KnetMiner

USER TUTORIAL

Keywan Hassani-Pak

ROTHAMSTED RESEARCH | 10 NOVEMBER 2017

About KnetMiner

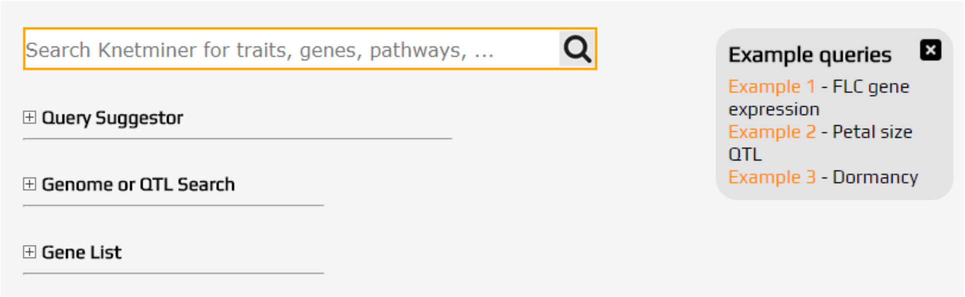
KnetMiner, with a silent "K" and standing for Knowledge Network Miner, is a suite of open-source software tools for integrating and visualising large biological datasets. The software mines the myriad databases that describe an organism's biology to present links between relevant pieces of information, such as genes, biological pathways, phenotypes and publications with the aim to provide leads for scientists who are investigating the molecular basis for a particular trait.

Knowledge networks or graphs provide a perfect data structure for heterogeneous, complex and interconnected biological information and are built using the open-source Ondex data integration platform. A knowledge network consists of labelled nodes, such as a gene, pathway, trait, publication, that are connected through labelled edges, such as encodes, interacts, published-in. Visit our wiki and read Hassani-Pak et al. (2016) to learn how we build knowledge networks.

KnetMiner performs over 70 graph queries of varying depths to find direct or indirect links between genes and user provided search terms. It is very fast using graph databases and graph-indexing techniques.

KnetMiner search interface

The search field of KnetMiner allows users to input any terms related to traits of interest. The terms can be high level descriptions of a phenotypic trait (e.g. disease resistance) or more specific terms such as biological processes and protein families (e.g. defense response to fungi or LRR). Search terms can be combined with OR, AND, NOT statements or put into "" for exact searches.



Search Knetminer for traits, genes, pathways, ...

⊞ Query Suggestor

⊞ Genome or QTL Search

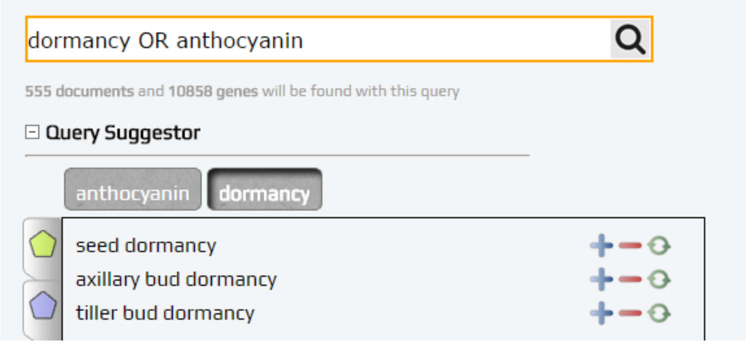
⊞ Gene List

Example queries ☒

Example 1 - FLC gene expression
Example 2 - Petal size QTL
Example 3 - Dormancy

Finding the right search terms

The query suggestion wizard 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 knowledge network.



dormancy OR anthocyanin

555 documents and 10858 genes will be found with this query

⊟ Query Suggestor

anthocyanin | dormancy

seed dormancy + − ⟳
axillary bud dormancy + − ⟳
tiller bud dormancy + − ⟳

Genome or QTL search

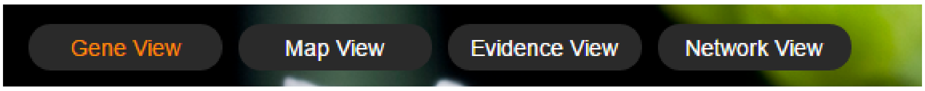
KnetMiner can be used in two modes "whole genome" and "within region". The default is whole genome mode which scores all genes of the genome that can be associated with the query terms and displays the top N (default N: 100). The latter mode (QTL mode) performs the same initial search but retains only genes that fall within the specified region. Genes that were outside the top 100 ranked genes in the in the whole-genome mode, could be inside the top 100 of the QTL mode. Entering the start and end position of a region will display the number of genes within those boundaries. **Note**: Some KnetMiner servers, e.g. wheat, do not have this feature activated yet as they are missing a physical map.

Gene list search

The gene list search allows users to enter a list of gene names or IDs (one per line). KnetMiner searches each gene-network for links to publications, annotations etc. containing the user search terms. The names or IDs need to match (partial match enabled) the gene names/ids stored in the knowledge network. KnetMiner can support different versions of IDs (old and new) if they have been integrated into the knowledge network. If you are not sure which IDs are supported, we suggest to perform a simple search without a gene list and check the Gene View output. The gene list search can cope with up to 100 genes. Searching larger gene lists may still be possible, but be aware that KnetMiner might become slow and unresponsive.

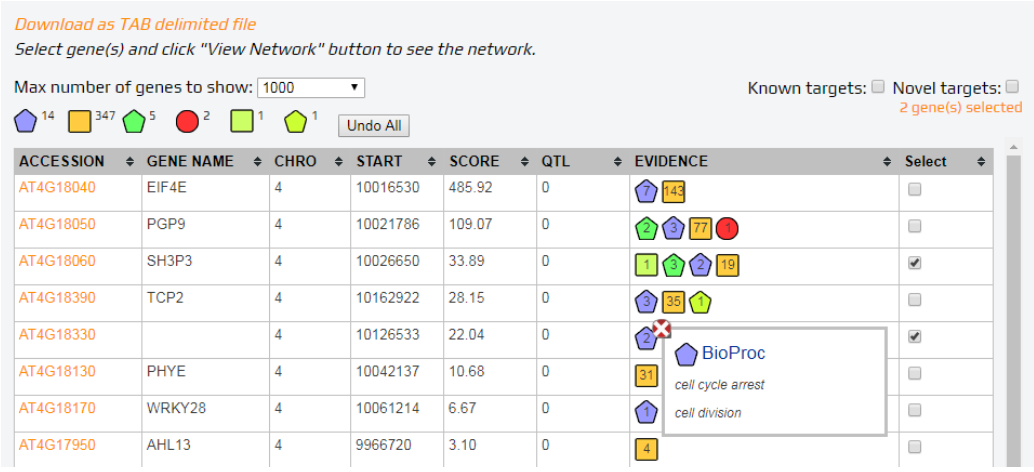
KnetMiner results views

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



Gene View

The *Gene View* uses a table to display up to 1000 candidate genes sorted by the KnetMiner relevance *score*. The various node types (GO, TO, phenotype, pathway, gene, publication etc.) matching the search terms are summarised in the legend. The legend is interactive and can be used as a filter. Clicking on one or multiple symbols in the legend filters the table to genes with matching symbols in the EVIDENCE column, e.g. genes with pathway AND phenotype information. The symbols in the EVIDENCE column are extendible and provide a short description string about the evidence. If the evidence is a publication, then the PubMed id is shown and linked to PubMed. In case of a "gene list search", only these genes are shown in the *Gene View*. Genes supplied by a user that are associated with the search terms are referred to as ***known targets***, whereas those user genes that are not associated with any search term (nil evidence) 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. A counter indicates how many genes have been selected. Clicking on a single gene or on View Network for a selection of genes opens the Network view.



Download as TAB delimited file

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

Max number of genes to show: 1000 ▾

Known targets: □ Novel targets: □

2 gene(s) selected

⬟ 14 ■ 347 ⬟ 5 ● 2 ■ 1 ⬟ 1 Undo All

ACCESSION	GENE NAME	CHRO	START	SCORE	QTL	EVIDENCE	Select
AT4G18040	EIF4E	4	10016530	485.92	0	⬟7 ■143	□
AT4G18050	PGP9	4	10021786	109.07	0	⬟2 ⬟3 ■77 ●1	□
AT4G18060	SH3P3	4	10026650	33.89	0	■1 ⬟3 ⬟2 ■19	☑
AT4G18390	TCP2	4	10162922	28.15	0	⬟3 ■35 ⬟1	□
AT4G18330		4	10126533	22.04	0	⬟2	☑
AT4G18130	PHYE	4	10042137	10.68	0	■31	□
AT4G18170	WRKY28	4	10061214	6.67	0	⬟1	□
AT4G17950	AHL13	4	9966720	3.10	0	■4	□

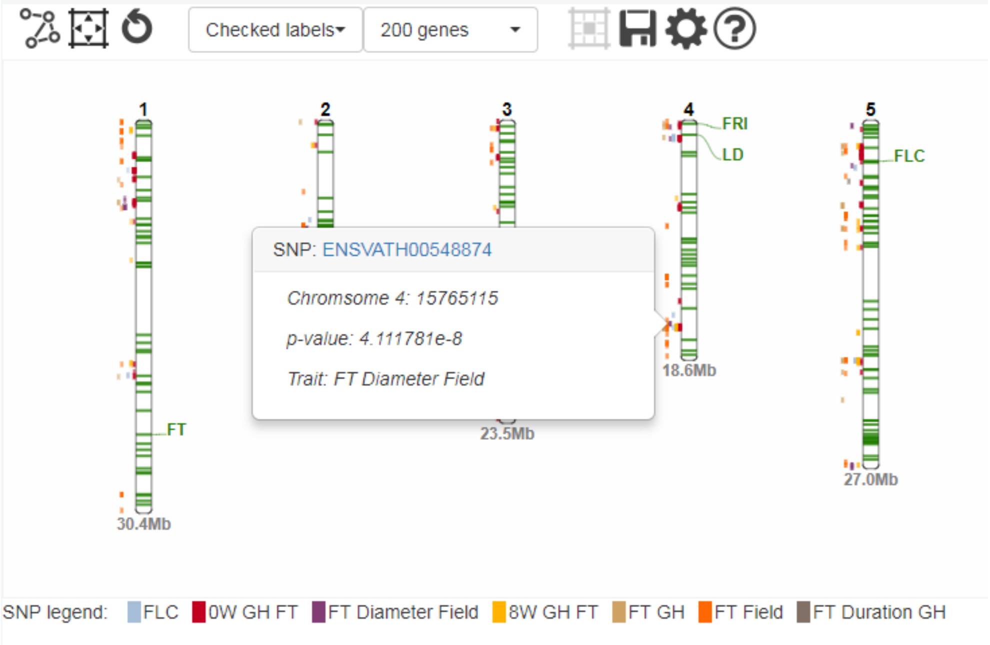
⬟ BioProc

cell cycle arrest

cell division

Map View

The *Map View* is a chromosome based display that shows all genes related to the search terms alongside the chromosomes and uses colour coding to distinguish genes with high (green), medium (orange) and low (red) scores. Integrated QTL and GWAS peaks related to the search terms are displayed on the left-hand side of the chromosome using a different colour for each study (see SNP legend). In case of a "within region" search, only the user-defined QTL and candidate genes that are within the specified loci are displayed. 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. The *Map View* can be exported in PNG format. Several parameters can be adjusted under Settings (e.g. p-value). Genes can be selected in the map and opened in the *Network View* by clicking the network icon (top left).



SNP: ENSVATH00548874

Chromsome 4: 15765115

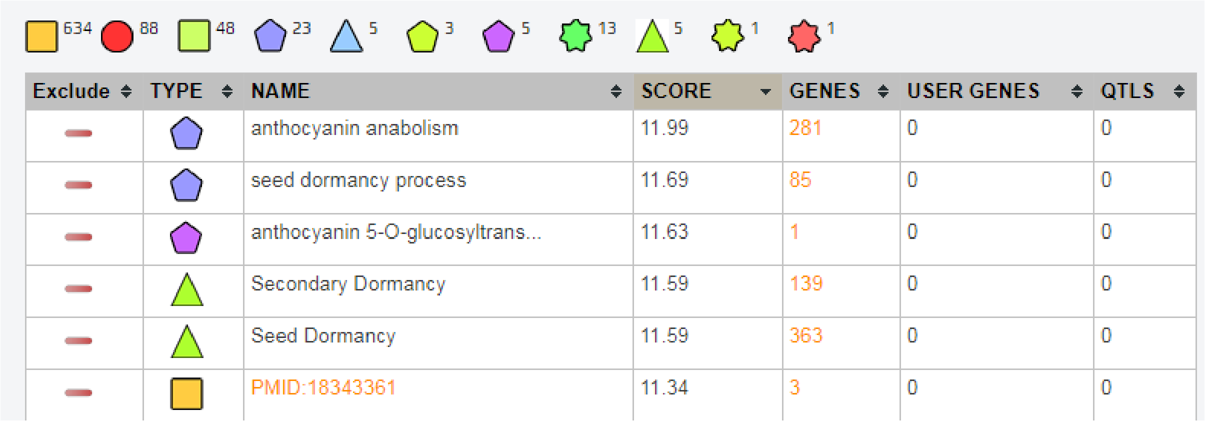
p-value: 4.111781e-8

Trait: FT Diameter Field

SNP legend: ■FLC ■0W GH FT ■FT Diameter Field ■8W GH FT ■FT GH ■FT Field ■FT Duration GH

Evidence View

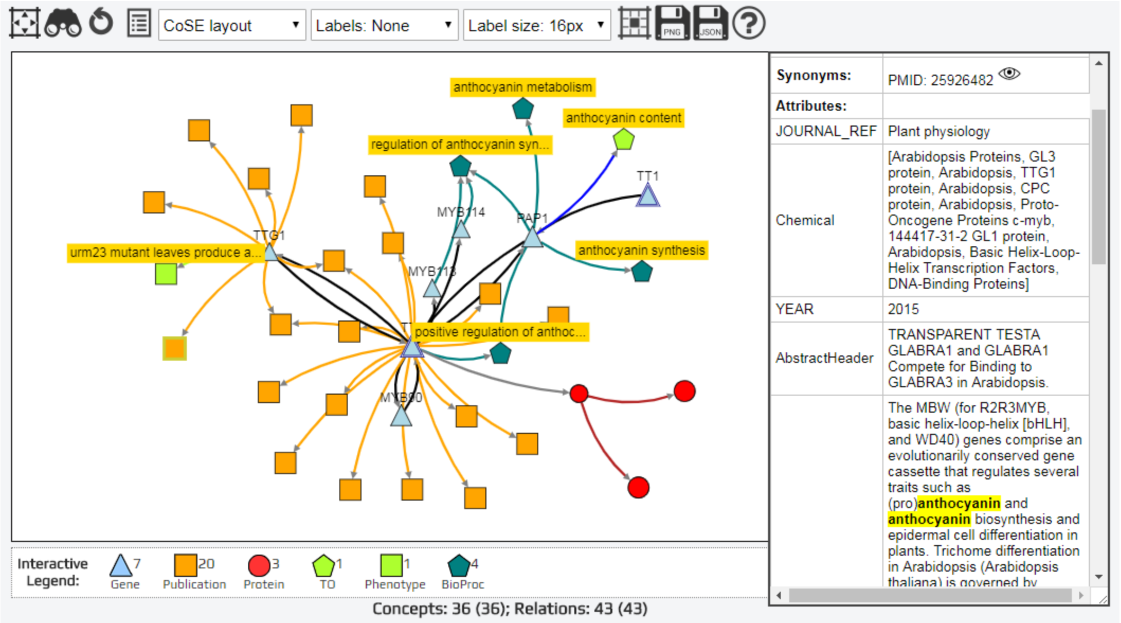
The *Evidence View* provides a document-centric view of the search results sorted by the query-relevance score (Lucene TF-IDF). All documents from the knowledge network containing the query terms are displayed. Click on the symbol in the Exclude column adds a NOT statement with that term to the existing search query. For every document the total number of all and user-provided genes, that are directly or in-directly connected to the document in the network, are displayed. 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.



Exclude	TYPE	NAME	SCORE	GENES	USER GENES	QTLS
−	⬟	anthocyanin anabolism	11.99	281	0	0
−	⬟	seed dormancy process	11.69	85	0	0
−	⬟	anthocyanin 5-O-glucosyltrans...	11.63	1	0	0
−	△	Secondary Dormancy	11.59	139	0	0
−	△	Seed Dormancy	11.59	363	0	0
−	🟨	PMID:18343361	11.34	3	0	0

Network View

The *Network View* displays knowledge networks of one or multiple genes selected in one of the previous views. The entry gene is displayed as a blue triangle with a double border. The legend explains the type of the other nodes. Every path starting from the entry gene and going to an evidence node provides a clue. Initially only the most important clues are displayed to the user. A user can interactively explore and extend the network with further clues. The Info Box panel can be opened to display all properties (e.g. name, description, title, abstract etc.) attached to nodes and edges. The *Map View* provides a right-click circular context menu on nodes and edges that allow a user to hide information or explore and expand the network with additional information. Users can also re-layout the network using a variety of layouts and use the interactive legend to add hidden, type-specific nodes to the visible network. Networks can be exported in JSON compatible with Cytoscape Desktop and as PNG images.



CoSE layout

Labels: None

Label size: 16px

Synonyms:	PMID: 25926482
Attributes:	
JOURNAL_REF	Plant physiology
Chemical	[Arabidopsis Proteins, GL3 protein, Arabidopsis, TTG1 protein, Arabidopsis, CPC protein, Arabidopsis, Proto-Oncogene Proteins c-myb, 144417-31-2 GL1 protein, Arabidopsis, Basic Helix-Loop-Helix Transcription Factors, DNA-Binding Proteins]
YEAR	2015
AbstractHeader	TRANSPARENT TESTA GLABRA1 and GLABRA1 Compete for Binding to GLABRA3 in Arabidopsis.
	The MBW (for R2R3MYB, basic helix-loop-helix [bHLH], and WD40) genes comprise an evolutionarily conserved gene cassette that regulates several traits such as (pro)**anthocyanin** and **anthocyanin** biosynthesis and epidermal cell differentiation in plants. Trichome differentiation in Arabidopsis (Arabidopsis thaliana) is governed by

Concepts: 36 (36); Relations: 43 (43)

KnetMiner Use Case

This application case shows the utility of KnetMiner for the functional analysis of a transcriptomics (RNA-seq) experiment in bread wheat (*Triticum aestivum*). Wheat is the third most-grown cereal crop in the world after maize and rice, and has a hexaploid genome 5 times the size of the human genome. The red colour of the grain is due to the presence of coloured compounds, called flavonoids, in the seed coat. White-grained wheat varieties can be bred that lack the red compounds of the seed coat. However, white grains are prone to germinate before harvest, a particular problem in countries such as the UK where cool, wet weather before harvest is common. This "pre-harvest sprouting" or PHS, results in a loss of grain quality and even a small proportion of sprouted grains can result a serious loss of value for the crop.

A transcriptome (RNA-seq) experiment was designed by the Phillips lab at Rothamsted to understand the transcriptional differences between red and white grains. The RNA-seq reads were mapped to the wheat reference genome (TGACv1) using HiSat. Transcript abundance was estimated using featureCount and differentially expressed genes identified with DeSeq2. In total **188** genes were differentially expressed ($p<0.05$) between red and white grain (inner endosperm) of which 44 have logFC>2.

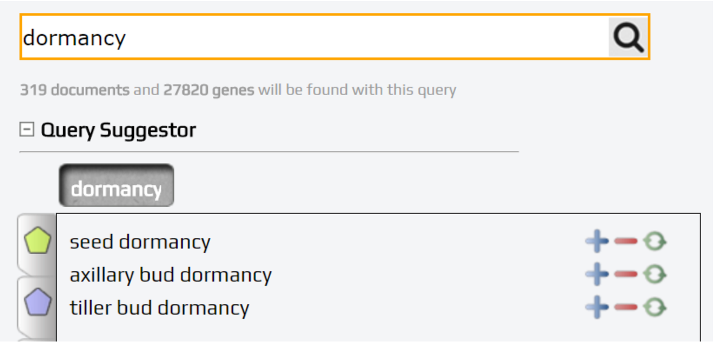
Having identified a list of differentially expressed genes (DEG), the questions scientists would consequently ask are:

- 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?

Exercise 1 - 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 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 **Query Suggester** provides alternative synonyms or more specific keywords. It also highlights key concept types that match the keywords.



dormancy

319 documents and 27820 genes will be found with this query

⊟ Query Suggestor

dormancy

seed dormancy

axillary bud dormancy

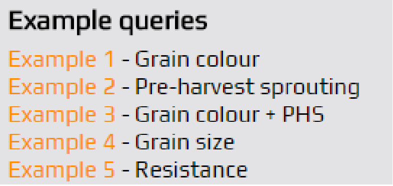
tiller bud dormancy

Task: 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. 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 like the ones involved in seed dormancy but different to bud dormancy. This returns more genes since the space between to keywords is treated as a Boolean OR. Change it to ***"seed dormancy"*** to make it more precise.

Exercise 2 - Exploring genes supplied by the user

We are now going to explore 44 differentially expressed wheat genes in red *versus* white grain ($p<0.05$, logFC>2) in the context of grain colour and PHS traits. The Wheat-KnetMiner has following three example queries that we are going to use:



Example queries

Example 1 - Grain colour
Example 2 - Pre-harvest sprouting
Example 3 - Grain colour + PHS
Example 4 - Grain size
Example 5 - Resistance

Example 1 - Grain colour

- color OR flavon* OR proanthocyanidin

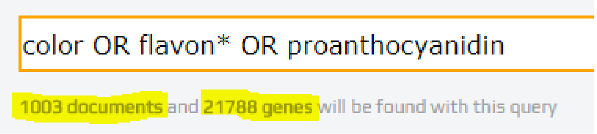
Example 2 - Pre-harvest sprouting (PHS)

- "seed germination" OR "seed dormancy"

Example 3 - Grain colour + PHS

- dormancy OR germination OR color OR flavon* OR proanthocyanidin

Task: Example 1 populates the search box with search terms and the Gene List with ids. Below the search box you can see two numbers. The first number indicates the nodes in the wheat knowledge network that match the search terms. The second number indicates the number of genes in the wheat genome that have direct or indirect links to these search terms. Press the search button.



color OR flavon* OR proanthocyanidin

1003 documents and 21788 genes will be found with this query

In Gene View, you can use the interactive legend and click on one or more symbols, e.g. pathways, phenotype etc. This will retain genes with the selected evidence types and filter other genes.

→ Which genes are part of the flavonoid biosynthesis pathway?

User genes that were not associated with the search terms appear in Gene View with a "0" in the EVIDENCE column.

→ How many user provided genes have known links to Example 1 search terms (known targets) and how many genes have no obvious links (novel targets)?

Repeat these steps for Example 2 – PHS.

Exercise 3 - Exploring gene-evidence networks

We are now going to select single or multiple genes and explore their gene-evidence networks, i.e. the information that links the genes with the search terms.

Task: Use Example 3 to perform a search. In the Gene View, select all user provided genes with the name CHS. Click on View Network.

→ Can you find the wheat genes (blue triangles with a double border) in the network and follow the path to the Arabidopsis ortholog? Where is the ortholog relation coming from?

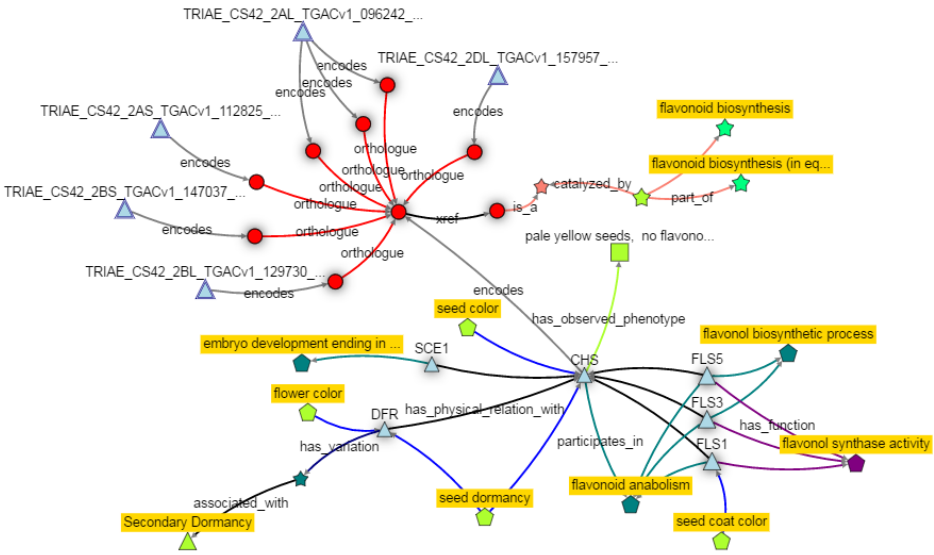
→ Which evidence was used to link the wheat CHS genes to seed colour?

→ Which evidence was used to link the wheat CHS genes to seed dormancy?

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

A: The figure below shows the gene-evidence network of **5** 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, SCI1 and FLS (1, 3, 5) genes.

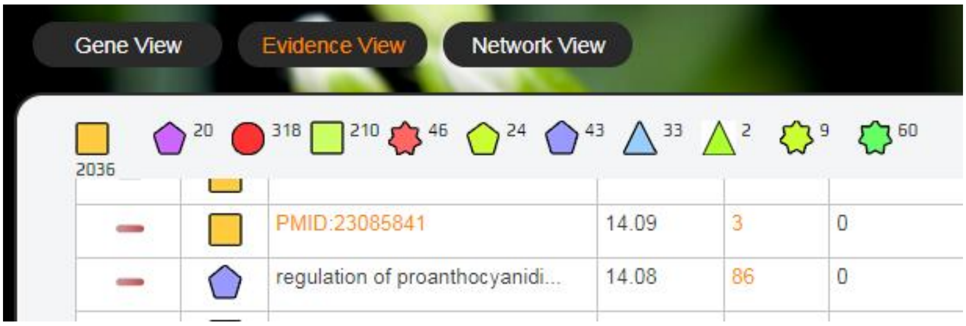
Initially, only those paths from the gene-evidence network are shown that link to the search terms and all other paths are hidden. This helps users to focus on the most important information first and to expand the network if additional information is required. In the wheat knowledge network, most functional 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 to predicts an involvement in a trait.



An excerpt from the gene-evidence network of 5 differentially expressed genes in white versus red samples. The ortholog of all 5 genes is the CHS gene from Arabidopsis.

Exercise 4 – Exploring all genes linked to a biological process or pathway

The *Evidence View* offers another way of exploring the knowledge network. The table displays all nodes in the knowledge network that contain the search terms. The columns *GENES* and *USER GENES* count the number of total genes and user-provided genes connected to the evidence nodes respectively.



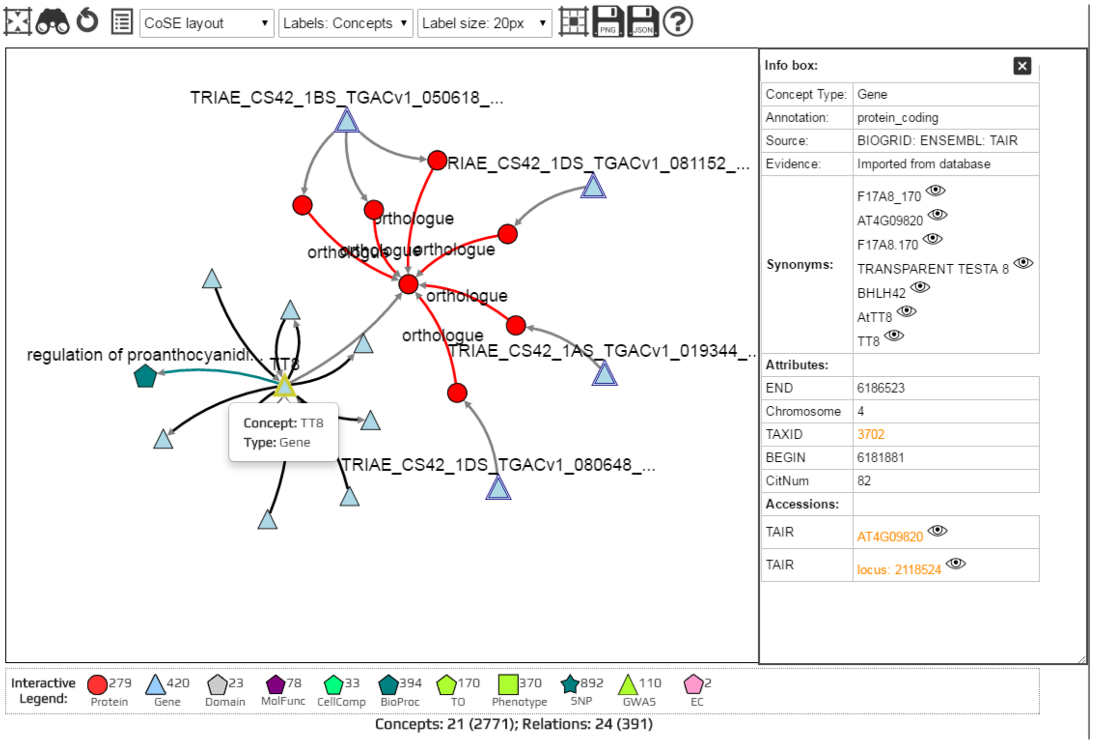
—	🟨	PMID:23085841	14.09	3	0
—	⬟	regulation of proanthocyanidi...	14.08	86	0

Task: 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*" (shown as a *Biological Process*) and click on the number **98** in column GENES. This will open a network of all wheat genes related to this GO term.

Hint: BioProc concepts are shown as **purple pentagon**.

→ Find all wheat genes in the network that are linked to this GO term through an Arabidopsis *TT8* ortholog relation (not PPI or sequence similarity)?

A: There are 4 wheat ***orthologs*** of ***TT8*** gene in Arabidopsis which is annotated to *regulation of proanthocyanidin synthesis*. All the other wheat orthologs *interact with* TT8 or TT16.


CoSE layout
Labels: Concepts
Label size: 20px

TRIAE_CS42_1BS_TGACv1_050618_...
RIAE_CS42_1DS_TGACv1_081152_...
rthologue
orthologue
orthologue
orthologue
orthologue
orthologue
orthologue
RIAE_CS42_1AS_TGACv1_019344_...
regulation of proanthocyanidi...
TT8
TRIAE_CS42_1DS_TGACv1_080648_...

Concept: TT8
Type: Gene

Info box:

Concept Type:	Gene
Annotation:	protein_coding
Source:	BIOGRID: ENSEMBL: TAIR
Evidence:	Imported from database
Synonyms:	F17A8_170 AT4G09820 F17A8.170 TRANSPARENT TESTA 8 BHLH42 AtTT8 TT8
Attributes:	
END	6186523
Chromosome	4
TAXID	3702
BEGIN	6181881
CitNum	82
Accessions:	
TAIR	AT4G09820
TAIR	locus: 2118524

Concepts: 21 (2771); Relations: 24 (391)