Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes

Baptiste Imbert, Jonathan Kreplak, Raphaël-Gauthier Flores, Grégoire Aubert, Judith Burstin, Nadim Tayeh

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 03 August 2023
DOI 10.3389/frai.2023.1191122

Check for updates

OPEN ACCESS

EDITED BY
Marie-Angélique Laporte,
Bioversity International, France

REVIEWED BY
Hai Wang,
China Agricultural University, China
Florence Amardeilh,
Elzeard, France

*CORRESPONDENCE
Baptiste Imbert
✉ baptiste.imbert@inrae.fr
Nadim Tayeh
✉ nadim.tayeh@inrae.fr

RECEIVED 21 March 2023
ACCEPTED 10 July 2023
PUBLISHED 03 August 2023

CITATION
Imbert B, Kreplak J, Flores R-G, Aubert G, Burstin J and Tayeh N (2023) Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes. *Front. Artif. Intell.* 6:1191122. doi: 10.3389/frai.2023.1191122

Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes

Baptiste Imbert[1]*, Jonathan Kreplak[1], Raphaël-Gauthier Flores[2,3], Grégoire Aubert[1], Judith Burstin[1] and Nadim Tayeh[1]*

[1]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, Dijon, France, [2]Université Paris-Saclay, INRAE, URGI, Versailles, France, [3]Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, Versailles, France

While the continuing decline in genotyping and sequencing costs has largely benefited plant research, some key species for meeting the challenges of agriculture remain mostly understudied. As a result, heterogeneous datasets for different traits are available for a significant number of these species. As gene structures and functions are to some extent conserved through evolution, comparative genomics can be used to transfer available knowledge from one species to another. However, such a translational research approach is complex due to the multiplicity of data sources and the non-harmonized description of the data. Here, we provide two pipelines, referred to as structural and functional pipelines, to create a framework for a NoSQL graph-database (Neo4j) to integrate and query heterogeneous data from multiple species. We call this framework Orthology-driven knowledge base framework for translational research (Ortho_KB). The structural pipeline builds bridges across species based on orthology. The functional pipeline integrates biological information, including QTL, and RNA-sequencing datasets, and uses the backbone from the structural pipeline to connect orthologs in the database. Queries can be written using the Neo4j Cypher language and can, for instance, lead to identify genes controlling a common trait across species. To explore the possibilities offered by such a framework, we populated Ortho_KB to obtain OrthoLegKB, an instance dedicated to legumes. The proposed model was evaluated by studying the conservation of a flowering-promoting gene. Through a series of queries, we have demonstrated that our knowledge graph base provides an intuitive and powerful platform to support research and development programmes.

KEYWORDS

graph database, orthology, ontology, quantitative genetics, gene expression, comparative omics, Ortho_KB, OrthoLegKB

Introduction

To accelerate plant research and manage costs, model species first emerged as a good strategy for studying plant development and stress response, thus providing the research community with data and knowledge. Databases such as TAIR for *Arabidopsis thaliana* (Berardini et al., 2015), MTGD for *Medicago truncatula Gaertn.* (Krishnakumar et al., 2015), miyakogusa-jp for *Lotus japonicus* (Sato et al., 2008) or RAP-DB for *Oryza sativa* (Ohyanagi et al., 2006) were created to centralize, organize and curate the information on model species while providing tools for their analysis. Meanwhile, researchers working on other species have been inferring information from closely-related model plants using orthology and synteny. In fact, orthologs, i.e. genes descending from a common ancestral gene by a speciation event, are likely to have similar and conserved functions (Linard et al., 2021). However, it can prove difficult to identify the correct ortholog of a gene among its homologs based only on sequence similarity because of duplication events. Synteny and collinearity, i.e., conservation of the content and the order of genes on chromosomal regions, respectively, can help identifying orthologous blocks and hence deciphering true orthologous genes (Drillon et al., 2020). Such information is already made available and exploited through platforms such as PLAZA (Van Bel et al., 2012, 2022), sometimes supplemented by tools giving access to gene expression data (Kamei et al., 2016).

With the advent of new technologies, the once daunting sequencing costs have been dramatically reduced (Shendure et al., 2017), allowing for the production of high-quality assembled genomes including for orphan species (Ye and Fan, 2021). These new resources, along with associated annotations, are often being hosted on dedicated websites and/or made available in repositories of well-known databases such as NCBI (Sayers et al., 2022), Ensembl Plants (Yates et al., 2022), Gramene (Tello-Ruiz et al., 2021) or Phytozome (Goodstein et al., 2012). The release of the genome sequences is significantly boosting the production of genetic data to inform on the control of phenotypic traits by genes and the production of -omic data (mostly represented by genomics, transcriptomics, proteomics and metabolomics) to characterize and quantify the different molecules from a biological entity. However, efforts are still uneven across the broad spectrum of species since conducting experiments spanning a wide range of genotypes, tissues and conditions to generate solid data can be very informative, but also expensive and hard to achieve. Also, some quantitative trait loci (QTL) controlling quantitative traits still display low resolution, either due to low marker density or to low recombination rate in the respective genomic regions, which can result in large number of genes within the confidence intervals and long lists of candidate genes. Comparing QTL positions across species can help pinpointing orthologous ones and thus refining the intervals of those with low resolution. Such comparative translational research has also the potential to transfer functional information from one species to another or to a group of species.

Databases are powerful tools to leverage already produced datasets, not only as a mean of storage but also of intelligent exploitation. For example, the Comparative Genomics (CoGe) platform currently allows for the comparison of datasets from a wide range of organisms, with nearly 58,000 genomes available (Lyons and Freeling, 2008). Using sequence homology and synteny, researchers can identify structural and nucleotide variations for their species of interest. Researchers can also use the LoadExp+ extension to import experimental data in various common formats, such as VCF for polymorphism or FASTQ for RNA-seq, process them, and display the results as tracks in the genome browsers (Grover et al., 2017). Nevertheless, the data are predominantly stored using relational database management system (RDBMS), distributed in category-specific tables. One problem than can arise with RDBMS is that connecting tables containing large datasets during querying requires several joining operations, which are expensive in terms of time and computational resources (Vicknair et al., 2010).

The intuitive idea of structuring intertwined data into a graph was propelled by the World Wide Consortium for semantic web through the Resource Description Framework (RDF), (W3C, 1994). In order to obtain a logical model in RDF, each piece of data is sliced into atomic statements stored as triples, i.e., (1) the subject of the resource to describe, (2) a property assigned to the resource, termed predicate and (3) the object, either a description or another resource. The subject and the object are nodes in the graph, while the predicate is an edge connecting the two nodes (Abuoda et al., 2022). The directional decomposition of information allows the use of ontologies that organize knowledge and greatly improve data sharing in scientific communities (Stevens et al., 2020). However, databases using this format, called triplestores or RDF stores, are characterized by an atomic granularity of nodes which can make database modeling tedious. In addition, deep traversal of the graph requires self-joining of all traversed triples which can make the cost of traversing edges logarithmic (Donkers et al., 2020).

Alongside RDF, labeled-property graph (LPG) databases have emerged, currently led by Neo4j, which are fundamentally designed to improve graph traversal by directly storing on disk all existing edges between nodes. A benchmark from Khayatbashi et al. (2022) comparing RDF triple-stores and LPG databases with twelve queries shows that Neo4j is in fact more efficient to traverse multiple layers of data. Neo4j databases offer high flexibility by adding key-value properties to nodes and edges to effectively compact information, consequently making the modeling easier to read and to incrementally improve (Donkers et al., 2020; Neo4j, 2023b). Considering these assets, Neo4j databases were found advantageous to manage dense networks of information required for systems biology. The Reactome database (Fabregat et al., 2018) and its plant counterpart Plant Reactome (Naithani et al., 2019) have already switched from an RDBMS database to a Neo4j database, since metabolic pathways are intrinsically connected as a graph structure. In fact, using a graph database dropped the average query time of Reactome by 93 % (Fabregat et al., 2018). While a graph is intuitive when representing a biological pathway, the value of such modeling extends to many applications, including translational research. For instance, orthologous relationships across genes required for translational research, could be modeled in a Neo4j database with an "IS_ORTHOLOGOUS_TO" relationship between the two "Gene" nodes. Information regarding the gene identifier or annotation could be stored as internal node properties, available for querying. As the system is adaptable, new layers of data can successively be added and articulated. Omics Database Generator (ODG) is the first LPG designed for translational research as defined by Guhlin et al. (2017). ODG is a Neo4j graph-database, developed primarily for annotation transfer to non-model species of bacteria and plants. The structure of ODG has been made available for researchers to import their own data. Indeed, the comparison of newly generated data with existing data can confirm hypotheses or help to generate new ones. This is especially useful when datasets do not yield results supporting the initial research hypothesis, end up being set aside and remain unpublished (Raciti et al., 2018). It is therefore crucial to use as many available and high-quality datasets as possible, whether published or unpublished, as valuable sources of knowledge. However, ODG does not offer support for the integration of annotated genetic data, which is necessary for crop improvement, and it is likely to be difficult for non-expert users to understand its model and its underlying potential (Misra et al., 2019; Kaur et al., 2021).

The legume family (Leguminosae or Fabaceae) is the third largest family of flowering plants, with about 750 genera and nearly 19,500 species (The Legume Phylogeny Working Group et al., 2013). The Leguminosae include many taxa of agricultural or other economic importance and significant research efforts are needed to advance legume breeding and address the new challenges imposed to agriculture, namely production under climate change, with less pesticides and fertilizers. *Pisum sativum* L. (pea), *Lens culinaris* Medik. (lentil) and *Vicia faba* L. (faba bean) are examples of grain legumes that produce protein-rich seeds and play a key role in sustainable cropping systems (Guiguitant et al., 2020; Rubiales et al., 2021; Semba et al., 2021). Because of their large genomes, sometimes up to 30 times larger than the genome of the model legume *M. truncatula* (Jayakodi et al., 2023), the creation of -omics data on these species has lagged behind. In addition, data on a given species were mostly produced by the research community in the country of production, as the dominant production areas are sometimes different. Several databases have been developed that attempt to inventory the diversity of published datasets and provide tools to analyse and visualize them, including Soybase (Grant et al., 2010), the Pulse Crop Database (Humann et al., 2019), KnowPulse (Sanderson et al., 2019) and the Legume Information System (Berendzen et al., 2021). However, there is still a lack of options to link multi-species datasets together for further study.

LegumeIP is a relational database, initially created to transfer knowledge from model to crop legume species, and recently transformed into an integrative platform to support translational research, hosting homology, gene annotation and expression data for 17 legume species in its latest version (Li et al., 2012, 2016; Dai et al., 2021). Some recently sequenced cool-season legumes are however missing, including *P. sativum* (Kreplak et al., 2019), *L. culinaris* (Ramsay et al., 2021) and *V. faba* (Jayakodi et al., 2023). In addition, the interface of LegumeIP is designed to facilitate pairwise comparisons, from model species to less studied crop species, making the current design unsuitable for simultaneous comparison of multi-species experiments.

Here, we developed Ortho_KB, a robust framework for translational research in diploid plant species. We developed a first pipeline to compute homology and define syntenic chromosomal regions across species. This method was chosen to identify putative orthologs among homologs, thus establishing links between corresponding genes and connecting chromosomes. We designed a second pipeline to execute custom scripts that reformat all heterogeneous data files, including -omics datasets, for input into the database. Users can integrate published and unpublished information related to their species of interest including gene-phenotype associations from QTL data and expression information from transcriptomic resources and use the provided framework to get the most out of their data. Ortho_KB provides an intuitive database model that can be queried using Cypher language, to extract meaningful information in comma-separated values (CSV) files for further analysis. The framework has been applied to a subset of legume species, resulting in a database called OrthoLegKB, a multi-species and multi-omics graph-based database for collecting, integrating and querying heterogeneous data. OrthoLegKB currently allows the comparison of genetic, and -omic data from 5 legume species, i.e., *P. sativum*, *V. faba*, *L. culinaris*, *Vigna radiata (L.) R.Wilczek* and *M. truncatula*. Finally, a use-case is described to demonstrate how the combination of quantitative genetics and expression data is possible in OrthoLegKB and can benefit translational research.

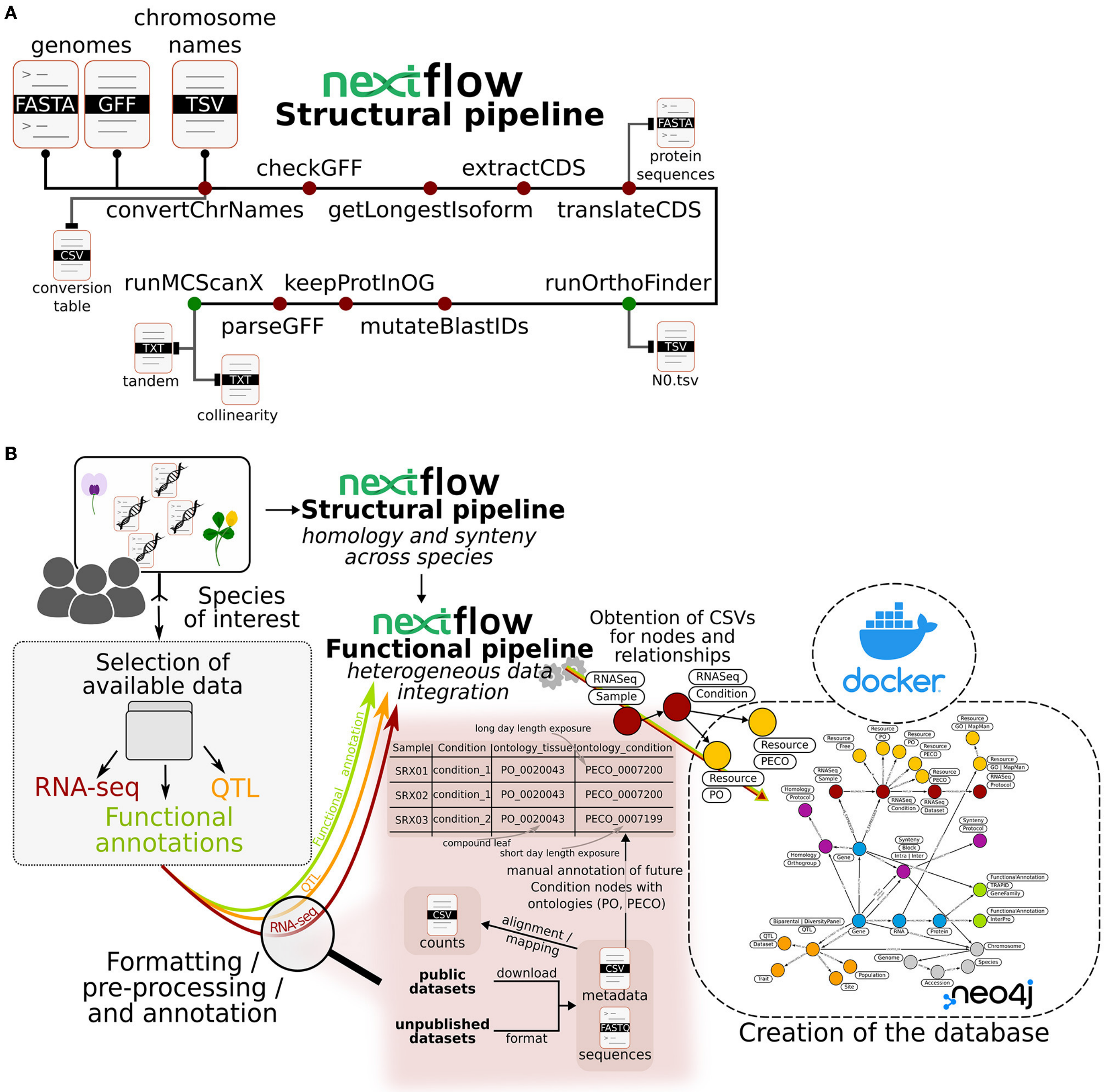
Materials and methods

Orthology and synteny

As illustrated in Figure 1A, in order to identify homologous genes and syntenic regions, genome FASTA and annotation files as well as an optional conversion table for chromosomes are used as input files. The conversion table must include the original chromosome ID and the desired ID in the database. Unique chromosome and scaffold IDs across species are more convenient for querying and are also required by synteny-visualization tools such as SynVisio (Bandi and Gutwin, 2020). The steps for synteny and orthology discovery are the following: (1) curate annotation files using the agat_convert_sp_gxf2gxf.pl parser from agat v0.9.1 by automatically removing duplicated features and/or IDs, inferring missing IDs or parent features; (2) filter annotation files to keep only the longest isoform using the agat_sp_keep_longest_isoform.pl script; (3) extract coding DNA sequences (CDS) using the agat_sp_extract_sequences.pl script (Dainat et al., 2022); (4) generate protein sequences using the Seqkit v2.3.0 translation module (Shen et al., 2016); (5) submit protein sequences in FASTA format to OrthoFinder v2.5.4 with its default parameters using Diamond v2.0.12 in ultra_sensitive mode for the alignment instead of BLAST (Emms and Kelly, 2015, 2019; Buchfink et al., 2021). Finally, to connect homologous chromosomal regions, the OrthoFinder output is used to obtain syntenic blocks. First, alignment files are filtered to retain only pairs of proteins that are part of the same orthogroup. Second, these filtered alignment files are provided to MCScanX along with a merge of annotation files from all considered species (Wang et al., 2012). A minimum number of 10 genes to form a collinear block is set by default in the pipeline. All above-mentioned steps were included in a single pipeline, called the structural pipeline, using Nextflow (Di Tommaso et al., 2017).

Functional gene annotation

Functional annotations were conducted by manually submitting CDS sequences to annotation tools either available on online platforms or to be run locally. The TRAPID online tool returned gene families, RNA families, and Gene Ontology (GO) terms associated with submitted genes (Bucchini et al., 2021).



Sample	Condition	ontology_tissue	ontology_condition
SRX01	condition_1	PO_0020043	PECO_0007200
SRX02	condition_1	PO_0020043	PECO_0007200
SRX03	condition_2	PO_0020043	PECO_0007199

FIGURE 1
Schematic representation of the pipelines used to build Ortho_KB, a NoSQL graph database framework for translational research. **(A)** The structural pipeline computing homology between genes and synteny across chromosomal regions from selected annotated genomes. All processes included in the pipeline, except those producing the mandatory final outputs, are represented by dark red circles. Processes producing the mandatory final outputs are represented by green circles. **(B)** General overview of the steps leading to the construction of an instance of Ortho_KB. Datasets that can be managed include RNA-seq data, QTL and functional annotations. As an example, we develop the treatment of an RNA-seq dataset from public or private origin. Alongside a regular extraction of counts, metadata of the samples must be annotated using ontologies to describe in particular the tissue of origin (Plant Ontology) and the experimental conditions to which the sample was subjected to (Plant Experimental Conditions Ontology). The functional pipeline will process inputed files and in this case the annotated metadata file will produce "Sample" and "Condition" nodes in the graph. This last node will also be connected by relationships to "Resource" nodes corresponding to the ontologies, thereby conserving the metadata information in the Neo4j graph database. The graph database is included in a Docker container, as shown on the right-hand side of the schema.

Genes that share sequence homology are gathered in gene families. "GeneFamily" nodes hold links to the Plaza website on which information regarding family-associated GO annotations and InterPro domains are available (Van Bel et al., 2022). To attribute summarized functions to genes, we assigned MapMan bins using the online Mercator4 (https://www.plabipd.de/portal/web/guest/mercator4) which resulted in a hierarchical annotation (Lohse et al., 2014; Schwacke et al., 2019). We also used eggNOG-mapper to obtain human-readable annotation and gene symbols from protein sequences (Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). A name is assigned to each gene when available in the literature. For instance, *MtrunA17Chr3g0135361* is annotated as *ELF3* for *EARLY FLOWERING 3*. Predicted proteins of each species were further annotated by locally running InterProScan v.5.53 with the "iprlookup" option (Jones et al., 2014; Blum et al., 2021), notably using databases such as Pfam (Mistry et al., 2021), Gene3D (Lees et al., 2012) or PANTHER (Mi and Thomas, 2009).

Genetic data extraction

The exact set of mandatory and optional information required to describe QTL data in Ortho_KB are described in the documentation available on the dedicated Git repository (see "Data availability statement" section). Briefly, the identifier, the trait name, and the associated markers are essential. A QTL arising from a study on a biparental mapping population is defined by a physical position on a chromosome between two flanking genetic markers and a peak marker within the confidence interval, if the information is available. A QTL from a genome-wide association study (GWAS) analysis is defined by a single marker location on a chromosome corresponding to the peak marker unless linkage disequilibrium data are available, then data is processed similarly to a QTL from a biparental population. Therefore, a QTL record might have information for one up to three markers. QTL data in the current version of OrthoLegKB were collected from published research articles (Supplementary Table S3).

Transcriptomic data extraction and expression quantifications

RNA-seq datasets were manually selected from NCBI. Sample IDs associated within a BioProject were collected in each case using esearch from entrez-direct v16.2 (Kans, 2013). The sample list was fed into nf-core/fetchngs pipeline v1.7 with the option "nf_core_pipeline rnaseq" to obtain all FASTQ files along with a metadata file (Patel et al., 2022). The nf-core/rnaseq v3.8 pipeline (Patel et al., 2023) was then run with the genome files, metadata file and FASTQ files with the arguments "skip_alignment", "pseudo_aligner salmon" and "salmon_quant_libtype A" to automatically assess strandedness (Patro et al., 2017). Salmon result files were finally processed into matrices for downstream analyses using tximport (Soneson et al., 2016). The "salmon.merged.gene_counts.tsv" file containing read counts and the "salmon.merged.gene_tpm.tsv" file containing the Transcript Per Million (TPM) normalized quantification were used for further processes. Samples listed in the metadata file originating from the nf-core/fetchngs pipeline were manually annotated to indicate the tissues used, the environmental conditions applied, and the experimental area (field, greenhouse, etc.), using the Plant Ontology (PO) and the Plant Experimental Condition Ontology (PECO) (Cooper et al., 2018).

Database construction and implementation

Graph database conceptualization

The current release of Ortho_KB was built as a NoSQL database framework to store and display data in a graph structure, using the Neo4j Community Edition v4.4.18 (Neo4j, 2023b).

We chose the Neo4j graph-database management system because of (1) its efficiency in handling highly connected data, (2) the graph algorithms already implemented and (3) the expressive Cypher query language it uses and (4) its capacity to import/export data using semantic web technologies. Entities, also called nodes, and edges, also referred to as relationships, were designed in a way to carry the biological information. Each gene or transcript is represented by a node and each gene is linked to its corresponding transcript by a relationship (e.g., gene *A* has a transcript RNA *A1*). Multiple properties can be stored and queried on nodes (e.g., RNA *A1* sequence length) and relationships (e.g., position of a protein domain on the protein sequence). In addition, one or more labels can be applied to nodes to group them into a set to facilitate querying. In this paper, labels are indicated by double quotation marks, for instance the "RNA" label for nodes of transcripts.

Input files processing

A Nextflow pipeline called functional pipeline, was created to process heterogeneous data from the previously described sources. The pipeline requires genome files, functional annotation files, RNA-seq files and QTL files to run. For the functional annotations, the pipeline includes a set of scripts to filter and format them into nodes and relationships, following the database model. For the GO annotations obtained from TRAPID, by default, only the most specific GO terms are retained for each gene by selecting those with parameter "is_hidden" equal to 0, resulting in a 90% reduction in the number of GO terms directly associated to genes. The GO W3C Web Ontology Language (OWL) file is downloaded and parsed to import the "is_a" and "part_of" predicates as relationships in the graph to allow graph traversal (W3C, 1994). Similarly, the provided annotation files from MapMan are used to create an ontology in TURTLE syntax using rdflib v.4.2.2 (Grimnes et al., 2023). For RNA-seq, salmon pseudo-counts are by default filtered to retain only genes for which the sum of TPM across samples is >5, to avoid creating many relationships for non-expressed genes. Gene expressions in all samples from the same condition are averaged, and both arithmetic and geometric means are stored on the edge between the "Gene" and the "Condition" nodes. For genetic data, previously formatted files are processed to identify genes included in the confidence interval of QTL using pybedtools v.0.9.0 (Quinlan and Hall, 2010; Dale et al., 2011).

Briefly, for all processes, the pipeline creates CSV files to populate the database and a summary file listing all CSV files to be imported in a format readable by Neo4j. All nodes and relationships that can be generated are described in Supplementary Tables S2, S3.

Database implementation

A Bash script was written to create and populate the database. It includes three steps. The first step prepares the import environment by building a Docker container. Running the

Docker container will start the database, by default available at http://0.0.0.0:7474/browser/. The second step performs the import to populate the Neo4j database using the neo4j-admin import command. The third step imports the PO, PECO, GO and MapMan ontologies using the n10s.onto.import.fetch method from the neosemantics (n10s) plugin (Barrasa, 2022). The import creates a node per term, connected to broader terms by a "SCO" relationship obtained from the property rdf:subClassOf. The nodes of the resulting subgraph are then labeled according to their source (PO, PECO etc.) and connected to the rest of the graph using a set of Cypher queries.

Plant species selection for OrthoLegKB

For this study, five species were chosen. These include the model legume *M. truncatula*, three cool-season legumes of agronomic importance, i.e., *P. sativum*, *L. culinaris*, *V. faba,* and a relatively distant warm-season legume species, *V. radiata*. All species belong to the Galegoids subclade, with the exception of *V. radiata*, which is part of the sister group, the Milletoids subclade. We selected the latest genomic data from *P. sativum* cultivar Cameor v.1 assembly (Kreplak et al., 2019), *M. truncatula* accession A17 v.5 assembly and v.1.9 annotation (Pecrix et al., 2018), *V. faba* accession Hedin/2 v.1.0 assembly (Jayakodi et al., 2023), *L. culinaris* cultivar CDC Redberry v.2.0 assembly (Ramsay et al., 2021) and *V. radiata subsp. radiata* cultivar VC1973A v.6 assembly (Ha et al., 2021). All genomes were assembled into chromosomes, generated using long-read technology, except for *P. sativum*. The *M. truncatula* annotation file was filtered to keep only features from EuGene and BioFileConverter. Gene prefixes were also modified using a custom script. Details on selected genome assemblies and genome statistics are available in Table 1.

Data visualization

The visualization of the graph model was created using Arrows (Neo4j, 2023a). The UpSet plot was created using UpSetR (Conway et al., 2017). Visualization of large-scale synteny was performed with the SynVisio online tool (Bandi and Gutwin, 2020) or with tailored R scripts, while microsynteny was plotted using the R package gggenomes (Hackl and Ankenbrand, 2023).

Hardware and query time

The server hosting OrthoLegKB is based on an OpenStack infrastructure, with 4 virtual CPUs and 8 Gb of RAM. For each query presented in the Results section, the average response time over five iterations was indicated.

Results

Ortho_KB is a framework for translational research in plant species

Studying a particular trait or gene often requires the collection of different types of information available on different websites and databases, for the species of interest as well as for close species. We have created Ortho_KB, a database framework built with successive pipelines to facilitate the exploration of all data relevant to a trait or gene of interest in a single environment. Ortho_KB provides a unique and multi-functional structure that can be populated with datasets of interest and then queried for comparative and functional genomics studies. The current Ortho_KB modeling aims at enabling translational research across a wide range of selected species by making data easily searchable and the process more straightforward. The framework relies heavily on orthology and synteny relationships to build bridges between species, and transfer and/or compare genetic and genomic information between them. A Nextflow pipeline, called the structural pipeline (Figure 1A), first identifies groups of homologous genes – orthogroups – based on protein sequence similarity. It then looks for conserved gene order between pairs of chromosomes, within or between species, to highlight collinear regions. Homologs in collinear regions are more likely to be orthologs and therefore have similar functions. A second Nextflow pipeline, called the functional pipeline, connects information from the first pipeline and additional data available from separate tables including gene annotation, gene expression and QTL positions (Figure 1B). All heterogeneous data are thus properly formatted for integration into the database.

Ortho_KB uses Neo4j graph-database management system

The Neo4j graph-database management system handles entities as nodes and their connections as relationships. In Ortho_KB, the data model revolves around "Gene" nodes, characterized by their start and stop positions on chromosomes (Figure 2). The "Gene" nodes are connected to their putative transcript ("RNA") nodes, themselves connected to the predicted proteins ("Protein") resulting from the translation of their RNA sequences. Homology and collinearity information computed using the structural pipeline create bridges across species at the gene and the chromosome levels, respectively. The current version of Ortho_KB includes 29 core categories of nodes tagged either by a single label, like "Gene" nodes or by a set of labels, like "RNASeq" supplemented by "Condition". They are connected by directed relationships, sometimes bearing additional properties (Figure 2). Individual nodes are defined by a unique identifier. For example, a "Gene" node is defined by a gene ID, matching the feature ID from the General Feature Format 3 (GFF3) annotation file, unique across species. Ortho_KB can be queried through the web Neo4j Browser, the terminal or other interfaces provided by Neo4j (Neo4j, 2023b).

Ortho_KB integrates different categories of data including gene annotation, genetic and transcriptomic resources

As shown in Figure 1, Ortho_KB gathers different categories of data.

In terms of functional annotation, complementary information sources are handled. These include TRAPID's gene families, GO annotations, MapMan bins and InterPro that are each integrated

TABLE 1 Specifications of species included in OrthoLegKB.

Species	Genotype	Assembly size (Mb)	Number of chromosomes	Protein coding genes	Assembly references
Lens culinaris	CDC Redberry	3,760	7	58,243	Ramsay et al., 2021
Medicago truncatula	A17	430	8	44,626	Pecrix et al., 2018
Pisum sativum	Cameor	3,920	7	46,905	Kreplak et al., 2019
Vicia faba	Hedin/2	11,900	6	34,221	Jayakodi et al., 2023
Vigna radiata	VC1973A	476	11	30,882	Ha et al., 2021

Values for "Protein coding genes" take into account a single isoform per gene.

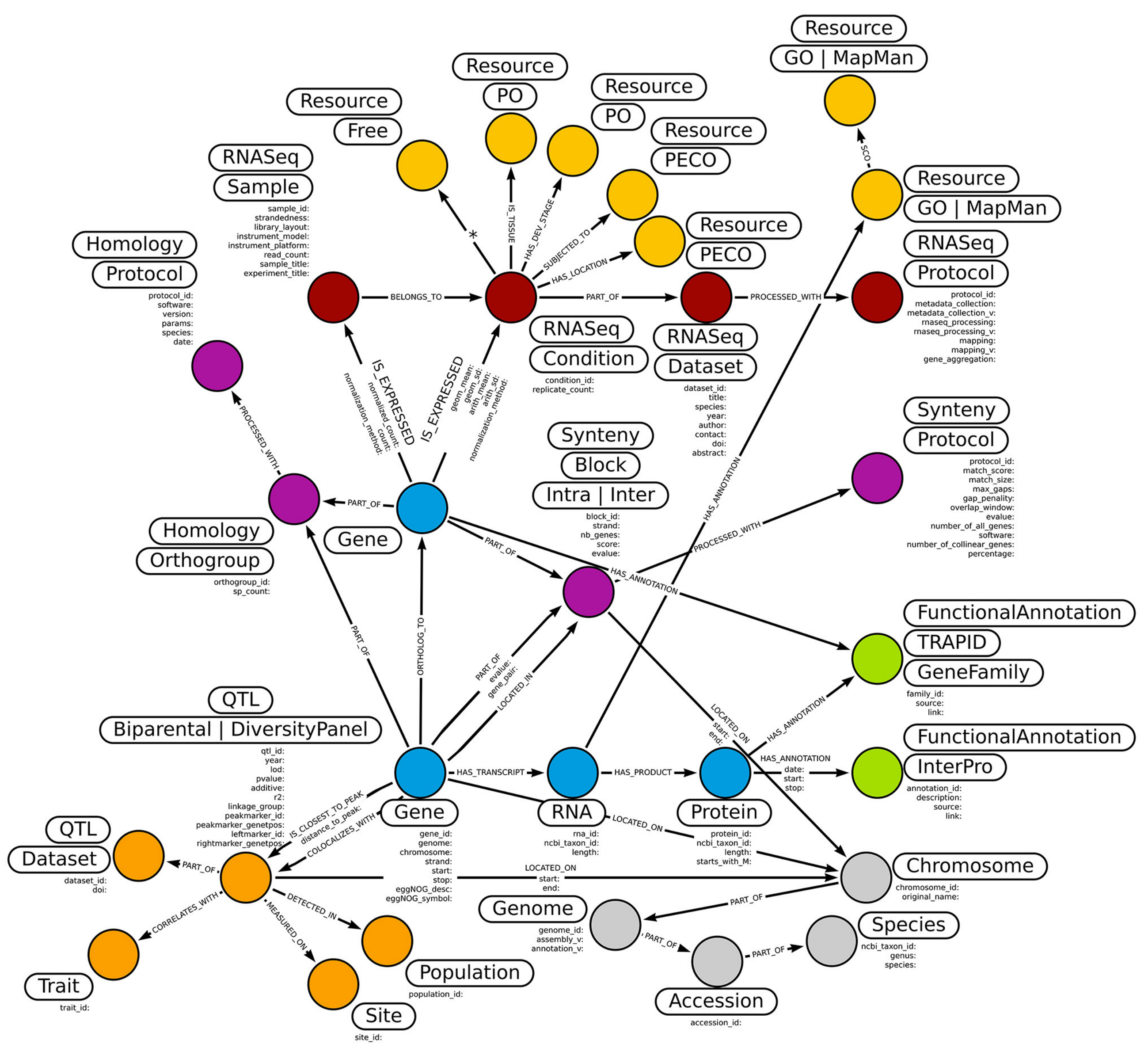


FIGURE 2
Overview of the Ortho_KB translational database model. In the graph model, colored circles represent the 29 core node types, which are entities with labels and properties. "Gene", "RNA", and "Protein" and related genomic nodes are shown in blue, "Homology" and "Synteny" and related nodes in mauve, ontology term nodes in yellow, the RNA-seq nodes in dark red, functional annotation nodes in light green, taxonomic nodes in light gray, and QTL-related nodes in orange. The category of each node is described by the associated labels, which are contained in elongated boxes near the nodes, and the properties correspond to the lists of elements placed below the labels. Nodes are connected to each other by relationships, represented by arrows, which can also store information as properties.

in a separate node type. TRAPID gene families and Mapman bins provide synthetic overviews of gene functions while GO annotations and InterPro provide detailed descriptions focusing on gene functions and protein domains, respectively (Figure 2).

Regarding genetic data, the model includes connections between genes and QTL information either resulting from QTL mapping in biparental populations or GWAS in diversity panels. Since the two mapping approaches are grouped with the "QTL" label, we added a second label, either "BiparentalPopulation" or "DiversityPanel" to differentiate them. All genes located within the confidence interval of a QTL are connected to the "QTL" node with a "COLOCALIZES_WITH" relationship. The closest gene to the peak marker is additionally connected to the "QTL" node with a "IS_CLOSEST_TO_PEAK" relationship with its distance to the peak marker included as a property. Additional information to describe a QTL are included in connected nodes such as the experimental geographical "Site", the studied "Population" and the "Trait" (Figure 2).

For transcriptomics, we have developed scripts to handle read counts. Read counts can either be generated using the pipeline of Patel et al. (2023) to optimize comparability of data (see "Materials and methods" section), or according to the method chosen by the user. The user is also free to integrate data previously analyzed with other methods. Replicates originating from the same biological condition are summarized into a condition that has to be manually annotated with ontology terms describing best the experimental conditions and biological material. The PECO and PO ontologies were selected for this purpose (Cooper et al., 2018). Using n10s inference, this model allows to traverse the ontology and unveil datasets from experiments performed in similar conditions (Barrasa, 2022). If no ontologies are available to appropriately describe conditions, free terms might be introduced (Figure 2).

OrthoLegKB was developed with Ortho_KB to provide a translational tool for grain legumes

To prove how Ortho_KB can serve translational approaches and the research goals of a scientific community, we chose to apply it to five diploid legume species belonging to the Galegoid (cool-season legumes) and Milletoid (warm-season legumes) clades creating the OrthoLegKB database.

To leverage data from all five species using comparative genomics, we started by searching for orthologs with the structural pipeline using the latest genome assemblies. The pipeline was run for 740 CPU hours with 20 CPUs allocated (7 h 40 in real time), with a maximum physical memory usage of 46 Gb. In total, 14,565 out of 29,428 total orthogroups (49.49 %) were shared by all five species and 8.24 % by all species excluding *V. radiata,* the only representative of the Milletoid clade (Figure 3).

Then, public datasets with QTL and RNA-Seq data were mined, annotated with the ontologies used in Ortho_KB and included in the database using the functional pipeline. The pipeline was run for 11 CPU hours (14 min in real time), with a maximum physical memory usage of 2 Gb. A list of these datasets is available in Supplementary Table S1. OrthoLegKB currently contains more than 815,000 nodes and close to 15,000,000 relationships associated to the different types of data. The exact number of nodes in each category can be found in Supplementary Table S4.

OrthoLegKB can be used to address various scientific questions including the conservation of the control of flowering time in legumes

As a use-case to demonstrate how to exploit OrthoLegKB, we searched for the orthologs of a previously-studied flowering time regulator, the *FLOWERING LOCUS T* from *M. truncatula* (*MtFTa1*) and sought evidences for potential conserved function across species. For this use-case, we have decided to work only on cool-season legume species. *MtFTa1* has been thoroughly studied (Hecht et al., 2007, 2011; Laurie et al., 2011; Cheng et al., 2021) and its physical position on *M. truncatula* chromosome 7 (Mt07) is known. It is identified as *Medtr7g084970* (Laurie et al., 2011; Cheng et al., 2021) or *MtrunA17Chr7_39606925_39618489* in the GFF3 of the Mt5.0 (r1.9) genome annotation version. The first step was the identification of candidate orthologous genes from *P. sativum*, *L. culinaris* and *V. faba*. Several candidates could be found across chromosomes through a single query in 11 ms (Figure 4). OrthoLegKB was then searched for syntenic blocks encompassing these candidate genes. Synteny between chromosome 3 from *P. sativum* (Ps03), chromosome 6 from *L. culinaris* (Lc06), chromosome 5 from *V. faba* (Vf05) and Mt07 at the *MtFTa1* locus was revealed highlighting the orthologs (Figure 5A). The syntenic blocks in *L. culinaris* and *V. faba* displayed each one orthologous *FTa1* gene, while two possible orthologous genes were detected in *P. sativum* namely *Psat3g090720* and *Psat3g090680* (Figure 5B). According to the conservation of protein length and domains' annotation information from the PANTHER database stored in OrthoLegKB, *Psat3g090720* seemed to be more similar to *MtFTa1* (Figure 6). In fact, *Psat3g090680* corresponds to *FTa2* described in Hecht et al. (2011). To examine any possible links with flowering control and thus function conservation, we searched for all QTL related to flowering contained in the previously identified syntenic blocks, allowing to also return QTL that did not include *FTa1* genes in their confidence intervals (Supplementary Table S5). As depicted in Figure 7, the query identified three QTL from Aguilar-Benitez et al. (2021) on Vf05, located at the same nucleotidic positions, that were linked to the number of days from the sowing until 50 % of the plants had visible open flowers (DF50_09-10(2)_1) and the number of days from the sowing until the appearance of the first flower (DF1_07-08(3)_1 and DF1_06-07(2)_1). On Ps03, a QTL from Gali et al. (2018) corresponding to the number of days to flowering (PR15_26_1) was found upstream of the *FTa1* locus (2018). Two QTL from Williams et al. (2022), closer to the *P. sativum* locus and associated to the number of days to flowering (DTF3_1) and number of nodes on the main stem to the first flower in long days (DTF3_3) were also identified. The *L. culinaris FTa1* gene was the only gene to be part of the confidence interval of a flowering-related QTL, qDTF.6-2_1. qDTF.6-2_1 is a number of days to flowering QTL from Haile et al. (2021), and close to the qDTFL-6A_1 from Yuan et al. (2021) related to the number of

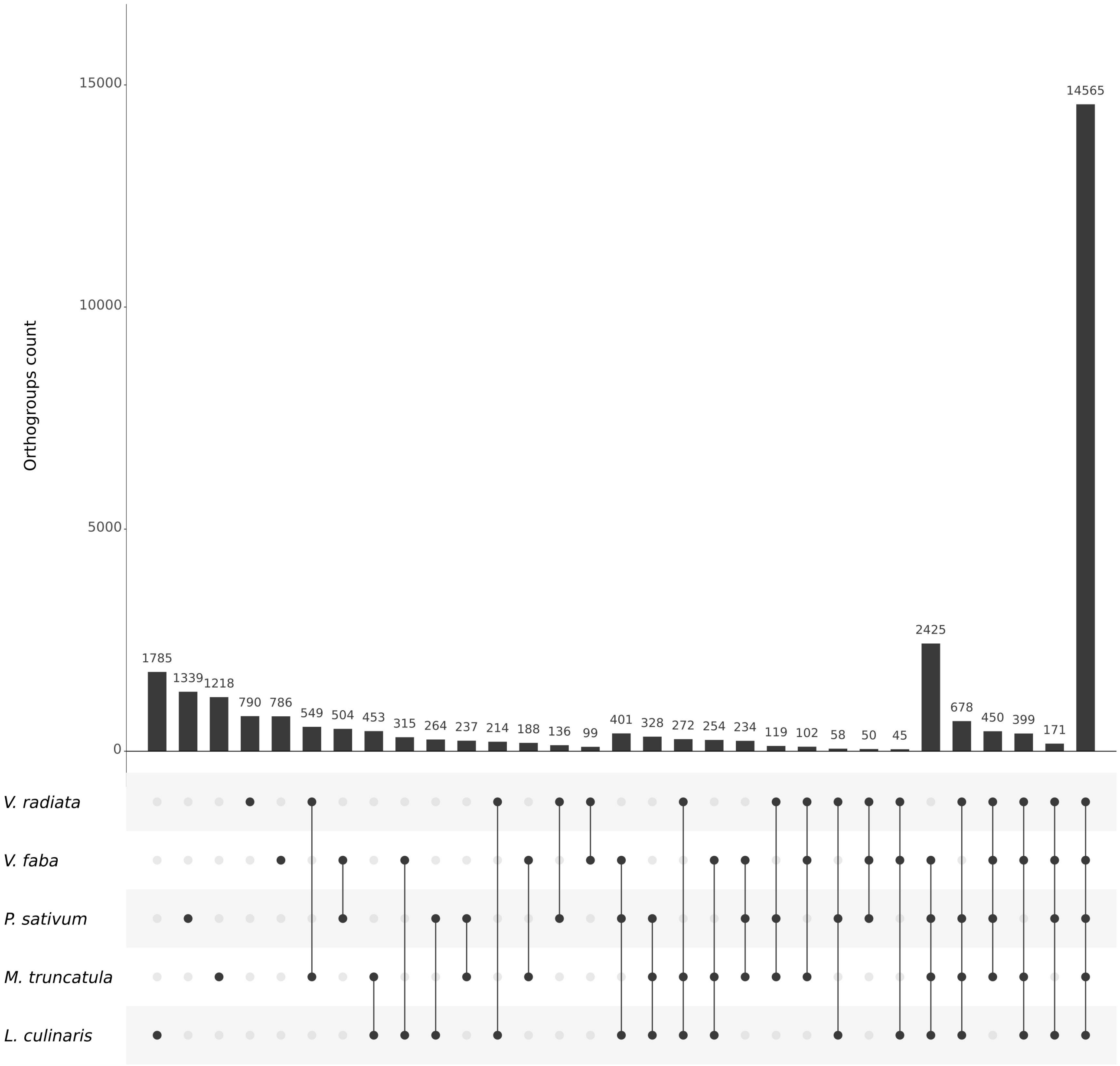
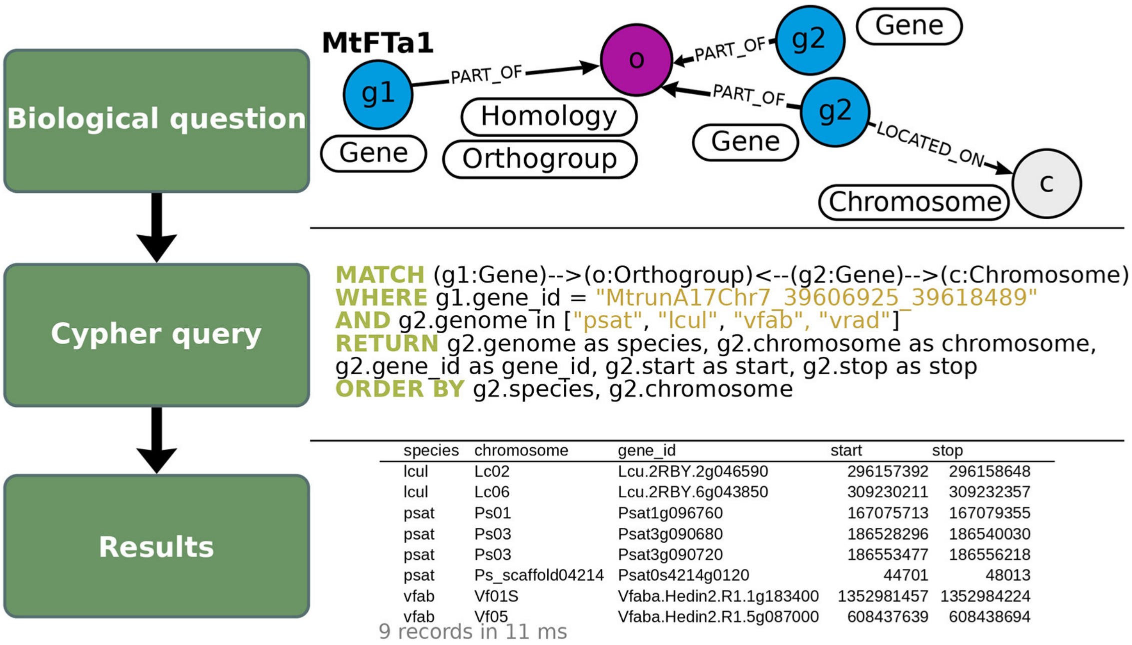


FIGURE 3
UpSet plot highlighting the number of orthogroups within and between legume species included in OrthoLegKB. The structural pipeline of Ortho_KB was used to identify the orthogroups. The bar plot shows the number of orthogroups for each possible set of species. The dots indicate the species associated with each bar.

days to flowering under low red/far-red light quality. Regarding expression, the *MtFTa1* gene is known to be mainly expressed in leaves and stems in *M. truncatula* (Laurie et al., 2011; Thomson et al., 2019). Therefore, we sought to investigate the top three tissues from the shoot system in which its orthologs were mostly expressed. Thanks to the inference allowed by the annotation of conditions with ontologies, we show that in the collection of RNA-seq datasets available for *M. truncatula* in OrthoLegKB, *MtrunA17Chr7_39606925_39618489* was mainly expressed in vegetative shoot apex, reproductive shoot apex and vegetative shoot system. *Psat3g090720* was expressed in the peduncle, stem and leaf tendrils. Transcripts from *Vfaba.Hedin2.R1.5g087000* were predominantly detected in adult vascular leaves, pods, and stems. For *L. culinaris*, all the experiments integrated were performed on leaves, in which the expression of *Lcu.2RBY.6g043850* was detected (Figure 8) and more particularly under far-red light conditions.

Discussion

This paper presented the use of knowledge graphs to integrate genetic and -omics data with the aim of facilitating translational research. The main philosophy was to provide a single environment where heterogeneous datasets from multiple species can be



```
MATCH (g1:Gene)-->(o:Orthogroup)<--(g2:Gene)-->(c:Chromosome)
WHERE g1.gene_id = "MtrunA17Chr7_39606925_39618489"
AND g2.genome in ["psat", "lcul", "vfab", "vrad"]
RETURN g2.genome as species, g2.chromosome as chromosome,
g2.gene_id as gene_id, g2.start as start, g2.stop as stop
ORDER BY g2.species, g2.chromosome
```

species	chromosome	gene_id	start	stop
lcul	Lc02	Lcu.2RBY.2g046590	296157392	296158648
lcul	Lc06	Lcu.2RBY.6g043850	309230211	309232357
psat	Ps01	Psat1g096760	167075713	167079355
psat	Ps03	Psat3g090680	186528296	186540030
psat	Ps03	Psat3g090720	186553477	186556218
psat	Ps_scaffold04214	Psat0s4214g0120	44701	48013
vfab	Vf01S	Vfaba.Hedin2.R1.1g183400	1352981457	1352984224
vfab	Vf05	Vfaba.Hedin2.R1.5g087000	608437639	608438694

9 records in 11 ms

FIGURE 4
Illustration of the query used to search for putative orthologs of *MtFTa1* in OrthoLegKB. Putative orthologs in pea (psat), lentil (lcul), faba bean (vfab) and mung bean (vrad; **top panel**) were queried in Cypher **(middle panel)**, and several properties were returned in CSV format **(bottom panel)**. Genes belonging to the same orthogroup as *MtFTa1* were selected and their positions on the respective chromosomes were returned. Note that relationships' names were not displayed in the query section to keep it concise but were specified when running the query. The number of records returned in the output table and the average response time of the query are shown in light gray below the table.

accessed and examined with quasi-instantaneous querying time, thus allowing to address relevant biological questions, generate hypotheses, and transfer information from a single or group of species to others. The current version of the framework handles genome annotations, QTL and transcriptomic data. Users can identify orthologs, highlight candidate genes for specific traits, pinpoint possible pleiotropy and reveal conserved functional synteny. Ortho_KB gives the opportunity to capitalize on both published and unpublished datasets for further valorisation. The interest of such a database was demonstrated by populating Ortho_KB to create OrthoLegKB, a database dedicated to research on legume crop species, and supported by a use-case study focusing on a flowering-time gene.

Ortho_KB leverages recent analytical workflows and ontology standards to host high-quality data and ensure comparability across datasets

The Ortho_KB framework was built with the hypothesis that homologous genes found in collinear regions are most likely to be orthologs. Collinearity mitigates the effects of genome duplication and fractionation and thus most likely pinpoints true orthologs (Tang et al., 2008). Besides bridges between genomes based on orthology, additional information layers were incrementally integrated and connected to gene entities, taking advantage of the modeling flexibility allowed by Neo4j. The integration of such information was planned following homogenization rules for quality purposes. For expression data, we chose to use a single pipeline to process all transcriptomic datasets and avoid prejudice related to discrepancies in bioinformatic analysis protocols including alignment procedure, GC bias treatment. A similar initiative was taken for the gene atlas dedicated to *M. truncatula* (Carrere et al., 2021). We further decided to integrate normalized expression but not differential expression (DE). In fact, since the aim with Ortho_KB is to explore gene expressions across multiple samples and experiments, including expression in the form of DE would restrict analyses to a specific imposed comparison. Yet, the support for differential expression might be provided in the near future. Several actively updated ontologies (PO, PECO) were further selected to best describe the various experimental conditions from which the transcriptomic data were obtained. Since it requires human expertise, the annotation of samples with ontologies remains manual in the current version of the framework.

Regarding QTL, and unlike trancriptomics data, the reprocessing approach in sake of comparability could not be established so far as the analysis requires access to metadata, which are often sparsely provided in the literature. However, as FAIR standards are gaining in popularity, a unified approach might be considered for genetic data analysis in an upcoming version (Wilkinson et al., 2016). To ensure that positions of QTL for similar traits can be compared within and between species, homogeneity in traits denominations is required. This constraint is difficult to meet as a trait can be measured or named differently. For example, the flowering time might be considered by some authors as the time

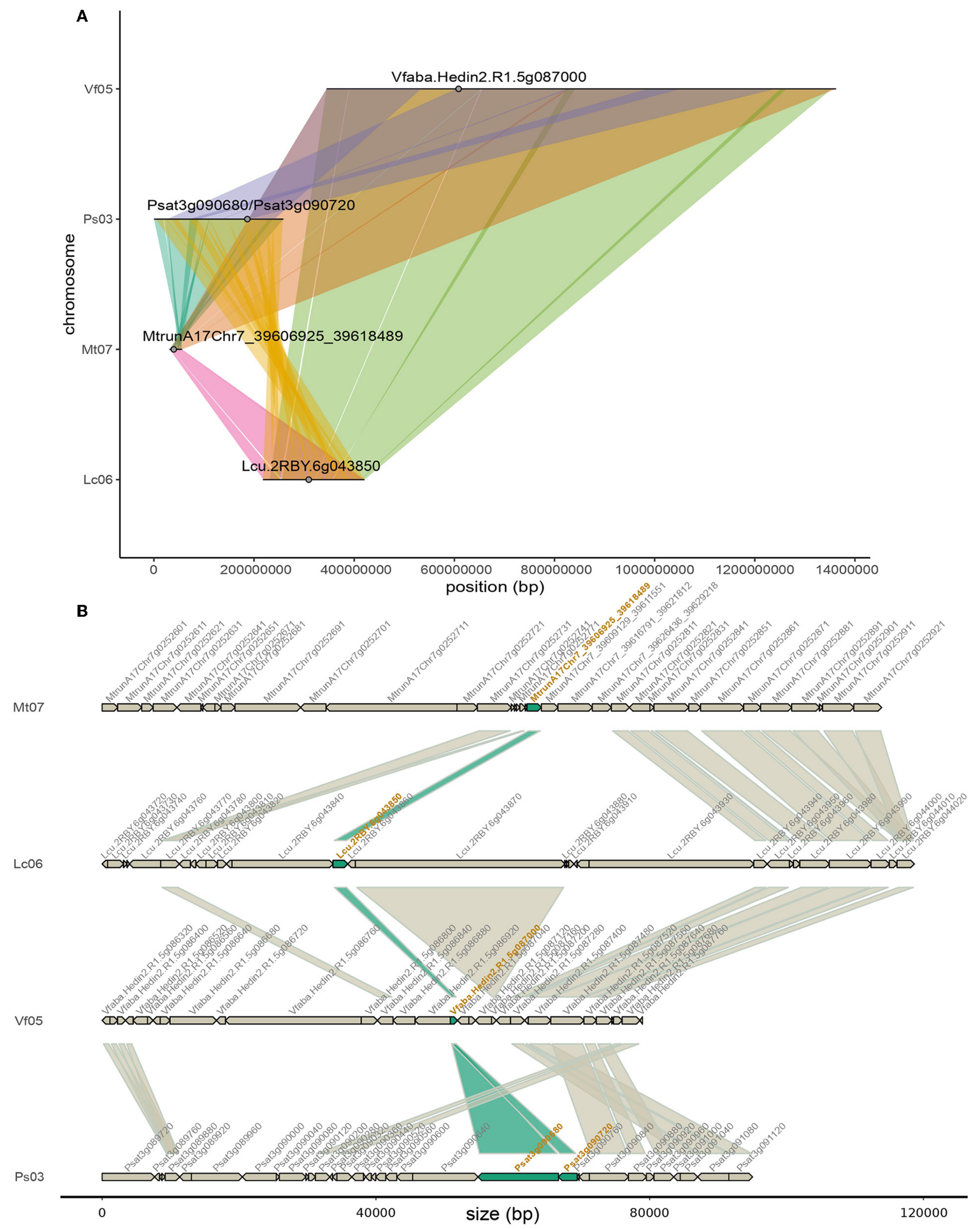
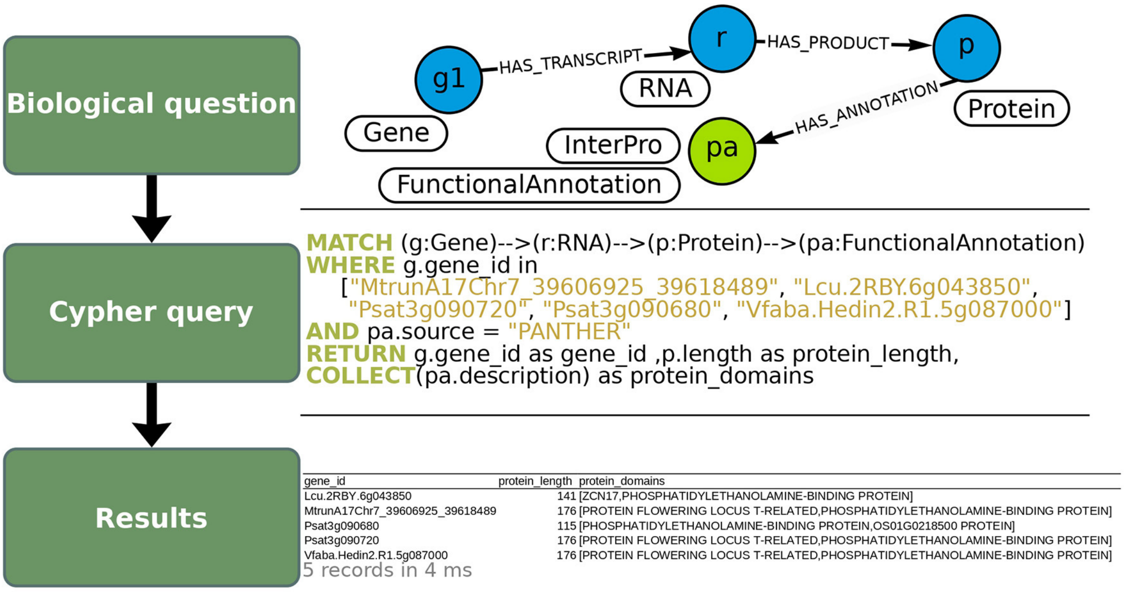


FIGURE 5
Macro- and micro-synteny of the chromosomal regions harboring *FTa1* or its orthologs in *M. truncatula, P. sativum, L. culinaris* and *V. faba*. **(A)** Macro-synteny at the chromosome level. *FTa1* and its orthologs are represented by gray dots on syntenic chromosome sections depicted as lines. Synteny between chromosomes is represented by ribbons. The positions of the two orthologs from *P. sativum* are shown even though they do not belong to any syntenic block in the database. **(B)** Micro-synteny of the *FTa1* loci. Genes are represented with arrows indicating the orientation of the open reading frames. Ribbons connect orthologous gene pairs. The IDs of *FTa1* orthologous genes are in orange and ribbons connecting them are filled in dark green. Since the four species have high genome size heterogeneity and variable intergenic sizes, intergenic regions were removed from the plot. Some gene names are not displayed due to space limitations. However, the gene sizes remain proportional.



```
MATCH (g:Gene)-->(r:RNA)-->(p:Protein)-->(pa:FunctionalAnnotation)
WHERE g.gene_id in
    ["MtrunA17Chr7_39606925_39618489", "Lcu.2RBY.6g043850",
     "Psat3g090720", "Psat3g090680", "Vfaba.Hedin2.R1.5g087000"]
AND pa.source = "PANTHER"
RETURN g.gene_id as gene_id ,p.length as protein_length,
COLLECT(pa.description) as protein_domains
```

```
gene_id                          protein_length protein_domains
Lcu.2RBY.6g043850                           141 [ZCN17,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
MtrunA17Chr7_39606925_39618489              176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
Psat3g090680                                115 [PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN,OS01G0218500 PROTEIN]
Psat3g090720                                176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
Vfaba.Hedin2.R1.5g087000                    176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
5 records in 4 ms
```

FIGURE 6
Extraction of protein domain annotations of *FTa1* and its orthologs using OrthoLegKB. "FunctionalAnnotation" nodes containing protein domain annotations **(top panel)** were queried in Cypher **(middle panel)**, for which several properties were returned in CSV format **(bottom panel)**. The nodes of protein domain annotations are connected to "Protein" nodes. Therefore, proteins corresponding to *FTa1* and its orthologous genes were selected, and their annotations from PANTHER were retrieved. Note that some relationships' names were not displayed in the query section to keep it concise but were specified when running the query. The number of records returned in the output table and the average response time of the query are shown in light gray below the table.

until the first plant has flowered, 50% of the plants have flowered or even 90% of them. Flowering time can also be expressed as the number of days between sowing and flowering or the number of degree-days. Nonetheless, a common vocabulary can be achieved with multi-species ontologies and needs to be developed. Such initiatives exist, such as the BBCH-scale framework to describe the phenological development stages of plants and serialized in RDF (Roussey, 2021), with instances for pea and faba bean but remain under-utilized. In the case of legumes, a higher-level ontology, not restricted to phenological stages, could use existing legume ontologies from the Crop Ontology, including the Lentil Ontology (CO_339) and the Faba bean Ontology (CO_365) (Shrestha et al., 2012). A general, consensus, ontology will however require manual work for the mapping of ontologies and its curation (Oellrich et al., 2015; Laporte et al., 2016; Cooper et al., 2018).

The graph model of Ortho_KB is intended to be regularly updated to enhance querying possibilities

The current version of Ortho_KB includes QTL and expression data but only allow the comparison of species based on single reference genomes. Lately, efforts on pangenomes and on the description of large diversity panels highlighted the importance of considering a wider set of accessions rather than a single representative one. As a first step toward the integration of structural variation, we intend to upgrade the graph model to allow hosting polymorphism variants in Ortho_KB. Since single nucleotide polymorphism (SNP) matrices constitute a large amount of data, the filtering and modeling will have to be thoroughly tested.

At the functional level, Ortho_KB presently provides solely transcriptomic data evidences. To provide complementary evidences regarding the function of genes of interest and their regulation at the post-transcriptional level, we plan to support the integration of proteomic data with Ortho_KB. This addition is also motivated by the ongoing standardization of proteomics output such as the mzTab format and downstream analyses (Griss et al., 2014; Ewels et al., 2020; Deutsch et al., 2022; Dubbelaar et al., 2022). Indeed, a recently published knowledge graph designed for clinical proteomic data namely the Clinical Knowledge Graph (CKG) accepts community-developed formats including mzTab and SDRF for metadata (Santos et al., 2022). Combining -omics layers can bring further evidence to a hypothesis and also open doors toward the understanding of complex underlying phenomena. Since, Ortho_KB was designed to be modular, one could even consider the inclusion of epigenomics information to gain insights on chromatin rearrangement during stress conditions for example. In this case, the integration of non-genic regions such as promoters and enhancers in the database could be evaluated.

Ortho_KB should be constantly evaluated to maintain performance and to facilitate its integration in the current databases ecosystem

As more biological data and data types are included, the Ortho_KB framework will have to be regularly fine-tuned to find

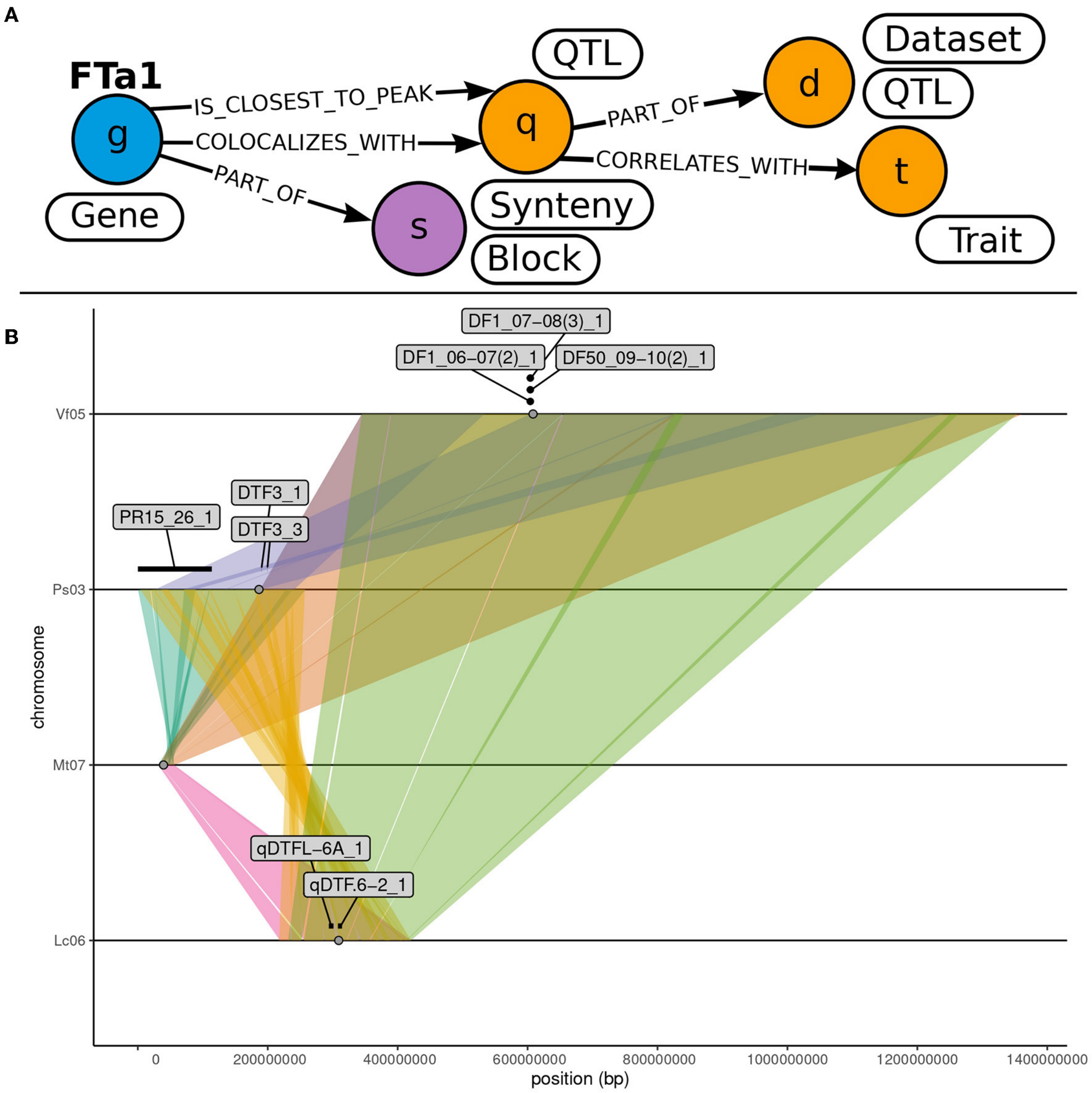
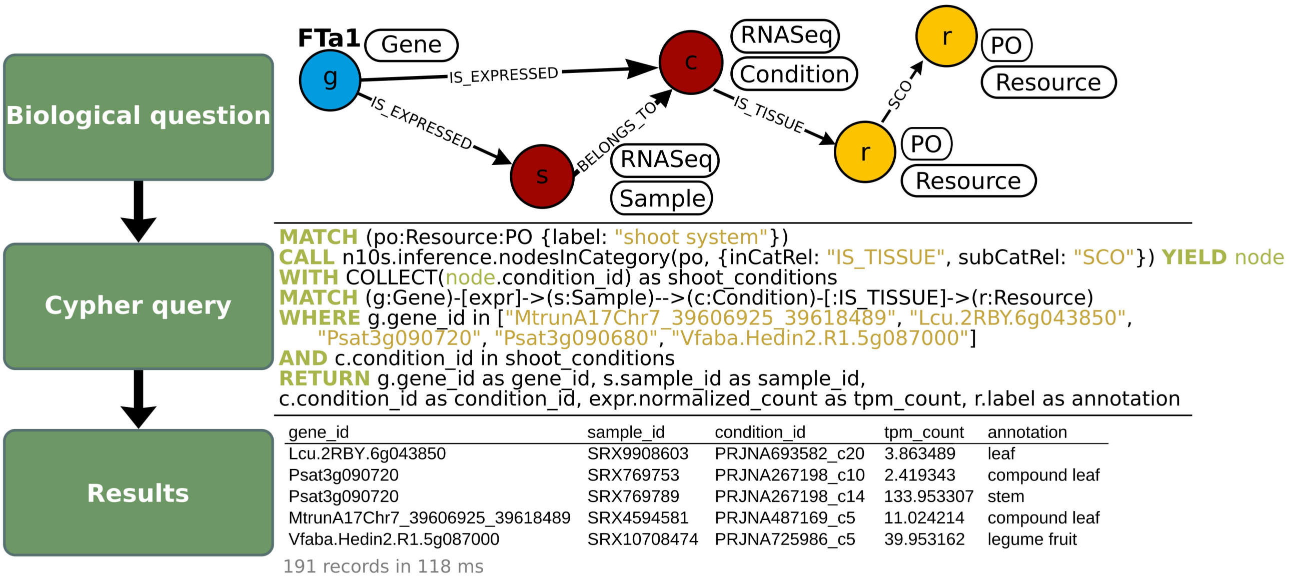


FIGURE 7
Identification of colocalising QTL with syntenic blocks hosting *MtFTa1* and its orthologs. **(A)** Illustration of the subgraph of OrthoLegKB queried to highlight QTL located near *FTa1* genes. "QTL" nodes contained within "Synteny" nodes including the *FTa1* gene were mined. Only QTL associated with flowering "Trait" were then kept. The query is available in Supplementary File S1. **(B)** Visualization of the colocalization between flowering QTL and syntenic blocks containing *FTa1* orthologs. Chromosome sections are represented by lines. Syntenic regions across chromosomes are represented by colored ribbons. *FTa1* and its orthologs are represented by gray dots. QTL labeled with their IDs are depicted by segments when information on both flanking markers is available or otherwise by simple dots.

the optimal graph model, but also in terms of the underlying configuration. In fact, for both the orthology backbone and the additional layers of the graph, single-property indexes have been created on properties that are regularly used as anchors to improve search performance at a small cost in storage space. Further guidance on the configuration of Neo4j has been previously published and will help to ensure high efficiency and scalability of Ortho_KB (Yoon et al., 2017). Several platforms already exist to study comparative genomics (Lyons and Freeling, 2008; Van Bel et al., 2022). The goal of Ortho_KB is different, since it mainly uses orthology and synteny as a way to transfer curated knowledge across species. Therefore, any created instance can be queried freely to answer complex tailored questions in a comprehensive manner.

As OrthoLegKB is primarily populated with published datasets, interoperability with already existing databases is essential. For RNA-seq, the NCBI Sequence Read Archive stores datasets according to defined rigorous standards (NCBI, 2023). QTL data, on the other hand, are typically scattered across multiple databases that store the information in different formats. Unlike the GWAS Catalog available for human (Sollis et al., 2023), no integrative databases store legumes QTL data in a unified format. Therefore, we plan to facilitate the integration of the content from existing legume databases. Other knowledge graph to understand the



```
MATCH (po:Resource:PO {label: "shoot system"})
CALL n10s.inference.nodesInCategory(po, {inCatRel: "IS_TISSUE", subCatRel: "SCO"}) YIELD node
WITH COLLECT(node.condition_id) as shoot_conditions
MATCH (g:Gene)-[expr]->(s:Sample)-->(c:Condition)-[:IS_TISSUE]->(r:Resource)
WHERE g.gene_id in ["MtrunA17Chr7_39606925_39618489", "Lcu.2RBY.6g043850",
    "Psat3g090720", "Psat3g090680", "Vfaba.Hedin2.R1.5g087000"]
AND c.condition_id in shoot_conditions
RETURN g.gene_id as gene_id, s.sample_id as sample_id,
c.condition_id as condition_id, expr.normalized_count as tpm_count, r.label as annotation
```

gene_id	sample_id	condition_id	tpm_count	annotation
Lcu.2RBY.6g043850	SRX9908603	PRJNA693582_c20	3.863489	leaf
Psat3g090720	SRX769753	PRJNA267198_c10	2.419343	compound leaf
Psat3g090720	SRX769789	PRJNA267198_c14	133.953307	stem
MtrunA17Chr7_39606925_39618489	SRX4594581	PRJNA487169_c5	11.024214	compound leaf
Vfaba.Hedin2.R1.5g087000	SRX10708474	PRJNA725986_c5	39.953162	legume fruit

191 records in 118 ms

FIGURE 8
Fetch of expression levels (in TPM) of *MtFTa1* and its orthologs in different tissues of the shoot system, in OrthoLegKB. Normalized expression level of *FTa1* and its orthologs in RNA-seq samples **(top panel)** were queried in Cypher **(middle panel)**, and several properties were returned in CSV format **(bottom panel)**. The expression of *FTa1* genes was queried at the "Sample" level using the "expr" variable. The tissue annotations from the "Condition" nodes connected to these "Samples" were filtered to have only "Condition" nodes connected to the "PO" node "shoot system" or any of its more specific child terms. Note that in the table, the order of the rows was rearranged to show the diversity of annotations under the "shoot system" term. The number of records returned in the output table and the average response time of the query are shown in light gray below the table. The full table is available as Supplementary Table S6.

role of genes are already available. The KnetMiner software was created to analyse genome-scale knowledge graphs, with a recent support for the Cypher graph query language (Hassani-Pak et al., 2021). This platform allows to build gene networks based on semantics and information primarily extracted from the literature, including genetic data, phenotypes associated to SNPs or biological pathways. It was recently applied to wheat, generating networks for the *TT2* gene involved in pre-harvest sprouting (Hassani-Pak et al., 2021). In the specific case of legumes, the AgroLD triplestore is to our knowledge the only phenomics agronomy-centered database aiming at an integrative storage of biological information in the form of a knowledge graph (Venkatesan et al., 2018; Larmande and Todorov, 2021). Since Neo4j can handle RDF import and export, thanks to the neosemantics plugin, data exchange between OrthoLegKB and AgroLD could be considered to take advantage of both technologies. This goal is further supported by the ongoing development of the RDF-star extension which could support properties on edges of the graph (Abuoda et al., 2022). This would bridge the gap between LPG and RDF technologies for improved interoperability (Hartig, 2014). While the SPARQL RDF query language is common to all triplestores, Cypher from Neo4j is only used by the proprietary. However, the open-source GraphQL initiative known as GQL is seen as a potential technology agnostic standardization query language for graph databases (Donkers et al., 2020). We envisage that the legume research community will participate in the data collection and provide feedback on OrthoLegKB for regular improvement.

Ortho_KB offered an opportunity to develop a valuable tool for translational research in legumes, OrthoLegKB

We decided to select legume species to showcase how the Ortho_KB framework can serve translational research. OrthoLegKB is currently centered on few members mostly diploid cool-season legumes as the identification of orthologs is more straightforward than in polyploid species. Still, having a high-quality assembly is crucial for synteny detection and therefore true orthologs identification. The *FTa* locus in *P. sativum* (*PsFTa*) is in fact incorrectly assembled and annotated in the current version of the Cameor genome. While the *FTa1* gene we identified in *P. sativum* (*Psat3g090720*) is consistent with results from Hecht et al. (2011), the other copy (*Psat3g090680*) was reported as *FTa2* in the same study, both genes displaying similar expression patterns in leaves and apices, but with a weaker expression for *FTa2* (Hecht et al., 2011). In our study, the incomplete annotation of *Psat3g090680* in the version 1 assembly of *P. sativum* cv. Cameor most-likely prevented the creation of orthogroups correctly encompassing the *FT* gene family, and the subsequent inclusion of *PsFTa* in syntenic blocks. The locus displayed increased synteny with the other studied species, this time including *PsFTa1* and *PsFTa2* when considering the more recent *P. sativum* genome assembly from the Zw6 accession (Yang et al., 2022). A new assembly of the Cameor genome is expected soon and should improve the assembly and annotation of this region. Furthermore, supplementing OrthoLegKB with transcriptomic data will provide stronger support when searching for *FT* orthologous genes, by comparing their expression profiles. Fortunately, more legume genomes, have been lately assembled in high-quality using high-throughput chromosome conformation capture sequencing or long-read technologies, namely chickpea (Garg et al., 2022) or common vetch (Xi et al., 2022), which might reveal to be novel sources of data for OrthoLegKB. Thus, the graph will encompass more connected datasets, including information on abiotic and biotic stress response and be useful to a larger part of the legume research community.

The Ortho_KB framework is for the plant community and beyond

As demonstrated for legumes, the Ortho_KB framework is suitable for translational research within plant families to address common biological questions. Therefore, Ortho_KB could for instance be used in Solanaceae to study late blight attacking potato, tomato but not eggplants nor pepper. Genomes were sequenced for all these diploid species with long-reads technologies (Pham et al., 2020; Wei et al., 2020; Su et al., 2021; Liao et al., 2022). With more caution regarding the identification of orthologs, this resource would also meet the needs of research across plant families, or the needs of polyploids in the Brassicaceae and Poaceae families. Precise study of gene expression bias would be then crucial to identify expressologs (Das et al., 2016). While the scope of Ortho_KB was limited to plants for annotation reasons, its concept could be adapted for the benefit of other communities. For example, the wealth of draft-assembled diploid genomes profiting the Chelicerates community was recently exploited to highlight the conservation of chemosensory genes through comparative genomics (Vizueta et al., 2018). As more qualitative assemblies and associated -omics data are generated across plant and animal groups, we can only anticipate that the need for integrative multi-species databases will increase and Ortho_KB can contribute in this regard.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found below in the article/Supplementary material. OrthoLegKB with its user-guide is available for legume translational research at http://ortholegkb.versailles.inrae.fr/browser/. The functional pipeline for synteny is available at: https://forgemia.inra.fr/geapsi/pipeline/specifics_syntenymcscanx. The pipeline to create the translational database is available at: https://forgemia.inra.fr/geapsi/pipeline/specifics_ortho_kb. All scripts used to create the figures presented, including microsynteny from OrthoLegKB data, are available at: https://forgemia.inra.fr/geapsi/ecp-paper/ortholegkb_data.

Author contributions

BI, JK, and NT contributed to the conception and design of this work. BI was responsible for developing the pipelines, running the use-case, and writing the manuscript. JK and R-GF helped BI to build the graph database and provided bioinformatic support. GA collected information on the datasets to be included in OrthoLegKB and participated in the development of the use-case. JB provided ideas and managed funding acquisition. NT contributed to the scientific management of this work and was involved in drafting and writing the manuscript. All authors participated to manuscript revision, read, and approved the submitted version.

Funding

BI is the recipient of a PhD grant from French Ministry of Higher Education and Research. The work was part of the SPECIFICS project (ANR-20-PCPA-0008) funded by the French Priority Research Programme Growing and Protecting crops Differently (PPR-CPA), part of the national investment plan operated by the French National Research Agency (ANR).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1191122/full#supplementary-material

References

Abuoda, G., Dell'Aglio, D., Keen, A., and Hose, K. (2022). Transforming RDF-star to property graphs: A preliminary analysis of transformation approaches – extended version. *arXiv [Preprint]*. arXiv: 2210.05781. doi: 10.48550/arXiv.2210.05781

Aguilar-Benitez, D., Casimiro-Soriguer, I., Maalouf, F., and Torres, A. M. (2021). Linkage mapping and QTL analysis of flowering time in faba bean. *Sci. Rep.* 11, 13716. doi: 10.1038/s41598-021-92680-4

Bandi, V., and Gutwin, C. (2020). *Interactive Exploration of Genomic Conservation in Proceedings of Graphics Interface 2020 GI 2020*. Toronto: Canadian Human-Computer Communications Society/Société canadienne du dialogue humain-machine, 74–83.

Barrasa, J. (2022). *Neosemantics (n10s)*. Available online at: https://github.com/neo4j-labs/neosemantics (accessed December 21, 2022).

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome: tair: making and mining the "gold standard" plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Berendzen, J., Brown, A. V., Cameron, C. T., Campbell, J. D., Cleary, A. M., Dash, S., et al. (2021). The legume information system and associated online genomic resources. *Legume Sci.* 3, 4. doi: 10.1002/leg3.74

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., and Kandasaamy, S., Mitchell, A., et al. (2021). The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., and Vandepoele, K. (2021). TRAPID 2.0: a web application for taxonomic and functional analysis of *de novo* transcriptomes. *Nucleic Acids Res.* 49, e101–e101. doi: 10.1093/nar/gkab565

Buchfink, B., Reuter, K., and Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Carrere, S., Verdier, J., and Gamas, P. (2021). MtExpress, a comprehensive and curated RNASEQ-based gene expression atlas for the model legume *Medicago truncatula*. *Plant Cell Physiol.* 62, 1494–1500. doi: 10.1093/pcp/pcab110

Cheng, X., Li, G., Krom, N., Tang, Y., and Wen, J. (2021). Genetic regulation of flowering time and inflorescence architecture by MtFDa and MtFTa1 in Medicago truncatula. *Plant Physiol.* 185, 18. doi: 10.1093/plphys/kiaa005

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinf.* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., and Mungall, C., Sinn, B. T., et al. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.* 46, D1168–D1180. doi: 10.1093/nar/gkx1152

Dai, X., Zhuang, Z., Boschiero, C., Dong, Y., and Zhao, P. X. (2021). LegumeIP V3: from models to crops—an integrative gene discovery platform for translational genomics in legumes. *Nucleic Acids Res.* 49, D1472–D1479. doi: 10.1093/nar/gkaa976

Dainat, J., Hereñú, D., Davis, E., Crouch, K., LucileSol, F., Agostinho, N., et al. (2022). NBISweden/AGAT: AGAT-v1.0.0. *Zenodo*. doi: 10.5281/zenodo.7255559

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27, 3423–3424. doi: 10.1093/bioinformatics/btr539

Das, M., Haberer, G., Panda, A., Das Laha, S., Ghosh, T. C., Schäffner, A. R., et al. (2016). Expression pattern similarities support the prediction of orthologs retaining common functions after gene duplication events. *Plant Physiol.* 171, 2343–2357. doi: 10.1104/pp.15.01207

Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L., et al. (2022). The proteomexchange consortium at 10 years: 2023 update. *Nucleic Acids Res.* 5, gkac1040. doi: 10.1093/nar/gkac1040

Di Tommaso, D., Chatzou, P., Floden, M., Barja, E. W., and Palumbo, P. P. E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Donkers, A., Yang, D., and Baken, N. (2020). "Linked data for smart homes: comparing RDF and labeled property graphs," in *LDAC*. Available online at: https://ceur-ws.org/Vol-2636/02paper.pdf

Drillon, G., Champeimont, R., Oteri, F., Fischer, G., and Carbone, A. (2020). Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol. Biol. Evol.* 37, 2747–2762. doi: 10.1093/molbev/msaa114

Dubbelaar, M., Leon-Bichmann, Heumos, L., Peltzer, A., bot, nf-core, Scheid, J., et al. (2022). *nf-core/mhcquant: mhcquant 2.4.0 – Maroon Gold Boxer. Zenodo.* doi: 10.5281/zenodo.7389537

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi: 10.1038/s41587-020-0439-x

Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., et al. (2018). Reactome graph database: efficient access to complex pathway data. *PLoS Comput. Biol.* 14, e1005968. doi: 10.1371/journal.pcbi.1005968

Gali, K. K., Liu, Y., Sindhu, A., Diapari, M., Shunmugam, A. S. K., Arganosa, G., et al. (2018). Construction of high-density linkage maps for mapping quantitative trait loci for multiple traits in field pea (Pisum sativum L.). *BMC Plant Biol.* 18, 172. doi: 10.1186/s12870-018-1368-4

Garg, V., Dudchenko, O., Wang, J., Khan, A. W., Gupta, S., Kaur, P., et al. (2022). Chromosome-length genome assemblies of six legume species provide insights into genome organization, evolution, and agronomic traits for crop improvement. *J. Adv. Res.* 42, 315–329. doi: 10.1016/j.jare.2021.10.009

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Grimnes, G. A., Higgins, G., Hees, J., Aucamp, I., Arndt, N., Sommer, A., et al. (2023). RDFLib/rdflib: RDFlib 6.3.1. *Zenodo*. doi: 10.5281/zenodo.7748890

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., et al. (2014). The mzTab Data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Prot.* 13, 2765–2775. doi: 10.1074/mcp.O113.036681

Grover, J. W., Bomhoff, M., Davey, S., Gregory, B. D., Mosher, R. A., Lyons, E., et al. (2017). CoGe LoadExp+: a web-based suite that integrates next-generation sequencing data analysis workflows and visualization. *Plant Direct* 1, 8. doi: 10.1002/pld3.8

Guhlin, J., Silverstein, K. A. T., Zhou, P., Tiffin, P., and Young, N. D. (2017). ODG: Omics database generator - a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. *BMC Bioinformatics* 18, 367. doi: 10.1186/s12859-017-1777-7

Guiguitant, J., Vile, D., Ghanem, M. E., Wery, J., and Marrou, H. (2020). Evaluation of pulse crops' functional diversity supporting food production. *Sci. Rep.* 10, 3416. doi: 10.1038/s41598-020-60166-4

Ha, J., Satyawan, D., Jeong, H., Lee, E., Cho, K., Kim, M. Y., et al. (2021). *A near-complete genome sequence of mungbean (Vigna radiata L.) provides key insights into the modern breeding program. Plant Genome.* 10, 121. doi: 10.1002/tpg2.20121

Hackl, T., and Ankenbrand, M. (2023). *Gggenomes: A Grammar of Graphics for Comparative Genomics*. Available online at: https://github.com/thackl/gggenomes (accessed March 20, 2023).

Haile, T. A., Stonehouse, R., Weller, J. L., and Bett, K. E. (2021). Genetic basis for lentil adaptation to summer cropping in northern temperate environments. *Plant Genome* 14, 144. doi: 10.1002/tpg2.20144

Hartig, O. (2014). *Reconciliation of RDF* and Property Graphs*. Available online at: http://arxiv.org/abs/1409.3288 (accessed March 13, 2023).

Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., et al. (2021). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol. J.* 19, 1670–1678. doi: 10.1111/pbi.13583

Hecht, V., Knowles, C. L., Vander Schoor, J. K., Liew, L. C., Jones, S. E., Lambert, M. J. M., et al. (2007). Pea LATE BLOOMER1 Is a GIGANTEA ortholog with roles in photoperiodic flowering, deetiolation, and transcriptional regulation of circadian clock gene homologs. *Plant Physiol.* 144, 648–661. doi: 10.1104/pp.107.096818

Hecht, V., Laurie, R. E., Vander Schoor, J. K., Ridge, S., Knowles, C. L., Liew, L. C., et al. (2011). The Pea *GIGAS* gene is a flowering locus t homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. *The Plant Cell* 23, 147–161. doi: 10.1105/tpc.110.081042

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Humann, J., Jung, S., Cheng, C. H., Lee, T., and Zheng, P., Frank, M., et al. (2019). *A resource for pea, lentil, faba bean, and chickpea genetics, genomics and breeding. Proceedings of the International Plant and Animal Genome Conference*, 3. Available online at: https://www.pulsedb.org/

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednár, P., et al. (2023). The giant diploid faba genome unlocks variation in a global protein crop. *Nature* 26, 1–8. doi: 10.1038/s41586-023-05791-5

Jones, P., Binns, D., Chang, H. Y., Fraser, M., and Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kamei, C. L. A., Severing, E. I., Dechesne, A., Furrer, H., Dolstra, O., Trindade, L. M., et al. (2016). Orphan crops browser: a bridge between model and orphan crops. *Mol. Breeding* 36, 9. doi: 10.1007/s11032-015-0430-2

Kans, J. (2013). *Entrez Direct: E-utilities on the Unix Command Line. National Center for Biotechnology Information (US)*. Available online at: https://www.ncbi.nlm.nih.gov/books/NBK179288/ (accessed March 20, 2023).

Kaur, P., Singh, A., and Chana, I. (2021). Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Arch Computat. Methods Eng.* 28, 4595–4631. doi: 10.1007/s11831-021-09547-0

Khayatbashi, S., Ferrada, S., and Hartig, O. (2022). "Converting property graphs to RDF: a preliminary study of the practical impact of different mappings," in *Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) GRADES-NDA '22* (New York, NY: Association for Computing Machinery), 1–9. doi: 10.1145/3534540.3534695

Kreplak, J., Madoui, M. A., Cápal, P., Novák, P., and Labadie, K., Aubert, G., et al. (2019). A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* 51, 1411–1422. doi: 10.1038/s41588-019-0480-1

Krishnakumar, V., Kim, M., Rosen, B. D., Karamycheva, S., Bidwell, S. L., Tang, H., et al. (2015). MTGD: the medicago truncatula genome database. *Plant Cell Physiol.* 56, e1. doi: 10.1093/pcp/pcu179

Laporte, M. -A., Valette, L., Arnaud, E., Cooper, L., Meier, A., Jaiswal, P., et al. (2016). *Comparison of Ontology Mapping Techniques to Map Plant Trait Ontologies*. Corvallis, OR: CEUR Workshop Proceedings. Available online at: https://ceur-ws.org/Vol-1747/IP17_ICBO2016.pdf

Larmande, P., and Todorov, K. (2021). "AgroLD: A knowledge graph for the plant sciences," in *Semantic Web - ISWC 2021 Lecture Notes in Computer Science*, eds A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, and P. Barnaghi (Cham: Springer International Publishing), 496–510. doi: 10.1007/978-3-030-88361-4_29

Laurie, R. E., Diwadkar, P., Jaudal, M., Zhang, L., Hecht, V., Wen, J., et al. (2011). The Medicago flowering locus T Homolog, MtFTa1, Is a Key Regulator of Flowering *Time*. 156, 18. doi: 10.1104/pp.111.180182

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., et al. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* 40, D465–471. doi: 10.1093/nar/gkr1181

Li, J., Dai, X., Liu, T., and Zhao, P. X. (2012). LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.* 40, D1221–1229. doi: 10.1093/nar/gkr939

Li, J., Dai, X., Zhuang, Z., and Zhao, P. X. (2016). LegumeIP 2.0–a platform for the study of gene function and genome evolution in legumes. *Nucleic Acids Res.* 44, D1189–1194. doi: 10.1093/nar/gkv1237

Liao, Y., Wang, J., Zhu, Z., Liu, Y., Chen, J., Zhou, Y., et al. (2022). The 3D architecture of the pepper genome and its relationship to function and evolution. *Nat. Commun.* 13, 3479. doi: 10.1038/s41467-022-31112-x

Linard, B., Ebersberger, I., McGlynn, S. E., Glover, N., Mochizuki, T., Patricio, M., et al. (2021). Ten years of collaborative progress in the quest for orthologs. *Mol. Biol. Evol.* 38, 3033–3045. doi: 10.1093/molbev/msab098

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data: Mercator: sequence functional annotation server. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant J.* 53, 661–673. doi: 10.1111/j.1365-313X.2007.03326.x

Mi, H., and Thomas, P. (2009). PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* 563, 123–140. doi: 10.1007/978-1-60761-175-2_7

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi: 10.1530/JME-18-0055

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Naithani, S., Gupta, P., Preece, J., D'Eustachio, P., Elser, J. L., Garg, P., et al. (2019). Plant Reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* 47, gkz996. doi: 10.1093/nar/gkz996

NCBI (2023). SRA Metadata and Submission Overview. Available online at: https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/ (accessed March 20, 2023).

Neo4j (2023a). Arrows. *Neo4j Graph Data Platform*. Available online at: https://neo4j.com/labs/arrows/ (accessed March 6, 2023).

Neo4j (2023b). *The Neo4j Graph Data Platform. Neo4j Graph Data Platform*. Available online at: https://neo4j.com/product/ (accessed January 30, 2023).

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods* 11, 10. doi: 10.1186/s13007-015-0053-y

Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., et al. (2006). The rice annotation project database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic Acids Res.* 34, D741–D744. doi: 10.1093/nar/gkj094

Patel, H., Beber, M. E., Han, D. W., Philips, E., Manning, J., Yates, J. A. F., et al. (2022). nf-core/fetchngs: nf-core/fetchngs v1.9 - Plutonium Prancer. *Zenodo*. doi: 10.5281/zenodo.7468050

Patel, H., Ewels, P., Peltzer, A., Botvinnik, O., Sturm, G., Moreno, D., et al. (2023). nf-core/rnaseq: nf-core/rnaseq v3.10.1 – Plastered Rhodium Rudolph. *Zenodo*. doi: 10.5281/zenodo.7505987

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., et al. (2018). Whole-genome landscape of Medicago truncatula symbiotic genes. *Nature Plants* 4, 1017–1025. doi: 10.1038/s41477-018-0286-7

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., et al. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* 9, giaa100. doi: 10.1093/gigascience/giaa100

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Raciti, D., Yook, K., Harris, T. W., Schedl, T., and Sternberg, P. W. (2018). *Micropublication* : incentivizing community curation and placing unpublished data into the public domain. *Database* 2018., e013 doi: 10.1093/database/bay013

Ramsay, L., Koh, C. S., Kagale, S., Gao, D., Kaur, S., Haile, T., et al. (2021). *Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. Plant Biol.* 24, 237. doi: 10.1101/2021.07.23.453237

Roussey, C. (2021). *BBCH-based Plant Phenological Description Ontology*. doi: 10.15454/TIMQHW

Rubiales, D., Annicchiarico, P., Vaz Patto, M. C., and Julier, B. (2021). Legume breeding for the agroecological transition of global agri-food systems: a european perspective. *Front. Plant Sci.* 12, 782574. doi: 10.3389/fpls.2021.782574

Sanderson, L. A., Caron, C. T., Tan, R., Shen, Y., and Liu, R., Bett, K. E., et al. (2019). KnowPulse: a web-resource focused on diversity data for pulse crop improvement. *Front. Plant Sci.* 10, 965. doi: 10.3389/fpls.2019.00965

Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Strauss, M., Geyer, P. E., et al. (2022). *A knowledge graph to interpret clinical proteomics data. Nat Biotechnol.* 40, 692–702. doi: 10.1038/s41587-021-01145-6

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., et al. (2008). Genome structure of the legume, lotus japonicus. *DNA Res.* 15, 227–239. doi: 10.1093/dnares/dsn008

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., et al. (2019). MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* 12, 879–892. doi: 10.1016/j.molp.2019.01.003

Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K., and Bloem, M. W. (2021). Legumes as a sustainable source of protein in human diets. *Global Food Security* 28, 100520. doi: 10.1016/j.gfs.2021.100520

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for fastA/Q file manipulation. *PLoS ONE* 11, e0163962. doi: 10.1371/journal.pone.0163962

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345–353. doi: 10.1038/nature24286

Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., et al. (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front. Physio.* 3, 326. doi: 10.3389/fphys.2012.00326

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985. doi: 10.1093/nar/gkac1010

Soneson, C., Love, M. I., and Robinson, M. D. (2016). *Differential analyses for* RNA-seq: transcript-level estimates improve gene-level inferences. doi: 10.12688/f1000research.7563.2

Stevens, I., Mukarram, A. K., Hörtenhuber, M., Meehan, T. F., Rung, J., Daub, C. O., et al. (2020). Ten simple rules for annotating sequencing experiments. *PLOS Computat. Biol.* 16, e1008260. doi: 10.1371/journal.pcbi.1008260

Su, X., Wang, B., Geng, X., Du, Y., Yang, Q., Liang, B., et al. (2021). A high-continuity and annotated tomato reference genome. *BMC Genomics* 22, 898. doi: 10.1186/s12864-021-08212-x

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., Paterson, A. H., et al. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tello-Ruiz, M. K., Naithani, S., Gupta, P., Olson, A., Wei, S., Preece, J., et al. (2021). Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.* 49, D1452–D1463. doi: 10.1093/nar/gkaa979

The Legume Phylogeny Working Group, Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., Kenicer, G., et al. (2013). Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species–rich clades. *TAXON* 62, 217–248. doi: 10.12705/622.8

Thomson, G., Taylor, J., and Putterill, J. (2019). The transcriptomic response to a short day to long day shift in leaves of the reference legume *Medicago truncatula*. *PeerJ* 7, e6626. doi: 10.7717/peerj.6626

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., et al. (2012). Dissecting plant genomes with the plaza comparative genomics platform. *Plant Physiol.* 158, 590–600. doi: 10.1104/pp.111.189514

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acid. Res.* 50, D1468–D1474. doi: 10.1093/nar/gkab1024

Venkatesan, A., Tagny Ngompe, G., Hassouni, N. E., Chentli, I., Guignon, V., Jonquet, C., et al. (2018). Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. *PLoS ONE* 13, e0198270. doi: 10.1371/journal.pone.0198270

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010). "A comparison of a graph database and a relational database: a data provenance perspective," in *Proceedings of the 48th Annual Southeast Regional Conference on - ACM SE '10* (Oxford, MS: ACM Press). doi: 10.1145/1900008.1900067

Vizueta, J., Rozas, J., and Sánchez-Gracia, A. (2018). Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biol. Evol.* 10, 1221–1236. doi: 10.1093/gbe/evy081

W3C (1994). Available online at: https://www.w3.org/ (accessed February 15, 2023).

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wei, Q., Wang, J., Wang, W., Hu, T., Hu, H., Bao, C., et al. (2020). A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic Res.* 7, 153. doi: 10.1038/s41438-020-00391-0

Wilkinson, M. D., and Dumontier, M., Aalbersberg, I.j., J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. doi: 10.1038/sdata.2016.18

Williams, O., Vander Schoor, J. K., Butler, J. B., Ridge, S., Sussmilch, F. C., Hecht, V. F. G., et al. (2022). The genetic architecture of flowering time changes in pea from wild to crop. *J. Exp. Bot.* 73, 3978–3990. doi: 10.1093/jxb/erac132

Xi, H., Nguyen, V., Ward, C., Liu, Z., and Searle, I. R. (2022). Chromosome-level assembly of the common vetch (Vicia sativa) reference genome. *Gigabyte* 2022, 1–20. doi: 10.46471/gigabyte.38

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., et al. (2022). *Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics*. *Nat Genet.* 10, 1553–1563. doi: 10.1038/s41588-022-01172-2

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., et al. (2022). Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* 50, D996–D1003. doi: 10.1093/nar/gkab1007

Ye, C. -Y., and Fan, L. (2021). Orphan crops and their wild relatives in the genomic era. *Mol. Plant* 14, 27–39. doi: 10.1016/j.molp.2020.12.013

Yoon, B. H., Kim, S-, K., and Kim, S. Y. (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics Inform* 15, 19. doi: 10.5808/GI.2017.15.1.19

Yuan, H. Y., Caron, C. T., Ramsay, L., and Fratini, R., de la Vega, M. P., Vandenberg, A., et al. (2021). Genetic and gene expression analysis of flowering time regulation by light quality in lentil. *Annal. Bot.* 128, 481–496. doi: 10.1093/aob/mcab083