

KnetMiner Tutorial

A web server for mining genes and networks controlling
complex traits

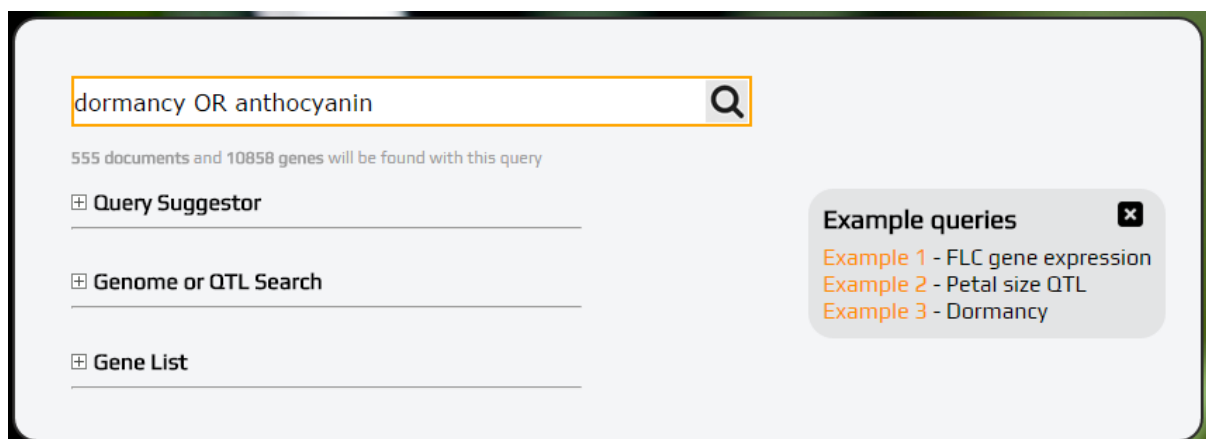
Author: Keywan Hassani-Pak

21 December 2016

About KnetMiner

KnetMiner provides an easy to use, web interface to visualisation and data mining tools for the discovery and evaluation of candidate genes from large scale integrations of public and private data sets. It addresses the needs of scientists who generally lack the time and technical expertise to review all relevant information available in the literature, from key model species and from a potentially wide range of related biological databases. We have previously developed genome-scale knowledge networks (GSKNs) for multiple crop and animal species (Hassani-Pak et al. 2016). The KnetMiner web server searches and evaluates millions of relations and concepts within the GSKNs in real-time to determine if direct or indirect links between genes and trait-based keywords can be established. KnetMiner accepts as user inputs: search terms in combination with a gene list and/or genomic regions. It produces a table of ranked candidate genes and allows users to explore the output in interactive genome and network map visualisation tools that have been optimised for web use on desktop and mobile devices. The KnetMiner web server and the GSKNs provide a step-forward towards systematic and evidence-based gene discovery. KnetMiner is available at: <http://knetminer.rothamsted.ac.uk>.

KnetMiner search interface



The screenshot displays the KnetMiner search interface. At the top, a search bar contains the text "dormancy OR anthocyanin" and a magnifying glass icon. Below the search bar, a message states "555 documents and 10858 genes will be found with this query". To the left, there are three expandable sections: "Query Suggestor", "Genome or QTL Search", and "Gene List", each with a plus icon and a horizontal line. To the right, a sidebar titled "Example queries" with a close button (X) lists three examples: "Example 1 - FLC gene expression", "Example 2 - Petal size QTL", and "Example 3 - Dormancy".

A Google-like search interface

The main search field of KnetMiner allows users to input any terms, for example related to a trait of interest. The search provides full support of the Lucene query syntax so that different terms can be combined with OR, AND, NOT statements. Real-time messaging provides the total number of resulting documents and genes while the user is typing the query. The terms can be high level

descriptions of a phenotypic trait (e.g. disease resistance) but also more specific terms such as biological processes or protein families (e.g. defense response to fungi or LRR). KnetMiner uses the provided input terms to search the organism specific knowledge network and to display genes and QTL associated with the input terms.

Query suggestions

KnetMiner contains a query suggestion wizard that helps users to refine their query by suggesting more specific terms or alternative synonyms. For example, using the query suggestion wizard on the term 'drought' would suggest other terms such as 'drought sensitivity' or 'response to dehydration'. The wizard allows adding, replacing or excluding the new terms from the query. The real-time messaging directly updates when the query changes to indicate if the new query would lead to a different number of resulting candidate genes. The suggested terms are derived from the underlying information network.



Whole genome vs. within QTL search

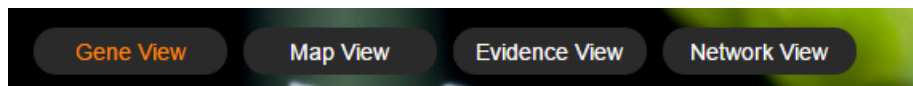
KnetMiner can be used in two modes "whole genome" and "within region". The former mode finds all genes of a genome that can be associated with the query terms, scores them and displays the top 100 (the whole list can be downloaded). In the latter mode when the user provides a list of QTL base pair positions, all genes that fall into the specified QTL boundaries will be displayed. Genes that were not displayed in the whole-genome mode because they were not within the top 100 ranked genes could be displayed in the QTL mode as long as they are related to the query terms and within the QTL region. Entering the start and end position of a QTL will display the number of genes in the genome that are covered by the QTL. Providing QTL information and restricting the search to the QTL genes will still conduct a full genome search but additionally apply a filter to discard all genes that are not located within the start and end boundaries of the QTL. This feature is only available if the KnetMiner organism has a sequenced genome and genes have therefor a physical location.

Adding gene lists to the search

Transcriptome analysis such as differential expression (DE) or co-expression (COE) analysis can provide a second route to gene discovery. Candidate genes from expression studies or other 'omics studies can be included in the search. KnetMiner will test if a gene from the given list has known associations with the input terms and overlaps with a QTL. A gene that is strongly associated with given input terms, and has one or more supporting QTL and is included in the user's gene list can be considered a strong candidate for experimental validation.

Display of search results in different views

The result of a search is essentially a list of candidate genes along with the supporting evidence. The web-interface provides different views that help to explore the search results and drill down into interesting candidate gene networks.



The Gene View

The Gene view uses a table to display up to 1000 candidate genes sorted by the KnetMiner relevance score and includes a column that summarizes the supporting evidence information. The evidence information is extendible and provides a short description string about the evidence. If the evidence is a publication, then the PubMed id is shown and linked to PubMed. A button allows excluding specific documents from the next search. User input genes are flagged in the "USER" column with a "yes". Clicking on a single gene or on a selection of genes opens the Network view.

Download as TAB delimited file

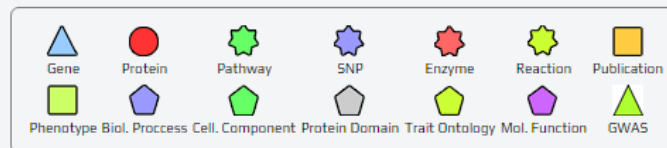
Select gene(s) and click "View Network" button to see the Oindex network.

Max number of genes to show:

Known targets: ☒ Novel targets: ☐

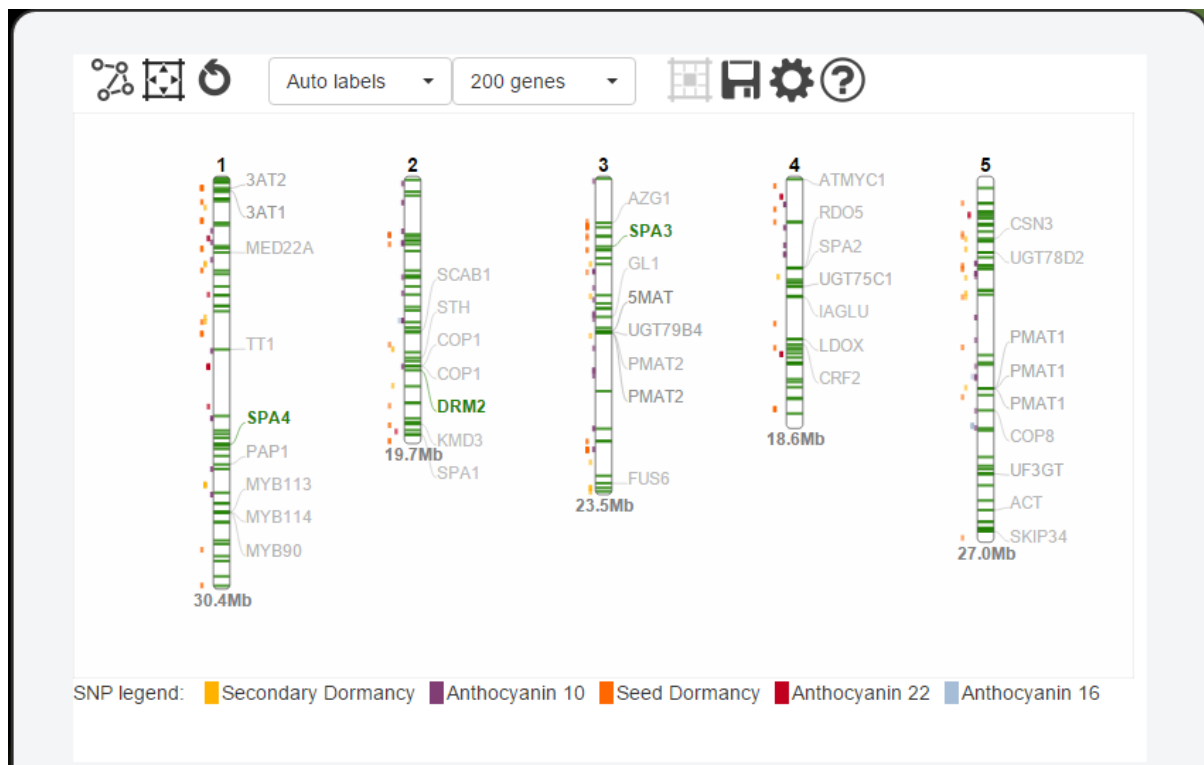
ACCESSION	CHRO	START	SCORE	USER	QTL	EVIDENCE	Select
AT5MAT	3	11398917	36.48	no	0		<input type="checkbox"/>
AT3G29680	3	11531442	27.28	no	0		<input type="checkbox"/>
AT1G03940	1	1009542	23.49	no	0		<input type="checkbox"/>
UGT79B4	3	11465851	17.73	no	0		<input type="checkbox"/>
AT1G03495	1	873184	17.02	no	0		<input type="checkbox"/>
AT5G39050	5	15634586	15.42	yes	0		<input checked="" type="checkbox"/>
PMAT1	5	15643603	14.46	yes	0		<input checked="" type="checkbox"/>
PMAT1	5	15641658	14.46	yes	0		<input checked="" type="checkbox"/>
AT4G14090	4	8122188	11.97	no	0		<input type="checkbox"/>
AT3G29670	3	11527872	11.77	no	0		<input type="checkbox"/>
AT5G54060	5	21936879	11.47	no	0		<input type="checkbox"/>

[View Network](#)



The Map View

The map view makes use of the newly developed and interactive [GenoMaps.js](https://github.com/Rothamsted/genomaps.js) (<https://github.com/Rothamsted/genomaps.js>) to display up to 1000 genes related to the search terms along the chromosomes and uses color coding to distinguish genes with high (green), medium (orange) and low (red) scores. User defined QTL and QTL/SNPs derived from the knowledge network are displayed on the left hand side of the chromosome. Searching within pre-defined QTL only displays candidate genes that are within the specified loci. This view not only illustrates effectively the overlap of genes and QTL but also the relative position of candidate genes w.r.t the QTL. Maps can be exported in PNG format. Genes that were selected by the user can be opened in the Network view by clicking the network icon.



The Evidence View

The Evidence view tab provides a document centric view of the search results sorted by query-relevance score (Lucene TF-IDF). All documents from the information network containing the query terms are displayed. A button allows users to exclude specific documents from the next search. For every document the total number of genes and the total number of user provided genes is displayed that are directly or in-directly connected to the document in the network. This is a very useful view to quickly get to genes that are for example involved in a specific pathway. Clicking on the number of genes will switch to the Network view which displays the document and how the genes are linked to it.



Legend: 100 (green square), 329 (orange square), 5 (blue triangle), 69 (red circle), 5 (purple pentagon), 23 (blue hexagon), 3 (green pentagon), 11 (green star), 8 (green triangle), 1 (yellow star), 1 (red star)

Exclude	TYPE	NAME	SCORE	GENES	USER GENES	QTLS
—	▲	Secondary dormancy was given ...	2.52	1236	0	0
—	▲	Number of days of seed dry st...	2.52	1364	0	0
—	▲	Results expressed as binary d...	2.28	1113	0	0
—	▲	Results expressed as binary d...	2.28	1016	2	0
—	▲	Results expressed as binary d...	2.28	1020	0	0
—	⬡	seed dormancy process	2.14	83	0	0
—	●	Q9LVW3	2.10	1	0	0
—	⬡	seed dormancy	2.01	1038	0	0
—	⬡	anthocyanin formation	1.96	296	3	0
—	■	Reduced seed dormancy phenoty...	1.95	2	0	0
—	■	Reduced seed dormancy phenoty...	1.95	2	0	0

Legend:

- Gene (blue triangle)
- Protein (red circle)
- Pathway (green star)
- SNP (blue star)
- Enzyme (red star)
- Reaction (yellow star)
- Publication (orange square)
- Phenotype (green square)
- Biol. Process (purple pentagon)
- Cell. Component (green pentagon)
- Protein Domain (blue hexagon)
- Trait Ontology (green triangle)
- Mol. Function (purple pentagon)
- GWAS (green triangle)

The Network View

The newly developed KnetMaps.js (<https://github.com/Rothamsted/knetmaps.js>) is used to display knowledge networks connecting genes with evidence documents. The network reveals the full information how genes are linked to a particular evidence concepts shown in the Gene View. Initially only the most important parts of the gene evidence network are displayed to the user. The Info Box panel can be opened to display all attributes (e.g. name, description, title, abstract etc.) attached to nodes and links in the network. KnetMaps provides a right-click menu on nodes and links that allow a user to hide information or explore and expand the network with additional information. Networks can be exported in Cytoscape compatible JSON, Oindex exchange format (OXL) and PNG.

A transcriptome (RNA-seq) experiment was designed by the Phillips lab at Rothamsted Research to understand the transcriptional differences between red and white grains. The RNA-seq reads were mapped to the wheat reference genome (Ensembl v21, cDNA transcripts) using BWA. Transcript abundance was estimated using eXpress and differentially expressed genes identified with edgeR. In total 214 genes were differentially expressed ($p < 0.05$) of which **104** had a considerable fold change ($\log_{2}FC > 2$) between red and white grain.

Having identified a list of differentially expressed genes (DEG), the questions scientists would consequently ask are:

- Do any of these DEG contribute to the expression of the grain colour trait?
- Do any of these DEG contribute to the expression of the PHS trait?
- Which biological processes and pathways are underlying these traits?
- Are there any common genes or mechanisms that regulate both traits?

The evidence sources users would need to navigate in order to answer these questions and evaluate whether any of these genes might have a role in that trait would include the GO terms, role in biochemical pathways, interaction networks, comparative information from related organisms, evidence of expression in tissue of interest, phenotype information, the scientific literature and other resources that might be specific to the domain of interest. Assembling a coherent view of how the bits of evidence might come together to “tell a story” about the biology that could explain how multiple genes might be implicated in a complex trait is demanding. The use case presented here demonstrates how KnetMiner can considerably reduce the data integration and exploration demands on the user by solving many of the technical challenges and providing the tools that allow biologists to focus on the biological story.

Choosing the right search terms

The use case presented here demonstrates the capabilities of KnetMiner for analysing a list of differentially expressed genes and to identify new targets or mechanisms that might help explain the as yet unknown basis for the link between grain colour and PHS.

Seed dormancy and germination are the underlying developmental processes that activate or prevent pre-harvest sprouting in many grains and other seeds. The user can provide this knowledge as a list of **keywords** into the search box. The KnetMiner **Query Suggester** can be used, on the one hand, to understand which evidence concepts from the knowledge network match the keywords and, on the other hand, to provide alternative synonyms or more specific keywords.

Exercise 1: Type the keyword *dormancy* into the search box. Try to replace it with a more specific keyword. Do you find more or less genes? What do you need to do to make it more precise?

A: The Query Suggester shows that the keyword ***dormancy*** matches Gene Ontology (GO), Trait Ontology (TO), gene, protein and publication evidence concepts from the wheat knowledge network (**Figure 2**). The TO and GO concepts are divided into terms specific for seed and bud dormancy. The term *grain dormancy* does not, however, occur in the knowledge network. As an alternative it is

possible to specialise the search keyword to ***seed dormancy*** as it can be assumed that processes involved in grain dormancy are similar to the ones involved in seed dormancy but different to bud dormancy. This returns more genes since the space between the keywords is treated as a Boolean OR. Change it to ***“seed dormancy”*** to make it more precise.

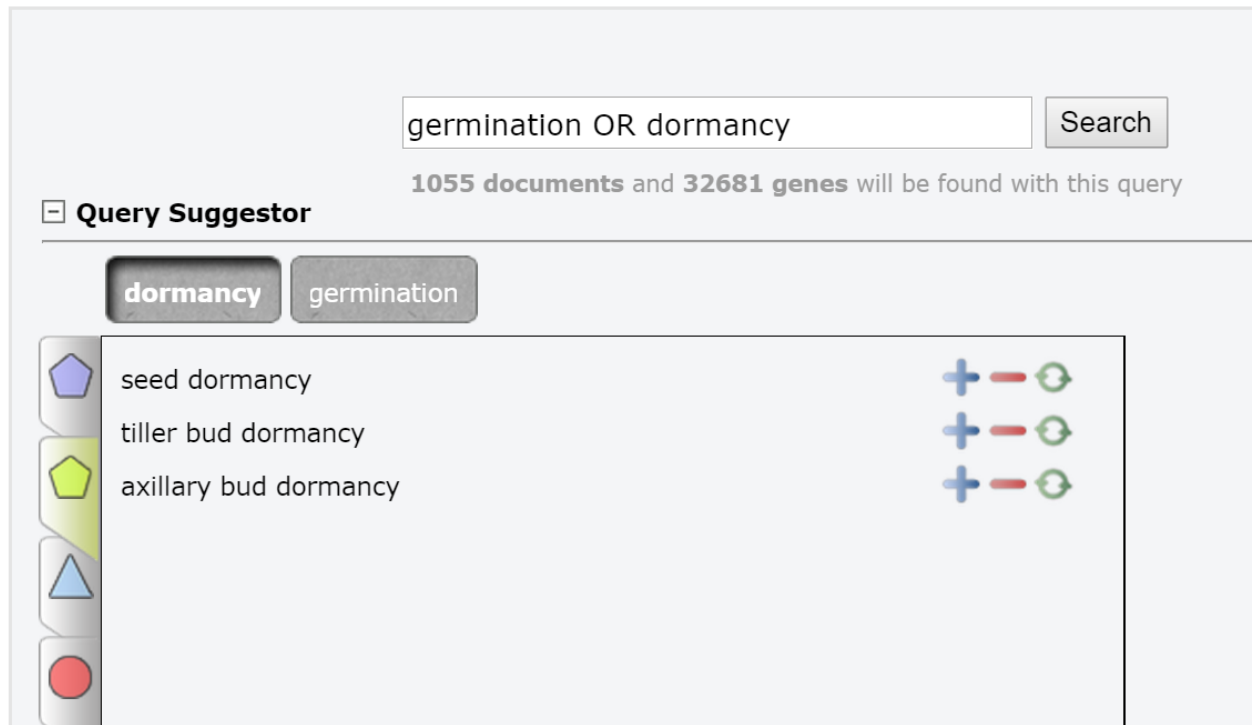


Figure 2: Screenshot of the Query Suggester. The header tabs list the different keywords and the left-hand tabs group the suggestions by evidence types, from top to bottom: Gene Ontology Biological Process, Trait Ontology, gene and protein.

The keyword *“grain color”* matches a TO concept from the wheat knowledge network with the synonyms *bran color* and *pericarp color*. Using *“grain color”* as a keyword, however, would miss many documents and genes that are related to *“seed color”* or other processes that might be influencing grain colour. Therefore, either a boolean operator can be used to search for both keywords *“seed color”* OR *“grain color”*, or the single keyword *“color”* can be used followed by a filter for irrelevant results. Additionally, the colour of the grain is known to be determined through proanthocyanidin (PA) a compound in the flavonoid pathway. These terms can, thus, be included to the grain colour related search terms.

We can search KnetMiner with each trait individually or by combining all search terms. Example search queries are provided on the right of the search box.:

1. Grain colour
 - color OR flavon* OR proanthocyanidin [Example 1]
2. Pre-harvest sprouting (PHS)
 - “seed germination” OR “seed dormancy” [Example 2]
3. Grain colour and PHS
 - dormancy OR color OR flavon* OR proanthocyanidin [Example 3]

Exercise 2: Use the search queries given in Example 1, 2 and 3. Below the search box you can see two numbers. The first number indicates the concepts (documents) in the network that match any of the search terms in any of their attributes. The second number indicates the number of wheat genes that have a biologically meaningful path to these documents.

How many documents are found for each search query?

Why do you think so many genes can be linked to these evidence documents?

Exploring genes supplied by the user

From the 104 differentially expressed wheat genes in red *versus* white grain, KnetMiner identifies 35 genes as being related to grain colour (Example 1) and 27 genes to traits related to germination or dormancy (Example 2). Interestingly, both these sets have 16 genes in common indicating that grain colour and dormancy could be controlled by similar genes. To understand the biological function of these genes and the mechanisms behind these traits, it is essential to analyse the gene-evidence networks which reveal the biological story (in the form of labelled relations) that link the wheat genes to the evidence information.

The Gene View table shows all wheat genes that were found to be related to the search terms. User provided genes are indicated through a “yes” in the *user* column. Sorting the table by this column puts the top scoring user genes at the top of the table even though other genes outside the user's gene list might have higher scores. The various evidence concepts including GO, TO, phenotype, pathway, gene, protein and literature are summarised in the *evidence* column.

The gene scoring function considers all evidence types equally and does not weight one higher than the other. The user, however, might want to look first at genes that have pathway and phenotypic evidence before looking at genes that have mostly publication as their source of evidence. This can currently only be achieved manually by scrolling through the gene list, looking at the evidence type symbols and selecting those genes that have the desired information.

Exercise 3: Use the keywords and gene list provided in Example 3 to perform a search. In the Gene View, select 8 user provided genes that have identical score and evidence information.

Why do you think they have the same score and evidence information?

Which evidence was used to link them to seed colour?

Which evidence was used to link them to seed dormancy?

Hint: Enable labels on TO terms that appear as green pentagon.

A: Figure 3 shows the gene-evidence network of 8 user provided genes that have identical score and evidence information. These genes share the same ortholog (*CHS*) in Arabidopsis. Phenotype data as provided by TAIR (green rectangle) and text-mining based relations (blue edge) reveal that *CHS* loss-of-function mutants show a yellow seed color: “*CHS RNAi plants generated using this method showed yellow seed color and a decrease in anthocyanin content--phenotypes typically observed in*

CHS loss-of-function mutants” (PMID: 19347568). Protein-protein interaction data shows that CHS interacts with DFR, CHI and FLS.

Initially, only those paths from the gene-evidence network are shown where there is a search term and all other concepts are hidden. This effect enables users to focus on the most important information and to expand the network if additional information is required. In the wheat knowledge network, most functional gene information is inferred through homology to Arabidopsis. The homolog itself does not always have direct evidence related to the trait, however, it might physically interact, e.g. based on protein-protein interaction evidence, with genes or proteins that are related. In these cases, KnetMiner exploits indirect information and predicts an involvement in a trait based on guilt-by-association principles.

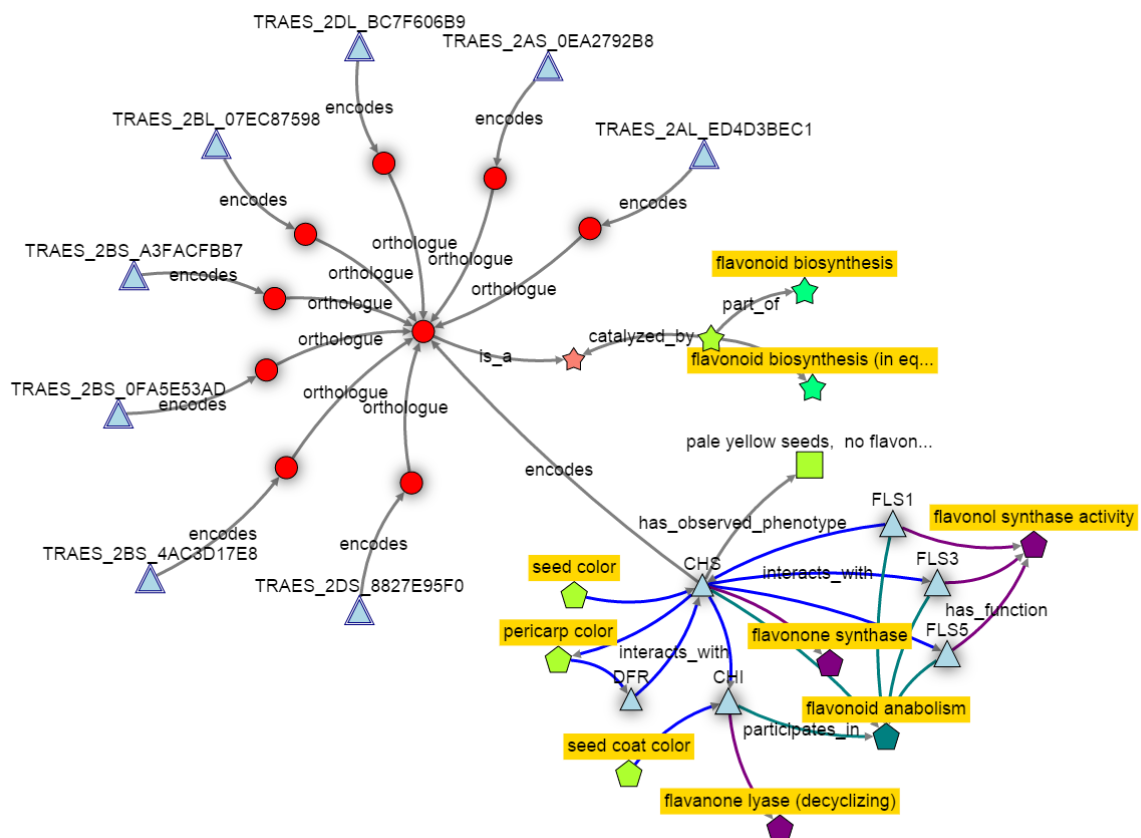


Figure 3: An excerpt from the gene-evidence network of 8 differentially expressed genes in white versus red samples. The ortholog of all 8 genes is the CHS gene from Arabidopsis.

Genes supplied by a user that are associated with the search terms and therefore have evidence documents are referred to as **known targets**, whereas those that are not associated with any search term and thus have nil evidence documents are referred to as **novel targets**. A checkbox at the top of the *Gene View* table allows a user to select all *known targets* or *novel targets* instantly in order to be studied further in the Network View.

The next step is to explore the gene-evidence network of the 16 genes that KnetMiner can relate to both traits grain colour and PHS ([Example 3](#)). In the Gene View, we select the check box **Known targets** and press **Show Network**.

Exercise 4: Which gene in the network is annotated to the GO terms “embryo development ending in seed dormancy” and “positive regulation of flavonoid anabolism”? None of the 16 user genes is orthologous to this gene, so why is it part of the evidence network?

Hint: Enable labels on GO terms that appear as teal pentagon.

A: ARR4 is annotated to the GO terms “embryo development ending in seed dormancy” (GO:0009793) and “positive regulation of flavonoid anabolism” (GO:0009963) based on “Inferred from Mutant Phenotype” (PMID:15634699) and “Inferred from Reviewed Computational Analysis” (PMID:22589469) evidence respectively. None of the differentially expressed wheat genes is directly orthologous to ARR4, however, evidence shows that it interacts with AHP1 (PMID:17545225) and AHP5 (PMID:18642946) which are the orthologs (Ensembl Compara) of **TRAES_4DS_8C9BC2BFA** in wheat. This is one of 37 differentially expressed genes that are higher ($\log_{2}FC = 3.4$) expressed in white grain than in red grain. AHP1 and ARR4 are components of cytokinin signalling network. The involvement of cytokinin in dormancy is usually related to the embryo, not the seed coat, and therefore providing a highly interesting candidate gene. This is only one of many examples that shows how gene-evidence networks produced by KnetMiner can be systematically explored by human domain experts to generate novel leads for follow-up research.

An alternative view for exploring search results

The *Evidence View* offers another way of exploring a user’s gene list. In contrast to the *Gene View*, it provides a document-centric organisation of the results. The columns *GENES* and *USER GENES* count the number of total genes and user-provided genes annotated to the evidence document respectively. This information can be used to calculate which documents are significantly overrepresented for a given set of genes. However, it needs to be noted that only documents that contain the search terms are listed in the *Evidence View*.

Exercise 5: Perform a search with Example 3. Go to the Evidence View tab and sort the table by column TYPE. Find the GO term “*regulation of proanthocyanidin synthesis*” and click on the number 98 in column GENES. This opens a network of all wheat genes related to this GO term.

Find the only wheat gene in the network that is linked to this GO term through an Arabidopsis *TT8* ortholog relation (not PPI or sequence similarity)?

A: The wheat gene **TRAES_1AS_4DA31F550** is a direct ortholog of *TT8* gene in Arabidopsis which is annotated to *regulation of proanthocyanidin synthesis*. All the other wheat orthologs interact with *TT8* or *TT16*.