A pioneering multi-omics data platform sheds light on the understanding of biological systems

The GLOMICAVE project, which ended in April 2024, has developed an innovative multi-omics data analysis platform that maximises the use of omics data at a massive level and supports large-scale omic experiments in animal, agrifood and environmental sectors.

Multi-omics research: an opportunity for our society

Research into omic disciplines can provide important breakthroughs in the knowledge and understanding of biological aspects and issues, representing a huge advancement in the analysis of cell functionality and biotechnological applications.

In recent years, massive increases in analytical throughput, together with reductions in costs and an increased storage capacity, have enabled multi-omics studies to be performed at a scale not previously imagined. These modern techniques allow the creation of large-scale biological datasets, including genomics, proteomics, metabolomics and epigenetics.

Despite all these advancements, **there has been less progress in linking genome information (genotype) to the complex variation in living organisms' observable traits (phenotype)**. There is an urgent need to understand this connection, which is key for addressing important societal challenges and needs across various sectors. It is particularly relevant in the case of complex organisms and environments, where systems biology benefits from omics data to decipher complex phenotypes.

An enhanced understanding of phenotypes will enable the development of new predictive models for living organisms, applicable in a variety of industrial sectors. The integration and exploitation of existing omics datasets and biological data types are needed to extract meaningful information about genotype-phenotype relationships and associations that will complete the picture of how biological models function and how phenotypes are established.

In this scenario, the GLOMICAVE project, coordinated by Eurecat Technology Centre, has developed a **multi-omics data platform combining and analysing different types of omics data** to obtain relevant information and discover new links between animal and vegetable genotype and phenotype.

The main innovation behind GLOMICAVE relies on demonstrating the integration of different omics databases with specific information available in other supports. This allows a better understanding of the many sources of genotype-phenotype associations and, as a result, a more accurate phenotypic prediction.

A look at the GLOMICAVE platform

Although each omics platform allows a particular comprehensive molecular survey for a given phenotype, a reductionist approach that analyses each omics layer in isolation cannot properly and fully assess the crosstalk between multiple molecular layers. Some important limitations and gaps are still preventing better data integration and genotype-phenotype validation:

- heterogeneity of data generated by the different omics platforms
- high rate of metabolites that remain unidentified in untargeted metabolomics analysis, preventing a large-scale integration
- poor interpretability of the results by different computational genotype-phenotype models
- the lack of knowledge integration in scientific literature and databases in an automated way
- non-biomedical sectors lack holistic omics science-based integration tools.

To overcome these problems and maximise the utility of pre-existing massive omics datasets, the GLOMICAVE project has developed an **innovative digital platform to process large-scale omics datasets using big data and artificial intelligence (AI)**, enhanced with automatic processing of scientific literature **by leveraging pre-existing data and improve the understanding of biological systems as a whole.**

Bioinformatics tools, like the GLOMICAVE platform, offer a great opportunity to maximise benefits for the industry and research community. These solutions can integrate and analyse data from various

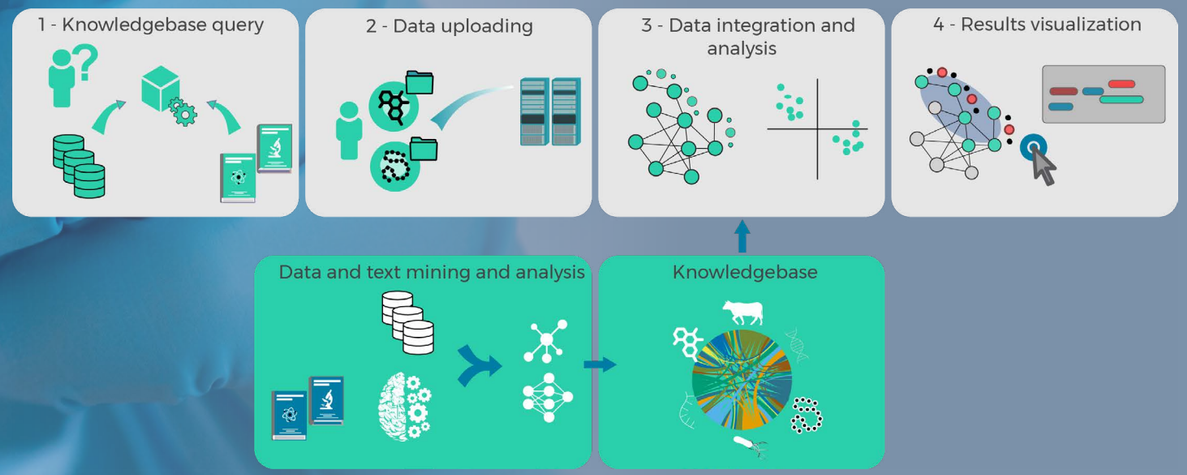
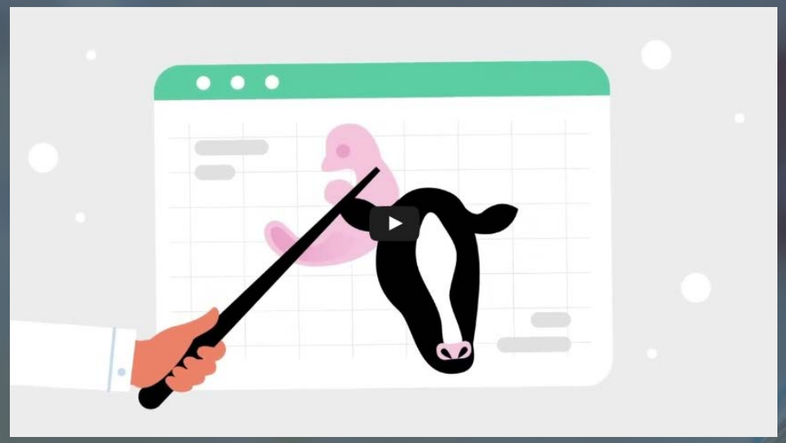


Figure 1: GLOMICAVE Platform structure.



GLOMICAVE project - Global omic data integration on animal, vegetal and environment sectors (youtube.com)

omics datasets and platforms to find data features linked to changes in phenotype.

The platform presents a novel and unique integration of structured data from omics datasets, including information without a pre-defined organised model, stored in text documents or similar sources in a variety of formats.

GLOMICAVE project has addressed the need to build systems that allow scaling analytical processes of primary data and supporting downstream large-scale omics experiments by maximising the utility of pre-existing massive omics datasets, going beyond existing data-driven tools.

A cloud-based genotype to phenotype platform based on big data and AI

Over the course of three and a half years, GLOMICAVE has developed its innovative platform, consisting of different analysis modules, as listed below. The solution aims to **assist experts and non-experts in identifying and understanding new links between genotype and phenotype, leading to their full comprehension and recommending the best course of action**.

Holistic integration of data

Technological and scientific findings are disseminated in text (scientific literature, patents, books, etc.) at an exponentially growing rate. Even experts in a specific field find it difficult to keep track of new findings because of this unrestrained growth of information.

A rapidly advancing technological area is natural language processing (NLP) algorithms, which are designed to computationally interpret what is known as unstructured data (data created with human language and understood only by humans).

In this sense, GLOMICAVE integrates **a module that has analysed, extracted and automatically interpreted natural language in thousands of publicly available scientific articles. The module now allows users to perform targeted queries through the platform to retrieve the expected and specific results regarding their questions**.

Structured data analysis

Most of the phenotype-genotype associations emerge from multifactorial mechanisms that require measuring the organisms' transcripts, proteins, metabolites and microbiota. **To provide a full comprehension of the genotype-phenotype relationships, we need to integrate as many omics data as possible**. However, the vast number of variables that modulate a phenotype make it highly difficult to find accurate associations.

In that sense, AI and big data analytics allow us to find hidden or unknown genotype-phenotype relationships that, by manual inspection, would be difficult to discover.

Innovative "autoAI" module

Despite the power and utility of AI, its application in multi-omics data analysis can be complex and difficult for research professionals without a bioinformatics background or without access to it.

To address this, the GLOMICAVE project has developed an **innovative "autoAI" module** where, through the web interface, users can apply AI techniques to their experimental data without programming knowledge. This approach democratises the application of AI by a group of users with a very large application potential but who are limited due to their lack of AI technical experience.

Extraction and harmonisation of multi-omics data

At the same time, AI has been applied to analyse data from public omics repositories, which store a vast amount of omics information and hidden patterns. The GLOMICAVE platform has been conceived to **extract explicit information and leverage the large amount of data in public texts and repositories**.

With advanced NLP, AI and BDA techniques, the project's platforms allow **the extraction of hidden information that would otherwise remain unnoticed**. With this, users can query, explore and fund relationships through the web interface.

Platform testing in livestock, agro-biology and water-environment sectors

In the final step of GLOMICAVE, we demonstrated our platform in the livestock, agro-biotechnology and environment sectors, addressing specific challenges in different business cases.

1. Dairy cattle fertility and meat quality

Current understanding of biological pathways in cattle reproduction, while growing, remains limited. Nowadays, there are no tools or biomarkers for an objective evaluation and selection of embryo transfer recipients which causes economic losses, as well as significant social and environmental impacts.

In the cattle fertility use case, the GLOMICAVE platform allowed researchers **to identify and validate efficient biomarkers to select competent recipients in dairy cattle and optimise reproduction technologies' efficiency**. This translated into minimising the chance of discarding fertile females, increasing the potential of pregnancy expectations, and promoting a reduction of the environmental footprint caused by current embryo transfer technologies.

Regarding meat quality, one of the most important challenges lies in understanding how the range of complex traits affects gene expression. Understanding this complex regulatory system is pivotal to translating results from genetic association studies.

The GLOMICAVE platform facilitated a better analysis and comprehension of the relationship between relevant phenotypes for the beef **industry, improving the understanding of the biological system of meat production. The outcomes can directly translate into increased profits for the beef sector, with further implications for animal welfare and the exploitation of sustainable food sources.**

2. Fruit quality and plant growth

Nearly two decades of post-genomics approaches have produced extensive multiple plant omics data, which has been stored in repositories and databases accessible for a wide range of research applications. However, multi-omics data integration is still limited and remains a significant challenge.

In terms of fruit biology, the GLOMICAVE consortium worked to better comprehend the factors influencing fruit growth and quality and manipulate them to improve **fruit traits**.

Our integrative platform helped in identifying metabolites and *in fine* metabolic pathways that are relevant to fruit quality or fruit affected by a specific pathogen. The dataset obtained allows us to compare healthy trees with trees infected by *Plum Pox Virus*, the causing agent of *Sharka* disease, a devastating viral disease of stone fruits with a relevant impact on agronomy and the economy.

Additionally, the GLOMICAVE platform facilitated the **integration of data on the root storage crop cassava** across the phenotypic, genomics and transcriptomics domains. Even though cassava crops are widely known worldwide, their genetic diversity is limited and narrow, making them vulnerable to different diseases.

With the GLOMICAVE platform, researchers could **identify biomarkers for yield and potentially discover new mechanistic insights into storage root formation. The data obtained could be used to improve yield and to finally generate designer plants**.

3. Water-environment scenarios

Lastly, the GLOMICAVE innovation has been validated in wastewater environment scenarios, where a deeper understanding of microbial communities and their role in wastewater treatment is crucial.

The global wastewater treatment industry faces increasing challenges due to population growth, urbanisation and industrialisation. Anaerobic digestion involves the decomposition of organic material by microorganisms in an oxygen-free environment. The process occurs in digesters, which are designed to maintain the optimal conditions for microbes to thrive and break down the feedstock efficiently.

In this sector, researchers can **develop predictive tools that correlate specific microbial biomarkers with system performance, enabling precise adjustments to improve its efficiency**.

More specifically, the research conducted in the project increased the knowledge available in **the application of omics technologies for bioenergy production in anaerobic digesters**. Moreover, the research results facilitated deeply analysing and understanding the function and interaction of microbial phosphorus removal and recovery.

GLOMICAVE

PROJECT SUMMARY

The GLOMICAVE project has developed an innovative multi-omics data analysis digital platform, relying on big data analytics and artificial intelligence and using large-scale publicly available and experimental omic datasets. The project aimed to maximise the utility of omic data at a massive level and discover new links between animal and vegetable genotype and phenotype, understanding biological systems as a whole.

PROJECT PARTNERS

A consortium of 15 partners in six European countries: five technology and research centres (Eurecat Technology Centre, project's coordinator, SERIDA, INRAE, ASINCAR and Forschungszentrum Julich); three universities (Aalborg University, the University of Minho and KU Leuven); three SMEs (TREE Technology, Eliance and AkiNaO); two large corporations (NEC Laboratories Europe and Águas do Norte); the ASEAVA animal cluster, and UNE as a standardisation body.

PROJECT LEAD PROFILE

Dr Núria Canela. Doctorate in Pharmacy from the University of Barcelona (2002) and Director of the Omic Sciences Unit at Eurecat Technology Centre. Her work focuses on applying omics methodologies to the study of biological systems.

Dr Xavier Domingo-Almenara received a PhD in Bioengineering (2016). He is a Principal Investigator at Eurecat Technology Centre where he leads the Bioinformatics group of the Eurecat's Centre for Omics Sciences.

PROJECT CONTACTS

Dr Núria Canela
GLOMICAVE scientific coordinator
Omic Sciences Unit of Eurecat Technology Centre. Av. Universitat Autònoma, 23. Parc Tecnològic del Vallés, 08290 Cerdanyola del Vallès – Barcelona

info@glomicave.eu
www.glomicave.eu

FUNDING

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952908.