AN INTEGRATED APPROACH TO ENHANCING FUNCTIONAL ANNOTATION OF SEQUENCES FOR DATA ANALYSIS OF A TRANSCRIPTOME

Matthew Morritt Hindle, BSc.(Hons) MSc.

Supervisors

Professor Christopher J. Rawlings
Biomathematics and Bioinformatics, Rothamsted Research

Dr Dimah Z. Habash
Plant Science, Rothamsted Research

Professor Charlie Hodgman
Multidisciplinary Centre for Integrative Biology, The University of Nottingham

This thesis was submitted to The University of Nottingham for the degree of Doctor of Philosophy

July 2012

Abstract

Given the ever increasing quantity of sequence data, functional annotation of new gene sequences persists as being a significant challenge for bioinformatics. This is a particular problem for transcriptomics studies in crop plants where large genomes and evolutionarily distant model organisms, means that identifying the function of a given gene used on a microarray, is often a non-trivial task. Information pertinent to gene annotations is spread across technically and semantically heterogeneous biological databases. Combining and exploiting these data in a consistent way has the potential to improve our ability to assign functions to new or uncharacterised genes.

Methods: The Ondex data integration framework was further developed to integrate databases pertinent to plant gene annotation, and provide data inference tools. The CoPSA annotation pipeline was created to provide automated annotation of novel plant genes using this knowledgebase. CoPSA was used to derive annotations for Affymetrix GeneChips available for plant species. A conjoint approach was used to align GeneChip sequences to orthologous proteins, and identify protein domain regions. These proteins and domains were used together with multiple evidences to predict functional annotations for sequences on the GeneChip. Quality was assessed with reference to other annotation pipelines. These improved gene annotations were used in the analysis of a time-series transcriptomics study of the differential responses of durum wheat varieties to water stress.

Results and Conclusions: The integration of plant databases using the Ondex showed that it was possible to increase the overall quantity and quality of information available, and thereby improve the resulting annotation. Direct data aggregation benefits were observed, as well as new information derived from inference across databases. The CoPSA pipeline was shown to improve coverage of the wheat microarray compared to the NetAffx and BLAST2GO pipelines. Leverage of these annotations during the analysis of data from a transcriptomics study of the durum wheat water stress responses, yielded new biological insights into water stress and highlighted potential candidate genes that could be used by breeders to improve drought response.

Acknowledgements

This thesis was only possible through of the work of many colleagues at Rothamsted Research. My thanks to all the members of the TRITIMED and Ondex SABR projects, whose contributions I acknowledge by references within this work.

"We are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness of sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size"
Bernard of Chartres

I am indebted to my supervisors over the last four years: Chris Rawlings, Dimah Habash, Jacob Köhler, and Charlie Hodgman. I particularly wish to thank Marcela Baudo, who together with Dimah Habash conceived and executed the time-series experiment that made this work possible. Michael Defoin-Platel provided invaluable collaboration and advice in the evaluation and refinement of CoPSA. I would also like to thank my fellow developers, who contributed elements of Ondex and ideas that were built upon or used during this research. I am pleased to acknowledge Andrea Splendiani, Artem Lysenko, Berend Hoekman, Catherine Canevet, Jan Taubert, Keywan Hassani-Pak, Matthew Pocock, and Rainer Winnenburg. I also acknowledge Mansoor Saqi for his suggestions throughout this project and Stephen Powers for his statistical advice, and contributions to the analysis of TRITIMED data.

Definitions of terms and abbreviations

AAO [*Gene*] Aldehyde Oxidase: a gene that encodes an enzyme that catalyzes the final step of ABA biosynthesis.
ABA [*Phytohormone*] Abscisic acid: one of the major plant hormones that forms a hub in a signalling network which regulates many cellular and plant wide processes.
ABAR/CHLH [*Gene*] cMagnesium-protoporphyrin IX chelatase H subunit: A protein that was put forward by Shen *et al.* (2006) as a candidate ABA receptor.
ABI [*Gene Family*] A family of ABA insensitive gene loci.
Accession [*Bioinformatics term*] A unique sequence of characters that within the scope of some definition uniquely define a database entry, independent of a database instance. If multiple accessions are merged together, then a new accession is created and the previous made obsolete. Only one accession should ever identify an entry, where obsolete accessions are present in an entry, this is sometimes referred to as the primary accession.
ANOVA [*Statistical methodology*] ANalysis Of VAriance: a statistical procedure by which observations are partitioned into groups representing sources of variation. **Application Programming Interface(s) (API)** [[*Computer Science term*] A defined set of publicly exposed functions in a program, that allow another program to make use of its resources. A good API exposes required functionality, while minimising complexity.
BiFC [*Molecular biology methodology*] Biomolecular florescence complimentation (BiFC): A methodology for confirming protein-protein interactions. Florescent protein fragments are attached to two or more proteins suspected of interacting. Interaction of these proteins will cause the fragments to reform and emit its florescence colour, thereby confirming the interaction.
CDPK [*Protein*] Calcium Dependent Protein Kinase: a group of proteins that are stimulated by Ca2+ to initiate their activity.
Drought Stress [*Agricultural Plant Science term*] The limitation on maximum potential crop yield imposed by a water limitation.
Endoplasmic Reticulum (ER) [*Cell organelle*] An organelle that forms an interconnected network of tubules, vesicles, and cisternae within a eukaryote cell. It is mainly involved in protein, lipid and steroid synthesis, and carbohydrate

and steroid metabolism.

Extensible Mark-up Language (XML) [*File format*] A structured text document, conforming to rules defined by W3C (2011). It allows documents to be machine-readable. **FCA** [*Gene*] A gene encoding a posttranscriptional regulator of transcripts involved in the flowering process (Macknight *et al.*, 1997).

G protein [*Protein family*] A family of proteins that bind to guanine nucleotides (GTP and GDP), and are involved in signalling.

GFP-florescence [*Molecular Biology methodology*]: Florescence labelling using a small (238 amino acids) Green Florescence Protein (GFP), which emits green light when exposed to blue light. The GFP gene can be fused into a target gene, which then may express a protein with florescence. The target genes localisation and expression in the cell can therefore be monitored (Phillips, 2001).

GPA1 [*Gene*] The sole member of the G-protein Gα-subunit family within the Arabidopsis genome.

GTG [*Protein family*] GPCR-type G proteins (GTG): A gene family including GTG1 and GTG2 which are candidate ABA binding proteins (Pandey *et al.*, 2009).

Homolog(y) [*Genetics term*] The relationship between two genes that are descended from a common ancestral genes (Fitch, 2000).

HSF [*Protein family*] Heat Shock Factors (HSFs). A family of transcription factors which regulate HSPs.

HSP [*Protein family*] Heat Shock Proteins (HSPs). A family of proteins that assist as chaperones in protein folding (Hu *et al.*, 2009).

NAC [*Protein family*] A superfamily of transcription factors, many of which are involved in hormonal regulation (Jensen *et al.*, 2010). **NCED** [*Gene family*] A gene family encoding 9-cis-epoxycarotenoid dioxygenase, which is an enzyme that is part of the ABA biosynthesis pathway.

MAPK [*Protein*] Mitogen-Activated Protein Kinase: a signalling molecule that forms the initial activator of the MAPK cascade. The second and third elements of this cascade are MAPK Kinase (MAPKK) and MAPKK Kinase (MAPKKK), respectively.

MYB [*Protein family*] A superfamily of transcription factors, which has the largest number of members in Arabidopsis (Yanhui *et al.*, 2006). They commonly involved in the regulation of developmental processes and defence response.

Object-Oriented (OO) [*Computer Science term*] A programming paradigm based around data objects, which consist of a group of fields and methods. OO is associated with a number of good practices. Encapsulation: access to an ob-

ject is restricted by public and private components. Abstraction: concepts or ideas independent of an implementation instance are separated. Modularity: where possible software is composed of separate interchangeable components. Inheritance: common properties and methods of objects are shared between implementations (e.g. Java class inheritance). Polymorphism: the ability to create variable, functions or objects that have multiple implementation forms (e.g. Java interfaces).

Ortholog(y) [*Genetics term*] The relationship between two homologous genes, whose common ancestor is the last universal common ancestor of the taxa from which the two sequences were obtained (Fitch, 2000).

OXL [*File format*] An XML based file format defined by Taubert *et al.* (2007) for lossless serialisation of an Ondex graph. It is the main file format for exchanging graphs within Ondex. **OST** [*Protein family*] Open stomata protein kinases (OST): contains the OST1 (also called SRK2E and SnRK2.6) gene in Arabidopsis which is regulated by ABA, the expression of which is strongly linked to stomatal aperture size (Mustilli et al., 2002).

PA [*Compound*] Phosphatidic acid (PA): A membrane lipid that is also a precursor for the biosynthesis many other lipids. It can also be involved in signalling.

Paralog(y) [*Genetics term*] The relationship between two homologous genes that has arisen as a result of duplication (Fitch, 2000).

PCo [*Statistical term*] A Principal Coordinate identified by a Principal COordinates analysis (PCO) analysis, and capturing variance within a dataset. PCos are numbered according to rank, starting from one, based on the quantity of variance they capture in the dataset.

PCO [*Statistical methodology*] Principal COordinates analysis: a methodology for exploring similarities and differences in multivariate data developed by Gower (1966). **PIP** [*Gene family*] Plasma membrane Intrinsic Proteins (PIP): a family of trans-membrane water channels (Phillips, 2001).

PPI [*Molecular biology term*] Protein-protein interaction (PPI). The transient or permanent physical binding of two proteins.

PP2C [*Gene family*] Protein phosphatases of category 2C (PP2Cs): Some members of this family are negative regulators of an ABA induced response (Santiago et al., 2009b).

RCAR [*Gene family*] Regulatory components of ABA receptor (RCAR): A family of 14 genes in Arabidopsis that bind to ABA and PP2Cs. Also known as PYR1 and PYLs (Ma et al., 2009, Santiago et al., 2009b).

RIL [*Genetics term*] Recombinant Inbred Line (RIL) a population that has bred without out-breeding from an F2 population. Two distinct genetic groups are crossed to create an F1 hybrid, which is back crossed with one of the parents to create the F2 population.

Semantic Drift [*Data Integration term*] The erroneous labelling of entities as equivalent that have different semantic meanings. For example: The labelling of a genetic loci as equivalent to a protein, based on a common accession in two databases.

SnRK [*Protein family*] Sucrose non-fermenting-1 (SNF1)-related protein kinases (SnRKs). A family encoded by 38 genes in Arabidopsis, that are named after their similarity to the SNF1 gene in yeast.

Water Stress [*Plant Science term*] The state wherein the water absorption of a plant does not amount to that which is lost through transpiration.

Y2H [*Molecular Biology methodology*] Yeast-2-hybrid (Y2H): A methodology for confirming protein-protein interactions. A transcription factor is fragmented into its constituent binding and an activation domain. A protein is attached to each respective fragments. Confirming the activity of the transcription factor, indirectly confirms the interaction of the proteins. The transcription factor chosen, is often linked to lethality genes, which confirm a PPI based on the death of the yeast.

Contents

Chapter 1. Introduction

Functional annotation of coding sequences persists in being a severe limiting factor in the analysis of both high-throughput genomics and transcriptomics data. The problem is aggravated in organisms like wheat, which has three genomes that together contain approximately 150,000 genes (Choulet *et al.*, 2010) and a large evolutionary distance to the plant model Arabidopsis thaliana (Gaut, 2002, Liu *et al.*, 2001). Identifying the function of a given sequence, from organisms like wheat, is frequently a non-trivial task. To compound the problem, information pertinent to gene annotations is spread across technically and semantically heterogeneous biological databases. Combining and exploiting these data in a unified way has the potential to improve our ability to predict novel-gene function (Lysenko *et al.*, 2010).

This thesis is divided into two parts. Part I comprises Chapters 2-3, and addresses the bioinformatics problems associated with functional annotation of sequence data. Part II (Chapters 6.1 and 6.3) demonstrates the utility of Part I by leveraging the new functional annotations for the analysis of a time-series transcriptome experiment in durum wheat, which studies the effect of drought in three phenotypically distinct cultivars. The bioinformatics problems addressed in Part I are motivated by the challenges of transcriptome analysis that emerge in the biological application case presented in Part II. The significantly regulated processes reported in Part II are consistent with existing models of water stress response and thus serve to validate the functional annotations predicted in Part I.

This introductory chapter provides the broad background to the challenges inherent in gene function annotation for use in transcriptome studies and presents an overview of the thesis structure. Subsequent chapters contain more detailed

introductory material pertinent to the work presented in that section of the thesis.

1.1 Part I: Addressing functional annotation poverty

The problem that motivates the bioinformatics research presented in this thesis is the analysis of a time-course gene expression dataset from durum wheat, which used the Affymetrix wheat genome array. This GeneChip contains 55,052 transcripts from all 42 chromosomes in wheat. Sequence information for this array comes from *Triticum aestivum* UniGene Build #38 (build date April 24, 2004). Also included are ESTs from the wheat species *T. monococcum*, *T. turgidum*, and *Aegilops tauschii*, and GenBank full-length mRNAs from all species through May 18, 2004 (Affymetrix, 2011c). Although this chip was designed using all the available wheat EST data that was in the public domain at the time, it unfortunately only represents approximately 30% of the expected gene content of wheat. The partial coverage of the wheat genome in this transcriptome dataset is the first source of poverty in the functional annotations available for wheat.

For those genes that are represented on the GeneChip, two important factors affect the efficacy of a given annotation set when interpreting the experimental data: the accuracy of the proposed annotation and the coverage of the genes for which correct annotations can be assigned.

1.1.1 Types of gene annotation

Functional annotations for genes are generally inferred from the predicted function of the protein which they encode and may be represented in a number

forms, the most common of which are free-text (unstructured), Enzyme Commission (EC) codes (NC-IUBMB, 1999), and Gene Ontology (GO) terms (Ashburner *et al.*, 2000). Free-text is of limited practical applications for systems wide analysis of high throughput transcriptome data. However, Natural Language Processing techniques are increasingly making the annotations encoded in free text accessible by identifying cross-references to other more structured annotations (Chapman and Cohen, 2009). Within this thesis the focus is on exploiting structured functional annotation sources, the most pervasive of which are EC and GO terms, which use non-redundant structured hierarchies and ontology's respectively and are therefore amenable to analysis by computational reasoning. Other less widespread functional classification systems include the MIPS Functional Catalogue (FunCat) (Ruepp *et al.*, 2004) and the classification hierarchy used within MAPMAN (Thimm *et al.*, 2004). Mapping of equivalence between annotation systems terms are often provided. The most comprehensive sets of translations are available for GO, which has a dedicated mapping file format for representing these relationships (The Gene Ontology Consortium, 2011a).

Additionally, databases which define groups of structurally similar genes such as Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.*, 2000, 2003) serve to provide a similar functional annotation resource. However, an important distinction is that by consequence of grouping by sequence they are structurally non-redundant; this is at the expense of functional redundancy (*i.e.* they define the relationship between genes rather than between functions). The COG database contains groups of similar protein sequences and has assigned functional annotations for each of these groups. If two different clusters have the same function, then they are redundantly represented in COG. Therefore, COG can be said to be structurally non-redundant and functionally redundant: it is a system of classification for protein sequence rather than function. The same is true for databases of functional domain families such as Pfam (Finn *et al.*, 2009).

Functional redundancy in these databases is common place as convergent evolution may lead to a single function or process (Koonin and Galperin, 2003) being encoded by very different sequences or protein structures. Depending on the granularity with which a function is defined, a single organism often has multiple protein-sequence solutions for a given function (*e.g.* for DNA binding (Riaño-Pachón *et al.*, 2007)).

The ability to define a function within a vocabulary is a prerequisite for leveraging function driven queries and statistical analysis. This can be achieved through a non-redundant controlled vocabulary or through statements which explicitly state the semantic relationship between terms in an ontology (*e.g.* "is a" and "part of" hierarchies). Therefore, systems like EC, FunCat and GO are important resource in defining functions independent of protein sequence or structure. When COG or Pfam families are annotated with terms from these classification systems, this allows powerful analysis that can construct cross database queries such as: "for a given set of sequences, what Pfam/COG families do they belong to, and which functions/processes are significantly enriched from the EC/FunCat/GO annotation of these families".

The Enzyme Commission (EC) provide a nomenclature for enzymes based on the reactions they catalyse. EC numbers classify enzyme reactions based on four levels of a hierarchy. The roots of the hierarchy are six broad enzyme classes, each subdivided into subclasses, and sub-subclasses. The fourth digit of an EC term is the serial number of the enzyme, the specificity and nature of which is set out in the guidelines of the EC (NC-IUBMB, 1999). Table 1.1 shows a breakdown of how an EC number is constructed. For a given nomenclature name the enzyme commission states that "a certain name designates not a single enzyme protein but a group of proteins with the same catalytic property" and they make no requirement for structural similarity for proteins annotated to a given EC name. EC is therefore a reaction based naming of proteins, and while this enables proteins to categorised independent of their sequence and physical

structure (*i.e.* two different sequences can have an identical EC term if they act on the same products and substrates in the same manor), it is not in the truest sense a functional annotation system. According to EC guidelines a protein classification should be assigned "based on the first enzyme-catalysed step that is essential to the subsequent transformations". Pragmatically, however, databases such as the ExPASy ENZYME (Bairoch, 2000) database, use EC numbers as identifiers of biochemical catalysis of a given reaction (defined by product and substrates), and allow multiple EC terms to be annotated to a given protein.

The Gene Ontology (GO) is a collection of three ontologies, which will be re-

Table 1.1: A breakdown of EC number hierarchy system using Glu-Glu dipeptidase (3.4.13.7) as an example.

EC Number	Digit	Reaction Specificity	Example
3.-.-.-	1	Reaction type	Hydrolases
3.4.-.-	2	Substrate class	Peptidases
3.4.13.-	3	Substrate sub-class	Dipeptidases
3.4.13.4	4	Substrate	Glu-Glu dipeptidase

ferred to here as GO categories. The *cellular component* category refers to parts of a cell or its external environment, and in terms of annotation allows gene products to be identified as "located in" a *cellular component*. The *molecular function* category describes the biochemical activities of a gene product, such as binding or catalysis". It describes the potential for a given activity, rather than the conditions under which it can be found. EC terms often have equivalents within the *molecular function* category. The *biological process* category describes participation by a gene product in operations or sets of molecular events with a defined beginning and end. Processes often involve physical or chemical transformations of entities, and are often described in terms of their end goal (*e.g.* proline biosynthetic process, GO:0006561). Pathway names in metabolic databases like KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi, 2006, Caspi *et al.*, 2008) and Reactome (Matthews *et al.*, 2009) often have equivalents within this

category. When combined, the three GO categories form a powerful resource for annotating a gene with a summary of its biological function. GO has therefore become a popular and valued resource for developers of gene and protein functional annotation tools.

For example: Okamoto *et al*. (2006) describes how the*Arabidopsis* gene CYP707A, encodes the endoplasmic reticulum (ER) membrane-bound Abscisic Acid 8'-Hydroxylase, which is involved in ABA biosynthesis, and expressed in the seed. Additionally they assert that this gene is involved in seed maturation. Using the GO categories, the processes this gene is involved in can be described by the terms "seed maturation" (GO:0010431), and "ABA biosynthesis" (GO:0009688). The latter is equivalent to the pathway PWY-695 in MetaCyc, which is located within a much larger pathway called carotenoid biosynthesis (ec00906) in KEGG. The function is captured by "(+)-abscisic acid 8'-hydroxylase activity" (GO:0010295), which has the equivalent EC term 1.14.13.93. The proteins cellular location is described by the "ER membrane" (GO:0005789) term.

GO terms form part of a directed acyclic graph (DAG), which encodes semantic relationships using "is a" and "part of", relationships. This enables semantic reasoning across annotation sets. For example, given any two genes annotated with GO functions it is possible using the relationships within the GO DAG to locate common ancestors, and thereby determine which properties two or more genes share in common. The inverse is also possible: identify all genes that inherit from a given ontology term such as catalytic activity (GO:0003824). A number of such methodologies for exploiting the GO DAG for reasoning are described in Chapter 3.

The Gene Ontology therefore forms a central role for the annotation and analysis of gene and protein sequence data. Its use is pervasive within the biological and bioinformatics literature, due to its simplicity (only two main types of relation types, and one less common "regulates" relation), maturity (GO was conceived by Ashburner (1998), and breadth of scope. The ability of GO to rep-

resent information also captured by other forms of functional annotation has led to the development of cross references to other structural and functional annotation schemes, to the Gene Ontology. This enables users of GO to incorporate gene annotations from other annotation schemas, or conversely GO annotations may be used to improve other annotation systems such as EC and FunCat.

1.1.2 Provenance of annotation

Data provenance is an important issue for gene annotation. The assignment of a given annotation to a gene, in a database, may have a complex history, involving multiple biological experiments, databases, and computational inferences. The confidence of any given annotation is dependent on the full history of data and methods it depends upon. For a given computational annotation in a higher-plant for example, it may be dependent on multiple experimental evidences and methods, on multiple genes in that family, from multiple species. It may also be dependent on computational and statistical evidences, such as sequence alignment and co-expression. In order to compute such confidence, it is essential that such provenance should be in a machine-readable form. Karp (1998) notes that most sequence databases retain little or no provenance information regarding the assignment of functions to sequences. Buneman *et al.* (2000) highlight a further potential problem: cycles of database interdependence can lead to a perpetual loops of inaccurate data. They highlight literature curation as one such source of error, where biological literature reference knowledge in databases, which themselves curate knowledge from published literature. The conclusion is that using knowledge in public databases can be potentially dangerous, without a full history of provenance for each statement, which should ultimately trace back a set of experimental evidence.

There are currently no widely adopted standards for describing the provenance of *in silico* pipelines in bioinformatics, and there is a pressing need for capturing this information (Stevens *et al.*, 2007). Defining some universal standards for such pipelines will be challenging, given the diversity of tools used in bioinformatics to generate pipelines, and the complex interdependencies between pipelines and evidences. Failure to consider provenance can result in the mis-annotation of genes and potential errors being propagated across databases, and reintroduced even when corrected.

A related issue is the potential loss of the resolving power of computation methods, due to iterative expansion of sequence clusters, beyond their original ability to resolve function. Bork and Koonin (1998) was one of the first to discuss this as a potential source of noise in public sequence database. Brenner (1999) went on to describe how such errors were the result of computational inference across sequences, where there is insufficient homology, or deficiencies in the alignment algorithms used. He emphasised the importance of labelling annotations in databases that were the result of computational predictions. This can prevent inferences based on computational prediction, and the potential chain of mis-annotation, as small errors are compounded, and the discriminatory power of a functionally annotated sequence cluster reduced Gilks *et al.* (2005, 2002) have presented a statistical framework for modelling how these chains of errors can percolate through sequence annotations, and reduce the function-resolving power of clusters.

There have been a number of efforts moving towards improved provenance tracking in database. The GOA file format, which stores GO gene annotations (The Gene Ontology Consortium, 2011a) contains a field to record the experimental evidence of a annotation. Table 1.2 shows evidence codes approved by GO. While useful for tracking the reliability of a GO annotation, Inferred from Electronic Annotation (IEA), is too broad a category to be useful beyond excluding non-primary evidence. To address this problem, UniProt (The UniProt

Consortium, 2010) records extend the IEA evidence by appending the computational method or database, from which the prediction is sourced. However, this only extends to referencing a simple controlled vocabulary of evidence, and omits basic provenance details such as database and program versions, pipeline structure and method parameters.

Table 1.2: Evidence codes for the provenance of GO annotations as defined by the Gene Ontology Consortium.(The Gene Ontology Consortium, 2011a)

Code	Type of evidence	Type category
NAS	Non-traceable Author Statement	Author Statement
TAS	Traceable Author Statement	
IEA	Inferred from Electronic Annotation	Automatically-assigned
IGC	Inferred from Genomic Context	Computational Analysis
ISA	Inferred from Sequence Alignment	
ISM	Inferred from Sequence Model	
ISO	Inferred from Sequence Orthology	
ISS	Inferred from Sequence or Structural Similarity	
RCA	Inferred from Reviewed Computational Analysis	
IC	Inferred by Curator	Curator Statement
ND	No biological Data available	
EXP	Inferred from Experiment	Experimental Evidence
IDA	Inferred from Direct Assay	
IEP	Inferred from Expression Pattern	
IGI	Inferred from Genetic Interaction	
IMP	Inferred from Mutant Phenotype	
IPI	Inferred from Physical Interaction	

1.1.3 The Affymetrix GeneChip

Part I of this thesis specifically addresses the annotation of genes measured by the Affymetrix GeneChip, which is used within the use case of part II. Within this section, a brief overview of the design and terminology of the sequences on the GeneChip is provided.

Affymetrix GeneChips are composed of a large number of cells, each of which

represents either a perfect match (PM) and mismatch (MM) probe (a homomeric mismatch at the 13th position for a PM probe: A->T or G->C) (Affymetrix, 2011a). MM probes act as controls for cross hybridisation. These unique probes are 25 nucleotide bases in length and are synthesised using photolithographic fabrication. This involves the use of a mask to selectively expose light onto a silicon wafer, which directs a light dependent chemical synthesis process. Hydroxyl groups are initially formed on the wafer by light passing through the mask. Nucleotides are then added to the sequence feature, one nucleotide at a time, with successive application of masks. This fabrication process is shown in Figure 1.1. Each of the tens-of-thousands of 25-mer sequence features, that the GeneChip array is targeted to quantify the expression of, is potentially hybridised to 11-20 PM probes. The full sequence that the PM probes are designed to measure the expression of is the *consensus sequence* and the subset of the sequence they recognise is the *target sequence*.

Functional annotations, for sequences on an Affymetrix GeneChip, usually cor-

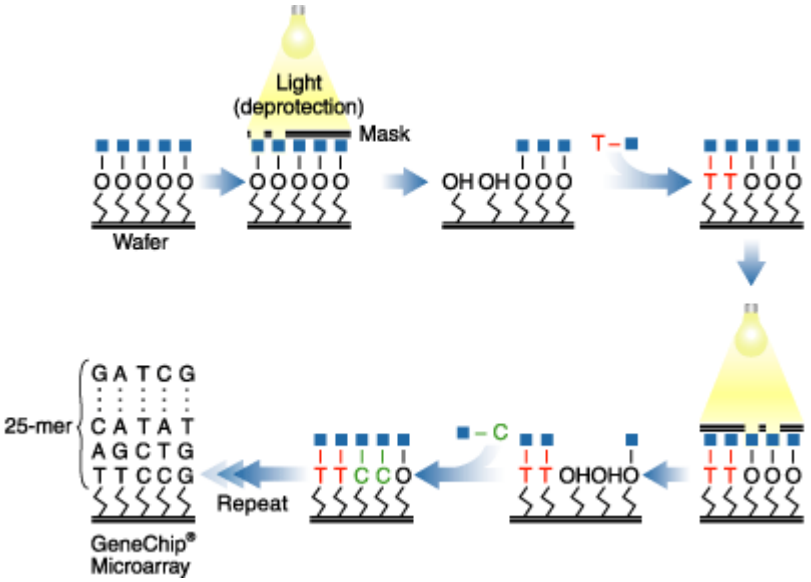


Figure 1.1: The photolithographic fabrication process used by Affymetrix to create microarray GeneChips.

respond to consensus or probe-set sequences. The relationship between these two sequence types is one-to-one and they share the same Affymetrix probe-set id. A probe-set id is appended by a basic extension nomenclature. The *_at*

and _*st* indicate whether the target sequence is sense or anti-sense respectively. These sequences are also sub-categorised by _*a*, _*s*, _*x*, or nothing. Sequences without a code appended to the id have been designated as uniquely identifying the set of corresponding target probes. Probe-set ids appended with _*a* recognise multiple alternative transcripts. Those appended with _*s* share common probes from multiple genes. Probes for ids appended with _*x* may cross hybridised unpredictably with multiple transcripts. These probe-set id categories are summarised graphically in Figure 1.2.

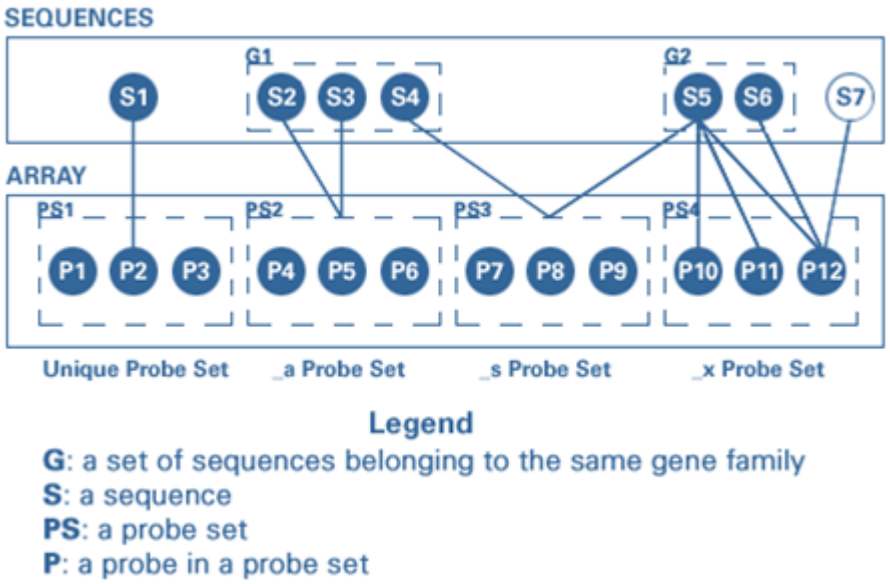


Figure 1.2: A graphical representation of the different probe set types on the Affymetrix chip Affymetrix (2011b).

1.1.4 Detecting putative functional orthologs

The presence of large protein families, with many paralogous proteins, make the process of detecting of the true ancestral functional ortholog, using sequence similarity alone, challenging. Algorithms like INPARANOID (Ostlund *et al.*, 2010a, Remm *et al.*, 2001), OrthoMCL (Chen *et al.*, 2006, Li *et al.*, 2003), OrthoMCL(Chen *et al.*, 2006, Li *et al.*, 2003), and PHOG (Datta *et al.*, 2009a,

Merkeev *et al.*, 2006) have attempted to identify the ancestral ortholog from multiple paralogous protein in the sequence alignment hits (using programs like BLAST). These approaches can improve the accuracy with which functional orthologs are detected. However, the most reliable computational approaches for predicting functional orthologs use phylogenetic analysis, however due to their computationally intensive requirements, and the requirement for human supervision in these approaches means they are often not suited to whole geneome analysis (Gabaldón, 2008).There are also limits to which sequence-based phylogenetic analysis can be used to reliably transfer function. Single residue changes can often have dramatic effects on the protein function and its participation in molecular processes. For example, Hanzawa *et al.* (2005) have shown that a single amino acid mutation in the TFL1 protein, in *Arabidopsis*, inverts the function from flowering repression, to that of an activator. Predicting, the importance of a residue for protein function using the phylogenetic approach, requires the alignment of multiple orthologous sequences. In many instances insufficient orthologous sequences exist, or too few of the functions of the putative orthologs have been experimentally characterized to establish they are functional orthologs. Karp (1998) notes that the danger in identifying putative functional othologs through sequence alignment is that it is very difficult to assign a measure of confidence that a given function can be transferred at a observed alignment score. Often, the alignment score threshold is defined arbitrarily, and the confidence unknown.

Structural similarity between proteins has been shown to be a more reliable method for interfering similar function, than sequence based similarity, however divergent and convergent evolution has led to some proteins which have different structures with similar functions, and similar structures with different functions, respectively (Hegyi and Gerstein, 1999). For example: Hegyi and Gerstein (1999) highlight the TIM-barrel structure, which is found in proteins encoding 16 evolutionary divergent enzymatic functions, which include

representations from four different EC classes: oxidoreductases, hydrolases, lyases, and iomerases.

Part of this thesis (Chapter 3) describes an automated, and high throughput, pipeline for transferring function based on the alignment with sequences from model organisms which have experimentally derived function. The sequences from which function is transferred are termed putative functional-orthologs. There is therefore uncertainty in transferring function based on sequence alignment alone, and functional assignments using this method should always be viewed as putative, even in well conserved protein families. Where multiple conserved domains exist, the probability is greater that a set of aligned sequences will share the same function (Hegyi and Gerstein, 2001). However, Hegyi and Gerstein (2001) has shown that even with multiple conserved domains, functional divergence in families frequently occurs. The divergence of function in a protein family will depend on the evolutionary pressures associated with it's role. Pathogen responsive proteins in sessile plants is an excellent of protein families under extreme evolutionary pressure. These proteins have a wide functional repertoire, which is constantly evolving in an evolutionary arms-race against plant pathogens Shan *et al.* (2007). The transference of function, across organisms, for pathogen responsive proteins is therefore a challenge.

1.1.5 The current status of annotations for Affymetrix plant species GeneChips

GO is one of the most important forms of structured annotation for interpreting high-throughput transcriptomics data. A low coverage or quality of GO annotation for the gene transcripts used to design a microarray adversely affects the power and validity of the later analysis. Low coverage of annotation

can result in much of the significant expression data being unexplained, and the resulting conclusions may be unrepresentative of the true biology of the system.

For Affymetrix GeneChip arrays, a standard annotation is provided by the NetAffx pipeline (Liu *et al.*, 2003). NetAffx links data from multiple resources using the SRS data linking system from Biowisdom (2009), and combines this with its own computational predictions. Annotations are derived from the NCBI databases UniGene, LocusLink (now EntrezGene), and Homologene databases (Sayers *et al.*, 2009)). UniGene is a database for identifying non-redundant sets of EST sequences that represent transcribed genes. This makes it a particularly important resource for partially-sequenced organisms such as wheat, where much of the array has been designed from assembled EST sequences. EntrezGene (Maglott *et al.*, 2005) contains unique and stable curated sequences and annotations derived from RefSeq (Pruitt *et al.*, 2005). Homologene (NCBI, 2011a) is a set of predictions of homologs and paralogs. SWISS-PROT, which is the curated component of UniProt, is also used to acquire annotations via the protein gi-number (NCBI, 2011b). Other annotations are acquired using InterPro (Mulder *et al.*, 2007) and pathway annotations using GenMAPP (Dahlquist *et al.*, 2002).

As well as linking existing annotations NetAffx produces electronically inferred annotations. HMMs are used to identify conserved regions using the GRAPA (Shigeta *et al.*, 2003) phylogenetic methodology, which identifies families and sub-families from conserved regions. Collections of annotated genes used to generate HMMs were created for SCOP, EC and G protein-coupled receptors (GPCR). In parallel to the NetAffx GRAPA pipeline, PSI-BLAST (Altschul and Koonin, 1998) is used to categorise kinases according to the Hanks and Quinn (1991) protocol, TMHMM (Krogh *et al.*, 2001) is used to identify transmembrane regions of proteins and BLASTx is used against GenBank (Benson *et al.*, 2007). Although NetAffx provides a recognised basis for annotation of GeneChip se-

quences, it has a number of shortcomings as the foundation of a more sophisticated approach to annotation. The provenance of the predicted annotations is not preserved within the resulting annotation files. The importance of provenance for annotation has been described in 1.1.2. In particular detailed information, on the parameters and thresholds, has not been fully described for the sequence comparison methods embedded in NetAffx. Consequently no measure of confidence is provided for any of the annotations. While some functional annotations provided by NetAffx may have been predicted by highly conserved sequence-similarity and strong experimental evidence, it is not possible to discern these from annotation derived from weak BLAST hits.

An alternative source of Affymetrix GeneChip annotations, which includes the wheat Microarray that is the subject of Part II, is the BLAST2GO Functional Annotation Repository (BLAST2GO-FAR) (Escobar, 2011), which is built using the BLAST2GO tool (Conesa and Götz, 2008, Conesa *et al.*, 2005). BLAST2GO is a tool for predicting GO and EC annotations based on sequence similarity. The structure of the BLAST2GO pipeline is shown in Figure 1.3, which is divided into five stages. Stage 1 allows users to identify target sequences by comparing query sequences using the NCBI BLAST sequence comparison methods with NCBI sequence databases (Sayers *et al.*, 2009) and/or to a custom sequence database. In Stage 2, the Stage 1 targets are then mapped to GO annotations within the GO database (The Gene Ontology Consortium, 2011b), or GOA format annotation files (The Gene Ontology Consortium, 2011a). Stage 3 of the pipeline is the critical and novel stage of BLAST2GO, and involves selecting terms from the set of GO annotation terms that occur within the annotation of all sequence hits. This is achieved by producing an annotation score (AS) for each of these candidate terms. This AS is calculated based on the sum of two other scores which are termed the direct term (DT) and second term (AT). The DT is simply the maximum scoring sequence hit for the given term, where the score of the hit is the product of the best GO evidence code weighting (provided

by a lookup table) and the sequence similarity. The exact nature of the sequence similarity score used by BLAST2GO is not clearly specified by their publications, but it can be speculated that either bitscore or sequence identity could perform this role. The formula for calculating DT is given in Equation 1.1.

The BLAST2GO score was calculated on a term t and a query sequence q, which they term the Direct Term (DT). The function $hits(t, q)$ returns the set of sequence hits that by inference could annotate t with the term q. The function $similarity(i)$ returns the sequence similarity of t to i. The function $evidence\_weight(i, t)$ returns the evidence code weight of the range $\{0..1\}$, which annotates the hit i with t.

$$DT(t, q) = \text{MAX}_{i \in hits(t,q)}\, similarity(t, i) \times evidence\_weight(i, t) \quad \text{(Equation 1.1)}$$

The second score, used in the calculation of the overall score, is defined ambiguously in the BLAST2GO publications. Conesa and Götz (2008) define it as the possibility of abstracting to a parent term, which is controlled by a weight ϵ . They define abstraction as the annotation to a parent node where several child nodes are present. The score AT is a product of the weight ϵ and the number of "terms that unify at the node". The most obvious interpretation of this is that it refers to the cardinality of the set, which is the intersection of all children of a given term with all other candidate terms. The implication of this is that in order to facilitate abstraction to parent terms they must also include parent terms in the scoring; this is confirmed by the subsequent inclusion of a "lowest term" selection function. Based on these assumptions Equation 1.2 gives the formula for AT.

The calculation of the score for the second term (AT) in BLAST2GO for a given term t within the set of all candidate terms C. The function $children(t, C)$ returns the terms in C that are children of t based on the GO hierarchy. The type of relationships in GO that are traversed to identify the child terms is not defined, but implicitly it can be assumed "is a" is used and optionally "part of". The weight

ϵ controls the possibility of abstraction, and the overall affect of AT on overall annotation.

$$AT(t, C) = \epsilon \, |children(t, C)| \qquad \text{(Equation 1.2)}$$

Finally, using these scoring functions the subset of selected terms is extracted from all candidate terms. This is achieved by scoring each candidate term and their parents (DT+ AT), retaining only those terms that score better than a set threshold, and then retaining the subset of "lowest node" (leaf terms).Stages 4 and 5 represent optional statistical analysis and visualisation of the resulting annotations.

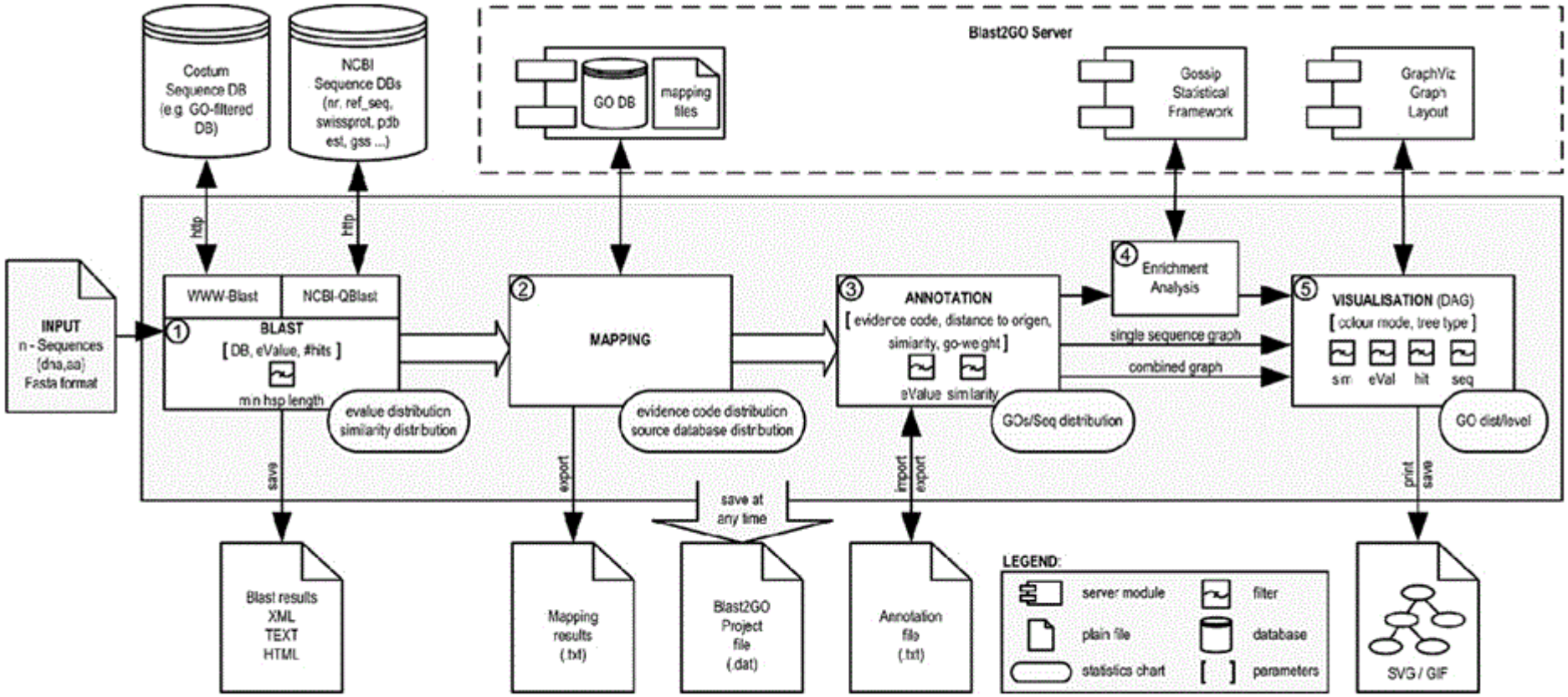


Figure 1.3: A schematic of the five step BLAST2GO pipeline. The five steps entailed BLAST of query sequence (1), mapping of query results to GO (2), applying annotation rules to select GO terms (3), statistical analysis (4), and visualisation (5) (Conesa *et al.*, 2005)

For the wheat annotation in BLAST2GO-FAR BLASTX, the BLAST2GO pipeline was run by Escobar (2011) using wheat Affymetrix consensus sequences against the NCBI GenBank NR (non-redundant) database. This contains GenBank CDS translations, Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR), and the Protein Research Foundation (PRF). In Stage 1 they used the following cut-offs: the first 20 hits per query, e-value less than 1×10^{-3}, and a minimal alignments length of 33 amino acids. These are low stringency parameters, and are likely to result in a high rate of false positive in the initial pool of candidates. However, the scoring metric acts to counteract these false positives, by selecting for the best alignments, highest confidence, and most consistent annotations.

The coverage of sequences on the Affymetrix chip annotated with at least one GO term in the functional annotation provided by NetAffx and BLAST2GO is shown in Figure 1.4. The low functional annotation coverage of both providers is a limitation for researchers who wishing to leverage the existing annotations for crop transcriptome analysis. It is evident from Figure 1.4 that more than 98% of the wheat microarray sequences have no annotations, from any category of GO, in the NetAffx annotation. The number of transcripts that were not annotates reduces considerably to 65% using BLAST2GO-FAR annotation. However, it is unclear from this data what the difference in quality is between NetAffy and BLAST2GO-FAR annotation. This issue is addressed in more detail within Chapter 3, where the annotations predicted by the pipeline developed in this thesis are compared. However, it is clear that low GO annotation coverage of microarray sequences is a problem for many of the Affymetrix plant species GeneChips. Therefore, there is a necessity to improve coverage of GO annotation for these arrays without adversely sacrificing the quality of annotation. This forms the primary motivation for Part I of this thesis, and is a prerequisite for the biological questions posed by the application use-case presented within Part II of this thesis.

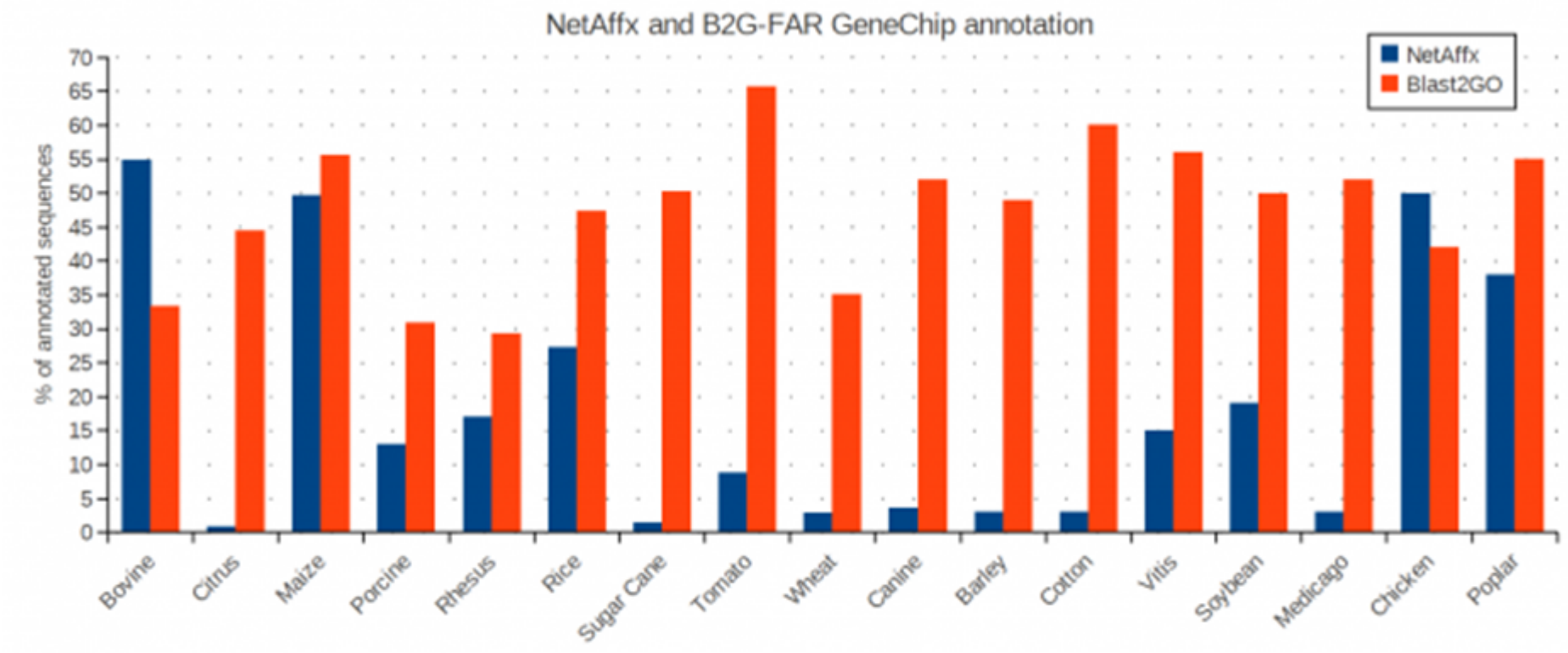


Figure 1.4: Existing GO annotation coverage of the target sequences on the NetAffx and Blast2GO Affymetrix plant microarrays. Data is for the September 2010 release of BLAST2GO-FAR (Escobar, 2011).

As previously discussed in this section, and is apparent from Figure 1.4, a limiting factor of interpreting transcriptome data in wheat is low coverage of functional annotation. Sequence based function prediction is aggravated by a lack of sequence data for durum wheat and a large evolutionary distance to the model organism *Arabidopsis*. The evolutionary closest source of sequence data is bread wheat, which contains an ancestral durum genome.

The genetics and evolution of the *Triticeae* tribe, is described within Chapter . One current estimate puts the number of genes on each diploid genome of bread wheat at 50,000 (Choulet *et al.*, 2010). GenBank currently has over 1 million EST sequences, however only 1,830 fully sequenced genes for bread wheat (October 2010). In Entrez wheat has 41,000 UniGene records which are clusters of sequences that are believed to be a single gene based on protein similarities, cDNA alignment, and genomic location data to originate from the same transcription locus. As previously stated the Affymetrix wheat chip is built from *Triticum aestivum* UniGene Build #38 (build date April 24, 2004), together with ESTs from across *Triticeae*. Some of these genes may be incomplete and misassembled, which further aggravates the assignment of function through sequence methods.

Feuillet and Muehlbauer (2009) estimate that the monocot wheat diverged from dicot *Arabidopsis* 150 million years ago (see Figure 1.5). It can be expected therefore that significant differences exist in the structural and functional repertoire of genes between these two species. In some instances this may include novel gene families or processes, not present in *Arabidopsis*.

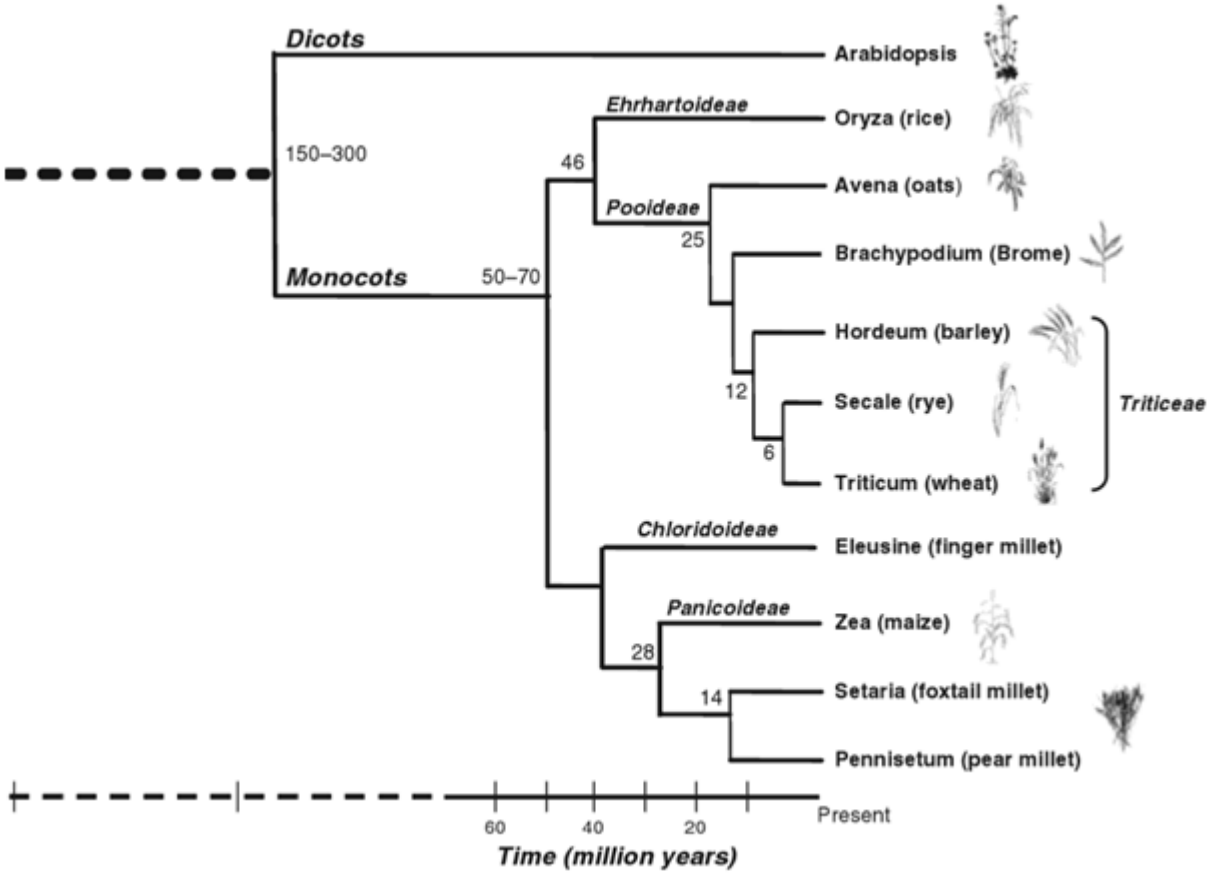


Figure 1.5: The phylogenetic relationship between grasses and the model organism *Arabidopsis*. Estimated divergence times in millions of years are indicated on branches (Feuillet and Muehlbauer, 2009).

1.2 Part II: Applying functional annotation to transcriptome analysis

The previous section presented the requirement for improved Affymetrix plant GeneChip annotations and in particular highlighted the difficulties associated with the wheat array annotation. Part II of this thesis entails the analysis of a time-series transcriptome data set from three differentially drought-tolerant cultivars of durum wheat subjected to water stress conditions. Part II comprises two chapters; Chapter describes the controlled environment experiment, the genetics of durum wheat, and the biological background of water stress in plants; Chapter details the statistical and bioinformatics analysis of the wheat

microarray, utilising the improved functional annotations derived from the pipeline described in Part I.

1.2.1 The water stress use-case

As discussed in Section 1.1, there is a general need for improved functional annotation of transcriptomic sequences in plants. This section outlines a specific use-case in the form of a transcriptome time-course response experiment in durum wheat (*Triticum turgidum subsp. Durum*), which was conducted at Rothamsted Research as part of the TRITIMED project. It begins by describing the biological motivation and importance of the experiments, and proceeds to outline the structure of the experiment and the types of annotation pertinent to its analysis.

Durum wheat is a tetraploid species of wheat that is widely grown agriculturally and used in the production of pasta and bread. It is widely grow in North and Central America, Russia, Europe, North Africa, and West Asia. In 2005 the EU accounted for 27% of the worlds production and North and Central America 34%, with most of the remainder produced in Asia (USDA, 2005). For individual countries the largest producers in terms of the proportion of world yield in 2005 were France (7%), Syria (8%), Turkey (8%), USA (10%), Italy (13%), and Canada (19%). Many of these regions are projected to have reduced precipitation, as a consequence of climate change (Neelin *et al.*, 2006). Within one of the major production areas of high quality grain (Mediterranean and western Asia) a trend of declining rainy days has been observed (Gupta *et al.*, 2009), which is shown in Figure 1.6. There is therefore a pressing need to better understand drought resistance in durum wheat and to provide candidate genes for targeted breeding. The development of drought resistant varieties will be a prerequisite in ensuring yield and therefore food sustainability in future years.

The methodology for the TRITIMED microarray experiment is described in

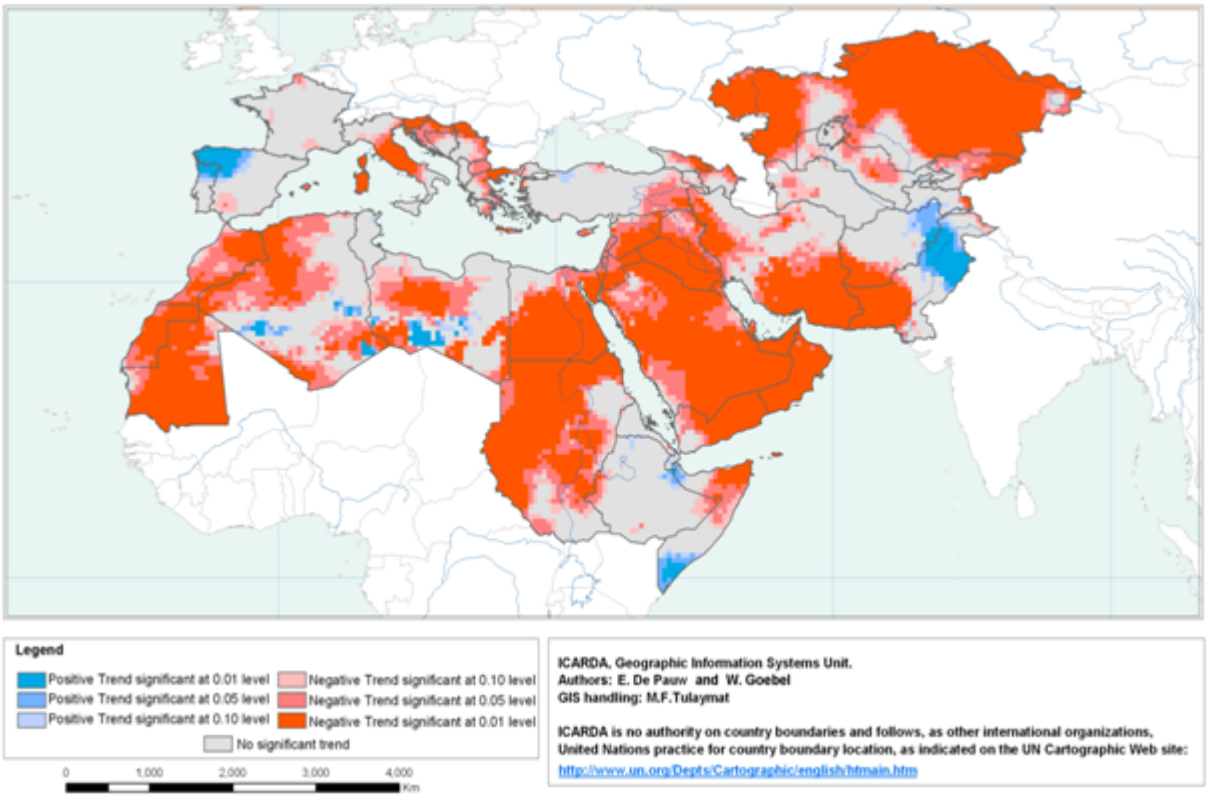


Figure 1.6: Trends for the number of rainy days in central and west Asia, north Africa, and the northern Mediterranean (1901-2002) (Gupta *et al.*, 2009)

Chapter . This section provides only a general overview of its structure and aims. The experiment consisted of time-series observation of water stress in a controlled environment for three cultivars of Durum wheat: Lahn, Cham1 and RIL2219. Lahn is a high yield variety but does not have good yield stability under drought. Cham1 has a lower yield than Lahn, but maintains high yield under drought conditions. RIL2219 was observed to have one of the best yield stabilities under drought of any of the recombinant inbred lines that resulted from crossing Cham1 and Lahn. The RIL2219 was observed to have higher yield stability than either parent. Five 24-hour interval transcriptome observations, using the Affymetrix wheat chip array, were conducted over five days of increasing water stress, from flag leaf tissue.

The experiment therefore measured two independent variables: time and cultivar. Time measurements were spread over five days and captured early re-

sponses to small changes in Relative Water Content (RWC) of the leaf, as well as late responses to dramatic changes in RWC. A biological overview of the known early and late molecular responses to drought are provided in Chapter . The molecular drought response is complex, and incorporates signalling, metabolic pathway, and structural changes to the plant. These responses are temporally and specially coordinated, and can result in system wide responses in the plant such as the initiation of senescence. The physiological consequences of drought also results in physical, heat, and salt stress, which have their own molecular sensing and response pathways. The TRITIMED experiment aims to characterise and observed these time dependent responses in the durum wheat, and compare and contrast how these responses vary across the three cultivars. The differences in the yield stabilities of the three cultivars could be a result of one or a combination of observed molecular differences: (a) distinct processes active between the varieties, (b) temporal shifts in processes, or (c) quantitative differences in process components. Molecular process annotation of genes is therefore important in considering process level changes in the transcriptome. In order to further dissect individual active processes, annotations of gene molecular functions are required. Both of these categories in the Gene Ontology are of particular importance in interpreting this data-set. EC annotations relate directly to enzyme molecular functions, and are relevant to the dissection of metabolic pathway processes.

Protein-protein interaction (PPI) networks together with protein-gene interactions (PGI) controls are thought to play a crucial role in a plants response to drought (Hirayama and Shinozaki, 2007). Plant hormones, particularly abscisic acid (ABA), are central to these regulatory signalling pathways, which coordinate gene expression. These signalling networks are described in detail within Chapter . There is therefore a strong motivation to incorporate these pathways, in the analysis of the TRITIMED dataset. Currently there is a number of PPI and PGI networks in *Arabidopsis* that Lysenko *et al.* (2010) have demonstrated are

amenable to data integration, and complimentary to each other. Regrettably, no database of experimentally validated PPI and PGI interactions exists for wheat. However, a large body of work exists in inferring these networks, from experimental evidences and comparative genomics (Carter, 2005). However, these methodologies often have high false positive rates, and require data that is not available for wheat. Usually multiple sources of complementary evidence are required to build high quality networks (De Smet and Marchal, 2010). For this reason, the scope of this thesis does not include PPI and PGI databases or inference methods. However, the biological process category within the Gene Ontology includes terms for signalling pathways (GO:0023033). Annotations of genes with these terms can provide some indication of enriched or active pathways within a microarray dataset. Although a topological analysis of these networks is not possible, as GO does not provided information on the adjacency of process. Also, identification of transcription factors is relatively straightforward given the conserved nature of families of DNA binding domains. There are many plant transcription factor databases, which contain annotation for wheat or closely related species. Those integrated in this project are described within Section 3.3.1. The annotation of transcription factors in a microarray dataset, such as TRITIMED, can reveal the quantity of regulation and the families of genes involved. This could have been extended to inference of PGI networks, using co-expression in the TRITIMED data. However, this is not included as part of this research project. The lack of genomic sequence data in wheat means it is not possible to identify cis-binding regions (Su *et al.*, 2010). The large 24 hour time interval between measurements in TRITIMED also makes identifying causality through network reconstruction techniques challenging. The time it takes for a transcriptome copy number change to affect a change in the protein levels varies, critically, depending on the protein, and the quantity of ribosomes in the cell (Piques *et al.*, 2009). Based on the work by Piques *et al.* (2009), in *Arabidopsis*, around three hours would be a reas-

onable time-interval to capture even the fastest protein translations in plants. However, the GO process ontology does includes signalling pathway annotations (GO:0023033), which provides some indication of enriched or active pathways within a microarray dataset. However, a topological analysis of these data is not possible when using GO, as they only indicate membership of signalling pathway, and it is not possible to reconstruct the full network without information on the relatedness of processes. Identification of transcription factors which bind to DNA promoter regions is relatively straightforward given the relatively conserved nature of families of these domains. There are also many plant transcription factor databases, which contain annotation for wheat or closely related species such as the grasses. Those integrated in this project are described within Chapter 3. The annotation of transcription factors in a microarray dataset, such as TRITIMED, can reveal the quantity of regulation and the families of genes involved. This could have been extended to inference of PGI networks, using co-expression in the TRITIMED data. However, this is not included as part of this research project, given the lack of genomic sequence data to identify cis-binding regions (Su *et al.*, 2010), and the large time interval between measurements in TRITIMED.

Analysis of the TRITIMED microarray data required statistical analysis to identify the genes that have significantly changing expression with regard to comparisons across the independent variables (time and cultivar). Statistical analysis was also used to identify the major trends in expression. Subsets of genes with statistically significant changes in transcript levels were statistically dissected, such as from groups in two-way-ANOVA (*e.g.* genes with interaction between independent variables), or from the greatest contributors to a given coordinate of variation in principal coordinates analysis. The statistical analysis was undertaken without regard to the potential gene function, so annotation of these subsets would be important in revealing the biological mechanism that underlies the observed changes in gene expression. The questions that were of in-

terest included whether genes contributing to a principal coordinate of variation within an experiment represented the activation of a particular process (*e.g.* Protein synthesis) or whether expression of genes with significant interaction between independent variables may be attributed to the action of a process. This is particularly important in this experiment as genes that have significant interaction over time and between cultivars may reveal processes that respond to drought and are differentially regulated between cultivars.

The analysis of the TRITIMED experiment therefore is highly dependent on the coverage and quality of sequences on the Affymetrix wheat array. A high coverage of quality annotations on the chip could reveal previously unknown processes active in the plant drought response. It also increases the probability that the analysis will identify the processes that are most responsible for the differences in yield stability between the varieties. For this reason the major effort described in Part I of this thesis was the development of methods to identify the annotations for biological functions and processes for the gene transcripts represented on the Affymetrix chip, and for verifying the quality of these proposed annotations. The predominant methods for Affymetrix GeneChip annotation make use of multiple sources of annotation information. It is essential, therefore, that the integration process combines information accurately and consistently for subsequent use in the annotation pipeline. Data integration, is however, recognised as a major challenge for the life sciences in general and no pre-eminent bioinformatics methods have emerged that solve it completely. Rather than develop an ad hoc point solution for this particular project it was considered more appropriate to build on earlier work at Rothamsted Research that created the Ondex system (Köhler *et al.*, 2006) as a general data integration framework for use in systems biology projects. As well as providing significant useful pre-built functionality, the general approach used by Ondex made it possible to study aspects of data integration processes and later evaluate them for their contribution to the annotation process. The general problems

of data integration and the approach used by Ondex, including the extensions developed to fulfil the requirements of this thesis, are introduced in Chapter 2. The remaining chapters in Part I of this thesis address the important topics of using integrated data sources to develop a new approach to gene function annotation and the evaluation of this new method alongside other annotation pipelines (Chapter 3).

Chapter 2. A data-integration framework for sequence annotation

The annotation of newly obtained gene sequences is a general problem in bioinformatics. It entails elements of both comparative sequence analyses and integration of functional annotation. Sequence comparisons allow the transfer annotations from genes in closely related *model* organisms. Data integration enables the collection and dissemination of potentially important functional information on genes from a large variety of resources. Data integration can enrich the more direct sequence-based methods. As has been described in Chapter 3.1.2(*a*), the research in this thesis has been motivated by the need to extract as much value as possible from the additional information sources because of the focus on the interpretation of transcriptome data from a partially-sequenced cereal crop species (wheat) for which relatively little direct annotation information can be found. In large part, the annotation problem in wheat arises because the wheat genome is so distant in evolutionary terms from the main model plant species *Arabidopsis thaliana* where most of the direct evidence of gene function has been obtained by the analysis of gene disruption and other functional genomics methods.

2.1 Aims and Objectives

The aims of this chapter are to:

- Describe the state-of-the-art in data integration.
- Outline the Ondex data integration platform.
- Describe the developments of Ondex needed to facilitate a functional annotation pipeline.

This will be achieved through:

- An overview of a cross section of tools and processes.
- A description of the Ondex architecture.
- A description of algorithms developed for workflow enactment and graph traversal, reduction, and query.

2.2 Introduction

Many studies and reviews have suggested that data fragmentation and heterogeneity can be a limiting factor in systems wide analysis of biological data sets (Köhler *et al.*, 2006, Lysenko *et al.*, 2010). Mochida and Shinozaki (2010) have recently suggested that further understanding of plant molecular systems for increasing crop yields will be dependent on integration of multi-omics data. Similarly, the prediction of functional annotations for new sequences are of-

ten limited by the quantity and quality of information available (Chapter 3). This is particularly true in comparative genomic approaches where predictions are dependent on sequence similarity to other proteins or genes of known or documented function. The success of these methods generally depends on the evolutionary distance, as determined from sequence similarity, between a new gene and a similar gene of known function. It is assumed that the more similar the gene or protein sequences the greater the probability they are functional orthologs. The diversity of functions within the sequences that are similar to the putative functional ortholog is also important, as it can reveal the degree of subfunctionalization within that gene family. Subfunctionalization can be indicative to the strength of the prediction, as gene families containing a greater diversity of functions as more problematic to predict function based on sequence similarity. For example: genes involved in the plant pathogen-host response have the greatest diversity of functions, given the high rate of evolution given that plants are sessile. Conversely, genes encoding enzymes that are part of central metabolism, like Glutamate Synthase, show a high degree of conservation even between evolutionary distant plants.

It has been demonstrated by Lysenko *et al.* (2010) that the functional information relevant to adding enriched annotations to the genes from plants species is distributed across many different databases. No single data source can therefore be relied upon to provide all potentially valuable annotation information. The use of data integration to build a knowledgebase of plant gene functional annotations was therefore a prerequisite for this project.

2.2.1 Data representation and integration

The primary goal of data integration should be to provide uniform access to a set of heterogeneous data sources (Calvanese *et al.*, 2009). This frees a user from

needing to locate the data, and understand the technical details how they are stored and accessed. Most importantly the user does not have to decipher the semantic idiosyncrasies of each data provider. This uniform access also enables computational queries and reasoning which can be composed against a single unified semantic schema. Data integration systems that have a focus on unified representations are referred to as global or mediated schemas.

Historically there are three types of integration strategy: link integration, view-based integration, and data warehousing (Stein, 2003). Link integration was spearheaded by the successful SRS system (Biowisdom, 2009), which linked together entities in databases based on shared unique names or accession identifiers. However, successful link integration is problematic because it does not account for semantic differences in concepts between databases, or in the ambiguity or inconsistency in the use of identifiers. This can lead to semantic drift when linking information on a given gene. Much has been made of the difference between view-based and warehouse-based data integration approaches (Sheth and Larson, 1990), however in practical terms, the most difficult challenges of both these data integration approaches are associated with creating and mapping information to a global schema. The choice of whether data are transformed and warehoused into a single repository, or accessed remotely via a federated system approach has been well studied and is generally considered a largely solved technical issue. Current research in data integration has therefore moved on, from a consideration of structural representation, to semantic representation and integration (Ziegler and Dittrich, 2004).

A unified semantic schema in practical terms, often takes the form of an existing or new ontology, hierarchy or controlled vocabulary developed according to the subject domain of interest. These can be described in a generic language like OWL (Group, 2009, Smith *et al.*, 2004), OBO (Smith *et al.*, 2007), or more often in a domain-specific description format. Incorporation of new data into a common semantic schema therefore consists of identifying how entries within a database

relate semantically to the defined concept-types and relation-types within the target semantic definition. These relationships can be explicitly stated if there is a defined ontology for a data-source, in the form of mappings between the originating and target ontologies. In the case of a data source whose structure does not reference any formal semantics, considerably more work is required to define the semantics and parse the data into an amenable form where the semantics can be analysed.

2.2.2 Data representations for biological data integration

Demir *et al.* (2002) helpfully observes that any representation of data is often a compromise between clarity, coverage and content. If we adopt a high-level approach to data representation that encompasses a wide coverage of different domains, we can isolate the data from the domain specialist and sacrifice the clarity of the data; we may also lose the ability to accurately represent the content. Conversely, data representations that emphasise data clarity are often limited in their coverage, and over simplify the domain. Data content captured by complex representations can be limiting in terms of both coverage and clarity (Figure 2.1).

There are several frameworks and exchange languages for modelling and integrating biological data. Pathway Analysis Tools for Integration and Knowledge Acquisition (PATIKA) is a two tier framework with a data integration server in the background and a simple web based service for querying the integrated data (Dogrusoz *et al.*, 2006). Data in PATIKA are stored as a graph based data model in a persistent Object Orientated (OO) database. Biological processes in PATIKA are represented as states and transitions between them. This is a process modelling representation that has early roots in computer science *e.g.* in the Petri nets formalisation as described by Choo (1982) and in

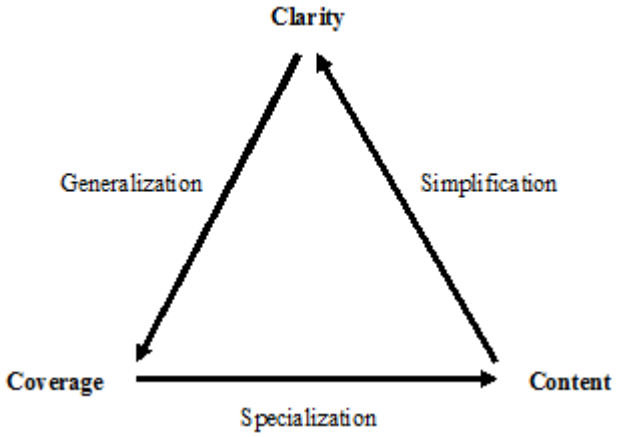


Figure 2.1: The compromises among Coverage, Content and Clarity when defining a data representation.

more recent time has been popular for developing quantitative models of biochemical pathways (Steggles *et al.*, 2007). In PATIKA a biological entity such as a protein may be represented multiple times within the data structure as distinct states, which are defined by an entity's transitions such as *phosphorylation*, *formation of a complex*, or *transport to a different cellular compartment*. Each state and transition is defined in a hierarchical tree based ontology (Figure 2.2) (Demir *et al.*, 2004). This data structure is powerful in describing cellular signal transduction pathways, but is complex and presents a significant challenge for communication with biologists as it diverges significantly from the now familiar pathway representations found in biological text books and the MetaCyc (Caspi *et al.*, 2008)and KEGG (Kanehisa and Goto, 2000) biochemical pathway databases. Another limitation is that the PATIKA data model is highly domain specific and could not be easily adapted to other non pathway information, and is exclusively tied to the PATIKA ontology.

The data integration platform BN++ uses the BioCore model, which is based on a PostgreSQL relational database and can represent metabolic pathways, transcription data, protein-protein interaction data, and signalling pathways (Sirava *et al.*, 2002). The BN++ platform is essentially a three-tier separation of integration, analysis, and visualization. Because the BN++ ontology is essen-

tially represented by UML and hard coded in its relational database, changes and expansion of the underlying ontology is problematic requiring expert technical knowledge of the system.

A recent warehousing data integration system is the PROtein Function, Evolu-

Figure 2.2: The hierarchical tree based relation Ontology in PATIKA, showing a subsection of the transitions ontology Dogrusoz *et al.* (2006).

tion, Structure and Sequence database (PROFESS) (Triplet *et al.*, 2010) database. PROFESS wraps a large number of protein related databases and makes them available via a single normalised SQL schema. A summary of the databases wrapped by PROFESS is shown in Figure 2.3. Users and applications can query the database via an SQL query, RESTful Web services, or a Web page.

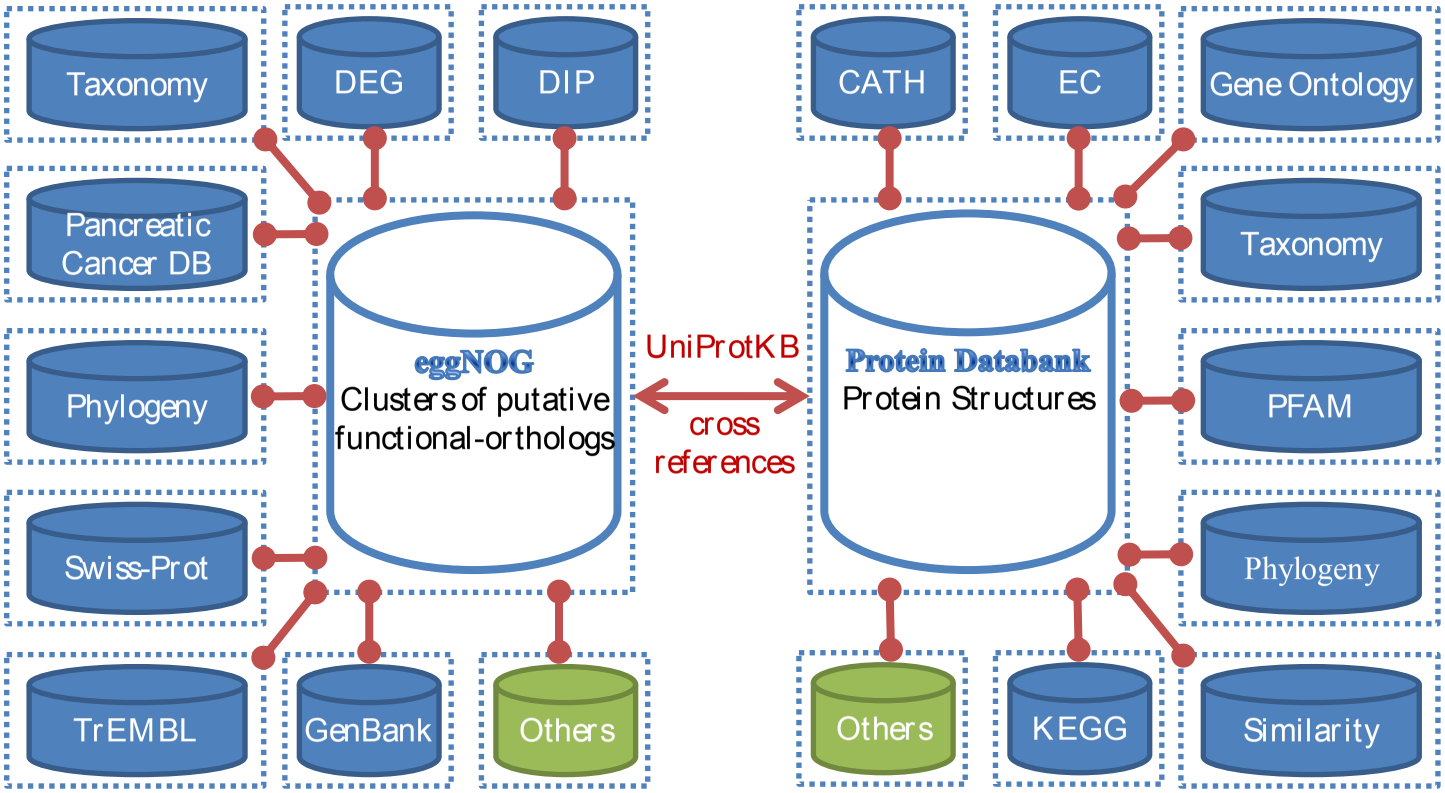


Figure 2.3: The PROFESS virtual relational data warehouse approach, showing all the databases wrapped by the framework. Broken lines around databases represent a wrapper service, red links show how the integrated databases reference each other. (Triplet *et al.*, 2010).

Other data integration systems of note are: VTT (Gopalacharyulu *et al.*, 2005), which uses a hybrid of a relational database and an XML server (allowing more flexibility in user contributed data) PRODORIC (Wingender, 2004), BioGRID (Stark, 2006), Biozone (Birkland and Yona, 2006a,b, Shafer *et al.*, 2006), and PathSys (Baitaluk *et al.*, 2006).

All of these approaches and systems share a common problem: they are not easily expanded to incorporate other domains, and a change to the semantic representation would require a major reengineering of the database. Furthermore, the semantics of the unified schema is not always explicitly stated and is dependent on the implicit semantic conversion encoded by the wrapper or parser code.

As well as data integration frameworks, a number of standardised information formats have been developed for use in the biological sciences. These standards facilitate the exchange of information, as well as providing a common target schema for integration. Their widespread adoption and Application Programming Interfaces (API)s represent a significant advantage over proprietary representations of integration frameworks. For the representation of biochemical pathways there are four main formats: SBML (Nishimura *et al.*, 2009), CellML (Miyazono *et al.*, 2009), PSI-MI (Hermjakob *et al.*, 2004) and BioPAX (BioPAX, 2011). SBML and CellML are modelling languages used for describing biochemical pathways with direct links to simulation tools. They describe the same process of states and transitions captured in the PATIKA system. The other data standards are primarily concerned with representing interactions between biological entities (mostly proteins) rather than how quantities of entities (typically metabolites or gene transcripts) change in a dynamic process model.

Systems Biology Mark-up Language (SBML) is a machine-readable format (XML) for representing models of cell signalling pathways, metabolic pathways, biochemical reactions, gene regulation, and many more dynamic processes. SBML optionally allows the referencing of other ontologies such as the Systems Bio-

logy Ontology (SBO) (Le Novère, 2006) and BioPAX. The usefulness of a given SBML file for data integration therefore depends upon the implementation. CellML is also an XML-based language that can represent systems of differential algebraic equations that model how quantities (*e.g.* of metabolites) flow through a process.

The Proteomics Standards Initiative Molecular Interaction (PSI-MI) format was developed for representing the experimental evidence that supports observations of molecular interactions. It is less expressive than SBML and BioPAX and lacks the capacity to describe inheritance hierarchies of entity and relation types. It has, however, been used by many databases providing molecular interaction data. The Biological PAthway eXchange (BioPAX) language (Demir *et al.*, 2010) is an expressive exchange format for describing biological pathways. It uses OWL Web ontology language for describing pathways, and therefore aims to be compatible with the Semantic Web framework for data integration (Ruttenberg *et al.*, 2007). Figure 2.4 shows the BioPAX schema for representing pathways, interactions and biological entities. The breadth of the language means is capable of expressing information stored as PSI-MI, and much of SBML.

Visualisation and abstraction provides a way of accommodating complex representation without sacrificing clarity and flexibility to the user. Data can be natively stored by a data integration framework in a complex representation and yet still made accessible to the end user through a visual abstraction to a more understandable form. An excellent example of this is Systems Biology Graphical Notation (SBGN), which allow the abstraction of SBML into three distinct graphical notation languages with associated visualisations. Its goal is to provide a standard format that is capable of describing a wide variety of models. Process Description Language (PDL) allows a user to see the temporal flow of biochemical interactions in a SMBL network. Entity Relationship Diagram (ERL) shows a relationship centric view, which represents the interactions of an entity independent of time. Activity Flow Language (AFL) focuses on the

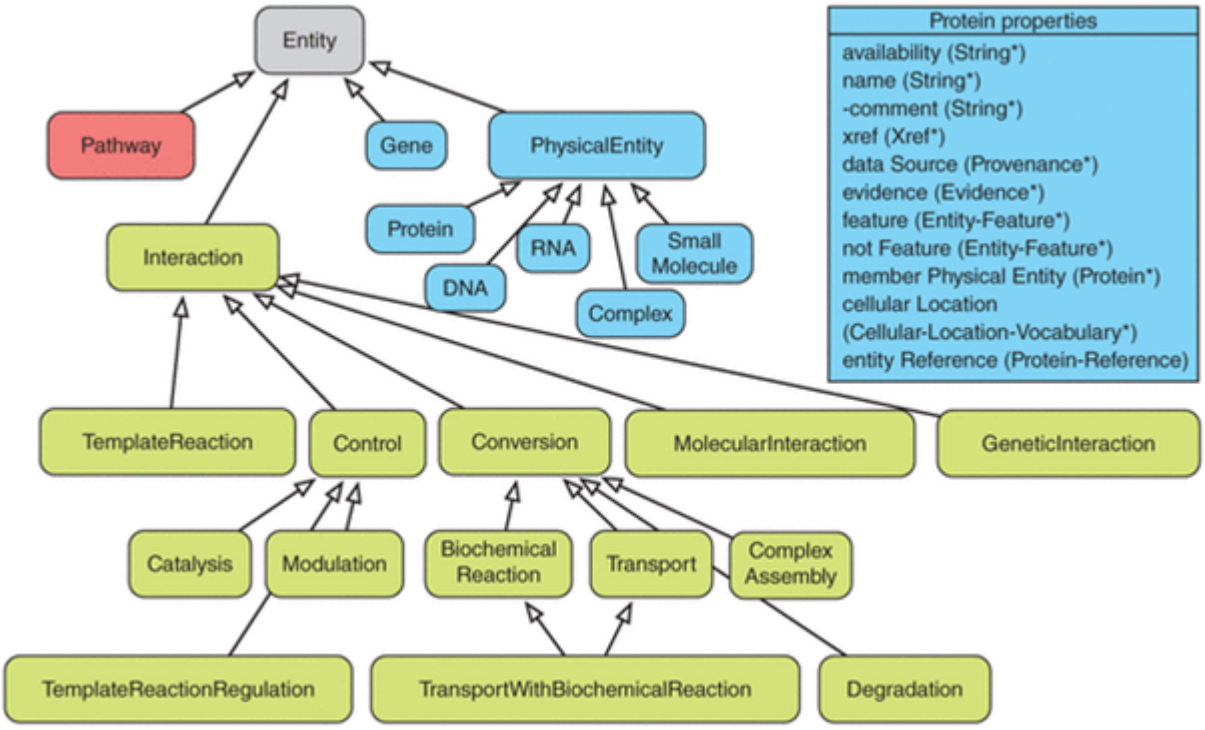


Figure 2.4: The BioPAX owl ontology. Classes are shown as boxes and arrows represent inheritance relationships. The three main names of types of classes in BioPAX are "Pathway" (red), "Interaction" (green) and "PhysicalEntity" and "Gene" (blue) (Demir *et al.*, 2004). An example of properties attached to an instance of a Protein is given top right.

flow of information between biochemical entities in a network. An example of a PDL representation of the insulin-signalling pathway created from an SBML model is shown in Figure 2.5.

Recently, Semantic Web (Ruttenberg *et al.*, 2007) approaches to data integration have become increasingly widespread. The term Semantic Web is was originally coined by (Berners-Lee and Hendler, 2001) to refer to their vision of how the world-wide-web should evolve to be fully machine-readable. The Semantic Web was defined with reference to three layers of enabling technologies that express the evolution of the Web (Figure 2.6), the first of which is the existing HTTP and HTML standards on which the Web was founded. The second enabling technologies of Semantic Web were standards which allowed documents to be self describing. This included eXtensible Markup Language (XML) and the Resource Description Framework (RDF). The third layer describes the cap-

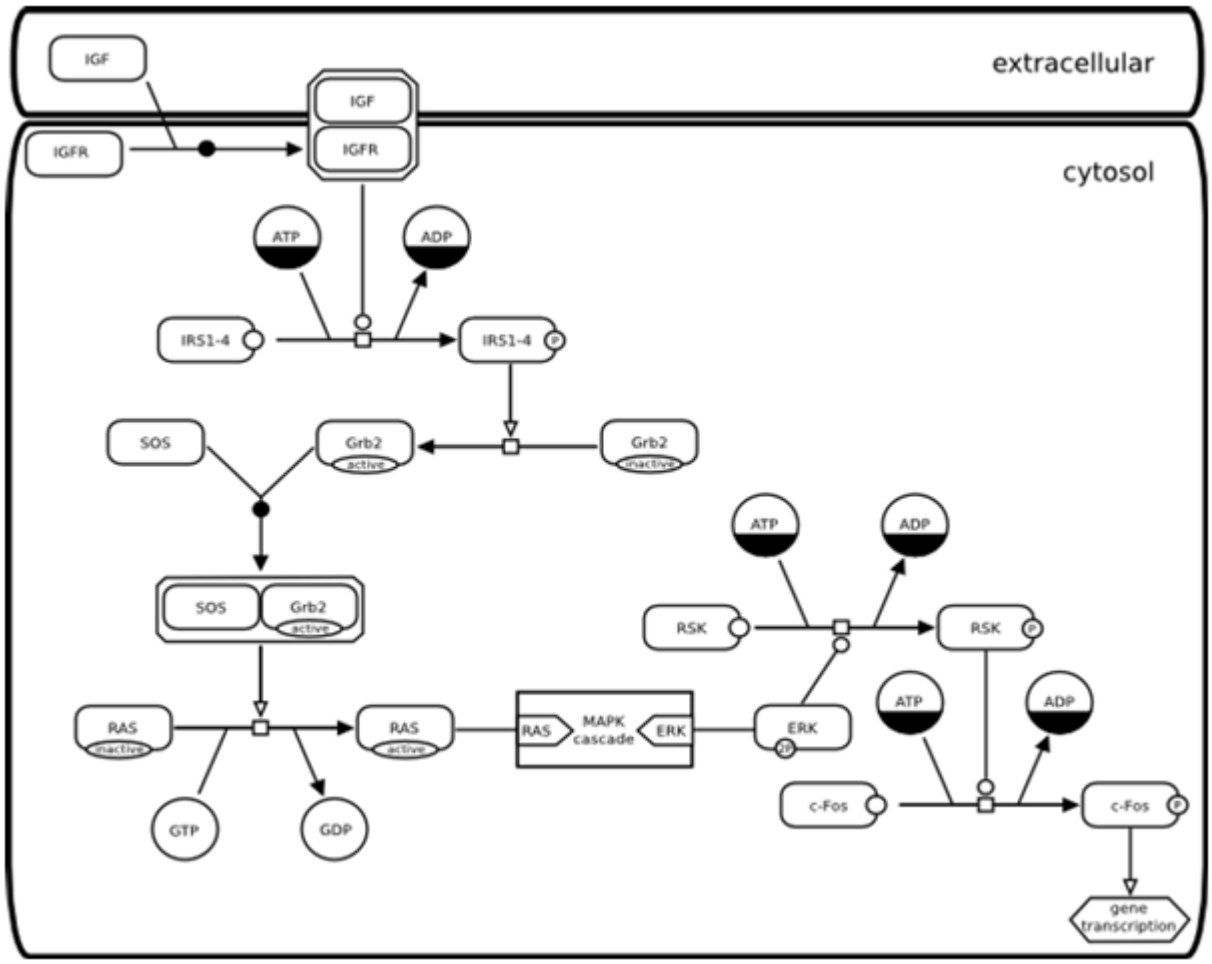


Figure 2.5: Systems Biology Graphical Notation (SBGN) rendering the insulin signalling pathway (SBGN, 2011).

ability to link documents through shared terminologies and ontologies. Here the term Semantic Web is used in a more constrained sense to refer to the set of RDF based technologies, which together with Web Ontology Language (OWL) have been used for data integration.

At the heart of the Semantic Web is the RDF (Beckett and McBride, 2004) language which is a framework for describing knowledge as a directed-labelled-graph of triples. Statements in an RDF graph are composed from subject-predicate-object triples, where the subject of the statement identifies the target of the statement, the predicate describes the trait to be described, and the object refers to the predicate property that the given subject has. Using a Uniform Resource Identifier (URI) allows each of the three elements of the triple to refer to a con-

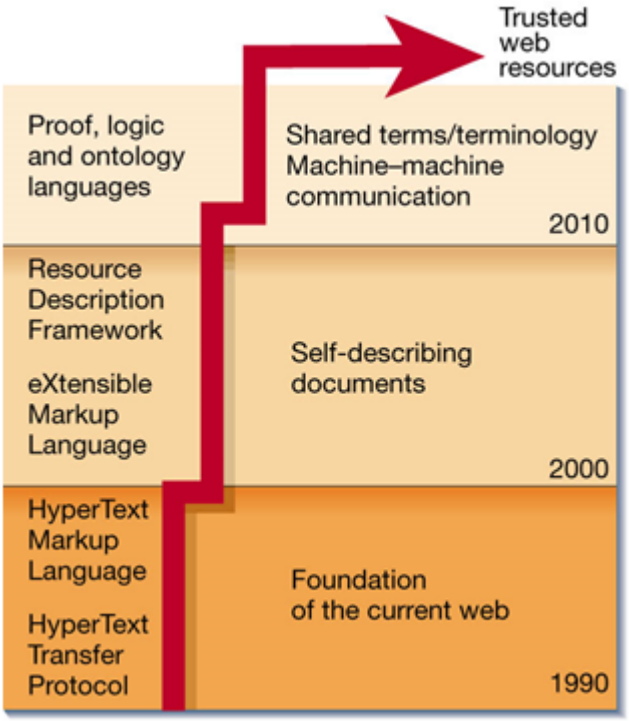


Figure 2.6: The layer cake of enabling technologies for machine readable documents, which evolves into a semantic Web (Berners-Lee and Hendler, 2001).

trolled vocabulary or preferably formal ontology definition. If two triplet statements reference the same URI, that means they refer to the same subject matter. In this way complex attributes and relationships for an entity can be described, using a consistent language and vocabulary that is applied across many entities. Defined controlled vocabularies and ontologies are therefore an essential prerequisite for unifying RDF documents. These are provided by the RDF Schema (RDFS), Simple Knowledge Organisation System (SKOS) (Isaac and Summers, 2009), and the Web Ontology Language (OWL).

OWL is an essential part of Semantic Web and has two major versions and a number of sub-languages. For the first version of OWL these were OWL-Lite, OWL-DL, and OWL-Full (Smith *et al.*, 2004) OWL-Lite is a subset OWL-DL that is restricted to cardinalities of 0 or 1 on constraints for attributes of concepts. OWL-DL is a subset of OWL-Full that allows the expressiveness of description logic but limits this to elements that are computationally complete and decidable. OWL-Full has maximum expressiveness with no computational guaran-

tees. In OWL2 there is a similar categorisation of 3 sublanguages according to syntactic restrictions each of which is more restrictive than OWL-DL: OWL2-EL, OWL2-QL, and OWL2-RL (Group, 2009). OWL provides formal semantics for defining individuals (*e.g.* the T-cell surface glycoprotein CD4) and their properties (AA sequence length = 458). Axioms are declared on classes of individuals ((*e.g.* glycoproteins), which define the types of relationships permitted between them. These axioms can then be used as a basis for reasoning.

Using RDF to describe knowledge with reference to shared ontologies is the fundamental strength of the Semantic Web approach and potentially enables compliant data to be queried across in a unified way. SPARQL is a language for defining queries across triples (Prud'hommeaux and Seaborne, 2008). It is capable of expressing required and optional graph patterns, as well as conjunctions and disjunctions.

2.2.3 Data representation and integration in Ondex

The Ondex data integration system is a warehousing based approach that relies on the semantic transformation through parsers of imported data to be compliant with a unified schema, formally declared in the Ondex metadata. The Ondex metadata can be defined by the user, however the recommended method is to use or extend the global schema defined by the Ondex development community. This enables interoperability between instances of Ondex warehouses. The Ondex metadata is composed of a simple hierarchy, with single inheritance, where classes within the hierarchy are considered sub-classes of their ancestors.

An Ondex data warehouse is therefore composed of two graphs: a directed graph and a tree. The directed graph describes data in terms of instances of concepts (represented by nodes) interconnected by directed relations (repres-

ented by edges), each of which have collection of associated name-value pair attributes. These concepts and relations reference a tree structured controlled vocabulary of metadata, which defines and describes concept-classes, relation-types, and attribute-names. For example: Figure 2.7 shows how the information "The enzyme Xanthoxin Dehydrogenase is encoded by the gene ABA2" can be represented in Ondex. The tree structure of the metadata allows Ondex to represent the rule that all Enzymatic proteins are also a type of protein. This allows enables queries based on more general terms, such as protein, rather than specifying every type protein. A user may import a protein-protein-interaction (PPI) database, such as IntAct (Aranda *et al.*, 2010), which does not contain explicit knowledge indicating that a given protein is an enzyme. It is still possible to identify that the enzyme *Xanthoxin Dehydrogenase* is identical to the protein of the same name in the IntAct database because the concept-class of one inherits from the other. A tree based metadata system is considerably less expressive than referencing an OWL or OBO ontology, as it does not permit any other relation than *is a* or allow for representing more complex relations where a concept-class inherits from multiple classes. However, a hierarchy represents a middle ground in complexity between a full ontology and a simple controlled vocabulary.

Ondex in relation to an RDF based representation has a number of similarities and differences. Two Ondex concepts connected by relations could be represented as an RDF subject-predicate-object triple. Similarly, attributes on concepts can be converted into a triple representation. Representing attributes on relations requires reifying Ondex relations into a new concept representing a relation and which is linked via two new properties. Attributes can then be connected via triples to the reified relation. In this way, Ondex and RDF representations are interchangeable, and import and export capabilities exist in Ondex to achieve this. However, by design there is no requirement for a concept in Ondex to be identified by a unique URI. A single concept may be represen-

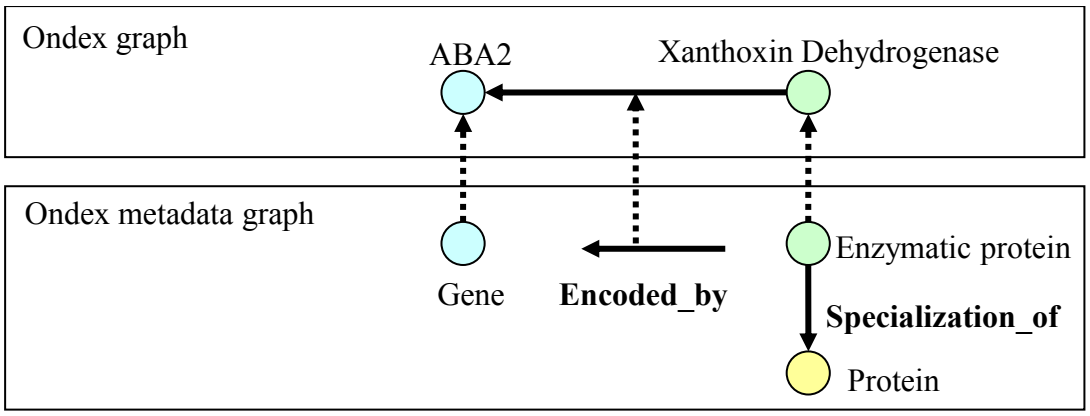


Figure 2.7: The Ondex data representation: concept and relation instances reference data types in the metadata. An enzyme taken from ABA biosynthesis is used as an example to show its instance representation in the Ondex graph, and the type references in the Ondex metadata.

ted multiple times within an Ondex graph, as concepts are redundantly created from each new data-source. This allows equivalent concepts to be connected through an *equivalent concept* relation at a later point, and retained indefinitely for provenance tracking, or collapsed into one abstract concept to simplify visualisation.

2.2.4 Ondex: an integration framework

The Ondex data representation forms the core of an integration and visualisation framework, which enables a user to build integration workflows, load pre-integrated knowledgebases, queries and analyses the knowledgebase with visualisation using graph layouts and overlays of quantitative data. At the heart of the Ondex framework philosophy is community development, and an object-oriented modular architecture that exposes a Java Application Programming Interface (API), which ensures plug-ins for parsing, export and analysis can be independently developed and incorporated into workflows. Ondex is implemented as a series of layers (Figure 2.8): minimizing complexity

and ensuring plug-ins are also straightforward to implement. The Ondex data storage structure is abstracted above the database engine. Therefore, any database can be used to store an Ondex Graph if an appropriate interface is written. Workflow plug-ins can also be implemented irrespective of the data storage engine. Currently data storage implementations exist for the Object Oriented (OO) transactional database BerkleyDB (Oracle, 2011), a relation format implemented using MySQL, and a fast Random Access Memory (RAM) storage based on hash-tables. For fast searching of attributes on the graph, which is a requirement for integration methods and text mining, a attribute search API provides an additional layer of abstraction on the Ondex graph API, the text search engine library Lucene (Foundation, 2011) currently implements this functionality. A graph query API was created as part of this project, and provides rule-based querying of the Ondex Graph API. The methodology for this is described in Section 2.2.4. Additionally a Decypher (TimeLogic, 2011) API was implemented as part of this thesis. It allows Ondex API concepts with sequence attributes to be submitted to Decypher Field Programmable Gate Array (FPGA) hardware, which implements hardware accelerated algorithms for the common BLAST and HMMR sequence tools.

The first stage of data integration in Ondex is parsing (Figure 2.9). This in-

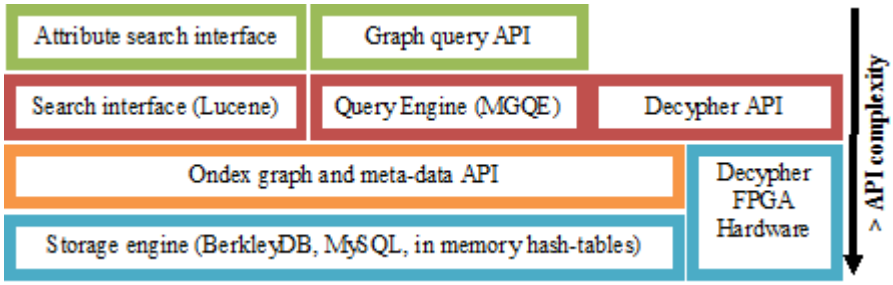


Figure 2.8: The major API layers within Ondex in order of increasing complexity. APIs are dependent on those beneath them in the stack.

volves the technical transformation of the data from its native format to the Ondex data format, and the semantic transformation of data to conform to the

Ondex controlled vocabulary, defined in the metadata. Each source of data is transformed into its own graph which is loaded into the Ondex structure. Parsers are written as plug-ins, which can be developed independently from the underlying Ondex backend source code by using the Java APIs. After databases have been parsed into the Ondex data format, the second stage is usually to identify equivalent concepts within the set of graphs parsed into the Ondex knowledgebase. It is expected that there will be redundancy among concepts and so an important step in the integration process is to identify the equivalent concepts among multiple graphs. The methodology for identifying equivalence depends on the nature of the data and in many instances finding equivalence using accession identifiers is sufficient. However, more advanced plug-ins are available that use fuzzy matching and stemming of names, alignment of similar graph motifs, or comparisons of attributes such as sequence using BLAST algorithms. For the data required for the applications described in Chapter 3, accession and sequence-based matching was sufficient. After equivalent concepts have been identified in the Ondex graph, subsequent workflows steps will depend on the needs of the application. In the example workflow described in Figure 2.9, a comparative alignment plug-in is used to create links between FASTA sequence concept and proteins in other databases. The resulting links and annotations are extracted by a query plug-in, and then exported as a tab-delimited file. A suite of generic plug-ins are provided in Ondex, and some of these have been described by Köhler *et al.* (2006) and (Taubert *et al.*, 2009). A full list of self documented plug-ins is available within the integrator menu option when loading the Ondex software.

By referencing a common Ondex data structure and semantics, integrated data can be used to both aggregate existing knowledge and infer new relationships in the data. The workflow described in 2.9 produces the Ondex meta-graph shown in Figure 2.10 by stage three. The Ondex meta-graph representation uses the standard Ondex graph layout methods to visualise the metadata con-

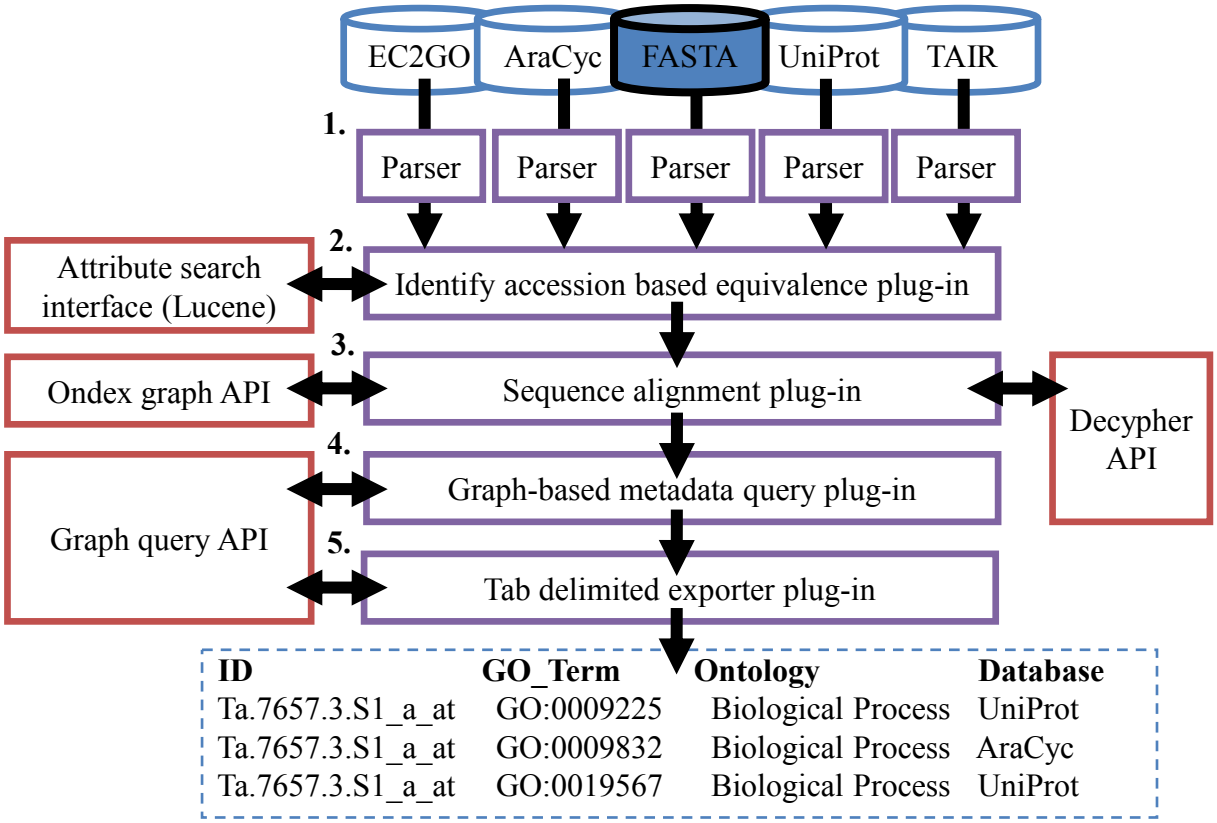


Figure 2.9: An example of a simple Ondex workflow to annotate novel sequences in a FASTA file. Separate plug-ins enacted by the Ondex workflow parse different databases into the Ondex graph format. The stages in the integration process are (1), identifying equivalent concepts across data sources based on accession (2), apply comparative genomics using sequence alignment (3), extracting an annotation pattern from the graph (4), and exporting the results for further analysis (5). Also shown are the APIs (red) used by each plug-in (purple) to access data in Ondex, or specialist hardware.

tent of the graphs in the knowledgebase. The nodes in the meta-graph represent concept classes and the edges represent the relationship types. In Figure 2.10 The *Arabidopsis* Information Resource (TAIR) database provides gene and protein annotations to GO and EC terms (Garcia-Hernandez *et al.*, 2002). UniProt contains GO and EC annotations but only to proteins, and is a rich source of accession cross references. AraCyc provides EC annotation of gene loci. EC2GO provides mappings from EC terms to GO functions and processes. Together the union of these GO functional annotations for proteins and genes provide an aggregation of existing knowledge. Using EC2GO mappings also enables AraCyc protein-EC annotations to be linked to GO annotations, by inferring that

the two relationships {protein, has class, EC} and {EC, has function, GO term}, are equivalent to {protein, has function, GO term}.

In order to enact sequential plug-ins such as shown in Figure 2.9, a way of

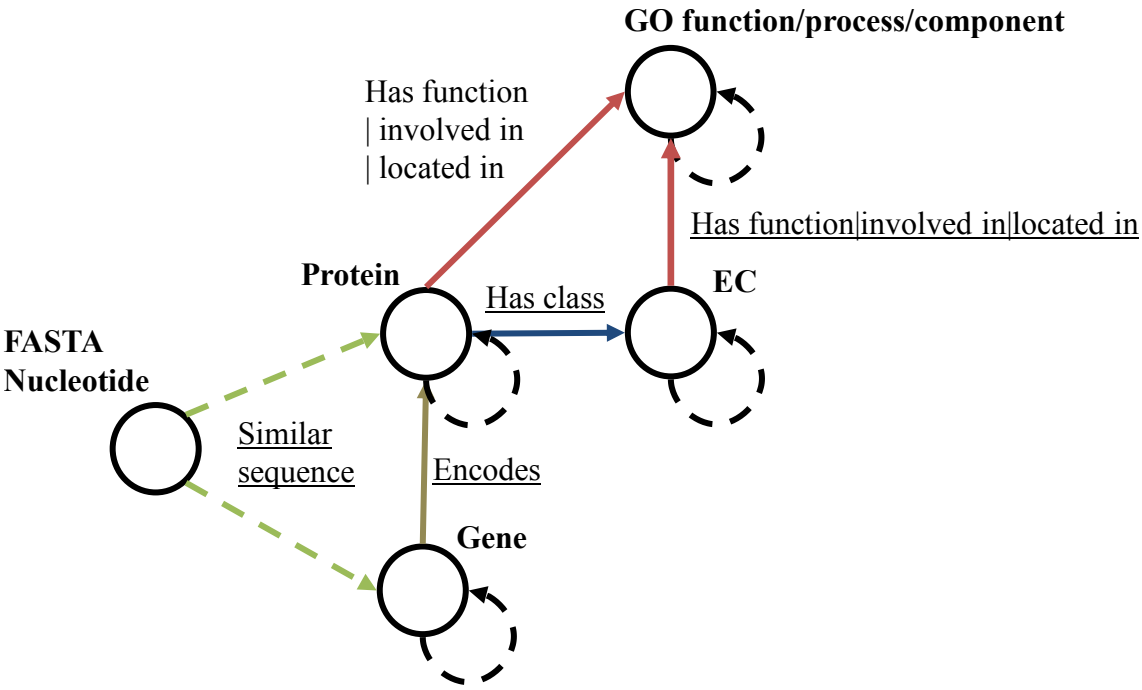


Figure 2.10: A subset of the meta-graph produced from the integration of the five databases in Figure 2.9. The circles represent concept classes and arrows relation types. The names in bold are classes of concepts and underlined names relation types. The broken black lines represent equivalence relations identified from accessions, solid coloured lines show relations imported from databases and dashed green lines are relations inferred from sequence similarities of attributes using BLASTx and BLASTn algorithms.

formally defining and sharing workflow descriptions was required. This was developed as part of this thesis and is described in Section 2.3.1. Additionally the process of extracting existing knowledge about a concept within an Ondex graph such as Figure 2.10 requires the ability to express queries relating one concept-class to another via a series of rules. For example for the graph in Figure 2.10 to achieve stage 4 of the workflow in Figure 2.9, it requires a query that can define a path from each query sequences, via genes, proteins and EC terms, to GO terms. This problem is addressed in Section 2.3.4 in the form of a Metadata-Graph based Query Engine (MGQE).

2.3 Methods

This section contains a description of the four key developments, which were a prerequisite to utilising Ondex for this research project. They form the basis for the annotation pipeline that will be described in Chapter 3 and address the requirements that have been outlined in this Chapter. (1) The development and architecture of an Ondex workflow enactor is described, which allows plug-ins to be sequentially executed on a shared Ondex graph. (2) This section also outlines the architecture of a parallel neighbour-hood-search library for Ondex. These methods forms a building block for (3) a graph-reduction algorithm for merging equivalent concepts, and (4) a Metadata-Graph based Query Engine (MGQE), which supports both general queries and simple inference methods.

2.3.1 Executing sequential processes in Ondex – a workflow enactor

Prior to this research project, all workflows in Ondex required hard coding via direct calls to the Java API, and failure to correctly specify parameters resulted in late failures at the point a plug-in failed. In order to provide a flexible framework, where workflows could be stored, shared and reproduced, a XML based file format was introduced. This lightweight format allowed a user to initialise Ondex graphs using specified database storage layers, and define the sequential enactment of plug-ins on a Graph. A structured system for defining the arguments that a plug-in required to run was also introduced. This enabled the a priori validation of Ondex workflow XML, and fail fast workflow management.

Figure 2.11 shows an example of a simple XML workflow that creates an On-

dex representation of UniProt, which is linked to the ExPASy enzyme database by finding equivalent EC terms based on the EC accession. ExPASy contains both Protein-EC term annotations, and the complete EC nomenclature hierarchy. The workflow described executes four separate plug-ins to achieve this. The steps are as follows and reference the numbering scheme within Figure 2.11:

1. A named Ondex graph is initialised using a memory-resident hash-table representation as the resulting graph is likely to be small.

   ReplaceExisting this parameter is redundant in the instance of a memory graph and can be set to true. It allows persistent graphs to be reloaded or replaced.

2. The ExPASy Enzyme database is parsed from the two flat-files enzclass.txt and enzyme.dat, which can be found at ftp://ftp.expasy.org/databases/enzyme the ExPASy ftp site.

   InputDir this specifies the directory which must contains the two flat-files

3. A UniProt XML file (The UniProt Consortium, 2010) is parsed into Ondex. This can be obtained by selecting a species of interest from the UniProt taxonomy databases, following the link to the full protein set, selecting the download link, then downloading the XML file .

   InputDir this specifies a directory from which to import all UniProt XML files

   Context information creates labelled sub-graph sets of proteins based on EC terms

   HideLargeScaleReferences excludes review like literature references ref-

erencing many proteins.

GoFile the latest Gene Ontology in version 1.2 OBO format

4. A plug-in that identifies equivalent EC concepts based on the EC accession, and creates a equivalence relation between them.

   ConceptClassRestriction specifies which concept-classes to attempt to find equivalence within

   CVRestriction determines the type of accession to match

   IgnoreAmbiguity when true it includes accessions flagged as ambiguous. As we are using EC terms on EC concepts this should never be the case.

5. Exports the whole graph using the Ondex data exchange format files (OXL), encoded in XML (Taubert *et al.*, 2007).

   ExportFile the file name of the OXL file to create

Figure 2.12 shows the design of the workflow enactor that takes a workflow XML file such as shown in Figure 2.11 and after verifying that the arguments are valid and complete, sequentially executes the plug-in elements. The workflow enactor is intrinsically linked with the plug-in architecture for detecting and loading plug-ins, which was developed together with fellow student Artem Lysenko. The plug-in argument definitions, required as part of the plug-in API, contain methods for verifying that the arguments for the plug-in are complete, and of the correct scope and type. For example: a plug-in can require as a parameter a single real number between 1 and 10, a list of URIs corresponding to files, or a collection of name value pairs. More complex objects can also be provided as serialised XML parameters which are converted to Java objects using the XStream library prior to validation by the plug-in (Walnes and Schaible, 2011). The philosophy of the workflow-enactor and plug-in registry is to fail

```
<?xml version="1.0"?>
<Ondex xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="ONDEXParamiters.xsd">

1.  <DefaultGraph name="ec_and_uniprot" type="memory">
        <Parameter name="ReplaceExisting">true</Parameter>
    </DefaultGraph>

2.  <Parser name="ec">
        <Parameter name="InputDir">importdata/uniprot12Feb2010</Parameter>
    </Parser>

3.  <Parser name="uniprot">
        <Parameter name="InputDir">importdata/uniprot12</Parameter>
        <Parameter name="ContextInformation">false</Parameter>
        <Parameter name="HideLargeScaleRef">true</Parameter>
        <Parameter name="GoFile">gene_ontology.1_2.obo</Parameter>
    </Parser>

4.  <Mapping name="lowmemoryaccessionbased">
        <Parameter name="ConceptClassRestriction">EC</Parameter>
        <Parameter name="CVRestriction">EC</Parameter>
        <Parameter name="IgnoreAmbiguity">true</Parameter>
    </Mapping>

5.  <Export name="oxl">
        <Parameter name="ExportFile">ec_and_uniprot.xml</Parameter>
    </Export>
</Ondex>
```

Figure 2.11: An example of a simple XML defined workflow. (1) A memory graph is created. (2) The EC hierarchy is parsed from ExPASy flat files. (3) UniProt is imported from an XML file with assistance from the latest GO OBO format definition. (4) An accession based mapping plug- optimised for low memory usage is then run to create links between EC concepts based on the EC accession. Finally (5) the graph is exported in the OXL format.

fast. This is essential as plug-ins within data integration workflows can be time consuming, it is therefore desirable to fail before the workflow plug-ins are executed.

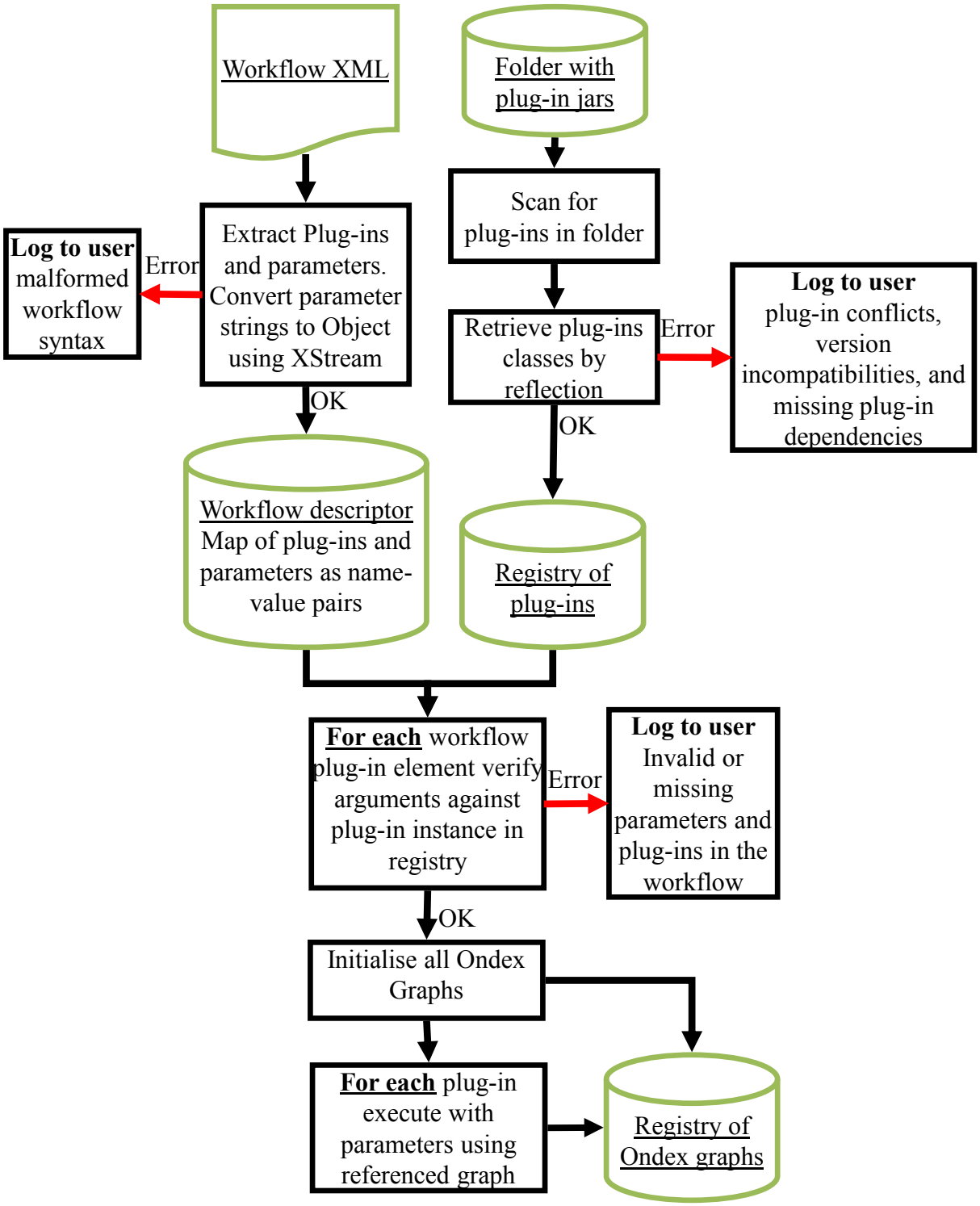


Figure 2.12: An overview of the information flow within the workflow enactment process. Black boxes are processes, and green boxes are information stores. The registry which detects and initialises plug-ins is also included because of the role of plug-ins in validating arguments.

2.3.2 Parallel connective sub-graph search

A requirement of the plug-ins that will be described in Sections 2.3.3 and 2.3.4, is the ability to find the sub-graph of nodes and edges within a graph that connect with a query-concept. In this search, the user should be able to define how nodes and edges in the graph should be traversed from using a rule-based approach.

A generic library was created to facilitate traversal of a connecting sub-graph using a breadth-first-search similar to that of Lau (2007), which supports parallelisation. Figure 2.13 shows an information flow diagram of the procedure. An Ondex graph with a labelled query concept forms the starting point. The traversal procedure (1) then identifies all the relations connecting the query concept that have not previously been encountered (recorded as a set of relations). These are then broken into candidates for traversal, which are composed as a triple in the form source concept, relation, target concept. This set of triples is placed into a work queue that is sorted according to depth (relations from the query concept). A thread-pool continually removes triples from the head of the queue (2) and passes them to a user-definable function, which evaluates if the triple may be traversed. For example Figure 2.13 shows the implementation of a simple rule to extract connected PPI data. If the user defined function returns true, then process (2) writes the verified triple into the results sub-graph, and applies the traversal procedure (1) to the target element of the triple. The thread is then returned to the thread-pool to continue processing jobs from the queue. When all threads are inactive in the thread-pool and the queue is empty, the resulting sub-graph is returned to the user.

When a user submits a large number of query concepts to be interrogated, then an alternative thread-pattern is executed. The procedure described in Figure 2.13 runs under a single thread but multiple procedures are executed in paral-

lel. Experience has shown this is more efficient where a moderate number of queries are required, as the thread related synchronisation costs are reduced.

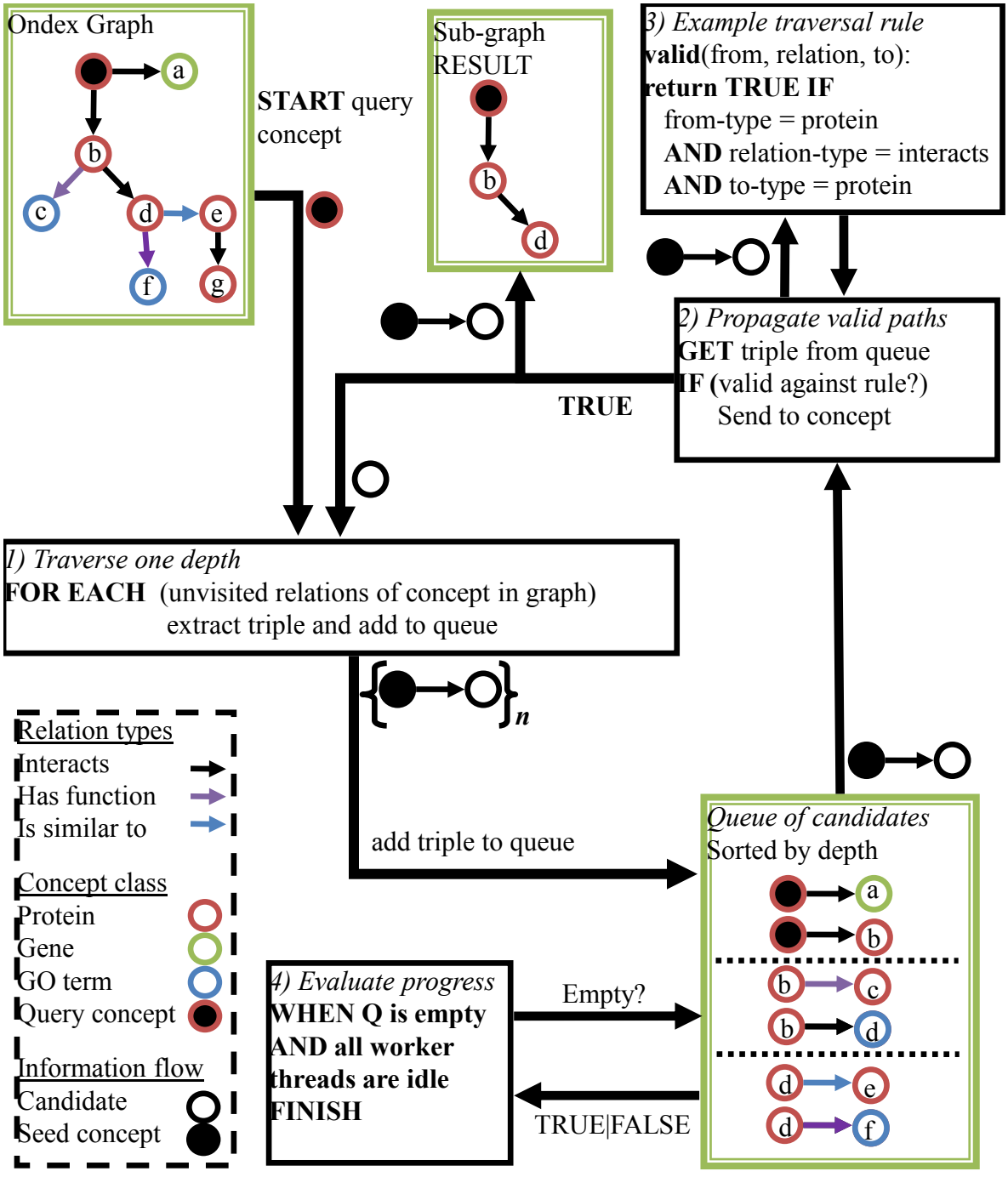


Figure 2.13: An information flow diagram of the sub-graph query search algorithm. The starting point for the algorithm is an Ondex graph with one concept labelled as the target of the query. The result is a sub-graph containing the query concept and any connecting nodes and edges that confirm to the traversal rule.

2.3.3 Graph transformation to remove redundancy

The integration workflow, to be described in Chapter 3, requires the integration of information from large sequence databases containing information about a large number of plant species. Given that data are parsed to create redundant concepts and relations in Ondex, the resulting graphs are expensive to store and slow to traverse. Information is also redundant at the attribute level, which is a particular problem with sequence data because of the data storage requirements. The query and traversal of such large graphs with millions of concepts, and tens of millions of relations is often not practical. Therefore, a requirement of this research project was to develop a method to remove concept redundancy within the graph, without losing information on provenance. Provenance is a particular concern because imported concepts are labelled with the database they originated from together with information about supporting evidence, such as experimental-type. In order to properly evaluate the contribution of information sources to the final annotation predictions (Chapter 6.3), and assign confidence values to predictions, provenance and experimental information must be retained within any non-redundant graph. The relation-collapse algorithm was developed to meet these needs.

The *sub-graph collapser* plug-in contains an algorithm for iteratively collapsing groups of equivalent concepts, connected by an *is a* relation, in an Ondex graph. Ondex parsers create redundant nodes and edges in the graph where the information in two or more databases intersect. These are mapped together through equivalence relations. The collapsing of equivalent concepts can be achieved by executing the *sub-graph collapser* plug-in, with a parameter defining the relation type that defines equivalence in the pipeline. An additional parameter allows attributes of concepts from collapsed concepts to be copied to the new super node.

The first step in the algorithm is the identification of clusters of fully equivalence-connected concepts in the graph. This was done using the algorithm described in Section 2.3.2. Once the clusters have been identified, they are iteratively collapsed by the creation of a super-concept to replace the cluster. Attributes of concepts within the cluster are transferred to the new super-concept and relations that connect the elements of the cluster to external concepts are reassigned to the new super-concept. Provenance is retained by concatenating data sources on the new super-node. Similarly, evidence and attributes are merged into a list on the new super-node or transferred to the new externally connecting edges. However, some information is lost through this approach, as it is no longer possible to attribute a particular source of evidence to a specific database. A work around of this, employed in this thesis, was to store a data-source and evidence pair as an element in the list. However, a more long-term solution would require changes in the Ondex data-structure, to accommodate this. Finally, all concepts in the cluster (excluding the super-node), together with all their connecting relations are removed from the graph. Three example clusters in Figure 2.14 (A) are shown being iteratively collapsed through (B) to (D).

2.3.4 Implementing a Meta-data based Graph Query Engine (MGQE)

The extraction of gene annotations, from an Ondex integrated graph, will be described in Chapter 3. A requirement for this process is an algorithm that can extract paths in the graph from a starting concept to a target annotation, using a priori encoded biological rules. The basis for such an algorithm was introduced in Section 2.3.2. However, at this stage, two important requirements are missing:

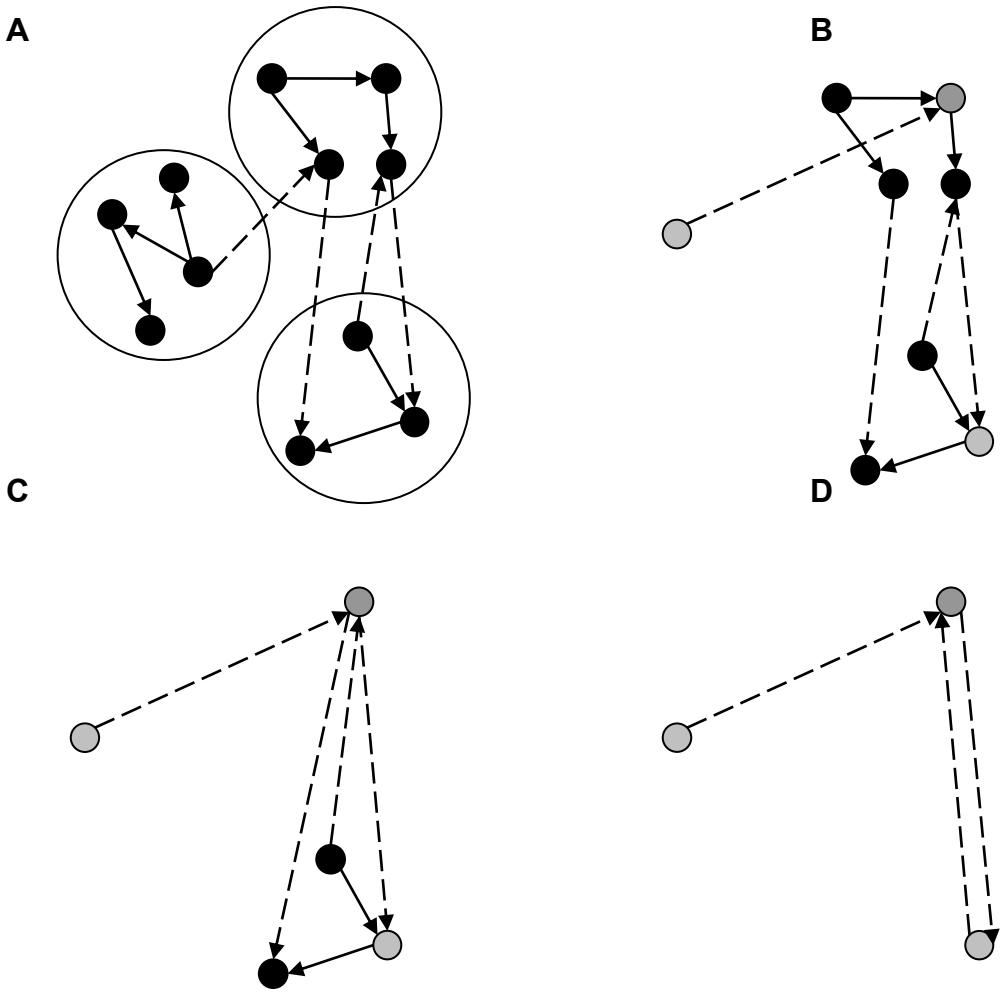


Figure 2.14: The sequence of operations conducted by the sub-graph collapse plug-in A-D to collapse nodes in an Ondex graph. Solid edges represent relations internal to a cluster, which are deleted. Dashed edges represent relations that interconnect clusters, which are reassigned to a cluster super-node. Black nodes are the original concepts and grey edges are the new super-nodes.

1. A formal syntax is needed for defining graph-traversal rules. Furthermore, defining a query should not require technical knowledge or programming skills, and must be accessible to a computer literate biologist, who is better placed to codify biological rules.

2. To retain the provenance of where a candidate annotation had originated from a sub-graph a full record of the path traversed across the graph must be retained and attached to the resulting target annotation.

Two example queries over an Ondex graph that codify a biological rule that have been used in Chapter 3 are provided below:

1. Find all *query sequences* that have a similar sequence to a *protein*, and where that *protein* has an annotated *EC class*.

2. Find all *query sequences* that have a recognised *protein domain*, and where that *protein domain* has an annotated *EC class*.

These rules can both be expressed declaratively in first order logic (FOL). However, there are many variants of FOL syntax, and formal logic languages are unlikely to be accessible to most biologists. Within computer science, a number of application specific languages have developed to provide formal query languages for databases. For relational databases the standard is the Structured Query Language (SQL), and for RDF graph queries SPARQL serves this purpose. However, these languages are often tied to specific data-structure implementations, and would require complex adaptors to make them function on an Ondex graph. Additionally, they do not represent a user-friendly syntax for the non-computer-scientist. A number of similarly motivated projects have endeavoured to represent SPARQL queries graphically (Hogenboom *et al.*, 2010).
In addition to supporting a query syntax that is user-friendly, the query engine results must retain the full provenance of the query. To retain the provenance of where a candidate annotation had originated from a sub-graph such as resulted from Section 2.3.2 is insufficient. A full record of the path traversed across the graph must be retained and attached to the resulting target annotation.
In order to address the first requirement, a graphical notation for defining valid

routes through a graph was defined. This represents concept-class and relation-type meta-data, represented as a graph, which are valid for traversal over the Ondex graph. A start and finish concept-class on the meta-data-graph indicate the query concept class and target annotation type respectively. Repetition of meta-data types and cycles are valid for the meta-data-graph, but not permitted in concept and relation instances in the Ondex graph. Figure 2.15 shows an example of such a meta-data-graph query. In order for a path in the Ondex graph to be valid it must contain the same order and types of concepts and relations permitted by the meta-graph. A validation of this is that a sub-graph composed of the paths returned by the query algorithm must have the same (or subset) meta-graph as the query. Additionally, attribute restrictions may be added to the meta-graph-query, for example: a requirement that a traversal across an "is similar sequence" relation must have a bitscore greater than a set value.

The meta-data-graph shown in Figure 2.15 is implemented on the assumption

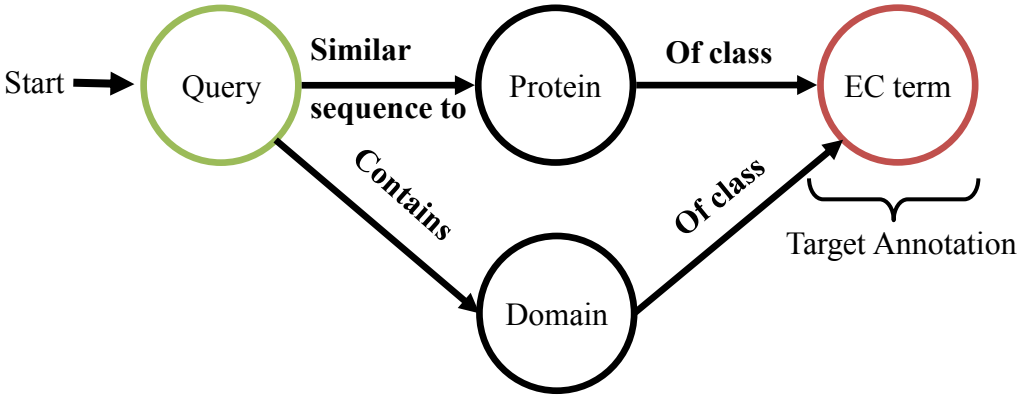


Figure 2.15: An example query showing permitted paths in the meta-graph from the query concept class to the target annotation class, traversing selected relation types.

that the Ondex graph is non-redundant, *i.e.* it has been processed by a graph transformation algorithm as described in section 2.3.3. The equivalent query for a redundant graph would include *is a* self relations for each concept in the meta-graph-query, which would allow free traversal across equivalent concepts.

The Meta-data based Graph Query Engine (MGQE) builds on the graph-traversal

algorithm described in Section 2.3.2. The algorithm executes in parallel against the set of concepts that fulfil the type requirements of the starting node of the meta-graph-query (*i.e.* for Figure 2.15 this corresponds to all concepts in the graph of concept-class type *query*). The traversal rule introduced in Section 2.3.2, instead of checking against a hard coded rule, verifies each from-concept, relation, to-concept triple against the meta-data-graph, encoded in a Java object-oriented (OO) model. The meta-data-graph is currently defined by a simple flat-file form, which is parsed into the native Java object model. However, creating a GUI to generate the same Java OO model would be relatively trivial. The meta-data-graph is implemented with a Java API that may be extended to perform more complex validations, as the full path history is submitted together with the next candidate concept to the traversal rule checker. In addition to the sub-graph traversal algorithm described in Section 2.3.2, and to fulfil the second requirement of tracking, the full provenance-path of the query results is stored. This replaces the sub-graph results described in Figure 2.13, and is implemented as a linked tree where each node references the previous concept/or relation in the path. This minimises the memory requirements of a large number of paths. The full set of paths is derived from recursion back along the path from the terminal nodes, selecting those that fulfil the type requirements of the finishing concept-class of the meta-graph-query.

An Ondex plug-in was developed to encapsulate the functionality of the meta-graph-query algorithm API. The resulting annotations from query concept to target annotation can be exported, together with the full path history (and associated provenance), in tab-delimited format. This functionality is used extensively within the pipeline described in Chapter 3. The plug-in can also be used to return a sub-graph of the elements included in the traversal. The meta-graph-query is currently provided in flat file-format by the user, however as stated it is intended that a GUI for defining queries will be developed.

2.4 Conclusions

This chapter has outlined the Ondex system, together with some key developments that were a requirement for the pipeline described in Chapter 3. A workflow framework for initialising graphs and executing sequential plug-ins was implemented. This forms the basis for all the Ondex workflows described in this thesis. Additionally, a persistent mechanism of describing workflows meant that workflows described in this thesis, are reproducible, and can be run against current data when they becomes available.

A number of methodologies were developed as part of this research project. The parallel connective sub-graph search was fundamental to the development of the procedure for removing redundancy, and underpinned the architecture of the Meta-data based Graph Query Engine (MGQE). The parallel computing engineered into these methods enabled workflows to run in tractable time. The reduction in graph space, by removing redundancy, allowed the workflows to be run completely in RAM on a server (128GB RAM), which dramatically reduced their running time. The parallelisation of these algorithms meant that on a server with 32 active processing threads, working with huge datasets became tractable. With all the improvements described, the workflows described in Chapter 3, took approximately two weeks on a high performance server, with 128GB RAM and 16 Intel hyper-threaded CPUs at 2300 MHz.

The meta-graph query algorithm, developed for extracting paths through a graph, was the most important prerequisite to using data-integration in Ondex for sequence annotation. It permitted the formal definition of biological rules that can be applied to an integration Ondex graph, and resulting candidate annotation exported for further analysis. These will be further outlined in Chapter 3. Using parallelisation greatly reduced the time required to extract annotations, and enabled regular annotation workflows to be re-run to test dif-

ferent parameters, and as new data became available. It also allows for queries to remain tractable as the size of the knowledgbase increases.

Chapter 3. CoPSA: Improving gene annotation through conjoint sequence alignment to an integrated knowledgebase

3.1 Introduction

Gene functional annotations are a prerequisite to the effective analysis of gene expression data. They allow significant over- or under-expressed genes to be classified by their associated functional groupings and biological processes. This can form the basis for further studies or as an aid to interpret other informatics and statistical analyses. Structured annotation systems allow generalization of annotation, which consequently enables in silico testing of general hypotheses against the dataset, for example: *The expression of genes that are components of photosynthetic processes varies as the stress progresses*. It also enables data driven hypothesis generation, where computational and statistical methods suggest processes that account for the observed gene expression. For example: the use of multiple testing to identify enriched processes, which form the basis for new hypothesis.

The annotation of genes from non-model plants, of which wheat is typical, presents a significant challenge. Their genomes are often large in size relative to *Arabidopsis* and they lack or are not amenable to high-throughput technologies for functional genomics. This has resulted in a scarcity of experimental validation of gene function. Given the continued reduction in the cost of sequencing technology, and initiatives like the International Wheat Genome Sequencing Consortium (IWGSC), the need to reliably transfer annotation from well-

characterised model organism genomes to large numbers of new sequences will continue to be a major requirement in crop bioinformatics.

This chapter presents a pipeline for the Conjoint Prediction of Sequence Annotation (CoPSA). CoPSA is built around the Ondex data-integration framework and is configured using data sources pertinent to plants; however the principles of data-integration and annotation transfer are applicable across all species. The performance of CoPSA is evaluated against other comparable pipelines, using the sequences used to design the Affymetrix wheat microarray as an example. The annotations generated by CoPSA were used as a basis for further analysis of the durum wheat microarray data. This is presented in Part II of this thesis.

There are many comparative genomic pipelines that propose functional annotations for sequences using the ever growing corpus of functional annotations from other species. The novelty of the CoPSA system lies in the integration of multiple sources of primary annotations.

3.1.1 Data sources and structures

The construction of a sequence-annotation pipeline requires the consideration of a number of possible strategies and parameters. A common strategy, and one that is adopted here, is to link a nucleotide sequence to existing sources of functional annotation based on a computational analysis of the primary structure of the sequence. This process is broken down into three main computational tasks, (a) sourcing functional annotations, (b) linking primary sequence to existing functional annotation, and (c) evaluating the evidence for transference of functional annotations.

A number of computational techniques exist to predict the function of a new sequence, based on its primary structure. Alignment to protein sequences with

a known function is a common approach. Nucleotide sequence alignments can be made directly against amino acid sequences using algorithms such as translated BLAST (Altschul *et al.*, 1990). A statistically significant bidirectional match makes it possible to infer potential functional-orthology between the new gene and its equivalent in a model organism. This putative functional-orthology relationship becomes the basis of transferring the annotation from gene in the model organism to the new gene. A more complex phylogenetic prediction of functional-orthology can be constructed using multiple sequence alignment at the computational cost of aligning all the sequences within the target organisms. Conserved sequence motifs across protein families can also be detected using Hidden Markov Models built from protein sequence databases with algorithms such as HMMR (Eddy, 2009, 1990).

The vocabulary and structure of protein and domain annotations, which form the basis for transference, are an important consideration for downstream analysis. Possible annotation types range from free-text descriptions, controlled vocabularies, hierarchies, and ontologies. Terms that form part of structured classification systems facilitate more powerful computational analysis. In practice the requirements for the classification system depends on the nature of the biological question and the required analysis. Common references for defining structured terms include the Enzyme Commission (EC) classification of enzymes (NC-IUBMB, 1999), Gene Ontology (GO) of gene function, process and cellular location categories (Ashburner *et al.*, 2000, Ashburner, 1998), the FunCat hierarchy of protein function (Ruepp *et al.*, 2004) and COG functional-ortholog groups (Tatusov *et al.*, 2000, 2003). Various domain specific ontologies have been developed to extend the Gene Ontology family with additional categories. The Plant Ontology Consortium (2002) has produced a number of categories including organ, tissue and cell type structures, developmental stages (Jaiswal *et al.*, 2005) and traits. When annotating genes to metabolic pathways, a number of database or tool specific pathway classification systems can be found

in the KEGG (Kanehisa *et al.*, 2010), MetaCyc (Caspi *et al.*, 2010) and Reactome (Matthews *et al.*, 2009) databases and their derivatives. These vocabularies for annotation have been described in detail within Chapter 3.1.2(*a*), a subset of which was selected for the biological use-case described in Part II. The target for annotation within this chapter are therefore GO, EC and a controlled vocabulary of transcription factor families.

In Chapter , the NetAffx and BLAST2GO pipelines were described in detail. The annot8r and ArrayIDer pipelines, while not providing downloadable annotation for the Affymetrix wheat GeneChip, provide tools capable of automated assignment of annotation. The annot8r is provided as a software-tool only, however ArrayIDer provides annotations for sequences from other species and microarray chips. Their methods are described here, as a point of comparison to CoPSA.

The annot8r pipeline uses BLAST to align nucleotide or protein sequences against the annotated portion of the UniProt database (Schmid and Blaxter, 2008). They focus on the annotation of species without sequenced genomes, where genes are assembled from transcript sequences, into database such as UniGene (Sayers *et al.*, 2011). As such their pipeline, while being developed and demonstrated on nematode sequences, addresses the same issues present for the wheat Affymetrix GeneChip. While the tool does not specifically include EST assembly, they indicate that within their overall pipeline they use the PartiGene pipeline for EST assembly (Parkinson *et al.*, 2004) and the prot4EST tools for optimally translating EST to proteins (Wasmuth and Blaxter, 2004). The annot8r pipeline annotates sequences with GO, EC, and KEGG pathways. BLAST was used to align sequences against the proportion of UniProt that has usable annotation, using user defined BLAST parameters. EC and KEGG pathway annotations for proteins are included from the ENZYME database (Gasteiger *et al.*, 2003), and UniProt annotations provided by KEGG (Kanehisa *et al.*, 2010), respectively. The user may specify at this stage to include annotations which

have the GO evidence code Inferred from Electronic Annotation (IEA), but inclusion is not recommended given the potential for accumulating error as discussed in Chapter . Confidence in an annotation transferred from a similar protein is based on two simple measures. The bitscore and e-value of the best hit supporting the annotation, and the proportion of all the protein hits that support the annotation. The limitations of the pipeline are that the exclusion of sequences without annotation at an early stage allows annotations to be transferred from suboptimal alignments. If a highly similar protein of unknown function is present in UniProt, for a given query sequence, it is excluded *a priori*. This skews the overall e-values (which are based on sequence diversity), and prevents the tool from accounting for the distance of the best protein alignment, to the best protein with annotation. The simple inclusion or exclusion option for IEA, without evidenced based weighting of annotations, means that if a user chooses to include IEA, then it may be included in annotation at the expense of experimental annotation. Conversely, if a user excludes IEA then it is at the expense of coverage.

3.1.2 Evaluating the quantity and quality of functional annotation

Evaluating the quality of annotation, in terms of quantity and precision, is essential for developing a functional annotation selection strategy for transferring knowledge by sequence similarity. The most common strategy in Bioinformatics is to use precision and recall to this end. Estimations of precision and recall can be calculated relative to a gold standard of very high quality annotations. Qualitative precision and recall metrics taking into account the hierarchical structure of the GO have been proposed by both Kiritchenko *et al.* (2005) and Pal and Eisenberg (2005), however these measures are based on single an-

notation comparisons. Comparisons of multiple sets of annotations, which is relevant to sequence annotation has been developed by Verspoor *et al.* (2006), based on the hierarchical measures proposed by Kiritchenko *et al.* (2005). The Verspoor *et al.* (2006) precision and recall metrics calculate the hierarchical distance to the closest term in the gold standard annotation for that gene for each GO term that annotates a gene. The resulting recall or precision for that gene is the mean of these scores for all terms.

Hierarchical precision (Equation 3.1), recall (Equation 3.2), and f-score (Equation 3.3) is defined for an annotation set $F(g)$ on a gene g, with reference to a gold standard set by $R(g)$ (Verspoor *et al.*, 2006). For the GO term t, the function $anc(t)$ returns a term and all its ancestors that are connected by *is a* and *part of* relations within the GO ontology.

$$hPrecision(F,R,g) = \frac{1}{|F(g)|}\sum_{a\in F(g)}\max_{r\in R(g)}\frac{|anc(r)\cap anc(a)|}{|anc(a)|} \qquad \text{(Equation 3.1)}$$

$$hRecall(F,R,g) = \frac{1}{|R(g)|}\sum_{r\in R(g)}\max_{a\in F(g)}\frac{|anc(r)\cap anc())|}{|anc(r)|} \qquad \text{(Equation 3.2)}$$

$$hF(hPrecision,hRecall) = \frac{2(hPrecision)(hRecall)}{hPrecision + hRecall} \qquad \text{(Equation 3.3)}$$

A prior assumption in the calculation of precision is that the gold standard annotation is complete, and consequently all surplus annotations in the predicted annotation are the result of false positives. Unfortunately, no such experimentally verified and complete functional annotation exists for every gene on the wheat Affymetrix chip. Consequently, indirect measures of the quality of annotation must be used as a proxy for the quality of novel predictions. The functional annotation evaluation metrics implemented in the AIGO library were used for this purpose. These metrics are described in Section 3.2.3. The recall calculation, however, can still be made using an incomplete set of high quality annotations, with the usefulness of the measure being proportional to the completeness of the gold standard.

3.1.2(*a*) Strategies for selecting the best functional annotations

A common method, for inferring gene function through sequence similarity, is to simply take the best hit against a model organism. However, the best hit from one species to the next may not reciprocally be the best hit. Whereas a slightly worse alignment, that is more reciprocal may be a better candidate for transference of function. Ideally a full phylogenetic tree should be reconstructed, based on multiple species, which allows the evolutionary history of the gene family and its subfunctionalizations to be leveraged in predicting functional-orthologs. In some pipelines, a pre-computed database of known functional-orthologs are used to transfer functional annotation; this approach is used in the NetAffx pipeline (Liu *et al.*, 2003). In the situations where the sequence is novel and does not exist in a database of established functional-orthologs, a transfer function based on reciprocal best hits is often used (Bork and Koonin, 1998, Tatusov *et al.*, 2003). This approach, however, only predicts ancestral functional-orthologs, and does not provide paralogous genes in the family that are sufficiently conserved that they may also have a shared function. It therefore results in a drastic reduction in coverage of annotation, in the target organism. More advanced techniques cluster proteins into tree structures of functional-orthologs and their paralogs Li *et al.* (2003), Ostlund *et al.* (2010b) or relate the novel sequence to an existing phylogenetic tree (Datta *et al.*, 2009b). However a severe limitation of these approaches is that they rely on reasonably complete sequence information in a species. For an incompletely sequenced organism like wheat, the reciprocal blast hit is vulnerable to producing false positives, because it is highly likely that a more similar sequence for the reciprocal hit exists in wheat, but has not yet been sequenced. As discussed in Chapter , wheat sequences on the GeneChip are mainly UniGenes, created from assembled EST sequences. They are often incomplete or inaccurate; this

may also result in incorrect alignments, and disrupt bidirectional results.

Provenance is also an important consideration in transferring functional annotations. The true functional-ortholog may be known, but only have electronically inferred annotation. Whereas another very similar sequence may have an experimentally validated function. In this case it may be appropriate to transfer the more trustworthy annotation. Each annotation predicted by CoPSA is accompanied by multiple provenances of information, which for BLAST derived annotations include the protein and species of origin, sequence alignment scores, gene ontology evidence code, and database sources. Database provenance for an annotation may in turn be a composite of multiple databases, for example: a sequence aligns to a protein sequence in Gramene, which has a pathway entry and EC term in AraCyc, which has a translation to GO term in EC2GO mapping. This chapter concerns the definition of a novel metric for transforming some of these types of provenance information into a confidence score. This confidence scoring function is restricted to BLAST derived annotation, as domain-derived annotation scores are not directly comparable and Gene Ontology (GO) evidence codes are absent in GO annotation of domains for the integrated databases.

3.1.3 Aims and Objectives

The broad aims of the work in this chapter are to:

- Improve the quality and quantity of gene-transcript annotation in non-model organisms.
- Provide metrics for the confidence in a given annotation.
- Preserve the provenance of annotations.

- Automate the above in a pipeline that can be re-run when new data become available.

This will be achieved through:

- Identifying similar genes in other organisms.
- Identifying functionally conserved domains.
- Sourcing annotations of multiple databases and evidences.
- Providing a unified data structure and semantics for extracting annotation.
- Extracting potential annotations by rule based reasoning over data.
- Selecting sub-sets of annotation based on metrics for *quality*.

This work will be evaluated by:

- Quantifying the contribution of sequence alignment methods, primary data-sources, and data integration.
- Comparison to other sequence based annotation pipelines using precision and recall.
- Empirical metrics of to compare the properties and semantic content of annotations.

3.2 Methods

With the aim of exploiting integrated data resources to enhance gene sequence annotation, CoPSA was developed as a four-stage process:

1. Data aggregation: selecting and extracting the required data resources and transforming them into a technically and semantically homogeneous knowledgebase.

2. Data integration: the identification of equivalence across data sources, including the data alignment of the redundancies created by data aggregation, in which equivalent concepts and relations are merged together.

3. Conjoint sequence alignment: identifying similarity from query sequences to proteins and presence of Pfam domains within sequences then inferring new gene-protein and protein-domain similarity relations in the knowledgebase.

4. Selection: Applying biological rules to traverse the integrated dataset and extract the best candidate annotations.

Methods concerning the data (1) aggregation, (2) integration, and (3) conjoint sequence alignment are described in Section 3.2.1 and methods concerning the (4) selection of candiatate annotations from putative functional-orthologs are described in Section 3.2.2.

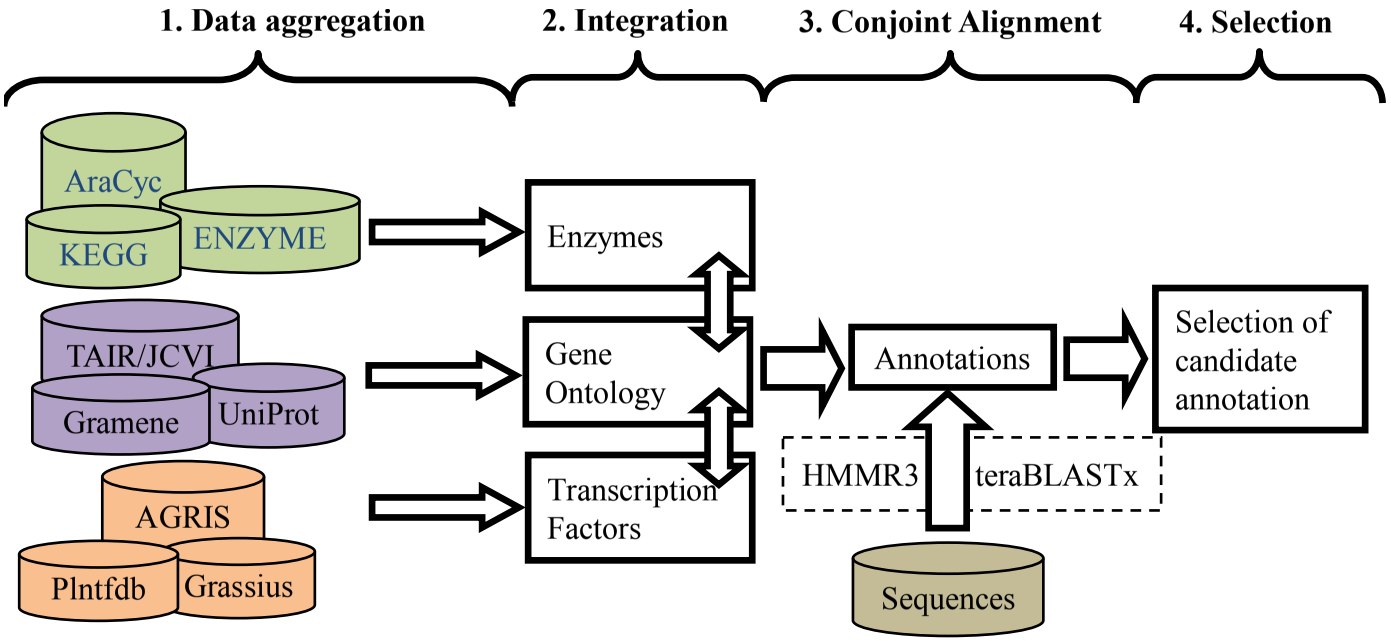


Figure 3.1: The integration of Enzyme, Gene Ontology and Transcription factor information from public data into a gene annotation pipeline using conjoint protein and domain alignment. An evaluation of the pipeline is against three other independent gene sequence based annotation pipelines.

3.2.1 Construction of a knowledgebase

3.2.1(*a*) Data aggregation

The aggregation of relevant data into Ondex targets three main types of annotation, which were identified as important to the annotation process (Chapter 3.1.2(*a*)). These consisted of information about enzymes (EC terms), transcription factor families, and GO annotations. These three annotation types were prioritised as important in Chapter 3.1.2(*a*) and form the basis for the analysis in Part II of this thesis. The source databases for these annotation types are summarised in Table 3.1. These annotation types were selected deliberately as complementary and with potential for developing the appropriate cross references via ec2go (The Gene Ontology Consortium, 2011c) and through Ondex-defined mappings. Additionally sequences and domain structural-properties were included, which allowed candidate annotations to be assigned to query sequences, based on HMMR and BLAST-derived similarities.

Table 3.1: Sources and versions of databases aggregated into Ondex for annotations

Database	Version (Retrieved)	Components	Species
KEGG	53 (15/02/10)	PATHWAY, BRITE, KO, GENES, LIGAND	*Viridiplantae* (see Appendix 1)
AraCyc	6 (14/10/09)	All	*Arabidopsis thaliana*
ENZYME	(22/02/10)	All	
TAIR/JCVI	9 (15/04/10)	GO annotations	*Arabidopsis thaliana*
Gramene	v29-Feb09 (15/04/10)	GO annotations	*Oryza sativa*
UniProt	2010_04 (01/04/10)	SwissProt/TrEMBL	*Saccharomyces cerevisiae* *Escherichia coli* *Viridiplantae*
GOA-*Arabidopsis* (TAIR)	(15/04/10)	GO annotations	*Arabidopsis*
GOA-*Oryza* (Gramene)	(15/04/10)	GO annotations	*Oryza*
GOA-*Solanaceae*	(15/04/10)	GO annotations	*Solanaceae*
Sol Genomics Network (SGN)			
external2go	(04/04/10)	GO annotations ec2go interpro2go pfam2go prosite2go	
AGRIS	(26/09/09)	AtTFDB, AtRegNet	*Arabidopsis thaliana*
Plntfdb	3 (26/09/09)	FASTA	*Viridiplantae* (see Table 3.2)
Grassius	(26/09/09)	FASTA	*Zea mays* *Orzya sativa* *Sorghum* *Saccharum*
Pfam A	24 (06/02/10)	Domains	
InterPro	24.0 (06/02/10)	Domain annotations	

Plant genes are assigned to metabolic pathways in public databases such as KEGG (Kanehisa *et al.*, 2010), AraCyc (Mueller *et al.*, 2003), Reactome (Matthews *et al.*, 2009) and *Arabidopsis* Reactome (Tsesmetzis *et al.*, 2008). However, because plant Reactome pathways are inferred using OrthoMCL and therefore poorly represented (Lysenko *et al.*, 2010), and *Arabidopsis* Reactome at the time of writing was out of date, being based on the manual integration of the outdated KEGG release 38 (April 2006) and AraCyc v3.5 (February 2007). Therefore, these two resources were excluded from the CoPSA aggregated process.

There is considerable semantic diversity in the definitions of metabolic pathways between databases, which vary in scope, structure and granularity. For example, in terms of their scope they differ in their inclusion of signalling pathways and post-translational modifications as reaction steps in metabolic pathways. When considering structure and granularity of pathway databases, for example, KEGG contains large pathway maps grouped in 11 large metabolic categories, whereas AraCyc has a deep tree-hierarchy of pathways that resolves down to leaves representing much smaller pathway units. Individual AraCyc pathways typically contain much smaller groups of reactions than KEGG (Karp *et al.*, 2002).

Metabolic pathways are commonly composed of units of metabolic reactions each of which has products, substrates, and other accessory molecules such as enzymes and cofactors. Common products and substrates link reactions together; some compounds form the spine of the pathway, whereas others act as accessory metabolites to the main chain of compound synthesis or catabolism. Computationally identifying the spine of the pathway is a challenge. For KEGG, the RPAIR database formally categorizes the role of reaction pairs in the reaction (Kanehisa *et al.*, 2006). In AraCyc the main reactions must be manually inferred from the pathway diagram, which would be non trivial to do computationally.

Integrating metabolic pathway data therefore requires a complete identifica-

tion of equivalent reactions in each database. However, there are no universal unique identifiers for reactions that make up pathways. Finding equivalent reactions between databases is extremely challenging, as the completeness and specificity of the compounds with which the reaction is described are dependent on the database. For example, where a "reduced electron acceptor" is required in a reaction but unknown, AraCyc will place a textual place-holder in the reaction, whereas KEGG omits the compound altogether. For example, in the conversion of phaseic acid to dihydroxyphaseic acid: AraCyc RXN-8154 (Equation 3.1) includes unspecified electron accepters and reducers.

$$\textit{phasic acid} + \textit{reduced electron acceptor} \rightleftharpoons \textit{dihydroxyphasic acid} + \textit{oxidized electron acceptor} \quad \text{(Equation 3.1)}$$

However, KEGG R07577 (Equation 3.2) omits both unknown electron acceptor from the equation. There are also disagreements on EC numbers assigned to reactions, or differences in their specificity.

$$\textit{phasic acid} \rightleftharpoons \textit{dihydroxyphasic acid} \quad \text{(Equation 3.2)}$$

Poolman *et al.* (2006) have identified a large amount of erroneous duplicated reactions in KEGG. They also identified ambiguity in metabolite identifiers and unbalanced reactions. These problems mean that finding equivalent reactions between pathway databases is a difficult task, which frequently requires expert assistance. For example: Radrich *et al.* (2010) have built high quality integrated SBML metabolic networks from KEGG and AraCyc using curator assisted semi-automated methods. Similarly, Tsesmetzis *et al.* (2008) used the Reactome framework to manually integrate KEGG and AraCyc and thereby identify differences and inconsistencies. Taubert *et al.* (2009) have previously described the use of graph visualisations in Ondex as a tool for compare the content of the KEGG and AraCyc databases.

For these reason, when developing the integration of pathway information in

the Ondex knowledge-base for CoPSA, only Enzyme Commission (EC) numbers were considered, as equivalence can more readily be found between EC entries in data sources. A full description of the role these EC numbers play in classifying enzyme function has been described in Chapter 3.1.2(*a*). The ENZYME parser in Ondex works on the flat files downloadable from ExPASy and was already available in Ondex. However, modifications were required to include protein and domain annotations in the database.

GO annotations for the TAIR and Gramene databases were extracted from the Gene Ontology Annotation (GOA) file format from the GO Website. An existing parser in Ondex was available for the GOA file format, but this required updating for this use-case. UniProt annotations were parsed from the UniProt XML files by adapting the existing Ondex parser to include GO annotations and UniProt Evidence Codes. The whole Gramene database, which contains GO annotations for rice and a large number of grain species, was imported into Ondex using the existing parser, which worked against the Gramene flat-file data export.

For the purposes of mapping Pfam domains to Gene Ontology terms, the Interpro2go database (Hunter *et al.*, 2009) was imported into Ondex. These data were parsed using the existing Ondex plug-in for the external2go format defined by the Gene Ontology Consortium. An InterPro XML parser was written to provide mappings between Pfam and InterPro accessions.

The data-sources selected for transcription-factor annotation were chosen in order to represent a broad range of species. AtTFDB (Davuluri *et al.*, 2003) is part of the *Arabidopsis* Gene Regulatory Information Server (AGRIS) (Palaniswamy *et al.*, 2006), and contains 2,661 (December 2011) putative *Arabidopsis thaliana* transcription factors identified through the presence of a DNA binding motif, sequence similarity to a known transcription-factor, and literature curation. Genes were divided into 50 transcription-factor families. For computational predictions, AtTFDB uses a combination of HMM profile search and

iterative BLAST (e-value $< 1\times10^{-5}$), however this provenance information was not preserved in the AtTFDB FASTA file. The Plant transcription-factor database (PlnTFDB) (Riaño-Pachón *et al.*, 2007) contained transcription-factor predictions based on HMM profile searches of 20 species (Table 3.2). PlnTFDB divided genes into 84 transcription-factor families using domain-based classification rules. The Grass Regulatory Information Services (Grassius) (Yilmaz *et al.*, 2009) provided transcription-factor predictions for four grass species (Table 3.2), of which only Saccharum is not included in PlnTFDB. Grassius uses a combination of BLAST (e-value $\leq 1\times10^{-5}$) with an InterPro-scan of DNA-binding domains. It uses a combination of PlnTFDB and AtTFDB protein-family classification systems.

Table 3.2: The species represented in the PlnTFDB database

Group	**Species**
Bangiophyceae	*Cyanidioschyzon merolae* *Galdieria sulphuraria*
Prasinophyceae	*Micromonas pusilla CCMP1545* *Micromonas sp. RCC299* *Ostreococcus lucimarinus* *Ostreococcus tauri*
Chlorophyceae	*Chlorella sp. NC64A* *Chlamydomonas reinhardtii* *Coccomyxa sp C-169*
Bryophyte	*Physcomitrella patens*
Lycopodiophyta	*Selaginella moellendorffii*
Monocot	*Oryza sativa subsp. indica* *Oryza sativa subsp. japonica* *Sorghum bicolor* *Zea mays*
Eudicot	*Arabidopsis lyrata* *Arabidopsis thaliana* *Carica papaya* *Populus trichocarpa* *Vitis vinifera*

3.2.1(*b*) Data integration

There were two stages to data integration in CoPSA: semantic alignment and identification of equivalence. Semantic alignment was mostly dealt with at the parser stage, where data within each database were transformed into concepts and relations that were consistent with Ondex metadata. Data from each source correspond to unique concepts and relations redundantly across data sources in the Ondex graph, requiring the creation of a new concept or relation, even if an equivalent exists from a previously parsed element in another database. This has a number of advantages, which have been outlined in Chapter 2. However, this duplication of data between sources is expensive in terms of resources, which is a particular concern for this Ondex application. It involves large sequence databases that are transformed into millions of concepts and tens of millions of relations. The identification of equivalent concepts within the graph (Figure 3.2) was a requirement both for executing cross database queries and for merging equivalent concepts to reduce memory requirements. Table 3.3 shows the identifiers that were used within this use-case to map equivalence between concepts in Ondex for each class of concept relevant to this application. More types of identifiers were present in these data sources, but these were not included as indicators of equivalence, either because they did not overlap with other data sources or because they were too ambiguous in the context of the concept class for this use-case.

Table 3.3: Identifiers used to map equivalence for Ondex classes and concepts.

Class of Concept	Identifier type	Example id	Reference
Protein	UniProt	P43291	(The UniProt Consortium, 2010)
	AGI locus	At1g10940	(TAIR, 2007)
	AGI models	At1g10940.1	
	SGN	1631	(Mueller *et al.*, 2005)
	TIGR locus	LOC_Os09g38320	(Yuan *et al.*, 2003)
	TIGR models	LOC_Os09g38320.1	
Biological Process	Gene Ontology	8150	(Ashburner *et al.*, 2000)
Molecular Function		5575	
Cellular Component		3674	
Enzyme Commission	Enzyme Commission	1.1.1.2	(NC-IUBMB, 1999)
Protein Domain	Pfam	PF01160	(Bateman, 2004)
	InterPro	IPR017442	(Hunter *et al.*, 2009)
	PROSITE	PS01252	(Hulo *et al.*, 2004)

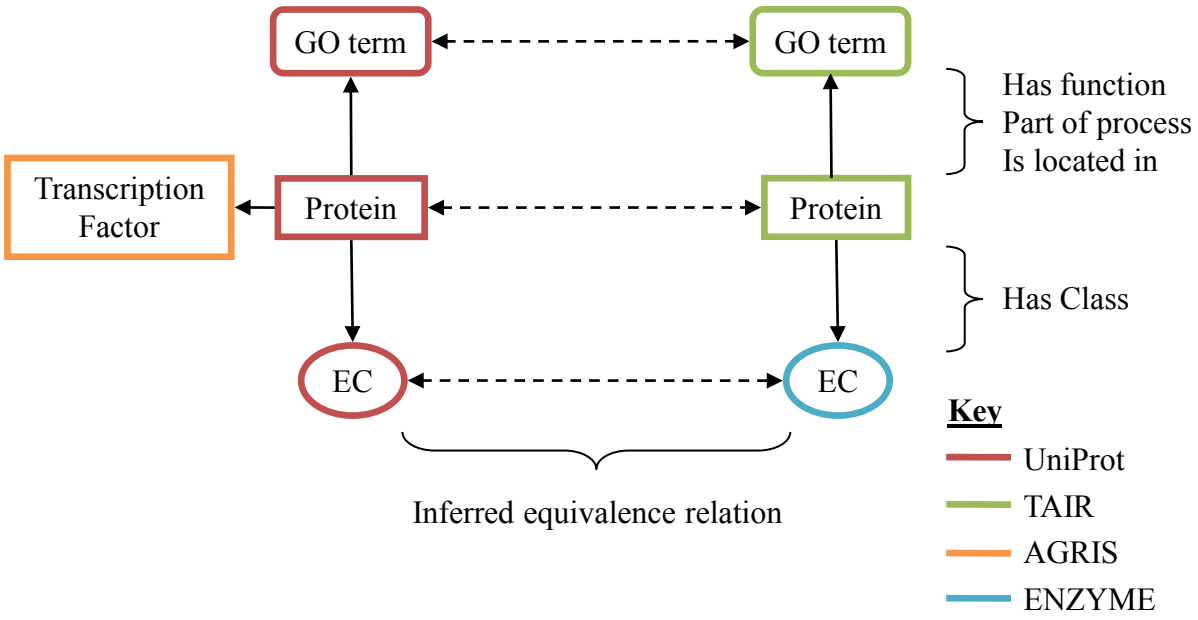


Figure 3.2: An example of the identification of concept equivalence on an Ondex graph containing four different data sources.

A protein is encoded by a single RNA transcript, which in turn is encoded by a single genomic locus. In the reverse relationship, a locus can yield multiple RNA transcripts, as a consequence of alternative splicing, each of which yields a different protein. As a consequence a database accession which identifies a protein, also unambiguously points to a single loci. However, a gene locus cannot be used in the same way to identify a protein. Some databases such as AraCyc, and the transcription factor databases described, only provide annotation to a gene locus. As the consensus sequences on a microarray represent RNA transcripts, if ambiguity is to be avoided, these consensus sequences cannot be inferred the function assigned to a gene loci. In some instances alternative splicing can result in differences in the function, process, and cellular location of a protein. For example: the loci Ppo-A1 in wheat encodes seven isoforms, only one of which is known to be a functional polyphenol oxidase (Sun *et al.*, 2011). However, without allowing inference to proteins from gene-loci, many potential annotations may be unusable. For this reason in this integration procedure, accessions which identify gene loci are permitted on proteins. This allows en-

zymes concept from AraCyc, which are identified only by an *Arabidopsis* Gene Identifier (AGI) locus (TAIR, 2007), to be mapped as equivalent to proteins in UniProt, based on the AGI locus.

Additionally 100% sequence identity of sequences was also used to identity identical proteins between databases. In some instances different gene loci code for identical protein sequences. In these instances using sequences as a means to identify equivalence may introduce ambiguity. However, in terms of function these proteins are likely to be identical, so this will not have an adverse effect on this pipeline.

GO and EC terms were unambiguously mapped using their respective types of accession. Obsolete terms (which were not found in the current release of GO or EC), were not translated to their current equivalent term (if any), as this was done at a later stage to preserve provenance for statistics on obsolescence.

After clusters of equivalent concepts were identified in the graph, and relations indicating equivalence created to link these together, the resulting graph is large and redundant. The size of the graph, which for this CoPSA pipeline was millions of concepts and tens of millions of relations, is a challenge for computation and storage. To address this issue, the redundancy removal methodology, as described in Chapter 2, was applied to the graph.

3.2.1(*c*) Conjoint Sequence-Alignment

The integrated Ondex knowledgebase described above is generic to the annotation of any plant gene sequence. However, the alignment methodology for consensus nucleotide sequences presented in this section has been specialised for use with plant Affymetrix GeneChip arrays. While the wheat Affymetrix array is the focus of the main use-case in Part II of this thesis, it was considered important to demonstrate and evaluate the CoPSA methods against all the available Affymetrix plant species GeneChip arrays. The arrays used in this analysis

are listed in Table 3.4.

Strictly, the annotation of an Affymetrix GeneChip concerns the sequences of

Table 3.4: The number of consensus sequences for each Affymetrix GeneChip (respective species only), and point at which NCBI taxonomony divides from *Arabidopsis thaliana*.

GeneChip	Consensus sequences	Taxonomy divides from *Arabidopsis* at:
Sugar Cane	8,224	Class: Liliopsida
Tomato	10,038	SubClass: asteroids
Vitis Vinifera	16,436	Class: Liliopsida
Maize	17,555	Class: Liliopsida
ATH1-121501	22,746	Class: Liliopsida
Barley1	22,782	Class: Liliopsida
Cotton	23,977	Order: Mavales
Citrus	30,219	Order: Sapindales
Rice	57,194	Class: Liliopsida
Medicago	61,035	Order: fabid, Fabales
Soybean	61,035	Order: fabid, Fabales
Wheat	61,115	Class: Liliopsida
Poplar	61,251	Order: fabid, Fabales

the probe sets used to fabricate the array. However, for most purposes it is more appropriate to use the consensus cDNA sequence, which is derived from assembled ESTs and used by Affymetrix as the basis for the selection of the array probe set sequence. This use of consensus sequences instead of probe-sets is consistent with the approach used by Affymetrix (Liu *et al.*, 2003) and others (Frickey *et al.*, 2008) for cross-species microarray sequence comparison. Dai *et al.* (2005) have made a strong case for re-annotating the GeneChip probes against current EST databases, as the original consensus sequences were based on more limited transcriptome knowledge than currently exists. A re-annotation of probes would reveal ambiguities where a probe-set that was designed to detect unique expression of an EST according to previous EST databases is actually hybridising with a previously unknown sequence. Despite some potential advantages in re-annotating the probe sets it was considered more appropriate for this project to re-annotate the original consensus-sequences because they provide a vital point of comparison to other published analyses, and annotation pipelines.

For sequence alignment between proteins and the consensus array sequences, a translated-alignment algorithm that is closely analogous to BLASTx (Altschul *et al.*, 1990) was used. For performance reasons, the Decypher Tera-BLASTx tool was utilised, which was run on a TimeLogic Field-Programmable Gate Array (FPGA). This allowed the six frame alignment of nucleotide sequences to millions of amino acid sequences within a reasonable time-frame (<48 hours using a SeqCruncher card). An e-value threshold of less than 1×10^{-4} and bitscore of greater than 50 was used to pre-filter hits. This means that the probability of finding at least one HSP by chance for that bitscore is 1×10^{-4}. Equation 3.3 shows the relationship between e-value and the probability of finding at least one High Scoring Pair (HSP) at the given score by chance (Karlin and Altschul, 1990)). The other parameters were six-frame-query translation, word size 4, banded gapped alignment, blosum62 matrix, open penalty -11 and extends penalty -1.

$$P = 1 - e^{-evalue} \qquad \text{(Equation 3.3)}$$

A plug-in module for Ondex was created to convert sets of array consensus transcript sequences and amino acid sequences from protein concepts captured in the CoPSA knowledgebase, into FASTA query and target files respectively. Tera-BLASTx was executed from within the Ondex plug-in via a generic Decypher interface that was written for this purpose. The tab-delimited file of sequence comparison hits was processed by the plug-in, which created a *has similar sequence* relation between each consensus-sequence-query concept and the protein hit. Properties such as e-value, bitscore, alignment length, and translation frame were stored as properties of the relationship.

The hmmscan program that is part of HMMER 3.0rc1 was used for profile searching (Eddy, 2009). Six frame translated consensus sequences were scanned against version 24 HMM profiles from Pfam A. The gathering thresholds (GA) were used as a score cut-off, which are manually defined by Pfam on a per model basis. An existing plug-in was adapted to work with the new version of

HMMER which created a *has similar sequence* between consensus sequences and Pfam protein domain concepts.

The product executing both the described alignment plug-ins was a graph containing relations connecting the query *Consensus Sequence* concepts to proteins and domains. A schematic of the types and properties of relations is shown in Figure 3.3. By consequence, proteins and domains linked the consensus sequences indirectly to a large range of information in the knowledgebase, some of which may form potential annotation of those sequences.

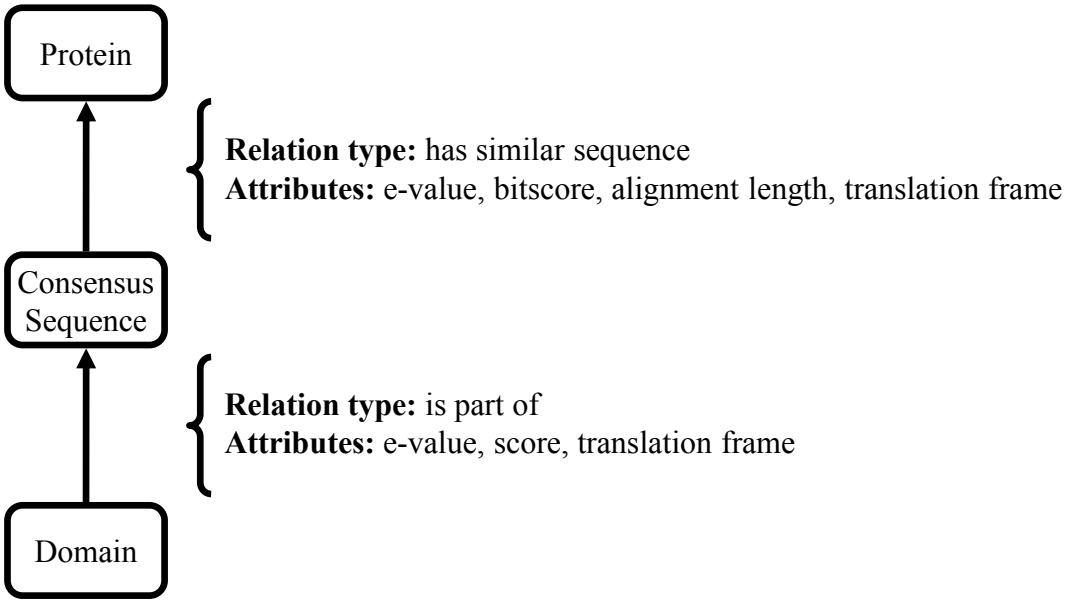


Figure 3.3: The hit results from conjoint sequence-alignment are stored as relations in the Ondex graph. Rectangles represent concept classes created by parsers of existing databases in Ondex. Arrows connecting concept classes are new relationships connecting them as a result of conjoint alignment. The attributes attached to each new relation type are listed.

3.2.1(*d*) An integrated knowledgebase

After the relevant data has been imported into an Ondex graph instance (Section 3.2.1(*a*)), integrated together to form a non-redundant knowledgebase (3.2.1(*b*)), and enriched with sequence similarity and domain annotations (Section 3.2.1(*c*)), the resulting graph is large and slow to work with in Ondex. Concept classes,

relation types, and attributes that were unrelated to the CoPSA workflow, were therefore removed from the graph. This mainly consisted of concepts representing publications, genes and KEGG Orthology (KO) groups, and relations connecting these components. Gene concepts were excluded as CoPSA utilises functional annotation of protein sequences and where appropriate gene locus functional annotation are transferred to the protein translation product. The biggest memory saving resulted from removing nucleotide and amino acid sequences, after the methods described in Section 3.3.3 were complete. The meta-graph for the resulting knowledgebase in Ondex, as shown by the Ondex visualisation tool, is shown in Figure 3.4. This is still a complex graph containing 14 classes of concept, and 20 types of relation. It contains over a million individual concepts and six million relations.

Concept class inheritance, previously described in Chapter 2 was not utilised in this graph. Instead an expanded notation was used where the relation type *is a* was used to denote the relationship between a concept and its parent concept. For example: for each enzyme concept, two concepts of the class *enzyme* and *protein* were created, and connected through an *is a* relation. This increased the complexity of the graph, but was necessary as at the time of writing not all Ondex plug-ins fully supported the class hierarchy functionality encoded by the metadata. However, these two forms of representing class hierarchy can be converted interchangeably.

Extraction of meaningful information from the knowledgebase shown in Figure 3.4 was a prerequisite for its utilisation in CoPSA. Its size and diversity in terms of concept classes and relation types meant that no informative visual layout of the whole graph was possible in Ondex. This semantic complexity was the motivation for the query engine previously described in Chapter 2. The semantic richness of the graph meant that multiple paths to target annotation, utilising different concept class and relation type combinations were possible. The query engine was therefore required to accept multiple declared routes through the

meta-graph to a given annotation concept class. The large number of concept and relation instances in the knowledgebase added the requirement that the query engine should be scalable to work with large graphs.

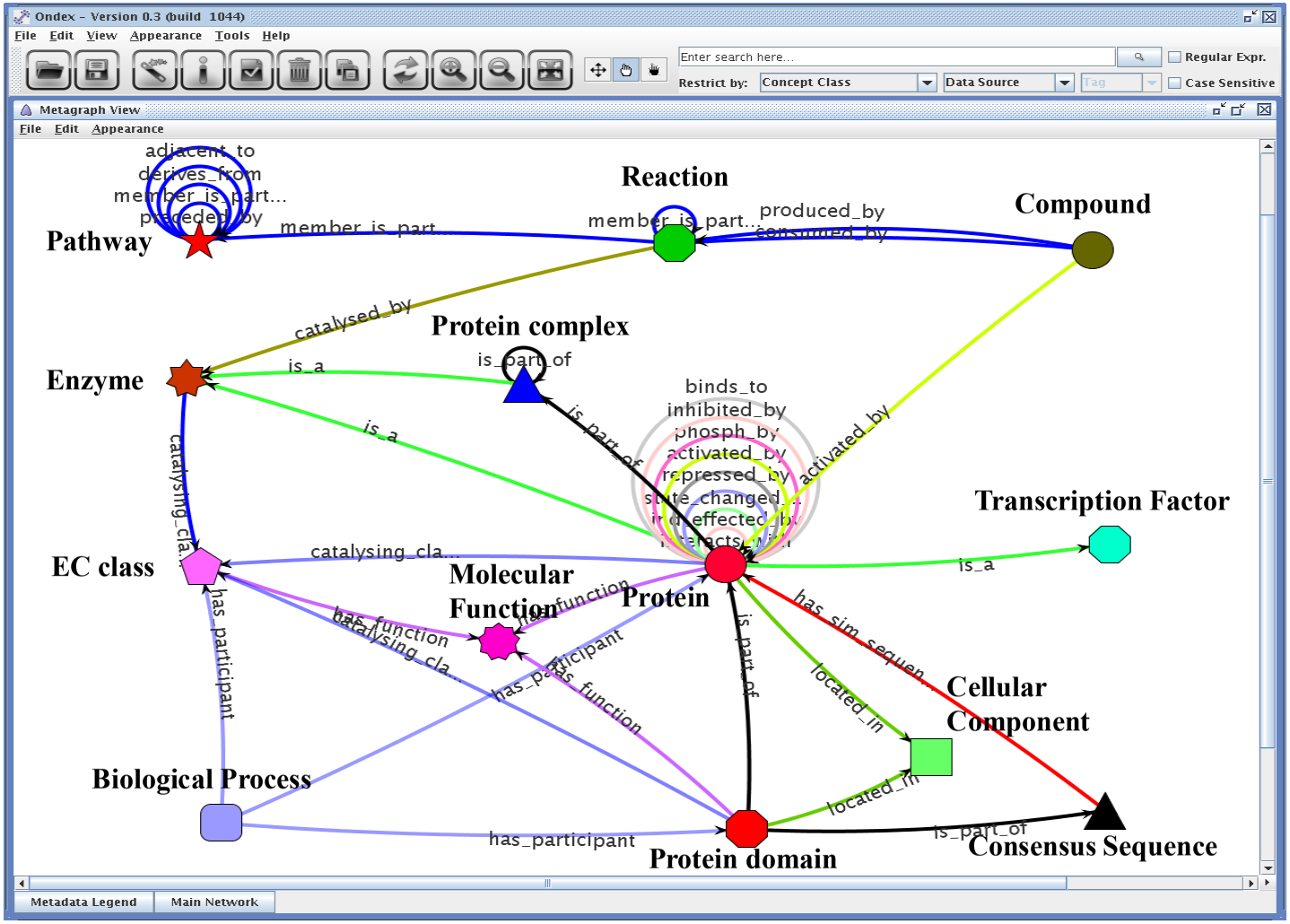


Figure 3.4: A screen shot from the Ondex visualisation software showing the meta-graph of the integrated knowledgebase for CoPSA annotation. Some concepts classes that were imported by parsers but unrelated to this use-case have been removed. Concept classes have been relabelled for clarity. HMMR and BLAST based relations are also included. A key to the concept and relation labels is found in Table 3.5 and Table 3.6 respectively.

Table 3.5 shows the quantity of each of the concept classes in the CoPSA knowledgebase. The protein concept class had the largest number of concept instances. These represent proteins and their sequences from all plant species, in addition to Algae, Yeast and *Escherichia coli*. Proteins were found in the majority of databases integrated in Section 3.2.1(*a*), and represent the biggest challenge in terms of computation and storage requirements on Ondex. Transcription factors are abundant in the knowledgebase, because they include three databases covering 22 species. Gene concepts are not present in the graph as functional annotation of protein sequences was targeted by the databases imported in Section 3.2.1(*a*). Where a functional annotation of a gene locus occurs in a database such as AraCyc, this is represented as a protein, as implicitly it is the translated product of the locus that performs the function. Enzyme concepts were less abundant than expected, as they were only created by AraCyc and KEGG parsers. Some parsers of databases like ENZYME import EC annotations of proteins but do not create Enzyme concepts. This does not affect the CoPSA pipeline however. Protein complexes are rare, which reflects the relatively small amount of information concerning them in functional annotation databases.

Table 3.6 shows the quantity of relations in the CoPSA knowledgebase. On average there are six relations for every concept, however 63% of these relations represent BLAST based sequence similarity, and connect 61,115 consensus sequences to 909,417 proteins. The remaining relations however are distributed across the concepts in the graph. Over 1.6 million annotations to GO terms are included in the graph, which are represented as relations to Proteins, EC terms and Pfam domains. EC annotations are less abundant, with only 74,881 annotations, which link EC terms to proteins, domains, and enzyme concepts. This was expected as EC represents a subset of the functionality captured in the GO ontologies. The Protein-Protein Interaction (PPI) and Protein-Concept Interaction (PCI) concepts in the graph are all derived from the KEGG database,

Table 3.5: A summary of the quantity of concepts and their classes in the Ondex integrated knowledgebase prepared for CoPSA annotation.

Ondex label	Concept description	Number of instances
Protein	Protein	909,417
Transcription Factor	Transcription factor	67,548
Consensus Sequence	Consensus sequence	61,115
Domain	Protein domain (Pfam)	30,716
Enzyme	Enzyme	24,003
Cellular Component	GO Cellular Component	6,650
Compound	Compound	4,563
Reaction	Biochemical reaction	3,892
Molecular Function	GO Molecular Function term	3,111
EC class	Enzyme Commission class	2,852
Biological Process	GO Biological Process term	2,717
Pathway	Metabolic Pathway	931
Protein complex	Protein complex	180
	Total number of concepts	**1,117,695**

they represent negligible useful-information and are therefore not utilised in this CoPSA pipeline.

3.2.1(*e*) Extracting functional annotations for sequences through graph traversal

As previously described in Section 3.2.1(*d*) the result of integration and conjoint alignment was a very large Ondex graph of concepts and relations, with numerous properties attached. Previous steps of data aggregation and integration created an Ondex database of annotations to protein and domain concepts. Conjoint alignment made these annotations accessible by introducing sequence similarity relations from the concepts that contain the query sequences. The next step in the workflow was to extract valid paths from the query sequence to annotations. Encoded biological rules governed which relations could be traversed in order to connect a candidate annotation to the query sequence. Figure 3.5 shows an example of three possible hypothetical paths from a consensus sequence to a GO function annotation. Simply taking all possible paths would

Table 3.6: A summary of the quantity of relations and their types in the Ondex integrated knowledgebase prepared for CoPSA annotation. Protein-Protein Interaction (PPI) and Protein-Concept Interaction (PCI) are included but not utilized in this CoPSA pipeline.

Concept class	Relation description	Number of instances
has_sim_sequence	Has similar sequence (alignment)	3,856,937
has_function	Has function (GO annotation)	694,147
has_participant	Has participant (GO annotation)	496,529
located_in	Is located in (GO annotation)	443,665
is_part_of	Is part of (protein→domain)	333,421
is_a	Is a (concept hierarchy)	91,547
catalysing_class	Catalysing class (protein→EC)	74,881
catalysed_by	Catalysed by	64,739
member_is_part_of	Member is part of	6,866
adjacent_to	Adjacent to (metabolic pathways)	588
preceded_by	Preceded by (metabolic pathways)	432
derives_from	Derives from (metabolic pathways)	388
ind_effected_by	Indirectly effected by (PPI)	35
activated_by	Activated by (PPI+PCI)	30
inhibited_by	Inhibited by (PPI+PCI)	20
binds_to	Binds to (PPI)	10
state_changed_from	State changed from (PPI)	5
phosph_by	Phosphorylated by (PPI)	4
repressed_by	Repressed by (PPI)	3
	Total number of relations	**6,064,247**

yield erroneous paths, as relations that indicate protein-protein interactions and metabolic pathway steps change the implied subject of the final annotation and the inference chain from the query to the final annotation becomes invalid.

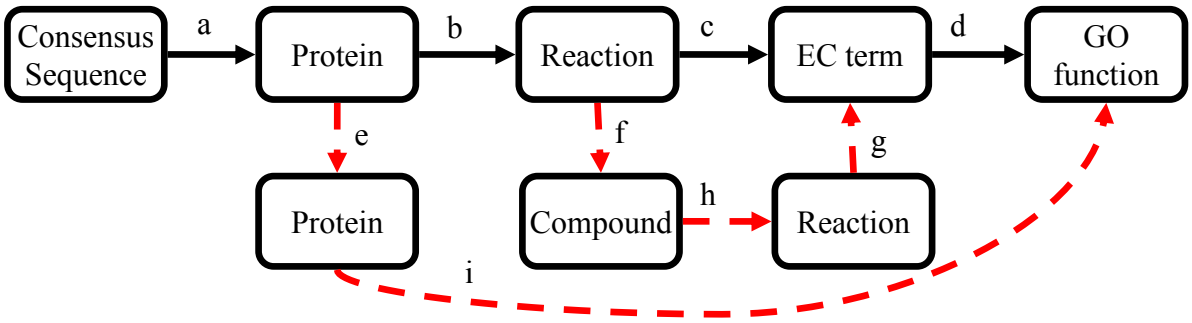


Figure 3.5: An example of valid and invalid paths through a graph. Key to relation types (a) has similar sequence (b) catalyses (c) of class (d) is equivalent to (e) interacts with (f) produced by (g) of class (h) consumed by (i) has function. The path {a,b,c,d} is valid, the paths {a,e,i} and {a,b,f,h,g,d} are invalid.

3.2.1(*f*) Encoding biological rules as meta-graphs

For the encoding of biological knowledge into a machine readable form, a declarative-query approach was adopted. This was described in detail in Chapter 2. Biological rules for traversing the knowledgebase were encoded as a graph of Ondex metadata (meta-graph). Traversal of the knowledgebase, from a query concept, was constrained by the query meta-graph, which contained a DAG of traversable concept classes and relations. The execution of the query in a traversal instance causes the algorithm, starting from the query concept, to pass through any part of the Ondex graph instance that is validated by the rules encoded in the meta-graph. For example: a very simple biological rule to define array sequences that are candidate transcription factors is shown below.

Every query sequence x that is "similar to" some "protein" concept y, where y "is a" "transcription factor" concept z.

The textual description above can also be expressed more formally in first-order logic (FOL), by defining which query sequences (x) are candidate transcription factors (z). This is defined in Equation 3.4.

Equation 3.4 encodes the rule which identifies when sequence x is a transcription factor in first-order-logic. The function type returns a constant defining the class/type of a concept or relation. The sets C and R define the complete set of concepts and relations in the graph respectively. The tuple $\{x, r, y\}$ indicates that in the graph x is linked to y through r.

$$\forall x, r', y, r'', z \; where \; (x \in C, r' \in R, y \in C, r'' \in R, z \in C) :$$
$$type(x) = \text{“}query\text{”} \; \wedge \; type(r') = \text{“}query\text{”} \; \wedge \{y, r'', z\} \wedge$$
$$type(y) = \text{“}is\ a\text{”} \wedge \{y, r''e, z\} \; \wedge \; type(z) = \text{“}transcriptionfactor\text{”}$$
$$\implies x \; \text{“}is\ a\text{”} \; transcriptionfactor$$

(Equation 3.4)

Figure 3.6 shows how the logic in the textual and FOL descriptions above was encoded as a query meta-graph, where each concept class and relation type restriction form an additional conjunction ($\wedge$) condition in FOL that restricts which sequences have the candidate transcription-factor target-annotation.

EC-codes annotations were extracted in a similar way with an additional 'or'

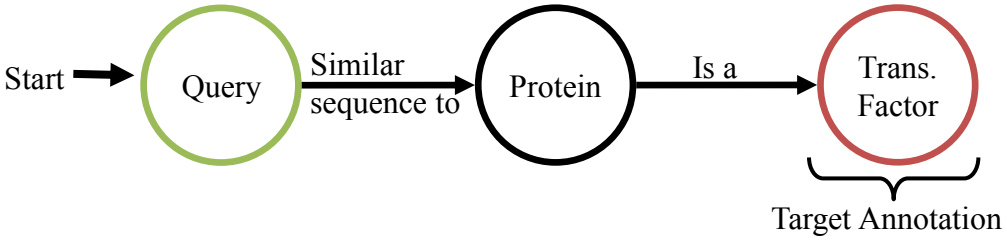


Figure 3.6: An example of valid and invalid paths through a graph. Key to relation types a has similar sequence b catalyses c of class d is equivalent to e interacts with f produced by g of class h consumed by i has function. The path $\{a, b, c, d\}$ is valid, the paths $\{a, e, i\}$ and $\{a, b, f, h, g, d\}$ are invalid.

clause to include annotations via domains: every sequence x that is "similar sequence to" a sequence of concept y, and y is a protein, and y has an "of class" relation to concept z which is of type EC term, or x contains (has part) a Pfam domain of the concept y, which has an "of class" relation to concept z which is of type EC term. This can be described in FOL using an inclusive disjunction ($\vee$), which is shown in Equation 3.5. This FOL equation describes valid the assignment of the EC term z to the sequence x via the intermediate concept y. The query graph equivalent that was used to encode this rule is shown in Figure 3.7.

$$\forall x, r', y, r'', z \; where \; (x \in C, r' \in R, y \in C, r'' \in R, z \in C) :$$
$$type(x) = \text{“}query\text{”} \wedge type(r') = \text{“}similar\ to\text{”} \wedge \{x, r', y\} \wedge$$
$$type(y) = \text{“}protein\text{”} \wedge type(r'') = \text{“}is\ a\text{”} \wedge \{y, r'', z\} \wedge type(z) = \text{“}EC\text{”}$$
$$\vee$$
$$type(x) = \text{“}query\text{”} \wedge type(r') = \text{“}haspart\text{”} \wedge \{x, r', y\} \wedge$$
$$type(y) = \text{“}domain\text{”} \wedge type(r'') = \text{“}of\ class\text{”} \wedge \{y, r'', z\} \wedge type(z) = \text{“}EC\text{”}$$
$$\implies z \text{“}describes\text{”} x$$

(Equation 3.5)

The rules that were encoded as query graphs in Figure 3.6 and Figure 3.7

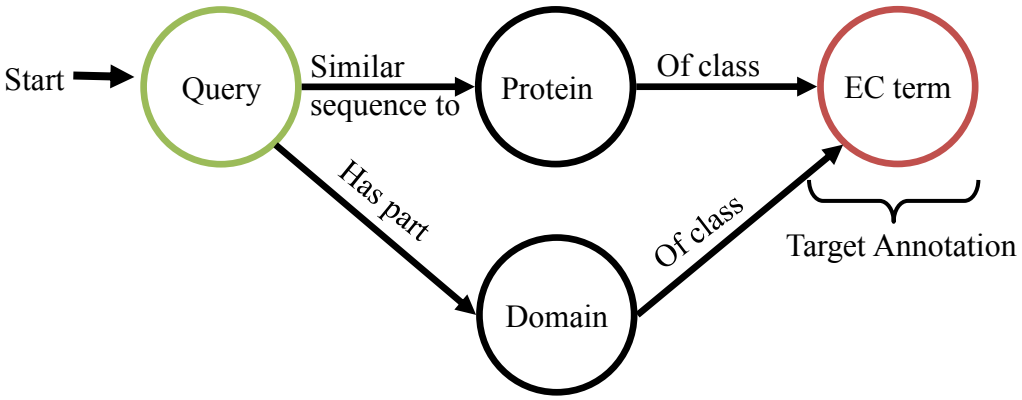


Figure 3.7: The encoding of a simple biological rule that defines valid inference paths from a query sequence to EC terms.

formed a sub-part of the rule that encodes the annotation of GO categories to an array sequence, as their products form intermediates that can be annotated via inference to GO terms themselves. Figure 3.8 shows the query graph for the concepts and relation types that encode the rules that govern the extraction of annotations from an array consensus sequence to entities from all three GO categories. Given the size of the rule, a textual description and FOL definition has been omitted but can be derived from the inclusive disjunction of the conjunction of steps from sequence to GO entity in Figure 3.8.

Execution of the query graph in Figure 3.6 yielded not only the annotation terms for each sequence but the full history of the path from which the term

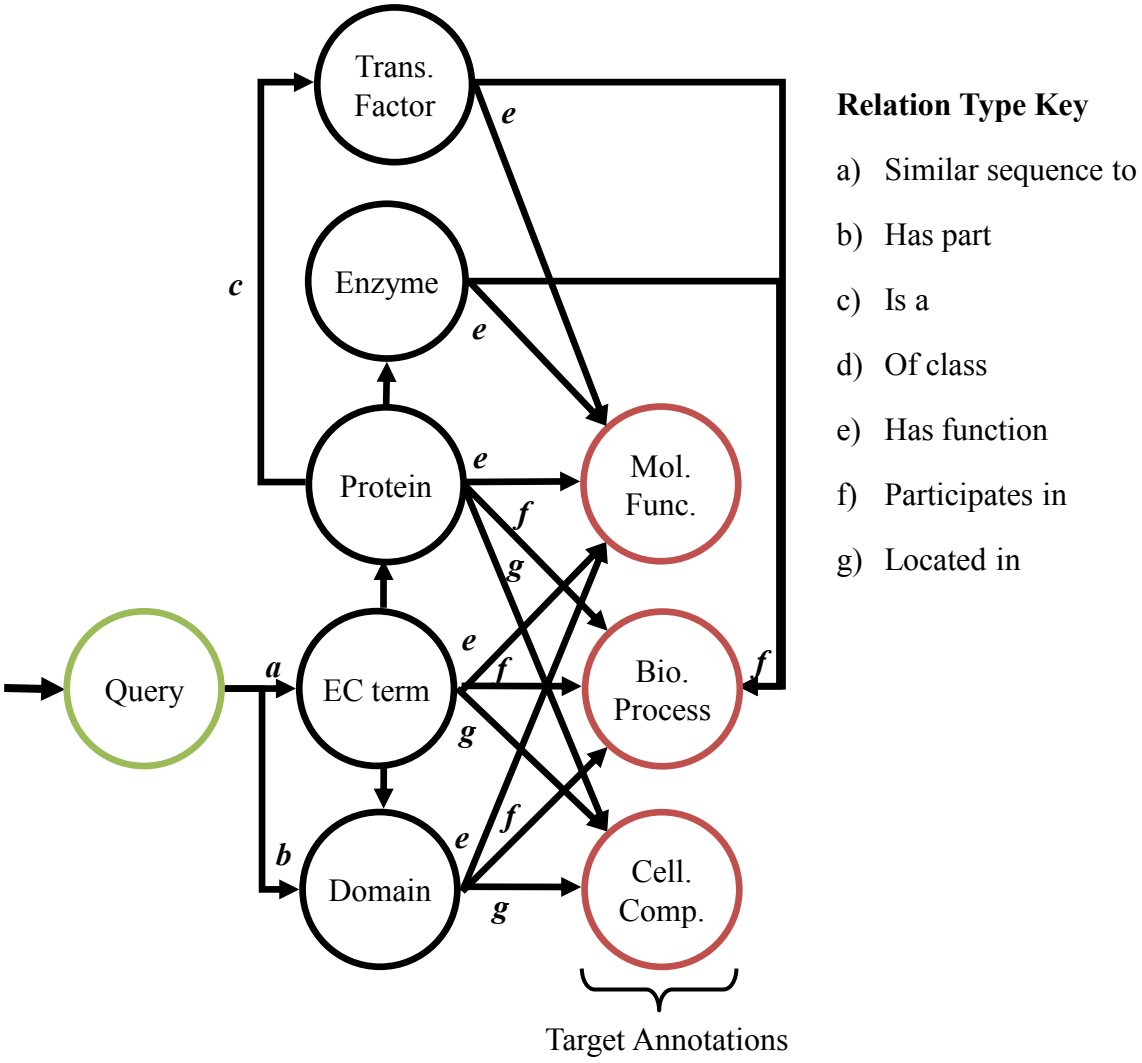


Figure 3.8: The encoding of biological rules that defines valid inference paths from a query sequence through to GO category entities (*molecular function*, *biological process* and *cellular component*).

was inferred. This was further exploited in evaluating provenance of the candidate annotation terms, based on the source of the elements traversed in the path and properties stored in the graph. This was used to evaluate the data sources of candidate annotation, retrieve scores on *similar sequence* relations, and GO evidence codes on *has function*, *participates in*, and *located in* relations.

3.2.1(*g*) Correcting for domain annotation ambiguity

An important complexity in domain annotation using InterPro arises because protein domain assignments are curated or inferred in the InterPro database from the existing annotations of the proteins that share that domain (Hunter

et al., 2009). InterPro therefore reflects the potential roles of the domain rather than its unambiguous function. This means the annotation of GO and EC terms, based on domains, may also be ambiguous depending on the specificity of the Pfam domain. Two separate strategies (exploiting the structure of the EC hierarchy and GO graph) were adopted in order to select the annotation appropriate to the specificity of the domain. In both instances the assumption is made that domains with a large number of conflicting annotations are more general, and require a more abstract annotation that unifies the annotations by using a term from a higher-level in the hierarchy.

In order to generalize conflicting terms, the Most Informative Common Ancestor (MICA) was used to find the most informative annotation term from the hierarchy of annotations. MICA was originally as proposed by Resnik (1999) and later applied by Lord *et al.* (2003) to the Gene Ontology. The MICA definition of *most informative* is based on Information content (IC), which is a useful proxy measure of the specificity of the GO terms and accounts for the variability in the granularity of the GO graph (Lord *et al.*, 2003). IC is based on the assumption that more specific (information rich) terms will be referenced less in a given set of annotation. The Shannon (2001) information-theoretic measure can be thus applied to give the IC, based on the probability of the occurrence of a term i. The calculation of IC from the probability p of the occurrence of a term i is given in Equation 3.6.

$$IC(i) = -\log_2 p(i) \qquad \text{(Equation 3.6)}$$

Within CoPSA the probability of the occurrence of a term i was defined based on the frequency of the annotation for that term and its set of ancestors (A), to the set of Pfam domains (D), within a given ontology category of the Gene Ontology (O). Ancestors of a term are defined as all terms in GO that subsume the term through the *is a* or *part of* relations. This is shown in Equation 3.7.

$$p(i) = \frac{\sum_{t \in A(i,O)+t} |D(t)|}{\sum_{j \in O} |GP(j)|} \qquad \text{(Equation 3.7)}$$

Generalization of domain terms involves the following steps: for each domain annotation, clusters of conflicting terms in an annotation-set were identified as terms that share at least one common ancestor, excluding the root term of that category. These conflicting terms were then removed from the annotation and replaced by the most informative term of the common ancestors. Where there are multiple candidates for MICA, all were taken. Figure 3.9 shows an example for the Carbamoyl-phosphate synthetase large chain, oligomerisation domain, CPSase_L_D3 (PF02787). In this instance the algorithm identified a cluster of five seemingly conflicting GO terms, and replaced them with the more general *ligase activity*. The limitation of this approach is that it does not always differentiate conflicting annotation from multi-functional domain annotations. However, owing to the topology of GO, many of these multi-functional and coherent annotations are divided at the root and therefore qualify for separate clusters, as was found in the given example for ATP binding.

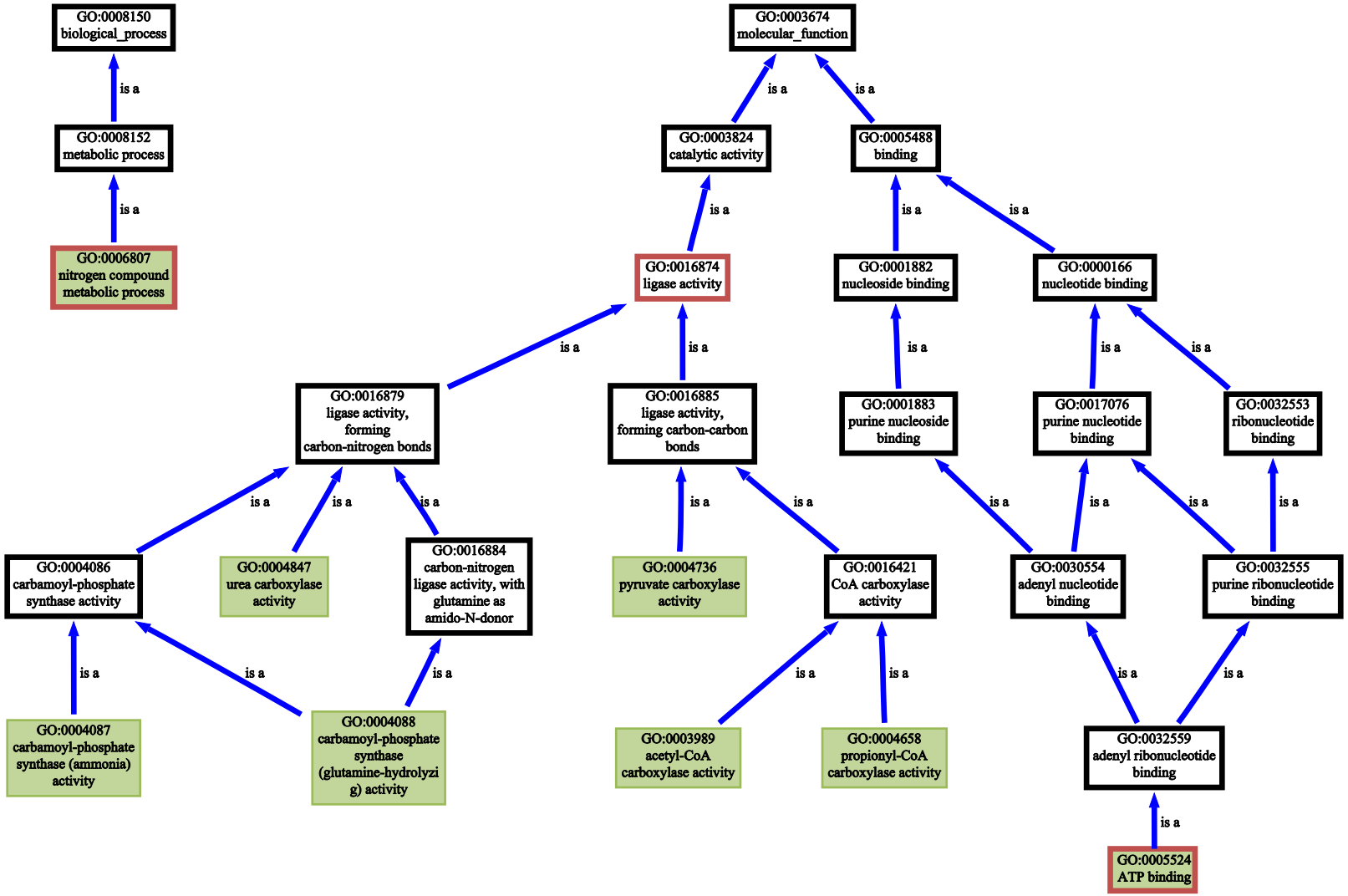


Figure 3.9: Generalization of conflicting annotations for the Pfam domain CPSase_L_D3 (PF02787) using MICA. Green nodes are annotations and red bordered nodes are the MICA selections.

Generalization of the terms for EC annotations was less complex given that EC is a hierarchy as opposed to a graph. A similar approach was therefore taken to that used for GO-term generalization, with the highest numerical level (most specific) common ancestor in the EC tree being used to generalize a set of EC terms. The generalization algorithm was executed on a per-domain basis, with the first-level EC terms being the greatest generalization possible. Where a set of EC terms that annotate a domain includes multiple first-level categories, generalization was performed for each tree, with the level-one EC term being the root of the tree. For example, the CPSase_L_D3 domain contains six candidate EC annotations, from the Ligase family of EC terms (Table 3.7). The highest level unifying EC category is Ligases (6.-.-.-), which maps to the GO Ligase activity term (GO:0016874). Figure 3.10 shows the EC hierarchy on which this conflict resolution is based.

Table 3.7: Annotations for the Pfam CPSase_L_D3 (PF02787) domain

EC code	Official name
6.3.4.6	Urea carboxylase
6.3.4.16	Carbamoyl-phosphate synthase (ammonia)
6.3.5.5	Carbamoyl-phosphate synthase (glutamine-hydrolyzing).
6.4.1.1	Pyruvic carboxylase
6.4.1.2	Acetyl-CoA carboxylase.
6.4.1.3	Propionyl-CoA carboxylase.

3.2.2 CoPSA metrics for selecting putative functional-orthologs

Previously in this Section the methodology for building an integrated knowledge-base and extracting annotation has been described. In the final part of this Section, the methodology for selecting the best annotations from the candidates is described. Five different methods for scoring candidate annotation sets (each set corresponding to inferences from a protein with a similar sequence) are

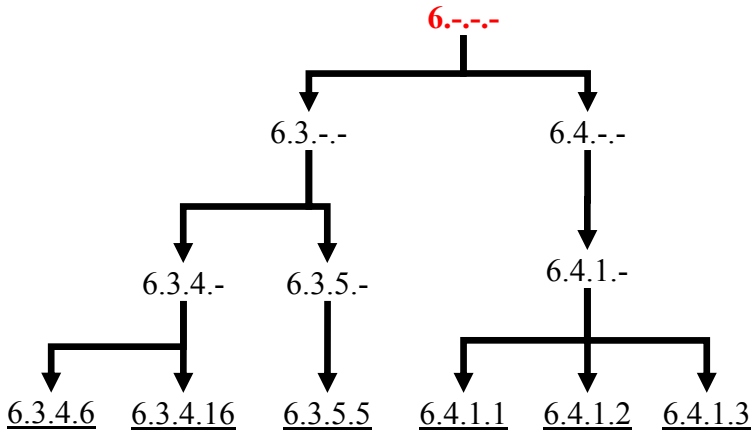


Figure 3.10: The generalization of conflicting EC terms for a domain annotation. Original annotations are underlined and the generalized candidate annotation is highlighted in red.

presented. Two of these methods consider the range of evidence supporting an annotation and the remaining three are pragmatic approaches to selecting annotation using the bitscore from BLASTx derived protein sequence alignments of query nucleotide to candidate proteins. After describing each method their performances are compared.

3.2.2(*a*) Best hit approach (CoPSA-BestHit)

The simplest approach to selecting the best annotation for a new sequence retains the hit with the highest BLAST bitscore and at least one functional annotation; all other candidate sequences with annotation are excluded. Where multiple top sequence alignments have the same bitscore, the union of their annotations is taken. This is a naive approach the limitations of which have been described in 3.1.2(*a*). This approach does not consider provenance of annotation and may therefore miss high quality hits of sub-optimal alignment.

3.2.2(*b*) Union above threshold (CoPSA-Union)

The most inclusive approach takes the non-redundant union of all possible annotations for each query sequence. It is then necessary to remove any GO structural redundancy from the resulting set. Structural redundancy is defined as a term that subsumes another by *is a* or *part of* relations in the GO hierarchy. BLAST runs in CoPSA with a cut-off threshold of $< 1 \times 10^{-4}$ and a bitscore cut-off of > 50. This approach is therefore analogous to executing a BLAST search at these thresholds and taking all annotations that meet these score thresholds. It is also unaware of provenance, and will take annotations with weak evidence, even when strong experimental evidence is available.

3.2.2(*c*) Best hits within a weighted threshold (CoPSA-ε)

A compromise between the best-hit and union approaches is to take the best-hit together with all hits within a threshold of similarity, this was determined by the user defined weight ε, where the threshold for transfer of the functional annotation of a putative functional-ortholog (g) to a sequence (p) in a functional annotation set (FA) was:

$$threshold(p, FA) = \max_{g \in FA} bitscore(p, g) \times \varepsilon \qquad \text{(Equation 3.8)}$$

For the purposes of these comparisons a ε value of 0.9 was used, which selected hits with less than 10% deviation from the bitscore of the best BLAST hit. This value was chosen as a pragmatic cut-off, which might reasonably be chosen by a user wanting to find BLAST hits broadly similar to a query sequence. Again this method is unaware of provenance.

3.2.2(*d*) Multiple weighted fitness measure (MWFM)

The annotation scoring metric which has been developed most fully and considered to be the preferred metric for CoPSA is the Multiple Weighted Fitness Measure (MWFM). In evaluating the relative confidence of a GO functional annotation, the MWFM method considers three sources of evidence:

1. Similarity of the target sequence to query
2. GO Evidence code (represents annotation evidence provenance)
3. Semantic similarity of a protein annotation set to other candidate annotation sets (captures diversity and therefore subfunctionalization in the gene family)

The first source of information for consideration is the similarity of the sequences under consideration. The assumption is that the higher the quality of the alignment the greater the likelihood that the two sequences lie within the same gene family and therefore share the same function. The unidirectional bitscore is used, because calculating a bidirectional hit was both computationally expensive, and less meaningful for incompletely sequenced wheat genome. It also makes the pipeline more flexible, as any sequence, irrespective of organism can be considered, even if it is the only sequence available in a given organism.

The second source for consideration is the providence of the annotation on the proteins that have been aligned to the query sequence. The experimental evidence for a GO annotation is captured by its GO evidence code. These were weighted based on the trustworthiness of the confidence, which are shown in Table 4.1. These weightings were adapted from the default values used by BLAST2GO, within the "b2gPipe.properties" file available as part of their tool download package . The IEA code for electronic annotation is given a much lower (0.2) weighting however compared to BLAST2GO (0.7). As with

BLAST2GO, the annotations that are backed by experimental observations are given the greatest confidence weighting.

The third piece of information considered by CoPSA-MWFM is the semantic similarity of function between the sets of annotation, where each set contains functional-annotations for a candidate protein aligned to the query sequence. This gives a measure of the diversity of function within the protein family. If the all the hits have very similar annotation then there is high confidence that it is a functionally conserved family, and the wheat sequence is unlikely to be a sub-function. If one of the best hits is very different to the rest of the family, the probability is that this is a subfunctionalization and the highest probability is with the major functions of the family.

For each of these sources of information, a confidence score between zero and one was assigned, with the overall confidence in a protein and its annotation being a product of these three considerations. These confidence metrics were defined in three functions: sequence structural similarity $sqs(p,g)$, mean evidence code weighting $mew(p,g,O)$, and mean semantic similarity to all other sets of annotation $mss(p,g,O)$. Where, g is the query sequence, p is the candidate protein, and O is a category of the Gene Ontology. The overall conference for an annotation is summarised in Equation 3.9.

$$confidence(p,g,O) = sqs(p,g) \times mew(p,g,O) \times mss(p,g,O) \quad \text{(Equation 3.9)}$$

The function $sqs(p,g)$ is the similarity of the protein p to the query sequence g expressed as a fraction of the highest scoring similarity within the set of all proteins P found to be similar to the query sequence g (for the given thresholds) (Equation 3.10).

$$sqs(p,g) = \frac{bitscore(g,p)}{\max_{i \in P} bitscore(g,i)} \quad \text{(Equation 3.10)}$$

The mean evidence weight was based on a lookup function $ecweight(t)$ that translates the evidence code of a functional annotation via a lookup table shown in (Table 3.8) to a weight between zero and one (these weightings are configurable by a user). The maximum scoring evidence code of all of the functional annotations F for the candidate protein p in the given category O was taken to indicate the highest potential quality of annotation that can be derived from the candidate protein (Equation 3.11).

$$mew(p,g) = \max_{a \in F(p,g,O)} ecweight(a) \qquad \text{(Equation 3.11)}$$

For calculations of semantic distance, the asymmetric GS2 measure (Ruths *et al.*,

Table 3.8: Confidence weighting of GO evidence codes, as used in MWFM.

Evidence	Name	Weight
NAS	Non-traceable Author Statement	0.9
TAS	Traceable Author Statement	
IEA	Inferred from Electronic Annotation	0.2
IGC	Inferred from Genomic Context	0.9
ISA	Inferred from Sequence Alignment	
ISM	Inferred from Sequence Model	
ISO	Inferred from Sequence Orthology	
ISS	Inferred from Sequence or Structural Similarity	
RCA	Inferred from Reviewed Computational Analysis	
IC	Inferred by Curator	0.9
ND	No biological Data available	0.2
EXP	Inferred from Experiment	1
IDA	Inferred from Direct Assay	
IEP	Inferred from Expression Pattern	
IGI	Inferred from Genetic Interaction	
IMP	Inferred from Mutant Phenotype	
IPI	Inferred from Physical Interaction	

2009) for comparison of term sets is used. This gives a measure of semantic similarity between the functional annotations of g inferred from p, with all other functional annotations of g, which are inferred from $P - \{p\}$, where P is the set

of all candidate proteins for the query sequence.

$$mss(p, g, O) = gs(p, P - \{p\}) \qquad \text{(Equation 3.12)}$$

GS2 is based on a rank function for a term t within the functional annotation of a set of genes G. The rank was simply a count of how many times a term appears within the functional annotations of a gene set and the ancestors $A(i, O)$ of those annotations.

$$rank(t, G, O) = \left| \left\{ g \in G | t \in \bigcup_{a \in F(g,O)} \{i\} + anc(i, O) \right\} \right| \qquad \text{(Equation 3.13)}$$

Based on the occurrence rank function it was then possible to derive the GS2 metric that compares for each set of annotations on a sequence, the similarity of a gene product derived annotation set p with the annotations inferred from the remaining gene products $P - p$.

$$gs(p, P - \{p\}) = \frac{1}{F(p, g, O))} \sum_{t \in F(p,g,O)} \frac{1}{anc(t, O)} \sum_{a \in anc(t,O))} \frac{rank(a, P - \{p\}, O)}{|P - \{p\}|} \qquad \text{(Equation 3.14)}$$

The three different measures of confidence are combined by calculating their product; thereby ascribing equal weight to each measure. The annotation set selected by MWFM is derived from the protein that constitutes the best compromise in semantic consensus with other protein annotations, while maximising the quality of evidence and sequence similarity against the query sequence.

3.2.2(*e*) Post-optimisation of MWFM to prioritise high quality evidence (MWFM-OE)

MWFM was used to select the protein with the optimum potential to transfer consistent, high quality evidence, from a protein with the highest similarity in another species. Each protein that MWFM selects to transfer annotation from

can have multiple functional annotations, each of which has different providence. For example: there may be strong experimental evidence for the catalytic role of a protein, but only electronically inferred evidence for protein-protein binding.

In order to improve confidence in the overall annotation of the protein, it is possible to exclude low quality annotation from being transferred from the MWFM selected protein to the query sequence. Simply removing all electronically inferred annotation prior to MWFM would have increased overall confidence, however for the cases where no experimental annotation could be found, coverage would have been greatly reduced. Removing the relatively low confidence annotation after the MWFM process means that query sequences that only have a good sequence alignment to a protein with electronic annotation, still have annotations suggested by CoPSA. Additionally, where a query sequence matches to a protein with a mixture of low confidence and high confidence annotation, only the high confidence annotation is preserved. This method therefore attempts to Optimise Evidence (OE) for each protein selected by MWFM, such that only the best annotations are reported.

Two intervals for annotation evidence confidence were set at $0.5 \geq x \leq 1$ and $0 \geq x \leq 0.5$, based on the evidence weighting reported in Table 3.8. A protein provided by MWFM for transfer of annotation to a query, was placed into one of the two intervals based on the weighting of its most confident annotation. All other annotations, lower than this interval, were not transferred to the query sequence. The upper interval corresponds to experimental and human curated evidence, the lower to electronically inferred annotation.

3.2.3 Quantitative and qualitative evaluation of GO functional annotations

This following section presents the metrics using to evaluate the qualitative and quantitative properties of functional annotations. These are vital in making comparisons between the functional annotations predicted by CoPSA, BLAST2GO, and Affymetrix, as well as for comparing the various scoring metrics previously proposed in this chapter. It is particularly important to recognise that these are indirect measures of annotation quality, as there are no experimentally verified annotations for the Affymetrix wheat microarray sequences, that could be used to provide precision and recall statistics. Although the high quality NetAffx annotations can form a point of comparison, using hierarchical recall. The metrics contained in this section are existing measures, or have been defined as part of the AIGO project. All these measures are present within the open source AIGO project, which has been implemented as an open source python library in collaboration with Michael Defoin-Platel.

3.2.3(*a*) Richness

Richness of annotation is simply the proportion of a gene Ontology category used in the functional annotation of a set of genes FA, for a given GO category O. Richness is calculated by taking the number of terms t within the category O used within the functional annotation FA and dividing this by the total number of terms in the gene ontology category. Equation 3.15 used to determine the richness of a functional annotation FA with respect to an ontology category O. The function $P(t)$ gives the number of genes that are annotated with t. Note that if a gene annotates a term, it is considered to annotate all the parents of

that term via the *is a* and *part of* relations.

$$richness(FA, O) = \frac{1}{|O|} |\{t|t \in O, |P(t) > 0|)\}| \qquad \text{(Equation 3.15)}$$

3.2.3(*b*) Structural specificity

The structural specificity of annotation is a very simple measure of the depth of the term in the GO hierarchy, as defined by the number of ancestors of the functional annotations of a gene in an ontology category. Specificity should be used with care as it makes an assumption that there is an equal semantic-distance in all *is a* and *part of* relations, which is not always correct. The semantic-distance represented by these relations is a measure of the difference in specificity between the parent and child term. The varies widely across the Gene Ontology, depending on the sub-domain of the ontology. For example: iron ion binding (GO:0005506) is 6 terms removed from the root in the GO DAG and has 10,168 genes annotated to it, whereas chloride channel inhibitor activity (GO:0019869) is 4 terms removed from the root and only has one gene annotated to it. Clearly, the latter is the more specific term but has the lower depth score.

3.2.3(*c*) Semantic coherence

The coherence of the annotations of a gene g within a FA is the mean semantic distance of its elements according to the asymmetric GS2 measure (Ruths *et al.*, 2009) that has previously been described in this chapter. Coherence is formally defined in Equation 3.16.

$$coherency(g, O) = \frac{1}{|FA(g, O))|} \sum_{i \in FA(g,O)} GS2sim(\{i\}, FA(g, O) - \{i\}) \qquad \text{(Equation 3.16)}$$

3.3 Results and discussion

This section presents the an analysis of CoPSA annotation, first with respect to the constructed knowledge-base and second with respect to the annotations. The quality of the final annotation is dependent on the coverage and quality of the background information available to the annotation pipeline and the effectiveness with which raw annotations can be extracted from these sources. Section 3.3.1 evaluates the potential annotations discovered by applying the query graph to the integrated knowledgebase. The relative contribution of the BLAST and HMMR methods in conjoint alignment were compared, as well as the efficacy of data aggregation and integration in sequence annotation. In the second part of this section these candidate annotations is subject to further quality-evaluation measures, with an assessment of the quality of the final processed annotations.

Section 3.3.3 is an evaluation of CoPSA annotations and candidate metrics for annotation ranking and selection. It is broken down into three aspects. Firstly, a consideration of confidence of annotation is made for each of the scoring metrics proposed within CoPSA. Confidence is assessed based on the original evidence of this annotation and the similarity of sequences. Secondly, a range of properties of the annotation are examined for each of the metrics and Blast2GO and NetAffx annotations. Thirdly, NetAffx is used as an incomplete but high quality source of gold standard annotations. The ability of CoPSA to correctly annotate genes with the same or similar annotations to NetAffx is assessed using the hierarchical recall metric proposed by Verspoor *et al.* (2006), which has been described in Section 3.3.3(*c*).

3.3.1 Analysis of the CoPSA knowledge-base

3.3.1(*a*) Evaluation of the quantity of annotation from conjoint alignment

The quantity of annotation for a given set of sequences is dependent on the ability to find similar protein sequences and domain regions, and the efficiency with which these candidate similar protein and domain regions can be used as annotations. This section begins by considering the quantity of annotation that can be derived from BLAST and HMMR alignment, then goes on to consider the efficiency with which CoPSA translates these into annotations.

Quantifying protein and domain candidates from conjoint alignment

A key step common to all the query graphs used in CoPSA, which have been described in Section 3.2.1, is the ability to identify similar protein sequences and Pfam domain regions, which are dependent on the BLAST and HMMR sequence alignment algorithms. As part of an annotation process, these steps were evaluated by first addressing the question: what was the relative contribution of protein sequence alignment (using BLAST) and domain identification (using HMMR) in identifying candidate annotation for the query consensus sequences, for each of the species GeneChips being analysed? A simple measure of coverage measures this contribution as the proportion of sequences that have at least one annotation for the given method. This can be presented using a Venn diagram, which is constructed from the coverage provided by each method, where the disjunctions show the two exclusive sources of annotation (BLAST or HMMR) and the conjunctions indicate when both methods conjointly provide at least one candidate annotation (BLAST and HMMR).
Figure 3.11 shows such a stacked Venn diagram for the coverage of 13 Affymetrix arrays, summarising the calculated proportion of sequences that have

exclusive annotation or are conjointly annotated with a domain or similar protein. The chart was created by looking up the provenance of the annotation on each sequence, which is retained by CoPSA. The tracking of provenance on graph queries is described in Chapter 2. This first step of the query, from query sequence to protein or domain concept, indicates the alignment method on which the final annotation is based. The methodology is used throughout this chapter in providing statistics on the providence of annotation proposed by CoPSA.

Figure 3.11 highlight that the first stage of CoPSA, identifying protein and domain similarity, is a limitation to further GO or EC annotation. Superficially, it appears that very few sequences are annotated using Pfam domains alone, and it could be concluded that protein sequence similarity (using BLAST) was sufficient to provide a potential link to function annotation. However, the simple coverage statistic shown in the Figure 3.11 does not show the potential for an identified Pfam domain or similar protein sequence to be linked to functional annotation. The quality of annotation derived from each method is also important, and is reflected in the specificity of function, and the breadth of functions captured for a microarray sequence. A Pfam domain based annotation system breaks up functional annotation into short sequence motifs, which may be detected multiple times on a given query sequence. A protein sequence similarity based annotation however is limited to annotating a query sequence with a whole protein, which may only contain a partial alignment, and consequently a limited common functionality. This is particularly relevant for query sequences with multiple functional regions that do not have similar sequence with that exact combination (or order) of domains. Therefore, the presence of both a similar protein sequence and identified domain provides greater potential to infer functional annotation and is an additional, and valuable, source of evidence.

Based on coverage alone it is difficult to judge the contribution of these two sequence-based annotation methods, to the overall pipeline. However, as key

components in the annotation pipeline, it was possible to gauge the maximum potential for annotation in each Affymetrix chip given the data and parameters used in this execution of CoPSA. From Figure 3.11 it was also apparent that conjoint alignment has the lowest potential for annotation in wheat, and not surprisingly the *Arabidopsis* ATH1 array was not limited at all by this step, with protein equivalents found for 98% of sequences. This indicates that the sequence annotation problem for wheat was much more limited by a shortage of putative functional-orthologs than other species. This is not surprising given the point of divergence between monocot and eudicot is estimated at 150-300 million years ago (Feuillet and Muehlbauer, 2009) (Chapter 3.1.2(*a*)), and the reliance on so much primary annotation from *Arabidopsis* when predicting wheat gene function.

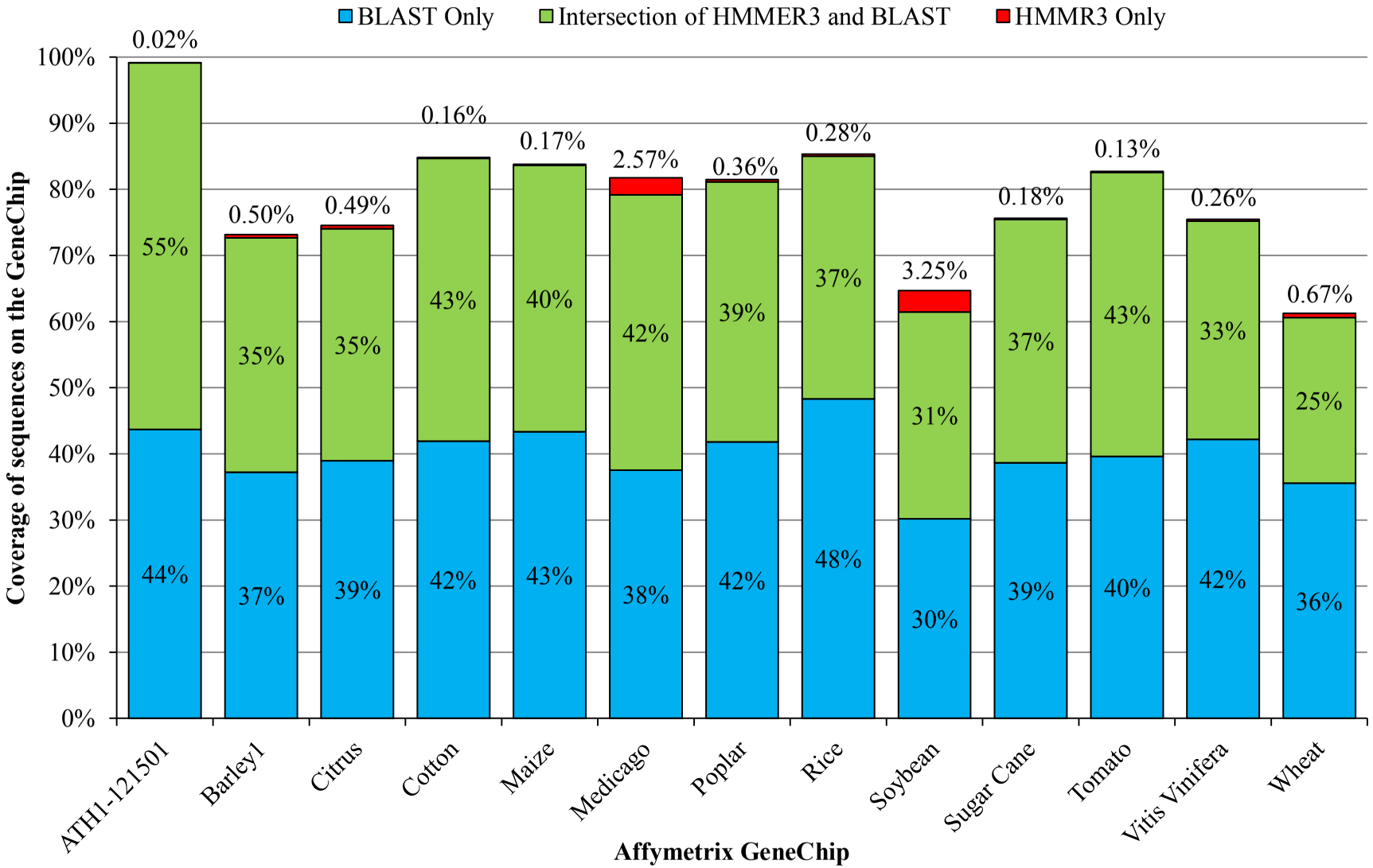


Figure 3.11: A stacked Venn showing the relative proportion of Affymetrix consensus sequences with hits in conjoint analysis. Green shows the proportion of the consensus sequences that have at least one BLAST homologous-sequence hit and a HMMR3 Pfam domain hit. Blue and red shows the proportion of sequences that have only BLAST or Pfam hits, respectively. The sum of all components of the stacked bar shows the overall proportion of consensus sequences on the chip that can be potentially annotated in this CoPSA run.

Once the related proteins and domains for a consensus GeneChip sequence have been identified by the query graph inference engine, the further inferences of gene ontology annotations become dependent on the volume and quality of GO terms that can be identified as annotations for these entities. With respect to genes with at least one GO annotation, Figure 3.12 shows the mean coverage of the plant microarray chips. This confirmed that the lower coverage of domain compared to protein annotations reported in the previous subsection, translates into a reduced ability to annotate GO functions on the chip. It is also apparent from Figure 3.12 that coverage also varies across the GO categories. For ranking of categories by coverage, *molecular function* was the most abundant of the categories to be annotated through HMMR and BLAST methods. However, the ranking of the remaining categories differed between methods, with *biological process* being more abundantly annotated through HMMR domain identification and *cellular component* through BLAST sequence alignment.

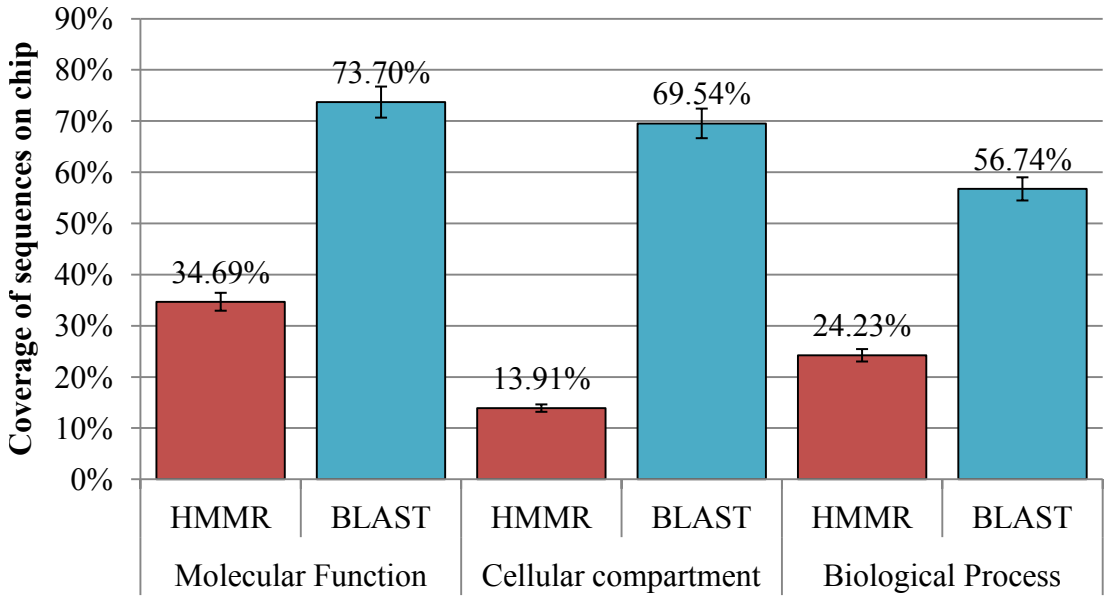


Figure 3.12: Mean sequence annotation coverage of Affymetrix chips for the three categories of the Gene Ontology, comparing for each BLAST and HMMR methods. Error bars are standard error of the mean.

Pfam functional domains detected by HMMR are more closely aligned with GO molecular function terms, and this probably explains the higher conversion efficiency seen in Figure 3.14. Some Pfam domains like *peroxidase* (PF00141) can be quite general and only resolve to a 3 digit EC term (*i.e.* 1.11.1.-), which could be involved in many potential pathway processes, so there can be more ambiguity in assigning a GO process to a Pfam domain. The lower Pfam domain to GO *cellular component* conversion efficiency may be caused by domains being active in many cellular locations. For example Glutamine Synthetase has a highly conserved domain (Pfam: Gln-synth_C) but may be expressed in the Golgi apparatus, cytosol or mitochondrion. The final cellular location is dependent on small changes in the non-domain region of the sequence (Bernard and Habash, 2009). However, some domains such as DNA binding and transmembrane proteins do clearly indicate *cellular component*. For these reasons, protein sequence similarities (identified using BLAST) with a query sequence were more efficient in obtaining cellular-component and molecular-process predictions for the query than through the identification of Pfam domains (using HMMR).

The coverage reported in Figure 3.12 is expanded in Figure 3.13 by a stacked Venn representation, showing the coverage of sequences annotated exclusive using protein sequence similarity (using BLAST) or Pfam domain identification (using HMMR), or by both methods (HMMR+BLAST). This shows not only the utility of each method on its own for annotating microarray chips in CoPSA, but its contribution with respect to a conjoint analysis. Coverage is provided, on a per-chip basis, to highlight the variability of each methods contribution across species. For the Affymetrix Soybean GeneChip, annotation to Pfam domain via HMMR provided candidate GO function annotations for 32% of the chip, and 3% of the sequences on the chip could annotated exclusively via this method. In contrast, for the wheat GeneChip, the contribution of HMMR is much lower relative to BLAST and the exclusive contribution by HMMR was negligible.

In terms of the quantity of sequences that can be annotated, the wheat GeneChip was the most difficult to annotate against GO (Figure 3.13). At least one similar protein or domain region was found for 61% of sequences on the chip; however, for GO *molecular function*, *biological process*, and *cellular component* at least one term could be found for 56%, 55%, and 43% of sequences, respectively.

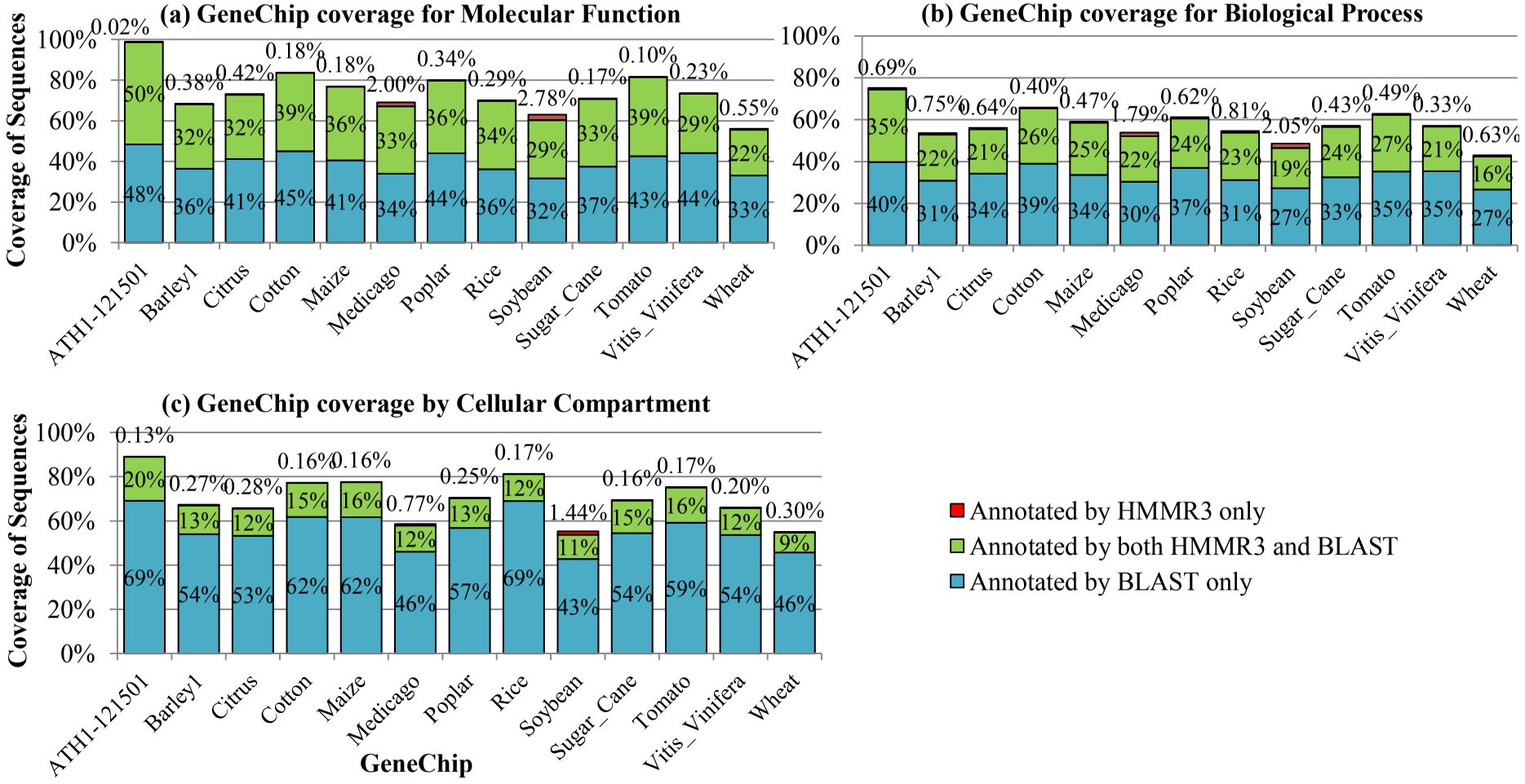


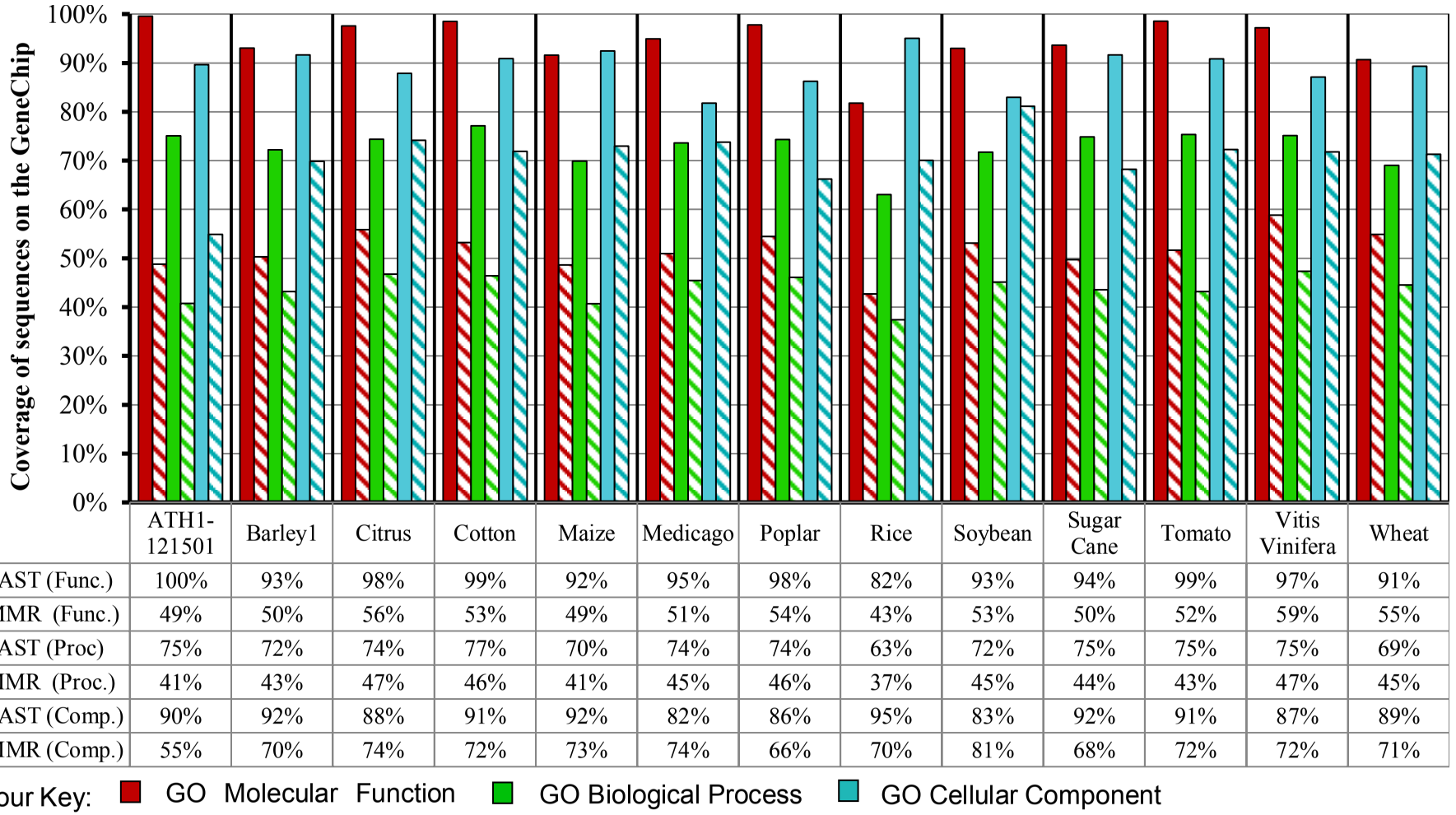
Figure 3.13: Proportion of probe-sets on each GeneChip with a GO (a) *molecular function*, (b) *biological process* and (c) *cellular component* annotation. Stacked Venn bars show the proportion uniquely and conjointly annotated by each method.

Previously, it was established that the lower domain annotation coverage obtained also resulted in lower coverage of GO annotations relative to protein based sequence similarity methods. However, this does not indicate protein sequence similarity based on BLAST is more effective for GO annotation, as sensitivity differences could have resulted from the individual parameter settings of methods. However, the efficacy with which domain or protein-similarity annotations result in a final CoPSA GO term prediction, does indicate the overall usefulness of the method. Figure 3.14 shows efficiency, measured as a percentage of domain or proteins-similarity annotations that also result in at least one GO annotation.

This lower conversion efficiency, together with the smaller quantity of sequences with identified Pfam domains (using HMMR) resulted in a far smaller proportion final candidate annotations through this method (Figure 3.12, Figure 3.13). Of those annotations based on identified Pfam domains, the majority are also annotated with protein sequence similarities (using BLAST). For the HMMR identified domains methodology there is a consistent pattern that GO function had the greatest conversion efficiency (Figure 3.14) and coverage (Figure 3.13), followed by GO *cellular location*, and then GO *biological process*.

As described previously for GO term coverage, the efficiency of a method in predicting GO annotations is species dependent. There is a marked reduction in the efficiency with which BLAST annotations can be converted to GO terms in non-*Arabidopsis* organisms, with the grasses being among the most difficult to annotate (Figure 3.14). The Rice Gene Chip has the lowest annotation-conversion efficiency. This indicates that evolutionary distance from *Arabidopsis* not only results in fewer protein-similarity predictions, but these annotations are less likely to result in GO annotations. This may indicate the less well-studied protein families in *Arabidopsis* are more prevalent in these organisms. Almost all the organisms show an improvement in Pfam to GO annotation conversion efficiency relative to *Arabidopsis*. This is a feature of HMMR

domain cut-offs, which are curated for each Pfam family. Domains that are less functionally characterised have more stringent cut-off thresholds, so are not as frequently annotated for evolutionary distant sequences. By consequence for species that are evolutionary distant from the model organisms, HMMR is restricted to more highly conserved domains, with a high efficiency for functional annotation.



	ATH1-121501	Barley1	Citrus	Cotton	Maize	Medicago	Poplar	Rice	Soybean	Sugar Cane	Tomato	Vitis Vinifera	Wheat
BLAST (Func.)	100%	93%	98%	99%	92%	95%	98%	82%	93%	94%	99%	97%	91%
HMMR (Func.)	49%	50%	56%	53%	49%	51%	54%	43%	53%	50%	52%	59%	55%
BLAST (Proc)	75%	72%	74%	77%	70%	74%	74%	63%	72%	75%	75%	75%	69%
HMMR (Proc.)	41%	43%	47%	46%	41%	45%	46%	37%	45%	44%	43%	47%	45%
BLAST (Comp.)	90%	92%	88%	91%	92%	82%	86%	95%	83%	92%	91%	87%	89%
HMMR (Comp.)	55%	70%	74%	72%	73%	74%	66%	70%	81%	68%	72%	72%	71%

Figure 3.14: The efficiency of converting protein or domain annotation into GO term annotation. Statistics are shown for each category of GO (differentiated by colour) and each sequence analysis method (differentiated by pattern). BLAST performance is measured by the proportion of sequences with protein annotation that can also be assigned at least one GO annotation. HMMR performance is measured by the proportion of sequences with domain annotation that can also be assigned at least one GO annotation.

Previously, the relative contribution of HMMR and BLAST to the CoPSA annotation pipeline was considered purely in terms of the numbers of sequences annotated (coverage). However, it is also important to consider their separate contributions in terms of the total candidate annotations within CoPSA, because there are multiple candidate annotations per sequence. Figure 3.15 shows for HMMR and BLAST the proportion of all candidate annotations that are exclusive to each method or are redundant with the other method. Exclusive content is defined as annotations for a sequence for which the same term or a child term is not present in the annotation for that sequence using the other method. A redundant annotation is one that shares the same GO term, or is a parent term of an annotation for that sequence by the other method. In many instances GO annotation via protein domain matching with HMMR was redundant with sequence similarity derived annotation via BLAST. However, the redundant evidence via both HMMR and BLAST added weight to the GO annotation, as it established that as well as sequence similarity to a protein of a given function, there is also the preservation of a functional domain. In a number of cases, HMMR also provided unique candidate annotation. For the *Arabidopsis* (ATH1) GeneChip there was also a large amount of consensus between HMMR and BLAST derived annotation, this was due to the success of BLAST in *Arabidopsis* with a 99% sequence coverage (Figure 3.11) and 100% sequence conversion efficiency (Figure 3.14). Additionally in *Arabidopsis* the highest proportion (40%) of the annotations were derived from HMMR. This almost certainly owes its strength to the position of *Arabidopsis* as a model organism and its role in providing reference annotations for gene function from which many of the Pfam domain models and annotations were originally derived.

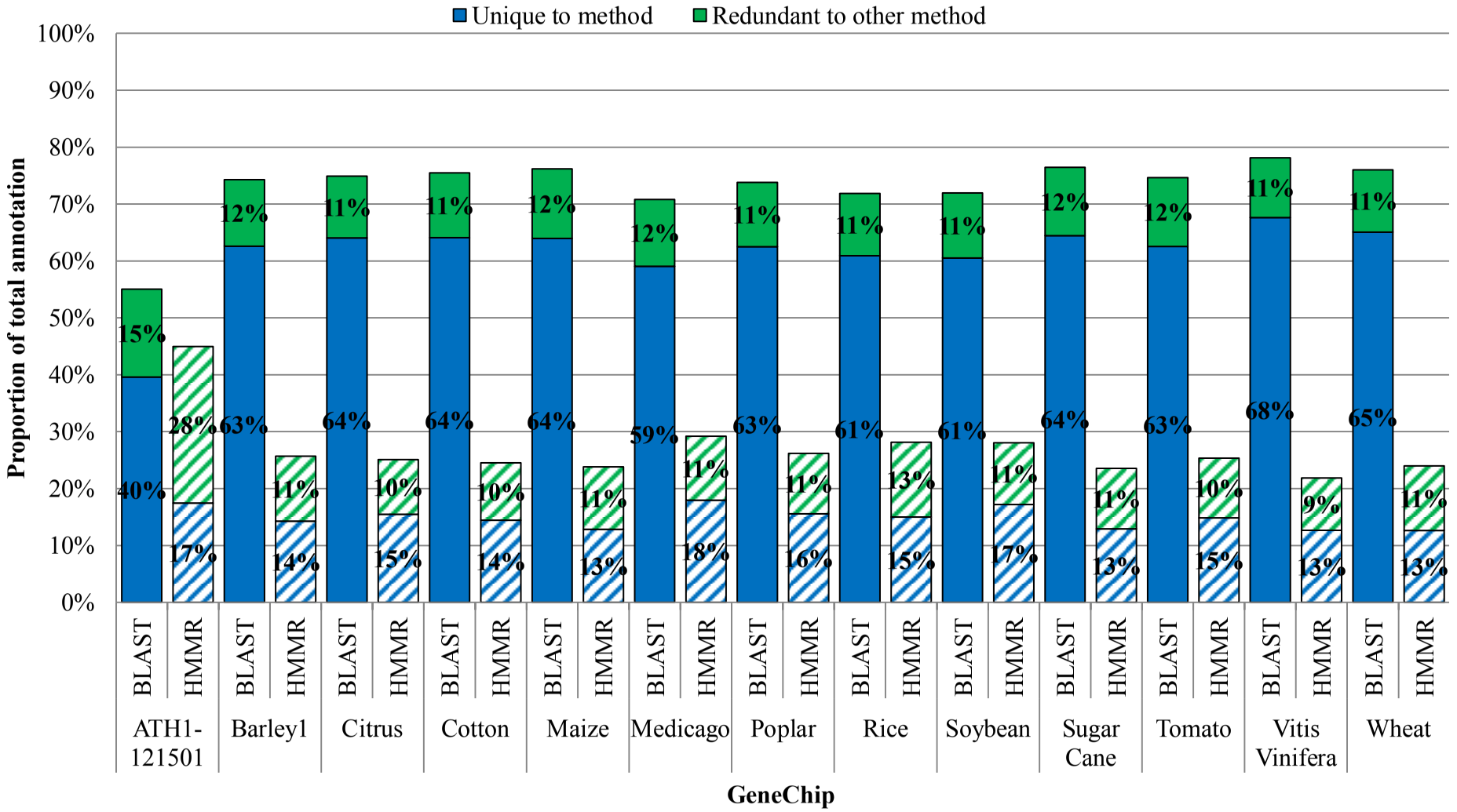


Figure 3.15: The distribution of sequence annotation sources for all unfiltered GO annotations, showing the unique predictions of BLAST and HMMR based annotations. The consensus indicates where there is agreement in GO annotation from both methods: *i.e.* an annotation on a sequence has the same or parent/child annotation on the same sequence from the other source.

Deriving EC annotation from conjoint alignment

As well as annotating GeneChip sequences with GO terms, CoPSA was also used to assign EC terms to sequences. An evaluation of the contribution of BLAST and HMMR methods to EC annotation is provided in addition to their previously described role in GO annotation. The utility of domain and protein-similarity annotations for EC number predictions may differ to that of GO for two reasons. (1) The original sources of EC primary database evidence may not overlap, despite the high compatibility between ontologies in terms of conceptual modelling. This could be caused by omissions or inadequacies in the EC2GO mapping utilised in CoPSA. (2) EC represents a functional subset of the GO ontology, which may be more or less amenable to annotation through a given method.

Figure 3.16 shows a stacked Venn diagram showing the proportion of sequences on each GeneChip annotated with EC terms. The Venn disjunctions shows the sequences exclusively annotated using protein sequence similarity (using BLAST) or identified protein domains (using HMMR). The Venn conjunction is the proportion of GeneChip sequences which have annotation provided by both methodologies. In an analysis of the same Affymetrix GeneChip arrays for plant species as undertaken in the previous two Sections, it was found that finding candidate EC annotations was overall less successful in terms of coverage than finding GO annotations. GO *molecular function*, *biological process*, and *cellular component* annotations on average covered 74% (standard error = 2.92%), 70% (standard error = 2.82%), and 58% (standard error = 2.20%) of consensus sequences for each chip respectively. Whereas EC annotation covers on average 40.62% of the chip (standard error = 1.48%).

On average 22.18% (standard error = 2.20%) of a chip could be annotated with EC using Pfam domain recognition with HMMR alone, and 39.14% (standard error = 1.43%) by protein sequence similarity with BLAST alone. This demonstrates a far greater role for HMMR in annotating EC terms, than was evident

for GO. There is also higher and more consistent contribution of HMMR to EC annotation, with on average 1.48% (standard error = 0.16%) of sequences being uniquely annotated with EC by HMMR.

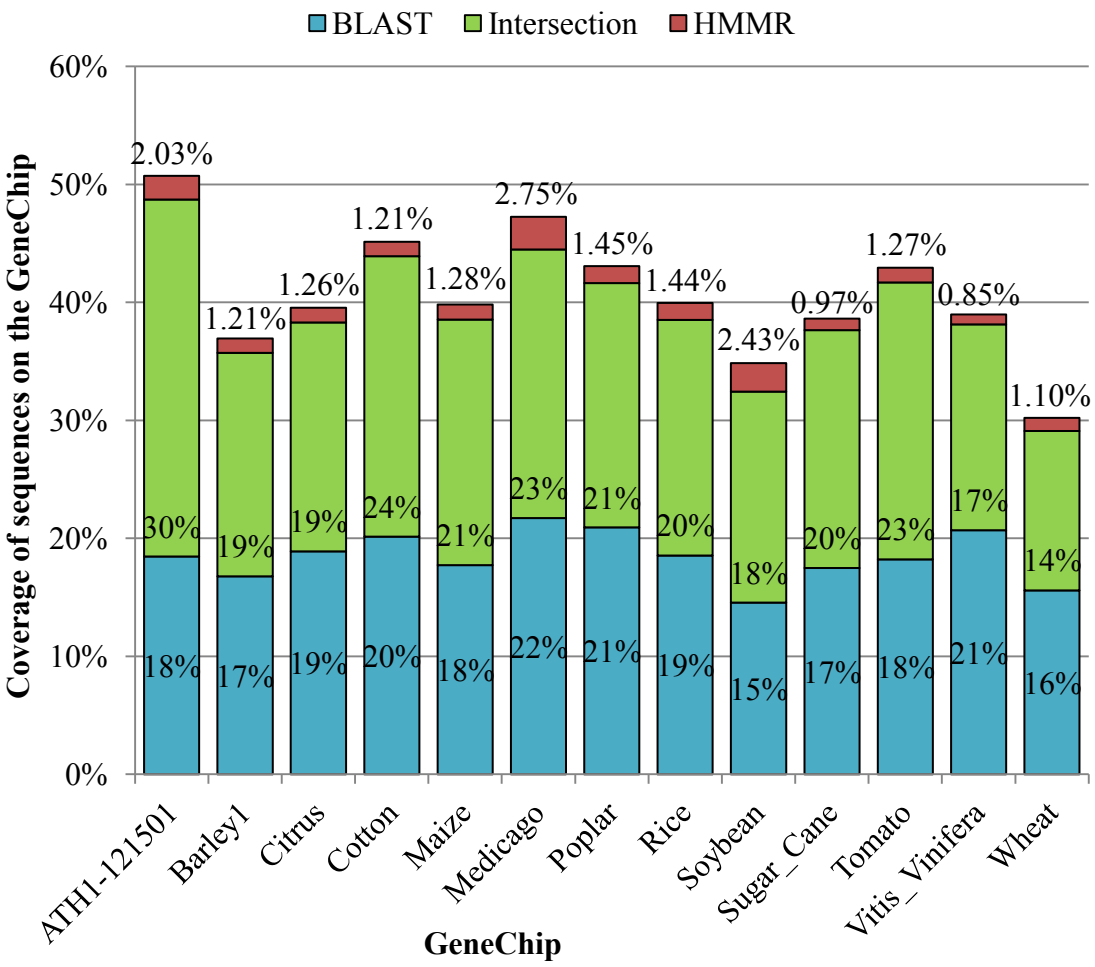


Figure 3.16: A stacked Venn diagram showing the coverage of EC annotation for sequences on each Affymetrix GeneChip. The sequences annotated exclusively by HMMR and BLAST are shown as well as the intersection containing sequences with annotation derived from both methods.

3.3.1(*b*) Evaluation of aggregated primary annotation

Aggregation of existing knowledge and inference of potentially new and unrealised knowledge are both important potential benefits of data integration. In order to quantify these benefits it is important to distinguish primary aggregated annotations, from knowledge that is the result of inference across mul-

tiple data sources. Aggregated primary annotation in CoPSA is the subset of functional annotations in public bioinformatics databases that can be extracted irrespective of data integration. The data aggregation benefit of a database is therefore defined as the singular contribution of that database to annotation, independent of knowledge present in other databases.

In order to separate these primary annotations from those that are the result of inference across multiple data sources, the systematic recording of provenance within the final annotation was required. This included information on the various elements, properties, and sources of data that the proposed CoPSA annotation relied upon. This provenance of information was stored as *evidence* attributes on graph nodes and edges in Ondex, and was the basis for determining the subset of primary annotations outlined in this section. This has been described in more detail in Chapter 2. Post filtering of annotations based on their provenance history, can therefore yield annotation that relies on knowledge from a single data source.

However, annotation provenance histories are permitted to include sequence data and Pfam HMM models sourced from other databases, as these are a minimum requirement for annotation. This is necessary as databases of primary annotation like GOA contain only links from protein accessions GO terms. Integration of sequence data, which is usually a trivial accession look-up, is therefore distinguished from the task of knowledge integration related to protein function.

Aggregating annotation to Gene Ontology

Figure 3.17 shows the unique, redundant and consistent contribution from GOA-*Arabidopsis thaliana*, GOA-*Orzya*, Gramene, and UniProtKB. These databases were aggregated and used by CoPSA to infer GO term annotations for consensus sequences. The raw number of primary aggregated GO annotations, for each database, in each chip, is represented by the total bar height. For

each CoPSA predicted GO term, for each sequence, in each database, the annotation is categorised according to three definitions denoting the presence or absence of equivalent annotation in other data sources. These are defined in Table 3.9. (1) The *unique to the method* category indicates no equivalent annotation exists in the other data sources. The presence of the same or similar annotation in other databases is categorised based on the ontology structure. (2) The category *identical to the other method*, indicates there is an identical annotation from another data source. (3) *Redundant to the other source*, indicates that a more specific annotation exists in another data source. (4) *Consistent with the other source*, identifies annotations that are consistent with more general terms present in other data sources. By implication the consistency of an annotation in one methodology in relation to another, also indicates the reverse relationship: the complementary redundancy of the shared annotation in the other methodology. Subdividing the aggregative contribution of databases according to these categories allows the contribution of a data source to be evaluated in relation to others. These categories are used through this section for comparing annotations.

It is apparent from Figure 3.17 that the most important data source in terms of total annotations and unique annotations is UniProtKB. However, the large proportion of unique content provided by the other databases demonstrates the utility of aggregating multiple primary data sources for maximising candidate annotations. Even the lowest content data source, GOA-*Oryza* (provided by Gramene in GOA format), contains 8767 unique annotations for rice.

GOA-*Oryza* is also provided by the Gramene database, and so the difference in the quantity of primary annotation was surprising. However, Ondex only has a parser for the version 29 Gramene flat-files (from Feb 2009), so the unique content in the GOA format (April 2010) is therefore due to the more up-to-date information in that format: an advantage of using a standard exchange-format. Gramene, however, cannot be considered having duplicated the information in

Table 3.9: Definition of categories used for comparative statistics of annotations derived from different methods or data sources. These categories apply both to GO or EC annotations, and parent and child relations are defined according to the topology of the graph or hierarchy respectively. For the GO hierarchy *is a* or *part of* indicates a parent child relationship. For EC a term is considered to be a child of the immediate more general level, *i.e.* 1.2.3.4 is a child of 1.2.3.-.

Annotation type	Definition
Identical to the other source	The term appears for the same sequence in the annotations of the other source
Consistent with the other source	A parent term (excluding the root) appears for the same sequence in the annotations of the other source
Redundant to the other source	A child term appears for the same sequence in the annotations of the other method
Unique to the source	No equivalent, parent, or child annotation can be found for the sequence in any of the annotations of the other source.

GOA-*Oryza* as it provides predicted GO annotations for all the grasses. The GOA-format export is, however, restricted to rice. For this reason, aggregating both sources for the annotation was a good compromise, and maximises candidate annotations. Similarly, UniProtKB contained many of the experimentally derived annotations from GOA as it sourced these annotations to create its own computationally predicted (IEA) annotations. However, given the cost of computing these predictions, it was inevitable they were not as up-to-date as the annotation in the directly downloaded GOA files. Aggregating UniProtKB with the GOA annotations enabled the annotation to benefit from the latest experimentally-derived annotations in addition to UniProtKB's extensive predicted annotations.

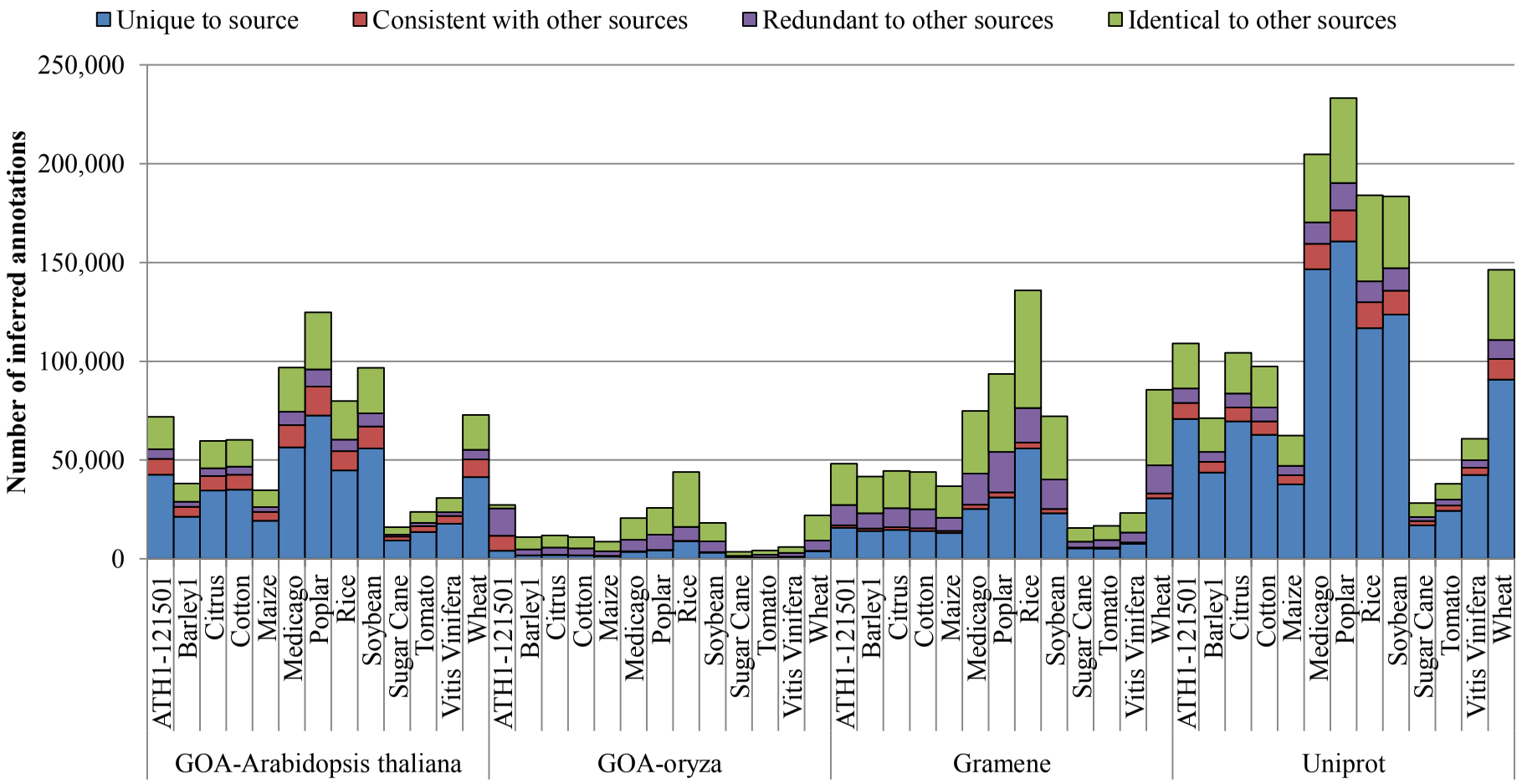


Figure 3.17: Provenance of inferred CoPSA sequence-GO annotations based on protein-GO aggregated annotations. An annotation is defined as a predicted GO term for a CoPSA sequence. A redundant annotation is one that occurs in another data source or has child terms in another data source for that sequence. A consistent annotation is one that contains a parent term (excluding the root terms) in another data source for that sequence. Unique annotations are terms that are not consistent or redundant to any other data source.

Figure 3.18 shows the proportion of sequences on each chip with at least one GO annotation (uniqueness with respect to other databases is not considered in this figure). This is reported for each of the primary database sources for the scenario where they are used independent of other sources for GeneChip annotation. Figure 3.18 is reported in addition to Figure 3.17 because it shows the impact on GeneChip coverage, not simply quantity of annotations. Differences in GeneChip coverage between primary annotation sources are indicative of the breadth of gene families annotated in that source.

The interpretation of each databases relative contribution changes when calculated in terms of the coverage of annotated sequences on each chip. Figure 3.18 shows that the dominant quantities of annotation from the UniProt database seen in Figure 3.17 is not reflected in a greater diversity of annotatable sequences. GOA- *Arabidopsis* and Gramene appeared to contain much less raw annotation Figure 3.17; however they have a surprisingly high GeneChip coverage. Likewise, the relatively smaller number of GOA-*Oryza* annotations translate into a much larger coverage of the chip. There are two factors which affect this difference between the number of raw annotations (Figure 3.17) and sequence coverage (Figure 3.18). Firstly, the primary data sources vary in the amount of annotations attached to each protein, with computational methods producing a greater quantity of annotation. Secondly, the transference of function via sequence similarity means that there was less benefit (in terms of GeneChip coverage) for primary annotations that were concentrated around well-known protein families and exhaustively annotated homologous proteins, compared to disparate annotations that covered novel or poorly annotated families. The contribution of a source of primary annotation to increasing coverage is therefore a function of both the quantity and diversity of original proteins annotated.

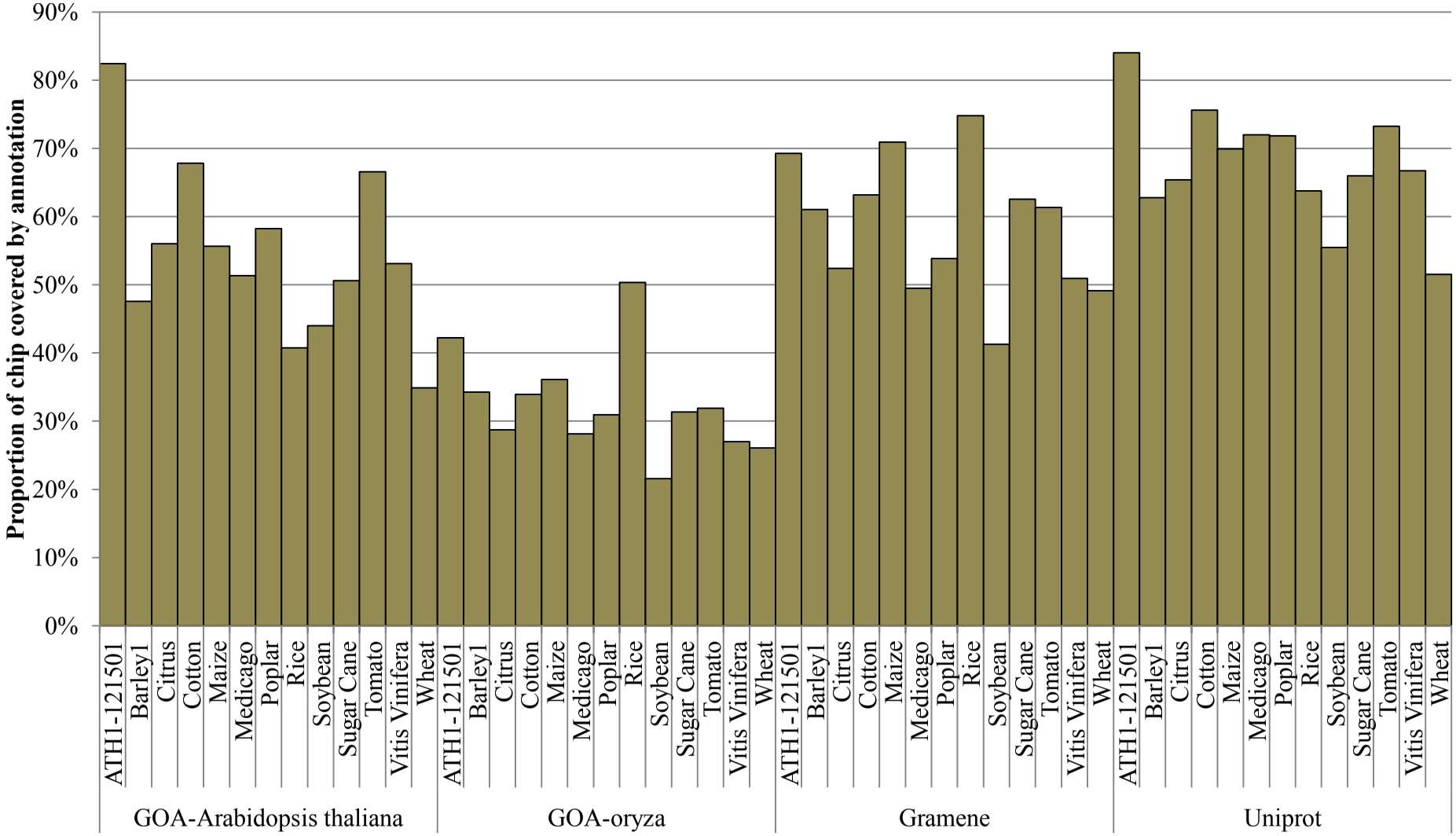


Figure 3.18: Proportion of sequences on each chip covered by GO annotation from primary data sources.

The analysis shown in Figure 3.17 compared the quantitative contribution of different databases to primary annotation. However, an evaluation of the original provenance of GO annotations within a primary source can give valuable information on the quality of the contribution of a data-source. Experimental validation of gene or protein function is of far greater value than computational predictions. The GO provides evidence codes that document the provenance of annotation and these were included as properties on the Ondex relations that store annotations from proteins to GO. The GO evidence codes and the categories used to type them are listed in (Table 3.10). Figure 3.19 shows the results of an evaluation of the provenance of the primary annotations by quantifying the number of proteins used for inference from each of the GO evidence code categories. It is evident that the majority of the experimental evidence was contributed by UniProtKB and GOA-*Arabidopsis* databases, whereas Gramene and GOA-*Oryza* contributed mainly computational predictions. This highlights the value of aggregating multiple data sources in order to incorporate as much experimental evidence as possible. GOA- *Arabidopsis* also contained Non-traceable Author Statement (NAS) and Traceable Author Statement (TAS), neither of which was present in UniProtKB derived CoPSA annotations. Conversely, UniProtKB contains additional experimental annotations not present in GOA-*Arabidopsis*. Gramene also contains a number of non-*Oryza* GO annotations of unstated provenance, which reflects the original table structure of the Gramene database where evidence is optional.

For GO annotations inferred from Pfam domain matches, no data aggregation occurred because pfam2go, interpro2go and prosite2go domain annotations are the only source of primary annotation used by to link domains directly with GO terms in CoPSA, and were all derived from the InterPro mapping by Hunter *et al.* (2009).

Table 3.10: Evidence codes for the provenance of GO annotations as defined by the Gene Ontology Consortium.(The Gene Ontology Consortium, 2011a)

Code	Type of evidence	Type category
NAS	Non-traceable Author Statement	Author Statement
TAS	Traceable Author Statement	
IEA	Inferred from Electronic Annotation	Automatically-assigned
IGC	Inferred from Genomic Context	Computational Analysis
ISA	Inferred from Sequence Alignment	
ISM	Inferred from Sequence Model	
ISO	Inferred from Sequence Orthology	
ISS	Inferred from Sequence or Structural Similarity	
RCA	Inferred from Reviewed Computational Analysis	
IC	Inferred by Curator	Curator Statement
ND	No biological Data available	
EXP	Inferred from Experiment	Experimental Evidence
IDA	Inferred from Direct Assay	
IEP	Inferred from Expression Pattern	
IGI	Inferred from Genetic Interaction	
IMP	Inferred from Mutant Phenotype	
IPI	Inferred from Physical Interaction	

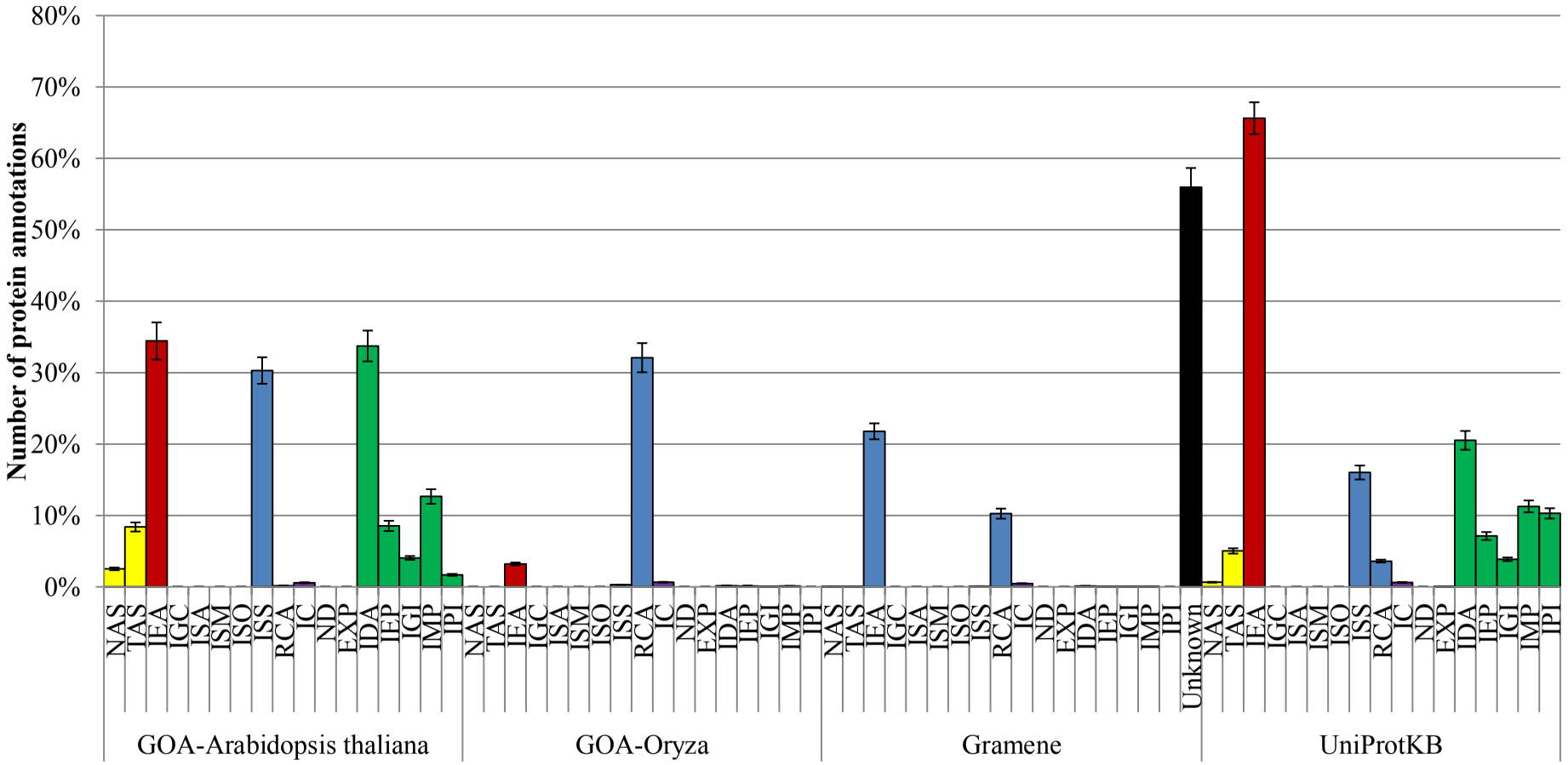


Figure 3.19: The mean proportion of chips with one or more annotations of the given evidence code for all chips analysed in each of the standard Gene Ontology evidence codes categories. Definitions of the GO evidence codes can be found in Table 3.10.

Aggregating annotation to EC

A separate evaluation of the role of aggregation in providing EC annotation is provided in addition to the previous assessment of GO. For the reasons previously stated within this section, EC annotation may differ substantially in source and content from GO. This can be explained by the two previously described factors. (1) The original sources of EC primary database evidence may not overlap with GO annotation. (2) EC represents a functional subset of the GO ontology, which for a given databases may be more or less amenable to annotation.

Figure 3.20 shows the contribution of the various primary data sources to CoPSA EC annotation using protein similarity detected by BLAST. Annotation of sequences with EC terms, based on primary annotation in databases, yielded far fewer annotations than CoPSA annotation of GO (Figure 3.17). This was to be expected as EC is more specialised than the Gene Ontology and only applies to a subset of proteins with catalytic function.

The estimates of redundant annotation were based on the EC hierarchy, where an EC term in a primary data source was classed as redundant if another data source annotated the same sequence to a more specific level and shared the same parent category. For example, the EC term "1.14.13.- " is redundant to the more specific term "1.14.13.93". Redundant annotation made up a minority of annotation but was more prevalent in AraCyc and KEGG databases.

Consistent annotation, as with the definition for GO, was based on the overlap of parent categories. An annotation is consensual to another if it shares all of the same parent categories. For example, "1.14.13.93" is consistent with "1.14.13.6" but not with "1.14.16.1". Most of the annotations fell into this consensual category, which highlights a key difference compared to GO. In GO, primary sources are composed of multiple annotations in the GO tree, whereas for EC usually only a single term is assigned to a sequence. On average in the primary annotations shown, 1.17 EC terms (standard error = 0.19) were annot-

ated per sequence, whereas for GO there was 2.72 terms per sequence (standard error = 0.86).

Unique annotations were defined as annotations that have no other consensual annotation in any of the other data sources. The only contributor of completely unique EC-term annotation was UniProt. This is almost certainly due to Uniprot providing curated mappings between GO and EC terms when populating their EC annotation.

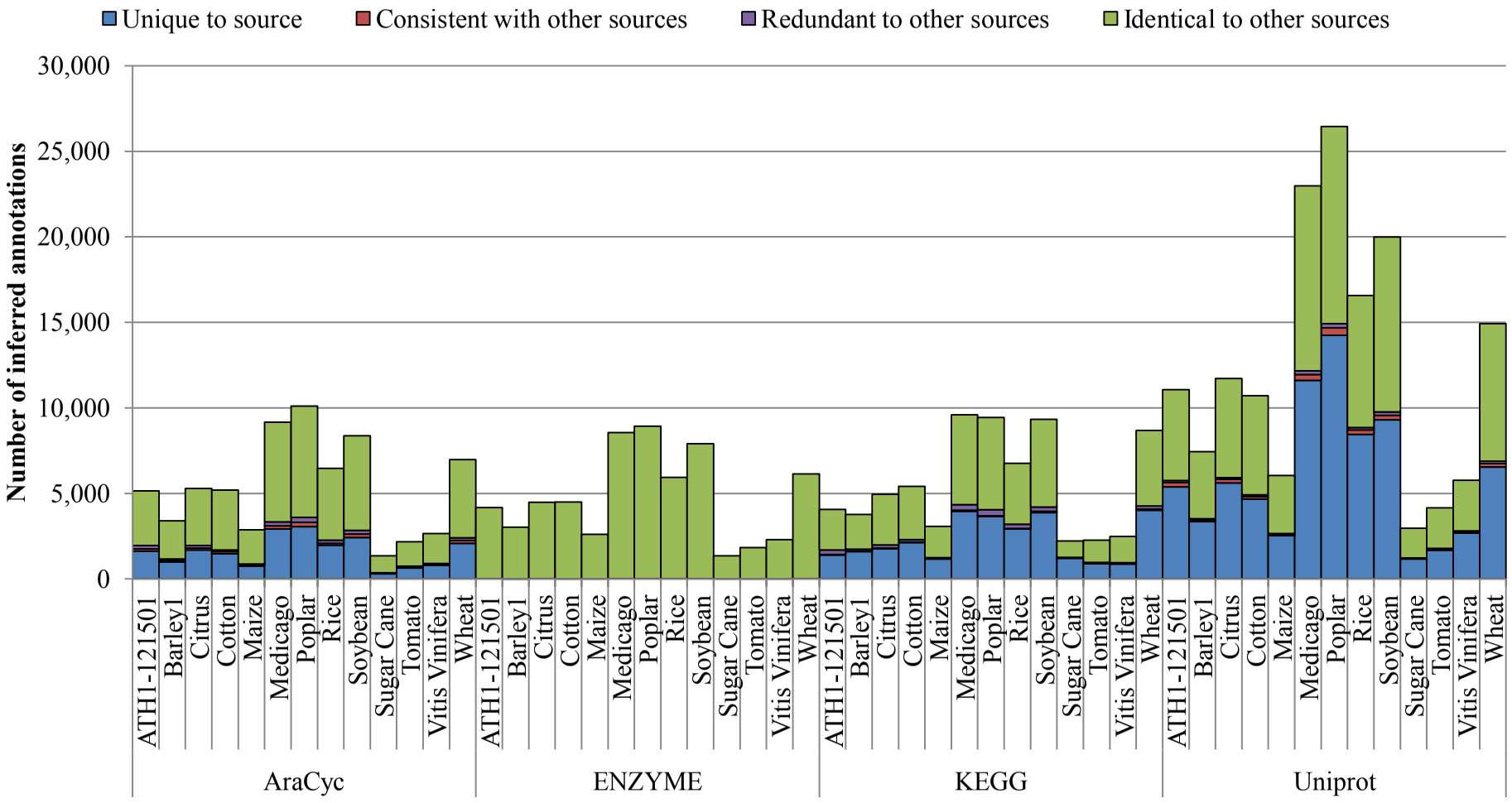


Figure 3.20: EC annotations based on existing primary annotation of proteins in data sources, inferred using BLAST. An annotation is a predicted EC term for a CoPSA sequence. A redundant annotation is one that occurs in another data source or has a child terms in another data source for that sequence. A consistent annotation is one that contains a parent terms (excluding the root terms) in another data source for that sequence. Unique annotations are terms that are not consistent or redundant to any other data source.

Figure 3.21 shows the proportion of sequences on each GeneChip with at least one EC term that was derived through aggregation of annotation from data-sources. These statistics are provided for each data source that contained direct protein to EC term annotation. Figure 3.21 shows that in terms of the usefulness of the primary annotation in providing coverage of sequences represented on the chips, UniProt was the most effective. The advantage of UniProt in providing protein to EC term annotation is more pronounced for GeneChip coverage than for the annotation counts presented in Figure 3.17. This indicates that UniProt not only provides a greater quantity of annotations, but also provides EC terms for a greater breadth of protein sequence types. Comparing the quantity of annotation in Figure 3.17, with the coverage in Figure 3.21, also reveals that the ENZYME database provides a similar improvement in the quantity of sequences if can annotation, relative to the total quantity of protein to EC term annotations (*i.e.* annotations are more widely spread across sequences).

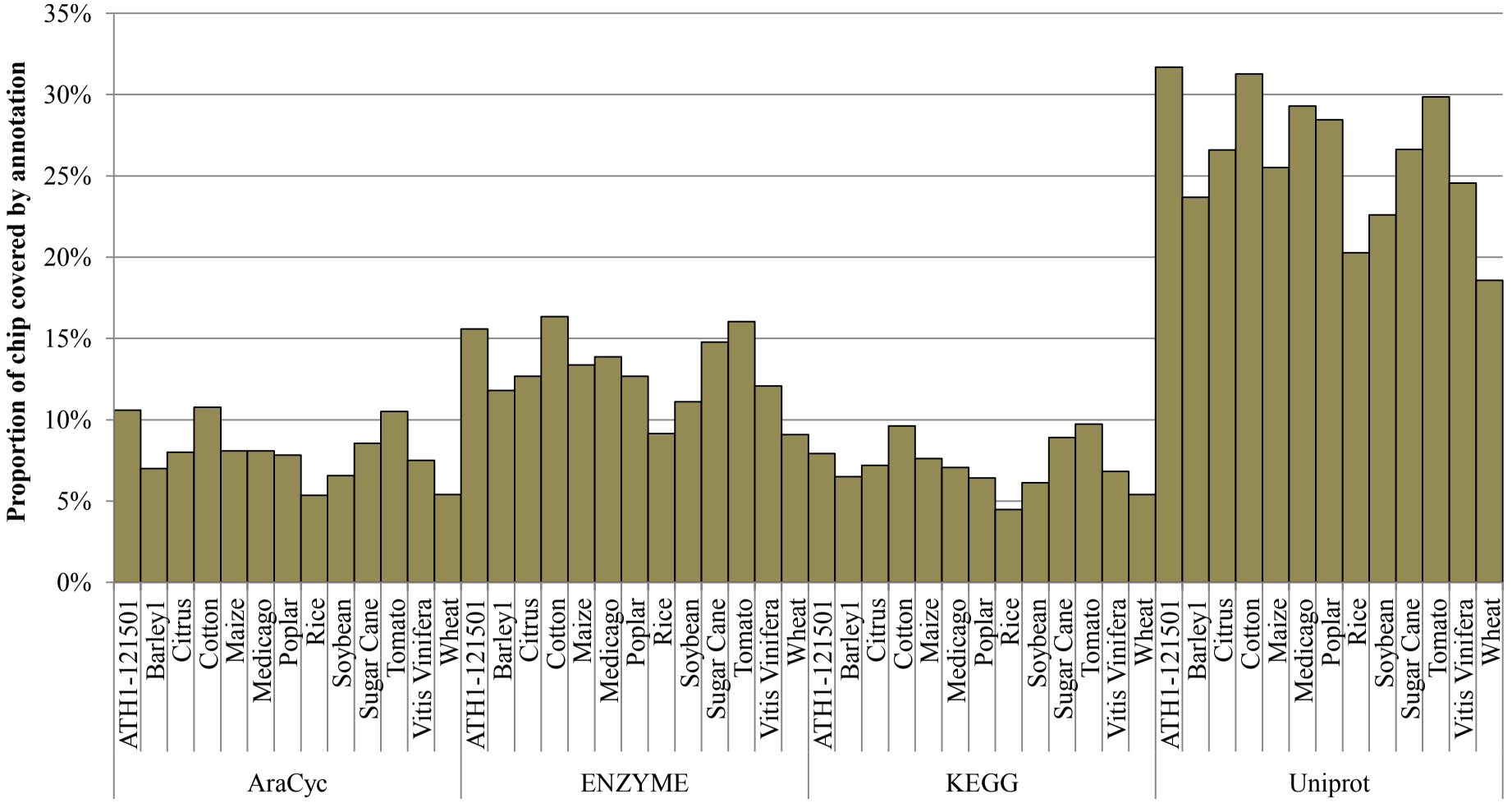


Figure 3.21: The proportion of sequences on each chip covered by EC annotation from primary data sources.

The quantity and coverage of annotation provided by a data source is not the only measure of usefulness. The specificity of EC annotation indicated by the level in the hierarchy is also an important measure of quality. Figure 3.22 shows the mean EC level of annotations on each chip provided by each data source. The specificity of annotation is species independent in ENZYME and largely uniform for UniProt. However, for the KEGG and AraCyc pathway databases, there is large variability, which follows a similar trend in both data sources. There does not appear to be an obvious explanation for this that is related to the evolutionary distance from *Arabidopsis*. The highest mean EC level is provided by annotations from the ENZYME databases. KEGG and UniProt on average contribute EC terms of a similar specificity. AraCyc aggregated EC annotations have the lowest mean specificity in CoPSA.

The higher specificity of KEGG compared to AraCyc is surprising, considering that AraCyc contains a greater proportion of curated terms. In order to explain this difference in mean specificity Figure 3.23 shows the distribution of the mean number of annotations for each chip, at each level in the EC hierarchy. It is apparent from this that AraCyc has a larger number of level 1 and 3 annotations than KEGG, but has less level 2 and 4 annotation. The majority of the increase in specificity of KEGG relative to AraCyc is as a result of more level 4 annotations. A possible explanation for this is the different species specific metabolic pathways found in AraCyc and KEGG. AraCyc contain only EC annotations for *Arabidopsis* enzymes, whereas the data imported from KEGG included EC predictions from all *Viridiplantae*. These predictions appear to have negligible impact on the quantity and coverage of EC annotations, but may contribute to an increased specificity in CoPSA predictions.

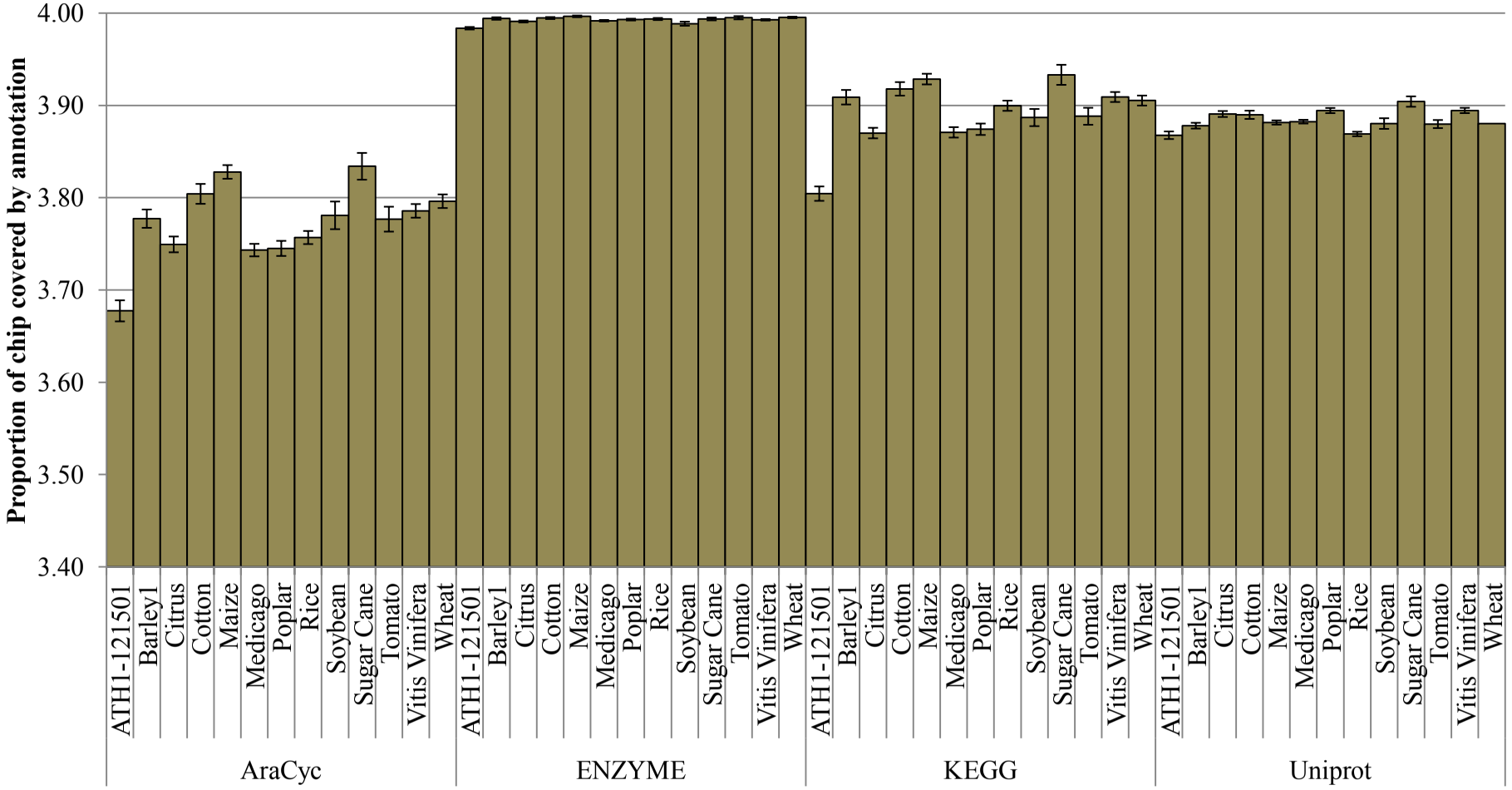


Figure 3.22: The mean EC term annotation level by database and chip. The error bars are standard error of the mean.

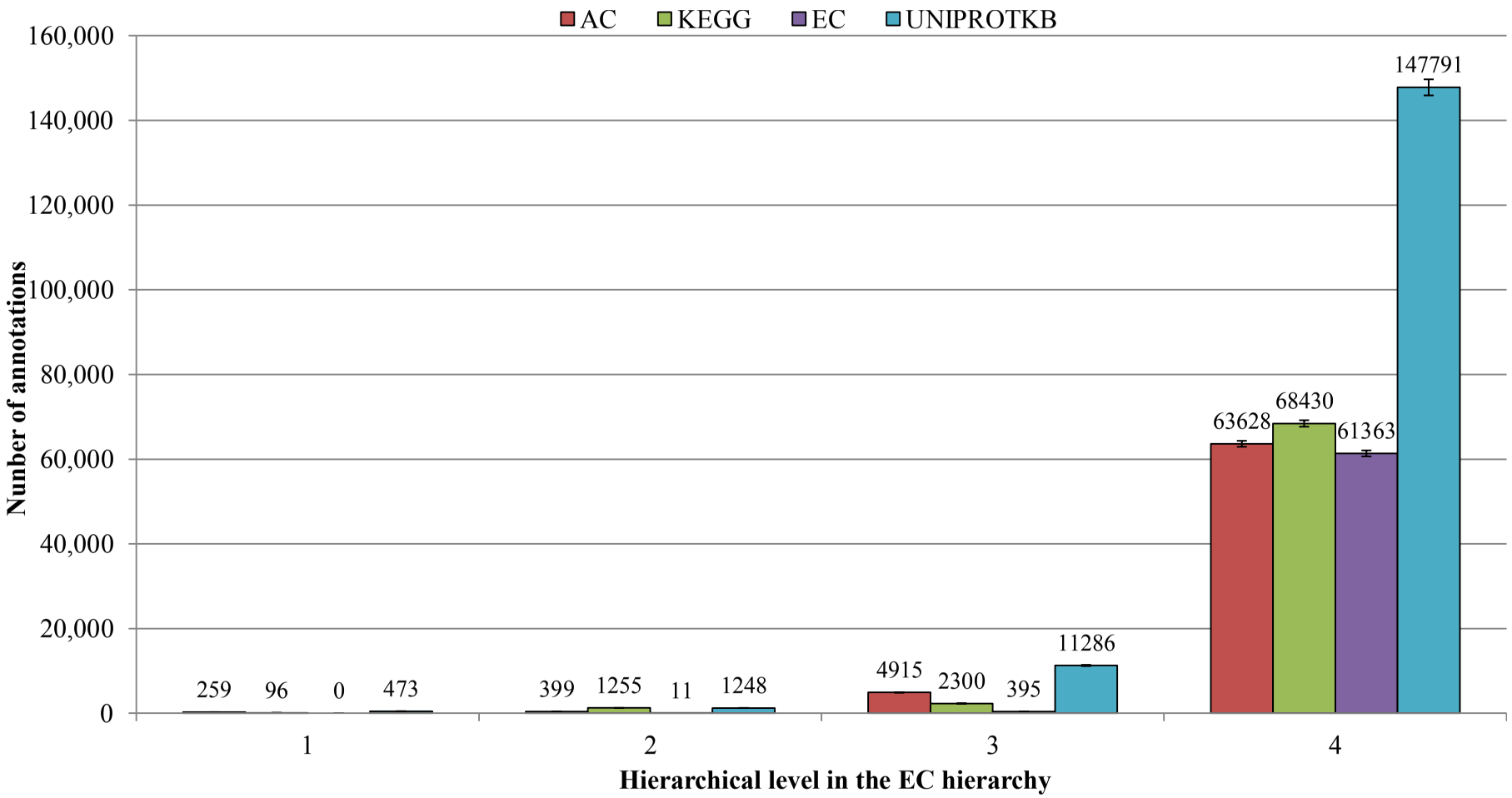


Figure 3.23: The mean distribution of levels (1, 2, 3 and 4) in the EC hierarchy across all 13 chips annotated by CoPSA. Error bars are standard error of the mean.

3.3.2 Evaluation of annotations inferred from data integration

Previously, CoPSA annotations that were derived from provenance based solely on a single data source have been discussed in Section 3.3.1(*b*). CoPSA also uses inference across multiple data-sources to create new candidate annotations that may not have been present in any of the original primary data sources. This subset of functional predictions, with an inference based provenance, is defined as: annotations created by CoPSA that are based on information distributed across more than one data-source. For example, where the ENZYME database provides an EC term annotation for a given protein, and the EC2GO data-source provides an equivalent translation of EC to a GO term. This GO term annotation is therefore a product of inference across these two data sources, and may not exist in an explicit form in any public database. In the same manor described in the introduction to Section 3.3.1(*b*), the provenance of information was stored as *evidence* attributes on graph nodes and edges in Ondex, and was the basis for determining the two provenance subsets: single-database aggregation and multi-database inference. This methodology of tracking provenance during annotation extraction has been described in more detail in Chapter 2. Post filtering of annotations based on their provenance history, can therefore yield annotation that relies on knowledge from a single data source (single-database aggregation) in addition to annotation reliant on multiple data sources (multi-database inference).

Within this Section, the relative contribution of inference over multiple data-sources is compared to that of aggregating existing annotations from individual data-sources. The contribution to the quantity of annotation and coverage of sequences on the chip is evaluated for both annotation using protein sequences (identified from BLAST) and Pfam domains (identified through HMMR). This comparison is made first for GO annotations, and then repeated for EC an-

notations. A comparison of provenance (multiple data-source inference verses annotation-aggregation) is therefore made against two annotations subsets (from protein sequences and Pfam domains) for both annotation types (GO and EC). CoPSA annotations identified using protein-sequence similarity are considered separately from those reliant on identified Pfam domains, because these two methodologies rely on different data from multiple sources (*e.g.* the UniProt and InterPro databases respectively). They therefore may differ in their reliance on the two provenance types in question. It is equally true that GO and EC annotations in CoPSA rely on different data (represented as different data types) in multiple sources (*e.g.* the GOA and ENZYME databases respectively). As in the previous section, in order to compare differences in the annotations derived from single data-source aggregation to that from multiple data-source inference, four annotation comparison categories were defined in Section 3.3.1(*b*) and are summarised in Table 3.9.

The following four sections address: (1) predicted GO annotations through protein similarity, (2) predicted GO annotations through domain identification, (3) predicted EC annotations through protein similarity, and (4) predicted EC annotations through domain identification. They have been structured in the same way, and address the contribution of single-database aggregation versus multi-database inference first with regard to the quantity of annotations provided, and then through the coverage of sequences on each GeneChip.

Annotation of GO terms based on protein sequence similarity using BLAST

This section addresses the subset of GO annotations derived from protein sequence similarity using BLAST. For these annotations, Figure 3.24 shows the relative contribution of multiple-database inference in comparison to using only aggregation of existing GO annotations of single data-source provenance. The results reported in this section are computed from the provenance recorded

in each annotation produced by CoPSA. The methodology for recording this provenance has been reported in Chapter 2. Subsets of the final annotation were extracted based on their reliance on a single or multiple dataset in the provenance history. The reporting of aggregation based annotations are equivalent to running CoPSA on each of the data-sources containing GO annotation in isolation, and combining the results afterwards. Annotations with integration provenance are those annotations that depend on multiple sources; this is not equivalent to applying CoPSA to all data-sources synchronously and subtracting the union of aggregated annotations. This is because a given annotation may be inferred from multiple sources, which are redundant to annotation already reported in a single data source. For example: using ENZYME in combination to EC2GO may yield annotations that are pre-existing in UniProt. For this reason, Figure 3.24 reports the proportion of annotations with aggregated and inference based provenance that are unique to each respective method, or which are more or less specific or identical to that which was reported in the complimentary method. The exact values for each of these comparison categories in each of the charted bars in Figure 3.24 are provided in Table 3.11. These comparative annotation categories have been previously described in detail in Section 3.3.1(*b*), and are summarised in Table 3.9.

It can be seen from Figure 3.24 that fewer novel annotations were derived from inference compared to aggregation alone, however an average of 1197 (standard error = 190) unique annotations per chip were found that were previously not explicitly annotated in any of the datasets. An additional average of 262 (standard error = 42) inference based annotations per chip improved on the specificity of aggregation based annotation (consistent with the other method). On average 10485 (standard error = 1742) inferred annotations per chip were identical to direct annotations. There were also an average of 798 (standard error = 126) inference based annotations that produced less specific annotation than were present for direct annotation. These consistent and identical annota-

tions predicted through inference methods lend support to the validity of these inference rules as well as adding evidence that corroborates the direct annotation. While the contribution of inference methods to the total CoPSA annotation was small in terms of quantity, their contribution is not negligible.

As previously stated there were relatively few inference based annotations (average per chip = 262) that improved on the specificity of GO term annotations compared to aggregating direct annotation. Inference based annotations that increased the specificity of a term (consistent with other method) accounted for only 2.94% (standard error = 0.71%) of the total inferred annotations. Reciprocally, aggregation based annotations did not improve much on the specificity provided by inference based methods, with only 0.67% (standard error = 0.15%) belonging to the consistent comparison group. This indicates that for GO annotations, inference provides identical or novel annotations, but rarely provides a more general or specific GO term for annotation.

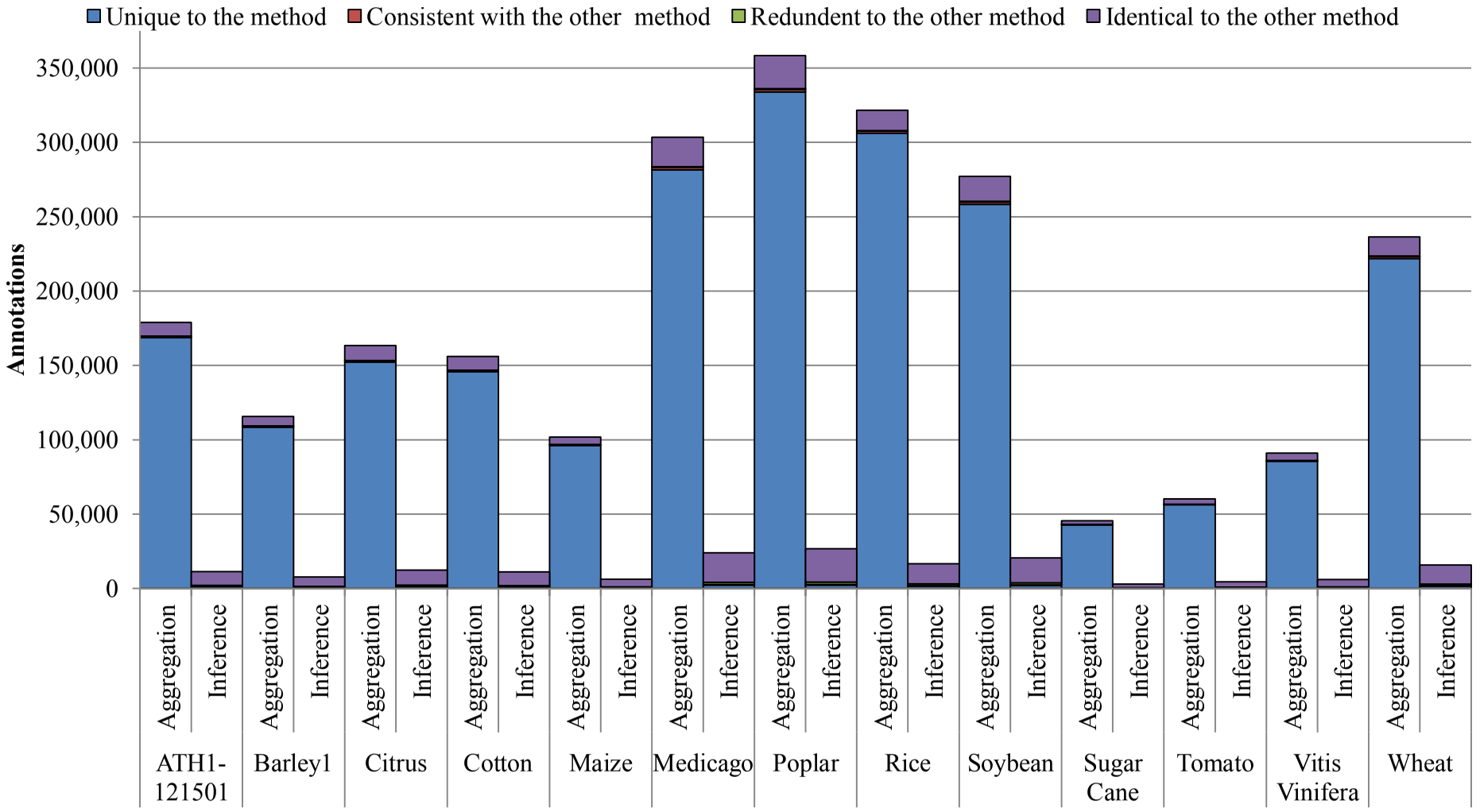


Figure 3.24: A comparison of the number of annotations derived from data aggregation verses inference over an integrated data set for GO annotation via BLAST. The categories of annotation in the legend are defined in Table 3.9. The annotation counts shown in this chart are provided in Table 3.11.

Table 3.11: Counts of GO annotations for each GeneChip, derived from aggregation and inference based provenance.

GeneChip	Method	Unique to the method	Consistent with the other method	Redundant to the other method	Identical to the other method	Total
ATH1-121501	Aggregation	168639	855	225	9160	**178879**
	Inference	1175	221	744	9160	**11300**
Barley1	Aggregation	108527	606	171	6370	**115674**
	Inference	747	168	497	6370	**7782**
Citrus	Aggregation	152244	816	259	10082	**163401**
	Inference	1294	264	710	10082	**12350**
Cotton	Aggregation	145827	785	215	9184	**156011**
	Inference	1045	219	698	9184	**11146**
Maize	Aggregation	96143	520	113	5041	**101817**
	Inference	566	107	437	5041	**6151**
Medicago	Aggregation	281613	1651	434	19795	**303493**
	Inference	2305	444	1400	19795	**23944**
Poplar	Aggregation	333895	1773	479	22295	**358442**
	Inference	2284	490	1565	22295	**26634**
Rice	Aggregation	306187	1389	349	13649	**321574**
	Inference	1495	348	1175	13649	**16667**
Soybean	Aggregation	258380	1567	481	16791	**277219**
	Inference	1891	485	1339	16791	**20506**
Sugar Cane	Aggregation	42777	243	71	2451	**45542**
	Inference	233	63	195	2451	**2942**
Tomato	Aggregation	56242	321	102	3629	**60294**
	Inference	433	101	277	3629	**4440**
Vitis Vinifera	Aggregation	85511	409	118	4994	**91032**
	Inference	564	121	357	4994	**6036**
Wheat	Aggregation	221968	1212	366	12868	**236414**
	Inference	1540	379	990	12868	**15777**

The quantity of raw annotations is important in studying the comparative merits of annotations of aggregation and integration provenance; however it does not indicate the distribution of annotation across sequences, which directly affects coverage and the efficacy of the annotation for microarray interpretation. Figure 3.25 shows the proportion of sequences on the chip that were annotated with GO, using protein sequence similarity and either single-database aggregation or multi-databases inference provenance exclusively. These coverage statistics are divided into sequences that received novel information not present in any of the direct annotation of the databases, and sequences that received annotation already present in direct annotation. In relation to the comparative definitions provided in Table 3.9, the category "additional information provided relative to other method" amalgamates annotations that are unique to the method, and consistent with the complementary method. The category "Only redundant information relative to the other method", amalgamates annotations redundant and identical to the other method. The four categories have been simplified into two broader categories in Figure 3.25. These highlight the impact of the aggregation and inference methods on overall annotation, rather than providing a more detailed qualitative comparison.

A comparison based on the number of sequences on the chip annotated (coverage) reveals that the impact of aggregating direct annotations, compared to inference over datasets, was greater. On average, inferred annotation covered in total 25.45% (standard error = 1.11%) of the sequences on a chip. An average of 3.73% (standard error = 1.03%) of sequences on the chip received new or more specific annotation from inference, that could not be predicted by using any single incorporated database. This indicates that the average of 1212 (standard error = 203) novel and 262 (standard error = 42) consistent annotation per chip previously reported, are widely dispersed across multiple sequences. Reciprocally, single-database aggregation of GO annotations contributed new or more specific annotation to on average 76.12% (standard error = 21.11%) the

GeneChip. This confirms the greater efficacy of single-database aggregation in providing a greater quantity and specificity of GO annotations for a GeneChip through protein-sequence similarity, compared to multi-database inference. However, the benefits of multi-database inference in providing new and more specific annotation is more variable, across databases than multi-database inference, with the *Arabidopsis* ATH1 (93.74%), and Soybean (59.86%) GeneChips receiving the most and least benefit respectively.

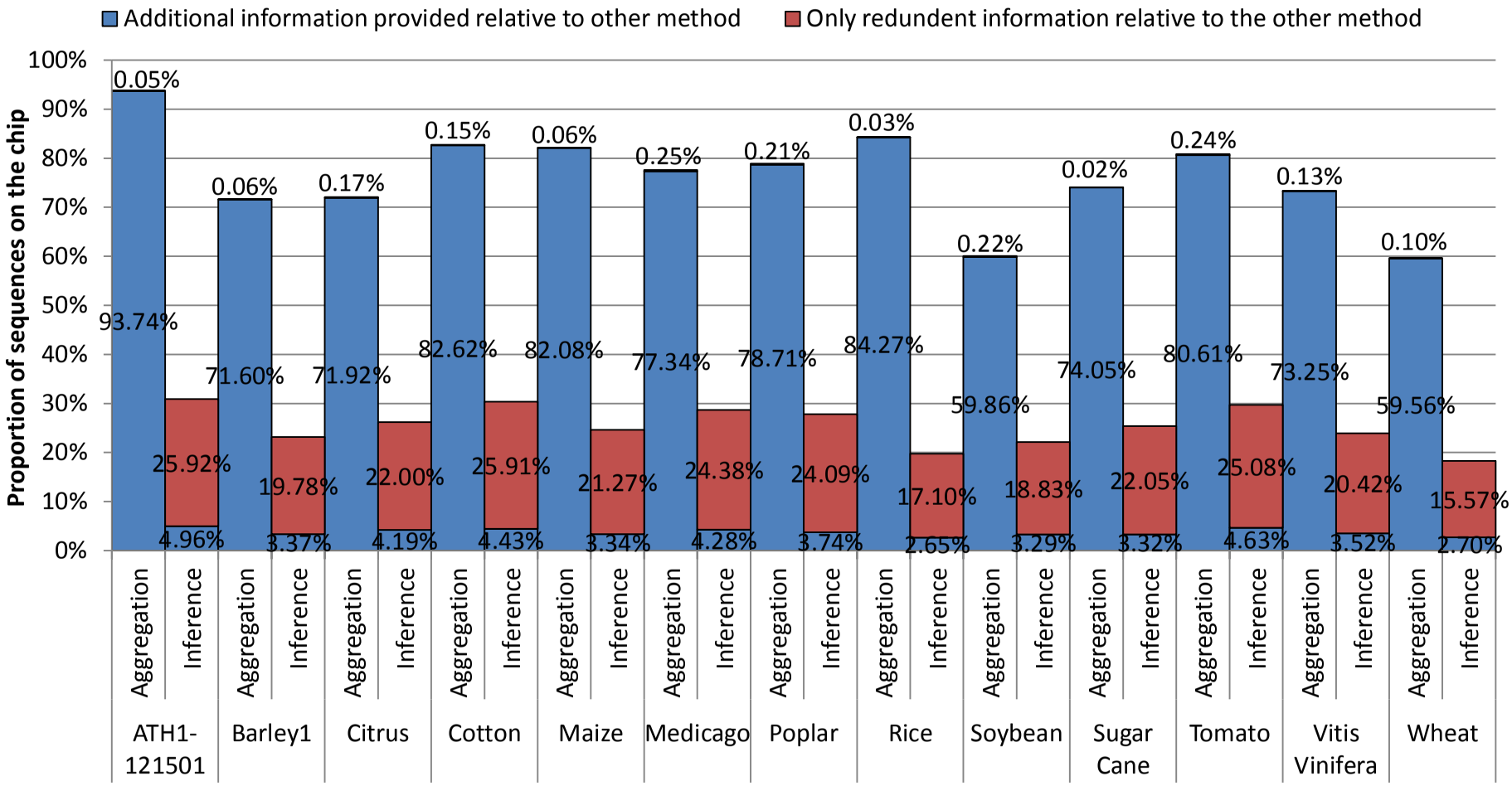


Figure 3.25: The proportion of sequence on the chip that can be annotated with GO terms using the aggregation of primary annotation and inference over integrated data using BLAST.

Annotation of GO terms based on detected protein domains using HMMR

The analysis described in the previous Section, dissected the data provenance of annotation derived from protein similarities (identified from BLAST). This analysis is repeated within this section for annotations derived from Pfam domains, which were identified using HMMR. It is replicated for both methods (using BLAST and HMMR), because domain and protein centric annotation rely on different data and sources. The CoPSA annotations created using domains may be more or less reliant on primary-aggregation or inference provenances. The statistics in this section were generated based on a categorisation of each annotation from the HMMR methodology into single-database aggregation or multi-database inference based provenance, in the same manor described in the previous section.

For the annotations derived from identified Pfam families (using HMMR), Figure 3.26 shows the relative contribution of annotations with multiple-database inference provenance in comparison to using only aggregation of existing annotations with single data source provenance. On average across GeneChips GO annotations derived from protein sequence similarity, described in the previous section, yielded 14.5 times (standard error = 0.40) more annotation from primary-aggregation than inference provenance. By comparison, CoPSA GO annotation using protein domains yielded only 4.3 times (standard error = 0.05) as many annotation derived from aggregation compared to inference. This indicates that inference across data sources, made a larger contribution when extracting GO annotations from Pfam domains (using HMMR) than from protein sequence similarities (using BLAST). The increased efficacy of inference is explained by EC annotations of Pfam domains that do not have an equivalent GO functional annotation in any integrated database.

The redundant annotation category represents GO terms that are predicted by an annotation subset, but that exist as more specific predictions in the other subset. Redundant GO term annotations across all GeneChips accounted for

on average 4.06% (standard error = 0.08%) and 28.33% (standard error = 0.80%) of single-database aggregation and multiple-database inference annotations respectively. The consistent annotation category represents GO terms which were more specific than the other category, but consistent with a parent (non-root) terms from the other annotation set. For CoPSA annotations aggregated from domains, on average per GeneChip 6.83% (standard error = 0.20%) are consistent with annotations by multiple-database inference methods. Reciprocally, for multiple-database inference, on average per GeneChip 16.08% (standard error = 0.27%) were consistent with aggregated annotations. Compared to the statistics for reciprocal annotations in protein similarity derived annotation (BLAST) (see Section 3.4.3.1), presented in the previous sections, domain based annotation benefits considerably more in specificity by multiple-database inference (2.94% versus 16.08%).

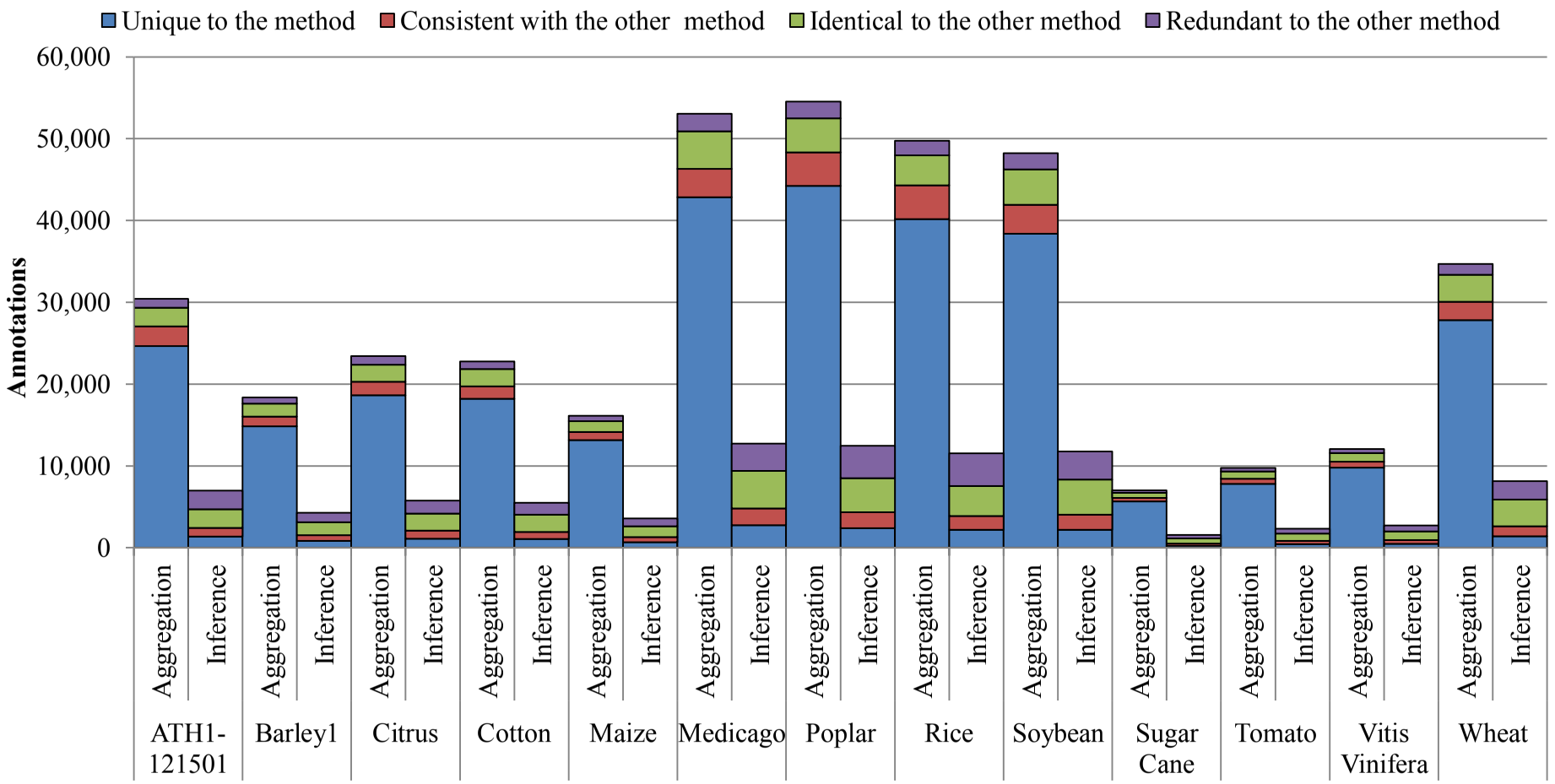


Figure 3.26: A comparison of the number of annotations derived from data aggregation verses inference over an integrated data set for GO annotation via Pfam domains (using HMMR). The categories of annotation in the legend are defined in Table 3.8. Annotations are filtered based on MICA terms per domain, which is described in Section 3.2.1(*g*).

As with the analysis procedure for annotations based on protein similarity, the distribution of annotation provenance across GeneChip sequences is also considered, which directly affects coverage and the efficacy of the annotation for microarray interpretation. Figure 3.27 shows the proportion of sequences on the chip that were annotated with GO using domains (identified using HMMR) and either single database aggregation or multi-database inference provenance exclusively. As in the previous section with Figure 3.25, these coverage statistics are subdivided into two broad comparative categories, contrasting new information with redundant information. These highlight the impact of the aggregation and inference methods on overall annotation, rather than providing a more detailed qualitative comparison.

In terms of *additional* coverage of sequences on the chip, shown in Figure 3.27, there were on average a greater proportion of of GeneChip sequences annotated, with GO, using multi-database inference, based on HMMR detected domains (based on 7.10% less inference than aggregation based annotations, standard error = 0.43%) than were derived from multi-database inference using protein sequence similarity (using BLAST) (based on 10.2% less inference than aggregation based annotation, standard error = 0.18%). This is despite GO annotation, using protein sequence similarity, in this CoPSA pipeline being a richer source of information than HMMR identified domain (see Section 3.3.1(*a*)).

Compared to protein similarity (BLAST) based annotation, described previously in this section, the role of multi-database inference using HMMR identified domains was far greater in increasing the proportion of GeneChip sequences with new or more specific GO annotations. For domain based annotation, on the average GeneChip, the proportion of sequences with additional information was 7.10% (standard error = 1.97%) and 35.05% (standard error = 9.72%) for single-database aggregation and multi-database inference, respectively. This is compared to 76.12% (standard error = 21.11%) and 25.45% (standard error = 9.72%) of sequences benefiting from single-database aggregation

and multi-database inference respectively in protein-similarity (using BLAST) based annotation.

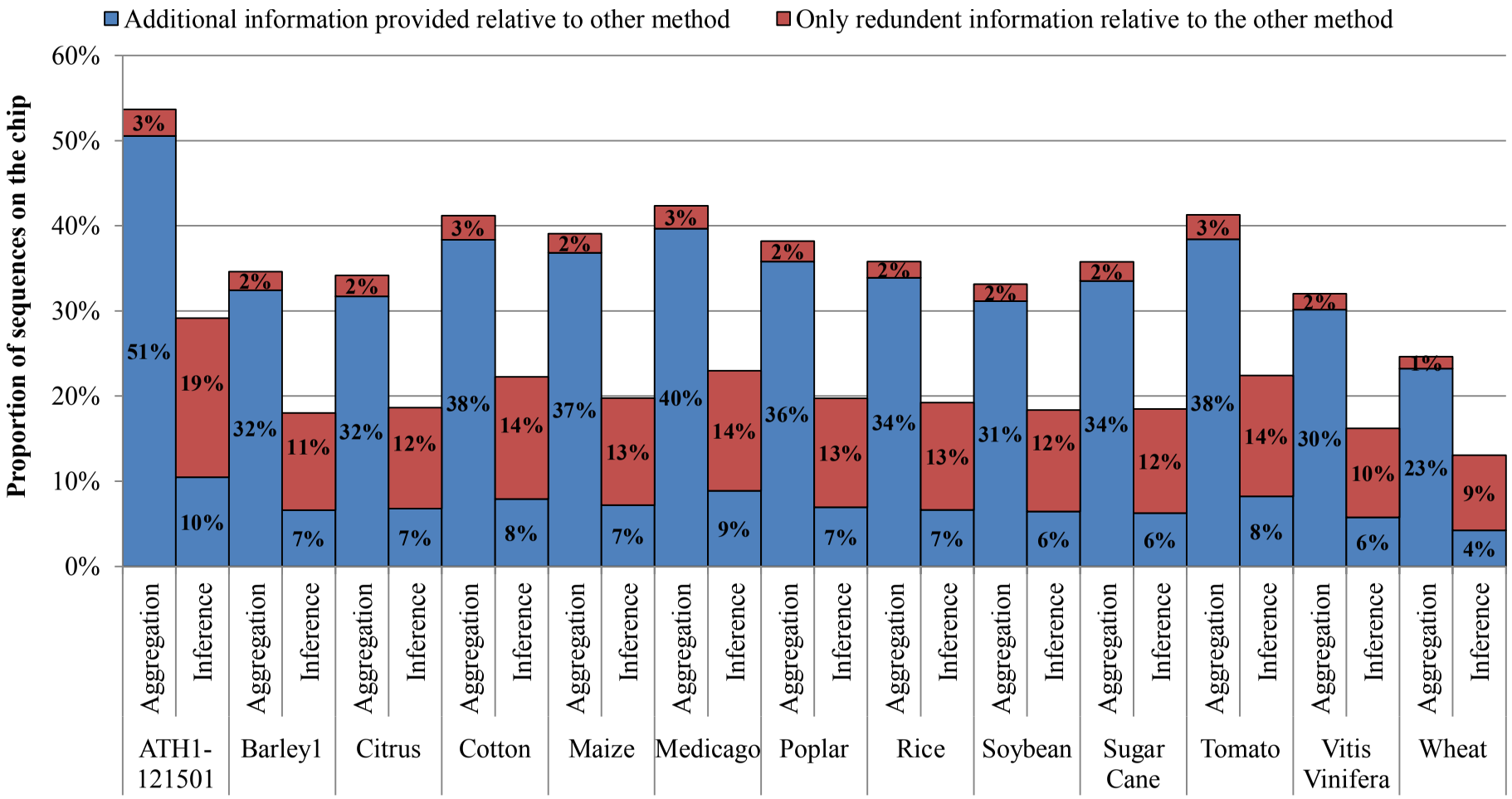


Figure 3.27: The proportion of sequence on the chip that can be annotated with GO terms using the aggregation of primary annotation and inference over integrated data using HMMR. GO terms are generalized according to domain using MICA as described in Section 3.2.1(*g*).

Annotation of EC terms based on protein sequence similarity using BLAST

Previously in this section, the relative contribution of single-database aggregation and multi-database inference to GO annotation was evaluated, using protein sequence similarity (BLAST) and Pfam domain detection (HMMR). Within this section the contribution of single-database aggregation and multi-database inference to EC annotation, using protein sequence similarity (BLAST) is evaluated. As previously stated in the introduction to Section 3.3.2, the evaluation is repeated for EC as well as GO, as the data types and data sources differ for these two annotation types. Therefore, it is expected that the contribution of single-database aggregation and multi-database inference may differ.

Figure 3.28 shows the relative contribution of single-database aggregation compared to multi-database inference, for predicted EC terms using protein sequence similarity (BLAST). The comparison categories used are the same as in the previous two sections and have been summarised in Table 3.9. The two categories of provenance of annotation are tracked in the same manner described within the introduction to Section 3.3.2.

Figure 3.28 reveals that for every GeneChip apart from Sugar Cane, the number of annotations contributed by multi-database inference was greater than from single-database aggregation. On average there was 19.38% (standard error = 5.38%) more annotation produced per GeneChip by multi-database inference compared to single-database aggregation. There were also on average 77.72% (standard error = 21.56%) more unique annotations produced by multi-database inference than single-database aggregation.

EC terms predicted through single-database aggregation tended to be more likely to increase the specificity than multi-database aggregation. On average there were nearly three times (279.03%) more annotations predicted by single-database aggregation that improved on the specificity provided by multi-database inference, compared to the reciprocal comparative group (consistent annotations within inference provenance group). This indicates that while producing

more novel annotations, the multiple-database inference were lower in the EC hierarchy than single-database inference.

Of all the annotations produced by both single-database aggregation and multi-database inference 58.18% were predicted identically in both methods. This high degree of agreement in the two methods may either be regarded as strong supporting evidence for the annotations, or indicate some existing inference by the data sources themselves: *i.e.* a database may use EC2GO to translate existing curated GO annotations into EC annotations, which leads to agreement when the procedure is duplicated in CoPSA.

There is a considerable amount of consistent annotation (increased specificity), which amounts to 14,766 single-database aggregated and 3549 multi-database inferred annotations, across all GeneChips. This shows that multi-database inference of annotation can increase the specificity of single-database aggregated EC annotation, and vice versa. However in this instance there was on average per GeneChip, 349.42% more consistent annotation from single-database aggregation. The combining of aggregated and inferred evidence however meant that overall genes could be annotated with more specific EC terms.

Interestingly, the improvement in specificity was dependent on the GeneChip, with Sugarcane gaining the least number of annotations (521) that increased in specificity when both provenances were considered. The Wheat GeneChip, which is the subject of the Part II use case, gained 1,978 and the most successful gains were by Poplar (2742).

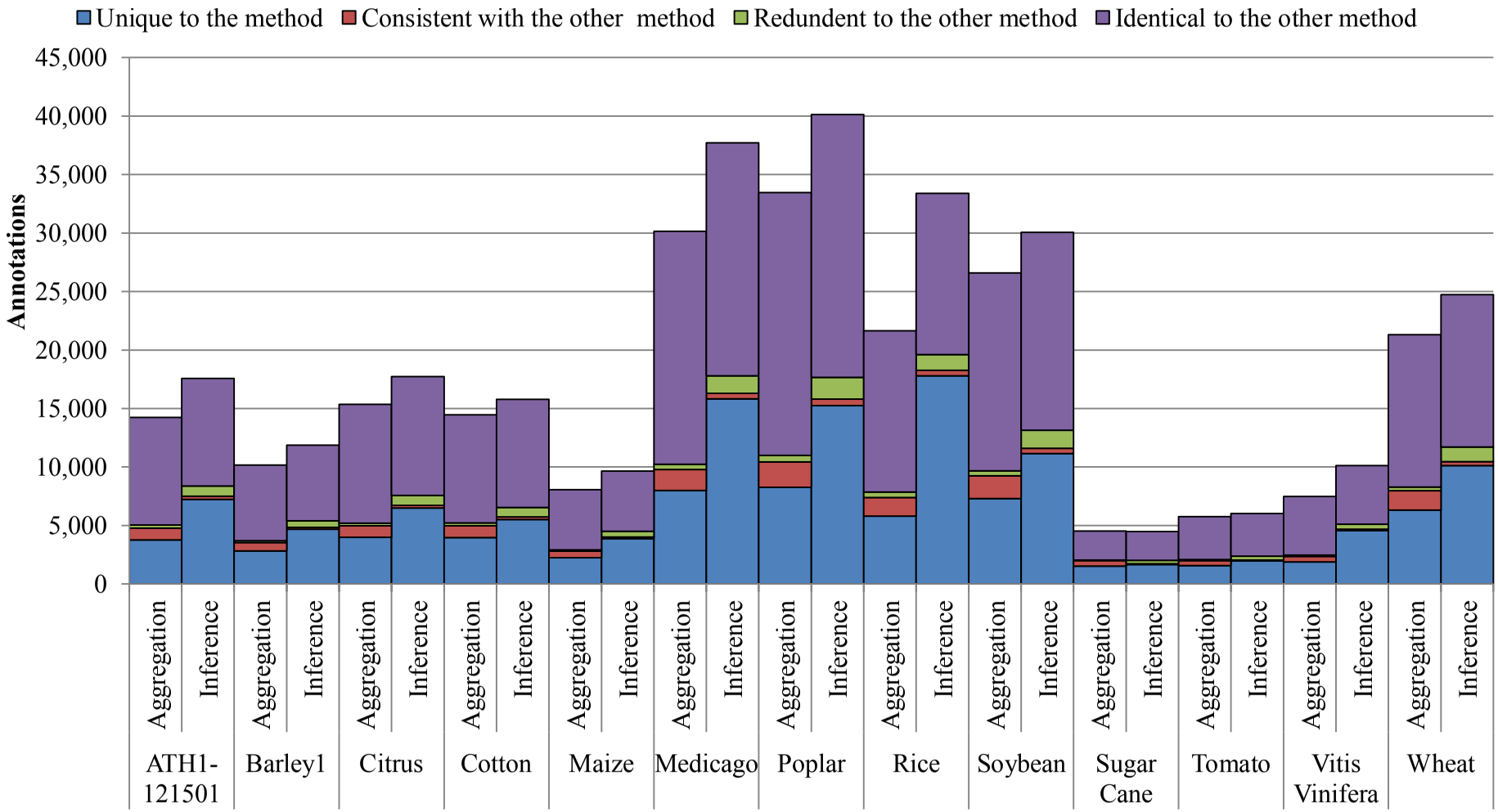


Figure 3.28: A comparison of the number of annotations derived from data aggregation verses inference over an integrated data set for EC annotation via BLAST. The categories of annotation in the legend are defined in Table 3.9.

Figure 3.29 shows the proportion of sequences on the chip that were annotated with EC, using protein sequence similarity (using BLAST) and either single database aggregation or multi-database inference provenance exclusively. As with Figure 3.25 and Figure 3.27, described in the previous two sections, these coverage statistics are divided into sequences that received novel information not present in any of the direct annotation of the databases, and sequences that received annotation already present in direct annotation. The parallel analysis of EC annotation provenance using GeneChip sequence coverage, in addition to raw annotation counts, can yield additional insights into the impact of provenance in the distribution of annotation across the chip.

It was previously shown in this section that multi-database inference contributes more to EC annotation counts than single-databases aggregation. Based on Figure 3.29, it is also apparent that the same holds true with regard to coverage of sequences on the GeneChip. On the average GeneChip 10.93% (standard error = 3.03%) of sequences could be annotated with EC by single-database aggregation, whereas 19.32% (standard error = 5.35%) of sequences could be annotated by multi-database inference.

This suggests that for EC annotation, through sequence similarity, incorporating a multi-database inference approach, is complementary to information derived from aggregating knowledge in existing databases. It leads to both greater quantities of annotations and a larger coverage of the GeneChips studied. While the previous three sections have demonstrated that inference over multiple-databases contributes to the final annotation quantity and coverage, EC annotation using protein sequence similarity, is the strongest demonstration of this yet. It is the first to show that in some instances data inferred across data-sources can exceed that obtained by single-database aggregation alone.

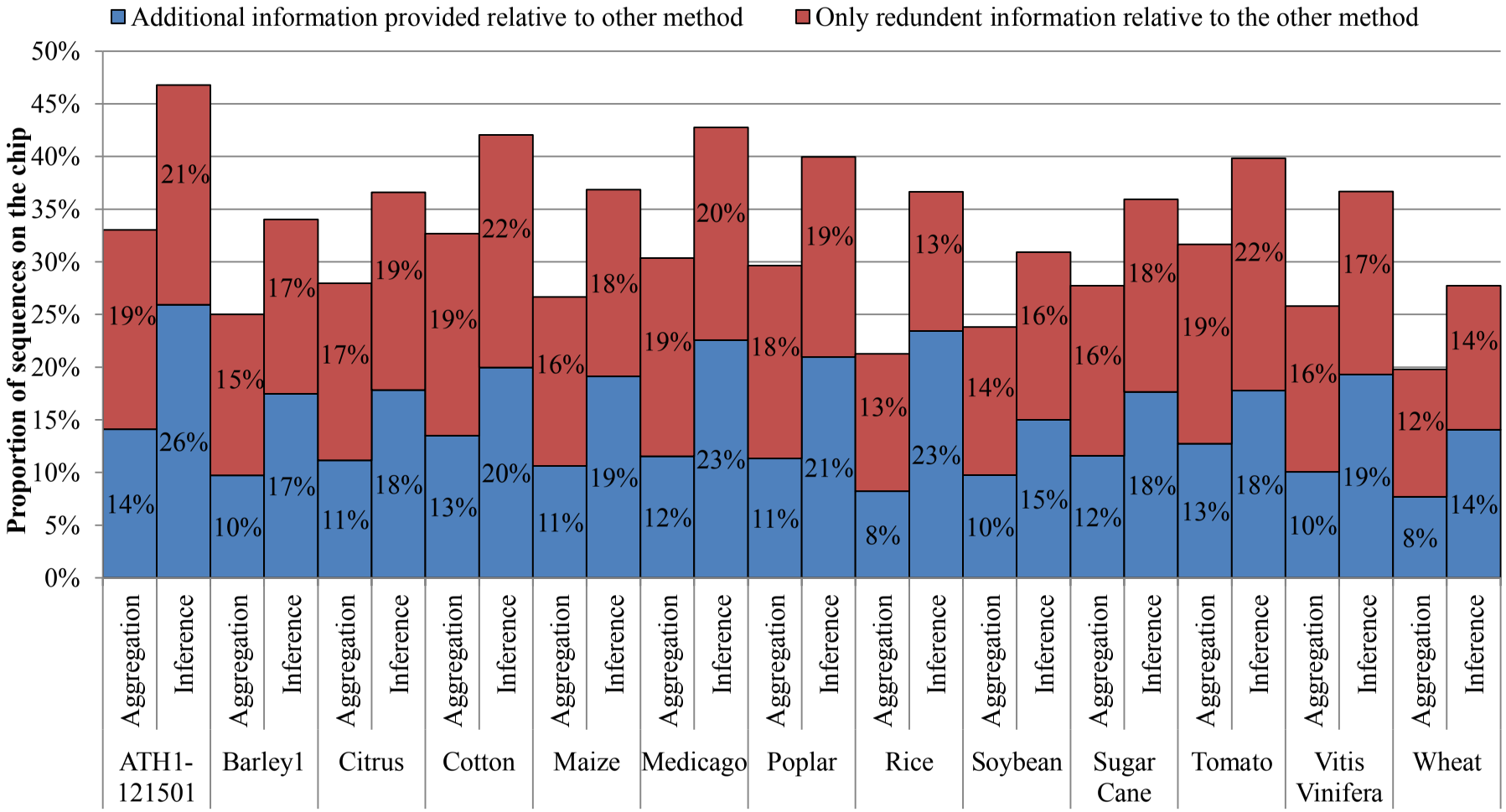


Figure 3.29: The proportion of sequence on the chip that can be annotated with EC terms using the aggregation of primary annotation and inference over integrated data for BLAST.

Annotation of EC terms based on protein domains detected using HMMR

The previous section compared single-database aggregation with multi-database inference, in providing EC annotation through protein sequence similarity using BLAST. This section makes the same comparison but for EC annotation using Pfam protein domains detected by HMMR. Figure 3.30 shows the relative contribution of single-database aggregation compared to multi-database inference, for predicted EC terms using Pfam domains identified through HMMR. The comparison categories used are the same as in the previous three sections and have been summarised in Table 3.9. The two categories of provenance of annotation are tracked in the same manner described within the introduction to Section 3.3.2.

Unlike protein sequence similarity based prediction of EC terms, multi-database inference yielded less EC annotation than single-database aggregation for protein domain based predictions. For the average GeneChip there were 68.60% (standard error = 19.03%) more EC annotations of single-database aggregation than multi-databases inference provenance. 37.90% of annotations predicted by both methods were identical. This was lower than was found using sequence similarity based predictions of EC annotations (58.18%), and reflects the lower yield of multi-database inference, which resulted in a far greater quantity of novel annotation from single-database aggregation.

There were a larger proportion of consistent terms (more specific) for multi-database inference, with 59.47% (standard error = 16.49%) more on average per GeneChip than from single-database aggregation. This indicates that multi-database inference based methods are more likely to increase the specificity of EC terms predicted for domain based annotation. This is the inverse of what was found for protein sequence based prediction of EC in the previous section.

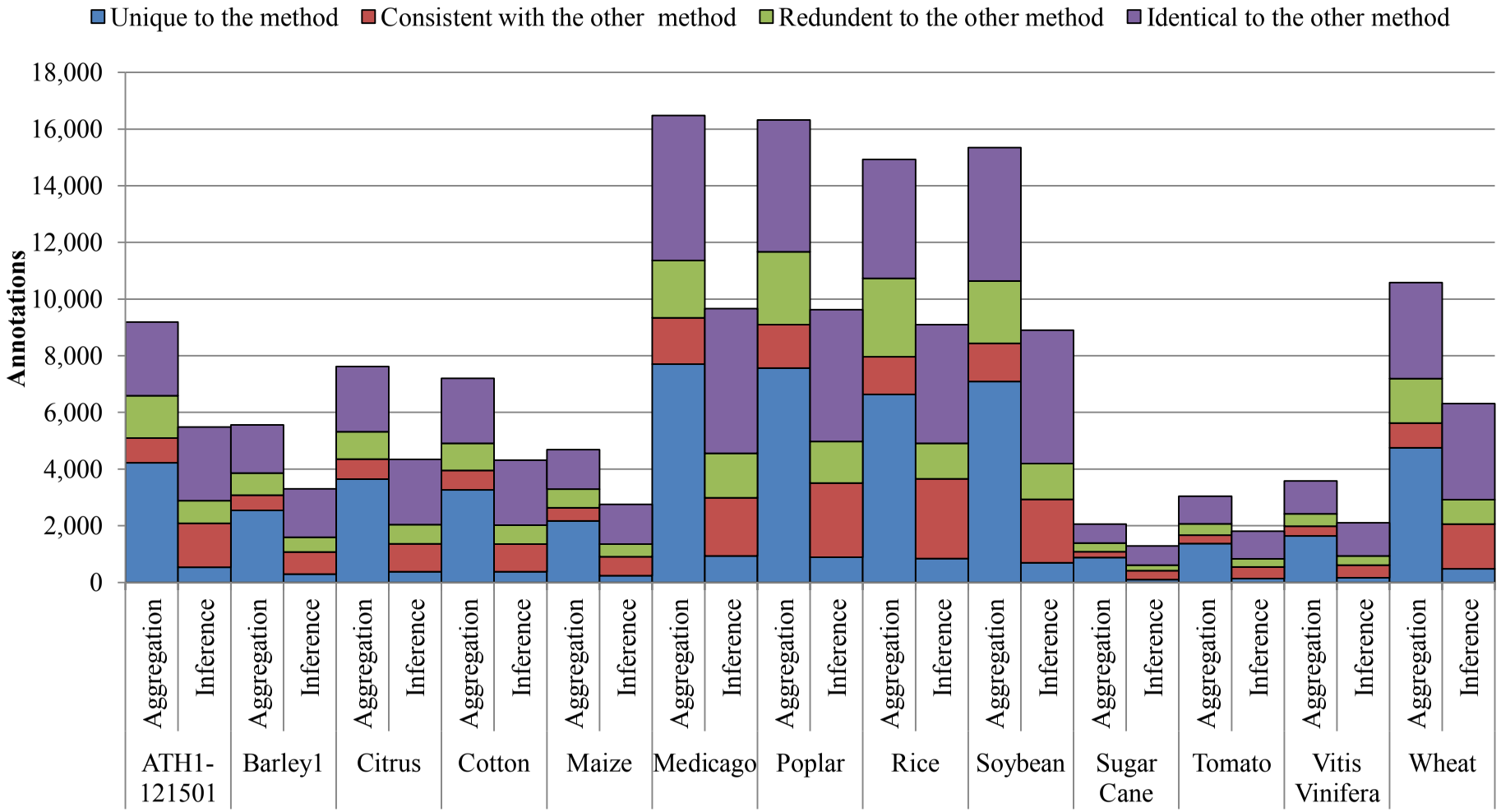


Figure 3.30: A comparison of the number of annotations derived from data aggregation verses inference over an integrated data set for EC annotation via HMMR. The categories of annotation in the legend are defined in Table 3.9. Annotations are filtered based on generalizing EC terms per domain, which is described in Section 3.2.1(*g*).

Figure 3.31 shows the proportion of sequences on the chip that were annotated with EC, using detected domains (using HMMR) and either single database aggregation or multi-database inference provenance exclusively. As with Figure 3.25, Figure 3.27, and Figure 3.29, which were described in the previously in this section, these coverage statistics are divided into sequences that received novel information not present in any of the direct annotation of the databases, and sequences that received annotation already present in direct annotation.
In the first part of this section a greater number of annotations were observed with provenance from single-database aggregation compared to using inference across multiple-database. Provenance from single-database aggregation was also responsible for the greater proportion of novel annotations. As expected Figure 3.31 shows that this translated into a greater coverage of GeneChips sequences with single-database aggregation providence. On average per GeneChip, annotations of single-database aggregation provenance covered 21.20% (standard error = 5.88%) of sequences, whereas multi-database inference covered only 15.33% (standard error = 4.25%). However, the 59.47% more annotations from multi-database inference that improved upon specificity (discussed earlier in this section), resulted in a larger proportion of sequences on the chip receiving additional information from this provenance. On average per chip 10.66% (standard error = 2.96%) of sequences received additional information from multi-database inference, whereas only 5.06% (standard error = 1.40%) of sequences benefited in this way from single-database aggregation.

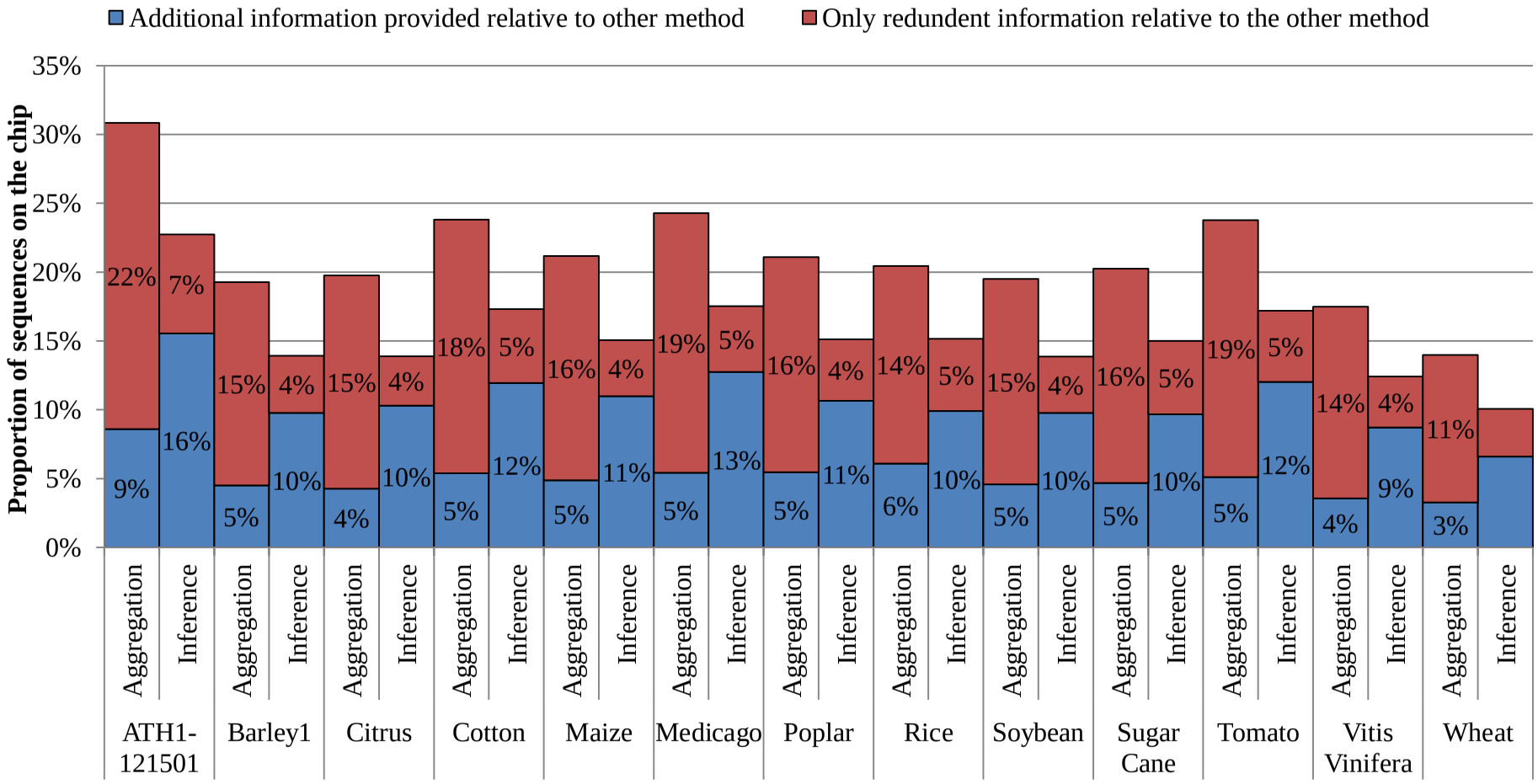


Figure 3.31: The proportion of sequence on the chip that can be annotated with EC terms using the aggregation of primary annotation and inference over integrated data for HMMR. EC terms are generalized according to domain as described in Section 3.2.1(*g*).

Comparison of CoPSA annotation coverage for the wheat GeneChip against other pipelines

CoPSA aims to provide improved annotation of the consensus sequences for Affymetrix GeneChips for non-model organism species, where experimentally validated GO terms are rare or non-existent. For the comparison of CoPSA to other sequence annotation pipelines, we will focus on wheat, which was one of the most challenging arrays to annotate, and is the subject of the use-case for this thesis. While it would be desirable, a comparison for all plant arrays is problematic, as comparable sequence annotations other than from NetAffx are not available for all arrays. Comparisons for wheat were made against the Affymetrix NetAffx version 30 annotation and the BLAST2GO annotation (downloaded 27/05/2009 from the B2G-FAR website (Escobar, 2011)).

Figure 3.32 shows a Venn diagram of the coverage of consensus-sequences annotated on the Wheat GeneChip with at least one GO *molecular function* term for Affymetrix-NetAffx, BLAST2GO pipelines, and CoPSA (using protein sequences and domains) pipelines. It is evident that CoPSA provided more candidate GO annotations than either of the other pipelines but there are significant overlaps with the annotations from the other pipelines. CoPSA was able to annotate 90% and 97% of the candidates annotated by NetAffx and BLAST2GO respectively. A total of 37% of the chip was annotated only by CoPSA. Conversely, BLAST2GO and NetAffx provided annotation on 0.8% of the chip that was not covered by CoPSA.

Figure 3.33 shows a Venn diagram of the coverage of consensus-sequences annotated on the Wheat GeneChip with at least one GO *biological process* term for Affymetrix-NetAffx, BLAST2GO pipelines, and CoPSA (using protein sequences and domains) pipelines. For GO biological-processes, CoPSA annotated 86% and 88% of the consensus sequences annotated by NetAffx and BLAST2GO respectively. 34.37% of the chip was exclusively annotated by CoPSA. BLAST2GO and NetAffx provided annotation on 2.4% of the chip that was not covered by CoPSA.

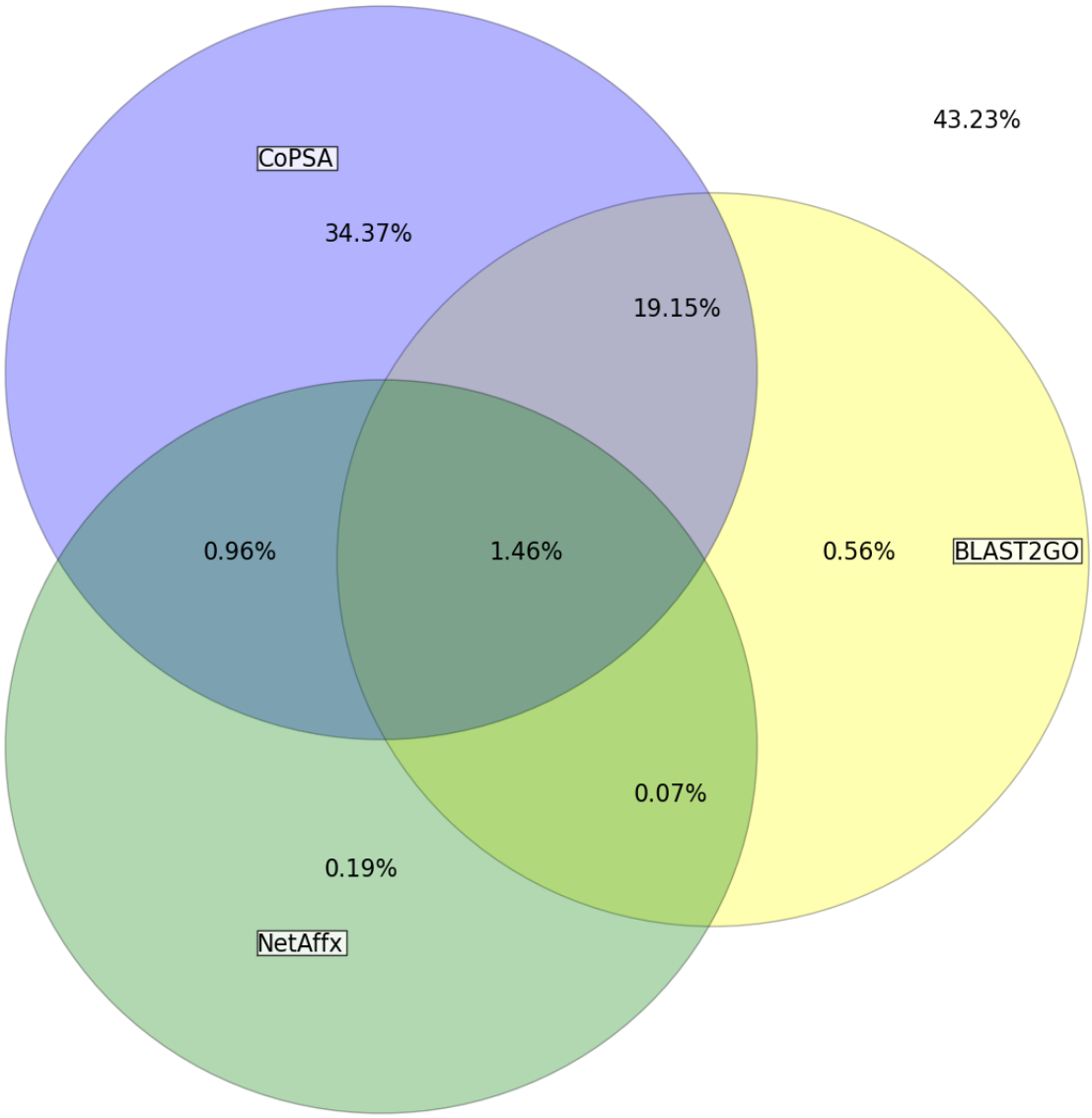


Figure 3.32: The wheat chip sequence coverage for GO molecular-function annotation, compared to BLAST2GO and NetAffx annotation. The value in the top right represents the sequences on the chip that were not annotated by any of the pipelines.

Figure 3.34 shows a Venn diagram of the coverage of consensus-sequences an-

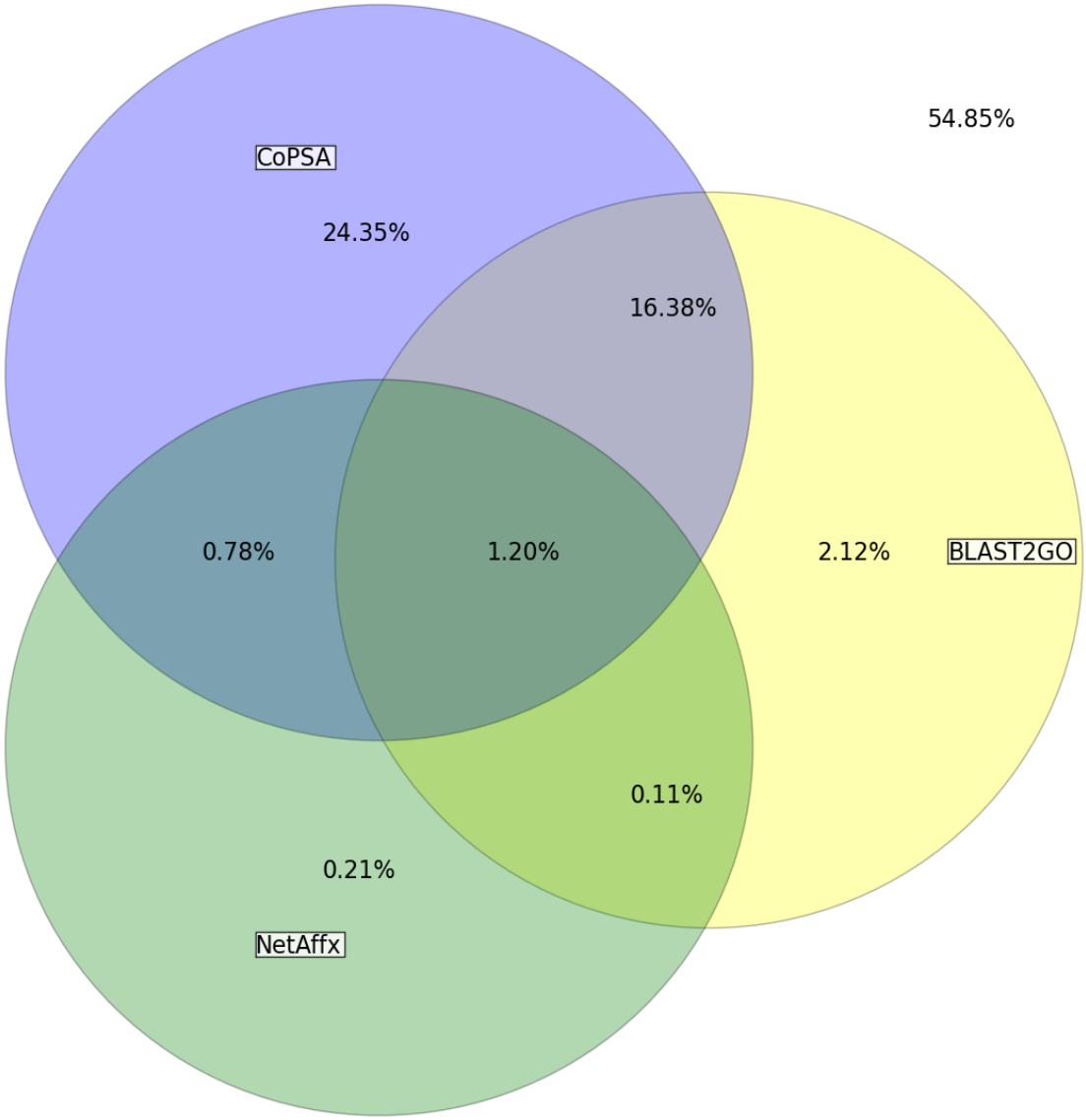


Figure 3.33: The wheat chip sequence coverage for GO biological-process annotation, compared to BLAST2GO and NetAffx annotation. The value in the top right represents the sequences on the chip that were not annotated by any of the pipelines.

notated on the Wheat GeneChip with at least one GO *cellular component* term for Affymetrix-NetAffx, BLAST2GO pipelines, and CoPSA (using protein sequences and domains) pipelines. For GO *cellular components terms*, CoPSA annotated 90% and 95% of the consensus sequences already annotated by NetAffx and BLAST2GO respectively. Unique annotation by CoPSA accounted for 24.35% of the chip. BLAST2GO and NetAffx provided annotation on 1.2% of the chip that was not covered by CoPSA.

Those annotations which overlap with other pipeline annotation predictions

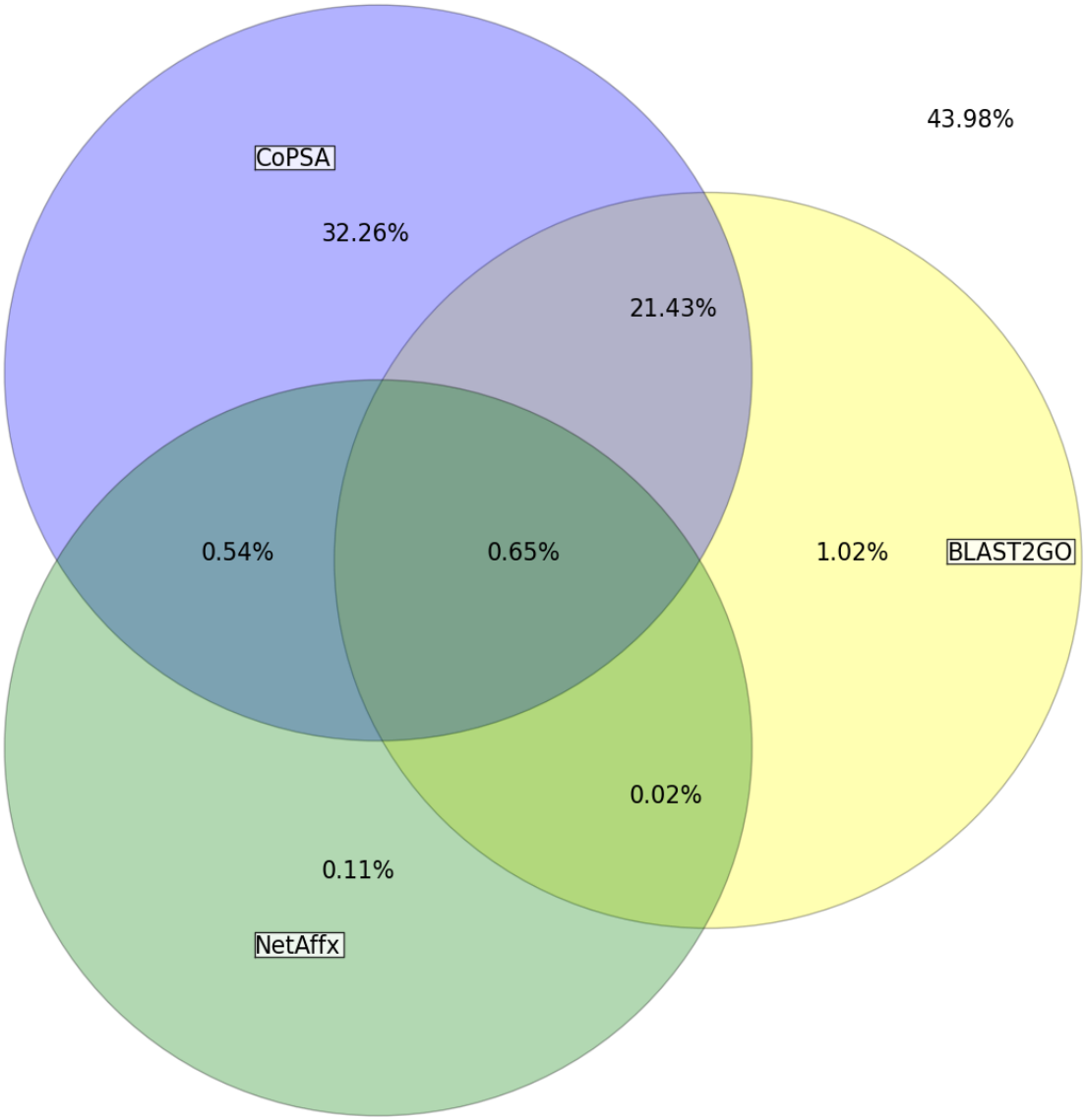


Figure 3.34: The wheat chip sequence coverage for GO cellular-component annotation, compared to BLAST2GO and NetAffx annotation. The value in the top right represents the sequences on the chip that were not annotated by any of the pipelines.

form an important part of validating CoPSA, and the next step was to compare the content of these annotation sets. The semantic similarity between these annotations is highly dependent on the selection of annotations from the candidates and this will be discussed later in this chapter.

3.3.2(*a*) Transcription factor annotation

Figure 3.35 shows the proportion of consensus sequences that were identified as a transcription factor by CoPSA, for each of the Affymetrix GeneChips studied. Previous work by 3.37 provides estimates for the number of transcription factors that can be expected in plant species based on other sequenced organisms (Figure 3.37). The number of transcriptions factors found by CoPSA was slightly higher than the 6% of the genome predicted by 3.37, with the exception of the ATH1 chip, which was 14.23% of the GeneChip. The CoPSA predictions for the Wheat GeneChip, which are used in the Part II use-case, are closest to the 6% observation. However, 6% is a very rough estimate, the number of transcription factors in a genome may vary according to species and Yilmaz *et al.* (2009) observes considerable variation in Figure 3.37. Many of the GeneChips studied detect only a subset of genes in the target genome, and therefore may not contain a representative number of transcription factors.

Figure 3.36 shows the number of transcription factors identified on each Gene-

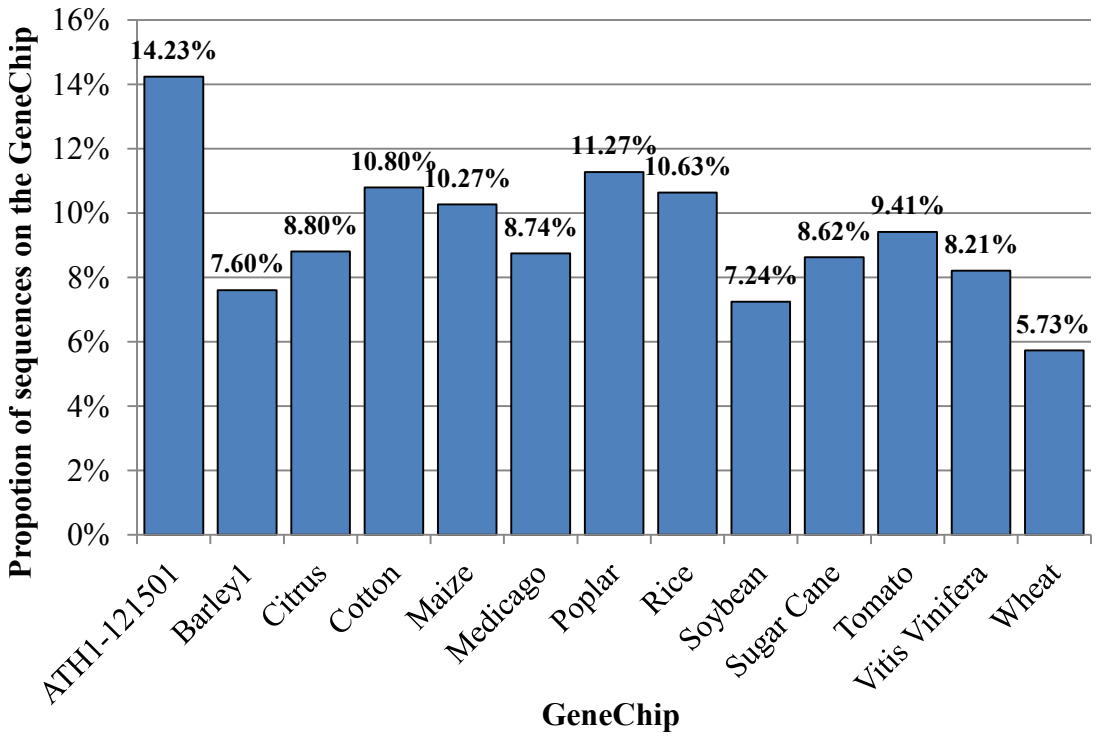


Figure 3.35: The proportion of consensus sequences on each GeneChip that were identified as potential transcription factors by CoPSA.

Chip by CoPSA. It is not readily comparable to the results by Yilmaz *et al.* (2009), which are shown in Figure 3.37, because most of the GeneChips were designed for partially sequenced organism and as such detect only a subset of transcription factors in the genome. Their prediction of just under 3,500 transcription factors, in a genome of around 60,000 genes for Maize is higher than the 1,803 found by CoPSA. However, the Maize GeneChip is built from NCBI's GenBank (September 29, 2004) and *Zea mays* UniGene Build 42 (July 23, 2004) databases, and is designed to detect expression of only 13,339 genes (AFFYMETRIX, 2004). If the Maize gene chip contains the same proportion of transcription factors as the genome, then the CoPSA predictions are in line with the predictions of Yilmaz *et al.* (2009). The Sugar Cane genome is expected to contain slightly greater than 2,000 transcription factors respectively. CoPSA identified 709 transcription factors on the Sugar Cane GeneChip. This GeneChip is built from *Saccharum officinarum* UniGene Build 5 (August 27, 2004) and GenBank mRNA (November 2, 2004), which is designed to pick up expression of 6,024 genes. Sugar Cane is expected to have in the region of 37,000 genes. The CoPSA predictions are therefore much higher than Yilmaz *et al.* (2009) projected. This could be caused by three factors (1) the proportion of transcription factors on the Sugar Cane GeneChip is not representative of their frequency in the genome, (2) Sugar Cane has an unusual overabundance of transcription factors, or (3) CoPSA has a high false positive rate of transcription factor prediction for this GeneChip. The first causative factor seems the most probable. CoPSA false positive rate is more likely to be related to methodologies, and databases. The false positive rate specific to a species, is likely to be cause by its distance from *Arabidopsis*.

Predictions for the ATH1 GeneChip were both too high in terms of coverage (14.23% shown in Figure 3.35) and the quantity of transcription factors (3,237 shown in Figure 3.36). Palaniswamy *et al.* (2006) predict 1,690 transcription factors for *Arabidopsis*. This seems to be a consequence of setting the parameters of CoPSA to pick up more distant putative functional-orthologs. The ap-

propriate alignment thresholds for transferring annotations from *Arabidopsis* to organisms like Poplar and Wheat, does not appear to be appropriate for annotating *Arabidopsis* sequences, and results in higher false positives. As previously stated, in this chapter, the aim of this configuration of the CoPSA pipeline is to demonstrate the annotation of non-model organism. It is therefore inappropriate to use CoPSA as a source of *Arabidopsis* TF annotations. It is included just for comparison purposes.

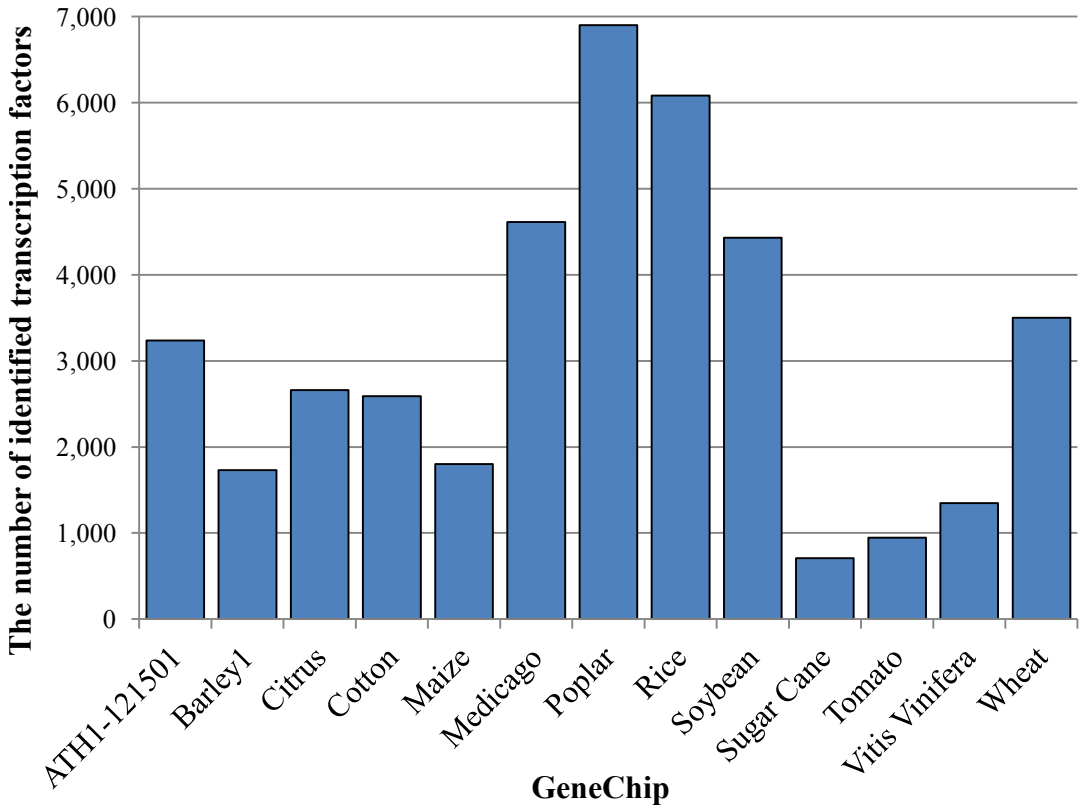


Figure 3.36: The number of transcription factors found by CoPSA on each GeneChip.

3.3.3 Analysis of CoPSA annotations and metrics

The evaluation of CoPSA and the proposed metrics is broken down into three aspects. Firstly, a consideration of confidence of annotation is made for each of the scoring metrics proposed within CoPSA. Confidence is assessed based

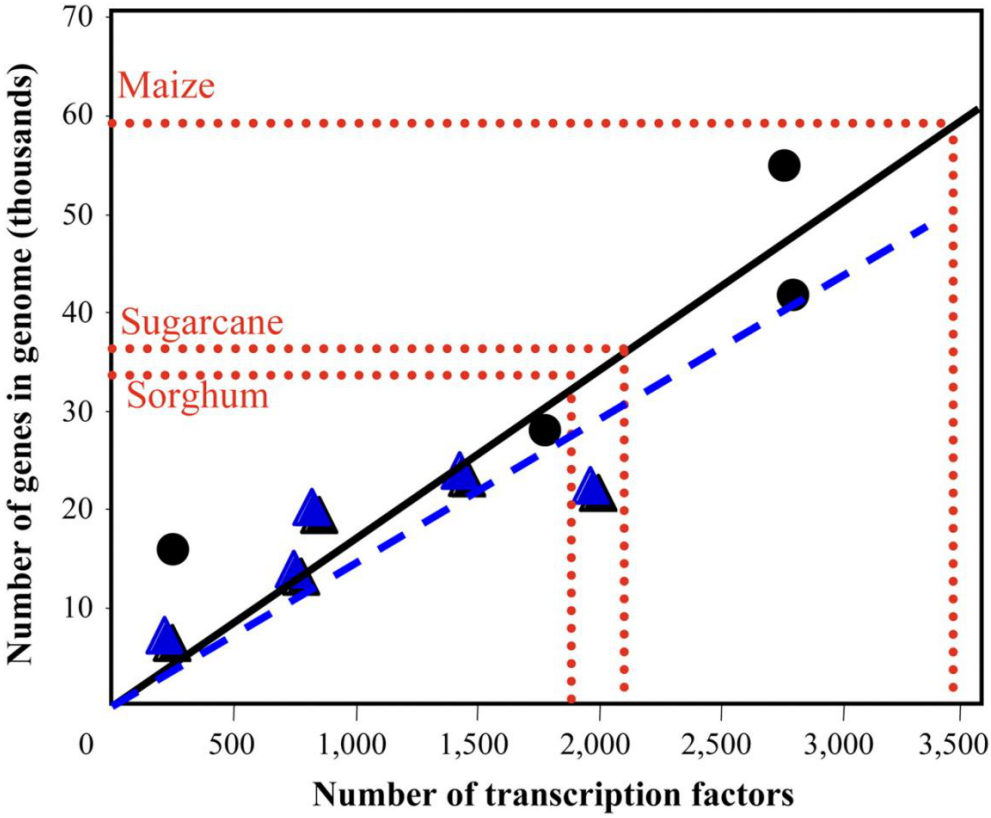


Figure 3.37: An estimation of the number of transcription factors in grasses by Yilmaz *et al.* (2009). Black circles represent *Arabidopsis*, rice, poplar, and *Chlamydomonas* (r^2=0.87). Blue triangles represent yeast, fruit fly, mouse, and human (r2=0.74).

on the original evidence of this annotation and the similarity of sequences. Secondly, a range of properties of the annotation are examined for each of the metrics and Blast2GO and NetAffx annotations. Thirdly, NetAffx is used as an incomplete but high quality source of gold standard annotations. The ability of CoPSA to correctly annotate genes with the same or similar annotations to NetAffx is assessed using the hierarchical recall metric proposed by Verspoor *et al.* (2006), which has previously been described in this Section 3.3.3(*c*).

3.3.3(*a*) Confidence in annotations from scoring strategies

The two variants of fitness metrics proposed around the Multiple Weighted Fitness Measures (MWFM) effectively make a compromise between sequence similarity, evidence codes, and degree of functional consensus with the other similar proteins. It is expected therefore that relative to the other confidence metrics

they compromise on selecting the most similar sequence, where a suboptimal sequence alignment provides better evidence or is more consistent with other sequences. Figure 3.38 was generated for each CoPSA post filtering method that was proposed in 3.2.2, by extracting the alignment of every protein used to transfer annotation to a query sequence, and then calculating the mean bitscore for all these proteins. The MWFM-OE method was identical in bitscore to the MWFM, as it filters at the post protein selection stage of annotation selection. It is therefore not included in Figure 3.38. If a method such as the MWFM compromises on protein-sequence alignment strength in favour of stronger experimental evidence, then it should result in a lower mean bitscore. Figure 3.38 therefore reveals an expected pattern of results with the BestHit and Epsilon methods resulting in the highest bitscores. The Epsilon approach affects the greatest average bitscore of proteins used in inference, because it selects for the highest bitscore while maximising the number of high scoring proteins used in analysis. The CoPSA-Union method average bitscore represents the mean of all proteins found at the given thresholds. It would therefore be expected that MWFM produces a lower average bitscore if it only selected for proteins that would be more distant in evolutionary terms but would carry stronger evidence terms; which in this instance would come from *Arabidopsis thaliana*, Algae, and Yeast. However, the average bitscore of MWFM is similar to CoPSA-Union and this suggests that a compromise has been found between evidence strength and sequence similarity.

Figure 3.38 showed that MWFM results in annotation being transferred from weaker alignments. This results from a compromise on sequence similarity in order to promote evidence-rich proteins for inference. Figure 3.39 was generated for each method by taking the evidence code for each annotation transferred, looking up the weighting for this evidence code using Table 3.8, and the then calculating the mean weighting for the method. If MWFM is compromising on sequence similarity in order to transfer more annotation with stronger

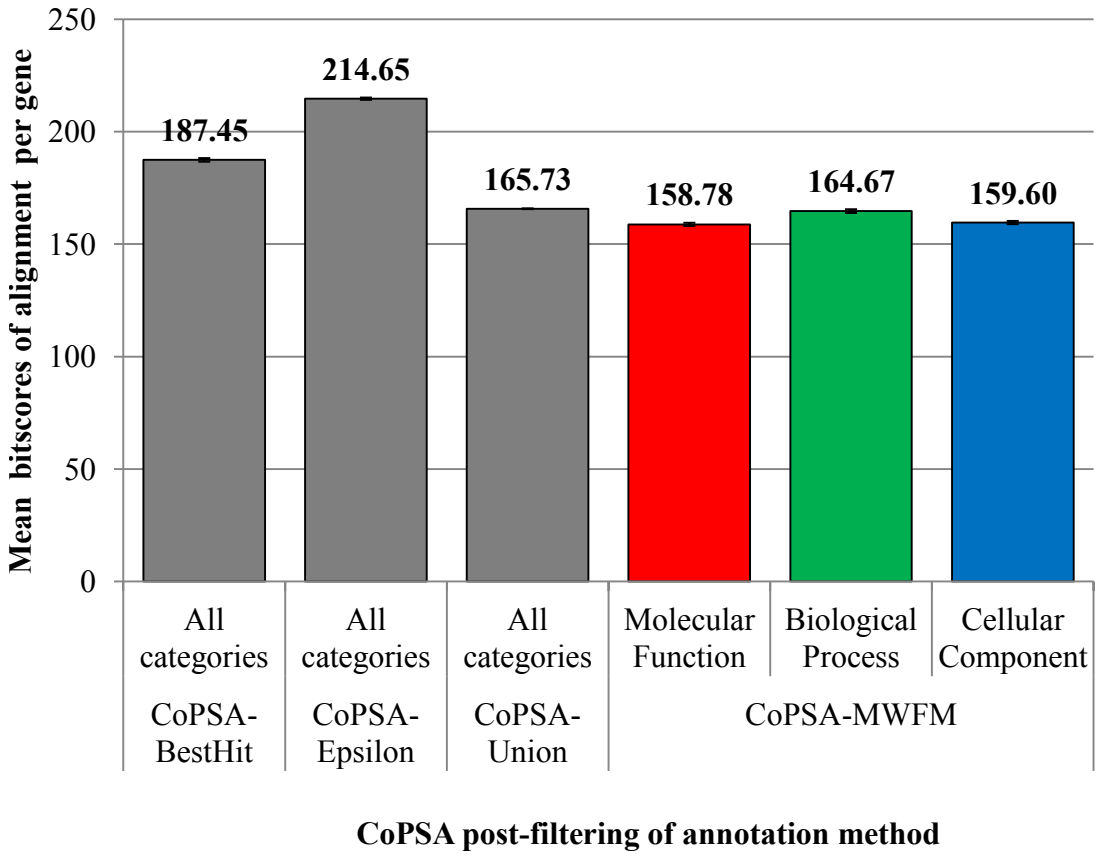


Figure 3.38: Average evidence weighting for annotations produced by the five strategies. For the two selection strategies that treat the three categories of the Gene Ontology separately, the respective values are shown.

evidence, then it would be expected that MWFM achieves a larger mean evidence weighting. As expected Figure 3.39 shows that MWFM this affects an increase in the mean annotation evidence code weighting relative to the BestHit and Epsilon strategies that act to maximise bitscore alone. The average evidence weight for annotations across all categories is 0.54 (standard error = 0.002) and 0.65 (standard error = 0.002), for CoPSA annotations based on MWFM and MWFM-OE respectively. This indicates that in terms of the average evidence weight both the MWFM and MWFM-OE outperform all three of the bitscore based confidence measures. The optimisation of evidence in MWFM-OE increases the confidence in annotation across all three aspects, without compromising on the bitscore.

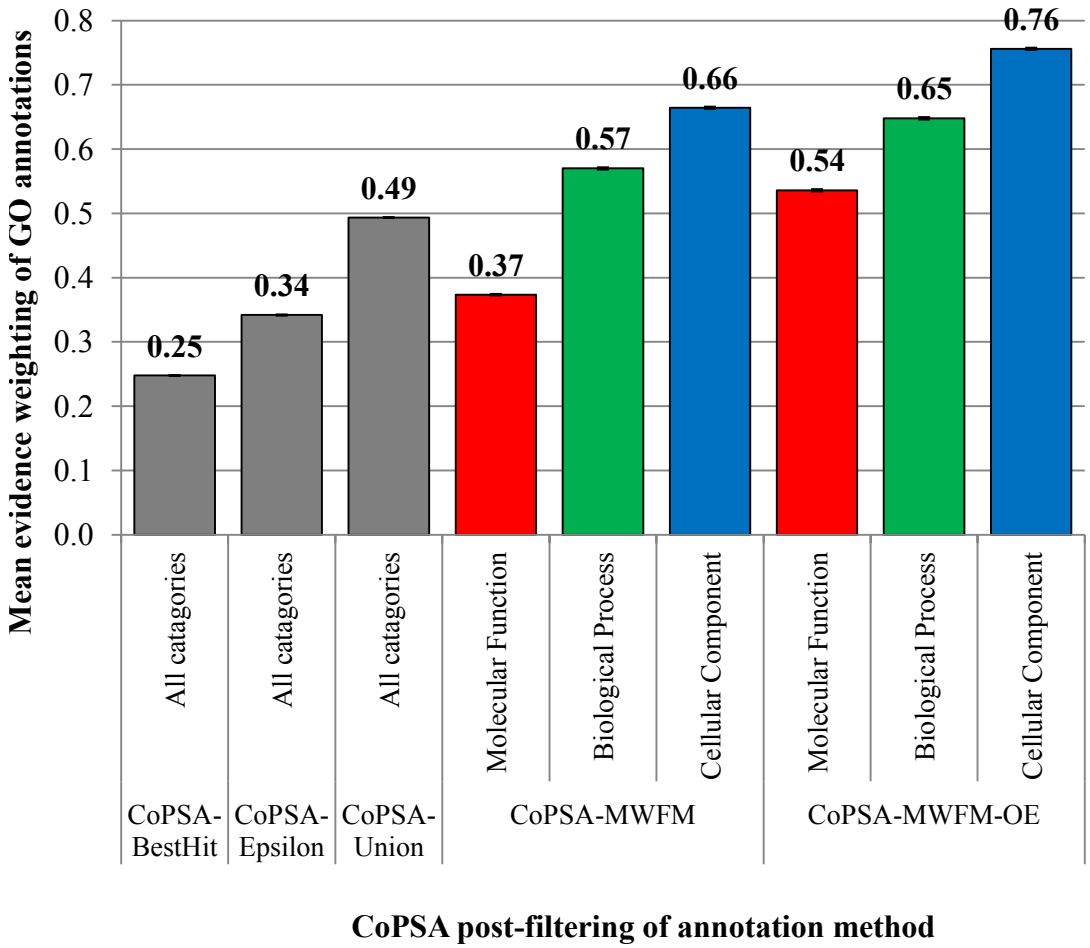


Figure 3.39: The mean bitscore of proteins that were used in annotation inference, for the four candidates for a CoPSA annotation selection strategy. For the MWFM selection strategy that treats the three categories of the Gene Ontology separately, the respective values are shown. Error bars are standard error of the mean.

3.3.3(*b*) Comparative properties of the annotation

A very simple measure that can be applied to an annotation pipeline is to compute the number of GO annotations proposed per-gene in each category. Figure 3.40 was calculated using the mean number of annotations proposed per query sequence, for each post-filtering method, and each of the GO categories. A low number of predictions per-gene in the GO *molecular function* or *biological process* categories (less than 2) indicates that the annotation method has a low recall. This is because GO is designed such that multiple aspects of a gene function can be described (*e.g.* catalytic activity and ATP binding), multiple processes that a gene is engaged in (*e.g.* regulation of catalytic activity and proteolysis),

and multiple cellular locations (*e.g.* integral to membrane and nitrate reductase complex). It is unusual for a gene function to be only applicable to one term from the function or process category.

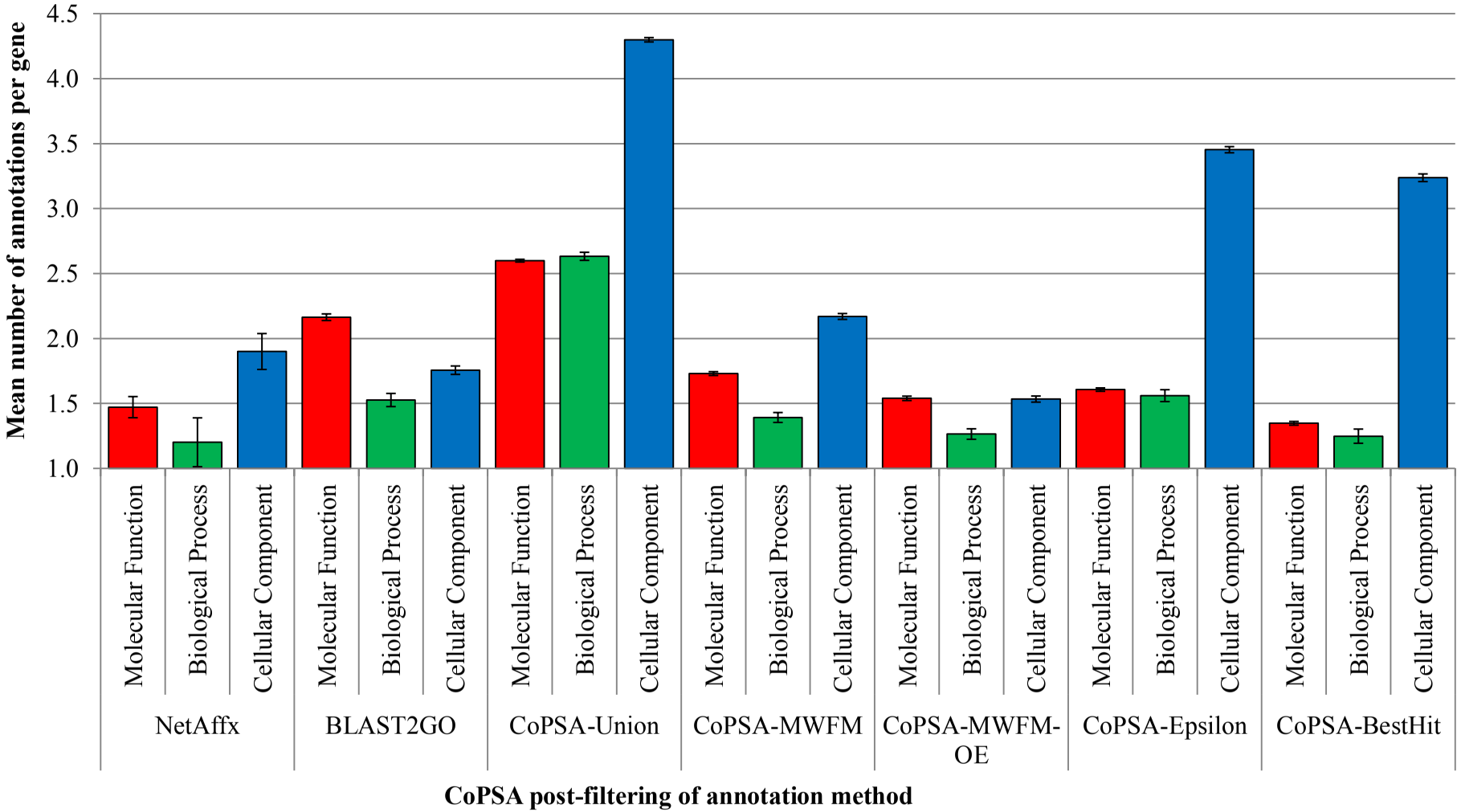


Figure 3.40: Mean number of annotations per gene. Error bars are standard error of the mean.

An example of multi-process assignment is shown by the graph in Figure , where CoPSA post-filtered by MWFM-OE correctly assigns two processes to the highly conserved Glutamate-Synthase-2 (GS2) gene, Ta.28435.1.S1_at. The process *Oxidation reduction*, describes the chemical process undertaken in the reaction, whereas *Glutamate biosynthetic process* describes the metabolic pathway-level process that the gene is involved in. NetAffx contains a correct free text description of the gene function, but fails to annotate any GO terms to the gene. Blast2go identifies the correct metabolic pathway process but does not assign the chemical process involved. However, it additionally annotates GS2 with the Glutamine biosynthetic process. While these pathways are adjacent, this gene is involved in Glutamine catabolism rather than biosynthesis. However, as it forms a cycle together with Glutamine synthetase, the catabolism can also form a precursor to the biosynthesis (Miflin and Habash, 2002), highlighting the complexity and ambiguity of process annotation.

As well as indicating low recall, the mean number of annotations per-gene (Figure 3.40), can indicate low precision if the number of annotations within a category is unfeasibly high. Identifying the point where the number of terms assigned to an annotation becomes improbable is not an easy task, and depends on the gene and structure of the gene ontology. Using the measure of semantic coherency described in Section 4.2.5, together with the number of annotations goes some way to addressing this issue, as it indicates the distribution of the annotations within the ontology. A small number of terms that are moderately spread throughout the GO tree, indicates that multiple aspects of the function or process are being described. Whereas the annotation of a large number of highly coherent terms (adjacent in the tree), indicates there is some disagreement in the sub-functionalization of a gene. Conversely, a low coherency may also indicate incompatible annotation. This ambiguity in interpreting semantic coherency, limits its usefulness but it can still be a useful tool when comparing pipelines, and identifying outliers. When comparing a high quality but low

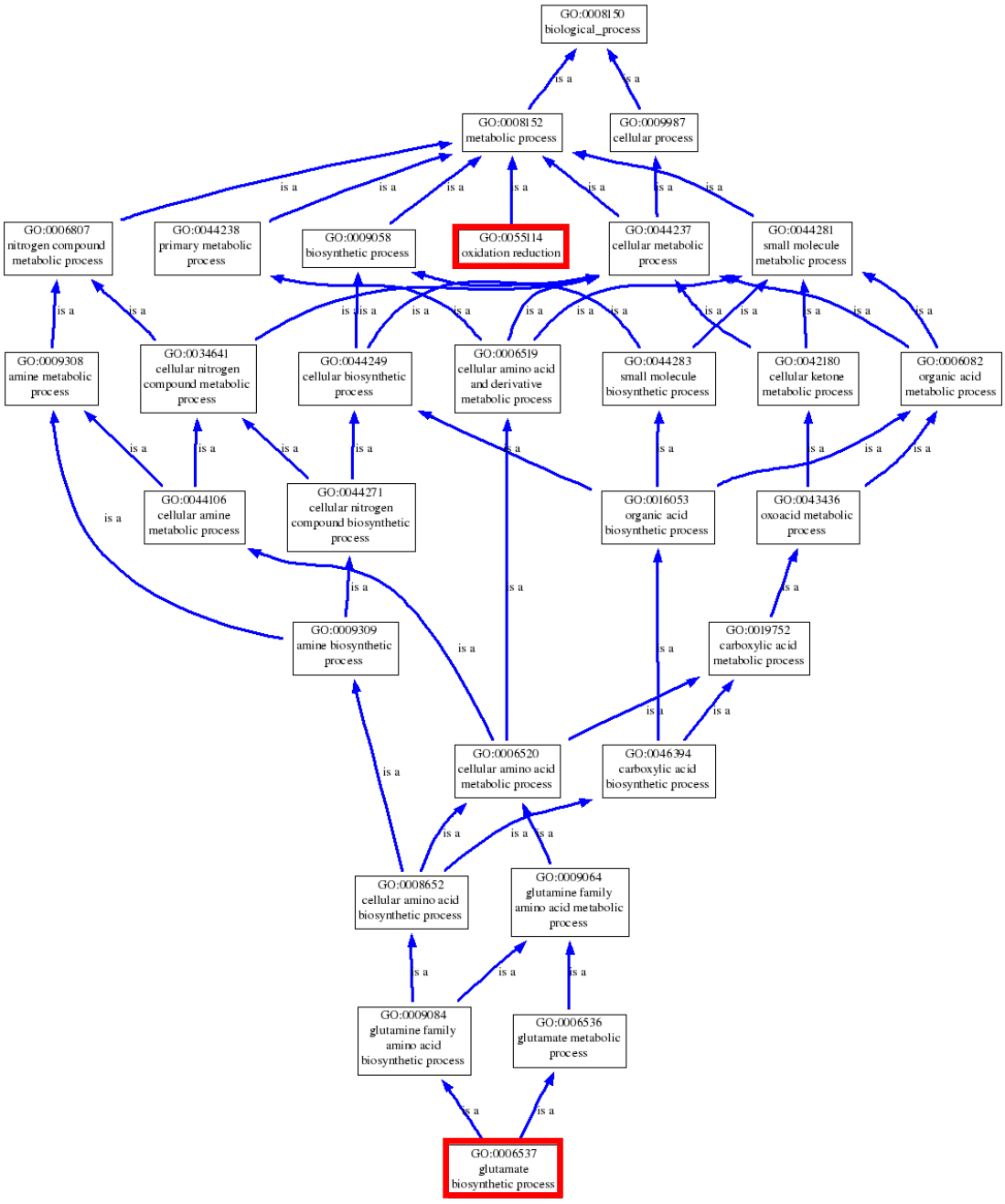


Figure 3.41: Highlighted in red within the Gene Ontology process category are two terms correctly annotated to Ta.28435.1.S1_at (Glutamate Synthase 2), by CoPSA-MWFM-OE. The annotation describes the chemical process of oxidation reduction, and a biochemical pathway process of glutamate biosynthesis.

coverage pipeline, against a high coverage pipeline, a similar semantic coherency indicates that the structure of the gene annotations for the latter pipeline follows a similar pattern of annotation to the first.

Figure 3.42 shows for each of the post filtering methods, the mean semantic coherency between annotations that were transferred to the wheat GeneChip query sequences. A comparison to BLAST2GO and NetAffx annotations for the wheat GeneChip has also been added to Figure 3.42. A mean value is provided for each GO category. However, coherency is dependent on the structure of each and so inter-category comparisons are not possible. From Figure 3.42 it is evident that the CoPSA-Union post-filtering methodology displays the lowest semantic coherency of all three categories. Furthermore, CoPSA-Union has a response pattern that is the least similar to the NetAffx and BLAST2GO pipelines. Since CoPSA-Union is the most inclusive pipeline, it might be expected to generate the highest number of false positive annotations. This result suggests that low semantic coherency may well be a product of large numbers of false positives. The nature of the GS2 metric means that increasing the number of terms, which are annotated to each gene, restricts the minimum and maximum coherency of sampled terms. The large number of annotation per gene for CoPSA-Union (Figure 3.40), acted as a moderator of low coherency, and contributed to the lower coherency through the lower abundance of small sets with high coherency.

CoPSA-MWFM produces a similar response pattern that is more similar to NetAffx and BLAST2GO in all categories, indicating the more selective MWFM weighting scheme acted to increase the coherency of annotation. Coherency within annotation sets, however, was not directly selected for by the MWFM weighting, as GS2 was used only for calculating semantic distance within categories. However, the main strategy of MWFM is that it acts to select the annotations from one putative functional-ortholog only; whereas Union aggregates annotations from across a range of putative functional-orthologs.

The OE extension to CoPSA-MWFM, which selected for experimentally validated annotations, showed increased coherency in all categories relative to basic MWFM filtering strategy. This is closely linked to a moderate decrease in the number of annotations per gene (Figure 3.40); indicating OE is acting to exclude the more distant annotations in a set.

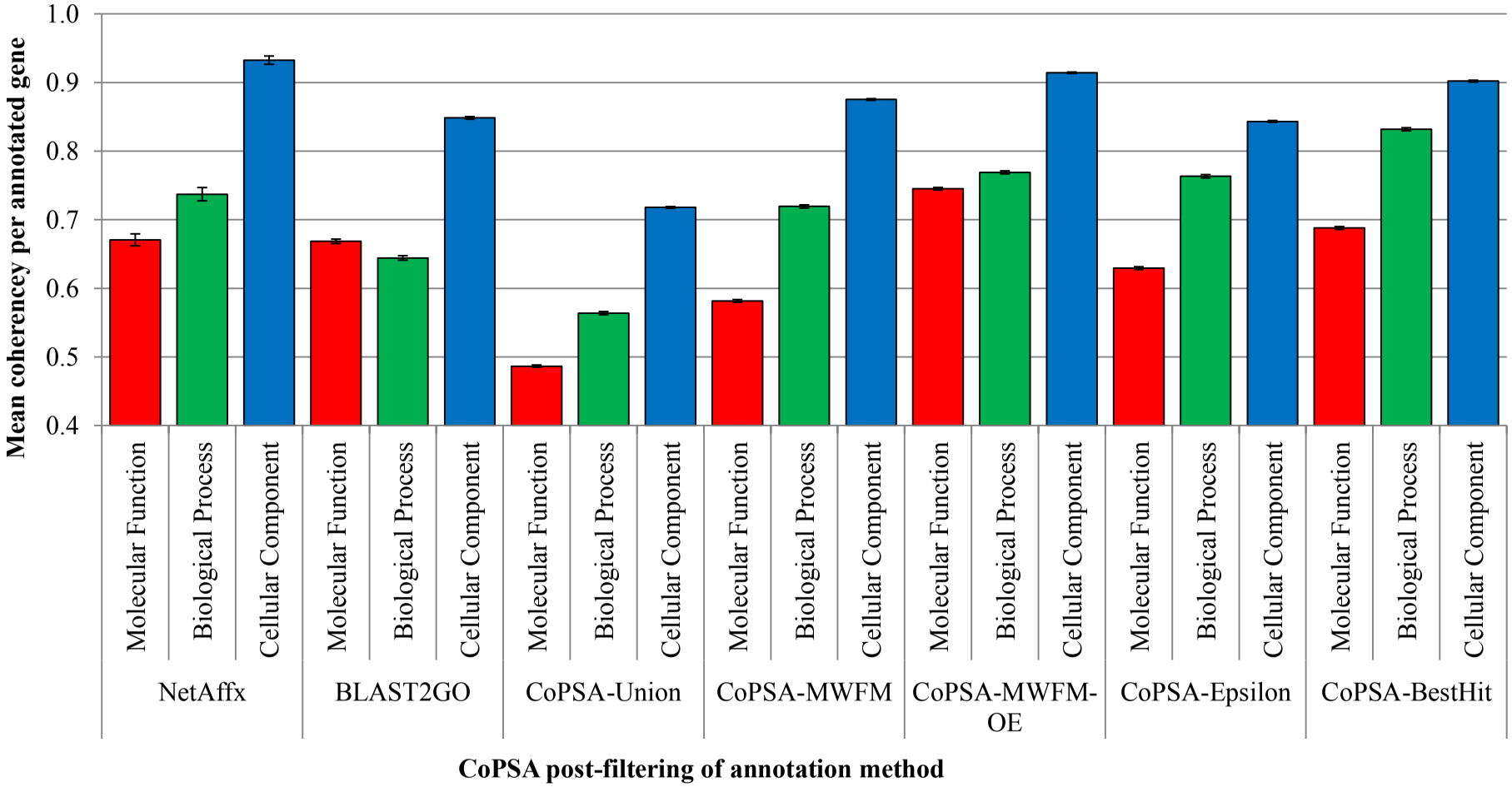


Figure 3.42: The mean coherence of annotations for genes with two or more annotations. Shown for the five filtering strategies in CoPSA plus NetAffx, and BLAST2GO. The error-bars are standard error of the mean.

Figure 3.43 shows a comparison of the different filtering strategies and pipelines in terms of the mean information content (IC) of their annotations for each of the GO categories within each of the methods. The method for calculating IC as been previously described in Section 3.2.1(g). With a given category, the mean IC values were calculated from the GO annotations of each gene, and then the mean of these means were taken across the genes. The mean IC score of the functional annotations produced by a method therefore reflects the IC of the annotation usage, as opposed to the IC of the GO terms utilised (the mean IC of the set of GO terms used for annotation by a method). The IC values observed were dependent on the size and structure of the GO category. Comparisons can therefore be made between methodologies within a given GO category but not between categories.

The CoPSA annotation filtering strategies all produce higher mean IC scores than NetAffx annotations and similar IC scores to BLAST2GO. Figure 3.44 shows the richness of annotations from each GO category, produced by each of the post-filtering methods and comparison pipelines. Richness is as defined in Section 3.2.3, which is simply the proportion of the gene ontology used in the functional annotation. Figure 3.45, shows the mean structural specificity of annotations from each GO category, produced by each of the post-filtering methods and comparison pipelines. Structural specificity is defined in Section 3.2.3, and is simply the number of ancestors for a GO term, which increases as GO terms become more specific (further from the root) in the GO Directed Acyclic Graph (DAG). IC scores (Figure 3.43) should be viewed together with the richness (Figure 3.44) and the structural specificity of terms used in the annotations (Figure 3.45). The structural specificity tends to be related to IC because, terms that are lower in the GO DAG, tend to be more rarely used. Because of the way IC is defined (Section 3.2.1(g)), terms that are higher in the DAG (closer to the root term) accumulate implicit annotation from their children. This means that IC scores drop as terms approach the root, and consequently the root contains no

information.

NetAffx has the lowest IC content of all the pipelines, but has a similar or better structural specificity. NetAffx also has the lowest coverage of the Gene ontology (richness). This indicates that NetAffx is mainly annotating using terms that are more specific in the GO hierarchy but have low IC score. This means that they are commonly used terms for annotation by all pipelines. This highlights a limitation of IC as a measure of annotation accuracy, because it penalises terms that are functionally informative, but frequently used, either because they represent a common function in the cell, or because they are well known or easily characterised (*e.g.* highly conserved proteins such as those found in central metabolism). In this instance, the low IC of NetAffx does not indicate it has poor quality of annotation, but rather a low coverage, and is highly conservative. The high quality of NetAffx annotation was confirmed by evaluating the annotation of a number of known genes, which included inspection of glutamate synthase annotations by an expert in the field (D Habash, data not shown). BLAST2GO has the highest structural specificity, and maintains a similar IC mean to CoPSA. This indicates that it is annotating specific terms in the tree that are more rarely annotated in the wheat GeneChip. However, the relatively low richness of annotation (Figure 3.44) relative to CoPSA indicated that high IC terms were selected at the expense of a more general coverage of the breadth of the ontology.

All the CoPSA annotation-selection strategies have the effect of increasing IC relative to the CoPSA-Union method. MWFM results in the greatest average IC content, whereas the OE adaption results in reduced IC. This indicates that selecting for experimentally derived annotations biases the annotation against more rarely used annotations. It implies experimental evidence transferred by protein sequence similarity is, more frequently, less specific (more conservative) than computational predictions, which may indicate the methods with high IC are assigning terms that are more specific than the evidence supports.

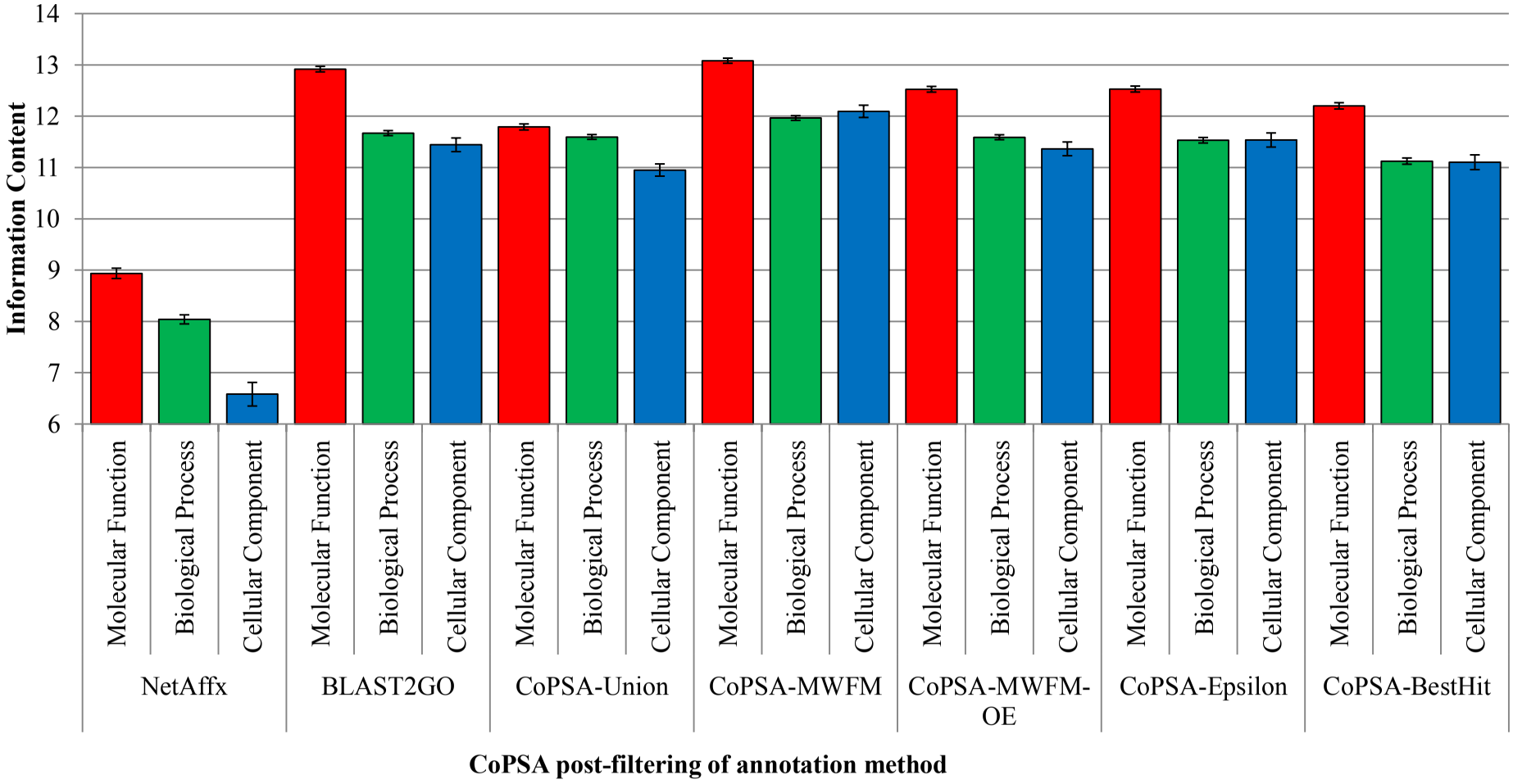


Figure 3.43: Mean Information content of wheat GO annotations from the five filtering strategies in CoPSA, plus NetAffx (AFFY) and BLAST2GO (B2G) pipelines. The error-bars are standard error of the mean.

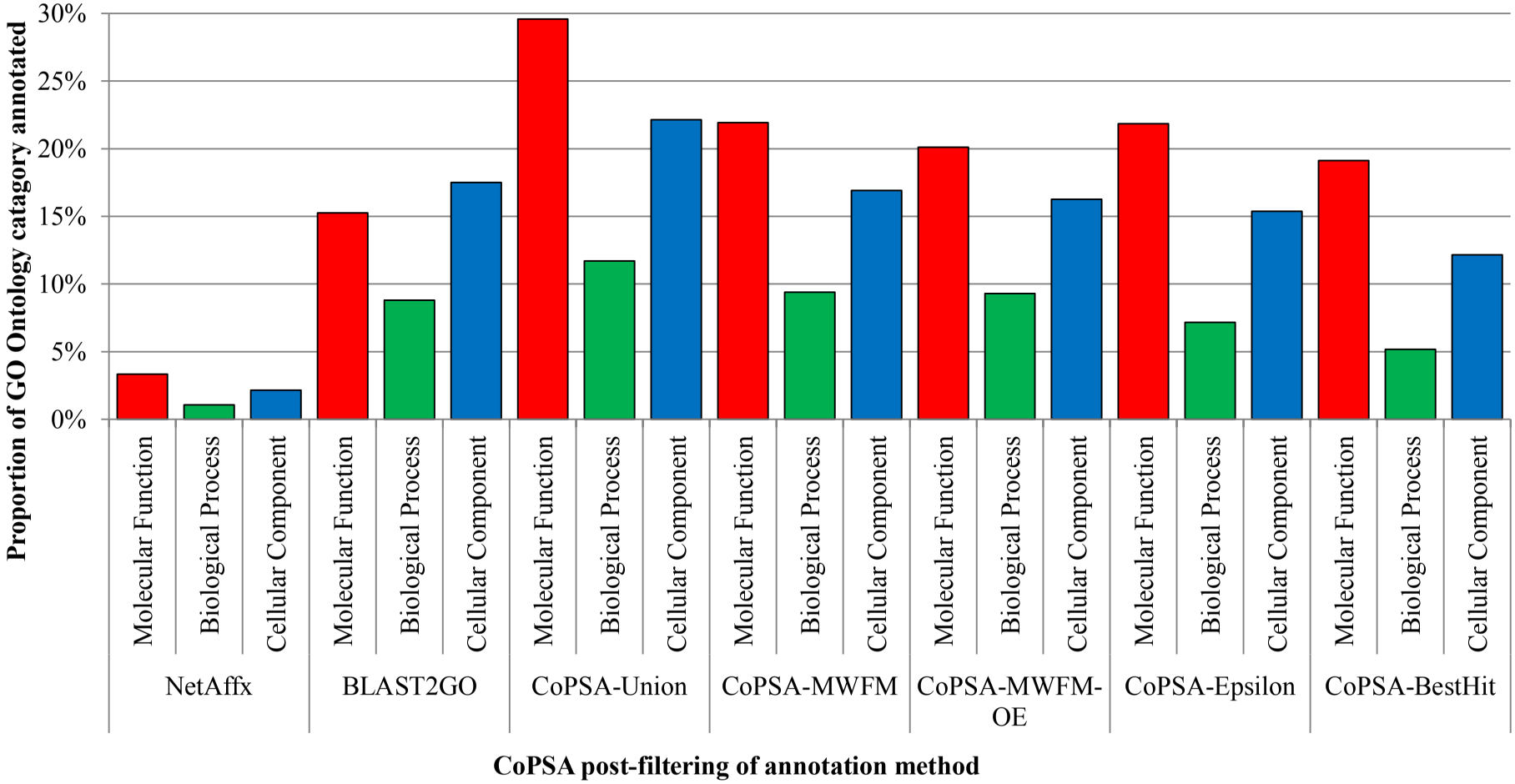


Figure 3.44: The richness of annotation, as indicated by the proportion of the Gene Ontology used to annotate genes in a given methodology.

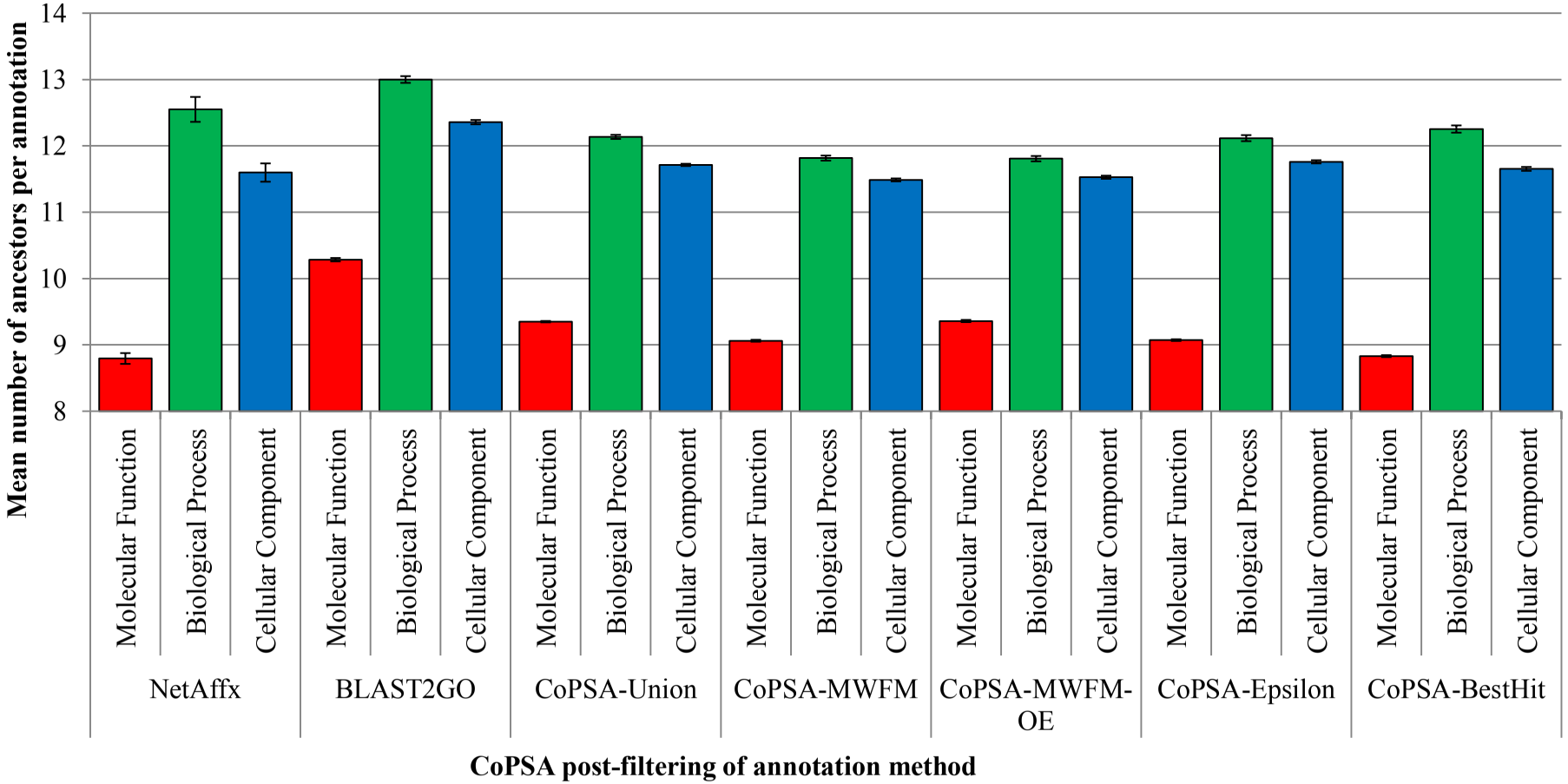


Figure 3.45: Mean specificity of GO term annotations in the five filtering strategies, plus NetAffx (AFFY) and BLAST2GO (B2G) pipelines. Error-bars are standard error of the mean.

3.3.3(*c*) Functional similarity of annotations to NetAffx

A well as examining the properties of the functional annotations produced by each post-filtering method and compared pipeline (Section 4.3.2), a comparison of the functional similarity of each functional annotation output to NetAffx is reported in this Section. As previously explained in Section , a full analysis of precision and recall against a gold standard that has been experimentally validated was desirable, but not possible, given the absence of any suitable gold standard. However, NetAffx was used as the next best substitute for a gold standard annotation set. The validity of this exercise is therefore dependent on the high quality of NetAffx annotations. There is no definitive proof of their quality in the absence of a gold standard annotation. However, based on the evaluation in the previous section, an inspection of the GO annotation for the glutamate synthase family of genes by D Habash, and through visual inspection of the annotation, there are many indications that they represent a conservative, high precision, but low coverage annotation set.

Figure 3.46 shows theVerspoor *et al.* (2006) hierarchical recall metric, calculated against NetAffx GO annotation, on the annotations from each post-filtering method and BLAST2GO. The GO categories are presented individually as the confidence in the annotations from each may differ. If this NetAffx is a reliable annotation set then Figure 3.46 demonstrates a significant endorsement of CoPSA annotations, as it consistently predicts the same or very similar annotation sets to NetAffx. BLAST2GO has consistently lower recall in all categories compared to CoPSA. It shows that for those genes annotated by NetAffx, the predictions of BLAST2GO are less similar to NetAffx than those from CoPSA. This may indicate that BLAST2GO has a higher false negative rate than CoPSA. The annotation selection strategies are designed to improve precision by excluding inaccurate annotations and this often has negative consequences for recall. However, none of the methods show more than a 10% reduction in recall. Out of all the methods CoPSA-MWFM-OE, which is the most stringent,

shows the greatest reduction in recall. This is expected as a more conservative approach to post-filtering is likely to trade a reduction in false positives against an increase in false negatives.

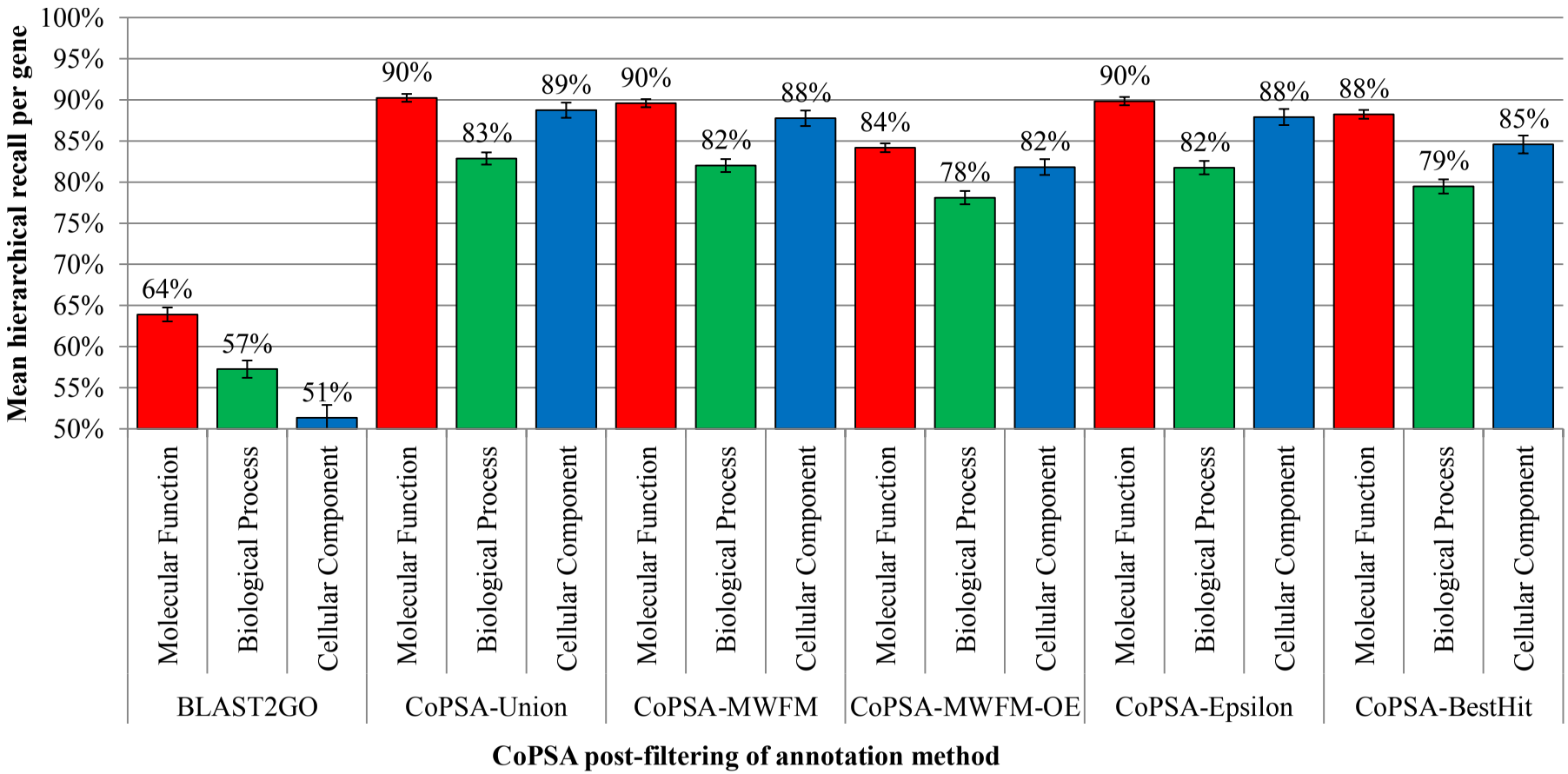


Figure 3.46: Hierarchical recall as defined by Verspoor *et al.* (2006), using NetAffx annotation as an incomplete gold standard. Mean recall is provided for those genes that are annotated within the NetAffx annotation.

3.4 Conclusions

In this chapter, the encoding of biological rules for inference over an Ondex graph has been shown to be an effective strategy for identifying a variety of biological function annotations for the consensus sequences used to design Affymetrix GeneChip arrays. These annotations incorporated elements derived from multiple data sources, alignment methodologies, and structures of annotation.

A conjoint strategy, incorporating both protein functional domain identification (HMMR with Pfam) and protein sequence alignment (BLAST), was shown to improve the quantity, coverage, and diversity of predicted annotations. Overall, BLAST proved the more sensitive method; however a small number of sequences could annotated only through the HMMR based method.

Aggregation of annotations from multiple data-sources was shown to increase both the number of sequences that could be annotated and the range of GO terms used in the annotation. An evaluation of the contribution made by different data sources, UniProt and GOA-*Arabidopsis* were shown to provide the greatest number of annotations, as well as being the richest source of annotations supported by experimental evidence. In addition, the Gramene-derived annotations provided a smaller but still substantial source of annotation. Omitting any of the above data sources from the data integration process would have resulted in a loss of annotation. This supports the view that the large and disparate collection of public biological databases (Cochrane and Galperin, 2009, Galperin, 2007) contain fragmented knowledge and that data aggregation through integration is therefore necessary to utilize this knowledge within systems biology (Molina *et al.*, 2010).

Data integration was also shown to be beneficial in the exploitation of unrealised information, components of which were spread across all the data sources.

This information was identified by the implementation of an inference-based query engine for Ondex which extracted new annotation relationships by encoding generic rules in a query graph. Protein sequence alignment (using BLAST) based annotations were found to be more readily translated by inference rules into GO annotations. A small number of unique GO annotations were also derived from identified Pfam domains (using HMMR). For predicting EC annotations, protein sequence alignment (using BLAST) was also the most effective strategy, however, with regard to the inference of GO annotations, a larger quantity of unique annotations could only be found using Pfam domains. Multi-database inference of annotations yielded less new information, relative to single-database aggregation, for protein sequence alignment based annotation from GO but contributed substantial amounts of unique and supporting data when using Pfam protein domain links to GO. For EC annotation using protein sequence alignment, inference contributed a greater quantity of annotation than data aggregation alone, which translated into a substantial increase in sequence coverage. For EC annotation based on identified Pfam domains, inference was less important for new-annotations, but proved to be important in increasing the specificity quality of annotations.

Compared to the BLAST2GO (Conesa and Götz, 2008) and NetAffx (Liu *et al.*, 2003) annotation pipelines, CoPSA increased the overall coverage of annotations for consensus sequences used the Affymetrix wheat array. CoPSA was able to uniquely annotate 37%, 34% and 24% of the wheat GeneChip sequences with GO categories of *molecular function*, *biological process* and *cellular component* respectively. These sequences had previously been un-annotated in their respective categories by any of the compared providers. There was a large degree of overlap with the other methods. For GO *molecular function* CoPSA was able to provide annotation for 90% and 97% of the sequences annotated with this category by NetAffx and BLAST2GO respectively. For GO *biological process* the overlap was 86% and 88%, and for *cellular component* 90% and 95%. This

confirms that in terms of coverage, CoPSA can provide annotation for more sequences than either of the providers compared. However, it does not address the functional similarity of the annotation provided. Nor does it address the reliability of the annotation. This aspect of the functional annotation provided by CoPSA was evaluated in Section 3.3.3 .

This chapter has concentrated on the improvements that the methods implemented in CoPSA have made to the number and specificity of annotations that could be identified from integrated data resources. This led to an increase in the number of sequences on the wheat Affymetrix array that could be assigned putative biological functions and was shown to be an improvement over existing methods. This result is important for the future analysis of data from the transcriptomics time-course experiment that will be presented in Part II.

CoPSA annotations were found to have a similar level of specificity to the high quality NetAffx annotations, and slightly lower specificity than BLAST2GO annotations. Lower specificity can be an indicator of more conservative annotation, as is the case with NetAffx. While, not conclusive, the similar specificity of CoPSA annotation indicates a pattern of annotation that is more consistent with conservative annotation.

CoPSA annotations are the richest in terms of utilising the breadth of the GO categories, which reflects the higher number of genes on the chip with annotation. It indicates that CoPSA has not simply assignedthe same set of function to more genes, as would be the case when using less stringent criteria. This suggests the greater body of annotation information created using data integration, has increased the ability of CoPSA to assign annotations to many more unknown or partial-characterised wheat genes than any of the standard annotation pipelines.

The annotations for genes using CoPSA had higher Information Content (IC) than NetAffx and similar, but slightly lower, IC when compared to BLAST2GO. It was also observed that the OE adaption of CoPSA-MWFM resulted in lower

IC scores. Together this points to an inverse relationship between IC and how conservative the annotation is. The would mean IC values observed in CoPSA-MWFM-OE therefore represented a middle ground between NetAffx and BLAST2GO, in the compromise between annotating rarer terms, and being more conservative in annotation.

The strongest endorsement of the accuracy of CoPSA predictions, come from the assessment of hierarchical recall as defined by Verspoor *et al.* (2006). This showed that CoPSA made the same or similar predictions for those genes annotated by NetAffx, giving a recall of 84-90% for function, 78-83% for process, and 82-89% for recall for CoPSA-Union compared to CoPSA-MWFM-OE respectively.

Having demonstrated that CoPSA annotations are a significant improvement over other annotation pipelines, the next step is to apply the CoPSA annotations in the analysis and interpretation of a time-course microarray data set.

Chapter 4. A time-series response to water stress in durum wheat

This chapter describes the background and methodology to the biological use case, which provided the primary motivation for development of the CoPSA annotation pipeline described and evaluated in Chapter 3, respectively. This begins by introducing water stress as a critical problem with global impact in agriculture and food production, and goes on to describe the primary molecular control of the drought and osmotic stress in plants as (Section 4.1). The importance and genetics of Durum wheat is then described (Section 4.1.2), with reference to its impact on transcriptome analysis. Finally, Section 4.1.3 describes a controlled environment experiment to study the effect of water stress over time on the transcriptome, which form the bases for the use-case for this thesis. Chapter 6.3 then describes how CoPSA annotations were used to build on this transcriptome analysis.

4.1 The Biological background to water stress

Drought is the major abiotic limitation to crop yield and is thus a serious problem globally with significant agricultural, economic, political and social impact. It is likely to become even more important if current predictions of the effects of climate change prove correct (Neelin *et al.*, 2006). Water stress, however, should not be viewed in isolation from other abiotic and biotic stresses which often occur simultaneously under field conditions.

4.1.1 Water stress in plants: Molecular response

It is usual in the field to observe multiple abiotic stresses acting concurrently on the plant. Significant physical similarities exist between the major abiotic stresses, with osmotic disruption being a common feature of drought, salinity and temperature stress (Zhu, 2002) . Consequently, water stress has the greatest commonality with the other stresses, at not only the physical but molecular scale. There is also significant overlap at the molecular scale, with the same genes, hormones, and signalling pathways being utilised in multiple abiotic stress responses. The overlap of the molecular response is termed cross-talk (Chinnusamy *et al.*, 2004).

An important component of the stress response occurs at the sensing stage. Stress sensors are a major target of study and have so far proved elusive. A candidate osmosensor for immediate osmotic stress signalling is ATHK1, and was identified by Urao *et al.* (1999) in *Arabidopsis*. Similar sequences for ATHK1 have been identified in other plant species (Chefdor *et al.*, 2006, Pareek *et al.*, 2006) ATHK1 encodes a protein with two transmembrane domains, which senses a change in osmotic potential within the cell and can trigger a number of cell signalling pathways, outlined in Figure 4.1.

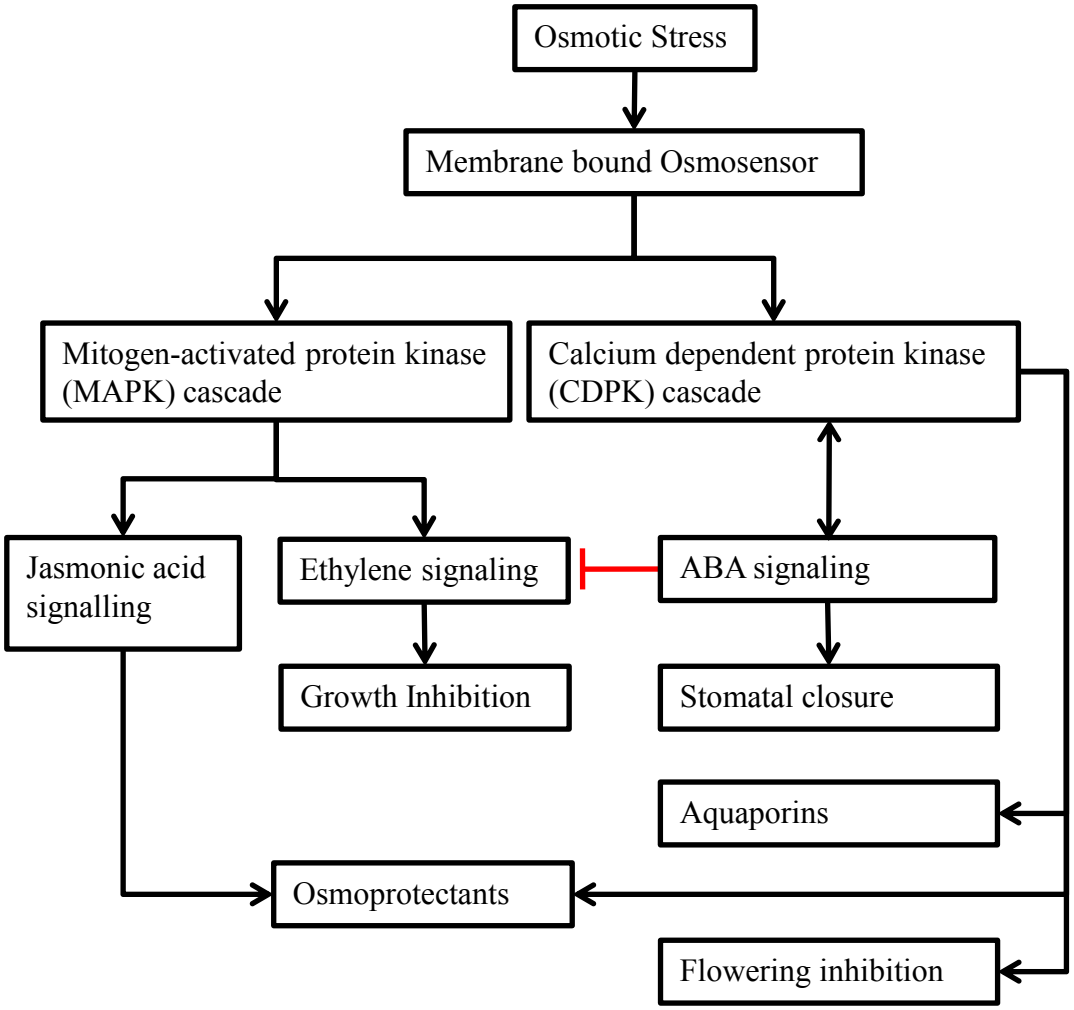


Figure 4.1: The osmotic stress sensing and signalling response. The red connection indicates negative inhibition of a pathway.

ATHK1 triggers a mitogen-activated protein kinase (MAPK) cascade: MAPKKK ⇒ MAPKK ⇒ MAPK (Agrawal *et al.*, 2003). Brader *et al.* (2007) demonstrated that an increase in expression of MPK4 in *Arabidopsis* coincides with an increase in the enzymes involved in jasmonic acid (JA) and ethylene biosynthesis pathways. Gao *et al.* (2004) has shown JA signalling may be involved in the production of the osmoprotectant glycinebetaine.

In addition to those pathways overviewed in Figure 4.1, Urao *et al.* (2000) and Suzuki *et al.* (2001) have identified three phosphorelay (signalling through the transfer of a phosphoryl group (Hoch and Varughese, 2001)) intermediates (ATHP1–3) and four potential phosphorelay response regulators (ATRR1–4). This repres-

ents an independent protein-protein interaction pathway connecting the *Arabidopsis* osmosensor (AtHK1) to gene signalling. A similar mechanism in Poplar was identified by Chefdor *et al.* (2006) and in Rice by Pareek *et al.* (2006).
At the metabolic level carbohydrates, betains, and proline are produced in response to water stress, correcting the immediate loss of osmotic pressure (Beck *et al.*, 2007). Proline has been linked to Auxin and Ca++ accumulation (Sadiqov *et al.*, 2002). Plant hormones play a crucial role in the sensing and signalling of water stress. ABA is an excellent example of the complexity of this signalling and plays a central role in many of the key signalling pathways from sense to response. For this reason in the following section, I examine in detail the role of ABA in water stress signalling.

4.1.1(*a*) The role of ABA in water-stress signalling

The role of ABA in water stress signalling has been known for a long time (Mizrahi *et al.*, 1970, Imber and Tal, 1970), but only recently has the full complexity of its interactions and regulatory mechanisms become known. Studies had suggested ABA was unique among plant hormones in the complexity of its regulation network. However, recent retractions and counter evidences in the field leave this in doubt. This section begins with a description of the regulation of ABA concentration and proceeds to describe its effects.

ABA biosynthesis and related pathways

ABA is highly mobile, and moves from the root to the shoot and surrounding soil. ABA movement from root to the phloem is highly pH-dependent and ABA leaches from the root into the surrounding soil of the plant forming equilibrium with the rhizosphere. This prevents excessive loss from the root during root-shoot signalling (Hartung *et al.*, 1996). The localized concentration of ABA is therefore controlled by its rate of biosynthesis. It is not surprising, there-

fore, that there is tissue-specific localization of enzymes involved in ABA biosynthesis. The AAO paralog group encodes Aldehyde Oxidase enzymes that catalyse the final non-redundant step of ABA biogenesis (Figure 4.2) which is also involved in the production of reactive oxygen species that play a role in ABA signal transduction (Yesbergenova *et al.*, 2005). AAO2 is expressed in the leaf (Seo *et al.*, 2000) whereas AAO4 is localized in the seeds; the other paralogs (AAO 1, 3) are likely to be important in the roots, given they are most highly expressed there (Seo *et al.*, 2000, Sekimoto *et al.*, 1998). Koiwai *et al.* (2004) went further and has identified by GFP-florescence, individual tissues and cells where AAO3 is concentrated. This implies that water stress perception is likely to be heterogeneous across cells and highly tissue specific in its characterization. This poses an interesting challenge to systems biology as significant tissue specific annotation of signalling pathways has not been readily available in public plant databases. However, the AREX databases has begun to make *Arabidopsis* tissue specific expression data available for the root (Brady *et al.*, 2007, Birnbaum *et al.*, 2003). When the viability of a signalling or metabolic pathway is dependent on the tissue proteins are located in, then tissue becomes an important consideration for interpreting gene expression data.

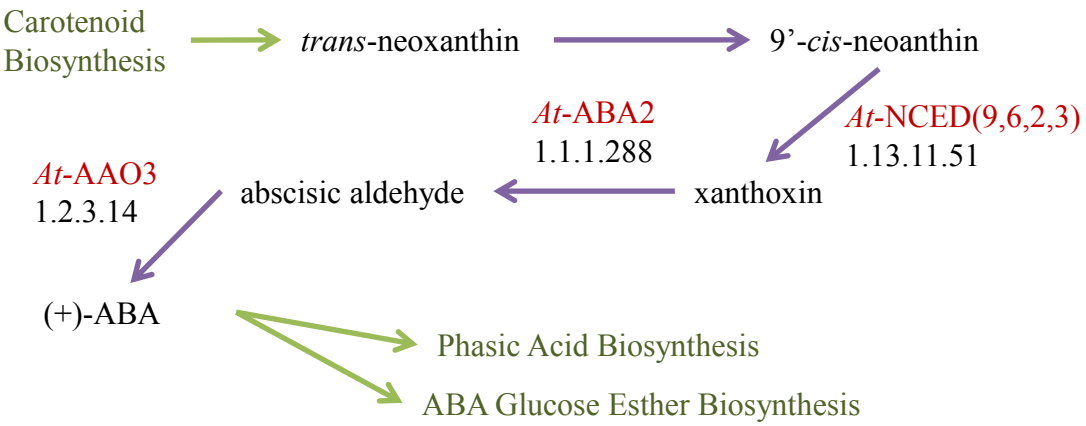


Figure 4.2: The ABA biosynthesis pathway from *Arabidopsis thaliana* (Source AraCyc, December 2011). Enzymes are represented in red, major biochemical products and substrates in black, and reactions as blue arrows.

In addition to the AAO mediated reaction, the enzyme 9-*cis*-epoxycarotenoid

dioxygenase (NCED) appears to catalyse an important reaction in controlling ABA biosynthesis (Figure 4.2). Iuchi *et al.* (2001) have observed a significant correlation between the expression of AtNCED3 with ABA levels and ABA mediated responses. Yang and Guo (2007) have identified the same response in leaves and roots for a NCED gene in Stylosanthes guianensi and the same has been shown in the leafs of *Solanum lycopersicum* (Thompson *et al.*, 2000). However, it is thought that *in planta* AAO is the prime regulatory enzyme (Nambara and Marion-Poll, 2005).

The molybdenum cofactor is involved in the final reaction of ABA biosynthesis and is synthesized in the Molybdenum Cofactor biosynthesis pathway shown in Figure 4.3. In addition to the AAO and NCED catalysed reactions of ABA biosynthesis the gene ABA3 which encodes the molybdenum-cofactor-sulpherase enzyme (catalysing the final stage of the pathway shown in Figure 4.3) has been observed to be over expressed in drought-stressed plants. The ZEP gene that codes for zeaxanthin epoxidase, which catalyses the first stage of ABA biosynthesis is also up regulated during water stress. The ABA2 gene (Figure 4.2) encoding the xanthoxin dehydrogenase enzyme however does not appear to be regulated during water stress (Nambara and Marion-Poll, 2005).

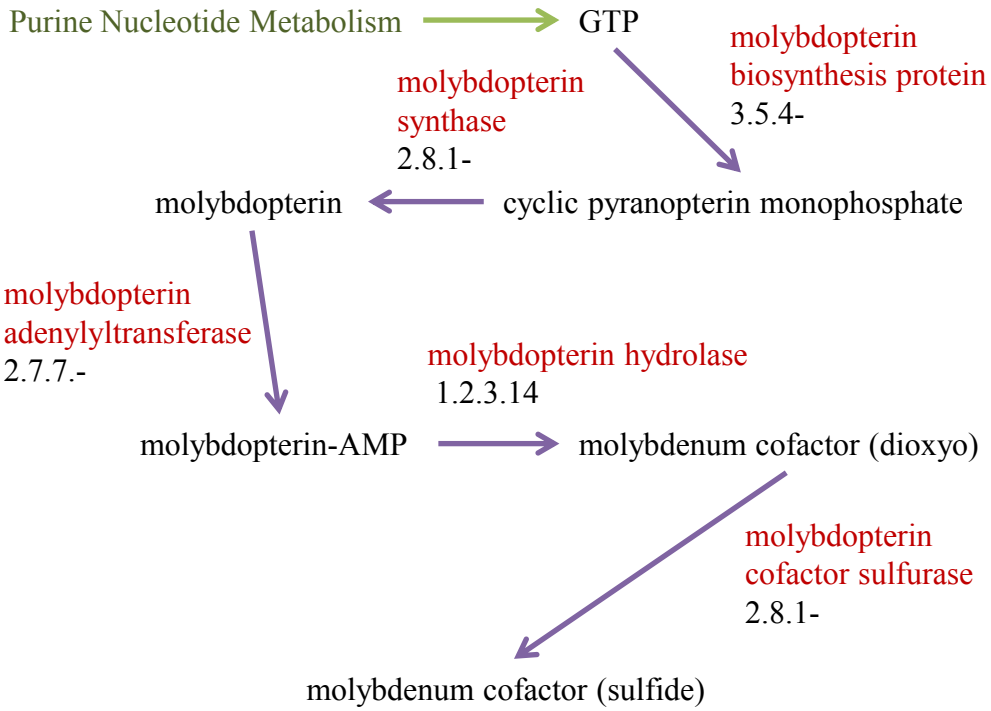


Figure 4.3: The molybdenum cofactor biosynthesis pathway from *Arabidopsis thaliana*, molybdenum cofactor is involved in the final stage of ABA biosynthesis (Figure 4.2) (Source AraCyc, December 2011). Enzymes are represented in red, major biochemical products and substrates in black, and reactions as blue arrows.

ABA catabolism is also important in water stress, as all 4 CYP707A genes in the downstream phaseic acid biosynthesis pathway (Figure 4.4) are up regulated by water stress (Kushiro *et al.*, 2004, Saito *et al.*, 2004), ABA concentration therefore seems to be co-regulated by its biosynthesis and catabolism. Interestingly, this seems to be a central node for plant hormone regulation, as CYP707A3 is positively regulated by both gibberellin and brassinolids (Saito *et al.*, 2004).

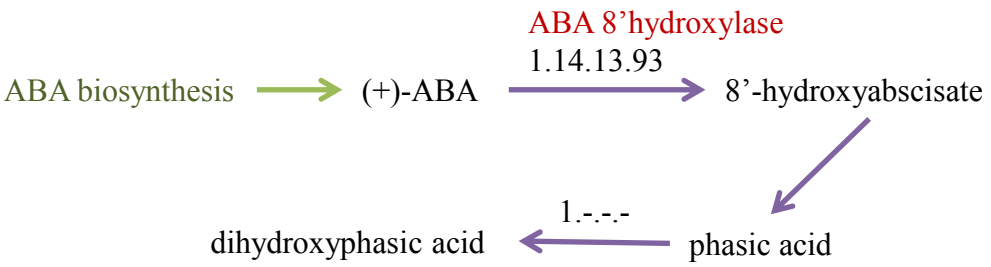


Figure 4.4: The phaseic acid biosynthesis pathway from *Arabidopsis thaliana* (Source AraCyc, December 2011)

As with ABA concentration, localization plays an important role in ABA signalling and significant work has been done to characterize the movement of ABA from its site of production to its site of action. It has long been thought that ABA plays a crucial role in the root to shoot signalling in water stress through xylem transport of ABA Zhang and Davies (1987), Davies and Zhang (1991), Sauter *et al.* (2001), Wilkinson and Davies (2002), Jiang and Hartung (2008). However, Christmann *et al.* (2007), in response to the observation by Holbrook *et al.* (2002) that the shoot drought ABA response can occur independently of a root ABA signal, has recently presented evidence that a hydraulic signal from root to shoot is far more important. This would indicate a model where ABA is synthesized in its locality as required in response to other signalling mechanisms. This seems consistent with the previously discussed work by Koiwai *et al.* (2004) where elevated ABA concentrations are shown to be cell and tissue specific, which requires either an efficient transport mechanism or localized production.

A key component to understanding ABA signalling activity is the identification of ABA receptor sites. In the last two years, considerable progress has been made towards identifying these sites of action, of which at least three have been identified. The *Vicia faba* (Broad bean) chloroplast protein Mg chelatase H subunit encoded by ABAR/CHLH was identified as a candidate ABA signalling receptor for stomatal closure, and ABA binding with a homolog in *Arabidopsis* has been observed (Shen *et al.*, 2006). Shang *et al.* (2010) have also shown ABAR interacting with WRKY transcription factor. However Müller and Hansson (2009) have failed to show any effect on ABA of barley from ABAR mutants. Liu *et al.* (2007) has also observed GCR2, a G protein–coupled receptor membrane protein binding to ABA. However recent studies have failed to show any loss in ABA sensitivity in GCR2 knockouts (Gao *et al.*, 2004, 2007). More recently, Risk *et al.* (2009) have show that GCR2 does not bind with ABA, cite

problems in protein purity and data-analysis in Liu *et al.* (2007) work.

Two promising new candidate families of proteins involved in ABA binding have recently emerged; the GPCR-type G proteins (GTG) and the regulatory components of ABA receptor (RCAR) proteins. Pandey *et al.* (2009) have identified GTG1 and GTG2 in *Arabidopsis*. They showed these transmembrane G-protein-coupled-type proteins were shown to bind to ABA, and double knock out mutants induced ABA hypersensitivity.

The RCAR family are the most promising candidates for ABA binding receptors to-date, and consist of 14 genes in *Arabidopsis*. RCAR1 was first identified by Ma *et al.* (2009), who demonstrated a physical interaction with ABA using three different methods: yeast-two-hybrid (Y2H), bimolecular florescence complementation (BiFC), and confocal microscopy of a green fluorescent protein (GFP) labelled RCAR1 protein. Ma *et al.* (2009) also showed multiple-RCAR knockouts were ABA insensitive. An ABA-RCAR complex has been observed with X-ray diffraction (Santiago, Rodrigues, Saez, Rubio, Antoni, Dupeux, Park, Márquez, Cutler and Rodriguez, 2009, Nishimura *et al.*, 2009), as well as a three-way complex (ABA-RCAR-PP2C) with Gα and phosphatase 2C (PP2C) (Yin *et al.*, 2009, Miyazono *et al.*, 2009). Melcher *et al.* (2009) have described a ligand-binding pocket in RCAR, which is flanked by a β-loop gate that closes upon ABA binding. This affects conformational changes which allows a PP2C protein to bind into an otherwise competitively inhibited active site. Santiago, Dupeux, Round, Antoni, Park, Jamin, Cutler, Rodriguez and Marquez (2009) observed this complex localised to the nucleus and the cytosol using GFP labelling. Figure 5.21 shows a schematic by Raghavendra *et al.* (2010) describing how this ABA sensing complex, fits in with the known ABA responsive signalling pathways. The ABA-RCAR-PP2C complex formation appears to prevent the inhibition of sucrose non-fermenting-1 (SNF1)-related protein kinases (SnRKs). PP2C have previously been implicated in SNF1 related kinase (SnRK) signalling. A number of studies have linked various SnRKs with stomatal aperture size (Li

et al., 2000, Mustilli *et al.*, 2002, Yoshida *et al.*, 2002, Fujii *et al.*, 2007). The gene OST1 in rice which encodes SnRK2 protein in rice, and is strongly correlated to stomatal aperture, has also been shown to be dependent on the functioning of the ABI1 gene encoding the PP2C (Yoshida *et al.*, 2002, Mustilli *et al.*, 2002). This confirms that this mechanism is conserved in other plant species, as well as *Arabidopsis*.

The first mode of action of the ABA-RCAR-PP2C signalling complex is shown in Figure 5.21(a). The competitive inhibition by the RCAR complex of the inhibitory action of PP2Cs on OST1, frees up OST1 to initiate stomatal closure. The action of OST1 on stomatal closure was elucidated by Mustilli *et al.* (2002). More recently it has been shown that in the absence of ABI1, OST1 phosphorylates, and consequently activates SLAC1 anion channel (Geiger *et al.*, 2009, Lee *et al.*, 2009). Ultimately, this activation results in the depolarisation of the guard cell, which leads to turgor loss through osmosis and consequent stomatal closure (Negi *et al.*, 2008, Vahisalu *et al.*, 2008). Sato *et al.* (2009) have also shown that OST1 phosphorylates the C-terminal region of the K^+ channel protein KAT1, which prevents the passage of K^+, which would otherwise repolarise the guard cell. Siegel *et al.* (2009) suggests that this, ABA-mediated signalling of stomatal closure, is co-regulated by Ca^{++} mediated signalling, which is multiplicative with the activity of OST1. They showed the action of ABA in stomatal closures is approximately 30% of normal, in the absence of elevated Ca^{++}. The presence of Ca^{++} inhibits K^+ channels, and activates slow anion channels such as SLAC1. The action of dependent phosphatase kinase (CDPK) links ABA with Ca^{++}, and connects to a wider signalling transcriptional regulation pathways. The gene ABI1 which encodes a PP2C has previously been shown by Leung *et al.* (1994) and Mori *et al.* (2006) to link ABA and Ca^{++} signalling, to affect stomatal closure. Ca^{++} activates CDPKs by binding to a calmodulin-like regulatory domain (Harmon *et al.*, 2000).

The second mode of action of the ABA-RCAR-PP2C signalling complex, in the

nucleus, is shown in Figure 5.21(b). In the absence of PP2C inactivation (by RCAR complex competitive inhibition) the OST1 protein phosphorylates ABA-responsive Element Binding Factor (ABFs), which bind to the ABA responsive prompter (ABRE) (Fujita *et al.*, 2009, Yoshida *et al.*, 2010). A key target for Ca^{++} signalling is also the AREB transcription factor domain. The CDPK AtCPK32 has been shown to regulate a family of transcription factors sharing the AREB domain, these are known to heighten sensitivity to ABA, and phosphorylate ABI5 (Choi *et al.*, 2005).

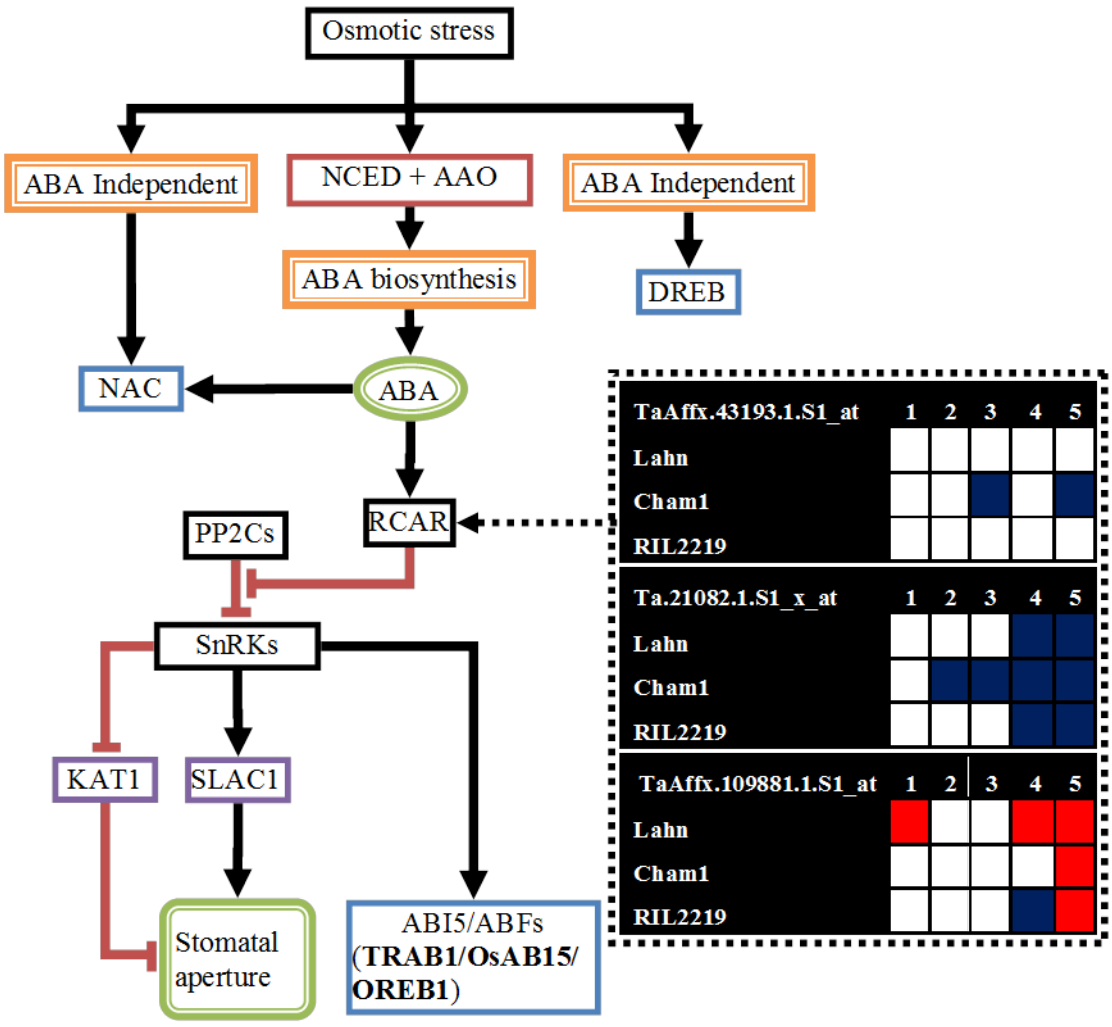


Figure 4.5: The RCAR, ABI1 (a PP2C), ABA complex (pink), and its role in cystosolic and nuclear sensing and signalling of ABA in the cytosol (a) and the nucleus (b) (Raghavendra *et al.*, 2010).

Downstream secondary messengers affected by ABA signalling form a complex network of cascades and regulations. In Figure 5.6 Hirayama and Shinozaki (2007) helpfully summarize the major proposed signalling pathways. They speculated even before the ABA-RCAR-PP2C complex was known, that PP2C was a likely control hub within ABA signalling.

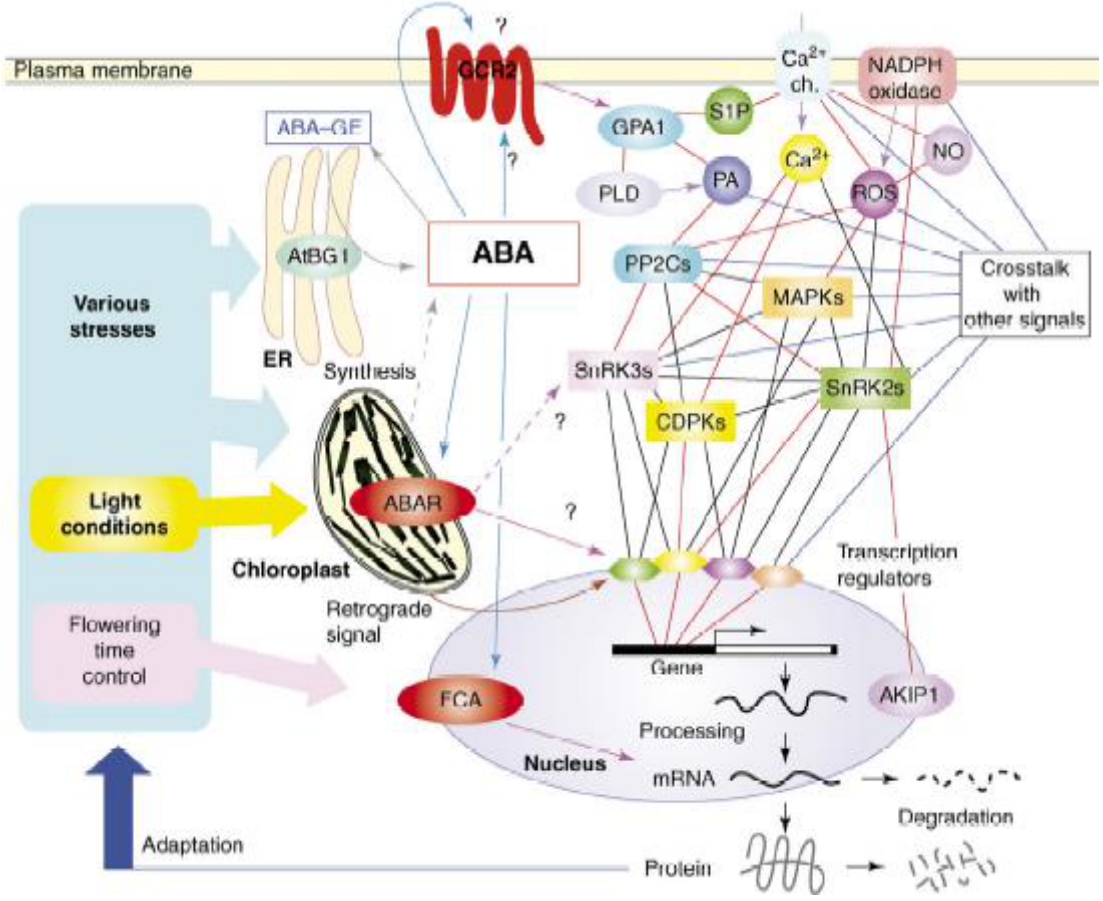


Figure 4.6: A summary of ABA downstream signalling pathways highlighting the three main ABA receptors (Hirayama and Shinozaki, 2007). Note: The work describing FCA as an ABA receptor has now been retracted (Razem *et al.*, 2006), as a result of work by Risk *et al.* (2008).

The role of CDPKs in aquaporin activity has been recognized since Maurel *et al.* (1995). Mariaux *et al.* (1998) and Jang *et al.* (2004) have both observed Plasma The plasma membrane intrinsic proteins (PIP) are aquaporin channel proteins, which are encoded by 13 known genes in *Arabidopsis thaliana*. The family has been found to contain both ABA-dependent and independent genes. PIP gene expression in response to ABA seems to be tissue and stress-specific;

intriguingly in roots, Jang *et al.* (2004) showed significant reductions in some PIP genes in response to ABA.

The MAPKs cascade pathway that is independent of ABA, briefly discussed in the introduction to this section, can also be triggered by higher ABA concentrations. It is not known whether this is a direct interaction, or as a result of indirect reactive oxygen species accumulation which also triggers a MAPK cascade (Zhang *et al.*, 2006).

Any useful representation of water stress response must encompass responses at the level of signal sensing, regulatory cascades, transcription factor signalling, and gene and metabolic pathway activation: to present an accurate system representation. In addition to the immediate propagation through the transduction scales described in Figure 4.7, signalling feedback and cross pathway inhibitions make the drought response system highly complex.

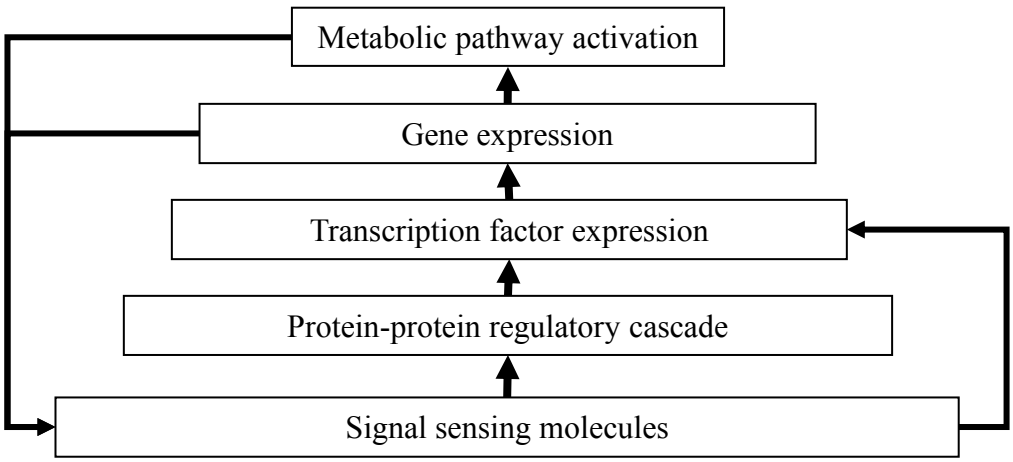


Figure 4.7: The signal cascade encompasses multiple transduction scales. Signal sensing molecules and hormones such as ABA and ATHK1 trigger protein-protein cascades and TF expression. TF expression influences the expression of a gene or group of genes. Genes encode signal sensing protein increase the severity of response and genes encoding enzymes alter the abundance of plant hormones.

Another important aspect in our understanding of the plant's response to water stress is the time frame for the molecular and physiological responses. Signalling processes and physiological responses do not occur in isolation, but

evolve and interact over time. For example: an early transcriptional response during mild stress may have consequences for later signalling responses to severe stress. An isolated snapshot of the transcriptome in time may therefore be inadequate to understand the plant's response. A real-time monitoring is often impractical because of cost implications; however a pragmatic interval for monitoring transcriptome responses can be selected which makes a balanced compromise between the cost of measurement, the frequency of gene changes, and the length of the water stress response.

4.1.2 The Genetics of durum wheat

Triticum turgidum subsp. Durum (durum wheat) is a tetraploid species of wheat that is widely grown as a crop and used in the production of pasta and bread. Durum wheat comprises approximately 8% of the worldwide wheat production. Taxonomically, it is a hybrid between a wild grass and primitive diploid wheat, and forms part of the hybrid with the grass T. tauschii that formed the common ancestor for the hexaploid bread wheat, *Triticum aestivum* (Nevo *et al.*, 2003). Figure 4.8 shows the genetic contribution Durum wheat has made to the current wheat species. One of the significant problems for researchers wanting to work on Durum wheat, is that there is little genome sequence data available in public sequence databases, with GenBank (Benson *et al.*, 2007) containing only 19,641 ESTs (October 2010). The closest related wheat species where significant amounts of sequence data are available is hexaploid bread wheat, which contains an ancestral durum genome. Bread wheat has an estimated genome size of 13.5Gb which is huge when compared to the model plant *Arabidopsis thaliana* genome size of 157Mb. At the time of writing, the number of bread wheat sequences in the public databases was just over 1 million EST sequences, and 1,830 fully sequenced genes within GenBank (October 2010). In Entrez,

wheat has 41,000 UniGene records which are clusters of sequences that are believed to originate from the same transcription locus, based on protein similarities, cDNA alignment, and genomic location data. This can be contrasted with *Arabidopsis* sequences in the version 10 release of TAIR which had 27,416 protein coding genes, 4827 pseudo-genes and transposable elements, and 1359 non coding RNAs. A recent preliminary study which sequenced 18.2 Mb of the wheat 3b chromosome found 175 gene and pseudo-gene models and from this they calculated a gene density of 1 gene per 104 kb, which led to a high estimate of 50,000 genes per diploid genome (Choulet *et al.*, 2010). The International Wheat Genome Sequencing Consortium (IWGSC) (Gill *et al.*, 2004) are currently in the process of building a physical map of the wheat genome, and has begun full sequencing of the 3b chromosome as a pilot, using Roche 454 next generation sequencing of BAC clones.

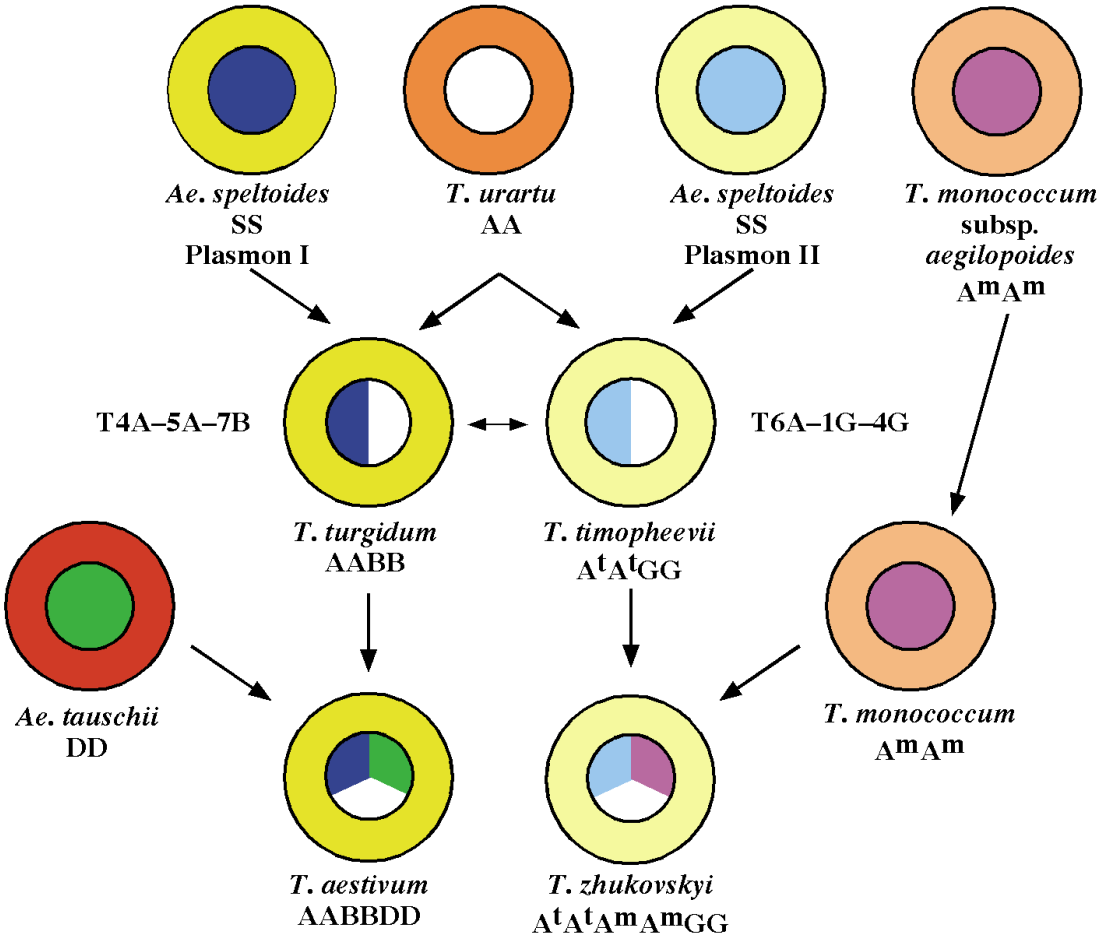


Figure 4.8: The current theory on the genetic history of modern wheat, showing the inheritance of genomes and the origin of polyploidy in modern wheat (Genetic and Center, 2011).

Durum wheat is tetraploid and consequently, ancestral genes are duplicated across A and B genomes as a homoeologue pair. If the sequences of a given homoeologue pair have not diverged in the regions targeted by the probes on the Affymetrix chip, and both genes are active in the transcriptome, then gene expression will be the product of multiple genomic loci. Redundant genes are sometimes eliminated or epigenetically silenced during the formation of the polyploid in order to stabilise a viable gene expression profile (Chen, 2007). Therefore, in some instances, only one gene from a single genome is active in the transcriptome. However it has been observed that for a large number of homoeologous genes in bread wheat, with single nucleotide polymorphism (SNP) differences that allow the homeologues to be identified, all three genes

are expressed (Mochida *et al.*, 2003). Where coding genes have 100% identity across genomes, then the proteins they encode are likely to be identical, and so in terms of function and process they are indistinguishable. However, where small polymorphisms between copies of a gene occur, and the microarray probe-sets are not able to differentiate, then part of the reported transcriptome profile of a gene may be in error. This aspect can be examined further when one needs to explore particular gene families with subsequent qPCR and other methodology and is beyond the scope of this thesis.

The issue of gene duplication is not an issue restricted to polyploidy. Gene duplications may occur within a genome from transposable elements, which are particularly active in wheat (Choulet *et al.*, 2010, Sabot *et al.*, 2005) and this result in paralogous genes. The complexity of genome duplication and polyploidy create problems for the interpretation of transcriptome data and these are compounded by incomplete knowledge of wheat gene sequences. The most important problem is that many microarray probe-sets record expressed transcripts of unknown genes or gene variants. It is imperative, therefore, that observed expression results from the microarray be experimentally validated using complementary methods such as qualitative real-time PCR (qRT-PCR).

A similar issue arises in the use of microarrays to compare individuals or genetic lines. Sequence polymorphism in genes varieties and germlines, can also result in variations in probe-sequence binding affinity, and consequently affect the reported expression. This could potentially be a source of error in comparing samples. However a study using the Affymetrix wheat microarray, which looked at 15 probe-sets that reported large differential expression across cultivar lines, confirmed using qRT-PCR that the differences were due to real expression (Wan *et al.*, 2009).

4.1.3 A controlled-environment time-series microarray experiment

This use case forms the motivation for developing improved functional annotation of microarray sequence targets for the analysis of a durum wheat water stress-response experiment. The data comes from a time-series microarray experiment designed to study water stress in a controlled environment for three cultivars of durum wheat. The experiment was conducted as part of the TRITIMED project at Rothamsted Research to understand gene transcript responses and identify candidate genes. It is included here as background to understanding the data set, for which an example analysis will be presented in Chapter 6.3 that demonstrates the utility of CoPSA annotations. The transcriptome responses were studied from three cultivars of durum wheat: Cham1, Lahn and RIL2219 (Figure 4.9). These cultivars were developed by the wheat breeder Miloudi Nachit (ICARDA, Syria). Cham1 is a drought-resistant variety widely grown in the Mediterranean basin and Lahn is a high yielding variety from Syria, which performs well under well-watered conditions but is susceptible to changes in temperature and water availability. Recombinant Inbred Line (RIL) 2219 is one of the 114 RILs that originating from the cross between Cham1 and Lahn. This RIL was identified as one of the more drought tolerant lines in terms of stability of yield when compared to either parent using multi-site field trials (15 sites) conducted over two years. The two parents and RIL2219 were subsequently studied for their responses to water stress imposed under controlled environment conditions.

The optimum conditions to produce severe water stress over 5 days where selected in preliminary studies over a period of two years prior to the controlled-environment experiment. The three cultivars were grown in 3.5L pots under controlled environment conditions of:

- Temperatures of 25°C during the days and 20°C during the nights
- A photoperiod of 14 hours at a 700-900 μmol^{-2}s^{-1} photon flux density
- 65% relative humidity
- A peat-free soil-compost enriched with slow-realise fertilizer (Osmocote, N:15%, P2O5: 11%, K2O:13%, MgO:2%, Bo:0.02%, Mo: 0.02%, Cu: 0.05%, Mn0.06%, Mn:0.06%, Zn: 0.0015%, Fe chelated: 0.15%)
- Hand watered

A random block design was used with four biological replicates per cultivar, treatment and time point. Plants from the different lines had similar developmental rates and plant heights; an important set of considerations to enable the study of stress without the complexity of added developmental-induced stress responses (Figure 4.9)

The drought stress was induced at one week post-anthesis, and the plants were deprived of water over five days. Samples, from 3 biological replicates, and from 3 repeated independent experiments, were taken each day from the total flag leaf and the RNA extracted. Total RNA was extracted using TRIZOL Reagent (Invitrogen) according to manufacturer recommendations. Total RNA was DNase treated with TURBO DNase enzyme (Ambion) and cleaned using RNeasy columns (Qiagen). RNA concentration was measured in a Nanodrop spectrophotometer and first strand cDNA was synthesised using SuperScript III enzyme (Invitrogen). Best quality RNA was applied to Affymetrix wheat gene expression arrays at Bristol University.

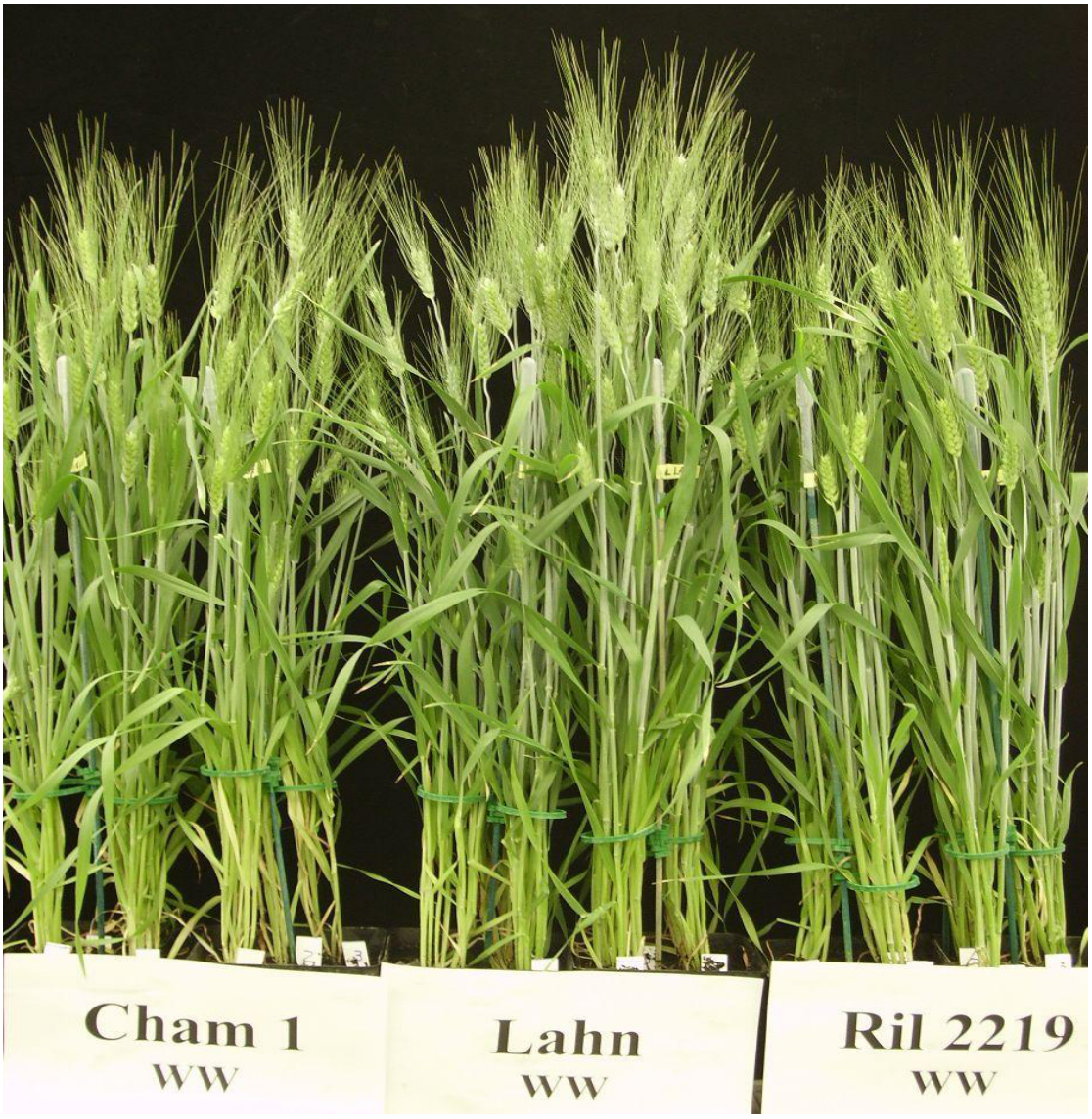


Figure 4.9: Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIL2219 (drought tolerant line) under well watered conditions in the TRITIMED experiment. (Source: Marcela Baudo).

Figure 4.10 shows the normalised fluorescence intensity from the probes on the wheat Affymetrix GeneChip array hybridised to RNA from the five time points using the GeneSpring microarray analysis tool kit. It clearly shows the complexity and variation within the data and the need for further dissection of genes. The statistical and bioinformatics analysis of this complex data set is described in Chapter 6.3, which incorporates the improved CoPSA annotations described in Chapters 3.

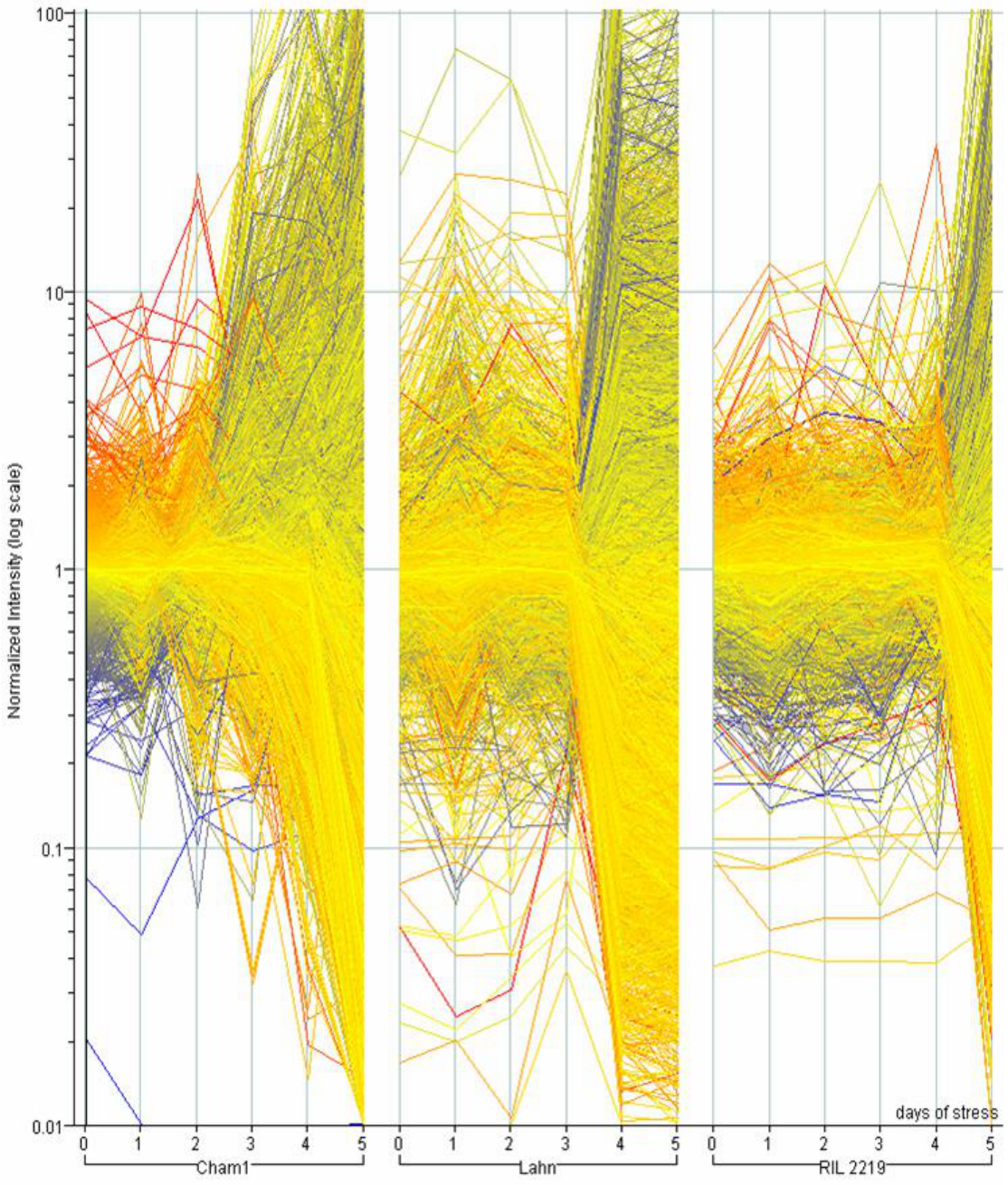


Figure 4.10: An overview of the expression of all significantly expressed genes (p>0.05) in the three time courses, for the three cultivars: Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line). (Source: Marcela Baudo, created in GeneSpring). The colour of each plotted genes relates to the colour chart and represents log of the expression relative the control.

Chapter 5. Utilizing CoPSA annotations for microarray analysis

This chapter describes how the Affymetrix wheat chip annotations, derived from CoPSA, were applied to the analysis of the microarray experiment described in Chapter 6.1. It is intended as a demonstration of how CoPSA annotation, presented in Chapter 3, can be used to facilitate the analysis of a transcriptomics experiment. This chapter also describes statistical analysis conducted by members of the TRITIMED project. The expression data were subjected to two-way ANalysis Of VAriance (ANOVA) and Principal COordinates (PCO) analysis, which were both completed by Stephen Powers. These statistical analyses were undertaken to identify subsets of the genes represented on the microarray that were changed significantly during the experiment. The influence of water stress in terms of Relative Water Content (RWC) required a transformation of the data, which was completed by Michael Defoin-Platel. The rest of the analysis was undertaken by myself.

5.1 Aims and Objectives

The aims of this chapter are to:

- Demonstrate the utility of an improved functional annotation set in the analysis of a transcriptome.
- Enrich statistical data-analysis with functional descriptions of gene-sets.

- Provide summaries of the functions and processes expressed during the experiment.

- Demonstrate, on a number of example processes, how functional annotation can be leveraged to drill down to individual pathways.

This will be achieved through:

- A summary of the dataset with respect to the significantly expressed ANOVA groups, and significantly enriched functions and processes within these.

- Describing the variation within the dataset with respect to Principal Coordinates (PCo), and the significantly enriched *molecular functions* and *biological processes* which contribute to these.

- Leveraging transcription-factor family annotation to summarise transcriptional control.

- Providing a summary of the highest level processes expressed at each time point in response to water stress.

- Drilling down to a more detailed view of genes connected to the ABA biosynthesis and RCAR signalling pathways. This is shown through an example analysis of ABA related genes, facilitated by improved functional annotation, from CoPSA.

5.2 Introduction

This introduction describes the statistical analysis undertaken for the TRITIMED project. Later within this chapter, the analysis was combined with the improved functional annotations from the CoPSA pipeline (Chapter 3) to demonstrate its utility.

5.2.1 Two-way ANOVA

From the TRITIMED controlled environment (CE) experiment, 63 samples were assayed, using Affymetrics technology, for gene expression. The samples comprised 6 replicates of well- watered (WW) and 3 replicates of stress (S) at 1-5 days for each of 3 wheat lines. The three wheat lines are composed of Lahn, a high yielding but drought susceptible line, Cham1, a drought resistant line, and RIL2219, a highly drought resistant line created from Lahn and Cham1. The gene expression data was filtered to exclude non-expressing genes (those with between 0.7 and 1.4 fold change in expression) and for rogue observations on the chip. After this initial processing there were 19,062 genes left for ANOVA analysis. The gene expression was first transformed into the log_2 scale in order to normalise the data. The test assesses the significance for the main effects, cultivar and time, and the interaction between these factors. The ANOVA was also repeated for the factors cultivar and Relative Water Content (RWC), after a *normalisation* with respect to RWC procedure described later in this section. The F-test, a variance ratio test, is used in ANOVA to measure the significance of the main effects. This significance of main effects for a given gene was tested by calculating the variance from the difference in lines across treatments and the residual variance. The F-statistic for a gene indicates significance if it is suf-

ficiently large to appear in the tail of the F-distribution, such that the probability of obtaining that F-statistic by chance is 0.05. Normal distribution in gene expression was checked by considering plots of these residuals. Tests to confirm normal distribution (an key assumption of ANOVA) were not applied, however visual inspection of residual plots supported the general robustness of the ANOVA.

The product of ANOVA was four non-redundant sets of genes that can be categorised as having significance with respect to time (2,877 genes), line (782 genes), line+time (4,621 genes), and line.time (10,363 genes). The final group records genes with an interaction between the two independent variables (factors). ANOVA also provides a Least Significant Difference (LSD) for each ANOVA comparison of means; these can be used to determine the level of difference between means required for significance. This is used within this study to test which times points in the line.time interaction group contain significant expression relative to the well watered control.

5.3 Principal COordinates analysis (PCO)

Principal Coordinates Analysis (PCO) is a methodology for exploring similarities and differences in multivariate data, which was developed by Gower (1966). A key advantage of PCO is that it does not assume a distribution, and is less computationally intensive than approaches such as Canonical Variates Analysis (CVA). PCO analyses variance for each sample in a similarity data matrix (genes were columns and samples rows), and allows the visual identification of differences due to the treatment combinations (lines, by days of stress, or RWC). A full similarity matrix was created, comprising the Euclidean distance between all replicates given the 19,062 genes measured on the GeneChip. A reduced matrix of mean similarities were derived from this original matrix per combin-

ation of line, and days of stress, or RWC, for each respective analysis. The PCO is then applied to this reduced matrix. The similarity between samples is based on the Euclidean distance between units.. Units are calculated, using Equation 5.1, for each gene k in 19,062 space, where the i and j are sample in each pair. x_{ki} is therefore the value of gene k in sample i and r_k is the range gene k. The similarity between two samples i and j is the Euclidean distance between a pair of units, which is given in Equation 5.2.

$$S_k = 1 - \{(x_{ki} - x_{kj})/r_k\}^2 \qquad \text{(Equation 5.1)}$$

$$S_{ij} = \sum_k \{S_k(x_{ki}, x_{kj})\} \qquad \text{(Equation 5.2)}$$

The result is a series of Principal Coordinates (PCos) that are ordered by the amount of variance in the data they encompass, with the first three PCos usually capturing the majority of the variance of the data. The first three PCos on the reduced (averaged across replicates) line.time similarity-matrix captured 52.84%, 9.22%, and 6.80% of variance, respectively. For the line.RWC reduced (averaged across replicates) similarity-matrix, the first three PCos captured 47.84%, 15.01%, and 9.95% of variance respectively. PCos are one-dimensional coordinates, which can be related back to the original data columns by linear regression analysis. Simple linear regressions of the principal coordinate scores on the corresponding gene expression data are used to derive F statistics. Equation 5.3 shows the linear regression model used to derive F-statistic, where PCo is the coordinate, i is the PCo index (*i.e.* for the PCos referenced in this chapter $i \in \{1..3\}$), j is the variable combination (*i.e.* one of 18 line.time or line.RWC combinations), and GE is the matrix of mean (over replicates) gene expression values. The largest F-statistic indicates the gene that correlates most strongly with the PCo score.

$$PCo_{ij} = \alpha + \beta \times GE_{kj} \qquad \text{(Equation 5.3)}$$

In this way, the means of replicates for each treatment combination were plotted for each PCo. The F-statistics from the correlation relating each PCo to the expression values of each gene were collated and sorted largest to smallest, for each PCo. The F-statistic is therefore inversely related to the probability that the expression gene of the gene contributes to that PCO. Genes were categorised as of primary importance to a PCo if they made up 5% of the sum of the F-statistics for that PCo.

5.4 Analysing stress in terms of Relative Water Content (RWC)

A key motivation of the experiment was to test the hypothesis that a transcriptome response of a cultivar to water stress affected its ability to maintain yield under drought. Results showed that RIL2219 had a delayed physiological response (leaf RWC) to water stress when compared to both parents after day three (Figure 5.1). This introduced an added complexity to the analysis of the data and necessitated the *normalisation* of gene expression data sets to allow for the exploration of differences in gene expression for all lines at the same level of water stress (RWC) measured in the leaves.

The sigmoid function shown in Equation 5.1 was used to determine RWC at a given time point t. The values a, b, c, and d were determined using a quasi-Newton gradient descent algorithm. However, given that the RWC in the drought resistant RIL2219 never fell below 70% the final two time points (day 4 and 5) Lahn and Cham1 were not comparable with RIL2219. Therefore, in order to find a reasonable solution, a projected data point of 30% RWC was added at day 10, based on an assumption of continuing decline in RWC. This value was projected based on consultation with the domain expert, D Habash.

$$irwc(t) = \frac{1}{1 + e^{\frac{t-c}{d}}}a + b \qquad \text{(Equation 5.1)}$$

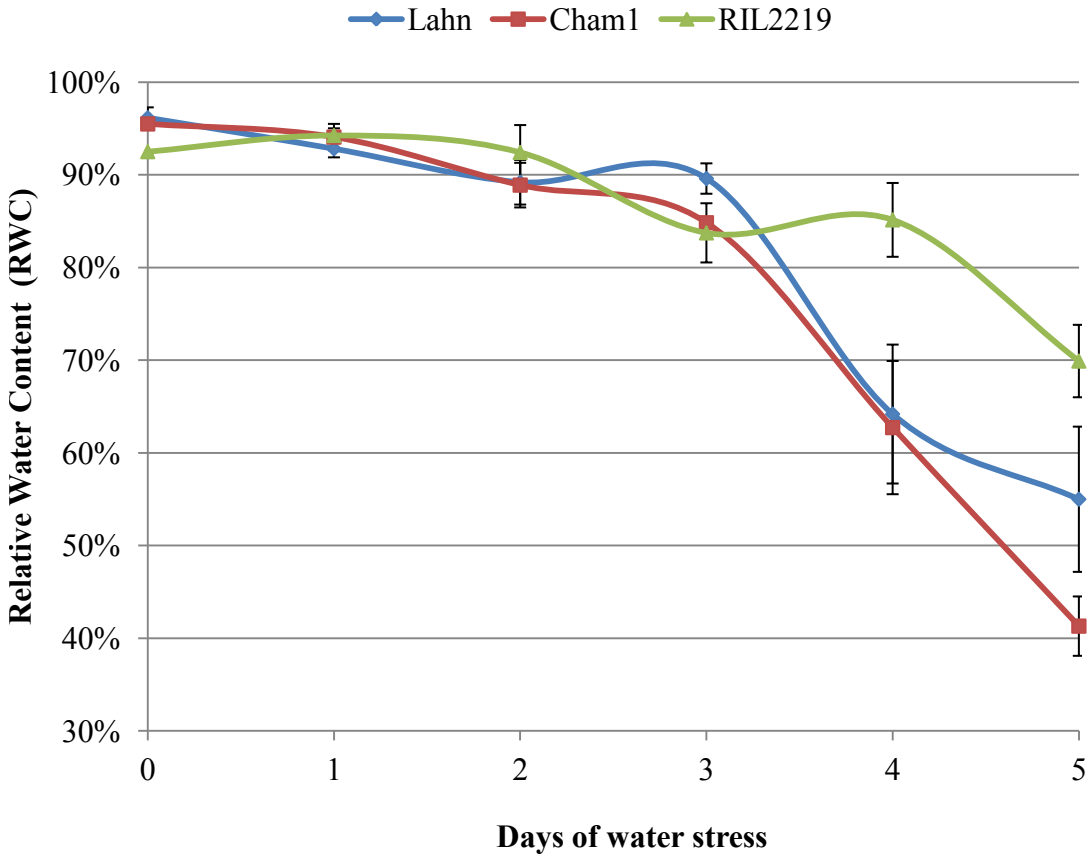


Figure 5.1: The mean Relative Water Content (RWC) over five days of stress in the three cultivars: Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line). Error bars are standard error of the mean. Raw data was provided by Marcela Baudo (2008).

From the interpolated RWC data, gene expression values were estimated for values of RWC from 40% to 90% in increments of 10%. Expression patterns between days were predicted by fitting a Lagrange polynomial to the data, this enabled the interpolation of expression according to RWC points, which fell in the interval between the daily expression measurements. This fitting function assumes that gene expression is tightly regulated (*i.e.* it does not modulate dramatically between days). However, given that the measurements are 24 hours apart and gene expression in plants can affect a change in protein levels after 3 hours (Piques *et al.*, 2009), this assumption for some genes may be incorrect. The function does allow gene expression to change rapidly, multiple times, in a positive or negative direction during the five days observed in the experiment. RWC does not fall much in the first two days of the experiment (Figure 5.1),

and the new extrapolated gene expression values are based on the interpolated RWC series, which falls within the interval of days 3, 4 and 5 of the experiment. The re-analysis of the experiment with regard to RWC is therefore an observation of the transcriptome from moderate to severe water stress, having already undergone 2 days of milder water stress. This analysis allowed for the exploration of the datasets both in terms of expression over the time course of stress and in terms of equivalent leaf RWC content.

ANOVA and PCO were then applied in the same manner as for line and time variables, to line and RWC variables. These statistical analyses of expression as a function of line and RWC are used throughout this chapter in the analysis of the progression of drought for the later time points (days 3, 4, and 5). They are viewed in conjunction with expression as a function of line and time which is more sensitive to the early onset of drought, before the onset of more dramatic physical changes in RWC.

5.5 Methods

The following section describe statistical procedures used to determine the statistical significance of an observing a CoPSA annotation for a subset of expressed genes.

5.5.1 Enrichment analysis for gene-sets annotated to GO terms

The analysis of the enrichment of functional annotation within gene sets utilised the Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) (Zheng and Wang, 2008). All Gene Ontology (GO) annotation were provided by the CoPSA annotation pipeline, using the multiple weighted fitness measure with per gene optimisation of evidence weightings (MWFM-OE) as was described

in Chapter 3.

A hyper-geometric tail, shown in Equation 5.1, was used to determine the probability that genes that annotate a GO term (t) in a set are enriched compared to the genes that annotate the same term in the whole chip. Where, k is the number of genes in the set to be tested, m is the total number of genes in the microarray, q is the number of genes in the tested set that annotate the term, and m the number of genes on the chip that annotate t. Hyper-geometric enrichment is a popular alternative to Fishers exact test (*e.g.* (Thibaud-Nissen *et al.*, 2006, Popescu *et al.*, 2009, Choudhury and Lahiri, 2011)) and has been used in many tools for GO term enrichment (Choudhury and Lahiri, 2011, Eden *et al.*, 2009, Zheng and Wang, 2008). It is computationally less demanding, but equivalent to, the one-tailed version of Fishers exact test (Rivals *et al.*, 2007). Depletion is not considered within this chapter; however, this can also be derived from a two-tailed Fishers exact test.

$$pvalue(t) = \sum_{i=q}^{m} \frac{\binom{m}{i}\binom{t-m}{k-i}}{\binom{t}{k}} \qquad \text{(Equation 5.1)}$$

Multiple repeated enrichment tests were conducted for GO terms not selected *a priori*, and therefore a correction for multiple testing must be applied to account for the expected proportion of falsely rejected null hypotheses (*i.e.* the probability of observing a GO term in the set is the same as on the chip) (Khatri and Drăghici, 2005). This False Discovery Rate (FDR) was controlled by using the (Benjamini and Yekutieli, 2001) p-value correction.

Enrichment is a useful analysis to determine *biological processes*, *molecular functions*, or *cellular components* that are overrepresented in a gene set, however there are a number of limitations to the analysis. It is constrained by the annotation itself, and low annotation coverage of a GO term inevitably results in an in-

creased likelihood of false negatives. In particular, low coverage of processes even when biological functions imply a common process lead to enrichment being missed. Also, current methods do not account for co-enrichment of biological functions processes, which can potentially be meaningful (Zheng and Lu, 2007). Co-enrichment of two biological functions or processes may have biological meaning within a group of proteins. For example: K^+ channel proteins which are also kinase binding may be co-enriched, which would indicate a significant specific-combination of these functions.

A contingency table for determining the set comparisons in enrichment is provided in Table 5.1. For the purposes of enrichment of GO annotated gene-sets within a given ANOVA group compared to an Affymetrix GeneChip C, $g \in C_t$ is the set of genes that are annotated with the given GO term t, $g \notin C_t$ are the genes that are not annotated by the GO term t. $g \in C_s$ are the significant genes that are found within the ANOVA group s, $g \notin C_s$ are the genes that are not found in the ANOVA group. The resulting 2x2 matrix therefore groups genes in 4 exclusive sets, and form the basis for comparative enrichment analysis.

Table 5.1: A contingency table for determining the comparisons for the enrichment of g within a set s for a term t within the set of genes on an Affymetrix GeneChip C.

	$g \in C_t$	$g \notin C_t$	$\cup g$
$g \in C_s$	$C_s \cap C_t$	$\frac{C_s}{C_t}$	C_s
$g \notin C_s$	$\frac{C_t}{C_s}$	$C(C_s \cup C_t)$	$\frac{C}{C_s}$
$\cup g$	C_t	$\frac{C}{C_t}$	C

5.5.2 Enrichment for transcription regulatory genes analysis

Fishers-exact-test was used for transcription factor analysis, as a smaller number of transcription factor families meant it was less computationally demand-

ing than GO gene-set enrichment. The contingency table for this is the same as that shown in Table 5.1, where s is the set of all transcriptional regulatory genes, and t is the given transcriptional regulatory family. Adaption was made for multiple testing using the correction by Benjamini and Hochberg (1995).

5.5.3 Identifying RCAR gene family members on the wheat GeneChip

The RCAR protein family, previously described in Chapter 6.1, is an ABA receptor, and therefore may play a critical role in ABA mediated water stress signalling. Unfortunately being a very recently functionally-characterised family, the public databases do not have an appropriately assigned EC number of GO term, even in *Arabidopsis*. It was therefore necessary to predict the orthologous gene family members on the wheat chip by a manually supervised sequence alignment procedure with the Arabidopsis RCAR proteins (PYL 1..13).
The translated, protein to nucleotide, alignment was performed using tBLASTn (NCBI version 2.2.24); this was done against a database of all consensus sequences on the wheat GeneChip, using the protein sequences of PYL 1 to 13 from UniProt release 2011_02 as the query. A conservative e-value of 1×10^{-7} was used, with default BLAST settings. A visual inspection of the hits revealed a cluster of strong hits ranging from e-values of 1×10^{-45} to 1×10^{-65}. Bitscores within this cluster ranged from 102 to 245. For each of these highly similar wheat sequences, GeneWise2 (Birney *et al.*, 2004) was used to translate the wheat nucleotide sequence to an amino acid sequence. GeneWise predicts a genes intron and exon structure using a similar protein structure as a template. For each of the nucleotide sequences in the orthologous wheat sequence cluster, the PYL gene with the highest bitscore was used as the template protein. A phylogram was generated from a ClustalW2 alignment, using the Neighbour-

Joining (NJ) method (Saitou and Nei, 1987, Studier and Keppler, 1988), which is a widely used (Gascuel and Steel, 2006) and fast method that clusters sequences by attempting to minimise the sum of the branch lengths. The NJ method does not require that all lineages have diverged by equal amounts. A key weakness is it provides only a single tree, and it cannot identify individual nucleotides that are informative or problematic to the tree construction (Sleator, 2011). However, as we are clustering RCAR genes, from a single gene family, we expect them to form a single tree. The full alignment of these RCAR genes is given in Appendix 7.

5.6 Results and Discussion

This section begins with a system-wide analysis of the water stress transcriptomes. Summary ANOVA statistics, principal coordinates analysis (PCO), and enrichment analysis facilitates the dissection of the microarray data into the principal variations and processes. The section then proceeds with a novel process-centric discussion of the early and late responses, pulling out enriched processes, and mechanisms currently thought to be associated with water stress. This analysis focuses on gene expression from a time analysis, but occasional reference is made the normalised RWC analysis. The time analysis includes early time-points that could be potentially important in explaining the overall yield stability in drought conditions of the Cham1 and RIL2219 cultivar lines. Because of space limitations it has not always been possible to include the analysis in terms of RWC and time, however where RWC provides additional insights it is included.

5.6.1 A systems-wide view of three water-stress transcriptomes

ANOVA dissected the datasets into four groups of significance:

1. Line alone contains line differences with no response to water-stress over time (these are constitutive genes which are useful for qPCR as controls for water stress).

2. Time alone contains genes responsive to water stress but with no difference between the three lines.

3. Time+line contains genes with a line and a time effect but without an interaction between these variables.

4. The line.time interaction ANOVA group contains genes with significant variation among the differences of the means, as line and time changes.

In addition to these four groups, there are two other sets of genes on the GeneChip, the first where no expression was reported, and the second has reported expression but no significant changes occurred over line or time were presence. The quantity of genes from the GeneChip present in each of these groups is given in Table 5.2.

The line.time group means that gene expression in this subset is a complex

Table 5.2: A summary of the number of genes in each ANOVA group

Group	Number of genes
Line alone significance (p-value<0.05)	781
Time alone significance (p-value<0.05)	2,877
Line+Time significance (p-value<0.05)	4,621
Line.time significance (p-value<0.05)	19,062
No significant (p-value<0.05)	420
No reported expression	27,291

interaction between line and time (*i.e.* these variables are not independent of each other). This was the most interesting ANOVA group because it identified genes with different expression profiles between varieties, and is a potential source of gene candidates that affect the yield-stability-under-drought trait. A 5% ($p<0.05$) Least Significant Difference (LSD) for each gene was used to identify *a posteriori* the genes that had significant differences compared to the well-watered control at each time-point in each cultivar. This enabled the dissection of line.time interaction into per time-cultivar significant subsets, for summary analysis.

Within the genes with significant line.time interaction, the progressive increase in water stress over time also resulted in an increased number of significantly expressed genes ($p<0.05$) relative to the control in all three cultivars (Figure 5.2). The number of expressed genes in Cham1 monotonically increases with days of progressive water stress. However, for the Lahn and RIL2219 cultivars, the quantity of significantly expressed genes modulated over time. In Lahn the number of significantly-expressed genes decreased for the first three days, and then rapidly increased. For the RIL2219 cultivar, which is highly yield stable under drought conditions, the number of genes expressed decreased until day two, slowly but progressively increased up until day four, and then rapidly increased on day five. The drought susceptible Lahn had a progressive and more measured increase in the number of genes expressed from day one, with a larger response observed much earlier than the other varieties. The amount of expression changes (fold change) seen across cultivars showed a marked increase at the last stages of the stress (Figure 5.3). Further, the modulation previously apparent for the number of genes changing expression levels over time in RIL2219 seen in Figure 5.2 was absent when mean fold change compared to the control was considered. The fall in the number of genes expressed during the second day in the line.time interaction group of RIL2219 does not accompany a fall in the mean expression. Within day 2, in Cham, similar quantities

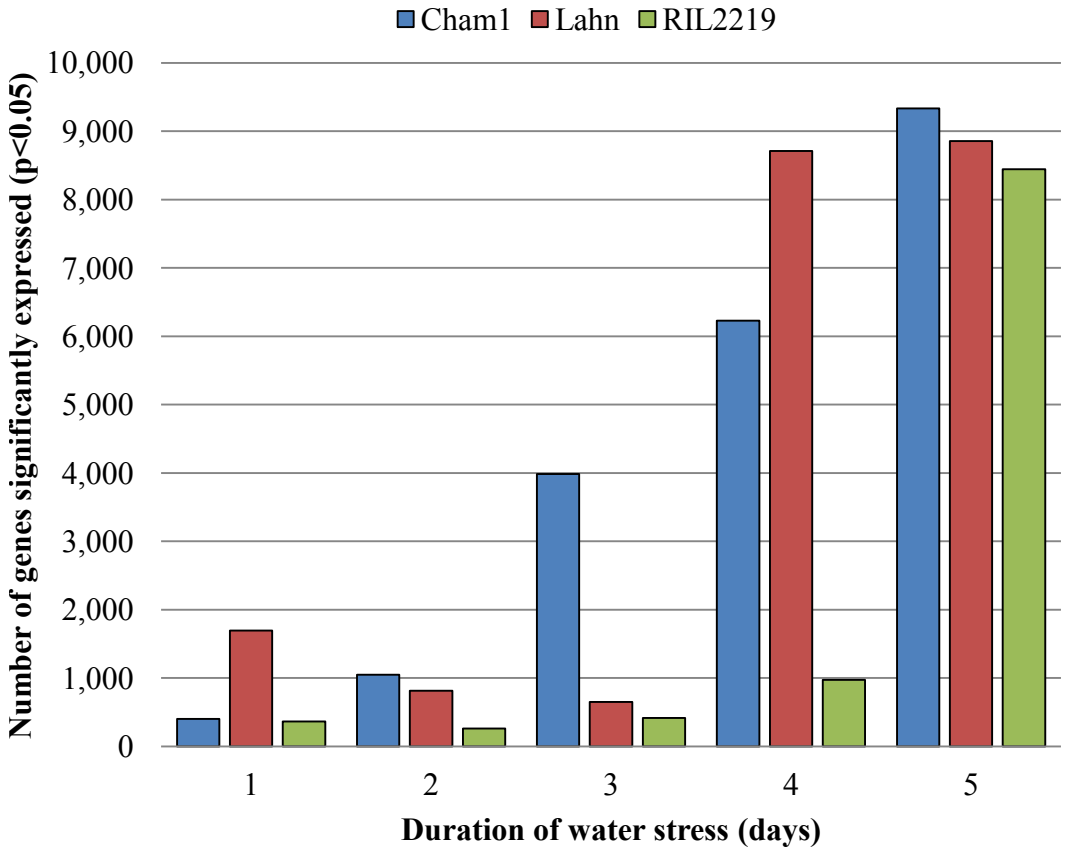


Figure 5.2: The number of genes with significant (p-value<0.05) line.time expression derived from the ANOVA LSD relative to stress free control for each day and in each of the three cultivars: Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line).

of genes were expressed as in the other lines, but at day 4, the amount of fold change dramatically drops and the quantity of genes increases. This indicates a trigger occurred on day three for Cham that was not seen in the other lines. The drought resistant RIL2219 and Cham1 both show delayed increases in the quantity of expression relative to the drought susceptible Lahn, which shows a dramatic increase after four days of stress. These preliminary views on the global changes in the transcriptome, hint at a relationship between the number of genes and their expression levels with yield stability under drought conditions.This global analysis shows that for all lines, there is a dramatic increase in the number and amount of fold expression in genes towards the end of the stress; at the severe stage where leaf RWC is below 70%. Further, the results also show that there are differences in the number genes regulated and their

mean fold change between the lines.

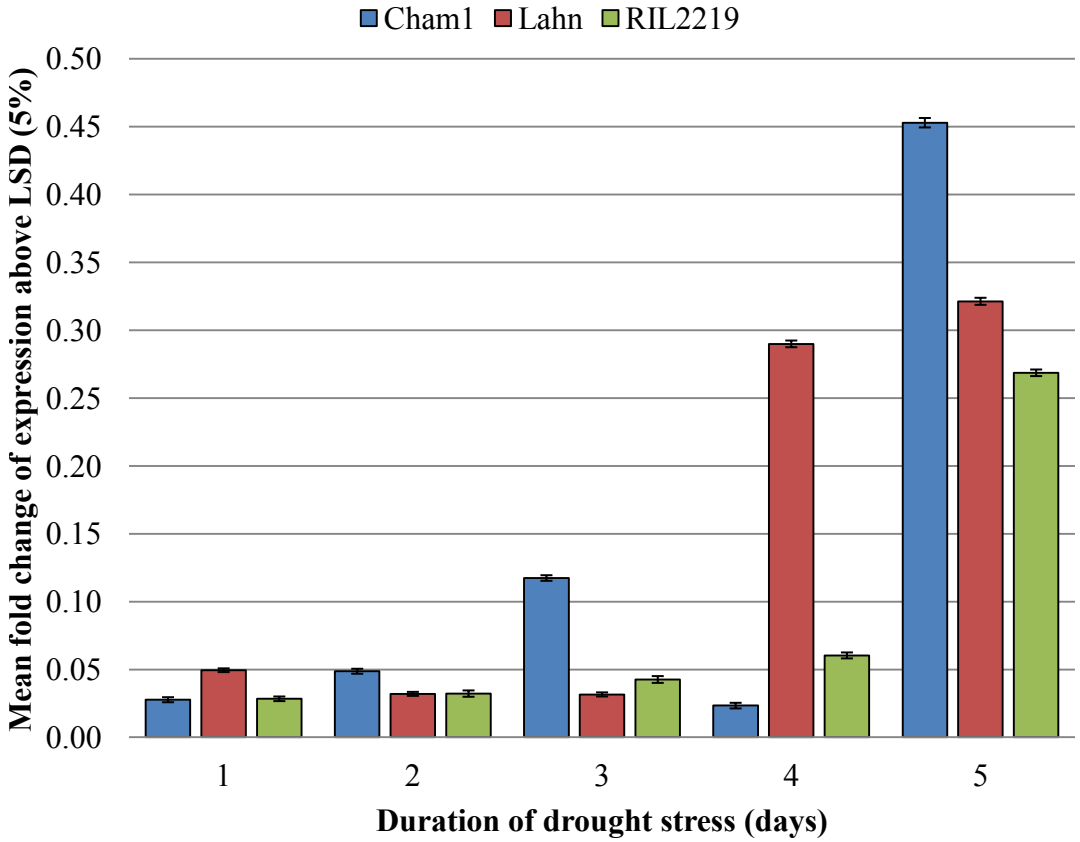


Figure 5.3: The mean fold change of expression above the Least Significant Difference (LSD) for genes with significant line.time interactions ($p<0.05$) for each day and in each of the three cultivars: Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line). Error bars are the standard error of the mean.

Differential regulation across cultivar lines

Observed differences in phenotype (such as drought tolerance) between lines can be the result of differential regulation of the same processes, or the activation of new processes by gene regulatory mechanisms. An analysis of the significantly regulated genes, common between cultivars, at each time-point, can reveal the degree to which common processes are being differentially regulated between cultivars.

The following proportional Venn diagrams (Figure 5.4-5.8) show a comparison of the genes common to all three cultivars within the line.time group that

were significantly regulated at each time-point. There were major differences in both the number and identity of those genes regulated at early and late time points. This indicates that there were qualitative differences in the stress response between the lines even during the first day of stress (Figure 5.4), when there was no observable difference in RWC (Figure 5.1). The 516 genes expressed only in the drought resistant cultivars Cham1 and RIL2219 are candidates for further studies.

These Venn intersections (Figure 5.4-5.8) indicate that for the regulation of genes in a per time-point comparison, the RIL2219 shares more genes in common with its Cham1 parent. Consistently throughout every time point the RIL2219 shared more regulated genes with its Cham1 parent that the Lahn parental cultivar.

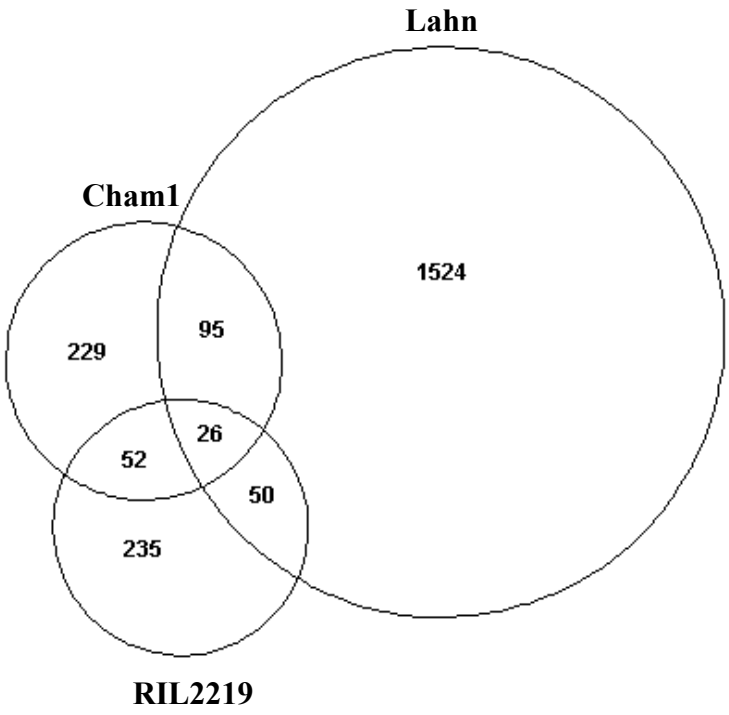


Figure 5.4: The intersection of significantly ($p<0.05$) expressed (relative to control) genes at day 1 in the three cultivars (Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line)), for genes with line.time significant expression.

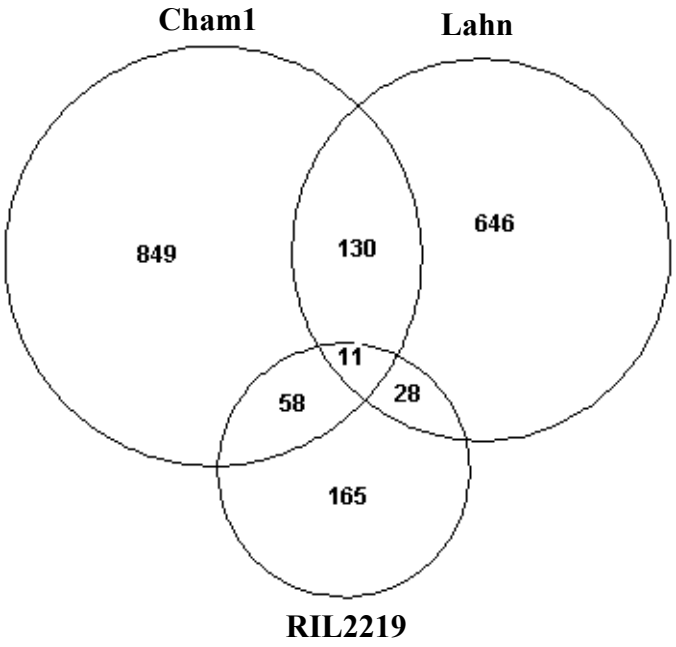


Figure 5.5: The intersection of significantly ($p<0.05$) expressed (relative to control) genes at day 2 in the three cultivars (Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line)), for genes with line.time significant expression.

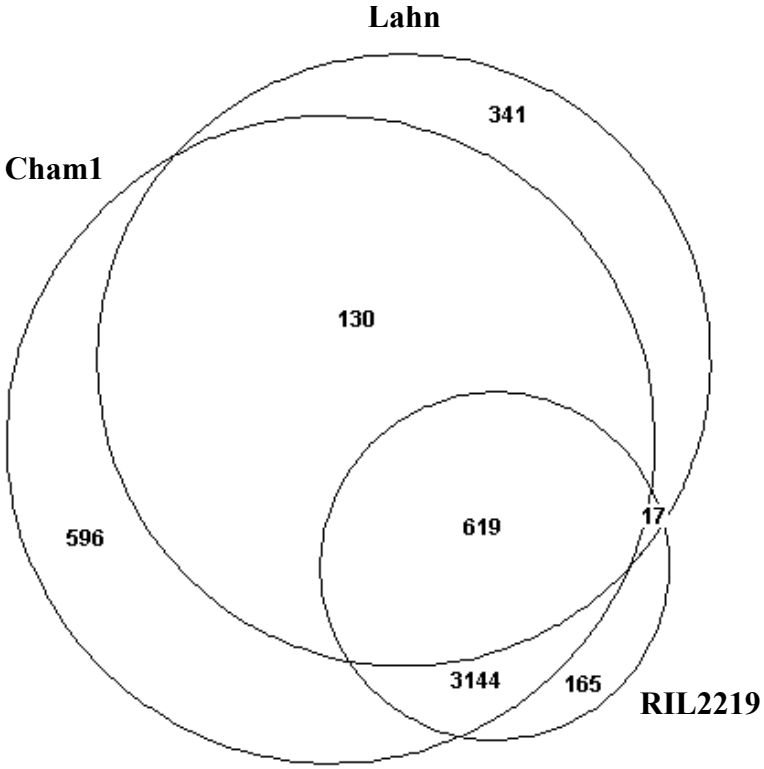


Figure 5.6: The intersection of significantly ($p<0.05$) expressed (relative to control) genes at day 3 in the three cultivars (Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line)), for genes with line.time significant expression.

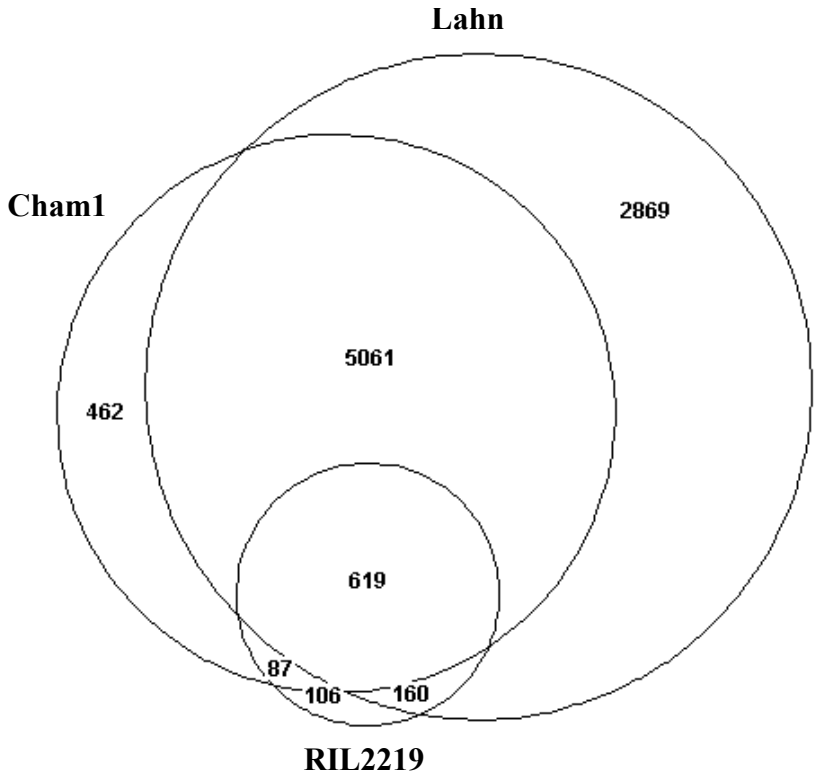


Figure 5.7: The intersection of significantly ($p<0.05$) expressed (relative to control) genes at day 4 in the three cultivars (Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line)), for genes with line.time significant expression.

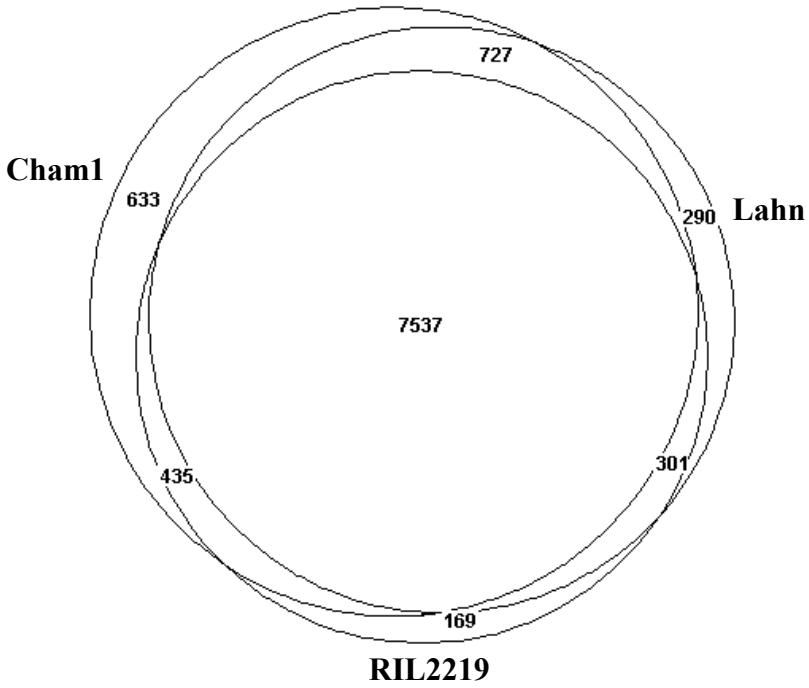


Figure 5.8: The intersection of significantly ($p<0.05$) expressed (relative to control) genes at day 5 in the three cultivars (Cham1 (drought tolerant), Lahn (high yielding but low drought tolerance) and RIl2219 (drought tolerant line)), for genes with line.time significant expression.

In the most drought resistant cultivar, RIL2219, the number of unique genes peaks at day one (Figure 5.4), with 235 significantly regulated genes. These unique early response genes are important candidates for conferring high yield-stability under drought, as they represent unique responses that occur before the more dramatic consequences of water stress develop in the plant. The enriched FO functions of these genes are given in Table 5.3. The presence of an endodeoxyribonuclease is surprising at such a minor change in RWC. These enzymes are often involved in DNA repair and are expressed in response to reactive oxygen species (Gutman and Niyogi, 2009, Choi *et al.*, 2005), which would be a consequence of severe osmotic stress. A gene encoding 3-methyl-2-oxobutanoate dehydrogenase (BCOAD) was also found to be significantly enriched in the expressed genes, after one day of water stress. BCOAD catalyses part of valine, leucine and isoleucine degradation pathways. The most abundant genes of enriched function after one day of water stress are 14 Heam oxy-

genase genes (HOs). HOs are significantly expressed after one day of stress, in the drought resistant RIL2219 cultivar, but are not present at this time point in any of the other cultivars. HOs catalyse the cleavage of heam to biverdin, iron and carbon monoxide and in plants are responsible for the production of the BV IXα precursor to the phytochome chromophore, phytochromobilin (Gisk *et al.*, 2010). These enzymes are therefore essential for photomorphogenesis, and therefore maintaining photosynthesis. Gisk *et al.* (2010) have also speculated that they may also play a role in responding to oxidative stress, as BV IXα has antioxidant properties. A folic acid transporter was also expressed within the RIL2219 Genes with enriched function, after one day of water stress. Folic acid is an essential cofactor for enzymes involved in the synthesis of purines, thymidylate, panthonetate, and methionine in plants (Neuburger *et al.*, 1996). Folic acid transport could therefore be affecting at a wide range of metabolic processes in the cell.

Table 5.3: Gene functions significantly enriched (p=0.05) in genes uniquely regulated at day 1 by RIL2219.

Go term	Description	Significance	Genes
GO:0000014	Single-stranded DNA specific endodeoxyribonuclease	0.05	1
GO:0003863	3-methyl-2-oxobutanoate dehydrogenase	0.05	1
GO:0004392	Heme oxygenase (decyclizing)	0.02	14
GO:0071614	Linoleic acid epoxygenase	0.005	6
GO:0008517	Folic acid transporter	0.05	1

5.6.2 Enriched processes with a temporal response to drought

The line.time group of genes that were identified as significant by ANOVA was the most interesting in terms of candidate gene identification (for improved drought resistance). However, line+time and time-only ANOVA groups also

captured drought responsive genes that vary over time but have no significant interaction. A combined analysis of these groups with the line.time group revealed all the significantly regulated drought responsive genes. Enrichment analysis of these time-responsive genes revealed processes that were regulated in response to drought.

Some of the most enriched processes within the genes that had significant time regulation are associated with regulation of transcription and translation. These include 1,271 genes that regulate gene expression with an enrichment p-value of 3.52×10^{-24} and 671 genes involved in translation with a p-value of 1.52×10^{-9}. This latter category included genes involved in ribosome biogenesis (243 genes, p=8.96×10^{-7}) and tRNA aminoacylation (80, p=7.04×10^{-8}). Post-translational processes such as protein folding (251 genes, p=3.61×10^{-16}), protein complex assembly (151 genes, p=0.0001) and protein transmembrane transport (76 genes, p=0.001) were also enriched. The metabolic machinery associated with these processes were also significantly enriched with 28 genes involved in purine ribonucleoside metabolism p=0.003), and 12 genes involved in lysine biosynthesis (p=0.02). There was also enrichment for 75 proteins associated with transmembrane import (p=0.00051).

There were 247 genes, involved in photosynthesis, which were highly enriched (p=2.58×10^{-26}) in this category, together with 6 genes specifically related to the negative regulation of photosynthesis (p=0.04). Chloroplast organization was also enriched (53 genes, p=2.81×10^{-5}) and together with protein targeting to chloroplast (29 genes, p=0.005×10^{-5}).

There was also evidence other stress responsive pathways, with enrichment for response to temperature stimulus (237 genes, p=0.02) of which 164 enriched genes where specifically associated with the cold stress response. Response to ozone (23 genes, p=0.01), inorganic substances (388 genes, p=0.002), and metal ions (352 genes, p=0.0002) were also enriched. These stress responses may be triggered as a result of cross-talk of pathways, and as a result of indirect

physiological effects of water stress.
Finally the other major enrichments were in seed development (352 genes, p=1.86×10^{-5}), and stomatal complex morphogenesis for which all 5 known genes on the chip that were associated with this process showed significantly regulated (5 genes, p=0).

5.6.3 Principle coordinates of variation

PCO was conducted on the individual replicates to explore the whole dataset and to identify top key candidate genes which explain the majority of variance in the data. In all instances a stronger cohesion was observed between replicates than the independent variable combinations. The plots of the PCos presented within this section are therefore based on a PCO analysis of the mean of the three replicates for each line.time combination, and line.RWC combination.
Visualising gene expression through PCO (Figure 5.9) revealed a clear separation of data points according to time, with early time points being more similar than later points. Most of this separation is contributed by the first Principal Coordinate (PCo1), with earlier time points clustering around higher values in the PCo. PCo1 accounts for 52.84% of variance, which strongly suggests that the major variation in gene expression is over time. There is a greater differentiation over time within the Cham1 and Lahn cultivars than the drought-tolerant RIL2219. This may indicate that the yield-stability of the RIL2219 cutivar, under water stress, is linked to a more moderated change in expression pattern over time, compared to the other varieties.
PCo2, however separate mostly according to cultivar and to a lesser extent time on a per cultivar basis, with Lahn clustering from 0.16 to -0.43, Cham1 from 0.15 to -0.56, and RIL2219 from 0.17 to -0.34. This arranges the cultivars in the same order as their respective yield-stabilities under drought-stress, which indicates

this PCo may be pulling out the underlying basis for these traits in the transcriptome. PCo2 accounts for 9.22% of variance in the data. PCo3 appears to separate according to cultivar and time, with the same yield stability order apparent in PCo2. PCo3 accounts for 6.80% of variance in the data.

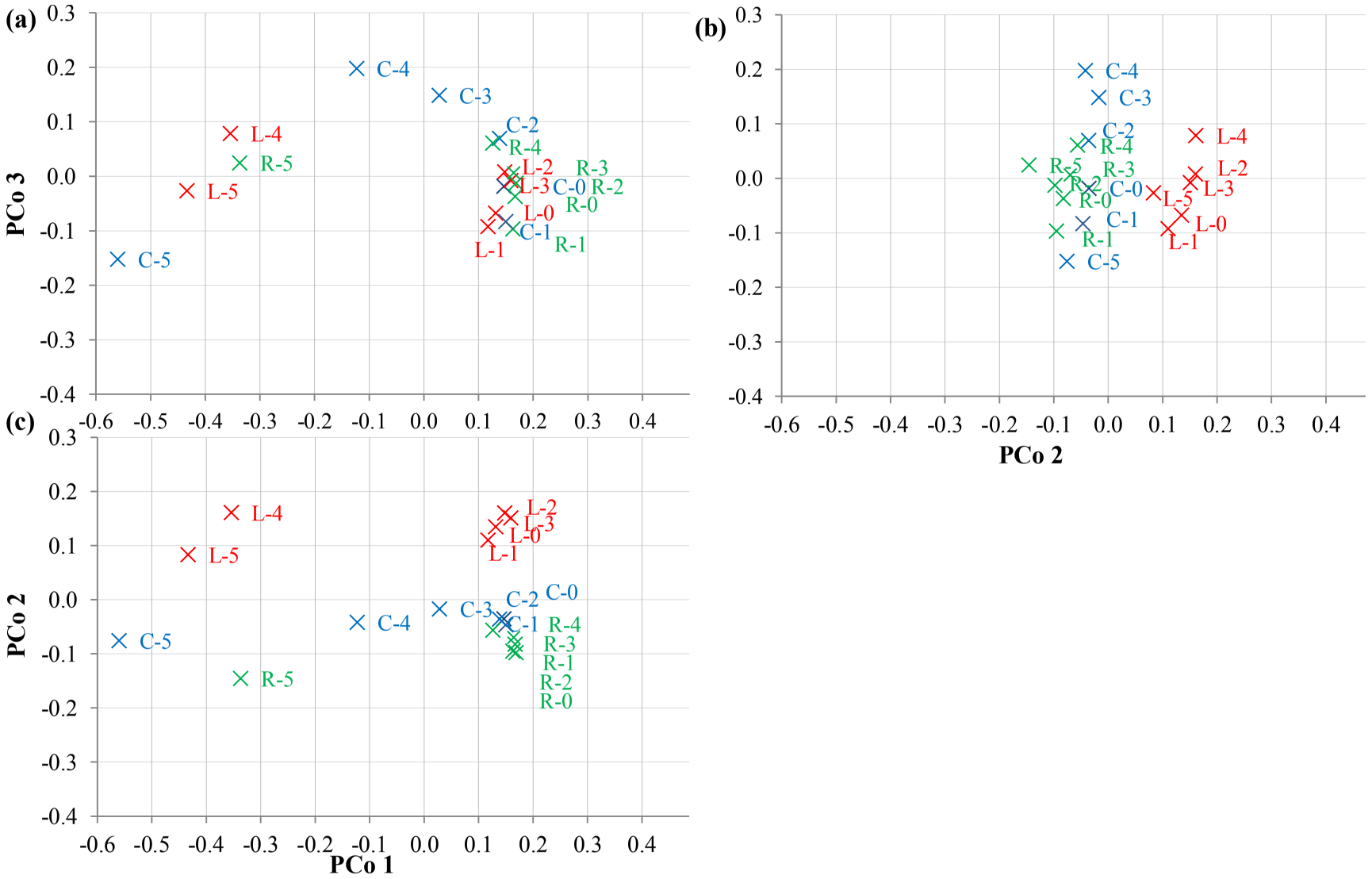


Figure 5.9: Principal Coordinates Analysis (PCO) of genes for PCo 1, 2, and 3 based on the mean of the replicates for all line against time combinations. The samples are identified by a combination of the cultivar (C = drought tolerant Cham1, L = high yielding but low drought tolerance Lahn and R = the drought tolerant line RIL2219) and the number of days of water stress (1, 2, 3, 4 and 5 days of stress). Data provided by Stephen Powers. (a) PCo1xPCo3 (b) PCo2xPCo3 (c) Pco1xPCo2

The results so far have highlighted that the progression of water stress over time effects a greater change in gene expression than observed between lines. The variation in RWC between lines shown in Figure 5.1 indicates that a time shift is present in the profile of gene expression between lines. Normalising the data with respect to leaf RWC (described in Section 5.4) allowed the dissection of differences within lines after day 3 of the stress. Figure 5.10 shows the results of PCO on the mean of the three replicates for each combination of line and RWC value. PCo1 very clearly shows the separation of lines and RWC values. There is a greater homogeneity in the distribution of cultivar time points in the PCo, as compared to PCo1 from Figure 5.10, which demonstrates that the procedure was effective in normalising the expression profiles against RWC. PCo1 accounts for 53.32% of variation in the data. As observed in the PCO analysis based on duration of water stress (time) (Figure 5.10), the RWC analysis clearly separates the cultivars in the same order as their yield-stability under drought both for PCo2 and PCo3.

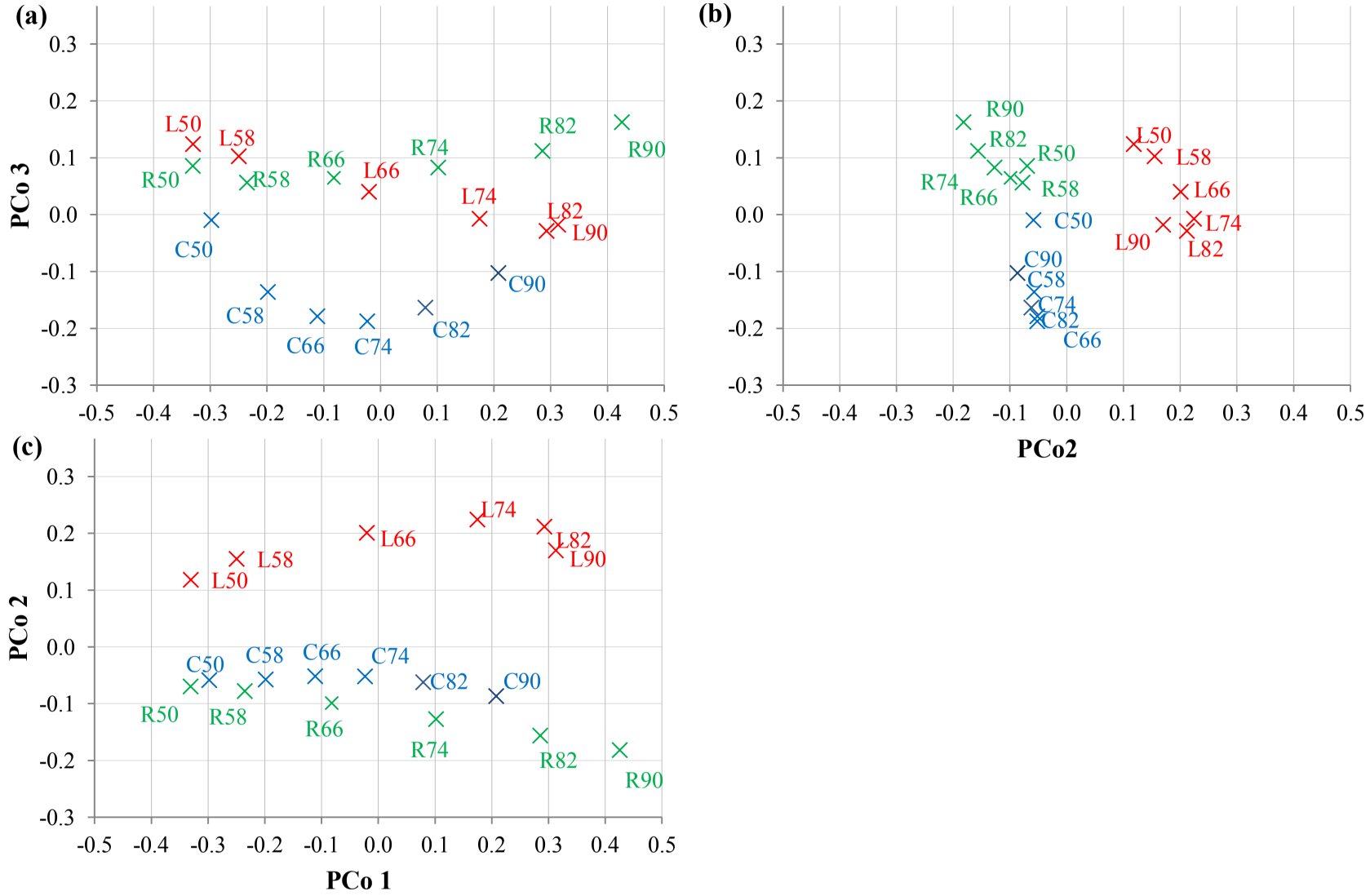


Figure 5.10: Principal Coordinates Analysis (PCO) of genes for PCo 1, 2, and 3 based on the mean of the replicates for all line against time combinations. The samples are identified by a combination of the cultivar (C = drought tolerant Cham1, L = high yielding but low drought tolerance Lahn and R = the drought tolerant line RIL2219) and RWC (90%, 82%, 74%, 66%, 58% and 50%). Data provided by Stephen Powers. (a) PCo1xPCo3 (b) PCo2xPCo3 (c) Pco1xPCo2.

Regression analysis of Principle Coordinates

Regression analysis, described in Section 5.3 and provided by Stephen Powers, was used to relate the gene expression values for each gene to each principal coordinate (PCo). It was therefore possible to rank the genes based on their contribution to the variance captured by the PCo. This section describes the GO *molecular functions* and *biological processes* that were represented within the genes that contributed the top 5% of the variation captured by each PCo. The PCo's discussed were calculated on the line.time expression data-set, the results of which as have been described previously in this section.

For PCo1 the genes most responsible for the variance separation were the ribosomal subunits which were identified in both functional (Table 5.4) and process (Table 5.5) annotations. For the genes that contributed 5% of the variance captured by PCo1, 47 out of the 69 genes were ribosomal subunits (Table 6.5). Ribosomal subunits involved in ribosome biogenesis (GO:0042254) were also found to be significantly enriched in the genes contributing the top 5% of variation for PCo1 (p-value = 3.17×10^{-9}). For all Ribosomal subunits in PCo1, the fold-change in expression compared to the control was greater than the least-significant-difference (LSD) (p-value <0.01). This confirms the previous observation in this section that PCo1 pulls out variation over time. The time-course expression pattern for these ribosomal subunit genes is shown in Figure 5.11. The second most common function within PCo1 was RNA binding, which is a co-function of a sub-set of ribosomal subunits. The remaining genes within PCo1 are mostly transcription factors, protein-protein interactions and enzymes.

The gene TaAffx.44105.1.S1_at is the second most important gene contributing to PCo1. It was reported by CoPSA as a transcription factor because of its weak similarity to At3g50685, which was erroneously annotated by AGRIS as a transcription factor (this annotation has since been removed in the most recent version of AGRIS, January 2011). A closer inspection reveals that the gene has the

greatest similarity to 02g023760 from *Sorghum bicolor* (67% amino acid identity, ClustalW2 alignment provided in Appendix 6). This gene family is functionally uncharacterised even in *Arabidopsis*. This is potentially an interesting target for further experimental research, given its importance in PCo1, and its functionally uncharacterised status. There is also one protein transporter that is the 9^{th} highest contributor to the PCo. A full rank of the genes contributing to the top 5% of variation captured by PCo1, together with their exact CoPSA predicted functions and processes can be found in Appendix 8 and 9.

GO *molecular function* and *biological process* annotation of genes in PCo1 indicates that overall the main contributor of gene expression variation during the progression of water stress over time, are processes controlling the synthesis and transport of proteins. As protein synthesis is fundamental to all processes in the cell, this suggests that the plant is undergoing a *global* reconfiguration of molecular processes in response to a progressively increasing water stress over time. The absence of ribosomal proteins, within the main contributors to PCo2 and PCo3, addressed later in this section, indicates that the regulation of protein synthesis is more of a constitutive response to stress, than a strong differentiator the stress response between cultivars.

In response to the observation of ribosomal proteins in PCo1, the role of ribosomal proteins that were significantly expressed in the ANOVA time-dependent groups (time-only, line+time, and line.time) was also studied. A significant enrichment (p value = 8.96×10^{-7}) was found for 243 genes involved in the ribosome biogenesis process within these ANOVA groups, compared the prevalence genes annotated with this process on the wheat GeneChip. The same genes, encoding ribosomal subunits, are responsible for the significant enrichment of the GO ribosomal biogenesis process. Repression of ribosomal biogenesis is a known response to oxidative stress (Novoa *et al.*, 2003), which is a physical consequence of water stress on the cell. The subsequent reduction in protein synthesis, as a result of reduced ribosome availability, prevents proteins being

abnormally folded during osmotic stress conditions and reduces the stress on the endoplasmic reticulum (Harding *et al.*, 2000). This suggests that this late inhibition of ribosome biosynthesis is a response to physical water stress on the cell. The delayed inhibition of protein synthesis in RIL2219 (Figure 5.11(a)) is indicative that physical stress to the cell is mitigated by other adaptations. This is consistent with the delayed fall in RWC observed in RIL2219, with an RWC of around 70% being consistent with the inhibition of protein synthesis (Figure 5.1).

Table 5.4: The high-level GO *molecular functions* of the top 5% contributing genes in PCo1 as annotated by CoPSA. Some functions have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms from different parts of the tree may appear twice. 67 out of the 70 genes within the 5% contributing genes have a predicted function and appear in this report. A full listing of the genes and their GO term annotations, from the top 5% of PCo1, is available in Appendix 8 and graphical as a DAG in Appendix 9.

Molecular Function	**Function name**	**Genes**
GO:0003735	Structural constituent of ribosome	47
GO:0003723	RNA binding	8
GO:0003700	Sequence-specific DNA binding transcription factor activity	7
GO:0003824	Catalytic Activity	7
GO:0005515	Protein binding	5
GO:0008565	Protein transporter	1
GO:0003755	Translation initiation factor activity	1

Table 5.5: The high-level GO *biological process* of the top 5% contributing genes in PCo1 as annotated by CoPSA. Some processes have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms from different parts of the tree may appear twice. 60 out of the 70 genes within the 5% contributing genes have a predicted process and appear in this report. A full listing of the genes and their GO term annotations from the top 5% of PCo1 is available in Appendix 8 and graphical as a DAG in Appendix 9.

Biological Process	**Process name**	**Genes**
GO:0006412	Translation	47
GO:0042254	Ribosome biogenesis	11
GO:0009416	Response to light stimulus	4
GO:0009965	Protein Folding	3
GO:0009651	Response to salt stress	2
GO:0042742	Defence response	2
GO:0000917	Barrier septum formation	1
GO:0006396	RNA Processing	1
GO:0009765	Photosynthesis	1
GO:0015031	Protein transport	1
GO:0015995	Chlorophyll biosynthetic process	1
GO:0009739	Response to gibberellin stimulus	1

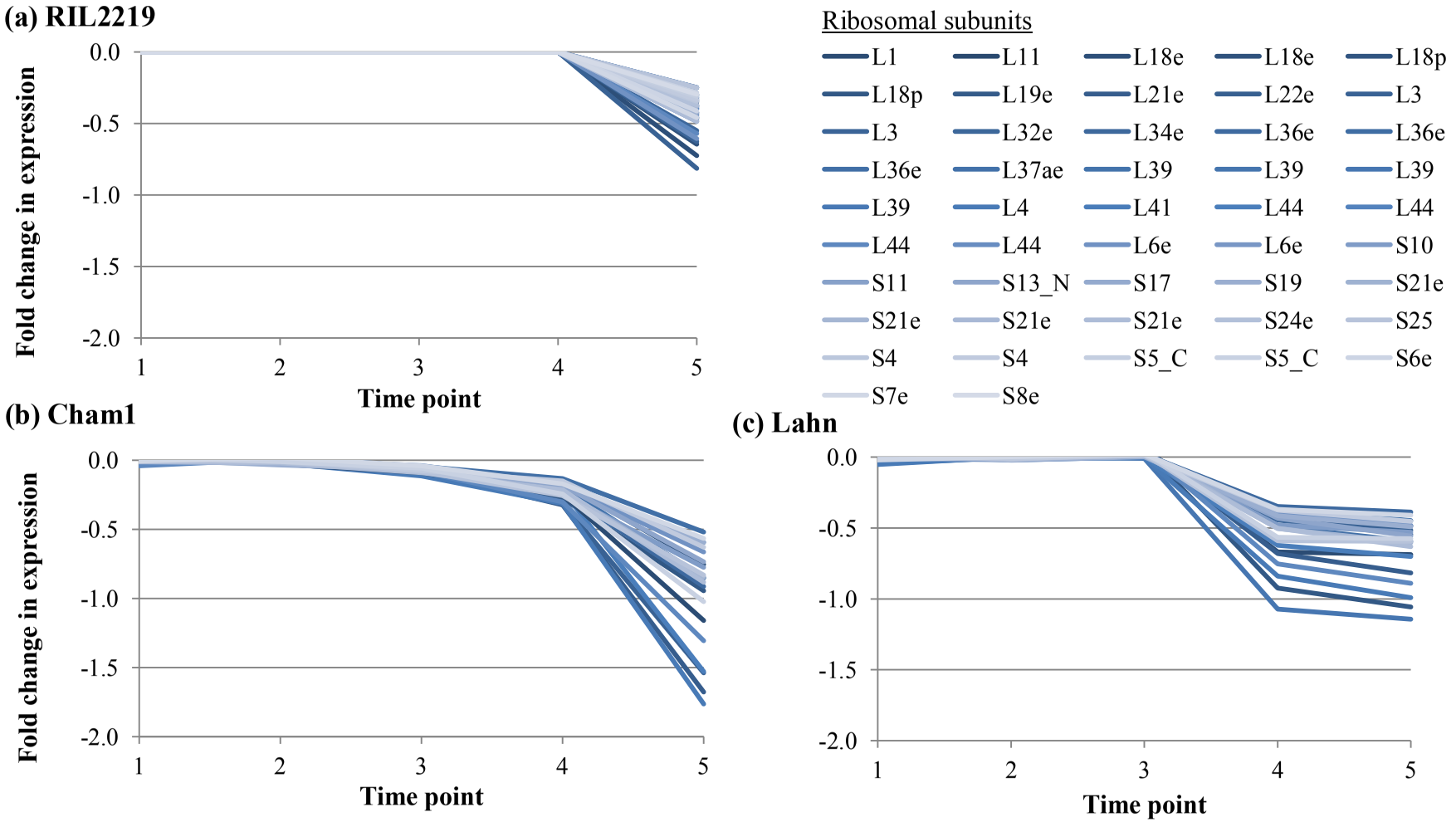


Figure 5.11: Ribosomal subunit genes in PCo1. The plots show the fold change in expression above the least significant difference (LSD) (p=0.01). Gene expression is shown for cultivars RIL2219 (a), Cham1, (b) and Lahn (c).

Table 5.6 and 5.7 show a summary of the GO functions and processes respectively, that were used by CoPSA to annotate genes within PCo2. There were 38 genes that contributed the top 5% variation captured within PCo2, of which 10 were identified as enzymes. This was higher in quantity and proportion than the 7 enzymes out of 70 genes in the top 5% of PCo1. Ubiquitin-protein ligase enzymes ubiquitinate proteins, to signal their degradation. The NADH dehydrogenase (ubiquinone) enzyme, catalyses the conversion of ubiquinone (necessary for ubiquitin-protein ligases) into ubiquinol. Regulation of these two proteins in PCo2, which mostly captures variation across lines, indicates that regulation of protein degradation is important in differentiating the lines. The enzyme disulfide oxidoreductase is involved in redox and therefore potentially regulates homeostasis, an important response to osmotic stress. There were 12 signalling related genes in the top 5% of this PCo, whose function included binding of nucleotides, proteins, cofactors, and lipids. As PCo2 mainly captures variation between the lines, this indicates that control of signalling related processes is important for differentiating differences in line. This hints that the drought resistant RIL2219 may be able to ameliorate the fall in RWC through a differential signalling strategy.

The *biological processes* that the genes in the top 5% of variation captured in PCo2, are diverse, however many of the processes are direct or indirect responses to water stress on the cell. The observed, responses to metal ions, and ion transport, may be processes responding to ion imbalances, caused by osmotic stress. They may also be involved in signalling, as Ca^{++} is an important in the ABA mediated signalling pathway (Chapter 6.1). Additionally, the presence of protein phosphorylation and ubiquitination processes indicates that this PCo has a strong signalling element.

Table 5.8 and 5.9 show a summary of the GO functions and processes respectively, that were used by CoPSA to annotate genes within PCo2. There were 41 genes that contributed the top 5% variation captured within PCo3. As with

Table 5.6: The high-level GO *molecular functions* of the top 5% contributing genes in PCo2 as annotated by CoPSA. Some functions have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms from different parts of the tree may appear twice. 20 out of the 38 genes within the 5% contributing genes have a predicted function and appear in this report. A full listing of the genes and their GO term annotations from the top 5% of PCo2 is available in Appendix 14.

Molecular Function	Function name	Genes
GO:0005515	Protein binding	6
GO:0009055	Electron carrier activity	3
GO:0005488	Binding (type not known)	2
GO:0030234	Enzyme regulator activity	2
GO:0008289	Lipid binding	2
GO:0016491	Oxidoreductase activity	2
GO:0015035	Protein disulfide oxidoreductase activity	2
GO:0004872	Receptor activity	2
GO:0004842	Ubiquitin-protein ligase activity	1
GO:0003824	Catalytic activity	1
GO:0048037	Cofactor binding	1
GO:0008137	NADH dehydrogenase (ubiquinone) activity	1
GO:0000166	Nucleotide binding	1
GO:0045735	Nutrient reservoir activity	1
GO:0004867	Serine-type endopeptidase inhibitor activity	1
GO:001615	Sucrose synthase activity	1

PCo2, there is a broad spectrum of functions captured by the top 5% of this PCo. As previously explained in this section, PCo3, mostly captures variation between lines. One of the most prominent gene functions in the top 5% of PCo3 is DNA binding, which accounts for 8 out of 41 of the genes within the top 5%. Given the absence of DNA binding in PCo2, PCo3 appears to captures the main variation of transcriptional control of across lines. The transcriptional control is distinct from the 8 DNA binding genes found in PCo1, which accounts mainly for variation across time. This indicates the presence of two distinct transcriptional clusters, which regulate time and line variation respectively. The variation in early and late transcription factor expression, and their gene families is further discusses in Section 5.6.4. 14 out of 41 genes within the top 5% of PCo3 are enzymes.

Trehalose phosphatase (EC: 3.1.3.12) has both a unique and mixed probe set (*x*

Table 5.7: The high-level GO *biological process* of the top % contributing genes in PCo2 as annotated by CoPSA. Some processes have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms in different parts of the tree may appear twice. 12 out of the 38 genes within the 5% contributing genes have a predicted process and appear in this report. A full listing of the genes and their GO term annotations from the top 5% of PCo2 is available in Appendix 14.

Biological Process	Process name	Genes
GO:0009058	Biosynthetic process	2
GO:0045454	Cell redox homeostasis	2
GO:0010252	Auxin homeostasis	1
GO:0006812	Cation transport	1
GO:0045454	Cell redox homeostasis	1
GO:0006879	Cellular iron ion homeostasis	1
GO:0042742	Defence response to bacterium	1
GO:0048527	Lateral root development	1
GO:0006869	Lipid transport	1
GO:0007140	Male meiosis	1
GO:0007067	Mitosis,	1
GO:0050819	Negative regulation of coagulation,	1
GO:0045910	Negative regulation of DNA recombination	1
GO:0009860	Pollen tube growth,	1
GO:0006468	Protein phosphorylation	1
GO:0016567	Protein ubiquitination	1
GO:0001558	Regulation of cell growth	1
GO:0046686	Response to cadmium ion	1
GO:0046686	Response to cadmium ion	1
GO:0009409	Response to cold	1
GO:0009408	Response to heat	1
GO:0006979	Response to oxidative stress	1
GO:0009639	Response to red or far red light	1
GO:0009651	Response to salt stress	1
GO:0009414	Response to water deprivation	1
GO:0048768	Root hair cell tip growth,	1
GO:0005985	Sucrose metabolic process	1
GO:0006511	Ubiquitin-dependent protein catabolic process	1

type probe-set) (see Chapter 3.1.2(*a*) for explanation of probe-set types) represented in the top 5% of variation in PCo3, which is involved in the biosynthesis of Trehalose. Trehalose has been shown to be involved in stress signalling, as well as the regulation of growth and pathogen response (Fernandez *et al.*, 2010). Iturriaga *et al.* (2009), have recently shown that over expression of the AtTPS1 gene, encoding this enzyme in Arabidopsis, resulted in an increase in trehalose. The over expression of AtTPS1, also inferred drought tolerance. The presence of this enzyme in a PCo that captures line variation, indicates that trehalose mediated signalling may be playing an important role in conferring a greater water stress resistance in these varieties.

The third and 17^{th} highest contributors to PCo3 are mixed (x type probe-set) and unique probe-sets for Ta.4760.1.S1_ at, which encodes a calcium-transporting ATPase activity that is also calmodulin binding. Further Given the importance of Ca^{++} in ABA mediated drought signalling through CDPKs, this could potentially be an interesting target for future study.

5.6.4 Control of transcription

Transcription factors control regulation of gene expression and in turn the quantity of RNA. This RNA may in itself regulate gene expression (Bonnet *et al.*, 2006), encode transcription factor that regulate other genes, or proteins with a diverse array of functions within the cell. Transcription factors therefore are powerful regulators of processes within the cell. A number of families have been shown to play a key role in regulating the plants response to drought (Shinozaki and Yamaguchi-Shinozaki, 2007). Other families of genes that control transcription via other means are also included in this analysis. For example: SET genes control transcription through histone methylation (Marmorstein, 2003). It was important to target transcription factors because of their

Table 5.8: The high-level gene functions of the top 5% contributing genes in PCo3 as annotated by CoPSA. Some functions have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms from different parts of the tree may appear twice. 35 out of the 41 genes within the 5% contributing genes have a predicted function and appear in this report. A full listing of the genes and their GO term annotations from the top 5% of PCo3 is available in Appendix 11.

Molecular Function	Function name	Genes
GO:0003677	DNA binding	8
GO:0022857	Transmembrane transporter activity	3
GO:0005488	Binding	2
GO:0005515	Protein binding	2
GO:0016787	Hydrolase activity	2
GO:0004805	Trehalose-phosphatase activity	2
GO:0005516	Calmodulin binding	2
GO:0005524	ATP binding	2
GO:0008270	Zinc ion binding	2
GO:0008381	Mechanically-gated ion channel activity	1
GO:0005388	Calcium-transporting ATPase activity	1
GO:0030599	Pectinesterase activity	1
GO:0004857	Enzyme inhibitor activity	1
GO:0008716	D-alanine-D-alanine ligase activity	1
GO:0004028	3-chloroallyl aldehyde dehydrogenase activity	1
GO:0004198	Calcium-dependent cysteine-type endopeptidase activity	1
GO:0008661	1-deoxy-D-xylulose-5-phosphate synthase activity	1
GO:0008716	D-alanine-D-alanine ligase activity	1
GO:0004321	Fatty-acyl-CoA synthase activity	1
GO:0016229	Steroid dehydrogenase activity	1
GO:0004197	Cysteine-type endopeptidase activity	1
GO:0005506	Iron ion binding	1
GO:0008825	Cyclopropane-fatty-acyl-phospholipid synthase activity	1
GO:0050342	Tocopherol O-methyltransferase activity	1
GO:0080064	4,4-dimethyl-9β,19-cyclopropylsterol-4α-methyl oxidase activity	1

Table 5.9: The high-level gene processes of the top 5% contributing genes in PCo3 as annotated by CoPSA. Some processes have been generalised up the GO tree in order to provide a simpler overview. Genes with multiple annotated terms in different parts of the tree may appear twice. 27 out of the 41 genes within the 5% contributing genes have a predicted process and appear in this report. A full listing of the genes and their GO term annotations from the top 5% of PCo3 is available in Appendix 11.

Biological Process	Function name	Genes
GO:0008610	Lipid biosynthetic process	1
GO:0016126	Sterol biosynthetic process	1
GO:0009409	Response to cold	2
GO:0042538	Hyperosmotic salinity response	2
GO:0006281	DNA repair	1
GO:0006754	ATP biosynthetic process	2
GO:0009624	Response to nematode	1
GO:0006839	Mitochondrial transport	1
GO:0009790	Embryo development	1
GO:0005992	Trehalose biosynthetic process	2
GO:0045449	Regulation of transcription	3
GO:0048366	Leaf development	1
GO:0010386	Lateral root primordium development	1
GO:0009723	Response to ethylene stimulus	1
GO:0009414	Response to water deprivation	1
GO:0009737	Response to abscisic acid stimulus	1
GO:0009252	Peptidoglycan biosynthetic process	2
GO:0055114	Oxidation-reduction process	2
GO:0006508	Proteolysis	2
GO:0009611	Response to wounding	1
GO:0016114	Terpenoid biosynthetic process	1
GO:0055085	Transmembrane transport	1
GO:0050982	Detection of mechanical stimulus	1
GO:0016567	Protein ubiquitination	1
GO:0006917	Induction of apoptosis	1

unique regulatory properties, and our poor understanding of the complexity of their interactions and involvement in water stress in wheat.

a) Transcriptional response to drought *per se*

Figure 5.12 shows the major families of transcription factors that change significantly over time. It is derived from a CoPSA TF family annotation of the genes in the ANOVA groups: time-only, time+line, and line.time (treated all together). It highlights the early responsive transcription factor families, which were defined as those that were significantly regulated within the first two days. These early responses represent signalling activity prior to any large changes in RWC (Figure 5.1). Table 5.10 summarises the numbers of genes that are regulated at early and late time-points, grouped according to the transcription factor families provided by CoPSA, that are currently thought to be drought responsive (Shameer *et al.*, 2009, Shinozaki and Yamaguchi-Shinozaki, 2007). All of these known transcription factors are expressed in large quantities with the time-responsive genes (Figure 5.12), and all appear within the top 50% of transcription factor families expressed, ranked by the number of genes expressed. Additionally the WRKY super-family contains members in-

Table 5.10: Summary of known transcription regulating gene families associated with drought, and the number of early and late regulated genes in the time responsive ANOVA groups. Major drought responsive transcription factors families comes from Shinozaki and Yamaguchi-Shinozaki (2007) and Shameer *et al.* (2009).

Protein family	Early responsive genes			Late responsive genes		
	Lahn	**Cham**	**RIL2219**	**Lahn**	**Cham**	**RIL2219**
AP2-EREBP (ERF)	0	0	1	29	29	24
BHLH	2	2	4	18	18	17
BZIP	3	3	5	26	26	25
HOMEOBOX	3	3	2	16	16	17
HSF	5	5	1	1	1	2
MYB	5	5	7	57	57	67
NAC	2	2	24	22	22	20

volved in senescence, and have been found in drought and a salt-stressed tissues (Eulgem *et al.*, 2000). The onset of senescence has been associated with water stress (Weaver *et al.*, 1998), and delayed senescence has been shown to improve drought tolerance (Rivero *et al.*, 2007). Within the early regulated genes, the only set of genes significantly enriched relative to their abundance on the chip were the Heat Shock Factors (HSF) which were enriched for Lahn (5 genes, p-value=0.02). For late time-points no significant enrichment was found for any transcription factor family within Lahn however there were 5 significantly enriched families in Cham1, and 4 in RIL2219 (Table 5.11 and Table 5.12 respectively). Those families enriched in RIL2219 were a subset those enriched in Cham1.

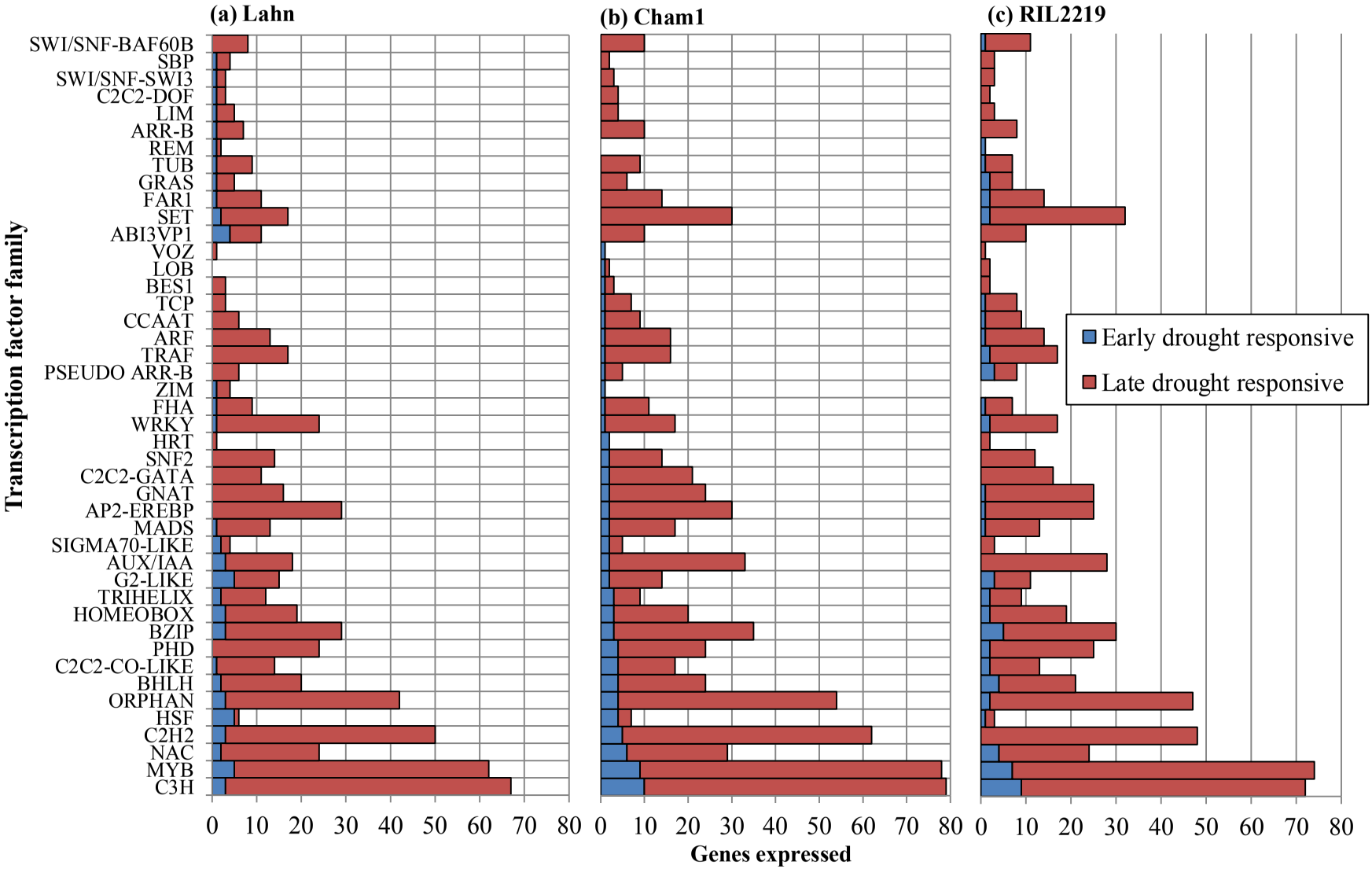


Figure 5.12: Transcription factors that change significantly over time, categorised as early (≤2days) and late (>2days) responsive, and grouped by family. Only transcription factor families with more than 5 expressed transcription factors are shown.

Enrichment analysis was performed on these time responsive genes, using Fishers-exact test with a correction for multiple testing devised by Benjamini and Hochberg (1995). These did not reveal any significantly enriched families within Lahn, however a number of transcription regulatory genes were enriched, the majority of which were not transcription factors, with the exception of the C2C2-GATA family in Cham1 (Table 5.11).

Table 5.11: Significantly enriched families of genes that regulate translation within Cham1.

Family	p-value	Genes
SET	5.60×10^{-4}	30
AUX/IAA	0.001	31
GNAT	7.27×10^{-3}	22
BHLH	0.01	20
C2C2-GATA	0.01	19
SWI/SNF-BAF60B	0.04	10

Table 5.12: Significantly enriched families of genes that regulate translation within RIL2219.

Family	p-value	Genes
SET	9.08×10^{-5}	30
GNAT	1.99×10^{-4}	24
AUX/IAA	0.002	28
BHLH	0.006	17
SWI/SNF-BAF60B	0.02	10

b) Cultivar specific transcriptional responses to drought

Figure 5.13 shows the transcription factor families that show significant regulation in the line.time ANOVA group which are a subset of those genes in Figure 5.12. As in Figure 5.12 these are broken down into summaries for each cultivar, by the early or late-responsive genes significantly expressed (p-value < 0.05) compared to the control. In order to reduce the number of families to a manageable level, only families with more than five transcriptions factors expressed at any time point are shown. Heat Shock Factors (HSFs) represent the largest

number of early responsive genes within the TRITIMED dataset, with the most present in Lahn, followed by Cham1 and RIL2219 respectively. This seems to closely follow RWC content and yield-stability-under-drought in these varieties. HSF were almost exclusively down regulated at early and late time points, with the exception of one in Cham1, which was significantly up-regulated during day 4.

A HSF is transcription factor that regulates a Heat Shock Protein (HSP). HSPs act as chaperones to ensure the correct folding of proteins, and are produced in response to heat, as a mechanism for refolding denatured proteins. In plants, the HSF family is particularly diverse and is expressed during heat and drought-stress, and seed development (von Koskull-Döring *et al.*, 2007). von Koskull-Döring *et al.* (2007) speculate that the HSFs are likely be involved in a number of other yet unknown signalling process. As the inability to adapt for heat is often a consequence of drought, since transpiration is reduced, the HSFs are therefore likely to be produced as an indirect response to heat stress. Additionally, an abnormal osmolyte potential in a cell, caused by water stress, also adversely affects correct protein folding (Hu *et al.*, 2009). Some HSFs are also known to responsive to senescence (Breeze *et al.*, 2008).

The largest group of transcription factors expressed in the line.time category are the MYB family, which are known to be involved in osmotic-stress response. MYBs have been shown to act as part of ABA mediated signalling (Abe *et al.*, 2003). All of the other transcription factor families known to be associated with drought are present in this ANOVA group and are shown in Table 5.13.

Table 5.13: Summary of known transcription regulating gene families associated with drought, and the number of early and late regulated genes in the line.time ANOVA group. Major drought responsive transcription factors families comes from Shinozaki and Yamaguchi-Shinozaki (2007) and Shameer *et al.* (2009).

Protein family	Early responsive genes			Late responsive genes		
	Lahn	**Cham**	**RIL2219**	**Lahn**	**Cham**	**RIL2219**
AP2-EREBP (ERF)	0	0	0	1	0	0
BHLH	2	3	4	10	11	10
BZIP	2	3	1	17	19	12
HOMEOBOX	2	2	1	10	9	11
HSF	4	3	1	1	1	1
MYB	1	8	5	34	38	37
NAC	2	5	1	12	13	13

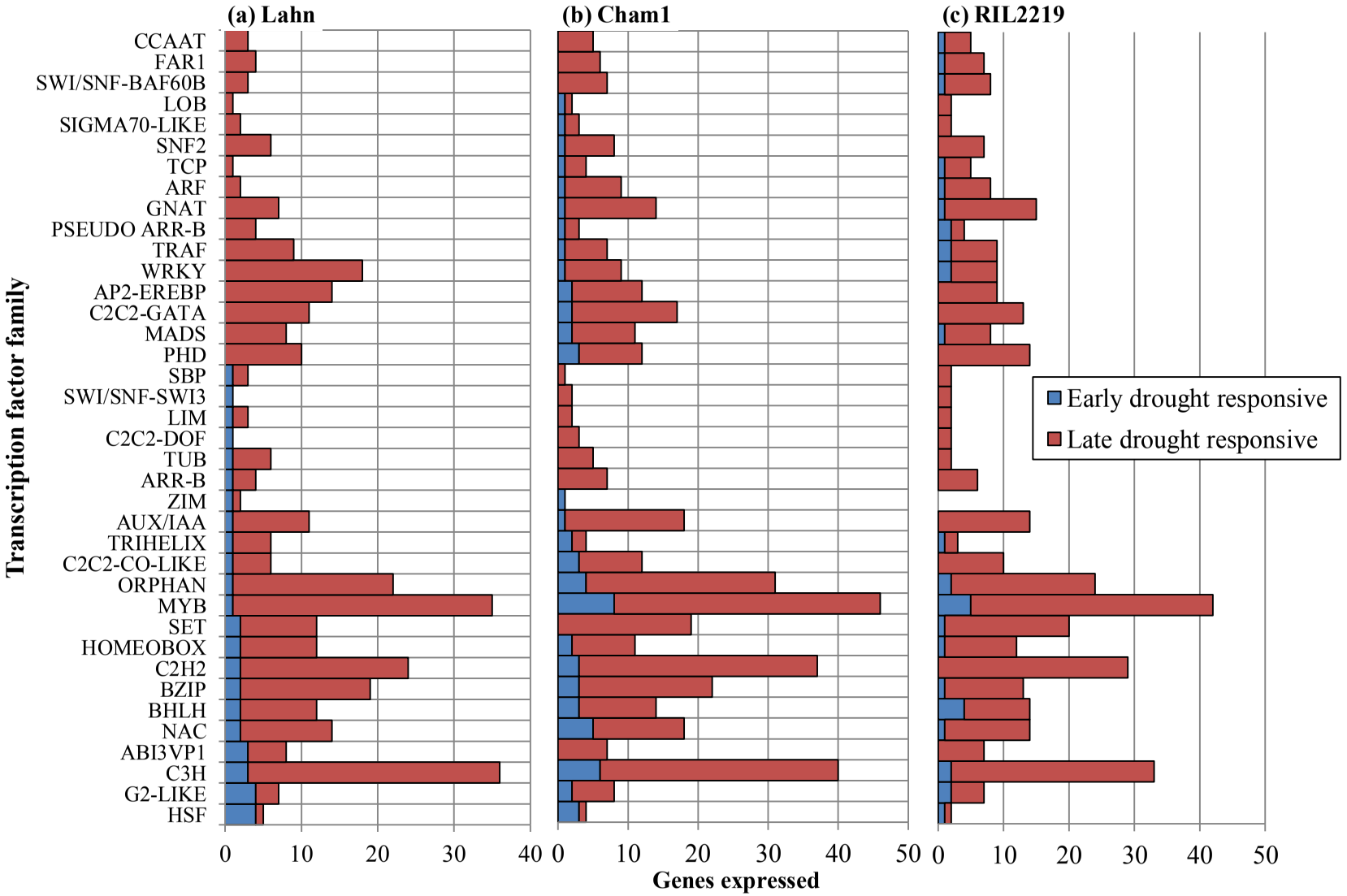


Figure 5.13: Transcription factors that change significantly over time, with a line.time interaction, categorised as early (<=2days) and late (>2days) responsive, and grouped by family. Only transcription factor families with more than 5 expressed transcription factors are shown.

5.6.5 Processes enriched in time-responsive genes

To summarise these results, the major enriched processes within significantly regulated time-responsive genes were broken down into six major categories. Figure 5.14 shows processes associated with the primary metabolic and genetic processes of the cell, which have been broken down into processes involved in the regulation of energy, carbon, and nitrogen metabolism; transcription; and translation. Figure 5.15 shows the processes involved in signalling; senescence and flowering; and general stress response processes. This figure was created by inspecting the GO *biological process* tree, derived from the annotation of significantly enriched genes (p-value < 0.05) at each time-point, and then generalising the more specific categories to provide a manageable overview of the data.

Figure 5.14 shows that photosynthesis genes are enriched in all but the third day of water-stress. This indicates that even during a mild water-stress, at an early time point, the transcriptome regulating photosynthesis is still significantly affected. Other processes related to energy storage and metabolism are enriched throughout the stress, with the exception of day three. Previous results have shown that the late regulation of ribosomal subunits is important in explaining the overall variation over time, and ribosomal biogenesis related genes are shown here to be significantly enriched during day 5 of stress. It is also apparent that the processes vital to transcription and translation are also significantly regulated throughout the progressive water-stress. This indicates a global reconfiguration of these mechanisms is initiated from the very early stress response, which culminates in the down regulation of primary ribosomal mechanism during the final two days of severe water stress.

Figure 5.15 shows the enrichment genes from of other important processes relating to signalling, senescence, and general stress responses. The presence of

four processes which are involved in the biosynthesis of ABA, ethylene, steroids, and trehalose during the first two days of stress, confirms that the plant is responding to mild water stress, and has initiated the regulation of compounds that mediate signalling in the plant. This early regulation of these processes is likely to impact the later sensitivity of the plant to stress, by effectively priming the signalling apparatus. Genes involved in processes related to the initiation of flowering and senescence are also significantly regulated from day one of the stress.

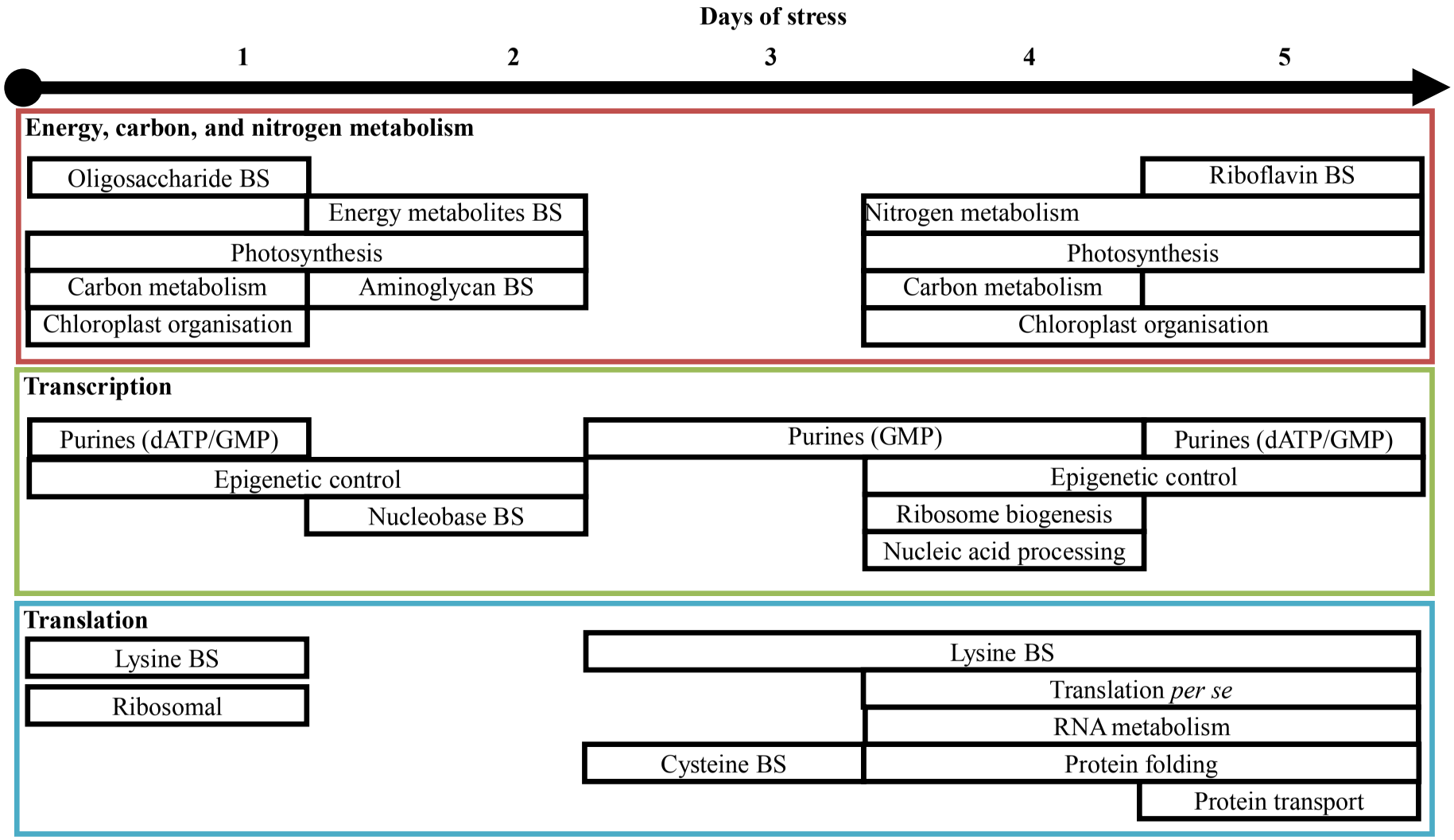


Figure 5.14: A summary of the major primary metabolism significantly enriched (p>0.05) processes that emerge over days of stress.

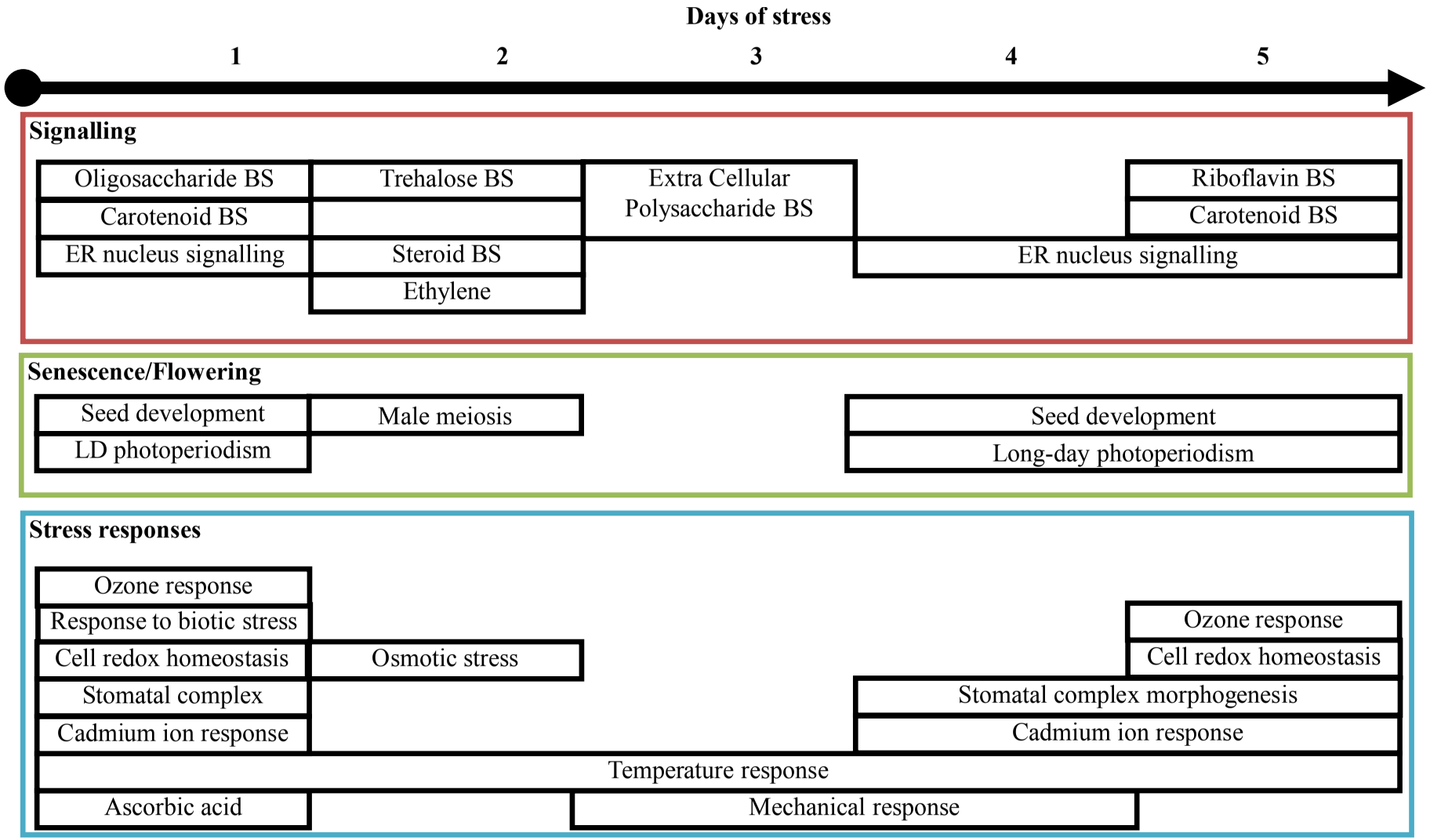


Figure 5.15: A summary of the physiological, signalling, and cross-talk significantly ($p>0.05$) enriched processes that emerge over days of stress. Abbreviation: biosynthesis (BS).

5.6.6 Hormonal control

Section 5.6.5 summarised the processes and signalling responses that were affected during the drought response time course identified by a gene enrichment analysis using all genes from the time-responsive ANOVA groups. Only ethylene biosynthesis genes were found in enough abundance for enrichment to be significant. However, within the line.time ANOVA group, which contains genes that have significant interactions between line and time, there are far more enriched hormonal processes. Table 5.14 shows significantly enriched hormonal processes in the line.time ANOVA group, compared to their abundance on the whole wheat GeneChip. Enrichment was performed on a per-day basis for those genes in that day that are significantly expressed compared to the control (change greater than the LSD at $p<0.05$). The methodology for calculating enrichment statistics has been described in Section 5.5.

In the ANOVA group from the Lahn cultivar there were no hormonal processes found to be enriched at any day of the time-course. However, in Cham1 jasmonic acid (JA) biosynthesis is significantly regulated during days 1-3 of the time-course. Within RIL2219 there was also early regulation of JA biosynthesis genes after one day of water stress, as well as Ethylene biosynthesis genes on day two, and Ethylene signalling on day three. Salicylic acid (SA) metabolism genes were found to be enriched at day three, and ABA signalling genes were enriched at day four. This points to the importance of hormonal regulation in differentiating RIL2219 from the other varieties, and highlights the pivotal role of hormones in regulating the drought response. Particularly interesting is the enrichment of ethylene pathway genes on day two in RIL2219, which were also enriched within the time responsive-only genes. This suggests that ethylene-mediated signalling is of pivotal importance in the drought response, and it differentiates the high-yield-stability RIL2219 from the other cultivars. The genes

involved in ethylene-mediated signalling are therefore a good source of candidate genes that could be suitable as gene-markers for selective breeding. The early regulation of JA biosynthesis, may also suggests that ethylene could be a candidate for regulating the later JA, SA, and ABA signalling response mechanisms.

This time series dataset clearly shows the global involvement of most hormone

Table 5.14: Enriched GO processes concerning plant hormones for genes within the line.time ANOVA group.

Day	RIL2219	Cham1	Lahn
1	JA biosynthesis (p=0.02)	JA biosynthesis (p=0.02)	None
2	Ethylene biosynthesis (p=0.07)	JA biosynthesis (p=0.01)	None
3	SA metabolism (p=0.01), Ethylene signalling (p=0.08)	JA biosynthesis (p=0.02)	None
4	ABA signalling (p=0.04)	None	None
5	None	None	None

groups in the wheat stress response. This is the first report of such a dissection of water stress response in wheat leaves post anthesis and the datasets are available for further mining. To demonstrate the uniqueness and utility of the time series approach, I have focused on the hormone ABA in the subsequent section.

Figure 5.16 shows an overview of the ABA biosynthesis pathway, together with the first step of ABA degradation. For the each enzyme in the pathway, the gene or genes on the wheat GeneChip, which correspond to the respective enzyme function (determined by the exact EC number), and appear within the line.time ANOVA category are reported. For each of these genes the main direction of their expression (if significant, p-value < 0.05) relative to the well-watered control is also given.

For the first step in the ABA biosynthesis pathway, catalysed by neoxanthin synthase (EC: 5.3.99.9), there were no sequences on the wheat GeneChip identified by CoPSA. For the second pathway step, catalysed by NCED, five can-

didate genes were found on the chip, of which TaAffx.76007.1.S1_at was found to have a line.time significant expression. In addition Ta.9731.1.S1_x_at and Ta.9731.2.S1_a_at were identified as having line+time significance. The remaining genes had no significant changes in expression over time; however one gene had significant variation across line only (TaAffx.59686.1.S1_at). For the second step in the pathway, catalysed by xanthoxin dehydrogenase (EC: 1.1.1.288), there was only one gene (TaAffx.29044.1.S1_at) matching this function on the wheat GeneChip, and this had a line.time significant interaction. For the third step in the pathway, catalysed by abscisic aldehyde oxidase (EC: 1.2.3.14), there were 9 genes on the wheat GeneChip matching this function. Two of these genes (Ta.6172.3.A1_a_at and TaAffx.92079.1.S1_at) had a line.time interaction, and the remainder had no significant changes in expression over time. Finally, the first degradation step of ABA catabolism, catalysed by ABA 8'-hydroxylase (EC: 1.14.13.93), had 5 genes on the wheat GeneChip matching this function. Of these only one had a line.time interaction. The remaining genes had no significant changes in expression over time; however one gene had significant variation across line only (TaAffx.48690.1.S1_at).

Figure 5.16 shows that the enzymes within the second (NCED), third (xanthoxin dehydrogenase), and forth (abscisic aldehyde oxidase (AAO)) stages of ABA biosynthesis were almost exclusively up-regulated from the third day of water stress. There were only two exceptions to this pattern: genes in, (a) the second stage of the pathway (NCED) expressed in RIL2219 at day three, and (b) the fourth stage (abscisic aldehyde oxidase (AAO)), which showed isoenzymes significantly expressed at day one, in the Lahn cultivar. It is impossible to quantify the effect of gene expression on the metabolites in the pathway from this data alone, given the complexities of post transcriptional modification, isoenzyme functional heterogeneity, and pathway flow dynamics. However, it could be speculated that the up-regulation of all the genes in the biosynthesis of ABA, may lead to an accumulation of ABA. This was further supported by the down

regulation of the first stage in the catabolism of ABA (ABA 8'-hydroxylase) at day 3 and 4. This certainly makes an interesting hypothesis for further biological validation.

Previous work had shown that xanthoxin dehydrogenase is not significantly regulated during water stress (Nambara and Marion-Poll, 2005). Figure 5.16 shows that within the TRITIMED data set, that this is not the case.

Figure 5.17 shows the expression of two genes on the wheat GeneChip that encode an NCED enzyme (corresponding LSDs provided in Table 5.15), which catalyses the second stage of ABA biosynthesis. These are constitutive responses to days of water stress, where the genes in each cultivar line are changing by the same magnitude and in the same direction, but the base expression level of the genes varies between cultivars. However, it is not possible to concretely link this observed difference in expression to a variance in true base expression level, as the effect is indistinguishable from systematic differences between lines which may affect the sensitivity of the probes to the quantity of a gene transcript (*e.g.* line polymorphisms). It seems probably however that a lack of interaction in these NCED genes, and the conserved regulation of expression throughout days of stress, in all three cultivars, means these genes are important as constitutive response to water stress. The late significant down-regulation (p-value<0.05) at the final day of water stress, is the inverse of the line.time interacting NCED enzymes, for which a significant up-regulation (p-value<0.05) was observed at days 3-5. This indicates the regulation of this step of ABA biosynthesis is additive, involving both constitutive and line specific regulation of isoenzymes over time.

Figure 5.18 shows for Lahn (drought susceptible), the expression over days of water stress, relative to the well watered control, for all line.time interacting genes in the ABA biosynthesis and initial degradation pathway previously shown in Figure 5.16. The LSD for comparing these relative expression values is given in Table 5.16. These relative expression values expand on the previous

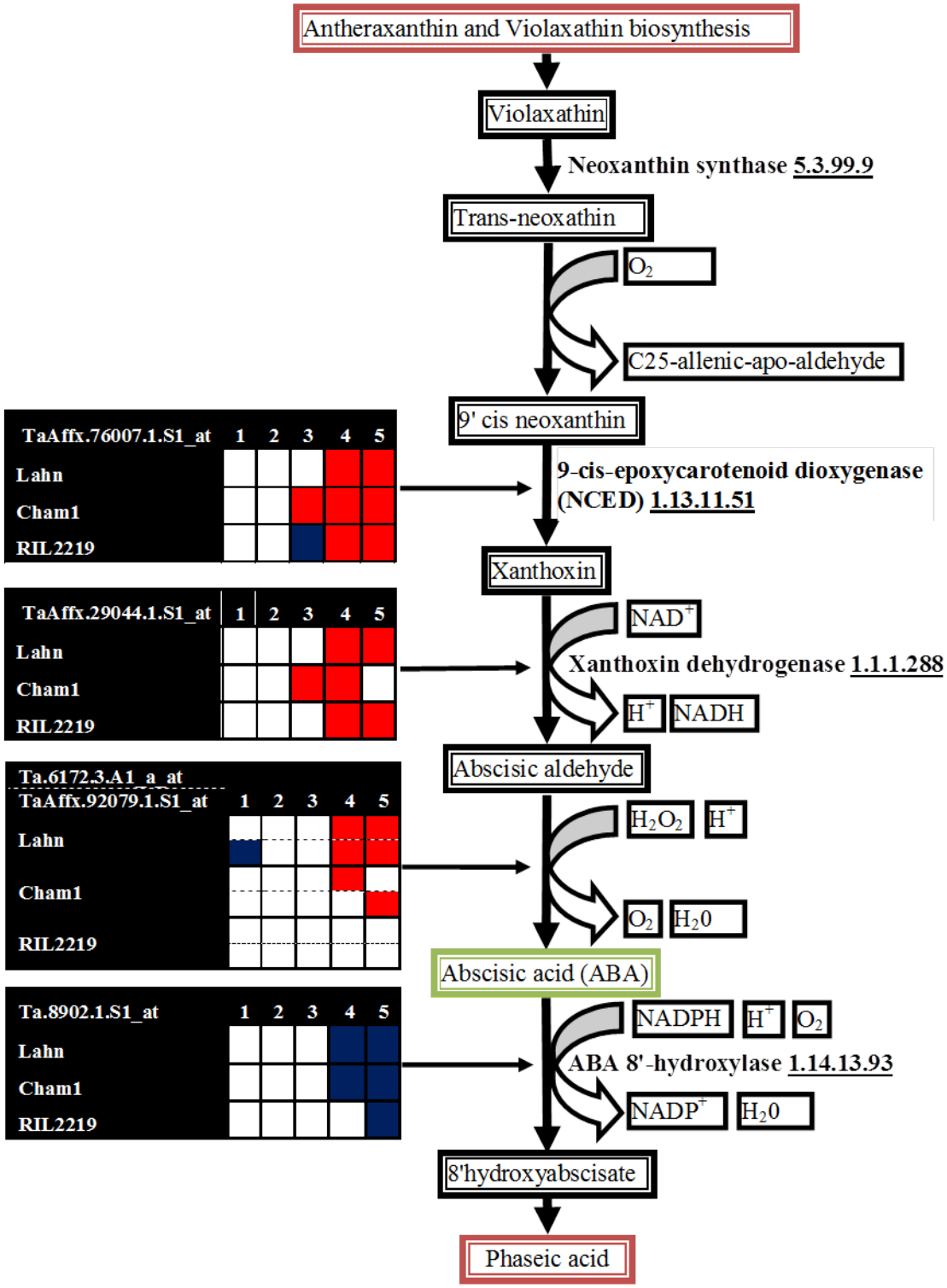


Figure 5.16: Metabolic steps within the ABA biosynthesis and catabolism pathway, with the direction of expression for line.time significant ($p<0.05$) enzymes identified by CoPSA reported. Red and blue indicate significant up and down regulation respectively, compared to the well-watered control. There are two isoenzymes reporting line.time significant expression for abscisic aldehyde oxidase (1.2.3.14), which are both reported.

Table 5.15: LSDs for comparing time-points to well-watered control for genes in ABA biosynthesis with a line+time interaction.

Enzyme	Gene	LSD
1.13.11.51 (NCED)	Ta.9731.1.S1_x_at	0.231129
1.13.11.51 (NCED)	Ta.9731.2.S1_a_at	0.201635

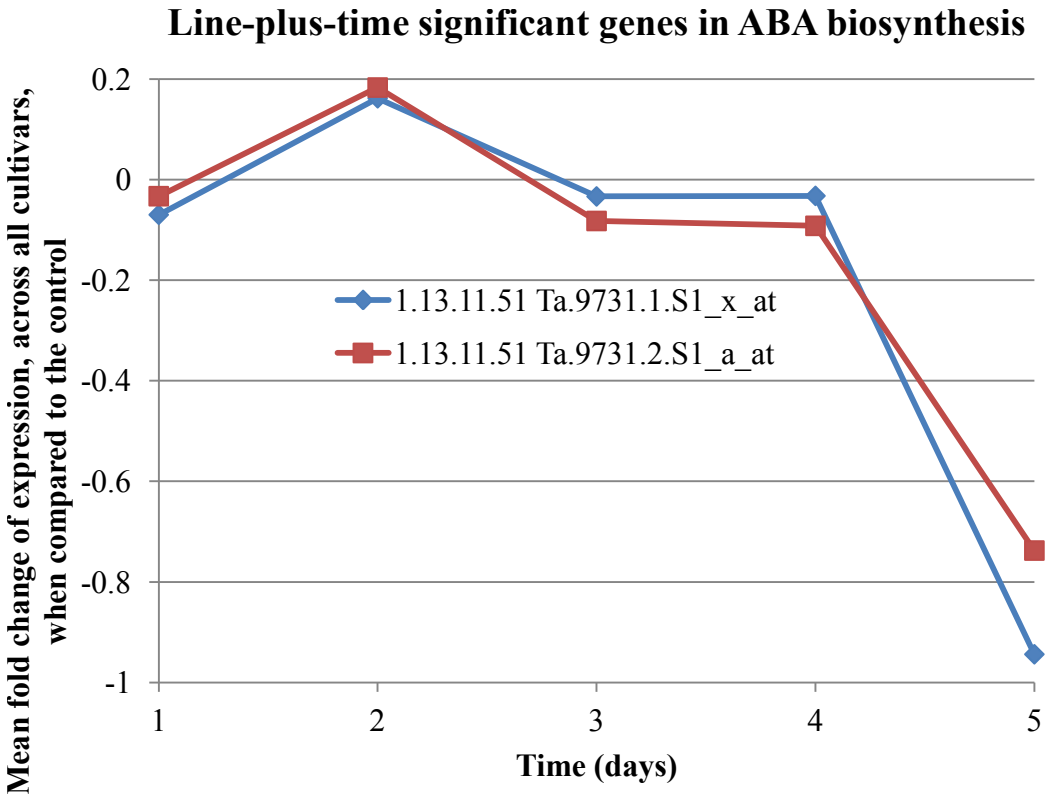


Figure 5.17: The mean expression of genes across lines, with line+time significance (p-value<0.05), that encode 9-cis-epoxycarotenoid dioxygenase (NCED) (EC: 1.13.11.51). The graph shows the gene variation in gene expression, over days of water stress, relative to the well-watered control. The least significant difference (LSD) relative to the control is given in Table 5.15

summary of direction of significant regulation, which were provided in Figure 5.16. It reveals that NCED has a greater magnitude of expression after 4 days of water-stress, compared to any of the later enzymes in the pathway. AOO also display a great magnitude of regulation at day four, however its magnitude of expression change relative to the control, is not as marked as NCED. It is difficult to assess the impact that this will have on ABA concentration, however previous work has indicated that the changes in expression of NCED and AAO are the most correlated with ABA (Nambara and Marion-Poll, 2005, Yang and

Guo, 2007). Therefore, it seems probable that the up-regulation of NCED and AAO in Lahn, after four days of water stress, will result in a reciprocal increase in ABA.

Figure 5.19 shows for Cham1 (drought resistant), the expression over days

Table 5.16: Least significant differences (LSDs) when comparing to control (p-value < 0.05), for comparing time-points to well-watered control for genes in ABA biosynthesis with a line.time interaction.

Enzyme	Gene	LSD
1.13.11.51 (NCED)	TaAffx.76007.1.S1_at	0.43
1.1.1.288 (Xanthoxin dehydrogenase)	TaAffx.29044.1.S1_at	0.32
1.2.3.14 (AAO)	Ta.6172.3.A1_a_at	0.09
1.2.3.14 (AAO)	TaAffx.92079.1.S1_at	0.27
1.14.13.93 (ABA 8'-hydroxylase)	Ta.8902.1.S1_at	0.23

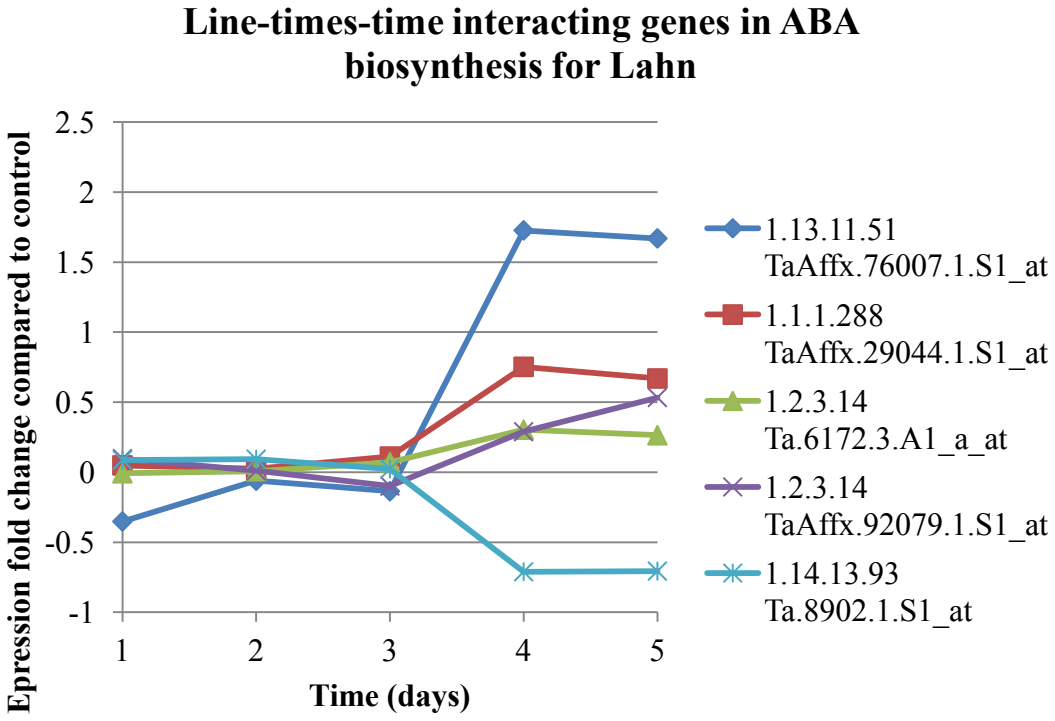


Figure 5.18: The expression of Lahn genes with line.time significance (p-value < 0.05), within the ABA pathway shown in Figure 5.16. The graph shows the gene variation in gene expression, over days of water stress, relative to the well-watered control. The least significant difference (LSD) relative to the control is given in Table 5.16.

of water stress, relative to the well-watered control, for all line.time interacting genes in the ABA biosynthesis and initial degradation pathway (LSDs are

given in Table 5.16). As with Lahn the expression of NCED and AAO genes have the greater magnitude of regulation, relative to the other enzymes, after four days of water stress. However, the regulation of on one of the AAO enzymes (Ta.6172.3.A1_a_at) at day five is not significant relative to the control, because of a down regulation. There is also a temporal shift compared to Lahn, with NCED and AOO genes showing significant expression a day earlier after only three days of water stress. Overall the regulation of these genes is more smoothly progressive than the dramatic changes in expression observed in Lahn.

Figure 5.20 shows for RIL2219 (drought resistant), the expression over days

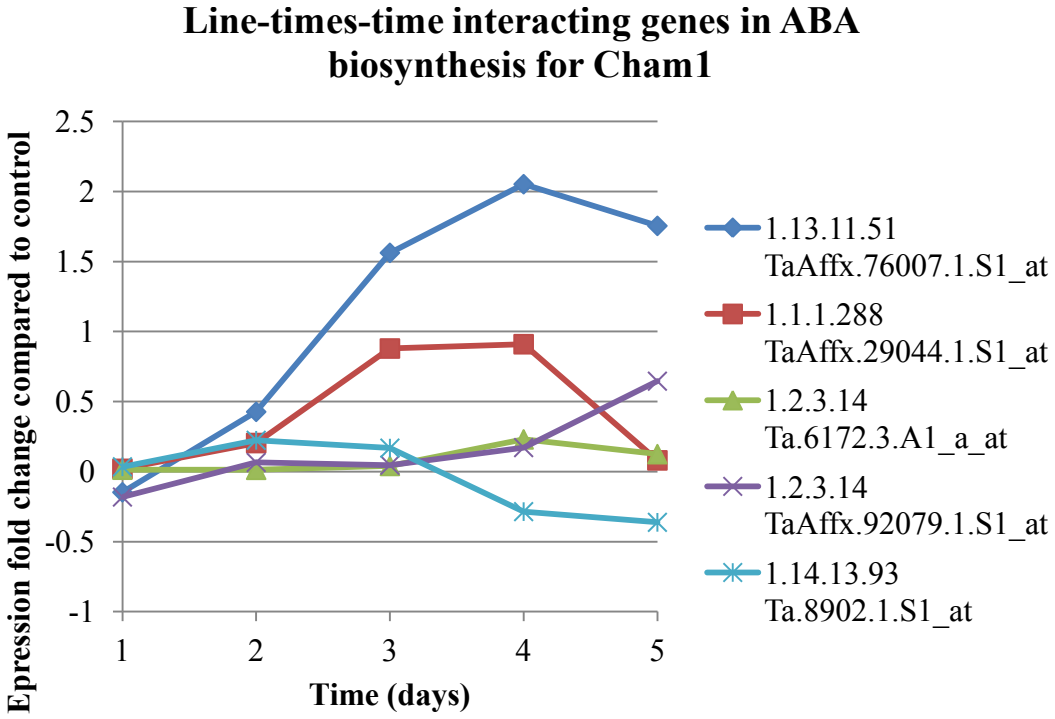


Figure 5.19: The expression of Cham1 genes with line.time significance (p-value<0.05), within the ABA pathway shown in Figure 5.16. The graph shows the gene variation in gene expression, over days of water stress, relative to the well-watered control. The least significant difference (LSD) relative to the control is given in Table 5.16.

of water stress, relative to the well-watered control, for all line.time interacting genes in the ABA biosynthesis and initial degradation pathway. NCED, as with Lahn is over expressed during the final two days of water stress. However, the difference between days 4 and 5 is more marked that Lahn or Cham1, and is

increasing in contrast to decreasing in the other cultivars. There are no significant expression changes in AAO over time, relative to the well-watered control. This could be indicative of a delayed increase in ABA regulation at days four and five. This is consistent with the lower RWC (Figure 5.1) observed in RIL2219, which indicates that RIL2219 has mitigated the fall in RWC enough to delay water stress, and the requirement to launch an ABA mediated response.

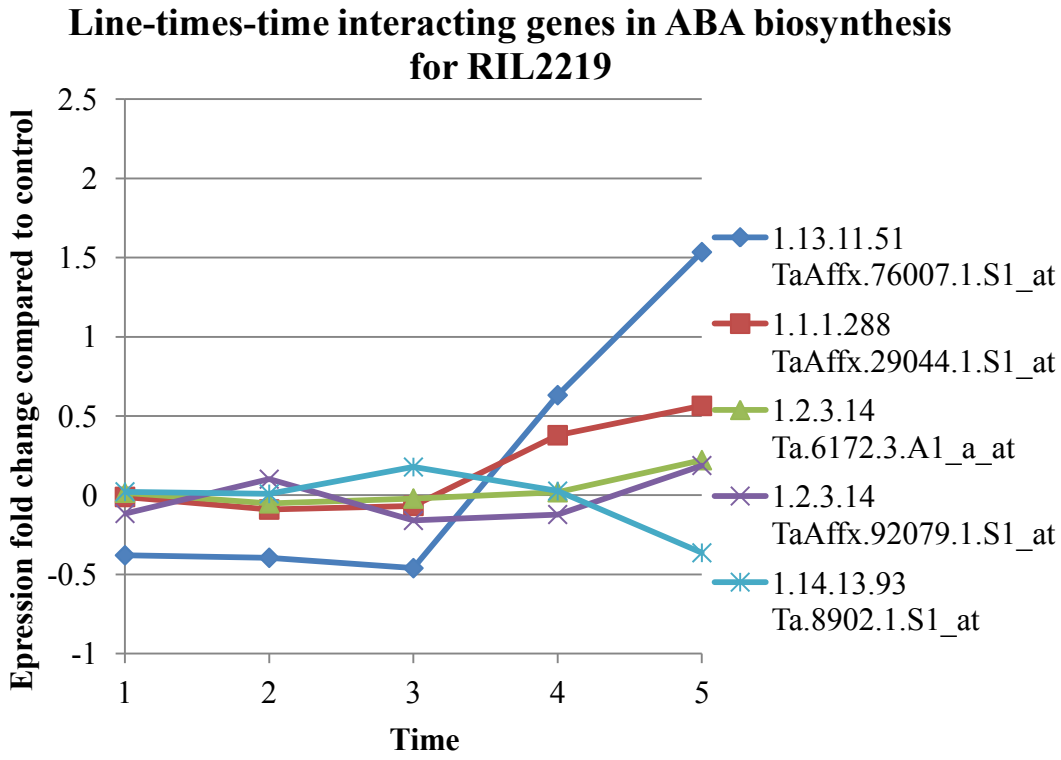


Figure 5.20: The expression of RIL2219 genes with line.time significance (p-value < 0.05), within the ABA pathway shown in Figure 5.16. The graph shows the gene variation in gene expression, over days of water stress, relative to the well-watered control. The least significant difference (LSD) relative to the control is given in Table 5.16.

ABA mediated signalling

As previously outlined in this section , as well as ABA biosynthesis, the transcription factor and protein-protein interaction mediated signalling is also integral to the ABA mediated drought response. Without the presence of an ABA receptor (RCAR), and the downstream components of the signalling pathway,

the regulation of ABA biosynthesis is ineffective at counteracting water stress (Ma *et al.*, 2009). Figure 5.21 shows a proposed model of the signalling pathway from osmotic stress to the control of stomatal aperture and transcription, based on current knowledge in this area. RCAR orthologs was identified on the wheat GeneChip according to the methodology described in Section 5.5. The direction of significant expression relative to the well watered control (p-value < 0.05) of 3 out of 5 RCAR genes on the wheat GeneChip that have line.time interaction are given. The probe-sets TaAffx.43193.1.S1_at, Ta.21082.1.S1_x_a, and TaAffx.109881.1.S1_at come from disparate parts of the RCAR family (Figure 5.22). There appears to be very little commonality in the directional pattern of their expression. There also appears to be large differences between the lines for TaAffx.43193.1.S1_at and TaAffx.109881.1.S1_at, however Ta.21082.1.S1_x_a is consistently down-regulated during day 4 and 5 in Lahn and RIL2219, and during days 2 to 5 in Cham1. However care should be taken with this probe-set as it is a mixed type, which may confuse expression of probes binding to other transcripts (*x* type probe-set, see Chapter 3.1.2(*a*) for an explanation of probe-set types), which may cross Out of the two remaining RCAR genes on the wheat GeneChip, TaAffx.131393.1.S1_at had line+time significant expression, and TaAffx.11433.1.S1_at did not register significant transcript binding, and was therefore excluded prior to ANOVA.

The gene expression relative to the well watered control, across days of stress, for the three RCAR genes with line.time interaction, in Lahn, Cham1, and RIL2219 is shown in Figure 5.23, Figure 5.24, and Figure 5.25 respectively. The Least Significant Differences (LSDs) corresponding to these genes are provided in Table 5.17. This expands on the brief summary of the direction of their regulation, previous provided in Figure 5.21. It is apparent from Figure 5.23 that most of the regulation for two RCAR genes (TaAffx.109881.1.S1_at and Ta.21082.1.S1_x_a) in Lahn are at days four and five. However a small but significant up-regulation in TaAffx.109881.1.S1_at occurs after only one day of stress, and then a much

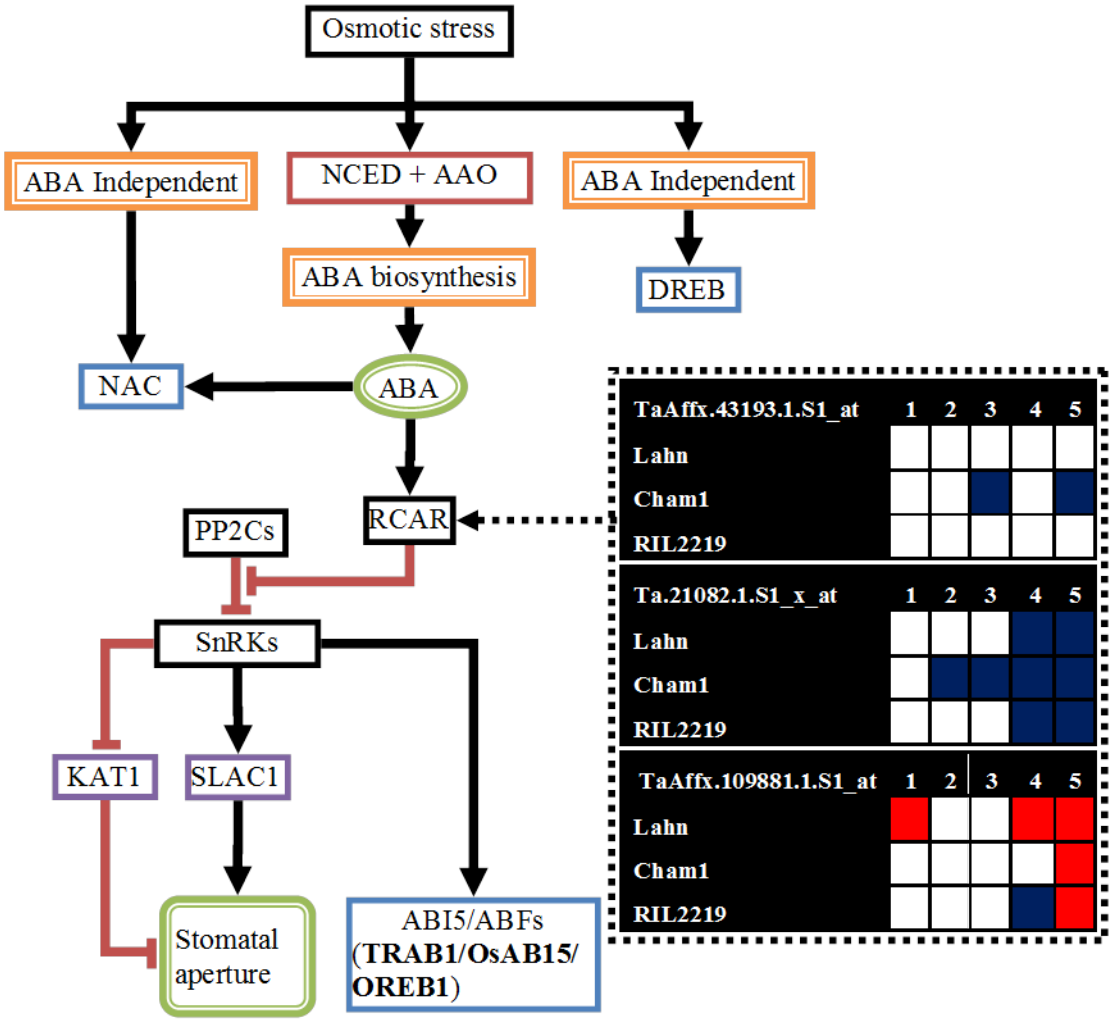


Figure 5.21: The role and direction of expression for line.time interacting RCAR genes within ABA water-stress signalling. Osmotic stress related signalling showing enzymes (red), transmembrane ion channels (purple), transcription factors (blue), and protein kinases (black). Red and blue indicate significant ($p<0.05$) up and down regulation respectively, compared to the well-watered control.

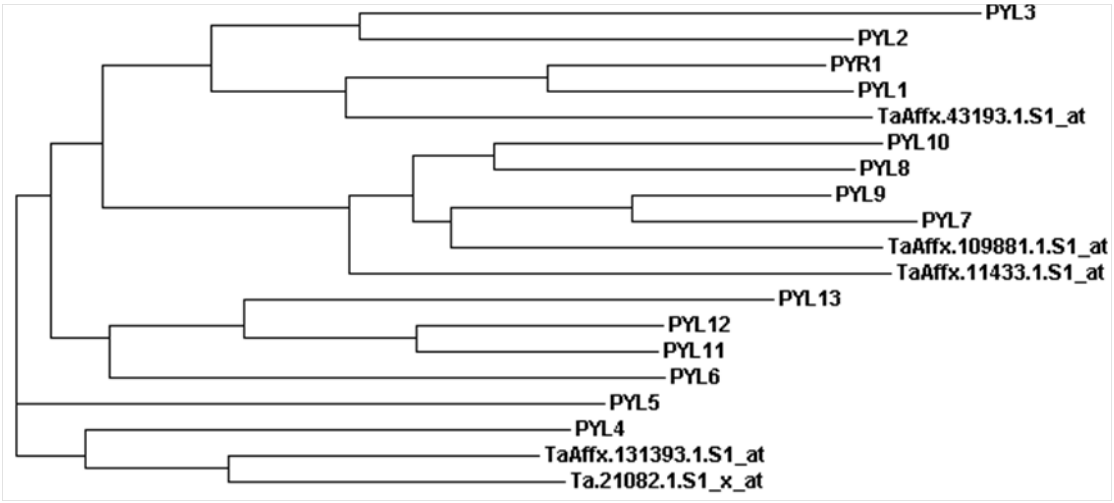


Figure 5.22: A phylogram showing the alignment of RCAR genes in Arabidopsis and orthologous sequences on the wheat GeneChip, using ClustalW2.

larger up-regulation on days 4 and 5. TaAffx.43193.1.S1_at has no significant regulation relative to the control, at any time point in Lahn.

The expression of Ta.21082.1.S1_x_a in Cham1 is down regulation (Figure

Table 5.17: Least Significant Differences (LSD) values (p-value < 0.05), (1) for the comparison of the gene expression recorded for a day to the well-watered control, and (2) for inter-comparison of expression between days. Significance p-values at 0.05 and 0.01 are provided for each of these comparison types.

Gene	LSD
TaAffx.43193.1.S1_at (PYL1)	0.12
Ta.21082.1.S1_x_at (PYL4)	0.14
TaAffx.109881.1.S1_at (PYL9)	0.13

5.24), as was the case in Lahn (Figure 5.23); however expression is reduced sooner, after only two days of stress, and recovers during days 4 and 5. In contrast to Lahn, TaAffx.43193.1.S1_at is significantly regulated at days 3 and 5 of water stress, and modulates between small-insignificant and significant down-regulation during days 2 to 5. Although the quantity of regulation in Lahn was not significant, relative to the control, the reduced up-regulation at days 2 and 4, compared to the peaks at days 3 and 4, is similar in the Cham1. The Cham1 however displays this expression pattern at a lower expression level than the well-watered control.

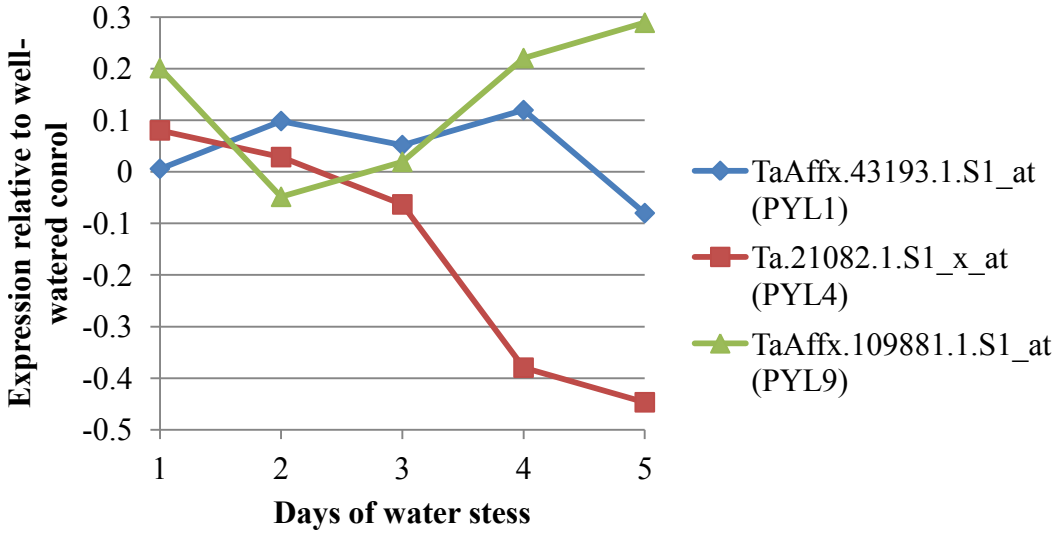


Figure 5.23: The expression of line.time significant genes in Lahn, for genes that are strongly orthologous with members of the RCAR family.

Overall the TaAffx.43193.1.S1_at and Ta.21082.1.S1_x_a RCAR genes both show a greater quantity of regulation at day 3 relative to the control, in a strong co-ordinated down-regulation. This is despite there being only minor difference in RWC at day 3 (Figure 5.1). This may suggest an early desensitisation of ABA-mediated-signalling in Lahn through RCAR down-regulation, at day three.

It might be expected that *if* the early regulation of RCAR genes in drought

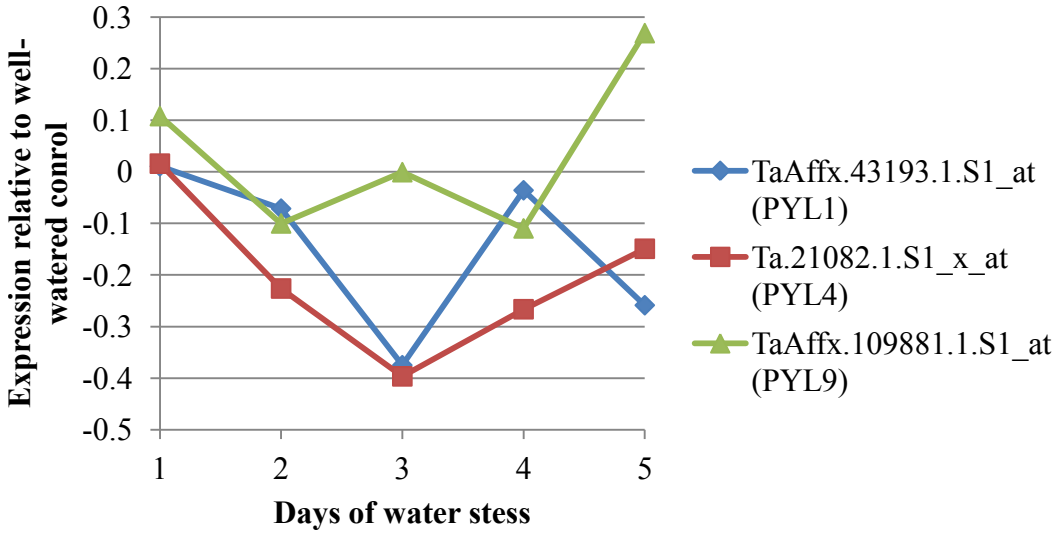


Figure 5.24: The expression of line.time significant genes in Cham1, for genes that are strongly orthologous with members of the RCAR family.

resistant Cham1 (Figure 5.24) acts to desensitise the plant to ABA mediated signalling, then the highly drought resistant RIL2219, would exhibit the same

early regulation. However, the opposite appears to be the case (Figure 5.25). Like Lahn TaAffx.43193.1.S1_at has no significant regulation relative to the control, at any time point. As with Lahn, TaAffx.109881.1.S1_at is not regulated until day 4 and 5 of water stress. The direction of magnitude of day five is highly similar to Lahn, however RIL2219 exhibits strong a down-regulation day four (Figure 5.25), unlike the strong up-regulation in Lahn (Figure 5.23). However, this can be accounted for by the overall delay of RWC fall in RIL2219. Lahn also exhibited a down regulation, much earlier at day one (Figure 5.23), but this was small and insignificant. The down-regulation of the expression of Ta.21082.1.S1_x_a at day 4 and 5, in RIL2219 is very similar to that observed in Lahn.

The expression of RCAR genes appear to modulate over time, with no clear

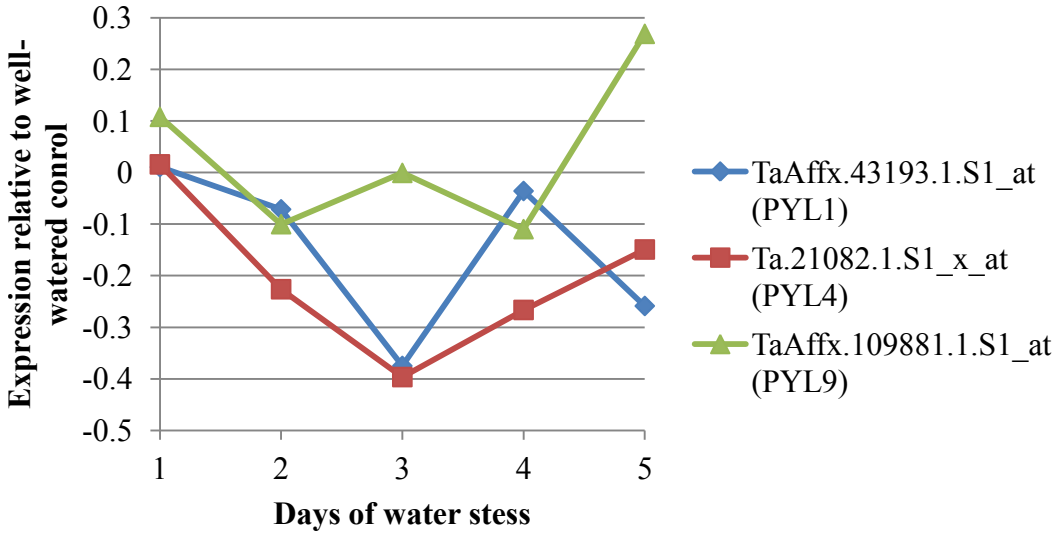


Figure 5.25: The expression of line.time significant genes in RIL2219, for genes that are strongly orthologous with members of the RCAR family.

linear relationship with time or RWC. It seems apparent that their regulation is as a result of a complex function. They may act multiplicatively to regulate ABA sensitivity for multiple GO *biological processes* and *cellular components*. Overall, Lahn and RIL2219 exhibit late regulation of Ta.21082.1.S1_x_a, with very similar profiles. Cham1 is unique in exhibiting significant regulation of TaAffx.43193.1.S1_at, across time.

There was no specific EC category for PP2Cs. However the EC category 3.1.3.16, which describes the function of phosphoprotein phosphatase, encompasses PP2Cs, as well as the closely related PP1s, PP2As and PP2Bs. CoPSA identified 120 of these genes on the wheat GeneChip, of which no expression was detected for 72. There was 31 of these genes with line.time interactions, 12 with only line+time significant regulation, 5 with significant differences between days, and 1 with significant differences between lines. None of the genes with detected expression, were not found within these ANOVA groups (*i.e.* none had emphno effect from the line or time variables).

The expression above significant (p-value < 0.05) of the 31 phosphoprotein phosphatases (PPs) (3.1.3.16) which have a line.time interaction is given for Lahn, Cham1 and RIL2219 in Figure 5.26, 5.27 and 5.28, respectively. A similar trend to that observed in RCAR emerges, with Lahn showing significant up and down regulation of PPs during days 4 and 5 of water stress. A small number of PPs are unregulated in RIL2219 at day 4, and most are unregulated during day 5. As with RCAR genes, Cham1 shows significant regulation of PPs during days 3, 4, and 5. Together with the observations of regulation of RCAR genes, this observation points to an early regulation of components of ABA mediated signalling in Cham1. As Lahn and Cham1 are the most synchronised in terms of RWC, it is likely that Cham1 is more responsive in regulating components of ABA-mediated signalling, in order to better respond to the fall in RWC. The similarity in expression of RCARs and PPs between Lahn and RIL2219, despite very different RWC levels, indicates that other adaptations in RIL2219 have mitigated the fall in RWC to such an extent that the early RCARs/PPs response in Cham1, has been delayed to days 4 and 5. This suggests that an earlier priming of the RCAR/PP2C may be part of the reason Cham1 and RIL2219, have improved yield stability in drought conditions.

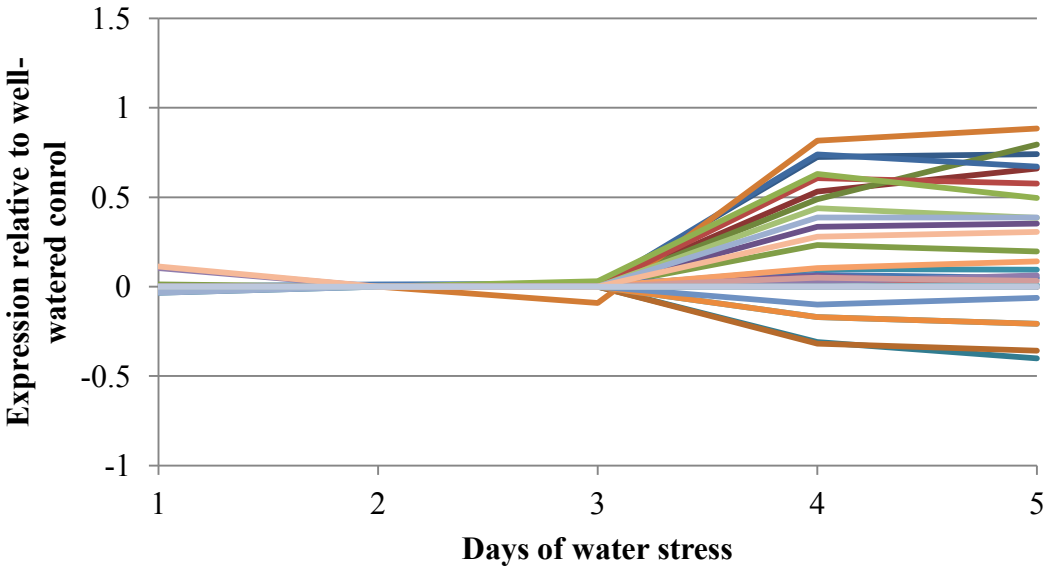


Figure 5.26: The expression above significant (p-value < 0.05) compared to the well-watered control, for 31 phosphoprotein phosphatase genes (EC: 3.1.3.16) in Lahn which have a line.time interaction.

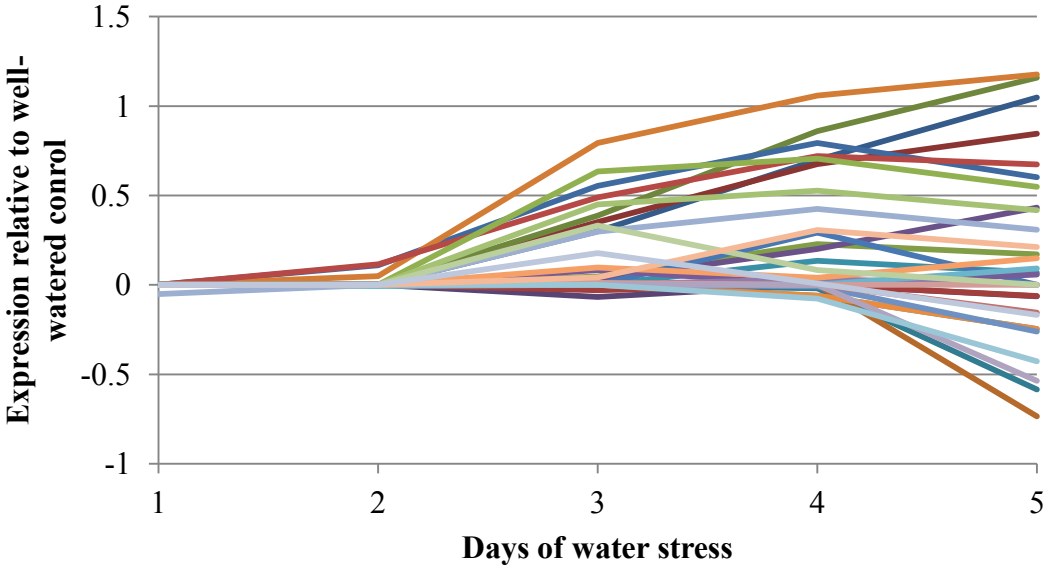


Figure 5.27: The expression above significant (p-value < 0.05) compared to the well-watered control, for 31 phosphoprotein phosphatase genes (EC: 3.1.3.16) in Cham1 which have a line.time interaction.

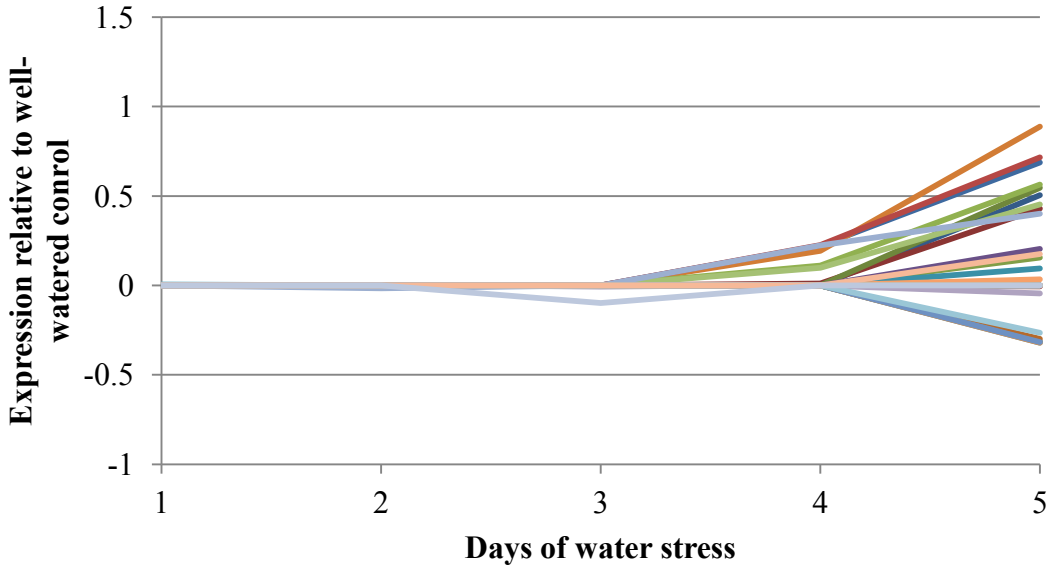


Figure 5.28: The expression above significant (p-value < 0.05) compared to the well-watered control, for 31 phosphoprotein phosphatase genes (EC: 3.1.3.16) in RIL2219 which have a line.time interaction.

5.6.6(*a*) Transcriptome responses to osmotic stress

The accumulation of proline is a known response to osmotic stress, and prevents shifts in redox potential. Cell redox homeostasis was observed to be an important process in early and late responses in time significantly regulated genes. Table 5.18 shows that proline metabolic process genes were enriched within the line.time ANOVA group for the RILL2219 and Cham1 cultivars. Metal-ion related processes play a role in signalling, maintaining osmotic potential, synthesis of metal containing compounds, and preventing toxic build-up of metals. Table 5.19 shows that a large number of metal ion related processes were enriched within the line.time ANOVA group. Cadmium was not

Table 5.18: Osmolyte enriched processes in the line.time group

Day	RIL2219	Cham1	Lahn
1			
2	Proline biosynthesis (p=0.04)		
3		peptidyl-proline hydroxylation (p=0.03)	
4	Proline metabolism (p=0.07)		
5			

present in the environment of the plants, so it seems probable that the Cd^{++} responses observed are a result of general metal-responsive genes. K^+ transporters such as KAT1 are important for control of stomatal aperture (Sato *et al.*, 2009), and their role in the ABA mediated water stress response has been described in Chapter 6.1. There enrichment during days 4 and 5 in RIL2219, 2 and 3 in Cham1, and 2 and 5 in Lahn indicate they are important throughout mild and severe water stress. The overabundance of K^+ transporters like KAT1 in the stomata is likely to reduce the sensitivity of the ABA mediated response. However further verification of KAT1 orthology and localisation would be required to further elucidate the mechanism here. Enrichment of these transporters highlights an important area for further work.

Table 5.19: Metal ion enriched processes in the line.time group.

Day	RIL2219	Cham1	Lahn
1	response to Zn^{++} (p=0.07)	response to Zn^{++} (p=0.01)	Na^+ transport (p=0.09), Zn^{++} transmembrane transport (p=0.08)
2		K^+ transmembrane transport (p=0.06)	K^+ transmembrane transport (p=0.04), Cd^{++} response (p=0.0005)
3		K^+ import and cellular transport (p=0.07)	
4	K^+ transmembrane transport (p=0.009)		Cd^{++} response (p=0.003), response to Zn^{++} (p=0.06)
5	K^+ transmembrane transport (p=0.04), Cd++ response (p=0.005)		K^+ transmembrane transport (p=0.006), Cd^{++} response (p=0.0008)

5.7 Conclusions

This chapter has shown that improved annotation using CoPSA has allowed a broad systems analysis of the TRITIMED durum wheat drought-response time-course microarray experiment and enabled dissection of particular pathways and candidate gene families. The uniqueness of the dataset has allowed the partition of the wheat leaf water response to early, intermediate and late phases of stress. Coupled with statistical models it allowed the identification of line specific responses and suggest candidate pathways and genes for future new studies. A number of novel biological insights were discovered as a consequence.

A high level of structure is present in the wheat leaf transcriptome response to water stress across all lines examined. A biological system in which regulation had broken down would exhibit a random expression of genes, and we would not expect it to exhibit enrichment of processes, across time-points and between lines. The high degree of order within the regulation and processes affected would indicate a controlled and progressive change in the plant genome and physiology. Whilst previous studies and reviews on early signalling have proposed a progression from sensing to final physiological effect (Gregersen and Holm, 2007, Shinozaki and Yamaguchi-Shinozaki, 2007, Zhou *et al.*, 2007), none have dissected the transcriptome transient to so many time points, identified time-specific components and further tested these changes across different wheat lines.

The combination of transcriptome dissection by statistical (ANOVA, PCO) and functional analysis (CoPSA annotations) facilitated the dissection of responses into various subgroups. One subgroup of ANOVA genes were not responsive to stress and these can be used as controls for further water stress studies (Section 5.6.1: 1,201 genes with line only or no significance).

PCO analysis was performed, to elucidate principle components of variation in gene expression for line and time, and line and RWC independent variables. CoPSA annotations were used to better understand the processes and functions that underlie the genes that contribute most to the top three PCos. These top contributing genes were identified through regression analysis. This revealed that the main contributor to variation across time (captured by PCo1) were late down-regulated ribosomal subunits, responsible for protein synthesis. This is in line with what others have observed (Novoa *et al.*, 2003), and it acts to reduce stress on the cell and prevent abnormal protein folding (Harding *et al.*, 2000). PCo2 and PCo3, which mainly highlighted variance between lines, contained processes that were related to metabolism, signalling and regulation of transcription. This highlights the importance of genes from these processes as candidates for breeding for improved yield stability in drought conditions.

Another group of ANOVA genes showed responses to stress across time (Section 5.6.1: 26,560 genes with time, time+line or time.line significance), these were used to build an overview of pathways, processes that were responsive to stress. This analysis was later expanded to elucidate transcription factor families that were early and late responsive in terms of time and RWC. Enrichment analysis of these time responsive genes was used to build a high level overview of water-stress responsive processes in Section 5.6.5. Figure 5.14 and **??** provided a visual summary of processes regulated by significantly enriched genes during each day of water stress. They demonstrated that the transcriptome response to stress is global, encompassing many processes, many of which regulate primary processes of transcription, translation and energy metabolism. It also showed that these processes are regulated from only one day of water stress, which indicates the plant is sensing the stress at this stage, and launching a progressive global regulation of processes. This type of progressive and global regulation from a mild to a sever water stress has not previously been shown. An overview for time responsive transcription factor families was also

provided. These time responsive processes provided potential genes for further biological verification as candidate genes for germplasm improvement.

The most interesting group of genes from ANOVA for wheat breeders was the line.time group (4,621), because these contained genes that were responding to water-stress but were differentially regulated between cultivars. There were the most likely to reveal genes responsible for inferring yield stability under drought in then Cham1 and RIL2219 cultivars. This group revealed line specific early and late responses and progressions. An overview for line.time responsive transcription factor families was provided in Section 5.6.4.

Transcription factor analysis revealed that a large number of transcription factor families were expressed in response to water stress (Section 5.6.4). Most of these were expressed at days 3, 4 and 5, however a large number were significantly regulated at days 1 and 2, at mild water-stress. This strongly indicates that transcriptional control is highly regulated and progressive from early stages of water-stress. The usual transcription factor families that are well studied in relation to drought were present in the data (Table 5.10). Particularly interesting was the early regulation of 24 NAC genes during the first two days of water-stress in RIL2219, which was more abundant that the two NACs expressed during these time points in Lahn and Cham1. NACs are known to regulate hormone signalling in a time dependent manner (Jensen *et al.*, 2010), and regulate senescence (Guo and Gan, 2006), which appears to be delayed in RIL2219. NACs also appear as an important family within the line.time ANOVA group. These are therefore strong candidates for explaining the RIL2219s superior yield stability under drought. Another interesting result were HSFs, of which 6 were significantly expressed differentially with time in Lahn and Cham1, however only 3 were expressed in RIL2219. HSFs were almost exclusively significantly down-regulated during days 1 and 2. Their role in regulating HSPs which assist in protein folding under heat and osmotic stress (Hu *et al.*, 2009), may be explained by the previous findings that protein synthesis is the

most regulated and enriched process in PCo1. If protein synthesis is down-regulated then HSFs are no longer required. However, their regulation after only one day of stress, before protein regulation is down-regulated is intriguing, and may indicate some more complex and previously unknown role in signalling. Some HSFs are known to responsive to senescence (Breeze *et al.*, 2008), their absence in RIL2219 is supportive of a model of delayed senescence in this cultivar. MYBs have been previously shown to be ABA inducible transcription factors (Abe *et al.*, 2003), and these where one of the most abundant families at late time points.

In addition to transcriptional factor signalling, in Section 5.6.6 hormonal processes were observed to be significantly enriched in RIL2219 and Cham1 within the line.time ANOVA. Jasmonic acid biosynthesis was enriched at early time points in both these cultivars. Jasmonic acid has recently been shown by Shan and Liang (2010) to regulate glutathione during water stress in plants, which is an antioxidant, protecting the plant from oxidative stress. Ethylene biosynthesis and signalling was enriched during day two and three respectively in RIL2219. Ethylene, as well as regulating growth and development, has been associated with the regulation of senescence and photosynthesis related genes (Grbić and Bleecker, 1995).

Given the importance of ABA in the water-stress response, a detailed analysis of genes involved its biosynthesis and downstream signalling was conducted in Section 5.6.6. An intriguing up-regulation of upstream genes encoding enzymes in the ABA biosynthesis, and down-regulation of downstream genes was observed. There were important cultivar differences, with Cham1 exhibiting an earlier up-regulation of upstream genes. The expression profiles of Lahn and RIL2219 appeared to be similar for this pathway. Previous work had shown that xanthoxin dehydrogenase is not significantly regulated during water stress (Nambara and Marion-Poll, 2005). The TRITIMED data set shows this is not the case at least in wheat, and the gene encoding xanthoxin dehydrogenase is sig-

nificantly up-regulated at days 4 and 5 in Lahn and RIL2219, and days 3 and 4 in Cham1.

Downstream ABA signalling was also considered in Section 5.6.6. The RCAR and PP genes which are important in forming the ABA-receptor complex, both demonstrated progressive increases in regulated over time. In both cases, up and down regulation was observed, indicating the regulation of these important signalling genes may be additive or multiplicative (*i.e.* multiple co-regulated genes acting together to regulate the water stress response). This supports the observation of Ma *et al.* (2009) that multiple knockouts of RCAR genes is required to induce stomatal-aperture insensitivity ABA. The enrichment of K^+ transmembrane transporters (Section 6.3.4), may indicate a downstream regulation of ABA-mediated signalling of stomatal closure (Sato *et al.*, 2009), the variation of expression of which may lead to stomatal insensitivity to ABA.

The utility of improved CoPSA annotation in understanding a complex transcriptome has been demonstrated throughout this chapter. They are of particular power when combined with statistical dissections and tests such as PCO, ANOVA, and enrichment analysis. Through improved annotation a potentially daunting 26,560 time responsive genes were categorised into their respective processes (Figure 5.14 and Figure 5.15). Through CoPSA annotation it was possible to further interrogate these data by asking specific question about transcriptional families (Section 5.6.4) and hormonal control (Section 5.6.6), which are known to be important to the water-stress response in plants. A high-level system wide view of the data-set, together with the power to drill down to details, is facilitated through structured-annotation with improved coverage and specificity. This improved understanding of this transcriptome data set has yielded a number of water-stress related gene targets for breeders. This information has now been transferred to the durum wheat breeders at CGIAR centre ICARDA for further validation and utility.

5.8 Further Work

The example of CoPSA enabled analysis, presented in this thesis, is limited to the protein functions and the metabolic functions they are involved in. Wheat breeders are primarily interested in plant traits, which are a consequence of many plant interacting processes (which are regulated over time, and between tissues). Quantitative Trait Loci (QTL) statistically relate the overall contribution of a genomic region to a trait of interest. There are many existing water stress QTLs for durum wheat (Habash *et al.*, 2009), and relating these regions to the biological processes summarised in this chapter would be a valuable contribution to understanding the molecular mechanism behind the trait, and could provide candidate marker genes in breeding. For non-model organisms, where there is not a complete genomic sequence, calculating the genomic location of genes within QTL coordinates is challenging. ESTs which have been mapped to chromosomal bins (Qi *et al.*, 2004, Conley *et al.*, 2004, Peng and Lapitan, 2005) could be used to propose candidate genes within QTL regions.

Recently, classical QTL analysis has been combined with transcriptome analysis through expression QTL (eQTL) studies. These work by measuring both gene expression and genetic variation in a large number of individuals. Statistical genetic methods, traditionally used in QTL analysis, can then be used to link genetic differences (associated to traits) to individual differences in expression. Nicolae *et al.* (2010) has shown these eQTLs, linked to traits, can also correspond to SNPs which are important for that trait. This directly feeds into marker assisted breeding strategies, which can then select for SNPs associated with desirable traits. The interaction between loci, which gives rise to a complex trait can also be inferred (Michaelson *et al.*, 2009). eQTLs are therefore a powerful tool in understanding the importance of expression of individual genes with respect to complex traits.

A further limitation of the TRITIMED transcriptome dataset, described in this chapter, is that only a subset of the transcriptome is measured by the wheat GeneChip. Wan *et al.* (2008) puts the coverage of the transcriptome measured by the wheat GeneChip at around 50% of genes in hexaploid wheat. The ever decreasing cost, and increasing read length, of sequencing technologies like Roche 454, makes this a more viable option for the futureCoram *et al.* (2008). Theoretically, transcriptome sequencing can achieve a 100% coverage of a transcriptome, although the large size of the wheat transcriptome makes this expensive.

Chapter 6. Summary conclusions and further work

6.1 Conclusions

This thesis was motivated by a need to improve annotation on the wheat GeneChip. The scarcity of annotation on the wheat GeneChip provided through Affymetrix NetAffy pipeline (Liu *et al.*, 2003) and Blast2GO-FAR (Escobar, 2011), potentially limits the analysis that can be done with transcriptomics data set. An incomplete annotation of the GeneChip presents the danger of a limited analysis that is unrepresentative of the true biological system, where genes with significant variation are simply unexplained.

Improved gene annotation coverage and specificity was achieved through a novel pipeline strategy presented in Chapter 3, which leveraged data integration methodologies developed in Chapter 2. Many existing sequence annotation pipeline tools like BLAST2GO (Conesa and Götz, 2008, Conesa *et al.*, 2005) leverage single data sources, that were used in annotation transfer, which was based on sequence alignment to identify putative functional-orthologs. Chapter 3 showed that the novel data integration approach incorporated into CoPSA improved the quantity of sequences on each Affymetrix plant GeneChips that could be annotated (coverage). It also improved the specificity with which genes could be annotated. This benefit was as a result of aggregating unique annotations from multiple data sources, as well as indirect inferences which revealed knowledge, some of which was not present in a single data source. CoPSA also incorporated a conjoint analysis that utilised both gene sequence alignment against proteins, and the identification of protein domains. Both of these approaches were shown to benefit from data integration

approaches.

As a consequence of these benefits derived from a novel data integration and conjoint methods approach, comparisons to NetAffx and BLAST2GO and pipelines showed that CoPSA was able to uniquely annotate genes that accounted for 34%, 24%, and 32% of the wheat chip for GO molecular function, biological process, and cellular component categories respectively. This represents a real benefit in terms of novel annotation, which was key to the success of the analysis presented in Part II.

The usefulness of any annotation is highly dependent on its quality. Benefits in terms of coverage and specificity of annotation must be balanced against the trustworthiness of annotation. Chapter 3 also showed that this increase in the coverage, quantity, and specificity of annotations can also be translated into improved quality of annotations. Comparison to wheat GeneChip annotations from NetAffx and BLAST2GO pipelines revealed that CoPSA makes very similar predictions to the high quality NetAffx pipeline. Metrics assessing the properties of the CoPSA annotations revealed that they show similar characteristics to other pipelines. This adds confidence to the novel CoPSA annotations that are not covered by other pipelines. Two novel filtering strategies for selecting high quality annotations from the CoPSA pipelines were proposed, and these were shown to perform better or similar to other standard filtering strategies. The success of these filtering methods demonstrated the benefit of recording provenance of annotation, which can be later used to derive confidence metrics. Provenance retained by the Meta-data based Graph Query Engine (MGQE) presented in Chapter 2, is a novel feature of a sequence based annotation pipeline, and was leveraged in the final annotation selection metrics in Chapter 3.

The transcriptome data set and its analysis, which was a key use-case to demonstrate the utility of the CoPSA annotation system, was presented in Chapters and . Regulation of *biological processes* were observed in the TRITIMED time-

series microarray experiment which were consistent with current knowledge in the field. This both demonstrated the utility of the current CoPSA annotation, in confirming existing models, and confirmed the quality of the annotations. A high error rate in the annotations would not have yielded observations of enriched processes which confirm current biological knowledge.

As well as confirming existing knowledge in the field of water-stress, analysis of the TRITIMED data set yielded novel biological insights. These concerned observations of processes and functions which where responsive to water stress, and differences between lines which may contribute to an explanation of the high yield stability under drought of the Cham1 and RIL2219 lines. Both observations yielded novel genes, which could potentially be used in targeted breeding of drought resistant cultivars. They could form candidate genes for investigating existing QTL loci in durum wheat (Habash *et al.*, 2009), which could help relate these genes to cultivar traits, and explain the molecular mechanism of the loci.

Combining statistical dissection and testing with CoPSA structured annotation, proved to be a powerful combination in interpreting the TRITIMED dataset. Principle coordinates analysis revealed three Principal Coordinates (PCos) that captured the main trends of variation in the gene expression. The first of these PCos captured the main expression changes across time, which was common to all three cultivars. It therefore represented constitutive responses to water-stress. Annotation of the genes with CoPSA data revealed that a large amount of variation could be explained by the late down-regulation of ribosomal subunits. This indicated that the down regulation of protein synthesis to prevent protein with abnormal folding and reduce stress on the cell is one of the main constitutive responses to water-stress. The observation of down regulation of Heat Shock Factors (HSFs) during early and late water-stress progression was intriguing. The absence of early regulated HSFs in RIL2219 may be important in its ability to mitigate drought. HSFs assist in protein folding under osmotic

stress, and their early regulation when protein synthesis is unaffected, suggests a novel role in signalling. One speculative explanation is their responsiveness to senescence (Breeze et al., 2008), which appears to be delayed in RIL2219.

CoPSA annotation of transcription factors on the wheat GeneChip enabled systems wide analysis of families expressed at early and late time-points (Section 5.6.4). This novel approach to interpreting transcriptional regulation in an evolving water-stress has never been performed for multiple cultivars of wheat. It revealed a global regulation of transcription, involving many transcription factor families, which included families with established associations with water-stress, but all included many families not usually associated with water-stress. The majority of transcriptional regulation was observed during late time points; however there were a number of early responsive transcription factors. This indicates that transcription is primed at a very early stage of water stress. ABA was chosen to demonstrate how CoPSA annotations could be used to drill down to detailed analysis. ABA biosynthesis and signalling is known to vitally important in a plants response to water stress (see Section 4.1.1(*a*)). Previous work had shown that the enzyme xanthoxin dehydrogenase which catalyses ABA biosynthesis is not significantly regulated during water stress (Nambara and Marion-Poll, 2005). However, the TRITIMED data set shows this is not the case, and the gene encoding xanthoxin dehydrogenase is significantly up-regulated at days 4 and 5 in Lahn and RIL2219, and days 3 and 4 in Cham1. Furthermore it was shown that all the genes upstream in ABA biosynthesis are regulated during the final three days of stress, and the catabolism step immediately after is strongly down regulated. This pointed to a late accumulation of ABA in response to severe water stress. The Cham1 cultivar exhibited an earlier response, which may explain its improved resistance to water stress. The RCAR-PP2C-ABA signalling receptor-complex was also examined, and a complex regulatory system (possibly additive) involving late up and down regulation during the final two day of drought stress in RIL2219 and Lahn, and the

final three days in Lahn. The early responsiveness of Lahn again points to a novel role of early ABA-mediated signalling in priming the plant, before severe water stress.

6.2 Acknowledgement of limitations

CoPSA is a sequence based functional annotations, which relies on identifying the putative functional-ortholog from which to transfer annotation. There are a number of inherent difficulties with this approach. The underlying assumption in this field is that the more similar the sequences, the greater the probability that they share function. This is not always the case, even with highly conserved sequences. Examples were given to support this in Section 1.1.4, and it was been discussed in some detail by Bork and Koonin (1998) and Karp (1998), the latter of which also discusses the dearth of provenance, for functional assignments, in public databases. The confidence with which the putative functional ortholog can be identified can be improved using phylogenetic techniques, however these approaches are computationally expensive, often require human supervision, and are still sequence based approaches. The biggest limitation with a sequence based approach is when a sequence only has weakly similar sequences in model organisms. Structural similarity between proteins has been shown to be a *more reliable* method for interfering similar function, however divergent and convergent evolution has led to some proteins which have different structures with similar functions, and similar structures with different functions, respectively (Hegyi and Gerstein, 1999). However, for most sequences a three dimensional structure is unavailable, and accurate and fast *in silico* prediction of three dimensional structure from protein sequence data has still not been realized in bioinformatics. Also, not all proteins expressed in the cell, are folded into complex globular structures, and remain

in a less structured state (Wright and Dyson, 1999). Given that sequences in non-model organisms, like wheat, are often incomplete, it seems unlikely this will be realized, in this area, in the immediate future. Given that sequence and structure based identification of putative functional-orthologs both have their limitations, we remain heavily reliant on experimental work to elucidate function in non-model organisms. Much of this experimental work already exists in unstructured resources, like text, and will require improvements in text-mining methods, and the adoption of structured publishing, to leverage it. Text-mining of functional annotation has been facilitated in recent years by the BioCreAtivE, grand challenges Yeh *et al.* (2005), Hirschman *et al.* (2005). Efforts to The statistical facilitate publishing experimental data directly in machine a readable form include Nano-Publications (Mons and Velterop, 2009) and Linked Data (Bizer *et al.*, 2009).

The example approach, presented in Chapter 6.3, was based on ANOVA for the analysis of significant expression, which assumes a normal distribution of gene expression. An inspection of the histogram plot of the residuals verses fitted values indicated a normal distribution, and Giles and Kipling (2003) previously reported normal distribution in 59 human GeneChips. However, Giles and Kipling (2003) also reported non-normal distributions among genes with low expression, and Hardin and Wilson (2009) have more recently reported non-normal distributions in Affymetrix data. Further analysis should include rigorous tests for non-normality, such as the Shapiro–Wilk test (SHAPIRO and WILK, 1965). Gao and Song (2005) have proposed a number of non-parametric tests, which are appropriate for for multi-factorial microarray expression data (factors are line and time in the experiment described in Chapter 6.3). They were able to elucidate treatment effects, without the assumption of a normal-distribution.

6.3 Further Work

Additional improvements could be made in the selection of high quality annotations from the CoPSA pipeline. The scoring metric selected for the applied annotation in Chapter 3, used consensus to identity annotations to transfer. Pragmatically, this works in the vast majority of cases as it excludes sub-functionalization within a gene family, and selects the majority function. However, an approach which more closely models the evolution of gene function would be to incorporate a phylogenetic analysis to identify the true ortholog ancestry of a gene, and then transfer only from the nearest informative ancestors. There are existing approaches to phylogeny that work on only sequence information, or with the expected sequencing of the wheat genome by the IWSC , it should be possible to incorporate synteny maps of the genomic sequence to identify orthologs. The filtering strategy used a single ortholog approach, *i.e.* it selected a gene, and then transferred all the annotation. Transferring annotation from multiple orthologs could improve annotation, by incorporating multi-functionalities of genes that have been verified in different species. However, for this approach to work a methodology for identifying the difference between *multi* and *sub*-functionalization would be required. This could be achieved through a phylogenetic approach to functional annotation. Additionally, a probabilistic approach based on the frequency of co-occurrence of function within the annotation sets of proteins, could be used to identify coherent annotations, sourced from *multiple* putative functional-orthologs.This would require a calculation of joint probabilities for sets of GO terms.

The existing CoPSA pipeline concentrates on the annotation of protein function, by exploiting protein sequence similarity and protein domains. Some sequences such as non-coding RNA are not translated to proteins, and consequently would not be assigned a function with the existing CoPSA pipeline.

There are existing efforts by groups such as (Dryanova *et al.*, 2008) to create pipelines in wheat to identify these sequences, which could be incorporated into the CoPSA pipeline.

One of the most exciting recent developments is the International Wheat Genome Sequencing Consortium (IWGSC) (Gill *et al.*, 2004), which aims to produce a high quality and fully assembled wheat sequence. This effort is currently mostly funded, and is currently producing genomic sequence data for the 3b chromosome. 85Gb of unassembled genomic sequence data of the wheat genome with 5x coverage has recently become available from a BBSRC project (Barker, 2010, 2011).

In addition to the work reported in this thesis. Information Theoretic approaches have been used by Stephen Powers to dissect the major components of variation in the data further. The GO and EC annotations created as part of this project, have also been used to better understand this alternative statistical dissection of the TRITIMED data. The transcription factors identified by CoPSA have been used by Michael Defoin-Platel as priors to predict regulatory networks, which have been inferred based on TRITIMED gene co-expression data. Enrichment of CoPSA derived GO annotations was observed in these regulatory clusters.

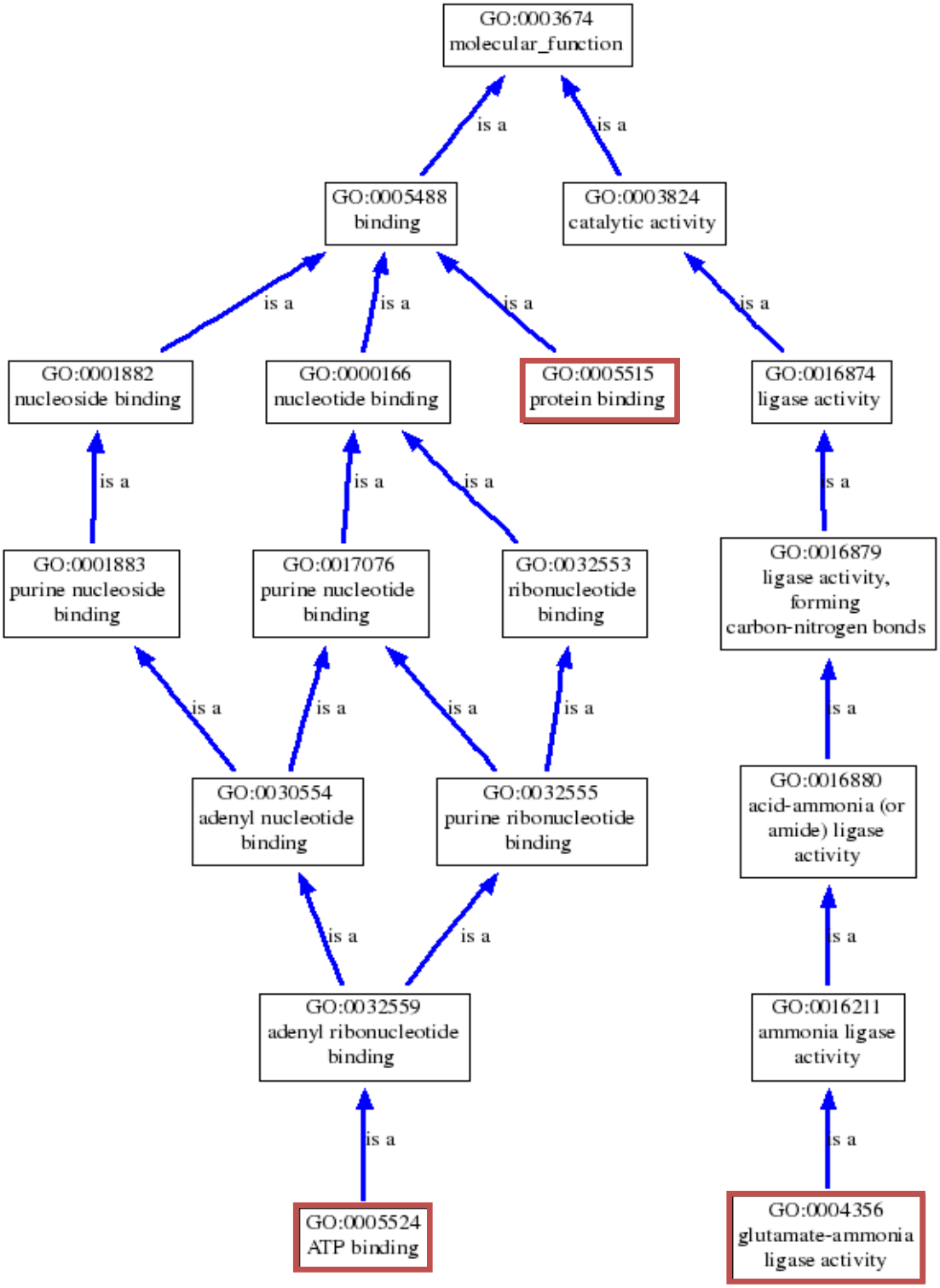
It became apparent when conducting the analysis in Chapter that visualisation of annotation for a large and complex transcriptome such the TRITIMED data set is a challenge. Ondex contains existing graph visualisation tools, which can lay out a complex network in an organic way. However, it is not always possible to provide generic visualisations that are accessible in a biological context. For example metabolic pathways require certain implicit assumptions about the main compounds and the important reactions, which are difficult to infer computationally, and require a specific visualisation schema. Reactome has created layouts to computationally generate metabolic pathways (Matthews *et al.*, 2009), however to accommodate multiple biological domains such as signalling, and physiological processes is a challenge. Existing tools like MAP-

MAN Thimm *et al.* (2004) have created user friendly visualisations that are accessible to a biologist. With a translation between GO, EC, and the MAPMAN ontology it would be possible to use CoPSA annotations within MAPMAN.

Appendix

Appendix 1: KEGG species imported into Ondex, as part of the CoPSA pipeline.

KEGG species code	Organism	Common name	Group
vvi	*Vitis vinifera*	Wine grape	Grape
osa	*Oryza sativa japonica*	Japanese rice	Grasses
sbi	*Sorghum bicolor*	Sorghum	Grasses
zma	*Zea mays*	Maize	Grasses
cre	*Chlamydomonas reinhardtii*	Green algae	
olu	*Ostreococcus lucimarinus*		Green algae
ppp	*Physcomitrella patens subsp. patens*	Mosses	
ath	*Arabidopsis thaliana*	Thale cress	Mustard
cma	*Cyanidioschyzon merolae*		Red algae
rcu	*Ricinus communis*	Castor bean	Spurge
pop	*Populus trichocarpa*	Black cottonwood	Willow



Appendix 2: Functions of GS2 gene in the gene ontology molecular function category highlighted in red.

Appendix 3: A comparison of significantly (p<0.05) expressed genes in RIL2219 with other cultivars

	Percentage of significant probes shared		
Time point	Lahn	Cham1	Both
1	20.94%	21.49%	7.16%
2	14.89%	26.34%	4.20%
3	12.44%	58.13%	8.37%
4	80.14%	72.63%	63.68%
5	92.85%	94.43%	89.28%

Appendix 4: A comparison of significantly (p<0.05) expressed genes in Cham1 with other cultivars.

	Percentage of significant probes shared		
Time point	Lahn	RIL2219	Both
1	30.10%	19.40%	6.47%
2	13.45%	6.58%	1.05%
3	79.81%	6.10%	0.88%
4	91.19%	11.33%	9.94%
5	88.56%	85.43%	80.77%

Appendix 5: A comparison of significantly (p<0.05) expressed genes in Lahn with other cultivars.

	Percentage of significant probes shared		
Time point	RIL2219	Cham1	Both
1	4.48%	7.14%	1.53%
2	4.79%	17.30%	1.35%
3	7.98%	45.09%	5.37%
4	8.94%	65.22%	7.11%
5	88.51%	93.33%	85.12%

Appendix 6: A ClustalW2 alignment of TaAffx.44105.1.S1_at with the protein encoded at the locus 02g023760 in *Sorghum bicolor* (XP_002462318.1

```
TaAffx.44105.1.S1_at  MASLALRPIMPAXXAAASTTTXLNRARRGSXTLLR-----RRQPVTCMAESSGGGNSTVE   55
02g023760             MASLALHPIIPAT-AASSTTLAVVVGSRHAATLHRCFRPRRRLITTCKAEPSGG-NSTVE   58
                      ******:**:** **:*** :  .  * :  ** * ** .** **.*** *****

TaAffx.44105.1.S1_at  LAAXAXGXASCAXVAWSLYXXKATGCGLPPGPGGSXGAAEGVSYLVVAGXVGWSLTTKVR   115
02g023760             LAAGAAGLASSAVVAWSLYTLKTTGCGLPPGPGGALGAAEGVSYLVVAGLVGWSATTKVR   118
                      *** * * **.* ****** *:***********:  ************ **** *****

TaAffx.44105.1.S1_at  TGSXLPAGPYGXLXAAEGVXYXXVVAIAAVXGLXFFXXGSLPGPXPXXXCFG           167
02g023760             TGSGLPAGPYGLLGAAEGVAYLTVAAIAVVFGLQFFQQGSIPGPLPSEQCFG           170
                      *** ******* * ***** * *.***.* ** ** **:*** * ***
```

Appendix 7: A ClustalW2 alignment of RCAR (PYL1-13 and PYR1) with candidates on the wheat chip. Translation of nucleotide consensus sequences was performed using GeneWise2.

```
TaAffx.131393.1.S1_at  ------------------------------------------------
  Ta.21082.1.S1_x_at   ------------------------------------------------
                PYL4   ---------------------MLAVHRPSSAVSDGDSVQIPMMIAS--F   26
                PYL5   -------------------MRSPVQLQHGSDATNGFHTLQPHDQTDG--P   29
               PYL12   -------------------MKTSQEQHVCG---------------------   11
               PYL11   -------------------METSQKYHTCG---------------------   11
               PYL13   -------------------MESS-KQKRCR---------------------   10
                PYL6   -------------------MPTSIQFQRSSTAAEAANATVRNYPHHHQKQ   31
               PYL10   ----------------------MNGDE------TKKVES-----------   11
                PYL8   --------------------MEANGIEN---LTNPNQER-----------   16
                PYL9   --------------------MMDGVEGGTAMYGGLETV-----------   18
                PYL7   -------------------MEMIGGDDTDTEMYGALVTA-----------   20
TaAffx.109881.1.S1_at  ------------------------------------EA-----------    2
  TaAffx.11433.1.S1_at ------------------------------------EEM----------    3
                PYL3   ------MNLAPIHDPSSSSTTTTSSSTPYGLTKDEFST-----------   32
                PYL2   --------------------MSSSPAVKGLTDEEQKT-----------   17
                PYR1   -------------------------MPSELTPEERSE-----------   12
                PYL1   MANSESSSSPVNEEENSQRISTLHHQTMPSDLTQDEFTQ-----------   39
  TaAffx.43193.1.S1_at ------------------------------LTAAEYAQ-----------    8

TaAffx.131393.1.S1_at  ----------VPPEVARHHXHAAPGGRCCCSAVVQRVAAPXADVWAVVX    39
  Ta.21082.1.S1_x_at   ---------------ARHHEHAEPGSGQCCSAVVQHVAAPAAAVWSVVR    34
                PYL4   QKRFPS---LSRDSTAARFHTHEV-GPNQCCSAVIQEISAPISTVWSVVR   72
                PYL5   IKRVCLTRGMHVPEHVAMHHTHDV-GPDQCCSSVVQMIHAPPESVWALVR   78
               PYL12   ------------------------------STVVQTINAPLPLVWSILR    30
               PYL11   ------------------------------STLVQTIDAPLSLVWSILR    30
               PYL13   ------------------------------SSVVETIEAPLPLVWSILR    29
                PYL6   VQKVSLTRGMADVPEHVELSHTHVVGPSQCFSVVVQDVEAPVSTVWSILS   81
               PYL10   -------------EYIKKHHRHELVESQCSSTLVKHIKAPLHLVWSIVR    47
                PYL8   -------------EFIRRHHKHELVDNQCSSTLVKHINAPVHIVWSLVR    52
                PYL9   -------------QYVRTHHQHLCRENQCTSALVKHIKAPLHLVWSLVR    54
                PYL7   -------------QSLRLRHLHHCRENQCTSVLVKYIQAPVHLVWSLVR    56
TaAffx.109881.1.S1_at  -------------DYMRRLHGHAPGENQCTSALVKHIKAPXHLVWSXVR    38
  TaAffx.11433.1.S1_at -------------EYVRRFHQHEPGANQCTSFIAKHIKAPLQTVWSVVR    39
                PYL3   -----------LDSIIRTHHTFPRSPNTCTSLIAHRVDAPAHAIWRFVR    70
                PYL2   -----------LEPVIKTYHQFEPDPTTCTSLITQRIHAPASVVWPLIR    55
                PYR1   -----------LKNSIAEFHTYQLDPGSCSSLHAQRIHAPPELVWSIVR    50
                PYL1   -----------LSQSIAEFHTYQLGNGRCSSLLAQRIHAPPETVWSVVR    77
  TaAffx.43193.1.S1_at -----------LLPTVEAYHRYAVGPGQCSXLVAQRIEAPPAAVWAIVR    46

TaAffx.131393.1.S1_at  RFDQPQAYKSFVRSCALLDXXG------GVGTLXEVRVVXGLPAASSRER    83
  Ta.21082.1.S1_x_at   RFDQPQAYKRFVRSCALVAGDX------GVGTLREVHVVSGLPAASSRER    78
                PYL4   RFDNPQAYKHFLKSCSVIGGDGD-----NVGSLRQVHVVSGLPAASSTER   117
                PYL5   RFDNPKVYKNFIRQCRIVQGDGL-----HVGDLREVMVVSGLPAVSSTER   123
               PYL12   RFDNPKTFKHFVKTCKLRSGDGG------EGSVREVTVVSDLPASFSLER    74
               PYL11   RFDNPQAYKQFVKTCNLSSGDGG------EGSVREVTVVSGLPAEFSRER    74
               PYL13   SFDKPQAYQRFVKSCTMRSGGGGGKGGEGKGSVRDVTLVSGFPADFSTER    79
                PYL6   RFEHPQAYKHFVKSCHVVIGDGRE-----VGSVREVRVVSGLPAAFSLER   126
               PYL10   RFDEPQKYKPFISRCVVQGK------KLEVGSVREVDLKSGLPATKSTEV    91
                PYL8   RFDQPQKYKPFISRCVVKG-------NMEIGTVREVDVKSGLPATRSTER    95
                PYL9   RFDQPQKYKPFVSRCTVIG-------DPEIGSLREVNVKSGLPATTSTER    97
                PYL7   RFDQPQKYKPFISRCTVNG-------DPEIGCLREVNVKSGLPATTSTER    99
TaAffx.109881.1.S1_at  SFDQPQRYKPFVSRCVVRGG------DLEIGSVREVNVKTXLPATTSTER    82
  TaAffx.11433.1.S1_at RFDKPQVYKRFVENCVMQG-------NIEPGCVREVTLKSGLPGKWSIER    82
                PYL3   DFANPNKYKHFIKSCTIRVNGNGI-KEIKVGTIREVSVVSGLPASTSVEI   119
                PYL2   RFDNPERYKHFVKRCRL-ISGDG-----DVGSVREVTVISGLPASTSTER    99
                PYR1   RFDKPQTYKHFIKSCSVEQN-----FEMRVGCTRDVIVISGLPANTSTER    95
                PYL1   RFDRPQIYKHFIKSCNVSED-----FEMRVGCTRDVNVISGLPANTSRER   122
  TaAffx.43193.1.S1_at RXDCPQVYXHFIRSCALRXDPEAG-DELRPGRLREVSVISGLPASTSTER    95
```

Continued on next page

```
TaAffx.131393.1.S1_at  LEILDDERHVLSFSVVGGEHRLRNYRSVTTVHPAPGES-----ASATLVV  128
  Ta.21082.1.S1_x_at   LEILDDESHVLXFRVVGGEHRLKNYLSVTTVHPSXAAP-----SSATVVV  123
                PYL4   LDILDDERHVISFSVVGGDHRLSNYRSVTTLHPSP--------ISGTVVV  159
                PYL5   LEILDEERHVISFSVVGGDHRLKNYRSVTTLHASD--------DEGTVVV  165
               PYL12   LDELDDESHVMVISIIGGDHRLVNYQSKTTVFVAAE-------EEKTVVV  117
               PYL11   LDELDDESHVMMISIIGGDHRLVNYRSKTMAFVAADT------EEKTVVV  118
               PYL13   LEELDDESHVMVVSIIGGNHRLVNYKSKTKVVASPEDM-----AKKTVVV  124
                PYL6   LEIMDDDRHVISFSVVGGDHRLMNYKSVTTVHESEEDSD---GKKRTRVV  173
               PYL10   LEILDDNEHILGIRIVGGDHRLKNYSSTISLHSETIDG-----KTGTLAI  136
                PYL8   LELLDDNEHILSIRIVGGDHRLKNYSSIISLHPETIEG-----RIGTLVI  140
                PYL9   LELLDDEEHILGIKIIGGDHRLKNYSSILTVHPEIIEG-----RAGTMVI  142
                PYL7   LEQLDDEEHILGINIIGGDHRLKNYSSILTVHPEMIDG-----RSGTMVM  144
TaAffx.109881.1.S1_at  LEQLDDDEHILSVKXVGGDHRLRNYSSIITVHPQSIDG-----RPGTLVI  127
TaAffx.11433.1.S1_at   LELLDDNEHILSVKFI-----------------------------------   98
                PYL3   LEVLDEEKRILSFRVLGGEHRLNNYRSVTSVNEFVVLEKDKKKRVYSVVL  169
                PYL2   LEFVDDDHRVLSFRVVGGEHRLKNYKSVTSVNEFLNQDSGK---VYTVVL  146
                PYR1   LDILDDERRVTGFSIIGGEHRLTNYKSVTTVHRFEKEN-----RIWTVVL  140
                PYL1   LDLLDDDRRVTGFSITGGEHRLRNYKSVTTVHRFEKEEEEE--RIWTVVL  170
TaAffx.43193.1.S1_at   LDLXXDARRAFGFXITGGEHRLRXYRSVTTVSELSXAAP-A--EICTVVL  142

TaAffx.131393.1.S1_at  ESYVVDVPPGNTPEDTRVFVDTIVKCNLQSLARTAEK-------------  165
  Ta.21082.1.S1_x_at   E-----------------------------------------------  124
                PYL4   ESYVVDVPPGNTKEETCDFVDVIVRCNLQSLAKIAENTAAESKKKMSL--  207
                PYL5   ESYIVDVPPGNTEEETLSFVDTIVRCNLQSLARSTNRQ------------  203
               PYL12   ESYVVDVPEGNTEEETTLFADTIVGCNLRSLAKLSEKMMELT--------  159
               PYL11   SYVVDVPEGNSEEETTSFADTIVGFNLKSLAKLSERVAHLKL--------  161
               PYL13   SYVVDVPEGTSEEDTIFFVDNIIRYNLTSLAKLTKKMMK----------  164
                PYL6   SYVVDVPAGNDKEETCSFADTIVRCNLQSLAKLAENTSKFS--------  215
               PYL10   SFVVDVPEGNTKEETCFFVEALIQCNLNSLADVTERLQAES-MEKKI--  183
                PYL8   SFVVDVPEGNTKDETCYFVEALIKCNLKSLADISERLAVQDTTESRV--  188
                PYL9   SFVVDVPQGNTKDETCYFVEALIRCNLKSLADVSERLASQDITQ-----  187
                PYL7   SFVVDVPQGNTKDDTCYFVESLIKCNLKSLACVSERLAAQDITNSIATF  194
TaAffx.109881.1.S1_at  SFVVDVXDGNTXDXTCXFVEAXXKXNXTSLXEXSXXLXVX----------  168
TaAffx.11433.1.S1_at   ------------------------------------------------
                PYL3   SYIVDIPQGNTEEDTRMFVDTVVKSNLQNLAVISTASPT----------  209
                PYL2   SYTVDIPEGNTEEDTKMFVDTVVKLNLQKLGVAATSAPMHDDE------  190
                PYR1   SYVVDMPEGNSEDDTRMFADTVVKLNLQKLATVAEAMARNSGDGSGSQV  190
                PYL1   SYVVDVPEGNSEEDTRLFADTVIRLNLQKLASITEAMNRNNNNNNSSQV  220
TaAffx.43193.1.S1_at   SYVVDVPDGNSEEDTRLFADTVVRLNLQKLKSVAEA-------------  179

TaAffx.131393.1.S1_at  ----------------
  Ta.21082.1.S1_x_at   ----------------
                PYL4   ----------------
                PYL5   ----------------
               PYL12   ----------------
               PYL11   ----------------
               PYL13   ----------------
                PYL6   ----------------
               PYL10   ----------------
                PYL8   ----------------
                PYL9   ----------------
                PYL7   CNASNGYREKNHTETNL                                  211
TaAffx.109881.1.S1_at  ----------------
TaAffx.11433.1.S1_at   ----------------
                PYL3   ----------------
                PYL2   ----------------
                PYR1   T---------------                                  191
                PYL1   R---------------                                  221
TaAffx.43193.1.S1_at   ----------------
```

Appendix 8: PCo1 from line.time analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score (PropF) for the PCo are shown. For each gene PropF is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
Ta.27765.3.S1_x_at	1.18×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
TaAffx.44105.1.S1_at	2.24×10^{-3}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.805.1.S1_at	3.27×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.27765.3.S1_at	4.26×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.2887.1.S1_x_at	5.22×10^{-3}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.2887.1.S1_at	6.17×10^{-3}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.1519.2.S1_x_at	7.09×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.1924.1.S1_x_at	7.99×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.2901.1.S1_at	8.89×10^{-3}	protein transporter activity(GO:0008565)	
Ta.1519.2.S1_at	9.78×10^{-3}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28291.1.S1_at	1.06×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.3258.1.S1_at	1.15×10^{-2}	transferase activity, transferring acyl groups other than amino-acyl groups, protochlorophyllide reductase activity, NADPH dehydrogenase activity(GO:0016747, GO:0016630, GO:0003959)	chlorophyll biosynthetic process, oxidation-reduction process(GO:0015995, GO:0055114)
Ta.9829.1.A1_s_at	1.23×10^{-2}		

Continued on next page

Ta.23479.1.S1_a_at	1.32×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.27117.2.S1_x_at	1.40×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.19112.1.S1_at	1.48×10^{-2}	unfolded protein binding, ATP binding(GO:0051082, GO:0005524)	protein folding(GO:0006457)
Ta.824.1.S1_a_at	1.56×10^{-2}	structural constituent of ribosome, RNA binding(GO:0003735, GO:0003723)	RNA processing, translation(GO:0006396, GO:0006412)
Ta.1924.2.S1_x_at	1.64×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.14564.1.S1_at	1.72×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28291.4.S1_x_at	1.80×10^{-2}	structural constituent of ribosome, 5S rRNA binding(GO:0003735, GO:0008097)	cell proliferation, ribosome biogenesis, leaf morphogenesis, translation(GO:0008283, GO:0042254, GO:0009965, GO:0006412)
Ta.27765.1.S1_x_at	1.88×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.14445.1.A1_x_at	1.95×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.3100.1.S1_x_at	2.03×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.1289.3.S1_at	2.10×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.27765.4.S1_x_at	2.17×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28855.2.S1_at	2.25×10^{-2}	structural constituent of ribosome, RNA binding(GO:0003735, GO:0003723)	translation(GO:0006412)
Ta.25124.2.S1_x_at	2.32×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.10538.1.A1_at	2.39×10^{-2}	peptidyl-prolyl cis-trans isomerase activity(GO:0003755)	protein folding, protein transport(GO:0006457, GO:0015031)
Ta.2739.4.S1_x_at	2.46×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.3789.3.A1_a_at	2.53×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28703.1.S1_at	2.60×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.25124.2.S1_a_at	2.67×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)

Continued on next page

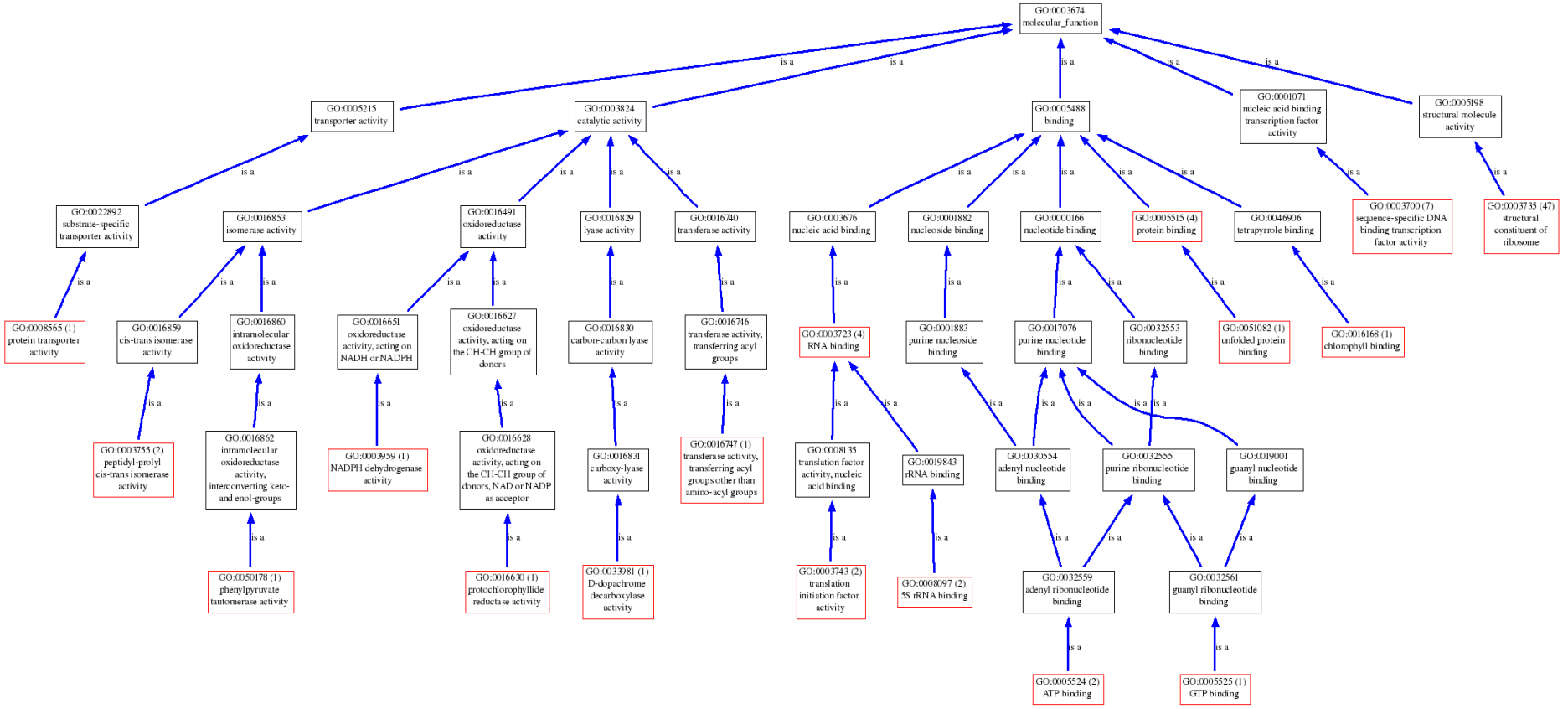
Ta.14324.1.S1_x_at	2.74×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.7195.2.S1_a_at	2.81×10^{-2}	protein binding, ATP binding(GO:0005515, GO:0005524)	protein metabolic process(GO:0019538)
TaAffx.30053.1.S1_at	2.88×10^{-2}		
Ta.28434.1.S1_x_at	2.94×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.23823.1.S1_a_at	3.01×10^{-2}	GTP binding(GO:0005525)	barrier septum formation(GO:0000917)
Ta.28855.1.S1_at	3.08×10^{-2}	structural constituent of ribosome, RNA binding(GO:0003735, GO:0003723)	translation(GO:0006412)
Ta.24128.1.S1_x_at	3.14×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.8669.1.S1_at	3.21×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
TaAffx.132352.1.S1_at	3.28×10^{-2}	D-dopachrome decarboxylase activity, phenylpyruvate tautomerase activity(GO:0033981, GO:0050178)	inflammatory response, response to other organism(GO:0006954, GO:0051707)
Ta.30628.3.S1_x_at	3.34×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.10364.1.S1_s_at	3.41×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to gibberellin stimulus(GO:0009739)
Ta.1698.1.S1_at	3.47×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.536.1.S1_at	3.54×10^{-2}	RNA binding, translation initiation factor activity(GO:0003723, GO:0003743)	

Continued on next page

Ta.28291.4.S1_at	3.60×10^{-2}	structural constituent of ribosome, 5S rRNA binding(GO:0003735, GO:0008097)	cell proliferation, ribosome biogenesis, leaf morphogenesis, translation(GO:0008283, GO:0042254, GO:0009965, GO:0006412)
Ta.28204.1.S1_x_at	3.67×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.1139.1.S1_x_at	3.73×10^{-2}	chlorophyll binding(GO:0016168)	nonphotochemical quenching, response to red light, photosynthesis, light harvesting, response to blue light, response to far red light(GO:0010196, GO:0010114, GO:0009765, GO:0009637, GO:0010218)
Ta.2841.1.S1_s_at	3.80×10^{-2}	translation initiation factor activity(GO:0003743)	
Ta.3756.1.S1_x_at	3.86×10^{-2}	peptidyl-prolyl cis-trans isomerase activity(GO:0003755)	protein folding, defense response to bacterium(GO:0006457, GO:0042742)
Ta.14324.3.S1_x_at	3.93×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.13368.1.S1_x_at	3.99×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.9496.3.S1_at	4.05×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.28426.1.S1_a_at	4.12×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.27785.1.S1_a_at	4.18×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.30803.2.S1_x_at	4.24×10^{-2}	protein binding, structural constituent of ribosome(GO:0005515, GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)

Continued on next page

Ta.22683.2.S1_at	4.31×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28204.3.S1_at	4.37×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.22849.1.S1_x_at	4.43×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.30787.1.S1_at	4.49×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.24405.1.S1_at	4.56×10^{-2}	protein binding, sequence-specific DNA binding transcription factor activity(GO:0005515, GO:0003700)	response to salt stress(GO:0009651)
Ta.15939.1.S1_at	4.62×10^{-2}		
Ta.3425.2.S1_x_at	4.68×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to salt stress(GO:0009651)
Ta.26908.1.A1_at	4.74×10^{-2}	protein binding(GO:0005515)	oxidation-reduction process(GO:0055114)
Ta.28338.1.S1_at	4.80×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.2605.2.S1_at	4.86×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.1551.1.S1_s_at	4.93×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.21954.1.S1_at	4.99×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.3100.2.S1_x_at	5.05×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)



Appendix 9: PCo1 from line.time analysis, showing the function of annotated genes highlighted in red with the number of genes with that exact term given in brackets. Only functions for genes that amount to 10% of the total f-score (PropF) for the PCo are shown. Visualisation created using AmiGO.

Appendix 10: PCo2 from line.time analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score (PropF) for the PCo are shown. For each gene PropF is the is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
Ta.8533.1.A1_s_at	2.30×10^{-3}	oxidoreductase activity(GO:0016491)	
Ta.11544.2.S1_a_at	4.36×10^{-3}	receptor activity(GO:0004872)	
Ta.2514.1.S1_a_at	6.41×10^{-3}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	protein ubiquitination(GO:0016567)
Ta.20390.2.S1_a_at	8.46×10^{-3}	protein binding, protein disulfide oxidoreductase activity, electron carrier activity(GO:0005515, GO:0015035, GO:0009055)	cellular iron ion homeostasis, cell redox homeostasis, response to oxidative stress(GO:0006879, GO:0045454, GO:0006979)
Ta.8183.2.S1_x_at	1.05×10^{-2}		
Ta.9941.1.S1_at	1.24×10^{-2}	NADH dehydrogenase (ubiquinone) activity(GO:0008137)	photorespiration(GO:0009853)
Ta.12503.1.S1_at	1.42×10^{-2}	binding, catalytic activity(GO:0005488, GO:0003824)	
TaAffx.53216.1.S1_x_at	1.59×10^{-2}	antiporter activity, protein disulfide oxidoreductase activity, electron carrier activity, glutathione disulfide oxidoreductase activity(GO:0015297, GO:0015035, GO:0009055, GO:0015038)	cell redox homeostasis, cation transport(GO:0045454, GO:0006812)
Ta.4274.1.S1_at	1.74×10^{-2}	sucrose synthase activity(GO:0016157)	biosynthetic process, sucrose metabolic process(GO:0009058, GO:0005985)

Continued on next page

TaAffx.128827.1.S1_s_at	1.89×10^{-2}		
Ta.22094.1.S1_at	2.03×10^{-2}	protein binding(GO:0005515)	auxin homeostasis, pollen tube growth, root hair cell tip growth, lateral root development, regulation of cell growth, response to cadmium ion, defense response to bacterium(GO:0010252, GO:0009860, GO:0048768, GO:0048527, GO:0001558, GO:0046686, GO:0042742)
Ta.26013.1.A1_a_at	2.17×10^{-2}	binding(GO:0005488)	
Ta.12503.1.S1_x_at	2.30×10^{-2}	cofactor binding, oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor(GO:0048037, GO:0016616)	
Ta.11544.3.S1_x_at	2.43×10^{-2}		
Ta.30529.2.A1_x_at	2.56×10^{-2}	nucleotide binding(GO:0000166)	
Ta.8477.1.S1_at	2.68×10^{-2}		
TaAffx.25219.1.S1_at	2.81×10^{-2}		
TaAffx.32208.1.S1_s_at	2.93×10^{-2}		
TaAffx.38326.1.S1_at	3.05×10^{-2}	protein binding(GO:0005515)	
Ta.20390.2.S1_x_at	3.17×10^{-2}	antiporter activity, protein disulfide oxidoreductase activity, electron carrier activity, glutathione disulfide oxidoreductase activity(GO:0015297, GO:0015035, GO:0009055, GO:0015038)	cell redox homeostasis(GO:0045454)
Ta.14176.1.S1_at	3.30×10^{-2}		
Ta.5965.1.S1_at	3.42×10^{-2}		

Continued on next page

Ta.3242.1.A1_a_at	3.54×10^{-2}		
Ta.14176.1.S1_x_at	3.66×10^{-2}		
Ta.7398.1.A1_at	3.78×10^{-2}		
Ta.1282.4.S1_at	3.90×10^{-2}	serine-type endopeptidase inhibitor activity, nutrient reservoir activity, lipid binding(GO:0004867, GO:0045735, GO:0008289)	lipid transport(GO:0006869)
TaAffx.110529.1.S1_at	4.01×10^{-2}	protein binding(GO:0005515)	protein phosphorylation(GO:0006468)
Ta.13966.3.S1_a_at	4.13×10^{-2}	calcium-dependent phospholipid binding, phospholipase inhibitor activity, calcium ion binding(GO:0005544, GO:0004859, GO:0005509)	negative regulation of coagulation, response to salt stress, response to heat, response to red or far red light, response to cold, response to water deprivation(GO:0050819, GO:0009651, GO:0009408, GO:0009639, GO:0009409, GO:0009414)
Ta.22488.1.S1_at	4.24×10^{-2}		
Ta.7823.1.S1_at	4.35×10^{-2}		
Ta.22794.1.S1_at	4.46×10^{-2}	protein binding, ubiquitin-protein ligase activity(GO:0005515, GO:0004842)	ubiquitin-dependent protein catabolic process, mitosis, response to cadmium ion, negative regulation of DNA recombination, male meiosis(GO:0006511, GO:0007067, GO:0046686, GO:0045910, GO:0007140)
Ta.11544.1.S1_a_at	4.57×10^{-2}		
Ta.11544.2.S1_x_at	4.67×10^{-2}	receptor activity(GO:0004872)	
Ta.8528.1.A1_at	4.77×10^{-2}		
TaAffx.38601.1.A1_at	4.87×10^{-2}		

Continued on next page

TaAffx.94115.1.S1_at	4.97×10^{-2}
Ta.28458.2.S1_at	5.07×10^{-2}

Appendix 11: PCo3 from line.time analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score (PropF) for the PCo are shown. For each gene PropF is the is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
Ta.21089.1.S1_at	2.12×10^{-3}	cyclopropane-fatty-acyl-phospholipid synthase activity, tocopherol O-methyltransferase activity(GO:0008825, GO:0050342)	lipid biosynthetic process(GO:0008610)
Ta.21211.1.S1_at	4.08×10^{-3}	iron ion binding, 4,4-dimethyl-9beta,19-cyclopropylsterol-4alpha-methyl oxidase activity(GO:0005506, GO:0080064)	sterol biosynthetic process(GO:0016126)
Ta.9000.1.S1_at	5.95×10^{-3}	hydrolase activity, hydrolyzing O-glycosyl compounds(GO:0004553)	carbohydrate metabolic process(GO:0005975)
Ta.22764.1.S1_x_at	7.80×10^{-3}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to cold, hyperosmotic salinity response(GO:0009409, GO:0042538)
Ta.9259.1.S1_at	9.53×10^{-3}	cysteine-type endopeptidase inhibitor activity, serine-type endopeptidase inhibitor activity(GO:0004869, GO:0004867)	DNA repair(GO:0006281)
TaAffx.132392.1.S1_at	1.12×10^{-2}		

Continued on next page

Ta.4760.1.S1_x_at	1.28×10^{-2}	calmodulin binding, calcium-transporting ATPase activity, ATP binding(GO:0005516, GO:0005388, GO:0005524)	ATP biosynthetic process, response to nematode(GO:0006754, GO:0009624)
Ta.3502.1.S1_at	1.43×10^{-2}	binding(GO:0005488)	mitochondrial transport(GO:0006839)
Ta.28432.1.S1_at	1.58×10^{-2}		embryo development(GO:0009790)
TaAffx.84503.1.S1_at	1.72×10^{-2}		
Ta.23417.1.S1_at	1.85×10^{-2}	binding(GO:0005488)	
Ta.9194.2.A1_at	1.97×10^{-2}	transmembrane transporter activity(GO:0022857)	
Ta.21055.1.S1_at	2.09×10^{-2}		
Ta.6470.1.S1_x_at	2.21×10^{-2}	trehalose-phosphatase activity, alpha,alpha-trehalose-phosphate synthase (UDP-forming) activity(GO:0004805, GO:0003825)	trehalose biosynthetic process(GO:0005992)
Ta.26178.1.A1_at	2.33×10^{-2}	hydrolase activity(GO:0016787)	
Ta.9497.1.S1_at	2.45×10^{-2}	DNA binding(GO:0003677)	regulation of transcription(GO:0045449)
Ta.4760.1.S1_at	2.57×10^{-2}	calmodulin binding, calcium-transporting ATPase activity, ATP binding(GO:0005516, GO:0005388, GO:0005524)	ATP biosynthetic process, response to nematode(GO:0006754, GO:0009624)
Ta.19289.1.S1_at	2.69×10^{-2}		
Ta.28513.1.S1_s_at	2.80×10^{-2}	sequence-specific DNA binding transcription factor activity, protein dimerization activity(GO:0003700, GO:0046983)	leaf development, lateral root primordium development, response to ethylene stimulus(GO:0048366, GO:0010386, GO:0009723)

Continued on next page

Ta.16038.1.S1_at	2.91×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to cold, response to water deprivation, response to abscisic acid stimulus, hyperosmotic salinity response(GO:0009409, GO:0009414, GO:0009737, GO:0042538)
Ta.29434.1.A1_at	3.02×10^{-2}	enzyme inhibitor activity, pectinesterase activity(GO:0004857, GO:0030599)	
Ta.4333.1.S1_at	3.13×10^{-2}		
Ta.10334.2.A1_at	3.24×10^{-2}	D-alanine-D-alanine ligase activity, ATP binding(GO:0008716, GO:0005524)	peptidoglycan biosynthetic process(GO:0009252)
Ta.435.1.S1_at	3.35×10^{-2}	3-chloroallyl aldehyde dehydrogenase activity(GO:0004028)	oxidation-reduction process(GO:0055114)
Ta.3677.1.S1_at	3.46×10^{-2}	calcium-dependent cysteine-type endopeptidase activity, sucrose transmembrane transporter activity(GO:0004198, GO:0008515)	proteolysis, response to wounding(GO:0006508, GO:0009611)
Ta.30607.1.A1_at	3.56×10^{-2}	sequence-specific DNA binding, sequence-specific DNA binding transcription factor activity(GO:0043565, GO:0003700)	regulation of transcription, DNA-dependent, transcription(GO:0006355, GO:0006350)
Ta.3366.1.S1_at	3.66×10^{-2}	1-deoxy-D-xylulose-5-phosphate synthase activity, oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor(GO:0008661, GO:0016624)	terpenoid biosynthetic process(GO:0016114)

Continued on next page

Ta.28335.1.S1_at	3.77×10^{-2}	transporter activity(GO:0005215)	transmembrane transport(GO:0055085)
Ta.6041.2.A1_x_at	3.87×10^{-2}	mechanically-gated ion channel activity(GO:0008381)	detection of mechanical stimulus(GO:0050982)
Ta.5457.1.S1_at	3.97×10^{-2}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	
TaAffx.122407.1.S1_at	4.07×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.4354.1.S1_s_at	4.17×10^{-2}	endonuclease activity, nucleic acid binding(GO:0004519, GO:0003676)	regulation of transcription, DNA-dependent(GO:0006355)
TaAffx.79552.1.S1_x_at	4.26×10^{-2}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	protein ubiquitination(GO:0016567)
TaAffx.118107.1.S1_at	4.36×10^{-2}	D-alanine-D-alanine ligase activity(GO:0008716)	peptidoglycan biosynthetic process(GO:0009252)
Ta.9194.1.S1_x_at	4.46×10^{-2}	transmembrane transporter activity(GO:0022857)	
Ta.1899.1.S1_x_at	4.55×10^{-2}	cysteine-type endopeptidase activity(GO:0004197)	proteolysis, induction of apoptosis(GO:0006508, GO:0006917)
Ta.10389.3.A1_at	4.65×10^{-2}	aldo-keto reductase activity, steroid dehydrogenase activity(GO:0004033, GO:0016229)	oxidation-reduction process(GO:0055114)
Ta.6470.1.S1_at	4.74×10^{-2}	trehalose-phosphatase activity, alpha,alpha-trehalose-phosphate synthase (UDP-forming) activity(GO:0004805, GO:0003825)	trehalose biosynthetic process(GO:0005992)
Ta.21064.2.S1_at	4.83×10^{-2}		
TaAffx.97935.1.S1_at	4.93×10^{-2}	fatty-acyl-CoA synthase activity, 4-coumarate-CoA ligase activity(GO:0004321, GO:0016207)	

Continued on next page

Ta.2462.1.S1_s_at	5.02×10^{-2}	protein binding, ATP binding, ATP:ADP antiporter activity(GO:0005515, GO:0005524, GO:0005471)

Appendix 12: PCo1 from line.RWC analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score for the PCo are shown. For each gene PropF is the is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
Ta.3659.1.S1_a_at	1.00×10^{-3}	GTP binding(GO:0005525)	response to stress(GO:0006950)
Ta.29505.1.S1_x_at	1.91×10^{-3}	microtubule binding, APG8-specific protease activity, APG8 activating enzyme activity, Atg8 ligase activity(GO:0008017, GO:0019786, GO:0019779, GO:0019776)	autophagy(GO:0006914)
Ta.824.1.S1_a_at	2.64×10^{-3}	structural constituent of ribosome, RNA binding(GO:0003735, GO:0003723)	RNA processing, translation(GO:0006396, GO:0006412)
Ta.14042.1.S1_at	3.34×10^{-3}		
Ta.1252.1.S1_at	4.03×10^{-3}	RNA binding(GO:0003723)	
Ta.2958.1.A1_at	4.69×10^{-3}		
Ta.8999.1.A1_at	5.35×10^{-3}	sequence-specific DNA binding transcription factor activity(GO:0003700)	multicellular organismal development(GO:0007275)
Ta.5636.2.S1_x_at	6.01×10^{-3}	hydrolase activity, acting on ester bonds(GO:0016788)	lipid metabolic process(GO:0006629)

Continued on next page

Ta.28146.1.S1_s_at	6.64×10^{-3}	binding(GO:0005488)	fatty acid beta-oxidation, ADP transport, ATP transport, indolebutyric acid metabolic process, mitochondrial transport(GO:0006635, GO:0015866, GO:0015867, GO:0080024, GO:0006839)
Ta.19158.2.S1_at	7.27×10^{-3}	serine-type endopeptidase activity(GO:0004252)	proteolysis(GO:0006508)
Ta.22212.1.S1_at	7.89×10^{-3}		
Ta.7238.1.S1_at	8.50×10^{-3}	protein binding(GO:0005515)	vernalization response, histone methylation, regulation of flower development(GO:0010048, GO:0016571, GO:0009909)
Ta.2427.1.S1_x_at	9.12×10^{-3}	DNA-directed RNA polymerase activity, DNA binding(GO:0003899, GO:0003677)	transcription(GO:0006350)
Ta.3262.1.S1_at	9.72×10^{-3}	ATP phosphoribosyltransferase activity, signal transducer activity, GTP binding(GO:0003879, GO:0004871, GO:0005525)	G-protein coupled receptor protein signalling pathway, small GTPase mediated signal transduction, intracellular protein transport, histidine biosynthetic process(GO:0007186, GO:0007264, GO:0006886, GO:0000105)
Ta.24826.2.S1_x_at	1.03×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.28308.1.S1_at	1.09×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.20786.1.A1_at	1.15×10^{-2}		

Continued on next page

Ta.1335.1.S1_at	1.21×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.10191.1.S1_at	1.26×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.19112.1.S1_at	1.32×10^{-2}	unfolded protein binding, ATP binding(GO:0051082, GO:0005524)	protein folding(GO:0006457)
TaAffx.18904.1.S1_at	1.37×10^{-2}	protein serine/threonine/tyrosine kinase activity, ATP binding(GO:0004712, GO:0005524)	signal transduction, protein phosphorylation(GO:0007165, GO:0006468)
Ta.2887.1.S1_at	1.42×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.621.1.S1_x_at	1.48×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
TaAffx.92235.1.A1_at	1.53×10^{-2}		
Ta.10660.1.A1_at	1.58×10^{-2}		
TaAffx.40189.1.S1_at	1.63×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.8167.2.S1_at	1.68×10^{-2}		
Ta.24129.1.S1_at	1.73×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.22874.1.S1_x_at	1.79×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.2958.2.S1_at	1.84×10^{-2}		
Ta.3141.1.S1_x_at	1.89×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.39.4.S1_x_at	1.94×10^{-2}	protein binding, structural constituent of ribosome(GO:0005515, GO:0003735)	translation(GO:0006412)

Continued on next page

Ta.3152.1.S1_at	1.99×10^{-2}	hydrogen ion transporting ATP synthase activity, rotational mechanism, hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances, ATPase activity, ATP binding(GO:0046933, GO:0016820, GO:0016887, GO:0005524)	ATP synthesis coupled proton transport(GO:0015986)
Ta.4599.1.S1_at	2.04×10^{-2}	asparagine-tRNA ligase activity, nucleic acid binding, ATP binding(GO:0004816, GO:0003676, GO:0005524)	asparaginyl-tRNA aminoacylation(GO:0006421)
Ta.10344.1.A1_x_at	2.08×10^{-2}		
Ta.10818.1.S1_x_at	2.13×10^{-2}	protein binding(GO:0005515)	regulation of transcription from RNA polymerase II promoter(GO:0006357)
Ta.2605.2.S1_x_at	2.18×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28677.1.S1_x_at	2.23×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28658.1.S1_at	2.28×10^{-2}	APG8-specific protease activity, APG8 activating enzyme activity, Atg8 ligase activity(GO:0019786, GO:0019779, GO:0019776)	autophagy(GO:0006914)

Continued on next page

Ta.7657.1.S1_s_at	2.33×10^{-2}	coenzyme binding, UDP-arabinose 4-epimerase activity, dTDP-4-dehydrorhamnose reductase activity, 3-beta-hydroxy-delta5-steroid dehydrogenase activity(GO:0050662, GO:0050373, GO:0008831, GO:0003854)	steroid biosynthetic process, nucleotide-sugar metabolic process, extracellular polysaccharide biosynthetic process, oxidation-reduction process(GO:0006694, GO:0009225, GO:0045226, GO:0055114)
Ta.24981.2.S1_a_at	2.38×10^{-2}	structural constituent of ribosome, rRNA binding(GO:0003735, GO:0019843)	translation(GO:0006412)
Ta.24128.1.S1_x_at	2.42×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28341.2.S1_at	2.47×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28493.1.S1_at	2.52×10^{-2}	serine-type carboxypeptidase activity(GO:0004185)	proteolysis(GO:0006508)
TaAffx.4262.1.A1_at	2.57×10^{-2}		
Ta.19269.1.S1_x_at	2.61×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.27765.3.S1_at	2.66×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.23247.1.S1_x_at	2.71×10^{-2}	ubiquitin-protein ligase activity(GO:0004842)	auxin metabolic process, cytokinin metabolic process(GO:0009850, GO:0009690)
Ta.12395.1.S1_at	2.75×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.28673.1.S1_at	2.80×10^{-2}	proton-transporting ATPase activity, rotational mechanism, hydrogen ion transporting ATP synthase activity, rotational mechanism(GO:0046961, GO:0046933)	ATP synthesis coupled proton transport(GO:0015986)

Continued on next page

Ta.13731.1.S1_at	2.84×10^{-2}	protein binding, peptidyl-prolyl cis-trans isomerase activity(GO:0005515, GO:0003755)	protein folding(GO:0006457)
Ta.10487.1.A1_at	2.89×10^{-2}	heme binding(GO:0020037)	
Ta.21954.1.S1_at	2.94×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28712.1.S1_x_at	2.98×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.28712.1.S1_at	3.03×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.24248.2.S1_x_at	3.07×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.25124.2.S1_x_at	3.12×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.8792.2.A1_a_at	3.16×10^{-2}	catalytic activity(GO:0003824)	
Ta.6046.2.S1_a_at	3.21×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.21750.1.S1_x_at	3.25×10^{-2}		
Ta.7657.3.S1_a_at	3.29×10^{-2}	coenzyme binding, UDP-arabinose 4-epimerase activity(GO:0050662, GO:0050373)	plant-type cell wall biogenesis, nucleotide-sugar metabolic process, arabinose biosynthetic process(GO:0009832, GO:0009225, GO:0019567)
Ta.21021.2.S1_at	3.34×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to stress(GO:0006950)
Ta.2887.1.S1_x_at	3.38×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.2299.1.S1_at	3.43×10^{-2}	nucleic acid binding(GO:0003676)	mRNA processing, RNA splicing(GO:0006397, GO:0008380)
Ta.20465.1.S1_at	3.47×10^{-2}	cysteine-type endopeptidase activity(GO:0004197)	proteolysis, induction of apoptosis(GO:0006508, GO:0006917)

Continued on next page

Ta.2576.1.S1_at	3.51×10^{-2}	GTP binding, translation elongation factor activity(GO:0005525, GO:0003746)	
Ta.14519.1.S1_x_at	3.56×10^{-2}	chromatin binding, Ran GTPase binding(GO:0003682, GO:0008536)	response to UV-B(GO:0010224)
Ta.4419.1.S1_at	3.60×10^{-2}	calcium-dependent cysteine-type endopeptidase activity(GO:0004198)	proteolysis(GO:0006508)
TaAffx.33265.1.S1_at	3.64×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	defense response to bacterium(GO:0042742)
Ta.28699.1.S1_x_at	3.69×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.28493.1.S1_x_at	3.73×10^{-2}	serine-type carboxypeptidase activity(GO:0004185)	proteolysis(GO:0006508)
Ta.22984.3.S1_at	3.77×10^{-2}	nucleotide binding, nucleic acid binding(GO:0000166, GO:0003676)	leaf development, leaf vascular tissue pattern formation, endonucleolytic cleavage involved in rRNA processing, sepal vascular tissue pattern formation, root development, cotyledon vascular tissue pattern formation, petal vascular tissue pattern formation(GO:0048366, GO:0010305, GO:0000478, GO:0080057, GO:0048364, GO:0010588, GO:0080056)
Ta.13226.1.S1_at	3.82×10^{-2}		

Continued on next page

Ta.9436.1.S1_a_at	3.86×10^{-2}	pyridoxal phosphate binding, transaminase activity(GO:0030170, GO:0008483)	biosynthetic process(GO:0009058)
Ta.4676.3.S1_a_at	3.90×10^{-2}	receptor activity(GO:0004872)	response to sucrose stimulus, response to hormone stimulus(GO:0009744, GO:0009725)
Ta.10740.1.S1_a_at	3.95×10^{-2}	acyl-CoA thioesterase activity, 4-hydroxybenzoyl-CoA thioesterase activity(GO:0016291, GO:0018739)	
Ta.9829.1.A1_s_at	3.99×10^{-2}		
Ta.20677.1.S1_x_at	4.03×10^{-2}	nucleotide binding, nucleic acid binding(GO:0000166, GO:0003676)	
Ta.2982.2.S1_x_at	4.07×10^{-2}	protein binding, ferrous iron transmembrane transporter activity, GTP binding, ATPase activity(GO:0005515, GO:0015093, GO:0005525, GO:0016887)	response to cadmium ion(GO:0046686)
Ta.28257.1.S1_x_at	4.11×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translation(GO:0042254, GO:0006412)
Ta.1184.1.A1_at	4.16×10^{-2}		
Ta.18890.1.S1_at	4.20×10^{-2}	APG8 activating enzyme activity(GO:0019779)	
Ta.6282.1.S1_at	4.24×10^{-2}	protein transporter activity(GO:0008565)	protein import into nucleus, docking(GO:0000059)
Ta.28365.1.S1_s_at	4.28×10^{-2}	structural constituent of ribosome(GO:0003735)	ribosome biogenesis, translational elongation(GO:0042254, GO:0006414)

Continued on next page

Ta.5262.1.S1_at	4.32×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.15265.1.S1_at	4.36×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.2153.1.S1_s_at	4.41×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.12514.1.A1_s_at	4.45×10^{-2}	binding(GO:0005488)	
Ta.2605.2.S1_at	4.49×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.11161.1.A1_at	4.53×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.19269.1.S1_at	4.57×10^{-2}	structural constituent of ribosome(GO:0003735)	translation(GO:0006412)
Ta.24299.1.S1_at	4.61×10^{-2}	protein binding, structural constituent of ribosome(GO:0005515, GO:0003735)	protein neddylation, aging, embryo development ending in seed dormancy, protein ubiquitination involved in ubiquitin-dependent protein catabolic process, response to UV-B, response to salicylic acid stimulus, response to auxin stimulus(GO:0045116, GO:0007568, GO:0009793, GO:0042787, GO:0010224, GO:0009751, GO:0009733)
Ta.28855.2.S1_at	4.65×10^{-2}	structural constituent of ribosome, RNA binding(GO:0003735, GO:0003723)	translation(GO:0006412)
Ta.14018.2.S1_a_at	4.69×10^{-2}	unfolded protein binding(GO:0051082)	protein folding(GO:0006457)
Ta.12406.1.S1_at	4.73×10^{-2}	protein binding, ubiquitin-specific protease activity, ubiquitin thiolesterase activity(GO:0005515, GO:0004843, GO:0004221)	ubiquitin-dependent protein catabolic process(GO:0006511)

Continued on next page

Ta.9580.1.S1_at	4.78×10^{-2}	uridine nucleosidase activity, purine nucleosidase activity, inosine nucleosidase activity, adenosine nucleosidase activity(GO:0045437, GO:0008477, GO:0047724, GO:0047622)	
Ta.13302.1.S1_at	4.82×10^{-2}		
TaAffx.98034.1.S1_at	4.86×10^{-2}	translation initiation factor activity(GO:0003743)	
Ta.1819.1.S1_at	4.90×10^{-2}	hydrogen ion transmembrane transporter activity, hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances, ATP binding(GO:0015078, GO:0016820, GO:0005524)	ATP synthesis coupled proton transport(GO:0015986)
Ta.4700.1.S1_at	4.94×10^{-2}		
Ta.10929.1.S1_at	4.98×10^{-2}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	senescence, response to starvation, autophagy(GO:0010149, GO:0042594, GO:0006914)

Appendix 13: PCo2 from line.RWC analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score for the PCo are shown. For each gene PropF is the is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
TaAffx.3401.1.S1_at	2.40×10^{-3}	protein binding, small GTPase regulator activity(GO:0005515, GO:0005083)	vesicle-mediated transport, intracellular protein transport(GO:0016192, GO:0006886)
Ta.8283.1.S1_at	4.56×10^{-3}	sequence-specific DNA binding transcription factor activity(GO:0003700)	
Ta.8362.1.A1_at	6.68×10^{-3}	protein tyrosine/serine/threonine phosphatase activity, protein tyrosine phosphatase activity(GO:0008138, GO:0004725)	protein dephosphorylation(GO:0006470)
Ta.3242.1.A1_a_at	8.75×10^{-3}		
TaAffx.133282.1.A1_at	1.07×10^{-2}		
Ta.28458.2.S1_at	1.27×10^{-2}		
TaAffx.38601.1.A1_at	1.46×10^{-2}		
TaAffx.120714.1.S1_at	1.65×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	embryo development ending in seed dormancy(GO:0009793)
Ta.30902.1.S1_at	1.84×10^{-2}		
Ta.27678.1.S1_at	2.02×10^{-2}	ATPase activity, ATP binding, DNA binding(GO:0016887, GO:0005524, GO:0003677)	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, transport(GO:0000184, GO:0006810)

Continued on next page

Ta.1282.4.S1_at	2.20×10^{-2}	serine-type endopeptidase inhibitor activity, nutrient reservoir activity, lipid binding(GO:0004867, GO:0045735, GO:0008289)	lipid transport(GO:0006869)
Ta.12373.2.S1_x_at	2.38×10^{-2}		
Ta.30781.1.S1_at	2.57×10^{-2}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	
TaAffx.119948.1.A1_at	2.74×10^{-2}		
Ta.5553.1.S1_at	2.92×10^{-2}	omega peptidase activity(GO:0008242)	tetrahydrofolylpolyglutamate metabolic process, glutamine metabolic process(GO:0046900, GO:0006541)
TaAffx.36930.1.A1_at	3.10×10^{-2}		
Ta.26161.1.A1_at	3.27×10^{-2}		
Ta.25854.1.S1_at	3.45×10^{-2}	amino acid transmembrane transporter activity(GO:0015171)	
Ta.21472.1.S1_at	3.62×10^{-2}		
TaAffx.42782.1.A1_at	3.79×10^{-2}		
Ta.22208.1.S1_at	3.96×10^{-2}	phosphatidylinositol-4,5-bisphosphate 5-phosphatase activity, phosphatidylinositol-4-phosphate phosphatase activity(GO:0004439, GO:0043812)	root hair cell tip growth(GO:0048768)
Ta.3907.1.S1_at	4.13×10^{-2}	potassium:hydrogen antiporter activity(GO:0015386)	metabolic process(GO:0008152)
Ta.3242.1.A1_at	4.30×10^{-2}		
TaAffx.56933.1.S1_at	4.46×10^{-2}		
Ta.20175.1.A1_at	4.63×10^{-2}		

Continued on next page

TaAffx.57631.1.S1_at	4.79×10^{-2}
TaAffx.97454.1.A1_at	4.95×10^{-2}
Ta.12298.1.A1_at	5.12×10^{-2}

Appendix 14: PCo3 from line.RWC analysis sorted by the genes contributing the most to the proportion of variance of this PCo. Only genes that amount to 10% of the total f-score for the PCo are shown. For each gene PropF is the is the proportion of the total f-score captured by the gene and the higher ranked genes, it therefore represents the proportion of the PCo that would be captured if the given gene was used as the cut-off point.

Gene	PropF	Molecular Function	Biological Process
Ta.16038.1.S1_at	2.61×10^{-3}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to cold, response to water deprivation, response to abscisic acid stimulus, hyperosmotic salinity response(GO:0009409, GO:0009414, GO:0009737, GO:0042538)
Ta.3452.3.S1_a_at	5.19×10^{-3}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	protein ubiquitination(GO:0016567)
Ta.23045.2.S1_x_at	7.73×10^{-3}	protein binding(GO:0005515)	aging, response to fungus, response to stress(GO:0007568, GO:0009620, GO:0006950)
Ta.27546.1.S1_at	9.98×10^{-3}	iron ion binding, oxidoreductase activity(GO:0005506, GO:0016491)	fatty acid biosynthetic process, oxidation-reduction process(GO:0006633, GO:0055114)
Ta.25228.1.S1_at	1.21×10^{-2}		
Ta.29434.1.A1_at	1.41×10^{-2}	enzyme inhibitor activity, pectinesterase activity(GO:0004857, GO:0030599)	
Ta.9830.1.A1_at	1.62×10^{-2}	calcium ion binding, heme oxygenase (decyclizing) activity, linoleic acid epoxygenase activity(GO:0005509, GO:0004392, GO:0071614)	oxylipin biosynthetic process, defense response(GO:0031408, GO:0006952)
Ta.26178.1.A1_at	1.81×10^{-2}	hydrolase activity(GO:0016787)	

Continued on next page

Ta.25997.1.A1_at	2.00×10^{-2}	pectate lyase activity(GO:0030570)	
TaAffx.128488.2.S1_s_at	2.18×10^{-2}	sequence-specific DNA binding transcription factor activity(GO:0003700)	response to cold, response to water deprivation, response to abscisic acid stimulus, hyperosmotic salinity response(GO:0009409, GO:0009414, GO:0009737, GO:0042538)
Ta.23045.1.S1_x_at	2.37×10^{-2}		
Ta.21146.1.S1_at	2.55×10^{-2}	protein kinase activity, choline kinase activity, ATP binding(GO:0004672, GO:0004103, GO:0005524)	response to wounding, protein phosphorylation(GO:0009611, GO:0006468)
Ta.6112.1.S1_at	2.73×10^{-2}		
TaAffx.97935.1.S1_at	2.91×10^{-2}	fatty-acyl-CoA synthase activity, 4-coumarate-CoA ligase activity(GO:0004321, GO:0016207)	
Ta.3502.1.S1_at	3.09×10^{-2}	binding(GO:0005488)	mitochondrial transport(GO:0006839)
Ta.7069.1.S1_at	3.26×10^{-2}	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor(GO:0016671)	cell redox homeostasis(GO:0045454)
Ta.23388.1.S1_at	3.43×10^{-2}	sequence-specific DNA binding transcription factor activity, oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen(GO:0003700, GO:0016702)	
TaAffx.114559.1.S1_x_at	3.60×10^{-2}	hydrolase activity(GO:0016787)	

Continued on next page

Ta.4760.1.S1_x_at	3.77×10^{-2}	calmodulin binding, calcium-transporting ATPase activity, ATP binding(GO:0005516, GO:0005388, GO:0005524)	ATP biosynthetic process, response to nematode(GO:0006754, GO:0009624)
Ta.28513.1.S1_s_at	3.94×10^{-2}	sequence-specific DNA binding transcription factor activity, protein dimerization activity(GO:0003700, GO:0046983)	leaf development, lateral root primordium development, response to ethylene stimulus(GO:0048366, GO:0010386, GO:0009723)
Ta.5457.1.S1_at	4.10×10^{-2}	protein binding, zinc ion binding(GO:0005515, GO:0008270)	
Ta.7571.1.S1_at	4.26×10^{-2}	ATP-dependent helicase activity, protein binding, nucleic acid binding, ATP binding(GO:0008026, GO:0005515, GO:0003676, GO:0005524)	
Ta.28335.1.S1_at	4.41×10^{-2}	transporter activity(GO:0005215)	transmembrane transport(GO:0055085)
Ta.4494.1.S1_x_at	4.57×10^{-2}	beta-amylase activity(GO:0016161)	response to water deprivation, starch catabolic process(GO:0009414, GO:0005983)
Ta.20707.2.S1_a_at	4.72×10^{-2}	organic anion transmembrane transporter activity(GO:0008514)	
Ta.28432.1.S1_at	4.87×10^{-2}		embryo development(GO:0009790)
TaAffx.114054.1.S1_at	5.01×10^{-2}	RNA binding, ribonuclease activity(GO:0003723, GO:0004540)	electron transport chain(GO:0022900)

Bibliography

ABE, H., URAO, T., ITO, T., SEKI, M., SHINOZAKI, K. AND YAMAGUCHI-SHINOZAKI, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling, *The Plant Cell Online* **15**(1): 63–78.

AFFYMETRIX (2011a). GeneChip expression analysis: Data analysis fundamentals. URL: `http://mmjggl.caltech.edu/microarray/data_analysis_fundamentals_manual.pdf`

AFFYMETRIX (2011b). Genechip® mouse expression set 430. URL: `http://www.affymetrix.com/support/help/faqs/mouse_430/faq_8.jsp`

AFFYMETRIX (2011c). Genechip® wheat genome array. URL: `http://media.affymetrix.com/support/technical/datasheets/wheat_datasheet.pdf`

AGRAWAL, G., IWAHASHI, H. AND RAKWAL, R. (2003). Rice MAPKs, *Biochemical and biophysical research communications* **302**(2): 171–80.

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. PMID: 2231712.

ALTSCHUL, S. F. AND KOONIN, E. V. (1998). Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases, *Trends in Biochemical Sciences* **23**(11): 444–447. PMID: 9852764.

ARANDA, B., ACHUTHAN, P., ALAM-FARUQUE, Y., ARMEAN, I., BRIDGE, A., DEROW, C., FEUERMANN, M., GHANBARIAN, A. T., KERRIEN, S., KHADAKE, J., KERSSEMAKERS, J., LEROY, C., MENDEN, M., MICHAUT, M., MONTECCHI-PALAZZI, L., NEUHAUSER, S. N., ORCHARD, S., PERREAU, V., ROECHERT, B.,

VAN EIJK, K. AND HERMJAKOB, H. (2010). The IntAct molecular interaction database in 2010, *Nucleic Acids Research* **38**(Database issue): D525–531. PMID: 19850723.

ASHBURNER, M. (1998). On the representation of "gene function" in databases.
URL: `http://www.geneontology.org/gene.ontology.discussion.shtml`

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. AND SHERLOCK, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium, *Nature Genetics* **25**(1): 25–29. PMID: 10802651.

BAIROCH, A. (2000). The ENZYME database in 2000, *Nucleic Acids Research* **28**(1): 304–305. PMID: 10592255.

BAITALUK, M., QIAN, X., GODBOLE, S., RAVAL, A., RAY, A. AND GUPTA, A. (2006). PathSys: integrating molecular interaction graphs for systems biology, *BMC Bioinformatics* **7**: 55. PMID: 16464251.

BARKER, G. (2010). Mining the wheat genome for useful polymorphisms.
URL: `http://www.wgin.org.uk/stakeholders/stakeholder2010/GaryBarkerWeb.pdf`

BARKER, G. (2011). Cerealsdb.
URL: `http://www.cerealsdb.uk.net/`

BATEMAN, A. (2004). The pfam protein families database, *Nucleic Acids Research* **32**: 138D–141.

BECK, E. H., FETTIG, S., KNAKE, C., HARTIG, K. AND BHATTARAI, T. (2007). Specific and unspecific responses of plants to cold and drought stress, *Journal of Biosciences* **32**(3): 501–510. PMID: 17536169.

BECKETT, D. AND MCBRIDE, B. (2004). RDF/XML syntax specification (Revised).
URL: `http://www.w3.org/TR/REC-rdf-syntax/`

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289–300.

BENJAMINI, Y. AND YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of statistics* **29**: 1165–1188.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. AND WHEELER, D. L. (2007). GenBank, *Nucleic Acids Research* **36**: D25–D30.

BERNARD, S. M. AND HABASH, D. Z. (2009). The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling, *New Phytologist* **182**(3): 608–620.

BERNERS-LEE, T. AND HENDLER, J. (2001). Publishing on the semantic web, *Nature* **410**(6832): 1023–1024.

BIOPAX (2011). BioPAX : Biological pathways exchange.
URL: `http://biopaxwiki.org`

BIOWISDOM (2009). SRS – specialist life science data integration.
URL: `http://www.biowisdom.com/2009/12/srs/`

BIRKLAND, A. AND YONA, G. (2006a). BIOZON: a hub of heterogeneous biological data, *Nucleic Acids Research* **34**(Database issue): D235–242. PMID: 16381854.

BIRKLAND, A. AND YONA, G. (2006b). BIOZON: a system for unification, management and analysis of heterogeneous biological data, *BMC Bioinformatics* **7**(1): 70.

BIRNBAUM, K., SHASHA, D. E., WANG, J. Y., JUNG, J. W., LAMBERT, G. M., GALBRAITH, D. W. AND BENFEY, P. N. (2003). A Gene Expression Map of the Arabidopsis Root, *Science* **302**(5652): 1956–1960.

BIRNEY, E., CLAMP, M. AND DURBIN, R. (2004). GeneWise and genomewise, *Genome Research* **14**(5): 988–995. PMID: 15123596.

BIZER, C., HEATH, T. AND BERNERS-LEE, T. (2009). Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems (IJSWIS)* .

BONNET, E., VAN DE PEER, Y. AND ROUZÉ, P. (2006). The small RNA world of plants, *New Phytologist* **171**(3): 451–468.

BORK, P. AND KOONIN, E. V. (1998). Predicting functions from protein sequences–where are the bottlenecks?, *Nature Genetics* **18**(4): 313–318. PMID: 9537411.

BRADER, G., DJAMEI, A., TEIGE, M., PALVA, E. T. AND HIRT, H. (2007). The MAP kinase kinase MKK2 affects disease resistance in arabidopsis, *Molecular Plant-Microbe Interactions: MPMI* **20**(5): 589–596. PMID: 17506336.

BRADY, S. M., ORLANDO, D. A., LEE, J.-Y., WANG, J. Y., KOCH, J., DINNENY, J. R., MACE, D., OHLER, U. AND BENFEY, P. N. (2007). A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns, *Science* **318**(5851): 801–806.

BREEZE, E., HARRISON, E., PAGE, T., WARNER, N., SHEN, C., ZHANG, C. AND BUCHANAN-WOLLASTON, V. (2008). Transcriptional regulation of plant senescence: from functional genomics to systems biology, *Plant Biology (Stuttgart, Germany)* **10**: 99–109. PMID: 18721315.

BRENNER, S. E. (1999). Errors in genome annotation, *Trends in Genetics: TIG* **15**(4): 132–133. PMID: 10203816.

BUNEMAN, P., KHANNA, S. AND TAN, W.-C. (2000). Data provenance: some basic issues, *Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science* pp. 87–93.

CALVANESE, D., DE GIACOMO, G., LEMBO, D., LENZERINI, M. AND ROSATI, R. (2009). Conceptual modeling for data integration, *in* A. T. Borgida, V. Chaudhri, P. Giorgini and E. Yu (eds), *Conceptual Modeling: Foundations and Applications - Essays in Honor of John Mylopoulos*, Vol. 5600 of *Lecture Notes in Computer Science*, Springer, pp. 173–197.

CARTER, G. W. (2005). Inferring network interactions within a cell, *Briefings in Bioinformatics* **6**(4): 380–389. PMID: 16420736.

CASPI, R. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Research* **34**: D511–D516.

CASPI, R., ALTMAN, T., DALE, J. M., DREHER, K., FULCHER, C. A., GILHAM, F., KAIPA, P., KARTHIKEYAN, A. S., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUELLER, L. A., PALEY, S., POPESCU, L., PUJAR, A., SHEARER, A. G., ZHANG, P. AND KARP, P. D. (2010). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases, *Nucleic Acids Research* **38**(suppl 1): D473–D479.

CASPI, R., FOERSTER, H., FULCHER, C. A., KAIPA, P., KRUMMENACKER, M., LATENDRESSE, M., PALEY, S., RHEE, S. Y., SHEARER, A. G., TISSIER, C., WALK, T. C., ZHANG, P. AND KARP, P. D. (2008). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome databases, *Nucleic Acids Research* **36**(Database issue): D623–D631. PMID: 17965431 PMCID: 2238876.

CHAPMAN, W. W. AND COHEN, K. B. (2009). Current issues in biomedical text mining and natural language processing, *Journal of Biomedical Informatics* **42**(5): 757–759. PMID: 19735740.

CHEFDOR, F., BÉNÉDETTI, H., DEPIERREUX, C., DELMOTTE, F., MORABITO, D. AND CARPIN, S. (2006). Osmotic stress sensing in populus: Components identification of a phosphorelay system, *FEBS Letters* **580**(1): 77–81.

CHEN, F., MACKEY, A. J., STOECKERT, JR, C. J. AND ROOS, D. S. (2006). Orthomcldb: querying a comprehensive multi-species collection of ortholog groups., *Nucleic Acids Res* **34**(Database issue): D363–D368.

CHEN, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids, *Annual Review of Plant Biology* **58**(1): 377–406.

CHINNUSAMY, V., SCHUMAKER, K. AND ZHU, J. K. (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants, *Journal of Experimental Botany* **55**(395): 225–236.

CHOI, H.-I., PARK, H., PARK, J. H., KIM, S., IM, M., SEO, H., KIM, Y., HWANG, I. AND KIM, S. Y. (2005). Arabidopsis calcium-dependent protein kinase AtCPK32 interacts with ABF4, a transcriptional regulator of abscisic acid-responsive gene expression, and modulates its activity, *Plant Physiology* **139**(4): 1750–1761. PMID: 16299177.

CHOO, Y.-I. (1982). Concurrency algebra and petri nets.
URL: http://resolver.caltech.edu/CaltechCSTR:1985.5190-tr-85

CHOUDHURY, A. AND LAHIRI, A. (2011). Comparative analysis of abscisic acid-regulated transcriptomes in arabidopsis., *Plant Biol (Stuttg)* **13**(1): 28–35.

CHOULET, F., WICKER, T., RUSTENHOLZ, C., PAUX, E., SALSE, J., LEROY, P., SCHLUB, S., LE PASLIER, M., MAGDELENAT, G., GONTHIER, C., COULOUX, A., BUDAK, H., BREEN, J., PUMPHREY, M., LIU, S., KONG, X., JIA, J., GUT, M., BRUNEL, D., ANDERSON, J. A., GILL, B. S., APPELS, R., KELLER, B. AND FEUILLET, C. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces, *The Plant Cell Online* **22**(6): 1686–1701.

CHRISTMANN, A., WEILER, E. W., STEUDLE, E. AND GRILL, E. (2007). A hydraulic signal in root-to-shoot signalling of water shortage, *The Plant Journal: For Cell and Molecular Biology* **52**(1): 167–174. PMID: 17711416.

COCHRANE, G. R. AND GALPERIN, M. Y. (2009). The 2010 nucleic acids research database issue and online database collection: a community of data resources, *Nucleic Acids Research* **38**: D1–D4.

CONESA, A. AND GÖTZ, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics, *International Journal of Plant Genomics* **2008**: 619832. PMID: 18483572.

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 (Oxford, England)* **21**(18): 3674–3676. PMID: 16081474.

CONLEY, E. J., NDUATI, V., GONZALEZ-HERNANDEZ, J. L., MESFIN, A., TRUDEAU-SPANJERS, M., CHAO, S., LAZO, G. R., HUMMEL, D. D., ANDERSON, O. D., QI, L. L., GILL, B. S., ECHALIER, B., LINKIEWICZ, A. M., DUBCOVSKY, J., AKHUNOV, E. D., DVORÁK, J., PENG, J. H., LAPITAN, N. L. V., PATHAN, M. S., NGUYEN, H. T., MA, X.-F., MIFTAHUDIN, GUSTAFSON, J. P., GREENE, R. A., SORRELLS, M. E., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., SIDHU, D.,

DILBIRLIGI, M., GILL, K. S., CHOI, D. W., FENTON, R. D., CLOSE, T. J., MCGUIRE, P. E., QUALSET, C. O. AND ANDERSON, J. A. (2004). A 2600-locus chromosome bin map of wheat homoeologous group 2 reveals interstitial gene-rich islands and colinearity with rice., *Genetics* **168**(2): 625–637.

CORAM, T. E., BROWN-GUEDIRA, G. AND CHEN, X. M. (2008). Using transcriptomics to understand the wheat genome, *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* **3**.

DAHLQUIST, K. D., SALOMONIS, N., VRANIZAN, K., LAWLOR, S. C. AND CONKLIN, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, *Nat Genet* **31**(1): 19–20.

DAI, M., WANG, P., BOYD, A. D., KOSTOV, G., ATHEY, B., JONES, E. G., BUNNEY, W. E., MYERS, R. M., SPEED, T. P., AKIL, H., WATSON, S. J. AND MENG, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, *Nucleic Acids Research* **33**(20): e175.

DATTA, R. S., MEACHAM, C., SAMAD, B., NEYER, C. AND SJÖLANDER, K. (2009a). Berkeley phog: Phylofacts orthology group prediction web server., *Nucleic Acids Res* **37**(Web Server issue): W84–W89.

DATTA, R. S., MEACHAM, C., SAMAD, B., NEYER, C. AND SJÃ¶LANDER, K. (2009b). Berkeley PHOG: PhyloFacts orthology group prediction web server, *Nucleic Acids Research* **37**(Web Server issue): W84–89. PMID: 19435885.

DAVIES, W. J. AND ZHANG, J. (1991). Root signals and the regulation of growth and development of plants in drying soil, *Annual Review of Plant Physiology and Plant Molecular Biology* **42**: 55–76.

DAVULURI, R., SUN, H., PALANISWAMY, S., MATTHEWS, N., MOLINA, C., KURTZ, M. AND GROTEWOLD, E. (2003). AGRIS: arabidopsis gene regulatory information server, an information resource of arabidopsis cis-regulatory elements and transcription factors, *BMC Bioinformatics* **4**(1): 25.

DE SMET, R. AND MARCHAL, K. (2010). Advantages and limitations of current network inference methods, *Nat Rev Micro* **8**(10): 717–729.

DEMIR, E., BABUR, O., DOGRUSOZ, U., GURSOY, A., AYAZ, A., GULESIR, G., NISANCI, G. AND CETIN-ATALAY, R. (2004). An ontology for collaborative construction and analysis of cellular pathways, *Bioinformatics* **20**(3): 349 –356.

DEMIR, E., BABUR, O., DOGRUSOZ, U., GURSOY, A., NISANCI, G., CETIN-ATALAY, R. AND OZTURK, M. (2002). Patika: an integrated visual environment for collaborative construction and analysis of cellular pathways, *Bioinformatics* **18**(7): 996 –1003.

DEMIR, E., CARY, M. P., PALEY, S., FUKUDA, K., LEMER, C., VASTRIK, I., WU, G., D'EUSTACHIO, P., SCHAEFER, C., LUCIANO, J., SCHACHERER, F., MARTINEZ-FLORES, I., HU, Z., JIMENEZ-JACINTO, V., JOSHI-TOPE, G., KANDASAMY, K., LOPEZ-FUENTES, A. C., MI, H., PICHLER, E., RODCHENKOV, I., SPLENDIANI, A., TKACHEV, S., ZUCKER, J., GOPINATH, G., RAJASIMHA, H., RAMAKRISHNAN, R., SHAH, I., SYED, M., ANWAR, N., BABUR, O., BLINOV, M., BRAUNER, E., CORWIN, D., DONALDSON, S., GIBBONS, F., GOLDBERG, R., HORNBECK, P., LUNA, A., MURRAY-RUST, P., NEUMANN, E., REUBENACKER, O., SAMWALD, M., VAN IERSEL, M., WIMALARATNE, S., ALLEN, K., BRAUN, B., WHIRL-CARRILLO, M., CHEUNG, K.-H., DAHLQUIST, K., FINNEY, A., GILLESPIE, M., GLASS, E., GONG, L., HAW, R., HONIG, M., HUBAUT, O., KANE, D., KRUPA, S., KUTMON, M., LEONARD, J., MARKS, D., MERBERG, D., PETRI, V., PICO, A., RAVENSCROFT, D., REN, L., SHAH, N., SUNSHINE, M., TANG, R., WHALEY, R., LETOVKSY, S., BUETOW, K. H., RZHETSKY, A., SCHACHTER, V., SOBRAL, B. S., DOGRUSOZ, U., MCWEENEY, S., ALADJEM, M., BIRNEY, E., COLLADO-VIDES, J., GOTO, S., HUCKA, M., LE NOVÈRE, N., MALTSEV, N., PANDEY, A., THOMAS, P., WINGENDER, E., KARP, P. D., SANDER, C. AND BADER, G. D. (2010). The biopax community standard for pathway data sharing., *Nat Biotechnol* **28**(9): 935–942.

DOGRUSOZ, U., ERSON, E. Z., GIRAL, E., DEMIR, E., BABUR, O., CETINTAS, A. AND COLAK, R. (2006). Patikaweb: a web interface for analyzing biological pathways through advanced querying and visualization, *Bioinformatics* **22**(3): 374 –375.

DRYANOVA, A., ZAKHAROV, A. AND GULICK, P. J. (2008). Data mining for miRNAs and their targets in the triticeae, *Genome / National Research Council Canada = GÃ©nome / Conseil National De Recherches Canada* **51**(6): 433–443. PMID: 18521122.

EDDY, S. R. (1990). Accelerated profile HMM searches.
URL: `ftp://selab.janelia.org/pub/publications/Eddy11/Eddy11-preprint.pdf`

EDDY, S. R. (2009). A new generation of homology search tools based on probabilistic inference, *Genome Informatics. International Conference on Genome Informatics* **23**(1): 205–211. PMID: 20180275.

EDEN, E., NAVON, R., STEINFELD, I., LIPSON, D. AND YAKHINI, Z. (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists., *BMC Bioinformatics* **10**: 48.

ESCOBAR, P. (2011). B2G-FAR: a species centered GO annotation repository.
URL: `http://bioinfo.cipf.es/b2gfar/home`

EULGEM, T., RUSHTON, P. J., ROBATZEK, S. AND SOMSSICH, I. E. (2000). The WRKY superfamily of plant transcription factors, *Trends in Plant Science* **5**(5): 199–206. PMID: 10785665.

FERNANDEZ, O., BÃ©THENCOURT, L., QUERO, A., SANGWAN, R. S. AND CLÃ©MENT, C. (2010). Trehalose and plant stress responses: friend or foe?, *Trends in Plant Science* **15**: 409–417.

FEUILLET, C. AND MUEHLBAUER, G. J. (2009). *Genetics and Genomics of the Triticeae*, Springer.

FINN, R. D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J. E., GAVIN, O. L., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R. AND BATEMAN, A. (2009). The pfam protein families database, *Nucleic Acids Research* **38**(Database): D211–D222.

FITCH, W. M. (2000). Homology a personal view on some of the problems, *Trends in Genetics: TIG* **16**(5): 227–231. PMID: 10782117.

FOUNDATION, T. A. S. (2011). Apache lucene.
URL: `http://lucene.apache.org/java/docs/index.html`

FRICKEY, T., BENEDITO, V. A., UDVARDI, M. AND WEILLER, G. (2008). AffyTrees: facilitating comparative analysis of affymetrix plant microarray chips, *Plant Physiology* **146**(2): 377–386. PMID: 18065560 PMCID: 2245853.

FUJII, H., VERSLUES, P. E. AND ZHU, J. (2007). Identification of two protein kinases required for abscisic acid regulation of seed germination, root growth, and gene expression in arabidopsis, *The Plant Cell Online* **19**(2): 485–494.

FUJITA, Y., NAKASHIMA, K., YOSHIDA, T., KATAGIRI, T., KIDOKORO, S., KANAMORI, N., UMEZAWA, T., FUJITA, M., MARUYAMA, K., ISHIYAMA, K., KOBAYASHI, M., NAKASONE, S., YAMADA, K., ITO, T., SHINOZAKI, K. AND YAMAGUCHI-SHINOZAKI, K. (2009). Three SnRK2 protein kinases are the main positive regulators of abscisic acid signaling in response to water stress in arabidopsis, *Plant and Cell Physiology* **50**(12): 2123–2132.

GABALDÓN, T. (2008). Large-scale assignment of orthology: back to phylogenetics?, *Genome Biol* **9**(10): 235.

GALPERIN, M. Y. (2007). The molecular biology database collection: 2008 update, *Nucleic Acids Research* **36**: D2–D4.

GAO, X. P., WANG, X. F., LU, Y. F., ZHANG, L. Y., SHEN, Y. Y., LIANG, Z. AND ZHANG, D. P. (2004). Jasmonic acid is involved in the water-stress-induced betaine accumulation in pear leaves, *Plant, Cell & Environment* **27**(4): 497–507.

GAO, X. AND SONG, P. X. K. (2005). Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments., *BMC Bioinformatics* **6**: 186.

GAO, Y., ZENG, Q., GUO, J., CHENG, J., ELLIS, B. E. AND CHEN, J. (2007). Genetic characterization reveals no role for the reported ABA receptor, GCR2, in ABA control of seed germination and early seedling development in arabidopsis, *The Plant Journal: For Cell and Molecular Biology* **52**(6): 1001–1013. PMID: 17894782.

GARCIA-HERNANDEZ, M., BERARDINI, T. Z., CHEN, G., CRIST, D., DOYLE, A., HUALA, E., KNEE, E., LAMBRECHT, M., MILLER, N., MUELLER, L. A., MUNDODI, S., REISER, L., RHEE, S. Y., SCHOLL, R., TACKLIND, J., WEEMS, D. C., WU, Y., XU, I., YOO, D., YOON, J. AND ZHANG, P. (2002). TAIR: a resource for integrated arabidopsis data, *Functional & Integrative Genomics* **2**(6): 239–253. PMID: 12444417.

GASCUEL, O. AND STEEL, M. (2006). Neighbor-joining revealed., *Mol Biol Evol* **23**(11): 1997–2000.

GASTEIGER, E., GATTIKER, A., HOOGLAND, C., IVANYI, I., APPEL, R. D. AND BAIROCH, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis, *Nucleic Acids Research* **31**(13): 3784–3788. PMID: 12824418.

GAUT, B. S. (2002). Evolutionary dynamics of grass genomes, *New Phytologist* **154**(1): 15–28.

GEIGER, D., SCHERZER, S., MUMM, P., STANGE, A., MARTEN, I., BAUER, H., ACHE, P., MATSCHI, S., LIESE, A., AL-RASHEID, K. A. S., ROMEIS, T. AND HEDRICH, R. (2009). Activity of guard cell anion channel SLAC1 is controlled by drought-stress signaling kinase-phosphatase pair, **106**(50): 21425–21430. PMID: 19955405 PMCID: 2795561.

GENETIC, W. AND CENTER, G. R. (2011). Current theory of the evolution of wheat. URL: `http://www.k-state.edu/wgrc/Extras/evolve.html`

GILES, P. J. AND KIPLING, D. (2003). Normality of oligonucleotide microarray data and implications for parametric statistical analyses., *Bioinformatics* **19**(17): 2254–2262.

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.

GILKS, W. R., AUDIT, B., DE ANGELIS, D., TSOKA, S. AND OUZOUNIS, C. A. (2005). Percolation of annotation errors through hierarchically structured protein sequence databases, *Mathematical Biosciences* **193**(2): 223–234.

GILL, B. S., APPELS, R., BOTHA-OBERHOLSTER, A., BUELL, C. R., BENNETZEN, J. L., CHALHOUB, B., CHUMLEY, F., DVOÅ™Ã¡K, J., IWANAGA, M., KELLER, B., LI, W., MCCOMBIE, W. R., OGIHARA, Y., QUETIER, F. AND SASAKI, T. (2004). A workshop report on wheat genome sequencing, *Genetics* **168**(2): 1087–1096.

GISK, B., YASUI, Y., KOHCHI, T. AND FRANKENBERG-DINKEL, N. (2010). Characterization of the haem oxygenase protein family in arabidopsis thaliana reveals a diversity of functions, *The Biochemical Journal* **425**(2): 425–434. PMID: 19860740.

GOPALACHARYULU, P. V., LINDFORS, E., BOUNSAYTHIP, C., KIVIOJA, T., YETUKURI, L., HOLLMEN, J. AND ORESIC, M. (2005). Data integration and visualization system for enabling conceptual biology, *Bioinformatics* **21**: i177–i185.

GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**(3-4): 325–338.

GRBIĆ, V. AND BLEECKER, A. B. (1995). Ethylene regulates the timing of leaf senescence in arabidopsis, *The Plant Journal* **8**(4): 595–602.

GREGERSEN, P. L. AND HOLM, P. B. (2007). Transcriptome analysis of senescence in the flag leaf of wheat (Triticum aestivum l.), *Plant Biotechnology Journal* **5**(1): 192–206. PMID: 17207268.

GROUP, W. O. W. (2009). OWL 2 web ontology language document overview.
URL: `http://www.w3.org/TR/owl2-overview/`

GUO, Y. AND GAN, S. (2006). AtNAP, a NAC family transcription factor, has an important role in leaf senescence, *The Plant Journal: For Cell and Molecular Biology* **46**(4): 601–612. PMID: 16640597.

GUPTA, R., KIENZLER, K., MARTIUS, C., MIRZABAEV, A., OWEIS, T., DE PAUW, E., QADIR, M., SHIDEED, K., SOMMER, R., THOMAS, R., SAYRE, K., CARLI, C., SAPAROV, A., BEKENOV, M., SANGINOV, S., NEPESOV, M. AND IKRAMOV, R. (2009). Research prospectus: A vision for sustainable land management research in central asia.
URL: `www.icarda.org/cac/files/sacac/SACAC-01_research-prospectus_eng.pdf`

GUTMAN, B. L. AND NIYOGI, K. K. (2009). Evidence for base excision repair of oxidative dna damage in chloroplasts of arabidopsis thaliana., *J Biol Chem* **284**(25): 17006–17012.

HABASH, D. Z., KEHEL, Z. AND NACHIT, M. (2009). Genomic approaches for designing durum wheat ready for climate change with a focus on drought., *J Exp Bot* **60**(10): 2805–2815.

HANKS, S. K. AND QUINN, A. M. (1991). Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members, *Methods in Enzymology* **200**: 38–62. PMID: 1956325.

HANZAWA, Y., MONEY, T. AND BRADLEY, D. (2005). A single amino acid converts a repressor to an activator of flowering., *Proc Natl Acad Sci U S A* **102**(21): 7748–7753.

HARDIN, J. AND WILSON, J. (2009). A note on oligonucleotide expression values not being normally distributed, *Biostatistics* **10**(3): 446–450.

HARDING, H. P., NOVOA, I., ZHANG, Y., ZENG, H., WEK, R., SCHAPIRA, M. AND RON, D. (2000). Regulated translation initiation controls stress-induced gene expression in mammalian cells, *Molecular Cell* **6**(5): 1099–1108. PMID: 11106749.

HARMON, A. C., GRIBSKOV, M. AND HARPER, J. F. (2000). CDPKs - a kinase for every ca2+ signal?, *Trends in Plant Science* **5**(4): 154–159. PMID: 10740296.

HARTUNG, W., SAUTER, A., TURNER, N. C., FILLERY, I. AND HEILMEIER, H. (1996). Abscisic acid in soils: What is its function and which factors and mechanisms influence its concentration?, *Plant and Soil* **184**: 105–110.

HEGYI, H. AND GERSTEIN, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome., *J Mol Biol* **288**(1): 147–164.

HEGYI, H. AND GERSTEIN, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins., *Genome Res* **11**(10): 1632–1640.

HERMJAKOB, H., MONTECCHI-PALAZZI, L., BADER, G., WOJCIK, J., SALWINSKI, L., CEOL, A., MOORE, S., ORCHARD, S., SARKANS, U., VON MERING, C., ROECHERT, B., POUX, S., JUNG, E., MERSCH, H., KERSEY, P., LAPPE, M., LI, Y., ZENG, R., RANA, D., NIKOLSKI, M., HUSI, H., BRUN, C., SHANKER, K., GRANT, S. G. N., SANDER, C., BORK, P., ZHU, W., PANDEY, A., BRAZMA, A., JACQ, B., VIDAL, M., SHERMAN, D., LEGRAIN, P., CESARENI, G., XENARIOS, I., EISENBERG, D., STEIPE, B., HOGUE, C. AND APWEILER, R. (2004). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data, *Nature Biotechnology* **22**(2): 177–183. PMID: 14755292.

HIRAYAMA, T. AND SHINOZAKI, K. (2007). Perception and transduction of abscisic acid signals: keys to the function of the versatile plant hormone ABA, *Trends in Plant Science* **12**(8): 343–351. PMID: 17629540.

HIRSCHMAN, L., COLOSIMO, M., MORGAN, A. AND YEH, A. (2005). Overview of biocreative task 1b: normalized gene lists., *BMC Bioinformatics* **6 Suppl 1**: S11.

HOCH, J. A. AND VARUGHESE, K. I. (2001). Keeping signals straight in phosphorelay signal transduction, *J. Bacteriol.* **183**(17): 4941–4949.

HOGENBOOM, F., MILEA, V., FRASINCAR, F. AND KAYMAK, U. (2010). RDF-GL: a SPARQL-Based graphical query language for RDF, *in* R. Chbeir, Y. Badr, A. Abraham and A. Hassanien (eds), *Emergent Web Intelligence: Advanced Information Retrieval*, Springer London, London, pp. 87–116.

HOLBROOK, N. M., SHASHIDHAR, V., JAMES, R. A. AND MUNNS, R. (2002). Stomatal control in tomato with ABA-deficient roots: response of grafted plants to soil drying, *Journal of Experimental Botany* **53**(373): 1503 –1514.

HU, C., PETTITT, B. AND ROESGEN, J. (2009). Osmolyte solutions and protein folding, *F1000 Biology Reports* **1**(41).

HULO, N., SIGRIST, C. J. A., LE SAUX, V., LANGENDIJK-GENEVAUX, P. S., BORDOLI, L., GATTIKER, A., DE CASTRO, E., BUCHER, P. AND BAIROCH, A. (2004). Recent improvements to the PROSITE database, *Nucleic Acids Research* **32**(Database issue): D134–137. PMID: 14681377.

HUNTER, S., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BORK, P., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LAUGRAUD, A., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., MULDER, N., NATALE, D., ORENGO, C., QUINN, A. F., SELENGUT, J. D., SIGRIST, C. J. A., THIMMA, M., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C. H. AND YEATS, C. (2009). InterPro: the integrative protein signature database, *Nucleic Acids Research* **37**(Database issue): D211–215. PMID: 18940856.

IMBER, D. AND TAL, M. (1970). Phenotypic reversion of flacca, a wilty mutant of tomato, by abscisic acid, *Science* **169**(3945): 592 –593.

ISAAC, A. AND SUMMERS, E. (2009). SKOS simple knowledge organization system primer.
URL: `http://www.w3.org/TR/skos-primer/`

ITURRIAGA, G., SUÃ¡REZ, R. AND NOVA-FRANCO, B. (2009). Trehalose metabolism: From osmoprotection to signaling, *International Journal of Molecular Sciences* **10**: 3793–3810.

IUCHI, S., KOBAYASHI, M., TAJI, T., NARAMOTO, M., SEKI, M., KATO, T., TABATA, S., KAKUBARI, Y., YAMAGUCHI-SHINOZAKI, K. AND SHINOZAKI, K. (2001). Regulation of drought tolerance by gene manipulation of 9-cis-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in arabidopsis, *The Plant Journal: For Cell and Molecular Biology* **27**(4): 325–333. PMID: 11532178.

JAISWAL, P., AVRAHAM, S., ILIC, K., KELLOGG, E. A., MCCOUCH, S., PUJAR, A., REISER, L., RHEE, S. Y., SACHS, M. M., SCHAEFFER, M., STEIN, L., STEVENS, P., VINCENT, L., WARE, D. AND ZAPATA, F. (2005). Plant ontology (PO): a controlled vocabulary of plant structures and growth stages, *Comparative and Functional Genomics* **6**(7-8): 388–397. PMID: 18629207.

JANG, J. Y., KIM, D. G., KIM, Y. O., KIM, J. S. AND KANG, H. (2004). An expression analysis of a gene family encoding plasma membrane aquaporins in response to abiotic stresses in arabidopsis thaliana, *Plant Molecular Biology* **54**: 713–725.

JENSEN, M. K., KJAERSGAARD, T., PETERSEN, K. AND SKRIVER, K. (2010). NAC genes: time-specific regulators of hormonal signaling in arabidopsis, *Plant Signaling & Behavior* **5**(7): 907–910. PMID: 20484991.

JIANG, F. AND HARTUNG, W. (2008). Long-distance signalling of abscisic acid (ABA): the factors regulating the intensity of the ABA signal, *Journal of Experimental Botany* **59**(1): 37–43.

KANEHISA, M. AND GOTO, S. (2000). KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Research* **28**(1): 27–30. PMID: 10592173.

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. AND HIRAKAWA, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Research* **38**(suppl 1): D355–D360.

KANEHISA, M., GOTO, S., HATTORI, M., AOKI-KINOSHITA, K. F., ITOH, M., KAWASHIMA, S., KATAYAMA, T., ARAKI, M. AND HIRAKAWA, M. (2006). From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Research* **34**(Database issue): D354–357. PMID: 16381885.

KARLIN, S. AND ALTSCHUL, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proceedings of the National Academy of Sciences of the United States of America* **87**(6): 2264–2268. PMID: 2315319.

KARP, P. D. (1998). What we do not know about sequence analysis and sequence databases., *Bioinformatics* **14**(9): 753–754.

KARP, P. D., RILEY, M., PALEY, S. M. AND PELLEGRINI-TOOLE, A. (2002). The MetaCyc database, *Nucleic Acids Research* **30**(1): 59–61. PMID: 11752254.

KHATRI, P. AND DRĂGHICI, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics* **21**(18): 3587–3595.

KÖHLER, J., BAUMBACH, J., TAUBERT, J., SPECHT, M., SKUSA, A., RÜEGG, A., RAWLINGS, C., VERRIER, P. AND PHILIPPI, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX, *Bioinformatics* **22**(11): 1383–1390.

KIRITCHENKO, S., MATWIN, S. AND FAMILI, A. F. (2005). Functional annotation of genes using hierarchical text categorization, *BioLINK SIG: Linking Literature, Information and Knowledge for Biology* .

KOIWAI, H., NAKAMINAMI, K., SEO, M., MITSUHASHI, W., TOYOMASU, T. AND KOSHIBA, T. (2004). Tissue-specific localization of an abscisic acid biosynthetic enzyme, aao3, in arabidopsis, *Plant Physiology* **134**(4): 1697–1707.

KOONIN, E. V. AND GALPERIN, M. Y. (2003). *Sequence - Evolution - Function*, Kluwer Academic.

KROGH, A., LARSSON, B., VON HEIJNE, G. AND SONNHAMMER, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, *Journal of Molecular Biology* **305**(3): 567–580. PMID: 11152613.

KUSHIRO, T., OKAMOTO, M., NAKABAYASHI, K., YAMAGISHI, K., KITAMURA, S., ASAMI, T., HIRAI, N., KOSHIBA, T., KAMIYA, Y. AND NAMBARA, E. (2004). The arabidopsis cytochrome p450 CYP707A encodes ABA 8[prime]-hydroxylases: key enzymes in ABA catabolism, *EMBO J* **23**(7): 1647–1656.

LAU, H. T. (2007). *A Java library of graph algorithms and optimizations*, Chapman & Hall.

LE NOVÈRE, N. (2006). Model storage, exchange and integration., *BMC Neurosci* **7 Suppl 1**: S11.

LEE, S. C., LAN, W., BUCHANAN, B. B. AND LUAN, S. (2009). A protein kinase-phosphatase pair interacts with an ion channel to regulate ABA signaling in plant guard cells, *Proceedings of the National Academy of Sciences* **106**: 21419–24.

LEUNG, J., BOUVIER-DURAND, M., MORRIS, P., GUERRIER, D., CHEFDOR, F. AND GIRAUDAT, J. (1994). Arabidopsis ABA response gene ABI1: features of a calcium-modulated protein phosphatase, *Science* **264**(5164): 1448–1452.

LI, J., WANG, X., WATSON, M. B. AND ASSMANN, S. M. (2000). Regulation of abscisic Acid-Induced stomatal closure and anion channels by guard cell AAPK kinase, *Science* **287**(5451): 300–303.

LI, L., STOECKERT, CHRISTIAN J, J. AND ROOS, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Research* **13**(9): 2178–2189. PMID: 12952885.

LIU, G., LORAINE, A. E., SHIGETA, R., CLINE, M., CHENG, J., VALMEEKAM, V., SUN, S., KULP, D. AND SIANI-ROSE, M. A. (2003). NetAffx: affymetrix probesets and annotations, *Nucleic Acids Research* **31**(1): 82–86. PMID: 12519953.

LIU, H., SACHIDANANDAM, R. AND STEIN, L. (2001). Comparative genomics between rice and arabidopsis shows scant collinearity in gene order, *Genome Research* **11**(12): 2029–2026. PMID: 11731491.

LIU, X., YUE, Y., LI, B., NIE, Y., LI, W., WU, W. AND MA, L. (2007). A g Protein-Coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid, *Science* **315**(5819): 1712–1716.

LORD, P. W., STEVENS, R. D., BRASS, A. AND GOBLE, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation, *Bioinformatics (Oxford, England)* **19**(10): 1275–1283. PMID: 12835272.

LYSENKO, A., HINDLE, M. M., TAUBERT, J., SAQI, M. AND RAWLINGS, C. J. (2010). Data integration for plant genomics–exemplars from the integration of arabidopsis thaliana databases, *Briefings in Bioinformatics* **11**(2): 265.

MA, Y., SZOSTKIEWICZ, I., KORTE, A., MOES, D., YANG, Y., CHRISTMANN, A. AND GRILL, E. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors, *Science* **324**(5930): 1064–1068.

MACKNIGHT, R., BANCROFT, I., PAGE, T., LISTER, C., SCHMIDT, R., LOVE, K., WESTPHAL, L., MURPHY, G., SHERSON, S., COBBETT, C. AND DEAN, C. (1997).

FCA, a gene controlling flowering time in arabidopsis, encodes a protein containing RNA-binding domains, *Cell* **89**(5): 737–745. PMID: 9182761.

MAGLOTT, D., OSTELL, J., PRUITT, K. D. AND TATUSOVA, T. (2005). Entrez gene: gene-centered information at NCBI, *Nucleic Acids Research* **33**(Database issue): D54–58. PMID: 15608257.

MARIAUX, J. B., BOCKEL, C., SALAMINI, F. AND BARTELS, D. (1998). Desiccation- and abscisic acid-responsive genes encoding major intrinsic proteins (MIPs) from the resurrection plant craterostigma plantagineum, *Plant Molecular Biology* **38**(6): 1089–1099. PMID: 9869415.

MARMORSTEIN, R. (2003). Structure of SET domain proteins: a new twist on histone methylation, *Trends in Biochemical Sciences* **28**(2): 59–62. PMID: 12575990.

MATTHEWS, L., GOPINATH, G., GILLESPIE, M., CAUDY, M., CROFT, D., DE BONO, B., GARAPATI, P., HEMISH, J., HERMJAKOB, H., JASSAL, B., KANAPIN, A., LEWIS, S., MAHAJAN, S., MAY, B., SCHMIDT, E., VASTRIK, I., WU, G., BIRNEY, E., STEIN, L. AND D'EUSTACHIO, P. (2009). Reactome knowledgebase of human biological pathways and processes, *Nucleic Acids Research* **37**(Database issue): D619–622. PMID: 18981052.

MAUREL, C., KADO, R. T., GUERN, J. AND CHRISPEELS, M. J. (1995). Phosphorylation regulates the water channel activity of the seed-specific aquaporin alpha-TIP., *The EMBO Journal* **14**(13): 3028–3035. PMID: 7542585 PMCID: 394363.

MELCHER, K., NG, L., ZHOU, X. E., SOON, F., XU, Y., SUINO-POWELL, K. M., PARK, S., WEINER, J. J., FUJII, H., CHINNUSAMY, V., KOVACH, A., LI, J., WANG, Y., LI, J., PETERSON, F. C., JENSEN, D. R., YONG, E., VOLKMAN, B. F., CUTLER, S. R., ZHU, J. AND XU, H. E. (2009). A gate-latch-lock mechanism for hormone signalling by abscisic acid receptors, *Nature* **462**(7273): 602–608.

MERKEEV, I. V., NOVICHKOV, P. S. AND MIRONOV, A. A. (2006). Phog: a database of supergenomes built from proteome complements., *BMC Evol Biol* **6**: 52.

MICHAELSON, J. J., LOGUERCIO, S. AND BEYER, A. (2009). Detection and interpretation of expression quantitative trait loci (eqtl)., *Methods* **48**(3): 265–276.

MIFLIN, B. J. AND HABASH, D. Z. (2002). The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops, *Journal of Experimental Botany* **53**(370): 979–987.

MIYAZONO, K.-I., MIYAKAWA, T., SAWANO, Y., KUBOTA, K., KANG, H., ASANO, A., MIYAUCHI, Y., TAKAHASHI, M., ZHI, Y., FUJITA, Y., YOSHIDA, T., KODAIRA, K., YAMAGUCHI-SHINOZAKI, K. AND TANOKURA, M. (2009). Structural basis of abscisic acid signalling, *Nature* **462**(7273): 609–614.

MIZRAHI, Y., BLUMENFELD, A. AND RICHMOND, A. E. (1970). Abscisic acid and transpiration in leaves in relation to osmotic root stress, *Plant Physiology* **46**(1): 169–171. PMID: 16657411 PMCID: 396553.

MÜLLER, A. H. AND HANSSON, M. (2009). The barley magnesium chelatase 150-kD subunit is not an abscisic acid receptor, *Plant Physiology* **150**(1): 157–166.

MOCHIDA, K. AND SHINOZAKI, K. (2010). Genomics and bioinformatics resources for crop improvement, *Plant and Cell Physiology* **51**(4): 497–523.

MOCHIDA, K., YAMAZAKI, Y. AND OGIHARA, Y. (2003). Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags, *Molecular Genetics and Genomics: MGG* **270**(5): 371–377. PMID: 14595557.

MOLINA, F., DEHMER, M., PERCO, P., GRABER, A., GIROLAMI, M., SPASOVSKI, G., SCHANSTRA, J. P. AND VLAHOU, A. (2010). Systems biology: opening new avenues in clinical research, *Nephrology Dialysis Transplantation* **25**(4): 1015–1018.

MONS, B. AND VELTEROP, J. (2009). Nano-publication in the e-science era, *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse* **523**.

MORI, I. C., MURATA, Y., YANG, Y., MUNEMASA, S., WANG, Y., ANDREOLI, S., TIRIAC, H., ALONSO, J. M., HARPER, J. F., ECKER, J. R., KWAK, J. M. AND

SCHROEDER, J. I. (2006). CDPKs CPK6 and CPK3 function in ABA regulation of guard cell s-type anion- and ca(2+)-permeable channels and stomatal closure, *PLoS Biology* **4**(10): e327. PMID: 17032064.

MUELLER, L. A., SOLOW, T. H., TAYLOR, N., SKWARECKI, B., BUELS, R., BINNS, J., LIN, C., WRIGHT, M. H., AHRENS, R., WANG, Y., HERBST, E. V., KEYDER, E. R., MENDA, N., ZAMIR, D. AND TANKSLEY, S. D. (2005). The SOL genomics network. a comparative resource for solanaceae biology and beyond, *Plant Physiology* **138**(3): 1310–1317.

MUELLER, L. A., ZHANG, P. AND RHEE, S. Y. (2003). AraCyc: a biochemical pathway database for arabidopsis, *Plant Physiology* **132**(2): 453–460.

MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BORK, P., BUILLARD, V., CERUTTI, L., COPLEY, R., COURCELLE, E., DAS, U., DAUGHERTY, L., DIBLEY, M., FINN, R., FLEISCHMANN, W., GOUGH, J., HAFT, D., HULO, N., HUNTER, S., KAHN, D., KANAPIN, A., KEJARIWAL, A., LABARGA, A., LANGENDIJK-GENEVAUX, P. S., LONSDALE, D., LOPEZ, R., LETUNIC, I., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., NIKOLSKAYA, A. N., ORCHARD, S., ORENGO, C., PETRYSZAK, R., SELENGUT, J. D., SIGRIST, C. J. A., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C. H. AND YEATS, C. (2007). New developments in the InterPro database, *Nucleic Acids Research* **35**(Database issue): D224–228. PMID: 17202162.

MUSTILLI, A.-C., MERLOT, S., VAVASSEUR, A., FENZI, F. AND GIRAUDAT, J. (2002). Arabidopsis ost1 protein kinase mediates the regulation of stomatal aperture by abscisic acid and acts upstream of reactive oxygen species production, *The Plant Cell Online* **14**(12): 3089–3099.

NAMBARA, E. AND MARION-POLL, A. (2005). Abscisic acid biosynthesis and catabolism, *Annual Review of Plant Biology* **56**: 165–185. PMID: 15862093.

NC-IUBMB (1999). Nomenclature committee of the international union of biochemistry and molecular biology (nc-iubmb), enzyme supplement 5 (1999), *European Journal of Biochemistry* **264**: 610–650.

NCBI (2011a). Homologene build procedure.
URL: http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html

NCBI (2011b). Sequence identifiers: A historical note.
URL: http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html

NEELIN, J. D., MÜNNICH, M., SU, H., MEYERSON, J. E. AND HOLLOWAY, C. E. (2006). Tropical drying trends in global warming models and observations, *Proceedings of the National Academy of Sciences* **103**(16): 6110–6115.

NEGI, J., MATSUDA, O., NAGASAWA, T., OBA, Y., TAKAHASHI, H., KAWAI-YAMADA, M., UCHIMIYA, H., HASHIMOTO, M. AND IBA, K. (2008). CO2 regulator SLAC1 and its homologues are essential for anion homeostasis in plant cells, *Nature* **452**(7186): 483–486.

NEUBURGER, M., RÉBEILLÉ, F., JOURDAIN, A., NAKAMURA, S. AND DOUCE, R. (1996). Mitochondria are a major site for folate and thymidylate synthesis in plants., *J Biol Chem* **271**(16): 9466–9472.

NEVO, E., KOROL, A. B. AND FAHIMA, T. (2003). *Evolution of wild emmer and wheat improvement: population genetics, genetic resources, and genome organization of wheat's progenitor, Triticum dicoccoides*, Springer, Berlin.

NICOLAE, D. L., GAMAZON, E., ZHANG, W., DUAN, S., DOLAN, M. E. AND COX, N. J. (2010). Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas., *PLoS Genet* **6**(4): e1000888.

NISHIMURA, N., HITOMI, K., ARVAI, A. S., RAMBO, R. P., HITOMI, C., CUTLER, S. R., SCHROEDER, J. I. AND GETZOFF, E. D. (2009). Structural mechanism of abscisic acid binding and signaling by dimeric PYR1, *Science* **326**(5958): 1373–1379.

NOVOA, I., ZHANG, Y., ZENG, H., JUNGREIS, R., HARDING, H. P. AND RON, D. (2003). Stress-induced gene expression requires programmed recovery from translational repression, *EMBO J* **22**(5): 1180–1187.

OKAMOTO, M., KUWAHARA, A., SEO, M., KUSHIRO, T., ASAMI, T., HIRAI, N., KAMIYA, Y., KOSHIBA, T. AND NAMBARA, E. (2006). CYP707A1 and CYP707A2, which encode abscisic acid 8'-Hydroxylases, are indispensable for proper control of seed dormancy and germination in arabidopsis, *Plant Physiology* **141**(1): 97–107.

ORACLE (2011). Oracle berkeley DB.
URL: `http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html`

OSTLUND, G., SCHMITT, T., FORSLUND, K., KÖSTLER, T., MESSINA, D. N., ROOPRA, S., FRINGS, O. AND SONNHAMMER, E. L. L. (2010a). Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis., *Nucleic Acids Res* **38**(Database issue): D196–D203.

OSTLUND, G., SCHMITT, T., FORSLUND, K., KÖSTLER, T., MESSINA, D. N., ROOPRA, S., FRINGS, O. AND SONNHAMMER, E. L. L. (2010b). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, *Nucleic Acids Research* **38**(Database issue): D196–203. PMID: 19892828.

PAL, D. AND EISENBERG, D. (2005). Inference of protein function from protein structure, *Structure (London, England: 1993)* **13**(1): 121–130. PMID: 15642267.

PALANISWAMY, S. K., JAMES, S., SUN, H., LAMB, R. S., DAVULURI, R. V. AND GROTEWOLD, E. (2006). AGRIS and AtRegNet. a platform to link cis-Regulatory elements and transcription factors into regulatory networks, *Plant Physiology* **140**(3): 818–829.

PANDEY, S., NELSON, D. C. AND ASSMANN, S. M. (2009). Two novel GPCR-Type g proteins are abscisic acid receptors in arabidopsis, *Cell* **136**: 136–148.

PAREEK, A., SINGH, A., KUMAR, M., KUSHWAHA, H. R., LYNN, A. M. AND SINGLA-PAREEK, S. L. (2006). Whole-Genome analysis of oryza sativa reveals similar architecture of Two-Component signaling machinery with arabidopsis, *Plant Physiology* **142**(2): 380–397.

PARKINSON, J., ANTHONY, A., WASMUTH, J., SCHMID, R., HEDLEY, A. AND BLAXTER, M. (2004). PartiGene–constructing partial genomes, *Bioinformatics (Oxford, England)* **20**(9): 1398–1404. PMID: 14988115.

PENG, J. H. AND LAPITAN, N. L. V. (2005). Characterization of est-derived microsatellites in the wheat genome and development of essr markers., *Funct Integr Genomics* **5**(2): 80–96.

PHILLIPS, G. J. (2001). Green fluorescent protein–a bright idea for the study of bacterial protein localization, *FEMS Microbiology Letters* **204**(1): 9–18. PMID: 11682170.

PIQUES, M., SCHULZE, W. X., HÖHNE, M., USADEL, B., GIBON, Y., ROHWER, J. AND STITT, M. (2009). Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in arabidopsis., *Mol Syst Biol* **5**: 314.

POOLMAN, M. G., BONDE, B. K., GEVORGYAN, A., PATEL, H. H. AND FELL, D. A. (2006). Challenges to be faced in the reconstruction of metabolic networks from public databases, *Systems Biology, IEE Proceedings* **153**(5): 379–384.

POPESCU, S. C., POPESCU, G. V., BACHAN, S., ZHANG, Z., GERSTEIN, M., SNYDER, M. AND DINESH-KUMAR, S. P. (2009). Mapk target networks in arabidopsis thaliana revealed using functional protein microarrays., *Genes Dev* **23**(1): 80–92.

PRUD'HOMMEAUX, E. AND SEABORNE, A. (2008). SPARQL query language for RDF.
URL: `http://www.w3.org/TR/rdf-sparql-query/`

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. PMID: 15608248.

QI, L. L., ECHALIER, B., CHAO, S., LAZO, G. R., BUTLER, G. E., ANDERSON, O. D., AKHUNOV, E. D., DVORÁK, J., LINKIEWICZ, A. M., RATNASIRI, A., DUBCOVSKY, J., BERMUDEZ-KANDIANIS, C. E., GREENE, R. A., KANTETY, R., LA ROTA, C. M., MUNKVOLD, J. D., SORRELLS, S. F., SORRELLS, M. E., DILBIRLIGI, M., SIDHU, D., ERAYMAN, M., RANDHAWA, H. S., SANDHU, D., BONDAREVA, S. N., GILL, K. S.,

MAHMOUD, A. A., MA, X.-F., MIFTAHUDIN, GUSTAFSON, J. P., CONLEY, E. J., NDUATI, V., GONZALEZ-HERNANDEZ, J. L., ANDERSON, J. A., PENG, J. H., LAPITAN, N. L. V., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., PATHAN, M. S., ZHANG, D. S., NGUYEN, H. T., CHOI, D.-W., FENTON, R. D., CLOSE, T. J., MCGUIRE, P. E., QUALSET, C. O. AND GILL, B. S. (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat., *Genetics* **168**(2): 701–712.

RADRICH, K., TSURUOKA, Y., DOBSON, P., GEVORGYAN, A., SWAINSTON, N., BAART, G. AND SCHWARTZ, J. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks, *BMC Systems Biology* **4**(1): 114.

RAGHAVENDRA, A. S., GONUGUNTA, V. K., CHRISTMANN, A. AND GRILL, E. (2010). ABA perception and signalling, *Trends in Plant Science* **15**: 395–401.

RAZEM, F. A., EL-KEREAMY, A., ABRAMS, S. R. AND HILL, R. D. (2006). The RNA-binding protein FCA is an abscisic acid receptor, *Nature* **439**(7074): 290–294.

REMM, M., STORM, C. E. AND SONNHAMMER, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons., *J Mol Biol* **314**(5): 1041–1052.

RESNIK, P. (1999). Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence Research* **11**: 95–130.

RIAÑO-PACHÓN, D. M., RUZICIC, S., DREYER, I. AND MUELLER-ROEBER, B. (2007). PlnTFDB: an integrative plant transcription factor database, *BMC Bioinformatics* **8**: 42. PMID: 17286856.

RISK, J. M., DAY, C. L. AND MACKNIGHT, R. C. (2009). Reevaluation of abscisic Acid-Binding assays shows that G-Protein-Coupled receptor2 does not bind abscisic acid, *Plant Physiology* **150**(1): 6–11.

RISK, J. M., MACKNIGHT, R. C. AND DAY, C. L. (2008). FCA does not bind abscisic acid, *Nature* **456**(7223): E5–E6.

RIVALS, I., PERSONNAZ, L., TAING, L. AND POTIER, M. (2007). Enrichment or depletion of a GO category within a class of genes: which test?, *Bioinformatics* **23**(4): 401–407.

RIVERO, R. M., KOJIMA, M., GEPSTEIN, A., SAKAKIBARA, H., MITTLER, R., GEPSTEIN, S. AND BLUMWALD, E. (2007). Delayed leaf senescence induces extreme drought tolerance in a flowering plant, *Proceedings of the National Academy of Sciences* **104**(49): 19631–19636.

RUEPP, A., ZOLLNER, A., MAIER, D., ALBERMANN, K., HANI, J., MOKREJS, M., TETKO, I., GÜLDENER, U., MANNHAUPT, G., MÜNSTERKÖTTER, M. AND MEWES, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Research* **32**(18): 5539–5545. PMID: 15486203.

RUTHS, T., RUTHS, D. AND NAKHLEH, L. (2009). GS2: an efficiently computable measure of GO-based similarity of gene sets, *Bioinformatics (Oxford, England)* **25**(9): 1178–1184. PMID: 19289444.

RUTTENBERG, A., CLARK, T., BUG, W., SAMWALD, M., BODENREIDER, O., CHEN, H., DOHERTY, D., FORSBERG, K., GAO, Y., KASHYAP, V., KINOSHITA, J., LUCIANO, J., MARSHALL, M. S., OGBUJI, C., REES, J., STEPHENS, S., WONG, G., WU, E., ZACCAGNINI, D., HONGSERMEIER, T., NEUMANN, E., HERMAN, I. AND CHEUNG, K. (2007). Advancing translational research with the semantic web, *BMC Bioinformatics* **8**(Suppl 3): S2.

SABOT, F., GUYOT, R., WICKER, T., CHANTRET, N., LAUBIN, B., CHALHOUB, B., LEROY, P., SOURDILLE, P. AND BERNARD, M. (2005). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations, *Molecular Genetics and Genomics* **274**: 119–130. 10.1007/s00438-005-0012-9.

SADIQOV, S. T., AKBULUT, M. AND EHMEDOV, V. (2002). Role of ca2+ in drought stress signaling in wheat seedlings, *Biochemistry* **67**(4): 491–497. PMID: 11996664.

SAITO, S., HIRAI, N., MATSUMOTO, C., OHIGASHI, H., OHTA, D., SAKATA, K. AND MIZUTANI, M. (2004). Arabidopsis CYP707As encode (+)-Abscisic acid 8′-Hydroxylase, a key enzyme in the oxidative catabolism of abscisic acid, *Plant Physiology* **134**(4): 1439–1449.

SAITOU, N. AND NEI, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution* **4**(4): 406–425. PMID: 3447015.

SANTIAGO, J., DUPEUX, F., ROUND, A., ANTONI, R., PARK, S., JAMIN, M., CUTLER, S. R., RODRIGUEZ, P. L. AND MARQUEZ, J. A. (2009). The abscisic acid receptor PYR1 in complex with abscisic acid, *Nature* **462**(7273): 665–668.

SANTIAGO, J., RODRIGUES, A., SAEZ, A., RUBIO, S., ANTONI, R., DUPEUX, F., PARK, S., MÁRQUEZ, J. A., CUTLER, S. R. AND RODRIGUEZ, P. L. (2009). Modulation of drought resistance by the abscisic acid receptor PYL5 through inhibition of clade a PP2Cs, *The Plant Journal: For Cell and Molecular Biology* **60**(4): 575–588. PMID: 19624469.

SATO, A., SATO, Y., FUKAO, Y., FUJIWARA, M., UMEZAWA, T., SHINOZAKI, K., HIBI, T., TANIGUCHI, M., MIYAKE, H., GOTO, D. B. AND UOZUMI, N. (2009). Threonine at position 306 of the KAT1 potassium channel is essential for channel activity and is a target site for ABA-activated SnRK2/OST1/SnRK2.6 protein kinase, *The Biochemical Journal* **424**(3): 439–448. PMID: 19785574.

SAUTER, A., DAVIES, W. AND HARTUNG, W. (2001). The long-distance abscisic acid signal in the droughted plant: the fate of the hormone on its way from root to shoot, *Journal of Experimental Botany* **52**(363): 1991–1997.

SAYERS, E. W., BARRETT, T., BENSON, D. A., BOLTON, E., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., FEDERHEN, S., FEOLO, M., FINGERMAN, I. M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., LANDSMAN, D., LIPMAN, D. J., LU, Z., MADDEN, T. L., MADEJ, T., MAGLOTT, D. R., MARCHLER-BAUER, A., MILLER, V., MIZRACHI, I., OSTELL, J., PANCHENKO, A., PHAN, L., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SHUMWAY, M.,

SIROTKIN, K., SLOTTA, D., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., WANG, Y., WILBUR, W. J., YASCHENKO, E. AND YE, J. (2011). Database resources of the national center for biotechnology information, *Nucleic Acids Research* **39**(suppl 1): D38–D51.

SAYERS, E. W., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., FEOLO, M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., LANDSMAN, D., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., MILLER, V., MIZRACHI, I., OSTELL, J., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SHUMWAY, M., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., YASCHENKO, E. AND YE, J. (2009). Database resources of the national center for biotechnology information, *Nucleic Acids Research* **37**(Database issue): D5–D15. PMID: 18940862 PMCID: 2686545.

SBGN (2011). Sbgn process description language: Level 1 examples.
URL: `http://www.sbgn.org/Documents/PD_L1_Examples`

SCHMID, R. AND BLAXTER, M. L. (2008). annot8r: GO, EC and KEGG annotation of EST datasets, *BMC Bioinformatics* **9**: 180–180. PMID: 18400082 PMCID: 2324097.

SEKIMOTO, H., SEO, M., KAWAKAMI, N., KOMANO, T., DESLOIRE, S., LIOTENBERG, S., MARION-POLL, A., CABOCHE, M., KAMIYA, Y. AND KOSHIBA, T. (1998). Molecular cloning and characterization of aldehyde oxidases in arabidopsis thaliana, *Plant and Cell Physiology* **39**(4): 433–442.

SEO, M., KOIWAI, H., AKABA, S., KOMANO, T., ORITANI, T., KAMIYA, Y. AND KOSHIBA, T. (2000). Abscisic aldehyde oxidase in leaves of arabidopsis thaliana, *The Plant Journal: For Cell and Molecular Biology* **23**(4): 481–488. PMID: 10972874.

SHAFER, P., ISGANITIS, T. AND YONA, G. (2006). Hubs of knowledge: using the functional link structure in biozon to mine for biologically significant entities, *BMC Bioinformatics* **7**: 71–71. PMID: 16480496 PMCID: 1421446.

SHAMEER, K., AMBIKA, S., VARGHESE, S. M., KARABA, N., UDAYAKUMAR, M. AND SOWDHAMINI, R. (2009). STIFDB-Arabidopsis stress responsive transcription factor DataBase, *International Journal of Plant Genomics* **2009**: 1–8.

SHAN, C. AND LIANG, Z. (2010). Jasmonic acid regulates ascorbate and glutathione metabolism in agropyron cristatum leaves under water stress, *Plant Science* **178**(2): 130–139.

SHAN, S., P., H. AND SHEEN, J. (2007). Endless hide-and-seek: Dynamic co-evolution in plant-bacterium warfare, *Journal of Integrative Plant Biology* **49**: 105–111.

SHANG, Y., YAN, L., LIU, Z., CAO, Z., MEI, C., XIN, Q., WU, F., WANG, X., DU, S., JIANG, T., ZHANG, X., ZHAO, R., SUN, H., LIU, R., YU, Y. AND ZHANG, D. (2010). The Mg-Chelatase h subunit of arabidopsis antagonizes a group of WRKY transcription repressors to relieve ABA-Responsive genes of inhibition, *The Plant Cell Online* **22**(6): 1909–1935.

SHANNON, C. E. (2001). A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**: 3.

SHAPIRO, S. S. AND WILK, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**(3-4): 591–611.

SHEN, Y., WANG, X., WU, F., DU, S., CAO, Z., SHANG, Y., WANG, X., PENG, C., YU, X., ZHU, S., FAN, R., XU, Y. AND ZHANG, D. (2006). The mg-chelatase h subunit is an abscisic acid receptor, *Nature* **443**(7113): 823–826.

SHETH, A. P. AND LARSON, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases, *ACM Comput. Surv.* **22**: 183–236.

SHIGETA, R., CLINE, M., LIU, G. AND SIANI-ROSE, M. A. (2003). GPCR-GRAPA-LIB–a refined library of hidden markov models for annotating GPCRs, *Bioinformatics* **19**(5): 667–668.

SHINOZAKI, K. AND YAMAGUCHI-SHINOZAKI, K. (2007). Gene networks involved in drought stress response and tolerance, *Journal of Experimental Botany* **58**(2): 221–227.

SIEGEL, R. S., XUE, S., MURATA, Y., YANG, Y., NISHIMURA, N., WANG, A. AND SCHROEDER, J. I. (2009). Calcium elevation-dependent and attenuated resting calcium-dependent abscisic acid induction of stomatal closure and abscisic acid-induced enhancement of calcium sensitivities of s-type anion and inward-rectifying k channels in arabidopsis guard cells, *The Plant Journal: For Cell and Molecular Biology* **59**(2): 207–220. PMID: 19302418.

SIRAVA, M., SCHAFER, T., EIGLSPERGER, M., KAUFMANN, M., KOHLBACHER, O., BORNBERG-BAUER, E. AND LENHOF, H. P. (2002). BioMiner–modeling, analyzing, and visualizing biochemical pathways and networks, *Bioinformatics* **18**: S219–S230.

SLEATOR, R. D. (2011). Phylogenetics., *Arch Microbiol* **193**(4): 235–239.

SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., GOLDBERG, L. J., EILBECK, K., IRELAND, A., MUNGALL, C. J., LEONTIS, N., ROCCA-SERRA, P., RUTTENBERG, A., SANSONE, S., SCHEUERMANN, R. H., SHAH, N., WHETZEL, P. L. AND LEWIS, S. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology* **25**: 1251–1255.

SMITH, M. K., WELTY, C. AND MCGUINNESS, D. L. (2004). OWL web ontology language, 2004.
URL: `http://www.w3.org/TR/owl-guide/`

STARK, C. (2006). BioGRID: a general repository for interaction datasets, *Nucleic Acids Research* **34**: D535–D539.

STEGGLES, L. J., BANKS, R., SHAW, O. AND WIPAT, A. (2007). Qualitatively modelling and analysing genetic regulatory networks: a petri net approach, *Bioinformatics* **23**(3): 336 –343.

STEIN, L. D. (2003). Integrating biological databases, *Nature Reviews. Genetics* **4**(5): 337–345. PMID: 12728276.

STEVENS, R., ZHAO, J. AND GOBLE, C. (2007). Using provenance to manage knowledge of in silico experiments, *Briefings in Bioinformatics* **8**(3): 183–194.

STUDIER, J. A. AND KEPPLER, K. J. (1988). A note on the neighbor-joining algorithm of saitou and nei., *Mol Biol Evol* **5**(6): 729–731.

SU, J., TEICHMANN, S. A. AND DOWN, T. A. (2010). Assessing computational methods of Cis-Regulatory module prediction, *PLoS Comput Biol* **6**(12): e1001020.

SUN, Y., HE, Z., MA, W. AND XIA, X. (2011). Alternative splicing in the coding region of Ppo-A1 directly influences the polyphenol oxidase activity in common wheat (Triticum aestivum l.), *Functional & Integrative Genomics* **11**(1): 85–93. PMID: 21046181.

SUZUKI, T., MIWA, K., ISHIKAWA, K., YAMADA, H., AIBA, H. AND MIZUNO, T. (2001). The arabidopsis sensor his-kinase, AHK4, can respond to cytokinins, *Plant and Cell Physiology* **42**(2): 107–113.

TAIR (2007). Arabidopsis nomenclature.
URL: `http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp`

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. AND NATALE, D. A. (2003). The COG database: an updated version includes eukaryotes, *BMC Bioinformatics* **4**: 41. PMID: 12969510.

TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A. AND KOONIN, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Research* **28**(1): 33–36. PMID: 10592175.

TAUBERT, J., HINDLE, M., LYSENKO, A., WEILE, J., KÃ¶HLER, J. AND RAWLINGS, C. J. (2009). Linking life sciences data using Graph-Based mapping, *in* N. W. Paton, P. Missier and C. Hedeler (eds), *Data Integration in the Life Sciences*, Vol. 5647, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 16–30.

TAUBERT, J., SIEREN, K. P., HINDLE, M., HOEKMAN, B., WINNENBURG, R., PHILIPPI, S., RAWLINGS, C. J. AND KÖHLER, J. (2007). The oxl format for the exchange of integrated datasets., *Journal Integrative Bioinformatics* **4**.

THE GENE ONTOLOGY CONSORTIUM (2011a). GO annotation file GAF 2.0 format guide.
URL: `http://www.geneontology.org/GO.format.gaf-2_0.shtml`

THE GENE ONTOLOGY CONSORTIUM (2011b). GO database guide.
URL: `http://www.geneontology.org/GO.database.shtml`

THE GENE ONTOLOGY CONSORTIUM (2011c). Mapping of gene ontology terms to enzyme commission entries.
URL: `http://www.geneontology.org/external2go/ec2go`

THE PLANT ONTOLOGY CONSORTIUM (2002). The plant ontology consortium and plant ontologies, *Comparative and Functional Genomics* **3**(2): 137–142. PMID: 18628842 PMCID: 2447263.

THE UNIPROT CONSORTIUM (2010). The universal protein resource (UniProt) in 2010, *Nucleic Acids Research* **38**(Database issue): D142–148. PMID: 19843607.

THIBAUD-NISSEN, F., WU, H., RICHMOND, T., REDMAN, J. C., JOHNSON, C., GREEN, R., ARIAS, J. AND TOWN, C. D. (2006). Development of arabidopsis whole-genome microarrays and their application to the discovery of binding sites for the tga2 transcription factor in salicylic acid-treated plants., *Plant J* **47**(1): 152–162.

THIMM, O., BLÄSING, O., GIBON, Y., NAGEL, A., MEYER, S., KRÜGER, P., SELBIG, J., MÜLLER, L. A., RHEE, S. Y. AND STITT, M. (2004). MAPMAN: a user- driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, *The Plant Journal* **37**(6): 914–939.

THOMPSON, A. J., JACKSON, A. C., SYMONDS, R. C., MULHOLLAND, B. J., DADSWELL, A. R., BLAKE, P. S., BURBIDGE, A. AND TAYLOR, I. B. (2000). Ectopic expression of a tomato 9-cis-epoxycarotenoid dioxygenase gene causes over-

production of abscisic acid, *The Plant Journal: For Cell and Molecular Biology* **23**(3): 363–374. PMID: 10929129.

TIMELOGIC (2011). Decypher fpga biocomputing systems.
URL: `http://www.timelogic.com/`

TRIPLET, T., SHORTRIDGE, M. D., GRIEP, M. A., STARK, J. L., POWERS, R. AND REVESZ, P. (2010). PROFESS: a PROtein function, evolution, structure and sequence database, *Database: The Journal of Biological Databases and Curation* **2010**.

TSESMETZIS, N., COUCHMAN, M., HIGGINS, J., SMITH, A., DOONAN, J. H., SEIFERT, G. J., SCHMIDT, E. E., VASTRIK, I., BIRNEY, E., WU, G., D'EUSTACHIO, P., STEIN, L. D., MORRIS, R. J., BEVAN, M. W. AND WALSH, S. V. (2008). Arabidopsis reactome: A foundation knowledgebase for plant systems biology, *The Plant Cell Online* **20**(6): 1426–1436.

URAO, T., MIYATA, S., YAMAGUCHI-SHINOZAKI, K. AND SHINOZAKI, K. (2000). Possible his to asp phosphorelay signaling in an arabidopsis two-component system, *FEBS Letters* **478**(3): 227–232.

URAO, T., YAKUBOV, B., SATOH, R., YAMAGUCHI-SHINOZAKI, K., SEKI, M., HIRAYAMA, T. AND SHINOZAKI, K. (1999). A transmembrane Hybrid-Type histidine kinase in arabidopsis functions as an osmosensor, *The Plant Cell Online* **11**(9): 1743–1754.

VAHISALU, T., KOLLIST, H., WANG, Y., NISHIMURA, N., CHAN, W., VALERIO, G., LAMMINMAKI, A., BROSCHE, M., MOLDAU, H., DESIKAN, R., SCHROEDER, J. I. AND KANGASJARVI, J. (2008). SLAC1 is required for plant guard cell s-type anion channel function in stomatal signalling, *Nature* **452**(7186): 487–491.

VERSPOOR, K., COHN, J., MNISZEWSKI, S. AND JOSLYN, C. (2006). A categorization approach to automated ontological function annotation, *Protein Science* **15**: 1544–1549.

VON KOSKULL-DÖRING, P., SCHARF, K. AND NOVER, L. (2007). The diversity of plant heat stress transcription factors, *Trends in Plant Science* **12**(10): 452–457.

W3C (2011). Extensible markup language (XML).
URL: http://www.w3.org/XML/

WALNES, J. AND SCHAIBLE, J. (2011). XStream.
URL: http://xstream.codehaus.org/

WAN, Y., POOLE, R. L., HUTTLY, A. K., TOSCANO-UNDERWOOD, C., FEENEY, K., WELHAM, S., GOODING, M. J., MILLS, C., EDWARDS, K. J., SHEWRY, P. R. AND MITCHELL, R. A. (2008). Transcriptome analysis of grain development in hexaploid wheat., *BMC Genomics* **9**: 121.

WAN, Y., UNDERWOOD, C., TOOLE, G., SKEGGS, P., ZHU, T., LEVERINGTON, M., GRIFFITHS, S., WHEELER, T., GOODING, M., POOLE, R., EDWARDS, K. J., GEZAN, S., WELHAM, S., SNAPE, J., MILLS, E. N. C., MITCHELL, R. A. C. AND SHEWRY, P. R. (2009). A novel transcriptomic approach to identify candidate genes for grain quality traits in wheat, *Plant Biotechnology Journal* **7**(5): 401–410. PMID: 19490503.

WASMUTH, J. AND BLAXTER, M. (2004). prot4EST: translating expressed sequence tags from neglected genomes, *BMC Bioinformatics* **5**(1): 187.

WEAVER, L. M., GAN, S., QUIRINO, B. AND AMASINO, R. M. (1998). A comparison of the expression patterns of several senescence-associated genes in response to stress and hormone treatment, *Plant Molecular Biology* **37**(3): 455–469. PMID: 9617813.

WILKINSON, S. AND DAVIES, W. J. (2002). ABA-based chemical signalling: the co-ordination of responses to stress in plants, *Plant, Cell & Environment* **25**(2): 195–210. PMID: 11841663.

WINGENDER, E. (2004). TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks, *In Silico Biology* **4**(1): 55–61. PMID: 15089753.

WRIGHT, P. E. AND DYSON, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm., *J Mol Biol* **293**(2): 321–331.

YANG, J. AND GUO, Z. (2007). Cloning of a 9-cis-epoxycarotenoid dioxygenase gene (SgNCED1) from stylosanthes guianensis and its expression in response to abiotic stresses, *Plant Cell Reports* **26**(8): 1383–1390. PMID: 17333017.

YANHUI, C., XIAOYUAN, Y., KUN, H., MEIHUA, L., JIGANG, L., ZHAOFENG, G., ZHIQIANG, L., YUNFEI, Z., XIAOXIAO, W., XIAOMING, Q., YUNPING, S., LI, Z., XIAOHUI, D., JINGCHU, L., XING-WANG, D., ZHANGLIANG, C., HONGYA, G. AND LI-JIA, Q. (2006). The MYB transcription factor superfamily of arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family, *Plant Molecular Biology* **60**(1): 107–124. PMID: 16463103.

YEH, A., MORGAN, A., COLOSIMO, M. AND HIRSCHMAN, L. (2005). Biocreative task 1a: gene mention finding evaluation., *BMC Bioinformatics* **6 Suppl 1**: S2.

YESBERGENOVA, Z., YANG, G., ORON, E., SOFFER, D., FLUHR, R. AND SAGI, M. (2005). The plant mo-hydroxylases aldehyde oxidase and xanthine dehydrogenase have distinct reactive oxygen species signatures and are induced by drought and abscisic acid, *The Plant Journal: For Cell and Molecular Biology* **42**(6): 862–876. PMID: 15941399.

YILMAZ, A., NISHIYAMA, M. Y., FUENTES, B. G., SOUZA, G. M., JANIES, D., GRAY, J. AND GROTEWOLD, E. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses, *Plant Physiology* **149**(1): 171–180.

YIN, P., FAN, H., HAO, Q., YUAN, X., WU, D., PANG, Y., YAN, C., LI, W., WANG, J. AND YAN, N. (2009). Structural insights into the mechanism of abscisic acid signaling by PYL proteins, *Nat Struct Mol Biol* **16**(12): 1230–1236.

YOSHIDA, R., HOBO, T., ICHIMURA, K., MIZOGUCHI, T., TAKAHASHI, F., ARONSO, J., ECKER, J. R. AND SHINOZAKI, K. (2002). ABA-Activated SnRK2 protein kinase is required for dehydration stress signaling in arabidopsis, *Plant and Cell Physiology* **43**(12): 1473–1483.

YOSHIDA, T., FUJITA, Y., SAYAMA, H., KIDOKORO, S., MARUYAMA, K., MIZOI, J., SHINOZAKI, K. AND YAMAGUCHI-SHINOZAKI, K. (2010). AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation, *The Plant Journal* **61**(4): 672–685.

YUAN, Q., OUYANG, S., LIU, J., SUH, B., CHEUNG, F., SULTANA, R., LEE, D., QUACKENBUSH, J. AND BUELL, C. R. (2003). The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists, *Nucleic Acids Research* **31**(1): 229–233. PMID: 12519988.

ZHANG, A., JIANG, M., ZHANG, J., TAN, M. AND HU, X. (2006). Mitogen-Activated protein kinase is involved in abscisic Acid-Induced antioxidant defense and acts downstream of reactive oxygen species production in leaves of maize plants, *Plant Physiology* **141**(2): 475–487.

ZHANG, H. J. AND DAVIES, W. J. (1987). Increased synthesis of ABA in partially dehydrated root tips and ABA transport from roots to leaves, *Journal of Experimental Botany* **38**(12): 2015–2023.

ZHENG, B. AND LU, X. (2007). Novel metrics for evaluating the functional coherence of protein groups via protein semantic network, *Genome Biology* **8**(7): R153. PMID: 17672896.

ZHENG, Q. AND WANG, X. (2008). GOEAST: a web-based software toolkit for gene ontology enrichment analysis, *Nucleic Acids Research* **36**(Web Server issue): W358–363. PMID: 18487275.

ZHOU, J., WANG, X., JIAO, Y., QIN, Y., LIU, X., HE, K., CHEN, C., MA, L., WANG, J., XIONG, L., ZHANG, Q., FAN, L. AND DENG, X. W. (2007). Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle, *Plant Molecular Biology* **63**(5): 591–608. PMID: 17225073 PMCID: 1805039.

ZHU, J. (2002). Salt and drought stress signal transduction in plants, *Annual Review of Plant Biology* **53**: 247–273. PMID: 12221975.

ZIEGLER, P. AND DITTRICH, K. R. (2004). Three decades of data integration - all problems solved?, *In 18th IFIP World Computer Congress (WCC 2004)* **12**: 3–12.