

Bioinformatics assisted breeding,
From QTL to candidate genes

Pierre-Yves Chibon

Thesis committee

Promotor

Prof. Dr R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Co-promotor

Dr H.J. Finkers
Senior Scientist, Wageningen UR Plant Breeding
Wageningen University & Research Centre

Other members

Prof. Dr P.C. de Ruiter, Wageningen University
Dr E. Schultes, Leiden University Medical Centre
Dr J.P.H. Nap, Hanze University of Applied Sciences, Groningen
Dr R.A. de Maagd, Plant Research International, Wageningen

This research was conducted under the auspices of the Graduate School: Experimental Plant Sciences (EPS)

Bioinformatics assisted breeding,
From QTL to candidate genes

Pierre-Yves Chibon

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M. J. Kropff,
in the presence of the
Thesis committee appointed by the Academic Board
to be defended in public
on Thursday, November 7th 2013
at 11 a.m. in the Aula.

Pierre-Yves Chibon

Bioinformatics assisted breeding, from QTL to candidate genes

PhD thesis Wageningen University, Wageningen, The Netherlands, 2013
With references, with summaries in English, French and Dutch.

ISBN: 978-94-6173-736-6

Contents

Chapter 1: General introduction	9
Chapter 2: Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on Linkage Group 16.....	27
Chapter 3: MQ ² : Visualizing multi-trait mapped QTL results.....	49
Chapter 4: Marker2sequence, mine your QTL regions for candidate genes	57
Chapter 5: Identification of transcription factor binding sites in tomato.....	61
Chapter 6: Annotex: Exploring the genome annotation	87
Chapter 7: General discussion.....	101
References.....	115
Summary	129
Samenvatting.....	133
Résumé.....	137
Acknowledgements	141
Curriculum vitae	145
Publications	147

Abbreviation table

API	Application Programming Interface
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA
cM	centiMorgan
CSV	Comma Separated Values
DART	Diversity Array Technology
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
FAO	Food and Agriculture Organization
FAQ	Frequently Asked Questions
FTP	File Transfer Protocol
GCMS	Gas Chromatography Mass Spectrometry
GO	Gene Ontology
HTTP	Hyper-Text Transfer Protocol
IL	Introgression Line
ITAG	International Tomato Annotation Group
JSON	JavaScript Object Notation
LCMS	Liquid Chromatography Mass Spectrometry
LG	Linkage Group
M2S	Marker2sequence
MAS	Marker Assisted Selection
MEME	Multiple Em for Motif Elicitation
MFLP	Microsatellite-anchored Fragment Length Polymorphism
MIME	Multipurpose Internet Mail Extensions
mRNA	Messenger RNA
NAR	Nucleic Acid Research
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PGSC	Potato Genome Sequencing Consortium
PPI	Protein-Protein Interaction
QTL	Quantitative Trait Loci
RAPD	Random Amplified Polymorphic DNA
RDF	Resource Description Framework
REST	Representation State Transfer
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
RSAT	Regulatory Sequence Analysis Tools
SNP	Single Nucleotide Polymorphism
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SSR	Single Sequence Repeat
TCP/IP	Transmission Control Protocol / Internet Protocol
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
URI	Unique Resource Identifier
URL	Unique Resource Locator
W3c	World Wide Web Consortium
WSDL	Web-Service Description Language
WWW	World Wide Web
XML	eXtensible Markup Language

Chapter 1: General introduction

Plant breeding is a key factor in the future.

A growing world population

According to the United Nations (UN), the world population was just above 2.5 billion persons in 1950; just under 6.2 billion in 2000 and passed 7 billion in 2010. Estimations from 2011 predict that more than 7.5 billion humans will be living on the planet in 2017. The world population will thus have tripled in less than 70 years. Provisions are that the world population will reach 9 billion in 2038; 10 billion people in 2057 and by the end of the century, in 2100, will be just below 11 billion (United Nations, Department of Economic and Social Affairs, Population Division (2011). World Population Prospects: The 2010 Revision, CD-ROM Edition - <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>). Maslow's hierarchy of needs puts access to food (one of the physiological needs) as one of the most important needs (Maslow 1943). The Food and Agriculture Organisation (FAO) believes that food safety will be one of the major challenges for the coming years: "Producing 70 percent more food for an additional 2.3 billion people by 2050 while at the same time combating poverty and hunger, using scarce natural resources more efficiently and adapting to climate change are the main challenges world agriculture will face in the coming decades" (<http://www.fao.org/news/story/en/item/35571/>). We, humans, depend on agriculture directly or indirectly for food but also fuel, clothing and we compete with it for housing.

As the world population increases, the competition on land for agriculture versus land for urban development will increase further but agricultural techniques and breeding will mitigate this. For example, between 1960 and 2000, the land used in agriculture world-wide has increased by 11% to reach 1.5 billion ha, while the world population has doubled (<http://www.fao.org/docrep/005/y4252e/y4252e06a.htm>). This low increase in land used for agriculture, is due to improved crops and agricultural techniques. These improvements have allowed, between 1961 and 1999, reducing by 56% the arable land required to produce any quantity of grain. Over this time period, the world average grain yield has increased from 1.4 T/ha to 3.05 T/ha (<http://www.fao.org/docrep/005/y4252e/y4252e06a.htm>). Plant breeding is therefore a key issue for the coming years.

A short history of plant breeding and its goals

Prehistoric visual selection of plants that facilitated the harvest or increased the productivity led to the first domesticated varieties (Harlan 1975). Since the domestication of the first plants 13,000 to 11,000 years ago, mankind has tried to develop plants, especially food plants, which better serve his needs. In recent years, this process has become a recognized scientific discipline named plant breeding (Allard 1999). The hybridizations and selection pressure applied by mankind over these 10,000 years has resulted in the domestication of wild varieties into hundreds of thousands of breeds, forming the basis of our current crops (McCouch 2004). This selection process however has reduced the genetic basis of the plants used for food production (Tester and Langridge 2010) leading to a situation where for instance in Russia, in 2006, more than 95% of all winter wheat varieties used are descendants of only two cultivars (Mba, Guimaraes et al. 2012). This narrow genetic base directly endangers food security as crops worldwide become susceptible to the same stresses (biotic or abiotic) and modern breeders use old, wild varieties to find genes to improve current crops (yield, resistance) (Gur and Zamir 2004). Breeders have two possibilities to improve current crops (McCouch 2004), either select for a superior individual among the existing possibilities or efficiently swap,

replace or recombine to build a biological system from an extending range of possibilities which includes wild and old varieties containing traits lost in the course of domestication (Gur and Zamir 2004). Modern breeding relies on the revolution that have brought advances in biotechnology, genomic and molecular marker development and application (Moose and Mumm 2008).

The evolution of modern breeding: from marker development to genome sequencing

Modern breeding integrating new biotechnological approaches started in the early 1980s with the production of the first transgenic plants using *Agrobacterium tumefaciens* transformation (Bevan, Flavell et al. 1983; Fraley, Rogers et al. 1983; Herrera-Estrella, Depicker et al. 1983). Genetic maps relying on molecular markers and allowing correlating genetic linkage between markers and quantitative traits appeared few years later (Edwards, Stuber et al. 1987; Paterson, Lander et al. 1988). The development of molecular markers has allowed predicting the results of a cross without waiting for the plant to express the phenotype by looking for the presence of specific molecular marker(s) associated with the phenotype (Eathington, Crosbie et al. 2007).

Over the last 30 years, the application of plant biotechnology, genomics and molecular breeding has led to the development of new cultivars with higher yield, more resistant to (a)biotic stresses, now used on a daily basis in our agriculture (Moose and Mumm 2008). These new cultivars together with agronomical practices are responsible for the yield increase observed by the FAO between 1961 and 1999.

Molecular marker technology has evolved over these 30 years, technics such as restriction fragment length polymorphism (RFLP) (Burr, Burr et al. 1988), random amplified polymorphic DNA (RAPD) (Williams, Kubelik et al. 1990; Molnar, James et al. 2000), simple sequence repeat (SSR) (Sundaram, Naveenkumar et al. 2008), diversity arrays technology (Dart) (Wittenberg, van der Lee et al. 2005), amplified fragment length polymorphism (AFLP) (Vos, Hogers et al. 1995; Brugmans, van der Hulst et al. 2003) and microsatellite-anchored fragment length polymorphism (MFLP) (Yang, Shankar et al. 2002) have been used to develop molecular markers. These methods are effective but laborious and time consuming while the next-generation sequencing technology allows detecting large numbers of DNA markers in a short time-frame (Yang, Tao et al. 2012).

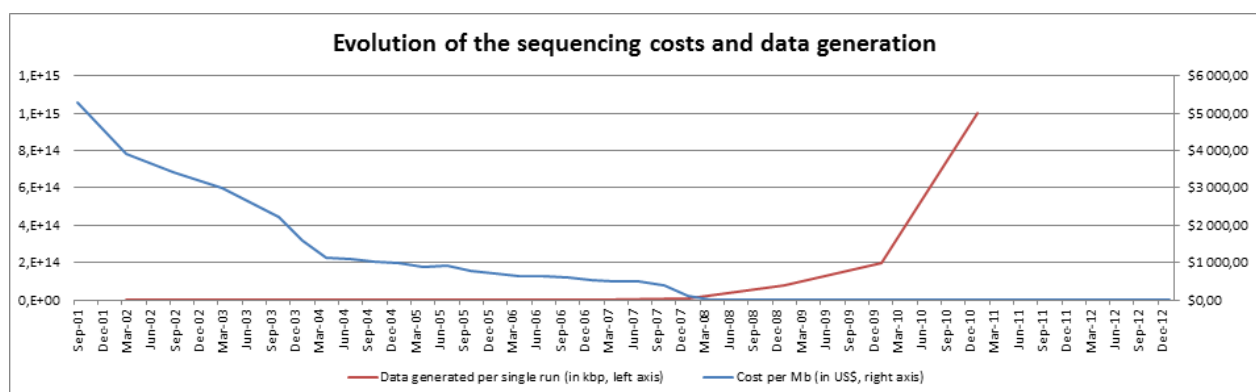


Figure 1: Evolution of the sequencing costs in US dollar per mega-base in parallel with the evolution of the data generated by a single run of a sequencer in kilo-bases. (Sources: <http://www.genome.gov/sequencingcosts/> and (Mardis 2011)).

Since the first application of 454 (Margulies, Egholm et al. 2005) and Solexa (Bennett 2004) sequencing technologies, next generation sequencing technologies (NGS) have evolved to produce millions of bases sequenced in a single run (Egan, Schlueter et al. 2012). Figure 1 presents in parallel the evolution of the number of bases sequenced in a single sequencing run using the NGS technologies and the evolution of the sequencing cost in US dollar per mega-base (1,000,000 bases). The amount of data sequenced in a single run went from 10^4 kilo-bases to 10^{12} kilo-bases in 2005 and up to 10^{14} kilo-bases in 2011, in parallel the cost to sequence 1 million bases dropped from more than US\$5,000 in 2001 to US\$0.06 in 2013.

With the development of the sequencing technology, the breeding paradigm is switching from a marker based system to a genomic base system and Marker Assisted Breeding is being replaced by Genomics Assisted Breeding. Genomics-assisted breeding is the holistic approach that tries to predict phenotypes from genotype information using genomic tools and strategies (Varshney, Graner et al. 2005; Varshney, Hoisington et al. 2006). The development of sequencing technologies has greatly improved genomics-assisted breeding by providing a way to generate large amounts of DNA-markers rapidly and at decreasing prices. Most of these DNA markers are single nucleotide polymorphisms (SNP) that have been established genome wide by the different sequencing technologies and projects. Many crops have had these SNPs made available from sequencing projects like: rice (McNally, Childs et al. 2009; Yamamoto, Nagasaki et al. 2010), maize (Barbazuk, Emrich et al. 2007), durum wheat (Trebbi, Maccaferri et al. 2011), potato (Hamilton, Hansey et al. 2011) and tomato (Hamilton, Sim et al. 2012). These SNP may then be integrated into genotyping platform such as SNP arrays (Stemers, Chang et al. 2006; Gupta, Rustgi et al. 2008) allowing scoring of thousands of markers in parallel and thus facilitating the construction of high-density genetic maps (Sim, Durstewitz et al. 2012).

While sequencing is becoming more efficient and cheaper, the number of genomes sequenced is greatly increasing to the point that projects dedicated to re-sequencing organisms are appearing to study the genetic diversity, for example the 1000 human genome project (<http://www.1000genomes.org/>) or the 150 tomato genome project (<http://www.tomatogenome.net/>). The pace of sequencing and of evolution of the technology are such that it is causing infrastructure problems to the bioinformaticians in the field (Stein 2010). However, if sequencing may be interesting for SNP calling, further use of the genome information require the sequences to be annotated and a “genome is only as good as its annotation” (Stein 2001). It is the annotation that links a DNA sequence to the biology of the organism (Stein 2001) using biological evidences collected by lab experiments.

The annotation of a genome consists of annotating the genome sequence in three levels: the nucleotide level, the protein level and the biological processes level. The nucleotide level contains, for example, known molecular markers, known and predicted genes with their introns and exons structure, repetitive elements, eventually duplication information and nucleotide variation (for example: Single Nucleotide Polymorphism, SNP). The protein level contains, among other, information about the proteins generated by the identified genes, eventually referring to known proteins from an external resource such as UniProt (The UniProt Consortium 2013) and with details such as the protein domains eventually referring to known protein domain databases such as InterPro (Hunter, Jones et al. 2012). The last level, process-level annotation is the most challenging part, which tries to link the genome sequence to biological processes. As a result, genes and proteins

will be annotated with Gene Ontology (Ashburner, Ball et al. 2000) terms (GO terms) describing the known or putative cellular locations, molecular functions and biological processes (Stein 2001).

The genome annotation often cross reference other resources (UniProt, InterPro, GO) allowing one to search for genes in other organisms involved in the same biological process or for proteins having the same protein domain. These cross-references in the annotation permit knowledge transfer from one species to another and also imply that when looking at a gene in a genome annotation, one has to query these other resources to retrieve more information about this gene, thus doing data integration which can be automated with bioinformatics tools. The genome annotation can then be used to investigate the genomic interval that a quantitative trait loci (QTL) mapping analysis links to a phenotype.

Quantitative trait locus links phenotype to the genome

If some parts of the natural variation of the plants are the results of “major genes”, much of the variation is the result of much more minor genetic changes in multiple genes (Kearsey 1998). QTL mapping is a statistical analysis linking phenotypic information (the trait of interest) with genotypic data (segregation of molecular marker over the individuals) to provide specific genomic regions linked with the studied trait (Miles and Wayne 2008). A QTL mapping study needs a population with as much variation as possible for the trait of interest. The mapping population is the result of the cross of two individuals having the most genetic diversity for the trait studied (i.e.: a different allelic composition) which will then segregate in the progeny producing different phenotypes. The resulting progeny can then be crossed again using one of the different crossing schemes (Darvasi 1998) to create the mapping population. This mapping population is then scored for as much molecular markers as possible providing a representation of the segregation in the population. The more markers, the lower the average distance between the markers, the more precise the genetic map is and the more accurate the mapping can be. The trait of interest is then measured in the population. The QTL mapping analysis consists of correlating the segregation of a marker with the measurement of the trait. The output of a QTL mapping analysis is therefore a region (also called QTL interval) on the genome, defined by molecular markers, which is statistically linked to the measured trait. For a simple trait, there might be a single QTL found while for a more complex trait there might be multiple QTL found each explaining a part of the variation that resulted in the measured phenotype (Remington and Purugganan 2003).

Molecular markers are unique genetic sites that can easily be scored and mapped in a segregating population. For most species, it is not difficult to find 10 to 50 segregating markers per linkage group (Kearsey 1998) and most markers will not influence the trait of interest but some will be correlated with it. QTL mapping relies on the principle that where such correlation occurs, the markers and the genes underlying the QTL will not segregate independently creating a “linkage disequilibrium” (Kearsey 1998). Relying on this linkage disequilibrium, differences in marker scores are associated with differences in phenotypes allows selecting plants at an early stage without having to wait for them to express (or not) the phenotype of interest. This association is the basis of the “marker assisted selection” (Johnson 2004).

QTL mapping analyses are commonly used by breeders to find regions of the genome involved in a specific trait. This region are then introgressed from a genome to another allowing to enhanced the

guest genome (McCouch 2004). These regions (commonly covering 10 to 30cM (Kearsey and Farquhar 1998)) may contain hundreds or thousands of genes (Chibon, Schoof et al. 2012) among which only few might be influencing the trait of interest (Miles and Wayne 2008), finding them has been described as the “greatest challenge facing geneticists in the twenty-first century” (Luo, Wu et al. 2002). By developing more molecular markers to enhance the resolution of the genetic map, the size of the QTL interval can be reduced significantly (to less than 1 cM) but Price (2006) mentions that this technique is not applicable to most QTL. Price (2006), however, also suggests that there will be circumstances under which mapped based cloning would help finding the candidate genes while avoiding the fine-mapping step. A third option, not considered by Price (2006), is to rely on known information to filter from a large pool of genes the potential genes of interest. This is the approach taken by Chibon, Schoof et al. (2012) (Chapter 3 of this thesis). All these approaches aim at reducing the QTL interval to reduce the list of genes and finally find out which are the genes in the QTL interval that influence the trait of interest. This gene maybe a gene encoding for an enzyme in a pathway that influences the trait measured (Kloosterman, Oortwijn et al. 2010), but it may also be that the QTL is the result of a regulatory element which influences the expression of other genes and in this way results in the expression of the phenotype of interest (Remington and Purugganan 2003; Khan, Chibon et al. 2012). Such regulatory element might be a transcription factor.

Transcription factor, a key to the gene regulatory network

In any given cell, at any given time, thousands of genes ensure the cell's function. To perform this task, genes must be expressed at a certain time and in a certain amount. This regulation is ensured by the presence of a gene regulatory network involving genes and transcription factors (Macneil and Walhout 2011). Transcription factors are proteins involved in the expression of other genes by binding to short DNA motifs, called transcription factor binding sites, in the promoter region of their target genes (Chen and Rajewsky 2007). A single transcription factor may influence the expression of several genes in the genome thus providing a coordination mechanism to control these genes (Lee and Young 2000). Finding the transcription factors as well as their binding site and thus their target genes is the first step in the understanding of the gene regulatory network of an organism.

Transcription factors may influence gene expression either positively (activate the transcription of a gene) or negatively (repress the transcription of a gene) (Latchman 1997). Understanding the regulatory network implies knowing which transcription factor regulates which genes, how and if these transcription factors are also regulated. The deregulation of some transcription factors can cause dramatic effects in any organism. In human for instance, leukemia and cancer arise among others due to the deregulation of transcription factors (Latchman 1997). Understanding the gene regulatory network is important for biologists as it helps explaining the development or non-development of certain traits in any organism but understanding the gene regulatory network is also important for breeders. In cases where a transcription factor leads to the detection of a QTL, introgressing the QTL region from a plant to another might not lead to the expected phenotype if for example the target genes are not of the right allele.

When analyzing a QTL to find the genes linked to the phenotype measured, transcription factors are elements which might be present in the QTL interval and be actually the influencing factor for the phenotype studied. Knowing the gene regulatory network allows studying the processes influenced by a transcription factor, eventually associating it to the phenotype studied. Finding precisely which

are the genes influencing a trait in a certain genome region is important as the same region may also contain genes that influence a different trait. If introgressed that region may improve a certain characteristic while damaging another, this phenomenon is called “linkage drag” (Tanksley, Young et al. 1989). But even knowing the exact gene responsible for the trait of interest can have undesirable side effects, such as pointed out by Powell, Nguyen et al. (2012), who demonstrated that the breeding for the *u* locus leading to a *uniform ripening* influences negatively the sugar content of the fruit leading to a perception of less sweet and flavorful tomatoes by the consumers (Klee and Tieman 2013).

Fine mapping and mapped based cloning are valid options to identify the gene underlying a QTL; however, they will not circumvent potential negative side-effects of the selection of a certain allele as eventually they do not provide any information about the gene. Using bioinformatics tools to aggregate and integrate the knowledge available for the genes of the QTL interval would provide information regarding the processes in which each gene is involved and give the possibility to a biologist or a breeder to learn more about the genes and how it influences the trait measured.

Data integration in (bio)informatics

The Internet is a network of computers relying on the same standard “Internet protocol suite” also known as TCP/IP (Transmission Control Protocol / Internet Protocol) to send requests and serve documents to each other. However, the Internet is just a network, the resources and services it carries compose what is called, the Web. The TCP/IP communications protocol tests were performed in 1975 and became the main communication protocol between computers in 1983.

In March 1989, Sir Tim Berners-Lee proposed a project of global hypertext while working at CERN. The work started in October 1990 and the outcome, a program called WorldWideWeb was made available in December 1990 within CERN and publicly on the Internet in the summer of 1991, that summer, the web was born.

First revolution, web 1.0

The apparition of the web in the early 1990 has been a revolution in (bio)informatics as for the first time it allowed, almost real-time, exchange of data and information between people.

However the web 1.0 was a static web where the information had to be manually generated or extracted. The web 1.0 relied on static html pages, written by hand and on which information had to be manually inserted and updated. Eventually, this led to a very stable web, where the information in one page was unlikely to change from one hour to the next. This web already supported sharing files allowing researchers to share data.

Sharing information on the web implies that one can start integrating by aggregating them from different locations. The web 1.0 had no specific mechanism or technology to perform data integration. Screen scrapping is the process of extract specific data from HTML page while ignoring graphics links and explanatory text targeted for humans to read (Stein 2002). In the web 1.0, screen scrapping was a valid approach as the content and structure of the HTML pages did not change too often.

From its launch in the summer of 1991, the Web has not stopped to expand. The number of resources and services on the Web has rapidly increased. Tilburg University and the ILK workgroup provide a website estimating the size of the World Wide Web (<http://www.worldwidewebsite.com/>) above 14 billion pages on April 8th 2013. Not only has the size of the Web changed but also its content and technology, all these changes cumulated led to the term “Web 2.0”.

Web 2.0, dynamic, social and chainable

The development of programming languages to render HTML or text content dynamically has led to the web as we know it today. It has allowed the creation of new programming languages generating dynamically HTML content on the fly, the creation of the HTTP cookies, used to track our browsing history but also allowing online shopping via the cart mechanism. It has led to the development of the social network and media sites, connecting people all across the world in a real-time fashion. Facebook started in 2003; Twitter, in 2006 but also blogging sites (such as Wordpress), Wikipedia and flickr to share images are examples of how we are consuming the Web nowadays (O'Reilly 2005).

Bioinformatics is not lagging behind regarding the development of new resources. The Nucleic Acid Research (NAR) journal publishes every year a special issue dedicated to databases available online, resources which they then keep up-to-date in their own database. The number of databases identified and recorded increases every-year and reached 1512 in the 2013 edition (Fernandez-Suarez and Galperin 2013). In addition to the NAR yearly database issue, a community driven project, Metabase is aiming at listing and organizing all databases resources available in bioinformatics. MetaBase took the form of a wiki, using the same software as Wikipedia, where anyone can edit or create a resource and help structure the information. As of August 2011, the wiki contained over 2,000 entries (Bolser, Chibon et al. 2012). Another token of this increase of bioinformatics resources is the number of publications on PubMed containing the word “database” in their title. Figure 2 shows this trend in PubMed since 1980 when only two articles mentioning “database” in their title were published, until 2012 when 1427 articles were published.

The increasing number of resources is also valid for genomic information as many of the sequenced genome were sequenced by consortium who tend to publish the data on their own resources (such as chicken (Chicken genome consortium 2004) or tomato (Tomato genome consortium 2012)).

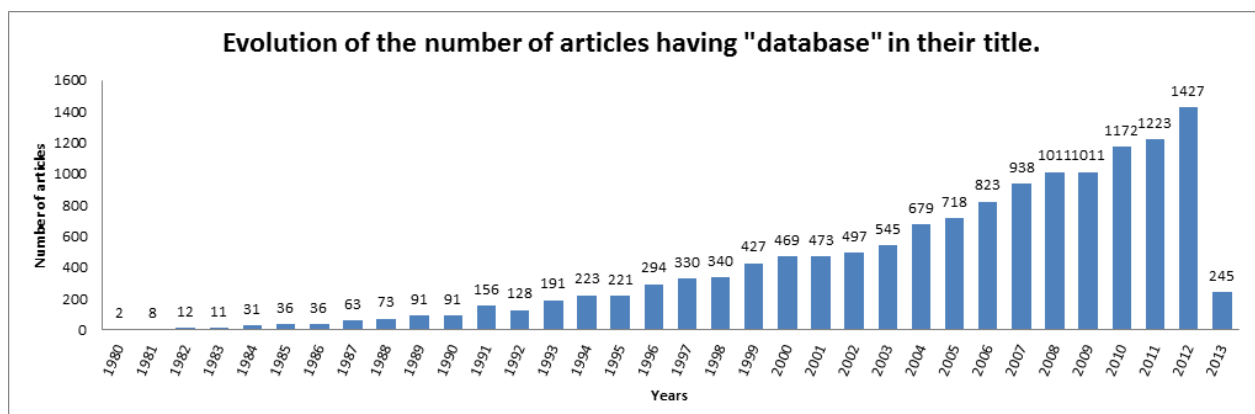


Figure 2: Evolution of the number of articles mentioning “database” in their title recorded in PubMed since 1980 (As of January 28th 2013).

Facing an increasing number of biological and bioinformatics resources, in a dynamic Web, data integration technology also had to adjust, screen scrapping turning into a “mediaeval torture” (Stein 2002). The web 2.0 is the achievement of data integration via SOAP (Simple Object Access Protocol) web-services (W3C April 27th, 2007) and REST APIs (Representational State Transfer Application Programming Interface) (Fielding and Taylor 2002).

SOAP web-services

Web-services are defined by the W3C as: “A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.” (W3C February 11th, 2004). In other words, web-services are systems designed for machine-to-machine interaction relying on HTTP and with a clearly described input and output formatted in XML (W3C January 24th, 2012). With the development of the Web 2.0, web-service is a solution for data integration allowing developers to retrieve data or perform an analysis from a remote machine. In 2006, Hull et al. found 3000 web-services publicly available in the field of molecular biology. Major data providers have set up web-services to give access to their data and resources, for example, the National Center for Biotechnology Information, better known as NCBI, provides a set of web-services to query its database (PubMed, PMC, Sequences, Gene, SNP) (Sayers and Miller 2010, January 21) but also to run analysis or tools such as BLAST (Camacho and Madden March 2nd, 2011).

Besides the clear advantages to the use of web-services (remote access to resources, designed for machine-to-machine interaction, clearly defined protocol) there are also some disadvantages, such as relying on a third party for your analysis and the lack of information about the services (Hull, Wolstencroft et al. 2006). The insufficient or just inexistent metadata about the services is one of the major drawbacks of web-services. The lack of information can be regarding the input that the service relies on or the analysis it runs. Wilkinson, Schoof et al. (2005) found 20 ways to represent a DNA sequence. It is based on this observation that they create the BioMoby project (Wilkinson, Schoof et al. 2005), a biological web-service interoperability initiative. To resolve the lack of metadata about the input of the web-services, BioMoby provides a registry where developers can register their web-service and define clearly the type of input and output their services consume. These inputs and outputs are defined using a flexible ontology structure format allowing them to be shared between services (Wilkinson, Schoof et al. 2005; Wilkinson, Senger et al. 2008). However, not all web-services are using the BioMoby framework and registry, making them eventually harder to discover and query. The BioCatalogue project tries to circumvent this void by providing an online resource listing all web-services available as well as a description of their input, output and action. The resource is also community driven, making anyone an editor of the website to add new services or provide more information about existing ones (Bhagat, Tanoh et al. 2010). Having web-services with clearly defined inputs and outputs implies that the input of one service may correspond to the output of another service. In this way services can be chained into a workflow. For example, Taverna (Oinn, Addis et al. 2004; Hull, Wolstencroft et al. 2006) is a tool that creates workflow by chaining multiple web-services in order to perform one main analysis composed of multiple steps each being a different web-service. Fisher, Hedeler et al. (2007) demonstrated the successful use of a workflow in a large-scale genotype to phenotype correlation to identify candidate genes involved in resistance against African trypanosomiasis in the mouse. However, Taverna itself does not provide any way to share the

constructed workflow in a way that someone can use to reproduce the work. MyExperiment is a resource to publish bioinformatics workflow and designed to share them with a network of researchers (De Roure, Goble et al. 2009; Goble, Bhagat et al. 2010). For example, the workflow used by Fisher, Hedeler et al. (2007) has been published on the MyExperiment website (<http://www.myexperiment.org/workflows/1661.html>).

SOAP web-services have been largely developed in the bioinformatics community but they are not the only type of services available to do data integration.

REST services

REST (Representational State Transfer) service or API (Application programming interface) correspond more to software design or architecture for services on the Web (Fielding and Taylor 2002). Among the criticisms for the SOAP web-services there is often the concern about the complexity of the process as well as performance concerns on the use of XML, which is enveloped into the SOAP message format resulting eventually into large to very large document which may become cumbersome to parse. REST services try to circumvent these problems. The UniProt Consortium (March 21st, 2012) explains in their FAQ that the data available via the UniProt website can be access via a REST interface, we will use UniProt as an example of REST API when possible.

To reduce the complexity, a REST API relies on Unique Resource Identifier that if called will return a representation of the object requested (Fielding and Taylor 2002). For example, requesting <http://www.uniprot.org/uniprot/P12345> will return the normal, HTML, page with the information about this protein. Depending on the protein, other format will be available simply by adding an extension to the url, for example <http://www.uniprot.org/uniprot/P12345.txt> to retrieve a text representation of the information about the protein or <http://www.uniprot.org/uniprot/P12345.xml> to retrieve a xml representation or <http://www.uniprot.org/uniprot/P12345.fasta> to retrieve the protein sequence.

Internet Media Type (previously known as MIME type) is meta-data provided by the web servers with the data, informing on the type of data delivered. Using this information the client can adjust its behavior. It is via these Internet Media Types that a web-browser can offer to start a music player when the user downloads a MP3 or start a spreadsheet program when the user downloads a CSV file. For UniProt, the page <http://www.uniprot.org/uniprot/P12345> has the type `txt/html`, the page <http://www.uniprot.org/uniprot/P12345.txt> has the type `txt/plain`, the page <http://www.uniprot.org/uniprot/P12345.xml> has the type `application/xml` and the page <http://www.uniprot.org/uniprot/P12345.rdf> has the type `application/rdf+xml`. One can thus adjust its parser according to the Internet Media Type returned.

Web-services have been and are still used to perform data aggregation and integration on the Web, whether they are being called directly or as part of a framework or a workflow, they are useful tools to do bioinformatics. However, they also have a number of disadvantages, some of which have been improved or worked on. Despite these efforts, they failed the data interoperability at a large scale (Wilkinson, Vandervalk et al. 2011). One of the main elements in this lack of global acceptance of the technology is the lack of semantics regarding the meaning of the element embedded in the XML input or output of these workflows. It is something that BioMoby (Wilkinson, Senger et al. 2008) tried to work on but without completely succeeding. Adding semantics to the data is intrinsic to the Semantic Web, also known as Web 3.0.

The semantic web, adding meaning to data

Why the Semantic Web

The Oxford dictionary defines Semantic as: “relating to meaning in language or logic”. The Web 1.0 as well as the Web 2.0 is designed for human consumption, with the notable exception of web-services the web is designed to be visualized, used and process by humans being able to associate concepts to chain of characters: words. Web-services are designed for machine-to-machine interaction but are mainly manually linked to each other (Wilkinson, Vandervalk et al. 2011) while with a correct description of the input and output and some reasoning one could foresee the situation where the user specifies his/her input and desired output and the programs builds the workflow calling the correct services to return the data desired. In order to do so, input and outputs have to be clearly defined in a way that a computer can “understand” and reason upon.

The Semantic Web was first mentioned in 2001 by Sir Tim Berners-Lee (Berners-Lee, Hendler et al. 2001) who is also the creator of the Web. The main idea is to transform the current web of human readable documents into a web of machine readable data, where machine to reason upon the information accessible to answer more sophisticated questions. Sir Tim Berners-Lee provides an example of what the Semantic Web could do: a person needs to make a series of appointments with a therapist, the program checks the treatment this person should receive, finds a list of therapists that could deliver this treatment, filters this list by distance to home and rating of these therapists on a trusted rating service and then gains access to each therapists’ agenda to find match between available appointment times and that person’s own agenda. If the resulting proposition of appointment is not satisfying, the search can then be made stricter regarding distance and time of the day in order to avoid the traffic jam occurring at the end of the day. Eventually, a solution could be found that would include rescheduling a couple of meetings rated less important than others in that person’s own agenda.

Transferred to bioinformatics this example could be converted to: a researcher is interested in drought resistance in potato. The program would find out that potato is a specific organism also known as *Solanum tuberosum*, that has been sequenced and whose genome annotation is available. Using a trait ontology, the program can find that drought resistance is related to the GO term GO:0042631 “cellular response to water deprivation” or the GO term GO:0009819 “drought recovery”. From there, the program can find using the genome annotation, all the genes related to one of these two GO terms but not only these two terms but also their children as the gene ontology is built as a tree where each child satisfies and specifies the condition of its parent. It can then, using the gene ontology, see in which pathways are the genes related to these GO terms and return the genes of potato involved in the same pathways. As proof for the biologists, the program could return for each assertion made the bibliographic references justifying the assertion made.

All these reasoning needs to be on the fly in order to remain valid, otherwise, you may end up with two persons having an appointment at the same time, or with an outdated gene annotation where a gene is wrongly annotated with a GO term, or wrongly missing the GO annotation.

The data representation for the Semantic Web

My cat is named Garfield

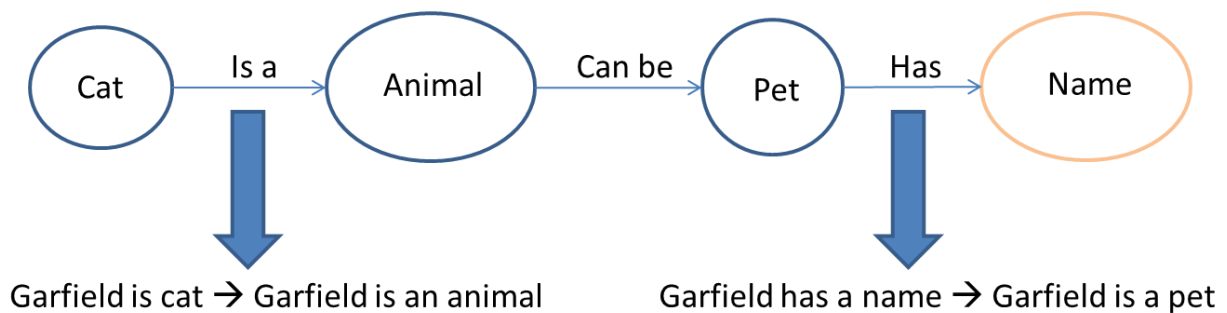


Figure 3: Knowledge that a computer needs to make assertion from the sentence “My cat is named Garfield”. The blue node represents concepts and the orange node represents an attribute. This example shows that reasoning can be based on hierarchy of concepts (cat is an animal) as well as presence of attributes (pet has name).

The Semantic Web is not a separate Web but an extension of the current one where both human and machine can extract the information available and reason upon them (Berners-Lee, Hendler et al. 2001). Transforming the Web of documents into a Web of data is the challenging part as it implies defining everything. Most humans will understand the sentence “My cat is named Garfield”, they know what a name is, they know what a cat is, that it is an animal, eventually a pet and that it can be named. A computer will only understand that something called a “cat” has an attribute “name” which is “Garfield”. To arrive to the same representation as a human, a computer will need to be taught as well that “Garfield is a cat”, “cat is an animal”, “animal can be a pet” and “name is an attribute of pet”. From these assertions, now the computer will be able to make the conclusion that if “Garfield is a cat” therefore “Garfield is an animal” and if “Garfield is an animal” and “Garfield has a name” therefore “Garfield is a pet”. Figure 3 represents graphically the knowledge needed for a computer to make these assertions.

In the example above the assertions used are using the same construction: subject, verb, complement. This is a simple way to phrase a piece of information. The Semantic Web used the same construction to represent information in the form of: subject, predicate, object (Figure 4). Subjects and objects are represented as the nodes of a graph. Predicates are represented as the edges; they are the properties that link two concepts. The graph can be further expanded as the object of a triple (an assertion) can be the subject of another triple. The Resource Description Framework (RDF) (W3C February 10th, 2004) is a framework to represent information on the web. It relies on a graph data model using triples, Unique Resource Identifier (URI) to uniquely identify concepts (subjects or objects). Object may have a specific data-type (for example, Boolean or date) or be literals (for example, number or string).

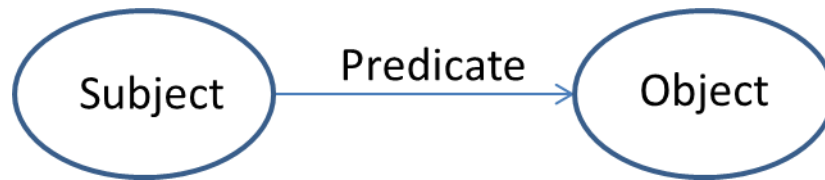


Figure 4: The underlying data structure of the Semantic Web: a triple. Each triple is consisting of a subject, a predicate (denote a property) and an object. The object of one triple can be the subject of another triple and vice-versa, leading to the creation of a graph of information.

Several formats are available to represent the graph of information built with RDF. The original format is XML based (W3C January 24th, 2012) and known as *rdf/xml* (W3C February 10th, 2004) and other formats have been developed, the N-triples format (W3C February 10th, 2004) which was originally designed for RDF test-cases, Notation3 (N3) (W3C March 28th, 2011) meant to be more human readable than XML and *turtle* (W3C February 19th, 2013) which provides some level of compatibility with N-triples. With the exception of N-triples, all these formats have a dedicated Internet Media Type (former MIME type) and can be used to represent a RDF graph.

From the importance of the URI and ontologies

In a RDF graph, subjects are always URI and objects can be either URI or literals. URI uniquely identify a concept, ideally it should also be Unique Resource Locator (URL) to which one would find more information about the concept represented by this URI. URI represents concepts but they also indicate where the concept was defined. English as well as French and other languages have homonyms where the same word has a different meaning, for example “bark” which refers both to the outer layer of a tree trunk as well as the sound a dog makes. They are also situations where a single word covers a large range of concepts, for example “ice-cream” (Figure 5). Ontologies are tools used within a community to specify the meaning this community attaches to a word.

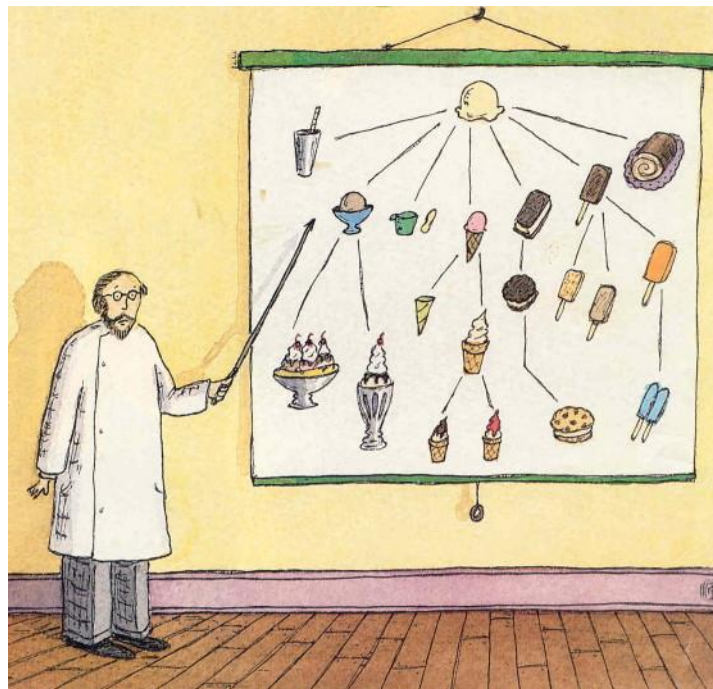


Figure 5: All the items on the boards are “ice-cream” but they are all different and can all be classified in a way that uniquely represents it, allowing anyone to order clearly from the ice-cream truck. (Source: An ontology of ice cream - Roz Chast for New Yorker cover, August 4th, 1986)

The word “ontology” is subject to debate, originally coming from philosophy where it refers to the subject of existence, it has been adopted by computer science which uses it in the context of knowledge sharing and where it is defined as: “An ontology is a specification of a conceptualization” (Gruber 1995). In other words, ontology is a description of the concepts and its relationships. Ontologies are often approached as a set of definitions allowing sharing knowledge between different parties. Gruber defines using an ontology as: “a commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology” (Gruber 1993). In bioinformatics, ontologies can be used to define the concept used such as gene sequence for example: is it the full sequence including introns and exons or just the coding sequence. In other fields such as plant breeding, ontologies can be used when collaborating on field experiments to make sure the same scales are applied when making the measurement and that each level of the scale matches between the different partners.

The Semantic Web relies on ontologies to define the concepts present in the RDF graphs in the form of URI. The URI represents not only the concept itself but also the ontology in which it is defined. Ontologies can then be mapped onto each other and a program will be able to reason and make assertions using the mapping information.

Querying the Semantic Web

RDF is a data structure to store information with a semantic context. The size of a RDF graph has no limits; the only limiting factor is the technology. Different graph database systems, also called triple stores, exist to store RDF graphs, such as Sesame (Broekstra, Kampman et al. 2002), Virtuoso (Erling and Mikhailov 2007) or Neo4j (Partner, Vukotic et al. 2013). The W3C defined a query language named SPARQL (SPARQL Protocol for RDF Query Language) specific to the graph structure of the RDF data (Prud'hommeaux and Seaborne 2008). This query language relies on pattern matching to extract the information from a triple store. Triple stores also offer a SPARQL endpoint, a service giving public access to a triple store via SPARQL allowing anyone to extract and use the data hosted in the triple store. The latest version of SPARQL (1.1) published in March 2013 (W3C March 21st, 2013) brings support for federated queries allowing to query different SPARQL endpoints over the WEB within a single query. These federated queries allow doing data aggregation and integration within a single query.

The semantic web offers a technology dedicated to data integration across multiple resources, providing different formats inter-compatible and defined in clear specifications from the W3C. The Semantic Web also provides a way to handle, via the use of ontologies and ontology mapping, the difference in vocabulary used to represent the same concept in different locations. For bioinformatics, the Semantic Web technology is one of the technologies allowing data integration over different resources while still preserving the provenance information. In addition, the semantic web relies on ontologies allowing easier collaboration between partners by setting a common vocabulary agreed upon by every partner and thus reducing the source of misunderstanding. Ultimately, the Semantic Web should allow inferring new patterns or associations automatically using reasoning algorithms and rules.

Data warehouse for data integration

The evolution of the web has changed the way data integration is performed on the fly. However, an alternative approach to on-the-fly data integration (which means that the integration is performed

when the request is sent) is data warehouse. Data warehouses are large databases built to integrate the data from different resources into one place which can then be queried locally without dependency on the network or the resource provider. There are several advantages to data warehouse: since the data is local, querying and retrieving information is normally fast. Another advantage is that doing a lot of queries will not impact the data provider and thus other users of this resource. Some tools such as Atlas (Shah, Huang et al. 2005) offer a framework to build a data warehouse, others such as SRS (Harte, Silventoinen et al. 2004) or LCB-DWH (Ameur, Yankovski et al. 2006) offer access to a data warehouse and provide analysis tools. For example SRS provides many of the EMBOSS tools (Rice, Longden et al. 2000). SRS and Atlas offer a structure to build a data warehouse but most data warehouses are built by bioinformatics departments according to their own needs by integrating their resources of interest.

Data warehouses, however, also have a number of downsides. The first issue is hardware. The larger the data warehouse, the bigger the hardware needs to be. This issue is becoming less and less of a problem with the recent development of the technologies, terabytes of hard-drives and gigabytes of RAM are more and more accessible. Another disadvantage of data warehouse is the maintenance costs. Building a data warehouse is a large task but keeping the data up to date has a costs in time and efforts. The different data providers have to be monitored for new data releases, these new releases have to be integrated into the data warehouse. This may be automated but might need some adjustments if the data structure has changed from a release to another, meaning that before starting the update one has to check if the data structure did not change, otherwise, for example, one might retrieve a gene identifier where a protein identifier was expected leading to the wrong integration of the data. Relying on tools such as Atlas might help but if the data structure changes, they also likely have to be adjusted. There are several ways to update a data warehouse, either via releases: every certain time the data provider provides a file with all the data stored at this date, or via incremental changes: every certain time the data provider provides a list of what has changed in the data stored (new entries, updates, deletion). For example, UniProt (The UniProt Consortium 2013) provides every month a copy of its data while RefSeq (Pruitt, Tatusova et al. 2012) provides daily changes as well as bi-monthly releases updates which contain and the full data and the changes between this release and the previous (Pruitt, Brown et al. 2002 (revised April 6, 2012)). However, the data changes in between releases and therefore, the information in the data warehouse can only be up to date if it is updated just after the data provider made a release.

The data warehouse approach of data integration offers a number of advantages and choosing between the data warehouse approach and the on-the-fly data integration approach is a trade-off between time, maintenance costs and dependency on the latest data (or tolerance for slightly outdated data).

Motivation and objectives of this thesis

In the context where the world population is increasing and where feeding this population is becoming a challenge, plant breeding is a key element to face this challenge. The development of Next Generation Sequencing technologies are helping speeding up the molecular breeding approaches but are also placing the breeders in the middle of a “tsunami of genome data” (Stein 2010). In addition, not only has the amount of data generated by a single experiment increased drastically, but the number of biological resources available on the web is also increasing every year

(Figure 2) coupled with the evolution of the web which has become an ever changing place. It is no longer possible for a human to know all the resources available on one topic, nor is it possible to work without the help of bioinformatics tools.

Nowadays, bioinformatics tools have to be able to handle the large amounts of data generated by the new biotechnologies, provide an overview of the results as well as the possibility to look in more details to a particular result and integrate information from the ever increasing number of resources available in order to provide the most accurate image of the state of the knowledge on a specific subject. It is within this context and with these goals in mind that this thesis was built up. The applications and analyses presented here are aiming for high-throughput datasets, visualization and data integration in order to provide the most accurate image on the subject and help biologists to build their hypotheses for their next experiments.

To improve crops in general breeders rely on wild varieties as a source of diversity containing new alleles, expressing new phenotypes. Using QTL mapping allows identifying the region of the genome containing the gene(s) responsible for this diversity. However, searching for the exact genes, the candidate genes, is much more time consuming and requires a lot of data aggregation and integration as well as a solid biological background to assess the probability that a gene might be of interest. With high-throughput QTL mapping such as performed by Khan, Chibon et al. (2012) (Chapter two of this thesis), the problem becomes even more overwhelming and require automation tools to help digging through the pile of data.

In parallel, the evolution of the web has come up with technologies dedicated to data integration. The semantic web technologies, if still under development, are raising and, we believe, are addressing a fundamental issue for bioinformatics.

In this thesis, we choose to address the question of candidate gene prediction using semantic web technologies. QTL are used on a daily basis by breeders and biologists to investigate the genetic reason underlying an observed phenotype. However, finding the specific genes (or even alleles) responsible for this phenotype in a single QTL remains a challenge. As the number of QTL found with high-throughput technologies increases, breeders and biologists need a bioinformatics approach of this question to help them reduce the list of potential candidate genes.

Searching for candidate genes means investigating a potentially large list of genes to find the one related to a specific trait. DAVID (Huang da, Sherman et al. 2009; Huang da, Sherman et al. 2009) and Ondex provide an interface to investigate gene lists. Thus, the list of the genes present in the QTL interval needs to be extracted first, this process is feasible via the genome browser (GBrowse, (Stein, Mungall et al. 2002)) but it would mean that to start looking for candidate genes, one would have to use at least three different tools: the QTL mapping software to compute the QTL and extract their interval, the genome browser (assuming one is available for their species) to extract the list of genes present in these intervals and then DAVID or Ondex to search these lists. DAVID integrates different resources and allows investigating large gene lists. However, DAVID is designed to extract the global trends in the gene list, finding what are the main processes linking these genes, what are the main GO terms, create clusters of genes in the list. However, for a QTL interval, the list of genes can be large and most genes will not be functionally related to the trait of interest. The main trends might be for house-keeping genes and the genes related to the trait of interest might be considered as noise by DAVID. Finally, DAVID only supports a limited set of organisms among which *Arabidopsis*

thaliana is the only plant, it would thus have to be extended. Ondex (Köhler, Baumbach et al. 2006) also integrates different resources, visualizes them in a graph model and offers some filtering on the graph to extract the main cluster or find the elements related to a specific entity. However, the Ondex graph visualization approach tends to clutter when confronted with large to very large datasets which renders the interpretation of the graph very complex. Ondex integrates the different data resources using its own data structure. A number of data sources and structures are supported for import (i.e.: GFF, OBO, SBML and for the data source: UniProt, TAIR, KEGG, Gramene) but it does imply that a single change in the data structure from these data providers and the code has to be adjusted. Ondex could be used to search a list for genes involved in a specific trait. However, like DAVID the visualization aims at showing the global trends.

The semantic web, although being more than 20 years old, is starting to appear in bioinformatics. Resources such as UniProt are providing their information in RDF as a standard format. The semantic web allows easier data integration. Each concept is uniquely identified, meaning that if a change occurs in the data structure, the concepts will change as well and thus there is no risk of retrieving incorrect information by error. The use of ontologies in the semantic technologies allows mapping a concept from a data provider to another concept of another provider if they used two different ontologies. Eventually, integrating multiple resources is equivalent to building a large meta-ontology that maps the same concept from the different resources. Via this mechanism, these different resources can be integrated on-the-fly and queried as one while still preserving the original structure from the data provider making maintenance easier and reducing the risk of false information.

While working on candidate gene prediction, we encountered other needs, such as the visualization of high-throughput QTL mapping experiments, or the ability to search the complete genome for genes matching certain criteria in their annotations. We also encountered cases where our biological knowledge was the limiting factor and we thus worked on predicting transcription factor binding sites in the tomato genome sequence, in order to integrate this information back into our tools to improve their outcome.

Outline of this thesis

This thesis presents analysis, visualization tools and research tools developed in the context of high-throughput analyses.

Chapter two is an example of a high-throughput QTL analyses. An apple segregating population has had been measured for its metabolites content. Several hundreds of metabolites were detected and further investigated in an untargeted QTL mapping analyses performed by MetaNetwork. More than 660 QTL were detected which led to the discovery of a hotspot of QTL in linkage group 16. The QTL of this hotspot have then been further analyzed using MapQTL, the metabolites have been annotated revealing that the expression of these metabolites with a QTL on linkage group 16 is related to the phenylpropanoid biosynthetic pathway.

Chapter three is built upon the outcome of chapter two. In chapter two, two different QTL mapping tools have been used for the QTL mapping analysis because MapQTL is able to perform QTL mapping on several hundreds of traits but cannot provide an overview of the output. MQ² has been created as an answer to this problem. MQ² is a visualization program designed for high-throughput QTL mapping analysis allowing biologists and breeders to keep using the QTL mapping tool they are used

to while still benefiting from a visualization tool providing them with an insight on the distribution of the QTL along the genetic map. Nowadays, MQ² could be used to generate Figure 1 of chapter two for example.

Chapter four presents a tool aiming at helping biologists and breeders to investigate the outcome of their QTL mapping analysis. As outlined in this introduction, it is important to understand the biological mechanisms as well as for breeding purposes to find the gene(s) underlying a QTL. Marker2sequence uses data integration from genome annotation to retrieve the list of all genes present in a region of the genome and aggregate as much known information as possible. It will provide a list of all these genes with the possibility to learn more about each of them as well as a possibility to search in the annotation of these genes to filter out genes of potential interest which might be used as hypothesis for further research.

If Marker2sequence presented in chapter four already provides some insight into the genes underlying a given QTL, one of its current limitations is the lack of integration of information regarding the regulatory gene network present in each genome. In **Chapter five** data from the genomic structure of the introgression lines built from the cross between *Solanum chmielewskii* LA1840 and *Solanum lycopersicum* cv. Moneyberg and gene expression data have been combined to study transcription factor binding sites. By comparing the gene expression between some progeny plants and the parent, we were able to retrieve a list of genes differentially expressed in the progeny. These genes are cis-genes (located in the introgression region) or trans-genes (located outside the introgression region). This study searched for DNA motifs present in the promoter region of trans-genes differentially expressed in genotypes with a *S. chmielewskii* introgression compared to genotypes without. These DNA motifs are potential TFBS of genes regulated by transcription factors present in the introgression regions. 17 DNA motifs, potential transcription factor binding sites have been found, which will need to be validated in the lab but are a first step in building the gene regulatory network of tomato.

Chapter six presents Annotex, a tool to explore genome annotation. Annotex provides a way to ask any question to a genome annotation or the network of information surrounding it. For a specific species, providing an input, an input type and an output type, Annotex browses its different resources to return the type of information asked related to the provided input. Using Annotex one can retrieve all the genes related to a protein, all the genes related to a GO term or all the proteins associated with a pathway. The lists returned could then be combined, subtracted, and intersected to provide a list combining the different factors.

Finally, **Chapter seven** is a general discussion on what was achieved with this thesis. It provides some insight on how this thesis is included in the work being done in the field and what remains to be done to reach an ideal workflow for high-throughput candidate gene predictions from high-throughput QTL mapping analyses.

Chapter 2: Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on Linkage Group 16

Sabaz Ali Khan^{1*}, Pierre-Yves Chibon^{1*}, Ric C.H. de Vos^{2,4,5}, Bert A. Schipper^{2,4,5}, Evert Walraven³, Jules Beekwilder², Thijs van Dijk¹, Richard Finkers², Richard G.F. Visser¹, Eric W. van de Weg², Arnaud Bovy^{1,5}, Alessandro Cestaro⁶, Riccardo Velasco⁶, Evert Jacobsen¹ and Henk J. Schouten²

¹Wageningen UR Plant Breeding, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands

²Wageningen University and Research Centre, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

³Wageningen University and Research Centre, Lingewal 1, 6668 LA Randwijk, The Netherlands

⁴Netherlands Metabolomics Centre, Einsteinweg 55, 2333 CC Leiden, The Netherlands

⁵Centre for BioSystems Genomics, P.O. Box 98, 6700 AB, Wageningen, The Netherlands

⁶Istituto Agrario San Michele all'Adige Research and Innovation Centre, Foundation Edmund Mach, Trento, Italy

*These authors contributed equally to this study.

Abstract

Apple (*Malus x domestica* Borkh.) is among the main sources of phenolic compounds in the human diet. The genetic basis of the quantitative variations of these potentially beneficial phenolic compounds was investigated. A segregating F1 population was used to map metabolite quantitative trait loci (mQTL). Untargeted metabolic profiling of peel and flesh tissues of ripe fruits was performed using liquid chromatography-mass spectrometry (LC-MS), resulting in the detection of 418 metabolites in peel and 254 in flesh. In mQTL mapping using MetaNetwork, 669 significant mQTL were detected: 488 in the peel and 181 in the flesh. Four linkage groups (LGs) i.e. LG1, LG8, LG13 and LG16 were found to contain mQTL hotspots, mainly regulating metabolites that belong to the phenylpropanoid pathway. The genetics of annotated metabolites was studied in more detail using MapQTL®. A number of quercetin conjugates had mQTL on LG1 or LG13. The most important mQTL hotspot with the largest number of metabolites was detected on LG16: mQTL for 33 peel-related and 17 flesh-related phenolic compounds. We located structural genes involved in the phenylpropanoid biosynthesis pathway, using the apple genome sequence. The structural gene leucoanthocyanidin reductase (LAR1) was in the mQTL hotspot on LG16, as were seven transcription factor genes. We believe this is the first time that a QTL analysis was performed on such a high number of metabolites in an outbreeding plant species.

Key words: *Malus x domestica* Borkh., untargeted and targeted mQTL mapping, genetical metabolomics, LC-MS, MetaNetwork, MapQTL.

Introduction

The fruit of apple (*Malus x domestica* Borkh) is a rich source of phytochemicals including phenolic compounds (Gerhauser 2008). There is increasing evidence that apple is an important source for various compounds that are beneficial for human health. For example, its consumption has been associated with a risk reduction of many human diseases, such as asthma, type-2 diabetes, thrombotic stroke, ischemic heart disease, and various cancers (Eberhardt, Lee et al. 2000; Mcghee, Hunt et al. 2005). Some of the major phenolic compounds isolated and identified from apple are procyanidins, anthocyanins, chlorogenic acid, hydroxycinnamic acid, flavan-3-ols such as; (-)-epicatechin, (+)-catechin, and gallatecatechin; phloridzin and quercetin glycosides (Mazza and Velioglu 1992; Lu and Foo 1997; Awad, De Jager et al. 2000; Treutter 2001).

The current study aims at elucidation of the genetic basis of metabolic variability in apple fruits. We initiated this study without any *a priori* with regard the specific metabolites groups. For that reasons we chose for large-scale LC-MS based metabolic profiling.

Metabolomics is defined as the large scale analysis of metabolites in an organism, and it concerns the simultaneous measurement of these metabolites in a given biological system (Dixon and Strack 2003). Metabolomics is developing as an important functional genomics tool in crop plants, including fruit trees (Carrari and Fernie 2006; Moco, Bino et al. 2006). Although QTL have been mapped in cultivated apples for different traits such as disease resistance (Calenge, Faure et al. 2004; Calenge and Durel 2006; Khan, Duffy et al. 2006), fruit quality (King, Lynn et al. 2001; Liebhard, Kellerhals et al. 2003; Davey, Kenis et al. 2006) and tree morphology (Kenis and Keulemans 2007), there is only

one report on the genetic mapping of a large number of metabolites in apple fruits, and, in that case, on volatiles (Dunemann, Ulrich et al. 2009).

The LC-MS metabolomics showed numerous metabolic compounds in the segregating F1 population, both in peel and in flesh of the fruits, allowing mQTL (metabolomic QTL) mapping. Standard QTL mapping software is designed to map individual traits, one by one, and is not suited to map hundreds of metabolites simultaneously. Therefore we decided to use the software MetaNetwork (Fu, Swertz et al. 2007). MetaNetwork enables simultaneous genome-wide screening of numerous traits. Keurentjes, Fu et al. (2006) used MetaNetwork to find mQTL for secondary metabolites in an *Arabidopsis thaliana* recombinant inbred line population.

In the current research metabolites from untargeted metabolic profiling were mapped. The majority of the mapped metabolites belong to phenylpropanoid pathway. A major mQTL hotspot was found on LG16. Only one structural gene *leucoanthocyanidin reductase (LAR1)* was detected in the mQTL hotspot, as were seven transcription factor genes. We believe that this is the first time that such a large mQTL mapping was performed in a highly heterozygous and cross pollinating crop species like apple.

Materials and methods

Plant Materials

For mQTL mapping, a segregating F1 population from the cross 'Prima' x 'Fiesta' (PF) was used. This population was also used for the first international reference linkage map of apple covering all chromosomes (Maliepaard, Alston et al. 1998). In this study a subset of 113 progenies and both parents was used. For the F1 population, two trees per genotype were present.

Harvesting and storage of the apples

Mature fruits of all genotypes were harvested in September and October, 2008 in a trial orchard in Elst, The Netherlands. The maturity of the fruits was assessed by checking the colour of the peel, the taste, and the browning of the seeds. For each progeny, more than ten fruits from each of the two trees were harvested separately, while for the two parents 'Prima' and 'Fiesta' fruits from five trees (five replicates) were harvested. The fruits were harvested randomly from different sides of each individual tree to level out possible differences due to environmental factors such as light. Fruits were over sampled in the field to forestall the possible damage or decay during transit. After harvesting, fruits were immediately stored at 0 °C in a cold storage room to minimize enzymatic activities. Once fruits for all of the genotypes were harvested, these were shifted to a storage room at 20 °C for seven days. This was done to mimic the storage conditions in a consumer's household.

Selection and grinding of apples

Samples of eight apples per genotype were selected. For the progeny genotypes, four apples from each of the two trees of one genotype were combined as one sample, giving one replicate per genotype. For each sample, the individual fruit was cut transverse wise to obtain a 1-cm thick round slice, and the round slice was peeled. The peel (1.4 mm thick) was chopped into small pieces and snap-frozen in a separate beaker with liquid nitrogen. The core was removed from the flesh and a slice (3.2 mm thick) of the flesh was also chopped into small pieces and snap-frozen. This was repeated for all of the eight apples of one genotype, and the samples from a tissue were pooled per

genotype. The samples were then ground using an IKA coffee grinder (model A11 basic). The powder for the flesh and peel was collected separately in 50-ml falcon tubes and stored at -80 °C. For the parents 'Prima' and 'Fiesta', the samples were treated separately in five replicates each and treated in the same way as described for the progenies.

Extract preparation

The aqueous-methanol extracts were prepared as described by De Vos, Moco et al. (2007), with minor modification (Keurentjes, Fu et al. 2006). Ice-cold 99.9% methanol (1.5 ml) acidified with 0.133% (vol/vol) formic acid, was added to each plant sample (final methanol concentration of 75%, assuming 90% water in the 500 ± 5 mg tissues). The ensuing steps, from sonication to the injection of the samples and separation using the Alliance 2795 HT system, were performed as described by De Vos, Moco et al. (2007). The separation was performed at 40 °C, by applying a 45 min gradient of 5-35% acetonitrile in water (acidified with 0.1% formic acid) at a flow rate of 0.19 ml/min. The compounds eluting from the column were detected online, first by a Waters 996 photodiode array detector at 200-700 nm and then by a Q-TOF Ultima MS (Waters) with an electron spray ionisation (ESI) source. Ions were detected in negative mode in the range of m/z 80 to 1,500 at a resolution of 10,000, using a scan time of 900 ms and an interscan delay of 100 ms. The desolvation temperature was 250 °C, with a nitrogen gas flow of 500 l/h, the capillary spray was 2.75 kV, the source temperature was 120 °C, the cone voltage was 35 V with 50 l/h nitrogen gas flow and the collision energy was 10 eV.

The mass spectrometer was calibrated as described by De Vos, Moco et al. (2007). MassLynx software version 4.0 (Waters) was used to control all instrumentation and for the calculation of accurate masses.

Pre-processing the dataset

Unbiased mass peak picking and baseline correction of the raw LC-MS data were performed using Metalign software (De Vos, Moco et al. 2007; Lommen 2009); www.metalign.nl) with a signal to noise ratio of 3 or higher. Thus, a total of 18,582 mass signals were extracted from the peel samples and 11,817 signals from the flesh samples. Both peaks lists were subsequently filtered for signals present in more than 10 samples, resulting in 4830 peel signals and 2826 flesh signals. A single metabolite may produce a number of mass peaks, due to natural isotopes, unavoidable fragmentation and adduct formation in the ESI source, resulting in data redundancy. Therefore, mass signals originating from the same metabolite were grouped, based on their corresponding retention time and intensity pattern over samples, using MSClust software (Tikunov, Laptinok et al. 2011) that can be freely downloaded from the Metalign website (www.metalign.nl). From the clustered mass signals, i.e. reconstructed metabolites, the most representative signal per mass peak cluster was taken for further data analyses. The metabolite signals were ¹⁰log transformed to normalise the variances among the metabolites.

Genetic linkage maps

Genetic linkage maps were available for both 'Prima' and 'Fiesta', representing the 17 linkage groups of apple. The maternal map consists of 562 markers and the paternal map consists of 452 markers, including DArT, AFLP, RFLP, NBS-LRR, SSRs, RAPD markers and some isozymes (Schouten, van de Weg et al. 2011).

In the untargeted mQTL mapping of metabolites in apple, the individual maps of 'Prima' and 'Fiesta' were used, as the MetaNetwork could not incorporate the integrated map of cross-pollinating crops such as apple. In the targeted mQTL mapping using MapQTL[®] 6.0, an integrated map of both parents was constructed and used for the analysis of the annotated metabolites. The integrated map contained 801 markers, spanning 1,348 cM.

Untargeted mQTL mapping of metabolites using MetaNetwork

MetaNetwork was designed for the mQTL analysis of homozygous recombinant inbred line (RIL) populations of inbreeding plants such as *A. thaliana* (Fu, Swertz et al. 2007), and was therefore not applicable for the analysis of a segregating F1 population of an out-crossing species. Therefore, we transformed the data, using single parental maps, giving 2x17=34 linkage groups. Hereby each parent was considered to be derived from a cross between two inbred lines, and the F1 progeny was considered to be the result of a backcross. The linkage phase information from the linkage map was used to assign F1 marker alleles to the respective parental inbred lines, thus giving the dichotomous marker scores as required by MetaNetwork. Missing marker data were imputed using information of flanking markers if they were within a 20-cM distance and in a non-recombinant segment.

MetaNetwork implements a two-part parametric model per trait, combining a non-parametric approach (Wilcoxon-Mann-Whitney test; (Brem, Yvert et al. 2002; Yvert, Brem et al. 2003) with a parametric test (ANOVA). The non-parametric test uses a user-defined spike value to distinguish qualitative segregation from quantitative differences. The value chosen as spike was 37, because this value was the noise level in the LC-MS analysis. MetaNetwork also allows setting a threshold for the significance of mQTL by performing permutation tests on samples. A bootstrap procedure was performed with a type I error of 5 % (default value of MetaNetwork) for finding an mQTL considering all genetic markers. This procedure rendered a $^{-10}\log(P)$ threshold of 3.8 for individual marker – trait combinations. This threshold was used for all analyses.

Annotation of metabolites

MetaNetwork revealed clusters of mQTL on the apple genome. We annotated the underlying LC-MS mass peaks of mQTL clusters by comparing their accurate mass and retention time with standards, with the metabolite databases Moto (Moco, Bino et al. 2006), KNapSack, Metabolome Japan, and the Dictionary of Natural Products. The results are shown in Tables S1 and S2.

Targeted mQTL mapping of annotated metabolites using MapQTL[®] 6.0

We aimed at studying the genetics of these annotated metabolites in more detail, such as revealing the allelic contributions to traits, and performing co-factors analysis to filter out the effect of strong mQTL. MetaNetwork appeared to be less suitable for these deeper analyses of individual metabolites. Therefore, we decided to use MapQTL[®] (Van Ooijen, 2009) for this. We applied interval mapping, followed by rMQM mapping with regression algorithm, Haldane's mapping function, with a mapping step size of 1, and independent LOD (logarithm of odds) test statistics was used. The threshold for mQTL significance was determined using a genome-wide permutation test with 1,000 iterations, which gave $\alpha = 0.005$ for the 17 chromosomes of apple, to obtain a 95% confidence interval. Two LOD support intervals were used to estimate the range in cM where the mQTL reside. Markers near mQTL peaks or at mQTL peaks were used as co-factors for rMQM mapping. This was followed by another round of co-factor selection by using markers from the newly found minor

mQTL from the rMQM. The results from this second round were recorded as the final result. An mQTL was named as minor QTL if its LOD score was close or just at threshold level.

Mapping of the metabolites that segregated as a monogenetic trait

The metabolites procyanidin dimer I, procyanidin dimer II, procyanidin trimer I, procyanidin trimer II, (+)-catechin and (-)-epicatechin had only one mQTL, segregating in a clear 3:1 ratio. This single locus explained a major part of the variation in the metabolite content (up to 81 %, Table S3). These metabolites were treated as monogenic traits and were integrated into the genetic linkage map by JoinMap 4.0 (Van Ooijn 2009). This was performed with the aim of locating the positions of the underlying genes more precisely.

Testing additional simple sequence repeats (SSR) loci for LG16

To map the monogenetically segregating metabolites more precisely, 17 additional SSR loci at the upper part of LG16 were tested for the 42 progenies that showed recombination in this genetic area. These SSRs along with their primers, have been previously published in other apple molecular marker linkage maps (Liebhard, Gianfranceschi et al. 2002; Kenis and Keulemans 2005; Silfverberg-Dilworth, Matasci et al. 2006; Celton, Tustin et al. 2009). The 17 SSR loci along with their primer sequences are listed in Table S7.

Locating structural genes of the phenylpropanoid pathway in the apple genome

To find the position of the orthologous genes on the 17 chromosomes of apple, the DNA sequences of the structural genes of *A. thaliana* (Table 2 in Lillo, Lea et al. 2008), were aligned to the entire genome sequence of the apple cv. 'Golden Delicious' (Velasco, Zharkikh et al. 2010).

Results

Untargeted mQTL mapping of metabolites showed 669 mQTL in peel and flesh

From the cross of 'Prima' x 'Fiesta', 113 progeny individuals were analyzed by accurate LC-MS. A total of 18,582 and 11,817 mass signals were detected in peel and flesh tissues respectively. Clustering of the mass signals based on their corresponding retention time and abundance profile across samples resulted in 672 centrotypes: 418 and 254 for peel and flesh, respectively. In the following sections these centrotypes are named metabolites.

In view of the genetic mapping, the distributions of these metabolites were studied. The $^{10}\log$ transformation appeared to provide Gaussian distributions in both peel and flesh for the majority of the metabolites follow (data not shown). In the untargeted mQTL mapping using MetaNetwork, a total of 669 mQTL were detected (Table 1), spread over all of the 17 linkage groups of the apple genome (Figure 1). Not all of the metabolites showed mQTL; 50% of the metabolites in peel and 44% in flesh exhibited statistically significant mQTL. Figure 2 shows that several mQTL had very high $-\log(P)$ values.

Table 1: MetaNetwork results of mQTL mapping in peel and flesh of apple from F1 mapping population from 'Prima' x 'Fiesta'

	'Prima' Peel	'Prima' Flesh	'Fiesta' Peel	'Fiesta' Flesh	Total Peel	Total Flesh	Total
Number of metabolites with at least one mQTL	184	77	169	67			
Number of mQTL	288	101	200	80	488	181	669
Number of markers with at least one mQTL	133 (28%)	62 (13%)	81 (24%)	50 (13%)	214	112	326

LG16 has a strong hotspot of mQTL in both parental genotypes and in both peel and flesh for phenolic compounds

The striking thing in Figs. 1 and 2 is the strong hotspot of mQTL on LG16. On other linkage groups such as LG1 and LG13, also many mQTL were detected, but these were not as strongly clustered as on LG16. The mQTL on LG16 clustered mainly around a single locus. Notable also is that the hotspot of mQTL on LG16 was present in both parents and in both tissues, in contrast to the mQTL hotspot on LG8, which was explicitly present in 'Prima' but absent in 'Fiesta' (Figure 1).

For peel 69 and for flesh 30 metabolites were annotated (Tables S1, S2). Of the annotated metabolites, 81 out of 99 were phenolic compounds belonging to the two groups of phenylpropanoids and polyphenols (Tables S1, S2).

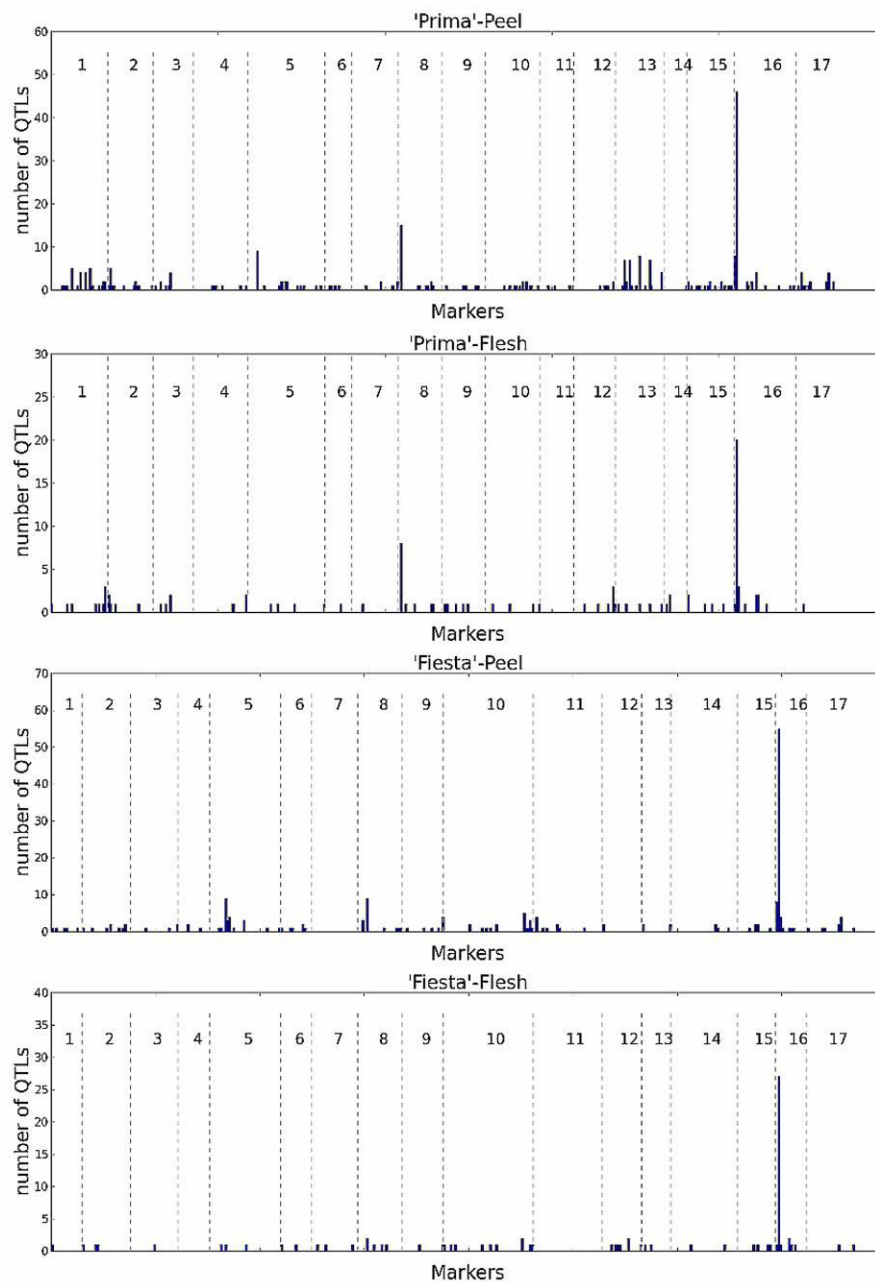


Figure 1: Number of mQTL over the apple genome. The linkage groups are separated by vertical dotted lines. In this figure, markers are ordered and positioned equidistantly, thus ignoring their genetic distances.

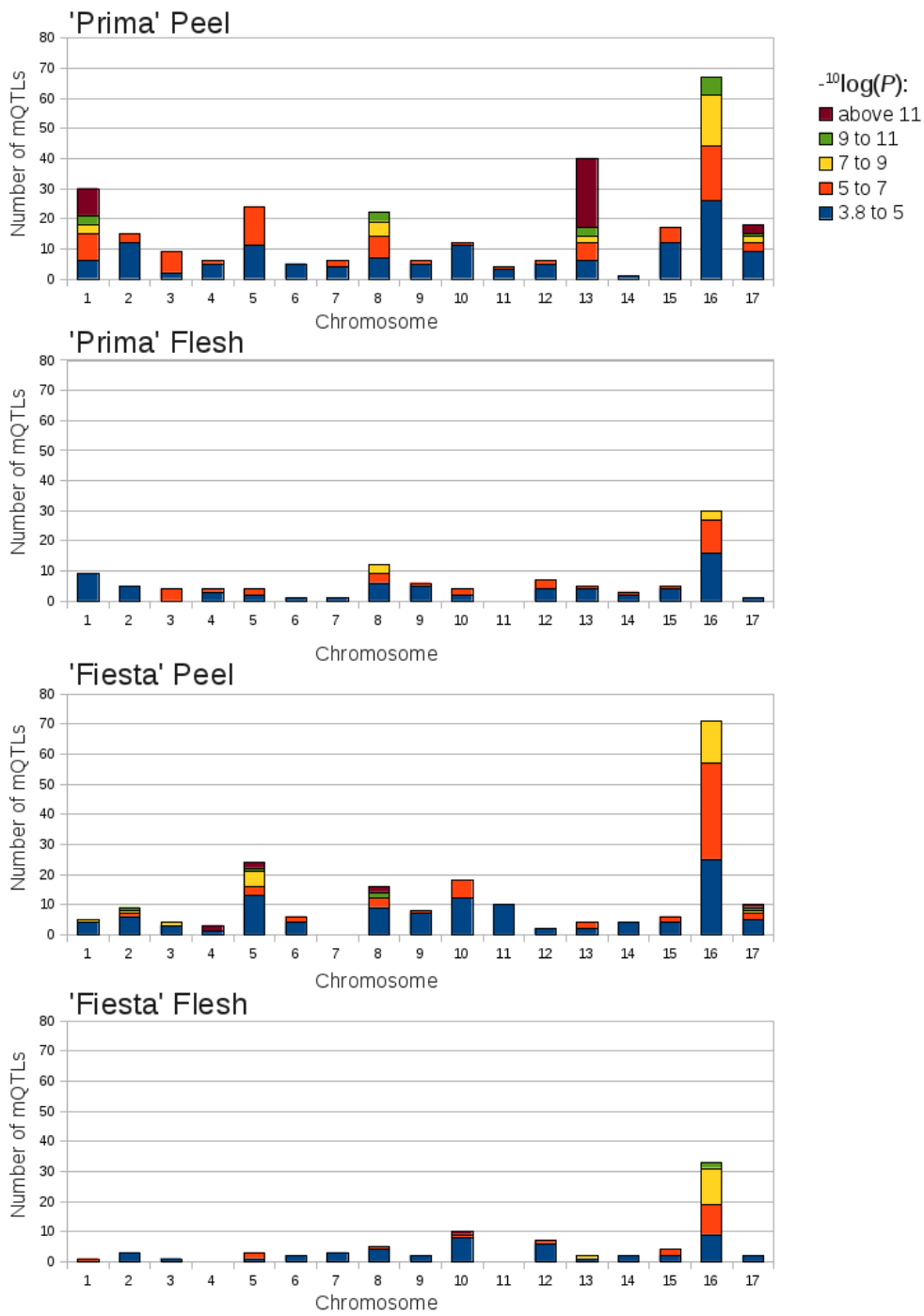


Figure 2: Significant mQTL with range of log p values over the apple genome. An mQTL was considered as significant if its log p value was higher than 3.8.

The mQTL hotspot on LG16 is not caused by the co-localizing major locus for pH

Maliepaard, Alston et al. (1998) previously mapped the pH of apple fruits on LG16 in the same segregating population. They observed monogenic inheritance for low versus high acidity and mapped the corresponding gene. We found that our current mQTL hotspot on LG16 was close to that gene (Figure 3). Remarkably, both parents had one dominant allele for low pH at the LG16 mQTL hotspot. They denoted the locus as *Ma* for malic acid, being the major acid in apple, although they measured pH rather than malic acid itself (Maliepaard, Alston et al. 1998). As the pH might influence different enzymatic processes and biochemical reactions in plant cells, differences in the pH may possibly have caused the mQTL hotspot. We evaluated this hypothesis. Both at low and high pH, high levels of the metabolites were found (Figure 4). Apparently, the dominant allele for high acidity was in repulsion to the allele for high level of metabolite content in both parents. Consequently, the occurrence of the hotspot was not a side effect of major differences in the pH.

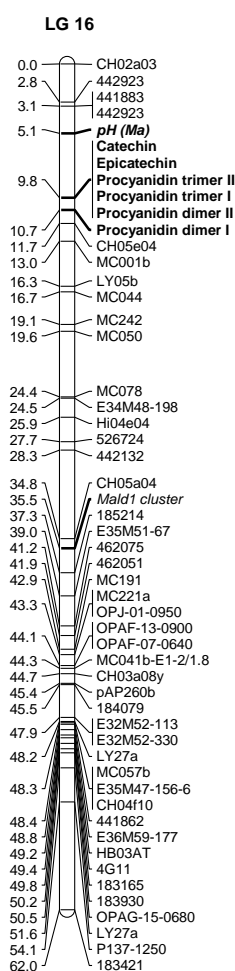


Figure 3: Mapping of (+)-catechin, (-)-epicatechin, several procyanidins and *pH (Ma)* on linkage group 16.

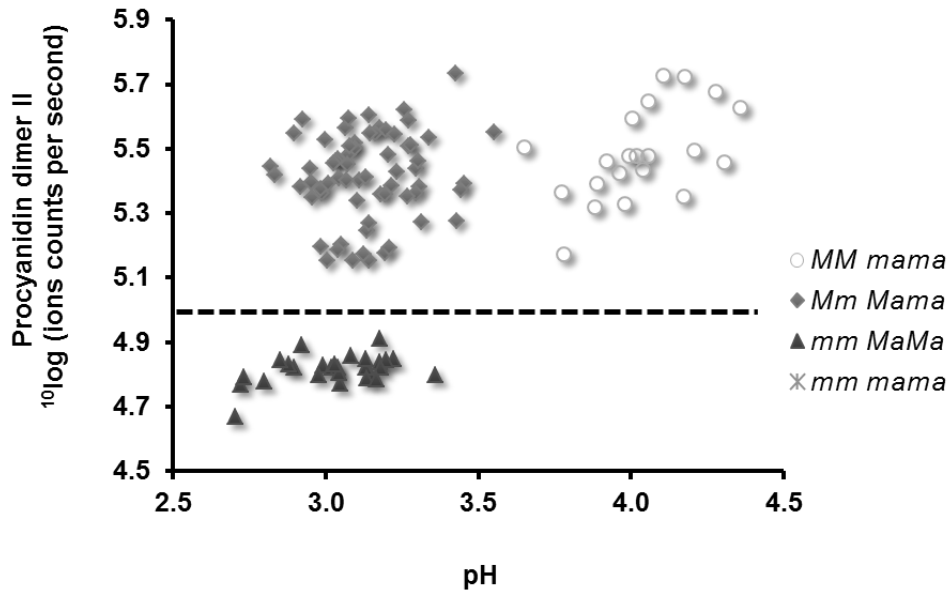


Figure 4: A scatter plot showing the distribution of F1 progenies of 'Prima' x 'Fiesta' over the four genotype classes for low/high pH and low/high procyanidin dimer II content, whereby procyanidin dimer II represents the metabolites that share the strong mQTL on LG16. The trait pH co-localizes to this hotspot. The dominant allele for high metabolite level is denoted as M, and for low pH (high acidity; presumably high level of malic acid) as Ma. As the dominant alleles M and Ma are in repulsion phase in both parents, giving as alleles in the gametes Mma and mMa, the progeny segregates into three genotypes, lacking the genotype mm mama. The horizontal dashed line represents the 3: 1 clear segregation for the procyanidin dimer II that shows that the two classes (i.e. mm and Mm + MM) show the full dominance.

Targeted mQTL mapping

The genetics of these annotated metabolites was studied in more detail using MapQTL[®] 6.0. Similar to the MetaNetwork analysis, the hotspot with the highest number of mQTL was detected on LG16. Among the 69 annotated metabolites in the peel, 33 had an mQTL on LG16 (Figure 5A, Table S3). The majority of these metabolites represented procyanidins of various chain lengths, including the monomeric building blocks of procyanidins, the flavan-3-ols (+)-catechin and (-)-epicatechin. The same region of LG16 showed mQTLs for quinic acid, phenolic esters, coumaroyl hexoside, kaempferol glycosides, and phloridzin. Interestingly, all these metabolites originate from the phenylpropanoid pathway (Figure 5). Most of the metabolites had only a single mQTL; however, a few metabolites were found to have some additional minor mQTL on other LGs (Figure 5). A restricted multiple QTL mapping (rMQM) analysis revealed in addition several minor mQTL (Tables S3, S4).

For the 69 annotated metabolites in the peel tissue, mQTL were located on five different linkage groups (Figure 5A, Table S3). LG1 contained specific mQTL for quercetin glycosides. There is also an mQTL for Kaempferol glycosides on LG1. Glycosides of isorhamnetin had an mQTL on LG13. Chlorogenic acid showed an mQTL on LG17. A distinguished group of mQTL, mapped on LG8, was formed by alcohol glycosides such as octane-di-ol hexoside and phenylethanol glycoside (Table S3).

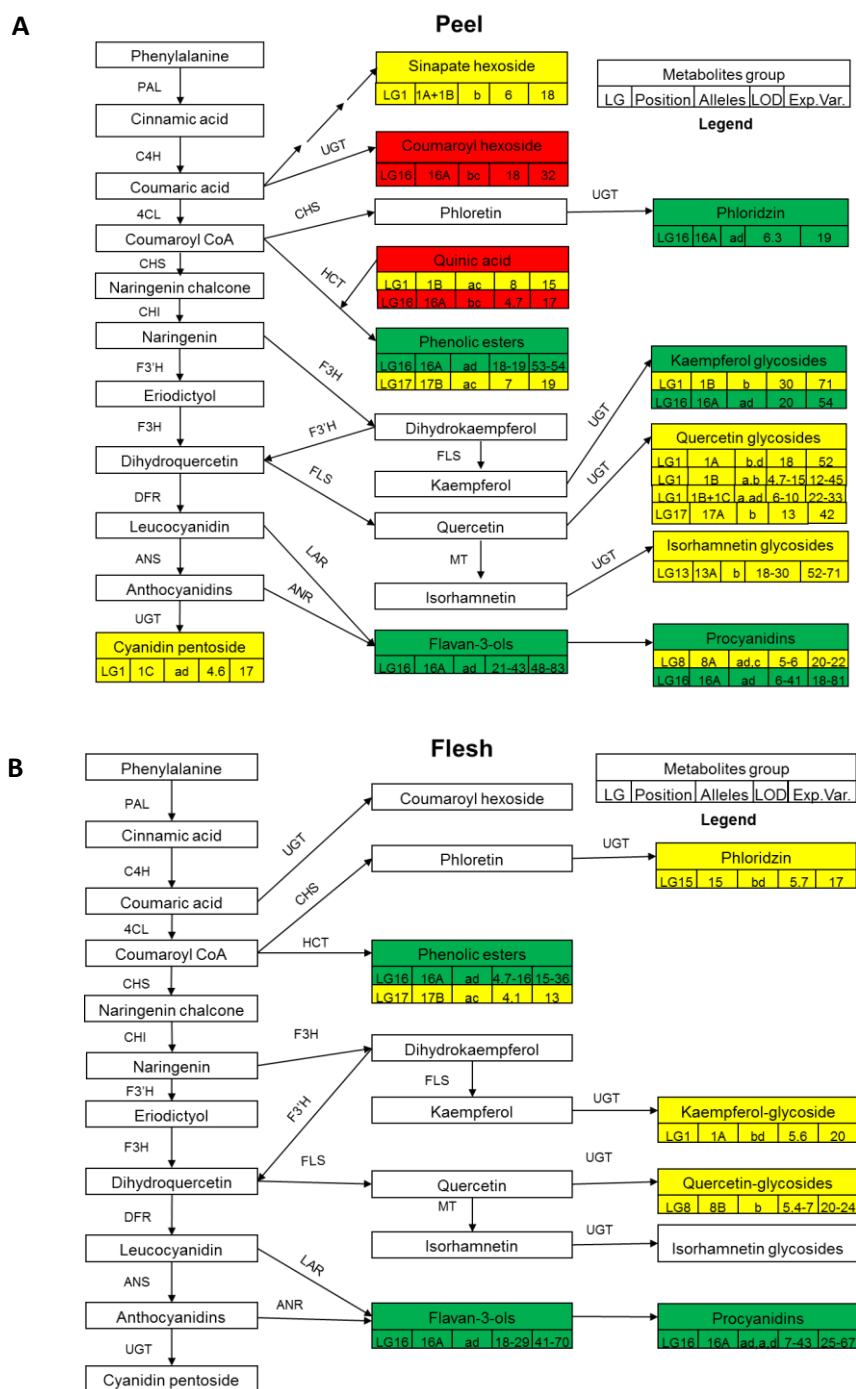


Figure 5: The phenylpropanoid pathway of phenolic compounds in two apple fruit tissues, peel (A) and flesh (B). The metabolites for which mQTL were found are presented in colored boxes. Colorless boxes show the metabolites that were not detected in our analysis or have no mQTL. Boxes with green color indicate mQTL of which the + alleles are in coupling phase. Boxes with yellow color show mQTL for metabolites other than on LG16A. The metabolites in the red box show a negative correlation with the metabolites in the green boxes, having an mQTL on LG16A. The linkage group (LG) where an mQTL was located is given. If different mQTL were present on different regions of a LG, these regions are distinguished with the letters A, B, C etc. The alleles 'a' and 'b' originate from the parent 'Prima', and the alleles 'c' and 'd' originate from the parent 'Fiesta', thus following JoinMap codes for outcrossers. As many metabolites in the phenylpropanoid pathway were mapped, for the purpose of simplicity, metabolites that belong to a similar group of compounds are shown as a group (e.g. phenolic esters is a group of several metabolites). Gene names are abbreviated as: phenylalanine ammonia-lyase (PAL), cinnamate-4-hydroxylase (C4H), 4-coumaroyl:CoA-ligase (4CL), chalcone isomerase (CHI), chalcone synthase (CHS), flavonone 3' hydroxylase (F3'H), dihydroflavonol 4-reductase (DFR), hydroxycinnamoyl-CoA quinate/shikimate hydroxycinnamoyl transferase (HCT), leucoanthocyanidin 4-reductase (LAR), UDP-glycosyltransferase (UGT), flavonol synthase (FLS) and anthocyanidins synthase (ANS).

The genetic loci controlling metabolites content in the peel also appear to control these in the flesh although less significantly

Like those in peel, mQTLs in flesh were detected on the same five linkage groups. Most of the metabolites showing mQTL in the peel also showed mQTL in the flesh; however, the number of mQTL in the flesh was lower than in the peel (Figs. 1, 2). Like in peel, in flesh also kaempferol glycosides had an mQTL on LG1 and LG16. In contrast to peel, quercetin rhamnoside is the only quercetin glycoside which had a clear mQTL on LG8. As in the peel, several octane-di-ol glycosides also had mQTL on LG8 (Table S4). LG16 contained mQTLs for procyanidins at the same genetic region as in the peel. In this genetic area on LG16, mQTL for phenolic esters, (+)-catechin, and (-)-epicatechin were also found (Figure 5B). Chlorogenic acid, which had an mQTL in the peel on LG17, had a minor mQTL in the same genetic region in the flesh (Tables S4). Glucuronic acid which is not part of the phenylpropanoid pathway also has an mQTL on LG16 (Table S4).

The levels of metabolites in the LG16 mQTL hotspot were controlled by a single, dominant locus present in both parents

The different metabolites were found to have a clear 3:1 segregation and therefore could be mapped. Surprisingly these metabolites mapped on one locus on LG16. Tables S3 and S4 show the effects of different parental allele pairs at the LG16 hotspot on metabolite levels in the progeny, allowing the detection of dominant, recessive or additive genetic effects. For the hotspot of mQTL on LG16, both 'Prima' and 'Fiesta' had one dominant and one recessive allele each (*Mm*). The combination of the two dominant alleles in the progeny (*MM*) occasionally showed a further increase in the metabolites level, indicating an additive effect or incomplete dominance in these cases.

As LG16 showed a cluster of many mQTL, we analyzed this particular hotspot in more detail. The metabolites (+)-catechin, (-)-epicatechin, two of the procyanidin dimers, and two procyanidin trimers gave one major mQTL per parent, showing two contrasting groups representing monogenic segregation. Figure 4 shows a typical example based on procyanidin dimer II, indicating a Mendelian 3:1 segregation of *Mm* x *Mm* ($\chi^2_{3:1} = 0.87$; $P > 0.05$), whereby the amount of this metabolite was apparently predominantly controlled by a single locus. Both parents were heterozygous and the *Mm* and *MM* offspring genotypes showed a similar average content for this metabolite, indicating full dominance both in peel and flesh (Figure 4). The effect of a single dominant allele was on average an increase of 0.62 compared to the recessive allele on the $^{10}\log$ scale (Figure 4). This resembles a 4.2-fold increase on a linear scale.

Procyanidins, phenolic esters, (+)-catechin, (-)-epicatechin, and kaempferol hexose rhamnose showed similar segregation patterns, apparently being controlled by the same dominant and recessive alleles of LG16 from both parents. Coumaroyl hexoside and quinic acid appeared to be controlled by the same locus, but in contrast to phenolic esters and other phenolic compounds, their level was negatively correlated to the other phenolic compounds that mapped at this hotspot (Tables S5, S6).

Graphical genotyping of the mQTL hotspot on LG16

The metabolites that segregated according to a 3:1 ratio and had only one mQTL behaved as monogenic traits, and could be mapped as genetic markers, which was true for (+)-catechin, (-)-

epicatechin, two procyanidin dimers, and two procyanidin trimers. In case where the relative metabolite level of a progeny was high, it was not clear whether that genotype had inherited the dominant allele from the mother, the father, or from both. Only in case of a low metabolite level, it was evident that both the mother and the father provided the recessive allele. Therefore, as for all dominant markers that segregate in a 3:1 fashion, the marker information could only be used for 25% of the progeny. In spite of this limitation, it was still worthwhile and helpful to locate the genetic window of the locus (Figure 6).

For more-detailed mapping of this locus on LG16, additional SSR markers in the LG16 mQTL hotspot were mapped. Several of the new SSR loci mapped in the genetic window (Figure 6). The names and allelic sizes of these markers are given in Table S7. Graphical genotyping of the metabolite ‘procyanidin dimer II’ revealed that the gene causing the mQTL hotspot on LG16 is located between the locus NH26a and the locus Ch05e04 in ‘Prima’, and between the locus Ch02a03 and locus Hi15a13 in ‘Fiesta’ (Figure 6).

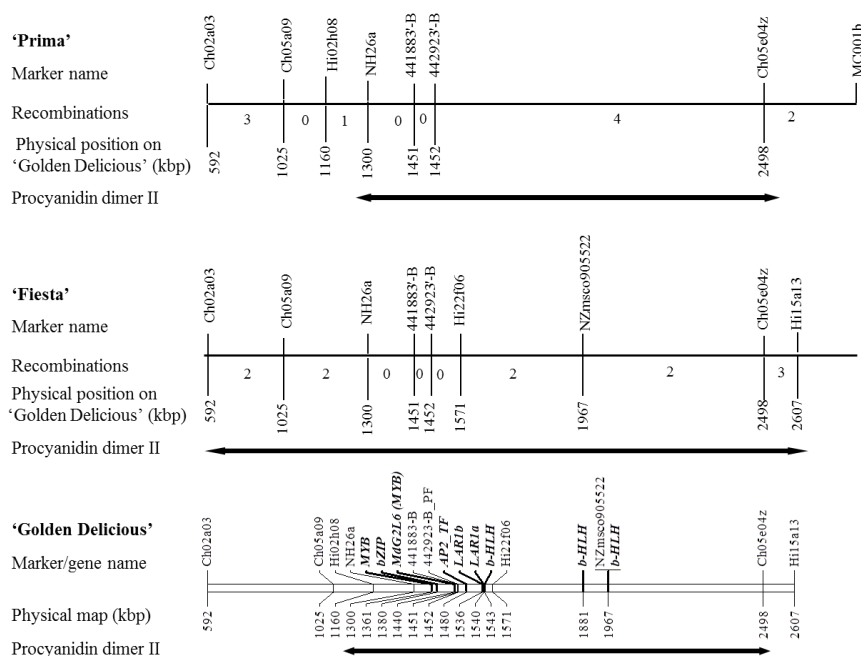


Figure 6: Genetic linkage maps of ‘Prima’ and ‘Fiesta’ and a physical map of the apple cv. ‘Golden Delicious’ for the mQTL hotspot region on LG16. Procyanidin dimer II was used as representative for the metabolites that mapped to the LG16 mQTL hotspot. The mQTL regions were genetically mapped as monogenic traits by means of graphical genotyping in both parents and are indicated as horizontal green arrows. The structural gene *leucoanthocyanidin reductase (LAR)* of the phenylpropanoid pathway appeared to be present in this region, according to the putative genes in the whole genome sequence of cv. ‘Golden Delicious’. Seven putative transcription factor genes including *MYB* and *bHLH* were also detected in this region. The structural gene *LAR* and Transcription factor genes are presented in bold text.

The structural gene *LAR* and seven transcription factors are at the mQTL hotspot

The results of the alignments of structural genes of *A. thaliana* against apple are shown in Table 2 and also in Fig. 5 and 6. Using the apple genome sequence (<http://genomics.research.iasma.it/gb2/gbrowse/apple/>), only the structural gene *LAR* was found in the mQTL hotspot on LG16 among the fifteen different structural genes of the phenylpropanoid pathway (Fig. 5, 6). A closer look revealed that the published 'Golden Delicious' genome sequence had at this locus at least five *LAR*-like sequences in overlapping contigs. However, a sequence homology search using the EMBOSS software package revealed that these sequences were identical or highly homologous, and actually only two different genomics sequences were found. These two sequences probably represent the two alleles for the *LAR* gene. Apart from *LAR*, seven putative transcription factor genes were also identified in the genetic window of the mQTL hotspot on LG16 (Table 3). Two of these belong to *MYB* class and three to basic *helix-loop-helix* (*bHLH*) class, one to *bZIP* and one to *AP2* class of transcription factor genes.

Table 2: Structural genes of the phenylpropanoid pathway in *Arabidopsis* and apple

Gene	Full name	Arabidops is locus	Known apple sequence	Homologues in <i>Arabidopsis</i>	Genetic positions in apple	At mQTL hotspot on LG16
4CL	4-coumarate-CoA ligase	At1g65060	GO565912, GR882782, GO577298, GO568847, etc	12	3 (LG1,3,7)	No
ANR	Anthocyanin reductase	At1g61720	AY830130	1	3 (LG5)	No
ANS	Anthocyanin synthase	At4g22870	AF117269	5	4 (LG6)	No
C3H	p-coumarate 3-hydroxylase Trans-cinnamate	At2g40890	TC28151 (http://compbio.dfci.harvard.edu/)	1	6 (LG8,15)	No
C4H	monooxygenase	At2g30490	GO549874	1	2 (LG3,11)	No
CHI	Chalcone isomerase	At3g55120	X68978	3	9 (LG1,11,14)	No
CHS	Chalcone synthase	At4g34850	X68977	4	3 (LG2,5,7)	No
DFR	Dihydroflavonol 4-reductase	At5g42800	AF117268	2	2 (LG8,12)	No
F3'H	Flavonoid 3'-monooxygenase	At5g07990	Apple_0223.261.C2.Contig645, Apple_0223.261.C1.Contig644 (http://titan.biotech.uiuc.edu/cgi- bin/ESTWebsite/estima_start?seqSet=apple)	1	2 (LG6,14)	No
F3H	Flavanone 3-hydroxylase	At3g51240	AF117270	2	2 (LG2,5)	No
FLS	Flavonol synthase	At5g08640	AF119095	2	2 (LG0)	No
HCT	Shikimate hydroxycinnamoyltransferase	O- At5g48930	Apple_0223.2950.C2.Contig4990, Apple_0223.850.C1.Contig1757	1	3 (LG9,17)	No
LAR	Leucoanthocyanidin reductase	*AJ550154	AY830131, AY830132 Apple_0223.263.C1.Contig648, Apple_0223.215.C2.Contig537,	0	5 (LG13,16)	Yes
PAL	Phenylalanine ammonium lyase UDP-dependent	At2g37040	Apple_0223.263.C2.Contig649	4	6 (LG1,4,8,12)	No
UGT	glycosyltransferase	At5g17050	AF117267	3	6 (LG0,1,7,9)	No

* Not found in *Arabidopsis* but in *Desmodium uncinatum*.

Table 3: Transcription factor genes at the mQTL hotspot on LG16

Putative gene	Apple sequence	Position on chromosome (bp)	Size (kbp)
<i>MYB</i>	MDP0000375685	1361220-1362093	1.3
<i>MYB</i>	MDP0000703817	1440436-1442198	1.7
<i>bHLH</i>	MDP0000319726	1543934-1555640	11.7
<i>bHLH</i>	MDP0000154272	1881558-1884164	2.6
<i>bHLH</i>	MDP0000261293	1967365-1970040	2.6
<i>AP2</i>	MDP0000939633	1475660-1476865	1.2
<i>bZIP</i>	MDP0000250967	1376596-1386527	9.9

Discussion

Locating genes that are responsible for variation in metabolite levels in the progeny

Many genes, involved in the biosynthesis of phenylpropanoids and flavonoids in different plant species have been identified and appear to occur throughout the plant kingdom (Dixon and Steele 1999; Winkel-Shirley 2001). In apple, most of the structural genes and several regulatory genes have been isolated (Kim, Lee et al. 2003; Takos, Ubi et al. 2006; Szankowski, Li et al. 2009). However, it is not yet clear which of these genes are critical for the variation in metabolite levels between tissues and genotypes. Genetic mapping of metabolites is a useful step to elucidate which genes are critical for this variation.

By combining the metabolomic data with genetic linkage maps, we detected 488 mQTL in peel and 254 mQTL in flesh, using the software MetaNetwork. To gain more insight into the biochemical pathways regulated by the detected mQTL, the centrotypes that showed highly probable mQTL were manually annotated using accurate mass, in-source MS/MS and UV/Vis-absorbance information. The annotation revealed several groups of metabolites in both peel and flesh, including phenylpropanoid esters and glycosides (such as coumaroyl-hexoside, chlorogenic acid and coumaroyl-quinic acid) and flavonoids (such as (+)-catechin, (-)-epicatechin, procyanidins, quercetin, and kaempferol glycosides) (Tables S1, S2). Except for the alcohol glycosides and glucuronic acid all of these metabolites originate from the phenylpropanoid and flavonoid biosynthetic pathway. This bias towards the phenylpropanoid pathway was at least partially caused by the sample preparation (aqueous-methanol extracts) and the LC-MS system used (C18-reversed phase LC and ESI ionisation in negative mode (Moco, Bino et al. 2006)). Other apple fruit metabolites, such as apolar steroids and volatiles could not be detected with these parameter settings and thus did not show up in our approach.

For many metabolites no mQTL was detected

For about half the number of the metabolites no mQTL was detected. A reason may be complex genetic regulation of metabolites by several loci. This would hamper detection of these mQTL. Another reason can be strong environmental effects compared to genetic effects for some metabolites. A third reason can be that levels of metabolites can be just above noise level for some progenies and within the noise for other progenies. That would hamper detection of mQTL too. A

metabolite is regarded as being present, in case 10 or more progeny genotypes show that metabolite above the noise level.

A flavonoid-mQTL hotspot is located on LG16

From MetaNetwork analysis, an mQTL hotspot on LG16 was found for both the parents and both the peel and flesh tissues (Figs. 1, 2). In the genetic window of the mQTL hotspot we detected 271 putative genes, using the published genome of apple (Velasco, Zharkikh et al. 2010). To find the putative underlying gene(s) causing this mQTL hotspot, the structural genes of the phenylpropanoid pathway were positioned on the chromosomes of apple, as shown in Fig. 5 and Table 2. In this case, the position of the structural genes was not determined by means of genetic mapping in a segregating population, but via the alignment of the known sequences of these structural genes in *A. thaliana* with the first draft whole genome sequence of the apple cv. 'Golden Delicious' (Velasco et al., 2010). Only the structural gene *LAR* was found in the mQTL hotspot on LG16. *LAR* catalyzes the conversion of leucocyanidins into the flavan-3-ols (+)-catechin and (-)-epicatechin, which are the building blocks of procyanidins (Fig. 5). Both the flavan-3-ols and the procyanidins showed an mQTL in this region. The *LAR* gene may explain the mQTL hotspot on LG16, as 23 procyanidins in peel and 13 procyanidins in flesh were mapped to this locus, as well as the two flavan-3-ols (+)-catechin and (-)-epicatechin in both tissues. This was also observed with the alleles involved: the level of flavan-3-ols was increased by the 'a' allele from 'Prima' and/or the 'd' allele from 'Fiesta' (Fig. 5). These same alleles also increased the level of procyanidins (Fig. 5, Tables S3, S4).

A contra-indication for *LAR* being the responsible gene for the hotspot is the presence of mQTL at this locus for the chlorogenic acid and coumaroyl quinic acid (phenolic esters), phloridzin, and kaempferol glycosides. These metabolites are upstream of the substrate for *LAR* (Fig. 5). Furthermore, their levels were simultaneously increased with flavan-3-ols and procyanidins (Fig. 5). A possible explanation for this is the presence of a transcription factor gene that regulates the structural genes for these metabolites. In view of this, we searched for transcription factor genes at the mQTL hotspot and detected seven transcription factor genes here. Some of these transcription factor genes belong to the *MYB* and *bHLH* types of transcriptional regulators. One of these transcription factor genes may be responsible for the mQTL hotspot for the phenolic esters and kaempferol glycosides and possibly also for the mQTL hotspot for flavan-3-ols, procyanidins and coumaroyl hexoside.

Coumaroyl hexoside was negatively correlated with procyanidins, indicating a key role for 4CL

The level of coumaroyl hexoside (Fig. 5A) was negatively correlated with the level of flavan-3-ols and procyanidins (Table S5, S6) in the progeny, as indicated by a red color in Fig. 5A. This was also indicated by the alleles that increase the levels of coumaroyl hexoside: whereas the levels of flavan-3-ols and procyanidins were elevated by presence of the marker alleles 'a' and 'd', the coumaroyl hexoside was elevated when these alleles were absent (Fig. 5A, Tables S3, S4). This may be explained by a strong sink effect for the production of flavan-3-ols and procyanidins, thus competing with coumaroyl hexoside for the substrate coumaric acid. This indicates that the enzyme 4CL is the critical factor. If this enzyme has a low activity, coumaroyl acid may accumulate, giving a stronger flow to the side branch that leads to coumaroyl hexoside. However, if 4CL would be more active, the downstream metabolites would be enhanced, at the cost of coumaric acid and coumaroyl hexoside.

Strikingly, the gene *4CL* is not at the mQTL hotspot but on other chromosomes (Table 2). This may indicate a feedback mechanism from downstream genes or downstream metabolites to *4CL*.

Fig 5a shows also that quinic acid was negatively correlated with flavan-3-ols, procyanidins and fenolic esters. This may be a result of a sink effect again. High activity of *4CL* may lead to high levels of coumaroyl CoA, and therewith to high levels of phenolic esters, but also to depletion of the other precursor of phenolic esters, i.e. quinic acid.

Structural genes of the phenylpropanoid pathway in other studies

Several structural genes for flavonoid biosynthesis have been described in apple. Takos, Ubi et al. (2006) described two *LAR* genes and an *Anthocyanidin reductase (ANR)* gene, detected in cDNA from the peel of the red apple cv. 'Cripps Red'. They named the two *LAR* genes *MdLAR1* and *MdLAR2*. BLAST results of the sequences of these genes from apple revealed that only *MdLAR1* is present in the LG16 mQTL hotspot. The *MdLAR2* is located on LG13. A major part of LG16 contains homoeologous sequences of LG13 due to whole genome duplication (Maliepaard, Alston et al. 1998; Velasco, Zharkikh et al. 2010). Park, Sugimoto et al. (2006) also identified *LAR* and *ANR* in apple fruits by statistically analyzing the expressed sequence tags (ESTs). *ANR* utilizes anthocyanidin and *LAR* use leucocyanidin as substrate. Both *ANR* and *LAR* participate in synthesis of flavan-3-ol monomers, whereas these monomers are the building blocks of procyanidin polymers (Xie, Sharma et al. 2003). In grape, *ANR* and *LAR* genes strongly influence procyanidin accumulation and composition during berry development (Bogs, Downey et al. 2005).

In pear (*Pyrus communis* L.), a closely related species to apple, (Fischer, Gosch et al. 2007) isolated cDNAs for the prominent genes in flavonoids biosynthesis mentioned in Fig. 5 via homology with the apple sequences. They found high homology to apple in the DNA and cDNA. Substrate specificities of the recombinant enzymes expressed in yeast were determined for physiological and non-physiological substrates and found to be in general agreement with the characteristic pear flavonoid metabolite pattern (Fischer, Gosch et al. 2007). In strawberry, another member of the rosaceous family, genes in the flavonoid pathway could be clearly classified into two groups according to their expression pattern; one having two transcription peaks at early and late stages (i.e., *FaANR*, *FaANS*, *FaCHI*, *FaFHT* and *FaLAR*), and the other showing an up-regulation trend with a single peak at the turning and/or ripening stage (i.e., *FaDFR*, *FaFGT*, *FaFLS* and *FaMYB* (Almeida, D'Amico et al. 2007). This shows that expression pattern for flavonoid genes can be different and fruit stage for the expression of certain flavonoid genes can be very critical.

Transcription factor genes of the phenylpropanoid pathway in other studies

Bogs, Jaffé et al. (2007) characterized a grapevine *MYB* transcription factor gene, *VvMYBPA1*. This regulatory gene was shown to be able to activate the *LAR* and *ANR* genes, and several other flavonoid pathway genes in grapevines (Bogs, Jaffé et al. 2007). Two other *MYB* genes, *VvMYBPA2* (Terrier, Torregrosa et al. 2009) and *VvMYB5b* (Deluc, Bogs et al. 2008) were also found to promote procyanidin biosynthesis in grapes. We compared *VvMYBPA1*, *VvMYBPA2* and *VvMYB5b* with the *MYB* gene in the hotspot on LG16, but found no convincing homology.

Three transcription factor genes from *Arabidopsis* that regulate procyanidin accumulation have been identified, i.e. the *MYB* transcription factor *TRANSPARENT TESTA 2* (Nesi, Jond et al. 2001), the *bHLH* transcription factor *TRANSPARENT TESTA 8* (Nesi, Debeaujon et al. 2000), and *TRANSPARENT TESTA*

GLABRA (*AtTTG1*; (Nesi, Debeaujon et al. 2000). For the latter gene, Brueggemann, Weisshaar et al. (2010) found a functional homologue in apple. Li, Flachowsky et al. (2007) reported the upregulation of mRNA for several structural enzymes of the flavonoid pathway in apple by overexpressing the maize leaf color (*Lc*) transcription factor gene. A BLAST analysis was performed to identify putative homologues of these transcription factor genes on the apple genome, particularly in the mQTL hotspot on LG16. None of the above transcription factor genes showed close homologues at this locus, and thus these cannot explain the mQTL hotspot on LG16.

Phenotypic buffering

Fu, Keurentjes et al. (2009) analyzed a segregating *A. thaliana* population for variation in transcript, protein and metabolite abundance. They mapped QTL for 40,580 molecular and 139 phenotypic traits, and found six QTL hotspots with major, system-wide effects. For the far majority of the 500,000 SNPs between the two parental lines, no or minor impact on the phenotype was detected. The authors interpreted this lack of dramatic changes by genetic variation as robustness of the system. The six hotspots are exceptions. These hotspots seem to correspond to a few molecular fragilities of an otherwise robust regulatory system (Fu *et al.*, 2009). In another study, Keurentjes, Fu et al. (2006) have described these hotspots in more detail for metabolites. Their results show striking similarities with our results in apple, although the *Arabidopsis* population contained only up to two alleles per gene because of the homozygous parents, whereas in the apple population up to four alleles were present per gene. Although a series of genes are involved in the pathway of phenolic compounds, a large extent of quantitative variation in these compounds is explained by one locus only, i.e. the hotspot on LG16.

Quercetin glycosides are not controlled by the flavonoid-mQTL hotspot on LG16

Quercetin glycosides are commonly found in apple fruits (Van der Sluis, Dekker et al. 2001). Although quercetin glycosides are part of the phenylpropanoid and flavonoid biosynthetic pathways, they did not exhibit any significant mQTL on LG16 (Fig. 5). Instead, these compounds had mQTL on LG1 in peel. On LG1, the structural gene UGT is located, which is responsible for the glycosidation of quercetins. This UGT gene may be responsible for the mQTL on LG1 (Fig. 5). Takos, Ubi et al. (2006) identified and characterised an UGT gene in apple, using a functional genomics approach. As UGT genes consist of a large gene family, further studies would be needed to verify which UGT gene would be responsible for the glycosidation of quercetin. In flesh, quercetin glycosides showed an mQTL on LG8 (Fig. 5).

Another quercetin derivative, i.e. isorhamnetin (a methoxylated form of quercetin) had a strong mQTL on LG13. Possibly, a gene for methoxylation is located on LG13 (Fig. 5). We observed that both quercetin metabolites (quercetin glycosides and isorhamnetin glycosides) were not depending on LG16, and we did not detect any free, unmodified quercetin in apple. Together, these observations suggest that the rate-limiting step in the formation of quercetin derivatives in apple is determined by the modifying enzymes (UGT, OMT), and that the flux of phenylpropanoid towards quercetin is adapted to the availability of modification opportunities.

Consequences of tight genetic linkage of the dominant alleles for high metabolites to the recessive alleles for pH

Figs. 3 and 4 illustrate that in both parents the dominant alleles for high levels of metabolites are genetically tightly linked to the recessive alleles for high pH. This has consequences for apple breeding. In Northern Europe, apples with a low pH are usually preferred to high pH, and therefore should contain the dominant allele for low pH. As the dominant allele for low pH is in repulsion phase to the dominant allele for high metabolites, at least in the genotypes we investigated, the selection for the dominant low pH allele implies the selection for the recessive allele for low levels of the phenolic compounds. Therefore, progeny that are more acidic, have higher chances of having lower levels of procyanidins and other phenolic compounds. This can be solved by the selection of progeny that have one dominant allele for low pH from one parent and one dominant allele for high metabolite levels from the other parent. This is the *MmMama* group in Fig. 4. In the southern countries of Europe and in Asia, consumers usually prefer a higher pH. In that case, the desired absence of the dominant allele for pH is automatically combined with the presence of the dominant allele for high levels of metabolites. This is the *MMmama* group in Figure 4. For apple, it takes usually six to eight years after sowing to have fruits that can be evaluated for pH and metabolites. Selection of the desired progeny for these fruit traits is feasible already at a very young stage, using DNA from leaves of seedlings and DNA markers (Figure 3).

Follow-up studies

In subsequent studies, the expression profiles of the *MdLAR1* candidate gene, other structural genes of the phenylpropanoid pathway, and the seven transcription factor genes found in the mQTL hotspot will be studied in progeny that have either low or high procyanidins levels. Final proof of their involvement needs complementation studies. Next, to increase the level of these beneficial metabolite(s), the most promising alleles may be inserted into existing, highly popular apple cultivars with low procyanidin levels, e.g. by means of a cisgenesis approach. Cisgenesis is defined as ‘the genetic modification of a recipient plant with natural gene(s) from a sexually compatible plant (Schouten, Krens et al. 2006; Schouten, Krens et al. 2006). Whatever the outcome of these follow-up studies, the knowledge obtained from the current study of the mQTL hotspot genes is already of use for the breeding of new cultivars with increased levels of these putatively beneficial metabolites through application in marker-assisted breeding.

Acknowledgements

We are thankful to Remmelt Groenwold for his assistance in the orchard.

This project was financially supported by PiDON, Inova Fruit B.V and the Higher Education Commission (HEC) of Pakistan. RdV acknowledges Centre for Biosystems Genomics and The Netherlands Metabolomics Centre, both initiatives under the auspices of the Netherlands Genomics Initiative, for additional financing.

Chapter 3: MQ²: Visualizing multi-trait mapped QTL results.

Pierre-Yves Chibon^{1,2,3}, Roeland E. Voorrips^{1,3}, Richard G.F. Visser^{1,2,3} and Richard Finkers^{1,3}

¹ *Wageningen UR Plant Breeding, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands*

² *Graduate School Experimental Plant Sciences, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands*

³ *Centre for BioSystems Genomics, Wageningen, 6708 PB, Wageningen, The Netherlands*

Adapted from the publication in Molecular Breeding, 2013.

DOI: 10.1007/s11032-013-9911-3

Abstract

Quantitative Trait Loci (QTL) mapping tools such as MapQTL and R/qtl allow easy and fast analysis of more than one trait at the same time. However, for experiments with large datasets, such as high-throughput metabolite QTL (mQTL) analysis, these tools do not provide an easy-to-inspect summary of the results. The ability to have an overview of the distribution of the identified QTL becomes a key factor. MQ2 fills this need by providing a command line tool and a web application that summarizes and visualizes the results of multi-trait QTL analysis. MQ2 can use the output of commonly used QTL analysis tools, such as MapQTL and R/qtl, as input. MQ2 can be used for free at: <http://www.plantbreeding.wur.nl/mq2/>

Keywords: multi-trait analysis, bioinformatics, high-throughput QTL analysis.

Results

Analysis of large datasets, such as high-throughput metabolomics experiments (Khan, Chibon et al. 2012), transcriptomics experiments (Hammond, Mayes et al. 2011) and genetical genomics experiments (Jansen and Nap 2001) involve quantitative trait loci (QTL) analysis of a large number of traits. When such experiments are analyzed with MapQTL (Van Ooijen 2009), it becomes cumbersome to obtain an overview of QTL positions for all these traits, to discern if there are QTL hotspots (i.e. loci that affect many different traits, such as transcription factors) and if so, where there are located in the genome. In MapQTL v6.0, one would have to go through all the traits one by one and extract this information manually. MQ² allows summarizing and visualizing the distribution of QTL from MapQTL (Van Ooijen 2009) analysis with a large number of traits. Using R/qtl (Broman, Wu et al. 2003) it is possible to obtain this overview more easily but some script writing is required.

MQ² is a tool for combining and visualizing QTL analyses of multiple traits, from the output of mapping software such as MapQTL and R/qtl. After reading the data files produced by the QTL mapping software, MQ² returns a visual representation of the distribution of QTL on the genetic map, several comma separated text (CSV) files containing detailed information for each QTL and a MapChart (Voorrips 2002) compatible file to visualize the 2-LOD interval for each QTL on the genetic linkage map. Markers flanking the 2-LOD interval are extracted and exported as well. These csv and MapChart files allow the user to visualize the QTL information using their own preferred program.

The workflow of MQ² is divided into three steps. The first step is to retrieve all significant QTL by parsing the MapQTL (version 5 or 6) project directory or the provided CSV file or Excel document. The user has to provide the LOD threshold. MapQTL (Van Ooijen 2009) and R/qtl (Broman, Wu et al. 2003) can calculate this LOD threshold using a permutation test. For each trait, MQ² extracts only the strongest QTL per linkage group. The second step is to identify, for all the QTL extracted in the first step, the marker closest to the peak. The third step is to calculate the number of QTL for each marker on the map.

MQ² is available as a web-based tool and as a command line application. Both versions produce the csv and MapChart files mentioned above; the web application additionally produces a visualization of QTL hotspots. Both components can run on any computing platform where a Python interpreter is available (version 2.6 and above, <http://www.python.org>); this includes all versions of Microsoft Windows, GNU/Linux and Mac OS X. Both components are licensed under the GPLv3 or any later

version; the source code can be found on github at <http://github.com/PBR/MQ2> and http://github.com/PBR/MQ2_Web.

The command line application has a built-in plug-in system allowing supporting multiple input formats. The plug-in system is based on the `straight.plugin` (<https://pypi.python.org/pypi/straight.plugin/>) library. MQ² supports MapQTL and CSV inputs by default using the standard python libraries. To support the reading of excel-compatible files, installation of the `python-xlrd` library (<https://pypi.python.org/pypi/xlrd/>) is required.

The web application is implemented with the Flask framework (<http://flask.pocoo.org/>), providing a web-interface to the Python library. From the front page, the user can upload the MapQTL project directory, CSV file or Excel document compressed in a zip file. This will create a session which uniquely identifies this data set. The MQ² session is valid for seven days after its last analysis, after which the data is removed from the system and the user will have to re-upload it. The user can access previous MQ² sessions by re-entering the session identifier.

Within a MQ² session, the user can specify which MapQTL session or which sheet of the Excel document to use and which LOD threshold to consider as significant for a QTL. If the data have already been extracted for a given LOD threshold, the application will inform the user that this analysis has already been done. Otherwise, MQ² will produce a result page with a visualization of the distribution of the QTL on the map, an overview of the parameters used and links to download the result files generated. The user can find the list of QTL associated with the selected marker by clicking on the visualization. The visualization of the distribution of QTL on the map is performed using the `Flot` javascript library (<http://www.flotcharts.org>) which supports Internet Explorer 7+, Chrome, Firefox 2+, Safari 3+ and Opera 9.5+.

New users can explore MQ² via the example data and the example session provided on the front page.

Usage

Running MQ² on MapQTL output

For each project, MapQTL generates a “.mqp” file containing information about the project. Within the same folder, MapQTL generates a folder that contains all QTL mapping results. The results folder has the same name as the project and ends with the extension: “.mpd”. To prepare the QTL mapping data for MQ², locate the “.mpd” folder associated with the MapQTL project and create a zip archive of this folder. Only the files with the extension “.mqo” are required, but the presence of other files does not affect the results.

Run MQ² on output from other QTL analysis software

To run MQ² on output from other software, for example `R/qtl` (Broman, Wu et al. 2003), the data needs to be provided either in a CSV file (using commas “,” as delimiters) or in an Excel document. The format of the input is very important for MQ² to work and should be as follow: the first column contains the markers, the second column contains the linkage group and the third column contains the position of these markers on the linkage group. Each subsequent column contains the LOD value

for each trait. The first row of the document contains the headers (Markers, Linkage Group, position, Trait name1, Trait name2, etc; Figure 1).

	A	B	C	D		E	F	G
1		chr	pos	pheno1	pheno2	Trait 3	Trait 4	
2	D1M430	1	0	0.6617472931	0.8342067823	0.8834137855	0.0825279093	
3	c1.loc2.5	1	2.5	0.7269621239	0.8504986784	0.9065070363	0.1279394177	
4	c1.loc5	1	5	0.7790906581	0.859657896	0.9115121893	0.178422737	
5	c1.loc7.5	1	7.5	0.8190823064	0.8628030898	0.9014637989	0.2313521931	
6	D1M318	1	9.8	0.8466105122	0.8613983719	0.8821573276	0.2803080129	
7	c1.loc10	1	10	0.8444246561	0.8564908068	0.8828229651	0.277641572	
8	c1.loc12.5	1	12.5	0.8124159915	0.7835078808	0.8750017498	0.2440978501	
9	D1M212	1	13.11	0.8034223114	0.7624578401	0.8690215932	0.2359481028	
10	c1.loc15	1	15	0.7251542941	0.6865992366	0.8189291853	0.203756772	
11	c1.loc17.5	1	17.5	0.6126848308	0.5792159868	0.7430835657	0.1598913733	
12	c1.loc20	1	20	0.4932165979	0.4668625949	0.657288466	0.1165394042	
13	c1.loc22.5	1	22.5	0.3722689578	0.354412031	0.5638913636	0.0764208998	
14	c1.loc25	1	25	0.2573679502	0.2481773121	0.4666323232	0.0425310026	
15	c1.loc27.5	1	27.5	0.1570299716	0.1551245542	0.3702857753	0.0175724016	
16	c1.loc30	1	30	0.0788310461	0.0814817524	0.2799152349	0.0033303283	
17	c1.loc32.5	1	32.5	0.0273121669	0.0311989858	0.199946517	0.0002533718	
18	c1.loc35	1	35	0.0029071875	0.0050429948	0.1333982027	0.0074296781	
19	D1M437	1	37.11	0.0011884124	0.0001815142	0.0886619641	0.0200728301	
20	c1.loc37.5	1	37.5	0.0012947144	0.000561299	0.085346705	0.0211487634	
21	c1.loc40	1	40	0.002080597	0.0080894819	0.0648385644	0.0286251425	
22	c1.loc42.5	1	42.5	0.0030256602	0.0243719726	0.0463018429	0.0368911452	
23	c1.loc45	1	45	0.0040882172	0.0485601416	0.0305365977	0.0455907583	
24	c1.loc47.5	1	47.5	0.0052174641	0.0790110201	0.0180528314	0.0543396188	
25	c1.loc50	1	50	0.006361623	0.1136544689	0.0090192889	0.0627832925	

Figure 1: Screenshot of a spreadsheet showing how the data should be formatted in order to be submitted to MQ². The first column contains the marker names, the second column contains the linkage groups, the third column contains the map positions on the linkage group (this column must be numeric). The following columns contain the LOD value calculated by the QTL mapping software for this trait at this position. The first row of the document contains the header and will be used to extract the trait names.

The CSV or Excel document should be compressed into a zip archive before uploading via the web interface. Note that MQ² can only analyze one CSV file or one Excel document at a time; however, the Excel document may contain multiple sheets.

Running MQ² via the web interface

To run MQ² via the web interface (Figure 2), upload the zip archive at the front page of the MQ² website. It will bring you to the MQ² session page. From here, you can run new analyses by selecting either a MapQTL session or the name of the Excel sheet to analyze and providing the LOD threshold. For each analysis, links are provided to a results page (Figure 3.1) giving a visual overview of the distribution of QTL. The list of QTL associated with a given marker can be found by clicking on the visualization of the distribution present on the results page (Figure 3.2).

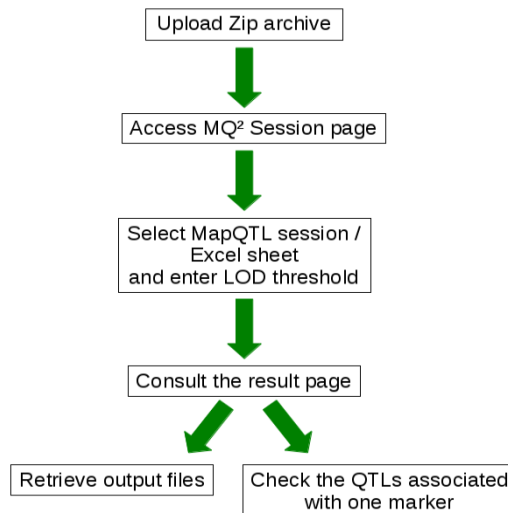


Figure 2: Workflow followed by the user when running the web application of MQ². The first step is to upload the ZIP archive which creates the MQ² session. The second step is to select a MapQTL session or an Excel sheet (depending on the input) and enter a LOD threshold and run the analysis. Finally the results can be visualized on the result page, where the output CSV file can also be downloaded and for each marker with QTL, the list of QTL associated can be retrieved by clicking on the visualization.

Running MQ² via the command line

Alternatively, the command line version can be run directly using either the zip archive, or the uncompressed MapQTL directory, CSV file or Excel document as input. A MapQTL session or Excel sheet name and the LOD threshold are also required. The command line utility can be called using the syntax: `MQ2 --file path/to/input.zip --lod <YourLodThreshold> --session <MapQTLSessionIdentifier>` or for excel documents: `MQ2 --file path/to/input.xls --lod <YourLodThreshold> --session <NameOfTheExcelSheet>`.

For example :

```
MQ2 --zipfile path/to/Demoset.zip -lod 3 --session 2
```

```
MQ2 --file path/to/Demoset.xls -lod 3 --session="Sheet1"
```

The output files are generated in the current working directory. All available options are shown using the command: `MQ2 --help`.

3.1

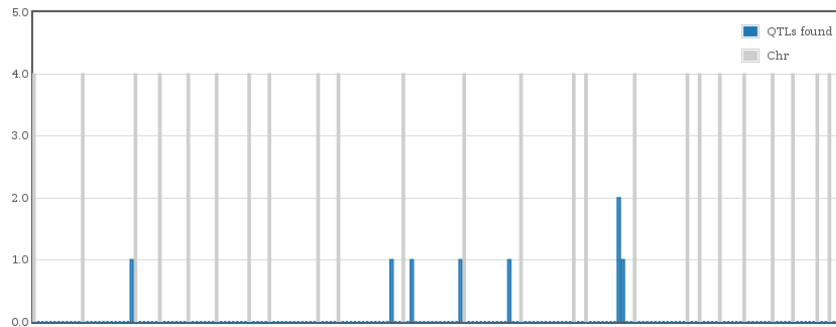
MQ² results

[Home](#) | [Return to session page](#)

Results of the analysis ran on 2013-05-16 at 11:40:17

Pre-visualisation

Number of QTLs for each marker on the map



The parameters of the experiment were:

LOD Threshold : 3.0
 Plugin used : MapQTL plugin
 Session : 2
 Experiment ID : 20130516114017_s2_t3
 Number of markers : 176
 Number of traits : 10

Output files

- [qtls.csv](#) -- List of all the QTLs from the QTL analysis output.
- [qtls_with_mk.csv](#) -- List of all the QTLs from the QTL analysis output with their closest marker on the map.
- [map_with_qtls.csv](#) -- Representation of the map in CSV with the number of QTLs found for each marker.
- [qtls_matrix.csv](#) -- Matrix giving for each marker trait combination the LOD values found by the QTL analysis.

3.2

MQ² results for marker E36M48-330

[Home](#) | [Return to session page](#) | [Return to result page](#)

Following is the list of QTLs found associated to this marker:

Trait_name	Group	Position	Locus	LOD	Variance	% Expl.	Additive	Dominance	Closest marker
A_trait07	P10b	40.55	E36M48-330	15.44	0.50	30.40	-0.63	0.14	E36M48-330
A_trait11	P10b	40.55	E36M48-330	27.45	0.35	47.50	-0.74	0.30	E36M48-330

2 QTLs found

Figure 3: Screenshot of the result page of the web application. The visualization displays for each marker in the genetic map the number of QTL associated (Fig 3.1). Each blue bar represents a markers and each light grey bar represents the limit of the designated linkage group. The QTL associated with a marker (Fig 3.2) can be accessed by clicking on one of the bars of the marker. Below the visualization are presented the parameters used for this analysis. At the bottom of the page are links to download each output file generated by MQ².

Discussion

MQ² provides a simple overview of the distribution of the QTL over the genome from a high-throughput QTL mapping analysis. Its main drawback is the assumption of one QTL per linkage group while extracting the QTL from the results. This assumption is the result of two factors. The first factor is the difficulty to determine the number of QTL in a linkage group from the LOD profile over that linkage group. Using the LOD information, it is difficult to distinguish what is a single QTL with a shoulder from two QTL profiles overlapping. This is also the reason why simple interval mapping analysis cannot determine if there are more than one QTL per linkage group and why more complex approaches such as MQM (Haley and Knott 1992) and restricted MQM have to be applied. It might be possible to model using the information from QTL mapping tool but that would be doing the work that QTL mapping tools already are doing and are acknowledged for. The second factor which influenced the decision of only considering one QTL per linkage group is that MQ² is targeted for high-throughput QTL analysis where hundreds of traits are considered at once. MQ² shows the global trend of the results, where most QTL are located and by assuming one QTL per linkage group it can do so reliably and fast (around a minute for more than 600 traits; data not shown). Once the global trend of the data is known, the biologists can go back to their QTL mapping tool to perform more in-depth analysis of these specific traits or regions (such as performed by Khan, Chibon et al. (2012)).

From a technical and licensing point of view, MQ² is written fully in python using modules from the standard libraries and two external libraries which are compatible with python 3. Its source code is compatible with python 2.6 to python 3.3. The web interface uses the flask web framework which is being ported to python 3. Thus MQ² will remain available and working for the coming years. In addition MQ² is open-sourced and licensed under the GPL version 3 (or any later version) allowing anyone to improve it and redistribute it, its plugin system providing a way to extend it easily to support new QTL mapping tools.

Conclusions

MQ² provides an interface to QTL mapping output allowing the visualization of multi-trait QTL mapping results in a fast and simple way, both for MapQTL projects and for other QTL mapping software provided their output is formatted as specified in a CSV file or an Excel document.

The separation between the web application and the command line interface allows the user to run the tool using the web instance (<http://www.plantbreeding.wur.nl/mq2/>) or to run the tool locally from the command line. MQ² can provide in a few seconds information on the distribution of the QTL along the map and potential QTL hotspots for a few hundred traits previously analyzed with a QTL mapping tool, making MQ² well suited for the modern high-throughput QTL mapping of ~omics data.

Acknowledgments

Funding: Wageningen UR Plant Breeding and the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

Chapter 4: Marker2sequence, mine your QTL regions for candidate genes

Pierre-Yves Chibon^{1,2,4}, Heiko Schoof³, Richard G.F. Visser^{1,4}, Richard Finkers^{1,4}

¹*Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ Wageningen, The Netherlands*

²*Graduate School Experimental Plant Science, P.O. Box 16, 6700 AA, Wageningen, The Netherlands*

³*Crop Bioinformatics, INRES, University of Bonn, 53115 Bonn, Germany*

⁴*Centre for BioSystems Genomics, P.O. Box 98, 6700 AA, Wageningen, The Netherlands*

Abstract

Summary: Marker2sequence aims at mining quantitative trait loci (QTL) for candidate genes. For each gene, within the QTL region, marker2sequence uses data integration technology to integrate putative gene function with associated Gene Ontology terms, proteins, pathways and literature. As a typical QTL region easily contains several hundreds of genes, this gene list can then be further filtered using a keyword based query on the aggregated annotations. Marker2sequence will help breeders to identify potential candidate genes for their traits of interest.

Availability: Marker2sequence is freely accessible at <http://www.plantbreeding.wur.nl/BreeDB/marker2seq/>. The source code can be obtained at <https://github.com/PBR/Marker2Sequence>

Supplementary information: supplementary figures are available at Bioinformatics online.

Introduction

Quantitative Trait Loci (QTL) are regions on the genome statistically associated with a phenotype. Plant Breeders aim to introgress these regions from a donor parent to improve a cultivar (Zamir 2001). However, a typical QTL region may contain over hundreds of genes, including genes negatively influencing the breeding goals. Complete genome sequences of many crop plants are becoming available, including the genome of important food crops such as tomato (Tomato genome consortium 2012) and potato (PGSC, Xu et al. 2011). The availability of structural and functional genome annotations makes it possible to investigate the QTL region for genes positively or negatively influencing the trait of interest. A tool that offers this functionality is GBrowse (Stein, Mungall et al. 2002). However, exploring several hundreds of putative candidate genes with GBrowse will still be a lot of work, since limited information is available for each gene.

Identification of putative candidate genes can be improved using data integration approaches. Biologically relevant knowledge about, for example, Gene Ontology (GO) (Ashburner, Ball et al. 2000), protein functions or metabolic pathways, can be combined with expert knowledge about the trait under investigation. The basic principle of Semantic Web technology is to integrate different types of data from different sources using standardized ontologies (Berners-Lee, Hendler et al. 2001). Important resources such as UniProt (Nicole Redaschi and UniProt Consortium 2009) and GO have become available in a Resource Description Framework (RDF, <http://www.w3.org/TR/rdf-primer/>) format allowing data integration against these resources.

Our research aims to develop a tool, called Marker2sequence (M2S), which plant breeders can use to identify the putative candidate gene for their QTL. We describe how M2S uses semantic data integration approaches to obtain available information for each gene model and combine this with a keyword based search function to mine for the appropriate candidate gene. M2S, by design, can work with any genome annotation, but we show the functionality of M2S the tomato genome annotation.

Design and implementation

Marker2sequence (M2S) is a web-based tool, using Java EE 6 and the Struts framework (v1.3.10). It runs on a Glassfish (v3.1) application server. M2S rely on the availability of a genome annotation and reference genetic linkage map in RDF format. A utility, gff2RDF, has been developed to perform the conversion of the tomato, potato and Arabidopsis genome annotation and linkage maps to the RDF format. It is available at <https://github.com/PBR/gff2RDF>. This utility extracts for all genes their location, their human readable description, their associated GO term identifier, and their associated UniProt protein identifiers (Jain, Bairoch et al. 2009) from the annotation. For Tomato, the EXPEN 2000 map was used as reference linkage map (Fulton, Van der Hoeven et al. 2002). The Jena library (Carroll, Dickinson et al. 2004) is used to build the RDF model and write it to disk. These graphs were loaded into a Virtuoso Open-source Edition (version 6.1.3) (Erling and Mikhailov 2007) triple store together with the Gene Ontology (version 2011_11_03) and UniProt (version 2011_10). Any triple store with a SPARQL end-point can be used with M2S.

M2S can handle three types of inputs (Supplementary Figure 1). The first two inputs are: two markers flanking the QTL region, or a list of markers spanning the QTL region. These markers should have a physical position on the genome sequence or a position on the reference linkage map. The third input requires a genomic region using the format: Chr:start..stop. Either input leads to a summary page divided into three sections; The top section show the alignment of the reference genetic map with the genomic information, which help to identify problems in the genetic map or the assembly of the genome. The lower section consists of three tabs. The first tab contains a list of all genes in the specified region, with their location and human readable description. The second tab lists all the markers, within the region. The third tab shows the genetic map for the specified region. Each list can be exported into a spreadsheet-compatible format or a pdf. The gene list can be searched using a keyword via the box in the middle section. This search is performed using SPARQL (<http://www.w3.org/TR/rdf-sparql-query/>) on all available resources and returns any gene with a matching keyword in any of the queried databases.

The details for each gene (Supplementary Figure 2) include information retrieved from the genome, the GO terms, the proteins (UniProt), pathways (UniPathway) and literature associated with these proteins. The GO terms are obtained from the genome annotation and, for tomato, from AFAWE (Jöcker, Hoffmann et al. 2008). This data integration aids the end-user to determine if this gene is a good candidate for the trait of interest.

Example

β -carotene content is a trait influencing the color of tomatoes (Lincoln and Porter 1950). Based on our QTL analysis, using data from the *Solanum lycopersicum* x *Solanum galapagense* LA0483 RIL population (Paran, Goldman et al. 1995), this compound has QTL on chromosome 3 (between TG130 and TG74) and 6 (between TG253 and TG314). Marker2sequence identified 2003 genes on chromosome 3 and 988 genes chromosome 6. A query with the keyword: “beta-carotene”, returns the gene Solyc03g007960.1.1 on chromosome 3 and the gene Solyc06g074240.1.1 on chromosome 6. Solyc03g007960.1.1 has the description “Carotene beta-hydroxylase”, is associated with the GO term for “carotene beta-ring hydroxylase activity”, the protein “Beta-carotene hydroxylase” and the pathway for “Carotenoid biosynthesis”. Solyc06g074240.1.1 is associated with the GO term for “carotenoid biosynthetic process”, the pathway for “Carotenoid biosynthesis” and more specifically

the part of “beta-carotene biosynthesis”. Information for each gene can be quickly mined using Marker2sequence and both genes are candidates for our trait of interest.

Conclusions

Marker2sequence provide plant breeders with a way to obtain all annotated gene models in a QTL region, to query, over multiple databases, within the QTL region of interest and an extensive summary for each gene model. Marker2sequence will help breeders to identify potential candidate genes for their traits of interest.

Acknowledgments

The authors would like to acknowledge the International Tomato Genome Consortium (<http://solgenomics.net/tomato/>) for their work on the tomato genome sequence and genome annotation.

Funding: This project was co-financed by the 6th framework EU project "High Quality *Solanaceous* crops for consumers, processors and producers by exploration of Biodiversity", contract number: FOOD-CT-2006-016214, Wageningen UR Plant Breeding and the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research

Chapter 5: Identification of transcription factor binding sites in tomato.

Pierre-Yves Chibon^{1,2,3}, Jos Hageman⁴, Yury Tikunov^{1,3}, Arnaud Bovy^{1,3}, Richard G.F. Visser^{1,2,3} and Richard Finkers^{1,3}

¹ *Wageningen UR Plant Breeding, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands*

² *Graduate School Experimental Plant Sciences, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands*

³ *Centre for BioSystems Genomics, Wageningen, 6708 PB, Wageningen, The Netherlands*

⁴ *Biometris-Applied Statistics, P.O. Box 100, 6700 AC Wageningen, The Netherlands*

Abstract

Transcription factors are proteins regulating gene expression by binding to the promoter regions of their target genes. In this study, 17 putative transcription factor binding sites have been identified by combining the unique genomic structure of selected introgression lines (IL) with gene expression measurements of these genotypes. Six genotypes (progeny) of an IL population and the mother (parent) of the cross have had their gene expression measured on an Affymetrix micro-array. By comparing the gene expression of the individual genotypes to the parent we were able to extract a list of genes differentially expressed in the six offspring plants. These differentially expressed genes are either located within the introgression segments (cis genes) or outside of the introgression segments (trans genes). The only genomic difference between the progeny plants and the parent are the introgression segments, leading to the question of regulation of the expression of the trans genes. The promoter regions of these differentially expressed trans genes have been analyzed for DNA motifs revealing 17 potential transcription factor binding sites. Twelve of these motifs resemble known transcription factor binding sites.

Introduction

In any given cell, at any given time, thousands of genes ensure the cell's function. To perform this task, genes must be expressed at a certain time and in a certain amount. This regulation is ensured by the presence of a gene regulatory network involving genes and transcription factors (Macneil and Walhout 2011). Transcription factors are proteins involved in regulating expression of other genes by binding to short DNA motifs, called transcription factor binding sites, in the promoter region of their target genes (Chen and Rajewsky 2007). A single transcription factor may influence the expression of several genes in the genome thus providing a coordinated mechanism to control these genes (Lee and Young 2000). Identifying transcription factors as well as their binding sites is a first step in the understanding of the gene regulatory network of an organism.

Different techniques allows predicting transcription factors in the genome, including sequence based (Robertson, Bilenky et al. 2006), and yeast one-hybrid assays (Barrasa, Vaglio et al. 2007). The results of these experiments are reported in literature and stored in databases such as EDGEDb (Barrasa, Vaglio et al. 2007), DBD (Wilson, Charoensawan et al. 2008), PlantTFDB (Zhang, Jin et al. 2011) and AnimalTFDB (Zhang, Chen et al. 2012). These resources only contain transcription factor information; they present the sequences, alignments and domains of these transcription factors. However, they do not provide any information regarding the genes regulated by these transcription factors nor the motifs to which they bind. Resources such as TRANSFAC (Wingender 2008) or JASPAR (Sandelin, Alkema et al. 2004) list transcription factors binding sites. They provide information about these short motifs that transcription factors recognize and bind to. Both TRANSFAC and JASPAR are not species specific resources. JASPAR only contains very few transcription factor binding sites from plant species (five TFBS of *Arabidopsis thaliana*, six TFBS of *Zea mays*, one TFBS of *Hordeum vulgare*, three TFBS of *Pisum sativum*, one TFBS of *Petunia x hybrida*, three TFBS of *Antirrhinum majus* and one motif of *Triticum aestivum* and *Nicotiana sp.*).

To predict transcription factor binding sites, multiple approaches have been described: using phylogeny, one may search for conserved domain in the promoter region of orthologous genes (Gumucio, Shelton et al. 1996; Hardison, Oeltjen et al. 1997). Using gene expression information, one

may search for common motifs in the promoter sequences of the co-expressed genes, either relying on known common motifs or using statistical models. Several studies (Ho Sui, Mortimer et al. 2005; Hestand, van Galen et al. 2008; Essaghir, Toffalini et al. 2010) mention the difficulties to identify transcription factor binding sites in sets of co-expressed genes. To circumvent these difficulties, tools such as oPOSSUM (Ho Sui, Mortimer et al. 2005) or CORE_TF (Hestand, van Galen et al. 2008) rely on a list of pre-computed predictive regulatory elements and compare this list with the list of gene sequences submitted. These approaches are reported to return a lot of false positive and false negative results (Hestand, van Galen et al. 2008). One reason for these false results is the level of noise present in the gene expression data of higher organisms (Essaghir, Toffalini et al. 2010). Using gene expression information, one may find sets of co-expressed genes that seem to be co-regulated. Without prior information, transcription factor binding sites can be predicted using statistical models. These methods rely on the computation of a background distribution of the nucleotides on the genome. This background allows them to predict if a short nucleotide sequence is appearing more often in the sets of co-regulated genes than it does at the genome level.

If oPOSSUM and CORE_TF rely on known TFBS to find motifs in a set of sequences, Multiple Em for Motif Elicitation (MEME) and RSAT are two examples of tools for *de novo* transcription factor binding sites prediction. MEME is one of the most widely used tool to identify motifs in related DNA or protein sequences (Bailey, Williams et al. 2006). Using statistical modeling, MEME is capable to detect recurring motifs in a set of DNA sequences. An extensive explanation of the MEME algorithm has been published by Bailey and Elkan (1994). Several tools have been built around MEME, for example, TOMTOM (Gupta, Stamatoyannopoulos et al. 2007) and GOMO (Buske, Boden et al. 2010). TOMTOM is a motif comparison tool allowing to search for matching motifs in known transcription factor binding site databases. It can be used to assess the novelty of the motif found. GOMO associates GO terms to provided DNA motifs. From the DNA motif provided, it scores the promoter region of each gene of the selected organism, ranks them by affinity with the motif and, using the GO terms associated with these genes, determines the GO terms associated with the motif. It can be used to estimate the biological processes in which the motif might be involved. The Regulatory Sequence Analysis Tools (RSAT) oligo-analysis (van Helden, Andre et al. 1998) is another *de novo* DNA motif prediction tool, originally designed for yeast sequences, it requires a fixed motif length and is known to return lots of false positive for genomes having mixtures of GC-rich and GC-poor promoter sequences.

In introgression lines (IL), each individual has a genotype similar to the mother of the original cross except for some introgression parts that are from the father of the original cross. Therefore, in a similar environment, phenotypic differences between the individuals of the IL population and the mother are assumed to be due to the introgression segment of the father. The phenotypic differences can occur because of genes present in the introgression segment (*cis*) or by (*cis*) genes regulating expression of genes elsewhere on the genome (*trans*). For this study, we aim to identify these *trans* genes in the ILs. The expression of *trans* genes being different in a similar genomic background and similar environment, might be the result of the action of a regulatory element such as microRNA or transcription factors. If some of the *trans* genes are influenced by a transcription factor, looking for motifs in their promoter regions might reveal known or new transcription factor binding sites.

In marker assisted selection, breeders rely on molecular markers to have a representation of the genome and when crossing two organisms, these markers can be used to assess the success of introgression. Breeders should choose the markers as close as possible to the target gene, if possible on the gene itself (Hospital 2001). Targeting the genes responsible for a phenotype, termed candidate genes, implies knowing these genes but manually finding them among the hundreds or thousands of genes present in a QTL interval is a cumbersome work. Marker2sequence (Chibon, Schoof et al. 2012) is aiming at facilitating this process. For a given region of the genome it retrieves all the genes present, integrates the information known for each of them in different resources and provides a way to search these genes and their annotation in order to filter out the genes of interest, potential candidate genes. However, Marker2sequence does not have access to regulatory information. If a transcription factor is present in the given QTL interval, Marker2sequence is not able to include in the list of potential candidate genes for the observed trait the genes regulated by this transcription factor. Marker2sequence is thus restricted to the detection of cis genes. One way, to find genes regulated by a transcription factor, is to identify transcription factor binding site (TFBS) in promoter regions. There have been studies that searched for TFBS in promoter regions of co-expressed genes (Vilo, Brazma et al. 1999; Long, Liu et al. 2004; Essaghir, Toffalini et al. 2010) but to our knowledge none used introgression lines with their particular genetic structure to enhance their predictions.

When studying gene expression in introgression lines, the genes expressed can be trans genes (outside of an introgression) or cis genes (within an introgression). By comparing the gene expression between individuals from the introgression lines and the individual of the same genetic background and without introgressions (recurrent parent of the IL lines), trans genes differentially expressed can be identified. These genes thus behave differently on a similar genetic background, maybe due to the presence of transcription factors in the introgressions. The current study is on tomato and relies on six progeny lines, which does not allow searching for conserved motif across orthologous genes. The transcription factor binding sites resources available do not contain information regarding transcription factor binding sites in tomato. Searching for TFBS using known TFBS from available resources is thus possible but would not allow finding any tomato specific TFBS. The promoter sequences of tomato have a highly variable GC content, making RSAT not suitable to find TFBS. This study therefore searched for *de novo* TFBS using MEME in promoter sequences of trans genes differentially expressed. The outcome of this study could then be integrated within Marker2sequence allowing it to consider genes outside of the QTL interval regulated by genes present within the QTL interval when searching for candidate genes.

Materials and Methods

Plant Materials

This study was performed using *Solanum lycopersicum* cultivar Moneyberg (SL) and 6 lines (provided by KeyGene, the Netherlands, Figure 1) carrying a single introgression of the *Solanum chmielewskii* LA1840 (SC) in the background of Moneyberg (Prudent, Causse et al. 2009; Do, Prudent et al. 2010). From this population, six lines were selected for their expression profile in flavonoids and volatile compounds (Bovy et al. unpublished). Figure 1 gives a schematic overview of the six introgression fragments which cover different parts of the genome.

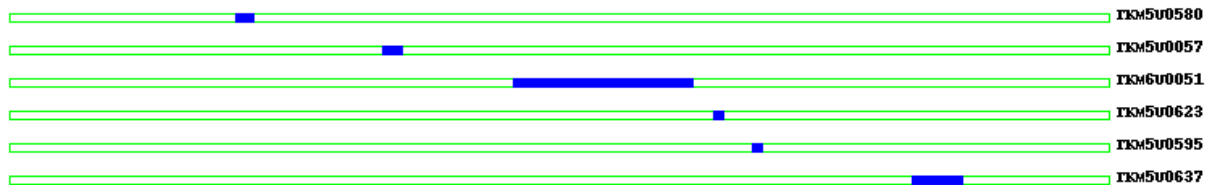


Figure 1: Graphical representation of the introgression of *S. chmielewskii* (in blue) in the Moneyberg background (in green). These six genotypes are the genotypes that have been selected to be analyzed with the EuTOM3 micro-array.

Gene expression analysis

Gene expressions have been measured for six genotypes, in four replicates, of the *S. chmielewskii* population: TKM5U0580, TKM5U0057, TKM5U0595, TKM5U623, TKM5U637 and TKM6U0051. The measurement was done with a custom Affymetrix array: EUtom3, described and used by Karlova, van Haarst et al. (2013). The cultivar Moneyberg was also included in the analysis providing a reference of gene expression without any introgression from *Solanum chmielewskii*.

Analysis of the gene expression differences

The expression of each introgression line has been compared to the expression of SL using a Dunnett test (Dunnett 1955) which allows comparing a set of groups to a reference group. In this experiment, the groups were the six different individuals from the *S. chmielewskii* IL population and the reference group was the expression measurement from SL. This analysis returned a list of genes differentially expressed (p-value lower than 0.01) in the IL lines compared to Moneyberg, thus genes whose expression was solely influenced by the presence of the introgression segment from SC in the genome.

Analysis of the promoter sequences

Each differentially expressed gene identified using the Dunnett test was mapped onto the tomato genome (Tomato genome consortium 2012) SL2.40 using blast (Altschul, Gish et al. 1990). The genes mapping outside of the introgressions (trans genes) were then selected. For each gene, the promoter region was extracted in FASTA format using a custom made script. The promoter region was defined as the region from 600 base-pair upstream to 100 base-pair downstream from the gene start (Veerla and Hoglund 2006). Special care was taken to extract the sequence in the right orientation.

Because promoter regions may contain repetitive elements which may introduce a bias compared to the background modeled, these regions should be masked before the DNA motif search is performed (Bailey, Williams et al. 2006). The sequences were masked using repeat masker (AFA, R et al. 1996-2010). The “cross_match” search engine has been used as it is said to be the most sensitive engine,

the sensitivity has been left to default and the DNA source was set to *Arabidopsis thaliana*. All other options were left to their default settings as well.

Benchmark of the DNA motif tool

In order to evaluate the sensitivity of the results returned by MEME, three datasets of 100 promoters containing 10, 20 and 30 times a given motif were analyzed by MEME. Each dataset had 100 promoters randomly selected from all the tomato promoters of the genome sequence. The motif inserted is: CCTATAAGG. This motif was allowed up to two random mutations every time it was inserted in a promoter.

Identification of DNA motifs

DNA motifs were searched using the MEME tool (Bailey, Williams et al. 2006), version 4.9.0. MEME relies on a background model to assess the distribution of the nucleotides in the DNA; from this it is able to find DNA motifs which are significantly different from the background. For this study, to generate the background model, all the promoter sequences from the tomato genome sequence have been extracted with the only filter that these sequences should not contain; any unknown nucleotide (marked as “N”), here as well the gene orientation has been taken into account. The background model was then computed using the “fasta-get-markov” program provided with MEME with an order of Markov model of 4.

The MEME analysis themselves have been run with the following settings, the motifs should have a length between 5 and 15 nucleotides, MEME should return the first 10 motifs found, the distribution model used was the default Zero Or One Per Sequence (ZOOPS). The maxsize parameter has been set to 1000000 to reserve enough memory to process the datasets.

To assess its performance, two controls have been analyzed. In the first control, 200 promoter sequences randomly picked from the tomato genome have been run through MEME. In the second control, 200 promoter sequences were randomly picked from the tomato genome and a defined motif was inserted at a random position in the sequence. The motif inserted was ten nucleotides long and was allowed to have up to two nucleotides varying (position and variation being random).

Finally, MEME was run on the promoter sequences of the genes found to be differentially expressed by the Dunnett test and cleaned by RepeatMasker.

Identification of related transcription factor binding sites

TOMTOM has been used on the significant motifs found by MEME to determine the likelihood of the motif to be a known transcription factor binding site. TOMTOM was run using the default settings, thus searched the motifs in the JASPAR (Sandelin, Alkema et al. 2004) and UniPROBE (Newburger and Bulyk 2009) databases.

The number of significant hits returned by TOMTOM for each motif found is presented in the results section.

Associating GO terms to motif

GOMO has been used to try to assess if the motif found is actually related to transcription or translation. GOMO was run using the “Multiple species” category which gave access to the

“*Arabidopsis thaliana* (plant)” database. The significant threshold used was the default one, q-value lower or equal to 0.05.

The main GO terms found to be associated with the significant motifs found by MEME are presented in the results section.

Analysis of the genes having a potential transcription factor binding site

Using Annotex (**Chapter 5**), the annotation of the genes found to share a potential transcription factor binding site has been retrieved and integrated to represent the most common GO terms, Pathway and Protein associated with these genes. The analysis has been performed using a custom-made python program that retrieved the information for each gene from Annotex using the JSON export function and integrate it.

Results

Results of the gene expression analysis

Six genotypes from the *S. chmielewskii* introgression line population and the mother of the original cross were analyzed for gene expression using the EUTOM3 Affymetrix micro-array (Figure 1). Using a Dunnett test, the different replicates were compared to the replicates of cv Moneyberg. The total number of differentially expressed genes was between 139 and 540, of which 81 to 482 were trans, depending on the introgression line (Table 1).

Table 1: Table presenting for each genotypes, the markers flanking the main introgression, the number of genes found to be differentially expressed by the Dunnett tests (p-value < 0.01) and for these genes how many are located within the introgression and how many are located outside the introgression. The last row indicates how many known transcription factors are differentially expressed in total and, between brackets, within the introgression (cis).

	TKM5U0580	TKM5U0057	TKM5U0595	TKM5U623	TKM5U637	TKM6U0051
Introgression (markers)	Seq-rs6990, seq-rs7635	Seq-rs9022, seq-rs7338	Seq-rs7994, Seq-rs5154	Seq-rs4549, seq-rs4547	Seq-rs7089, seq-rs5387	Seq-rs2904, seq-rs5505
Genes differentially expressed	139	191	220	249	278	504
In an introgression: Cis Genes	58	73	124	16	100	22
Outside an introgression: Trans Genes	81	118	96	233	178	482
Transcription factor differentially expressed (cis TF)	2 (2)	10 (4)	12 (5)	15 (2)	8 (4)	0 (0)

The position of the genes found to be differentially expressed was compared to the position of the introgression segment on the genome. The number of cis and trans genes was counted (Table 1). Trans genes are genes outside the introgressions from *S. chmielewskii*, thus on a Moneyberg background but behave differentially compared to Moneyberg. Focusing on potential transcription

factors, we searched for transcription factor binding sites in the promoter regions of the trans genes differentially expressed.

Results of the MEME benchmark

In order to assess the sensitivity of MEME we created three datasets each of 100 random promoter extracted from the genome sequence and in which a motif were inserted in 10, 20 and 30 promoter sequences (Table 2). The motif was allowed up to two random point mutations for each insertion.

Table 2: Table describing the results of the small benchmark of MEME performed on 100 random promoter sequences from the tomato genome in which a motif was inserted in 10, 20 or 30 sequences. The motif was allowed to have up to two point mutations for each insertion. The second column gives the number of significant motifs found in this dataset; the third column provides information on whether or not the inserted motif was found and if so where in the results.

	Number of motifs significant	Motif inserted
10 motifs inserted	3	Not found
20 motifs inserted	0	First result but not significant
30 motifs inserted	2	The Most significant motif

The motif could not be found in the dataset containing 10 motifs in 100 sequences, but could be found in the datasets containing 20 and 30 motifs.

Results of the MEME analysis

For each genotype, the promoter sequence of the trans genes differentially expressed has been extracted from the tomato genome sequence and sequence repeats were masked with RepeatMasker. MEME analysis identified in total 17 DNA motifs overrepresented in promoter sequences of the trans genes (Table 3).

The results of the MEME analyses are presented in Table 3 and Supplementary Table 1. A more detailed description of these results will be given here, focusing on the genotypes TMK5U0580 and TKM5U0057 as an example. These two genotypes have been chosen as they have relatively small introgressions and a single motif was identified in the promoter region of their trans genes, and these motifs score differently in TOMTOM and GOMO.

The genotype TMK5U0580 has 58 cis genes and 81 trans genes found differentially expressed by the Dunnett test. The genotype TKM5U0057 has 73 cis genes and 118 trans genes found differentially expressed by the Dunnett test. So TMK5U0580 has 81 genes potentially regulated by a regulatory element present in its introgression and TKM5U0057 has 118 genes potentially regulated by a regulatory element present in its introgression. These trans genes putatively contain novel transcription factors binding sites.

MEME identified one significant motif in the promoter region of the 81 trans genes of line TKM5U0580. This motif is present in 36 of the 81 genes evaluated. TOMTOM did not find any similar transcription factor binding sites (with E-value < 1). GOMO associated, to this motif, five different GO terms, including "Transcription factor activity".

MEME found one significant motif in the promoter region of the 118 trans genes of TKM5U0057. This motif is present in 18 of the 118 genes considered. TOMTOM found 9 similar transcription factor binding sites (with E-value < 1). GOMO associated, to this motif, three different GO terms including "Transcription factor activity".

TKM5U0580 has 36 genes with a common potential transcription factor binding site (Table 3). These 36 genes are annotated with a total of 14 different GO terms, two different pathways and 58 different proteins. TKM5U0057 has 18 genes with a common potential transcription factor binding site. These 18 genes are annotated with a total of five different GO terms and 23 different proteins. The annotation of the genes containing in their promoter the motifs found is presented in Table 4 and Supplementary Table 2.

Using RT-PCR, the expression of known transcription factors was measured (data not shown). The total number of cis and trans transcription factors differentially expressed was counted (Table 1). The introgression segment of TMK5U0580 contains two differentially expressed cis transcription factors, which may regulate the 81 trans genes found differentially expressed. TKM5U0057 has four differentially expressed cis transcription factors which may regulate the 118 trans genes differentially expressed.

Table 3: Table summarizing the results of the motifs found by MEME analysis for each genotype. These motifs have been found in the trans genes found differentially expressed and thus eventually regulated by a shared transcription factor. The motifs were then annotated with TOMTOM and GOMO providing some information whether the motifs are similar to known motifs and whether they are related to transcription or translation processes. The line presenting the number of significant hits on TOMTOM includes the number of significant hits with the thresholds of E-value < 10 (default), E-value < 5 and E-value < 1 (where the E-value is the expected number of false positives in the matches).

	TKM5U0580	TKM5U0057	TKM5U0595	TKM5U623	TKM5U637	TKM6U0051
Significant motifs	1	1	1	4	6	4
E-value of these motifs	1.3 e-9	3.5 e-13	1.5 e-9	2.3e-45 1.3e-10 4.4e-4 2.4e-8	8.5e-16 1.1e-6 8.9e-4 5.1e-8 1.1e-4 1.2e-3	1.2e-79 9.2e-30 1.4e-9 7.3e-7
Number of promoter sequences with the motifs	36	18	28	104 18 21 44	72 10 22 26 16 15	133 95 29 22
Number of significant hits on TOMTOM with the threshold: E-value <10 E-value < 5 E-value < 1	4 / 1 / 0	33 / 27 / 9	23 / 16 / 8	21 / 14 / 9 29 / 21 / 8 36 / 23 / 3 5 / 2 / 1	16 / 11 / 4 11 / 6 / 3 16 / 9 / 5 8 / 5 / 3 30 / 25 / 14 4 / 1 / 1	26 / 22 / 16 7 / 2 / 0 19 / 11 / 1 22 / 15 / 6
Transcription or Translation related GO found in GOMO	✓	✓	✓	✓ x ✓ x	✓ x x ✓ x ✓	✓ x x x

Table 4: The annotation of the genes having a common potential transcription factor binding site for the genotypes TKM5U0580 and TKM5U0057. It presents the list of GO terms, pathways and proteins associated with these genes and their frequency. The 36 genes of TKM5U0580 are thus associated with 58 different proteins and two of these 36 genes are associated with "Uncharacterized mitochondrial protein AtMg01410". (The information for all motifs in all genotypes can be found in Supplementary Table 2)

TKM5U0580 (36 genes)		TKM5U0057 (18 genes)	
GO terms			
Ribosome	1/14	RNA processing	2/5
oxidation-reduction process	1/14	regulation of transcription, DNA-dependent	1/5
zinc ion binding	1/14	RNA binding	1/5
ATP binding	1/14	DNA binding	1/5
receptor activity	1/14		
cell redox homeostasis	1/14		
intrinsic to membrane	1/14		
DNA binding	1/14		
enzyme regulator activity	1/14		
heme binding	1/14		
Membrane	1/14		
proton-transporting ATP synthase complex, catalytic core F(1)	1/14		
Mitochondrion	1/14		
copper ion binding	1/14		
Pathways			
Pigment biosynthesis; anthocyanin biosynthesis	1/2		
Protein modification; protein ubiquitination	1/2		
Proteins			
Uncharacterized mitochondrial protein AtMg01410	2/58	Putative CCA tRNA nucleotidyltransferase 1	2/23
Uncharacterized mitochondrial protein AtMg01110	2/58	Putative CCA tRNA nucleotidyltransferase 2	2/23
Uncharacterized 8.8 kDa protein in rps12-tRNA-Val intergenic region	1/58	CCA-adding enzyme	2/23
ATP synthase gamma chain 1, chloroplastic	1/58	Probable poly(A) polymerase	2/23
Flavonoid 3',5'-hydroxylase	1/58	CCA tRNA nucleotidyltransferase, mitochondrial	2/23
NADH-ubiquinone oxidoreductase chain 2	1/58	F-box protein At3g07870	1/23
Disease resistance protein At4g27190	1/58	Ethylene-responsive transcription factor ERF008	1/23
ATP synthase gamma chain 2, chloroplastic	1/58	Ethylene-responsive transcription factor ERF018	1/23
TMV resistance protein N	1/58	Ethylene-responsive transcription factor RAP2-9	1/23
RING-H2 finger protein ATL51	1/58	CCA tRNA nucleotidyltransferase 1, mitochondrial	1/23

Discussion

The aim of this experiment was to analyze the promoter regions of trans genes differentially expressed for potential transcription factor binding sites. In a “classic” transcriptomics experiment, gene expression is measured in different genotypes or conditions and genes differentially expressed are then extracted and analyzed with regards to the conditions or phenotypes (Figure 2 a). In this experiment the workflow differed after the genes differentially expressed have been retrieved. The position of these genes has then been assessed on the tomato genome sequence and each gene differentially expressed has been classified either as cis genes (located within an introgression region) or as trans genes (located outside of an introgression region). The trans genes behave differentially while in a similar genomic background and might thus be under the influence of a regulatory element located within the introgressions. They have thus been further analyzed for DNA motifs, potential transcription factor binding sites, in their promoter region (Figure 2 b).

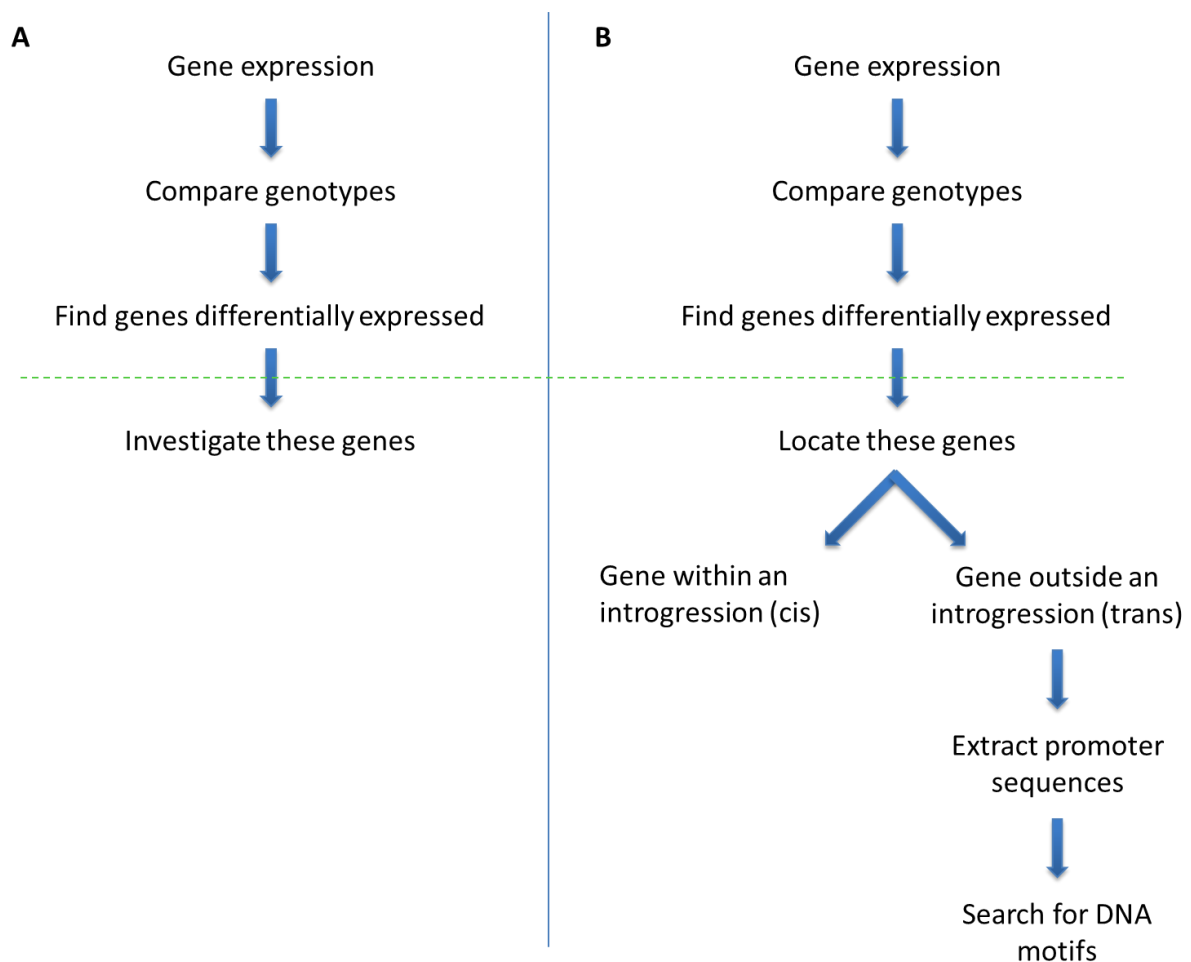


Figure 2: Schematic representation of the workflow followed in (a) a “classic” transcriptomics experiment used to find out which genes are differentially expressed between different genotypes or condition and (b) the current experiment where the differentially expressed genes have been mapped onto the genome sequence, determined whether they are cis or trans genes and the trans genes considered for a DNA motif search.

The presence or absence of the DNA motifs splits the promoters into different groups. A functional analysis of these groups would provide insight in the presence of regulatory elements within the introgression segment that regulates the expression of genes outside the introgression and whether these regulatory elements influence a complete pathway or biological process. TMK5U0580 and

TKM5U0057 have both had a single potential transcription factor binding site found. The current level of annotation of the tomato genome does not allow drawing any firm conclusions from the integrated data. However, from the list of GO terms associated to these genes in both genotypes, there is a clear link to DNA binding and translation regulation, but overall, the annotation is too incomplete at this stage to speculate which processes these genes belong to and thus which processes are regulated by regulatory elements present in the introgressions.

From our benchmark, we concluded that MEME considers the nucleotides composing the motif as well as the number of occurrences of a motif to assess its significance. From our benchmark it appears that a motif should be present at least in 20% of the promoters to be found, but depending on the set of sequences analyzed it might require a higher frequency to be found significant. This implies that MEME might miss motifs which occur in a low frequency. Grouping promoters before performing the motif search might circumvent this, for example, by creating sub-sets of similar promoters it might increase the frequency of motifs involved in a small process which are otherwise lost in the larger sets. The promoters could be split using the genome annotation (the GO terms or pathways associated with the genes), however, the genome annotation might not be sufficient to perform this clustering as we have seen it is not yet sufficient to provide information regarding the biological processes a gene set, sharing a DNA motif, is involved in. Another approach for the clustering would be to build the cluster based on gene expression but this would require more gene expression analysis and according to Yeung, Medvedovic et al. (2004) this is strongly dependent on the number of micro-arrays used for the clustering. In our case we only had four replicates per genotype which is not sufficient to perform a good clustering based on the gene expression; Yeung, Medvedovic et al. (2004) mentions the optimal is around 50 to 100 (on yeast). In their study, van Helden, Andre et al. (1998) split the promoters into different clusters sequence based. This last approach might help creating smaller sets of promoters to run MEME on, but the length of the sequence considered might also just render the clustering inefficient.

Tompa, Li et al. (2005) assessed the prediction of MEME and 12 other DNA motif prediction tools. From this benchmark one can conclude that the sensitivity of MEME on the negative control varies between 0.98 and 0.99, meaning that MEME tends to find one to two motifs in sequence that do not contain any known motif. This implies that there are likely false positive motifs in our results. One of the conclusions of the study of Tompa, Li et al. (2005), is that these 13 transcription factor binding site prediction tools seem to be complementary to each other in their approaches and they recommended using MEME with Weeder (Pavesi, Mereghetti et al. 2004). However, Weeder seems not to be maintained anymore, the last release available appears to be from 2009 while MEME is still maintained; the version 4.9.0 has been released in October 2012 and updated in January 2013. It would be interesting to re-perform their benchmark and see how current tools (e.g.: MEME, RSAT) perform, which due to time constraints we have not been able to do for this chapter.

From the results of the TOMTOM analysis on each of the 17 putative identified transcription factor binding sites, it can be estimated that 12 of the motifs were already identified in other species, three motifs are found similar to only one known transcription factor binding site, and thus eventually less common. The remaining two motifs are the motifs found in TKM5U0580 and the second motif found in TKM6U0051. The motif found in TKM5U0580 has an E-value of $1.3e-9$ and is associated by GOMO with the GO term "Transcription factor Activity", however it is the only motif found for this genotype. The second motif found in TKM6U0051 has an E-value of $9.2e-30$ and is not associated with any

transcription related GO terms, on the other hand, it is one of four motifs found. According to the sensitivity of MEME found on negative control data sets (Tompa, Li et al. 2005), these motifs might be false positives or tomato specific motifs and thus would be valid candidates to investigate further.

Only one motif was found in TKM5U0057 and TKM5U0595 however, the results from TOMTOM associate both motifs with nine and eight known transcription factor. These two motifs combined with the first motifs found in TKM5U623 and TKM6U0051 would be good motifs to confirm in the lab in order to assess the quality of the results produced by MEME as these four motifs have low E-value (especially the last two), are found by TOMTOM in a high number of known transcription factor binding sites and are all associated via GOMO to the GO term "Transcription factor activity". These four motifs should be confirmed in the lab as they combine all the criteria to be involved in transcription regulation.

Some motifs, such as the second motif found for TKM5U0637, are similar to known transcription factor binding sites according to TOMTOM (three in this case), however, GOMO does not associate any transcription factor related GO term. TOMTOM is ran against the JASPAR and UniPROBE databases which are not species specific while GOMO is ran against sequences from the *Arabidopsis thaliana* genome. This may explain these results; these motifs are similar to known transcription factor binding sites but have not been identified in *Arabidopsis thaliana*.

Including transcription factor expression data measured by RT-PCR (data not shown), provides a list of transcription factors differentially expressed and potentially regulating the genes found in the present study (Table 1). The total number of transcription factors found differentially expressed varies from 15 in TKM5U623 which has 274 genes differentially expressed in total to 0 in TKM6U0051 which has 504 genes differentially expressed. This implies that TKM6U0051 although having the largest introgression segment (Figure 1) has no transcription factors differentially expressed on the array. This reflects that either the micro-array did not contain probes for all known transcription factors or that some transcription factors of tomato are still unknown. The tomato genome annotation lists a number of transcription factors, some of which might be missing on the array. Alternatively, the expression of trans genes might also be the result of other regulatory mechanisms such as microRNA.

In this analysis, we only focused on potential transcription factor binding sites while for the gene expression analysis mRNA is transformed into cDNA which is placed on the micro-array. This implies that some level of post-transcriptional regulation might already have happened and regulatory elements such as microRNA should also be considered. MicroRNAs are small (19 to 25 nucleotide) (Lee, Kim et al. 2004), non-coding, single stranded pieces of RNA binding to specific messenger RNA and preventing their translation into proteins (Mack 2007). Several articles have been published in which microRNAs in tomato have been investigated (Yin, Li et al. 2008; Zuo, Zhu et al. 2012; Karlova, van Haarst et al. 2013), combining information about the localization of these microRNA genes with the list of genes differentially expressed would allow finding microRNA genes outside of the introgressions and potentially regulated by genes within the introgressions or microRNA within the introgression and regulating genes outside of the introgression. Other regulatory systems such as DNA methylation influence the expression of genes as well (Saze, Tsugane et al. 2012). These genes will not appear as differentially expressed in this experiment as their expression is repressed pre-

transcriptionally and as such no mRNA is created, transformed into cDNA and present on the microarray.

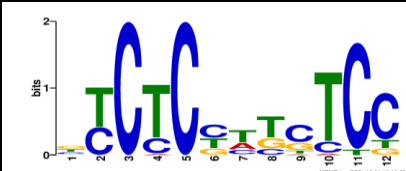
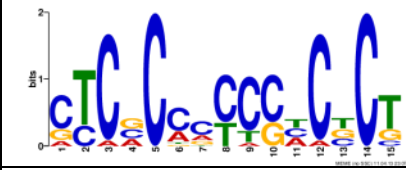
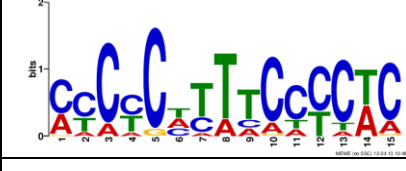
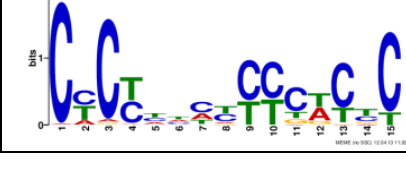
The use of introgression lines allowed us to investigate the interaction between the genes present in this integration and the other genes of the genomes. However, only a portion of the genome was covered by the introgressions. Being able to re-run this analysis on a larger set of genotypes, each having a small introgression segment would help finding more potential transcription factor binding sites. Reducing the introgression segment would help reducing the list of potential transcription factors influencing the genes outside of the introgression. Finally, being able to study these IL at different growth stages, i.e.: having replicates of each genotype at different growth stages could provide insight in the gene regulatory network over time, for example during fruit ripening.



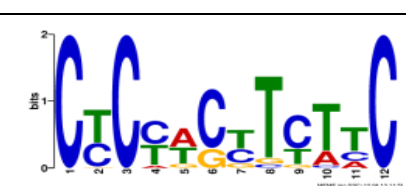
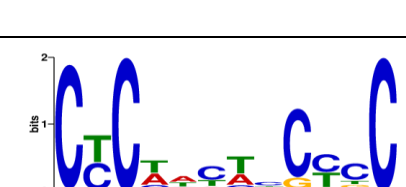


Combining introgression lines with gene expression data also means that a small subset of the data generated is used for transcription factor binding sites. With this experiment we showed that it is possible to search for transcription factor binding sites by combining gene expression and genomic structure but we believe that this should not be the first goal of the experiment while other techniques such as chip-chip or chip-seq exist and provide much more information regarding the gene regulatory network of an organism.

Conclusion

In this study we identified 17 different DNA motifs in the promoter regions of trans genes of a set of six introgression lines. 12 of these motifs are similar to known transcription factors, four of these seem to be well-known and could be used to assess the validity of the results. Using the tomato genome annotation, version SL2.40, it was not possible to assess the processes in which the genes differentially expressed might be involved. Further experiments and improvement of the genome annotation should help in this aspect. The potential transcription factor binding sites will need to be confirmed in the lab by further experimentation. If the relation between transcription factor and transcription factor binding sites can be found, it would allow expanding the gene regulatory network information on tomato, information which might then be integrated into Marker2sequence (Chibon, Schoof et al. 2012) and used to give additional insight into potential candidate genes regulating a QTL.

Supplementary Table 1: Summary of the output of the MEME analysis. It presents for each genotype, the significant motifs (E-value < 0.05) found with their e-value and their number of sites as well as the number of significant hits found for this motif on TOMTOM (E-value < 10 (default) / E-value < 5 / E-value < 1, where the E-value is the expected number of false positives in the matches) and the significant GO terms associated with GOMO. The TOMTOM hits provide information on the number of known transcription factor to which this motif is sequence-wise close. The different E-value are used to provide information on how reliable these hits to known transcription factor are, the lower the e-value, the lower the number of false positive returned. GOMO is used to provide information on what type of GO term could be associated to this DNA motif using the *Arabidopsis thaliana* sequence as reference. The combination of TOMTOM and GOMO provides information regarding the novelty of the motif and the probability that this motif is involved in transcription regulation.

Genotype	Significant motif found	Text motif	E-value	Number of sites	Number of significant hit on TOMTOM	Significant GO term associated with GOMO
TKM5U0580		[GTC][TC]C[TC]C[CTG]][TA][TG][CG]TCC	1.3e-009	36	4 / 1 / 0	Transcription factor Activity Nucleus Chloroplast Microtubule motor activity Protein binding
TKM5U0057		[CG][TC]C[GCA]C[CA] [CA][CT]C[CG][ACT]C [TGC]C[TG]	3.5e-013	18	33 / 27 / 9	Chloroplast Transcription factor activity Microtubule motor activity
TKM5U0595		[CA][CT]C[CT]C[ATC] [TC][TA]TCC[CT]C[TA] C	1.5E-009	28	23 / 16 / 8	Transcription factor activity Chloroplast Regulation of transcription Nucleus Leaf development
TKM5U0623		C[CT]C[TC][TCA][TAC]][CAT][CT][CT][CT]C TA][CT][TC]C	2.3e-045	104	21 / 14 / 9	Transcription factor activity Nucleus Chloroplast Plasma membrane Leaf development

		[GC]G[GT][GA][GA]T TG]GGG[TG]T[GC]GG	1.3e-010	18	29 / 21 / 8	-
		C[TAG]CCG[CG][CA][GT][CA]C[CGA][AG][CG][CA][AC]	4.4e-004	21	36 / 23 / 3	Chloroplast envelope Chloroplast stroma Mitochondrial inner membrane Chloroplast thylakoid membrane Translation
		C[TC]C[CT][AT][CG][T C]T[CT][TA][TC]C	2.4e-008	44	5 / 2 / 1	ATP binding Chloroplast envelope Carotenoid biosynthetic process Chloroplast stroma
TKM5U0637		C[CT]C[TAC][ATC][CT][TA][CT]C[CT]CC	8.5e-016	72	16 / 11 / 4	Transcription factor activity Nucleus Chloroplast envelope Microtubule motor activity Chloroplast stroma
		[GA][GC]A[GC][AG][GC]TG[GC]CGA[GT][C T][GT]	1.1e-006	10	11 / 6 / 3	Chloroplast Mitochondrion
		G[GT][GT]T[TG][GTC] G[GC]G[TG][TC][AGT]GGG	8.9e-004	22	16 / 9 / 5	-

		[GC]C[ACG]G[AC][TA]][GT][CA]TGC[AT][GC]][CG][TGC]	5.1e-008	26	8 / 5 / 3	Mitochondrion Structural constituent of ribosome Translation Ribosome Chloroplast thylakoid membrane
		G[CA]CACGTGTC	1.1 e-004	16	30 / 25 / 14	Chlorophyll binding Chloroplast thylakoid membrane Plastoglobule Response to abscisic acid stimulus Monoxygenase activity
		T[CT]TC[TA][CA][CGA TA]CT[CT][TCA]CT[CT]][CG]	1.2 e-003	15	4 / 1 / 1	Transcription factor activity Nucleus Plasma membrane Protein serine/threonine kinase activity Protein binding
TKM6U0051		C[CT]CC[TA][CT][CT][CTA]C[TC][TC][CT][CT]][CT]C	1.2 e-079	133	26 / 22 / 16	Transcription factor activity Nucleus Microtubule motor activity Chloroplast thylakoid membrane Chloroplast envelope
		[TC]TCT[CT][TC][AT]T [TC]TTCT[TC]C	9.2 e-030	95	7 / 2 / 0	Chloroplast Plasma membrane Nucleus Transmembrane receptor protein tyrosine kinase signaling pathway Protein serine/threonine kinase activity

		<p>G[GTACA][CG][AGT]C CG[GA][CT]G[AG]C</p>	<p>1.4 e-009</p>	<p>29</p>	<p>19 / 11 / 1</p>	<p>Mitochondrion Chloroplast Nucleotide binding</p>
		<p>GTGG[TG][TG][GT][G T]G[GT]G[GC][TAG]G G</p>	<p>7.3 e-007</p>	<p>22</p>	<p>22 / 15 / 6</p>	<p>Chloroplast Mitochondrion</p>

Supplementary Table 2: The annotation of the genes having a common potential transcription factor binding. It presents the list of GO terms, pathways and proteins associated with these genes and their frequency.

	TKM5U0580 (36 genes)		TKM5U0057 (18 genes)		TKM5U0595 (28 genes)			
GO Terms	Ribosome (CC)	1/14	RNA processing (BP)	2/5	ATP binding	1/6		
	oxidation-reduction process (BP)	1/14	regulation of transcription, DNA-dependent (BP)	1/5	transcription repressor activity	1/6		
	zinc ion binding (MF)	1/14	RNA binding (MF)	1/5	protein binding	1/6		
	ATP binding (MF)	1/14	DNA binding (MF)	1/5	regulation of transcription, DNA-dependent	1/6		
	receptor activity (MF)	1/14			DNA binding	1/6		
	cell redox homeostasis (BP)	1/14			Exocyst	1/6		
	intrinsic to membrane (CC)	1/14						
	DNA binding (MF)	1/14						
	enzyme regulator activity (MF)	1/14						
	heme binding (MF)	1/14						
Pathways	Pigment biosynthesis; anthocyanin biosynthesis	1/2						
	Protein modification; protein ubiquitination	1/2						
Proteins	Uncharacterized mitochondrial protein AtMg01410	2/58	Putative CCA tRNA nucleotidyltransferase 1	2/23	B3 domain-containing transcription factor NGA3	1/16		
	Uncharacterized mitochondrial protein AtMg01110	2/58	Putative CCA tRNA nucleotidyltransferase 2	2/23	AP2/ERF and B3 domain-containing transcription factor ARF14	1/16		
	Uncharacterized 8.8 kDa protein in rps12-tRNA-Val intergenic region	1/58	CCA-adding enzyme	2/23	97 kDa heat shock protein	1/16		
	ATP synthase gamma chain 1, chloroplastic	1/58	CCA tRNA nucleotidyltransferase, mitochondrial	2/23	AP2/ERF and B3 domain-containing transcription factor RAV1	1/16		
	Flavonoid 3',5'-hydroxylase	1/58	Probable poly(A) polymerase	2/23	F-box protein At3g07870	1/16		
	NADH-ubiquinone oxidoreductase chain 2	1/58	F-box protein At3g07870	1/23	B3 domain-containing protein Os02g0764100	1/16		
	Disease resistance protein At4g27190	1/58	Ethylene-responsive transcription factor ERF008	1/23	AP2/ERF and B3 domain-containing protein Os05g0549800	1/16		
	ATP synthase gamma chain 2, chloroplastic	1/58	Ethylene-responsive transcription factor ERF018	1/23	Heat shock protein 105 kDa	1/16		
	TMV resistance protein N	1/58	Ethylene-responsive transcription factor RAP2-9	1/23	Putative AP2/ERF and B3 domain-containing protein Os01g0140700	1/16		

	RING-H2 finger protein ATL51	1/58	CCA tRNA nucleotidyltransferase 1, mitochondrial	1/23	AP2/ERF and B3 domain-containing protein Os01g0141000	1/16		
	TKMSU623 Motif 1 (104 genes)		TKMSU623 Motif 2 (18 genes)		TKMSU623 Motif 3 (21 genes)		TKMSU623 Motif 4 (21 genes)	
GO Terms	protein binding	3/29	oxidation-reduction process	1/3	cytoplasm	2/8	ribosome	2/11
	oxidoreductase activity	2/29	protein phosphorylation	1/3	oxidation-reduction process	1/8	RS domain binding	1/11
	calcium ion binding	2/29	nucleus	1/3	protein binding	1/8	adenylosuccinate synthase activity	1/11
	integral to membrane	2/29			adenylosuccinate synthase activity	1/8	cytoplasm	1/11
	lipid metabolic process	2/29			DNA binding	1/8	phosphatidylethanolamine binding	1/11
	nucleus	1/29			asparagine-tRNA ligase activity	1/8	signal transduction	1/11
	intracellular transport	1/29			integral to membrane	1/8	nucleic acid binding	1/11
	signal transduction	1/29					calcium ion binding	1/11
	phosphatidylethanolamine binding	1/29					AMP-activated protein kinase activity	1/11
	DNA binding	1/29					integral to membrane	1/11
Pathways	Secondary metabolite biosynthesis; flavonoid biosynthesis.	2/5			Purine metabolism; AMP biosynthesis via de novo pathway; AMP from IMP: step ½	1/1	Purine metabolism; AMP biosynthesis via de novo pathway; AMP from IMP: step ½	1/1
	Alkaloid biosynthesis; vindoline biosynthesis; vindoline from tabersonine: step 5/6	1/5						
	Pigment biosynthesis; anthocyanin biosynthesis	1/5						
	Alkaloid biosynthesis; scopolamine biosynthesis	1/5						
Proteins	1-aminocyclopropane-1-carboxylate oxidase homolog 3	2/134	Growth-regulating factor 2	1/17	Transcriptional activator Myb	1/27	Adenylosuccinate synthetase	1/55
	1-aminocyclopropane-1-carboxylate oxidase homolog 1	2/134	LRR receptor-like serine/threonine-protein kinase RCH1	1/17	Oleosin Bn-V	1/27	Calcineurin B-like protein 2	1/55
	Feruloyl CoA ortho-hydroxylase 1	2/134	Probable phospholipid hydroperoxide glutathione peroxidase 6, mitochondrial * (1/17) Glutathione peroxidase 1	1/17	Myb-related protein B	1/27	Pre-mRNA-splicing factor SF2	1/55
	Feruloyl CoA ortho-hydroxylase 2	2/134	Probable LRR receptor-like serine/threonine-protein kinase At1g34110	1/17	Myb-related protein A	1/27	CBL-interacting protein kinase 10	1/55
	Naringenin,2-oxoglutarate 3-	2/134	Probable glutathione peroxidase	1/17	Adenylosuccinate synthetase	1/27	CBL-interacting serine/threonine-protein	1/55

Identification of transcription factor binding sites in tomato

dioxygenase		8		2, chloroplastic		kinase 19	
Flavanone 3-dioxygenase	2/134	Glutathione peroxidase	1/17	Oleosin Bn-III	1/27	Oleosin Bn-V	1/55
Thalianol synthase	1/134	Probable phospholipid hydroperoxide glutathione peroxidase	1/17	Catalase	1/27	UPF0098 protein PYRAB11530	1/55
GDSL esterase/lipase At3g14820	1/134	Probable LRR receptor-like serine/threonine-protein kinase At4g26540	1/17	Adenylosuccinate synthetase isozyme 1	1/27	Serine/arginine-rich splicing factor 1A	1/55
2'-deoxymugineic-acid 2'-dioxygenase	1/134	Leucine-rich repeat receptor-like tyrosine-protein kinase At2g41820	1/17	Oleosin Zm-I	1/27	UPF0098 protein CPn_0877/CP_0992/CPj0877/CpB0906	1/55
Calcineurin B-like protein 2	1/134	LRR receptor-like serine/threonine-protein kinase GSO1	1/17	BRI1 kinase inhibitor 1	1/27	60S ribosomal protein L4	1/55

	TKMSU637 Motif 1 (72 genes)		TKMSU637 Motif 2 (10 genes)		TKMSU637 Motif 3 (22 genes)		TKMSU637 Motif 4 (26 genes)	
GO Terms	RNA binding	2/13			regulation of transcription, DNA-dependent	1/3	embryo development	2/3
	RNA processing	2/13			ATP binding	1/3	integral to membrane	1/3
	integral to membrane	2/13			serine-type endopeptidase inhibitor activity	1/3		
	regulation of transcription, DNA-dependent	1/13						
	zinc ion binding	1/13						
	retrograde transport, endosome to Golgi	1/13						
	lipid metabolic process	1/13						
	membrane	1/13						
	serine-type endopeptidase inhibitor activity	1/13						
	binding	1/13						
Pathways	Protein modification; protein ubiquitination	2/3						
	Lipid metabolism; fatty acid biosynthesis	1/3						
Proteins	CCA tRNA nucleotidyltransferase, mitochondrial	2/72	Late embryogenesis abundant protein D-29	1/1	ABSCISIC ACID-INSENSITIVE 5-like protein 2	1/17	Protein LE25	1/4
	CCA-adding enzyme	2/72			ABSCISIC ACID-INSENSITIVE 5-like protein 3	1/17	11 kDa late embryogenesis abundant protein	1/4
	Probable poly(A) polymerase	2/72			ABSCISIC ACID-INSENSITIVE 5-like protein 1	1/17	Late embryogenesis abundant protein D-113	1/4
	Putative CCA tRNA nucleotidyltransferase 1	2/72			ABSCISIC ACID-INSENSITIVE 5-like protein 6	1/17	18 kDa seed maturation protein	1/4
	Putative CCA tRNA nucleotidyltransferase 2	2/72			ABSCISIC ACID-INSENSITIVE 5-like protein 7	1/17		
	ABSCISIC ACID-INSENSITIVE 5-like protein 2	1/72			ABSCISIC ACID-INSENSITIVE 5-like protein 4	1/17		
	ABSCISIC ACID-INSENSITIVE 5-like protein 3	1/72			ABSCISIC ACID-INSENSITIVE 5-like protein 5	1/17		
	ABSCISIC ACID-INSENSITIVE 5-like protein 1	1/72			ABSCISIC ACID-INSENSITIVE 5-like protein 8	1/17		
	ABSCISIC ACID-INSENSITIVE 5-like protein 6	1/72			DEAD-box ATP-dependent RNA helicase 34	1/17		
	ABSCISIC ACID-INSENSITIVE 5-like protein 7	1/72			K(+) efflux antiporter 3, chloroplastic * (1/17) Eukaryotic initiation factor 4A-III	1/17		

	TKMSU637 Motif 5 (72 genes)		TKMSU637 Motif 6 (10 genes)					
Pathways	lipid metabolic process	1/4	intracellular	1/2				
	transferase activity, transferring hexosyl groups	1/4	nitrogen compound metabolic process	1/2				
	integral to membrane	1/4						
	transferase activity, transferring glycosyl groups	1/4						
	Glycan biosynthesis; glycogen biosynthesis	1/1						
Proteins	Galactinol synthase	1/6	Glutamine synthetase cytosolic isozyyme 1-1	1/11				
	Galactinol synthase 1	1/6	Glutamine synthetase cytosolic isozyyme	1/11				
	Uncharacterized protein R707	1/6	Glutamine synthetase nodule isozyyme	1/11				
	Late embryogenesis abundant protein D-29	1/6	Glutamine synthetase N-1	1/11				
	Glycogenin-1	1/6	Putative poly [ADP-ribose] polymerase 3	1/11				
	Glycogenin-2	1/6	Glutamine synthetase root isozyyme 1	1/11				
			Glutamine synthetase cytosolic isozyyme 1	1/11				
			Glutamine synthetase cytosolic isozyyme 2	1/11				
			Glutamine synthetase	1/11				
			Poly [ADP-ribose] polymerase 3	1/11				

	TKM6U0051 Motif 1 (133 genes)		TKM6U0051 Motif 2 (95 genes)		TKM6U0051 Motif 3 (29 genes)		TKM6U0051 Motif 4 (22 genes)	
GO Terms	binding	4/30	protein binding	3/30	oxidation-reduction process	1/6	transferase activity, transferring acyl groups other than amino-acyl groups	1/1
	protein binding	3/30	metabolic process	2/30	inorganic phosphate transmembrane transporter activity	1/6		
	protein phosphorylation	3/30	xyloglucan:xyloglucosyl transferase activity	1/30	zinc ion binding	1/6		
	oxygen binding	2/30	carbohydrate metabolic process	1/30	transferase activity, transferring acyl groups other than amino-acyl groups	1/6		
	heme binding	2/30	nucleus	1/30	catalytic activity	1/6		
	zinc ion binding	2/30	endopeptidase inhibitor activity	1/30	integral to membrane	1/6		
	kinase activity	2/30	metal ion transport	1/30				
	oxidation-reduction process	1/30	kinase activity	1/30				
	cell redox homeostasis	1/30	alpha-amylase inhibitor activity	1/30				
	heat shock protein binding	1/30	transmembrane receptor protein kinase activity	1/30				
Pathways	Protein modification; protein ubiquitination	4/7	Secondary metabolite biosynthesis; dhurrin biosynthesis; dhurrin from L-tyrosine: step 3/3	1/4	Amino-acid biosynthesis; L-isoleucine biosynthesis; L-isoleucine from 2-oxobutanoate: step 2/4	1/3		
	Amino-acid biosynthesis; L-isoleucine biosynthesis; L-isoleucine from 2-oxobutanoate: step 2/4	1/7	Amino-acid biosynthesis; L-tyrosine biosynthesis; L-tyrosine from L-arogenate (NADP(+) route): step 1/1	1/4	Amino-acid biosynthesis; L-valine biosynthesis; L-valine from pyruvate: step 2/4	1/3		
	Amino-acid biosynthesis; L-valine biosynthesis; L-valine from pyruvate: step 2/4	1/7	Pigment biosynthesis; anthocyanin biosynthesis	1/4	Protein modification; protein ubiquitination	1/3		
	Pigment biosynthesis; anthocyanin biosynthesis	1/7	Protein modification; protein ubiquitination	1/4				
Proteins	Mitochondria fission protein 1	2/162	Pentatricopeptide repeat-containing protein At3g63370	1/171	E3 ubiquitin-protein ligase AIP2	1/33	Thylakoid lumenal protein Atlg03610, chloroplastic	1/1
	Cytochrome P450 86B1	2/162	Scarecrow-like protein 34	1/171	RING finger protein 126-B	1/33		
	Cytochrome P450 94A1	2/162	Scarecrow-like protein 33	1/171	RING finger protein 126-A	1/33		
	Cytochrome P450 94A2	2/162	Scarecrow-like protein 30	1/171	Scarecrow-like protein 34	1/33		

Identification of transcription factor binding sites in tomato

Cytochrome P450 52A6	2/162	Scarecrow-like protein 31	1/171	Protein SAMHD1 homolog	1/33		
Cytochrome P450 52A5	2/162	Cytochrome P450 86B1	1/171	Scarecrow-like protein 33	1/33		
Cytochrome P450 704C1	2/162	Serine/threonine-protein kinase BRI1-like 2	1/171	Scarecrow-like protein 30	1/33		
Probable LRR receptor-like serine/threonine-protein kinase Atlg56140	2/162	U-box domain-containing protein 27	1/171	Scarecrow-like protein 31	1/33		
Cytochrome P450 86A1	2/162	Serine/threonine-protein kinase BCK1/SLK1/SSP31	1/171	Scarecrow-like protein 21	1/33		
Cytochrome P450 86A2	2/162	U-box domain-containing protein 25	1/171	RING finger protein 126	1/33		

Chapter 6: Annotex: Exploring the genome annotation

Pierre-Yves Chibon^{1,2,3}, Arnaud Bovy^{1,3}, Richard G.F. Visser^{1,2,3} and Richard Finkers^{1,3}

¹ Wageningen UR Plant Breeding, Wageningen University and Research Centre, PO Box 386, 6700 AJ Wageningen, The Netherlands

² Graduate School Experimental Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands

³ Centre for BioSystems Genomics, Wageningen, The Netherlands

Abstract

The number of sequenced and annotated genomes has increased drastically over the last few years as well as the number of online resources providing information about genes, proteins or pathways. As more and more databases are created, cross-references between them become more and more important as well. These cross-references may indicate that two elements from two different databases are in fact the same, or that two elements from two different databases are related biologically or functionally. As a result, the time spent by a researcher to decipher which genes are involved in a certain pathway and/or related to a specific biological process has increased dramatically. Tools such as Marker2sequence allow exploration of a specified region of the genome to find genes related to a specific process or pathway. However, they do not allow a genome-wide search for all genes fitting specified criteria (such as genes related to a specific GO term or protein). Annotex offers this functionality by integrating four public resources, namely: UniProt, GO, Rhea and ChEBI with the genome annotations of potato and tomato. Annotex allows querying any biological entity or annotation (gene, protein, pathway, GO term, publications, biochemical reaction, metabolite) from almost any biological entity or annotation (gene, protein, pathway, GO term, metabolite). By integrating these different resources, Annotex also shows the current state of data and resources with regards to data integration.

Annotex is freely available at: <http://www.plantbreeding.wur.nl/BreeDB/annotex/>

Introduction

Every year, the number of biological databases available on the Web is increasing. The journal *Nucleic Acid Research* (NAR) publishes a yearly issue dedicated to databases. In its latest issue, NAR announced 1512 biological databases, 132 more than in 2012 (Fernández-Suárez and Galperin 2013). We can distinguish two distinct classes of databases. The first type of databases contains biological information such as genes (Maglott, Ostell et al. 2005; Flicek, Ahmed et al. 2013), proteins (The UniProt Consortium 2013), metabolites (Wishart, Tzur et al. 2007; Matos de, Alcantara et al. 2010), or biochemical reactions (Kanehisa and Goto 2000; Caspi, Altman et al. 2010; Alcantara, Axelsen et al. 2012). Biological evidences link together elements from different databases. For example, genes encode for proteins. The link between, for example, a gene and its corresponding protein is called cross-reference in database terms. The second set of databases contains information about biological concepts. These ontologies are a sort of dictionaries, used to define concepts and relations between them (Gruber 1993). One example is the Gene Ontology (Ashburner, Ball et al. 2000) which is used as the central place to define concepts related to cellular location, biological processes and molecular functions. Genes and proteins are annotated using the GO terms, to provide insight in their location and their action. Ontologies are built as a hierarchical tree where each term verifies and specifies its parents. For example, the GO term GO:0006952 “Defense response” is a child of the GO term GO:0006950 “Response to stress”. So the GO term GO:0006952 is a more specific classification of response to stress (Figure 1). This means that the lower down the tree one goes, the more specific the terms are. It also allows annotating genes and proteins with generic terms when not enough information allows attributing them to more specific ones. It also means that when searching for the genes associated with a GO term, one may want to search for the gene associated with this GO term

but also the genes associated with all the children of this GO term as they all verify the condition of their parent.

- I GO:0008150 biological process [497680 gene products]
- I GO:0050896 response to stimulus [88966 gene products]
- ▼ **GO:0006950 response to stress [38751 gene products]**
 - I GO:0033554 cellular response to stress [18674 gene products]
 - I GO:0006952 defense response [11863 gene products]
 - I GO:0033555 multicellular organismal response to stress [396 gene products]
 - I GO:0003299 muscle hypertrophy in response to stress [50 gene products]
 - I GO:0009271 phage shock [25 gene products]
 - R GO:0080134 regulation of response to stress [6306 gene products]
 - I GO:0034059 response to anoxia [43 gene products]
 - I GO:0009409 response to cold [1028 gene products]
 - I GO:0009413 response to flooding [2 gene products]
 - I GO:0034405 response to fluid shear stress [123 gene products]
 - I GO:0009408 response to heat [1690 gene products]
 - I GO:0009635 response to herbicide [73 gene products]
 - I GO:0051599 response to hydrostatic pressure [24 gene products]
 - I GO:0055093 response to hyperoxia [107 gene products]
 - I GO:0001666 response to hypoxia [1703 gene products]
 - I GO:0035902 response to immobilization stress [11 gene products]
 - I GO:0002931 response to ischemia [31 gene products]
 - I GO:0035900 response to isolation stress [0 gene products]
 - I GO:0051409 response to nitrosative stress [182 gene products]
 - I GO:0006970 response to osmotic stress [1931 gene products]
 - I GO:0006979 response to oxidative stress [4330 gene products]
 - I GO:0042594 response to starvation [3185 gene products]
 - I GO:0006991 response to sterol depletion [69 gene products]
 - I GO:0035966 response to topologically incorrect protein [1172 gene products]
 - I GO:0009414 response to water deprivation [489 gene products]
 - I GO:0009611 response to wounding [5907 gene products]

Figure 1: Representation of the Gene Ontology tree for the GO term GO:0006950. This term has two parents and 26 children which themselves may have children. In total, the GO term GO:0006950 is a parent for 495 GO terms. On the right side of each GO term the number of genes (from UniProtKB, MGI, RGD, TAIR and SGN, etc) associated with this GO term or one of its children is presented.. On the left side of the GO term is presented the type of relation between the term and its parent, here two types are present: "I" for an "is a" relationship and "R" for a "regulates" relationship. It can be read as "GO:0080134 regulates GO:0006950". (Source: http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0006950)

Next generation sequencing has made sequencing a genome much quicker and cheaper (Mardis 2011; Kircher 2012). As a result, the number of genomes sequenced is rapidly increasing, including a number of important crop species and feed animals. Examples include rice (Goff, Ricke et al. 2002; Yu, Hu et al. 2002), grapevine (Jaillon, Aury et al. 2007), maize (Schnable, Ware et al. 2009), apple (Velasco, Zharkikh et al. 2010), potato (PGSC, Xu et al. 2011), tomato (Tomato genome consortium 2012), chicken (Chicken genome consortium 2004), swine (Groenen, Archibald et al. 2012), cow (Sequencing, Consortium et al. 2009) and sheep (Archibald, Cockett et al. 2010). Each of these genomes has been annotated with, (putative) genes models, regulatory elements, and polymorphisms. Genes may be annotated with cross-references to proteins, protein domains and gene ontology (GO) terms. The quality of this annotation is becoming a key factor in their

exploitation: “the value of the genome is only as good as its annotation” (Stein 2001) as the link it provides is a source of information when gathering information about a gene. A better annotation implies better links for each gene and a more accurate representation of the function of the gene in the organism.

Several tools exist to browse and query a genome annotation. The first example is GBrowse (Stein, Mungall et al. 2002; Donlin 2009). GBrowse allows browsing of genome sequencing data with its associated annotations. It allows zooming into specific regions and allows searching the annotations. However, searching in GBrowse is only limited to annotations loaded in the GBrowse database and not all the other data associated to the stored information via cross-references. The second example is Marker2sequence (M2S) (Chibon, Schoof et al. 2012), which offers functionality to query genome annotations for all genes in a specified region (e.g. a quantitative trait locus). M2S allows only searching the genome annotations, and also includes linked data, associated via cross-references in the search. However, M2S only allows searching delimited genome regions. The third example is Ondex (Köhler, Baumbach et al. 2006). Ondex offers a way to integrate and visualize different resources (GO, TAIR, KEGG SGD, UniProt). However, Ondex relies on visualizing the relationships in the data in a graph. The interpretation of such a graph becomes very complex when large datasets, such as genome annotation, are loaded. None of these tools is able to, for example, easily interrogate the relationship between a gene and a metabolite.

As there are many databases available online, biologists and breeders are facing the challenge to know which relevant information is available where, and how to use it. One solution would be to access all this information from a central place. This central place would provide the information either by a local copy of the different databases (so called data-warehouse) or by being able to integrate the information from the different databases on the fly. For the latter option, the different resources need to be interoperable. One solution to achieve this is using semantic web technology. The semantic web technology is designed to uniquely identify each concept and element and map (link) them (Berners-Lee, Hendler et al. 2001). Using this technology it would be easy to uniquely identify a biological element, explicit that this element is also present in other resources and provide complementary information by linking to other databases with an explicit relation. For example, this technology could be used to identify a gene present in the NCBI Entrez database, explicit that this gene is also described in the EMBL database and that it encodes for a known protein described in UniProt. Several major database providers, such as EBI and the Swiss Bioinformatics Institute, have seen the potential of the semantic web and do provide their resources, such as RHEA, Uniprot, GO in a semantic web compatible format.

Researchers have expressed the need to, for example, easily interrogate the relationship between a gene and a metabolite. In a more generalized schema, a researcher might want to ask any question starting from one entity and obtain all results about another entity (Figure 2). This, together with the availability of several of the major databases in semantic format, has led to the development of Annotex as an easy to use tool for asking such questions.

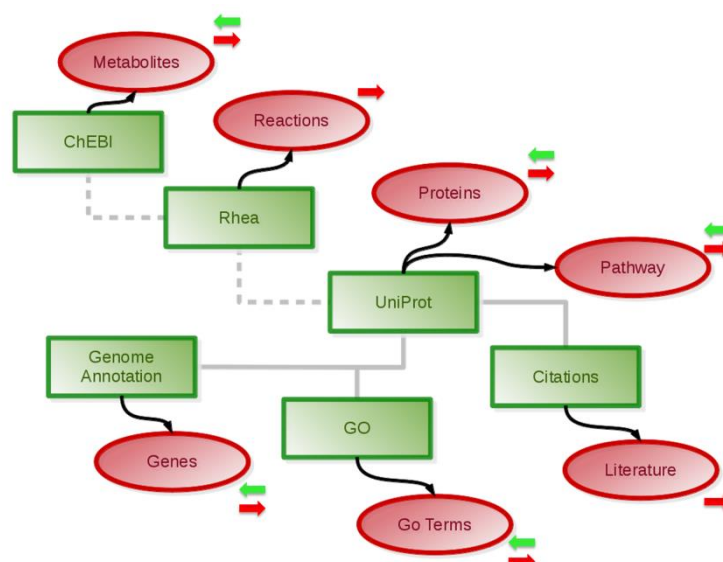


Figure 2: Scheme representing the data integration performed by Annotex. The green boxes represent the different data sources and the red ovals represent the type of information they provide. The plain lines represent direct integration between the databases, while the dashed lines represent indirect integration. Indirect mapping is used here to describe a situation where the databases are using two different identifiers to represent the same concept/entity. The green arrows represent the type of information that can be provided as input, while the red arrows represent the type of information that can be asked as output. These arrows show how the system allows querying any type of information from almost any input.

Materials and methods

Annotex integrates protein information from UniProt (Feb 2013), biochemical compounds from ChEBI (version of February, 13 2013), biochemical reactions from Rhea (version of February, 13 2013), gene ontology from GO (version of February, 12 2013) and genome annotations of tomato (version 2.3) (Tomato genome consortium 2012) and potato (PGSC, Xu et al. 2011) (Figure 2).

Annotex integrates the genome annotations of potato and tomato but can be expanded to support any genome, provided that the genome annotation is available in the right format. Both potato and tomato genome annotations have been converted to a RDF representation using the gff2RDF tool (<http://github.com/PBR/gff2RDF>, (Chibon, Schoof et al. 2012)) which uses the URI provided by UniProt and GO to link the annotation to these resources.

The databases (Rhea, ChEBI and UniProt) and genome annotations have been uploaded to a Virtuoso Open-Source Edition (version 6.1) (Erling and Mikhailov 2007) and Annotex performs the integration via SPARQL (Prud'hommeaux and Seaborne 2008) queries. The web interface was implemented in Java as part of the BreeDB framework (<http://www.wageningenur.nl/en/show/BreeDB.htm>).

Results

Annotex provides a flexible interface which allows linking any type of data to retrieve any other type of data and thus exploring the genome annotations integrated within the network of knowledge built from other resources.

Annotex follows a two-step approach. When a user chooses an organism, enters an input, specifies its type and selects the desired type of output (Figure 3), Annotex searches for elements of the

specified type containing the specified input in their name or identifier (depending on the type of the input). If only one result is found, Annotex directly moves to the result page. If several results are found, the user is given the choice to disambiguate the input among the different possibilities (see first example in the examples section below). Once the user has disambiguated the input, Annotex shows the result page.

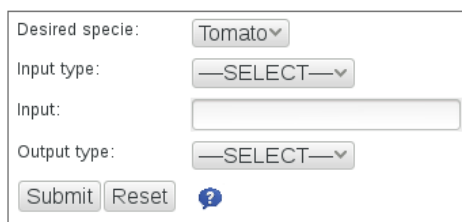


Figure 3: Screenshot of the input box for Annotex. In the box, the user can specify the species to query (Tomato or Potato), the type of input given, the input itself (being an identifier or a name) and specify which type of output to retrieve. The available input types are: Gene, Protein, Metabolite, GO, GO and child, Pathway. The available output types are: Gene, Protein, Metabolite, GO, Pathway, Reaction, Article.

Annotex is implemented as a web application with an input page, which is divided into two parts. The upper part contains an introduction to Annotex with a scheme outlining how the data is integrated and what type of information is available (Figure 2). The lower part contains the input box itself (Figure 3). This input box asks for four different inputs. The first parameter is a selection box listing the organisms available. The second parameter is another selection box listing the types of the information entered by the user as third parameter. The types available are: Gene, Protein, Metabolite, GO, GO and children, and Pathway. The third parameter is a text field in which an identifier from one of the integrated databases can be inserted (i.e.: the tomato gene identifier: Solyc06g074240 or the UniProt protein identifier Q38933), or a keyword (i.e.: molecule or GO term “beta-carotene”). The fourth and last parameter is a selection box specifying the type of output desired. The different outputs possible are: Gene, Protein, Metabolite, GO, Pathway, Reaction and Article. Biochemical reactions are difficult to represent in a string of characters, therefore, they are not supported as input type for Annotex.

By offering the input types “GO and children” and “GO” Annotex provides a way to retrieve all the genes, proteins, metabolites, articles which are associated with the specified GO term and all its children or with the specified GO term only.

The result page of Annotex provides a link to the specific page of each result on their provider (i.e.: UniProt for proteins, ChEBI for metabolites). In addition the result page offers the possibility to download the output either as CSV or JSON. CSV can be opened in any spreadsheet program and is a format familiar to the biologists while JSON is a computer-friendly format familiar to bioinformaticians.

The amount of information integrated from UniProt is so large that searching for all the proteins having the specified keyword in their names is too slow to be integrated on a website. The only way to use Annotex with a protein as input is, therefore, by providing to Annotex the UniProt identifier of this protein.

Examples

Annotex can answer a number of questions such as 1) find all genes involved in a pathway or 2) find genes related to a specific compound or 3) retrieve all publications related to a gene or 4) retrieving the GO term associated with a gene or a protein. In this section, two use-cases for Annotex will be discussed.

Example 1: Beta-carotene in tomato

Beta-carotene is an important compound in tomato. Besides its health beneficial properties (Lincoln and Porter 1950), it is also the compound responsible for the red color of the fruit (Ray, Moureau et al. 1992; Gady, Vriezen et al. 2012). This compound has therefore a dual interest for tomato breeders from a sales and marketing point of view.

Previously, Quantitative Trait Analysis (QTL) studies have identified loci for beta-carotene content in tomato on chromosome 2, 3 and 6. For the QTL located on chromosome 3 and 6 (Paran, Goldman et al. 1995), the functional genes underlying the QTL are known: Solyc03g007960.1.1 on chromosome 3 (Galpaz, Ronen et al. 2006) and Solyc06g074240.1.1 on chromosome 6 (Kilambi, Kumar et al. 2013). For the search for “Gene” from the “Metabolite” “beta-carotene” in the tomato organism, Annotex finds six different compounds having the words “beta-carotene” in their name (Figure 4), it will thus ask for disambiguation of the term “beta-carotene”. The compound of interest presently, is the compound with the ChEBI identifier 17579 for which the name is simply “beta-carotene”. For this specific compound, Annotex returns seven genes on the tomato genome (Figure 5). Of these seven genes, one is on chromosome three, one is on chromosome four, two are on chromosome six, one is on chromosome eight, one is on chromosome ten and the last one is on chromosome 12. It should be noted that the gene found on chromosome three is the known candidate gene (Galpaz, Ronen et al. 2006). Additionally, of the two genes on chromosome six, Solyc06g074240.1.1 is the gene responsible for the presence of the QTL on this chromosome as well. Of the other four genes, three (on chromosome four, ten and twelve) are linked, via their proteins, to the article of Bonk et al. (1997) which studied four enzymes of the carotenoid pathway *in vitro*. The genes on chromosomes ten and twelve are related to the publication of Guzman et al. (2010) who studied carotenoids in *Capsicum* spp. (pepper). The gene on chromosome four is involved in retinol metabolism which is relying on beta-carotene. It is also related, according to the literature found via Annotex, with rice (Yu, Wang et al. 2005) and (Ohyanagi, Tanaka et al. 2006), sorghum (Paterson, Bowers et al. 2009) and is involved in the biosynthesis of carlactone, a strigolactone-like plant hormone (Alder, Jamil et al. 2012) and of neoxanthin (product of the degradation of beta-carotene) in potato (Al-Babili, Huguency et al. 2000).

Identifier	Name
17579	beta-carotene
27793	beta-carotene 5,6-epoxide
35309	(5S,6R)-beta-carotene 5,6-epoxide
67188	9-cis-beta-carotene
67227	beta-carotene 15,15'-epoxide
67228	15,15'-dihydroxy-beta-carotene

Figure 4: List of compounds found by Annotex in ChEBI having “beta-carotene” in their name. This is how the disambiguation page looks. By clicking on the link on the left column, the user specifies which of these compounds is of interest.

Gene

Input Metabolite 17579			
Gene	Scaffold	Position	Description
Solyc03g007960.1.1	SL2.31ch03	2449187 - 2451252	Carotene beta-hydroxylase (AHRD V1 ***** Q0GGX1_SOLLC)%3B contains Interpro domain(s) IPR006694 Fatty acid hydroxylase
Solyc04g040190.1.1	SL2.31ch04	31103094 - 31104596	Beta-lycopene cyclase (AHRD V1 ***- A6YS01_SOLLC)%3B contains Interpro domain(s) IPR010108 Lycopene cyclase%2C beta and epsilon
Solyc06g036260.1.1	SL2.31ch06	22432028 - 22433582	Beta-carotene hydroxylase (AHRD V1 ***** Q9S6Y1_SOLLC)%3B contains Interpro domain(s) IPR006694 Fatty acid hydroxylase
Solyc06g074240.1.1	SL2.31ch06	42289963 - 42291459	Lycopene beta-cyclase (AHRD V1 ***- B7U386_ACTCH)%3B contains Interpro domain(s) IPR010108 Lycopene cyclase%2C beta and epsilon
Solyc08g066650.1.1	SL2.31ch08	52619865 - 52622706	15 15%26apos beta carotene dioxygenase (AHRD V1 *- *- Q2S6A1_SALRD)%3B contains Interpro domain(s) IPR004294 Carotenoid oxygenase
Solyc10g079480.1.1	SL2.31ch10	60348153 - 60349655	Beta-lycopene cyclase (AHRD V1 ***- A6YS01_SOLLC)%3B contains Interpro domain(s) IPR010108 Lycopene cyclase%2C beta and epsilon
Solyc12g008980.1.1	SL2.31ch12	2286570 - 2291525	Lycopene beta cyclase (AHRD V1 ***** C1N7E6_MICPS)%3B contains Interpro domain(s) IPR010108 Lycopene cyclase%2C beta and epsilon

Figure 5: List of all the genes found by Annotex to be related to the ChEBI compound “17579” (beta-carotene). Each gene identifier on the left column is a link to a page presenting more information about the gene. This list can be downloaded in CSV or JSON format for further processing.

Using Annotex, seven genes could be found related to beta-carotene in the tomato genome annotation. Of these seven genes, three were expected from the QTL mapping analysis and four were found on other chromosomes where no QTL were found using this population and marker combination. For biologists Annotex offers thus a way to find genes related to a specific trait or metabolic process while for breeders it brings potential new genes which might be of interest as potential candidate genes to validate and, eventually, introgress. It might also show important associations that had not been considered before, such as genes involved in the production of other compounds relying on the one in the search term.

Example 2: Intersect or differentiate different genes based on multiple criteria in potato

Potato is the third food crop in terms of food consumption world-wide (Visser, Bachem et al. 2009). Between 1845 and 1852, the pathogen *Phytophthora infestans* impacted so much the Irish potato production that approximately 1 million people died of starvation and more than 1 million people immigrated. Still today *P. infestans* is the major disease in potato worldwide. Breeding for resistance in *Solanum tuberosum* is therefore an important economic goal. *Phytophthora infestans* is not the only pathogen of potato whose interactions are studied. Other diseases caused by bacteria (such as *Erwinia carotovora* (Reiter, Pfeifer et al. 2002) or *Ralstonia solanacearum* (Esposito, Ovchinnikova et al. 2008)) and fungi (McArthur and Knowles 1992; McArthur and Knowles 1993) are also of importance and thus studied.

Annotex can be queried to retrieve the list of genes associated with a GO term. When searching with Annotex for genes related to the GO term “response to stress” (GO:0006950) or one of its children, 714 potato genes are returned. Of these 467 genes are associated with “defense response” (GO:0006952) or one of its children. Among the children of “defense response” are “defense response to fungus” (GO:0050832) and “defense response to bacterium” (GO:0042742). Using Excel or any programming language, different lists corresponding to different GO terms can be integrated and the intersection between all these lists can be retrieved. In this way, one can find that three genes (PGSC0003DMG400008096, PGSC0003DMG400019435, PGSC0003DMG400019437) share the common annotation of being related to “defense to fungus” and “defense to bacterium”. These three genes are therefore very interesting candidates to study when considering genes in potato involved in diverse resistance mechanisms.

Annotex can also be used to evaluate the differences between two lists. “Defense response” or its children shows relationships to 467 genes; however, “defense response” alone is related to 431 genes itself, leaving 36 genes related to one of its children. Using Annotex, we can see to which child of “defense response” these 36 genes are related. First, retrieve the list of the 467 genes related to “defense response” and its children. Retrieve the list of 431 genes related to “defense response” itself. Then, using Excel or a programming language, analyze the difference between these two lists. Finally, for each gene on this list, query Annotex to retrieve the GO terms they relate to. In this way, it appeared that “innate immune response” (GO:0045087) relates to 33 of the 36 genes. The other three genes are the ones mentioned above that are related to “defense response to fungus” (GO:0050832) and “defense response to bacterium” (GO:0042742).

By these two examples, two different workflows relying on the data integrated by Annotex have been shown (Figure 6). The first workflow searches for the biochemical reactions involving the metabolite beta-carotene, then the proteins involved in these biochemical reactions and finally returns the genes related to these proteins according to the genome annotation (Figure 6a). The second workflow has two different ways. Using the genome annotation it can return the list of genes related to a specific GO term and using the Gene Ontology and the genome annotation it can retrieve the children of a GO term and the list of genes associated with at least one of these children (Figure 6b).

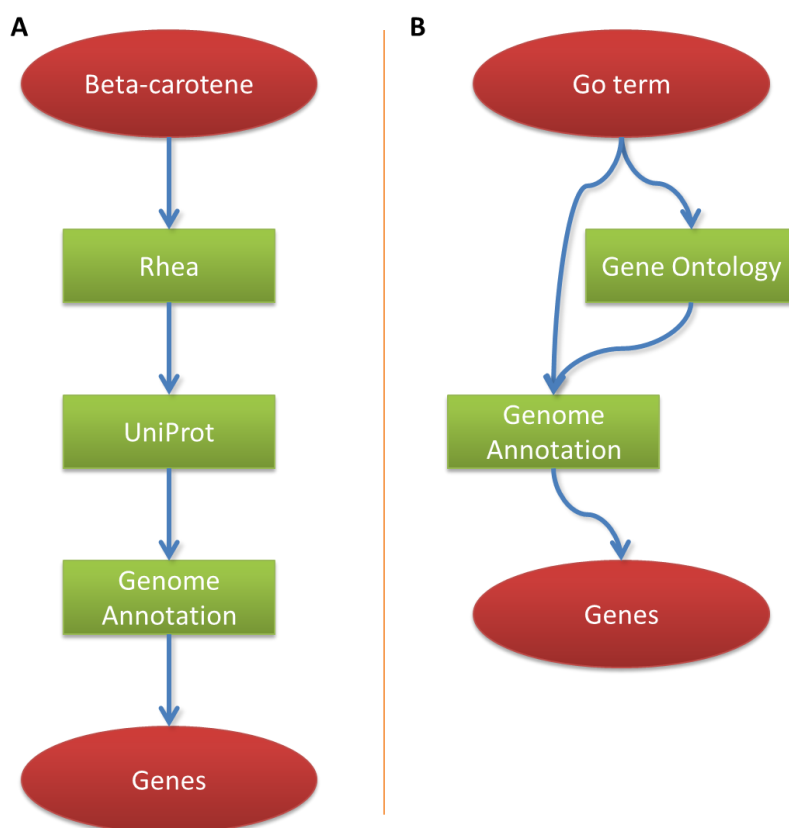


Figure 6: Workflow followed in the examples. Example 1 (part A) starts from a defined metabolite (beta-carotene), searches for all biochemical reactions involving this compound using Rhea, extract the proteins involved in these reactions and using the genome annotation, retrieve the list of genes related to these proteins. Example 2 (part B) uses the genome annotation to retrieve the list of genes associated with a single GO terms, or using the Gene Ontology, retrieves the children of the GO terms and uses the genome annotation to retrieve the list of genes associated with at least one of these GO terms.

Discussion

Annotex is a unique tool showing how data integration can be used to filter out genes of interest using their relations to information provided outside of the genome annotation.

Except the genome annotations, all the resources integrated by Annotex are not species-specific, which allows to return information of interest which might not have been the primary search target of the user. In this case, *Arabidopsis thaliana* being a model organism for plants, the user might have looked for this type of literature but the annotation of the gene Solyc04g028390.1.1 also links to two articles from human biology (Freimuth, Raftogianis et al. 2000; Allali-Hassani, Pan et al. 2007) where it is said that in humans sulfotransferase enzymes are involved in the metabolism of drugs and hormones, functions that might be related in plants as Varin, DeLuca et al. (1992) concluded that there might be an evolutionary link for sulfotransferases between plant and animal. Links to literature outside of the field of study might in this case provide insight into the function of this protein.

Annotex offers the possibility to download the output of a search in CSV or JSON, allowing biologists to approach a question from different angles. Biologists and breeders can retrieve the list of entities matching the provided criteria and look at the intersection or difference between these lists using a spreadsheet program, R (R Core team 2013), Galaxy (Goecks, Nekrutenko et al. 2010) or the VirtualPlant platform (Katari, Nowicki et al. 2010). Bioinformaticians can query Annotex for a number of criteria and return the intersection of these different criteria using any programming language and the JSON export function. Future plans include expanding the current user interface of Annotex to allow anyone to build such queries. The user should be able to provide a list of criteria to match or not match and retrieve the output desired.

However, the results of Annotex are limited by the current state of interoperability and the content of the resources used.

Biased resources

In the second example, we have shown that “Defense response” or its children shows relationships to 467 genes; however, “defense response” alone is related to 431 genes itself, leaving 36 genes related to one of its children. We therefore went to the level above, “Response to stress” and using a custom script, we queried the genes associated with each of the 504 children of “Response to stress” in potato (Figure 7).

In potato, the 504 child GO terms of “response to stress” are related to 689 different genes. The distribution of these 689 genes over the 504 GO terms is completely unbalanced (Figure 7). This shows a flaw in the potato genome annotation: there is a clear disequilibrium for defense response and DNA repair mechanisms (six of the 13 GO terms are related to DNA repair) in the annotation. This bias could result from either a lack of knowledge (no biological evidences) or a lack of availability of this knowledge (there are biological evidences but, for example, “hidden” in the literature) which is thus not included in the genome annotation pipeline. Improving the genome annotation, using collaborative tools such as Orcae (Sterck, Billiau et al. 2012) or relying on known experiments and data to iteratively improve the genome annotation would benefit Annotex and improve its results. However, having new resources to improve the content of existing resources only makes the situation of the spreading of the information worse: the original information is in one place, the improved information in another.

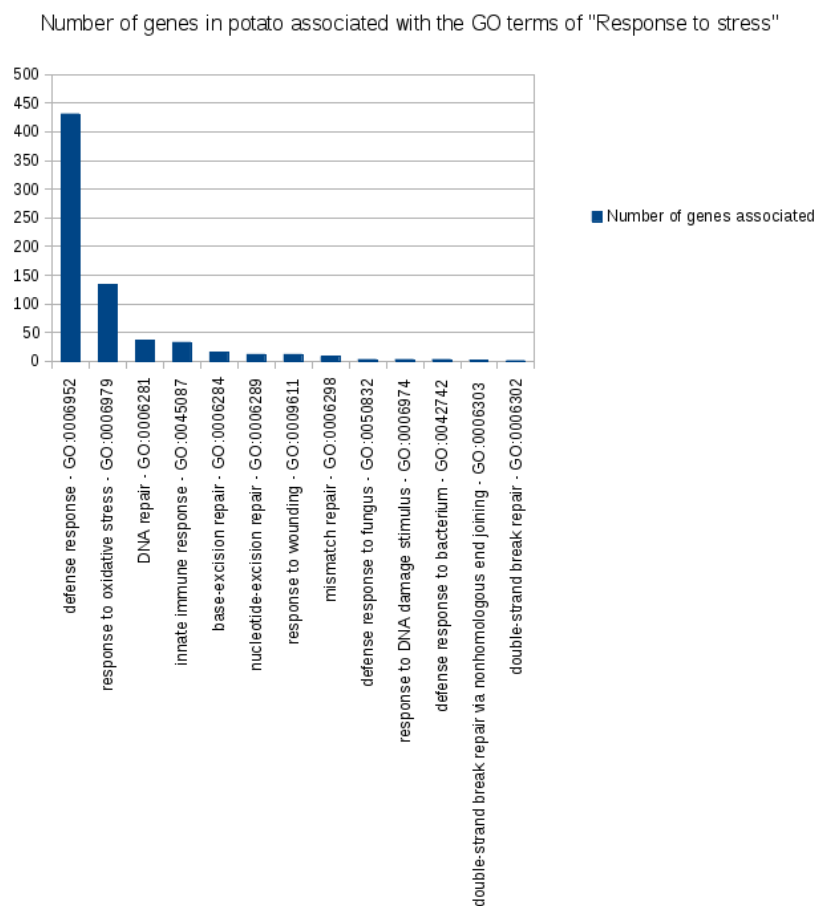


Figure 7: Of all the 504 children of the GO term “response to stress” only 13 are related to at least one gene in the potato genome annotation. This graph shows the distribution of the 689 potato genes related to these 13 GO terms.

Spreading of the information

Many resources provide pathway information: UniPathway (Morgat, Coissac et al. 2012), KEGG (Kanehisa and Goto 2000), MetaCyc (Caspi, Altman et al. 2010), Solcyc (Pujar, Caspi et al. 2009), WikiPathway (Kelder, van Iersel et al. 2012). All the resources have their own specificities and access rules. UniPathway, Solcyc and WikiPathway can be freely downloaded while KEGG and MetaCyc are only available under upon registration (paid for KEGG). SolCyc is species specific, WikiPathway provide pathways information per species, MetaCyc and KEGG provide generic pathways as well as species specific and UniPathway provides only generic pathways.

Incomplete resources

Rhea, the database used to map biochemical compounds to proteins via the chemical reactions in which they are involved is not well covering the plant reactions and thus provides limited mapping. For example, compounds such as flavonoid (ChEBI:47916) or phenylalanine (ChEBI:28044) are present in ChEBI but are not linked to any reactions in Rhea. In fact, for flavonoid, Annotex proposes nine different compounds for disambiguation, containing “flavonoid” in their name. From these nine compounds, none is found associated with a biochemical reaction in Rhea. For phenylalanine, of the 23 compounds containing “phenylalanine” proposed for disambiguation by Annotex, only two are linked to biochemical reactions, proteins and genes: L-phenylalanine zwitterion (ChEBI:58095) and D-phenylalanine zwitterion (ChEBI:57981). Other compounds such as quercetin, (Femia, Caderni et al. 2003), isobutylthiazole (Baldwin, Scott et al. 1998) or methylbutanol (Baldwin, Scott et al. 1998),

have been studied in tomato but are not present at all in ChEBI. Table 1 summarizes for 24 compounds known and studied in tomato if they are present in ChEBI and Rhea. 15 of these compounds were found in ChEBI but only 6 in Rhea, showing the lack of integration of plant metabolites in ChEBI and the lack of biochemical reactions involving these compounds in Rhea.

Table 1: Table presenting for a list of compounds known in tomato whether they are found in ChEBI and in Rhea. Of these 24 compounds, 15 were found in ChEBI and 6 in Rhea.

	Present in ChEBI	Present in Rhea
Benzyl alcohol	✓ ChEBI: 17987	✓
Citric acid	✓ ChEBI: 30769	✗
Eugenol	✓ ChEBI: 4917	✓
Flavonoid	✓ ChEBI:47916	✗
Fructose	✓ ChEBI: 28757	✗
Glucose	✓ ChEBI: 37624	✓
Glutamic acid	✓ ChEBI: 18237	✗
Guaiacol	✓ ChEBI: 28591	✗
Hexanal	✗	✗
Hexanol	✗	✗
Isobutylthiazole	✗	✗
Isobutylthiazole	✗	✗
Isoleucine	✓ ChEBI: 24898	✗
Leucine	✓ ChEBI: 25017	✗
Malic acid	✓ ChEBI: 6650	✗
Methyl salicylate	✓ ChEBI: 31832	✗
Methylbutanal	~ 3-methylbutanal ChEBI: 16638	✓
Methylbutanol	✗	✗
Methylbutanol	✗	✗
Phenylacetaldehyde	✓ ChEBI: 16424	✓
Phenylalanine	✓ ChEBI: 28044	✗
Phenylethanol	✗	✗
Quercetin	✗	✗
Sucrose	✓ ChEBI: 17992	✓

Adding a more plant oriented resource such as SolCyc (Pujar, Caspi et al. 2009) to link chemical compounds to proteins would improve the mapping from biological compounds to proteins. SolCyc contains in its latest version 1613 biochemical reactions all identified in tomato. However, the latest version of SolCyc only has 18 cross-references to ChEBI and two to UniProt, revealing another problem in some resources, the lack of cross-references. Another problem of this resource is the cross-references used on proteins. SolCyc uses references to SGN Unigene as cross-references for proteins. Meaning that to obtain more information about the protein, one has to retrieve information about the unigene, find its equivalent on the tomato genome annotation and find the corresponding proteins related to this gene.

The spread of the information across multiple resources is likely correlated with the incompleteness of some of these resources, the more spread the resources are, the more spread the data is and therefore the higher the risk of incompleteness. To solve this problem, the plant community needs to agree on a central database to store its information in. The community should work on adding the metabolites into ChEBI, adding the biochemical reactions into Rhea or a pathway database with the appropriate cross-references to ChEBI for the metabolites and UniProt for the enzymes. While this

centralization of the information is clearly needed, the community should work on improving existing solutions so that they become a central entity rather than trying to develop yet another one which would spread the information a little more.

Conclusions

Annotex performs data integration relying on semantic web technology and the cross-reference across the resources integrated. Simply providing an input and the desired type of information to retrieve, one can browse the network of information that is built around a genome annotation while not being limited to genome annotations. Annotex also allows finding genes that might be of interest and that might not have been picked up by simply looking at their gene description. Finally, Annotex, by providing the possibility to export its output, allows approaching a question from different angles, reducing the amount of data to handle by filtering out part of it using known information. The examples have shown that Annotex can already be used to retrieve useful information; however, the resources it is using could clearly be improved, either by integrating new, complementary resources or by working with the resource providers to include more plant related information. The plant community needs to work together to build central places of information across the plant genomes.

Acknowledgements

Funding: Wageningen UR Plant Breeding and the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

Chapter 7: General discussion

Biology is facing an enormous data explosion at different levels because of the different high throughput technologies commonly known under the name ~omics technologies. Metabolite analysis using GCMS (James and Martin 1952) and LCMS (Arpino 1989) have been improved (Hsieh and Korfmacher 2006) and nowadays return few hundreds clustered peaks (Khan, Chibon et al. 2012) corresponding to as many metabolites detected. Micro-array technologies have evolved from few hundred to several hundred thousands of genes analyzed in a single experiment (Augenlicht and Kobrin 1982; Augenlicht, Wahrman et al. 1987). The latest generation of Sanger sequencers are able to produce up to 96 sequences of 400 to 900 base pairs (~85.000 base pairs per run) while new sequencing technologies such as 454 are able to generate five hundred millions bases in a few hours (Pettersson, Lundeberg et al. 2009) allowing to sequence a complete genome in a single run (Margulies, Egholm et al. 2005). Plant phenotyping is also changing. Historically, phenotyping has been done by breeders who are trained and gained experience in targeting traits of interest for the breeding goals. Nowadays, biotechnologies are rising to provide automatic phenotyping platforms (Iyer-Pascuzzi, Symonova et al. 2010). Sozzani and Benfey (2011) provide a review of different phenotyping platforms able to record different phenotypes (i.e.: leaves, roots or height) (PHENOPSIS (Granier, Aguirrezabal et al. 2006), PHENODYN (Sadok, Naudin et al. 2007) or GERMINATOR (Joosen, Kodde et al. 2010)) and software (LeafAnalyser (Weight, Parnham et al. 2008) or RootTrace (French, Ubeda-Tomas et al. 2009)) to analyze them. These technologies are already available to biologists and breeders and are still being improved, potentially increasing the number of observations returned. Being able to handle this tsunami of data is becoming a pressing challenge where bioinformatics plays a key role. In parallel to the increase of data generated in a single experiment, the number of resources providing information about genes, proteins, metabolites or pathways for one or multiple species is also increasing (Chapter 1 Figure 2). The amount and spreading of the information available leave biologists and breeders with the increasing challenge of finding and integrating information for their research. Genomes are being sequenced and annotated. When investigating a QTL interval the list of genes present in this interval can be retrieved and the genome annotation provides for each of these genes relations to proteins, GO terms and/or protein domains, each available from different resources.

This thesis is focused on the development of three bioinformatics tools to assist scientists and breeders in the interpretation of their research data, namely MQ², Marker2sequence and Annotex (Figure 1). Large-scale QTL mapping analysis where few hundred metabolites were measured can be analyzed in a single experiment. This leads to the discovery of a specific region of the genome related to the expression of a large number of metabolites. Further analysis of the metabolites associated to this region as well as the genes present revealed the presence of a transcription factor binding site regulating the expression of the phenylpropanoid biosynthesis pathway (**Chapter 2**). This experiment showed the need for a good methodology to summarize QTL results when hundreds of traits are analyzed. This has led to the development of MQ² (**Chapter 3**). MQ² provides visualization for large-scale QTL mapping experiments, showing co-localization and eventually QTL hot spots. The second tool developed within this thesis is centered on the filtering of the genes in the QTL interval with the aim to identify candidate genes. Marker2sequence was designed to integrate information (name, description, related GO term and proteins) about each gene present in a specified genomic region (for example, the QTL interval) and offer a way to filter out potential gene(s) of interest based on their annotation (**Chapter 4**). In **Chapter 2**, we concluded that a transcription factor in the QTL interval could influence the expression of our trait(s) of interest (demonstrated by Kloosterman,

Anithakumari et al. (2012)). Marker2sequence (**Chapter 3**) only searches for annotations for putative candidate genes within the QTL interval. In order to include putative candidates, information about, for example, transcription factors and the genes they regulate need to be incorporated. Since transcription factors, and their binding sites, are poorly annotated in tomato, we combined the genomic structure of six selected genotypes from an Introgression Line population with gene expression information and found 17 potential transcription factor binding sites in tomato (**Chapter 5**). If confirmed and linked to transcription factors this would be the beginning of the construction of the gene regulatory network of tomato which could then be integrated into Marker2sequence allowing it to include genes outside of the QTL interval (trans) but regulated by genes (i.e.: transcription factor) within the QTL interval (cis). However, Marker2sequence is focused on the provided QTL interval in relation with a phenotype. When breeding for a specific phenotype or investigating a biological process several genes may be involved and knowing which genes provide additional information on the genomic regions that are involved in this process is crucial. This realization led to the development of a third tool, namely Annotex (**Chapter 6**). Annotex allows finding genes related to a certain pathway, GO term, protein or metabolite using cross-references from the different resources it integrates. These cross-references can indicate either that two elements in two different resources with two different identifiers are, in fact, the same biological entity or are related (at a biological level).

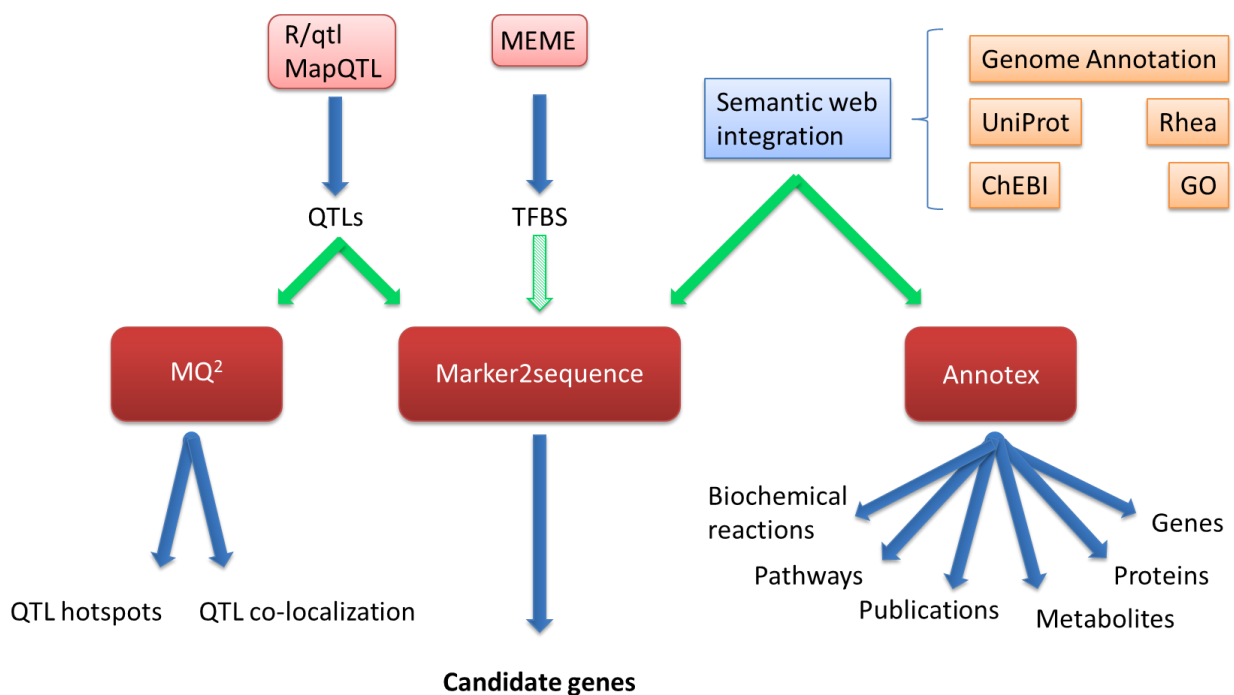


Figure 1: Overview of the bioinformatics tools developed in this thesis (dark red boxes). The developed tools rely on the input provided by the user (green arrows), which were results of different analysis tools (light red boxes). The blue arrows present the possible outcome of the tools. The spotted green arrow represents information which could be and should be integrated into Marker2sequence but is not at this point. The orange boxes are different resources integrated using the semantic web technologies and used by Marker2sequence and Annotex.

Visualizing High-throughput QTL mapping with MQ²

QTL mapping on the output of micro-array data can return thousands of eQTL (expression QTL) (Kloosterman, Anithakumari et al. 2012). QTL mapping on the output of metabolite expression

analysis (such as LCMS) can return hundreds of mQTL (metabolite QTL) (Khan, Chibon et al. 2012). Current QTL mapping software can perform such large scale QTL mapping analysis, however, none of them provides a summary of the distribution of the identified QTL. With R/qtl (Broman, Wu et al. 2003) some knowledge of R programming is required to generate such a figure. QTL cartographer (Basten, Weir et al. 2004) requires some knowledge of gnuplot (<http://www.gnuplot.info/>) as well as some data formatting. MapQTL provides visualization of the output but only for one trait at a time. With none of these tools, there is an easy way to visualize the co-localization of the QTL.

MQ² was developed to fill in this need. It aims at parsing the output from MapQTL, R/qtl and QTL cartographer and presents a visual overview of the distribution of the QTL along the genetic map. Using MQ², biologists and breeders can easily identify co-localizing QTL and identify potential QTL hot-spots. This is important as these QTL hot-spots may reveal the presence of regulatory elements such as transcription factors (Khan, Chibon et al. 2012; Kloosterman, Anithakumari et al. 2012).

In **Chapter 2**, an LCMS analysis was described on peel and flesh samples of individuals from an apple population. The metabolite expression levels have then been correlated with marker scores measured on this population leading to the detection of mQTL for peel and flesh separately. The mQTL found provide an insight in the genetic basis for the diversity of metabolite expressions. Apple is a source for various health beneficial compounds (Khan, Chibon et al. 2012) and understanding the genetic basis of the expression of these compounds can help breeding for these compounds and to develop new, healthier apple varieties. In this analysis, 669 mQTL were identified, 488 in peel and 181 in flesh respectively. This first QTL mapping analysis was performed with MetaNetwork (Fu, Swertz et al. 2007). MetaNetwork is an R library designed to make high-throughput QTL mapping easy and provide a correlation network of the traits analyzed. However, when using it on our data set we faced some problems: MetaNetwork had to be adjusted to work with our datasets containing zeros and missing values; as only a compiled R package had been released, we spent some time extracting the source code and rebuild the package on a newer version of R to use it (since the sources have been made available on github). Finally, we have not been able to use the correlation network as this part of the analysis never worked due to the size of our dataset. Overall, MetaNetwork did cost more time than MapQTL would have, which has been used for the fine mapping of a subset of the traits. Running the few hundreds metabolites in MapQTL would have taken a few minutes, gathering the information by hand to obtain a representation of the co-localization of the QTL was the tedious part which is now solved by MQ².

Next steps for MQ²

MQ² strong point is in providing a distribution of the QTL along the genetic map. However, this visualization could be enhanced, for example, by including the significance of the identified QTL. In such a visualization, categories showing how many QTL have a LOD value between the LOD threshold provided and the LOD threshold plus two, or plus four or above LOD threshold plus five, similarly to the Figure 2 of **Chapter 2**. Categorizing the QTL per significance will help finding positions with highly significant QTL. A concentration of highly significant QTL may reveal the presence of an important genomic region containing genes strongly correlated with the expression of the traits.

MQ² fills a need for the visualization of large-scale QTL mapping results. It helps identification of QTL co-localization and finding potential QTL hot-spots which can lead to the discovery of regulatory elements. However, if MQ² currently fills a need, QTL mapping tools should include such visualization

by default, the next version of MapQTL should include facilities to help large-scale QTL mapping. R/qtl should provide facilities to analyze all traits of a dataset and visualize the results.

Marker2sequence: From QTL to candidate genes

Once a genomic region has been linked to a phenotype via the presence of a QTL, the next question might be which gene putatively affects the trait under investigation. In daily practice, breeders introgress regions containing QTL into elite breeding germplasm. A larger QTL interval means a higher number of genes in which the one (or more) gene(s) which are responsible for the expression of the phenotype are “hidden”. For a breeder, a larger region to introgress also means a higher risk of linkage drag (introduction of genes negatively influencing the desired phenotype). Fine mapping approaches, in which the size of the mapping population is increased to increase the recombination and reduce the QTL interval is often used (Yang, Zhang et al. 2012). Increasing the mapping population is, however, a tedious task. By relying on known information about the QTL interval Marker2sequence aids in the identification of gene(s) potentially related to the trait of interest. For a given genomic region, Marker2sequence extracts the list of all genes in the region relying on the genome annotation and aggregates for all these genes the information known about them using the cross-references present in the genome annotation.

Genome exploration tools

Marker2sequence is not the only tool to explore a specific region of the genome. GBrowse (Stein, Mungall et al. 2002; Donlin 2009) is a visualization tool for genomic data. It provides a way to browse genome annotations, localize them and compare them to each other and eventually place additional elements such as genetic markers or gene expression information for the genes on these sequences. GBrowse allows zooming in or out of a region, searching the genome information for keywords. However, this search is limited to the element (gene, markers) name and description. The advantage of GBrowse is that it provides a much better visualization of a genome region than Marker2sequence. The added feature of Marker2sequence is that it allows downloading this list of genes, accessing additional information for each of these genes and filtering out genes of interest by searching through the cross-references of their annotation, from the databases with which it is integrated. This expanded search of Marker2sequence allows searching a genomic region for genes related to a pathway or a protein or a GO term. Marker2sequence will filter out genes if the keyword searched is present in their name or related GO term, protein, pathway or publications. This search is more appropriate when searching a list of genes for genes involved in a specific phenotype.

Ondex (Köhler, Baumbach et al. 2006) is another tool for data integration and visualization, it can integrate data from different resources (such as UniProt, GO, KEGG, SGD) but offers a graph-based visualization which clutters when confronted with large datasets. In addition, in the case of QTL intervals, the graph is composed of a multitude of sub-graphs centered on a gene with sometimes links between the sub-graphs when genes share a common annotation (i.e.: GO term or protein). This type of visualization is not ideal for a QTL region. If a large part of the genes in the set are related, this visualization would provide information on the main process, GO and proteins shared between these genes, but in a QTL interval, out of hundreds or thousands of genes, only few are related to the trait of interest and might thus be related in their annotation. Ondex is therefore not suitable as a visualization and exploration tool for QTL. Marker2sequence provides only a visualization of the alignment of the genetic map with the genome sequence if the input provided

relies on a genetic map. The network of information retrieved for each gene in the region is not presented. The list of genes present in the genome region is given in a table. For each gene, a dedicated page presents information about the gene: description, location, chromosome, the associated GO term and proteins with their name and linked to the UniProt page, the pathways of these proteins, linked to their UniPathway page and the publications of these proteins, linked to PubMed. This information is provided as text which can be downloaded and has no problem of scaling when the amount of information increases.

Gene filtering via data integration

Marker2sequence provides a user-friendly interface to dive into a specific genomic region. It retrieves the list of all genes in this region and allows searching through this list, which in the case of a QTL can include hundreds to thousands genes, using a keyword-based search system. This search system filters out genes containing the exact keyword provided. The search is case insensitive but is an exact-match search, meaning a search for “caroten” might return more results than a search for “carotenoid” as the first search will match genes related to the “carotenoid biosynthesis” pathways as well as genes related to the GO term GO:0006629 “lipid metabolic process” which contains “carotenes” in its description. This is a limitation of the current search engine and should be considered when filtering the gene list with Marker2sequence. To improve the search, a future version will want to look into stemming. Stemming is the process of reducing the query word to its root (Porter October, 2001), for example: “testing” would become “test” and “carotene” would become “caroten”. By reducing the length of the word, we reduce its specificity and eventually increase the number of potential matches of the keyword.

Marker2sequence should be able to filter out, from a list, genes related to a trait. If the trait is a biological element such as a metabolite (ie: beta-carotene) or a protein, it will likely be sufficiently described in the genome annotations or in the elements integrated by the cross-references of the genome annotation that Marker2sequence can find and return genes whose annotation mention it. More abstract traits, such as “Plant height” or “yield” or “fruit color”, are not mentioned in the annotation (proteins, GO terms, pathway), and thus cannot be found by Marker2sequence. This is one of the biggest shortcomings of the current search function which restrict Marker2sequence to be a filtering tool using keyword search rather than a predictive tool. In order to be able to search for genes related to “Plant height” one has to have previous knowledge of the processes involved and should know, a-priori, what the possible annotations of genes involved in such process are. For example, potato flesh color is related to the expression of the carotenoid pathway (Brown, Edwards et al. 1993). If someone investigates a QTL for flesh color in Marker2sequence and searches the QTL interval for “flesh color” there will not be any gene returned, however, if the same interval is searched for “carotenoid” the genes related to the “carotenoid biosynthesis” pathway will be returned and these genes are involved in flesh color. Being able to search the gene list for genes related to a trait is one of the major up-coming tasks of Marker2sequence. The first step to achieve this is to develop ontologies precisely defining the traits and that the whole community agrees upon.

Linking traits to genes and vice-versa

Projects, such as cropontology (<http://www.cropontology.org>) (Shrestha, Matteis et al. 2012), Solanaceae Phenotype (SP) ontology (Menda, Buels et al. 2008) and the plant ontology consortium (The Plant Ontology Consortium 2002) aim at providing crop-related ontologies to the breeding community. Platforms such as the OBO Foundry (Smith, Ashburner et al. 2007) provide a central

place where the community can gather, discuss and reach a consensus on the best way to build, extend or complete existing ontologies. Currently, cropontology provides 29 ontologies including 22 ontologies related to phenotypes and traits. As of June 25th 2013, these 29 ontologies contain a total of 11,237 terms of which 2,690 terms (23%) have at least one cross-reference. The front page of the project lists the 18 contributors that have created the 29 ontologies, but in total 77 persons have registered to create or update the ontologies. Cropontology is a unique project dedicated to plant science, providing a platform for the community to build, maintain and expand their ontologies. Consortia such as the SUNRISE consortium (<http://www.sunrise-project.fr/>) or Virtual Lab of Plant Breeding (<http://www.vlpb.nl/>) are already working on integrating these ontologies in their data management systems giving to their data a clear and defined semantic and allowing better communication and better integration of the resources between the partners of the consortium.

Once the ontologies have been built, the second step will be to integrate them with current genome annotation, in the same way that the gene ontology is currently integrated. The integration of these ontologies into genome annotation is a key element as it will allow making the ontologies more broadly known which will generate discussions and debates within the community which in return will lead to improved ontologies. The integration into the existing genome annotation should happen in two steps. The first step is to manually annotate genes known to be related to given traits. For example “beta-carotene hydroxylase” (bch) is known to be involved in flesh color (Brown, Kim et al. 2006), and has been mapped on the Y locus (Thorup, Tanyolac et al. 2000; Wolters, Uitdewilligen et al. 2010) on chromosome 3 (Bonierbale, Plaisted et al. 1988). This gene could therefore be linked to the trait: SP:0000188 “tuber flesh color” from the *Solanaceae* phenotype ontology. By expanding the genome annotations, the value of these ontologies will grow, and researchers will become enthusiastic to start using and improving these ontologies. In a second step, this process could be automated. Using scientific literature as source of information the relationships between, for example, phenotypes and genes, could be extracted by standard text-mining tools (Korbel, Doerks et al. 2005). Nanopublications (Mons, van Haagen et al. 2011) can be an approach to represent these relationships. Nanopublications correspond to small assertions published as such, in form of triples (subject verb complement) containing the assertion itself (gene X is regulated by the transcription factor Y) and provenance information (typically the article from which these assertions are extracted) and supporting information (would provide some general information on when or where this assertion was found, for example, in which organism). These assertions can be anything and thus could include assertions about the relation between a gene and a trait. Nanopublications are based on semantic web technologies, meaning that in order to make the assertion: “gene bch is related to flesh color”, the gene bch should be defined with a unique identifier, for example gene identifier from the genome annotation. The relation “is related to” should also be defined in an ontology, “skos:related” (<http://www.w3.org/2009/08/skos-reference/skos.html#related>) is a generic term that could be used as predicate in this case, a more specific term could also be used but it will have to be defined in an ontology and will vary from case to case depending on the relation between the gene and the trait. Manual annotation of the genome by linking genes with terms from different ontologies can be turned into a nanopublication, which then becomes a publication, reference-able, and thus provide credits to the person that created the statement. The resulting nanopublication would indicate that bch is related to tuber flesh color, provide the source and the supporting information and publish it as such. Eventually, this should lead to resolving cross-references between clearly defined biological trait ontology concepts, such as “plant height” or “fruit sugar content” with

entities, such as genes or metabolites. An added value of a well-defined ontology is that it could then be re-used by tools such as Marker2sequence to find the genes related to a specified trait of the trait ontology. The current free-text search functionality can then be upgraded into an ontology based search functionality. Nanopublications can then be cited, incentivizing researcher to participate in the process. An application could be built to provide an infrastructure to manually annotate gene with terms from ontologies. This application could be including gamification where users earn points by annotating the genome with ontology terms and the top participants are then displayed in the leaderboard. This leads to competitions between the participants who play the game, the leaderboard could also be set to display the top 10 institutes assisting in improving the annotations. With such a system, users provide for themselves a number of citable assertions and enhance genome annotations. This idea of gamification is described for genome annotations but could work just as well for any other entities (i.e.: proteins, pathways), however a tight integration with the data providers would be needed in order to avoid further spreading of the information.

Crowd-sourcing is the idea of outsourcing a task to a crowd of people, Wikipedia is the classic example where anyone can become involved in building an encyclopedia. Dai, Tian et al. (2013) discussed that the limitations to crowd-sourcing in biology is the lack of participation from the community. They have developed an extension for MediaWiki to reward the participation of the members of the community. Our approach, by using gamification, rewards each contributor and institute for their involvement and by using nanopublications, lets each participant create nanopublications which are associated with them and that are citable to by their peer, just like any scientific publication.

Both the nanopublications project and the cropontology project are community based projects supported by pharmacology consortia. One of the challenges is to gather a community of contributors, developing the resource. The Gene Ontology project has been able to develop a community by developing a core group of members that fund staff to work on GO. This core group serves as mentor to the associates which can contribute to the ontology via a member of the core group. This model, although not being truly following the open-source model, has the benefit of motivating institutes to join the core group and thus finance staff to work on the project as it will help the institute to steer the project toward areas of its own interest. For example, if such a model was applied to the cropontology project, it would become a consortium with core members and associates, and then being a core member would allow steering the development of ontologies toward areas of interest, for example improving the *Solanaceae* traits ontology rather than the Musa anatomy ontology. As associate, you may propose improvement to the Musa anatomy ontology but as it will have to go through a core member it might be harder to have your changes accepted. An alternative is to adopt a model closer to the open-source model where one group, institute or expert is in charge of one or several ontologies and anyone can contribute to them to improve them. Changes are proposed to the person in charge which will accept or decline them after discussion. This model has the advantage of being simpler; however, it is harder to secure funding for the person in charge of the ontology. Institutes would contribute to the ontology until it satisfies their needs and securing funding afterwards is more problematic.

The semantic web in Marker2sequence

Marker2sequence relies on semantic web technologies to integrate the data previously aggregated into a single triple store (semantic web database). The advantage of semantic-web technologies is

that they are designed for data integration. Each concept conveys a clear semantic defined in a specific ontology. However, in the current implementation of Marker2sequence, it evades one of the main objectives of the semantic-web, namely: on the fly data integration. There are multiple reasons for this: The first reason is that when Marker2sequence was developed, the SPARQL 1.1 specifications, allowing querying multiple triple store in a single query, had not been published and were therefore not implemented in the different programming libraries and triple store. The SPARQL 1.0 specifications allow querying the different triple-store separately, integrate the information in a graph and re-query it to extract the relevant parts. This approach is too time consuming for a web-based tool and was therefore not implemented. Another reason to build our own triple store was that at the time of the implementation of Marker2sequence, no SPARQL endpoint was available to query the UniProt database, however, UNIPROT was available in the semantic web compatible format. Nowadays this SPARQL end-point for UniProt exists, the SPARQL 1.1 specifications have been released and libraries and triple-store have adjusted or are adjusting to them. Marker2sequence could therefore be adjusted to retrieve its information from UniProt directly rather than using the local graph, allowing it to access the most up to date information. However, the cost in time to perform a SPARQL query against a remote SPARQL endpoint remains high and may vary according to the load of the server which is a drawback for a web application. On the other hand, Marker2sequence relies directly on the ontologies from UniProt for its data integration, thus, switching to the UniProt SPARQL endpoint to retrieve part of our information would require a rewrite of the SPARQL queries but not of the whole application.

The gene regulatory network expands the gene list from a QTL analysis

The gene regulatory network is the network of interactions at the gene level which regulates positively or negatively the expression of genes. In **Chapter 4**, gene expression information has been combined with the genomic structure of introgression line (IL) genotypes and information from the genome annotation to predict potential transcription factor binding sites (TFBS). 17 potential TFBS have been predicted. These TFBS need to be validated in the lab. The *S. chmielewskii* population used in that experiment has also been analyzed for transcription factor expression using RT-PCR. The expression of the transcription factor could be integrated provided a list of transcription factors potentially regulating the transcription factor binding could be found (**Chapter 4**). Correlating the transcription factors to their target would allow building the gene regulatory network of tomato. This information could then be integrated into Marker2sequence allowing it to expand the list of potential candidate genes to the genes outside the QTL interval but regulated by genes (transcription factor) within the QTL interval, thus enabling Marker2sequence to consider regulatory genes such as transcription factors as candidate genes. Protein-protein interaction (PPI) is built-in in Marker2sequence using data from the IntAct database (Kerrien, Aranda et al. 2012). However, we have not activated the PPI due to speed issues for a web application. This speed issue could be circumvented by leaving the option to the user to include the PPI information to his search (although limited by the time out of the server), or having the possibility to make Marker2sequence asynchronous by letting the user provide an email to which the results would be sent once the analysis finished. With PPI support activated, Marker2sequence will include into the list of genes within a QTL region, genes that are outside the region but interact with genes inside the region at the protein level. Integrating regulatory network information into Marker2sequence has thus been started. PPI network is present but needs some optimization before being activated and gene

regulatory network information have been started but needs to be validated in the lab and expanded as **Chapter 4** only covers a small set of transcription factors.

Improving Marker2sequence

For a future better performance of Marker2sequence several new features should be included: (1) improve the keyword search by including a stemming library (while still give the option to not use it); (2) work on annotating genes, proteins and pathways with terms from the trait ontologies, then add next to the keyword search the possibility to find genes from a trait defined in a trait ontology; (3) as seen in **Chapter 2**, regulatory elements can influence greatly the output of a QTL mapping experiment, it is therefore important to include regulatory network information in Marker2sequence, enabling it to return genes regulating biological elements involved in the expression of the trait of interest. Finally, (4) Marker2sequence has been designed as a web application to be integrated into the BreeDB framework. As its filtering algorithm becomes more complex it might be worthwhile investigating turning it into a desktop application or an asynchronous web-service (which would allow keeping some integration with BreeDB). This would also allow making use of SPARQL 1.1 and querying directly, for example, the UniProt SPARQL. This will also ensure that we are always using the latest available information.

Genome-wide filtering: Annotex

Marker2sequence offers a way to filter genes from a specific gene set using their annotation. However, this gene set can only be as long as one chromosome. This allows identification of genes related to a trait in a specific chromosome; however, Marker2sequence cannot answer such question on a whole-genome scale. Annotex offers this functionality by directly allowing searching the entire genome annotation for genes related to a specific entity (GO term, protein, pathway, and metabolite). Annotex integrates UniProt, GO, ChEBI, Rhea and the genome annotations of tomato and potato. From our tests we concluded that if ChEBI contains a number of metabolites present in plants, Rhea lacks the chemical reactions in which there are involved. The cross-references between metabolites and proteins in Annotex are affected by which database is used. Annotex queries the UniPathway database (Morgat, Coissac et al. 2012), as this database is made available in a semantic web format by the UniProt consortium, however, this resource contains the annotation of the pathways rather than their chemical reactions and thus cannot be used to map metabolites to pathways. Other pathway resources such as SolCyc (Pujar, Caspi et al. 2009) could be included in Annotex, but they require a good coverage of cross-references with ChEBI and Rhea, which is lacking at the moment (only 18 cross-references to ChEBI and two to UniProt in the last version of LycoCyc). WikiPathways (Pico, Kelder et al. 2008; Kelder, van Iersel et al. 2012) is another resource for pathway information, however, there are currently very few pathways supported for plants. In fact, as of June 2013, only three plant species have pathways in this resource: *Arabidopsis thaliana* (14 pathways), *Oryza sativa* (14 pathways) and *Zea mays* (11 pathways). *Solanum lycopersicum*, *Glycine max*, *Populus trichocarpa* and *Vitis vinifera* are on the list of species, but the pathway information has not yet been made available. Once available, there will be a clear incentive to assess the feasibility to include WikiPathways in Annotex as well. We can conclude that the current pathway resources contain cross-references but vary from species to species. Expanding the cross-references on the existing resources would allow easier integration of these different resources. Thus questions such

as, in which pathway is this compound involved or what are the genes associated to this particular biochemical reaction would become easier to answer.

Improving Annotex

Annotex as a proof-of-concept tool is already easy-to-use. However, to improve Annotex, there are several critical points to address. These points mainly cover the data and the cross-references between the data. My recommendations for improving the data, and therefore Annotex, are: (1) work with the community to improve the coverage of plant metabolites in ChEBI. ChEBI is well known and species a-specific, making it well suited to be expanded with plant metabolites. (2) Either improve the coverage of plant pathways in Rhea or work with the community to find a central place for plant specific pathways and use cross-references to major databases (ChEBI, UniProt) for the compounds involved in the reactions; (3) similarly to Marker2sequence, work in getting genes, proteins, pathway and metabolites annotated with terms from trait ontologies; (4) as regulatory network information is added (being at the gene or protein level), and Annotex includes finding genes involved with a trait via the trait ontology, it might be necessary to reconsider the current web-application approach for either a desktop application or an asynchronous web-service.

Future perspectives

The major bottlenecks that were hampering this research are: (1) the spread of plant resources over multiple databases with no consensus on a central one. The multiplicity of the pathway resources is a clear example of a situation that should be avoided. (2) For Annotex and Marker2sequence the unavailability of information regarding the gene regulatory network of tomato. A number of transcription factors are known, the actions of some of these have been published (Bemer, Karlova et al. 2012) but there is no central place providing this information in a clear and meaningful way. (3) The only ontology that has been widely acknowledged and is widely used is the Gene Ontology. Many resources use it and many publications rely on it, either to improve it (Khodiyar, Hill et al. 2011), improve the annotation of certain genes or proteins using it (Amthauer and Tsatsoulis 2010), build upon it (Renfro, McIntosh et al. 2012) or use it as a component in the analysis of the results of an experiment (Mohammadi, Saraee et al. 2011). However popular the gene ontology is, it does not provide any information about the phenotype influenced by the gene or the protein it annotates. (4) The major databases have built a powerful network of cross-references; however, more specialized resources tend to not be able (because of time and workload) to include their resources in the larger network of information, while it is critical to use these resources for data integration.

As most of these bottlenecks are more generic than just the work presented in this thesis, future research could try to tackle them. This could be achieved in the following ways: (1) The lack of integration of other ontologies in genome annotation could be tackled by the idea presented earlier of a community based effort to annotate manually genes of interest with their corresponding terms in the trait ontology but also eventually the plant ontology and in fact any ontologies available. Integrating nanopublications and game theory, this system would reward each contributor by providing him/her with a number of reference-able assertions and would display the top contributors and institute, creating both a rewarding and challenging environment. (2) As more genes and proteins are annotated with terms from other ontologies manually, we will be able to start building tools to automate this annotation. Proteins involved in the same pathway are probably involved in the expression of the same phenotype. Using similar algorithms as the ones used to automatically

annotate genes with terms from the Gene Ontology, we could annotate genes and proteins with terms from other ontologies. (3) Resources such as SolCyc should be able to be cross-referenced to ChEBI by simply searching for an exact match in the molecule name from SolCyc present in ChEBI. Molecules that are not part of ChEBI could then be added, improving the coverage of plant metabolites in this resource. Similar work could be done for the proteins with the additional filter that the protein of UniProt matching the name should also match the organism (e.g.: tomato or potato as shown in our case). Improving the cross-references in SolCyc would allow easier integration of this resource in the broader network of knowledge.

Conclusions

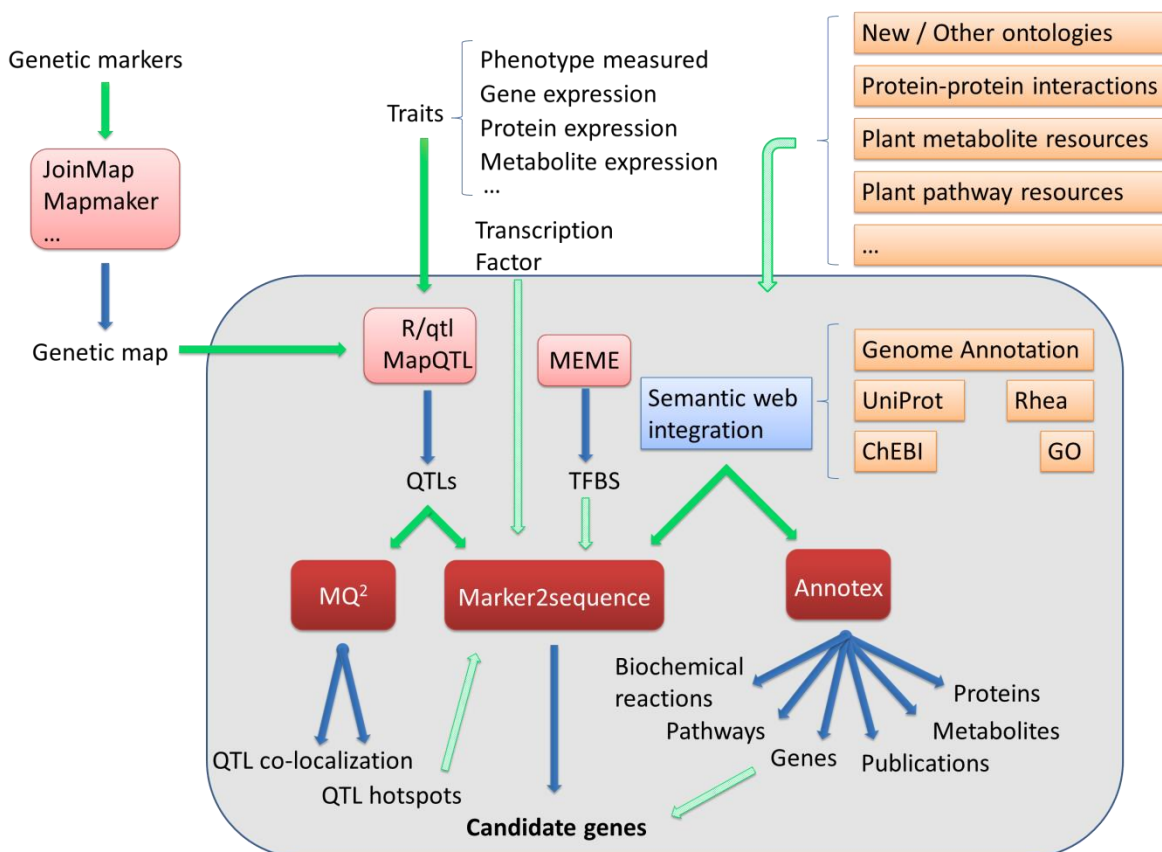


Figure 2: Global representation of the work performed in this thesis (in the grey box) with regards to its environment (outside the box). The dark red boxes are the three bioinformatics tools developed for plant biologists and plant breeders within this thesis. The green arrows represent the input the tools use, while the blue arrows represent the “output” they produce. The light red boxes are examples of tools, not developed within the context of this thesis, which can be used to generate the appropriate inputs. The spotted green arrows represent inputs that are not used at the moment but which should be considered to improve the tools.

This thesis has resulted in the development of three tools (Figure 2) for plant biologists and plant breeders, allowing them to visualize the results of high-throughput QTL mapping analysis, search a specific region of the genome for candidate genes or search the complete genome annotation for genes related to a certain biological element. These three tools offer new possibilities to deal and face the tsunami of data that new biotechnologies generate. The original goal of this thesis was to develop candidate gene prediction tool(s) but we created a tool to filter out potential genes of interest using known information about them. The prediction of candidate genes relies on the integration of more data than just the genome annotation and the proteins. Regulatory mechanisms

have to be taken into account as shown in **Chapter 1**. These mechanisms may be at the transcriptomics, proteomics or metabolomics level, and each has to be considered. The lack of integration of trait ontologies also prevents a direct prediction of genes involved with the trait as there is no direct association between the concept of the trait (“fruit color” or “plant height”) and biological elements (genes, proteins, pathways). These new hypotheses need to be investigated and provide ample starting points for new (PhD) research projects, both from a bioinformatics perspective and from a biological perspective.

References

- AFA, S., H. R., et al. (1996-2010). "Repeat Master Open-3.0." <http://www.repeatmasker.org>.
- Al-Babili, S., P. Huguency, et al. (2000). "Identification of a novel gene coding for neoxanthin synthase from *Solanum tuberosum*." *FEBS Letters* **485**(2-3): 168-172.
- Alcantara, R., K. B. Axelsen, et al. (2012). "Rhea--a manually curated resource of biochemical reactions." *Nucleic Acids Res* **40**(Database issue): D754-760.
- Alder, A., M. Jamil, et al. (2012). "The path from beta-carotene to carlactone, a strigolactone-like plant hormone." *Science* **335**(6074): 1348-1351.
- Allali-Hassani, A., P. W. Pan, et al. (2007). "Structural and chemical profiling of the human cytosolic sulfotransferases." *PLoS Biol* **5**(5): e97.
- Allard, R. W. (1999). *Principles of plant breeding*. New York [etc.], Wiley.
- Almeida, J. R. M., E. D'Amico, et al. (2007). "Characterization of major enzymes and genes involved in flavonoid and proanthocyanidin biosynthesis during fruit development in strawberry (*Fragaria x ananassa*)." *Archives of Biochemistry and Biophysics* **465**(1): 61-71.
- Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." *J Mol Biol* **215**(3): 403-410.
- Ameur, A., V. Yankovski, et al. (2006). "The LCB Data Warehouse." *Bioinformatics* **22**(8): 1024-1026.
- Amthauer, H. A. and C. Tsatsoulis (2010). "Classifying genes to the correct Gene Ontology Slim term in *Saccharomyces cerevisiae* using neighbouring genes with classification learning." *BMC Genomics* **11**: 340.
- Archibald, A. L., N. E. Cockett, et al. (2010). "The sheep genome reference sequence: a work in progress." *Anim Genet* **41**(5): 449-453.
- Arpino, P. (1989). "Combined liquid chromatography mass spectrometry. Part I. Coupling by means of a moving belt interface." *Mass Spectrometry Reviews* **8**(1): 35-55.
- Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* **25**(1): 25-29.
- Augenlicht, L. H. and D. Kobrin (1982). "Cloning and screening of sequences expressed in a mouse colon tumor." *Cancer Res* **42**(3): 1088-1093.
- Augenlicht, L. H., M. Z. Wahrman, et al. (1987). "Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro." *Cancer Res* **47**(22): 6017-6021.
- Awad, M. A., A. De Jager, et al. (2000). "Flavonoid and chlorogenic acid levels in apple fruit: Characterisation of variation." *Scientia Horticulturae* **83**(3-4): 249-263.
- Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.
- Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." *Nucleic Acids Res* **34**(Web Server issue): W369-373.
- Baldwin, E. A., J. W. Scott, et al. (1998). "Relationship between Sensory and Instrumental Analysis for Tomato Flavor." *J Am Soc Hortic Sci* **123**(5): 906-915.
- Barbazuk, W. B., S. J. Emrich, et al. (2007). "SNP discovery via 454 transcriptome sequencing." *Plant J* **51**(5): 910-918.
- Barrasa, M. I., P. Vaglio, et al. (2007). "EDGEdb: a transcription factor-DNA interaction database for the analysis of *C. elegans* differential gene expression." *BMC Genomics* **8**: 21.
- Basten, C. J., B. S. Weir, et al. (2004). "QTL cartographer, version 1.17." *Department of Statistics, North Carolina State University, Raleigh, NC*.
- Bemer, M., R. Karlova, et al. (2012). "The tomato FRUITFULL homologs TDR4/FUL1 and MBP7/FUL2 regulate ethylene-independent aspects of fruit ripening." *Plant Cell* **24**(11): 4437-4451.
- Bennett, S. (2004). "Solexa Ltd." *Pharmacogenomics* **5**(4): 433-438.
- Berners-Lee, T., J. Hendler, et al. (2001). "The Semantic Web: Scientific American." *Scientific American* **284**: 34-43.
- Bevan, M. W., R. B. Flavell, et al. (1983). "A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation." *Nature* **304**(5922): 184-187.
- Bhagat, J., F. Tanoh, et al. (2010). "BioCatalogue: a universal catalogue of web services for the life sciences." *Nucleic Acids Res* **38**(Web Server issue): W689-694.

- Bogs, J., M. O. Downey, et al. (2005). "Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves." *Plant Physiology* **139**(2): 652-663.
- Bogs, J., F. W. Jaffé, et al. (2007). "The grapevine transcription factor *VvMYBPA1* regulates proanthocyanidin synthesis during fruit development." *Plant Physiology* **143**(3): 1347-1361.
- Bolser, D. M., P.-Y. Chibon, et al. (2012). "MetaBase—the wiki-database of biological databases." *Nucleic Acids Res* **40**(D1): D1250-D1254.
- Bonierbale, M. W., R. L. Plaisted, et al. (1988). "RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato." *Genetics* **120**(4): 1095-1103.
- Bonk, M., B. Hoffmann, et al. (1997). "Chloroplast import of four carotenoid biosynthetic enzymes in vitro reveals differential fates prior to membrane binding and oligomeric assembly." *Eur J Biochem* **247**(3): 942-950.
- Brem, R. B., G. Yvert, et al. (2002). "Genetic dissection of transcriptional regulation in budding yeast." *Science* **296**(5568): 752-755.
- Broekstra, J., A. Kampman, et al. (2002). *Sesame: A Generic Architecture for Storing and Querying {RDF} and {RDF Schema}*. Proceedings of the first Int'l Semantic Web Conference (ISWC 2002), Sardinia, Italy, Springer Verlag.
- Broman, K. W., H. Wu, et al. (2003). "R/qtl: QTL mapping in experimental crosses." *Bioinformatics* **19**(7): 889-890.
- Brown, C. R., C. G. Edwards, et al. (1993). "Orange Flesh Trait in Potato - Inheritance and Carotenoid Content." *J Am Soc Hortic Sci* **118**(1): 145-150.
- Brown, C. R., T. S. Kim, et al. (2006). "Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism." *Am J Potato Res* **83**(5): 365-372.
- Brueggemann, J., B. Weisshaar, et al. (2010). "A WD40-repeat gene from *Malus × domestica* is a functional homologue of *Arabidopsis thaliana* *TRANSPARENT TESTA GLABRA1*." *Plant Cell Reports* **29**(3): 285-294.
- Brugmans, B., R. G. van der Hulst, et al. (2003). "A new and versatile method for the successful conversion of AFLP markers into simple single locus markers." *Nucleic Acids Res* **31**(10): e55.
- Burr, B., F. A. Burr, et al. (1988). "Gene mapping with recombinant inbreds in maize." *Genetics* **118**(3): 519-526.
- Buske, F. A., M. Boden, et al. (2010). "Assigning roles to DNA regulatory motifs using comparative genomics." *Bioinformatics* **26**(7): 860-866.
- Calenge, F. and C. E. Durel (2006). "Both stable and unstable QTLs for resistance to powdery mildew are detected in apple after four years of field assessments." *Molecular Breeding* **17**(4): 329-339.
- Calenge, F., A. Faure, et al. (2004). "Quantitative Trait Loci (QTL) analysis reveals both broad-spectrum and isolate-specific QTL for scab resistance in an apple progeny challenged with eight isolates of *Venturia inaequalis*." *Phytopathology* **94**(4): 370-379.
- Camacho, C. and T. Madden. (March 2nd, 2011, July 15th, 2011). "SOAP-based BLAST Web Service." Retrieved April 8th, 2013, from <http://www.ncbi.nlm.nih.gov/books/NBK55699/>.
- Carrari, F. and A. R. Fernie (2006). "Metabolic regulation underlying tomato fruit development." *Journal of Experimental Botany* **57**(9): 1883-1897.
- Carroll, J. J., I. Dickinson, et al. (2004). Jena: implementing the semantic web recommendations. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. New York, NY, USA, ACM. **10**: 74-83.
- Caspi, R., T. Altman, et al. (2010). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic Acids Res* **38**(Database issue): D473-479.
- Celton, J. M., D. S. Tustin, et al. (2009). "Construction of a dense genetic linkage map for apple rootstocks using SSRs developed from *Malus* ESTs and *Pyrus* genomic sequences." *Tree Genetics and Genomes* **5**(1): 93-107.

- Chen, K. and N. Rajewsky (2007). "The evolution of gene regulation by transcription factors and microRNAs." *Nat Rev Genet* **8**(2): 93-103.
- Chibon, P.-Y., H. Schoof, et al. (2012). "Marker2sequence, mine your QTL regions for candidate genes." *Bioinformatics* **28**(14): 1921-1922.
- Chicken genome consortium (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." *Nature* **432**(7018): 695-716.
- Dai, L., M. Tian, et al. (2013). "AuthorReward: increasing community curation in biological knowledge wikis through automated authorship quantification." *Bioinformatics* **29**(14): 1837-1839.
- Darvasi, A. (1998). "Experimental strategies for the genetic dissection of complex traits in animal models." *Nat Genet* **18**(1): 19-24.
- Davey, M. W., K. Kenis, et al. (2006). "Genetic control of fruit vitamin C contents." *Plant Physiology* **142**(1): 343-351.
- De Roure, D., C. Goble, et al. (2009). "The design and realisation of the Virtual Research Environment for social sharing of workflows." *Future Gener Comp Sy* **25**(5): 561-567.
- De Vos, R. C. H., S. Moco, et al. (2007). "Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry." *Nature Protocols* **2**(4): 778-791.
- Deluc, L., J. Bogs, et al. (2008). "The transcription factor *VvMYB5b* contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries." *Plant Physiology* **147**(4): 2041-2053.
- Dixon, R. A. and C. L. Steele (1999). "Flavonoids and isoflavonoids - A gold mine for metabolic engineering." *Trends in Plant Science* **4**(10): 394-400.
- Dixon, R. A. and D. Strack (2003). "Phytochemistry meets genome analysis, and beyond." *Phytochemistry* **62**(6): 815-816.
- Do, P. T., M. Prudent, et al. (2010). "The influence of fruit load on the tomato pericarp metabolome in a *Solanum chmielewskii* introgression line population." *Plant Physiol* **154**(3): 1128-1142.
- Donlin, M. J. (2009). "Using the Generic Genome Browser (GBrowse)." *Curr Protoc Bioinformatics* **Chapter 9**: Unit 9 9.
- Dunemann, F., D. Ulrich, et al. (2009). "QTL mapping of aroma compounds analysed by headspace solid-phase microextraction gas chromatography in the apple progeny 'Discovery' x 'Prima'." *Molecular Breeding* **23**(3): 501-521.
- Dunnett, C. W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." *J Am Stat Assoc* **50**(272): 1096-1121.
- Eathington, S. R., T. M. Crosbie, et al. (2007). "Molecular Markers in a Commercial Breeding Program." *Crop Sci* **47**(Supplement_3): S-154-S-163.
- Eberhardt, M. V., C. Y. Lee, et al. (2000). "Antioxidant activity of fresh apples." *Nature* **405**(6789): 903-904.
- Edwards, M. D., C. W. Stuber, et al. (1987). "Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action." *Genetics* **116**(1): 113-125.
- Egan, A. N., J. Schlueter, et al. (2012). "Applications of next-generation sequencing in plant biology." *Am J Bot* **99**(2): 175-185.
- Erling, O. and I. Mikhailov (2007). *RDF Support in the Virtuoso DBMS*. Proceedings of the 1st Conference on Social Semantic Web CSSW.
- Esposito, N., O. G. Ovchinnikova, et al. (2008). "Host and Non-Host Plant Response to Bacterial Wilt in Potato: Role of the Lipopolysaccharide Isolated from *Ralstonia solanacearum* and Molecular Analysis of Plant-Pathogen Interaction." *Chem Biodivers* **5**(12): 2662-2675.
- Essaghir, A., F. Toffalini, et al. (2010). "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data." *Nucleic Acids Res* **38**(11): e120.
- Femia, A. P., G. Caderni, et al. (2003). "Effect of diets fortified with tomatoes or onions with variable quercetin-glycoside content on azoxymethane-induced aberrant crypt foci in the colon of rats." *Eur J Nutr* **42**(6): 346-352.

- Fernandez-Suarez, X. M. and M. Y. Galperin (2013). "The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection." Nucleic Acids Res **41**(Database issue): D1-7.
- Fernández-Suárez, X. M. and M. Y. Galperin (2013). "The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection." Nucleic Acids Res **41**(D1): D1-D7.
- Fielding, R. T. and R. N. Taylor (2002). "Principled design of the modern Web architecture." ACM Transactions on Internet Technology **2**(2): 115-150.
- Fischer, T. C., C. Gosch, et al. (2007). "Flavonoid genes of pear (*Pyrus communis*)." Trees - Structure and Function **21**(5): 521-529.
- Fisher, P., C. Hedeler, et al. (2007). "A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis." Nucleic Acids Res **35**(16): 5625-5633.
- Flicek, P., I. Ahmed, et al. (2013). "Ensembl 2013." Nucleic Acids Res **41**(D1): D48-D55.
- Fraley, R. T., S. G. Rogers, et al. (1983). "Expression of bacterial genes in plant cells." Proc Natl Acad Sci U S A **80**(15): 4803-4807.
- Freimuth, R. R., R. B. Raftogianis, et al. (2000). "Human sulfotransferases SULT1C1 and SULT1C2: cDNA characterization, gene cloning, and chromosomal localization." Genomics **65**(2): 157-165.
- French, A., S. Ubeda-Tomas, et al. (2009). "High-throughput quantification of root growth using a novel image-analysis tool." Plant Physiology **150**(4): 1784-1795.
- Fu, J., J. J. B. Keurentjes, et al. (2009). "System-wide molecular evidence for phenotypic buffering in Arabidopsis." Nature Genetics **41**(2): 166-167.
- Fu, J., M. A. Swertz, et al. (2007). "MetaNetwork: a computational protocol for the genetic study of metabolic networks." Nat Protoc **2**(3): 685-694.
- Fu, J., M. A. Swertz, et al. (2007). "MetaNetwork: A computational protocol for the genetic study of metabolic networks." Nature Protocols **2**(3): 685-694.
- Fulton, T. M., R. Van der Hoeven, et al. (2002). "Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants." Plant Cell **14**(7): 1457-1467.
- Gady, A. L., W. H. Vriezen, et al. (2012). "Induced point mutations in the phytoene synthase 1 gene cause differences in carotenoid content during tomato fruit ripening." Mol Breed **29**(3): 801-812.
- Galpaz, N., G. Ronen, et al. (2006). "A chromoplast-specific carotenoid biosynthesis pathway is revealed by cloning of the tomato white-flower locus." Plant Cell **18**(8): 1947-1960.
- Gerhauser, C. (2008). "Cancer chemopreventive potential of apples, apple juice, and apple components." Planta Medica **74**(13): 1608-1624.
- Goble, C. A., J. Bhagat, et al. (2010). "myExperiment: a repository and social network for the sharing of bioinformatics workflows." Nucleic Acids Res **38**(Web Server issue): W677-682.
- Goecks, J., A. Nekrutenko, et al. (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol **11**(8): R86.
- Goff, S. A., D. Ricke, et al. (2002). "A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica)." Science **296**(5565): 92-100.
- Granier, C., L. Aguirrezabal, et al. (2006). "PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit." New Phytol **169**(3): 623-635.
- Groenen, M. A. M., A. L. Archibald, et al. (2012). "Analyses of pig genomes provide insight into porcine demography and evolution." Nature **491**(7424): 393-398.
- Gruber, T. R. (1993). "A translation approach to portable ontology specifications." Knowl Acquis **5**(2): 199-220.
- Gruber, T. R. (1995). "Toward principles for the design of ontologies used for knowledge sharing." Int J Hum-Comput Stud **43**(5-6): 907-928.

- Gumucio, D. L., D. A. Shelton, et al. (1996). "Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes." *Mol Phylogenet Evol* **5**(1): 18-32.
- Gupta, P. K., S. Rustgi, et al. (2008). "Array-based high-throughput DNA markers for crop improvement." *Heredity (Edinb)* **101**(1): 5-18.
- Gupta, S., J. A. Stamatoyannopoulos, et al. (2007). "Quantifying similarity between motifs." *Genome Biol* **8**(2): R24.
- Gur, A. and D. Zamir (2004). "Unused natural variation can lift yield barriers in plant breeding." *PLoS Biol* **2**(10): e245.
- Guzman, I., S. Hamby, et al. (2010). "Variability of Carotenoid Biosynthesis in Orange Colored Capsicum spp." *Plant Sci* **179**(1-2): 49-59.
- Haley, C. S. and S. A. Knott (1992). "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers." *Heredity (Edinb)* **69**(4): 315-324.
- Hamilton, J. P., C. N. Hansey, et al. (2011). "Single nucleotide polymorphism discovery in elite North American potato germplasm." *BMC Genomics* **12**: 302.
- Hamilton, J. P., S.-C. Sim, et al. (2012). "Single Nucleotide Polymorphism Discovery in Cultivated Tomato via Sequencing by Synthesis." *Plant Gen* **5**(1): 17-29.
- Hammond, J. P., S. Mayes, et al. (2011). "Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in Brassica rapa." *Plant Physiol* **156**(3): 1230-1241.
- Hardison, R. C., J. Oeltjen, et al. (1997). "Long Human–Mouse Sequence Alignments Reveal Novel Regulatory Elements: A Reason to Sequence the Mouse Genome." *Genome Res* **7**(10): 959-966.
- Harlan, J. R. (1975). *Crops and man*. Madison, American society of agronomy.
- Harte, N., V. Silventoinen, et al. (2004). "Public web-based services from the European Bioinformatics Institute." *Nucleic Acids Res* **32**(Web Server issue): W3-9.
- Herrera-Estrella, L., A. Depicker, et al. (1983). "Expression of chimaeric genes transferred into plant cells using a Ti-plasmid-derived vector." *Nature* **303**(5914): 209-213.
- Hestand, M. S., M. van Galen, et al. (2008). "CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes." *BMC Bioinformatics* **9**: 495.
- Ho Sui, S. J., J. R. Mortimer, et al. (2005). "oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes." *Nucleic Acids Res* **33**(10): 3154-3164.
- Hospital, F. (2001). "Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs." *Genetics* **158**(3): 1363-1379.
- Hsieh, Y. and W. A. Korfmacher (2006). "Increasing speed and throughput when using HPLC-MS/MS systems for drug metabolism and pharmacokinetic screening." *Curr Drug Metab* **7**(5): 479-489.
- Huang da, W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic Acids Res* **37**(1): 1-13.
- Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." *Nat Protoc* **4**(1): 44-57.
- Hull, D., K. Wolstencroft, et al. (2006). "Taverna: a tool for building and running workflows of services." *Nucleic Acids Res* **34**(Web Server issue): W729-732.
- Hunter, S., P. Jones, et al. (2012). "InterPro in 2011: new developments in the family and domain prediction database." *Nucleic Acids Res* **40**(Database issue): D306-312.
- Iohnson, R. (2004). "Marker-assisted selection."
- Iyer-Pascuzzi, A. S., O. Symonova, et al. (2010). "Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems." *Plant Physiol* **152**(3): 1148-1157.
- Jaillon, O., J. M. Aury, et al. (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." *Nature* **449**(7161): 463-467.

- Jain, E., A. Bairoch, et al. (2009). "Infrastructure for the life sciences: design and implementation of the UniProt website." *BMC Bioinformatics* **10**: 136.
- James, A. T. and A. J. Martin (1952). "Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid." *Biochem J* **50**(5): 679-690.
- Jansen, R. C. and J. P. Nap (2001). "Genetical genomics: the added value from segregation." *Trends Genet* **17**(7): 388-391.
- Jöcker, A., F. Hoffmann, et al. (2008). "Protein function prediction and annotation in an integrated environment powered by web services (AFAWE)." *Bioinformatics* **24**(20): 2393-2394.
- Joosen, R. V., J. Kodde, et al. (2010). "GERMINATOR: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination." *Plant J* **62**(1): 148-159.
- Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Res* **28**(1): 27-30.
- Karlova, R., J. C. van Haarst, et al. (2013). "Identification of microRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis." *J Exp Bot* **64**(7): 1863-1878.
- Katari, M. S., S. D. Nowicki, et al. (2010). "VirtualPlant: a software platform to support systems biology research." *Plant Physiol* **152**(2): 500-515.
- Kearsey, M. J. (1998). "The principles of QTL analysis (a minimal mathematics approach)." *J Exp Bot* **49**(327): 1619-1623.
- Kearsey, M. J. and A. G. L. Farquhar (1998). "QTL analysis in plants; where are we now?" *Heredity (Edinb)* **80**(2): 137-142.
- Kelder, T., M. P. van Iersel, et al. (2012). "WikiPathways: building research communities on biological pathways." *Nucleic Acids Res* **40**(D1): D1301-D1307.
- Kenis, K. and J. Keulemans (2005). "Genetic linkage maps of two apple cultivars (*Malus x domestica* Borkh.) based on AFLP and microsatellite markers." *Molecular Breeding* **15**(2): 205-219.
- Kenis, K. and J. Keulemans (2007). "Study of tree architecture of apple (*Malus x domestica* Borkh.) by QTL analysis of growth traits." *Molecular Breeding* **19**(3): 193-208.
- Kerrien, S., B. Aranda, et al. (2012). "The IntAct molecular interaction database in 2012." *Nucleic Acids Res* **40**(D1): D841-D846.
- Keurentjes, J. J. B., J. Y. Fu, et al. (2006). "The genetics of plant metabolism." *Nature Genetics* **38**(7): 842-849.
- Khan, M. A., B. Duffy, et al. (2006). "QTL mapping of fire blight resistance in apple." *Molecular Breeding* **17**(4): 299-306.
- Khan, S. A., P.-Y. Chibon, et al. (2012). "Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16." *J Exp Bot* **63**(8): 2895-2908.
- Khodiyar, V. K., D. P. Hill, et al. (2011). "The representation of heart development in the gene ontology." *Dev Biol* **354**(1): 9-17.
- Kilambi, H. V., R. Kumar, et al. (2013). "Chromoplast-specific carotenoid-associated protein appears to be important for enhanced accumulation of carotenoids in hp1 tomato fruits." *Plant Physiology* **161**(4): 2085-2101.
- Kim, S. H., J. R. Lee, et al. (2003). "Molecular cloning and analysis of anthocyanin biosynthesis genes preferentially expressed in apple skin." *Plant Science* **165**(2): 403-413.
- King, G. J., J. R. Lynn, et al. (2001). "Resolution of quantitative trait loci for mechanical measures accounting for genetic variation in fruit texture of apple (*Malus pumila* Mill.)." *Theoretical and Applied Genetics* **102**(8): 1227-1235.
- Kircher, M. (2012). "Analysis of high-throughput ancient DNA sequencing data." *Methods Mol Biol* **840**: 197-228.
- Klee, H. J. and D. M. Tieman (2013). "Genetic challenges of flavor improvement in tomato." *Trends Genet* **29**(4): 257-262.

- Kloosterman, B., A. M. Anithakumari, et al. (2012). "Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source--sink tissues in a segregating potato population." *BMC Plant Biol* **12**: 17.
- Kloosterman, B., M. Oortwijn, et al. (2010). "From QTL to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy." *BMC Genomics* **11**(1): 158.
- Köhler, J., J. Baumbach, et al. (2006). "Graph-based analysis and visualization of experimental results with ONDEX." *Bioinformatics* **22**(11): 1383-1390.
- Korbel, J. O., T. Doerks, et al. (2005). "Systematic Association of Genes to Phenotypes by Genome and Literature Mining." *PLoS Biol* **3**(5): e134.
- Latchman, D. S. (1997). "Transcription factors: An overview." *Int J Biochem Cell Biol* **29**(12): 1305-1312.
- Lee, T. I. and R. A. Young (2000). "Transcription of eukaryotic protein-coding genes." *Annu Rev Genet* **34**: 77-137.
- Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." *EMBO J* **23**(20): 4051-4060.
- Li, H., H. Flachowsky, et al. (2007). "Maize *Lc* transcription factor enhances biosynthesis of anthocyanins, distinct proanthocyanidins and phenylpropanoids in apple (*Malus domestica* Borkh.)." *Planta* **226**(5): 1243-1254.
- Liebhart, R., L. Gianfranceschi, et al. (2002). "Development and characterisation of 140 new microsatellites in apple (*Malus x domestica* Borkh.)." *Molecular Breeding* **10**(4): 217-241.
- Liebhart, R., M. Kellerhals, et al. (2003). "Mapping quantitative physiological traits in apple (*Malus x domestica* Borkh.)." *Plant Molecular Biology* **52**(3): 511-526.
- Lillo, C., U. S. Lea, et al. (2008). "Nutrient depletion as a key factor for manipulating gene expression and product formation in different branches of the flavonoid pathway." *Plant, Cell and Environment* **31**(5): 587-601.
- Lincoln, R. E. and J. W. Porter (1950). "Inheritance of Beta-Carotene in Tomatoes." *Genetics* **35**(2): 206-211.
- Lommen, A. (2009). "Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing." *Analytical Chemistry* **81**(8): 3079-3086.
- Long, F., H. Liu, et al. (2004). "Genome-Wide Prediction and Analysis of Function-Specific Transcription Factor Binding Sites." *In Silico Biol. (Gedrukt)* **4**(4): 395-410.
- Lu, Y. and L. Y. Foo (1997). "Identification and quantification of major polyphenols in apple pomace." *Food Chemistry* **59**(2): 187-194.
- Luo, Z. W., C. I. Wu, et al. (2002). "Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes." *Genetics* **161**(2): 915-929.
- Mack, G. S. (2007). "MicroRNA gets down to business." *Nat Biotech* **25**(6): 631-638.
- Macneil, L. T. and A. J. Walhout (2011). "Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression." *Genome Res* **21**(5): 645-657.
- Maglott, D., J. Ostell, et al. (2005). "Entrez Gene: gene-centered information at NCBI." *Nucleic Acids Res* **33**(Database issue): D54-58.
- Maliepaard, C., F. H. Alston, et al. (1998). "Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers." *Theoretical and Applied Genetics* **97**(1-2): 60-73.
- Mardis, E. R. (2011). "A decade's perspective on DNA sequencing technology." *Nature* **470**(7333): 198-203.
- Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* **437**(7057): 376-380.
- Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* **437**(7057): 376-380.
- Maslow, A. H. (1943). "A theory of human motivation." *Psychological review* **50**(4): 370.
- Matos de, P., R. Alcantara, et al. (2010). "Chemical Entities of Biological Interest: an update." *Nucleic Acids Res* **38**(Database issue): D249-254.

- Mazza, G. and Y. S. Velioglu (1992). "Anthocyanins and other phenolic compounds in fruits of red-flesh apples." *Food Chemistry* **43**(2): 113-117.
- Mba, C., E. Guimaraes, et al. (2012). "Re-orienting crop improvement for the changing climatic conditions of the 21st century." *Agriculture & Food Security* **1**(1): 7.
- McArthur, D. and N. R. Knowles (1993). "Influence of Vesicular-Arbuscular Mycorrhizal Fungi on the Response of Potato to Phosphorus Deficiency." *Plant Physiol* **101**(1): 147-160.
- McArthur, D. A. and N. R. Knowles (1992). "Resistance Responses of Potato to Vesicular-Arbuscular Mycorrhizal Fungi under Varying Abiotic Phosphorus Levels." *Plant Physiol* **100**(1): 341-351.
- McCouch, S. (2004). "Diversifying selection in plant breeding." *PLoS Biol* **2**(10): e347.
- McGhie, T. K., M. Hunt, et al. (2005). "Cultivar and growing region determine the antioxidant polyphenolic concentration and composition of apples grown in New Zealand." *Journal of Agricultural and Food Chemistry* **53**(8): 3065-3070.
- McNally, K. L., K. L. Childs, et al. (2009). "Genomewide SNP variation reveals relationships among landraces and modern varieties of rice." *Proc Natl Acad Sci U S A* **106**(30): 12273-12278.
- Menda, N., R. M. Buels, et al. (2008). "A community-based annotation framework for linking solanaceae genomes with phenomes." *Plant Physiology* **147**(4): 1788-1799.
- Miles, C. and M. Wayne (2008). "Quantitative trait locus (QTL) analysis." *Nature Education* **1**(1).
- Moco, S., R. J. Bino, et al. (2006). "A liquid chromatography-mass spectrometry-based metabolome database for tomato." *Plant Physiology* **141**(4): 1205-1218.
- Mohammadi, A., M. H. Saraei, et al. (2011). "Identification of disease-causing genes using microarray data mining and Gene Ontology." *BMC Med Genomics* **4**: 12.
- Molnar, S. J., L. E. James, et al. (2000). "Inheritance and RAPD tagging of multiple genes for resistance to net blotch in barley." *Genome* **43**(2): 224-231.
- Mons, B., H. van Haagen, et al. (2011). "The value of data." *Nat Genet* **43**(4): 281-283.
- Moose, S. P. and R. H. Mumm (2008). "Molecular plant breeding as the foundation for 21st century crop improvement." *Plant Physiol* **147**(3): 969-977.
- Morgat, A., E. Coissac, et al. (2012). "UniPathway: a resource for the exploration and annotation of metabolic pathways." *Nucleic Acids Res* **40**(D1): D761-D769.
- Nesi, N., I. Debeaujon, et al. (2000). "The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in Arabidopsis siliques." *Plant Cell* **12**(10): 1863-1878.
- Nesi, N., C. Jond, et al. (2001). "The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed." *Plant Cell* **13**(9): 2099-2114.
- Newburger, D. E. and M. L. Bulyk (2009). "UniPROBE: an online database of protein binding microarray data on protein-DNA interactions." *Nucleic Acids Res* **37**(Database issue): D77-82.
- Nicole Redaschi and UniProt Consortium (2009). UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web. *3rd International Biocuration Conference*, Nature preceedings.
- O'Reilly, T. (2005, 2005-09-30). "What is Web 2.0." Retrieved April 8th 2013, from <http://oreilly.com/web2/archive/what-is-web-20.html>.
- Ohyanagi, H., T. Tanaka, et al. (2006). "The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information." *Nucleic Acids Res* **34**(Database issue): D741-744.
- Oinn, T., M. Addis, et al. (2004). "Taverna: a tool for the composition and enactment of bioinformatics workflows." *Bioinformatics* **20**(17): 3045-3054.
- Paran, I., I. Goldman, et al. (1995). "Recombinant Inbred Lines for Genetic-Mapping in Tomato." *Theor Appl Genet* **90**(3-4): 542-548.
- Park, S., N. Sugimoto, et al. (2006). "Identification of genes with potential roles in apple fruit development and biochemistry through large-scale statistical analysis of expressed sequence tags." *Plant Physiology* **141**(3): 811-824.
- Partner, J., A. Vukotic, et al. (2013). *Neo4j in Action*, Manning Publications Company.

- Paterson, A. H., J. E. Bowers, et al. (2009). "The Sorghum bicolor genome and the diversification of grasses." *Nature* **457**(7229): 551-556.
- Paterson, A. H., E. S. Lander, et al. (1988). "Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms." *Nature* **335**(6192): 721-726.
- Pavesi, G., P. Mereghetti, et al. (2004). "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." *Nucleic Acids Res* **32**(Web Server issue): W199-203.
- Pettersson, E., J. Lundeberg, et al. (2009). "Generations of sequencing technologies." *Genomics* **93**(2): 105-111.
- PGSC, X. Xu, et al. (2011). "Genome sequence and analysis of the tuber crop potato." *Nature* **475**(7355): 189-195.
- Pico, A. R., T. Kelder, et al. (2008). "WikiPathways: Pathway Editing for the People." *PLoS Biol* **6**(7): e184.
- Porter, M. F. (October, 2001). "Snowball: A language for stemming algorithms." Retrieved July 11th, 2013, from <http://snowball.tartarus.org/texts/introduction.html>.
- Powell, A. L. T., C. V. Nguyen, et al. (2012). "Uniform ripening Encodes a Golden 2-like Transcription Factor Regulating Tomato Fruit Chloroplast Development." *Science* **336**(6089): 1711-1715.
- Price, A. H. (2006). "Believe it or not, QTLs are accurate!" *Trends Plant Sci* **11**(5): 213-216.
- Prud'hommeaux, E. and A. Seaborne (2008). "SPARQL Query Language for RDF."
- Prudent, M., M. Causse, et al. (2009). "Genetic and physiological analysis of tomato fruit weight and composition: influence of carbon availability on QTL detection." *J Exp Bot* **60**(3): 923-937.
- Pruitt, K., G. Brown, et al. (2002 (revised April 6, 2012)). *The Reference Sequence (RefSeq) Database*. Internet, Bethesda (MD): National Center for Biotechnology Information (US).
- Pruitt, K. D., T. Tatusova, et al. (2012). "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." *Nucleic Acids Res* **40**(Database issue): D130-135.
- Pujar, A., R. Caspi, et al. (2009). "Plant Metabolic Pathways in MetaCyc and SolCyc." *Nat Precedings* <http://dx.doi.org/10.1038/npre.2009.3192.1>.
- R Core team (2013). R: A language and environment for statistical computing., R Foundation for Statistical Computing.
- Ray, J., P. Moureau, et al. (1992). "Cloning and characterization of a gene involved in phytoene synthesis from tomato." *Plant Mol Biol* **19**(3): 401-404.
- Reiter, B., U. Pfeifer, et al. (2002). "Response of endophytic bacterial communities in potato plants to infection with *Erwinia carotovora* subsp. *atroseptica*." *Appl Environ Microbiol* **68**(5): 2261-2268.
- Remington, D. L. and M. D. Purugganan (2003). "Candidate Genes, Quantitative Trait Loci, and Functional Trait Evolution in Plants." *Int J Plant Sci* **164**(S3): S7-S20.
- Renfro, D. P., B. K. McIntosh, et al. (2012). "GONUTS: the Gene Ontology Normal Usage Tracking System." *Nucleic Acids Res* **40**(Database issue): D1262-1269.
- Rice, P., I. Longden, et al. (2000). "EMBOSS: The European Molecular Biology Open Software Suite." *Trends Genet* **16**(6): 276-277.
- Robertson, G., M. Bilenky, et al. (2006). "cisRED: a database system for genome-scale computational discovery of regulatory elements." *Nucleic Acids Res* **34**(Database issue): D68-73.
- Sadok, W., P. Naudin, et al. (2007). "Leaf growth rate per unit thermal time follows QTL-dependent daily patterns in hundreds of maize lines under naturally fluctuating conditions." *Plant Cell Environ* **30**(2): 135-146.
- Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." *Nucleic Acids Res* **32**(Database issue): D91-94.
- Sayers, E. and V. Miller. (2010, January 21, May 31th, 2012). "Overview of the E-utility Web Service (SOAP)." Retrieved April 8th, 2013, from <http://www.ncbi.nlm.nih.gov/books/NBK43082/>.

- Saze, H., K. Tsugane, et al. (2012). "DNA Methylation in Plants: Relationship to Small RNAs and Histone Modifications, and Functions in Transposon Inactivation." *Plant Cell Physiol* **53**(5): 766-784.
- Schnable, P. S., D. Ware, et al. (2009). "The B73 maize genome: complexity, diversity, and dynamics." *Science* **326**(5956): 1112-1115.
- Schouten, H. J., F. A. Krens, et al. (2006). "Cisgenic plants are similar to traditionally bred plants: International regulations for genetically modified organisms should be altered to exempt cisgenesis." *EMBO Reports* **7**(8): 750-753.
- Schouten, H. J., F. A. Krens, et al. (2006). "Do cisgenic plants warrant less stringent oversight? [6]." *Nature Biotechnology* **24**(7): 753.
- Schouten, H. J., W. E. van de Weg, et al. (2011). "Diversity arrays technology (DArT) markers in apple for genetic linkage maps." *Molecular Breeding*: 1-16.
- Sequencing, T. B. G., A. Consortium, et al. (2009). "The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution." *Science* **324**(5926): 522-528.
- Shah, S., Y. Huang, et al. (2005). "Atlas - a data warehouse for integrative bioinformatics." *BMC Bioinformatics* **6**(1): 34.
- Shrestha, R., L. Matteis, et al. (2012). "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." *Front Physiol* **3**.
- Silfverberg-Dilworth, E., C. L. Matasci, et al. (2006). "Microsatellite markers spanning the apple (*Malus x domestica* Borkh.) genome." *Tree Genetics and Genomes* **2**(4): 202-224.
- Sim, S.-C., G. Durstewitz, et al. (2012). "Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato." *PLoS ONE* **7**(7): e40563.
- Smith, B., M. Ashburner, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nat Biotechnol* **25**(11): 1251-1255.
- Sozzani, R. and P. N. Benfey (2011). "High-throughput phenotyping of multicellular organisms: finding the link between genotype and phenotype." *Genome Biol* **12**(3): 219.
- Stemers, F. J., W. Chang, et al. (2006). "Whole-genome genotyping with the single-base extension assay." *Nat Meth* **3**(1): 31-33.
- Stein, L. (2001). "Genome annotation: from sequence to biology." *Nat Rev Genet* **2**(7): 493-503.
- Stein, L. (2002). "Creating a bioinformatics nation." *Nature* **417**(6885): 119-120.
- Stein, L. (2010). "The case for cloud computing in genome informatics." *Genome Biol* **11**(5): 207.
- Stein, L. D., C. Mungall, et al. (2002). "The generic genome browser: a building block for a model organism system database." *Genome Res* **12**(10): 1599-1610.
- Sterck, L., K. Billiau, et al. (2012). "ORCAE: online resource for community annotation of eukaryotes." *Nat Meth* **9**(11): 1041-1041.
- Sundaram, R. M., B. Naveenkumar, et al. (2008). "Identification of informative SSR markers capable of distinguishing hybrid rice parental lines and their utilization in seed purity assessment." *Euphytica* **163**(2): 215-224.
- Szankowski, I., H. Li, et al. (2009). Metabolic engineering of flavonoid biosynthesis in apple (*Malus x domestica* Borkh.). *Acta Horticulturae*. **814**: 511-516.
- Takos, A. M., B. E. Ubi, et al. (2006). "Condensed tannin biosynthesis genes are regulated separately from other flavonoid biosynthesis genes in apple fruit skin." *Plant Science* **170**(3): 487-499.
- Tanksley, S. D., N. D. Young, et al. (1989). "RFLP Mapping in Plant Breeding: New Tools for an Old Science." *Nat Biotech* **7**(3): 257-264.
- Terrier, N., L. Torregrosa, et al. (2009). "Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggests additional targets in the pathway." *Plant Physiology* **149**(2): 1028-1041.
- Tester, M. and P. Langridge (2010). "Breeding technologies to increase crop production in a changing world." *Science* **327**(5967): 818-822.
- The Plant Ontology Consortium (2002). "The Plant Ontology Consortium and plant ontologies." *Comp Funct Genomics* **3**(2): 137-142.

- The UniProt Consortium (2013). "Update on activities at the Universal Protein Resource (UniProt) in 2013." *Nucleic Acids Res* **41**(D1): D43-D47.
- The UniProt Consortium. (March 21st, 2012, March 21st, 2012). "How can I access resources on this web site programmatically?" Retrieved April 8th,, 2013, from <http://www.uniprot.org/faq/28>.
- Thorup, T. A., B. Tanyolac, et al. (2000). "Candidate gene analysis of organ pigmentation loci in the Solanaceae." *Proc Natl Acad Sci U S A* **97**(21): 11192-11197.
- Tikunov, Y. M., S. Laptinok, et al. (2011). "MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data." *Metabolomics*: 1-5.
- Tomato genome consortium (2012). "The tomato genome sequence provides insights into fleshy fruit evolution." *Nature* **485**(7400): 635-641.
- Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." *Nat Biotech* **23**(1): 137-144.
- Trebbi, D., M. Maccaferri, et al. (2011). "High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.)." *Theor Appl Genet* **123**(4): 555-569.
- Treutter, D. (2001). "Biosynthesis of phenolic compounds and its regulation in apple." *Plant Growth Regulation* **34**(1): 71-89.
- Van der Sluis, A. A., M. Dekker, et al. (2001). "Activity and concentration of polyphenolic antioxidants in apple: Effect of cultivar, harvest year, and storage conditions." *Journal of Agricultural and Food Chemistry* **49**(8): 3606-3613.
- van Helden, J., B. Andre, et al. (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." *J Mol Biol* **281**(5): 827-842.
- Van Ooijen, J. W. (2009). *MapQTL 6.0. Software for the mapping of quantitative trait loci in experimental populations of diploid species*. Wageningen, The Netherlands, Kyazma BV.
- Van Ooijn, J. (2009). "JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations " *Kyazma B.V., Wageningen, The Netherlands*.
- Varin, L., V. DeLuca, et al. (1992). "Molecular characterization of two plant flavonol sulfotransferases." *Proc Natl Acad Sci U S A* **89**(4): 1286-1290.
- Varshney, R. K., A. Graner, et al. (2005). "Genomics-assisted breeding for crop improvement." *Trends Plant Sci* **10**(12): 621-630.
- Varshney, R. K., D. A. Hoisington, et al. (2006). "Advances in cereal genomics and applications in crop breeding." *Trends Biotechnol* **24**(11): 490-499.
- Veerla, S. and M. Hoglund (2006). "Analysis of promoter regions of co-expressed genes identified by microarray analysis." *BMC Bioinformatics* **7**(1): 384.
- Velasco, R., A. Zharkikh, et al. (2010). "The genome of the domesticated apple (*Malus x domestica* Borkh.)." *Nat Genet* **42**(10): 833-839.
- Velasco, R., A. Zharkikh, et al. (2010). "The genome of the domesticated apple (*Malus x domestica* Borkh.)." *Nature Genetics* **42**(10): 833-839.
- Vilo, J., A. Brazma, et al. (1999). "Discovery of putative transcription factor binding sites from microarray-based gene expression profiles." *Nat Genet* **23**: 79-80.
- Visser, R. F., C. B. Bachem, et al. (2009). "Sequencing the Potato Genome: Outline and First Results to Come from the Elucidation of the Sequence of the World's Third Most Important Food Crop." *Am J Potato Res* **86**(6): 417-429.
- Voorrips, R. E. (2002). "MapChart: software for the graphical presentation of linkage maps and QTLs." *J Hered* **93**(1): 77-78.
- Vos, P., R. Hogers, et al. (1995). "AFLP: a new technique for DNA fingerprinting." *Nucleic Acids Res* **23**(21): 4407-4414.
- W3C. (April 27th, 2007). "SOAP Version 1.2." Retrieved April 8th, 2013, from <http://www.w3.org/TR/soap>.
- W3C. (February 10th, 2004). "RDF Test cases." 1.9. Retrieved April 9th, 2013, from <http://www.w3.org/TR/rdf-testcases/>.

- W3C. (February 10th, 2004). "RDF/XML Syntax Specification (Revised)." Retrieved April 9th, 2013, from <http://www.w3.org/TR/REC-rdf-syntax/>.
- W3C. (February 10th, 2004). "Resource Description Framework (RDF): Concepts and Abstract Syntax." Retrieved April 9th, 2013, from <http://www.w3.org/TR/rdf-concepts/>.
- W3C. (February 11th, 2004). "Web Services Glossary." Retrieved April 8th, 2013, from <http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice>.
- W3C. (February 19th, 2013). "Turtle." Retrieved April 9th, 2013, from <http://www.w3.org/TR/turtle/>.
- W3C. (January 24th, 2012). "Extensible Markup Language (XML)." Retrieved April 8th, 2013, from <http://www.w3.org/XML/>.
- W3C. (March 21st, 2013). "SPARQL 1.1 Overview." Retrieved April 9th, 2013, from <http://www.w3.org/TR/sparql11-overview/>.
- W3C. (March 28th, 2011). "Notation3 (N3): A readable RDF syntax." Retrieved April 9th, 2013, from <http://www.w3.org/TeamSubmission/n3/>.
- Weight, C., D. Parnham, et al. (2008). "LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation." *Plant J* **53**(3): 578-586.
- Wilkinson, M., H. Schoof, et al. (2005). "BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case." *Plant Physiol* **138**(1): 5-17.
- Wilkinson, M. D., M. Senger, et al. (2008). "Interoperability with Moby 1.0--it's better than sharing your toothbrush!" *Brief Bioinform* **9**(3): 220-231.
- Wilkinson, M. D., B. Vandervalk, et al. (2011). "The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation." *J Biomed Semantics* **2**(1): 8.
- Williams, J. G., A. R. Kubelik, et al. (1990). "DNA polymorphisms amplified by arbitrary primers are useful as genetic markers." *Nucleic Acids Res* **18**(22): 6531-6535.
- Wilson, D., V. Charoensawan, et al. (2008). "DBD--taxonomically broad transcription factor predictions: new content and functionality." *Nucleic Acids Res* **36**(Database issue): D88-92.
- Wingender, E. (2008). "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation." *Brief Bioinform* **9**(4): 326-332.
- Winkel-Shirley, B. (2001). "Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology." *Plant Physiology* **126**(2): 485-493.
- Wishart, D. S., D. Tzur, et al. (2007). "HMDB: the Human Metabolome Database." *Nucleic Acids Res* **35**(Database issue): D521-526.
- Wittenberg, A. H., T. van der Lee, et al. (2005). "Validation of the high-throughput marker technology DARt using the model plant *Arabidopsis thaliana*." *Mol Genet Genomics* **274**(1): 30-39.
- Wolters, A. M., J. G. Uitdewilligen, et al. (2010). "Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers." *Plant Mol Biol* **73**(6): 659-671.
- Xie, D. Y., S. B. Sharma, et al. (2003). "Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis." *Science* **299**(5605): 396-399.
- Yamamoto, T., H. Nagasaki, et al. (2010). "Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms." *BMC Genomics* **11**: 267.
- Yang, H., M. Shankar, et al. (2002). "Development of molecular markers using MFLP linked to a gene conferring resistance to *Diaporthe toxica* in narrow-leafed lupin (*Lupinus angustifolius* L.)." *Theor Appl Genet* **105**(2-3): 265-270.
- Yang, H., Y. Tao, et al. (2012). "Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in *Lupinus angustifolius* L." *BMC Genomics* **13**: 318.
- Yang, Q., D. Zhang, et al. (2012). "A sequential quantitative trait locus fine-mapping strategy using recombinant-derived progeny." *J Integr Plant Biol* **54**(4): 228-237.
- Yeung, K. Y., M. Medvedovic, et al. (2004). "From co-expression to co-regulation: how many microarray experiments do we need?" *Genome Biol* **5**(7): R48.

- Yin, Z., C. Li, et al. (2008). "Identification of conserved microRNAs and their target genes in tomato (*Lycopersicon esculentum*)."
Gene **414**(1–2): 60-66.
- Yu, J., S. Hu, et al. (2002). "A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica)."
Science **296**(5565): 79-92.
- Yu, J., J. Wang, et al. (2005). "The Genomes of *Oryza sativa*: a history of duplications."
PLoS Biol **3**(2): e38.
- Yvert, G., R. B. Brem, et al. (2003). "Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors."
Nature Genetics **35**(1): 57-64.
- Zamir, D. (2001). "Improving plant breeding with exotic genetic libraries."
Nat Rev Genet **2**(12): 983-989.
- Zhang, H., J. Jin, et al. (2011). "PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database."
Nucleic Acids Res **39**(Database issue): D1114-1117.
- Zhang, H. M., H. Chen, et al. (2012). "AnimalTFDB: a comprehensive animal transcription factor database."
Nucleic Acids Res **40**(Database issue): D144-149.
- Zuo, J., B. Zhu, et al. (2012). "Sculpting the maturation, softening and ethylene pathway: The influences of microRNAs on tomato fruits."
BMC Genomics **13**(1): 7.

Summary

Over the last decade, the amount of data generated by a single run of a NGS sequencer outperforms days of work done with Sanger sequencing. Metabolomics, proteomics and transcriptomics technologies have also involved producing more and more information at an ever faster rate. In addition, the number of databases available to biologists and breeders is increasing every year. The challenge for them becomes two-fold, namely: to cope with the increased amount of data produced by these new technologies and to cope with the distribution of the information across the Web. An example of a study with a lot of ~omics data is described in **Chapter 2**, where more than 600 peaks have been measured using liquid chromatography mass-spectrometry (LCMS) in peel and flesh of a segregating F_1 apple population. In total, 669 mQTL were identified in this study. The amount of mQTL identified is vast and almost overwhelming. Extracting meaningful information from such an experiment requires appropriate data filtering and data visualization techniques. The visualization of the distribution of the mQTL on the genetic map led to the discovery of QTL hotspots on linkage group: 1, 8, 13 and 16. The mQTL hotspot on linkage group 16 was further investigated and mainly contained compounds involved in the phenylpropanoid pathway. The apple genome sequence and its annotation were used to gain insight in genes potentially regulating this QTL hotspot. This led to the identification of the structural gene leucoanthocyanidin reductase (LAR1) as well as seven genes encoding transcription factors as putative candidates regulating the phenylpropanoid pathway, and thus candidates for the biosynthesis of health beneficial compounds. However, this study also indicated bottlenecks in the availability of biologist-friendly tools to visualize large-scale QTL mapping results and smart ways to mine genes underlying QTL intervals.

In this thesis, we provide bioinformatics solutions to allow exploration of regions of interest on the genome more efficiently. In **Chapter 3**, we describe MQ^2 , a tool to visualize results of large-scale QTL mapping experiments. It allows biologists and breeders to use their favorite QTL mapping tool such as MapQTL or R/qtl and visualize the distribution of these QTL among the genetic map used in the analysis with MQ^2 . MQ^2 provides the distribution of the QTL over the markers of the genetic map for a few hundreds traits. MQ^2 is accessible online via its web interface but can also be used locally via its command line interface. In **Chapter 4**, we describe Marker2sequence (M2S), a tool to filter out genes of interest from all the genes underlying a QTL. M2S returns the list of genes for a specific genome interval and provides a search function to filter out genes related to the provided keyword(s) by their annotation. Genome annotations often contain cross-references to resources such as the Gene Ontology (GO), or proteins of the UniProt database. Via these annotations, additional information can be gathered about each gene. By integrating information from different resources and offering a way to mine the list of genes present in a QTL interval, M2S provides a way to reduce a list of hundreds of genes to possibly tens or less of genes potentially related to the trait of interest. Using semantic web technologies M2S integrates multiple resources and has the flexibility to extend this integration to more resources as they become available to these technologies.

Besides the importance of efficient bioinformatics tools to analyze and visualize data, the work in **Chapter 2** also revealed the importance of regulatory elements controlling key genes of pathways. The limitation of M2S is that it only considers genes within the interval. In genome annotations, transcription factors are not linked to the trait (keyword) and to the gene it controls, and these relationships will therefore not be considered. By integrating information about the gene regulatory network of the organism into Marker2sequence, it should be able to integrate in its list of genes, genes outside of the QTL interval but regulated by elements present within the QTL interval. In tomato, the genome annotation already lists a number of transcription factors, however, it does not

provide any information about their target. In **Chapter 5**, we describe how we combined transcriptomics information with six genotypes from an Introgression Line (IL) population to find genes differentially expressed while being in a similar genomic background (i.e.: outside of any introgression segments) as the reference genotype (with no introgression). These genes may be differentially expressed as a result of a regulatory element present in an introgression. The promoter regions of these genes have been analyzed for DNA motifs, and putative transcription factor binding sites have been found.

The approaches taken in M2S (**Chapter 4**) are focused on a specific region of the genome, namely the QTL interval. In **Chapter 6**, we generalized this approach to develop Annotex. Annotex provides a simple way to browse the cross-references existing between biological databases (ChEBI, Rhea, UniProt, GO) and genome annotations. The main concept of Annotex being, that from any type of data present in the databases, one can navigate the cross-references to retrieve the desired type of information.

This thesis has resulted in the production of three tools that biologists and breeders can use to speed up their research and build new hypothesis on. This thesis also revealed the state of bioinformatics with regards to data integration. It also reveals the need for integration into annotations (for example, genome annotations, protein annotations, and pathway annotations) of more ontologies than just the Gene Ontology (GO) currently used. Multiple platforms are arising to build these new ontologies but the process of integrating them into existing resources remains to be done. It also confirms the state of the data in plants where multiples resources may contain overlapping. Finally, this thesis also shows what can be achieved when the data is made inter-operable which should be an incentive to the community to work together and build inter-operable, non-overlapping resources, creating a bioinformatics Web for plant research.

Samenvatting

In de afgelopen tien jaar, is de hoeveelheid data die in één run door een next generation sequencer gegenereerd wordt vele malen groter dan wat mogelijk was met behulp van Sanger sequensen. Metabolomics, proteomics en transcriptomics technologieën zijn ook verder ontwikkeld en genereren meer en meer data. Bovendien neemt het aantal beschikbare databanken, beschikbaar voor biologen en veredelaars, elk jaar toe. De uitdaging voor de mens wordt tweeledig, namelijk: om te gaan met de hoeveelheid data die gegenereerd wordt door deze nieuwe technologieën en om te gaan met het groeiend aantal plekken op het internet waar informatie beschikbaar is. Een voorbeeld van een studie met veel ~omics data is beschreven in Hoofdstuk 2, waar meer dan 600 pieken gemeten zijn met vloeistof chromatografie massaspectrometrie (LC-MS) in de schil en het vruchtvlees van individuen van een splitsende F1 appel-populatie. In totaal werden in deze studie voor 669 metaboliëten kwantitatieve trait loci (mQTL) geïdentificeerd. Het aantal mQTL is overweldigend en het extraheren van zinvolle informatie uit een dergelijk experiment vereist geschikte technieken voor data filtering en data visualisatie. Het visualiseren van de verdeling van mQTL op de genetische kaart leidde tot het identificeren van QTL hotspots op de koppelingsgroepen 1, 8, 13 en 16. De mQTL hotspot op koppelingsgroep 16 werd verder onderzocht en bevatte vooral verbindingen die betrokken zijn bij de fenylpropanoïde metabolische route. De genomsequentie van appel en de bijbehorende annotatie zijn gebruikt om inzicht te krijgen in genen die mogelijk betrokken zijn bij regulatie van deze QTL hotspot. Dit heeft geleid tot de identificatie van het structurele gen leucoanthocyanidin reductase (LAR1) en zeven genen die coderen voor mogelijke transscriptiefactoren als kandidaat genen voor regulatie van de fenylpropanoïde metabolische route, en dus als kandidaten voor de biosynthese van gezondheidsbevorderende verbindingen. Echter, deze studie toont ook aan dat er een knelpunt is in de beschikbaarheid van biologo-vriendelijke hulpmiddelen voor de visualisatie van resultaten van grootschalige QTL mapping resultaten en dat het ontbreekt aan slimme manieren om genen in een QTL interval te onderzoeken.

In dit proefschrift beschrijven we oplossingen vanuit de bio-informatica voor het efficiënter onderzoeken van interessante regio's op het genoom. In hoofdstuk 3 beschrijven we MQ2, een hulpmiddel om de resultaten van grootschalige QTL mapping experimenten te visualiseren. Biologen en veredelaars kunnen hun favoriete QTL mapping software, zoals MapQTL of R/qtl gebruiken om hun data te analyseren en vervolgens MQ2 gebruiken om de verdeling van de QTL t.o.v. een genetische kaart te visualiseren. MQ2 kan met gemak gegevens van vele honderden QTL visualiseren en is zowel beschikbaar als een web-applicatie en als een commandline-applicatie. In hoofdstuk 4 beschrijven we Marker to Sequence (M2S), een hulpmiddel voor het filteren van mogelijke kandidaatgenen in een QTL interval. M2S geeft een lijst met genen weer voor het gedefinieerde genoom-interval en biedt zoekmogelijkheden aan om te filteren op genen die een opgegeven trefwoord in hun annotatie hebben. Genoom-annotaties bevatten vaak verwijzingen naar andere bronnen, zoals Gene Ontology-gegevens (GO) of eiwitgegevens in de UniProt databank. Via deze verwijzingen kunnen extra gegevens verzameld worden over elk gen. Door het integreren van informatie uit verschillende bronnen met behulp van semantisch web-technologieën, wordt het zoeken met trefwoorden efficiënter. Ook kunnen door het gebruik van semantisch web-technologie gemakkelijk nieuwe gegevensbronnen toegevoegd worden aan het systeem. M2S biedt de mogelijkheid om een lijst van honderden genen te verkleinen tot tientallen mogelijke kandidaat genen gerelateerd aan het kenmerk waarnaar gekeken wordt.

Naast het belang van efficiënte bio-informatica-tools voor het analyseren en visualiseren van data laat de studie in hoofdstuk 2 ook zien wat het belang is van regulatie van sleutelgenen in een

metabolische route. De beperking van M2S is dat het alleen genen in een QTL-interval analyseert. In genoom-annotaties worden transcriptiefactoren niet gekoppeld aan een eigenschap en aan een gen of de genen die worden gereguleerd. Deze relaties worden daarom niet geëvalueerd in M2S. Door het integreren van informatie over een genregulatie netwerk in M2S moet het mogelijk zijn om via genen, die buiten het QTL interval liggen, de onderliggende regulerende genen te identificeren, die in het QTL-interval tot expressie komen. Hoewel de genoom-annotatie van tomaat een lijst met transcriptiefactoren bevat, ontbreekt informatie over welke genen door deze transcriptiefactoren gereguleerd kunnen worden. In hoofdstuk 5 beschrijven we hoe data van een genexpressie-experiment met zes introgressielijnen gebruikt worden om genen op te sporen die tot expressie komen, terwijl de genoomlocatie(s) buiten het introgressie segment liggen. Deze expressie wordt hiervoor vergeleken met het referentie-genotype (dat deze introgressies niet bevat). De hypothese is dat deze genen differentieel tot expressie komen vanwege regulatie door genen (bijvoorbeeld transcriptiefactoren) die in het introgressie-segment liggen. De promotorregio's van deze genen zijn geanalyseerd op het bovengemiddeld voorkomen van DNA-motieven, die wij aanmerken als mogelijke binding sites voor transcriptiefactoren.

De aanpak gekozen in M2S (hoofdstuk 4) is gericht op een specifieke regio van het genoom, namelijk het QTL-interval. In hoofdstuk 6 beschrijven we een generalisatie van deze aanpak en de ontwikkeling van Annotex. Annotex biedt een simpele manier om wederzijdse referenties tussen biologische databanken (ChEBI, Rhea, UniProt, GO) en genoom-annotaties te verkennen. Het belangrijkste concept van Annotex is dat het mogelijk is om van elk type data (bijvoorbeeld gen of eiwit) informatie te krijgen over elk andere type (bijvoorbeeld metabolische route of literatuur).

Dit proefschrift heeft geresulteerd in de ontwikkeling van drie methodes die biologen en veredelaars kunnen gebruiken om hun onderzoek te versnellen en te helpen bij het formuleren van nieuwe onderzoeksvragen. Dit proefschrift laat ook de stand van zaken zien met betrekking tot data-integratie en de noodzaak van integratie van meer data (bijvoorbeeld eiwit-annotatie en metabolische routes) dan alleen Gene Ontology in de genoom-sequentie. Er zijn verscheidene platforms ontstaan die nieuwe ontologieën bouwen, maar het gebruik hiervan in bestaande databanken is een punt van aandacht. Ook bevatten meerdere databanken (deels) dezelfde informatie. Tot slot laat dit proefschrift zien wat bereikt kan worden wanneer data inter-operabel wordt gemaakt, wat een stimulans moet zijn voor het verder bouwen van een inter-operabel web van bio-informatica-tools voor plantenonderzoek.

Résumé

Au cours de la dernière décennie, la quantité de données générées par une simple analyse de séquenceur de nouvelle génération remplace des jours de travail avec les séquenceurs Sanger. Les analyses métaboliques, protéomiques et transcriptomiques ont aussi évoluées pour générer de plus en plus de données à un rythme de plus en plus soutenu. De plus, le nombre de bases de données disponibles pour les biologistes et sélectionneurs augmente chaque année. Le défi devient alors double : arriver à gérer la quantité de données produites par ces nouvelles technologies et la distribution de l'information à travers le web. Un exemple d'expérience avec de grande quantité de données génomiques est présente dans le **chapitre 2**, où plus de 600 pics ont été mesurés par LCMS dans la pulpe et la peau d'une population F_1 de pomme. Au total, 669 mQTL ont été identifiés dans cette expérience, nombre trop important pour être traités manuellement. Extraire des informations utiles de telles expériences requiert analyse et visualisation. La visualisation de la distribution des mQTL sur la carte génétique a mené à la découverte de points de concentration de QTLs sur les groupes de liaisons 1, 8, 13 et 16. Le point de concentration de QTL sur le groupe de liaison 16 a été analysé de manière plus poussée et contient principalement des composés liés à la voie métabolique des phénylpropanoïdes. Le génome de la pomme et son annotation ont été utilisés pour comprendre l'action des gènes associés à ce point de concentration. Cette analyse a mené à la découverte du gène « Leucoanthocyanidin reductase » (LAR1) ainsi que sept gènes encodant pour des facteurs de transcription, potentiel gène candidat à la régulation de la voie métabolique des phénylpropanoïdes et donc gène candidat pour la synthèse de composé bénéfique pour la santé. Cette étude a cependant aussi mis en avant le manque d'outils permettant à des biologistes de visualiser les résultats d'analyse QTL à grande échelle et permettant d'explorer les gènes associés à des QTLs.

Dans cette thèse, nous présentons des outils de bioinformatique permettant une exploration efficace de régions du génome d'intérêt. Dans le **chapitre 3**, nous présentons MQ², un outil pour la visualisation des résultats d'analyse QTL à grande échelle. Il permet à des biologistes ou des sélectionneurs d'utiliser leur outil d'analyse QTL préféré tel que MapQTL ou R/qtl et en utilisant MQ², de visualiser la distribution des QTL sur la carte génétique utilisée lors de l'analyse. MQ² peut présenter la distribution des QTLs sur les marqueurs de la carte génétique pour quelques centaines de caractères. MQ² est disponible en ligne mais peut aussi être installé sur l'ordinateur de l'utilisateur et alors être utilisé par la ligne de commande. Dans le **chapitre 4**, nous présentons Marker2sequence (M2S), un outil permettant d'extraire des gènes d'intérêt depuis la liste de tous les gènes associés à un QTL. M2S donne accès à la liste de tous les gènes d'une région spécifique d'un génome et fournit une fonction de recherche permettant d'extraire les gènes contenant dans leur annotation le mot clé spécifié par l'utilisateur. L'annotation d'un génome contient souvent des références à d'autres ressources telles que l'ontologie des gènes « Gene Ontology » (GO) ou des protéines de la base de données UniProt. Des informations supplémentaires sur chaque gène du génome peuvent ainsi être rassemblées par ces références. En intégrant les informations de différentes ressources et en offrant une façon de les chercher, M2S permet de réduire une liste contenant quelques centaines de gènes, à éventuellement une liste de quelques gènes potentiellement liés au caractère étudié. M2S réalise l'intégration des données depuis différentes ressources en se basant sur les technologies du web sémantique. Cela offre une flexibilité permettant d'étendre l'intégration à de nouvelles ressources au fur et à mesure qu'elles s'ouvrent à ces technologies.

En plus de révéler l'importance d'outils de bioinformatique pour l'analyse et la visualisation des données, l'étude en **chapitre 2** a aussi révélé l'importance des éléments régulant l'expression des

gènes d'une voie métabolique. Une des limites d'utilisation de M2S vient du fait que M2S ne considère que les gènes présents dans l'intervalle spécifié. Dans l'annotation des génomes, les facteurs de transcriptions ne sont pas liés avec les caractères (mot clé) dont ils régulent l'expression, ces régulations ne sont donc pas prises en compte par M2S. En intégrant des informations sur les réseaux de régulations des gènes d'un organisme dans Marker2sequence, il devrait alors être capable d'intégrer dans la liste de gènes considérés, gènes hors de l'intervalle QTL mais régulés par des gènes présents dans le dit intervalle. Pour la tomate, l'annotation du génome liste déjà un certain nombre de facteurs de transcription, cependant, elle ne fournit pas d'information sur leur cibles. Dans le **chapitre 5**, nous avons combiné des mesures d'expression génique de six génotypes d'une lignée d'introgession pour trouver des gènes ayant une expression différente entre différents génotypes tout en étant dans une section du génome identique (c'est-à-dire, hors intégration) par rapport au génotype de référence (qui n'a aucune intégration). Ces gènes s'expriment peut-être différemment par la présence d'éléments régulateurs dans les intégrations du génome. Les promoteurs de ces gènes ont été analysés et 17 motifs, potentiels sites d'attache pour les facteurs de transcription, ont été trouvés.

L'approche prise avec M2S (**chapitre 4**) se concentre sur une région spécifique du génome, l'intervalle du QTL. Dans le **chapitre 6**, nous avons généralisé cette approche en développant Annotex. Annotex fournit une interface simple pour parcourir les références existantes entre différentes bases de données biologiques (ChEBI, Rhea, UniProt et GO) et l'annotation d'un génome. L'idée principale d'Annotex étant qu'à partir de n'importe quel type de données (gènes, protéines, etc.) présentes dans une des bases de données, l'utilisateur peut parcourir les références pour extraire des données d'un type désiré (gènes, protéines, voix métabolique, bibliographie, etc.).

Cette thèse a produit trois outils de bioinformatique que biologistes et sélectionneurs peuvent utiliser pour accélérer leurs recherches et créer de nouvelles hypothèses de recherche. Cette thèse a aussi révélé l'état de la bioinformatique en général par rapport à l'interopérabilité des bases de données. Cela a montré le besoin pour l'intégration dans les annotations (par exemple, les annotations de génomes, de protéines ou de voix métaboliques) de nouvelles ontologies, en plus de l'ontologie des gènes (GO) déjà utilisé. Plusieurs plateformes émergent pour aider à la création de ces ontologies, mais l'intégration de celles-ci dans les ressources existantes reste à faire. Cette thèse a aussi démontré l'état des bases de données biologiques pour les plantes, où plusieurs bases de données ont des informations qui se recouvrent. Enfin, cette thèse a montré ce qui pourra être fait quand ces ressources seront interopérables ce qui est une invitation pour la communauté scientifique à travailler en commun pour construire des ressources interopérables, qui ne se recouvrent pas et ainsi créer un réseau bioinformatique pour la biologie des plantes.

Acknowledgements

I consider this thesis, although being an account of my personal work, as the result of team work. This team is composed of all the people that supported me over the past four years, on a professional level, a personal level or both.

The first persons I want to thank are those without whom, this thesis would never have happened. I would like to thank my promoter, Richard Visser who offered me this PhD position four years ago. In a way a pretty unusual position, bioinformatician in a group of biologists. I also would like to thank Richard Finkers, my supervisor for the last 6 years. You have followed me and supported me through three degrees but this last one will remain the most challenging and rewarding. I sincerely would like to thank you both for all your support and help over these four years but also in the last months of the thesis while I was “programming” in the most difficult language (for me): scientific English.

Over these 6 years I have spent in the Plant Breeding group, I have met quite a few colleagues but many, many more friends.

From the French team that was there when I arrived, Anoma, Nicolas, Mathieu, Estelle, Benoit and Antoine. To the Dutch team making sure the rules of “no more than two French people in one office at the same time” was followed: Jeroen and Bjorn and our Belgium cyclist, Björn. Krissana and her world famous toffee cake. There have been many occasion of fun and parties, the Batavierenrace (and its night team), your defenses, Björn's wedding and your babies. This time I am the one inviting but rest assure this will not be the last celebration we will do.

Plant Breeding is an ever changing department as PhD students come and go. Over the years I have made many friends that I would like to thank for just being whom they are and for the time we spent together. Thank you, Chris for your availability, over these six years, every time I had a question. Thank you, Roeland for all your testing and constructive feedbacks. Thank you, Yury and Arnaud for being present every time I had a question about some biological data or metabolites. Thank you Arwa for all the discussion we have had about politics, culture or just for fun. Thank you Mirjam for all the fun we had, also about the propositions and promise, I will shut up (when you ask me). Thank you Yusuf for your always present good mood and all the fun we had. Thank you Mariame for all your support over these years and always being present when needed. Thank you Marian for the good times and fun we had in the lab or preparing Nasim's graduation party and for being there. Thank you Animesh for your smile and the time we were neighbors in the Bornesesteeg. Thank you Anitha for all the good times we had, in the lab or in the PhD council or outside of it. Thank you, Natalia, for cheering us up, up to the last days of the PhD. Thank you Theo for all these conversations we had about linux, BSD or FOSS in general, in Wageningen or at FOSDEM.

Finally I would like to thank Nasim, for being the friend she is and giving me the honor of being her paranymp with Marian at her defense.

I have also been in contact with the Bioinformatics department from WUR and PRI. We had very interesting work conversation every Thursday, quite often followed by some not-so-much-work-related discussion at the coffee machine. Thanks to Harm, Jan, Henri, Elio, Erwin (I shall remember that looking at seashells can be dangerous), Sandra, Sander, Aalt-Jan, Judith, Joachim, Felipe, Pieter,

Thomas, Edouard, Lin, Jose, Jan-Peter, Gabino and Jack (who passed away in May 2012) for all these times.

During my master study in Wageningen, I met some of my closest friends. After the exams of December 2008 we decided to go for a drink at the Forum building and we kept the idea in January making it a weekly event every Thursday, only to move the location to the “Vlaamsche Reus” on the market square. This group became known as the “Thursday group”, following the “Thursday” religion whose motto is “Any day can be a Thursday”. Being one of the founders of this tradition, I am glad to see it kept alive by the new generation of bioinformaticians. So if you drop by the Vlaamsche Reus on a Thursday, you will probably see a group of people talking about computers, programming language and how many spaces a tab should be, they are the Thursday group.

Of this group my generation consist of my closest friends Erik and (the other) Erik, Frank (aka the black sheep), Saulo, Aalt-Jan, Harm-Jan, Lex and more recently arrived: Paul, Heleen and Man.

If my defense is on a Thursday it is not quite a coincidence!

Frank, my friend, you crossed the ocean but I know we will keep in touch, looking forward visiting you.

Heleen became a full Thursday member but she is also a colleague. She managed to deal with me at work and still agree on getting a drink after work on Thursdays. She has been there for me on so many occasions. I owe her the apartment where I lived for the last two years and I am really grateful to her for this. Thank you, Heleen, for all your help and support over these years.

During these four years, I have been two years in the PhD council of my graduate school, EPS. I have had a lot of fun and this I owe to Jenny, Tila, Martine, Padraic.

Via the PhD council I also met Wilma and later Pieter who became close friends, thank you to both of you for the nice evenings and conferences we went to.

There are also a number of friends from groups that I have not mentioned before that I would like to thank. Thank you, Marion for being there all these years, 10 years now that we know each other. Thank you Warren for all the adventures we had, Morocco was a unique experience that I will never forget. Thank you Karsten for all these years in Wageningen, the parties, the dinners, the evenings (I will need to train more on commodore 64). Thanks also to Karin for welcoming me as a friend every time. Thanks to Andreas & Jenny for being such nice neighbor, always there to help, talk to and play a game of chess, backgammon or GO.

Thanks also to the French girls from Lille, Caro, Berthe, Maelle and Flori for all the great time we had while you were here, I am looking forward to the next time.

There is also a group of friends that I would like to thank. Although I have not been interacting with them at a professional level, I have learned much from them and with them. Knowledge that I have then been able to apply within my thesis, and which led me to advocate good software practices also in bioinformatics. Thank you, Toshio, Ralph, Luke, Ian, Ryan, Ricky, Aurélien, Máirín, Patrick and Tom. We were friends hacking together, we are now colleagues. I am very happy to be working with you and looking forward to keep learning from you and with you in the coming years.

Acknowledgements

I would like to thank my two paranymphs Juliane and Benoit. You are both close friends that I know I can rely on. Juliane we have met via the PhD council and even discovered that we shared some friends later in the persons of Bjorn and Alexander. You have always been present and a close friend. I am looking forward showing you (both) the grape harvest and other parts of France.

Benoit, you have been my student before being my colleague. We have had many good evenings, playing games or watching movies or just for a barbecue. Congratulations on your wedding, and again good luck for your PhD. Maybe one day I can convince you to use a decent Linux distribution.

Enfin, je voudrais remercier mes parents et ma sœur qui m'ont poussé à faire ce doctorat et m'ont ensuite supporté pendant ces quatre années. Les explications sur ce que je faisais n'étaient pas toujours les plus claires mais vous avez toujours été là pour moi et je vous en remercie.

Merci aussi à mon parrain et ma marraine pour leur présence et soutien à chaque étape de ma vie.

Et Sarah, je tiens à te remercier, toi aussi, pour avoir accepté que je fasse ce doctorat même si on s'est rapidement rendu compte que cela impliquerait vivre séparément pendant plusieurs années. Merci pour ton soutien dans les périodes de doutes ou de découragement. Merci de ta présence pour les périodes de joies et de succès. Nous voilà maintenant réunis et avec une vie de couple à créer.

I am happy to share this day with you all.

Thank you, Merci, Bedankt.

Pierre-Yves

Curriculum vitae

Pierre-Yves Chibon was born on August 14th, 1985 in Paris, France. From 1985 until 2003 he studied in Paris until passing his “baccalauréat” at 17 years old. The following September, Pierre-Yves moved to Toulouse to study agriculture, food sciences and related industries in the “École d’Ingénieur de Purpan (EIP)”, (former “École Supérieure d’Agriculture de Purpan (ESAP)”). In 2006, Pierre-Yves received a bachelor in agriculture, food sciences and related industries and in January 2007, he moved to Wageningen as an exchange student within the Erasmus program. After six months in Wageningen, Pierre-Yves stayed for a double degree in bioinformatics and agriculture between his French university and Wageningen University. In June 2007, Pierre-Yves joined the Plant Breeding department of Wageningen University for his master thesis. After his Msc in Bioinformatics, Pierre-Yves worked six months in the Plant Breeding department after which he started his PhD under the supervision of Prof. Dr. Richard Visser and Dr. Richard Finkers. The results of his research are presented in this thesis.

Nowadays, Pierre-Yves Chibon has been hired by Red Hat as a software engineer to work for the Fedora project.

Publications

Bolser, D. M., P.-Y. Chibon, et al. (2012). "MetaBase—the wiki-database of biological databases." *Nucleic Acids Res* **40**(D1): D1250-D1254.

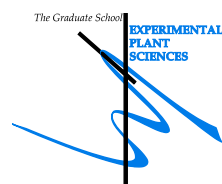
Chibon, P.-Y., H. Schoof, et al. (2012). "Marker2sequence, mine your QTL regions for candidate genes." *Bioinformatics* **28**(14): 1921-1922.

Chibon, P.-Y., R.Voorrips, et al. (2013). "MQ²: visualizing multi-trait mapped QTL results." *Mol Breeding* 1380-3743: 1-5.

Khan, S. A., P.-Y. Chibon, et al. (2012). "Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16." *J Exp Bot* **63**(8): 2895-2908.

Kloosterman, B., A. M. Anithakumari, et al. (2012). "Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source--sink tissues in a segregating potato population." *BMC Plant Biol* **12**: 17.

**Education Statement of the Graduate School
Experimental Plant Sciences**



Issued to: Pierre-Yves Chibon
Date: 7 November 2013
Group: Plant Breeding, Wageningen University & Research Centre

1) Start-up phase <ul style="list-style-type: none"> ▶ First presentation of your project Candidate gene prediction from large-scale -Omics QTL data-sets. ▶ Writing or rewriting a project proposal PhD Proposal 'Candidate gene prediction through data integration for high-throughput QTL data-sets' ▶ Writing a review or book chapter ▶ MSc courses ▶ Laboratory use of isotopes 	<p align="right"><i>date</i></p> <p align="right">Oct 13, 2009</p> <p align="right">Dec 17, 2009</p>
<i>Subtotal Start-up Phase</i> 7.5 credits*	
2) Scientific Exposure <ul style="list-style-type: none"> ▶ EPS PhD student days EPS PhD student days (Utrecht University) EPS PhD student days (Wageningen University) International PhD Retreat (Paris, France) ExPeCtationS Career Day, Wageningen University ▶ EPS theme symposia EPS theme 3 'Metabolism and Adaptation' Symposium (University of Amsterdam) ▶ NWO Lunteren days and other National Platforms NWO-ALW meeting Lunteren NWO-ALW meeting Lunteren ▶ Seminars (series), workshops and symposia Workshop: Semantic Web Applications and Tools for Life Sciences CBSG tomato cluster workshop PSG Seminar Gabino Sanchez Perez ▶ Seminar plus ▶ International symposia and congresses EU-SOL workshop in Munich EU-SOL conference in Natal CBSG tomato cluster meeting CBSG tomato cluster meeting NBIC conference in Lunteren CBSG tomato cluster meeting NBIC conference in Lunteren SWAT4LS in Paris CBSG tomato cluster meeting ▶ Presentations Poster Presentation EU-SOL Natal : From QTLs to potential candidate genes. Poster Presentation CBSG Wageningen: BreeDB: Exploring the tomato Genome Poster Presentation Intern. PhD Retreat (Paris): From QTLs to potential candidate genes. Poster NBIC Conference: From QTLs to potential candidate genes. Poster SWAT4LS Conference: Using semantic web technology to accelerate plant breeding. Presentation at the CBSG genome mining course Presentation at the Data Management course from Wageningen Business school Presentation at the Data Management course from Wageningen Business school Presentation at the John Ines Center (Computational and System biology department) Presentation meeting Maastricht Presentation at the 100 year PBR conference Presentation at the CBSG genome mining course Presentation at the CBSG Tomato cluster meeting ▶ IAB interview ▶ Excursions 	<p align="right"><i>date</i></p> <p align="right">Jun 01, 2010</p> <p align="right">May 20, 2011</p> <p align="right">Jul 06-08, 2011</p> <p align="right">Nov 18, 2011</p> <p align="right">Mar 22, 2013</p> <p align="right">Apr 04-05, 2011</p> <p align="right">Apr 02-03, 2012</p> <p align="right">Nov 20, 2009</p> <p align="right">Oct 26, 2010</p> <p align="right">Mar 12, 2013</p> <p align="right">Jun 30-Jul 02, 2010</p> <p align="right">Nov 13-17, 2010</p> <p align="right">Oct 07, 2010</p> <p align="right">Jan 31-Feb 01, 2011</p> <p align="right">Apr 19-20, 2011</p> <p align="right">Feb 29, 2012</p> <p align="right">Apr 24-25, 2012</p> <p align="right">Nov 29-30, 2012</p> <p align="right">Feb 10-13, 2013</p> <p align="right">Nov 13-17, 2010</p> <p align="right">Jan 31-Feb 01, 2011</p> <p align="right">Jul 06-08, 2011</p> <p align="right">Apr 24-25, 2012</p> <p align="right">Nov 29-30, 2012</p> <p align="right">Oct 27, 2011</p> <p align="right">Dec 14, 2011</p> <p align="right">Mar 09, 2012</p> <p align="right">Apr 17, 2012</p> <p align="right">Jun 06, 2012</p> <p align="right">Nov 14, 2012</p> <p align="right">Dec 13, 2012</p> <p align="right">Feb 13, 2013</p>
<i>Subtotal Scientific Exposure</i> 23.0 credits*	
3) In-Depth Studies <ul style="list-style-type: none"> ▶ EPS courses or other PhD courses Managing Life Science Information CBSG Genome Mining Course Ethics and Philosophy in Life Science CBSG Genome Mining Course ▶ Journal club Literature discussion "Plant Breeding" ▶ Individual research training MPMI-Cologne in the group of Heiko Schoof 	<p align="right"><i>date</i></p> <p align="right">Oct 17-21 & 28, 2011</p> <p align="right">Oct 27, 2011</p> <p align="right">Jun 20-22, 2012</p> <p align="right">Dec 13, 2012</p> <p align="right">2009-2012</p> <p align="right">May 03-13, 2010</p>
<i>Subtotal In-Depth Studies</i> 9.8 credits*	
4) Personal development <ul style="list-style-type: none"> ▶ Skill training courses Supervising Msc Students Techniques for Writing and Presenting a Scientific paper Data Management Course, Wageningen Business School Data Management Course, Wageningen Business School ▶ Organisation of PhD students day, course or conference ExPeCtationS days 2011 ▶ Membership of Board, Committee or PhD council Board Member of PhD Council Board Member of PhD Council 	<p align="right"><i>date</i></p> <p align="right">Jun 17-18, 2010</p> <p align="right">Dec 06-09, 2011</p> <p align="right">Dec 14, 2011</p> <p align="right">Mar 09, 2012</p> <p align="right">2011</p> <p align="right">2010</p> <p align="right">2011</p>
<i>Subtotal Personal Development</i> 5.3 credits*	
TOTAL NUMBER OF CREDIT POINTS*	
45,6	

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

* A credit represents a normative study load of 28 hours of study.

Artwork of the cover and invitation: María “tatica” Leandro (tatica@tatica.org). This artwork is licensed under CC-BY-SA (<http://creativecommons.org/licenses/by-sa/2.0/>).

Printed by Wöhrmann Print Service (WPS), Zutphen, The Netherlands.