Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2021), pp. 1–9 doi: 10.1111/pbi.13583

KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species

Keywan Hassani-Pak[1,*], Ajit Singh[1], Marco Brandizi[1], Joseph Hearnshaw[1], Jeremy D. Parsons[1], Sandeep Amberkar[1], Andrew L. Phillips[1], John H. Doonan[2] and Chris Rawlings[1]

[1]*Rothamsted Research, Harpenden, UK*
[2]*IBERS, Aberystwyth University, Aberystwyth, UK*

Received 30 July 2020;
revised 17 December 2020;
accepted 16 March 2021.
**Correspondence* (Tel +44 (0)75 905 104 13; email keywan.hassani-pak@rothamsted.ac.uk)

Keywords: gene discovery, gene network, knowledge graph, knowledge discovery, exploratory data mining, data integration, candidate gene prioritization, information visualization, systems biology, bioinformatics.

Abstract

The generation of new ideas and scientific hypotheses is often the result of extensive literature and database searches, but, with the growing wealth of public and private knowledge, the process of searching diverse and interconnected data to generate new insights into genes, gene networks, traits and diseases is becoming both more complex and more time-consuming. To guide this technically challenging data integration task and to make gene discovery and hypotheses generation easier for researchers, we have developed a comprehensive software package called KnetMiner which is open-source and containerized for easy use. KnetMiner is an integrated, intelligent, interactive gene and gene network discovery platform that supports scientists explore and understand the biological stories of complex traits and diseases across species. It features fast algorithms for generating rich interactive gene networks and prioritizing candidate genes based on knowledge mining approaches. KnetMiner is used in many plant science institutions and has been adopted by several plant breeding organizations to accelerate gene discovery. The software is generic and customizable and can therefore be readily applied to new species and data types; for example, it has been applied to pest insects and fungal pathogens; and most recently repurposed to support COVID-19 research. Here, we give an overview of the main approaches behind KnetMiner and we report plant-centric case studies for identifying genes, gene networks and trait relationships in *Triticum aestivum* (bread wheat), as well as, an evidence-based approach to rank candidate genes under a large *Arabidopsis thaliana* QTL. KnetMiner is available at: https://knetminer.org.

Introduction

Genomics is undergoing a revolution. Unprecedented amounts of data are being generated to gain deeper insight into the complex nature of many traits and diseases (Boyle *et al.,* 2017; Stephens *et al.,* 2015). Paradoxically, the vast growing landscape of diverse and interconnected data can often hinder scientists from translating complex and sometimes contradictory information into biological understanding and discoveries. Searching for relevant information amongst larger and more complex data can take longer and so risks information being overlooked or subjective biases being introduced. Even after the gathered information is complete, it can be demanding to assemble a coherent view of how each piece of evidence might come together to 'thread a story' about the biology that can explain how genes and gene networks might be implicated in a complex trait or disease. New tools are needed to provide scientists with a more fine-grained and connected view of the scientific literature and databases, rather than the conventional information retrieval tools currently at their disposal.

Scientists are not alone facing these challenges: knowledge searches form a core part of the duties of many professions. Studies have highlighted the necessity but significant challenge for search systems to give feedback and generate confidence, explainability and accountability (Russell-Rose *et al.,* 2018). Search duration also influences human choice about whether to continue the task (Sweis *et al.,* 2018). When implemented well, search systems can accelerate research by cutting both the time and the cost of reviewing genes, traits and molecules of interest before initiating expensive experiments. Additionally, search systems offer a framework for the prioritization of future research, which can highlight gaps in knowledge.

Knowledge graphs (KG) are increasingly used to make search and information discovery more efficient (Fensel *et al.,* 2020). KGs are contributing to various Artificial Intelligence (AI) applications including link prediction, node classification, and both recommendation and question answering systems (Ali *et al.,*n.d.; Sheth *et al.,* 2019). KGs model heterogeneous knowledge domains by integrating information into advanced unified data schemas (i.e. ontologies) and leverage that to apply formal and

Please cite this article as: Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., Phillips, A. L., Doonan, J. H. and Rawlings, C. (2021) KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol J.*, https://doi.org/10.1111/pbi.13583

statistical inference methods to derive new knowledge (Ehrlinger and Wöß, 2016). Compared to more traditional data models, knowledge graphs are very flexible at integrating and searching connected heterogeneous data, where data schemas are not established a-priori (Yoon *et al.,* 2017), and often subject to frequent changes. KGs in various forms have been widely adopted in many disciplines, ranging from social sciences to engineering, physics, computer science, design and manufacturing. Many research laboratories, like us, are building biological KGs aimed at supporting crop improvement (Hassani-Pak *et al.,* 2016; Xiaoxue *et al.,* 2019), drug target discovery (Mohamed *et al.,* 2019), disease gene prioritization (Alshahrani and Hoehndorf, 2018; Messina *et al.,* 2018) and COVID-19 research (Reese *et al.,* 2021).

The integrated, semi-structured and machine readable nature of KGs provides an ideal basis for the development of sophisticated knowledge discovery and data mining (KDD) applications (Holmes, 2014; Sacchi and Holmes, 2016). Exploratory data mining (EDM), a sub-discipline of knowledge discovery, requires an extensive exploration stage, using both intelligent and intuitive techniques, before predictive modelling and confirmatory analysis can realistically and usefully be applied (De Bie, 2013; De Bie and Spyropoulou, 2013). Furthermore, it is considered important to include the end user into the 'interactive' knowledge discovery process with the goal of supporting human intelligence with artificial intelligence (Holzinger and Jurisica, 2014). Several reports have described the benefits attained by leveraging the unique human cognitive capabilities we have, both within the fields of pattern recognition and higher-order reasoning, to interpret complex biological data and help extract biologically meaningful interpretations (Isenberg *et al.,* 2013; Lee *et al.,* 2012). Visualizing biological information in a concise format and user-centred design can help achieve this (Fox and Hendler, 2011; Pavelin *et al.,* 2012).

There are, however, a few important research challenges that need resolving before KDD and EDM techniques can optimally be applied to KGs. These include the formalization of concepts such as an 'interesting pattern' found in a genome-scale KG, since 'interestingness' is subjective and will depend on the user's perspective. The concept of 'explaining a specific biological story' using a minimum set of non-redundant and relevant patterns from the KG also needs to be formalized. These theoretical insights need to be turned into useful, scalable and interactive tools, suitable for use by non-experts and tested against real biological problems.

We have previously described our approaches (Figure 1) to build genome-scale KGs with the KnetBuilder (https://github.com/Rothamsted/knetbuilder) data integration platform (Hassani-Pak *et al.,* 2016), to extend KGs with novel gene–phenotype relations from the literature (Hassani-Pak *et al.,* 2010), to publish KGs as standardized and interoperable data based on FAIR principles (Brandizi *et al.,* 2018a) and to visualize biological knowledge networks in an interactive web application (Singh *et al.,* 2018). Our data integration approach to build KGs is based on an intelligent data model with just enough semantics to capture complex biological relationships between genes, traits, diseases and many more information types derived from curated or predicted information sources. In this paper, we describe the KnetMiner gene discovery platform (knetminer.org) for searching large genome-scale KGs and visualizing interesting subgraphs of connected information about the biology of genes, gene networks, traits and diseases. KnetMiner is customizable and portable and therefore provides a cost-effective delivery platform for application to new species and datasets. We provide use cases to demonstrate how KnetMiner has helped scientists to tell the story of complex traits and diseases in *Arabidopsis thaliana* and *Triticum aestivum* (bread wheat). The methods section describes the algorithms behind core features of KnetMiner, that is, identifying interesting subgraphs and using these to rank candidate genes.

Results

KnetMiner can assist in various stages of a typical gene discovery project: from the early stages of literature review and hypothesis generation to later stages of biological understanding and hypothesis validation. The user-centric web interfaces support user journeys for the exploration of complex connected data. An initial simple search interface triggers a sophisticated search process and steps the user from generation to publication of interactive gene networks (Figure 2). We have selected two biological case studies that show the application of KnetMiner in gene–trait discovery and candidate gene prioritization in a model and non-model species. A detailed description of the latest KnetMiner features is available in the File S1 or online user tutorial.

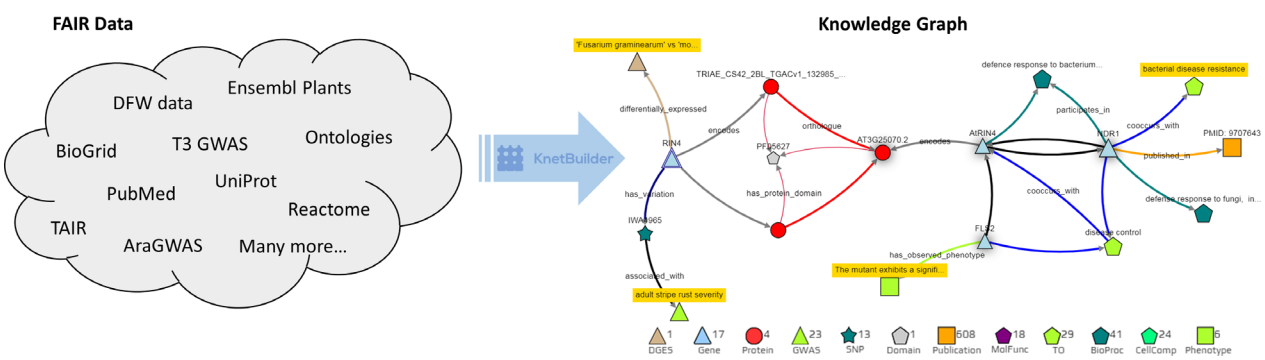

Figure 1 Diverse and heterogenous FAIR data sets are harmonized into a knowledge graph using the KnetBuilder software. Genome-scale knowledge graphs contain all the genes of an organism with links to functional information across species (Hassani-Pak *et al.,*2016). The Illustration shows a single gene knowledge network containing many biological labels and relation types.

Gene network discovery

KnetMiner is being used extensively to drive gene–trait discovery research in the publicly funded 'Designing Future Wheat' programme (https://designingfuturewheat.org.uk/), see for example (Adamski *et al.,* 2020; Alabdullah *et al.,* 2019; Harrington *et al.,* 2020). Wheat (*Triticum aestivum*) is the third most-grown cereal crop in the world after maize and rice, and has a hexaploid 15 Gb genome which is 5 times the size of the human genome (The International Wheat Genome Sequencing Consortium (IWGSC) *et al.,* 2018). As an example, white-grained wheat varieties lack the red compounds (flavonoids) of the seed coat and are milder in flavour. However, white grains are prone to pre-harvest sprouting (PHS) which causes the grain to germinate before harvest and results in a loss of grain quality. It has been known for some time that PHS is associated with grain colour (Nilsson-Ehle, 1914), and that the red pigmentation of wheat grain is controlled by *R* genes on the long arms of chromosomes 3A, 3B and 3D (Sears, 1944). However, after decades of research, it remains unclear whether there is a potential link between the grain colour gene R (Myb) and other phenotypes such as PHS.

We used KnetMiner to search for TRAESCS3D02G468400 (https://knetminer.org/wheatknet/genepage?list=TRAESCS3D02G468400) – the wheat *R* gene (the orthologue of Arabidopsis *TT2*) on chromosome 3D, and to explore its knowledge network as generated by KnetMiner. Our generated *TT2* network has a total of 823 connected nodes of 11 different types (see Table S1) including wheat-specific information sources but also cross-species information from model organisms such as Arabidopsis and rice. Furthermore, a range of relation types are present in the

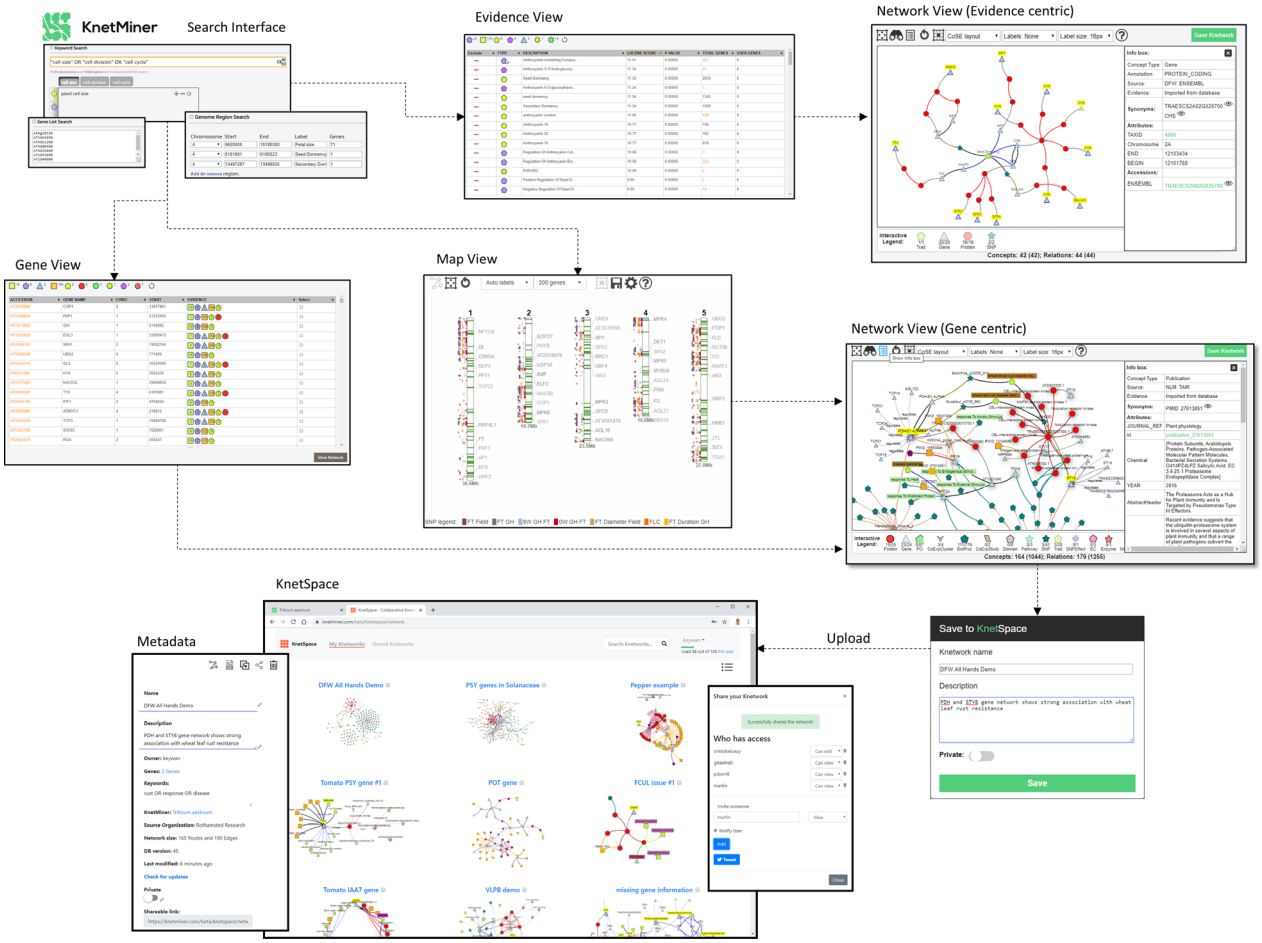

Figure 2 User journeys in KnetMiner. Users start with a search for keywords, genes and regions. KnetMiner provides search term suggestions and real-time query feedback. From a search, a user is presented with the following views: *Gene View* is a ranked list of candidate genes along with a summary of related evidence types. *Map View* is a chromosome-based display of QTL, GWAS peaks and genes related to the search terms. *Evidence View* is a ranked list of query-related evidence terms and enrichment scores along with linked genes. By selecting one or multiple elements in these three views, the user can get to the *Network View* to explore a gene-centric or evidence-centric knowledge network related to their query and the subsequent selection. The *Network View* has a set of features for exploring highly connected and rich biological information. For example, the 'Info Box' shows the properties of nodes and edges and provides hyperlinks to core databases. The 'Interactive Legend' allows users to add or hide information on a whole network level, and, a right-click radial menu allows to do this for individual nodes and edges. The 'Save Knetwork' button allows users to save a gene network, along with metadata and layout information to their workspace (named KnetSpace). In KnetSpace, users can manage all gene networks saved from various KnetMiner resources, edit networks, share with other users and publish online, for example https://knetminer.com/beta/knetspace/network/b21cd8b5-9eb8-4713-aefc-4785f4c8c8f7.

network including homologies, transcription-factor target relations, protein–protein interactions, phenotypic observations and correlations from mutant and genetic studies, as well as, curated or auto-generated links to ontology terms and publications.

To reach the final network visualization, KnetMiner first searches for interesting subgraphs generated from the user-provided genes and keywords (see Methods—Graph Interestingness). In this example, the *TT2* gene search was performed without additional keywords, KnetMiner therefore only shows paths in the gene network that lead to traits and phenotypes. This trims the network from 823 nodes down to 245 nodes including 101 Traits, 48 Phenotypes, 72 SNPs, 22 Genes and 2 Protein nodes (Figure 3a). This network is ultimately displayed in the Network View which provides interactive features to hide or add specific evidence types from the network. Graphical nodes are displayed in a meaningful combination of shapes, colours and sizes to distinguish different types of evidence. A shadow effect on nodes indicates that more information is available but has been hidden. However, the auto-generated network is not yet telling a story that is specific to our traits of interest and it is limited to evidence that is phenotypic in nature.

To highlight extra ways to use Knetminer and to further refine and extend the search for evidence that links *TT2* to grain colour and PHS, we can provide additional keywords relevant to the traits of interest. Seed germination and dormancy are the underlying developmental processes that activate or prevent pre-harvest sprouting in many grains and other seeds. The colour of the grain is known to be determined through accumulation of proanthocyanidin, an intermediate in the flavonoid pathway, found in the seed coat. These terms and phrases can be combined using Boolean operators (AND, OR, NOT) and used in conjunction with a list of genes. Thus, we search for TRAESCS3D02G468400 (https://knetminer.org/wheatknet/genepage?list=TRAESCS3D02G468400&keyword=dormancy%20germination%20color%20flavonoid%20proanthocyanidin) and the keywords: *'seed germination' OR 'seed dormancy' OR colour OR flavonoid OR proanthocyanidin*. This time, KnetMiner filters the extracted *TT2* gene network (823 nodes) down to a smaller subgraph of 68 nodes and 87 relations in which every path from *TT2* to another node corresponds to a line of evidence to phenotype or molecular characteristics based on our keywords of interest (Figure 3b).

This auto-generated subgraph visualizes complex information in a concise and connected format, helping facilitate biologically meaningful conclusions between *TT2* and phenotypes such PHS (see Table S2). The subgraph indicates that *TT2* in wheat is predicted to regulate the transcriptional activation of *MFT*. If you click on the 'cooccurs_with' edge between *MFT* and the linked trait nodes, you can see in the 'Info Box' that *MFT* has been linked in a recent publication to grain germination and seed dormancy in wheat (Li *et al.,* 2014; Nakamura *et al.,* 2011). The evidence sentences and hyperlinks to publications can be accessed by clicking links provided in the Info Box. The graph also reveals that the *MFT* ortholog in Arabidopsis is linked to decreased germination rate in the presence of ABA (Xi *et al.,* 2010) and positive regulation of seed germination. To investigate potential links between grain colour and other phenotypes, the TT2 gene network can be expanded with two clicks using the Interactive Legend (see User Tutorial), to add interacting genes in wheat or model species along with their phenotypic information. For example, the Arabidopsis *TT2* ortholog is shown to interact with *TTG1* which has links to phenotypes such as lateral root number and root hair length in Arabidopsis (Bahmani *et al.,* 2016; Bipei Zhang, 2017). Root hairs are tubular outgrowths from specific epidermal cells that function in nutrient and water absorption (Larry Peterson and Farquhar, 1996).

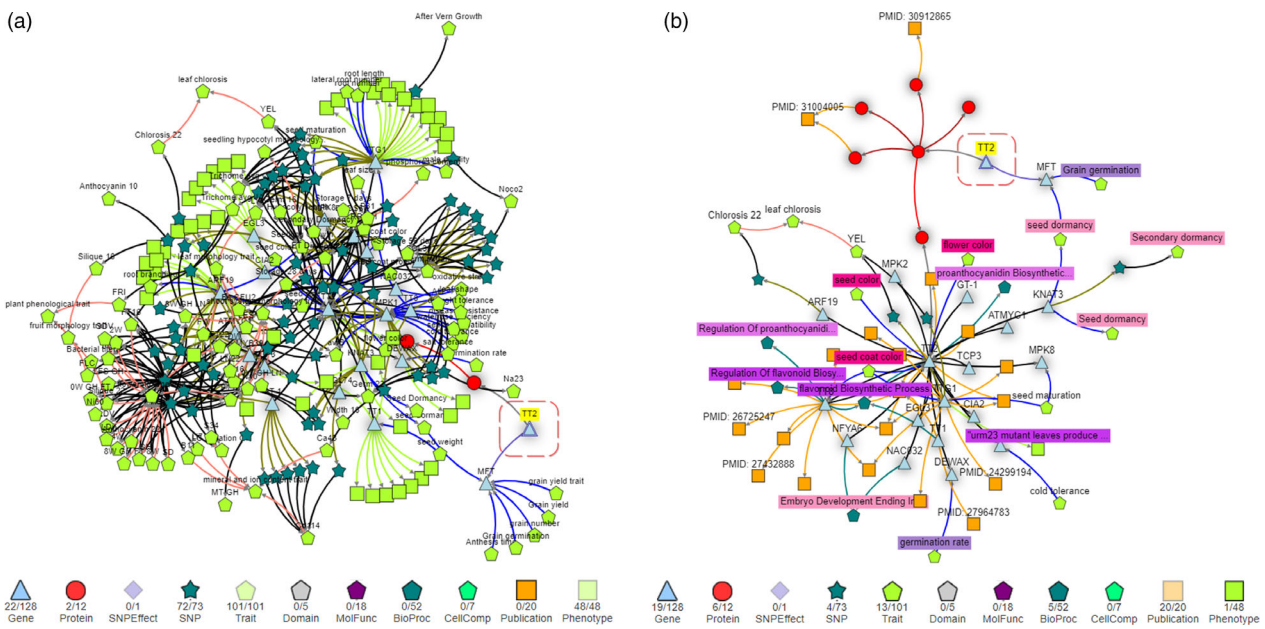

Figure 3 KnetMiner has two modes for generating gene networks depending on whether keywords are provided. (a) TT2 gene network without additional keywords shows all paths to traits and phenotypes. (b) TT2 gene network with keywords relevant to PHS and grain colour traits shows a smaller and more specific network including publications, traits, phenotypes, GO terms, pathways and more.

Overall, the exploratory link analysis has generated a potential linkage between grain colour and PHS due to *TT2-MFT* gene interaction and suggested a new hypothesis between two traits (PHS and root hair density) that were not part of the initial investigation and previously thought to be unrelated. Furthermore, it raises the possibility that *TT2* mutants might have more root hairs and higher nutrient and water absorption, and therefore cause early germination of the grain. More data and experiments will be needed to address this hypothesis and close the knowledge gap.

Candidate gene prioritization

Forward genetics studies, such as a genome-wide association study (GWAS) or quantitative trait loci (QTL) mapping, aim to identify regions in the genome where the genetic variation correlates with variation observed in a quantitative trait (e.g. general intelligence, days to flowering) (Atwell *et al.,* 2010; Polderman *et al.,* 2015; Sonah *et al.,* 2015). They are based purely on statistical tests and do not use biological understanding in considering candidates. It is often difficult to elucidate which exact marker is actually biologically significant, particularly in the face of epistatic and epigenetic effects which are often not considered. GWAS and QTL regions can encompass many seemingly unrelated genes. Candidate gene analysis aims to identify the most likely cause for the phenotypic variation. The identification of candidate genes underlying QTL is not trivial; therefore, genetic studies often stop after QTL mapping or perform a basic search for genes with potentially interesting annotations.

For example, in a recent QTL study in Arabidopsis, a region on chromosome 4 was identified that contained overlapping QTLs for multiple petal traits (Abraham *et al.,* 2013). As this QTL overlapped with the *ULTRAPETALA1* (*ULT1*) locus, a known floral meristem regulator with a role in petal development (Fletcher, 2001), the authors tested whether *ULT1* might be responsible for this QTL. However, the authors stated that among the ecotypes used in the study none showed any polymorphic sites within the *ULT1* coding or 2kb upstream region; and the T-DNA insertional mutation of *ULT1* showed no significant effect on petal form either. Taken together, the evidence suggested that *ULT1* was not responsible for the petal size QTL, and the causal gene remained unidentified as is the case in many other GWAS and QTL studies. Therefore, to explore this further, we analysed an overlapping petal size QTL (manuscript in preparation) using a more sophisticated and evidence-based search to see whether the authors may have missed something. The biological processes underpinning the size of plant tissues and organs are likely to be related to changes on a cellular level. We therefore used, as inputs to KnetMiner, the location of a petal size QTL (chromosome 4, 9.92–10.18 Mb) and the keywords *'cell size' OR 'cell cycle' OR 'cell division'*. KnetMiner identified 71 genes in the QTL region and ranked them according to their relevance to the keywords (Figure 4a) (see Methods—Gene Ranking).

The top five highest ranked genes by KnetMiner included a poorly studied gene (AT4G18330) with no links to publications in Arabidopsis and a few high-level GO annotations. However, the KnetMiner subgraph for AT4G18330 indicated that the yeast ortholog YER025W (eIF-2-gamma) interacts with cell division cycle proteins such as CDC123 (Figure 4b). Although no knockouts were available for this gene, a polymorphism in the regulatory region was associated with altered cellular and petal phenotypes consistent with a role in petal size (manuscript in preparation). The ability to both systematically and visually evaluate different layers of evidence arising from orthologs to interactions is highly advantageous; it is quick to view, and as such, the most relevant genes can immediately be investigated further.

Methods

Graph pattern mining

We have previously described our tools and methods to build FAIR genome-scale Knowledge Graphs (KG) using the KnetBuilder and rdf2neo data integration platforms (Brandizi *et al.,* 2018a, 2018b;

(a)

ACCESSION	GENE NAME	CHRO	START	EVIDENCE
AT4G18040	EIF4E	4	10016530	7 1 164
AT4G18390	TCP2	4	10162922	2 1 10 2 106
AT4G18050	PGP9	4	10021786	4 3 80 1
AT4G17980	NAC071	4	9978703	3 42
AT4G18330		4	10126533	10 2 33 1
AT4G18130	PHYE	4	10042137	38
AT4G18060	SH3P3	4	10026650	2 3 2 20
AT4G18170	WRKY28	4	10061214	2 17
AT4G18300		4	10118784	14
AT4G18340		4	10130050	1 1 12
AT4G17880	MYC4	4	9933326	12

(b)

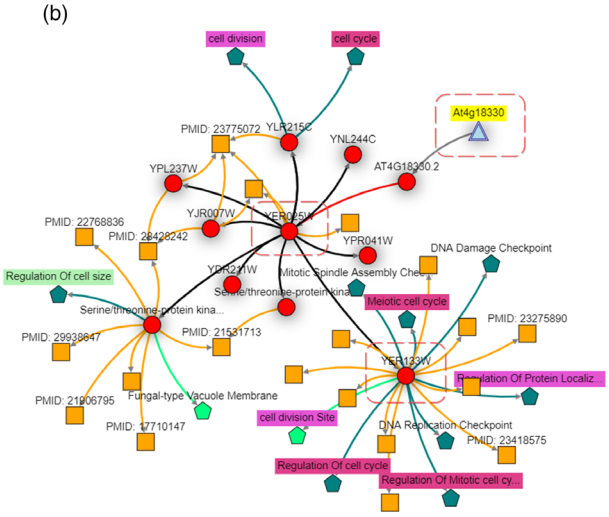

Figure 4 From QTL to candidate genes to gene networks. (a) Candidate genes within a petal size QTL are ranked based on their KnetScore. The Evidence column summarizes the related information found across species. AT4G18330 is linked to 10 biological processes, 2 cellular components, 33 publications and 1 protein related to 'cell size' OR 'cell cycle' OR 'cell division'. (b) Interactive gene network for AT4G18330 and 'cell size'-related keywords. All publications are linked to the yeast ortholog. The yeast ortholog YER025W interacts with several cell size, cell division and cell cycle-related proteins.

Hassani-Pak *et al.,* 2016). Here, we elaborate how KnetMiner uses the KG to extract biologically meaningful subgraphs that tell the story of complex traits and diseases. Biologically, plausible patterns in the KG are collections of paths through the connected information that most biologists would generally agree to be informative when studying the function of a gene. Searching a KG for such patterns is akin to searching for relevant sentences containing evidence that supports a particular point of view within a book. Such evidence paths can be short; for example, Gene A was knocked out and phenotype *X* was observed; or alternatively, the evidence path can be longer; for example, Gene A in species *X* has an ortholog in species *Y*, which was shown to regulate the expression of a disease-related gene (with a link to the paper). In the first example, the relationship between gene and disease is directly evident and experimentally proven, while in the second example, the relationship is indirect and less certain but still biologically meaningful. There are many evidence types that should be considered for evaluating the relevance of a gene to a trait. In a KG context, a gene is considered to be, for example, related to 'early flowering' if any of its biologically plausible graph patterns contain nodes related to 'early flowering'. In this context, the word 'related' does not necessarily mean that the gene in question will have an effect on 'flowering time', but it means that there is a valid piece of evidence that a domain expert should consider when judging whether the gene is related to 'flowering time'.

We use the notion of a semantic motif to define a plausible path through the KG. Our semantic motifs start with a gene node and end with other nodes representing biological entities, ontology terms, publications etc. For example, a path that travels from a Gene node to a GO-term, through an ortholog relation, is biologically plausible (orthologs have often the same function), while travelling through a paralog relation is not (paralogs often adapt new functions). KnetMiner instances can have a bespoke set of semantic motifs reflecting the data model of the KG built for one particular species or domain of interest. We are working towards migrating KnetMiner to support the Cypher graph query language and the Neo4j graph database as a practical and expressive way to define the graph searches that capture the semantic motifs of interest. Table S3 contains example Cypher queries used in the public wheat KnetMiner along with summary statistics for each query. The KnetMiner gene search and subgraph generation are essentially based on these well-defined graph queries. Not every gene will necessarily match all semantic motifs; however, the ones it contains are extracted and their union is taken to produce a gene-centric subgraph (GCS). For example, the wheat KG has over 114 000 GCSs (one for each wheat gene) with sizes of min = 1, max = 6220 and mean = 181 nodes.

Nodes that are included in a GCS are presumed to be transferable to the gene of interest; in contrast, concepts that are excluded from a GCS (although still part of the KG) are presumed to be irrelevant to the gene in question. Notably, if a semantic motif fails to capture an important biological motif, then downstream knowledge mining applications would not be able to exploit this information.

Graph interestingness

Even a single GCS with hundreds of nodes can be complex and challenging to comprehend when shown to a user, let alone if combining GCSs for tens to hundreds of genes. There is therefore a need to filter and visualize the subset of information in the GCSs that is most interesting to a specific user. However, the interestingness of information is subjective and will depend on the biological question or the hypothesis that needs to be tested. A scientist with an interest in disease biology is likely to be interested in links to publications, pathways and annotations related to diseases, while someone studying the biological process of grain filling is likely more interested in links to physiological or anatomical traits. To reduce information overload and visualize the most interesting pieces of information, we have devised two strategies. (1) In the case of a combined gene and keyword search, we use the keywords as a filter to show only paths in the GCS that connect genes with keyword-related nodes, that is nodes that contain the given keywords in one of their node properties. In the special case where too many publications remain even after keyword filtering, we select the most recent N publications (default $N = 50$). Nodes not matching the keyword are hidden but not removed from the GCS. (2) In the case of a simple gene query (without additional keywords), we initially show all paths between the gene and nodes of type phenotype/trait, that is any semantic motif that ends with a trait/phenotype, as this is considered the most important relationship to many KnetMiner users.

Gene ranking

We have developed a simple and fast algorithm to rank genes and their GCS for their importance. We give every node in the KG a weight composed of three components, referred to as SDR, standing for the Specificity to the gene, Distance to the gene and Relevance to the search terms. Specificity reflects how specific a node is to a gene in question. For example, a publication that is cited (linked) by hundreds of genes receives a smaller weight than a publication which is linked to one or two genes only. We define the specificity of a node x as: $S(x) = \log\frac{N}{n}$ where n is the frequency of the node occurring in all N GCS. Distance assumes information which is associated more closely to a gene can generally be considered more certain, versus one that is further away, for example, inferred through homology and other interactions increases the uncertainty of annotation propagation. A short semantic motif is therefore given a stronger weight, whereas a long motif receives a weaker weight. Thus, we define the second weight as the inverse shortest path distance of a gene g and a node x: $D(g,x) = \frac{1}{|v_g \to v_x|}$. Both weights S and D are not influenced by the search terms and can therefore be precomputed for every node in the KG. Relevance reflects the relevance or importance of a node to user-provided search terms using the well-established measure of inverse document frequency (IDF) and term frequency (TF) (Salton and Yang, 1973). TF*IDF forms the basis of the Lucene search engine library (https://lucene.apache.org/), used in KnetMiner. We define the relevance of node x to a search term t as $R(t,x) = \mathrm{TF} \times \mathrm{IDF}(t,x)$, where $R = 0$ when no match is found and $R = 1$ when the user does not provide any keywords. The three measures (S, D and R) have unique and uncorrelated characteristics. Each node in KnetMiner is given a combined SDR weight. Therefore, for a given GCS $X_g = \{x_1, x_2, \ldots, x_n\}$ and search terms t, we define the *KnetScore* of a gene as:

$$KnetScore(t, X_g) = \sum_{x_i \in X_g \cap x_i \ni t} S(x_i) * D(g,x) * R(t, x_i)$$

The sum considers only GCS nodes that contain the search terms. In the absence of search terms, we sum over all nodes of

the GCS with $R = 1$ for each node. The computation of the KnetScore (*SDR* weights) requires graph traversals and string searches over the KG. Performing these operations on-the-fly would slow down the responsiveness of the application. Therefore at initialization, KnetMiner pre-processes the KG and builds indices to speed up the *SDR* weight calculation. The pre-indexing time depends on a number of factors including number of available cores, the KG size, number of genes and number of semantic motifs. With the indices in place, the SDR weight can be computed in constant time O(1). A KnetMiner search that returns n genes and m evidence nodes can rank all genes in linear time O (n + m).

Discussion

Scientists spend a considerable amount of time searching for new clues and ideas by synthesizing many different sources of information and using their expertise to generate hypotheses. Gene discovery is often hampered by the challenges of data integration, and new approaches are needed to improve the efficiency, reproducibility and objectivity of the process that leads to new ideas and hypotheses. KnetMiner provides a sophisticated search across a semantically rich knowledge graph built from large scale integration of public and private data sets. It addresses the needs of scientists who may lack the time and the broad expertise that is necessary to connect, explore and compare the wealth of genetic, 'omics, and phenotypic information available in the literature and a wide range of related biological databases from key model and non-model species.

KnetMiner is commonly used by scientists in academia and industry to accelerate gene–trait discovery research. In several biological studies, KnetMiner enabled the identification of hidden relationships between important agronomic traits and potential candidate genes. The presented case studies have shown practical applications of KnetMiner to the understanding of challenging and complex traits in wheat and Arabidopsis. KnetMiner was used in 2014 to investigate traits such as height of biomass willows (Hanley and Karp, 2014) and has more recently become part of a wider roadmap for gene function characterization in crops (Adamski *et al.,* 2020; Harrington *et al.,* 2020). Public KnetMiner resources (e.g. Arabidopsis, wheat and rice) give a flavour of the capabilities that are in KnetMiner. While we have so far mostly concentrated on customizing KnetMiner for plant sciences and crop improvement, the software we have developed is generic and KGs and KnetMiner can readily be built for many species. Compared to biological discovery platforms available for specific species (Carvalho-Silva *et al.,* 2019; Miller *et al.,* 2017; Mungall *et al.,* 2017), KnetMiner is species-agnostic and therefore provides a more cost-effective delivery platform for application to new data sets. For example, we are working on the development of KnetMiner resources for pest insects and fungal pathogens, and, have recently repurposed KnetMiner for mining COVID-19 biomedical data and literature (Hutson, 2020). KnetMiner is available as a Docker image from DockerHub and can easily be deployed with a provided sample KG.

Different KnetMiner views for exploring the search output have been developed; each view has a different aim and helps address different questions. The main design principle was to divide the visualization into two steps. First, to present the results in formats that are intuitive and familiar to biologists, such as tables and chromosome views, allowing them to explore the data, make choices as to which gene to view, or refine the query if needed. These initial views help users to reach a certain level of confidence with the selection of potential candidate genes. However, they do not tell the biological story that links candidate genes to traits and diseases. In a second step, to enable the stories and their evidence to be investigated in full detail, the Network View visualizes highly complex information in a concise and connected format, helping facilitate biologically meaningful conclusions. Consistent graphical symbols are used for representing evidence types throughout the different views, so that users develop a certain level of familiarity, before being exposed to networks with complex interactions and rich content.

The methods (graph pattern mining, graph interestingness and gene ranking) that power the KnetMiner user interface are also available as API calls and can be used to embed visualizations of gene-centric subgraphs in third party web applications or to integrate graph analytics and gene ranking in custom workflows. For example, the KnetMiner REST API is used in Ensembl Plants (Bolser *et al.,* 2017), The Triticeae Toolbox (Blake *et al.,* 2016) and GrainGenes (Blake *et al.,* 2019) to link gene sequences to rich gene knowledge graphs. The graph database backend, as well as the FAIR-based data management policies, is another development in which we are investing our efforts, which have the main advantage of allowing us to build a data asset that has the potential to be useful to a wealth of applications, complementary to KnetMiner. The SPARQL and Cypher endpoints have the benefit of providing a layer of access to data that have a more general use than gene-centric knowledge exploration and which, for instance, could be obtained with scripts accessing APIs, workflow tools like Galaxy (Afgan *et al.,* 2018) or data analytics workbenches like Jupyter (Kluyver *et al.,* 2016). This is facilitated by adhering to the well-known good practice of the FAIR principles, which includes the adoption of common data schemas and ontologies (Garcia *et al.,* 2017).

Conclusion

KnetMiner is an integrated, intelligent, interactive gene and gene network discovery platform, designed to help scientists understand the biological stories of complex traits and diseases across species. The challenges faced in gene discovery for crop improvement are different from work done in drug target discovery in humans. A plethora of diverse crop, insect and pathogen data sets need to be interconnected with well-curated model species data and the scientific literature information. KnetMiner has been designed to support a diversity of species, data sets and use cases. We see the real value of KnetMiner being in the areas of multi-dimensional data integration and global optimization where, by comparison, a human cannot hold that much data depth or breadth together and in balance. An expert scientist might read and understand a single journal article better and follow a single thread or conclusion accurately, but it is the integration and balance of the entirety of the indexed knowledge graph where KnetMiner shines brightest.

Acknowledgements

This work was supported by the UKRI Biotechnology and Biological Sciences Research Council (BBSRC) through the Designing Future Wheat ISP (BB/P016855/1), DiseaseNetMiner TRDF (BB/N022874/1), ONDEX SABR funding (BB/F006039/1) and National Capability in Crop Phenotyping (BB/J004464/1). CR and KHP are additionally supported by strategic funding to

Rothamsted Research from BBSRC. JHD also acknowledges support from the National Science Foundation (cROP project 1340112). We acknowledge all the past and present members of the KnetMiner Bioinformatics team at Rothamsted for their scientific inputs, software testing and technical support: Emma Bailey, Dan Smith, Robert King, David Hughes, Monika Mistry, Minja Zorc, Fengyuan Hu, Jan Taubert and William Brown. We acknowledge all our collaborators who contributed to the development of the KnetMiner resources and software in the past including Martin Castellote, Maria Esch, Vasiliki Koutra, Haolin Li, Philipp Bayer, Ramil Mauleon, Cristobal Uauy, Jean-Luc Jannink, Clay Birkett, Uwe Schulz, Steve Hanley, Francis Newson and Richard Holland.

Conflicts of interest

The authors declare that they have no competing interests.

Author contributions

KHP designed the approach as part of his dissertation with CR, collected results and drafted the manuscript. KHP, AS, MB, JH and the KnetMiner team implemented the KnetMiner framework and maintain its public instances. SA and JDP helped to build the Arabidopsis and wheat knowledge graphs. AP and JHD contributed towards the biological use cases. All authors read, reviewed and approved the final manuscript.

Software availability

Project name: KnetMiner—Knowledge Network Miner.
Project home page: https://knetminer.org
Source code: https://github.com/Rothamsted/knetminer
Docker image: https://hub.docker.com/r/knetminer/knetminer
Deployment instructions: https://github.com/Rothamsted/knetminer/wiki/
Knowledge Graph Endpoints: http://knetminer.org/data
Operating system(s): Platform independent.
Programming language: Java and JavaScript.
Other requirements: Docker.
Licence: MIT.
Any restrictions to use by non-academics: database licence needed.

References

Abraham, M.C., Metheetrairut, C. and Irish, V.F. (2013) Natural variation identifies multiple loci controlling petal shape and size in Arabidopsis thaliana. *PLoS One,* **8**, e56743.

Adamski, N.M., Borrill, P., Brinton, J., Harrington, S.A., Marchal, C., Bentley, A.R., Bovill, W.D. *et al*. (2020) A roadmap for gene functional characterisation in crops with large genomes: Lessons from polyploid wheat. *eLife,* **9**, https://doi.org/10.7554/eLife.55646.

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J. *et al*. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**(W1), W537–W544.

Alabdullah, A.K., Borrill, P., Martin, A.C., Ramirez-Gonzalez, R.H., Hassani-Pak, K., Uauy, C., Shaw, P. *et al*. (2019) A co-expression network in hexaploid wheat reveals mostly balanced expression and lack of significant gene loss of homeologous meiotic genes upon polyploidization. *Front. Plant Sci.* **10**, 1325.

Ali, M., Hoyt, C.T., Domingo-Fernández, D., Lehmann, J. and Jabeen, H. (n.d.). *BioKEEN: A library for learning and evaluating biological knowledge graph embeddings*. https://doi.org/10.1101/475202

Alshahrani, M. and Hoehndorf, R. (2018) Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics,* **34**, i901–i907.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D. *et al*. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature,* **465**, 627.

Bahmani, R., Kim, D.G., Kim, J.A. and Hwang, S. (2016) The density and length of root hairs are enhanced in response to cadmium and arsenic by modulating gene expressions involved in fate determination and morphogenesis of root hairs in Arabidopsis. *Front. Plant Sci.* **7**, 1763.

Bipei Zhang, A.S. (2017) TRANSPARENT TESTA GLABRA 1-dependent regulation of flavonoid biosynthesis. *Plants,* **6**, https://doi.org/10.3390/plants6040065.

Blake, V.C., Birkett, C., Matthews, D.E., Hane, D.L., Bradbury, P. and Jannink, J.-L. (2016) The Triticeae Toolbox: combining phenotype and genotype data to advance small-grains breeding. *Plant Genome,* **9**, https://doi.org/10.3835/plantgenome2014.12.0099.

Blake, V.C., Woodhouse, M.R., Lazo, G.R., Odell, S.G., Wight, C.P., Tinker, N.A., Wang, Y. *et al*. (2019). GrainGenes: centralized small grain resources and digital platform for geneticists and breeders. *Database: J. Biol. Databases Curat*. https://doi.org/10.1093/database/baz065

Bolser, D.M., Staines, D.M., Perry, E. and Kersey, P.J. (2017) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods Mol. Biol.* **1533**, 1–31.

Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.

Brandizi, M., Singh, A., Rawlings, C. and Hassani-Pak, K. (2018a) Towards FAIRer biological knowledge networks using a hybrid linked data and graph database approach. *J. Integr. Bioinform.* **15** https://doi.org/10.1515/jib-2018-0023.

Brandizi, M., Singh, A., Rawlings, C. and Hassani-Pak, K. (2018b) Getting the best of Linked Data and Property Graphs: rdf2neo and the KnetMiner Use Case. *SWAT4LS Proceedings*. https://doi.org/10.6084/m9.figshare.7314323.v1.

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M. *et al*. (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **47**(D1), D1056–D1065.

De Bie, T. (2013) Subjective Interestingness in Exploratory Data Mining. In *Advances in Intelligent Data Analysis XII. IDA 2013. Lecture Notes in Computer Science* (Vol. **8207**, pp. 19–31). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41398-8_3

De Bie, T. and Spyropoulou, E. (2013) A Theoretical Framework for Exploratory Data Mining: Recent Insights and Challenges Ahead. In *Lecture Notes in Computer Science* (pp. 612–616).

Ehrlinger, L. and Wöß, W. (2016) Towards a Definition of Knowledge Graphs. *Proceedings of SEMANTiCS 2016*, **48**. http://ceur-ws.org/Vol-1695/

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I. *et al*. (2020) Introduction: What Is a Knowledge Graph? In *Knowledge Graphs* (pp. 1–10). Cham: Springer. https://link.springer.com/book/10.1007/978-3-030-37439-6

Fletcher, J.C. (2001) The ULTRAPETALA gene controls shoot and floral meristem size in Arabidopsis. *Development,* **128**, 1323–1333.

Fox, P. and Hendler, J. (2011) Changing the equation on scientific data visualization. *Science,* **331**, 705–708.

Garcia, L., Giraldo, O., Garcia, A. and Dumontier, M. (2017) Bioschemas: schema.org for the life sciences. *Proceedings of SWAT4LS*.

Hanley, S.J. and Karp, A. (2014) Genetic strategies for dissecting complex traits in biomass willows (Salix spp.). *Tree Physiol.* **34**, 1167–1180.

Harrington, S.A., Backhaus, A.E., Singh, A., Hassani-Pak, K. and Uauy, C. (2020) The wheat GENIE3 network provides biologically-relevant information in polyploid wheat. *G3,* **10**, 3675–3686. https://doi.org/10.1534/g3.120.401436

Hassani-Pak, K., Castellote, M., Esch, M., Hindle, M., Lysenko, A., Taubert, J. and Rawlings, C. (2016) Developing integrated crop knowledge networks to advance candidate gene discovery. *Appl. Transl. Genom.* **11**, 18–26.

Hassani-Pak, K., Legaie, R., Canevet, C., van den Berg, H.A., Moore, J.D. and Rawlings, C.J. (2010) Enhancing data integration with text analysis to find proteins implicated in plant stress response. *J. Integr. Bioinform.* **7**, https://doi.org/10.2390/biecoll-jib-2010-121.

Holmes, J.H., & (2014). Knowledge discovery in biomedical data: theory and methods. In *Methods in Biomedical Informatics* (pp. 179–240). https://doi.org/10.1016/B978-0-12-401678-1.00007-5

Holzinger, A. and Jurisica, I. (2014) Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. In *Lecture Notes in Computer Science* (pp. 1–18).

Hutson, M. (2020) Artificial-intelligence tools aim to tame the coronavirus literature. *Nature*, https://doi.org/10.1038/d41586-020-01733-7.

Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M. and Möller, T. (2013) A systematic review on the practice of evaluating visualization. *IEEE Trans. Visual Comput. Graph.* **19**, 2818–2827.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K. *et al*. (2016) *Jupyter Notebooks - a publishing format for reproducible computational workflows*. https://doi.org/10.3233/978-1-61499-649-1-87.

Larry Peterson, R. and Farquhar, M.L. (1996) Root hairs: Specialized tubular cells extending root surfaces. *Botanical Review; Interpreting Botanical Progress,* **62**, 1–40.

Lee, B., Isenberg, P., Riche, N.H. and Carpendale, S. (2012) Beyond mouse and keyboard: expanding design considerations for information visualization interactions. *IEEE Trans. Visual Comput. Graph.* **18**, 2689–2698.

Li, Q., Fan, C., Zhang, X., Wang, X., Wu, F., Hu, R. and Fu, Y. (2014) Identification of a soybean MOTHER OF FT AND TFL1 homolog involved in regulation of seed germination. *PLoS One,* **9**, e99642 https://doi.org/10.1371/journal.pone.0099642.

Messina, A., Fiannaca, A., La Paglia, L., La Rosa, M. and Urso, A. (2018) BioGraph: a web application and a graph database for querying and analyzing bioinformatics resources. *BMC Syst. Biol.* **12**(Suppl 5), 98.

Miller, J., Town, C., Stuerzlinger, W. and Provart, N.J. (2017) ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *Plant*, http://www.plantcell.org/content/29/8/1806.short

Mohamed, S.K., Nováček, V. and Nounu, A. (2019) Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btz600.

Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S. *et al*. (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**(D1), D712–D722.

Nakamura, S., Abe, F., Kawahigashi, H., Nakazono, K., Tagiri, A., Matsumoto, T., Utsugi, S. *et al*.(2011) A wheat homolog of MOTHER OF FT AND TFL1 acts in the regulation of germination. *Plant Cell*, **23**, 3215–3229.

Nilsson-Ehle, H. (1914). *Zur Kenntnis der mit der keimungsphysiologie des weizens in zusammenhang stehenden inneren faktoren*.

Pavelin, K., Cham, J.A., de Matos, P., Brooksbank, C., Cameron, G. and Steinbeck, C. (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comput. Biol.* **8**, e1002554.

Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M. and Posthuma, D.(2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709.

Reese, J.T., Unni, D., Callahan, T.J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K.A. *et al*. (2021) KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns (New York, N.Y.),* **2**, 100155.

Russell-Rose, T., Chamberlain, J. and Azzopardi, L. (2018) Information retrieval in the workplace: A comparison of professional search practices. *Inf. Process. Manage.* **54**, 1042–1057.

Sacchi, L. and Holmes, J.H. (2016) Progress in biomedical knowledge discovery: a 25-year retrospective. *Yearbook Med. Inform.* **25**(S 01), S117–S129.

Salton, G. and Yang, C.S. (1973) On the specification of term values in automatic indexing. *J. Document.* **29**, 351–372.

Sears, E.R. (1944) Cytogenetic studies with polyploid species of wheat. II. Additional chromosomal aberrations in *Triticum vulgare*. *Genetics*, **29**, 232.

Sheth, A., Padhee, S. and Gyrard, A. (2019) Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Comput.* **23**, 67–75.

Singh, A., Rawlings, C.J. and Hassani-Pak, K. (2018) KnetMaps: a BioJS component to visualize biological knowledge networks. *F1000Research,* **7**, 1651.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I. and Belzile, F. (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* **13**, 211–221. https://doi.org/10.1111/pbi.12249.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R. *et al*. (2015) Big data: astronomical or genomical? *PLoS Biol.* **13**, e1002195.

Sweis, B.M., Abram, S.V., Schmidt, B.J., Seeland, K.D., MacDonald, A.W. 3rd, Thomas, M.J. and Redish, A.D. (2018) Sensitivity to "sunk costs" in mice, rats, and humans. *Science,* **361**, 178–181.

The International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators, Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J. *et al*. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.

Xi, W., Liu, C., Hou, X. and Yu, H. (2010) MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in Arabidopsis. *Plant Cell,* **22**, 1733–1748.

Xiaoxue, L., Xuesong, B., Longhe, W., Bingyuan, R., Shuhan, L. and Lin, L. (2019) Review and trend analysis of knowledge graphs for crop pest and diseases. *IEEE Access,* **7**, 62251–62264.

Yoon, B.-H., Kim, S.-K. and Kim, S.-Y. (2017) Use of graph database for the integration of heterogeneous biological data. *Genom. Inform.* **15**, 19–27.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 Occurrence of various information types in the wheat TT2 (TRAESCS3D02G468400) gene-centric subgraph.

Table S2 Examples of relation types and properties in the keyword-filtered TT2 (TRAESCS3D02G468400) subgraph.

Table S3 Example of semantic motifs (in Cypher language) used in KnetMiner with number of matches found in the Wheat Knowledge Graph (Release 45).

File S1 KnetMiner v4.0 user tutorial.

