Filling the gap between sequence and function

a bioinformatics approach

Joachim W. Bargsten

Thesis committee

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

Co-promotor
Dr J.P.H. Nap
Researcher
Plant Research International
Wageningen University

Other members
Prof. Dr C.J.F. ter Braak, Wageningen University
Prof. Dr G.C. Angenent, Wageningen University
Prof. Dr M.E. Schranz, Wageningen University
Prof. Dr D. de Ridder, Wageningen University

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

Filling the gap between sequence and function

a bioinformatics approach

Joachim W. Bargsten

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 Tuesday 28 October 2014
at 11 a.m. in the Aula.

Joachim W. Bargsten
Filling the gap between sequence and function: a bioinformatics approach,
182 pages.

PhD thesis, Wageningen University, Wageningen, NL (2014)
With references, with summaries in Dutch and English

ISBN 978-94-6257-076-4

Contents

Chapter 1

Introduction

1 Approaching bioinformatics

Bioinformatics has become a key discipline in modern biology (Hagen, 2003; Searls, 2010; Ouzounis, 2012). Major factors in this development have been various technological advances, allowing to create vast amounts of biological data, sometimes referred to as »big data« (O'Driscoll et al., 2013). Such data should be stored, interpreted and integrated to answer biological research questions appropriately and to generate new ideas and new research leads. As a result of these advances, biology has become a more quantitative and a much more data-driven science (Schneider and Jungck, 2013). Bioinformatics is at the interface of data, computer science and biological research. In recent years, the development and application of bioinformatics methods has led to many applications in different branches of biology, such as medicine or plant breeding. Originally defined as and aimed at the study of informatic processes in biotic systems, it has developed into computational methods for (comparative) analysis of genome and other »omics« data (Hogeweg, 2011). Bioinformatics has a dual nature, not only in the combination of biology and computer science, but also in serving as a tool for biologists on the one hand and as a separate research field, sometimes referred to as »computational biology« (Searls, 2010) on the other hand. Like in statistics, most bioinformatics approaches can be applied in multiple settings and are independent of particular species or biological models. This inherent flexibility of the tools of bioinformatics has contributed to their wide-spread use. Despite such flexibility, methods generally need to be adapted due to the particular aspects of the biological research topic under study, as well as the nature and quality of data available. Unlike human research with one organism as the central focus of attention, plant bioinformatics generally deals with different species that each present their own data, challenges and issues. This thesis presents a showcase of plant bioinformatics, with examples of genome annotation, comparative genomics, gene function prediction and the analysis of network topology for gene function prediction, for which core methods are used and developed. These topics will be outlined in the remainder of this introduction and described in more detail in the subsequent chapters. Many analyses center around two commercially interesting crops of the Solanaceae or nightshade family, tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*). The specifics of the bioinformatics for these two crops will be introduced at the end of this introduction.

2 Genome annotation

A genome sequence as such does not allow biological insight into its structure and the function of the elements it contains, such as genes, non-coding RNA or transposons. The annotation of a genome assumes proper assembly, which is not always guaranteed (Florea et al., 2011; Schatz et al., 2012). Next, it involves identifying genes, developing gene models and assigning functions to these. Proper annota-

tion is an important step for further research. Although a staggering amount of sequence data is now available, genome annotation is still a challenge (Yandell and Ence, 2012). The annotation of any genome can be divided into two distinct processes: structural and functional genome annotation. *Structural* genome annotation is the process of identifying genes, their intron-exon structures and all other components that are present in a genome. To annotate structural elements in a genome, a wealth of methods and pipelines is available. *Functional* genome annotation is the process of attaching metadata to structural annotations to enable a biologically appropriate interpretation of an identified feature. Metadata can be extremely diverse and range from the molecular function of an individual element to a complete biological pathway characterization of several elements (Yandell and Ence, 2012). Metadata tend to focus on function in the form of Gene Ontology terms (Yandell and Ence, 2012).

2.1 Structural genome annotation

Structural annotation of a genome consists of a few consecutive steps. First, all repeats are identified (RepeatMasker http://www.repeatmasker.org), in addition to other non-coding and often repetitive elements, such as rRNA, tRNA and other non-coding RNA (tRNAscan-SE (Schattner et al., 2005); RFAM (Burge et al., 2013)). Repeats and repetitive elements form a major fraction of eukaryotic genomes (Zhi et al., 2006). In the next step, all elements are marked and excluded from further analyses (masked). Unless these elements are effectively masked, subsequent gene annotations may contain portions of transposons, viruses and other disturbing elements (Cantarel et al., 2008). The masked genome sequence is the input for most structural annotation programs that predict genes. There are three main approaches in gene prediction: (*i*) ab initio, that predicts gene models solely based on probabilistic models (AUGUSTUS (Stanke and Waack, 2003); SNAP (Korf, 2004); GeneID (Parra et al., 2000); mGene (Schweikert et al., 2009)) (*ii*) evidence-based, that incorporates experimental evidence such as cDNA, EST or RNA-seq data and (*iii*) homology-based gene prediction that uses sequence information of related organisms based on evolutionary conservation (fig. 1.1). In both, evidence-based and homology-based gene prediction, sequence reads are aligned to the genomic sequence to detect exons and exon-intron boundaries and develop gene models (GMAP (Wu and Watanabe, 2005); Exonerate (Slater and Birney, 2005); BLAST (Altschul et al., 1997)). The part of the genome identified as protein-coding is translated using the appropriate translation table. With the decline of DNA sequencing costs, evidence-based genome annotation is more and more based on RNA-seq data. New annotations are incorporating RNA-seq data (Trinity (Grabherr et al., 2011); GSNAP (Wu and Nacu, 2010); TopHat (Trapnell et al., 2009)) and homology-based annotation approaches (Holt and Yandell, 2011). In the last step, the various results of different annotation programs are combined (integrated) to the structural genome annotation (fig. 1.1) (JIGSAW (Allen and Salzberg, 2005); EvidenceModeler (Haas et al., 2008)) that serves as input for the functional genome annotation.

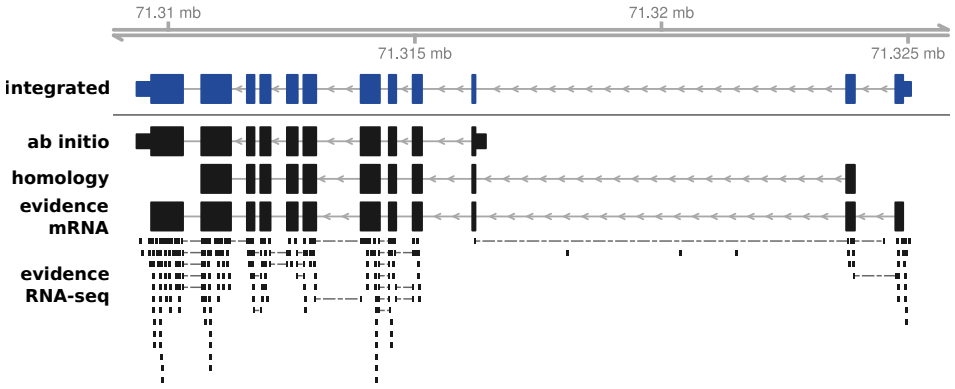


Figure 1.1: Schematic representation of the gene annotation process. Different types of predictions (ab initio, homology-based and evidence-based) are integrated into a full gene model (blue).

2.2 Functional genome annotation

Functional genome annotation is very diverse (Yandell and Ence, 2012), as »function« in a biological context is a broad concept with many layers of complexity. An integral part of such annotation is generally the prediction of protein domains (Friedberg, 2006). Identified protein domains can be matched against clusters of homologous protein domains (protein families; Punta et al., 2012), enabling (if available) the transfer of functional information. Several databases, such as Pfam (Punta et al., 2012) and InterPro (Jones et al., 2014) are available to perform this type of function transfer. An important aspect of functional genome annotation featuring in this thesis is connecting genes and proteins with functional information in the form of Gene Ontology-terms. The Gene Ontology (GO) offers structured controlled vocabulary (ontology) terms that mark the de facto standard for the annotation of function (du Plessis et al., 2011). GO consists of three separate and independent ontologies that describe in a species-independent manner domain knowledge how genes associate with biological processes (BP), cellular components (CC) and molecular functions (MF). GO is organized as a directed acyclic graph (fig. 1.2) to define relationships between terms of the ontology in a way that a computer can easily deal with (Gene Ontology Consortium, 2000). MF terms describe activities of molecules (fig. 1.2A), such as the enzyme catalytic activity or binding activity at the molecular level. MF terms generally do not represent molecules or molecule complexes, nor do they contain information about the location or the biological context of the activity described (Gene Ontology Consortium, 2000). On top of MF, BP terms give information about biological context (fig. 1.2B), e.g. »response to cold« (Gene Ontology Consortium, 2000). BP terms are a concatenation of multiple molecular events and therefore tend to describe higher, more abstract levels of function. Ideally, the GO terms associated with a gene or protein, are determined by biological experimentation, such as creating

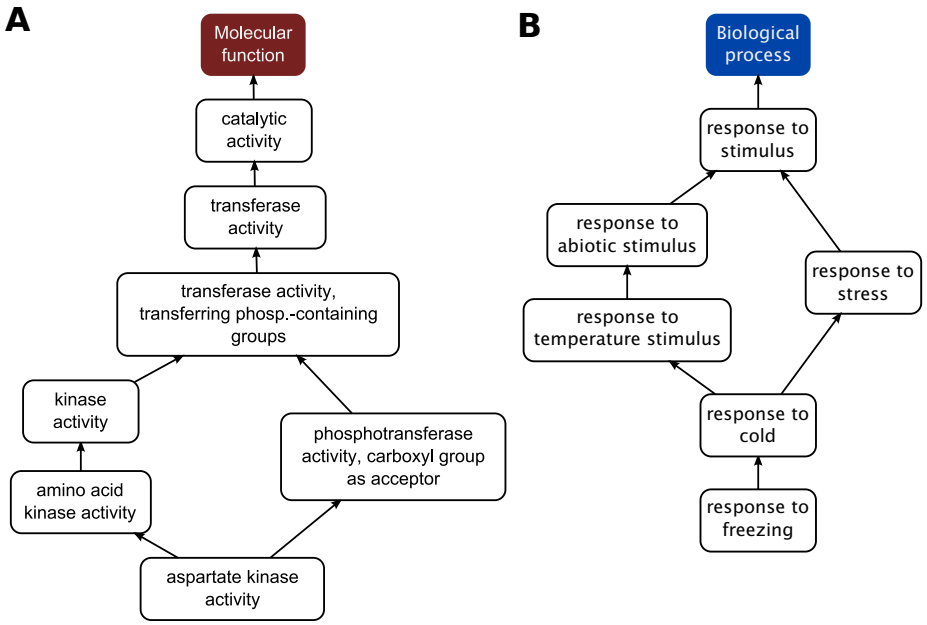


Figure 1.2: Example of graphs of a molecular function and biological process annotation. (A) molecular function. (B) biological process. The arrows indicate a »is a« relationship.

mutant phenotypes or using gene expression. Such experimentation is, however, generally time-consuming and can be challenging and/or costly, especially relative to the efforts needed to generate genomic sequence data (Lee et al., 2007; Sboner et al., 2011). As a result, there is currently a lot of genome data without any experimental backing, verification or curation (Clark and Radivojac, 2011).

The large gap between experimental annotation and available (sequence) data has motivated the development of computational functional annotation. Various approaches are put forward in the literature, such as Blast2GO (Conesa et al., 2005), BMRF (Kourmpetis et al., 2010) or Argot2 (Falda et al., 2012). Generally (and often indirectly), the function of a protein is inferred from sequence similarity (if only in part) to a protein with experimental annotation (Radivojac et al., 2013). Databases, such as UniProt (UniProt Consortium, 2014), provide the link between a sequence and functional annotation, which can be used to transfer function to an unannotated protein, given sufficient sequence similarity (homology-based transfer of function). The starting point is that sufficiently similar sequences are homologous (conserved in evolution) and are therefore likely to have similar or identical functions (Friedberg, 2006). Inferring homology from sequence similarity is not a trivial task and has been subject of debate since its discovery (Dalquen and Dessimoz, 2013). This difficulty exists partly due to the fact that, despite the correlation of sequence similarity and function, the 3D structure of a protein is neglected. Hence, function prediction methods face two types of error, the func-

tion was not predicted (false negative) or the predicted function is incorrect (false positive) (Friedberg, 2006; Furnham et al., 2012). The latter one poses one of the biggest problems in function prediction. For complete genomes, error rates for correct prediction of enzymatic function can be up to 40% depending on the type of function predicted (Lee et al., 2007; Schnoes et al., 2009). As result, an immense amount of cut-offs and methods are available to improve sequence-based function prediction. Still, a high error rate remains in current predictions and databases (Jones et al., 2007). In particular, the effect of annotation errors can be amplified by error propagation in databases. Erroneous functional annotation in a database might be used to newly annotate unannotated proteins. Applied in an iterative fashion, a chain of misannotations is created (error percolation), lowering the quality of the database (Gilks et al., 2002). Error percolation makes it difficult to discover and trace back the error. However, with the standardization by the Gene Ontology, this situation improves. One major reason is the attachment of evidence codes to annotations. Evidence codes give information about the source of the annotation, making it possible to distinguish between experimental, e.g. »inferred from mutant phenotype« (IMP) or »inferred from direct assay« (IDA), and electronically inferred, e.g. »inferred from sequence similarity« (ISS), annotations. This source of information gives the possibility to spot errors more easily. Recent research results suggest that the quality of the GO-database is increasing, despite its rapid growth (Skunca et al., 2012). Further improvements could include reliability or quality information as supplement for annotated functions.

Whereas sequence similarity is a useful proxy for MF, sequence information has lower information content for the prediction of the biological context captured in BP terms. BP terms represent more abstract functionality, namely a concatenation of multiple steps at the cellular and organismal level which is hardly contained in the sequence. Homology-based transfer of function has therefore a low prediction performance for BP terms (Radivojac et al., 2013). Experiments comparing multiple species show a performance difference of 20% between MF and BP (Nehrt et al., 2011; Altenhoff et al., 2012). To be able to predict BP terms with biological relevance sufficiently accurately, additional data with other information context is required. Biological networks are such an additional data source (Sharan et al., 2007). It has been shown that network data contain information about biological process-related protein function and therefore help to improve the performance of function prediction algorithms (Sharan et al., 2007; Vital-Lopez et al., 2012), including guilt-by-association (Oliver, 2000) or BMRF (Kourmpetis et al., 2010). Network data can be created from different data sources, might contain time-point information or be computationally inferred. In the context of function prediction, networks are generally created from experimental data, for example from yeast two-hybrid experiments, co-expression data or direct protein-interaction measurements (Sharan et al., 2007).

In general, a network (also called a graph) is a set of components (nodes) that are connected by links (edges) (fig. 1.3). In biological networks, these components represent genes, proteins, or, more abstract, molecules (Barabási and Oltvai,

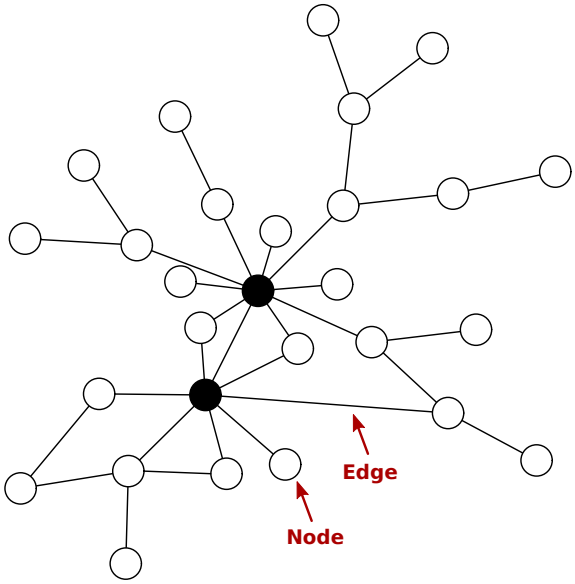


Figure 1.3: Example of a biological network with scale-free topology. Proteins (nodes) that interact with each other share a connection (edge). Few nodes (hub nodes; black filled circle) have a high amount of edges, whereas the majority of proteins (empty circle) possesses a small amount of edges.

2004; Zhu et al., 2007). In the context of function prediction, nodes correspond almost exclusively to proteins. To perform a function, proteins often form complexes or interact with other proteins. This concept can be transferred naturally to biological processes, making biological networks a perfect match for BP prediction. By using experimentally characterized interaction partners, the function of an uncharacterized protein can be predicted (Jansen et al., 2003; Barabási and Oltvai, 2004; Hu et al., 2010). On a global level, proteins fall into two classes, proteins with pre-existing annotations and unannotated proteins. Statistical frameworks require these pre-existing annotations (training data) to propagate the functional annotation on the basis of the connections in the network to unannotated proteins (Sharan et al., 2007; Pavlidis and Gillis, 2012). The framework explored and augmented in this thesis is the Bayesian Markov Random Fields (BMRF) approach, that was specifically developed for network-based prediction of protein function (in terms of BP) (Kourmpetis et al., 2011). BMRF requires initial training data to perform function prediction. The initial training data can consist of (experimental) BP annotation from e.g. the Gene Ontology. Given initial experimental annotation and network information, BMRF calculates the probability of a protein or gene belonging to a BP term. This thesis presents multiple new approaches to improve the performance of BMRF, notably by addressing the quality of the input (chapters 4 and 5).

3 Comparative genomics

The goal of comparative genomics is to give insight into functions of genomic elements and evolution of organisms. Both aspects, function and evolution, are connected by the central paradigm that conserved genomic elements are functionally important (Koonin and Galperin, 2003; Alföldi and Lindblad-Toh, 2013). The opposite is, however, not necessarily true. Genomic elements can have a (biological) function without being conserved (Alföldi and Lindblad-Toh, 2013). Two or more genomic elements are called (evolutionary) conserved if their sequences show significant similarity (Durbin et al., 1998), which is used to establish homology (Koonin and Galperin, 2003). With the availability of completely sequenced genomes, it becomes possible to compare the sequences to detect conserved elements, unravel their evolutionary relationships and pinpoint differences in genomes, such as losses, duplications or rearrangements. These steps are in essence the definition of comparative genomics. Once compared, inferences about function and detection of unique genomic elements are possible. Genomic elements can be very diverse and commonly comprise transcribed protein coding and non-protein coding sequences, *cis*-acting elements and chromatin structures (Zheng et al., 2004). Comparative genomics can be applied at many different levels, starting at the level of a single individual to large populations, multiple species or even to the tree of life (Brown, 2007). Even though comparative genomics provides useful tools in understanding function, it is limited by the evolutionary distance of and the lack of knowledge on organisms that are compared. Comparative genomics loses its power when comparing very distant organisms. It becomes difficult to detect conserved elements and their functions are likely to be different (Stojanovic, 2007). But also for closely related species, the transfer of function requires an existing body of experimentally studied genomic elements. The lack of experimental data leads to a large amount of conserved elements without any functional information (Galperin and Koonin, 2010). Yet, the growing number of sequenced genomes makes comparative genomics more powerful to analyze the evolutionary history of genomes. In particular in the analysis of the plant genomes, comparative genomics has become an important tool. Notable highlights are the breeding for traits such as resistance or yield (Krieger et al., 2010; Ranjan et al., 2012) and tracing back the history of plant domestication (Morrell et al., 2011).

3.1 Concepts of evolutionary genomics

A large part of comparative genomics is devoted to elucidating evolutionary relationships between genomic elements and species. Evolutionary relationships can be used to make inferences about function, but they are also of interest by themselves (Koonin and Galperin, 2003). The most basic relationship between two genomic elements is homology, denoting »common decent« of two entities. Homologs are categorized into orthologs and paralogs. Orthologs are related via speciation, whereas paralogs are related via duplication (Koonin, 2005). Orthologs typically

retain the same function following speciation, while paralogs are likely to diverge with new functions through point mutations and domain recombinations (Chen et al., 2007; Altenhoff et al., 2012). Hence, orthology and, on a finer grained level, protein domains are widely used to infer (molecular) function (Kuzniar et al., 2008; Engelhardt et al., 2011; Kristensen et al., 2011; Hunter et al., 2012). With the concept of orthology and paralogy, it becomes possible to outline the evolution of genes and gene families in multiple species. The relationships between genes of different organisms are commonly represented by a phylogenetic tree. To construct a phylogenetic tree, several basic steps are necessary (*i*) multiple sequence alignment (MAFFT (Katoh and Standley, 2013); Clustal Omega (Sievers et al., 2011)) (*ii*) manual curation of the alignment (*iii*) estimation of the phylogenetic tree (RAxML (Stamatakis, 2006); PhyML (Guindon et al., 2010); MrBayes (Ronquist and Huelsenbeck, 2003)) (*iv*) visualization (Archaeopteryx (Han and Zmasek, 2009); Dendroscope (Huson and Scornavacca, 2012)). The reconstructed tree allows insights into the evolution of genomic complexity and lineage-specific adaptations (Koonin, 2005; Guo, 2013) with implications for (molecular) gene function (De Smet and Van de Peer, 2012). More importantly, phylogenetic trees allow a detailed reconstruction of evolutionary distances, gene losses and duplications.

Studying complete gene families and their variation throughout the plant kingdom, lineage-specific developments and the lack thereof, can give useful hints towards the evolutionary mechanism behind gene family development (De Smet and Van de Peer, 2012; De Smet et al., 2013). An example is the Snf2 subfamily DRD1. The DRD1 subfamily shows a high variation throughout the plant kingdom. This high variation might play a role in species-specific adaption of stress response to the environment. In particular plant genomes that were subject to large scale duplications and losses, provide a valuable resource for this kind of studies (chapter 3). These studies contribute to the ultimate goal to elucidate the relationship between genotype and phenotype (Brown, 2007).

3.2 Synteny and chromosomal rearrangements

Originally, synteny was used to denote genes that remain on the same chromosome within or between organisms. Nowadays, synteny has shifted its meaning to specify regions of common evolutionary ancestry (paralogous or orthologous regions), leading to an ambiguous usage (Passarge et al., 1999). Despite its ambiguity, synteny analyses are frequently performed to study genomes. Here, (shared or conserved) synteny refers to two or more homologous genes that reside on the same chromosome in two or more species, following earlier definitions of Nadeau (1989) and Ehrlich et al. (1997). Synteny in itself is not restricted to the same gene order, making it necessary to introduce »collinearity« as additional concept. Collinearity indicates that syntenic genes also show the same ordering on the chromosome when compared between organisms (Coghlan et al., 2005; Tang et al., 2008a; Wang et al., 2012a). Differences in synteny and/or collinearity are termed chromosomal rearrangements and can be categorized in inversions, translocations, duplications or losses (Van de Peer, 2004). The concepts of synteny and collinearity are mostly

applied on a large scale, i.e. to analyze complete genomes and no single genes. In particular for the analysis of plant genomes, these concepts are extremely useful. Plant genomes have an enormous diversity and can differ to a factor of thousand in genome size. To effectively study larger genomes, knowledge about chromosomal rearrangements that can relate well-studied small genomes with unstudied large genomes is paramount (Bowers et al., 2003; Tang et al., 2008a). Various software applications facilitate the analysis of chromosomal rearrangements. Basic steps include (*i*) sequence alignment (BLAST (Altschul et al., 1997); LASTZ (Harris, 2007); MUMmer (Kurtz et al., 2004)), (*ii*) detection of homologous genes or regions (MUMmer; reciprocal best BLAST hits (Tatusov, 2001; Kuzniar et al., 2008); InParanoid (Remm et al., 2001); OrthoMCL (Li et al., 2003)) and (*iii*) clustering of homologous parts to syntenic and collinear segments (DAGchainer (Haas et al., 2004); MCscan (Wang et al., 2012a); ADHoRe (Vandepoele, 2002)). With these steps completed, the genomes of two or more organisms can be compared and their relationships can be deciphered.

4 Plant bioinformatics: from model species to actual crops

Since the publication of the *Arabidopsis thaliana* genome sequence in 2000 (Arabidopsis Genome Initiative, 2000), plant biology has undergone significant changes. A wealth of sequenced plant genomes – more than 80 at this point – has become available. However, the experimental annotation of function is lacking far behind. This situation is similar to non-plant bioinformatics, but the focus on a multitude of species, either as model, crop or both, poses special challenges for plant bioinformatics. In particular, the research on angiosperms (flowering plants) is a complex task. Angiosperms show remarkable differences in genome size, resulting from whole genome duplications and large scale gene losses. Arabidopsis, for example, has undergone three genome duplications accompanied by heavy gene loss (Simillion et al., 2002). In addition to these large rearrangements, numerous small scale rearrangements and repetitive elements were induced by mobile elements. All these events led to a severe genome reshuffling, obscuring evolutionary traces (Tang et al., 2008a). As result, 80% of a plant's genome may consist of repetitive elements (Brenchley et al., 2012; Kim et al., 2014). With the study of synteny between organisms, rearrangement events can be traced back. Often such rearrangements can prevent proper pairing during meiosis and therefore support reproductive isolation, eventually leading to speciation (Widmer et al., 2009). Particularly in breeding, the knowledge about rearranged regions plays an important role in introgression of favorable traits from a wild cultivar into a crop plant. Thus the delineation of chromosomal rearrangements may not only shed light on genome diversity and evolution, but also contribute to novel strategies for modern plant breeding (chapter 2). To progress further in plant breeding, not only rearrangements are of interest, but also the functional aspect, i.e. connecting the genotype with the phenotype.

Despite the rapid growth of published plant genomes (Goodstein et al., 2012), genome annotation and prediction of function stay challenging in plants, as they are equally affected by rearrangements and reshuffling throughout evolution. In addition, accurate prediction of the function of genes or proteins is hindered by the lack of experimental data. This lack of experimental data arises from rapid growth of sequenced genomes on the one side and slow progress of experimental validation on the other side. As consequence, the quality of structural and functional annotation is affected. The annotation quality and coverage varies considerably between different plants. Model plants, such as Arabidopsis (*Arabidopsis thaliana*) or rice (*Oryza sativa*), have a relatively high amount of experimental data and consequently a high quality annotation. Non-model-plants are often lacking experimental data (Yandell and Ence, 2012) (fig. 1.4). Including all species, ~98% of all functional annotations are computationally inferred (du Plessis et al., 2011). Plants have a similar percentage (~97%). After rice, which is the second-best annotated plant with less than 2000 MF and BP annotations, annotation coverage is declining fast. Thus, from a bioinformatics perspective, providing reliable function predictions is paramount. One way to improve this situation is to incorporate network data. Also here Arabidopsis is the most studied plant species, covering ~66% of the proteins. Networks can be obtained from STRING (Franceschini et al., 2013), BioGRID (Chatr-Aryamontri et al., 2013) and other sources (Brandão et al., 2009; Arabidopsis Interactome Mapping Consortium, 2011; Mutwil et al., 2011; Orchard et al., 2014). Other plant species are not covered extensively. STRING, as most exhaustive resource, covers 11 plant species (fig. 1.5); nearly all plant network resources are focused on Arabidopsis (Braun et al., 2013). This situation requires generating network data from in-house experiments or public raw data archives, such as the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). Alternatively, it is feasible to use comparative approaches to transfer network data from Arabidopsis (Mutwil et al., 2011). Most algorithms – also BMRF – need initial training (seed) data to perform function prediction (Hastie et al., 2003; Kourmpetis et al., 2011). Since experimentally verified seed data is sparse in nearly all plants, we combined BMRF with a sequence-based function prediction algorithm. This setup allowed us to create seed data for BMRF and perform network-based function prediction, even though the species' experimental annotation was sparse. To be able to predict for sparsely annotated species is crucial when applied to newly sequenced or non-model plants.

One reason for the sparse annotation in plants is that, in terms of plant breeding, many plants are (commercially) interesting by themselves and not a proxy for a central model organism. Examples are the vegetable crops tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*), the principal crops featuring in this thesis. They belong to the Solanaceae or nightshade family, which includes more than 3000 species (chapter 2). The family is economically one of the most important, accounting for 10% of the worldwide gross production value of crop plants in 2011 (http://faostat.fao.org). In terms of vegetable crops it ranks number one, with tomato and potato as most abundant representatives, followed

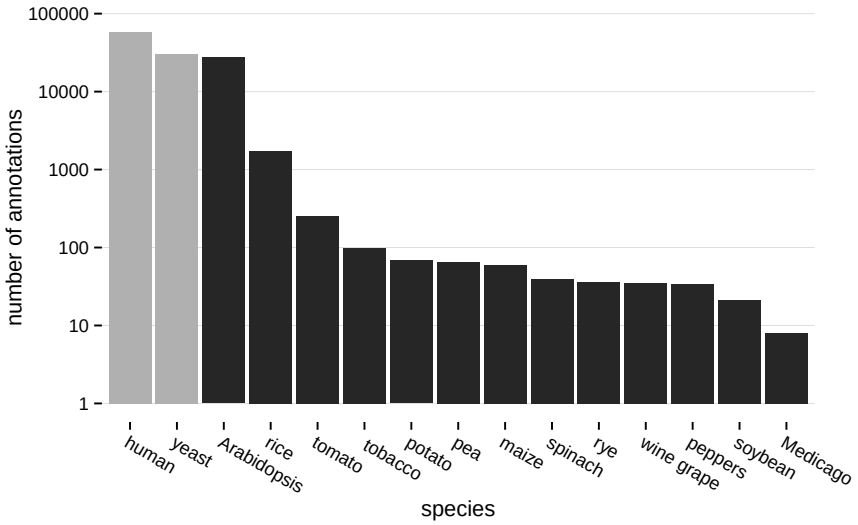


Figure 1.4: Overview of experimental annotation in selected species. Listed are all experimental annotations (molecular function and biological process) of the GO-database. Human and yeast (gray) are shown as reference.

by pepper and eggplant (Foolad, 2007; Xu et al., 2011). The Solanaceae show remarkable adaptability to diverse climatic conditions, ranging from wet rainforests to dry and arid environments, and exhibit a huge phenotypic diversity from tiny annual herbs to large forest trees (Knapp, 2002). With the advent of whole genome sequences of potato and tomato, bioinformatics allows comparison on different levels, such as genome-wide comparisons, gene family analyses and protein function assessments. Such comparisons are the prerequisite for exploring and exploiting the differences and similarities of this plant family in future breeding.

Both plants are similar in genome size. Potato and tomato have a size of approx. 850 megabases (Mb) and approx. 900 Mb, respectively (Xu et al., 2011; Sato et al., 2012). They contain mostly large collinear regions, disrupted by several large and multiple small inversions. Overall, they show a nucleotide divergence of 8% (Sato et al., 2012). Despite the small difference in genome sequence, both plants show remarkably different phenotypic traits, such as fruit size and tuber production (Xu et al., 2011; Sato et al., 2012). The genome sequences enable to connect phenotypic properties back to the genome. Expansion of gene families to introduce new protein functionality is a common phenomenon and Solanaceae are no exception (Xu et al., 2011; Sato et al., 2012; Guo, 2013). For example, homologs of the flowering-inducing gene FLOWERING LOCUS T evolved into key players of tuber formation in potato (Abelenda et al., 2014). In this thesis we investigate the expansion of the stress-related Snf2 gene family with its unique development in tomato and potato (chapter 3). Due to their close relationship and economic importance, tomato and potato provide a unique opportunity to study

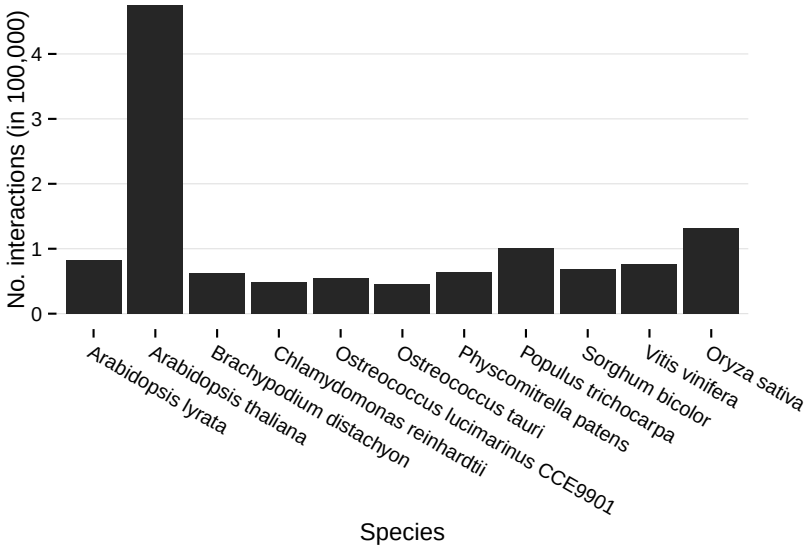


Figure 1.5: Number of plant-related interactions in the STRING database. In total, 11 plants are represented in STRING. Arabidopsis possesses the highest number of interactions, covering approx. 66% of its proteome. Other plants have a significant lower number of interactions in STRING.

evolutionary relationships of gene families and connections to (complex) traits. In addition, both plants can function as models for flowering plants of the asterid clade. The asterid clade represents 25% of all flowering plants (Xu et al., 2011) and tomato and potato might provide a platform for further fundamental and applied research in this clade.

Both plants have been annotated by the respective genome sequencing consortia. In case of tomato, the iTAG (International Tomato Annotation Group) used an in-house pipeline, incorporating ab initio, evidence-based and homology-based methods. All results were integrated with Eugene (Foissac et al., 2008) and manually curated, resulting in 34,727 gene models. Even though no comprehensive evaluation is available, the structural iTAG annotation of tomato is considered to be of high quality (Sato et al., 2012). Similar to iTAG, the PGSC (Potato Genome Sequencing Consortium) used an in-house pipeline to annotate the potato genome. The basic steps are the same as in the structural tomato annotation, but in terms of complexity, with complexity translating to the number of tools, data sources and internal validation used, the PGSC pipeline is much simpler. In total, 39,031 gene models were predicted. Due to inconsistencies between different pipelines, the iTAG decided to redo the potato annotation, allowing a direct comparison of potato, tomato and their structural annotations. The potato annotation conducted by iTAG resulted in 35,004 gene models, 5027 models less than the PGSC annotation. With Arabidopsis (TAIR10; Lamesch et al., 2012) as reference, the iTAG

potato annotation matched 92% of the predicted gene models to Arabidopsis gene models, whereas the PGSC annotation matched only 69% (Sato et al., 2012). This result clearly indicates differences. However, the rate of correctly predicted annotations can only be determined experimentally. Neither the iTAG potato annotation, nor the PGSC potato annotation has been evaluated in this respect. Despite extensive research on tomato and potato, the amount of experimentally verified functional annotation available in the UniProt-GOA database (Dimmer et al., 2012) is negligible. Tomato has less than 500 and potato less than 100 annotations, MF and BP aggregated. Computationally inferred functional annotations are currently exclusively performed via sequence-based transfer of function. A significant improvement can be expected by incorporating complementary data, such as RNA-seq derived network data.

5 Outline of this thesis

The research presented in this thesis aims to shape and develop the approaches in plant bioinformatics for computational function prediction. In chapter 2 the structural homology is presented in euchromatin regions of tomato, potato and pepper with special attention to the long arm of chromosome 2. It shows that the local gene vicinity is largely preserved, despite many small-scale synteny perturbations. These results indicate a high frequency of chromosomal rearrangements accompanying the evolution in the *Solanum* genus.

The adequate identification of chromosome organization is, among others, required for the efficiency and success of introgressive hybridization breeding. In the near future, technological advances in sequencing technology will allow sequencing large numbers of complex genomes relatively fast and cheaply. This will undoubtedly speed up identification of compatible genomes for introgression breeding, the rearrangement phylogeny within the Solanaceae, and reconstruction of the ancestral *Solanum* karyotype. The results described in chapter 2 are a first step in mining of structural genetic diversity towards the development of genome-based breeding tools. Chapter 3 surveys the Snf2 gene family in the plant kingdom. Members of the Snf2 gene family can affect (a)biotic stress response in plants via chromatin remodeling. The Snf2 gene family shows high variation across the plant genomes analyzed with unexpected expansions of the DRD1 subfamily in the tomato genome. The results point towards a novel role of DRD1 members in developmental or stress regulation in tomato. Chapter 4 explores a new way of combining the protein function prediction methods BMRF and Argot2 to gain additional performance in plants that are sparsely annotated. Newly sequenced and non-model plants often lack experimental annotation, required to perform accurate function predictions. The approach of combining a sequence- and network-based method is able to supply and improve function predictions in such environments. Chapter 5 shows that removing proteins from a protein-protein interaction network can improve the prediction performance of the network-based function prediction algorithm BMRF. Results show that highly connected (hub) proteins can impede

function prediction performance. These proteins tend to connect functionally different network modules, which results in additional noise. As a consequence, the removal of hub proteins increases the signal and improves function prediction performance. Chapter 6 discusses the implications of this work on plant science and outlines the perspective for future research.

Chapter 2

Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper

Abstract

We have analyzed the structural homology in euchromatin regions of tomato, potato and pepper with special attention for the long arm of chromosome 2 (2L). Molecular organization and collinear junctions were delineated using multi-color BAC FISH analysis and comparative sequence alignment. We found large-scale rearrangements including inversions and segmental translocations that were not reported in previous comparative studies. Some of the structural rearrangements are specific for the tomato clade, and differentiate tomato from potato, pepper and other solanaceous species. Although local gene vicinity is largely preserved, there are many small-scale synteny perturbations. Gene adjacency in the aligned segments was frequently disrupted for 47% of the ortholog pairs as a result of gene and LTR retrotransposon insertions, and occasionally by single gene inversions and translocations. Our data also suggests that long distance intra-chromosomal rearrangements and local gene rearrangements have evolved frequently during speciation in the *Solanum* genus, and that small changes are more prevalent than large-scale differences. The occurrence of sonata and harbinger transposable elements and other repeats near or at junction breaks is considered in the light of repeat-mediated rearrangements and a reconstruction scenario for an ancestral 2L topology is discussed.

Bargsten[*], J. W., Peters[*], S. A., Szinay, D., van de Belt, J., Visser, R. G. F., Bai, Y., and de Jong, H. (2012). *Plant Journal*, 71(4):602–614. ([*]These authors contributed equally)

1 Introduction

The Solanaceae or nightshade family is a large group of more than 3000 species that includes tuber or fruit-bearing vegetables (tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), pepper (*Capsicum annuum*) and eggplant/aubergine (*Solanum melongena*)), and plants of horticultural (petunia (*Petunia hybrida*)) and medicinal (tobacco (*Nicotiana tabacum*)) importance (Knapp, 2002; Sesso et al., 2003). The family is economically the third most important, and ranks number one in terms of vegetable crops (Foolad, 2007). The Solanaceae show remarkable adaptability to diverse climatic conditions, ranging from wet rainforests to dry and arid environments, and exhibit a huge phenotypic diversity from tiny annual herbs to large forest trees (Knapp, 2002). In contrast, cultivated *Solanum* and *Capsicum* crops have a strikingly narrowed genetic basis through domestication, resulting in loss of desirable traits, including those that confer (a)biotic stress tolerance. In addition, rapidly changing climate conditions and increasing competing claims for arable lands will increase the demand for new varieties that tolerate harsh environmental conditions, confer resistance against pathogens and at the same time have better productivity and nutritional quality. The need to compensate for such genetic losses requires introgression of alien chromatin from wild relatives to the crops, a process referred to as introgressive hybridization. In general, the wild relatives of solanaceous crops provide a gene pool that is sufficiently rich for crop improvement. However, their use as donor in introgressive breeding is limited (Rick et al., 1987; Singh, 2006; Bai and Lindhout, 2007). Crossing barriers and linkage drag are well-known phenomena that limit the use of germplasm for introgressive hybridization (Rieseberg and Willis, 2007; Bedinger et al., 2011).

The transfer of alien chromatin containing the genetic information for a desirable trait depends on homeologous recombination between the donor chromosome and its corresponding counterpart in the crop, which, amongst other factors, is determined by their level of collinearity. Severe problems may occur in those cases where the donor chromosome and its homeolog differ as a result of large-scale rearrangements (inversions or translocations). Heterozygosity for such rearrangements may lead to failure of synapsis and/or illegitimate crossovers at meiosis. As a consequence, genes are unlikely to recombine, and so are transmitted as a single locus, a phenomenon known as linkage drag. To facilitate identification of compatible donor species or genotypes for error-free homeologous introgression of important agronomic traits such as (a)biotic stress tolerance, elucidating the genome organization is imperative. This is usually accomplished by analyzing collinearity, synteny and linkage at both the chromosome and gene level.

The synteny concept was introduced in 1971, and pertains to the preserved co-location of homologous genes on chromosomes between species, irrespective of genetic linkage and gene order (Ehrlich et al., 1997; Passarge et al., 1999). Conservation of both synteny and order of homologs determine conserved linkage of genes. Both synteny and conserved linkage have been used to investigate solanaceous genome organization, gene diversification and evolutionary ancestry (Ku

et al., 2000; Fulton, 2002; Wu et al., 2006; Wang et al., 2008). The identification of ancestral relationships between homologous genes and their distinction into orthologs and paralogs has become more feasible with the advent of high-throughput sequencing that facilitates comparisons of entire genomes. Until now, sequence comparisons for conserved syntenic segments in *Solanum* have mainly been obtained for relatively small orthologous regions. In general, the order and sequence of orthologs was found to be conserved, despite a few small-scale differences and positive gene selections (Doganlar et al., 2002b; Wang et al., 2008).

Analysis of chromosome structure in Solanaceae has been based on several lines of research. For example, light microscopy observations on pachytene chromosomes of F_1 hybrids showed normal synapsis along the chromosomes. Furthermore, linkage maps of intra- and inter-specific hybrids were found to be largely collinear (Pertuzé et al., 2002; Chetelat and Ji, 2007; Moyle, 2008). Both similarity in chromosome morphology and marker collinearity supported the notion that *Solanum* species have evolved primarily by genic change rather than by large-scale chromosomal rearrangements. Nonetheless, genetic linkage analyses indicated that tomato and potato are differentiated by a series of whole-arm inversions of chromosomes 5, 9, 10, 11 and 12 (Bonierbale et al., 1988; Tanksley et al., 1988; Livingstone et al., 1999; Doganlar et al., 2002a; Pertuzé et al., 2002). Furthermore, electron microscopy studies on somatic hybrids of tomato and potato (de Jong et al., 1993) and F_1 hybrids from inter-specific crosses revealed substantial changes in chromosome structure among *Solanum* species (Anderson et al., 2010). Most of these structural changes, however, were found in the heterochromatin, with comparatively few genes and low recombination, and thus would have little effect on the collinearity of linkage maps (Anderson et al., 2010).

Although genetic mapping studies provided valuable starting points for unraveling plant genome organization, they are inaccurate in regions where crossover recombination is suppressed or even absent, as in the distal heterochromatin and the large pericentromere regions, for example. Furthermore, there are insufficient DNA polymorphisms for simple markers that are locus-specific across species, causing markerless gaps in linkage maps that could leave chromosome rearrangements undetected. Such problems were recognized within the framework of the tomato genome sequencing project (Szinay et al., 2008; Peters et al., 2009). For example, integrated mapping revealed genetic intervals comprising hundreds of tomato genes in euchromatic regions with a marker coverage insufficient to support microsynteny analysis. Some studies used a sequence-based comparison for microsyntenic analysis in Solanaceae (Fulton, 2002; Van der Hoeven, 2002; Datema et al., 2008; Wang et al., 2008), but genome-wide comparative sequence analysis is still limited for *Solanum*, as comparable sequence clades have not yet emerged. Alternatively, genome-wide cross-species fluorescence in situ hybridization (FISH) has provided a foundation for comprehensive comparative maps, as well as rapid and reliable detection of genetic elements that are associated with traits of interest. For example, FISH has been used to analyze the organization of the short arm of chromosome 6 (6S) in tomato and potato, which contains the *Mi* resistance homolog cluster. An

additional inversion was revealed on 6S (Iovene et al., 2008; Tang et al., 2008b) that was not reported in synteny studies with molecular markers (Tanksley et al., 1992; Grube et al., 2000), probably due to lack of marker coverage and suppression of recombination (Liharska et al., 1996; Bai et al., 2004; Seah et al., 2004).

Recently, the complete genome sequences of both tomato and potato have become available, and this allowed us to take advantage of a combined cytogenetic-based macrosyntenic approach and a comparative sequence-based microsyntenic approach. Here, we address a number of issues including content and gene organization in *Solanum* chromosomes. In addition, we present a detailed analysis of chromosome rearrangements in tomato (*S. lycopersicum*), potato (*S. tuberosum*) and pepper (*Capsicum annuum*), with special attention to regions containing stress tolerance and disease resistance homologs, and discuss the rearrangements in the light of chromosome evolution.

2 Materials and Methods

2.1 Scaffold selection

S. lycopersicum Heinz 1706 (build 2.40) and *S. tuberosum* DM scaffolds were obtained from ftp://ftp.solgenomics.net/tomato_genome/wgs/assembly and http://potatogenomics.plantbiology.msu.edu, respectively. The scaffolds of tomato and potato were aligned against genetic and physical maps from the Sol Genomics Network (http://solgenomics.net) to verify the chromosomal location. Genome sequences were aligned using MUMmer 3.22 (http://mummer.sourceforge.net) with the tomato scaffold as a reference. Subsequently, the coords file output generated by the MUMmer script »NUCmer« was converted into ClustalW format and used as input for the generic synteny browser GBrowse_syn (http://gmod.org/wiki/Synteny) for visualization and further analysis. Tomato Heinz 1706 BACs and potato BAC sequences from the RH clone library were aligned to scaffolds using blastn.

2.2 BAC selection, growth and DNA preparation

For solanaceous species-derived stress tolerance genes, a genomic location was determined by best blastn hits against tomato scaffolds from the Sol Genomics Network (SGN; http://solgenomics.net). Translated coding regions from non-solanaceous species-derived stress tolerance genes were used in a tblastx screen against tomato unigene sequences. Unigenes were then used for a blastn screen against tomato scaffolds to determine sequence coordinates. Subsequently, BAC clones containing stress tolerance gene homologs were identified by blastn of a scaffold sequence interval against a tomato BAC end sequence database, and linked to the EXPEN2000 genetic map by screening the genomic intervals against the genetic marker database. Subsequently, BAC clones of tomato cv. Heinz 1706 *Hin*dIII, *Eco*RI and *Mbo*I libraries were grown, and BAC DNA was isolated as described previously (Peters et al., 2009).

2.3 Chromosome preparations

All FISH experiments were performed on tomato cv. Heinz 1706 ($2n = 2x = 24$). Pachytene preparations from young anthers containing pollen mother cells and spreads of extended DNA fibers from young leaves were made as described by Zhong et al. (1996) and Budiman et al. (2004).

2.4 Fluorescence in situ hybridization (FISH)

Two-color and multi-color FISH of BAC clones to pachytene chromosomes were performed as described by Zhong et al. (1996). Slides were examined under an Axioplan 2 imaging photomicroscope (Zeiss, Jena, Germany) equipped with epi-fluorescence illumination and small band filter sets for DAPI (4',6-diamino-2-phenylindole) and for FITC (fluorescein-5-isothiocyanate), Cy3 (cyanine 3), Cy5 (cyanine 5), DEAC (7-diethylaminocoumarin-3-carboxylic acid) and Cy3.5 (cyanine 3.5) fluorescence. Capturing of selected images and image processing were performed as previously described (Szinay et al., 2008).

2.5 Sequence annotation

Interspersed repeats were identified through similarity searches to the *Magnoliophyta* section of the Repbase repeat database (release 2008-08-01) (Jurka et al., 2005) and the Institute for Genomic Research *Lycopersicon* repeats version 3.1 and Solanaceae repeats version 3.1 (currently available via http://plantrepeats.plantbiology.msu.edu) using RepeatMasker 3.2.5 (http://www.repeatmasker.org) and cross_match 0.990319 (http://www.phrap.org). In addition, LTR retrotransposons were predicted ab initio using LTR Finder (Xu and Wang, 2007; http://tlife.fudan.edu.cn/ltr_finder). Ab initio gene prediction on the repeat masked sequences was performed using Genscan (Burge and Karlin, 1997) using *Arabidopsis thaliana* gene models. Alignments of tomato and potato ESTs and genetic markers were generated using blastn 2.2.17 (Altschul et al., 1997). Tomato and potato EST sequences and marker sequences were downloaded from the Sol Genomics Network (http://solgenomics.net).

2.6 Orthology analysis

For ortholog detection, tomato ITAG annotation 2.31 (http://solgenomics.net/gbrowse/bin/gbrowse/ITAG1_genomic) and potato PGSC annotation 3.4 (Xu et al., 2011) were used to obtain the sequences of the predicted proteins. In each case, only the longest transcripts were taken into consideration. Reciprocal best hits were obtained using blastp (Altschul et al., 1997) and results were clustered into orthologous groups by InParanoid version 4.1 (Remm et al., 2001) with a score cut-off of 40. Further processing of annotation data to fit the coordinates on the selected segment was performed using custom Perl and R scripts in combination with BioPerl (http://www.bioperl.org/wiki/Main_Page).

The coordinates and identity of ortholog group members, together with the mapped positions of BAC sequences, LTR retrotransposons and unigenes and coordinates for mapped annotated proteins were subsequently stored in a gff format for visualization of the local gene repertoire with SynBrowse (Pan et al., 2005).

The predicted orthologs were clustered into ortholog groups and subsequently classified into genes with conserved or disrupted linkage based on the linear order of orthologs.

3 Results

3.1 Cytogenetic macrosynteny between tomato, potato and pepper

Nucleotide sequences and translations from annotated coding regions of 32 genes that have been implicated in stress tolerance (Jenks et al., 2007) were collected from the Genbank and Refseq sequence databases (http://www.ncbi.nlm.nih.gov), and were used for blastn- or tblastx-based similarity searches. We identified 19 stress tolerance gene homologs in 19 tomato BACs, and verified the cytogenetic mapping position on tomato and potato chromosome 2, 6, 10 and 11 pachytene complements using BAC FISH. Eight BACs, of which six contain a stress tolerance homolog, display a single clear fluorescent signal on chromosome 2. This chromosome can easily be distinguished from the other chromosomes by its acocentric structure and the large nucleolar organizing region. Previously, cytogenetic studies of tomato pachytene chromosomes revealed long continuous stretches of less condensed euchromatin in both chromosome arms, flanked by highly condensed heterochromatin at the telomere ends and the centromeres (Ganal et al., 1991; Jong et al., 2000; Chang et al., 2008). Based on these morphological properties, the foci observed were in the euchromatic part on the long arm of tomato chromosome 2 (2L) (fig. 2.1). Genetically, these BACs have been mapped at an interval between 82 and 143 cM based on blastn hits to SGN EXPEN2000 markers (http://solgenomics.net/cview/map.pl?map_version_id=52). The genetic map order of all BACs is consistent with the linear cytogenetic mapping order, except for H146O6, which appears to be inverted (figs. 2.1 and 2.S1). Using common markers, we mapped the tomato BACs to the euchromatic portion of potato 2L (see below) in an interval between 33 and 50 cM on the potato TXB genetic map (Koo et al., 2008). Strikingly, the cytogenetic mapping order on potato and tomato pachytene is clearly different, suggesting multiple rearranged segments. On potato and pepper chromosome 2L, the BAC FISH mapping order appears collinear (fig. 2.1).

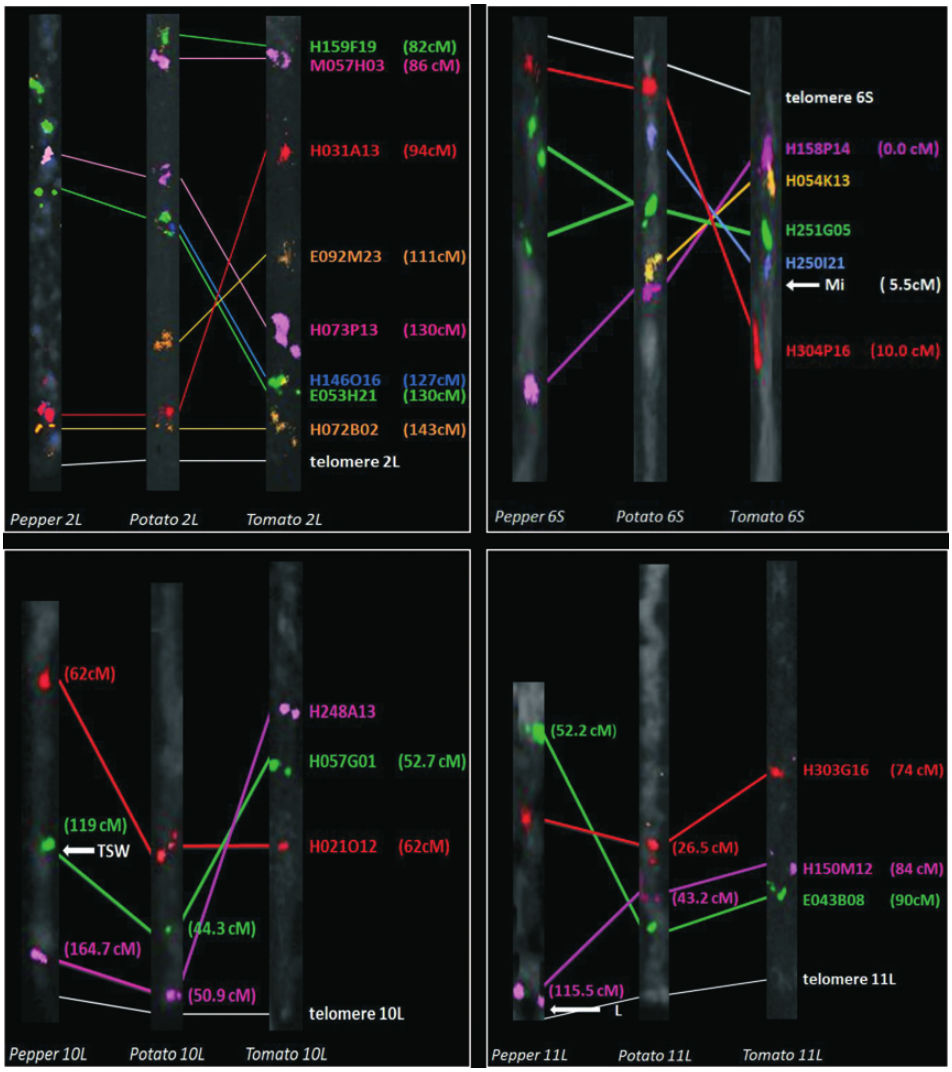


Figure 2.1: Comparative FISH analysis of tomato BACs on tomato, potato and pepper 2L, 6S, 10L and 11L. Identifiers for tomato BACs are indicated on the right of each pachytene chromosome. White arrows indicate the positions of *Mi*, *L* and *TSW* loci. Corresponding cytogenetic mapping positions of *Hin*dIII (H), *Eco*RI (E) and *Mbo*I (M) tomato BACs, and telomere positions on tomato, potato and pepper pachytenes are shown by coloured connecting lines and white connecting lines, respectively. The positions of SGN markers from the tomato EXPEN2000, potato TXB1992 and pepper AC99 genetic maps (http://solgenomics.net/cview/map.pl?map_id=11) associated with the tomato BACs are indicated (cM).

In addition to 2L, cross-species FISH analysis previously revealed major differences in chromosome organization for 6S, including the *Mi* disease resistance homolog cluster. Mapping results in the 1–10 cM interval were interpreted as a large paracentric inversion that covers the 4.5 Mb euchromatin part with breakpoints close to the top arm telomere and at the border of the pericentromeric heterochromatin. The observations also suggested rearrangements in the middle part of the short arm. However, the precise differences in chromosomal organization were not resolved at the time (Tang et al., 2008b; Peters et al., 2009). Here we aim to further elucidate the topology of 6S. A distal inversion was detected in pepper and potato 6S compared to tomato. In addition, one BAC displayed two foci on pepper 6S, which could be due to a duplicated segment or a breakpoint in pepper, or a deletion in potato (fig. 2.1).

Differences in the mapping order of markers on tomato, potato and pepper genetic maps also indicate rearrangements near the *TSW* and *L* loci. These loci reside on the long arm of chromosomes 10 (10L) and 11 (11L), respectively, and have been implicated in tomato spotted wilt virus and tobacco mosaic virus resistance in *Capsicum* spp. (Jahn et al., 2000; Yang et al., 2009). We selected three anchored tomato BACs in the 52–62 cM interval and the 74–90 cM interval on the EXPEN2000 genetic map for FISH analysis (figs. 2.S2 and 2.S3). These tomato BACs each display a focus in the euchromatic portion of tomato, potato and pepper pachytene chromosomes, and the order of foci on potato and pepper is collinear. However, the order of foci is inverted on tomato 10L, indicating rearrangements in this part of the tomato genome. On tomato and potato 11L the order is collinear, but cytogenetic mapping indicates a translocated segment with a reversed orientation in 11L of pepper (fig. 2.1), in agreement with the comparative mapping in the 57–115 cM interval (Yang et al., 2009) (fig. 2.S3).

3.2 Comparative sequence alignment

To further delineate differences in the chromosomal organization, we selected tomato scaffold sequences based on blastn hits using tomato BAC ends and genetic marker sequences (http://solgenomics.net). On 2L, the 82–143 cM interval corresponds to a sequence of approx. 7 Mb from tomato scaffold SL2.40sc03665 (http://solgenomics.net/sequencing/agp.pl), which spans H159F19 and H072B02 on the FISH map (fig. 2.1). Tomato BAC ends and markers anchoring BACs to the EXPEN2000 map show blastn matches to five potato scaffolds (DM244, DM441, DM1213, DM12 and DM4) from *Solanum tuberosum* group Phureja DM1-3 516 R44 (hereafter referred to as *S. tuberosum* DM), which have a total length of approx. 7 Mb. Both DM scaffold order and orientation on the potato 2L FISH map are consistent with the genetic marker order, and the »accessioned golden path« (AGP) map for potato chromosome 2 (Xu et al., 2011). Subsequently, a comparative alignment revealed the coordinates of the junction breaks of translocated segments for scaffold DM4, DM12 and DM244. In addition, two segments of approx. 3 Mb and 980 kb originating from scaffold DM4, two DM12 segments of 300 and 690 kb, and a small 20 kb segment of DM1213 align in the opposite orientation (fig. 2.2), which suggests inversions have occurred.

The identity plot also displays multiple rearrangements between corresponding segments of tomato and potato 6S, 10L and 11L (fig. 2.2b–d; table 2.S1). We observed large inversions for practically the entire 4.5 Mb of tomato 6S, with several segments that appear translocated compared to potato 6S. In addition, the first 0.5 Mb just downstream of the tomato 6S telomere position appears almost entirely missing in potato, leaving only a small collinear fragment of 60 kb. For a 2.9 Mb segment from tomato 10L, we found two collinear fragments of 375 kb and 1.35 Mb. Two fragments of approx. 470 and 710 kb aligned in opposite orientation to fragments from a 7.5 Mb potato segment, of which 4.14 Mb did not align. A comparative alignment of a 1.9 Mb tomato 11L segment to a corresponding 2.5 Mb potato segment showed three inversions of 144, 295 and 190 kb, and deletions of approx. 230, 140 and 90 kb. Although cytogenetic mapping did not reveal the inversions in tomato 11L, the FISH mapping order is explained by the blastn hit positions (fig. 2.2d). Because genomic sequences are currently lacking for pepper, we cannot confirm the structural rearrangements between tomato and pepper 6S, 10L and 11L by sequence comparison.

3.3 Topology of tomato and potato 2L segments

To exclude the possibility that the observed collinearity breaks in the comparative sequence alignment arose from aberrantly assembled sequences, we validated the borders of collinear scaffold segments by comparative FISH analysis using BACs that span a junction break. The FISH mapping shows single clear foci for the junction break BACs on tomato 2L (fig. 2.3). The mapping order is in agreement with the order and BLAST position of BACs on scaffold 3665, and thus confirms correct assembly of the 7 Mb region. Remarkably, several tomato BACs that span alignment breaks displayed two FISH signals on 2L of both potato accessions G254 and RH89039 (fig. 2.3, lanes 1, 3, 5 and 7). In addition, the FISH map position of H138J12 is just north of E129C17 (fig. 2.3, lanes 3 and 4), in agreement with the inverted orientation of a 300 kb DM12 segment. H028F18 co-localizes with H015P22 on potato 2L from accession RH8903916 (fig. 2.3, lane 3), in contrast to two distinct foci on tomato (fig. 2.3, lane 4). This is in agreement with two overlapping blastn hits on DM4 and two hits on scaffold SL2.40sc03665, which map several megabases apart. Furthermore, M046B12 and E129C17 both show a single focus on tomato and potato (fig. 2.3, lanes 3, 4, 7 and 8), and both H015P22 and H028F18 co-localize on potato pachytene (fig. 2.3, lane 3). Apparently, a large 3 Mb segment from DM4 has been translocated and has a reversed orientation in tomato. H088K05 and H040C22/H138P10 each display a single focus in a similar order on both tomato and potato pachytene (fig. 2.3, lanes 5 and 6). This confirms the orientation and position of DM441 and a translocated 690 kb segment from DM12. The two signals for H160D06 (fig. 2.3, lanes 5 and 7) are in agreement with BLAST hits to DM244 and DM4 sequences. For E022J22, we also observed one signal on tomato (fig. 2.3, lane 2) and two foci on potato (fig. 2.3, lanes 1 and 5), consistent with two BLAST hits to DM244 and DM12, which are 1 Mb apart on the physical map. The corresponding cytogenetic spacing between two

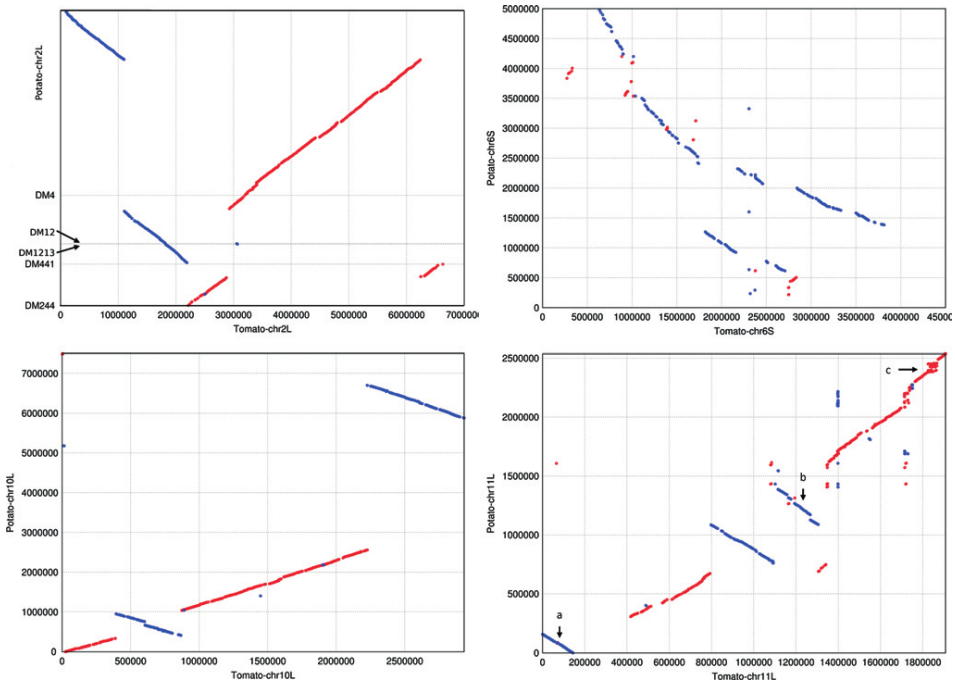


Figure 2.2: Identity plots of *S. lycopersicum* and *S. tuberosum* group Phureja DM1-3 516 R44 segments of 2L, 6S, 10L and 11L. Sequences aligned in forward and reversed orientation are represented by red and blue lines, respectively. Tomato and potato chromosome labels are indicated on the x and y axes. Segment positions aligning to BAC H303G16, H150M12 and E043B08 are marked in the bottom right plot by arrows (a) to (c), respectively.

foci from E022J22 appeared to be comparable to that observed between H160D06 and E022J22 (fig. 2.3, lane 5). This result adds to the notion that 660 and 386 kb segments from DM244 have an inverted orientation, of which the 660 kb segment is also translocated in tomato. Furthermore, mapping confirms that potato accessions G254 and RH890391 share a similar 2L topology with potato DM. Based on these results, we reconstructed the tomato and potato chromosomal organization (fig. 2.4).

3.4 Orthologs on tomato and potato 2L and linkage conservation

The recent tomato and potato genome sequencing efforts and available genetic map information enable analysis of the gene repertoire and relative order and orthology detection, a prerequisite to investigate the extent of conserved linked genes in repositioned homeologous segments. In this respect, Bonierbale et al. (1988) have demonstrated that cDNA markers are largely collinear along the tomato and potato chromosomes. In particular, the segments from tomato and potato 2L have several EST-derived genetic markers in common, and thus appear to be conserved syntenic (fig. 2.4). To assess the conserved linkage, we used the complete set of ITAG2 and PGSC release 3.4 annotated protein sequences (see section 2) to identify and map the main orthologs in syntenic tomato and potato 2L segments, respectively.

Ortholog detection

The syntenic 2L segments contain 893 predicted genes in tomato and 820 in potato, respectively. Use of InParanoid (Remm et al., 2001) identified 664 ortholog groups, of which 623 comprise an ortholog pair localized in corresponding collinear segments, consistent with a conserved syntenic nature. For 25 ortholog groups, 14 tomato and 11 potato ortholog genes mapped outside homeologous segments (table 2.S2). Within the 623 ortholog groups, we detected 721 tomato genes and 679 potato genes, apparently indicating that gene duplications have occurred. The mean tomato and potato gene copy (paralog) numbers for the 623 ortholog groups are 1.16 and 1.09, respectively (table 2.S2). There are 36 ortholog groups with more than two members, of which 12 groups have multiple tomato paralogs, 16 groups have multiple potato orthologs, and eight groups have multiple tomato and potato paralogs. In two ortholog groups, tomato genes Solyc02g090350 and Solyc02g085990 have adjacent paralogs. Thus, although it appears that tomato and potato share a comparable basic set of genes overall, the order and number of gene copies is substantially different.

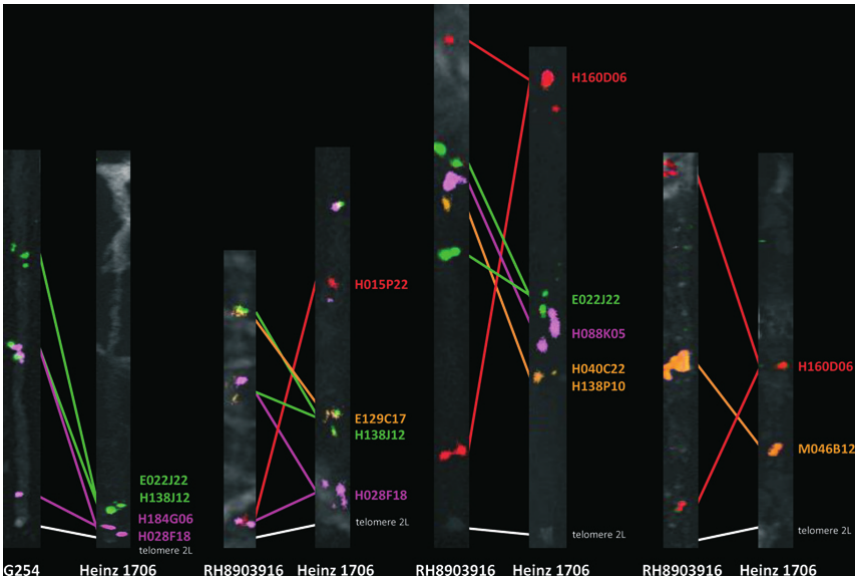


Figure 2.3: FISH analysis of collinearity breaks between tomato and potato 2L. Tomato BACs were hybridized on pachytene complements from tomato Heinz 1706 (lanes 2, 4, 6 and 8) and pachytenes from potato G254 (lane 1) and RH8903916 (lanes 3, 5 and 7). Corresponding cytogenetic mapping positions of *HindI*II (H), *EcoRI* (E) and *MboI* (M) tomato BACs on tomato and potato pachytenes are shown by coloured connecting lines. Telomere positions on the 2L are indicated by white connecting lines.

Gene adjacency, orientation and unclustered genes

We found 335 ortholog pairs that have a conserved linkage, and 288 pairs that have disrupted gene adjacency. For example, comparison of tomato genes with their potato orthologs just downstream of the F18 junction in scaffold DM12 showed that local gene vicinities are preserved, but gene adjacency is disrupted. Tomato and potato orthologs were frequently interrupted by putative LTR retrotransposons (fig. 2.S4). In total, we predicted 28 LTRs in tomato scaffold 3665 and 56 LTRs in the collinear potato DM scaffolds.

Of the identified orthologs, eight orthologous gene pairs show irregular positioning compared to the layout of the collinear segments (fig. 2.5; table 2.S3). The orthologs corresponding to the gene models located on potato scaffold DM1213 are shifted by 1.23 Mb upstream and inverted in tomato (fig. 2.5, genes e and f; fig. 2.2). Here, the provisional state of the current potato scaffold order appears to be the most straightforward explanation for this aberrant position, although a biological rearrangement cannot be excluded. In tomato, this domain contains two AP2-domain transcription factors. ITAG RepeatMasker annotation (see section 2) shows that these AP2 homologs are flanked by copies of LTRs that possibly originate from the same transposon. Furthermore, there are three orthologous gene

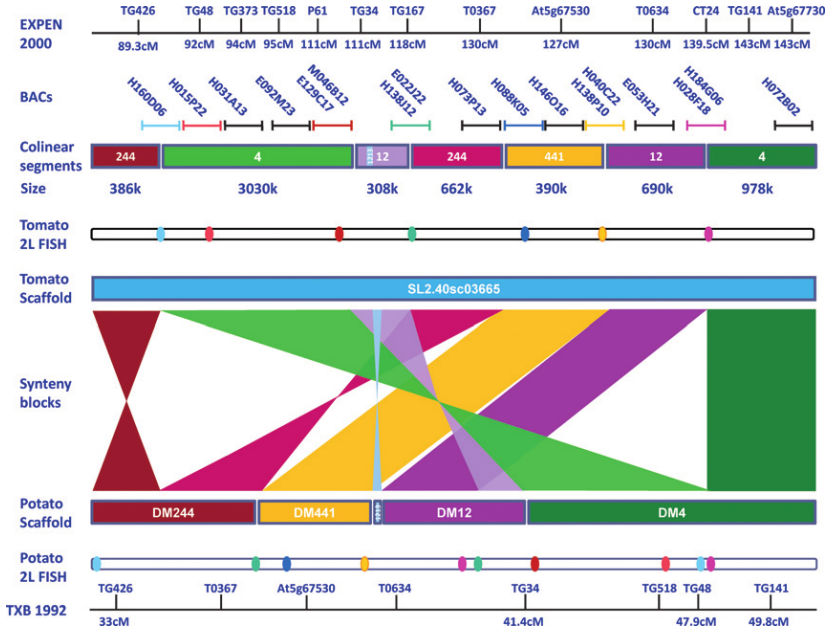


Figure 2.4: Comparative map for tomato and potato 2L. Chromosome 2 markers in tomato and potato scaffolds are indicated on the top and bottom bars with cM positions from the tomato EXPEN2000 and potato TXB1992 genetic maps. The order and relative positions of markers without a genetic position are derived from blastn hits to potato scaffolds. Aligned segments from potato DM scaffolds to tomato scaffold SL2.40sc03665 are represented by colored rectangles. Syntenic blocks are represented by colored polygons in the middle section, and correspond to the position and orientation of rearranged potato segments. The relative positions of tomato BACs flanking or spanning collinearity breaks are labeled with their identifier above each DM segment. The relative positions and order of the tomato BACs on the tomato and potato FISH map are indicated by corresponding colored dots. Some BACs spanning a collinearity break have one corresponding position on the tomato FISH map and two corresponding positions on the potato FISH map.

pairs that remain as relicts of the proposed splitting of scaffold DM244 into separate contigs (fig. 2.5, genes b, c and d). In addition, we found two pairs of orthologs with an inverted orientation to the syntenic segment (fig. 2.5, genes a and h), and one ortholog pair that shows transposition of the tomato ortholog to a region approx. 2.8 Mb distal of its expected position (fig. 2.5, gene g). A substantial proportion of the 241 and 173 putative genes, respectively, in the tomato and potato aligned segments remained unclustered (fig. 2.5), and probably are the result of stringent clustering cut-off values or aberrantly predicted genes in the gene annotations, or they may be species-specific. Furthermore, some unclustered genes are probably false negatives, taking into account that InParanoid has a false-negative rate of approx. 3% when omitting the use of an outgroup (Remm et al., 2001).

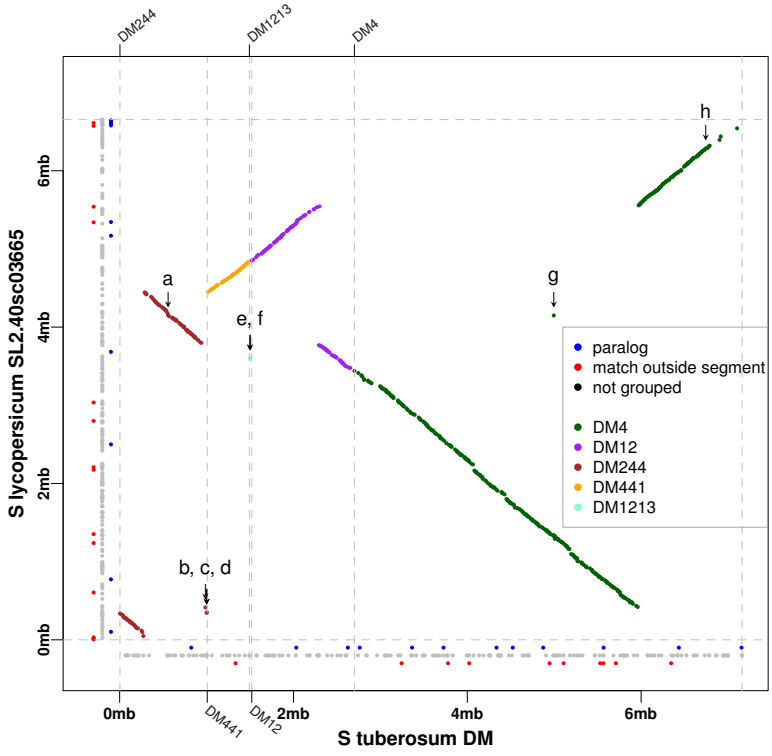


Figure 2.5: Orthologous gene matches for potato versus tomato. The first 6.7 Mb of tomato scaffold SL2.40sc03665 and potato scaffolds DM4 (green), DM12 (purple), DM244 (red), DM441 (yellow) and DM1213 (blue) are shown. The midpoint of a gene model is represented by a dot. An orthologous gene pair is indicated by a dot in the first quadrant, unclustered genes are shown at −100 000, paralogs are shown at −50 000, and clustered genes without a matching counterpart in the selected segment are shown at −500 000 along the corresponding axis. Ortholog pairs with an irregular position are indicated by arrows (a) to (h).

3.5 A rearrangement pathway model

There is compelling evidence that recombination plays a much larger role in the evolution of plant genomes than previously appreciated. Recombination in plants is highly variable, and includes (*i*) meiotic recombination between homologous chromosomes, (*ii*) intra-strand crossing over between direct and inverted repeats, (*iii*) unequal crossing-over between misaligned repeats on homologous chromosomes, and (*iv*) illegitimate or non-homologous recombination (Gaut et al., 2007).

Although the mechanisms that underlie complex genome rearrangements in plants are quite diverse and not fully understood, there is accumulating evidence that links rearrangement breakpoints and repeats (Coghlan et al., 2005). Thus, rearrangements in tomato and potato 2L may be explained in conformance with the type and location of repeats. We detected similarity to class I and II transposons in 20 kb bins near the synteny breaks, in particular sonata, harbinger, Long Interspersed Elements (LINEs), Short Interspersed Elements (SINEs), unknown retrotransposons and telomere-like related sequences (figs. 2.S5 and 2.S6). Furthermore, comparative sequence analysis revealed two copies of a 4.5 kb inverted repeat on tomato 2L flanking the junction between the 690 and 978 kb syntenic blocks, and spanning the synteny junction between the 308 and 662 kb segments. Another repeat of 267 bp is located near the 386–3030 kb segment junction and the 308–662 kb segment junction (fig. 2.S7). We did not detect similarity to known transposable elements (TEs), and therefore these repeats are not likely to be of transposon origin. Currently, we do not have direct evidence for their involvement in the rearrangements, but, taking into account their location near synteny breaks, we hypothesize that transposon element/repeat-mediated recombination may be explanatory for the 2L rearrangements as follows. We assume the existence of unichromosomal breakpoints, as we have not found any 7 Mb segmental duplication in other parts of the tomato genome, and according to Pevzner and Tesler (2003), that the unichromosomal breakpoints are related and inter-dependent. Furthermore, the 2L rearrangements are probably tomato lineage-specific. This notion is substantiated by several genome mapping studies (Bonierbale et al., 1988; Tanksley et al., 1988, 1992; Livingstone et al., 1999; Thorup et al., 2000; Doganlar et al., 2002a; Pertuzé et al., 2002; Ashrafi et al., 2009) and the collinear BAC FISH maps for 2L of *S. lycopersicum*, *S. lycopersicoides*, *S. pennellii* and *S. chilense* (H.d.J., unpublished results). Therefore, we assumed the potato and pepper organization to be ancestral, and thus the rearrangements needed to transform the tomato into potato chromosome topology can be used to reconstruct an ancestral karyotype. Taking this evolution direction into account, we propose a rearrangement pathway model in which intra-strand crossing-over and ectopic recombination give rise to segmental rearrangements (fig. 2.6). First, two segments from DM244 become inverted by ectopic recombination mediated by copies of the same transposon elements that are positioned near the synteny breaks, or by non-homologous recombination. After the second reversal, a chromatid strand folds back, generating local pairing of the 276 bp inverted repeats, subsequent to which, inter-chromatid crossing-over occurs, leading to a third inversion. After breakage and fusion, this generates an intermediate configuration in which the 4.5 kb repeat is inverted. The 4.5 kb inverted repeat element then mediates a second intra-strand crossing-over event, resulting in a fourth inversion. Indeed, when using GRIMM (Tesler, 2002), a single and most parsimonious scenario was predicted, which consisted of a trajectory involving four recombination events (table 2.S4) that resembled the conversion steps presented in the rearrangement model proposed above.

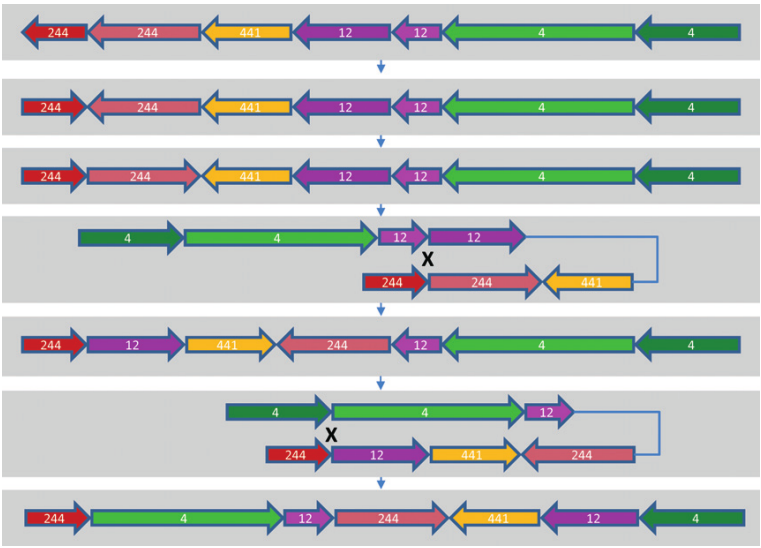


Figure 2.6: Proposed rearrangement scenario involving four structural conversions given the initial potato order (upper panel) and final tomato order (bottom panel) of 2L segments. Positions of intra-strand crossing-overs between segments are marked »X«. Arrows indicate the relative order and orientation of 2L segments.

4 Discussion

4.1 Unravelling structural differences in *Solanum* and *Capsicum*

In recent decades, marker-based mapping studies provided broad knowledge of the structural and molecular organization of various plant genomes, including tomato and potato. It is apparent that plant genomes share extensive conserved linkage despite their diversity in size and complexity (Bonierbale et al., 1988; Tanksley et al., 1992; Bennetzen, 2000a). These studies also indicated large-scale differences in genomic organization, most of which were found in heterochromatic domains. Nevertheless, comparative maps have limitations that arise from low marker density and low placement accuracy that complicate local resolution of chromosomal organization. Tang et al. (2008b) and Peters et al. (2009) used BAC FISH painting to unravel complex rearrangements in euchromatin regions and elucidated discrepancies between genetic and physical maps. In this paper, we provide more detailed mapping at the sequence level in order to identify the coordinates of collinear segments and subsequent selection of BAC targets to validate the collinearity breaks. We have shown that large-scale structural differences in *Solanum* and *Capsicum* are not only confined to heterochromatin portions, but frequently occurred in the euchromatic portion of 2L, 6S, 10L and 11L. Thus, although genetic maps and

comparative sequence analysis each have limitations in power and resolution, their combined usage is indispensable for unsurpassed knowledge of chromosome organization at the structural and sequence level.

4.2 Large-scale rearrangements

Comparative sequence alignment indicated the location of collinearity breaks between tomato and potato DM 2L. Using BACs that span junction breaks, FISH revealed double signals marking the borders of collinear segments in potato accessions G254 and RH890316. In total, six large-scale rearrangements were discovered on a 7 Mb euchromatic region of 2L between tomato cv. Heinz 1706 and potato G254 and RH8903916. Although chromosome 2 markers TG34 and TG48 in the 89–143 cM intervals show an inverted order (Tanksley et al., 1992), these structural differences between tomato and potato were not apparent from linkage maps.

Linkage maps between tomato and several wild relatives show a similar order of 2L genetic markers (Tanksley et al., 1992; Fulton et al., 1997; Pertuzé et al., 2002; Ashrafi et al., 2009; http://solgenomics.net), but they appear to be inverted for pepper, eggplant/aubergine (Livingstone et al., 1999; Thorup et al., 2000; Doganlar et al., 2002a; http://solgenomics.net) and potato (this paper). Furthermore, potato, pepper and eggplant share extensive marker collinearity in the homeologous 2L segments. Taken together, these results suggest that the rearrangements in 2L are specific for the tomato clade, and thus occurred after the split from the common ancestor of tomato and potato.

A tomato-specific rearrangement was also reported for chromosome 10, for which the tomato, potato and pepper homeologs apparently differ by a paracentric inversion (Tanksley et al., 1988, 1992; Livingstone et al., 1999). Comparative linkage map studies in sister species of tomato, *S. lycopersicoides* and *S. sitiens* (Pertuzé et al., 2002) showed a 10L configuration that was similar to that of potato and pepper (*Capsicum*), and this provided further evidence for the notion that the paracentric inversion was already fixed in the common ancestor for the tomato lineage. The FISH mapping presented here confirms two inversions in tomato 10L compared to the potato and pepper chromosome organization, and is consistent with the tomato lineage-specific nature of the rearrangements.

The structural organization appears less similar in tomato and potato 11L. The comparative sequence alignment indicates multiple breakpoints corresponding to three inversions and three deletions. Although an accurate sequence alignment for pepper 11L is currently lacking, cytogenetic mapping revealed at least one translocated segment with a reversed orientation. By comparing tomato and pepper genetic maps, a minimum of 22–32 breakages of tomato chromosomes would be necessary to transform the order and position of tomato genes to that observed in pepper (Tanksley et al., 1988; Livingstone et al., 1999). We have not investigated all putative chromosome rearrangements, and it is conceivable the number of rearrangements is an under-estimate and a consequence of a low density of genetic markers. This notion is supported by our observation that the previously reported hidden inversion for the 4.5 Mb short arm of chromosome 6 (Tang et al., 2008b)

actually appears more complex. A comparative sequence alignment revealed multiple reversals, translocated segments and a deletion in potato 6S. From the present results and those of previous studies, we propose a genome landscape in which evolution on the structural level for the majority of the 12 chromosomes in Solanaceae was far more dynamic than currently appreciated.

4.3 Chromosomal rearrangements and reproductive isolation

The role of the large structural rearrangements in 2L with respect to inter-generic and intra-generic reproductive isolation and speciation of Solanaceae, remains unclear. Rearrangements can impede proper pairing of homeologous chromosomes and reduce recombination, and may also cause decreased fitness or even sterility (Noor et al., 2001; Livingstone and Rieseberg, 2004; Rieseberg and Willis, 2007; Bedinger et al., 2011). For example, inter-specific hybrids between *S. lycopersicum* and *S. pennellii* are highly fertile, and light microscopy analysis demonstrated near normal levels of meiotic pairing and crossing-over (Tanksley et al., 1992). However, high-resolution electron microscopic analysis of chromosomes at pachytene showed frequent unusual synaptic configurations (Anderson et al., 2010). The latter observation might seem in contradiction with fertility and the unaffected recombination between tomato and *S. pennellii*. However, it is possible that relatively small structural differences between homeologous chromosomes may be tolerated, but relatively large rearrangements may disrupt meiotic synapsis or recombination and may result in infertility. For example, inter-specific hybrids between *S. lycopersicoides* and *S. sitiens* are fertile and show recombination rates similar to tomato (Pertuzé et al., 2002). In contrast, inter-generic hybrids of *S. lycopersicum* with *S. lycopersicoides* or *S. sitiens* are sterile and show genome-wide suppressed recombination (Chetelat et al., 1997). Moreover, recombination is completely abolished for 10L, which may be explained by a large paracentric inversion between the L-type (*Lycopersicon* spp.) and S-type (*S. lycopersicoides* and *S. sitiens*) genomes (Pertuzé et al., 2002).

Although we have not investigated the molecular nature of inter-specific barriers between tomato and potato, the rearrangements presented here may very well impede proper synapsis. In line with this are results obtained from somatic hybrids of tomato and potato, in which irregular synapsis was also frequently observed (de Jong et al., 1993). The latter may perhaps involve absent, repositioned or protected blocks of genes that suppress homeologous pairing (Bedinger et al., 2011). In this respect, a phenomenon known as transmission ratio distortion is often observed in inter-specific crosses between tomato and wild relatives. Possibly, selection against particular allelic combinations that are associated with »transmission ratio distortion« loci underlies hybrid incompatibility (Moyle and Graham, 2006). For example, strong reproductive barriers have been observed between *S. lycopersicum* and the tomato-like nightshades *S. ochrantum* and *S. juglandifolium*, which have been placed phylogenetically and morphologically in an intermedi-

ate position between tomato and potato (Spooner et al., 2005). Remarkably, we physically mapped markers associated with distortion loci *trd2.2*, *sd10.1*, *sd10.2*, and *trd11.1* from *Solanum* species (Pertuzé et al., 2002; Albrecht and Chetelat, 2009) to the rearranged segments in 2L, 10L and 11L discussed here (figs. 2.S1 to 2.S3), suggesting a relationship between hybrid incompatibility, transmission ratio distortion and large rearrangements.

4.4 Small-scale rearrangements and changes in gene repertoire

Overall, tomato and potato have a 7% difference in gene copy number in rearranged 2L segments (table 2.S2). Studies on gene repertoire and gene order in tomato and pepper indicated a change in locus number for approx. 12% of the loci, accompanied by an extensively modified linear order of genes and many chromosome rearrangements (Tanksley et al., 1988). These differences in copy number are in line with the slightly larger genome size of tomato compared to potato and the expanded pepper genome compared to tomato.

Our observation that small rearrangements are more frequent than large-scale differences seems consistent with earlier observations made for other plant genomes (Bennetzen, 2000a). In particular, micro-collinearity and conserved linkage between orthologs is apparent, but we nevertheless found many small exceptions. Observations in yeast and *Drosophila*, for example, show that direct repeats of LTR transposon copies may act in reciprocal recombination, giving rise to gene loss. Reciprocal recombination between inverted repeats from LTR retrotransposons may result in gene inversions, while recombination between repetitive elements on different chromosomes may lead to reciprocal translocation (Bennetzen, 2000b; Gray, 2000). In some cases, we suspect the micro-collinearity in 2L has been disrupted by TEs. For example, two translocated AP2-like transcription factors in tomato 2L with a disrupted conserved linkage are flanked by LTRs from a single retrotransposon, suggesting its involvement in the relocation and inversion of these genes. Strikingly, the inversion of AP2-like transcription factors in tomato maps to near the *fw2.2* and *fs2.2* loci controlling fruit weight and bell-shaped fruit morphology in the heirloom tomato cv. Yellow Stuffer and garden pepper (Grandillo et al., 1999; van der Knaap and Tanksley, 2003). The inverted context of genes functioning in signalling pathways has been reported to affect traits and cause phenotypic changes in *Drosophila* (Hoffmann et al., 2004). However, we currently have no functional or phenotypic evidence indicating an inversion-induced difference in gene interaction or regulation in tomato.

4.5 Analyses of chromosomal rearrangement junctions and rearrangement scenario

Currently, it is unclear what might have caused the large rearrangements in 2L. TEs are known to induce large inversions, deletions and translocations, mediated directly via their transposition mechanisms, or indirectly via homologous recombination, or by ectopic or non-allelic homologous recombination (Bennetzen, 2000b). For example, in maize (*Zea mays*), pairs of TEs spaced beyond 100 kb are efficient chromosome breakers, generating deletions and inversions via alternative transposition (Huang and Dooner, 2008; Zhang et al., 2009). Recombination in plants is not limited to homologous chromosomes only. For example, inter-chromatid recombination can result in deletions and inversions, and such recombination between different genomic regions can lead to large chromosomal rearrangements (Gaut et al., 2007). Furthermore, it is important to realize that meiotic as well as mitotic rearrangements in plants can be passed to progeny. In Arabidopsis, elevated somatic recombination rates have been observed and found to be positively correlated with DNA damage and stress, suggesting that genomic flux caused by recombination plays an important role in environmental stress adaption. Direct evidence for genomic flux involving large-scale rearrangements caused by repeat-mediated recombination has not been reported for plants. However, indirect evidence has been found in Brassicaceae species, in which rearrangements are to a large extent located near repetitive sequences (Ziolkowski, 2003; Lysak et al., 2006).

In general, any chromosomal rearrangement involves a breakage and a subsequent repair of the chromosome ends or fusion to another chromosome end. Of the different types of rearrangements in plant genomes, inversions probably occur the most frequently, and can range size from a few kb up to hundreds of genes in length (Coghlan et al., 2005). Our results are in agreement with this notion, showing that inversion is the predominant rearrangement type in tomato 2L, 6S, 10L and 11L. Previously, Livingstone et al. (1999) inferred the most recent ancestral genome using lineage-specific rearrangements, the phylogeny of tomato, potato and pepper (Spooner et al., 1993), and comparative maps. We reasoned that, as the potato and pepper 2L organization appear similar, with pepper being considered as an outgroup, and because the rearrangements are tomato lineage-specific, the potato/pepper topology may be considered ancestral. In the proposed model, the inversion rearrangement was preferred above other types, and the reconstruction was directed from potato/pepper towards the tomato 2L organization. Indeed, when calculating a most parsimonious rearrangement scenario, only four reversion steps were needed to transform the potato 2L organization into the tomato 2L organization. However, we currently cannot exclude other rearrangement pathways. Tandem repeats and (retro)transposons are present also outside the 2L junction breaks, and as we have no indications about the relative order of the proposed inversion steps, the proposed model should be considered as a working hypothesis.

The efficiency and success of introgressive hybridization breeding on the basis of DNA-based selection depends, among others, on adequate identification of chromosome organization. The implication of technological advances with respect

to next-generation sequencing technology for the use of extant germplasm resources is that large numbers of complex genomes can be sequenced relatively fast and cheaply. This will undoubtedly speed up identification of compatible genomes for introgression breeding, the rearrangement phylogeny within the Solanaceae, and reconstruction of the ancestral *Solanum* karyotype. The present project is a first step in mining of structural genetic diversity and the development of genome-based breeding tools.

Acknowledgements

This work is supported by Technological Top Institute Green Genetics grant number 2CC037RP, financial aid from Rijk Zwaan, Syngenta AG and Monsanto, and by the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

5 Supporting Information

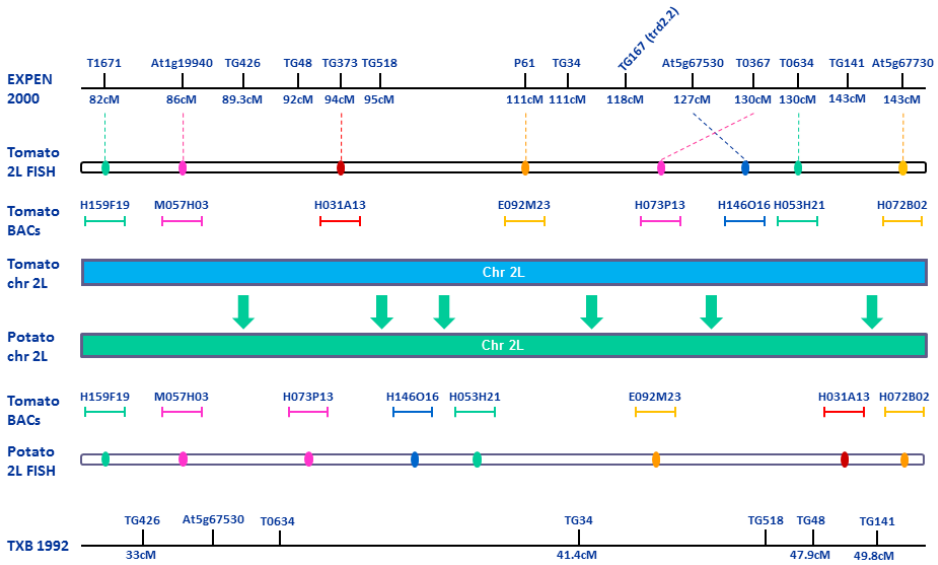


Figure 2.S1: Integrated map of tomato and potato chromosome 2L segments. In the middle section tomato BAC positions on tomato and potato chromosomes are displayed and correspond to the tomato and potato BAC FISH map positions. Chromosome 2 markers that have a genetic positions on the Tomato-EXPEN2000 and Potato-TXB 1992 genetic maps are displayed at the top and bottom. The order and relative position of markers without a genetic position is derived from blastn hits to potato scaffolds. Arrows indicate approximate positions of chromosome collinearity breaks.

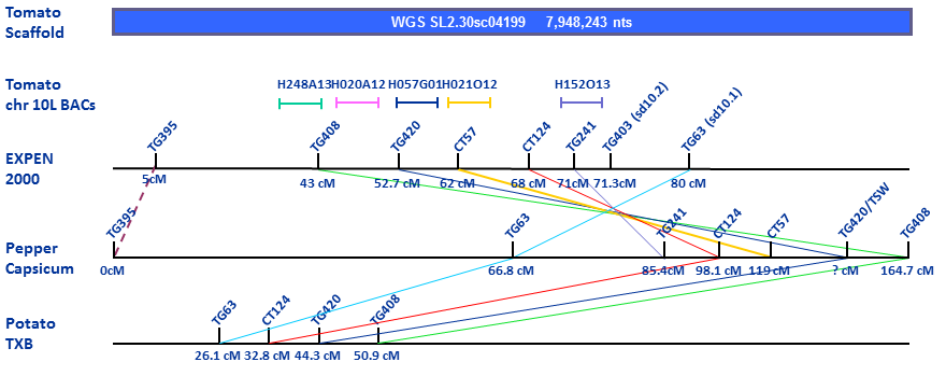


Figure 2.S2: Comparative map of tomato, potato and pepper chromosome 10L.

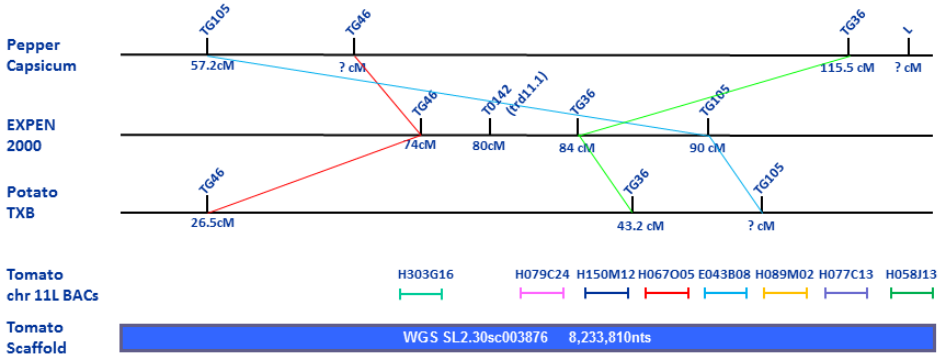


Figure 2.S3: Comparative map of tomato, potato and pepper chromosome 11L. Map positions of genetic markers and L resistance gene are according to Yang et al. (2009).

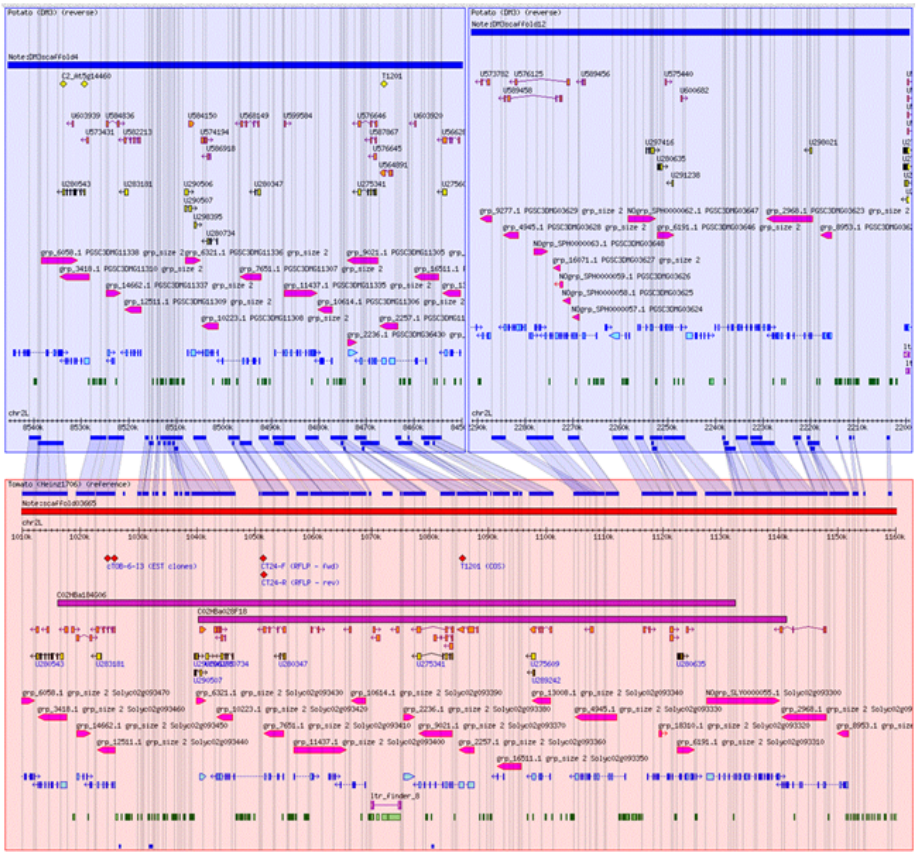


Figure 2.S4: Tomato synteny browser snapshot of a 150 kb region at the F18 junction. Shaded connectors represent NUCmer sequence alignments which connect segments between DM4 (blue bar top left panel) and DM12 potato scaffolds (blue bar top right panel), and tomato scaffold SL2.31sc03665 (red bar bottom panel). Red and yellow diamonds represent EXPEN2000 markers and SGN markers, respectively. SGN tomato unigenes, SGN potato unigenes, predicted proteins from ITAG and BGI annotations, Genscan predicted genes, LTR finder predicted retrotransposons, and RepeatMasker predicted repeats are represented as orange, yellow, silver, blue, pink and green colored gbrowse glyphs, respectively, with arrow heads displaying the orientation. Each annotated protein is depicted with its identifier, ortholog group id and ortholog group size. Purple bars (bottom panel) represent tomato BACs H084G06 and H028F18 spanning the F18 junction between DM4 and DM12 scaffold segments.

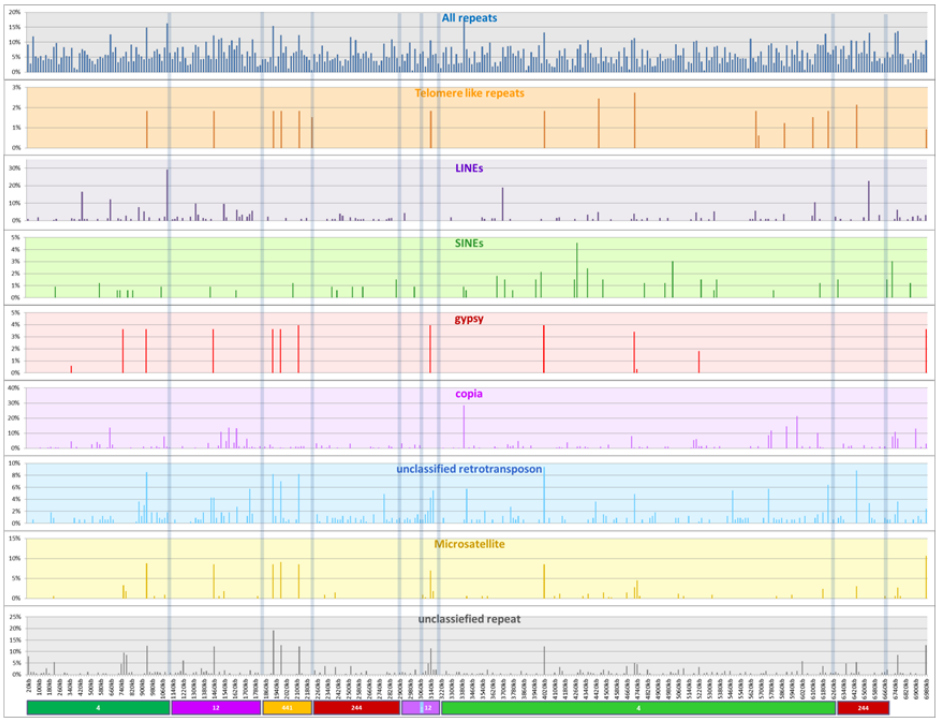


Figure 2.S5: Distribution of class I transposon-related repeats in chromosome 2L. The histograms reflect transposon content in 20 kb bins. Vertical blue bars represent 20 kb intervals at rearrangement junctions. Transposon families are depicted as headers in each histogram. The tomato chromosome 2L topology is represented by the bottom bar.

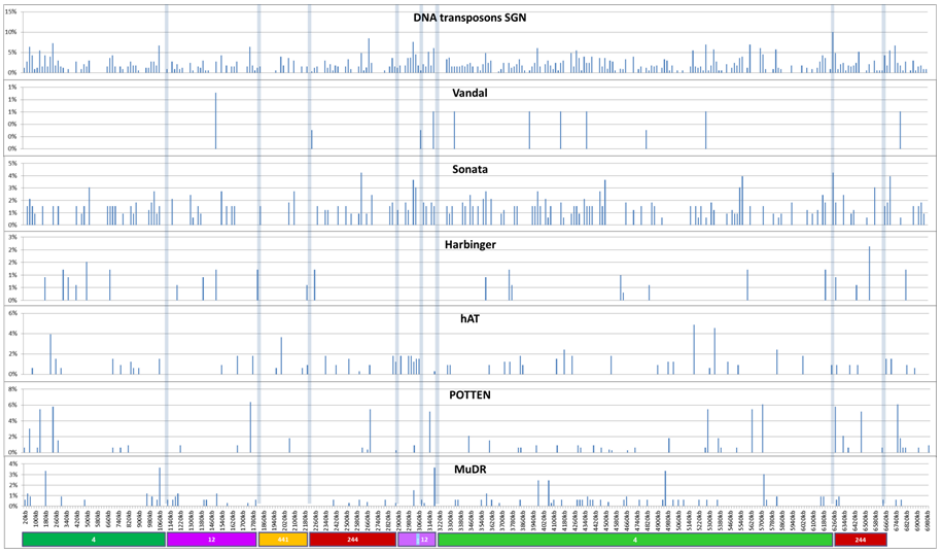


Figure 2.S6: Distribution of class II transposon-related repeats in chromosome 2L. The histograms reflect transposon content in 20 kb bins. Vertical blue bars represent a 20 kb interval at rearrangement junctions. Transposon families are depicted as headers in each histogram. The tomato chromosome 2L topology is represented by the bottom bar.

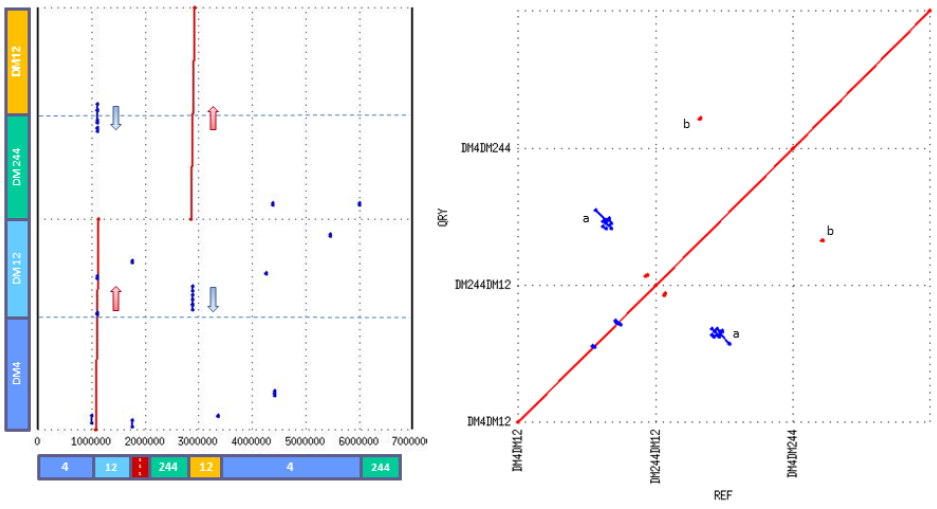


Figure 2.S7: Identity plot of synteny junctions. Left panel: alignment plot of 60 kb sequences flanking the DM4DM12 junction and DM244DM12 junction (y-axis) to chromosome 2L segment of *S. lycopersicum* (x-axis). Segments junctions are indicated by dashed horizontal blue lines. A 4.5 kb inverted repeat is indicated by red and blue colored arrows. Tomato segments syntenic to potato are indicated below the x-axis. Right panel: Identity plot of 60kb sequences flanking synteny junctions. A 4.5 kb inverted repeat and a 276 bp repeat are denoted by a and b respectively.

Table 2.S1: Large-scale rearrangements between tomato, potato and pepper. Rearrangement types are indicated as reversal (r), translocation (t), deletion (d) and insertion (i), or not determined (n.d.).

tomato chromosome	EXPEN 2000 position (cM)	SL2.40 scaffold	Match Size (Mb)	Rearrangement type	
				potato	*pepper*
2L	89.3 - 143	SL2.40sc03665	6.7	r, t	n.d.
6S	1 - 10	SL2.40sc04474	4	r, t, d	n.d.
10L	52.7 - 62	SL2.40sc04199	2.5	r, t, d	r
11L	74 - 90	SL2.40sc03876	1.8	r, t, i	r, t

Table 2.S2: Predicted ortholog pairs and paralogs in tomato and potato chromosome 2L. Average gene copy numbers are defined as number of tomato (Sly) or potato (Stu) genes per ortholog group and are indicated between brackets.

	Sly. ortho grps	Stu ortho grps	Sly multi Stu single	Stu multi Sly single	Sly multi Stu multi
grps	623	623	12	16	8
Sly genes	721 (1.16)	0	80 (6.66)	16 (1)	38 (4.74)
Stu genes	0	679 (1.09)	12 (1)	51 (3.20)	29 (3.6)
Sly paralogs	98	0	68	0	30
Stu paralogs	0	56	0	35	21
Sly orthologs	623	0	12	16	8
Stu orthologs	0	623	12	16	8

Table 2.S3: Outliers identified by collinearity analysis of orthologous gene pairs.

ARROW	ORTHO GRP	POS SLY	SCF SLY	GENE ID SLY	STR. SLY	POS STU	SCF STU	ID STU	STR. STU
a	1052	4149794	SL2.31ch02	Solyc02g091580	-	557410.5	PGSC0003 DMB000000244	PGSC0003 DMG400030613	-
b, c, d	14277	413379	SL2.31ch02	Solyc02g086430	+	985880.5	PGSC0003 DMB000000244	PGSC0003 DMG400030591	-
	7553	345960	SL2.31ch02	Solyc02g086370	+	1000291	PGSC0003 DMB000000244	PGSC0003 DMG400030590	-
	16299	350999	SL2.31ch02	Solyc02g086380	-	1002504.5	PGSC0003DMB 000000244	PGSC0003 DMG400038447	+
e, f	16164	3597511.5	SL2.31ch02	Solyc02g090770	-	1498191.5	PGSC0003 DMB000001213	PGSC0003 DMG400004538	-
	5822	3601451.5	SL2.31ch02	Solyc02g090780	-	1502431	PGSC0003 DMB000001213	PGSC0003 DMG400004539	-
g	16446	4148981	SL2.31ch02	Solyc02g091570	-	4994785	PGSC0003 DMB000000004	PGSC0003 DMG400040733	-
h	14592	6285504.5	SL2.31ch02	Solyc02g094440	+	6743466	PGSC0003 DMB000000004	PGSC0003 DMG400020198	-

Table 2.S4: Parsimonious trajectory for rearrangements between potato and tomato chromosome 2L output by GRIMM (http://nbcr.sdsc.edu/GRIMM/grimm.cgi; Tesler, 2002). The order of segments are displayed as positive or negative integers representing a forward or reversed segment orientation respectively. The minimal amount of steps needed to transform the order and position of tomato segments to that observed in potato involves 4 inversions (reversals). The repositioned segments are underlined.

STEP	DESCRIPTION	AFFECTED CHROMOSOMES	
		before	*after*
1	(potato chr2L) reversal	1 2 3 4 5 6 7	-1 2 3 4 5 6 7
2	reversal	-1 2 3 4 5 6 7	-1 -2 3 4 5 6 7
3	reversal	-1 -2 3 4 5 6 7	-1 -4 -3 2 5 6 7
4	reversal (tomato chr2L)	-1 -4 -3 2 5 6 7	-1 -6 -5 -2 3 4 7

Chapter 3

Snf2 family gene distribution in higher plant genomes reveals DRD1 expansion and diversification in the tomato genome

Abstract

As part of large protein complexes, Snf2 family ATPases are responsible for energy supply during chromatin remodeling, but the precise mechanism of action of many of these proteins is largely unknown. They influence many processes in plants, such as the response to environmental stress. This analysis is the first comprehensive study of Snf2 family ATPases in plants. We here present a comparative analysis of 1159 candidate plant Snf2 genes in 33 complete and annotated plant genomes, including two green algae. The number of Snf2 ATPases shows considerable variation across plant genomes (17-63 genes). The DRD1, Rad5/16 and Snf2 subfamily members occur most often. Detailed analysis of the plant-specific DRD1 subfamily in related plant genomes shows the occurrence of a complex series of evolutionary events. Notably tomato carries unexpected gene expansions of DRD1 gene members. Most of these genes are expressed in tomato, although at low levels and with distinct tissue or organ specificity. In contrast, the Snf2 subfamily genes tend to be expressed constitutively in tomato. The results underpin and extend the Snf2 subfamily classification, which could help to determine the various functional roles of Snf2 ATPases and to target environmental stress tolerance and yield in future breeding.

Bargsten, J. W., Folta, A., Mlynárová, L., and Nap, J.-P. (2013). *PLoS ONE*, 8(11):e81147.

1 Introduction

In eukaryotes, genomic DNA is organized into chromatin, which is physically restricting the access of regulatory proteins to the genome (Eisen et al., 1995). The access to the genome can be changed by chromatin modifying activities, altering histone tails or the histone cores covalently; and chromatin remodeling activities, altering DNA–histone interactions non-covalently (Eisen et al., 1995). Both provide important epigenetic mechanisms to regulate gene expression (Flaus and Owen-Hughes, 2011). The associated ATP-dependent changes in nucleosome organization catalyzed by Snf2-family ATPases accounts for a large part of chromatin remodeling activities (Flaus and Owen-Hughes, 2011).

Snf2 ATPases show broad functional diversity and are involved in a variety of genome-wide processes involving DNA, such as transcription, replication, repair and recombination. As ATPase they provide a motor that can translocate and move a complex directionally on double-stranded DNA (Flaus and Owen-Hughes, 2011). In general, Snf2 family ATPases form large complexes with interacting partners (Knizewski et al., 2008), although few Snf2 family members can act alone (Hauk et al., 2010; Lall, 2011). Swapping the ATPase region of two different Snf2 family ATPases in different complexes can also exchange their functionality (Fan et al., 2005). The Snf2 ATPases therefore shape the functionality of a complex.

A first analysis of Snf2 family ATPases based on 30 sequences resulted in a classification of eight distinct subfamilies (Eisen et al., 1995). Snf2 family ATPases are characterized by seven helicase motifs (Eisen et al., 1995; Flaus et al., 2006; Flaus and Owen-Hughes, 2011). The sequence spanning these motifs is called the Snf2 family ATPase region (fig. 3.S1). The conserved ATPase region averages at about 400 amino acids (Eisen et al., 1995) and is supposed to catalyze the translocase activity. A new survey of 1300 Snf2 family ATPases extended the classification to six groups (Snf2-like, Swr1-like, SSO1653-like, Rad54-like, Rad5/16-like and distantly-related Snf2 members) and 24 subfamilies (Flaus and Owen-Hughes, 2011). The division into groups and subfamilies is based on phylogenetic analyses of the Snf2 family ATPase region. In many family members additional (accessory) domains are present, reflecting the sequence-based subfamily classification (Flaus et al., 2006; Knizewski et al., 2008). Not all subfamilies occur in every species or kingdom. An example is the DRD1 (defective in RNA-directed DNA methylation) subfamily occurring only in plant species (Kanno et al., 2004; Matzke et al., 2006).

In plants, functional annotation of Snf2 family members is most advanced in Arabidopsis. The Arabidopsis genome encodes 41 Snf2 family gene loci (http://www.chromdb.org; http://www.snf2.net). Encoded genes are distributed over six groups and 18 subfamilies. The specific function of the majority of the Snf2 proteins in plants is unknown (Knizewski et al., 2008), apart from the general contribution to DNA repair and recombination in development (Flaus and Owen-Hughes, 2011; Sang et al., 2012). Different Snf2 ATPases, including members of the Snf2 and DRD1 subfamilies, have been shown to play a role in plant stress responses. Hence, the exploitation of such genes provides the basis for further

functional characterization and could help develop plants that are better able to withstand environmental variation and/or (a)biotic stress. This may result in higher yields in less favorable environments.

We here present the first comprehensive analysis of Snf2 family members within the plant kingdom, to investigate phylogenetic relationships and infer putative specific functions of individual family members. Plant genomes show a high variability of the number of Snf2 genes, ranging from 17 to 63 members. The tomato (*S. lycopersicum*) genome shows gene expansions of the DRD1 subfamily with distinct expression patterns, suggesting further subfunctionalization of the duplicated members.

2 Materials and Methods

2.1 Genome sequence data, databases and software

Tomato (*S. lycopersicum*) assembly release 2.40 and iTAG annotation release 2.3 (Sato et al., 2012) were retrieved from the SolGenomics Network (SGN; http://www.solgenomics.net). The potato (*S. tuberosum* group Phureja DM1-3 516R44 (CIP801092)) genome assembly v3 and annotation v3.4 (Xu et al., 2011) were retrieved from the Potato Genome Sequencing Consortium (PGSC; http://www.potatogenome.net). Where available, SGN Unigene builds of other solanaceous species were used (http://www.solgenomics.net; accessed on 7 October 2011). Other green plant genome data were taken from Phytozome (Goodstein et al., 2012) (http://www.phytozome.net; version 7). The rice (*O. sativa*) annotation of Phytozome was enhanced by incorporating the annotation of the Rice Annotation Project Database (Itoh et al., 2007; Tanaka et al., 2008). In addition, protein sequences from ChromDB (http://chromdb.org; accessed on 7 October 2011), UniRef100 (http://www.uniprot.org; accessed on 7 October 2011) and RefSeq (Pruitt et al., 2005) (accessed on 7 October 2011) were used. Arabidopsis genome data were obtained from TAIR (http://www.arabidopsis.org). Snf2 family analysis of Arabidopsis and rice was taken from the general Snf2 family protein resource (http://www.snf2.net) for reference (Flaus et al., 2006). Taxonomy information was obtained from the Tree-of-Life project (http://tolweb.org) and Phytozome.

2.2 Phylogenetic Analysis

Data preparation, conversion and filtering were performed with BioPerl (Stajich et al., 2002), Bio::Phylo (Vos et al., 2011) and custom Perl scripts. For the Snf2 gene calling in potato, potato protein sequences were determined by aligning all candidate Snf2 ATPase protein sequences against the potato genome using tblastn (Altschul et al., 1997) (E-value < 10). Hits were clustered into genomic regions with single linkage clustering (distance cut-off of 15kb) using C Clustering Library/Algorithm::Cluster (de Hoon et al., 2004). Final gene models were pre-

dicted with Exonerate (Slater and Birney, 2005) using the parameters »–model protein2genome –showvulgar no –showalignment no –showtargetgff yes« in the respective regions. Predicted potato gene models, unigenes, cDNAs and transcript sequences were translated using ESTScan2 (Lottaz et al., 2003) (additional parameter »-l 200«) with the tomato hexamer frequency model obtained from SGN (http://www.solgenomics.net).

Domain detection was performed with HMMER v3.0 (Finn et al., 2011) and InterProScan (Zdobnov and Apweiler, 2001) using InterPro Database version 35.0 (15 December 2011). Domain profiles were obtained from Pfam (Finn et al., 2010) and SMART (Letunic et al., 2009). A domain detection threshold of 1e-3 was used. It was adjusted with Arabidopsis as reference. To create an HMM model of the ATPase region, seed sequences were selected from UniProt, plant section, with the requirement of having the SNF2_N and Helicase_C domains present. Protein sequences smaller than 200 aa or with »putative«, »uncharacterized« or »predicted« in the description were excluded. The ATPase region was selected manually by identifying its conserved motifs Q-N (according to Flaus et al. (2006)) in the multiple alignment of the seed sequences. The model itself was trained with HMMER v3.0 (Finn et al., 2011), using hmmbuild with default parameters. A bitscore-based threshold of 200 was used to filter for Snf2 candidates. It was adjusted with Arabidopsis as reference.

Protein alignments were carried out with MAFFT v6.717b (Katoh et al., 2005) using the E-INS-i mode with a maximum of 1000 iterations. Phylogenetic trees were estimated with RAxML v7.7.5 (Stamatakis, 2006; Stamatakis et al., 2008) using the fast bootstrapping mode and the JTT matrix model (parameters were »-x 12345 -p 12345 -f a -m PROTGAMMAJTTF«).

Gene duplications and losses were evaluated with Notung (Chen et al., 2000). Intrinsically disordered regions were analyzed with FoldIndex (Prilusky et al., 2005) using a score cut-off of -0.2. Phylogenetic trees were visualized with Dendroscope v3 (Huson et al., 2007) or E.T.E. (Huerta-Cepas et al., 2010).

2.3 Expression data and analysis

Publicly available RNA-seq datasets from tomato (*Solanum lycopersicum* cv. Heinz 1706; data SRA049915) were retrieved from the SRA database (http://www.ncbi.nlm.nih.gov/sra). Sequence reads were mapped against the tomato reference genome (v. 2.40) with GSNAP (Wu and Nacu, 2010). The number of fragments per kb of exon per million fragments mapped (FPKM-values) were estimated for each gene model with cufflinks (Trapnell et al., 2010) on the basis of the iTAG 2.3 annotation and in-house enhanced gene models, where applicable. Conversions between SAM and BAM formatted alignments were performed with SAMtools (Li et al., 2009). Genes were categorized in three classes of expression: lowly expressed (FPKM 5), moderately expressed (5 < FPKM 200) and highly expressed (FPKM > 200). These categories are similar to a recent analysis of maize RNA-seq data (Hansey et al., 2012), however without the more stringent cut-off proposed. For comparison, the cut-off based on the 95% confidence level was also used for analysis.

2.4 RT-PCR analysis

Tomato cultivar Heinz plants were grown in a controlled greenhouse at 23 °C in long-day conditions (16 h light/8 h darkness). Seedlings were grown on 1/2 MS (Murashige & Skoog) agar plates supplemented with 1% sucrose in a growing chamber at 25 °C in long-day conditions. Total RNA was isolated from 10-day-old seedlings, as well as from flowers, leaves and green mature fruits from greenhouse-grown plants using the E.Z.N.A.™ Plant RNA Mini Kit (Omega Bio-Tek, Inc., USA) followed by on column DNase treatment (Qiagen, RNase-free DNase Set). One microgram of RNA was used for cDNA synthesis using the iScript™ cDNA Synthesis Kit (Bio-Rad Laboratories, Inc., USA) according to the recommendations of the manufacturer. Primers were designed with Primer3Plus (http://www.bioinformatics.nl/primer3plus; Untergasser et al., 2007) and checked for uniqueness in the tomato genome v. 2.40/ ITAG annotation v. 2.3 with the short-sequence blastn search of the BLAST 2.2.22+ toolkit (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Primers used are listed in table 3.S1. All primer pairs were validated by generating positive PCR reactions on genomic DNA. For RT-PCR, 2.5 µL of 10-times diluted cDNA was used. In all cases, actin was used as a reference gene (Løvdal and Lillo, 2009). The conditions used for all RT-PCR were: 95 °C for 4 min, followed by 25 to 35 cycles of 95 °C for 30 s, 60 °C for 30 s, 72 °C for 90 s and final extension at 72 °C for 7 min.

The activity of the primers was tested in a series of PCR reactions on genomic DNA with different concentrations of each primer. The concentration with highest band intensity was determined as the best primer concentration. The specificity of all primer pairs was established in a series of PCR reactions with tomato genomic DNA or cDNA to have only one single band of expected size (data not shown).

3 Results

3.1 Variable numbers of Snf2 family members in plant genomes

Snf2 family members in the predicted proteomes of 33 plant genomes including two green algae, were identified (table 3.S2). To prevent the inclusion of peptide fragments in the gene predictions, a cut-off of 200 amino acids (aa) was used, given that the conserved ATPase region has a length of about 400 aa (Eisen et al., 1995). All protein sequences longer than 200 aa were analyzed for the presence of the SNF2_N and Helicase_C domain. To be considered present, domains required a match in the protein sequence with an E-value smaller than 1e-3. Protein sequences containing at least one SNF2_N domain and one Helicase_C domain were listed as candidate Snf2 ATPase. To improve accuracy, a HMM model spanning the conserved ATPase region was created. The initial result set was filtered with this model and only candidates with a bitscore of at least 200 were used for fur-

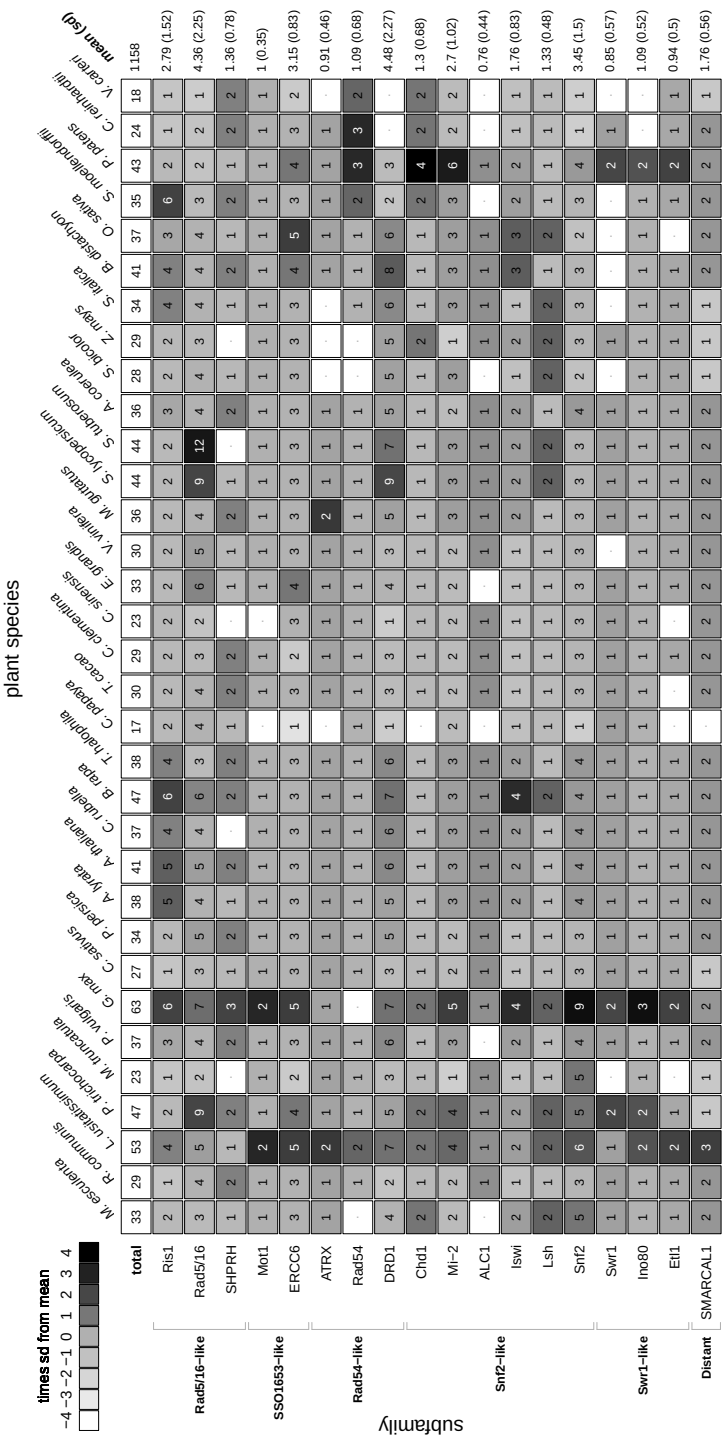
ther analyses. For Arabidopsis, all (41) previously known Snf2 genes (ChromDB; Gendler et al., 2008) were identified (fig. 3.1). In total, 1159 family members were identified (fig. 3.1).

The total number of candidate Snf2 ATPases in plant genomes (fig. 3.S2) shows considerable variation, ranging from 17-63 genes, with an interquartile range of 11, settled between 32 (Q1) and 43 (Q3). The papaya (*Carica papaya*) genome has only 17 candidate Snf2 family members, whereas in soybean (*Glycine max*, 63 members) and flax (*Linum usitatissimum*, 53 members) show an elevated number of family members. We identified 44 candidate Snf2 family members in the tomato genome (fig. 3.1), whereas the potato genome would carry only 23 candidate members that are also present in the official potato genome annotation. Given that both genomes are closely related in the *Solanum* genus, the surprising difference motivated an identification and re-calling of Snf2 genes in the potato genome. The re-calling identified 21 unannotated candidate Snf2 genes in the potato genome, in addition to the 23 from the first analysis. In other plant annotations, the number of potential Snf2 members was comparable between the genome annotation from Phytozome (Goodstein et al., 2012) and the re-calling (data not shown). Hence, all subsequent analyses were carried out with the set of 44 Snf2 family members in potato, the tomato annotation from ITAG and the annotation from Phytozome in all other cases.

3.2 Phylogenetic analysis

To infer evolutionary and potentially functional relationships of all plant candidate Snf2 genes, a phylogenetic tree was estimated on the basis of the conserved ATPase region of the protein sequence, including 30 aa flanking sequence on both sides to compensate for inaccuracies in domain prediction. To provide a more complete survey with focus on the *Solanum* genus, also transcriptome and unigene data (table 3.S2) were included. Each Snf2 subfamily was labeled according to the name of the Arabidopsis Snf2 subfamily in the relevant branch of the estimated tree. The unrooted tree summarizing the evolutionary relationships is presented in fig. 3.2.

Figure 3.1 ***(facing page)*****:** Distribution of Snf2 family members in plant genomes. Groupings and subfamilies on the left are named according to the Arabidopsis subfamily classification (Knizewski et al., 2008). Species names on the top are organized on the basis of their phylogenetic relationship according to Phytozome (Goodstein et al., 2012). Snf2 candidate member *Cre09.g390000.t1.1* (*Chlamydomonas reinhardtii*) could not be assigned to any subfamily and was excluded. Subfamily counts are shaded according to the deviation from the subfamily mean in standard deviations (sd). The total count is given on the top right cell. Mean and standard deviations per subfamily are indicated in the last column.


plant species
M. esculenta
R. communis
L. usitatissimum
P. trichocarpa
M. truncatula
P. vulgaris
G. max
C. sativus
P. persica
A. lyrata
A. thaliana
C. rubella
B. rapa
T. halophila
C. papaya
T. cacao
C. clementina
C. sinensis
E. grandis
V. vinifera
M. guttatus
S. lycopersicum
A. coerulea
S. bicolor
Z. mays
S. italica
B. distachyon
O. sativa
S. moellendorffii
P. patens
C. reinhardtii
V. carteri
mean (sd)
1158
2.79 (1.52)
4.36 (2.25)
1.36 (0.78)
1 (0.35)
3.15 (0.83)
0.91 (0.46)
1.09 (0.68)
4.48 (2.27)
1.3 (0.68)
2.7 (1.02)
0.76 (0.44)
1.76 (0.83)
1.33 (0.48)
3.45 (1.5)
0.85 (0.57)
1.09 (0.52)
0.94 (0.5)
1.76 (0.56)
total
Ris1
Rad5/16
SHPRH
Mot1
ERCC6
ATRX
Rad54
DRD1
Chd1
Mi−2
ALC1
Iswi
Lsh
Snf2
Swr1
Ino80
Etl1
SMARCAL1
Rad5/16−like
SSO1653−like
Rad54−like
Snf2−like
Swr1−like
Distant
subfamily
times sd from mean
−4 −3 −2 −1 0 1 2 3 4

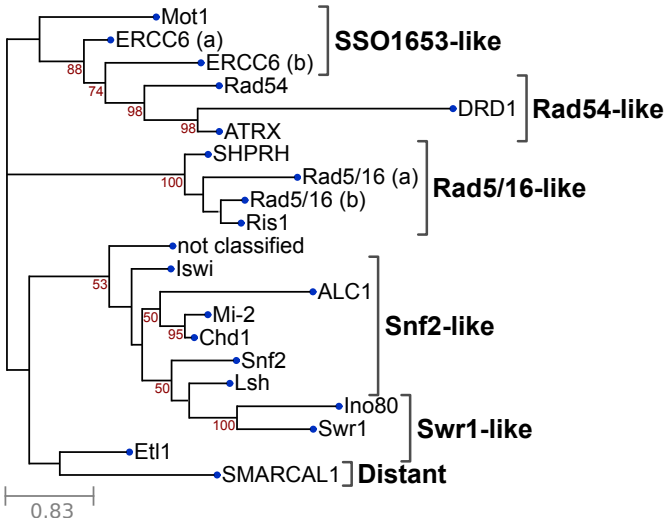


Figure 3.2: Unrooted phylogenetic tree of all candidate Snf2 genes in plant genomes. The full tree from which this subset was extracted is presented in fig. 3.S3. The subfamily branches were collapsed to a single node that represents the first split that is part of the subfamily branch. Confidence values (50-100) are indicated at the relevant splits of the branches. The tree is based on 100 bootstrap replicates. The leaf tagged »not classified« indicates candidate Snf2 members that are not part of a known subfamily, including *Cre09.g390000.t1.1* (*Chlamydomonas reinhardtii*) and members of sequence databases.

All 18 subfamilies identified are present in the tree and the overall tree topology of plant Snf2 genes is in agreement with earlier analyses (Flaus et al., 2006), although members of the subfamilies Rad 5/16 and ERCC6 were distributed over two different branches. In green algae, only 3 of the 18 subfamilies are not present (DRD1, ALC1 and Ino80), suggesting a high conservation of Snf2 ATPases in the plant kingdom. The distribution of genes over the various Snf2 subfamilies per plant species is presented in fig. 3.1. For this estimation, only whole genome data were included. Half of the subfamilies occur in relatively small numbers (mean < 2), whereas 19 of 33 plant species miss one or more of these subfamilies. Four subfamilies (mean ≥ 3) are large: DRD1, Rad 5/16, Snf2 and ERCC6. Largest is the plant-specific DRD1 subfamily (148 members, mean 4.48), followed by the Rad 5/16 subfamily (144 members, mean 4.36) and the Snf2 subfamily (114 members, mean 3.45). Eight Snf2 candidate members originating form ChromDB, RefSeq and UniRef100 and the Snf2 candidate member *Cre09.g390000.t1.1* (*Chlamydomonas reinhardtii*) could not be assigned to any subfamily (not classified). These members were not taken into account. More plant genomes will have to be sequenced to ascertain whether the Snf2 family member distribution reflects any phylogenetic bias in genome sequencing.

3.3 Snf2 family members involved in stress responses: DRD1 and Snf2

We focused further analyses on the two subfamilies reported to be connected to stress responses in plants, the DRD1 and Snf2 subfamilies (Huettel et al., 2007; Mlynárová et al., 2007; Walley et al., 2008; López et al., 2011) and on tomato and potato. Functional annotation of these subfamilies is guided by the functional information available for Arabidopsis genes.

DRD1 subfamily

In Arabidopsis, the DRD1 subfamily has six members. Tomato has eleven members and potato seven. To characterize the phylogenetic relationships between the DRD1 subfamily members of plant species in the Asterid clade (potato, tomato and *Mimulus guttatus*) and Arabidopsis as model plant at a high resolution, the further analysis was focused on these four plants. According to the species tree (fig. 3.S2), Mimulus is most close to the two solanaceous plants of interest. It has five DRD1 members.

In the unrooted phylogenetic tree based on the data from these four species (fig. 3.3), the DRD1 members could be grouped in three distinct branches, labeled a, b and c, each containing two Arabidopsis members. *AtCHR42* and *AtCLSY1* are in branch a, *AtCHR31* and *AtCHR40* in branch b, whereas *AtDRD1* and *AtCHR34* are in branch c. In all three branches, DRD1 members from tomato, potato and Mimulus are present. The tree shows that *AtCHR42* and *AtCLSY1* are in-paralogs (Koonin, 2005) with one ortholog in tomato, potato and Mimulus (fig. 3.3; branch a). Likewise, *AtDRD1* and *AtCHR34* are in-paralogs with also one ortholog in tomato, potato and Mimulus (fig. 3.3, branch c). It is apparent from the tree that branch b is the most complex. In addition to the two members of Arabidopsis in branch b, Mimulus has 3, potato 7 and tomato 9 members. The number of members in branch c is relatively stable in other plant species, ranging from 1 to 3 (mean 1.49, sd 1, tomato and potato excluded). This indicates a relative expansion of DRD1 ATPases in the tomato and potato genomes.

The potato/tomato members establish a separate sub-branch without members of either Arabidopsis or Mimulus suggesting independent evolution of DRD1 members in tomato and potato. Such evolution requires, the occurrence of a gene duplication in the common ancestor of all four species (labeled »ancient duplication« in fig. 3.3), followed by independent gene losses in all four species. The high confidence value (99 from 100) for the ancient duplication supports this scenario. Also analysis with Notung (Chen et al., 2000) supports the mutual gene loss scenario (details not shown). The evolutionary history of solanaceous DRD1 genes suggests specific functions for such genes in tomato and potato.

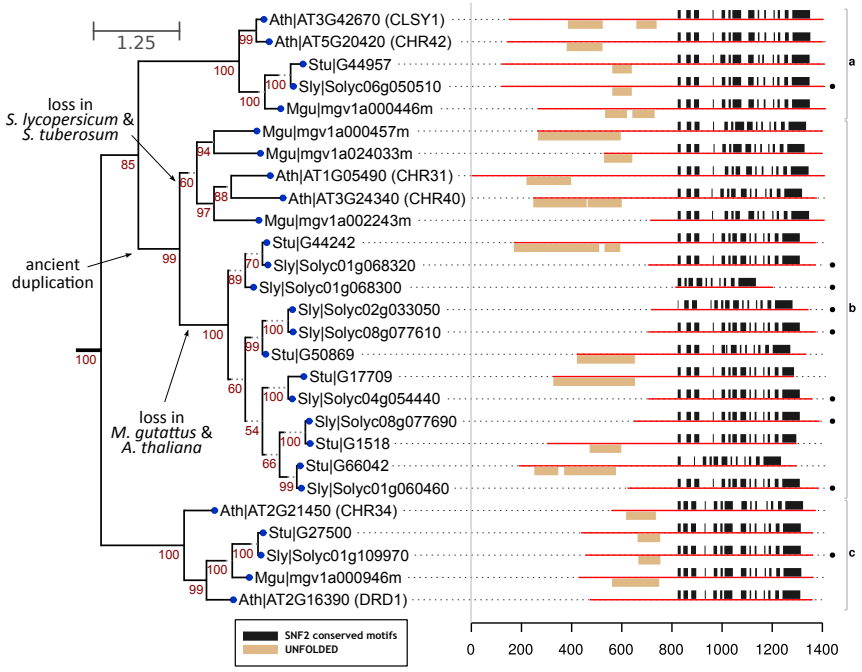


Figure 3.3: Analysis of the DRD1 subfamily in tomato, potato, Mimulus and Arabidopsis. The left side shows a detailed view of the DRD1 subfamily branch of an unrooted tree based on 1000 bootstraps of Snf2 data from *Arabidopsis thaliana* (Ath), *Mimulus guttatus* (Mgu), *Solanum lycopersicum* (Sly) and *Solanum tuberosum* (Stu). Confidence values (50-100) are given at the relevant branches of the tree. Identifiers give the name of the organism in three-letter abbreviations together with gene identifiers. The individual branches identified are indicated by letters in lowercase on the right side. To increase readability, some branch edges have been extended by dotted grey lines. These grey dotted lines are therefore not part of the estimated branch length. The right side shows structural elements (domains and unfolded regions) in the protein sequence of the DRD1 subfamily members in Arabidopsis, Mimulus, tomato and potato. Besides the ATPase region no other domains are present in these genes. A black dot at the right end of the figure indicates the expression of the respective gene in tomato based on the analysis of RNA-seq data.

To infer potential functions of the DRD1 subfamily members, we investigated the presence of additional structural/functional elements in the protein sequences. The DRD1 subfamily members of the four species here investigated had no accessory domains (fig. 3.3). In many cases, the N-terminal region of DRD1 subfamily members shows a predicted disordered region. In Arabidopsis, this applies to all DRD1 subfamily members, except for the AtDRD1 protein (fig. 3.3).

Snf2 subfamily

In Arabidopsis, the Snf2 subfamily has four members, while only three were found in tomato, potato and Mimulus. The tree estimated on data from these four species again shows three distinct branches (fig. 3.S4), labeled a, b and c, respectively. The Arabidopsis genes *AtCHR12* and *AtCHR23* cluster together (fig. 3.S4, branch a), in addition to single genes of the other species. It shows that *AtCHR12* and *AtCHR23* are in-paralogs with one ortholog in tomato, potato and Mimulus. The two Arabidopsis genes are likely to be the result of a gene duplication event specific to the Arabidopsis genus. The other Arabidopsis genes form one-to-one ortholog relationships with the respective tomato, potato and Mimulus genes (fig. 3.S4). The evolutionary history of the Snf2 subfamily is therefore overall much less eventful than the history of the DRD1 subfamily.

AtCHR12 and AtCHR23 (branch c) carry an unfolded region at the C-terminal end which is not present in any of the other members of the branch (fig. 3.S4). The difference in length of the proteins in this subfamily is remarkable. Whereas branch a consists of relatively short proteins of approx. 1100 amino acids, branch b is characterized by very large proteins, the largest one (AtSYD) carrying 3574 amino acids. AtSYD has a considerably larger C-terminal end compared to all orthologs in its branch and compared to all members in the subfamily. Yet it only shows an unfolded region in the C-terminal end and no other functional or structural domains.

3.4 Expression analysis of DRD1 and Snf2 subfamilies

Expression characteristics could also help elucidating the biological function of DRD1 and Snf2 subfamily members. We evaluated the expression profile of these genes in tomato public-domain RNA-seq libraries (Sato et al., 2012) for flowers, roots, leaves and various stages of fruit of tomato cv Heinz 1706 (table 3.S3). The FPKM-values of all libraries were calculated and visualized as heat map for the DRD1 and Snf2 subfamilies (fig. 3.S5). All three Snf2 subfamily members of tomato are moderately expressed in the majority of the libraries analyzed. No tissue specificity and/or developmental control are apparent, suggesting a constitutive expression.

In contrast, expression of members of the DRD1 subfamily is more heterogeneous. The highest and most diversely expressed DRD1 subfamily genes are *Solyc01g109970* (branch c) and *Solyc06g050510* (branch a). *Solyc01g109970* is constitutively expressed in all libraries with FPKM values from 5 (leaves) to 37 (fully

ripe fruit). Expression of the *Solyc06g050510* gene is similar, with the highest FPKM-value of 30 in roots, mature green fruits, immature fruits and 3-cm fruits. The lowest expression shows this gene in breaker and fully ripe fruits (FPKM around 7). The gene *Solyc01g068320* shows low specific expression in flower and flower bud tissue. The other five members that constitute the solanaceous-specific expansion of branch b in tomato show extremely low expression.

To confirm these expression characteristics, semi-quantitative RT-PCR was performed on leaves, flowers and mature fruits. To be able to extend the analysis to early stages of plant development, 10-day-old in-vitro-grown seedlings were included. RT-PCR analysis of the three Snf2 genes confirmed expression in all four tissues analyzed, in concordance with the RNA-seq analysis (fig. 3.4). It also largely confirmed the RNA-seq results of the DRD1 subfamily genes (fig. 3.4). *Solyc08g077690* is expressed in all tissues examined at the highest level shown by any member in this branch. Expression of *Solyc01g068320* is restricted to flower and fruit tissue, the latter at lower levels. For *Solyc01g068300* RT-PCR shows a relatively easily detectable product in all tissues except seedlings. Also expression of *Solyc02g033050*, *Solyc01g060460* and *Solyc08g077610* is detectable by RT-PCR in all tissues. However, the level of expression is low to very low, approaching the lower limit of reliable detection. Gene *Solyc04g054440* is very lowly expressed in possibly only fruits. The highly variable expression patterns of the various DRD1 subfamily genes indicate that the putative function of the encoded DRD1 proteins is likely to be subtle in terms of time or location.

4 Discussion

The Snf2 family of ATPases is a large family of chromatin remodeling enzymes that have versatile roles in a variety of fundamental processes in growth and development. In plants, little is known about the function of individual members of this family, although notably in Arabidopsis functional relationships with gene regulation, DNA recombination, DNA repair and stress tolerance have been reported (Bezhani et al., 2007; Mlynárová et al., 2007; Walley et al., 2008). Here, we present the first comprehensive comparative analysis of all Snf2 genes in 33 sequenced and annotated plant genomes, including two green algae. We have identified and analyzed 1159 potential candidate Snf2 family ATPases, of which all but one could be placed in previously established groups and subfamilies and represent genuine plant Snf2 genes. The variation in numbers of Snf2 genes is large, ranging from 17 in papaya to 63 in soybean. This suggests a broad functional diversification of this gene family in the plant kingdom. The high member counts in flax and soybean may originate from recent whole-genome duplications in both species (Schmutz et al., 2010; Wang et al., 2012b).

Our results for rice show considerably more differences when compared to another recent study of Snf2 family genes (Li et al., 2011), in which 39 putative Snf2 family genes are identified. The overall tree presented (Li et al., 2011) does not seem to agree well with the subfamily classification. An example is a branch

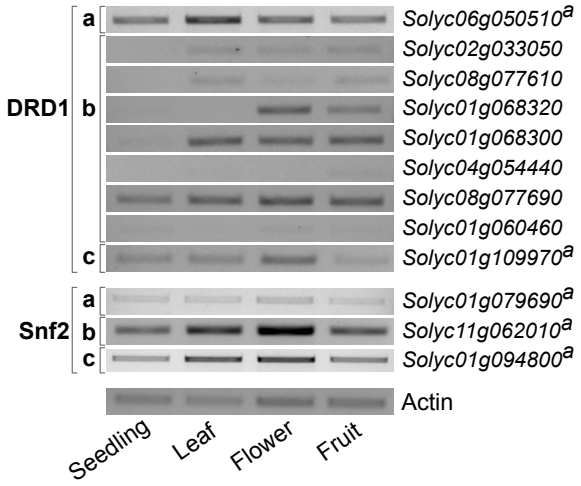


Figure 3.4: RT-PCR expression analysis of DRD1 and Snf2 subfamilies of tomato Snf2 ATPase genes. The tissue used is indicated on the x-axis. The individual genes are indicated of the right, the branches identified on the left. The expression of the actin gene (25 cycles) was used as control (lower panel). The number of PCR cycles used for the analysis of the individual gene was adjusted to generate a detectable amount of PCR product. For most of genes, 35 cycles were used. Genes marked with superscript a (^{a}) were amplified with 29 cycles. For the actin gene 25 cycles were used.

containing rice genes (*Os02g0114000* (Snf2), *Os01g0779400* (Ris1), *Os05g0150300* (Iswi), *Os05g0392400* (DRD1) and *Os07g0497000* (Mi-2)) that are distributed in five different subfamilies according to our classification. Possible explanations for the differences are phylogenetic tree modeling based on the complete protein sequence rather than the conserved region, and/or the use of another rice annotation (Tanaka et al., 2008; http://rapdb.dna.affrc.go.jp).

Surprising sources of error in Snf2 family member identification are the publicly available genome assemblies and annotations. Our example in potato highlights the better performance of gene calling within a protein family opposed to automatic gene calling. Half of the Snf2 family members are absent from the current genome annotation of potato. Assembly and calling of Snf2 genes may be troublesome for the partly automated pipelines in place for overall genome assembly and annotation, despite manual curation effort. Here we show increased sensitivity of candidate Snf2 family gene identification by iterative rounds of homology-based gene prediction. This approach minimizes errors in the predicted coding region that would affect the multiple sequence alignment and phylogenetic reconstruction considerably. For Arabidopsis and rice, the plant species with the richest set of annotation and experimental data, inferred gene models were consistent with the currently available high-quality annotations (not shown). Therefore, the annotation of the potato Snf2 family is likely to have improved markedly with the

homology-based prediction routine put in place and is recommended for future analyses. The accuracy of the prediction of the proper coding region is not likely to be improved with the help of (family-) specific gene models or better hexamer models. Such homology-based prediction will not safeguard against errors in assembly.

Not anticipated from the earlier analyses of Snf2 family genes (Flaus et al., 2006) is the relative expansion of the DRD1 (148 genes), Rad5/16 (144 genes) and Snf2 (114 genes) subfamilies in plant genomes. So far, members two of these subfamilies have been associated with environmental stress responses in Arabidopsis, possibly indicating the relative importance of chromatin remodeling in combating environmental stress in plants. The most abundant subfamily, DRD1, has evolved from apparent non-existence in non-plant species (http://www.snf2.net) and lower plants, such as *Volvox carteri* and *Chlamydomonas reinhardtii*, to the largest and most diverse subfamily in current-day higher plants. It indicates that the DRD1 protein has become an important and possibly diversified asset in the regulation of plant growth and development. Within the expanded DRD1 subfamily, tomato has one of the highest member count of all genomes analyzed, whereas potato, even if higher than average, does not reach this high member count. However, the expansion within this subfamily was not uniform, and while some seem to be unique for Solanaceae (fig. 3.3, branch b), in other cases, the genome of Arabidopsis carries two genes whereas potato and tomato have only one.

The DRD1 subfamily tree suggests a complex evolutionary history involving a series of independent gene losses, duplication and genomic reshuffling events (recombination, transposition) resulting in a relative expansion of genes in notably tomato. It suggests that the DRD1 subfamily has gained additional functionality in tomato. The results suggest that the relative expansion has been specific for the Solanaceae, although more solanaceous genomes (*S. pennellii*, *N. tabacum*, *S. pimpinellifolium*) are required to validate the specificity of this expansion for Solanaceae in general, or for a given species in particular.

It is supposed that the conserved ATPase domain is responsible for the energy release of DRD1 proteins, whereas other parts of the protein specify interaction partners, DNA specificity and/or sub-nuclear localization. The presence of a disordered region that may be characteristic for the expanded branch b. The differences in structure, if any, are so subtle or complex that it is difficult to associate particular sequence determinants with function. The unfolded regions occur regularly at approximately the same position in the N-terminal regions of DRD1 proteins. Such unfolded regions may help or direct protein-protein or protein-nucleic acid interactions (Ward et al., 2004; Uversky and Dunker, 2010; Bolanos-Garcia et al., 2012). Disordered regions in the DRD1 genes may therefore interface the ATPase domain to other proteins or DNA/RNA molecules (Nguyen Ba et al., 2012). This may help to specify interaction partners, whereas the lack of accessory domains indicates that ATPase-mediated remodeling is the main enzymatic function of these DRD1 subfamily members. New interaction partners could determine involvement of DRD1 proteins in new biological processes or conditions.

Given the complex evolution and expression pattern of DRD1 genes in tomato, it is not as straightforward as for the Snf2 subfamily to transfer the function of Arabidopsis genes to the orthologous tomato genes. In Arabidopsis, several genes of this subfamily are important components of RNA-directed DNA methylation (RdDM), the pathway in which specific genomic loci are targeted for methylation by 24 bases small interfering RNAs (siRNA) (Huettel et al., 2007; Mahfouz, 2010). RdDM operates in many organisms and requires common components such as DNA methyltransferases, histone modifying enzymes and RNAi proteins.

The genes of branch a, *AtCLSY1* and *AtCHR42* were found in the Pol-IV polymerase protein complex (Law et al., 2010), the RNA polymerase thought to initiate the biogenesis of the targeting siRNAs (Pikaard et al., 2008). In the same complex, *AtCHR31* and *AtCHR40* (branch b) are also present, suggesting they play a role in the same RdDM pathway (Law et al., 2011). In addition to siRNAs, RdDM is also associated with the accumulation of so-called intergenic noncoding (IGN) transcripts that involves the plant specific RNA-polymerase Pol-V (Wierzbicki et al., 2008). *DRD1* (branch c) was identified in a protein complex critical for the production of Pol-V dependent IGN transcripts (Law et al., 2010). Recently, this gene was also established as an important player in plant immunity. Its knockout mutant showed increased susceptibility to the fungal pathogen *Plectosphaerella cucumerina* (López et al., 2011). The second gene of Arabidopsis branch c, *At2g21450*, was shown to be modulated during early embryogenesis, suggesting a role after fertilization (Xiang et al., 2011). Related functions affecting small RNA accumulation and cytosine methylation have been shown for *RMR1*, an Snf2 ortholog in *Zea mays* (maize), in the context of paramutation (Hale et al., 2007). As five out of six Arabidopsis DRD1 genes and *RMR1* are implicated in RdDM pathways, a similar function of this subfamily in tomato is likely.

Why tomato would need so much more active DRD1 genes than Arabidopsis? Possibly the continued selection for traits in tomato as agricultural crop has been the driving force for such developments. The functions assigned so far in Arabidopsis point in the direction of protection against biotic and abiotic stresses. The comprehensive analysis here presented shows the evolution and presence of Snf2 genes in plants. Closer evaluation of, e.g. DRD1 subfamily members, could make suitable targets for breeding and plant improvement.

Acknowledgements

The authors would like to thank Dr. Berend Snel and Michael Seidl (Utrecht University, the Netherlands) for the help with the interpretation of the phylogenetic analyses. We also would like to thank Dr. Andrew Flaus (NUI Galway, Ireland) for his valuable comments.

5 Supporting Information

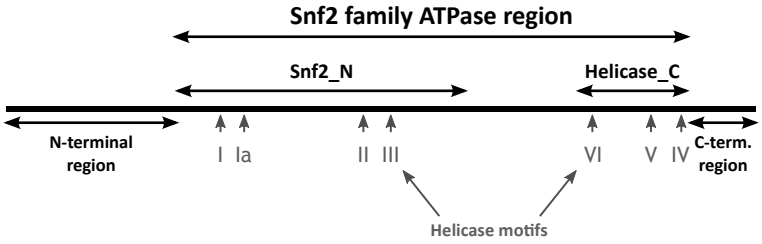


Figure 3.S1: Schematic layout of Snf2 family ATPases. The conserved Snf2 family ATPase region is part of the protein and consists of two Pfam domains, Snf2_N and Helicase_C, in which seven helicase motifs are present. The average size of the Snf2 family ATPase region is approx. 400aa (Eisen et al., 1995). In individual proteins, the N-terminal or C-terminal region can be very small (Flaus and Owen-Hughes, 2011).

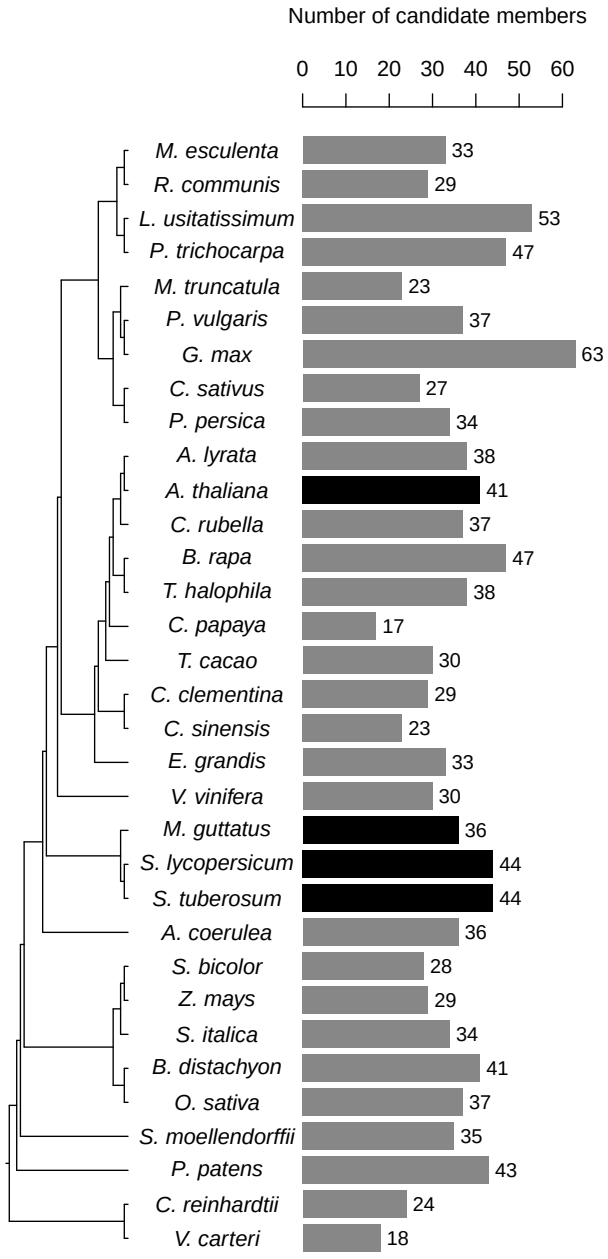


Figure 3.S2: The number of candidate Snf2 genes in annotated plant genomes. The total number of genes estimated for a genome is plotted above the bar in the histogram. Plant species included are organized on the basis of the position in the tree of life (shown at the left). The four species given most attention in this study (Arabidopsis, potato, tomato and *Mimulus guttatus*) are given in black.

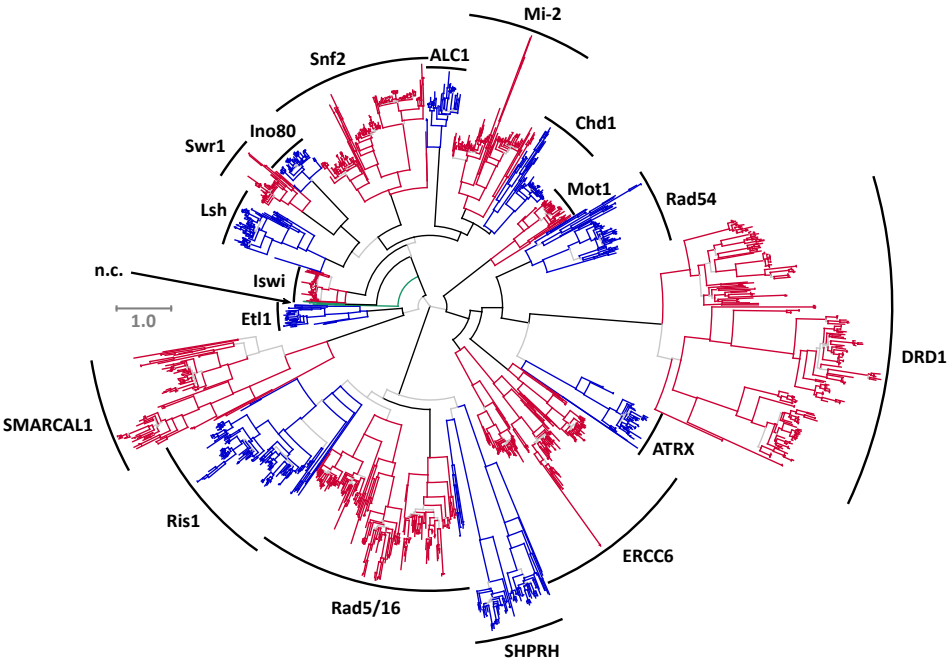


Figure 3.S3: Full phylogenetic tree of all plant Snf2 candidates. The tree is based on the plant data listed in table 3.S2 and calculated with 100 bootstraps due to computational constraints. Branches with a confidence lower than 50 are marked in grey. Members not classified (n.c.) into any subfamily are indicated in light green. To increase readability, the colors of subfamily branches alternate between blue and red.

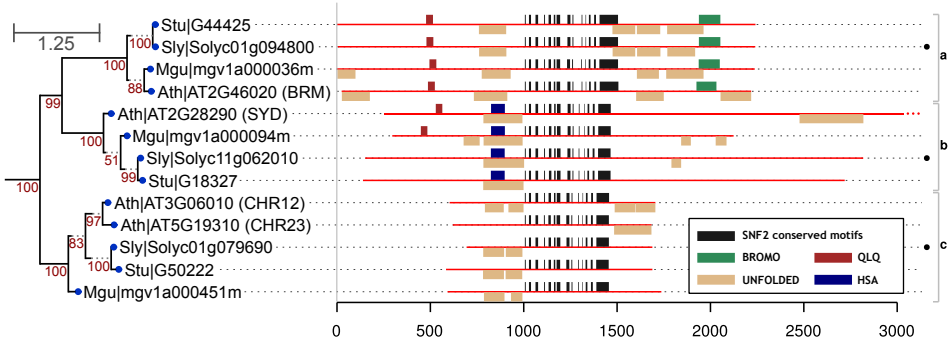


Figure 3.S4: Analysis of the Snf2 subfamily in tomato, potato, Mimulus and Arabidopsis. The left side shows a detailed view of the DRD1 subfamily branch of an unrooted tree based on 1000 bootstraps of Snf2 data from *Arabidopsis thaliana* (Ath), *Mimulus guttatus* (Mgu), *Solanum lycopersicum* (Sly) and *Solanum tuberosum* (Stu). Confidence values (50-100) are given at the relevant branches of the tree. Identifiers give the name of the organism in three-letter abbreviations together with gene identifiers. The individual branches identified are indicated by letters in lowercase on the right side. To increase readability, some branch edges have been extended by dotted grey lines. These grey dotted lines are therefore not part of the estimated branch length. The right side shows structural elements in the protein sequence of the Snf2 subfamily members in Arabidopsis, Mimulus, tomato and potato. The individual branches identified are indicated by letters in lowercase. Besides the ATPase region, BROMO (protein-histone interaction), QLQ (protein-protein interaction) and HSA (DNA-binding) domains are present in several members. A black dot at the right end of the figure indicates the expression of the respective gene in tomato based on the analysis of RNA-seq data.

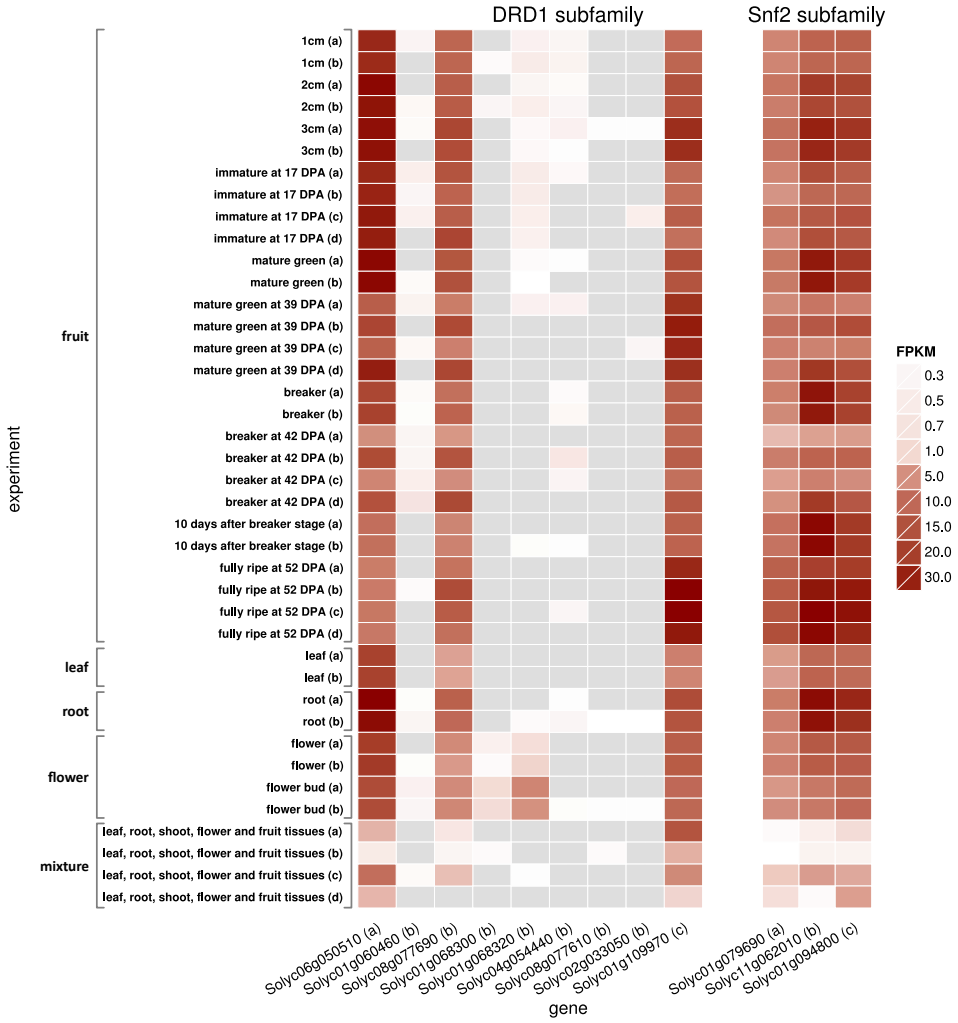


Figure 3.S5: Heat map of the RNA-seq expression data of the tomato DRD1 & Snf2 subfamily genes. The expression is indicated as fragments per kb exon model per million mapped reads-value (FPKM-value). No cut-off was applied. Grey areas correspond to FPKM-values of 0. Gene identifiers are indicated on the x-axis with the corresponding branch name given between brackets. The biological material used to generate the RNA-seq libraries is given on the y-axis. Replicates are indicated by lowercase letters. Details on the RNA-seq libraries used are given in table 3.S3.

Table 3.S1: Primers used for RT-PCR analysis. The primer sequence of the forward (F) and reversed (R) primer is given for each gene identifier.

direction	sequence	gene ID
F	GAAACAGAGAAGCGCATAGTTTT	Solyc01g068300
R	GTTTTGGAGGTTGGTTACAAGAA	Solyc01g068300
F	GGAAATTTAAATGACTGTCAGATGG	Solyc01g068320
R	CAAGTGAATTACAGTGTCCCTTATAC	Solyc01g068320
F	GAATCTATCAGTTTCGCCGATG	Solyc02g033050
R	GCTTACGTTCTTTACATTTTCGCTAC	Solyc02g033050
F	GAGACATAAGTGGCTGTGAGATG	Solyc04g054440
R	CACTACATCTATGAACAAATGGTGA	Solyc04g054440
F	CGGTGATGCAGAGTGGAG	Solyc08g077610
R	GAATATCCCTAAGCTCTTCCAACG	Solyc08g077610
F	GAGCAAGTACATCTTCCCTCCA	Solyc08g077690
R	AGGATGAACAGAGATTAGAGACACC	Solyc08g077690
F	GAAGAAGGGAAAAAGGAGTCAAA	Solyc01g060460
R	TAACCATCCCCATCTTCTCC	Solyc01g060460
F	CCACTTGATGTTGATGTTCCTG	Solyc06g050510
R	ACCTTTTCCCTTAGAACCTCTCC	Solyc06g050510
F	GGATGGACAGGAAACTAACAACA	Solyc01g109970
R	CACTACCAACATTGTCACACACA	Solyc01g109970
F	CCAAAATAAAAAGGAAACGCAGT	Solyc01g094800
R	CCCAACTTCTCTCTATCTTTTCTTTTC	Solyc01g094800
F	CTGTAATGGCGTCTCCTGCT	Solyc11g062010
R	GATTTCCACTGTTGCCTCAAG	Solyc11g062010
F	GTTCAGGCTTGGCATGGAA	Solyc01g079690
R	CCGATAAGTGTGATGTCTCTC	Solyc01g079690

Table 3.S2: Plant data included in the analyses. Sources are the Phytozome annotation (indicated as genome), SGN unigenes (indicated as unigene), de-novo assembled transcriptomes (indicated as transcript) and reference databases (indicated as database). The differences in Snf2 members between the annotation (first value) and the homology-based re-analysis here presented (second value) are indicated for potato (*Solanum tuberosum*).

species/ database	genome size (Mb)	no. of predicted gene models	no. of Snf2 members	data type	ref.
Antirrhinum majus	n/a	n/a	0	unigene	Bombarely et al., 2011
Aquilegia coerulea	302	24823	36	genome	Goodstein et al., 2012
Arabidopsis lyrata	230	32670	38	genome	Feuillet et al., 2010
Arabidopsis thaliana	125	27416	41	genome	Feuillet et al., 2010
Brachypodium distachyon	300	26552	41	genome	Feuillet et al., 2010
Brassica rapa	530	40905	47	genome	Feuillet et al., 2010
Capsella rubella	250	26521	37	genome	Feuillet et al., 2010
Capsicum annuum	2700	n/a	0	unigene	Bombarely et al., 2011
Carica papaya	372	27769	17	genome	Feuillet et al., 2010
Chlamydomonas reinhardtii	112	17114	25	genome	Goodstein et al., 2012
ChromDB (plants only)	n/a	8618	377	database	Gendler et al., 2008
Citrus clementina	296	25385	29	genome	Goodstein et al., 2012
Citrus sinensis	382	25379	23	genome	Feuillet et al., 2010
Coffea arabica	n/a	n/a	0	unigene	Bombarely et al., 2011
Coffea canephora	n/a	n/a	1	unigene	Bombarely et al., 2011
Cucumis sativus	367	21646	27	genome	Feuillet et al., 2010
Eucalyptus grandis	600	36376	33	genome	Feuillet et al., 2010
Glycine max	1100	46367	63	genome	Feuillet et al., 2010

Continued on next page

Table 3.S2 – Continued

species/ database	genome size (Mb)	no. of predicted gene models	no. of Snf2 members	data type	ref.
Ipomoea batatas	n/a	n/a	0	unigene	Bombarely et al., 2011
Linum usitatissimum	350	43471	53	genome	Goodstein et al., 2012
Manihot esculenta	770	30666	33	genome	Feuillet et al., 2010
Medicago truncatula	500	50962	23	genome	Feuillet et al., 2010
Mimulus guttatus	430	26718	36	genome	Feuillet et al., 2010
Nicotiana benthamiana	n/a	n/a	0	unigene	Bombarely et al., 2011
Nicotiana sylvestris	n/a	n/a	0	unigene	Bombarely et al., 2011
Nicotiana tabacum	n/a	n/a	4	unigene	Bombarely et al., 2011
Oryza sativa	433	55986	37	genome	Feuillet et al., 2010
Petunia hybrid cultivar	n/a	n/a	0	unigene	Bombarely et al., 2011
Phaseolus vulgaris	487	26374	37	genome	Goodstein et al., 2012
Physcomitrella patens	480	32273	43	genome	Goodstein et al., 2012
Populus trichocarpa	485	40668	47	genome	Feuillet et al., 2010
Prunus persica	220	27864	34	genome	Feuillet et al., 2010
RefSeq (plants only)	n/a	519211	195	database	Pruitt et al., 2012
Ricinus communis	400	31221	29	genome	Feuillet et al., 2010
Selaginella moellendorffii	213	22285	35	genome	Goodstein et al., 2012
Setaria italica	515	35471	34	genome	Feuillet et al., 2010
Solanum dulcamara	n/a	14288	12	transcript	Bombarely et al., 2011
Solanum lycopersicum	900	34727	44	genome	Sato et al., 2012
Solanum melongena	1100	n/a	0	unigene	Bombarely et al., 2011

Continued on next page

Table 3.S2 – Continued

species/ database	genome size (Mb)	no. of predicted gene models	no. of Snf2 members	data type	ref.
Solanum peruvianum	n/a	17280	34	transcript	Bombarely et al., 2011
Solanum tuberosum	840	39031	23/44	genome	Feuillet et al., 2010
Sorghum bicolor	770	27608	28	genome	Feuillet et al., 2010
Thellungiella halophila	243	26351	38	genome	Goodstein et al., 2012
Theobroma cacao	430	46143	30	genome	Argout et al., 2011
UniRef100 (plants only)	n/a	591965	46	database	Suzek et al., 2007
Vitis vinifera	475	26346	30	genome	Jaillon et al., 2007
Volvox carteri	131	14971	18	genome	Goodstein et al., 2012
Zea mays	2500	39656	29	genome	Feuillet et al., 2010

Table 3.S3: RNA-seq libraries included in the analysis. Data are from the short read archive (SRA; http://www.ncbi.nlm.nih.gov/sra). The library and sample IDs refer to the run and sample identifiers in SRA, respectively.

name	library ID	sample ID
1cm (a)	SRR404317	SRS291272
1cm (b)	SRR404318	SRS291272
2cm (a)	SRR404319	SRS291273
2cm (b)	SRR404320	SRS291273
3cm (a)	SRR404321	SRS291274
3cm (b)	SRR404322	SRS291274
immature at 17 DPA (a)	SRR346617	SRS265321
immature at 17 DPA (b)	SRR346618	SRS265321
immature at 17 DPA (c)	SRR346619	SRS265321
immature at 17 DPA (d)	SRR346620	SRS265321
mature green (a)	SRR404324	SRS291275
mature green (b)	SRR404325	SRS291275
mature green at 39 DPA (a)	SRR346621	SRS265322
mature green at 39 DPA (b)	SRR346622	SRS265322
mature green at 39 DPA (c)	SRR346623	SRS265322
mature green at 39 DPA (d)	SRR346624	SRS265322
breaker (a)	SRR404326	SRS291276
breaker (b)	SRR404327	SRS291276
breaker at 42 DPA (a)	SRR346625	SRS265323
breaker at 42 DPA (b)	SRR346626	SRS265323
breaker at 42 DPA (c)	SRR346627	SRS265323
breaker at 42 DPA (d)	SRR346628	SRS265323
10 days after breaker stage (a)	SRR404328	SRS291277
10 days after breaker stage (b)	SRR404329	SRS291277
fully ripe at 52 DPA (a)	SRR346629	SRS265324
fully ripe at 52 DPA (b)	SRR346630	SRS265324
fully ripe at 52 DPA (c)	SRR346631	SRS265324
fully ripe at 52 DPA (d)	SRR346632	SRS265324
leaf (a)	SRR404309	SRS291268
leaf (b)	SRR404310	SRS291268
root (a)	SRR404311	SRS291269
root (b)	SRR404312	SRS291269
flower (a)	SRR404313	SRS291270
flower (b)	SRR404314	SRS291270
flower bud (a)	SRR404315	SRS291271
flower bud (b)	SRR404316	SRS291271
leaf, root, shoot, flower and fruit tissues (a)	SRR346633	SRS265325
leaf, root, shoot, flower and fruit tissues (b)	SRR346634	SRS265325
leaf, root, shoot, flower and fruit tissues (c)	SRR346635	SRS265325
leaf, root, shoot, flower and fruit tissues (d)	SRR346636	SRS265325

Dataset S1. Text file with custom predicted gene models of *Solanum tuberosum*.

Dataset S2. Text file of the multiple alignment of all plant Snf2 candidates.

Dataset S3. Phylogenetic tree of all plant Snf2 candidates in NEWICK format.

The supplemental datasets are available as part of the online publication (DOI 10.1371/journal.pone.0081147).

Chapter 4

Biological process annotation of proteins across the plant kingdom

Abstract

Accurate annotation of protein function is key to understanding life at the molecular level, but automated annotation of functions is challenging. We here demonstrate the combination of a method for protein function annotation that uses network information to predict the biological processes a protein is involved in with a sequence-based prediction method. The combined function prediction is based on co-expression networks and combines the network-based prediction method BMRF with the sequence-based prediction method Argot2. The combination shows significantly improved performance compared to each of the methods separately, as well as compared to Blast2GO. The approach was applied to predict biological processes for the proteomes of rice, barrel clover, poplar, soybean and tomato. The novel function predictions are available at www.ab.wur.nl/bmrf. Analysis of the relationships between sequence similarity and predicted function similarity identifies numerous cases of divergence of biological processes in which proteins are involved, in spite of sequence similarity. This shows that the integration of network-based and sequence-based function prediction is necessary to optimize the analysis of evolutionary relationships. Examples of potential divergence are identified for various biological processes, notably for processes related to cell development, regulation, and response to chemical stimulus. Such divergence in biological process annotation for proteins with similar sequences should be taken into account when analyzing plant gene and genome evolution.

Bargsten, J. W., Severing, E. I., Nap, J.-P., Sanchez-Perez, G. F., and van Dijk, A. D. J. (2013). *Current Plant Biology*, in press.

1 Introduction

The amount of plant genome data grows disproportional to the amount of available experimental data on these genomes (du Plessis et al., 2011; De Bodt et al., 2012; Goodstein et al., 2012; Schatz et al., 2012; Van Bel et al., 2012). To connect this ever increasing amount of genome data to plant biology, structural gene annotation followed by function annotation is imperative. For example, the identification of candidate genes involved in a trait-of-interest greatly benefits from gene function annotation (Monclus et al., 2012). In the context of the study of genome evolution, gene function annotations are necessary in order to enable comparison between sets of genes with different evolutionary histories, e.g. those retained vs. those lost after duplication (De Smet et al., 2013). To annotate gene or protein function, experimental data, if available, can be used to annotate gene or protein function. However, the scarcity of experimental data highlights the attractiveness of computational approaches to assist in gene function annotation (Rhee and Mutwil, 2014). Indeed, newly sequenced genomes are in general accompanied by a function annotation which heavily relies on computational predictions. Such automated annotations are delivered by a variety of approaches, often without much knowledge about their reliability. For studying plant genomes and plant genome evolution, reliable function annotation is therefore a major challenge.

One way to annotate proteins without experimental data is to infer function from sequence data (du Plessis et al., 2011). The de facto standard to capture function annotation today is the Gene Ontology (GO), in particular, the Molecular Function (MF) and Biological Process (BP) sub-ontologies (Gene Ontology Consortium, 2000). MF describes activities, such as catalytic or binding activities, that occur at the molecular level, whereas BP describes a series of events accomplished by one or more ordered assemblies of molecular functions (Gene Ontology Consortium, 2000). Compared to MF, terms in the BP ontology are generally associated with more conceptual and abstract levels of function. The prediction of BP terms can depend on the cellular and organismal context (Radivojac et al., 2013). Therefore, BP terms tend to be poorly predicted by methods based on sequence similarity only, such as BLAST (Altschul et al., 1990; Radivojac et al., 2013). The reliability of BP predictions increases with advanced approaches that employ e.g. phylogenetic frameworks (Martin et al., 2004; Clark and Radivojac, 2011) or network data such as protein-protein-interactions (Vazquez et al., 2003).

We recently developed a protein function prediction method for BP terms called Bayesian Markov Random Field (BMRF) (Kourmpetis et al., 2010), which uses network data as input. In BMRF, each protein is represented as a node in the network, and connections in the network indicate functional relationships between proteins. Networks can be based on e.g. protein-protein interactions or co-expression data. BMRF uses existing BP annotations for proteins in the network to infer biological processes for unannotated proteins in that network. To do so, BMRF uses a statistical model describing how likely neighbors are to participate in the same BP; this constitutes the Markov Random Field. Existing BP annotations

are used as »seed« or »training« data, providing a set of initial labels for the Markov Random Field. Parameters in the statistical model are trained using a Bayesian approach by performing simultaneous estimation of the model parameters and prediction of protein functions. Importantly, BMRF can transfer functional information beyond direct interactions. Therefore, it is able to generate function predictions for proteins that are only linked with other proteins with unknown function.

In the Critical Assessment of Function Annotations (CAFA) protein function prediction challenge (Radivojac et al., 2013) BMRF obtained particularly good performance in human (first place) and Arabidopsis (second place) for BP term prediction (Radivojac et al., 2013). In these species, BMRF performance benefits from the wealth of existing function annotation, i.e. experimental data. Because of its dependence on training data, function annotation for species with more sparse function annotation is challenging for BMRF. To improve the prediction performance in sparsely annotated species, we present here a strategy to combine BMRF with the sequence-based function prediction method Argot2 (Falda et al., 2012). Argot2 was among the top performing sequence-based algorithms in the CAFA category »eukaryotic BP«. In its computational approach Argot2 is complementary to BMRF, because it is purely sequence-based.

We demonstrate that the combination of Argot2 and BMRF has a markedly better function prediction performance than each method separately. This integrated method was applied to predict BP terms for proteins in five plant species, *Medicago truncatula* (barrel clover), *Oryza sativa* (rice), *Glycine max* (soybean), *Populus trichocarpa* (poplar) and *Solanum lycopersicum* (tomato), using microarray co-expression networks as input. Numerous new proteins were associated with specific biological processes, such as seed development in rice or nitrogen fixation in Medicago. By comparison between sequence divergence and predicted function divergence, numerous cases of putative neo-functionalization involving various biological processes were identified. This new method and the resulting set of predicted gene functions will be of great value in capitalizing on the large amount of plant genome data that is currently being generated for the study of the evolution of genome and gene function.

2 Materials and Methods

2.1 Function prediction methods and their integration

BMRF uses network data as input. Each protein is represented as a node in the network, and connections in the network indicate functional relationships between proteins. A statistical model (Markov Random Field) describes how involvement of a protein in a particular BP influences the probability that its neighbors in the network are also involved in that BP. The parameters in the statistical model describe for each BP how strongly neighbors influence each other. Parameter values are trained using a Bayesian approach by performing simultaneous estimation of

the model parameters and prediction of protein functions. This strategy needs a set of known protein functions as initial labeling of the network. Argot2 is a purely sequence-based prediction method, using searches of the UniProt and Pfam databases as input. To combine these two methods, two strategies were applied. In the first integration method, for each biological process, ranks for the different proteins were obtained from both BMRF and Argot2, by ordering the proteins based on their score for that process. These ranks were added to obtain a final ranking, which was used as the prediction score for that biological process. In a second integration strategy, initial predictions were generated with Argot2. These were supplied to BMRF as training data, meaning that the initial labeling of the nodes in the network was based on the Argot2 predictions.

2.2 Sequence and domain data

Sequence data for Arabidopsis, rice, soybean and *Medicago truncatula* were obtained from the Phytozome database v8.0 (Goodstein et al., 2012). Poplar sequence data were downloaded from the JGI (ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/annotation/v1.1), annotation version 1.1. Tomato sequence v2.4 and annotation v2.3 data (Sato et al., 2012) were retrieved from the SGN network (http://www.solgenomics.net). Arabidopsis InterPro domains were retrieved from TAIR10 (Lamesch et al., 2012). Domains of transcript isoforms were merged into one set per gene.

2.3 Function annotation data

Annotations from the Gene Ontology project, version 1.1418 (Gene Ontology Consortium, 2000), and from Gramene (Youens-Clark et al., 2011), were used as input for training and cross-validation. Annotations from Oryzabase version 4 (Kurata and Yamazaki, 2006) were used as an independent validation set. Only genes for which no annotation was available in the data from the Gene Ontology project were used for validation. In all cases, only Biological Process (BP) terms with evidence codes IDA (inferred from direct assay), IGI (genetic interaction) and IMP (mutant phenotype) were used.

2.4 Network data

Co-expression networks based on microarray data for Arabidopsis, rice, *Glycine max*, *Medicago truncatula* and poplar were obtained from PlaNet (Mutwil et al., 2011). For tomato, a recently published microarray-based co-expression network (Fukushima et al., 2012) was used. The probe ids of the tomato co-expression network were obtained from Affymetrix (http://www.affymetrix.com) and mapped with BLAST v2.2.26 (Altschul et al., 1990) to the tomato protein sequences. Further network data for Arabidopsis and rice was obtained from FunctionalNet (http://www.functionalnet.org/) (Lee et al., 2010) and STRING (Szklarczyk et al., 2011). Arabidopsis yeast-two-hybrid data were acquired from literature

(Arabidopsis Interactome Mapping Consortium, 2011). The rice-Arabidopsis interspecies network was generated by using BLAST (cut-off on E-value of 1e-4). BMRF requires all proteins to be part of the input network. Thus, proteins not contained in the input network were removed. In all cases, the longest isoform of alternatively spliced variants was used.

2.5 Validation setup

Performance assessment was performed with rice. HMMER version 3 (http://hmmer.org/) search against Pfam (Finn et al., 2010) and BLAST (Altschul et al., 1990) alignment against UniProt (The UniProt Consortium, 2012) were used to generate the input for Argot2 (Falda et al., 2012). In the context of the validation setup, all rice proteins were removed from the UniProt database to avoid Argot2 using information from those proteins.

For comparison, sequence similarity-based annotation was carried out with Blast2GO (Conesa et al., 2005). Rice protein sequences were queried against the non-redundant part of GenBank (NR) (Benson et al., 2013), using an E-value cut-off of 1e-4. In the context of the validation setup, hits to monocot proteins in NR were removed from the BLAST results before supplying them to Blast2GO.

Prediction runs of different method and network combinations were assessed with 100 cross-validation runs. In each run, randomly, a subset (n=200) of proteins was chosen and the annotation was removed (masked). For every run, predicted functions were compared with the masked ones. Only biological process terms with at least three masked proteins were used in the performance assessment in order to allow for sufficient statistics. In the performance assessment, negative cases consisted of gene-BP associations which were not annotated as such in the experimental data.

Performance was assessed by the area under the receiver operating characteristic curve (AUC) and the F-score. The AUC is the area under the curve of 1-specificity vs. sensitivity, and is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Hanley and McNeil, 1982). Specificity is the fraction of proteins experimentally known not to perform a given function which are indeed not predicted to do so, whereas sensitivity (or recall) is the fraction of proteins experimentally known to perform a given function which are indeed predicted to do so. F-score is based on the precision-recall (precision vs. sensitivity) curve. Precision is the fraction of proteins predicted to perform a given function which are indeed experimentally known to do so. The F-score is equal to the harmonic mean of precision and recall, and the maximum value of the F-score (F_{max}-score) was used for each biological process.

To obtain a finite set of predictions, functions of a protein were assigned by using an F-score-based cut-off. The F-score was calculated per GO term and its maximum (F_{max}-score), calculated with Arabidopsis data as previously described (Kourmpetis et al., 2011), was used to set a cut-off on the posterior probability. The threshold obtained with Arabidopsis data was used in the other species, because

in those species, too few annotations are available to obtain a species-specific threshold. All performance measures were calculated with the R-package ROCR (Sing et al., 2005) and custom R-scripts.

2.6 Application setup

Function annotations predicted for barrel clover, poplar, rice, soybean and tomato were compared with existing predictions in terms of coverage of proteins and number of predicted functions per protein. Barrel clover, poplar and rice biological process predictions were obtained from the official genome annotations version Mt3.5v5 (Young et al., 2011), v1.1 (Tuskan et al., 2006) and v7.0 (Ouyang et al., 2007), respectively. Soybean annotation was obtained from Phytozome (Goodstein et al., 2012). Tomato function annotation data was extracted from the ITAG annotation v2.3 (Sato et al., 2012).

To determine the total number of proteins and total number of GO terms for which annotations were obtained, the annotation of each protein was expanded by including the parent GO terms of all assigned GO terms. For the calculation of the number of annotations per protein, only the leaf-terms of the Gene Ontology were included.

2.7 Evolutionary and functional distance calculation

Groups of orthologs were predicted with OrthoMCL (Li et al., 2003). To calculate functional divergence, BMRF posterior probabilities for each protein were interpreted as vector. The Euclidean distance for each combination of proteins within a group of orthologs was calculated. The mean of distances within a group (inner group distance) was used to rank groups of orthologs. For the PAP26 example, only groups with existing experimental annotation in Arabidopsis were taken in to account. The PAP26 tree was estimated with RAxML version 7.2.8-ALPHA (Stamatakis, 2006) using the PROTGAMMAJTTF substitution model and 1000 bootstraps. Expression data for PAP26 was obtained from the AtGenExpress developmental set (Schmid et al., 2005); publicly available RNA-seq datasets from tomato (*Solanum lycopersicum* cv. Heinz 1706; data SRA049915) were retrieved from the SRA database (http://www.ncbi.nlm.nih.gov/sra). Reads were mapped with GSNAP (Wu and Nacu, 2010) against the tomato reference genome (v. 2.40, Sato et al., 2012) and the expression was determined with cufflinks (Trapnell et al., 2010) with default parameters. Soybean expression data was obtained from SoyBase (Severin et al., 2010). Rice expression data was obtained from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu). All expression experiment data were z-score normalized and percentile ranked to facilitate comparison. Replicates were merged by averaging over the expression for each gene.

3 Results

3.1 Method development and evaluation

We previously developed the protein function prediction method BMRF and used it to annotate protein function in *Arabidopsis thaliana* (Kourmpetis et al., 2011). This method relies, besides on network data, on existing function annotation as input. For Arabidopsis, we demonstrated that the amount of available annotation (training) data was sufficient to achieve a good prediction performance (Kourmpetis et al., 2011). However, for crop species, much less annotation data is available as input. To increase the overall function prediction performance for plants with sparse experimental data, we explored combining BMRF with the sequence-based method Argot2.

Argot2 and BMRF were tested separately (standalone setting) or in two combinations (fig. 4.1). Performance assessment focused on rice, the crop with the largest amount of annotation data available: 415 proteins with experimental evidence for a biological process. The rice network used as input for BMRF was obtained from a combination of microarray-based co-expression data, data from STRING (Szklarczyk et al., 2011) and FunctionalNet (Lee et al., 2010) (table 4.S1). Of the 415 proteins with experimental evidence, 394 were present in the network, and were used for validation of predicted functions. Function prediction performance was assessed on the basis of cross-validation, leaving out randomly selected proteins with known function and comparing the predictions with those data. The area under the receiver operator characteristic curve (AUC) was used to compare the performance of the predictions that come as ordered lists of predicted proteins per biological process. In the standalone setting (fig. 4.1A,B) with rice sequence and network data, BMRF and Argot2 both have a low performance, with AUC (average ± standard deviation) of 0.6±0.12 and 0.67±0.11, respectively (tables 4.1 and 4.S2). These values are considerably lower than the AUC previously obtained with BMRF for Arabidopsis (0.75) (Kourmpetis et al., 2011) due to the small amount of training data (annotated gene functions) that is available for rice. Assuming information from Arabidopsis would improve the performance of rice protein function predictions in BMRF, we connected proteins in an available Arabidopsis network (table 4.S1) to proteins in the rice network based on sequence similarity using BLAST. With this rice-Arabidopsis interspecies network in addition to the networks of both species separately (fig. 4.1C), BMRF performed slightly better than Argot2 (AUC 0.70±0.12). The precise value of the BLAST E-value cut-off used to create the interspecies network did not influence the performance of BMRF (data not shown). Both methods use complimentary information about biological processes (network input for BMRF, sequence input for Argot2). Therefore, we tested combining the two. Argot2 and BMRF can be combined in multiple ways. We used a simple rank-based approach to predict biological processes by ordering Argot2 and BMRF results separately and then combining their ranks to produce a final rank (fig. 4.1D). This integration was performed for each

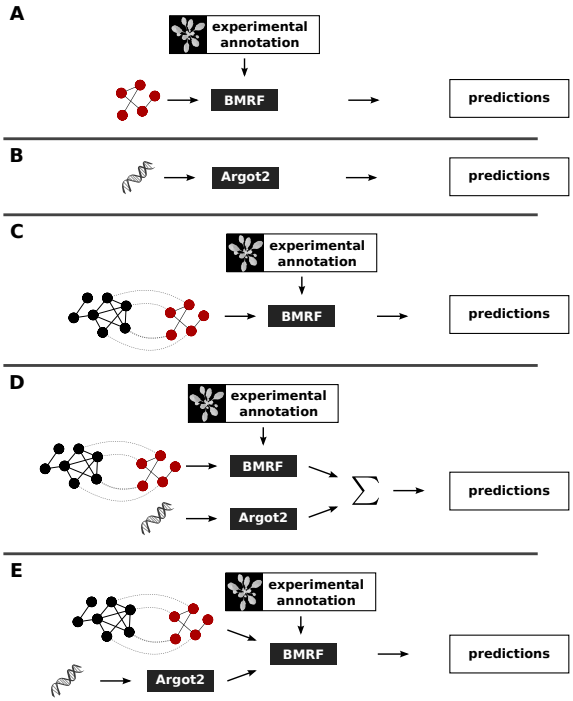


Figure 4.1: Strategies for predicting protein function. BMRF (A,C) and Argot2 (B) were used in a standalone setting or in two different combinations (D,E). Combining BMRF and Argot2 was done by combining the results of each of the two methods (D), and by using Argot2 predictions as input for BMRF (E). The rice network is indicated in red, the Arabidopsis network in black and interspecies connections in grey dashed lines. Sequence-based input is indicated by a DNA-helix symbol.

biological process separately by sorting the proteins based on their score for that process and using the sum of the ranks induced by this ordering for BMRF and for Argot2. This integration of Argot2 and BMRF did not improve results compared to standalone BMRF (table 4.1). Performance was markedly improved, however, by generating initial predictions with Argot2 and supplying these to BMRF as training data (seed data; fig. 4.1E). In this integration method, the initial labeling of proteins in the network (i.e. the seed data for BMRF), was based on the Argot2 predictions. Argot2 uses an algorithm-specific score to rank its results and requires a threshold for such a score. To assess the influence of different thresholds on the performance of BMRF, BMRF was seeded with 5 different output sets of Argot2 (table 4.S3). The best performance was achieved with the default threshold of 5.

The results above indicate that our integrated method performed markedly better than each of the two methods separately. As additional assessment of performance, we predicted annotations with the often-used method Blast2GO (Conesa et al., 2005). The resulting AUC of Blast2GO was 0.72±0.13, and the

Table 4.1: Prediction performance for rice protein function of various combinations of methods and input datasets.

	Network	Method[a]	AUC[b]
(A)	Rice only	BMRF	0.60 (0.12)
(B)	Rice only	Argot2	0.67 (0.11)
(C)	Arabidopsis & rice combined	BMRF	0.70 (0.12)
(D)	Arabidopsis & rice combined	Blast2GO	0.72 (0.13)
(E)	Arabidopsis & rice combined	Argot2 + BMRF	0.71 (0.12)
(F)	Arabidopsis & rice combined	Argot2 → BMRF	0.83 (0.15)

[a] Methods analyzed were BMRF, Argot2, Blast2GO, Argot2 + BMRF (rank sum) and Argot2 → BMRF (seeding). Rice network was used separately (rice only), or it was connected to an Arabidopsis network based on sequence similarity (combined).
[b] Area under the curve; mean (standard deviation).

AUC of the combined Argot2-BMRF predictions was 0.83±0.15 which is significantly ($p < 10^{-15}$; Mann–Whitney U) better than Blast2GO (fig. 4.2A). The small number of experimentally verified annotations (true positives) and high number of unannotated proteins (true negatives) could introduce a skew in the cross-validation sets, leading to a bias in the AUC performance assessment (Davis and Goadrich, 2006). The F-score (harmonic mean of precision and recall) does not suffer from this skew and the final prediction performance was therefore also assessed with the maximum F-score (F_{max}-score). In agreement with the AUC evaluation, the F_{max}-scores of Argot2-seeded BMRF (0.56±0.24) were significantly better ($p < 10^{-15}$; Mann–Whitney U) than Blast2GO (0.51±0.23). Visual inspection of a histogram of AUC values and of F_{max}-score values for different BP terms in different cross-validation runs confirms the performance difference between the combined Argot2-BMRF predictions and Blast2GO (fig. 4.2B,C). To obtain independent validation in addition to the cross-validation performed above, the Argot2-seeded BMRF predictions were compared to annotations available in the Oryzabase database (Kurata and Yamazaki, 2006), which were not present in our input data (71 proteins). The AUC of 0.88±0.13 we obtained was similar to the AUC obtained in the cross-validation, confirming the performance assessment. Overall, the performance evaluation demonstrates that Argot2-seeded BMRF is an effective way to predict BP protein function in sparsely annotated plant genomes.

3.2 Application to crop species

Argot2-seeded BMRF using PlaNet (Mutwil et al., 2011) co-expression networks as input (table 4.S4) was applied to predict BP protein functions in a selection of model and crop plants comprising *Oryza sativa* (rice), *Medicago truncatula* (barrel clover), *Glycine max* (soybean), *Populus trichocarpa* (poplar) and *Solanum lycopersicum* (tomato). The posterior probability of a protein associated with a certain GO term was estimated for all GO terms and all proteins in the network. In or-

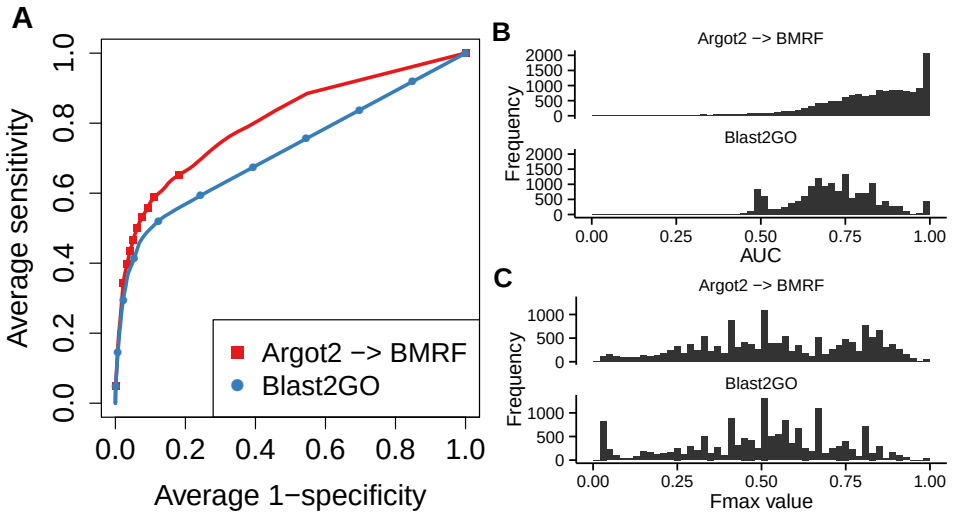


Figure 4.2: Performance assessment of function prediction on rice proteins. (A) Receiver operator characteristic curve showing 1-specificity vs. sensitivity of the predictions of Argot2-seeded BMRF and Blast2GO. Specificity and sensitivity were averaged over all cross-validation runs. Dots indicate evenly spaced intervals of the underlying prediction score, line represents complete curve. Performance is summarized as AUC which is the area under these curves. (B) Histogram of AUC values per GO term of every cross-validation run calculated for Argot2-seeded BMRF and Blast2GO. (C) Histogram of F_{max} values per GO term of every cross-validation run calculated for Argot2 seeding BMRF and Blast2GO.

der to answer a question such as »does protein X perform biological process Y«, a finite set of predictions is needed. To obtain such finite set, an F-score-based cut-off was applied to the posterior probability. As Arabidopsis has the highest coverage of experimental data, this cut-off was adjusted per GO term by comparing Arabidopsis predictions with available experimental data, as previously described (Kourmpetis et al., 2011): for each GO term, a threshold on the posterior probability was defined that results in the maximum F-score for that GO term. All predictions are available online (http://www.ab.wur.nl/bmrf/). The online resource can be queried for predictions of proteins or for GO terms of interest, and the results can be downloaded in bulk. Queries can be based on protein identifiers, biological process GO identifiers, or text descriptors of biological processes (fig. 4.3). The fraction of proteins out of the complete proteome annotated with at least one biological process (annotation coverage) varies considerably between the species: rice shows the highest annotation coverage (99%), followed by poplar (77%). Soybean (43%) and barrel clover (39%) show lower coverage. Tomato has the lowest coverage (12%). Such differences in annotation coverage can have at least two reasons. First, although for every biological process every protein in the input network will have an associated posterior probability, these probabilities can

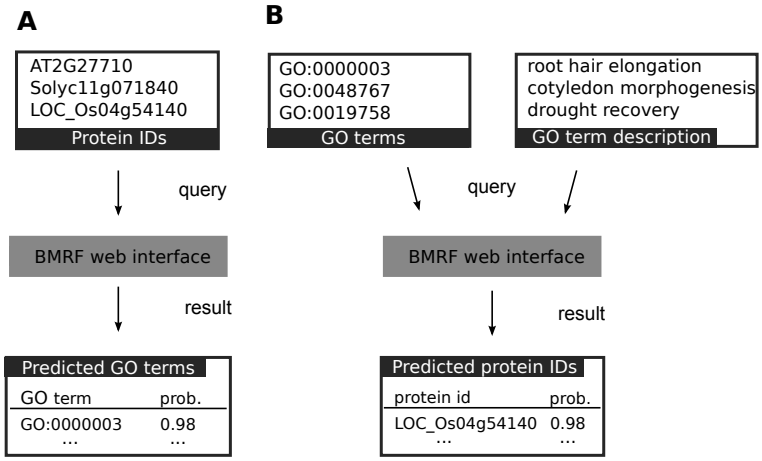


Figure 4.3: Use case scenarios for the web interface. Argot2-seeded BMRF results can be queried in two ways. (A) Protein identifiers as query input. The result consists of predicted GO terms for each protein. (B) GO terms (or GO term descriptions) as query input. The result consists of predicted protein identifiers for the relevant GO term(s) and associated posterior probabilities (prob).

be below the F-score-based cut-off. This means that not necessarily every protein in the input network will be annotated. In addition, because BMRF only predicts functions for proteins in the input network, the maximum possible annotation coverage is limited by the number of proteins in the respective network. This limit is reflected by the tomato annotation coverage, as the tomato network is the smallest with 4355 proteins. With exception of soybean, the annotation coverage correlates with the number of proteins in the respective network (table 4.S4).

To investigate differences between available gene function annotation data and Argot2-seeded BMRF, we compared the results with existing protein function predictions from the reference genomes of barrel clover (Young et al., 2011), poplar (Tuskan et al., 2006), tomato (Sato et al., 2012), rice (Ouyang et al., 2007) and soybean (Schmutz et al., 2010). Except for tomato, the existing annotations have a much lower coverage than the above mentioned coverage obtained by Argot2-seeded BMRF (table 4.S5). The increase of percentage of number of proteins with at least one biological process predicted by our approach varied per species. The percentage increase ranged from ~60% for rice (24,160 in existing annotation vs. 38,998 in our annotation) to over 100% for poplar (13,682 vs. 32,119).

To complement the above presented results on coverage, which focused on the question how many proteins obtain at least one annotation, we also compared the number of predicted functions per protein. The average number of GO terms per protein in the available experimental annotation data for Arabidopsis is 4.4. As additional experimental evidence is supposed to accumulate, this number should be regarded as a lower bound of the average real number of GO terms a protein should

be annotated with. Existing sets of predicted annotations for the plant species included here are considerably below this bound, whereas our set of predictions is relatively close to this bound (table 4.S5). Note that in this assessment, only the most granular level of the Gene Ontology is taken into account (i.e. only leaf-node terms are considered, and not more general parent terms). For those proteins for which existing annotations are available, these annotations are to a large extent a subset of what we predict (~80% of the existing annotations is also predicted by Argot2-seeded BMRF; data not shown). The higher annotation coverage in combination with the good prediction performance demonstrates the appreciable added value of the Argot2-seeded BMRF strategy for obtaining gene function annotations.

3.3 Predicted protein functions: showcases

To illustrate the potential of the functions predicted, we screened all predictions for newly annotated biological processes that are considered particularly relevant for the individual species (table 4.S6). Biological processes considered comprise: seed development for rice and soybean; nitrogen fixation for barrel clover; fruit development for tomato; and lignin related processes for poplar. Inspection of the selected predictions shows that the functions of proteins tend to become more specific: broadly defined functions are replaced by or augmented with more specific biological processes. For example, the rice protein LOC_Os10g38080, was previously annotated with anatomical structure morphogenesis, and is annotated by Argot2-seeded BMRF with seed (coat) development. LOC_Os10g38080 is a subtilisin homolog, which according to available RNA-seq data is expressed in amongst other reproductive organs and seeds (Ouyang et al., 2007). As additional evidence for the Argot2-seeded BMRF prediction, in Arabidopsis subtilisin and related proteases are involved in seed coat development (Rautengarten et al., 2008). An example for an annotation for a previously completely unannotated protein is LOC_Os05g02520, a cupin domain containing protein, which was annotated by Argot2-seeded BMRF with seed maturation.

3.4 Divergence and conservation of biological processes in ortholog groups

The set of function predictions delivered above allows to compare function annotation between different plants, a task which is much less easily performed with existing annotations that are derived from various methods and that have a much lower coverage than our approach. Such comparison between orthologous genes in different plants allows to assess the limits of orthology-based function prediction, and to analyze gene function evolution.

To characterize ortholog groups with functional predictions that differ from expectations based on sequence similarity, orthologs and paralogs were identified with OrthoMCL (Li et al., 2003), resulting in 25,347 groups (table 4.S7). Group members for which no functions were predicted were removed. To assess the simi-

larity of function predictions within ortholog groups, the mean functional distance within each ortholog group (dubbed »inner group distance«) was calculated (see section 2). In case the predicted biological processes in such a group are different despite high sequence similarity, this would be indicative of evolutionary divergence by, e.g. neo-functionalization. To identify such cases, groups with at least four different organisms (6,073) were ranked by their largest inner group distance and the most divergent groups (n=100) were selected. In those groups, biological processes that were significantly overrepresented (more present than randomly expected) were obtained. A variety of biological processes was found (fig. 4.S1), indicating the widespread occurrence of changes in biological processes proteins are involved in. Most prominent are processes related to cell development, regulation, and response to chemical stimulus. For the latter group, biological processes involved are shown in fig. 4.4A.

Among the top ranking groups (with highest »inner group distance«) involved in those processes, we chose as example a phosphatase with existing experimental annotation in Arabidopsis, PURPLE ACID PHOSPHATASE 26 (PAP26). PAP26 plays a role in the phosphate metabolism (Hurley et al., 2010) and phosphate starvation (Hurley et al., 2010) in Arabidopsis. The majority of the proteins with function predictions in the orthologous group (five out of seven) are indeed predicted by Argot2-seeded BMRF to be involved in phosphate metabolism or the response to phosphate starvation. However, additional function predictions differ. Populus and soybean proteins are predominantly annotated with cell death related terms; Arabidopsis with pollination and pollen germination processes; tomato with DNA repair and rice with microtubule cytoskeleton organization. This diversity in function is not reflected by orthology predictions and phylogenetic relationships of the group members (fig. 4.4B,C). Independent expression data indicates that Arabidopsis PAP26 is expressed in a housekeeping-like manner, but the expression pattern varies between paralogs in other species, e.g. soybean, and to a lesser extent orthologs, e.g. between tomato and soybean (fig. 4.4D). The different expression patterns give credibility to the variation in function predictions of Argot2-seeded BMRF. This indicates that PAP26, although its molecular function presumably is invariant, is involved in various biological processes in various plant species. More generally, the analysis of functional divergence presented here highlights the potential of using our set of predicted gene functions for large scale comparisons between various plant species.

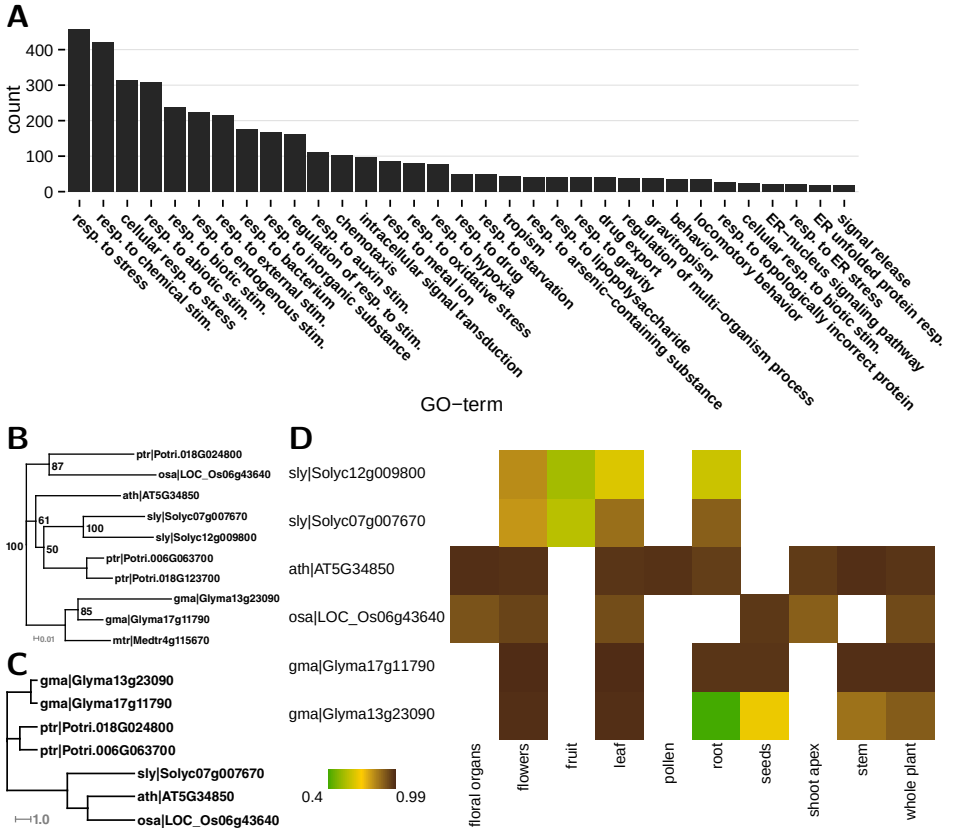


Figure 4.4: Comparison between sequence divergence and functional divergence. (A) Overview of the most frequent GO terms in the top 100 most functionally divergent ortholog groups that are represented by »response to chemical stimulus« (fig. 4.S1). (B-C) Phylogenetic relations of Arabidopsis PURPLE ACID PHOSPHATASE 26 orthologs. Trees contain Arabidopsis (ath), soybean (gma), tomato (sly), Populus (ptr) and rice (osa) PAP26 orthologs. (B) Unrooted phylogenetic tree based on sequence data. The tree was calculated with 1000 bootstraps. Confidence values are indicated at the branches in per cent. (C) Distance tree based on our function predictions. Missing identifiers were not part of the co-expression network and are therefore not part of the functional distance tree. (D) Expression ranking of PURPLE ACID PHOSPHATASE 26 orthologs and paralogs in different tissue clusters. The heatmap color represents a mean percentile rank of normalized expression studies aggregated by averaging to ten tissue clusters (table 4.S8). Missing data is indicated in white. An overview of the aggregated studies is available in table 4.S8.

4 Discussion

Finding associations between proteins and biological processes is a major challenge in non-model plants. Most experimental studies are aimed towards model organisms, hence experiment-based function annotation is sparse in the remainder of sequenced plant genomes. High-throughput experiments to define protein functions are overall less informative than those provided by low-throughput experiments (Schnoes et al., 2013). Moreover, the experimental setup in large-scale approaches might restrict the type of function annotation that can be obtained. An example is the characterization of overexpressed rice genes in Arabidopsis (Sakurai et al., 2011) to infer function. Here, the problem is that the biological process of a protein is often bound to the local environment or a specific condition and a different (plant) environment might change the outcome. Another large scale analysis of gene families in Arabidopsis used prokaryotic gene information to predict function (Gerdes et al., 2011). This semi-manual approach yielded good results for conserved gene families; however, gene families with low conservation were not covered.

Several computational approaches to protein function annotation exist, albeit mostly not targeted to plants, or to model plant species only (Lee et al., 2011). An integrated platform such as Phytozome (Goodstein et al., 2012) provides a consistent set of Gene Ontology annotations for various plant species and hence overcomes the above-mentioned problem that annotations associated with genomes are obtained by various methods. However, Phytozome only provides sequence-based predictions. The recently published MORPH algorithm ranked genes for their membership of Arabidopsis and tomato pathways, based on a set of known genes from the target pathway, a collection of expression profiles, and interaction and metabolic networks (Tzfadia et al., 2012). Approaches such as PlaNet construct networks based on expression data (Mutwil et al., 2011), but such networks do not directly lead to gene function annotation. Similarly, a recently presented text mining approach generated networks in Arabidopsis and not gene function annotations (Blanc and Wolfe, 2004). Here we provide a structured approach to extract gene function information from networks and combine that with sequence-based information.

The combination of sequence- and network-based function prediction obtained by seeding BMRF with Argot2, offers a significant benefit over applying these methods separately. We validated the method in rice and demonstrated greatly improved performance compared to each of the methods separately and compared to Blast2GO. This performance assessment was performed using two complementary indicators, AUC and F-score, which both gave consistent results. Existing annotations provided for the plant genomes to which we applied our method have been obtained by various, mostly sequence-based approaches. A clear description of the methods and input data is often lacking, leading to the risk of error propagation and circular reasoning (du Plessis et al., 2011; Engelhardt et al., 2011). Our approach has the benefit of applying a standard method to the vari-

ous genomes. Moreover, for many proteins which so far were not associated with any biological process, we now provide predictions of biological processes. Nevertheless, the combination of Argot2 and BMRF is indirectly constrained by the experimental data in databases such as UniProt (Dimmer et al., 2012) or Pfam (Finn et al., 2010), and by the proteins covered in available networks. It will however be straightforward to integrate newly available datasets such as additional co-expression networks or novel gene function annotations in the framework presented. An additional limitation of our current approach is that the structure of the Gene Ontology is not taken into account in the prediction process. Most existing computational methods for gene function prediction suffer from this drawback. It is feasible to make a set of GO term predictions consistent with the GO-structure (Kourmpetis et al., 2013) and we plan to apply this method to Argot2-seeded BMRF predictions in the future.

BMRF output consists of a list of probabilities for each gene to be associated with each biological process. This allows to rank proteins in order of their likelihood of association with a biological process of interest. However, it can also be important to have a finite set of predictions. To provide that, we applied a cut-off to the probabilities, based on Arabidopsis, the only species from which enough data was available. It is difficult to assess how valid the application of this cut-off in other plant species is. However, the average number of predictions per protein that we obtain in each of the species based on the cut-off that was applied is close to the observed average for Arabidopsis, giving some credibility to this cut-off. For one species, tomato, the number of predicted BP terms per protein is somewhat higher than the experimentally observed number for Arabidopsis. Hence, Argot2-seeded BMRF possibly suffers from overprediction in this case. This could possibly be caused by the higher density (number of interactions compared to number of proteins) of the tomato network. However, in any case, the probabilities associated with the predictions allow narrowing down the prediction results to the most reliable ones, if so desired.

With the consistent annotation of multiple plant genomes that we performed, the relation between homology and biological process predictions can be analyzed. Ortholog groups with divergent functions indicated cases where conclusions based on sequence similarity might be inappropriate. Such inappropriate conclusions may be more common than generally acknowledged. For example, about half of a collection of Arabidopsis loss-of-function mutants had only low or moderate phenotypic similarity with mutants of putative orthologs in tomato, rice or maize (Lloyd and Meinke, 2012). Large scale evolutionary comparisons between plant species, for example aimed at identifying patterns in retention of duplicated genes (Guo et al., 2013; Jiang et al., 2013) or functional biases in single-copy genes (De Smet et al., 2013), are currently performed based on function annotations obtained using sequence similarity. Such studies will benefit from the gene annotations presented here, which overcome the limitations of purely sequence-based annotation of gene functions.

In the example of the PAP26 homologs, homology captures the molecular function, but at the biological process level there is divergence. Our integrated sequence- and network-based function annotation method allows to predict such divergent biological processes. Differences in expression between the different PAP26 homologs in different species provide additional evidence for our function predictions. More generally, the results on biological process divergence are in line with the concept that evolution acts in particular by »tinkering« with genes, coopting available components of a genome for new processes.

The combination of sequence-based and network-based predictions is a huge improvement for sparsely annotated plant genomes. With the advent of RNA-seq (Marguerat and Bähler, 2010) co-expression network-based protein function prediction can become a preferred method. Combined with additional analysis, such as genome-wide association studies (GWAS), potential candidate genes for traits-of-interest could be identified more reliably. Such candidate genes will be of great help in applications related to plant breeding. The ability to associate unannotated proteins to particular biological processes will spark experimental work and be essential for the advancement of understanding of gene function in plant genome evolution.

Acknowledgements

We thank Dr. Yiannis Kourmpetis for helpful discussions. This work was supported by the FP7 »Infrastructures« project TransPLANT (award 283496) and by the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

5 Supporting Information

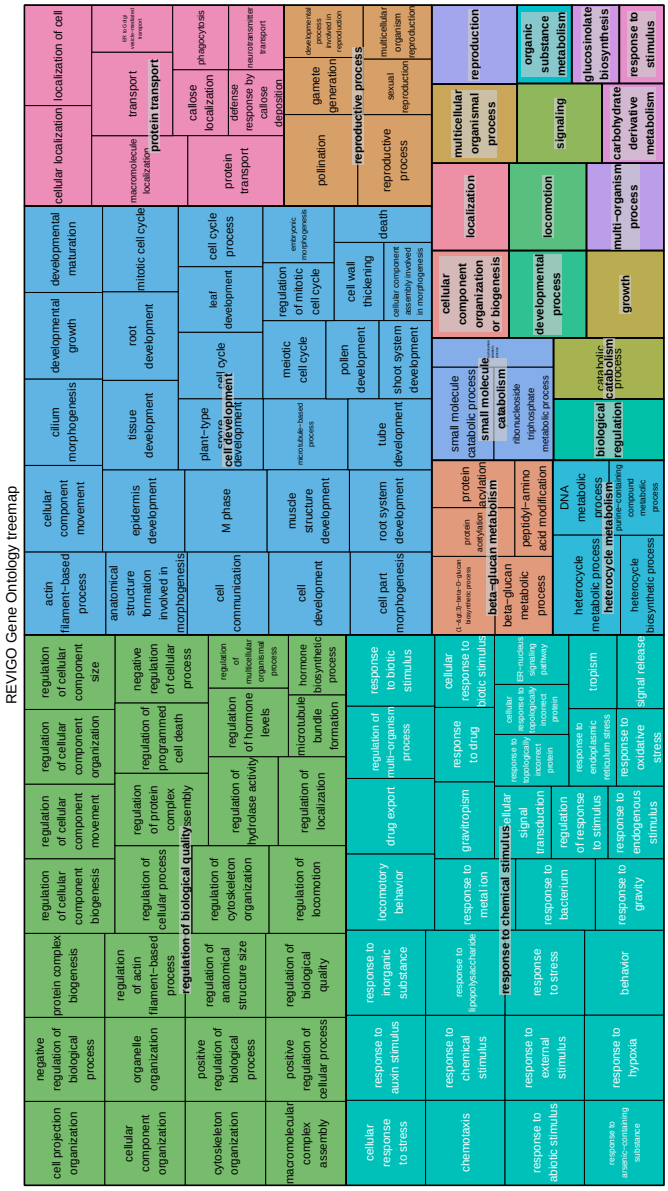


Figure 4.S1: Treemap of enriched GO-terms in the top 100 most divergent orthologous groups. Treemap was generated using REVIGO (Supek et al., 2011).

Table 4.S1: Overview of networks used in the validation setup. Networks used in BMRF were obtained for Arabidopsis (Ath) and rice (Osa) from PlaNet (Mutwil et al., 2011), STRING (Szklarczyk et al., 2011), FunctionalNet (RiceNet & AraNet) (Lee et al., 2010) and the Arabidopsis interactome (Arabidopsis Interactome Mapping Consortium, 2011).

Network	Total		Annotation		
	proteins	*edges*	*proteins*	*GO terms*	*annot.*
Arabidopsis					
STRING	10397	125877	3518	3119	99524
Interactome	4866	11374	1737	2436	53513
Rice					
STRING	6061	45015	206	558	4619
RiceNet	18377	588221	323	735	7406
Planet & STRING & RiceNet	38998	2163770	394	757	9114
Interspecies (combined)					
STRING (Osa, Ath), RiceNet, AraNet, Planet (Osa, Ath), Ath interactome; subnetworks connected with BLAST	60637	2970769	4813	3364	127601

Table 4.S2: Comparison of dependence of BMRF performance on rice input networks. BMRF was tested on different rice input networks with 100-fold cross-validation. The performance of Argot2 was compared on the same cross-validation sets as BMRF. Rice networks used in BMRF were obtained from PlaNet (Mutwil et al., 2011), STRING (Szklarczyk et al., 2011) and FunctionalNet (RiceNet) (Lee et al., 2010). For other plant species, in general only co-expression networks are available. Importantly, for rice, the performance did not depend much on the input network used.

Network	Algorithm	AUC[a]
RiceNet	BMRF	0.64 (0.11)
	Argot2	0.64 (0.09)
STRING	BMRF	0.65 (0.14)
	Argot2	0.65 (0.13)
Planet, STRING and RiceNet	BMRF	0.60 (0.12)
	Argot2	0.67 (0.11)

[a]Area under the curve; mean (standard deviation).

Table 4.S3: Prediction performance of Argot2-seeded BMRF at different Argot2 thresholds.

Threshold	AUC[a]
5	0.83 (0.15)
200	0.81 (0.16)
500	0.81 (0.17)
1000	0.79 (0.17)
2000	0.77 (0.17)

[a]Area under the curve; mean (standard deviation).

Table 4.S4: Size of input networks for the application setup.

Network	proteins in the proteome	proteins in network	edges
Rice[a]	39,049	38,998	2,163,770
Medicago[a]	44,135	17,464	425,587
Soybean[a]	54,175	25,113	951,419
Poplar[a]	41,335	32,119	566,012
Tomato[b]	34,727	4,355	910,171

[a]Planet co-expression network (Mutwil et al., 2011)
[b]Tomato microarray co-expression network (Fukushima et al., 2012)

Table 4.S5: Prediction coverage compared to existing predictions. Existing sets of function annotations were compared to our function annotations. Columns spanned by »proteins« and »GO-terms« indicate number of proteins for which biological process annotations are provided, and number of GO-terms for which proteins are predicted, respectively; Columns spanned by »GO-terms per protein« indicate how many GO-terms are on average predicted per protein. For the latter, only leaf GO-terms are taken into account.

ORGANISM	PROTEINS		GO-TERMS		GO-TERMS PER PROTEIN	
	Argot2/BMRF set	*existing set*	*Argot2/BMRF set*	*existing set*	*Argot2/BMRF set*	*existing set*
Medicago	17,464	9,178	1,475	874	3.4	1.1
Rice	38,998	24,160	1,892	86	3.5	1.2
Poplar	32,119	13,682	1,949	1,254	3.3	1.2
Tomato	4,355	8,457	923	955	4.8	1.2
Soybean	23,620	14,511	1,663	1,098	3.8	1.2

Table 4.S6: Top five predictions for the respective plant relevant biological processes. Cases discussed in the main text are underlined.

organism	GO-term	protein ID	prob.
Glycine max	seed development (GO:0048316)	Glyma05g30000	0.999
		Glyma09g02750	0.999
		Glyma15g13640	0.999
		Glyma03g01860	0.998
		Glyma08g21610	0.998
Medicago truncatula	nodulation (GO:0009877)	Medtr3g040210	0.988
		Medtr3g040300	0.987
		Medtr3g040320	0.986
		Medtr4g081350	0.986
		Medtr4g039010	0.984
Oryza sativa	seed coat development (GO:0010214)	LOC_Os01g64850	1
		LOC_Os10g38080	1
		LOC_Os04g47160	1
		LOC_Os02g53850	1
		LOC_Os02g53970	1
	seed maturation (GO:0010431)	LOC_Os05g02520	0.999
		LOC_Os02g15169	0.998
		LOC_Os02g16820	0.998
		LOC_Os01g74480	0.996
		LOC_Os08g03410	0.995
	seed development (GO:0048316)	LOC_Os01g64850	1
		LOC_Os10g38080	1
		LOC_Os04g47160	1
		LOC_Os02g53850	1
		LOC_Os02g53970	1
Populus trichocarpa	cytokinin metabolic process (GO:0009690)	Potri.011G159600	0.967
		Potri.011G137800	0.956
		Potri.008G084800	0.86
		Potri.001G462200	0.852
		Potri.002G178300	0.851
	cytokinin biosynthetic process (GO:0009691)	Potri.011G137800	0.956
		Potri.012G142000	0.767
		Potri.011G001300	0.711
	lignin metabolic process (GO:0009808)	Potri.008G064000	1
		Potri.007G023300	1
		Potri.005G247700	1
		Potri.001G219300	1
		Potri.009G034500	1

Continued on next page

Table 4.S6 – Continued

organism	GO-term	protein ID	prob.
Populus trichocarpa	lignin biosynthetic process (GO:0009809)	Potri.016G112100	1
		Potri.006G087500	1
		Potri.008G064000	1
		Potri.009G034500	1
		Potri.005G200700	1
Solanum lycopersicum	fruit development (GO:0010154)	Solyc01g073860	1
		Solyc01g073820	1
		Solyc01g073880	1
		Solyc04g007000	0.989
		Solyc11g073210	0.987

Table 4.S7: Overview of OrthoMCL groups. Organisms present in a group are indicated by a dot. The number of groups possessing the marked organisms are given either directly from the OrthoMCL analysis or after filtering for proteins present in the interspecies network. Note that if due to the filtering no proteins from a given species remain, the total number of species present in the group will decrease. Groups with four or more organisms after filtering were used in further analysis.

Organisms						distinct organisms	OrthoMCL groups	groups in net
Ath[a]	*Gma*[b]	*Mtr*[c]	*Osa*[d]	*Ptr*[e]	*Sly*[f]			
●	●	●	●	●	●	6	7055	336
●	●	●	●	●		5	247	497
	●	●	●	●	●	5	313	35
●	●	●	●		●	5	84	140
●	●		●	●	●	5	2193	1253
●		●	●	●	●	5	24	24
●	●	●		●	●	5	1156	32
●	●	●	●			4	20	302
●	●			●	●	4	477	91
●			●	●	●	4	82	127
	●	●		●	●	4	384	7
	●		●	●	●	4	133	162
●		●	●	●		4	2	79
●	●		●	●		4	96	2046
	●	●	●		●	4	30	33
	●	●	●	●		4	78	86
		●	●	●	●	4	9	9
●	●	●		●		4	275	85
●	●		●		●	4	31	692
●		●	●		●	4	6	18
●	●	●			●	4	72	16
●		●		●	●	4	18	3
●	●	●				3	71	60
●	●		●			3	22	1577
●	●			●		3	133	446
	●		●		●	3	16	149
			●	●	●	3	37	28
	●			●	●	3	179	47
	●	●			●	3	105	11
●		●	●			3	1	99
	●	●		●		3	282	45

Continued on next page

Table 4.S7 – Continued

Organisms						distinct organisms	OrthoMCL groups	groups in net
Ath[a]	*Gma*[b]	*Mtr*[c]	*Osa*[d]	*Ptr*[e]	*Sly*[f]			
		●	●	●		3	3	42
	●	●	●			3	45	66
●			●		●	3	13	108
	●		●	●		3	38	431
●				●	●	3	130	33
		●		●	●	3	10	1
●			●	●		3	39	478
		●	●		●	3	6	5
●		●		●		3	7	46
●		●			●	3	7	10
●	●				●	3	53	94
	●		●			2	47	501
	●			●		2	242	358
●			●			2	38	566
	●	●				2	976	169
			●		●	2	67	37
			●	●		2	49	274
●					●	2	45	34
		●	●			2	14	64
				●	●	2	186	29
●	●					2	58	379
●				●		2	143	336
		●		●		2	27	61
	●				●	2	92	38
		●			●	2	49	2
●		●				2	13	51
		●				1	2329	1151
					●	1	1135	109
			●			1	2080	2431
●						1	841	986
	●					1	1706	1453
				●		1	1228	1128
						Total	25347	20006

[a] *Arabidopsis thaliana*
[b] *Glycine max*
[c] *Medicago truncatula*
[d] *Oryza sativa*
[e] *Populus trichocarpa*
[f] *Solanum lycopersicum*

Table 4.S8: Overview of the tissue types used in the expression ranking of PAP26 homologs.

ORGANISM	TISSUE CLUSTER	LIBRARY	REFERENCE
Arabidopsis thaliana	floral organs	ATGE_34	Schmid et al., 2005
	floral organs	ATGE_35	
	floral organs	ATGE_36	
	floral organs	ATGE_37	
	floral organs	ATGE_40	
	floral organs	ATGE_41	
	floral organs	ATGE_42	
	floral organs	ATGE_43	
	flowers	ATGE_31	
	flowers	ATGE_32	
	flowers	ATGE_33	
	flowers	ATGE_39	
	leaf	ATGE_1	
	leaf	ATGE_10	
	leaf	ATGE_12	
	leaf	ATGE_13	
	leaf	ATGE_14	
	leaf	ATGE_15	
	leaf	ATGE_16	
	leaf	ATGE_17	
	leaf	ATGE_19	
	leaf	ATGE_20	
	leaf	ATGE_21	
	leaf	ATGE_25	
	leaf	ATGE_26	
	leaf	ATGE_5	
	pollen	ATGE_45	
	pollen	ATGE_73	
	root	ATGE_3	
	root	ATGE_9	
	shoot apex	ATGE_29	
	shoot apex	ATGE_4	
	shoot apex	ATGE_6	
	shoot apex	ATGE_8	
	stem	ATGE_2	
	stem	ATGE_27	
	stem	ATGE_28	

Continued on next page

Table 4.S8 – Continued

ORGANISM	TISSUE CLUSTER	LIBRARY	REFERENCE
Arabidopsis thaliana	whole plant	ATGE_22	
	whole plant	ATGE_23	
	whole plant	ATGE_24	
	whole plant	ATGE_7	
Glycine max	flowers	Flower	Severin et al., 2010
	leaf	young_leaf	
	root	Root	
	seeds	pod.shell.10DAF	
	seeds	pod.shell.14DAF	
	seeds	seed.10DAF	
	seeds	seed.14DAF	
	seeds	seed.21DAF	
	seeds	seed.25DAF	
	seeds	seed.28DAF	
	seeds	seed.35DAF	
	seeds	seed.42DAF	
	stem	Nodule	
	whole plant	one.cm.pod	
Oryza sativa	floral organs	OSN_AD	Ouyang et al., 2007
	floral organs	OSN_AE	
	flowers	OSN_AB	
	flowers	OSN_AC	
	leaf	OSN_AA/ OSN_CA[a]	
	seeds	OSN_AF	
	seeds	OSN_AG	
	seeds	OSN_AH/ OSN_BH[a]	
	seeds	OSN_AK	
	shoot apex	SRR042529	
	whole plant	SRX016110	

Continued on next page

Table 4.S8 – Continued

ORGANISM	TISSUE CLUSTER	LIBRARY	REFERENCE
Solanum lycopersicum	flowers	SRR404314/ SRR404313[a]	Sato et al., 2012
	flowers	SRR404316/ SRR404315[a]	
	fruit	SRR404318/ SRR404317[a]	
	fruit	SRR404320/ SRR404319[a]	
	fruit	SRR404322/ SRR404321[a]	
	fruit	SRR404325/ SRR404324[a]	
	fruit	SRR404327/ SRR404326[a]	
	fruit	SRR404329/ SRR404328[a]	
	fruit	SRR404333/ SRR404331[a]	
	fruit	SRR404336/ SRR404334[a]	
	fruit	SRR404339/ SRR404338[a]	
	leaf	SRR404310/ SRR404309[a]	
	leaf	SRR412748/ SRR412747[a]	
	root	SRR404312/ SRR404311[a]	

[a]Biological replicate

Chapter 5

Less is more: pruning nodes from a biological network can improve prediction of protein function

Abstract

Incorporation of biological networks by using algorithms such as Bayesian Markov Random Field (BMRF) is valuable for the prediction of biological processes that proteins are involved in. The topological properties that influence prediction performance in such networks are however largely unknown. Here we evaluate the performance of BMRF upon progressive removal of highly connected hub nodes (pruning). Three different protein-protein interaction networks with data from Arabidopsis, human and yeast were analyzed. All three show that the average prediction performance can improve significantly. Hub nodes apparently hamper prediction. The functional similarity between hub nodes connected to non-hub nodes is smaller than that of non-hub – non-hub connections. The prediction of more specific biological processes is more likely to benefit from node pruning. A major issue in performance of BMRF by pruning is the amount of annotation used in and/or left for the prediction. Because the prediction relies on existing annotation, the optimal number of pruned nodes varies between networks. As a result, the optimal pruning size will have to be determined for each network separately.

Bargsten, J. W., Sanchez-Perez, G. F., Nap, J.-P. and van Dijk, A. D. J., in preparation.

1 Introduction

Prediction of the functions of a protein is a major challenge in current biology. Due to the avalanche of sequenced genomes, the number of protein sequences is increasing exponentially. To date, more than 98% of all functions assigned are predicted without being experimentally verified (du Plessis et al., 2011; Rhee and Mutwil, 2014). Experimental annotation of function is not only lagging far behind, experimental annotations presented in publications are also not machine readable, thus not easily accessible. To make annotations machine-readable, the Gene Ontology Consortium (Gene Ontology Consortium, 2000) created a controlled vocabulary (ontology) of GO-terms to standardize annotation. The vocabulary is divided into three domains, molecular function (MF), biological process (BP) and cellular component (CC). Notably BP terms are challenging to predict, because the sequence of a protein is only a moderate proxy for the transfer of functional annotation on this level (Radivojac et al., 2013).

Additional information about the relationship between proteins (or genes), in the form of protein-protein-interactions (PPIs) or co-expression data, provides a valuable resource for predicting BP terms (Sharan et al., 2007; Ryan et al., 2013). Nearly all function prediction that uses such information is based on the premise that modular structures such as protein complexes, signaling cascades or transcriptional regulatory circuits, represent biological properties, and as a consequence allow the transfer of functional information (Gillis and Pavlidis, 2012; Mitra et al., 2013)

The sum of interactions between biological units (proteins, genes or otherwise) is commonly referred to as »biological network«. A biological unit in such a network is referred to as »node« and a connection between two nodes is referred to as »edge«. Biological networks possess characteristic topological properties (Barabási and Oltvai, 2004). Typically, biological networks can be decomposed into multiple organizational levels. In most cases, three organizational levels can be distinguished. At the lowest level, nodes are aggregated into network motifs, which aggregate into network modules at the second highest level. At the highest level, modules are connected by a small number of highly connected nodes, generally known as »hub« nodes (Vital-Lopez et al., 2012). To capture the role of a node in the network, so-called centrality measures are used. The simplest centrality measure to calculate is the number of connections of a node, also known as node degree (Borgatti and Everett, 2006). Other common centrality measures are betweenness, eccentricity, local cluster coefficient or closeness (Borgatti and Everett, 2006).

The property that a small fraction of nodes (proteins, genes) are hubs that possess a high number of connections, whereas the majority of nodes have only a few, is termed »scale-free-like« topology of the network (Barabási and Bonabeau, 2003; Barabási and Oltvai, 2004). This scale-free-like topology is interwoven with the modular structure of a biological network. Whereas a modular structure of a network is crucial for function prediction, a scale-free-like topology could hamper

function prediction (Gillis and Pavlidis, 2012). Although the biological units represented by hubs may be essential for the proper functioning of a biological system (Jeong et al., 2001; He and Zhang, 2006), from an information-theoretical perspective they contain a relatively low amount of functional information (Gillis and Pavlidis, 2012). The information that such hubs affect most biological processes in a cell is not useful in the context of function prediction. Hubs or nodes with a high value for a centrality measure may therefore impede the prediction of protein function. At present, it is unclear to what extent network topology in general, and the presence of hub nodes in particular, can influence the performance of a prediction algorithm. The impact of hub nodes may also depend on the type of prediction algorithm used. Algorithms for the prediction of protein function(s) can be divided into direct and module-assisted approaches. Direct approaches, such as guilt-by-association (Walker et al., 1999; Oliver, 2000), BMRF (Kourmpetis et al., 2010) or FunctionalFlow (Nabieva et al., 2005), as well as module-assisted approaches, assume that proteins that are closer to one another in the PPI network are more likely to have similar function. In contrast to direct approaches, module-assisted approaches first identify coherent groups of genes and then assign functions to all the genes in each group (Sharan et al., 2007). In both types of procedures, simple algorithms, such as guilt-by-association, depend on the local environment only, but other algorithms, including BMRF, need the complete network to achieve high performance. The impact of removing nodes from the network may therefore vary from algorithm to algorithm. Direct approaches could benefit most, if disturbing nodes are excluded from the input data, because prediction algorithms rely on the edges and nodes without prior clustering and filtering. To assess and possibly improve the performance of a network in protein function prediction, it should be investigated what influences the performance of a prediction algorithm. Some nodes may increase the noise and their removal could increase the performance in function prediction.

We previously developed the network-based algorithm termed Bayesian Markov Random Field (BMRF) for the prediction of BP-terms (Kourmpetis et al., 2010, 2011; chapter 4). BMRF is an example of a direct approach, using interactions in a network to infer function. BMRF assumes that the function of a protein is dependent on the functions of its immediate neighbors (Markov property). This assumption is implemented as a Markov Random Field (MRF), using a Bayesian approach. BMRF requires an initial set of known annotations (seed set). Model parameters and protein functions are estimated iteratively, until convergence. BMRF allows, in contrast to guilt-by-association, the transfer of functional information beyond direct interactions and is able to provide reliable function predictions even for proteins that are only linked with other proteins of unknown function (chapter 4).

We here use BMRF to assess the influence of the presence of hub nodes on the performance of the prediction of protein function by extensive node removal (node pruning) based on multiple pruning strategies and different centrality parameters. It was shown previously that much of the functional information can be

contained in a relatively small subset of a biological network (Gillis and Pavlidis, 2012), but it has yet to be established if node pruning can improve the performance in the prediction of the function of proteins in biological processes. The results show significant boosts in the protein function prediction performance of BMRF by removing hub nodes. Such a pruning approach gives the possibility to remove apparently noisy elements in the network and increase prediction performance without the need for additional (biological) data or experiments.

2 Materials and Methods

2.1 Network data

The biological networks used in this study were obtained and combined from different sources. Biological networks for Arabidopsis, yeast and human were obtained from BioGRID (Chatr-Aryamontri et al., 2013) and STRING (Franceschini et al., 2013). Edges of STRING networks were required to have a confidence score of at least 700, as used previously (Radivojac et al., 2013). Only a small fraction of edges in BioGRID networks had score information, therefore the edges were taken as is and no cutoff was applied. A combined network was created for every species by creating the union of the corresponding BioGRID and STRING network.

Pruned networks were generated from the combined network. Evaluating the performance for all possible removal steps would be computationally prohibitive. We therefore pruned the networks at distinct pruning steps on an exponential scale. As a result, the first pruning steps (only few nodes removed) are densely spaced, whereas later pruning steps (high amount of nodes removed) are much more sparsely spaced. Pruning steps were performed with a step-function (explained in detail in text 5.S6). The step function was selected because of convenience. It results in 100 steps with at most 2848 nodes pruned, the pruning limit of yeast. The total number of steps varies from analysis to analysis, but in all cases pruning was stopped when a minimum of 50,000 annotations or 1,500 proteins per network was reached. This minimum annotation coverage is based on previous experience (chapter 4).

2.2 Experimental annotation data

Experimental annotation was acquired from the Gene Ontology (Gene Ontology Consortium, 2000; retrieved Oct. 2013 from http://www.geneontology.org). The data was filtered for biological process (BP) terms and matched to the corresponding network nodes (proteins). Comparing functional profiles between two genes/proteins is not trivial. To be able to compare functions between different nodes, semantic similarity was used. In the context of this study, we defined a semantic similarity measure as a function that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in meaning between the two (sets of) terms (Pesquita et al., 2009). Semantic sim-

ilarity was calculated using the GOSemSim R-package (Yu et al., 2010) using the Wang measure (Wang et al., 2007) and best-match averaging (Jiang and Conrath, 1997; Pesquita et al., 2009).

2.3 Network centralities

For the calculation of network-related properties, the R-package igraph was used (http://igraph.org). Different network centralities (degree, betweenness, closeness, local cluster coefficient and eccentricity) were calculated for every node. Nodes were ordered (descending) by the value of the corresponding network centrality and stepwise removed, starting with highest ranked. BMRF is not able to retain unconnected nodes. Therefore, »orphan« nodes that, as a result of the pruning lost all edges, were removed too. This is taken into account in the node count used. Each random pruning setup was created by randomizing the order of the degree-based ranking. The node pruning was performed according to the randomized order.

2.4 Domain data

In addition to network and seed annotation, BMRF utilizes protein domain information. The domain information was obtained from the *Saccharomyces* Genome Database (http://www.yeastgenome.org; retrieved Jul. 2013) for yeast, from UniProt (UniProt Consortium, 2014; http://www.uniprot.org; retrieved Oct. 2013) for human and from the Arabidopsis Information Resource (Lamesch et al., 2012; ver. TAIR10 from http://www.arabidopsis.org) for Arabidopsis. Only domain data derived from the InterPro (Jones et al., 2014) and Pfam (Punta et al., 2012) databases were used. Domains of transcript isoforms were merged into one set per gene.

2.5 Validation setup

Each pruning step was cross-validated with 100 cross-validation samples. The cross-validation setup is similar to the setup presented earlier (chapter 4). Prediction runs of pruned networks were assessed for the independently created samples. A random sample (n=200) of proteins was chosen and the annotation was removed (masked). Only BP terms with at least three masked proteins were used in the performance assessment to have sufficient statistics. In the performance assessment, negative cases consisted of gene-BP associations which were not annotated as such in the experimental data (closed world assumption; Dessimoz et al., 2013). Every run, the performance was calculated from the comparison of predicted and masked functions with the experimental annotation data.

2.6 Assessment of performance

Performance was assessed by the area under the receiver operating characteristic (specificity vs. sensitivity) curve (AUC). The AUC is interpreted as the probability that a prediction algorithm will rank a randomly chosen positive instance higher than a randomly chosen negative one (Hanley and McNeil, 1982). Specificity is the fraction of proteins experimentally known not to perform a given function which are not predicted to do so, whereas sensitivity is the fraction of proteins experimentally known to perform a given function which are predicted to do so. The performance is compared between both the pruned and unpruned network and the degree-based and randomly pruned network. In all cases, the difference in prediction performance is expressed as fraction of the maximum AUC, which is 1 for the theoretically perfect prediction algorithm.

This way assessed, the increase in prediction performance could be underestimated, because removal of nodes implies removal of annotation from the network. The performance of BMRF is sensitive to the amount of annotation (training data) (chapter 4). Moreover, the coverage of experimental annotation is different between random and degree-based removal (fig. 5.S1), which would result in bias. To reduce such biases in the assessment of the prediction performance, the functional annotation not present in the maximum pruned state was masked in all pruning steps. This way, all pruning steps have the same annotation coverage.

3 Results

The input networks were created on the basis of public domain protein-protein-interaction (PPI) data for three different organisms, yeast, human and Arabidopsis. Overall characteristics of these networks are given in table 5.1. The node distribution is given in fig. 5.1. The distribution is typical for the scale-free topology of biological networks: most nodes have few connections whereas a small fraction of nodes has many such edges (fig. 5.1). Therefore we removed highly connected nodes (hubs) stepwise to investigate the influence of hub nodes on the prediction performance. It is anticipated that edges between hub nodes and non-hub nodes, or between hub nodes and different network modules, can result in connections that are not related with respect to function (fig. 5.2) and therefore can hamper predictions based on such connections.

Nodes were ranked by their degree (number of connections) and removed stepwise according to their rank. All three networks were pruned till either a minimum of 50.000 annotations or 1.500 proteins, whatever limit was reached first. The maximum number of nodes removed was 5683 for Arabidopsis, 12856 for human and 2848 for yeast (table 5.2). To assess the performance differences, we used a 100-fold cross-validation with 200 proteins per test set. The area under the receiver operator characteristic (AUC) was used as measure of performance. By removing high-degree nodes (hubs) stepwise, BMRF shows a clear performance increase compared to the not pruned network (fig. 5.3). When more nodes are removed,

Table 5.1: Descriptive statistics of the three networks used in this study.

	Arabidopsis	human	yeast
no. of proteins (nodes)	14,846	29,313	6,396
no. of connections (edges)	142,413	603,605	293,578
no. of experimental annotations	12,168	49,011	11,378
no. of proteins with experimental annotations	4,572	9,228	4,282
no. of experimental annotations per protein	2.66	5.31	2.66
correlation of degree difference vs. semantic similarity ($p < 1e\text{-}16$)	-0.25	-0.11	-0.09
annotation coverage[a] (%)	31	31	67

[a]no. of experimentally annotated proteins/no. of proteins in network

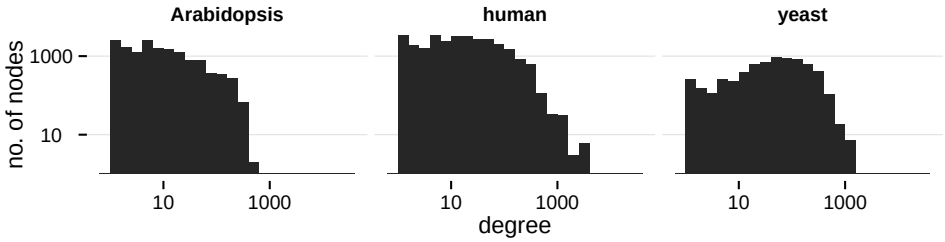


Figure 5.1: Distribution of nodes in the networks of Arabidopsis, human and yeast. The histogram (note that both axes are on logarithmic scale) shows the scale-free-like topology characteristic for this type of biological networks.

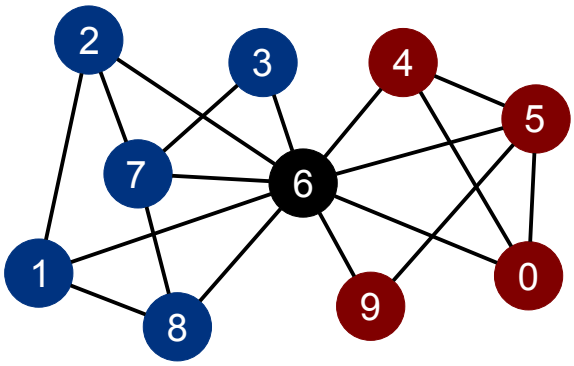


Figure 5.2: Simplified example of a biological network. Functionally different (blue vs. red) network modules are connected by a hub node (node 6). The hub node connects two modules that represent different functions (»bridging«). Each node corresponds to a gene/protein, interactions are indicated by edges. In function prediction, the interactions are used as proxy for functional similarity. Function prediction algorithms could transfer functions via the hub node between these two modules, possibly introducing incorrect function predictions. Incorrect functions could also be transferred from the hub node to the modules directly.

the performance for Arabidopsis and human reaches an optimum. The optimum occurs for human at 5683 pruned nodes (19% of all nodes) and for Arabidopsis at 2848 pruned nodes (19% of all nodes). In contrast, the performance curve for yeast continues to rise till the final pruning step at 2848 pruned nodes (45% of all nodes). Further pruning in yeast would violate the minimal requirements for proper cross-validation. It is therefore not possible to determine if additional pruning would show an optimum for the performance. At the optimum, the improvement of the performance (AUC difference of the median) is 0.013 for the Arabidopsis data, 0.021 for the human data and 0.212 for the yeast data (table 5.S1 and fig. 5.4A). The AUC difference data show a high variation for certain BP-terms (fig. 5.4A), reflecting a relatively high uncertainty in the prediction performance for an individual GO-term.

Unexpectedly, despite correcting for the coverage of experimental annotation (see section 2), random pruning of nodes still increased the performance of BMRF compared to the unpruned network (fig. 5.3). For Arabidopsis and human the performance curve continued to rise till the final pruning step, whereas for yeast, random pruning reaches its optimum at the second last step (2675 nodes pruned). At the optimum, random pruning shows a performance increase of 0.017 for Arabidopsis, 0.010 for human and 0.109 for yeast (table 5.S1 and fig. 5.4A).

To properly evaluate the performance, we compared degree-based and random pruning in two ways. The optimum of degree-based pruning was compared with the optimum of random pruning (fig. 5.3). Degree-based pruning achieves higher performance for human (0.011) and yeast (0.104), whereas for Arabidopsis,

Table 5.2: Overview of the maximum pruning states. Network information about the maximum pruned state is shown for Arabidopsis, human and yeast.

species	pruned nodes	no. of pruning steps	no. of nodes[a]	no. of BP-terms	limitation[b]
Arabidopsis	5683	111	1,748	51,706	BP-terms
human	12,856	124	1,619	85,639	nodes
yeast	2848	100	1,683	54,920	BP-terms

[a]no. of nodes that remain after pruning
[b]limiting threshold (50,000 BP-terms or 1,500 remaining proteins in network)

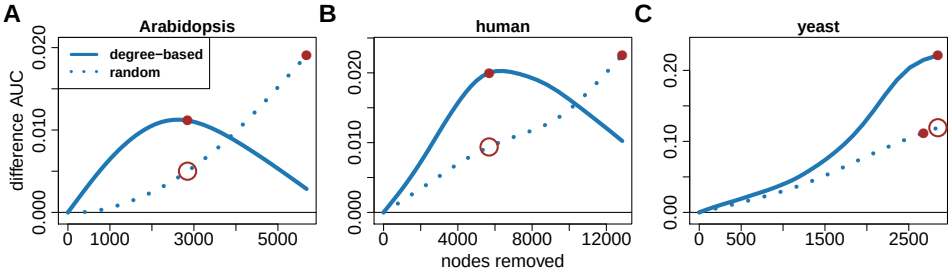


Figure 5.3: Performance curves of the pruning analyses in Arabidopsis (A), human (B) and yeast (C). The results expressed as difference AUC of degree-based pruning (solid line) and random pruning (dotted line) are plotted in the respective panels, calculated with the help of function smoothing. An optimum is indicated by red dot. Open circles indicate the performance of random pruning at the optimum of degree-based pruning. The performance in this figure might differ slightly from the boxplot (fig. 5.4), due to the application of function smoothing. The exact performance values are given in table 5.S1.

random pruning performs better (0.005). In addition, the performance of random and degree-based pruning was compared at the optimum of degree-based pruning. In this comparison, degree-based removal performs better for all three networks (fig. 5.3): 0.007 for Arabidopsis, 0.017 for human and 0.106 for yeast. These comparisons show that degree-based pruning tends to perform better than random pruning.

The performance increase of random pruning compared to the unpruned network, as well as the small difference between random and degree-based pruning for Arabidopsis and human, motivated more detailed investigations. Analysis of the masked annotation reveals that random removal retains a higher amount of experimental annotations per node (fig. 5.S1) in the later removal stages. Such annotation bias for high-degree nodes has an effect on the prediction performance of BMRF upon masking of annotation. Moreover, the fraction of nodes shared between random and degree-based removal increases on progressing node pruning (fig. 5.S2). Additional analyses with the largest network (human) were performed using a combined set of nodes used for masking annotation in such a way that degree-based and random removal were assessed on the same set of annotations and cross-validation sets. These analyses showed an increase in performance of degree-based pruning, but negligible increase of performance of random pruning (table 5.S1). This demonstrates that no issues other than the distribution of annotation data affect the performance of random pruning. The performance increases seen in random pruning are biased by the distribution of annotation. They should be considered an artifact of the experimental set-up that should not affect the conclusions of results of degree-based pruning.

To see if the impact of unequal annotation removal could be reduced further, the performance was assessed at a lesser number of nodes pruned for Arabidopsis and human. We selected 2848 pruned nodes as maximum number of nodes pruned (the maximum number used for yeast) to retain larger numbers of nodes with experimental annotation. Indeed, BMRF shows a higher performance difference between random and degree-based removal (fig. 5.S3). With only 2848 nodes pruned, degree-based pruning performs better than random pruning. Results show a AUC difference of 0.015 for human, 0.037 for Arabidopsis and 0.104 for yeast. Compared to the unpruned network, the performance increase is 0.042 for Arabidopsis, 0.023 for human and 0.212 for yeast (table 5.S1 and fig. 5.4B).

In the best unbiased set-up developed, pruning high-degree nodes improves the performance of BMRF in the prediction of protein function in biological processes. To assess if edges between hub nodes and non-hub nodes can result in functionally unrelated connections (fig. 5.2) that affect performance (Cao et al., 2013), the functional information contained in the edges of a biological network was analyzed to see how this is related to a difference in node degree. Experimentally annotated proteins were compared using semantic similarity (see section 2) as proxy for functional similarity or difference. Proteins with a high difference in degree tend to be weakly (yeast R=-0.09; human R=-0.11; Arabidopsis R=-0.25), but significantly ($p < 1e\text{-}16$) functionally different (fig. 5.5; table 5.1). The high significance indicates that such a correlation exists.

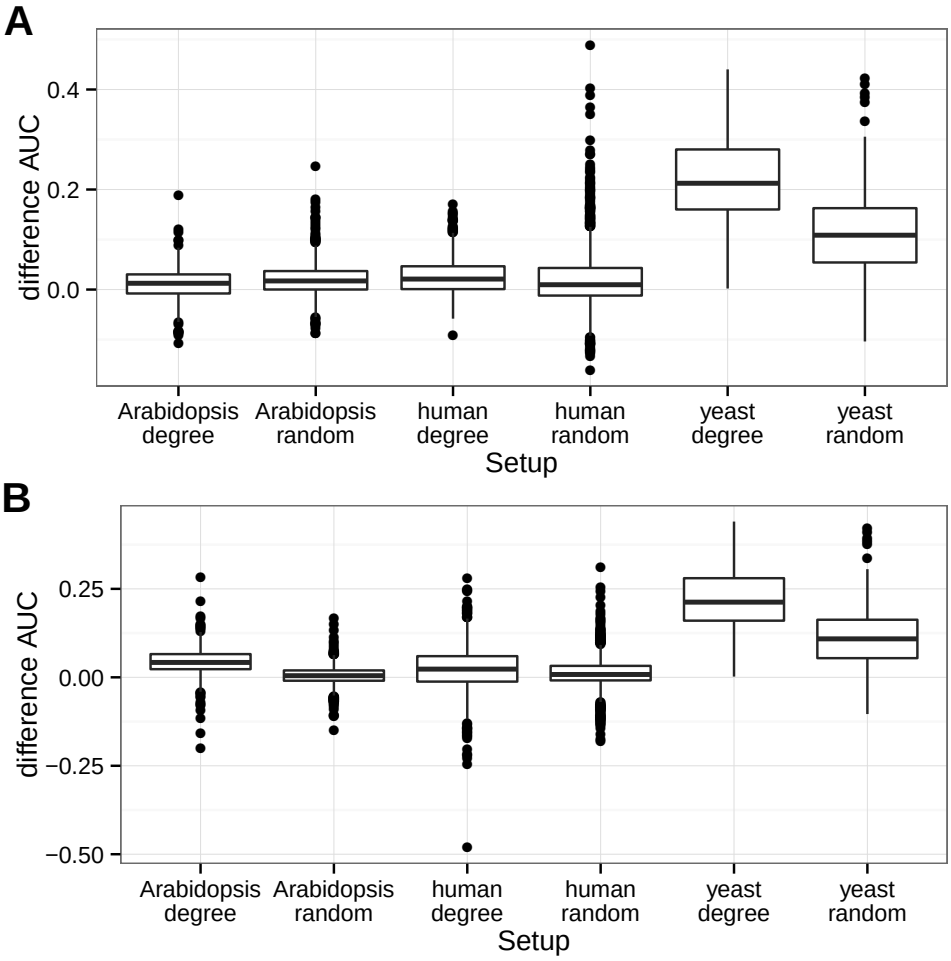


Figure 5.4: Function prediction performance of pruned networks upon degree-based (degree) and random (random) pruning for Arabidopsis, human and yeast. Shown is the box plot of the difference in AUC compared to the unpruned network at the optimum of the performance curve (fig. 5.3) for all GO BP terms in the annotation. The boxplot represents the mean and quartiles of the AUC differences (y-axis). The plot was created according to Krzywinski and Altman (2014). The AUC difference was calculated per BP term. Outliers are represented by black dots. The x-axis shows combinations (setups) of a pruning strategy (degree, random) and an organism (Arabidopsis, human, yeast). The corresponding quantitative data are given in table 5.2. (A) The performance at the optimum pruning step. (B) Same as (A), but the maximum number of nodes pruned is 2848 for all three networks. The exact performance values are given in table 5.S1.

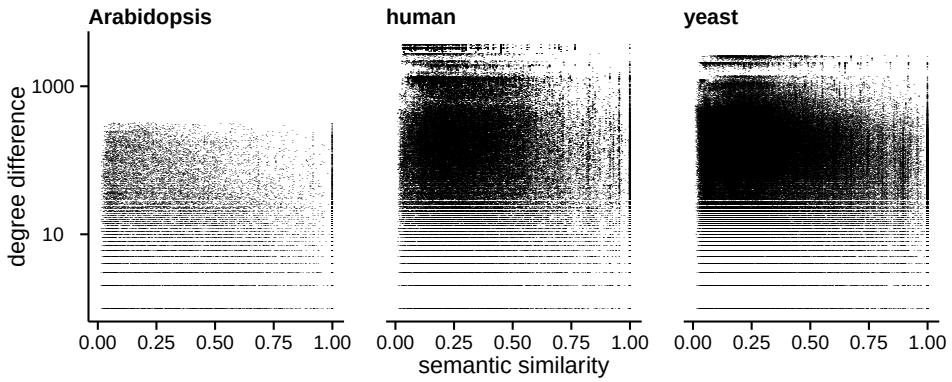


Figure 5.5: Relationship between semantic similarity and degree for the three networks. Semantic similarity is plotted as function of the degree difference of two proteins that are connected. Proteins with a high difference in degree, i.e. a hub node connected to a non-hub node, tend to have lower semantic similarity (top-right is sparse). The overall correlation between degree difference and semantic similarity is given in table 5.1.

To investigate how the performance improvement translates to single GO-terms, the prediction performance for human and yeast was analyzed in more detail. Human and yeast data were chosen because the extensive annotation that is available results in many shared GO terms. The analyses show that the same biological processes, tend to share the same trends in performance difference (R=0.42; fig. 5.6). This trend could however be influenced by the depth or frequency of GO-terms. Therefore, the frequency of a GO-term in a network was related to the performance difference in human and yeast. No, or at most a negative, association was found between performance and the number of proteins with a particular function (human: R=-0.05, p=0.137, fig. 5.S4A; yeast: R=-0.27, p=4.534e-8, fig. 5.S4B). A negative association indicates that more rare GO-terms show a higher gain in performance. Also, the depth of a GO-term was related to the performance difference. The relation is weak (human R=0.13, p=0.023; yeast R=0.17, p=0.002), indicating that the prediction of more specific GO-terms is likely to benefit more from network pruning than more general GO-terms.

In addition to node degree, other centrality measures are available to rank nodes and identify hubs in networks (Borgatti and Everett, 2006). A different centrality may give different results. We therefore evaluated the four centralities betweenness, eccentricity, local cluster coefficient and closeness in combination with the prediction performance of BMRF for the network data in yeast. Yeast was selected because of the clear performance increase obtained with degree-based removal presented above and the relatively small network size requiring less computational efforts. For all centralities tested, the performance of BMRF increases with progressing node removal (fig. 5.S5A), but the overall results do not differ from the results obtained for the centrality measure node degree.

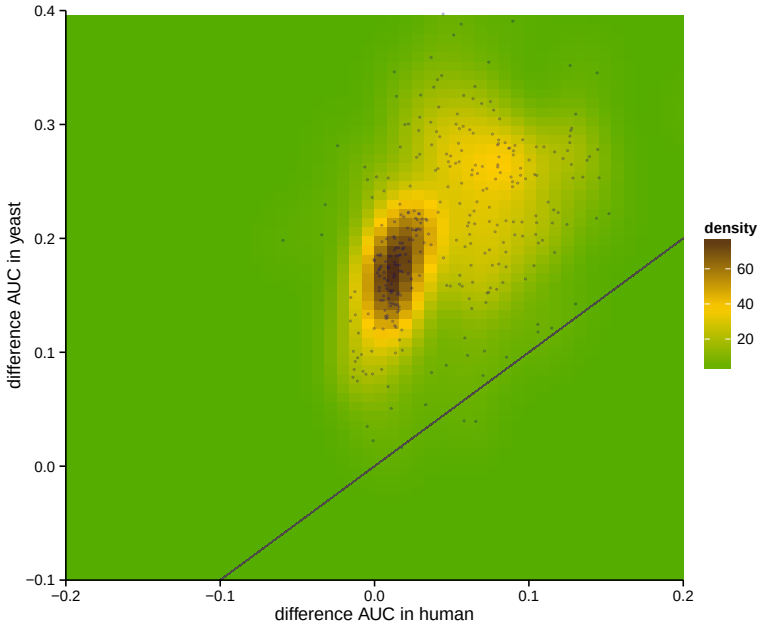


Figure 5.6: GO terms differ in performance gain. Comparison of performance gain in the yeast and human network per GO-term. Each point corresponds to the difference in AUC of a GO-term in human and yeast upon node pruning. Based on the point distribution, a density was calculated and added as background layer. Dense regions (brown) show a high concentration of GO-terms. Yeast shows for almost all GO-terms a higher performance gain (points above diagonal line). A positive performance difference indicates that a particular GO-term increased its performance on node pruning. Biological processes that show a performance increase in yeast show also a performance increase in human (R=0.42). The difference shown is based on the AUC at 2848 nodes removed.

4 Discussion

We evaluated the effect of node pruning on the performance of the prediction of the association of GO terms for biological processes with proteins with BMRF (Kourmpetis et al., 2010) as network-based function prediction algorithm and networks based on protein-protein interaction data. The results demonstrate that centrality-based node removal improves the prediction performance significantly, irrespective of the centrality measure taken. As the measure »degree« (the number of connections to other proteins) is the easiest to calculate and is the most intuitive, we conclude that degree gives the best trade-off between performance gain and simplicity; its use is therefore recommended. Using degree as centrality measure, the performance increase obtained with BMRF ranges from 0.02 to 0.20 difference in AUC. BMRF depends a lot on the amount of annotation available and therefore the improvements obtained depend strongly on the amount of an-

notation used. As the volume of experimental annotation is likely to continue to increase, BMRF-based network pruning as here presented may develop into an attractive improvement for network-based function prediction methods.

Although it is known that (noise in) network topology can influence prediction performance (Nabieva et al., 2005; Gillis and Pavlidis, 2012; Pavlidis and Gillis, 2012), only few attempts are presented in the literature that evaluate network pruning as method for the improvement of the prediction of function. Earlier analyses using edge pruning and a direct guilt-by-association approach indicated that a biological network can be reduced four-fold in size and still retain most of its functional information. It showed that a small number of edges can have a major impact on prediction performance (Gillis and Pavlidis, 2012). Here we show that removal of a large number of hub nodes actually improves prediction performance using a considerably more sophisticated algorithm than guilt-by-association, BMRF (Kourmpetis et al., 2010).

In the context of function prediction, the presence of hub nodes can apparently be unfavorable for performance. Possibly the participation of hub nodes in multiple processes results in a broad and unspecific functional profile which reduces the information content, or hubs may show a higher number of false positive annotations (Gillis and Pavlidis, 2012). The data shown here on specific vs. more general GO terms supports that indeed a broad, less specific function profile of hub nodes may explain the performance increase upon pruning. Assessing the possibility that hubs are associated with higher numbers of false-positive annotations will require more analyses.

A third possibility explaining the results of hub pruning could be that the many connections to a large number of functionally unrelated proteins result in hub nodes bridging functionally distinct modules. This way, network-based function prediction algorithms may extend the functional annotations beyond functionally related module members, which could propagate incorrect functional annotations. To what extent such bridging occurs and what the impact can be, depends on the nature of the relationship between hub and non-hub nodes. A more detailed assessment of this will require an artificial setup derived from a biological network. For this setup, potential »module-hub-module«-bridges need to be identified and analyzed. Such analysis may complement the results presented here. For a first insight, the direct interplay between hub and non-hub nodes was analyzed on the level of correlation between semantic similarity and degree. Semantic similarity (for definition used, see section 2) correlated, albeit weakly, highly significantly with the difference in degree of two connected proteins (table 5.1). The high significance is taken to show that hub nodes connected to non-hub nodes show a different functional profile than two hub nodes connected or two non-hub nodes connected. This property may be related to the bridging of modules by hub nodes, because a functional difference of hub and non-hub nodes is a necessary condition for bridging. Similar functional profiles between hub and non-hub nodes would render the to be bridged modules similar, too. Due to the weak association of the functional difference between hub and non-hub nodes, only a subset of the network is likely to

show this difference. Moreover, the (low) correlation varies considerably between the three networks analyzed. The variation could point towards differences in the sources of the network data. The networks are compiled from multiple studies (with potential differences in research focus) and detection methods (Ryan et al., 2013). This multitude of interaction data sources introduces uncertainty, which is likely to be responsible for the variation. To reduce such variation, the interaction data could be separated by their detection method. This would allow to test the effect of network pruning in relation to the detection method. It is expected to have lower variation in such a setup.

The properties and topology of a network affect the prediction performance of node pruning. The improved prediction performance may be the result of different issues coming together in the topology of a network. Due to the massive amount of unannotated proteins, the small performance increase can translate to a high number (in absolute terms) of improved annotations. However, more networks will have to be analyzed to see if there is any general issue or characteristic of the network involved and if that issue can be used for improving the performance even more.

The effect of hub-node pruning clearly depends on the type of biological process (BP) considered. The comparison of the performance for human and yeast shows that the increase of shared BP-terms is correlated. This indicates that the pruning is affecting BP terms independent of the organism. Some BP terms respond very well to pruning in general, whereas other terms have no or a negative response: not all BP benefit equally from the network pruning. A low performance in several networks (of multiple organisms) could be used as indication that such BP-terms rely on or even require the presence of hubs. However, by comparing the BP performance of yeast and human, no relation could be made to GO-term depth or specificity. A biologically intuitive pattern to be able to classify or predict the response of BP to pruning stays elusive.

A large part of the network pruning analysis was devoted towards the assessment and reduction of potential biases. All biases addressed in this analysis fall into the category annotation biases. Annotation is used to measure the performance, thus biases introduced by unequal and missing annotation coverage can influence the results. An important potential bias that is discussed often is that for many proteins the BPs in which these proteins function are not (yet) known (Huttenhower et al., 2009; Dessimoz et al., 2013; Gillis and Pavlidis, 2013). Here we consider an unknown annotation as a negative annotation, because of the lack of an established assessment method. Our approach can result in lower performance, because a correctly predicted, yet unknown BP will be considered as wrong in the evaluation. However, BMRF was used in all evaluations, so no or negligible bias is expected with respect to the methodology of analysis here presented.

In addition to the potential bias of unknown annotations, we identified and addressed other potential biases. First, the removal of nodes implies removal of annotation from the network. By pruning the network, the change in performance could reflect the reduction of annotation and not the removal of hub nodes. Sec-

ond, the coverage of experimental annotation is different between random and degree-based removal (fig. 5.S1). Removing nodes randomly decreases the annotation slower than degree-based removal. To reduce the effect of these two biases, the annotation of the network nodes was masked prior to pruning. In addition, we evaluated performance at a fixed cut-off of 2848 nodes for human, yeast and Arabidopsis to investigate the influence of the amount of annotation. In the latter set-up, a higher amount of annotation is available and as a consequence, indeed a better performance was achieved. The results show that masking is subject to a trade-off. BMRF performs best in situations where the annotation coverage is high and under-performs in environments with poor annotation coverage. Evaluation on masked setups might therefore underestimate the true performance of BMRF.

The last major bias we considered has a more subtle effect on the prediction performance. The masking of annotation reduced the bias, but random removal still showed an unexpected increase in performance, despite all precautions taken *a priori* to prevent biased assessment related to the amount of annotation. This behavior of random pruning could indicate that the results of degree-based pruning should be interpreted with a lot of caution, or that the prevention taken was not enough. The performance curve of random pruning differs from degree-based pruning (fig. 5.3), suggesting different mechanisms could be in play. From a theoretical point of view, biological networks should be resistant to random pruning of nodes (Barabási and Bonabeau, 2003). In case of an uneven distribution of experimental annotation, random pruning will remove a large fraction of unannotated, low-degree nodes. As a result, the fraction of randomly removed hub-nodes decreases with the size of the network. Therefore, the largest performance increase of random removal is seen in the network of yeast, where a relatively high amount of nodes is pruned. To test whether the behavior of the random pruning approach still suffers from differences in annotation (and removal of annotation upon pruning), the ranked lists were combined in such a way that the same sets of annotation could be pruned in both random and degree-based pruning. This allowed using the same cross-validation sets for random and degree-based pruning. In this direct comparison, random pruning did not show improvement of predictions, whereas degree-pruning did (results not shown). This demonstrates that indeed the amount and relative composition of annotation terms interferes with BMRF performance. More analysis and research will be necessary to see if equal-annotation-set pruning is feasible as routine analysis in biological networks and how it compares to the analyses presented here. The analysis of network pruning presented here should be considered as a first step into improving network-based function prediction. Many future approaches are conceivable improve both assessment and methodology of network pruning.

The first improvement will result from the ongoing annotation efforts of the Gene Ontology Consortium. As annotation will accumulate in the future, the annotation- and network coverage of organisms (likely notably of model organisms) will reach higher levels. In such case, node pruning will become much more powerful. A large gap in terms of network data and annotation coverage exists

between model- and non-model organisms. Once the network and annotation data in non-model organisms also reaches higher levels, a much more detailed analysis of the network-pruning-effect becomes possible. In addition, also the quality of existing experimental annotation is likely to increase (Skunca et al., 2012), leading to a more comprehensive and consistent starting point for network pruning.

To increase not only the amount of annotation, but also the amount of available network data, data from multiple species could be combined. There are many possibilities to combine networks from multiple species. However, integration of such networks can lead to contrary effects. It was demonstrated that the same network across species can result in different phenotypes (McGary et al., 2010; Ideker et al., 2011) and, vice versa, different networks across species result in similar responses (Erwin and Davidson, 2009; Ideker et al., 2011). Without further assessment, the combination of cross-species networks and network pruning might lead to unexpected results. Yet, the basic property of hub nodes impeding function prediction performance may be preserved even in cross-species networks. Hence, incorporating more networks and species in the analyses is an interesting option for increasing the performance of BMRF with network pruning.

Overall, the positive effect of pruning on prediction performance demonstrates that hub nodes can hamper prediction performance. The positive effect of hub pruning does not seem to depend on the particular organism or network (Mossa et al., 2002; Gillis and Pavlidis, 2012; Winterbach et al., 2013). Analysis of pruned networks and their hubs is therefore a relatively easy way to improve network-based function predictions with BMRF. In practical applications, it will be essential to determine the optimal pruning step for any given network. Given an optimum pruning step, we suggest doing two separate prediction runs. The first run is with the full network. This allows the prediction of hub node functions. And the second run is a subset of the network, containing only non-hub nodes and edges. The functions predicted for the subset are used as higher quality non-hub node annotations. This strategy can be applied to most network-based function prediction algorithms without further modification. It would make network pruning an attractive option for future protein function prediction.

Acknowledgements

This work was supported by the FP7 »Infrastructures« project TransPLANT (award 283496) and by the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

5 Supporting Information

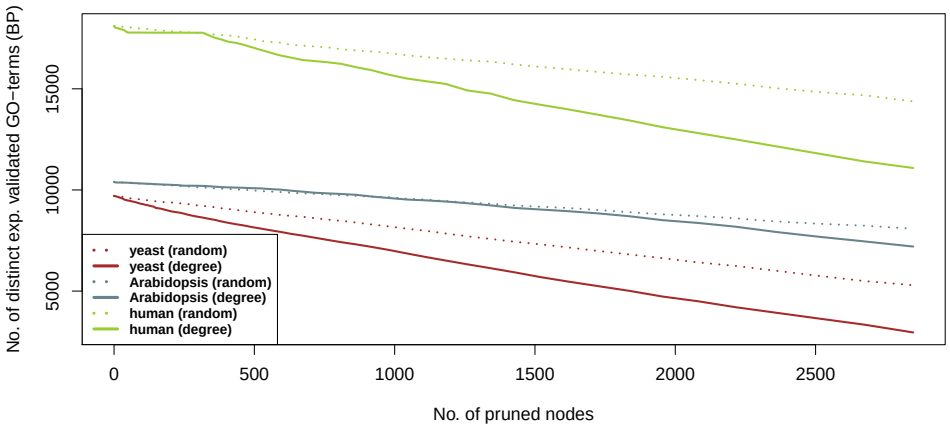


Figure 5.S1: Impact of node removal on annotation coverage. Degree-based node pruning removes more annotation than random pruning. For biological processes (BP), experimental annotation (exp. validated GO-terms) is concentrated at nodes with a high degree, thus the number of GO-terms per protein decreases faster when nodes are removed by degree compared to nodes removed randomly.

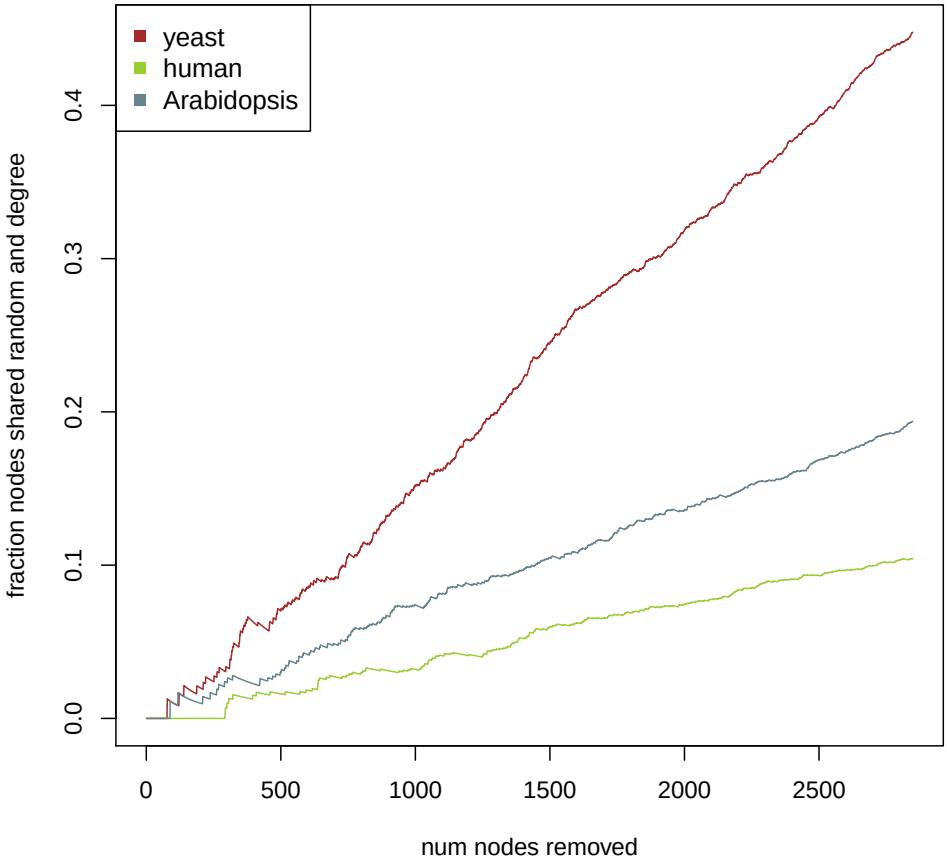


Figure 5.S2: Intersection of random and degree-based pruning sets. The number of nodes shared between random and degree-based removal increases in later pruning stages. Yeast shows the highest overlap, followed by Arabidopsis and human. The maximal removal stage is 2848.

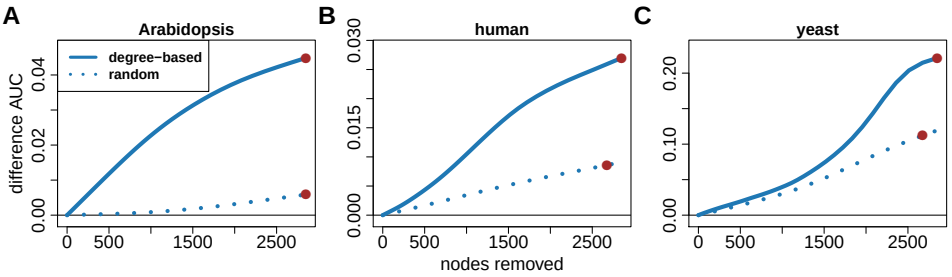


Figure 5.S3: Performance curves of the pruning setups. The networks were pruned degree-based and randomly. The calculations were performed in Arabidopsis (A), human (B) and yeast (C). The maximum number of removed nodes is 2848 nodes. All species reach the highest performance difference at 2848 pruned nodes. Optima are indicated by filled red circles. The performance in this figure might differ slightly from the boxplot (fig. 5.4), due to the application of function smoothing. The exact performance values are given in table 5.S1.

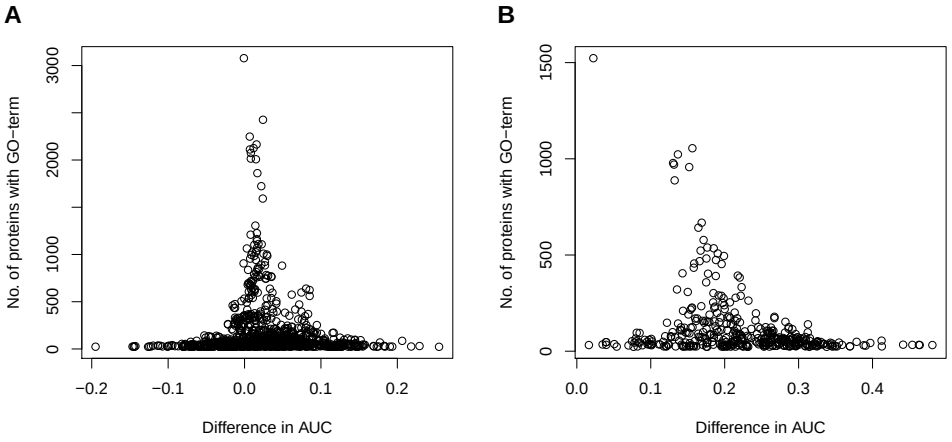


Figure 5.S4: Comparison of GO-term frequency and performance difference in human and yeast. The frequency of a GO-term in a network was related to the performance difference in human and yeast. (A) The difference in AUC (positive = performance increase) is not correlated with the number of proteins performing a particular function in human (R=-0.05, p=0.13). (B) The difference in AUC (positive = performance increase) is weakly correlated with the number of proteins performing a particular function in yeast (R=-0.27, p=5e-8). The correlation is negative, rare GO-terms tend to have a higher performance increase.

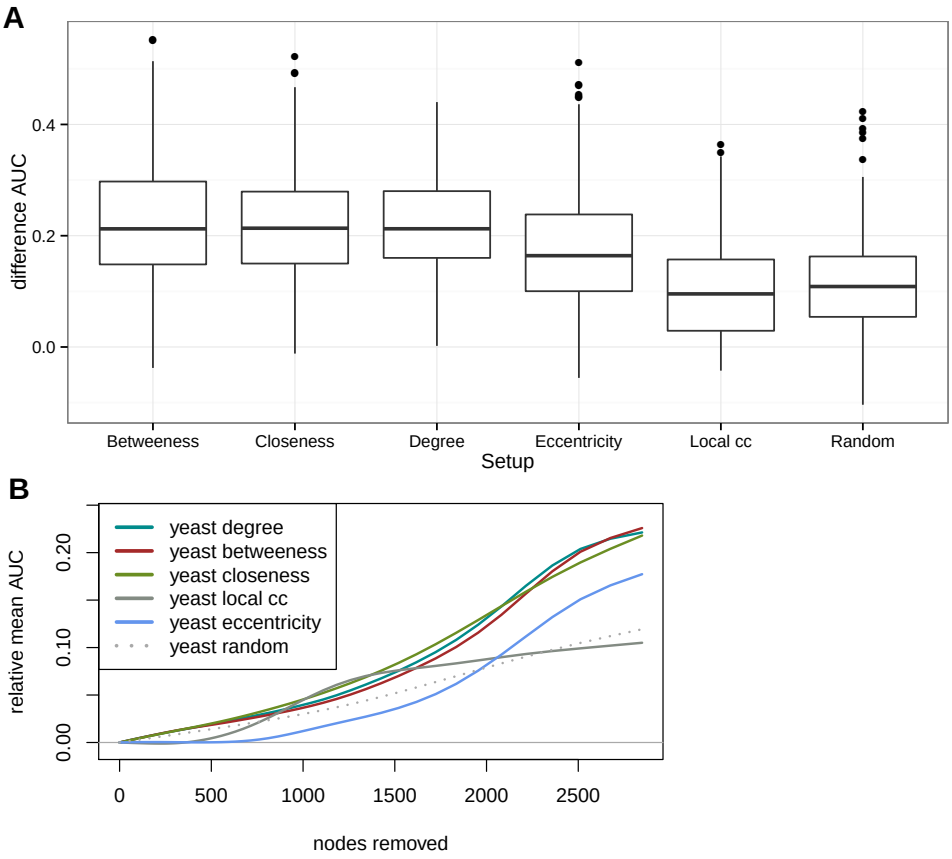


Figure 5.S5: Function prediction performance of different centralities in yeast. The yeast network was pruned by 6 different setups, corresponding to the centralities betweenness, eccentricity, local cluster coefficient, closeness and degree. Shown is the performance difference compared to the unpruned network (A) Shown is the box plot of the difference in AUC at the optimum of the performance curve for all GO BP terms in the annotation. The boxplot represents the mean and quartiles of the AUC differences. The plot was created according to Krzywinski and Altman (2014). The AUC was calculated per BP term. Outliers are represented by black dots. The corresponding quantitative data are given in table 5.2. (B) The performance (y-axis) increases on progressing node removal, reaching the highest value at 2848 nodes (maximum pruned state). The performance in this figure might differ slightly from the boxplot in subfigure (A) due to the application of function smoothing. The exact performance values are given in table 5.S1.

Table 5.S1: Overview of the pruning performances. The overview is divided by analysis. For each pruning strategy and organism (setup), the AUC values of the unpruned (unpr.) and optimum (opt.) states are listed. AUC differences (diff.) between optimum and unpruned AUC are shown in separate columns. Additionally, the AUC values and differences for the smoothed performance curve (smooth) are listed. Corresponding to the AUC values, the number of nodes pruned (nodes pr.) are shown for the optimum and the final pruning step (final).

setup	median AUC (unpr.)	nodes pr. (opt.)	nodes pr. (final)	median AUC diff.	median AUC (opt.)	AUC diff. (smooth)	smooth AUC (opt.)	nodes pr. (opt. smooth)
pruning till annotation/node threshold								
Arabidopsis degree	0.6776	2848	5683	0.0125	0.6973	0.0112	0.7010	2675
Arabidopsis random	0.6827	5683	5683	0.0172	0.7008	0.0191	0.6998	5683
human degree	0.7437	5683	12856	0.0210	0.7717	0.0203	0.7625	6443
human random	0.7898	12856	12856	0.0096	0.8097	0.0225	0.8080	12856
yeast degree	0.6087	2848	2848	0.2125	0.8307	0.2213	0.8324	2848
yeast random	0.7371	2675	2848	0.1087	0.8474	0.1193	0.8470	2848
random performance at degree optimum								
Arabidopsis random	0.6827	2848	5683	0.0052	0.6890	0.0191	0.6998	5683
human random	0.7898	5683	12856	0.0035	0.8007	0.0225	0.8080	12856
yeast random	0.7371	2848	2848	0.1067	0.8473	0.1193	0.8470	2848
final pruning step 2848								
Arabidopsis degree	0.7375	2848	2848	0.0421	0.7870	0.0448	0.7792	2848
Arabidopsis random	0.7139	2848	2848	0.0046	0.7174	0.0059	0.7196	2848
human degree	0.6947	2675	2848	0.0232	0.7202	0.0270	0.7185	2848
human random	0.7238	2848	2848	0.0080	0.7298	0.0091	0.7284	2848
yeast degree	0.6087	2848	2848	0.2125	0.8307	0.2213	0.8324	2848
yeast random	0.7371	2675	2848	0.1087	0.8474	0.1193	0.8470	2848
same annotation for random and degree								
human random	0.7065	3438	9930	0.0022	0.7083	0.0015	0.7014	4151
human degree	0.7063	8820	9930	0.0303	0.7388	0.0351	0.7385	9391
pruning by other centralities (yeast)								
Degree	0.6087	2848	2848	0.2125	0.8307	0.2213	0.8324	2848
Betweenness	0.6244	2848	2848	0.2124	0.8484	0.2259	0.8460	2848
Closeness	0.6347	2848	2848	0.2135	0.8434	0.2180	0.8469	2848
Local cc	0.7435	2675	2848	0.0954	0.8457	0.1051	0.8444	2848
Eccentricity	0.6750	2848	2848	0.1641	0.8464	0.1773	0.8469	2848
Random	0.7371	2675	2848	0.1087	0.8474	0.1193	0.8470	2848

Text 5.S6: Definition of the pruning-step function.

Due to computational constraints, it was not possible to test all pruning steps. Therefore, we selected a function to create distinct steps on an exponential scale to cover the space of pruned network states. The function has the form:

$$f(x) = 10^{0.1+3\frac{x}{110}}$$

with $x = 0, 1, 2, \ldots, N$; the steps were rounded to integer number and duplicated values were removed.

Chapter 6

General Discussion

The research presented in this thesis focuses on deriving function from sequence information, with the emphasis on plant sequence data. The connection between sequence information and function was approached on the level of chromosome structure (chapter 2) and of gene families (chapter 3) using combinations of existing bioinformatics tools. The applicability of using interaction networks for function prediction was demonstrated by first markedly improving an existing method (chapter 4) and by exploring the role of network topology in function prediction (chapter 5). Taken together, the combination of methods and results indicate the potential as well as the current state-of-the-art of function prediction in (plant) bioinformatics.

The number of sequenced genomes is growing fast. About 1100 human genomes are now in the public domain, or announced, and analyses focus on structural variations in connection with any phenotype or disease of interest (Abecasis et al., 2012). About 38 million single nucleotide polymorphisms (SNPs) are available for analysis of the human genome, but results so far illustrate the high complexity of complex traits (Lupski et al., 2011).

Also in plants, the number of sequenced genomes is increasing. The attention is shifting from the few model plants (Bevan and Walsh, 2005; Paterson et al., 2005; Morrell et al., 2011) to agronomically important plant species. Challenge is to translate the findings in model species to real crops. After the model species *Arabidopsis thaliana* (Arabidopsis), many more plant species have been sequenced (Hamilton and Buell, 2012). These are often considered »model« for a particular trait-of-interest not offered by Arabidopsis (Paterson et al., 2005; Morrell et al., 2011). In addition, projects involving re-sequencing of segregating populations or selections of biological variation are bridging the gap between genomics and plant breeding. There are now more than 80 plant species sequenced, of which the largest known genome is carried by *Pinus taeda* (Neale et al., 2014). Its draft genome of 23.2 Gb was released in the beginning of 2014. Issues are genome size, genome complexity (repeats) and ploidy. In addition, the number of resequenced plant genomes is increasing. The first major effort was to capture the genetic variation in Arabidopsis (Cao et al., 2011; Schmitz et al., 2013) in the form of the Arabidopsis 1001 Genomes Project. The project yielded the first comprehensive catalog of common SNPs as well as small- and large-scale rearrangements, such as deleted, duplicated or non-reference regions. This effort marks the first step into deciphering the adaptation of Arabidopsis to diverse environments. Following the footsteps of human GWAS studies, the project aims to cover most of the common variants. Currently this spans over 1,000 strains with about 216,000 tagged SNPs, covering approx. 90% of all common variants (Cao et al., 2011). Similar efforts were completed for rice (Subbaiyan et al., 2012; Xu et al., 2012), soybean (Lam et al., 2010) and maize (Lai et al., 2010; Hufford et al., 2012). Currently, the 150 Tomato Genome ReSequencing (www.tomatogenome.net) project aims at identifying and exploring the genetic variation in tomato. The results will allow to study recombination and identification of alleles that have been lost during domestication (www.tomatogenome.net; Causse et al., 2013).

Yet, it is getting clear that having the sequence of a plant genome alone is not sufficient. The real challenge is what to do with all that sequence data. How to make sense of the bulge of sequence data and how to use it most efficiently for crop improvement is a key challenge (Jackson et al., 2011). Data generation may nowadays be straightforward and affordable, the interpretation and analytical approach towards understanding of function is, however, far from trivial (Mardis, 2011). The quest for function is not only motivated by the desire to know, but also by the need to use. Plant breeding is directed towards improving plants for human benefit (Xu, 2009). Its efforts tend to focus on yield, disease resistance, agronomical performance and quality of for example fruits or grains. Ways to utilize sequence data are still at the beginning. One way of utilization is to understand the origins and domestication of crop plants. Understanding the origins and domestication of crop plants is considered essential for the identification and use of the appropriate genetic resources and loci of agronomical interest for crop improvement (Morrell et al., 2011). Plant genome sequencing should help such understanding and contribute to the improvement of plants for human use.

The function implied in plant sequence data can be assessed at different levels of organization. Laboratory and field experimentation are often preferred methods, but they are costly and slow (Lee et al., 2007). Sequencing was the biggest cost factor in sequencing projects 14 years ago. Nowadays, due to the reduction in sequencing costs, the biggest cost factors are downstream analyses, including annotation of function. It has been estimated that future sequencing projects need to allocate more than 50% of their financial resources to analyses not directly connected to sequencing itself (Sboner et al., 2011). Current projects, however, focus on the sequence production and (partly) neglect the importance of follow-up experiments (Sboner et al., 2011). In the absence of experimental data, computational analysis is the method explored in this thesis. In general terms it involves comparison; comparison with known sequences, known functions or known interactions. Behind almost all approaches of comparison are the concepts of evolution, selection and descent.

A first level of comparison for functional inference is comparison of genome structure (Zheng et al., 2004). This is a first step in the genetic underpinning of plant breeding and selection in plant breeding (chapter 2). The earlier development of (molecular) markers and marker-based mapping studies provided insight in the structural organization of plant genomes, including tomato and potato (Bennetzen, 2000a). Such maps tend to have a low marker density and a limited accuracy that prevent good local resolution of chromosomal organization (chapter 2 ;Bennetzen, 2000a; Liu et al., 2012). Genome sequencing offers the highest information density feasible and genome comparison will show most if not all details of structural variation and possibly the evolutionary history of plant genomes.

In chapter 2 of this thesis, a first step into resolving the rearrangement phylogeny within the Solanaceae is presented on the basis of extensive sequence comparisons. Syntenic loci have been identified for selected chromosome regions in *Solanum* and *Capsicum* genomes. We have identified collinear segments and

collinearity breaks. These breaks do not only occur in heterochromatic portions, but also frequently in the euchromatic portion of the chromosome regions we have investigated. The combination of genetic maps and comparative sequence analysis provide a valuable resource for resolving the chromosome organization at the structural level. Several structural differences observed between tomato and potato have not been seen in existing linkage maps. Some rearrangements point towards specific recombination events in the tomato clade that occurred after the split from the last common ancestor of tomato and potato.

The results from the synteny study pave the way for DNA-based selection in introgression breeding. Similar analyses of more genomes may identify hotspots for recombination and/or cases of linkage drag that are difficult to prevent. Such knowledge of structural issues and functions of plant genomes may help future selections of (more) suitable parents in a (pre)breeding program. The results can also motivate strategies to mine structural genetic diversity to develop genome-based breeding tools that will accelerate breeding for targets-of-interest. Current efforts, such as the Plant Genome Database (Duvick et al., 2008), Plant Genome Duplication Database (Lee et al., 2013b), CoGepedia (Lyons and Freeling, 2008), PLAZA (Van Bel et al., 2012) or Phytozome (Goodstein et al., 2012), try to integrate, centralize and visualize structural properties of plants. Future sequencing projects may integrate their data seamlessly into these platforms. This allows to automate and standardize the most common set of analyses, including structural rearrangements and structural genome annotation.

A level of function more directly related to sequence data than the structural functionality outlined in chapter 2, are the functions as defined by the domains of the Gene Ontology project (Gene Ontology Consortium, 2000): cellular components (location), molecular function (catalytic activity, binding and the like) and biological process (e.g. tuber formation), defined as the sets of operations or molecular events with a defined beginning and end, that are pertinent to the functioning of integrated living units such as cells, tissues, organs or whole organisms (www.geneontology.org).

In the context of translating functions from model plants to agronomical relevant crops, gene family analysis is an option (Martinez, 2013). A comparative genomic analysis of the Snf2 gene family is presented in chapter 3 of this thesis. Snf2 family ATPases function in large protein complexes. They are responsible for energy supply during chromatin remodeling and influence many processes in plants. Analyses in the model species Arabidopsis show a possible link with the response to environmental stress (Kanno et al., 2004; Huettel et al., 2007), which would be a possible lead for translation to non-model plants such as tomato or potato. The precise molecular mechanism of action of many of these proteins remains however largely unknown (chapter 3; Knizewski et al., 2008).

Chapter 3 of this thesis presents the first comprehensive study of Snf2 family ATPases in available plant genomes. The number of Snf2 ATPases shows considerable variation across plant genomes, suggesting a broad functional diversification within this gene family. Some subfamilies of the Snf2 gene family are re-

markably stable whereas others show expansion and contraction in several plants. One of these subfamilies, the plant-specific DRD1 subfamily, is non-existent in lower eukaryote genomes, yet it developed into the largest Snf2 subfamily in plant genomes. It shows the occurrence of a complex series of evolutionary events. Its expansion, notably in tomato, suggests novel functionality in processes connected to chromatin remodeling. Members of this subfamily could make suitable targets for breeding and plant improvement to target environmental stress tolerance and yield in future breeding, for example in reducing QTL×environment interactions.

Gene family analysis as presented in chapter 3 is an example of downstream analysis that assumes proper genome assembly and annotation. Experience has shown, however, that such assumptions must be considered carefully. Unfortunately, information about the quality of data is scarce in structural genome annotations. When gene models are taken for granted (which is common practice in comparative genomics), but wrong, it can lead to misinterpretations (Jones et al., 2007; Lee et al., 2007; Schnoes et al., 2009). Also, the amount of missing annotation – functions and properties not (yet) associated with a gene – should be estimated and taken into account (Dessimoz et al., 2013). The analysis pipeline for such biological data generally consists of a chain of tools, in which the output of one tool is the input for the next tool. Errors introduced in the beginning of the analytical chain continue, with potentially severe effects on the final results. In this thesis, tools were combined largely by hand because that gives most flexibility and best error control. Work-flow management approaches such as Galaxy (Goecks et al., 2010), or integrated approaches as the commercial package CLCBio (http://www.clcbio.com/), may reduce the work load of combining tools. However, they do not safeguard against error propagation. By the nature of error propagation, such errors will only be discovered by an end-user (Hamilton and Buell, 2012), when using the final result of the analysis for interpretation, further experimentation or, in case of plant breeding, selection.

In case of the identification of Snf2 family members, publicly available genome assemblies and annotations were indeed identified as sources of error. An example was encountered in potato, where a considerable number of Snf2 family members was absent from the at-that-time actual genome annotation and could only be detected by iterative rounds of homology-based gene prediction. Such a careful approach minimizes errors in the gene models and facilitates downstream analyses. However, due to the lack of community-based platforms, the improved annotation remains hidden as supplementary information of the respective publication (Bargsten et al., 2013). When annotation information is scattered among multiple publications, it requires a lot of manual work to piece everything together. In particular non-model plants suffer from this situation.

Better situations could be supplied by bioinformatics approaches spanning multiple genome projects that were recently put into practice (Loveland et al., 2012; Sterck et al., 2012; Lee et al., 2013a) and deserve more funding. Such community-based efforts now focus on structural genome annotation, but conquering functional annotation should be considered as one of the main future

challenges of plant genome data management and maintenance. The integration of analyzed gene families into databases is one essential requirement for systematic analyses of function (Martinez, 2011). Integrated gene family data can provide a useful resource for plant breeders, first for the correct identification of orthologs and, second, to generate agronomically relevant leads. An example lead would be the expansion of the stress-related DRD1 subfamily in tomato (chapter 3).

Gene family analysis is only a step into inferring potential functions of gene family members. The analyses in chapter 3 show that it is not feasible to infer more precise functional characteristics based on sequence data alone. The function particularly difficult to infer is the biological process in which a gene product is involved. The more detailed the biological process is desired to be, the more difficult this is. For breeders, knowledge of the biological process would seem to be a very useful description of function. To improve the inference of function, additional information sources than sequence data alone are necessary (Rentzsch and Orengo, 2009). The integration of multiple, complementary data sources presented in ways that help interpretation is one of the many challenges for future bioinformatics. Rich and often complementary information sources include biological networks of either protein interactions, co-expression of genes or otherwise.

In chapter 4 of this thesis, it is demonstrated that such biological networks can improve the inference of function significantly, especially in absence of experimental data. Function prediction methods, such as BMRF, require existing (training) data commonly in the form of experimentally annotated gene functions to achieve a high performance. This training data is needed to propagate function information via the connections of a biological network to unannotated proteins. Plants, with the exception of Arabidopsis and rice, lack such training data. To be able to apply BMRF in such environments, we combined BMRF with the sequence-based method Argot2 (Falda et al., 2012). Argot2 was used to create a high quality training data (seed) set for BMRF. The combination of sequence- and network-based function prediction obtained by seeding BMRF with Argot2 offers significant benefits over applying these methods separately. In this constellation, not only the prediction performance for biological processes was improved markedly, but also plant species with a low amount of training data could be annotated consistently. This opens up opportunities to generate predictions for many so far unannotated proteins. In the context of the study, we predicted functions for *Oryza sativa* (rice), *Medicago truncatula* (barrel clover), *Glycine max* (soybean), *Populus trichocarpa* (poplar) and *Solanum lycopersicum* (tomato).

Even though the sequence-based prediction algorithm can be easily substituted by other methods, such as Blast2GO (Conesa et al., 2005), we chose Argot2 because of its performance in CAFA (Radivojac et al., 2013). This flexibility combined with the availability of RNA-seq-based co-expression networks (Marguerat and Bähler, 2010) will make this network-based method a powerful and attractive addition to sequence-based protein function prediction. The approach strongly depends on accurate input data. First, function annotation transferred via homology from model organisms, such as Arabidopsis, to an unannotated protein

need to be correct. Correct in terms of annotation in the model organism and in terms of detected homology between an annotated and unannotated protein. Errors in this step lead to error propagation in BMRF, as it merely propagates the seed annotation via the network to unannotated proteins. A careful selection of the sequence-based method is therefore important. Second, the quality of the biological network used by BMRF directly affects the predictions. In analyses as presented and developed in chapter 4, network data is used as input without any filtering steps. Network data generated by experiments, such as yeast two-hybrid or co-expression, often come with a potentially high level of noise (Zhu et al., 2013) and disturbing topological properties. Noise is inherent in biological networks. In particular in lowly expressed genes, stochastic effects can become prominent, leading to random connections in networks (Raser and O'Shea, 2005). In addition, the overlay of two fundamentally different network topologies, modular and scale-free-like, could affect function prediction negatively. Function prediction algorithms, such as BMRF, may improve their performance upon removal of such disturbing elements. These effects are arguably the downside of experimentally acquired data. For example it has been shown in genome assembly that breaking up reads into shorter k-mers is very advantageous (Compeau et al., 2011). Another example is the Snf2 gene family analysis (chapter 3). Focusing on the core region of a protein, ignoring more than 50% of the sequence, is sufficient to reconstruct the complete gene family tree accurately. The concept of noise in combination with data necessity could imply that the prediction of protein function is improved when noisy data are identified and removed.

In chapter 5 of this thesis, the influence of noise on prediction performance is evaluated by assessing network topology and removing nodes. Proteins that are highly connected in a network were identified as disturbing elements. These elements could connect unrelated proteins, misleading function prediction algorithms. Such proteins tend to be subject to intensive investigation due to their essential role for an organism and tend to involved in a high number of distinct biological processes (Jeong et al., 2001; He and Zhang, 2006). Such characteristics can lead to prediction bias and often contain – following the definition of information by Shannon (Shannon, 1948) – low functionally relevant information (Gillis and Pavlidis, 2012). The results as presented in chapter 5 show that identifying and removing such proteins (and their connections) improves the performance of BMRF significantly. The identification and removal of noise intrinsic in biological data also positively affects the actual computation time and storage space needed to cope with the growing data volumes. Computation and storage time is often overlooked as challenge in bioinformatics.

As soon as noise or redundant data are detectable and removed, performance and computation time and space are improved. However, every data source requires an analysis of its nature with respect to information content, possible downstream analyses and possible errors. If one of these properties is unknown, the data should be stored in its raw format. As soon as the intersection, the common denominator of all analyses, is found, raw data can be transformed/com-

pressed. But here, at the intersection between raw data and transformed data, the error propagation starts, at least from a computational perspective. Once the raw data is deleted, it is not possible to trace back potential errors introduced by the transformation or compression. An example would be the bias in Illumina transcriptome sequencing caused by non-random hexamer priming (Hansen et al., 2010). Once the gene expression is determined and the raw data is not available anymore, it is hardly possible to detect such bias. In biological networks, the definition of noise strongly depends on the analysis applied. Thus, a general approach to remove noise from networks is likely to be non-existent. The raw data should be kept available to allow multiple views on the same data.

In conclusion, different approaches to filter network data should be considered a promising future direction for bioinformatics. With further experiments it will be possible to explore and utilize the biological networks in many different contexts, including protein function prediction, disease gene prioritization and network-based genome-wide association studies (Yu et al., 2013). One interesting aspect in protein function prediction is the usage of tissue- or time-specific networks. These networks would allow studying function on a finer-grained level.

The work presented in this thesis shows how the computational methods of bioinformatics can contribute to the biologically relevant interpretation of the large volumes of data nowadays generated by genome and transcriptome sequencing. In this way, the field of bioinformatics adds value to the production of large volumes of data. The future of plant breeding is likely to see the sequencing of all commercial species, subspecies and pathogens (Eggen, 2012). The next step will be to integrate the data of the genome, the transcriptome, the epigenome and all other omics levels that can be defined and measured, with the phenotype of interest. Such integration and subsequent generation of usable knowledge will be a future challenge for notably bioinformatics, in combination with the field of biology known as »systems biology«.

Overall, the technical requirements for such integration are already met. First efforts into integrating plant genomes are visible in projects, such as Phytozome (Goodstein et al., 2012), Gramene (Monaco et al., 2014), SolGenomics (Bombarely et al., 2011) and Ensembl Plants (Kersey et al., 2014). The core goal will be to extend these projects to integrate all the different data sources. This will allow exploring new ways for elucidating and understanding complex functionality. There is no standard approach to relate the integrated data to complex biological systems. Therefore, bioinformatics and biology will continue to merge and eventually become so tightly integrated that the term »biology« may be sufficient again.

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Abelenda, J. A., Navarro, C., and Prat, S. (2014). Flowering and tuberization: a tale of two nightshades. *Trends in Plant Science*, 19(2):115–122.

Albrecht, E. and Chetelat, R. T. (2009). Comparative genetic linkage map of *Solanum* sect. *Juglandifolia*: evidence of chromosomal rearrangements and overall synteny with the tomatoes and related nightshades. *Theoretical and Applied Genetics*, 118(5):831–847.

Alföldi, J. and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Research*, 23(7):1063–1068.

Allen, J. E. and Salzberg, S. L. (2005). JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, 21(18):3596–3603.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, 8(5):e1002514.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Anderson, L. K., Covey, P. A., Larsen, L. R., Bedinger, P., and Stack, S. M. (2010). Structural differences in chromosomes distinguish species in the tomato clade. *Cytogenetic and Genome Research*, 129(1-3):24–34.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333(6042):601–607.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M. et al. (2011). The genome of *Theobroma cacao*. *Nature Genetics*, 43(2):101–108.

Ashrafi, H., Kinkade, M., and Foolad, M. R. (2009). A new genetic linkage map of tomato based on a *Solanum lycopersicum* × *Solanum pimpinellifolium* RIL population displaying locations of candidate pathogen response genes. *Genome*, 52(11):935–956.

Bai, Y. and Lindhout, P. (2007). Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Annals of Botany*, 100(5):1085–1094.

Bai, Y., van der Hulst, R., Huang, C. C., Wei, L., Stam, P., and Lindhout, P. (2004). Mapping Ol-4, a gene conferring resistance to *Oidium neolycopersici* and originating from *Lycopersicon peruvianum* LA2172, requires multi-allelic, single-locus markers. *Theoretical and Applied Genetics*, 109(6):1215–1223.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5):60–69.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Bargsten, J. W., Folta, A., Mlynárová, L., and Nap, J.-P. (2013). Snf2 family gene distribution in higher plant genomes reveals DRD1 expansion and diversification in the tomato genome. *PLoS ONE*, 8(11):e81147.

Bedinger, P. A., Chetelat, R. T., McClure, B., Moyle, L. C., Rose, J. K. C., Stack, S. M., van der Knaap, E. et al. (2011). Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. *Sexual Plant Reproduction*, 24(3):171–187.

Bennetzen, J. L. (2000a). Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell*, 12(7):1021–1029.

Bennetzen, J. L. (2000b). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, 42(1):251–269.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(Database issue):D36–42.

Bevan, M. and Walsh, S. (2005). The Arabidopsis genome: a foundation for plant research. *Genome Research*, 15(12):1632–1642.

Bezhani, S., Winter, C., Hershman, S., Wagner, J. D., Kennedy, J. F., Kwon, C. S., Pfluger, J. et al. (2007). Unique, shared, and redundant roles for the Arabidopsis SWI/SNF chromatin remodeling ATPases BRAHMA and SPLAYED. *Plant Cell*, 19(2):403–416.

Blanc, G. and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*, 16(7):1679–1691.

Bolanos-Garcia, V. M., Wu, Q., Ochi, T., Chirgadze, D. Y., Sibanda, B. L., and Blundell, T. L. (2012). Spatial and temporal organization of multi-protein assemblies: achieving sensitive control in information-rich cell-regulatory systems. *Philosophical Transactions of the Royal Society, Series A*, 370(1969):3023–3039.

Bombarely, A., Menda, N., Tecle, I. Y., Buels, R. M., Strickler, S., Fischer-York, T., Pujar, A. et al. (2011). The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Research*, 39(Database issue):D1149–1155.

Bonierbale, M. W., Plaisted, R. L., and Tanksley, S. D. (1988). RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics*, 120(4):1095–1103.

Borgatti, S. P. and Everett, M. G. (2006). A Graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–438.

Brandão, M. M., Dantas, L. L., and Silva-Filho, M. C. (2009). AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinformatics*, 10:454.

Braun, P., Aubourg, S., Van Leene, J., De Jaeger, G., and Lurin, C. (2013). Plant protein interactomes. *Annual Review of Plant Biology*, 64:161–187.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D'Amore, R., Allen, A. M., McKenzie, N. et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710.

Brown, J. R., editor (2007). *Comparative genomics: basic and applied research*. CRC Press, Boca Raton, USA.

Budiman, M. A., Chang, S.-B., Lee, S., Yang, T. J., Zhang, H.-B., de Jong, H., and Wing, R. A. (2004). Localization of *jointless-2* gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. *Theoretical and Applied Genetics*, 108(2):190–196.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R. et al. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41(Database issue):D226–D232.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C. et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188–196.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D. et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43(10):956–963.

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE*, 8(10):e76339.

Causse, M., Desplat, N., Pascual, L., Le Paslier, M.-C., Sauvage, C., Bauchet, G., Bérard, A. et al. (2013). Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, 14:791.

Chang, S.-B., Yang, T.-J., Datema, E., van Vugt, J., Vosman, B., Kuipers, A., Meznikova, M. et al. (2008). FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome Research*, 16(7):919–933.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J. et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue):D816–D823.

Chen, F., Mackey, A. J., Vermunt, J. K., and Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383.

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, 7(3-4):429–447.

Chetelat, R., Cisneros, P., Stamova, L., and Rick, C. (1997). A male-fertile *Lycopersicon esculentum* × *Solanum lycopersicoides* hybrid enables direct backcrossing to tomato at the diploid level. *Euphytica*, 95(1):99–108.

Chetelat, R. T. and Ji, Y. (2007). Cytogenetics and evolution. In Mattoo, A. K., editor, *Genetic improvement of solanaceous crops, volume 2*, 77–112. Science Publishers, Enfield, USA.

Clark, W. T. and Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. *Proteins*, 79(7):2086–2096.

Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H., and Stein, L. (2005). Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics*, 21(12):673–682.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

Dalquen, D. A. and Dessimoz, C. (2013). Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biology and Evolution*, 5(10):1800–1806.

Datema, E., Mueller, L. A., Buels, R., Giovannoni, J. J., Visser, R. G. F., Stiekema, W. J., and van Ham, R. C. G. J. (2008). Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. *BMC Plant Biology*, 8:34.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning*, 233–240. ACM Press, New York, USA.

De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N., and Inzé, D. (2012). CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytologist*, 195(3):707–720.

de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

de Jong, J. H., Wolters, A. M. A., Kok, J. M., Verhaar, H., and van Eden, J. (1993). Chromosome pairing and potential for intergeneric recombination in some hypotetraploid somatic hybrids of *Lycopersicon esculentum* (+) *Solanum tuberosum*. *Genome*, 36(6):1032–1041.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8):2898–2903.

De Smet, R. and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology*, 15(2):168–176.

Dessimoz, C., Škunca, N., and Thomas, P. D. (2013). CAFA and the Open World of protein function predictions. *Trends in Genetics*, 29(11):609–610.

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B. et al. (2012). The UniProt-GO Annotation database in 2011. *Nucleic Acids Research*, 40(Database issue):D565–570.

Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N., and Tanksley, S. D. (2002a). A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. *Genetics*, 161(4):1697–1711.

Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N., and Tanksley, S. D. (2002b). Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. *Genetics*, 161(4):1713–1726.

du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C. et al. (2008). PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research*, 36(Database issue):D959–D965.

Eggen, A. (2012). The development and application of genomic selection as a new breeding paradigm. *Animal Frontiers*, 2(1):10–15.

Ehrlich, J., Sankoff, D., and Nadeau, J. H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–296.

Eisen, J. A., Sweder, K. S., and Hanawalt, P. C. (1995). Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Research*, 23(14):2715–2723.

Engelhardt, B. E., Jordan, M. I., Srouji, J. R., and Brenner, S. E. (2011). Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Research*, 21(11):1969–1980.

Erwin, D. H. and Davidson, E. H. (2009). The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10(2):141–148.

Falda, M., Toppo, S., Pescarolo, A., Lavezzo, E., Di Camillo, B., Facchinetti, A., Cilia, E. et al. (2012). Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics*, 13 Suppl 4(Suppl 4):S14.

Fan, H.-Y., Trotter, K. W., Archer, T. K., and Kingston, R. E. (2005). Swapping function of two chromatin remodeling complexes. *Molecular Cell*, 17(6):805–815.

Feuillet, C., Leach, J. E., Rogers, J., Schnable, P. S., and Eversole, K. (2010). Crop genome sequencing: lessons and rationales. *Trends in Plant Science*, 16(2):77–88.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39 Suppl 2:W29–37.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L. et al. (2010). The Pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–222.

Flaus, A., Martin, D. M. A., Barton, G. J., and Owen-Hughes, T. (2006). Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Research*, 34(10):2887–2905.

Flaus, A. and Owen-Hughes, T. (2011). Mechanisms for ATP-dependent chromatin remodeling: the means to the end. *FEBS Journal*, 278(19):3579–3595.

Florea, L., Souvorov, A., Kalbfleisch, T. S., and Salzberg, S. L. (2011). Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS ONE*, 6(6):e21400.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., de Peer, Y. et al. (2008). Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics*, 3(2):87–97.

Foolad, M. (2007). Current status of breeding tomatoes for salt and drought tolerance. In Jenks, M., Hasegawa, P., and Jain, S., editors, *Advances in Molecular Breeding Toward Drought and Salt Tolerant Crops*, 669–700. Springer Netherlands, Dordrecht, NL.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J. et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue):D808–D815.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242.

Fukushima, A., Nishizawa, T., Hayakumo, M., Hikosaka, S., Saito, K., Goto, E., and Kusano, M. (2012). Exploring tomato gene functions based on coexpression modules using graph clustering and differential coexpression approaches. *Plant Physiology*, 158(4):1487–1502.

Fulton, T. M. (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, 14(7):1457–1467.

Fulton, T. M., Nelson, J. C., and Tanksley, S. D. (1997). Introgression and DNA marker analysis of *Lycopersicon peruvianum*, a wild relative of the cultivated tomato, into *Lycopersicon esculentum* , followed through three successive backcross generations. *Theoretical and Applied Genetics*, 95(5-6):895–902.

Furnham, N., de Beer, T. A. P., and Thornton, J. M. (2012). Current challenges in genome annotation through structural biology and bioinformatics. *Current Opinion in Structural Biology*, 22(5):594–601.

Galperin, M. Y. and Koonin, E. V. (2010). From complete genome sequence to »complete« understanding? *Trends in Biotechnology*, 28(8):398–406.

Ganal, M. W., Lapitan, N. L., and Tanksley, S. D. (1991). Macrostructure of the tomato telomeres. *Plant Cell*, 3(1):87–94.

Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J., and Anderson, L. K. (2007). Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics*, 8(1):77–84.

Gendler, K., Paulsen, T., and Napoli, C. (2008). ChromDB: the chromatin database. *Nucleic Acids Research*, 36(Database issue):D298–302.

Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Gerdes, S., El Yacoubi, B., Bailly, M., Blaby, I. K., Blaby-Haas, C. E., Jeanguenin, L., Lara-Núñez, A. et al. (2011). Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genomics*, 12 Suppl 1(Suppl 1):S2.

Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S., and Ouzounis, C. A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641–1649.

Gillis, J. and Pavlidis, P. (2012). »Guilt-by-Association« is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444.

Gillis, J. and Pavlidis, P. (2013). Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics*, 14(Suppl 3):S15.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T. et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(Database issue):D1178–1186.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X. et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652.

Grandillo, S., Ku, H. M., and Tanksley, S. D. (1999). Identifying the loci responsible for natural variation in fruit size and shape in tomato. *Theoretical and Applied Genetics*, 99(6):978–987.

Gray, Y. H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics*, 16(10):461–468.

Grube, R. C., Radwanski, E. R., and Jahn, M. (2000). Comparative genetics of disease resistance within the Solanaceae. *Genetics*, 155(2):873–887.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Guo, H., Lee, T.-H., Wang, X., and Paterson, A. H. (2013). Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. *Plant Physiology*, 162(2):769–778.

Guo, Y.-L. (2013). Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *Plant Journal*, 73(6):941–951.

Haas, B. J., Delcher, A. L., Wortman, J. R., and Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O. et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, 9(1):R7.

Hagen, J. (2003). The statistical frame of mind in systematic biology from quantitative zoology to biometry. *Journal of the History of Biology*, 36(2):353–384.

Hale, C. J., Stonaker, J. L., Gross, S. M., and Hollick, J. B. (2007). A novel Snf2 protein maintains trans-generational regulatory states established by paramutation in maize. *PLoS Biology*, 5(10):e275.

Hamilton, J. P. and Buell, C. R. (2012). Advances in plant genome sequencing. *Plant Journal*, 70(1):177–190.

Han, M. V. and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131.

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2012). Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS ONE*, 7(3):e33071.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. PhD thesis, Pennsylvania State University.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The elements of statistical learning*. Springer, New York, USA, corrected edition.

Hauk, G., McKnight, J. N., Nodelman, I. M., and Bowman, G. D. (2010). The chromodomains of the Chd1 chromatin remodeler regulate DNA access to the ATPase motor. *Molecular Cell*, 39(5):711–723.

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2(6):e88.

Hoffmann, A. A., Sgrò, C. M., and Weeks, A. R. (2004). Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution*, 19(9):482–488.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, 7(3):e1002021.

Holt, C. and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12:491.

Hu, P., Jiang, H., and Emili, A. (2010). Predicting protein functions by relaxation labelling protein interaction network. *BMC Bioinformatics*, 11 Suppl 1:S64.

Huang, J. T. and Dooner, H. K. (2008). Macrotransposition and other complex chromosomal restructuring in maize by closely linked transposons in direct orientation. *Plant Cell*, 20(8):2019–2032.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python environment for tree exploration. *BMC Bioinformatics*, 11:24.

Huettel, B., Kanno, T., Daxinger, L., Bucher, E., van der Winden, J., Matzke, A. J. M., and Matzke, M. (2007). RNA-directed DNA methylation mediated by DRD1 and Pol IVb: a versatile pathway for transcriptional gene silencing in plants. *Biochimica et Biophysica Acta*, 1769(5-6):358–374.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J. et al. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44(7):808–811.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T. et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(Database issue):D306–312.

Hurley, B. A., Tran, H. T., Marty, N. J., Park, J., Snedden, W. A., Mullen, R. T., and Plaxton, W. C. (2010). The dual-targeted purple acid phosphatase isozyme AtPAP26 is essential for efficient acclimation of Arabidopsis to nutritional phosphate deprivation. *Plant Physiology*, 153(3):1112–1122.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8:460.

Huson, D. H. and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, 61(6):1061–1067.

Huttenhower, C., Hibbs, M. A., Myers, C. L., Caudy, A. A., Hess, D. C., and Troyanskaya, O. G. (2009). The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. *Bioinformatics*, 25(18):2404–2410.

Ideker, T., Dutkowski, J., and Hood, L. (2011). Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144(6):860–863.

Iovene, M., Grzebelus, E., Carputo, D., Jiang, J., and Simon, P. W. (2008). Major cytogenetic landmarks and karyotype analysis in *Daucus carota* and other *Apiaceae*. *American Journal of Botany*, 95(7):793–804.

Itoh, T., Tanaka, T., Barrero, R. A., Yamasaki, C., Fujii, Y., Hilton, P. B., Antonio, B. A. et al. (2007). Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome Research*, 17(2):175–183.

Jackson, S. A., Iwata, A., Lee, S.-H., Schmutz, J., and Shoemaker, R. (2011). Sequencing crop genomes: approaches and applications. *New Phytologist*, 191(4):915–925.

Jahn, M., Paran, I., Hoffmann, K., Radwanski, E. R., Livingstone, K. D., Grube, R. C., Aftergoot, E. et al. (2000). Genetic mapping of the *Tsw* locus for resistance to the *Tospovirus Tomato spotted wilt virus* in *Capsicum* spp. and its relationship to the *Sw-5* gene for resistance to the same pathogen in tomato. *Molecular Plant-Microbe Interactions*, 13(6):673–682.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N. et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A. et al. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.

Jenks, M. A., Hasegawa, P. M., and Jain, S. M., editors (2007). *Advances in molecular breeding toward drought and salt tolerant crops*. Springer Netherlands, Dordrecht, NL.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, 19–33.

Jiang, W.-k., Liu, Y.-l., Xia, E.-h., and Gao, L.-z. (2013). Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. *Plant Physiology*, 161(4):1844–1861.

Jones, C. E., Brown, A. L., and Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8:170.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H. et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Jong, J. H., Zhong, X.-B., Fransz, P. F., Wennekes-van Eden, J., Jacobsen, E., and Zabel, P. (2000). High resolution FISH reveals the molecular and chromosomal organization of repetitive sequences of individual tomato chromosomes. In Olmo, E. and Redi, C., editors, *Chromosomes Today*, 267–275. Birkhäuser Basel, Basel, CH.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467.

Kanno, T., Mette, M. F., Kreil, D. P., Aufsatz, W., Matzke, M., and Matzke, A. J. M. (2004). Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. *Current Biology*, 14(9):801–805.

Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Hughes, D. S. T. et al. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research*, 42(Database issue):D546–D552.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., Seo, E. et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics*, 46(3):270–278.

Knapp, S. (2002). Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *Journal of Experimental Botany*, 53(377):2001–2022.

Knizewski, L., Ginalski, K., and Jerzmanowski, A. (2008). Snf2 proteins in plants: gene silencing and beyond. *Trends in Plant Science*, 13(10):557–565.

Koo, D.-H., Jo, S.-H., Bang, J.-W., Park, H.-M., Lee, S., and Choi, D. (2008). Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2. *Genetics*, 179(3):1211–1220.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39(1):309–338.

Koonin, E. V. and Galperin, M. (2003). *Sequence - evolution - function: computational approaches in comparative genomics.* Kluwer Academic, Boston, USA.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5:59.

Kourmpetis, Y. A., van Dijk, A. D., and Ter Braak, C. J. (2013). Gene Ontology consistent protein function prediction: the FALCON algorithm applied to six eukaryotic genomes. *Algorithms for Molecular Biology*, 8(1):10.

Kourmpetis, Y. A. I., van Dijk, A. D. J., Bink, M. C. A. M., van Ham, R. C. H. J., and ter Braak, C. J. F. (2010). Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS ONE*, 5(2):e9293.

Kourmpetis, Y. A. I., van Dijk, A. D. J., van Ham, R. C. H. J., and ter Braak, C. J. F. (2011). Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. *Plant Physiology*, 155(1):271–281.

Krieger, U., Lippman, Z. B., and Zamir, D. (2010). The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nature Genetics*, 42(5):459–463.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., and Koonin, E. V. (2011). Computational methods for Gene Orthology inference. *Briefings in Bioinformatics*, 12(5):379–391.

Krzywinski, M. and Altman, N. (2014). Points of Significance: Visualizing samples with box plots. *Nature Methods*, 11(2):119–120.

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the*

National Academy of Sciences of the United States of America, 97(16):9121–9126.

Kurata, N. and Yamazaki, Y. (2006). Oryzabase. An integrated biological and genome information database for rice. *Plant Physiology*, 140(1):12–17.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S., and Leunissen, J. A. M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z. et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, 42(11):1027–1030.

Lall, S. (2011). A bottle opener for TBP. *Nature Structural & Molecular Biology*, 18(8):865.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W. et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, 42(12):1053–1059.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R. et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(Database issue):D1202–D1210.

Law, J. A., Ausin, I., Johnson, L. M., Vashisht, A. A., Zhu, J.-K., Wohlschlegel, J. A., and Jacobsen, S. E. (2010). A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Current Biology*, 20(10):951–956.

Law, J. A., Vashisht, A. A., Wohlschlegel, J. A., and Jacobsen, S. E. (2011). SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genetics*, 7(7):e1002195.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995–1005.

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., Stein, L. et al. (2013a). Web Apollo: a web-based genomic annotation editing platform. *Genome Biology*, 14(8):R93.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature Biotechnology*, 28(2):149–156.

Lee, I., Seo, Y.-S., Coltrane, D., Hwang, S., Oh, T., Marcotte, E. M., and Ronald, P. C. (2011). Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18548–18553.

Lee, T.-H., Tang, H., Wang, X., and Paterson, A. H. (2013b). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research*, 41(Database issue):D1152–D1158.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Research*, 37(Database issue):D229–232.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

Li, X. Y., Wang, C., Nie, P. P., Lu, X. W., Wang, M., Liu, W., Yao, J. et al. (2011). Characterization and expression analysis of the SNF2 family genes in response to phytohormones and abiotic stresses in rice. *Biologia Plantarum*, 55(4):625–633.

Liharska, T., van Wordragen, M., van Kammen, A., Zabel, P., and Koornneef, M. (1996). Tomato chromosome 6: effect of alien chromosomal segments on recombinant frequencies. *Genome*, 39(3):485–491.

Liu, Y., He, Z., Appels, R., and Xia, X. (2012). Functional markers in wheat: current status and future prospects. *Theoretical and Applied Genetics*, 125(1):1–10.

Livingstone, K. and Rieseberg, L. (2004). Chromosomal evolution and speciation: a recombination-based approach. *New Phytologist*, 161(1):107–112.

Livingstone, K. D., Lackney, V. K., Blauth, J. R., van Wijk, R., and Jahn, M. K. (1999). Genome mapping in *Capsicum* and the evolution of genome structure in the Solanaceae. *Genetics*, 152(3):1183–1202.

Lloyd, J. and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. *Plant Physiology*, 158(3):1115–1129.

López, A., Ramírez, V., García-Andrade, J., Flors, V., and Vera, P. (2011). The RNA silencing enzyme RNA polymerase V is required for plant immunity. *PLoS Genetics*, 7(12):e1002434.

Lottaz, C., Iseli, C., Jongeneel, C. V., and Bucher, P. (2003). Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, 19 Suppl 2:ii103–112.

Løvdal, T. and Lillo, C. (2009). Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Analytical Biochemistry*, 387(2):238–242.

Loveland, J. E., Gilbert, J. G. R., Griffiths, E., and Harrow, J. L. (2012). Community gene annotation in practice. *Database*, 2012:bas009.

Lupski, J. R., Belmont, J. W., Boerwinkle, E., and Gibbs, R. A. (2011). Clan genomics and the complex architecture of human disease. *Cell*, 147(1):32–43.

Lyons, E. and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal*, 53(4):661–673.

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related *Brassicaceae* species. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):5224–5229.

Mahfouz, M. M. (2010). RNA-directed DNA methylation: mechanisms and functions. *Plant Signaling & Behavior*, 5(7):806–816.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.

Marguerat, S. and Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 67(4):569–579.

Martin, D. M. A., Berriman, M., and Barton, G. J. (2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178.

Martinez, M. (2011). Plant protein-coding gene families: emerging bioinformatics approaches. *Trends in Plant Science*, 16(10):558–567.

Martinez, M. (2013). From plant genomes to protein families: computational tools. *Computational and Structural Biotechnology Journal*, 8:e201307001.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A. J. M. (2006). RNA-directed DNA methylation and Pol IVb in Arabidopsis. *Cold Spring Harbor Symposia on Quantitative Biology*, 71:449–459.

McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6544–6549.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

Mlynárová, L., Nap, J. P., and Bisseling, T. (2007). The SWI/SNF chromatin-remodeling gene AtCHR12 mediates temporary growth arrest in *Arabidopsis thaliana* upon perceiving environmental stress. *Plant Journal*, 51(5):874–885.

Monaco, M. K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S., Amarasinghe, V. et al. (2014). Gramene 2013: comparative plant genomics resources. *Nucleic Acids Research*, 42(Database issue):D1193–D1199.

Monclus, R., Leplé, J.-C., Bastien, C., Bert, P.-F., Villar, M., Marron, N., Brignolas, F. et al. (2012). Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in Populus spp. *BMC Plant Biology*, 12:173.

Morrell, P. L., Buckler, E. S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. *Nature Reviews Genetics*, 13(2):85–96.

Mossa, S., Barthélémy, M., Eugene Stanley, H., and Nunes Amaral, L. (2002). Truncation of power law behavior in »scale-free« network models due to information filtering. *Physical Review Letters*, 88(13):138701.

Moyle, L. C. (2008). Ecological and evolutionary genomics in the wild tomatoes (*Solanum* sect. *Lycopersicon*). *Evolution*, 62(12):2995–3013.

Moyle, L. C. and Graham, E. B. (2006). Genome-wide associations between hybrid sterility QTL and marker transmission ratio distortion. *Molecular Biology and Evolution*, 23(5):973–980.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R. et al. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*, 23(3):895–910.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–i310.

Nadeau, J. H. (1989). Maps of linkage and synteny homologies between mouse and man. *Trends in Genetics*, 5:82–86.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., Cardeno, C. et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(3):R59.

Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, 7(6):e1002073.

Nguyen Ba, A. N., Yeh, B. J., van Dyk, D., Davidson, A. R., Andrews, B. J., Weiss, E. L., and Moses, A. M. (2012). Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science Signaling*, 5(215):rs1.

Noor, M. A., Grams, K. L., Bertucci, L. A., and Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21):12084–12088.

O'Driscoll, A., Daugelaite, J., and Sleator, R. D. (2013). »Big data«, Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, 46(5):774–781.

Oliver, S. (2000). Guilt-by-association goes global. *Nature*, 403(6770):601–603.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H. et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(Database issue):D358–D363.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F. et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research*, 35(Database issue):D883–887.

Ouzounis, C. A. (2012). Rise and demise of bioinformatics? Promise and progress. *PLoS Computational Biology*, 8(4):e1002487.

Pan, X., Stein, L., and Brendel, V. (2005). SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17):3461–3468.

Parra, G., Blanco, E., and Guigó, R. (2000). GeneID in Drosophila. *Genome Research*, 10(4):511–515.

Passarge, E., Horsthemke, B., and Farber, R. A. (1999). Incorrect use of the term synteny. *Nature Genetics*, 23(4):387.

Paterson, A. H., Freeling, M., and Sasaki, T. (2005). Grains of knowledge: genomics of model cereals. *Genome Research*, 15(12):1643–1650.

Pavlidis, P. and Gillis, J. (2012). Progress and challenges in the computational prediction of gene function using networks. *F1000 Research*, 1:14.

Pertuzé, R. A., Ji, Y., and Chetelat, R. T. (2002). Comparative linkage map of the *Solanum lycopersicoides* and *S. sitiens* genomes and their differentiation from tomato. *Genome*, 45(6):1003–1012.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443.

Peters, S. A., Datema, E., Szinay, D., van Staveren, M. J., Schijlen, E. G. W. M., van Haarst, J. C., Hesselink, T. et al. (2009). *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant Journal*, 58(5):857–869.

Pevzner, P. and Tesler, G. (2003). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, 13(1):37–45.

Pikaard, C. S., Haag, J. R., Ream, T., and Wierzbicki, A. T. (2008). Roles of RNA polymerase IV in gene silencing. *Trends in Plant Science*, 13(7):390–397.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I. et al. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16):3435–3438.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(Database issue):D130–135.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–504.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N. et al. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40(Database issue):D290–301.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K. et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227.

Ranjan, A., Ichihashi, Y., and Sinha, N. R. (2012). The tomato genome: implications for plant breeding, genomics and evolution. *Genome Biology*, 13(8):167.

Raser, J. M. and O'Shea, E. K. (2005). Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013.

Rautengarten, C., Usadel, B., Neumetzler, L., Hartmann, J., Büssis, D., and Altmann, T. (2008). A subtilisin-like serine protease essential for mucilage release from Arabidopsis seed coats. *Plant Journal*, 54(3):466–480.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052.

Rentzsch, R. and Orengo, C. A. (2009). Protein function prediction–the power of multiplicity. *Trends in Biotechnology*, 27(4):210–219.

Rhee, S. Y. and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19(4):212–221.

Rick, C., DeVerna, J., Chetelat, R., and Stevens, M. (1987). Potential contributions of wide crosses to improvement of processing tomatoes. *Acta Horticulturae*, 200:45–56.

Rieseberg, L. H. and Willis, J. H. (2007). Plant speciation. *Science*, 317(5840):910–914.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Ryan, C. J., Cimermančič, P., Szpiech, Z. A., Sali, A., Hernandez, R. D., and Krogan, N. J. (2013). High-resolution network biology: connecting sequence with function. *Nature Reviews Genetics*, 14(12):865–879.

Sakurai, T., Kondou, Y., Akiyama, K., Kurotani, A., Higuchi, M., Ichikawa, T., Kuroda, H. et al. (2011). RiceFOX: a database of Arabidopsis mutant lines overexpressing rice full-length cDNA that contains a wide range of trait information to facilitate analysis of gene function. *Plant & Cell Physiology*, 52(2):265–273.

Sang, Y., Silva-Ortega, C. O., Wu, S., Yamaguchi, N., Wu, M.-F., Pfluger, J., Gillmor, C. S. et al. (2012). Mutations in two non-canonical Arabidopsis SWI2/SNF2 chromatin remodeling ATPases cause embryogenesis and stem cell maintenance defects. *Plant Journal*, 72(6):1000–1014.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T. et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology*, 12(8):125.

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33(Web Server issue):W686–689.

Schatz, M. C., Witkowski, J., and McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome Biology*, 13(4):243.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B. et al. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*, 37(5):501–506.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A. et al. (2013). Patterns of population epigenomic diversity. *Nature*, 495(7440):193–198.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L. et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–183.

Schneider, M. V. and Jungck, J. R. (2013). Editorial: International, interdisciplinary, multi-level bioinformatics training and education. *Briefings in Bioinformatics*, 14(5):527.

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12):e1000605.

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Computational Biology*, 9(5):e1003063.

Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Ong, C. S., Philips, P. et al. (2009). mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 19(11):2133–2143.

Seah, S., Yaghoobi, J., Rossi, M., Gleason, C. A., and Williamson, V. M. (2004). The nematode-resistance gene, *Mi-1*, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. *Theoretical and Applied Genetics*, 108(8):1635–1642.

Searls, D. B. (2010). The roots of bioinformatics. *PLoS Computational Biology*, 6(6):e1000809.

Sesso, H. D., Liu, S., Gaziano, J. M., and Buring, J. E. (2003). Dietary lycopene, tomato-based food products and cardiovascular disease in women. *Journal of Nutrition*, 133(7):2336–2341.

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., Muehlbauer, G. J. et al. (2010). RNA-seq atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biology*, 10:160.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, 3:88.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R. et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7:539.

Simillion, C., Vandepoele, K., Van Montagu, M. C. E., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13627–13632.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

Singh, R. J., editor (2006). *Genetic resources, chromosome engineering, and crop improvement: vegetable crops, volume 3*. CRC Press, Boca Raton, USA.

Skunca, N., Altenhoff, A., and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, 8(5):e1002533.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

Spooner, D. M., Anderson, G. J., and Jansen, R. K. (1993). Chloroplast DNA evidence for the interrelationships of tomatoes, potatoes, and pepinos (Solanaceae). *American Journal of Botany*, 80(6):676–688.

Spooner, D. M., Peralta, I. E., and Knapp, S. (2005). Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. *Taxon*, 54(1):43–61.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G. et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, 57(5):758–771.

Stanke, M. and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2):ii215–ii225.

Sterck, L., Billiau, K., Abeel, T., Rouzé, P., and Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. *Nature Methods*, 9(11):1041.

Stojanovic, N., editor (2007). *Computational genomics: current methods*. Horizon Bioscience, Poole, UK.

Subbaiyan, G. K., Waters, D. L. E., Katiyar, S. K., Sadananda, A. R., Vaddadi, S., and Henry, R. J. (2012). Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. *Plant Biotechnology Journal*, 10(6):623–634.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE*, 6(7):e21800.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288.

Szinay, D., Chang, S.-B., Khrustaleva, L., Peters, S., Schijlen, E., Bai, Y., Stiekema, W. J. et al. (2008). High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. *Plant Journal*, 56(4):627–637.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T. et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):D561–D568.

Tanaka, T., Antonio, B. A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H. et al. (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Research*, 36(Database issue):D1028–1033.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008a). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.

Tang, X., Szinay, D., Lang, C., Ramanna, M. S., van der Vossen, E. A. G., Datema, E., Lankhorst, R. K. et al. (2008b). Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics*, 180(3):1319–1328.

Tanksley, S. D., Bernatzky, R., Lapitan, N. L., and Prince, J. P. (1988). Conservation of gene repertoire but not gene order in pepper and tomato. *Proceedings of the National Academy of Sciences of the United States of America*, 85(17):6419–6423.

Tanksley, S. D., Ganal, M. W., Prince, J. P., de Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M. et al. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics*, 132(4):1141–1160.

Tatusov, R. L. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1):22–28.

Tesler, G. (2002). GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492–493.

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(Database issue):D71–D75.

Thorup, T. A., Tanyolac, B., Livingstone, K. D., Popovsky, S., Paran, I., and Jahn, M. (2000). Candidate gene analysis of organ pigmentation loci in the Solanaceae. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11192–11197.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L. et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N. et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604.

Tzfadia, O., Amar, D., Bradbury, L. M. T., Wurtzel, E. T., and Shamir, R. (2012). The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways. *Plant Cell*, 24(11):4389–4406.

UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(1):D191–D198.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, 35(Web Server issue):W71–74.

Uversky, V. N. and Dunker, A. K. (2010). Understanding protein non-folding. *Biochimica et Biophysica Acta*, 1804(6):1231–1264.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiology*, 158(2):590–600.

Van de Peer, Y. (2004). Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics*, 5(10):752–763.

Van der Hoeven, R. (2002). Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell*, 14(7):1441–1456.

van der Knaap, E. and Tanksley, S. D. (2003). The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato. *Theoretical and Applied Genetics*, 107(1):139–147.

Vandepoele, K. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Research*, 12(11):1792–1801.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700.

Vital-Lopez, F. G., Memišević, V., and Dutta, B. (2012). Tutorial on biological networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4):298–325.

Vos, R. A., Caravas, J., Hartmann, K., Jensen, M. A., and Miller, C. (2011). BIO::Phylo-phyloinformatic analysis using Perl. *BMC Bioinformatics*, 12:63.

Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Research*, 9(12):1198–1203.

Walley, J. W., Rowe, H. C., Xiao, Y., Chehab, E. W., Kliebenstein, D. J., Wagner, D., and Dehesh, K. (2008). The chromatin remodeler SPLAYED regulates specific stress signaling pathways. *PLoS Pathogens*, 4(12):e1000237.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281.

Wang, Y., Diehl, A., Wu, F., Vrebalov, J., Giovannoni, J., Siepel, A., and Tanksley, S. D. (2008). Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. *Genetics*, 180(1):391–408.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-h. et al. (2012a). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7):e49.

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., Yang, L. et al. (2012b). The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. *Plant Journal*, 72(3):461–473.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337(3):635–645.

Widmer, A., Lexer, C., and Cozzolino, S. (2009). Evolution of reproductive isolation in plants. *Heredity*, 102(1):31–38.

Wierzbicki, A. T., Haag, J. R., and Pikaard, C. S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, 135(4):635–648.

Winterbach, W., Mieghem, P. V., Reinders, M., Wang, H., and de Ridder, D. (2013). Topology of molecular interaction networks. *BMC Systems Biology*, 7(1):90.

Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V., and Tanksley, S. D. (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics*, 174(3):1407–1420.

Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881.

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.

Xiang, D., Venglat, P., Tibiche, C., Yang, H., Risseeuw, E., Cao, Y., Babic, V. et al. (2011). Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in Arabidopsis. *Plant Physiology*, 156(1):346–356.

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., Dong, Y. et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, 30(1):105–111.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G. et al. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189–195.

Xu, Y. (2009). *Molecular plant breeding*. CAB International, Oxfordshire, UK.

Xu, Z. and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35(Web Server issue):W265–W268.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329–342.

Yang, H.-B., Liu, W. Y., Kang, W.-H., Jahn, M., and Kang, B.-C. (2009). Development of SNP markers linked to the *L* locus in *Capsicum* spp. by a comparative genetic analysis. *Molecular Breeding*, 24(4):433–446.

Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., Derwent, P., Dharmawardhana, P. et al. (2011). Gramene database in 2010: updates and extensions. *Nucleic Acids Research*, 39(Database issue):D1085–D1094.

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A. et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524.

Yu, D., Kim, M., Xiao, G., and Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics & Informatics*, 11(4):200–210.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

Zdobnov, E. M. and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848.

Zhang, J., Yu, C., Pulletikurti, V., Lamb, J., Danilova, T., Weber, D. F., Birchler, J. et al. (2009). Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize. *Genes & Development*, 23(6):755–765.

Zheng, X. H., Lu, F., Wang, Z.-Y. Z.-Y., Zhong, F., Hoover, J., and Mural, R. (2004). Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*, 21(6):703–710.

Zhi, D., Raphael, B. J., Price, A. L., Tang, H., and Pevzner, P. A. (2006). Identifying repeat domains in large genomes. *Genome Biology*, 7(1):R7.

Zhong, X.-B., Jong, J. H., and Zabel, P. (1996). Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence in situ hybridization (FISH). *Chromosome Research*, 4(1):24–28.

Zhu, L., You, Z.-H., and Huang, D.-S. (2013). Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. *Neurocomputing*, 121:99–107.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024.

Ziolkowski, P. A. (2003). Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome. *Nucleic Acids Research*, 31(4):1339–1350.

Summary

The research presented in this thesis focuses on deriving function from sequence information, with the emphasis on plant sequence data. Unravelling the impact of genomic elements, in most cases genes, on the phenotype of an organism is a major challenge in biological research and modern plant breeding. An important part of this challenge is the (functional) annotation of such genomic elements. Currently, wet lab experiments may provide high quality, but they are laborious and costly. With the advent of next generation sequencing platforms, vast amounts of sequence data are generated. This data are used in connection with the available experimental data to derive function from a bioinformatics perspective.

The connection between sequence information and function was approached on the level of chromosome structure (chapter 2) and of gene families (chapter 3) using combinations of existing bioinformatics tools. The applicability of using interaction networks for function prediction was demonstrated by first markedly improving an existing method (chapter 4) and by exploring the role of network topology in function prediction (chapter 5). Taken together, the combination of methods and results presented indicate the potential as well as the current state-of-the-art of function prediction in (plant) bioinformatics.

Chapter 1 introduces the basis for the approaches used and developed in this thesis. This includes the concepts of genome annotation, comparative genomics, gene function prediction and the analysis of network topology for gene function prediction. A requirement for the study of any new organism is the sequencing and annotation of its genome. Current genome annotation is divided into structural identification and functional categorization of genomic elements. The de facto standard for categorizing functional annotation is provided by the Gene Ontology. The Gene Ontology is divided into three domains, molecular function, biological process and cellular component. Approaches to predict molecular function and biological process are outlined. Accurate function prediction generally relies on existing input data, often of experimental origin, that can be transferred to unannotated genomic elements. Plants often lack such input data, which poses a big challenge for current function prediction algorithms. In unravelling the function of genomic elements, comparative genomics is an important approach. Via the comparison of multiple genomes it gives insights into evolution, function as well as genomic

structure and variation. Comparative genomics has become an essential toolkit for the analysis of newly sequenced organisms. Often bioinformatics methods need to be adapted to the specific needs of plant genome research. With a focus on the commercially important crop plants tomato and potato, specific requirements of plant bioinformatics, such as the high amount of repetitive elements and the lack of experimental data, are outlined.

In chapter 2, the structural homology of the long arm of chromosome 2 (2L) of tomato, potato and pepper is analyzed. Molecular organization and collinear junctions are delineated using multi-color BAC FISH analysis and comparative sequence alignment. We identify several large-scale rearrangements including inversions and segmental translocations that were not reported in previous comparative studies. Some of the structural rearrangements are specific for the tomato clade, and differentiate tomato from potato, pepper and other solanaceous species. There are many small-scale synteny perturbations, but local gene vicinity is largely preserved. The data suggests that long distance intra-chromosomal rearrangements and local gene rearrangements have evolved frequently during speciation in the *Solanum* genus, and that small changes are more prevalent than large-scale differences. The occurrence of transposable elements and other repeats near or at junction breaks may indicate repeat-mediated rearrangements. The ancestral 2L topology is reconstructed and the evolutionary events leading to the current topology are discussed.

In chapter 3, we analyze the Snf2 gene family. As part of large protein complexes, Snf2 family ATPases are responsible for energy supply during chromatin remodeling, but the precise mechanism of action of many of these proteins is largely unknown. They influence many processes in plants, such as the response to environmental stress. The analysis is the first comprehensive study of Snf2 family ATPases in plants. Some subfamilies of the Snf2 gene family are remarkably stable in number of genes per genome, whereas others show expansion and contraction in several plants. One of these subfamilies, the plant-specific DRD1 subfamily, is non-existent in lower eukaryote genomes, yet it developed into the largest Snf2 subfamily in plant genomes. It shows the occurrence of a complex series of evolutionary events. Its expansion, notably in tomato, suggests novel functionality in processes connected to chromatin remodeling. The results underpin and extend the Snf2 subfamily classification, which could help to determine the various functional roles of Snf2 ATPases and to target environmental stress tolerance and yield in future breeding with these genes.

In chapter 4, a new approach to improve the prediction of protein function in terms of biological processes is developed that is particularly attractive for sparsely annotated plant genomes. The combination of the network-based prediction method Bayesian Markov Random Field (BMRF) with the sequence-based prediction method Argot2 shows significantly improved performance compared to each of the methods separately, as well as compared to Blast2GO. The approach was applied to predict biological processes for the proteomes of rice, barrel clover, poplar, soybean and tomato. Analysis of the relationships between sequence sim-

ilarity and predicted function similarity identifies numerous cases of divergence of biological processes in which proteins are involved, in spite of sequence similarity. Examples of potential divergence are identified for various biological processes, notably for processes related to cell development, regulation, and response to chemical stimulus. Such divergence in biological process annotation for proteins with similar sequences should be taken into account when analyzing plant gene and genome evolution. This way, the integration of network-based and sequence-based function prediction will strengthen the analysis of evolutionary relationships of plant genomes.

In chapter 5 the influence of network topology on network-based function prediction algorithms is investigated. The analysis of biological networks using algorithms such as Bayesian Markov Random Field (BMRF) is a valuable predictor of the biological processes that proteins are involved in. The topological properties and constraints that determine prediction performance in such networks are however largely unknown. This chapter presents analyses based on network centrality measures, such as node degree, to evaluate the performance of BMRF upon progressive removal of highly connected hub nodes (pruning). Three different protein-protein interaction networks with data from Arabidopsis, human and yeast were analyzed. All three show that the average prediction performance can improve significantly. The chapter paves the way for further improvement of network-based function prediction methods based on node pruning.

Chapter 6 discusses the results and methods developed in this thesis in the context of the vast amount of generated sequencing data. Sequencing or resequencing a (plant) genome has become fairly straightforward and affordable, but the interpretation for subsequent use of this sequence data is far from trivial. The topics addressed in this thesis, annotation of function, analysis of genome structure and identifying genomic variation, focus on this main bottleneck of biological research. Issues discussed in connection with this work and its future are data accuracy, error propagation, possible improvements and future implications for biological research in crop plants. In particular the shift of costs from sequencing to downstream analyses, with functional genome annotation as essential step, is covered. One of the biggest challenges biology and bioinformatics will face is the integration of results from such downstream analyses and other sources into a complete picture. Only this will allow understanding of complex biological systems.

Samenvatting

Het onderzoek dat in dit proefschrift beschreven wordt, richt zich op het bepalen van de functie van sequentie-informatie, met de nadruk op sequentiedata van planten. Het ontrafelen van de rol en werking van elementen uit het genoom, doorgaans genen, op het fenotype van een organisme is nog steeds een grote uitdaging in biologisch onderzoek en voor moderne plantenveredeling. Een belangrijk onderdeel van deze uitdaging is de (functionele) annotatie van dergelijke genoomelementen. Experimenten in het laboratorium (»natte experimenten«) kunnen hoge kwaliteit leveren, maar zijn doorgaans tijdrovend en kostbaar. Door de opkomst van moderne methoden om sequenties te bepalen worden enorme hoeveelheden sequentiedata gegenereerd. Deze gegevens worden samen met de beschikbare experimentele gegevens gebruikt om functies af te leiden en te voorspellen met behulp van bioinformatica. Met combinaties van bestaande bioinformatica methoden worden in dit proefschrift de relaties tussen sequentie-informatie en functie onderzocht op de niveaus van chromosoomstructuur (hoofdstuk 2) en genfamilies (hoofdstuk 3). De bruikbaarheid van interactienetwerken voor het voorspellen van functies wordt gedemonstreerd door eerst een bestaande methode aanzienlijk te verbeteren (hoofdstuk 4) en vervolgens door het analyseren van de rol van de topologie van de netwerken voor dergelijke functievoorspellingen (hoofdstuk 5). De combinaties van methoden en resultaten als hier gepresenteerd geven zicht op de mogelijkheden alsook de huidige stand van ontwikkeling van de voorspelling van functies met behulp van (op planten gerichte) bioinformatica.

Hoofdstuk 1 introduceert de achtergrond voor de benaderingen die in dit proefschrift worden toegepast en ontwikkeld. Dit omvat de begrippen genoomannotatie, vergelijkende genoomanalyse, voorspelling van de functie van genen en de analyse van netwerktopologie voor gen-functie voorspelling. Voorwaarde voor dit type onderzoek aan een nieuw organisme is de sequentie en annotatie van het genoom van dat organisme. Genoomannotatie wordt momenteel onderverdeeld in structurele identificatie en functionele indeling van genoomelementen. De de facto standaard voor het indelen van functionele annotatie is de gen-ontologie. Deze is verdeeld in drie domeinen: moleculaire functie, biologisch proces en cellulaire component. Methoden worden besproken waarmee de domeinen moleculaire functie en biologisch proces voorspeld kunnen worden. Nauwkeurige functievoorspelling

hangt over het algemeen af van de beschikbaarheid van data van doorgaans experimentele oorsprong, die verbonden kunnen worden aan genoomelementen zonder annotatie. Voor planten zijn dergelijke data vaak nog niet beschikbaar. Dit is een grote uitdaging voor de huidige algoritmen die functies voorspellen. Voor het ontrafelen van de functie van genoomelementen is vergelijkende genoomanalyse belangrijk. Het vergelijken van verschillende genomen geeft inzicht in evolutie en functie, alsook in genoomstructuur en variatie. Vergelijkende genoomanalyse is een essentieel instrument voor de analyse van nieuw gesequencete organismen. Bioinformatica methoden moeten meestal worden aangepast aan de specifieke karakteristieken van het genoomonderzoek aan planten. Specifieke kenmerken voor de bioinformatica van planten, met focus op de commercieel belangrijke gewassen tomaat en aardappel, zijn het goed kunnen omgaan met de grote hoeveelheden repetitieve elementen en met het niet beschikbaar zijn van experimentele data.

In hoofdstuk 2 wordt de structurele homologie van de lange arm van chromosoom 2 (2L) van tomaat, aardappel en peper geanalyseerd. Moleculaire organisatie en co-lineaire chromosomale verbindingen worden geanalyseerd met behulp van meer-kleuren BAC FISH en gedetailleerde sequentievergelijkingen. We identificeren diverse grote herschikkingen, met inbegrip van inversies en grote translocaties, die niet werden beschreven in eerdere studies met vergelijkingen tussen genomen. Sommige structurele herschikkingen zijn specifiek voor de tomatengroep, en onderscheiden tomaat van aardappel, peper en andere soorten van de nachtschadefamilie (Solanaceae). Er zijn veel kleine verschillen in syntenie, maar de lokale gen-omgeving is grotendeels bewaard gebleven. De gegevens suggereren dat tijdens de soortvorming in het geslacht *Solanum* regelmatig zowel intra-chromosomale herschikkingen op grote afstand, als lokale herschikkingen van genen plaats hebben gevonden. Kleine veranderingen komen vaker voor dan grote verschillen. De aanwezigheid van transposons en andere repetitieve sequenties op of dichtbij chromosomale breekpunten kan erop duiden dat de herschikkingen afhankelijk zijn van dergelijke repetitieve sequenties. De 2L topologie van de voorouders is gereconstrueerd en de mogelijke evolutionaire gebeurtenissen die hebben geleid tot de huidige topologie worden besproken.

In hoofdstuk 3 analyseren we de SNF2 gen familie. ATPases van de SNF2 familie zijn onderdeel van grote eiwitcomplexen en verantwoordelijk voor energievoorziening tijdens de reorganisatie (het remodelleren) van chromatine. Het precieze werkingsmechanisme van veel van deze eiwitten is grotendeels onbekend, maar ze beïnvloeden veel processen in planten, zoals de reactie op omgevingsstress. De analyse in hoofdstuk 3 is de eerste uitgebreide studie van SNF2 familie ATPases in planten. Sommige subfamilies van de SNF2 gen-familie hebben een opmerkelijk stabiel aantal genen per genoom, terwijl andere families in verschillende planten juist relatief meer (dus expansie) of juist minder (dus contractie) genen hebben. De plant-specifieke DRD1 subfamilie ontbreekt in lagere eukaryote genomen, maar heeft zich ontwikkeld tot de grootste SNF2 onderfamilie in plantengenomen. Dit duidt op een complexe serie evolutionaire gebeurtenissen. Deze uitbreiding van het aantal genen suggereert dat vooral in tomaat nieuwe biologische functionaliteit is

ontstaan in processen gerelateerd aan chromatine reorganisatie. De resultaten breiden de bestaande indeling van SNF2 subfamilie uit. Dit helpt om de verschillende functionele rollen van SNF2 ATPases te bepalen. Toekomstige veredeling met deze genen kan hiermee bijdragen aan betere tolerantie tegen omgevingsstress en op die manier aan meer opbrengst.

In hoofdstuk 4 wordt een nieuwe aanpak voorgesteld voor een betere voorspelling van de functie van eiwitten in termen van biologische processen. Dit is met name aantrekkelijk voor eiwitten in plantengenomen met weinig annotatie. De combinatie van de op netwerken gebaseerde voorspellingsmethode Bayesiaanse Markov Random Field (BMRF) met de op sequentievergelijkingen gebaseerde voorspellingsmethode Argot2 geeft significant verbeterde voorspellingen vergeleken met elk van de methoden afzonderlijk en ook vergeleken met Blast2GO. De methode is gebruikt om voor alle bekende eiwitten (proteomen) van rijst, rupsklaver, populier, soja en tomaat de betrokkenheid bij biologische processen te voorspellen. Analyse van de overeenkomsten tussen DNA sequentie en voorspelde functie laat talrijke gevallen zien waarin eiwitten bij verschillende biologische processen betrokken zijn terwijl ze qua sequentie erg op elkaar lijken. Voor diverse biologische processen zijn voorbeelden van dergelijke mogelijke divergenties gevonden. Dit betreft vooral processen betrokken bij cel-ontwikkeling, regulatie en reactie op een chemische stimulus. Voor eiwitten met gelijkende sequenties moet daarom in de analyse van de evolutie van plantengenen en -genomen rekening gehouden worden met dergelijke verschillen in de functionele annotatie van biologische processen. Op die manier kan de integratie van netwerk-gebaseerde en sequentie-gebaseerde voorspelling van functie de analyse van evolutionaire verwantschappen van plantengenomen versterken.

In hoofdstuk 5 wordt de invloed van de topologie van het netwerk onderzocht op de werking van op netwerk-gebaseerde algoritmen voor functievoorspelling. De analyse van biologische netwerken met behulp van algoritmen zoals Bayesiaanse Markov Random Field (BMRF) leidt tot waardevolle voorspellingen van de biologische processen waarin eiwitten een rol spelen. De topologische eigenschappen van, en beperkingen in, zulke netwerken die de kwaliteit van dergelijke voorspellingen bepalen zijn echter grotendeels onbekend. Dit hoofdstuk presenteert een analyse op basis van netwerkparameters, zoals knooppuntverknoping, om de prestaties van BMRF te evalueren bij geleidelijke verwijdering van zeer sterk verknoopte knooppunten. Drie verschillende eiwit-eiwit interactienetwerken met data van zandraket, mens en gist zijn geanalyseerd. Alle drie laten zien aan dat gemiddeld de voorspellingen aanzienlijk kunnen verbeteren. Dit kan leiden tot verdere verbetering van netwerk-gebaseerde functievoorspelling op basis van het verwijderen van knooppunten.

Hoofdstuk 6 bespreekt de resultaten en methoden die zijn ontwikkeld in dit proefschrift in de context van de enorme hoeveelheid gegenereerde sequentie data. Het sequencen of her-sequencen een (planten)genoom is al met al simpel en betaalbaar geworden, maar de interpretatie van deze data voor uiteindelijke toepassing is verre van triviaal. De onderwerpen die in dit proefschrift ter sprake komen,

annotatie van functie, analyse van genoomstructuur en het identificeren van genoomvariatie, richten zich allemaal op dit belangrijkste knelpunt van biologisch onderzoek. Kwesties besproken in verband met dit werk en de toekomst daarvan zijn de nauwkeurigheid van de gegevens, foutenpropagatie, mogelijke verbeteringen en toekomstige implicaties voor biologisch onderzoek in gewassen. Vooral de verschuiving van de kosten van sequencen naar analyse, met functionele genoomannotatie als essentiële stap, wordt belicht. Een van de grootste uitdagingen voor biologie en bioinformatica is de integratie van de resultaten van dergelijke dataanalyses en andere informatie in een compleet plaatje. Alleen dat zal het mogelijk maken om complexe biologische systemen te begrijpen.

Acknowledgements

In retrospect, I have to admit that in the beginning of my PhD I was feeling like a headless chicken, running around without a clear plan of what to do and where to go. Full of confidence I can now say: this did not change. Thus, getting so far was only possible with the support and guidance of friends, colleagues, my family and, of course, Deborah.

But first things first. I would like to thank my promotor Richard Visser for making this journey possible. I am always amazed how he can supervise so many PhD students and still have plenty of time for meetings with me. I concluded that he either has a very efficient secretary or a doppelgänger. Thank you very much for providing optimal PhD growing conditions and letting me follow my research interests.

In my day-to-day PhD life, Jan-Peter was the most important person. He certainly deserves his own paragraph. I have to say that Jan-Peter is an excellent teacher, especially if it comes to writing articles, giving talks and finding inconsistencies. Figuratively speaking, if my research was a water tank, Jan-Peter would be the water that makes it immediately visible where the tank is leaking. However, I had to learn to conduct such leakage tests not in the living room, but in the bathroom. Otherwise I would not only know where the leaks are, but I also would have a big mess in the living room. All this water floating around in the wrong place can be annoying. For future PhD students I therefore advice to pour »Jan-Peter« over your research only in the bathroom. In summary, given that a good student-supervisor relationship is essential to finish a PhD, I am glad that Jan-Peter was my supervisor. The thesis is finished and I have learned a lot. Thank you! I hope you liked working with me, too, even though I missed 80% of my deadlines. I guess this was my way of messing up your living room.

Besides Jan-Peter, I would like to thank Aalt-Jan, Gabino and Sander. At the same time, you were the Bermuda Triangle and the supporting pillars of my PhD. Bermuda Triangle, because of the overwhelming amount of ideas and research questions offered by you so one can get lost in finding the right direction. Pillars, because you were always there to discuss my research and to provide novel insights and angles on how to solve a problem. I liked working with you and I enjoyed your company very much.

The same is true for the remaining members of my bioinformatics ecosphere, namely Edouard, Jan, Henri, Pieter, Pierre, Heleen, Sandra, Saulo, Sven, Bas, Elio, Linda, Judith, Luca and Ke. You are great colleagues to work with, I learned plenty of practical tips, tricks and insights. The coffee breaks were essential for improving my Dutch, my political incorrectness or both. A subset of my colleagues, the »sysop-team«, deserves special thanks for keeping up the systems. Without you, my PhD would have been a nightmare. I should not forget Felipe, my PhD mate. I had a lot of fun talking crap with you, in the office and in private. It is a pity that you already left Wageningen. Outside of my bioinformatics ecosphere, I have to thank Adam, Ludmila and Hans. It was to nice to collaborate with you.

I would like to thank my paranymphs Harm and Abhishek. Harm, it was a pleasure having you as »external supervisor« and colleague. I enjoyed the coffee discussions with you and, you might believe it or not, the discussions we had were quite helpful in times I was without plan. Abhishek was one of the first friends I made in Wageningen. Together with Willemijn, Alexandra, Ioanna and Arjan I had a lot of fun and sometimes I even miss having you as housemates. Especially the huisuitjes were memorable.

At this point I will go against Jan-Peter's advice to always put the important stuff first. The most important result of my PhD is certainly not the PhD title, nor the scientific achievements, but meeting Deborah, the love of my life. I was not expecting to meet, nor to start a family with you and, in retrospect, it was pure luck that we met. Lucky me! Thank you very much, my love, for all your patience, love and our son, Oliver. Last but not least, I would like to thank my family, Hannelore, Dietrich, Vinzenz (and Paulina) for supporting all my decisions and being the safe harbor I can always return to.

As a matter of fact, I will miss some people that should have been mentioned here. Please do not take it personally, headless chickens are known for missing quite some grains of corn.

Curriculum Vitae

Joachim W. Bargsten was born on the 9th of August 1982 in Stade, Germany. In 2002, he obtained his high school degree at the Gymnasium Athenaeum in Stade, Germany. From 2003 to 2009, Joachim studied bioinformatics at the Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany. During his studies, he conducted internships at the Leibniz Institute of Plant Biochemistry (IPB), Halle (Saale), Germany, focussing on the alignment of high resolution mass spectra (Dr. Steffen Neumann) and at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, focussing on the improvement of search results in biological databases (Prof. Dr. Falk Schreiber & Dr. Matthias Lange). The work at the IPK resulted in his diploma thesis titled »user profile based ranking of search engine results from integrated biological databases«. He completed his studies with the degree »Diplom Bioinformatiker«. In 2010, Joachim started his PhD in bioinformatics at Wageningen University, initially working on comparative genomics of tomato and potato, but later broadening his research to functional annotation of genomic elements in plants (Prof. Dr. Richard Visser & Dr. Jan-Peter Nap).

Publications

Bargsten, J. W., Folta, A., Mlynárová, L., and Nap, J.-P. (2013). Snf2 family gene distribution in higher plant genomes reveals DRD1 expansion and diversification in the tomato genome. *PLoS ONE*, 8(11):e81147.

Bargsten, J. W., Severing, E. I., Nap, J.-P., Sanchez-Perez, G. F., and van Dijk, A. D. J. (2013). Biological process annotation of proteins across the plant kingdom. *Current Plant Biology*, in press.

D'Agostino, N., Golas, T., van de Geest, H., Bombarely, A., Dawood, T., Zethof, J., Driedonks, N., Wijnker, E., **Bargsten**, J. W., Nap, J.-P., Mariani, C., and Rieu, I. (2013). Genomic analysis of the native European *Solanum* species, *S. dulcamara*. *BMC Genomics*, 14:356.

Lange, M., Spies, K., **Bargsten**, J., Haberhauer, G., Klapperstück, M., Leps, M., Weinel, C., Wünschiers, R., Weissbach, M., Stein, J., and Scholz, U. (2010). The LAILAPS search engine: relevance ranking in life science databases. *Journal of Integrative Bioinformatics*, 7(2):110.

Peters*, S. A., **Bargsten***, J. W., Szinay, D., van de Belt, J., Visser, R. G. F., Bai, Y., and de Jong, H. (2012). Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *Plant Journal*, 71(4):602–614.

*These authors contributed equally to this paper.

Rogowski, K. J., Folta, A., **Bargsten**, J. W., Nap, J.-P., and Mlynárová, L. (2013). Unexpectedly rapid IS1 transposition into an Arabidopsis chromatin remodeling gene. *Transgenic Research*, 22(4):869–871.

Shahin, A., van Kaauwen, M., Esselink, D., **Bargsten**, J. W., van Tuyl, J. M., Visser, R. G. F., and Arens, P. (2012). Generation and analysis of expressed sequence tags in the extreme large genomes *Lilium* and *Tulipa*. *BMC Genomics*, 13:640.

This thesis was funded by the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI), and by the FP7 »Infrastructures« project TransPLANT (award 283496).

Printed by
Gildeprint, Enschede, NL

Education Statement of the Graduate School
Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Joachim Bargsten**
Date: **28 October 2014**
Group: **Plant Breeding and PRI-Bioscience**
University: **Wageningen University & Research Centre**

1) Start-up phase	*date*
► **First presentation of your project**	
Genome deciphering and comparative bioinformatics of Solanaceous genomes	Oct 21, 2010
► **Writing or rewriting a project proposal**	
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*1.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student day, Utrecht University	Jun 01, 2010
EPS PhD student day, Wageningen University	May 20, 2011
► **EPS theme symposia**	
Theme 4: Genome Biology, Radboud University Nijmegen	Dec 07, 2012
Theme 2: Interactions between Plants and Biotic Agents, Utrecht University	Jan 24, 2013
► **NWO Lunteren days and other National Platforms**	
NBIC Conference 2010, Lunteren	Mar 29-30, 2010
NBIC Conference 2011, Lunteren	Apr 19-20 2011
NBIC Conference 2012, Lunteren	Apr 24-25 2012
NBIC Conference 2013, Lunteren	Apr 16-17, 2013
BioRange meeting	Oct 11, 2010
BioRange meeting	Oct 10, 2011
BioRange meeting	Oct 29, 2012
CBSG Summit 2010	Mar 15-16, 2010
CBSG Clustermeeting Tomato	Jul 2010
CBSG Summit 2011	Jan 31 2011
CBSG Midterm Review	Mar 2011
TTI Green Genetics Conference	Sep 21, 2011
CBSG Clustermeeting Arabidopsis/Brassica	Oct 06, 2011
► **Seminars (series), workshops and symposia**	
Illumina Next GenerationSequencing, Seminar	Mar 03 2010
Invited seminar BGI: 'Genomics in China'	Apr 01 2010
Invited seminar Matteo Brilli: 'Development of mathematical model for carb. util.'	Apr 29 2010
Invited seminar Daniel Schubert: 'Polycomb-group proteins'	May 11 2010
Invited seminar Paul Birch	May 20, 2010
NBIC Galaxy Seminar	Sep 2010
WEES Seminar Bas Haring: 'The value of biodiversity'	Sep 16, 2010
WEES Seminar Fiona Jordan: 'Culture evolves! How phylogenetic thinking is transforming anthropology and linguistics'	Nov 18, 2010
Rob Goldbach memorial lecture by David C. Baulcombe	Oct 10, 2012
► **Seminar plus**	
► **International symposia and congresses**	
ECCB10, Ghent, Belgium	Sep 26-29, 2010
Comparative & Regulatory Genomics in Plants, Ghent, Belgium	Apr 11-12, 2011
Benelux Bioinformatics Conference, Radboud University Nijmegen	Dec 10, 2012
ISMB/ECCB, ICC Berlin, Germany	Jul 19-23, 2013
► **Presentations**	
Poster presentation ECCB10	Sep 26-29, 2010
Oral presentation, NBIC BioRange Project Meeting	Oct 11, 2010
Oral presentation, Research group evolutionary genomics, Utrecht	Jun 14, 2011
Oral presentation CBSG Clustermeeting Arabidopsis/Brassica	Oct 06, 2011
Oral presentation, NBIC BioRange Project Meeting	Oct 10, 2011
Poster presentation NBIC conference 2012	Apr 24-25, 2012
Oral presentation, NBIC BioRange Project Meeting	Oct 29, 2012
Oral presentation, Theme 4 - Genome Biology	Dec 07, 2012
Oral presentation BBC12	Dec 10, 2012
Oral presentation NBIC conference 2013	Apr 16-17, 2013
Oral presentation, Biointeractions group	Jun 10, 2013
Oral presentation ECCB/ISMB 2013	Jul 19-23, 2013
► **IAB interview**	
Meeting With a member of the International Advisory Board of EPS	Nov 14, 2012
► **Excursions**	
Subtotal Scientific Exposure	*23.5 credits**

CONTINUED ON NEXT PAGE

3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	Autumn School - Biomolecular Modelling	Nov 29-Dec 17, 2010
	Comparative genomics, Utrecht (NBIC PhD School)	Jun 27-Jul 01, 2011
	Generalized Linear Models	Jun 13-13, 2013
	Linear Models	Jun 05-07, 2013
	Bayesian Statistics	Oct 17-18, 2013
	Mixed Linear Models	Jun 20-21, 2013
►	**Journal club**	
	member of a literature discussion group at Bioinformatics	Feb 2010-Feb 2014
►	**Individual research training**	
	Subtotal In-Depth Studies	*10.2 credits**

4) Personal development		*date*
►	**Skill training courses**	
	NBIC PhD Retreat	Mar 28, 2010
	NBIC PhD Retreat	Apr 18, 2011
	Techniques for Writing and Presenting a Scientific Paper	Feb 05-08, 2013
	Project- & Time Management	Oct 09,23-Nov 20, 2013
►	**Organisation of PhD students day, course or conference**	
►	**Membership of Board, Committee or PhD council**	
	Subtotal Personal Development	*3.1 credits**

TOTAL NUMBER OF CREDIT POINTS*	**38.3**

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.*