Proposals to establish genome annotation systems with phenome data and information

COST Action: EU-LI-PHE
Action Number: CA22112
Deliverable: D2.2
Work Package: WP2, Task 2.2: Expand genome information with phenotype data
Date: 31/09/2025
Prepared by: Claudia Kasper, Agroscope Switzerland and Emily Clark, The Roslin Institute, University of Edinburgh

1. Motivation

Findable, Accessible, Interoperable and Reusable (FAIR) data sharing has gained considerable attention due to its undeniable benefits for the scientific community, and interest from the livestock industry. In animal science, a wide spectrum of datasets are generated by diverse disciplines, including omics data (genomic, transcriptomic, proteomic and metabolomic), functional information (e.g. genome annotation, ChIP-seq and ATAC-seq) and phenotype data (e.g. production traits, nutrition, product quality, body composition and behaviour). Although omics data, particularly sequence data, are routinely deposited in publicly accessible repositories, the sharing of phenotype data remains rare, less standardised and often overlooked. Integrating phenotype data into existing genomic platforms would improve access to, and discovery and reuse of, livestock data. It would also facilitate the integration of phenotype and omics layers and promote the adoption of common standards, vocabularies and best practices for data management. This integration would allow researchers to improve genome annotation, generate new hypotheses, perform cross-study analyses and design studies with greater statistical power. It could achieved through displaying linked phenotype information and variation data in the genome browser Ensembl for example, which is currently available for humans via the G2P plug in for the Variant Effect Predictor (VEP). However, because data submission to the public repositories is currently fragmented this is difficult to achieve. Sequence data are deposited in repositories such as most commonly for Europe to the European Nucleotide Archive (ENA)[1] and globally to the National Center for Biotechnology Information (NCBI)[2], while phenotype datasets may be scattered across Zenodo, Dryad, or Figshare. Genotype data from SNP arrays are often inaccessible due to commercial or privacy concerns, and linking records (e.g. nucleic acid sequences from ENA) to individual phenotypes remains challenging. Metadata is often sparse, which further reduces reusability. The lack of unified standards and inconsistent application of ontologies result in heterogeneous data descriptions that hinder integration. Integrating phenotype data into existing public repositories through a standardised submission pathway would address these challenges directly, creating

[1] https://www.ebi.ac.uk/ena/browser/home
[2] https://www.ncbi.nlm.nih.gov/

a more coherent, comprehensive and sustainable ecosystem for sharing data in European animal science and globally.

2. Phenotype data availability

In this context, we define 'phenotype data' as the (qualitative or quantitative) measured or estimated values of an individual for a specific trait. This distinction is important because many platforms refer to phenotypes when they are actually referring to traits, such as those associated with specific QTLs on QTL platforms. In this report, 'phenotype data' refers to various types of individual-level observation that do not fall into the 'omics' categories. These may be single values per individual, or repeated assessments of the same trait over time or throughout an individual's lifetime. Examples include body composition at slaughter, methane production over a lactation period, and individually measured feed intake. Phenotypic data can exist in various formats, including images, videos and infrared spectra. However, they are often converted into two-dimensional tabular formats for analysis, particularly when genotype is associated with phenotype. These datasets can vary greatly in size, ranging from simple N × 1 vectors for single traits to complex N × k matrices (e.g. infrared spectra), where k can be in the thousands. Other examples include longitudinal data, such as feed intake or milk yield records, which involve many repeated measurements from the same individual over time.

Such datasets originate from various livestock species, including cattle, pigs, sheep, goats, poultry and horses, each of which has its own recording systems, trait definitions and industry practices. To ensure comparability, trait ontologies such as the Animal Trait Ontology for Livestock (ATOL)[3] and the Vertebrate Trait Ontology (VT)[4] are essential; however, their consistent application remains limited. Beyond the phenotype values themselves, rich metadata (e.g. management conditions, diet composition, housing, measurement protocols and environmental context) is critical for interpreting and reusing the data. As such developing sets of minimal reporting criteria are essential for phenotyping experiments including appropriate ontologies that can recorded at the point of data collection.

Sources of phenotypic data are diverse, ranging from experimental stations and abattoir records to on-farm sensors, computer vision systems, automated feeders and imaging platforms. However, the availability of data is often limited by commercial confidentiality, intellectual property concerns, or the absence of clear frameworks for data governance and sharing. These barriers result in a fragmented landscape where the availability, accessibility and interoperability of phenotypic datasets can vary significantly, meaning that many valuable resources remain underutilised.

3. User groups and use cases

3.1. Research

Integrated genome–phenome resources create opportunities for researchers to advance fundamental and applied science. Harmonised datasets support genome annotation, genotype–phenotype mapping and comparative analysis across studies and species. They also facilitate hypothesis generation, pilot studies and robust power and sample size estimation for experimental design. Beyond providing access to data, integrated platforms support the development of community-driven workflows and tools. Examples include nf-core pipelines[5], which promote reproducibility and collaboration. Examples of use cases include linking methane emission phenotypes to genomic data to identify causal variants of climate-relevant

[3] https://www.umrh.inrae.fr/ontologies/visualisation/public/
[4] https://www.animalgenome.org/bioinfo/projects/vt/
[5] https://nf-co.re/

traits or testing new genome annotation algorithms with standardised benchmark datasets to ensure comparability across projects.

3.2. Industry

Breeding organisations and industry stakeholders can use integrated data to improve genetic evaluation and breeding value estimation. Access to large-scale, standardised datasets is particularly valuable in helping the industry to incorporate sustainability traits, such as feed efficiency, methane emissions and welfare indicators, into breeding programmes. Aggregated data can also support the assurance of product quality in the production of milk, meat, and eggs. Ideally, industry is not only a user of data, but also a contributor. By sharing raw records or summary statistics under standardised pseudonymisation procedures, companies can strengthen collective resources while safeguarding sensitive information. Examples of use cases include aggregating across-farm feed intake data to refine breeding value predictions or combining abattoir-derived carcass composition data with genomic information to improve breeding indices across populations.

3.3. Policy and regulation

Policy and regulatory bodies represent a third user group with their own specific requirements. At the national level, agricultural and environmental authorities require reliable data to support greenhouse gas inventories, animal welfare monitoring and sustainability reporting. Similarly, at the European level, institutions such as the European Food Safety Authority (EFSA)[6], the European Environment Agency (EEA)[7] and the European Commission (e.g. Directorate-General (DG) AGRI[8] and DG CLIMA[9]) depend on robust datasets to inform food safety assessments, agricultural policy and climate strategies. At the international level, organisations such as the Food and Agriculture Organization of the United Nations (FAO)[10] and the International Committee for Animal Recording (ICAR)[11] rely on harmonised data to underpin global livestock sustainability assessments and recording standards. Examples of use cases include monitoring compliance with CAP eco-schemes using integrated resources, and providing validated datasets for international reporting on livestock greenhouse gas emissions.

4. Integration of data from different sources

Integration is essential to connect heterogeneous phenotype records, including individual trait measurements, imaging data, and sensor-derived time series, with genomic and other omics resources. Because livestock data vary across species, recording systems, and trait definitions, integration must address the following dimensions:

- Technical: shared data models, formats, and APIs
- Semantic: ontologies and controlled vocabularies to align trait definitions and metadata
- Procedural: standard operating procedures, consistent measurement units, and protocols
- Computational: methods that combine signals across assays and studies.

Clear governance and access rules are also required to enable the linking of sensitive or commercial data without compromising confidentiality.

[6] https://www.efsa.europa.eu/en
[7] https://www.eea.europa.eu/en
[8] https://commission.europa.eu/about/departments-and-executive-agencies/agriculture-and-rural-development_en
[9] https://commission.europa.eu/about/departments-and-executive-agencies/climate-action_en
[10] https://www.fao.org/home/en
[11] https://www.icar.org/

A range of analytical strategies exist to integrate phenotype and omics datasets, each addressing different layers of biological and technical complexity. Harmonisation and meta-analysis align study datasets through standardised metadata and trait mappings, allowing effects to be estimated across studies. Co-localisation tests whether association signals (e.g. GWAS and eQTL) share a causal variant. Transcriptome-wide association studies (TWAS) combine GWAS data with eQTL summary statistics to test the association between genetically predicted gene expression and phenotypes. Meanwhile, Mendelian randomisation (MR) uses genetic instruments to infer the causal relationship between molecular intermediates and traits. For high-dimensional data such as images, spectra and sensor streams, feature extraction and representation learning create stable, comparable features that can be linked to genotypes and other omics data. Livestock resources such as FarmGTEx[12] demonstrate how genotype–transcriptome–phenotype layers can be combined. Initiatives such as FAANG data portal[13] provide a foundation for metadata standards and interoperability, although further development is still required.

5. Platform requirements

In order to become widely used and attractive, a genome–phenome integration platform must address several key challenges. These include incomplete metadata, limited ontology uptake across species, difficulties in linking phenotypes with corresponding genotypes/omics, and uneven data accessibility due to commercial sensitivity. To overcome these barriers, the platform should provide the following:

- Tools for FAIRification, including validation against schemas, ontology lookups and measurement unit normalisation.
- Tools for resolving and linking individual animal IDs to connect individuals across repositories.
- Analysis pipelines for harmonisation, co-localisation, TWAS and MR, with reproducible provenance.
- Additionally, a policy and permission layer is needed to manage data usage agreements, offer various access levels (ranging from open to controlled) and maintain data usage audit trails.

User-friendliness is equally important for broad adoption. An attractive, widely used platform should have an intuitive landing page with clear options, such as 'Browse data' and 'Upload data', in prominent positions. Users should be guided through a short, step-by-step process with concise instructions, with links to more detailed guidelines provided where necessary. Accepted data formats should be communicated early on to avoid frustration. While ontologies and controlled vocabularies are essential for interoperability, they can be resource-intensive and difficult to navigate. Here, AI-based assistance, similar to OntoGPT[14], DIVE[15] or LexMapr[16], could help to reduce the burden and improve usability. Well-established resources such as dbGaP[17] demonstrate that clear workflows, accessible search functions and transparent access controls can make a complex platform appealing and functional for a broad user base.

Finally, the platform must be designed with longevity and sustainability in mind. Experience has shown that many portals become obsolete when funding ends or maintenance is discontinued. Therefore, ensuring long-term hosting, technical support and integration with

[12] https://www.farmgtex.org/
[13] https://data.faang.org/home
[14] https://zenodo.org/records/17154584
[15] https://tacc.utexas.edu/research/tacc-research/dive/
[16] https://www.cineca-project.eu/blog-all/lexmapr-a-rule-based-text-mining-tool-for-ontology-term-mapping-and-classification
[17] https://dbgap.ncbi.nlm.nih.gov/home

established infrastructures, such as ELIXIR[18], will be essential for lasting impact. To this end the ELIXIR Domestic Animals Genome and Phenome Focus Group was established in 2024[19].

6. Integration of genome annotations and phenotype data using Ensembl as an example

For integratation of genome annotations specifically to phenotype data one option is developing a plug-in for the Ensembl Variant Effect Predictor (VEP). There is currently an Ensembl VEP plug-in for human clinical phenotype data (VEP-G2P) that uses G2P allelic requirements to assess variants in genes for potential phenotype involvement[20]. This could be used a model but relies on having the allele frequency data and suitable phenotype information, across a sufficient number of traits available for livestock e.g. from the FarmGTEx[12] project.

Ensembl already displays multiple sequence to consequence predictions including SIFT and CADD[21]. SIFT predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids. SIFT scores are available for the main livestock species (Figure 1). The Combined Annotation Dependent Depletion (CADD) tool scores the predicted deleteriousness of single nucleotide variants by integrating multiple annotations including conservation and functional information into one metric. Phred-style CADD raw scores are displayed in Ensembl and variants with higher scores are more likely to be deleterious. Chicken and pig CADD scores are currently visible in Ensembl with turkey and cattle coming soon.

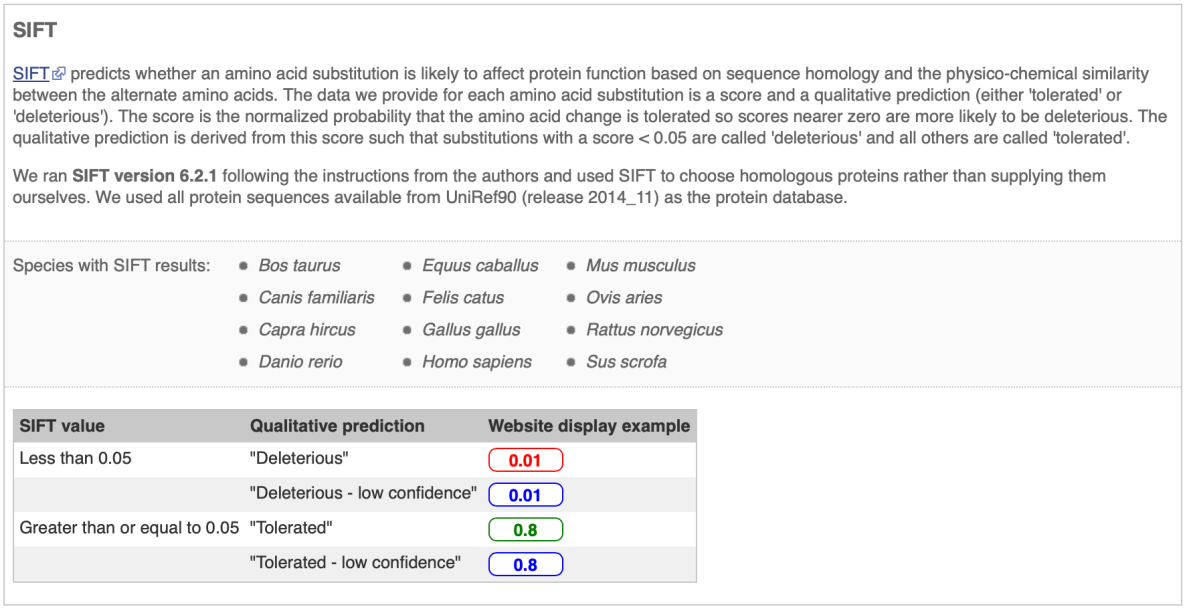

Figure 1: Livestock species with SIFT scores displayed in the Ensembl genome browser (screenshot 26/08/25)

[18]https://elixir-europe.org/
[19]https://elixir-europe.org/focus-groups/domestic-animals-genome-phenome
[20]https://github.com/Ensembl/VEP_plugins/blob/release/115/G2P.pm
[21]https://www.ensembl.org/info/genome/variation/prediction/protein_function.html

7. Concept

Figure 2 illustrates the proposed concept. It shows how phenotype and omics data from public and private sources feed into a genome-phenome platform, which provides tools, standards and user-friendly access. There, the data are combined via an integration layer to facilitate harmonisation, feature extraction and cross-layer association. The platform serves the needs of researchers, industry and policymakers, creating impact through improved genome annotation, data reuse and sustainable livestock production.

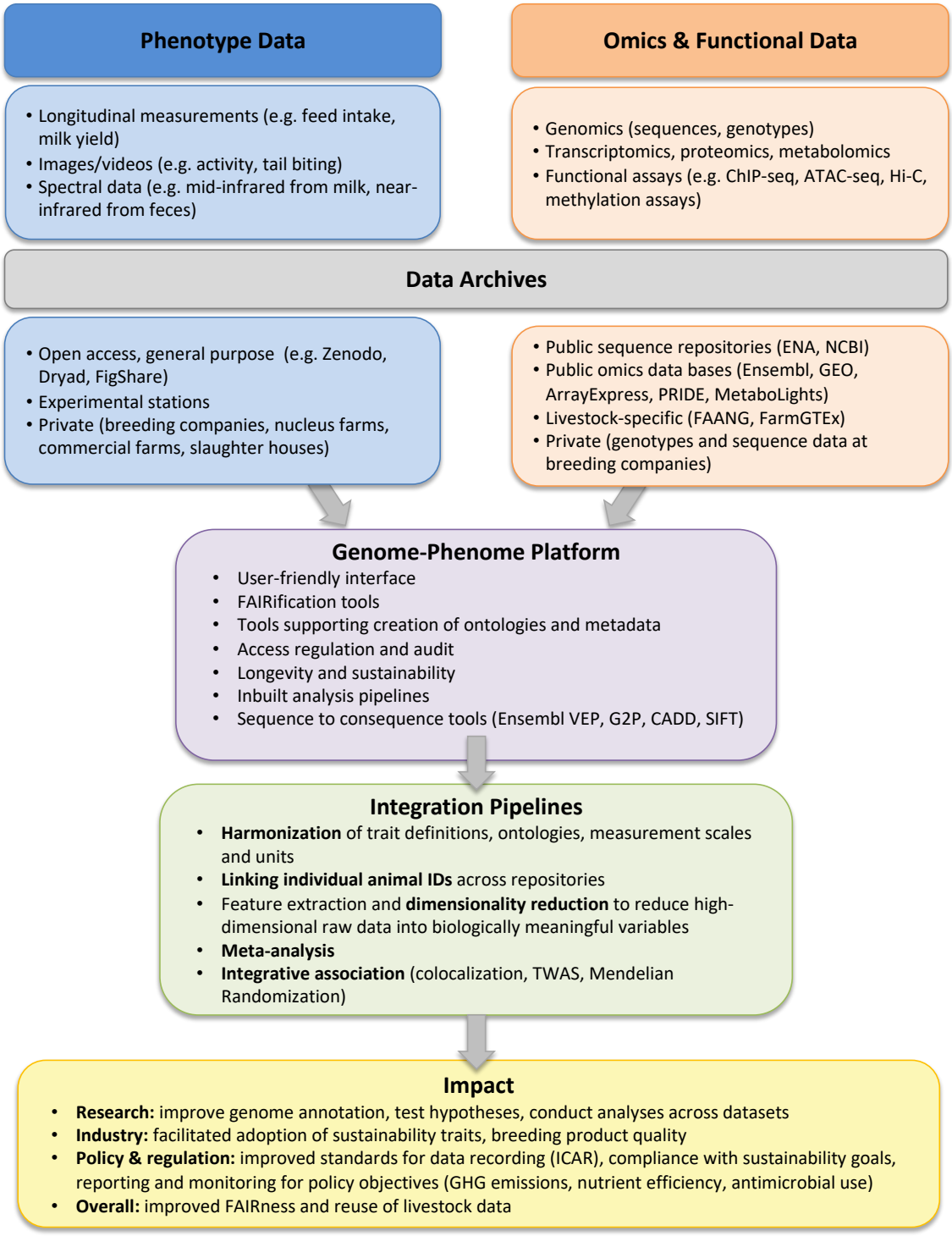

Figure 2: Flow chart describing a route to integrate genome annotations with phenotype data through the different data types, archives, platforms, pipelines to deliver impact.

8. Conclusions

Integration of phenotype data with genomic annotations is possible but requires significant effort towards standardising the data submission pathways for phenotyping data, data recording and accessibility. There are challenges to overcome particularly around sharing of proprietory data to be able to link genome and phenome datasets at sufficient scale to be effective. Once these are overcome however there are a number of options, for example, through the VEP tools in the Ensembl genome browser, and new machine learning alghorithms, to link genomic annotation systems with phenotype data and information. There is also considerable momentum and enthusiasm within the livestock genomics and phenomics communities to do this.

Acknowledgements: This work was supported by COST (European Cooperation in Science and Technology) Action CA22112 and an STSM awarded to Dr Claudia Kasper. We thank the members of Working Group 2 for discussions that helped to formulate this deliverable.