

KnetMiner v2.1

USER TUTORIAL

Keywan Hassani-Pak

ROTHAMSTED RESEARCH | 08 NOVEMBER 2018

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 of 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 Oindex 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.

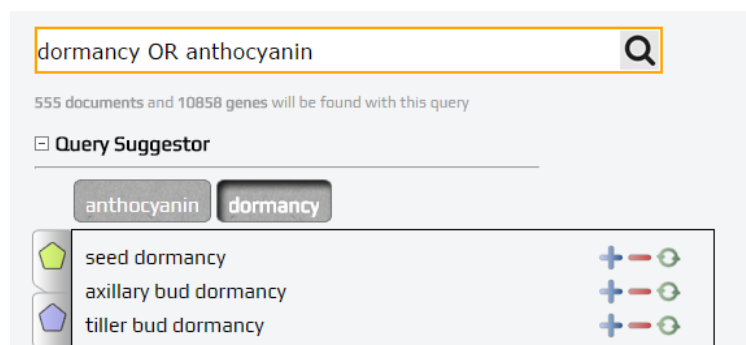
KnetMiner search interface

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. heat tolerance) 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.

Additionally, a feedback mechanism was implemented that constantly returns the number of resulting documents and genes while the user is typing the query. This feature is activated once the query term is at least 3 characters long and is updated at each additional keyboard event. This feature provides several benefits to users: 1) helps to detect spelling mistakes, 2) gives a hint if the query term is too general or too specific before the user executes the search and 3) motivates the user to examine their query and explore different spelling, language or more complex query statements (AND, OR, NOT, "").

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 water deprivation'. 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.

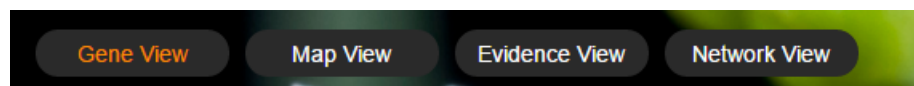


Genome, region and gene list search

KnetMiner can be used to search the whole genome, specific regions or a gene list. The default mode is whole genome search which scores all genes of the genome by their association with the query. The region 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 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. The gene list search allows users to enter a list of gene names or accessions. The names or accessions need to match (partial match enabled) the gene names/ids stored in the knowledge network. Different versions of gene accessions (old and new) can be used if they exist as such in the knowledge graph. KnetMiner will show which of these genes can be linked to the search terms using the knowledge graph.

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 identified candidate genes sorted by the KnetScore. 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 of 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

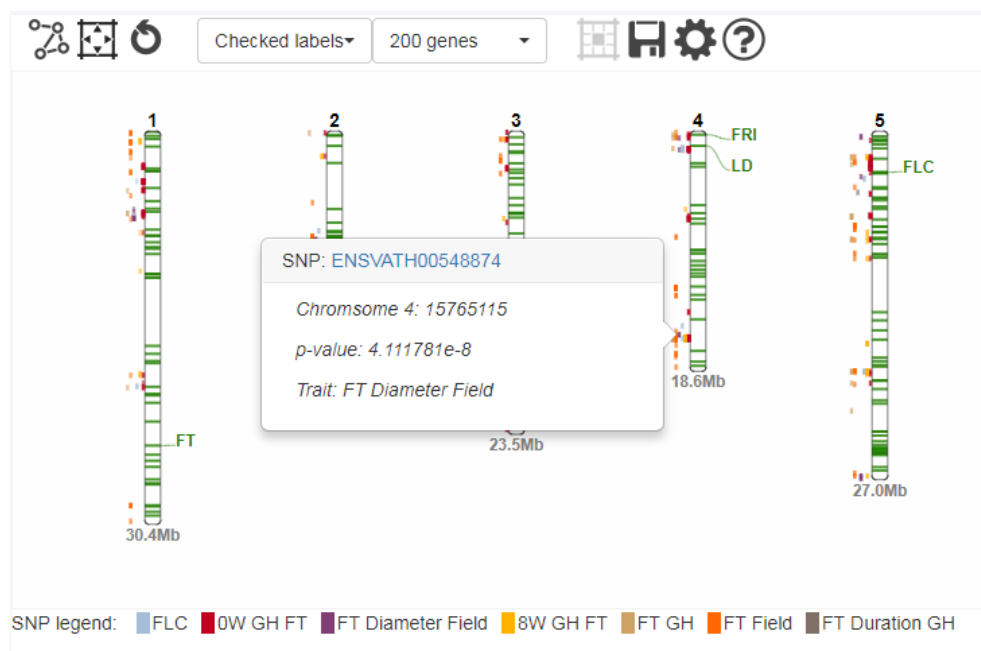
Known targets: Novel targets:
2 gene(s) selected

14 347 5 2 1 1 Undo All

ACCESSION	GENE NAME	CHRO	START	SCORE	QTL	EVIDENCE	Select
AT4G18040	EIF4E	4	10016530	485.92	0	43	<input type="checkbox"/>
AT4G18050	PGP9	4	10021786	109.07	0	77	<input type="checkbox"/>
AT4G18060	SH3P3	4	10026650	33.89	0	1 19	<input checked="" type="checkbox"/>
AT4G18390	TCP2	4	10162922	28.15	0	35	<input type="checkbox"/>
AT4G18330		4	10126533	22.04	0	BioProc	<input checked="" type="checkbox"/>
AT4G18130	PHYE	4	10042137	10.68	0	31 cell cycle arrest	<input type="checkbox"/>
AT4G18170	WRKY28	4	10061214	6.67	0	cell division	<input type="checkbox"/>
AT4G17950	AHL13	4	9966720	3.10	0	4	<input type="checkbox"/>

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 information 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 as summarized in the SNP legend. In case of a region search, only the user-defined regions and candidate genes within these regions 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).



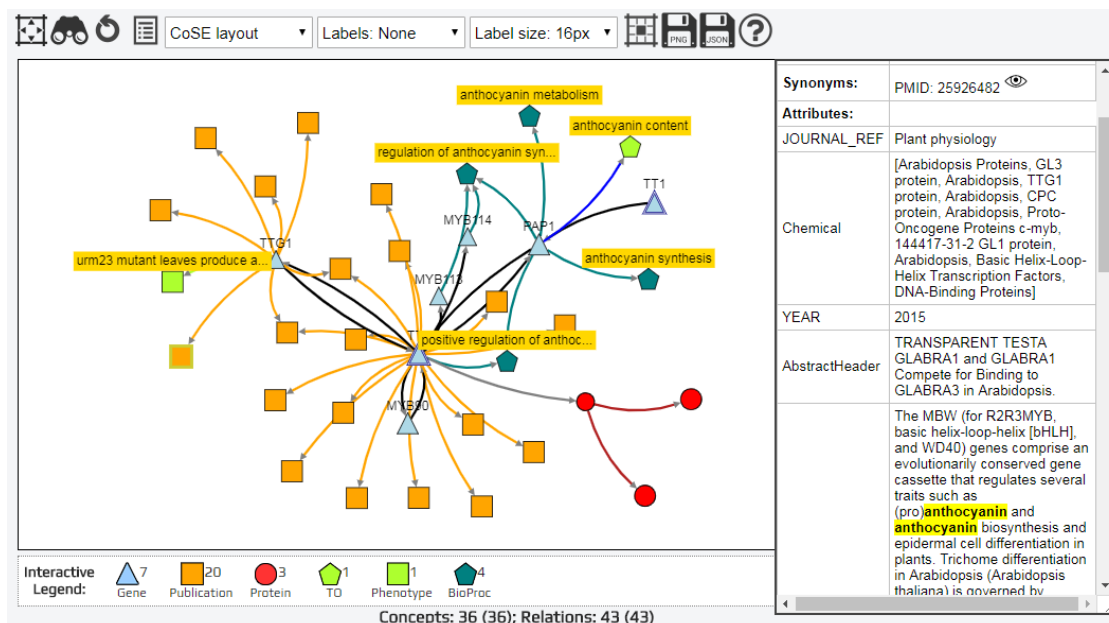
Evidence View

The *Evidence View* provides a document-centric view of the search results sorted by the query-relevance score. All nodes (documents) from the knowledge graph containing the query terms are displayed. A function is provided to exclude a specific document from the existing search by appending a NOT statement with the document's identifier. For every document, the number of genes linked to it in the knowledge graph 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 selected document concept in the centre of the network and all shortest paths that connect the document to the linked genes.

Exclude	TYPE	NAME	SCORE	GENES	USER GENES	QTLs
-	Gene	anthocyanin anabolism	11.99	281	0	0
-	Gene	seed dormancy process	11.69	85	0	0
-	Gene	anthocyanin 5-O-glucosyltrans...	11.63	1	0	0
-	Phenotype	Secondary Dormancy	11.59	139	0	0
-	Phenotype	Seed Dormancy	11.59	363	0	0
-	Publication	PMID:18343361	11.34	3	0	0

Network View

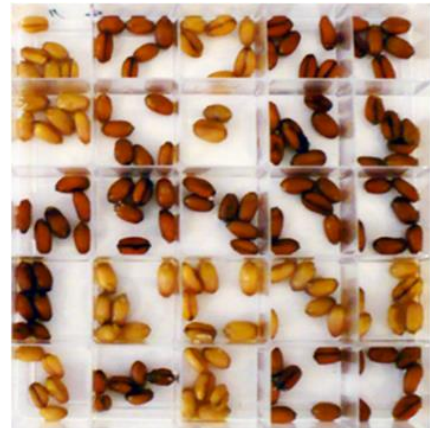
The *Network View* displays knowledge networks of one or multiple genes selected in one of the previous views. The entry gene is displayed as a blue triangle with a **double border**. Every path starting from the entry gene and going to an evidence node provides a clue. Initially only the most important clues are displayed to the user. A user can interactively explore and extend the network with further clues. This is achieved by annotating (server-side) all nodes and edges in a gene knowledge graph based on their relevance to the search query. When the network is initially visualised, only those nodes and edges are shown that were set to be visible, and a shadow effect is added to nodes that can be further expanded. This effect enables users to focus on the most important information and to expand the network if additional information is required. Additionally, the node size itself is increased to visually distinguish query related nodes from unrelated nodes. A circular context menu on nodes and edges allows users to hide or expand concepts and relations of certain types. An interactive legend allows users to add hidden, type-specific nodes to the visible network. Networks can be exported in JSON compatible with Cytoscape Desktop and as PNG images.



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 (bran). These flavonoids give wholemeal bread not only its colour, but also a slightly bitter taste which is disliked by many people. White-grained wheat varieties lack the red compounds of the seed coat and are milder in flavor. 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 and that the red pigmentation of wheat grain is controlled by R genes on the long arms of chromosomes 3A, 3B and 3D. In the last decade, the genetic basis of the relationship between grain colour and PHS has been studied and molecular characterisation showed the R gene is a Myb-type transcription factor responsible for transcriptional activation of genes (CHS, CHI, F3H and DFR) in the flavonoid biosynthesis pathway. However, the link between the R (Myb) gene and PHS is still unclear.

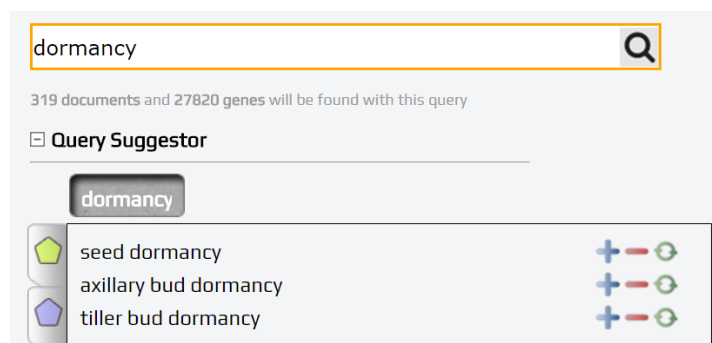


Here we demonstrate the utility of KnetMiner for analysing candidate genes from reverse genetics or transcriptomics studies and answering questions such as:

1. Do any of these genes contribute to the expression of trait A (e.g. grain colour)?
2. Do any of these genes contribute to the expression of trait B (e.g. PHS trait)?
3. Which biological processes and pathways are underlying these traits?
4. Are there common genes or mechanisms that influence both traits?
5. Which other processes and traits will be affected by loss-of-function mutants?

Exercise 1 - Choosing the right search terms

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

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 knowledge 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: Click on Example 3 and perform a search. In Gene View, focus on the *TT2* gene (TRAESCS3D01G468400). Check its evidence column and click on it to see the 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 path was used to link the wheat *TT2* gene to pericarp and seed color?

→ Which evidence path was used to link the wheat *TT2* gene to seed dormancy or grain germination?

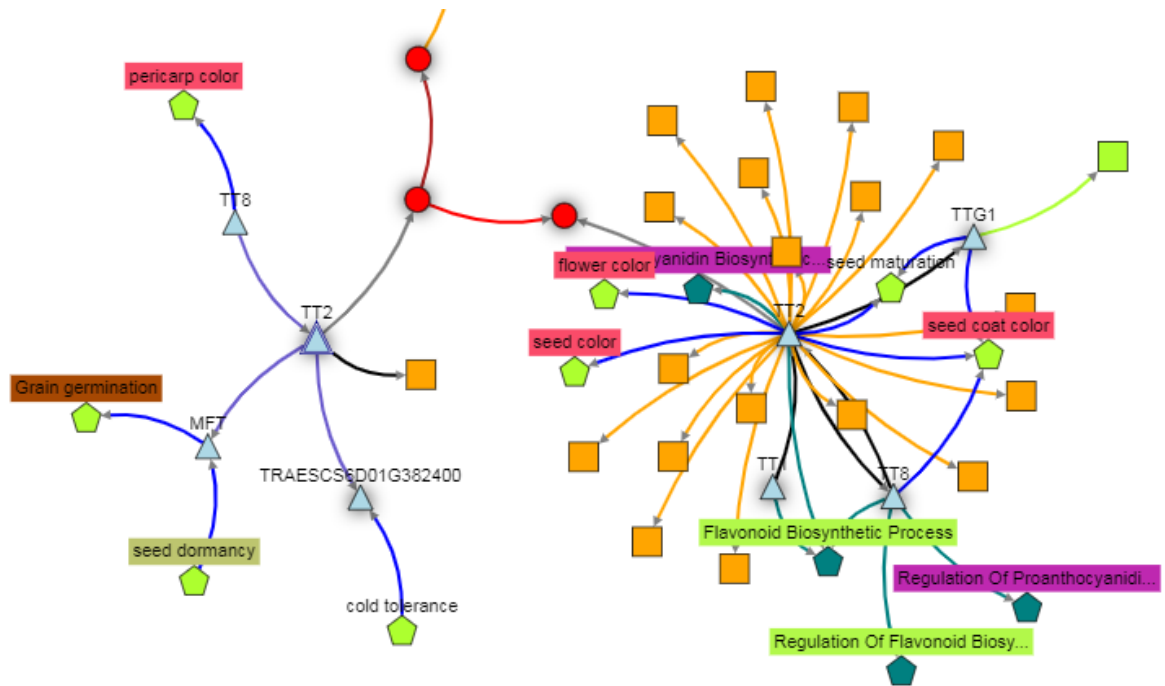
→ Which other traits can be affected by *TT2* loss-of-function mutants?

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

A: The knowledge network of wheat *TT2* (R MYB) contains gene regulatory information, protein-protein interactions, phenotypic information in the form of mutant/genetic studies or text-mining, links to relevant ontology terms and publications, and, similar information from Arabidopsis and other species. A more thorough exploration of the information (i.e. node and edge properties) captured in *TT2* network tells the following detailed biological story:

*TT2 (R Myb) on chromosome 3D in wheat is predicted (p -value=0.01) to regulate the transcriptional activation of MFT according to data from the analysis of 850 RNA-seq samples in wheat using GENIE3. The *TT2 3B* homeologue is not predicted to regulate MFT, and the *TT2 3A* homeologue is not annotated in the latest version of the wheat genome, MFT has been recently linked to grain germination [“Recent studies in both Arabidopsis and wheat have uncovered a new role of MOTHER OF FT AND TFL1 (MFT) in seed germination”] and seed dormancy [Mapping analysis showed that MFT on chromosome 3A (MFT-3A) colocalized with the seed dormancy quantitative trait locus (QTL) QPhs.ocs-3A.]. The MFT ortholog in Arabidopsis has a 3' UTR variant that has been associated with (p -value=5.5x10⁻⁵) increased germination rate after 56 days of dry storage.*

To discover which other traits will be affected by *TT2* loss-of-function mutants, with a simple click, the user expands the initial *TT2* knowledge graph to add all other genes that are regulated by or interact with *TT2*. Other wheat genes regulated by *TT2* do not show any surprising phenotypes, however, the Arabidopsis *TT2* interacts with *TTG1* - a gene controlling root hair density and root hair length in Arabidopsis root hairs are tubular outgrowths from specific epidermal cells that function in nutrient and water absorption. This interesting clue, provides an immediate idea to phenotype *TT2* knock-outs also under the ground for root related traits, and poses a bigger, more speculative hypothesis that pre-harvest sprouting could be caused by increased root hairs due to higher nutrient and water absorption.



The gene knowledge network of the TT2 (R Myb) gene.

Exercise 4 - Exploring Map View

- Click on Example 5
- Press Search
- Go to Map View: Explore chromosome, genes, SNP, QTL information
- Select ADA2A on Chr5A and BPM4 on Chr2D
- Click on the Network button (top left)
- Explore the network
- Add more information using the interactive legend

