

Development and application of analysis modules in MADIBA, a Web-based toolkit for the interpretation of microarray data

by

Philip John Law

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae Bioinformatics

in the Faculty of Natural and Agricultural Science

University of Pretoria

Pretoria

July 2008

Acknowledgements

I would like to thank the National Bioinformatics Network for funding this project, and providing me the opportunity to visit the Scottish Crop Research Institute (SCRI).

From the SCRI, I wish to thank Dr Leighton Pritchard for his advice on bioinformatics and computational topics and Prof Paul Birch and Dr Ian Toth for valuable discussions on *Pectobacterium*.

I would also like to thank Jeremy Baxter (Rhodes University) for advice on statistical matters, and Dr Sanushka Naidoo and Dr Bridget Campbell from the Molecular Plant-Pathogen Interactions group for the *Arabidopsis* and millet data respectively.

Special thanks to Natasha Wood for all the support and encouragement.

Finally, I would like to thank Prof Braam Louw and Dr Clotilde Claudel-Renard for initiating the MADIBA project, and my supervisors Prof Dave Berger and Prof Fourie Joubert for their assistance and guidance.



Table of Contents

ACKNOWLEDGEMENTS	I
TABLE OF CONTENTS	II
ABBREVIATIONS	V
LIST OF TABLES	VII
LIST OF FIGURES	VIII
PREFACE	X
1 CHAPTER 1 – INTRODUCTION/BACKGROUND	1
1.1 Introduction to microarrays	1
1.1.1 Basics about microarrays	1
1.1.2 Clustering	2
1.1.3 Post-processing of microarray data	4
1.2 Tools for the post-processing of microarray data	5
1.2.1 FatiGO	7
1.2.2 Genevestigator	10
1.2.3 GeneXPress	13
1.2.4 Genome2D	15
1.2.5 GoMiner	17
1.2.6 MAPPFinder	19
1.2.7 WebGestalt	22
1.2.8 MADIBA	24
1.3 Conclusion	26
2 CHAPTER 2 – MADIBA IMPLEMENTATION	29
2.1 Introduction	29
2.2 Data sources	29
2.3 Database implementation	30
2.4 User interface	31
2.5 Data submission	32
2.6 MADIBA modules to analyse gene clusters	33



2.6.1	Front page	35
2.6.2	Gene Ontology module	36
2.6.3	Metabolic Pathways module	39
2.6.4	Chromosomal Localisation module	41
2.6.5	Transcription Regulation module	43
2.6.6	Organism Specific Characteristics module	45
2.6.7	Output	51
2.6.8	Contact form	51
2.7	Data maintenance	51
2.8	Documentation	52
2.9	Conclusion	52
3	CHAPTER 3 – APPLICATION TO BIOLOGICAL DATA	54
3.1	Introduction	54
3.2	Application to <i>Plasmodium falciparum</i>	54
3.2.1	<i>Plasmodium falciparum</i> introduction	54
3.2.2	MADIBA <i>Plasmodium falciparum</i> data analysis	56
3.2.3	Conclusion	59
3.3	Application to <i>Arabidopsis thaliana</i>	59
3.3.1	<i>Arabidopsis thaliana</i> as a model species	59
3.3.2	Plant defence responses	61
3.3.3	MADIBA <i>Arabidopsis thaliana</i> data analysis	70
3.3.4	MADIBA defence pathways analysis	73
3.3.5	PCA on <i>Ralstonia solanacearum</i> data	80
3.3.6	Conclusion	81
3.4	Application to rice	82
3.4.1	Importance of rice	82
3.4.2	Rice as a model species	83
3.4.3	MADIBA rice data analysis	84
3.4.4	MADIBA pearl millet data analysis	88
3.4.5	Conclusion	93
3.5	Application to <i>Pectobacterium atrosepticum</i>	94
3.5.1	<i>Pectobacterium atrosepticum</i> introduction	94
3.5.2	Quorum sensing	95
3.5.3	MADIBA <i>Pectobacterium atrosepticum</i> data analysis	98



3.5.4	Conclusion	105
3.6	Concluding remarks	105
	CHAPTER 4 – CONCLUDING DISCUSSION	107
	SUMMARY	112
	REFERENCES	114

Abbreviations

BLAST	:	Basic Local Alignment Search Tool
BTH	:	Benzothiadiazole
CGI	:	Common Gateway Interface
DAG	:	Directed Acyclic Graph
DRASTIC	:	Database Resource for the Analysis of Signal Transduction in Cells
ET	:	Ethylene
FDR	:	False Discovery Rate
GO	:	Gene Ontology
HSL	:	Homoserine Lactone
JA	:	Jasmonic Acid
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
MADIBA	:	MicroArray Data Interface for Biological Annotation
MeJ	:	Methyl Jasmonate
MPPI	:	Molecular Plant-Pathogen Interactions
<i>N</i> -AHL	:	<i>N</i> -acyl homoserine lactone
NASCArrays	:	Nottingham Arabidopsis Stock Centre Arrays
<i>Pba</i>	:	<i>Pectobacterium atrosepticum</i>
PC	:	Principal Component
PCA	:	Principal Component Analysis
PHP	:	PHP: Hypertext Preprocessor
QS	:	Quorum Sensing
RSAT	:	Regulatory Sequence Analysis Tools
SA	:	Salicylic Acid
SQL	:	Structured Query Language
SREPP	:	Soft-Rotting Enterobacterial Plant Pathogen

STEM	:	Short Time-series Expression Miner
TAIR	:	The Arabidopsis Information Resource
TF	:	Transcription Factor
TFBS	:	Transcription Factor Binding Site
TIGR	:	The Institute for Genomic Research

List of Tables

Table 1.1: Table of the basic specifications of various tools for post-processing microarray data	27
Table 1.2: General features of the discussed tools	28
Table 2.1: A summary table of the various tools that were used MADIBA.....	53
Table 3.1: A portion of the table of <i>p</i> -values from a GO analysis	58
Table 3.2: The five closest results from the PCA Experiment Comparer to the data from the susceptible <i>R. solanacearum</i> interaction	83
Table 3.3: Table of the salicylic acid responsive millet genes that were used in the MADIBA analysis	88
Table 3.4: Table showing the best hits salicylic acid responsive millet genes from the rice database in MADIBA	89
Table 3.5: Set of millet cDNA fragments that were identified as being responsive to methyl jasmonate only	92

List of Figures

Figure 1.1: A generic microarray experiment.....	2
Figure 1.2: Example of the output from a clustering algorithm.....	3
Figure 1.3: Front page of the FatiGO tool	9
Figure 1.4: The Meta-Analyser tool of Genevestigator.....	12
Figure 1.5: View of the tree browser view and the birdseye view from GeneXPress	15
Figure 1.6: View of the Genemap of Genome2D	17
Figure 1.7: Screenshot of GoMiner’s GUI	19
Figure 1.8: The MAPPFinder Browser.....	22
Figure 1.9: User interface of WebGestalt	24
Figure 1.10: Front page of MADIBA.....	26
Figure 2.1: UML class diagram showing the four generic tables of the MADIBA databases .	30
Figure 2.2: A schematic representation of the flow of data through MADIBA	34
Figure 2.3: Screenshot of the page after submission.....	35
Figure 2.4: Pie charts of MADIBA, showing annotation commonalities	36
Figure 2.5: Screenshots of the Gene Ontology module.....	37
Figure 2.6: Screenshots of the Metabolic Pathways module	40
Figure 2.7: Screenshot of the Chromosomal Localisation module	42
Figure 2.8: Screenshot of MADIBA’s version of gbrowse	43
Figure 2.9: Example outputs from the Transcription Regulation module	44
Figure 2.10: Examples from the Organism Specific module.....	46
Figure 2.11: Schematic dataflow illustrating the PCA Experiment Comparer.....	50
Figure 3.1: Metabolic pathways analysis of cluster 6 of the <i>Plasmodium falciparum</i> data	57
Figure 3.2: An analysis of the biological process ontology of the cluster 6 of the <i>P. falciparum</i> data	58
Figure 3.3: Brief summary of plant defence signalling pathways in <i>Arabidopsis thaliana</i>	65

Figure 3.4: Screenshot of a portion of the results of the regulation of the genes from the DRASTIC database	75
Figure 3.5: Clusters proposed by Glazebrook <i>et al.</i>	77
Figure 3.6: Plots of the top two principal components, using the data from the Glazebrook <i>et al.</i> mutant study	79
Figure 3.7: Box plot of the Q^2 values for the first three principal components after a PCA using the susceptible <i>R. solanacearum</i> interaction data.....	82
Figure 3.8: An alignment of the genomes of six major grass crop species	84
Figure 3.9: Pie chart from MADIBA, showing common annotations in the cluster.....	85
Figure 3.10: gbrowse view of a portion of chromosome 10 of the genes from the rice BTH treatment	87
Figure 3.11: Screenshot of the model profiles overview interface of STEM using the <i>Pectobacterium atrosepticum</i> data	100
Figure 3.12: Output of the Chromosomal Localisation for <i>Pba</i>	102
Figure 3.13: STEM profile 10 of the <i>Pba</i> data	103
Figure 3.14: Portion of the results as obtained by the dyad-analysis program of RSAT	104

Preface

Microarray and suppression subtractive hybridization (SSH) technologies have made it possible to observe changes in the gene expression of various organisms, under a range of conditions, and at different time points. While these experiments can improve our understanding of the biology of an organism, data mining is essential for inferring significant biological information, such as the identification of new biological mechanisms. Many algorithms and software have been developed for analysing gene expression, although the extraction of relevant information from experimental data is still a substantial challenge.

MADIBA (MicroArray Data Interface for Biological Annotation) is an integrated, online tool that facilitates the assignment of biological meaning to gene expression clusters, thus assisting researchers in interpreting their results and understanding the meaning of the co-expression of a cluster of genes. MADIBA automates this post processing stage by performing a number of diverse analyses and generating graphical representations for easy interpretation. While other tools exist, they are designed for specific model organisms such as human, mouse and rat. MADIBA is instead aimed at the socially and economically important *Plasmodium* species, with a specific aim being to identify new drug targets; a bacterial plant pathogen that causes major economic losses of potato; as well as plant species with the aim to improve crops of relevance in Africa, such as pearl millet. Specifically, MADIBA at present contains the genomic data for *Plasmodium falciparum*, *Pectobacterium atrosepticum* (*Pba*), *Arabidopsis thaliana* and rice (*Oryza sativa* ssp *japonica* cv Nipponbare).

Tools within the MADIBA web interface allows rapid analyses for the identification of over-represented Gene Ontology terms; visualising of implicated genes on KEGG metabolic pathways; their chromosomal localisations; putative common transcriptional regulatory elements in the upstream sequences; and an analysis specific to the organism being analysed, for example to identify potential drug targets in Plasmodium or for gaining insights into improving crops. Specifically in the Arabidopsis Characteristics module, genes that could ultimately be used in improving plant defences were identified by determining how they respond to different treatments. The genes' response (either up- or down-regulated) to the different treatments was obtained from the information in the DRASTIC database. A more complex approach was developed and named PCA Experiment Comparer, which compared the gene expression levels of the experiments in NASCArrays with a submitted set of genes.

MADIBA was initially designed for use on *P. falciparum* data, and written by Dr. Clotilde Claudel-Renard in Perl and PHP. During this MSc study, much of the Perl code was converted to Python, in particular the Metabolic Pathways module. In addition, the Gene Ontology module was rewritten using a different approach, the output module added, and *A. thaliana*, rice, and *Pba* genome data were implemented into MADIBA. A paper on MADIBA was published in 2008 in BMC Genomics (volume 9, page 105), titled “MADIBA: A web server toolkit for biological interpretation of Plasmodium and plant gene clusters”. In addition to this information, this MSc dissertation includes information on *Pba*, as well as new analyses in the Arabidopsis and rice sections, specifically focussing on plant defences.

The primary aim of MADIBA is to enable biologists to analyse their microarray data in an integrated fashion. All a user needs to do is submit a set of co-expressed genes, and the relevant data will be retrieved and a series of diverse analyses performed on it. In this way, researchers do not need to be concerned with data consistency and different formats for different analyses. MADIBA is freely available and can be accessed on the web using a JavaScript enabled browser at <http://www.bi.up.ac.za/MADIBA/>.

Chapter 1 of this dissertation is a literature survey dealing with downstream microarray analyses and a discussion of a selection of programs that are currently available; *Chapter 2* describes with the implementation and design of the MADIBA system; *Chapter 3* discusses the biological application of data from the various organisms that have been implemented in MADIBA, described below in further detail; *Chapter 4* provides some general concluding remarks; and is followed by a summary.

Specifically in this dissertation, MADIBA was used to study two biological systems under investigation by the Molecular Plant-Pathogen Interactions group at the University of Pretoria, namely the identification of signalling pathways in *Arabidopsis thaliana* in both the resistant and susceptible interactions when infected with *Ralstonia solanacearum*; and a comparison of salicylic acid and methyl jasmonate treatments in pearl millet prior to infection with the rust fungus. In addition, case studies using *Plasmodium falciparum* data were used to demonstrate MADIBA’s functionality, and for *Pba*, data from an *expI* mutant experiment was used to identify genes involved in quorum sensing. Preceding each case study is a literature survey of the biological system under study.

A note on references: Numbers that appear in round brackets, e.g. (1), indicate internet references, usually the location of a downloadable resource that was used. These references are listed before the journal citations in the reference list.

Chapter 1 – Introduction/Background

1.1 Introduction to microarrays

1.1.1 Basics about microarrays

Microarrays are typically glass slides which contain DNA molecules attached at fixed locations called spots or features. There may be thousands of spots on an array, and each spot may contain millions of copies of genes or fragments of genes (Causton *et al.*, 2003).

To study gene expression levels using microarrays, the most popular method is to compare the gene expression levels of two different samples which come from different conditions, for example, cell types at different cell cycle stages, with each sample labelled with a different coloured dye (Stekel, 2003; Causton *et al.*, 2003). These so-called two-colour microarrays generally use the fluorescent cyanine (Cy) dyes, namely Cy3 and Cy5. Once the mRNA from each sample has been extracted from the tissue of interest, it is converted to cDNA (using reverse transcriptase) and labelled with the respective fluorescent Cy dye (Causton *et al.*, 2003; Stekel, 2003). These labelled samples are then hybridised to the probes on the glass slide to form heteroduplexes *via* Watson-Crick base pairing (Stekel, 2003). The hybridised microarray is scanned by a laser at wavelengths to excite the dyes, although typically when using two colour arrays, two lasers are used – one for each dye. The output of the scanner is two monochrome images, one from each of the lasers. When these images are combined, it is possible to create the usual red-green false colour images (Stekel, 2003), where the Cy3 dye is represented a green colour and Cy5 is represented as red. The amount of fluorescence emitted by the spot when excited by the laser corresponds to the amount of nucleic acid bound to each spot. If the sample labelled with Cy3 is abundant, the spot will be green, and red if the sample labelled with Cy5 is abundant (Causton *et al.*, 2003). If both samples occur with equal abundance, the spot will appear yellow and if neither are present, it will not fluoresce and will appear black (Causton *et al.*, 2003). Figure 1.1 illustrates the process of a typical microarray experiment.

Microarrays are an extremely powerful tool for monitoring gene expression levels for thousands of genes. Since there may be many thousands of different DNA molecules bonded to an array, it is possible to measure the expression of those thousands of genes simultaneously (Stekel, 2003). The identification of patterns in the gene expression can

potentially be used to rationalise a wide range of diverse phenomena, from explaining disease states to responses to stimuli (Causton *et al.*, 2003).

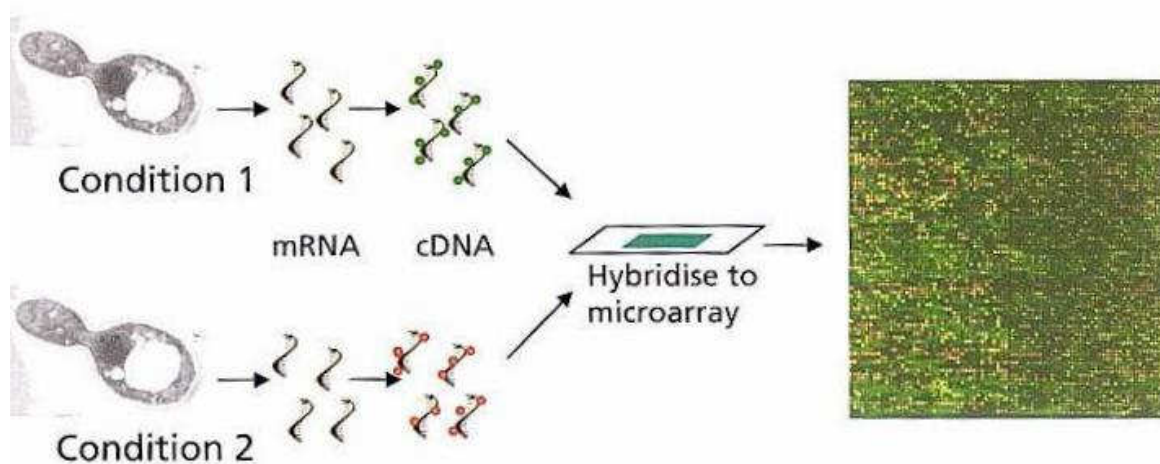


Figure 1.1: A generic microarray experiment (Causton *et al.*, 2003). After the mRNA is extracted, it is labelled with one of the two different dyes. Both samples are allowed to hybridise with the microarray where the labelled samples bind to their complementary sequences. Afterwards, the microarray is scanned with a laser to excite the dyes to determine the abundance of the samples on each spot.

1.1.2 Clustering

Generally, gene expression data are normalised, filtered and finally genes with similar expression profiles are clustered into groups. The biological hypothesis behind this is that similarly expressed genes have a common biological characteristic, for example participation in the same biological process, or regulation by a common transcription factor (Segal *et al.*, 2004; Zhu, 2003). Clustering methods can also be used to tentatively annotate genes with unknown function based on the concept of “guilt by association”, where the function is inferred from known genes with similar expression patterns (Jupiter and Vanburen, 2008). Identifying these groups of genes creates an informative description of the biology and presents a comprehensive overview that is occurring in a particular dataset. This makes it possible to locate those areas of biology that warrant a more detailed investigation (Doniger *et al.*, 2003).

Eisen *et al.* (1998) were among the first to apply a clustering algorithm to microarray expression data, and showed that by using clustering, it was possible to group together genes that were known to have similar functions. Hierarchical clustering, using uncentred correlation distance and centroid linkage, was used to analyse yeast microarray data, under various stages in the organism’s life cycle as well as different abiotic conditions (Figure 1.2)

(Eisen *et al.*, 1998). As a result of the early availability of free clustering and visualisation software for expression data, this is currently an extremely popular and commonly used clustering algorithm (D'haeseleer, 2005).

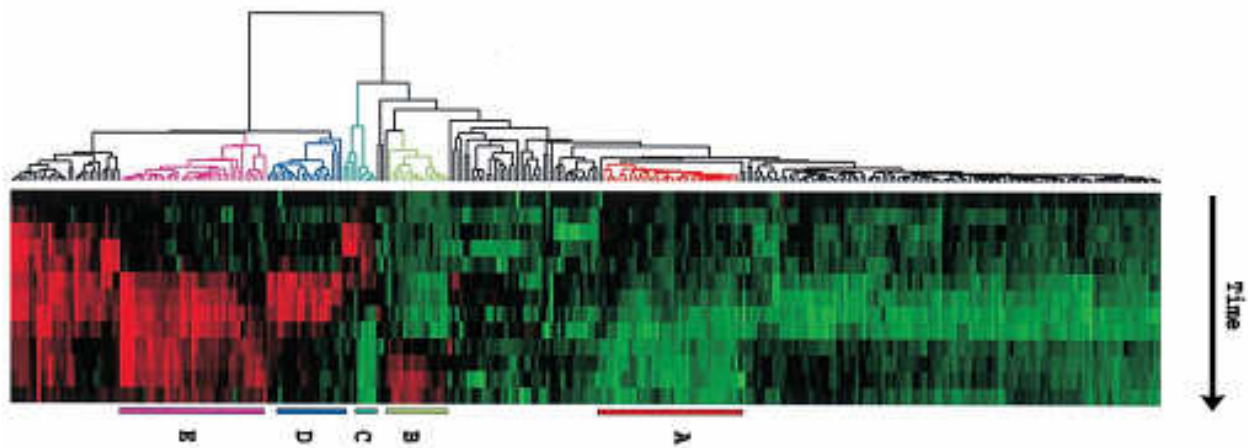


Figure 1.2: Example of the output from a clustering algorithm (Eisen *et al.*, 1998). Shown is the dendrogram, illustrating the similarities of the genes, together with a heatmap showing how the expression of the genes changes over time. Heatmaps are measures of expression levels (vectors) where the colour scales ranges from saturated green for log ratios ≤ -3.0 and lower, to saturated red for log ratios greater or equal to 3.0 . Black indicates unchanged expressions. Each gene's expression is represented by a column of coloured boxes, and each time point is represented by a row.

Some other general trends have been found regarding clustering real expression data including that single linkage affords extremely poor performance, and so should not be used; complete linkage seems to outperform average linkage; Euclidian distance and Pearson correlation seem to work reasonably well as distance measures, particularly for log-ratio data and absolute-valued data, e.g. Affymetrix data, respectively (D'haeseleer, 2005).

Ultimately, there is no “one-size-fits-all” solution with regard to clustering. The interpretation of the clusters is highly subjective, as no precise definition of what a cluster should be exists (Causton *et al.*, 2003). As a result, a cluster may be an arbitrary shape and size. Each clustering algorithm imposes its own set of biases on the data, and as such, each algorithm can give widely differing results on noisy real-world data, such as expression data. Statisticians argue that clustering is often misapplied and the results over-interpreted, and that clustering is best suited to determining relations between only a small number of variables, rather than deriving patterns from thousands of genes (Vos, 2005). In addition, hierarchical clustering has

been criticised as there are no compelling reasons why a hierarchical structure should be imposed on gene expression data (Causton *et al.*, 2003).

Despite essentially becoming a standard in visualising expression data, it has been shown that *k*-means clustering and Self Organising Maps (SOM) outperform hierarchical clustering, with hierarchical methods being particularly unreliable for obtaining good high-level clusters and provide sub-optimal results for large data sets (D'haeseleer, 2005).

1.1.3 Post-processing of microarray data

The analysis of clusters is largely dependant on access to previously described features of the genes being studied (Diehn *et al.*, 2003). Various publicly available resources exist, which catalogues diverse attributes of genes, ranging from their localisation within the genome, to the enzymatic function of the proteins they encode, to their position in a metabolic pathway. These resources include SwissProt, LocusLink, UniGene, GenBank, Protein Information Resource (PIR), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Ensembl, amongst many others (Diehn *et al.*, 2003). In addition, genes may be annotated by assigning them to the functional categories of the Munich Information Center for Protein Sequences (MIPS) classification, or to the Gene Ontology terms, by using descriptions from public databases, such as those mentioned above (Chung *et al.*, 2004). These resources provide exceptional depth and coverage with regard to the functional data available for a given gene, but are not designed to effectively explore the biological knowledge associated with hundreds or thousands of genes in parallel (Dennis *et al.*, 2003).

Traditionally, scientists have investigated genomic molecular biology in a “one-gene-at-a-time” approach. However, with the recent emergence of high throughput technologies, such as microarrays, this has led to the rapid growth of genome-scale datasets and a rapidly increasing number of publicly available genomic sequences (Baerends *et al.*, 2004). The availability of complete genome sequences, and genome-wide measurements of gene expression, provides the means to understand function, expression and regulation of the genome (Zhu, 2003).

Microarray methods provide a global evaluation of changes in the gene expression of a cell at a given instant. Although the genome is mostly invariant in each cell of an organism, the genes can have different expression patterns related to environmental conditions, in response to a particular treatment, condition, or developmental program (Lelandais *et al.*, 2004).

Microarray expression analysis has become one of the most widely used techniques for the assessment of mRNA transcript levels on a genomic scale, allowing tens of thousands of genes to be assayed in a single experiment (Khatri *et al.*, 2002). A transcriptome state can thus be defined as the expression levels of all the genes expressed in a cell population at any given time (Lelandais *et al.*, 2004). These expression patterns can possibly implicate unknown genes in various cellular processes and assign putative functions, as well as facilitating the identification of genes that are co-expressed in a transcriptional unit (Zhu, 2003). Consequently, for a given genome, different transcriptome states can be observed, depending on complex regulatory networks and homeostasis. These patterns, or expression profiles, depict subsets of transcripts that reflect the gene activity at a given moment in time. These profiles represent a “genetic fingerprint” that characterises the cell or tissue being studied and provides a base upon which to start an investigation (Khatri *et al.*, 2002).

With the increased use of microarrays in genome-wide transcription profiling and the general research focus shifting from single genes to large gene sets, understanding what is occurring in the underlying biology presents a huge challenge for biologists. However, genome-scale approaches generate large amounts of data, and an efficient and flexible solution is required in order to access and interpret the data (Doniger *et al.*, 2003). Even a simple task such as retrieving the functional information for a set of genes can be time consuming. To further complicate matters, without the assistance of appropriate bioinformatics tools, visualising and statistically analysing the data can be challenging and a nontrivial task for biologists (Zhang *et al.*, 2005).

1.2 Tools for the post-processing of microarray data

Although microarray technology has been used for many years, and has seen exponential growth, the analysis of the data remains a challenge to many investigators. The difficulty primarily lies in interpreting the list of differentially expressed genes, and how to plan new experiments given that knowledge (Jupiter and Vanburen, 2008).

Genome-wide gene expression analyses offer unique opportunities to study the interactions of genes in metabolic pathways, and to characterise gene regulatory networks. Although it was traditionally thought that these networks act as a linear sequence of events, it is now known that the processes may consist of a network of events, with cross-talk among pathways caused by the induction or suppression genes that belong to more than one pathway (Zhu, 2003). Microarray experiments have enabled the study and identification of gene regulatory

networks, such as transcriptional regulation (Zhu, 2003). In addition, by exploiting the large-scale gene expression datasets, mostly from *Saccharomyces cerevisiae* and *Escherichia coli*, this has led to the discovery of global structures governing metabolic and regulatory networks (Zimmermann *et al.*, 2004). It has also been observed that genes that encode transcription factors are frequently observed among genes that are rapidly suppressed by environmental and developmental stimuli (Zhu, 2003). Transcription factors often have crucial functions in regulating multiple pathways, especially the genes that are involved in several related pathways (Zhu, 2003). Multiple-genome comparisons have also yielded interesting observations on the modularity and connectivity distributions of gene expression data (Zimmermann *et al.*, 2004).

The introduction of high-throughput methodologies has generated experimental data at rates that exceed knowledge growth. While researchers are beginning to appreciate the statistical rigours required for the analysis of genome-scale datasets, the rate-limiting step in knowledge growth occurs at the transition from statistical significance to biological discovery (Dennis *et al.*, 2003). The analysis of the large amount of data presents a tremendous challenge to biologists and new tools are needed to help gain biological insights from these experiments. To interpret the biology of these genetic profiles, investigators must analyse the data in the context of other information such as the biological, biochemical, and molecular function of the translated proteins (Khatri *et al.*, 2002). This is particularly challenging for a human analyst because large quantities of largely irrelevant data often buries the useful information (Khatri *et al.*, 2002).

There are many tools that are available for the analysis of data generated from whole genome microarray experiments, with each tool able to analyse the data at different biological levels. A greater understanding of the biological mechanisms within organisms becomes possible with the availability of complete genome data, in combination with high-throughput screening methodologies such as microarrays. In addition, numerous databases provide annotation at different biological levels. These include databases on the annotation of genes according to the Gene Ontology (GO) nomenclature (Ashburner *et al.*, 2000), metabolic pathways as in KEGG (Kanehisa *et al.*, 2004) and BioCyc (Krieger *et al.*, 2004), or Transcription Factor Binding Sites (TFBS) in TRANSFAC (Matys *et al.*, 2003) to annotate promoters.

Several currently available tools provide an interpretation of gene clusters but are often specialised in their analyses. For example, FatiGO (Al Shahrour *et al.*, 2004), GeneLynx

(Lenhard *et al.*, 2001) and Gostat (Beissbarth and Speed, 2004) are powerful tools for GO term identification; GoMiner (Zeeberg *et al.*, 2003), MAPPFinder (Doniger *et al.*, 2003) and DAVID (Dennis *et al.*, 2003) propose GO and metabolic pathway interpretation; Genome2D (Baerends *et al.*, 2004) visualises transcriptional elements; GeneXPress (Segal *et al.*, 2004) identifies DNA binding sites that are unique to the genes in each cluster; and offers multiple visualisations of the data; MiCoViTo (Lelandais *et al.*, 2004) proposes metabolic pathways and incorporates transcription regulation visualisation; metaSHARK (Pinney *et al.*, 2005) predicts enzyme-coding genes from unannotated genome data and places them on generic metabolic pathways; and WebGestalt (Zhang *et al.*, 2005) uses data obtained from different public resources and offers an integrated platform to perform various analyses such as a GO analysis, metabolic pathways and chromosomal distributions.

With the advent of whole genome microarray chips, and with the increasingly large number of publicly available genome sequences, such tools will be indispensable for the interpretation of large complex datasets, particularly those from transcriptome studies. In other words, these tools are used to create a comprehensive overview and interpretation of the expression profiles.

Several of the abovementioned tools, namely FatiGO, Genevestigator, GeneXPress, Genome2D, GoMiner, MAPPFinder, and WebGestalt will be discussed and for each, the approach and methodology used will be reviewed. Subsequent to this will be a brief introduction to MADIBA, the topic of this study.

1.2.1 FatiGO

FatiGO (Al Shahrour *et al.*, 2004; 13) is a tool that attempts to map biological knowledge onto sets of genes by extracting Gene Ontology (GO) terms that are significantly over- or under-represented in a given set of genes from a genome scale experiment, such as microarrays. FatiGO currently includes the GO associations for the genes of several diverse organisms (currently *Homo sapiens*, *Mus musculus*, *Bos taurus*, *Gallus gallus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Streptomyces coelicolor*) as well as the genes whose proteins are contained in the SwissProt database. As far as possible, FatiGO attempts to use curated associations of genes to GO terms.

FatiGO differs from most other technologies in that it extracts annotation information by using ontologies. In its simplest representation, ontologies provide a structured description of biological information that is extremely useful for the management of data. One of the most widely accepted biological ontologies is the Gene Ontology (Ashburner *et al.*, 2000) which organises information for molecular function, biological processes and cellular components. While it is possible to utilise biological information extracted from literature, this approach has serious drawbacks, most notably that pre-processing the information is absolutely necessary, or else the volume of data will be excessive for common online usage. Ontologies can thus be used as a quick and efficient information mining tool for the identification and validation of clusters of co-expressing genes studied. By using GO terms, interactivity becomes feasible, and additionally, GO terms have a clear biological meaning. GO represents the biological knowledge as a directed acyclic graph (DAG) where higher nodes represent more general concepts and deeper terms are more precise.

For a group (cluster) of genes, FatiGO extracts the GO terms which are deemed relevant with respect to a reference set of genes, which is typically the rest of the genes from the experiment. Figure 1.3 shows the initial submission page. Genes are submitted to FatiGO by uploading a list of gene identifiers, including identifiers from HGNC, EMBL, UniProt/SwissProt, UniProtKB/TrEMBL, Ensembl, RefSeq, EntrezGene, Affymetrix, Agilent, PDB, and many more. It is only required that the identifiers be annotated in Ensembl, and any gene that is not will not be used in the further analyses. These terms are considered to be relevant by the application of Fisher's exact test, and FatiGO also takes into account the multiple-hypothesis testing nature of the statistical contrast performed. The p -values calculated from Fisher's exact test for 2×2 contingency tables are adjusted using three different methods of multiple testing. These include the step-down minP method, which controls the family wise error rate (the probability of making a Type I error over a family of tests); and two variations on the false discovery rate (FDR) test, which calculates the expected number of false rejections among the rejected hypothesis. The FDR-type tests include the Benjamini and Hochberg method which controls the FDR only under independence and some specific types of positive dependence of the test statistics, and the Benjamini and Yekutieli method which offers strong control under arbitrary dependency of test statistics.

To determine the significance of the GO terms, a Nested Inclusive Analysis (NIA) is used where Fisher's exact test is performed recursively on all the terms from level 3 to 9 of the GO

hierarchy, until the deepest significant level is obtained. Only the results from this level are reported for each branch. Since genes are often annotated at different levels, the NIA is used instead of directly using the annotation of the genes at the deepest level possible. In addition, a level of abstraction is chosen and genes annotated at deeper levels are assigned to this level. This improves the efficiency of the test because there are fewer terms to test and more genes annotated to each term. However, the selection of the level is arbitrary.

FatiGO has since continued development and is now part of a suite of tools known as Babelomics (Al-Shahrour *et al.*, 2006; Al-Shahrour *et al.*, 2005), which includes FatiGO, FatiGO+ (finds differential distributions of biological terms, using GO terms, KEGG pathways, InterPro motifs, TRANSFAC motifs, and CisRed motifs, between two groups of genes), Tissues Mining Tool (finds tissues where the genes of two groups display differential distributions), Marmite (finds differential distributions of bio-entities extracted from PubMed between two groups of genes), FatiScan (detects blocks of functionally related genes (GO, KEGG, etc.) with coordinated behaviour across a list of ranked genes using a segment test.), GSEA (detects blocks of functionally related genes with over- or under-expression using the Gene Set Enrichment Analysis) and MarmiteScan (uses chemical and disease-related information to detect related blocks of genes in a gene list with associated values).

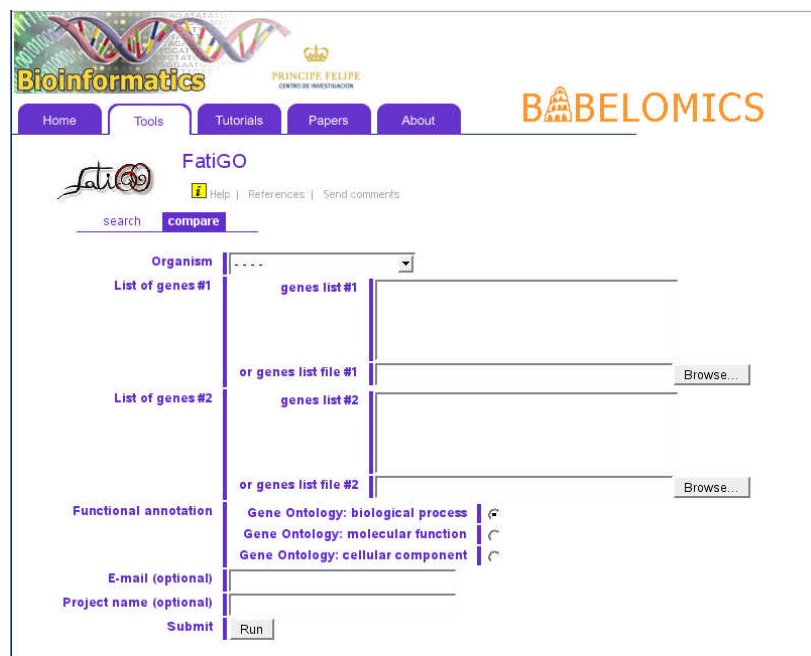


Figure 1.3: Front page of the FatiGO tool (Al Shahrour *et al.*, 2004). Shown are the boxes for gene identifier submission, selection of an ontology, and organism under investigation.

1.2.2 Genevestigator

Genevestigator (Zimmermann *et al.*, 2004; Laule *et al.*, 2006; Zimmermann *et al.*, 2005; 14) is an online tool that allows biologists to query several microarray gene expression experiments by mining Affymetrix GeneChip data. Genevestigator consists of a gene expression database and provides a number of functions that allows the data to be queried and analysed. Users are able to retrieve the expression patterns of genes through a chosen selection of environmental conditions, growth stages and organs, as well as identify genes specifically expressed during selected stresses. The objective of this application is to direct gene function discovery and aid in the design of new experiments by providing plant biologists with contextual information on the expression of genes.

Genevestigator consists of a MySQL database, and is accessed through PHP script pages. The database contains information on the experimental and annotation data, pre-processed data as well as tables for the control of the workflow. Genevestigator was originally conceived for use on *Arabidopsis thaliana*, but has since progressed to include other organisms. Currently, there are five organisms that are available for data analysis in Genevestigator: *A. thaliana*, human, mouse, rat and barley. The annotation data are primarily obtained from public data repositories. For example, the *A. thaliana* data were obtained from experiments from the Grussem Laboratory, the Functional Genomics Center Zurich, NASCArrays, ArrayExpress at EBI and GEO at NCBI.

Genevestigator has several tools for the analysis of the data, and each can be grouped as possessing either a gene-centric approach, which reports the signal intensities for individual genes; or a genome-centric approach, which reports lists of genes fulfilling chosen criteria. The gene-centric approach tries to answer questions such as “How is gene X expressed in a series of conditions?”, whereas the genome-centric approaches try to determine which genes are expressed in selected conditions. The tools available in Genevestigator include:

- Digital Northern, which retrieves the intensity values for the input genes. The user is able to choose only those experiments which exactly fit single or several criteria, such as a specific organ, growth stage or environmental factor.
- Gene Correlator compares the signal intensities of two genes throughout a set of chosen experiments, with a Pearson’s correlation coefficient used to determine the relationship between the expression signals of the two genes.

- Gene Atlas similarly provides the average signal intensity of a gene of interest in all organs or tissues annotated in the database. Conversely, Genevestigator can also output a list of genes which have signal intensity that is above a chosen threshold in selected organs.
- The Gene Chronologer is based on the Boyes growth stage ontology (Boyes *et al.*, 2001) and outputs the signal intensities of a gene of interest at certain representative sections of the organism's life cycle. For example, for *Arabidopsis* there are 10 stages, while in mouse there are 7 embryonic and 5 post-natal stage groups. In addition, the user is able to obtain a list of genes with an expression above a chosen threshold at a given growth stage.
- Response View provides a similar functionality as Gene Atlas and Gene Chronologer, except it is based on stress response annotations. For each condition several representative experiments are chosen, and for each stress factor, the corresponding control is given allowing direct comparison.
- The Meta-Analyzer utility (Figure 1.4) is designed to study the gene expression profiles of several genes simultaneously in the context of environmental stress, organs and growth stages. The output is a heat map of the normalised signal intensities, which have been clustered using single, average or complete linkage hierarchical clustering. This tool is particularly useful for comparing members of gene families and to identify clusters of similarly expressed genes.
- The Database and Documentation sections provide users with technical and annotation information about the experiments in the database, as well as practical information about the tools, including details about the statistical procedures, probe set specificity, normalisation, interpretation of the data, and precautions to avoid over-interpretation.

Each tool in Genevestigator attempts to utilise the best available source of data for processing, while unsuitable data are ignored. For example data from RNA extracted from whole adult plants will be unsuitable for use in relating to specific organs, although it may be used in other calculations. Missing information does not impact the analysis as corresponding arrays are not used in subsequent analyses. In addition, ambiguous data or annotations are ignored.

Genevestigator uses the concept of a “meta-profile”, which can be defined as a representative vector of expression under a given condition, and underlies many of the analyses. This meta-

profile can then be used to show how strongly a gene is expressed under different conditions or in different organs. This concept is used in Genevestigator to characterise expression across anatomy, development, stimuli (diseases or drug treatments) and mutations.

Since Genevestigator is an analysis tool and not a data repository, only a reduced version of the data are stored. However, links to the full MIAME compliant data are provided. Nonetheless, Genevestigator still contains a coherent dataset contained in a reference expression database, and utilises a meta-analysis system. It aims to allow biologists to study the expression and regulation of genes in a broad variety of contexts by summarising information from hundreds of microarray experiments, into easily interpretable results. This type of meta-analysis is core to understanding the spatial and temporal regulation of genes, to identify or validate biomarkers, and to find out which expression pathways are commonly affected in different diseases and conditions.

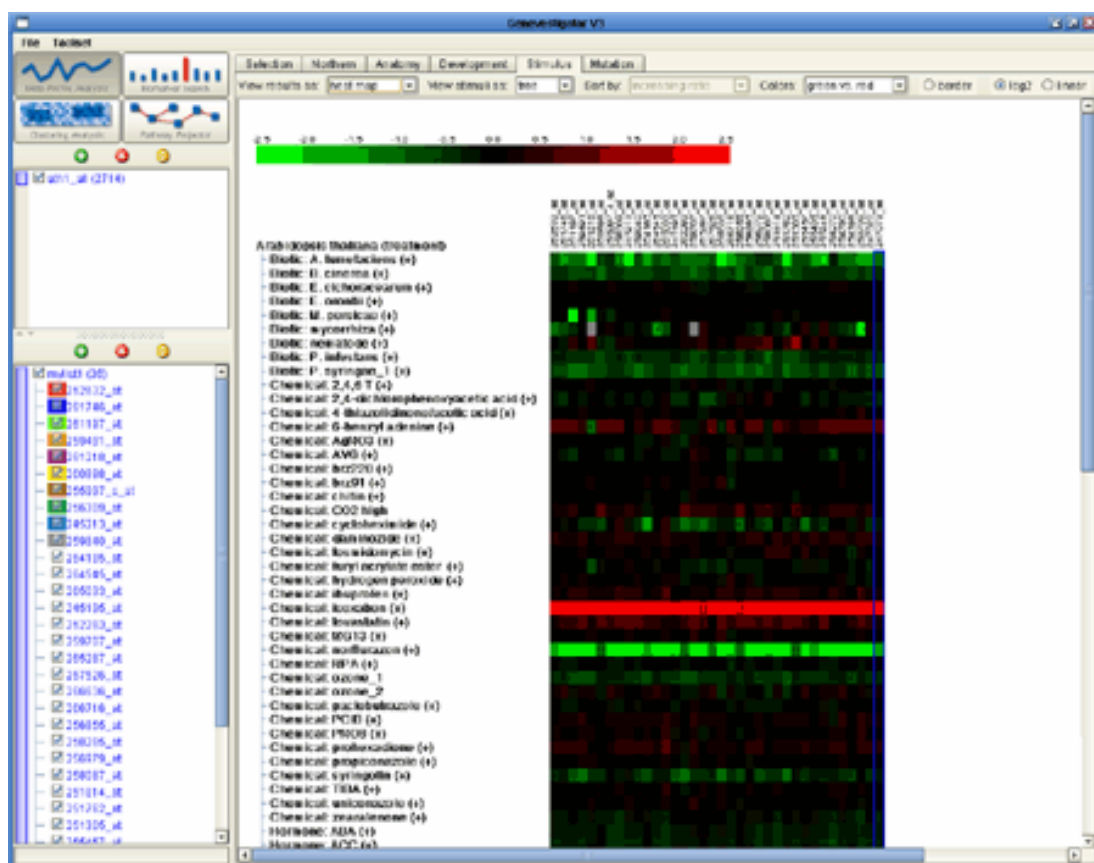


Figure 1.4: The Meta-Analyser tool shows the response of several genes to a large compendium of stimuli. Here, a cluster of *Arabidopsis* genes is illustrated that shows a strong induction response to isoxaben and a strong repression response to norflurazon.

1.2.3 GeneXPress

GeneXPress (Segal *et al.*, 2004; 15) is a Java application that was designed to be a general-purpose visualisation and analysis tool, to “decipher” and extract useful information from gene expression experiments. The program consists of a suite of tools that are able to identify biological processes represented by each cluster, identify DNA binding sites that are unique to the genes in each cluster, and examine multiple visualisations of the expression and sequence data.

For the input of the expression and the corresponding cluster information, there are three possible formats that GeneXPress can use. They are: tab-delimited files, GeneXPress's own format (a *.gxp* file) and files generated with TreeView, a commonly used software package for the visualisation of expression data.

GeneXPress uses XML-based file formats, so it is easy to convert the output of cluster and motif-finding algorithms to such a format and use them in GeneXPress. The XML files that GeneXPress utilises for input are a *.gxp* file for expression and cluster information, a *.gxa* file for gene annotation information, and a *.gxm* file for gene motif information.

Through statistical analysis of clusters relative to databases of gene function annotations, such as GO or KEGG, GeneXPress is able to associate each cluster with one or more biological processes. The data can be obtained from various other sources, including protein sequence motifs (e.g. InterPro), protein complexes, or any user defined gene group that contains annotations for the genes. This gene annotation information is loaded into GeneXPress either as a tab-delimited file or as a *.gxa* file. Given two gene sets, GeneXPress is able to identify pairs of gene sets that have a statistically significant overlap between their member genes. The procedure whereby this takes place is that, with two given gene sets, every pair of genes in both sets is compared using a statistical test that is based on hypergeometric distribution. Each comparison is associated a hypergeometric *p*-value and GeneXPress provides the option to define a cut-off *p*-value to filter that data so that only significant overlaps are reported. Such an analysis is used when a gene set from the cluster is compared against a collection of gene sets that represent biological processes, such as the GO database. Thus, it is possible to identify the biological processes represented by the genes in each cluster. The graphical view is a matrix of the two collections of gene sets, where each coloured entry indicates that the two gene sets have a statistically significant overlap. The intensity of each coloured spot represents the fraction of genes in the overlap. It is possible to save the image, as well as

generate a table providing the image data that is sortable by the annotations, clusters or p -values.

By using a similar analysis for motifs, GeneXPress can identify motifs that are present in the promoter regions of the genes in each cluster. The loaded *.gxm* file contains the motifs and also points to a FASTA file containing the promoter regions of all the genes. Each motif is encoded using the Position Specific Scoring Matrix (PSSM) representation, which is used to score each putative binding-site for its fit to the motif.

GeneXPress provides several visualisations that allow both global and detailed views of expression profiles, promoter regions and motifs. The primary views that are available are the *birdseye view*, the *tree browser view* and the *cluster view*. The *birdseye view* shows the entire expression data, with a dendrogram illustrating how the genes and experiments are grouped. The *tree browser view* allows the user to browse through the data hierarchically, viewing a node, its children, and its parents. This view allows the user to focus on a certain subtree (Figure 1.5). The *cluster view* shows in detail the expression of the genes in the cluster that currently selected in the *birdseye* or *tree browser* views. The annotations and experimental descriptions of the cluster are also displayed in the cluster view. There is also an option to sort the genes by performing a hierarchical clustering on the current cluster. A *sequence view* is also available when performing a motif analysis. This view shows the sequence of the clusters and as well as highlighting the matching motif consensus sequence positions on the clusters.

It is possible to control the various properties of each of the views, including the colour and maximum intensity of the induced and repressed expression values, the colour for missing values, the width and height of the pixels, and the colour and width of padding lines on top of the expression data.

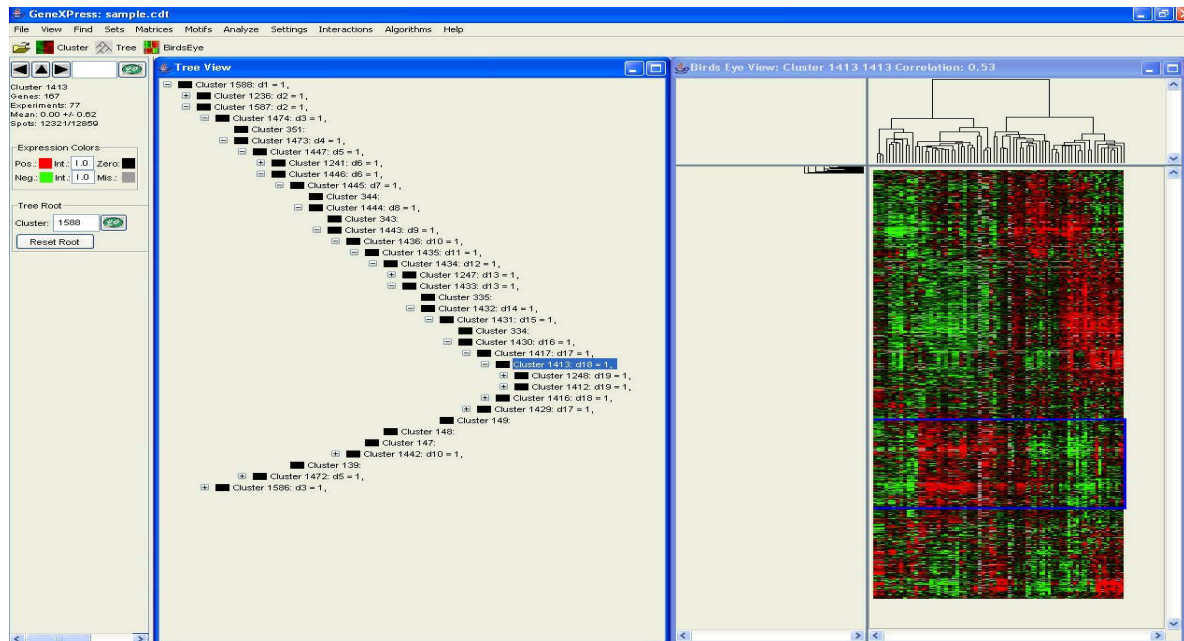


Figure 1.5: View of the tree browser view (left) and the birdseye view (right) from GeneXpress (Segal *et al.*, 2004). A cluster selected in the tree browser view is highlighted (in blue) on the birdseye view. This cluster is illustrated in greater detail in the cluster view (not shown).

1.2.4 Genome2D

Genome2D (Baerends *et al.*, 2004; 16) is a Windows-based software tool for the visualisation of a bacterial transcriptome on a linear chromosome map, constructed from annotated genome sequences. Programmed in Borland Delphi 6, it is easy to use and easily accessible because of its low system requirements. Due to its object-orientated design, it is possible to easily extend Genome2D, so new algorithms and tools can easily be implemented. Genome2D has an update function which automatically connects through the Internet to receive updates.

Genome2D facilitates the analysis of transcriptome data by using different colour ranges to depict differences in gene expression levels on a genome map. This output format enables the visual inspection of the transcriptome data, and can quickly reveal transcriptional units and involvement of possible transcriptional operator sites, without prior knowledge on expression level cut-off values.

Genome2D's drawing module visualises a bacterial genome with all its individual genes in a single computer screen – the Genemap (Figure 1.6). This comprehensive genome map enables quick identification of biologically relevant information such as gene orientation, operon structure, transcriptional terminators or regulator binding sites. Using a simple tab-delimited input file, subsets of genes can be visualised by single or multiple colourings. This colour

input file consists of two columns: one with the name of gene to be coloured and the other indicating the respective colour or a value. This value can be a gene-expression ratio or differences in transcription levels. When values are used, the Genemap colours the genes using a colour gradient, where the degree of up- or down-regulation is depicted by the intensity of the colour. This feature enables easy and rapid identification of genes that are transcriptionally linked, e.g. operons. In a multiple transcriptome analysis experiment, such as a time-course experiment, all the datasets can be loaded as separate input files and subsequently shown in animation. Through this, the changes in gene expression can be readily recognised. The images and data tables from Genome2D can easily be exported for further use in other presentation programs. The image of the Genemap can be saved as a WMF (Windows MetaFormat) or as a BMP file.

In addition to the visualisation abilities, Genome2D contains a toolbox of bioinformatics utilities including several data-extraction and conversion algorithms. The combination of both extraction and visualisation allows subsequent rounds of analyses to be performed, and so an increase in complexity is achieved. This facilitates rapid, easy and intuitive analysis of genomics data.

These features include BLAST support routines, which allow the user to perform a BLAST search locally or at the NCBI. Also included are search algorithms that are able to produce a weight matrix (consensus sequence/motif) and screen the genome for specific sequences or patterns, such as binding sites of transcriptional regulators. Genomic tools are provided for the analysis of the data, with functions such as randomly cutting a chromosome into fragments, randomising the genome sequence for statistical analysis, selecting a number of random genes to create a gene list, and extraction of the coding and non-coding regions. A section of Genome2D was designed primarily for the detection of K-boxes (ComK-binding sites), although most of the available routines can be also be used for any box or pattern analysis. Genome2D includes two proteomics tools: trypsin digestion and pI calculation. Finally, Genome2D also allows for the reformatting of several formats, including GenBank files to Excel or FASTA files, and FASTA databases to a set of single entry files.

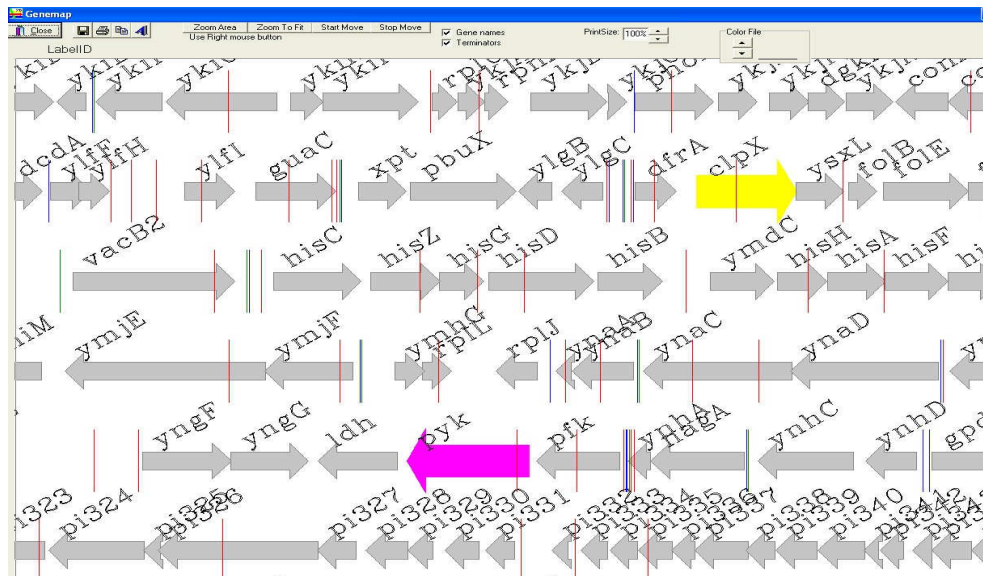


Figure 1.6: View of the Genemap, showing a portion of the *L. lactis* IL1403 genome. The genes (arrows) can be coloured according to user specifications. The boxes are drawn as lines with the *cre*-boxes in red, the -35 boxes in green and -10 boxes in blue.

1.2.5 GoMiner

GoMiner (Zeeberg *et al.*, 2003; 17) is a Java-based program package that organises lists of genes of interest for biological interpretation in the context of the Gene Ontology. Such genes could include up- or down-regulated genes from a microarray experiment. GoMiner accesses the information, *via* the internet, from the GO database, which is updated monthly. For users running a Unix-based operating system, an automated script has been written to enable the installation of the local database. This is particularly useful when a high-speed internet connection is not available. GoMiner can accommodate either the default GO hierarchy (the GO Consortium's database of categories and gene associations) or any user customised version.

GoMiner was developed as a tool for the interpretation of biological information, and facilitates the analysis, organisation and visualisation of results. The GO Ontology is used in order to gain insights into the biological meaning of the gene lists that are produced as a result of the microarray experiments. Instead of using a gene-by-gene approach to analyse the data, GoMiner classifies the genes into biologically coherent categories and assesses these categories.

GoMiner receives input in two files, each with a list of genes: one with the total set of genes on the array, and the other with a list of genes to be queried. The latter list is a subset of the

total set and includes those genes that the user flags as interesting, such as those exhibiting altered expression levels. With the gene list, GoMiner is able to calculate the enrichment or depletion of categories with genes that have changed expression.

The output of these analyses appears in a graphical user interface (GUI), which takes the form of a three-panel window (Figure 1.7). These panels are:

1. A left-hand panel that lists the genes, the databases from which their identities were derived, and whether the gene has changed (under-expressed, over-expressed, or unchanged).
2. A middle panel showing a tree visualisation that is similar to the AmiGO browser. Genes that are displayed in the tree are tagged with green down arrows, red up arrows, or grey circles to indicate under-expressed, over-expressed and unchanged genes respectively. After each category are a series of calculations, which (from left to right) indicate the total number of genes in the category (in black); the relative enrichment of under-expressed genes (green); the relative enrichment of over-expressed genes (red); and the relative enrichment of all changed genes (blue). Fischer's exact p -value is calculated for all the enrichments.
3. A right-hand panel showing all appearances of a gene selected from the left or middle panel, within the GO hierarchy. There is also a tab which switches the view to display all the data values, and can be sorted according to any of these.

By clicking a gene in the tree view, the user is able to submit the gene as a query to an external data resource, such as LocusLink, PubMed, NCBI 3D Structures and BioCarta and KEGG, as implemented by the NCI Cancer Genome Anatomy Project (CGAP). GoMiner also provides a second visualisation, a compact, interactive directed acyclic graph (DAG). Nodes in the DAG can be "moused-over" to list flagged genes or clicked to highlight all possible pathways connecting it to the root. The DAG shows in compact form the spanning hierarchy for all flagged genes and can be searched to locate a specific node. The summary results are downloadable to tab-delimited text files, which can be used further in a spreadsheet program for statistical analysis. The DAG can also be exported as a SVG (scalable vector graphic) file.

A command-line version of GoMiner has also been developed to complement the GUI version. This allows GoMiner to be integrated with other tools *via* scripts or pipes. An

example would be allowing GoMiner to interact with its companion program, MatchMiner, as a pre-processor to obtain gene names for input. A high-throughput version of GoMiner is also available, which allows the output data stream to be coupled with integrated downstream analyses. This allows for the automated recognition of interesting results that are buried within a large number of exploratory experiments.

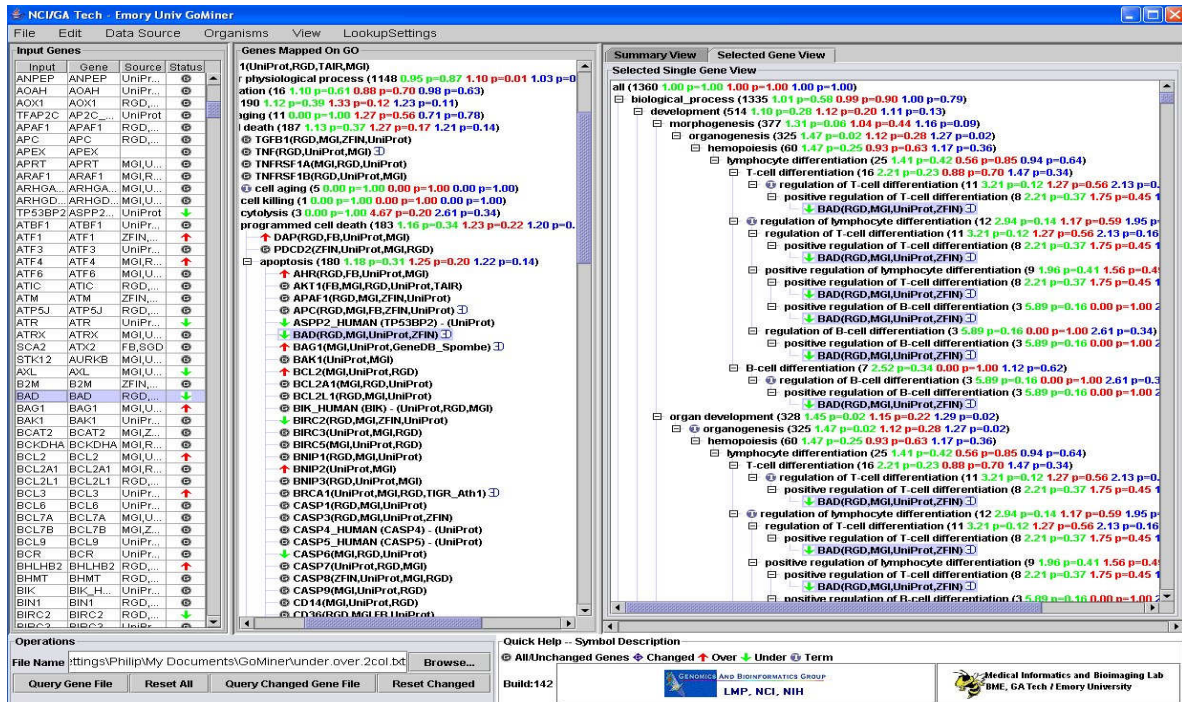


Figure 1.7: Screenshot of GoMiner’s GUI. The green down arrows, red up arrows, or grey circles indicate under-expressed, over-expressed and unchanged genes respectively.

1.2.6 MAPPFinder

MAPPFinder (Doniger *et al.*, 2003; 18) is an accessory program that works together with GenMAPP and the annotations from the GO Consortium to identify global biological trends in gene expression data, thus integrating the GO analysis with biological pathways. The package is currently written in Visual Basic, and therefore only runs on Windows systems. GenMAPP is a program that was designed to view and analyse gene expression data, which is represented on MAPPs. MAPP initially stood for MicroArray Pathway Profiler, due to its development for the analysis of microarray experiments, but now has been renamed to Map Annotator and Pathway Profiler, to include other high throughput experiments. These GenMAPP-produced files graphically show biological pathway relationships between genes and/or gene products. Examples of the types of MAPPs that can be represented are metabolic pathways, signal transduction cascades, subcellular locations, gene families, or lists of genes

associated with Gene Ontology categories. Each MAPP contains symbols that depict genes and gene products, as well as other objects, such as receptors, ligands, membranes and ribosomes, and arrows representing the relationships between the objects. Each gene object on a MAPP is identified by a gene identifier from one of the GenMAPP accepted gene ID systems, such as Affymetrix, SwissProt, InterPro or LocusLink. This gene identifier can also be linked to a gene expression dataset which allows GenMAPP to colour the gene, according to user-defined criteria.

To be useful in identifying gene expression changes across a biological system, the available information on pathways needs to be expanded. This is required since GenMAPP and other pathway programs, such as KEGG, focus on well-defined metabolic pathways, and would benefit greatly from a broader base of pathway information. To resolve this problem, MAPPFinder was developed which uses information from the GO Consortium. MAPPFinder provides the ability to perform the analysis on genes associated with GO terms, as well as on MAPPs that are local to the user's computer. MAPPFinder dynamically links the gene expression data onto the GO hierarchy and for each node (GO term or local MAPP), performs a series of calculations. These are: 1) the number of genes meeting the user-defined criteria; 2) the number of genes measured at this node; 3) the number of genes associated with this node; 4) the % genes meeting the criterion ($\text{genes meeting the criterion} / \text{genes measured} * 100$); and 5) the % genes measured in the node ($\text{genes measured} / \text{genes associated} * 100$). For the GO terms only, the same calculations are done for the cumulative total of the number of genes that meets the user's criteria in a GO term along with its descendants (children, grandchildren, etc). This "nested" number of genes gives a more accurate and complete representation of the changes in gene expression that are associated with a particular GO term. The final calculation is the Z score, which is a standard statistical test under the hypergeometric distribution. A positive Z score indicates that there are more genes meeting the criterion in a GO term or MAPP than would be expected by random chance. A negative Z score indicates the opposite, that is, there are fewer genes than expected. These values are all displayed after each node on the MAPPFinder Browser (Figure 1.8). This browser is similar in appearance to the AmiGO Browser and the GO hierarchy or local MAPP folder structure can be navigated in a similar fashion.

Included in the MAPPFinder Browser are several filters that can be used to further refine the tree to highlight only the nodes that meet specific criteria. It is possible to filter by the

percentage of genes changed, the number of genes changed, and the Z score. All the nodes that pass the filters are highlighted in yellow. It is also possible to search for a specific GO term, MAPP, or keyword using the Word Search option. If a match is found, that node will be coloured blue and if the node was already coloured yellow from the above filters, the node will be coloured green. Clicking on a GO term in the MAPPFinder Browser window opens a MAPP in GenMAPP that lists all the genes associated with that GO term so that the user can view the pathway.

MAPPFinder accepts gene expression data from a GenMAPP Expression Dataset (.*gex*) file, which is the output file from GenMAPP after the expression data has been entered. This file contains the selected colour sets and criteria that will be used to filter the data.

The results from a MAPPFinder analysis are exported as a tab-delimited text file (.*txt*) that can be viewed in a spreadsheet program to allow the user to perform additional filtering and sorting of the results. This aids the user in identifying MAPPs that are of particular interest.

MAPPFinder results are specific to a particular build of the GenMAPP Gene Database. In particular, the GO data are constantly updated, and with each update, new nodes may be added and existing nodes moved within the GO or made obsolete. As a result, the date of the GO data being used is displayed, with each version of the GenMAPP Gene Database containing a specific build of the GO. Thus each time the Gene Database is updated, the MAPPFinder results must be recalculated.

Currently, MAPPFinder is only available for *C. elegans*, *Drosophila*, human, mouse, rat, yeast and zebrafish. There are, however plans to expand it further to include other species in the GO Consortium.

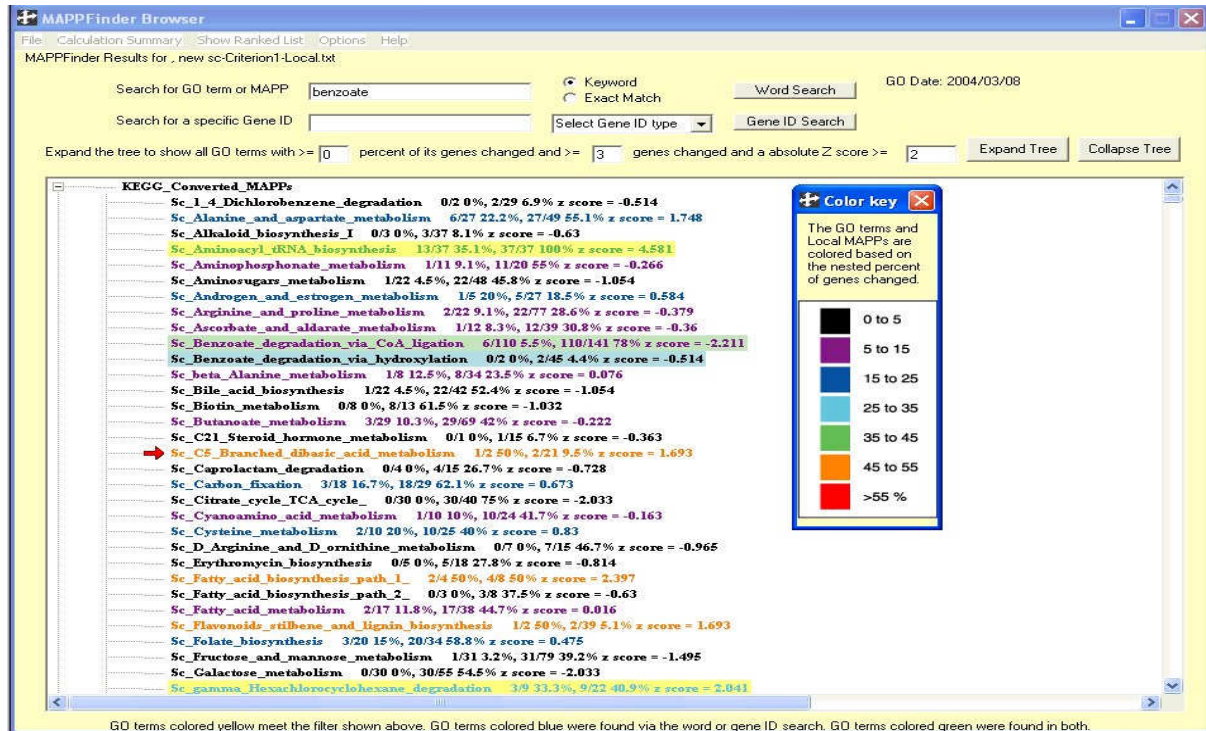


Figure 1.8: The MAPPFinder Browser showing the results after the analysis on a number of local MAPPs.

The terms highlighted in yellow are those that passed the filters on top, the terms in blue are those that were found in the text search, and the terms in green met both criteria. Also included is a colour chart which defines the colours of the terms.

1.2.7 WebGestalt

WebGestalt (Zhang *et al.*, 2005; 28) is a “WEB-based GEne SeT AnaLysis Toolkit” and uses data obtained from different public resources to offer an integrated platform to perform various analyses such as a GO analysis, metabolic pathways and chromosomal distributions. WebGestalt is a web-based integrated data mining system which aims to aid biologists in exploring large sets of genes.

WebGestalt utilises the GeneKeyDB database schema which in turn uses the ORACLE relational database. This database uses a strong gene- and protein-centric viewpoint. Gene and gene product information is primarily taken from NCBI LocusLink, Ensembl, Swiss-Prot, HomoloGene, Unigene, CGAP, UCSC, GO Consortium, KEGG BioCarta and Affymetrix, and is currently implemented for human and mouse. The web pages were implemented using PHP scripts.

WebGestalt is composed of four main modules: gene set management, information retrieval, organisation/visualisation and statistics.

Gene set management – This module uploads, saves, retrieves and deletes gene sets (Figure 1.9). It accepts gene sets by files, GO categories or chromosome location ranges. Gene identifiers that can be used include EntrezGene IDs, Swiss-Prot IDs, Ensembl IDs, Unigene IDs, and Affymetrix probeset IDs. The management module also performs Boolean operations to generate the union, intersection and difference between two gene sets. By recursively applying these operations it is possible to combine information from any number of sets of genes.

Information retrieval – The information retrieval module provides rapid access to the existing information for all genes in a gene set. Currently 20 attributes can be retrieved, which include nomenclature, identifiers to different databases, map and function information. This data can be exported as a tab-delimited text file, or Excel file.

Organisation and visualisation – While the information retrieval module provides rapid access to large sets of data, it does not really help biologists explore the information associated with the gene sets. This module is intended to assist the biologist in exploring large genes under various biological contexts. The tools available include: GO Tree (organises the gene sets based on the GO DAG using several visualisations, including an expandable tree, bar chart, and enriched tree); KEGG Table and Maps, and BioCarta Table and Maps (identifies pathways involved in biological studies); Protein Domain Table (organises gene based on PFAM protein domains); Tissue Expression Bar Chart (organise a gene set based on large-scale, publicly available gene expression data derived from a variety of tissue and organ types); Chromosome Distribution Chart (visualise chromosome distribution of the genes); and PubMed Table and GRIF Table (organise genes according to their co-occurrence in publications, based on the gene-publication association information retrieved from the LocusLink database.).

Statistics – The statistics module recommends and performs statistical tests to suggest biological areas that are significant and warrant further investigation. To identify functional categories with significantly enriched gene numbers in a gene set, the gene set of interest is compared to a reference gene set. If the gene set of interest is a subset of the reference gene set, a hypergeometric test is used to evaluate significance. If the gene sets are independent, Fisher's exact test is used instead.

WebGestalt's advantages include the ability to retrieve a large variety of information for all genes in a gene set, a multitude of visualisations under different biological conditions, assistance in selecting the appropriate statistical tests, a simple and intuitive user interface, and the ability to use Boolean operations on selected gene sets. Modules within WebGestalt can also easily be used by third-party applications, as implemented by WebQTL.

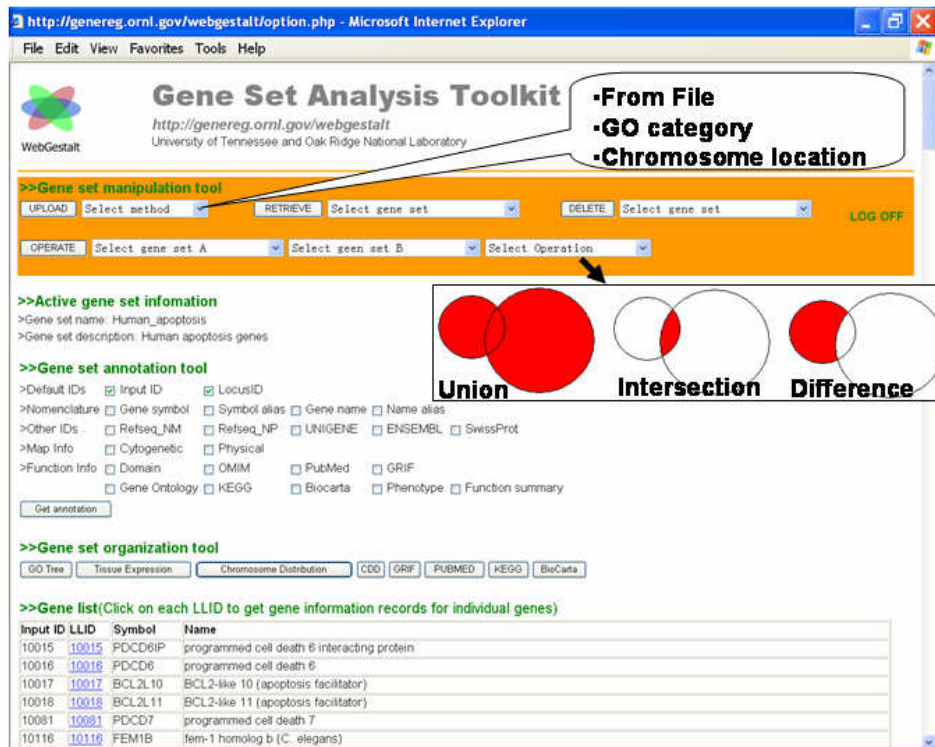


Figure 1.9: User interface of WebGestalt. This page shows the gene set management tools, active gene set information, gene set retrieval, gene set organisation, and a gene list of the active genes. Gene set management tool also allows the user to apply Boolean operators (union, intersection and difference) to the gene sets.

1.2.8 MADIBA

MADIBA (MicroArray Data Interface for Biological Annotation) (Law *et al.*, 2008; 4) is the topic of this dissertation, and a detailed description of the implementation of MADIBA and its analysis modules is provided in the following chapter. A screenshot of the front page of MADIBA is shown in Figure 1.10.

MADIBA is a web-based tool, which subjects a cluster of genes to five diverse analyses, namely: 1) a search of over-represented GO terms in the cluster; 2) mapping of the cluster's gene products onto metabolic pathways using the KEGG representation; 3) visualisation of the chromosomal localisation; 4) a search of over-represented motifs in the upstream

sequences of the genes and 5) an organism specific analysis. This tool aims to assist researchers in the identification of possible reasons for the common expression of a cluster of genes.

MADIBA has currently been implemented for *Plasmodium falciparum*, *Oryza sativa* (rice), *Arabidopsis thaliana*, and *Pectobacterium atrosepticum* (strain SCRI1043; *Pba*). Malaria is a devastating disease, particularly in Africa, so understanding how its causative agent, *Plasmodium*, functions is essential. Rice and *A. thaliana* are model species for monocotyledonous and dicotyledonous plants respectively (Rensink and Buell, 2004), and plant analyses are useful particularly for gaining insights into improving crops in both developed and developing countries, for example orphan crops such as cassava, cowpea and pearl millet, which are important for food security in Africa. In addition, *Plasmodium* is related to plants as the apicoplast (apicomplexan plastid) is reminiscent of the chloroplast (Marechal and Cesbron-Delauw, 2001; Ralph *et al.*, 2001). *Pba* is a destructive bacterial pathogen that infects potatoes (Bell *et al.*, 2004).

While other tools similar to MADIBA, such as WebGestalt, FatiGO and GoMiner exist, MADIBA differs in that it has a wider range of analyses which can be performed in an integrated fashion, for example, it performs a GO analysis as well as a Transcription Regulation analysis. Of the previously mentioned tools, MADIBA is most similar to WebGestalt (Zhang *et al.*, 2005), which also obtains information from different data sources and provides an integrated set of analysis tools to assist researchers in mining this gene set. WebGestalt, however, does not provide information on transcription regulation, and currently only works for human and mouse data. MADIBA is unique in the diverse organisms it is able to analyse – a bacterium (*Pectobacterium atrosepticum*), a eukaryotic pathogen (*Plasmodium falciparum*), and plants, both a monocotyledon (rice) and a dicotyledon (*Arabidopsis thaliana*).

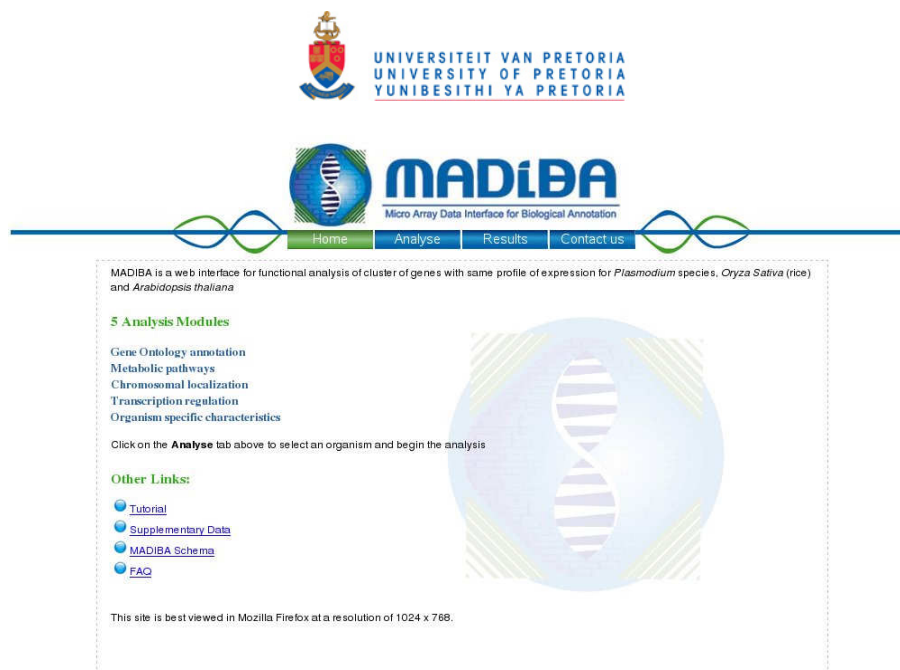


Figure 1.10: Front page of MADIBA. From this page, the user is able to begin a new analysis or retrieve previous results, as well as view the tutorial and supplementary data.

1.3 Conclusion

While several tools have been discussed, this is by no means a complete list of available tools. Each tool takes a different approach, whether a web-based or a downloadable tool, and which features to focus on. Some, are annotative tools and do not directly provide a visualisation, while other tools are exploratory and give the results in the form of a graphical representation. This is sometimes preferable to a purely statistical approach since with a visualisation, it is possible to quickly identify patterns in the data. Visualisation of information enables the researcher to deal with the overwhelming amount of information associated with a gene set by taking advantage of human pattern processing abilities (Zhang *et al.*, 2005).

Many of the tools were designed for specific needs by the developers of the tool, and cannot readily be applied generically. For example, Genome2D was specifically designed to meet the requirements of the research group to investigate bacterial genomes. On the other hand, GeneXPress is not specific to any organism, but takes any gene expression data from the user. Indeed, through the individual differences, a broad range of approaches were observed, all which are used to tackle a similar problem, that is, the identification of the biological meaning of large amounts of data. In Table 1.1, a basic summary of the specifications of each of the discussed tools is presented, and Table 1.2 contains information about the general features of each tool.

Most of the tools mentioned have in some way required the data to be pre-clustered to identify significant terms in the smaller gene set. However, such clustering methods suffer from limitations, most notably that visualisations are usually only in terms of the group members, with no information about its nearest neighbours; and that the quality of the relationships are not easily inferred (Jupiter and Vanburen, 2008). Two tools with a different approach include ClutrFree and StarNet. ClutrFree (Bidaut and Ochs, 2004) is a desktop-based tool which attempts to provide a flexible and generic platform that allows the user to compare different annotation and analysis approaches to a microarray data set. Pattern recognition allows visualisation of the relationships in a directed graph (tree) that assists the user in deriving biological conclusions. StarNet (Jupiter and Vanburen, 2008) is a tool that visually explores the correlation networks radiating from a selected gene. Using this technique presents the possibility of deriving and inferring transcription regulatory networks.

Analysis from a whole genome experiment can be an extremely complex procedure, but the benefits that can be gained are enormously beneficial in biological discovery. In particular, investigating changes in the entire genome after a certain treatment will aid in the understanding of the complexities involved in a biological system. With the advent of whole genome microarrays, tools such as those discussed will become even more significant in understanding cellular functions and determining what the data means biologically.

Table 1.1: Table of the basic specifications of the discussed tools. Client-server indicates whether or not the tools need to connect to a database on a remote server.

Tool	Platform	Client-server	Input	Output
<i>FatiGO</i>	Web	Yes	Gene list	GO enrichments
<i>Genevestigator</i>	Web	Yes	Gene list	Visualisation of genes satisfying a given condition
<i>GeneXPress</i>	Java	No	Gene expression, motif and annotation files	Visualisation of expression data: tree, cluster and birdseye view
<i>Genome2D</i>	Windows	No	Genome sequence	Genome map, with selected genes and genetic elements marked.
<i>GoMiner</i>	Java	Yes*	Gene list	Tree visualisation and a directed acyclic graph (DAG)
<i>MAPPFinder</i>	Windows	No	Expression data from GenMAPP	Tree visualisation of clusters that meet user's criteria.
<i>WebGestalt</i>	Web	Yes	Gene list	Visualisation depending on biological context
<i>MADIBA</i>	Web	Yes	Gene list or sequence file	Visualisation depending on selected module

* Database is generally accessed via the internet, but can also be installed locally.



Table 1.2: General features of the discussed tools

Tool	Functional annotation	GO	Metabolic pathway	Genomic localisation	Transcription regulation	Organism
<i>FatiGO</i>	•	•				Human, mouse, bovine, chicken, rat, zebrafish, <i>Drosophila</i> , yeast, <i>C elegans</i> , <i>Arabidopsis</i> , <i>Streptomyces coelicolor</i> .
<i>Genevestigator</i>	•	•				<i>Arabidopsis</i> , human, mouse, rat and barley.
<i>GeneXPress</i>	•	•			•	Non-specific.
<i>Genome2D</i>	•			•	•	Bacteria.
<i>GoMiner</i>	•	•	•	•		Organisms in GO.
<i>MAPPFinder</i>	•	•	•		•	<i>C elegans</i> , human, mouse, zebrafish, <i>Drosophila</i> , rat, yeast.
<i>WebGestalt</i>	•	•	•	•		Human, mouse
<i>MADIBA</i>	•	•	•	•	•	<i>Plasmodium</i> , rice, <i>Arabidopsis</i> , <i>Pectobacterium</i>

Chapter 2 – MADIBA Implementation

2.1 Introduction

In this chapter, the details of how MADIBA is implemented will be discussed. This will include information on the data currently used in MADIBA, as well as information pertaining to what each analysis module does and how the analyses are performed.

2.2 Data sources

The downloaded data currently consists of data from the PlasmoDB (Fraunholz and Roos, 2003) database for *Plasmodium falciparum* (release 5.4), TIGR (Yuan *et al.*, 2005) for rice (*Oryza sativa* ssp *japonica* cv *Nipponbare*) (Osa1 database release 5), TAIR (Rhee *et al.*, 2003) for *Arabidopsis thaliana* data (TAIR7) and Sanger-SCRI (Bell *et al.*, 2004) for *Pectobacterium atrosepticum* strain SCRI1043 (*Pba*). Stored are the gene name, functional annotation, GO identifiers, chromosomal localisations, and the enzymatic annotations (EC identifiers) from the above data sources, as well as the EC numbers proposed by the KEGG Orthology results.

Data pre-calculated by programs before being stored in a PostgreSQL (version 8.1.4) database include putative metabolic enzyme predictions using PRIAM (Claudel-Renard *et al.*, 2003). These predictions are calculated based on enzyme profiles from position specific weight matrices (PSSM). Also stored are the 1500 nucleotides upstream of the start codon (ATG) based on previous experimental research on *P. falciparum* promoters (Crabb and Cowman, 1996; Dechering *et al.*, 1999; Horrocks *et al.*, 1998; Lanzer *et al.*, 1992), and 1000 nucleotides upstream of the rice and *A. thaliana* genes, as made available by TIGR and TAIR respectively. For *Pba*, the sequence from the end of the previous gene to the start codon was taken, with a maximum of 400 nucleotides. The MADIBA system also identifies putative orthologues between *Plasmodium falciparum* and human proteins, rice and *A. thaliana* proteins, and *Pba* and *Dickeya dadantii* 3937, by performing a reciprocal BLASTP search, with an e-value cut-off of 10^{-15} . In addition, for rice and *A. thaliana*, a BLASTP of all expressed proteins was performed against the proteome of the organism in question (a self BLAST), in order to determine paralogues, or protein ‘families’, for each gene. The BLAST results with an e-value less than 10^{-3} were stored in the database. Sequence information was obtained from the original data source (PlasmoDB, TIGR, TAIR, or SCRI), and BLAST searches were performed using a local version of NCBI-BLAST.

2.3 Database implementation

In MADIBA, each organism possesses its own PostgreSQL database, and is used to store the downloaded and pre-calculated data. Although there are separate databases, each database has a similar structure, consisting of four primary tables: the `annotation`, `databases`, `homologues` and `promo_annot` tables (Figure 2.1).

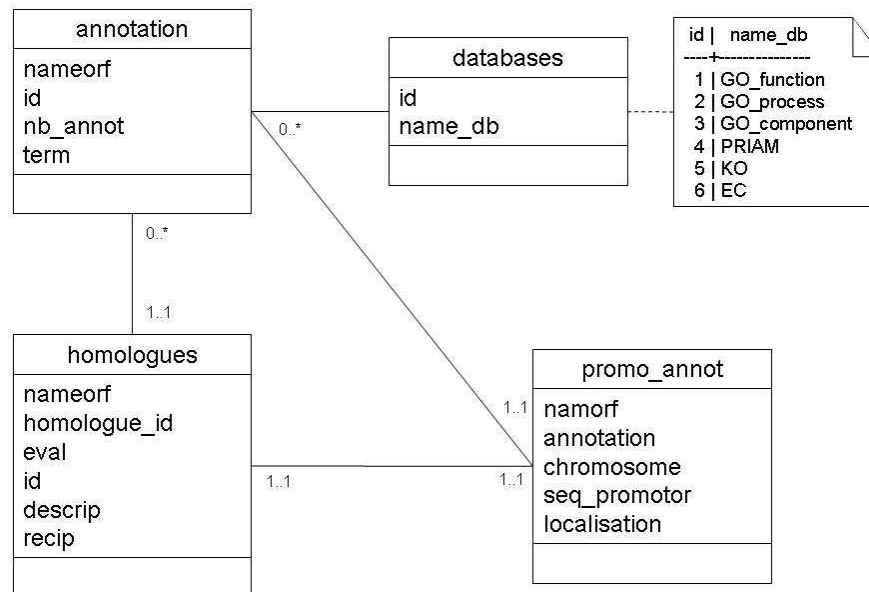


Figure 2.1: UML class diagram showing the four generic tables for the MADIBA databases. The `annotation` table holds the information on the GO and enzymatic annotations, as indicated by the rows in the `databases` table; the `homologues` table holds information on the best BLAST hit for a gene; and the `promo_annot` table contains the functional annotation of the gene, its location on the chromosome and the sequence upstream of the gene.

The `annotation` table contains the annotations for all the genes in the genome with annotations from all three of the GO ontologies, as well as the enzymatic annotations from PRIAM, KEGG and the original annotations. The annotation types are distinguished by using an integer identifier, which is a number from 1 to 6 to indicate the different annotations from the three GO annotations, PRIAM, KEGG and the original annotation. These relations are stored in the `databases` table.

The `homologues` table contains the results of the top BLASTP hit of each protein in the organism's proteome against a model organism – *Plasmodium falciparum* against human, *A.*

thaliana against rice, rice against *A. thaliana*, and *Pba* against *Dickeya dadantii*. In this table, the identifier and functional annotation of the top hit, the e-value, the percent identity (the number of matches in the query and match), and whether or not the hit was a reciprocal hit are stored. Only the best hits with a minimum e-value of 10^{-15} are stored. If no hit meets this requirement, a null value is stored. A reciprocal hit means that if protein A in organism 1 matches protein B in organism 2, the top hit of a BLAST search of protein B in organism 1 will be protein A.

The `promo_annot` table contains the functional annotation of each gene, which chromosome it is found on, as well as the location of the gene on the chromosome. Also stored in this table is the upstream sequence which is used in the Transcriptional Regulation module.

In addition to these standard tables, there are also extra tables that were used in special cases. For example, in *A. thaliana* and rice, there is an extra table containing all the BLAST results from the self-BLAST. These data are then used in the Organism Specific module where the user defines what the paralogue cut-off should be. Similarly, in the *P. falciparum* database, there is a table containing the BLAST hits of the apicoplast genes with the *A. thaliana* proteome.

The databases were populated through Python scripts which parsed the raw data and utilised PyGreSQL (22), a Python module that interfaces to PostgreSQL databases, to connect to the database.

2.4 User interface

The MADIBA interface was written in PHP 5.1.1 and Python 2.4.4, and is provided using an Apache HTTP server (version 2.0.59) on a Sun V880 server, running Solaris 9. The Python pages utilise the standard `cgi` module to allow the Python scripts to be executed as CGI scripts, as well as the `cgitb` module for debugging purposes. The `cgitb` module acts an exception handler and any exceptions that occur in any of the Python pages will be caught and logged for further investigation (Beazley, 2006). The website is accessible using any JavaScript enabled browser at <http://www.bi.up.ac.za/MADIBA>.

2.5 Data submission

A cluster of genes is submitted to MADIBA, either by uploading a file, or directly pasting a set of nucleotide sequences, in FASTA format. Alternatively, a list of gene identifiers can be submitted. The gene clusters are obtained from any clustering algorithm, such as hierarchical or *k*-means, since MADIBA does not perform any clustering.

For *P. falciparum*, *A. thaliana* and *Pba* sequences, a BLASTN search is performed to find similar genes in the relevant organism's genome. For rice sequences, a BLASTX search is performed to allow the possibility of entering gene clusters from the *indica* as well as the *japonica* subspecies. In addition, this will potentially allow orthologous gene clusters from other cereals to be analysed, such as pearl millet. After the BLAST search, the top five hits for each submitted sequence is presented, allowing users to select which of the hits they wish to continue the analyses with, and this list of genes is stored. The BLAST runs are executed using the `NCBIStandalone` class from the BLAST library in BioPython 1.4.4 (10), a set of Python tools for computational biology.

The gene identifiers that are currently accepted by MADIBA are those as determined by the original annotation source, that is, PlasmoDB for *P. falciparum*, TIGR for rice, TAIR for *A. thaliana*, and the SCRI for *Pba*.

The Plasmodium gene nomenclature are in the form `PFX#_#####` where *XX* is the chromosome number and `#####` is an identifier based on the order of the genes. However, some identifiers are also in the form `MALxPy.zz`, where *x* is the chromosome, and *y* and *zz* are roughly based on gene order.

The TIGR gene nomenclature (27) for rice genes follows the convention `LOC_OsXXg#####` where `LOC_Os` indicates an *Oryza sativa* Locus, *XX* is the chromosome number from 01-12, *g* indicates it is a gene, and the hashes are a five-digit code which indicates the gene's position on the chromosome and are numbered from top/north to bottom/south of chromosome. Additionally, different versions of a gene product, e.g. a differentially spliced gene, are denoted by a full stop followed by a number (1, 2, etc). However, when submitting rice identifiers, this information is not required and all splice forms will be retrieved in such cases. An example identifier is `LOC_Os03g17740.1`.

The *A. thaliana* gene nomenclature (26) is similar to the rice nomenclature, where genes are named as ATXg##### where AT means *Arabidopsis thaliana*, X is the chromosome number (1, 2, 3, 4, 5) or M or C for mitochondrial or chloroplast genes respectively, and the hashes are a five-digit number indicating the gene's chromosomal order. Again, an indication of an alternatively spliced gene may be present. In addition, *A. thaliana* gene symbols (usually based on function) may also be submitted. Examples of acceptable inputs are AT1G08680, AT1G08680.1 and ZIGA4.

Acceptable *Pba* gene identifiers are ECA followed by a four digit number, or the gene names, such as ExpI.

The list of genes that is submitted, whether as sequences or identifiers, is stored for one week on the server's file system, and is used by each analysis module to retrieve the necessary information required by that module from the database. In addition, a unique identifier is provided to allow users to later access and retrieve their data. By entering this identifier, users can access a gene list that they had previously submitted, and thus do not have to resubmit any data. In addition, most of the results of analyses will be present, so these analyses will not have to be rerun. These include the results from the Transcription Regulation module, which can take a long time to run. The unique identifiers consist of two parts: an initial alphabetic component and a numeric component. The alphabetic component indicates the organism that is being analysed, where ATH indicates *Arabidopsis thaliana*, OSA indicates rice and PF indicates *Plasmodium falciparum* and PBA indicates *Pectobacterium atrosepticum*. The numeric component is a random 7-digit number that is generated using the `randint` function from Python's built-in `random` module, which in turn uses the Mersenne Twister as the core generator (Beazley, 2006).

2.6 MADIBA modules to analyse gene clusters

Figure 2.2 illustrates the architecture and basic data flow of an analysis in MADIBA as described in the next section.

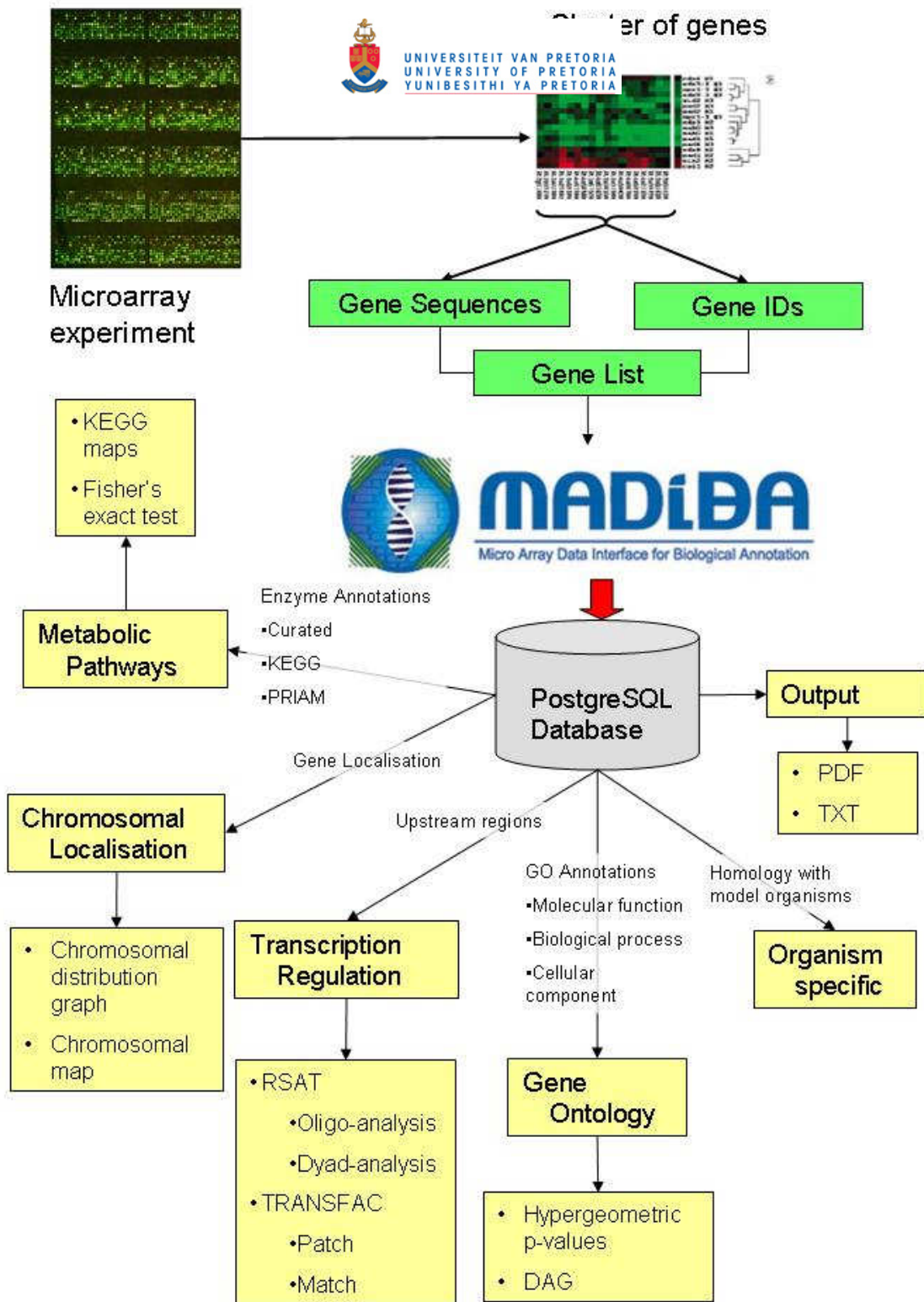


Figure 2.2: A schematic representation of the flow of data through MADIBA. After a microarray experiment, the data are normalised and then clustered, since it is hypothesised that the genes in a cluster have common biological implications. A cluster of genes is submitted to MADIBA, either as nucleotide sequences, or gene identifiers. This list of genes can then be subjected to five analysis modules – Gene Ontology Analysis, Metabolic Pathways Analysis, Transcription Regulation Analysis, Chromosomal Localisation Analysis and an Organism Specific Analysis. Also shown are the data that are required by each of the analysis modules. The results from the analyses can be exported as a PDF file, or as plain text.



Home Analyse Results Contact us

Analysis Menu

A

METABOLIC PATHWAYS GENE ONTOLOGY TRANSCRIPTION REGULATION CHROMOSOMAL LOCALISATION PLASMODIUM CHARACTERISTICS OUTPUT

Selected BLAST Results

B

Your ID is: PF07_1867055
Results are kept for one week

C

PlasmoDB ID	PlasmoDB Annotation
PF10_0324	hypothetical protein
PF14_0360	hypothetical protein
PFC0600w	hypothetical protein
PF11695c	small nuclear ribonucleoprotein (snRNP), putative
PF10_0130	hypothetical protein
PF10_0289	adenosine deaminase, putative
PF10_0323	hypothetical protein
PFC0285c	T-complex protein beta subunit, putative
PF11_0436	hypothetical protein
PFA0145c	aspartyl-tRNA synthetase
PF14_0589	valine - tRNA ligase, putative
PFE0660c	uridine phosphorylase, putative
PF07_0046	50S ribosomal protein L1, putative
PF13_0014	40S ribosomal protein S7 homologue, putative
PF14_0744	hypothetical protein
PF11745c	hypothetical protein

Figure 2.3: Screenshot of the page after submission. Highlighted are the analysis menu with the links to the five analysis modules and the output module (block A), the unique identifier that is provided to the user for later retrieval (block B), and the list of genes that are to be used in subsequent analyses (section C).

2.6.1 Front page

MADIBA is accessible through a simple and user friendly web interface. Figure 2.3 shows the front page after submission of a set of sequences. Block A illustrates the analysis menu with the links to the five analysis modules and the output module. Each analysis module is independent of the others and is accessed individually. Block B shows the unique identifier that is provided to the user (see above), and section C lists the genes that are to be used in subsequent analyses. If gene identifiers were used, the list will also indicate the genes that could not be found in the database.

Below this list of genes are two pie charts (Figure 2.4) – one comparing the number of genes that are annotated as hypothetical, expressed or have some functional annotation; and the other indicating how many genes have the same functional annotation. Three or more genes have to have the same annotation to be included in the diagram. However, annotations are compared using an exact text match, so annotations that are slightly different will not be considered to be the same. A fuzzy text comparison could be used to enhance this feature. Both diagrams are dynamically generated using PHP scripts which utilise the GD library (2),

and are created by retrieving the annotations for each gene from the database and comparing them.

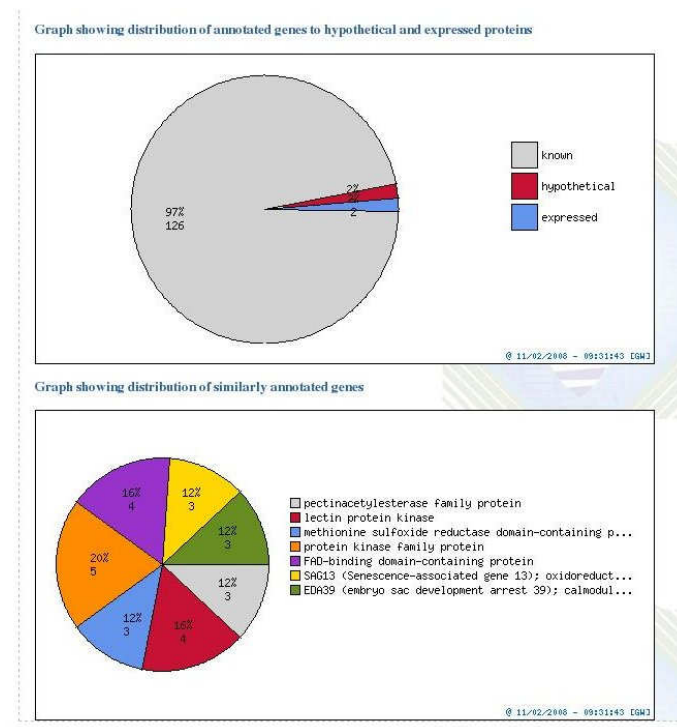


Figure 2.4: The two pie charts illustrating the number genes that are annotated as hypothetical, expressed or have some functional annotation (top); and indicating whether the submitted genes have a common functional annotation (bottom). Three or more genes have to have the same annotation to be included in the diagram.

2.6.2 Gene Ontology module

This analysis module extracts the Gene Ontology (GO) annotations according to the molecular function, biological process and cellular component ontologies. Due to the complex nature of the GO hierarchy, it was decided not to use the “ontology level” approach and instead represent each ontology as separate Directed Acyclic Graphs (DAG). Each DAG is drawn to show, in a single view, the genes from the cluster, and the GO terms that they are annotated to. Each GO term that is found in the cluster is drawn to show its position in the GO hierarchy, in a manner similar to AmiGO’s graphical representation (1) and GO::Termfinder (Boyle *et al.*, 2004). The user is able to select which genes should be visualised, to prevent overly complex graphs (Figure 2.5a).

Once this set of genes, along with the ontology to be analysed is submitted, the GO annotations for the selected genes in the required ontology are retrieved from the database.

The annotations for the genes that were not selected are also retrieved so that the statistics will be consistent for any selection of genes. This means that the p -values are calculated in terms of the whole cluster, irrespective of the genes that are selected by the user. The annotations are stored in a Python dictionary with the GO annotations as the keys, and the genes that they are annotated to as the values.

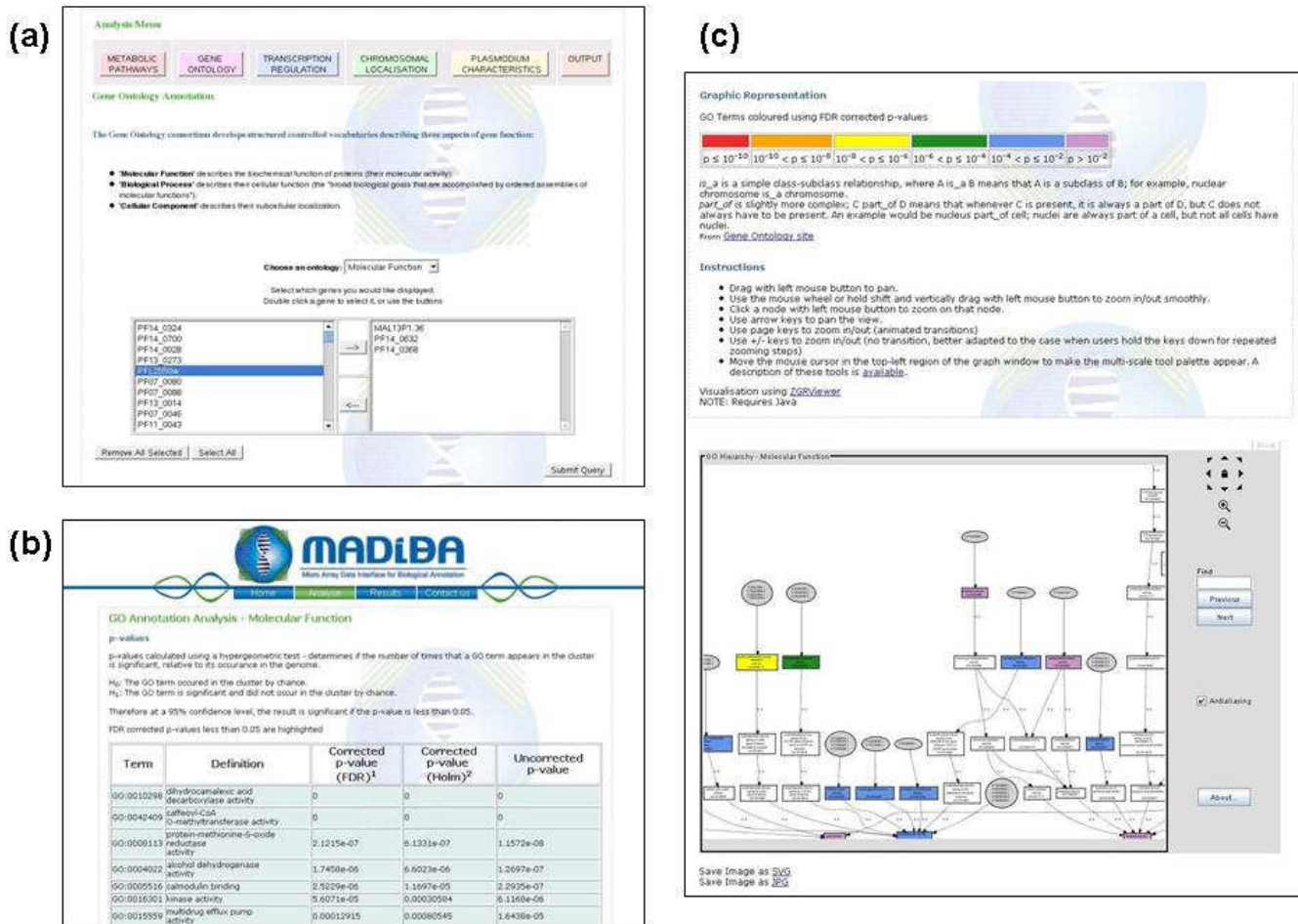


Figure 2.5: Screenshots of the Gene Ontology module. (a) The initial selection page, where users can select which genes should be analysed. Genes are selected or deselected by double-clicking a gene name or using the buttons. (b) The top of the results page showing the hypothesis being tested, followed by a table of p -values, calculated from the hypergeometric distribution to determine significance of each GO term. The FDR-, and Holm-corrected p -values, as well as the uncorrected p -value are shown. (c) Below definition of GO terms is the Directed Acyclic Graph (DAG) showing all the GO terms and their positions in the GO hierarchy. GO terms are coloured based on their FDR-corrected p -value, as indicated by the legend above. Users are able to download the DAG in either SVG or JPG format.

For each GO term, a hypergeometric p -value is calculated, and a multiple hypothesis correction performed to evaluate the significance of the annotation. The adjustment methods used are a False Discovery Rate (FDR) adjustment (Benjamini and Hochberg, 1995), and a Holm adjustment (Holm, 1979), which works under the same assumptions as a Bonferroni correction, but is statistically more powerful (Aickin and Gensler, 1996). The uncorrected and corrected p -values are assembled into a table and any GO terms with a FDR-corrected p -value less than 0.05 is highlighted (Figure 2.5b). Other statistical tests may be implemented in the future to provide greater flexibility to the user in calculating the enrichment of the GO terms. In addition, other adjustment methods are being considered, including the q -value, which is a measure of significance in terms of the FDR instead of the False Positive Rate (Storey and Tibshirani, 2003). A tab delimited text file with each gene and the GO terms associated is provided should users wish to perform their own analyses. The statistical calculations (hypergeometric test and multiple hypothesis corrections) were computed by the R statistics package (2.6.0) (5), which was accessed using the RPy (1.0-RC3) package (24), a Python interface to R.

To draw the DAG, a tree traversal is performed, for each GO term, to the root node of whichever ontology is being analysed, that is, the molecular function, biological process or cellular component term. These GO trees were obtained from the Gene Ontology website, and parsed using NetworkX (0.33) (19), a Python package for the creation and manipulation of graphs and networks. This network information is part of the pre-calculated data, and is made available for future use by using the standard `pickle` module, which serialises the network data structure so that it can be stored as a text file (Beazley, 2006). This means that each time the tree structure is required the data can easily be “unpickled” and used, instead of having to constantly re-parse the data. Once the DAG has been generated, the nodes are laid out using the `dot` program from Graphviz (version 2.16.1) (3), a graph layout and visualisation package. Graphviz is accessed using PyGraphviz (version 0.35) (19), a Python interface to Graphviz. Edges are also labelled according to a node’s relation with its child node – either an *is_a* or *part_of* relation. *is_a* is a simple class-subclass relationship, where A *is_a* B means that A is a subclass of B; for example, nuclear chromosome *is_a* chromosome. *part_of* is slightly more complex; C *part_of* D means that whenever C is present, it is always a part of D, but C does not always have to be present. An example would be nucleus *part_of* cell; nuclei are always part of a cell, but not all cells have nuclei (8). Nodes on the DAG are coloured according to its FDR corrected p -value, where red indicates p -value $\leq 10^{-10}$; orange, $10^{-10} < p$ -value $\leq 10^{-8}$;

yellow, $10^{-8} < p\text{-value} \leq 10^{-6}$; green, $10^{-6} < p\text{-value} \leq 10^{-4}$; blue, $10^{-4} < p\text{-value} \leq 10^{-2}$; purple, $p\text{-value} > 10^{-2}$. The graphs are rendered as both a JPG and a SVG, either of which can be downloaded by the user. The graph is also displayed to the user using the ZGRViewer applet (7), which is a tool to display SVG images (Figure 2.5c). The applet was used as the GO graphs often tend to be extremely large and difficult to view without some form of viewer.

2.6.3 Metabolic Pathways module

When this module is accessed, a list of all the KEGG metabolic pathways is presented, along with an indication of how many enzymes encoded by genes from the input cluster were found in each pathway (Figure 2.6a). Each pathway in the list is linked to its diagram where the protein products of the genes in the cluster are highlighted. This module compares the enzymatic annotation from three different annotation sources, namely the curated annotation from the original data source (PlasmoDB, TIGR, TAIR or SCRI); the semi-automatic KEGG annotation and the automatic PRIAM annotation (Claudel-Renard *et al.*, 2003). The use of these three diverse and independent annotations increases the robustness of the analysis.

To perform this analysis, it is necessary to determine which of the genes in the cluster have been annotated as an enzyme, that is, if the `id` column in the `annotation` table is 4, 5, or 6 (KEGG, PRIAM, or original annotators respectively). These results are stored in a Python dictionary, with the enzymes as the keys, and the number of annotations as the values. A value that was pre-calculated is the list of enzymes that are found in all the pathways, and a value that is determined in each analysis is a list of the enzymes that are present in the genome of the organism of interest.

Different colours are used to indicate the agreement between the three annotation methods, where yellow indicates that the enzyme was annotated by all three annotation sources; red, by any two annotations; blue, by KEGG only; purple, by PRIAM only; and green, by the original annotators only. In addition, any enzyme found in the genome annotation, but not in the cluster is coloured grey (Figure 2.6b). The KEGG maps are coloured using the Python Imaging Library (version 1.1.6) (23), and the coordinates of the enzymes on the maps are given by KEGG in the `conf` files that accompany the image files. The colouring is performed by creating a coloured block with a degree of transparency and having the same dimensions as the enzyme box on the map. The coloured box is then pasted on top of the pathway image at coordinates provided in the `conf` file for this particular pathway. A set of blank template pathways is used to create a new set of pathways for each user.

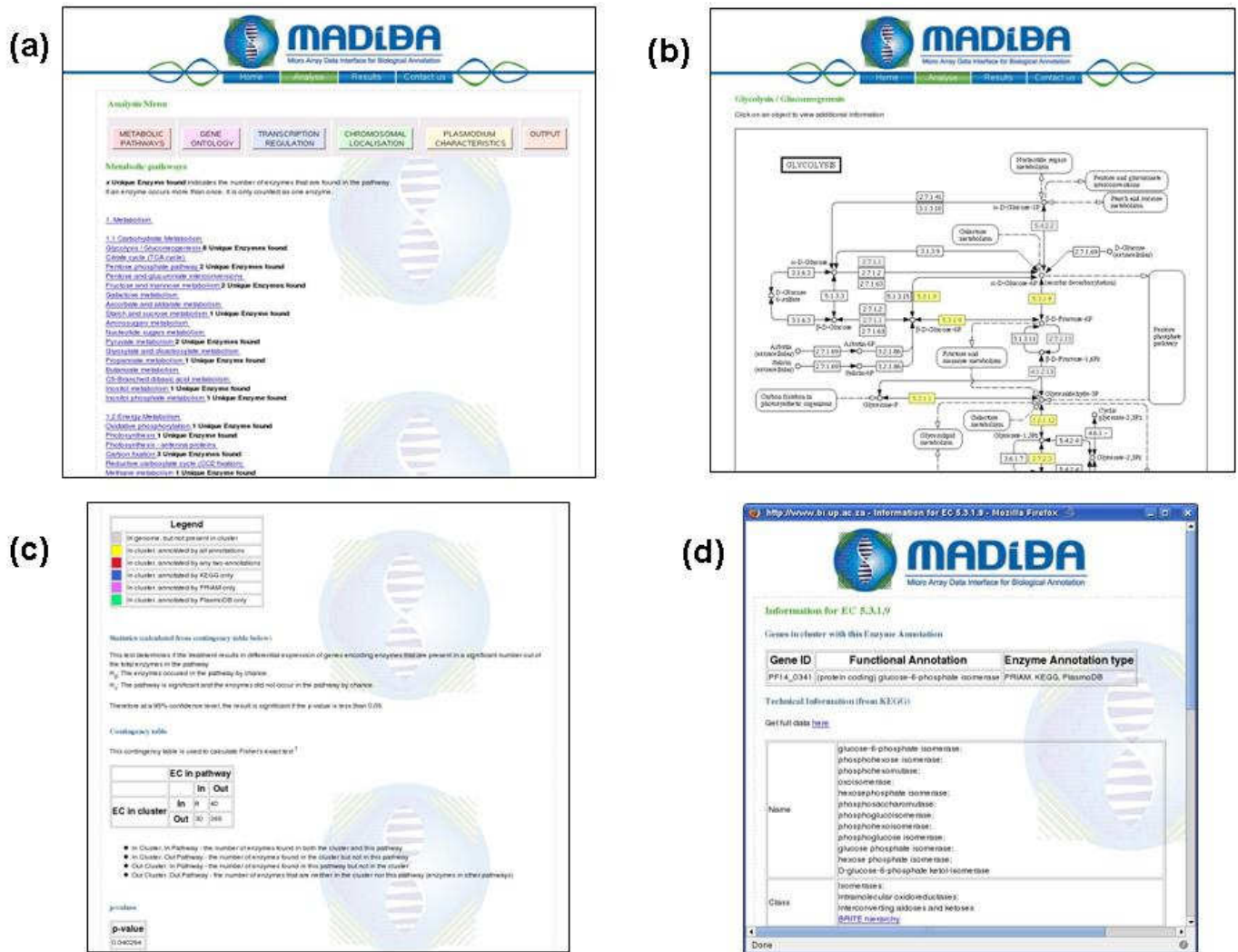


Figure 2.6: Screenshots of the Metabolic Pathways module. (a) Main page of the module, listing the names of all the metabolic pathways, as well as the number of enzymes that are found in each. (b) Example of a metabolic pathway (glycolysis), with the coloured blocks indicating enzymes that are found in the submitted cluster. (c) Blocks in the metabolic pathway diagrams are colour coded according to the number of annotations (PRIAM, KEGG or original annotators) (top), and a p-value is calculated using Fisher's exact test to determine the significance of the pathway. (d) Each element on the pathway is clickable and brings up a window with additional information about that particular compound.

Once the image creation is complete, the p -value for the pathway is calculated for each pathway using Fisher's exact test (Fisher, 1935) to indicate the significance of the pathways, using a 2×2 contingency table where the rows indicate pathway membership and the columns indicate cluster membership (Chung *et al.*, 2004) (Figure 2.6c). This means that the table elements are *In Cluster, In Path* – the number of enzymes found in both the cluster and this pathway; *In Cluster, Out Path* – the number of enzymes found in the cluster but not in this

pathway; *Out Cluster, In Path* – the number of enzymes found in this pathway but not in the cluster; and *Out Cluster, Out Path* – the number of enzymes that are neither in the cluster nor this pathway (enzymes in other pathways). The Fisher's test was performed using R, which is accessed through RPy.

A HTML image map is laid over the image to make the components clickable, which links to the KEGG website and provides additional information on the particular enzyme or compound (Figure 2.6d). This information is obtained by screen-scraping the relevant KEGG site using BeautifulSoup 3.0.3 (9), a Python HTML/XML parser. Clicking on an enzyme that is present in the cluster also provides information as to which annotations were used to describe it, and which genes from the cluster encode it.

2.6.4 Chromosomal Localisation module

This module permits the identification of co-expressed genes on the same chromosomal region. It provides a bar chart showing the distribution of the genes in the cluster across the chromosomes, that is, the number of genes on each chromosome (Figure 2.7). This chart is drawn by simply retrieving the chromosome that each gene in the cluster is found on, calculating the occurrence of each chromosome and sending that information to the image generating functions.

In addition, a schematic visualisation of the genes along the chromosomes is provided, where each chromosome is drawn as a horizontal bar, and each gene is represented by a vertical blue line. The size of the bar is proportional to the size of the chromosomes, and the genes are drawn relative to its actual position on the chromosome. Localisation data was obtained from the original annotation source (PlasmoDB, TIGR, TAIR or SCRI). A mouse-over effect was included to this diagram to allow easier identification of a gene at a particular position, and was accomplished using `wz_tooltip` (12), a JavaScript library. The chromosomal localisation is drawn by using the locations obtained from the original annotation source. A scaling factor is necessary to fit the representation onto a screen, so the *P. falciparum* is reduced by a factor of 254000 and the plants by a factor of 3175000. Assuming that one base pair is the size of one pixel, this is the value required to fit the chromosome onto the screen. Both the chromosomal distribution bar chart and chromosomal schematic are dynamically generated using the PHP GD library.

For *Pba*, the circular genome is represented as a line circle, with the genes represented by blue lines perpendicular to it. Since there is only the single “chromosome”, the chromosomal distribution bar is not drawn.

To further assist in identification of genes on the chromosomes, especially those that are close to one another in the representation, gbrowse (version 1.62) (Stein *et al.*, 2002), a generic genome browser was implemented. When a user clicks on a gene on the chromosomal representation, the gff file required by gbrowse is generated, and contains all the genes from the cluster that occur on that chromosome. Once the necessary files are generated, gbrowse is launched, and the initial view will be of the gene that was clicked (Figure 2.8).

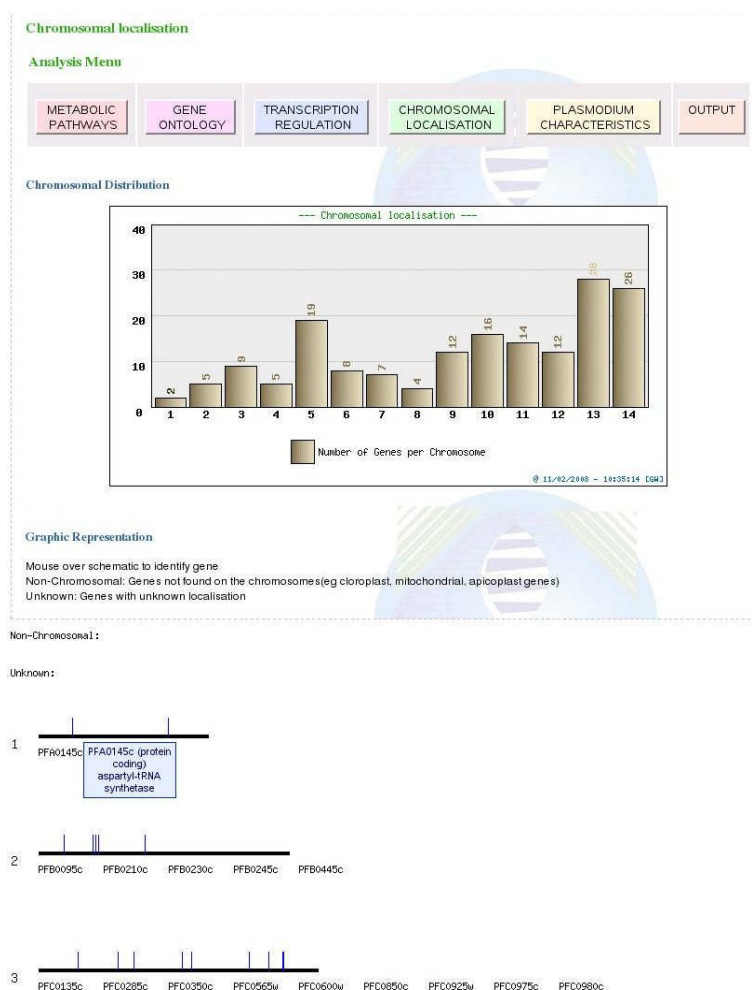


Figure 2.7: Screenshot of the Chromosomal Localisation module, showing a bar chart (top) of the distribution of the genes in the cluster across the chromosomes, i.e. the number of genes on each chromosome. In addition, a schematic visualisation of the genes (b) along the chromosomes is provided, where each chromosome is drawn as a horizontal bar, and each gene is represented by a vertical blue line. The size of the bar is relative to the size of the chromosomes. A mouse-over effect was added to assist in the identification of the genes



MADIBA Chromosome Browser

Showing 2.619 kbp from chr1, positions 175,782 to 178,400

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the ScrollZoom buttons to change magnification and position.

[\[Hide banner\]](#) [\[Hide instructions\]](#) [\[Bookmark this view\]](#) [\[Link to an image of this view\]](#) [\[Publication quality image\]](#) [\[Help\]](#)

Search [Reset] [Flip] ScrollZoom: <<< Show 2.619 kbp >>>

Overview of chr1

Genes in Cluster

RT1G01480.1
RCS2 (1-Amino-cyclopropane-1-carboxylate synthase 2)

RT1G01480.2
RCS2 (1-Amino-cyclopropane-1-carboxylate synthase 2)

Data Source
MADIBA Chromosome Browser

Tracks [\[Hide\]](#)

External tracks italicized Genes in Cluster

Overview track

Image Width 450 640 800 1024

Key position Between Beneath

Track Name Table Alphabetic Varying

Set Track Options... Update Image

Upload your own annotations: [\[Help\]](#)

Upload a file [Browse...] [Upload] [New...]

Add remote annotations: [\[Help\]](#)

Enter Remote Annotation URL [Update URLs]

Figure 2.8: Screenshot of MADIBA's version of gbrowse. Highlighted is the gene that the user clicked, and shows the exon structure of that gene.

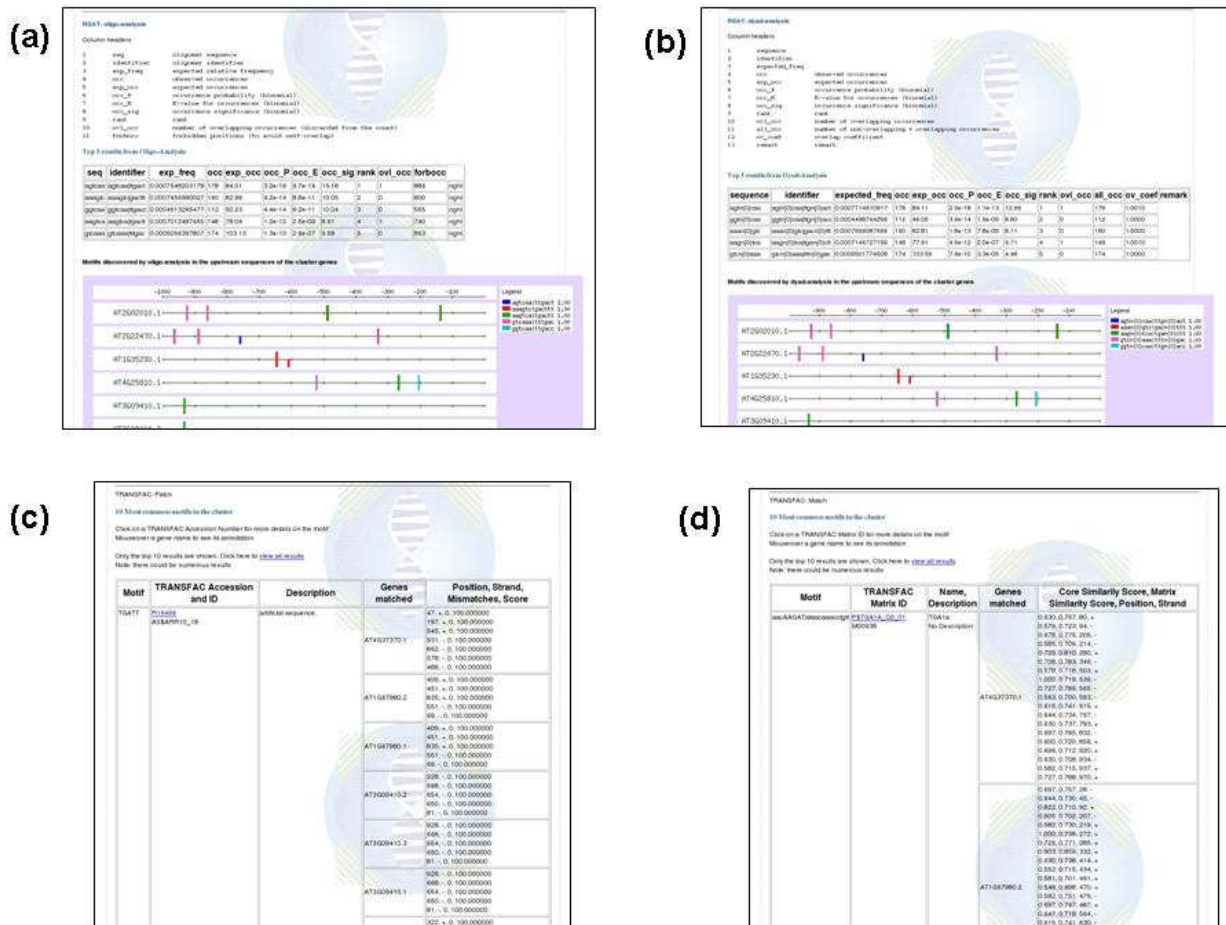
2.6.5 Transcription Regulation module

This module presents an approach of motif identification without any prior knowledge, by automatically detecting potential Transcription Factor Binding Sites (TFBS) in the promoter sequences of co-regulated genes, using Regulatory Sequence Analysis Tools (RSAT) (van Helden, 2003), specifically using the oligo-analysis and dyad-analysis programs (Figure 2.9a-b). Oligo-analysis calculates the occurrence of words (oligo-nucleotides) in a set of sequences, and determines which are over-represented based on a background model. Dyad-analysis detects overrepresented spaced dyads (oligo-nucleotide pairs which are separated by a variable spacer region) in a set of sequences. For each analysis, the five most significant motifs are reported, as well as a link to all the results. RSAT performs a number of statistics to provide an indication of the significance of the detected motif. The p-value represents the probability of the motif being significant when it is not (a false positive). The e-value is an estimation of the expected number of false positives for a series of test, and is often called the Bonferoni-corrected p-value. In MADIBA, a threshold of e-value < 1 was selected.

The upstream regions are also searched for known TFBS in the TRANSFAC database (Professional version 11.1) using the built-in Patch and Match programs (Matys *et al.*, 2006) (Figure 2.9c-d). Patch uses predefined binding site entries and performs a pattern-based

binding site search, while Match uses positional weight matrices derived from alignments of binding sites (i.e. matrix-based search). The ten most common motifs found by each tool are presented. For each identified binding sites, a link is provided to additional information from TRANSFAC about the factors that bind to that site. If the sequence of the binding site is available, it is possible to BLAST it against the genome of the organism under study to see if that same factor is present in the genome. This is useful as the binding sites and binding factors in TRANSFAC are often obtained from other organisms, such as tobacco (*Nicotiana tabacum*).

Both RSAT and TRANSFAC are accessed using system calls, and the results are parsed in Python scripts.



2.6.6 Organism Specific Characteristics module

In the *P. falciparum* cluster analyses, a component of the aim for this module is to identify putative new drug targets. Thus, a list of the genes without human homologues, with their respective annotations is generated (Figure 2.10a). Also, if any genes similar to the apicoplast are present, its closest homologue to *A. thaliana* is identified. Due to its vegetal nature, the apicoplast may provide a target for herbicide-like drugs which will not affect the human host (Marechal and Cesbron-Delauw, 2001; Ralph *et al.*, 2001).

For *Pba* the results from a reciprocal BLAST search against *Dickeya dadantii* are stored.

For the plant analyses, the closest *A. thaliana* orthologue of each rice gene, and *vice versa*, is given, in an effort to identify similar genes (Figure 2.10b). This was accomplished by implementing a reciprocal BLASTP search, with a stringent e-value cut-off of 10^{-15} to identify highly probable orthologous proteins. In addition, a list of all similar genes, based on sequence similarity is returned, representing the paralogues, or protein ‘families’, for each gene. These results are determined by performing a self BLAST, and the user is able to determine the most relevant results by choosing their own e-value cut-off, minimum percent coverage (how much of the query matched the subject) and minimum percent identity (how much of the match corresponded). All the BLAST results from this analysis with an e-value less than 10^{-3} were stored in the database.

As an addition to the Arabidopsis Characteristics module, two analyses were developed to identify similar expression profiles in a submitted set of genes to various experiments. A simple approach that was used identified how a set of genes were expressed under the various experimental treatments which are stored in the DRASTIC database (Button *et al.*, 2006). A more complex approach, named PCA Experiment Comparer, used the expression data from NASCArrays (Craigon *et al.*, 2004) to compare the expression profiles from different experiments to the expression profile of the genes in the submitted cluster. Further details on these approaches are provided below.

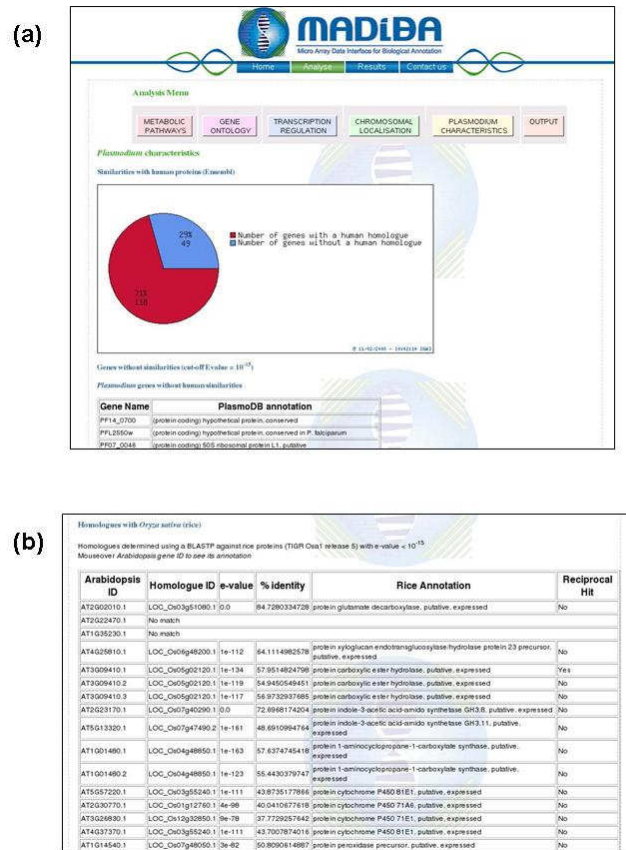


Figure 2.10: Examples from the Organism Specific module. Shown is an output from *P. falciparum* (a) illustrating the percentage of genes that have human homologues; and an example from *A. thaliana* (b) showing the best BLAST results against rice.

2.6.6.1 DRASTIC

To determine which conditions an experiment has a similar expression pattern to, the regulation of a set of genes was retrieved from the experiments contained in the DRASTIC database (11; Button *et al.*, 2006). DRASTIC (Database Resource for the Analysis of Signal Transduction in Cells) is a gene expression database that was developed to record a plant's response to various treatments, including exposure to pathogens (Button *et al.*, 2006). The database contains the treatments affecting several different plant species, but with most of the emphasis on *A. thaliana*. DRASTIC contains lists of genes with their up- or down-regulation in response to various pathogens and other treatments, such as cold, drought and salt. The data includes genomic, EST, northern and microarray data that is obtained from peer-reviewed publications, and is all human curated to ensure accuracy and to standardise the gene nomenclature (G. Lyon, SCRI, personal communication).

Within MADIBA, a subsection of the Arabidopsis Characteristics module was developed to implement DRASTIC. A copy of DRASTIC was kindly provided by Dr Gary Lyon of the Scottish Crop Research Institute (SCRI). Using this database, it was possible to retrieve the regulation (either up- or down-regulated) for each gene in the submitted cluster, to various conditions. These data were then compiled into a table to show the regulation of the genes in the cluster and possibly reveal common annotations, where a green block indicates up-regulation and a red block indicated down-regulation.

2.6.6.2 PCA Experiment Comparer

Since the DRASTIC database is relatively small, due to the data being manually curated, the dataset may not reveal all patterns. In addition, information for only a few of the genes present in *A. thaliana* are contained in the database, which itself is not updated very regularly. Thus, an alternative approach was to compare the expression profile of the genes in the cluster to the expression profiles of previously acquired data. The plant's response to a particular treatment could then be tentatively transferred to the expression of the genes in the cluster. For example, if the expression profile from an ET treatment experiment was most similar to a JA treatment, it could be inferred that both these treatments elicit a similar gene response.

To this end, a large set of microarray experiments was obtained from NASCArrays (Nottingham Arabidopsis Stock Centre Arrays) (Craigon *et al.*, 2004; 20). NASCArrays is a data repository which stores information on Affymetrix experiments on *A. thaliana*, including the expression values for each slide as well as sample preparation and experiment information. The \log_2 -ratios for each experiment in NASCArrays were stored within the MADIBA database. A \log_2 -ratio means that the ratio of the expression level in the test case to the expression level in the corresponding control case was calculated and then \log_2 transformed. If multiple replicate slides were used, the median of all the replicate values was taken. A problem arose in that the slide names in the NASCArrays dataset do not have a standardised naming scheme, making it extremely difficult to parse them. As a result, the slides that did not explicitly indicate which were the control and test cases had to be discarded. Unfortunately this resulted in a lot of experiments being discarded. Currently in the "normalised NASCArrays database" there is data present for 301 experiments. In the original data file (the NASCArrays "supercluster" file, downloaded on 22 February 2008) there was information on 2906 slides. Assuming that there were three replicate slides per experiment, this means that over two-thirds of the slides were discarded, simply because the control and

experimental slides could not be easily distinguished. In addition, any genes that were not represented in all the experiments (i.e. had missing data) were removed. This ensures that the data for each gene is consistent.

With the normalised NASCArrays database, it was possible to obtain the expression profile for a set of genes, over a number of different experiments. However, even when reducing the dataset to only the genes in the cluster, this is still an extremely large amount of data that would be difficult to interpret. One approach to the problem was to reduce the dimensionality of the data such that only the vital information is retained and the redundant information is discarded. In addition to making the data simpler to handle, if the number of dimensions were to be reduced to two or three dimensions, it would be possible to visualise the data, making it easier for a researcher to identify patterns in the data.

Dimension reduction is possible as, in general, not all the measured variables are “important” for understanding the underlying phenomena of interest (Fodor, 2002). The complex, high dimensional data can be governed by only a few simple variables, often called the “hidden causes” or “latent variables” (Carreira-Perpiñán, 1997). The remaining “unimportant” data may be discarded as many of the variables will be correlated with each other (by some linear combinations or other functional dependence), thus making them redundant (Carreira-Perpiñán, 1997). Therefore in many situations it should be possible to discard this unneeded information and produce a more economical representation of the data (Carreira-Perpiñán, 1997).

One common and relatively simple dimension reduction technique is Principal Component Analysis (PCA). PCA reduces the data’s dimensionality by finding orthogonal linear combinations (the Principal Components – PCs) of the original variables with the largest variance. The first PC is the linear combination with the largest variance; the second PC is the linear combination with the second largest variance and is orthogonal (perpendicular) to the first PC; and so on (Fodor, 2002). Thus PCA aims to identify the dimensions with the largest variances that are the most important (most “principal”) (Shlens, 2005). There are as many PCs as the number of original variables, although for most datasets, the first few PCs will explain most of the variance, so the rest can be disregarded with minimal loss of information (Fodor, 2002). Usually the first two or three PCs are taken so that the data can be visualised. Essentially, PCA uses these linear combinations of vectors to transform, or rotate, the original vector space into a new vector space such that the axes of the new coordinate system are

oriented along the directions of greatest variability (Wold *et al.*, 1987; Fodor, 2002). Mathematically, PCA can be defined as $X = TP^T$, where P^T is a projection matrix, which projects X down to a k -dimensional space, and gives the object coordinates in this plane, T . The columns in T , t_a , are called the score vectors and the rows in P^T , p_a , are called loading vectors, with the vectors t_a and p_a being orthogonal (Wold *et al.*, 1987).

PCA is not a very robust technique, as the covariance matrix for the PCA is derived independently each time, so this could result in the same data generating different plots. Thus in any multivariate statistics approach, particularly in applications such as PCA and neural networks, validation is necessary to determine how well the model fits the data. Statistics such as the R^2 statistic can be used for this purpose. In addition, cross-validation is needed to discover how well the model will handle new data, and to avoid over-fitting of the data. This can be calculated using the Q^2 statistic, which is a validation criteria obtained from an estimate of R^2 from leave-one-out cross-validation (LOOCV) (Hawkins *et al.*, 2003), and is interpreted as the amount of variance that can be represented by the PCA model (Stacklies *et al.*, 2007). A poor (low) Q^2 indicates that the model describes noise and that the model may not be related to the true data structure. Thus, the better (higher) the value of Q^2 (to a maximum of 1), the better the model prediction (Stacklies *et al.*, 2007).

In PCA Experiment Comparer, the expression values (\log_2 -ratios) of the genes in the cluster are required. Once the set of genes with the associated \log_2 -ratios is submitted, the \log_2 -ratios for each gene in the submitted cluster, across all the experiments, are retrieved from the normalised NASCArrays database. These values are combined together and the PCA is performed on this reduced dataset. Once the PCA is performed, the data can be inspected by plotting the top two or three PCs against each other. To determine which of the other experiments closest matches the submitted data, a distance measure is used to estimate how “far” any one of the experiments in the experiment dataset is from the submitted data. A Euclidian distance measure was used although numerous other measures of similarity are available, such as the Pearson correlation and the Spearman rank correlation (Yona *et al.*, 2006). The closest 10 experiments are reported. For each slide, there is a link to the NASCArrays site for additional information on the experiment. This process is illustrated in Figure 2.11.

The PCA calculations were performed using the `pca` function in the `pcaMethods` library (21; Stacklies *et al.*, 2007), a Bioconductor package. The calculation is done by a singular

value decomposition (SVD) of the data matrix, and not by using eigenvector decomposition on the covariance matrix, as this is generally the preferred method for numerical accuracy (Vos, 2005). The actual method used in `pcaMethods` is an implementation of the `SVDimpute` algorithm, which is a variation of a singular value decomposition of the data matrix that allows for missing values (Troyanskaya *et al.*, 2001). In all the visualisations, the loadings of the eigenvectors were used, so that the correlation between the variables (the genes) can be observed. For the 3D plots, the `scatterplot3d` library (25) was used. The data were all initially mean centred.

The Q^2 statistic for the PCA model was implemented by using the `Q2` function, also from the `pcaMethods` library. The data were divided into five groups, and the entire cross-validation was repeated three times. The results of this analysis were visualised as a box plot (also known as a box-and-whisker plot), using the built-in `boxplot` function in R.

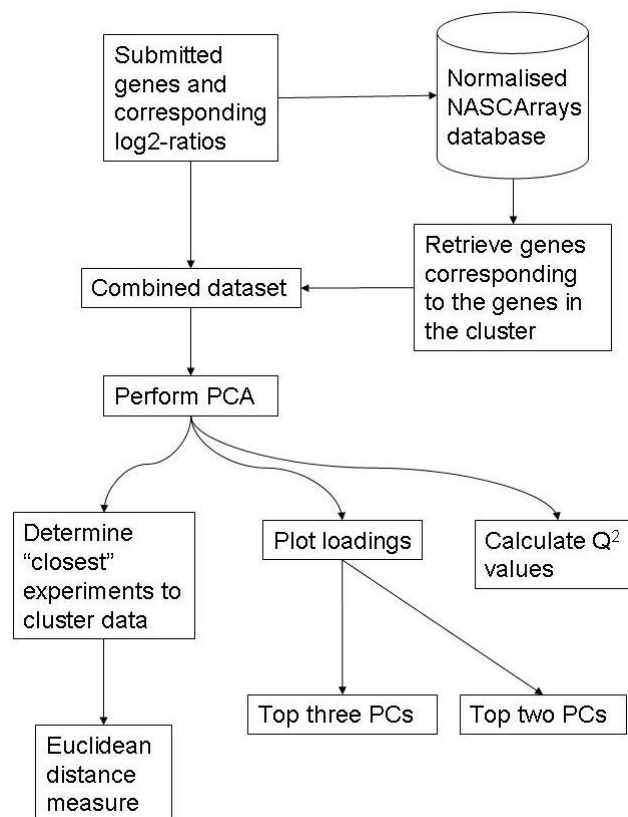


Figure 2.11: Schematic dataflow illustrating the PCA Experiment Comparer. A set of genes and their corresponding \log_2 -ratios are submitted by the user. The expression ratios for these same genes across all the various experiments are retrieved from the normalised NASCArrays database, and combined with the user submitted data. The PCA is performed on this dataset, and the loadings of the top two and three PCs are plotted. At the same time, the “closest” experiments are calculated using a Euclidian distance measure and the Q^2 values for the PCs are calculated.

2.6.7 Output

Results from MADIBA can be exported in either plain text (*.txt*) or PDF formats. The user selects the required set of results which is then generated for immediate download. The requested information is obtained by connecting to the relevant analysis script using the Python `urllib` module, and then screen-scraping the page using Beautiful Soup (9) to obtain the relevant information. The text files are created using Python's standard file object `write` method, and PDF output files are created using ReportLab 2.0 (6).

2.6.8 Contact form

A contact form is also provided for the user to send comments or problems back to the administrator. This form was written in PHP and uses the standard `mail` function, and sends the information contained in the form – the person's name, email address, subject and comment. In addition, the form also contains the referring page (that is, the page which the user came from) to assist in identifying any problems that may have occurred on that page.

2.7 Data maintenance

In order to ensure that the server does not become cluttered, it is necessary to remove the old results. As mentioned previously, a user's data are stored on the server for one week. The maintenance task is performed by setting a Python script to run in the server's crontab file. This ensures that the task will be performed at a consistent interval (currently every Sunday at midnight). The Python script that is run utilises a number of functions from the standard `os` module, including `getmtime`, to obtain the last time a path was modified, and the `remove` function to remove files. The `rmtree` function from the `shutil` module is used to remove directories. In addition, the standard `time` module is required to determine if the time given by `getmtime` and the current time are at least one week apart. If so, the files in the following directories are removed:

- `/usr/local/MADIBA/temp/<uid>`
- `/usr/local/apache2/htdocs/MADIBA/temp/<uid>`
- `/usr/local/apache2/conf/gbrowse.conf/MADIBA/`

In addition, as the organisms' genome annotations are revised, it is important to update the data within MADIBA, and this can be done in a semi-automated manner using pre-built Python scripts.

2.8 Documentation

From the main page, it is possible to access a tutorial, which explains what MADIBA does as well as provide and explain the outputs from a typical analysis. In addition, a basic help page can be accessed which contains the solutions to some common problems, such as gene identifier formats and errors that the user may encounter.

Also, for each organism, the gene clusters that were used in the original study are accessible (Law *et al.*, 2008). This includes data from a *Plasmodium falciparum* life cycle expression analysis (Le Roch *et al.*, 2003); an investigation of hypothetical genes in *P. falciparum* that are thought to be involved in the sexual development of the parasite (Young *et al.*, 2005); an *Arabidopsis thaliana* salt stress experiment (Ma *et al.*, 2006); expression of rice cells in response to flagellin (Fujiwara *et al.*, 2004); and a rice transcriptome study on cold, drought, salinity and abscisic acid treatments (Rabbani *et al.*, 2003). The full set of clusters from each of the above experiments is provided, and a brief summary of some of the results can also be obtained. A demo set of genes can be accessed on the submission page, so that the user does not have to go through the above system to obtain a set of test genes.

2.9 Conclusion

MADIBA currently implements five analysis modules – a Gene Ontology, Metabolic Pathways, Chromosomal Localisation, Transcription Regulation, and Organism Specific analysis modules. Table 2.1 summarises the tools that were utilised in each of the analysis modules. Each organism in MADIBA (currently *Plasmodium falciparum*, *Oryza sativa*, *Arabidopsis thaliana* and *Pectobacterium atrosepticum*) has its own PostgreSQL database containing sequence and annotation information. Some data are pre-calculated, such as BLAST searches against a related organism, to reduce the time required for computations. A user can submit a cluster of genes to MADIBA either as a set of nucleotide sequences in FASTA format, or as a list of gene identifiers. If sequences are submitted, a BLASTN search is performed (BLASTX for rice sequences), and the user can select which of the hits to continue the analyses with. This list of genes is submitted to each analysis module as it is accessed.

The MADIBA web interface was written using Python and PHP. A tutorial explaining the functionality of MADIBA, as well as several test datasets for each organism are available for users. MADIBA is accessible at <http://www.bi.up.ac.za/MADIBA/>.

Table 2.1: A summary table of the various tools that were used in each of the analysis modules.

Module	Tool	Function
<i>General</i>	PHP and Python <code>cgi</code> module GD graphics library PostgreSQL PyGreSQL	Web-page generation Generation of pie charts MADIBA Database Python library for interfacing with PostgreSQL database
<i>Data submission</i>	BioPython Python <code>random</code> module	Perform BLAST searches on submitted sequences Generation of unique user identifier
<i>Gene Ontology</i>	Python <code>pickle</code> module NetworkX Graphviz/PyGraphviz R/RPy ZGRViewer	Load the stored GO hierarchy Perform tree traversal of the GO hierarchy Layout and render the graph as an image Calculate hypergeometric statistics Java applet for easier viewing of graph
<i>Metabolic Pathways</i>	Python Imaging Library Beautiful Soup R/RPy	Colour nodes of the metabolic pathways Read additional information for the KEGG site Calculate statistics
<i>Chromosomal Localisation</i>	GD graphics library wz_tooltip gbrowse	Draw the histogram and chromosomal representation Mouseover effect Genome browser
<i>Transcription Regulation</i>	RSAT – oligo-analysis and dyad-analysis TRANSFAC – Patch and Match	Tools to identify over-represented motifs in the upstream regions. Accessed using Python
<i>Organism Specific</i>	BLAST	Calculate orthologues
<i>Output</i>	Beautiful Soup ReportLab	Generate pdf or txt files of selected analyses
<i>Contact form</i>	PHP <code>mail</code> function	Allow user to send an email to the administrator from the MADIBA site
<i>Data maintainance</i>	Python	Various functions to delete old data

Chapter 3 – Application to Biological Data

3.1 Introduction

In this chapter, for each of the implemented organisms, the application of MADIBA to biological data will be demonstrated. Firstly, the importance of studying the malaria parasite *Plasmodium falciparum* will be discussed, followed by an application of a *P. falciparum* dataset in MADIBA only as a short proof of concept. Following this, an extended discussion of *Arabidopsis thaliana* and rice is provided, with a strong focus on plant defences. Finally, *Pectobacterium atrosepticum* will be discussed, concentrating on quorum sensing. The data for all these analyses can be found in the appendices, available online at <http://www.bi.up.ac.za/MADIBA/doc/appendix/>.

3.2 Application to *Plasmodium falciparum*

3.2.1 *Plasmodium falciparum* introduction

Malaria is one of the most significant tropical parasitic diseases and one of the top three killers among communicable diseases. Malaria is a particularly devastating disease, where it is estimated that each year, approximately 2.2 billion people are exposed to the threat of malaria (Snow *et al.*, 2005), resulting in a conservative estimate of 300-600 million clinical cases attributable to the disease (Sachs and Malaney, 2002; Snow *et al.*, 2005). Approximately 70% of these cases are concentrated in Africa (Sachs and Malaney, 2002; Snow *et al.*, 2005). Of these 300-600 million cases, malaria will kill more than one million people a year, and the figure is possibly closer to three million when the role of malaria in death related to other diseases is included (Sachs and Malaney, 2002; Snow *et al.*, 2005). In areas of stable endemic transmission, the cause of death of about 25% cases involving children aged 0 to 4 has been attributed directly to malaria (Sachs and Malaney, 2002). This means the every 40 seconds a child dies of malaria, resulting in a daily loss of more than 2000 young lives worldwide (Sachs and Malaney, 2002). Malaria is so destructive that it has been suggested that certain genetic polymorphisms, such as sickle cell trait, were selected for due to their protective effect against malaria. However, this can be fatal if the same allele is inherited from both parents. The implications are that the chance of death from malaria was so high that it justified introducing a potentially fatal mutation into the gene pool (Sachs and Malaney, 2002).

In general, where malaria prospers the most, human societies have prospered least. Comparisons of malaria maps shows that the disease has been geographically restricted to the tropical and subtropical zones, which closely frames the poorest areas of the world (Sachs and Malaney, 2002; Snow *et al.*, 2005). A comparison of the average GDP (adjusted so that there is equal purchasing power) in malarious and non-malarious countries in 1995 showed a greater than five-fold difference in GDP. This poverty-malarial infections correlation can be explained in several possible ways, with the most probable being that the causality runs both ways (Sachs and Malaney, 2002). By personal expenditure on preventative methods such as bed nets or insecticides, and with government funding on control programs, a decrease in malaria transmission may be obtained. Indeed, malaria has been essentially eliminated in wealthier countries, such as the United States, Italy and Greece, as a result of socioeconomic development and intensive anti-malarial interventions (Sachs and Malaney, 2002). However, economic development is not enough. Even wealthy countries such as Oman and the United Arab Emirates have not been able to eliminate the disease (Sachs and Malaney, 2002). Causation in the other direction, from malaria to poverty, is indicative of the economic burden of the disease. The cost of malaria is often divided into two categories: private and non-private medical costs. Private medical costs refer to the personal expenditures on the prevention, diagnosis, treatment of the disease, such as bed nets, doctor's fees and the cost of anti-malarial drugs. Non-private medical care refers to public expenditures by a government on factors such as health facilities, education and research (Sachs and Malaney, 2002). Besides these expenditures, there is also the loss of income as a result of mortality (foregone incomes). In addition to these economic costs, there are other effects that malaria has on the population. This includes effects through changes in household behaviour in response to the disease, such as schooling, demography, migration and saving, as well as the macroeconomic costs that arise in response to the disease and include the impact on trade, tourism and foreign investment (Sachs and Malaney, 2002).

Thus malaria is one of the most devastating diseases, particularly in Africa, so it is critical to understand how its causative agent, *Plasmodium falciparum*, functions. In addition, *P. falciparum* is related to plants as the apicoplast (apicomplexan plastid) is reminiscent of the chloroplast (Marechal and Cesbron-Delauw, 2001; Ralph *et al.*, 2001).

3.2.2 MADIBA *Plasmodium falciparum* data analysis

The application of MADIBA to *P. falciparum* data was not the focus of this dissertation, but is presented here primarily for proof of concept purposes. The results of an oligonucleotide array profiling the expression of human and mosquito stages of the malaria parasite's lifecycle (Le Roch *et al.*, 2003) were chosen to demonstrate the functionalities of MADIBA. The experiment performed by Le Roch *et al.* was aimed primarily at determining the patterns in the expression of genes in the *P. falciparum* genome during the different phases in its lifecycle. After a robust *k*-means clustering was performed by the group, 15 clusters were proposed, and these clusters were then subsequently analysed in this study with MADIBA.

The Gene Ontology module automatically allocated annotations to the gene clusters with terms including immune evasion, in cluster 1, and cell invasion in cluster 15. This makes sense as the genes in cluster 1 were found to be expressed during the sporozoite stage of the *P. falciparum* lifecycle, where the parasite attempts to infect the host by invading the liver cells. Similarly, cluster 15 contained genes that were highly transcribed in the schizont stage, where the parasite leaves the infected red blood cells to invade more red blood cells (Le Roch *et al.*, 2003). In addition, the genes in cluster 6 were correctly identified as being involved in hexose metabolism. The genes in this cluster were expressed during the trophozoite stage, where the parasite is in its activated, feeding phase (Le Roch *et al.*, 2003). The Metabolic Pathways analysis module successfully showed that in this cluster, six of the nine enzymes in the glycolysis pathway were found, with a *p*-value of 0.04, as calculated by using Fisher's exact test (Figure 3.1). This result is further supported by the indication that all the enzymes in the pathway were identified by all three annotation sources, as indicated by the yellow boxes, and by using the GO analysis, it was shown that the anaerobic glycolysis term had a highly significant *p*-value (Figure 3.2 and Table 3.1). Using the module specific for *P. falciparum* characteristics allowed the identification of genes in cluster 3 as interesting drug or vaccine targets, such as PF10_0303, an ookinete surface antigen. Cluster 3 is noted as having genes that are highly expressed during the gametocyte stage, so targeting the parasite during this stage, before the sexual cycle and large scale mitosis begins would be valuable in the treatment of malaria.

Young *et al.* (2005) performed an analysis on the transcriptome of *P. falciparum* in an attempt to identify genes that are likely to be involved in the sexual development of the parasite. Young *et al.* used an algorithm called ontology-based pattern identification (OPI) on the data,

which grouped a set of 246 genes. These genes were found to have expression patterns specific for the gametocyte lifecycle stage, and so are most likely involved in the sexual development process. Applying RSAT from the Transcriptional Regulation module of MADIBA on this data set identified several overrepresented motifs in the upstream regions of the genes. In particular, oligo-analysis identified the motif GATGAA, which had an expected occurrence of 96.6 based on the background model, but occurred 228 times (e-value for occurrence 10^{-26}). Similarly with dyad analysis, the motif ATCN(7)TCA was found to occur 154 times, as opposed to its expected occurrence of 41.3 (e-value for occurrence 3.5×10^{-32}). While these patterns did not match those found by the authors, it is possible that these motifs may still be relevant, particularly since relatively little information regarding the transcription regulation of genes involved in pathogenesis and development in *P. falciparum* is known (Young *et al.*, 2005)

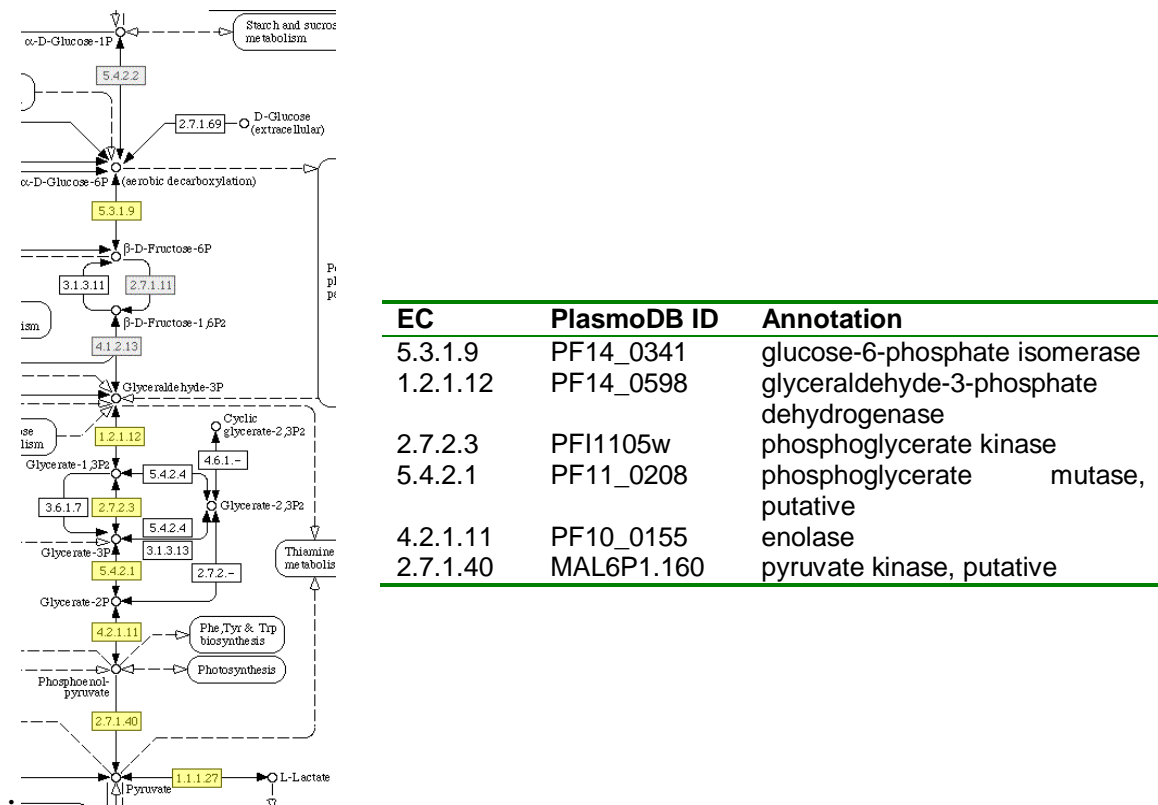


Figure 3.1: Analysis of cluster 6 of the *Plasmodium falciparum* data (Le Roch *et al.*, 2003) revealed that it was noticeably involved in glycolysis. In the KEGG map for glycolysis (left), it could be seen that almost all of the enzymes involved are present in the cluster. Additionally, all of the enzymes were annotated by all the three annotation sources – the curated annotation from the original data source (PlasmODB, TIGR or TAIR); the semi-automatic KEGG annotation and the automatic PRIAM annotation, as indicated by the yellow boxes. The table on the right lists the enzymes that were found in the glycolysis pathway and the gene from the cluster that encodes it.

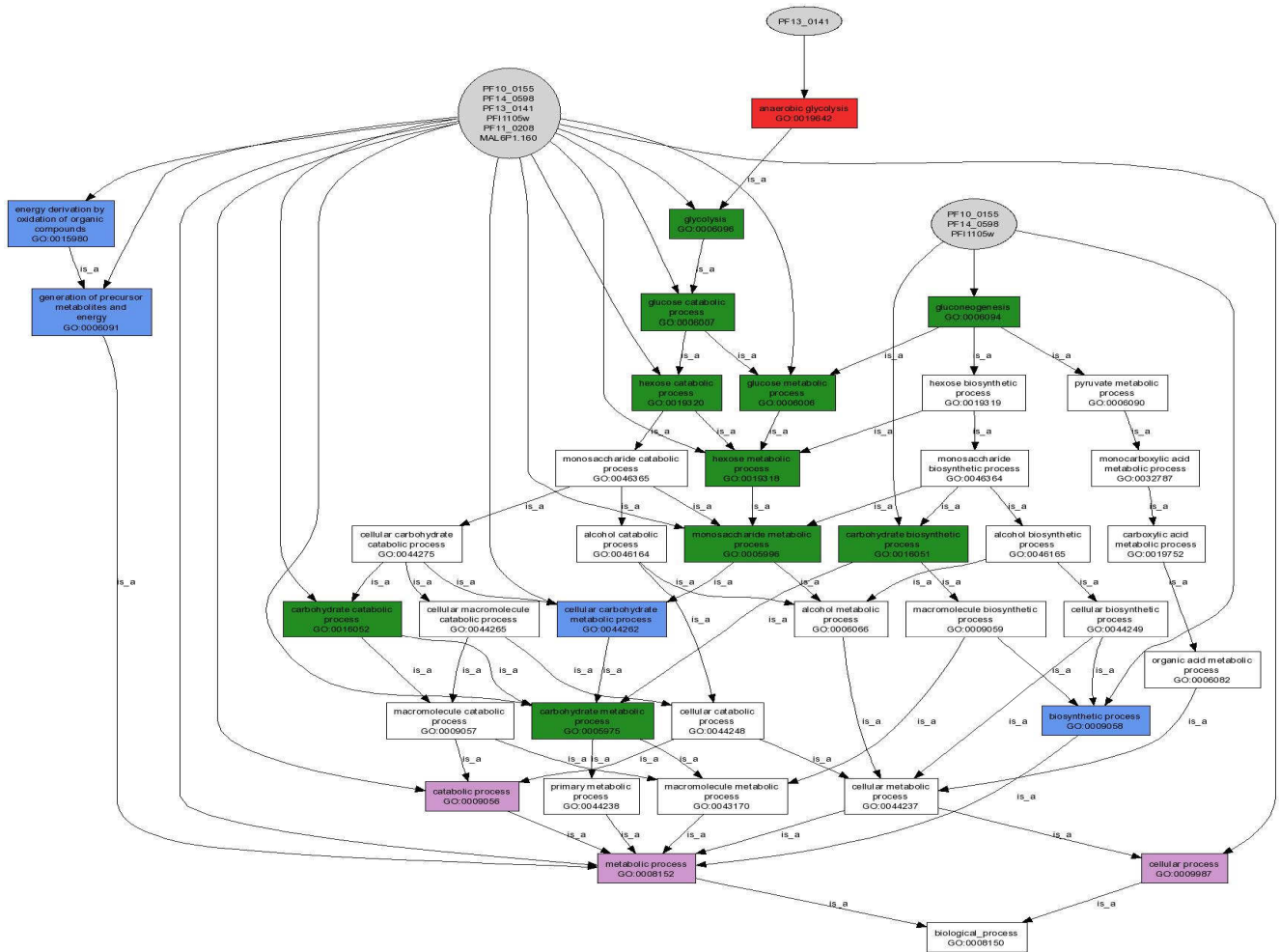


Figure 3.2: Results from the Gene Ontology module. An analysis of the biological process ontology of the cluster 6 of the *P. falciparum* data (Le Roch *et al.*, 2003) revealed that anaerobic glycolysis was the most significant term. The DAG was reduced to show only the terms that are most relevant to glucose metabolism. The grey ellipses contain the genes that are annotated to the connected GO term and the colour of the GO terms indicates different levels of significance, as indicated by the legend (bottom left).

Table 3.1: A portion of the table of *p*-values that accompanied the DAG in Figure 3.2, showing the *p*-value calculated from the hypergeometric distribution, along with the Holm and FDR multiple hypothesis corrections.

Term	Definition	Corrected p-value (FDR)	Corrected p-value (Holm)	Uncorrected p-value
GO:0019642	anaerobic glycolysis	0	0	0
GO:0006096	glycolysis	3.6675×10^{-6}	1.0971×10^{-4}	7.2180×10^{-7}
GO:0005996	monosaccharide metabolic process	5.3464×10^{-6}	1.6747×10^{-4}	1.1091×10^{-6}
GO:0019318	hexose metabolic process	5.3464×10^{-6}	1.6747×10^{-4}	1.1091×10^{-6}
GO:0006094	gluconeogenesis	6.7072×10^{-6}	2.1795×10^{-4}	1.4627×10^{-6}
GO:0016051	carbohydrate biosynthetic process	6.7072×10^{-6}	2.1795×10^{-4}	1.4627×10^{-6}

3.2.3 Conclusion

Analysis of clusters of expression data from *P. falciparum* using MADIBA showed that it is possible to easily identify what the biological implications of a co-expressed set of gene are. By simply submitting a set of gene names, it was possible to retrieve a wealth of information relating to those genes. The varied tools provide access to a wide range of analyses that will allow a researcher to derive some conclusions from the data.

3.3 Application to *Arabidopsis thaliana*

Plant analyses are useful particularly for gaining insights into improving crops in both developed and developing countries. *Arabidopsis thaliana*, which formed a major part of this study, will be discussed as a model species in this section, followed by a discussion on plant defences. The primary aim of this analysis was to determine which defence signalling pathways are activated upon infection by *Ralstonia solanacearum* (bacterial wilt). The current hypothesis suggested that it is the salicylic acid (SA) pathway, so it was hoped that MADIBA can be used to replicate this result.

3.3.1 *Arabidopsis thaliana* as a model species

Over twenty years ago, plant biologists began to search for a model organism that could be used for a detailed analysis of plant genetics and molecular biology. Plants that could effectively be regenerated in culture, such as petunia and tomato, were logical candidates, but attention gradually shifted towards *Arabidopsis thaliana*, a small plant from the mustard family (Meinke *et al.*, 1998). This shift gained momentum with the release of a detailed genetic map in the early 1980s, followed by transformation protocols and the demonstration that *A. thaliana* had a small enough genome (about 120Mb) to allow detailed molecular analysis (Meinke *et al.*, 1998). *A. thaliana* is a member of the Cruciferae (Brassicaceae) family with a broad natural distribution throughout Europe, Asia and North America. Various different ecotypes are available for molecular and genetic studies, but the Columbia and Landsberg ecotypes are the accepted standards (Meinke *et al.*, 1998).

A. thaliana has many advantages for use in genetic analysis, the most important being its short generation time, with the plant's entire life cycle, from seed germination to maturation of the first seed, taking approximately 6 weeks (Meinke *et al.*, 1998). In addition, *A. thaliana* has a relatively small nuclear genome, a large number of offspring and is small in size, with plants reaching only 2-10cm in diameter and up to 20cm in height, depending on growth

conditions. It is possible for the plants to be grown in Petri plates as well as in pots (Meinke *et al.*, 1998). The ability to save the genetic stock as seeds also minimises the effort required for storage over long periods of time.

The complete sequencing of a flowering plant's nuclear genome would allow a comparison of the differences and similarities between plants and other eukaryotes, at a genetic level. Furthermore, it would provide a foundation for a detailed characterisation of plant genes, thus allowing for further studies into plant development and environmental responses (The Arabidopsis Genome Initiative, 2000). As a result, the complete sequencing of the *A. thaliana* genome began in 1996, and was largely completed in 2000 by an international collaboration called the Arabidopsis Genome Initiative (AGI) (Meinke *et al.*, 1998; The Arabidopsis Genome Initiative, 2000). The 125MB genome is organized into five chromosomes and an estimated 26 000 genes. At the time, about 70% of the genes were classified according to sequence similarity to proteins of known function in other organisms, with only 9% of the genes characterised experimentally, and approximately 30% of the predicted gene products could not be assigned to functional categories (The Arabidopsis Genome Initiative, 2000). Since then, although the number of genes that have been manually annotated has increased, the number of genes with no annotation has not changed much (Clare *et al.*, 2006). As of January 2006, 37% of the genes had no annotation or had an annotation to the GO term "molecular function unknown" (Clare *et al.*, 2006).

With the *A. thaliana* genome sequence available, it becomes possible to exploit this information to learn more about the plant's biology. For example, by studying the *A. thaliana* genome sequence, it was determined that the plant had undergone several rounds of complete genome duplications in the past, as well as determine the fates of duplicated genes (Seoighe and Gehring, 2004). In addition, as the first complete genome sequence of a plant, the sequence provides a basis for a more detailed comparison of conserved processes in all eukaryotes, identifying the set of plant-specific gene functions and establishing a method for identifying potential genes for crop improvements (The Arabidopsis Genome Initiative, 2000).

A. thaliana is the most widely-studied plant today and so, it serves as a model organism for the understanding of the complex processes involved in plant growth and development. With the large quantity and diversity of information being generated, The Arabidopsis Information Resource (TAIR) was developed (Rhee *et al.*, 2003) (<http://arabidopsis.org>). An on-line

resource, TAIR contains fully-annotated genes and gene products, using a controlled vocabulary, as well as a data retrieval system and analysis and visualisation tools (Rhee *et al.*, 2003).

3.3.2 Plant defence responses

3.3.2.1 Overview

Even with advances in the control of plant diseases, the global food supply is still under threat from a multitude of pathogens, with up to 20% of the yield in developed countries lost to disease (Anderson *et al.*, 2005). The impact of crop loss is even more detrimental in developing countries. The use of pesticides and other chemicals provides some protection, but the disadvantages can include adverse environmental effects and the emergence of resistant strains. Furthermore, such chemical controls are often too expensive for use by farmers in developing countries. It is for these reasons that much effort has been invested in understanding the “built-in” defence responses.

Plant defence responses to microbial pathogens have been extensively studied for many years, as a result of crop damage from pathogen attack (McDowell and Woffenden, 2003), with much research gone into understanding a plant’s response to pathogens (Gurr and Rushton, 2005). The ultimate goal is to produce crops with increased and durable resistance to a variety of diseases (Murray *et al.*, 2002). Plants have an effective collection of inducible defence responses including genetically programmed suicide of infected cells, as well as tissue reinforcement and the expression of genes involved in defence (McDowell and Woffenden, 2003).

No real attempt to “genetically dissect” these responses was made until the advent of the well-characterised *A. thaliana* model system. This was largely due to the intractability of genetic analysis of other hosts, caused partly by long generation times and large, polyploid or repetitive genomes. Moreover, *A. thaliana* is susceptible to a wide range of pathogens, including bacterial, fungal and viral pathogens (Glazebrook *et al.*, 1997). Before the genome sequence was known, mutants of *A. thaliana* that are defective in almost every aspect of plant growth, development and reproduction were generated using forward genetics. This approach entails random EMS mutagenesis, selecting a phenotype, genetically characterising the mutant, and finally cloning the mutant gene by map-based cloning. However, with the genome sequence known, it has also been possible to develop mutants using reverse genetics

approaches, such as RNAi and VIGS. The existence of a large collection of mutants with defects in defence-related signalling pathways made it possible to use the mutants to determine which pathways are controlling an observed response, as well as place and order genes within signal transduction networks (Glazebrook *et al.*, 1997; Glazebrook, 2001). Analysis of these mutants generated in *A. thaliana* is beginning to give researchers an idea as to the organisation of the complex transduction pathways that result in the defence responses that protect plants from pathogen infection (Glazebrook *et al.*, 1997).

In nature, plants are continuously challenged by fungi, bacteria, viruses and nematodes, yet relatively few successfully infect the plant (Gurr and Rushton, 2005). This is due to the various layers of defence strategies that plants possess that include strengthening of structural barriers by lignification of host cell walls, as well as enzymatic and chemical defences that interfere with pathogen metabolism. These defences may be the synthesis of reactive oxygen species (ROS), nitric oxide (NO), or the expression of genes that encode glucanases, chitinases, thionins, defensins and glutathione-*S*-transferases and other pathogenesis related (PR) genes (Thatcher *et al.*, 2005). Of the eleven classes of PR proteins, most have been assigned probable functions. They target the pathogen cell wall (PR-2, -3, -4, -8, -11), pathogen membrane (PR-1, -5), pathogen RNA (PR-10), undefined pathogen proteins (PR-6) or display peroxidase activity (PR-9) (Gurr and Rushton, 2005). These defence responses have to be strictly regulated as activation of defence responses impacts negatively on plant growth, as it is metabolically expensive (Glazebrook, 2005; McDowell and Woffenden, 2003).

In recent years, the focus of studies of the genes controlling expression of defence responses in *A. thaliana* has shifted from the identification of involved genes, to the ordering of these genes within the branches of signal transduction networks. It is now clear that resistance is mediated through at least three genetically distinct pathways (Glazebrook, 2001). These pathways can be distinguished by the downstream signalling molecules: salicylic acid (SA), jasmonic acid (JA) and ethylene (ET). In addition to possessing different signalling molecules, these pathways result in an increase in the expression of different sets of genes. An increase in SA results in the expression of a subset of PR genes including *PR-1*, *PR-2* and *PR-5*, whereas the induction of JA and ET signal transduction pathways induces a different subset of PR genes, including *PDF1.2* and *Thi2.1* (Thatcher *et al.*, 2005).

SA-dependent and JA/ET-dependent responses are used differently against pathogens with different invasive techniques (Murray *et al.*, 2002; Thatcher *et al.*, 2005). Biotrophic pathogens keep their host alive and cause minimal cell damage, in order to extract food from host cells using specialised feeding structures, known as haustoria (Murray *et al.*, 2002; Glazebrook, 2005). These pathogens do not cause host cell death, as they can only live and multiply on another living organism, and include the oomycete *Peronospora parasitica* and fungal pathogen, *Erysiphe orontii* (Glazebrook, 2005). Resistance to biotrophic pathogens tends to rely on SA-dependent responses and defence occurs through programmed cell death in the host (Glazebrook, 2005). Necrotrophic pathogens, however, kill host tissue by producing cell wall degrading enzymes or toxins, leading to host tissue maceration, thus obtaining its energy from the dead host cells. Therefore, necrotrophs would benefit from host cell death, and so defence against necrotrophic pathogens tends to rely on a different defence response, namely JA/ET-dependent signalling (Glazebrook, 2005). Pathogens of this type include *Botrytis cinerea* and *Alternaria brassicicola*. However, some pathogens behave both as biotrophs and necrotrophs, depending on growth conditions or stages in the developmental cycle. Such pathogens are termed hemi-biotrophs and include the bacterium *Pseudomonas syringae* (Glazebrook, 2005) and *Phytophthora infestans*, the oomycete that causes potato blight (Strange and Scott, 2005). Similarly, some bacterial pathogens such as *Pectobacterium atrosepticum* and *Ralstonia solanacearum* are generally considered necrotrophs due to the large amounts of cell wall degrading enzymes that they produce, although it has been suggested that they may also have a biotrophic phase during early infections (Toth and Birch, 2005).

3.3.2.2 Gene-for-gene resistance

The simplest model of plant-pathogen recognition is based on the interaction of pathogen elicitors with plant receptors and the subsequent transduction of this interaction into a defence response. This interaction can be subdivided into non-host resistance and host-specific recognition (Thatcher *et al.*, 2005). Non-host resistance is a type of resistance shown by all members of a plant species to a specific pathogen. It is the less understood resistance mechanism, even though it is the most common form of resistance (Mysore and Ryu, 2004). In host-specific recognition (more commonly known as the “gene-for-gene” model), resistance occurs when the pathogen carries an avirulence (*avr*) gene that corresponds to a particular resistance gene (*R* gene) in the host. Generally, each *R* gene confers resistance only to pathogens carrying the corresponding *avr* gene. Essentially, when corresponding *R* and *avr*

genes are present, the result is disease resistance (an incompatible reaction) and if either is absent, the result is disease (a compatible reaction) (Gurr and Rushton, 2005).

In both non-host resistance and host-specific recognition, reactive oxygen species (ROS) are produced, a hypersensitive response (HR, a programmed cell death) occurs, and lignification takes place (Mysore and Ryu, 2004). However, it is questionable whether these similarities involve the same signal transduction pathways. Even with the similarities, there are significant differences, and it is possible that both non-host resistance and host-specific recognition have completely separate signal transduction pathways that merely have significant cross-talk between the two pathways that converge at a later stage (Mysore and Ryu, 2004).

Avirulent pathogens often trigger the hypersensitive response (HR), a programmed cell death of the plant cells in contact with the pathogen. This is thought to limit biotrophic pathogens access to water and nutrients, and thus limit pathogen growth (Glazebrook, 2005). During the HR, there is a influx of Ca^{2+} and H^+ , and an efflux of K^+ and Cl^- , and these are thought to be signals for the generation of ROS such as the superoxide anion ($\text{O}_2^{\bullet-}$) and hydrogen peroxide (H_2O_2), and reactive nitrogen species such as nitric oxide (NO) (Thatcher *et al.*, 2005). During HR, H_2O_2 not only has a microcidal effect, but also stimulates cell wall lignification, thus strengthening the cell wall against the pathogen. It has been found that NO alone is not able to activate hypersensitive cell death during HR, but only when there is an appropriate balance between NO and ROS, in particular H_2O_2 (Thatcher *et al.*, 2005).

Several signalling molecules are involved in the downstream responses, and these include salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) (Thatcher *et al.*, 2005). Signal transduction networks are proving to be tightly regulated and accommodate considerable cross talk between the SA-dependent and JA/ET-dependent responses in synergistic or antagonistic fashions (Thatcher *et al.*, 2005). Based on current knowledge of signalling pathways, the signal transduction networks can be divided into five categories: SA-dependent resistance, JA-dependent resistance and ET-dependent resistance, systemic acquired resistance (SAR), induced systemic resistance (ISR) (Thatcher *et al.*, 2005). Figure 3.3 illustrates a summary of these pathways.

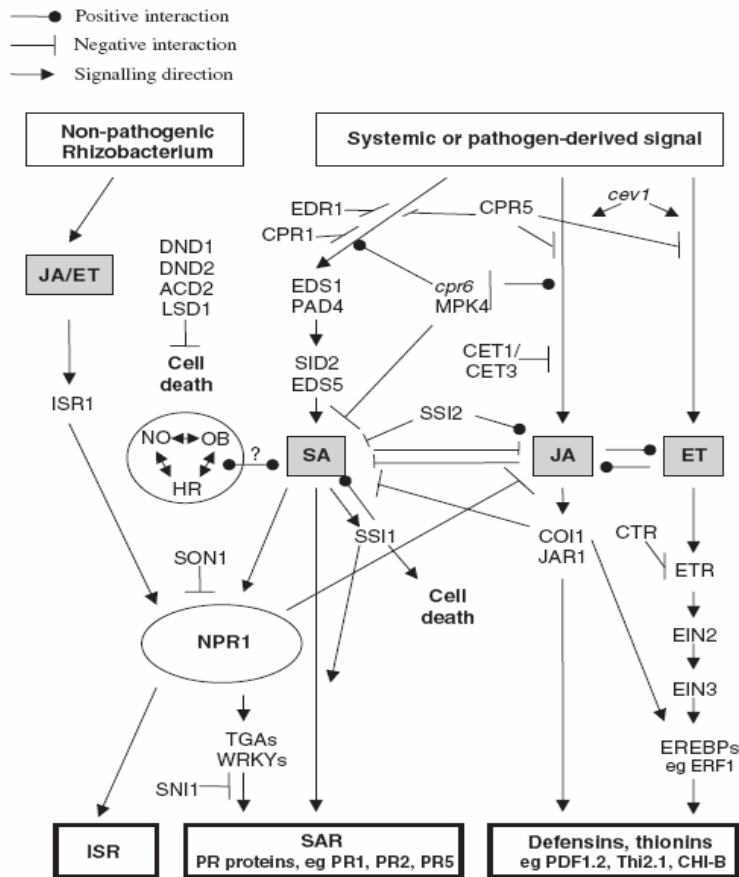


Figure 3.3: Brief summary of plant defence signalling pathways in *Arabidopsis thaliana* (Thatcher *et al.*, 2005). Shown are the primary genes that have been found to be involved in SA-, JA- and ET-dependent signalling pathways, as well as SAR and ISR. Capitals indicate wild-types and italics indicate mutants. ISR: Induced Systemic Resistance, SAR: Systemic Acquired Resistance, OB: Oxidative Burst.

3.3.2.3 Salicylic acid (SA) dependent pathway

Salicylic acid has been shown to play a central role in local plant defence responses, and several *A. thaliana* mutants have been isolated that have defects in SA signalling and have been positioned within the signalling pathway. These include *eds5-1* (*enhanced disease susceptibility*) and *sid* (*salicylic acid induction deficient*) mutants, *sid1* and *sid2*, which are unable to accumulate SA after pathogen infection (Thatcher *et al.*, 2005). Thus, *EDS5* and *SID2* are suggested to function upstream of SA accumulation. *SID2* encodes isochorismate synthase, and in *sid2* mutants, the production of SA is reduced (Glazebrook, 2005). This suggests that the majority of SA is produced from isochorismate, although a small amount may be produced via the phenylalanine pathway (Glazebrook, 2005; Thatcher *et al.*, 2005). Both *eds5* and *sid2* mutants result in reduced *PR-1* gene expression, although *PR-2* and *PR-5*

expression levels are normal (Thatcher *et al.*, 2005). *SID1/EDS5* expression also increases with SA treatment, suggesting a positive feedback mechanism (Thatcher *et al.*, 2005).

Phytoalexin-deficient mutants (*pad1*, *pad2*, *pad3*, *pad4*) have defects in phytoalexin and SA accumulation, and reduced *PR* expression (Thatcher *et al.*, 2005). Phytoalexins are antimicrobial compounds in *A. thaliana*, which include camalexin (Glazebrook, 2005). *PAD3* encodes a camalexin biosynthetic enzyme and so, *pad3* mutants, which produce very little camalexin, are commonly used for pathogen studies (Glazebrook, 2005). *PAD4* encodes a protein similar to EDS1 and is predicted that *PAD4* may lead to the synthesis and degradation of a molecule that is involved in defence signalling (Thatcher *et al.*, 2005). EDS1 and *PAD4* act upstream of SA to promote SA accumulation. *EDS5* expression requires *PAD4* and *EDS1* and so is placed downstream of them (Glazebrook, 2005).

A protein with a key regulatory function in SA signalling is NPR1 (*non-expressor of PR genes 1*), also known as NIM1 (Thatcher *et al.*, 2005; Glazebrook, 2001). NPR1 and SAI1 (*salicylic acid insensitive 1*) both act downstream of SA to promote the expression of *PR-1*, *BGL2* and *PR-5*. *npr1* is impaired in its ability to express *PR* genes in the presence of avirulent pathogens (Thatcher *et al.*, 2005; Glazebrook *et al.*, 2003). When cellular SA levels are low, NPR1 exists in an oligomeric form in the cytoplasm (Eulgem, 2005). When SA levels rise, the oligomers dissociates into NPR1 monomers which move into the nucleus (Glazebrook, 2005; Maleck *et al.*, 2000). *NPR1* encodes an ankyrin-repeat containing protein, a domain that is often involved in protein-protein interactions. It has been found that a subclass of basic region/leucine zipper (bZIP) transcription factors, called TGAs, interact specifically with NPR1 (Thatcher *et al.*, 2005; Glazebrook *et al.*, 2003; Maleck *et al.*, 2000). Furthermore, NPR1 interacts with TGA2, which in turn binds to a SA-responsive promoter element in the *PR-1* gene. Using triple knockout mutants of *TGA2*, *TGA5* and *TGA6*, *PR* gene expression and pathogen resistance induced by SA was blocked and this mimicked mutations in *NPR1* (Thatcher *et al.*, 2005).

With regard to transcription factors involved in SA-mediated defence responses, an important family is the plant-specific WRKYs, which bind to the W-box motif (TTGAC) (Thatcher *et al.*, 2005; Maleck *et al.*, 2000). The *NPR1* promoter contains several W-box motifs and when they are mutated in promoter::reporter fusions, promoter activity is abolished, suggesting WRKY transcription factors are required for *NPR1* expression (Thatcher *et al.*, 2005; Maleck *et al.*, 2000). Additionally, analysis of *PR-1* co-regulated genes induced during SAR shows an

over-representation of W-box or W-box-like motifs in their promoters (Thatcher *et al.*, 2005; Maleck *et al.*, 2000).

Researchers often use NahG plants to dissect SA signalling as the *NahG* transgene blocks SA accumulation by encoding a SA-degrading enzyme, salicylate hydroxylase. This enzyme converts SA to catechol, and thus affects SA-dependent signalling (Thatcher *et al.*, 2005; Glazebrook, 2001). However the assumption that *NahG* phenotypes result from an absence of SA may not be entirely correct (Glazebrook *et al.*, 2003). It has been suggested that catechol, the product of SA degradation increases susceptibility, with *NahG* affecting the expression of many more genes than *sid2*, thus implying that *NahG* may influence signalling pathways in addition to the SA-dependent pathway (Thatcher *et al.*, 2005; Glazebrook *et al.*, 2003; Glazebrook, 2001).

Ordering of genes in the SA-dependent pathway is complex as there are several feedback loops (Glazebrook, 2005). For example, cell death promotes SA production, but SA production promotes cell death (Glazebrook, 2005). In addition, *PAD4* and *EDS1* are required for SA production, but the expression of these genes is enhanced by SA (Glazebrook, 2005).

3.3.2.4 Jasmonic acid (JA) dependent pathway

Jasmonates are produced from the major plant plasma membrane lipid, linolenic acid (Thatcher *et al.*, 2005). Biological roles of JA and some of its biosynthesis intermediates include fruit ripening, fertility and root growth, and responses to wounding, insects, microbial pathogens and abiotic stress (Thatcher *et al.*, 2005). JA and methyl jasmonate (MeJ) induce defence-related genes, such as the defensin *PDF1.2* (*PR-12*) and the thionin *Thi2.1* (*PR-13*) (Glazebrook, 2005; Thatcher *et al.*, 2005). *PDF1.2* and *Thi2.1* gene expression is not induced after SA treatment, although *PDF1.2* also requires ethylene (ET) (Glazebrook, 2005). JA also induces genes that regulate its own synthesis (*DAD1*, *LOX2*, *AOS*, *OPR3*, and *JMT*) (Thatcher *et al.*, 2005).

Several mutants have been isolated that activate or suppress the JA-mediated signalling pathways. These include *fad* (*fatty acid deficient*), *dad1* (*defective anther dehiscence 1*), *dde1* (*delayed dehiscence 1*) and *opr3* (*12-oxophytodienoic acid reductase 3*) mutants (Thatcher *et al.*, 2005). All of the mutants are defective in the enzymes that they encode, and these enzymes are required for JA biosynthesis (Thatcher *et al.*, 2005).

All known activities of JA in *A. thaliana* require the function of *COI1* (Glazebrook, 2005). The *coi1* (*coronatine insensitive*) mutant is insensitive to MeJ and coronatine (a bacterial toxin that mimics the action of MeJ) and encodes a LRR-containing F-box protein that may recruit repressors and target them for removal by ubiquitination (Glazebrook, 2005; Thatcher *et al.*, 2005). The *coi1* mutant blocks JA signalling and increases susceptibility to necrotrophic fungal pathogens (Thatcher *et al.*, 2005).

A mutation in the MAP kinase *mpk4* blocks expression of *PDF1.2* and *Thi2.1* in response to MeJ and expresses SA-mediated defences. Additionally, impairment of JA signalling is independent of elevated SA levels. It is thought that *MPK4* may impact SA signalling upstream of SA by affecting the balance between SA-dependent and JA-dependent signalling (Thatcher *et al.*, 2005).

Mutants with activated JA-mediated defence pathways include *cev1* and *cet* mutants (Thatcher *et al.*, 2005). *cev1* expresses *PDF1.2*, *Thi2.1* and *CHI-B* and has increased levels of JA and ET. The *cet* (*constitutive expressor of thionin*) mutants exhibit activated JA-dependent gene expression, increased levels of JA and show spontaneous lesion formation (Thatcher *et al.*, 2005). It is thought that the lesion formations in the various *cet* mutants may be a result of cell death pathways that are independent of SA (Thatcher *et al.*, 2005).

3.3.2.5 Ethylene (ET) dependent pathway

Ethylene (ET) is a plant hormone that is involved in plant growth and development and is regulated in response to both biotic and abiotic stresses (Thatcher *et al.*, 2005). Mutants isolated with defects in ET responses include *ein* (*ET insensitive*) and *etr* (*ET resistant*) mutants. Components of the ET signalling pathway include the nuclear-localised transcription factor EIN3, which activates *ERF1* (*ET response factor 1*), a member of plant specific ethylene-responsive element binding protein (EREBP) family, which in turn binds to GCC-box promoter elements to activate defence genes such as *PDF1.2* and *CHI-B* (Thatcher *et al.*, 2005). The GCC-box motif is associated with ET and pathogen-induced gene expression and found in many pathogen-responsive genes. *ERF1* expression can be induced by both ET and JA, although intact signalling components from both pathways are simultaneously required for expression, since mutations that block either pathway prevent *ERF1* expression (Thatcher *et al.*, 2005). Microarray experiments have shown that *ERF1* regulates the expression and integrates the signals of both ET and JA responsive genes indicating that *ERF1* acts at a downstream intersection between the ET and JA signalling pathways (Glazebrook, 2005;

Thatcher *et al.*, 2005). *ERF1* expression requires ET and JA as well as *COI1* and *EIN2* (Glazebrook, 2005).

3.3.2.6 Systemic Acquired Resistance (SAR)

Following pathogen attack, the early defence signalling events are often amplified through the generation of secondary signalling molecules, such as SA, JA, and ET (Thatcher *et al.*, 2005). These may activate defences both locally at the infection site and systemically in non-infected tissues. The activation of the HR triggers a systemic response known as systemic acquired resistance (SAR). SAR is associated with an increase in SA locally, and then accumulated systemically throughout the plant. This leads to the expression of defence genes, such as *PRs* in distant uninfected tissues (Thatcher *et al.*, 2005). These activated defence genes result in the susceptible, uninfected tissue gaining resistance to the pathogen (Thatcher *et al.*, 2005). SAR results in a long-lasting, systemic resistance to subsequent infection by other pathogens, including bacteria, fungi and viruses (Thatcher *et al.*, 2005).

Very few mutants unique in SAR have been identified. Two such mutants include *dir1-1* (*defective in induced resistance*) and *cdr1* (*constitutive disease resistance*). *DIR1* encodes an apoplastic lipid transfer protein that leads to the production or transduction of a mobile signal from locally infected tissue to systemic tissue to induce SAR (Thatcher *et al.*, 2005). *CDR1* encodes an apoplastic aspartic protease and may be involved in the activation of SAR through the generation of a mobile signal (Thatcher *et al.*, 2005).

3.3.2.7 Interactions between pathways

Although SA-dependent and JA/ET-dependent pathways induce different sets of *PR* genes and provide resistance against different pathogens, there is both synergism and antagonism between the pathways (Thatcher *et al.*, 2005; Schenk *et al.*, 2000). The signalling molecules, SA, JA and ET, each result in lesion formation mimicking HR cells death. However, there are differences in lesion formation and these differences may be a result of synergistic or antagonistic effects between the SA, JA and ET signalling pathways (Thatcher *et al.*, 2005).

SA and JA pathways often seem to act antagonistically, with SA shown to have an inhibitory effect on JA biosynthesis (Thatcher *et al.*, 2005). Pathogen-induced SA accumulation suppresses JA formation and JA-responsive gene expression. Conversely, JA is reported to negatively regulate SA signalling (Thatcher *et al.*, 2005). The *mpk4* mutation blocks the JA-inducible expression of *PDF1.2* and causes the constitutive activation of SA-dependent

signalling, suggesting that the block in JA signalling may relieve the suppression of SA signalling (Thatcher *et al.*, 2005). The mechanisms that result in the negative cross-talk between JA and SA signalling are not well understood, and it has been suggested that this is due to cross-talk at multiple points (Glazebrook, 2005).

However, SA and JA also act synergistically to induce defence-associated genes (Schenk *et al.*, 2000). It has been shown that *PDF1.2* expression and NPR1-independent expression of *PR-1* in *ssi1* (*suppressor of SA insensitivity 1*) requires both SA and JA/ET signalling pathways, suggesting that *SSI1* plays a role in regulating the cross-talk between SA and JA/ET signalling pathways (Thatcher *et al.*, 2005). The JA and ET pathways often work synergistically. Most genes induced by ET in microarray studies were also induced by MeJ (Maleck *et al.*, 2000).

Generally, it was found that in response to virulent bacterial pathogen attack, SA and JA signalling oppose each other, SA and ET signalling also tend to oppose each other, and JA and ET signalling usually acts together (Thatcher *et al.*, 2005).

3.3.3 MADIBA *Arabidopsis thaliana* data analysis

For this analysis, the expression of *A. thaliana* plants that were treated with *Ralstonia solanacearum* (bacterial wilt) was used. This microarray data was produced by the Molecular Plant-Pathogen Interactions (MPPI) group at the University of Pretoria. *R. solanacearum* is a soil-borne plant pathogen that naturally infects roots and specifically invades the xylem vessels. This bacterium has an unusually wide host range, being able to infect over 200 host species, belonging to more than 50 botanical families (Salanoubat *et al.*, 2002). MADIBA was used to analyse these data to determine which pathways are activated in response to a *R. solanacearum* infection, in both the susceptible and resistant interactions.

To determine the expression of the susceptible interaction, *R. solanacearum* isolate K was used to infect *A. thaliana* ecotype Col-5 (Naidoo, 2008). From this experiment, 133 genes were found to be significantly responsive, with 76 genes being up-regulated and 57 genes down-regulated. The up- and down-regulated genes were analysed separately.

In the analysis of the up-regulated genes, it was found in the Metabolic Pathways module that the fatty acid metabolism, phenylalanine metabolism and stilbene, coumarine and lignin biosynthesis pathways all contained three enzymes each. However, the annotations in these pathways were primarily derived from PRIAM, so this may not be a particularly reliable

result. This is because the PRIAM enzyme predictions are completely automatic, based solely on similarity to known enzymes and have not been tested experimentally. In addition, the enzyme position specific scoring matrices (PSSM) that are used to determine similarity were derived primarily from bacterial enzymes (Claudel-Renard *et al.*, 2003), and so may not be reliable for plants.

In the Gene Ontology module, there were several terms pertaining to glutamate enzyme activity in the molecular function ontology. These included glutamate-5-kinase and glutamate decarboxylase activity (GO:0004349 and GO:0004351, respectively). In the biological process ontology, as expected from a response to bacterial wilt, several terms involving water loss were significant, including cellular response to water deprivation (GO:0042631), response to desiccation (GO:0009269), response to water deprivation (GO:0042631), and hyperosmotic salinity response (GO:0042538). Interestingly the term jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signalling pathway (GO:0009871) appeared, although with a relatively poor *p*-value (*p*-value = 0.0018). The ethylene (ET) term make sense as it has been suggested that ET promotes wilting, as delayed wilting symptoms were observed in an *ein2-1* mutant, which is insensitive to ET signalling (Hirsch *et al.*, 2002). Other terms included fatty acid oxidation (GO:0001561), tyrosine catabolic process (GO:0006572) and glyoxysome organisation and biogenesis (GO:0010111).

In the Transcription Regulation module, oligo-analysis, a tool within RSAT, found that the motifs ACACGT and ATAAAT were over-represented (e-values = 0.012 and 0.2 respectively). Similarly, the dyad-analysis program found that the motifs CACN(2)GTC and ACAN(0)CGT were over-represented (e-values = 0.02 and 0.3 respectively).

When analysing the genes that were down-regulated in the susceptible interaction, it was found in the Metabolic Pathways module that the only significant pathway was the carbon fixation pathway (3 enzymes, *p*-value = 0.0043). This could mean that when a plant is infected by a pathogen, carbon fixation is reduced, with more energy focussed towards fighting the pathogen.

In the Gene Ontology module, the molecular function ontology found some interesting terms including hydrolase activity, specifically catalysing the transmembrane movement of substances (GO:0016820), transmembrane receptor protein serine/threonine kinase activity (GO:0004675) and MAP kinase activity (GO:0004707). The last two terms in particular are

known parts of signal transduction cascades, so it is possible that these genes are down-regulated when the pathogen interferes with these signals to prevent an effective defence response. It has been shown that some pathogens are able to interfere with the host responses. For example, experiments have revealed that *Pseudomonas syringae* pv. *tomato* is able to manipulate the abscisic acid signalling pathway in *A. thaliana*, resulting in a compromised defence response (de Torres-Zabala *et al.*, 2007). Thus it is possible that pathogens are able to manipulate aspects of the host response, such as plant hormone homeostasis and signal cascades, to suppress defence responses (de Torres-Zabala *et al.*, 2007). In the biological process ontology, the signalling aspect was confirmed with the activation of MAPK activity during osmolarity sensing term (GO:0000169). Also, as predicated by the Metabolic Pathways module, the terms carbon utilisation by fixation of carbon dioxide (GO:00015977) and carbon utilization (GO:0015976) were significant. The cellular component ontology identified several significant terms that indicated that most of the genes are localised in the chloroplast including the chloroplast ribulose biphosphate carboxylase complex (GO:0009573), chloroplast thylakoid (GO:0009534, GO:0009535 and GO:0009579) and photosystem I reaction centre (GO:0009538). These terms are most likely related to the carbon fixation terms seen above. Another significant term suggested that the gene products were located extrinsic to the membrane (GO:0019898), again hinting at some possible involvement in signal transduction.

Using oligo-analysis, it was found that the motifs CGTTCA and GGTCCA were over-represented (e-values = 0.48 and 0.52 respectively). Dyad-analysis identified the motifs AATN(20)GGG and CTGN(2)GAC as over-represented (e-values = 0.02 and 0.3 respectively). While all the motifs detected did not match any currently known TFBS, they still nonetheless could be significant.

To determine how the gene expression differed in the resistant interaction, *R. solanacearum* isolate CK was used to infect the *A. thaliana* ecotype Kil-0, and this experiment was performed on a 27000 probe whole genome array (Naidoo, 2008). In the data from the resistant interaction, 76 genes were found to be significantly different, with 53 genes up-regulated and 23 down-regulated. The current hypothesis is that the salicylic acid (SA) signalling pathway is involved in the resistant response in *A. thaliana* to *R. solanacearum* (Deslandes *et al.*, 2003; Noutoshi *et al.*, 2005), so it was hoped that MADIBA could be used to confirm this result.

When analysing the up-regulated genes in MADIBA, only the Gene Ontology module showed significant results. Terms that were found in the molecular function ontology included chitin binding (GO:0008061) and chitinase activity (GO:0004568), as well as alcohol dehydrogenase activity (GO:0004022) and calmodulin binding (GO:0005516). Terms in the biological process ontology corroborated some of these terms, including cellulose and pectin-containing cell wall organization and biogenesis (GO:0009664) and response to chitin (GO:0010200). Some defence and *R. solanacearum* response terms included cellular response to water deprivation (GO:0042631), response to reactive oxygen species (GO:0000302), oxygen and reactive oxygen species metabolic process (GO:0006800), response to oxidative stress (GO:0006979), cell death (GO:0008219), in addition to defence response to bacterium, incompatible interaction (GO:0009816), defence response to bacterium (GO:0042742), and response to bacterium (GO:0009617). Unfortunately, no terms relating to SA signalling were found, although curiously several terms relating to JA and ET signalling were. These included jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signalling pathway (GO:0009871), jasmonic acid and ethylene-dependent systemic resistance (GO:0009861), response to ethylene stimulus (GO:0009723), response to jasmonic acid stimulus (GO:0009753) and response to abscisic acid stimulus (GO:0009737). Some other response terms included embryonic development (GO:0009790), response to high light intensity (GO:0009644) and response to cold (GO:0009409).

When the down-regulated genes were analysed, again only the Gene Ontology module found results. These terms included carbon utilisation (GO:0015976), cellular calcium ion homeostasis (GO:0006874), cellulose and pectin-containing cell wall modification during multidimensional cell growth (GO:0009831), cellulose and pectin-containing cell wall loosening (GO:0009828).

3.3.4 MADIBA defence pathways analysis

Since MADIBA does not currently possess a means to identify genes which have a part in defence signalling, and are thus potentially useful in improving plant defences, a sub-module of the Arabidopsis Characteristics module was considered. Initial thoughts were to derive a plant defence signalling pathway map from literature studies, such as in Figure 3.3 and highlight the genes (nodes) that are present in the pathway, in a similar fashion to the Metabolic Pathways module. However, this approach was abandoned as it was concluded that there is insufficient information to draw an accurate representation. In addition many of the

interactions happen only after infection with certain pathogens and different combinations of signalling components are activated under different conditions. Further, the ordering of the genes in the signalling cascade was primarily performed by using mutant studies, so this means that such networks may not necessarily occur in this manner in a wild-type plant. Thus an alternative approach was required.

3.3.4.1 DRASTIC

The first approach was to investigate the up- or down-regulation of a group of genes in response to some experimental condition. The aim was to identify which conditions an experiment has a similar expression pattern to, and this was accomplished by obtaining the regulation of a set of genes from the experiments contained in the DRASTIC database (11; Button *et al.*, 2006). Using this database, the regulation (either up- or down-regulated) for each gene in the submitted cluster, to various conditions, was obtained and compiled into a table, where a green block indicates up-regulation and a red block indicated down-regulation.

When DRASTIC was applied to the up-regulated set of genes from the susceptible interaction with *R. solanacearum*, 55 were found that had some response to a treatment. Of these, 25 were annotated as up-regulated in response to *Pseudomonas syringae* pv *tomato* DC3000 avrRpt2, and one was marked as down-regulated (Figure 3.4). However, this result is the opposite that would be expected from a susceptible interaction. A possible reason for this is that there is a delayed expression of certain genes in the susceptible interaction, which would normally be immediately expressed in a resistant interaction, that is, there is a common expression of genes between in the susceptible and resistant interactions.

In the down-regulated genes from the susceptible interaction, the only significant experiment found was a methyl jasmonate treatment. Of the 57 genes that were submitted 12 results were found, with 8 marked as up-regulated and the remaining 4 were down-regulated. This is again the opposite of what would be expected, as in the susceptible interaction, SA signalling is repressed, resulting in increased MeJ signalling. However, in this case, the genes in the gene set were down-regulated in the *R. solanacearum* experiment.

This same process was repeated for the resistant interaction. In the up-regulated gene set, of the 55 genes submitted, 9 genes were marked as up-regulated in response to *P. syringae* pv. *tomato* DC3000 avrRpt2; 5 up and 2 down in response to an *Alternaria brassicicola* infection; and 5 up and 2 down in response to a MeJ treatment. The first result is what would be

expected, as both the experiment in DRASTIC and the data from the *R. solanacearum* experiment are from resistant interactions. However, the last two responses are again the opposite of what would be expected. In the down-regulated gene set, there were no significant responses found.

Responses to Treatments - DRASTIC

Data from the manually curated DRASTIC¹ database showing responses of genes to various experimental conditions
Click on a response to get more information

	Yariv phenylglycoside (beta-D-Glc)3	low oxygen	light (HIGH)	ABA (abscisic acid)	Pseudomonas syringae pv tomato	Myzus persicae	ozone	Colletotrichum higginsianum
At3g57520					up	up		
At4g11650		up			up			
At5g59320			up	up				
At5g06760	down			up				
At1g03090		up				up		
At2g39800				up				
At3g02550		up						
At2g42890					up			
At4g15530					up	down		
At5g45350	up							
At2g34500				up	up			
At5g46180					up			
At1g52890					up			
At1g02205								
At4g16260			down		up			
At2g28200								
At1g75170					up			
At3g26100								
At5g53970						down		
At4g37430				up	up			
At2g15970				up				
At5g66170		up		up	up			
At5g66760					up			
At1g43160					up			

Figure 3.4: Screenshot of a portion of the results of the regulation of the genes, retrieved from the DRASTIC database, in the up-regulated gene set from the susceptible interaction with *R. solanacearum*. It can be seen that many of the genes in the cluster have been annotated as being up-regulated in response to *Pseudomonas syringae pv tomato* DC3000 avrRpt2 (a green block indicates up-regulation and a red block indicated down-regulation). Each block can be clicked for additional information, such as reference data.

3.3.4.2 PCA Experiment Comparer

Since the DRASTIC database is relatively small, due to the data being manually curated, the dataset may not reveal all patterns. Thus, an alternative approach was to compare the expression profile of the genes in the cluster to the expression profiles of a variety of other experiments, namely the data in NASCArrays (Nottingham Arabidopsis Stock Centre Arrays) (Craigon *et al.*, 2004; 20). However, this dataset is extremely large, making it difficult for a researcher to interpret. Thus, to reduce the number of variables, Principal Component Analysis (PCA) was applied to the data.

As an initial test to determine if a PCA could be used to correctly group similar expression profiles, the data from Glazebrook *et al.* (2003) were used. In this experiment, slides containing approximately 8000 probes were used to determine the expression of wild-type plants against several signalling-defective mutant plants, including *eds3*, *eds4*, *eds5*, *eds8*, *pad1*, *pad2*, *pad4*, *NahG*, *npr1*, *sid2*, *ein2* and *coi1* (Glazebrook *et al.*, 2003). This dataset was used as it explored the global expression of wild-type and several signalling-defective mutant plants in response to *P. syringae* pv. *maculicola* strain ES4326 infection, allowing the placement of regulatory genes in the defence signalling network. Hierarchical clustering was performed by Glazebrook *et al.* on both the mutants and genes, and four clusters were identified as being biologically significant (Figure 3.5). The mutants could be roughly grouped into three groups: one group consisting of *eds4*, *eds5*, *sid2* and *npr1-3*, which were affected only SA signalling; another consisting of *pad2*, *eds3*, *npr1-1*, *pad4* and *NahG* affecting SA signalling as well as another unknown process; and the last comprised of *eds8*, *pad1*, *ein2*, and *coi1*, which affected ethylene and jasmonate signalling (Glazebrook *et al.*, 2003).

A PCA was applied on the \log_2 -ratios of the genes in all the clusters, and across all the mutants, to determine if the same clustering could be obtained as by the authors. The results of the PCA are shown in Figure 3.6. As can be seen, a relatively similar result was obtained, with similar groupings. Thus it suggested that a PCA could be applied to group expression data into similar profiles. The Q^2 values for this PCA model were 0.329 for the first PC and 0.318 for the second. While these values are fairly low, it is most likely due to the relatively small dataset, and possibly noisy microarray data.

When applying the PCA Experiment Comparer method on defence response data, again the clusters from Glazebrook *et al.* (2005) were used for the proof of concept. The dataset that was used in this exploratory analysis were the mutants that were only affected by SA signalling, that is, *eds4*, *eds5-1*, *eds5-3*, *npr1-3* and *sid2*. All the genes from the four clusters were used.

In the *eds4* mutant, the closest match was to experiment observing the transcriptome changes in *Arabidopsis* during pathogen and insect attack. In particular, the closest slide was from a green peach aphid (*Myzus persicae*) infestation.

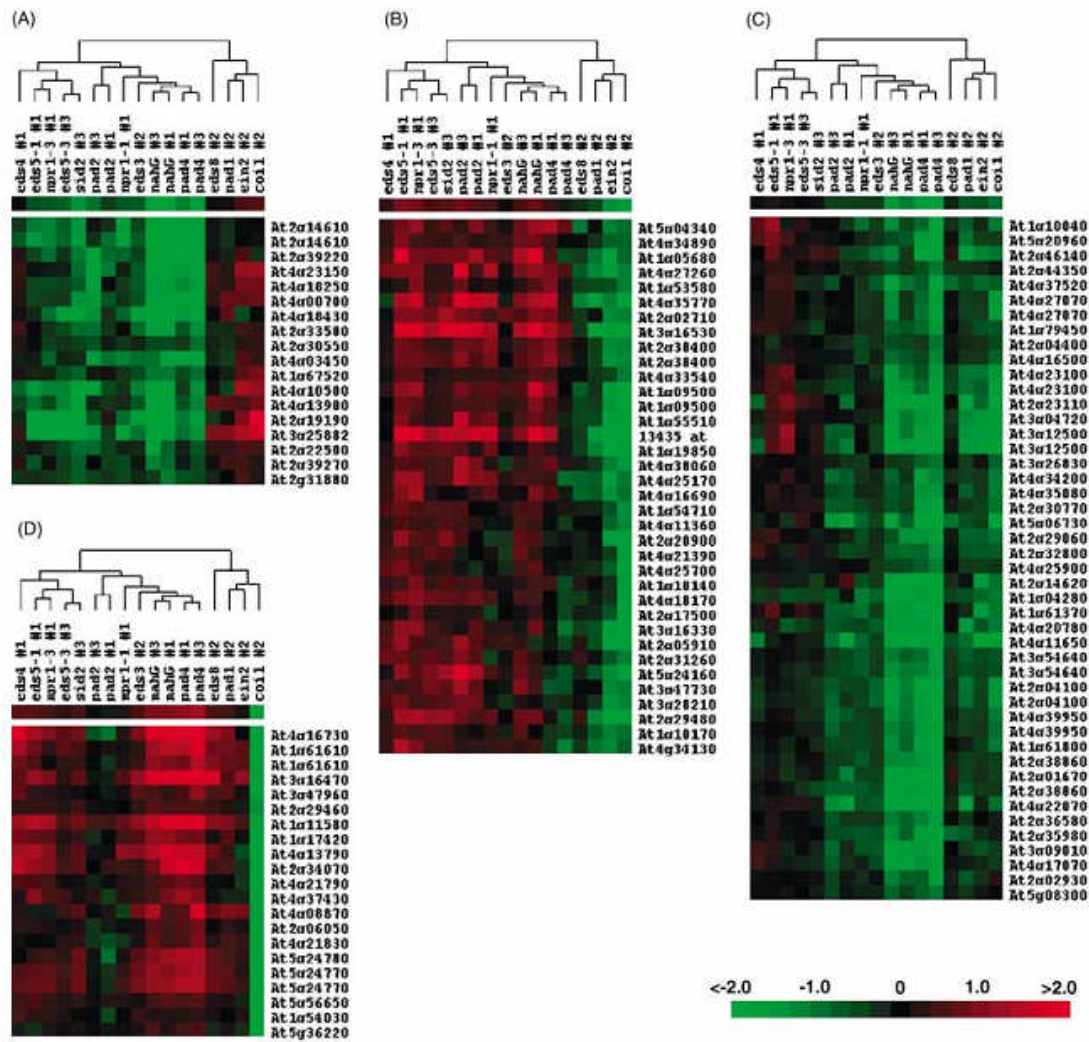


Figure 3.5: Clusters proposed by Glazebrook *et al.*, showing the gene name on the right, and the mutant along the top (Glazebrook *et al.*, 2003). Each coloured block represents the \log_2 -ratio values (calculated as $\log_2(\text{infected mutant}/\text{infected wild-type})$). The lengths of the branches on the dendrogram indicate the correlation between the datasets with shorter branch lengths meaning closer correlations.

Close matches with the *eds5-1* mutant interestingly included a response of wild-type plants to a *Pseudomonas syringae* pv *tomato* DC3000 infiltration after 2 hours, as well as treatment with gibberellin GA4. The former is significant as infection with the virulent strain of *P. syringae* DC3000 would result in suppression of the SA pathway and induction of the JA and ET pathways, as would be expected from this SA-deficient signalling mutant (*eds5-1*). In the latter experiment, gibberellins are known to promote the degradation of the plant growth repressor DELLA proteins (Navarro *et al.*, 2008). In addition, it has been shown that DELLAs suppress SA signalling and biosynthesis, and is also involved in JA perception and signalling, suggesting that DELLAs increase susceptibility to virulent biotrophs and resistance to

necrotrophs (Navarro *et al.*, 2008). Therefore through this mechanism, gibberellin increases the susceptibility to necrotrophs by degrading DELLAs and thus altering the levels of SA and JA (Navarro *et al.*, 2008). However, this is the opposite result of what would be expected from the *eds5-1* mutant which is deficient in SA signalling. Possible reasons could be that the above proposed model is not accurate, or simply that GA4 acts differently than the gibberellins that were used by Navarro *et al.* Other close experiments to the *eds5-1* mutant included a whole genome expression experiment on Arabidopsis' response to the application of herbicidal levels of 2,4-D, as well as a potassium starvation treatment. These could be indications of crosstalk with stress responses.

In the *eds5-3* mutant, a close experiment was again the *Pseudomonas syringae* pv *tomato* DC3000 infiltration after 2 hours. The *eds5-3* mutant's expression was similar to a *cpr5* mutant experiment. The *cpr5* mutant has been shown to cause constitutive expression both an NPR1-dependent and an NPR1-independent signalling pathways (Bowling *et al.*, 1997), that is SA- and JA/ET-dependent signalling pathways respectively. This match could have resulted because both the *eds5-3* and *cpr5* mutants caused increased expression in the JA signalling pathway. Other experiments were responses to cold and a drought stress treatment, again possibly indicative of stress response crosstalk.

When analysing the data from the *npr1-3* mutant, the experiments matched were similar to those of the *eds5-1* mutant. These included the response to 2,4-D application and potassium starvation. Interestingly, these data included two responses to *Pseudomonas syringae* pv *tomato* DC3000 infiltration – at 2 hours and at 24 hours. Again, a similar experiment was a response to cold treatment.

Finally, in the *sid2* mutant close experiments were again overlaps of experiments previously seen. These included the *Pseudomonas syringae* pv *tomato* DC3000 infiltration, harvested after 24 hours, and the *cpr5* mutant.

When calculating Q^2 for the above datasets, it was generally found that the values were generally fairly poor, averaging between 0.1 and 0.3. This suggested that there was possibly a lot of noise in the data making it difficult to detect clear patterns. One possible explanation for the noise is the comparison of different types of experiments to each other that may not in fact be comparable, for example, comparing a gene knockout experiment with a whole genome treatment. Nonetheless, the use of PCA did identify some patterns in the expression profiles.

The amount of overlap between all the mutants is not surprising as both the PCA and the Glazebrook *et al.* hierarchical clustering showed that these mutants have similar expression patterns. In addition, since all the mutants are suspected of being involved in the SA signalling pathways (and thus are deficient in SA signalling) this explains the large number of treatments involved in the JA- and ET-dependent signalling pathways.

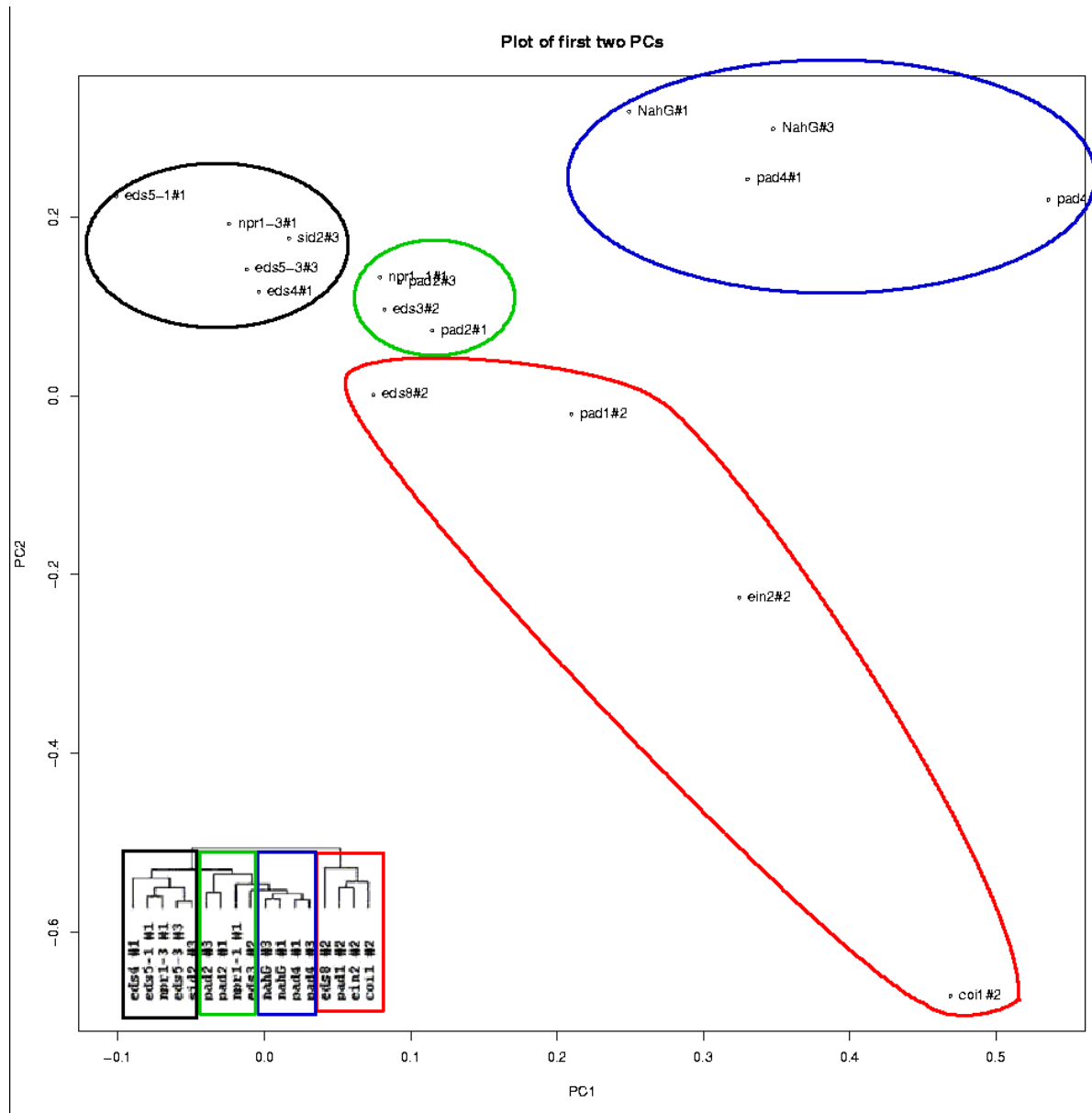


Figure 3.6: Plots of the top two principal components, using the data from the Glazebrook *et al.* (2003) mutant study. The inset in the bottom left is the dendrogram representing the clusters that were elucidated using hierarchical cluster (Glazebrook *et al.*, 2003). The areas bordered with the same colours represent the same set of mutants. As can be seen, the PCA resulted in similar results, with same groupings formed, even though the distances between the mutant experiments is different.

3.3.5 PCA on *Ralstonia solanacearum* data

The data from the susceptible *R. solanacearum* interaction was analysed using the PCA Experiment Comparer in an attempt to see if it was possible to determine which signalling pathway is activated in response to *R. solanacearum*. It was hypothesised that the data from this experiment would match the data from other susceptible interactions with other bacterial pathogens. The top five results after submitting the data are shown in Table 3.2. Several interesting results were obtained including a *Pseudomonas syringae* DC3000 infiltration in a WRKY knockout mutant, as well as a MeJ treatment. This last result confirmed the MADIBA GO analysis that also indicated that terms with JA and MeJ were significant. In a susceptible interaction, the SA pathway is generally repressed by the pathogen to reduce the defence response, and so the JA pathway is up-regulated. In addition, since it is known that some pathogens such as *P. syringae* act as hemi-biotrophs (act both as biotrophs and necrotrophs) (Glazebrook, 2005), it could be possible that the gene set used was particular to a phase that stimulated the JA signalling pathway.

The other results were a match to a *cpr5/scv1* double mutant, which has wild-type susceptibility to *Peronospora parasitica*, slightly enhanced susceptibility to *Pseudomonas syringae* and resistance to *Botrytis cinerea* (Anderson *et al.*, 2004); as well as drought stress and a hydrogen peroxide treatment on plants over-expressing Zat12. The drought stress experiment is understandable due to the nature of a *R. solanacearum* infection which causes wilting symptoms (Deslandes *et al.*, 2003). Terms relating to wilting and water loss were also identified in the MADIBA GO analysis. Zat12 is thought to be involved in high light and cold acclimation, and so together with the drought stress and hydrogen peroxide treatments, could also indicate some crosstalk between the different signalling pathways.

The box plot of the Q^2 values for the top three principal components is shown in Figure 3.7. While the average value is fairly low (all less than 0.5), these values are still higher than the Q^2 values that were calculated for the test cases.

When the PCA Experiment Comparer was applied to the resistant interaction data, it was expected that these data would match experiments related to SA signalling (Deslandes *et al.*, 2003; Noutoshi *et al.*, 2005). However, the only experiment related to SA signalling was a BTH treatment (NASCARRAYS-392), and specifically the slide that matched was a wild-type plant, harvested 24 hours after spraying with 60 μ M BTH. Since BTH is a SA analogue (Shimono *et al.*, 2007), this could be an indication of increased SA signalling in the resistant

interaction. Unfortunately, the other close experiments had little to do with defence signalling, being involved in diverse experiments such as circadian gene expression in response to different light conditions, response to cold, cell cycle experiments and an investigation into the transcriptome of post-germinative *Arabidopsis* embryos. Interestingly, many of these experiments matched terms found using the MADIBA GO analysis module.

3.3.6 Conclusion

Some novel ways were suggested for identifying genes involved in the defence response. One way was to identify how each gene responds to various treatments, and the DRASTIC database was used for this purpose. Unfortunately, no definitive results could be determined when using this information, and in many cases, the opposite result of what was expected was found. Since the data in DRASTIC is manually curated from literature, the amount of data are not sufficient to draw inferences as to whether or not a gene is involved in a defence response.

Using the NASCArrays data provided a greater set of works to perform this analysis on, and applying a PCA on it allowed for the data to be interpreted. While this approach did work to an extent, numerous other concerns arose. One in particular was the issue of comparing different types of experiments to each other that may not in fact be comparable, for example, comparing a gene knockout experiment with a whole genome treatment, or gene expressions from roots compared with the gene expression from shoots. In addition, because there was not a standard format for naming of slides, many values had to be discarded, including many that were related to plant defence.

The PCA Experiment Comparer method showed that it is possible to group some unknown data from an experiment, and potentially identify which other experimental conditions the data matches. When applying this data to *R. solanacearum* data, several interesting results were found, particularly in the data from the susceptible interaction experiment. This method provided numerous insights into the data, showing similarity to another susceptible interaction, as well as a MeJ treatment, as was expected. An analysis of the genes in the resistant interaction showed a match to a BTH treatment, which may be an indication of increased SA signalling, thus supporting the hypothesis that SA is involved in the resistant interaction. When applying this method, only genes that were determined to be significantly up- or down-regulated were used. An alternative approach may be to use all genes, including those that are constitutively expressed, that is, those genes whose expression does not

significantly change. This may potentially provide more information upon which to base a comparison.

Besides using a PCA, it may be possible to utilise more involved methods, such as principal components discriminant function analysis (PC-DFA) and support vector machines (SVM) instead, which are more robust and give an estimate of strongly the data clusters together (i.e. how likely the clustering is true).

While numerous other cross validation methods, such as parameter bootstrapping, could be used, the Q^2 statistic was used as a basic suggestion to the relevance of the data. Thus this value attempted to provide an indication of the significance of the results.

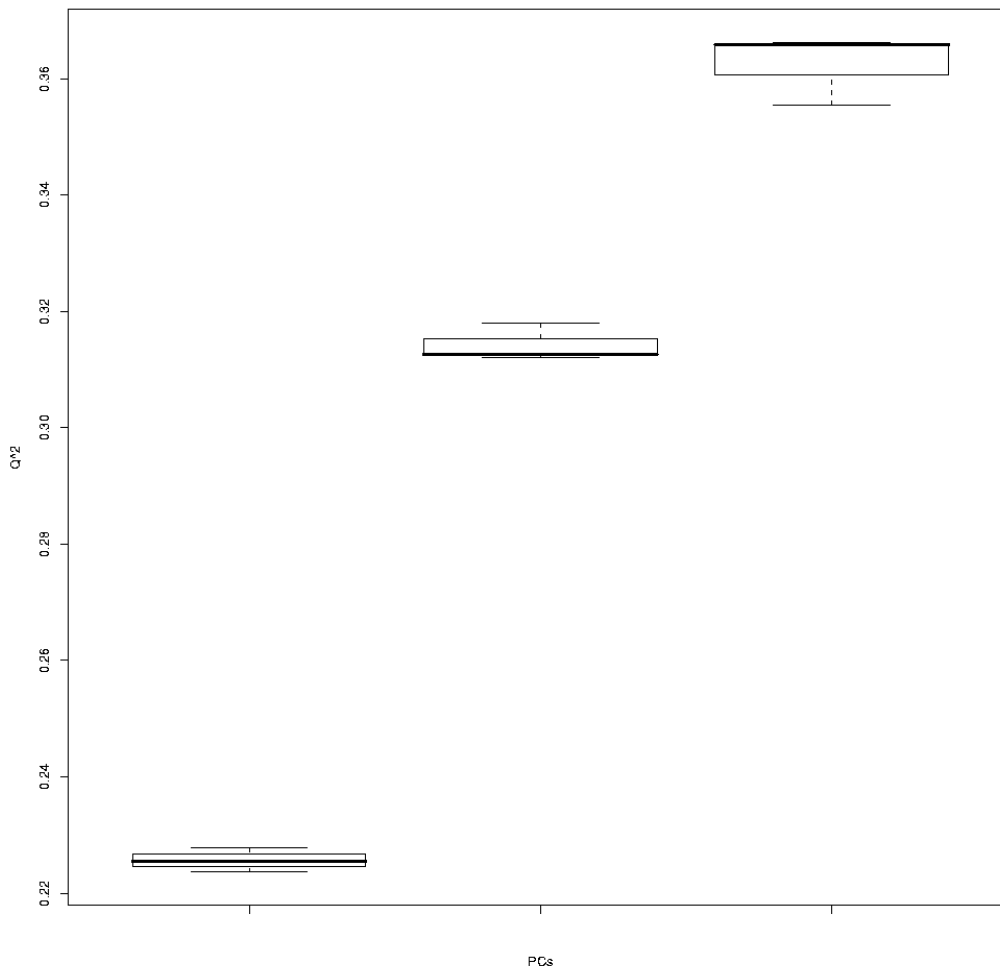


Figure 3.7: Box plot of the Q^2 values for the first three principal components after a PCA using the susceptible *R. solanacearum* interaction data. The plot was determined after three complete cross-validation runs. Each box shows the median, the first and third quartiles, and the maximum and minimum calculations. Outliers are represented as circles (none present in this figure).

Table 3.2: The five closest results from the PCA Experiment Comparison of the susceptible *R. solanacearum* interaction. Provided are the experiment's reference number in NASCArrays (20), the distance indicating how "far away" the experiment is from the *R. solanacearum* data, the title of the experiment, the slide name, the treatment of the particular slide matched, and a description of the experiment. The distance measure is a calculation of the Euclidian distance from the experiment to the submitted data and thus is a relative measure having no scale.

NASCArrays Reference	Distance	Title	Slide	Treatment	Description
398	0.00756	Group II-A WRKY transcription factors and early leaf senescence	Ulker_WRKY-KO-30-Pst-DC3000	Vacuum infiltration with <i>Pseudomonas syringae</i> DC3000 (1x10 ⁷ cfu/ml), harvested 6 hours after treatment	Comparison of the gene expression profiles of 3-weekold wild type and WRKY T-DNA knockout mutants grown in a growth chamber under long day growth conditions and subsequently challenged for 6 hours with the virulent bacterial pathogen <i>P. syringae</i> DC3000.
174	0.00917	AtGenExpress: Methyl Jasmonate time course in wildtype	RIKEN-GODA22A	10 μ M MeJ for 3 hours	Wild-type seedlings were treated with methyl jasmonate for 30 min, 1 and 3 hours.
141	0.01109	AtGenExpress: Stress Treatments (Drought stress)	AtGen_6-4411_Droughtstres s-Shoots-6.0h_Rep1	Plants were stressed by 15 min. dry air stream (clean bench) until 10% loss of fresh weight; then incubation in closed vessels in the climate chamber.	Wild type plants (col-0) were grown for 16 days, and drought stress treatments started at 3 hours of light period; samples taken at 0.5, 1, 3, 6, 12, 24 hours after treatment.
355	0.01175	Mutant array	Yang_1-9_CPR5SCV1-1_Rep1_ATH1	<i>cpr5/scv1</i> double mutant	Two week-old Arabidopsis aerial tissues from Columbia-0 and <i>cpr5</i> , <i>cpr5npr1</i> , <i>cpr5scv1</i> , <i>cpr5npr1svi1</i> , and <i>npr1</i> lines were collected for analysis.
338	0.01272	Hydrogen peroxide stress and Zat12 over-expression in Arabidopsis	Mittler_2-7_Zat12_Rep1_AT H1	20 mM hydrogen peroxide for 1 hour	Investigation of the transcriptome of transgenic Arabidopsis seedlings (5-day-old) constitutively expressing the zinc-finger protein Zat12 (At5g59820) under the control of the 35S-CaMV promoter (Zat12). The transcriptome of these seedlings was compared to that of wild type seedlings grown under the same conditions (WT) and to that of wild type seedlings grown under the same conditions and subjected to a hydrogen peroxide stress (WT+H ₂ O ₂).

3.4 Application to rice

Arabidopsis thaliana and rice are model species for dicotyledonous and monocotyledonous plants respectively (Rensink and Buell, 2004). Analysis of rice can lead to crop improvements in both developed and developing countries. These crops can include economically important crops, such as wheat and maize, as well as orphan crops such as cassava, cowpea and pearl millet, which are important for food security in Africa. In this section, rice will be discussed as model species, followed by a case study using MADIBA. This will be followed by an application of MADIBA on pearl millet, a related member of the grass family.

3.4.1 Importance of rice

Cereals are one of the greatest sources of food for the world's population. Rice (*Oryza sativa*) is one of three cereals produced annually at worldwide levels of over half a billion tons (608 million metric tons of rice was produced worldwide in 2004 – UN Food and Agriculture Organisation). Unlike the other cereals, more than 90% of rice is consumed by humans, with approximately half of the world's population deriving a significant caloric intake from its consumption (Goff, 1999). Since the human population is predicted to rise over the next several years, it is likely that the demand for rice will also increase. Rice production has increased significantly over the past several years as a result of new varieties and improved technologies. However, the increase in production is no longer keeping pace with the growth in the number of consumers (Goff, 1999; Sasaki and Burr, 2000). This is primarily due to the lack of new land, water and labour to increase the cultivation of rice, and so, larger yields per plant will be needed to meet the higher demand (Sasaki and Burr, 2000; Ronald and Leung, 2002). In addition, the increasing affluent portions of the population will want better quality rice.

Applications of molecular techniques will assist in achieving these crop improvements. EST cloning however, does not provide sufficient information in order to base crop improvements (Goff, 1999). Map based sequence information is required to fully exploit the full potential of the rice sequence since of the location of a specific gene in the genome allows the identification of candidate genes that control specific traits. To this end, the rice genome was sequenced. Through multiple sequencing projects, there are currently four draft sequences available – three focussed in the Nipponbare cultivar from the temperate *japonica* subspecies, and one focussed on the 93-11 variety from the tropical *indica* subspecies (Rensink and Buell, 2004). All four drafts are publicly available to the academic community (Buell, 2002). With

the genome sequenced, it should be possible to identify all genes or proteins that are expressed under a condition of interest, and using comparative genomic analyses, identify the genes that are predicted to be involved in the trait of interest. Through this, it will be possible to use reverse genetics to confirm the prediction (by over-expression or knock-out mutants), assess the correlation of the gene with the trait in the population and so aid in the engineering of new crop varieties, and allele mining of various germplasms to identify useful variants (Ronald and Leung, 2002).

3.4.2 Rice as a model species

Two features makes rice attractive as a model species: 1) it is a crop species, so improvements can have practical applications, and 2) it represents the taxonomically distinct monocotyledons (Rensink and Buell, 2004). While *Arabidopsis thaliana* is well established as a model species for plant biology, in particular dicotyledonous plants, it serves as a poor model for cereals and other monocotyledons. Furthermore, rice has a considerably smaller genome size as compared to the other cereals, with an estimated size of 420-450Mb (Sasaki and Burr, 2000). Despite having the smallest genome of the cereals, the rice genome is still three times the size of the *A. thaliana* genome (Sasaki and Burr, 2000). In a comparative study between rice and *A. thaliana*, it was found that rice contains a homologue for approximately 81% of the proteins in the *A. thaliana* genome (Sasaki and Burr, 2000). This suggests a considerable overlap between the genes that are required for basic functions in monocotyledonous and dicotyledonous plants. The complete sequence of both plants presents opportunities for comparative genomics and investigations into the divergence of plants.

As a member of the Poaceae (Gramineae) family, rice is closely related to the other major cereals: sorghum, maize, barley and wheat. Since these cereals have much larger genome sizes (1000, 3000, 5000 and 16000Mb respectively), the smaller genome size of rice means a higher gene density, that is, a higher chance of encountering a gene (Goff, 1999). It is estimated that the rice genome contains one gene every 5.3kb (Rensink and Buell, 2004; Yuan *et al.*, 2005). Despite the difference in the size of the various cereals, it is predicted that the cereals display close synteny, which means that the genes are arranged in a similar general order within the genome (Rensink and Buell, 2004; Moore *et al.*, 1995). Figure 3.8 illustrates this concept. Although the synteny between the cereal genomes is not as absolute as previously thought, local regions of co-linearity will still be important in positioning clones in

larger cereal genome projects, as well as identifying agricultural regions of interest (Goff, 1999).

The study of individual genes also demonstrates that there is a considerable homology among the various cereal gene families (Gale and Devos, 1998). This conservation of gene and protein sequences suggests that studies on the function of a gene in rice could lead to the elucidation of functions of orthologous genes and proteins of other cereals. It is thus possible for the identity of a gene in any cereal to be matched to the corresponding rice gene (Gale and Devos, 1998). The structural similarities in the various cereal genomes have led to the proposal that cereal genomes arose from a common ancestor and can be viewed as a single genetic system. Thus the use of the rice genome will become important as the base species in comparative genomics in cereals.

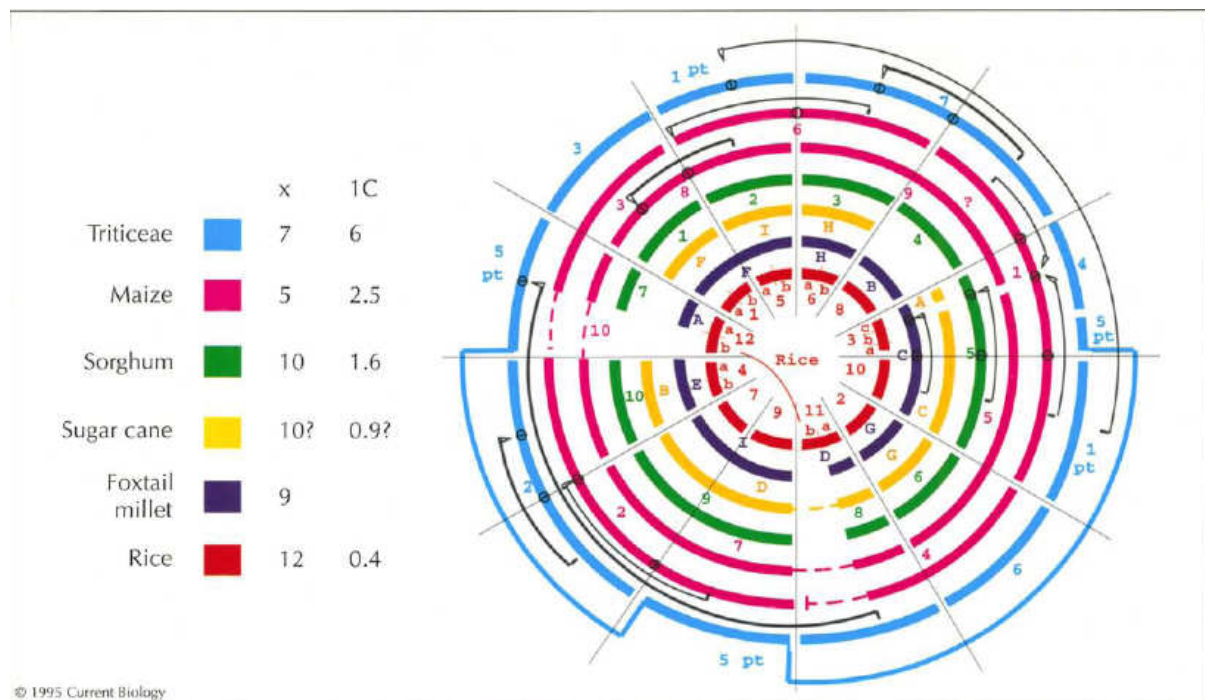


Figure 3.8: An alignment of the genomes of six major grass crop species, drawn using 12 rice linkage segments, whose order reflects the circularized ancestral grass genome. The rice linkage segments are drawn as radiating lines and formed into chromosomes (coloured and numbered lines). The thin dashed lines correspond to duplicated segments (Moore *et al.*, 1995).

3.4.3 MADIBA rice data analysis

In this section, the objective was to determine whether the data from a rice SA treatment could easily be interpreted using MADIBA. Once this proof of concept was shown, it was hoped to use MADIBA on SA responses in pearl millet (section 3.4.4).

The data for the rice analysis was obtained from an experiment where rice plants were exposed to benzothiadiazole (BTH) (Shimono *et al.*, 2007). BTH is a functional analogue of SA and is used to protect various plants from infection diseases by activating the salicylic acid (SA) signalling pathway (Shimono *et al.*, 2007; Murray *et al.*, 2002). When applied to plants at high dosages, it induces constitutive activation of defence response. By contrast, when applied at relatively low dosages, the plants' defence responses do not activate immediately but only become apparent after pathogen infection (Shimono *et al.*, 2007). It has also been shown that over-expression of BTH- and SA-inducible WRKY transcription factor (TF) genes in rice increased resistance to rice blast fungus (*Magnaporthe grisea*) and *Rhizoctonia solani* (Shimono *et al.*, 2007).

44000 genes from BTH-treated rice plants were screened using 60-mer oligo DNA microarrays, and a statistical analysis by Shimono *et al.*, using analysis of variance with a false discovery rate (*q*-value) < 0.05, identified 326 significant up-regulated genes. These included several WRKY TF genes and many defence related genes (Shimono *et al.*, 2007).

The sequences for these 326 genes were obtained from GenBank, and submitted to MADIBA. A BLASTX search was performed on the sequences with a maximum e-value of 1. The default top hit was taken for each gene, and generally the matches had good e-values (e-value < 10⁻¹⁰). When the annotations of the genes were compared, it could be seen that the set of genes contained several genes annotated as glutathione transferases, DNA binding proteins, protein binding proteins, cytochrome P450 and numerous WRKY genes (Figure 3.9).

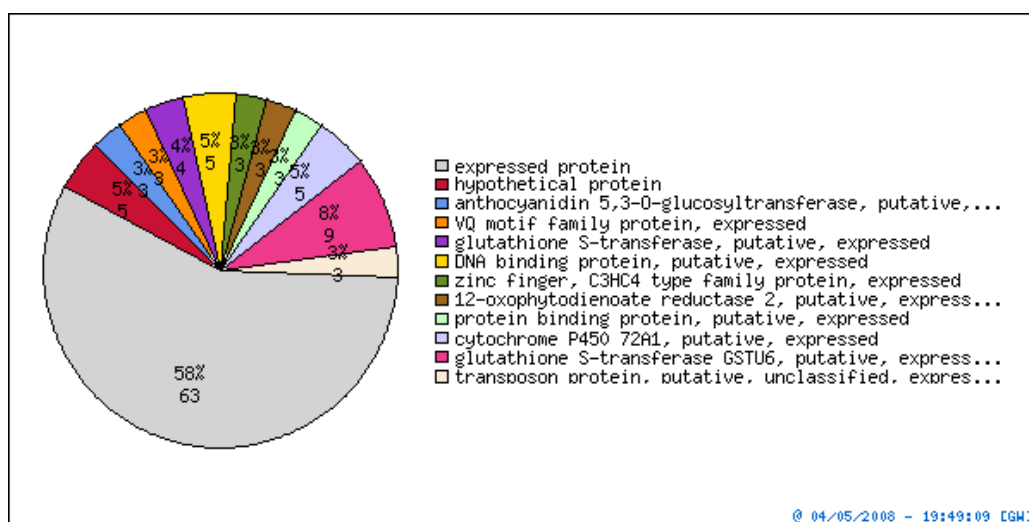


Figure 3.9: Pie chart from MADIBA, showing common annotations in the cluster. Of interest are the glutathione transferases and cytochrome P450, as both are involved in plant defences.

Using the Metabolic Pathways module, a number of the genes' products appear in several metabolic pathways. These included the starch and sucrose metabolism pathway (6 unique enzymes), tyrosine metabolism, phenylalanine metabolism and flavanoid biosynthesis pathways (4 unique enzymes in each). However, none of these appear to be significant, with only the flavanoid biosynthesis pathway having a p -value less than 0.05 (p -value = 0.03514). However, in this pathway, all the enzymes that were present in the cluster were annotated by PRIAM only, so this result may not be completely reliable. As mentioned above, PRIAM annotations are performed automatically, using sequence similarity to find enzymes of the same family, and are not proven experimentally.

After analysing the set of genes using the Gene Ontology module and in the molecular function ontology, it was found that the term glutathione transferase activity (GO:0004364) was highly significant (FDR corrected p -value = 1.1636×10^{-11}), as well as several terms relating to transferase activity including UDP-glycosyltransferase activity (GO:0008194) and UDP-glucosyltransferase (GO:0035251) activity. When analysing the cluster using the biological process ontology, a number of unusual results were found. The GO term with the lowest p -value was the toxin catabolic process term (GO:0009407, p -value = 1.954×10^{-10}), as well as several other terms related to jasmonic acid, for example jasmonic acid biosynthetic process (GO:0009695) and response to jasmonic acid stimulus (GO:0009686). Terms related to salicylic acid did occur, but had much lower p -values. This is curious as BTH treatments are known to induce the SA pathway. It is possible that this discrepancy is due to incomplete GO annotations of the rice genome. However, it has been reported in *Arabidopsis thaliana* that systemic acquired resistance (SAR) is mediated by jasmonic acid, and not SA as previously thought (Truman *et al.*, 2007). Other terms that appeared related to responses to abiotic stress, such as cold and salt, and to cell growth stimuli. Looking at the cellular component ontology revealed that most of the gene products occur in the nucleolus and in the chloroplast.

The Transcription Regulation module was used to try to identify transcription regulator binding sites, using RSAT and TRANSFAC. Analysis by the Patch program in the TRANSFAC subsection of the Transcription Regulation module showed that a large proportion of the genes (134 out of a total of 326) contained a motif (ATTTAC) that is functionally important in the promoter of PR-1a, a well characterised pathogenesis related protein (Buchel *et al.*, 1999). Using the oligo-analysis tool of RSAT, the highest ranking

motif was CGCCGC, with an expected occurrence of 27.05, and actual occurrence of 282 and an e-value of 2×10^{-15} . None of the top 5 motifs presented matched the W-box binding site of the WRKY transcription factors ((C/T)TGAC(T/C) (Ulker and Somssich, 2004)), although while searching through all the motifs that were found, the motif CTGACC was found, which exactly matches the WRKY TFBS. This motif occurred 88 times, with an expected occurrence of 46.59 and e-value of 8.7×10^{-5} , ranked 606th. This low position is probably due to the fact that not all the genes that were submitted were WRKY TF. It is interesting to note that most of the motifs that were found by both the oligo-analysis and dyad-analysis were extremely CG rich. This is probably due to the fact that *A. thaliana* was used as the background model, since rice is currently not available in RSAT as a background model. The choice of background model is important as the statistics that RSAT calculates are dependent on the base composition of the intergenic regions of the organism being studied. However, the intergenic base composition of rice is not the same as *A. thaliana*. The average GC content in rice is approximately 44%, while it is 36% in *A. thaliana* (Rensink and Buell, 2004; Saccone and Pesole, 2003). In fact, it has been reported that in general, the average GC content in grasses is higher than other plants (Saccone and Pesole, 2003). Nonetheless, several of the motifs may still be valid.

In the Chromosomal Localisation module, it could be seen that most of the genes occurred on chromosomes 1 and 3. In addition, many of the glutathione S-transferases were located close to each other on chromosome 10 (Figure 3.10).

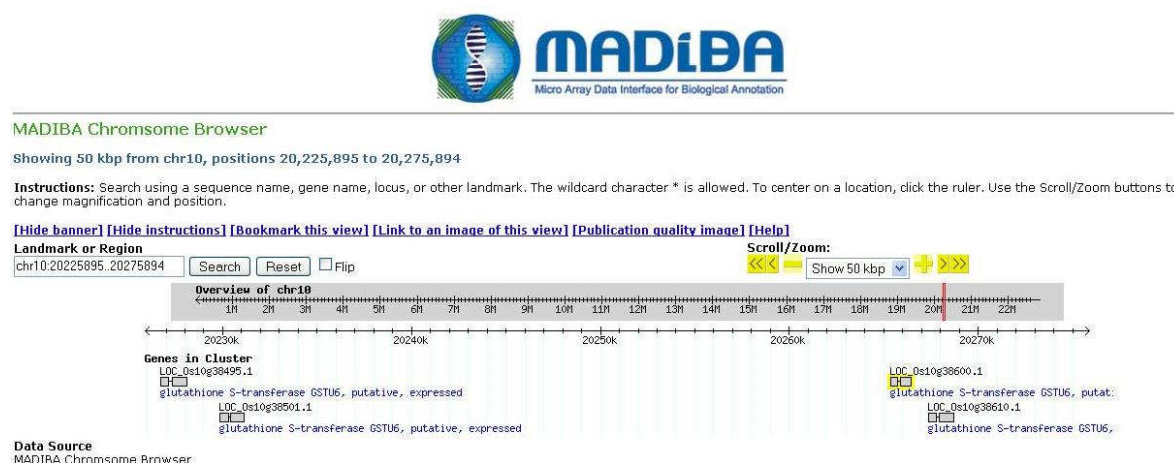


Figure 3.10: View of a portion of chromosome 10 of the genes, showing that a number of glutathione S-transferases were located close to each other.

3.4.4 MADIBA pearl millet data analysis

Pearl millet is an indigenous African crop, and is an important crop in the semiarid tropics of Africa and Asia. Thus, as a staple food source, it is necessary to identify means to prevent diseases that may potentially reduce yields of this crop (Crampton, 2006).

To test whether it is possible to use MADIBA on other cereals, data from a SSH (suppression subtractive hybridisation) experiment was used, where pearl millet plants were treated with either methyl jasmonate (MeJ) or salicylic acid (SA), prior to infection with the leaf rust fungus *Puccinia substriata* (Crampton, 2006). This study was performed by the MPPI group at the University of Pretoria, using a 2000 probe microarray derived from a SSH cDNA library, where pearl millet plants were treated with the elicitors chitin, flagellin and wounding. In the experiment, it was found that the prior treatment of SA resulted in increased resistance to the pathogen, whereas MeJ did not confer any significant resistance (Crampton, 2006). This result suggested that the SA signalling pathway is activated in response to rust infection. MADIBA was used to compare the responsive genes after SA treatment with those that were responsive after MeJ treatment.

3.4.4.1 SA responsive genes

From these data, 19 cDNA fragments that were responsive to SA but not MeJ were identified (Table 3.3). All these cDNAs were up-regulated in response to SA in at least one time point when compared to t_0 . In addition, a BLAST search of these cDNAs against GenBank revealed that they have functions that are known to be involved in SA biosynthesis and defence (UDP-salicylic acid glucosyltransferase, heat-shock protein 70), signal transduction in response to pathogens (calcium binding EF hand protein, serine carboxypeptidase, S-adenosylmethionine decarboxylase), and cellular detoxification in response to pathogens (glutaredoxin, multi-drug efflux protein, peroxidase).

Table 3.3: Table of the salicylic acid responsive genes that were used in the MADIBA analysis

cDNA	Blast ID	Putative Function	Function	E-value
5-B6	P12783	Phosphoglycerate kinase, cytosolic		1.00×10^{-49}
6-A4		No significant similarity		
16-B9		No significant similarity		
7-A7		No significant similarity		
14-B12	P49105	Glucose-6-phosphate isomerase	Basic metabolism	3.00×10^{-60}
13-D2	XP466501	Rhodanese-like domain-containing protein	Basic metabolism	3.00×10^{-21}
8-B2	BAD34358	Putative UDP-salicylic acid	Defence	5.00×10^{-8}

		glucosyltransferase		
2-F11	CAA05547	Putative HSP70	Defence	1.00x10 ⁻³⁹
3-B6	AAP51748	Serine carboxypeptidase	Defence	2.00x10 ⁻³³
6-H2	CAA69075	S-adenosylmethionine decarboxylase	Defence	8.00x10 ⁻³⁷
1-G9	AY104653	Glutaredoxin	Oxidative burst	3.00x10 ⁻⁴⁶
12-C6	NP_919535	Putative peroxidase	Oxidative burst	3.00x10 ⁻⁵³
6-B6	AB007405	Alanine aminotransferase	Photorespiration	
1-H12	AAM15963	Putative phosphoenolpyruvate carboxylase	Photosynthesis	5.00x10 ⁻¹⁶
12-F9	P12329	Chlorophyll a/b binding protein 1, chloroplast precursor (LHCII type I CAB-1)	Photosynthesis	4.00x10 ⁻¹⁰
7-A8	D63581	Elongation factor 1 alpha	Protein synthesis	2.00x10 ⁻¹⁰
5-B12	AK101337	Putative calcium binding EF-hand protein	Signalling	1.00x10 ⁻⁵³
8-D7	XM_478265	Putative MATE efflux protein family protein	Stress	2.00x10 ⁻³⁷
16-B8	BAD28236	Putative ASR2	Stress	2.00x10 ⁻¹¹

The sequences of these 19 cDNAs were submitted to MADIBA, and a BLASTX search was performed against the rice proteome, with a maximum e-value of 1. cDNA 6-A4 was discarded as it did not have any similarity to any rice protein, and the cDNAs 16-B9 and 7-A7 had fairly low e-values (0.266 and 0.285, respectively) but were retained for the analysis. The default top hit of the others were taken. Table 3.4 shows the hits that were used.

Table 3.4: Table showing the best hits from the rice database in MADIBA that were used in the further analyses.

cDNA	MADIBA best hit	Annotation	e-value
8-B2	LOC_Os09g34250.1	Indole-3-acetate beta-glucosyltransferase, putative, expressed	3.742x10 ⁻⁹
1-G9	LOC_Os04g42930.1	OsGrx_C2.2 - glutaredoxin subgroup I, expressed	4.265x10 ⁻⁵²
8-D7	LOC_Os07g31884.1	Transparent testa 12 protein, putative, expressed	1.024x10 ⁻³⁸
2-F11	LOC_Os01g62290.2	Heat shock cognate 70 kda protein, putative, expressed	1.047x10 ⁻¹⁹
3-B6	LOC_Os10g01134.1	Serine carboxypeptidase 1 precursor, putative, expressed	1.436x10 ⁻⁵⁰
5-B12	LOC_Os06g14324.1	Calcium binding EF-hand protein, putative, expressed	5.549x10 ⁻⁶⁶
12-C6	LOC_Os10g02040.1	Peroxidase 54 precursor, putative, expressed	1.686x10 ⁻⁵⁴
6-H2	LOC_Os04g42090.5	S-adenosylmethionine decarboxylase proenzyme, putative, expressed	0.000
16-B8	LOC_Os02g33820.1	Abscisic stress ripening protein 1, putative, expressed	5.947x10 ⁻⁹
5-B6	LOC_Os06g45710.1	Phosphoglycerate kinase, cytosolic, putative, expressed	2.360x10 ⁻⁴⁹
14-B12	LOC_Os03g56460.3	Glucose-6-phosphate isomerase, cytosolic A, putative, expressed	3.116x10 ⁻⁵⁹
7-A8	LOC_Os03g08060.2	Elongation factor 1-alpha, putative, expressed	2.089x10 ⁻¹¹
6-B6	LOC_Os10g25130.1	Alanine aminotransferase 2, putative, expressed	8.167x10 ⁻⁷⁸
13-D2	LOC_Os02g38240.1	Rhodanese family protein, putative, expressed	1.615x10 ⁻²²

6-A4	None	None	None
16-B9	LOC_Os03g40390.1	Expressed protein	0.266
1-H12	LOC_Os02g14770.3	Phosphoenolpyruvate carboxylase 1, putative, expressed	0.000
7-A7	LOC_Os11g40840.1	Receptor protein kinase CLAVATA1 precursor, putative, expressed	0.285
12-F9	LOC_Os01g41710.1	Chlorophyll a-b binding protein 2, chloroplast precursor, putative, expressed	1.037x10 ⁻⁶

A comparison of the genes that were responsive to BTH in the previous rice experiment, and the genes that were responsive to SA in this millet experiment, showed only an overlap of one gene, namely LOC_Os09g34250, a putative indole-3-acetate beta-glucosyltransferase. However, the product of this gene has been annotated as a UDP-glucose:salicylic glucosyltransferase by the original annotators of the gene and protein in GenBank. This indicates that there are differences in the annotations that are defined in the TIGR Osa1 database and the original annotations. Nonetheless, it is possible that this gene is an important component in the SA signalling pathway.

After analysing the genes with the Metabolic Pathways module, no real significant pathways were seen. The pathways with the most unique enzymes were the carbon fixation pathway (three enzymes) and the glycolysis/gluconeogenesis pathway (two enzymes).

Analysis of the GO terms in the molecular function ontology revealed that the most significant terms were alanine transaminase activity (GO:0004021) and glycine transaminase activity (GO:0047958). Previous experiments (Wu *et al.*, 2006) have shown that a γ -aminobutyrate transaminase is induced by SA and abscisic acid but not JA in rice. It was also found that hot pepper plants (*Capsicum annuum L. cv. Bugang*) encode an alanine transaminase whose expression is induced by SA and ethylene but not by MeJ (Kim *et al.*, 2005). It is possible that this enzyme, as well as being involved in metabolic reactions, may be a component in the plant defence signalling pathway. Other significant terms include transmembrane receptor protein serine/threonine kinase activity (GO:004675) and chlorophyll binding (GO:0016168). In the biological process ontology, the most significant terms seem to orient around photosynthesis (GO:0009769 and GO:0015979) and photorespiration (GO:0009853), as well as glycolysis (GO:0006069), responses to toxins (GO:0009036) and detection of bacterium (GO:0016045). Less significant was a term dealing with the response to abscisic acid (GO:0009737), another plant hormone known to be involved in plant defence signalling as well as abiotic stresses (Rabbani *et al.*, 2003). In the cellular component

ontology, the most significant terms seemed to be localised in the thylakoid membrane and light harvesting complex, further emphasising the photosynthesis aspect from the other ontologies. It should be noted that in both the biological process and cellular component ontologies, 10 of the genes were not annotated with annotations from that ontology. Further in the molecular function ontology, 7 genes did not have any GO annotation. This means that a significant portion of the submitted genes (over half in the cases of the biological process and cellular component ontologies) could not be used in this analysis.

A search in the TRANSFAC section of the Transcription Regulation module showed that many motifs were found that were related to chlorophyll binding proteins, including cab140 (chlorophyll a/b-protein 140) (binding sites CTCA and TAGCC), cab11 (CATCC), as well as PR-1a (ATTTAC).

The high occurrence of photosynthesis related terms was intriguing, possibly suggesting that a relation between defence signalling and photosynthesis exists. It has been shown previously that light and its interactions with photosynthesis related processes impact strongly on the susceptibility of plants to infections (Bechtold *et al.*, 2005). In addition, light intensity may affect which pathways operate, as low light intensities are required for the expression of the SA signalling pathway, but does not operate at very high light intensities (Bechtold *et al.*, 2005).

3.4.4.2 MeJ responsive genes

To contrast the SA responsive genes, a set of cDNAs that were responsive only to MeJ were selected, and the analyses run on them. Table 3.5 shows the cDNAs that were selected for the analysis, as well as the results after submitting to MADIBA. A BLASTX was again performed with a maximum e-value of one. As can be seen, several of the cDNAs that were submitted did not have any significant similarity to rice. It could be possible that these gene fragments may be unique to pearl millet. Increasing the maximum e-value did result in more hits found, but due to the poor matches, these genes were not used in the further analyses. This meant that out of the 24 cDNAs that were selected, 9 were discarded. Curiously, a gene that was included in this analysis was a Pathogenesis Related (PR) protein (clone 14-A1), which is traditionally thought to be involved in SA signalling (Thatcher *et al.*, 2005; Glazebrook, 2001). However, this gene was significantly expressed when the pearl millet plants were treated with MeJ, but not as much when treated with SA.

A comparison of the genes that are responsive to MeJ and the genes that were responsive to BTH showed that again there was only one gene in common, namely LOC_Os01g28450.1, a pathogenesis-related protein PRB1-2 precursor. This could mean that important components that are pathogenesis-related may be used in both the SA and JA pathways.

In the Metabolic Pathways module, no pathways were indicated, although this was to be expected as this experiment was not a metabolic perturbation, but rather a signal transduction experiment.

Analysis with the GO module showed that in the biological process ontology, significant terms included response to jasmonic acid stimulus (GO:0009753, FDR corrected p -value = 0.009), as expected, in addition to response to ethylene stimulus (GO:0009723), systemic acquired resistance (GO:0009627) and response to salicylic acid stimulus (GO:0009751). These last two terms are most likely a result of the inclusion of the PR protein in this analysis. Numerous other responses to stimuli were detected, including response to chemical stimulus (GO:0042221, p -value = 1.326×10^{-4}), desiccation (GO:0009269), salt stress (GO:0009651), response to oxidative stress (GO:0009751), response to sucrose stimulus (GO:0009744), response to stress (GO:0009651), heat (GO:0009408) and light (GO:0009408). This again could be suggestive of crosstalk between the various signalling pathways.

Table 3.5: Set of cDNA fragments that were identified as being responsive to methyl jasmonate (MeJ) only, and the results after submitting their sequences to MADIBA.

cDNA	Putative Function	MADIBA best hit	E-value
4-A1	Pathogenesis related protein 1	LOC_Os01g28450.1 protein pathogenesis-related protein PRB1-2 precursor, putative, expressed	1.23×10^{-20}
16-E11	Pore-forming toxin-like protein Hfr-2	No similarities found	
19-H3	Putative disease resistance protein	LOC_Os12g39620.5 protein disease resistance protein, putative, expressed	1.97×10^{-72}
15-G10	Manganese superoxide dismutase	No similarities found	
13-G1	Putative dehydration-responsive protein RD22	LOC_Os08g38810.2 protein RAFTIN1a protein, putative, expressed	1.52×10^{-18}
1-D3	Putative farnesyl-pyrophosphate synthetase	LOC_Os01g50760.1 protein farnesyl pyrophosphate synthetase, putative, expressed	8.03×10^{-64}
7-E2	Putative inorganic pyrophosphatase	LOC_Os06g08080.1 protein pyrophosphate-energized vacuolar membrane proton pump, putative, expressed	6.68×10^{-48}
7-G5	Glyceraldehyde 3-phosphate dehydrogenase, phosphorylating	LOC_Os08g03290.2 protein glyceraldehyde-3-phosphate dehydrogenase, cytosolic, putative, expressed	1.39×10^{-63}
10-C3	Putative transcription factor	LOC_Os02g54160.1 protein ethylene	1.65×10^{-9}

	EREBP1	response element binding protein, putative, expressed		
1-E5	Putative ubiquitin-associated (UBA) protein	LOC_Os02g38050.1 protein PB1 domain containing protein, expressed		5.02x10 ⁻²⁵
14-C1	Putative pyruvate dehydrogenase kinase 1	LOC_Os07g44330.3 protein protein kinase, mitochondrial precursor, putative, expressed		2.65x10 ⁻⁶⁵
1-G12	Triose phosphate/phosphate translocator	No similarities found		
18-D6	Putative photosystem I reaction centre subunit II, chloroplast precursor	LOC_Os08g44680.1 protein photosystem I reaction center subunit II, chloroplast precursor, putative, expressed		1.58x10 ⁻²⁶
3-F10	Rice homologue of Tat binding protein	LOC_Os07g49150.1 protein 26S protease regulatory subunit 4, putative, expressed		4.55x10 ⁻⁵⁸
2-D3	No significant similarity	No similarities found		
4-A9	No significant similarity	No similarities found		
5-B9	No significant similarity	No similarities found		
5-F7	No significant similarity	No similarities found		
5-H11	No significant similarity	No similarities found		
6-C3	No significant similarity	LOC_Os01g03390.1 protein Bowman-Birk type bran trypsin inhibitor precursor, putative, expressed		7.48x10 ⁻⁴
6-E1	No significant similarity	LOC_Os06g01210.1 protein plastocyanin, chloroplast precursor, putative, expressed		4.80x10 ⁻¹⁷
6-G9	No significant similarity	No similarities found		
6-H12	No significant similarity	LOC_Os01g69100.1 protein expressed	protein expressed	2.11x10 ⁻⁶
7-A10	No significant similarity	LOC_Os01g69100.1 protein	protein expressed	1.50x10 ⁻⁴

3.4.5 Conclusion

Data from a rice expression experiment to BTH response was used to show the functionality of MADIBA. The results did not show SA responsiveness as was expected, although this possibly could be a result of incomplete annotations of the rice genes. Even so, RSAT was able to identify the WRKY TFBS (W-box) in the upstream regions.

Data from a pearl millet experiment was applied to MADIBA. In this way, it was shown that it is possible to use orthologous cereal data with the rice database. In the set of SA responsive genes, there were a large number of genes relating to photosynthesis. Thus, it could be possible that photosynthesis and light, and the formation of reactive oxygen species play a significant role in plant defences. The MeJ responsive genes did not produce any remarkable results, although it did confirm many of the genes that were in the cluster as being part of jasmonic acid and ethylene signalling.

The results from many of the plant analyses (both rice and *A. thaliana*) did not only return results pertaining to defence responses, as expected, but also numerous responses to abiotic

stresses, such as cold and light. This could be an indication the expression of overlapping gene sets, as well as common signalling molecules, resulting in crosstalk between the different response pathways.

3.5 Application to *Pectobacterium atrosepticum*

3.5.1 *Pectobacterium atrosepticum* introduction

The Enterobacteriaceae are a large family of rod-shaped Gram-negative bacteria which contains some of the most devastating pathogens, including many human and animal pathogens, including *Escherichia coli*, *Salmonella enterica* and species of *Yersinia* and *Shigella* (Toth *et al.*, 2006). However, this family also includes several plant pathogens, including species of *Erwinia* (Toth *et al.*, 2006; Toth *et al.*, 2003). The *Erwinia* genus was first described in 1917 to include all members of Enterobacteriaceae that cause disease on plants. However, this has resulted in nomenclatural differences, and on the basis of 16S rDNA sequence analysis, it has been suggested the taxonomy be restructured with *Erwinia carotovora* ssp. *atrosepticum* renamed to *Pectobacterium atrosepticum* (*Pba*), *Erwinia carotovora* ssp. *carotovora* to *Pectobacterium carotovorum* ssp. *carotovorum* (*Pcc*) and *Erwinia chrysanthemi* to *Dickeya dadantii*. However, this new nomenclature has not been widely accepted by the research community (Toth *et al.*, 2003).

The soft-rotting enterobacterial plant pathogens (SREPPs), such as those mentioned above, are major bacterial plant pathogens of potato and other crops, and cause tissue maceration, termed soft rot disease, through the production of plant cell wall degrading enzymes (Toth *et al.*, 2003). Worldwide, these pathogens result in the loss of millions of US dollars as a result of crop losses, particularly in potato, the fourth largest crop in the world (Toth *et al.*, 2006). *Pectobacterium atrosepticum* (*Pba*), which is the aetiological agent of potato black leg, will be the focus of this section. Other notable members of the group include *Pectobacterium carotovorum* ssp. *carotovorum* (*Pcc*) which infects a wide range of crops, including Brussel sprouts, carrot, celery, cucumber, turnip and potato; and *Dickeya dadantii* (formally known as *Erwinia chrysanthemi*), which is more frequent in sub-tropical regions and infects carnations, maize, pineapple, potato and African violet (Toth *et al.*, 2003). While much is currently known about how the bacteria attack plants and protect themselves against plant defences the processes underlying the establishment of infection, differences in host range and their ability to survive when not causing disease is still largely a mystery (Toth *et al.*, 2003).

With the complete sequencing of the *Pba* genome (strain SCRI1043) (Bell *et al.*, 2004), comparative genomics has led to several interesting discoveries. It was found that *Pba* has a genome that is of a similar size to genomes of enterobacterial animal pathogens and shares a common set of genes. In *Pba*, several of these genes are regulators that have been altered for the control of genes associated exclusively with disease in plants. These genes are apparently acquired by horizontal gene transfer that involved interactions with plants (Toth *et al.*, 2006).

SREPPs synthesise and secrete large quantities of plant cell wall degrading enzymes that are responsible for the soft rot phenotype, earning them the epithet “brute force” pathogens (Toth and Birch, 2005). This is in contrast to classic “stealth” pathogens such as *Pseudomonas syringae* which possesses a large variety of Type III secreted effector proteins and phytotoxins to manipulate and suppress host defences (Toth and Birch, 2005). However, as a result of whole genome sequencing, it has been found that there are components of stealth pathogenesis within *Pba*, including a Type III secretion system and phytotoxins, suggesting that stealth and brute force should not be regarded as mutually exclusive modes of pathogenesis (Toth *et al.*, 2006; Toth and Birch, 2005). Thus *Pba* reveals the capacity for a pathogen to be highly destructive, as well as subtly manipulate host defences (Toth and Birch, 2005). In addition, it also possesses the ability to thrive in association with plant hosts on which it does not cause disease, suggesting that it is more versatile than many animal-pathogenic enterobacteria, which are highly specific in their hosts (Toth *et al.*, 2006; Toth and Birch, 2005).

3.5.2 Quorum sensing

The primary weapon used by soft-rotting enterobacterial plant pathogens (SREPPs) is the coordinated production of high levels of multiple exoenzymes (enzymes that are secreted by the bacteria and function outside the cell), including pectinases, cellulases and proteases, which are used to breakdown plant cell walls in order to release nutrients for bacterial growth (Toth *et al.*, 2003). Pectinases are the main exoenzyme responsible for disease development, and results in the breakdown of pectins in the middle lamella and plant cell walls causing tissue collapse, cell damage and cell leakage (Toth *et al.*, 2003). While exoenzymes are not unique to SREPPs, the ability to co-ordinately produce large amounts of these exoenzymes at critical stages of infection makes them formidable pathogens (Toth *et al.*, 2003). This is accomplished through a complex set of regulatory networks and secretion systems within the pathogen.

The term quorum sensing (QS) refers to the ability of bacteria to regulate gene expression according to the accumulation of signalling molecules that are made by every cell in the population (Barnard and Salmond, 2007; Toth *et al.*, 2004). In this way, the gene expression of related regulatory systems is coupled to the accumulation of a diffusible chemical signalling molecule (Barnard and Salmond, 2007). SREPP utilise two types of QS signalling molecules: *N*-acyl homoserine lactones (*N*-AHL, or AI-1 signals), synthesised by the LuxI family of proteins; and the AI-2 signal, synthesised by the LuxS homologues (Barnard and Salmond, 2007; Toth *et al.*, 2004). The AI-2 system has not been fully elucidated in SREPP, so the focus here will be on the *N*-AHL system.

N-AHL signals are characterised by an invariant homoserine lactone ring to which a variable acyl side chain is attached (Barnard and Salmond, 2007). *Pba* is known as a class II strain as the QS signal used is 3-oxo-C6-HSL (*N*-(3-oxohexanoyl)-L-homoserine lactone), but very little amounts of 3-oxo-C8-HSL, and vice versa for class I strains (such as *Pcc* strains EC153 and SCC3193) (Barnard and Salmond, 2007). In both class I and II stains, these molecules are synthesised by a *luxI* homologue. Specifically in *Pba* strain SCRI1043, the homologue is known as *expI* (Barnard and Salmond, 2007; Toth *et al.*, 2004). *expI* is known to be involved in the regulation of plant cell wall degrading enzymes as well as virulence (Toth *et al.*, 2004).

The QS signal molecule is constitutively expressed in small amounts by each cell in the population and accumulates as the population increases and indicates their presence to other cells in the population (Toth *et al.*, 2003; Barnard and Salmond, 2007; Toth *et al.*, 2004). Upon reaching a critical population limit (a point at which the population is said to be “quorate”, estimated to be about 10^6 cells/mL) a synchronous, community wide response is indicated by the triggering of the expression of certain genes (Toth *et al.*, 2004; Toth *et al.*, 2003). Behaviour associated with *N*-AHL production includes pathogenesis, biofilm formation, antibiotic production, antibiotic production, exoenzyme production and plasmid production (Toth *et al.*, 2004). The classical explanation of how the QS signal is able to reach a high enough signal, is that the concentration of the signal is purely dependent upon the sheer number of cells in a population that are producing the signal within a confined space (Barnard and Salmond, 2007). Essentially this means that the level of signal will be proportional to the size of the population, if the signal is not allowed to escape quickly from the vicinity (Barnard and Salmond, 2007).

In order to respond to the QS signal, other bacteria in the population require a sensor which will detect and transmit the information into a cellular response. In SREPP, the QS signal is detected by a member of the LuxR family of DNA-binding transcriptional regulators (Barnard and Salmond, 2007). These are thought to bind directly to the *N*-AHL signal molecule and undergo a conformational change, which then modulates their DNA-binding ability. It is also possible for there to be more than one *luxR* homologue (Barnard and Salmond, 2007). In *Pba*, the *luxR* homologues are called *expR* and *virR* (Toth *et al.*, 2003), the latter being found to be central to the QS-dependent regulation of the production of various virulence determinants (Barnard and Salmond, 2007). *VirR* appears to function as a repressor only in the absence of the 3-oxo-C6-HSL signal. This means that at low cell densities, expression of *Nip*, *Svx* (both secreted pathogenicity factors) and plant cell wall degrading enzymes (cellulases, proteases, pectate lyases) is repressed by *VirR* and is alleviated upon accumulation of the signal molecule (Barnard and Salmond, 2007). *ExpR* is also only fully activated in the presence of the QS signal, and is a transcriptional activator which induces the production of exoenzymes (Toth *et al.*, 2003). In addition it has an auto-inducing effect on *expI*, resulting in accelerated production of pathogenicity factors (Toth *et al.*, 2003).

N-AHLs are particularly sensitive to pH and become unstable under alkaline conditions, undergoing rapid hydrolysis (Toth *et al.*, 2003; Barnard and Salmond, 2007). Thus the 3-oxo-C6-HSL QS signal acts as a built-in pH sensor (Barnard and Salmond, 2007). This is possibly an element in a plant's defences against a SREPP infection since one of the first plant responses to infection is an alkalinisation of the infection site to a pH greater than 8.2 (Toth *et al.*, 2003; Barnard and Salmond, 2007).

While QS is obviously a key factor in the control of gene expression, it is not the only regulatory system. For example, full expression of the plant cell wall degrading enzymes requires the presence of plant cell wall breakdown products, such as 2-keto-3-deoxygluconate (KDG). This is due to the presence of the *KdgR* regulator, which binds to the DNA target sites of these genes (Barnard and Salmond, 2007). The repression is abolished by the presence of KDG, and thus expression of plant cell wall degrading enzymes is induced. The production of these enzymes being under both QS and *KdgR* control makes sense since being quorate does not necessarily mean that the bacteria are located anywhere near plant tissue, and the secretion of plant cell wall degrading enzymes without the presence of a substrate would be a waste of resources (Barnard and Salmond, 2007). In addition, these plant cell wall degrading

enzymes are potent activators of plant defence responses, so if the population was small, it would make sense to postpone production until absolutely required (Barnard and Salmond, 2007). Similarly, the action of the pectinases on pectin in the cell walls results to a range of breakdown products which in turn results in the induction of exoenzymes, thus acting as a positive feedback mechanism to accelerate exoenzyme production (Toth *et al.*, 2003).

QS may be more subtle than previously thought as potato plants that were genetically modified to produce *N*-AHL were found to be more susceptible to infection (Barnard and Salmond, 2007; Toth *et al.*, 2004). In addition QS can be stimulated by other factors such as available oxygen and nitrogen, temperature, iron deprivation, plant degradation intermediates, DNA-damaging agents, among many others (Toth *et al.*, 2003). Although SREPPs do not need QS for successful colonisation of the plant host, it is required for disease development in suitable environmental conditions through the induction and delivery of the major virulence factors (Toth and Birch, 2005).

The QS regulatory network allows the bacteria to couple the accumulation of a small, diffusible signalling molecule to the regulation of gene expression. This system serves as an integration point for a variety of different regulatory networks, and simplifies the components of the gene regulatory system (Barnard and Salmond, 2007). It also reduces the number of transcriptional regulators that are required to bring about the required changes in expression of downstream target genes (Barnard and Salmond, 2007). Separate regulators, such as KdgR can be layered into the regulatory hierarchy to allow differential regulation of subsets of genes in response to different environmental cues (Barnard and Salmond, 2007). Quorum sensing thus represents an elegant solution to the complex problem that bacteria face – that of sensing and responding appropriately to a large number of external signals (Barnard and Salmond, 2007).

3.5.3 MADIBA *Pectobacterium atrosepticum* data analysis

Data for this analysis was from an *expI* mutant experiment in Pba strain SCRI1043. This experiment was performed at the Scottish Crop Research Institute (SCRI) and the data were kindly provided by Dr Ian Toth and Dr Leighton Pritchard. *expI* has been found to be central to quorum sensing (QS), so the aim was to identify genes involved in this biological phenomenon. ExpI synthesises the AHL, the QS signalling molecule, and it has been found that virulence is significantly reduced in strains with a mutation in the *expI* gene (Liu *et al.*, 2008). In addition, virulence was restored following complementation with the *expI* gene *in*

trans (Liu et al., 2008). In the previous analyses, the data were pre-processed using some form of clustering. However, the *expI* data had not clustered and thus STEM was used to group the genes together, prior to analysing selected clusters with MADIBA.

Short Time-series Expression Miner (STEM) (Ernst and Bar-Joseph, 2006) is a Java program designed specifically to analyse short time series microarray gene expression data (3-8 time points). In time series experiments, thousands of genes are being profiled simultaneously while the number of time points is small. In these cases, many genes may randomly have the same expression pattern. Furthermore, there are few full time series repeats to increase the statistical power. Popular current methods for time series data, such as clustering and self-organising maps (SOM), ignore the temporal dependencies among successive time points. STEM performs an analysis by taking advantage of the fact that there are a large number of genes and a number of time points that is too small to identify significant temporal expression profiles and the genes associated with these profiles (Ernst and Bar-Joseph, 2006). The clustering algorithm selects a set of distinct and representative temporal expression profiles (called “model profiles”), which are selected independently of the data. The algorithm then assigns each gene passing a filter to the model profile that most closely matches the gene’s expression profile, as determined by a correlation coefficient. The filter conditions used include if a gene does not show a sufficient response to experimental conditions (Minimum Absolute Expression Change); there are too many missing values (Maximum Number of Missing Values); or the gene expression pattern over repeats is too inconsistent (Minimum Correlation between Repeats). Since the model profiles were selected independently from the data, the algorithm is able to determine which profiles have a statistically significant higher number of genes assigned by using a permutation test (Ernst and Bar-Joseph, 2006). Significant model profiles can be analysed by themselves or grouped together based on similarity to form clusters of significant profiles.

The *expI* experiment consisted of a time course with 6 time points, namely 0, 0.5, 4, 12, 20, and 72 hours post-infection (hpi), with inoculations performed on sterilised potato plants. When these data were applied to STEM, only 1127 out of a total of 5051 genes passed the filter, using the default filter options. These default values were Minimum Absolute Expression Change=1, Maximum Number of Missing Values=0, and Minimum Correlation between Repeats=0. It is possible that the Minimum Absolute Expression Change variable was set too high which resulted in most of the genes being discarded. The expression values

were normalised using log normalisation ($\log_2(v_i) - \log_2(v_0)$) and the data were clustered using the STEM clustering method. Interestingly, STEM is fairly sensitive to the number of decimal places in the data. When the data was rounded to two decimal places (the original data had seven decimal places) a different clustering result was obtained. Although the only difference was a slight rearrangement in the significance of the clusters, this could mean that the STEM clustering method is not that robust. However it could be argued that a microarray scanner is not sensitive enough to read down to seven decimal places. The result of the clustering using all decimal places is shown in Figure 3.11.

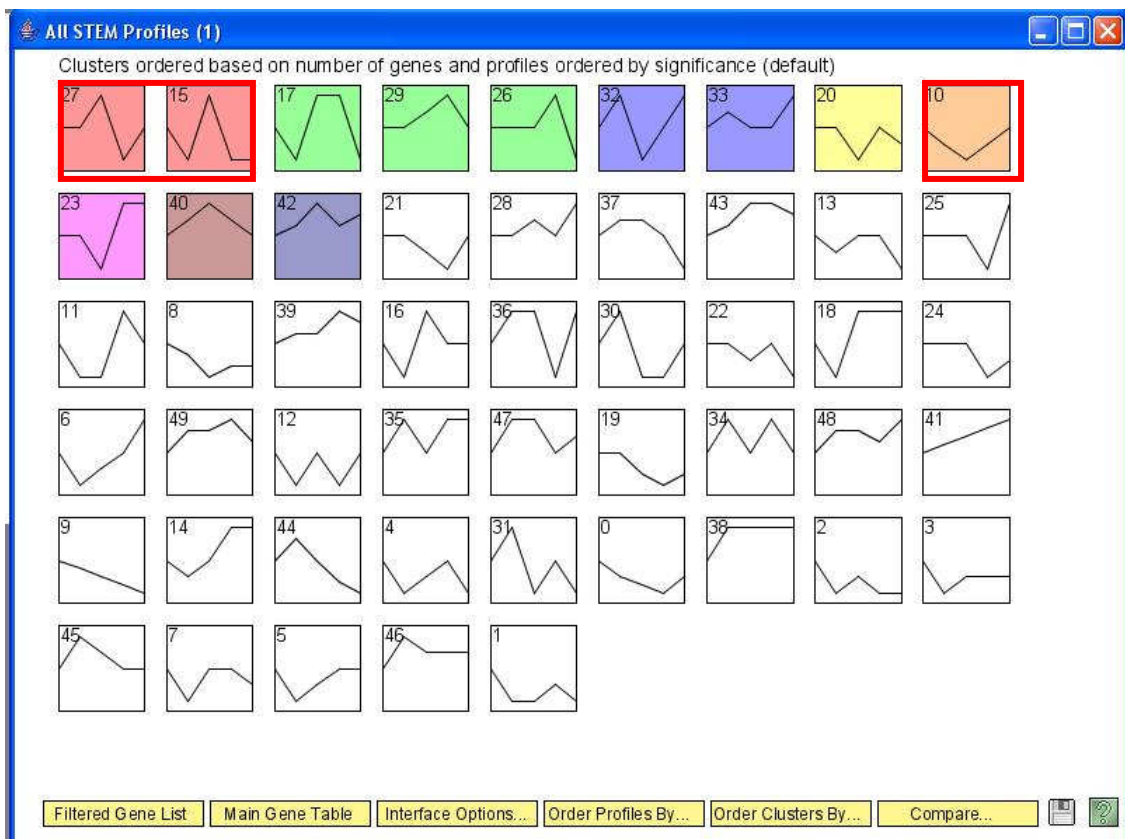


Figure 3.11: Screenshot of the model profiles overview interface of STEM (Ernst and Bar-Joseph, 2006), using the *Pba* data. Each box represents a profile, and the number in the corner is the profile ID number. The coloured profiles have a statistically significant number of genes assigned, with non-white profiles of the same colour representing profiles grouped into a single cluster. Clicking on a profile box brings up detailed information about the profile (Figure 3.13). The profiles surrounded by a red box were analysed in this section.

Two analyses were performed on the STEM clusters. First the most significant cluster was analysed using MADIBA as a proof of concept to show that MADIBA could be used to

analyse *Pba* data. Next, a cluster was selected to answer the biological question of which genes are affected by ExpI, and more broadly which genes are involved in quorum sensing.

STEM cluster 1 was analysed as a proof of concept for the *Pba* data analysis, and consisted of profiles 27 and 15, with a total of 222 genes. After submitting the gene names from this cluster to MADIBA, it could be seen that there were common annotations involving phage regulatory proteins and type III secretion proteins. Using the Metabolic Pathways module, it was found that there were numerous enzymes present in the purine (10 enzymes) and pyrimidine (4 enzymes) metabolism pathways. In the purine metabolism pathway, most of the enzymes in the cluster were involved in the metabolism of GNP (GMP, GDP and GTP) as well as some terms involved in DNA and RNA elongation. These predictions are fairly reliable as the enzymes were coloured red or yellow, indicating that two or three annotations agreed respectively. The *p*-value for this pathway was not particularly significant (*p*-value = 0.1722) but was mostly likely because the set of enzymes only made up a small part of a large map. A similar prediction could be seen in the pyrimidine metabolism pathway, except involving UNP and CNP. Other pathways present with multiple enzymes present included glycolysis and gluconeogenesis (2 enzymes), starch and sucrose metabolism (3 enzymes), biotin metabolism (3 enzymes).

After analysing the data with the Gene Ontology module, it was found that many of the terms in the molecular function ontology were related to replication, such as alpha DNA polymerase activity (GO:0003889), as well as rRNA binding (GO:0019843) and several tRNA activity related annotations (GO:0016439 and GO:0008883). This result was further corroborated in the biological process ontology where the most significant terms were mRNA polyadenylation (GO:0006378) and catabolism (GO:0006402), translation (GO:0006412), RNA processing (GO:0006396), tRNA modification and aminoacylation (GO:0006400, GO:0006426, and GO:0006429) and chromatin silencing (GO:0006342). Other terms included defence responses (GO:0006952) and GTPase mediated signal transduction (GO:0007264). The cellular component ontology added additional support for the cluster involving replication with the localisation primarily on the ribosome (GO:0005840, GO:0015934), or nuclease complexes (GO:0005971, GO:0009318 and GO:0030529).

The Transcription Regulation module did not identify any motifs that are known to be involved in QS. However, many of the genes with similar functions in *Pba* (and bacteria in general) belong to a single operon, and thus would have only one transcription factor binding

site to induce the transcription of all of the genes. Since tools such as RSAT and TRANSFAC try to identify over-represented motifs, such approaches may not be appropriate for prokaryotes that express genes under operonic control. Nonetheless, some of the motifs that were found may potentially be involved in other molecular processes. The top motif found by oligo-analysis was GGCTGA (e-value= 7.7×10^{-3}), and the top result found by dyad-analysis was TCAN(1)CCA (e-value=0.041).

Using the Chromosomal Localisation module, a circular diagram is drawn representing the circular chromosome of *Pba* (Figure 3.12). In it, it is possible to see that some of the genes are located close to each other in an operon-type structure.

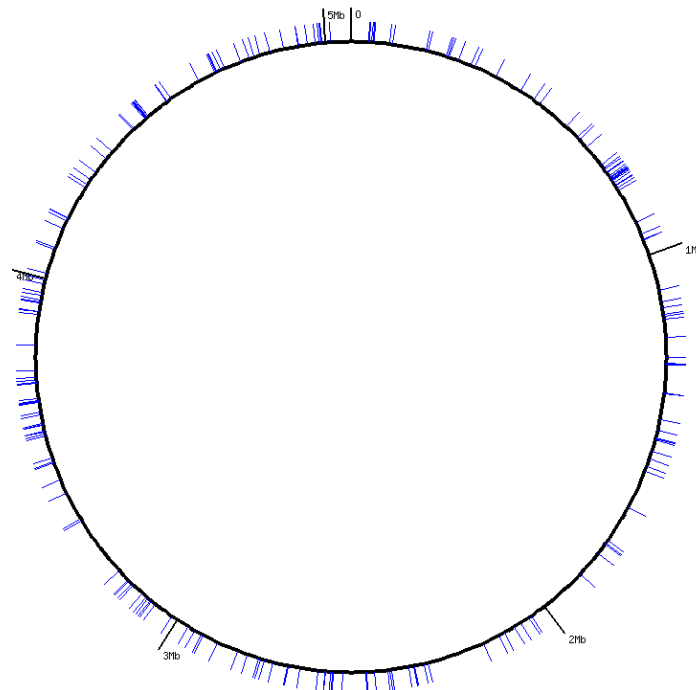


Figure 3.12: Output of the Chromosomal Localisation for *Pba*. The circle represents the circular chromosome of *Pba*, and the blue lines represent a gene. The genes shown are the genes from the STEM cluster 1. Some genes can be seen to be located close together, and may be a possible indication of co-expression.

In the Pectobacterium Specific module, it was found that most of the annotations matched those for *Dickeya dadantii* (formally *Erwinia chrysanthemi*). This similarity is due to the two organisms being closely related to other, and possibly also because a gene's annotation in one organism was used to annotate a gene in the other.

Since the most significant cluster did not reveal anything particularly interesting relating to QS, it was decided to analyse the clusters that one would expect from genes under control of ExpI. Since this experiment was performed on an *expI* mutant, genes under the control of ExpI would be down-regulated when compared to wild-type plants. Profile 10 (cluster 5) was selected as it was the most significant profile which matched this requirement of generally down-regulated genes (Figure 3.13). This group of genes contained 129 genes, although when inspecting the annotations, it was found that approximately one third of the genes do not have a functional annotation, being marked as hypothetical.

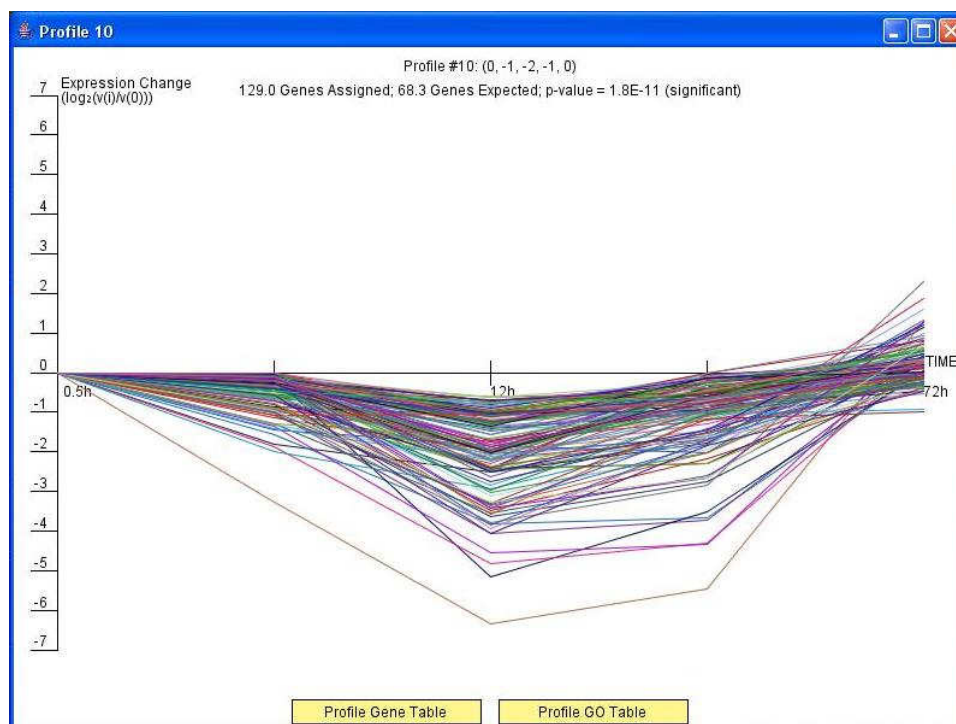


Figure 3.13: Profile 10 of the *Pba* data. Each line represents the expression of a gene across the time points. Along the top are the statistics on the number of genes assigned to the profile, the number of genes expected and the enrichment *p*-value. The time points are 0.5, 4, 12, 20, and 72 hpi. 0hr is not shown as it was used as the control in the normalisation process.

When analysing these data in the Metabolic Pathways module, it was found that the only significant genes were involved in the starch and sucrose metabolism, and the nitrogen metabolism pathways (4 enzymes each). It is possible that these pathways are induced during QS to produce energy for the infection process.

In the Gene Ontology analysis, interestingly the molecular function ontology showed that cellulase activity (GO:0008810), as well as hydrolase activity, acting on ester bonds (GO:0016788) were significant terms. These terms are noteworthy as an aspect of QS is the

production of large amounts of plant cell wall degrading enzymes. In the biological process, several terms relating to QS were found, most notably the quorum sensing term itself (GO:0009372), in addition to further terms relating to cell wall degrading enzyme activity, such as polysaccharide catabolic process (GO:0000272). Also present were terms involved in the secretion and transport of these exoenzymes, including extracellular transport (GO:0006858) and protein secretion by the type II secretion system (GO:0015628). This last result was confirmed in the cellular component ontology with the terms proteinaceous extracellular matrix (GO:0005578) and type II protein secretion system complex (GO:0015627). It is known that the plant cell wall degrading enzymes delivered by *Pba*, mainly cellulases and pectinases, are delivered using this secretion system (Toth and Birch, 2005).

Analysing this cluster with the Transcription Regulation module, the oligo-analysis program of RSAT found that the motifs ATAAAT, ATAATA and GATAAA were significant (e-value = 0.088, 0.49 and 0.9 respectively). The dyad-analysis program only found one motif, AATN(6)AAT, that was significant (e-value = 0.57) (Figure 3.14). While these motifs did not match any motifs that are known to be involved in QS, such as the *lux* or *esaR* boxes (von Bodman *et al.*, 2003), these still could be involved in the quorum sensing process.

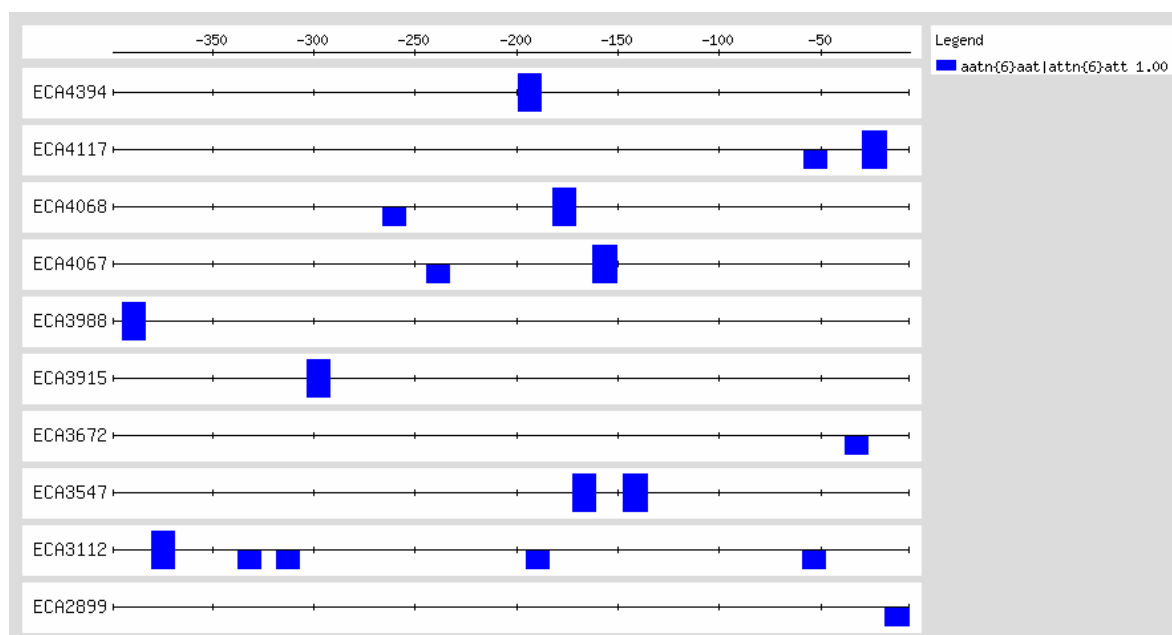


Figure 3.14: Portion of the results as obtained by the dyad-analysis program of RSAT. Shown is the only motif found by dyad-analysis, and is drawn showing the location of that motif in the upstream region of the genes.

3.5.4 Conclusion

The data that were used in this analysis was from an *expI* mutant in *Pba*, which is known to be involved in quorum sensing. STEM was used to cluster the data, and two clusters were selected for analysis with MADIBA. The most significant cluster (cluster 1) was used as a proof of concept to ensure that MADIBA worked correctly for *Pba* data. It was found that in this cluster most of the genes were involved in transcription and translation. The profiles in this cluster showed down-regulation at the later time points in comparison to the wild-type, so it is possible that gene transcription is activated late in the infection when the bacterium has become quorate, signalling the production of plant cell wall degrading enzymes.

Profile 10 was also analysed, as it fitted the profile of genes that are under the control of ExpI, that is, are down-regulated compared to the wild-type. A MADIBA analysis of the genes in this cluster showed significant terms involved in quorum sensing, including the production of cell wall degrading enzymes and transport of these exoenzymes through the type II secretion system. Since genes in profile 10 showed significant involvement in QS, further examination of the genes in this cluster may reveal more genes that are implicated in QS, especially those that are currently annotated as hypothetical.

3.6 Concluding remarks

In this chapter, MADIBA was applied to numerous datasets to show its functionality and ease of use. Various organisms were analysed, namely *Plasmodium falciparum*, *Arabidopsis thaliana*, rice, and *Pectobacterium atrosepticum*. The Plasmodium data were used as a proof of concept, and was analysed using published microarray data. The Arabidopsis section focussed on plant defences, with data from the MPPI group where *A. thaliana* plants were infected with *Ralstonia solanacearum*. A method termed PCA Expression Comparer was developed to compare the expression of a submitted dataset to other experimental treatments, and showed that as expected, that the susceptible interaction had increased JA induction. The converse was also shown where the resistant interaction showed induced SA signalling, with a match to a BTH treatment. In the rice section, published data on a BTH experiment on rice plants was used to show functionality. As an extension to the rice section, pearl millet data from the MPPI group was used, where SA and MeJ treatments were used to identify resistance to the rust fungus. Pearl millet is a related cereal and showed that MADIBA could be used on other monocotyledonous crops. Finally data from an *expI* mutant of *Pectobacterium atrosepticum* were used to identify which genes are involved in quorum sensing. Using MADIBA, a cluster

containing several terms involved in quorum sensing and plant cell wall degrading enzymes was identified.

Chapter 4 – Concluding Discussion

While numerous other tools similar to MADIBA, such as WebGestalt, FatiGO and GoMiner exist, MADIBA differs in that it has a wider range of analyses which can be performed in an integrated fashion, for example, it performs a GO analysis as well as a Transcription Regulation analysis. In addition, MADIBA is unique in the organisms it is able to analyse – a eukaryotic pathogen (*Plasmodium falciparum*), a bacterium (*Pectobacterium atrosepticum*), a monocotyledonous plant (*Oryza sativa*) and a dicotyledonous plant (*Arabidopsis thaliana*).

MADIBA has been designed to be generic and easily expandable, so that any new organisms that are required by the community can readily be incorporated into the database, with only a fully annotated genome necessary. For example, MADIBA could be applied to potato, the host of *Pba*, once this genome has been sequenced, to identify host responses. The current analysis modules are continually being improved to assist the user in identifying the reasons for the co-expression of a set of genes. Also, due to its modular nature, any new analysis can easily be added to MADIBA at a later stage. Since it is a web application, this makes MADIBA platform-independent and can be accessed from anywhere in the world. Furthermore, the database can be updated, so that the latest information is available to the user. MADIBA is highly dependent on the quality of the genomes' annotations, so as the annotations are improved, so will the results returned by MADIBA. Furthermore, as the genome annotations are revised, it is important to update the data within MADIBA, and will be done in a semi-automated manner using pre-built Python scripts.

All the statistics performed on the data are analysed in terms of the entire genome. However, since whole genome microarray slides are not always used, a proposed improvement is to analyse the data in terms of only the genes that were on the slide. Also, with the increased number of statistical methods being adopted, the variety of statistical analyses available could be widened, such as by including GSEA (Subramanian *et al.*, 2005) (through R) or rank tests, and so, provide a greater level of flexibility for the user.

The number of genes in the cluster that are initially submitted is important, as this may have an affect on the quality of the results returned by MADIBA. If too many genes are submitted, a significant result may get masked by other possibly insignificant results. However, if too few genes are submitted, the calculations and statistics may indicate that the result is not

significant at all. This is particularly evident in the Gene Ontology module, where the size of the cluster affects the statistics.

In the Metabolic Pathways module, the significance of a pathway is determined by using Fisher's exact test, which determines the membership of an enzyme in the pathway and the cluster. While this result can be useful in determining the importance of a metabolic pathway, its interpretation can be difficult. For example, in a large metabolic pathway, if enzymes are only found in a small section of the pathway, the *p*-value will be insignificant. However, that portion of the pathway may in fact have an important function, for example, a specific aspect of glycolysis.

In the Organism Specific module, orthology was inferred by using the reciprocal best BLAST hit. Since this is not the most accurate or reliable method for determining orthology, the implementation of tools such as Ortholuge (Fulton *et al.*, 2006) and GreenPhyl (Conte *et al.*, 2007) are being considered. These tools take into account phylogenetic information in addition to sequence similarity, increasing the confidence of an orthology prediction. Also, in general orthologous genes had similar annotations. While it is possible that orthologous genes have similar functions, such as in rice and *Arabidopsis thaliana* which have a large amount of shared genes (Sasaki and Burr, 2000), it is also likely that the *A. thaliana* data would have been used to putatively annotate the rice genome. The annotation would then have been putatively transferred from the one organism to the other, without any experimental evidence. A similar situation is seen with *Pba* and *Dickeya dadantii*. Thus, in its current format, this module is not particularly informative, and while some details can be gained from homologous and orthologous proteins, this is not a particularly valuable analysis to perform. Further improvements could be applied to make this module more useful, such as by a feature to detect leucine rich repeats (LRRs), which are important in plant defence (Di Matteo *et al.*, 2003), or host-pathogen interactions. Some such improvements were added in the Arabidopsis Characteristics module with the implementation of the DRASTIC database and the PCA Experiment Comparer.

MADIBA is useful in the study of *Plasmodium falciparum*, especially as the gene regulation in this parasite is poorly understood (Daily *et al.*, 2007). MADIBA successfully identified useful genes and pathways that could possibly be used in gaining knowledge on the molecular workings of this parasite.

Diseases of plants are an enormous problem for agriculture worldwide, with effects ranging from spots on leaves to catastrophes where entire fields of crops are destroyed. An example of the latter includes the potato blight (caused by *Phytophthora infestans*) that struck Europe in the 1840s (Strange and Scott, 2005). The high reliance of the Irish population on potatoes resulted in about a million Irish dying of starvation and caused the emigration of more than a million people. Thus it is obvious that crop species are a vital source of human nutrition, and understanding the mechanisms that plants use to defend themselves against pathogens may lead to novel strategies to enhance disease resistance in crops. A tool that matches the expression profile of a group of genes in a study to the expression profiles in publicly available microarray data thus becomes of value.

Once it is determined which genes are induced and which pathway is involved, this information can be used to develop potential crop protection strategies. Such strategies could include genetically modifying a transcription factor that is master regulator of a certain pathway, inserting an antimicrobial protein that is highly induced by a pathway, or use a spray that activates natural plant defences. An example of the last is spraying plants with benzothiadiazole (BTH), a SA analogue, which induces the SA pathway (Murray *et al.*, 2002). In this way, an effective means to enhance disease resistance in agriculturally important crops can be found. However, the extent of the conservation between these basic common pathways in *Arabidopsis thaliana* and other plants is unknown, although recent evidence suggests that the level of conservation of down-stream defence signalling components may be substantial (Anderson *et al.*, 2005).

In the *Arabidopsis* analyses, MADIBA was used to investigate the signalling pathways in *Arabidopsis thaliana* in both the resistant and susceptible interactions when infected with *Ralstonia solanacearum* (bacterial wilt) (Naidoo, 2008). The DRASTIC and PCA Experiment Comparer analyses were implemented as a subsection in the *Arabidopsis* Characteristics module and used to analyse the data from these experiments. It was expected that the data from the resistant interaction would match experiments related to SA signalling (Deslandes *et al.*, 2003; Noutoshi *et al.*, 2005). While a BTH treatment did match the data, confirming the hypothesis that SA is involved in the resistant interaction with *R. solanacearum*, several other diverse experiments were found including responses to light and cold, as well as several cell cycle experiments. This could suggest that the resistant response induces abiotic stress responses and possibly affect the cell cycle. Possible future experiments could involve

infection at different growth stages, or in conjunction with an abiotic stress, such as cold, to determine if these factors affect the defence response.

Conversely, in the susceptible interaction with *R. solanacearum*, JA-signalling was expected to be induced due to impaired SA-signalling in the susceptible plants and the general antagonistic nature of the two signalling pathways. This was indeed the case, and was further confirmed by the MADIBA GO analysis. These data also matched a susceptible interaction with *Pseudomonas syringae* DC3000 in a WRKY knockout mutant. The WRKY knockout possibly made the plant more susceptible since WRKY transcription factors are important for *NPRI* expression, a key regulatory protein in the SA signalling pathway (Thatcher *et al.*, 2005).

While the DRASTIC database did not provide much information, it could still be valuable for a researcher to easily identify how certain genes are regulated in response to various treatments. The PCA Experiment Comparer allowed a comparison of expression profiles from a large number of experiments. This analysis is particularly powerful as a user with the log₂-ratios for a cluster of genes is able to determine what other experimental conditions can result in a similar expression profile.

In the rice analysis, data from rice treated with BTH was used as a proof of concept. In addition, data from a comparison of salicylic acid and methyl jasmonate treatments in pearl millet prior to infection with the rust fungus was used. In the experiment, it was found that the SA treatment resulted in increased resistance to the pathogen, whereas MeJ did not confer any significant resistance, suggesting that the SA signalling is involved in response to rust infection (Crampton, 2006). MADIBA was used to analyse the responsive genes after SA treatment, as well as those that were responsive after MeJ treatment. In the SA responsive genes, several terms related to defence were found, in addition to a large number of genes involved in photosynthesis, possibly suggesting a relation between defence signalling and photosynthesis. While it has been determined that light and the resultant photosynthetic processes can impact the susceptibility of plants to infections (Bechtold *et al.*, 2005), further investigations could involve determining if there are differences in susceptibility when infection occurs during the day or at night. A MADIBA analysis on the MeJ responsive genes confirmed that the responsive genes are involved in JA- and ET-signalling, although the GO analysis also suggested a considerable involvement in stress responses. Experiments to test this could involve treatment with JA or ET prior to some form of stress, such as drought.

Pectobacterium atrosepticum recently became the first fully sequenced enterobacterial plant pathogen, and currently remains the only published genome sequence for this group of pathogens (Bell *et al.*, 2004). With the availability of whole genome microarrays for *Pba*, it is possible to investigate quorum sensing to identify previously undefined gene sets linked to this major pathogenicity regulon. MADIBA was used to analyse microarray data from an *expl* mutant and a cluster was identified that contained several quorum sensing related terms. However, many of the other members of the cluster did not possess a functional annotation, and these unknown genes could be studied experimentally to possibly reveal greater detail about the mechanisms of quorum sensing in enterobacterial plant pathogens. MADIBA could also be used to further analyse data to help target key metabolic and regulatory pathways that are affected during the infection process, as well as identify genomic regions and functionally related pathways in *Pba* that are activated or suppressed during disease development.

Thus, it is hoped that MADIBA will make analysing data easier for researchers, so that less time is spent examining the data and more time deriving conclusions. MADIBA can assist in assigning putative functions to unknown cereal genes (from GO data or orthologous annotations) and providing data on the protein product, enzymatic pathways, common promoter elements, transcription regulation and localisation of genes on the chromosomes. Such information can be utilised in identifying genes with quantitative traits for use as a functional marker (such as the *Dwarf8* gene in maize from which a functional marker can be developed for plant height and flowering time (Andersen and Lubberstedt, 2003)), or key pathway regulators, and so be useful in understanding an organism's biology.

Summary

Microarray technology makes it possible to identify changes in gene expression of an organism, under various conditions. The challenge to researchers that employ microarray expression profiling is that once pre-processing is completed, and a cluster of co-expressed genes obtained, is to derive biological meaning from this data. Data mining is thus essential for deducing significant biological information such as the identification of new biological mechanisms or putative drug targets. While many algorithms and software have been developed for analysing gene expression, the extraction of relevant information from experimental data is still a substantial challenge, requiring significant time and skill.

MADIBA (MicroArray Data Interface for Biological Annotation) facilitates the assignment of biological meaning to gene expression clusters by automating the post-processing stage. A relational database has been designed to store the data from gene to pathway for *Plasmodium falciparum*, *Oryza sativa* (rice), *Arabidopsis thaliana*, and *Pectobacterium atrosepticum* (*Pba*).

As input, the user submits a cluster of genes, either the gene identifiers or the gene sequences. Tools within the web interface allow rapid analyses for the identification of the Gene Ontology terms relevant to each cluster; visualising the metabolic pathways where the gene products are implicated, their genomic localisations, putative common transcriptional regulatory elements in the upstream sequences, and an analysis specific to the organism being studied. The user has the option of outputting selected results of the analyses, either in PDF or plain text formats.

MADIBA is an integrated, online tool that will assist researchers in interpreting their results and understand the meaning of the co-expression of a cluster of genes. Functionality of MADIBA was used to analyse a number of gene clusters from several experiments – expression profiling of the *Plasmodium falciparum* life cycle, a *Ralstonia solanacearum* infection of *Arabidopsis thaliana*, a rice treatment with BTH, a millet SA- and MeJ-treatment experiment, and an *expI* mutant experiment in *Pectobacterium atrosepticum*. Data from the *Plasmodium falciparum* and rice were used to illustrate MADIBA's functionality. For the *A. thaliana* analyses, the DRASTIC database was implemented to identify how genes respond to various treatments. In addition, a method named PCA Experiment Comparer was developed, which compares the expression values of the numerous experiments in NASCArrays. Using

the *A. thaliana*-*R. solanacearum* interaction data several related experiments matched in both the susceptible and resistant interactions. In the millet analyses, besides defence related genes being identified, several genes also involved in photosynthesis were found, possibly suggesting a relation between light and defence signalling. The *Pba* data identified genes involved in quorum sensing, as well as some associated genes with no known function that may also be related to this regulatory process.

With the advent of whole genome microarray chips and an increasing number of organisms being sequenced, tools such as MADIBA will become even more significant in understanding the underlying biology. MADIBA provides access to several genomic data sources and analyses, allowing users to quickly annotate and visualise the results. MADIBA is freely available and can be accessed at <http://www.bi.up.ac.za/MADIBA/>.

References

- (1) AmiGO [<http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>]
 - (2) GD Graphics Library [<http://www.boutell.com/gd>]
 - (3) Graphviz - Graph Visualization Software [<http://www.graphviz.org/>]
 - (4) MADIBA [<http://www.bi.up.ac.za/MADIBA>]
 - (5) R Programming Language [<http://www.r-project.org/>]
 - (6) ReportLab [<http://www.reportlab.org/>]
 - (7) ZGRViewer [<http://zvtm.sourceforge.net/zgrviewer.html>]
 - (8) An Introduction to the Gene Ontology [<http://www.geneontology.org/GO.doc.shtml>]
 - (9) Beautiful Soup [<http://www.crummy.com/software/BeautifulSoup>]
 - (10) BioPython [<http://biopython.org>]
 - (11) Database Resource for the Analysis of Signal Transduction in Cells (DRASTIC) [<http://www.drastic.org.uk>]
 - (12) DHTML JavaScript Tooltips [http://www.walterzorn.com/tooltip/tooltip_e.htm]
 - (13) FatiGO [<http://fatigo.bioinfo.cipf.es/>]
 - (14) Genevestigator [<https://www.genevestigator.ethz.ch/>]
 - (15) GeneXPress [<http://GeneXPress.stanford.edu>]
 - (16) Genome2D [<http://molgen.bio1.rug.nl/genome2d>]
 - (17) GoMiner [<http://discover.nci.nih.gov/gominer>]
 - (18) MAPPFinder and GenMAPP [<http://www.GenMAPP.org>]
 - (19) NetworkX [<https://networkx.lanl.gov>]
 - (20) Nottingham Arabidopsis Stock Centre Arrays (NASCArrays) [<http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl>]
 - (21) pcaMethods - Bioconductor library [<http://www.bioconductor.org/packages/bioc/html/pcaMethods.html>]
 - (22) PyGreSQL [<http://www.pygresql.org/>]
 - (23) Python Imaging Library [<http://www.pythonware.com/products/pil/>]
 - (24) RPy [<http://rpy.sourceforge.net/>]
 - (25) Scatterplot 3D - R package [<http://cran.r-project.org/web/packages/scatterplot3d/index.html>]
 - (26) TAIR gene nomenclature [<http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp>]
 - (27) TIGR gene nomenclature [http://www.tigr.org/tdb/e2k1/osa1/tigr_gene_nomenclature.shtml]
 - (28) WebGestalt [<http://bioinfo.vanderbilt.edu/webgestalt/>]
- Aickin, M. and Gensler, H. (1996) Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health*, **86**, 726-728.
- Al Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578-580.
- Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, **34**, W472-W476.
- Al-Shahrour, F., Minguez, P., Vaquerizas, J. M., Conde, L. and Dopazo, J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res*, **33**, W460-W464.
- Andersen, J. R. and Lubberstedt, T. (2003) Functional markers in plants. *Trends Plant Sci*, **8**, 554-560.

- Anderson, J. P., Thatcher, L. F. and Singh, K. B. (2005) Plant defence responses: conservation between models and crops. *Funct Plant Biol*, **32**, 21-34.
- Anderson, L. K., Gong, L. X. and Dong, X. (2004) scv1 is a suppressor of cpr5-mediated disease resistance. At: *15th International Conference on Arabidopsis Research*.
- 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. *Nat Genet*, **25**, 25-29.
- Baerends, R. J., Smits, W. K., de, J. A., Hamoen, L. W., Kok, J. and Kuipers, O. P. (2004) Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data. *Genome Biol*, **5**, R37.
- Barnard, A. and Salmond, G. (2007) Quorum sensing in Erwinia species. *Anal Bioanal Chem*, **387**, 415-423.
- Beazley, D. M. (2006) In: *Python Essential Reference*, Sams Publishing.
- Bechtold, U., Karpinski, S. and Mullineaux, P. M. (2005) The influence of the light environment and photosynthesis on oxidative signalling responses in plant-biotrophic pathogen interactions. *Plant Cell Environ*, **28**, 1046-1055.
- Beissbarth, T. and Speed, T. P. (2004) GStat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464-1465.
- Bell, K. S., Sebahia, M., Pritchard, L., Holden, M. T. G., Hyman, L. J., Holeva, M. C., Thomson, N. R., Bentley, S. D., Churcher, L. J. C., Mungall, K., Atkin, R., Bason, N., Brooks, K., Chillingworth, T., Clark, K., Doggett, J., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Norbertczak, H., Ormond, D., Price, C., Quail, M. A., Sanders, M., Walker, D., Whitehead, S., Salmond, G. P. C., Birch, P. R. J., Parkhill, J. and Toth, I. K. (2004) Genome sequence of the enterobacterial phytopathogen Erwinia carotovora subsp. atroseptica and characterization of virulence factors. *Proc Natl Acad Sci U S A*, **101**, 11105-11110.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc*, **57**, 289-300.
- Bidaut, G. and Ochs, M. F. (2004) ClutrFree: cluster tree visualization and interpretation. *Bioinformatics*, **20**, 2869-2871.
- Bowling, S. A., Clarke, J. D., Liu, Y., Klessig, D. F. and Dong, X. (1997) The cpr5 mutant of Arabidopsis expresses both NPR1-dependent and NPR1-independent resistance. *Plant Cell*, **9**, 1573-1584.
- Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R. and Gortlach, J. (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell*, **13**, 1499-1510.
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710-3715.
- Buchel, A. S., Brederode, F. T., Bol, J. F. and Linthorst, H. J. (1999) Mutation of GT-1 binding sites in the Pr-1A promoter influences the level of inducible gene expression in vivo. *Plant Mol Biol*, **40**, 387-396.
- Buell, C. R. (2002) Current status of the sequence of the rice genome and prospects for finishing the first monocot genome. *Plant Physiol*, **130**, 1585-1586.
- Button, D. K., Gartland, K. M., Ball, L. D., Natanson, L., Gartland, J. S. and Lyon, G. D. (2006) DRASTIC--INSIGHTS: querying information in a plant gene expression database. *Nucleic Acids Res*, **34**, D712-D716.

- Carreira-Perpiñán, M. Á. (1997) A review of dimension reduction techniques. *Technical Report CS-96-09*.
- Causton, H. C., Quackenbush, J. and Brazma, A. (2003) In: *Microarray gene expression data analysis: a beginner's guide*, Blackwell Publishing.
- Chung, H. J., Kim, M., Park, C. H., Kim, J. and Kim, J. H. (2004) ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*, **32**, W460-W464.
- Clare, A., Karwath, A., Ougham, H. and King, R. D. (2006) Functional bioinformatics for *Arabidopsis thaliana*. *Bioinformatics*, **22**, 1130-1136.
- Claudiel-Renard, C., Chevalet, C., Faraut, T. and Kahn, D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucl Acids Res*, **31**, 6633-6639.
- Conte, M. G., Gaillard, S., Lanau, N., Rouard, M. and Perin, C. (2007) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res*,
- Crabb, B. S. and Cowman, A. F. (1996) Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. *Proc Natl Acad Sci U S A*, **93**, 7289-7294.
- Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*, **32**, D575-D577.
- Crampton, B. (2006) Elucidation of defence response mechanisms in pearl millet. *PhD Thesis*.
- D'haeseleer, P. (2005) How does gene expression clustering work? *Nat Biotechnol*, **23**, 1499-1501.
- Daily, J. P., Scanfeld, D., Pochet, N., Le Roch, K., Plouffe, D., Kamal, M., Sarr, O., Mboup, S., Ndir, O., Wypij, D., Levasseur, K., Thomas, E., Tamayo, P., Dong, C., Zhou, Y., Lander, E. S., Ndiaye, D., Wirth, D., Winzeler, E. A., Mesirov, J. P. and Regev, A. (2007) Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients. *Nature*, **450**, 1091-1095.
- de Torres-Zabala, M., Truman, W., Bennett, M. H., Lafforgue, G., Mansfield, J. W., Rodriguez, E. P., Bogre, L. and Grant, M. (2007) *Pseudomonas syringae* pv. tomato hijacks the *Arabidopsis* abscisic acid signalling pathway to cause disease. *EMBO J*, **26**, 1434-1443.
- Dechering, K. J., Kaan, A. M., Mbacham, W., Wirth, D. F., Eling, W., Konings, R. N. and Stunnenberg, H. G. (1999) Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol Cell Biol*, **19**, 967-978.
- Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. and Lempicki, R. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, **4**, R60.
- Deslandes, L., Olivier, J., Peeters, N., Feng, D. X., Khounlotham, M., Boucher, C., Somssich, I., Genin, S. and Marco, Y. (2003) Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. *Proc Natl Acad Sci U S A*, **100**, 8024-8029.
- Di Matteo, A., Federici, L., Mattei, B., Salvi, G., Johnson, K. A., Savino, C., De Lorenzo, G., Tsernoglou, D. and Cervone, F. (2003) The crystal structure of polygalacturonase-inhibiting protein (PGIP), a leucine-rich repeat protein involved in plant defense. *Proc Natl Acad Sci U S A*, **100**, 10124-10128.
- Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., Rees, C. A., Cherry, J. M., Botstein, D., Brown, P. O. and Alizadeh, A. A. (2003) SOURCE: a

- unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res*, **31**, 219-223.
- Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, **4**, R7.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-14868.
- Ernst, J. and Bar-Joseph, Z. (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, **7**, 191.
- Eulgem, T. (2005) Regulation of the Arabidopsis defense transcriptome. *Trends in Plant Science*, **10**, 71-78.
- Fisher, R. A. (1935) The logic of inductive inference. *J Roy Stat Soc*, **98**, 39-82.
- Fodor, I. K. (2002) A survey of dimension reduction techniques.
- Fraunholz, M. J. and Roos, D. S. (2003) PlasmoDB: exploring genomics and post-genomics data of the malaria parasite, Plasmodium falciparum. *Redox Rep*, **8**, 317-320.
- Fujiwara, S., Tanaka, N., Kaneda, T., Takayama, S., Isogai, A. and Che, F. S. (2004) Rice cDNA microarray-based gene expression profiling of the response to flagellin perception in cultured rice cells. *Mol Plant Microbe Interact*, **17**, 986-998.
- Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M. and Brinkman, F. S. (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.
- Gale, M. D. and Devos, K. M. (1998) Comparative genetics in the grasses. *Proc Natl Acad Sci U S A*, **95**, 1971-1974.
- Glazebrook, J. (2001) Genes controlling expression of defense responses in Arabidopsis--2001 status. *Curr Opin Plant Biol*, **4**, 301-308.
- Glazebrook, J. (2005) Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu Rev Phytopathol*, **43**, 205-227.
- Glazebrook, J., Chen, W., Estes, B., Chang, H. S., Nawrath, C., Metraux, J. P., Zhu, T. and Katagiri, F. (2003) Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping. *Plant J*, **34**, 217-228.
- Glazebrook, J., Rogers, E. E. and Ausubel, F. M. (1997) Use of Arabidopsis for genetic dissection of plant defense responses. *Annu Rev Genet*, **31**, 547-569.
- Goff, S. A. (1999) Rice as a model for cereal genomics. *Curr Opin Plant Biol*, **2**, 86-89.
- Gurr, S. J. and Rushton, P. J. (2005) Engineering plants with increased disease resistance: what are we going to express? *Trends Biotechnol*, **23**, 275-282.
- Hawkins, D. M., Basak, S. C. and Mills, D. (2003) Assessing model fit by cross-validation. *J Chem Inf Comput Sci*, **43**, 579-586.
- Hirsch, J., Deslandes, L., Feng, D. X., Balague, C. and Marco, Y. (2002) Delayed symptom development in ein2-1, an Arabidopsis ethylene-insensitive mutant, in response to bacterial wilt caused by Ralstonia solanacearum. *Phytopathology*, **92**, 1142.
- Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat*, **6**, 65-70.
- Horrocks, P., Dechering, K. and Lanzer, M. (1998) Control of gene expression in Plasmodium falciparum. *Mol Biochem Parasitol*, **95**, 171-181.
- Jupiter, D. C. and Vanburen, V. (2008) A visual data mining tool that facilitates reconstruction of transcription regulatory networks. *PLoS ONE*, **3**, e1717.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**, D277-D280.
- Khatri, P., Draghici, S., Ostermeier, G. C. and Krawetz, S. A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266-270.

- Kim, K. J., Park, C. J., An, J. M., Ham, B. K., Lee, B. J. and Paek, K. H. (2005) CaAlaAT1 catalyzes the alanine: 2-oxoglutarate aminotransferase reaction during the resistance response against Tobacco mosaic virus in hot pepper. *Planta*, **221**, 857-867.
- Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y. and Karp, P. D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, **32**, D438-D442.
- Lanzer, M., de Bruin, D. and Ravetch, J. V. (1992) A sequence element associated with the Plasmodium falciparum KAHRP gene is the site of developmentally regulated protein-DNA interactions. *Nucleic Acids Res*, **20**, 3051-3056.
- Laule, O., Hirsch-Hoffmann, M., Hruz, T., Gruissem, W. and Zimmermann, P. (2006) Web-based analysis of the mouse transcriptome using Genevestigator. *BMC Bioinformatics*, **7**, 311.
- Law, P. J., Claudel-Renard, C., Joubert, F., Louw, A. I. and Berger, D. K. (2008) MADIBA: A web server toolkit for biological interpretation of Plasmodium and plant gene clusters. *BMC Genomics*, **9**, 105.
- Le Roch, K. G., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., De la Vega, P., Holder, A. A., Batalov, S., Carucci, D. J. and Winzeler, E. A. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503-1508.
- Lelandais, G., Marc, P., Vincens, P., Jacq, C. and Vialette, S. (2004) MiCoViTo: a tool for gene-centric comparison and visualization of yeast transcriptome states. *BMC Bioinformatics*, **5**, 20.
- Lenhard, B., Hayes, W. S. and Wasserman, W. W. (2001) GeneLynx: a gene-centric portal to the human genome. *Genome Res*, **11**, 2151-2157.
- Liu, H., Coulthurst, S. J., Pritchard, L., Hedley, P. E., Ravensdale, M., Humphris, S., Burr, T., Takle, G., Brurberg, M.-B., Birch, P. R. J., Salmond, G. P. C. and Toth, I. K. (2008) Quorum Sensing coordinates brute force and stealth modes of infection in the plant pathogen Pectobacterium atrosepticum. *PLoS Pathogens*, **4**, e1000093.
- Ma, S., Gong, Q. and Bohnert, H. J. (2006) Dissecting salt stress pathways. *J Exp Bot*, **57**, 1097-1107.
- Maleck, K., Levine, A., Eulgem, T., Morgan, A., Schmid, J., Lawton, K. A., Dangl, J. L. and Dietrich, R. A. (2000) The transcriptome of Arabidopsis thaliana during systemic acquired resistance. *Nat Genet*, **26**, 403-410.
- Marechal, E. and Cesbron-Delauw, M. F. (2001) The apicoplast: a new member of the plastid family. *Trends in Plant Science*, **6**, 200-205.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucl Acids Res*, **31**, 374-378.
- Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E. and Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, **34**, D108-D110.
- McDowell, J. M. and Woffenden, B. J. (2003) Plant disease resistance genes: recent insights and potential applications. *Trends Biotechnol*, **21**, 178-183.
- Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. and Koornneef, M. (1998) Arabidopsis thaliana: a model plant for genome analysis. *Science*, **282**, 662-682.
- Moore, G., Devos, K. M., Wang, Z. and Gale, M. D. (1995) Cereal genome evolution: grasses, line up and form a circle. *Curr Biol*, **5**, 737-739.

- Murray, S. L., Denby, K. J., Berger, D. K. and Loake, G. J. (2002) Disease resistance signalling in Arabidopsis: applications in the study of plant pathology in South Africa. *S Afr J Sci*, **98**, 161-165.
- Mysore, K. S. and Ryu, C. M. (2004) Nonhost resistance: how much do we know? *Trends Plant Sci*, **9**, 97-104.
- Naidoo, S. (2008) Microarray expression studies in the model plant Arabidopsis thaliana infected with the bacterial pathogen Ralstonia solanacearum. *PhD Thesis*.
- Navarro, L., Bari, R., Achard, P., Lison, P., Nemri, A., Harberd, N. P. and Jones, J. D. (2008) DELLAs control plant immune responses by modulating the balance of jasmonic acid and salicylic acid signaling. *Curr Biol*, **18**, 650-655.
- Noutoshi, Y., Ito, T., Seki, M., Nakashita, H., Yoshida, S., Marco, Y., Shirasu, K. and Shinozaki, K. (2005) A single amino acid insertion in the WRKY domain of the Arabidopsis TIR-NBS-LRR-WRKY-type disease resistance protein SLH1 (sensitive to low humidity 1) causes activation of defense responses and hypersensitive cell death. *Plant J*, **43**, 873-888.
- Pinney, J. W., Shirley, M. W., McConkey, G. A. and Westhead, D. R. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucl Acids Res*, **33**, 1399-1409.
- Rabbani, M. A., Maruyama, K., Abe, H., Khan, M. A., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2003) Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. *Plant Physiol*, **133**, 1755-1767.
- Ralph, S. A., D'Ombrain, M. C. and McFadden, G. I. (2001) The apicoplast as an antimalarial drug target. *Drug Resist Updat*, **4**, 145-151.
- Rensink, W. A. and Buell, C. R. (2004) Arabidopsis to rice. Applying knowledge from a weed to enhance our understanding of a crop species. *Plant Physiol*, **135**, 622-629.
- Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, **31**, 224-228.
- Ronald, P. and Leung, H. (2002) The rice genome: the most precious things are not jade and pearls. *Science*, **296**, 58-59.
- Saccone, C. and Pesole, G. (2003) In: *Handbook of comparative genomics: principles and methodology*, John Wiley & Sons.
- Sachs, J. and Malaney, P. (2002) The economic and social burden of malaria. *Nature*, **415**, 680-685.
- Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J. C., Cattolico, L., Chandler, M., Choisine, N., Claudel-Renard, C., Cunnac, S., Demange, N., Gaspin, C., Lavie, M., Moisan, A., Robert, C., Saurin, W., Schiex, T., Siguier, P., Thebault, P., Whalen, M., Wincker, P., Levy, M., Weissenbach, J. and Boucher, C. A. (2002) Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature*, **415**, 497-502.
- Sasaki, T. and Burr, B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr Opin Plant Biol*, **3**, 138-142.

- Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C. and Manners, J. M. (2000) Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proc Natl Acad Sci U S A*, **97**, 11655-11660.
- Segal, E., Kaushal, A., Yelensky, R., Pham, T., Regev, A., Koller, D. and Friedman, N. (2004) GeneXPress: a visualization and statistical analysis tool for gene expression and sequence data. At: *Eleventh International Conference on Intelligent Systems for Molecular Biology*.
- Seoighe, C. and Gehring, C. (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet*, **20**, 461-464.
- Shimono, M., Sugano, S., Nakayama, A., Jiang, C. J., Ono, K., Toki, S. and Takatsuji, H. (2007) Rice WRKY45 plays a crucial role in benzothiadiazole-inducible blast resistance. *Plant Cell*, **19**, 2064-2076.
- Shlens, J. (2005) A tutorial on principal components analysis.
- Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. and Hay, S. I. (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, **434**, 214-217.
- Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. (2007) pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164-1167.
- Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*, **12**, 1599-1610.
- Stekel, D. (2003) In: *Microarray bioinformatics*, Cambridge University Press.
- Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, **100**, 9440-9445.
- Strange, R. N. and Scott, P. R. (2005) Plant disease: a threat to global food security. *Annu Rev Phytopathol*, **43**, 83-116.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.
- Thatcher, L. F., Anderson, J. P. and Singh, K. B. (2005) Plant defence responses: what have we learnt from *Arabidopsis*? *Funct Plant Biol*, **32**, 1-19.
- The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.
- Toth, I. K., Newton, J. A., Hyman, L. J., Lees, A. K., Daykin, M., Ortori, C., Williams, P. and Fray, R. G. (2004) Potato plants genetically modified to produce N-acylhomoserine lactones increase susceptibility to soft rot erwiniae. *Mol Plant Microbe Interact*, **17**, 880-887.
- Toth, I. K., Pritchard, L. and Birch, P. R. (2006) Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu Rev Phytopathol*, **44**, 305-336.
- Toth, I. K., Bell, K. S., Holeva, M. C. and Birch, P. R. J. (2003) Soft rot erwiniae: from genes to genomes. *Mol Plant Pathol*, **4**, 17-30.
- Toth, I. K. and Birch, P. R. J. (2005) Rotting softly and stealthily. *Curr Opin Plant Biol*, **8**, 424-429.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520-525.
- Truman, W., Bennett, M. H., Kubigsteltig, I., Turnbull, C. and Grant, M. (2007) *Arabidopsis* systemic immunity uses conserved defense signaling pathways and is mediated by jasmonates. *Proc Natl Acad Sci U S A*, **104**, 1075-1080.

- Ulker, B. and Somssich, I. E. (2004) WRKY transcription factors: from DNA binding towards biological function. *Curr Opin Plant Biol*, **7**, 491-498.
- van Helden, J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res*, **31**, 3593-3596.
- von Bodman, S. B., Ball, J. K., Faini, M. A., Herrera, C. M., Minogue, T. D., Urbanowski, M. L. and Stevens, A. M. (2003) The quorum sensing negative regulators EsaR and ExpREcc, homologues within the LuxR family, retain the ability to function as activators of transcription. *J Bacteriol*, **185**, 7001-7007.
- Vos, W. J. (2005) Microarray discoveries. *National Bioinformatics Network workshop - Statistics of high throughput biology*.
- Wold, S., Esbensen, K. and Geland, P. (1987) Principal component analysis. *Chemometr Intell Lab*, **2**, 37-52.
- Wu, C., Zhou, S., Zhang, Q., Zhao, W. and Peng, Y. (2006) Molecular cloning and differential expression of an gamma-aminobutyrate transaminase gene, OsGABA-T, in rice (*Oryza sativa*) leaves infected with blast fungus. *J Plant Res*, **119**, 663-669.
- Yona, G., Dirks, W., Rahman, S. and Lin, D. M. (2006) Effective similarity measures for expression profiles. *Bioinformatics*, **22**, 1616-1622.
- Young, J. A., Fivelman, Q. L., Blair, P. L., de la Vega, P., Le Roch, K. G., Zhou, Y., Carucci, D. J., Baker, D. A. and Winzeler, E. A. (2005) The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol*, **143**, 67-79.
- Yuan, Q., Ouyang, S., Wang, A., Zhu, W., Maiti, R., Lin, H., Hamilton, J., Haas, B., Sultana, R., Cheung, F., Wortman, J. and Buell, C. R. (2005) The Institute for Genomic Research Osa1 rice genome annotation database. *Plant Physiol*, **138**, 18-26.
- Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C. and Weinstein, J. N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, **4**, R28.
- Zhang, B., Kirov, S. and Snoddy, J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, **33**, W741-W748.
- Zhu, T. (2003) Global analysis of gene expression using GeneChip microarrays. *Curr Opin Plant Biol*, **6**, 418-425.
- Zimmermann, P., Hennig, L. and Grissem, W. (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci*, **10**, 407-409.
- Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Grissem, W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*, **136**, 2621-2632.