Dissertation

An integrated data platform for gene mining and biological knowledge discovery

Keywan Hassani-Pak

2017

Dissertation submitted to the faculty of technology at Bielefeld University to obtain the degree of Doctor of Science (Dr. rer. nat.).

Title

KnetMiner - An integrated data platform for gene mining and biological knowledge discovery

Author

Keywan Hassani-Pak
Rothamsted Research
AL5 2JQ, UK
keywan.hassani-pak@rothamsted.ac.uk

Supervisors

Prof. Dr. Ralf Hofestädt, Bielefeld University, Germany
Prof. Dr. Christopher Rawlings, Rothamsted Research, UK

Dates

Submitted: January 2nd, 2017
Doctoral exam: May 4th, 2017

Printed on non-ageing paper (ISO 9706)

Data isn't information. Information, unlike data, is useful. While there's a gulf between data and information, there's a wide ocean between information and knowledge. What turns the gears in our brains isn't information, but ideas, inventions, and inspiration. Knowledge - not information - implies understanding. And beyond knowledge lies what we should be seeking: wisdom.

Clifford Stoll

In High-Tech Heretic: Reflections of a Computer Contrarian (2000), 185-186

Abstract

Discovery of novel genes that control important phenotypes and diseases is one of the key challenges in biological sciences. Now, in the post-genomics era, scientists have access to a vast range of genomes, genotypes, phenotypes and ‘omics data which - when used systematically - can help to gain new insights and make faster discoveries. However, the volume and diversity of such un-integrated data is often seen as a burden that only those with specialist bioinformatics skills, but often only minimal specialist biological knowledge, can penetrate. Therefore, new tools are required to allow researchers to connect, explore and compare large-scale datasets to identify the genes and pathways that control important phenotypes and diseases in plants, animals and humans.

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. The KnetMiner approach is based on 1) integration of heterogeneous, complex and interconnected biological information into a knowledge graph; 2) text-mining to enrich the knowledge graph with novel relations extracted from literature; 3) graph queries of varying depths to find paths between genes and evidence nodes; 4) evidence-based gene rank algorithm that combines graph and information theory; 5) fast search and interactive knowledge visualisation techniques. Overall, KnetMiner is a publicly available resource (http://knetminer.rothamsted.ac.uk) that helps scientists trawl diverse biological databases for clues to design better crop varieties and understand diseases. The key strength of KnetMiner is to include the end user into the “interactive” knowledge discovery process with the goal of supporting human intelligence with machine intelligence.

Acknowledgements

This work would not have been possible without the support of several people.

First, and foremost, I would like to thank my supervisor Prof. Chris Rawlings at Rothamsted Research for his support during the Ondex and KnetMiner projects, and for guiding me to be a creative and independent scientist. I would also like to show my greatest appreciation to Prof. Ralf Hofestädt at University of Bielefeld for giving me the opportunity to do an independent PhD ("freie promotion") with him and examining my work. I'm very grateful to Prof. Achim Dobermann, Director of Rothamsted Research, for his warm encouragement to complete my PhD journey by granting me permission to take a sabbatical to write my dissertation.

I want to thank my colleagues Ajit Singh, Marco Brandizi, Martin Castellote, Maria Esch, Fengyuan Hu, Minja Zorc, Mike Phillips, Monika Mistry and Lisa Lill for their contributions to the KnetMiner project from 2011 to 2017. I enjoyed working with everyone of you on this exciting project. I'm also grateful to Jan Taubert, Matthew Hindle, Artem Lysenko, Catherine Canevet and all other Ondex team members, from 2008 to 2011, for the development of the Ondex software which provided the foundation of my work. I also appreciate the help of the Rothamsted Bioinformatics team for proofreading my thesis and providing valuable comments.

Finally, my very special thanks goes to my wonderful family - my wife, my children Mehdi, Omid and Mia; my mum, my step dad who sadly passed away in 2016, and my parents in law - for their love and continuous support.

Contents

List of Abbreviations

AHD	Arabidopsis Hormone Database
AMPRIL	Arabidopsis multiparent recombinant inbred lines
API	Application programming interfaces
CO	Crop Ontology
CPM	Counts per million
DEG	Differentially expressed genes
DOI	Digital Object Identifier
ECO	Evidence and Conclusion Ontology
EDF	Evidence Document Frequency
ES	Evidence sentences
EVA	European Variation Archive
FC	Fold change
GAF	Gene Association Format
GO	Gene Ontology
GSKN	Genome-scale knowledge network
GWAS	Genome-wide association studies
IDF	Inverse document frequency
IGF	Inverse Gene Frequency
IP	Inner product
IR	Information retrieval
LD	Linkage disequilibrium
MAGIC	Multiparent Advanced Generation Inter-Cross
MeSH	Medical Subject Headings
MGQE	Metadata-based Graph Query Engine
NER	Named Entity Recognition
PHS	Pre-harvest sprouting
POM	Project object model
PPI	Protein interaction
QTL	Quantitative trait loci
RAM	Random-access memory
REML	REsidual Maximum Likelihood
SNP	Single nucleotide polymorphisms

TAIR	The Arabidopsis Information Resource
TO	Trait Ontology

1 INTRODUCTION

The development of improved agricultural crops is a critical societal challenge, given current global developments such as population growth, climate and environmental change, and the increasingly scarcity of inputs (fuel, fertilizer, etc.) needed for agricultural productivity. To meet this challenge, we will need to design improved crop varieties, with higher yields, robustness to biotic (e.g. pathogens, pests) and abiotic shock. Furthermore, there is a need to accelerate the breeding programmes needed to implement these designs. The use of forward genetics, reverse genetics and “omics” technologies to understand genotype-phenotype relationships will be critical to achieving this goal.

In the recent past, during the genomics era, scientists developed technologies to sequence and assemble the chromosomes of an organism and predict the gene content. Now, in the post-genomic era, next generation sequencing technologies have been developed and this has led to an explosion of more genomic data alongside a wealth of gene expression, protein expression, genetic and biological data, which are used by scientists to decipher the complex human, animal and plant systems and understand the molecular basis of phenotypes and disease conditions. The interpretation of such data has considerable potential as an adjunct to plant and animal breeding, however, it is not yet easy to interrogate these data and obtain clear, objective answers that can be applied in practice. For many scientists with expertise in biology, biochemistry or genetics, this “omics” data explosion is often seen as a challenge that only those with specialist bioinformatics or data analytics skills, but often only minimal specialist biological knowledge, can penetrate. Therefore, new high-quality tools for data integration and interpretation urgently need to be developed to allow researchers to connect, explore and compare the relevant large and small-scale datasets available for many species. Once we fully understand how biomedical or agronomic phenotypes are regulated and how diseases emerge, it should be possible to manipulate these processes and mechanisms and go on to devise new ways to improve crop and animal productivity and reduce disease levels and thereby improve human health and global food security.

Genetics and ‘omics studies designed to identify gene-phenotype relationships often identify large numbers of potential candidate genes. At some stage, every scientist will need to choose which genes to investigate further in the lab. Often, this choice is done subjectively, based on hunches or (potentially selective) prior experience and generally without a robust

scientific justification. Data-driven systematic methods that search and filter the wealth of available data and evidence in order to objectively prioritize candidate genes based on validated algorithms will be of great value to life science researchers. Such methods and tools will save them valuable time and help to provide an evidence-based justification for why certain genes were considered and others not.

The objective of this PhD project was to develop a tool that will allow researchers without specialist bioinformatics skills to explore and compare the wealth of existing open-access data from multiple species with their own experimental results in order to identify gene-trait relationships through the exploration of biological databases. An approach was taken that effectively connects heterogeneous information types, mines the information and then returns the results in an accessible, explorable, as well as scalable, format that can be easily manipulated, displayed and interrogated. The aim was to create a novel *in silico* environment from which new scientific insights and biological discoveries can be made. The resulting software is called KnetMiner - Knowledge Network Miner. Knowledge networks or graphs provide a perfect data structure for heterogeneous, complex and interconnected biological information and consist of labelled nodes, such as a gene, pathway, trait, publication, that are connected through labelled edges, such as encodes, interacts, published-in. KnetMiner was developed in stages to address the three overarching challenges i.e. biological data integration, gene mining and knowledge discovery. The KnetMiner software and the knowledge resources are freely available and provide a first step towards systematic and evidence-based gene discovery in order to facilitate crop improvement. The chapters of this thesis will describe the development and application of KnetMiner.

1.1 Overview

Chapter 2 gives an introduction to the techniques used by biologists and breeders to link phenotype to gene(s). The main focus is on the accuracy of each method in regard to the number of potential candidate genes that they may reveal. I explain why complementary computational methods are needed to accelerate the identification of causal genes and describe the types of evidence that need to be considered for candidate gene prioritisation and knowledge discovery tasks.

Chapter 3 gives an overview of the Ondex data integration and network generation platform. I present datasets, methods and workflows for the construction of genome-scale knowledge networks for several crop species including wheat, barley, potato, tomato, maize, poplar and Brassica.
Chapter 4 presents the development, implementation and validation of a text-mining plugin for the Ondex platform. This text-mining plugin was developed to extend the previously constructed knowledge networks with novel gene-phenotype relations derived from the scientific literature.

Chapter 5 describes how the genome-scale knowledge networks can be mined for relevant pieces of evidence and proposes a new method for candidate gene prioritization based on biological knowledge mining. Proof-of-concept and validation of the methodology is presented using a wheat dataset of known gibberellin genes.

Chapter 6 presents a new web application, named KnetMiner, making big data available to scientists and breeders through an easy-to-use, user-targeted application. The KnetMiner platform is applicable to all species but the prototypes presented here use data from crop and animal species. I give a technical overview of the development and implementation of the KnetMiner web application and describe its configuration and deployment.

Chapter 7 demonstrates KnetMiner with two different use cases based on the analysis of QTL/GWAS data in Arabidopsis and for the analysis of differentially expressed genes in wheat. The results highlight the power of KnetMiner to support scientists and breeders with biological knowledge discovery and crop improvement.

Chapter 8 provides an overall conclusion and presents a summary of future work.

2 BACKGROUND

2.1 Connecting genotype to phenotype

In the past 50 years, science has tried to understand the relative importance and influence of genes and/or environment on shaping phenotypic traits (Polderman et al. 2015). Many biomedical and agronomic traits are complex and their expression is determined by a number of both genes and environmental factors. Complex traits have no apparent simple Mendelian basis for their variation. They may be the result of a single gene strongly influenced by environmental factors or the result of a number of genes of equal or differing effect; most likely a combination of both multiple genes and environmental factors. Discovering those genes that determine a particular biological phenotype in crops, animals or humans is referred to as the genotype to phenotype challenge.

Perfect examples of **complex traits in humans** are general intelligence (IQ) and height. Studies have shown that IQ and height are highly heritable and polygenic traits involving many genes with small effect sizes. Height is approximately 80-90% heritable and at least 40 loci have been associated with human height (Visscher 2008). Surprisingly, these loci explain only about 5% (of the expected 80%) of phenotypic variance and no gene (variant) has been discovered so far that contributes more than 0.5cm in height per gene despite studies of tens of thousands of people (Lango Allen et al. 2010). The exact heritability of IQ is more controversial but is estimated to be about 40-50% (Davies et al. 2011). The influence of the environment on the development of complex traits is more challenging to quantify. Meta-analysis studies in data collected from young children have shown that environmental factors such as iodine deficiency can result in reduction of 12.5 IQ points (Qian et al. 2005). Beside genetic and environmental factors, studies have shown large IQ differences between monozygotic twins due to epigenetic effects (i.e. DNA methylation) which resulted in differences in gene expression (Yu et al. 2012). Epigenetics is therefore seen as an important regulatory link between nature and nurture and can provide the key to transform the genetic information into phenotype (Tammen, Friso, and Choi 2013).

Furthermore, as is becoming apparent in diseases such as cancer, a complex phenotype may be the consequence of groups of seemingly independent genes interacting through a **network** of different biological relationships. A mutation in a gene may change the three dimensional structure of the protein which may affect the biological interaction network that

rewires a phenotype. From these studies, and others like them, emerges a growing belief that searching for individual or small numbers of functional genes may not be the best approach and that a network biology approach is more appropriate for bridging the genotype to phenotype gap (Benfey and Mitchell-Olds 2008; Carter, Hofree, and Ideker 2013; Y.-A. Kim, Yoo-Ah, and Przytycka 2013). In particular, Kitano has argued (Kitano 2004) that some complex diseases (e.g. cancer) are difficult to treat because there are networks of genes and products which interact to increase the robustness of the system. Intervention at any single point in the network is therefore unlikely to have a major effect.

Systematic genome-wide approaches and meta-analyses of all relevant studies are needed to determine how genetics, epigenetics, and environment interact to produce complex biomedical and agronomic traits. Identification of causal genes would facilitate the translation of research results into important clinical and commercial outcomes, including identifying new biomarkers for animal or human diseases that can lead to new diagnostics; and helping to select new varieties of crop or livestock animals with improved productivity or resistance to stresses such as disease. Searching for these causal genes in human, crop or animal genomes is, however, like searching for a needle in a haystack and gathering the evidence that supports the choice of one gene over another is even more daunting.

2.2 Genetic methods for dissecting complex traits

The genetic variation found in a population of individuals is an experimental result that can be used to inform many areas of biology (Koornneef, Alonso-Blanco, and Vreugdenhil 2004). In plant and animal breeding, genetic variation is a key concept by which natural genetic diversity is characterised and exploited for human gain. Even if the underlying biological mechanisms are not completely understood, genetic variants can be associated with phenotypic variation, and used as markers for phenotypic prediction in breeding populations. Forward (classical) genetic approaches are designed to identify regions (loci) of the genome that are linked with a particular trait. Many traits of agronomic and medical importance are not monogenic, but are determined by the action of many genes each having a small effect on the phenotype. This often results in a trait being quantitative (rather than discrete) in nature, such as yield of grain in cereal crops, or carcass weight in livestock animals. Quantitative genetics uses populations and families and applies statistical techniques to identify these regions in the genome, which are referred to as Quantitative Trait Loci (QTL) (Kearsey 1998).

Many comprehensive reviews are available describing forward genetics methods for correlating genotype and phenotype, for example see (Weigel 2012; Mauricio 2001) for reviews of methods used in plants or a comparison between two different genetic mapping strategies in soya bean (Sonah et al. 2015). For a review of molecular marker technology in plant sciences see (Henry 2012). The focus of this chapter is not *per se* the description of forward genetics approaches but rather a review of their resolution and number of identified candidate loci.

2.2.1 QTL mapping (genetic linkage)

Typically, QTL mapping is performed using segregating biparental populations. Commonly, low-density marker coverage on a few hundred members of the population (lines) is sufficient to identify many QTLs. For instance, a panel of 342 microsatellite markers were used to map QTL for carcass weight and other production traits in cattle (Zimin et al. 2009). The multigenic nature of complex traits means that many QTL may be identified in a forward genetics screen. For example, a recent study in *Brassica napus* identified 47 QTLs which were relevant for seed yield (Shi et al. 2009) and a similar number of 50-60 QTLs were reported to control seed oil and protein content in soya bean (Eskandari, Cober, and Rajcan 2013a, [b] 2013). In the bioenergy crop Poplar five QTL hotspots for biomass yield were identified (Rae et al. 2009) and various QTL studies in pig have discovered more than 400 fatness QTLs (Rothschild, Hu, and Jiang 2007).

These estimated QTL intervals can span over several cM, a genetic distance based on recombination frequencies and translates into large genomic regions with tens to hundreds of candidate genes. The recombination frequency is not distributed uniformly along the chromosomes. In humans, for instance, recombination rate varies in a range of about 0.1 to 4 cM per Mb (Kong et al. 2002). In cattle, there is an approximate correspondence of 1 cM to 10^6 base pairs and one gene every 127kb. Therefore, even the intervals between highly dense markers would contain in the region of 1.2M base pairs and with QTL intervals typically in region of at 20-40 cM so we could expect each QTL to overlay about 200-400 genes. This limited resolution is mainly the result of low recombination frequencies in biparental mapping populations, and not the effect of low marker density.

To increase the recombination frequency of biparental mapping populations, experimental populations can be created from multiple parents such as MAGIC (for multiple advanced generation intercross) and AMPRIL (for Arabidopsis multiparent recombinant inbred lines) populations (Kover et al. 2009; Xueqing Huang et al. 2011). The MAGIC population was recently used to investigate the genetic basis of variation in seed size and number (Gnan, Priest, and Kover 2014). The study identified 9 QTL for seed number and 8 for seed size. QTL mapping accuracy increases with the MAGIC population to within 300kb, or an equivalent of 60 genes.

These studies show that typical QTLs in both plants and animals generally encompass quite sizeable parts of the genome - typically several hundred genes. While QTL mapping improves the chances of finding the right gene (or genes), reducing the options down from 22,000 in cattle or 100,000 in wheat, to hundreds of genes for a particular QTL, it is still a daunting and expensive task to evaluate every potential candidate gene in the laboratory or in a field experiment.

2.2.2 GWAS (genetic association)

Genome-wide association studies (GWAS) associate phenotype with genotype at a genome-wide level using “unrelated” individuals (Hirschhorn and Daly 2005). The limitation of family-based mapping populations can be overcome by the use of unrelated genotypes that have accumulated much higher number of recombination events since their last common progenitor (Sonah et al. 2015). GWAS can have different **designs**, a simple design is to group individuals in large case-control groups. The control group may contain individuals that are healthy or show a certain phenotype, while the case groups includes individuals with a disease or a different phenotype. The study design of quantitative traits can vary and include more complex groupings. All individuals in each group are genotyped for a large number of markers to provide a high coverage of the genomes. The commonly used marker in GWAS are single nucleotide polymorphisms (SNP). For example, there is one SNP every 100 nucleotides between elite inbred lines of maize (Ching et al. 2002). In Arabidopsis, about 216,000 SNPs, or one every 0.5 kb, have been typed in over 1,000 accessions (Horton et al. 2012). Using modern SNP-arrays, a large panel of these SNPs can be used as markers. For each of these SNPs, it is then investigated if there is a statistically significant difference between the alleles in the case and control groups using for example a simple chi-squared test, or more sophisticated statistical tests for quantitative traits. Every

SNP receives a certain p-value from the statistical test. These associations then need to be evaluated to show whether they contribute to the trait of interest directly, or are linked/ in linkage disequilibrium (LD) to a QTL that contributes to the trait of interest. The negative logarithm of the p-values is often used to create so called Manhattan plots that visualise significant peak SNPs along the chromosomes (e.g. see Figure 1 in (Hui Li et al. 2012)). SNPs above a certain threshold (e.g. -log(P-value)>8) are often considered as significant.

In contrast to simple traits, GWA studies of complex traits often identify **many significant associations** along the genome. Identifying causal genes (rather than causal SNP) from GWAS requires estimations of the LD in the association population. For example, LD estimates in the global Arabidopsis population are reported to extend over not more than about 5 to 10 kb, or one to two genes, which is very convenient for GWAS (S. Kim et al. 2007). This means for every significant SNP a region +/- LD can be considered as a QTL and all genes within this region can be considered potential candidate genes. Studies in soya bean have shown that for several simple Mendelian traits the SNP physically closest to the causal gene is not always the most highly associated, or peak SNP (Sonah et al. 2015). For instance, the SNP closest to the causal gene for pubescence colour in soya bean showed the fourth greatest association. In all cases that were examined where the causal gene was known, it was found that the peak SNP was located within 100 kb of this gene and sometimes much closer, but in no case was the causal SNP captured in the gene itself. Similar findings have been reported with GWAS performed in other plant species such as Arabidopsis (Atwell et al. 2010), rice (Xuehui Huang et al. 2010) and maize (Hui Li et al. 2012).

Furthermore, GWAS are prone to a high false-positive rate of genotype-phenotype associations due to effects of the population structure and the large number of statistical tests. Epistasis and other factors can additionally lead to false-negatives where loci with known effects are not detected by the statistical tests applied in GWAS. Therefore, individual studies that report statistically significant associations between genes and phenotypes need to be approached with great caution until they have been replicated in multiple large samples (Chabris et al. 2012).

Although QTL intervals derived from GWAS encompass much smaller regions of the genome compared to QTLs from biparental mapping populations, they still produce many significant candidate SNPs. Consequently the biological interpretation of candidate SNPs to

elucidate the biological processes and pathways that they influence remains a major challenge.

2.3 Genomics and other omics technologies

Omics technologies provide the key to characterize and use genetic variation information efficiently. For example, high throughput genomic sequencing provides the means to characterize individuals and populations, to understand the genetic repertoire that they contain, to associate individuals, haplotypes and specific loci with desired characteristics and to track the transmission of parent material through successive genetic crosses. Other "omics" technologies – for example, for measuring gene expression, the presence/absence of metabolites, automatic imaging for morphological changes, etc. – can all used to quantify different aspects of response to growth and development of an organism, as well as, natural or experimental changes. They provide a toolbox to complement genetic studies by enhancing our knowledge and understanding of gene function and the translation of genotype to phenotype.

The transcriptional regulation of genes is influenced by genetic (e.g mutations, deletions, insertions, copy number variation etc.), epigenetic (e.g. methylation) and environmental factors (e.g. biotic or abiotic stresses). Changes in gene expression level consequently lead to changed concentrations of proteins in the cell that can impact biological pathways and other molecular interactions that ultimately more directly influence phenotype. High-throughput technologies such as Microarrays or RNA-sequencing make it possible to measure the abundance of the entire transcriptome (all expressed genes) of the cell. Experiments can be designed to study the effect of different treatments or environments on the same genotype or to compare gene expression in different genotypes. The aim of such studies is to identify those genes that show a statistically significant change in gene expression level between certain conditions.

The number of differentially expressed genes (DEG) in transcriptomics experiments can be very large, depending on the effect size of the treatment or environmental change. Understanding the biological mechanisms implicated by a treatment or environmental change requires functional information about the DEG. Computational approaches have therefore been developed to summarise the representation of different functional classes in the DEG. The information on gene function comes from annotations of the reference

genome and is generally captured as Gene Ontology (GO) terms (Ashburner et al. 2000). This type of analysis is known as gene set enrichment analysis whereby a gene set is analysed for overrepresented functional annotations compared to a background set (e.g. the entire genome). Enrichment analyses are popular because they are simple to run and do not require *a priori* knowledge about the experiment. They can help with a global, initial data analysis. However, the precision of gene function annotation is a problem since too many times, the detail is missing in the ontology and so the function assignment is too general to be helpful. Therefore, gene enrichment results tend to reveal very high-level biological processes that are not necessarily helpful in generating precise hypothesis.

Analysis of QTL genes differs from the analysis of DEG. The aim of QTL analysis is to identify the causal loci or alleles that control the variation in the phenotype. The majority of genes between two significant markers in a QTL analysis may be unrelated to the phenotype and only one or a few will be causal. In contrast all DEG in omics experiments are "somehow" related to the phenotype of interest, and therefore, the DEGs needs to studied as a whole. The combination of QTLs and DEG provides key inputs to generate precise hypotheses about the biological processes and networks linking genotype to phenotypes.

2.4 Bioinformatics approach to gene discovery

2.4.1 Life Science databases

Currently, over 1500 different Life Science databases are available and documented with publications in Nucleic Acid Research Databases (Galperin, Rigden, and Fernández-Suárez 2015). The majority of them are open access and contain structured and unstructured data such as sequences, gene expression, protein interaction, quantitative traits, ontologies, literature or pathways. Bioinformatics approaches that systematically integrate and mine the wealth of biological knowledge available in myriad of databases provide another route to gene discovery. The key information types and databases for *in silico* gene discovery in plants are elaborated below.

Ontologies

A major advance in data interoperability in the biosciences in recent years has been the growing use of ontologies to unambiguously identify and describe biological concepts. Ontology terms are used to annotate identified objects such as genes, experiments, and biological materials in a consistent way. An ontology is both a controlled vocabulary of terms,

often with associated synonyms, definitions, etc., and a set of semantic relationships between terms. These relationships support greater interoperability through the extension of existing ontologies, the ability to combine annotations that have been applied at different levels of specificity (based on relevance to the current question and/or availability of data), and the ability to reason over a data set and extract implicit knowledge that hides between the annotation and the semantics. Ontologies are needed both to formally define the semantics for the primary data under consideration, but also to define the metadata - the information that describes the data provenance, the measurement method and scale used - so that the data can be correctly interpreted and the definition of the gene function or trait remains consistent across interdisciplinary data resources. The use of ontologies also supports, through the use of synonyms, the mapping of annotated terms between different natural languages. One of the most comprehensive and best used ontologies in Life Sciences is the Gene Ontology (The Gene Ontology Consortium 2014) comprising over 43,000 terms and over 6.5 Million gene annotations that use these ontology terms (01/09/2016).

Genotype and genetics data

Genetic variants that are linked to phenotypes via QTL mapping, GWAS or other genetics studies provide a key data resource for gene-phenotype discovery. Access to public databases that contain such information is invaluable, however, this information is often hidden in the literature in an unstructured manner; which makes it very hard to retrieve and integrate. This has been recognised in the animal sciences and a major database AnimalQTLdb (Hu, Park, and Reecy 2016) has been established that stores results from genetics experiments. Incentives have been set that require submission of data to AnimalQTLdb as part of a journal's publication policy. AnimalQTLdb has developed to become a major genetic resource and provides a trait ontology that allows scientists to annotate QTL data with standardized ontology terms. Database curators integrate data from different genetic maps into genome based coordinates. QTL locations can therefore be downloaded in centiMorgan (cM), a genetic distance measure, and if a genome sequence is available in base pair (bp) coordinates. Such data are often available in data formats such as GFF3, SAM or BED. AnimalQTLdb contains 106,028 QTL for 1,768 traits based on 1,712 publications in 7 species (Release 30, Aug 2016). Unfortunately, an equivalent resource at similar scale does not exist for plant species although Gramene (Ni et al. 2009; Monaco et al. 2014), GnpIS (Steinbach et al. 2013) or Triticeae Toolbox (Blake et al. 2016) provide limited QTL databases for rice, barley, wheat and several other crops. QTL positions in crop

databases are often only available in cM based on genetic maps of the specific mapping population because the genome sequences are not yet available.

Genetic variants that do not have reported links to phenotypes might initially be considered less important to gene discovery. However, knowledge about published genetic variants and their effect on protein level can inform candidate gene prioritization since variants of genes with major effects can be given higher weight than genes with no reported variants or minor variant effects. The European Variation Archive (EVA) provides access to all types of genetic variants, ranging from single nucleotide polymorphisms to large structural variants from any eukaryotic organism. EVA uses the Variant Effect Predictor (Yourshaw et al. 2015) of Ensembl to annotate variant consequences. The variant consequences are described using Sequence Ontology terms.

Reverse genetics approaches are based on disrupting genes of known sequence and studying the effect of the disruption on the phenome (Gilchrist and Haughn 2010). Reverse genetics resources consist of plant material (i.e. seeds) with a certain knockout gene that can be grown and used for functional characterisation of the disrupted gene. For several plant species, e.g. Arabidopsis, rice and wheat, reverse genetics resources have been generated that allow scientists to study the function of many genes more effectively (Kleinboelting et al. 2012; Chen et al. 2012; An et al. 2005). The data from such resources is often available in custom tabular formats and could be used in gene prioritization tasks to rank genes higher for which gene knockouts with associated phenotype data exist.

Phenotype data

Genotypic data is stable for a given plant or animal. In contrast, phenotypic characterisation data is highly heterogeneous resulting from the experimental parameters applied on a given sample. The development of standards for capturing phenotypic data has been challenging since “phenotype” is a broad concept that covers all observable traits stored as descriptive data, numeric observations including time series, molecular data and image data. Phenotypic information can be obtained from dedicated phenotyping platforms, from farmers’ fields, or from ecological diagnostics in natural environments. Phenotyping platforms measure a wide range of structural and functional plant traits at the same time as collecting accurate metadata on the environment and experimental setup (Fiorani and Schurr 2013). Traits are measured at different spatial scales, from the field level (e.g. crop yield) to the cell

(e.g. cell wall polysaccharide composition) and over widely varying temporal scales, from seconds (e.g. photosynthetic response) to months (e.g whole season biomass).

Phenotype data itself (without being associated to genotype) is important in upstream processes involved in trait discovery and QTL mapping but less to gene discovery *per se*. Once phenotype data can be related to genotype, gene or mutants then it becomes a relationship of high importance. Reported gene-phenotype knowledge is one of the most valuable pieces of evidence in candidate gene prioritization. Such information is dispersed in many heterogeneous formats and locations. The public database UniProt contains a subsection 'disruption phenotype' that describes the *in vivo* effects caused by knockout or knockdown of a gene ("UniProt Website" n.d.). The Arabidopsis Information Resource (TAIR) provides phenotypic information for a range of genotypes with mutations in individual genes ("TAIR Website" n.d.). NCBI has the GeneRIF database ("Gene RIF Website" n.d.) that contains concise phrases describing a gene function that is sometimes used to add phenotypic descriptions. The majority of phenotypic information is, however, available in an unstructured form in the scientific literature and is therefore difficult to integrate with other knowledge resources such as ontologies. Text-mining techniques are required to extract and integrate such information effectively (see Chapter 4).

Due to the heterogeneous nature of phenotype data, a variety of ontologies have been developed for phenotypic data and experimental metadata, of which many are species-specific. For example, available ontologies for plants and crops include the Plant Ontology, the Crop Ontology, the Plant Trait Ontology and the Environment Ontology. The utility of such ontologies to annotate plant genomes are still limited. Even in model species such as Arabidopsis, most phenotypic descriptions are in free text which makes automated reasoning over such data very difficult. On the other hand, in other species such as Drosophila, the phenotype ontology is systematically used to annotate genes and alleles enabling more powerful search queries (Osumi-Sutherland et al. 2013).

Gene expression data

Gene expression data can be used as evidence to confirm the expression of candidate genes in tissues, organs, during developmental stages, under treatments of interest or in particular genotypes. For example a grain specific trait and QTL would require any causal gene to be expressed at some stage during grain development and potentially only expressed in certain individuals of a mapping population and not in others. Several gene

expression databases exist such as the Gene Expression Atlas (Petryszak et al. 2014) or the Gene Expression Omnibus (Edgar 2002). Reference-species resources such as TAIR have annotated Arabidopsis genes with Plant Ontology (Monaco et al. 2014) terms that describe in which tissues and during which developmental stages a gene is expressed. Other databases such as ATTED-II (Obayashi et al. 2009) analyse large amounts of expression datasets to compute clusters of coexpressed genes. Such co-expression data provides weak, speculative evidence that these genes are co-regulated and therefore could share a similar biological function or act together to control a phenotype.

Interaction data

Protein-protein interaction (PPI) data provides very useful knowledge for candidate gene discovery. In contrast to co-expression data, PPI data provides evidence about the physical interaction of proteins in the cell. A large number of methods have been developed over the years to study protein-protein interactions, e.g. affinity-tagged proteins, the two-hybrid system and some quantitative proteomic techniques (Berggård et al. 2007). Interaction most likely means that the proteins are involved in the same biological process and higher level traits although they might have different functions. Public PPI databases can be searched to identify previously reported interactions for a given bait protein. BioGRID (Chatr-aryamontri et al. 2014) and IntAct (Orchard et al. 2014) databases are populated by data either curated from the literature or from direct data depositions. Data access and download are provided for many species and in different data formats such as PSIMI-XML, PSIMI-TAB, BioPAX or RDF. Other PPI databases such as STRING (Szklarczyk et al. 2010) provide integrated and computationally inferred interaction data.

Functional annotation data

Functional annotation of genes and gene products provides a key resource for candidate gene discovery. Gene Ontology annotations capture the knowledge that we have about the molecular function of genes in a systematic and cross-species comparable manner. GO provides a controlled vocabulary to describe biological processes, molecular functions and cellular components. GO annotations require the provision of evidence codes that describe the experimental or computational methods used to establish the gene function. The Evidence and Conclusion Ontology (ECO) is used to describe the evidence in a formalised manner and help to distinguish high quality annotations (e.g. inferred through mutant phenotypes) from low quality annotations (e.g. inferred through electronic annotations). As the best studied plant species *Arabidopsis thaliana* has about 50,000 (25%) GO annotations

of experimental evidence (“GO Statistics” n.d.). The majority of annotations in non-model species are electronically inferred through sequence based comparisons with model species. The common data type for functional gene annotations is the Gene Association Format (GAF). Many functional or structural bioinformatics databases provide mappings to GO terms e.g. EC2GO, Pfam2GO and InterPro2GO. Biological pathways provide a more fine-grained knowledge about the enzymes, chemical reactions and small molecules that form the elements of biosynthetic pathways. Popular pathway databases such as KEGG (Ogata et al. 1999), Reactome (Fabregat et al. 2016) and BioCyc (Caspi et al. 2013) provide curated pathway information for model species and computationally inferred pathways for non-model species. A common file format for pathway data is the Biological Pathway Exchange (BioPAX) format.

Orthology data

The function of the vast majority of genes in non-model species remains uncharacterised. Any effort to prioritize candidate genes without any evidence about their function is difficult or even impossible. Genes that have been well characterised in other species provide a reliable source of putative evidence assuming this knowledge can be transferred from one species to another. The principal idea supporting cross-species annotation transfer is that the function of proteins is, to some extent, conserved through evolution. Thus, two orthologs in two closely related species are likely to share the same function. But the level of conservation of protein function across species largely depends on the evolution of these species, including the evolution of their proteins, of their biochemical pathways and of their higher level biological traits. Orthologous relationships can be established when comparing the genomes of two or more species. Identification of orthologous gene sets typically involves phylogenetic tree analysis, heuristic algorithms based on sequence conservation, synteny analysis, or some combination of these approaches (Trachana et al. 2014; Kristensen et al. 2011). Some of the prominent databases of orthologous genes include Ensembl (Herrero et al. 2016), OrthoDB (Kriventseva et al. 2015) OMA (Altenhoff et al. 2015) and Phytozome (Goodstein et al. 2011). The common data standard for orthology data provision is OrthoXML (Schmitt et al. 2011).

In addition to using orthology data for cross-species annotation transfer, a more direct approach exploiting sequence database search with the BLAST (Altschul et al. 1990) or Smith-Waterman (T. F. Smith and Waterman 1981) algorithms can be used to infer putative gene function. This is a common shortcut taken by many scientists and bioinformatics tools

such as Blast2GO (Gotz et al. 2008). Such data can be used for exploratory analysis but is prone to a high false positive rate. In the context of prioritizing genes it should be given a much lower weight than more accurate orthology inference methods.

2.4.2 Data integration and biological networks

The assembly of such diverse information is a technically challenging task for biologists and bioinformatician who also find it hard to evaluate the different sources of evidence and select from them the most plausible functional candidate genes. Even when this functional information gathering task is complete, 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 from QTLs or DEGs might be implicated in a complex trait is challenging. Bioinformatics approaches and public data resources can help to bridge the genotype to phenotype gap and prioritise candidate genes (Willet and Wade 2014). Using such *in silico* approaches, scientists can integrate multiple heterogeneous types of information and provide means to interrogate the information in a more systematic and informed way.

As described above, the types of biological information that are useful for gene discovery and candidate gene prioritization can include known gene-phenotype links, gene-disease associations, gene expression and co-expression, allelic information and effects of genetic variation, links to scientific literature, homology relations, protein-protein interactions, gene regulation, protein pathway memberships, gene-ontology annotations, protein-domain information and other domain specific information. Such data is typically highly connected, e.g. through common references to named biological entities, and semi-structured, e.g. because some data can be found in databases and other in free text. Furthermore, these data types are not static because new types of data are constantly emerging from advances in high-throughput experimental platforms. These characteristics of Life Science data make networks, consisting of nodes and links between them, represent a flexible data model that can capture some of the complexity and interconnectedness in the data (Huber et al. 2007). In addition, networks are often considered as the layer that connects genotype to phenotype (Carter, Hofree, and Ideker 2013).

In summary, different routes to gene discovery exist that can utilise genetics, omics and bioinformatics approaches. All these approaches can identify hundreds of potential candidate genes for specific traits. Especially in crop species, experimental validation from

lab to greenhouse to field is a slow process that can last several years. Following a wrong lead would waste significant effort, time and money. Therefore, it is important that only candidate genes with the highest level of evidence are considered for experimental validation. One of the key challenges is therefore to prioritise candidate genes and components of interaction networks that, if perturbed through potential interventions, have a positive impact on the biological outcome in the whole organism without producing negative side effects.

2.5 Related tools

Data integration is recognised as a challenge of general importance in the Life Sciences, a number of biological data warehouse solutions have been constructed to facilitate data integration and information retrieval from diverse biological data, e.g. InterMine (R. N. Smith et al. 2012), BioMart (Yates et al. 2016), LAILAPS (Esch et al. 2015) and Ondex (Köhler et al. 2006). The majority of biological data warehouse solutions use relational databases to store information and only a few systems such as Ondex use networks as their internal data structure.

Once the data have been integrated, advanced data analytics tools are needed for data mining and knowledge discovery in order to identify gene-phenotype relationships and prioritise these results. A number of web-based resources for **prioritizing** candidate genes by exploiting multiple information types have therefore been developed (Moreau and Tranchevent 2012; Bornigen et al. 2012). For example, BioGraph is based on a data warehouse approach and uses unsupervised data mining for the exploration and discovery of biomedical information (Liekens et al. 2011). In total, BioGraph contains 532,889 distinct relations among 71,042 biomedical concepts, supported by 61,570 literature references. The biological knowledge graph, which includes many indirect relationships, is used for gene prioritization and hypothesis generation. The main limitations of existing gene prioritization tools such as BioGraph is that they are restricted to the analysis of human data and that the data integration process is not easily reproducible and adaptable to other species. PosMed-Plus (Makita et al. 2009) was the first tool to prioritize candidate genes for two plant species (*Arabidopsis thaliana* and rice) using a knowledge-based approach and including literature co-occurrence and cross-species information. Similarly important to predictions is the visualisation of complex interconnected information to scientists and breeders. Appropriate data visualisation can substantially increase the yield of downstream studies.

One of the most popular tools for network visualisation in Life Sciences is Cytoscape (P. Shannon et al. 2003).

The software, called KnetMiner (Knowledge Network Miner), developed as part of this PhD thesis addresses several key shortcomings of biological knowledge warehouse and mining approaches i.e. irreproducible data acquisition and integration, infrequent database updates, lack of extension to new species and new data types, limited knowledge network exploration and visualisation capabilities. As part of this work the Ondex software was extended and the novel KnetMIner software was developed. The software was formerly known as QTLNetMiner because of it's original purpose to prioritise candidate genes within QTL regions. Once the capabilities had expanded to mine the entire genome or any gene list, we chose to rename it to KnetMiner. The silent "K" stands for Knowledge and not for Keywan as some people interestingly assume. The software and knowledge resources are free and open-source.

3 BUILDING GENOME-SCALE KNOWLEDGE NETWORKS

Life Sciences data are dispersed in various databases and heterogeneous data formats which makes a systematic interrogation of the data technically challenging. Genome-scale knowledge networks (GSKN) provide a centralised and unified representation of heterogeneous but interconnected datasets that can enable more effective knowledge mining. This chapter introduces the Ondex software and presents data sets and methods for building knowledge networks for major crops such as wheat and barley. The results section describes global characteristics of GSKNs and illustrates on one example the value of Linked Data. The principles of this work are generic and can be extended with more datasets or to other species. Some parts of this chapter have been published in (Hassani-Pak et al. 2016).

3.1 Background

The discovery of the hypotheses linking genotype to phenotype and identification of the candidate genes increasingly involves the integration of multiple heterogeneous types of information. This information is spread across many different databases (Rigden, Fernández-Suárez, and Galperin 2016) that can include known gene-phenotype or gene-disease associations, gene expression and co-expression, allelic information and effects of genetic variation, links to scientific literature, homology relations, protein-protein interactions, gene regulation, protein pathway memberships, gene-ontology annotations, protein-domain information and other domain specific information. Such data is typically highly connected, semi-structured and the data types are not static as new types of data are constantly emerging from advances in high-throughput experimental platforms.

These characteristics make networks, consisting of nodes and links between them, a natural data structure for the representation complex and interconnected biological data. Compared to relational databases, networks provide better query performance on highly connected data (many join statements are slow). In addition, networks provide more flexibility to model the data as data is not forced into a structure like a relational table, and attributes can be added and removed easily. This is especially useful for semi-structured data where a representation in relational database would result in lots of NULL column values.

In contrast to homogeneous networks, where all nodes have the same type (e.g. protein-protein interaction networks), heterogeneous information networks, referred to as knowledge networks, are networks where nodes and links can have various types (Sun and

Han 2012). Biological knowledge networks are composed of nodes which represent biological entities such as genes, transcripts, proteins and compounds, as well as, other entities such as protein domains, ontology terms, pathways, literature and phenotypes. The links in the network correspond to relations between entities and are described using terms which reflect the semantics of the biological or functional relationship such as *encodes*, *interacts*, *controls*, *expressed*, *part_of, is_a*, *published_in* etc. A knowledge network is referred to as genome-scale knowledge network (GSKN) when it contains the entire known genome (all genes) of an organism as nodes in the network. A centralised GSKN that is build from dispersed, heterogeneous data can significantly facilitate both computer-aided data mining and manual data exploration.

There are different ways of representing information in knowledge networks. Information such as gene position can be added as an attribute of the Gene node. However, when the nature of the information is more complex, it should be represented as linked data. Linked nodes are connected through relations of well defined types. These triples can then be exploited for analysis in a more systematic way. For example, SNP information could either be represented in a compact manner as a series of attributes on a Gene node or in an expanded way by using separate SNP nodes and creating links of type *has_a* to create triples. The latter approach provides more power for reasoning and allows linking specific SNPs to traits, for example, based on the results of a genome wide association study.

Ondex provides a framework for building integrated knowledge networks from heterogeneous datasets (Köhler et al. 2006). In Ondex terminology, the nodes of a network are called concepts and the links between them are called relations. For achieving a certain integration or analysis task in Ondex, public and private data sources containing the desired type of information need to be selected. The Ondex framework uses a graph-based data model and provides an API to get data into that data model. The Ondex network data structure is based on a labelled and directed multi-graph that is relatively flexible and allows information and metadata from diverse biological databases to be captured. Ondex networks can be exported in several formats such as the Ondex exchange format OXL (Taubert et al. 2007), RDF (Splendiani et al. 2012) or Cytoscape-compatible JSON. Networks can be visualised and inspected using tools like Cytoscape (P. Shannon et al. 2003) or the Ondex frontend itself (Figure 3.1). An Ondex integration workflow can be specified in an XML-defined language to achieve a reproducible integration and analysis goal. A workflow can include various operations to import data (parsers), identify equivalent nodes (mapping

methods), remove unwanted information (filters) and simplify the network structure (transformers). Workflows can be generated and executed either via a graphical user interface (Ondex Integrator) or via the command line interface (Ondex CLI). The Ondex Scripting Console provides a means to parse custom TAB data types for integration into Ondex where no dedicated Ondex parsers are yet available. The scripting syntax is based on a domain specific language developed in Lysenko 2012.

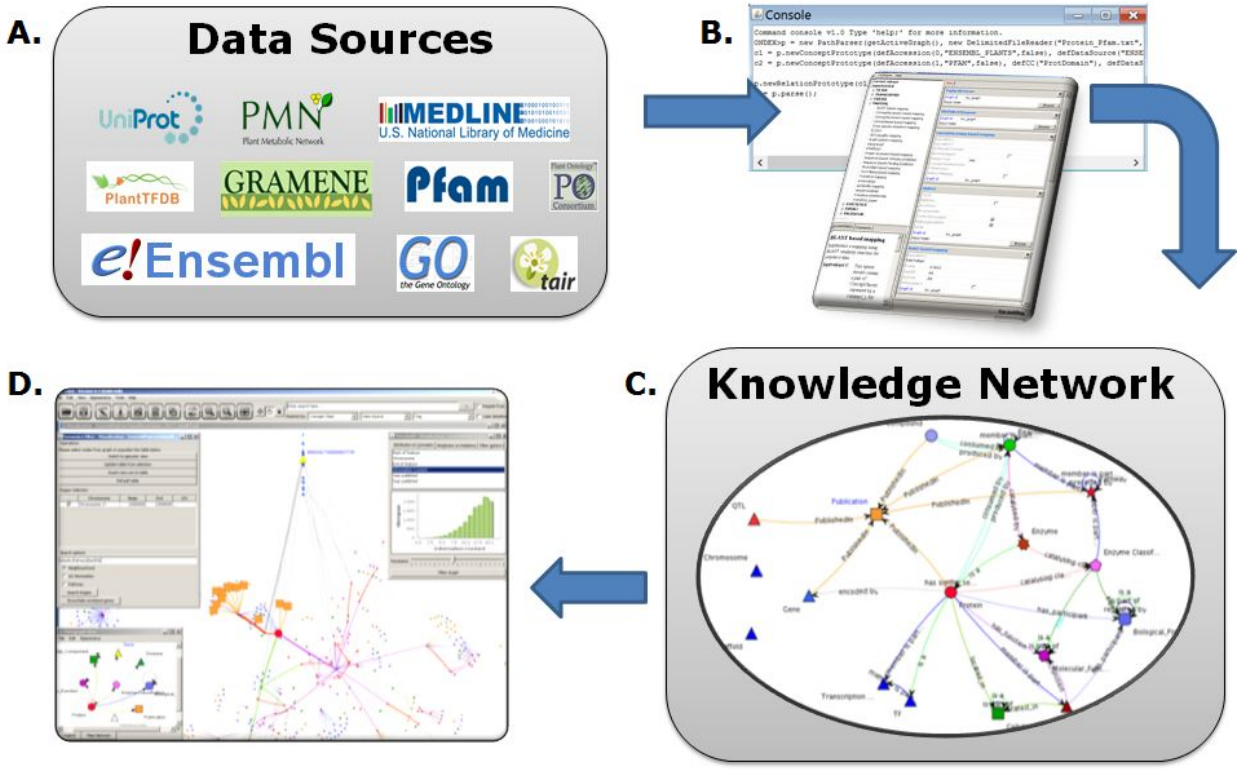

Figure 3.1. Public data sources that can be integrated into Ondex (A) using the Ondex Integrator and the Ondex Console (B). Following the data integration workflow, the integrated knowledge network (C) is loaded into the Ondex frontend for visualisation and exploration (D).

Since its release, Ondex has undergone various phases of development. Recent work extended the Ondex Visualisation Toolkit (OVTK) with an on-demand information retrieval capability using web-service based scripts that add the retrieved information to a visualised network (Horn et al. 2014). This enables an exploratory analysis to start with a small network and then gradually, on-demand, move to a larger network. The OVTK is a stand-alone, Java-based toolkit that cannot be embedded in websites. A web-enabled version of the OVTK, called Ondex Web, was developed to allow Ondex networks to be embedded in web-pages (Taubert et al. 2014). Furthermore, a Cytoscape plugin, called OndexView, was

developed that allows for concise graphical representations of integrated knowledge networks (Weile et al. 2011). Some extensive work was made to evaluate the utility of semantic web technologies (RDF, SPARQL) within the umbrella of Ondex (Splendiani et al. 2012; Canevet et al. 2010). Two studies showed the contributions of Ondex towards Bayesian data integration (Weile et al. 2012) and towards logic-based modelling (Lesk et al. 2011). Finally, Ondex was used as the main platform for biological network analysis (Lysenko et al. 2011; Defoin-Platel, Hassani-Pak, and Rawlings 2011) and in a biological study to identify candidate virulence genes in the fungus *Fusarium graminearum* (Lysenko et al. 2013).

Ondex has been under continuous technical development with major contributions by Jan Taubert (Taubert 2011), Artem Lysenko (Lysenko 2012) and Matthew Hindle (Hindle 2012). The developments have been focused on various aspects, such as improvements to application programming interfaces (APIs), webservices, workflow engine, plugins and visualisation components. I was part of the Ondex team from 2008-2011 and was responsible for the development of novel applications of Ondex to candidate gene discovery for biomass related traits in Willow (*Salix viminalis*). As part of this project, I extended Ondex with several new Java-based plugins and workflows that were essential to meet the requirements of trait-based candidate gene discovery. Some of these plugins included Ondex parsers for new data types such as Medline XML, UniProt XML, FASTA, GFF3, GAF, OrthoXML and GeneRIF. These provided the building blocks for the construction of larger data integration workflows as described in this chapter.

3.2 Methods

The methods section is composed of four parts: i) the general principles of data integration in Ondex, ii) building crop-specific knowledge networks (CropNet), iii) building reference networks of model species and integration with CropNet and finally iv) steps involved to update genome-scale knowledge networks.

3.2.1 Ondex approach to data integration

To illustrate the Ondex approach to data integration, an example is used of integrating a small knowledge network from three different data sources UniProt, Gene Ontology (GO) and PubMed. The goal is to merge these different data sources in order to gain insights from

studying the overlaps. The main steps towards achieving this goal include parsing data, mapping equivalent concepts and collapsing redundant concepts (Figure 3.2).

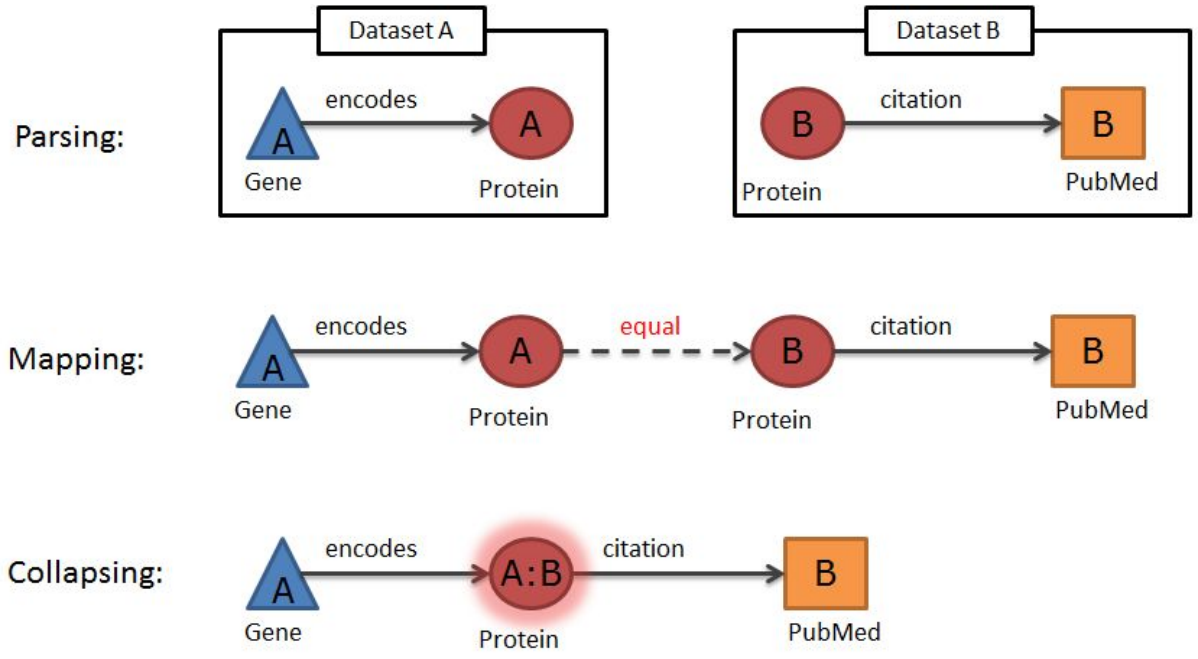

Figure 3.2: The Ondex workflow involves parsing, mapping and collapsing the data. Ondex input datasets A and B are merged via common concepts (e.g. Protein). The mapping step creates relations of type *equal* between "identical" concepts. The collapsing is a network transformation that merges identical concepts into a single concept to avoid redundancy. The merged concepts contain a summary of all the data provenances.

Parsing

Ondex has a metamodel that describes what and how data are captured in an Ondex network. The metamodel is the core semantic framework for the data model and is based on an ontology for describing Ondex *Concept Classes, Relation Types, Data Sources, Attribute Names* and *Evidence Types*. Every external data source is parsed so that it becomes a network that is an instance of the metamodel. Parsers are plugins in Ondex that read each dataset and produce independent networks within Ondex. Concepts and relations of an Ondex network will capture the parsed content within their attributes (key-value pairs). Some attributes such as *Concept Class, Relation Type, Evidence Type* and *Data Source* are requirement and others such Concept Name, Concept Accession or general attributes are optional. Attribute keys and values can be described by the metamodel, e.g *ConceptClass=Protein*, *DataSource=UniProtKB, ConceptAccession=P08684* and

ConceptName=P450. Relations can have evidence terms and numerical weights which can quantify the strength of the relationship, e.g. p-value, blast e-value or bitscore.

An alternative way of getting data into Ondex is by using the Ondex Scripting Console. The scripting code in Table A.3.1 shows how to parse a TAB file that was downloaded from Ensembl BioMart into Ondex. The file contains gene IDs in one column and gene-related SNP data in the other columns. The code creates *Gene* and *SNP* concepts for every line in the TAB file and a relation of type *has_variance* that connects the two concepts. Concepts sharing the same accessions within the parsed dataset are automatically merged into one representative concept. The scripting functionality is not “type strict” in regard to the use of the Ondex metadata. It allows new metadata to be defined that is not part of the Ondex metadata model. Validation steps need to be done manually by the user to confirm that the metadata is consistently used. These scripts can currently not be executed as part of larger integration workflows. Therefore the output network of a script needs to be manually saved in OXL format once it has been visually validated in Ondex. The OXL files can then be parsed using the OXL parser and included as part of larger data integration workflows.

As a general rule, biological entities that have gene identifiers as their primary accessions (e.g. AT1G35540) need to be represented as *Gene* concepts and entities that have protein identifiers (e.g. AT1G35540.1) need to be represented as *Protein* concepts. For example, translating a protein-protein interaction (PPI) dataset that is based on gene identifiers in the raw data into a *Protein-Protein* network would cause semantic problems in the downstream data integration process. For this reason, such a PPI dataset needs to be treated as if it were a *Gene-Gene* interaction network. Although this might not correspond to the biological truth, it is the only way of modelling the given dataset correctly, without contravening against good practices of semantic data integration.

Mapping

Importing UniProt, GO and PubMed into Ondex creates individual Ondex networks for each dataset. Each of these networks may have information overlapping with the other networks that may provide more information about particular concepts or relations. For example, UniProt has references to GO and PubMed, but lacks the hierarchy of GO and the abstracts of publications that are necessary for proper querying and analysis. This missing information is contained within the Gene Ontology OBO and PubMed XML files. Thus, the GO concepts in the UniProt network need to be mapped to the Gene Ontology network. This process of

mapping can be done in a variety of ways in Ondex . For instance, "*accession based mapping*" can be used to create a relation of type *equivalent* between two concepts when they share a common unique identifier such as GO or PubMed IDs. Two concepts can be mapped (with default accession-based mapping parameters) following best practises of data integration in Ondex when they have:

- Same Concept Class (case sensitive)
- Different Concept Data Source (case sensitive)
- Same Accession Data Source (case sensitive)
- Same Accession value (case insensitive)

Alternative mapping approaches can be used for mapping concepts with no common identifiers. These include "*name based mapping*" that maps based on shared names / synonyms or "*sequence based mapping*" that maps based on the similarity of sequence attributes.

Collapsing

With an accession based mapping approach, which requires exact matching of IDs, one can be confident in the reliability of the mapping results. Therefore, after relating concepts together based on mapping, all mapped concepts can then be collapsed to a single concept. In Ondex, this can be done with a network transformer such as the "*Relation collapser*". The provenance of the data is stored within the *Data Source* attribute of each Ondex concept. Once two or more concepts have been collapsed this attribute will be assigned a summary of all the data provenances. This is how Ondex keeps track of the origin of the data. These simple steps interconnect several data sources into one integrated knowledge network. The mapping and collapsing steps ensure that no concept occurs more than once in the network. Avoiding redundancy in the construction of the knowledge network makes successive data mining of the integrated resources significantly easier. The resulting knowledge network is content rich with light semantics and provenance on concepts.

3.2.2 Integration of crop specific data

In this section, the datasets and methods are described to build crop-specific knowledge networks (CropNet). The methods and workflows are demonstrated on data from barley and wheat but are similarly applicable to other crop species. A full overview of all data sources and Ondex parsers used for building the knowledge network are given in Table 3.1.

Genes and Proteins

The starting point of building a genome-scale knowledge network for a certain species is to collect information and relations of its genes and gene products. Information about location of genes and the transcripts they encode can be derived from GFF3 files ("The Sequence Ontology - Resources - GFF3" n.d.) and sequence information can be obtained in FASTA format. The Ondex parser "*FASTA and GFF3*" takes standard GFF3 and protein FASTA files as inputs and produces a network of *Gene* and *Protein* concepts connected via relations of type *encodes*. The parser only considers lines of the GFF3 file that are of type "gene" and ignores other features such as "mRNA", "CDS" and "exon". Information such as chromosome, start and end are added as attributes of the Gene concepts. The protein ID (excluding the "." and the integer suffix) in the FASTA files must match the gene IDs in the GFF3 file to establish the correct relation between Gene and Protein concepts. Optionally a TAB file can be provided to the parser which explicitly specifies the gene-protein mapping method. The parser requires a taxonomy identification argument (TAXID) that is added to all genes and proteins contained in the GFF3 and FASTA files. This attribute can be used to distinguish the main organism (e.g. wheat) from other species included in the knowledge network. Wheat and barley gene models and protein sequences were downloaded from Ensembl and parsed using the FASTA-GFF3 Ondex parser. This created the Gene-Protein network in which concepts are connected via relations of type *encodes* (Figure 3.3a).

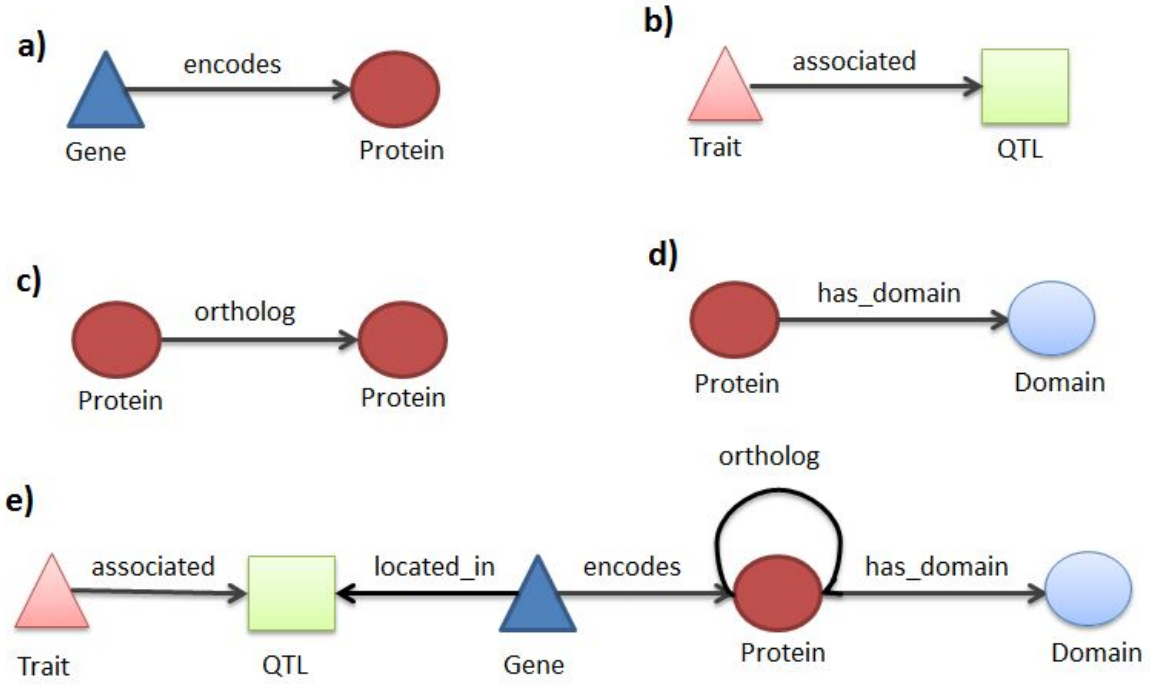

Figure 3.3: Overview of individual information types and relations (a-d) that can be integrated to build the CropNet (e).

The ‘TSV file parser’ can be used to add simple information such as gene location or synonymous names to the Gene concepts. The parser creates a concept for every line in a TAB file and adds information from the columns as attributes of the concept. This is generally a useful way for adding missing attributes to concepts in the knowledge network, however, it cannot create links (relations) to new *Concept Classes*. The wheat POPSEQ dataset provides estimated gene locations in centiMorgans (cM) based on a whole genome sequencing approach (Chapman et al. 2015). The “*TSV file parser*” was used to add the POPSEQ-based cM coordinates to the wheat *Gene* concepts.

Genetics and Genome Variation

The next integration goal is to incorporate genome variation and genetics data into the knowledge networks. In order to ensure that SNP and QTL data can be associated with genes it is important that the datasets are based on the same physical or genetic maps. In ideal case, every gene in the network will have a chromosome, start and stop position based on base pair (bp), and similarly QTL intervals would be defined using base pair. Otherwise, it is important that QTL intervals are transferred to the same genetic or physical map that specifies the gene positions before being incorporated into the knowledge network. An ideal resource is the AnimalQTLdb (Hu, Park, and Reecy 2016) that provides livestock QTL data in standardised GFF3 format. In crops, such a resource is not yet available and instead the manually curated QTL data is often provided in custom TAB format. Data about SNPs that are within or in close proximity of genes and their consequences can be downloaded from Ensembl BioMart. We use the Ondex Scripting Console to translate the raw TAB data into Gene-SNP and QTL-Trait networks (Figure 3.3b). The *Gene-SNP* network can be mapped to the Gene-Protein network based on common ENSEMBL gene ids (e.g. TRAES_2AL_65B19CC73). The *QTL-Trait* network can be incorporated in two ways: 1) it can be added to the Gene-Protein network without connecting the *QTL* concepts to it’s underlying *Gene* concepts as long as QTL and genes contain the attributes chromosome, start and stop or 2) explicit relations can be created between *Gene* and *QTL* concepts of type “*is_part*”.

Orthology and Protein Domains

The next step was to enrich and extend the wheat Gene-Protein network with new relations based on sequence analysis including orthology to other species, protein domains and sequence similarity to protein databases. Such information may sometimes be available as pre-computed datasets for download from public databases such as Ensembl, Phytozome or

OMA. It can also be locally computed using tools like OMA Standalone, InterProScan, HMMer, Smith-Waterman or Blast. We have downloaded precomputed wheat protein domain data and orthology relations to Arabidopsis and barley from Ensembl BioMart. The sequence alignments to the UniProt database were created by running a local TimeLogic® DeCypherSW™ algorithm (Active Motif Inc., Carlsbad, CA) using all wheat protein sequences as query and all reviewed UniProt plant proteins (excluding Arabidopsis) as the database while taking the top 10 hits per query sequence (E-value<0.01). The Ondex Scripting Console was used to parse these additional wheat related datasets, transform them into Ondex networks and export them in OXL format (Figure 3.3c+d). These steps create new concepts of type *Protein* and *Protein Domain* and new relations of type *ortholog, has_domain* and *has_similar_sequence* that can be interlinked with the *Protein* concepts of wheat Gene-Protein network based on ENSEMBL protein ids (e.g. TRAES_2AL_65B19CC73.1).

All these steps create a knowledge network for a crop of interest (**CropNet**) which contains genes, proteins, genetic, orthology and protein domain information (Figure 3.3e), but lacks knowledge about the biological role of these genes.

3.2.3 Integration of model species data

A similar analytical approach was implemented next to build a **reference knowledge network (RefNet)**. The first goal is to identify suitable model species, i.e. well studied species with a range of high quality annotation and interaction data. Data from Arabidopsis and other well-studied plants was taken to provide high quality functional gene information. Several curated Arabidopsis and plant datasets can be retrieved from public databases such as TAIR, Gramene and UniProt. Using similar principles as before, an Arabidopsis basic network was first developed consisting of *Gene-Protein* relations that can then be enriched and extended with functional and interaction data. These data include GO annotations, ontologies, pathways, phenotypes, protein-protein interactions (PPI) and links to relevant publications. Table 3.1 shows the *Concept Classes* and *Relation Types* that were created for each individual dataset. The **RefNet** was created by interconnecting these individual datasets based on mapping and collapsing equivalent *Gene* or *Protein* concepts. The PPI dataset provided by TAIR is based on Arabidopsis gene identifiers and not protein ids, therefore, it was translated into a *Gene-Gene* interaction network. Next, all reviewed plant proteins (excluding Arabidopsis) with their GO annotations and literature citations were

retrieved from UniProt and added to the RefNet. It is important that the *Protein* concepts in the RefNet have the same accessions as provided in the orthology and sequence similarity based datasets of **CropNet**, since these will be used to interconnect the two networks.

Gene-phenotype information is the most valuable piece of evidence in trait-based gene discovery. Phenotypic information are available in dispersed locations including UniProt, TAIR and GeneRIF databases. These were incorporated using the Ondex Scripting Console. In the knowledge network, expert-curated phenotypic information are represented as *Phenotype* concepts linked to genes or as attributes of protein concepts from UniProt. The majority of phenotypic information is available in an unstructured form in the scientific literature which makes it difficult to integrate with other concepts in the knowledge network. A description of how to extract and integrate meaningful phenotypic information from publications when building the Arabidopsis knowledge network is presented in Chapter 4.

Lastly, the **CropNet** and the **RefNet** are brought together and linked through a final workflow. The integration can be based on common concept accessions whereby duplicated concepts are mapped and collapsed. For example, CropNet contains both wheat *Protein* concepts and *Protein* concepts from the reference species. These can function as anchors for linking the two networks through accession-based mapping steps by mapping and collapsing *Protein* concepts based on shared TAIR and UniProt accessions. Additionally, *Protein Domain* concepts from the CropNet can be connected to corresponding GO terms in the RefNet. This step exploits public GO mapping files (“Index of /external2go” n.d.) as the input to the Ondex mapping plugin *External2GO*. This mapping method creates relations of type *equal* between semantically similar but non-duplicated concepts and therefore the relationships are kept and not collapsed. The result of this final integration workflow are the **genome-scale knowledge networks** for wheat or barley as presented in the results section. All workflows and datasets for building the wheat GSKN are available online (“Wheat Release Notes” n.d.).

Table 3.1. Overview of some knowledge types and Ondex parsers that are used to create the crop and reference knowledge networks. The column Metagraph shows the network semantics produced by an individual parser using following notation: Concept Class--[Relation Type]-->Concept Class.

Knowledge Type	Source file data-type	Ondex Parser Name	Metagraph
Gene-protein	gene.gff3	fastagff	Gene--[encodes]-->Protein

	protein.fa		
UniProt Protein Annotation	uniprot.xml	uniprotkb	Protein--[pub_in]-->Publication Protein--[participates_in]-->BioProc Protein--[has_function]-->MolFunc Protein--[located_in]-->CelComp Protein--[cat_c]-->EC
GO Annotations	gene_association.tair.gz GAF2Ondex.txt	gaf	Gene--[has_function]-->MolFunc Gene--[not_function]-->MolFunc Gene--[participates_in]-->BioProc Gene--[participates_not]-->BioProc Gene--[pub_in]-->Publication Gene--[located_in]-->CelComp Gene--[not_located_in]-->CelComp
Pathway	biopax.owl	biocyc	Reaction--[cat_c]-->EC Reaction--[catalyzed_by]-->Enzyme Reaction--[part_of]-->Path Path--[part_of]-->Transport Transport--[catalyzed_by]-->Enzyme Comp--[consumed_by]-->Transport Comp--[produced_by]-->Transport Comp--[consumed_by]-->Reaction Comp--[produced_by]-->Reaction Enzyme--[activated_by]-->Comp Enzyme--[inhibited_by]-->Comp Enzyme--[activated_by]-->Protein Protein--[is_a]-->Enzyme Protein--[is_part_of]-->Protcmplx Protein--[produced_by]-->Reaction Protein--[consumed_by]-->Reaction Protcmplx--[is_a]-->Enzyme Protcmplx--[consumed_by]-->Reaction Protcmplx--[produced_by]-->Reaction
Literature	medline.xml	medline	Publication
Interactions	interactions_biogrid.oxl	oxl	Gene--[pub_in]-->Publication Gene--[genetic]-->Gene Gene--[physical]-->Gene
GWAS	Gene-SNP-Phenotype.oxl	oxl	Gene--[has_variation]-->SNP SNP--[associated_with]-->Trait
Phenotype	Gene-Phenotype.oxl	oxl	Gene--[pub_in]-->Publication Gene--[has_observ_pheno]-->Phenotype
Gene Ontology	go-basic.obo	genericobo	MolFunc---[is_a]--MolFunc MolFunc--[part_of]-->MolFunc CelComp--[is_a]-->CelComp CelComp--[part_of]-->CelComp BioProc--[neg_reg]-->BioProc BioProc--[pos_reg]-->BioProc BioProc--[is_a]-->BioProc BioProc--[part_of]-->BioProc BioProc--[regulates]-->BioProc
Trait Ontology	to-basic.obo	genericobo	TO--[is_a]-->TO TO--[part_of]-->TO
Homology	Arabidopsis_Plants.oxl	oxl	Protein--[h_s_s]-->Protein

Homology	Inparanoid_Arabidopsis_ Yeast.oxl	oxl	Protein--[ortholog]-->Protein Protein--[paralog]-->Protein

3.2.4 Updating knowledge networks

Three Ondex workflows were designed to build a genome-scale knowledge network for plant species. The first workflow integrates crop-specific (e.g. wheat) information. The second workflow integrates publicly available data from model species such as Arabidopsis into a reference network. The final workflow links the crop-specific and the reference networks.

Public databases such as protein entries in UniProt, publications in PubMed and GO annotations are not static and are updated on a daily or monthly basis. For example, the number of publications in PubMed that contain the word Arabidopsis has risen by nearly 20,000 new articles in the last 5 years (see Figure 3.4). In recent times, GO annotations, nucleotide and protein sequence repositories have had a similar sharp rise in their number. Therefore, it is increasingly important to keep such fast growing information up-to-date on a frequent and regular basis. On the other hand, crop-specific datasets such as genome assemblies, gene models and QTL data change less frequently and can therefore be updated on demand.

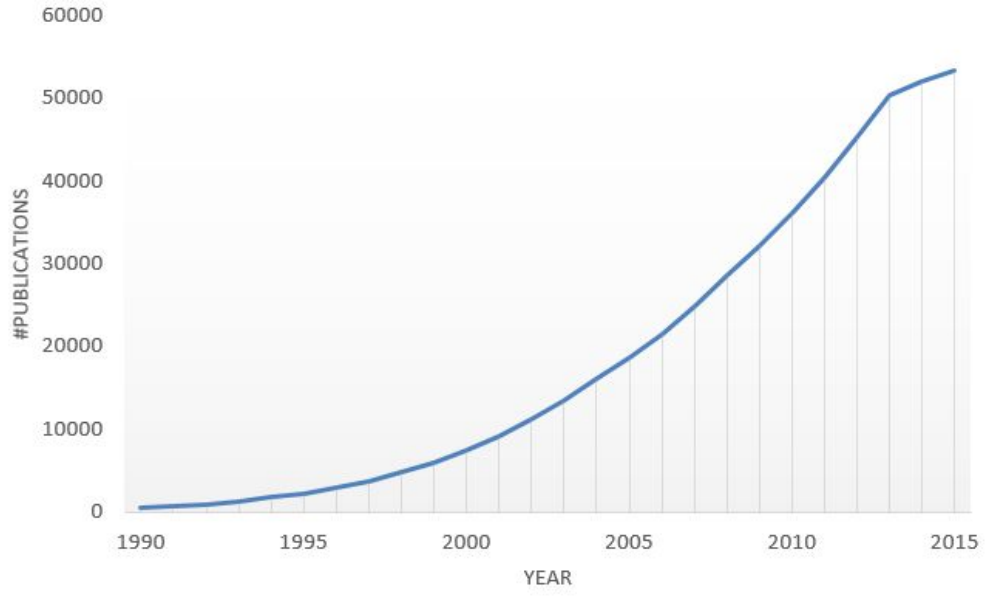

Figure 3.4: Number of Publications in PubMed that contain the word Arabidopsis. Data from PubMed.

Our approach to data integration is based on building a centralized knowledge store using data snapshots available at the time of integration. We have developed prototype scripts for

rebuilding the knowledge store in a semi-automated manner. The focus has been on automating the update of datasets that frequently change and are publicly accessible in standardised formats for which Ondex Parsers are available, i.e. ontologies, publications and GO annotations. We have therefore automated most of the RefNet data download and integration steps that include:

- Backup of old RefNet datasets
- Download and integrate new RefNet datasets including
    - Genes and Proteins (FASTA, GFF3 format)
    - UniProt Plants (XML format)
    - Gene Ontology (OBO format)
    - Trait Ontology (OBO format)
    - Arabidopsis Gene Annotations (GAF)
    - PubMed abstracts (XML format)
    - BioGRID interactions (TAB format)
    - Pathway information (OWL format)
- Re-run workflow that integrate new RefNet with existing CropNet
- Export new GSKN in OXL or other formats

The update and integration scripts make use of Ondex-CLI which is a lightweight version of Ondex that runs on the command line and not via a graphical user interface. About 20Gb of random-access memory (RAM) is required when running Ondex-CLI with workflows and datasets that create the final GSKN. This process currently takes about ~5 hours to update datasets and re-build knowledge networks for species like wheat and barley. The results are manually inspected by studying the metagraph and integration logs.

3.3 Results

3.3.1 Comparison of GSKNs

Genome-scale knowledge networks (GSKN) were developed for major plant and crop species including Arabidopsis, wheat, barley, maize, Brassica, potato, Solanaceae and poplar. The networks differ in the crop-specific information included (genes, SNP, QTL, traits, publications). Nevertheless, they all contain the identical reference networks consisting of Arabidopsis and other plant species from UniProt-SwissProt Plants. The size of the GSKN can vary depending on the genome size and data integrated for that particular organism (Table 3.2). The wheat and barley genome releases that were integrated contain

99,386 and 79,379 genes respectively. The current version of the wheat GSKN contains about 450k concepts and 1.7 million relations and the barley GSKN is slightly smaller with 420k concepts and 1.3 million relations (26/05/2016).

Table 3.2: The size of the individual knowledge networks in terms of total number of crop or plant specific genes, concepts and relations (as retrieved on 26/05/2015).

Knowledge Network	Number of Genes	Number of Concepts	Number of Relations
Arabidopsis	29,507	221,334	823,287
Wheat	99,386	448,653	1,669,464
Barley	79,379	419,896	1,257,049
Maize	39,469	335,882	1,367,956
Brassica	59,225	428,803	1,361,537
Potato	35,119	315,455	1,061,972
Solanaceae	69,846	386,092	1,296,974
Poplar	41,335	381,307	1,536,244

The type and amount of information held in a knowledge network varies from species to species. Tables 2.3 and 2.4 provide an overview of the number and type of *Concept Classes* and *Relation Types* in the Arabidopsis, barley and wheat knowledge networks. The Arabidopsis reference network is included in both the barley and wheat networks. A total of 130,815 genes are included in the wheat GSKN. Of this total, 31,429 and 99,386 genes are Arabidopsis and wheat specific, respectively. Similarly in barley, 80,662 genes are from barley and the remaining genes are from Arabidopsis. The 5 *Chromosome* concepts across all networks only represent Arabidopsis since wheat and barley networks have stored the chromosome information within the *Gene* attributes and not as explicit concepts. The Arabidopsis GSKN does not contain protein domains, orthology relations or sequence similarity to protein databases which can explain the smaller numbers seen in some *Concept Classes*. The wheat network contains nil wheat specific QTL, SNP and trait information, whereas in barley we integrated QTL and SNP datasets from Gramene and Ensembl.

Table 3.3: Total number of concepts and type per Concept Class included in the Arabidopsis, barley and wheat knowledge networks.

Concept Class	Arabidopsis	Barley	Wheat
Biological Process	27,525	27,486	27,525
Cellular Component	3,787	3,787	3,787
Compound	2,980	5,457	2,980
EC	2,391	1,789	1,754
Enzyme	15,150	26,698	15,150
Gene	31,429	112,091	130,815
Molecular Function	9,919	9,866	9,919
Pathway	587	676	587
Phenotype	6,489	6,489	6,489
Protein Complex	187	192	187
Protein Domain	0	7,032	9,417
Protein	57,301	136,735	177,378
Publication	62,270	61,329	61,305
QTL	0	285	0
Reaction	3,097	5,612	3,097
RNA	1,296	1,296	1,296
SNP	0	16,030	0
Thing	187	192	187
TO	1,314	1,314	1,314
Trait	0	30	0
Transport	54	96	54
Total	**225,963**	**424,482**	**453,241**

Table 3.4: Total number of relations and type per Relation Type included in the Arabidopsis, barley and wheat knowledge networks.

Relation Type	Name	Arabidopsis	Barley	Wheat
ac_by	activated by	135	138	135
ca_by	catalysed by	15,150	26,698	15,150
cat_c	catalysing class	11,561	13,718	11,561
control	control	0	285	0
cooc_wi	co-occurs with	97,076	97,076	97,076
cs_by	consumed by	6,799	12,412	6,799
enc	encodes	32,498	111,877	131,277
equ	equal	3,398	2,234	2,212
h_s_s	has similar sequence	37,219	201,763	153,655
has_domain	has domain	0	60,465	253,873
has_function	has function	81,964	107,348	182,568
has_observ_pheno	has observed phenotype	6,489	6,489	6,489

has_part	has participant	7	7	7
has_variation	has variation	0	23,898	0
in_by	inhibited by	250	266	250
is_a	is a	89,227	100,687	89,227
is_part_of	is part of	299	302	299
interacts_with	interacts with	6,481	6,481	6,481
located_in	located in	86,859	92,968	114,087
neg_reg	negative regulation	2,525	2,525	2,525
not_function	not function	172	172	172
not_located_in	not located in	713	713	713
occ_in	occurs in	96,435	96,435	96,435
ortho	ortholog	0	20,001	198,721
part_of	part of	10,619	12,904	10,619
participates_in	participates in	118,144	133,260	169,642
participates_not	participates not	199	199	199
pd_by	produced by	7,761	14,167	7,761
pos_reg	positive regulation	2,512	2,512	2,512
pub_in	published in	160,291	160,327	160,291
regulates	regulates	2,942	2,942	2,942
Total		**877,725**	**1,311,269**	**1,723,678**

3.3.2 Search and visualisation of GSKN in the Ondex frontend

Due to the large size of GSKNs visualisation and interaction with such size networks is not simple. To open a GSKN in the Ondex frontend takes about 5 minutes and requires at least 6Gb of RAM. Again due to the large size, the Ondex network is initially hidden in the main visualisation window.

Although the main network is too large to be displayed, the Metagraph window summarizes the different *Concept Classes* and *Relation Types* present in the knowledge network and their relationships. Figure 3.5 shows the metagraph of the barley GSKN. The knowledge network consists of 22 different *Concept Classes* representative of both biological entities (Gene, Protein, Protein Complex, Compound, SNP) and general entities (Biological Process, Pathway, Phenotype, Publication). The metagraph nicely visualises relationships between *Concept Classes*, for example, 'Biological Process (GO)' has incoming relations from multiple *Concept Classes*, such as Gene, Protein, RNA, Protein Domain and Enzyme

Classification. The Metagraph window allows users to make subsets of the main network visible/invisible.

In addition, Ondex can be used to search, filter and annotate networks. For example, one can search for genes or phenotypes of interest and apply a neighbourhood search or a shortest path search to identify a potential link between selected concepts. Such smaller subnetworks can be gradually extended using the context-sensitive right-click menus in Ondex that allow additional links of certain types to be added to the network.

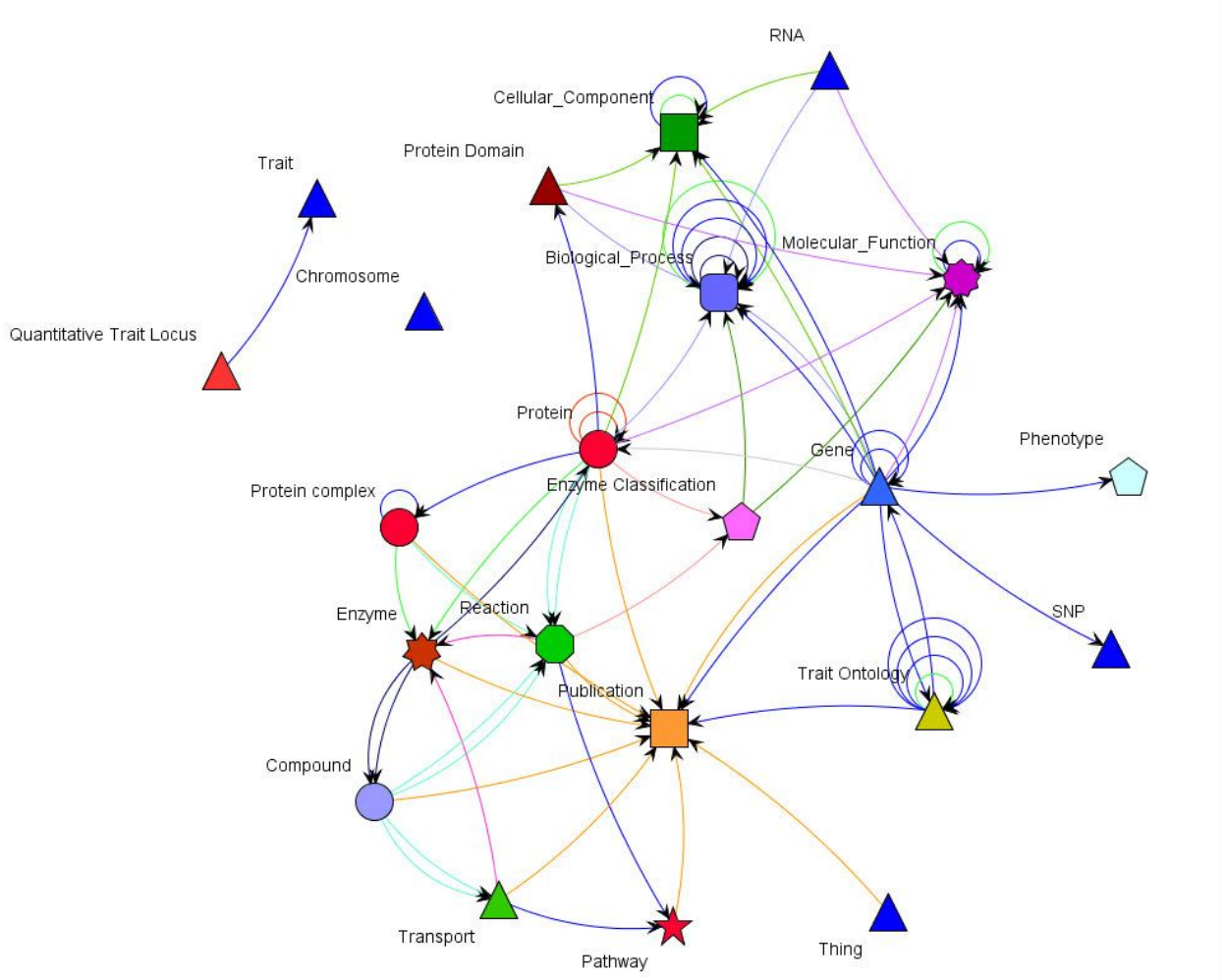

Figure 3.5: The metagraph of the barley genome-scale knowledge network. Different node shapes and colors represent different Concept Classes. Relation Types are omitted here for clarity reasons.

3.3.3 Application of GSKN to gene discovery and crop improvement

The main driver for building GSKN was that selected data sources contain fundamental relationships that upon integration and consecutive analysis, can yield chains of functional associations among more distant concepts. In this case, the application was to identify chains of functional associations between traits (phenotypes) and causal genes. Here, we present an example network, extracted from the barley GSKN, to provide a proof-of-concept and demonstrate the potential application of Linked Data to gene and knowledge discovery. The barley GSKN was searched and filtered to identify a potential relationship between barley gene MLOC_10687.2 and increased/decreased seed size (Figure 3.6). The results show that this gene is located within QTLs for seed width (AQDE021) and leaf water potential (AQGZ019). It encodes a protein that has a DNA-binding WRKY domain and is orthologous to *TTG2* in Arabidopsis. Evidence in Arabidopsis indicates that *TTG2* mutants have smaller seeds and that *TTG2* is involved in seed coat development and epidermal cell fate specification. PubMed references are provided within the knowledge network to strengthen the association (PMID:22251317 and PMID:15598800).

This example highlights the potential benefits of data integration and linked data to establish associations between distant concepts such as traits/QTL on one side and genes/biological processes on the other side. The original information was dispersed across several heterogeneous databases (Gramene, Ensembl, TAIR, GO and PubMed) and only by interconnecting them in a semantically consistent manner it is possible to search the information effectively. This and other similar examples provide a proof-of-concept for data integration needs in life sciences. Tools for knowledge mining and discovery still need to be developed that can exploit integrated knowledge networks more effectively in order to predict candidate genes for key agronomic traits in a systematic manner.

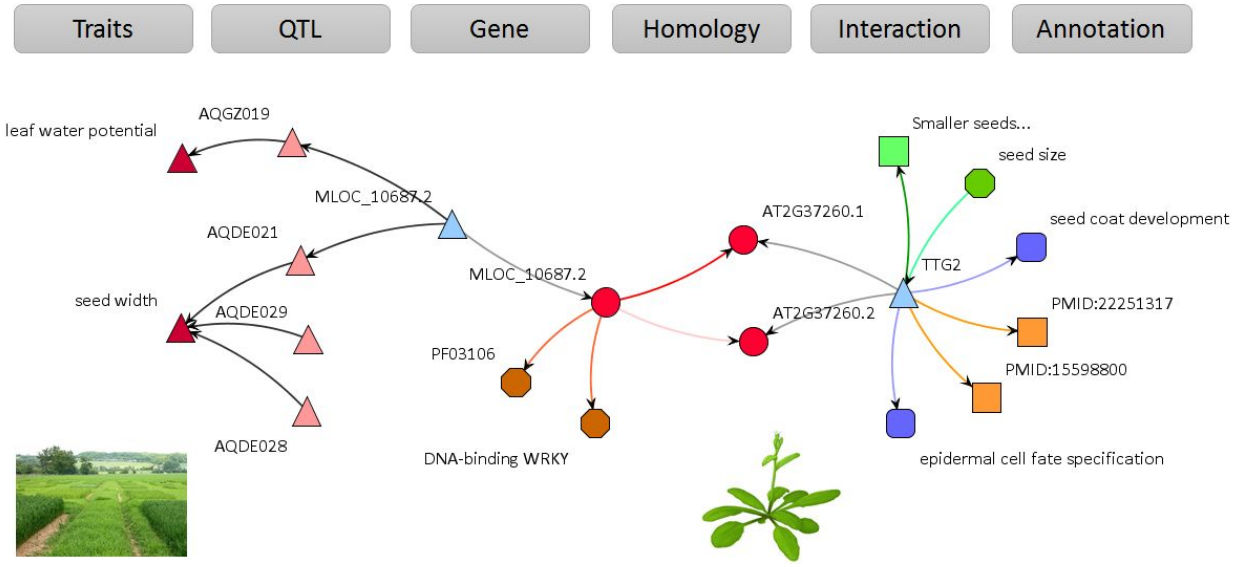

Figure 3.6: A heterogeneous network that links CropNet information on the left (Traits, QTL and Gene) to RefNet information on the right (Homology, Interaction and Annotation).

3.4 Discussion

Navigating the heterogeneous data landscape is a technically challenging task for many biologists and bioinformaticians who have to spend excessive time pre-processing data for integrative analysis. Therefore, knowledge discovery is often hampered by the challenges of data integration and new approaches are needed to improve the efficiency, reproducibility and objectivity of these processes. Knowledge networks provide a solution for the representation of heterogeneous but interconnected information. Eleven knowledge networks have been created that are used for the identification of candidate genes and the generation of research hypotheses in the species Arabidopsis, poplar, wheat, barley, potato, tomato, Brassica, maize, pig, cattle and chicken. In addition, several new knowledge networks are near completion for model and commercial insect species such as Drosophila and bumblebee. The flexibility of the network construction approach has meant that it was possible to extend its application to other plant and animal species, often as part of national or international collaborations. Each time a new knowledge network was built, the Ondex plugins were refined and the workflows improved.

The data integration approach is selective and does not attempt to integrate every dataset that is in the public domain. Instead, the method focuses on integrating datasets that add value to a particular application case (e.g. candidate gene discovery). Selected data sources usually contain fundamental relationships that upon integration and consecutive analysis,

can yield chains of functional associations among more distant concepts. The data integration method also does not attempt to utilise raw data, but instead, it integrates processed data. For example, NGS data is integrated as processed data in the form of SNP or gene expression information. Resulting networks are provided in OXL format that can be manually explored by humans or mined automatically by computers.

Several shortcomings in Ondex can make it challenging for new users to fully exploit the power of Ondex data integration. One limitation of the data structure is that multiple relations of the same type between two concepts are not permitted in Ondex. For example, two proteins A and B cannot have multiple relations of type *interacts* between them based on different evidence. This situation can be modelled in Ondex by using a single *interacts* relation that contains multiple evidence types such as yeast-two-hybrid or affinity purification. In Ondex, there is a lack of user-friendly monitoring and reporting tools to provide useful information during the parsing, mapping, filtering and network transformation stages. There is also a lack of documentation, which can make it difficult to understand the semantics and metagraphs that the Ondex parsers create. Without this understanding, larger workflows and mapping parameter definitions can become cumbersome to design and test. The Ondex Integrator contains a plugin for exporting the Ondex network as an XML based report. This plugin can provide a quick way of retrieving the network metadata information. Nevertheless, effective integration often requires a step-by-step approach, whereby each step is evaluated by opening and investigating the Ondex network in the Ondex Visualisation Toolkit. A step-by-step integration approach, allows the user to systematically build a knowledge network that meets their particular investigation requirements. Due to the large size of GSKN, the exploration of such networks in Ondex is tedious, slow and requires powerful computers with sufficient RAM. Finally, the Ondex Integrator is missing a highly-configurable Ondex parser for custom tabular data types. Before TAB files can be included in an automated Ondex integration workflow, they first must be manually parsed into Ondex using the Ondex Console.

The examples used here showed how to build knowledge networks for species with a sequenced genome. In many cases, especially for non-model organisms, whole genome sequences or gene models are not available. For these species, a transcriptome assembly (for example from RNA-seq data) can alternatively be used to build the knowledge network. A knowledge network that is based on a transcriptome assembly differs in that it won't have gene locations with genome-wide coordinates, nor will it have standardised entity identifiers

across public databases. Data integration workflows and methods are therefore far more challenging for transcriptomes and the approach needs to be customised far more.

The process of building knowledge networks has been partially automated to include data download and data integration of RefNet using the Ondex-CLI. Using these automated steps, new versions of the knowledge networks are created on a regular (monthly) basis. This approach rebuilds knowledge networks from scratch instead of updating the parts that have changed. Mechanisms are not currently available to determine which parts of the knowledge network have changed after an update. Future work could develop an analytical approach for identifying genes that have new links or updated annotations of interest. This would allow for the development of automated services to inform users when new information about their genes of interest becomes available.

In conclusion, heterogenous data sources can be integrated to form a knowledge network. Heterogenous data sources can contain explicit references to each other, thereby making it relatively straight-forward to interconnect the data. In situations, where direct references don't exist between the data sources, other approaches need to be exploited for interconnecting the different datasets. One such approach can be the analysis of the scientific literature to establish connections between different biological entities. *Publication* concepts provide an unstructured source of evidence that can be exploited to interconnect information in an Ondex network. This information can even create novel relationships between concepts that are not yet present in structured databases. Chapter 4 presents how the Ondex data integration framework was further enhanced with novel text-mining capabilities to facilitate the interlinking of datasets lacking direct references.

Appendix

Table A3.1: An example script to parse a TAB file into Ondex and create gene-SNP relations using the Ondex Scripting Console :

Allele	Variation ID	Distance to transcript	Gene stable ID	Strain name	Consequence to transcript
C/A	tmp_morex_contig_1558931_656	473	MLOC_10372	barke WGS	downstream_gene_variant
G/A	tmp_morex_contig_1558931_667	462	MLOC_10372	barke WGS	downstream_gene_variant

```
p = new PathParser(getActiveGraph(), new DelimitedFileReader("biomart_export.txt", "\\t+",1));

c1 = p.newConceptPrototype(defAccession(3,"IBSC",false), defDataSource("ENSEMBL"), defCC("Gene"));
c2 = p.newConceptPrototype(defAccession(1,"IBSC",false), defDataSource("ENSEMBL"), defCC("SNP"), defName(9),
defAttribute(0,"Allele","TEXT", false),
defAttribute(2,"Distance","INTEGER", false),
defAttribute(4,"Strain_Name","TEXT",false),
defAttribute(5,"Transcript_Consequence","TEXT",false);
p.newRelationPrototype(c1, c2, defRT("has_variation"), defEvidence("ENSEMBL"));
s = p.parse();
```

4 EXTENDING ONDEX WITH TEXT MINING CAPABILITIES

One of the strongest evidences for candidate gene discovery is a known link between gene and phenotype. The majority of such gene-phenotype information is available in an unstructured form in the scientific literature. Automated approaches are needed to extract and integrate phenotypic information from publications and link these to the corresponding genes. Such approaches will create novel, structured relationships between concepts and therefore improve the ability to reason over the data. As part of this thesis, the Ondex data integration software was extended with additional text mining capabilities to improve the richness of knowledge networks and enhance gene discovery. The initial research on this topic was published in (Hassani-Pak et al. 2010).

This chapter describes the motivation and the requirements for use of text mining methods as an important contribution to data integration in the biological sciences. First, the design and implementation of the text mining approach taken is described and then its application on two biological use cases is outlined. The first use case demonstrates how the developed methodology can be utilised to build a weighted association network which is benchmarked against a gold standard dataset. The second use case presents how knowledge networks can be extended with novel gene-trait relationships using the trait ontology as the input for the text mining.

4.1 Background

A tremendous wealth of knowledge is contained within the scientific literature in the form of unstructured free text. PubMed comprises over 25 million publication citations. On average, Pubmed grows at a rate of 500,000 publications per year. It is the source of the most up-to-date research results in the biomedical and life sciences. Searching PubMed for descriptors relevant to a scientist's research discipline such as disease, species or phenotype can retrieve thousands of publications. Reading publications, extracting facts and using the facts to create knowledge is a time-consuming task. Automated solutions that can extract relevant facts from text, integrate these facts seamlessly into the data and visualise the data intelligently are therefore in high demand. Such solutions can help humans or computers to make novel connections between previously unrelated biological concepts (Rebholz-Schuhmann, Oellrich, and Hoehndorf 2012).

Text mining is the discipline of analysing unstructured free text in order to extract structured facts and knowledge from it. Many excellent reviews of, and introductions to text mining approaches have been published and these reviews categorize text mining approaches into three main types: co-occurrence-based, rule-based and machine-learning-based approaches (Cohen and Hunter 2008; Krallinger, Valencia, and Hirschman 2008; Ananiadou et al. 2006). **Co-occurrence-based** methods search for concepts that occur in the same unit of text (typically sentence or abstract) and create relationships between them. **Rule-based** systems are more sophisticated as they apply linguistic and semantic analyses to find explicit statements that explain the relationship between concept classes (e.g. <gene> *controls* <phenotype> or <protein> *positively regulates* <pathway>). **Machine-learning-based** methods differ in that they require a training information set consisting of labelled sentences. A trained classifier is then used to identify similar associations in a larger body of text or text corpus. The two main challenges that any type of text mining method must deal with are the issues of ambiguity and variability of language. **Variability** means that there are different ways of expressing the same concept in written text. For example, the trait *grain colour* can also be expressed as *bran colour* or *pericarp colour* (synonyms). Additional synonyms include the variability associated with regional spelling differences of words. For example, British spelling of *colour* and the American spelling of *color* are synonyms. **Ambiguity** means that certain words or phrases can have different meanings. For example, the word *ear* can refer to the sense organ that detects sound but for plants an ear is the top part of a grain plant such as wheat. Additionally, EAR is a three letter acronym of the gene name, Ethylene-responsive element binding factor-associated Amphiphilic Repression.

In recent years, a plethora of stand-alone text mining systems have been developed (Leitner et al. 2013), mostly to support database curators in finding evidence text for particular information of interest, such as protein-protein interactions or functional gene annotations (Lu and Hirschman 2012; Mao et al. 2014). The input to text mining systems is generally a text corpus. This can be fields from a database (e.g. comment fields from UniProt or GeneRIFs from NCBI), abstracts from PubMed or full-text journal articles, for example. The user then selects concept types such as gene, drug, disease, treatment etc. and the text mining method annotates the text corpus with entities that correspond to these classes. The output format is mostly based on tables that contain the associated facts (tuples or triples) in different columns, with links to database identifiers and extracted evidence text. In addition to user interface based systems, Java based libraries and frameworks have recently

emerged providing APIs that enable language processing functionality to be embedded in diverse applications (Cunningham et al. 2013; “Apache UIMA” n.d.). Such frameworks allow text mining workflows to be created that consist of elementary components, for example text segmentation, sentence boundary detection, entity detection and relation extraction .

The aim of this work was to extend Ondex with text processing capabilities in order to augment the knowledge network with additional facts (knowledge) extracted from PubMed abstracts. The following requirements were considered as important:

- Detection of entity names in publications. Entities and publications are both concepts of the knowledge network. Some concepts may contain several synonyms and others very few. It is important that any type of entity (*Concept Class*) can be recognised.
- Building weighted association networks from the data produced in the first requirement and the data already available in the network (*published in* relations). Edges between concepts are required to contain statistical confidence scores and evidence text.
- Text mining features need to be implemented as Ondex plugins in order to be invoked as part of fully automated Ondex workflows.

Several possibilities for the design of such a system were considered.

- Export data from Ondex into a stand-alone text mining system and import the results back again. The advantage of this approach is that existing text mining tools could be used for the text analysis. However, the downsides are the limited flexibility, incompatibility of data formats and the manual intervention that would be necessary for moving the data back and forwards.
- Use text mining web-services. The advantage would be that no manual intervention would be necessary. However, the disadvantages are again the limited flexibility and data incompatibility as well as the potential risks of slow speed of a web-service.
- Integration of text mining features into Ondex. This would give text mining methods direct access to the Ondex graph API. The methods could be implemented as flexible Ondex plugins that can be invoked as part of automated Ondex workflows.

Tight integration of text mining features into the Ondex data integration framework outweighed the benefits of the other approaches and therefore a set of basic text processing features were implemented that matched the requirements specified above. The implementation in Ondex was developed on top of Apache Lucene (“Apache Lucene - Welcome to Apache Lucene” n.d.). Lucene provides a library for information retrieval and

document ranking. This chapter describes the design and implementation of a co-occurrence-based text mining plugin for Ondex that can be used as a simple baseline for the future development of more sophisticated rule-based or machine-learning-based systems that exploit frameworks such as GATE (Cunningham et al. 2013) and Apache UIMA ("Apache UIMA" n.d.) for advanced text analysis.

4.2 Methods

A genome-scale knowledge network as developed in the previous chapter usually contains a large number of *Publication* concepts linked to other concepts through high quality, manually curated references provided in public databases. The section will first describe how an additional corpus of text can be added to a GSKN and then explain how new links can be established between *Publication* concepts and other concepts (e.g. ontology terms) in a knowledge network. When consecutive co-occurrence associations are made as described in the final methods part, it is important to discern the source of relations (text-mining derived relations or from human-based curation), eventually providing more weight to manually curated data.

4.2.1 Document retrieval and indexing in Ondex

Three alternative strategies are available for adding a collection of publications, referred to as a corpus, to an Ondex knowledge network. The first is to provide a PubMed XML file ("MEDLINE®PubMed® XML Element Descriptions and Their Attributes" 2005) to the Ondex Medline parser, which can come from a PubMed search using a keyword (e.g. "Arabidopsis" or "disease resistance"). Secondly, if the Ondex graph already contains concepts with PubMed IDs, the Medline parser's eFetch parameter can be used to retrieve the corresponding XML entries using NCBI web services (which download approximately 1,000 publications/minute). Thirdly, a list of PubMed IDs can be provided to the Medline parser, which will also retrieve the corresponding XML entries using web services. It is possible to easily combine any of these strategies together as a single parsing tool. The Medline parser reads the XML files and creates unconnected Publication concepts with attributes such as PubMed ID (PMID), Digital Object Identifier (DOI), title, abstract, authors, journal, year, Medical Subject Headings (MeSH) and Chemical terms. Subsequently, the 'Ondex accession based mapping' is used to map publications previously available in the graph to

the new Publication concepts that contain titles and abstracts. To avoid redundancy, the 'Ondex relation collapser' plugin is used to merge the mapped concepts into a single concept. These steps create the corpus which consists of connected *Publication* concepts that are linked through *published_in* relations with other concepts in the network and a set of unconnected *Publication* concepts that were newly added to the knowledge network.

The Lucene search engine library is part of the Ondex API and allows knowledge networks to be indexed and searched efficiently. The graph indexing methods translate all concept types including *Publication* concepts of the Ondex network into Lucene documents, where different Lucene fields represent different concept attributes. All text is first converted to lowercase and non-alphanumeric characters and stop words (e.g. "the", "of", "a") are removed, before the text gets tokenized (broken up) into words and added to the index. Typographical variants are thereby excluded, so that words like *Kcnip3* and *KCNIP3* are stored as *kcnip3* and *KCNIP-3* or *kcnip_3* are stored as *kcnip 3* in the index.

4.2.2 Mapping publications to concepts in the knowledge network

With the ontologies, databases and publications integrated with Ondex, the next task is to detect a biological concept (name or synonym) in the title or abstract of a publication. This is known as Named Entity Recognition (NER), which we have implemented using different Lucene search methods. The standard search method considers *exact* occurrence of the concept names in the abstract or title of the publication. Two other search methods have also been implemented: *fuzzy search* matches documents that contain terms similar to the specified query term based on the Levenshtein algorithm and *proximity search* supports finding patterns, i.e., ordered words appearing within a specific distance of one another. Concept names and synonyms are used as Lucene query terms, normalised the same way as the index, to search the title and abstract fields of the publication documents that are available in the index. Matching documents are ranked and scored using a modified version of the Lucene TF*IDF (Robertson 2004) based scoring function. The Lucene scoring method was modified to give a higher weight to publications containing the query in their title rather than in their abstract. This recognises the assumption that reference to a concept in a title will be more informative than in an abstract. If several synonyms of the same concept are found in the publication, the highest Lucene score is recorded.

The entity recognition step has been implemented as an Ondex mapping method. Having recognised a concept in the publication, a relation of type *occurs_in* is created, indicating that the given publication is related to the identified concept (Figure 4.1.a). In order to provide context or evidence for the relation, each abstract is split into sentences and each sentence containing the matching query is stored as evidence text. Furthermore, the Lucene score is added as a weight to the publication-concept mapping.

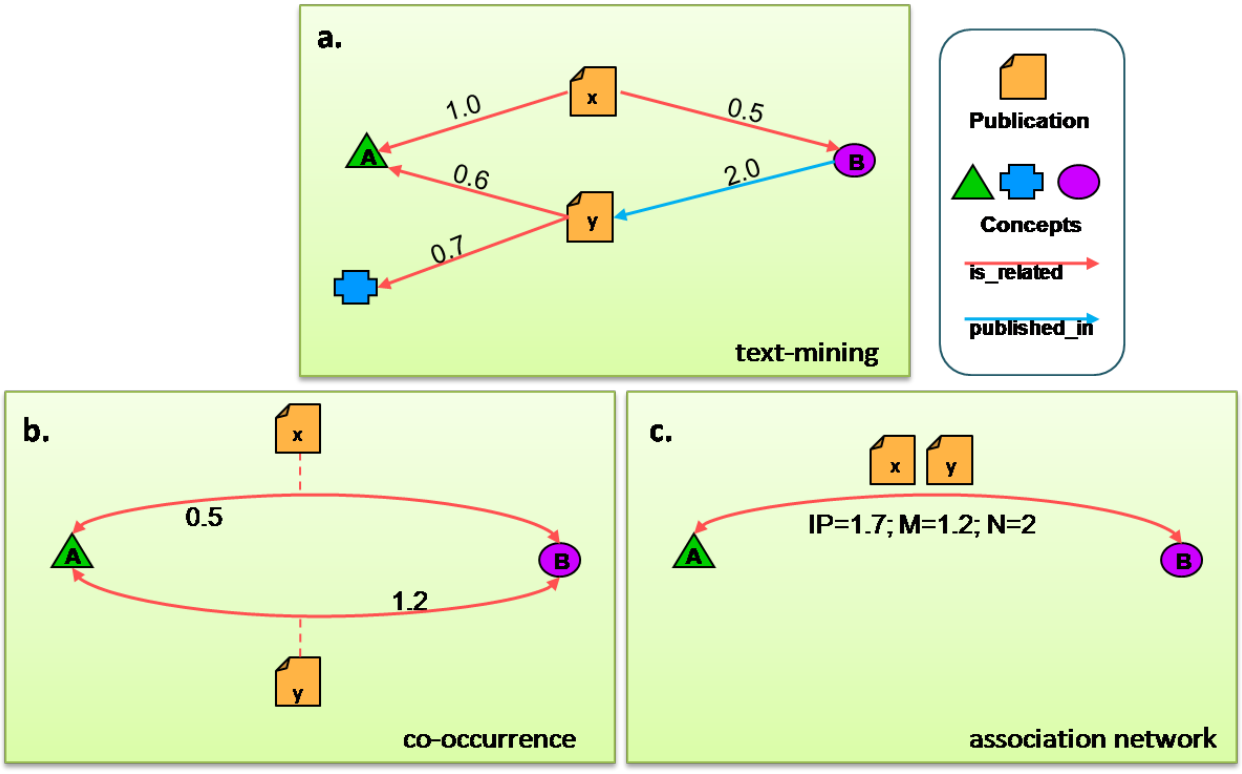

Figure 4.1: a) The outcome of the data integration and entity recognition based mapping step. Blue relations are manually curated citations from databases and are given a fixed high score. Red relations are based on automated text mining methods and are given Lucene scores. b) The co-occurrence step creates a direct link between the co-cited concepts within each publication (qualifier of the relation) and calculates the score as the product of the two original tf-idf scores. c) The final association step combines all relations from the previous step into one single relation and different scores are calculated to determine the strengths of the association.

4.2.3 Using co-occurrence to build weighted association networks

The final component of the text mining plugin is a transformation step that simplifies the Ondex knowledge network, computes association weights and combines text-based

evidence. This enables filtering of weak associations to make subsequent data mining and visualisation more effective.

If more than one concept is linked to the same publication (abstract or title) then this step directly connects them with a new *cooccurrs_with* relation. For example, if two concepts A and B are cited in the same publication, a relation of type *cooccurrs_with* is created and the product of the two original Lucene scores assigned as the combined score of the relation (Figure 4.1.b). The combined score represents the relevance of a document d for the query "a AND b", with a, b being possible concept names of A and B. It is often the case that two concepts may be cited together more than once in different publications. For each pair A-B, the following quantities were calculated and assigned to the *cooccurrs_with* relation: (i) the inner product of the scores (**IP** = $S_i\ x_i\ y_i$) where the index i ranges over the co-citations of the pair at hand, with S_i being the sum function, with x_i being the A score and y_i being the B score in the i^{th} co-citation; (ii) **M** = $Max_i(x_i\ y_i)$, with i ranging over the co-citations of the pair at hand; (iii) **N** = number of documents in which A and B were co-cited. An illustration of all three metrics is shown in Figure 4.1.c. During the derivation of co-occurrence data, the manually curated *published_in* relations are processed in an equivalent way to the *occurrs_in* relations but with a fixed score. To reflect the confidence in the curated *published_in* relations, this score is set at 2.0 (arbitrary high number). In cases where a concept is linked to a publication with both types of relation (*occurs_in* and *published_in*), the higher score is considered.

4.3 Results

4.3.1 Proof-of-concept and evaluation of the text mining approach

This section presents a proof-of-concept study that was published in (Hassani-Pak et al. 2010). The study was repeated using the latest versions of the public databases. The manual evaluation, however, is time-consuming and therefore the results of the 2009 analysis are shown. The numbers in brackets indicate the size of the 2015 database versions.

The database UniProtKB-SwissProt (release 15.8) was searched using the keyword "*Arabidopsis*" (TaxID: 3702) and the set of 8,582 proteins was downloaded in UniProt-XML format (14,095 proteins on 9/9/2015). Using the Ondex UniProt parser, this subset was loaded as *Protein* concepts into the knowledge network. The subset also served as the

Arabidopsis protein name dictionary to be used for text mining. This set of UniProt proteins contained 13,502 curated links to published papers represented as *published_in* relations to be incorporated into the co-occurrence analyses. PubMed was used to retrieve all Medline articles that contained the keywords *"Arabidopsis thaliana"* in their abstract, title or MeSH header. On August 28th 2009, this resulted in 28,653 articles being retrieved (53,455 publications on 9/9/2015). This PubMed subset was downloaded in XML format, added to the network using the Ondex Medline parser and integrated with the publications from UniProt. A custom Plant Stress Ontology in tabular format was developed by collaborators at Warwick University. The ontology encompasses 33 concepts related to biotic and abiotic stresses such as the fungal disease Botrytis, Ethylene and Drought. This constituted the second dictionary that was added to the network using the Ondex Console.

The three steps taken to create the knowledge network in the Ondex Integrator are illustrated as a metagraph in Figure 4.2. The integrated input data for *Arabidopsis* protein that already had manually curated links to publications are shown by the orange *published_in* relations (Figure 4.2 a). The NER step of the text mining plugin then linked proteins and stresses to individual publications from the integrated corpus (see blue *is_related* relations in Figure 4.2 b). The NER step also assigned Lucene scores and evidence sentences extracted from the publications to each blue relation. In the final step, co-occurring protein-stress pairs were identified and connected with *is_related* relations (Figure 4.2. c). Publications serve as evidence of the association between co-occurring pairs of concepts. All *is_related* relations between protein-stress pairs are annotated with three different scores (see Methods section).

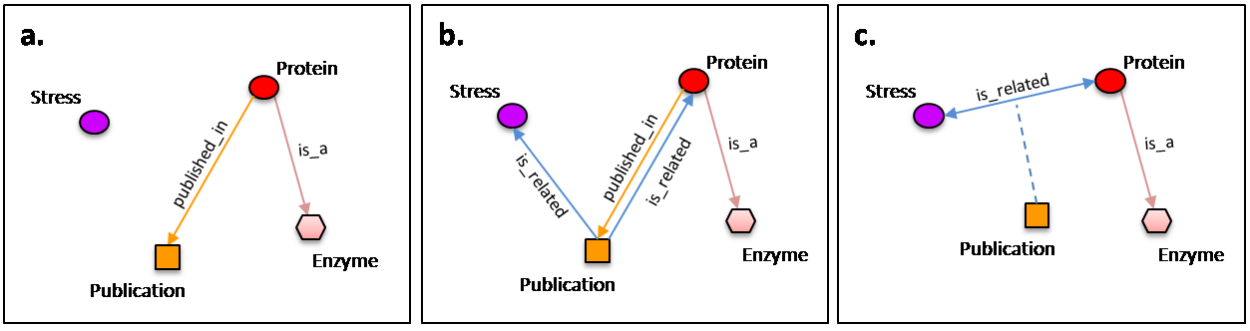

Figure 4.2: A metagraph illustrating the three steps towards creating a protein-stress association network using the Ondex text mining plugin that include a) Integrated input data, b) Named Entity Recognition (NER) and c) from co-occurrence to weighted association networks.

4.3.1.1 Mapping concepts to the corpus

In total, 52,430 protein and stress concepts were recognised from the corpus comprising 19,884 publications. Approximately 2.6 concepts were recognised per publication. Lucene scores on *occurs_in* relations indicate how significant a protein or stress concept is to a document in the *Arabidopsis* corpus. In order to understand how Lucene scores are distributed, a histogram was plotted for the whole dataset (Figure 4.3). A long-tailed distribution is observed that is characterised by a peak at approximately 0.2 and a long tail that extends to a maximum value of approximately 15.0 (not shown). Few Lucene scores (tf-idf) were greater than 3.9, however, therefore for presentation purposes we elected to combine observations greater than this value into an additional category "More". The data are a very close fit to a mixed distribution of two lognormal components. This observation can be explained by the fact that queries occurring in the title of a document receive an enhanced Lucene score. Choosing a tf-idf cut-off value of 0.956, separates the data very clearly into two subgroups those representing NER results based on the abstract (tf-idf < 0.956) and of the titles (tf-idf > 0.956). Thus for this particular dataset, at this cut-off value there is a 98.85% chance of classifying a tf-idf score correctly that is the result of a query matching either abstract or title of the document.

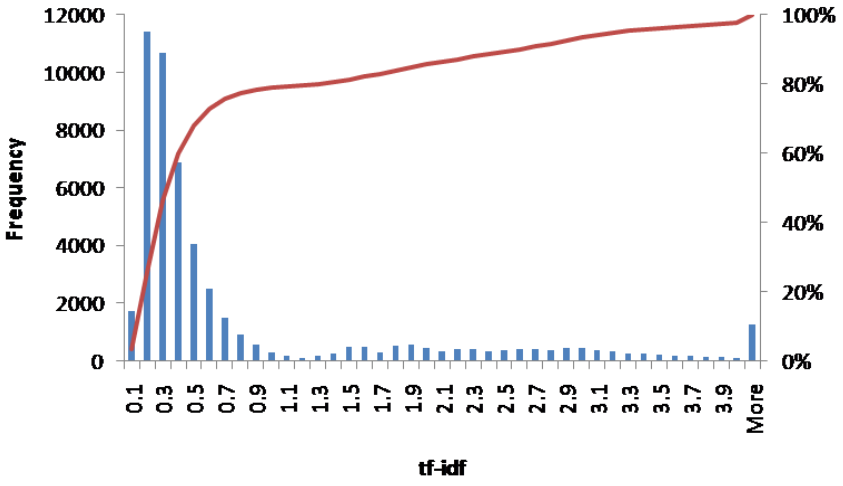

Figure 4.3: The distribution and cumulative frequency distribution (line) of Lucene scores (tf-idf) from 52,430 protein and stress concept relations identified in the *Arabidopsis* corpus by the NER search method (for details see Methods section). The 'More' category comprises about 1700 observations with tf-idf scores ranging between 4.0 and 15.0.

4.3.1.2 Weighted association networks

In a co-occurrence network, protein and stress concepts can be connected via one or several publications. In order to make the protein-stress relations more evident, the structure of the text mining based network is transformed to an association network. The resulting association network, after filtering out unconnected nodes, contained 3,145 proteins linked to 32 stresses by 10,777 relations. In other words, 36.7% of reviewed *Arabidopsis* proteins from UniProt (3,145 of 8,582) were co-cited with at least one Stress term from the Plant Stress Ontology database. On average, each co-cited protein was related to approximately 3.4 stresses and each stress related to 337 proteins.

Three different confidence scores were assigned to protein-stress associations. The **IP** score ranged between 0.01 and 347.26, the **M** score ranged between 0.01 and 26.86 and the **N** score ranged from 1 to 600. The highest IP score was found between “Light” and “Phytochrome A” (photoreceptor), while the lowest score was between “Hormone” and “ADP-glucose synthase” (a protein known to be regulated by hormones in rice cells (Zhu et al. 2011)). In the majority of cases, the IP and M scores were numerically lower than the co-citation number, N, but some opposite cases were observed. For example, the association between *ACBP4* and ethylene only had one co-citation, but the IP and M scores each had 13.51.

Comparison of the three scoring metrics over all 10,777 protein-stress pairs showed that IP and N are the most strongly correlated (r = 0.79) and IP and M to a lesser extent (r = 0.53). In Figure 4.4 it can be seen that IP is correlated with N but with a large variance especially for small N. For example, for N = 5 the IP score is very variable and ranges between 0.1 and 13.0. The M score on the other hand does not seem to be correlated with N, in the sense that a similar proportion at each N is greater than a given value.

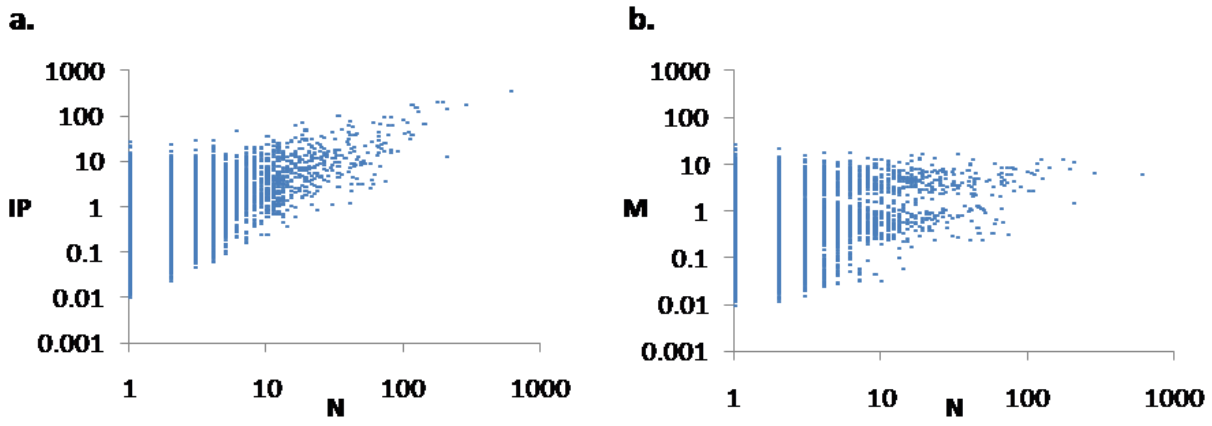

Figure 4.4: Comparison of Inner product of the scores (IP) vs number of co-citations (N) (a); and the Max scores (M) vs number of co-citations (N) for all 10,777 protein-stress pairs (b).

Because of the large size of integrated graphs, it can be useful to filter associations using confidence scores. The difficulty is finding a method to remove false positives while retaining true positives, thereby improving the signal-to-noise ratio while retaining sensitivity. Figure 4.5 illustrates an association network for all protein-stress pairs that are co-cited five times or more. The co-citation number is the simplest way to potentially reduce noise in such association networks. (Jenssen et al. 2001) examined the accuracy and type of interactions found among genes mentioned more than once or more than five times together. They found a decrease in the number of false positives as the number of co-occurrences increased. In this study, we found that sorting and filtering by IP and M metrics was in general more accurate than by simple co-citation frequency at reducing noise in the network, as both IP and M consider the frequency of terms in the corpus. None of the metrics seems to be superior overall, however, and the selection of the best metric may depend on the individual use case. Considering several metrics at the same time when analysing protein-stress associations seems to be the method of choice to highlight key associations and filter noise.

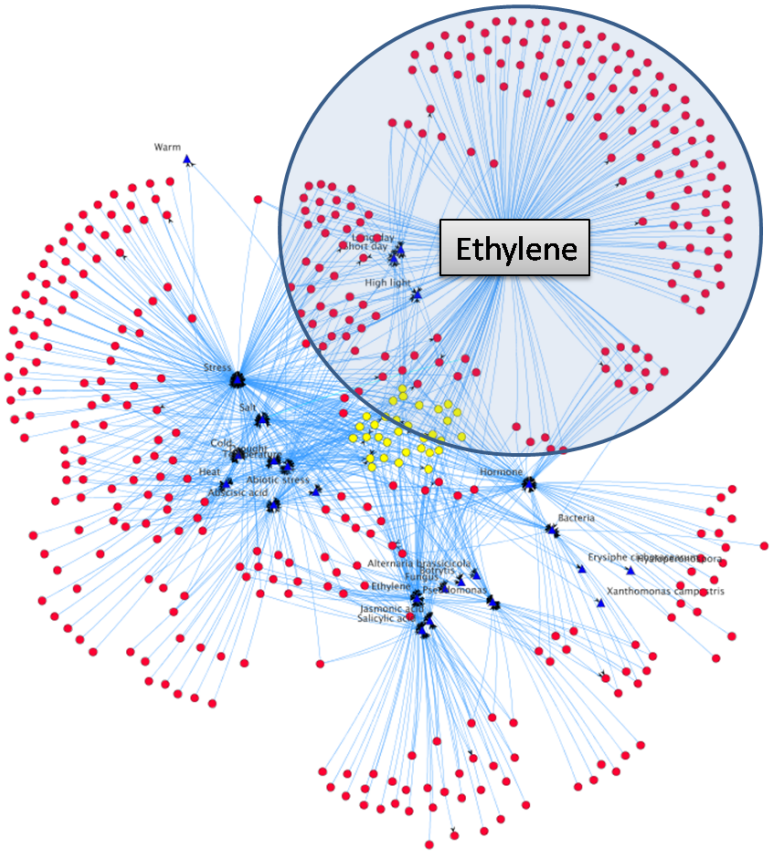

Figure 4.5: An example protein-stress association network based on 5 or more co-citations. The network contains 444 proteins (red circles), 25 stresses (blue triangles) and 1133 relations (blue edges). Proteins in the centre of the network (yellow circles) are implicated in several plant stress responses. The Ethylene association sub-network is highlighted.

4.3.1.3 Validation of ethylene-protein associations

To test the validity of the association (co-citation) network, we decided to focus on ethylene, a major plant hormone included in our Plant Stress Ontology. The ethylene association network contained 533 proteins and the same number of relations, with IP scores ranging from 0.016 to 202.35 and M scores from 0.016 to 13.91 (see Figure 4.5). To compare our predictions with a manually curated set, we chose the Arabidopsis Hormone Database (AHD; (Peng et al. 2009) as a gold standard. In AHD, 31 proteins are related to ethylene response based on manual curation of the published literature. Our text mining derived ethylene network contained 22 of the 31 AHD proteins giving a recall rate of 71.0%. The 9 proteins omitted from the network were not in the UniProtKB-SwissProt database at the time of the analysis (release 15.8, 2009), and were not included in our protein dictionary. Considering this fact, the actual recall rate was 100%.

In a second validation, we compared our network predictions to all 166 ethylene related proteins of AHD (including those extracted from GO) and achieved a recall rate of 44.8%. Table 4.1 shows the top 10 proteins (sorted by M score) from our analyses that are linked to ethylene but were not found in AHD. The four different confidence scores N, M and IP display the strength of the association. To classify these as true or false positives, evidence sentences from the literature were manually inspected. This can be done within Ondex as the association network contains the publications (titles and abstracts) and annotates relations with evidence sentences extracted from the publication. For example:

- *the interaction of ACBP4 and AtEBP may be related to AtEBP-mediated defence possibly via ethylene and/or jasmonate signalling.* [PMID: 18836139]
- *protein phosphatase 2C ABI1 modulates biosynthesis ratio of ABA and ethylene.* [PMID: 19705149]
- *a specific interaction of ETR1 with the histidine-containing transfer protein AHP1, supporting the idea that a phosphorelay module is involved in ethylene signalling* [PMID: 18384742]

Preliminary analyses of this manually validated subset of text mining based associations that are not included in AHD indicate that the inner product (IP) and the maximum scores (M) are highly significant ($P<0.0001$) correlates of a physiologically meaningful link between the protein and the stress.

Table 4.1: Top 10 ethylene related proteins (sorted by M) that the text mining analysis predicted but were not found in AHD. Each row displays the top hit PubMed ID and the year of publication. The next three columns show the weights of the protein-stress association according to the different scoring metrics, N, M and IP (see Methods section). The last column indicates whether the association is correct or not according to expert evaluation; the association is considered correct when True = yes.

ACCESSION	PROTEIN NAME	PUBMED	YEAR	N	M	IP	TRUE
AT3G05420	ACBP4	18836139	2008	1	13.51	13.51	yes
AT1G31812	ACBP6	18836139	2008	2	11.57	17.14	yes
AT3G03190	ATGSTF6	14617075	2003	7	7.36	15.75	yes

AT4G26080	ABI1	19705149	2009	10	6.66	12.22	yes
AT3G21510	AHP1	18384742	2008	3	6.60	6.70	yes
AT1G75040	PR-5	15988566	2005	12	5.18	5.47	yes
AT2G45820	Remorin	9159183	1997	4	5.04	6.77	no
AT3G11410	PP2CA	19705149	2009	1	5.00	5.00	yes
AT1G09570	Phytochrome A	8703080	1996	11	4.79	8.47	no
AT1G04240	IAA3	19213814	2009	3	4.54	5.14	yes

4.3.2 Extending Ondex workflows with text mining

The text-mining component is an essential part of the data integration workflow which creates the genome-scale knowledge networks. It extracts facts from the scientific literature in order to establish novel links between genes and Trait Ontology terms that did not exist before. Here we describe the datasets and methods for extending the plant knowledge networks using text mining. The animal versions are based on similar principles but using different datasets.

The results presented here are based on a corpus composed of Arabidopsis-related publications from PubMed and TAIR. PubMed was searched for articles that contain the keyword “Arabidopsis” in their abstract, title or MeSH header (52,561 publications as of 22/06/2015). Additionally the Arabidopsis TAIR gene-publication file is used (ftp://ftp.arabidopsis.org/home/tair/User_Requests/Locus_Published_20130305.txt) which contains references to 22,201 publications. This set added 255 (1.14%) citations to the corpus as the majority of the citations already existed in the initial PubMed corpus. Arabidopsis gene names and synonyms for 27,416 genes were downloaded from Phytozome. An alternative dataset could consist of non-Arabidopsis plant proteins and publications from UniProtKB-SwissProt containing 22,596 proteins with 18,519 curated literature references to 7,962 publications (as of 22/06/2015). The Gramene Trait Ontology (TO) (Jaiswal et al. 2002) encompassing over 1300 trait terms (names and synonyms) is used as the second dictionary for plant phenotypic descriptions, such as glutinous endosperm, disease resistance, plant height, shoot branching, photosensitivity and flowering time (Figure 4.6).

Figure 4.6: Excerpt of the Gramene Trait Ontology.

An Ondex workflow was designed to parse all datasets and execute the text mining steps (Code 3.1). As the first step in the text mining process, the '*tmbased*' mapping method creates *occurs_in* relations between Gene and Publication concepts as well as TO and Publication concepts based on occurrence of names or synonyms in the title or abstract of the publication. The second step '*coocurrence*' transforms the graph into a weighted association network which has direct links between Gene and TO concepts (Figure 4.7). These two steps connect 5553 Arabidopsis genes to 409 TO terms based on 18,341 co-citations. Each *cooccurs_with* relation is assigned IP, M and N scores as defined in the Methods section. Since the text mining plugin was published in 2010, it has been extended with a new score which counts the number of evidence sentences (ES) when two concept names co-occur on sentence level.

```
<?xml version="1.0" encoding="UTF-8"?>
<Ondex version="3.0">
  <Workflow>
    <Graph name="memorygraph">
      <Arg name="GraphName">default</Arg>
      <Arg name="graphId">default</Arg>
    </Graph>
    <Parser name="phytozome">
      <Arg name="InputDir">phytozome/Arabidopsis/</Arg>
      <Arg name="TaxID">3702</Arg>
      <Arg name="AccDataSource">TAIR</Arg>
      <Arg name="ChromosomeNumber">5</Arg>
      <Arg name="PreferredSynonyms">false</Arg>
      <Arg name="graphId">default</Arg>
    </Parser>
    <Parser name="genericobo">
      <Arg name="OboType">TO</Arg>
      <Arg name="Obsoletes">false</Arg>
      <Arg name="InputFile">ontologies/to.obo</Arg>
      <Arg name="graphId">default</Arg>
```

```
    </Parser>
    <Parser name="medline">
      <Arg name="InputFile">pubmed/pubmed_result_arabidopsis.xml</Arg>
      <Arg name="ImportCitedPMIDs">true</Arg>
      <Arg name="graphId">default</Arg>
    </Parser>
    <Mapping name="tmbased">
      <Arg name="OnlyPreferredNames">false</Arg>
      <Arg name="UseFullText">false</Arg>
      <Arg name="Search">exact</Arg>
      <Arg name="graphId">default</Arg>
      <Arg name="ConceptClass">Gene</Arg>
      <Arg name="ConceptClass">TO</Arg>
    </Mapping>
    <Transformer name="coocurrence">
      <Arg name="TargetConceptClass">Publication</Arg>
      <Arg name="graphId">default</Arg>
    </Transformer>
</Workflow>
</Ondex>
```

Code 3.1: Ondex workflow with 5 steps that parse Phytozome (“phytozome” parser), Trait Ontology (“genericobo” parser) and PubMed (“medline” parser). Map Gene and TO concepts to publications (“tmbased” mapping) and transform the graph into a weighted association network (“coocurrence” transformer).

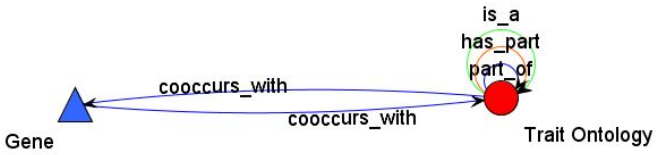

Figure 4.7: Metagraph after running the above workflow and filtering Publication concepts

To demonstrate the output of the text mining we focus on flowering time related traits that are represented in the Trait Ontology. Figure 4.8 shows that most Arabidopsis genes co-occur with the term ‘flowering time’ (TO:0002616) and its child term ‘days to flower’ (TO:0000344). The more specific terms such as days to tassel flowering, days to silk, male flowering and female flowering did not co-occur with any Arabidopsis gene names.

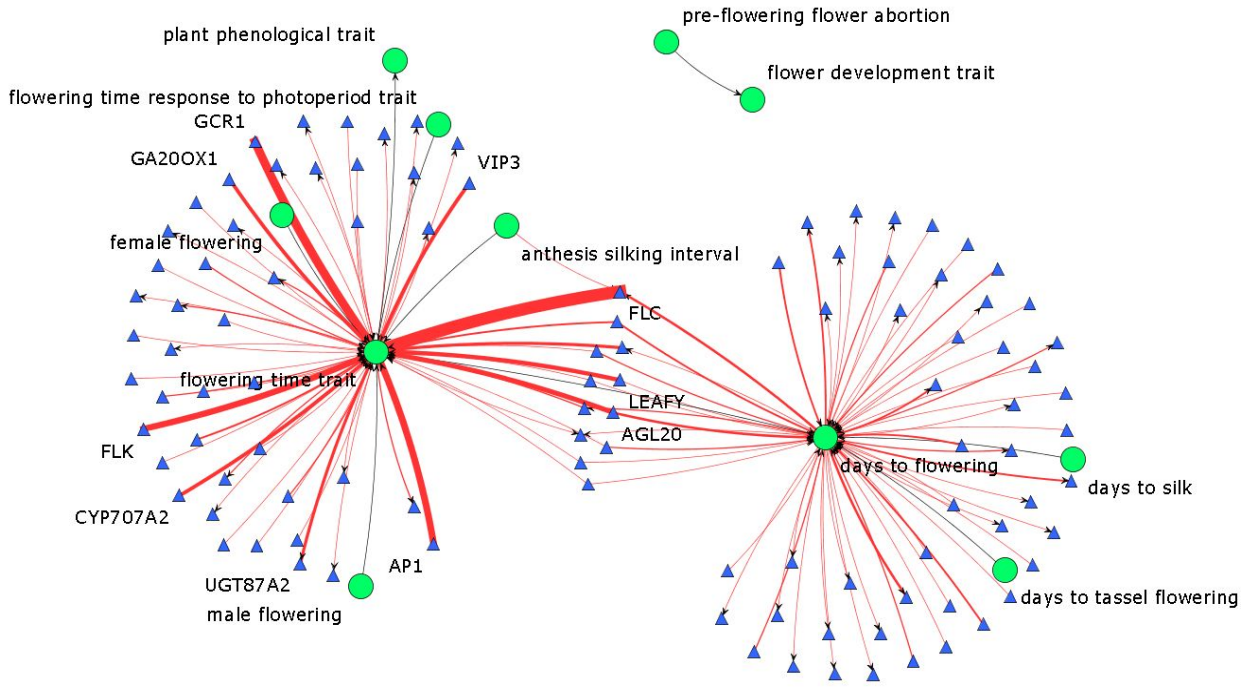

Figure 4.8: Weighted association network of Trait Ontology concepts (green circles) and Arabidopsis genes (blue triangle). Red relations represent *cooccurrs_with* relations; the widths reflects the IP score of the association. Gene labels are shown for genes that are connected through relations with highest IP scores. Grey relations between ontology terms represent *is_a* relations.

4.4 Discussion

A major challenge for those working with high-throughput 'omics datasets is to contextualise new results by comparing them with information, either from structured databases or the scientific literature. The development of data integration and text mining methods have, however, largely been conducted independently. This chapter described the design and implementation of a Java-based text processing plugin for the Ondex data integration framework. The plugin consists of information retrieval (IR), named entity recognition (NER) and co-occurrence methods. The requirements were to recognise the occurrence of specific concepts in the literature and to build weighted association networks that can integrate text mining based relations and curated literature references. Additionally, it was required to demonstrate that the entire process from data integration to text mining can be covered within simple and reproducible Ondex workflows.

The Java-based text mining plugin that has been developed features information retrieval techniques and dictionary-based NER to recognise concepts (names and synonyms) in an application case dependent corpus. In cases where the data integration step has used a source containing curated literature references (e.g. such as from TAIR or UniProt) these links are incorporated into the text mining and are given a stronger weight. Using co-occurrence relationships between *Concept Classes* of interest the plugin can generate weighted association networks. Our work demonstrates that basic text mining features can bring added value to data integration and be used as a simple baseline when developing more sophisticated systems. In the future, we hope to include machine-learning-based and rule-based approaches for entity recognition and relation extraction respectively. These should extract more information from the literature and have increased sensitivity and specificity. New text processing frameworks such as GATE and Apache UIMA (Cunningham et al. 2013; “Apache UIMA” n.d.) will thereby play a crucial role.

The two application cases demonstrated the usability of the text mining plugin as part of simple Ondex workflows. Focusing on gene-trait and protein-stress association (co-occurrence) networks from *Arabidopsis*, we showed that workflows incorporating text mining produced meaningful results consistent with manually curated data. We identified proteins involved in ethylene response in *Arabidopsis*. Validating against the manually curated *Arabidopsis* Hormone Database (AHD) showed that our method produced a recall of 71.0%. The lack of a proper gold standard dataset, made it impossible to calculate precision values and F-measures (ratio between recall and precision). For example, our approach identified many more significant associations that are not present in AHD (yet) but were considered plausible by domain experts. Overall, our recall rates are similar to other dictionary or rule-based approaches applied, for example, in cancer research that achieve NER rates of 60-70% (Kang et al. 2013) but lower than machine learning NER approaches that achieve recall rates of 80-90% (Spasić et al. 2014).

It is well known that text mining results can be noisy. The challenge lays in distinguishing significant information from noise. Improving the signal-to-noise ratio is a demanding task. We used four different scoring metrics IP, M, N and ES to score the confidence of associations and made a first attempt to compare them. Many alternative weighting schemes for the definition of association confidence between two concepts exist, and it is easy to envisage how a weighted mean of multiple scores may compensate for the weaknesses in any particular scoring scheme. Presuming large enough training sets are available in the

future, optimisation methods could be applied to pick the best weighting schemes or to train machine learning based methods. Meanwhile, interactive filters combined with visualisation methods in Ondex provide domain experts with tools to explore the results from both text mining and data integration in an intuitive and semi-automated way.

Our text mining approach is currently applied to titles and abstracts from PubMed articles. The approach, however, is general and could be employed with full documents if these were available. Since there are unlimited ways of expressing the same thing in free text, dictionary-based NER techniques are more successful if they can choose from a greater number of synonyms. In this respect, the Trait Ontology is a rather immature ontology for text mining applications. The number of synonyms is low compared with the Gene Ontology (GO). For example the trait term ‘shoot branching’ (TO:0002639) contains no synonyms, whereas the biological process term ‘shoot branching’ (GO:0010223) contains several (secondary shoot formation, auxiliary shoot formation, axillary shoot formation, axillary shoot system formation, shoot branching). This could be one reason why we only found one third of the TO terms to be co-cited with Arabidopsis gene names. Other alternatives such as the Crop Ontology have similar issues and therefore stronger efforts are needed in the future to better integrate ontologies and make ontologies compatible with text mining applications.

In conclusion, in this chapter it was shown that there are significant benefits to being able to combine data integration and text mining. The text mining plugin developed here extends the data integration framework Ondex with basic text processing functionality. It gives Ondex the capability to create novel links between heterogeneous data sources using the scientific literature. It is flexible and computationally undemanding, and therefore practical to integrate into workflows for building enhanced knowledge networks. The enhanced knowledge networks can be exploited, as described in the next chapters, through automated data mining and manual data exploration steps to accelerate biological discovery and hypotheses generation processes.

5 SEARCHING KNOWLEDGE NETWORKS AND RANKING GENES

A genome-scale knowledge network (GSKN) can be very large and highly connected. Efficient methods are needed to search such networks, extract biologically plausible paths through the network and use these to identify and rank potential candidate genes. This chapter explains the requirements to be taken into account when searching a GSKN. It then presents a newly developed method, called KNETscore, for ranking candidate genes based on the notion of gene-evidence networks. Proof-of-concept and validation of the methodology are presented in the results, before the limitations and future work are discussed at the end.

5.1 Background

As described in Chapter 3, GSKNs are labelled and directed multi-graphs that include all genes and proteins of an organism as concepts and link them directly or indirectly with a variety of other concept types such as proteins, pathways, publications, ontology terms, etc. GSKNs are indexed by converting concepts and relations into Lucene documents using fields to represent their attributes (see Chapter 4). Searching a knowledge network with search terms becomes equivalent to searching a collection of text documents. Due to the large number of documents in GSKNs, measures are required to rank documents based on their importance to the search terms. The measure of inverse document frequency (IDF) is a well-established method in information retrieval (Sparck Jones 1972). IDF is based on counting the number of documents d in the collection D being searched which contain (or are indexed by) the term t. The inverse document frequency is defined as:

$$idf(t, D) = log \frac{|D|}{|\{d \in D : t \in d\}|}$$

A search term which occurs in many documents is seen as a bad discriminator and is given less weight than one which occurs in few documents. The IDF is a measure of term specificity, it does not rank documents, because all documents containing a search term will have the same IDF score. This can be coupled with TF (the frequency of a term within the document itself, the more the better) which is defined as:

$$tf(t, d) = \frac{|t \in d|}{|d|}$$

TF*IDF which involves multiplying the IDF measure by the TF measure can be used for ranking documents by their relevance to the search terms. It emerged from extensive empirical studies of combinations of weighting factors (Salton and Yang 1973) and has proved very robust and difficult to beat, even by much more refined models and theories (Robertson 2004).

Nodes in a GSKN containing a certain search term will be referred to as evidence concepts. Once evidence concepts are identified, the aim is to find the genes that are linked to these evidence concepts. The links do not need to be individual direct relations, but can also be formed from a series of transitive relations which define indirect paths through the network. However, given the density and high connectivity of GSKNs, a path that connects two concepts nearly always exists. Therefore, one objective of this chapter is to develop a method that can distinguish biologically plausible paths from non-plausible paths.

Once the genes are identified that are directly or indirectly linked to the evidence concepts, the second objective of this chapter is to develop a method for scoring candidate genes based on the supporting evidence. It is important that the gene scoring method reflects how relevant a search term (e.g. flowering time) is to a gene in a collection (genome). A high score should be given to genes with frequent and specific evidence concepts. The scoring method needs to be generic and compatible with any underlying GSKN. Additionally, it is required that the scoring method is not only accurate but can also be computed rapidly for thousands of potential candidate genes in large plant and animal genomes.

5.2 Methods

5.2.1 Gene-evidence networks and semantic motifs

We define a **gene-evidence network** as a restricted gene neighborhood network that only contains biologically plausible paths for any given gene. Biologically plausible paths are ones that allow evidence (knowledge) to be transferred to a gene of interest, for example through ortholog or protein-protein interaction relationships. The difference between a gene-evidence network and an unrestricted gene neighbourhood network is explained below (Figure 5.1).

The direct neighbourhood (n=1) of a gene would retrieve all concepts with in-going or out-going relations, for example, the proteins it encodes, linked publications, SNPs etc. Increasing the neighbourhood to n=2 would extract protein orthologs, protein-protein

interactions and GO annotations, however it would also add unrelated new genes that were cited in the publications from n=1. At n=3 we could see a rapid expansion of the gene neighbourhood as new information such as annotations of orthologous and interacting proteins would be added but also information about those new genes seen at n=2, or other unrelated proteins annotated to the GO concepts at n=2. Biological useful knowledge can be transferred to the target gene through even longer paths. For example, a path of length n=5 such as "<gene> *encodes* <protein> *ortholog* <protein> *encodes* <gene> *interacts* <gene> *involved_in* <biological_process>" can provide weak but still useful information for candidate gene discovery. There are fairly obvious traversals through the network of n=5 and even more that can be useful, provided these biologically plausible paths are known. Since GSKNs are highly connected, well formed searches are needed to avoid an exponential growth of the graph traversal. This is not only problematic because of the sheer volume of information but also because it violates the definition of a gene-evidence network which requires it only to contain evidence that is biologically plausible when transferred to a gene of interest.

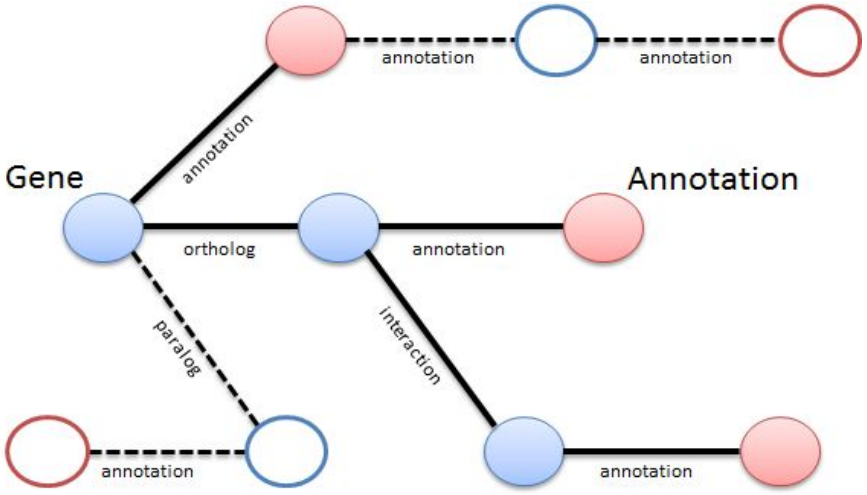

Figure 5.1: Illustration of a gene-evidence network as derived through "biologically plausible" semantic motifs. Blue nodes represent *Gene* concepts; red nodes are annotations such as *GO, TO, EC, Pathway, Publication* concepts. A path that goes via bold edges is valid (biologically meaningful path that allows annotations to be transferred to the seed gene). A path that goes via dashed edges is invalid. A gene-evidence network contains only filled nodes, whereas a gene-neighbourhood network would also contain the unfilled nodes.

To create gene-evidence networks, biologically plausible paths that can subsequently be extracted have to be formally specified in a knowledge network. Therefore, we use the notion of **semantic motif** to define a path through the metagraph that matches a particular

biologically plausible metamodel. Semantic motifs (Biemann et al. 2016) can start with gene concepts and end with other biological entities or functional annotations. For example, valid motifs going from gene concepts to annotations such as GO concepts can include paths that go via orthologs and interaction partners but exclude any path that goes via paralogs (Figure 5.1). Semantic motifs can be formally defined and extracted from the knowledge networks using the Metadata-based Graph Query Engine (**MGQE**) which was implemented by Matthew Hindle (Hindle 2012). The method uses a parallelized implementation of the breadth-first search algorithm starting with a set of seed concepts and at each depth retaining concepts and relations that match a particular metamodel (*Concept Class* or *Relation Type*). The MGQE query syntax is somewhat idiosyncratic but it is effectively a low level form of a deductive database query language (Ramakrishnan, Raghu, and Ullman 1995) and can be illustrated with the help of the example in Figure 5.1 which contains the following three semantic motifs (not dashed paths):

<1* Gene> *annotation* <3^ GO>
<1* Gene> *ortholog* <2 Gene> *annotation* <3^ GO>
<1* Gene> *ortholog* <2 Gene> *interacts* <2 Gene> *annotation* <3^ GO>

The MGQE query syntax for specifying these semantic motifs is shown in Table 5.1.

Table 5.1: MGQE query syntax for specifying three semantic motifs. Left column contains the *Concept Class* where * means “start” and ^ means “end”. Right column shows the permitted transitions between *Concept Class* pairs (see text for more details).

Concept Class	**Transition**
1* Gene 2 Gene 3^ GO	1-3 annotation 1-2 ortholog 2-2 interacts 5 2-3 annotation

The first column defines the Concept Classes that constitute valid semantic motifs. Concept Classes are labelled with ‘*’ or ‘^’ to indicate the start or end of a semantic motif respectively. For example the Concept Class ‘*Gene*’ appears twice because it can either be the start node (1*) or an intermediate node (2) of a semantic motif. The second column defines valid transitions between Concept Class pairs. Two valid transitions are possible from the starting node *Gene*, transition 1-3 is of type *annotation* and transition 1-2 of type *ortholog*. The first transition (1-3) leads into an end GO node which therefore results in obtaining the first, and

also shortest, valid semantic motif. The next valid transitions are the 2-x transitions where x can be an intermediate or end Concept Class from the first column. In case x is an intermediate node then again further valid transitions are required in order to create a complete semantic motif. When the Concept Classes on each side of the transition are identical (e.g. 2-2), this transition would be executed recursively. It is possible, however, to deny multiple identical transitions by restricting the maximum path length. This can be specified as part of the transition definition, for example '5', and is defined as the total number of nodes and edges in a path. A transition is only applied if the specified path length limit has not been exceeded.

The advantage of the MGQE syntax is that it does not require repetitive statements to be made. For example, adding a fourth semantic motif (in red) to the above list:

<1 Gene*> *annotation* <3 GO^>
<1 Gene*> *ortholog* <2 Gene> *annotation* <3 GO^>
<1 Gene*> *ortholog* <2 Gene> *interacts* <2 Gene> *annotation* <3 GO^>
<1 Gene*> *ortholog* <2 Gene> *co-cited* <2 Gene> *annotation* <3 GO^>

would only require a single new transition (last line) to be added to the MGQE semantic motif definition as shown in Table 5.2

Table 5.2: Same as Table 5.1 but additionally relations of type co-cited between Gene-Gene pairs are permitted. Recursions are restricted by setting the maximum path length to 5.

Concept Class	**Transition**
1* Gene 2 Gene 3^ GO	1-3 annotation 1-2 ortholog 2-2 interacts 5 2-2 co-cited 5 2-3 annotation

As part of this work, a minor extension was made to the original MGQE implementation in order to support multiple relation types between same Concept Classes (e.g. 2-2 ortholog and 2-2 co-cited).

5.2.2 Extracting gene-evidence networks in wheat

A list of semantic motifs (Table 5.3) was created that contains biologically plausible paths to be included in the gene-evidence networks for wheat. This list contained 57 distinct paths that can be projected onto the metagraph of the wheat knowledge network. The two halves of the table are nearly identical, however the first half uses the *has_similar_sequence (h_s_s)* Relation Type in the fourth column, while the bottom half uses the *ortho* Relation Type in the same column. This is because the wheat knowledge network contains two different types of homology relations.

Table 5.3: The 57 semantic motifs that were chosen to be biologically plausible paths in the wheat knowledge network. Each line corresponds to a single semantic motif path. To be read from left to right, white columns indicate the *Concept Class* and grey columns the *Relation Type* of the path.

	1		2		3		4		5	
Gene	enc	Protein								
Gene	enc	Protein	h_s_s	Protein						
Gene	enc	Protein	h_s_s	Protein	pub_in	Publication				
Gene	enc	Protein	h_s_s	Protein	has_function	MolFunc				
Gene	enc	Protein	h_s_s	Protein	participates_in	BioProc				
Gene	enc	Protein	h_s_s	Protein	located_in	CelComp				
Gene	enc	Protein	h_s_s	Protein	has_domain	ProtDomain				
Gene	enc	Protein	h_s_s	Protein	has_domain	ProtDomain	has_function	MolFunc		
Gene	enc	Protein	h_s_s	Protein	has_domain	ProtDomain	participates_in	BioProc		
Gene	enc	Protein	h_s_s	Protein	has_domain	ProtDomain	located_in	CelComp		
Gene	enc	Protein	h_s_s	Protein	cat_c	EC				
Gene	enc	Protein	h_s_s	Protein	cat_c	EC	equ	MolFunc		
Gene	enc	Protein	h_s_s	Protein	is_a	Enzyme				
Gene	enc	Protein	h_s_s	Protein	is_a	Enzyme	ca_by	Reaction		
Gene	enc	Protein	h_s_s	Protein	is_a	Enzyme	ca_by	Reaction	part_of	Path
Gene	enc	Protein	h_s_s	Protein	enc	Gene				
Gene	enc	Protein	h_s_s	Protein	enc	Gene	pub_in	Publication		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	has_function	MolFunc		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	participates_in	BioProc		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	located_in	CelComp		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	has_obser_pheno	Phenotype		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	cooc_wi	TO		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene		
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	pub_in	Publication
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	has_function	MolFunc
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	participates_in	BioProc
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	located_in	CelComp
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	has_obser_pheno	Phenotype
Gene	enc	Protein	h_s_s	Protein	enc	Gene	it_wi	Gene	cooc_wi	TO
Gene	enc	Protein	ortho	Protein						
Gene	enc	Protein	ortho	Protein	pub_in	Publication				

Gene	enc	Protein	ortho	Protein	has_function	MolFunc				
Gene	enc	Protein	ortho	Protein	participates_in	BioProc				
Gene	enc	Protein	ortho	Protein	located_in	CelComp				
Gene	enc	Protein	ortho	Protein	has_domain	ProtDomain				
Gene	enc	Protein	ortho	Protein	has_domain	ProtDomain	has_function	MolFunc		
Gene	enc	Protein	ortho	Protein	has_domain	ProtDomain	participates_in	BioProc		
Gene	enc	Protein	ortho	Protein	has_domain	ProtDomain	located_in	CelComp		
Gene	enc	Protein	ortho	Protein	cat_c	EC				
Gene	enc	Protein	ortho	Protein	cat_c	EC	equ	MolFunc		
Gene	enc	Protein	ortho	Protein	is_a	Enzyme				
Gene	enc	Protein	ortho	Protein	is_a	Enzyme	ca_by	Reaction		
Gene	enc	Protein	ortho	Protein	is_a	Enzyme	ca_by	Reaction	part_of	Path
Gene	enc	Protein	ortho	Protein	enc	Gene				
Gene	enc	Protein	ortho	Protein	enc	Gene	pub_in	Publication		
Gene	enc	Protein	ortho	Protein	enc	Gene	has_function	MolFunc		
Gene	enc	Protein	ortho	Protein	enc	Gene	participates_in	BioProc		
Gene	enc	Protein	ortho	Protein	enc	Gene	located_in	CelComp		
Gene	enc	Protein	ortho	Protein	enc	Gene	has_obser_pheno	Phenotype		
Gene	enc	Protein	ortho	Protein	enc	Gene	cooc_wi	TO		
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene		
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	pub_in	Publication
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	has_function	MolFunc
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	participates_in	BioProc
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	located_in	CelComp
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	has_obser_pheno	Phenotype
Gene	enc	Protein	ortho	Protein	enc	Gene	it_wi	Gene	cooc_wi	TO

This list was subsequently translated into the MGQE query syntax (Table 5.4) which is used to query each seed gene in the wheat knowledge network for the existence of any of these 57 paths. Not every gene node will necessarily contain all 57 paths, however, the ones it contains are extracted and their union is taken to produce an individual gene-evidence network for each of the 99,386 wheat genes. Gene-evidence networks provide the core elements for identifying and ranking candidate genes as described next.

Table 5.4: Definition of 57 different semantic-motifs for extracting gene-evidence networks from the wheat knowledge network. Both columns use the formal MGQE syntax as described in [Citation error].

Concept Class (*=start ^=end)	**Transition**
1* Gene 2^ Publication 3^ MolFunc 4^ BioProc 5^ CelComp 7^ Protein	1-10 enc 1-7 enc 10-10 h_s_s 5 10-7 h_s_s 5 10-10 ortho 5 10-7 ortho 5

8^	Gene	10-2	pub_in
9	Gene	10-3	has_function
10	Protein	10-4	participates_in
11	ProtDomain	10-5	located_in
12	EC	10-9	enc
13^	Phenotype	10-8	enc
14^	ProtDomain	10-11	has_domain
15^	EC	10-12	cat_c
16^	TO	10-14	has_domain
17	Enzyme	10-15	cat_c
177^	Enzyme	9-2	pub_in
18	Reaction	9-3	has_function
188^	Reaction	9-4	participates_in
19^	Path	9-5	located_in
		9-9	it_wi 8
		9-8	it_wi 8
		9-13	has_observ_pheno
		11-3	has_function
		11-4	participates_in
		11-5	located_in
		12-3	equ
		9-16	cooc_wi
		10-17	is_a
		10-177	is_a
		17-18	ca_by
		17-188	ca_by
		18-19	part_of

5.2.3 Gene Ranking

Here we describe the development of a method for scoring and ranking gene-evidence networks based on their relevance to certain search terms. The scoring method consists of three components TF*IDF, IGF and EDF that are first explained individually and then brought together to form the gene scoring function.

5.2.3.1 Inverse Gene Frequency (IGF)

The TF*IDF score reveals how specific a search term is to a document; it does not imply how specific the document is to a gene in question. A publication can for example receive a high TF*IDF score because not many other documents in the knowledge network contain the same search term. If the publication is, however, cited (linked) by hundreds of genes it should receive a smaller weight than a publication which is linked to one or two genes only.

It was therefore necessary to develop a second metric that incorporates the specificity of a document to a gene as the number of genes a document is linked to. It is important to remember that a search term can occur in one or many documents and that a document can occur in one or many gene-evidence networks ('occur in' is taken as shorthand for is part of a gene's evidence network).

Assume there are N genes (or N gene-evidence networks) in the collection (knowledge network), and that document d_i occurs in n_i of them. Then a measure similar to inverse document frequency, called **IGF**, can be defined that weights a document d_i according to its specificity to the genes in the collection:

$$igf(d_i) = log\ \frac{N}{n_i}$$

which can be redefined as a probability (C. E. Shannon 1948) that a random gene g would contain the document:

$$P(d_i) = P(d_i\ occurs\ in\ X_g)\ \approx \frac{n_i}{N}$$

$$igf(d_i)\ =\ -log\ P(d_i)$$

Note that IGF can be computed independently from the search terms because it only requires knowledge of gene-evidence networks and is not influenced by the input terms themselves. It is frequently assumed that term or document weights are additive (Robertson 2004). So that a sum or mean weight of a set of documents can be defined as:

$$igf_{sum}(d_1..d_n) = \sum_{i=1}^{n} igf(d_i) \qquad \text{(SUM)}$$

$$igf_{mean}(d_1..d_n) = \frac{1}{n}\sum_{i=1}^{n} igf(d_i) \qquad \text{(MEAN)}$$

A simple gene scoring function could thus be based on the sum or mean of IGF weights of all evidence documents of a gene. Given two scenarios: gene A with one evidence document that has a high IGF weight vs. gene B with many evidence documents that have low IGF weights. Taking the mean would always rank gene A higher than gene B. Taking the sum might rank gene B above gene A if the sum of IGF scores of a B > A. Both methods are taken forward and evaluated in the results section.

Because of our assumption that TF*IDF and IGF weights are independent and additive, it would be possible to multiply TF*IDF by IGF to retrieve a single measure of document specificity.

5.2.3.2 Evidence Document Frequency (EDF)

Documents can therefore now be scored according to their specificity to search terms (TF*IDF) and their specificity to genes (IGF). A final measure is required that can specify the relevance of a search term to the gene-evidence network as a whole. The more evidence documents there are that contain the search terms the higher the relevance of the gene to the search term will be. This needs to be normalised by the total number of documents in a gene-evidence network since larger networks are more likely to contain more evidence documents. Similar to the term-frequency measure which counts the number of terms in a document and normalises by the length of the document, **EDF** can be defined as the frequency of evidence documents in a gene-evidence network.

Thus, given a gene-evidence network for a gene g with a set of documents $X_g=\{d_1,d_2,\ldots,d_n\}$ that are directly or indirectly linked with g. The evidence document frequency of a gene g and a term t can be defined as:

$$edf(t, X_g) = \frac{|t \in d : d \in X_g|}{|X_g|}$$

where $|t \in d : d \in X_g|$ is the number of evidence documents for gene g that contain the term t.

A simple gene scoring function could be solely based on EDF. An example given two scenarios: gene A with a small gene-evidence network of size 2 and gene B with a large gene-evidence network of size 100. Both of them have only one evidence document that contains the search term. Gene A would get a EDF score of 0.5 (1/2) while gene B would get a score of 0.01(1/100). For this reason, gene A would have a higher relevance to the search term compared to gene B.

5.2.3.3 Gene scoring function (KNETscore)

All three measures TF*IDF, IGF and EDF have unique characteristics. Combining them can provide one single score that reflects the relevance and specificity of a query term to a gene in a collection.

Thus, given a gene g and its gene-evidence network that consists of a set of documents $X_g=\{d_1,d_2,\ldots,d_n\}$. A novel gene scoring function, called **KNETscore**, was developed that computes the relevance of a gene in the knowledge network to a search term t as follows:

$$KNETscore_{mean}(t,X_g) = edf(t,X_g) * \frac{1}{|t \in d_i : d_i \in X_g|} \sum_{t \in d_i : d_i \in X_g} tf * idf(t,d_i) * igf(d_i) \qquad \text{(MEAN)}$$

$$KNETscore_{sum}(t,X_g) = edf(t,X_g) * \sum_{t \in d_i : d_i \in X_g} tf * idf(t,d_i) * igf(d_i) \qquad \text{(SUM)}$$

The rest of this chapter will present results from the gene scoring method for both MEAN and SUM and evaluate the ranking approach using a set of known candidate genes.

5.3 Results

5.3.1 Characteristics of gene-evidence networks

A gene-evidence network represents biologically meaningful knowledge about a gene in question and is different to a gene neighbourhood network. Figure 5.2 shows the difference between those two types of networks. The gene-evidence network has less concepts than the unrestricted gene-neighbourhood network although it contains longer paths of length n=5, while the gene-neighbourhood networks was restricted to n=4. The result of a gene-evidence networks considers a gene to be, for example, related to ‘early flowering’ if any of its evidence concepts are 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 human domain expert should consider when judging if the gene is related to ‘flowering time’.

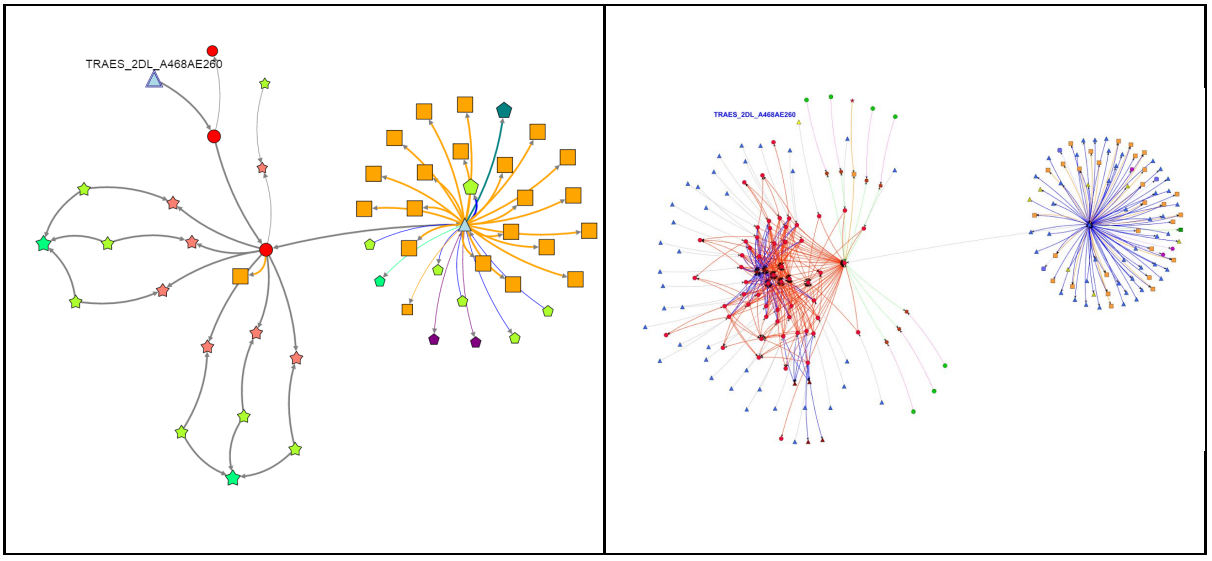

Figure 5.2: Wheat gene TRAES_2DL_A468AE260. On the left the gene-evidence network is shown and on the right the gene-neighbourhood network for n=4.

The size distribution of gene-evidence networks varies between the different GSKNs. Arabidopsis gene-evidence networks have on average 26.3 concepts where size two networks "<Gene> *encodes* <Protein>" are very rare since most genes are linked to various annotations (Figure 5.3 A). In contrast, the mean gene-evidence size in wheat GSKN has 42.9 concepts but nearly 10% of genes have a gene-evidence network of size two; these are genes for which the only available information are the proteins they encode (without homology or protein domain data) (Figure 5.3 B). The average gene-evidence networks in wheat are double the size of Arabidopsis networks because of homology and protein domain information that is available in the wheat GSKN but missing in Arabidopsis.

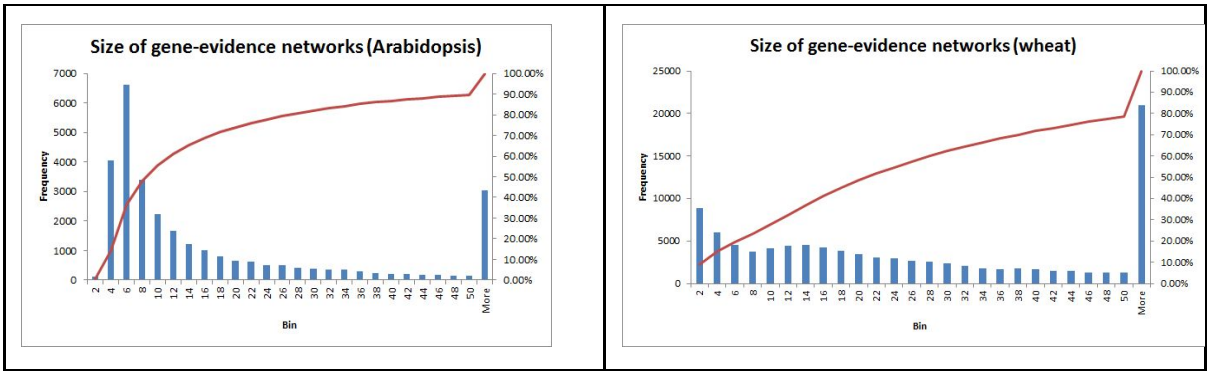

Figure 5.3: Size distribution of gene-evidence networks (#unique concepts) in Arabidopsis and wheat knowledge networks

An evidence concept (e.g. GO concept) can be part of several gene-evidence networks. Figure 5.4 shows the number of genes that are connected to evidence concepts in

Arabidopsis and wheat knowledge networks. Both frequency distributions are characterized through a long tail which contains a large number of occurrences far from the 'head' of the distribution. The big majority of evidence concepts are connected to only one gene; these include Protein concepts that are encoded by Gene concepts. The wheat distribution shows a small peak at 3, this is because of the hexaploid nature of wheat which means that evidence concepts are often related to all three homoeologous wheat genes. About 10% of evidence concepts occur in more than 20 gene-evidence networks, these frequently include GO and Publication concepts.

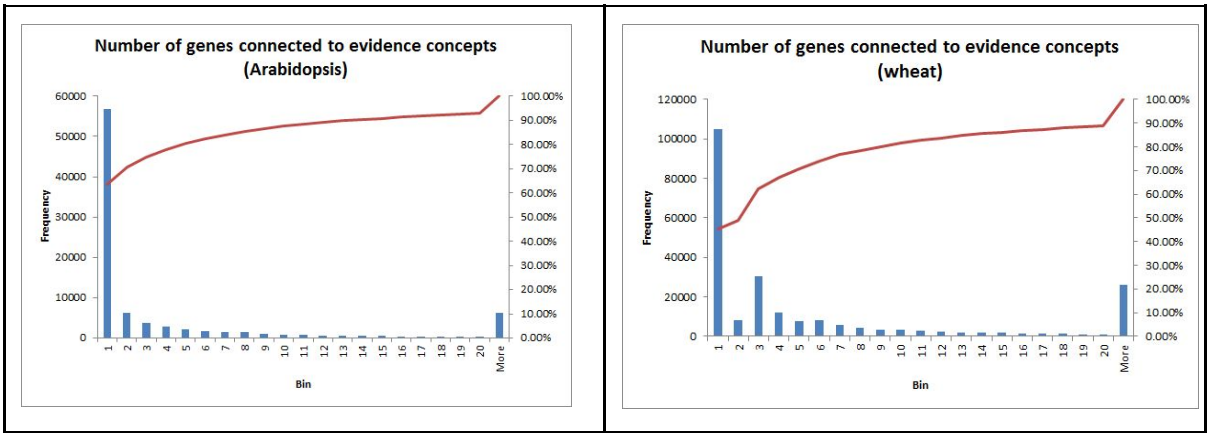

Figure 5.4: Size distribution of genes connected to evidence concepts in Arabidopsis and wheat knowledge networks

5.3.2 Validation of gene scoring method

A gene scoring method (KNETscore) that consists of three components TF*IDF, IGF and EDF was developed. Figure 5.5 shows an example of A) a high scoring gene and B) a low scoring gene for a certain search term. Both gene-evidence networks have about the same number of evidence documents. However, because the network size of A is much smaller it results in a higher EDF score, because the evidence documents are more specific to gene A results in a higher IGF score and because the search terms are more specific to the documents in A it results in a higher TF*IDF score. For these reasons, gene A ranks high while gene B ranks much lower, even though they have the same number of evidence documents.

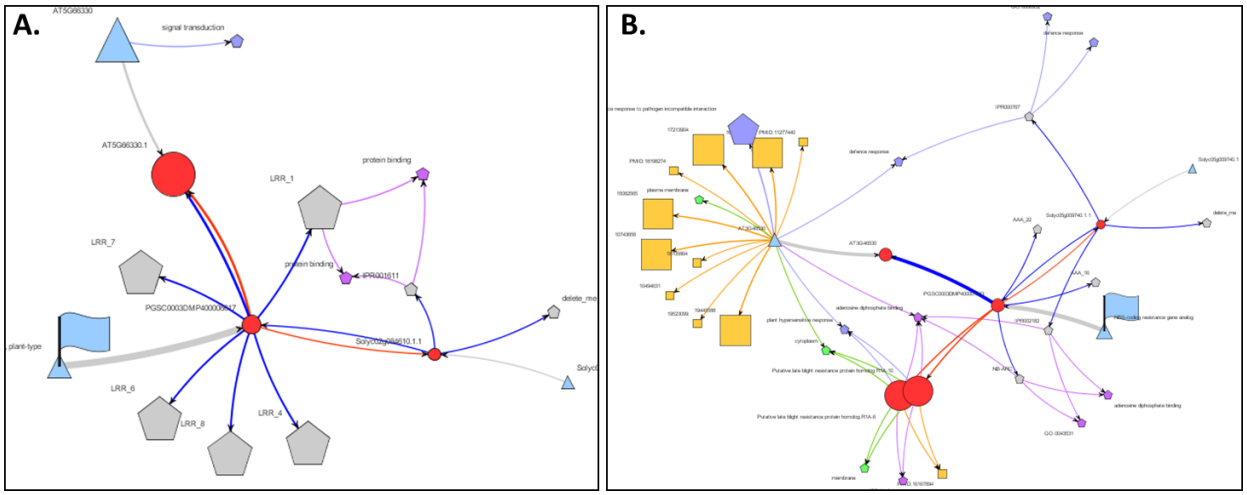

Figure 5.5: A) Example of a high scoring gene and B) example of low scoring gene. Flagged blue triangles represent gene A and B along with their gene-evidence networks. Concepts that contain the search term have a larger size.

To show the usability of KNETscore and compare the differences between MEAN and SUM scores, we focused on a set of 79 wheat genes known to be involved in Gibberellin (GA) synthesis (Hedden and Kamiya 1997). GA is an important plant hormone that regulates growth and influences various developmental processes, including stem elongation, germination, dormancy, flowering and senescence. Searching the wheat knowledge network for the term “gibberellin” identifies 605 documents that contain the term “gibberellin” which are part of 8751 gene-evidence networks. All these gene-evidence networks were scored and ranked using the KNETscore SUM and MEAN methods. Global comparison of MEAN and SUM ranks of all 8751 “gibberellin” related genes shows a correlation of R^2=0.77.

We then studied how many of the known 79 GA genes (reference set) were found and how they ranked using the two approaches respectively. The search identified 72 (91.1%) of the known GA genes. The top 25 genes of the SUM and MEAN ranking (after ranking all 8751 genes) contain 25 and 22 of the GA reference set respectively. The top 100 genes of the SUM and MEAN ranking contain 46 (63.9%) and 48 (66.7%) of the GA reference set respectively. Table 5.5 shows the scores of the top 25 ranked wheat genes for the search term “*gibberellin*” based on KNETscore SUM, and adds the corresponding rank and score of KNETscore MEAN.

Table 5.5: Scores of top 25 ranked genes for search term “gibberellin” based on KNETscore SUM with the corresponding KNETscore MEAN rank and scores.

Gene ID	Sum_Rank	Mean_Rank	Sum_Score	Mean_Score
TRAES_5BL_D412D28CC	1	16	458.17	5.39
TRAES_5DL_3E77D28A6	2	17	452.87	5.33
TRAES_4AL_FABDF4EDA	3	18	452.87	5.33
TRAES_3B_763D7ABA2	4	22	450.27	5.30
TRAES_3B_A2E5CB642	5	19	441.72	5.32
TRAES_1BL_32506F819	6	20	441.72	5.32
TRAES_1AL_3A716350F	7	21	441.72	5.32
TRAES_1AS_B90725283	8	3	390.02	6.29
TRAES_1AS_570581E09	9	10	382.04	5.88
TRAES_5BL_8123B1AD3	10	8	375.24	6.05
TRAES_1DS_A44358D5B	11	11	373.74	5.75
TRAES_1BS_2C29ED3EF	12	12	373.74	5.75
TRAES_3B_7ABEA6AAD	13	9	366.51	5.91
TRAES_3AL_14A36F545	14	7	359.33	6.09
TRAES_2AL_B8AB48108	16	26	342.11	4.44
TRAES_1BL_A1CF1385F	17	13	331.69	5.62
TRAES_2BL_FF2BB4801	18	30	331.58	4.20
TRAES_3B_0CC70372F	19	14	329.16	5.58
TRAES_1AL_C6975BBBD	20	15	322.63	5.56
TRAES_3AS_3A79F81AF	21	27	319.18	4.43
TRAES_2AL_85471F53F	22	28	317.28	4.41
TRAES_2BL_9E115B19F	23	31	305.89	4.08
TRAES_2DL_66F9CEA3C	24	29	302.86	4.21
TRAES_3B_791A6E8DF	25	32	292.87	4.07

Gene TRAES_5BL_D412D28CC (TaGA20ox1B) is the top ranked gene using the SUM score and ranks 16 using the MEAN score. Gene TRAES_2AL_65B19CC73 (TaGA3ox4A) is the top ranked gene using the MEAN score and ranks 79 using the SUM score. The gene-evidence network of TaGA3ox4A has 9 concepts and 5 contain the term “gibberellin”, whereas the gene-evidence network of TaGA20ox1B contains over 120 concepts of which 85 contain the term “gibberellin”. TaGA3ox4A gets ranked top using MEAN scores because it’s evidence concepts are on average very specific and TaGA20ox1B gets ranked top using SUM scores because it has a relatively high number of evidence concepts while they may or may not be very specific.

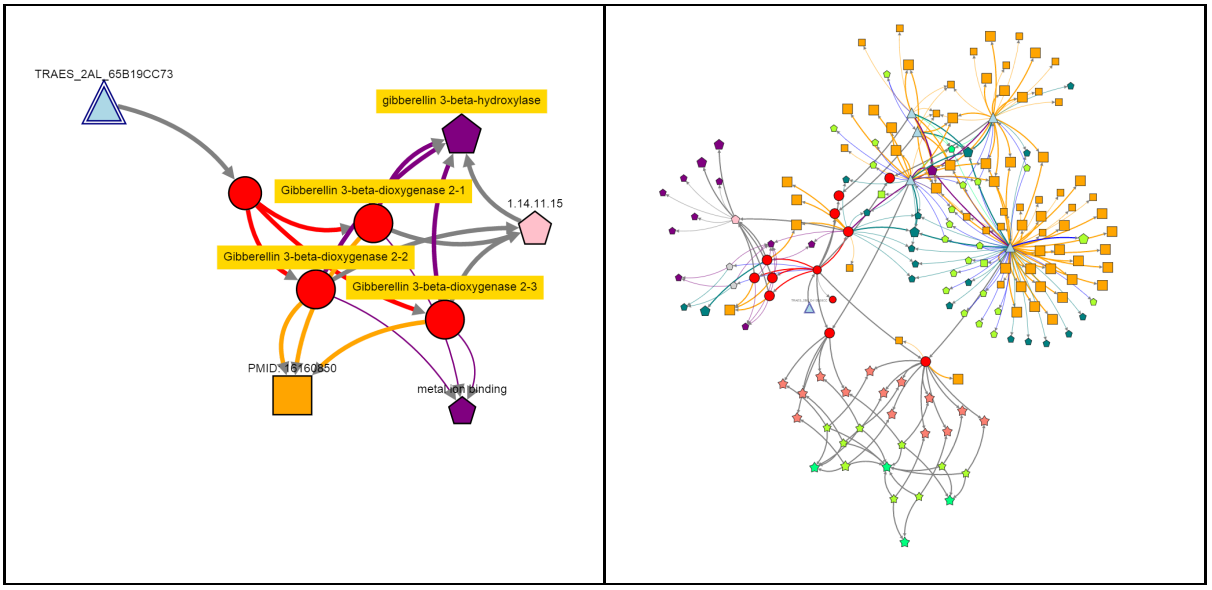

Figure 5.6: On the left, gene TRAES_2AL_65B19CC73 (TaGA3ox4A) that scores highest using KNETscore MEAN. On the right, gene TRAES_5BL_D412D28CC (TaGA20ox1B) that scores highest using KNETscore SUM. Concepts with the search term “gibberellin” are slightly larger.

The lowest ranked gene from the GA evaluation set is TRAES_3B_F6CE72D8D (TaGID2B) with a rank of 1280 using MEAN and 2887 using SUM (Figure 5.7). Interestingly it receives an identical score of 0.2 in both MEAN and SUM. The score is identical because TaGID2B has only one GA related evidence document which leads to SUM and MEAN being the same formula. The evidence document is the GO concept ‘response to gibberellin’ which receives a relatively high TF*IDF score of 3.29 for the search term ‘gibberellin’ (the range of TF*IDF scores among all 605 documents goes from 0.53 to 7.14 with a mean of 2.33). However, since this GO concept is linked to 815 other genes (as expected since GA is a key plant hormone involved in many processes) it will receive a relatively small IGF score (log(99386/815)). In addition it is the only evidence concept among 34 total concepts which leads to a small EDF score (1/34). Although this gene has been in the GA reference set, the evidence in the wheat knowledge network is not specific and sufficient enough to rank it higher.

$$KNETscore('gibberellin', TRAES\_3B\_F6CE72D8D)$$
$$= EDF * IGF * TFIDF$$
$$= 1/34 * log(99386/815) * 3.29$$
$$= 0.2$$

5.4 Discussion

5.4.1 Gene-evidence networks

This chapter introduced the definition of semantic motifs and presented how they are used to find gene-evidence networks. Semantic motifs can be specified and extracted from the knowledge network using the Meta-data based Graph Query Engine (MGQE) that is part of Ondex. The MGQE query syntax requires knowledge about the metagraph to correctly define Concept Classes and Relation Types that constitute biologically plausible paths. The MGQE enables many complex graph queries to be defined in a very concise manner without the need of writing long and complicated SQL or SPARQL like query statements. The examples showed 57 distinct semantic motifs defined within a single query file. The disadvantage of, the concise query syntax of MGQE is that it can become cumbersome to decipher all graph queries that are represented within it. Additional shortcoming of MGQE are that it is not based on a defined deductive logic with associated theoretical basis and has not query planning and optimisation strategies.

Concepts that are included in a gene-evidence network are presumed to be transferable to the gene of interest, in contrast, concepts that are excluded from a gene evidence network (although still part of the GSKN) 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 won't be able to exploit this information.

One current limitation of MGQE is the way it specifies the maximum number of recursive links (hops) a transition can make if the *from* and *to* Concept Classes are the same. The current implementation in Ondex allows the number of hops to be restricted by specifying the total path length, but not by specifying the number of hops itself. This has inconsistency implications when creating gene-evidence networks as hops of type 'interacts with' are denied if they occur in the reference datasets (path length > 8) but are possible if they occur in the wheat specific datasets (path depth < 8).

The semantic motif search considers the directionality of a relation as irrelevant and the relation will be traversed as long as the *from* and *to* Concept Classes, as well as, the Relation Type matches the specification. This approach works well for all cases when the *from* and *to* Concept Classes are different. However, in cases when they are the same, as in the Gene Ontology, it might be required to consider the directionality of the relation. For

example, when a gene is annotated to a GO concept one might want to include all its parent terms (*is_a* relations) but not its child terms into the gene-evidence network. Additionally, in cases when *from* and *to* Concept Classes are the same, the current implementation requires a maximum path length to be set to avoid loops.

In future, more expressive graph query languages such as cypher ("Cypher Query Language" n.d.) or SPARQL (Hancock 2004) are needed to overcome some of these shortcomings. This would enable general statements about biologically plausible hops to be made regardless of when and where in the query path they occur.

Gene-evidence networks are an important requirement for candidate gene discovery. Searching individual gene-evidence networks can, however, identify many potential candidate genes. Methods are required to rank candidate genes and help users to focus on genes with most important evidence information.

5.4.2 Gene scoring method

Searching gene-evidence networks for keywords such as 'gibberellin' or 'flowering' can retrieve hundreds to thousands of genes that are 'somehow' related to the search terms. Methods are required that can sort the list and present the most relevant genes at the top and the less important genes at the bottom. This task of sorting a gene list based on its relevance to a search terms is placed somewhere between information retrieval (document ranking) and candidate gene prioritisation.

The gene scoring method, KNETscore, that was developed as part of this thesis builds on the TF*IDF measure that has been well established in the field of information retrieval for more than 40 years (Sparck Jones 1972). It uses TF*IDF to rank documents by their relevance to a search term, and additionally, considers the properties of gene-evidence networks such as the specificity of documents to a gene (IGF) and the frequency of evidence concepts (EDF). Taking together they provide a measure to differentiate between genes that are highly relevant or less relevant to a search term. Omitting IGF or EDF from the score would have two major consequences 1) genes that have the same evidence documents would always score equally and 2) evidence documents that are linked to hundreds of genes would be given the same weight as documents that are specific to one or two genes.

We have compared two alternative strategies for combining document relevance weights in cases where a gene has two or more evidence concepts 1) taking the sum (SUM) or 2) taking the mean (MEAN) of all TF*IDF * IGF scores. The score ranges between MEAN (0.01-7.19) and SUM (0.01- 450.1) are very different, as expected. The ranking results of the GA reference genes however, show a strong positive correlation (R^2=0.87). Our evaluation has shown that our method ranks 63.9% and 66.7% of 72 GA reference genes within the top 100 genes using SUM and MEAN respectively. Specific examples have shown that the SUM function gives higher scores to well-studied genes. In contrast, the MEAN doesn't take into account the total number of evidence concepts, instead it looks at average specificity over all concepts. MEAN is therefore a more useful measure for identifying novel or newly studied genes that do not yet have large evidence networks or do not have many evidence concepts. MEAN scores are in general easier to interpret since they solely reflect the average specificity whereas SUM scores reflect a combination of both specificity and total number of evidence concepts.

It is an important requirement that the scores can be computed rapidly. Evidence document retrieval and TF*IDF calculation can be achieved in constant time (O(1)) since all documents are indexed for direct retrieval via Lucene. To compute the second component, IGF, it is necessary to know the number of genes that are connected to an evidence document. This can be determined through graph traversal (depth-first search or breadth-first search) starting at the root node (evidence concept) and exploring as far as possible each branch and ensuring it matches a certain semantic motif. The time complexity of DFS or BFS algorithms can be expressed as O(|V|+|E|) where |V| is the number of nodes and |E| the number of edges in the knowledge graph. The graph traversal needs to be performed n-times for all n evidence documents that were retrieved through the search. To compute the last component, EDF, the gene-evidence network needs to be generated using graph traversal. So the total run time complexity for computing the score of a single gene can be expressed as O(2n(|V|+|E|)). It would be very slow to compute this every time a search is performed. However, by developing two additional pre-build indices (HashMaps) that contain the necessary information to compute IGF and EDF, the computation of the score can be achieved in constant time O(1). More details about the implementation are given in the next chapter.

Currently an equal weight is given to all evidence documents. Future work would be to investigate and incorporate specific domain knowledge into the scoring. For example, genes

that have SNPs with causative phenotypes could be ranked higher than genes without variation data. The choice of priority datasets will be application dependent and therefore it would be ideal to make this configurable.

6 DESIGN AND IMPLEMENTATION OF KNETMINER

Having developed genome-scale knowledge networks (GSKNs) and methodology for searching and ranking candidate genes, the next aim was to build user interfaces and data visualisation methods that can give researchers and breeders the means to interrogate knowledge networks themselves. This chapter describes the design and implementation of KnetMiner - a web-based application that was developed specifically for candidate gene discovery.

6.1 Background

In the beginning of the Ondex project (2011), the only way to visualise and explore GSKNs was through the stand-alone Ondex frontend also known as the Ondex Visualisation Toolkit (OVTK). Integrated networks in the form of OXL files were loaded into OVTK and a series of generic graph operations such as filters, annotators and layout algorithms were available to study the networks. Those graph operations made sense to a bioinformatician, however, they were not intuitive enough for use by biologists. The first idea was to develop an Ondex frontend plugin with a task-focused user interface that automates several graph operations and thus facilitates searching of GSKNs. This attempt resulted in a “Genomics” Ondex plugin which consisted of a set of graph filters and annotators, as well as, a novel layout algorithm (Figure 6.1). The plugin allowed users to enter query terms and specify genomic regions of interest. It then filtered Gene concepts based on the genomic positions, extracted a small gene neighbourhood and searched it for the query terms. The result was a subnetwork that was rendered in Ondex using a graph layout algorithm that positioned genes according to their genomic coordinates at the top of the screen and the rest of the network was placed underneath using the GEM graph layout. It was clear, however, that this solution was impractical for two reasons. First, due to large RAM requirements (at least 8Gb RAM) and slow processing time it was tedious to work with it, and second, due to constraints in visualising genomic data with Ondex (designed for network visualisation). Therefore, a faster, more scalable and more user-friendly solution was needed that could overcome existing limitations.

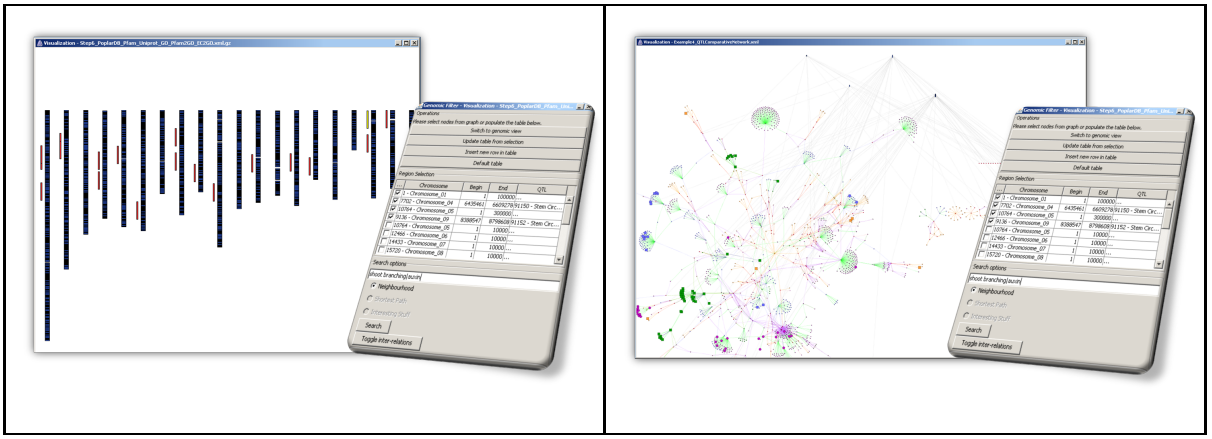

Figure 6.1: The ‘Genomic’ plugin for Ondex to filter the knowledge network and visualise the subnetworks using a loci based layout algorithm. A) Gene (black lines) and QTL (red lines) concepts are placed based on their genomic positions. B) The ‘Genomic’ plugin filters genes based on specified regions, extracts gene neighbourhood networks and visualises genes based on position (top of screen) and the neighbourhood using the GEM layout (bottom of screen).

6.2 Objectives

The aim was to develop a software resource that targets two types of end users. The first type, those with a strong domain expertise and experimental data for a particular trait who want to identify novel targets or generate new hypotheses about biological processes involved in the trait. The second type, those biologists that are beginning to study a new trait of interest and need to efficiently review the existing knowledge without having to manually navigate from database to database as is currently the case.

The **first** requirement identified for both user types is to have a solution that enables GSKNs to be interrogated with user data (search terms, QTL and gene list) without requiring any technical expertise about data integration or retrieval. The **second** requirement is that the solution needs to be scalable and fast even when the knowledge networks are very large. The **third** requirement is that different types of outputs, including genome and network visualisations, should be simply interchanged and coordinated to increase the user’s understanding. The **final** requirement is to provide a modular and portable solution that enables developers to build instances for new species and to deploy them on any IT infrastructure that meets the software requirements.

It was decided that a web-accessible application, consisting of a client that can run in a web browser and a separate backend server, is the ideal design pattern to suit these

requirements. Demanding computation can be performed on server-side and visualisation of the output using flexible web technologies on client-side. The project was given the name KnetMiner and a set of key objectives were defined:

- **Objective 1:** Develop a fast application/web server with methods to query a large knowledge network with user provided data (search terms, QTL and gene list) and to generate bespoke files to be visualised in client applications.

- **Objective 2:** Develop a network visualisation tool that can be embedded in websites and has functions for the exploration of data rich knowledge networks.

- **Objective 3:** Develop a visually appealing web application with a simple submission page for user data (keywords, QTL and gene list) with different visualisations for search results such as tables, networks and genome coordinate based maps (gene, QTL). These must all be easy to navigate between, for example using a tabbed interface, and with extensive cross-referencing.

- **Objective 4:** Enhance the client application through the provision of advanced query support (AND, OR, NOT), real-time user feedback, query term suggestions and visual presentation of evidence information.

- **Objective 5**: Provide a modular solution and facilitate the management of the build and deployment processes.

6.3 KnetMiner System Overview

A KnetMiner web application instance is divided into client and server subsystems (Figure 6.2). The client component, called **KnetMiner-Client**, is deployed in an Apache Tomcat container and holds the application submission and presentation interfaces. The client is mostly based on JavaScript, jQuery and DHTML for data presentation, with optional dependencies to Java Applet and Flash. The client machine sends HTTP requests via Ajax to a Java servlet which passes the requests via a socket connection to the database. The database server, called **KnetMiner-Server**, holds the knowledge network in a memory-based Ondex graph database. The application logic and data processing of the KnetMiner-Server are implemented as a Java multithreaded server. In a single threaded

server, the incoming requests were processed in the same thread that accepted the client connection which meant long-running requests made the server unresponsive for a long period. In contrast, in a multithreaded server connections are handed off to a worker thread that will process the request and enable the server to accept new requests in the meantime, making it more responsive. The KnetMiner-Server produces query-dependent views (OXL, JSON, TAB files) of the knowledge networks which are passed on to the JavaScript methods that requested them in order to be presented in the web browser.

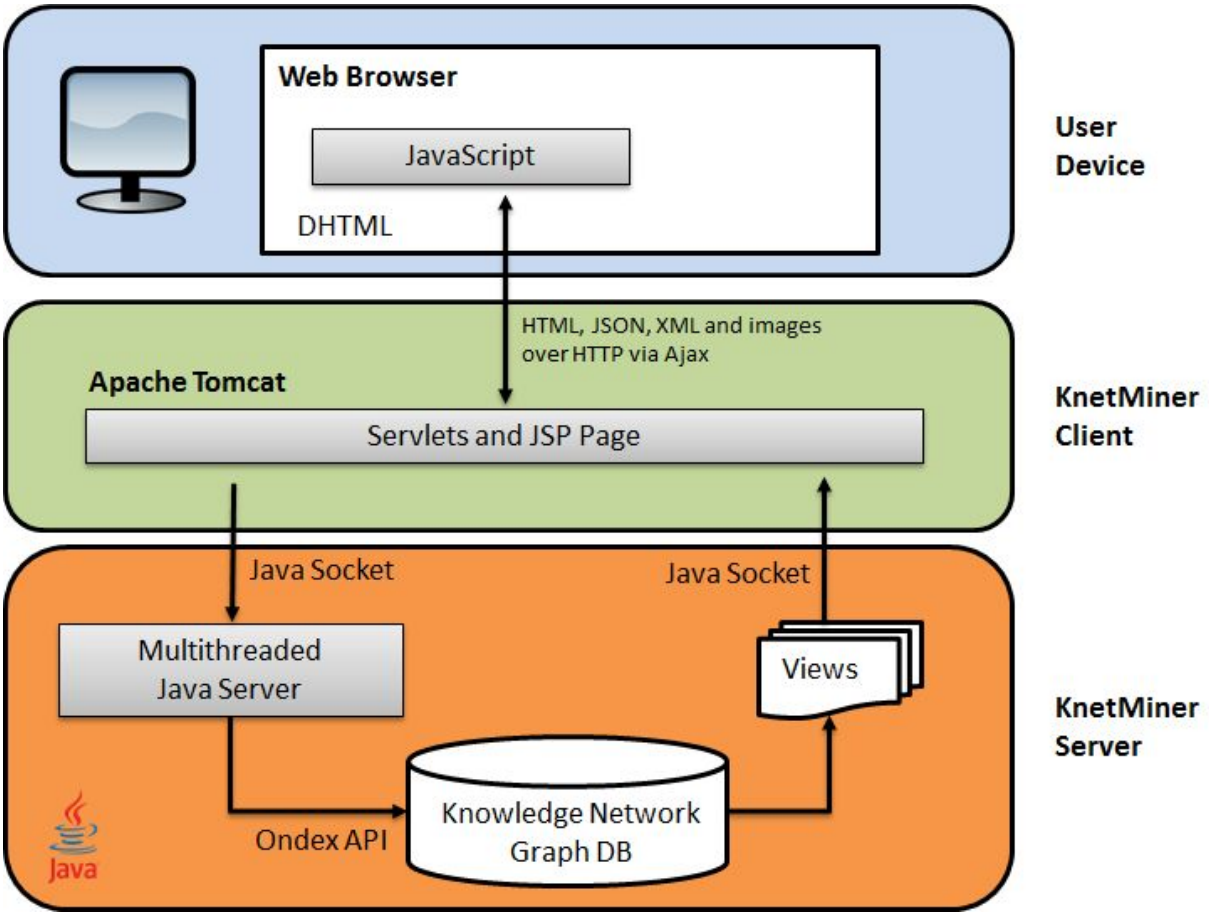

Figure 6.2: The architecture of a single KnetMiner web server based on a client-server design model.

6.4 The KnetMiner-Server

6.4.1 Pre-processing the knowledge network

The KnetMiner-Server is a Java Archive (JAR) that requires as an argument the path to the GSKN (OXL file) that has already been built using Ondex. Initially, when the JAR is run, it pre-processes the GSKN in order to allow fast responses to client requests and to generate basic statistics of the database content. The steps involved in the pre-processing include:

1. Parsing GSKN OXL file into a memory-based graph database
2. Indexing the GSKN using Lucene. The index is stored in a new sub folder
3. Traversal of the entire GSKN using the MGQE (see Chapter 5) for all *Gene* concepts and semantic motifs (packaged with the server JAR file)
4. Building two indices with the results of the network traversal: **gene-evidence map** and **evidence-gene map**.
5. Retrieve total number of genes, concept, relations in the GSKN, as well as, the minimum, maximum and average size of gene-evidence networks.

The pre-processing of the GSKN requires about 10GB RAM on a server with 4 cores and takes about 10 minutes for a network with 0.5 Million concepts, 1.5 Million relations, 100,000 seed genes and 57 semantic motifs.

6.4.2 Incoming request types

Once the pre-processing has completed, the KnetMiner-Server is ready to receive messages from client applications. Methods have been implemented to support different types of requests (Table 6.1) that come via socket connections. Every client request is parsed to identify the request type which then triggers the corresponding subroutines as described in more detail below.

Table 6.1: Methods (services) provided by the KnetMiner-Server.

Request type	Input	Output	Description
keyword	Keywords QTL Gene list	Gene file (TAB) Evidence file (TAB) Genomic file (XML)	Search the knowledge network. Rank all genes and evidences. Produce output files.
gene_net	Keyword Gene list	Ondex-OXL CytoscapeJS-JSON	Extract gene evidence networks. Annotate the network to highlight important information. Produce network output files.
evidence_net	Ondex Concept ID	Ondex-OXL CytoscapeJS-JSON	Extract paths in gene-evidence networks that end with given concept. Produce network output files.
counthits	Keywords	Number	Number of genes and evidence concepts 'matching' a query
countloci	QTL	Number	Number of genes within a QTL region
synonyms	Keywords	File (TAB)	Searches the knowledge network for query terms. Identifies synonymous concept names.

A request of type **keyword** takes keywords, QTL and a gene list and produces a ranked list of candidate genes and evidence information. The steps involved in this process include:

1. Parse query terms,QTL and gene ids (if provided).
2. Search the Lucene index to identify and rank matching concepts.
3. Perform a lookup in the evidence-gene map to retrieve genes.
4. Score genes based on their importance to the search terms (see Chapter 5).
5. Determine overlaps with user provided QTL and gene list.
6. Export results as TAB and GViewer-XML files to the web data folder.
7. Return the file path.

A request of type **gene_net** takes a list of gene ids and search terms, and returns a network as output. The steps involved in preparing the network include:

1. Extract gene-evidence networks for given input genes from the knowledge network.
2. Annotate the gene-evidence networks with hide or show attributes.
3. Export the sub network in Ondex-OXL and CytoscapeJS-JSON format.
4. Return the file path.

The request of type **evidence_net** takes as input an Ondex concept and returns a network of genes connected to the given Ondex concept via valid semantic motifs. The steps involved include following subroutines:

1. Identify gene-evidence networks that contain the given concept. This is efficiently done through a lookup in the evidence-gene map to retrieve all gene ids.
2. Extract gene-evidence networks for given gene ids.
3. Filter gene-evidence networks by retaining paths that end with the given concept id.
4. Export the sub network in Ondex-OXL and CytoscapeJS-JSON format.
5. Return the file path.

The request type **synonyms** takes as input a string of search terms and produces a list of synonyms for each term. The steps involved in this process include:

1. Divide search terms into major tokens (split AND, OR, NOT)
2. Search the knowledge network for every token using Lucene
3. Organise the hits into Concept Class categories
4. Extract top scoring concept names per Concept Class category for each search term
5. Export the results in TAB format
6. Return the file path

The request of type **count_hits** takes as input a string and returns the number of total concepts, evidence concepts and genes matching the search term. The steps involved in producing counts efficiently include:

1. Search the Lucene index to identify evidence concepts.
2. Perform a lookup in the evidence-gene map.
3. Determine the number of distinct genes the evidence concepts are linked to.
4. Return all three numbers from previous three steps.

The request of type **count_loci** takes as input a QTL region and returns the number of genes within the given region. The steps involved in producing the counts include:

1. Retrieve all *Gene* concepts from knowledge network (filter by TAXID)
2. Get chromosome, start and end values (in base pairs) for each gene. If this is not available get chromosome and centimorgan (cM) values.
3. Count how many genes are within the given QTL boundaries.

6.5 KnetMiner Client Subsystem

The KnetMiner-Client is deployed as a WAR file (Web application ARchive) in a Tomcat container and is what the user sees in the web browser. The client is implemented based on DHTML, CSS, JavaScript, jQuery and Java Servlets. Here we describe the two major client components, the user query interface and the data visualisation interfaces, of KnetMiner.

6.5.1 User query interface

The query interface was designed to provide a simple submission page for user data (keywords, QTL and gene list) and to support the refinement of search queries. The user interface is divided into four sections A, B, C and D (Figure 6.3) and the question marks next to each section provide documentation or example queries. The search terms provided in A) are the only required user input, while the other fields in the form are optional (QTL and gene list). Pressing the *Search* button uses all information provided in A, C and D to submit a server request.

This sections provides an overview of the different components that are available in the user query interface.

Figure 6.3: The KnetMiner user query interface. A) Search term input, B) Query suggester, C) QTL input and D) Gene list input.

6.5.1.1 A Google-like search interface

The main search field of KnetMiner allows users to input any search terms as lists of keywords, for example related to a trait of interest. The search provides full support for the Lucene query syntax so that different terms can be combined with the logical operators OR, AND, NOT to create more complex query statements. The terms can be high level

descriptions of a phenotypic trait (e.g. disease resistance) but also more specific terms such as biological processes, protein families or gene names (e.g. defense response to fungi, LRR or SNC1). KnetMiner sends a ***keyword*** request to the server when the user clicks the *Search* button. The server processes the request and returns several output files that are visualised in different tabs of the results page (more detailed explanation will follow in the visualisation section).

Additionally a feedback mechanism was implemented that constantly (in real-time) returns the number of resulting documents and genes while the user is typing the query (Figure 6.4). This feature is activated once the query term is at least 3 characters long and is updated at each additional keyboard event. It uses the ***count_hits*** function to send data to and retrieve from the server asynchronously (in the background) without interfering with the display of the existing page. 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 examine their query and explore different spelling, language or more complex query statements (AND, OR, NOT).

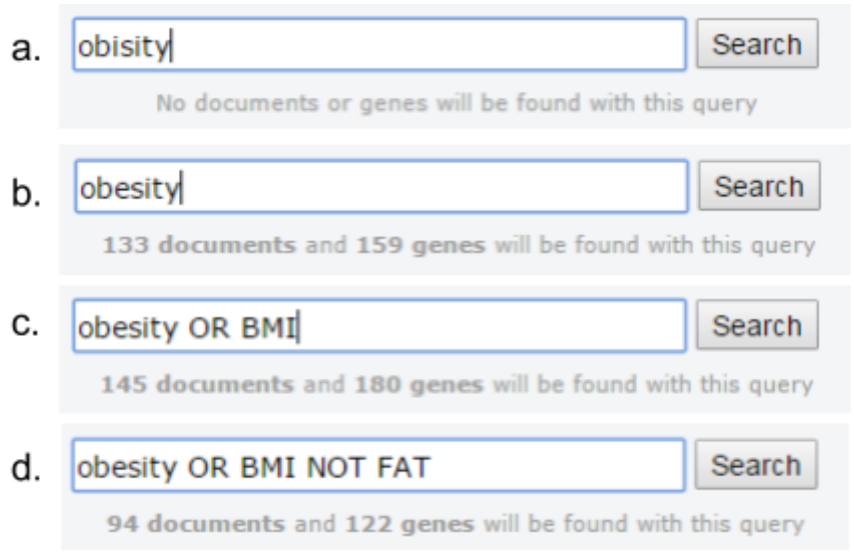

Figure 6.4: The search interface of the cow KnetMiner and the user feedback mechanism. a) The query 'obisity' contains a spelling mistake which prompts a feedback that no results can be found. b) The user corrects the query to 'obesity' and is given as feedback that 133 documents and 159 genes can be found. c) The user extends the search to 'obesity or BMI' finding 12 more documents and 21 more genes. d) The query is made more specific by excluding any evidence document that contains the word 'FAT' which results in 94 documents and 122 genes.

6.5.1.2 Query suggestions

There are many ways by which a single trait can be referred to in the literature. The first difficulty users are faced with when using an information retrieval system is therefore to know which terms to include in a query. A common strategy for users is to start with a simple query and gradually refine it. KnetMiner contains a query suggestion wizard that helps users to refine their query by suggesting more specific terms or alternative synonyms (Figure 6.5). The suggested terms are derived from the underlying knowledge network. 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.

By opening the query suggestion wizard, the query string is sent to the KnetMiner server and a request of type **synonyms** is made. A server function first tokenizes the entire string into its main components (splitting by AND, OR, NOT). The knowledge network is then searched and concepts containing the tokenized terms in their concept names are identified. All synonymous terms are retrieved and ranked by the Lucene score. Per *Concept Class*, the top 25 concept names for every term are returned in a text output file to the client. A client-side JavaScript function renders the data in a table-like frame, grouping the information by query tokens and *Concept Classes* (e.g. *Gene, Pathway* and *Biological Process*). The *Concept Classes* are represented with the same symbols that are consistently used throughout the whole KnetMiner application. This visual aspect aims preparing users to the meaning of the different evidence types that are present in the knowledge networks, so that they are well versed before they start exploring the gene-evidence networks.

Figure 6.5: The output of the 'Query Suggester' for the search terms 'drought OR heat'. The tabs at the top contain the different query tokens (i.e. heat and drought). The tabs on the left contain suggested synonyms per *Concept Classes*, i.e. *Trait Ontology, Gene Ontology (BP), Protein* and *Gene* (from top to down). The drought and Gene Ontology (BP) tabs are selected in the screenshot.

6.5.1.3 Adding QTL data to the search

KnetMiner provides an optional field to input one or many QTL regions. This feature is only available if the species for which the KnetMiner instance has been built has a sequenced genome and genes have a physical location defined by a coordinate system (base pair or centimorgan). Entering the chromosome, start and end position of a QTL will automatically display the number of genes that are within the QTL boundaries (**count_loci** request). An option is provided to restrict the search to the provided QTL genes. Genes that rank low when searching the whole genome, might rank high when the search space is reduced to the QTL genes. The QTL information that the user provided will be visualised in the Gene and Map views as part of the search output.

Note that genetic marker names as an input for start and end position of a QTL are not yet supported. Users are required to independently identify base pair positions for their QTL intervals by aligning known genetic markers to a genome sequence assembly with physical distance measured in base pairs (bp).

6.5.1.4 Adding gene lists to the search

Users may wish to include candidate genes from expression studies or other 'omics studies in the search. KnetMiner provides an input form for entering gene names or accessions (one per line). The user gene list will be visualised in the Gene, Evidence and Map views as part of the search output.

The gene list (optional) is incorporated into the **keyword** request that is sent to the server. The server first tests if the given gene names or ids match any *Gene* concept in the knowledge network and, second, flags all genes in the main search output that were part of the user gene list.

An option is provided that adds user genes to the results output regardless of whether they were related to the search terms (*Map gene list without restrictions*). The "unrelated" user genes can not be assigned a score and evidence information, and will therefore appear at the bottom of the Gene View results table. This feature was added to allow users to explore gene-evidence networks of any genes of interest.

6.5.2 Visualisation of search results

Different views for exploring the search output were developed; each has a different aim and helps address different questions. The main design principle was to divide the visualisation into two steps in contrast to the original *Genomics* plugin (see Background) which immediately exposed users to networks. First, it was decided 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 or to 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 provide the full evidence path that resulted in the prediction of the candidate genes. In a second step, to enable the evidence path to be investigated in full detail, a network visualisation component allows users to study the gene-evidence networks of selected genes. 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.

This section describes the four different views **Map View**, **Gene View**, **Evidence View** and **Network View** in which the search results are presented.

6.5.2.1 Map view

The Map view visualises candidate genes and QTL according to their coordinates on a chromosome-based map (Figure 6.6). It is intended to provide an overview of the genomic locations of a specific set of features associated with a specific search rather than as a way to view all features on the genome. The Map view displays the top 100 scoring genes as triangles and uses color coding to distinguish genes with high (green), medium (orange) and low (red) scores. User defined QTL and QTL retrieved from the knowledge network (because it is matching the query terms) are displayed as rectangles with a variable length that corresponds to the QTL interval. This view not only illustrates effectively the overlap of genes and QTL but also the relative position of candidate genes with respect to the QTL. Chromosomes themselves can also be colour-coded to create banding pattern effects that illustrate for example genome rearrangements or differences between hetero- and euchromatic regions as are commonly used in traditional cytogenetic maps.

The current implementation of the KnetMiner Map view makes use of GViewer - a customizable Flash movie that is part of the free and open-source GMOD tools (www.gmod.org). This means that a Flash plugin needs to be installed to be able to see the map. GViewer has two input files: a base map which defines the number and sizes of the chromosomes, and an annotation file which defines the position of features such as genes and QTL. The base maps in the different KnetMiner instances are static and are created manually as part of the configuration of the client application. The annotation file, however, is generated by the KnetMiner server (see *keyword* request) and dynamically loaded when generating the Map view.

The Flash dependency and the fact that GViewer is not under active development (while still having several major bugs) means that it is not an ideal solution for modern web applications any longer.

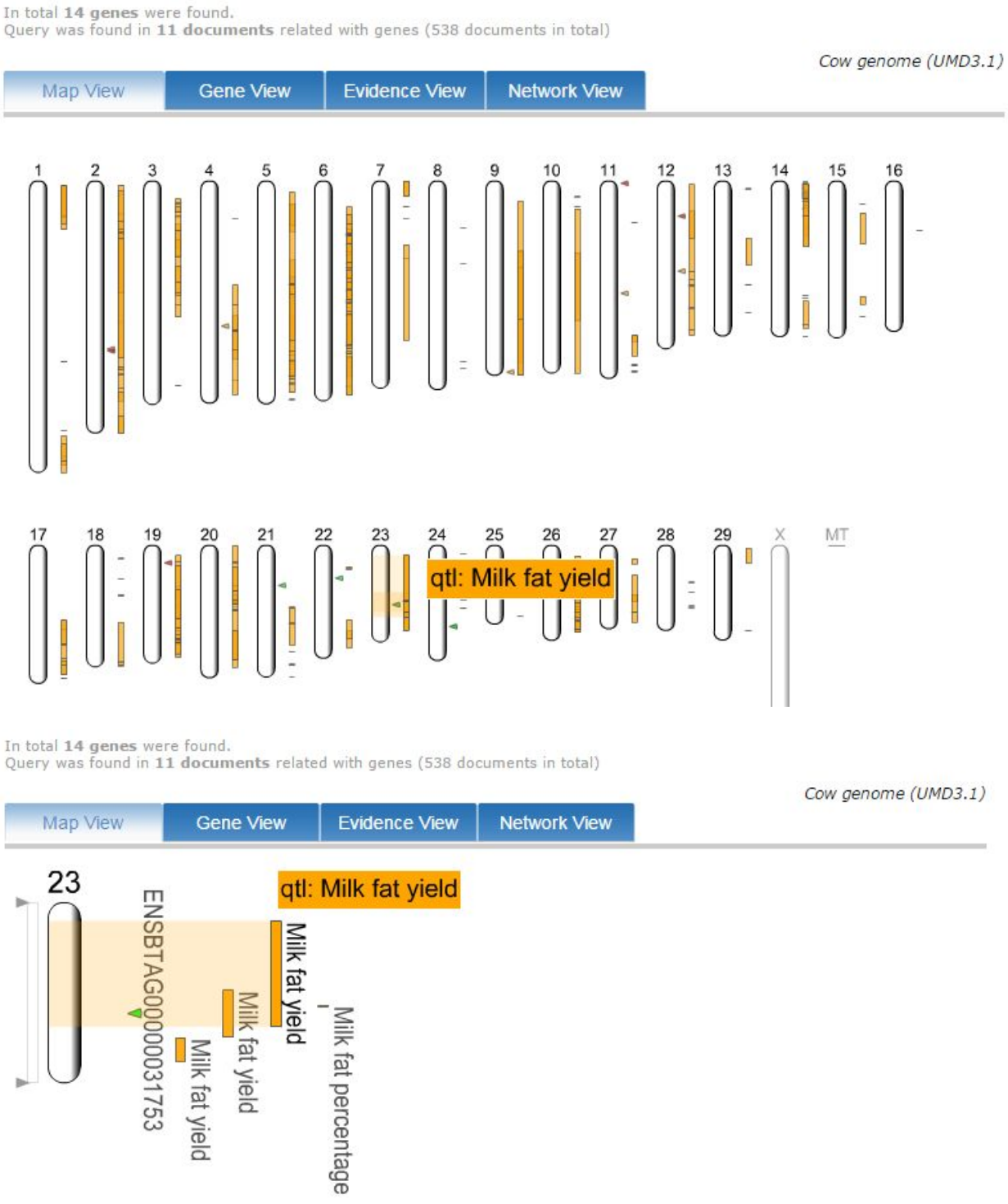

Figure 6.6: The Map View. Shows genes (triangles) and QTL (rectangles) from the knowledge network that are related to the search terms. Enables zooming into selected chromosomes (below).

6.5.2.2 Gene view

The Gene view displays the identified candidate genes in a table. The genes are sorted by their KNETscore, the most relevant ones at the top and the less important ones at the bottom (Figure 6.7). The top 100 genes are shown by default, but an option is provided to display up to 1000 genes and the full table can also be downloaded. The first column displays the gene accession or if available, the gene name. The next two columns provide information about the location of the gene. The fourth column shows the computed relevance score (see Methods) which is used to sort the table. The fifth column indicates the number of distinct QTL that overlap with this gene (including QTL from the knowledge network and from the user input). The sixth column indicates whether the gene is part of the user provided gene list (yes/no). Finally, the last column summarizes the supporting evidence concepts that contain the search terms.

The evidence concepts are grouped according to their evidence types (*Concept Classes*) and these are illustrated utilising the same graphical symbols as presented in the knowledge networks or metagraphs. An integer in the centre of the *Concept Class* symbol counts the instances of this class (number of concepts). For example, an orange rectangle with the number 18 and a green pentagon with the number 2 mean that the gene-evidence network of this gene has 18 concepts of type Publication and 2 concepts of type Trait Ontology which contain the search term. The evidence images are clickable and extend to provide one representative description string for each evidence concept. If the evidence is a publication then the PubMed id is shown and linked via URL to PubMed.

Two checkboxes at the top of the table ‘Known targets’ and ‘Novel targets’ make it easier to select multiple genes when user genes were provided during the search. ‘Known targets’ selects all user genes that have some evidence concepts, while, ‘Novel targets’ selects all user genes that have no evidence concepts. A slightly different network visualisation approach has been developed for novel genes which initially shows routes to GO and TO concepts and hides the rest of the network. More details are available in the first use case of Chapter 7.

In summary, the Gene view table is built by parsing the server-side generated output file and rendering it using client-side JavaScript functions. It provides sortable columns, appealing graphical images and is rich in detail. All together it enables domain experts to make

effective selections of suitable candidate genes that can be explored further in the Network View; with the ultimate goal to identify strong candidate genes for experimental validation. The gene-evidence networks can be generated by either clicking on a gene name or selecting a set of genes and clicking the *View Network* button underneath of the table which will send a request of type **gene_net** to the server.

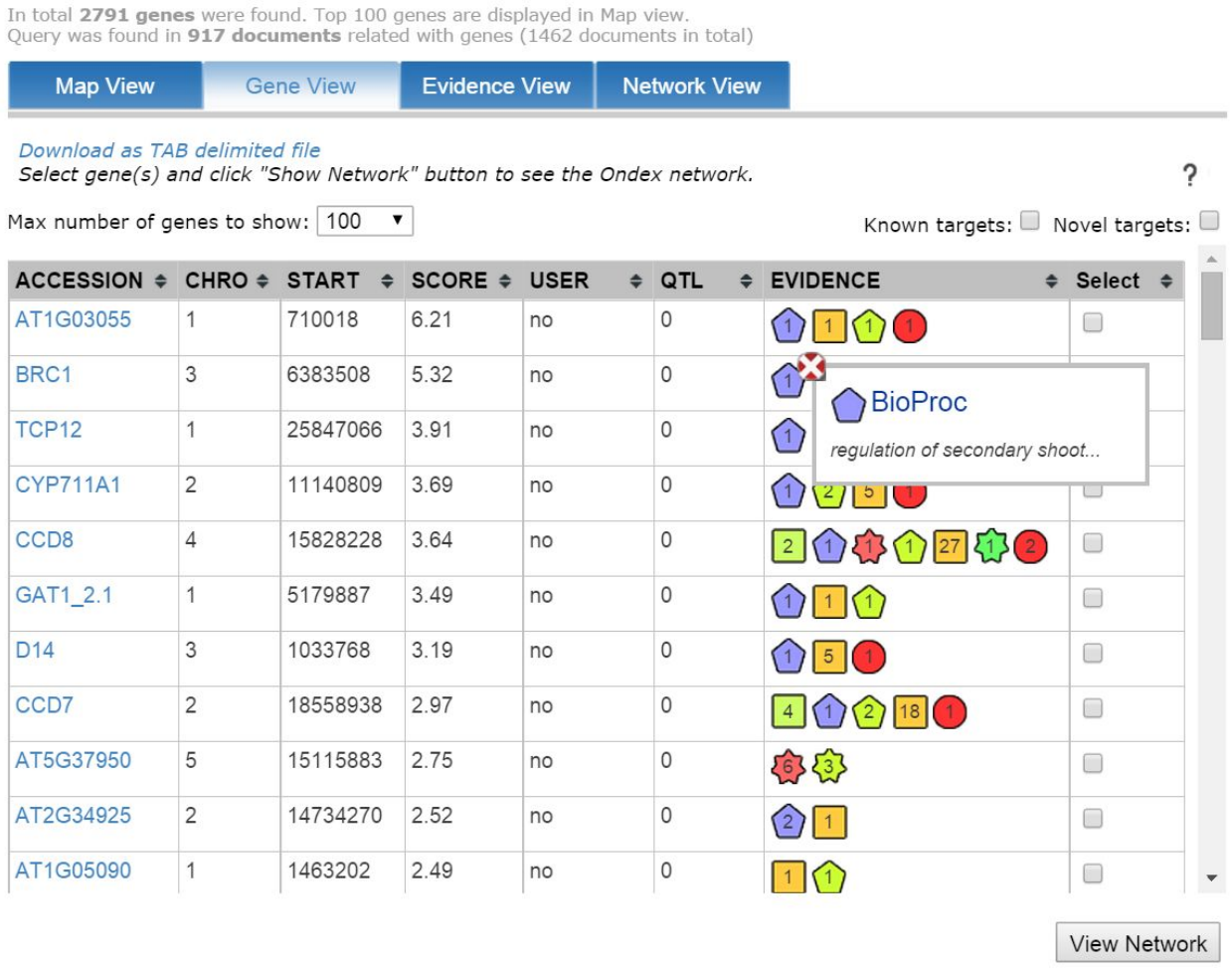

Figure 6.7: The Gene view displays genes and evidence concepts from the knowledge network that are related to the search terms. Networks can be shown for single or multiple genes.

6.5.2.3 Evidence view

The Evidence view provides a document-centric table of the search results sorted by the Lucene score (Figure 6.8). A legend at the top of the table illustrates the total number of evidence documents found per evidence type. The table shows all concepts from the knowledge network containing the query terms. An action button allows users to exclude

specific documents from the next search by adding a 'NOT Concept ID' to the user query statement. For every concept, the total number of genes and the total number of user-provided genes is displayed that are directly or indirectly connected to this evidence concept in the network. This is a very useful view to quickly get to genes that are, for example, involved in a specific pathway or to identify concepts that are enriched in the user-provided gene list. Clicking on the number of genes in column five will send a request of type **evidence_net** to the server and generate a network which shows the concept and how the genes are linked to it.

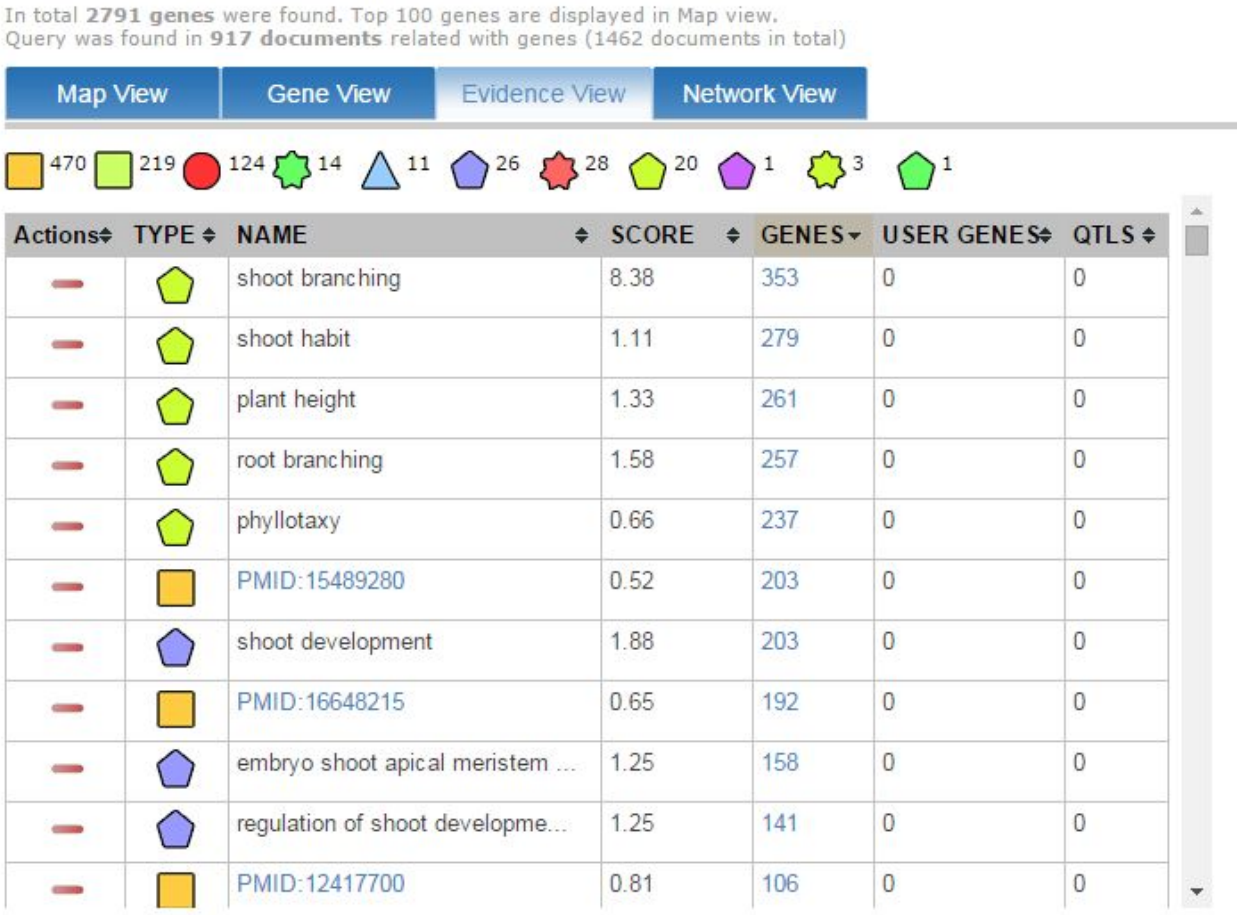

Figure 6.8: The Evidence view presents the evidence concepts from the knowledge network that contain the search terms. The column GENES contains the number of genes directly or indirectly linked to that concept. Clicking on the number will visualise a network.

6.5.2.4 Network view

The initial version of KnetMiner used Ondex Web (Taubert et al. 2014) as its main network viewer. Ondex Web is a modified version of the Ondex Visualisation Toolkit (OVTK) and was

especially developed to meet the requirements of KnetMiner. Features such as context-sensitive menus and annotation tools provide users with intuitive ways to explore and manipulate the appearance of heterogeneous biological networks. Ondex Web is open source, written in Java and can be embedded in websites as an applet.

Since Ondex Web was released, the security requirements imposed by modern web browsers and Java (starting with Java 7 Update 51) mean that it is not possible to run Java applications that are not signed by a trusted authority or that are missing permission attributes. A signed certificate for Ondex Web has been acquired and the permission information is available for use on PCs with Java and Java enabled web browsers. However, Ondex Web does not work on Apple iOS devices and will never be compatible with touch devices such as tablets or smartphones. Furthermore, the large size of the applet (80 Mb) means very slow loading time when it is started the first time (subsequent calls are faster because it gets cached). All these features of Ondex Web cause problems for both users and developers and a replacement solution was sought.

The development of new web software technologies and the rise of mobile touch devices motivated an investigation into alternatives that would be free, easy to use, well supported, not require dependencies such as Java, run on any OS and any web browser, and would be touch-enabled. Several JavaScript based libraries that would allow rendering and visualization of networks were evaluated including Vis.js, Arbor.js, Sigma.js, D3.js and CytoscapeJS. CytoscapeJS was ultimately chosen because of its powerful graph API, appealing graph visualisation, its large community of developers and its popularity within the bioinformatics community. A new network viewer, called **KnetMaps**, especially optimised for the visualisation and exploration of heterogeneous knowledge networks was designed by me and implemented by Ajit Singh.

KnetMaps displays the gene-evidence networks that are requested by the previous views (Figure 6.9). It uses CytoscapeJS, a fully featured open-source graph library written in JavaScript, to render a JSON file that is produced by the KnetMiner-Server. KnetMaps has touchscreen compatibility and can be used on tablets, touch PC's and smartphones running MS Windows, Apple iOS and Google Android operating systems. Touch gestures such as tap, hold and drag have been incorporated and these significantly enhance the user experience when directly interacting with the network visualisation.

Gene-evidence networks are labelled and directed multi-graphs. This means that concepts and relations can have different semantics (Gene, Protein, Pathway, Phenotype, etc.). KnetMaps can visualise heterogenous networks and can incorporate additional search specific **visualisation effects** on top of it. Concepts (nodes) are displayed using different symbols and colours (detailed in the Legend below the network). Relations (edges) use various colours depending on the relation type. All the genes are displayed as blue triangles but the gene(s) originally selected for viewing have a double border to visually distinguish them from other genes. The KnetMiner-Server annotates the nodes and edges in a gene-evidence network regarding 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 concepts that have hidden concepts connected to them. This effect enables users to focus on the most important information and to expand the network if additional information is required. Additionally, the node symbol size is increased to visually highlight nodes that have attributes which contain the user's search terms.

Right-clicking a concept or relation opens a circular **context menu** with features like Item Info (to display specific information about the selected concept or relation in a sliding overlay panel), Show Links (to show hidden elements in its neighbourhood), Hide (to hide the selected concept or relation), Hide by Type (to hide all the concepts or relations of a particular type, i.e., the same type as the selected concept or relation), Label on/off (to toggle the visibility of the Label on/off for the selected concept or relation) and Label on/off by Type (to toggle the visibility of Labels on/off for all concepts or relations of a particular type).

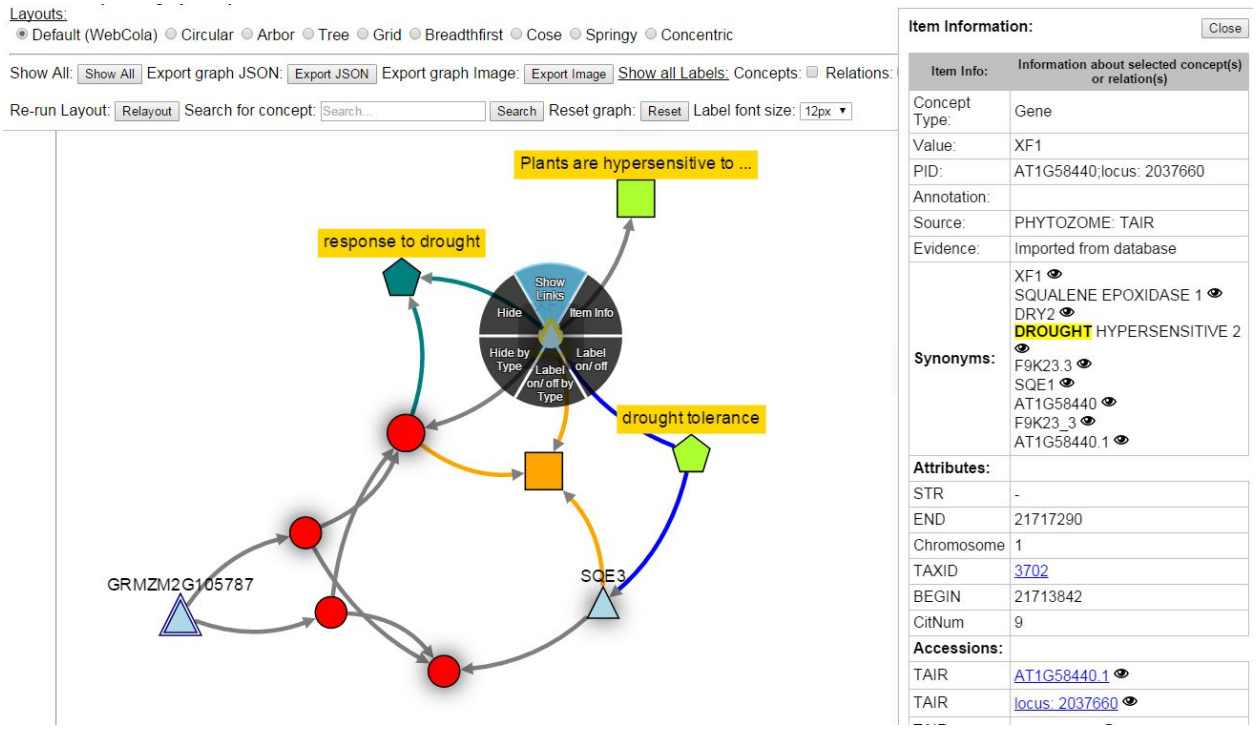

Figure 6.9: The KnetMaps can display gene-evidence networks extracted from the knowledge network. A blurred effect on concepts indicates hidden relations in its neighborhood that can be expanded via the context menu. The Item Information panel on the right shows the rich information content of the concepts and relations. The top configuration panel allows the selection of the layout algorithm and other global networks settings to be made.

Gene-evidence networks are content-rich which means that concepts and relations not only have a type but also have various attributes such as synonyms, accession numbers, cross-references and other data. The **Item Info** panel on the right is used to display the content of the selected concept or relation. It automatically slides open if users right-click a concept or relation and select “Item Info” option. Once open, it also automatically updates the displayed content if users select another concept or relation. The panel displays information such as concept/relation type, labels, Annotations, Attributes (such as publication abstracts, title, authors, amino-acid sequence, TAX ID, etc.) and Accessions (with links which cross-reference to TAIR, Ensembl, UniProtKB, PubMed, KEGG, IPRO, PFAM, etc., where relevant). A user’s search query terms, where found, are highlighted in the Item Info and in the network. The eye-shaped icons next to concept synonyms and accessions in the Item Info enable users to update the concept label (i.e., the preferred concept name) in the network with this new value. This is a useful feature when preparing publication ready images requiring customization.

A **configuration** panel on the top allows users to change the graph layout using a variety of graph layout algorithms, supported by CytoscapeJS, which have been incorporated in the KnetMaps such as WebCola (default), circular, Arbor, Grid, Cose and Concentric. It also provides options to show/hide labels or to change the label font size on all concepts and relations. Finally, the network can also be exported in JSON and OXL formats or as a PNG image.

6.6 Development of new KnetMiner Instances

Every species or group of closely related species has its own knowledge network and is deployed as a separate instance of KnetMiner. This model, in contrast to deploying one large application, provides better flexibility (instances can be deployed independently) and robustness (crash of one instance does not affect other instances of KnetMiner). The disadvantage is that it is challenging to make some classes of change across the set of KnetMiner instances.

The steps involved in developing a new instance of a KnetMiner server-client application include cloning the source code, customization of several configuration files, building new client and server packages and deploying them on a server. This section gives a detailed description of these steps. It requires Java, Maven, Git, Tomcat and a genome-scale knowledge network to be available.

6.6.1 KnetMiner project model

The KnetMiner source code is freely available from GitHub under the GNU LGPL license agreement.

```
git clone https://github.com/KeywanHP/KnetMiner
```

The KnetMiner project contains a *common* project and several independent modules (one for each KnetMiner species). The *common* project contains most of the source code for both the client and the server. The modules consist mostly of configuration files that enable each KnetMiner instance to be customised.

Maven is used to manage each project's build based on the concept of a project object model (POM). The parent project (*KnetMiner/pom.xml*) contains all general specifications including dependencies and inherits these to every module. All modules have the *common* project as a dependency. Building the parent project (*mvn package*) automatically builds the common project and packages all modules.

```
<modules>
        <module>arabidopsis</module>
        <module>common</module>
        <module>poplar</module>
        <module>pig</module>
        <module>tomato</module>
        <module>potato</module>
        <module>wheat</module>
        <module>rice</module>
        <module>barley</module>
        <module>chicken</module>
        <module>cow</module>
        <module>boleracea</module>
        <module>maize</module>
</modules>
```

6.6.2 Configuration of KnetMiner client and server

The simplest way of configuring a new instance (module) of KnetMiner is by cloning the git repository and using an existing module as a template. The client has considerably more options for customization as shown in Table 6.2. The server provides several configuration files as shown in Table 6.3 but only few of them need to be modified.

Table 6.2: Client configuration files and parameters

File	Description
`client/src/main/resources/`**`config.xml`**	Specifies the server host and port to which a socket connection will be established ('localhost' can be used if client and server run on the same machine).

	`<entry key="ServerHost">localhost</entry>` `<entry key="ServerPort">8189</entry>`

`client/src/main/webapp/html/javascript/`**`utils-config.js`**	Specifies global variables that are used in the JavaScript code.
	
	URL of the web folder in which the output files are written by the server
	`var data_url = "http://localhost:8080/species_data";`
	
	URL of the Ondex Web applet
	`var applet_url = "http://ondex.rothamsted.ac.uk/OndexWebBeta";`
	
	Disables the QTL Search and the Map view in set to false
	`var reference_genome = true;`
	
	Shows the TAXID in the *Gene View* if set to true
	`var multiorganism = false;`
	
	Adjusts scaling in GViewer if first chromosome is not the longest
	`var longest_chr = 196087864;`

`client/src/main/webapp/html/`**`index.jsp`**	Provides parameters to customize the website: *title*: HTML title of the website *image*: Path to the top-right header image (logo) *chromosomes*: Comma separated values; used in drop down menu of the *QTL Search* interface. *assembly*: String that is displayed in the *Map view*.
`client/src/main/webapp/html/`**`release.html`**	Can contain custom HTML that will be shown when users click on *Release Notes*.
`client/src/main/webapp/html/image/`**`organism.png`**	An image that will be shown at the top right of the website (the file name needs to remain *organism.png*).
`client/src/main/webapp/html/data/`**`basemap.xml`**	Specifies the number, length and appearance of the chromosomes as displayed in the Map view. The specification can be found here: http://gmod.org/wiki/Flash_GViewer_Documentation#The_BaseMap

client/src/main/webapp/html/data/**sampleQuery.xml**	Specifies example queries that will be provided under the help (?) section of the user query interface and automatically prefill the query form. The format is:
	
	`<sampleQueries>`
	`  <query>`
	`  <name></name>`
	`  <description></description>`
	`  <term></term>`
	`  <withinRegion></withinRegion>`
	`  <region>`
	`  <chromosome></chromosome>`
	`  <start></start>`
	`  <end></end>`
	`  <label></label>`
	`  </region>`
	`  <gene></gene>`
	`  <gene></gene>`
	`  </query>`
	`<sampleQueries>`

client/**pom.xml**	Specifies the maven build and dependency options. Contains the name of the WAR output file.

Table 6.3: Server configuration files and parameters

File	Description
server/src/main/resources/**config.xml**	Specifies the main configuration options of the server application such as: • `<entry key="DataPath">` Path to a web folder which stores the temporary output files (i.e. /var/www/species_data). Needs to match the client's 'var data_url" • `<entry key="SpeciesTaxId">` Specifies the species taxonomy ID and needs to be identical to the species TAXID in the knowledge network. • `<entry key="ServerPort">` The server port of the Socket connection. It has to match client's ServerPort. Different ports need to be used when

	multiple application servers are running on the same machine. ● `<entry key="reference_genome">` Needs to match the client's “var reference_genome” variable
`server/src/main/resources/`**`chromosomes.xml`**	Provides a mapping between the chromosome names [String] used in the Map View (GViewer) and QTL Search, with the ones used in in the knowledge network [Integer].
`server/src/main/resources/`**`SemanticMotifs.txt`**	Contains semantic motifs as specified in Chapter 5.
`server/src/main/scripts/`**`startup.sh`**	Shell script to start the KnetMiner application server. Contains the filename of the knowledge network (OXL) and the maximum RAM allocation (-Xmx).
`server/src/main/scripts/`**`shutdown.sh`**	Shell script that terminates a running KnetMiner server process.
`server/`**`pom.xml`**	Specifies the maven build and dependency options.

6.6.3 Deployment of KnetMiner client and server

Once configured, the new KnetMiner client and server packages can be generated with the command:

```
mvn package
```

This will compile and package the maven module into ready-to-deploy files inside the subfolders named *target*. The **client** package consists of a WAR file that can be deployed on a Tomcat web server and opened in a web browser (e.g. http://localhost:8080/KnetMinerMySpecies). The **server** package consists of a zip-archive which includes a Java JAR file called *knetminer-server.jar*. This can be copied to a server and the Java multithreaded server can be started with the command:

```
java -Xmx10G -jar knetminer-server.jar MySpeciesNetwork.oxl
```

Note this requires that the server has at least 10GB of RAM and that the knowledge network (*MySpeciesNetwork.oxl*) is available in the execution folder.

After successful deployment, the KnetMiner website can be opened and a query submitted, the server log files will indicate that a request has been received and output files are created in the specified data folder (http://localhost:8080/species_data).

6.7 Discussion

The analysis and visualisation of large integrated datasets such as GSKNs requires scalable software solutions. We have taken the methodology developed in the previous chapter and wrapped it into a scalable client-server software resource, called KnetMiner, that gives anyone easy access to the integrated datasets. KnetMiner is a web application that **efficiently** interrogates the GSKN with user data such as search terms, QTL and gene list. The data visualisation components have been developed specifically for facilitating candidate gene discovery and hypothesis generation for research. The main key benefits of KnetMiner are:

- A visually appealing web application with a simple submission page for user data (keywords, QTL and gene list). The user is **guided** and **supported** when writing the search terms through features such as real-time user feedback and query term suggestions. No technical knowledge (metagraph, query statements) is required.

- The output is **dynamic** and **rich in detail** including different visualisations such as tables, networks and genome coordinate-based maps that are easy to navigate through a tabbed interface, and with extensive cross-referencing.

- A lightweight JavaScript-based network viewer, called KnetMaps, that is optimised for visualising and exploring data rich knowledge networks and incorporates search-specific **visualisation effects**.

- The underlying knowledge networks are **regularly updated** to include the latest database releases.

- The software platform is **configurable** and **portable** so that developers can easily build instances for new species and deploy them on any IT infrastructure that meets the software requirements.

Future work will investigate new methods for network analysis and data visualisation that are specifically directed to accelerating gene discovery research. For example, KnetMiner currently lacks the support for gene or annotation enrichment analysis (Glass and Girvan 2014) which could provide another view of the knowledge network without being restricted to user provided search terms. Additionally, we would like to investigate how gene expression data can be incorporated into the search and the network visualisation in order to further improve candidate gene scoring, decision making and hypotheses generation processes. Gene expression data could be either provided by the user directly or automatically retrieved from gene expression databases such as the Gene Expression Atlas (Petryszak et al. 2014).

The KnetMiner Map view uses currently the GMOD GViewer which requires Flash. We are currently in the process to replace it with a more modern and lightweight visualisation component and will, therefore, evaluate existing JavaScript libraries for genome coordinate-based visualisation of features such as gene, QTL or SNP data (Gómez et al. 2013). Our objective is to make KnetMiner as user-friendly as possible and compatible with mobile touch devices.

The query suggestion wizard benefits from the fact that many concepts in the knowledge network contain names and synonyms. Further work is needed to improve the selection process of the most suitable query suggestions by removing redundancies and taking advantage of the ontology structure when it is available. For example, the parent-child relations of an ontology (e.g. GO or TO) can be exploited to provide more specific or general query suggestions. Some other ideas and approaches for improving query suggestion workflows for the Life Sciences were discussed in our paper (Esch et al. 2014).

In summary, this chapter has successfully shown the implementation and benefits of a client-server design model for exploring large knowledge networks. Multiple instances of KnetMiner have been developed and deployed for species such as Arabidopsis, poplar, wheat, barley, potato, tomato, Brassica, maize, pig, cattle and chicken. Several of the KnetMiner instances have been developed as part of national and international collaborations, for example with the Roslin Institute (UK), National Agricultural Technology

Institute (Argentina), IPK Gatersleben (Germany) or the University of Western Australia. These have shown that the KnetMiner instances can be configured for diverse species. Work is in progress to develop new instances of KnetMiner for insects and pathogens. The next chapter will present applications of KnetMiner to real biological problems.

7 APPLICATIONS OF KNETMINER IN GENE DISCOVERY RESEARCH

In several biological studies, KnetMiner enabled the interpretation of hidden relationships between important agronomic traits and causal candidate genes. For example, it was used to investigate traits such as height of biomass willows (Hanley and Karp 2014) or to pinpoint the causal genes in a Arabidopsis petal size QTL (Koumproglou *et al*., submitted). This chapter presents two very different applications of KnetMiner. The first application case shows the utility of KnetMiner to help with the interpretation of transcriptomics (RNA-seq) experiments using an example dataset from 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 second application case presents the utility of KnetMiner for candidate gene prioritisation in GWAS and QTL data using an example dataset from *Arabidopsis thaliana* (Guillaume *et al*., submitted). Several of the identified gene-phenotype relationships are currently being validated using gene knockout or knockdown experiments in different species.

7.1 Using KnetMiner to interpret a transcriptomics study in wheat

7.1.1 Introduction

The majority of bread produced in the UK or US is from red-grained wheat[1] (Figure 7.1.1). 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 can be bred that lack the red compounds of the seed coat and are milder in flavor. Wholemeal bread made from white-grained varieties has therefore been found to be more appealing to people. 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. For this reason, white-grained wheats are mainly grown in warmer, drier parts of the world such as Australia, necessitating the costly importation of grain by UK millers and bakers.

[1] http://wholegrainscouncil.org/whole-grains-101/whole-white-wheat-faq

Figure 7.1.1: Lines of an Avalon x Cadenza doubled-haploid population segregating for red & white grains. The colour has been intensified by staining in 1M NaOH (Figure kindly provided by Andy Phillips, Rothamsted Research).

Major loci controlling grain colour in wheat are the *R Myb* homoeologous genes on the long arms of chromosomes 3A, 3B and 3D. Previous small-scale studies have shown that *R Myb* transcription factor gene regulates the transcriptional activation of four genes (*CHS, CHI, F3H and DFR*) in the flavonoid biosynthesis pathway (Figure 7.1.2) (Eiko and Noda 2005).

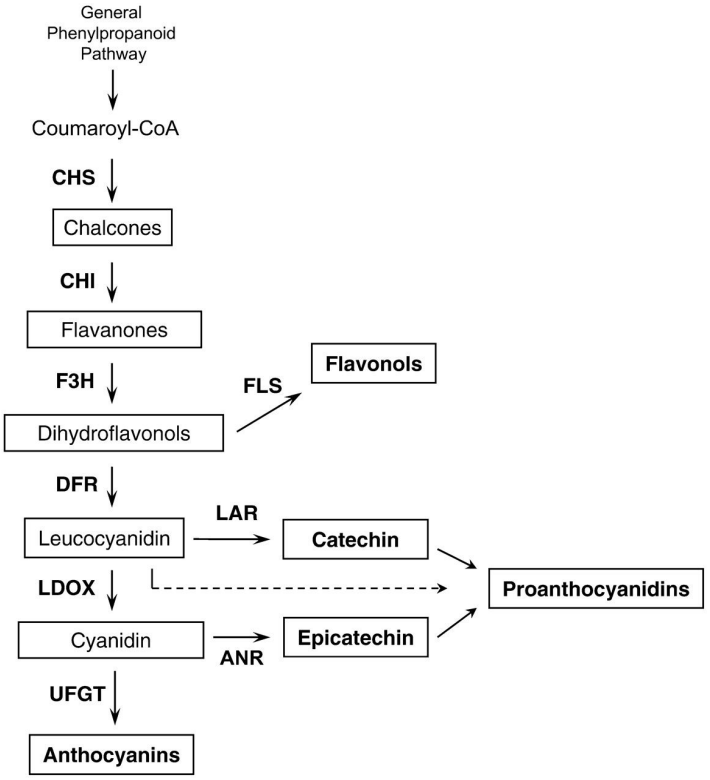

Figure 7.1.2: Scheme of the flavonoid pathway leading to synthesis of anthocyanins, flavonols, and proanthocyanidin (PA). The red pigment has been shown to be PA. The enzymes involved in the pathway are shown as follows: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone-3β-hydroxylase; DFR, dihydroflavonol-4-reductase; LDOX, leucoanthocyanidin dioxygenase; FLS, flavonol synthase; LAR, leucoanthocyanidin reductase; ANR, anthocyanidin reductase; and UFGT, UDP-Glc:flavonoid-3-O-glucosyltransferase. (Bogs et al. 2007)

A global transcriptome (RNA-seq) experiment was designed by the Phillips lab at Rothamsted Research to understand the transcriptional differences between red and white grains. The red and white lines used were near-isogenic lines NILs of a white wheat variety called Holdfast with introgression of a red R allele (Flintham 2000). The mRNA was extracted from isolated inner pericarp tissues (a tissue sample including the integuments) from developing grain of red (RI) and white (WI) lines. Three biological replicates per sample were included. The RNA was sequenced using Illumina HiSeq 2000.

As part of this thesis, the RNA-seq reads were mapped to the wheat reference genome (Ensembl v21, cDNA transcripts) using BWA (Heng Li and Durbin 2010). Transcript abundance was estimated using eXpress (Roberts and Pachter 2013) and differentially expressed genes identified with edgeR (Robinson, McCarthy, and Smyth 2009). In total 214

genes were differentially expressed ($p<0.05$) of which 104 had a considerable fold change (logFC > 2) between red and white grain (Figure 7.1.3). Of these 104 genes, 67 genes were lower in expression and 37 genes more highly expressed in the inner pericarp of white compared to red grain and might therefore be under the direct (or indirect) control of the R Myb transcription factor.

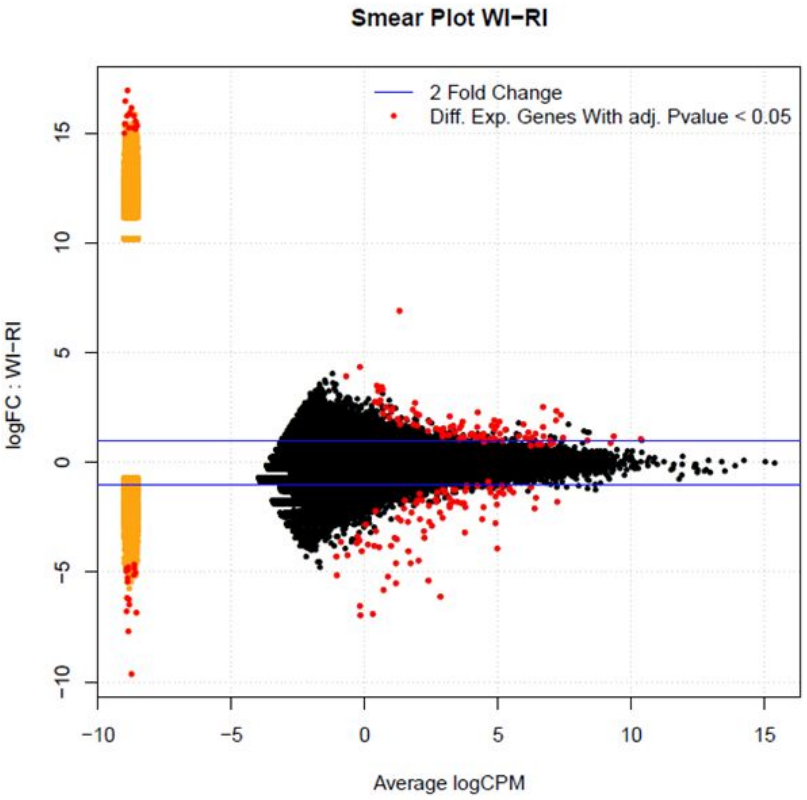

Figure 7.1.3: The fold change (FC) of a gene in red versus white grain is plotted as a function of average counts per million (CPM). Red dots indicate differentially expressed genes with $p<0.05$ and a minimum two-fold expression differences (blue lines).

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?
- Which other processes besides of grain colour and PHS are affected by the R (Myb) loss-of-function mutants?

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. Even when this functional information gathering task is complete, 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.

7.1.2 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 colour and PHS. These traits will first be analysed separately and afterwards together.

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. For example, the keyword *dormancy* matches Gene Ontology (GO), Trait Ontology (TO), gene, protein and publication evidence concepts from the wheat knowledge network (Figure 7.1.4). 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. When using *"germination" as a keyword*, it was necessary for similar reasons to be more specific and use "seed germination" as a keyword. The term *pre-harvest sprouting* appears to be part of the Gene Ontology and is suggested as a synonym for *seed germination*.

Figure 7.1.4: 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. In summary, KnetMiner was used with the following search queries:

1. Grain colour
    a. color OR flavon* OR proanthocyanidin
2. Pre-harvest sprouting (PHS)
    a. "seed germination" OR "seed dormancy"
    b. seed AND (germination OR dormancy)
3. Grain colour and PHS

a. “seed germination” OR “seed dormancy” OR color OR flavon* OR proanthocyanidin

Additionally, the IWGSC gene ids of the 104 differentially expressed wheat genes were entered into the *Gene List* and the option “Map gene list without restrictions” was selected.

7.1.3 General features for exploring genes supplied by the user

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.

As it was shown in Chapter 5, 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.

It can often be observed, especially in wheat, that several genes have exactly the same score and the same evidence information. This is a characteristic of genes that have nearly identical gene-evidence networks. In a hexaploid species such as bread wheat, there are three homoeologous copies of each gene which often results in all homoeologues having very similar gene-evidence networks and therefore identical scores and evidence information.

Figure 7.1.5 shows the gene-evidence network of 8 genes provided as potential candidate genes that have identical score and evidence information. The evidence from sequence homology indicates that these wheat genes are encoding CHS from the flavonoid pathway. Additional phenotype data as provided by TAIR (green rectangle) and text-mining based relations (blue edge with *) 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" (Higuchi et al. 2009). Protein-protein interaction data shows that CHS interacts with DFR, CHI and FLS.

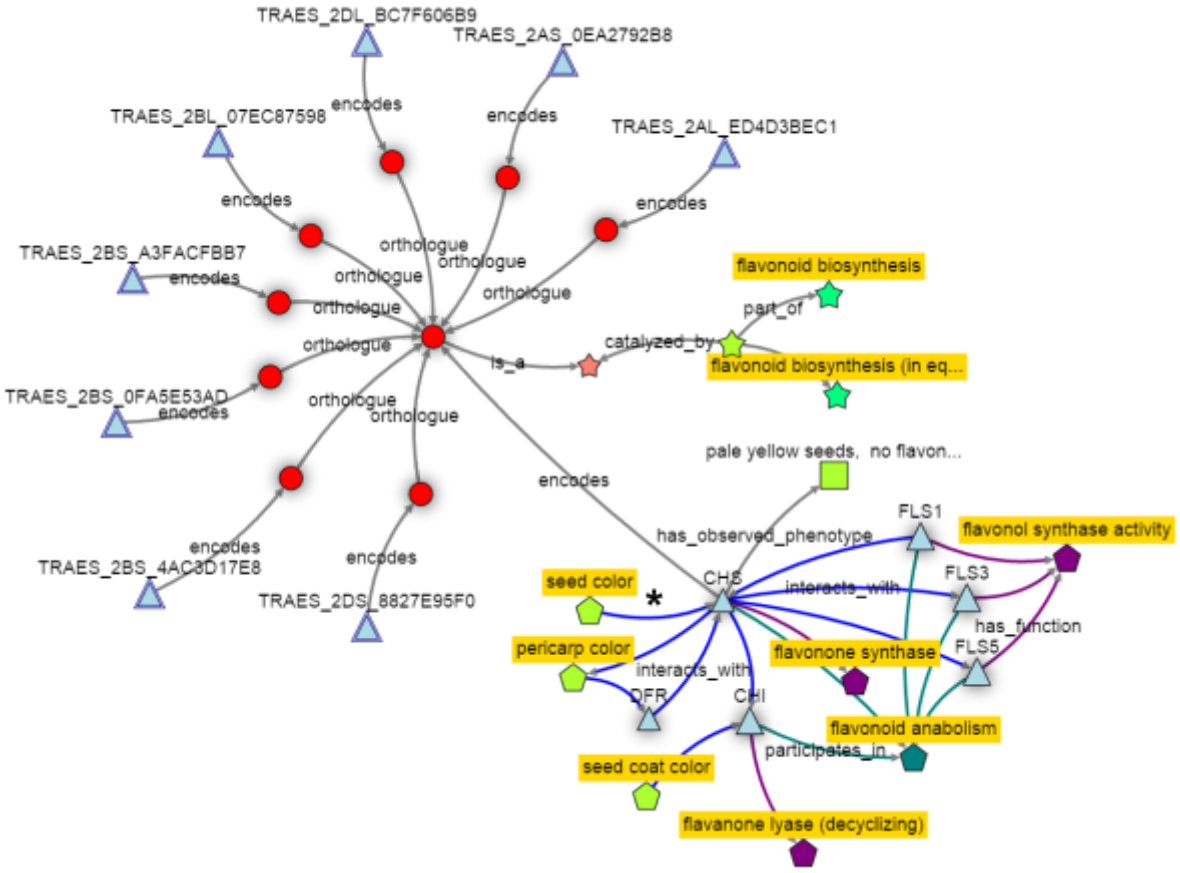

Figure 7.1.5: An excerpt from the gene-evidence network of 8 differentially expressed genes in white versus red samples. The ortholog of all 8 genes is the CHS gene from Arabidopsis.

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. Two distinct strategies are used to visualise the network of *known targets* versus the network of *novel targets*.

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*.

7.1.4 Candidate gene discovery for grain colour and pre-harvest sprouting traits

From the 104 differentially expressed wheat genes in red *versus* white grain, KnetMiner identifies 35 genes as being related to grain colour and 27 genes to traits related to germination or dormancy (Table 7.1). 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.

Table 7.1 shows 46 differentially expressed wheat genes in the red vs. white grain comparison (out of 104 genes with $p<0.05$ and $|logFC| > 2$) that KnetMiner can relate to grain colour or pre-harvest sprouting traits. The second column indicates the corresponding ortholog(s) in Arabidopsis or when not available the best Blast hit to other plants (UniProt).

Gene Id	Ortholog	Grain color	Pre-harvest sprouting
TRAES_4DS_8C9BC2BFA	AHP1, AHP2, AHP3, AHP5	1	1
TRAES_7AS_556F7B49E	ASK2/SKP1B	1	1
TRAES_2DS_8827E95F0	CHS	1	1
TRAES_2BS_A3FACFBB7	CHS	1	1
TRAES_2DL_BC7F606B9	CHS	1	1
TRAES_2BS_4AC3D17E8	CHS	1	1
TRAES_2AL_ED4D3BEC1	CHS	1	1
TRAES_2BL_07EC87598	CHS	1	1
TRAES_2AS_0EA2792B8	CHS	1	1
TRAES_2BS_0FA5E53AD	CHS	1	1
TRAES_3B_96D744B6B	DFR	1	1
TRAES_3AL_197871859	DFR	1	1

TRAES_7AL_8D7C375FF	FZR1, FZR2	1	1
TRAES_3B_0B9FADF42	IMD3	1	1
TRAES_2DL_C0E026879	STH/BBX25, STO/BBX24	1	1
TRAES_3B_82E1F5484	tny	1	1
TRAES_6AL_9032339D6	AT4G28570	1	0
TRAES_2BL_A06E8F248	AT5G45280, AT4G19410	1	0
TRAES_6DS_5B0F73A26	CHS-E (UniProt)	1	0
TRAES_2BL_B2B3C624C	ELI3	1	0
TRAES_4DL_5A3D8F519	F3'5'H (UniProt)	1	0
TRAES_2DL_F47B9B20E	F3H	1	0
TRAES_2BL_E3C1E6450	F3H	1	0
TRAES_2DL_4C86F28DC	F3H (UniProt)	1	0
ANON1	ANON1	1	0
ANON2	ANON2	1	0
ANON3	ANON3	1	0
TRAES_2DS_DE2E9E2D5	LBD37, LBD38, LBD39	1	0
TRAES_1AL_19AF54D53	LTL1, AT5G33370	1	0
TRAES_5BL_B2F45B45A	LTP6	1	0
TRAES_3DL_4D42B475B	MYB11, MYB111, MYB12	1	0
TRAES_2DL_2050A1ADC	PAL1, PAL2, PAL3, PAL4	1	0
TRAES_2DL_F4216BDB8	PME5	1	0
ANON4	ANON4	1	0
ANON5	ANON5	1	0
TRAES_3B_05835EDC0	BGAL10	0	1
TRAES_4DL_3E8E652D3	HSP/HSC (UniProt)	0	1
TRAES_1BS_27474466D	LCR69	0	1
TRAES_7AS_CC48D0C77	LCR69	0	1
TRAES_3DL_B9C57507A	NACA3	0	1
TRAES_3DS_0A0650113	OVA9	0	1
TRAES_3B_607315E21	PGY2	0	1
TRAES_3B_7A9E4CA37	TFL1/MFT	0	1
TRAES_2DL_83168C1E0	ZB8 (UniProt)	0	1
TRAES_2BL_5A50FDA1A	ZB8 (UniProt)	0	1
TRAES_2AL_9D78F85E2	ZB8 (UniProt)	0	1

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 (Figure 7.1.6). KnetMiner was therefore used with the search terms 3a. (see above) and the 16 common gene ids as parameters. 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, rice and other plant species. 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.

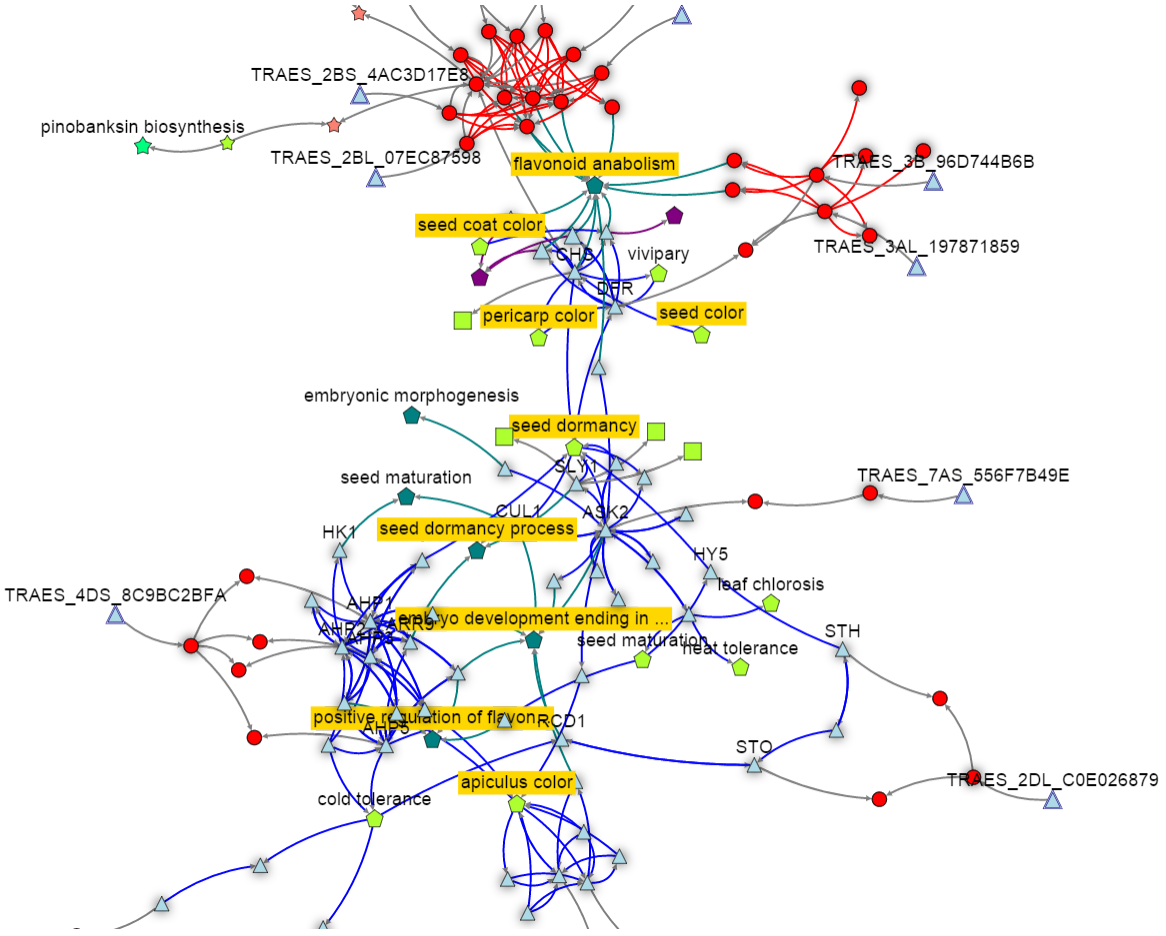

Figure 7.1.6: KnetMiner was used with the search terms: “dormancy OR color OR flavon* OR proanthocyanidin” and the common 16 genes for grain colour and PHS from Table 1 were selected and the gene-evidence network was visualised in the KnetMaps. Concept labels containing any of the search terms are highlighted yellow.

CHS (*TT4*) and DFR (*TT3*) in Arabidopsis are well-characterised enzymes in the flavonoid biosynthesis pathway and are linked to grain colour traits based on phenotype data and literature information, i.e. "*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*." (Higuchi et al. 2009). KnetMiner also links both genes to dormancy based on co-occurrence of the gene names with the term *seed dormancy* in (Martínez-Andújar et al. 2011); and the extracted evidence sentence: "*In another study, induction of the NCED6 gene in transgenic seeds of nondormant mutants tt3 and tt4 reestablished seed dormancy*."

Another interesting gene in the network is ARR4 which 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 (logFC = 3.4) expressed in white grain than in red grain. AHP1 and ARR4 are components of cytokinin signalling network (Hwang et al. 2012). 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.

7.1.5 Exploring novel candidate genes unrelated to initial search terms

The previous examples have shown the utility of KnetMiner for identifying and ranking candidate genes provided by a user based on the relevance to trait-based search terms, i.e. 46 differentially expressed genes that were related to grain colour or PHS traits. However, KnetMiner can also be used to discover the function of genes provided by a user that are not related to the initial search terms, i.e., the remaining 58 differentially expressed genes in the red *versus* red grain comparison.

These genes (referred to as novel targets) appear in the *Gene View* with a score of 0 with nil evidence documents because they are not related to any of the search terms. The gene-evidence networks of novel target genes can be studied individually or simultaneously by selecting *novel targets* at the top of the Gene View table and clicking *Show Network*. The network contains the selected genes and routes to connected GO or TO terms, but hides information such as publications in order to reduce the size of the visible network. Figure 7.1.7 shows the network of three homoeologous wheat genes that are more highly expressed in red grain (logFC=-5.58, p=6.6E-24) and appear to encode a transcription factor that regulates, among others, calcium signalling processes.

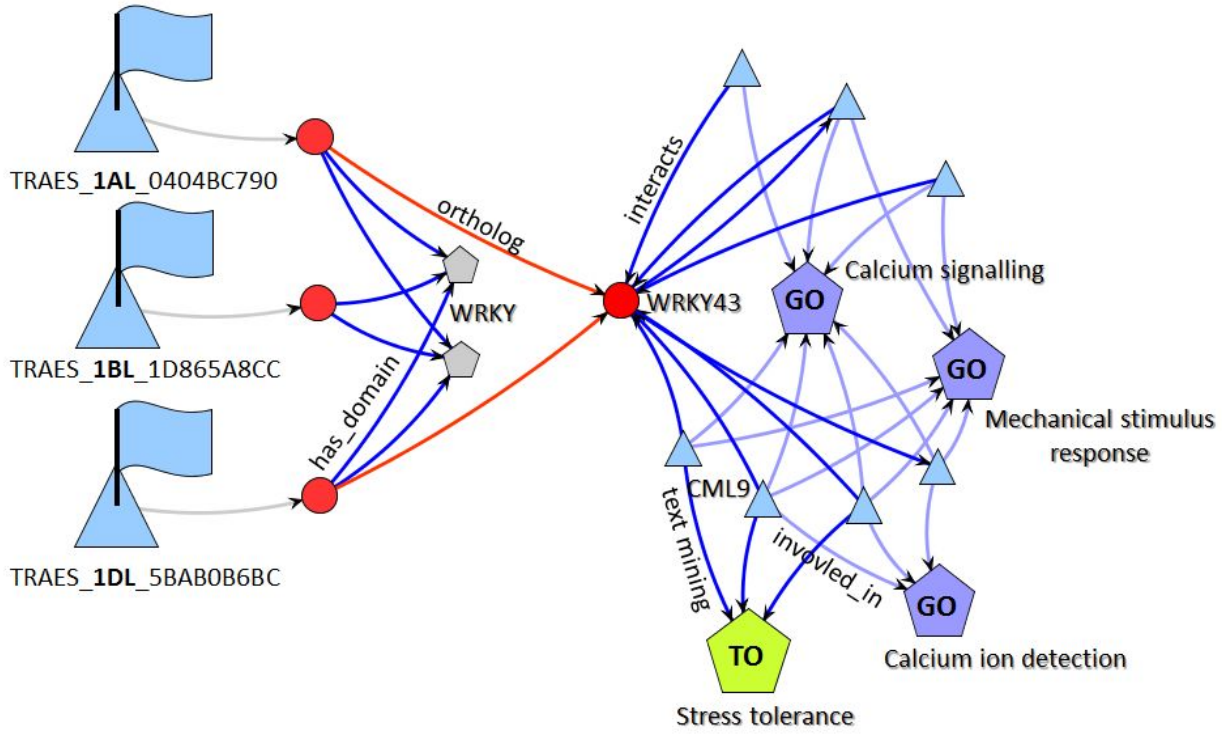

Figure 7.1.7: Gene-evidence network of homoeologous wheat genes containing the WRKY domain and are orthologs of the WRKY43 transcription factor in Arabidopsis. WRKY43 interacts with several gene products such as CML9 that are involved in calcium signalling and have evidence associating them with stress tolerance.

Selecting a large number of genes for network visualisation can result in very large networks despite the automatic data reduction steps that hide certain evidence types such as publications. To improve clarity, concepts and relations can be annotated based on specific attributes or network properties. The annotated network of the 58 novel target genes is shown in Figure 7.1.8. Important GO and TO concepts that are used to annotate several of the novel targets have an increased size and appear in the centre of the network. The

analysis of the network shows that further processes controlled directly or indirectly by the R *Myb* transcription factor in wheat, appear to be related to zinc binding, salt tolerance, lipid transport, cell wall differentiations and flower development.

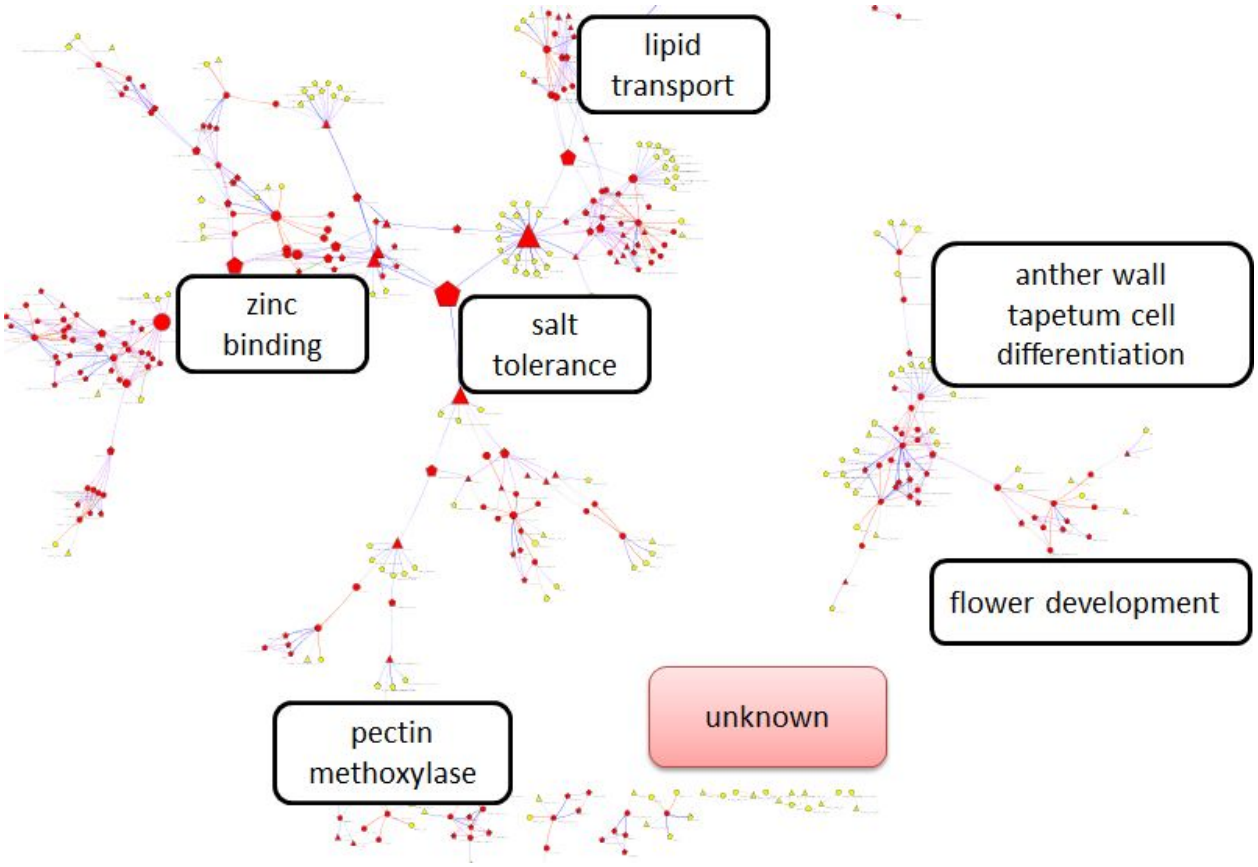

Figure 7.1.8: Annotated gene-evidence network of 58 novel target genes (created in Ondex).

7.1.6 Summary

It is known that grain colour in wheat is the result of flavonoid accumulation in the inner integument (testa). Mutations in the R (*Myb*) gene yield white grains (positive trait) and can cause pre-harvest sprouting (negative trait). R (*Myb*) is known to regulate several genes in the flavonoid pathway but it is not well understood which other genes are regulated by the R (*Myb*) transcription factor that can cause the grain to end dormancy and to start germination. A transcriptomics experiment was designed to identify genes that are differentially expressed between red and white grains, but means are required that can support the explanation of how these genes can control complex traits such as grain colour and PHS.

Provided with a list of genes and trait-based search terms (referred to as an initial, informal hypothesis), KnetMiner can rapidly search and evaluate a vast amount of heterogeneous relations and evidence types to determine if direct or indirect links between the genes and the hypothesis can be established. It produces a table of ranked candidate genes and allows users to explore their very rich gene-evidence networks. These networks provide an opportunity to explain how genes and biological processes are contributing to the original hypothesis or phenotype. In addition, they allow a user to identify potential new links to areas that have not been considered before. Such networks can contain complex interactions that require appropriate visualisation tools to navigate the highly interlinked information. In conclusion, KnetMiner gives domain experts (biologists) the required tools to systematically dissect a complex trait, identify trait-related candidate genes, and to refine an original hypothesis or define new hypotheses through the exploration of biological knowledge networks.

Many of the genes identified here encode known enzymes of the flavonoid biosynthetic pathway: chalcone synthase (CHS), flavanone 3-hydroxylase (F3H), dihydroflavonol 4-reductase (DFR). Expression of several of these genes was confirmed by quantitative reverse transcription polymerase chain reaction (QRT-PCR). The biological validation of gene-phenotype relationships identified by KnetMiner is currently being explored by reverse genetics tools in wheat such as RNAi (Travella 2006), TILLING (Chen et al. 2012) or CRISPR/Cas9 (Shan et al. 2014) to generate knock-down or knockout lines and study the phenotype.

The KnetMiner Evidence View contains enrichment information for documents that match the search terms. Future work will include the development of an Enrichment View that provides integrated tools for gene and annotation enrichment analysis regardless whether they are related to the search terms or not. The enrichment analysis would take into account the gene-evidence networks and look for enrichment of any type of evidence document including GO, TO, pathways etc.

7.2 Using KnetMiner to interpret GWAS and QTL studies in Arabidopsis

7.2.1 Introduction

Arabidopsis thaliana is a small flowering plant that is commonly used as a model organism in plant biology. Although it is not of major agronomic importance, Arabidopsis offers significant

advantages for basic research in genetics and molecular biology. It has a relatively small genome size (135Mb) consisting of 5 diploid chromosomes and 27,416 coding genes and a short life cycle of about 6 weeks from germination to seed maturation. Such advantages have made Arabidopsis a model organism for studies of a large number of plant traits. Recently, the Multiparent Advanced Generation Inter-Cross (MAGIC) population was developed as a resource for identifying quantitative trait loci (QTL) in Arabidopsis (Kover et al. 2009). About 500 MAGIC lines have been resequenced at low coverage in order to obtain about 500k single nucleotide polymorphisms (SNPs) for each line. Peter Eastmond and his lab at Rothamsted Research have grown the MAGIC lines in a glasshouse experiment and measured the content of different fatty acids (chemical phenotypes). The trait measurements were normalised using REsidual Maximum Likelihood (REML).

As part of this thesis, the data from this study was analysed using *genome_scan* (v4.0) in order to identify significant genotype-phenotype associations and QTL across the 5 Arabidopsis chromosomes (http://mus.well.ox.ac.uk/19genomes/magic.html). The *genome_scan* output for a given trait is a table of SNPs and p-values indicating the significance of a polymorphism to that phenotype. These can be visualised in so-called Manhattan graphs that plot the logP value of every SNP sorted by genomic coordinates. Figure 7.2.1 shows the results for palmitic acid content (the first fatty acid produced during fatty acid synthesis). Many statistically significant SNPs can be identified even after choosing logP>5 (blue line) or even logP>7 (red line) as a significance threshold. Similarly for a mapped QTL the identified genomic region can encompass tens to hundreds of genes.

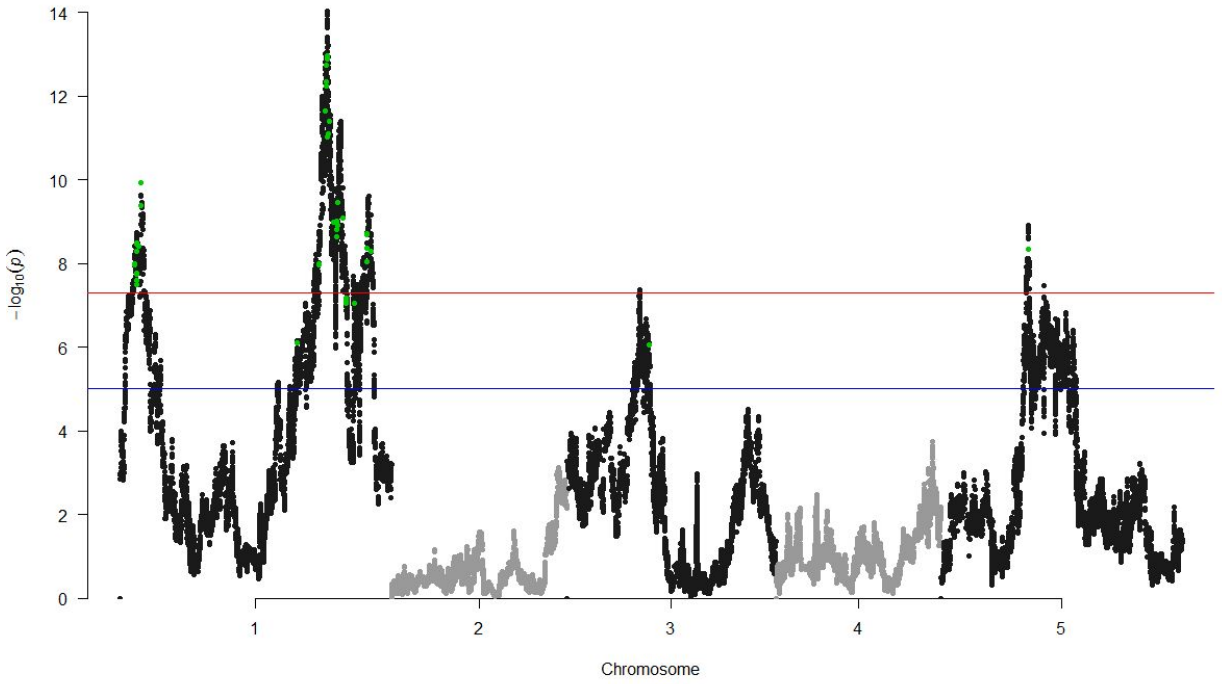

Figure 7.2.1: Manhattan plots showing significant SNPs for palmitic acid content (the first fatty acid produced during fatty acid synthesis). Significant loci (logP > 7) are found on chromosomes 1, 3 and 5. Green dots represent SNPs in proximity to known fatty acid genes. The plot was created using the R package qqman (Turner 2014).

Having identified statistically significant SNPs and QTLs for a given trait, the questions consequently asked by the users of such data are:

- Do the SNPs occur within or in the neighbourhood of trait-related candidate genes?
- Do the QTLs contain any trait related candidate genes?
- Which biological processes and pathways underlie these traits?
- Which other phenotypes are influenced by these genes (alleles)?

Evaluating the functional candidacy of every potential candidate gene requires a user to navigate heterogeneous evidence sources including functional gene annotations, phenotype data, scientific literature, gene expression information, protein-protein interaction and other relevant datasets to genetics. Explaining how complex, polygenic traits are influenced by the genes (alleles) identified in the genetics study is an even harder challenge that requires as the first step the assembly of a knowledge network. This can quickly become a time-consuming and resource-intensive task that can be prone to information being missed

and subjective biases being introduced. Here, we demonstrate how KnetMiner can be used to prioritise candidate genes that resulted from GWAS and QTL studies in a reproducible, effective and systematic manner.

7.2.2 Identifying candidate genes in GWAS output

KnetMiner for Arabidopsis was used with the search term "*fatty acid OR lipid"* which returned 6932 genes ranked by score. The score ranges from 0.01 to 4.13 based on the relevance of the gene to fatty acid pathways, processes, phenotypes etc. The results were downloaded from the *Gene View* in tabular format. The downloaded file contains for every gene the chromosome, start and stop information. The KnetMiner list was compared with a list of 774 expert curated Arabidopsis lipid genes (http://aralip.plantbiology.msu.edu/data/aralip_data.xlsx) that are largely restricted to enzymes. From the 774 expert curated lipid genes 630 (81.4%) occur in the KnetMiner list.

A custom Python script was written that checks for every significant SNP (logP > 6) in the GWAS output if it is located within or 1000bp down/up-stream of a candidate gene provided by KnetMiner. If this is true the SNP id in the GWAS output is changed to the corresponding name of Arabidopsis candidate gene and the candidate gene is recorded in a separate file. In total, 63 Arabidopsis genes were identified with significant genetic variation (alleles) in the MAGIC lines that can alter total lipid (palmitic acid) content. SNPs that are within or in proximity of the 63 genes provided by KnetMiner are shown as green dots in Figure 7.2.1. The Manhattan plot shows that many of the peak SNPs are in the neighbourhood of potential candidate genes.

Functional analysis of genes that may influence traits underlying lipid content abnormalities in some MAGIC lines could be studied one by one, however, the bigger biological picture is usually more evident when the significant genes are studied simultaneously in an integrated, connected manner. KnetMiner can, therefore, be used once again and supplied with the same search terms "*fatty acid OR lipid"* and the list of 63 significant genes as the *User Gene List*. The results are identical to the first search, but this time the genes provided by the user are indicated with a "yes" in the *Gene View*. These can be selected individually or by using the *target genes* checkbox. Pressing the *View Network* button generates a connected and integrated gene-evidence network for all selected candidate genes (Figure 7.2.2). The heterogeneous evidence concepts in the network include Gene Ontology annotations,

AraCyc pathways, phenotype data, literature references, protein-protein interactions, relations to UniProt based on BLAST, relations to Trait Ontology based on text-mining. Initially only those paths from the gene-evidence network are shown where there is a search term and all other concepts are hidden, but can be displayed if additional information is required. The gene-evidence network can be explored to determine biological processes and pathways underlying this complex, polygenic trait. The analysis showed that many of the genes were involved in processes and pathways related to fatty acid metabolism.

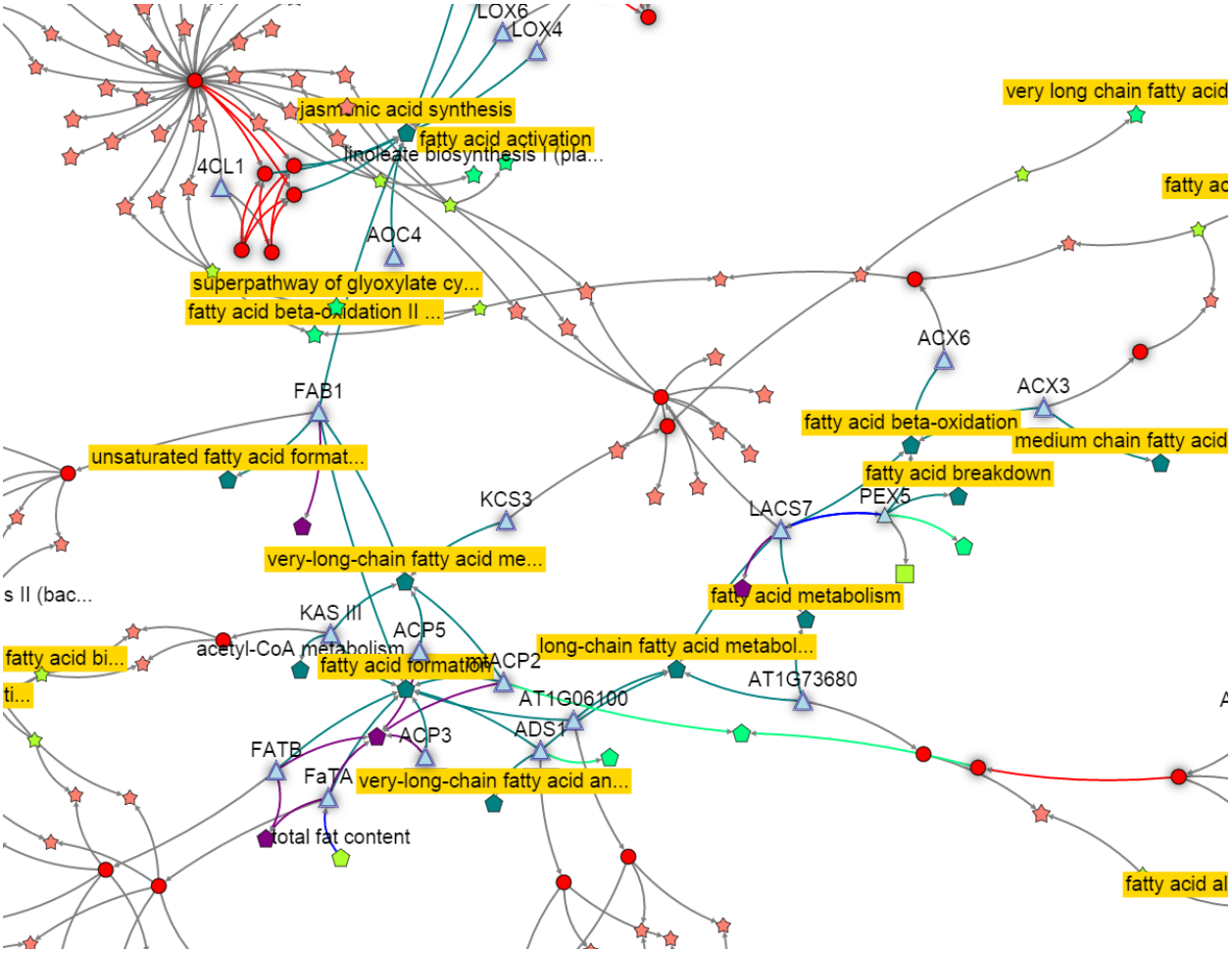

Figure 7.2.2: Integrated gene-evidence network of significant lipid genes that were indicated with green dots in the previous Manhattan plot.

7.2.3 Identifying candidate genes in QTL mapping output

This section presents an alternative workflow for identifying candidate genes in QTLs that are associated with a phenotype. The previous workflow analysed every SNP from the association mapping individually. Alternatively, it is possible to define regions that are above

a certain p-value as a QTL. The borders where the red threshold line cuts the peaks of the graph in Figure 7.2.1 can be defined as a QTL. For example, choosing logP > 7 as a statistical threshold, it is possible to define 2 large QTL region on chromosome 1, one small region on chromosome 3, and one on chromosome 5.

- QTL 1: Chromosome 1, 935413- 2901633 [562 genes]
- QTL 2: Chromosome 1, 21722576 - 28378561 [1698 genes]
- QTL 3: Chromosome 3, 8132487 - 8175461 [10 genes]
- QTL 4: Chromosome 5, 9529165 - 9862793 [74 genes]

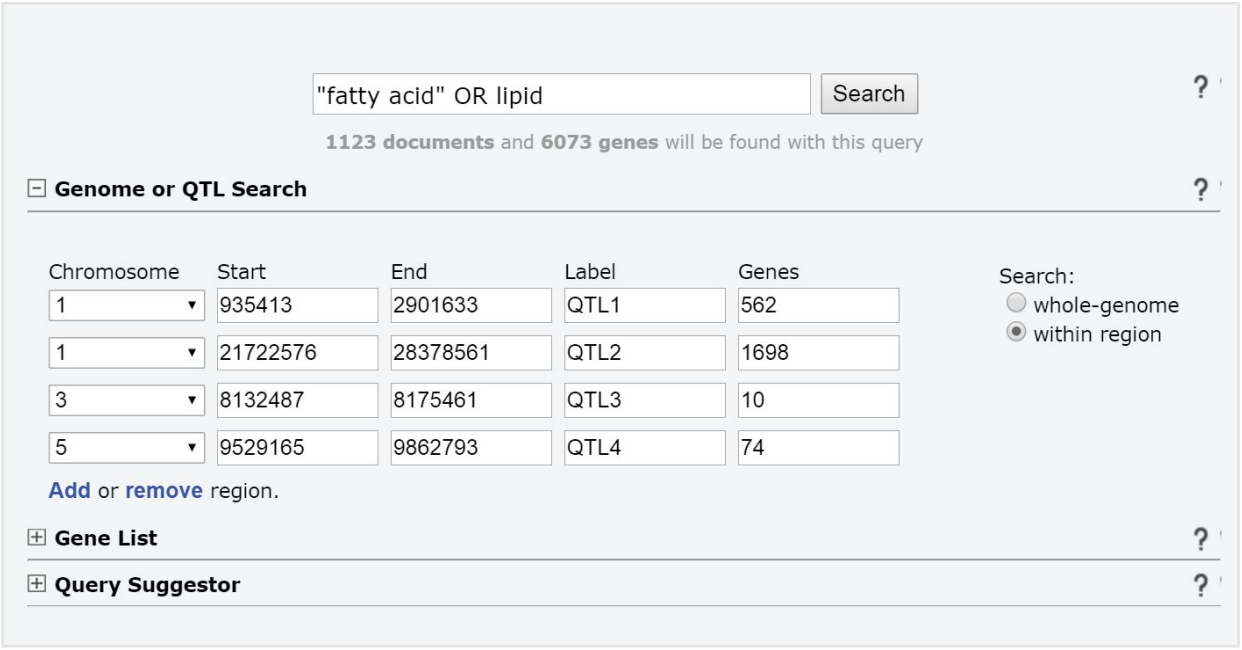

Figure 7.2.3: Definition of search terms and multiple QTL in the KnetMiner search interface.

The defined genomic regions contain in total 2,344 potential candidate genes. KnetMiner evaluates every gene whether it can be directly or indirectly related to the search terms, and ranks the genes based on the computed relevance score (see Chapter 5).

The *Gene View* contains the ranked candidate genes with their location and the evidence concepts. The top scoring gene from each QTL are AT1G64400, KCS2 and AT5G27600. No trait-related candidate genes were found in the QTL on chromosome 3.

ACCESSION	CHRO	START	SCORE	USER	QTL	EVIDENCE	Select
AT1G64400	1	23915598	4.17	no	0	3 11 7 1 4	☑
KCS2	1	1119699	2.74	no	0	1 13 6 1 1 5	☑
ADS1	1	1843568	2.30	no	0	3 5 1 1	☐
AT1G60660	1	22342414	2.26	no	0	2	☐
AT5G27600	5	9742576	1.89	no	0	1 4 10 3 1 1 2	☑
AT1G08640	1	2748398	1.87	no	0	1 1 2	☐
CYP86A7	1	23632178	1.81	no	0	1 1 3 1 1	☐

Figure 7.2.4: QTL genes are ranked and the evidence concepts are summarised in the Gene View. The selected genes *LACS3* (AT1G64400), *KCS2* (AT1G04220) and *LACS7* (AT5G27600) are the top scoring gene of each QTL.

The gene-evidence network of the three top candidate genes can be easily generated and viewed in the KnetMaps. Exploring the relations between the genes can identify common processes and pathways that might explain the complex nature of the trait and justify why several QTL were identified. Figure 7.2.5 shows the gene-evidence network of *LACS3* (AT1G64400), *KCS2* (AT1G04220) and *LACS7* (AT5G27600). It is evident that all three genes encode enzymes (long-chain-fatty-acid CoA ligases) that catalyse reaction RXN-7904 of the fatty acid activation pathway and RXN-9644 of the linoleate biosynthesis I (plants) pathway. The *Item Info* of KnetMaps provides external links to the original data sources, for example to PlantCyc, so that further details such as chemical equations of the reaction can be easily reached from KnetMiner.

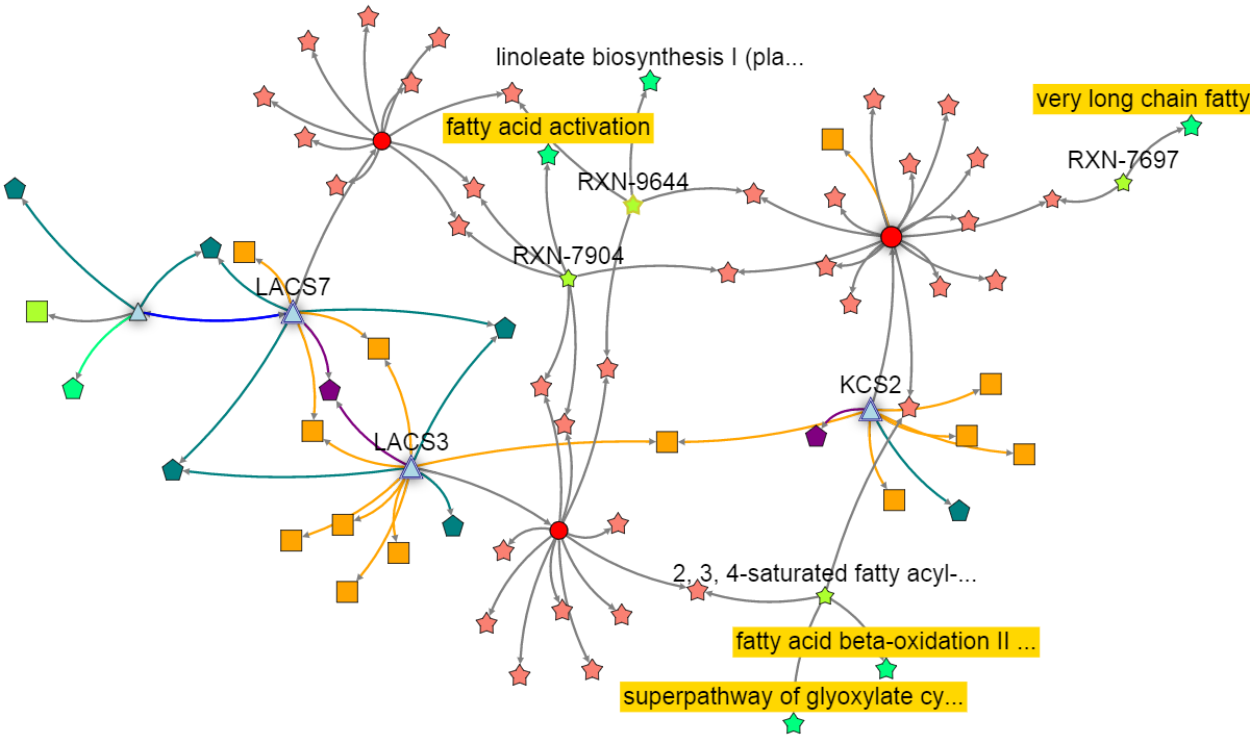

Figure 7.2.5: The gene-evidence network of the top ranked genes *LACS3, KCS2 and LACS7* from each QTL. All three genes encode enzymes that catalyse reaction RXN-7904 of the fatty acid activation pathway and RXN-9644 of the linoleate biosynthesis I (plants) pathway.

Interestingly, the highest scoring gene in KnetMiner is *LACS3* (AT1G64400) which is not part of the 63 candidate genes identified in the SNP-based analysis of the GWAS data. *LACS3* is located on chromosome 1 from 23,915,598 to 23,919,783. The closest SNPs are about 8kb downstream and 4kb upstream of *LACS3*. Although these two SNPs are significant (logP = 8.3) for total lipid content, their distance exceeded the arbitrary maximum distance of 1kb that was defined as the selection criteria for candidate genes. This shows that although the SNP based approach is potentially very accurate, important candidate genes can easily be missed because of uneven or low marker (SNP) density, or because of arbitrary threshold definitions such as logP or maximum distance to candidate gene.

The *Evidence View* and the *Map View* provide further complementary views that can help to visualise the relative position of candidate genes within a QTL or to systematically dissect the biological processes and pathways underlying QTL genes. The column QTL in the *Evidence View* indicates how many of the QTL genes are related to that specific evidence (e.g GO concept, pathway, reaction).

7.2.4 Summary

GWAS or QTL studies are only the beginning of every gene discovery investigation to determine genes, biological processes and pathways underlying a complex, polygenic trait. The challenges of evaluating the functional candidacy of potential candidate genes include data acquisition, integration, mining and visualisation. This use case showed the utility of KnetMiner in facilitating the interpretation of genetic experiments. Two different approaches for using KnetMiner were illustrated: a SNP-based and a QTL-based approach.

The SNP-based approach was divided in two steps. First, KnetMiner was used to create a genome-wide list of candidate genes for a given trait. This list was refined outside of KnetMiner with a script that takes as input the GWAS (SNP-Pvalue) output and the KnetMiner genes in order to evaluate if any genes have at least one significant SNP within a 1kb radius from the start and end of the gene. This approach identified 63 Arabidopsis genes with significant genetic variation (alleles) in the MAGIC lines that can alter total lipid (palmitic acid) content. The analysis showed that many of these genes were involved in processes and pathways related to fatty acid metabolism.

The QTL-based approach consists of a single step. KnetMiner was used with 4 distinct QTL regions for palmitic acid content and trait-related search terms as parameters. The defined genomic regions contained in total 2,344 potential candidate genes. In contrast to the previous 2-step approach, KnetMiner directly evaluated and ranked all QTL genes. The top candidate gene per QTL was *LACS3* (this gene was not found in the previous SNP-based approach), *KCS2* and *LACS7*. The integrated networks of the top scoring gene from each QTL were studied and interesting connections were found such as all three genes encode enzymes (long-chain-fatty-acid CoA ligases) that catalyse reaction RXN-7904 of the fatty acid activation pathway and RXN-9644 of the linoleate biosynthesis I (plants) pathway. Some of these genes (alleles) are currently subject of experimental validation.

In conclusion, KnetMiner has a user-friendly interface that facilitates the biological interpretation of GWAS and QTL data. It has a predictive component that ranks candidate genes and a explorative component that enables domain experts to generate hypotheses that can explain the translation of the genotype to the phenotype via network biology. Future work would investigate a more seamless integration of GWAS input data into the KnetMiner user interface and the development of analytical tools for the exploration of public genetics

resources such as AnimalQTLdb (Hu, Fritz, and Reecy 2007) or Triticeae Toolbox (Blake et al. 2016).

8 CONCLUSION

Biological knowledge discovery is often hampered by the challenges of data integration and new approaches are needed to improve the efficiency, reproducibility and objectivity gene discovery. 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 connect, explore and compare the wealth of genetic, genomic, transcriptomic, proteomic and phenotypic information available in the literature, from key model species and from a potentially wide range of related biological databases.

The first major achievement of this work was the development of genome-scale knowledge networks (GSKNs) for 11 species including crops such as wheat, maize and willow. This was achieved by extending the Ondex data integration platform with text mining capabilities (Hassani-Pak et al. 2010) and by optimising the process of building knowledge networks (Hassani-Pak et al. 2016). The process is pragmatic in that it allows a network of appropriate complexity to be developed and updated without an excessive technical and semantic burden to the user. Feasibility studies have shown that knowledge networks provide a suitable data structure for effective gene mining and biological knowledge discovery. GSKNs can encompass millions of labelled nodes, semantic links and manifold attributes. The current version of the Arabidopsis GSKN integrates Arabidopsis gene and proteins with multiple information types including gene-SNP-phenotype associations, protein-protein interactions and annotations to GO, EC, pathway, protein domain and publications. We conducted a study to evaluate the suitability of annotation transfer from model species such as Arabidopsis to non-model species such as wheat or rice (Defoin-Platel, Hassani-Pak, and Rawlings 2011). For example, the Arabidopsis GSKN integrates homology relations to the yeast interactome and yeast GO annotations that can be useful in understanding developmental traits in plants. All crop GSKNs link into the Arabidopsis knowledge network via orthology relations which can be exploited for the transfer of phenotypic information. Importantly, GSKNs not only contain information from structured databases but also novel gene-phenotype relationships extracted from unstructured PubMed abstracts by our own text mining tools as described in Chapter 4. In the past, despite of appreciating the value in GSKNs, biologists and breeders were unable to take great advantage of these resources because of the slow and cumbersome process to interrogate them using the available Ondex standalone application.

The second major achievement was the development of new web-based tools for mining and visualising large knowledge networks. The KnetMiner web server searches, evaluates and scores 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. Modified measures of information content are used to rank potential candidate genes for their relevance to the trait. KnetMiner accepts as user inputs: search terms in combination with a gene list and/or genomic regions. It produces as outputs: (i) ranked candidate genes and supporting evidence tables, (ii) interactive network maps to visualise and explore gene-knowledge networks and (iii) interactive chromosome maps with genes, SNP, QTL, GWAS data. All components have been optimised for web use on desktop and mobile touch devices. Feasibility studies in different crop species demonstrated that KnetMiner can enable biological knowledge discovery which was not easily possible before in these species. For example, it supported the discovery of an inferred relationship between a gene and a plant height phenotype in willow (Hanley and Karp 2014), a gene that might be controlling grain color and pre-harvest sprouting in wheat (manuscript submitted) and a gene controlling petal size in Arabidopsis (manuscript submitted). These and other examples have shown that the KnetMiner web server and the GSKNs are important tools for biologists and breeders wanting to interpret the results of genetic and omics studies.

In summary, the main key benefits of KnetMiner are:

- The user is **guided** and **supported** when writing the search terms through features such as real-time user feedback and query term suggestions. No technical knowledge (metagraph, query statements) is required.

- The output is **dynamic** and **rich in detail** including different interactive visualisations such as tables, network and genome maps that are easy to navigate through a tabbed interface, and with extensive cross-referencing.

- Support for **non-model** diploid and polyploid species and different information types to connect genes to phenotypes including functional annotation, genetic association, homology, protein-protein interactions and text-mining.

- The underlying knowledge networks are built automatically and are **regularly updated** to include the latest database releases as described in Chapter 3.

- The software platform is **configurable** and **portable** so that developers can easily build instances for new species and deploy them on any IT infrastructure that meets the software requirements.

Future work will investigate new methods for knowledge network mining and data visualisation that are specifically directed to accelerating gene discovery research. For example, KnetMiner currently lacks the support for gene or annotation enrichment analysis (Glass and Girvan 2014) which could provide another view of the knowledge network without being biased by the user provided search terms. Additionally, we would like to investigate how gene expression data can be incorporated into the search and the network visualisation in order to further improve candidate gene scoring. Gene expression data could be either provided by the user directly or automatically retrieved via APIs from gene expression databases such as the Gene Expression Atlas (Petryszak et al. 2014). Future work would also investigate a more seamless integration of GWAS input data into the KnetMiner user interface and the development of analytical tools for the exploration of public genetics resources such as AnimalQTLdb (Hu, Fritz, and Reecy 2007) or Triticeae Toolbox (Blake et al. 2016).

Currently the proposed gene ranking algorithm gives equal weight to all integrated evidence documents. However, not all integrated datasets are of same quality. Future work would be to investigate a weighting scheme for different types of evidence. For example, genes that have causative SNPs linked to phenotypes could be ranked higher than evidence that is inferred through sequence homology; or curated gene-phenotype evidence can be weighted stronger than text-mining based evidence. Since this choice is generally dependent on the application, the evidence weighting scheme could be made configurable by the user. Finally, the proposed gene ranking approach computes a score based on frequency and specificity of evidence documents, however it does not determine how significant the score is, i.e. the probability of obtaining a given score by chance. Similar to the approach of BLAST using score and e-value, our future work would develop an e-value for KnetMiner gene scores which would estimate the number of hits retrieved by chance for a given search term and a GSKN with certain network properties.

In summary, KnetMiner is an intuitive tool that makes gene discovery fun and efficient for biologists and breeders. Some KnetMiner servers (e.g. for Arabidopsis, wheat, poplar) have

been running for several years and new KnetMiner servers are about to emerge soon. KnetMiner is used by different labs at Rothamsted Research and elsewhere to accelerate gene discovery pipelines for crop breeding and crop improvement. While we have so far mostly concentrated on crop species, the approaches we have taken are generic and GSKNs and KnetMiner servers can readily be built for other species. This PhD thesis described the version of the KnetMiner software that was available at the time of writing. We are constantly improving the usability of the software, adding new features and extending the knowledge mining approaches. The latest version of the KnetMiner software and documentation will be available at: http://knetminer.rothamsted.ac.uk/.

References

Altenhoff, Adrian M., Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, et al. 2015. “The OMA Orthology Database in 2015: Function Predictions, Better Plant Support, Synteny View and Other Improvements.” *Nucleic Acids Research* 43 (Database issue):D240–49.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3):403–10.

Ananiadou, Sophia, Ananiadou Sophia, Douglas B. Kell, and Tsujii Jun-ichi. 2006. “Text Mining and Its Potential Applications in Systems Biology.” *Trends in Biotechnology* 24 (12):571–79.

An, Gynheung, An Gynheung, Jeong Dong-Hoon, Jung Ki-Hong, and Lee Sichul. 2005. “Reverse Genetic Approaches for Functional Genomics of Rice.” *Plant Molecular Biology* 59 (1):111–23.

“Apache Lucene - Welcome to Apache Lucene.” n.d. Accessed September 9, 2015. https://lucene.apache.org/.

“Apache UIMA.” n.d. Accessed September 9, 2015. http://uima.apache.org/.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1):25–29.

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. “Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines.” *Nature* 465 (7298):627–31.

Benfey, Philip N., and Thomas Mitchell-Olds. 2008. “From Genotype to Phenotype: Systems Biology Meets Natural Variation.” *Science* 320 (5875):495–97.

Berggård, Tord, Berggård Tord, Linse Sara, and James Peter. 2007. “Methods for the Detection and Analysis of Protein–protein Interactions.” *Proteomics* 7 (16):2833–42.

Biemann, Chris, Biemann Chris, Krumov Lachezar, Roos Stefanie, and Weihe Karsten. 2016. “Network Motifs Are a Powerful Tool for Semantic Distinction.” In *Understanding Complex Systems*, 83–105.

Blake, Victoria C., Clay Birkett, David E. Matthews, David L. Hane, Peter Bradbury, and Jean-Luc Jannink. 2016. “The Triticeae Toolbox: Combining Phenotype and Genotype Data to Advance Small-Grains Breeding.” *The Plant Genome* 9 (2). https://doi.org/10.3835/plantgenome2014.12.0099.

Bornigen, D., Tranchevent L.-C., F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, and Y. Moreau. 2012. “An Unbiased Evaluation of Gene Prioritization Tools.” *Bioinformatics* 28 (23):3081–88.

Canevet, Catherine, Canevet Catherine, Canevet Catherine, Lysenko Artem, Splendiani Andrea, Pocock Matthew, and Rawlings Chris. 2010. “Analysis and Visualisation of RDF Resources in Ondex.” *Nature Precedings*. https://doi.org/10.1038/npre.2010.5430.

Carter, Hannah, Matan Hofree, and Trey Ideker. 2013. “Genotype to Phenotype via Network Analysis.” *Current Opinion in Genetics & Development* 23 (6):611–21.

Caspi, Ron, Caspi Ron, Altman Tomer, Billington Richard, Dreher Kate, Foerster Hartmut, Carol A. Fulcher, et al. 2013. “The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases.” *Nucleic Acids Research* 42 (D1):D459–71.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan Beauchamp, David Cesarini, Matthijs van der Loos, Magnus Johannesson, et al. 2012. “Most Reported Genetic Associations with General Intelligence Are Probably False Positives.”

Psychological Science 23 (11):1314–23.
Chapman, Jarrod A., Martin Mascher, Aydın Buluç, Kerrie Barry, Evangelos Georganas, Adam Session, Veronika Strnadova, et al. 2015. “A Whole-Genome Shotgun Approach for Assembling and Anchoring the Hexaploid Bread Wheat Genome.” *Genome Biology* 16 (January):26.
Chatr-aryamontri, A., Breitkreutz B.-J., R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, et al. 2014. “The BioGRID Interaction Database: 2015 Update.” *Nucleic Acids Research* 43 (D1):D470–78.
Chen, Liang, Linzhou Huang, Donghong Min, Andy Phillips, Shiqiang Wang, Pippa J. Madgwick, Martin A. J. Parry, and Yin-Gang Hu. 2012. “Development and Characterization of a New TILLING Population of Common Bread Wheat (Triticum Aestivum L.).” *PloS One* 7 (7):e41570.
Ching, Ada, Katherine S. Caldwell, Mark Jung, Maurine Dolan, Oscar S. Smith, Scott Tingey, Michele Morgante, and Antoni J. Rafalski. 2002. “SNP Frequency, Haplotype Structure and Linkage Disequilibrium in Elite Maize Inbred Lines.” *BMC Genetics* 3 (October):19.
Cohen, K. Bretonnel, and Lawrence Hunter. 2008. “Getting Started in Text Mining.” *PLoS Computational Biology* 4 (1):e20.
Cunningham, Hamish, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. “Getting More out of Biomedical Documents with GATE’s Full Lifecycle Open Source Text Analytics.” *PLoS Computational Biology* 9 (2):e1002854.
“Cypher Query Language.” n.d. Neo4j Graph Database. Accessed November 22, 2017. https://neo4j.com/developer/cypher/.
Davies, G., A. Tenesa, A. Payton, J. Yang, S. E. Harris, D. Liewald, X. Ke, et al. 2011. “Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic.” *Molecular Psychiatry* 16 (10):996–1005.
Defoin-Platel, Michael, Keywan Hassani-Pak, and Chris Rawlings. 2011. “Gaining Confidence in Cross-Species Annotation Transfer: From Simple Molecular Function to Complex Phenotypic Traits.” *Aspects of Applied Biology / Association of Applied Biologists* 107 (April):79–87.
Edgar, R. 2002. “Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository.” *Nucleic Acids Research* 30 (1):207–10.
Eiko, Himi, and Kazuhiko Noda. 2005. “Red Grain Colour Gene (R) of Wheat Is a Myb-Type Transcription Factor.” *Euphytica/ Netherlands Journal of Plant Breeding* 143 (3):239–42.
Esch, Maria, Jinbo Chen, Christian Colmsee, Matthias Klapperstück, Eva Grafahrend-Belau, Uwe Scholz, and Matthias Lange. 2015. “LAILAPS: The Plant Science Search Engine.” *Plant & Cell Physiology* 56 (1):e8.
Esch, Maria, Jinbo Chen, Stephan Weise, Keywan Hassani-Pak, Uwe Scholz, and Matthias Lange. 2014. “A Query Suggestion Workflow for Life Science IR-Systems.” *Journal of Integrative Bioinformatics* 11 (2):237.
Eskandari, Mehrzad, Elroy R. Cober, and Istvan Rajcan. 2013a. “Genetic Control of Soybean Seed Oil: I. QTL and Genes Associated with Seed Oil Concentration in RIL Populations Derived from Crossing Moderately High-Oil Parents.” *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 126 (2):483–95.
———. 2013b. “Genetic Control of Soybean Seed Oil: II. QTL and Genes That Increase Oil Concentration without Decreasing Protein or with Increased Seed Yield.” *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 126 (6):1677–87.
Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 44 (D1):D481–87.
Fiorani, Fabio, and Ulrich Schurr. 2013. “Future Scenarios for Plant Phenotyping.” *Annual*

Review of Plant Biology 64 (February):267–91.
Flintham, John E. 2000. “Different Genetic Components Control Coat-Imposed and Embryo-Imposeddormancy in Wheat.” *Seed Science Research* 10 (01):43–50.
Galperin, Michael Y., Daniel J. Rigden, and Xosé M. Fernández-Suárez. 2015. “The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection.” *Nucleic Acids Research* 43 (Database issue):D1–5.
“Gene RIF Website.” n.d. Accessed September 5, 2015. http://www.ncbi.nlm.nih.gov/gene/about-generif.
Gilchrist, Erin, and George Haughn. 2010. “Reverse Genetics Techniques: Engineering Loss and Gain of Gene Function in Plants.” *Briefings in Functional Genomics* 9 (2):103–10.
Glass, Kimberly, and Michelle Girvan. 2014. “Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets.” *Scientific Reports* 4 (February):4191.
Gnan, S., A. Priest, and P. X. Kover. 2014. “The Genetic Basis of Natural Variation in Seed Size and Seed Number and Their Trade-Off Using Arabidopsis Thaliana MAGIC Lines.” *Genetics* 198 (4):1751–58.
Gómez, John, Leyla J. García, Gustavo A. Salazar, Jose Villaveces, Swanand Gore, Alexander García, Maria J. Martín, et al. 2013. “BioJS: An Open Source JavaScript Framework for Biological Data Visualization.” *Bioinformatics* 29 (8):1103–4.
Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, et al. 2011. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40 (D1):D1178–86.
“GO Statistics.” n.d. Accessed November 22, 2017. http://geneontology.org/page/current-go-statistics.
Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo, and A. Conesa. 2008. “High-Throughput Functional Annotation and Data Mining with the Blast2GO Suite.” *Nucleic Acids Research* 36 (10):3420–35.
Hancock, John M. 2004. “SPARQL (SPARQL Protocol and RDF Query Language).” In *Dictionary of Bioinformatics and Computational Biology*.
Hanley, Steven J., and Angela Karp. 2014. “Genetic Strategies for Dissecting Complex Traits in Biomass Willows (Salix Spp.).” *Tree Physiology* 34 (11). Oxford University Press:1167–80.
Hassani-Pak, Keywan, Martin Castellote, Maria Esch, Matthew Hindle, Artem Lysenko, Jan Taubert, and Chris Rawlings. 2016. “Developing Integrated Crop Knowledge Networks to Advance Candidate Gene Discovery.” *Applied & Translational Genomics*, November. https://doi.org/10.1016/j.atg.2016.10.003.
Hassani-Pak, Keywan, Roxane Legaie, Catherine Canevet, Hugo A. van den Berg, Jonathan D. Moore, and Christopher J. Rawlings. 2010. “Enhancing Data Integration with Text Analysis to Find Proteins Implicated in Plant Stress Response.” *Journal of Integrative Bioinformatics* 7 (3). https://doi.org/10.2390/biecoll-jib-2010-121.
Hedden, Peter, and Yuji Kamiya. 1997. “GIBBERELLIN BIOSYNTHESIS: Enzymes, Genes and Their Regulation.” *Annual Review of Plant Physiology and Plant Molecular Biology* 48 (June):431–60.
Henry, Robert J. 2012. *Molecular Markers in Plants*. John Wiley & Sons.
Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, et al. 2016. “Ensembl Comparative Genomics Resources.” *Database: The Journal of Biological Databases and Curation* 2016 (February). https://doi.org/10.1093/database/bav096.
Higuchi, Mieko, Takeshi Yoshizumi, Tomoko Kuriyama, Hiroko Hara, Chika Akagi, Hiroaki Shimada, and Minami Matsui. 2009. “Simple Construction of Plant RNAi Vectors Using

Long Oligonucleotides." *Journal of Plant Research* 122 (4):477–82.
Hindle, Matthew M. 2012. "An Integrated Approach to Enhancing Functional Annotation of Sequences for Data Analysis of a Transcriptome." University of Nottingham.
Hirschhorn, Joel N., and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews. Genetics* 6 (2):95–108.
Horn, Fabian, Horn Fabian, Rittweger Martin, Taubert Jan, Lysenko Artem, Rawlings Christopher, and Guthke Reinhard. 2014. "Interactive Exploration of Integrated Biological Datasets Using Context-Sensitive Workflows." *Frontiers in Genetics* 5. https://doi.org/10.3389/fgene.2014.00021.
Horton, Matthew W., Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, et al. 2012. "Genome-Wide Patterns of Genetic Variation in Worldwide Arabidopsis Thaliana Accessions from the RegMap Panel." *Nature Genetics* 44 (2):212–16.
Huang, Xuehui, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, et al. 2010. "Genome-Wide Association Studies of 14 Agronomic Traits in Rice Landraces." *Nature Genetics* 42 (11):961–67.
Huang, Xueqing, Maria-João Paulo, Martin Boer, Sigi Effgen, Paul Keizer, Maarten Koornneef, and Fred A. van Eeuwijk. 2011. "Analysis of Natural Allelic Variation in Arabidopsis Using a Multiparent Recombinant Inbred Line Population." *Proceedings of the National Academy of Sciences of the United States of America* 108 (11):4488–93.
Huber, Wolfgang, Vincent J. Carey, Li Long, Seth Falcon, and Robert Gentleman. 2007. "Graphs in Molecular Biology." *BMC Bioinformatics* 8 Suppl 6 (September):S8.
Hu, Zhi-Liang, Eric Ryan Fritz, and James M. Reecy. 2007. "AnimalQTLdb: A Livestock QTL Database Tool Set for Positional QTL Information Mining and beyond." *Nucleic Acids Research* 35 (Database issue):D604–9.
Hu, Zhi-Liang, Carissa A. Park, and James M. Reecy. 2016. "Developmental Progress and Current Status of the Animal QTLdb." *Nucleic Acids Research* 44 (D1):D827–33.
Hwang, Ildoo, Hwang Ildoo, Sheen Jen, and Müller Bruno. 2012. "Cytokinin Signaling Networks." *Annual Review of Plant Biology* 63 (1):353–80.
"Index of /external2go." n.d. Accessed September 5, 2015. http://geneontology.org/external2go/.
Jaiswal, Pankaj, Doreen Ware, Junjian Ni, Kuan Chang, Wei Zhao, Steven Schmidt, Xiaokang Pan, et al. 2002. "Gramene: Development and Integration of Trait and Gene Ontologies for Rice." *Comparative and Functional Genomics* 3 (2):132–36.
Jenssen, Tor-Kristian, Jenssen Tor-Kristian, Lægreid Astrid, Komorowski Jan, and Hovig Eivind. 2001. "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression." *Nature Genetics* 28 (1):21–28.
Kang, Ning, Kang Ning, Singh Bharat, Afzal Zubair, Erik M. van Mulligen, and Jan A. Kors. 2013. "Using Rule-Based Natural Language Processing to Improve Disease Normalization in Biomedical Text." *Journal of the American Medical Informatics Association: JAMIA* 20 (5):876–81.
Kearsey, M. 1998. "The Principles of QTL Analysis (a Minimal Mathematics Approach)." *Journal of Experimental Botany* 49 (327):1619–23.
Kim, Sung, Kim Sung, Plagnol Vincent, Tina T. Hu, Toomajian Christopher, Richard M. Clark, Ossowski Stephan, Joseph R. Ecker, Weigel Detlef, and Nordborg Magnus. 2007. "Recombination and Linkage Disequilibrium in Arabidopsis Thaliana." *Nature Genetics* 39 (9):1151–55.
Kim, Yoo-Ah, Kim Yoo-Ah, and Teresa M. Przytycka. 2013. "Bridging the Gap between Genotype and Phenotype via Network Approaches." *Frontiers in Genetics* 3. https://doi.org/10.3389/fgene.2012.00227.
Kitano, Hiroaki. 2004. "Cancer as a Robust System: Implications for Anticancer Therapy."

Nature Reviews. Cancer 4 (3):227–35.
Kleinboelting, Nils, Gunnar Huep, Andreas Kloetgen, Prisca Viehoever, and Bernd Weisshaar. 2012. “GABI-Kat SimpleSearch: New Features of the Arabidopsis Thaliana T-DNA Mutant Database.” *Nucleic Acids Research* 40 (Database issue):D1211–15.
Köhler, Jacob, Jan Baumbach, Jan Taubert, Michael Specht, Andre Skusa, Alexander Rüegg, Chris Rawlings, Paul Verrier, and Stephan Philippi. 2006. “Graph-Based Analysis and Visualization of Experimental Results with ONDEX.” *Bioinformatics* 22 (11):1383–90.
Kong, Augustine, Daniel F. Gudbjartsson, Jesus Sainz, Gudrun M. Jonsdottir, Sigurjon A. Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, et al. 2002. “A High-Resolution Recombination Map of the Human Genome.” *Nature Genetics* 31 (3):241–47.
Koornneef, Maarten, Carlos Alonso-Blanco, and Dick Vreugdenhil. 2004. “Naturally Occurring Genetic Variation in Arabidopsis Thaliana.” *Annual Review of Plant Biology* 55:141–72.
Kover, Paula X., William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M. Ehrenreich, Michael D. Purugganan, Caroline Durrant, and Richard Mott. 2009. “A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis Thaliana.” *PLoS Genetics* 5 (7):e1000551.
Krallinger, Martin, Alfonso Valencia, and Lynette Hirschman. 2008. “Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology.” *Genome Biology* 9 Suppl 2 (September):S8.
Kristensen, David M., Yuri I. Wolf, Arcady R. Mushegian, and Eugene V. Koonin. 2011. “Computational Methods for Gene Orthology Inference.” *Briefings in Bioinformatics* 12 (5):379–91.
Kriventseva, Evgenia V., Fredrik Tegenfeldt, Tom J. Petty, Robert M. Waterhouse, Felipe A. Simão, Igor A. Pozdnyakov, Panagiotis Ioannidis, and Evgeny M. Zdobnov. 2015. “OrthoDB v8: Update of the Hierarchical Catalog of Orthologs and the Underlying Free Software.” *Nucleic Acids Research* 43 (Database issue):D250–56.
Lango Allen, Hana, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, et al. 2010. “Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height.” *Nature* 467 (7317):832–38.
Leitner, Florian, Leitner Florian, Krallinger Martin, and Alfonso Valencia. 2013. “BioCreative Meta-Server and Text-Mining Interoperability Standard.” In *Encyclopedia of Systems Biology*, 106–10.
Lesk, Victor, Jan Taubert, Chris Rawlings, Stuart Dunbar, and Stephen Muggleton. 2011. “WIBL: Workbench for Integrative Biological Learning.” *Journal of Integrative Bioinformatics* 8 (2):156.
Liekens, Anthony M. L., Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, and Jurgen Del-Favero. 2011. “BioGraph: Unsupervised Biomedical Knowledge Discovery via Automated Hypothesis Generation.” *Genome Biology* 12 (6):R57.
Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 26 (5):589–95.
Li, Hui, Li Hui, Peng Zhiyu, Yang Xiaohong, Wang Weidong, Fu Junjie, Wang Jianhua, et al. 2012. “Genome-Wide Association Study Dissects the Genetic Architecture of Oil Biosynthesis in Maize Kernels.” *Nature Genetics* 45 (1):43–50.
Lu, Z., and L. Hirschman. 2012. “Biocuration Workflows and Text Mining: Overview of the BioCreative 2012 Workshop Track II.” *Database* 2012 (0):bas043–bas043.
Lysenko, Artem. 2012. “Integration Strategies and Data Analysis Methods for Plant Systems Biology.” University of Nottingham.

Lysenko, Artem, Michael Defoin-Platel, Keywan Hassani-Pak, Jan Taubert, Charlie Hodgman, Christopher J. Rawlings, and Mansoor Saqi. 2011. “Assessing the Functional Coherence of Modules Found in Multiple-Evidence Networks from Arabidopsis.” *BMC Bioinformatics* 12 (May):203.

Lysenko, Artem, Martin Urban, Laura Bennett, Sophia Tsoka, Elzbieta Janowska-Sejda, Chris J. Rawlings, Kim E. Hammond-Kosack, and Mansoor Saqi. 2013. “Network-Based Data Integration for Selecting Candidate Virulence Associated Proteins in the Cereal Infecting Fungus Fusarium Graminearum.” *PloS One* 8 (7):e67926.

Makita, Yuko, Norio Kobayashi, Yoshiki Mochizuki, Yuko Yoshida, Satomi Asano, Naohiko Heida, Mrinalini Deshpande, et al. 2009. “PosMed-plus: An Intelligent Search Engine That Inferentially Integrates Cross-Species Information Resources for Molecular Breeding of Plants.” *Plant & Cell Physiology* 50 (7):1249–59.

Mao, Yuqing, Kimberly Van Auken, Donghui Li, Cecilia N. Arighi, Peter McQuilton, G. Thomas Hayman, Susan Tweedie, et al. 2014. “Overview of the Gene Ontology Task at BioCreative IV.” *Database: The Journal of Biological Databases and Curation* 2014 (August). https://doi.org/10.1093/database/bau086.

Martínez-Andújar, Cristina, M. Isabel Ordiz, Zhonglian Huang, Mariko Nonogaki, Roger N. Beachy, and Hiroyuki Nonogaki. 2011. “Induction of 9-Cis-Epoxycarotenoid Dioxygenase in Arabidopsis Thaliana Seeds Enhances Seed Dormancy.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (41):17225–29.

Mauricio, R. 2001. “Mapping Quantitative Trait Loci in Plants: Uses and Caveats for Evolutionary Biology.” *Nature Reviews. Genetics* 2 (5):370–81.

“MEDLINE®PubMed® XML Element Descriptions and Their Attributes.” 2005, December. U.S. National Library of Medicine. http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html.

Monaco, Marcela K., Joshua Stein, Sushma Naithani, Sharon Wei, Palitha Dharmawardhana, Sunita Kumari, Vindhya Amarasinghe, et al. 2014. “Gramene 2013: Comparative Plant Genomics Resources.” *Nucleic Acids Research* 42 (Database issue):D1193–99.

Moreau, Yves, and Léon-Charles Tranchevent. 2012. “Computational Tools for Prioritizing Candidate Genes: Boosting Disease Gene Discovery.” *Nature Reviews. Genetics* 13 (8):523–36.

Ni, Junjian, Anuradha Pujar, Ken Youens-Clark, Immanuel Yap, Pankaj Jaiswal, Isaak Tecle, Chih-Wei Tung, et al. 2009. “Gramene QTL Database: Development, Content and Applications.” *Database: The Journal of Biological Databases and Curation* 2009 (May):bap005.

Obayashi, Takeshi, Shinpei Hayashi, Motoshi Saeki, Hiroyuki Ohta, and Kengo Kinoshita. 2009. “ATTED-II Provides Coexpressed Gene Networks for Arabidopsis.” *Nucleic Acids Research* 37 (Database issue):D987–91.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 27 (1):29–34.

Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, et al. 2014. “The MIntAct Project--IntAct as a Common Curation Platform for 11 Molecular Interaction Databases.” *Nucleic Acids Research* 42 (Database issue):D358–63.

Osumi-Sutherland, David, Steven J. Marygold, Gillian H. Millburn, Peter A. McQuilton, Laura Ponting, Raymund Stefancsik, Kathleen Falls, Nicholas H. Brown, and Georgios V. Gkoutos. 2013. “The Drosophila Phenotype Ontology.” *Journal of Biomedical Semantics* 4 (1):30.

Peng, Zhi-Yu, Xin Zhou, Linchuan Li, Xiangchun Yu, Hongjiang Li, Zhiqiang Jiang, Guangyu

Cao, et al. 2009. “Arabidopsis Hormone Database: A Comprehensive Genetic and Phenotypic Information Database for Plant Hormone Research in Arabidopsis.” *Nucleic Acids Research* 37 (Database issue):D975–82.
Petryszak, Robert, Tony Burdett, Benedetto Fiorelli, Nuno A. Fonseca, Mar Gonzalez-Porta, Emma Hastings, Wolfgang Huber, et al. 2014. “Expression Atlas Update--a Database of Gene and Transcript Expression from Microarray- and Sequencing-Based Functional Genomics Experiments.” *Nucleic Acids Research* 42 (Database issue):D926–32.
Polderman, Tinca J. C., Benyamin Beben, Christiaan A. de Leeuw, Patrick F. Sullivan, Arjen van Bochoven, Peter M. Visscher, and Posthuma Danielle. 2015. “Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies.” *Nature Genetics* 47 (7):702–9.
Qian, Ming, Dong Wang, William E. Watkins, Val Gebski, Yu Qin Yan, Mu Li, and Zu Pei Chen. 2005. “The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China.” *Asia Pacific Journal of Clinical Nutrition* 14 (1):32–42.
Rae, Anne M., Nathaniel Robert Street, Kathryn Megan Robinson, Nicole Harris, and Gail Taylor. 2009. “Five QTL Hotspots for Yield in Short Rotation Coppice Bioenergy Poplar: The Poplar Biomass Loci.” *BMC Plant Biology* 9 (February):23.
Ramakrishnan, Raghu, Ramakrishnan Raghu, and Jeffrey D. Ullman. 1995. “A Survey of Deductive Database Systems.” *The Journal of Logic and Algebraic Programming* 23 (2):125–49.
Rebholz-Schuhmann, Dietrich, Anika Oellrich, and Robert Hoehndorf. 2012. “Text-Mining Solutions for Biomedical Research: Enabling Integrative Biology.” *Nature Reviews. Genetics* 13 (12):829–39.
Rigden, Daniel J., Xosé M. Fernández-Suárez, and Michael Y. Galperin. 2016. “The 2016 Database Issue of Nucleic Acids Research and an Updated Molecular Biology Database Collection.” *Nucleic Acids Research* 44 (D1):D1–6.
Roberts, Adam, and Lior Pachter. 2013. “Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments.” *Nature Methods* 10 (1):71–73.
Robertson, Stephen. 2004. “Understanding Inverse Document Frequency: On Theoretical Arguments for IDF.” *Journal of Documentation* 60 (5):503–20.
Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2009. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1):139–40.
Rothschild, Max F., Zhi-Liang Hu, and Zhihua Jiang. 2007. “Advances in QTL Mapping in Pigs.” *International Journal of Biological Sciences* 3 (3):192–97.
Salton, Gerard, and C. S. Yang. 1973. *On the Specification of Term Values in Automatic Indexing*. Journal of Documentation.
Schmitt, Thomas, David N. Messina, Fabian Schreiber, and Erik L. L. Sonnhammer. 2011. “Letter to the Editor: SeqXML and OrthoXML: Standards for Sequence and Orthology Information.” *Briefings in Bioinformatics* 12 (5):485–88.
Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3):379–423.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11):2498–2504.
Shan, Qiwei, Yanpeng Wang, Jun Li, and Caixia Gao. 2014. “Genome Editing in Rice and Wheat Using the CRISPR/Cas System.” *Nature Protocols* 9 (10):2395–2410.
Shi, Jiaqin, Ruiyuan Li, Dan Qiu, Congcong Jiang, Yan Long, Colin Morgan, Ian Bancroft, Jianyi Zhao, and Jinling Meng. 2009. “Unraveling the Complex Trait of Crop Yield with Quantitative Trait Loci Mapping in Brassica Napus.” *Genetics* 182 (3):851–61.

Smith, Richard N., Jelena Aleksic, Daniela Butano, Adrian Carr, Sergio Contrino, Fengyuan Hu, Mike Lyne, et al. 2012. “InterMine: A Flexible Data Warehouse System for the Integration and Analysis of Heterogeneous Biological Data.” *Bioinformatics* 28 (23):3163–65.

Smith, T. F., and M. S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1):195–97.

Sonah, Humira, Louise O’Donoughue, Elroy Cober, Istvan Rajcan, and François Belzile. 2015. “Identification of Loci Governing Eight Agronomic Traits Using a GBS-GWAS Approach and Validation by QTL Mapping in Soya Bean.” *Plant Biotechnology Journal* 13 (2):211–21.

Sparck Jones, Karen. 1972. “A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL.” *Journal of Documentation* 28 (1):11–21.

Spasić, Irena, Jacqueline Livsey, John A. Keane, and Goran Nenadić. 2014. “Text Mining of Cancer-Related Information: Review of Current Status and Future Directions.” *International Journal of Medical Informatics* 83 (9):605–23.

Splendiani, Andrea, Splendiani Andrea, Chris J. Rawlings, Kuo Shao-Chih, Stevens Robert, and Lord Phillip. 2012. “Lost in Translation: Data Integration Tools Meet the Semantic Web (Experiences from the Ondex Project).” In *Lecture Notes in Electrical Engineering*, 87–97.

Steinbach, Delphine, Michael Alaux, Joelle Amselem, Nathalie Choisne, Sophie Durand, Raphaël Flores, Aminah-Olivia Keliet, et al. 2013. “GnpIS: An Information System to Integrate Genetic and Genomic Data from Plants and Fungi.” *Database: The Journal of Biological Databases and Curation* 2013 (August):bat058.

Sun, Yizhou, and Jiawei Han. 2012. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, et al. 2010. “The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored.” *Nucleic Acids Research* 39 (Database):D561–68.

“TAIR Website.” n.d. Accessed September 5, 2015. ftp://ftp.arabidopsis.org/home/tair/User_Requests/Locus_Germplasm_Phenotype_20130122.

Tammen, Stephanie A., Simonetta Friso, and Sang-Woon Choi. 2013. “Epigenetics: The Link between Nature and Nurture.” *Molecular Aspects of Medicine* 34 (4):753–64.

Taubert, Jan. 2011. “ONDEX – A Data Integration Framework for the Life Sciences.” University of Bielefeld.

Taubert, Jan, Keywan Hassani-Pak, Nathalie Castells-Brooke, and Christopher J. Rawlings. 2014. “Ondex Web: Web-Based Visualization and Exploration of Heterogeneous Biological Networks.” *Bioinformatics* 30 (7):1034–35.

Taubert, Jan, Klaus Peter Sieren, Matthew M. Hindle, Berend Hoekman, Rainer Winnenburg, Stephan Philippi, Chris Rawlings, and Jacob Kohler. 2007. “The OXL Format for the Exchange of Integrated Datasets.” *Journal of Integrative Bioinformatics* 4(3):62. https://doi.org/10.2390/biecoll-jib-2007-62.

The Gene Ontology Consortium. 2014. “Gene Ontology Consortium: Going Forward.” *Nucleic Acids Research* 43 (D1):D1049–56.

“The Sequence Ontology - Resources - GFF3.” n.d. Accessed September 5, 2015. http://www.sequenceontology.org/gff3.shtml.

Trachana, Kalliopi, Kristoffer Forslund, Tomas Larsson, Sean Powell, Tobias Doerks, Christian von Mering, and Peer Bork. 2014. “A Phylogeny-Based Benchmarking Test for Orthology Inference Reveals the Limitations of Function-Based Validation.” *PloS One* 9 (11):e111122.

Travella, S. 2006. “RNA Interference-Based Gene Silencing as an Efficient Tool for

Functional Genomics in Hexaploid Bread Wheat." *Plant Physiology* 142 (1):6–20.
"UniProt Website." n.d. Accessed September 5, 2015. http://www.uniprot.org/help/disruption_phenotype.
Visscher, Peter M. 2008. "Sizing up Human Height Variation." *Nature Genetics* 40 (5):489–90.
Weigel, Detlef. 2012. "Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics." *Plant Physiology* 158 (1):2–22.
Weile, Jochen, Katherine James, Jennifer Hallinan, Simon J. Cockell, Phillip Lord, Anil Wipat, and Darren J. Wilkinson. 2012. "Bayesian Integration of Networks without Gold Standards." *Bioinformatics* 28 (11):1495–1500.
Weile, Jochen, M. Pocock, S. J. Cockell, P. Lord, J. M. Dewar, Holstein E.-M., D. Wilkinson, D. Lydall, J. Hallinan, and A. Wipat. 2011. "Customizable Views on Semantically Integrated Networks for Systems Biology." *Bioinformatics* 27 (9):1299–1306.
"Wheat Release Notes." n.d. Accessed September 5, 2015. https://ondex.rothamsted.ac.uk/QTLNetMinerWheat/html/release.html.
Willet, Cali E., and Claire M. Wade. 2014. "From the Phenotype to the Genotype via Bioinformatics." In *Methods in Molecular Biology*, 1–16.
Yates, Andrew, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, et al. 2016. "Ensembl 2016." *Nucleic Acids Research* 44 (D1):D710–16.
Yourshaw, Michael, S. Paige Taylor, Aliz R. Rao, Martín G. Martín, and Stanley F. Nelson. 2015. "Rich Annotation of DNA Sequencing Variants by Leveraging the Ensembl Variant Effect Predictor with Plugins." *Briefings in Bioinformatics* 16 (2):255–64.
Yu, Chih-Chieh, Yu Chih-Chieh, Furukawa Mari, Kobayashi Kazuhiro, Shikishima Chizuru, Cha Pei-Chieng, Sese Jun, et al. 2012. "Genome-Wide DNA Methylation and Gene Expression Analyses of Monozygotic Twins Discordant for Intelligence Levels." *PloS One* 7 (10):e47081.
Zhu, Guohui, Nenghui Ye, Jianchang Yang, Xinxiang Peng, and Jianhua Zhang. 2011. "Regulation of Expression of Starch Synthesis Genes by Ethylene and ABA in Relation to the Development of Rice Inferior and Superior Spikelets." *Journal of Experimental Botany* 62 (11):3907–16.
Zimin, Aleksey V., Arthur L. Delcher, Florea Liliana, David R. Kelley, Michael C. Schatz, Puiu Daniela, Hanrahan Finnian, et al. 2009. "A Whole-Genome Assembly of the Domestic Cow, Bos Taurus." *Genome Biology* 10 (4):R42.