Peer Community Journal

Section: Animal Science

Opinion / Perspective

Published
2025-12-01

Cite as
Guillaume Devailly, Sonia E Eynard, Chloé Cerutti, Arthur Durante, Jean-Noël Hubert, Keyvan Karami, Noémien Maillard, Denis Milan, Mathilde Perret, Frédérique Pitel, Annie Robic, Juliette Riquet, Stacy Rousse, Elena Terenina and Julie Demars (2025) *The future of systems genetics in farm animal sciences, a route out of the data jungle*, Peer Community Journal, 5: e132.

Correspondence
guillaume.devailly@inrae.fr
julie.demars@inrae.fr

Peer-review
Peer reviewed and recommended by PCI Animal Science, https://doi.org/10.24072/pci.animsci.100353

The future of systems genetics in farm animal sciences, a route out of the data jungle

Guillaume Devailly [#,1], Sonia E Eynard [#,1], Chloé Cerutti [1], Arthur Durante [1], Jean-Noël Hubert [1], Keyvan Karami [1], Noémien Maillard [1], Denis Milan [1], Mathilde Perret [1], Frédérique Pitel [1], Annie Robic [1], Juliette Riquet [1], Stacy Rousse [1], Elena Terenina [1], and Julie Demars [1]

Volume **5** (2025), article e132

https://doi.org/10.24072/pcjournal.653

Abstract

Farm animal species are under intense selection on relatively small population sizes. Genetic and genomic selection has provided remarkable genetic gains in the last century. Nevertheless, current methods aiming to link genome to phenome in such populations remain limited, notably due to the difficulty to identify causal variants for complex traits. The diversity of species as well as breeds in livestock has diluted the number of genomic datasets available for each genome as compared to model organisms or human diseases. In this article, we propose a systems genetics approach as an opportunity to go beyond current limits and find a way out of the data jungle, taking advantage of novel computational development allowing integration of omics datasets from different analyses across species. A major challenge is that systems genetics requires careful but efficient data and metadata management, as well as rigorous statistical strategies on which approach to use. Here, we highlight examples of the broad contribution systems genetics can bring to farm animal sciences, particularly across species, notably in the genome-to-phenome field within the larger scope of agricultural challenges, including adaptation to environmental changes and animal welfare.

[1]Université de Toulouse, ENVT, INP, INRAE, GenPhySE, Castanet-Tolosan, France, [#]Equal contribution

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Introduction

Systems genetics has first been described by Civelek and Lusis, 2014 as an approach allowing for the understanding of complex traits, using intermediate phenotypes, from different omics technologies. This approach was forged from the massive development of high-throughput omics technologies and computing facilities making the treatment of generated data accessible. This approach was first geared towards model species analysis, with the aim to translate discoveries to human science, in particular to improve knowledge on diseases. Other recent studies have used systems genetics strategies to establish a link between genome and phenome in humans (Allayee et al., 2023; Weeks et al., 2023). Additionally, there has been a recent bloom in cross-species systems genetics as a promising way to characterize the molecular architecture of complex traits using data coming from experiments on multiple species (Jurrjens et al., 2023). Yet, this concept remains mostly geared towards the assessment of complex traits in humans, with the help of knowledge on model species, such as mice. There is a limited number of such studies focusing on farm animal species despite twenty years of livestock genomics research.

Farm animal species have limited effective population sizes and are under intense selection pressure in controlled environments in opposition to wild animals, model species or even humans. Farm animal species in particular are diligently studied as models to understand the genetic mechanisms underlying complex phenotypes. Unlike humans, they often show large family sizes, leading to a strong structure in the population, generally resulting in multiple isolated populations, referred to as breeds. This can be observed at the genomic level by a strong linkage disequilibrium (LD, see lexicon). On the other hand, unlike in wild species, these structural challenges have the advantage of often providing substantial and accurate records of their evolutionary history and phenotypic performances, allowing a certain control over experimental design. Finally, unlike model species, farm animal species mostly inhabit natural environments, considerably more representative than laboratory conditions of the realistic ecological interactions between all genera and representing a large diversity of population histories. The combination of such benefits and challenges related to the study of farm animal species makes it crucial to define suitable strategies to apply systems genetics approaches (Loy et al., 2024; Weller, 2016).

Sustaining the growth of farm animals production for feeding the still increasing worldwide population is a crucial challenge for the agronomic sector. However, this challenge encompasses dealing with the impact of climate change, eutrophication and deforestation, managing competition between feed and food production, sanitary considerations linked to animal production, welfare improvements, and the economic considerations for consumers and producers (farmers and breeders). Reduction of meat consumption is often advocated as a means to reduce human carbon footprint (Ritchie, 2020; Van Zanten et al., 2018), but meat production and consumption is still growing worldwide (FAO, 2022). Changes in animal production practices are essential to achieve sustainable animal production and answer socio-economic requests. Scientific knowledge is expected to provide solutions and innovations to reach this goal.

To date, genetic selection in livestock has already contributed to considerable improvements in production traits. For example, in poultry the increase in performance has been remarkable in both broilers and layers (Aggrey et al., 2020), with growth rates in broiler chickens increased by 400% in 50 years (Zuidhof et al., 2014), cattle milk production increased by 50% in the past 25 years (Lu, 2017; Von Keyserlingk et al., 2013), and pigs growth rate increased of more than 300% in 45 years (Bidanel et al., 2020). In fact, the economic importance of farm animal species

has already led to the development of significant genomic resources, as illustrated by the proportion of high-quality reference genome assemblies for domesticated animal species (Li et al., 2021; Morris et al., 2020; Rice et al., 2023; Talenti et al., 2022; Warr et al., 2020). Therefore, to adapt selection criteria to new challenges such as climate and social changes, the integration of new and more complex phenotypes as well as novel genomic data has already opened new avenues. Genomic selection, in which the genetic value of breeding animals is predicted from their genotypes using a reference population of phenotyped and genotyped animals, has been replacing selection schemes based on the animal's phenotypes. It has resulted in accelerated genetic gains, but comes with its own challenges, such as the risk of deterioration of secondary traits (Misztal and Lourenco, 2024), loss of genetic diversity (Doublet et al., 2019; Eynard et al., 2016) or the difficulties in accounting for the functional relevance of genetic variants (Boichard et al., 2015). Better understanding the molecular determinism of traits might alleviate some of these limitations. Linking genomes to phenomes in the post-genomic era is becoming more and more feasible and affordable for farm animals thanks to the increasing knowledge of livestock genomes through international consortia initiatives (FAANG (Clark et al., 2020), AG2PI (Tuggle et al., 2022), FarmGTEX (Fang et al., 2025; Guan et al., 2025; Liu et al., 2022; Teng et al., 2024)).

Although many genomic regions have already been associated with a broad range of phenotypes of agronomic interest in livestock, the molecular architecture of such complex traits still remains to be elucidated. Systems genetics strategies relying on multi-directional information flows between species and populations, integrating data from domestic animals in community-led projects (Clark et al., 2020; Destoumieux-Garzón et al., 2021), driven by biological relevance, may be an option to link genome to phenome. In this article we propose to briefly revisit the historical approaches, their limits, and to offer a new perspective breaking down the current compartmentalization of knowledge (by species and/or population). Finally, we discuss opportunities and challenges to leverage cross-species systems genetics for the future of farm animal science.

Historical approaches and their limits

So far, the most common historical approach often consists in analysing data in silos. Occurring naturally through time, data silos correspond to a collection of data, often concentrating on a focal species, owned by one organization and not necessarily made accessible to others (Tuly et al., 2025). This data structure makes the extrapolation to systems genetics difficult if not impossible. In general data silos are built and analysed in two ways, *i)* based on candidate genes, which can be seen as a bottom-up approach, starting from genome and for which the purpose is to identify genes based on information regarding their function (Zhu and Zhao, 2007) and *ii)* based on the analysis of the association between phenotypes and genotypes, which can be seen as a top-down approach, starting from the phenome such as GWAS and fine mapping of Quantitative Trait Locus (QTL, Box 1).

Having access to datasets, often in the form of silos, linking genome to phenome remains a challenge. The majority of standard associations and linkage genetic approaches mostly relying on data silos have struggled so far to identify causal variants for complex traits (Burnett et al., 2020; Tam et al., 2019), due to multiple limitations:

i) Limited statistical power and the resulting risk of detection bias: on the one hand potential false-negative, absence of detection and on the other hand, false-positive findings. For example,

restricted datasets can be sufficient for the analysis of binary traits whereas extensive or even combined datasets are required to provide significant results on polygenic traits; *ii)* The LD value around the causal mutation (i.e. the size of the genomic segment associated with the mutation). For example, experimental designs based on familial datasets often result in the detection of large candidate regions including a high number of putative causal variants (Uffelmann et al., 2021). Dedicated crosses such as back-crosses can be attempted in order to break the linkage disequilibrium between the different polymorphisms and the causal mutation. In fact, as the size of the LD decreases, the QTL localization interval also decreases. However, when analysing data with low LD structure, it is necessary to have a high density of SNPs in order to achieve a statistically significant association; *iii)* Lack of annotation of reference genomes, making it difficult to evaluate variant effects. Indeed, the majority of genetic variants (>80%) affecting complex traits are located outside coding regions (Edwards et al., 2013; Maurano et al., 2012; Tam et al., 2019; Xiang et al., 2019), which impairs our understanding of the genotype-phenotype link; *iv)* Other particular cases of genetic determinisms such as non-additive or non-Mendelian determinism like genomic imprinting (with the example of the *IGF2* gene shown by Van Laere et al., 2003, or the single base mutation on the *CLPG* gene (Freking et al., 2002)), allelic heterogeneity (Bellinge et al., 2005), combination of several alleles through epistasis (Demars et al., 2022) and indirect genetic effects (Baud et al., 2022) further complicate the identification of causal variants (Fisher and Lewis, 2008); *v)* The difficulty of genotyping structural variants (SVs). Most large-scale studies focus exclusively on SNPs and discard structural variation occurring on the genome between individuals and across species. Pangenomes open up new avenues to take this crucial variability into account in the identification of causal variation (Secomandi et al., 2025).

Box 1 - Genetics lexicon

GWAS (Genome Wide Association Study) - Study focusing on the identification of statistical association between a trait of interest within a dedicated group of individuals and the genomic variants (e.g. SNPs) that they carry. A GWAS generally explores the observed population-level variation in allele frequency at the nucleotide resolution depending on the performance of individuals for a given trait.

LD (Linkage Disequilibrium) - LD is the non-random association of alleles at multiple points in the genome. High LD is seen as a higher frequency of association between alleles than what is expected as random and is often the result of a physical proximity between the positions, and sometimes the preferential association between alleles from two distant loci.

QTL (Quantitative Trait Locus) - Specific genomic position, locus, correlating with variation in a trait of interest within a population. QTLs are often identified through GWAS (or linkage analyses within familial pedigrees) and are a first step to pinpoint causal genes and/or mutations.

Pangenome - Union of all genomes of a species. The pangenome refers to the entire set of genes within a species containing sequences shared between all individuals of the species. Pangenomes across species are referred to as super-pangenomes.

SNP (Single Nucleotide Polymorphism) - Genomic variant caused by the change of a single base position in the genome.

Regardless of the aforementioned challenges, the growing availability of large datasets in different species, combined with the constant evolution of methods, brings new opportunities

to identify molecular mechanisms involved in the variability of complex traits. The initial approach to benefit from data silos consists in combining outputs from different data types and studies into meta-analysis. In a second time, systems genetics approaches propose to take advantage of different omics datasets, including genetics/genomics, epigenomics, transcriptomics, and metabolomics, to deepen our knowledge of biological processes and identify both candidate genes and variants (Civelek and Lusis, 2014; Threadgill, 2006). Cross-species systems genetics strategies may pave the way to systematic interpretation across diverse types of data over a multitude of organisms (humans, domesticated animals and wildlife species) (Kelley, 2020; Minnoye et al., 2020).

New perspectives

Two approaches

We propose to group current and future work in farm animal systems genetics in two main complementary approaches (Williams and Auwerx, 2015) (Figure 1). These approaches are based on the FAIR (Findability, Accessibility, Interoperability, Reusability) principles of data accessibility (Wilkinson et al., 2016), which became increasingly important in animal science in recent years. FAIR data sharing principles have contributed to make a broad range of datasets available to a larger research community through public databases, limiting the need to generate *de novo* extensive datasets.

One approach, here called “phenotype-driven”, focuses on specific animal experimentation carefully designed to study one or several traits of interest. While the interest was in the past mainly in productivity, it now tends to include sustainability, welfare, animal behaviour and the environmental impact of farm animals, as mentioned in the last report of the Food and Agriculture Organization (Hendriks et al., 2025). Experimental designs dedicated to the analysis of traits of interest produce rich and heterogeneous datasets on the same animals, composed of pedigree, deep phenotyping, multi-omics characterisation, including genotyping or sequencing, epigenomics, transcriptomics, metabolomics. The individual exploitation and integration of these multi-omics datasets aim to produce new knowledge regarding both the molecular architecture and biological mechanisms underlying the traits of interest (Akiyama, 2021; Hasin et al., 2017; Kreitmaier et al., 2023; Lim et al., 2023).

Let us extrapolate and describe an optimal design for such a study. One could want to investigate the impact of husbandry practices for housing outdoor rather than indoor on farm animals such as pigs, poultry or rabbits. We could design an experiment where animals would be raised either indoors in conventional breeding facilities, or with outdoor access in traditional or innovative systems. Complex traits, such as health, welfare and productive capacities, could be analysed by contrasting molecular phenotypes (i.e transcriptomes of different tissues, epigenomes, microbiomes) between systems. One could correlate breed diversity, through their genetic background, to the collected molecular phenotypes. Integrating such heterogeneous data would help identify biological mechanisms linked to our traits of interest which could ultimately contribute to a better design of selection plans or farm management.

Going one step further, a novel perspective lies in the cross-species integration of such research results to identify universal or specific biological mechanisms driving population evolution for traits of interest. To date, most examples of cross-species systems genetics are developed to study traits of interest for humans. Ashbrook et al., 2019 combined GWAS on human cohorts

and a linkage study on a mouse model to accurately identify novel candidate genes (*APBB1IP*) impacting schizophrenia. Similarly, Calabrese et al., 2017 integrated a GWAS in humans and co-expression network in mouse to identify two genes involved in bone mineral density in human and Komljenovic et al., 2019 identified genes, processes and networks linked to aging are conserved throughout evolution in four species. Recently, other studies have emerged integrating transcriptomics data across species at the single cell level (Song et al., 2023).

The other approach, here called "data-driven", consists in using already available data to investigate new relevant questions. In this article we propose to call this approach "data-driven", but it has previously infamously been qualified as "data parasitism" (Longo and Drazen, 2016). Rich datasets are already available for farm animals, either publicly or within individual organizations and they can be re-visited to investigate new research prospects, either on their own or jointly with other datasets. An example of this approach would be the use of genomic selection datasets to identify deleterious recessive mutations, looking for regions violating Hardy-Weinberg equilibrium, as proposed by several groups (Ben Braiek et al., 2024a,b; Cole et al., 2025; Hayes and Daetwyler, 2019; Jourdain et al., 2023; VanRaden et al., 2011), who identified candidate genes for recessive deleterious mutations in sheep and cattle. The newly generated knowledge can benefit the selection schemes by avoiding risky mating, but will also enrich functional gene annotation, and could provide new animal models for known human diseases. For example, Pan et al., 2021 proposed an enriched pig genome functional annotation as a tool for the biological interpretation of complex traits and diseases in humans. The availability of artificial intelligence, especially through deep learning and machine learning, applied to data shared in a FAIR way will greatly benefit this data-driven approach. Indeed, it is becoming possible to predict regulatory sequence activity using cross-species strategies (Li et al., 2024; Minnoye et al., 2020) where the genome annotation for a focal species can be predicted accurately using much more robust annotations in model species or in humans. DNA language models are continuously developing to predict more and more precisely the effect of variants (Benegas et al., 2023; Jagota et al., 2023), highlighting a renewed need for open datasets to feed the development of innovative methods.

In summary, the "phenotype-driven" approach will often work on deep, heterogeneous datasets produced on the same set of animals, while the "data-driven" approach will integrate more homogeneous data types on a large number of animals. Both approaches depend on an extensive re-use of datasets made available using the FAIR principles, which will contribute to the limitation of animal experimentation. In recent studies the same number of animals have produced richer datasets than before, allowing for a deeper analysis per animal providing more knowledge.

The era of data abundance

Farm animal genomics has now entered an era of data abundance. High quality reference genomes are available for all species of economic interests, and the field is now moving into breed-specific, or even animal-specific, reference genomes, pangenomes and even super-pangenomes (Gong et al., 2023; Secomandi et al., 2025; Smith et al., 2023). Animal genotypes are relatively cheap and easy to acquire thanks to mid- and high-density SNP arrays, genotyping by sequencing (GBS) technologies or low-pass sequencing (Lloret-Villas et al., 2023). Genomic selection datasets can be accessible for researchers and genotypes are being increasingly imputed to the whole-genome sequence. These imputations can be of extremely high accuracy if the reference haplotypes are available (Wragg et al., 2024), as for mainstream livestock breeds in cattle, pigs or poultry. In addition, more and more whole-genome sequence data are made available and

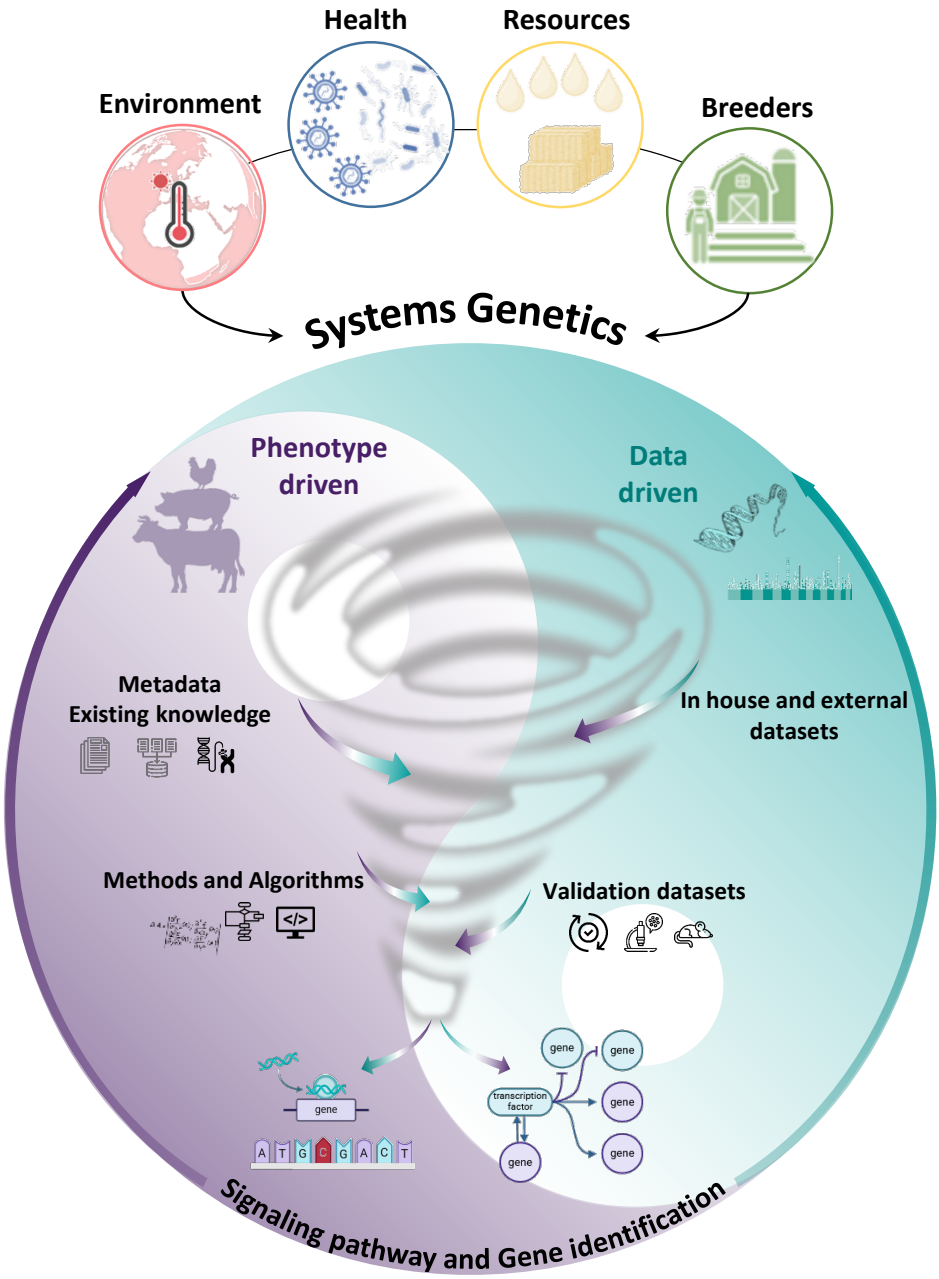

Figure 1 – Systems genetics strategies as a lever for linking genome to phenome to face agriculture challenges. The agricultural sector is evolving to adapt to the growing constraints (climate change, its impact, such as the development and spread of pathogens, or the availability of natural resources like water and feed) for a more sustainable agriculture for farmers, breeders and society. While socio-economic changes are essential to achieve sustainable animal production, scientific knowledge is expected to provide solutions. Systems genetics approaches, "phenotype-driven" and "data-driven", may contribute to gradually improving the link between genome and phenome. These strategies strongly rely on existing knowledge through the availability of datasets (in house and/or external) and their metadata, following FAIR principles. Appropriate methods and development to integrate multiple scales of heterogeneous data are needed, as well as validations using different means like new datasets, other species, in-vitro experiments or other algorithms. The novel knowledge generated by identifying molecular pathways and/or mutations involved in a particular function of a trait will progressively help understand its underlying biological architecture.

many studies highlighted the expected benefit of using a pangenome graph rather than a single reference genome for alignment and GWAS (Li et al., 2022; Secomandi et al., 2025; Teng et al., 2024). Pangenomes allow for a better representation of the genome diversity and facilitate the analysis of large genomic variations, especially SVs which are crucial for understanding the biological mechanisms of many complex traits. In addition, functional genomic data are broadly available thanks to large scale data collection projects led by laboratories (Jin et al., 2021; Pan et al., 2021; Teng et al., 2024) and consortia (FAANG (Clark et al., 2020), AG2PI (Tuggle et al., 2022), FarmGTEX (Fang et al., 2025; Guan et al., 2025; Liu et al., 2022; Teng et al., 2024)). Altogether, high-throughput data generation leads to rich datasets that may not be entirely analysed and presented within a single publication.

However, a specificity of farm animal science is that it is confronted with a diversity of species and breeds with subpopulations reproductively isolated at the national or local levels, therefore affecting the distribution of genetic variation. Mainstream commercial breeds can remain inaccessible to the scientific community due to its economic interests. On the contrary, local, traditional breeds may not have the economic interest necessary to convince funders even though they tend to raise more and more attention due to their specific characteristics and capacities to be adapted to specific environments. Reinforcing public/private partnership is central to the effort to build a systems genetics approach, as already evidenced in plants (Lozada et al., 2022). This starts at the phenotypic level, with the availability of accurate phenotypes but rarely shared among the community because of private interests. Indeed, while more and more omics datasets are subjected to FAIR principles (Wilkinson et al., 2016) with associated metadata, even for commercial breeds, it is not always the case for phenotypic data. One drawback of data abundance is the financial and ecological cost (Price et al., 2024) for analysing and storing such datasets, which can overcome the acquisition cost. This era of data abundance manifests itself more as a lush jungle, hard to navigate, than an ordered warehouse. To make the most of it, we believe that researchers embarking in systems genetics need to hold a good knowledge and tracking of the literature, of the available databases and datasets, as well as excellent data cleaning and data stewardship skills.

Navigating the metadata jungle

Whilst relevant datasets for systems genetics are now widely shared by researchers, metadata can often hinder their re-use by other teams in several ways.

1. Finding datasets in a fragmented landscape: data types of the same nature are shared through distinct repositories. For example, DNA sequence data can be deposited in the NCBI's SRA (SRA), in the EBI's ENA (ENA), in DDBJ's DRA (DRA), or in CNCB's GSA (GSA), among others. Knowing the main repositories and querying each one of them to find relevant data is essential.
2. While within each repository metadata are reasonably homogeneous, metadata schemes are often heterogeneous. Fields may differ and it may be impossible to perform a field-for-field equivalence. Mandatory fields in one database can be deemed optional or absent in another. Fields may not use the same controlled vocabulary (ontology, Box 2), or may not use them at all. Reusing public data requires dedicated time for metadata cleaning and harmonization.

Yet, data repositories are actively fighting these issues. For example, the INSDC (Arita et al., 2020) (INSDC) synchronized efforts of the NCBI, EBI and DDBJ in sharing DNA sequence, so that data deposited in one repository is visible and accessible through the other repositories. We can also mention Animal Trait Ontology for Livestock (Animal Trait Ontology) or CorrDB (CorrDB)

as ontologies relevant to farm animal studies. Databases such as Cell Ontology (Cell Ontology) offer a valuable platform for single cell omics data analysis. In addition, scientific communities are building secondary portals with well-defined scopes allowing the sharing of curated datasets. For example, we can cite IHEC, a portal providing human and mouse epigenomic datasets (Bujold et al., 2016), FlyBase, a repository of genetic and molecular datasets for *Drosophila melanogaster* (Gramates et al., 2022), or FAANG, a global network helping researchers to standardize experiments, coordinate data sharing and provide infrastructures for data analysis (Harrison et al., 2021). Each project needs to find the balance between exhaustiveness and curation levels on data quality and metadata standards. For these reasons, navigating the metadata jungle often requires a thorough literature review to identify relevant and maintained databases, followed by necessary but laborious metadata harmonisations between the retrieved datasets.

Box 2 - Ontology for metadata and further analyses

Ontology aims at unifying the various scales of biology through an integrated common dictionary. Ontology is essential for all the levels of phenotypes going from macroscopic phenotypes, such as traits like weight, to molecular information such as biological pathways. Indeed, meta-analyses comparing and combining datasets rely on common vocabulary. For macroscopic phenotypes, Animal QTLdb strives to collect all publicly available trait mapping data, i.e. QTL (phenotype/expression, eQTL), candidate gene and association data (GWAS), and copy number variations (CNV) mapped to livestock genomes, in order to facilitate locating and comparing discoveries within and between species (QTLdb). For molecular phenotypes, the DAVID knowledgebase provides a comprehensive set of functional annotation tools for researchers to understand the biological meaning behind large lists of genes (DAVID) and QTLbase2 curates and compiles genome-wide QTL summary statistics for many human molecular traits across over 95 tissue/cell types and multiple biological conditions (QTLbase2). These efforts for common ontologies at various levels open the way to the integration of different datasets for meta-analyses and even cross-species studies.

Analysis superficiality and the methodological bloom

Systems genetics relies on large, heterogeneous datasets that need to be handled with dedicated methods (Roesch et al., 2023). Statistical detection power is constrained by the limited number of available samples for a specific data type. Noisy datasets can impair the robustness of statistical analyses. Outliers, miss-labelling, miss-alignments, and/or miss-annotations can lead to highly significant false-positive results. Methods less sensitive to outliers are often less powerful and can result in false-negative discoveries. A common way to reduce the bias lies in the normalization of datasets, for example by transforming, scaling and standardizing different data units into a known uniform scale. Normalization is not trivial to apply and the method needs to be chosen with care. Some packages such as 'moments', in R (Komsta and Novomestky, 2022), might help researchers make the most informed choice. Network approaches allow heterogeneous data integration and visualization, as do other multi-block data integration approaches (see the popular WGCNA (Langfelder and Horvath, 2008), mixOmics (Rohart et al., 2017), and MOFA2 (Argelaguet et al., 2018)). The identification of transcription factors responsible for the regulation of gene expression modules can be inferred from co-expression gene networks (Chen et al., 2018), and can lead to cross-species general biological pathway discoveries. The approach

proposed by Kuijjer et al., 2019, LIONESS, or by Weighill et al., 2022, EGRET, offer the possibility to identify species-specific or genotype-specific pathways in large aggregated datasets. This field of study is highly active and multiple resources are available, as reported by Ben Guebila et al., 2022. Machine learning methods allow powerful data augmentation, including cross-species predictions to increase the amount of usable data (Kelley, 2020; Minnoye et al., 2020).

One prevalent issue when working with large scale heterogeneous datasets is the statistical "double dipping/double filtering" phenomenon, that we will illustrate through two approaches:
i) Statistical differences between clusters - when comparing two clusters one can be tempted to use the same data for clustering and for testing the differences between clusters, making results potentially entirely artifactual (Song et al., 2025). One strategy, called data thinning (Neufeld et al., 2024), aims to avoid such statistical limitations by dividing the data in two sets, one used to fit the model and estimate parameters and the other to validate it. Other approaches consist in developing models taking into account such redundancy in the data, as in Gao et al., 2024, which proposes the use of a selective inference to test differences in means across groups while controlling for type I error by taking into account that the null hypothesis was extracted from the initial dataset;
ii) Controlling for false discovery rate (FDR) - as for all statistical tests involving multiple testing FDR approach should be favoured compared to simple p-values. In addition, once a statistical procedure results in a list of features under an FDR threshold, it is unlikely that arbitrary subsets of this list have the same FDR. In fact, significance and FDR threshold need to be re-computed for any given dataset or combinations of datasets, or lists, and cannot be straightforwardly compared or concatenated across experiments. In omics data analysis, this issue can occur when applying a fold change threshold on a list of differentially expressed genes (Enjalbert-Courrech and Neuvial, 2022), or even when crossing two lists of features found significant in two distinct datasets. Dedicated methods to overcome this limitation are developed and implemented, especially in the field of systems genetics, for a large spectrum of analyses. We can list some examples of such: (m)ashr (Stephens, 2017) uses shrinkage-based estimates adapted to each dataset and sanssouci (Durand et al., 2020) uses user-defined or data-driven post-hoc inference for multiple testing.
It can be challenging for individual researchers and teams to keep up with the systems genetics methodological bloom. While testing new methods, in the absence of a full understanding of the underlying hypotheses and procedures, an option could be the use of controlled negative datasets (e.g obtained by random permutations of existing datasets) to detect false-positive outcomes (Han et al., 2009; Kunert-Graf et al., 2021). On the contrary, positive controlled datasets can also be artificially constructed (e.g. through simulations), and are especially useful when conducting benchmarking studies (Syed et al., 2021; Wharrie et al., 2023). Another approach could be to validate results from one subset of a dataset to another independent replicate dataset, although this approach often requires double the amount of data (or a reduced statistical power of the study by two) (Gallitto et al., 2023; Ioannidis, 2005). Rigorous statistical analysis practices are necessary to avoid, or at least limit, the so-called "reproducibility crisis" in the field of animal systems genetics where the results of a specific analysis cannot be repeated or transcribed.

Case studies in systems genetics

Researchers working in animal science hold powerful models to contribute to the growth of systems genetics. Systems genetics has been applied to establish new animal models in the pathophysiology of specific diseases (Calabrese et al., 2017), identifying convergent non-invasive phenotypic proxies (like hair rather than blood to measure cortisol level to test for stress) (Burnett et al., 2015) or determine conservation at the molecular pathways and gene family level despite differences at the gene level (Komljenovic et al., 2019). The success of systems genetics approaches relies on the availability of portals or databases integrating various types of analyses such as GeneNetwork (Sloan et al., 2016) and FarmGTEX portals (Fang et al., 2025; Guan et al., 2025; Liu et al., 2022; Teng et al., 2024). GeneNetwork started as a database to centralize phenotypes and genotypes on the historic BXD mice (Peirce et al., 2004; Rosen et al., 2007; Wang et al., 2003) that has been intensively studied for decades, but since then it has grown to present data from other datasets and species. The philosophy behind GeneNetwork is to gather and curate coherent datasets to allow cross-dataset integration and easy validation of findings. GeneNetwork provides various analysis types, but allows researchers to decide for crucial analysis parameters which can drastically change the results, and therefore still requires some training. Nonetheless, the ability to analyse multiple curated systems genetics datasets with a single tool is a phenomenal accelerator of discovery. The FarmGTEX open consortium (Fang et al., 2025; Guan et al., 2025; Liu et al., 2022; Teng et al., 2024) provides publicly available data that have been carefully gathered and re-analysed to provide additional information not directly extractable for the individual studies. As an example, FarmGTEX members gathered an extensive amount of RNA-seq data in farm animal species from multiple tissues to detect SNPs. Using broader genetic datasets also gathered by the consortium, the RNA-seq SNPs were used to impute polymorphisms along the whole sequences. Various expression and splicing QTLs were then detected, and compared between tissues, breeds and species, with the aim to both improve functional annotation of the animal genomes and help interpreting phenotypic GWAS results (Den Berg et al., 2020; Leal-Gutiérrez et al., 2020; Zhang et al., 2023).

Meta-analyses may become very powerful for complex traits analyses that require a large number of individuals, or for traits sensitive to population structure. Meta-analyses can be performed at different levels of the systems genetics approach, like by combining datasets of the same nature coming from multiple independent studies or by integrating several genomics or transcriptomics datasets. As an example, Duarte et al., 2019 performed a GWAS pathway-based meta-analysis in cattle for feed-efficiency traits. While independent GWAS for Residual Feed Intake (RFI) in beef cattle from 10 studies were inconsistent in highlighting common genomic regions associated with the phenotype, the meta-analysis using all datasets slightly improved the results by decreasing false-positive rates. Such an approach made it possible to identify gene sets involved in the same pathway to explain the studied phenotype. The valine, leucine and isoleucine degradation pathway was found to be significantly related to RFI showing that such pathway-based GWAS meta-analysis can be an appropriate method to uncover biological insights by combining useful information from different studies (Duarte et al., 2019). Another example of intra-species systems genetics approach is a recent study on merino sheep hair follicles by Zhao et al., 2021. While GWAS related to hair and skin were previously available, the authors generated multi-omics data on hair follicles (histology, gene expression - including lncRNA and circRNA - DNA methylation) across sheep developmental stages and highlighted intricate gene

regulatory networks governing the development of hair follicles. Evidence has shown that integrating such regulatory elements, as well as targeting highly conserved variants during evolution and intermediate phenotypes such as metabolites, helps better predict traits of interest (Xiang et al., 2019). Methods to perform phenotype integration will also allow the re-use of heterogenous phenotype records that are increasingly available (Dahl et al., 2023; Vasseur et al., 2022).

The future of systems genetics will come from cross-species analyses to benefit from the extensive knowledge gathered for model organisms. A few studies contributed to improve knowledge on molecular mechanisms and genes underlying pathological human phenotypes by combining datasets on human cohorts and on model organisms (Ashbrook et al., 2019; Calabrese et al., 2017; Sabik et al., 2020). Notably, Komljenovic et al., 2019 used a cross-species systems genetics approach to detect key genes and biological pathways involved in aging. The initial assumption for their study was a conservation of core genes involved in aging across evolution targeting species from different phylogenetic clades including *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*. A first step aimed at performing differential expression analyses in appropriate tissues in each species, with a challenge experiment as a validation dataset. To integrate independent results, a second step was based on orthology to highlight functional enrichment and identify core genes related to aging. After building co-expression networks of genes identified from this second step, a network-level cross-species integration strategy was applied. Finally, genes involved in shared modules were used as targets to search for genomic regions identified from various GWAS datasets available for phenotypes related to aging. The outcome of this study showed that *i)* evolutionarily conserved modules of aging existed across diverse taxa and *ii)* cross-species networks were enriched for genes that encompass genetic variants associated with aging-related human traits. In farm animals, the current effort to make larger sets of animal multi-omics data discoverable and reusable should improve the feasibility and emergence of cross-species phenotype-centered omics studies (IAnimal (Fu et al., 2023), PigBioBank (Zeng et al., 2024)). Multi-populations and multi-species analyses based on domesticated animal omics will contribute to the identification of causal biological mechanisms and variants for complex phenotypes.

The development of artificial intelligence methods might reinforce the cross-species systems genetics strategy by allowing the use of large public datasets in model organisms including humans, mice and rats to predict different molecular levels in farm animals when data are unavailable. This is particularly true for epigenomic datasets that are still lagging behind for farm animal species. Several methods and algorithms show that prediction of regulatory sequences between species is becoming possible (Kelley, 2020; Li et al., 2022; Minnoye et al., 2020; Mourad, 2023).

Conclusion

In this article we highlighted some success and discussed the limits of past approaches linking animal genomes to phenomes. We see in systems genetics the opportunity to overcome these limits, re-using and integrating available datasets on a large scale. However, we remain aware of the many challenges of this new approach, in terms of data management, metadata integration, and statistical developments. Lastly, we highlighted a few studies that paved the way for future animal systems genetics studies (Box 3). We are confident that this approach will contribute to closing the genome-to-phenome gap, allow deeper understanding of animal genome functions

and their evolution under selection, and ultimately, contribute to providing solutions for sustainable animal breeding. Indeed, systems genetics approaches might be one piece of the puzzle to answer the major challenges faced by agriculture.

Box 3 - Community efforts for systems and translational genomics in farm animal species

First of all it is essential to acknowledge the massive efforts that have already been made into the building of platforms to assess resources, the availability of pipelines for analysis, and the building of relevant ontology. However some key points remain crucial to build a stronger community that would benefit from the expansion of the field of systems and translational genomics.

Below we list some potential action points to further promote research in farm animal systems genetics.

Things to work on

- Make efforts to reuse existing ontology and databases, and aggregate them when possible
- Make data, metadata, methods and results FAIRer and FAIRer
- Work towards an uniformization of data types and analysis practices
- Combine both public and private data to significantly empower systems genetics approaches

How can we reach these goals

- By building up on already existing collaborations and consortia such as FAANG (https://www.faang.org/), or AG2PI, or community pipelines such as nf-core (Langer et al., 2025)
- Through the organization of dedicated sessions in congresses and symposia
- By promoting exchange of scientists between laboratories
- Through training (courses, workshops or hackathons) under expert guidance

Acknowledgments

We would like to thank the reviewers, Chris Tuggle and Karel Novak, for their valuable comments and suggestions during the PCI Animal Science recommendation process. Preprint version 3 of this article has been peer-reviewed and recommended by PCI Animal Science (https://doi.org/10.24072/pci.animsci.100353, Rafat, 2025).

Funding

This work benefited from the support of the Animal Genetics Division INRAE; of the French National Research Agency (Agence Nationale de la Recherche, ANR) under the projects PIPETTE (ANR-18-CE20-0018), FeedSeq (ANR-22-CE20-0040); of the European Union's Horizon 2020 research and innovation program under the grant agreement N°101000236 (GEroNIMO) as part of the EuroFAANG (https://eurofaang.eu) initiative; of the priority research programmes and equipments PEPR Agroecology and Numeric (22-PEAE-0015) and PEPR AgroDiv (22-PEAE-0005); of the meta-program INRAE DIGIT-BIO; of the Occitanie region.

Conflict of interest disclosure

The authors declare that they do not have financial conflicts of interest in relation to the content of the article.

Data, scripts, code, and supplementary information availability

There was no dataset generated or analysed in this study.

References

Aggrey S, Zhou H, Tixier-Boichard M, Rhoads D (2020). *Advances in Poultry Genetics and Genomics (1st ed.)* Burleigh Dodds Science Publishing. https://doi.org/10.1201/9781003047735.

Akiyama M (2021). Multi-omics study for interpretation of genome-wide association study. *Journal of Human Genetics* **66**, 3–10. https://doi.org/10.1038/s10038-020-00842-5.

Allayee H, Farber CR, Seldin MM, Williams EG, James DE, Lusis AJ (2023). Systems genetics approaches for understanding complex traits with relevance for human disease. *Elife* **12**, e91004. https://doi.org/10.7554/eLife.91004.

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology* **14**, e8124. https://doi.org/10.15252/msb.20178124.

Arita M, Karsch-Mizrachi I, Cochrane obotINSDC (Guy) (2020). The international nucleotide sequence database collaboration. *Nucleic Acids Research* **49**, D121–D124. https://doi.org/10.1093/nar/gkaa967.

Ashbrook DG, Cahill S, Hager R (2019). A cross-species systems genetics analysis links APBB1IP as a candidate for schizophrenia and prepulse inhibition. *Frontiers in behavioral neuroscience* **13**, 266. https://doi.org/10.3389/fnbeh.2019.00266.

Baud A, McPeek S, Chen N, Hughes KA (2022). Indirect genetic effects: A cross-disciplinary perspective on empirical studies. *Journal of Heredity* **113**, 1–15. https://doi.org/10.1093/jhered/esab059.

Bellinge R, Liberles DA, Iaschi S, O'brien P, Tay G (2005). Myostatin and its implications on animal breeding: a review. *Animal genetics* **36**, 1–6. https://doi.org/10.1111/j.1365-2052.2004.01229.x.

Ben Braiek M, Moreno-Romieux C, André C, Astruc JM, Bardou P, Bordes A, Debat F, Fidelle F, Granado-Tajada I, Hozé C (2024a). Searching for homozygous haplotype deficiency in Manech Tête Rousse dairy sheep revealed a nonsense variant in the MMUT gene affecting newborn lamb viability. *Genetics Selection Evolution* **56**, 16. https://doi.org/10.1186/s12711-024-00886-7.

Ben Braiek M, Szymczak S, André C, Bardou P, Fidelle F, Granado-Tajada I, Plisson-Petit F, Sarry J, Woloszyn F, Moreno-Romieux C (2024b). A single base pair duplication in the SLC33A1 gene is associated with fetal losses and neonatal lethality in Manech Tête Rousse dairy sheep. *Animal Genetics* **55**, 644–657. https://doi.org/10.1111/age.13459.

Ben Guebila M, Weighill D, Lopes-Ramos CM, Burkholz R, Pop RT, Palepu K, Shapoval M, Fagny M, Schlauch D, Glass K, et al. (2022). An online notebook resource for reproducible inference, analysis and publication of gene regulatory networks. *Nature methods* **19**, 511–513. https://doi.org/10.1038/s41592-022-01479-2.

Benegas G, Batra SS, Song YS (2023). DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences* **120**, e2311219120. https://doi.org/10.1073/pnas.2311219120.

Bidanel JP, Silalahi P, Tribout T, Canario LL, Ducos A, Garreau H, Gilbert H, Larzul C, Milan D, Riquet J (2020). Cinquante années d'amélioration génétique du porc en France: bilan et perspectives. *INRAE Production Animales* **33**. https://doi.org/10.20870/productions-animales.2020.33.1.3092.

Boichard D, Ducrocq V, Fritz S (2015). Sustainable dairy cattle selection in the genomic era. *Journal of Animal Breeding and Genetics* **132**, 135–143. https://doi.org/10.1111/jbg.12150.

Bujold D, de Lima Morais DA, Gauthier C, Côté C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T (2016). The international human epigenome consortium data portal. *Cell systems* **3**, 496–499. https://doi.org/10.1016/j.cels.2016.10.019.

Burnett KG, Durica DS, Mykles DL, Stillman JH (2020). Building Bridges from Genome to Phenome: Molecules, Methods and Models—An Introduction to the Symposium. *Integrative and Comparative Biology* **60**, 261–266. https://doi.org/10.1093/icb/icaa073.

Burnett TA, Madureira AM, Silper BF, Tahmasbi A, Nadalin A, Veira DM, Cerri RL (2015). Relationship of concentrations of cortisol in hair with health, biomarkers in blood, and reproductive status in dairy cows. *Journal of Dairy Science* **98**, 4414–4426. https://doi.org/10.3168/jds.2014-8871.

Calabrese GM, Mesner LD, Stains JP, Tommasini SM, Horowitz MC, Rosen CJ, Farber CR (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell systems* **4**, 46–59. https://doi.org/10.1016/j.cels.2016.10.014.

Chen L, Fish AE, Capra JA (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS computational biology* **14**, e1006484. https://doi.org/10.1371/journal.pcbi.1006484.

Civelek M, Lusis AJ (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* **15**, 34–48. https://doi.org/10.1038/nrg3575.

Clark EL, Archibald AL, Daetwyler HD, Groenen MA, Harrison PW, Houston RD, Kühn C, Lien S, Macqueen DJ, Reecy JM (2020). From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biology* **21**, 1–9. https://doi.org/10.1186/s13059-020-02197-8.

Cole JB, Baes CF, Eaglen SA, Lawlor TJ, Maltecca C, Ortega MS, VanRaden PM (2025). Invited review: Management of genetic defects in dairy cattle populations. *Journal of Dairy Science* **108**, 3045–3067. https://doi.org/10.3168/jds.2024-26035.

Dahl A, Thompson M, An U, Krebs M, Appadurai V, Border R, Bacanu SA, Werge T, Flint J, Schork AJ (2023). Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nature Genetics* **55**, 2082–2093. https://doi.org/10.1038/s41588-023-01559-9.

Demars J, Labrune Y, Iannuccelli N, Deshayes A, Leroux S, Gilbert H, Aymard P, Benitez F, Riquet J (2022). A genome-wide epistatic network underlies the molecular architecture of continuous color variation of body extremities. *Genomics* **114**, 110361. https://doi.org/10.1016/j.ygeno.2022.110361.

van Den Berg I, Xiang R, Jenko J, Pausch H, Boussaha M, Schrooten C, Tribout T, Gjuvsland AB, Boichard D, Nordbø Ø (2020). Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. *Genetics Selection Evolution* **52**, 1–16. https://doi.org/10.1186/s12711-020-00556-4.

Destoumieux-Garzón D, Bonnet P, Teplitsky C, Criscuolo F, Henry PY, Mazurais D, Prunet P, Salvat G, Usseglio-Polatera P, Verrier E (2021). Animal board invited review: OneARK: Strengthening the links between animal production science and animal ecology. *Animal* **15**, 100053. https://doi.org/10.1016/j.animal.2020.100053.

Doublet AC, Croiseau P, Fritz S, Michenet A, Hozé C, Danchin-Burge C, Laloë D, Restoux G (2019). The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. *Genetics Selection Evolution* **51**, 52. https://doi.org/10.1186/s12711-019-0495-1.

Duarte D, Newbold CJ, Detmann E, Silva F, Freitas P, Veroneze R, Duarte MdS (2019). Genome-wide association studies pathway-based meta-analysis for residual feed intake in beef cattle. *Animal genetics* **50**, 150–153. https://doi.org/10.1111/age.12761.

Durand G, Blanchard G, Neuvial P, Roquain E (2020). Post hoc false positive control for structured hypotheses. *Scandinavian journal of Statistics* **47**, 1114–1148. https://doi.org/10.1111/sjos.12453.

Edwards SL, Beesley J, French JD, Dunning AM (2013). Beyond GWASs: illuminating the dark road from association to function. *The American Journal of Human Genetics* **93**, 779–797. https://doi.org/10.1016/j.ajhg.2013.10.012.

Enjalbert-Courrech N, Neuvial P (2022). Powerful and interpretable control of false discoveries in two-group differential expression studies. *Bioinformatics* **38**, 5214–5221. https://doi.org/10.1093/bioinformatics/btac693.

Eynard SE, Windig JJ, Hiemstra SJ, Calus MP (2016). Whole-genome sequence data uncover loss of genetic diversity due to selection. *Genetics Selection Evolution* **48**, 33. https://doi.org/10.1186/s12711-016-0210-4.

Fang L, Teng J, Lin Q, Bai Z, Liu S, Guan D, Li B, Gao Y, Hou Y, Gong M (2025). The Farm Animal Genotype–Tissue Expression (FarmGTEx) Project. *Nature genetics*, 1–11. https://doi.org/10.1038/s41588-025-02121-5.

FAO (2022). *Food Outlook – Biannual Report on Global Food Markets*. https://doi.org/10.4060/cb9427en. URL: https://app.dimensions.ai/details/publication/pub.1148571664 (visited on 06/16/2025).

Fisher SA, Lewis CM (2008). Power of genetic association studies in the presence of linkage disequilibrium and allelic heterogeneity. *Human Heredity* **66**, 210–222. https://doi.org/10.1159/000143404.

Freking BA, Murphy SK, Wylie AA, Rhodes SJ, Keele JW, Leymaster KA, Jirtle RL, Smith TP (2002). Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome research* **12**, 1496–1506. https://doi.org/10.1101/gr.571002.

Fu Y, Liu H, Dou J, Wang Y, Liao Y, Huang X, Tang Z, Xu J, Yin D, Zhu S (2023). IAnimal: a cross-species omics knowledgebase for animals. *Nucleic acids research* **51**, D1312–D1324. https://doi.org/10.1093/nar/gkac936.

Gallitto G, Englert R, Kincses B, Kotikalapudi R, Li J, Hoffschlag K, Bingel U, Spisak T (2023). External validation of machine learning models-registered models and adaptive sample splitting. *bioRxiv*, 1–12. https://doi.org/10.1101/2023.12.01.569626.

Gao LL, Bien J, Witten D (2024). Selective inference for hierarchical clustering. *Journal of the American Statistical Association* **119**, 332–342. https://doi.org/10.1080/01621459.2022.2116331.

Gong Y, Li Y, Liu X, Ma Y, Jiang L (2023). A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *Journal of Animal Science and Biotechnology* **14**, 73. https://doi.org/10.1186/s40104-023-00860-1.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T (2022). FlyBase: a guided tour of highlighted features. *Genetics* **220**, iyac035. https://doi.org/10.1093/genetics/iyac035.

Guan D, Bai Z, Zhu X, Zhong C, Hou Y, Zhu D, Li H, Lan F, Diao S, Yao Y (2025). Genetic regulation of gene expression across multiple tissues in chickens. *Nature Genetics* **57**, 1298–1308. https://doi.org/10.1038/s41588-025-02155-9.

Han B, Kang HM, Eskin E (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS genetics* **5**, e1000456. https://doi.org/10.1371/journal.pgen.1000456.

Harrison PW, Sokolov A, Nayak A, Fan J, Zerbino D, Cochrane G, Flicek P (2021). The FAANG data portal: Global, open-access,"FAIR", and richly validated genotype to phenotype data for high-quality functional annotation of animal genomes. *Frontiers in Genetics* **12**, 639238. https://doi.org/10.3389/fgene.2021.639238.

Hasin Y, Seldin M, Lusis A (2017). Multi-omics approaches to disease. *Genome biology* **18**, 1–15. https://doi.org/10.1186/s13059-017-1215-1.

Hayes BJ, Daetwyler HD (2019). 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences* **7**, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024.

Hendriks SJ, Schmitt O, Boyle L (2025). Rethinking sustainability: recognizing animal welfare's critical role. *Animal Frontiers* **15**, 3–7. https://doi.org/10.1093/af/vfaf005.

Ioannidis JP (2005). Why most published research findings are false. *PLoS medicine* **2**, e124. https://doi.org/10.1371/journal.pmed.0020124.

Jagota M, Ye C, Albors C, Rastogi R, Koehl A, Ioannidis N, Song YS (2023). Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biology* **24**, 182. https://doi.org/10.1186/s13059-023-03024-6.

Jin L, Tang Q, Hu S, Chen Z, Zhou X, Zeng B, Wang Y, He M, Li Y, Gui L (2021). A pig BodyMap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription. *Nature communications* **12**, 3715. https://doi.org/10.1038/s41467-021-23560-8.

Jourdain J, Barasc H, Faraut T, Calgaro A, Bonnet N, Marcuzzo C, Suin A, Barbat A, Hozé C, Besnard F (2023). Large-scale detection and characterization of interchromosomal rearrangements in normozoospermic bulls using massive genotype and phenotype data sets. *Genome Research* **33**, 957–971. https://doi.org/10.1101/gr.277787.123.

Jurrjens AW, Seldin MM, Giles C, Meikle PJ, Drew BG, Calkin AC (2023). The potential of integrating human and mouse discovery platforms to advance our understanding of cardiometabolic diseases. *Elife* **12**, e86139. https://doi.org/10.7554/eLife.86139.

Kelley DR (2020). Cross-species regulatory sequence activity prediction. *PLoS computational biology* **16**, e1008050. https://doi.org/10.1371/journal.pcbi.1008050.

Komljenovic A, Li H, Sorrentino V, Kutalik Z, Auwerx J, Robinson-Rechavi M (2019). Cross-species functional modules link proteostasis to human normal aging. *PLoS computational biology* **15**, e1007162. https://doi.org/10.1371/journal.pcbi.1007162.

Komsta L, Novomestky F (2022). *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.14.1. URL: https://CRAN.R-project.org/package=moments.

Kreitmaier P, Katsoula G, Zeggini E (2023). Insights from multi-omics integration in complex disease primary tissues. *Trends in Genetics* **39**, 46–58. https://doi.org/10.1016/j.tig.2022.08.005.

Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K (2019). Estimating sample-specific regulatory networks. *Iscience* **14**, 226–240. https://doi.org/10.1016/j.isci.2019.03.021.

Kunert-Graf JM, Sakhanenko NA, Galas DJ (2021). Optimized permutation testing for information theoretic measures of multi-gene interactions. *BMC bioinformatics* **22**, 1–11. https://doi.org/10.1186/s12859-021-04107-6.

Langer BE, Amaral A, Baudement MO, Bonath F, Charles M, Chitneedi PK, Clark EL, Di Tommaso P, Djebali S, Ewels PA, et al. (2025). Empowering bioinformatics communities with Nextflow and nf-core. *Genome Biology* **26**, 228. https://doi.org/10.1186/s13059-025-03673-9.

Langfelder P, Horvath S (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1–13. https://doi.org/10.1186/1471-2105-9-559.

Leal-Gutiérrez JD, Rezende FM, Reecy JM, Kramer LM, Peñagaricano F, Mateescu RG (2020). Whole genome sequence data provides novel insights into the genetic architecture of meat quality traits in beef. *Frontiers in Genetics* **11**, 538640. https://doi.org/10.3389/fgene.2020.538640.

Li J, Wang J, Zhang P, Wang R, Mei Y, Sun Z, Fei L, Jiang M, Ma L, E W (2022). Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nature Genetics* **54**, 1711–1720. https://doi.org/10.1038/s41588-022-01197-7.

Li R, Yang P, Dai X, Asadollahpour Nanaei H, Fang W, Yang Z, Cai Y, Zheng Z, Wang X, Jiang Y (2021). A near complete genome for goat genetic and genomic research. *Genetics Selection Evolution* **53**, 1–17. https://doi.org/10.1186/s12711-021-00668-5.

Li Z, Zhang Y, Peng B, Qin S, Zhang Q, Chen Y, Chen C, Bao Y, Zhu Y, Hong Y (2024). A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. *Nucleic Acids Research* **52**, 13447–13468. https://doi.org/10.1093/nar/gkae912.

Lim KS, Cheng J, Tuggle C, Dyck M, Canada P, Fortin F, Harding J, Plastow G, Dekkers J (2023). Genetic analysis of the blood transcriptome of young healthy pigs to improve disease resilience. *Genetics Selection Evolution* **55**, 90. https://doi.org/10.1186/s12711-023-00860-9.

Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, Li B, Xiang R, Chamberlain AJ, Pairo-Castineira E (2022). A multi-tissue atlas of regulatory variants in cattle. *Nature Genetics* **54**, 1438–1447. https://doi.org/10.1038/s41588-022-01153-5.

Lloret-Villas A, Pausch H, Leonard AS (2023). The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. *Genetics Selection Evolution* **55**, 33. https://doi.org/10.1186/s12711-023-00809-y.

Longo DL, Drazen JM (2016). Data sharing. *New England Journal of Medicine* **374**, 276–277. https://doi.org/10.1056/NEJMe1516564.

Loy JD, Klabnik JL, O'Boyle NJ (2024). Genomics for the Modern Beef and Dairy Practitioner. *Veterinary Clinics: Food Animal Practice* **40**, ix–x. https://doi.org/10.1016/j.cvfa.2024.06.001.

Lozada DN, Bosland PW, Barchenger DW, Haghshenas-Jaryani M, Sanogo S, Walker S (2022). Chile pepper (Capsicum) breeding and improvement in the "multi-omics" era. *Frontiers in plant science* **13**, 879182. https://doi.org/10.3389/fpls.2022.879182.

Lu C (2017). Dairy, science, society, and the environment. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.316.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195. https://doi.org/10.1126/science.1222794.

Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V, Makhzami S, Seltenhammer M, Karras P (2020). Cross-species analysis of enhancer logic using deep learning. *Genome research* **30**, 1815–1834. https://doi.org/10.1101/gr.260844.120.

Misztal I, Lourenco D (2024). Potential negative effects of genomic selection. *Journal of animal science* **102**, skae155. https://doi.org/10.1093/jas/skae155.

Morris KM, Hindle MM, Boitard S, Burt DW, Danner AF, Eory L, Forrest HL, Gourichon D, Gros J, Hillier LW (2020). The quail genome: insights into social behaviour, seasonal biology and infectious disease response. *BMC biology* **18**, 1–18. https://doi.org/10.1186/s12915-020-0743-4.

Mourad R (2023). Semi-supervised learning improves regulatory sequence prediction with unlabeled sequences. *BMC bioinformatics* **24**, 186. https://doi.org/10.1186/s12859-023-05303-2.

Neufeld A, Dharamshi A, Gao LL, Witten D (2024). Data thinning for convolution-closed distributions. *Journal of Machine Learning Research* **25**, 1–35. URL: http://jmlr.org/papers/v25/23-0446.html.

Pan Z, Yao Y, Yin H, Cai Z, Wang Y, Bai L, Kern C, Halstead M, Chanthavixay G, Trakooljul N (2021). Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nature Communications* **12**, 5848. https://doi.org/10.1038/s41467-021-26153-7.

Peirce JL, Lu L, Gu J, Silver LM, Williams RW (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC genetics* **5**, 1–17. https://doi.org/10.1186/1471-2156-5-7.

Price E, Feyertag F, Evans T, Miskin J, Mitrophanous K, Dikicioglu D (2024). What is the real value of omics data? Enhancing research outcomes and securing long-term data excellence. *Nucleic Acids Research* **52**, 12130–12140. https://doi.org/10.1093/nar/gkae901.

Rafat SA (2025). A novel approach for integration of omics databases from various analyses across domestic species. *Peer Community in Animal Science*, 100353. https://doi.org/10.24072/pci.animsci.100353.

Rice ES, Alberdi A, Alfieri J, Athrey G, Balacco JR, Bardou P, Blackmon H, Charles M, Cheng HH, Fedrigo O (2023). A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC biology* **21**, 267. https://doi.org/10.1186/s12915-023-01758-0.

Ritchie H (2020). Less meat is nearly always better than sustainable meat, to reduce your carbon footprint. *Our world in data*. URL: https://ourworldindata.org/less-meat-or-sustainable-meat.

Roesch E, Greener JG, MacLean AL, Nassar H, Rackauckas C, Holy TE, Stumpf MP (2023). Julia for biologists. *Nature methods* **20**, 655–664. https://doi.org/10.1038/s41592-023-01832-z.

Rohart F, Gautier B, Singh A, Lê Cao KA (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology* **13**, e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Rosen GD, Chesler EJ, Manly KF, Williams RW (2007). An informatics approach to systems neurogenetics. *Neuroinformatics* **1**, 287–303. https://doi.org/10.1007/978-1-59745-520-6_16.

Sabik OL, Calabrese GM, Taleghani E, Ackert-Bicknell CL, Farber CR (2020). Identification of a core module for bone mineral density through the integration of a co-expression network and GWAS data. *Cell reports* **32**. https://doi.org/10.1016/j.celrep.2020.108145.

Secomandi S, Gallo GR, Rossi R, Rodríguez Fernandes C, Jarvis ED, Bonisoli-Alquati A, Gianfranceschi L, Formenti G (2025). Pangenome graphs and their applications in biodiversity genomics. *Nature Genetics*, 1–14. https://doi.org/10.1038/s41588-024-02029-6.

Sloan Z, Arends D, Broman KW, Centeno A, Furlotte N, Nijveen H, Yan L, Zhou X, Williams RW, Prins P (2016). GeneNetwork: framework for web-based genetics. *Journal of Open Source Software* **1**, 25. https://doi.org/10.21105/joss.00025.

Smith TP, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, Low WY, Pausch H, Demyda-Peyrás S, Prendergast J (2023). The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome biology* **24**, 139. https://doi.org/10.1186/s13059-023-02975-0.

Song D, Chen S, Lee C, Li K, Ge X, Li JJ (2025). Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses. In: *International Conference on Research in Computational Molecular Biology*. Cold Spring Harbor Laboratory, pp. 400–404. https://doi.org/10.1101/2023.07.21.550107.

Song Y, Miao Z, Brazma A, Papatheodorou I (2023). Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications* **14**, 6495. https://doi.org/10.1038/s41467-023-41855-w.

Stephens M (2017). False discovery rates: a new deal. *Biostatistics* **18**, 275–294. https://doi.org/10.1093/biostatistics/kxw041.

Syed H, Otto GW, Kelberman D, Bacchelli C, Beales PL (2021). MOPower: an R-shiny application for the simulation and power calculation of multi-omics studies. *bioRxiv*, 2021–12. https://doi.org/10.1101/2021.12.19.473339.

Talenti A, Powell J, Hemmink JD, Cook EA, Wragg D, Jayaraman S, Paxton E, Ezeasor C, Obishakin E, Agusi E (2022). A cattle graph genome incorporating global breed diversity. *Nature communications* **13**, 910. https://doi.org/10.1038/s41467-022-28605-0.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484. https://doi.org/10.1038/s41576-019-0127-1.

Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, PigGTEx Consortium, Bai L, Cai Z, Zhao B (2024). A compendium of genetic regulatory effects across pig tissues. *Nature genetics* **56**, 112–123. https://doi.org/10.1038/s41588-023-01585-7.

Threadgill DW (2006). Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics. *Mammalian Genome* **17**, 2–4. https://doi.org/10.1007/s00335-005-0153-5.

Tuggle CK, Clarke J, Dekkers JC, Ertl D, Lawrence-Dill CJ, Lyons E, Murdoch BM, Scott NM, Schnable PS (2022). The Agricultural Genome to Phenome Initiative (AG2PI): creating a shared vision across crop and livestock research communities. *Genome biology* **23**, 1–11. https://doi.org/10.1186/s13059-021-02570-1.

Tuly SR, Ranjbari S, Murat EA, Arslanturk S (2025). From Silos to Synthesis: A comprehensive review of domain adaptation strategies for multi-source data integration in healthcare. *Computers in Biology and Medicine* **191**, 110108. https://doi.org/10.1016/j.compbiomed.2025.110108.

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D (2021). Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59. https://doi.org/10.1038/s43586-021-00056-9.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836. https://doi.org/10.1038/nature02064.

Van Zanten HH, Herrero M, Van Hal O, Röös E, Muller A, Garnett T, Gerber PJ, Schader C, De Boer IJ (2018). Defining a land boundary for sustainable livestock consumption. *Global change biology* **24**, 4185–4194. https://doi.org/10.1111/gcb.14321.

VanRaden P, Olson K, Null D, Hutchison J (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *Journal of dairy science* **94**, 6153–6161. https://doi.org/10.3168/jds.2011-4624.

Vasseur F, Westgeest AJ, Vile D, Violle C (2022). Solving the grand challenge of phenotypic integration: allometry across scales. *Genetica* **150**, 161–169. https://doi.org/10.1007/s10709-022-00158-6.

Von Keyserlingk M, Martin N, Kebreab E, Knowlton K, Grant R, Stephenson M, Sniffen C, Harner Iii J, Wright A, Smith S (2013). Invited review: Sustainability of the US dairy industry. *Journal of dairy science* **96**, 5405–5425. https://doi.org/10.3168/jds.2012-6354.

Wang J, Williams RW, Manly KF (2003). WebQTL: web-based complex trait analysis. *Neuroinformatics* **1**, 299–308. https://doi.org/10.1385/NI:1:4:299.

Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, Chow W, Eory L, Finlayson HA, Flicek P (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* **9**, giaa051. https://doi.org/10.1093/gigascience/giaa051.

Weeks EM, Ulirsch JC, Cheng NY, Trippe BL, Fine RS, Miao J, Patwardhan TA, Kanai M, Nasser J, Fulco CP (2023). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nature Genetics* **55**, 1267–1276. https://doi.org/10.1038/s41588-023-01443-6.

Weighill D, Guebila MB, Glass K, Quackenbush J, Platig J (2022). Predicting genotype-specific gene regulatory networks. *Genome research* **32**, 524–533. https://doi.org/10.1101/gr.275107.120.

Weller JI (2016). *Genomic selection in animals*. Wiley Online Library. https://doi.org/10.1002/9781119213628.

Wharrie S, Yang Z, Raj V, Monti R, Gupta R, Wang Y, Martin A, O'Connor LJ, Kaski S, Marttinen P (2023). HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. *Bioinformatics* **39**, btad535. https://doi.org/10.1093/bioinformatics/btad535.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3**, 1–9. https://doi.org/10.1038/sdata.2016.18.

Williams EG, Auwerx J (2015). The convergence of systems and reductionist approaches in complex trait analysis. *Cell* **162**, 23–32. https://doi.org/10.1016/j.cell.2015.06.024.

Wragg D, Zhang W, Peterson S, Yerramilli M, Mellanby R, Schoenebeck JJ, Clements DN (2024). A cautionary tale of low-pass sequencing and imputation with respect to haplotype accuracy. *Genetics Selection Evolution* **56**, 6. https://doi.org/10.1186/s12711-024-00875-w.

Xiang R, Berg Ivd, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, Bolormaa S, Liu Z, Rochfort SJ, Reich CM (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* **116**, 19398–19408. https://doi.org/10.1073/pnas.1904159116.

Zeng H, Zhang W, Lin Q, Gao Y, Teng J, Xu Z, Cai X, Zhong Z, Wu J, Liu Y (2024). PigBiobank: a valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs. *Nucleic acids research* **52**, D980–D989. https://doi.org/10.1093/nar/gkad1080.

Zhang X, Wang H, Yang M, Liu R, Zhang X, Jia Z, Li P (2023). Natural variation in ZmNAC087 contributes to total root length regulation in maize seedlings under salt stress. *BMC Plant Biology* **23**, 392. https://doi.org/10.1186/s12870-023-04393-7.

Zhao B, Luo H, He J, Huang X, Chen S, Fu X, Zeng W, Tian Y, Liu S, Li Cj (2021). Comprehensive transcriptome and methylome analysis delineates the biological basis of hair follicle development and wool-related traits in Merino sheep. *BMC biology* **19**, 1–18. https://doi.org/10.1186/s12915-021-01127-9.

Zhu M, Zhao S (2007). Candidate gene identification approach: progress and challenges. *International journal of biological sciences* **3**, 420. https://doi.org/10.7150/ijbs.3.420.

Zuidhof M, Schneider B, Carney V, Korver D, Robinson F (2014). Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 2005. *Poultry science* **93**, 2970–2982. https://doi.org/10.3382/ps.2014-04291.