voted as appropriate or flagged as requiring improvement by users of the service, this is expected initially to mainly be members of EuroFAANG and the wider FAANG community. Users with specific knowledge about the term will be able to provide suggested improvements or propose brand new ontology terms as required. Provenance of these suggested changes will be tracked. Collaborative editing and voting for approval will lead to terms being put forward to the ontologies to be updated or included, this final step will be delivered in subsequent versions of the tool.

Importantly, awaiting agreement on ontology improvements need not delay data submissions by BovReg, as the descriptions and synonyms of ontology terms can be updated independently of the use of their ontology codes in FAANG BovReg submissions. Additionally, if new terms are deemed more appropriate, submissions can be updated to use the improved term.

4. Conclusions

This assessment of the current state of ontology usage with the first submissions from the BovReg project has confirmed the need for improvements to ontology definitions for cattle genomics. This catalogue of ontologies provides a key initial set of ontologies for the FAANG Ontology Improvement Tool. This tool will be managed by EMBL.EBI to collate ontology corrections and additions from BovReg scientists and the wider FAANG and 1000 BGC communities to improve ontologies for use in cattle and other farm animal genomics.

5. References

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, et al. 2016. “The Ontology for Biomedical Investigations.” PloS One 11 (4): e0154556.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2008. “ChEBI: A Database and Ontology for Chemical Entities of Biological Interest.” Nucleic Acids Research 36 (Database issue): D344–50.

Diehl, Alexander D., Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, et al. 2016. “The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability.” Journal of Biomedical Semantics 7 (1): 44.

Gkoutos, Georgios V., Eain C. J. Green, Ann-Marie Mallon, John M. Hancock, and Duncan Davidson. 2005. “Using Ontologies to Describe Mouse Phenotypes.” Genome Biology 6 (1): R8.

Gremse, Marion, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. 2011. “The BRENDA Tissue Ontology (BTO): The First All-Integrating Ontology of All Organisms for Enzyme Sources.” Nucleic Acids Research 39 (Database issue): D507–13.

Harrison, P. W., J. Fan, D. Richardson, L. Clarke, D. Zerbino, G. Cochrane, A. L. Archibald, C. J. Schmidt, and P. Flicek. 2018. “FAANG, Establishing Metadata Standards, Validation and Best Practices for the Farmed and Companion Animal Community.” Animal Genetics 49 (6): 520–26.

Malone, James, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. “Modeling Sample Variables with an Experimental Factor Ontology.” Bioinformatics 26 (8): 1112–18.

Mungall, Christopher J., Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. 2012. “Uberon, an Integrative Multi-Species Anatomy Ontology.” Genome Biology 13 (1): R5.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, et al. 2020. “NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools.” Database: The Journal of Biological Databases and Curation 2020 (January). https://doi.org/10.1093/database/baaa062.