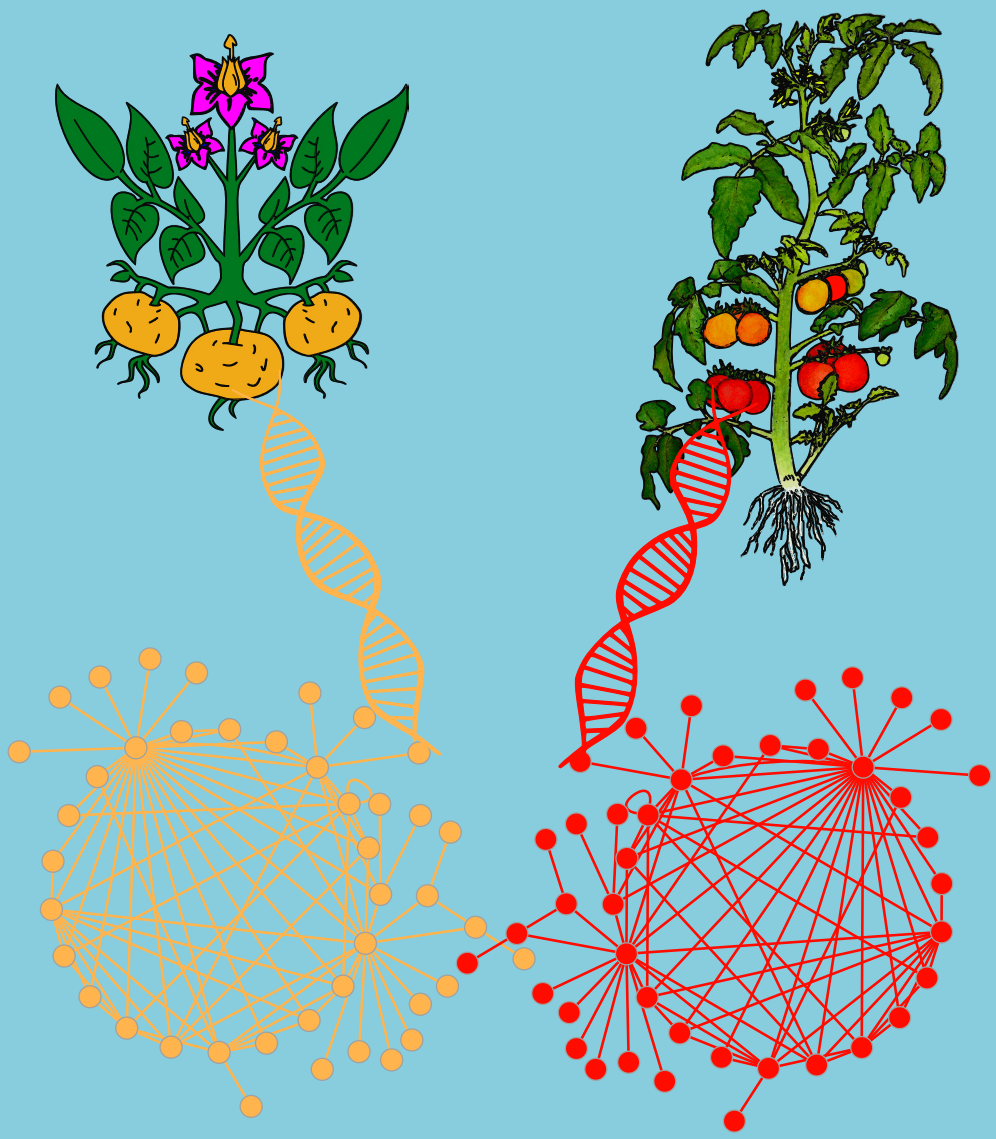
Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature



Gurnoor Singh

Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature

Gurnoor Singh

Thesis committee

Promotor
Prof. Dr. Richard G. F. Visser
Professor of Plant Breeding
Wageningen University & Research

Co-promotor
Dr. Christian W. B. Bachem
Assistant professor of Plant Breeding
Wageningen University & Research

Dr. Richard Finkers
Senior Scientist, Plant Breeding
Wageningen University & Research

Other members
Prof. Dr. V.A.P. Martins dos Santos, Wageningen University & Research
Dr. Willem Jan J. Knibbe, Wageningen University & Research
Prof. Dr. Jan L. Top, Wageningen University & Research
Prof. Dr. Dick de Ridder, Wageningen & Plant Research

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature

Gurnoor Singh

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 9 December 2019
at 4 p.m. in the Aula.

Gurnoor Singh

Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature
111 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summary in English

DOI: 10.18174/505685
ISBN: 978-94-6395-201-9

To my beloved mother,

Contents

CHAPTER 1	**General Introduction**	**5**
CHAPTER 2	**Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait**	**17**
CHAPTER 3	**QTLTableMiner^{++}: semantic mining of QTLtables in scientific articles**	**37**
CHAPTER 4	**Linked Data platform for Solanaceae species**	**53**
CHAPTER 5	**Prediction of candidate genes with QTL regions for tomato, using pbg-ld, functional annotations and evolutionary analysis**	**71**
CHAPTER 6	**General discussion and future prospects**	**85**
Supplementary results		**93**
References		**109**
Summary		**127**
Acknowledgements		**129**
Curriculum Vitae		**131**
Publication list		**132**
Education Statement		**133**

Chapter 1

General Introduction

With the rapid increase in the human population and food demands, there is a societal challenge to increase agricultural productivity. One approach to address this challenge is to breed new crop varieties that yield more even under unfavourable conditions like drought or extreme pathogen challenges. However, designing a breeding program is a laborious and time consuming effort that often lacks the capacity to generate new cultivars quickly in response to the required traits. Recent advances in biotechnology and genomics data science have the potential to accelerate and precise breeding programs [1].

The availability of annotated reference genome assemblies, for many crop species (including tomato [2], potato [3], rice [4], wheat [5], maize [6], brassica [7], and cucumber [8]) has enabled researchers and plant breeders to elucidate traits linkage at the level of genome annotations. Focusing more-and-more on mining genome annotations can help in identifying candidate genes that positively / negatively affect a trait that breeders aim to improve. Traditionally, plant breeders are introgressing chromosomal regions containing genes, positively affecting a trait of interest, after detection of so-called quantitative trait loci (QTL) into their elite breeding lines. However, a QTL region can easily contain 1000s of genes, encompassing genes that negatively influence the trait of interest or other pivotal traits. Breeding using the actual causative gene is, therefore, a much better approach.

One way to pursue this challenge is via analysis of gene expression or gene expression networks [9]. A second strategy used is comparative genomics approaches between crop species to infer candidate genes [10]. A third strategy utilizes the integration of large-scale genomic information stored in the scientific literature and molecular databases. These outlined "big data" strategies can be applied if data is made interoperable with the usage of semantic web technologies and text-mining. Thereby, providing an innovative eScience solution, which can speed-up the design of new cultivars in the future and provide novel insights into biological systems.

Scientific data: unstructured and structured data

Science relies only on data, whether it is to generate and test a hypothesis, or whether it is to verify a prediction through an experiment. After the revelation of the first map of a human genome, the fields of molecular biology and bioinformatics have generated an enormous amount of genomic data at an accelerating pace. [11]. This scientific data can occur in two major forms, unstructured data, available as text in scientific literature and patents, or structured data, available in molecular databases [12]. To obtain a complete overview of the relevant knowledge, scientists rely on these two major sources of information to conduct further research. Generic workflows of today's genome sciences research are composed by the use of analytical tools and bioinformatic pipelines which combines data/text mining technologies in a synergistic fashion [13].

However, to improve the usage of available data, there is a need to semantically collect, organize and integrate information from these two kinds of information resources.

However, both of them, have a significant difference in the structure and distribution of information. On the one hand, information which is published in biological databases is distributed over a multitude of independent databases, that may not be able to communicate with one another. On the other hand, information which is published as text in scientific literature is distributed over a multitude of independent text documents having varying representations. With the use of big data infrastructures, text mining and data mining techniques, an organized knowledge infrastructure can be formed that combines information from both these resources.

Unstructured data: Scientific literature

Scientific literature is a form of unstructured data, that accumulates up-to-date knowledge in any field of research. It plays a vital role in manually understanding the state-of-the-art in any area of interest. Despite the availability of various data repositories for plant research, a wealth of information is currently available only as (free) text in scientific publications. One of the most important tasks in a researcher's work and career is keeping up to date with the ever-increasing scientific literature, placing new outputs into context, and investigating the implications in their field. However, as the number of scientific publications is growing at an exponential rate, there is a need for using the power of machines, to automatically extract novel results and discoveries from literature. This section highlights the various resources where biomedical text corpora are made to be freely available.

MEDLINE and PubMed

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a premier bibliographic database of life sciences and biomedical information. This database is developed and maintained by the United States National Library of Medicine (NLM). Currently, it includes citations from over 5600 selected scholarly journals, which contribute to over 24 million references of articles published since 1966 till present [14]. These include bibliographic information for articles from academic journals that cover medicine, nursing, pharmacy, dentistry, veterinary medicine, plant sciences, genetics, genomics, metabolomics, proteomics, biotechnology, bioinformatics, and health care [15]. To make MEDLINE content more accessible for researchers, and the general public through the internet, NLM developed a searching system called PubMed. PubMed search service provides access to both MEDLINE and PreMEDLINE (new records that are in the process of being added to MEDLINE) records. The PubMed search service not only simplifies searching but it also links MEDLINE users to publisher's websites to retrieve the full text for the journal articles identified in the search.

The user interface (UI) with which users interact with PubMed is displayed in Figure 1.1. In general, a user queries PubMed or other similar systems whenever scientific articles in regards to particular information are in need. Offered a set of retrieved

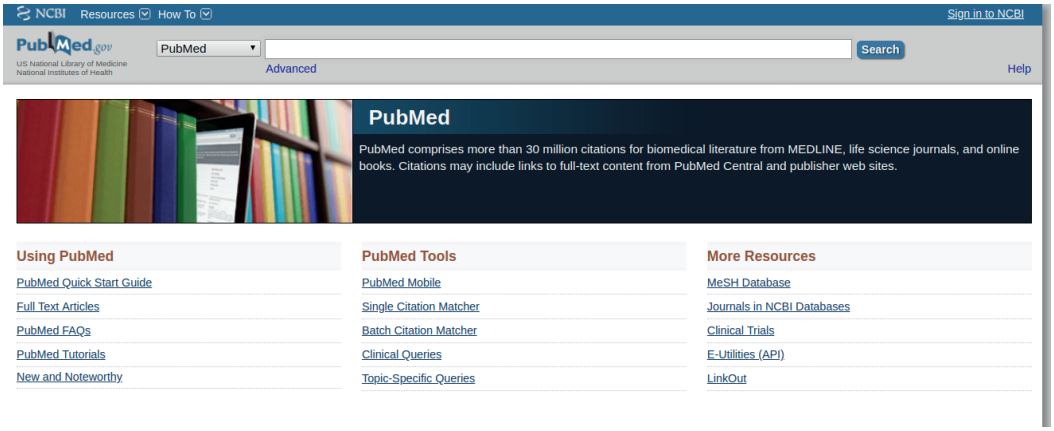


Figure 1.1: **PubMed homepage**

documents, the user can browse the result set and subsequently click to view abstracts or full-text articles, issue a new query, or abandon the current search. Particularly from a search perspective, PubMed takes as input natural language or free-text keywords and returns a list of citations that match input keywords ignoring the stop-words. Its search strategy has two major characteristics: first, by default, it adds boolean operators into user queries and uses automatic term mapping (ATM). Specifically, the boolean operator AND is inserted between multi-term user queries to require retrieved documents to contain all the user keywords. For example, if a user issued the query "Semantic search", the boolean operator AND would be automatically inserted between the two words as "Semantic AND search". Secondly, to increase the specificity of its searching techniques, PubMed automatically compares and maps keywords to pre-indexed MeSH terms (MeSH is a controlled vocabulary thesaurus of NLM) through its ATM process. Therefore, with the help of MeSH terms, PubMed does a precise subject related search, as these MeSH terminologies provide a consistent way to retrieve information that may use different terminology for the same concepts [16].

Other scientific text resources

Europe PubMed Central (Europe PMC) is a document repository providing full-text open access to scientific articles. Europe PMC does contain over 3 million full-text articles [17]. This repository also provides full text articles in PDF files as well as in XML format that complies with the Journal Article Tag Suite (JATS) schema. JATS is commonly used by publishers and archives to exchange journal meta-data content.

Google Scholar is also used frequently as a scientific text resource. Google Scholars can not only recover peer-reviewed articles, but also other scholarly texts, such as MSc/PhD theses, books, and pre-print repositories. A comparative study [18] suggests that Google

Scholar often returns larger retrieval sets, but a substantial number are link-outs to PubMed records. Google Scholar allows sorting of the articles according to the time of their release but does not provide other advanced search functions offered by PubMed and many other specialized biomedical systems [19].

Patents

The scientific text also exists in the form of patents. A patent is a set of exclusive official rights given to an inventor/inventors to protect and claim their inventory knowledge. Patents contain protected knowledge of a specific domain in condensed information related to the claimed knowledge. Therefore, patents are a valuable resource for text mining [20]. Patents in life sciences encompass innovation in gene sequencing, candidate gene discovery, diagnostic tests, or therapeutic delivery systems. For example, many pharmaceutical and seed companies have patented a group of genes, proteins, and metabolites that are of commercial importance [21]. Many patent offices provide web based patent retrieval systems. The Open Patent Services (OPS) are web based patent retrieval web services of the European Patent Office (EPO) [22]. OPS web services were firstly published in 2006 and have been revised several times since then. OPS services are available on the EPO website. Google has also created search functionalities for about 7 million patents from the United States Patent and Trademark Office (USPTO). Google Patents provides the search over patents in PDF format [23]. A search can be taken out by providing keywords for a full text search, giving the patent number, entering keywords for a title, inventor, assignee, U.S. classification, international classification, selecting the document status being an issued patent or an application, selecting the patent type, or by restricting to a period of time. FreePatentOnline is also a famous IR system for patents [24].

Natural Language Processing

Natural Language Processing (NLP) or text mining is a field of artificial intelligence that focuses on enabling the machine to understand and analyze (unstructured) data in the form of text [25]. This technology helps researchers to analyze, explore and manage free text in large text repositories as well as gain new insights from it.

There are three fundamental aspects to NLP: information retrieval (IR), information processing, and information extraction (IE). IR refers to the recovery of documents from a text repository based on a search system that takes the user's query as an input. Information processing is the major aspect of text mining, in which text is tokenized, classified and semantically annotated (based on ontologies and controlled vocabularies) for identification. This way it is easier for a computer to make sense out of the text. Finally, IE is the extraction of ideas and concepts from a text. Table 1.1 highlights the major basic components for processing textual information with a generic NLP workflow.

Table 1.1: **A list of major components in a NLP workflow**

NLP Components	Description
Tokenization	Tokenization is the process of chopping words into pieces, called tokens. The tokens are usually words in a sentence.
Part of Speech Tagging	Labelling of words as subject-predicate-object based on the role a word plays in the sentence. It reduces ambiguities.
Noun phrase chunking	Focuses on the identification of basic structural relations between groups of words.
Named Entity Recognition	Identification and classification of entities in the text with the help a defined dictionary.
Syntactic analysis	Analyse the component of a sentence i.e. establishes the connection between different parts of each sentence. This is done in the simplest case through co-occurrence and statistical analysis or with different syntactic parsing methods.

Named Entity Recognition

Named Entity Recognition (NER) is the methodology to identify named entities in the text. NER is a prerequisite for most of NLP based applications [26]. NER consists of three different problems, firstly the recognition of an entity in text, secondly the assignment of a class to this entity (gene, protein, metabolites, traits, etc), and finally the selection of a preferred term for naming the object in case their synonyms exist. The latter is especially important if the recognized entities are to be combined with information from other resources, such as databases.

There are different ways of identifying NER such as:

- Ontology/Controlled vocabulary based NER: In information science, a controlled vocabulary (CV) is a list of terms that have been enumerated explicitly. Each term in a CV has an unambiguous, non-redundant definition and is not connected to each other. On the other hand, an ontology is a controlled vocabulary expressed in an ontology representation language. This language has a grammar for using vocabulary terms to express relationships, similarities & differences in between terms.

  Overall, domain-specific ontologies and CV represent current knowledge of a domain and therefore, can be used to annotate entities in text. Ontologies and CV provides a definite mapping and identity to the named entities.

- Rule-based NER: Regular expressions and logical interactions are used to identify

entities. This allows the identification of more complex variants of terms than in dictionary-based NER. Rule-based NER increases the intricacy of the mapping of the terms.

- Supervised Machine learning methods like conditional random fields (CRF) can also be used for finding NER in a given data set [27]. A set of text with identified entities is taken as 'training set' to train an algorithm which is further applied onto a larger data set.

Information Extraction

It is well quoted in [28] that "information extraction (IE) forms the basis for text mining the same way as NER forms the basis for information extraction". IE is used in the identification of explicit entities in a given text and further to extract information related to these entities in the same text [29]. In biological literature, IE generally concentrates on finding information about genes, proteins, metabolites, traits, diseases and drugs, and relationships between these entities.

Structured data: Databases in plant genomics

Massive amount of plant genomics research data is modeled in databases, as structured data. The popular databases in this field can be classified into two main categories, large-scale public repositories or community-specific / project-specific databases [30]. Large-scale public repositories are usually developed and maintained by government agencies or international consortia. Most of the large-scale public repository contains data from all life sciences and is not limited to plant-sciences. On the other hand community-specific or project-specific databases focus on a particular model organism for example *Arabidopsis thaliana*, rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*) or maize (*Zea mays*)). Table 1.2 enlist the most popular databases in both categories.

Semantic Web Technologies

Due to the advancement in sequence technologies, the amount of omics data in public data repositories is growing exponentially [32]. Moreover, this data is heterogeneous and distributed in multiple resources that may not allow interoperability. Semantic web technologies allow harmonized aggregation of heterogeneous data with a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.[42] The Semantic Web is therefore regarded as an integrator across different content, information applications and systems. With the usage of semantic web technologies, we can integrate and query, the exponential volume of available omics data with a top-down approach, for gaining insights into the molecular mechanism.

Table 1.2: **A list of popular databases in plant sciences**

Large-scale public repositories

Database Name	Description	Reference
GenBank	An international nucleotide sequence repository developed and maintained by the National Center for Biotechnology Institute (NCBI).	[31]
The European Nucleotide Archive (ENA)	An international nucleotide sequence repository developed and maintained by EMBL-EBI Europe.	[32]
Genomic Expression Archive (GEA)	A genomic expression data archive containing functional genomics data, developed and maintained by DDBJ Japan.	[33]
The Universal Protein Resource (UniProt)	A collection of protein sequences, protein annotation and proteomes of various species (including plants). IS maintained by EMBL-EBI Europe SIB Switzerland, and the Protein Information Resource (PIR).	[34]
Ensembl Plants	An integrated repository containing plants genomics data i.e genome sequences, gene models, functional annotation, and polymorphic loci.	[35]
Plant Genome DataBase Japan (PGDBj)	An intergrative repository containing information of plant genomic data from numerous databases and literature.	[36]
ELIXIR Core Data Resources	The European data resources (including ENA, Ensembl Plants, UniProt, PRIDE, and PDB).	[37]

Community-specific / project-specific databases

Database Name	Description	Reference
Sol GenomicsNetwork (SGN)	A central repository genomic, genetic, phenotypic and taxonomic information for members of the Solanaceae family	[38]
The Arabidopsis Information Resource (TAIR)	A resource containing molecular, biological and genetical data of *Arbidopsis thaliana*.	[39]
Maize Genetics and Genomics Database (MaizeGDB)	A central repository for maize sequence, stock, phenotype, genotypic and karyotypic variation, and chromosomal mapping data.	[40]
Rice (*Oryza sativa*) genome annotation database (Osa1)	Contains structural and functional annotation of the model species rice(*Oryza sativa*)	[41]

Sir Tim Berners-Lee, the creator of the most successful means to global information sharing i.e. the World Wide Web (WWW) network, first mentioned about semantic web technologies in the year 2001 [43]. The idea behind semantic web technologies aimed at transforming the internet from a network of human-readable web pages into a semantic web of interlinked data which is machine-readable. Similar to the internet, semantic web information space is characterized by Uniform Resource Identifiers (URIs), resources, protocols (HTTP, SMTP, FTP, etc.), and data formats (RDF). Here, a Uniform Resource Identifier (URI) is a sequence of characters used to identify an entity or physical resource. HTTP, SMTP, FTP is a protocol (request/response) standard to transfer information over a computer network.

RDF: The Resource Description Framework (RDF) is a default framework to represent semantic web data [44]. Building blocks of RDF frameworks our knowledge triples, which form a graph of linked data. Each RDF triple consists of a subject, linked with an object, via a predicate. A subject and a predicate are always uniquely identifiable with a definitive URI, while an object can be another subject with a URI or a literal. Every subject represents an entity such as a person, a place, a data file, biological entities like genes/proteins/drugs/. A literal is a data value that can either be numerical, text or a timestamp.

The collection of RDF statements describing data is called the RDF graph. The collection of RDF graphs is called the RDF dataset. RDF graphs can be defined in several formats: for example Extensible Markup Language for RDF (RDF/XML), Terse RDF Triple Language (TURTLE), Notation 3 (N3), JSON-LD, etc.

SPARQL: Simple Protocol and RDF Query Language (SPARQL) is a programming language used to extract data from RDF graphs. SPARQL queries are sent from a client to a service known as a SPARQL endpoint, using the data access protocols (HTTP protocol).

Ontologies: One of the key elements of modeling data in RDF graphs is the use of identifiers to define terms and relationships between them. For semantic web technologies, ontologies were idealized to specify not only the definition of a controlled set of terms but also their relations with each other in a single domain. In the field of life sciences, a major advance in data interoperability has occurred with the growing use of ontologies to unambiguously identify and describe biological concepts. Ontology terms are used to annotate entities such as genes, proteins, metabolites, drugs, traits, QTLs, experiments, etc. in a consistent way. An ontology often contains vocabulary of terms, their definitions, associated synonyms, and a set of semantic relationships between terms. These relationships provide interoperability and the ability to combine annotations that have been applied at different levels of specificity (in heterogeneous data). Ontologies are used in semantic web for both, annotate data entities under consideration, and define the metadata of the data.

Metadata: Metadata is information that describes the data, i.e. provenance of data, the measurements/methods used to retrieve this data, etc. Meta-data is used to make the correct interpretation of data.

FAIR Data Principles

In the domain of life sciences, scientists recognize that the findability, accessibility, interoperability, and reusability of data by interlinking different datasets from various resources is important to give a complete view of biological activities. To achieve this efficiently, data needs to be organized according to be FAIR data principles [45]. Briefly, according to these principles every data element should have a unique persistent identifier, with a searchable metadata ("Findable"); these identifiers should resolve to (meta)data using an open standard protocol ("Accessible"); the (meta)data should use a representation language that utilizes widely accepted domain-specific ontologies ("Interoperable"); and finally, the data should be well described with cross-references and with available license information ("Reusable"). FAIR data principles can be used to facilitate development of a global landscape for integrating genomics datasets for better predictions and reproducible analysis [46].

Research objective

The main objective of this PhD project was to improve the integration of genomic data for knowledge discovery in *Solanacae* Species, using genome annotations available in molecular databases and scientific literature. This genomic data integration can be used in the identification of candidate gene(s) for (crop) QTL regions, via utilizing available knowledge of genome annotations from literature and molecular databases. Hence, this research is an asset for those (plant) breeding companies that aim to effectively improve their current cultivars, because breeding can be more precise (e.g. the gene candidate for the trait, or the regulator of the gene) of interest. Although the primary target of our research is for improving plant breeding, our research can also be relevant to breeding in general, as well as more fundamental research fields, where genetics is utilized to uncover the mechanisms of traits of interest.

Solanaceae Species

Solanaceae are members of flowering plants (Eudicots) that have high economic importance. Members of the family are cultivated for their edible fruits and tubers, for example, tomatoes (*Solanum lycopersicum*), pepper (*Capsicum spp*), eggplant (*Solanum melongena*), and potato (*Solanum tuberosum*); remedial properties, for example tobacco (*Nicotiana tabacum*) and mandrake (*Mandragora officinarum*); or ornamental flowers such as petunia (*Petunia x hybrida*).

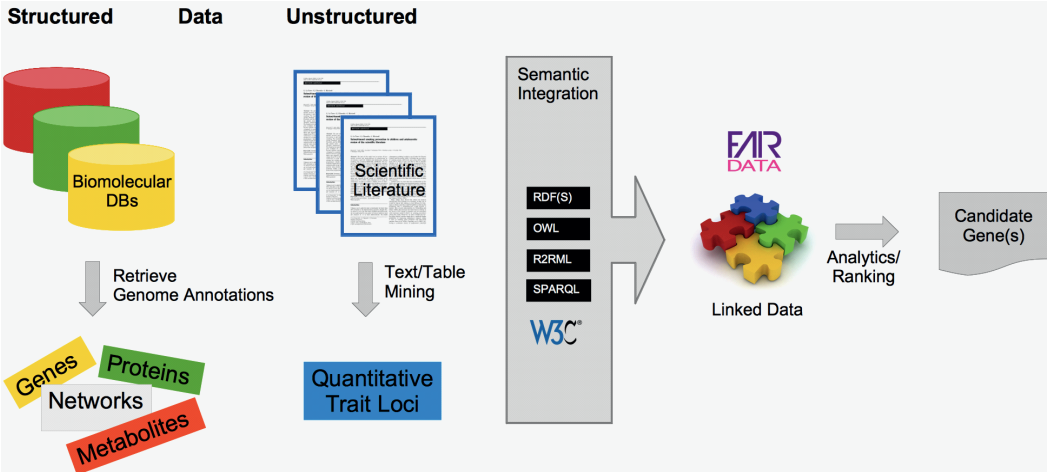


Figure 1.2: **Prototype architecture for the integration of genomic data to assist knowledge discovery using genome annotations. [47]**

Potato (*Solanum tuberosum L.*) is one of the most important staple crops for human nutrition. Underground stems called stolons under suitable environmental conditions form potatoes in the form of tubers. In addition to its culinary versatility, potato is a cost-effective product and plays a major role in meeting the ever-increasing food demands of the world. Its tubers are a good source of starch, proteins, vitamin C, folate, and carotenoids [48]. Different potato genotypes produce tubers of different properties, like shape, size, color, starch content, and nutritional value.

Tomato (*Solanum lycopersicum*) is one of the most consumed fruits in the world, as well as the second most consumed edible products of the Solanaceae species. Tomatoes are a globally important dietary source of lycopene, beta-carotene, vitamin C, and fiber. In addition to its agricultural value and due to its diploid genetics and inbreeding potential, tomato is a widely used model species for fundamental research on subjects including fruit development and pathogen response. [49].

Road Map

This thesis is compiled of 6 chapters. Chapter 2 introduces a supervised Natural language processing (NLP) model, developed using IBM Watson, to extract knowledge networks containing genotypic-phenotypic associations of potato tuber flesh color from the scientific literature. Chapter 3 illustrates QTLTableMiner^{++} (QTM), a table mining tool that extracts and semantically annotates QTL information buried in (heterogeneous) tables

of plant science literature. QTM was further used to extract QTL information from QTL studies of tomatoes and potatoes. Chapter 4 presents the development of a linked-data platform called Solanace linked data platform (pbg-ld), which provides semantically integrated genotypic and phenotypic data on Solanaceous species. This platform combines both unstructured data from scientific literature and structured data from publicly available biological databases using the Linked Data approach. Chapter 5 describes a seamlessly integrative workflow for the prediction of candidate genes within QTL regions using our linked-data platform pbg-ld, data about functional annotations and evolutionary analysis. Finally, Chapter 6 highlights the general discussion and future prospects of this research.

Chapter 2

Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait

Gurnoor Singh [1, *], Evangelia A. Papoutsoglou [1, *], Frederique Keijts-Lalleman [2], Bilyana Vencheva [2], MarkRice [2], Richard G.F. Visser [1], Christian W.B. Bachem [1], and Richard Finkers [1]

[1] Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands
[2] IBM Netherlands, Amsterdam, the Netherlands
[*] Both authors contributed equally to this manuscript.

to be submitted

Abstract

Introduction

A wealth of scientific information is available only as (free) text in scientific publications. As opposed to biological databases, text in literature is a source of unstructured information. Hence, textual information remains difficult for machines to process and analyze information from it. Natural language processing (NLP) or text mining, is a field of artificial intelligence that focuses on the power of a machine to understand and analyze text. NLP can render textual information to be computationally accessible; as well as support information extraction and lead to knowledge network construction.

Methodology

In this pilot study, we have developed a supervised NLP model using IBM Watson, to extract knowledge networks containing genotypic-phenotypic associations of potato tuber flesh color from the scientific literature. Initially, we used Watson Knowledge Studio (WKS) to develop a domain-specific NLP model for finding biological entities (genes, proteins, metabolites, and traits) that relate to tuber flesh color and relationships among them. We manually annotated a training corpus of 34 full-text scientific papers, indicating all instances of such biological entities and their relationships in each sentence. WKS uses the manual annotations, related dictionaries, and a model type system to generate a supervised NLP model capable of extracting knowledge networks for tuber flesh color. Subsequently, we assessed our NLP model by deploying it on a larger test corpus, containing about 4000 PubMed abstracts related to the Solanaceae taxon, published from the years 2000 to 2016.

Results

The resulting knowledge networks contained both previously known as well as contemporaneously unknown leads to subsequently discovered biological phenomena relating to the flesh color trait. Such leads included the link between potato tuber flesh color and the zeaxanthin epoxidase (ZEP) enzyme/gene that already became a 2nd order relation in 2007 and a 1st order relation to tuber flesh color in 2010. However, the relation between tuber flesh color and ZEP was experimentally substantiated later in the year 2011. The results illustrate that, indeed, supervised NLP shows a lot of promise to speed up knowledge discovery, data integration and hypothesis generation in scientific research.

Keywords: NLP, plant science literature, IBM Watson, text mining, relationship extraction, knowledge networks

Introduction

Scientific publications accumulate knowledge and developments in any field of research. One of the most important tasks in a researcher's work and career is keeping up to date with the ever-increasing scientific literature, placing new outputs into context, and investigating the implications in their field. However, as the number of scientific publications are growing at an exponential rate, there is a need for using the power of machines, to automatically extract novel results and discoveries from literature.

Potato (*Solanum tuberosum L.*) is one of the most important staple crops for human nutrition. In addition to its culinary versatility, potato is a cost-effective product and plays a major role in meeting the ever-increasing food demands of the world. Its tubers are a good source of starch, proteins, vitamin C, folate, and carotenoids [48]. Different potato genotypes produce tubers of different properties, like shape, size, color, starch content, and nutritional value.

One of the most extensively studied traits in potato is tuber flesh color. Carotenoids are considered to be the primary determinant of tuber flesh color [50]. Carotenoids play essential roles in photosynthesis, while in non-photosynthetic tissues, they exert a broad range of functions acting as pigments, antioxidants, and precursors of signaling molecules, including volatiles [51]. Previous studies have shown that β-carotene and zeaxanthin are the prominent components that determine potato tuber flesh color. In recent years, several candidate genes like BCH / CHY2 and ZEP have been found to relate to the tuber flesh color. BCH / CHY2 are the genes related to the production of β-carotene while ZEP is considered responsible for the accumulation of zeaxanthin [52]. Scientific evidence for the association of tuber flesh color with genetic and molecular entities is mentioned in the scientific literature or biological databases. For example Acharjee et al. previously published networks of experimentally found biological entities that related to tuber flesh color in the years 2011 and 2016 [52], [53]. In this research, we try to automate the process of extracting knowledge of molecular entities (genes/proteins/metabolites) that are influencing changes in tuber flesh color from scientific publications.

Compared to structured information (as in databases), textual information is huge, noisy, and redundant. Therefore, to understand and analyze textual information in a meaningful way, there is a need to use artificial intelligence to establish a machine read, extracting and analysis knowledge in the form of textual information. Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling the machine to understand and analyze (unstructured) data in the form of text [25]. Despite the availability of various data repositories for plant research, a wealth of information currently remains hidden within the scientific literature. Hence, there is continuous growth, in scope and importance, of information extraction via NLP for scientific literature. NLP can render these texts to be computationally accessible; as well as support information extraction, knowledge network (KN) construction, and hypothesis generation.

In the past years, many NLP based research studies have been conducted on molecular

biological literature [54], [55]. On the one hand, NLP studies focused majorly on rule-based named entity recognition (NER) i.e. identifying and annotating biological entities such as genes or proteins [56], [57], metabolites [58], [59], traits [60], QTLs [61], diseases [62], and drugs [63] in literature. On the other hand, a few NLP studies pay attention to extracting associations (relationships and event) between these biological entities, while using NER systems under the hood [57], [64]. Automated approaches to mining knowledge concerning the association of an entity to its phenotypes are required to further advance in the field of precision breeding [65]. Rule-based NLP is more widely used in mining knowledge from biological context than machine learning based NLP [66], [67]. However, construction and regularization of rules is a complex task in rule-based NLP. Often the rule-based NLP user tends to do overfitting of the rules in the training set, which affects performance in the test set. Dictionaries and ontologies are used as building blocks in rule-based NLP. In comparison to rule-based NLP, supervised NLP methodology can also be used. In supervised NLP, a domain specialist annotates the training set of documents manually. These manually annotated documents supported by dictionaries and ontologies are used by the algorithm to produce context-specific rules. Finally, these rules are used to perform NLP on the unannotated test set.

IBM Watson is a trailblazer in applying NLP and machine learning solutions to mine knowledge from huge corpora of texts available online [68], [69]. Watson Knowledge Studio is a cloud-based application to train an NLP model based on the context and linguistic nuances of a specific literature domain. In addition to annotating entities of interest in a given text (named entity recognition), Watson is also able to perform relationship extraction; that is, to label the connections between the detected entities of interest.

In this research, we used the Watson platform to develop a model based on domain-relevant literature to find biological entities (genes, proteins, metabolites, and traits) and relationships that relate to tuber flesh color, an agronomically important trait. Later, we tested our model by deploying it on a larger testing corpus, containing selected PubMed abstracts that are related to the Solanaceae family, published from the years 2000 to 2016. We shaped the Watson outputs into knowledge networks (KNs) and, over this year range, tracked the closeness of our trait of interest to relevant entities, marking the time points where significant developments were made.

This proof of concept (although limited in size) is an example of how literature mining, enabled by tools like Watson, could help plant scientists obtain a clearer “big picture” about specific (narrow) areas in their field of expertise. Facts easily missable in the expanding sea of literature could come to light, be automatically organized into KNs, and ultimately accelerate research in a process with little human intervention.

Methodology

Experimental corpora

To make a supervised NLP model, we assembled the scientific articles into 2 corpora of the training set and the test set. The training set consisted of open source full-text articles, while the test set was built from PubMed abstracts.

The training corpus is a collection of 34 full-text scientific articles (see supplementary table 6.1) which focus on tuber flesh color and known biological entities like metabolites and proteins involved in the carotenoid pathway, for example, beta-carotene hydroxylase and zeaxanthin epoxidase [53]. The training set was manually annotated with Watson Knowledge Studio (WKS). WKS uses these manual annotations to generate a supervised NLP model that can capture phenotypic tuber traits and the associating genes, proteins, and metabolites. Later, we assessed the capabilities of this supervised NLP model to make KN on this training set as well as on a larger test set.

The test set consists of 4023 abstracts from PubMed from the years 2000 to 2016. These abstracts are plant genetics-based articles that focus on 4 major Solanaceous crops (tomato, potato, eggplant, and capsicum). To limit the scope of the NLP model to find direct genomic associations related to tuber flesh color, no pathogen related articles were included in the test set. Our developed NLP model is capable of extracting KN for the tuber flesh color trait. However, the articles in the test sets talk about a variety of different topics in plant genetics and do not limit their scope to only the tuber flesh color trait. This test set challenges the NLP model to a more real world application, as opposed to a restricted use case in our training set.

Additionally, to assess the information content per section from articles we divided the training set into subsets based on the section they were coming from. The test set of abstracts have also been divided into subsets based on their year of publication. This was done to study the evolution of knowledge over time.

Watson Knowledge Studio and Watson Explorer

IBM's Watson™Knowledge Studio (WKS) is a proprietary text mining solution that makes machine learning models to interpret linguistic nuances, meaning, and relationships specific to a domain. It provides a platform of user-friendly tools for the manual annotation of domain-specific literature. Further, it uses manual annotations to create an annotator, which is a custom machine learning model that understands the language of the domain.

WKS was used to create a supervised NLP based model tailored to annotate of potato-specific scientific literature [70]. The annotator received input from a human (i.e. domain expert), about what type of content should be captured from the text. Based on the requirements, a type system is formulated to guide both humans and machines. WKS's type system is a prime component that defines rules of annotation and identifies the subjects of interest to the domain-specific user. The type system establishes how

content can be captured by defining the types of entities that can be labeled, and how relationships between different pairs of entities can be marked. Figure 2.1 illustrates the type system used by us to capture a set of traits, associated genes, proteins, and metabolites in potato-specific literature.

An entity is a blueprint of a set of objects that belong to a specific type, serving as a class or category. For our type system, which is used specifically to identify genomic relationships in potato-specific literature, three types of entities have been defined i.e. firstly Gene/Protein, secondly Trait, and thirdly Metabolite. Here, genes and proteins are clubbed as a single entity in our model. This is due to the fact that generally, no added insight is provided by knowing whether the text refers to the gene or the protein/enzyme that it encodes.

If there is a relationship between entities present in the text, it is also possible to define and capture this relationship in the type system of a model annotator. Relationships are directional in WKS (so a relation from A to B is not the same as a relation from B to A). There were 7 kinds of relationships defined in our type system. Remaining consistent with the above entities and the reasons for their selection, the relations are similarly simple and all-encompassing in nature, which is why many of the relations have the "related to" label. The exceptions ("encodes", "part of") were included because the high number of these relation mentions in the corpus allowed for WKS to produce models that could also identify them in the text.

In WKS, each entity is supported by an entity-specific dictionary. Dictionaries are used to do a pre-annotation step of, automatic entity recognition, before the actual textual corpus is manually annotated by a human. Therefore, in order to not capture a lot of noise, all dictionaries are made small and are limited to the scope of molecular entities previously known to be associated with tuber flesh color or carotenoid pathway. Preferred names in these dictionaries are selected from a known molecular database or ontology. To elaborate, the Gene/Protein dictionary contains 183 genes/proteins from the carotenoid biosynthesis pathways. Similarly, the Metabolites dictionary is made by using 85 metabolites from the same pathway. While terms in the Trait dictionary consist of 56 potato-related trait terms taken from the Solanaceae Phenotype Ontology [71].

The Watson Explorer is an analytical software platform that uses an NLP model generated from manually annotated scientific articles (by WKS) as an input. Watson Explorer's outputs are text documents in XML/CAS files, containing annotations of the entities and their relationships that have been extracted, as well as their documents (and document position) of origin. We use these XML/CAS files to build KN.

Modeling decisions

To train our NLP model to capture KN of only genotypic-phenotypic entities and their relationships, the type system underwent a number of major changes and revisions in an iterative process. With trial-and-error optimization, entities and relationships were introduced as well as discarded, based on how well the knowledge is captured and

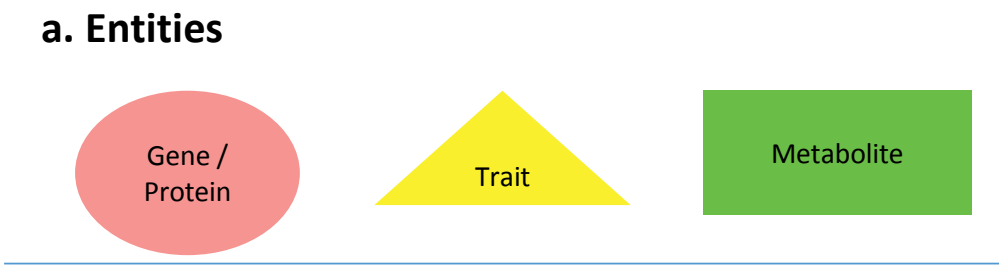


Figure 2.1: **Watson Knowledge Studio (WKS) configurations of the type system for a customized NLP annotator.** a) 3 types of entities in the type system. b) 7 types of relationships defined in the type system of an annotator.

presented in the KN.

Some modeling decisions important to be mentioned are presented below.

- Biological entities that were tested but not included in the final model:
  - biochemical processes
  - metabolic pathways
  - trait values
  - organism names, species names, and genotypes

  While these biological entities occur in the text and contain sources of knowledge to understand the biological mechanisms involved in the phenotypes, the number of mentions in the text were insufficient for WKS to adequately train a model. We therefore choose not to include these entities in the type system of our NLP

model. Further, including these entities in our model shifts the focus away from the research question of mining genotypic-phenotypic relationships in text.

- Combination of genes and proteins to a single entity:

  Initially, we kept genes and proteins as two separate entities. However, during manual annotation difficulties were encountered in distinguishing between the two, as they are frequently used interchangeably in the text. Furthermore, for subject matter experts, there is little information to be lost by combining them. Hence, in our type system, genes and proteins are a single entity.

- Annotation rule for metabolites (specific metabolite mentions vs generic mentions)

  Metabolites are included in scientific literature in different forms. Their mentions may consist of specific composite terms (e.g. petunidin-3-p-coumaroyl-rutinoside-5-glucoside) to more generic ones (e.g. carotenoids). According to our type system, we annotated all forms of metabolite mentions as in this way we can capture both knowledge triples with specific entities as well as knowledge triples with generic entities. In our analysis, a knowledge triple is defined as a data structure consisting of two entities and a label for their underlying relationship.

- Annotation rules for genes

  As is the case with metabolites, genes may be introduced in different formats. To name a few, sometimes the full name is present (zeaxanthin epoxidase), sometimes the short form of it (ZEP), and other times there is a species indicator as a prefix (LeZEP). We chose to annotate all these cases to train the model.

Building and visualization of knowledge networks (KNs)

For the construction of a KN only entities with relationships were used. The mention of an entity by itself, with no connections, was not included in the KN. With the help of a Python script, we filtered out data of entities and relationships data from XML/CAS files. This script captured relationships as knowledge triples in easily parsable CSV files containing the relationship ID, relationship type, original mention of each entity, entity label, entity type, document in which this sentence occurred, sentence position and position of the source and target nodes.

As various entities appear with a variety of spellings in the corpus (e.g. β-carotene, b-carotene, beta-carotene), we also included a normalization step, attributing an additional preferred label to each entity. This was done manually on the list of individual entities that had been extracted. In the normalization process we first converted all spellings of entities and relationships to American English uppercase characters. Additionally, Prefixes relating to species were removed from gene names, for example, the StAN1 referring to anthocyanin 1 in *Solanum tuberosum* (potato), was converted to AN1. Similarly, suffixes indicating individual members of gene families were also removed, for example BCH1

and BCH2 (both referring to forms of beta-carotene hydroxylase), were converted to beta-carotene hydroxylase.

For metabolites, EC number references were converted to the full name of an enzyme. Further, apostrophes and # notations were also removed, e.g. flavonoid-3',5'-hydroxylase becomes flavonoid-3,5-hydroxylase, 9#-cis-neoxanthin becomes 9-cis-neoxanthin. Lastly, all abbreviations were expanded to the long form, for example, NCED2 into 9-cis-epoxycarotenoid dioxygenase. While the above steps reduce the specificity a particular entity (for example we labeled BCH1 and BCH2 as BCH), as is always the case with tokenization, this simplification boosts network connectivity, despite the loss of information.

Finally, Cytoscape [72] version 3.7.1 was used to visualize these KNs. Cytoscape can plot KNs using CSV files as input.

Results

To confirm that our domain-specific NLP model performed as intended, and extracted KN with the focus on tuber flesh color from scientific literature, we deployed this tool on 2 different corpora, i.e. the training set with full-text articles and the test set with PubMed abstracts only.

Case 1: Analysis of training corpus (Full-text articles)

Watson retrieved a KN with a total of 293 nodes with 551 unique edges from the training set of 34 articles. Out of these 293 nodes, there are a total of 159 genes/proteins, 112 metabolites and 22 traits (Figure 2.2). Carotenoids (entity of the type metabolites) was the primary centroid of this network having 76 first-order neighbors. In order to evaluate the nodes and connections of this KN, we analyzed the overall structure based on the currently known experimental knowledge of tuber flesh color. Our KN contains nodes and edges, which show scientifically credible links between nodes and the trait of interest; tuber flesh color. Most genes/proteins and metabolite entries in this network are part of the carotenoid biosynthesis pathway, which includes beta-carotene biosynthesis, xanthophyll cycle, abscisic acid biosynthesis, lutein biosynthesis, etc.

The trait under study, tuber flesh color, has 38 first-order neighbors, comprising 11 genes/proteins and 27 metabolites. These genes/proteins and metabolites are also enlisted in Table 2.1. Previously conducted research studies have found that ZEP, BCH/CHY are associated with white, yellow and orange flesh color. AN1, a gene responsible for production of anthocyanin, is associated to purple flesh color. All these genes occur as direct neighbors of tuber flesh color in our network.

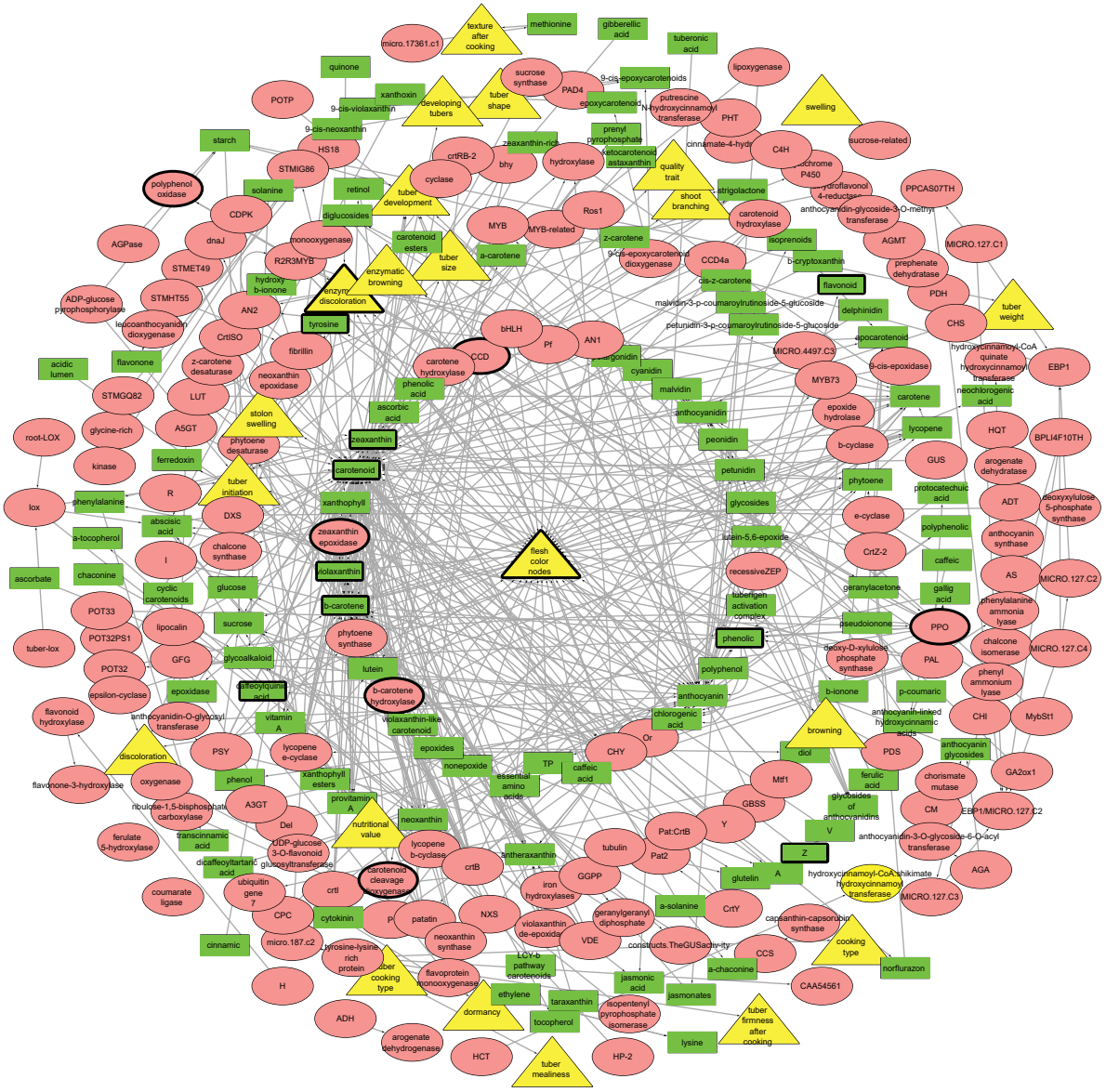


Figure 2.2: **A KN representing knowledge triples found in the training set of 34 full articles.** Yellow nodes refer to a trait entity, red nodes refer to gene/protein entities, and green nodes represent the metabolite entities. The centroid of this network is tuber flesh color. Nodes with bold outlines indicate that these entities have an experimentally proved association with tuber flesh color (trait of interest). This experimental evidence of these entities with tuber flesh color is reported in the articles [52], [53].

Watson's NLP model retrieved the total number of entities in the training set with a precision of 97.65%, a recall of 88.91% and an F1 score of 93.07% against the manually annotated training corpus. Supplementary table 6.2 represents a confusion matrix showing the total number of entities per document, number of true positives (TP), number of false negatives (FN) and number of false positives (FP). Precision and recall were calculated as TP / (TP + FP) and TP / (TP + FN) respectively.

Additionally, to compare the difference in volume and quality of information extracted from abstracts vs full-text representation of an article, our NLP model was applied separately on only the abstracts of the training corpus.

The comparison between abstract vs full-text representation of an article highlights a quantitative difference between 2 representations of a scientific article (abstract-only vs full text). We hypothesized that the abstract would concretely and concisely present the core outputs of a publication, whereas the introduction section would mainly recapitulate established theories and relevant biological connections but without contributing new knowledge. Finally, the results and discussion sections would combine, in greater detail, the significant contributions of the article, and at the same time make further suggestions for future experimentation. We found supporting evidence for this hypothesis, as the abstract-only network still includes the entities experimentally shown to be most important for tuber flesh color. In sets A and B, Table 2.1 lists the direct neighbors of tuber flesh color node in the KNs of full text representation (Figure 2.2) and abstracts only (Figure 2.3).

The difference between these two sets (Table 2.1; SET A - SET B) is also shown. These 20 entities occur as direct neighbors of flesh color in full-text KN, but do not occur as direct neighbors in the abstract-only KN. Of these 20 entities, 6 entities (AN1, lutein, lutein-5,6-epoxide, polyphenol, phytoene synthase, violaxanthin) are still present in the KN of abstracts (Figure 2.1), even though they are not direct neighbors of the tuber flesh color node. These entities are second-order neighbors of tuber flesh color and first-order neighbors of carotenoids, BCH, CHY or ZEP. Furthermore, recessive ZEP is also represented in the abstract-only KN. Since the recessive allelic variant of ZEP is similar to dominant ZEP form, these nodes are not represented as separate entities or indeed other details of gene/protein characteristics, such as chemical isomers and trait measures. The remaining 12 entities (nonepoxide, peonidin, anthocyanidin, petunidin, pelargonidin, cyanidin, pf, malvidin, epoxides, glycosides) are not represented in the abstract-only KN. These entities are associated with key metabolites causing changes in flesh color, however, they do not influence the trait directly. Hence, our results illustrate that the most important nodes in the full-text network are still present in the reduced network from the abstracts.

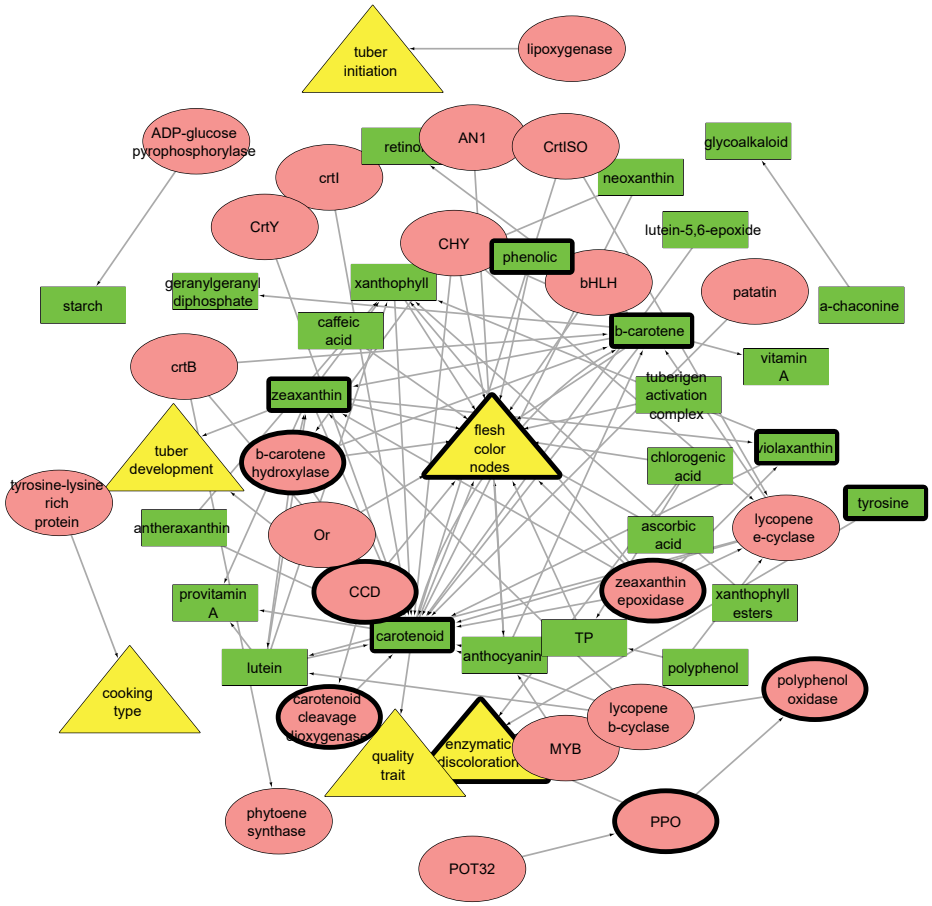


Figure 2.3: **A KN representing knowledge triples found in the training set of 34 articles, abstracts only**. Yellow nodes refer to a trait entity, red nodes refer to gene/protein entities, and green nodes represent the metabolite entities. The centroid of this network is tuber flesh color. Nodes with bold outlines indicate that these entities have an experimentally proved association with tuber flesh color (trait of interest). This experimental evidence of these entities with tuber flesh color is reported in the articles [52], [53].

Table 2.1: **Sets representing first order (direct) neighbors of flesh color nodes.** Set A represents first-order neighbors of tuber flesh color nodes found in full-text articles. Set B represents first-order neighbors of tuber flesh color nodes found in abstracts of articles of the training set. The difference between these sets (SET A - SET B) represents all entities that are first-order neighbors of tuber flesh color in full-text articles, but they not in abstracts alone. Certain entities are bold to indicate that these nodes have an experimentally proved association with tuber flesh color (trait of interest). Metabolites are marked in green, and genes/proteins in red. This experimental evidence of these entities with tuber flesh color is reported in the articles [52], [53].

Set A	**Set B**	**Set A - Set B**
AN1	anthocyanin	AN1
anthocyanidin	ascorbic acid	anthocyanidin
anthocyanin	b-carotene	carotene hydroxylase
ascorbic acid	b-carotene hydroxylase	cyanidin
b-carotene	bHLH	epoxides
b-carotene hydroxylase	caffeic acid	essential amino acids
bHLH	carotenoid	glycosides
caffeic acid	CCD	lutein
carotene hydroxylase	chlorogenic acid	lutein-5,6-epoxide
carotenoid	CHY	malvidin
CCD	Or	nonepoxide
chlorogenic acid	phenolic	pelargonidin
CHY	TP	peonidin
cyanidin	tuberigen activation complex	petunidin
epoxides	xanthophyll	Pf
essential amino acids	zeaxanthin	phenolic acid
glycosides	zeaxanthin epoxidase	phytoene synthase
lutein		polyphenol
lutein-5,6-epoxide		recessiveZEP
malvidin		violaxanthin
nonepoxide		violaxanthin-like carotenoid
Or		
pelargonidin		
peonidin		
petunidin		
Pf		
phenolic		
phenolic acid		
phytoene synthase		
polyphenol		
recessiveZEP		
TP		
tuberigen activation complex		
violaxanthin		
violaxanthin-like carotenoid		
xanthophyll		
zeaxanthin		
zeaxanthin epoxidase		

Case 2: Analysis of testing corpus (PubMed abstracts)

To assess how the NLP model performed over an unknown corpus, we deployed this model on a testing corpus of 4023 abstracts from PubMed articles. From this testing corpus, Watson retrieved a KN with a total of 681 nodes and 976 unique edges (Figure 2.4a). Carotenoid (entity of the type metabolites) was again the primary centroid of this network, with 107 first-order neighbors. Our trait under study, tuber flesh color, has 21 first-order neighbors, comprising 9 genes / proteins and 12 metabolites.

While our model is tailored toward potato tuber flesh color (range between white to orange), additional traits and their respective biological associations were detected as well. For example, the KN from the test set also detected those genes/proteins and metabolites which influence other traits such as enzymatic discoloration, tuber initiation, tuber development, tuber maturation, cooking types, stolon swelling, flower development, etc (Figure 2.4b). This illustrates that the information content extends beyond the specific use case. Moreover, Watson uses our NLP model to extract information influencing tuber flesh color in a wider context than the use case only, without the requiring further specific training.

Identifying emerging candidate with time analysis

To assess the accumulation of knowledge over time, the abstracts of the test set were organized in subsets ordered chronologically (i.e. by the date of their publication). Starting from the year 2000 and incrementing yearly (i.e. all publications up to 2000, all publications up to 2001, . . . , all publications up to 2016), subsets were formed. Each of these subsets was used to construct a separate KN. A network of a given year is always a subset of a KN from the following years and a superset of the previous years.

To study the development of entity connections with regard to our trait of interest (tuber flesh color), we worked backwards. The most recent collection was the most complete, so the nodes widely concerning tuber flesh color were chosen (color, flesh, flesh color, flesh trait, orange flesh color, tuber color, tuber flesh, tuber flesh color, white flesh color, yellow-orange flesh color) and are henceforth referred to as flesh color nodes. We focused our attention on the nodes that eventually ended up directly connected to a flesh color node. Then, we tracked the distance of these selected nodes to each individual flesh color node, and the changes over time. Supplementary table 6.3 shows an example of such a table for changes occurring between 2009 and 2010. Scripts were finally written to parse the collections for all years in the corpus. Based on these year-by-year summaries, a master summary table was made (Table 2.2).

Table 2.2 shows that the literature already contained significant indications as to the relevance of specific genes that were found to be important for potato flesh color [52]. Most prominently, both beta-carotene hydroxylase (BCH) and zeaxanthin epoxidase (ZEP) were in close proximity (2nd order neighbors) from 2007 and made the transition to direct neighbors of flesh color nodes in 2010, before experimental evidence was published in

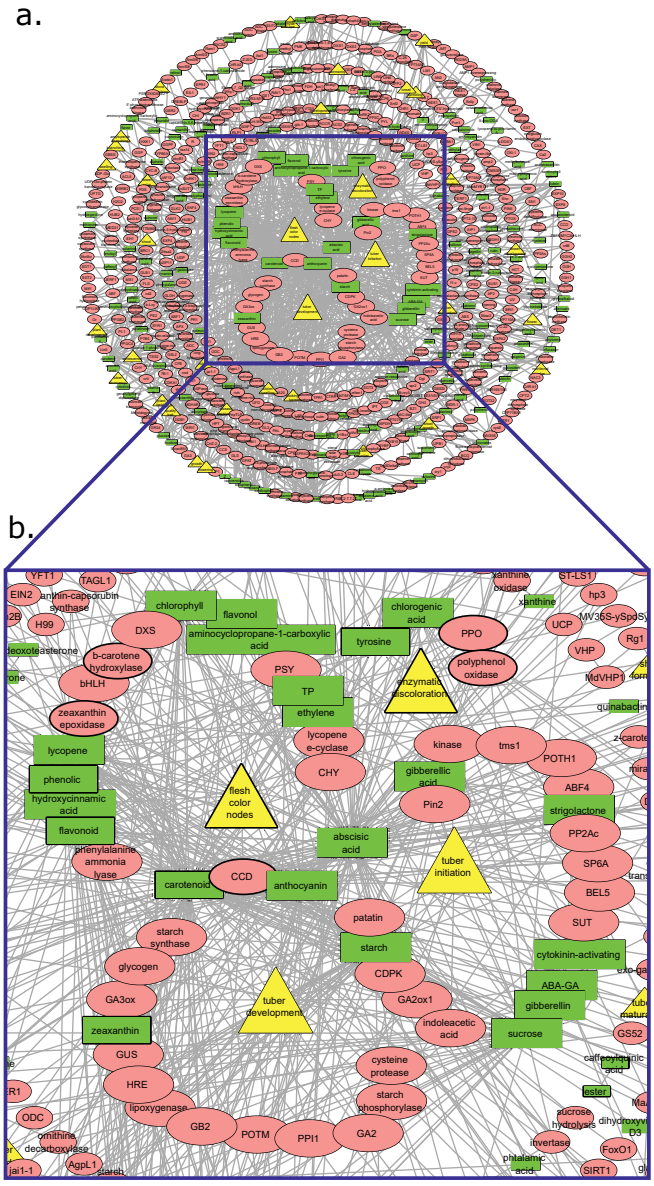


Figure 2.4: **A KN representing knowledge triples found in the test set of 4023 PubMed articles** Yellow nodes refer to a trait entity, red nodes refer to gene entities, and green nodes represent the metabolite entities. a) complete zoomed-out knowledge network b) zoomed-in snapshot of the knowledge network focusing on tuber flesh color and additional traits with their respective biological associations.

2011. We conclude that having such information available, can provide key indications of scientifically relevant links prior to such links having experimentally been substantiated or published.

Equally, false positives such as lycopene, a metabolite not found in potato tubers, arises in the KN as a first-order neighbor. While for most domain experts it is clear that lycopene is the compound responsible for flesh color in tomato, and therefore trivial to eliminate from the knowledge network as a significant player, it does reinforce the requirement for domain specialists to apply their knowledge to these results.

Table 2.2: **Overview of yearly changes in the network**, based on the individual year summaries (for example supplementary table 6.3). Each column represents a year, with an eventual neighbor of flesh color nodes listed in each row. The distances are the shortest path, at the time, from the node indicated to any flesh color node.

node year	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
CCD	x	x	x	x	3	3	3	3	3	3	1	1	1	1	1	1	1
CHY	x	x	x	x	x	x	x	2	2	2	1	1	1	1	1	1	1
DXS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PSY	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1
TP	x	x	x	x	x	x	x	x	x	3	3	2	2	2	2	2	1
abscisic acid	x	x	4	2	2	2	2	2	2	1	1	1	1	1	1	1	1
aminocyclopropane-1-carboxylic acid	x	x	x	x	x	5	5	5	5	1	1	1	1	1	1	1	1
anthocyanin	x	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1
beta-carotene hydroxylase	x	x	4	4	4	4	4	2	2	2	1	1	1	1	1	1	1
bHLH	x	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1
carotenoid	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
chlorophyll	x	x	x	x	x	3	3	3	1	1	1	1	1	1	1	1	1
ethylene	x	x	x	x	x	5	5	5	5	3	1	1	1	1	1	1	1
flavonoid	x	x	x	1	1	1	1	1	1	1	1	1	1	1	1	1	1
flavonol	x	x	x	x	x	x	x	x	x	x	x	x	x	x	1	1	1
hydroxycinnamic acid	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1	1
lycopene	3	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1
lycopene e-cyclase	x	x	x	x	x	x	2	2	2	2	1	1	1	1	1	1	1
phenolic	x	x	x	x	3	3	2	2	2	2	2	2	2	2	2	2	1
phenylalanine ammonia lyase	x	x	x	x	x	x	x	x	x	x	x	x	3	3	3	3	1
zeaxanthin epoxidase	x	x	3	3	3	3	3	2	2	2	1	1	1	1	1	1	1

Based on the above, particular attention has been given to BCH and ZEP, and their transitions between 2006-2007 and 2009-2010.

Discussion

This work served as a pilot to study the benefits of using NLP platforms like Watson, for performing knowledge discovery over plant science literature. With the exponential increase in the amount of scholarly publications and the sheer volume of available biological literature, researchers are finding it increasingly difficult to keep up-to-date with all information relevant to their field. Assembling knowledge from available literature in a single network could be used to generate new hypotheses, or aid researchers in assembling a better overall picture about the components surrounding their area of interest. However, unlike a human research expert, it is more challenging for a machine to comprehend biological insights from complicated sentences and text structures of scientific literature. Every NLP model has a limited scope of research questions it can address. The developed type system of our NLP model cannot capture and reflect all biological complexities in KNs. However, our developed NLP model is intended to only mine genotypic-phenotypic information from scientific literature into KNs, so that this knowledge can be structured-data, easily readable by both machines and humans.

Further, only generic relationships ("is related to") of association between these entities were captured, the degree of association between 2 entities (positive, negative, inexplicit) was ignored in our model. The performance of our model, nevertheless, is satisfactory for the pilot study and addresses the above stated research objective. In order to optimize the efficiency for the process of manual annotation of the training set, we restricted ourselves to a limited training corpus of 34 full-text articles. Although the training we provided to Watson was limited, it was still sufficient to enable our model to extract similar knowledge from the test set, which was a collection of documents referring to different crops, traits and processes.

While making the testing corpus for our NLP model, we including literature from other Solanaceae crop species (tomatoes, capsicum, eggplant) as well. Mining and assembling information from all of these different literature resources into a single KN was a controversial decision too. Many genes and metabolites are involved in a similar bio-mechanism across these crops species. However, in some cases that may introduce noise, whereas in others it may be a source of ideas. There is a certain tradeoff to be observed here: the wider the scope of the processed documents, the higher the margin for noise, but also the potential. The premise for this trial, after all, was that newly published research in a broad domain of science would indiscriminately be funneled into an NLP model, to produce networks that can assist humans.

A similar balance exists, when it comes to the parts of documents that are used for text analysis. Abstracts are an easily accessible and summarized form of significant information from an article. However, it is important to note that different journals

prescribe different formats for their abstracts and other sections of scientific articles they publish. Therefore, the quality of minable information mentioned in an abstract also depends on the journal as well as type of article. Abstracts of articles such as reviews, scientific methods, or articles that cover wide scope of certain topic, might not provide comprehensive minable scientific leads. For example, Nature contributions may not always formally describe all scientific leads in their abstract, results are more frequently mentioned in the main text. It is worth mentioning that there were instances where the NLP approach failed to meet expectations. In cases where biological entities were abbreviated, or associations between 2 entities was mentioned in more than 1 sentence, our Watson's NLP model could not predict these entities and relationships. Watson type-system includes facilities to co-refer abbreviated entries or pronoun to its original form, however, due to less number of instances in our training corpus, Watson's NLP model was not able to capture these entities and relations. However, Watson is not unique in this respect. In fact, most NLP tools used in biological text capable of recognizing relations in text suffer from the same flaw. Biological abbreviations are haphazard. Many times two biological concepts have a same abbreviation. For example. an abbreviation MIC might mean minimal inhibitory concentration or MIC gene which is a Major histocompatibility complex (MHC) class I chain related (MIC) gene. Training on a bigger corpus might increase accuracy to predict the right entity mentions.

Overall, Watson produced a model that powered the construction and time analysis of meaningful KNs under restricted-effort conditions. Therefore we believe that a more intensive effort would yield excellent results, and could play an important role in bringing together diverse information from large literature corpora and hypothesis generation. Presently our resulting KN contain unweighted edges. In the future, we would like to enhance this by having an edge network. Edges can have a weighted-score which is based on experimental knowledge from databases and number of times a particular relationship occurred in text and how biological relevant a relationship is. This way, text mining can be used to compare established knowledge and emerging knowledge.

Conclusions

The results of our analysis strongly indicate that NLP methods, such as those provided by IBM Watson, can be deployed on plant science literature as a powerful tool for the construction of networks that distill and integrate knowledge to facilitate future research.

Chapter 3

QTLTableMiner^{++}: semantic mining of QTLtables in scientific articles

Gurnoor Singh [1], Arnold Kuzniar [2], Erik van Mulligen [3], Anand Gavai [2], Christian W. B. Bachem [1], Richard G.F. Visser1 [1], and Richard Finkers [1]

[1] Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands
[2] Netherlands eScience Center (NLeSC), Amsterdam, the Netherlands
[3] Department of Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands

Published: 25 May 2018 in *BMC Bioinformatics*
DOI: 10.1186/s12859-018-2165-7

Abstract

Introduction

A quantitative trait locus (QTL) is a genomic region that correlates with a phenotype. Most of the experimental information about QTL mapping studies is described in tables of scientific publications. Traditional text mining techniques aim to extract information from unstructured text rather than from tables. We present QTLTableMiner^{++} (QTM), a table mining tool that extracts and semantically annotates QTL information buried in (heterogeneous) tables of plant science literature.

Methodology

QTM is a command line tool written in the Java programming language. This tool takes scientific articles from the Europe PMC repository as input, extracts QTL tables using keyword matching and ontology-based concept identification. The tables are further normalized using rules derived from table properties such as captions, column headers and table footers. Furthermore, table columns are classified into three categories namely column descriptors, properties and values based on column headers and data types of cell entries. Abbreviations found in the tables are expanded using the Schwartz and Hearst algorithm. Finally, the content of QTL tables is semantically enriched with domain-specific ontologies (e.g. Crop Ontology, Plant Ontology and Trait Ontology) using the Apache Solr search platform and the results are stored in a relational database and a text file.

Results

The performance of the QTM tool was assessed by precision and recall based on the information retrieved from two manually annotated corpora of open access articles, i.e. QTL mapping studies in tomato (*Solanum lycopersicum*) and in potato (*S. tuberosum*). In summary, QTM detected QTL statements in tomato with 74.53% precision and 92.56% recall and in potato with 82.82% precision and 98.94% recall. Hence, QTM is a unique tool that aids in providing QTL information in machine-readable and semantically interoperable formats.

Keywords: Quantitative trait locus, QTL, Plant breeding, Table mining, Ontologies, Semantic interoperability

Background

Modern genetic analysis in crop plants aims to understand the contribution of individual genes and loci in the development of agronomic traits. Quantitative variation results from the combined action of multiple genes and environmental factors. With the help of molecular marker studies, it is possible to detect genomic regions that are statistically associated with variation in non-Mendelian phenotypic traits, also termed as quantitative trait loci (QTL) [73].

Detecting QTLs can help in the development of precision breeding programs. However, elucidating QTL regions for genes that are causative to a trait of interest is a tedious and time-consuming process because a single QTL region commonly entails hundreds of genes, including those that might negatively influence the trait [74]. Leveraging upon knowledge available in both scientific literature and molecular biology databases can help in narrowing down the QTL regions to candidate genes associated with traits of interest.

QTL studies have widely been published in scientific articles, in particular in tables or supplementary materials. However, there is no established repository where experimental data on plant-specific QTL studies can be submitted. In the past, there have been several attempts to create manually curated databases with QTL information; for example, AnimalQTLdb [75], MaizeGDB [76], Gramene QTL database [77] and SGN/solQTL [78]. Manual curation of such database systems is a laborious task. There is a need to retrieve QTL data from publications efficiently, which can further reduce the cost of QTL database curation and QTL knowledge discovery process.

Using tables is the most common way to represent (semi-)structured data (e.g. results of QTL mapping experiments) on the web or in the scientific literature [79]. As QTL information is mostly published in tables rather than in the main text of articles, traditional text-mining techniques are not suited for this task [80]. There are several challenges associated with table-mining. The information in a table can be easily interpreted by human but not by a machine. For example, when parsing an article in text, HTML or PDF formats, it is difficult for a machine to determine which cells are part of a header and which cells contain data. Moreover, tables can have different orientations (horizontal *versus* vertical layout). Furthermore, tables can have nested structure including rows/columns with multiple headers.

Several commercial and open source table-mining tools have been developed including Tabula [81], Google Tables [82] [83], TableMiner$^+$ [84] and the domain-specific QTLMiner [80]. While Tabula and QTLMiner extract tables from PDF documents, Google Tables and TableMiner$^+$ process web pages. TableMiner$^+$ makes use of contextual information, for example, in table captions, footers and column headers, to improve the identification of relevant tables in web pages. In contrast, the Google's system does not use author-defined table properties, such as column headers, captions and footers, but rather assigns class-labels to columns using a machine-learning approach combined with maximum likelihood estimation over web-derived knowledgebase. QTLMiner [80] was the first tool focused on mining QTLs from tables of plant science literature. Briefly, QTLMiner first converts

articles in PDF to HTML documents, identifies trait-related tables, extracts relevant data and finally stores the results in a relational database. QTLMiner lacks wider applicability as its performance to extract information from tables of a literature, depends mainly on the conversion of articles from PDF to HTML file, which is done by commercially available web service from BCL [85]. Secondly, QTLMiner could only extract QTL statements only when a table in literature occur in a particular format and lacks the capability to mine this information from heterogeneous tables.

Current tools that extract tabulated information from PDF or HTML documents have difficulty with parsing tables correctly because table structures are (semantically) not described using these formats. Although, scientific articles are distributed in PDF format, it is inconvenient to use these PDF documents for automated information extraction as they lack machine readability and a logical structure specifying which content constitutes a paragraph, table, figure, header or footer etc. Therefore, even if massive amounts of unstructured data are held in the form of PDF documents, automated extraction of tables, figures or other structured information can be very difficult. Similarly, HTML file represents a layout of a web page and is not focused on describing data. Therefore, our tool uses XML files as they represent information in a logical structure that is machine-readable.

QTLTableMiner^{++} (QTM) is a Java-based command-line tool that extracts and semantically annotates QTL information from tables of scientific articles. QTM takes articles in a syntactically interoperable format, XML, as an input. The Europe PMC [17] repository provides full-text open access articles in the XML format that complies to the Journal Article Tag Suite (JATS) schema. JATS is commonly used by publishers and archives to exchange journal content.

QTM filters (candidate) trait tables (i.e. those with phenotypic information) out of all tables in an article. In these tables, a QTL statement refers to a relationship between pheno- and genotypic entities. QTM extracts QTL statements and semantically annotates the biological entities in these statements with domain-specific ontologies using the Apache Solr search platform [86]. Finally, QTM outputs the results both in a relational database and in a text file (CSV). In summary, QTM is a unique tool that aids in providing QTL information in machine-readable and semantically interoperable formats.

Implementation

Table mining workflow

Figure 3.1 illustrates the overall workflow implemented by the QTM tool. This workflow consists of three parts, which are described in more details below.

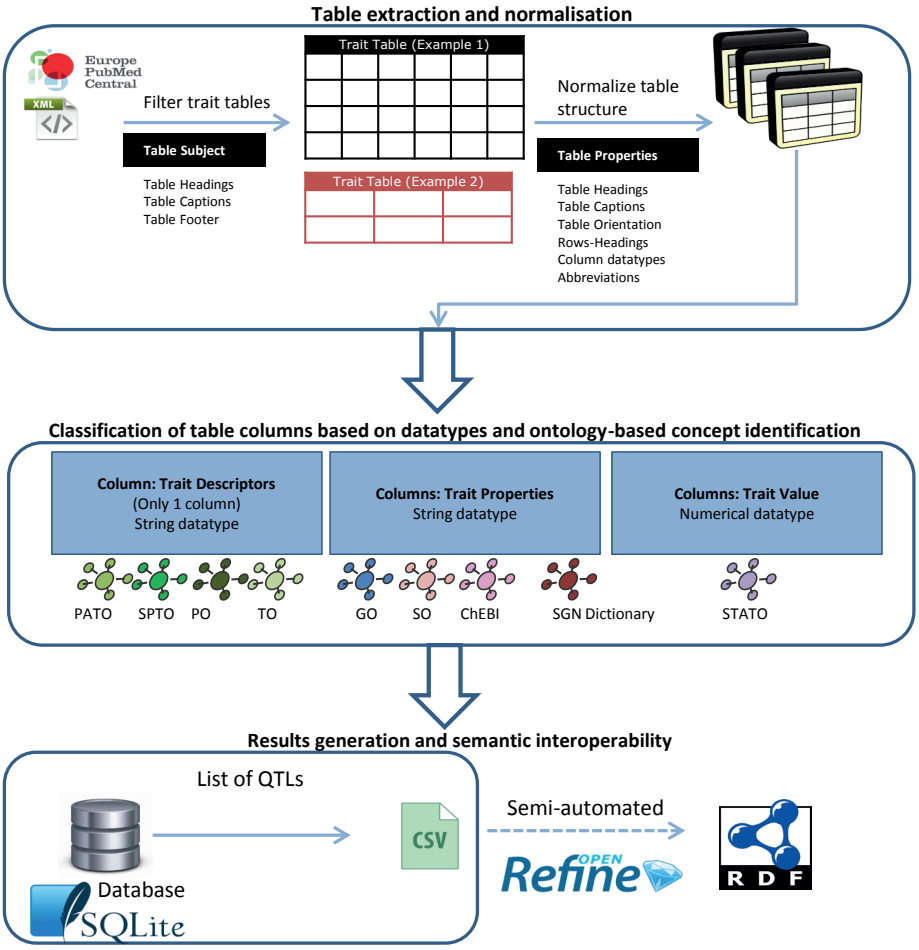


Figure 3.1: **QTLTableMiner^{++} workflow including semantic transformation using OpenRefine**.

Table extraction and normalization

First, the QTM tool retrieves open access articles in *XML* format from the Europe PMC repository [17] using the programmatic web interface (RESTful API). Then it detects tables in the articles using the `<table-wrap> .. </table-wrap>` XML tags and filters trait-related tables using keyword matching against table subjects derived from captions, headings and footers.

Tables are usually heterogeneous in structure (Fig. 3.2a). For example, they can have horizontal/vertical orientation, nested structure or headings that refer to more

than one row or column. Although the XML output includes tables in (semi-)structured forms, further normalization of the tables is required to query over them. Therefore, we developed normalization rules based on table properties (e.g. captions, footers, column headers, data types and abbreviations). We use the Schwartz and Hearst abbreviation-expansion (S & H) algorithm to identify and expand all abbreviations found in table headings and cell entries [87].

After the normalization step, each table has a single row of headings including expanded abbreviations and each cell is identified by a pair of row/column indices.

Ontology-based concept identification and classification of table columns

QTM uses the Apache Solr search platform (version 6.2.1, [86]) to semantically annotate biological entities and statistical concepts found in tables using domain-specific dictionaries or ontologies. In particular, the Solanaceae Phenotype Ontology (SPTO) [88] [89], Plant Ontology (PO) [90] [91], Phenotypic Quality Ontology (PATO) [92] [93] and Trait Ontology (TO) [94] were used to identify plant-specific phenotypic information whereas Gene Ontology (GO) [95] and Sequence Ontology (SO) [96] were used to identify genotypic information. Further, small chemical compounds were annotated using the Chemical Entities of Biological Interest database/ontology (ChEBI) [97] [98]. Plant-specific genetic markers and gene or transcript IDs were obtained from the Sol Genomics Network (SGN) [99] [100]. STATistics Ontology (STATO) [101] was used to annotate the quantitative (statistical) results of QTL mapping experiments.

Table columns were classified according to the column properties into three categories: i) trait descriptors refer to a trait, phenotype or QTL in the column headings with alphanumeric data type (using SPTO, PO, PATO and TO); ii) trait properties refer to chemical compounds, genes, transcripts or genetic markers in all other columns with alphanumeric data type (using ChEBI, GO and SO); and iii) trait values are columns that contain exclusively numerical data types (using STATO).

Results generation and semantic interoperability

The last steps of the workflow involve extracting QTL statements from the trait tables and writing the annotated results into a relational database (SQLite v3.11.0) [102]. The database schema consists of six tables: *ARTICLE*, *TRAIT_TABLE*, *ABBREVIATION*, *QTL*, *COLUMN_ENTRY* and *CELL_ENTRY* (see 6.1 in Supplementary Materials). In addition, the results stored in the *QTL* table are written into a text file (CSV).

Furthermore, the extracted QTL data were transformed into semantically interoperable RDF-based formats using the OpenRefine software [103]. The resulting RDF graphs including the SQLite database and CSV files were deposited at the Zenodo repository according to the FAIR (Findable, Accessible, Interoperable and Re-usable) Data guiding principles [45] (doi:10.5281/zenodo.1215044, [104]).

Performance evaluation and validation

Experimental design

We assessed the performance of the QTM tool using two manually annotated corpora of 30 open access articles each. The first set contains QTL mapping studies of tomato (*Solanum lycopersicum*) whereas the second set is focused on potato (*Solanum tuberosum*). Although the presented version of the tool uses a tomato-specific dictionary to annotate genes, transcripts and genetic markers, it can be adopted for use on other crop species. QTM is expected to detect and semantically annotate biological entities such as genes and markers in the set 'tomato'. However, QTM can also perform well on other species. For this, we use the second set of articles, i.e. set 'potato', for which QTM is expected to detect QTL statements without annotating biological entities such as genes and markers.

By our manual curation, the set 'tomato' included 66 trait tables with 2326 rows, 292 abbreviations, 757 biological entities and 405 QTL statements whereas the set 'potato' included 71 trait tables with 1292 rows, 207 abbreviations, 200 biological entities and 196 QTL statements. Specifically, precision and recall measures were obtained at four distinct levels of i) trait table, ii) abbreviation, iii) biological entities, and iv) QTL statement. Each result set was classified into four disjoint classes of the confusion matrix (i.e. true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN)). Precision and recall were calculated as TP / (TP + FP) and TP / (TP + FN), respectively.

Runtime and memory usage

The runtime and memory usage of the QTM tool were collected using three sets of articles (N=10, 20 or 30) derived from the tomato-specific corpus.

Results

Workflow demonstration on exemplary articles

QTM takes one or more PubMed Central identifiers (PMCIDs) as input and returns a list of QTL statements, further exemplified by an article (PMC4266912) in Fig. 3.2. In this article, there are three trait tables (i.e. Table 1, Table 2 and Table 3) with a total 35 rows out of which only 8 were QTL statements (Table 3). The tool detected 7 out of 8 QTL statements.

In each QTL statement, biological entities such as traits, genes and markers were annotated using ontological terms. In the 7 QTL statements detected, there were 7 unique traits linked to 7 genes and 7 SNP-based markers. In particular, QTM annotated a subset of traits (3 out of 7) while it detected all genes and markers (7 out of 7).

Importantly, QTL statements from multiple tables can be combined using the ontology-based annotation and the S & H abbreviation-expansion algorithm. For example, the state-

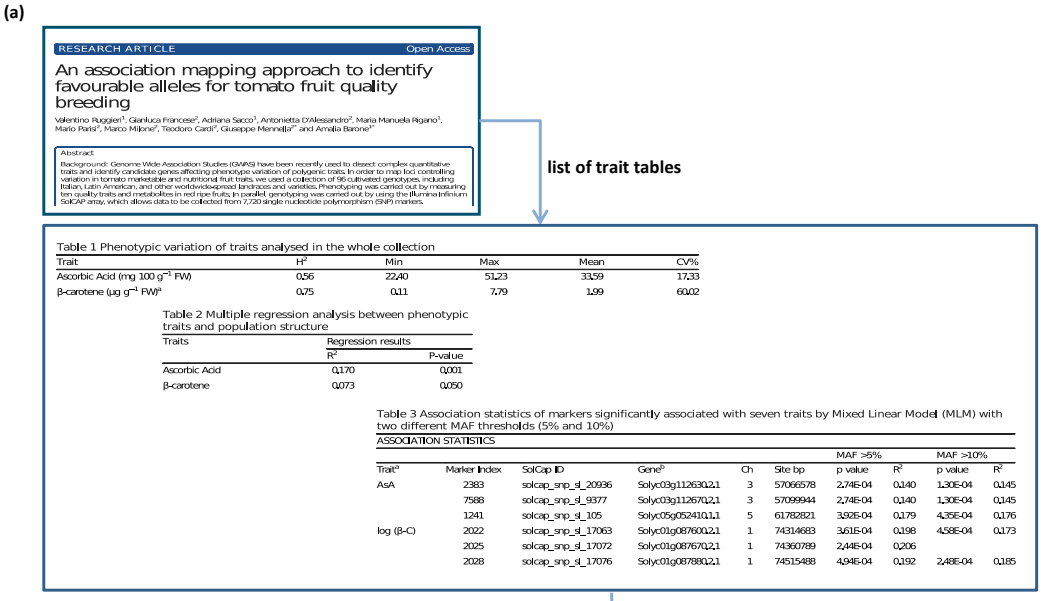


Figure 3.2: QTLTableMiner^{++} workflow exemplified on an article. a) Input article (PMC4266912) with three trait tables (*Table 1-3*, only the top-two rows *per* table are shown), b) trait statements identified in these tables, c) output list of QTL statements.

ments including terms such as *ascorbic acid* (CHEBI:22652) and *β-carotene* (TO:000269) were combined from the three tables (Fig. 3.2b). Note that both terms were abbreviated as *AsA* and *β*-C in this article (*Table 3*). Finally, QTM outputs all QTL statements detected in the article(s) (Fig. 3.2c).

Performance evaluation on both tomato and potato datasets

Detection of trait tables

The QTM tool recovered almost all trait-related tables for both manually curated corpora (Fig. 3.3). All trait tables were correctly identified except Table 1 of PMC2652058 (in tomato) and Table 1 and Table 2 of PMC3023753 (in potato). In fact, these three tables eluded detection due to missing words such as trait, QTL or phenotype in their descriptions and/or bodies.

The detection of trait tables reached 100% precision for both sets whereas the recall was slightly lower (98.48% for tomato and 97.18% for potato). Confusion matrices for the detection of trait tables for set 'tomato' and set 'potato' are provided in supplementary tables 6.4 and 6.8 respectively.

Detection of trait-specific abbreviations

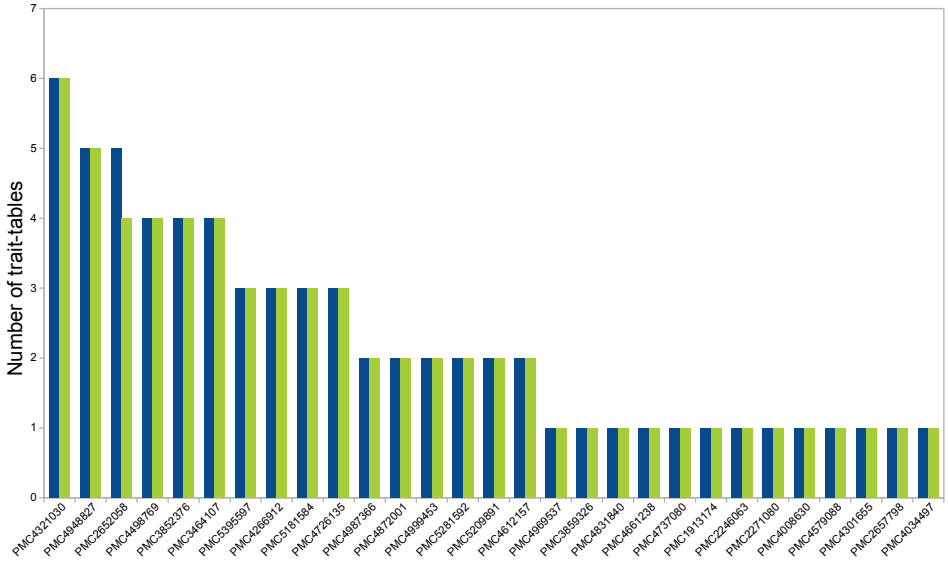
Detecting abbreviations is a prerequisite for reliable annotation of biological entities (e.g. traits, genes and markers) using standardized terms from domain-specific dictionaries or ontologies.

QTM detected abbreviations in the trait tables found in 10 out of 20 articles in set 'tomato' and in 12 out of 19 articles in the set 'potato' (Fig. 3.4). As the S & H algorithm is a rule-based approach, QTM performs in all or nothing (binary) manner. This means that if the statements mentioning abbreviations were written in the algorithm required formations (long form (abbreviation) or abbreviation (long form)), QTM was able to detect all the abbreviations and *vice versa*.

QTM identified 159 out of 292 abbreviations (recall of 54.45%) for tomato and 147 out of 207 abbreviations (recall of 71.01%) for potato in the trait tables. All abbreviations were true positives (100% precision). Confusion matrices for the detection of abbreviation for set 'tomato' and set 'potato' are available in supplementary tables 6.5 and 6.9 respectively.

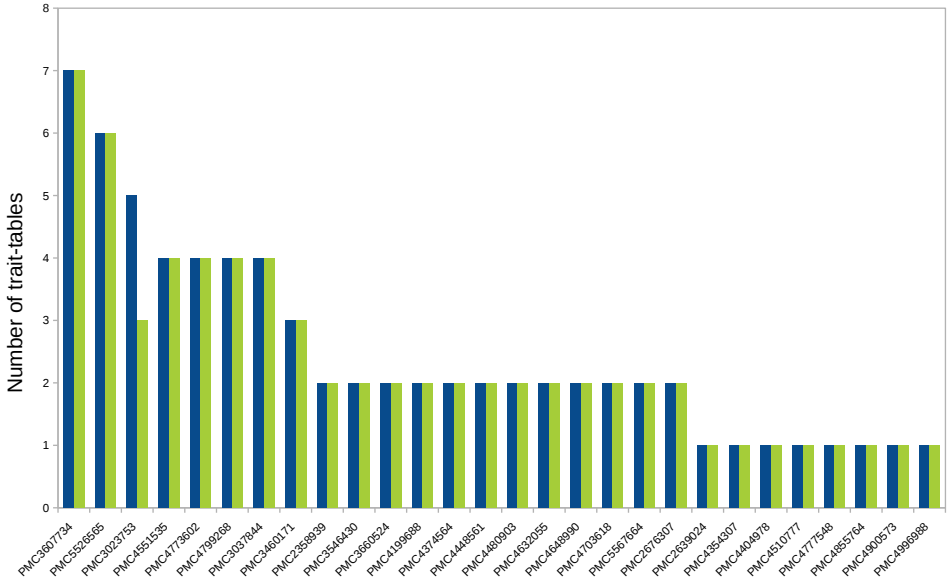
Annotation of biological entities

QTM identifies and semantically annotates biological entities such as genes, genetic markers, proteins, metabolites or traits. In the set 'tomato', QTM detected 468 out of 757 biological entities, of which 393 were TP, 82 were FP, and 288 were FN with a recall of 57.71% and a precision of 82.74%. Similarly, in the set 'potato' QTM detected 73 biological entities out of the total 200. There were a total of 62 TP, 3 FP, 127 were FN. Here, the



(a) Set ‘tomato’

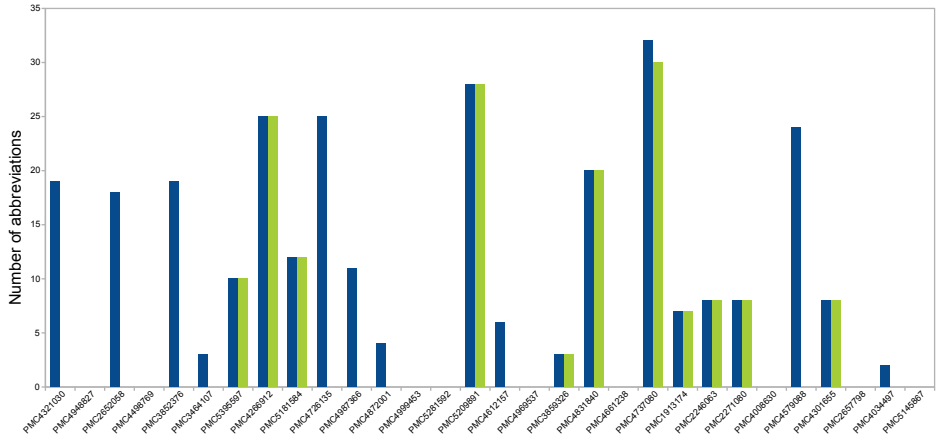
actual detected



(b) Set ‘potato’

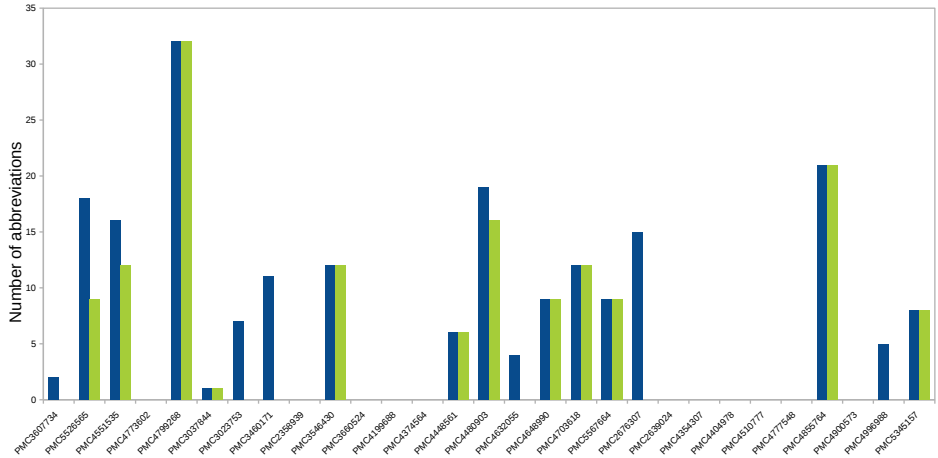
actual detected

Figure 3.3: Bar graphs of the numbers of QTL tables detected per article for the manually curated set ‘tomato’ (a) and set ‘potato’ (b) using the QTLTableMiner^{++}.



(a) Set 'tomato'

actual detected



(b) Set 'potato'

actual detected

Figure 3.4: Bar graphs of the numbers of abbreviations detected per article for the manually curated set 'tomato' (a) and set 'potato' (b) using the QTLTableMiner^{++}.

recall was low (35.53%) but the precision was high (95.89%). These results are shown in the Figure 3.5. Confusion matrices for the detection of biological entities in tables for set 'tomato' and set 'potato' are provided in supplementary tables 6.6 and 6.10 respectively.

Detection of QTL statements

The main objective of QTM is to find QTL statements in tables. In the set 'tomato', QTM detected 529 QTL statements while the actual number of QTL statements were only 405. There were a total of 398 TP, 136 FP and 32 FN. Here, there is an increase in the number of FP statements detected due to the fact that QTM has difficulties in dealing with columns with special characters. For example, in Table 1 of PMC4987366, QTM reads column Genotype as a column with alphanumeric data type due to the presence of characters '**', and thereby associates traits with the given genotype. Nevertheless, QTM performed with a precision of 74.53% and recall of 92.56% in set 'tomato'. Similarly, in the set 'potato' QTM detected 233 QTL statements while the actual number of QTL statements were total 196. There were a total of 188 TP, 39 FP and 2 FN, thus QTM performed with a high precision of 82.82% and a recall of 98.94%. These results are shown in the Figure 3.6. Confusion matrices for the detection of QTL statements in tables for set 'tomato' and set 'potato' are provided in supplementary tables 6.7 and 6.11 respectively.

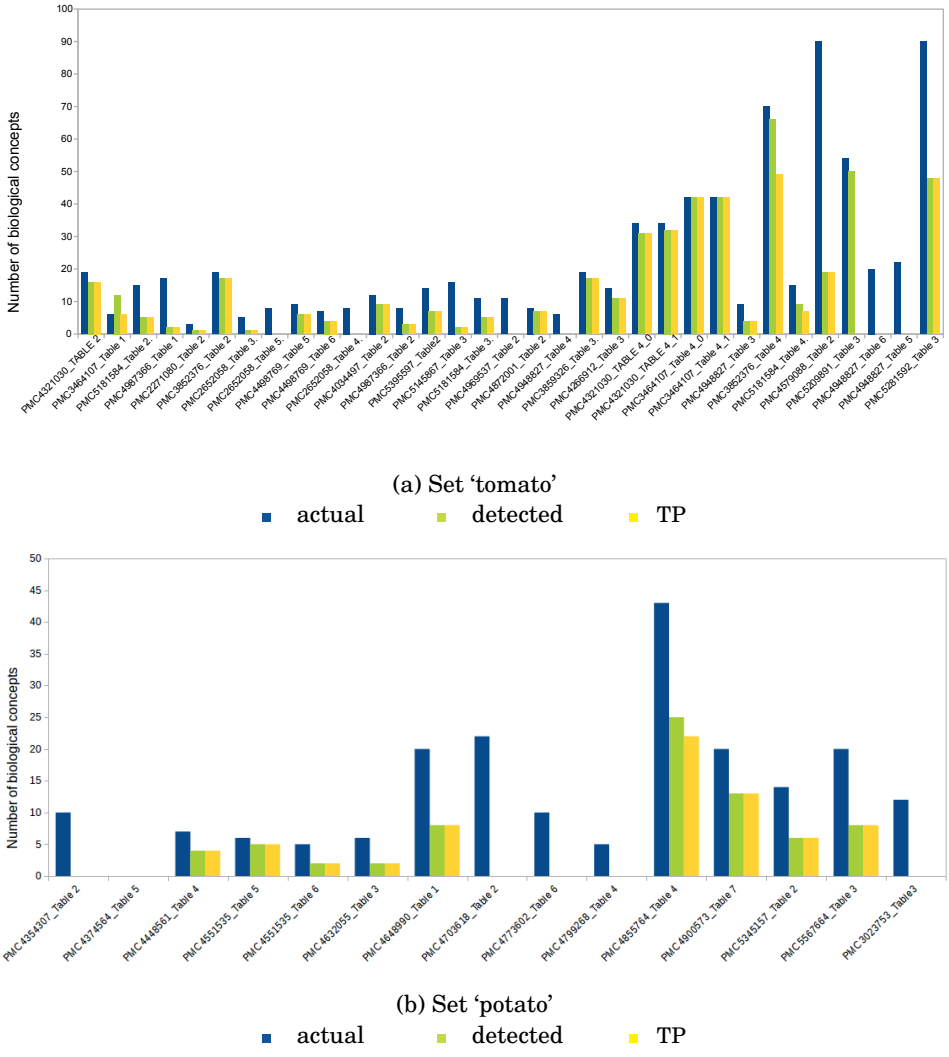
Table 3.1 tabulates the precision and recall obtained for each task described above.

Table 3.1: **Benchmark results of the QTLTableMiner^{++} tool on different tasks.**

Detection	**Precision (%)**		**Recall (%)**	
	Tomato	***Potato***	***Tomato***	***Potato***
QTL tables	100	100	98.55	97.18
Abbreviations	100	100	54.45	71.01
Biological entities	82.74	95.89	57.71	35.53
QTL statements	74.53	82.82	92.56	98.94

Runtime and memory use

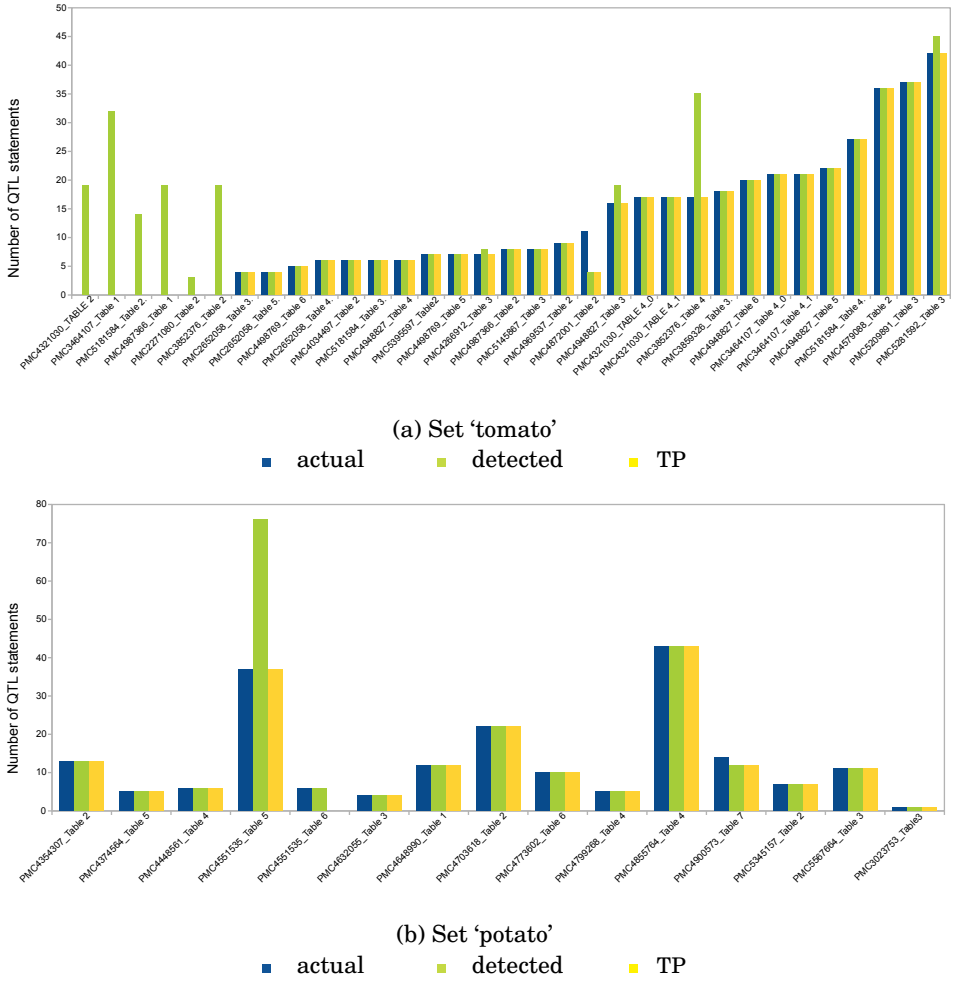
Table 3.2 summarizes the runtime and memory use of the QTM tool for three sets of full-text articles (using a commodity hardware with Intel Core i5 CPU, 4GB RAM, 228GB SSD, Ubuntu Linux 16.04.3 LTS). The results indicate that both the runtime and memory use increase approx. linearly with the amount of input.



(a) Set 'tomato'

(b) Set 'potato'

Figure 3.5: Bar graphs of the numbers of biological entities detected in trait tables for the manually curated set 'tomato' (a) and set 'potato' (b) using the QTLTableMiner^{++}.



(a) Set 'tomato'

(b) Set 'potato'

Figure 3.6: Bar graphs of the numbers of QTL statements detected in trait tables for the manually curated set 'tomato' (a) and set 'potato' (b) using the QTLTableMiner^{++}.

Table 3.2: **Scalability of the QTLTableMiner^{++} tool in terms of runtime and memory use.**

Number of articles	**Number of tables**	**Number of rows in tables**	**Runtime (HH:MM:SS)**	**Max. memory (MB)**
10	42	1562	00:04:10	19
20	58	2090	00:06:56	23
30	66	2326	00:07:58	30

Discussion

QTM extracts QTL statements from tables of scientific articles as well as enables (re)publishing these statements in machine-readable and semantically interoperable RDF-based formats. Although it is possible to include review papers as an input for this tool, more accurate information can be obtained in the primary-data papers. Review papers frequently contain abbreviated references to the original papers and not the primary data.

Although, this tool was used to extract trait tables from plant-specific literature, the approach is also applicable to other domains. For example, a similar approach was used by Mulwad *et al.* [105] and Milosevic *et al.* [106] to retrieve health-related indicators about patients (e.g. the body mass index or BMI) from clinical literature. An important component in the QTM workflow is the use of the RESTful API of the Europe PMC, which provides open access articles in the (semi-)structured XML format. The resulting XML output complies with the JATS schema, which is a *de facto* standard for archiving and interchanging scientific articles. One drawback of using Europe PMC is that it is mainly focused on the biomedical literature while the plant literature is not covered extensively in this repository. As a result, we had to restrict our input set of articles (60 in total). Recently, publishers such as Springer or Elsevier have released Web APIs, which provide access to articles in JATS-compliant XML format. Therefore, our tool can be extended to use these APIs in the near future.

In total, QTM detected 529 QTL statements associated with 73 traits in tomato and 233 QTL statements associated with 16 traits in potato. In the set 'tomato' the five most common traits associated with the detected QTL statements were pH (SP:0000170), fruit shape (SP:0000038), compound leaf (SP:0000177), fruit (SP:00000378), and stem (SP:0000193). Whereas, in the set 'potato' the five most common traits associated with the detected QTL statements were anthocyanin content (SP:0000016), fruit shape (SP:0000038), fructose content (SP:0000386), stem strength (TO:0000051), and plant fresh weight (TO:0000442).

QTM performed better in the detection of biological entities for the set 'tomato' in comparison to the set 'potato' because the dictionaries used to annotate genes and genetic markers were tomato specific. The QTM algorithm has two distinctive features: i) the classification of table columns according to column properties into trait descriptors, trait properties and trait values, as well as ii) the ontology-based concept identification and annotation. We also present an approach to transform the extracted QTL information into the form of triples (<*subject*> <*predicate*> <*object*>), where <*subject*> refers to a trait descriptor, <*predicate*> is the column heading and <*object*> refers to the cell value in that column. QTM outputs a list of QTL statements both in a CSV file and in a SQLite database. Using the Linked Data approach, the resulting QTL statements can be integrated with genome-sequencing and annotation data to develop new or improve upon existing precision breeding programs. Combining the information available in scientific literature and molecular biology databases will help in narrowing down the QTL regions

to detect candidate genes associated with traits of interest.

Conclusions

QTM is a tool that aids in extracting QTLs from literature and in sharing these valuable data assets in machine-readable and semantically interoperable formats, and as such can help in formulating strategies for breeding crops of interest.

Chapter 4

Linked Data platform for Solanaceae species

Gurnoor Singh [1, *], Arnold Kuzniar [2, *], Matthijs Brouwer [1], Carlos M. Ortiz[2], Christian W. B. Bachem[1], Richard Finkers [1], Richard G.F. Visser [1]

[1] Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands
[2] Netherlands eScience Center (NLeSC), Amsterdam, the Netherlands
[*] Both authors contributed equally to this manuscript.

to be submitted

Abstract

Introduction

Genetics research is focusing more and more on mining fully sequenced genomes and their annotations to identify the causal genes associated with traits (phenotypes) of interest. However, a complex trait is typically associated with multiple quantitative trait loci (QTLs), each comprising of many genes, that can positively or negatively affect the desired trait of interest. To help breeders in ranking candidate genes, we developed an analytics platform called pbg-ld that provides semantically integrated geno- and pheno-typic data on Solanaceae species.

Methodology

This platform combines both unstructured data from scientific literature and structured data from publicly available biological databases using the Linked Data approach. In particular, QTLs were extracted from tables of full-text articles from the Europe PMC repository using QTLTableMiner++, while the genomic annotations were obtained from the Sol Genomics Network (SGN), UniProt, and Ensembl Plants databases. These datasets were transformed into Linked Data graphs, which include cross-references to many other relevant databases such as Gramene, Plant Reactome, InterPro and KEGG Orthology (KO), etc. Users can query and analyze the integrated data through a web interface or programmatically via the SPARQL and RESTful services (APIs).

Results

We illustrate the usability of pbg-ld for querying the genome annotations, comparing genome graphs, retrieving candidate genes and proteins for a trait of interest using GO annotations and text-search, and by studying similar traits of interests in 2 plant species.

Keywords: Prioritization of Candidate Genes, Semantic Web, Linked data, Plant Breeding, QTLs

Introduction

The availability of annotated reference genome assemblies for several crop species including tomato [2], potato [3], brassica [7], and cucumber [8] has enabled plant breeders and researchers to elucidate a trait's linkage to a genomic location(s). Mining genome annotations can help in identifying candidate genes that positively or negatively affect a trait of interest, which plant breeders aim to improve. However, genome annotations are commonly available across multiple databases and file formats (e.g., in the Generic Feature Format (GFF)), which hampers integrated data analyses.

Traditionally, plant breeders identified chromosomal regions using genetic markers that are statistically associated with traits of interest. These genomic regions are called quantitative trait loci (QTLs). A QTL region can easily contain thousands of genes including those that negatively influence the trait of interest [107]. Therefore, detecting the causative gene for breeding is of major importance. There are three major approaches to address the challenge of candidate gene prediction in crop species: i) the analysis of gene expression data or co-expression networks [9], ii) comparative genomics [10], and iii) integrate information stored in scientific literature and in molecular biology databases such as the ELIXIR Core Data Resources [37] (including ENA [32], Ensembl Plants [35] and UniProt [34]) and the Sol Genomics Network (SGN) [38]. To address the need for improved access to integrated plant data, we developed the Solanaceae Linked Data platform (pbg-ld) [108] that combines QTLs from scientific literature and genome annotations from public databases using the Linked Data approach [109]. Our approach is to create a (semantic) web of data rather than that of hypertext (HTML) documents using Uniform Resource Identifiers (URIs) and Resource Description Framework (RDF) [15]. A URI is an HTTP-based resource identifier assigned to an entity whereas RDF is a generic graph-based data model for describing entities and their relationships. In addition, publishing data according to FAIR Data Principles [45] further increases the degree of discoverability and (re-)usability of research data. Briefly, according to these principles every data element should have a unique persistent identifier, with a searchable metadata ("Findable"); all identifiers should resolve to (meta)data using an open standard protocol ("Accessible"); the (meta)data should use a representation language that utilizes widely accepted domain-specific ontologies ("Interoperable"); and finally, the data should be well described with cross-references and with available license information ("Reusable"). Further, FAIR Data Point is an exemplary implementation that allows data owners to expose their data assets in compliance with the FAIR Data Principles via a RESTful API [110].

In plant sciences, several controlled vocabularies and ontologies have been developed to standardize domain-specific terms and/or represent the current knowledge of the domain in a machine-readable form. For example, the Solanaceae Phenotype Ontology (SPTO) [88], Crop Ontology [89], Plant Ontology (PO) [91], Phenotypic Quality Ontology (PATO) [93] and Trait Ontology (TO) [94] are used to identify plant-specific phenotypic information while Gene Ontology (GO) [111], Sequence Ontology (SO) [96] and FALDO [112]

are used to identify genotypic information. Similarly, the Chemical Entities of Biological Interest database/ontology (ChEBI) [98] is focused on 'small' chemical compounds.

There are several plant-specific databases available that provide geno- and phenotypic data. For example, Ensembl Plants is a widely used integrated resource on plant genomes. Similarly, UniProt is a database of protein sequences, function annotations and proteomes of various species including plants. Both Ensembl Plants and UniProt release their data in RDF-based format. The Arabidopsis Information Resource (TAIR) [39] is a resource to analyze and to compare molecular, biological, and genetical data of the model species Arabidopsis thaliana. Further, the Sol Genomics Network (SGN) [38] provides genomic, genetic and phenotypic information for members of the Solanaceae family. Plant Genome DataBase Japan (PGDBj) [36] is an integrated web resource for plant genome-related information from literature and public databases. However, the TAIR, SGN, and PGDBj do not distribute their data in a semantically interoperable (RDF) format. The Planteome [113] database provides gene annotations and phenotypes with the help of reference ontologies such as PO, TO, GO and ChEBI. Planteome is a user-friendly tool to query traits of interest, germplasm, and putative candidate genes. However, it lacks QTLs, genetic markers and links to publicly available databases such as Ensembl Plants.

We present the pbg-ld platform that provides semantically integrated geno-/phenotypic data on Solanaceae species such as the (wild) tomato and potato species. The resulting (linked) datasets are made available through a web interface or programmatic services (SPARQL and RESTful APIs). The use of these data-access points is illustrated in the results section. Pbg-ld is a plant-oriented resource that aids breeders in detecting candidate genes for complex traits using the knowledge available in scientific literature and public databases.

Data generation and ingestion pipeline

Figure 4.1 illustrates the data generation and ingestion pipeline used by the pbg-ld platform. Geno- and pheno-typic data from three Solanaceae species: i) reference sequence tomato (*S. lycopersicum*), ii) wild tomato (*S. pennellii*) and iii) the reference sequence potato (*S. tuberosum*) is collected and integrated into this pipeline.

Data sources

To facilitate the integration of geno- and pheno-typic data of Solanaceae species, we used data from, both, semi-structured as well as structured resources. Semi-structured data resources include scientific articles in XML file format obtained from EuropePMC, or GFF (General Feature Format) based text files, which consists of one line per genomic feature, where each line contains about 9 columns of data. Data from these semi-structured resources were classified as Non-RDF data and were subsequently transformed into a (structured) RDF-data. On the other hand, structured data resources contained data

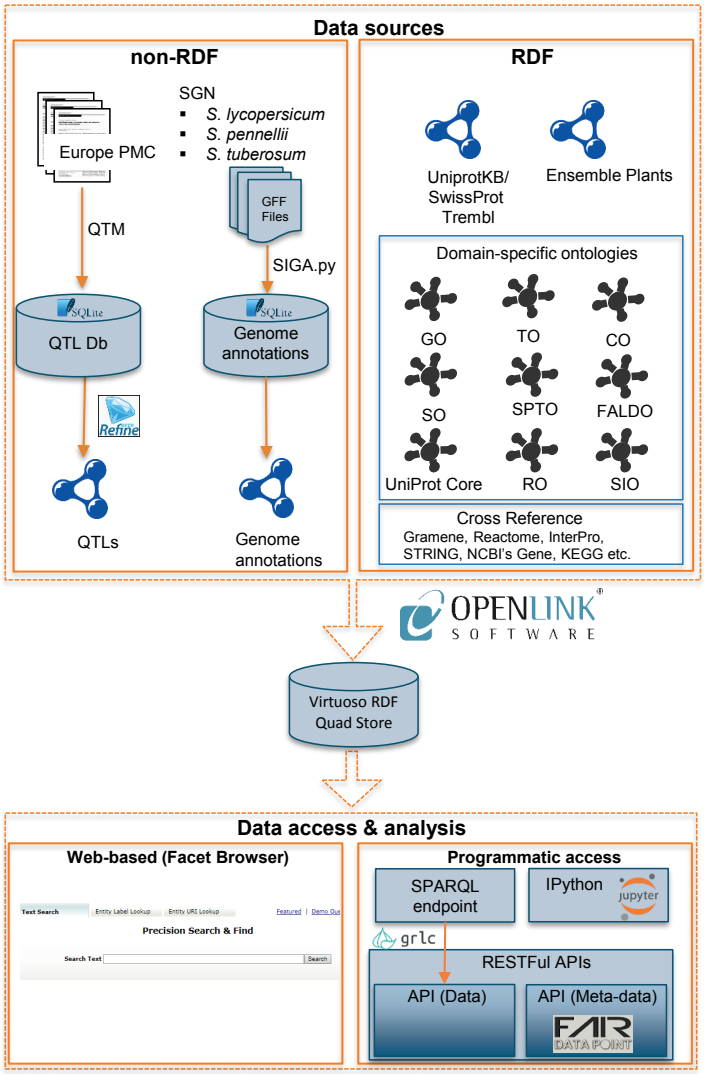


Figure 4.1: Data generation and ingestion pipeline. All data originate from either non-RDF or RDF sources. Several tools are used to retrieve and transform non-RDF data into RDF graphs: QTM is used to extract tomato and potato QTLs from Europe PMC articles; OpenRefine is used to transform the QTLs into RDF according to the specified data model. Similarly, SIGA.py tool converts the genome annotations, as provided by the Sol Genomics Network (SGN) in GFF files into RDF graphs with gene models and markers. In addition, the UniProt (proteomes) and Ensembl Plants (gene models) distribute their data in RDF format. All RDF graphs including domain-specific ontologies (in OWL) and database cross-references were stored and integrated with Virtuoso RDF Quad Store. The resulting linked datasets are made available for queries and analyses through data-access layer: i) Linked data browser, ii) SPARQL endpoint, iii) grlc-based Web API [114] and iv) FAIR Data Point (FDP) metadata (RESTful) service.

in the form of RDF structure, for example, genome annotation in Ensemble Plants and Uniprot, as well as domain ontologies.

QTL

QTL studies have widely been published in scientific articles, particularly in tables or supplementary materials. However, there is no established repository where experimental data on plant-specific QTL studies can be submitted. Therefore, QTL information is classified as non-RDF data and extracted from XML based scientific literature and processed to RDF graphs using the QTL TableMiner++ (QTM) tool [61] version (v1.1.0) [115]. QTM extracted 324 QTLs from a total of 21 Solanaceae–specific full-text articles in the Europe PMC literature repository. 234 of these QTLs (i.e., 93 in tomato and 64 in potato) were associated with exact chromosomal locations based on flanking markers while the remaining 90 QTLs were associated with peak markers and/or candidate genes.

SGN

SGN provides genome annotations in GFF files for Solanaceae species. The GFF files were transformed into RDF graphs using the SIGA.py command-line tool (v0.5.1) [116] (supplementary figure 6.2 shows the overall architecture). The gene models and the genetic markers of (wild) tomato (*S. lycopersicum* and *S. pennellii*) and potato (*S. tuberosum*) were downloaded from the SGN's FTP server (`ftp://ftp.solgenomics.net/genomes/`). For *S. lycopersicum*, the genome annotations comprising of Gene models, SGN markers, SolCAP markers, were taken from GFF files of the ITAG 2.4 released on 23-02-2014 as input files [117]. For *S.pennelli*, the genome annotations comprising of Gene model (spenn_v2.0) released on 27-08-14, and SGN markers released on 10-08-14 were taken as input [118]. Similarly, for *S. tuberosum*, genome annotations of PGSC_DM (diploid/double monoploid) version 4.03 released on 04-09-2013 on was taken as input [119].

Ensembl Plants and UniProt

Ensembl Plants is an genome-centeric integrated resource for plant sciences. Genome annotations of *S. lycopersicum* (release ITAG2.4 genome annotation based on SL2.50 genome assembly) [120] and *S. tuberosum* (release PGSC_DM 3.0 genome annotation based on SolTub3.0 genome assembly) were taken from the Ensembl Plants database (release 33) [121] in RDF format. Similarly, from Uniprot, the proteomes of *S. lycopersicum* [122] and *S. tuberosum* [123] in the RDF/Turtle format.

Ontologies

In this linked data platform, we integrated the following domain-specific ontologies: GO [111], SPTO [88], Crop ontology [89], SO [96], FALDO [112], TO [94], UniProt Core [124],

Semanticscience Integrated Ontology (SIO) [125], Relation Ontology (RO) [126], PO [91], PATO [93].

Linked data deployment

OpenLink's Virtuoso Universal Server (version 7.20.3217, open-source edition) was used to store and connect the data graphs in the RDF Quad Store. The pbg-ld platform including the associated RESTful web services, namely the grlc-based API for data and the FAIR Data Point API for metadata, were deployed on the academic HPC Cloud operated by SURFsara (using a virtual machine preinstalled with Ubuntu 18.04.1 LTS). Pbg-ld pipeline is made as a modular and re-usable pipeline with the help of Docker [127] and Ansible [128].

Data access & analysis

Pbg-ld provides access to the (meta)data through a web-based user interface (Virtuoso Faceted Browser [129]) and programmatic interfaces such as SPARQL [130] and RESTful APIs. Using the web-based user interface, a user can query the RDF triples in three different ways through i) a free-text search box, ii) an entity label search box or iii) an entity URI search box. There is a SPARQL endpoint provided for a user to write and execute SPARQL queries on the RDF graphs available in the pbg-ld platform. Further, with the help of grlc tool [114], we published customized RESTful APIs, built on the top of pbg-ld's datasets to provide easy to use programmatic access to our SPARQL endpoint. Data consumers who do not know the SPARQL query language can use these APIs to query the platform. This way grlc helps us hides the complexities or intricacies of SPARQL. Supplementary Table 6.12 provides a list of RESTful APIs available in pbg-ld. Lastly, a FAIR Data Point service is provided to expose machine-readable descriptions (metadata) about datasets in the pbg-ld platform. To show a valuable use-case of the pbg-ld platform we have developed exemplary Jupyter (IPython) Notebooks.

Results

Genome annotations on the faceted web browser

Pbg-ld allows the user to access and analyze data with the help of a facets browser. Figure 4.2 exemplifies a query for trait-gene associations using "fruit shape" as a search term. Here, this term (partially) matches several standardized trait names in the SPTO and TO ontologies (e.g., SP:0000038 [131] and TO:0002628 [132]). By selecting either one, pbg-ld returns seven QTLs associated with the trait of interest (i.e. "fruit shape"). In Figure 4.2, one such a QTL is selected for further analysis, i.e. QTL:4321030_4_14 [133]. QTM extracted this QTL from table 4 of a Europe PMC article PMC4321030 [134]. This QTL

occurs on chromosome 11 is marker by flanking markers C2_At2g14260 and TG400 on chromosome 11, therefore pbg-ld finds out the list of all the gene inside this QTL region. In Figure 4.2, one such gene (Solyc11g038340.1) in QTL:4321030_4_14 is selected. Pbg-ld web interfaces contain direct links to allow the user to further browse the annotation, properties, and sequence of this gene at the SGN database, Ensembl Plants database, UniProt. For example Figure 4.2, shows the sequence of this selected gene in the SGN's genome browser (JBrowser).

Figure 4.2: Browsing trait-gene associations for the trait “fruit shape”, using the pbg-ld faceted browser.

Exemplary data queries via SPARQL and API

(I) **SPARQL query to list QTLs, associated gene IDs and GO annotations related to an example trait "fruit shape" (SP:0000038).**

SPARQL Query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX so: <http://purl.obolibrary.org/obo/so#>
PREFIX go: <http://www.geneontology.org/formats/oboInOwl#>
SELECT
   str(?qtl_id) AS ?qtl_id
   str(?sgn_gene_id) AS ?sgn_gene_id
   GROUP_CONCAT(DISTINCT str(?go_id); SEPARATOR="\n") AS ?go_id
   GROUP_CONCAT(DISTINCT str(?go_term); SEPARATOR="\n")AS ?go_terms
   GROUP_CONCAT(str(?go_cat); SEPARATOR="\n")AS ?go_cat
WHERE {
   GRAPH <http://europepmc.org/articles> {
      ?qtl a obo:SO_0000771 ;
         obo:RO_0003308 ?trait ;
         so:overlaps ?gene ;
         dcterms:identifier ?qtl_id .
      FILTER(?trait = obo:SP_0000038)}
   GRAPH <http://solgenomics.net/genome/Solanum_lycopersicum> {
      ?gene so:transcribed_to ?transcript ;
         dcterms:identifier ?sgn_gene_id .
      ?transcript rdfs:comment ?annot ;
         dcterms:identifier ?sgn_trans_id}
   GRAPH <http://purl.obolibrary.org/obo/go.owl> {
      ?go rdfs:label ?go_term ;
         go:id ?go_id;
         go:hasOBONamespace ?go_cat .
       FILTER regex(?go_cat, 'biological_process|molecular_function')}
   FILTER CONTAINS(?annot,?go_id)
}
GROUP BY ?qtl_id ?sgn_gene_id ?annot
ORDER BY DESC(STRLEN(str(?annot)))
LIMIT 5
```

qtl_id	sgn_gene_id	go_id	go_term	go_category
QTL:4321030_4_14	Solyc11g040340.1	GO:0003824	carbohydrate metabolic process	molecular_function
		GO:0004553	catalytic activity	molecular_function
		GO:0005975	hydrolase activity, hydrolyzing O-glycosyl compounds	biological_process
QTL:4321030_4_14	Solyc11g040390.1	GO:0006520	NADP binding	biological_process
		GO:0008152	cellular amino acid biosynthetic process	biological_process
		GO:0008652	cellular amino acid metabolic process	molecular_function
		GO:0050661	metabolic process	biological_process
		GO:0055114	oxidation-reduction process	biological_process
QTL:4321030_4_14	Solyc11g039840.1	GO:0008121	metal ion binding	molecular_function
		GO:0046872	oxidation-reduction process	biological_process
		GO:0055114	ubiquinol-cytochrome-c reductase activity	molecular_function
QTL:4321030_4_14	Solyc11g061750.1	GO:0003677	DNA binding	molecular_function
		GO:0006355	regulation of transcription, DNA-template	biological_process
QTL:4321030_4_14	Solyc11g043120.1	GO:0004428	1-phosphatidylinositol 4-kinase activity	molecular_function
		GO:0004430	Binding	biological_process
		GO:0005488	obsolete inositol or phosphatidylinositol kinase activity	molecular_function
		GO:0016773	phosphatidylinositol binding	molecular_function
		GO:0035091	phosphatidylinositol-mediated signaling	molecular_function
		GO:0048015	phosphotransferase activity, alcohol group as acceptor	molecular_function

Figure 4.3: Input and output of a sparql query to list QTL, associated gene IDs, and GO annotations related to a trait (example fruit shape (SP:0000038).

Similar to the manually browsed query in Figure 4.2 with a user interface, Figure chap4:Fig3 exemplifies a way to write the query of trait-gene associations. This query yields the QTLs, and candidate genes with their GO annotations for the trait fruit shape (i.e. represented in the SPTO ontology with the id SP:0000038). Here GO annotations of a molecular function, or biological process are only retrieved.

(II) **Input and output of a SPARQL query to list genes/proteins annotated**

with Gene Ontology (GO) terms that relate to both, “fruit” and “ripening”. Searching for text matches (regular expressions) in the SPARQL query has been done with the help of virtuoso’s bif:contains predicate.

SPARQL Query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX uniprot: <http://purl.uniprot.org/core/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX so: <http://purl.obolibrary.org/obo/so#>
PREFIX go: <http://www.geneontology.org/formats/oboInOwl#>

SELECT
   DISTINCT
   str(?gene_name) AS ?gene_name
   str(?sgn_gene_id) As ?sgn_gene_id
   str(?uniprot_acc) AS ?uniprot_acc
   concat(?go_id, ' ', ?go_term,) AS ?go_term
   str(?go_cat) AS ?go_category
WHERE {
   GRAPH <http://www.uniprot.org/proteomes/Solanum_lycopersicum> {
      ?prot uniprot:classifiedWith ?go ;
           uniprot:encodedBy/skos:prefLabel ?gene_name
   }
   GRAPH <http://plants.ensembl.org/Solanum_lycopersicum> {
      ?prot dc:identifier ?uniprot_acc ;
           rdfs:label ?uniprot_id ;
           dc:description ?uniprot_des ;
           ^<http://rdf.ebi.ac.uk/terms/ensembl/CHECKSUM>
?ensembl_prot_id .
       ?ensembl_transcript_id so:translates_to ?ensembl_prot_id ;
           so:transcribed_from/dc:identifier ?sgn_gene_id .
   }
   GRAPH <http://purl.obolibrary.org/obo/go.owl> {
      ?go ?p ?o ;
         rdfs:label ?go_term ;
         go:id ?go_id ;
         go:hasOBONamespace ?go_cat .
      ?o bif:contains '( fruit AND ripening )' .
      FILTER regex(?go, obo:GO_)
   }
}
ORDER BY ?gene_name
```

gene_name	sgn_gene_id	uniprot_acc	go_id	Go_term	go category
ACO1	Solyc07g049530.2	P05116	GO:0009835	fruit ripening	biological_process
ACO3	Solyc09g089580.2	P10967	GO:0009835	fruit ripening	biological_process
ACO4	Solyc07g049550.2	P24157	GO:0009835	fruit ripening	biological_process
ACS2	Solyc01g095080.2	P18485	GO:0009835	fruit ripening	biological_process
ACS3	Solyc02g091990.2	Q42881	GO:0009835	fruit ripening	biological_process

Figure 4.4: Input and Output of a sparql query to count number of proteins in cultivated tomato (S.lycopersicum) and its wild relative (S.pennellii) according to the SGN database.

SPARQL query in Figure 4.4 highlights a way to do textual search over the annotations of genes/proteins. With the help of bag-of-words based regular expressions, we query genes and proteins containing GO annotation with the words “fruit” and “ripening”. The resulting output is the list of genes/proteins involved in the biological process called fruit ripening (i.e. represented in the GO ontology with the id GO:0009835).

(III) **An APIs based Comparison among tomato genome graphs**

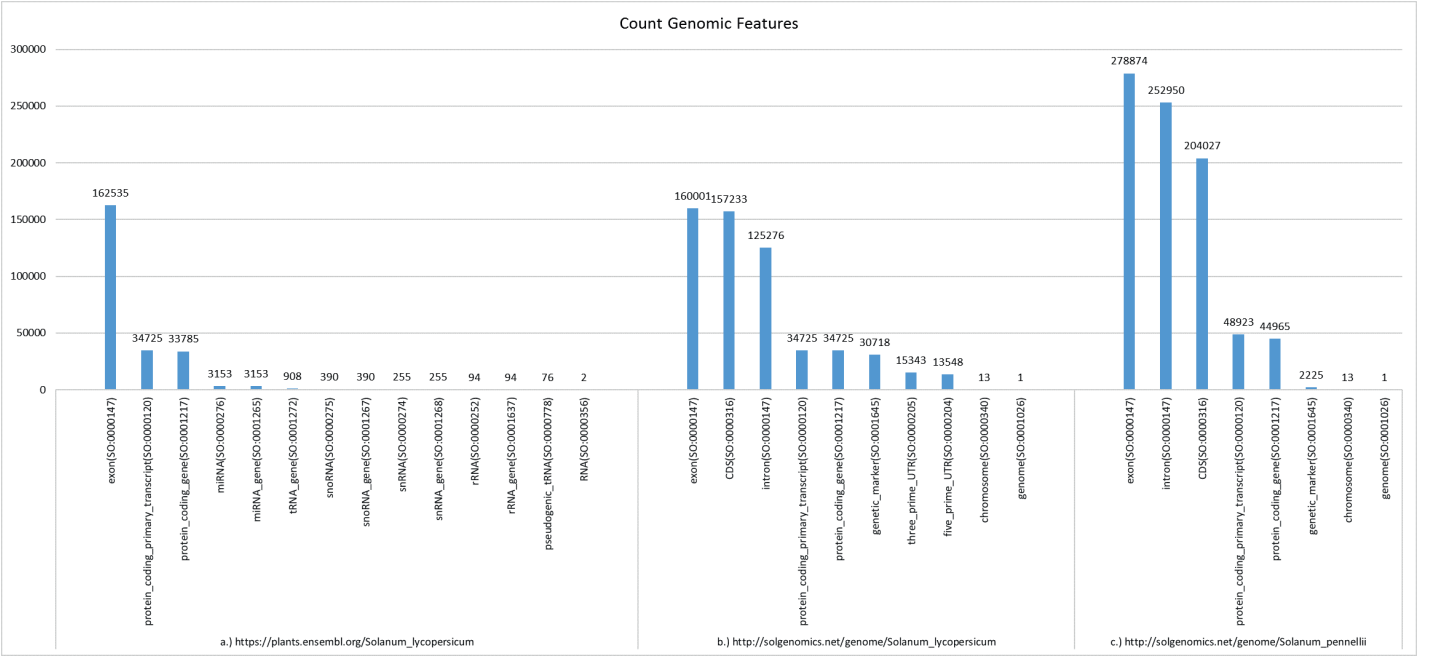


Figure 4.5: Column Chart to count the total number of genomic feature 3 genome graphs. These three genomic graphs are `https://plants.ensembl.org/Solanum_lycopersicum`, `http://solgenomics.net/genome/Solanum_lycopersicum`, `http://solgenomics.net/genome/Solanum_pennellii`.

Pbg-ld can be used to check data features, data consistency and data quality across various biological databases. This can be done either by manually writing a SPARQL query on the SPARQL endpoint of pbg-ld or accessing the countfeatures API of pbg-ld, with the genomic graph as a parameter. For example, these cited URI counts the genomic features annotated in the *S. lycopersicum* genome according to Ensembl Plants [135] and SGN [136], and in the *S. pennelli* genome in SGN [137]. We compare the differences between the genomic features of the tomato graphs in a column chart in Figure 3. While studying the genomic features a.) *S. lycopersicum* according to Ensembl Plants and SGN, it's evident that there are a total of 33785 protein_coding_genes in Ensemble Plants, whereas there are 34725 protein_coding_genes in SGN graph. There are about 940 unique genes in the SGN database that have still not been mentioned in the Ensembl Plants database. Furthermore, the results also highlight that genetic marker are included in SGN but not in Ensembl Plants while the latter database contains RNAs. On the one hand, where pbg-ld can be used to compare databases, pbg-ld can also be used to compare genomic data of different species in the same family. *S. pennellii* is a wild tomato species that is relatively distant from the domesticated *S. lycopersicum*. Because of *S. pennelli*'s extreme stress tolerance, unusual morphology, and a genome sequence 119 Mb more than *S. lycopersicum*, it is an important donor of germplasm for the cultivated tomato. While comparing the genomic features b.) *S. lycopersicum* in SGN vs c.) *S. pennelli* in SGN on Figure 3, it is depicted that the number of genomic features in S.pennellii are averagely 1.5 times the number of genomic features of than S.lycopersicum.

Comparitive genomics use-case: Using Pbg-ld to study tomato fruit shape and tuber-shape

In this section, we exemplify the use of the pbg-ld endpoints (APIs) with a Jupyter Notebook [51] to study the difference in the genetic mechanism underlying fruit shape in tomatoes and tuber shape in potatoes.

Tomato fruits have a round shape while potato tubers can have both a round shape as well as an elongated shape. The candidate gene *Solyc10g076180* (SlOFP20, a member of the OVATE Family Protein (OFP)) on chromosome 10 of the reference tomato genome (Heinz 1706) is responsible for round fruits. However, this gene does not have an ortholog in the reference potato genome (DM), which has very elongated tubers [138]. In this use-case, first, we query and compare the QTL regions on chromosome 10 in tomato and potato. This QTL region is associated with round shape in tomato fruits and predominantly elongated shape in potatoes. We classify these genes in 3 tables a) genes that are unique in tomato b) genes that are unique in potato c) genes that are common in both tomato and potato (see Table 4.1). Further, we check the GO annotations as well as orthologs in all these genes. 3 genes (*Solyc10g076170.1*, *Solyc10g076190.1*, *Solyc10g076180.1*) are

present in class a) which indicates that they are unique in tomato. Out of these 3 genes; *Solyc10g076170.1* is an obsolete entry in and has been removed from UniProt and other databases. *Solyc10g076190.1* is a peroxidase gene (that is also common in class b) and class c)) in our table and *Solyc10g076180.1* is the only unique ovate family gene member in this pool. Genes in class b) are all peroxidase genes and class c) contains both some peroxidases and a lipid transport gene.

It is clear from our analysis that the candidate gene *Solyc10g076180.1* does not have any database entry, of an ortholog in the potato DM reference genome in the same QTL region on Chromosome 10. However, with the help of our tool we tried to explore this further, and retrieve a homolog based knowledge network for our candidate gene. This homolog network is retrieved with a nested query, in which we first locate all paralogs of *Solyc10g076180.1* gene in the tomato genome and then find orthologs of these genes in the potato genome (see Figure 4.6). With the help of this nested query analysis, we were able to find 10 ovate OFP genes in potato, which were not mentioned as orthologs to *Solyc10g076180.1* in any database. Out of these 10 OFP genes in potato, *PGSC0003DMG400028155* is located on chromosome 10 in the region 56030393-56031156. However, this region is 6.7Mb away from the studied QTL region and thus seems unlikely to harbor the determinant candidate gene.

Table 4.1: **Table comparing genes in QTLs of (Tomato) fruit shape and (Potato) tuber shape.** Three classes represent (a) Genes unique in Tomato; (b) Genes unique in Potato (c) Genes mapped in both tomato and potato. Each row contains a geneID, GO annotations, orthologs inside QTL region, and orthologs outside the QTL regions. The query clearly depicts that only 3 genes are unique in tomato i.e *Solyc10g076190.1*, *Solyc10g076170.1*, *Solyc10g076180.1*. *Solyc10g076190.1* is a peroxidase genes, *Solyc10g076170.1* is an obsolete gene entry according to UniProt database and *Solyc10g076180.1* is the ovate gene responsible for roundness in tomatoes.

a) Genes uniquie in tomato

Tomoto Genes	GO annotations	Potato orthologs inside QTL region	Potato Orthologs outside the QTL region
Solyc10g076180.1	GO:0003677 [DNA binding]; GO:0045892 [negative regulation of transcription DNA-templated];	none	none
Solyc10g076190.1	GO:0004601 [peroxidase activity]; GO:0005576 [extracellular region]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding]; GO:0042744 [hydrogen peroxide catabolic process]; GO:0046872 [metal ion binding];	none	PGSC0003DMG400011948
Solyc10g076170.1	none	none	none

b) Genes unique in potato

Potato Genes	GO annotations	Tomato orthologs inside QTL region	Tomato Orthologs outside the QTL Region
PGSC0003DMG400006679	GO:0004601 [peroxidase activity]; GO:0005576 [extracellular region]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding]; GO:0042744 [hydrogen peroxide catabolic process]; GO:0046872 [metal ion binding];	none	none
PGSC0003DMG400006680	GO:0004601 [peroxidase activity]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding];	none	none
PGSC0003DMG400006681	GO:0004601 [peroxidase activity]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding];	none	none
PGSC0003DMG400020795	GO:0004601 [peroxidase activity]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding];	none	none

c) Genes mapped in both tomato and potato

Tomato Genes	GO annotations	Potato orthologs inside QTL region	Potato Orthologs outside the QTL Region
Solyc10g076200.1	GO:0006869 [lipid transport]; GO:0008289 [lipid binding]; GO:0016020 [membrane];	PGSC0003DMG400040954	PGSC0003DMG400011955
Solyc10g076210.1	GO:0004601 [peroxidase activity]; GO:0005576 [extracellular region]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding]; GO:0042744 [hydrogen peroxide catabolic process]; GO:0046872 [metal ion binding]	PGSC0003DMG400020799; PGSC0003DMG400020800	none
Solyc10g076220.1	GO:0004601 [peroxidase activity]; GO:0005576 [extracellular region]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding]; GO:0042744 [hydrogen peroxide catabolic process]; GO:0046872 [metal ion binding];	PGSC0003DMG400020799; PGSC0003DMG400020800	none
Solyc10g076230.1	GO:0004601 [peroxidase activity]; GO:0006979 [response to oxidative stress]; GO:0020037 [heme binding];	PGSC0003DMG400020798	none

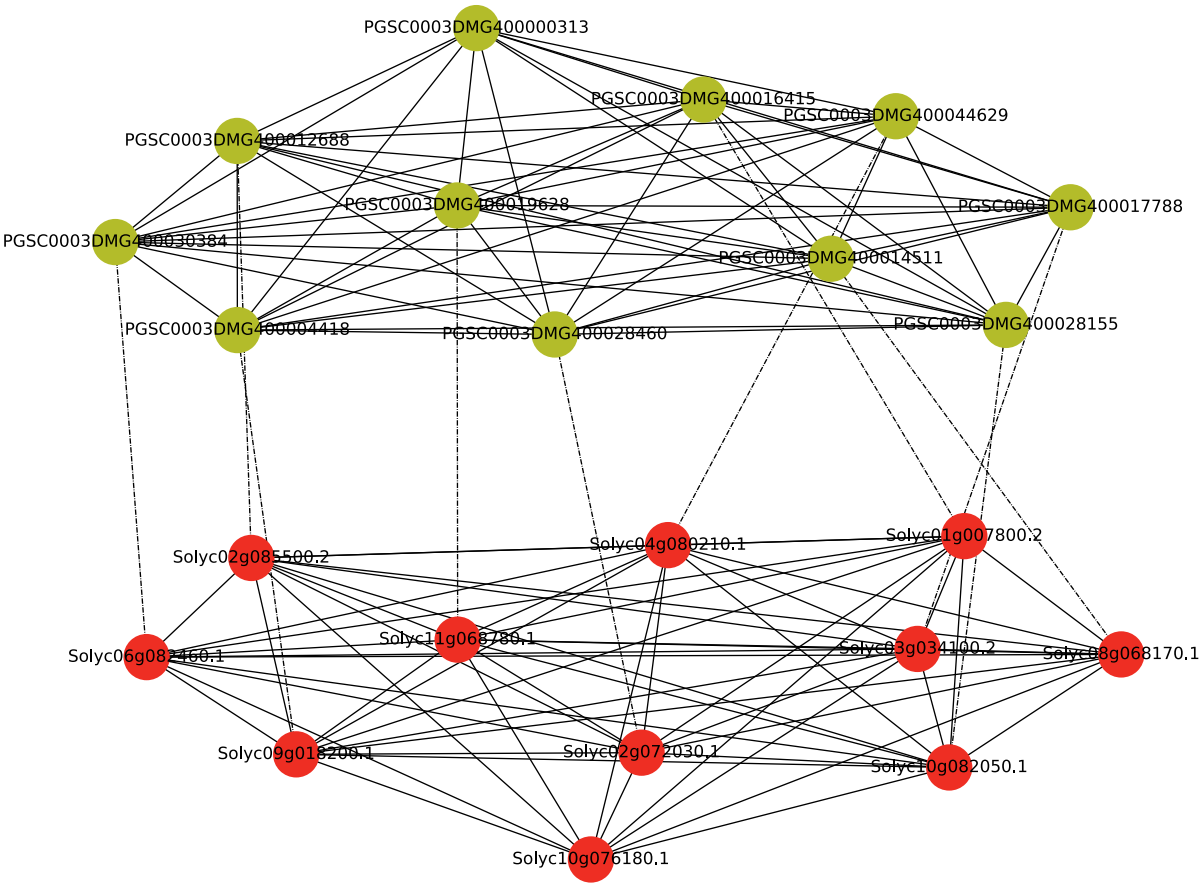


Figure 4.6: A knowledge network to represent all homologs of *Solyc10g076180.1* in tomato and potato. Here tomato genes are in red color, while potato genes are in green color. Solid edges represent paralogous relation, and dotted edges represent orthologous relation.

Discussion

The main objective of developing the pbg-ld platform was to improve the FAIRness of candidate gene identification in Solanaceae species by providing (semantically) integrated genomics and QTL datasets available in public resources (i.e., UniProt, Ensembl Plants, SGN and Europe PMC) from a central endpoint.

After selecting various datasets and information relevant to candidate gene discovery, a critical step in the knowledge discovery process is the transformation of data into a suitable data infrastructure. Biological data is complex and highly connected, for example e.g. there is huge ambiguity in the names of genes, proteins, and transcripts, hence semantic model with correct identifiers is required to differentiate them. Pbg-ld addresses the challenges of providing a semantic layer over most used datasets for candidate gene discovery in tomato and potato. A critical step in our approach was the transformation of (semi-)structured or non-RDF data sources to inter-linked RDF graphs using existing and newly developed tools such as the QTM and SIGA.py. Further, FDP provides meta-data explanation, which makes the user aware of the originating graph(s) to perform queries and interpret the result. Lastly, different data access points provide flexibility for users who wish to analyze and/or visualize data on this platform.

Data sets are not static and constantly emerging over time. Pbg-ld combines open data from different third party resources, like EuropePMC, SGN, UniProt and Ensemble Plants. However, as these data sets are not static a significant improvement in Pbg-ld could be to automate the process of regular updates of the data sets. Another improvement in Pbg-ld could be to provide visualization of our interoperable data graphs as knowledge graphs. Currently, Pbg-ld uses Openlinks virtuoso universal server's faceted browser to browse RDFgraphs from the RDF store. A data graph visualization that includes, all relevant information available from the literature and databases, about a particular entity, can be a nice user friendly tool to browse the information of interest.

To conclude, Pbg-ld is an integrated resource for Solanaceae species that provides access to available knowledge about genome annotations in public databases and scientific literature in a more robust way. This resource aids in the identification of candidate genes for complex traits using available knowledge in the databases and literature.

Chapter 5

Prediction of candidate genes with QTL regions for tomato, using pbg-ld, functional annotations and evolutionary analysis

Gurnoor Singh [1], Matthijs Brouwer [1], Arnold Kuzniar [2], Yury M Tikunov [1], Arnaud G. Bovy, Richard Finkers [1], Richard G.F. Visser [1]

[1] Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands
[2] Netherlands eScience Center (NLeSC), Amsterdam, the Netherlands

to be submitted

Abstract

Introduction

Predicting candidate genes for QTL regions is a key objective in plant genetics and breeding. However, a single QTL region can contain many genes. Mining candidate genes from such a QTL region could be done using existing knowledge of structural and functional gene annotations. Here we present a seamlessly integrative workflow for predicting candidate genes for metabolic traits of tomato within QTL regions using our linked-data platform Pbg-ld, gene prediction algorithms that use functional annotations and evolutionary analysis.

Results

We test this workflow on 5 QTLs related to metabolic traits in tomatoes. The workflow was able to correctly predict candidate genes for 4 out of the 5 selected QTLs. Hence, this use-case is an exemplary proof of concept study for using such linked-data workflows for predicting candidate genes in QTL regions using available knowledge.

Source code and Data: `https://github.com/PBR/QTL-prioritisation`

Keywords: Prioritization of Candidate Genes, Linked data, Plant Breeding, QTLs

Introduction

Tomato is one of the most consumed fruits in the world. The metabolic composition of a tomato is directly associated with its nutritional value, taste, aromas, and quality [49]. Metabolomic research studies in the past, have been able to predict functional characteristics of metabolites in the life-cycle of a plant. For example, volatiles are known to play an important role in the defense mechanism of plants against pathogens, where they serve as airborne signaling molecules to induce a defense response in other plant parts or neighboring plants [139]. Similarly, other metabolites like soluble solids (glucose, fructose, and sucrose) contribute to the sweetness of a fruit [140]. Lycopene is a carotenoid compound found in tomatoes which contributes to the nutritional value of a tomato and the red pigment in tomatoes responsible for fruit-color [141]. Similarly, terpenoids play a role in attracting pollinators [142].

While functional genomics research studies have tried to assess the structure and function of genes and proteins that relate to the expression and concentration of metabolites in plants, QTL-mapping research studies try to identify the genomic locations which associate with the changes in the expression or concentration of a metabolite (the trait of interest). However, it is still challenging to predict a causal or candidate gene which is directly associated with the trait of interest. The size of a QTL region can vary enormously depending on the number of markers used and the genome size of the plant under investigation but easily can range from hundreds of kb to several Mb [143] and a single QTL region can contain very many genes [107]. One way to mine candidate genes from a QTL region could be done using the existing knowledge of structural and functional annotations of genes. These genome annotations are mostly available across multiple databases and file formats (e.g., in the Generic Feature Format or GFF), which hampers integrated data analyses. Linked data approaches and semantic web technologies should be used to integrate heterogeneous knowledge [144].

Big-data utilities like Solanaceae-centric Linked Data platform (Pbg-ld) provides an interface to query genotypic and phenotypic data using available knowledge for tomatoes and potatoes [145]. Pbg-ld contains knowledge from both unstructured data of scientific literature and structured data of publicly available biological databases. Pbg-ld data types include QTL data from EuropePMC [17], genomic annotations (i.e., gene models, genetic markers and proteins) from the Solanaces Genomic Network [146], Ensembl Plants [35] and UniProt [34], and domain specific ontologies like Gene Ontology, Trait Ontology, Sequence Ontology and Solanace Phenotypic Ontology. Further, pbg-ld allows easy data access by faceted browser (user friendly interface), SPARQL queries and restful API. Data in pbg-ld is published according to FAIR Data Principles [45] to increase the degree of discoverability and (re-)usability of the data. Therefore, pbg-ld supports genomic analysis to predict candidate genes for complex traits in Solanaceous species.

Finding candidate genes within QTL regions for the trait of interest, using computational approaches is a major challenge in plant bioinformatics. Several tools have been developed in the past that tried to prioritize candidate genes based on existing knowledge.

QTLSearch is a software tool that searches for candidate causal genes in QTL studies by combining Gene Ontology annotations across many species and leveraging hierarchical orthologous groups [147]. QTG-Finder is a recently published article that uses a machine learning model to prioritize candidate genes in using function annotation, co-function network, and paralog copy number [148]. However, both these tools have been developed in *Arabidopsis thaliana* and rice (*Oryza sativa*) and it is difficult to use and test these tools in other species like tomato and potato. QTLSearch uses the HOGProp algorithm that requires access to hierarchical orthologous groups available in OMA browser [149], to score candidate genes based on trait-related GO terms. As the OMA browser data graph is cross-referenced in Uniprot, which is part of pbg-ld, the QTLSearch algorithm can be tested for tomato/potato data with the help of pbg-ld.

This study aimed to develop, illustrate, and analyze a seamlessly integrative workflow that uses linked genomic-data and prioritization pipelines to predict candidate genes within QTL regions for metabolic traits of tomato.

Workflow

Figure 5.1 illustrates a prediction to candidate genes workflow within a QTL region for the trait of interest, using function annotations and evolutionary genomics data. Input to this workflow is either a QTL region (containing physical location or a genetic location) or a trait of interest. If the input parameter is a trait of interest, the Pbg-ld database retrieves all QTL locations for that trait in tomato. This QTL information occurs in tables of scientific literature and is compiled in Pbg-ld using the QTLTableMiner++ (QTM) tool [61].

After receiving the QTL inputs, this workflow queries the set of all genes occurring within this QTL region. For every gene, this workflow retrieves a set of all GO terms as well as all orthologs and paralogs of these genes. These genome annotations are served as input to the QTLSearch pipeline which uses the Hogprop algorithm. HogProp algorithm assigns scores to every Gene within a QTL region. It uses GO terms which relate to the trait of interest and GO terms which relate to the genes within a QTL region to assess the distance of these functional annotations along gene phylogenies. This workflow has been developed in a (IPython) Notebook. It is a modular framework in which data is fetched from multiple data sources and can also accommodate new analysis modules as they are being developed by our group or the scientific community.

Test-case for QTLs related to metabolic traits

To test the usability of our workflow in predicting candidate genes for metabolic traits, we selected 5 QTLs for different metabolic traits (See Table 5.1). Out of the selected 5 QTLs for metabolic traits, 3 QTLs which relate to the following traits, soluble solids, lycopene beta-cyclase activity and phenolic compounds (2-phenylethanol, phenylacetaldehyde),

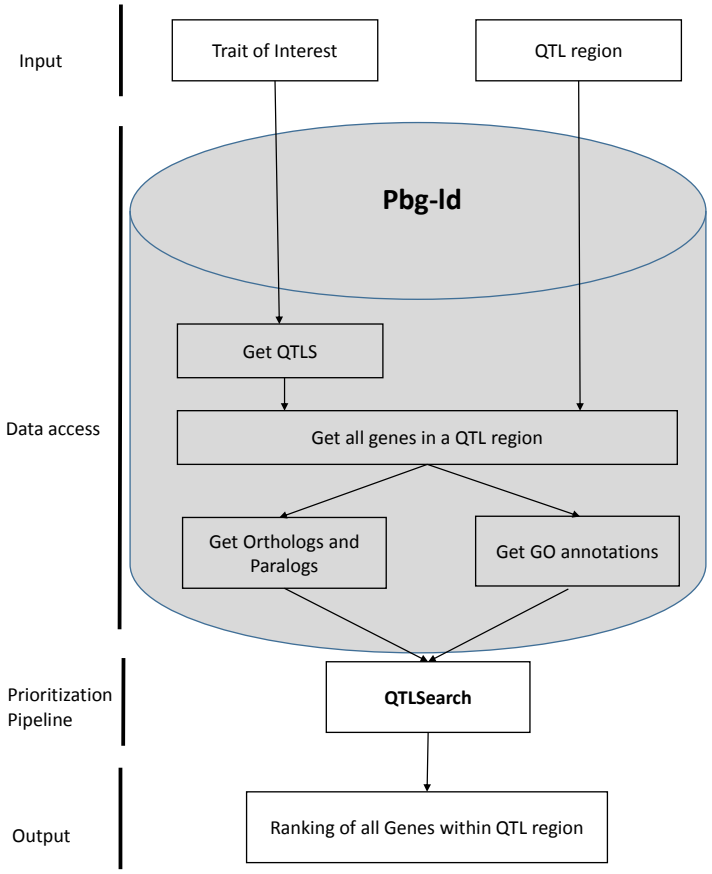


Figure 5.1: A workflow to predict candidate genes in a QTL region for the traits of interest

have known candidates. Further, these candidate genes are already annotated with related GO terms in publically available databases like UniProt. While, for one of the QTL regions which relates to terpenoids, Terpene synthase is a known candidate gene, however, Terpene synthase is not annotated with GO terms which show an association with the trait of interest (i.e. terpenoids). The 2 GO terms related to this gene are DNA binding and DNA methylation. Lastly, for the QTL region selected for volatile compounds (i.e. 3-methylbuthanal, 3-methylbuthanol) there no well-known candidate genes in the QTLs that had experimentally proven significance.

Table 5.1: **A selected set of 5 QTLs for metabolic traits in tomatoes.** These 5 QTLs are used as test-cases to analyse the prediction power of the underlying workflow

Traits of interest	GO annotations	Chromosome	Location	Candidate Genes	References
Total soluble solids (Brix)	GO:0006094, GO:0046370, GO:0046369, GO:0005985, GO:0015770	9	3474710	Lin5	[150]
Carotenoid compounds (Lycopene beta-cyclase activity)	GO:0045436, GO:0016117	6	Solyc06g073470 Solyc06g083850.3	Soly06g074240.1	[151]
Polyphenolic compounds (2-phenylethanol & phenylacetaldehyde)	GO:0016747, GO:0102387, GO:0018449, GO:0004029, GO:0008957, GO:1990055, GO:0050177, GO:0018814	8	55068565-63267130	LePAR	[152]
Terpenoid compounds	GO:0003677, GO:0045893	1	86142248-86467672	Terpense synthase	[142]
Volatile compounds (3-methylbuthanal, 3-methylbuthanol)	GO:0046568, GO:0018455, GO:0052676	3	69685329-71362039	?	?

Results & Discussion

This section summarizes the results and discusses the usability of our approach in detecting candidate genes with QTLs regions for the following traits of interest.

Total Soluble solids

One of the most extensively studied metabolic traits in tomato is the total soluble solids content in fruits (i.e. TSS or Brix) [150]. Brix is mainly made of Glucose, Fructose, and Sucrose and therefore, we selected 5 GO terms related to these metabolites and Brix trait (see Table 1) and fed it to the QTLSearch pipeline in our workflow. Previously known studies have identified multiple QTLs that are associated with the Brix trait in tomatoes [134]. Out of the many know QTL locations, the most significant QTL is located

of chromosome 9, containing Lin5 as the popularly known candidate gene for Brix trait [150]. Table 5.2 highlights the top 10 genes predicted from our workflow for the Brix QTL region from 3374710 to 3574710 on Chromosome 9. Lin7 and Lin5 were the top predicted genes related to this trait. Both these genes are from a homologous family and are known to be associated with Brix. In the list of top 10 genes, these 2 genes score significantly higher than all other genes. We conclude that our pipeline performed well to predict the candidate genes for this QTL.

Table 5.2: Top 10 candidate genes found for the Brix / soluble solids trait on Chromosome 9 by QTLSearch

Gene ID	Alias	UniProt ID	Protein Description	chromosome number	Location	Prioritization score
Solyc09g010090.2	LIN7	Q8L4N2	Cell-wall invertase	9	3480545-3484159	0.202943
Solyc09g010080.2	lin5	Q9LD97	Beta-fructofuranosidase insoluble isoenzyme 1	9	3475480-3479343	0.172502
Solyc09g010020.2	-	K4CR31	1-aminocyclopropane-1-carboxylate oxidase	9	3447416-3449839	0.043606
Solyc09g010040.1	101248415	K4CR33	1-aminocyclopropane-1-carboxylate oxidase	9	3454295-3455561	0.043606
Solyc09g010000.2	101249481	K4CR29	1-aminocyclopropane-1-carboxylate oxidase	9	3444303-3445806	0.037065
Solyc09g009900.2	-	K4CR19	Pollen-specific protein	9	3382755-3386531	0.034483
Solyc09g009910.2	101252320	K4CR20	Lipid A, ABC transporter permease	9	3386287-3395367	0.034483
Solyc09g009920.1	101252019	K4CR21	F-box family protein	9	3397388-3398827	0.034483
Solyc09g009930.1	101251713	K4CR22	Unknown Protein	9	3400106-3400900	0.034483
Solyc09g009940.2	101251215	K4CR23	Signal recognition particle protein	9	3401811-3408955	0.034483

Table 5.3: Top 10 candidate genes found for the lycopene beta-cyclase by QTLSearch

Gene ID	Alias	UniProt ID	Protein Description	chromosome number	Location	Prioritization score
Solyc06g074240.1	NSY	K4C9E2	Chromoplast-specific lycopene beta-cyclase	6	45898227-45899723	7.399091
Solyc06g073570.2	101245261	K4C976	Cytochrome P450	6	45361777-45364885	0.554559
Solyc06g076160.2	101248306	K4C9X6	Cytochrome P450	6	47289151-47291972	0.471375
Solyc06g082730.2	101262224	K4CAC8	Cytochrome P450	6	48445567-48448356	0.340569
Solyc06g074420.1	-	K4C9G0	Cytochrome P450	6	46053420-46054877	0.209152
Solyc06g076520.1	HSP17.7	Q9SYU8	class I heat shock protein	6	47546790-47547254	0.003917
Solyc06g076540.1	101266525	K4CA14	class I heat shock protein	6	47551057-47551521	0.003917
Solyc06g076560.1	HSP17.6	Q9SYV0	class I heat shock protein	6	47559714-47560178	0.003917
Solyc06g076570.1	Sl20.0shsp	G5DGD4	class I heat shock protein	6	47564101-47564565	0.003917
Solyc06g075260.2	101257482	K4C9P1	Vicilin	6	46706558-46708609	0.003330

Carotenoid compounds

Carotenoids compounds are the primary determinants of tomato fruit color [153]. Carotenoids exert a broad range of functions which associate to photosynthesis, the formation of pigments, antioxidant activities, and being precursors to signaling molecules, including volatiles [51]. Lycopene is a major carotenoid in tomato [154]. Lycopene occurrence with a matrix of many bioactive components, like vitamin C, vitamin E, other carotenoids (a-carotene, beta-carotene, gamma-carotene, lutein), and flavonoids are associated with the color of a tomato. Lycopene beta-cyclase is a key enzyme occurring at the branch point of the carotenoid biosynthesis pathway and responsible for converting lycopene to beta-carotene. Lycopene beta-cyclase activity is also related to the total carotenoid content accumulated in the tomato fruit. The major QTL region which is related to Lycopene beta-cyclase activity is found to be located on Chromosome 6 between the region 45280179-49150528 [155]. Here we analyzed the prediction of candidate genes for Lycopene beta-cyclase activity with the help of our developed workflow. 2 GO terms that relate to lycopene beta-cyclase activity were selected for inclusion in our workflow. Previous known studies suggest that lycopene cyclase (LCY) is a known candidate gene related to this trait. In some databases, lycopene cyclase (LCY) is also annotated as neoxanthin synthase (NSY) as these are genes that are closely related carotenogenic enzymes belonging to the same family. Table 5.3 shows the results from the workflow, containing the top 10 genes predicted for this QTL region on Chromosome 6. NSY was ranked at the top of the list and has a score significantly higher than all other genes. Here also, we can conclude that our workflow performed as expected.

2-phenylethanol & phenylacetaldehyde

Phenolic derivative compounds like 2-phenylethanol, phenylacetaldehyde have a great impact on the aroma of a tomato [152]. Several QTL locations related to phenolic compounds have been identified in the past out of which, a major QTL region on chromosome 8 mapped by the markers, TG330-CT77 and TG330-CT148 is associated with the accumulation of 2-phenylethanol and phenylacetaldehyde (having genomic coordinated 55068565-63267130) [156]. Additionally, two putative proteins, 2-phenylacetaldehyde reductases proteins (LePAR1 and LePAR2) are known candidates, which catalyze the conversion of 2-phenylacetaldehyde to 2-phenylethanol [157]. Both these proteins are members of a reductase/dehydrogenase family. Table 5.4 illustrates the top 10 genes predicted from our pipeline for these phenolic compounds on Chromosome 8. *Solyc08g068190.2* was the top predicted gene related to this trait. Although we are not sure if this gene is the same as LePAR1 and LePAR2, this gene belongs to the same aldehyde dehydrogenase family. Therefore our workflow could detect the causal gene within this QTL region.

Terpenoids

A major QTL related to Terpenoids has been mapped on chromosome 1 with the genomic coordinates of 86142248-86467672 [158]. Proteins of the Terpene synthase (TPS) family and TPS gene are the expected candidate genes associated with Terpenoids. 5 of the TPS-a subclade genes (TPS31, TPS32, TPS33, and TPS35) occur in close proximity within this QTL. Table 5.5 highlights the top 10 genes predicted from our pipeline for this QTL. Our results suggest that the gene *Solyc01g095030.2*, which is a MYB transcription factor, is the causal gene for this QTL region. This is possibly the wrong prediction. The reason for our pipeline to give a wrong prediction here could be that there is no term present in the GO ontology that directly related to Terpenoids. Further, because of this missing GO annotation terms, Terpene synthase is not been well annotated with its function, which makes it difficult for our workflow to detect it as a high ranking gene for the Terpenoids trait.

Volatile compounds (3-methylbuthanal, 3-methylbuthanol)

Volatile compounds like 3-methylbuthanal, 3-methylbuthanol influence the flavor, sensory changes, and defense mechanism of tomato fruits [159]. A major QTL related to the volatile compounds (3-methylbuthanal, 3-methylbuthanol) has been mapped on chromosome 3 with the genomic coordinates of 69685329-71362039 [160]. However, it is not known which candidate gene in this QTL is responsible for changes in the concentration of these volatile compounds. Our results suggest that the lactate dehydrogenase (LDH) gene is possibly a candidate gene in this trait.

Out of the total 5 QTLs, our workflow performed significantly well in detecting candidate genes for the QTLs of soluble solids, lycopene beta-cyclase activity, and phenolic compounds. Our workflow did not perform well in the detection of candidate genes within the QTL for terpenoids on chromosome 1. This is most probably due to the fact that this QTL region is not well annotated, and there are no GO terms related to Terpenoids. Lastly, our workflow predicts a candidate gene called LDH, for the previously unknown QTL region associated with volatile compounds.

QTLSearch, a prediction pipeline for candidate genes in QTL regions is based on existing knowledge and evolutionary data (orthologs and paralogs). While the performance of QTLSearch is high with well-annotated data, it fails to perform well in detecting candidate genes for QTL regions where little is known. Hence, it's still very challenging to infer about candidate genes with a less annotated QTL region.

Conclusions

Linked-data platforms like Pbg-ld can help in accessing, querying and analyzing genomic data for Solanaceous species. This platform provides access to available knowledge about

genome annotations in public databases and scientific literature. This tool can robustly be used with other candidate-gene prediction pipelines.

Table 5.4: Top 10 candidate genes found for the 2-phenylethanol, phenylacetaldehyde by QTLSearch

Gene ID	Alias	UniProt ID	Protein Description	chromosome number	Location	Prioritization score
Solyc08g068190.2	101257095	K4CM43	Aldehyde dehydrogenase	8	57303048-57306002	3.702432
Solyc08g076790.2	101246651	K4CN39	Cinnamoyl-CoA reductase-like protein	8	60704895-60707948	0.009002
Solyc08g068600.2	101264847	K4CM83	Decarboxylase family protein	8	57730921-57733032	0.004473
Solyc08g068610.2	778255	K4CM84	Decarboxylase family protein	8	57740004-57742160	0.004473
Solyc08g068620.1		K4CM85	Decarboxylase family protein	8	57747891-57749811	0.004473
Solyc08g068630.2	101265155	K4CM86	Decarboxylase family protein	8	57763544-57765666	0.004473
Solyc08g068640.2	101265757	K4CM87	Decarboxylase family protein	8	57774707-57776533	0.004473
Solyc08g068670.2	101265461	K4CM89	Decarboxylase family protein	8	57798879-57800980	0.004473
Solyc08g068680.2	AADC1A	Q1KSC6	Decarboxylase family protein	8	57812621-57814771	0.004473
Solyc08g068690.1	101266245	K4CM91	N-acetyltransferase	8	57820371-57821093	0.003834

Table 5.5: Top 10 candidate genes found for the terpenoids by QTLSearch

Gene ID	Alias	UniProt ID	Protein Description	chromosome number	Location	Prioritization score
Solyc01g095030.2	101257705	K4AZP3	MYB transcription factor	1	86401425-86409205	20.423493
Solyc01g094820.2		K4AZM2	ARID/BRIGHT DNA-binding domain-containing protein	1	86227211-86231284	3.715449
Solyc01g094800.2	101245796	K4AZM0	Chromodomain-helicase-DNA-binding protein 1	1	86208090-86220086	1.800694
Solyc01g094760.2	101260082	K4AZL6	Origin recognition complex 5 subunit	1	86177893-86180953	1.258382
Solyc01g094810.2		K4AZM1	Ubiquitin-conjugating enzyme	1	86221218-86224535	0.241966
Solyc01g095080.2	ACS2	P18485	1-aminocyclopropane-1-carboxylate synthase	1	86453267-86456196	0.145123
Solyc01g095010.2	101256219	K4AZP1	Diacylglycerol kinase catalytic region	1	86376984-86382216	0.096375
Solyc01g094930.2	101247858	K4AZN3	CAAX prenyl protease 1	1	86329801-86339860	0.060635
Solyc01g094700.2	101254424	K4AZL1	Glycerol-3-phosphate acyltransferase 4	1	86144883-86147820	0.026316
Solyc01g094770.2		K4AZL7	Unknown Protein	1	86182420-86186401	0.026316

Table 5.6: Top 10 candidate genes found for the Volatile compounds by QTLSearch

Gene ID	Alias	UniProt ID	Protein Description	chromosome number	Location	Prioritization score
Solyc03g122130.2		K4BN11	L-lactate dehydrogenase	3	70079537-70082001	0.008994
Solyc03g122140.2	101255867	K4BN12	L-lactate dehydrogenase	3	70082466-70085406	0.008994
Solyc03g122170.2		K4BN15	L-lactate dehydrogenase	3	70091655-70095510	0.008994
Solyc03g122320.1	104646525	K4BN30	Unknown Protein	3	70188302-70188649	0.007054
Solyc03g122330.1		K4BN31	Unknown Protein	3	70191124-70191474	0.007054
Solyc03g121600.2		K4BMV9	Choline dehydrogenase	3	69697420-69700296	0.006757
Solyc03g121610.2	101266497	K4BMW0	Receptor-like kinase	3	69700738-69706470	0.006757
Solyc03g121620.1	101266204	K4BMW1	Harpin-induced protein-like	3	69713500-69714303	0.006757
Solyc03g121630.2	101265907	K4BMW2	Protein FAR1-RELATED SEQUENCE 4	3	69717526-69729866	0.006757
Solyc03g121640.2	101265626	K4BMW3	chaperonin	3	69735709-69743120	0.006757

Chapter 6

General discussion and future prospects

To feed the several billion people living on this planet, efficiency and precision in breeding, of new crop varieties having more yield, disease resistance, and stress tolerance is important [161]. Food demands are expected to increase by 50% in 2030 [162]. Traditional crop breeding activities are still based on phenotype selection, which is a laborious and time consuming effort cannot often generate new cultivars quickly in response to the required traits [163]. Recent advances in biotechnology and genomics data science have the potential to accelerate and precise breeding programs greatly. However, molecular and genomic data sets of a crop species are often distributed over multiple independent data sources and scientific literature, and thus there is a need to semantically collect, organize and integrate information from these different kinds of information resources. This research focuses on the objectives of integrating heterogeneous genomic data of Solanaceae species for knowledge discovery. To objectively prioritize the selection of candidate genes for the traits of interest, this research can help in formulating data-driven algorithms that search published evidence from the wealth of available data. Thereby, it addresses the needs of scientists to efficiently explore and compare the wealth of genetic, genomic, and phenotypic information available in the literature, and the biological databases. Combining this information narrows down the genomics regions (QTL regions or the list of candidate genes) which are associated with traits of interest. This study can be of great value in developing or improving future precision breeding programs.

Genomic knowledge discovery is often confronted by the challenges of data integration from a multitude of independent databases and research articles. For discovering candidate genes with the help of large scale data integration, there is a need to organize candidate data resources according to the FAIR data principals. The core development in this research provides a linked data platform that semantically organizes and integrates genotypic and phenotypic data on Solanaceae species according to these principals. This progress in digital science helps genomic datasets to be more findable, accessible, interoperable and reusable.

Complimentary to our developments, some similar plant-specific software and databases provide genotypic and phenotypic data sets in a semantically integrated way. KNETMiner is an open source software that integrates plant-specific biological data sets into a knowledge graph[164]. These biological data sets contain information related to genes, biological pathways, phenotypes and publications for many important crop species like wheat, barley, potato, tomato, maize, poplar, and brassica. Additionally, KNETMiner has an evidence-based gene ranking algorithm that ranks and visualizes this integrated data based on gene annotations. Although, KNETMiner provides integrated data sets for many important crops, the quality of data related to genome annotations for some crops like tomato, is not at par with the data published in legacy databases. Similarly, Planteome database [113] provides gene annotations and phenotypes with the help of reference ontologies such as PO, TO, GO and ChEBI. Planteome is a user-friendly tool to query traits of interest, germplasm, and putative candidate genes. However, it lacks QTLs, genetic markers and links to publicly available databases such as Ensembl Plants. Therefore, our linked data platform is a unique resource for Solanaceae species that

provides access to available knowledge about genome annotations in public databases and scientific literature.

Knowledge discovery from scientific articles is often hampered by the unstructured, human-readable form of textual information. Textual information is not machine-readable, hence, difficult for machines to process and analyze information from it. Natural language processing(NLP) or text mining, renders textual information to be computationally accessible. In this research, we focused on two key challenges of knowledge discovery from scientific articles. Firstly, we developed a supervised NLP model, to extract a knowledge network of genotypic-phenotypic associations from sentences of scientific articles. Secondly, we developed QTLTableMiner++(QTM), a table mining tool that extracts and semantically annotates QTL information from tables of these articles. A huge amount of scientific information is available only in scientific literature, this study is helpful to process and analysis information from scientific literature [165].

Another major achievement of this research is the mining and construction of, genotype-phenotype based knowledge networks from scientific literature with the help of IBM Watson. A knowledge network here is defined as a data structure of three data nodes, consisting of two biological entities and a relationship between them. With the help of these knowledge networks, we are able to capture heterogeneous, complex and interconnected biological information represented in scientific literature as structured data, which is easily readable by both humans and machines. Hence, the developed supervised NLP model to capture knowledge from scientific literature into knowledge networks can be used by the scientific community to convert unstructured data to structured data.

Lastly, data integration and data reusability are two different yet strongly interlinked concepts. Data integration is the concept to connect modular data nodes with each other for a particular study. Reusability is the quality of those modules, for being redeployable and reusable building blocks for other studies [166]. Pbg-ld, the linked data platform developed by us, provides data for seamless integration with other data sets and analytical pipelines. This is well illustrated in Chapter 5 of this thesis, where we use data from the Pbg-ld and the OMA browser to prioritize candidate genes within QTL regions. Hence, both the above mentioned concepts of data integration and data reusability, are addressed in this research study.

Future Prospects

Some directions in which future research developments can be focused upon are presented below.

Precision Agriculture and Farm data train:

The ultimate goal of this research and other similar researches in the agricultural section is to produce more crops using fewer resources while reducing any negative effects on the environment and society. Improvement in data handling and upcoming

technologies are resulting in the emergence of precision agriculture. Just like precision medicine, where doctors choose an optimal treatment for patients based on a combination of data sources (e.g., a DNA test and clinical data). In precision agriculture, a farmer would be able to choose an optimal strategy based on a concoction of various data studies (e.g., omics data, environmental conditions, weather conditions) or a breeder would be able to make better phenotypic calculations, with the help of additional data collection from the farm. For example, estimating yield stability from larger data sets of numerous fields is better than, estimating from a single controlled field experiment.

The biggest challenge towards this objective is that of data integration. FAIR data principles lay the foundation in which each data node should be published. However, an important step for making data FAIR is to define a set of metadata elements that formulate a set of minimal information checkpoints that a data provider should add to his or her data to be FAIR, for example, BrAPI defines minimum standards for plant breeding applications [167]. Meta-data should be rich, standardized and collected from the source in lines with FAIR principles. Once the meta-data is defined, collecting various forms of agriculture data becomes more feasible with a standardised and usable way. Farm Data Train is a research project that has a wider scope of establishing a shared infrastructure for connecting various information nodes of agricultural data, by making every data node FAIR. Decision making for managing a farm usually depends on data from various independent resources. This data is not reusable as it is produced and managed by either the manufacturing company of various farm machines or farmers itself. By making every data node FAIR, this research project establishes a distributed and federated infrastructure that could enable the use and the reuse of FARM data for the benefit of farmers and crop breeders. In the prototype architecture of this project, every data node is called a farm-station, which publishes data with FAIR data principles. The data-access rules in between two stations are called a farm-track, while farm-trains are the analytical pipelines that communicate between 2 farm-stations. In this way, the designed digital infrastructure ensures every data governing body to manage, safeguard, and share their data with other stakeholders.

Using k-mers for better prediction of candidate genes for traits:

Prediction of correct candidate genes or genetic loci that are associated with the traits of interest is an important issue for both plant breeders and researchers. Most methods which use read mapping, rely on the availability of a reference genome. However, using other data-sets like using Whole Genome Sequence (WGS) reads directly without alignment can also have several advantages. Many non-model crop species or complex polyploid species do not have a reference genome available yet. Additionally, even if a reference genome is available, the assembly quality, experimental conditions, or completeness of the reference could be sub-optimal and effect read mappings and genetic analysis in further experiments. Hence, reference genome based approaches for genetic analysis and crop improvement can cause limited and incomplete results [168]. Working directly

with WGS reads, for example use of subsequences (k-mers) instead of the full reads allows to bypass sequencing errors in the reads and allows unambiguous comparisons between samples thanks to their invariable length. This has been well illustrated in the prediction of Sen3 markers for wart disease resistance in potatoes [169], with the help of Comparative Subsequence Sets Analysis (CoSSA) workflows. Having WGS data from an experiment as linked-data graphs in Pbg-ld can make CoSSA workflows more robust, efficient and combinable with other knowledge base data sets.

Machine-readable scientific articles prepared for NLP:

An important step to improve NLP based knowledge discovery from scientific articles, is to use machine readable input file formats for these articles. This step provides data clean from the source and ready to be processed for NLP. Although scientific articles are distributed in PDF, it is inconvenient to use this file format for automated information extraction as they lack machine readability and a logical structure specifying which content constitutes a paragraph, table, figure, header, footer, etc. Therefore, even if massive amounts of unstructured data are held in the form of PDF documents, automated extraction of sentences, tables or figures is very challenging. Similarly, HTML web pages are also used for the distribution of scientific articles. However, HTML files only represent a layout of a web page and are not focused on describing data or the meta-data provided in this resource. QTM uses XML files as input as they represent information in a logical structure that is machine-readable. The Europe PMC repository provides full-text open access articles in the XML format. One drawback of using Europe PMC is that it is mainly focused on the biomedical literature while the plant literature is not covered extensively in this repository. However, publishers such as Springer Nature or Elsevier have released web based APIs, which provide access to their articles in XML format. Currently, these XML files only make the meta-data of a scientific article machine readable. A significant improvement in this could be to have biological entities pre-annotated in the text of these articles itself. This would facilitate knowledge mining over scientific literature, with the need of doing Named Entity Recognition over it. The more the input data is clean, semantic and machine readable from the source, the easier it is to work with it.

A key challenge in today's research is also the accessibility and sharing of supporting data for a research article [170]. In the future, scientific articles and their supporting data should be more machine-readable and FAIR. This will allow users of that article to analyze, test and query the results of an article in a better way.

Scaling up NLP based genotypic-phenotypic knowledge extraction:

In this research, IBM Watson was used in the development of a supervised NLP model which can extract knowledge networks containing genotypic-phenotypic associations related to potato tuber flesh color, from the scientific literature. However, due to limited time availability, only 1 supervised NLP model was trained with articles related to tuber-flesh color. Hence, this model is capable of hunting genotypic-phenotypic associations

related to tuber flesh color, but will not perform well in other plant species or traits. However, further training of these supervised NLP models should for sure overcome this problem and should enable the establishment of a system that can search for genotypic-phenotypic association for any trait across public databases, patents, and scientific literature. This system can be complementary to the Watson for drug discovery system, which mines drugs related associations with the help of many NLP models that search through vast amounts of textual data in the form of laboratory data, clinical reports, patents and scientific publications [171].

Using QTM algorithm over supplementary data of an article:

QTM tool works only on the tables from the main-text of an article. However, data related to QTL studies are also published in supplementary materials of an article. Therefore, it is important to extract data from QTL studies from supplementary materials as well. If supplementary tables of an article are also available in XML formats, QTM can easily address this challenge.

Including data from other reference genomes species in pbg-ld:

Pbg-ld currently contains the gene models and the genetic markers based GFF files of reference tomato genome (*S. lycopersicum*), wild tomato (*S. pennellii*), and the reference sequence potato (*S. tuberosum*). However, other than these genomes, the SGN database, for instance also includes GFF files from other Solanaceae and closely related genomes, such as reference genome of Pepper (*Capsicum annuum*) [172], Eggplant (*Solanum melongena L.*) [173] etc. Converting these GFF files to RDF graphs and adding then to the infrastructure of Pbg-ld improves its power to do more comparative genomics. Nevertheless, a big challenge in doing this type of comparative genomics using linked-data knowledge graphs is to have the mapping of entity (genes, proteins, etc) identifiers among these various species. There is a chance of having different identifiers in two related species, for example, genomes of the reference sequence potato *S.tuberosum* and wild type species *M6* have both been sequenced, however, both use different sets of identifiers and the mapping of genes between these species is not available. However, having the data about cross references in these genes for both the genomes can give us better insights into underpinning certain gene functions with comparative studies.

Automatic data production and updates in Pbg-ld:

Currently Pbg-ld combines open data from different third party resources, like EuropePMC, SGN, UniProt and Ensemble. A significant improvement in Pbg-ld could be to automate the data production and have regular updates. For example, QTL information in Pbg-ld, is extracted from tables of scientific articles in EuropePMC via the QTM tool. However, QTL data graphs are static and require manual updates. Ideally, the QTL data graph should be updated automatically, whenever a new article of QTL studies is published in EuropePMC. A researcher is always interested in retrieving the most newly

published scientific articles in the domain of his/her research interests. Further, introducing Pbg-ld user profiles and notifying users about new data sets with alert functionality to provide information about the most recently updated data sets can enhance usability substantially.

Plugin for visualization of semantic knowledge networks:

Pbg-ld uses Openlinks virtuoso universal server's faceted browser to browse RDF graphs from the RDF store. This faceted browser displays all semantic triple related to an entity on multiple webpages. However, genome biologists prefer to visualize existing knowledge extracted from published literature and databases, in the form of knowledge networks [174]. A Cytoscape [72] based plugin or a BioJS component [175] for having an interactive visualization of biological knowledge networks can be a big asset with the pbg-ld's faceted browser.

Conclusions

To conclude, this research provides knowledge discovery tools and an *in-silico* genomic data infrastructure, which integrates data from molecular databases and scientific literature. This research facilitates the prediction and prioritization of candidate genes for the traits of interests as well as contributes towards designing a precise and improved breeding program.

Supplementary results

Figures

Chapter 3: QTLTableMiner^{++}: semantic mining of QTLtables in scientific articles

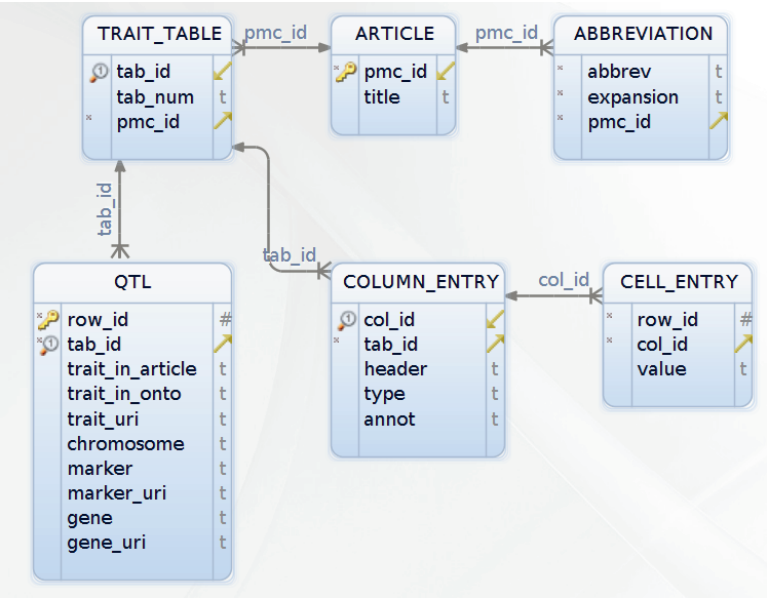


Figure 6.1: **ER diagram of the QTM database**

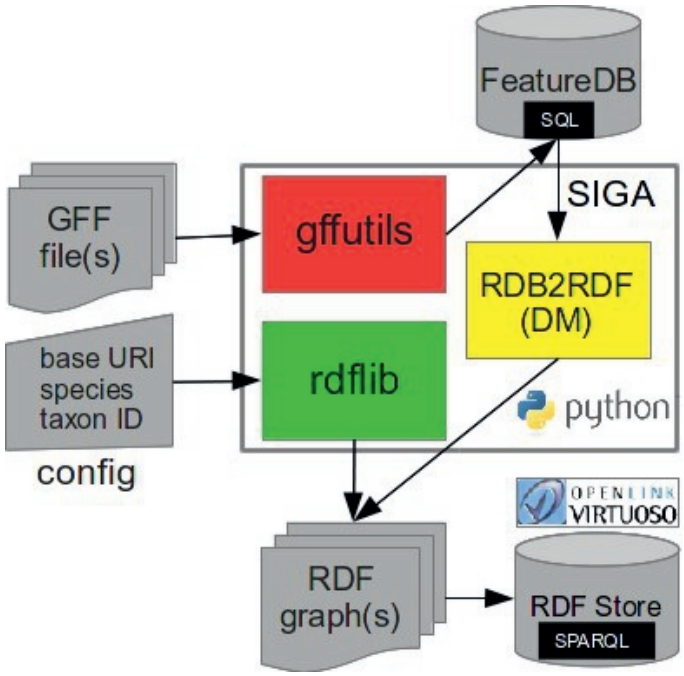


Figure 6.2: SIGA.py software architecture

Tables

Chapter 2: Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait

Table 6.1: **List of 34 articles used in training set for IBM Watson**

Index	article title	year	Reference
1	Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array.	2008	[176]
2	From QTL to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy.	2010	[177]
3	Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development.	1996	[178]
4	Temporal dynamics of tuber formation and related processes in a crossing population of potato (Solanum tuberosum).	2003	[179]
5	Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism.	2006	[50]
6	Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis.	2007	[180]
7	The Metabolic and Developmental Roles of Carotenoid Cleavage Dioxygenase4 from Potato.	2010	[181]
8	RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato.	1988	[182]
9	Flavonoid profiling and transcriptome analysis reveals new gene–metabolite correlations in tubers of Solanum tuberosum L.	2010	[183]
10	Silencing of beta-carotene hydroxylase increases total carotenoid and beta-carotene levels in potato tubers.	2007	[184]
11	Metabolic Engineering of Potato Carotenoid Content through Tuber-Specific Overexpression of a Bacterial Mini-Pathway.	2007	[185]
12	Metabolic engineering of high carotenoid potato tubers containing enhanced levels of beta-carotene and lutein	2005	[186]
13	Genetic Engineering of a Zeaxanthin-rich Potato by Antisense Inactivation and Co-suppression of Carotenoid Epoxidation.	2002	[187]
14	Overexpression of zeaxanthin epoxidase gene enhances the sensitivity of tomato PSII photoinhibition to high light and chilling stress.	2008	[188]
15	Carotenogenesis during tuber development and storage in potato.	2004	[189]
16	Regulatory control of high levels of carotenoid accumulation in potato tubers.	2011	[190]

Continued on next page

Table 6.1 – continued from previous page

Index	article title	year	Reference
17	Inheritance of Carotenoid Content in Tetraploid × Diploid Potato Crosses.	2011	[191]
18	Effect of cultivar, location and method of cultivation on the content of chlorogenic acid in potatoes with different flesh colour.	2013	[192]
19	Degradation kinetics and colour of anthocyanins in aqueous extracts of purple- and red-flesh potatoes (Solanum tuberosum L.).	2007	[193]
20	Effect of peeling and three cooking methods on the content of selected phytochemicals in potato tubers with various colour of flesh.	2013	[194]
21	Differences in anthocyanin content and antioxidant activity of potato tubers with different flesh colour.	2011	[195]
22	Effects of postharvest curing treatment on flesh colour and phenolic metabolism in fresh-cut potato products.	2015	[196]
23	Influence of flesh colour, year and growing area on carotenoid and anthocyanin content in potato tubers.	2013	[197]
24	Effect of natural and growing conditions on the contentof phenolics in potatoes with different flesh colour.	2010	[198]
25	Orange Flesh Trait in Potato: Inheritance and Carotenoid Content.	1993	[199]
26	Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six Solanum species.	2008	[200]
27	Genetic analysis of pigmented tuber flesh in potato.	2009	[201]
28	Inheritance of anthocyanin pigmentation in the cultivated potato: A critical review.	1991	[202]
29	Antioxidant activities, phenolic and β-carotene contents of sweet potato genotypes with varying flesh colours.	2007	[203]
30	Carotenoid Content and Color in Diploid Potatoes.	2001	[204]
31	Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase	2006	[205]
32	Enhancing beta-carotene content in potato by rnai-mediated silencing of the beta-carotene hydroxylase gene.	2007	[206]
33	Genome-wide QTL and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (Solanum tuberosum L.).	2014	[207]
34	The Incidence and Effect on Total Tuber Carotenoids of a Recessive Zeaxanthin Epoxidase Allele (Zep1) in Yellow-fleshed Potatoes.	2012	[208]

Table 6.2: **Confusion matrix for displaying the entity detection per article for the full training set of 34 articles. Here precision was 0.9765, recall was 0.8891 and F1 score was 0.9307**

Document ID	Total entity per article	TP	FP	TN	FN
10.1007/s10142-008-0083-x	188	182	1	0	5
10.1186/1471-2164-11-158	280	248	5	0	27
10.1046/j.1365-313X.1996.9050745.x	126	113	0	0	13
10.1111/j.1744-7348.2003.tb00284.x	289	275	2	0	12
10.1007/BF02872013	148	140	5	0	3
10.1007/s00122-007-0560-y	221	220	0	0	1
10.1104/pp.110.158733	303	294	5	0	4
ISBN: 0016-6731	38	32	0	0	6
10.1093/jxb/erp394	298	278	9	0	11
10.1186/1471-2229-7-11	200	179	2	0	19
10.1371/journal.pone.0000350	230	225	1	0	4
10.1093/jxb/eri016	293	288	3	0	2
10.1006/mben.2002.0234	253	249	1	0	3
10.1111/j.1399-3054.2007.01016.x	257	187	44	0	26
10.1093/jxb/erh121	292	283	6	0	3
10.1111/j.1365-3040.2011.02301.x	375	355	14	0	6
10.21273/JASHS.136.4.265	275	244	19	0	12
10.17221/460/2013-PSE	128	122	1	0	5
10.1016/j.foodchem.2005.11.002	143	99	0	0	44
10.1016/j.foodchem.2012.11.114	359	232	3	0	124
10.17221/265/2011-PSE	222	144	5	0	73
10.1016/j.foodchem.2014.08.011	257	206	10	0	41
10.1016/j.jfca.2013.07.001	271	247	2	0	22
10.17221/49/2010-PSE	168	126	3	0	39
10.21273/JASHS.118.1.145	195	142	3	0	50
10.1111/j.1439-0523.2008.01420.x	139	134	0	0	5
10.1007/s00122-009-1024-3	99	66	6	0	27
10.1007/BF02853712	23	20	0	0	3
10.1016/j.foodchem.2006.09.033	220	132	3	0	85
10.21273/JASHS.126.6.722	222	181	2	0	39
10.1186/1471-2229-6-13	272	218	0	0	54
10.1007/BF02986245	274	266	1	0	7
10.1007/s00122-014-2349-0	340	309	1	0	30
10.1007/s12230-012-9250-7	152	136	1	0	15
Total	7550	6572	158	0	820

Table 6.3: **Summary table of the single-year difference in connections between flesh color and its eventual neighbors.**

2009→2010		flesh color-like nodes						min
		flesh	flesh color	tuber flesh	tuber flesh color	white flesh color	yellow-orange color	
eventual direct neighbours to flesh color-like nodes	CCD	3→1	x→3	6→3	x→3	x→1	x→2	3→1
	CHY	2	x→1	5→3	x→2	x→3	x→3	2→1
	DXS	1	x→3	5→3	x→3	x→3	x→3	1
	PSY	1	x→3	5→3	x→3	x→3	x→2	1
	TP	3	x→5	7→4	x→5	x→4	x→4	3
	abscisic acid	1	x→3	5→2	x→3	x→2	x→3	1
	aminocyclopropane-1-carboxylic acid	1	x→4	5→4	x→4	x→3	x→3	1
	anthocyanin	3	x→4	1	x→5	x→5	x→5	1
	b-carotene hydroxylase	2	x→1	5→3	x→1	x→3	x→3	2→1
	bHLH	5→4	x→4	1	x→5	x→5	x→5	1
	carotenoid	1	x→2	4→2	x→2	x→3	x→2	1
	chlorophyll	1	x→3	5→3	x→3	x→3	x→3	1
	ethylene	3	x→5	7→5	x→5	x→4	x→1	3→1
	flavonoid	1	x→3	3	x→3	x→3	x→3	1
	flavonol	x	x	x	x	x	x	
	hydroxycinnamic acid	1	x→4	5→4	x→4	x→3	x→4	1
	lycopene	2	x→3	5→3	x→3	x→2	x→1	2→1
	lycopene e-cyclase	2	x→1	5→2	x→3	x→3	x→3	2→1
	phenolic	2	x→3	4→3	x→3	x→4	x→3	2
	phenylalanine ammonia lyase	x	x	x	x	x	x	
	zeaxanthin epoxidase	2	x→2	5→1	x→3	x→3	x→3	2→1

Table 6.4: **Confusion matrix for number of QTL tables of 30 articles in set 'tomato'. Here precision was 1 and recall was 0,985.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN
4321030	6	6	6	0	0	0
4948827	5	5	5	0	0	0
2652058	5	4	4	0	0	1
4498769	4	4	4	0	0	0
3852376	4	4	4	0	0	0
3464107	4	4	4	0	0	0
5395597	3	3	3	0	0	0
4266912	3	3	3	0	0	0
5181584	3	3	3	0	0	0
4726135	3	3	3	0	0	0
4987366	2	2	2	0	0	0
4872001	2	2	2	0	0	0
4999453	2	2	2	0	0	0
5281592	2	2	2	0	0	0
5209891	2	2	2	0	0	0
4612157	2	2	2	0	0	0
4969537	1	1	1	0	0	0
3859326	1	1	1	0	0	0
4831840	1	1	1	0	0	0
4661238	1	1	1	0	0	0
4737080	1	1	1	0	0	0
1913174	1	1	1	0	0	0
2246063	1	1	1	0	0	0
2271080	1	1	1	0	0	0
4008630	1	1	1	0	0	0
4579088	1	1	1	0	0	0
4301655	1	1	1	0	0	0
2657798	1	1	1	0	0	0
4034497	1	1	1	0	0	0
5145867	1	1	1	0	0	0
Total	66	65	65	0	0	1

Table 6.5: **Confusion matrix for number of abbreviations in tables of 30 articles in set 'tomato'. Here precision was 1 and recall was 0,545.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN
4321030	19	0	0	0	0	19
4948827	0	0	0	0	0	0
2652058	18	0	0	0	0	18
4498769	0	0	0	0	0	0
3852376	19	0	0	0	0	19
3464107	3	0	0	0	0	3
5395597	10	10	10	0	0	0
4266912	25	25	25	0	0	0
5181584	12	12	12	0	0	0
4726135	25	0	0	0	0	25
4987366	11	0	0	0	0	11
4872001	4	0	0	0	0	4
4999453	0	0	0	0	0	0
5281592	0	0	0	0	0	0
5209891	28	28	28	0	0	0
4612157	6	0	0	0	0	6
4969537	0	0	0	0	0	0
3859326	3	3	3	0	0	0
4831840	20	20	20	0	0	0
4661238	0	0	0	0	0	0
4737080	32	30	30	0	0	2
1913174	7	7	7	0	0	0
2246063	8	8	8	0	0	0
2271080	8	8	8	0	0	0
4008630	0	0	0	0	0	0
4579088	24	0	0	0	0	24
4301655	8	8	8	0	0	0
2657798	0	0	0	0	0	0
4034497	2	0	0	0	0	2
5145867	0	0	0	0	0	0
Total	292	159	159	0	0	133

Table 6.6: **Confusion matrix for number of biological concepts identified in tables of 30 articles in set 'tomato'. Here precision was 0,745 and recall was 0,926.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN	
4321030	2	19	16	16	0	0	3
3464107	1	6	12	6	6	0	0
5181584	2	15	5	5	0	0	10
4987366	1	17	2	2	0	0	15
2271080	2	3	1	1	0	0	2
3852376	2	19	17	17	0	0	2
2652058	3	5	1	1	0	0	4
2652058	5	8	0	0	0	0	8
4498769	5	9	6	6	0	0	3
4498769	6	7	4	4	0	0	3
2652058	4	8	0	0	0	0	8
4034497	2	12	9	9	0	0	3
4987366	2	8	3	3	0	0	5
5395597	2	14	7	7	7	0	0
5145867	3	16	2	2	0	0	14
5181584	3	11	5	5	0	0	6
4969537	2	11	0	0	0	0	11
4872001	2	8	7	7	0	0	1
4948827	4	6	0	0	0	0	6
3859326	3	19	17	17	0	0	2
4266912	3	14	11	11	0	0	3
4321030	4	34	31	31	0	0	3
4321030	4,1	34	32	32	0	0	2
3464107	4	42	42	42	0	0	0
3464107	4,1	42	42	42	0	0	0
4948827	3	9	4	4	0	0	5
3852376	4	70	66	49	17	0	4
5181584	4	15	9	7	2	0	6
4579088	2	90	19	19	0	0	71
5209891	3	54	50	0	50	0	4
4948827	6	20	0	0	0	0	20
4948827	5	22	0	0	0	0	22
5281592	3	90	48	48	0	0	42
Total		757	468	393	82	0	288

Table 6.7: **Confusion matrix for number of QTLs identified in tables of 30 articles in set 'tomato'. Here precision was 0,827 and recall was 0,577.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN	
4321030	2	0	19	0	19	0	0
3464107	1	0	32	0	32	0	0
5181584	2	0	14	0	14	0	0
4987366	1	0	19	0	19	0	0
2271080	2	0	3	0	3	0	0
3852376	2	0	19	0	19	0	0
2652058	3	4	4	4	0	0	0
2652058	5	4	4	4	0	0	0
4498769	5	5	5	5	0	0	0
4498769	6	6	6	6	0	0	0
2652058	4	6	6	6	3	0	1
4034497	2	6	6	6	0	0	2
4987366	2	6	6	6	0	0	0
5395597	2	7	7	7	0	0	
5145867	3	7	7	7	0	0	0
5181584	3	7	8	7	1	0	0
4969537	2	8	8	8	0	0	0
4872001	2	8	8	8	0	0	0
4948827	4	9	9	9	0	0	0
3859326	3	11	4	4	0	0	7
4266912	3	16	19	16	3	0	0
4321030	4	17	17	17	0	0	0
4321030	4	17	17	17	0	0	0
3464107	4	17	35	17	18	0	17
3464107	4	18	18	18	0	0	0
4948827	3	20	20	20	0	0	0
3852376	4	21	21	21	0	0	0
5181584	4	21	21	21	0	0	0
4579088	2	22	22	22	0	0	0
5209891	3	27	27	27	0	0	0
4948827	6	36	36	36	2	0	5
4948827	5	37	37	37	0	0	0
5281592	3	42	45	42	3	0	0
Total		405	529	398	136	0	32

Table 6.8: **Confusion matrix for number of QTL tables identified for 30 articles in set 'potato'. Here precision was 1 and recall was 0,972.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN
3607734	7	7	7	0	0	0
5526565	6	6	6	0	0	0
3023753	5	3	3	0	0	2
4551535	4	4	4	0	0	0
4773602	4	4	4	0	0	0
4799268	4	4	4	0	0	0
3037844	4	4	4	0	0	0
3460171	3	3	3	0	0	0
2358939	2	2	2	0	0	0
3546430	2	2	2	0	0	0
3660524	2	2	2	0	0	0
4199688	2	2	2	0	0	0
4374564	2	2	2	0	0	0
4448561	2	2	2	0	0	0
4480903	2	2	2	0	0	0
4632055	2	2	2	0	0	0
4648990	2	2	2	0	0	0
4703618	2	2	2	0	0	0
5567664	2	2	2	0	0	0
2676307	2	2	2	0	0	0
2639024	1	1	1	0	0	0
4354307	1	1	1	0	0	0
4404978	1	1	1	0	0	0
4510777	1	1	1	0	0	0
4777548	1	1	1	0	0	0
4855764	1	1	1	0	0	0
4900573	1	1	1	0	0	0
4996988	1	1	1	0	0	0
5345157	1	1	1	0	0	0
3347998	1	1	1	0	0	0
Total	71	69	69	0	0	2

Table 6.9: **Confusion matrix for number of abbreviations detected in tables of 30 articles in set 'potato'. Here precision was 1 and recall was 0,710.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN
3607734	2	0	0	0	0	2
5526565	18	9	9	0	0	9
3023753	16	12	12	0	0	4
4551535	0	0	0	0	0	0
4773602	32	32	32	0	0	0
4799268	1	1	1	0	0	0
3037844	7	0	0	0	0	7
3460171	11	0	0	0	0	11
2358939	0	0	0	0	0	0
3546430	12	12	12	0	0	0
3660524	0	0	0	0	0	0
4199688	0	0	0	0	0	0
4374564	0	0	0	0	0	0
4448561	6	6	6	0	0	0
4480903	19	16	16	0	0	3
4632055	4	0	0	0	0	4
4648990	9	9	9	0	0	0
4703618	12	12	12	0	0	0
5567664	9	9	9	0	0	0
2676307	15	0	0	0	0	15
2639024	0	0	0	0	0	0
4354307	0	0	0	0	0	0
4404978	0	0	0	0	0	0
4510777	0	0	0	0	0	0
4777548	0	0	0	0	0	0
4855764	21	21	21	0	0	0
4900573	0	0	0	0	0	0
4996988	5	0	0	0	0	5
5345157	8	8	8	0	0	0
3347998	0	0	0	0	0	0
Total	207	147	147	0	0	60

Table 6.10: **Confusion matrix for number of biological concepts identified in tables of 30 articles in set 'potato'. Here precision was 0,959 and recall was 0,355.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN	
4354307	2	10	0	0	0	0	10
4374564	5	0	0	0	0	0	0
4448561	4	7	4	4	0	0	3
4551535	5	6	5	5	0	0	1
4551535	6	5	2	2	0	0	3
4632055	3	6	2	2	0	0	4
4648990	1	20	8	8	0	0	12
4703618	2	22	0	0	0	0	22
4773602	6	10	0	0	0	0	10
4799268	4	5	0	0	0	0	5
4855764	4	43	25	22	3	0	18
4900573	7	20	13	13	0	0	7
5345157	2	14	6	6	0	0	8
5567664	3	20	8	8	0	0	12
3023753	3	12	0	0	0	0	12
Total		200	73	70	3	0	127

Table 6.11: **Confusion matrix for number of QTLs identified in tables of 30 articles in set 'potato'. Here precision was 0,828 and recall was 0,989.**

pmc_id	number of actual QTL Tables	number of predicted QTL tables	TP	FP	TN	FN	
4354307	2	13	13	13	0	0	0
4374564	5	5	5	5	0	0	0
4448561	4	6	6	6	0	0	0
4480903	3	37	76	37	39	0	0
4551535	6	6	6	0	0	0	0
4632055	3	4	4	4	0	0	0
4648990	1	12	12	12	0	0	0
4703618	2	22	22	22	0	0	0
4773602	6	10	10	10	0	0	0
4799268	4	5	5	5	0	0	0
4855764	4	43	43	43	0	0	0
4900573	7	14	12	12	0	0	2
5345157	2	7	7	7	0	0	0
5567664	3	11	11	11	0	0	0
3023753	3	1	1	1	0	0	0
Total		196	233	188	39	0	2

Table 6.12: **List of RESTful APIs of pbg-ld with the help of grlc**

Number	Endpoint (path)	Description	Response fields
1.	/countFeatures	Count genomic features given a (input) genome graph.	feature_id, feature_uri, feature
2.	/getGeneAnnotations	Get annotations from SGN given a (input) sgn_ and gene_id.	gene_uri, gene_annotations
3.	/getGenesInQTL	Get genes contained in a QTL, given (input) qtl_id	gene_id, gene_uri
4.	/getQTLinArticle	Get QTLs in an article, given (input) pmc_id.	qtl_id, qtl_uri
5.	/getQTLs	Get QTLs associated with a trait, given (input) trait_id.	qtl_id,qtl_uri
6.	/ getTraitIds	Get trait_ids given a (input) trait name.	trait_id, trait_term, trait_uri
7.	/summarizeQTLs	Summarize QTLs extracted from articles, given (input) pmc_id	qtl_id, associated_trait, chromosomal_location

References

[1] S. van Nocker and S. E. Gardiner, "Breeding better cultivars, faster: Applications of new technologies for the rapid deployment of superior horticultural tree crops," *Horticulture research*, vol. 1, p. 14 022, 2014.

[2] T. G. Consortium *et al.*, "The tomato genome sequence provides insights into fleshy fruit evolution," *Nature*, vol. 485, no. 7400, p. 635, 2012.

[3] P. G. S. Consortium *et al.*, "Genome sequence and analysis of the tuber crop potato," *Nature*, vol. 475, no. 7355, p. 189, 2011.

[4] R. G. S. P. International, "The map-based sequence of the rice genome.," *Nature*, vol. 436, no. 7052, p. 793, 2005.

[5] R. Appels, K. Eversole, C. Feuillet, B. Keller, J. Rogers, N. Stein, C. J. Pozniak, F. Choulet, A. Distelfeld, J. Poland, *et al.*, "Shifting the limits in wheat research and breeding using a fully annotated reference genome," *Science*, vol. 361, no. 6403, eaar7191, 2018.

[6] Y. Jiao, P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C.-S. Chin, *et al.*, "Improved maize reference genome with single-molecule technologies," *Nature*, vol. 546, no. 7659, p. 524, 2017.

[7] X. Wang, H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J.-H. Mun, I. Bancroft, F. Cheng, *et al.*, "The genome of the mesopolyploid crop species brassica rapa," *Nature genetics*, vol. 43, no. 10, p. 1035, 2011.

[8] S. Huang, R. Li, Z. Zhang, L. Li, X. Gu, W. Fan, W. J. Lucas, X. Wang, B. Xie, P. Ni, *et al.*, "The genome of the cucumber, cucumis sativus l.," *Nature genetics*, vol. 41, no. 12, p. 1275, 2009.

[9] L. Astola, H. Stigter, A. D. van Dijk, R. van Daelen, and J. Molenaar, "Inferring the gene network underlying the branching of tomato inflorescence," *PloS one*, vol. 9, no. 4, e89689, 2014.

[10] H. Shinozuka, N. O. Cogan, G. C. Spangenberg, and J. W. Forster, "Quantitative trait locus (qtl) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (lolium perenne l.)," *BMC genetics*, vol. 13, no. 1, p. 101, 2012.

[11] M. Kanehisa and P. Bork, "Bioinformatics in the post-sequence era," *Nature genetics*, vol. 33, no. 3s, p. 305, 2003.

[12] J. E. Risse, "Text mining for metabolic reaction extraction from scientific literature," 2014.

[13] M. D. Yandell and W. H. Majoros, "Genomics and natural language processing.," *Nature reviews. Genetics*, vol. 3, no. 8, pp. 601–10, Aug. 2002, ISSN: 1471-0056. DOI: 10.1038/nrg861. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12154383.

[14] Y. F.-Y. Edith Motschall, "Searching the medline literature database through pubmed: A short guide," *Onkologie*, vol. 28, pp. 517–522, 2005.

[15] D. A. LINDBERG, "Internet access to the national library of medicine," *Effective Clinical Practices*, vol. 3, no. 5, pp. 256–260, 2000.

[16] Z. Lu, W. Kim, and W. J. Wilbur, "Evaluation of query expansion using mesh in pubmed," *Information retrieval*, vol. 12, no. 1, pp. 69–80, 2009.

[17] The Europe PMC Consortium, "Europe PMC: a full-text literature database for the life sciences and platform for innovation," *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D1042–D1048, 2015, ISSN: 0305-1048. DOI: 10.1093/nar/gku1061. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383902/.

[18] M. Shultz, "Comparing test searches in pubmed and google scholar," *Journal of the Medical Library Association: JMLA*, vol. 95, no. 4, p. 442, 2007.

[19] M. Krallinger, A. Valencia, and L. Hirschman, "Linking genes to literature: Text mining, information extraction, and retrieval applications for biology," *Genome biology*, vol. 9, no. 2, S8, 2008.

[20] B. Müller, R. Klinger, H. Gurulingappa, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. M. Friedrich, "Abstracts versus full texts and patents: A quantitative analysis of biomedical entities," in *Information Retrieval Facility Conference*, Springer, 2010, pp. 152–165.

[21] S. Mukherjea and B. Bamba, "Biopatentminer: An information retrieval system for biomedical patents," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, VLDB Endowment, 2004, pp. 1066–1077.

[22] E. P. Office, *Open patent services (ops)*. [Online]. Available: https://ops.epo.org (visited on 08/01/2019).

[23] *Google patents*. [Online]. Available: `http://www.google.com/patents` (visited on 08/01/2019).

[24] *Fpo*. [Online]. Available: `http://www.freepatentsonline.com` (visited on 08/01/2019).

[25] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[26] P. Agarwal and D. B. Searls, "Literature mining in support of drug discovery," *Briefings in bioinformatics*, vol. 9, no. 6, pp. 479–492, 2008.

[27] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius, "Named entity recognition with combinations of conditional random fields," in *Proceedings of the second biocreative challenge evaluation workshop*, 2007.

[28] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: From information retrieval to biological discovery," *Nature reviews genetics*, vol. 7, no. 2, p. 119, 2006.

[29] M. Zimmermann, J. Fluck, L. T. Thi, C. Kolarik, K. Kumpf, and M. Hofmann, "Information extraction in the life sciences: Perspectives for medicinal chemistry, pharmacology and toxicology," *Current Topics in Medicinal Chemistry*, vol. 5, no. 8, pp. 785–796, 2005.

[30] S. Y. Rhee and B. Crosby, *Biological databases for plant research*, 2005.

[31] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers, "Genbank.," *Nucleic acids research*, vol. 46, no. D1, pp. D41–D47, 2018.

[32] P. W. Harrison, B. Alako, C. Amid, A. Cerdeño-Tárraga, I. Cleland, S. Holt, A. Hussein, S. Jayathilaka, S. Kay, T. Keane, *et al.*, "The european nucleotide archive in 2018," *Nucleic acids research*, vol. 47, no. D1, pp. D84–D88, 2018.

[33] Y. Kodama, J. Mashima, T. Kosuge, and O. Ogasawara, "Ddbj update: The genomic expression archive (gea) for functional genomics data," *Nucleic acids research*, vol. 47, no. D1, pp. D69–D73, 2018.

[34] S. Pundir, M. J. Martin, and C. O'Donovan, "Uniprot protein knowledgebase," in *Protein Bioinformatics*, Springer, 2017, pp. 41–55.

[35] D. M. Bolser, D. M. Staines, E. Perry, and P. J. Kersey, "Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomic data," in *Plant Genomics Databases*, Springer, 2017, pp. 1–31.

[36] A. Nakaya, H. Ichihara, E. Asamizu, S. Shirasawa, Y. Nakamura, S. Tabata, and H. Hirakawa, "Plant genome database japan (pgdbj)," in *Plant Genomics Databases*, Springer, 2017, pp. 45–77.

[37] C. Durinx, J. McEntyre, R. Appel, R. Apweiler, M. Barlow, N. Blomberg, C. Cook, E. Gasteiger, J.-H. Kim, R. Lopez, *et al.*, "Identifying elixir core data resources," *F1000Research*, vol. 5, 2016.

[38] L. A. Mueller, T. H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M. H. Wright, R. Ahrens, Y. Wang, *et al.*, "The sol genomics network. a comparative resource for solanaceae biology and beyond," *Plant physiology*, vol. 138, no. 3, pp. 1310–1317, 2005.

[39] M. Garcia-Hernandez, T. Berardini, G. Chen, D. Crist, A. Doyle, E. Huala, E. Knee, M. Lambrecht, N. Miller, L. A. Mueller, *et al.*, "Tair: A resource for integrated arabidopsis data," *Functional & integrative genomics*, vol. 2, no. 6, pp. 239–253, 2002.

[40] C. J. Lawrence, Q. Dong, M. L. Polacco, T. E. Seigfried, and V. Brendel, "Maizegdb, the community database for maize genetics and genomics," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D393–D397, 2004.

[41] Q. Yuan, S. Ouyang, A. Wang, W. Zhu, R. Maiti, H. Lin, J. Hamilton, B. Haas, R. Sultana, F. Cheung, *et al.*, "The institute for genomic research osa1 rice genome annotation database," *Plant physiology*, vol. 138, no. 1, pp. 18–26, 2005.

[42] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens, "The semantic web in action," *Scientific American*, vol. 297, no. 6, pp. 90–97, 2007.

[43] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[44] *Resource description framework (rdf): Concepts and abstract syntax*. [Online]. Available: `https://www.w3.org/TR/rdf-concepts/` (visited on 08/01/2019).

[45] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[46] M. Corpas, N. V. Kovalevskaya, A. McMurray, and F. G. Nielsen, "A fair guide for data providers to maximise sharing of human genomic data," *PLoS computational biology*, vol. 14, no. 3, e1005873, 2018.

[47] A. Kuzniar, A. Gavai, L. Ridder, L. O. B. da Silva Santos, G. Singh, R. Visser, and R. Finkers, *candYgene: enabling precision breeding through FAIR Data*, Sep. 2015. DOI: `10.5281/zenodo.30554`. [Online]. Available: `https://doi.org/10.5281/zenodo.30554`.

[48] M. Sulli, G. Mandolino, M. Sturaro, C. Onofri, G. Diretto, B. Parisi, and G. Giuliano, "Molecular and biochemical characterization of a potato collection with contrasting tuber carotenoid content," *PloS One*, vol. 12, no. 9, e0184143, 2017.

[49] A. Ballester, Y. Tikunov, J. Molthoff, S. Grandillo, M. Viquez-Zamora, R. de Vos, R. de Maagd, S. van Heusden, and A. Bovy, "Identification of loci affecting accumulation of secondary metabolites in tomato fruit of a solanum lycopersicum x solanum chmielewskii introgression line population.," *Frontiers in plant science*, vol. 7, p. 1428, 2016.

[50] C. Brown, T. Kim, Z. Ganga, K. Haynes, D. De Jong, M. Jahn, I. Paran, and W. De Jong, "Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism," *American Journal of Potato Research*, vol. 83, no. 5, pp. 365–372, 2006.

[51] G. Giuliano, "Plant carotenoids: Genomics meets multi-gene engineering," *Current Opinion in Plant Biology*, vol. 19, pp. 111–117, 2014.

[52] A. Acharjee, B. Kloosterman, R. C. de Vos, J. S. Werij, C. W. Bachem, R. G. Visser, and C. Maliepaard, "Data integration and network reconstruction with omics data using random forest regression in potato," *Analytica Chimica Acta*, vol. 705, no. 1-2, pp. 56–63, 2011.

[53] A. Acharjee, B. Kloosterman, R. G. Visser, and C. Maliepaard, "Integration of multi-omics data for prediction of phenotypic traits using random forest," *BMC Bioinformatics*, vol. 17, no. 5, p. 180, 2016.

[54] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: Success, failure and the future," *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 132–144, 2015.

[55] N. Harmston, W. Filsell, and M. P. Stumpf, "What the papers say: Text mining for genomics and systems biology," *Human Genomics*, vol. 5, no. 1, p. 17, 2010.

[56] J. Baran, M. Gerner, M. Haeussler, G. Nenadic, and C. M. Bergman, "Pubmed2ensembl: A resource for mining the biological literature on genes," *PloS One*, vol. 6, no. 9, e24716, 2011.

[57] R. Ding, C. N. Arighi, J.-Y. Lee, C. H. Wu, and K. Vijay-Shanker, "Pgenn, a gene normalization tool for plant genes and proteins in scientific literature," *PLoS One*, vol. 10, no. 8, e0135305, 2015.

[58] W. Choi, B. Kim, H. Cho, D. Lee, and H. Lee, "A corpus for plant-chemical relationships in the biomedical domain," *BMC Bioinformatics*, vol. 17, no. 1, p. 386, 2016.

[59] D. Galea, I. Laponogov, and K. Veselkov, "Exploiting and assessing multi-source data for supervised biomedical named entity recognition," *Bioinformatics*, vol. 34, no. 14, pp. 2474–2482, 2018.

[60] L. Endara, H. Cui, and J. G. Burleigh, "Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing," *Applications in Plant Sciences*, vol. 6, no. 3, e1035, 2018.

[61] G. Singh, A. Kuzniar, E. M. van Mulligen, A. Gavai, C. W. Bachem, R. G. Visser, and R. Finkers, "Qtltableminer++: Semantic mining of qtl tables in scientific articles," *BMC bioinformatics*, vol. 19, no. 1, p. 183, 2018.

[62] H. Cho, W. Choi, and H. Lee, "A method for named entity normalization in biomedical articles: Application to diseases and plants," *BMC Bioinformatics*, vol. 18, no. 1, p. 451, 2017.

[63] G. Jang, T. Lee, S. Hwang, C. Park, J. Ahn, S. Seo, Y. Hwang, and Y. Yoon, "PISTON: predicting drug indications and side effects using topic modeling and natural language processing," *Journal of Biomedical Informatics*, vol. 87, pp. 96–107, 2018.

[64] U. Hahn, K. B. Cohen, Y. Garten, and N. H. Shah, "Mining the pharmacogenomics literature—a survey of the state of the art," *Briefings in Bioinformatics*, vol. 13, no. 4, pp. 460–494, 2012.

[65] V. Sharma, W. Law, M. J. Balick, and I. N. Sarkar, "Harnessing biomedical natural language processing tools to identify medicinal plant knowledge from historical texts," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2017, 2017, p. 1537.

[66] H. V. Cook and L. J. Jensen, "A guide to dictionary-based text mining," *Bioinformatics and Drug Discovery*, pp. 73–89, 2019.

[67] C. Kim, V. Zhu, J. Obeid, and L. Lenert, "Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke," *PloS One*, vol. 14, no. 2, e0212778, 2019.

[68] D. A. Ferrucci, "Introduction to "This is Watson"," *IBM Journal of Research and Development*, vol. 56, no. 3.4, pp. 1–1, 2012.

[69] Y. Chen, J. E. Argentinis, and G. Weber, "IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research," *Clinical Therapeutics*, vol. 38, no. 4, pp. 688–701, 2016.

[70] G. Singh and E. Papoutsoglou, *Cytoscape session for the potato knowledge graph extracted with IBM Watson's supervised NLP model*, Jul. 2019. DOI: 10.5281/zenodo.3275105.

[71] N. Menda, R. M. Buels, I. Tecle, and L. A. Mueller, "A community-based annotation framework for linking solanaceae genomes with phenomes," *Plant Physiology*, vol. 147, no. 4, pp. 1788–1799, 2008.

[72] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[73] R. C. Jansen, *Mapping of quantitative trait loci by using genetic markers: an overview of biometrical models used*. University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute (GBB): EPRINTS-BOOK-TITLE, 1994.

[74] P. Y. Chibon, R. E. Voorrips, R. G. F. Visser, and R. Finkers, "MQ2: Visualizing multi-trait mapped QTL results," *Molecular Breeding*, vol. 32, no. 4, pp. 981–985, 2013, ISSN: 13803743. DOI: 10.1007/s11032-013-9911-3. [Online]. Available: http://dx.doi.org/10.1007/s11032-013-9911-3.

[75] Z.-L. Hu, E. R. Fritz, and J. M. Reecy, *AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond*, eng, 2007. DOI: 10.1093/nar/gkl946.

[76] C. J. Lawrence, L. C. Harper, M. L. Schaeffer, T. Z. Sen, T. E. Seigfried, and D. A. Campbell, *MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research*, eng, 2008. DOI: 10.1155/2008/496957.

[77] J. Ni, A. Pujar, K. Youens-Clark, I. Yap, P. Jaiswal, I. Tecle, C.-W. Tung, L. Ren, W. Spooner, X. Wei, S. Avraham, D. Ware, L. Stein, and S. McCouch, "Gramene QTL database: development, content and applications," *Database*, vol. 2009, bap005, 2009. DOI: 10.1093/database/bap005. [Online]. Available: +%20http://dx.doi.org/10.1093/database/bap005.

[78] I. Y. Tecle, N. Menda, R. M. Buels, E. van der Knaap, and L. A. Mueller, "solQTL: a tool for QTL analysis, visualization and linking to genomes at SGN database," *BMC Bioinformatics*, vol. 11, no. 1, p. 525, 2010, ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-525. [Online]. Available: https://doi.org/10.1186/1471-2105-11-525.

[79] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "WebTables," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 538–549, 2008, ISSN: 21508097. DOI: 10.14778/1453856.1453916. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1453856.1453916.

[80] J. Peng, X. Shi, Y. Sun, D. Li, B. Liu, F. Kong, and X. Yuan, "QTLMiner: QTL database curation by mining tables in literature," *Bioinformatics*, vol. 31, no. 10, pp. 1689–1691, 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv016. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btv016.

[81] *Tabula: A tool for liberating data tables locked inside pdf files.* [Online]. Available: http://tabula.technology/ (visited on 04/01/2018).

[82] *Google tables homepage*. [Online]. Available: https://research.google.com/tables (visited on 04/01/2018).

[83] P. Venetis, A. Halevy, and J. Madhavan, "Recovering semantics of tables on the web," *Proceedings of the VLDB Endowment*, vol. 4, pp. 528–538, 2011, ISSN: 21508097. DOI: 10.14778/2002938.2002939. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002939.

[84] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu, "Understanding tables on the web," in *Conceptual Modeling*, P. Atzeni, D. Cheung, and S. Ram, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 141–155, ISBN: 978-3-642-34002-4.

[85] *Bcl technologies: Pdf conversion*. [Online]. Available: http://www.pdfonline.com/corporate/ (visited on 04/01/2018).

[86] *Apache solr: Solr is the popular, blazing-fast, open source enterprise search platform built on apache lucene*. [Online]. Available: https://lucene.apache.org/solr/ (visited on 04/01/2018).

[87] A. S Schwartz and M. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," vol. 4, pp. 451–62, Feb. 2003.

[88] *Spto: Solanaceae phenotype ontology*. [Online]. Available: http://bioportal.bioontology.org/ontologies/SPTO?p=classes&conceptid=root (visited on 04/01/2018).

[89] R. Shrestha, L. Matteis, M. Skofic, A. Portugal, G. McLaren, G. Hyman, and E. Arnaud, "Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice," *Frontiers in Physiology*, vol. 3, p. 326, 2012, ISSN: 1664-042X. DOI: 10.3389/fphys.2012.00326. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fphys.2012.00326.

[90] *Po: Plant ontology*. [Online]. Available: https://raw.githubusercontent.com/Planteome/plant-ontology/master/po.owl (visited on 04/01/2018).

[91] L. Cooper, R. L. Walls, J. Elser, M. A. Gandolfo, D. W. Stevenson, B. Smith, J. Preece, B. Athreya, C. J. Mungall, S. Rensing, M. Hiss, D. Lang, R. Reski, T. Z. Berardini, D. Li, E. Huala, M. Schaeffer, N. Menda, E. Arnaud, R. Shrestha, Y. Yamazaki, and P. Jaiswal, "The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses," *Plant and Cell Physiology*, vol. 54, no. 2, e1, 2013. DOI: 10.1093/pcp/pcs163. [Online]. Available: +%20http://dx.doi.org/10.1093/pcp/pcs163.

[92] *Pato ontology*. [Online]. Available: https://raw.githubusercontent.com/pato-ontology/pato/master/pato.owl (visited on 04/01/2018).

[93] R. L. Walls, B. Athreya, L. Cooper, J. Elser, M. A. Gandolfo, P. Jaiswal, C. J. Mungall, J. Preece, S. Rensing, B. Smith, and D. W. Stevenson, "Ontologies as integrative tools for plant science.," eng, *American journal of botany*, vol. 99, no. 8, pp. 1263–1275, 2012, ISSN: 1537-2197 (Electronic). DOI: 10.3732/ajb.1200222.

[94] *To: Trait ontology*. [Online]. Available: https://raw.githubusercontent.com/Planteome/plant-trait-ontology/master/to.owl (visited on 04/01/2018).

[95] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000, ISSN: 1061-4036. DOI: 10.1038/75556. [Online]. Available: http://europepmc.org/articles/PMC3037419.

[96] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The Sequence Ontology: a tool for the unification of genome annotations," *Genome Biology*, vol. 6, no. 5, R44, 2005, ISSN: 1474-760X. DOI: 10.1186/gb-2005-6-5-r44. [Online]. Available: https://doi.org/10.1186/gb-2005-6-5-r44.

[97] *Chebi: Chemical entities of biological interest database/ontology*. [Online]. Available: http://purl.obolibrary.org/obo/chebi.owl (visited on 04/01/2018).

[98] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013.," eng, *Nucleic acids research*, vol. 41, no. Database issue, pp. D456–63, 2013, ISSN: 1362-4962 (Electronic). DOI: 10.1093/nar/gks1146.

[99] *Sol genomics network*. [Online]. Available: https://solgenomics.net/ (visited on 04/01/2018).

[100] N. Fernandez-Pozo, N. Menda, J. D. Edwards, S. Saha, I. Y. Tecle, S. R. Strickler, A. Bombarely, T. Fisher-York, A. Pujar, H. Foerster, A. Yan, and L. A. Mueller, "The Sol Genomics Network (SGN) - from genotype to phenotype to breeding," *Nucleic Acids Research*, vol. 43, no. D1, p. D1036, 2015. DOI: 10.1093/nar/gku1195. [Online]. Available: +%20http://dx.doi.org/10.1093/nar/gku1195.

[101] *Stato: Statistics ontology*. [Online]. Available: https://raw.githubusercontent.com/ISA-tools/stato/dev/releases/latest_release/stato.owl (visited on 04/01/2018).

[102] *Sqlite: Self-contained, high-reliability, embedded, full-featured, public-domain, sql database engine.* [Online]. Available: https://www.sqlite.org/ (visited on 04/01/2018).

[103] *Openrefine: Tool that allows you to load data, understand it, clean it up, reconcile it, and augment it with data coming from the web.* [Online]. Available: https://github.com/OpenRefine/OpenRefine (visited on 04/01/2018).

[104] A. Kuzniar and G. Singh, *Quantitative trait loci in solanaceae species*, 2018. DOI: 10.5281/zenodo.1215044. [Online]. Available: https://doi.org/10.5281/zenodo.1215044.

[105] V. Mulwad, T. Finin, and A. Joshi, "Interpreting medical tables as linked data for generating meta-analysis reports," *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014*, no. August, pp. 677–686, 2014. DOI: 10.1109/IRI.2014.7051955.

[106] N. Milosevic, C. Gregson, R. Hernandez, and G. Nenadic, "Extracting patient data from tables in clinical literature Case study on extraction of BMI , weight and number of patients," vol. 5, no. Biostec, pp. 1–6, 2016. DOI: 10.5220/0005660102230228.

[107] P.-Y. Chibon, H. Schoof, R. G. Visser, and R. Finkers, "Marker2sequence, mine your qtl regions for candidate genes," *Bioinformatics*, vol. 28, no. 14, pp. 1921–1922, 2012.

[108] *Solanacea Linked Data platform*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/fct/ (visited on 08/01/2019).

[109] T. Berners-Lee, *Linked data*, 2006. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html (visited on 08/01/2019).

[110] L. O. da Silva Santos, M. Wilkinson, A. Kuzniar, R. Kaliyaperumal, M. Thompson, M. Dumontier, and K. Burger, "FAIR Data Points Supporting Big Data Interoperability," in, 2016, pp. 270–279, ISBN: 9781847040442.

[111] "Expansion of the Gene Ontology knowledgebase and resources.," eng, *Nucleic acids research*, vol. 45, no. D1, pp. D331–D338, 2017, ISSN: 1362-4962 (Electronic). DOI: 10.1093/nar/gkw1108.

[112] *FALDO Ontology*. [Online]. Available: http://biohackathon.org/resource/faldo.rdf (visited on 08/01/2019).

[113] L. Cooper, A. Meier, M.-A. Laporte, J. L. Elser, C. Mungall, B. T. Sinn, D. Cavaliere, S. Carbon, N. A. Dunn, B. Smith, B. Qu, J. Preece, E. Zhang, S. Todorovic, G. Gkoutos, J. H. Doonan, D. W. Stevenson, E. Arnaud, and P. Jaiswal, "The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics.," eng, *Nucleic acids research*, vol. 46, no. D1, pp. D1168–D1180, 2018, ISSN: 1362-4962 (Electronic). DOI: 10.1093/nar/gkx1152.

[114] A. Meroño-Peñuela and R. Hoekstra, "grlc Makes GitHub Taste Like Linked Data APIs," in *The Semantic Web*, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, and C. Lange, Eds., Cham: Springer International Publishing, 2016, pp. 342–353, ISBN: 978-3-319-47602-5.

[115] G. Singh and A. Kuzniar, *QTLTableMiner++:(v1.1.0)*, 2019. DOI: 10.5281/zenodo.1193640. [Online]. Available: https://doi.org/10.5281/zenodo.1193640.

[116] A. Kuzniar, *SIGA.py*, 2017. DOI: 10.5281/ZENODO.1076438. [Online]. Available: https://zenodo.org/record/1076438.

[117] *Solanum_lycopersicum(ITAG2.4)*. [Online]. Available: ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/ (visited on 08/01/2019).

[118] *Solanum_pennellii(ITAG2.4)*. [Online]. Available: ftp://ftp.solgenomics.net/genomes/Solanum_pennellii (visited on 08/01/2019).

[119] *Solanum_pennellii(ITAG2.4)*. [Online]. Available: ftp://ftp.solgenomics.net/genomes/Solanum_tuberosum/ (visited on 08/01/2019).

[120] *EnsemblPlant:Solanum lycopersicum*. [Online]. Available: http://plants.ensembl.org/Solanum_lycopersicum (visited on 08/01/2019).

[121] *EnsemblPlant:Solanum lycopersicum*. [Online]. Available: http://plants.ensembl.org/Solanum_tuberosum (visited on 08/01/2019).

[122] *Uniprot: Solanum Lycopersicum*. [Online]. Available: http://www.uniprot.org/proteomes/Solanum_lycopersicum (visited on 08/01/2019).

[123] *Uniprot: Solanum Lycopersicum*. [Online]. Available: http://www.uniprot.org/proteomes/Solanum_tuberosum (visited on 08/01/2019).

[124] *Uniprot RDF Core Ontology*. [Online]. Available: https://www.uniprot.org/core/ (visited on 08/01/2019).

[125] *Semanticscience Integrated Ontology*. [Online]. Available: http://semanticscience.org/ontology/sio.owl (visited on 08/01/2019).

[126] *RO: Relation Ontology*. [Online]. Available: http://purl.obolibrary.org/obo/ro.owl (visited on 08/01/2019).

[127] C. Boettiger, "An introduction to docker for reproducible research," *ACM SIGOPS Operating Systems Review*, vol. 49, no. 1, pp. 71–79, 2015.

[128] R. H. Ansible, *Ansible is simple it automation*. [Online]. Available: https://www.ansible.com/ (visited on 08/01/2019).

[129] A. Kuzniar and G. Singh, *Pbg-ld (virtuoso faceted browser)*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/fct/.

[130] A. Kuzniar and G. Singh, *Pbg-ld (sparql endpoint)*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/sparql/.

[131] *SPTO: Fruit shape*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/describe/?url=http%7B%5C%%7D3A%7B%5C%%7D2F%7B%5C%%7D2Fpurl.obolibrary.org%7B%5C%%7D2Fobo%7B%5C%%7D2FSP_0000038&sid=42&urilookup=1 (visited on 08/01/2019).

[132] *Trait Ontology: Fruit shape*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/describe/?url=http%7B%5C%%7D3A%7B%5C%%7D2F%7B%5C%%7D2Fpurl.obolibrary.org%7B%5C%%7D2Fobo%7B%5C%%7D2FTO%7B%5C_%7D0002628%7B%5C&%7Dsid=5%7B%5C&%7Durilookup=1 (visited on 08/01/2019).

[133] *QTL:4321030_4_14*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl/describe/?url=http%7B%5C%%7D3A%7B%5C%%7D2F%7B%5C%%7D2Flocalhost%7B%5C%%7D3A8890%7B%5C%%7D2Fgenome%7B%5C%%7D2FSolanum%7B%5C_%7Dlycopersicum%7B%5C%%7D2Fqtl%7B%5C%%7D2F4321030%7B%5C_%7D4%7B%5C_%7D14%7B%5C&%7Dsid=5 (visited on 08/01/2019).

[134] J. E. Haggard, E. B. Johnson, and D. A. S. Clair, "Multiple qtl for horticultural traits and quantitative resistance to phytophthora infestans linked on solanum habrochaites chromosome 11," *G3: Genes, Genomes, Genetics*, vol. 5, no. 2, pp. 219–233, 2015.

[135] *Api access to count features of s. lycopersicum in ensemble plants graph*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl:8088/api/candYgene/queries/countFeatures?graph=http://plants.ensembl.org/Solanum_lycopersicum&endpoint=http://pbg-ld.candygene-nlesc.surf-hosted.nl/sparql (visited on 08/01/2019).

[136] *Api access to count features of s. lycopersicum in sgn graph*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl:8088/api/candYgene/queries/countFeatures?graph=http://solgenomics.net/genome/Solanum_lycopersicum&endpoint=http://pbg-ld.candygene-nlesc.surf-hosted.nl/sparql (visited on 08/01/2019).

[137] *Api access to count features of s. pennellii in sgn graph*. [Online]. Available: http://pbg-ld.candygene-nlesc.surf-hosted.nl:8088/api/candYgene/queries/countFeatures?graph=http://solgenomics.net/genome/Solanum_pennellii&endpoint=http://pbg-ld.candygene-nlesc.surf-hosted.nl/sparql (visited on 08/01/2019).

[138] S. Wu, B. Zhang, N. Keyhaninejad, *et al.*, "A common genetic mechanism underlies morphological diversity in fruits and other plant organs," *Nature communications*, vol. 9, no. 1, p. 4734, 2018.

[139] V. Shulaev, P. Silverman, and I. Raskin, "Airborne signalling by methyl salicylate in plant pathogen resistance," *Nature*, vol. 385, no. 6618, p. 718, 1997.

[140] K. Luengwilai, O. E. Fiehn, and D. M. Beckles, "Comparison of leaf and fruit metabolism in two tomato (solanum lycopersicum l.) genotypes varying in total soluble solids," *Journal of Agricultural and Food Chemistry*, vol. 58, no. 22, pp. 11 790–11 800, 2010.

[141] P. Di Mascio, S. Kaiser, and H. Sies, "Lycopene as the most efficient biological carotenoid singlet oxygen quencher," *Archives of biochemistry and biophysics*, vol. 274, no. 2, pp. 532–538, 1989.

[142] V. Falara, T. A. Akhtar, T. T. Nguyen, E. A. Spyropoulou, P. M. Bleeker, I. Schauvinhold, Y. Matsuba, M. E. Bonini, A. L. Schilmiller, R. L. Last, *et al.*, "The tomato terpene synthase gene family," *Plant physiology*, vol. 157, no. 2, pp. 770–789, 2011.

[143] L. Chen, Y. An, Y.-x. Li, C. Li, Y. Shi, Y. Song, D. Zhang, T. Wang, and Y. Li, "Candidate loci for yield-related traits in maize revealed by a combination of metaqtl analysis and regional association mapping," *Frontiers in Plant Science*, vol. 8, p. 2190, 2017.

[144] M. Brandizi, A. Singh, and K. Hassani-Pak, "Getting the best of linked data and property graphs: Rdf2neo and the knetminer use case.," in *SWAT4LS*, 2018.

[145] A. Kuzniar, *pbg-ld: Linked Data Platform for Plant Breeding & Genomics*, Oct. 2018. DOI: 10.5281/zenodo.1458169. [Online]. Available: https://doi.org/10.5281/zenodo.1458169.

[146] N. Fernandez-Pozo, N. Menda, J. D. Edwards, S. Saha, I. Y. Tecle, S. R. Strickler, A. Bombarely, T. Fisher-York, A. Pujar, H. Foerster, *et al.*, "The sol genomics network (sgn)—from genotype to phenotype to breeding," *Nucleic acids research*, vol. 43, no. D1, pp. D1036–D1041, 2014.

[147] A. Warwick Vesztrocy, C. Dessimoz, and H. Redestig, "Prioritising candidate genes causing qtl using hierarchical orthologous groups," *Bioinformatics*, vol. 34, no. 17, pp. i612–i619, 2018.

[148] F. Lin, J. Fan, and S. Y. Rhee, "Qtg-finder: A machine-learning based algorithm to prioritize causal genes of quantitative trait loci in arabidopsis and rice," *bioRxiv*, p. 484 204, 2019.

[149] A. Schneider, C. Dessimoz, and G. H. Gonnet, "Oma browser—exploring orthologous relations across 352 complete genomes," *Bioinformatics*, vol. 23, no. 16, pp. 2180–2182, 2007.

[150] E. Fridman, Y. Liu, L. Carmel-Goren, A. Gur, M. Shoresh, T. Pleban, Y. Eshed, and D. Zamir, "Two tightly linked qtls modify tomato sugar content via different physiological pathways," *Molecular Genetics and Genomics*, vol. 266, no. 5, pp. 821–826, 2002.

[151] F. Bouvier, A. D'Harlingue, R. A. Backhaus, M. H. Kumagai, and B. Camara, "Identification of neoxanthin synthase as a carotenoid cyclase paralog," *European Journal of Biochemistry*, vol. 267, no. 21, pp. 6346–6352, 2000.

[152] Y. Tadmor, E. Fridman, A. Gur, O. Larkov, E. Lastochkin, U. Ravid, D. Zamir, and E. Lewinsohn, "Identification of malodorous, a wild species allele affecting tomato aroma that was selected against during domestication," *Journal of Agricultural and Food Chemistry*, vol. 50, no. 7, pp. 2005–2009, 2002.

[153] R. Marti, S. Rosello, and J. Cebolla-Cornejo, "Tomato as a source of carotenoids and polyphenols targeted to cancer prevention," *Cancers*, vol. 8, no. 6, p. 58, 2016.

[154] J. Shi, Y. Kakuda, and D. Yeung, "Antioxidative properties of lycopene and other carotenoids from tomatoes: Synergistic effects," *Biofactors*, vol. 21, no. 1-4, pp. 203–210, 2004.

[155] F. X. Cunningham, B. Pogson, Z. Sun, K. A. McDonald, D. DellaPenna, and E. Gantt, "Functional analysis of the beta and epsilon lycopene cyclase enzymes of arabidopsis reveals a mechanism for control of cyclic carotenoid formation.," *The Plant Cell*, vol. 8, no. 9, pp. 1613–1626, 1996.

[156] M. C. Rousseaux, C. M. Jones, D. Adams, R. Chetelat, A. Bennett, and A. Powell, "Qtl analysis of fruit antioxidants in tomato using lycopersicon pennellii introgression lines," *Theoretical and Applied Genetics*, vol. 111, no. 7, pp. 1396–1408, 2005.

[157] D. M. Tieman, H. M. Loucas, J. Y. Kim, D. G. Clark, and H. J. Klee, "Tomato phenylacetaldehyde reductases catalyze the last step in the synthesis of the aroma volatile 2-phenylethanol," *Phytochemistry*, vol. 68, no. 21, pp. 2660–2669, 2007.

[158] J. Zhang, J. Zhao, Y. Xu, J. Liang, P. Chang, F. Yan, M. Li, Y. Liang, and Z. Zou, "Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor," *Frontiers in plant science*, vol. 6, p. 1042, 2015.

[159] S. A. Socaci, C. Socaciu, C. Mureşan, A. Fărcaş, M. Tofană, S. Vicaş, and A. Pintea, "Chemometric discrimination of different tomato cultivars based on their volatile fingerprint in relation to lycopene and total phenolics content," *Phytochemical analysis*, vol. 25, no. 2, pp. 161–169, 2014.

[160] J. L. Rambla, A. Medina, A. Fernandez-del-Carmen, W. Barrantes, S. Grandillo, M. Cammareri, G. Lopez-Casado, G. Rodrigo, A. Alonso, S. Garcia-Martinez, *et al.*, "Identification, introgression, and validation of fruit volatile qtls from a red-fruited wild tomato species," *Journal of experimental botany*, vol. 68, no. 3, pp. 429–442, 2016.

[161] D. J. Gray, Z. T. Li, and S. A. Dhekney, "Precision breeding of grapevine (vitis vinifera l.) for improved traits," *Plant science*, vol. 228, pp. 3–10, 2014.

[162] M. W. Rosegrant and S. A. Cline, "Global food security: Challenges and policies," *Science*, vol. 302, no. 5652, pp. 1917–1919, 2003.

[163] J. Wan, "Perspectives of molecular design breeding in crops," *Zuo wu xue bao*, vol. 32, no. 3, pp. 455–462, 2006.

[164] K. Hassani-Pak and C. Rawlings, "Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes," *Journal of integrative bioinformatics*, vol. 14, no. 1, 2017.

[165] P. M. Davis and W. H. Walters, "The impact of free access to the scientific literature: A review of recent research," *Journal of the Medical Library Association: JMLA*, vol. 99, no. 3, p. 208, 2011.

[166] G. Schlosser, "Modularity and the units of evolution," *Theory in Biosciences*, vol. 121, no. 1, pp. 1–80, 2002.

[167] P. Selby, R. Abbeloos, J. E. Backlund, M. Basterrechea Salido, G. Bauchet, O. E. Benites-Alfaro, C. Birkett, V. C. Calaminos, P. Carceller, G. Cornut, *et al.*, "Brapi—an application programming interface for plant breeding applications," 2019.

[168] J. D. Montenegro, A. A. Golicz, P. E. Bayer, B. Hurgobin, H. Lee, C.-K. K. Chan, P. Visendi, K. Lai, J. Doležel, J. Batley, *et al.*, "The pangenome of hexaploid bread wheat," *The Plant Journal*, vol. 90, no. 5, pp. 1007–1013, 2017.

[169] C. Prodhomme, D. Esselink, T. Borm, R. G. Visser, H. J. Van Eck, and J. H. Vossen, "Comparative subsequence sets analysis (cossa) is a robust approach to identify haplotype specific snps; mapping and pedigree analysis of a potato wart disease resistance gene sen3," *Plant methods*, vol. 15, no. 1, p. 60, 2019.

[170] "Supporting data," *Nature Medicine*, vol. 16, no. 2, pp. 131–131, Feb. 2010. DOI: 10.1038/nm0210-131. [Online]. Available: https://doi.org/10.1038/nm0210-131.

[171] E. Smalley, *Ai-powered drug discovery captures pharma interest*, 2017.

[172] A. M. Hulse-Kemp, S. Maheshwari, K. Stoffel, T. A. Hill, D. Jaffe, S. R. Williams, N. Weisenfeld, S. Ramakrishnan, V. Kumar, P. Shah, *et al.*, "Reference quality assembly of the 3.5-gb genome of capsicum annuum from a single linked-read library," *Horticulture Research*, vol. 5, no. 1, p. 4, 2018.

[173] H. Hirakawa, K. Shirasawa, K. Miyatake, T. Nunome, S. Negoro, A. Ohyama, H. Yamaguchi, S. Sato, S. Isobe, S. Tabata, *et al.*, "Draft genome sequence of eggplant (solanum melongena l.): The representative solanum species indigenous to the old world," *DNA research*, vol. 21, no. 6, pp. 649–660, 2014.

[174] C. Vehlow, D. P. Kao, M. R. Bristow, L. E. Hunter, D. Weiskopf, and C. Görg, "Visual analysis of biological data-knowledge networks," *BMC bioinformatics*, vol. 16, no. 1, p. 135, 2015.

[175] J. Gomez, L. J. Garcia, G. A. Salazar, J. Villaveces, S. Gore, A. Garcia, M. J. Martin, G. Launay, R. Alcantara, N. Del-Toro, *et al.*, "Biojs: An open source javascript framework for biological data visualization," *Bioinformatics*, vol. 29, no. 8, pp. 1103–1104, 2013.

[176] B. Kloosterman, D. De Koeyer, R. Griffiths, B. Flinn, B. Steuernagel, U. Scholz, S. Sonnewald, U. Sonnewald, G. J. Bryan, S. Prat, *et al.*, "Genes driving potato tuber initiation and growth: Identification based on transcriptional changes using the poci array," *Functional & integrative genomics*, vol. 8, no. 4, pp. 329–340, 2008.

[177] B. Kloosterman, M. Oortwijn, T. America, R. de Vos, R. G. Visser, C. W. Bachem, *et al.*, "From qtl to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy," *BMC genomics*, vol. 11, no. 1, p. 158, 2010.

[178] C. W. Bachem, R. S. Van Der Hoeven, S. M. De Bruijn, D. Vreugdenhil, M. Zabeau, and R. G. Visser, "Visualization of differential gene expression using a novel method of rna fingerprinting based on aflp: Analysis of gene expression during potato tuber development," *The plant journal*, vol. 9, no. 5, pp. 745–753, 1996.

[179] C. CELIS-GAMBOA, P. Struik, E. Jacobsen, and R. Visser, "Temporal dynamics of tuber formation and related processes in a crossing population of potato (solanum tuberosum)," *Annals of Applied Biology*, vol. 143, no. 2, pp. 175–186, 2003.

[180] J. S. Werij, B. Kloosterman, C. Celis-Gamboa, C. R. De Vos, T. America, R. G. Visser, and C. W. Bachem, "Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, qtl, and expression analysis," *Theoretical and Applied Genetics*, vol. 115, no. 2, pp. 245–252, 2007.

[181] R. Campbell, L. J. Ducreux, W. L. Morris, J. A. Morris, J. C. Suttle, G. Ramsay, G. J. Bryan, P. E. Hedley, and M. A. Taylor, "The metabolic and developmental roles of carotenoid cleavage dioxygenase4 from potato," *Plant Physiology*, vol. 154, no. 2, pp. 656–664, 2010.

[182] M. W. Bonierbale, R. L. Plaisted, and S. D. Tanksley, "Rflp maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato.," *Genetics*, vol. 120, no. 4, pp. 1095–1103, 1988.

[183] C. Stushnoff, L. J. Ducreux, R. D. Hancock, P. E. Hedley, D. G. Holm, G. J. McDougall, J. W. McNicol, J. Morris, W. L. Morris, J. A. Sungurtas, *et al.*, "Flavonoid profiling and transcriptome analysis reveals new gene–metabolite correlations in tubers of solanum tuberosum l.," *Journal of Experimental Botany*, vol. 61, no. 4, pp. 1225–1238, 2010.

[184] G. Diretto, R. Welsch, R. Tavazza, F. Mourgues, D. Pizzichini, P. Beyer, and G. Giuliano, "Silencing of beta-carotene hydroxylase increases total carotenoid and beta-carotene levels in potato tubers," *BMC Plant Biology*, vol. 7, no. 1, p. 11, 2007.

[185] G. Diretto, S. Al-Babili, R. Tavazza, V. Papacchioli, P. Beyer, and G. Giuliano, "Metabolic engineering of potato carotenoid content through tuber-specific overexpression of a bacterial mini-pathway," *PLoS One*, vol. 2, no. 4, e350, 2007.

[186] L. J. Ducreux, W. L. Morris, P. E. Hedley, T. Shepherd, H. V. Davies, S. Millam, and M. A. Taylor, "Metabolic engineering of high carotenoid potato tubers containing enhanced levels of beta-carotene and lutein," *Journal of Experimental Botany*, vol. 56, no. 409, pp. 81–89, 2004.

[187] S. Romer, J. Lubeck, F. Kauder, S. Steiger, C. Adomat, and G. Sandmann, "Genetic engineering of a zeaxanthin-rich potato by antisense inactivation and co-suppression of carotenoid epoxidation," *Metabolic engineering*, vol. 4, no. 4, pp. 263–272, 2002.

[188] N. Wang, W. Fang, H. Han, N. Sui, B. Li, and Q.-W. Meng, "Overexpression of zeaxanthin epoxidase gene enhances the sensitivity of tomato psii photoinhibition to high light and chilling stress," *Physiologia plantarum*, vol. 132, no. 3, pp. 384–396, 2008.

[189] W. Morris, L. Ducreux, D. Griffiths, D. Stewart, H. Davies, and M. Taylor, "Carotenogenesis during tuber development and storage in potato," *Journal of Experimental Botany*, vol. 55, no. 399, pp. 975–982, 2004.

[190] X. Zhou, R. McQuinn, Z. Fei, A.-M. A. WOLTERS, J. Van Eck, C. Brown, J. J. Giovannoni, and L. Li, "Regulatory control of high levels of carotenoid accumulation in potato tubers," *Plant, Cell & Environment*, vol. 34, no. 6, pp. 1020–1030, 2011.

[191] K. G. Haynes, B. A. Clevidence, D. Rao, and B. T. Vinyard, "Inheritance of carotenoid content in tetraploid x diploid potato crosses," *Journal of the American Society for Horticultural Science*, vol. 136, no. 4, pp. 265–272, 2011.

[192] K. Hamouz, J. Lachman, K. Pazderuu, K. Hejtmankova, *et al.*, "Effect of cultivar, location and method of cultivation on the content of chlorogenic acid in potatoes with different flesh colour," *Plant, Soil and Environment*, vol. 59, no. 10, pp. 465–471, 2013.

[193] L. F. Reyes and L. Cisneros-Zevallos, "Degradation kinetics and colour of anthocyanins in aqueous extracts of purple-and red-flesh potatoes (solanum tuberosum l.)," *Food Chemistry*, vol. 100, no. 3, pp. 885–894, 2007.

[194] J. Lachman, K. Hamouz, J. Musilova, K. Hejtmankova, and other, "Effect of peeling and three cooking methods on the content of selected phytochemicals in potato tubers with various colour of flesh," *Food Chemistry*, vol. 138, no. 2-3, pp. 1189–1197, 2013.

[195] K. Hamouz, J. Lachman, *et al.*, "Differences in anthocyanin content and antioxidant activity of potato tubers with different flesh colour," *Plant, Soil and Environment*, vol. 57, no. 10, pp. 478–485, 2011.

[196] Q. Wang, Y. Cao, L. Zhou, C.-Z. Jiang, Y. Feng, and S. Wei, "Effects of postharvest curing treatment on flesh colour and phenolic metabolism in fresh-cut potato products," *Food chemistry*, vol. 169, pp. 246–254, 2015.

[197] K. Hejtmankova, Z. Kotikova, K. Hamouz, V. Pivec, *et al.*, "Influence of flesh colour, year and growing area on carotenoid and anthocyanin content in potato tubers," *Journal of food composition and analysis*, vol. 32, no. 1, pp. 20–27, 2013.

[198] K. Hamouz, J. Lachman, K. Hejtmankova, *et al.*, "Effect of natural and growing conditions on the contentof phenolics in potatoes with different flesh colour," *Plant, Soil and Environment*, vol. 56, no. 8, pp. 368–374, 2010.

[199] C. Brown, C. Edwards, C.-P. Yang, and B. Dean, "Orange flesh trait in potato: Inheritance and carotenoid content," *Journal of the American Society for Horticultural Science*, vol. 118, no. 1, pp. 145–150, 1993.

[200] J. Sliwka, I. Wasilewicz-Flis, H. Jakuczun, and C. Gebhardt, "Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six solanum species," *Plant Breeding*, vol. 127, no. 1, pp. 49–55, 2008.

[201] Y. Zhang, C. S. Jung, and W. S. De Jong, "Genetic analysis of pigmented tuber flesh in potato," *Theoretical and Applied Genetics*, vol. 119, no. 1, pp. 143–150, 2009.

[202] H. De Jong, "Inheritance of anthocyanin pigmentation in the cultivated potato: A critical review," *American Potato Journal*, vol. 68, no. 9, pp. 585–593, 1991.

[203] C. C. Teow, V.-D. Truong, R. F. McFeeters, R. L. Thompson, K. V. Pecota, and G. C. Yencho, "Antioxidant activities, phenolic and beta-carotene contents of sweet potato genotypes with varying flesh colours," *Food chemistry*, vol. 103, no. 3, pp. 829–838, 2007.

[204] W. Lu, K. Haynes, E. Wiley, and B. Clevidence, "Carotenoid content and color in diploid potatoes," *Journal of the American Society for Horticultural Science*, vol. 126, no. 6, pp. 722–726, 2001.

[205] G. Diretto, R. Tavazza, R. Welsch, D. Pizzichini, F. Mourgues, V. Papacchioli, P. Beyer, and G. Giuliano, "Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase," *BMC plant biology*, vol. 6, no. 1, p. 13, 2006.

[206] J. Van Eck, B. Conlin, D. Garvin, H. Mason, D. Navarre, and C. Brown, "Enhancing beta-carotene content in potato by rnai-mediated silencing of the beta-carotene hydroxylase gene," *American Journal of Potato Research*, vol. 84, no. 4, p. 331, 2007.

[207] R. Campbell, S. D. Pont, J. A. Morris, G. McKenzie, S. K. Sharma, P. E. Hedley, G. Ramsay, G. J. Bryan, and M. A. Taylor, "Genome-wide qtl and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (solanum tuberosum l.)," *Theoretical and applied genetics*, vol. 127, no. 9, pp. 1917–1933, 2014.

[208] P. McCord, L. Zhang, and C. Brown, "The incidence and effect on total tuber carotenoids of a recessive zeaxanthin epoxidase allele (zep1) in yellow-fleshed potatoes," *American journal of potato research*, vol. 89, no. 4, pp. 262–268, 2012.

Summary

One of the major global challenges of today is to meet the food demands of an ever-increasing population (food demand will increase by 50% in 2030). One approach to address this challenge is to breed new crop varieties that yield more even under unfavorable conditions e.g. have improved tolerance to drought and/or resistance to pathogens. However, designing a breeding program is a laborious and time consuming effort that often lacks the capacity to generate new cultivars quickly in response to the required traits. Recent advances in biotechnology and genomics data science have the potential to accelerate and precise breeding programs greatly. As large-scale genomic data sets for crop species are available in multiple independent data sources and scientific literature, this thesis provides innovative technologies that use natural language processing (NLP) and semantic web technologies to address challenges of integrating genomic data for improving plant breeding.

Firstly, in this research study, we developed a supervised Natural language processing (NLP) model with the help of IBM Watson, to extract knowledge networks containing genotypic-phenotypic associations of potato tuber flesh color from the scientific literature. Secondly, a table mining tool called QTLTableMiner^{++} (QTM) was developed which enables knowledge discovery of novel genomic regions (such as QTL regions), which positively or negatively affect the traits of interest. The objective of both above mentioned, NLP techniques was to extract information which is implicitly described in the literature and is not available in structured resources, like databases. Thirdly, with the help of semantic web technology, a linked-data platform called Solanaceae linked data platform(pbg-ld) was developed, to semantically integrates geno- and pheno-typic data of Solanaceae species. This platform combines both unstructured data from scientific literature and structured data from publicly available biological databases using the Linked Data approach. Lastly, analysis workflows for prioritizing candidate genes with QTL regions were tested using pbg-ld. Hence, this research provides *in-silico* knowledge discovery tools and genomic data infrastructure, which aids researchers and breeders in the design of a precise and improved breeding program.

Acknowledgements

These four years of PhD journey has been a tremendous learning experience. Besides the development of my scientific and research skills, this journey played a vital role in improving my interpersonal skills. I could not have embarked upon this journey without the help and support of countless people over these four years. Therefore, I am very grateful to all those people, who in diverse ways contributed to the successful completion of my PhD research.

I would like to express my gratitude towards my supervisors **Prof. Dr. Richard G. F. Visser**, **Dr. Christian W. Bachem** and **Dr. Richard Finkers** for giving me an opportunity to pursue my PhD in their group. I am highly grateful for their inspiration, invaluable guidance, moral support and encouragement during my doctoral studies. Thank you for your scientific inspiration and continuous support. I would like to express my sincere thank Dr. Arnold Kuzniar for his valuable support from the Netherlands eSCience Center. His assistance and motivation throughout my PhD were highly valuable.

Further, I would also like to express my gratitude towards **Dr. Erik M. van Mulligen**, **Dr. Anand Gavai**, **Matthijs Brouwe** and **Evangelia A. Papoutsoglou** for their support, assistance, guidance and useful suggestions in all our collaborative projects. I am highly thankful to all the members of the department of plant breeding & research for providing a nice and interactive working environment. Special thanks to my dear friends **Jarst van Belle**, **Pauline van Haperen** and colleagues **Patrick Hendrickx**, **Martijn van Kaauwen**, **Brian Lavrijssen**, **Danny Esselink**, **Dr. Yury M Tikunov** and **Dr. Arnaud G. Bovy** for all the good times we shared, for stimulating discussions, for their help and support during my thesis and the nice coffee breaks we shared. A big thank's to my office colleagues, **Sri Sunarti**, **Sara Bergonzi**, **Lorena Ramirez Gonzales**, **Naser Askari**, **Cynara Romero**, and **Christos Kissoudis** for the a tremendous amount of fun, and happiness we shared in our office. Also, a special thanks to Evangelia and Pauline for being my paranymphs.

I would also like to thank some of my mentors, **Prof. Dr. Tiratha Raj Singh**, **Dr. Dipankar Sengupta**, **Dr. Philipp Senger** and **Prof. Martin Hofmann-Apitius** who encouraged me during my Bachelor's and Master's days until today.

I am deeply grateful to all my friends for always being there on my side and uplift-

ing my spirits. Special thanks go to my freinds: **Rohit Soni**, **Harpreet Singh Sondh**, **Amardeep Kaur Ahluwalia**, **Jelle van Crij**, **Sandeep Sarde**, **Bandan Chakrabortty**, **Sneha Gulati**, **Suraj Jamge**, **Tanvi Taparia**, **Kavya Yalamanchili**, **Priyanka Singh**, **Sidharam Pujari**, **Swarit Jaisal**, **Pranika Singh Rana**, **Meemansa Sood**, **Gurleen Kaur**, **Shivani Tiwari**, **Arka Mallick**, **Tanya Singhania**, **Gaurav Garg**, **Simona Foldesova**, **Mamur Mallaev**, **Jaspreet Kaur Nagpal**, and **Linda Kuerten**.

I dedicate this work to my mother **Prof. Dr. Kiran Preet Kaur**. Thank you Mumma, for always believing in me. It is your impulse that makes me achieve a tag of a Dr. in my name. I owe tons of thanks to my family members for all their love and support. I am especially grateful to my father **Jasbir Singh**, for supporting me in all of my pursuits and inspiring me to follow my dreams. I am privileged to be your son. Special thanks to my grandparents (**Narindar Singh Aneja**, **Davinder Kaur**), my brother and his family (**Dr. Sargundeep Singh**, **Dr. Bhavini Oberoi** and my lovely nephew Hishal), my inlaws (**Onkar Singh**, **Balvir Kaur** and **Gurminder Singh**), my aunts (**Dr. Jiwanjot Kaur**, **Jaipreet Kaur**, and **Rajneet Kaur Puri**).

Last but not the least, to the woman of my life **Jaspinder Saini**. You have been there all the time for me. I am highly thankful to you for your unconditional love, your huge support and loving me. Without you, I would have never accomplished this. Thank you so much!

Curriculum Vitae

Gurnoor Singh was born on 26 April 1990, in the city of Chandigarh, India. He did his school education from Yadavindra Public School. Since childhood, Gurnoor was passionate about playing chess, mathematics, and computer science. After pursuing a Bachelor's degree in Bioinformatics from the Jaypee University of Information Technology, Solan, India, he pursued his Master's degree in Life Science Informatics from the Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany. During his studies in Bonn, he had the opportunity to work as a research software engineer at Fraunhofer SCAI, Sankt Augustin, Germany. During this time he gained insights into software development, text mining and knowledge discovery. In Fraunhofer SCAI, he also accomplished his master's thesis entitled, "Efficient Integration of Semantic Search, Document Retrieval, and Information Extraction in Life Sciences" under the supervision of Dr. Philipp Senger and Prof. Dr. Martin Hofmann-Apitius.

In June 2015, he started his PhD in the department of Plant breeding research at the Wageningen University & Research, The Netherlands under the supervision of Prof. Dr. Richard G. F. Visser, Dr. Christian W. Bachem and Dr. Richard Finkers. The doctoral program was part of the project CandYGene, funded by the Netherlands Organisation for Scientific Research through the Netherlands eScience Center, Amsterdam. The results of his research are presented in this thesis.

Nowadays, Gurnoor Singh is working as a data steward in the Center for Molecular and Biomolecular Informatics (CMBI), RadboudUMC, Nijmegen.

Publication list

Published articles in a journal

1. **G. Singh**, A. Kuzniar, E. M. van Mulligen, A. Gavai, C. W. B. Bachem, R. G. F. Visser, and R. Finkers, "**Qtltableminer++: Semantic mining of qtl tables in scientific articles**", BMCbioinformatics, vol. 19, no. 1, p. 183, 2018.

Articles in preparation

1. **G. Singh**, E. A. Papoutsoglou1, F. K. Lalleman, B. Vencheva, M Rice, Richard G.F. Visser, C. W.B. Bachem, and R Finkers, "**Extracting knowledge networks from plantscientific literature: Potato tuber flesh color asan exemplary trait.**" Submitted (**Chapter 2 in this thesis**)

2. **G. Singh**, A. Kuzniar, M. Brouwer, C. M. Ortiz, C. W. B.Bachem, Y. M. Tikunov, A. G. Bovy, R. Finkers, R G.F. Visser, "**Linked Data platform for Solanaceae species**" In prep. (**Chapter 4 in this thesis**)

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: Gurnoor Singh
Date: 09 December 2019
Group: Laboratory of Plant Breeding
University: Wageningen University & Research

1) Start-Up Phase	date	cp
► **First presentation of your project**		
Prediction of candidate genes for traits using interoperable genome annotations and literature	16th June, 2015	1.5
► **Writing or rewriting a project proposal**		
Prediction of candidate genes for traits using interoperable genome annotations and literature	September, 2015	6.0
► **Writing a review or book chapter**		
► **MSc courses**		
Subtotal Start-Up Phase		7.5

2) Scientific Exposure	date	cp
► **EPS PhD student days**		
EPS PhD student days Get2Gether, Soest	28th & 29th January, 2016	0.6
EPS PhD student days Get2Gether, Soest	9th & 10th February, 2017	0.5
BioSB 2016 PhD retreat	18th April, 2016	0.2
BioSB 2017 PhD retreat	3rd April, 2017	0.2
► **EPS theme symposia**		
EPS Theme 4 Symposium 'Genome Biology', University of Amsterdam	15th December, 2015	0.3
EPS Theme 4 Symposium 'Genome Biology', Wageningen University & Research	16th December, 2016	0.3
EPS Theme 3 Symposium 'Metabolism and Adaptation', Wageningen University & Research	13th March, 2018	0.3
► **Lunteren Days and other national platforms**		
Annual meeting 'Experimental Plant Science', Lunteren	11th & 12th April, 2016	0.6
Annual meeting 'Experimental Plant Science', Lunteren	10th & 11th April, 2017	0.6
Annual meeting 'Experimental Plant Science', Lunteren	9th & 10th April, 2018	0.6
Annual Dutch Bioinformatics & Systems Biology Conference (BioSB 2016), Lunteren	19th & 20th April, 2016	0.6
Annual Dutch Bioinformatics & Systems Biology Conference (BioSB 2017), Lunteren	4th & 5th April, 2017	0.6
3rd National eScience Symposium, Amsterdam	8th October, 2015	0.3
4th National eScience Symposium, Amsterdam	12th October, 2017	0.3
► **Seminars (series), workshops and symposia**		
98th Dies Natalis Symposium Wageningen, Wageningen	9th March, 2016	0.3
IBM Watson Knowledge Studio Workshop, Amsterdam	9th November, 2017	0.3
BYOD BrAPI workshop, Gent, Belgium	30th May - 1st June, 2017	0.9
Farmhack, Hackathon, Dairy Campus, Leeuwarden	24th & 25th November, 2017	0.6
► **Seminar plus**		
► **International symposia and congresses**		
Semantic web for Life Science, Cambridge, UK	7th- 10th December, 2015	1.2
15th European Conference on Computational Biology, Den Haag	3rd - 7th September, 2016	1.2
Benelux Bioinformatics Conference, Leuven, Belgium	13th & 14th December, 2017	0.5
17th European Conference on Computational Biology, Athens, Greece	8th & 12th September, 2018	1.2
► **Presentations**		
Semantic web for Life Science (Poster and Flash Presentation)	9th December, 2015	1.0
Dies Symposium (Presentation)	9th March, 2016	1.0
15th European Conference on Computational Biology (Poster)	3rd - 7th September, 2016	1.0
EPS Theme 4 Symposium 'Genome Biology' (Presentation)	16th December, 2016	1.0
BioSB 2017 (Presentation)	4th April, 2017	1.0
Benelux Bioinformatics Conference 2017 (Presentation)	14th December, 2017	1.0
17th European Conference on Computational Biology (Poster)	10th September, 2018	1.0
► **IAB interview**		
► **Excursions**		
Subtotal Scientific Exposure		19.2

3) In-Depth Studies	date	cp
► **Advanced scientific courses & workshops**		
Managing and Integrating Life Science Information, Utrecht	30th Nov - 4th Dec 2015	3.0
Basic statistics course PE & RC, Wageningen	17th - 24th May, 2016	1.5
Statistical Techniques in Biological and Medical Sciences, India	13th- 18th June, 2016	1.8
Quantitative & Predictive modelling, Wageningen	26-30th June, 2017	3.0
Version Control and Collaboration with Git and GitHub, Utrecht	6th September, 2017	0.3
► **Journal club**		
► **Individual research training**		
Subtotal In-Depth Studies		9.6

CONTINUED ON NEXT PAGE

4) Personal Development		date	cp
►	**General skill training courses**		
	EPS Introduction Course, Wageningen	29th September, 2016	0.3
	Presenting with Impact, Wageningen	6th - 20th March, 2017	1.0
	Project and Time Management, Wageningen	7 - 26 October, 2017	1.5
	Pitch Training for Dies Symposium, Wageningen	17th February, 2016	0.2
►	**Organisation of meetings, PhD courses or outreach activities**		
	BYOD BrAPI workshop, Gent, Belgium	30th May - 1st June, 2017	1.5
	Volunteer 17th European Conference on Computation Biology Athens, Greece	9th & 10th September, 2018	0.6
►	**Membership of EPS PhD Council**		
	Subtotal Personal Development		5.1

TOTAL NUMBER OF CREDIT POINTS*	41.4
Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.	
** A credit represents a normative study load of 28 hours of study.*	

The research described in this thesis was funded by the Netherlands Organisation for Scientific Research through the Netherlands eScience Center (NWO, grant number: 27014204).

Cover design and Thesis layout: Gurnoor Singh

Printed by: Proefschriftmaken (Proefschriftmaken.nl)