

SSHscreen and SSHdb: software for microarray-based screening and sequence management of cDNA libraries

by

Nanette Coetzer

Submitted in partial fulfilment of the requirements of the degree

Magister Scientiae Bioinformatics

in the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

November 2009



Declaration

I, Nanette Coetzer, declare that the dissertation which I hereby submit for the degree *MSc Bioinformatics* at the University of Pretoria is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: Date:

Acknowledgements

I want to thank the following people from my heart:

- Firstly, God for the opportunity that I had to do my MSc degree, for the support system in the form of key persons that He placed in my life, and for the special way in which He guided me during these years.
- My parents, Gerhard and Yda, for their continuous support, prayers, motivation and encouragement through all my studies. My sisters, Eugenie and Gerda, for their friendship and for always wanting the best for me. My grandmother, Ouma Nancy, for her heartening and wisdom in difficult times.
- Barry Christie for believing in me, encouraging me and for all his prayers and fellowship, Dr. Armin Deffur for his inputs and positive motivation and all my friends, especially the friends from Every Nation Pretoria East, for moral boosting and coffee dates when I really needed it.
- Prof. Dave Berger for his unique guidance, support, assistance and that he always believed in me.
- My fellow students at the Bioinformatics and Computational Biology Unit. In particular Charles Hefer for his valuable inputs, help, advice and friendship as well as Prof. Fourie Joubert for his guidance and encouragement.
- From the Molecular Plant-Pathogen Interactions group, specifically Inge Gazendam, Dr. Bridget Crampton and Dr. Sanushka Naidoo for their cowpea, pearl millet and *Ara-bidopsis* SSH library data that I could use during my MSc study, and also the rest of the lab – it is a pleasure being part this group.
- Finally, the National Bioinformatics Network and the National Research Foundation for funding this project.

Contents

Declaration	i
Acknowledgements	ii
List of Abbreviations	viii
List of Tables	x
List of Figures	xi
Preface	xiii
Chapter 1. Introduction/Background	1
1.1. Identification of differentially expressed genes	1
1.1.1. Using cDNA sequences in transcriptome profiling	1
1.1.2. Molecular techniques to identify differentially expressed genes	2
1.1.3. The SSH technique	4
1.1.4. Screening SSH cDNA libraries	6
1.1.5. cDNA microarrays	7
1.1.6. Software for microarray data analysis	8
1.1.7. Validation of selected transcripts	9
1.2. Statistical principles in the data analysis of two-colour microarrays using limma	10
1.2.1. Experimental design	10
1.2.2. Limma functions for reading two-colour data	11
1.2.2.1. The targets file	11
1.2.2.2. Intensity data files	11
1.2.2.3. The GAL file	12
1.2.2.4. The spot types file	12
1.2.3. Limma functions for data exploration	15
1.2.4. Limma functions for pre-processing two-colour microarray data	17
1.2.4.1. Background correction	17
1.2.4.2. Normalization	19
1.2.4.3. Within-array normalization methods	20
1.2.4.4. Using control spots for with-in array normalization	21

1.2.4.5.	Between-array normalization methods	24
1.2.5.	Limma functions for the identification of differentially expressed genes	26
1.2.5.1.	Linear models	26
1.2.5.2.	Gene-wise linear models from experimental designs	27
1.2.5.3.	Within-array duplicate spots	29
1.2.5.4.	Linear model fit	30
1.2.5.5.	The ordinary t-statistic	30
1.2.5.6.	Variations on the ordinary t-statistic	31
1.2.5.7.	Empirical Bayes analysis	33
1.2.5.8.	The moderated t-statistic	34
1.2.5.9.	The p-value associated with the moderated t-statistic	34
1.2.5.10.	The B-statistic (posterior odds)	35
1.2.5.11.	Estimation of hyperparameters	37
1.2.5.12.	Comparison of the B-statistic and the moderated t-statistic	38
1.2.5.13.	The moderated F statistic	38
1.2.5.14.	A summary of the advantages of the moderated t-statistic	38
1.2.5.15.	Summary of statistical variables used in section 1.2.5	39
1.2.6.	Limma functions for handling multiple testing and output	40
1.2.6.1.	The multiple testing problem	40
1.2.6.2.	Error rates	40
1.2.6.3.	FPR	41
1.2.6.4.	FWER	41
1.2.6.5.	FDR	42
1.2.6.6.	Top tables	43
1.3.	The background to SSHscreen	43
1.3.1.	Using MS Excel to calculate enrichment ratios ER1 and ER2	43
1.3.2.	SSHscreen as a package in R	44
1.4.	Databases for management of cDNA sequences	46
1.4.1.	Similarity searches against sequence databases using BLAST	46
1.4.2.	Tools currently available for managing in-house sequencing projects	46
1.5.	Conclusion	47
Chapter 2. SSHscreen and SSHdb, generic software for microarray-based gene discovery: application to the stress response in cowpea		48
2.1.	Note	49
2.2.	Authors' contributions	49
2.3.	Abstract	49
2.4.	Background	50

2.5.	Methods	52
2.5.1.	Plant materials and treatments	52
2.5.2.	RNA extraction	53
2.5.3.	Construction of cDNA library using SSH	53
2.5.4.	Fabrication of SSH library on glass slide array	54
2.5.5.	Screening SSH library on microarrays	55
2.5.6.	SSHscreen software analysis of microarray data	57
2.5.7.	Sequencing	58
2.5.8.	Annotation and management of sequences using SSHdb	58
2.5.9.	Quantitative PCR	58
2.6.	Results	60
2.6.1.	Construction of cowpea drought expression SSH library and overview of SSHscreen/SSHdb data analysis pipeline	60
2.6.2.	Screening the cowpea SSH libraries using SSHscreen 2.0.0	63
2.6.3.	Annotation and management of cowpea SSH library sequences using SSHdb	73
2.6.4.	Cowpea SSH library contains genes known to play a role in plant response to stress	74
2.6.5.	Verification of SSHscreen Enrichment Ratios using qPCR Representative	80
2.7.	Discussion	81
2.8.	Conclusion	87
2.9.	Acknowledgements	87
Chapter 3. Application of SSHscreen-SSHdb pipeline in Pearl Millet		88
3.1.	Note	88
3.2.	Introduction	88
3.3.	Methods	90
3.3.1.	Construction of cDNA library using SSH	90
3.3.2.	Screening SSH library on microarrays	90
3.3.2.1.	Slide layout and probes	90
3.3.2.2.	Experimental design and targets	90
3.3.2.3.	SSHscreen analysis of the forward library	91
3.3.2.4.	Limma analysis of the reverse library	93
3.3.2.5.	Verification and evaluation of SSHscreen results	93
3.3.3.	Sequencing	94
3.3.4.	Management and annotation of clones in the SSH library using SSHdb	94
3.4.	Results	95
3.4.1.	Comparison of within-array normalization methods	95
3.4.2.	Comparison of different control-spot normalization methods	96
3.4.3.	Taking the print-tip effect into account in control-spot normalization	100

3.4.4.	Top Table statistics	100
3.4.5.	SSHscreen forward library ER3 analysis	101
3.4.6.	SSHscreen forward library ER1 analysis	103
3.4.6.1.	ER1 analysis using SSHscreen 2.0.0 in R	103
3.4.6.2.	ER1 versus ER2 plot in MS Excel	104
3.4.6.3.	Predicting the ER3 value from ER1 and ER2	104
3.4.7.	Using limma to verify the SSHscreen values	106
3.4.8.	Reverse library analysis using limma	107
3.4.9.	Management and annotation of the pearl millet SSH library using SSHdb	108
3.4.9.1.	The forward library	108
3.4.9.2.	The reverse library	110
3.5.	Discussion	112
Chapter 4.	Application of SSHscreen-SSHdb pipeline in <i>Arabidopsis</i>	115
4.1.	Note	115
4.2.	Introduction	115
4.2.1.	The plant pathosystem being studied	115
4.2.2.	Introduction/overview of plant defense	117
4.3.	Methods	118
4.3.1.	Construction of cDNA library using SSH	118
4.3.2.	Screening SSH library on microarray	119
4.3.2.1.	Slide layout and probes	119
4.3.2.2.	Experimental design and targets	119
4.3.2.3.	SSHscreen software analysis	120
4.3.3.	Sequencing	122
4.3.4.	Management and annotation of clones in SSH library	122
4.3.4.1.	SSHdb	122
4.3.4.2.	Selection of redundant partner groups for further analysis	122
4.3.4.3.	Other bioinformatics tools	123
4.4.	Results	123
4.4.1.	SSHscreen ER3 analysis of both libraries	123
4.4.2.	Biological annotation of the forward library	126
4.4.2.1.	SSHdb output	126
4.4.2.2.	EasyGO analysis	127
4.4.2.3.	TAIR bulk data retrieval tools	128
4.4.2.4.	MADIBA	132
4.4.3.	Biological annotation of the reverse library	132
4.4.3.1.	SSHdb output	132
4.4.3.2.	EasyGO analysis	133

Contents	vii
4.4.3.3. TAIR bulk data retrieval tools	133
4.4.3.4. MADIBA	136
4.5. Discussion	137
Chapter 5. Concluding discussion	142
Summary	147
Bibliography	149
Appendix	160
SSHscreen R Documentation	160
High-throughput screening of SSH cDNA libraries using DNA microarray data	160
Description	160
Usage	160
Arguments	160
Details	163
Value	163
Note	165
Author(s)	165
References	165

List of Abbreviations

ABA	- Abscisic Acid
AFLP	- Amplified fragment length polymorphism
BLAST	- Basic Local Alignment Search Tool
cDNA	- Complementary DNA
CRAN	- Comprehensive R Archive Network
DD	- Differential display
DNA	- Deoxyribonucleic Acid
ER	- Enrichment Ratio
EST	- Expressed sequence tag
ET	- Ethylene
FDR	- False Discovery Rate
FNR	- False Negative Rate
FPR	- False Positive Rate
FWER	- Family Wise Error Rate
GAL	- GenePix Array List
GO	- Gene Ontology
GUI	- Graphical User Interface
HR	- Hypersensitive Response
JA	- Jasmonic Acid
limma	- Linear Models for Microarray data
MADIBA	- MicroArray Data Interface for Biological Annotation
MIAME	- Minimal Information About a Microarray Experiment
mRNAs	- Messenger RNA
MPPI	- Molecular Plant-Pathogen Interactions
MS	- Microsoft
NCBI	- National Center for Biotechnology Information
PCR	- Polymerase Chain Reaction
qPCR	- Quantitative PCR
qRT-PCR	- Quantitative Reverse Transcriptase RCR

RAP-PCR	- RNA-fingerprinting by arbitrarily primed PCR
RDA	- Representational difference analysis
RNA	- Ribonucleic Acid
ROS	- Reactive Oxygen Species
RT-PCR	- Reverse Transcriptase PCR
SA	- Salicylic Acid
SAGE	- Serial analysis of gene expression
SAR	- Systemic Acquired Resistance
SSH	- Suppression subtractive hybridization
ST	- Subtracted Tester
TAIR	- Arabidopsis Information Resource
UD	- Unsubtracted Driver
UP	- University of Pretoria
UT	- Unsubtracted Tester

List of Tables

1.1.	An example of a limma top table.	37
1.2.	Numbers of correct and incorrect conclusions of n hypothesis tests.	42
2.1.	Table of oligonucleotide primers used in this study.	55
2.2.	Targets file used for ER3 analysis of cowpea SSH libraries.	68
2.3.	Targets file used for ER3 analysis of cowpea SSH libraries.	68
2.4.	Top tables produced by SSHscreen for the forward and reverse cowpea libraries.	72
2.5.	Top tables produced by SSHscreen for the forward and reverse cowpea libraries.	75
2.6.	Table 2.5 continue.	76
3.1.	Number (%) of genes in the top table, using different cut-off criteria, different prior guesses of the number of differentially expressed genes, as well as different control-spot within-array normalization method.	101
3.2.	SSHscreen 2.0.0 output: up/down regulation top table (forward library).	103
3.3.	SSHdb output: Annotated ER3 up/down-regulation top table for pearl millet (forward library).	109
3.4.	SSHdb output: Annotated limma up/down-regulation top table for pearl millet (reverse library).	111
4.1.	Groups 1-30 of the annotated ER3 up/down-regulation top table for <i>Arabidopsis</i> (forward library).	130
4.2.	Groups 31-60 of the annotated ER3 up/down-regulation top table for <i>Arabidopsis</i> (forward library).	131
4.3.	Annotated ER3 up/down-regulation top table for <i>Arabidopsis</i> (reverse library).	135

List of Figures

1.1.	Forward and reverse libraries SSH libraries.	4
1.2.	A schematic outline of the SSH metod.	5
1.3.	Diagram of a two-colour microarray experiment.	7
1.4.	Example of a targets file, spot types file and part of a GAL file.	13
1.5.	R code showing the limma functions for reading the required files.	14
1.6.	An example of the layout of a microarray slide.	15
1.7.	GenePix background calculation.	16
1.8.	Image plots showing the variation of the red and green background values.	17
1.9.	MA-plots after background correction.	19
1.10.	MA-plot showing three different trend lines.	21
1.11.	An illustration of print-tip loess normalization.	23
1.12.	MA-plots after up-weighting within-array normalization.	24
1.13.	Density plots illustrating <i>loess</i> within-array and <i>A-quantile</i> between-array normalization.	25
1.14.	Example experimental designs for two-colour microarrays.	27
1.15.	Components of the <i>MArrayLM</i> object in limma.	32
1.16.	The volcano plot.	36
1.17.	SSHscreen 'ER1' analysis and 'ER3' analysis.	45
2.1.	Screenshot of the SSHdb 'Sequence Database' view.	61
2.2.	Schematic representation of the flow of data through the SSHscreen-SSHdb pipeline.	62
2.3.	SSHscreen R script used for ER3 analysis of cowpea SSH libraries	66
2.4.	Example of Microarray pseudocolour images following hybridization.	67
2.5.	MA plots after normalization of the forward and reverse cowpea SSH libraries.	69
2.6.	ER3 versus inverse ER2 plot produced by SSHscreen for the cowpea forward library.	70
2.7.	ER3 versus inverse ER2 plot produced by SSHscreen for the cowpea reverse library.	71
2.8.	Schematic representation of SSHdb.	74
2.9.	ER3 versus inverse ER2 plot for sequenced clones to illustrate that redundant partners cluster together. The ER	78
2.10.	Regulation (a) and abundnace (b) of selected cowpea genes (qPCR verification).	79
3.1.	Pearl Millet for grain.	89
3.2.	Experimental design of the pearl millet microarray experiment.	91

3.3. Box-plots of the M-values (before and after normalization) of the control spots in each print-tip group.	97
3.4. MA-plots after normalization: comparison of within-array normalization methods.	98
3.5. Plots giving the correlation between within-array normalization methods using control spots.	99
3.6. SSHscreen 2.0.0 output: ER3 versus inverse ER2 plot (forward library).	102
3.7. SSHscreen 2.0.0 output: ER1 versus ER2 plot (forward library).	105
3.8. ER1 versus ER2 plot produced with MS Excel for the millet forward subtraction library.	106
3.9. MA-plots before and after normalization of the pearl millet reverse library (limma output).	107
3.10. Functional categorization pie-chart for the pearl millet forward SSH cDNA library.	108
3.11. Functional categorization pie-chart for the pearl millet reverse SSH cDNA library.	110
4.1. A differential response was obtained between <i>A. thaliana</i> ecotypes Kil-0 and Be-0 in response to <i>R. solanacearum</i>	116
4.2. A model giving an overview of the plant immune system and illustrating the quantitative output thereof.	118
4.3. Experimental design of the <i>Arabidopsis</i> microarray experiment.	119
4.4. <i>Arabidopsis</i> SSHscreen analysis input files.	120
4.5. SSHscreen output: MA-plots after normalization for the <i>Arabidopsis</i> ER3 analysis of the forward and reverse libraries respectively.	125
4.6. SSHscreen output: ER3 versus inverse ER2 plots for the <i>Arabidopsis</i> ER3 analysis of the forward and reverse libraries respectively.	126
4.8. TAIR functional categorization by annotation for: GO biological process (<i>Arabidopsis</i> forward library analysis).	128
4.7. EasyGO Gene Ontology output, on the aspect <i>biological process</i> for GO:0050896 (<i>Arabidopsis</i> forward library analysis).	129
4.9. EasyGO Gene Ontology output, on the aspect <i>biological process</i> for GO:0008150 (<i>Arabidopsis</i> reverse library analysis).	134
4.10. TAIR functional categorization by annotation for: GO biological process (<i>Arabidopsis</i> reverse library analysis).	136

Preface

The ultimate aim of functional genomics is to increase one's ability to understand genome function. In order to achieve this aim, many gene discovery and functional annotation projects are underway. Although there are several alternative approaches such as cDNA-AFLP, DD-RT-PCR and RNA-Seq, SSH remains a popular approach for gene discovery from non-model organisms, for which an annotated genome sequence is not available. A recent search with the keywords 'suppression subtractive hybridization' in the title of research articles on PubMed produced 1213 hits, which confirmed the technique's popularity.

The Molecular Plant-Pathogen Interactions (MPPI) research group at the University of Pretoria (UP) chose to apply SSH to gene discovery in cowpea (*Vigna unguiculata* (L.) Walp), pearl millet (*Pennisetum glaucum* (L.) R. Br.) and *Arabidopsis* (*Arabidopsis thaliana* (L.) Heynh) ecotype Kil-0, where the objective of each library was to enrich for genes expressed during drought stress, biotic stress and treatment with the bacterial wilt pathogen *Ralstonia solanacearum*, respectively. This dissertation describes two software innovations that facilitate gene discovery using SSH in the form of the "SSHscreen-SSHdb pipeline", that was used to screen these libraries and to manage the resulting sequencing and annotation information.

A method for screening SSH libraries using Microsoft Excel calculations, is outlined in van den Berg *et al.*, 2004. SSHscreen version 1.0.1 was developed as an R package by Dr. Wiesner Vos (while at the Department of Statistics, Oxford University), in collaboration with Prof. Dave Berger (Department Plant Science, FABI, UP) (Berger *et al.*, 2007), including more sophisticated normalization and statistical analysis steps using the limma package from the BioConductor project. In this MSc study, substantial improvements to the functionality of SSHscreen were added, leading to the latest version, SSHscreen 2.0.0, available at <http://microarray.up.ac.za/SSHscreen/>. The necessity for a database system to manage the resulting data and sequence information, lead to the development of SSHdb, a web-based tool for the management and annotation of cDNA sequences in a SSH cDNA library, which can be accessed at <http://sshdb.bi.up.ac.za/>.

Chapter 1 of this dissertation is a literature survey dealing with methods to identify

differentially expressed genes, the statistical principles in the data analysis of two-colour microarrays using limma and databases for the management of cDNA sequences.

Chapter 2 describes the development and validation of the SSHscreen-SSHdb pipeline using a cowpea drought expression SSH cDNA library.

Chapters 3 and *4* aim to demonstrate specific features of the SSHscreen-SSHdb pipeline using two different case studies. *Chapter 3* focuses mainly on the flexibility and the use of different SSHscreen argument options using the pearl millet case study and *Chapter 4* on the biological annotation of individual genes identified by the pipeline with SSHdb and other bioinformatics tools, using the *Arabidopsis* case study.

Chapter 5 provides a general concluding discussion and is followed by a summary.

The *Bibliography* of the complete dissertation is at the end of the document. Although *Chapter 2* has been submitted to the journal *Plant Methods*, the references are included in the *Bibliography* at the end to ensure consistency with the rest of the dissertation layout.

The *Appendix* contains the SSHscreen R documentation.

The primary aim of this MSc study was the development of the two software tools, SSHscreen and SSHdb, which forms part of a pipeline for gene discovery using SSH. SSHscreen, used for quantitative screening of clones in a SSH library, was improved, and SSHdb was developed as a web-based tool to manage the cDNA sequences.

In this dissertation, it was hypothesized that using the SSHscreen-SSHdb pipeline, differential gene expression can be quantified and defense-related genes identified, following stress response (i.e. drought or pathogen challenge) in Pearl Millet, Cowpea, *Arabidopsis* ecotype Kil-0, in a quick, easy and efficient way.

Chapter 1

Introduction/Background

Functional genomics attempts to make use of the vast wealth of data produced by genome projects, for example genome sequencing projects, to describe gene functions and interactions. Transcriptomics is the branch of functional genomics that focuses on the subset of the genome that is 'expressed', since unlike the genome, the transcriptome can vary with external environmental conditions. Hence, the regulatory mechanisms and transcriptional networks underlying particular biological processes can be studied using either high-throughput transcriptome profiling techniques, such as Serial Analysis of Gene Expression (SAGE), DNA microarrays or RNA-Seq, or other molecular techniques which identifies differentially expressed genes between different populations, such as cDNA amplified fragment length polymorphism (cDNA-AFLP) or suppression subtractive hybridization (SSH). Techniques like these are often used in gene discovery projects, when studying non-model organisms.

1.1. Identification of differentially expressed genes

1.1.1. Using cDNA sequences in transcriptome profiling

The first step in a gene discovery project is usually to construct a gene library in the laboratory. A genomic library represents all the DNA sequences found in the genome of a particular organism (Alberts *et al.*, 1997), whereas a cDNA library refers to a complete, or near complete, set of all the mRNAs present in a particular tissue or cell line of interest. Soltis *et al.*, 2002, Dowd *et al.*, 2004 and Collett *et al.*, 2004 constructed cDNA libraries in order to study gene expression changes in non-model plants.

ESTs can be used as a tool for gene discovery and expression analysis, allowing the rapid characterization of thousands of cDNAs at a minimal cost. It is created by single pass sequencing of the 5' and/or 3' ends of randomly isolated gene transcripts after converted into cDNA. A fraction of the resulting sequence data is normally erroneous and the length of a typical EST is approximately 200-900 nucleotides, representing only a portion of a coding

sequence. Despite these limitations, EST databases can be an effective and reliable source of gene expression data (Adams *et al.*, 1991; Alba *et al.*, 2004). Luo *et al.*, 2005 used an Expressed sequence tag (EST) library to identify resistance genes in peanut in response to *Aspergillus parasiticus* infection under drought stress.

To study the molecular regulation in particular biological processes in more depth, it is necessary to identify and clone the relevant subsets of differentially expressed genes of interest so that it can be studied in detail. A variety of molecular techniques are available to isolate and characterize cDNA fragments that are differentially expressed under specific conditions (Diatchenko *et al.*, 1996; Bachem *et al.*, 1998). Three main categories of such techniques include RNA-fingerprinting techniques, sequencing-based approaches and PCR-based cDNA subtractive hybridization methods.

1.1.2. Molecular techniques to identify differentially expressed genes

Differential display reverse transcriptase PCR (DD-RT-PCR) uses oligonucleotide primer pairs to define mRNA sub-populations for comparison. The first primer is always anchored to the 3' poly-A tail of an mRNA molecule whereas the other primer is short and non-specific in sequence, so that it anneals at different positions in relation to the first primer. The resulting mRNA sub-populations are reverse transcribed, amplified and visualized on poly-crylamide gels. Side-by-side comparison of the band patterns of related samples, lead to the identification of differentially expressed cDNA fragments which can be isolated from the gel for sequencing (Liang and Pardee, 1992).

RNA-fingerprinting by arbitrarily primed PCR (RAP-PCR) also displays the products of cDNA synthesis after amplification by PCR on a gel as a fingerprint. An arbitrarily chosen primer initiates first strand cDNA synthesis by reverse transcriptase at sites in the RNA that best match the primer, whereas an extension of the same arbitrarily primer initiates second strand cDNA synthesis using *Taq* polymerase on the first strand cDNA product at sites where matching is less stringent. When comparing the band patterns of separate RNA populations on the gel, differences in the pattern reflect abundance differences in individual RNAs (Welsh *et al.*, 1992; Ralph *et al.*, 1993).

DD-RT-PCR and RAP-PCR are collectively referred to as RNA-fingerprinting. This is a relatively fast way to identify differentially expressed genes, however limitations include problems with reproducibility and the generation of a high percentage of false positives. It is also limited by its ability to capture low abundance clones and gives inaccurate results when only a few genes are expected to vary (Diatchenko *et al.*, 1996; Zegzouti *et al.*, 1997).

cDNA amplified fragment length polymorphism (cDNA-AFLP) is another PCR based genetic fingerprinting technique and it largely overcomes these limitations, except for the ability

of resulting cDNA libraries to capture low-abundance transcripts. Following cDNA synthesis, two restriction enzymes are used to digest the cDNA, whereafter adaptors are ligated to the ends of the double stranded restriction fragments. Two PCR primers with a complementary sequence to the adaptor and restriction site fragments, and with higher annealing temperatures than the annealing temperatures used in the above mentioned RNA-fingerprinting techniques, are used to selectively amplify subsets of the cDNA populations. Electrophoretic separation of amplicons on polyacrylamide gels can be performed for visualization and comparison of the band patterns (Bachem *et al.*, 1998).

Serial analysis of gene expression (SAGE) combines differential display and cDNA sequencing approaches. Short diagnostic sequence tags are extracted from mRNA molecules, to be concatenated, cloned and sequenced. The output is a list of short sequence tags, together with the number of times it is observed. By comparing these tags to sequence databases, it is usually possible to determine with a reasonable level of confidence the original mRNA that the tag was extracted from. Although SAGE has the advantage that it allows a quantitative analysis, it is labour-intensive and requires a large-scale foundation of sequence information. SAGE also suffers from its limited ability to capture low abundance transcripts (Velculescu *et al.*, 1995; Alba *et al.*, 2004).

cDNA representational difference analysis (cDNA-RDA) is a PCR-based technique where the difference between cDNA sequences from two samples are analyzed using subtractive DNA hybridization. The two samples involved are the tester and the driver, usually representing the treated and the control samples respectively. In cDNA-RDA the first step is to amplify the mRNA representations to ensure that there are enough tester and driver material to start with. An adapter is added to the tester population, and the two populations (tester and driver) are mixed together. After denaturation and hybridization, the result will include tester cDNA bound to driver cDNA, tester cDNA bound to itself and driver cDNA bound to itself. The ends of the fragments are filled so that the tester cDNA bound to itself have an adapter at each end on each strand. After running a PCR reaction with primers that can recognize a sequence on the adapter, the tester cDNA bound to itself will be exponentially enriched, whereas the tester cDNA bound to driver cDNA will only be linearly enriched. The aim is to enrich for differentially expressed transcripts in the treated sample, represented by tester cDNA bound to itself. Thus a few rounds of subtractive hybridization and PCR amplification are necessary to fulfill this aim. cDNA-RDA allow the cloning of rare differentially expressed transcripts to a sufficiently higher extent when compared to the methods described above (O'Neill and Sinclair, 1997).

SSH is a newer and highly effective PCR-based cDNA subtractive hybridization method. In a single procedure, SSH combines normalization and subtraction. Normalization equalizes

the abundance of cDNAs within the tester population, and subtraction excludes the common sequences between the tester and the driver populations. Two libraries can be constructed by SSH, a forward library which enriches for target genes up-regulated in response to treatment and a reverse library which enriches for target genes down-regulated in response to the same treatment. The reverse subtractive library is one where the tester would be the untreated sample and the driver the treated sample (Figure 1.1). Two issues when constructing a cDNA library with SSH, is firstly the necessity to determine if the experimental aim requires a wide or narrow subtraction and secondly it is important to make sure that there is sufficient material to make both forward and reverse cDNA subtraction libraries (Diatchenko *et al.*, 1996).

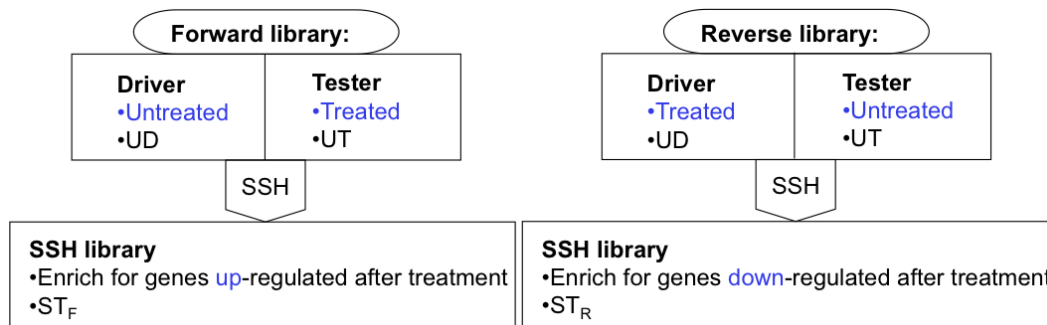


Figure 1.1: Forward and reverse SSH libraries. The reverse library differs from the forward library in that the starting material is switched. In the forward library, the driver is the untreated sample and the tester is the treated sample, whereas in the reverse library, the driver is the treated sample and the tester is the untreated sample. The resulting forward (ST_F ; forward library subtracted tester) and reverse (ST_R ; reverse library subtracted tester) SSH libraries enrich for target genes up- and down-regulated after treatment respectively.

1.1.3. The SSH technique

Figure 1.2 on the next page outlines the SSH method. Firstly the tester and driver populations, also called unsubtracted tester (UT) and unsubtracted driver (UD), are digested with a four-base cutting restriction enzyme *RsaI*, yielding blunt ends. The tester population is divided into two sub-populations, which are then ligated with two different adaptors. An excess of driver cDNA fragments is added to both the tester samples whereafter the respective samples are heat-denatured and allowed to hybridize. After this first hybridization, the resulting molecules are numbered (a), (b), (c) and (d).

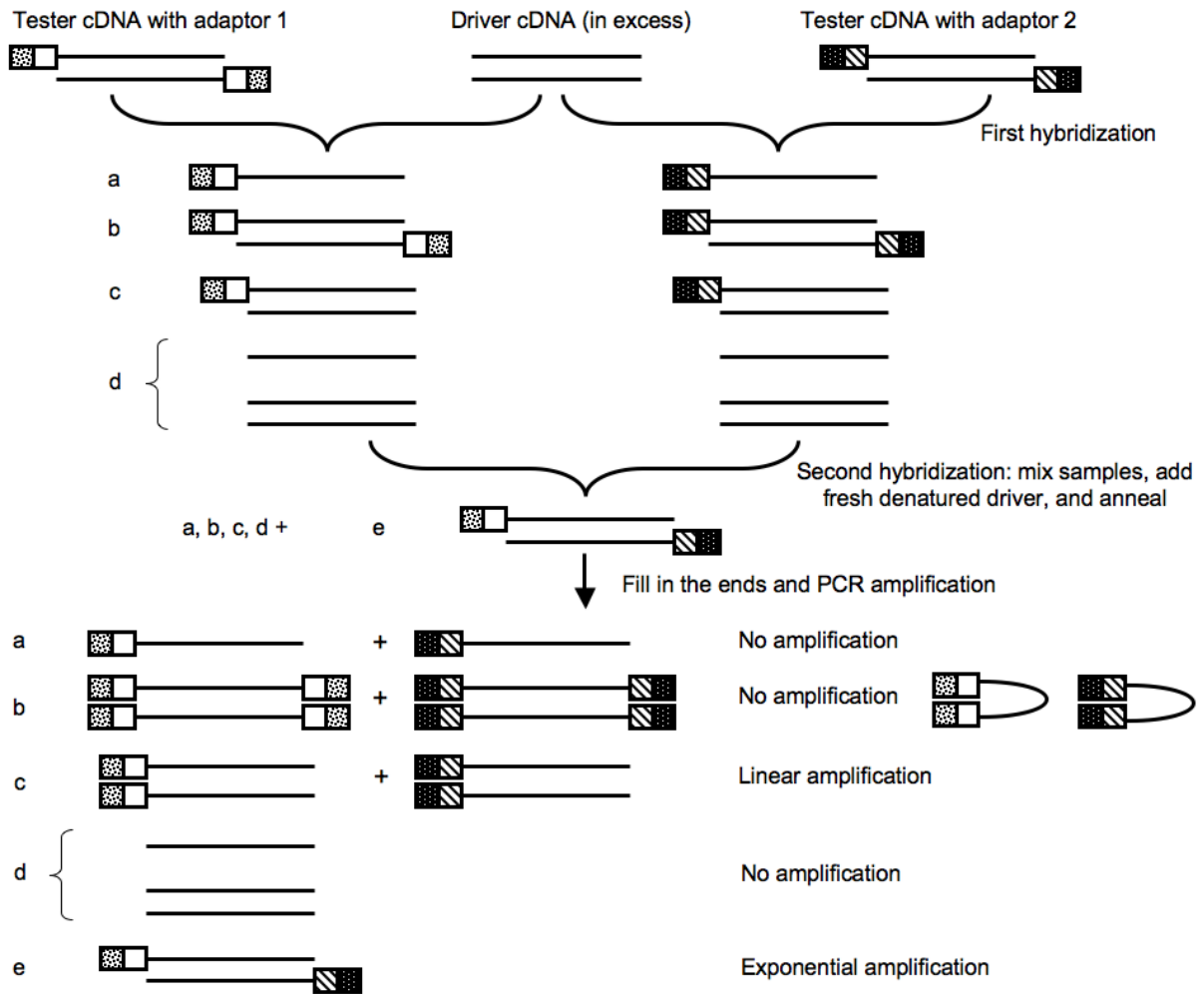


Figure 1.2: A schematic outline of the SSH method (Diatchenko *et al.*, 1996). Solid lines represent the *RsaI* digested tester or driver cDNA. Dark shaded boxes represent the outer part of adaptor 1 and 2 longer strand and corresponding PCR primer sequences. Brighter boxes represent the inner part of the adaptors and corresponding nested PCR primer sequences.

The aim of the first hybridization is to generate single stranded tester molecules (a). These molecules (a) are normalized and subtracted. Normalized, in that concentrations of high and low abundance cDNA become roughly equal. This happens due to the fact that reannealing, which is the generation of molecules (b) is faster for more abundant molecules. It is subtracted because molecules (c), which results from fragments present in the tester and driver populations, are in effect canceled out. The two samples resulting from the first hybridization as well as fresh denatured driver are mixed together and allowed to anneal, where only the normalized and subtracted single stranded tester cDNAs are able to re-associate.

The second hybridization result in molecules (a), (b), (c), (d) and a new combination (e). The aim of the second hybridization is to enrich for tester hybrids with different adaptor sequences (e). Adding a second portion of denatured driver further enrich hybrids (e) which are the differentially expressed genes. After filling in the ends of the molecules by adding two primers corresponding to the outer part of the two different adaptors, the resulting sample is amplified by PCR. Exponential amplification can only occur with tester hybrids having different adapter sequences (e). Thus the aim of the PCR amplification is to produce an ending sample called the subtracted tester (ST), which is cDNA enriched and exponentially amplified for differentially expressed transcripts. Apart from the advantage that SSH includes a normalization step that enables the detection of low abundance differentially expressed transcripts, it also yields cDNA fragments that can be used to generate a cDNA library that can be used in subsequent cDNA microarray expression profiling (Diatchenko *et al.*, 1996; Yang *et al.*, 1999).

1.1.4. Screening SSH cDNA libraries

SSH cDNA libraries do not always yield differentially expressed genes due to the nature of the SSH technique and therefore it is standard approach to screen such libraries in order to identify clones that are most likely to be differentially expressed. Therefore, as a test for quality control, patterns of gene expression can be compared with methods such as cDNA-AFLP or inverse northern blot analysis (Birch *et al.*, 1999; Mahalingam *et al.*, 2003). Apart from the fact that normalization of radioactivity membrane blots is difficult, other disadvantages are that both these methods are tedious and do not allow the level of enrichment of a transcript to be quantified. cDNA microarrays on the other hand, provide a rapid and high throughput method for the quantitative screening of an SSH cDNA library (Yang *et al.*, 1999; van den Berg *et al.*, 2004).

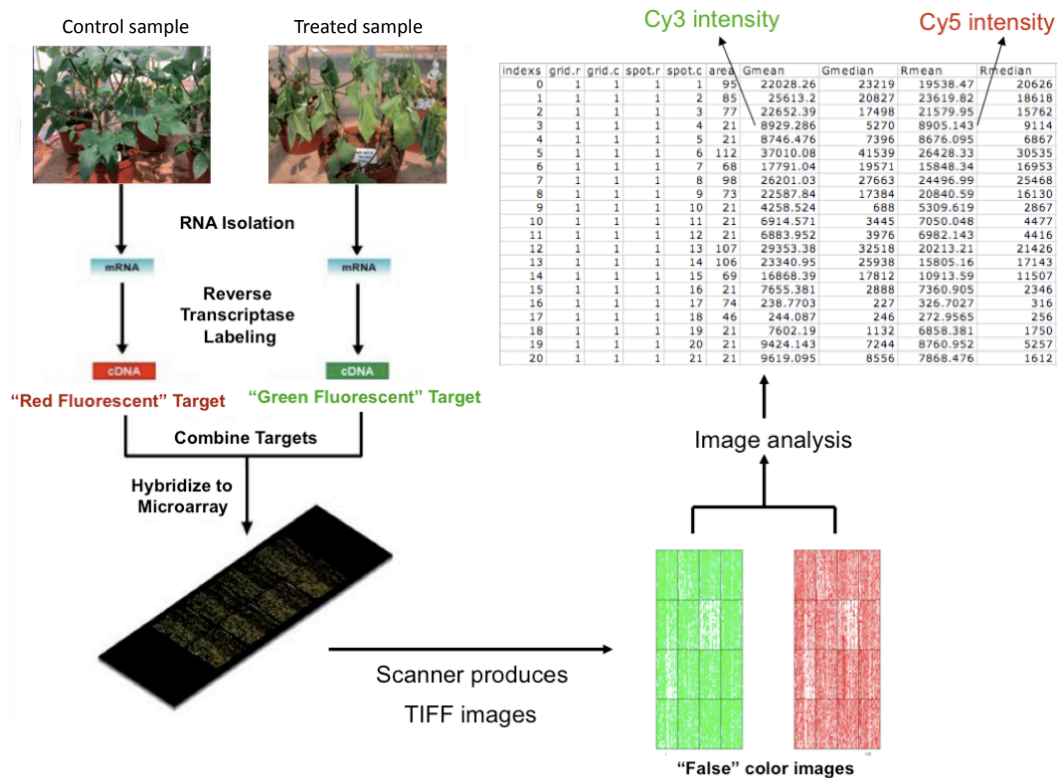


Figure 1.3: Diagram of a two-colour microarray experiment. cDNA prepared from the two samples to be compared (for example cells from a drought treated cowpea plant and cells from an untreated cowpea plant) are labeled using different fluorescent dyes, mixed in equal proportions and hybridized to the arrayed probes. After competitive hybridization, the slides are imaged using a specialized scanner to measure the amount of hybridized target at each probe. These measurements are reported as intensities by image analysis software. Before the differentially expressed genes can be identified, the data is usually adjusted using background subtraction and normalization methods, since various sources of variation need to be accounted for.

1.1.5. cDNA microarrays

Microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. Several competing microarray gene expression platforms have emerged, of which one-colour platforms such as Affymetrix GeneChips and two-colour platforms such as spotted DNA microarrays have gained increasing use and acceptance. Both these platforms have matured into complex technologies as biologists have teamed with statisticians to address the problems associated with the manipulation of large data sets (Hardiman, 2004; Park *et al.*, 2003). This dissertation focuses on two-colour microarray experiments.

Two-colour spotted microarrays can be used to measure the difference in quantity of specific nucleic acid transcripts of interest present in two samples, in effect comparing two conditions. These arrays consist of thousands of different single-stranded nucleic acid molecules, known as the probes and printed in a high-density array on a glass microscope slide using a robotic arrayer. The probes are obtained from PCR-amplifications of cDNA clones.

Figure 1.3 on the preceding page is a diagram of a typical two-colour microarray experiment. cDNA prepared from the two samples to be compared, for example cells from a drought treated cowpea plant versus cells from an untreated (water-treated) cowpea, are labeled using different fluorescent dyes, then mixed in equal proportions and hybridized to the arrayed probes. After competitive hybridization, the slides are imaged using a specialized scanner to measure the amount of hybridized target at each probe. These measurements are reported as intensities by image analysis software packages such as Genepix, Spot, ArrayVision, Imagene and TIGR_Spotfinder. The log ratio of the pixel intensities from the Cyanine Dye 3 (Cy3; green) and Cyanine Dye 5 (Cy5; red) channels for each spot is intended to be indicative of the relative abundance of the corresponding molecule in the two target samples (Smyth, 2005).

Before these data can be used in research it is usually adjusted, since various sources of variation need to be accounted for. This modification of the intensity data is commonly referred to as pre-processing and includes processes such as background correction and normalization. The objective of a microarray study is often the identification of differentially expressed genes in order to identify candidate genes that might play a role in particular biological processes related to the biological question under investigation.

1.1.6. Software for microarray data analysis

R (R Development Core Team, 2009) is a powerful statistical programming environment, made freely available through the internet under the General Public License (GPL). It provides an environment in which one can perform statistical analyses and produce graphics (Dalgaard, 2002). Rooted in R, BioConductor (Gentleman *et al.*, 2004) is a widely used open source and open development software project, for the analysis and comprehension of data arising from high-throughput experimentation in genomics and molecular biology. Limma (Linear Models for Microarray Data) is available as part of the BioConductor project and it can be installed from the R Project CRAN (Comprehensive R Archive Network) repository (Smyth, 2005). Limma provides the tools for assessing designed experiments, thereby disclosing differential expression by fitting a linear model to the expression data for each gene. Using these linear models, it is possible to analyze complex experiments involving comparisons between many RNA targets simultaneously. Shrinkage methods such as empir-

ical Bayes are used to borrow information across genes, making the analyses stable even for experiments with a small number of arrays (Smyth, 2004). Limma can be used in conjunction with other R packages, for example *affy* (Methods for Affymetrix Oligonucleotide Arrays) or *affyPLM* (Fitting Probe Level Models) for Affymetrix data, and *marray* for two-colour microarray data. *Marray* is a powerful package for quality assessment and normalization before applying *limma* functions to the data for ranking the genes in terms of differential expression. However, *limma* itself also provides input and normalization functions that support features especially useful for the linear modeling approach and is based on a more general separation between within-array and between-array normalization than *marray*. *Limma* as a command-driven package is very powerful, although the R packages *limmaGUI* (Linear Models for Microarray Data Graphical User Interface) and *affylmGUI* (Affymetrix Linear Modeling Graphical User Interface) are also available, which provide graphical user interfaces to the most commonly used functions in *limma* (Smyth *et al.*, 2008). Compared to other alternatives, the flexibility of *limma* is exceptional.

The TM4 suite (Saeed *et al.*, 2003) is an alternative to using GenePix (or any other image analysis software) for image analysis together with R and *limma* for microarray data analysis, and additional web-based tools for exploration (for example gene ontologies). It consists of four main applications: Microarray Data Manager (MADAM), TIGR_Spotfinder for segmentation/quantitation, Microarray Data Analysis System (MIDAS) for data analysis and Multiexperiment Viewer (MeV) for visualization and exploration. It also includes a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database for the organization of experiments. TM4 is free, open-source software released under the Open Source Initiative (OSI) certified Artistic license.

GEPAS (Herrero *et al.*, 2003) is a collection of web-based tools including tools for scanning slides, quantitation, normalization, quality checking, plotting, cluster analysis, classification and comparison of gene lists. The drawback, since it is a web-based tool, is that it can be slow in handling large data sets.

1.1.7. Validation of selected transcripts

Quantitative PCR (qPCR) is a powerful tool for the accurate quantification of mRNA expression levels in cells of different populations (Toegel *et al.*, 2007). Microarray results of selected transcripts can be verified by qPCR on cDNA templates, or by quantitative Reverse Transcriptase-PCR (qRT-PCR) on RNA templates.

1.2. Statistical principles in the data analysis of two-colour microarrays using limma

1.2.1. Experimental design

Experimental design before conducting a microarray experiment is crucial. It includes the choice and collection of samples; the choice of probes and array platform; the choice of controls, RNA extraction method, amplification method, labeling method, and hybridization procedures; the allocation of replicates; and the scheduling of the experiments (Smyth, 2005). In this regard, (Wit and McClure, 2004), emphasize the importance of replication, which implies the repetition of a certain experiment in order to decrease the uncertainty introduced in the experiment by systematic and random variations.

The hypothesis tested for each gene g when comparing two conditions using two-colour cDNA microarrays, is

- H_0 : gene g is not differentially expressed between the two conditions
- H_1 : gene g is differentially expressed between the two conditions

Ideally, each condition should be represented by multiple independent biological samples (biological replicates) in order to conduct statistical tests. Biological replicates represent RNA samples obtained from independent biological sources, and technical replicates represent repeated sampling of the same biological material. If only technical replicates are available, statistical testing is still possible but the scope of any conclusions drawn may be limited. Two typical constraining factors in deciding on the number of microarrays to use in an experiment, are the costs of the physical microarray and the amount of RNA available for performing the hybridization. With two-colour arrays, samples can be compared directly on the same microarray or indirectly by hybridizing each sample with a common reference sample.

The central idea behind limma, is to fit a linear model to the expression data for each gene. The expression data can be log-ratios from two-colour microarrays or log-intensities from one-channel technologies such as Affymetrix. Empirical Bayesian methods are used to borrow information across genes, providing stable results even with a small number of replicates per gene.

Analyzing two-colour spotted microarrays, a range of limma functions covering the data analysis process are available. These limma functions can be divided into 5 main categories which are in turn discussed below: functions for reading in the data; functions for exploratory data analysis; pre-processing functions including background subtraction and normalization;

the linear model and differential expression functions; and functions handling multiple testing and output.

1.2.2. Limma functions for reading two-colour data

Data importation methods should be flexible, since data comes in different formats where data is scattered across a number of fields in various files. To partly deal with this problem, limma requires the preparation of a targets file and a spot types file for each analysis. These files can be created in Microsoft (MS) Excel, but should be saved as tab delimited text files in the same directory, together with the image analysis output files (for example *.gpr* files from GenePix) and sometimes also a GAL file.

1.2.2.1. The targets file

The targets file (for example Figure 1.4a) describes, for each array, which RNA target was labeled with the Cy3 and Cy5 dyes respectively before hybridization. Each row corresponds to an individual array. The targets file should include columns labeled *Cy3* and *Cy5*, specifying the labeled RNA sample, as well as a column named *FileName* giving the names of the files containing the image analysis output (Smyth *et al.*, 2008). For ImaGene (image analysis software) the *FileName* column is split into *FileNameCy3* and *FileNameCy5*, since ImaGene stores red and green intensities in separate files. Other columns are optional. The limma function *readTargets()* reads the targets file (Figure 1.5a shows the R code).

1.2.2.2. Intensity data files

The *RGList* (Red-Green list) is a class in R defined by limma, used to store raw intensities as they are read in from the image analysis output files. The *read.maimages()* function extracts the foreground and background intensities from a series of image analysis output files, and assembles them into the components of the *RGList* (Figure 1.5b shows the R code). Usually the mean feature pixel intensities and the median feature background intensities are used, depending on the image analysis program specified by the user. The *RGList* object is designed to obey many analogies with matrices. In the *RGList*, rows correspond to spots and columns to arrays. It has components *R* (red channel foreground intensities), *Rb* (red channel background intensities), *G* (green channel foreground intensities), *Gb* (green channel background intensities), *weights* (spot quality weights), *genes* (gene names, gene IDs and spatial positions on the array), *targets* (information from the targets file), *source* (the image analysis program) and *printer* (information about the process used to print spots on a microarray for example the number of grid rows and columns, number of spots per grid, number of duplicate spots and the spacing between duplicate spots) (Smyth, 2005).

Spot quality weights

Limma calculates a weight (a value between 0 and 1) for each spot, calculated as function of the flags (from the image analysis program) associated with that spot. It indicates the reliability of the acquired intensities for each spot (Smyth *et al.*, 2008). This forms the *weights* component in the *RGList*.

1.2.2.3. The GAL file

In some cases the *genes* component in the *RGList* will not be set (after reading in the intensity data), if there is no probe information in the image analysis output files. In this case, the probe information needs to be read in separately through a GAL file. GAL files (for example Figure 1.4c) are produced by image analysis software, for example *GenePix* or *Spot*, when the array images are scanned with a GenePix 4000B scanner (Axon Instruments, USA), such as the one installed at the ACGT microarray facility at UP. The GAL file contains data columns labeled *Block*, *Column*, *Row*, *ID* and *Name*. Other columns are optional. This information can also be referred to as the “gene list”. The limma function *readGAL()* reads the GAL file (Figure 1.5c shows the R code).

Printer layout

The “printer layout” refers to the arrangement of spots and blocks of spots on the arrays (for example Figure 1.6). Each block corresponds to a print tip on the print-head of the printer/arrayer, and the number of spots in each block refers to the number of times the print-head was lowered onto the array. The *limma* function *getLayout()* determines the printer layout from the GAL file, and the *printer* component of the *RGList* is set accordingly (Figure 1.5d shows the R code).

1.2.2.4. The spot types file

The spot types file (for example Figure 1.4b) allows the user to identify different types of spots from the gene list, with rows corresponding to types of spots and the following columns: *SpotType* gives the name of the spot type, *ID* is a regular expression matching the ID column in the gene list, *Name* is a regular expression matching the Name column in the gene list, and *colour* is the *R* name for the colour to be associated with the spot type. This information is used to set the “control status” of each spot on the arrays so that plots may highlight different types of spots in an appropriate way. The limma function *readSpotTypes()* reads the spot types file (Figure 1.5e shows the R code).

The control status

The limma function `controlStatus()` specifies the type/status of each spot on the array by searching for patterns in the gene list, according to the description in the spot types file (Figure 1.5f shows the R code). This function, adds an extra column to the *genes* component of the *RGList*, called *Status*. In this way, different types of controls spots and the genes of interest can be distinguished.

a

	A	B	C	D
1	SlideNumber	FileName	Cy3	Cy5
2	40	40_up_NC_gpr.gpr	ST	UT
3	42	42_up_NC_gpr.gpr	UT	ST
4	58	58_up_NC_gpr.gpr	UD	UT
5	114	114_up_NC_gpr.gpr	UT	UD

b

	A	B	C	D
1	SpotType	ID	Name	Color
2	cDNA	*-*	cDNA	blue
3	control_Gus	control_Gus	Gus	yellow
4	control_Bar	control_Bar	Bar	red
5	control_Luc	control_Luc	Luc	green
6	control_ITS	control_ITS	ITS	brown
7	blank	BLANK	blank	NA

c

	A	B	C	D	E
1	Block	Column	Row	Name	ID
2	1	1	1	cDNA	M56_01-A1
3	1	2	1	cDNA	M56_01-A2
4	1	3	1	cDNA	M56_01-B1
5	1	4	1	cDNA	M56_01-B2
6	1	5	1	cDNA	M56_01-C1
7	1	6	1	cDNA	M56_01-C2
8	1	7	1	cDNA	M56_01-D1
9	1	8	1	cDNA	M56_01-D2
10	1	9	1	cDNA	M56_01-E1
11	1	10	1	cDNA	M56_01-E2

Figure 1.4: Example of a targets file, spot types file and part of a GAL file. The targets file (a) and the spot types file (b) can be prepared and viewed in MS Excel. These files need to be saved as tab-delimited text files before submitting to limma. The GAL file (c shows only part of a GAL file), also in tab-delimited text format, is produced by image analysis software, for example *GenePix* or *Spot* when the image is scanned with a GenePix 4000B scanner (Axon Instruments, USA).

```

> setwd("/Users/nanette/data/pearl_millet")
a > targets <- readTargets()
> RG <- read.maimages(targets$FileName, source="genepix", verbose = TRUE,
  path=path, wt.fun=wtflags(0))
b Read /Users/nanette/Documents/Bioinformatics/SSHscreen/Datasets/pearl_millet/
  summary_15_01_08/Forward/limma/ER3/58_up_NC_gpr.gpr
  Read /Users/nanette/Documents/Bioinformatics/SSHscreen/Datasets/pearl_millet/
  summary_15_01_08/Forward/limma/ER3/114_up_NC_gpr.gpr
c > RG$genes <- readGAL()
> RG$printer <- getLayout(RG$genes)
> RG$printer
$ngrid.r
[1] 6
d $ngrid.c
[1] 2
$nspt.r
[1] 6
$nspt.c
[1] 32
e > spottypes <- readSpotTypes()
> RG$genes$Status <- controlStatus(spottypes, RG)
Matching patterns for: ID
Found 1920 cDNA
Found 336 blank
f Found 12 control_Gus
  Found 12 control_Bar
  Found 12 control_Luc
  Found 12 control_ITS
  Setting attributes: values Color
  
```

Figure 1.5: R code showing the limma functions for reading the required files. Before any analysis, the working directory or environment needs to be appropriately set using the limma function *setwd()*. The following limma functions are used to read in the files for data analysis of two-colour microarrays: (a) *readTargets()* reads the targets file, (b) *read.maimages()* extracts the foreground and background intensities from a series of image analysis output files, (c) *readGAL()* reads the GAL file, (d) *getLayout()* determines the printer layout from the GAL file, (e) *readSpotTypes()* reads the spot types file and (f) *controlStatus()* specifies the type/status of each spot on the array according to the description in the spot types file.

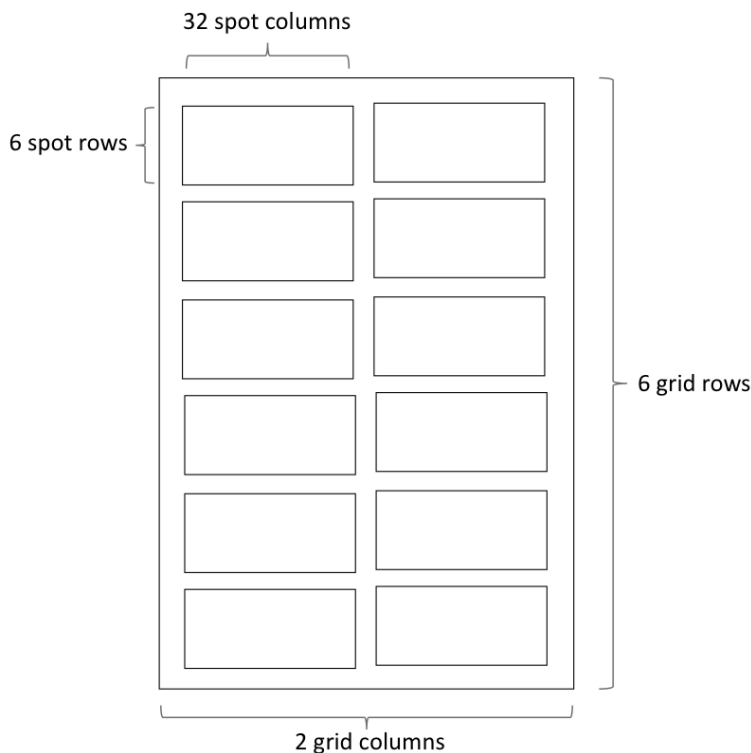


Figure 1.6: An example of the layout of a microarray slide. The printer layout for this microarray slide can be determined by the GAL file using the limma function `getLayout()` (Figure 1.5d shows the R code). The result is stored in the `printer` component in the `RGList`. There are 6 grid rows (`RG$printer$ngrid.r`) and 2 grid columns (`RG$printer$ngrid.c`). Within each block, there are 6 spot rows (`RG$printer$nspt.r`) and 32 spot columns (`RG$printer$nspt.c`).

1.2.3. Limma functions for data exploration

It is advisable to display the raw data in various ways as a quality check and also to check for unexpected effects (Smyth *et al.*, 2008). Image analysis produces four sets of probe-level data per microarray, the red and green foreground intensity measurements and also the red and green local background noise levels. The latter is measured from areas in the glass slide not containing probe and Figure 1.7 on the next page shows the region on the slide that GenePix uses for background calculation. The limma function `imageplot()` allows the user to get an idea of the variation of the red and green background values over the arrays, as illustrated in Figure 1.8 on page 17. Image plots can be used to explore any spatial effects across the microarray slides.

The red and green intensities are used to measure the relative abundance of each probe sequence in the two target samples by calculating the fold change, in the form of an expression

ratio of these intensities, i.e.

$$\text{Fold change for gene } g = \frac{(Cy_5 \text{ intensity})_g}{(Cy_3 \text{ intensity})_g} = \frac{R_g}{G_g}. \quad (1.1)$$

The logarithms of the expression ratios rather than the ratios themselves are mostly used in calculations. This is because effects on intensity of microarray signals tend to be multiplicative and the log-transformation converts these multiplicative effects into additive effects, which is easier to model (Cui and Churchill, 2003). Another advantage of using the log-transformation is that up- and down-regulated genes are treated symmetrically. Therefore, when comparing two samples using two-colour cDNA microarrays and the two samples are respectively labeled with $Cy3$ and $Cy5$ dyes, the log-2 fluorescence intensity ratio is calculated for each spot on the microarray. This is called the M-value for gene g , i.e.

$$M_g = \log_2 \left[\frac{(Cy5 \text{ intensity})_g}{(Cy3 \text{ intensity})_g} \right] = \log_2 \left(\frac{R_g}{G_g} \right). \quad (1.2)$$

MA-plots can give the user an idea of the behavior of the microarray data for each slide before normalization (raw data) and after normalization. An MA-plot can be produced with the `limma` function `MAplot()`. It is a scatter plot of log intensity ratio $M = \log_2(\frac{R}{G})$ versus the average log intensity $A = \log_2(\frac{R+G}{2})$ for each spot (A is in units of 2-fold increase in brightness), where R and G represent the red and green fluorescence intensities respectively for a specific spot (Smyth *et al.*, 2003). Figure 1.12 on page 24 is an example of MA-plots after within-array normalization for each slide.

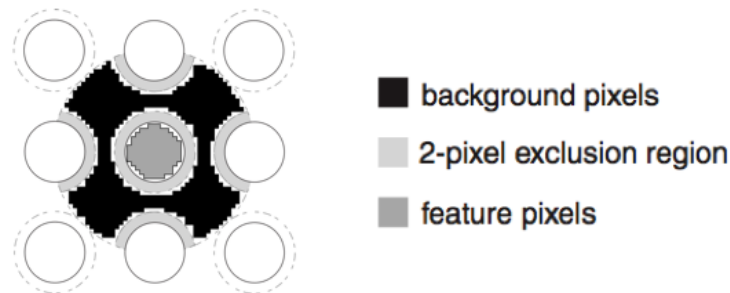


Figure 1.7: GenePix background calculation. The black region represents the pixels used for computing the background, the dark gray region represents the pixels used for the feature intensities, and the light gray region represents excluded pixels.

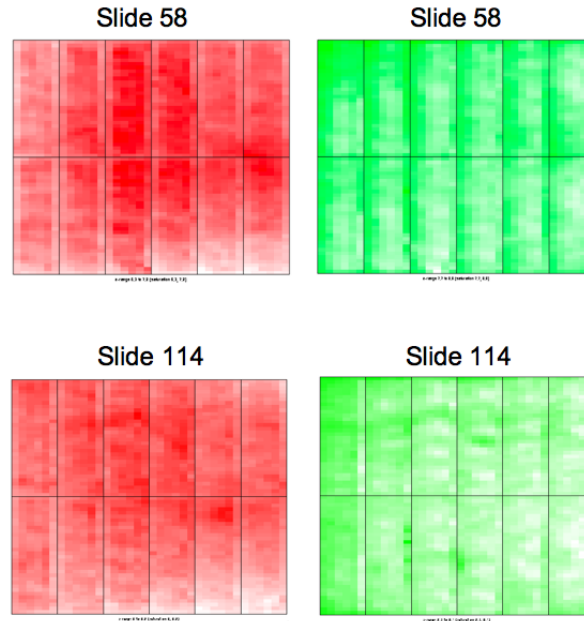


Figure 1.8: Image plots showing the variation of the red and green background values for two microarray slides, numbered 58 and 114.

1.2.4. Limma functions for pre-processing two-colour microarray data

1.2.4.1. Background correction

It is possible that some target attach to the array even when there is no probe available. This can be due to non-specific hybridization and the noise in the optical detection system. Therefore the existence of persistent background signal is a common problem. This background signal is measured irrespective of any true signal.

Most of the solutions to counteract this effect assume that the background effect is additive, that is, that the observed signal S is a sum of the true signal T and the background signal B , i.e.

$$S = T + B. \quad (1.3)$$

Because the background (B in figure 1.3) in the spot cannot be measured, the only measurement available is the background value near the spot. Figure 1.7 on the preceding page illustrates which regions the image analysis program GenePix Pro uses to calculate the background signal B and foreground signal S .

The simplest background correction method simply subtracts this value B from the observed foreground signal S to get an estimate of the true signal (Wit and McClure, 2004). This method for background correction, is called *subtract*. However this may cause other

problems such as negative corrected intensities and high variability of low intensity log-ratios. After background correction, a new *RGList* object is created in which components *R* and *G* are background corrected and components *Rb* and *Gb* are removed from the list.

The *limma* function to do the background adjustment, is *backgroundCorrect()*. The user can specify the desired background correction method. These include *subtract*, *movingmin*, *minimum*, *half*, *edwards* and *normexp*.

Subtract subtracts the estimated background value from the observed foreground value.

Movingmin replaces the background estimates by the minimum background of a moving 3×3 grid of spots (the spot self and its eight neighbors), before subtracting the background intensities from the foreground intensities. Since the interpretation of a negative gene expression value is not clear, the remaining methods are all designed to produce positive corrected intensities.

Half subtracts the background from the foreground observed value and sets any intensity, which is less than 0.5, equal to 0.5.

Minimum sets any intensity, which is zero or negative after background subtraction, to half the minimum of the positive corrected intensities for that array.

Edwards uses a log-linear interpolation method to adjust lower intensities.

Normexp fits a convolution of normal and exponential distributions to the foreground intensities, using the background intensities as a covariate. The model suggests that the observed intensities (*S* in equation 1.3), result from a convolution of the true signal (*T* in equation 1.3) and a background noise component (*B* in equation 1.3). The true signal is assumed to be exponentially distributed, i.e. $T \sim Exponential(\alpha)$, and the background noise is assumed to have a normal distribution, i.e. $B \sim N(\mu, \sigma^2)$. The expected signal given the observed foreground, i.e. $E(T|S)$, becomes the corrected intensity. Estimates of the mean, μ , and variance, σ^2 , of the normal distribution as well as the rate parameter, α , of the exponential distribution are needed to calculate this expectation. This results in a smooth monotonic transformation of the background subtracted intensities such that all the corrected intensities are positive. Figure 1.9 on the next page shows a comparison between the *subtract* and *normexp* background correction methods.

An *offset* (a constant value) can be added to the intensities before log-transforming, so that the log-ratios are shrunk towards zero at the lower intensities. This may eliminate or reverse the usual “fanning” of log-ratios at low intensities associated with local background subtraction (Figure 1.9). In other words, it will stabilize the variability of the M-values as a function of intensity. Smyth *et al.*, 2008, encourages the *normexp* background correction method with an offset, since the empirical Bayes methods implemented in the *limma* package

for assessing differential expression will yield most benefit by reducing the dependence of variability on intensity as far as possible. According to Ritchie *et al.*, 2007, this method (the *normexp* method with an offset) is found to give the lowest false discovery rate, compared to all other background correction methods.

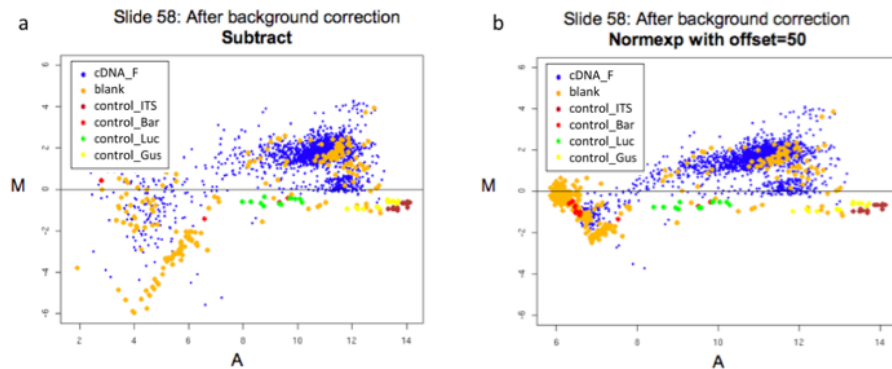


Figure 1.9: MA-plots after background correction. (a) shows that there is a high variability in the low intensity M-values after using the *subtract* method. This is known as the “fanning” of log-ratios at low intensities associated with local background subtraction. (b) shows that using the *normexp* background correction method with an offset of 50, stabilizes the variability of the M-values as a function of intensity.

1.2.4.2. Normalization

Normalization is intended to remove from the expression measures any systematic trends, which arise from the microarray technology rather than from biological differences between the probes or between the target RNA samples hybridized to the arrays. Sources of variation causing these trends may include different efficiencies of reverse transcription, labeling, or hybridization reactions, physical problems with the arrays, reagent batch effects, and laboratory conditions (Yang *et al.*, 2001; Yang *et al.*, 2002).

There are two stages of normalization. Normalization within arrays, where the M-values for each array are normalized separately, and normalization between arrays, which normalize log-ratios so that it can be compared across arrays. The limma functions *normalizeWithinArrays()* and *normalizeBetweenArrays()* are used to perform normalization.

The *MAList* (M-value A-value expression list) is a class in R defined by limma, containing all components found in the *RGList* (see a description of the *RGList* on page 11), except that the *R* and *G* values (the background corrected intensities) are replaced with *M* and *A* values on the log-2 scale. *M* and *A* calculations from the *R* and *G* values are given on page

16. The M and A values are adjusted after within-array normalization and once again after between-array normalization.

1.2.4.3. Within-array normalization methods

The different within-array normalization methods available in *limma* are *median*, *global loess*, *print-tip loess*, *composite*, *control* and *robust spline*. All these methods assume that there is a relationship between dye bias and spot intensity and aim to minimize this correlation. When fitting an overall trend line through the data points on a MA-plot, as estimated by loess regression (the orange line in Figure 1.10), it is clear that there is a gradual trend from green-bias at low intensities to red-bias at high intensities. A loess curve is a locally weighted smooth curve plotted through a set of data points using polynomial regression (Smyth *et al.*, 2003). In *loess* normalization, each M -value is adjusted by subtracting from it the value of the estimated loess curve. *Median* normalization simply subtracts the weighted median (the blue line in Figure 1.10) from the M -values for each array.

Global loess normalization

With *global loess* normalization, each M -value is normalized by subtracting from it, the corresponding value of the global loess curve (loess curve fitted through all the data points), i.e.

$$N = M - loess(A)$$

where $loess(A)$ is the global loess curve as a function of A .

Print-tip loess normalization

Print-tip loess normalization uses individual loess curves for each print-tip group and the M -values are normalized by subtracting from it, the corresponding value of the print-tip group loess curve, i.e.

$$N = M - loess_i(A)$$

where $loess_i(A)$ is the loess curve as a function of A for the i^{th} print-tip group (Smyth and Speed, 2003; Yang *et al.*, 2001). Figure 1.11 on page 23 illustrates *print-tip loess* normalization. The use of all genes per tip-group for normalization offers stability in terms of numbers of spots and flexibility in terms of estimating tip-group specific trends.

Robust spline loess normalization

Robust spline normalization is an empirical Bayes compromise between *print-tip loess* and *global loess* normalization, with 5-parameter regression splines used in place of the loess curves.

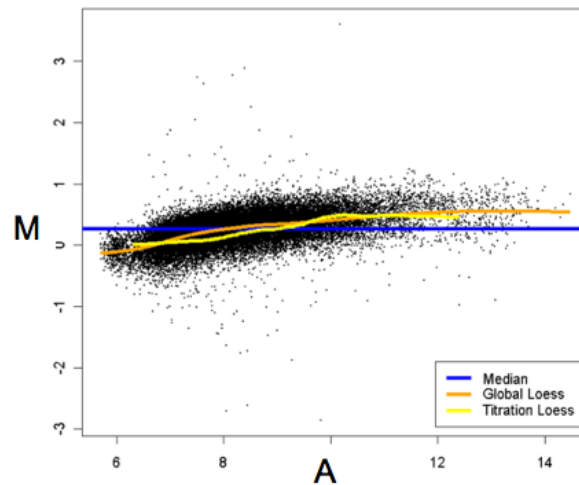


Figure 1.10: MA-plot showing three different trend lines (Smyth and Speed, 2003). The horizontal blue line shows the median of the M-values. The continuous orange curve shows the overall trend line as estimated by loess regression. The yellow curve shows the loess curve through a set of control spots known to be not differentially expressed.

1.2.4.4. Using control spots for within array normalization

Loess normalization assumes that the bulk of the probes on the array are not differentially expressed. Accordingly, it is necessary that there should be a substantial body of probes, which do not change expression levels. This assumption is valid when analyzing whole genome arrays, for example Zhou *et al.*, 2007, used *global loess* normalization in the identification of drought and high-salinity stress responsive genes using a rice whole genome oligomer microarray. If genes have been selected for being specifically expressed in one of the RNA sources, such as a cDNA library constructed with SSH, the best strategy is to include on the arrays a series of non-differentially expressed control spots, preferably spanning as wide a range of intensities as possible, such as a specially designed microarray sample pool (MSP) titration series. A MSP titration series means that the entire clone library was pooled and then titrated at a series of different concentrations (Yang *et al.*, 2002; Smyth *et al.*, 2003). Theoretically all labeled cDNA sequences should hybridize to this mixed probe sample, whereafter one can then use the *control*, *composite* or *up-weighting* within-array normalization methods to exploit the fact that the control spots are known to be non-differentially expressed.

Control-spot loess normalization

The *control* normalization method fits a global loess curve through a suitable set of control spots (as described in the previous paragraph) and applies that curve to all the other spots. In this case the estimated value of the loess curve through the control spots are subtracted from each spot's original M-value, i.e.

$$N = M - loess_{MSP}(A)$$

where $loess_{MSP}(A)$ is the loess curve through the MSP spots. The loess curve through the control spots offers security that the curve is not biased by differentially expressed genes.

Composite loess normalization

Composite loess normalization also relies on a suitable set of control spots, known to be non-differentially expressed and spanning the whole range of intensities (such as a whole library titration series, as explained above). It uses a compromise between the tip-group curves and the global titration series curve (for example a global MSP curve), i.e.

$$N = M - p(A)loess_{MSP}(A) - \{1 - p(A)\}loess_i(A)$$

where $loess_{MSP}(A)$ is the loess curve through the MSP spots and $p(A)$ is the proportion of spots on the array with A -values less than A . The idea behind this is that normalization will be increasingly based on the global MSP curve rather than the individual tip-group curves at higher intensities where the individual curves are less reliable due to the smaller number of spots (Yang *et al.*, 2002; Smyth *et al.*, 2008).

Up-weighting loess normalization

The *up-weighting loess* normalization method uses the spot quality weights found in the *RGList* (see page 11 for a description of the *weights* component in the *RGList*). This means that spots with zero weight i.e. spots which are flagged out, will be normalized, but such spots will not have any influence on the normalization of other spots. Therefore, when the arrays contain a series of spots which are known in advance to be non-differentially expressed, these spots can be given more weight in the normalization process, while other spots can be down-weighted using the limma function *modifyWeights()* (Smyth *et al.*, 2008). This *up-weighting* normalization method can be used with *global loess* or *print-tip loess* normalization. Figure 1.12 on page 24 gives an example of MA-plots after using *up-weighting print-tip loess* normalization where the cDNA clones were given zero weight and the control spots double weight.

There are some notable cases where *print-tip loess* normalization is not appropriate. For example, Agilent arrays do not have print-tip groups and *print-tip loess* normalization is unreliable for small arrays with less than, 150 spots per print-tip group. In these cases *global loess* normalization or *robust spline* normalization should rather be used.

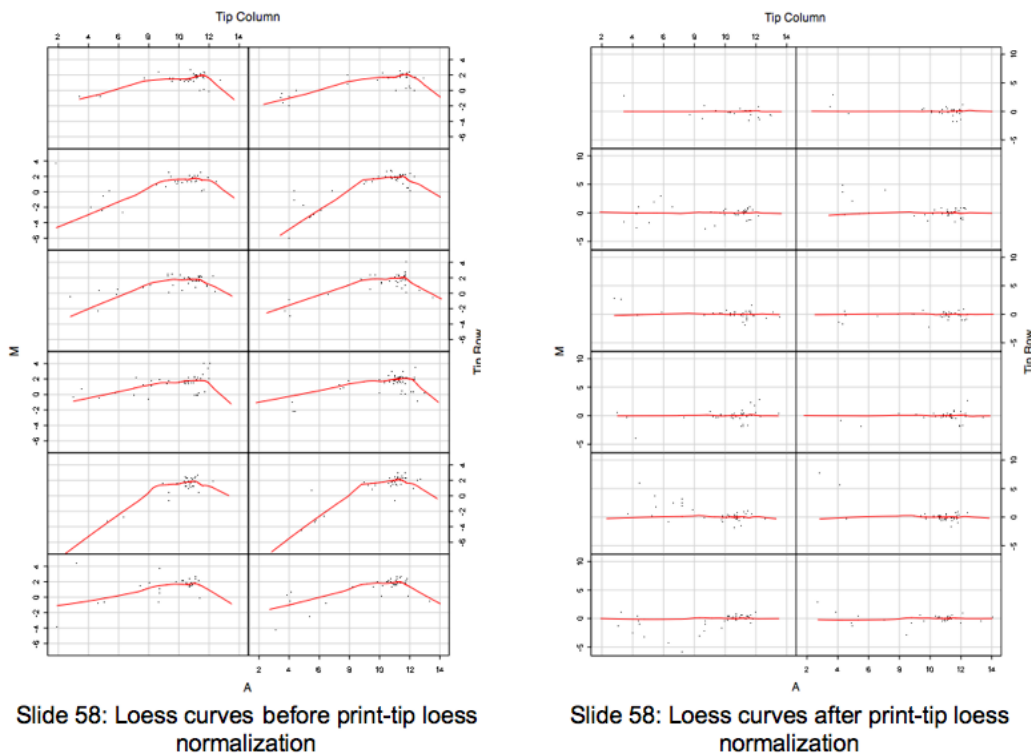


Figure 1.11: An illustration of print-tip loess normalization. A separate loess curve is fitted for each print-tip group. The curve is straightened and shifted towards the $M = 0$ axis and all data points are adjusted accordingly.

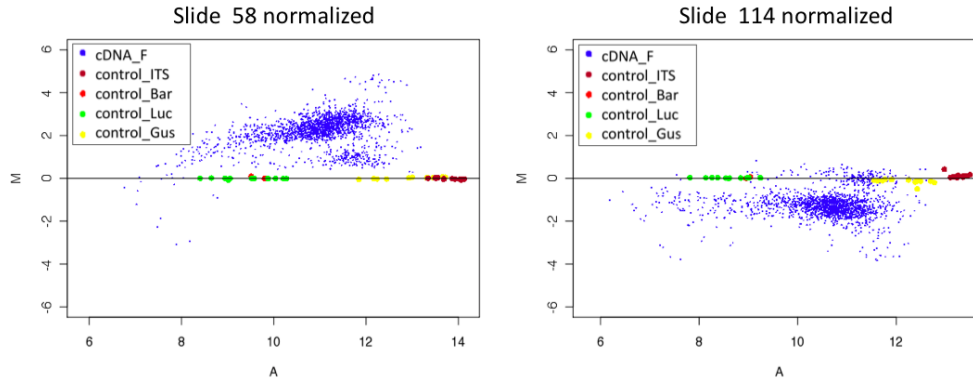


Figure 1.12: MA-plots after using *up-weighting print-tip loess* normalization, where the cDNA clones were given zero weight and the spiked-in control spots full weight. Using this method ensures that the control spots (coloured spots) are placed on the $M = 0$ line in the MA-plots, thereby adjusting the cloud of cDNAs (blue spots) accordingly for each slide.

1.2.4.5. Between-array normalization methods

The different between-array normalization methods available in *limma* are *scale*, *quantile*, *Aquantile*, *Gquantile*, *Rquantile*, *Tquantile* and *vsn*.

Scale normalization

Scale normalization may be important, since large scale differences between multiple slides can cause some slides to skew the average of log-ratios across slides. One common method of scale normalization is to divide each intensity measure by the total of the intensities on the slide, so that all slides then have the same total intensity. The *scale* between-array normalization method in *limma*, uses a more robust estimate for scale, the median-absolute-deviation (MAD). It scales the log-ratios to have the same MAD across arrays (Yang and Thorne, 2003).

Quantile normalization

The goal of *quantile normalization* is to impose the same empirical distribution of intensities to each array. Therefore *quantile* ensures that the intensities have the same empirical distribution across arrays and across channels.

Aquantile ensures that the A-values (average expression) have the same empirical distribution across arrays, leaving the M-values unchanged. Figure 1.13 on the next page is an illustration of *Aquantile* normalization, using density plots to display the distributions of the red and green channels separately for two microarrays. In Figure 1.13, (a) is the density plot before normalization, (b) shows that the distribution of A-values after *loess* within-array

normalization is the same for the red and green channels for individual arrays and (c) shows that the distribution of A-values for the red and green channels after *loess* within-array and *Aquantile* between-array normalization, is the same within and between the arrays. The biological assumption behind *Aquantile* normalization is that the distribution of A-values is similar across all arrays. This assumption is generally valid when arrays are technical or biological replicates (hybridizing the same samples to the arrays). *Tquantile* performs quantile normalization separately for the groups indicated by the targets file. *Gquantile* ensures that the green channel has the same empirical distribution across arrays, leaving the M-values unchanged. This method might be used when the green channel is a common reference throughout the experiment. In such a case the green channel represents the same target throughout, so it makes sense to force the distribution of intensities to be the same for the green channel on all the arrays, and to adjust to the red channel accordingly. *Rquantile* ensures that the red channel has the same empirical distribution across arrays, leaving the M-values unchanged.

Vsn normalization

Vsn normalization uses the *vsn* package in R. This method combines background correction and normalization into one single procedure, whereas the other methods consider background correction and normalization as separate tasks. This allows that information across arrays can be shared to estimate the background correction parameters, which are otherwise estimated separately for each array (Smyth, 2005).

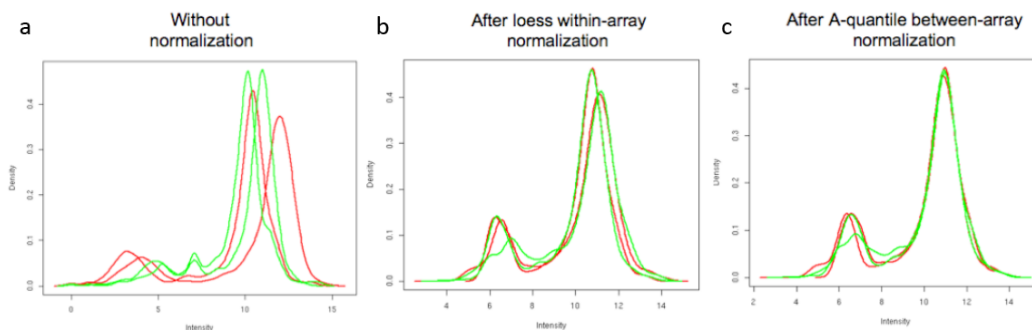


Figure 1.13: Density plots for two arrays illustrating *loess* within-array and *A-quantile* between-array normalization. These density plots (intensity versus density) display the intensity distributions of the red and green channels separately (a) without normalization, (b) after *loess* within-array normalization and (c) after *A-quantile* between-array normalization.

1.2.5. Limma functions for the identification of differentially expressed genes

1.2.5.1. Linear models

The design of any microarray experiment can be represented in terms of a gene-wise linear model. It is assumed that all the microarrays involved in the experimental design are spotted with the same set of probes, however different RNA samples can be hybridized to each slide. After normalization, it is assumed that a set of n microarrays will yield a response vector y_g for the g^{th} gene. For two-colour data, these responses are log-ratios or M-values as described on page 16.

The gene-wise linear model is of the form

$$y_g = X\alpha_g + \epsilon_g \quad (1.4)$$

where y_g is a response vector of M-values for gene g , X is a design matrix (see description on page 27), α_g is a coefficient vector, and ϵ is the vector of 'errors', also known as the residuals, which are uncorrelated random variables each with expected value 0 and residual variance σ_g^2 . The linear model describes the total variation of the expression data for gene g (y_g in equation 1.4), by partitioning it into systematic variation $X\alpha_g$ and random variation ϵ_g . It is necessary to model the systematic part of the data so that it can be distinguished from the random variation. The linear model is specified by the design matrix, X .

It is assumed that the expected value of y_g is $X\alpha_g$ i.e.

$$E(y_g) = X\alpha_g \quad (1.5)$$

and the variance of y_g is $W_g\sigma_g^2$ i.e.

$$var(y_g) = W_g\sigma_g^2 \quad (1.6)$$

where W_g is a known non-negative definite weight matrix.

Having observed or calculated values for y_g and X , the aim when fitting the linear model to the responses for each gene, is to estimate the values of the parameters α_g (the M-values/coefficients for the contrasts of interest) and σ_g^2 (the residual variance). Thus, it is of interest to obtain coefficient estimators $\hat{\alpha}_g$ and estimators s_g^2 for σ_g^2 , for each gene g .

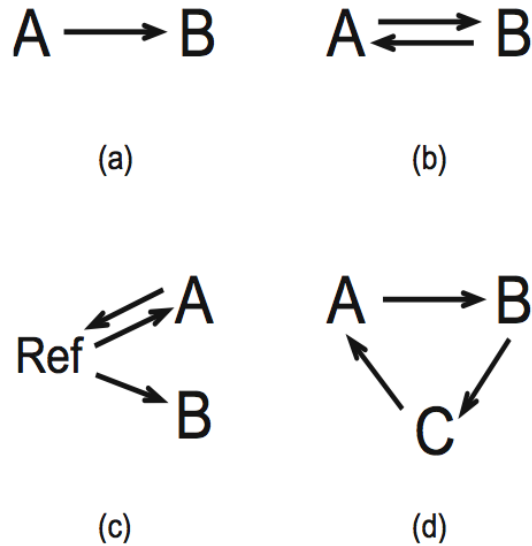


Figure 1.14: Example experimental designs for two-colour microarrays (Smyth, 2004). Each arrow represents a microarray. The arrow points towards the RNA sample which is labeled red and the sample at the base of the arrow is labeled green. The symbols A, B and C represent RNA samples to be compared.

1.2.5.2. Gene-wise linear models from experimental designs

The design matrix

Kerr and Churchill, 2001, used arrow notation to display the experimental designs of two colour microarray experiments. Figure 1.14 displays some examples of simple designs. Each arrow represents a microarray. The arrow points towards the RNA sample which is labeled red and the sample at the base of the arrow is labeled green. The symbols A, B and C represent RNA samples to be compared.

A design matrix can be used to represent such an experimental design in terms of a linear model. The design matrices constructed below for designs (a) to (d) in Figure 1.14, have rows corresponding to the arrays in the experiment and columns corresponding to contrast estimates, which is the difference between the two RNA samples being compared (Smyth, 2004).

In experiment (a), only one microarray compares RNA samples A and B. A log-2 ratio (M-value) for each gene can be calculated i.e.

$$y_g = \log_2(R_g) - \log_2(G_g) = \log_2\left(\frac{R_g}{G_g}\right) \quad (1.7)$$

where R_g and G_g are the red and green intensities for gene g . When only one array is used,

there is no need for a linear model since calculating the log-2 ratio for each gene will give an estimate for the only contrast, $B - A$, on the log-scale.

Design (b) is a dye-swap experiment leading to a very simple linear model of the form

$$y_g = X\alpha_g + \epsilon_g \quad (1.8)$$

where y_g is a vector of responses for each array, that is y_{g1} and y_{g2} are log-ratios from arrays 1 and 2 respectively. The design matrix is

$$X = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

where the negative number indicates the dye swap. The regression coefficient $\hat{\alpha}_g$ (see equation 1.8) estimates the contrast $B - A$ on the log-scale.

Design (c) compares samples A and B indirectly through a common reference RNA sample. An appropriate design matrix for this experiment is

$$X = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

which produces a linear model in which the first coefficient (corresponding to the 1st column in the design matrix) estimates the difference between A and the reference sample ($A - Ref$) while the second (corresponding to the 2nd column in the design matrix) estimates the difference of interest, $B - A$. Note that the number of coefficients being estimated should be equal to the number of RNA samples minus 1. Therefore for design (c), two coefficients are estimated.

Design (d) is a simple loop design comparing three samples. Different design matrices can be chosen corresponding to different parameterizations. One choice is

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

so that the coefficients correspond to the differences $B - A$ and $C - B$ respectively (corresponding to the two columns in X respectively).

The contrast matrix

Certain contrasts of the coefficients are assumed to be of biological interest and can be written in the form of another linear model. The contrasts of interest are defined by

$$\beta_g = C^T \alpha_g \quad (1.9)$$

where β_g correspond to the difference between the RNA samples of biological interest to the user, C is a contrast matrix (C^T is the transpose of matrix C) and α_g the coefficients already estimated using the response vector y_g and the design matrix X (see calculations of $\hat{\alpha}_g$ in examples (a) to (d) on page 27). We assume that it is of interest to test whether individual contrast values β_{gj} ('M-value' for contrast j gene g) are equal to zero. For example, with design (d) above, the experimenter might want to make all the pair-wise comparisons $B - A$, $C - B$ and $C - A$, which correspond to the contrast matrix

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Therefore, from the coefficient estimators $\hat{\alpha}_g$ and the contrast matrix C , the contrast estimators $\hat{\beta}_g$ can be determined. For each gene g and contrast of interest j (the columns in the contrast matrix) a contrast value, β_{gj} on the log-scale, can be calculated per gene, such that $\hat{\beta}_g = C^T \hat{\alpha}_g$ (Smyth, 2004).

1.2.5.3. Within-array duplicate spots

The section on *gene-wise linear models from experimental designs* on page 27 assumed only one replicate per gene on each array, or just averaging over the within-array duplicate spots. However, these spots may give valuable information. Spots that are spatially close together on the same array are likely to be highly correlated since they share many common causes for example local effects on the array surfaces, as well as hybridization and labeling effects (Smyth *et al.*, 2005). The limma function `duplicateCorrelation()` extracts more information from within-array replicate spots in microarray experiments, by estimating the strength of the correlation between them. By fitting a separate linear model to the expression data for each gene, but with a common value for the between-replicate correlation, the method greatly improves the precision with which the gene-wise variances are estimated and thereby improves inference methods designed to identify differentially expressed genes

(Smyth *et al.*, 2005). This approach can also be combined with empirical Bayes methods for moderating the gene-wise variances between genes (see section on Empirical Bayes analysis on page 33). This function can estimate the correlation between duplicate spots (regularly spaced replicate spots on the same array) or between technical replicates from a series of arrays.

1.2.5.4. Linear model fit

Gene-wise linear models using the limma function *lmFit()* are fitted through the expression data for each gene in order to estimate the contrasts of interest i.e. $\hat{\beta}_{gj}$. A *MArrayLM* object is created after the linear model fit. It is a list-based class in R, for storing the results after fitting gene-wise linear models to a batch of microarrays. Components of the *MArrayLM* object (Figure 1.15 on page 32) include:

- coefficients** a matrix containing fitted coefficients or contrasts of interest j for each gene g , i.e. $\hat{\beta}_{gj}$
- sigma** a vector containing the residual standard deviation for each gene g i.e. s_g
- stdev.unscaled** a matrix containing unscaled standard deviations of the coefficients or contrasts in interest j for each gene g i.e. $\sqrt{\nu_{gj}}$
- df.residual** residual degrees of freedom for each gene g i.e. d_g

The residual standard deviation for each gene, s_g , is calculated such that s_g^2 is an estimator for the residual variance σ_g^2 (see description of residual variance on page 26). After calculating the unscaled standard deviations of the coefficients $\sqrt{\nu_{gj}}$, the gene-wise standard error (SE_g) can be calculated i.e.

$$SE_g = s_g \sqrt{\nu_{gj}}. \quad (1.10)$$

1.2.5.5. The ordinary t-statistic

With the information in the *MArrayLM* object, a gene-wise ordinary t-statistic can be calculated as

$$t = \frac{\hat{\beta}_{gj}}{SE_g} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{\nu_{gj}}} \quad (1.11)$$

which are assumed to follow a t-distribution with d_g degrees of freedom. The ordinary t-statistic can be used to assess differential expression of gene g for contrast j , by testing the hypothesis

- $H_0 : \beta_{gj} = 0$
- $H_1 : \beta_{gj} \neq 0$

This gene-specific t-test is not affected by heterogeneity in variance across genes, because it only uses information from one gene at a time. However, because the number of replicates for each condition is usually small it may have low power, which is the probability that a real effect can be identified by a statistical test. Although the t-test is not affected by heterogeneity in variance across genes, it is a disadvantage that the variances estimated from each gene are not stable. For example, if the estimated variance for one gene is small i.e. small SE_g , by chance, the t-statistic can be large even when the corresponding $\hat{\beta}_{gj}$ value is small.

The t-distribution, first derived by Student (William Sealy Gosset), 1908, is a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. A gene-wise t-test will only be appropriate if the observations (replicate expression values of a gene) are normally distributed. Although this is not always the case, all calculations for identifying differentially expressed genes in limma are based on this assumption. The distributional assumptions made by Smyth, 2004 about the data can be summarized by

$$\hat{\beta}_{gj} | \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, \nu_{gj} \sigma_g^2)$$

and

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2$$

where d_g is the residual degrees of freedom for the linear model for gene g . Under these assumptions the ordinary t-statistic in equation 1.11, follows an approximate t-distribution on d_g degrees of freedom. The residual standard deviation for each gene, s_g in equation 1.11, is the same as the estimate for the residual variance of the error term in the linear model equation 1.4.

Since the standard error is hard to estimate and subject to inconsistent fluctuations when sample sizes are small, there are a few modified versions of the t-test, attempting to obtain more stable standard error estimates.

1.2.5.6. Variations on the ordinary t-statistic

The global t-test, uses an estimate of the standard error that is pooled across all genes, SE , but it is assumed that the variance is homogeneous between different genes. The global t-test ranks genes in the same order than when simply calculating the average M-value for gene g across the arrays, since it does not adjust for individual gene variability. The global t-test statistic is

$$t = \frac{\hat{\beta}_{gj}}{SE}. \quad (1.12)$$

In the *significance analysis of microarrays* (SAM) (Tusher *et al.*, 2001) version of the t-test, a small positive constant is added to the denominator, preventing it from getting too small. The SAM test statistic is

$$S = \frac{\hat{\beta}_{gj}}{(SE_g + c)} \quad (1.13)$$

where the constant c can be taken to be the 90th percentile SE_g value, that is the value below which 90% of the SE_g values may be found. With this modification, genes with small $\hat{\beta}_{gj}$ values will not be selected as significant.

Another modification of the t-test, is the regularized t-test which combines information from gene-specific and global average standard error estimates by using a weighted average of the two as the denominator for a gene-specific t-test.

```

> fit
An object of class "MArrayLM"
$coefficients
[1] 1.8097996 1.9702763 2.1004593 0.3788751 1.9712102
955 more elements ...

$stdev.unscaled
[1] 0.7036972 0.7036972 0.7036972 0.7071068 0.7036972
955 more elements ...

$sigma
[1] 3.763279 3.358991 1.972522 0.562123 1.753468
955 more elements ...

$df.residual
[1] 3 3 3 1 3
955 more elements ...
  
```

Figure 1.15: Components of the *MArrayLM* object in limma. This is output in the R computing environment, after fitting the linear model, using the limma function *lmFit()*. The output is stored in an object called *fit*.

1.2.5.7. Empirical Bayes analysis

In the Bayesian approach to statistics, the problem of estimating some probability is based on measurements of the data, a model for these measurements called the likelihood, and some model for the prior beliefs about the system called the prior. The prior distributions usually include some hyperparameters. Empirical Bayes methods employ the complete set of empirical data to make inferences about the prior and then use this result in the likelihood to produce estimates of individual measurements.

Empirical Bayes methods can be utilized to take advantage of the parallel nature of the inference in microarrays. This approach allows compensating possibilities for borrowing information from all the genes, which can assist in inference about each gene individually. Lonnstedt and Speed, 2002, took a parametric empirical Bayes approach using a simple mixture of normal models and a conjugate prior and derived an expression for the posterior odds of differential expression for each gene (the posterior odds is also known as the B-statistic; see detail on page 35). The posterior odds expression has proved to be a useful means of ranking genes in terms of evidence for differential expression. Smyth, 2004, reformulated this posterior odds statistic in terms of a moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations. This is the difference between the moderated t-statistic described on the following page and the ordinary t-statistic described on page 30. The moderated t-statistic results in a shrinkage of the gene-wise residual sample variances towards a common value, resulting in far more stable inference, even when the number of arrays are small.

Given a series of related parameter estimates and standard errors, after the linear model fit with *lmFit()* (as described on page 30), the limma function *eBayes()* compute moderated t-statistics, moderated F-statistics, and log-odds of differential expression (B-statistic) by empirical Bayes shrinkage of the standard errors towards a common value. These statistics will be added as extra components to the *MArrayLM* object produced during the linear model fit, including:

t vector of moderated t-statistics (see equation 1.14 on the next page)

p.value vector of p-values corresponding to the moderated t-statistics

s2.prior estimated prior value for the residual variance σ^2 i.e. s_0^2

df.prior degrees of freedom associated with s2.prior i.e. d_0

s2.post vector giving the posterior residual variance \tilde{s}_g^2 (see equation 1.15 on the following page)

lods vector giving the log-odds of differential expression (the B-statistic or the posterior odds of differential expression; see equation 1.19 on page 36)

F vector of moderated F-statistics for testing all contrasts simultaneously equal to zero
F.p.value vector giving p-values corresponding to F

1.2.5.8. The moderated t-statistic

The moderated t-statistic (t column in the top table; Table 1.1) is the ratio of the M-value to its standard error, as for the ordinary t-statistic described on page 30. Although it has the same interpretation as an ordinary t-statistic, the standard errors have been moderated across genes, i.e. shrunk towards a common value. It has the effect of borrowing information from the ensemble of genes to aid with inference about each individual gene (Smyth, 2004; Bruland *et al.*, 2007). The moderated t-statistic is defined by

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{\nu_{gj}}} \quad (1.14)$$

where \tilde{s}_g is the posterior residual standard deviation for gene g (see equation 1.15), $\hat{\beta}_{gj}$ is the contrast of interest j for each gene g as defined for the ordinary t-statistic on page 30 and $\sqrt{\nu_{gj}}$ is the unscaled standard deviation of the contrast in interest j for each gene g as defined for the ordinary t-statistic (also see the *MArrayLM* object description on page 30).

The posterior residual standard deviations \tilde{s}_g is defined as

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \quad (1.15)$$

where s_g^2 is an estimator for the residual variances with d_g degrees of freedom and s_0^2 is a prior estimator for the residual variance with d_0 degrees of freedom.

This statistic represents a hybrid classical/Bayes approach in which the posterior variance (equation 1.15) has been substituted into the classical t-statistic in place of the usual sample variance. The moderated t-statistic is shown to follow a t-distribution under the null hypothesis $H_0 : \beta_{gj} = 0$ (see hypothesis on page 30) with degrees of freedom $d_g + d_0$ (Smyth, 2004). The larger degrees of freedom for \tilde{t}_{gj} compared to t_{gj} reflect the extra information, which is borrowed from the ensemble of genes for inference about each individual gene.

1.2.5.9. The p-value associated with the moderated t-statistic

The p-value (*p-value* in the top table; Table 1.1) is associated with the calculated value of the moderated t-statistic. In a microarray experiment, for n genes there are n pairs of mutually exclusive hypotheses, one for each gene.

In the top table in Table 1.1 on page 37 the first gene has an observed moderated t-statistic $t = 6.58$ with an associated p-value of $4.56e - 11$. According to the definition of a p-value in

Steyn *et al.*, 1994, it can be thought of as the probability of getting such an extreme or more extreme observed t-statistic ($t \geq 6.58$), by chance, if the gene is actually not differentially expressed (if H_0 is true).

1.2.5.10. The B-statistic (posterior odds)

A simplified version

The B-statistic is the posterior log odds (or logit) of the event that gene g is differentially expressed. The *odds* in favor of an event, is described to be the quantity $\frac{p_{gj}}{1-p_{gj}}$, where p is the probability of the event. Therefore, if p_{gj} is the probability that gene g is differentially expressed, then $1 - p_{gj}$ will be the probability that gene g is not differentially expressed and according to the definition of *log odds*, a simplified version of the B-statistic is

$$B_{gj} \approx \log\left(\frac{p_{gj}}{1-p_{gj}}\right). \quad (1.16)$$

Suppose for example that $B_{gj} = 1.5$. Then the odds of differential expression is $e^{1.5} = 4.48$, i.e. four and a half to one. Solving for p_{gj} in equation 1.16, gives

$$p_{gj} = \frac{4.48}{4.48 + 1} = 0.82.$$

This result indicates that there is a 82% probability that this gene is differentially expressed. A B-statistic of zero corresponds to a 50-50 chance that the gene is differentially expressed, since $\log_{10}(50/50) = \log_{10}(1) = 0$ (Smyth *et al.*, 2008).

The B-statistic

For any given contrast j , limma assumes that a β_{gj} is non-zero with known probability, p_j (the expected proportion of truly differentially expressed genes), i.e.

$$P(\beta_{gj} \neq 0) = p_j. \quad (1.17)$$

For those genes with $\beta_{gj} \neq 0$ (differentially expressed genes), prior information on the coefficient is assumed such as an unscaled variance of ν_{0j} . The B-statistic (B column in the top table; Table 1.1) is the log-odds of differential expression. The odds that the g^{th} gene has non-zero β_{gj} , that is the odds that the g^{th} gene is differentially expressed, is

$$O_{gj} = \frac{p_j}{1-p_j} \left(\frac{\nu_{gj}}{\nu_{gj} + \nu_{0j}} \right)^{1/2} \left(\frac{\tilde{t}_{gj}^2 + d_0 + d_g}{\tilde{t}_{gj}^2 \frac{\nu_{gj}}{\nu_{gj} + \nu_{0j}} + d_0 + d_g} \right)^{(1+d_0+d_g)/2} \quad (1.18)$$

where p_j is the expected proportion of truly differentially expressed genes (see equation 1.17), ν_{gj} is the unscaled standard deviations of the coefficients, ν_{0j} is the prior unscaled standard deviation of the non-zero coefficients (genes assumed to be differentially expressed), \tilde{t}_{gj} is the moderated t-statistic as specified in equation 1.14 on page 34 (\tilde{t}_{gj}^2 is squared value of the observed moderated t-statistic), d_g is the degrees of freedom for gene g and d_0 the prior degrees of freedom (see components of the *MArrayLM* object on page 30). Therefore the B-statistic defined as

$$B_{gj} = \log(O_{gj}) \quad (1.19)$$

is on a scale that is easy to work with and is useful for ranking genes in order of evidence for differential expression (Smyth, 2004).

Volcano plot

A volcano plot (M-values versus B-statistics) summarizes both the fold-change and the log-odds of differential expression. Genes with statistically significant differential expression, will lie above a horizontal threshold line and genes with large fold-change values will lie outside a pair of vertical threshold lines. Therefore significant genes identified will tend to be located in the upper left or upper right parts of the plot.

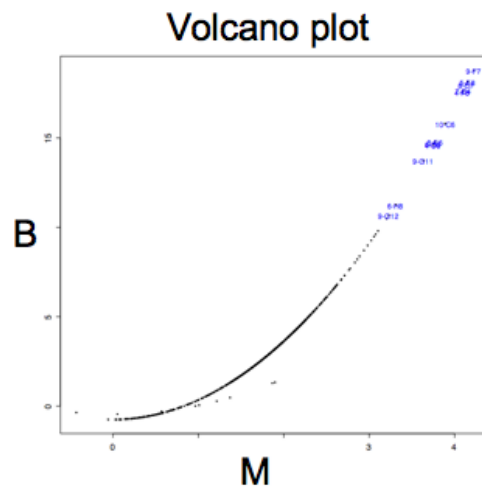


Figure 1.16: The volcano plot. The *volcano plot* is an effective graph summarizing both the fold-change (M-values on the x-axis) and the B-statistic (on the y-axis). Genes with statistically significant differential expression according to the gene-specific B-statistic will lie above a horizontal threshold line. Genes with large fold-change values will lie outside a pair of vertical threshold lines. Therefore significant genes identified will tend to be located in the upper left or upper right parts of the plot (Cui and Churchill, 2003).

Table 1.1: An example of a limma top table (only the first 22 rows are displayed).

Block	Column	Row	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
7	11	5	M56_09-F7	4,22	11,72	6,58	4,6E-11	3,2E-08	14,66
7	10	3	M56_05-E8	4,16	11,66	6,48	9,3E-11	3,2E-08	13,99
7	21	4	M56_08-C7	4,14	11,96	6,45	1,1E-10	3,2E-08	13,82
3	8	4	M56_07-D4	4,10	12,19	6,39	1,6E-10	3,2E-08	13,46
7	28	2	M56_04-F8	4,09	12,03	6,38	1,8E-10	3,2E-08	13,37
5	22	5	M56_10-C6	3,89	11,52	6,06	1,3E-09	2,0E-07	11,47
5	26	4	M56_08-E6	3,77	11,83	5,88	4,2E-09	5,0E-07	10,40
9	23	3	M56_06-D9	3,75	12,52	5,85	4,9E-09	5,0E-07	10,24
7	24	2	M56_04-D8	3,75	12,10	5,85	5,1E-09	5,0E-07	10,22
11	13	5	M56_09-G11	3,64	10,82	5,67	1,4E-08	1,3E-06	9,24
7	32	3	M56_06-H8	3,31	11,98	5,16	2,4E-07	2,0E-05	6,59
11	14	5	M56_09-G12	3,23	9,04	5,04	4,7E-07	3,5E-05	5,98
9	20	1	M56_02-B10	3,20	10,88	4,99	5,9E-07	4,0E-05	5,76
5	12	2	M56_03-F6	3,10	11,80	4,84	1,3E-06	8,3E-05	5,02
5	12	3	M56_05-F6	3,08	11,38	4,80	1,6E-06	9,6E-05	4,82
1	14	5	M56_09-G2	3,06	11,85	4,77	1,9E-06	1,0E-04	4,69
7	10	2	M56_03-E8	3,02	11,30	4,72	2,4E-06	1,3E-04	4,45
7	25	2	M56_04-E7	2,98	11,93	4,65	3,3E-06	1,6E-04	4,16
3	30	2	M56_04-G4	2,94	12,20	4,58	4,6E-06	2,1E-04	3,85
9	6	1	M56_01-C10	2,89	11,49	4,50	6,7E-06	3,0E-04	3,50
1	16	5	M56_09-H2	2,86	11,65	4,46	8,2E-06	3,4E-04	3,32
7	9	3	M56_05-E7	2,83	11,25	4,42	1,0E-05	4,0E-04	3,13

1.2.5.11. Estimation of hyperparameters

The statistics β_{gj} and \tilde{t}_{gj} depend on the hyperparameters d_0 , s_0^2 , ν_{0j} and p_j . A fully Bayesian approach would be to allow the user to choose these parameters, instead Smyth, 2004, takes an empirical Bayes approach in which these parameters are estimated from the data. Consistent closed form estimators are derived for the hyperparameters in the model. These estimators have robust behavior even for small numbers of arrays and allow for incomplete data arising from spot filtering or spot quality weights.

Smyth, 2004, estimates d_0 and s_0^2 from the observed sample variances s_g^2 ; and ν_{0j} from the moderated t-statistics \tilde{t}_{gj} assuming d_0 and p_j to be known.

Note that the data contain considerably more information about d_0 and s_0^2 than about the ν_{0j} or p_j , because all the genes contribute to estimation of d_0 and s_0^2 whereas only those which are differentially expressed contribute to estimation of ν_{0j} and p_j , and even that indirectly as the identity of the differentially expressed genes is unknown (Smyth, 2004). Therefore the estimation of ν_{0j} and p_j is rather unstable in that estimates on the boundaries $p_j = 0$, $p_j = 1$

or $\nu_{0j} = 0$ have positive probability and these boundary values lead to degenerate values for the posterior odds statistics B_{gj} .

In limma, the user is allowed to set the value for p_j (the default used in limma is $p_j = 0.01$) and also place limits on the possible values for the standard deviation of log-2 fold changes for differentially expressed genes, $\sqrt{\nu_{0j}s_0}$. By default the lower and upper limits are set at 0.1 and 4.

1.2.5.12. Comparison of the B-statistic and the moderated t-statistic

The posterior odds require estimates of all four hyperparameters, including a prior guess from the user of the expected proportion of differentially expressed genes p_j . The moderated t-statistic on the other hand does not require knowledge of the proportion of differentially expressed genes, nor does it make any assumptions about the magnitude of differential expression, in that it only depend on estimation of two hyperparameters, d_0 and s_0^2 . Therefore the posterior odds (B-statistic) is useful for ranking genes in order of evidence for differential expression, however it is advisable to base gene selections on the p-value associated with the moderated t-statistic (Smyth, 2004).

The B-statistic is automatically adjusted for multiple testing by assuming that 1% of the genes, or some other percentage specified by the user, are expected to be differentially expressed. If there are no missing values in the data, then the moderated t-statistics and the B-statistics will rank the genes in exactly the same order (Smyth *et al.*, 2008).

1.2.5.13. The moderated F statistic

The moderated t-statistic leads naturally to the moderated F-statistic, which can be used to test hypotheses about any set of contrasts simultaneously. For example we can test all contrasts for a given gene equal to zero, i.e. $H_0 : \beta_g = 0$, where β_g is the vector of all estimated effects of the contrasts of interest. Appropriate quadratic forms of moderated t-statistics follow F-distributions. The moderated F statistic is equivalent to a one-way ANOVA for each gene except that the residual mean squares and residual degrees of freedom have been moderated across genes (similar to methodology used for the moderated t-statistic) (Smyth, 2004).

1.2.5.14. A summary of the advantages of the moderated t-statistic

Smyth, 2004, highlights four advantages of the moderated t-statistic:

- Compared to the B-statistic, the number of hyperparameters which need to be estimated are reduced (see the section on hyperparameters on the previous page).

- The B-statistic requires prior knowledge of the proportion of differentially expressed genes as described on the preceding page; this is not required when calculating the moderated t-statistic.
- The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom ($d_g + d_0$ degrees of freedom; see page 34) compared to the ordinary t-statistic (d_g degrees of freedom; see page 30).
- The moderated t-statistic inferential approach, extends to accommodate tests involving two or more contrasts, through the use of moderated F-statistics (as described on the previous page).

1.2.5.15. Summary of statistical variables used in section 1.2.5

y_g	Response vector of M-values for gene g
α_g	Vector of coefficients for gene g
β_g	Vector of contrasts for gene g
ϵ_g	Vector of residuals for gene g
X	Design matrix
C	Contrast matrix
$\hat{\alpha}_g$	Vector of coefficient estimators for gene g
$\hat{\beta}_g$	Vector of contrast estimators for gene g
σ_g^2	Vector of residual variances for gene g
s_g^2	Vector of residual variance estimators for gene g
s_0^2	Vector of prior values for the residual variances for gene g
s_g	Vector of residual standard deviations for gene g
\tilde{s}_g	Vector of posterior residual variances for gene g
SE_g	Vector of standard errors for gene g
d_g	Vector of residual degrees of freedom for gene g (associated with s_g^2)
d_0	Vector of prior degrees of freedom for gene g (associated with s_0^2)
p_j	Expected proportion of truly differentially expressed genes
ν_{gj}	Matrix of unscaled standard deviations of the coefficients/contrasts of interest j for gene g
ν_{0j}	Matrix of unscaled standard deviations of the non-zero coefficients for gene g
$\hat{\beta}_{gj}$	Matrix of fitted/estimated contrasts of interest j for gene g
\tilde{t}_{gj}	Matrix of moderated t-statistics of the coefficients/contrasts of interest j for gene g
B_{gj}	Matrix of B-statistics (posterior odds) of the coefficients/contrasts of interest j for gene g

1.2.6. Limma functions for handling multiple testing and output

1.2.6.1. The multiple testing problem

Performing a hypothesis test for each gene on a microarray simultaneously, will increase the probability of finding one of the tests to be significant; that is, the p-values tend to be exaggerated (Dalgaard, 2002; Feise, 2002). Applying this consideration to the example with a t -statistic of 6.58 and a p-value of $4.56e - 11$ (see example on page 34), the probability of getting such an extreme observed t -statistic ($t \geq 6.58$), by chance, if the gene is actually not differentially expressed (if H_0 is true), will increase if many tests are performed. Therefore, the p-value should in fact be larger. In other words, if each hypothesis is rejected at some fixed posterior probability or fixed p-value and the number of hypotheses grows, then it becomes more and more likely that at least one null hypothesis will be falsely rejected, being a false positive (Feise, 2002; Wit and McClure, 2004).

1.2.6.2. Error rates

In a microarray setting where thousands of tests are considered simultaneously, generally the null hypothesis is either correctly rejected, indicating that a truly differentially expressed gene is differentially expressed, which is a true positive (T_P), or correctly accepted, indicating that a truly non-differentially expressed gene is not differentially expressed, which is a true negative (T_N). However, Table 1.2 on page 42 describes the two possible errors when testing a hypothesis. The null hypothesis can be wrongly rejected which is a false positive (F_P), or wrongly accepted which is a false negative (F_N).

There is a trade-off between the probabilities of the two types of errors. By reducing the significance level α (the cut-off value for the p-value; a p-value smaller than α is assumed to be significant) and thereby reducing the probability of a false positive, one tends to reduce the so-called power of a test, which in turn increases the probability of a false negative.

Microarray experiments are often performed with a small number of biological replicates, resulting in low statistical power for detecting differentially expressed genes and consequently high false positive rates. According to Wei *et al.*, 2004, with careful experimental design it is possible to maximize the statistical power of the test while balancing resource allocation. For example using twice as many independent biological replicates, are preferable to dye swapped technical replicates and will increase the power to detect biologically significant gene expression differences.

- The false positive rate, FPR, is defined as the probability of a false positive.
- The false negative rate, FNR, is defined as the probability of a false negative.
- The power of a test is defined as: $1 - FNR$

1.2.6.3. FPR

Most traditional methods for controlling these error rates, focus on controlling the FPR, which is the expected fraction of false positives i.e.

$$FPR = E\left[\frac{F_P}{n_0}\right].$$

The p-value itself controls the FPR by comparing it to a pre-specified significance level α . Thus for a single hypothesis test, the probability of a false positive is less than α (Wit and McClure, 2004).

1.2.6.4. FWER

The *family wise error rate* (FWER) is the probability that among all the genes that are not differentially expressed, at least one is incorrectly classified as differentially expressed, thus

$$FWER = P(F_P > 0).$$

Methods for controlling the FWER are often used in practice. However, the FWER is a very conservative error rate and especially when working with a large number of tests, it doesn't make sense to require that the probability of making even only one false rejection should be small. *Bonferroni*, *Holm*, *Hochberg* and *Hommel's* methods for multiple testing are designed to give strong control of the FWER.

The *Bonferroni* correction multiply the p -values by the number of comparisons, thus classifying all genes that have an associated p -value less than $\frac{\alpha}{n}$ as differentially expressed. In this case, if the resulting adjusted p -value is larger than 1, the adjusted p -value is set to 1. The unmodified *Bonferroni* correction however, is dominated by *Holm's* method, which is valid under arbitrary assumptions.

In *Holm's* method, only the smallest p -value needs to be corrected by the full number of tests, say n (thus multiply the smallest p -value with n), the second smallest p -value needs to be corrected by $n-1$, etc. This pattern is followed unless the corrected p -value is smaller than the previous answer, since the order of the p -values should be unaffected by the adjustment.

Hochberg's and *Hommel's* methods are valid when the hypothesis tests are independent or when they are non-negatively associated. *Hommel's* method is more powerful than *Hochberg's*, but the difference is usually small and the *Hochberg* p -values are faster to compute. (Wit and McClure, 2004)

Table 1.2: Numbers of correct and incorrect conclusions of n hypothesis tests (Wit and McClure, 2004). The null hypothesis can be wrongly rejected which is a false positive (F_P), or wrongly accepted which is a false negative (F_N).

	Declared “not differentially expressed”	Declared “differentially expressed”	Total
Truly “not differentially expressed”	T_N	F_P	n_0
Truly “differentially expressed”	F_N	T_P	$n - n_0$
Total	$n - S$	S	n

T_N : the number of true negatives (genes truly not differentially expressed and correctly declared as being not differentially expressed)

F_P : the number of false positives (genes truly not differentially expressed but wrongly declared as differentially expressed)

F_N : the number of false negatives (genes truly differentially expressed but wrongly declared as not differentially expressed)

T_P : the number of true positives (genes truly differentially expressed and correctly declared as being differentially expressed)

n : Total number of genes in the data set

n_0 : the total number of truly not differentially expressed genes

$n - n_0$: the total number of truly differentially expressed genes

S : the total number of genes declared to be differentially expressed

$n - S$: the number of genes declared to be not differentially expressed

1.2.6.5. FDR

The *false discovery rate* (FDR) is the expected number of non differentially expressed genes among those that are declared “differentially expressed”, thus

$$FDR = E\left[\frac{F_P}{S}\right]$$

(Table 1.2). In other words, the FDR is the expected proportion of false discoveries amongst the rejected hypotheses. The FDR is a less stringent condition than the FWER, and is in spirit closer to the FPR. (Ge *et al.*, 2003; Wit and McClure, 2004)

Methods controlling the FDR are very popular, including *Benjamini & Hochberg* (BH) and *Benjamini & Yekutieli* (BY). These methods tend to be more suited for studies of an exploratory, rather than confirmatory nature.

1.2.6.6. Top tables

The *limma* function *topTable()* extracts a Table of the top-ranked genes from a linear model fit after empirical Bayes analysis. There will be a top table for each contrast j that the user is interested in and within each top table, each row corresponds to a different gene g . Table 1.1 on page 37 is an example of a *limma* top table.

The *Block*, *Column*, *Row* and *ID* columns in the top table specify the spatial positions of the probes on the arrays linked to the corresponding gene IDs.

The *logFC* (M-value) column in the top table is the contrast estimate $\hat{\beta}_{gj}$ for each gene. For example, if the contrast j compares two treatment conditions, the *logFC* or M-value is the \log_2 fold-change between those two conditions for gene g . For detail on $\hat{\beta}_{gj}$ see page 30.

The *AveExp* (A-value) column in the top table is the mean log-expression level for that gene across all channels and all arrays in the linear model fit. It is the mean log-expression for all arrays in the experiment, even if only a smaller number of arrays are involved in the contrast of interest.

The t and B columns are the moderated t-statistic for each gene and the log-odds that that gene is differentially expressed respectively. For details on the moderated t-statistic and the B-statistic, see the sections on pages 34 and 35.

The *P.Value* column gives the original p -value for each gene associated with the moderated t-statistic, whereas the column *adj.P.Val* gives the adjusted p -value after an adjustment for multiple testing has been performed. For details on the adjustment for multiple testing, see the section on page 40.

1.3. The background to SSHscreen

1.3.1. Using MS Excel to calculate enrichment ratios ER1 and ER2

Initially, van den Berg *et al.*, 2004, used *MS Excel* to calculate ER1 and ER2 values for each clone after the resulting SSH forward library (*ST*) was probed onto a set of glass microarray slides. As indicated in Figure 1.17 on page 45, ER1 and ER2 can respectively be presented as $ST : UD$ and $ST : UT$. Hence, to determine an ER1 value for each clone, two dye-swapped slides were hybridized with *ST* and *UD*, including replicate spots on each slide. An ER1 value can be calculated for each clone by \log_2 -transforming the value of the *ST* fluorescence intensity divided by the *UD* fluorescence intensity, after averaging the fluorescence intensities of the replicate spots in the *Cy3* and *Cy5* channels separately on each slide. Global normalization of the data for the cyanine dye effect, was performed using a control gene set to calculate normalization functions c and c' for the pair of dye swap slides,

and van den Berg *et al.*, 2004, calculated an ER1 value for each clone, incorporating the dye swap data as well as the normalization functions c and c' , using the formula

$$\frac{1}{2}[(\log_2(\frac{ST(Cy3)}{UD(Cy5)}) - c) - (\log_2(\frac{UD(Cy3)}{ST(Cy5)}) - c')]$$

which simplifies to

$$\frac{1}{2}[\log_2(\frac{ST(Cy3)}{UD(Cy5)} * \frac{ST(Cy5)}{UD(Cy3)}) - c + c'].$$

The same procedure was followed in order to determine an ER2 value for each clone. Two dye-swapped slides were hybridized with ST and UT, also including replicate spots on each slide. An ER2 value can be calculated for each clone by log-2 transforming the value of the ST fluorescence intensity divided by the UT fluorescence intensity, after averaging the fluorescence intensities of the replicate spots in the *Cy3* and *Cy5* channels separately on each slide. Following the ER2 calculation which was done in the same fashion than the ER1 calculation, van den Berg *et al.*, 2004, showed the ER1 values plotted against the ER2 values (see Figure 3.8 on page 106). The diagonal line indicates clones derived from transcripts of equal abundance in UD and UT (i.e. $ER1 = ER2$ which implies that $UD = UT$). Clones lying above the diagonal line represent transcripts that were induced upon treatment ($ER1 > ER2$ implying that $UD < UT$) i.e. up-regulated clones, while those lying below the line indicate transcripts that escaped the subtraction ($ER1 < ER2$ implying that $UD > UT$) i.e. down-regulated clones. Clones above the diagonal line with positive ER2 values (i.e. $ST > UT$) represent rare transcripts, whereas clones above the diagonal line with negative ER2 (i.e. $ST < UT$) values are regarded as abundant, hence these clones have been reduced in relative concentration during normalization.

1.3.2. SSHscreen as a package in R

In 2004 Dr. Wiesner Vos (while at the Department of Statistics, Oxford University) in collaboration with Prof. Dave Berger (Department Plant Science, FABI, UP) developed SSHscreen as a software package in the *R* computing environment using limma functions from the BioConductor project to screen SSH libraries. SSHscreen also calculates enrichment ratios ER1, ER2 and ER3 for each clone in the library. ER3 is the comparison $UT : UD$ and $\log_2(UT/UD)$ directly describes the regulation after treatment (since this is a comparison of the two original un-subtracted samples).

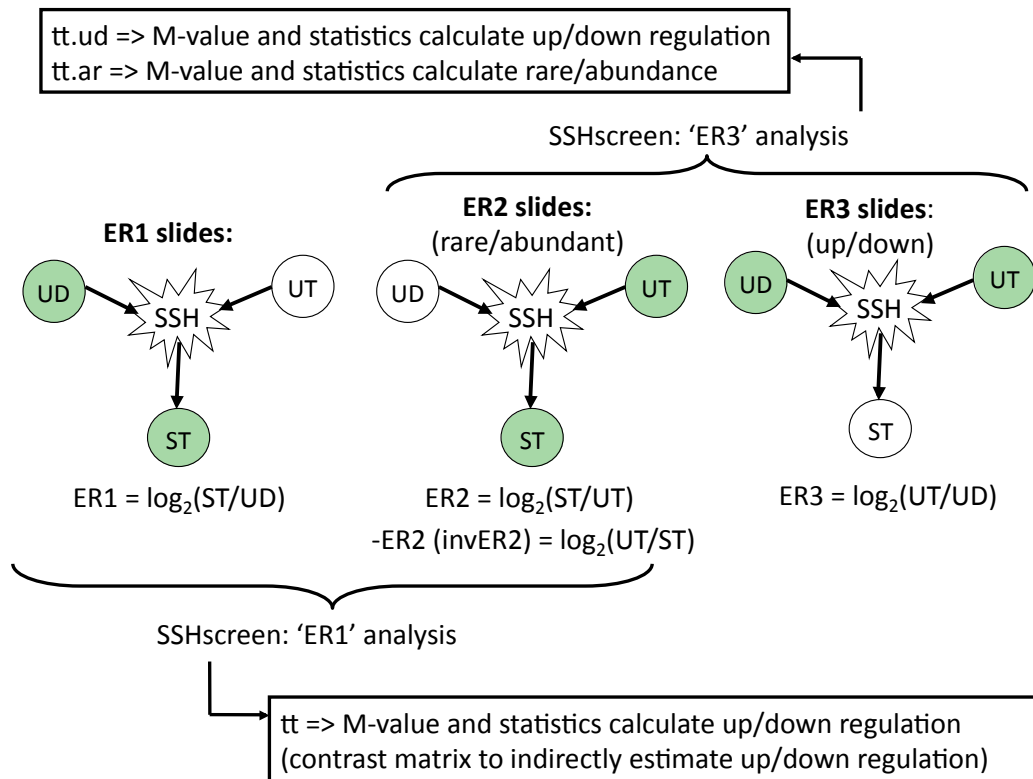


Figure 1.17: SSHscreen 'ER1' analysis and 'ER3' analysis. In SSHscreen, either an 'ER1' analysis or an 'ER3' analysis can be performed (by specifying `method = "ER1"` or `method = "ER3"` in the SSHscreen function). For the 'ER3' analysis, ER2 and ER3 slides must be available and for the 'ER1' analysis, ER2 and ER1 slides must be available. The 'ER3' analysis exports two top tables, `tt.ud` and `tt.ar`. The 'ER1' analysis exports only one top table.

In SSHscreen, either an 'ER1' analysis or an 'ER3' analysis can be performed (by specifying `method = 'ER1'` or `method = 'ER3'` in the SSHscreen function). Figure 1.17 shows the difference between the 'ER1' and 'ER3' analyses.

For the 'ER3' analysis, ER2 and ER3 slides must be available. After an 'ER3' analysis, top tables `tt.ud` and `tt.ar` are produced. Top Table `tt.ud` calculates whether clones are up/down-regulated and `tt.ar` whether clones are rare/abundant in the treated sample. In both cases clones are ranked in terms of the statistical significance, as calculated using linear models and empirical Bayes models, across all replicates. To both top tables, the calculated ER3 and invER2 values are also added. SSHscreen produces an ER3 versus inverse ER2 plot allowing one to visually screen all the clones in the SSH cDNA library (see Figure 3.6 on page 102).

For the 'ER1' analysis, ER2 and ER1 slides must be available. After an 'ER1' analysis, only one top table is exported. This top table indirectly calculates whether clones are up/down-regulated, using a contrast matrix (for details on the contrast matrix, see the section on page 29). Again, clones are ranked in terms of the statistical significance and the calculated ER1 and ER2 values for each clone are added to the resulting top table. SSHscreen produces an ER1 versus ER2 plot to visually screen the all the clones in the SSH cDNA library (see Figure 3.7 on page 105).

1.4. Databases for management of cDNA sequences

1.4.1. Similarity searches against sequence databases using BLAST

After sequencing differentially expressed gene fragments from a cDNA library, the putative identity of these genes can be inferred by BLAST searches against sequence databases. The SSH technique is especially useful for non-sequenced organisms, for example pearl millet. Since the pearl millet genome has not been sequenced and very few gene sequences for pearl millet are available, putative identities are obtained from other plant sequences. A useful organism in this case is cereal rice, since pearl millet is also a member of the grass family, and therefore these plants show a close relationship in evolutionary terms. An important concept is synteny, which means that the position and order of chromosome segments are highly conserved. Therefore, the identity of a gene sequence from any cereal can be matched to a rice gene. In addition, if a matching rice gene has not been well annotated, a match can sometimes be found with a gene from the model plant *Arabidopsis*, since this genome is better annotated due to the greater amount of biochemical studies that have been carried out in *Arabidopsis*.

1.4.2. Tools currently available for managing in-house sequencing projects

To identify and characterize the anonymous differentially expressed sequences in SSH/cDNA libraries, integrated bioinformatics tools for sequence management and annotation are needed. Various automated in-house pipelines were prepared previously to process and annotate EST/cDNA sequences. These pipelines often exploit public software and collect data in customized Structured Query Language (SQL) databases according to specific needs (Lazzari *et al.*, 2005; Paschall *et al.*, 2004; Smith *et al.*, 2008). CAS (cDNA Annotation System) provides a new standard for large-scale annotation, in which the initial automatic annotations are manually investigated, whereafter computational methods are iteratively modified

and improved based on results of manual curation (Kasukawa *et al.*, 2003). SSHSuite is an example of integrated package handling and storing large-scale SSH data (Weckx *et al.*, 2004).

1.5. Conclusion

Critical issues in transcriptome profiling includes the isolation and analysis of RNA (which usually includes the construction of a cDNA library), the quantitative global gene expression profiling of known/novel transcripts using high-throughput technology platforms (for example cDNA microarrays), and the validation of selected transcripts using quantitative real-time reverse transcriptase PCR (RT-PCR) or northern blots.

SSH is a method for the construction of enriched cDNA gene libraries, particularly useful for non-sequenced organisms. In a single procedure, SSH combines normalization and subtraction. cDNA microarrays provide a rapid and high throughput method to screen resulting SSH cDNA libraries, but without sufficient statistical knowledge molecular biologists struggle to analyze the resulting microarray data. Therefore the need exists for easy-to-use software packages that can be employed for this purpose. SSHscreen was developed as an R package using limma functions from the BioConductor project to analyze the resulting microarray data to firstly identify clones in the cDNA libraries that were significantly differentially expressed, and secondly determine if they were rare or abundant in the original treated sample.

This approach makes it possible to select a subset of clones from the SSH/cDNA library for further investigation, such as detailed expression profiling using a custom cDNA microarray, DNA sequencing, northern blotting or real-time RT-PCR. Sequencing these clones is of great value since they can then be annotated with putative functions, using sequence similarity searches such as BLAST.

In order to efficiently and effectively manage SSHscreen enrichment ratio information as well as sequence data from a SSH library, an automated integrated bioinformatics approach is needed. An approach like this should be able to store SSHscreen/limma toptables, handle sequences, and also perform alignment searches as well as store, manage and retrieve the resulting alignment information. Due to the nature of the SSH technique, there may be redundancy in the library and the automatic grouping of clones with the same sequences may be valuable.

Chapter 2

SSHscreen and SSHdb, generic software for microarray-based gene discovery: application to the stress response in cowpea

Nanette Coetzer^{1*}, Inge Gazendam^{2,3*}, Dean Oelofse², D.K. Berger^{3§}

¹ ACGT Computational Biology and Bioinformatics Unit, University of Pretoria, 0002, South Africa

² Germplasm Development Division, Agricultural Research Council-Vegetable and Ornamental Plant Institute, Private Bag X293, Pretoria, 0001, South Africa

³ Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, 0002, South Africa

* These authors contributed equally to this work

§ Corresponding author

Email addresses:

NC: nanette.coetzer@gmail.com

IG: IGazendam@arc.agric.za

DO: DOelofse@arc.agric.za

DKB: Dave.Berger@fabi.up.ac.za

2.1. Note

The content of this chapter has been submitted as a manuscript to the “Methodology” section of the *Plant Methods* journal. To be consistent with the dissertation layout, the figures are imbedded in the text and the references are included in the Bibliography section at the end of the dissertation. Since the laboratory work was done by Inge Gazendam and the data analysis was carried out by myself, we are joint first authors on the manuscript.

2.2. Authors’ contributions

NC developed the SSHscreen and SSHdb software, analyzed the microarray and sequence data and drafted the manuscript. IG constructed the SSH library, performed the microarray hybridizations and qPCR experiments, participated in the microarray and sequence data analysis and drafted the manuscript. DO contributed to conceiving the study, the design of the biological experiments and assisted in writing the manuscript. DKB was instrumental in the design of the study, development of the software, interpretation of data, and drafted the manuscript. All authors have read and approved the final manuscript.

2.3. Abstract

Background

Suppression subtractive hybridization is a popular technique for gene discovery from non-model organisms without an annotated genome sequence, such as cowpea (*Vigna unguiculata* (L.) Walp). We aimed to use this method to enrich for genes expressed during drought stress in a drought tolerant cowpea line. However, current methods were inefficient in screening libraries and management of the sequence data, and thus there was a need to develop software tools to facilitate the process.

Results

Forward and reverse cDNA libraries enriched for cowpea drought response genes were screened on microarrays, and the R software package SSHscreen 2.0.0 was developed (i) to normalize the data effectively using spike-in control spot normalization, and (ii) to select clones for sequencing based on the calculation of enrichment ratios with associated statistics. Enrichment ratio 3 values for each clone showed that 62% of the forward library and 34% of the reverse library clones were significantly differentially expressed by drought stress (adjusted p-value < 0.05). Enrichment ratio 2 calculations showed that > 88% of the clones in both

libraries were derived from rare transcripts in the original tester samples, thus supporting the notion that suppression subtractive hybridization enriches for rare transcripts. A set of 118 clones were chosen for sequencing, and drought-induced cowpea genes were identified, the most interesting encoding a late embryogenesis abundant *Lea5* protein, a glutathione S-transferase, a thaumatin, a universal stress protein, and a wound induced protein. A lipid transfer protein and several components of photosynthesis were down-regulated by the drought stress. Reverse transcriptase quantitative PCR confirmed the enrichment ratio values for the selected cowpea genes. SSHdb, a web-accessible database, was developed to manage the clone sequences and combine the SSHscreen data with sequence annotations derived from BLAST against the GenBank database. The self-BLAST function within SSHdb grouped redundant clones together and illustrated that the SSHscreen plots are a useful tool for choosing anonymous clones for sequencing, since redundant clones cluster together on the enrichment ratio plots.

Conclusions

We developed the SSHscreen-SSHdb software pipeline, which greatly facilitates gene discovery using suppression subtractive hybridization by improving the selection of clones for sequencing after screening the library on a small number of microarrays. Annotation of the sequence information and collaboration is further enhanced through a web-based SSHdb database, and we illustrated this through identification of drought responsive genes from cowpea, which can now be investigated in gene function studies. SSH is a popular and powerful gene discovery tool, and therefore this pipeline will have application for gene discovery in any biological system, particularly non-model organisms. SSHscreen 2.0.0 is available from <http://sshdb.bi.up.ac.za/> and SSHdb can be accessed at <http://sshdb.bi.up.ac.za/>.

2.4. Background

A range of techniques are available for gene discovery. Expressed sequence tag (EST) sequencing of cloned cDNAs is a common approach with the advantage that if full-length cDNAs are cloned they can be directly employed for further gene function experiments (Ralph *et al.*, 2008). Next generation sequencing, such as 454 technologyTM, has been employed for sequencing cDNA libraries (Cheung *et al.*, 2006), and the term RNA-Seq has been dubbed for this approach when applied at deep enough coverage to compare transcript counts between one or more biological states (Wang *et al.*, 2009). Previous methods, such as serial analysis of gene expression (SAGE), are also based on counting short sequence tags (Velculescu *et al.*, 1995). Although these methods provided exceptional quantitative

analysis, they are labour-intensive and currently very costly. Additionally, they are most effective if an annotated genome sequence is available.

Many research laboratories that are investigating non-model crops without genome sequence resources or have research questions that do not require a full genome analysis have the option of applying different “RNA fingerprinting” techniques for gene discovery. Examples of these techniques are differential display RT-PCR (DD-RT-PCR), RNA-fingerprinting by arbitrarily primed PCR (RAP-PCR) and cDNA amplified fragment length polymorphism (cDNA-AFLP) where cDNA sub populations are amplified and visualized on polyacrylamide gels, whereafter differentially expressed transcripts are isolated from the gel for sequencing (Liang and Pardee, 1992; Welsh *et al.*, 1992; Bachem *et al.*, 1998). These methods have limitations such as bias based on choice of initial primer sets, problems with reproducibility, generation of false positives, and reliance on time-consuming polyacrylamide gel electrophoresis and gel extraction to obtain sequence information. Another limitation of the above methods is the difficulty to capture low abundance clones.

A third alternative for gene discovery are PCR-based cDNA subtractive hybridization methods. These methods exclude common cDNA sequences between the two or more samples and, thus enrich for target sequences of interest, which are subsequently cloned. These methods include representational difference analysis (RDA) and suppression subtractive hybridization (SSH) (O’Neill and Sinclair, 1997; Diatchenko *et al.*, 1996; Morissette *et al.*, 2008). SSH has the advantage of enriching for rare transcripts. A recent search with the keywords ‘suppression subtractive hybridization’ in the title of research articles in PubMed produced 1213 hits (data not shown), indicating that SSH remains a popular method for the construction of enriched cDNA libraries. We chose to apply SSH to gene discovery in the non-model crop cowpea, and in this work we describe two software innovations that facilitate gene discovery using SSH.

Subsequent to gene cloning methods such as SSH, integrated bioinformatics tools for sequence management and annotation are needed. Various automated in-house pipelines have been developed to process and annotate EST/cDNA sequences exploiting public software, and collecting data in customized SQL databases according to specific needs (Paschall *et al.*, 2004; Lazzari *et al.*, 2005; Smith *et al.*, 2008). The cDNA Annotation System (CAS) is a useful tool for large-scale annotation, which can be implemented on a single desktop. Automatic annotations of sequences can subsequently be manually investigated and curated (Kasukawa *et al.*, 2003). SSHSuite is an example of a workstation package capable of handling and storing cDNA sequences from a SSH library (Weckx *et al.*, 2004).

In this study we chose to apply SSH to gene discovery from cowpea (*Vigna unguiculata* (L.) Walp.) plants. Cowpea is a tropical legume crop with a high protein content, since it

is able to fix nitrogen, and is used as a protein substitute for meat products (Singh *et al.*, 2003). The crop is fully utilised by people in Africa as leaves and seeds are consumed, and the plants are used for grazing and the feeding of livestock. Since many lines are drought tolerant, cowpea can be grown under the harshest growing conditions, and in the poorest soils, and is, therefore, an important crop for subsistence and small-holder farmers (Quass, 1995). Breeding efforts to improve yield of cowpea under different production systems is ongoing (Singh *et al.*, 2003), and lines with differential drought tolerance have been identified (Dingkuhn *et al.*, 2006; Agbicodo *et al.*, 2009; Muchero *et al.*, 2009). Promising QTLs for drought tolerance in cowpea have recently been reported (Muchero *et al.*, 2009).

Cowpea can be classified as an orphan crop, which means that it is important for food security in many developing countries, however limited research funding has been devoted to it (Varshney *et al.*, 2009). Genomics resources for cowpea are starting to be developed with sequencing of a methyl-filtered genomic library (Timko *et al.*, 2008), as well as an EST dataset (Varshney *et al.*, 2009). The availability of a cowpea breeding line that exhibited drought tolerance in the field prompted us to investigate gene expression in this line in response to drought stress. Based on previous experience of using SSH for gene discovery in other orphan crops, banana and pearl millet (van den Berg *et al.*, 2007; Crampton *et al.*, 2009), we encountered bottlenecks in the process. Consequently, in this study we developed improvements to the gene discovery pipeline, through the software SSHscreen 2.0.0, an R package, which quantitatively describes each clone in the library in terms of up/down regulation and rarity/abundance in the treated sample. We then validated the enrichment ratio calculations from the microarray screening and SSHscreen 2.0.0 analysis for selected drought-responsive cowpea clones using quantitative PCR (qPCR). SSHscreen facilitated the efficient choice of clones to be sequenced, which then led to the development of a web-based sequence database SSHdb, which facilitated the management and annotation of the SSH cDNA library clones. We, therefore, report development of the SSHscreen-SSHdb pipeline, a useful resource for any research group embarking on gene discovery using SSH.

2.5. Methods

2.5.1. Plant materials and treatments

Cowpea (*V. unguiculata* L. Walp) breeding lines IT96D-602 and Tvu7778 were provided by the Dr BB Singh of the International Institute of Tropical Agriculture (IITA) (Dingkuhn *et al.*, 2006). Seeds were germinated and plants were grown in a glasshouse under 11h day length, 28°C and 18°C day and night temperatures, respectively, and watering three times

weekly. At six weeks, five replicate plants of each variety were divided into two groups, one that was subjected to drought stress by withholding watering and the other that was kept to the control watering scheme.

2.5.2. RNA extraction

RNA was isolated from cowpea leaves using Tri-reagent (Sigma) and Polyvinyl pyrrolidone (PVP) (Ambion's Plant RNA isolation aid). Contaminating genomic DNA was removed with the Turbo DNA-free kit (Ambion) and the RNA cleaned up with the Plant RNeasy kit (Qiagen, Hilden, Germany).

2.5.3. Construction of cDNA library using SSH

Differential expression analysis by means of SSH (Diatchenko *et al.*, 1996) was employed to prepare a cDNA drought expression library for cowpea. Messenger RNA (mRNA) was isolated from 50 μ g pools of stressed IT96D-602 RNA (tester) (9 and 12 days without water) and control Tvu7778 RNA (driver) (9 and 12 days) using an Oligotex mRNA purification kit (Qiagen). cDNA was synthesised from mRNA using the cDNA synthesis system (Roche Diagnostics, Basel, Switzerland). Prior to subtraction, unsubtracted tester (UT) and unsubtracted driver (UD) cDNA samples were prepared as described (Timko *et al.*, 2008). Subtractive hybridisation was performed on *RsaI* (Roche Diagnostics) -digested tester and driver cDNA fragments using the PCR-Select cDNA subtraction kit (BD Biosciences Clontech, Palo Alto, CA), as previously described (van den Berg *et al.*, 2004). Both forward and reverse subtractions were performed. After subtraction the products were amplified by a primary PCR and a nested secondary suppression PCR to generate differentially expressed cDNA fragments (termed ST_F and ST_R for the forward and reverse libraries, respectively). Replicate PCR reactions were pooled, size fractionated and cloned into the pGEM-T Easy cloning vector and transformed into *Escherichia coli* JM109 following the manufacturers' instructions (Promega, Madison, WI). Transformed colonies were selected by blue-white selection on 100 μ g/ml ampicillin LB-agar selection media (spread with X-Gal and IPTG) and stored as 25% glycerol stocks at -70°C in sterile 96-well culture plates (Corning, NY). In addition, unsubtracted PCR products from the tester cDNA (drought stressed IT96D-602) (termed UT) and driver cDNA (control Tvu7778) (termed UD) were also prepared to be used for SSHscreen analysis as described in (Berger *et al.*, 2007).

2.5.4. Fabrication of SSH library on glass slide array

Inserts of the cowpea drought expression cDNA library were amplified with PCR directly from overnight bacterial cultures in 96-well format (Thermo-Fast, ABGene, Epsom, UK) in 100 μ l reactions with 1U Biotaq DNA polymerase (Bioline) and the SP6 and T7 primers (Table 2.1 on the next page). The PCR plate was sealed with a silicon mat (Corning). Reactions were incubated in a PTC-100 Thermocycler (MJ Research) at 94°C for 5 min; 30 cycles of (94°C for 30s, 50°C for 30s and 72°C for 1 min); and 72°C for 5 min.

The PCR products were purified with Montage PCR purification plates on a vacuum manifold (Microsep) and resuspended in 50 μ l SDW. The suspensions were transferred to 96-well storage plates, covered with well caps (Nunc, Roskilde, Denmark) and stored at -20°C. The purified PCR products were dried down in a vacuum centrifuge at 45°C, resuspended in 50% dimethyl sulfoxide (DMSO), transferred to 384-well spotting plates and stored at -70°C until microarray spotting.

The control genes *gfp* (717bp fragment in pGEM-T Easy, positions 1603-2319 of GenBank accession number AF078810), *globin* (human *beta-globin*; 474bp fragment in pBluescriptSK, positions 50-523 of NM_000518) and *nptII* (812bp fragment in pGEM-T Easy, positions 142-953 of V00618) were purchased from the Nottingham Arabidopsis stock centre [NASC, <http://arabidopsis.org.uk>]. They were transformed into *E. coli* JM109 (Promega). An *its* clone in pGEM-T Easy (193bp fragment from the internal transcribed spacer 2 of the rRNA genes from *Leptographium elegans*) was also used as a control gene. It matches to positions 268-458 of AF343675.1. Plasmids were isolated from cultures using the Qiaspin miniprep plasmid isolation kit (Qiagen). PCR products of the four control genes were prepared using the T7 and SP6 primers (PCR product sizes: *gfp* (893bp), *nptII* (988bp), *its* (369bp)) or the T7 and M13R primers for *globin* (677bp). Montage purified PCR products of twelve 100 μ l PCR reactions each were pooled, concentrated and transferred to 12 wells each of a 384-well spotting plate. An equal volume of DMSO was added so that the final concentrations in 50% DMSO ranged from 70 – 100ng/ μ l. Five two-fold serial dilutions were also prepared for each PCR fragment (*gfp*: 180 – 11.25ng/ μ l; *globin*: 100 – 6.25ng/ μ l; *nptII*: 150 – 9.375ng/ μ l; and *its*: 130 – 8.125ng/ μ l), transferred to an additional 10 wells per fragment, an equal volume of DMSO added, and spotted on the glass slides.

Glass slides were spotted with the cowpea drought expression library (4160 clones in total from the forward and reverse libraries) and controls using the Array Spotter Generation III (Molecular Dynamics, Sunnyvale, CA) at the University of Pretoria, Pretoria, South Africa [<http://microarray.up.ac.za>]. Each sample spot was duplicated on the slide. The slides were allowed to dry overnight in the protective atmosphere of the spotter, after which the

DNA was cross-linked under ultraviolet (UV) light for 3 min. The slides were stored in a desiccator covered in foil at room temperature.

Table 2.1: Table of oligonucleotide primers used in this study.

Primer code	Forward primer (5' – 3')	Reverse primer (5' – 3')	Source (library clone number or GenBank accession number)	Expected product length (bp)
SP6	ATTTAGGTGACACTATAG			
T7	TAATACGACTCACTATAGGG			
M13R	CAGGAAACAGCTATGACC			
GST	GCTGGTGAAGGTGTTGGATA	CCACGATGGTCTGCTACTTA	25B06-F	199
THAU	AAGGTTCAAGTTGCGCCACAG	AATCCGTCCACGTTGCTCAC	33E07-F	147
LEA	CCGTCTCCTTCTCCTCAGT	TGCACCATCTCTTGTCACAG	07F09-F	163
26S	GGAATCGAGAGCTCCAAGTG	GTTGATTCGGCAGGTGAGTT	38G04-R	198
CHL	CTCATCCACGCTCAGAGCAT	CTGGACGAAGAAGCCGAACA	44C07-R	240
LTP	GCATCAGCGGTATCAACCTC	CCTCCTTGCCATCTCTTCTC	36F07-R	147
			Consensus from: AC135505_Mt (exons only), DQ192668, DQ355800,	
GAPC	ATCAGCCAAGGACTGGAGAG	ACGGAATGCCATACCAGTCA	PEAGAPCI	130
Globin	GGAGAAGTCTGCCGTTACTG	GCCATGAGCCTTCACCTTAG	NM_000518	175

2.5.5. Screening SSH library on microarrays

SSH cDNA fragments (ST_F , ST_R , UT and UD), purified by PCR Minelute cleanup kit (Qiagen), were digested with *RsaI* (10U per microgram DNA) in the appropriate buffer overnight at 37°C. The fragments were separated from the adaptor fragments by electrophoresis on a 1.5% low melting point agarose gel (Seaplaque, FMC Bioproducts) in 0.5x TAE and purified from the gel using the Qiaquick gel extraction kit (Qiagen).

The control fragments were excised from their plasmids using restriction digestion to exclude any T7 and SP6 primer binding sites (*KpnI/XbaI* for *globin* (product of 548bp); *NcoI/PstI* for *gfp* (768bp); *EcoRI* for *nptII* (830bp) and *its* (211bp)). Restriction fragments were purified with the Qiaquick gel extraction kit (Qiagen). Each target sample of SSH cDNA fragments (200ng) were spiked with equal amounts of a control fragment pool made up of different quantities of four control fragments (45ng *globin*, 45ng *its*, 4.5ng *nptII* and 0.45ng *gfp*) for within-slide normalization. Spiking with equal amounts of fifteen- or three-fold

dilutions of the control fragment pool were tested and also gave sufficient hybridization for within-slide normalization (data not shown).

Targets were labelled by direct Cy-dUTP incorporation by Klenow enzyme (Fermentas, Vilnius, Lithuania). Each SSH fragment sample was labelled with both dyes (Cy3 and Cy5) for a dye-swap experiment of each slide. The protocols and data analysis techniques described in (Berger *et al.*, 2007) were followed, with some modifications. DNA to be labelled, in a volume of $12\mu\text{l}$, was denatured at 95°C for 5 min and placed on ice. The following were added to the pairs of denatured DNA samples to yield a total reaction volume of $20\mu\text{l}$: $2\mu\text{l}$ of 10x Klenow buffer (Fermentas); $2\mu\text{l}$ 10x Hexanucleotide mix (Roche Diagnostics); $2\mu\text{l}$ Klenow enzyme ($5\text{U}/\mu\text{l}$; Fermentas); $2\mu\text{l}$ of a dNTP mix containing 1nmol each of dATP, dCTP and dGTP, 0.74nmol dTTP and 0.27nmol of either Cy3-dUTP or Cy5-dUTP (Amersham Biosciences). The labelling reaction was incubated overnight (17-20h) at 37°C . The labelled DNA was cleaned up from unincorporated dye using the Qiaquick PCR purification kit (Qiagen). Dye incorporation was measured using a NanoDrop ND-1000 UV-Vis Spectrophotometer (Nanodrop Technologies, Wilmington, DE).

Labelled SSH targets were combined in pairs using equal amounts of Cy3 and Cy5 dye incorporation for each target in each pair required for SSHscreen analysis. Each labelled target DNA mix was dried down in a vacuum centrifuge at 45°C and resuspended in $50\mu\text{l}$ hybridisation solution (50% formamide, 25% 4X Microarray hybridisation buffer (Amersham Biosciences), 25% SDW). Labelled targets in hybridisation solution were denatured at 95°C for 2 min and placed on ice.

Glass slides arrayed with the SSH cDNA libraries were pre-treated in 1% bovine serum albumin (BSA; Roche Diagnostics), 3.5X SSC (525mM sodium chloride and 52.5mM sodium citrate) and 0.2% sodium dodecyl sulphate (SDS) at 60°C for 20 min. After rinsing in SDW at room temperature, the slide was dried by centrifugation in a 50ml tube at $1000\times g$ for 4 min at room temperature in a swing-out rotor (Eppendorf 5810R centrifuge). The slide was placed in a locally manufactured hybridisation chamber (HybUP, NB Engineering, Pretoria, South Africa) with $20\mu\text{l}$ SDW in the reservoirs on either side. Labelled and denatured target was applied to the slide and gently overlaid with a cover slip. The chamber was sealed and incubated in a water bath at 42°C for 16h. Slides were washed for 4 min at 42°C with $1\times$ SSC (150mM NaCl, 15mM sodium citrate)/ 0.2% SDS, twice with $0.1\times$ SSC (15mM NaCl, 1.5mM sodium citrate)/ 0.2% SDS and three washes of $0.1\times$ SSC for 1 min at room temperature. After dipping the slide in SDW at room temperature and centrifuged to dry, it was immediately scanned using a GenePix 4000B scanner (Axon Instruments, Foster City, CA).

GenePix Pro 5.1 software (Axon Instruments) was used to automatically locate all the

spot positions from the scanner-generated TIFF images and associate them with each specific clone in a GenePix Array List (GAL file)(available at <http://sshdb.bi.up.ac.za>). The GAL file links the information from the arraying process to the analysis, since it provides identification information for each spot printed on the slide. Bad quality spots (irregularly shaped or with hybridisation artefacts; signal/noise ratio < 3) were flagged for exclusion during data analysis and the array of circles were manually adjusted for a better fit. GenePix Pro 5.1 was used to extract the dye intensity data of each spot and save the data for each slide in a GenePix Results file (.gpr).

2.5.6. SSHscreen software analysis of microarray data

The SSHscreen 2.0.0 package, written as a single function in the R programming language, was used for analyzing the resulting microarray data to calculate ER3 values ($\log_2(UT/UD)$) for the forward library clones; and $\log_2(UD/UT)$ for the reverse library clones). ER3 values quantify the amount of up-regulation of the clones in each library (Berger *et al.*, 2007). Inverse ER2 values ($\log_2(UT/ST_F)$) for the forward library clones; and $\log_2(UD/ST_R)$ for the reverse library clones) reflect the relative abundance of transcripts for each gene in the unsubtracted samples (Berger *et al.*, 2007). The original version of SSHscreen is described in (Berger *et al.*, 2007). Improvements to the functionality were added to the original R code, the documentation was updated and the latest version was packaged as SSHscreen 2.0.0. SSHscreen can be downloaded at <http://microarray.up.ac.za/SSHscreen/>, together with a demo data set and an example R script. R version 2.8.1 and limma version 2.16.5 were used (www.bioconductor.org). For details on how to use SSHscreen and to view a full description of all the possible argument options, type `help(SSHscreen)` at the R command line (after loading the SSHscreen library). The data from the hybridization experiments to the cowpea SSH library arrays were submitted to SSHscreen 2.0.0 with the Targets file, Spot types file (Tables 2.2 and 2.3 on page 68, respectively), the .gpr files and the GAL file, and used for analysis using the limma functions executed in R for background subtraction, within slide normalization, between-slide normalization, average dye intensities from replicate spots and combining dye swap data to calculate the ER3 and inverse ER2 values with associated statistics (see Figure 2.3 for the R script). The outputs of SSHscreen: Top tables (Table 2.4), MA-plots (Figure 2.5) and a graphical representation of each clone on ER-plots (Figures 2.6 and 2.7) were used to select clones for sequencing.

2.5.7. Sequencing

Selected cowpea drought expression library clones were sequenced using the T7 Promoter primer by Inqaba Biotec (SA) or Macrogen (USA). Colonies were sent on LB-agar plates containing $100\mu\text{g}/\text{ml}$ ampicillin.

2.5.8. Annotation and management of sequences using SSHdb

SSHdb (available at <http://sshdb.bi.up.ac.za>) was developed as a web-based tool for sequence management of clones in SSH libraries. The SSHdb interface was written using Turbogears (Ramm *et al.*, 2006), a Python web application framework. Currently, a central MySQL database is used to store sequence, top table and annotation information. SQLAlchemy (Copeland, 2008), an object relational mapper for Python and toolkit for SQL, is used to query the database.

For each input sequence in FASTA format, SSHdb removed the vector and adaptor fragments after BLASTN (Altschul *et al.*, 1990) searches were performed against the NCBI UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>). Further BLASTN searches were carried out against all sequences already uploaded in the database, so that redundant partners in the library (using a BLASTN e-value cut-off value of $10\text{e-}10$) could be identified. For each redundant partner group, the longest sequence in the group was selected by default as the representative clone. Multiple sequence alignments, generated by ClustalW (Thompson *et al.*, 1994), for individual redundant partner groups could be viewed and downloaded from SSHdb. For each representative clone, SSHdb performed nucleotide-nucleotide and translated sequence comparisons using BLASTN and BLASTX searches against a local installation of the NCBI non-redundant nucleotide and peptide databases (nt/nr) (Altschul *et al.*, 1990). For cases where the e-value of the top BLASTX hit was small enough (smaller than $10\text{e-}10$), this hit was automatically selected as the default priority annotation. The top 10 BLASTX and BLASTN hits were stored in the database. SSHdb provides two major views of the data, the SSH database view, which shows the annotated representative clones (see Figure 2.1), or the SSH toptable view, which shows the enrichment ratio data for each clone in each library (see Table 2.4 on page 72). SSHdb can be updated as additional clones are sequenced.

2.5.9. Quantitative PCR

qPCR primer pairs (20-mers) were designed from selected cowpea sequences to amplify products between 120 and 250bp in length from the SSH cDNA fragment pools (UT, UD, ST_F , and ST_R) (Table 2.1). qPCR reactions containing 1x Sensimix (Quantace, Celtic

Molecular Diagnostics South Africa), SYBR Green, 2.5mM MgCl₂, the appropriate primer pair (200nM each) and cDNA template in a total volume of 25 μ l were set up and run on a Rotor-Gene (Corbett Research). The enzyme was activated by a hold at 95°C for 10 min followed by 45 cycles of 95°C for 15 s, annealing at 56°C for 30 s and extension at 72°C for 6 s. SYBR green fluorescence was measured after the extension step of every cycle. qPCR was performed on serial ten-fold dilutions of a mix of UT and UD cDNAs (templates ranging from 0.5pg to 50ng) to construct standard curves for each primer pair. The quantification cycle (Cq) values from the qPCR fluorescent profiles were converted to input nanograms of template using the standard curves. Average nanogram quantities for each gene was normalized relative to the data for the respective sample's reference gene content.

For ER3 verification, qPCR was performed in duplicate on 50ng each of cDNA from the two cowpea cultivars before subtraction (UT and UD). Glyceraldehyde-3-phosphate dehydrogenase C-subunit (*gapC*) was used as a reference gene. A consensus sequence between the *gapC* genes of *Medicago truncatula* [GenBank:AC135505_Mt, exons only], *G. max* [GenBank:DQ192668_Gmax1 and DQ355800_Gmax2] and *Pisum sativum* [GenBank:PEAGAPCI] was used to design the *gapC* reference gene primers (Table 2.1). An expression ratio of $\log_2(\text{ng in UT} / \text{ng in UD})$ was calculated for each gene.

For ER2 verification, normalization of qPCR results between unsubtracted and subtracted cDNA samples (i.e. UT and *ST_F*; UD and *ST_R*) required the spiking of cDNA samples with equal amounts of an alien gene. qPCR was performed in duplicate on 10ng of the UT, *ST_F*, UD and *ST_R* cDNA templates, each spiked with 50pg of the human *beta-globin* fragment (prepared using the Globin forward and reverse primers, Table 2.1). Average input nanogram quantities were calculated for each gene and normalized with the *globin* spike content. Expression ratios of $\log_2(\text{ng in UT} / \text{ng in } ST_F)$ and $\log_2(\text{ng in UD} / \text{ng in } ST_R)$ were calculated for each gene.

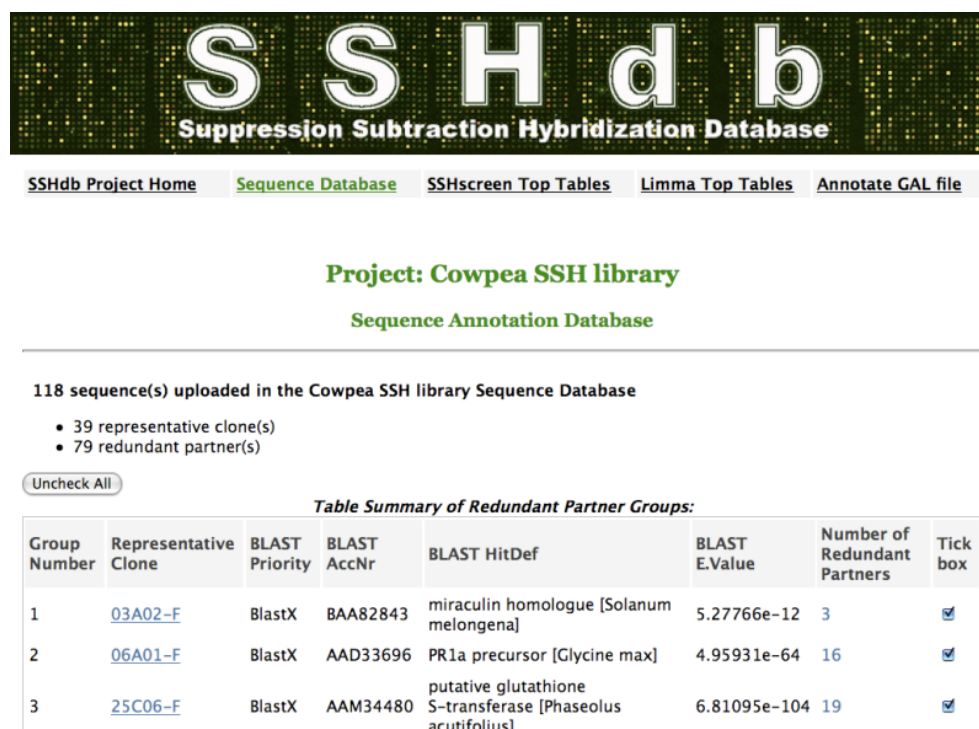
RT-qPCR reactions were performed in duplicate using 100ng of RNA isolated prior to SSH library construction from drought-treated IT96D-602 cultivar and control Tvu7778 at two time points, 9 and 12 days, separately. The Sensimix One-step RT-qPCR kit (Quantace) was used with a reverse transcription step at 49°C for 30 min inserted before the cycling profile described above. The quantification cycle (Cq) values from the RT-qPCR fluorescent profiles were converted to input nanograms of template using the standard curves.

2.6. Results

2.6.1. Construction of cowpea drought expression SSH library and overview of SSHscreen/SSHdb data analysis pipeline

We developed a pipeline for quantitative screening and sequence management of clones from a SSH cDNA library. The pipeline is particularly useful for gene discovery in non-sequenced organisms. As an example, we used a cowpea (*V. unguiculata*, (L.) Walp) drought expression library where the objective was to identify and isolate genes responding to drought stress in cowpea. Figure 2.2 on page 62 gives an outline of the pipeline. SSH (Diatchenko *et al.*, 1996) was used to enrich for genes that were differentially expressed between drought stressed and unstressed cowpea plants. Cowpea breeding lines from the International Institute of Tropical Agriculture (IITA) that were previously shown to be drought tolerant (line IT96D-602) and drought susceptible (line Tvu7778) were used (Spreeth *et al.*, 2004). The “tester” for forward library construction was from drought stressed line IT96D-602 (cDNA pooled from plants 9 and 12 days after water was withheld), whereas the “driver” for forward library construction was from control treated line Tvu7778 (cDNA pooled from plants grown for the same time on a normal watering regime). These time points were chosen for maximum drought stress symptoms, before leaves were too senesced for RNA extractions. The aim of this wide subtraction was to be sure to capture sufficient differentially expressed transcripts to illustrate the efficacy of the SSHscreen/SSHdb software. This could include not only genes that are induced/repressed by drought stress in drought tolerant IT96D-602 only, but also those that are constitutively expressed at higher/lower levels in IT96D-602 compared to the drought sensitive line Tvu7778. Good quality forward and reverse subtracted cDNA fragments were generated (data not shown) and used to construct a cDNA library with a total of 4160 cDNA clones (2144 in the forward and 2016 in the reverse library), which were amplified by PCR and spotted onto glass slides for screening and selection of clones for sequencing (Berger *et al.*, 2007)(Figure 2.2). Subtracted and unsubtracted cDNA samples from cowpea used to construct the SSH libraries were prepared as Cy3- and Cy5-labelled targets and hybridized to the microarrays. These cDNA samples were UT (unsubtracted tester), UD (unsubtracted driver), ST_F (forward library subtracted tester), and ST_R (reverse library subtracted tester)(Figure 2.2). The R package SSHscreen version 2.2.0, available from <http://microarray.up.ac.za/SSHscreen/>, was developed in this study to analyze the resulting microarray data using limma functions, thereby quantitatively screening the library for significantly differentially expressed clones (Berger *et al.*, 2007)(Figure 2.2). SSHscreen analysis of the microarray data was used to assist in the selection of 118 clones for sequencing, based on their statistics of differential expression (Figure 2.2). The SSHscreen data output

(top tables with the statistics of differential expression for each clone), as well as the selected sequences in FASTA format were uploaded to SSHdb. SSHdb was developed as a web-based database for sequence management and annotation of clones in SSH libraries and can be accessed at <http://sshdb.bi.up.ac.za/> (Figure 2.2). A screenshot of the SSHdb interface is given in Figure 2.1, showing data for some of the clones in the cowpea SSH library. BLAST analysis that was carried out when sequences were uploaded to SSHdb was used to combine clones with the same sequence into redundant partner groups, as well as identify putative annotations for each group, and this is illustrated in Figure 2.1. Six genes identified from the cowpea SSH library were selected and used to validate the microarray/SSHscreen results with an independent technique – qPCR (Figure 2.2).



SSHdb
Suppression Subtraction Hybridization Database

SSHdb Project Home Sequence Database SSHscreen Top Tables Limma Top Tables Annotate GAL file

Project: Cowpea SSH library
Sequence Annotation Database

118 sequence(s) uploaded in the Cowpea SSH library Sequence Database

- 39 representative clone(s)
- 79 redundant partner(s)

Table Summary of Redundant Partner Groups:

Group Number	Representative Clone	BLAST Priority	BLAST AccNr	BLAST HitDef	BLAST E.Value	Number of Redundant Partners	Tick box
1	03A02-F	BlastX	BAA82843	miraculin homologue [Solanum melongena]	5.27766e-12	3	<input checked="" type="checkbox"/>
2	06A01-F	BlastX	AAD33696	PR1a precursor [Glycine max]	4.95931e-64	16	<input checked="" type="checkbox"/>
3	25C06-F	BlastX	AAM34480	putative glutathione S-transferase [Phaseolus acutifolius]	6.81095e-104	19	<input checked="" type="checkbox"/>

Figure 2.1: The SSHdb screenshot shows a summary of some of the redundant partner groups of the sequenced clones in the cowpea SSH library (<http://sshdb.bi.up.ac.za/>). For each group, the representative clone ID, the priority BLAST annotation and the number of redundant partners in the group are given, as well as a tick box allowing individual groups to be marked so that corresponding sequence and/or annotation information can be exported. By clicking on the representative clone ID, the user can view and select the preferred annotation from the top 10 BLASTX or BLASTN hits, download the multiple sequence alignment of the clones in that group and change the representative clone if required.

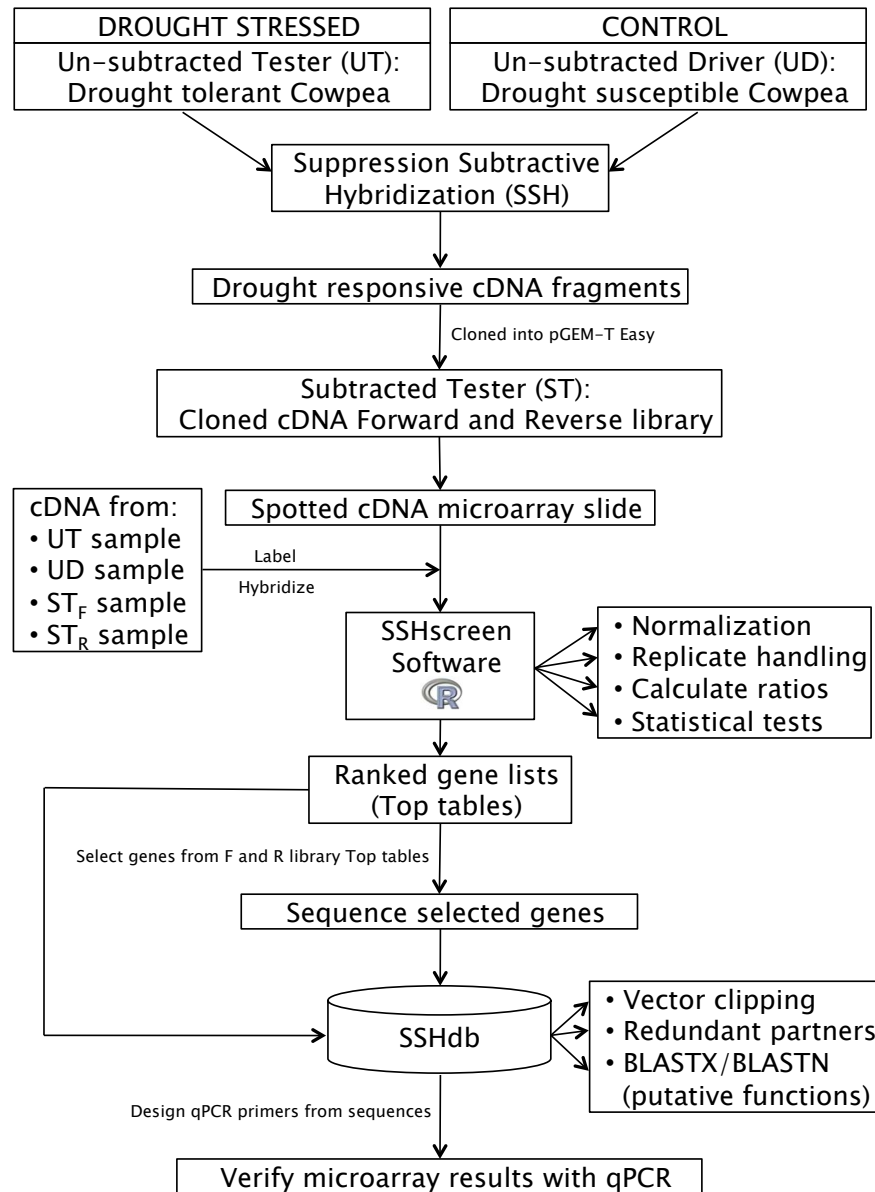


Figure 2.2: Schematic representation of the flow of data through the SSHscreen-SSHdb pipeline. SSH was used for the construction of a cowpea drought expression library. Tester cDNA were prepared from drought stressed IT96D-602 cowpea leaf RNA and driver cDNA from control Tvu7778 cowpea leaf RNA. The subtracted library was spotted onto glass slides and hybridised with a mix of differently labelled subtracted and unsubtractd cDNA from the two cowpea cultivars. The R package SSHscreen 2.0.0 was used to analyze the microarray data, using limma functions for pre-processing the data as well as to identify statistically significant differentially expressed genes. A subset of clones was selected for sequencing. Available FASTA sequences as well as top tables (output from SSHscreen) were uploaded to the web-based database, SSHdb, to manage and annotate clones in the library. Expression results for selected genes were verified using qPCR.

2.6.2. Screening the cowpea SSH libraries using SSHscreen 2.0.0

SSHscreen facilitates the screening of an SSH library using cDNA microarrays (Berger *et al.*, 2007). Each clone is quantitatively described in terms of up/down regulation (Enrichment ratio 3 (ER3) values; $\log_2(UT/UD)$) and rarity/abundance (Enrichment ratio 2 (inverse ER2) values; $\log_2(UT/ST)$) in the treated sample; and a measure of statistical significance for each result is provided in the form of a moderated t-statistic with an associated p-value (Smyth, 2005). SSHscreen is built around the limma R package from the BioConductor project (Smyth, 2005), which provides the functionality for importing and analyzing gene expression microarray data. (Berger *et al.*, 2007) described the implementation of the original version of SSHscreen (version 1.0.1). An improved version, SSHscreen 2.0.0 was developed to analyze the data from the cowpea SSH libraries in this study.

High quality microarray images were obtained from hybridization of pairs of Cy-labelled cDNA targets (UT, UD, ST_F or ST_R) to the cowpea drought expression microarrays (see pseudocolour image; Figure 2.4). For example, strong hybridization of Cy3 targets from ST_F to probes from the forward library spotted in the top six rows of each array block can be observed as green spots in Figure 2.4 on page 67a, whereas Cy5 targets from UT hybridize predominantly to probes from the reverse library as red spots (rows 7-11 of each array block) (Figure 2.4a), as expected. The opposite hybridization pattern is observed in a dye swap slide (Figure 2.4b), as expected.

Within-slide normalization of two-colour microarray data is an important consideration to account for systematic bias due to differences between the Cy3 and Cy5 dyes (Smyth and Speed, 2003). Commonly, loess normalization is applied (Smyth and Speed, 2003), however this is based on the assumption that most of the genes on the array are not differentially expressed. This is legitimate for most whole genome microarray experiments, however it is not appropriate when the array is constructed from an SSH library, which selects for differentially expressed genes. Therefore, spike-in control spot-based normalization was applied in SSHscreen analysis of the cowpea SSH libraries (Smyth and Speed, 2003). Serial dilutions of four “alien” control probes (green fluorescent protein (*gfp*), human beta-globin (*globin*), bacterial neomycin phosphotransferase II (*nptII*) and a fungal rRNA gene internal transcribed spacer (*its*); see methods) were spotted on the glass slides. These probes were chosen, since matching sequences are unlikely to be present in the cowpea cDNA samples. Importantly, a “spike-in” control mix of restriction enzyme fragments of the genes corresponding to the four control probes *gfp*, *globin*, *nptII* and *its* was prepared in which each of the four genes was

present at a different concentration. The spike-in control mix was added in equal amounts to each cDNA target sample prior to labelling.

SSHscreen 2.0.0 data analysis to calculate ER3 and inverse ER2 values was carried out using the R script and associated data files, provided in Figure 2.3 on page 66 and tables 2.2 and 2.3, by first weighting all spots that had been flagged as poor quality (signal/noise < 3) by the GenePix 3.1 image analysis programme (<http://www.axon.com>) so that these spots would not be used to calculate the normalization factors. Background correction used the normexp method in limma (Ritchie *et al.*, 2007) with an offset of 50 to damp the variation of the log-ratios for very low intensity spots towards zero. This approach is encouraged specifically when using empirical Bayes methods from the limma package (Smyth, 2004). The dilution series of control spots on each array which have hybridized to the spike-in controls (added in equal amounts to the pairs of target cDNAs) can be observed in the raw pseudocolour images as yellow spots in row 12 of most array blocks (Figure 2.4). Data from these control spots were used to apply the up-weighting print-tip loess within-array normalization method of limma in SSHscreen 2.0.0, which essentially applies full weight to all the control spots and zero weight to the spots of the probes from the SSH library. Thereafter, a loess curve was fitted through these control spots that span a range of intensities to normalize the data within each slide and thereby remove the systematic errors due to the dye effects (Smyth and Speed, 2003). Between-slide normalization was carried out using the Aquantile method, which is based on the assumption that the distribution of A values (Average expression; $(\log_2(UT * UD)/2)$) is similar across all arrays.

The quality of the data and success of the normalization can be seen by inspection of the MA-plots from SSHscreen analysis after background subtraction and normalization, as shown in Figure 2.5 on page 69. Successful within-slide normalization can be seen for each slide in Figure 2.5, since the control spots (colours other than blue or yellow) were placed on M=0 line in the MA-plots after normalization. Clones of the forward and reverse libraries are illustrated by blue and yellow dots in panels a-h and i-p, respectively. Dye swap slides show consistent clouds of data points above and below the M=0 line, as expected (compare panel a with b, for example). Excellent consistency of replicates can also be seen (compare each pair of panels on the left hand side (e.g. a/b) with their replicates on the right hand side (e.g. c/d; Figure 2.5).

The results of the SSHscreen 2.0.0 analysis were visualised by ER3 versus inverse ER2 plots for the forward and reverse libraries (Figure 2.6 on page 70 and Figure 2.7 on page 71, respectively). Most of the genes in these plots fall in quadrant I, where ER3 > 0 and inverse ER2 < 0, meaning up-regulated by drought stress and rare in the unsubtractd drought stressed cDNA for the forward library (92%; Figure 2.6), and down-regulated by

drought stress and rare in the control cDNA for the reverse library (52%; Figure 2.7). The criterion we chose to score genes as statistically significant differentially expressed (ER3 analysis: UT versus UD comparison) was that the adjusted p-value should be less than 0.05 after the linear model fit and empirical Bayes calculations. The p-value reflects the probability of rejecting the null hypothesis that there is no differential expression between the drought stressed (UT) and control (UD) samples for the forward library, and vice versa for the reverse library. We adjusted for multiple testing using Benjamini & Hochberg's method (Benjamini and Hochberg, 1995) for controlling the false discovery rate. There were 62% (1337/2146) significantly differentially expressed clones in the forward library and 34% (688/2018) in the reverse library using the stringent criterion of adjusted p-value < 0.05 . Only the most significant 300 for each library are marked in Figures 2.6 and 2.7. The quality of the subtraction process was reflected in the low number of clones that had negative ER3 values for the forward (8%; Figure 2.6) and reverse (48%; Figure 2.7) libraries.

SSHscreen also provides an alternative statistic to choose differentially expressed genes, namely the B-statistic. The B-statistic (Smyth, 2004; Lonnstedt and Speed, 2002) can be interpreted as the log-odds that a specific gene is differentially expressed. This means that a B-statistic of zero corresponds to a 50-50 chance of differential expression, and accordingly a user is generally interested in genes with a positive B-statistic. For the cowpea ER3 analysis, 67% of the clones in the forward and 52% of the clones in the reverse library had positive B-statistics, which includes more clones than the stringent criterion of adjusted pvalue < 0.05 . Importantly, the B-statistic calculation in SSHscreen/limma requires the user to make a prior guess of the number of differentially expressed genes in each library. Since a SSH library is enriched for differentially expressed genes, this value was set at 50% for this study, in contrast to the default of 1% in limma, which is designed for whole genome microarray data in which most genes are assumed not to be differentially expressed.

Table 2.6 on page 70 shows the top 20 cowpea clones sorted by p-value for the forward and reverse libraries extracted from the top tables that are generated from the ER3 analysis in SSHscreen. The most significant up regulated forward library clone, 46D03-F, has a log-2 fold change of 2.9 (equivalent to the ER3 value). Taking the antilog of the log with base 2, it can be shown that this clone is ~ 8 -fold up-regulated. With a similar calculation it can be shown that the most significant down-regulated reverse library clone, 45C07-R, with a log-2 fold change of 2.4, is ~ 5 -fold down-regulated. The top table also reports the Average expression (A value) and statistics associated with the ER3 value, namely a moderated t-statistic, a p-value, an adjusted p-value and a B-statistic (Table 2.4 on page 72). A top table for the ER2 analysis can also be generated by SSHscreen, which reports the statistics

of whether the clones represent rare or abundant transcripts in the original treated sample (data not shown).

```
# Cowpea
# 12 slides
# ER3 analysis with F and R library

library(SSHscreen)

cowpeaSSH <- SSHscreen(path = "/Users/Nanette/SSHscreen/
cowpea_F_and_R", source = "genepix", norm.plot = TRUE, mfrow =
c(4,2), legend = TRUE, bc.method = "normexp", wa.method =
"printtiploess", ba.method = "Aquantile", irregular = TRUE,
ndups = 2, spacing = 1, spot.ave = FALSE, method = "ER3",
toplist = "all", adjust = "fdr", negflags = 0, offset = 50,
weights = TRUE, library = "both", sort = "p", cutoff = 0.05,
proportion = 0.5)

write.table(cowpeaSSH$tt.ud.F, file="/Users/Nanette/SSHscreen/
cowpea_F_and_R/cowpea$tt.ud.F.txt", sep="\t")
write.table(cowpeaSSH$tt.ar.F, file="/Users/Nanette/SSHscreen/
cowpea_F_and_R/cowpea$tt.ar.F.txt", sep="\t")

write.table(cowpeaSSH$tt.ud.R, file="/Users/Nanette/SSHscreen/
cowpea_F_and_R/cowpea$tt.ud.R.txt", sep="\t")
write.table(cowpeaSSH$tt.ar.R, file="/Users/Nanette/SSHscreen/
cowpea_F_and_R/cowpea$tt.ar.R.txt", sep="\t")
```

Figure 2.3: R script used for ER3 analysis of both forward and reverse cowpea SSH libraries together after loading the limma 2.16.5 and SSHscreen 2.0.0 libraries in R 2.8.1. The working directory contained the Targets file (Table 2.2; saved as a tab delimited file), SpotTypes file (Table 2.3; saved as a tab delimited file), the 16 Genepix Results (.gpr) files, and the GenePix Array List (GAL) file.

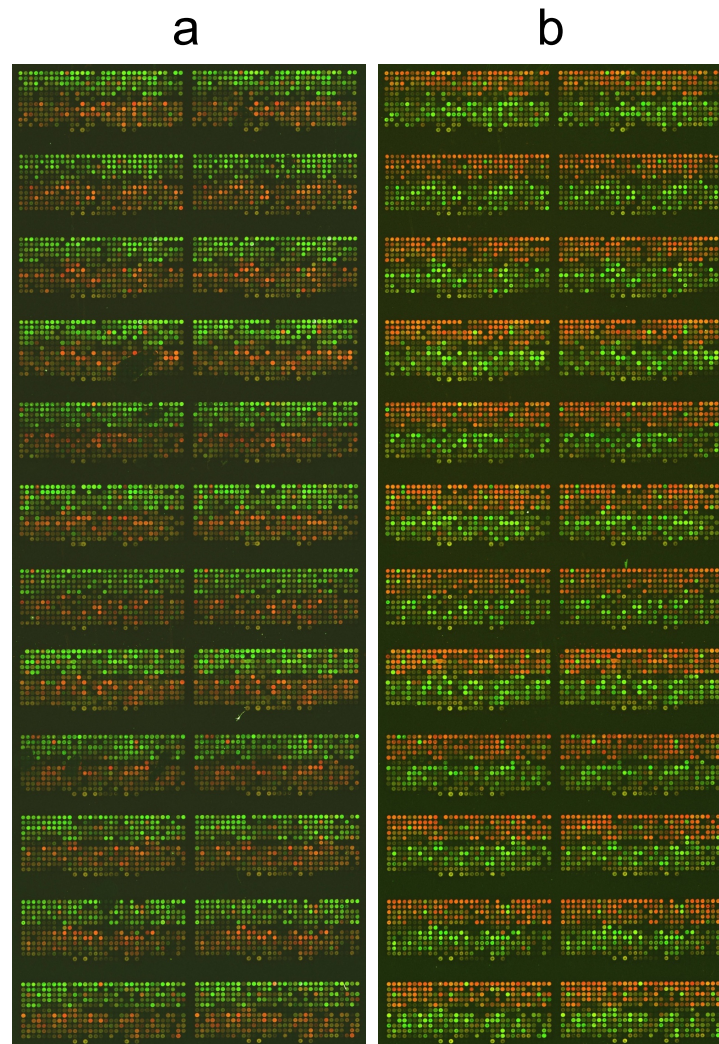


Figure 2.4: (a) Example of a cowpea microarray image following hybridization with differentially labelled cDNA samples, and scanning with a GenePixTM 4000B scanner (Axon Instruments). In this particular example, subtracted tester (ST_F) (cDNA prepared from pooled RNA extracted from IT96D-602 cowpea plants drought stressed for 9 and 12 days, and subtracted with cDNA prepared from RNA isolated from control Tvu7778 plants) was labelled with CyanineTM-3 dye, (green pseudocolour). Unsubtracted tester cDNA (prepared from pooled RNA extracted from IT96D-602 cowpea plants drought stressed for 9 and 12 days) was labelled with CyanineTM-5 dye (red pseudocolour). (b) Dye swap of the experiment in Figure 2.4a. Subtracted tester (ST_F) cDNA was labelled with CyanineTM-5 dye, and unsubtracted tester cDNA was labelled with CyanineTM-3 dye. These differentially labelled cDNA samples were hybridised to the cowpea microarray slide and scanned with a GenePixTM 4000B scanner (Axon Instruments).

Table 2.2: Targets file listing the raw microarray slide data used for the SSHscreen ER3 analysis described in Figure 2.3 (tab-delimited text file).

SlideNumber	FileName	Cy3_F	Cy5_F	Cy3_R	Cy5_R
34	F034_UT_ST.gpr	ST	UT		
22	F022_ST_UT.gpr	UT	ST		
17	R017_UT_ST.gpr			ST	UT
672	R672_ST_UT.gpr			UT	ST
671	F671_UT_ST.gpr	ST	UT		
674	F674_ST_UT.gpr	UT	ST		
670	R670_UT_ST.gpr			ST	UT
19	R019_ST_UT.gpr			UT	ST
673	F673_UD_UT_R673_UT_UD.gpr	UT	UD	UD	UT
18	F018_UT_UD_R018_UD_UT.gpr	UD	UT	UT	UD
669	F669_UD_UT_R669_UT_UD.gpr	UT	UD	UD	UT
668	F668_UT_UD_R668_UD_UT.gpr	UD	UT	UT	UD

Table 2.3: Spot types file listing the raw microarray slide data used for the SSHscreen ER3 analysis described in Figure 2.3 (tab-delimited text file).

SpotType	Name	ID	Color
blank	BLANK	*	NA
cDNA_F	*-F	*	blue
cDNA_R	*-R	*	yellow
control_Globin	control_Globin	*	brown
control_NPTII	control_NPTII	*	red
control_GFP	control_GFP	*	green
control_ITS	control_ITS	*	purple

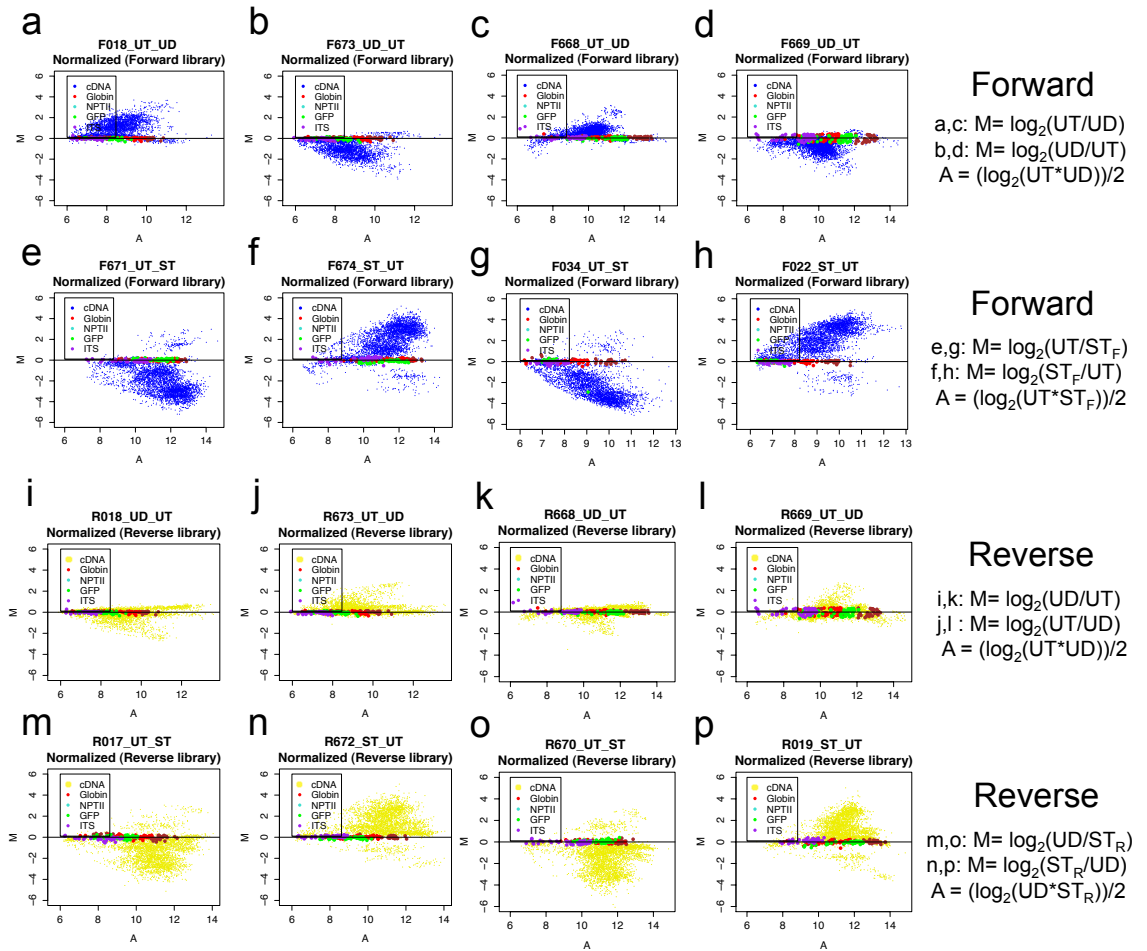


Figure 2.5: M versus A plots for microarray slides after within and between slide normalization in SSHscreen 2.0.0. For each comparison of interest, there were four technical replicates of which two were dye-swaps: plots a-d (forward library, UT versus UD), e-h (forward library, UT versus ST_F), i-l (reverse library, UD versus UT), and m-p (reverse library, UD versus ST_R). Forward and reverse library clones are indicated by blue and yellow dots, respectively. Control spots are indicated as red, light blue, green or mauve dots. M and A values were calculated as described in van den Berg *et al.* (2007), for example (a) $M = \log_2(\text{Cy5 labelled sample} = UT) / (\text{Cy3 labelled sample} = UD)$; $A = (\log_2(UT*UD))/2$; and for example (e) $M = \log_2(\text{Cy5 labelled sample} = UT) / (\text{Cy3 labelled sample} = ST_F)$; $A = (\log_2(UT*ST_F))/2$.

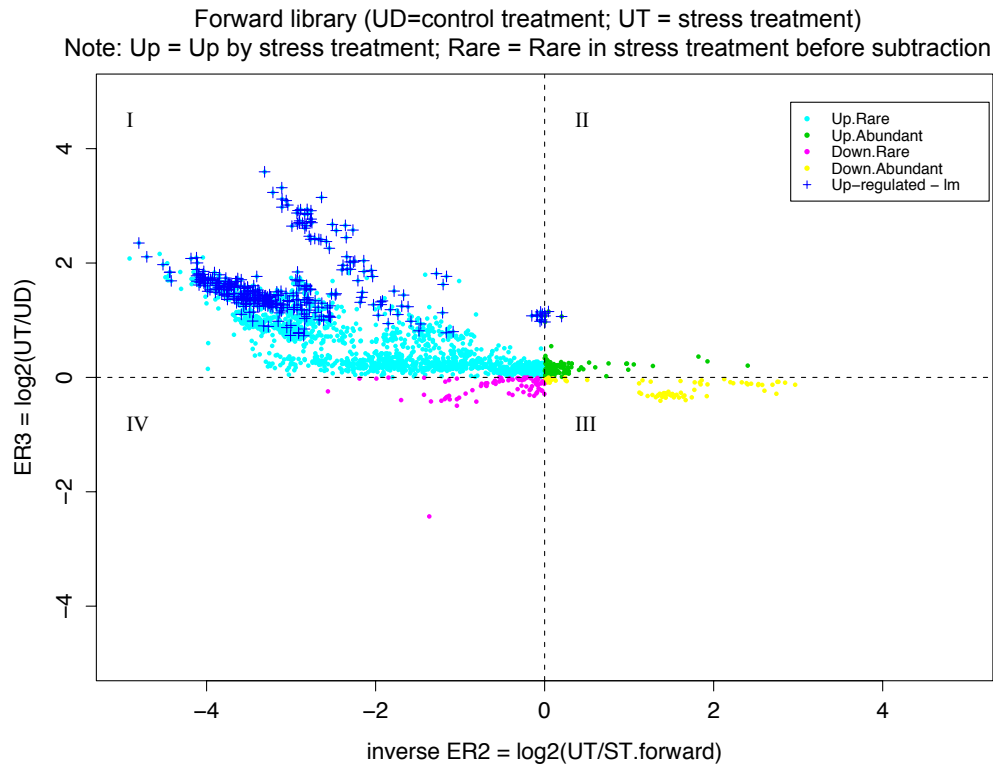


Figure 2.6: ER3 versus inverse ER2 plot produced by SSHscreen for the cowpea forward library. The ER3 versus inverse ER2 plot allows one to visually screen SSH cDNA library clones from the forward library. ER3 for the forward library was calculated as the log-2 ratio of the unsubtracted tester (UT; drought stressed sample) divided by the unsubtracted driver (UD; control sample). Inverse ER2 was calculated as the log-2 ratio of the untreated tester (UT; drought stressed sample) divided by the forward library subtracted tester (ST_F ; SSH library enriched for up-regulated genes). Data points were classified as: up-regulated by stress treatment/rare (Up.Rare) transcripts (quadrant 1; $ER3 > 0$ and $\text{inverse } ER2 < 0$), up-regulated by stress treatment/abundant (Up.Abandant) transcripts (quadrant 2; $ER3 > 0$ and $\text{inverse } ER2 > 0$), down-regulated by stress treatment/rare (Down.Rare) transcripts (quadrant 3; $ER3 < 0$ and $\text{inverse } ER2 > 0$) and down-regulated by stress treatment/abundant (Down.Abandant) transcripts (quadrant 4; $ER3 < 0$ and $\text{inverse } ER2 < 0$). The top 300 statistically significant clones are represented on the plot (adjusted p -value < 0.05).

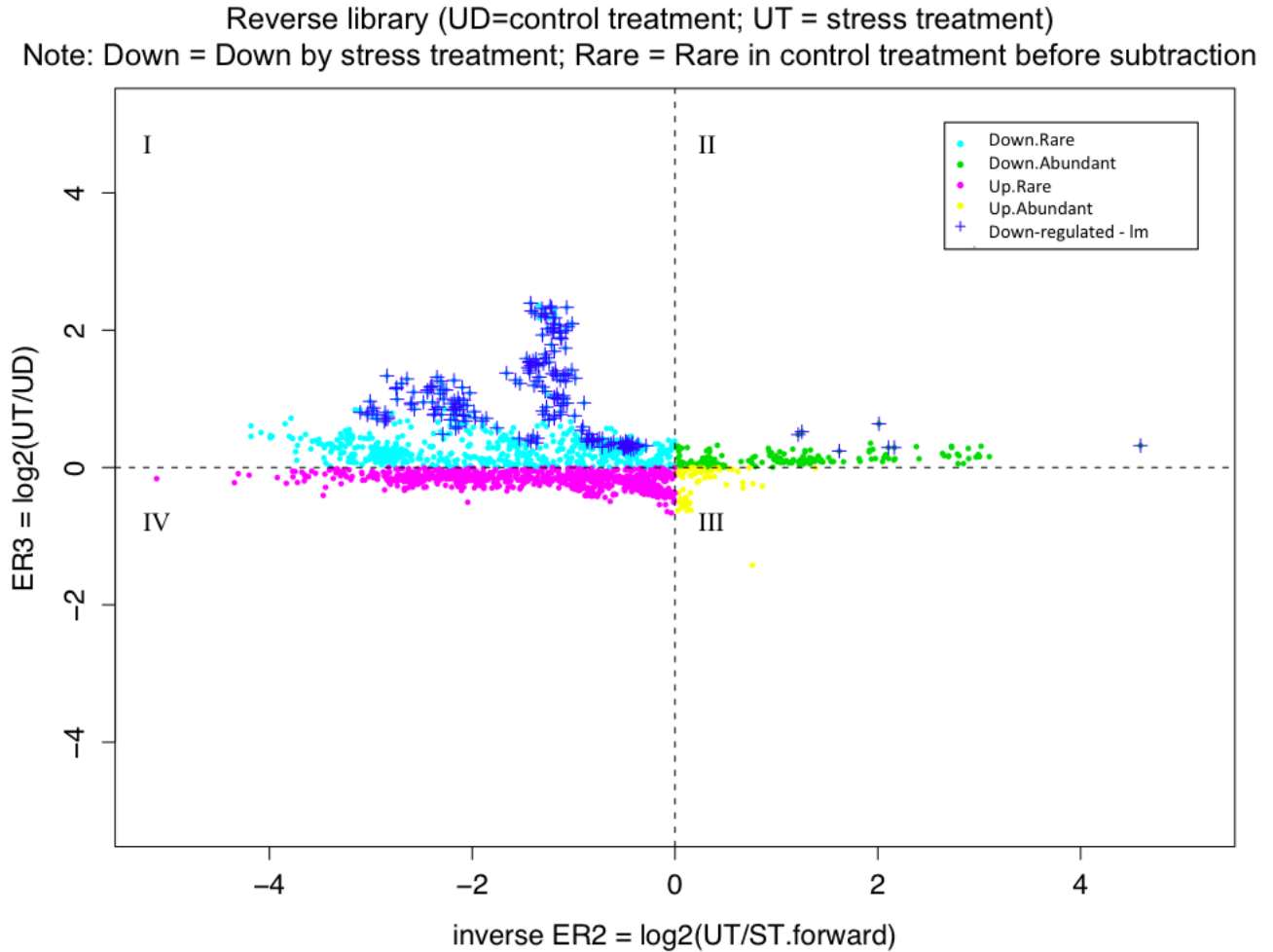


Figure 2.7: ER3 versus inverse ER2 plot produced by SSHscreen for the cowpea reverse library. ER3 for the reverse library was calculated as the log-2 ratio of the unsubtracted driver (UD; control) divided by the unsubtracted tester (UT; drought stressed). Inverse ER2 was calculated as the log-2 ratio of the untreated driver (UD; control sample) divided by the reverse library subtracted tester (STR; SSH library enriched for down-regulated genes). Data points were classified as: down-regulated by stress treatment/rare (Down.Rare) transcripts (quadrant 1; ER3>0 and inverse ER2<0), down-regulated by stress treatment/abundant (Down.Abandant) transcripts (quadrant 2; ER3>0 and inverse ER2>0), up-regulated by stress treatment/rare (Up.Rare) transcripts (quadrant 3; ER3<0 and inverse ER2>0) and up-regulated by stress treatment/abundant (Up.Abandant) transcripts (quadrant 4; ER3<0 and inverse ER2<0). The top 300 statistically significant clones are represented on the plot (adjusted p-value<0.05).

Table 2.4: Top tables produced by SSHscreen for the forward and reverse cowpea libraries. Only the top 20 statistically significant up- and down-regulated genes (before sequencing) are shown (sorted by B statistic).

Forward library top table: up/down regulation

ID	logFC(ER3)*	AveExpr	t	P.Value	adj.P.Val	B	invER2
46D03-F	2.91	10.89	34.37	5.4E-12	1.1E-08	19.51	-2.77
25B07-F	3.09	11.13	31.07	1.5E-11	1.6E-08	19.04	-3.06
13C10-F	2.71	10.62	29.27	2.8E-11	1.9E-08	18.73	-2.75
07E11-F	2.86	10.18	26.11	9.0E-11	3.2E-08	18.10	-2.84
13B02-F	2.77	10.59	25.71	1.1E-10	3.2E-08	18.01	-2.76
07E08-F	2.67	8.97	25.70	1.1E-10	3.2E-08	18.00	-2.51
26F04-F	2.98	11.46	25.60	1.1E-10	3.2E-08	17.98	-3.11
25A05-F	2.94	10.82	25.26	1.3E-10	3.2E-08	17.90	-2.81
25B08-F	2.86	11.25	24.99	1.4E-10	3.2E-08	17.83	-2.80
13C01-F	2.73	10.41	24.74	1.6E-10	3.2E-08	17.77	-2.77
05E09-F	2.58	10.86	23.98	2.1E-10	4.0E-08	17.58	-2.27
25C06-F	2.92	11.08	23.75	2.4E-10	4.0E-08	17.52	-2.92
25B06-F	3.60	11.17	22.99	3.3E-10	5.2E-08	17.31	-3.31
13F02-F	2.61	10.99	22.68	3.8E-10	5.2E-08	17.22	-2.84
13A08-F	2.65	10.84	22.54	4.0E-10	5.2E-08	17.18	-2.82
13C12-F	2.93	11.03	22.53	4.0E-10	5.2E-08	17.17	-2.89
33A07-F	3.32	10.42	22.27	4.5E-10	5.3E-08	17.10	-3.11
33C06-F	3.23	10.61	22.21	4.7E-10	5.3E-08	17.08	-3.22
06H12-F	2.41	10.54	21.83	5.5E-10	5.8E-08	16.96	-2.65
08B07-F	1.88	10.01	21.76	5.7E-10	5.8E-08	16.94	-2.40

*ER3 for Forward library calculated as $\log_2(\text{drought stressed}/\text{control})$

Reverse library top table: up/down regulation

ID	logFC(ER3)*	AveExpr	t	P.Value	adj.P.Val	B	invER2
45C07-R	2.36	11.01	26.14	2.9E-10	5.5E-07	13.66	-1.35
36E04-R	2.25	10.81	22.19	1.4E-09	7.4E-07	13.20	-1.38
36B11-R	2.33	11.31	21.49	1.9E-09	7.4E-07	13.10	-1.22
44C07-R	2.39	11.69	20.93	2.4E-09	7.4E-07	13.02	-1.42
37B05-R	2.33	10.99	20.55	2.8E-09	7.4E-07	12.96	-1.07
45B02-R	2.26	11.05	20.32	3.2E-09	7.4E-07	12.92	-1.20
35F03-R	2.28	11.24	20.11	3.5E-09	7.4E-07	12.88	-1.41
35H05-R	2.35	11.22	20.08	3.5E-09	7.4E-07	12.88	-1.23
35E05-R	2.10	10.79	19.99	3.7E-09	7.4E-07	12.86	-1.02
45E02-R	2.29	10.88	19.88	3.9E-09	7.4E-07	12.84	-1.19
45G05-R	2.18	10.27	19.26	5.2E-09	9.0E-07	12.73	-1.24
35H04-R	2.24	11.22	18.85	6.4E-09	1.0E-06	12.65	-1.29
36A03-R	2.02	10.15	18.21	8.8E-09	1.3E-06	12.51	-1.22
16D06-R	1.65	11.02	17.88	1.1E-08	1.4E-06	12.44	-1.28
16C10-R	1.98	11.01	16.74	1.9E-08	2.3E-06	12.17	-1.20
35A01-R	2.19	10.92	16.73	2.0E-08	2.3E-06	12.16	-1.27
23G10-R	1.74	10.59	16.68	2.0E-08	2.3E-06	12.15	-1.08
45C04-R	2.17	10.50	16.11	2.8E-08	2.9E-06	12.00	-1.34
37C11-R	1.99	11.20	16.06	2.9E-08	2.9E-06	11.98	-1.08
16D05-R	2.09	11.12	15.85	3.2E-08	3.1E-06	11.92	-1.01

*ER3 for Reverse library calculated as $\log_2(\text{control}/\text{drought stressed})$

2.6.3. Annotation and management of cowpea SSH library sequences using SSHdb

The top tables (e.g. Table 2.4) and plots (Figures 2.6 and 2.7) from SSHscreen analysis of the forward and reverse libraries were used to effectively select clones for sequencing based on the criteria of most significant differential expression and least likelihood of sequencing the same gene fragment twice. This was achieved by choosing those clones with the lowest adjusted p-value calculated from the ER3 values. Selection of clones that were spatially separated on the SSHscreen ER plots (Figures 2.6 and 2.7) increased the likelihood of sequencing non-redundant clones. Sequence data for 118 clones, as well as SSHscreen top table data for the entire array, were uploaded to SSHdb for interpretation and management of the data.

Figure 2.8 on the next page gives a schematic representation of the flow of data through SSHdb. For each input cowpea sequence (in FASTA format), SSHdb removed the vector and adaptor fragments by performing BLASTN searches against the NCBI UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>). Next, similarity searches were carried out against all sequences already uploaded in the database, in order to identify clones with the same sequence i.e. redundant partners in the library, using a BLASTN e-value cut-off of $10e-10$. Thirty nine of the 118 sequenced clones were unique, implying that 67% of these sequences were redundant partners (Table 2.5/2.6, starting on page 75). The largest group had 19 redundant partners. For each of the 39 redundant partner groups, the longest sequence in the group was selected by default as the representative clone. The choice of representative clone could be reviewed by downloading from SSHdb the multiple sequence alignments of redundant partner groups with two or more members (generated by ClustalW). Following the identification of redundant partner groups, annotation was performed on the representative clones, with BLASTN and BLASTX (Altschul *et al.*, 1990) against the NCBI non-redundant nucleotide database (nt) and the NCBI non-redundant peptide database (nr), thereby inferring putative functions for each group (Table 2.5/2.6). For cases where the e-value of the top BLASTX hit was low enough (less than $10e-10$), this hit was automatically selected as the default priority annotation. The top 10 BLASTX and BLASTN hits were stored in the database. For each redundant partner group, SSHdb allowed the top BLAST results to be viewed and in several cases the priority annotation was changed after manual inspection. SSHdb linked the selected BLAST annotations to SSHscreen top table entries and it was possible to export different combinations of annotation information for selected subsets of clones (for example, this allowed the construction of Figure 4, see later). One could export selected clones as FASTA files with the functional annotation as part of the header, which was particularly useful in preparing the sequences for submission to GenBank, or as a tab delimited text file containing various columns of available annotation information linked

to the selected clones (Table 2.5/2.6). SSHdb also provides the option to export annotated SSHscreen top tables or Genepix Array List (GAL) files.

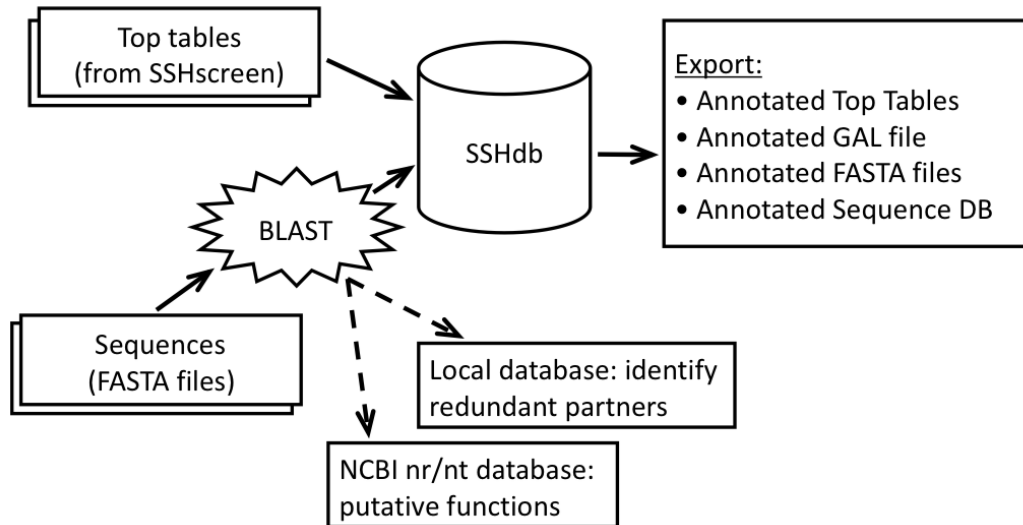


Figure 2.8: Schematic representation of SSHdb. Top tables from SSHscreen, as well as available FASTA sequences for individual clones can be uploaded to SSHdb. For each input FASTA sequence, BLAST searches against the local database are performed in order to identify redundant partners in the library. For each redundant partner group, a representative clone is selected and BLAST searches against the NCBI non-redundant (nr) and nucleotide (nt) databases are performed in order to annotate each group with putative functions. Output from SSHdb includes annotated top tables, annotated GAL files and annotated FASTA files. The user can also export a tab-delimited file of the annotated sequence database, containing all available information about each sequenced clone in the library.

2.6.4. Cowpea SSH library contains genes known to play a role in plant response to stress

Table 2.5/2.6 is a summary of the annotations for the cowpea SSH libraries that were extracted from SSHdb's sequence database. The data for the forward library is sorted by ER3 values, which represents the amount of up-regulation of the transcript in drought stressed IT96D-602 cowpea plants compared to the control treatment. Several genes with known roles in the stress response in other plants were present at high frequency in the cowpea SSH forward library with positive ER3 values, such as glutathione S-transferase (GST), a late

Table 2.5: Top tables produced by SSHscreen for the forward and reverse cowpea libraries. Only the top 20 statistically significant up- and down-regulated genes (before sequencing) are shown (sorted by B statistic).

Forward library											
Group number	Representative clone ID	SSHscreen annotations				SSHdb annotations				RedPartner	
		logFC (ER3)*	adj.P.Val	B	inVER2	Length (bp.) Vector free sequence	BLAST Priority	BLAST AccNr	BLAST Hit Def		BLAST Eval
1	25C06-F	2.92	4.E-08	17.52	-2.92	705	BLASTX	AAM34480	glutathione S-transferase [<i>Phaseolus acutifolius</i>] [GST]	7.E-104	19
2	46D11-F	2.1	1.E-05	10.03	-4.04	543	BLASTX	AAD33696	PR1a precursor [<i>Glycine max</i>] [PR1]	5.E-64	16
3	33D04-F	2.08	5.E-05	8.13	-4.91	780	BLASTX	DQ269446	thaumatin-like [<i>G. max</i>] [THAU]	2.E-66	5
4	26H07-F	1.69	1.E-06	13.97	-2.9	524	BLASTX	P09762	wound-induced protein 2 [<i>Solanum tuberosum</i>]	3.E-63	2
5	07F09-F	1.15	5.E-07	14.86	0.05	473	BLASTX	AAB38782	late embryogenesis abundant 5 [<i>G. max</i>] [LEA]	1.E-27	0
6	46E07-F	1.1	2.E-04	6.36	-0.81	538	BLASTX	AAU14999	MIN19-like protein [<i>Pisum sativum</i>]	7.E-73	0
7	29F12-F	1.06	1.E-04	7.02	-3.47	404	BLASTX	BAA82840	miraculin [<i>Youngia japonica</i>] [MIR]	2.E-10	3
8	26B01-F	0.82	7.E-04	5.08	-2.92	360	BLASTX	BAB33033	ammonium transmembrane transporter [<i>Arabidopsis thaliana</i>]	2.E-60	0
9	26B07-F	0.5	9.E-04	4.85	-3.45	516	BLASTX	NP_193087	universal stress protein (USP) [<i>A.thaliana</i>]	1.E-63	0
10	29D02-F	0.3	3.E-02	1.1	-1.96	359	BLASTX	NP_563888	zinc finger (MYND type) family protein / F-box [<i>A.thaliana</i>]	1.E-14	0
11	30B02-F	0.25	3.E-02	0.93	-2.64	487	BLASTX	NP_564894	calmodulin-binding protein 60-A [<i>Phaseolus vulgaris</i>]	9.E-10	0
12	14B10-F	0.22	1.E-01	-0.45	-1.75	276	BLASTX	AAN65367	cDNA, clone: GMFL01-19-A06 [<i>G. max</i>]	3.E-11	0
13	46G04-F	0.17	1.E-01	-0.68	-1.12	161	BLASTN	AK285943	26S ribosomal RNA [<i>Medicago sativa</i>]	1.E-48	0
14	31A10-F	-0.09	3.E-01	-1.44	1.84	356	BLASTN	AF479110		5.E-176	0

*ER3 for forward library calculated as $\log_2(\text{drought stressed/control})$

Reverse library											
Group number	Representative clone ID	SSHscreen annotations				SSHdb annotations				RedPartner	
		logFC (ER3)*	adj.P.Val	B	inVER2	Length (bp) Vector free Sequence	BLAST Priority	BLAST AccNr	BLAST HitDef		BLAST Eval
15	16C12-R	1.99	5.E-06	11.6	-1.13	653	BLASTX	AAA50172	photosystem II chlorophyll <i>ab</i> -binding (type I) [<i>G. max</i>] [CHL]	3.E-120	10
16	16B08-R	1.53	2.E-05	10.47	-1.24	570	BLASTX	Q01516	Fructose-bisphosphate aldolase 1, chloroplastic [<i>P. sativum</i>]	4.E-92	2
17	38A07-R	1.32	2.E-05	10.29	-2.35	432	BLASTX	NP_181539	LHCb4.3 (light harvesting complex PSI) [<i>A.thaliana</i>] [LHC]	2.E-43	4
18	17C04-R	1.17	3.E-03	5.3	-2.75	439	BLASTX	ACA23202	Kunitz trypsin inhibitor p20-1 [<i>G. max</i>]	9.E-14	1
19	37F06-R	0.94	4.E-04	7.73	-0.9	369	BLASTX	AAQ24882	ribulose-1,5-bisphosphate carboxylase rbcS1 [<i>G. max</i>]	1.E-23	0
20	36G04-R	0.76	1.E-04	8.95	-2.94	669	BLASTX	AAQ74628	lipid transfer protein II [<i>Vigna radiata</i>] [LTP]	1.E-30	3
21	38A10-R	0.75	5.E-03	4.39	-2.08	422	BLASTX	ABF06565	non-specific lipid transfer-like protein [<i>Prosopis juliflora</i>]	2.E-16	0
22	21F04-R	0.46	2.E-02	2.44	-1.67	492	BLASTX	1609235A	chlorophyll <i>ab</i> binding protein 8 [<i>Solanum lycopersicum</i>]	1.E-61	1
23	42G12-R	0.43	2.E-01	-0.05	-3.2	505			No significant similarity		0

Table 2.6: Table 2.5 continue.

Reverse library		SSHscreen annotations				SSHdb annotations			RedPartner		
Group number	Representative clone ID	logFC (ER3) ^a	adj.P.Val	B	inVER2	Length (bp) (Vector free Sequence)	BLAST Priority	BLAST AccNr	BLAST HitDef	BLAST Eval	RedPartner
24	16F09-R	0.3	3.E-03	5.15	-0.48	521	BLASTN	EU196765	23S chloroplast rRNA [<i>P. vulgaris</i>]	0.E+00	1
25	35F12-R	0.29	1.E-01	0.49	-2.78	523	BLASTX	Q43019	lipid-transfer protein 3 [<i>Prunus dulcis</i>]	2.E-31	0
26	15F11-R	0.19	2.E-01	0.09	2.99	620	BLASTN	AF479110	26S ribosomal RNA [<i>M. sativa</i>]	5.E-176	1
27	24B10-R	0.18	3.E-01	-0.43	-2.82	569	BLASTN	EU196765	16S chloroplast rRNA [<i>P. vulgaris</i>]	0.E+00	5
28	36A11-R	0.04	8.E-01	-1.48	-2.39	611	BLASTX	AAU27676	carbonic anhydrase [<i>V. radiata</i>]	1.E-100	0
29	43B08-R	-0.05	7.E-01	-1.4	-0.69	144			No significant similarity		0
30	36A10-R	-0.09	4.E-01	-1.05	-4.31	766	BLASTX	AAG13810	RNA-binding protein Virp1a [<i>S. lycopersicum</i>]	7.E-49	1
31	18C04-R	-0.1	3.E-01	-0.52	-2.02	305	BLASTX	P49972	Signal recognition particle SRP54 [<i>S. lycopersicum</i>]	5.E-23	0
32	38E10-R	-0.1	3.E-01	-0.58	-2.53	254	BLASTX	NP_190638	leucine-rich repeat family protein [<i>A.thaliana</i>]	2.E-26	0
33	16B07-R	-0.11	2.E-01	-0.1	-3.65	723	BLASTX	NP_565656	aldo/keto reductase family protein [<i>A.thaliana</i>]	1.E-109	0
34	21H08-R	-0.13	5.E-01	-1.09	-2.48	746	BLASTX	NP_568178	proline-rich family protein [<i>A.thaliana</i>]	2.E-23	0
35	43E12-R	-0.13	3.E-01	-0.55	-0.28	147	BLASTX	AAI86330	prolyl endopeptidase [<i>A.thaliana</i>]	1.E-13	0
36	21D09-R	-0.14	1.E-01	0.3	-3.77	437	BLASTX	ABI03547	ubiquitin-interacting factor 7 [<i>A.thaliana</i>]	2.E-14	0
37	24G10-R	-0.16	1.E-01	0.62	-1.97	323	BLASTX	Q8SSU8	Phytoene synthase, chloroplastic [<i>Daucus carota</i>]	3.E-02	0
38	43G10-R	-0.16	7.E-02	1.02	-0.36	115	BLASTX	AAK15493	brassinosteroid biosynthetic protein LKB [<i>P. sativum</i>]	3.E-13	0
39	37B02-R	-0.34	2.E-02	2.72	-0.82	541	BLASTN	EU196765	23S chloroplast rRNA [<i>P. vulgaris</i>]	0.E+00	0
40	38A04-R	-0.4	9.E-04	6.84	-0.16	695	BLASTN	AF479105	26S ribosomal RNA [<i>Luglans nigra</i>] [26S]	0.E+00	5

^aER3 for reverse library calculated as $\log_2(\text{control/drought stressed})$

Each redundant partner Group (column 1) is represented by a Representative clone ID (column 2). By default SSHdb selects the clone that is the longest sequence from the group. Columns 3-6 give the regulation (ER3), statistical support for the ER3 value (adj.P.Val and B) and abundance (inVER2) for each group calculated by SSHscreen. Columns 7-12 are annotations added by SSHdb for each group. Length is the length of the representative clone sequence after vector and adaptor fragments are removed. Priority indicates whether the BLASTN or BLASTX hit was selected (this can be changed by the user), BLAST Accession number (AccNr), Hit Definition (HitDef) and E-value (Eval) are for the selected BLAST hit (from the top 10 hits) from either BLASTX/BLASTN results. RedPartners are the number of redundant partner clones in that group (grouped by SSHdb after performing local BLAST searches with a e-value cutoff of $1E^{-10}$).

embryogenesis abundant 5 protein (LEA), miraculin (MIR), thaumatin (THAU), pathogenesis related protein 1 (PR1), cowpea responsive to dehydration 2 (CPRD2), and a universal stress response protein (Table 2.5/2.6). Photosynthesis related genes had positive ER3 values in the reverse library screening indicating that their transcripts were up-regulated in the control treatment, which means they were down-regulated in the drought-stressed IT96D-602 cowpea plants (Table 2.5/2.6). Table 2.5/2.6 illustrates the usefulness of the output from SSHdb, showing the 13 redundant partner groups from the forward library and 26 redundant partner groups from the reverse library. For each group, the representative clone's ID is given, together with the number of redundant partners in that group. Also, each representative clone is labelled with its ER3 value, adjusted p-value, B-statistic and inverse ER2 value calculated by SSHscreen, as well as with a putative function corresponding to the priority selected BLAST result for each group added by SSHdb. The provision of BLASTN results (as well as BLASTX results) is very useful, since several of the priority annotations were BLASTN hits to rRNA of chloroplast or nuclear origin, indicating that some of the highly abundant non-coding RNA had been retained in the mRNA preparation and was cloned in the SSH library. This is most likely due to priming on U-rich tracts within non-coding RNA or self-priming of rRNA during cDNA synthesis (Bloom *et al.*, 2009).

Interestingly, inspection of the ER plots (Figure 2.6 and 2.7) indicates that the majority of the genes (>88%) that were cloned in both the forward and reverse SSH libraries have negative inverse ER2 values (present in quadrants I and IV). This indicates that most of the forward library clones were rare in the drought stressed IT96D-602 cowpea plants, and thus were enriched relative to other transcripts in this sample by the normalization step of the SSH process (Figure 2.6). This is because if the inverse ER2 value ($\log_2(UT/ST_F)$) is negative, then the amount of molecules of the gene is greater in ST_F (i.e. after subtraction) than in UT (before subtraction). The same is true for the reverse library clones, indicating the transcripts are rare relative to other transcripts in the control plants (Figure 2.7).

Figure 2.9 on the following page shows the value of the ER plots to aid in the choice of non-redundant clones for sequencing. To illustrate this, we plotted the ER3 versus inverse ER2 values for a selection of clones from the eight largest redundant partner groups in the library (42 clones from the forward library and 26 from the reverse library). As indicated by the colour coding in Figure 2.9, clones from the same redundant partner groups clustered together. Drought stress up-regulated clones (ER3 > 0) encoding GST (mauve), THAU (red), PR1 (blue) and MIR (yellow) formed clusters that were relatively distinct, thus the choice of a few clones within each region is likely to capture the sequences for most genes in the library. Redundant partners of drought stress down-regulated clones (ER3 < 0 in Figure 2.9) also clustered together, namely lipid transfer protein (LTP; purple), LHCB4.3

light harvesting complex PSII (LHC; orange) and chlorophyll a/b-binding protein (CHL < 90 bp; green) (CHL > 170 bp; dark green). Clones encoding 26S rRNA (26S; blue) also clustered together with ER3 values close to 0. This indicates that 26S rRNA transcripts are present in similar quantities in the stressed and control cowpea plants, as expected, although non-coding RNA was not expected to be captured in either library.

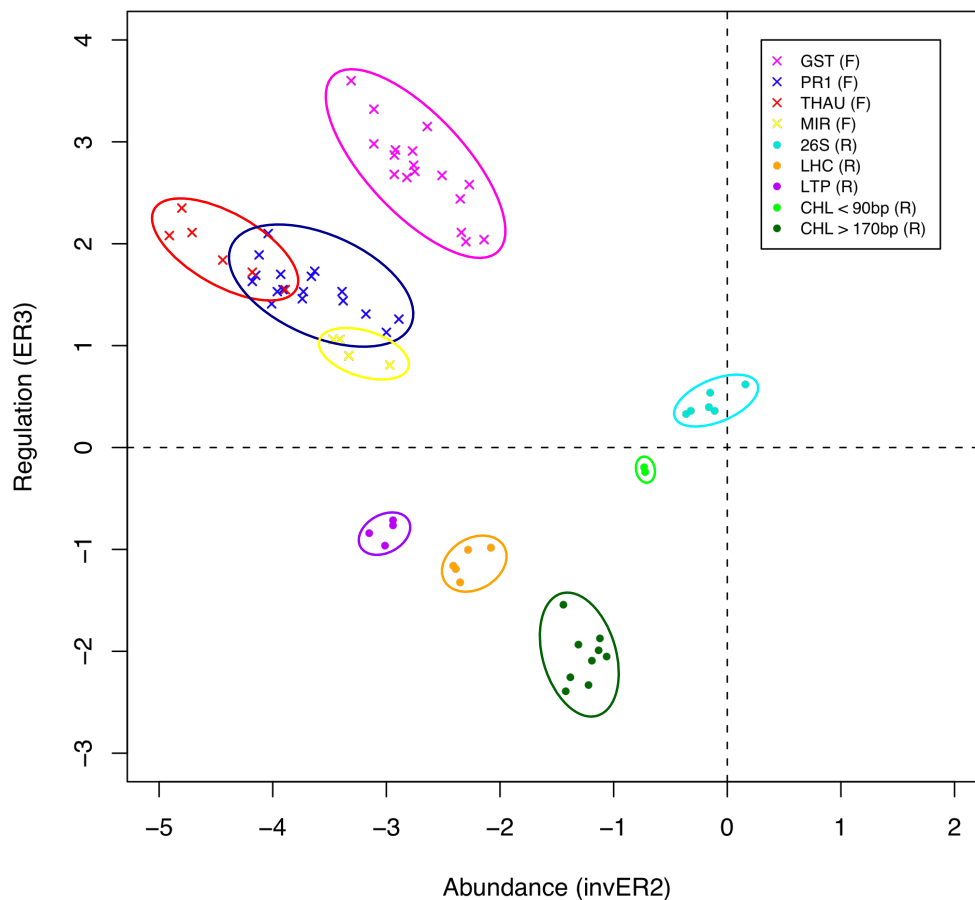


Figure 2.9: ER3 versus inverse ER2 plot for sequenced clones to illustrate that redundant partners cluster together. The ER

values of clones from the eight largest redundant partner groups in the library were plotted. Clones from the same redundant partner groups clustered together. Groups are colour coded and labelled as follows: glutathione S-transferase GST (mauve), pathogenesis related protein 1a (PR1; blue), Thaumatin (THAU; red), miraculin (MIR; yellow), 26S rRNA (26S; blue), light harvesting complex PSII (LHC; orange), lipid transfer protein (LTP; purple), chlorophyll a/b-binding protein (CHL < 90 bp; green) (CHL > 170 bp; dark green).

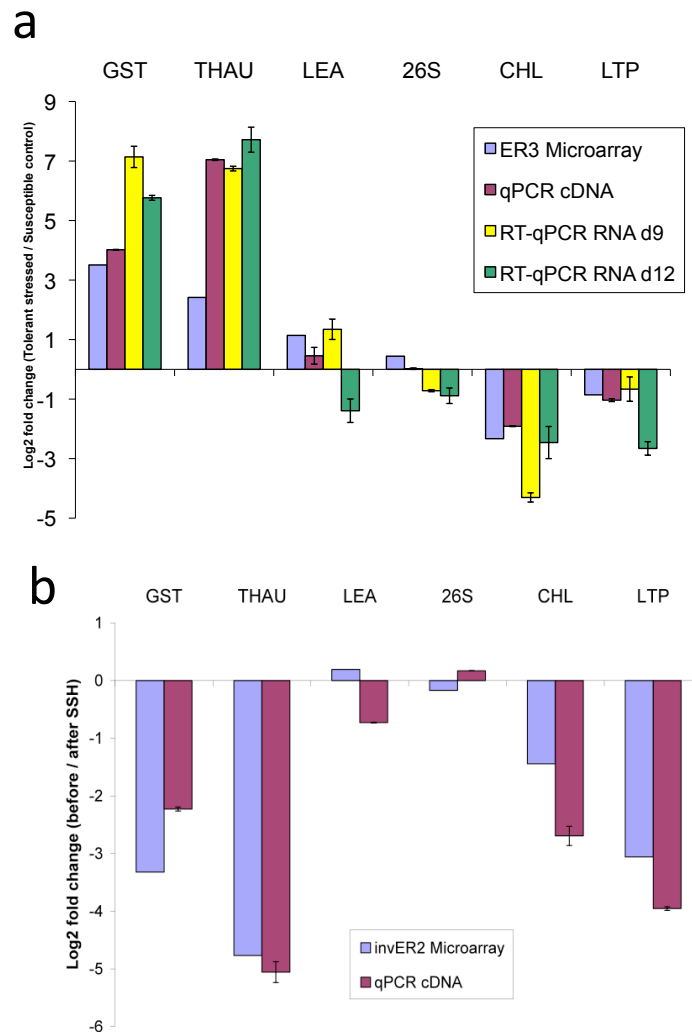


Figure 2.10: (a) Regulation of selected cowpea genes (qPCR verification). Confirmation of differential expression in drought-stressed tolerant cowpea (IT96D-602) versus control susceptible cowpea (Tvu7778) observed in microarray studies. The expression ratios for each gene in the microarray experiment are indicated by blue bars and qPCR on cDNA by red bars. RT-qPCR using total RNA isolated from leaves after 9 and 12 days of stress treatment are indicated in yellow and green bars, respectively. (Error bars = standard deviation of replicate qPCR experiments.) (b) Abundance of selected cowpea genes (qPCR verification). Confirmation that transcripts of selected genes had low abundance (i.e. rare) before subtraction. The log₂ ratios before and after SSH (unsubtracted (UT) / subtracted (ST) ratios) are presented. Negative log₂ ratios indicate that cDNAs have greater signals in ST compared to UT, indicating that they were rare in UT and have been enriched by the SSH process. Results from the microarray experiment are indicated by blue bars and the qPCR results by red bars. (Error bars = standard deviation of replicate qPCR experiments.)

2.6.5. Verification of SSHscreen Enrichment Ratios using qPCR Representative

SSH library clones of six cowpea genes were selected for verification of the SSHscreen enrichment ratios using qPCR. These were three up-regulated genes from the forward library (GST, THAU and LEA; Table 2.5/2.6), two down-regulated genes from the reverse library (CHL and LTP; Table 2.5/2.6), and 26S rRNA which was not differentially expressed (Table 2.5/2.6). qPCR corroborated the direction of gene regulation (ER3 value) calculated by SSHscreen analysis of the microarray data for all six selected genes (Figure 2.10a; compare blue to purple, yellow and green bars). Firstly, qPCR was carried out on the unsubtracted material used to construct the SSH libraries, the same material used to determine the ER3 values. The unsubtracted tester (UT) cDNA sample was a mixture of cDNA from drought stressed cowpea IT96D-602 at 9 and 12 days; and the unsubtracted driver (UD) was a mixture of cDNA from control cowpea Tvu7778 at 9 and 12 days. After normalization of the qPCR data using the glyceraldehyde-3-phosphate dehydrogenase C-subunit (gapC) gene, an expression ratio was calculated ($\log_2(\text{drought stressed cowpea}/\text{control cowpea})$). Good correlation between the ER3 values and the qPCR expression ratios was seen for all six genes (Figure 5a; compare blue bars with purple bars). GST, THAU and LEA were up-regulated, CHL and LTP were down-regulated and 26 S rRNA unchanged (Figure 2.10a). Secondly, since the libraries were constructed from mixtures of cDNA at two time points, reverse transcriptase quantitative PCR (RT-qPCR) was carried out on the RNA samples from the individual time points before they were pooled for SSH library construction (Figure 2.10a, yellow and green bars). GST and THAU were up-regulated, and CHL and LTP were down-regulated at both time points, thus corroborating the ER3 values (Figure 2.10a). Interestingly, LEA was up-regulated at 9 d and down-regulated at 12 d, and thus the transcript abundance measured in the mixtures used to make the SSH library is likely to be an average between the two (Figure 2.10a). RT-qPCR analysis of 26S rRNA at the two time points gave expression ratios that are essentially unchanged between the two treatments (Figure 2.10a).

The SSH process aims to equalize the proportion of genes in the final subtracted sample before cloning by enriching for rare transcripts and suppressing the amplification of highly abundant transcripts (Diatchenko *et al.*, 1999). A rare gene before subtraction should be in increased amounts in the subtracted sample and vice versa. The inverse ER2 value ($\log_2(UT/ST_F)$ for the forward library; $\log_2(UD/ST_R)$ for the reverse library) provides a measure of this, since clones with inverse ER2 value < 0 are rare before subtraction, since $UT < ST$. The cowpea drought expression forward and reverse libraries contained mostly

rare clones, with inverse ER2 values < 0 (see Figures 2.6 and 2.7; blue bars in Figure 2.10b). qPCR was also used to verify the SSHscreen inverse ER2 values using the same cDNA samples, and all five genes that were tested (GST, THAU, LEA, CHL and LTP) gave negative \log_2 (before/after subtraction) values, and closely mirrored the inverse ER2 values, confirming that they were rare in the unsubtracted samples (compare purple with blue bars; Figure 2.10b). 26S rRNA transcripts are expected to be abundant in any plant cell, however the amount of rRNA in the cDNA sample derived from the mRNA isolation step is unlikely to be representative, since it is present due to false priming. Importantly, normalization of the qPCR data for verification of the ER2 values cannot be done with an endogenous housekeeping gene, since no product should be present in the same abundance before and after subtraction. Therefore, equal amounts of an alien gene fragment (human *beta-globin*) were spiked into the cDNA samples and effectively used for normalization of the qPCR data.

2.7. Discussion

SSH remains a popular approach for gene discovery based on its advantages of enriching for genes that are differentially expressed between treatments, as well as the recovery of rare transcripts (Diatchenko *et al.*, 1996; Hillmann *et al.*, 2009). SSH has proven particularly useful as a first step in genomics research of non-model organisms that do not have genome sequence information (van den Berg *et al.*, 2007; Crampton *et al.*, 2009). In this study, we have developed two software tools, SSHscreen and SSHdb, which greatly facilitate gene discovery using SSH (Figure 2.2; Figure 2.8). Furthermore, we have demonstrated functionality of the SSHscreen-SSHdb pipeline with the application to the identification of drought-responsive genes from the non-model crop cowpea. Our approach represents a significant improvement compared to commonly used approaches in which SSH libraries are screened qualitatively using inverse dot blots, and sequence information is stored and managed on an individual researchers desktop.

SSH libraries constructed using either a commercial kit or homemade protocols, have the limitation that they often contain clones derived from transcripts that escaped subtraction (i.e. false positives), clones derived from highly abundant RNA species, such as rRNA, and some redundancy (i.e. the same inserts in several clones) (van den Berg *et al.*, 2004). For example, even though we performed the subtraction effectively, the forward and reverse SSH libraries constructed from cowpea plants in this study were calculated to have 9% and 46% false positives, respectively (negative ER3 values). Despite using mRNA for library construction, 6% of the clones in the reverse library were 26S rRNA, mostly likely due to self-priming or priming on U rich regions by the oligo-dT primer (Bloom *et al.*, 2009). Approximately

67% of the sequenced clones were redundant. This means that sequencing all the clones from an SSH library would be a very inefficient use of resources, since many false positives and redundant clones would be sequenced. Commonly, this is overcome by first screening the SSH library clones as colonies or PCR products on nylon membranes using inverse dot blots (Hein *et al.*, 2004). This, however does not provide accurate quantification, and the choice of clones to use for normalization is difficult. This study provides an alternative approach of using a simple R package SSHscreen 2.0.0 to apply appropriate control spot normalization methods, and calculate differential expression ratios with statistical support after screening the SSH clones on a small number of microarray slides.

SSHscreen 2.0.0 allows the user to apply functions of limma (Smyth, 2005), designed for analysis of microarray expression data, to calculate ER3 values, which reflect whether a clone is derived from transcripts that are up-regulated in the original tester sample (e.g. drought stressed cowpea plants for the forward library). A moderated t-statistic is calculated for each ER3 value (Smyth, 2004), using functions from the limma package (Smyth, 2005). This procedure in effect borrows information from the ensemble of genes to aid with inference about individual genes, taking advantage of the parallel structure whereby the same model is fitted to the data for each gene. Furthermore, the choice of clones for sequencing can be determined by choosing those that have a positive ER3 value and a user-defined threshold of a p-value adjusted for multiple testing (Benjamini and Hochberg, 1995). Importantly, normalization of the microarray data was achieved by spiking the pairs of Cyanine dye-labelled targets with equal amounts of a mix of four alien genes; *globin*, *gfp*, *nptII* and *its*. Implementation of up-weighting print-tip loess control spot normalization in SSHscreen 2.0.0 effectively normalized the dye effect. The forward and the reverse libraries were spotted on the same slides, and thus use of the command `library="both"` allowed separate analyses for the two libraries in one run. These additional functionalities were not available in SSHscreen 1.0.1 (Berger *et al.*, 2007).

Our previous study inferred a measure of the up-regulation of a clone in the SSH library (UT/UD ratio) based on data calculated from two sets of microarray slides, as follows: $\log_2(UT/UD) = [\text{Enrichment ratio 1 (ER1)}(\log_2(ST/UD))] - [(\text{ER2})(\log_2(ST/UT))]$ (van den Berg *et al.*, 2004). The current SSHscreen 2.0.0 approach supersedes the ER1 calculation, since ER3 ($\log_2(UT/UD)$) provides a direct measure of the difference in transcript number between UT and UD on a single set of microarray slides.

Enrichment ratio 2 is calculated in SSHscreen 2.0.0 as an inverse ER2 value ($\log_2(UT/ST)$) for ease of interpretation in the ER3 versus inverse ER2 plots, since it arranges rare \rightarrow abundant transcripts from left \rightarrow right on the plot (Figure 2.6 and 2.7). It gives a measure of whether a clone in the library represents a transcript that was rare or abundant in the

original tester sample, based on the theory of the SSH process that normalizes the relative amount of transcripts in the final subtracted tester sample that is cloned (Diatchenko *et al.*, 1996). SSHscreen 2.0.0 provides a plot of the ER3 versus inverse ER2 values, which provides another tool in the selection of clones for sequencing. As shown in the current cowpea study (Figure 2.9), redundant clones clustered on the ER3 versus inverse ER2 plot, thus these plots can be used to choose clones for sequencing that are spatially separated. Interestingly, this plot was able to distinguish between longer and shorter clones of CHL (Figure 2.9). It should be noted that clusters do overlap (Figure 2.9), so although this plot serves to improve the efficiency of selecting unique clones, some redundant clones will be chosen.

Furthermore, we validated the ER3 and ER2 calculations derived from microarray hybridization signals using an independent technique, qPCR. Three cowpea genes, encoding GST, THAU, and LEA were significantly up-regulated more than 2-fold in the drought-stressed cowpea plants compared to the control plants (ER3 value > 1 ; adjusted p-value < 0.05). qPCR of the UT and UD cDNA mixes prior to subtraction, as well as RT-qPCR of RNA from the individual time points used to make the UT and UD mixes confirmed the up-regulation of these three genes (Figure 2.10a). Interestingly, RT-qPCR showed that LEA was up-regulated at 9 d after initiation of drought stress and down-regulated at 12 d after drought stress, and thus the microarray and qPCR of the UT/UD mixes represent the average (Figure 2.10a). Similarly, ER3 values for two selected down-regulated genes, CHL and LTP were confirmed by qPCR and RT-qPCR. The 26S rRNA escaped subtraction in the construction of the reverse subtraction library and thus is observed to be at equal quantities in the UT and UD samples prior to subtraction (i.e. ER3 ~ 0) and this was also confirmed by the qPCR results (Figure 2.10a). The inverse ER2 values for all five selected differentially expressed genes were negative, indicating that their transcripts were rare in the original tester samples and had been enriched during the normalization step of the SSH process. qPCR confirmed this, indicating that the microarray hybridizations accurately reflect the relative amount of gene fragments in the target cDNA mixes (Figure 2.10b).

The output from SSHscreen is a priority list of clones to sequence, and thus the next step is efficient management of the sequence information in the context of the SSHscreen results. The tool SSHSuite was developed to manage sequence information from SSH libraries (Weckx *et al.*, 2004), however each user is required to install the software as well as the complete NCBI sequence database on a Linux workstation. Since it also lacked several functionalities, such as linkage to SSHscreen top table information, grouping of redundant clones, and customized export of data, we chose to develop SSHdb as a web-based tool with no software requirements for the user except an internet browser. Another advantage of

our approach is that the complete NCBI sequence database is mirrored at a single site and, therefore, can be updated centrally.

SSHdb proved very effective in managing the sequence information for a set of sequences obtained from the forward and reverse drought-stressed cowpea libraries. Each clone was annotated with putative identities based on BLAST similarity searches of sequenced clones against the NCBI non-redundant nucleotide and peptide databases (nt/nr). Several features of SSHdb make it particularly effective for non-model organisms for which there is not an annotated genome sequence available. Providing BLASTN, as well as BLASTX hits, allows the identification of clones derived from non-coding RNA, which escaped the subtraction. This is a common problem in SSH library construction, as seen in our study with 6% clones derived from 26S rRNA. The top ten BLAST hits sorted by E value are stored in the database, and the user is given the choice of choosing the representative annotation. Very often with non-model organisms the top hit is to a sequence that is not functionally annotated (e.g. “hypothetical protein”, “expressed sequence”), whereas the second hit is to an annotated sequence, which can then provide the user with a working hypothesis of the putative identity of the clone. This was our experience for some of the cowpea clones in this study. In other studies, due to the poorly annotated rice genome in GenBank, we found the same problem with SSH clones from non-model monocots, pearl millet and banana, that had top hits to unannotated rice genes, whereas more useful hits within the top 10 were to sequences from other plants with annotations (van den Berg *et al.*, 2007; Crampton *et al.*, 2009).

A useful feature of SSHdb is that it can identify redundant clones in the library, and Figure 2.9 confirmed that redundant clones cluster together on the SSHscreen ER3 versus inverse ER2 plots. The SSHscreen data for each clone can be inspected in the SSHscreen toptable view, and annotated toptables or GAL files can be exported from SSHdb. This is particularly useful in cases where the same array is to be used later for gene expression profiling in a more in-depth study, for example over a time course of drought stress. Such an experiment could be analyzed for differentially expressed genes using limma in R, for example, which would benefit from an annotated GAL file so that it could immediately be seen if differentially expressed clones had been sequenced. In this study, another feature of SSHdb was used to export the representative sequences of each redundant partner group in FASTA format with the correct header information, so that they could be submitted easily to dbEST at GenBank.

SSHdb is not limited to the management and analysis of sequences from SSH libraries, since it can organise any sequence dataset in FASTA format, including cDNA sequences from next generation sequencing projects. The cDNA Annotation System (CAS) is another generic tool for analysis of cDNA sequences (Kasukawa *et al.*, 2003), however it requires the

complete NCBI database to be loaded and up-dated on individual desktops, and thus is less user-friendly for collaborative projects such as ours in which the co-workers are at different institutions.

Confirming the importance of the cowpea genes identified in this study as role players in the drought response is beyond the scope of this paper, which focuses on a comprehensive description of the SSHscreen-SSHdb pipeline. However some inferences can be made by comparison with studies of stress responses in other plants. A glutathione S-transferase, a late embryogenesis abundant protein 5, and a universal stress response protein have clear links to drought stress responses. Glutathione S-transferase (EC 2.5.1.18; GST, group 1, Table 2.5/2.6) is an enzyme that catalyses the conjugation of reduced glutathione, via its sulfhydryl group, to the electrophilic centers on various substrates (Dixon *et al.*, 2002). Glutathione is a tripeptide present in the intracellular space of plants and other organisms, functioning to keep sulfhydryl groups reduced and to remove toxic metabolites. The induction of GST during drought stress in cowpea may protect the plant cells from a build up of toxic compounds, thus contributing to its drought tolerance.

Late embryogenesis abundant (LEA, group 5, Table 2.5/2.6) proteins were initially discovered in desiccating plant seeds but have subsequently been described in various plants and plant tissues. They are associated with abiotic stress tolerance in plants, namely desiccation, salt and cold stress (Tunnacliffe and Wise, 2007). Their structure changes during dehydration from an unordered conformation, lacking in tertiary structure, to a folded structure which may protect the cell from collapse, stabilising membranes or protecting other proteins by acting as chaperones during periods of water stress. Most LEA proteins fall into three main groups, but two unnumbered groups were discovered in cotton: Lea5 and Lea14 (Galau *et al.*, 1993). These two are the only cloned cotton mRNAs encoding LEA's that are highly induced in drought-stressed leaves. They are predicted to be more hydrophobic and possibly more structured than LEA groups 1 - 3 (Tunnacliffe and Wise, 2007). LEA from the cowpea drought expression library in this study has the characteristic Lea5 motif (Pfam family PF03242 [<http://www.sanger.ac.uk/Software/Pfam>]), and is most similar to the drought-induced cotton Lea5 and a Lea5 protein identified in desiccating seeds of soybean (GenBank AAB38782).

The cowpea drought stressed forward library also contained a clone that matched a universal stress protein from *Arabidopsis thaliana* (TAIR:AT5G54430.1). These plant proteins have sequence similarity to UspA that has been well characterized in bacteria. Bacterial UspA is a small serine and threonine phosphoprotein that is induced by several stress treatments, and strains with mutations in this gene are less stress tolerant (Freestone *et al.*, 1997). This may represent an ancient conserved stress mechanism at the cellular level.

(Iuchi *et al.*, 1996) identified genes induced after 5h of dehydration in detached leaves of cowpea line IT84S-2246-4, and named them “cowpea clones responsive to dehydration” (CPRD). One of these genes (CPRD2) was also isolated in our study (Table 2.5/2.6).

Several pathogenesis-related genes were induced during drought stress in cowpea, namely a THAU, PR1, and a wound induced protein (WIN2). Overlap in the responses to biotic and abiotic stresses has been documented (Fujita *et al.*, 2006). This may reflect a structural stabilizing role that these proteins may confer to protect against water loss and cellular damage by either stress. THAU, for example has the unique property of being a very sweet protein with a distinct protein structure made up of beta-sheets with a high content of beta-turns and very few alpha-helices.

The reverse library was dominated by clones encoding components of photosynthesis, such as chlorophyll a/b binding proteins (groups 15, 17 and 22, Table 2.5/2.6) (de Bianchi *et al.*, 2008; Liu *et al.*, 2008), ribulose-1,5-bisphosphate carboxylase small subunit rbcS1, and the chloroplast genes fructose-bisphosphate aldolase 1 and phytoene synthase (Table 2.5/2.6). This reflects a reduction in photosynthesis during drought stress. Orthologues were also down-regulated in leaves of *P. vulgaris* under progressive drought stress (Kavar *et al.*, 2008). They include carbonic anhydrase and the photosynthesis-related genes encoding ribulose 1,5-bisphosphate carboxylase (large and small subunits), chlorophyll a/b-binding protein CP24 precursor and photosystem I light-harvesting chlorophyll a/b-binding protein. Chlorophyll a/b-binding proteins are part of the light-harvesting complex that act as antennae to capture light excitation energy and deliver it to photosystems I and II. In *Arabidopsis*, *cab* genes were also more than 5-fold repressed under drought stress (Seki *et al.*, 2002).

Three different lipid transfer proteins (LTP; group 20, 21 & 25, Table 2.5/2.6) were cloned in the reverse library. Plant LTPs show a highly conserved secondary structure, forming a hydrophobic pocket capable of carrying a fatty acid, phospholipid or acyl-CoA, and have been shown *in vitro* to transfer lipids between membranes (Arondel *et al.*, 2000). Drought responsive LTPs have been described in *Solanum pennellii* (Treviño and O’Connell, 1998). Down-regulation of LTP during drought stress possibly indicates a need to suppress LTP mediated signalling.

The SSHscreen-SSHdb pipeline could be improved in future by developing a GUI version of SSHscreen, taking the user through a step-by-step analysis of the microarray data, similar to the limmaGUI version of limma (Wettenhall and Smyth, 2004). Additionally, an integrated web-based package incorporating SSHscreen and SSHdb functionality could be developed, similar to WebArray (Xia *et al.*, 2005).

2.8. Conclusion

Although there are several alternative approaches such as cDNA-AFLP, DD-RT-PCR and RNA-Seq, SSH remains a popular approach for gene discovery from non-model organisms for which an annotated genome sequence is not available. It is particularly useful for laboratories focused on a particular research question without access to resources to conduct whole transcriptome sequencing using next generation technologies. We have developed the software SSHscreen 2.0.0 which facilitates the quantitative screening of clones in an SSH library from any biological system, and provides the user with a range of statistics to make effective choices of which clones to sequence. The sequence information is then stored and annotated in a web-accessible database, SSHdb, which project collaborators can readily access and interpret for future gene function studies. SSHscreen can be downloaded from <http://microarray.up.ac.za/SSHscreen/>. SSHdb is available at <http://sshdb.bi.up.ac.za/>.

2.9. Acknowledgements

We acknowledge funding from the National Bioinformatics Network (NC), the National Research Foundation (NC, DKB) and a parliamentary grant to the Agricultural Research Council (ARC)-Roodeplaat (IG). We thank Dr. BB Singh for providing cowpea seed and Dr. BD Wingfield (FABI) for providing the *Leptographium elegans* clone. We thank Jan-Peter Nap, Wageningen University and Research Centre, for critical reading of the manuscript.

Chapter 3

Application of SSHscreen-SSHdb pipeline in Pearl Millet

3.1. Note

The construction of the pearl millet SSH library and the microarray screening experiments were carried out by Dr. Bridget Crampton (Molecular Plant-Pathogen Interactions (MPPI) research group, Department of Plant Science, FABI, UP). The data capturing in GenePix and complete data analysis were carried out by myself as part of this MSc dissertation.

3.2. Introduction

Pearl millet (*Pennisetum glaucum* (L.) R. Br.) is an important staple crop in the drought prone semi-arid regions of Africa and Asia. In the USA, Australia and South America, pearl millet is grown mainly for animal feed (Goldman *et al.*, 2003). It is one of the most drought resistant grains (Figure 3.1 on the next page) in commercial production and suffers less from diseases and insect pests than sorghum, maize, or other grains. Being a non-model crop, very little information regarding the pearl millet genome sequence is available. Currently only 2919 pearl millet ESTs have been deposited in GenBank, in contrast with the 189,099 ESTs from cowpea and 1,891,703 ESTs from *Arabidopsis*.

Studies on the pearl millet transcriptome are being done by the Molecular Plant-Pathogen Interactions (MPPI) group at the University of Pretoria (UP). The aim of this programme is to find novel mechanisms of defense against biotic stress. The main biotic stresses of pearl millet are downy mildew and rust diseases.

A pearl millet forward and reverse SSH library, which was enriched for genes either up- or down-regulated in pearl millet leaves at various time points following wounding or treatment with elicitors, has been constructed previously and screened using a high-throughput DNA microarray method (van den Berg *et al.*, 2004). Clones from these SSH libraries were spotted

as probes on microarray slides and hybridizations were done with different combinations of SSH cDNA samples (*UD*, *UT* and *ST*; samples that were used in the construction of the SSH libraries). In this dissertation, these microarrays and SSHscreen version 2.0.0 were used to screen each SSH library for truly differentially expressed genes. After sequencing subsets of genes from the forward and reverse libraries, SSHdb was used to annotate these genes with putative functions so that they could be characterized and studied further.

Using the same elicitor-treated SSH libraries, Crampton *et al.*, 2009, reported on the induction of defense response pathways in pearl millet, in response to infection with the leaf rust fungus *Puccinia substriata*. This study suggests that the salicylic acid (SA) defense pathway is involved in rust resistance, since pretreatment of pearl millet with SA significantly reduced infection levels, in contrast to pretreatment with methyl jasmonate (MeJ). Crampton *et al.*, 2009, hybridized labeled targets for each treatment (SA, MeJA or *P. substriata*) to these SSH libraries arrayed on glass microarray slides. A number of candidate genes were identified that are specifically regulated in response to SA (and not to MeJA), that could thus play a role in resistance to *P. substriata*.



Figure 3.1: Pearl Millet for grain. (a) Maturing Pearl Millet grain, University of Georgia, College of Agricultural and Environmental Sciences (<http://commodities.caes.uga.edu/grasses/grain.htm>). (b) Pearl Millet has small seeds. The image was taken after harvest before cleaning for food use, by J. Wilson, United States Department of Agriculture - Agricultural Research Service (USDA-ARS).

3.3. Methods

3.3.1. Construction of cDNA library using SSH

Dr. Bridget Crampton, part of the MPPI Laboratory (Department of Plant Science, FABI, UP), used SSH to construct a forward and a reverse library of pearl millet (van den Berg *et al.*, 2004). The forward library enriches for target genes that are up-regulated in response to treatment of the plants with elicitors, the mimic attack by bacteria (flagellin), fungi (chitin) and insects (wounding). The reverse library enriches for target genes down-regulated in response to elicitor-treatment. Pearl millet leaves were pricked with a fine needle, and either chitin or an enriched flagellin extract was applied to the abaxial surface of the leaf. A SSH library was prepared by subtracting elicitor-treated pearl millet plants and water-treated (untreated) pearl millet plants. 1920 cDNAs were cloned using SSH.

3.3.2. Screening SSH library on microarrays

3.3.2.1. Slide layout and probes

The library screening was done using six cDNA microarray slides, spotted with the cloned SSH cDNAs. Each microarray slide consists of 12 blocks (print-tip groups).

For the ER1 and ER2 slides (the orange and green arrows in Figure 3.2: slides 36, 38, 40 and 42), each block has 6 rows and 32 columns of spots. On these slides, the forward library clones, ST_F , were spotted in rows 1 - 5 of each block; and control and blank spots were spotted in row 6. No reverse library clones are present on the ER1 and ER2 slides.

The ER3 slides (the blue arrows in Figure 3.2: slides 58 and 114), each has 11 rows and 32 columns of spots. On these slides, the forward library clones, ST_F , were spotted in rows 1-5 of each block and the reverse library clones, ST_R , were spotted in rows 6-10. Control and blank spots were spotted in row 11.

The *gus*, *luc*, and *bar* genes and a fungal rDNA internal transcribed spacer (ITS) fragment were printed in rows 6 and 11 respectively to serve as controls for normalization (van den Berg *et al.*, 2004). Four control genes per print-tip group were printed.

3.3.2.2. Experimental design and targets

Figure 3.2 gives the experimental design of the pearl millet SSH microarray experiment. Each arrow (each representing a microarray slide) connects the two labeled cDNA targets that were hybridized to that slide.

The forward SSH cDNA library was screened using six microarray slides, hybridized with different combinations of *UD*, *UT* and ST_F samples (slides 36, 38, 40, 42, 58 and 114 in

Figure 3.2), to calculate enrichment ratios ER1, ER2 and ER3 for each gene (Berger *et al.*, 2007).

Since there are no reverse library clones present on the ER1 and ER2 slides (see the paragraph above), it makes sense that the ST_R sample was not hybridized to any slides (there are no arrows from the ST_R sample in Figure 3.2). Thus for the reverse SSH cDNA library only enrichment ratio ER3 could be calculated for each gene.

In conclusion, for the forward library SSHscreen '*ER1*' and '*ER3*' analyses were performed using only the forward library clones on all six slides, and since this was not possible for the reverse library an independent limma script was used to identify the truly differentially expressed genes using only the reverse library clones on slides 58 and 114 (untreated versus treated).

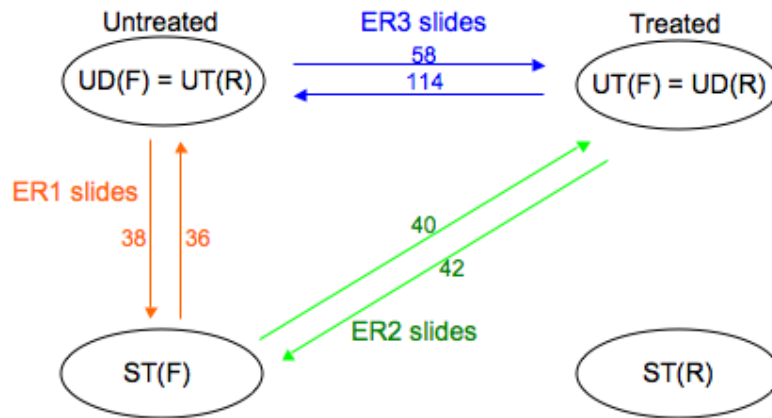


Figure 3.2: Experimental design of the pearl millet microarray experiment. Each arrow represents a microarray slide and the labeled ovals represent RNA samples. The RNA sample to which the arrow points is labeled with Cy5 dye (red) and the sample at the base of the arrow is labeled with Cy3 dye (green). F indicates the forward library and R the reverse library. For the forward library: $ER3 = \log_2(UT_F/UD_F)$, $ER2 = \log_2(ST_F/UT_F)$ and $ER1 = \log_2(ST_F/UD_F)$.

3.3.2.3. SSHscreen analysis of the forward library

GenePix was used to extract the dye intensity data of each spot on each slide into six GenePix Results files (*.gpr* files), one for each microarray slide. Two separate directories were created, one for the ER1 analysis and one for the ER3 analysis. A spot types file and a targets file were created for each analysis, and together with the relevant image analysis output files and the GenePix Array List (GAL) file, stored in the corresponding directories.

The SSHscreen ER3 analysis for the pearl millet forward library microarray slides, was performed by the R command:

```
> SSHscreen(path=~data/millet/ER3vsER2_F", source="genepix", negflags=0,
norm.plot=TRUE, mfrow=c(2,2), legend=TRUE, bc.method="normexp", offset=50,
wa.method="printtiploess", ba.method="Aquantile", weights=TRUE, irregular=TRUE,
ndups=2, spacing=1, spot.ave=FALSE, method="ER3", adjust="fdr", sort="p", cut-
off="0.05", library="F", proportion=0.75)
```

The *path* argument specifies the directory containing the input files; *source* specifies the image analysis program which produced the image analysis output, in this case GenePix was used; *negflags=0* changes the spot quality weights, of all spots receiving a negative flag from the image analysis program, to 0; *norm.plot=TRUE* indicates that MA plots before and after normalization will be produced and saved to the working directory, specified by the *path* argument; *mfrow=c(2,2)* specifies that the plot layout for the MA-plots should be 2×2, i.e. 4 plots (one plot for each microarray slide) on one output window; *legend=TRUE* will include a legend of plotting symbols and colours in the MA-plots; *bc.method="normexp"* indicates that the normexp method should be used for background correction; *offset=50* adds a value of 50 to the intensities before log-transforming; *wa.method="printtiploess"* indicates that the print-tip loess method should be used for within-array normalization; *ba.method="Aquantile"* indicates that the A-quantile method should be used for between-array normalization; *weights=TRUE* assigns zero weight to the cDNAs and double weight to the control spots during normalization; *irregular=TRUE* sorts the gene list by gene ID, since the spacing between duplicate spots is irregular; *ndups=2* specifies that each gene is printed twice on each array; *spacing=1* indicates that consecutive spots are duplicates (after sorting the gene list by gene ID using the *irregular* argument); *spot.ave=FALSE* ensures that duplicate spots on each array will be analyzed separately using limma's function; *method="ER3"* performs a ER3 versus inverse ER2 analysis, using the ER3 and ER2 slides (Figure 3.2) and *method="ER1"* performs a ER1 versus ER2 analysis, using the ER1 and ER2 slides (Figure 3.2); *adjust="fdr"* indicates the p-values should be adjusted for multiple testing using FDR; *sort="p"* will sort the genes in the top table in ascending order according to the adjusted p-values; *cutoff=0.05* ensures that only genes with an adjusted p-value smaller than 0.05 will be included in the top table (this argument is associated with the statistic specified in *sort*); *library="F"* specifies that only a forward library analysis should be done and *proportion=0.75* indicates the assumed proportion of differentially expressed genes in the library. Each argument with all its possible

options and detail, is described in the SSHscreen R documentation (provided in the appendix on page 160).

Output from the SSHscreen ER3 analysis (forward library) were two top tables: *tt.ud* (up/down-regulation; direct comparison of *UT* and *UD*) and *tt.ar* (rare/abundance; direct comparison of *UT* and *ST_F*). A ER1 analysis was also performed on these slides. The ER1 analysis indirectly compares *UT* and *UD* and one up/down-regulation top table, *tt*, was generated. These top tables were uploaded to SSHdb.

3.3.2.4. Limma analysis of the reverse library

For the reverse library limma was used to compare *UT* and *UD* samples, since the reverse library SSH clones were spotted only on slides 58 and 114. For this analysis, the same methods for background correction and normalization were used than for the forward library. A different spot types file and targets file was created, and the original GAL file was manually edited to only include the reverse library probes. A limma top table (direct comparison of *UT* and *UD*) was generated for the reverse library and uploaded to SSHdb. Note that no ER3 and inverse ER2 values were produced in this analysis.

3.3.2.5. Verification and evaluation of SSHscreen results

The SSHscreen forward library results were verified with a separate limma analysis for each pair of slides (i.e. ER3, ER2 and ER1 slides). For each analysis an R script was written and the same argument options (for example background correction and normalization methods) that were used in SSHscreen, were used in separate limma functions. Correlation analyses were performed using MS Excel and R to compare the SSHscreen and limma top tables.

After the forward library *ER1* and *ER3* analyses were performed, the relation $ER3 = ER2 - ER1$ were investigated using MS Excel calculations (see results on page 104).

The impact of different SSHscreen argument values were investigated using the pearl millet data. The approach was to run the SSHscreen function with different values for the a specific argument, while keeping the rest of the argument values fixed, and then compare the output (top tables, MA-plots and ER-plots). In order to compare different within-array normalization methods, various R scripts were written using limma and other functions to generate the plots in Figures 3.4 and 3.5.

Intensity data for the control spots on slides 58 and 114 were extracted from the *.gpr* files using R, and the box plots in Figure 3.3 were generated. The percentage of genes in the top table (genes classified as being significantly differentially expressed) were reported for different cut-off criteria to construct Table 3.1 on page 101.

3.3.3. Sequencing

Selected clones from the pearl millet forward and reverse SSH libraries were sequenced using the T7 Promoter primer by Inqaba Biotec (SA) or Macrogen (USA). Currently 174 sequences are available in FASTA format, 108 forward library clones and 66 reverse library clones.

3.3.4. Management and annotation of clones in the SSH library using SSHdb

All available FASTA sequences were uploaded to SSHdb, which classified each input sequence as part of a redundant partner group and stored the top 10 *BLASTX* and *BLASTN* hits of each group. For each redundant partner group, the *BLAST* results were viewed in SSHdb. When a hit with a good E-value provided a better or more complete description than the default hit selected by SSHdb, the priority annotation was changed.

SSHscreen final top tables for the forward library (up/down-regulation and rare/abundant top tables), containing all the genes in the library ranked in terms of statistical significance, were uploaded to SSHdb. For the reverse library, the final limma top table (up/down-regulation top table) was uploaded to SSHdb. SSHdb linked the top table entry of each sequenced clone, to the priority annotation of the redundant partner group it belongs to. Annotated top tables were exported from SSHdb as tab delimited text files.

Tables 3.3 on page 109 and Figure 3.4 on page 111 are subsets of the information from the annotated SSHdb top tables. Only one representative clone per redundant partner group of the sequenced clones was included in these tables; and the *function category* column was added later (see description in next paragraph).

The representative (sequenced) clone from each redundant partner group, of the forward and reverse libraries respectively, were classified into functional groups (Figure 3.10 on page 108 and 3.11 on page 110), following the strategy described below. BLASTN searches were performed against the model plant *A. thaliana* using MADIBA (www.bi.up.ac.za/MADIBA) to find orthologs. The resulting Atg numbers were submitted to the TAIR functional categorization tool (www.Arabidopsis.org/tools/) to look up the *biological process* GO term for each gene. Each gene was then classified with manual curation, as part of one of the main functional groups: defense, metabolism, protein metabolism, photosynthesis and other.

3.4. Results

3.4.1. Comparison of within-array normalization methods

Figure 3.4 on page 98 gives the MA-plots for the ER3 slides (slides 58 and 114) after normalization with 8 different within-array normalization methods.

MA-plots (a)-(d) were produced without using control spots in the normalization strategy: (a) gives MA-plots of the raw data (no normalization); (b) gives MA-plots after *median* normalization, which simply subtracts the weighted median from the M-values for each slide; (c) gives the MA-plots after global loess normalization, which fits a global loess curve through all the spots on the slide and then subtracts this curve from the M-values for each slide; and (d) gives MA-plots after print-tip loess normalization, which fits print-tip loess curves through the spots in each print-tip group (on the slide) and then subtract these curves from the M-values per print-tip group for each slide. These normalization methods have the effect of centering the cloud of spots around the x-axis (where $M=0$).

MA-plots (e)-(h) were produced using control spots as part of the normalization strategy: (e) gives the MA-plots after control-spot loess normalization, which fits a loess curve only through the set of control spots and applies that curve to all the other spots by subtracting it from the original M-values (in effect, the loess curve through the control spots is shifted to the $M=0$ axis and the cloud of cDNAs are adjusted accordingly); (f) gives the MA-plots after composite loess normalization, which uses a compromise between the print-tip loess curves and the global control-spot loess curve; (g) gives the MA-plots after up-weighting print-tip loess normalization, using spot quality weights when fitting a separate loess curve to each print-tip group (the cDNAs are given zero weight and the control spots double weight) and (h) gives the MA-plots after up-weighting global loess normalization, using spot quality weights when fitting one global loess curve through the data points (the cDNAs are given zero weight and the control spots double weight).

Since SSH enriches for differentially expressed genes, most of the genes in the forward library are expected to be up-regulated by the treatment (positive M-values in slide 58 and the negative M-values in the dye-swap slide 114). Therefore the assumption underlying loess normalization that most of the probes on the array are not differentially expressed, does not hold when working with SSH cDNA libraries. A solution to this is to give more weight (or in this case all the weight) to a set of non-differentially expressed control spots during normalization. This set of control spots should span the intensity range and exhibit a relatively constant expression level across biological samples. On these pearl millet slides, 4 suitable control genes in each of the 12 print-tip groups were printed: *gus*, *luc*, *bar* and *ITS*. After normalization, these control spots are expected to lie on the $M=0$ line.

Looking at the MA-plots in Figure 3.4, control-spot loess (e) and up-weighting print-tip loess (g) normalization give results closest to what is expected. Up-weighting global loess normalization (h) results in an unexpected pattern where the control spots don't cover the entire range of A-values (for example $10 < A < 12$). Composite loess normalization (f) also relies on the set of control spots, but shares the weight between these control spots and the cDNAs for each print-tip group when fitting the loess curves. Median (b), print-tip loess (c) and global loess (d) normalization are not satisfactory for this data, since the assumption that a substantial body of probes do not change expression levels doesn't hold and as a result the data is over-normalized. Since it is known that the control spots do not change expression levels and expected that most of the cDNAs are differentially expressed, the best strategy would be to assign all weight to the control spots and zero weight to the cDNAs as in MA-plots (e), (g) and (h) in Figure 3.4.

3.4.2. Comparison of different control-spot normalization methods

Figure 3.5 on page 99 shows that the results after up-weighting print-tip loess and control-spot loess normalization correlates well: (a) shows that the correlation coefficient of the M-values (log fold changes) after normalization is 0.995 and (b) shows that the correlation coefficient of the order of the genes (ranks) is 0.997. It also shows that the results after up-weighting print-tip loess and up-weighting global loess normalization doesn't correlate as well as expected: (c) shows that the correlation coefficient of the M-values after normalization is 0.64 and (d) shows that the correlation coefficient of the ranks is 0.73. This was a surprising result and can be argued as follows.

Control-spot loess normalization fits a loess curve through a given set of control spots and apply this curve to all the cDNA spots. The *loess* function from the *stats* package is used for local polynomial regression fitting, with the *span* parameter, controlling the degree of smoothing, set to 0.75. This parameter specifies the proportion of data to be used in the local regression moving window. The larger this number, the smoother the loess curve will be. Up-weighting global loess normalization fits a global loess curve through all the spots, but gives double weight to the control spots and zero weight to the cDNAs. Thus in effect, this also fits a loess curve through the set of control spots. However in global loess and print-tip loess normalization, the *loessFit* function from the *limma* package, a fast version of locally weighted regression, is used where the *span* parameter has a value of 0.3 only. This implies that the global loess curve fitted through only the control spots will be less smooth, and regions on the x-axis (A-values) where no control spots spans that area will particularly be influenced. This explains the unexpected pattern around $A=11$ in Figure 3.4 (h). In the same fashion, the up-weighting print-tip loess method fits a loess curve for each print-tip,

giving double weight to the control spots and zero weight to the cDNAs. However, since these arrays have only 4 control spots in each print-tip group (each type of control spot – *gus*, *luc*, *bar* and ITS, is printed in one spot each per print-tip), the *span* parameter will not have such a big influence on the outcome. The 12 fitted loess curves for each microarray slide, ensure that all the control spots lie almost perfectly on the x-axis and the rest of the data points in each print-tip group will be adjusted accordingly.

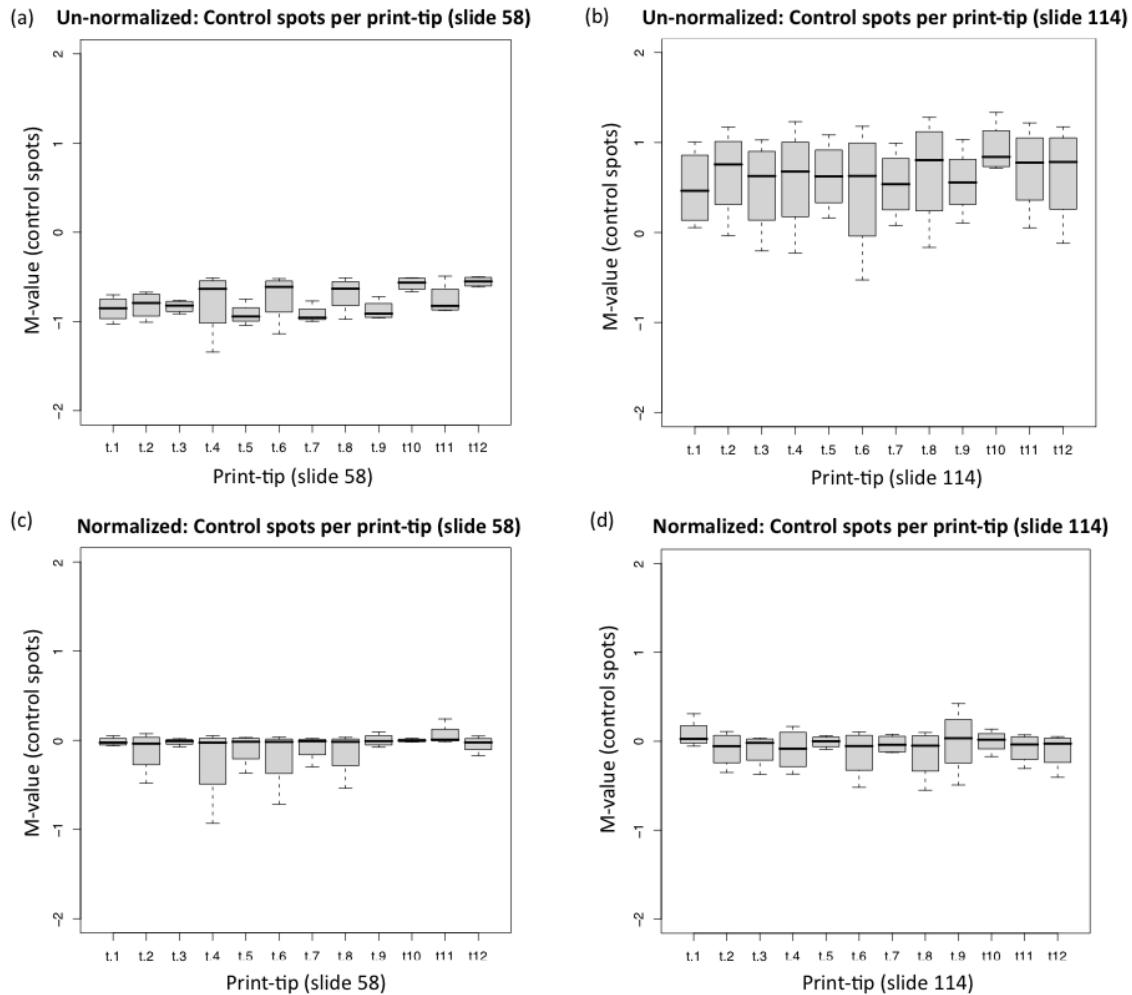


Figure 3.3: Box-plots of M-values (before and after normalization) of the control spots in each print-tip group for two slides. For each of the 12 print-tip groups on slide 58 and slide 114, a box plot of the M-values (fold changes) of the control spots was constructed. (a) and (b) show the print-tip groups before normalization. (c) and (d) show that up-weighting print-tip loess normalization shifted the median (dark middle line) of each box-plot closer to $M=0$. During normalization, the cDNAs data in each print-tip was moved accordingly.

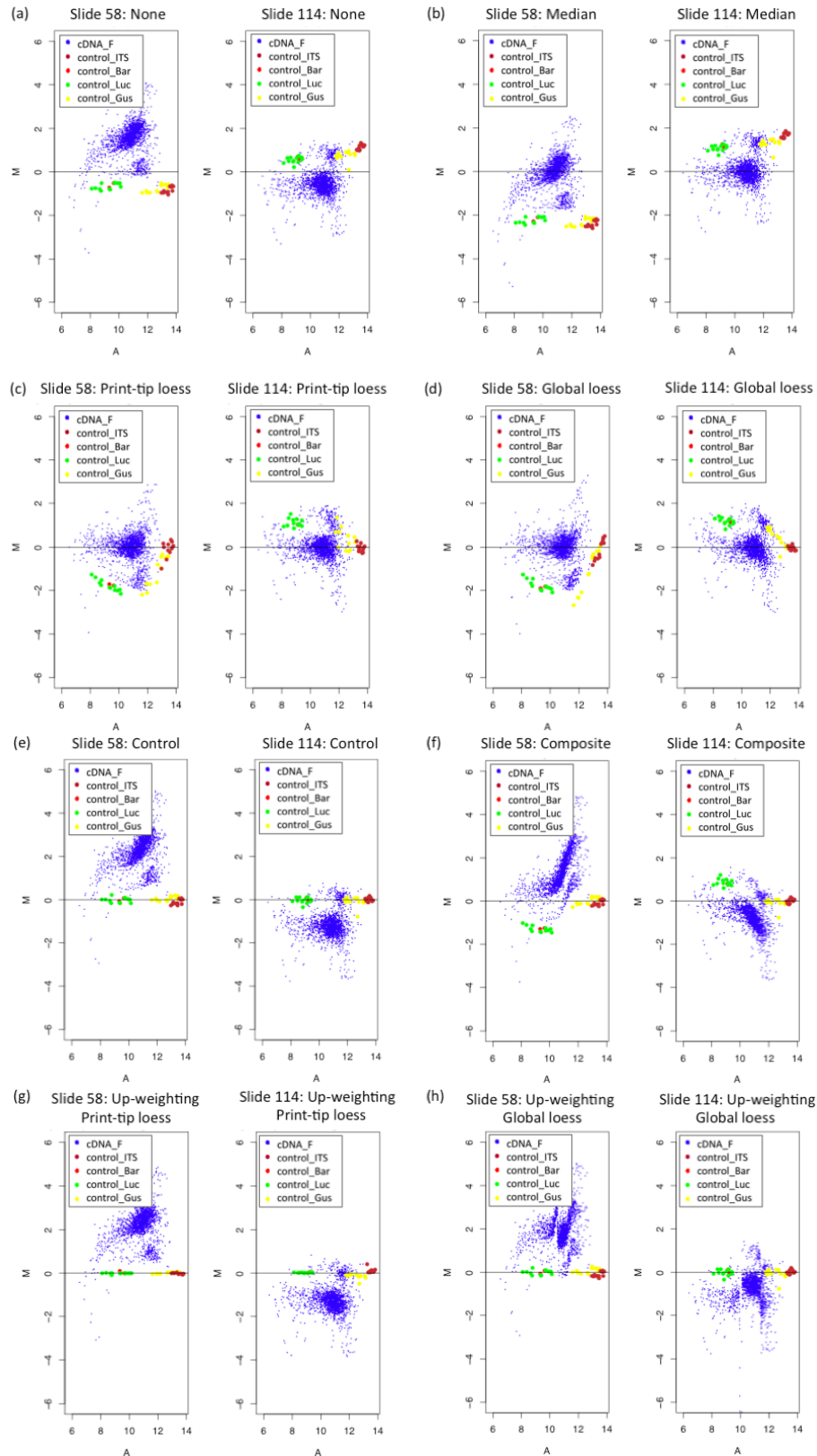


Figure 3.4: M versus A plots illustrating 8 different within-array normalization methods of which the first 4 don't use information from control spots and the last 4 do: (a) no normalization, (b) median normalization, (c) global loess normalization, (d) print-tip loess normalization, (e) control-spot loess normalization, (f) composite normalization, (g) up-weighting print-tip loess normalization and (h) up-weighting global loess normalization.

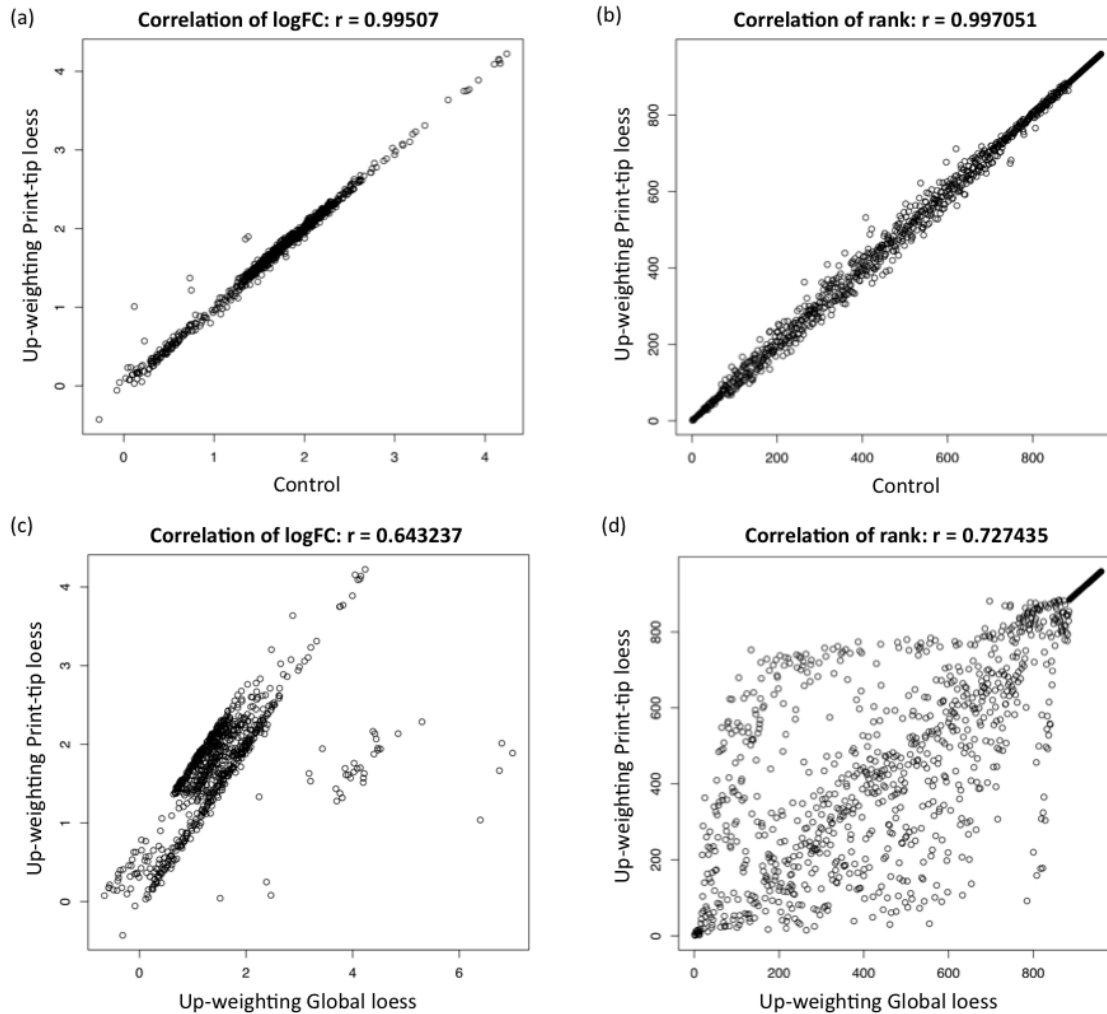


Figure 3.5: Correlation (r = correlation coefficient) between normalization methods using control spots: Forward library, ER3 analysis, up/down-regulation top table. (a) is a plot of the correlation between the M-values (log fold changes) in two top tables after normalizing with the up-weighting print-tip loess and control-spot loess methods respectively. (b) is a plot of the correlation between the order of the genes (ranks) in two top tables after normalizing with the up-weighting print-tip loess and control-spot loess methods respectively. (c) is a plot of the correlation between the M-values (log fold changes) in two top tables after normalizing with the up-weighting print-tip loess and up-weighting global loess methods respectively. (d) is a plot of the correlation between the order of the genes (ranks) in two top tables after normalizing with the up-weighting print-tip loess and up-weighting global loess methods respectively.

3.4.3. Taking the print-tip effect into account in control-spot normalization

Figure 3.3 on page 97 gives a box-plot of the M-values (before and after normalization) of only the control spots in each of the 12 print-tip groups on slide 58 (a and c) and slide 114 (b and d). The control spots are expected to have constant expression levels across all intensity values (i.e. lie in a horizontal line on the MA-plots in Figure 3.4) and to be non-differentially expressed (i.e. lie on the x-axis where $M=0$ in the MA-plots in Figure 3.4). There are 4 control genes per print-tip group: *gus*, *luc*, *bar* and *ITS*. Looking at the box-plots of expression levels in Figure 3.3, there are slight differences between print-tip groups that could be worthwhile to take into account. Overall, the spread of the boxplots in slide 114 is larger (larger variation between expression levels) than that for slide 58. Also, the control spots in slide 58 are more green (negative M-values) than expected and in slide 114 more red (positive M-values) than expected.

Using the up-weighting print-tip loess normalization will decrease the variation in expression levels of the control spots in each print-tip (resulting in smaller boxes after normalization, see Figure 3.3) and shifts the median (dark middle line) of each box-plot closer to $M=0$. All cDNA spots in the corresponding print-tip groups will also be moved accordingly.

3.4.4. Top Table statistics

Table 3.1 on the following page again compares the three control-spot normalization methods. It shows the effect of different top table statistics and corresponding cut-off values that can be selected to classify a subset of genes as *significantly differentially expressed* (present in the top table). The prior guess of the number (proportion) of differentially expressed genes in the library, necessary to calculate the B-statistic is also investigated.

From Table 3.1 it is clear that the percentage of genes in the top table for different cut-off criteria, using control-spot loess normalization (column 1) and up-weighting print-tip loess normalization (column 2), are comparable. For example, for an adjusted p-value < 0.05 (the generally preferred cut-off criterion), 70.1% of the genes in the library are included in the top table using control-spot loess normalization and 74.58% using up-weighting print-tip loess normalization. Since the SSH technique enriches for differentially expressed genes, it is expected that most genes ($> 50\%$) should appear in the top table. In contrast to the good results in columns 1 and 2 of Table 3.1, results for the up-weighting global loess normalization (column 3) are suspect and no genes come up as significantly differentially expressed according the adjusted p-value. This result can be explained using the same reasoning as above (see page 96).

The percentage of genes with a moderated t-statistic between -2 and 2, for data with 5

degrees of freedom, normally corresponds to percentage of genes with an adjusted p-value < 0.05 . Table 3.1 columns 1 and 2 confirms this. A positive B-statistic ($B > 0$) corresponds to a 50/50 chance that a gene is differentially expressed, but calculating the B-statistic requires a prior guess specifying the proportion of expected differentially expressed genes p_j (see equation 1.17 on page 35 and the section on estimation of hyperparameters on page 37). The default proportion p_j in limma is 0.01 (1%). For a SSH library the proportion p_j should be much higher, since most of the genes are expected to be differentially expressed. Table 3.1 shows that the percentage of genes included in the top table increases exponentially as p_j increases. The default value for this expected proportion of differentially expressed genes in SSHscreen is 0.75 (the *proportion* argument in SSHscreen).

Table 3.1: Proportion of genes included in the top table (out of all the genes in the forward library), using different cut-off criteria, different prior guesses of the number of differentially expressed genes, as well as different control-spot within-array normalization methods.

Cut-off criterion⁺ and prior guess of proportion of DE genes[§]	control-spot loess (%)	up-weighting printtiploess (%)	up-weighting global loess (%)
adjusted p-value < 0.05	70,10	74,58	0,00
$-2 < \text{moderated } t < 2$	72,81	74,79	40,21
$B > 0$ (prop = 0.01)	5,83	7,92	0,42
$B > 0$ (prop = 0.15)	48,23	60,63	9,27
$B > 0$ (prop = 0.3)	67,60	73,02	28,33
$B > 0$ (prop = 0.5)	77,92	78,44	63,23
$B > 0$ (prop = 0.75)	92,08	91,04	92,08
$B > 0$ (prop = 0.9)	92,08	92,08	92,08

+ The cut-off criterion includes a statistic and a corresponding cut-off value for genes to be classified as 'significantly differentially expressed'

§ Prior guess of the proportion of differentially expressed (DE) genes only influences the B-statistic

3.4.5. SSHscreen forward library ER3 analysis

SSHscreen input files for the ER3 analysis include a targets file, a spot types file and a GAL file (given in Figure 1.4 on page 13), as well as the *.gpr* files for slides 40, 42, 58 and 114. Executing SSHscreen (see the R command on page 121), generated two top tables and two ER-plots; one pair indicating up/down regulation and the other for rarity/abundance. Up-weighting print-tip loess normalization was used, and an adjusted p-value < 0.05 was selected as a cut-off criterion.

Figure 3.6 on the next page gives the ER3 versus inverse ER2 plot showing the significantly up-regulated genes (the genes marked with blue crosses) for the forward library, as

calculated by the linear model, and Table 3.2 on the following page shows the top 30 entries of the up/down regulation top table. A cut-off criterion of the adjusted p-value < 0.05 was implemented.

The ER-plot and top table from the rarity/abundance analysis are not shown.

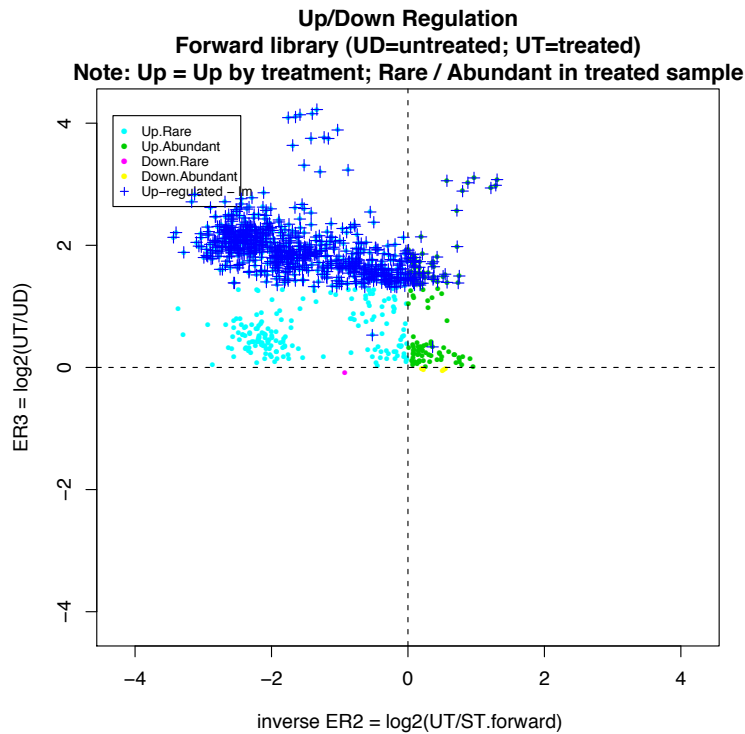


Figure 3.6: The ER3 versus inverse ER2 plot produced with SSHscreen 2.0.0, allow one to visually screen SSH cDNA library clones from the forward library. ER3 is calculated as the log-2 ratio of the un-subtracted tester (elicitor treated sample) divided by the un-subtracted driver (untreated sample). Inverse ER2 is calculated as the log-2 ratio of the un-subtracted tester (elicitor treated sample) divided by the forward/reverse library subtracted tester (SSH library enriched for up/down-regulated genes respectively). The plot shows the significantly up- and down-regulated genes, as calculated by the linear model (the marked genes). A cut-off criterion of the adjusted p-value < 0.05 was implemented. The up-weighting print-tip loess within-array normalization method was used i.e. print-tip loess normalization, assigning zero weight to the cDNAs and double weight to the control spots. Data points are classified as: up-regulated/rare (Up.Rare) transcripts (quadrant 1; $ER3 > 0$ and $invER2 < 0$), up-regulated/abundant (Up.Abandant) transcripts (quadrant 2; $ER3 > 0$ and $invER2 > 0$), down-regulated/rare (Down.Rare) transcripts (quadrant 3; $ER3 < 0$ and $invER2 > 0$) and down-regulated/abundant (Down.Abandant) transcripts (quadrant 4; $ER3 < 0$ and $invER2 < 0$).

Table 3.2: Top Table produced by SSHscreen 2.2.0 for the pearl millet forward SSH library, showing the top 30 statistically significant up-regulated genes.

Forward library top table: up/down regulation (SSHscreen)										
Block	Column	Row	ID	Status	logFC	AveExpr	t	P.Value	adj.P.Val	B
7	11	5	09-F7	cDNA	4.22	11.72	6.58	4.6E-11	3.2E-08	14.66
7	10	3	05-E8	cDNA	4.16	11.66	6.48	9.3E-11	3.2E-08	13.99
7	21	4	08-C7	cDNA	4.14	11.96	6.45	1.1E-10	3.2E-08	13.82
3	8	4	07-D4	cDNA	4.10	12.19	6.39	1.6E-10	3.2E-08	13.46
7	28	2	04-F8	cDNA	4.09	12.03	6.38	1.8E-10	3.2E-08	13.37
5	22	5	10-C6	cDNA	3.89	11.52	6.06	1.3E-09	2.0E-07	11.47
5	26	4	08-E6	cDNA	3.77	11.83	5.88	4.2E-09	5.0E-07	10.40
9	23	3	06-D9	cDNA	3.75	12.52	5.85	4.9E-09	5.0E-07	10.24
7	24	2	04-D8	cDNA	3.75	12.10	5.85	5.1E-09	5.0E-07	10.22
11	13	5	09-G11	cDNA	3.64	10.82	5.67	1.4E-08	1.3E-06	9.24
7	32	3	06-H8	cDNA	3.31	11.98	5.16	2.4E-07	2.0E-05	6.59
11	14	5	09-G12	cDNA	3.23	9.04	5.04	4.7E-07	3.5E-05	5.98
9	20	1	02-B10	cDNA	3.20	10.88	4.99	5.9E-07	4.0E-05	5.76
5	12	2	03-F6	cDNA	3.10	11.80	4.84	1.3E-06	8.3E-05	5.02
5	12	3	05-F6	cDNA	3.08	11.38	4.80	1.6E-06	9.6E-05	4.82
1	14	5	09-G2	cDNA	3.06	11.85	4.77	1.9E-06	1.0E-04	4.69
7	10	2	03-E8	cDNA	3.02	11.30	4.72	2.4E-06	1.3E-04	4.45
7	25	2	04-E7	cDNA	2.98	11.93	4.65	3.3E-06	1.6E-04	4.16
3	30	2	04-G4	cDNA	2.94	12.20	4.58	4.6E-06	2.1E-04	3.85
9	6	1	01-C10	cDNA	2.89	11.49	4.50	6.7E-06	3.0E-04	3.50
1	16	5	09-H2	cDNA	2.86	11.65	4.46	8.2E-06	3.4E-04	3.32
7	9	3	05-E7	cDNA	2.83	11.25	4.42	1.0E-05	4.0E-04	3.13
5	27	3	06-F5	cDNA	2.78	10.53	4.33	1.5E-05	5.7E-04	2.77
9	21	5	10-C9	cDNA	2.76	11.00	4.31	1.7E-05	6.1E-04	2.67
5	11	2	03-F5	cDNA	2.71	11.49	4.23	2.4E-05	8.0E-04	2.35
5	24	5	10-D6	cDNA	2.71	11.44	4.23	2.4E-05	8.0E-04	2.34
7	8	2	03-D8	cDNA	2.68	11.04	4.17	3.0E-05	9.8E-04	2.12
7	32	5	10-H8	cDNA	2.67	10.34	4.16	3.2E-05	1.0E-03	2.08

3.4.6. SSHscreen forward library ER1 analysis

3.4.6.1. ER1 analysis using SSHscreen 2.0.0 in R

The same argument options were selected for the SSHscreen ER1 analysis of the forward library (see the R command on page 121), except *method* was changed to 'ER1'. For this analysis, a different targets file was created specifying details for microarray slides 36, 38, 40 and 42 (Figure 3.2 on page 91 for the experimental design). Executing the SSHscreen function in R for the ER1 analysis, produced one top table and one ER-plot after an indirect comparison of the UD and UT samples, using a contrast matrix (see the section on page 29). Therefore significantly up/down-regulated clones can be marked in the ER1 versus ER2 plot, as shown in Figure 3.7. This plot shows the top 100 up-regulated genes, marked with blue crosses (the SSHscreen *toplist* argument was used to select the top 100 genes). Data points on the diagonal line in Figure 3.7 represents genes that are unchanged by the treatment. On the diagonal line $ER1 = ER2$, i.e. $\log_2(ST/UD) = \log_2(ST/UT)$ which implies that $UD = UT$. Genes above the diagonal line, where $ER1 > ER2$ and thus $UT > UD$, represent genes up-regulated in the treated sample. Genes marked in red (with positive ER2 values) are

rare and genes marked in green (with negative ER2 values) are abundant. Although the top 100 genes are marked in Figure 3.7, these genes are not statistically significant differentially expressed. When using a p-value cut-off of 0.05, only 3 genes are significant and for a cut-off of 0.1 only 7 genes are significant. This is in contrast with output from the ER3 analysis.

3.4.6.2. ER1 versus ER2 plot in MS Excel

Before SSHscreen was written, MS Excel was used to do a ER1 versus ER2 analysis using the same pearl millet SSH cDNA libraries (van den Berg *et al.*, 2004). Slightly different results were expected since ArrayVision, instead of GenePix, was used to calculate the signal densities. But control spots were also used as part of the normalization strategy. Later the ER1 analysis was written in R, in a more sophisticated and automated way, as the first version of SSHscreen. The ER3 analysis was added later. Figure 3.8 on page 106 gives the ER1 versus ER2 plot after the MS Excel ER1 analysis (van den Berg *et al.*, 2004) and by eye it correlates well with the SSHscreen 2.0.0 ER1 versus ER2 plot in Figure 3.7 on the next page.

3.4.6.3. Predicting the ER3 value from ER1 and ER2

The relation $ER1 - ER2 = ER3$, can algebraically be shown, i.e.

$$\begin{aligned}
 & ER1 - ER2 \\
 &= \log_2(ST/UD) - \log_2(ST/UT) \\
 &= \log_2(ST/UD * UT/ST) \\
 &= \log_2(UT/UD) \\
 &= ER3 \text{ (= M-value giving an indication of up/down regulation)}
 \end{aligned}$$

Following this relation, from the ER1 analysis which calculates an ER1 and an ER2 value for each gene, an ER3 value can indirectly be calculated. Although this makes sense in theory, it is not the case in practice. The correlation coefficient between newly calculated ER3 values (from the ER1 analysis) using the above relation and the original ER3 values (from the ER3 analysis) is 0.11 and the correlation coefficient between the ranks of the two top tables (both top tables are ranked in terms of up/down regulation) is 0.10. The conclusion that follows is that using this pearl millet data, it was not possible to predict the ER3 values from only the ER1 and ER2 slides. This could be due to the fact that the ER1, ER2 and ER3 slides are different hybridizations from different slides.

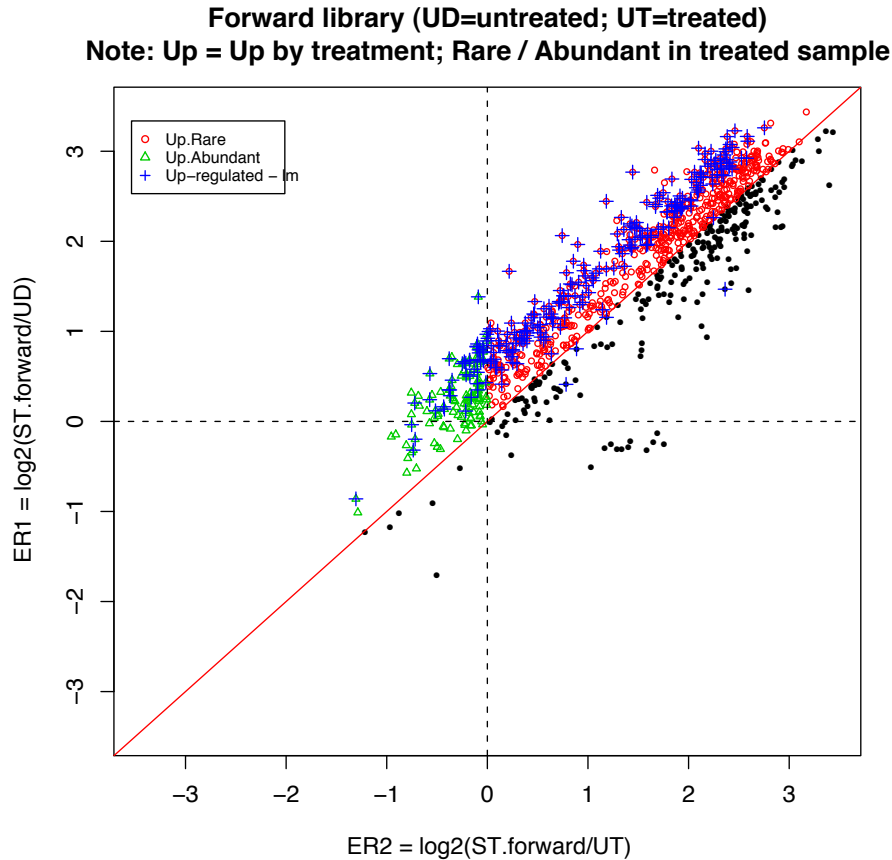


Figure 3.7: ER1 versus ER2 plot produced with SSHscreen 2.0.0 for the forward subtraction library. Data points on the diagonal line represents genes that are unchanged by the treatment. On the diagonal line $ER1 = ER2$, i.e. $\log_2(\text{ST}/\text{UD}) = \log_2(\text{ST}/\text{UT})$ which implies that $UD = UT$. Genes above the diagonal line, where $ER1 > ER2$ and thus $UT > UD$, represent genes up-regulated in the treated sample. Genes marked in red (with positive ER2 values) are rare and genes marked in green (with negative ER2 values) are abundant. The plot shows the top 100 up- and down-regulated genes. The up-weighting print-tip loess within-array normalization method was used i.e. print-tip loess normalization, assigning zero weight to the cDNAs and double weight to the control spots.

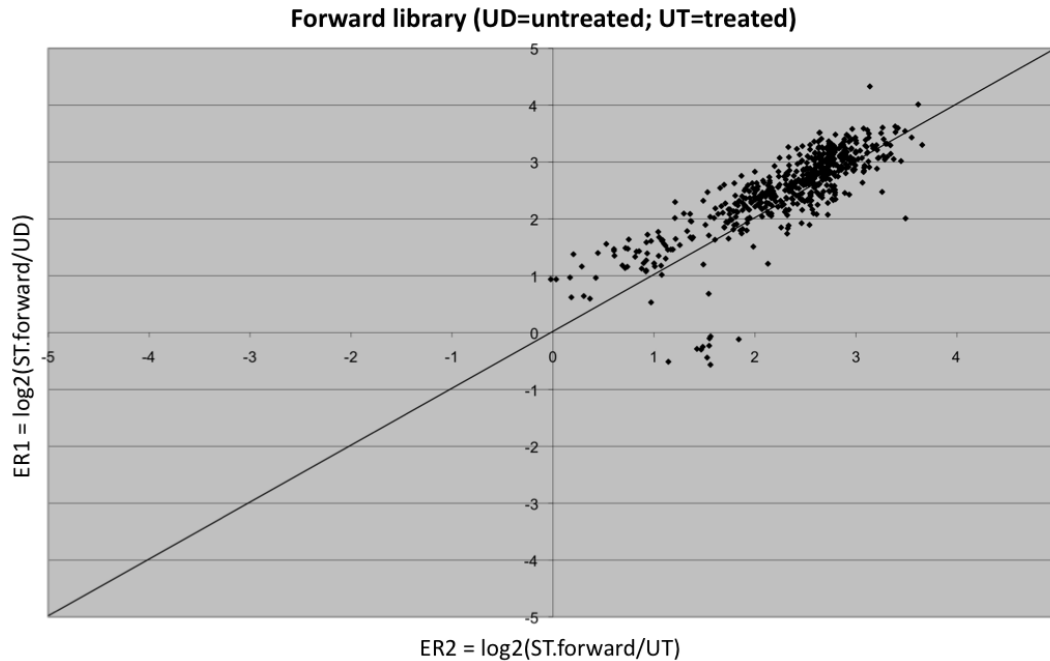


Figure 3.8: ER1 versus ER2 plot produced with MS Excel for the millet forward subtraction library (van den Berg *et al.*, 2004). Before SSHscreen was written, the ER values were calculated and plots were drawn using MS Excel. Data points on the diagonal line represents genes that are unchanged by the treatment. On the diagonal line $ER1 = ER2$, i.e. $\log_2(ST/UD) = \log_2(ST/UT)$ which implies that $UD = UT$. Genes above the diagonal line, where $ER1 > ER2$ and thus $UT > UD$, represent genes up-regulated in the treated sample. Genes with positive ER2 values are rare and genes with negative ER2 values are abundant. Control spots were used as part of the normalization strategy.

3.4.7. Using limma to verify the SSHscreen values

Separate limma analyses were performed on the ER1 slides (dye-swapped slides 36 and 38), ER2 slides (dye-swapped slides 40 and 42) and ER3 slides (dye-swapped slides 58 and 114) (Figure 3.2 on page 91). For each direct comparison, an R script with separate limma functions were executed, using the same argument values for background correction, normalization etc. than for the SSHscreen analyses (see the R command with argument values on page 121). The output for the three analyses, were in the form of three top tables where the *logFC* column (M-values) of each analysis represented the ER1, ER2 and ER3 values respectively. As expected, these ER values correlated perfectly (with a correlation coefficient of 1) with the SSHscreen ER values after the *ER1* and *ER3* analyses respectively.

3.4.8. Reverse library analysis using limma

Limma was used to do a direct comparison of the treated (UT) versus the control (UD) samples for the reverse library, using the ER3 slides (slides 58 and 114). Since only ER3 slides were available for the reverse library, SSHscreen could not be used to do the analysis and accordingly no rarity/abundance information could be calculated. The GAL file and image analysis output files were manually edited to only include the reverse library probes. The same argument values than for the forward library, were used for the reverse library analysis (normexp background correction with an offset of 50, the up-weighting print-tip loess within-array normalization method and A-quantile between-array normalization). Figure 3.9 gives the MA-plots for the two slides, before normalization (a) and after normalization (b) with the up-weighting print-tip loess method (the cDNAs were given zero weight and the control spots double weight before fitting a separate loess curve for each print-tip group). With 75% as a prior guess of the proportion of differentially expressed genes in the reverse library, all the genes had a positive B-statistic and 30% (286 out of 960 genes) of the genes had an adjusted p-value smaller than 0.05. With 50% as a prior guess, 60% of the genes had a positive B-statistic and this didn't influence the proportion of genes having an adjusted p-value smaller than 0.05. The final limma top table (mentioned first) was uploaded to SSHdb.

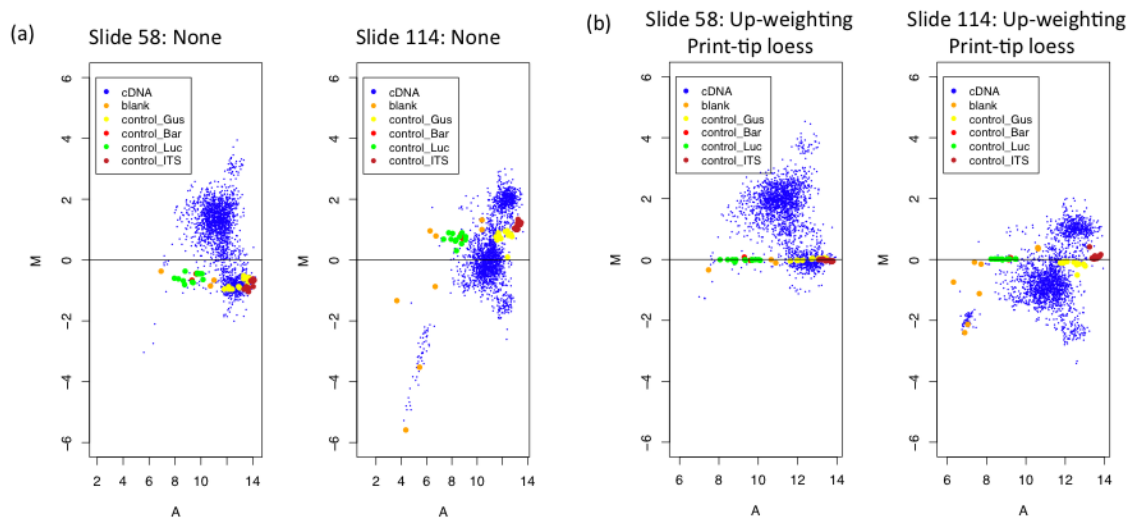


Figure 3.9: Limma output: M versus A plots before and after normalization with the up-weighting print-tip loess normalization method, of the pearl millet reverse library.

3.4.9. Management and annotation of the pearl millet SSH library using SSHdb

3.4.9.1. The forward library

Table 3.3 on the next page is a summary of the annotated ER3 up/down regulation top table exported from SSHdb for the forward library. In total 103 forward library clones were sequenced and SSHdb BLAST analyses classified these sequences into 33 redundant partner groups. Only one representative clone of each redundant partner group is included in Table 3.3. Of all the groups in Table 3.3, 24% (groups 1-8) are well-characterized defense response genes encoding: a wound-induced proteinase inhibitor (WIP1), a pathogenesis-related protein1 (PR1), a heat shock protein 70 (HSP70), a beta-glucosyltransferase (SAGTase), a *S*-adenosylmethionine decarboxylase (SAMDC), a multidrug and toxin extrusion transporter protein (MATE), a calcium-binding EF-hand protein (EF-hand), and an ethylene response element binding protein (EREBP).

Figure 3.10 gives the functional categorization of the pearl millet forward library (see strategy on page 94). 24% of the 33 redundant partner groups are annotated as defense-related, and 39 out of the 103 sequenced clones (including all redundant partners) are part of this group. This is a good result, since the forward pearl millet SSH library is enriched for genes up-regulated in response to treatment with elicitors and therefore the proportion of defense-related genes is expected to be high in this library. Further, 30% of the groups were categorized as metabolism genes, 9% as protein-metabolism genes, 9% as photosynthesis related genes and 27% other (Table 3.3 on the following page for details).

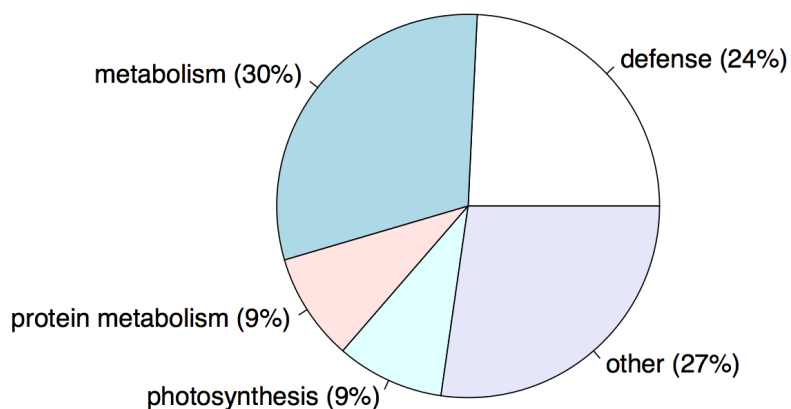


Figure 3.10: Functional categorization for the pearl millet forward SSH cDNA library. This pie chart summarizes the *function category* column in Table 3.3 on the following page.

Table 3.3: SSHdb output: Annotated ER3 up/down-regulation top table for pearl millet (forward library).

Pearl Millet - Forward library												
SSHscreen annotations					SSHdb annotations							
Group number	Representative clone ID	ER3*	adj.P.Val	B	invER2 (VecFree)	Length (VecFree)	BLAST Priority	BLAST AccNr	BLAST HitDef*	BLAST Eval	Number of redundant partners	Function category*
1	M56.10-D2	2,38	2,9E-03	6,14	-1,89	322	X	ACG30796	Bowman-Birk wound-induced proteinase inhibitor (WIP1)	9,0E-12	24	defense
2	M56.07-F1	2,07	5,1E-03	4,62	-1,14	243	X	ACJ62489	pathogenesis-related protein 1 (PR1) [Zea mays]	5,2E-20	3	defense
3	M56.02-F11	1,95	7,0E-03	4,09	-1,21	135	X	ACG43177	heat shock cognate 70 kDa protein 2 (HSP70) [Zea mays]	1,0E-36	2	defense
4	M56.08-B2	2,35	3,2E-03	6	-1,13	140	X	ACG39535	indole-3-acetate beta-glucosyltransferase (SAGTase) [Zea mays]	3,4E-11	1	defense
5	M56.06-H2	2,26	3,7E-03	5,51	-2,23	624	X	NP_001105713	S-adenosyl methionine decarboxylase 2 (SAMDC) [Zea mays]	6,2E-57	1	defense
6	M56.08-D7	2,13	4,2E-03	4,91	-0,06	590	X	NP_188997	MATE efflux family protein (MATE) [Arabidopsis thaliana]	1,3E-21	0	defense
7	M56.05-B12	2,04	5,4E-03	4,48	-0,95	139	X	ACG30702	calcium binding EF-hand protein (EF-hand) [Zea mays]	3,8E-15	0	defense
8	M56.10-C3	1,6	2,0E-02	2,77	-1,59	141	X	ACG38354	ethylene response element binding protein (EREBP) [Zea mays]	1,6E-08	0	defense
9	M56.06-H8	3,31	2,0E-05	11,9	-1,52	376	N	EU953063	glyceraldehyde-3-phosphate dehydrogenase [Zea mays]	1,5E-16	8	metabolism
10	M56.06-H9	0,76	2,7E-01	0,6	-2,14	298	X	ACG27752	plastocyanin [Zea mays]	3,3E-14	4	metabolism
11	M56.07-G5	2,07	5,0E-03	4,64	-2,82	408	X	ABK96988	glyceraldehyde-3-phosphate dehydrogenase [Urochloa decumbens]	8,4E-63	2	metabolism
12	M56.01-H12	1,37	4,1E-02	2,01	-0,89	219	X	AAM15963	C4 phosphoenolpyruvate carboxylase [Setaria italica]	3,2E-22	1	metabolism
13	M56.10-H1	2,38	2,9E-03	6,12	-0,5	391	X	ACG36802	indole-3-glycerol phosphate lyase [Zea mays]	4,7E-21	0	metabolism
14	M56.07-E2	2,31	3,5E-03	5,77	-2,31	287	X	ACA63883	vacuolar proton-inorganic pyrophosphatase [Hordeum vulgare]	1,3E-45	0	metabolism
15	M56.05-B6	2,18	4,0E-03	5,16	-2,12	648	X	P12783	Phosphoglycerate kinase, cytosolic	5,2E-49	0	metabolism
16	M56.01-D3	1,91	7,6E-03	3,93	-2,24	353	X	ACG34051	farnesyl pyrophosphate synthetase [Zea mays]	8,0E-61	0	metabolism
17	M56.07-A8	1,84	9,3E-03	3,64	-1,43	160	N	AB284983	bmST1 sulfotransferase [Bombyx mori]	4,8E-02	0	metabolism
18	M56.06-B6	1,38	4,1E-02	2,03	-1,01	135	X	ACG34631	alanine aminotransferase 2 [Zea mays]	2,3E-15	0	metabolism
19	M56.09-A5	0,57	4,1E-01	0,32	-2,14	367	X	NP_001105698	chlorophyll a/b-binding apoprotein CP26 precursor [Zea mays]	3,9E-34	5	photosynthesis
20	M56.01-G12	2,13	4,2E-03	4,92	-2,3	365	N	Z26595	triose phosphate/phosphate translocator [Zea mays]	1,3E-07	1	photosynthesis
21	M56.04-H6	1,09	1,1E-01	1,27	-0,34	191	N	EU959354	chlorophyll a-b binding protein 2 mRNA [Zea mays]	1,5E-05	1	photosynthesis
22	M56.01-F1	2,18	4,1E-03	5,12	-2,37	338	X	BAD53005	translation initiation factor [Oryza sativa]	1,5E-14	6	protein metabolism
23	M56.01-E5	2,19	4,0E-03	5,17	-2,05	479	X	ABB18390	ubiquitin-associated protein [Triticum aestivum]	5,8E-17	0	protein metabolism
24	M56.02-E10	1,43	3,4E-02	2,2	-1,54	135	X	BAB62890	aspartic proteinase 1 [Glycine max]	1,2E-16	0	protein metabolism
25	M56.07-A2	2,17	4,1E-03	5,07	-2,09	373	X	NP_566493	nodulin MtN3 family protein [Arabidopsis thaliana]	5,9E-17	5	other
26	M56.08-C7	4,14	3,2E-08	18,59	-1,58	405	N	DQ490950	16S ribosomal RNA gene [Festuca arundinacea]	0,0E+00	4	other
27	M56.06-H12	2,11	4,5E-03	4,81	-1,83	322	N	EU970028	hypothenical protein [Zea mays]	6,2E-34	1	other
28	M56.08-F10	1,75	1,2E-02	3,29	-1,62	348	N	BT054386	ZM_BF0017P07 mRNA [Zea mays]	1,2E-07	1	other
29	M56.06-F5	2,78	5,7E-04	8,36	-2,29	368	X	ACG29767	hypothenical protein [Zea mays]	5,7E-11	0	other
30	M56.03-F10	1,65	1,7E-02	2,93	-0,47	425	X	BAA04615	Tat binding protein [Oryza sativa]	2,9E-56	0	other
31	M56.07-A7	1,35	4,4E-02	1,96	0,1	246	N	EU972163	dynammin-related protein [Zea mays]	8,7E-11	0	other
32	M56.07-G12	0,8	2,5E-01	0,66	-2,46	184	N	EU965728	hypothenical protein [Zea mays]	1,5E-05	0	other
33	M56.07-E11	0,31	6,5E-01	0,08	-2,03	489	X	NP_567869	electron carrier [Arabidopsis thaliana]	1,1E-41	0	other

* Positive ER3 values indicate up-regulation in pearl millet plants treated with elicitors compared to the control treatment

\$ BLAST annotation selected from the top 10 blastX and blastN hits

The 'function category' column was added later

3.4.9.2. The reverse library

Table 3.4 on the next page is a summary of the annotated limma up/down regulation top table exported from SSHdb for the reverse library. In total 56 reverse library clones were sequenced and SSHdb BLAST analyses classified these sequences into 28 redundant partner groups. Only one representative clone of each redundant partner group is included in Table 3.4 on the following page.

Figure 3.11 gives the functional categorization of the pearl millet reverse library (see strategy on page 94). 43% of the 28 redundant partner groups are in the *other* category. 18% are annotated as photosynthesis genes and 9 out of the 56 sequenced clones (including all redundant partners) are part of this group. The chlorophyll a/b binding proteins are components of the light-harvesting complex (LHC) in photosystem I and II in the chloroplasts of cells, which are responsible for harvesting light energy during photosynthesis (Liu *et al.*, 2008). It is expected that genes taking part in these biological processes are down-regulated after treatment of plants with elicitors so that defense-related processes can be switched on. 18% was categorized as metabolism genes, 14% as defense genes and 7% as oxidative stress (Table 3.4 on the following page for details).

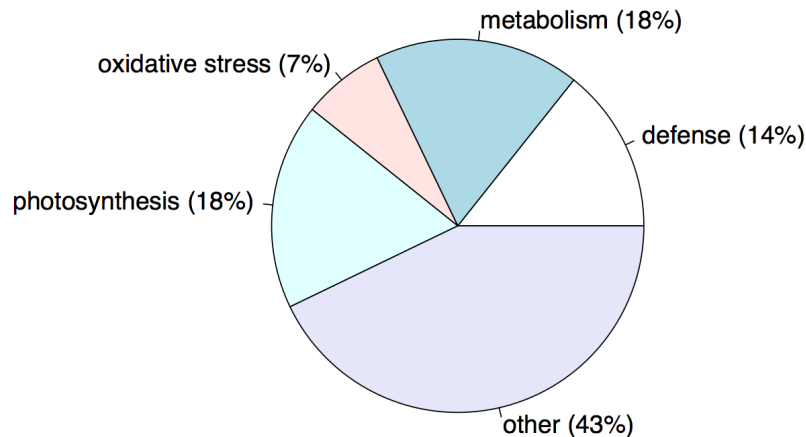


Figure 3.11: Functional categorization for the pearl millet reverse SSH cDNA library . This pie chart summarizes the *function category* column in Table 3.4 on the following page.

Table 3.4: SSHdb output: Annotated limma up/down-regulation top table for pearl millet (reverse library).

Pearl Millet - Reverse library												
Limma annotations					SSHdb annotations							
Group number	Repr. clone ID	logFC*	AveExpr	adj.P.Val	B	Length (VecFree)	BLAST Priority	BLAST AccNr	BLAST HitDef	BLAST Eval	Number of redundant partners	Function category#
1	M56_18-H8	1,56	11,19	4,60E-02	2,82	535	X	ACG47089	abscisic stress ripening protein 1 [Zea mays]	5,90E-18	3	defense
2	M56_11-F11	1,26	11,49	9,30E-02	1,92	492	X	ABF74001	disease resistance protein [Arabidopsis thaliana]	1,10E-36	0	defense
3	M56_13-C4	0,78	11,76	3,20E-01	0,86	248	X	AAAB21529	thionin [Hordeum jubatum]	1,50E-14	0	defense
4	M56_16-C11	0,45	11,36	5,40E-01	0,41	382	X	AAW48295	pore-forming toxin-like protein Hfr-2 [Triticum aestivum]	3,70E-13	0	defense
5	M56_16-E11	1,93	11,6	1,90E-02	4,21	645	X	AAK07429	beta-glucosidase [Musa acuminata]	6,30E-47	1	metabolism
6	M56_18-D6	0,81	11,58	3,10E-01	0,9	421	X	NP_001104897	pyruvate dehydrogenase (lipoamide) kinase1 [Zea mays]	6,50E-26	1	metabolism
7	M56_14-C1	2	10,64	1,60E-02	4,48	394	X	NP_001105368	phosphohexose isomerase1 [Zea mays]	1,10E-59	0	metabolism
8	M56_15-B3	1,62	12,26	4,20E-02	3,02	376	X	EU953063	glyceraldehyde-3-phosphate dehydrogenase	1,50E-16	0	metabolism
9	M56_14-B2	1,15	11,63	1,30E-01	1,62	479	X	NP_001105386	adenylosuccinate synthetase [Zea mays]	9,10E-78	0	metabolism
10	M56_19-H3	1,01	10,81	1,80E-01	1,3	207	X	P18123	Catalase isozyme 3	5,20E-20	1	oxidative stress
11	M56_15-G7	1,84	11,62	2,50E-02	3,81	470	X	ACG32994	peroxidase 54 [Zea mays]	1,00E-55	0	oxidative stress
12	M56_13-B12	1,21	10,77	7,10E-02	2,25	220	X	NP_001105545	light-harvesting chlorophyll a/b binding protein[Zea mays]	5,70E-06	4	photosynthesis
13	M56_11-F3	1,5	10,96	5,20E-02	2,63	416	X	ACG37564	photosystem II protein [Zea mays]	5,00E-47	0	photosynthesis
14	M56_13-E8	1,21	11,49	1,10E-01	1,76	216	X	ACG28217	chlorophyll a-b binding protein 2 [Zea mays]	1,10E-09	0	photosynthesis
15	M56_13-C10	0,87	11,44	2,70E-01	1	184	X	CAD89270	photosystem I reaction centre PSI-D subunit	2,00E-24	0	photosynthesis
16	M56_11-A3	0,83	10,83	2,90E-01	0,94	291	X	NP_001105698	chlorophyll a/b-binding appoprotein CP26	3,90E-34	0	photosynthesis
17	M56_14-H11	1,11	11,44	1,40E-01	1,53	524	X	ABA99659	retrotransposon protein	9,30E-17	7	other
18	M56_18-E3	2,02	11,72	1,50E-02	4,58	328	N	X57662	glycine-rich RNA-binding protein [Sorghum vulgare]	1,10E-07	3	other
19	M56_20-E8	2,04	10,7	1,50E-02	4,65	461	X	AAV25411	hypothetical protein [Arabidopsis thaliana]	6,60E-15	2	other
20	M56_13-D8	1,54	10,5	4,70E-02	2,75	395	X	BT042103	ZM_BFB0133G19 mRNA [Zea mays]	4,10E-20	2	other
21	M56_13-G2	1,53	11,22	4,90E-02	2,7	253	X	NP_001057513	Os06g0320500 [Oryza sativa]	4,50E-36	2	other
22	M56_16-D10	2	10,62	1,50E-02	4,5	453	X	ACG25125	hypothetical protein [Zea mays]	1,20E-08	1	other
23	M56_18-H6	1,93	10,7	1,90E-02	4,18	220	X	AY596553	SCUTLR1058C02 [Saccharum officinarum]	1,50E-03	1	other
24	M56_12-A3	2	10,7	1,60E-02	4,47	536	X	EEC83811	hypothetical protein [Oryza sativa]	7,20E-17	0	other
25	M56_15-G10	1,71	10,94	3,40E-02	3,32	237	X	P93371	Actin-93	4,30E-33	0	other
26	M56_17-B8	1,49	10,8	5,40E-02	2,59	324	N	EU970028	hypothetical protein [Zea mays]	6,20E-34	0	other
27	M56_13-G7	1,04	10,55	1,70E-01	1,36	189	N	EU963925	hypothetical protein [Zea mays]	1,10E-15	0	other
28	M56_17-C5	0,66	11,2	4,20E-01	0,67	294	N	EU971788	370677 CP12-1 mRNA [Zea mays]	3,70E-01	0	other

* Positive logFC values indicate down-regulation in pearl millet plants treated with elicitors compared to the control treatment

\$ BLAST annotation selected from the top 10 blastX and blastN hits

The 'function category' column was added later

3.5. Discussion

This pearl millet case study was aimed at contributing to the identification of plant genes activated and deactivated in defense response against biotic stress. In parallel with this, the flexibility of the SSHscreen-SSHdb pipeline was illustrated and the effect of different SSHscreen argument values were examined.

The SSH technique enriches for differentially expressed genes and accordingly most of the genes in the forward library are expected to be up-regulated by the treatment. Therefore the assumption underlying loess normalization that most of the probes on the array are not differentially expressed, does not hold when working with SSH cDNA libraries. A good solution to this is to make use of non-differentially expressed control spots in the normalization strategy. This set of control spots should span the intensity range and exhibit a relatively constant expression level across biological samples. In this study it was not known which pearl millet 'housekeeping genes' would be expressed the same in all treatments, so four 'alien' genes not present in the pearl millet libraries were used. They were spiked into the Cy dye labeling mixes in equal amounts and also spotted as control spots on the array.

The preferred normalization strategy for a SSHscreen analysis, having a suitable set of control spots, is the up-weighting print-tip loess method where the cDNAs are given zero weight and the control spots double weight, when fitting a separate loess curve for each print-tip group (Figure 3.4 on page 98 (g) shows the MA-plot after up-weighting print-tip loess normalization). Control-spot loess normalization also gives satisfactory results, but does not take the print-tip effect into account (Figures 3.5 and 3.3). Up-weighting global loess normalization is generally acceptable, but should not be used for normalization of this data set due to the too small value for the smoothing *span* parameter when the loess curve is fitted through the whole set of unequally spaced control spots (Figure 3.4 on page 98 (h) shows the MA-plot after up-weighting global loess normalization).

Deciding on the number of statistically differentially expressed genes that should appear in the top table, can be based on different top table statistics. A B-statistic > 0 corresponds to a 50-50 chance that a gene is differentially expressed. However, calculating the B-statistic, requires a prior guess specifying the proportion of expected differentially expressed genes p_j (Smyth, 2004). For an SSH cDNA library the proportion p_j should be much higher than 1% which is the default in limma. According to Table 3.1 on page 101, the number of genes included in the top table increases exponentially as p_j increases. The default proportion of expected differentially expressed genes in a SSHscreen analysis is 75%.

Separate limma analyses for the forward library were performed on the ER1 slides, ER2 slides and the ER3 slides. As expected, the M-values (calculated by limma) correlated

perfectly with the SSHscreen ER values after the *ER1* and *ER3* analyses respectively for the forward library. This confirms that the limma functions were correctly implemented within the SSHscreen code. Since only ER3 slides were available for the reverse library, only a limma analysis for the reverse library could be performed and hence no ER2 values could be calculated for the reverse library.

Using a contrast matrix, the ER1 analysis in SSHscreen indirectly compares the treated (UT) and untreated (UD) samples to get a measure of up/down regulation after treatment (M-value in the top table). This measure of up/down regulation for each gene is in fact an estimate of the ER3 value, calculated from the ER1 and ER2 values (after an *ER1* analysis) using the theoretical relation $ER3 = ER1 - ER2$. However, after comparison of the *ER1* analysis top table (M-values) with the *ER3* analysis up/down regulation top table (ER3 values), it was concluded that it was not possible to accurately predict the ER3 value from only ER1 and ER2 using this pearl millet data. This could be due to the fact that the ER1, ER2 and ER3 slides are different hybridizations from different slides. However, since the *ER3* analysis is available in SSHscreen 2.0.0, the *ER1* analysis is no longer necessary.

The forward pearl millet SSH library is enriched for genes up-regulated in pearl millet leaves at various time points following wounding or treatment with elicitors. Therefore, defense-related genes were expected in the screening of this library for truly up-regulated genes. Sequencing and annotation of 103 forward library clones resulted in 33 redundant partner groups of which 24% are well characterized defense response genes (Table 3.3). The main defense genes in the library are described below. Of the 174 sequenced clones in the library, 14% were annotated as Bowman-Birk wound-induced proteinase inhibitor (WIP1; Table 3.3, group 1). It is strongly wound-induced in contrast to the other members of the Bowman-Birk proteinase inhibitor family, which occur in seeds and are regulated during development (Rohrmeier and Lehle, 1993). According to Eckelkamp *et al.*, 1993, wounding in maize resulted in the systematic accumulation of a transcript coding for a Bowman-Birk trypsin inhibitor-related protein. Crampton *et al.*, 2009, used the same elicitor treated libraries (than analyzed in this chapter) enriched for genes regulated in response to biotic stress and reported on the induction of defense response pathways in pearl millet, in response to infection with the leaf rust fungus *Puccinia substriata*. According to their study, a MATE transporter protein and HSP70 were up-regulated when pearl millet was treated with salicylic acid (SA), but not methyl jasmonate (MeJA). Screening this elicitor treated pearl millet library, revealed a significantly up-regulated MATE efflux family protein (Table 3.3, group 6) and a heat shock cognate protein (HSP70; Table 3.3, group 3). A member of the MATE transporter family, EDS5, is an essential component of SA-dependent signaling for disease resistance in *Arabidopsis* (Nawrath *et al.*, 2002). Kanzaki *et al.*, 2003,

showed using Virus-induced gene silencing (VIGS) that HSP70 is an essential component of the plant defense signal transduction pathway. Ethylene response element binding protein (EREBP; Table 3.3, group 8) is a homeobox gene which encodes a transcription factor. In *Arabidopsis*, Büttner and Singh, 1997, reported that AtEBP is an ethylene inducible GCC box DNA-binding protein, that interacts with an ocs element binding protein, where ocs elements are a group of promoter sequences required for the expression of both pathogen genes in infected plants and plant defense genes. Pathogenesis-related (PR) genes have very different expression patterns in different plant species, but the induction of particularly PR1 by pathogens and chemicals in dicots have often been used as markers of SAR onset (Lawton *et al.*, 1996). PR1 was significantly up-regulated (Table 3.3, group 2) in the elicitor treated pearl millet library and 4 copies of this gene were sequenced. Lastly, S-adenosyl methionine decarboxylase (SAMDC; Table 3.3, group 5) is a key enzyme involved in the polyamine (PA) biosynthetic pathway. Wi *et al.*, 2006, showed that the overexpression of a carnation SAMDC gene generates a broad-spectrum tolerance to abiotic stresses in transgenic tobacco plants.

The reverse library is enriched for genes down-regulated in pearl millet leaves at various time points following wounding or treatment with elicitors. After screening of the library for truly down-regulated genes taking part in biological processes suppressed after treatment with elicitors so that defense-related processes can be switched on, sequencing and annotation of 56 reverse library clones resulted in 28 redundant partner groups. One of the largest categories (other than *other* with 43%) for the reverse library, was photosynthesis with 18% (Table 3.4). Further, 18% of the groups were categorized as metabolism genes, 14% as defense genes and 7% as oxidative stress. Botha *et al.*, 2006, isolated and analyzed ESTs from SSH libraries in order to study the response of wheat (*Triticum aestivum*) to RWA (Russian wheat aphid) feeding. It was proposed that expression of transcripts for photosynthetic activity enable resistant plants to overcome stress. However, in contrast, elicitors (small proteins secreted by RWA) are not recognized by receptor proteins in susceptible plants, a delayed activation of the systematic acquired resistance (SAR) causes that the plants have no time to activate the appropriate machinery for cell maintenance, which leads to the loss of energy production and cell death as a result of chlorophyll breakdown and a decrease in photosynthesis. In this pearl millet case study, almost 1/5 of the selected reverse library clones were significantly down-regulated photosynthesis genes and only 1/10 of the selected forward library clones were significantly up-regulated photosynthesis genes. This result suggests that the treated pearl millet plants probably fail to maintain photosynthetic function.

Chapter 4

Application of SSHscreen-SSHdb pipeline in *Arabidopsis*

4.1. Note

The *Arabidopsis* SSH libraries analyzed in this chapter, were prepared by S. Naidoo and A. McLeod, while in the MPPI research group, Department of Plant Science, FABI, UP. The microarray screening experiments were done by D. Theron at the ACGT Microarray Facility, UP. The complete data analysis was done by myself as part of this MSc dissertation.

4.2. Introduction

4.2.1. The plant pathosystem being studied

Arabidopsis thaliana (L.) Heynh, is a small flowering plant, part of the family *Brassicaceae*. The project to sequence the *A. thaliana* genome was completed in December 2000 (Bevan *et al.*, 2001). The genome encodes approximately 30 000 genes (www.Arabidopsis.org) and the Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for this model higher plant.

Ralstonia solanacearum is a gram-negative bacterium causing bacterial wilt in various tropical and sub-tropical plants (such as tomato, tobacco and eggplant) by penetrating the host through openings, such as wounds in the root system. After invading intercellular spaces of roots, it multiplies before invading xylem vessels and then produces exopolysaccharide (EPS), which leads to wilt of the infected plant. A distinguishing characteristic of *R. solanacearum* is its innate ability to adjust to the stressful and nutrient poor xylem environment (Hikichi *et al.*, 2007).

A. thaliana is a known host for the plant pathogen *R. solanacearum*. Being a model organism, *A. thaliana* provides a good starting point for the identification of candidate genes involved in resistance to pathogens, in this case *R. solanacearum*. Further characterization of

the identified genes can be used to improve resistance in susceptible hosts, such as *Eucalyptus*, via genetic engineering.

The research question posed in this study is to get a better understanding of the molecular mechanisms of plant resistance to the bacterial wilt pathogen, *R. solanacearum*. The resistant interaction between the *A. thaliana* ecotype Killean (Kil-0) and the *R. solanacearum* isolate BCCF 402 (CK) was used to prepare a SSH library by subtracting cDNA from infected Kil-0 plants and uninfected Kil-0 plants at various time-points. Figure 4.1 shows the differential response that was obtained between *A. thaliana* ecotypes Kil-0 and Be-0 in response to *R. solanacearum*.

Since *R. solanacearum* is considered to be a necrotrophic pathogen, it is expected that the most prominent pathway in resistance against it would be the jasmonic acid (JA) / ethylene (ET) signaling pathway. When considering a biotrophic pathogen on the other hand, the salicylic acid (SA) pathway is expected to be the most prominent (Thomma *et al.*, 1999).

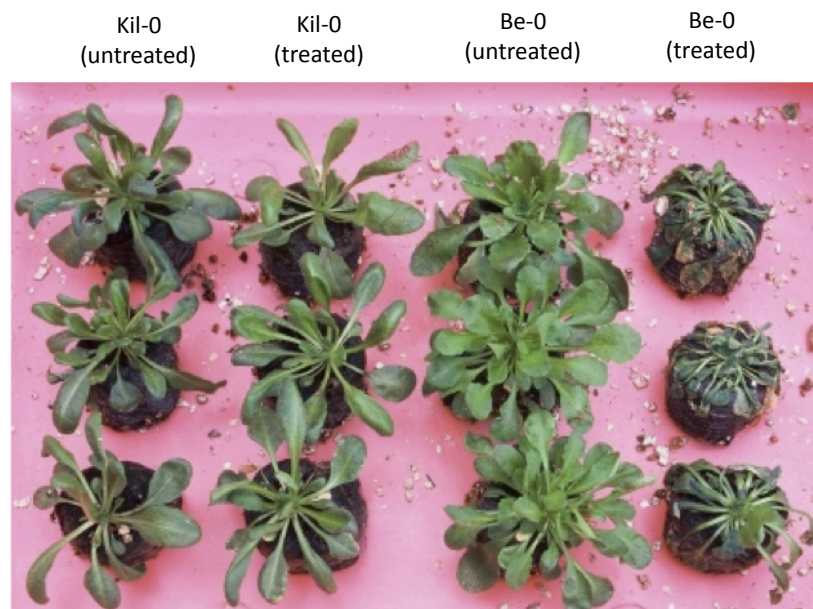


Figure 4.1: A differential response was obtained between *A. thaliana* ecotypes Kil-0 and Be-0 in response to *R. solanacearum*. From left to right: *A. thaliana* ecotype Kil-0 untreated plants; *A. thaliana* ecotype Be-0 untreated plants; *A. thaliana* ecotype Kil-0 treated plants (showing resistance to *R. solanacearum*); and *A. thaliana* ecotype Be-0 treated plants (showing susceptibility to *R. solanacearum*).

4.2.2. Introduction/overview of plant defense

Jones and Dangl, 2006, represents the plant immune system as a four phased 'zigzag' model (Figure 4.2). The basal defense system relies on the recognition of pathogen-associated molecular patterns (PAMPs) such as flagellin, lipopolysaccharides, peptidoglycans from bacteria and mannans of yeast by pattern recognition receptors (PRRs) (phase 1 in Figure 4.2). This results in PAMP-triggered immunity (PTI) which serves as an early warning for the activation of defense-related genes (phase 2 in Figure 4.2). Pathogens have evolved to overcome recognition by the PRR via type III secretion systems (TTSS) that releases effector molecules, known as avirulence genes (Avr), resulting in effector-triggered susceptibility (ETS) (phase 3 in Figure 4.2). As a countermeasure to these Avr proteins, plants have evolved to synthesize R-genes. These R-genes, of which the nucleotide-binding site plus leucine rich repeat (NB-LRR) protein is the most prominent class, interact with the Avr genes to activate effector-triggered immunity (ETI) (phase 4 in Figure 4.2).

Both the detection of PAMPs by PRRs and the interaction of R-Avr result in the activation of a signaling cascade that induces defense response genes. These pathways include salicylic acid (SA), jasmonic acid (JA), ethylene (ET) and abscisic acid (ABA), resulting in the production of pathogenesis related proteins and the initiation of various processes such as the hypersensitive response (HR) and systemic acquired resistance (SAR) to limit further invasion by the pathogen (Ryals *et al.*, 1996; Heath, 2000).

It has been shown that SA is the systemic signal required for the induction of SAR, but also that JA and ET both lead to a broad-spectrum, systemic resistance against microbial pathogens. The ET/JA pathway has an antagonistic effect on the SA pathway, indicating that control and maintenance of multiple pathways is necessary for the maintenance of disease resistance (Dong, 1998).

ABA is a plant hormone that regulates various signaling pathways in abiotic stress such as drought, cold stress as well as salinity and also has been shown to be involved in plant development (Thatcher *et al.*, 2005). It has been shown that ABA can have a positive or negative influence on the expression of plant defense genes depending on the particular plant-pathogen interaction.

Oxidative burst is the rapid release of reactive oxygen species (ROS), taking place in early plant defense mechanisms. ROS is mainly superoxide (O_2^-) and hydrogen peroxide (H_2O_2). According to Bolwell, 1999, ROS has numerous roles, including direct killing of the pathogen, involvement in structural changes in the cell wall, the induction of defense gene expression as well as promotion of programmed cell death (PCD) during hypersensitive response (HR).

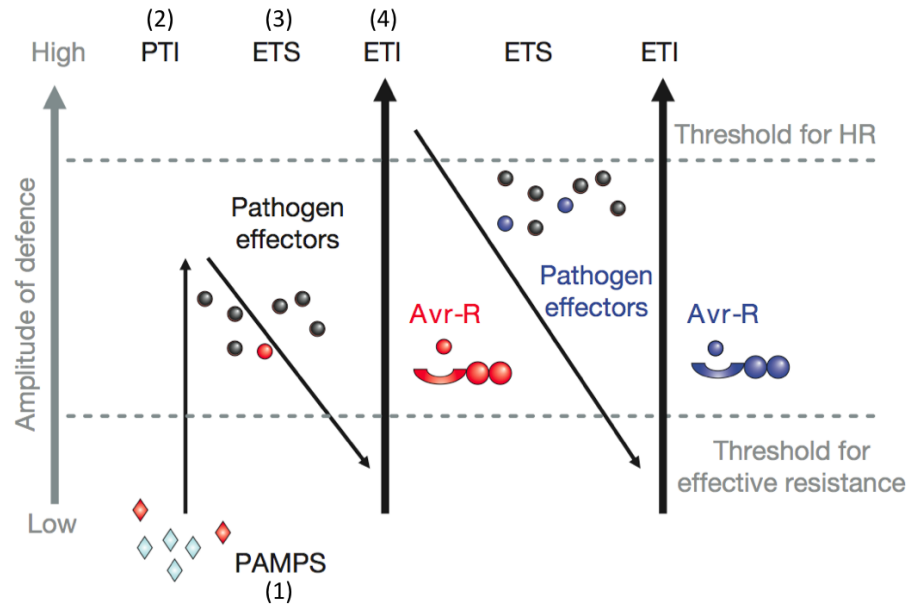


Figure 4.2: A model giving an overview of the plant immune system and illustrating the quantitative output thereof (Jones and Dangl, 2006). The ultimate amplitude of disease resistance or susceptibility is proportional to $PTI - ETS + ETI$. In phase 1, plants detect microbial/pathogen-associated molecular patterns (MAMPs/PAMPs, red diamonds) via pattern recognition receptors (PRRs) to trigger PAMP-triggered immunity (PTI). In phase 2, successful pathogens deliver effectors that interfere with PTI, or otherwise enable pathogen nutrition and dispersal, resulting in effector-triggered susceptibility (ETS). In phase 3, one effector (indicated in red) is recognized by a NB-LRR (nucleotide-binding site plus leucine rich repeat) protein, activating effector-triggered immunity (ETI), an amplified version of PTI that often passes a threshold for induction of hypersensitive cell death (HR). In phase 4, pathogen isolates are selected that have lost the red effector, and perhaps gained new effectors through horizontal gene flow (in blue) - these can help pathogens to suppress ETI. Selection favours new plant NB-LRR alleles that can recognize one of the newly acquired effectors, resulting in ETI.

4.3. Methods

4.3.1. Construction of cDNA library using SSH

The Molecular Plant-Pathogen Interactions (MPPI) group at the University of Pretoria (UP) constructed an SSH library for *Arabidopsis* (S. Naidoo and A. McLeod, unpublished). The resistant interaction between the *Arabidopsis* ecotype Killean (Kil-0) and the *R. solanearum* isolate BCCF 402 (CK) was used to prepare a SSH library by subtracting infected Kil-0 plants and uninfected Kil-0 plants at various time-points. A forward and a reverse

library were constructed, enriching for up- and down-regulated genes respectively. All together, 2304 cDNAs were cloned using SSH.

4.3.2. Screening SSH library on microarray

4.3.2.1. Slide layout and probes

In order to do SSH screening, these libraries were spotted on 8 cDNA microarray slides. Each slide consists of 24 blocks (print-tip groups), each with 7 rows and 32 columns of spots. On each slide, the forward library clones (ST_F) were spotted in the top half of each block (rows 1 - 3) and the reverse library clones (ST_R) were spotted in the bottom half of each block (rows 4 - 6). Row 7 in each block contains some control and blank spots. Unfortunately the DNA homologous to these control spots was not spiked into the labeling reactions and accordingly the control spots cannot be used for within-array normalization.

4.3.2.2. Experimental design and targets

Subtracted and un-subtracted cDNA samples that were used in the construction the *Arabidopsis* SSH libraries (ST_F , ST_R , UD and UT) were prepared as Cy3- and Cy5-labeled targets and hybridized to the microarrays.

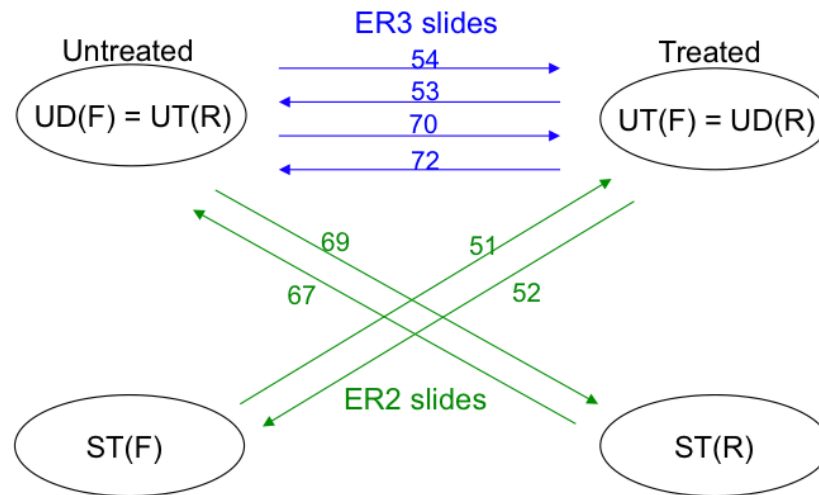


Figure 4.3: Experimental design of the *Arabidopsis* microarray experiment. Each arrow represents a microarray and the labeled ovals represent RNA samples. The RNA sample to which the arrow points are labeled with Cy5 dye (red) and the sample at the base of the arrow is labeled with Cy3 dye (green). F indicates the forward library and R the reverse library. For the forward library: $ER3 = \log_2(UT_F/UD_F)$ and $ER2 = \log_2(ST_F/UT_F)$. For the reverse library: $ER3 = \log_2(UT_R/UD_R)$ and $ER2 = \log_2(ST_R/UT_R)$.

Figure 4.3 gives the experimental design of the *Arabidopsis* SSH microarray data. Each arrow (each representing a microarray slide) connects the two cDNA samples that were hybridized to that slide. ER2 slides and ER3 slides were available, allowing a 'ER3' analysis to be performed on each library. ER1 slides were not included in the study.

(a)

◇	A	B	C	D
1	SpotType	Name	ID	Color
2	cDNA_F	cDNA*_F	*	blue
3	cDNA_R	cDNA*_R	*	yellow
4	blank	BLANK	*	orange
5	control_Sp	control_Sp*	*	green
6	control	control_*	*	red
7	ScoreCard	ScoreCard_*	*	purple

(b)

◇	A	B	C	D	E	F
1	SlideNumber	FileName	Cy3_F	Cy5_F	Cy3_R	Cy5_R
2		51 F_51_ST_UT.gpr	ST	UT		
3		52 F_52_UT_ST.gpr	UT	ST		
4		53 F_53_UT_UD.gpr	UT	UD	UD	UT
5		54 F_54_UD_UT.gpr	UD	UT	UT	UD
6		67 R_67_ST_UT.gpr			ST	UT
7		69 R_69_UT_ST.gpr			UT	ST
8		70 R_70_UT_UD.gpr	UD	UT	UT	UD
9		72 R_72_UD_UT.gpr	UT	UD	UD	UT

(c)

◇	A	B	C	D	E
1	Block	Row	Column	Name	ID
2		1	1	1 cDNA_Atha_AF1-A1_F	M64_AF1-A1_F
3		1	1	2 cDNA_Atha_AF3-A1_F	M64_AF3-A1_F
4		1	1	3 cDNA_Atha_AF1-B1_F	M64_AF1-B1_F
5		1	1	4 cDNA_Atha_AF3-B1_F	M64_AF3-B1_F
6		1	1	5 cDNA_Atha_AF1-C1_F	M64_AF1-C1_F
7		1	1	6 cDNA_Atha_AF3-C1_F	M64_AF3-C1_F
8		1	1	7 cDNA_Atha_AF1-D1_F	M64_AF1-D1_F
9		1	1	8 cDNA_Atha_AF3-D1_F	M64_AF3-D1_F
10		1	1	9 cDNA_Atha_AF1-E1_F	M64_AF1-E1_F
11		1	1	10 cDNA_Atha_AF3-E1_F	M64_AF3-E1_F
12		1	1	11 cDNA_Atha_AF1-F1_F	M64_AF1-F1_F
13		1	1	12 cDNA_Atha_AF3-F1_F	M64_AF3-F1_F
14		1	1	13 cDNA_Atha_AF1-G1_F	M64_AF1-G1_F
15		1	1	14 cDNA_Atha_AF3-G1_F	M64_AF3-G1_F
16		1	1	15 cDNA_Atha_AF1-H1_F	M64_AF1-H1_F
17		1	1	16 cDNA_Atha_AF3-H1_F	M64_AF3-H1_F

Figure 4.4: *Arabidopsis* SSHscreen ER3 analysis input files: (a) is the spot types file, (b) is the targets file for the 'ER3' analysis when *library="both"* and (c) is the first 16 entries of the GAL file. Clones from the forward library are distinguished from the reverse library by a '_F' or a '_R' at the end of the geneIDs and gene names.

4.3.2.3. SSHscreen software analysis

GenePix was used to extract the dye intensity data of each spot on each slide into 8 GenePix Results files (.gpr files), one for each microarray slide. A spot types file and a

targets file were constructed for the ER3 analysis, and together with the image analysis output files (*.gpr* files) and GAL file, stored in one directory.

Figure 4.4 on the preceding page gives the SSHscreen input files (except for the *.gpr* files) for the ER3 analysis. Figure 4.4 (a) is the spot types file, Figure 4.4 (b) the targets file and Figure 4.4 (c) a part of the GAL file. From the *Name* and *ID* columns in the GAL (and *.gpr* files), it is clear that a specific naming convention was followed. For example M64_AF1-A1_F is a cDNA clone with gene-ID AF1-A1 from the forward library spotted with print run M64. Reverse library clones can be distinguished in that the names and IDs end with '_R'. Also, the spot types file distinguish between forward and reverse library clones using *cDNA_F* and *cDNA_R* as different spot types. Since both libraries are analyzed in one run, the targets file must have columns *Cy3_F*, *Cy5_F*, *Cy3_R* and *Cy5_R*. This is necessary since slides were *ST* is one of the targets (ER2 slides), are only useful for either a forward or a reverse library analysis (because *ST_F* and *ST_R* are totally different samples (libraries); Figure 4.3 on page 119). ER3 slides are useful for a forward and a reverse library analysis, since $UT_F = UD_R$ and $UD_F = UT_R$ (note that the Cy3 and Cy5 columns are switched for the two libraries).

The SSHscreen ER3 analysis, was performed by the following R command:

```
> SSHscreen(path=~data/Arabidopsis/ER3vsER2", source="genepix", negflags=0,
norm.plot=TRUE, mfrow=c(3,2), legend=TRUE, bc.method="normexp", offset=50,
wa.method="printtiploess", ba.method="Aquantile", weights=FALSE, irregular=TRUE,
ndups=2, spacing=1, spot.ave=FALSE, method="ER3", adjust="fdr", sort="B", cut-
off="none", library="both", proportion=0.5)
```

Each argument with all its possible options and detail, is described in the SSHscreen R documentation (provided in the appendix on page 160). Argument values that were different from the pearl millet description on page 92 are described in this paragraph. *mfrow=c(3,2)* specifies that the plot layout for the MA-plots should be 3×2 , i.e. 6 plots per output window (one plot for each microarray slide when *method='ER3'*; note that the forward and reverse library slides are plotted in separate devices); *wa.method="printtiploess"* indicates that the print-tip loess method should be used for within-array normalization; *weights=FALSE* doesn't allow the use of control spots during normalization (the up-weighting print-tip loess method is not used); *sort="B"* will sort the genes in the top table in descending order according to the B-statistic; *cutoff="none"* will include all clones in the top table (no cut-off value is set); *library="both"* specifies that both the forward and the reverse library analysis

should be done in one run; and $proportion=0.5$ indicates that the assumed proportion of differentially expressed genes in the library is 50%.

Output from the SSHscreen ER3 analysis included two top tables for the forward library and two for the reverse library. These top tables are stored in R objects: *tt.ud.F* (up/down-regulation, forward library), *tt.ar.F* (rare/abundance, forward library), *tt.ud.R* (up/down-regulation, reverse library) and *tt.ar.R* (rare/abundance, reverse library). These top tables were uploaded to SSHdb.

4.3.3. Sequencing

Selected clones from the *Arabidopsis* forward and reverse SSH libraries were sequenced using the T7 Promoter primer by Inqaba Biotec (SA) or Macrogen (USA). Currently, 260 sequences are available in FASTA format and were uploaded to SSHdb.

4.3.4. Management and annotation of clones in SSH library

4.3.4.1. SSHdb

All available FASTA sequences were uploaded to SSHdb, which classified each input sequence as part of a redundant partner group and stored the top 10 *BLASTX* and *BLASTN* hits of each group. For each redundant partner group, the *BLAST* results were viewed in SSHdb. When a hit with a good E-value provided a better or more complete description than the default hit selected by SSHdb, the priority annotation was changed.

SSHscreen final top tables, containing all the genes in the library and ranked in terms of significance, were uploaded to SSHdb. SSHdb linked the top table entry of each sequenced clone, to the priority annotation of the redundant partner group it belongs to. The up/down-regulation annotated top tables for the forward and reverse libraries were exported from SSHdb as tab delimited text files.

4.3.4.2. Selection of redundant partner groups for further analysis

The annotated top tables exported from SSHdb (tab delimited text files) were opened in MS Excel. For each top table, the rows corresponding to non-sequenced clones were deleted. This resulted in 135 forward library clones and 125 reverse library annotated clones. The next aim was to select only one representative clone for each redundant partner group. The representative clone for each group was selected based on the best ER3 value. The result was 85 forward library groups and 66 reverse library groups. For the purpose of drawing conclusions for this dissertation, the tables were sorted by ER3 value in descending order and all redundant partner groups with $ER3 < -0.7$ were deleted. Groups with hit definitions *unknown* or *hypothetical protein*, as well as hits from other species than plants were also

deleted. Further analyses were performed on the resulting 60 groups in the forward library and 40 groups in the reverse library.

In order to get the corresponding Atg number for each group, FASTA sequences of the representative clones were exported from SSHdb and imported to MADIBA (Law *et al.*, 2008), which used the TAIR database to link Atg numbers to the sequences.

The forward library selected groups are summarized in Table 4.1 on page 130 (only 44 groups are shown) and the reverse library groups in Table 4.3 on page 135.

4.3.4.3. Other bioinformatics tools

The Atg numbers of the forward and reverse libraries respectively were submitted to various online bioinformatics tools, to perform gene ontology enrichment analyses on each cluster. The following tools were used:

- MADIBA (www.bi.up.ac.za/MADIBA)
- EasyGO (www.bioinformatics.cau.edu.cn/easygo)
- TAIR bulk data retrieval tools (www.Arabidopsis.org/tools/bulk/go/index.jsp)

4.4. Results

4.4.1. SSHscreen ER3 analysis of both libraries

The aim with the SSH forward and reverse library construction, was to identify genes up- and down-regulated respectively in the resistant interaction between *A. thaliana* ecotype Kil-0 and *R. solanearum* isolate BCCF 402 (CK). The SSHscreen ER3 analysis was used to generate for each library, a top table sorted in terms of differential expression between infected and uninfected Kil-0 plants.

SSHscreen ER3 analyses of the forward and the reverse libraries were performed in one run, specifying *library="both"* in SSHscreen 2.0.0 (see the R code on page 121). For each library, separate MA-plots (Figure 4.5 on page 125), ER-plots (Figure 4.6 on page 126) and top tables were output.

For the ER3 slides, comparing UT and UD, there were four technical replicates of which two were dye-swaps: slides 53 and 54, and slides 70 and 72. For the ER2 slides, comparing UT and ST, there were two technical replicates of which slides 51 and 52 were a dye-swap for the forward library, and slides 67 and 69 were a dye-swap for the reverse library (Figure 4.3 on page 119).

Figure 4.5 on page 125 gives the MA-plots after print-tip loess within-array normalization. Since there was no set of non-differentially expressed control spots spanning the whole

intensity range, the up-weighting print-tip loess within-array normalization method could not be used and accordingly the *weights* argument in SSHscreen was set to FALSE (see the R code on page 121). Hence, the data may be over-normalized, in that the cloud of data points after normalization is centered around the $M=0$ axis in slides 53, 54, 70 and 72, and not predominantly above (as expected for the forward library in slides 53 and 70) or below (as expected for the reverse library in slides 53 and 70) the $M=0$ line. Slides 54 and 72 are dye-swaps of slides 53 and 70, and therefore the reverse behaviour is expected. In slides 53 and 70, $M = \log_2(\text{infected}/\text{uninfected})$, indicating that genes with $M > 0$ is up-regulated by *R. solanacearum* infection and genes with $M < 0$ down-regulated. Despite over-normalization, the statistically significant up/down-regulated genes in response to *R. solanacearum* infection could still be identified. The *proportion* argument in SSHscreen (the prior guess of the proportion of differentially expressed genes in the library, influencing only the B-statistic) was set to 0.5, instead of 0.75 as for previous SSHscreen analyses.

Figure 4.6 on page 126 gives the ER3 versus inverse ER2 plots for both libraries, allowing one to visually screen SSH cDNA library clones. Data points in different quadrants, indicated with different colours, are annotated as up-regulated and rare; up-regulated and abundant; down-regulated and rare; or down-regulated and abundant. Figure 4.6 (a) shows that most of the genes in the forward library were abundant in the un-subtracted bacterial wilt infected sample (inverse ER2 values > 0). Since the data is actually over-normalized due to the lack of control spots, about 50% of the genes appear to be up-regulated (ER3 values > 0) and 50% down-regulated (ER3 values < 0) by bacterial wilt treatment. Figure 4.6 (b) shows that most of the genes in the reverse library were rare (inverse ER2 values < 0). About 50% of the genes appear to be down-regulated (ER3 > 0 ; note that positive ER3 values indicates down-regulation in the reverse library) and 50% up-regulated (ER3 < 0). The 300 most significant genes for each library were marked.

Four top tables were generated by the SSHscreen ER3 analyses: *tt.ud.F*, *tt.ar.F*, *tt.ud.R* and *tt.ar.R*. The up/down regulation top tables were used for further study. Using an adjusted p-value (Benjamini & Hochberg's (1995) method) cut-off of 0.05, only 2 genes seemed to be significant in the forward library and 50 in the reverse library (out of 1153 genes in each library). However, a t-statistic cut-off of ± 2 or an un-adjusted p-value cut-off of 0.05 resulted in about 200 significant genes for each library. The top tables were uploaded to SSHdb.

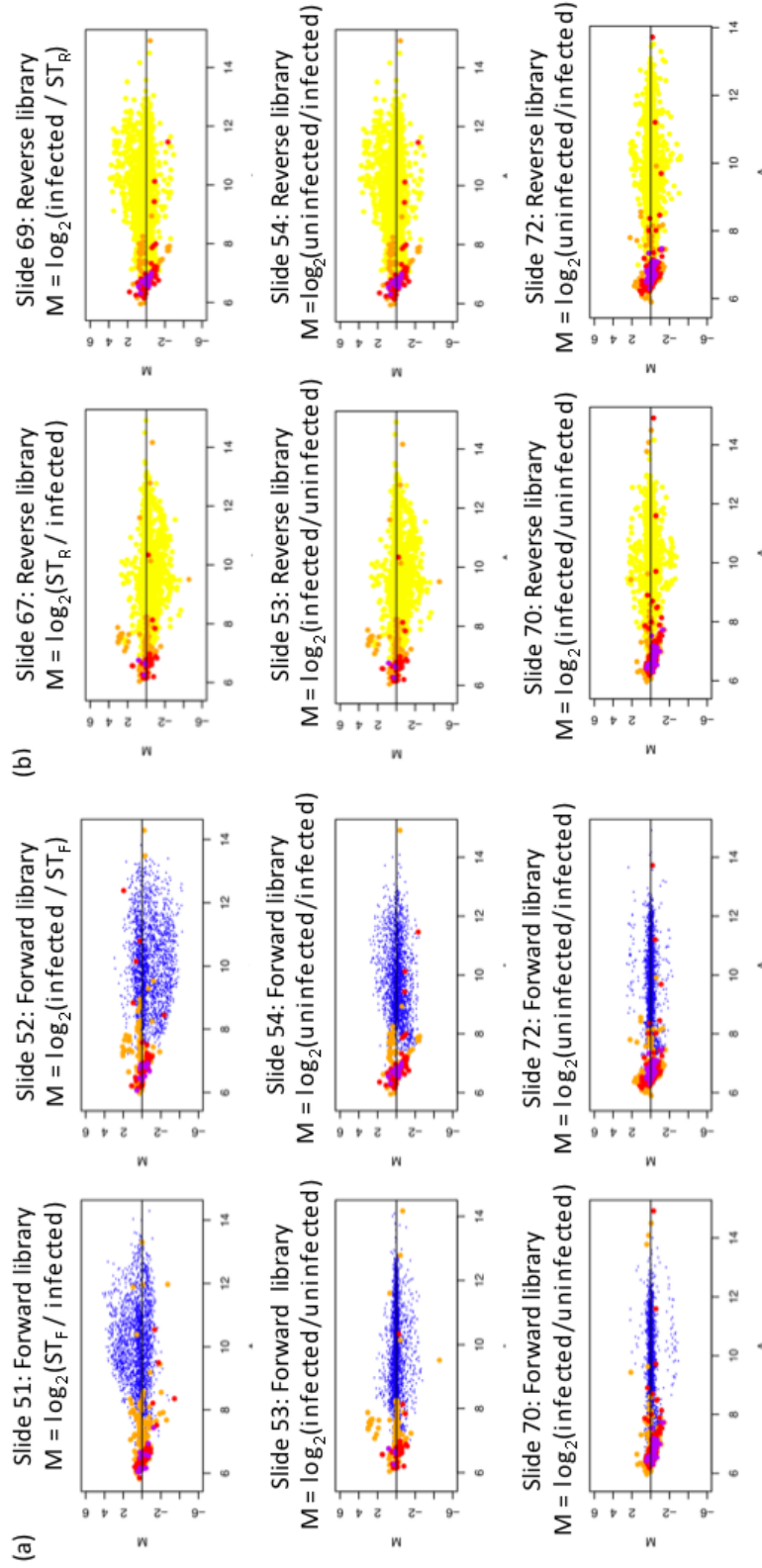


Figure 4.5: SSHscreen output: MA-plots after print-tip loess normalization for the *Arabidopsis* ER3 analysis of the forward and reverse libraries respectively. M versus A plots for the ER3 microarray slides (UT versus UD) and the ER2 microarray slides (UT versus ST). Blue data points represent forward library cDNAs (a) and yellow data points represent reverse library cDNAs (b). ST_F is the forward library subtracted tester enriched for up-regulated genes after *R. solanacearum* infection and ST_R is the reverse library subtracted tester enriched for down-regulated genes after *R. solanacearum* infection.

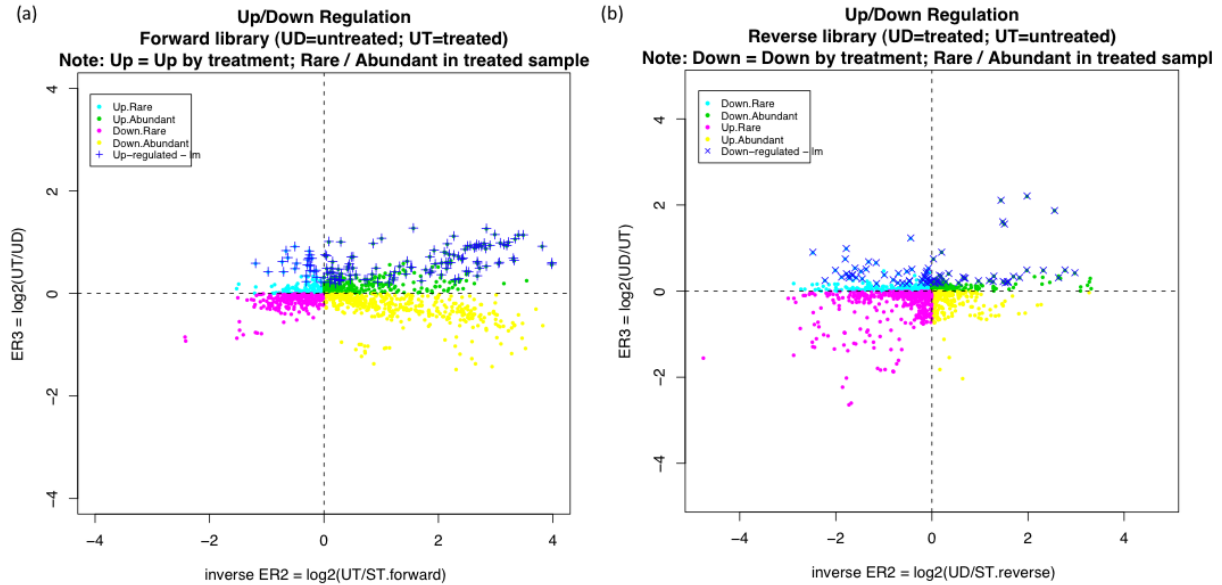


Figure 4.6: SSHscreen output: ER3 versus inverse ER2 plots for the *Arabidopsis* ER3 analysis of the forward and reverse libraries respectively. The ER3 versus inverse ER2 plots allow one to visually screen SSH cDNA library clones from the forward library (a) and the reverse library (b). ER3 is calculated as the log-2 ratio of the un-subtracted tester (bacterial wilt infected sample) divided by the un-subtracted driver (uninfected sample). Inverse ER2 is calculated as the log-2 ratio of the un-subtracted tester (bacterial wilt infected sample) divided by the forward/reverse library subtracted tester (SSH library enriched for up/down-regulated genes respectively). The top 300 up/down-regulated clones are marked with blue crosses. Data points are classified as: up-regulated/rare (Up.Rare) transcripts (quadrant 1; $ER3 > 0$ and $invER2 < 0$), up-regulated/abundant (Up.Abandant) transcripts (quadrant 2; $ER3 > 0$ and $invER2 > 0$), down-regulated/rare (Down.Rare) transcripts (quadrant 3; $ER3 < 0$ and $invER2 > 0$) and down-regulated/abundant (Down.Abandant) transcripts (quadrant 4; $ER3 < 0$ and $invER2 < 0$). The top 300 statistically significant up- and down-regulated clones are marked.

4.4.2. Biological annotation of the forward library

4.4.2.1. SSHdb output

Using SSHdb, an annotated up/down-regulation top table for the forward library was exported. This Table was edited and reduced to 60 redundant partner groups using MS Excel (see explanation of method on page 122). Table 4.1 on page 130 gives a summary of the first 30 groups and Table 4.2 on page 131 of the next 30 groups. Positive $ER3$ values in the forward library indicate up-regulation by *R. solanearum* infection and negative $ER3$ values down-regulation, since $ER3 = \log_2(UT/UD) = \log_2(infected/uninfected)$ for the forward library. Two extra columns were added: one giving the corresponding Atg number

for each gene (obtained from MADIBA) and the other giving GO-annotation information from the GO-term GO:0050896 (response to stimulus), since this GO-term was significantly enriched in the cluster of forward library genes (with a p-value of $1.6e - 10$) according to EasyGO.

4.4.2.2. EasyGO analysis

EasyGO (Zhou and Su, 2007) is a web-server to perform Gene Ontology based functional interpretation on groups of genes or GeneChip probe sets. Currently it supports 11 agronomical plants, 3 farm animals, and the model plant *A. thaliana*.

In this study, EasyGO was used to find enriched GO terms in the cluster of the 60 forward library Atg numbers. A Gene Ontology search on the aspect *biological process* gave a few significant GO terms. Of these, GO:0050896 (response to stimulus) had the smallest p-value, $1.6e-10$, indicating that this GO term did not occur in the cluster by chance. 50% (30/60) of the query list (input Atg numbers) mapped to this GO term.

Figure 4.7 on page 129 gives an overview of the GO terms in the next level, the level below GO:0050896 (response to stimulus), each with a p-value < 0.05 . GO:0006950 (response to stress) had the largest number of entries in the query list (19) as well as the smallest p-value ($8.8e - 07$). The other levels included, in order of significance, GO:0009628 (response to abiotic stimulus), GO:0042221 (response to chemical stimulus), GO:0009605 (response to external stimulus), GO:0009719 (response to endogenous stimulus) and GO:0009607 (response to biotic stimulus). Genes mapping to these GO-terms are marked in the last column of Table 4.1 on page 130 and Table 4.2 on page 131.

The 19 stress-related genes mapping to GO:0006950 are marked with an 's' in the last column of tables 4.1 and 4.2. These genes, in order of significance, encode: a 60S acidic ribosomal protein P0 (RPP0B) (group 11); a cell wall-modifying enzyme / hydrolase, acting on glycosyl bonds (TCH4) (group 37); a disease resistance RPP8-like protein 4 (group 19); the beta subunit of the chloroplast chaperonin 60 (group 53); squalene epoxidase 3 (SQE3) / an oxidoreductase (group 6); a member of heat shock protein 70 family (group 46); a distinct nitric oxide synthase (NOS1) that regulates growth and hormonal signaling in plants (group 60); a major latex-like protein (MLP31) (group 8); an oxygen-evolving enhancer protein (PSBP-1) (group 3); a putative ribonucleoprotein chloroplast precursor, RNA binding protein (group10); a secreted purple acid phosphate precursor (PRP1) (group 31); a allene oxide cyclase (group 40); an alpha-tubulin chain (group 5); an elongation factor 1-beta 2 (group 44); a vesicle-associated membrane protein (AtVAMP7C) (group 36); a small glycine-rich RNA binding protein (AtGRP7) (group 47); an endo chitinase-like protein which is essential for tolerance to heat, salt and drought stresses (group 7); a glutathione peroxidase

8 (GPX8) (group 50) and a chloroplast lipoxygenase required for wound-induced jasmonic acid accumulation (group 16).

4.4.2.3. TAIR bulk data retrieval tools

The Atg numbers were also submitted to the TAIR functional categorization tool (Rhee *et al.*, 2003). Figure 4.8 is the functional categorization for the GO aspect *biological process*. *Other metabolic processes* and *other cellular processes*, were the GOslim categories with the largest number of annotations to terms in the category (together 40% of the annotations mapped to these categories). The next two categories were *response to abiotic or biotic stimulus* (15%) and *response to stress* (14%). These percentages were calculated with the following formula:

$$\left(\frac{\# \text{ of annotations to terms in this GOslim category} \times 100}{\# \text{ of total annotations to terms in this ontology}} \right) = \%$$

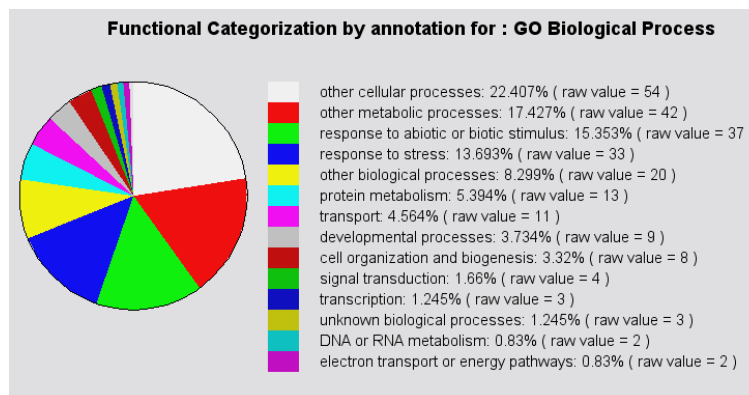


Figure 4.8: TAIR functional categorization by annotation for: GO biological process (*Arabidopsis* forward library analysis). Except for *other cellular processes* and *other metabolic processes*, the GOslim categories with the largest number of annotations to terms in the category were *response to abiotic or biotic stimulus* (15%) and *response to stress* (14%).

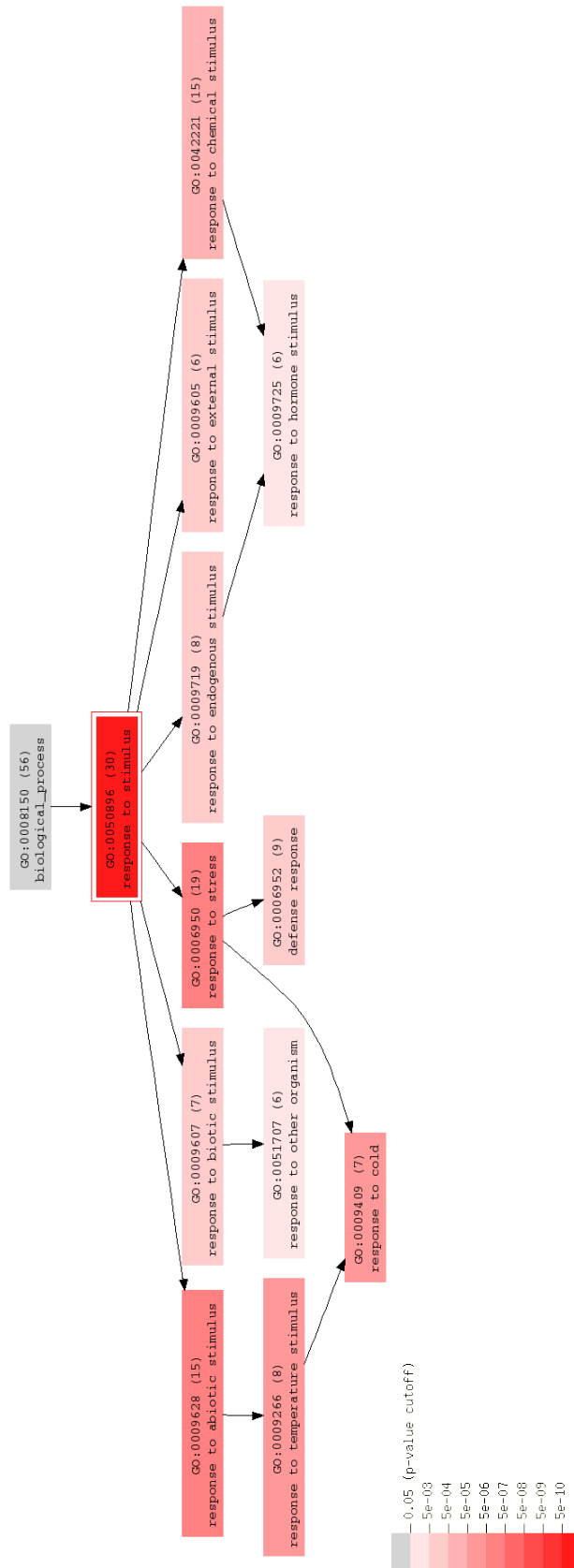


Figure 4.7: EasyGO Gene Ontology output, on the aspect *biological process* for GO:0050896 (*Arabidopsis* forward library analysis). The Atg numbers in tables 4.1 and 4.2 were submitted to EasyGO. The number of genes in each GO category is shown in brackets. A Gene Ontology search on the aspect *biological process* gave a few significant GO terms, of which GO:0050896 had the smallest p-value ($1.6e-10$), indicating that this GO term did not occur in the cluster by chance. The GO terms in the next level, each had a significantly small p-value. GO:0006950 (response to stress) had the largest number of entries in the query list (19) as well as the smallest p-value ($8.8e-07$). The other levels included (in order of significance) GO:0009628 (response to abiotic stimulus), GO:0042221 (response to chemical stimulus), GO:0009605 (response to external stimulus), GO:0009719 (response to endogenous stimulus) and GO:0009607 (response to biotic stimulus).

Table 4.1: Groups 1-30 of the annotated ER3 up/down-regulation top table for *Arabidopsis* (forward library).

Arabidopsis - Forward library										SSHscreen annotations				SSHdb annotations				Response to stimulus	
Group number	Representative clone ID	ER3	adj.P.Val	B	invER2	Length (VecFree)	BLAST Priority	BLAST AccNr	BLAST HitDef [§]	BLAST Eval	Number of redundant partners	Atg number ⁺	GO:0050896 [#]						
1	M64_AF1-G6_F	1.07	3.0E-01	2.38	3.34	559	X	AAB71969	similar to auxin-induced protein(aldol/keto reductase family); putative receptor protein kinase	1.3E-40	6	AT1G60710.1	c						
2	M64_AF3-F7_F	0.89	3.0E-01	2.72	2.63	450	X	AAK92807	PSBP-1(OXYGEN-EVOLVING ENHANCER);poly(U) binding	1.0E-18	6	AT2G26730.1	s/b						
3	M64_AF6-E10_F	1.08	3.0E-01	3.14	1.01	470	X	NP_172153	Lhcb2 protein	8.6E-102	2	AT1G06680.1	s/b						
4	M64_SF2-B11_F	0.66	3.0E-01	2.38	-0.28	704	X	AAD28771	putative tubulin alpha-2/alpha-4 chain	1.3E-70	2	AT2G05070.1	s/a						
5	M64_AF6-G1_F	-0.22	3.9E-01	1.49	1.68	152	X	AAQ81585	SQE3 (SQUALENE EPOXIDASE 3); oxidoreductase	2.1E-86	2	AT4G14960.2	s/c/ex/en						
6	M64_SF2-E7_F	1.01	3.0E-01	2.14	0.29	782	X	NP_568033	POM1 (POM-POM1); chitinase	1.2E-123	1	AT14G37760.1	s/c/a/en						
7	M64_SF6-C3_F	0.92	3.0E-01	2.37	-0.51	425	X	NP_172076	MPLP3 (MPLP-LIKE PROTEIN 31)	3.3E-46	1	AT1G05850.1	s/b						
8	M64_AF6-E12_F	0.72	3.3E-01	1.92	2.35	749	X	NP_177241	PLD1ALPHA1(1);phospholipase D	2.6E-36	1	AT1G23130.1	c/en						
9	M64_AF1-B5_F	0.64	3.0E-01	2.65	3.33	701	X	NP_188194	60S acidic ribosomal protein P0 (RPP0B)	1.4E-133	1	AT3G15730.1	s/a						
10	M64_SF3-F4_F	0.21	4.9E-01	1.24	-0.28	726	X	AAIM66970	putative RNA-binding protein	4.4E-93	1	AT3G09200.2	s/a						
11	M64_AF5-G11_F	-0.17	4.4E-01	1.34	0.54	137	X	NP_001078125	ATEYA(EYES ABSENT HOMOLOG); tyrosine phosphatase	8.3E-18	1	AT2G35320.1	s/a						
12	M64_AF5-E12_F	-0.19	5.3E-01	1.16	0.84	361	X	NP_565803	KEU (KEULE); protein transporter	4.4E-59	1	AT2G35320.1	s/a						
13	M64_SF2-D9_F	-0.36	3.0E-01	2.01	0.33	736	X	NP_563905	LHCA4 (Photosystem I light harvesting complex gene 4)	5.8E-80	1	AT1G12360.1	s/a						
14	M64_AF4-G3_F	-0.4	3.0E-01	2.44	2.61	605	X	NP_190331	LHCA1; chlorophyll binding	7.9E-14	1	AT3G47470.1	s/a						
15	M64_AF3-A8_F	-0.45	3.5E-01	1.63	1.2	252	X	NP_191049	lipoxigenase AtLOX2	2.4E-28	1	AT3G54890.1	s/c/a/ex/b/en						
16	M64_SF3-B10_F	1.01	3.0E-01	2.48	0.08	755	X	AAI32689	basic helix-loop-helix (bHLH) family protein	6.7E-140	0	AT3G45140.1	s/c/a/ex/b/en						
17	M64_AF3-D1_F	0.77	3.4E-01	1.71	2.19	464	X	NP_564583	cytochrome b6f complex subunit (petM)	3.2E-78	0	AT1G51140.1	s						
18	M64_SF2-D3_F	0.7	3.3E-01	1.94	-0.63	461	X	NP_565623	disease resistance protein (CC-NBS-LRR class)	2.9E-59	0	AT2G26500.1	s						
19	M64_AF2-F9_F	0.67	3.0E-01	2.07	2.41	571	N	NM_124238	Serine carboxypeptidase-like 20	0.0E+00	0	AT5G48620.1	s						
20	M64_AF3-C6_F	0.61	3.0E-01	2.08	3.06	614	X	Q8L7B2	calmodulin-like calcium-binding protein	2.6E-118	0	AT4G12910.1	s						
21	M64_SF5-B1_F	0.42	3.4E-01	1.81	1.24	621	X	NP_181642	NCEAD4(NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 4)	8.0E-105	0	AT4G19170.1	s						
22	M64_AF2-F7_F	0.34	3.4E-01	1.72	2.68	628	X	ABW17197	alanine aminotransferase 2	2.6E-113	0	AT4G19170.1	s						
23	M64_AF4-D2_F	0.41	7.0E-01	0.88	1.58	587	X	NP_193652	zinc finger-like protein	2.4E-81	0	AT1G23310.2	s						
24	M64_AF4-F7_F	0.27	3.5E-01	1.62	0.75	425	X	3EXL_A	Pyruvate Dehydrogenase(E1p)	2.2E-42	0	AT2G04280.1	s						
25	M64_AF6-H10_F	0.18	5.5E-01	1.09	2.29	596	X	AAK68811	photosystem I subunit III precursor	1.0E-10	0	AT3G52800.1	c						
26	M64_SF3-A1_F	0.15	4.1E-01	1.42	1.03	349	X	CAB52747	phosphoribulokinase/uridine kinase-related	2.9E-50	0	AT1G31330.1	c						
27	M64_SF3-G1_F	0.13	8.9E-01	0.7	0.08	407	X	NP_001077855	aspartate aminotransferase Asp2	7.5E-43	0	AT1G80380.3	c						
28	M64_SF1-H2_F	0.08	9.0E-01	0.67	-0.11	262	N	NM_100157	ferredoxin-dependent glutamate synthase	1.1E-126	0	AT1G02780.1	c						
29	M64_AR1-A1_R	0.06	9.1E-01	0.66	0.02	223	X	AAW91546		8.5E-34	0	AT5G19550.1	c						
30	M64_SF6-C11_F	0.05	8.0E-01	0.75	0.48	539	X	BAE98673		1.1E-60	0	AT5G04140.2	a						

§ BLAST annotation selected from the top 10 blastX and blastN hits
+ Atg numbers retrieved from MADIBA (www.bi.up.ac.za/MADIBA)

s=stress; c=chemical stimulus; a=abiotic stimulus; b=biotic stimulus; ex=external stimulus; en=endogenous stimulus

Table 4.2: Groups 31-60 of the annotated ER3 up/down-regulation top table for *Arabidopsis* (forward library).

Arabidopsis - Forward library										SSHdb annotations										SSHscreen annotations										Response to stimulus									
SSHscreen annotations										SSHdb annotations										SSHscreen annotations										Response to stimulus									
Group number	Representative clone ID	ER3	adj.P.Val	B	invER2	Length (VecFree)	BLAST Priority	BLAST AccNr	BLAST HitDef*	BLAST Eval	Number of redundant partners	Atg number*	Response to stimulus																										
31	M64_SF5-A10_F	0.05	8.9E-01	0.68	0.14	249	X	NP_180287	PAP1(PURPLE ACID PHOSPHATASE 1);acid phosphatase	6.3E-42	0	AT2G27190.1	s/ex																										
32	M64_SF6-D7_F	0.04	8.9E-01	0.68	0.51	681	X	NP_186896	SAMDC (S-ADENOSYLMETHIONINE DECARBOXYLASE)	6.7E-130	0	AT3G02470.3	c/en																										
33	M64_AF4-G6_F	0.02	9.6E-01	0.64	1.4	238	N	NM_124830	FQR1 (FLAVODOXIN-LIKE QUINONE REDUCTASE 1)	3.0E-37	0	AT5G54500.1	c/en																										
34	M64_AF1-E2_F	0.01	9.7E-01	0.64	0.38	211	X	NP_199101	lipin family protein	5.6E-29	0	AT5G42870.1																											
35	M64_SF1-G2_F	0.01	9.8E-01	0.64	-0.57	475	X	AAAM64443	nucellin-like protein	1.7E-58	0	AT4G33490.1																											
36	M64_SF2-H2_F	0.01	9.7E-01	0.64	0.55	353	N	NM_122141	ATVAMP714 (Vesicle-associated membrane protein 714)	0.0E+00	0	AT5G22360.1	s/a																										
37	M64_SF5-G1_F	0.01	9.8E-01	0.64	0.33	357	X	NP_200564	TCH4(TOUCH4); hydrolase; xyloglucosyl transferase	1.4E-65	0	AT5G57560.1	s/c/a/ex/en																										
38	M64_AF4-A1_F	0.01	9.6E-01	0.65	0.63	421	X	NP_566638	ATNHDI1(Na/H antiporter 1);sodium:hydrogen antiporter	6.1E-45	0	AT3G19490.1																											
39	M64_SF4-B9_F	0.01	9.8E-01	0.64	0.12	361	X	NP_850615	ARCS(accumulation & replication of chloroplast 5);GTPase	9.5E-46	0	AT3G19720.2																											
40	M64_SF4-D1_F	0	9.9E-01	0.63	-0.35	421	X	2DIO_A	Allene Oxide Cyclase 2	3.6E-61	0	AT3G25770.1	s/a/ex/b																										
41	M64_SF5-B2_F	-0.03	9.5E-01	0.65	-0.07	664	X	AAAM64533	contains similarity to plastid ribosomal protein L19	5.9E-51	0	AT5G47190.1																											
42	M64_SF1-A1_F	-0.04	9.1E-01	0.67	-0.1	434	X	BAF02012	sodium-dicarboxylate cotransporter-like	4.7E-69	0	AT5G47560.1																											
43	M64_SF3-D10_F	-0.04	8.7E-01	0.69	0.02	357	X	NP_187124	IAA16 (indoleacetic acid-induced 16); transcription factor	2.0E-27	0	AT3G04730.1	c/en																										
44	M64_SF3-B7_F	-0.05	8.9E-01	0.68	-0.52	294	X	NP_568375	elongation factor 1B alpha-subunit 2 (eEF1Balpha2)	1.5E-19	0	AT5G19510.1	s/b																										
45	M64_AF1-G11_F	-0.05	8.7E-01	0.69	0.91	384	X	NP_193113	CYP83A1 (CYTOCHROME P450 83A1); oxygen binding	3.4E-59	0	AT4G13770.1	a																										
46	M64_AF5-E9_F	-0.05	8.6E-01	0.7	0.74	234	N	NM_001125684	HSC70-1(heat shock cognate 70 kDa protein1);ATP binding	5.1E-83	0	AT5G02500.1	s/c/a/b																										
47	M64_AF6-F1_F	-0.05	8.4E-01	0.72	0.3	234	N	NM_179686	ATGRP7(COLD,CIRCADIAN RHYTHM,AND RNA BINDING 2)	1.4E-126	0	AT2G21660.2	s/c/a																										
48	M64_SF2-H9_F	-0.06	8.5E-01	0.71	-0.37	592	X	NP_565656	aldo/keto reductase family protein	5.5E-110	0	AT2G27680.1																											
49	M64_AF1-D2_F	-0.09	7.9E-01	0.76	1.18	244	X	AAAM65107	geranylgeranyl pyrophosphate synthase	1.7E-15	0	AT4G36810.1																											
50	M64_AF2-A2_F	-0.12	8.8E-01	0.69	0.39	164	X	NP_564813	glutathione peroxidase	2.4E-25	0	AT1G63460.1	s/c																										
51	M64_AF4-H7_F	-0.15	6.0E-01	0.99	1.01	396	X	AAAF02847	Similar to NAM protein	5.9E-32	0	AT1G56010.2	c/en																										
52	M64_AF3-C2_F	-0.16	7.8E-01	0.79	1.96	449	X	NP_200555	protein binding / protein transporter	2.2E-79	0	AT5G57460.1																											
53	M64_SF6-H9_F	-0.17	4.8E-01	1.25	-0.22	540	X	NP_175945	CPN60B(chaperonin 60 beta);ATP binding/protein binding	6.4E-93	0	AT1G55490.2	s/a/b																										
54	M64_AF6-A5_F	-0.2	5.7E-01	1.02	1.48	404	X	NP_199675	cyclin family protein	5.4E-33	0	AT5G48640.1																											
55	M64_AF3-B11_F	-0.24	3.9E-01	1.53	2.51	601	X	NP_196232	phox (PX) domain-containing protein	1.7E-106	0	AT5G06140.1	a/ex																										
56	M64_AF3-G3_F	-0.25	4.8E-01	1.25	0.68	473	X	NP_567965	CLB6(CHLOROPLAST BIOGENESIS6);diphosphate reductase	1.8E-68	0	AT4G34350.1	c																										
57	M64_AF1-C2_F	-0.33	5.7E-01	1.05	2.43	245	X	NP_172330	STGB(SIGMA FACTOR B);DNA-directed RNA polymerase	5.5E-38	0	AT1G08540.1	a																										
58	M64_AF4-F3_F	-0.39	3.0E-01	2.32	-0.68	118	N	AK226546	mRNA for ribulose biphosphate carboxylase like protein	6.1E-13	0	AT3G23590.1																											
59	M64_AF2-A1_F	-0.4	5.6E-01	1.08	1.21	631	X	NP_190187	ADF1 (ACTIN DEPOLYMERIZING FACTOR 1)	1.1E-64	0	AT3G46010.1																											
60	M64_AF6-B9_F	-0.41	3.8E-01	1.53	3.4	355	X	NP_190329	nitric-oxide synthase	2.1E-61	0	AT3G47450.2	s/c																										

\$ BLAST annotation selected from the top 10 blastX and blastN hits
+ Atg numbers retrieved from MADIBA (www.bi.up.ac.za/MADIBA)

s=stress; c=chemical stimulus; a=abiotic stimulus; b=biotic stimulus; ex=endogenous stimulus

4.4.2.4. MADIBA

MADIBA (Law *et al.*, 2008) GO analysis of the forward library Atg numbers was performed. For each GO term in the *biological process* ontology, MADIBA performed a hypothesis test, to test whether that GO term occurred in the cluster of Atg numbers by chance, taking into account the number of occurrences of the GO-term as well as the size of the cluster. A p-value was calculated for each GO term using a hypergeometric test, and GO-terms with FDR corrected p-values less than 0.05 were considered. Of these, a few interesting GO terms are listed below. Group numbers in brackets relate to Table 4.1 on page 130 and Table 4.2 on the previous page. These genes are described in the discussion on page 137.

- Regulation of salicylic acid metabolic process (GO:0010337): At1G05850 (AtCTL1 chitinase POM1; group 7)
- Jasmonic acid biosynthetic process (GO:0009695): At3G45140 (lipoxygenase AtLOX2; group 16) and At3G25770 (allene oxide cyclase AOC2; group 40)
- Response to jasmonic acid stimulus (GO:0009753): At4G37760 (squalene epoxidase SQE3, an oxidoreductase; group 6) and At3G45140 (lipoxygenase AtLOX2; group 16)
- Positive regulation of abscisic acid mediated signaling (GO:0009789): At3G15730 (phospholipase-D PLDalpha1; group 9)
- Response to wounding (GO:0009611): At4G37760 (squalene epoxidase SQE3, an oxidoreductase; group 6), and At3G45140 (lipoxygenase AtLOX2; group 16)
- Cellulose and pectin-containing cell wall organization and biogenesis (GO:0009664): At5G57560 (TCH4; group 37)
- Response to biotic stimulus (GO:0009607): At1G23130 (major latex-like protein MLP; group 8)

4.4.3. Biological annotation of the reverse library

4.4.3.1. SSHdb output

Using SSHdb, an annotated up/down-regulation top table for the reverse library was exported. This Table was edited and reduced to 40 redundant partner groups using MS Excel (see explanation of method on page 122). Table 4.3 on page 135 gives a summary of the 40 groups. Positive *ER3* values in the reverse library indicate down-regulation by *R. solanearum* infection, and negative *ER3* values up-regulation, since $ER3 = \log_2(UT/UD) = \log_2(uninfected/infected)$ for the reverse library. One extra column was added with the genes' corresponding Atg numbers extracted from MADIBA.

4.4.3.2. EasyGO analysis

EasyGO was used to find enriched GO terms in the cluster of the 40 reverse library Atg numbers. A Gene Ontology search on the aspect *biological process* gave a few significant GO terms, of which GO:0009987 (cellular process) had the smallest p-value ($5.4e-03$), indicating that this GO term did not occur in the cluster by chance. 67.5% (27/40) of the query list (input Atg numbers) mapped to this GO term. Figure 4.9 on the next page shows that GO:0044237 (cellular metabolic process) in the next level and GO:0015979 (photosynthesis) another level deeper had significantly small p-values (< 0.05), indicating that the cellular metabolic process, photosynthesis, was scaled down significantly after treatment of plants with bacterial wilt. The other significant GO-term was GO:0009628 (response to abiotic stimulus), however this term had a much larger p-value than the first.

The six photosynthesis related genes in Table 4.3, in order of significance, encode: a ferric chelate reductase (FRO2) (group 5), a chloroplast localized glyceraldehyde-3-phosphate dehydrogenase (group 2), Lhcb1.1: a component of the LHCIIb light harvesting complex associated with photosystem II (group 9), a mitochondrial ribosomal protein similar to ATP synthase delta chain (group 28), Lhcb2.2: the light-harvesting chlorophyll a/b-binding (LHC) proteins that constitute the antenna system of the photosynthetic apparatus (group 24), a chlorophyll a/b binding protein (CP29.1) similar to light-harvesting complex PSII (LHCB4.2) (group 15).

4.4.3.3. TAIR bulk data retrieval tools

The Atg numbers were also submitted to the TAIR functional categorization tool. Figure 4.10 on page 136 is the functional categorization for GO aspect *biological process* for the reverse library. 44% of the GO annotations mapped to *other cellular processes* and *other metabolic processes*. The GOslim categories with the 3rd and 4th largest number of annotations to terms in the category were *response to abiotic or biotic stimulus* (12%) and *other biological processes* (9%). These percentages were calculated with this formula:

$$\left(\frac{\# \text{ of annotations to terms in this GOslim category} \times 100}{\# \text{ of total annotations to terms in this ontology}} \right) = \%$$

Using the GO aspect, *cellular component*, 20% of the cellular component GO annotations mapped to *chloroplast*. This makes sense, since from the EasyGO analysis (results described

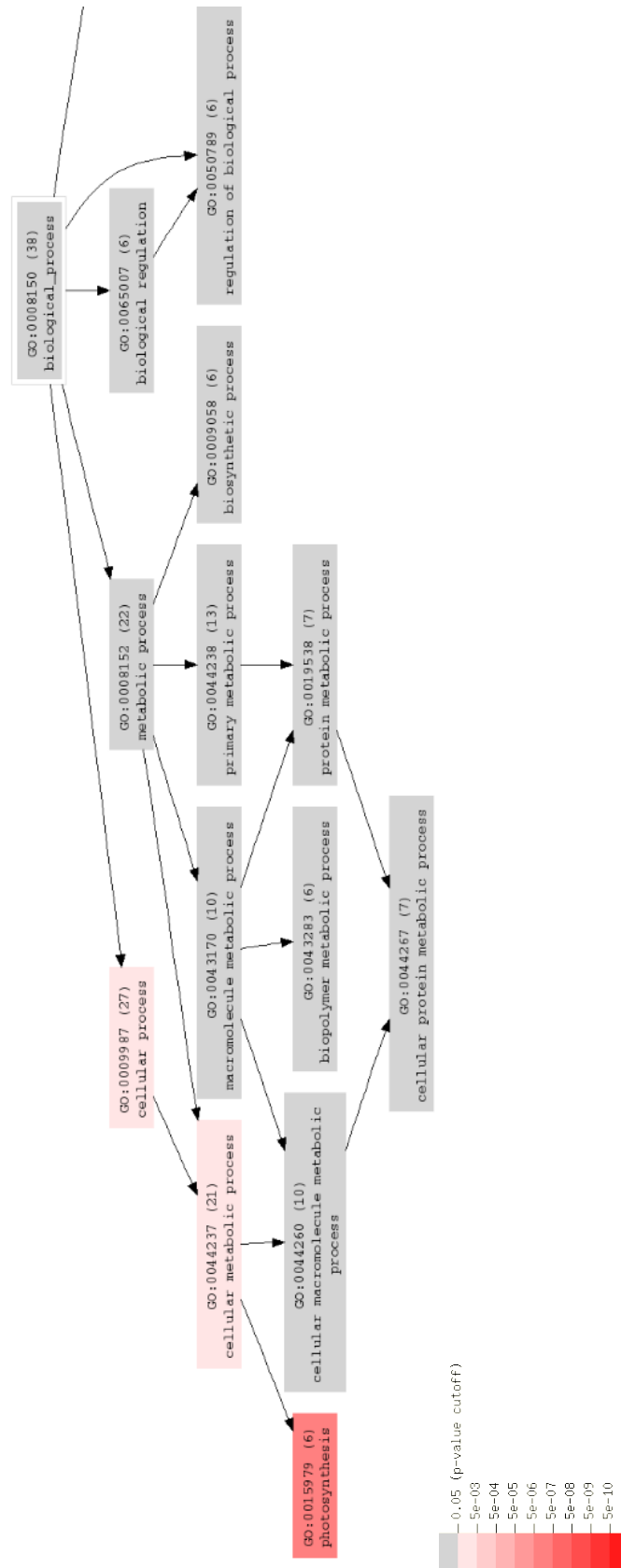


Figure 4.9: EasyGO Gene Ontology output, on the aspect *biological process* for GO:0008150 (*Arabidopsis* reverse library analysis). The graph extends to the right hand side (not shown; no significant terms). The Atg numbers in Table 4.3 was submitted to EasyGO. The number of genes in each GO category is shown in brackets. A Gene Ontology search on the aspect *biological process* gave a few significant GO terms, of which GO:0009987 (cellular process) had the smallest p-value ($5.4e - 03$), indicating that this GO term did not occur in the cluster by chance. 67.5% (27/40) of the query list (input Atg numbers) mapped to this GO term. GO:0044237 (cellular metabolic process) in the next level and GO:0015979 (photosynthesis) another level deeper had significantly small p-values (< 0.05), indicating that the cellular metabolic process, photosynthesis, was scaled down significantly after treatment of plants with bacterial wilt.

Table 4.3: Annotated ER3 up/down-regulation top table for *Arabidopsis* (reverse library).

SSHscreen annotations										SSHdb annotations									
Group number	Representative clone ID	ER3	adj.P.Val	B	invER2	Length (VecFree)	BLAST Priority	BLAST AccNr	BLAST HitDef [§]	BLAST Eval	Number of redundant partners	Atg number [*]							
1	M64_SR4-D7_R	-0.62	3.3E-01	1.77	0.98	605	X	NP_193652	NCE44 (NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 4)	2.6E-113	6	AT4G19170.1							
2	M64_SR1-A1_R	0.3	9.3E-01	0.67	2.68	361	X	AAD10210	glyceraldehyde 3-phosphate dehydrogenase B subunit	1.5E-20	3	AT1G42970.1							
3	M64_SR2-C7_R	2.21	2.8E-05	5.76	1.98	540	X	NP_565623	cytochrome b6f complex subunit (petM)	2.9E-59	2	AT2G26500.1							
4	M64_SR2-F2_R	-0.08	8.3E-01	0.89	-0.25	526	X	NP_192703	ATP synthase delta chain, chloroplast	9.0E-72	2	AT3G59650.1							
5	M64_SR1-G8_R	-0.25	6.0E-01	1.24	-2.83	622	X	BAA98161	PRO2-like protein; NADPH oxidase-like	1.6E-42	2	AT5G49740.1							
6	M64_SR1-G8_R	-0.34	2.8E-01	1.93	0.47	330	X	NP_001031954	TIF3B1(EUKARYOTIC TRANSLATION INITIATION FACTOR 3B)	1.1E-57	2	AT5G27640.2							
7	M64_SR4-A11_R	-0.6	2.4E-01	2.15	-0.03	489	X	NP_187514	alpha 1,4-glycosyltransferase family protein	3.1E-20	2	AT3G09020.1							
8	M64_SR4-E7_R	-0.63	1.7E-01	2.52	0	431	X	NP_198154	ATCYS2D2 (Arabidopsis thaliana cysteine synthase D2)	3.7E-66	2	AT5G28020.6							
9	M64_SR5-D11_R	1.56	2.0E-04	5.49	1.51	357	X	AAP44089	chlorophyll a/b binding protein	5.4E-25	1	AT1G29920.1							
10	M64_AR2-D12_R	0.31	3.6E-01	1.69	-1.72	338	X	NP_190187	ADF1 (ACTIN DEPOLYMERIZING FACTOR 1)	1.1E-64	1	AT3G46010.1							
11	M64_AR6-E10_R	0.19	4.7E-01	1.45	-0.71	223	X	AAA32813	plasma membrane proton pump H+ ATPase	1.8E-36	1	AT2G18960.1							
12	M64_AR6-H4_R	0.14	6.5E-01	1.16	-0.23	625	X	Q944G9	probable fructose-bisphosphate aldolase 2, chloroplastic	2.1E-113	1	AT4G38970.1							
13	M64_AR2-B8_R	0.06	9.1E-01	0.7	-0.55	224	X	NP_564844	protein kinase family protein	3.9E-15	1	AT1G65190.1							
14	M64_AR4-F6_R	-0.38	2.7E-01	1.97	-2.31	593	X	ABD36807	glutathione S-transferase	2.1E-72	1	AT1G78370.1							
15	M64_SR3-E10_R	-0.5	8.1E-01	0.92	0.26	331	X	AAM12979	chlorophyll a/b-binding protein CP29	5.8E-35	1	AT5G01530.1							
16	M64_AR6-A2_R	0.49	1.7E-01	2.5	2.77	456	X	BAD94215	omega-3 fatty acid desaturase, chloroplast precursor	6.8E-12	0	AT3G11170.1							
17	M64_AR6-H2_R	0.37	4.1E-01	1.61	0.16	533	X	NP_199330	cbby protein-related	1.7E-95	0	AT5G45170.1							
18	M64_SR6-F1_R	0.35	4.3E-01	1.56	0.19	869	X	NP_178112	heat shock protein binding / unfolded protein binding	1.1E-142	0	AT1G79940.2							
19	M64_AR6-G10_R	0.3	1.2E-01	2.91	-0.07	214	X	NP_194823	potassium channel tetramerisation domain-containing protein	1.7E-34	0	AT4G30940.1							
20	M64_AR5-G3_R	0.16	8.1E-01	0.91	1.02	237	X	2ISQ_A	O-Acetylserrine Sulphydrylase	5.0E-39	0	AT4G14880.2							
21	M64_SR3-B9_R	0.12	8.3E-01	0.87	-0.17	380	X	NP_172244	leucine-rich repeat transmembrane protein kinase	1.4E-65	0	AT1G07650.1							
22	M64_SR1-F9_R	0.12	8.0E-01	0.93	-0.01	387	N	NM_203056	zinc finger (C3HC4-type RING finger) family protein	1.1E-113	0	AT1G43850.1							
23	M64_AR1-G6_R	0.1	9.0E-01	0.73	-0.47	415	X	AAM61612	putative thioredoxin-like U5 small ribonucleoprotein	4.5E-32	0	AT5G08290.1							
24	M64_SR1-E2_R	0.09	9.1E-01	0.71	0.16	767	X	AAD28771	Lhcb2 protein	1.3E-70	0	AT2G05070.1							
25	M64_SR3-B7_R	0.08	8.9E-01	0.74	-0.01	622	X	ABO20848	cytochrome P450-like TBP protein	1.0E-53	0	AT1G16780.1							
26	M64_AR6-B6_R	0.05	9.3E-01	0.65	0.45	302	X	NP_195650	SHM4 (SERINE HYDROXYMETHYLTRANSFERASE 4)	4.3E-65	0	AT2G02170.2							
27	M64_AR4-G10_R	0.04	9.6E-01	0.61	-1.25	463	X	NP_191524	BRL1 (BRASSINOSTEROID INSENSITIVE 1); Kinase	9.5E-54	0	AT4G39400.1							
28	M64_AR4-G10_R	0.04	9.6E-01	0.61	-1.25	463	X	NP_191524	mitochondrial ribosomal protein L51/S25/C1-B8 family protein	5.4E-62	0	AT4G09650.1							
29	M64_SR2-H5_R	0.03	9.7E-01	0.59	0.05	444	X	AAM64661	cystatin-like protein	2.4E-20	0	AT5G47550.1							
30	M64_AR4-F7_R	0.03	8.2E-01	0.97	-1.49	68	N	NM_113510	ATPRX Q; antioxidant/ peroxiredoxin	1.5E-28	0	AT3G26060.1							
31	M64_AR1-G10_R	0.01	9.9E-01	0.58	3.1	403	X	NP_198682	APE1(ACCLIMATION OF PHOTOSYNTHESIS TO ENVIRONMENT)	9.3E-42	0	AT5G38660.1							
32	M64_AR3-H2_R	0.01	9.7E-01	0.59	-0.83	409	X	BAA19751	hydroxypyruvate reductase	6.1E-61	0	AT1G68010.1							
33	M64_AR1-G10_R	0.01	9.9E-01	0.58	-0.53	497	X	AAN31832	chloroplast translation elongation factor EF-Tu precursor	5.5E-76	0	AT4G20360.1							
34	M64_SR2-H8_R	-0.03	9.9E-01	0.58	0.02	354	X	NM_122141	ATVAMP714 (Vesicle-associated membrane protein 714)	0.0E+00	0	AT5G22360.1							
35	M64_AR6-B11_R	-0.04	9.5E-01	0.61	-0.29	726	X	Q8L7B2	Serine carboxypeptidase-like 20	2.6E-118	0	AT4G12910.1							
36	M64_AR2-A1_R	-0.05	9.7E-01	0.59	-2.16	376	X	AAF81356	Identical to wall-associated kinase 1	2.3E-47	0	AT1G21250.1							
37	M64_AR4-F1_R	-0.07	9.0E-01	0.72	-2.12	712	X	BAD93915	Carbonic anhydrase, chloroplast precursor	5.8E-95	0	AT3G01500.4							
38	M64_AR5-H6_R	-0.1	9.5E-01	0.63	-0.25	172	N	NM_129411	LTP1 (nonspecific lipid transfer protein 1)	9.9E-90	0	AT1G79680.1							
39	M64_AR1-H2_R	-0.14	8.8E-01	0.8	-1.54	410	X	NP_566016	ATCS (CITRATE SYNTHASE 4); citrate (SI)-synthase	1.2E-13	0	AT2G44350.2							
40	M64_AR1-H10_R	-0.23	8.1E-01	0.91	0.28	168	X	POC7R1	Pentatricopeptide repeat-containing protein	1.7E-23	0	AT1G47580.1							
40	M64_AR5-C9_R	-0.23	5.4E-01	1.36	-1.41	748	X	CAA74896	sigma factorB	3.4E-128	0	AT1G64860.1							

§ BLAST annotation selected from the top 10 blastX and blastN hits
+ Atg numbers retrieved from MADIBA (www.bi.up.ac.za/MADIBA)

above) several photosynthesis related genes were present in this cluster (from the reverse library).

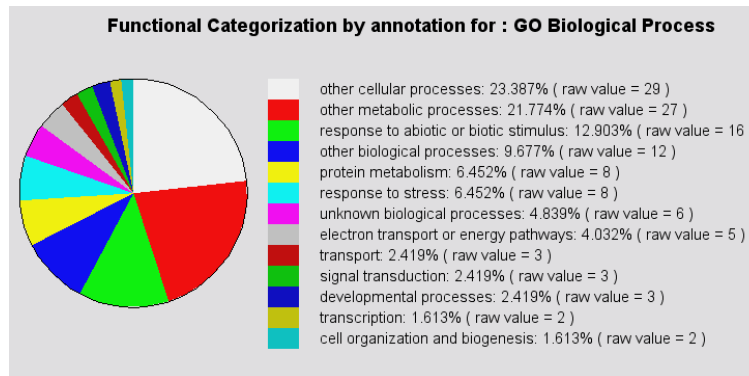


Figure 4.10: TAIR functional categorization by annotation for: GO biological process (*Arabidopsis* reverse library analysis). 44% of the GO annotations mapped to *other cellular processes* and *other metabolic processes*. The GOslim categories with the 3rd and 4th largest number of annotations to terms in the category were *response to abiotic or biotic stimulus* (12%) and *other biological processes* (9%).

4.4.3.4. MADIBA

MADIBA GO analysis of the reverse library Atg numbers was performed. For each term in the *biological process* ontology, MADIBA performed a hypothesis test, to test whether the GO term occurred in the cluster by chance, taking into account the number of occurrences of the GO-term as well as the size of the cluster. A p-value is calculated for each GO term using a hypergeometric test, and GO-terms with FDR corrected p-values less than 0.05 were considered. Of these, the five most significant GO terms ($10e - 06 < p\text{-value} < 10e - 04$) as well as two other terms are listed below (last two). Group numbers in brackets relate to Table 4.3 on the previous page. These genes are described in the discussion on the following page.

- Cell surface receptor linked signal transduction (GO:0007166); Response to salicylic acid stimulus (GO:0009751): At1G21250 (wall-associated kinase WAK1; group 36)
- Detection of brassinosteroid stimulus (GO:0009729); Response to UV_B (GO:0010224); Brassinosteroid homeostasis (GO:0010268): At4G39400 (BR1 kinase; group 27)
- Photosynthesis (GO:0015979): At4G09650 (mitochondrial ribosomal protein; group 28), At1G29920 (chlorophyll a/b-binding protein; group 9), At5G01530 (chlorophyll a/b-binding protein CP29; group 15) and At2G05070 (LHCB2 protein; group 24)

- Cysteine biosynthetic process from serine (GO:0006535): At4G14880 (O-Acetylserine Sulphydrylase; group 20)
- Proton transport (GO:0015992): At18960 (plasma membrane proton pump H⁺ ATPase; group 11)
- Photorespiration (GO:0006352): At1G68010 (hydroxypyruvate reductase ; group 32)
- Glycolysis (GO:0006096): At1G42970 (glyceraldehyde 3-phosphate dehydrogenase B sub-unit; group 2)

4.5. Discussion

The causal agent of bacterial wilt, *R. solanacearum*, is a devastating pathogen to numerous plant species. This case study aimed at contributing to the identification of signals that are activated and deactivated in response to invading pathogens, by identifying the genes that are differentially expressed in the resistant *A. thaliana* ecotype Kil-0.

Low-abundance genes that are slightly up-regulated in response to *R. solanacearum* may have a pivotal role in resistance, but generally escape detection due to the over-expression of other genes. However, due to a combination of the fact that the SSH technique includes a normalization step that enables the detection of low-abundance differentially expressed transcripts and the use of microarray technology for quantitative screening of the library, these low-abundance differentially expressed genes are likely to be identified.

SSHscreen was used to calculate ER2 and ER3 values for each gene in the forward and reverse the libraries. This allowed the visual screening of SSH cDNA library clones using ER-plots (Figure 4.6 on page 126), to get an idea of the quality of the library. Since SSH enriches for up/down-regulated genes, it is expected that the bulk of data points in the ER3 versus inverse ER2 plot lies above the ER3=0 line. However in this case where the non-differentially expressed control-spots could not be used for normalization, the data was over-normalized and the cloud of data points were centered around the ER3=0 axis after print-tip loess normalization. Nonetheless, the statistically significant up/down-regulated genes in response to *R. solanacearum* treatment could still be identified, however there were very few genes with good statistics (in general the adjusted p-values were not very low).

Interesting candidate genes were identified in the forward library, being up-regulated in response to treatment with *R. solanacearum*. However further characterization of these genes in gene function studies will be needed to determine whether these defenses are integral for resistance against *R. solanacearum* in Kil-0. A few of these up-regulated genes, that were annotated with one/more of the enriched GO-terms for the cluster of 60 forward library genes (according to the MADIBA software), are described below.

AtCTL1 is an endo chitinase-like protein, associated with tolerance to heat, salt and drought stresses (Kwon *et al.*, 2006) (Table 4.1, group 7). Chitinases are glycosyl hydrolases that catalyze the degradation of chitin, which is one of the most abundant biopolymers in nature. Because plants do not contain chitin it has been assumed that the role of the plant chitinases is in plant defense (attacking chitin), since chitin is a common constituent of cell walls in fungi and itself elicits plant defense. Plant chitinases have been shown to have functions in interaction with symbiotic bacteria and developmental processes. It also plays a role in stimulation of embryo and seed development. In resistance to *R. solanacearum*, a bacterium without chitin in cell walls, some chitinases act as PR proteins to directly inhibit pathogen growth in vivo by activating the SA-signaling pathway to induce SAR (Rogers and Ausubel, 1997). However, Zhong *et al.*, 2002 showed in their study that the AtCTL1 gene was expressed in all organs during normal plant growth and development, but it was not induced by wounding, salicylic acid, pectin fragments, or ethylene – indicating that AtCTL1 is a development-associated rather than a pathogenesis-related chitinase-like protein. According to TAIR GO analysis, AtCTL1 is annotated with the GO-term *regulation of salicylic acid metabolic process* (GO:0010337).

Phospholipase D alpha 1 (PLDalpha1) is a positive regulator of abscisic acid (ABA) mediated stomatal movements (Table 4.1, group 9). The hormone ABA decreases water loss by regulating opening and closing of stomata and (Mishra *et al.*, 2006) show that PLDalpha1 plays a significant role in a bifurcating signaling pathway that regulates plant water loss. PLDalpha1 also plays an important role in seed deterioration and aging in *Arabidopsis*. The GO-term *positive regulation of abscisic acid mediated signaling* (GO:0009789) mapped to PLDalpha1.

Squalene epoxidase 3 (SQE3) is an oxidoreductase (Table 4.1, group 6). Squalene epoxidase converts squalene into oxidosqualene, the precursor of all known angiosperm cyclic triterpenoids, which include membrane sterols, brassinosteroid phytohormones, saponins, other defense compounds, cuticular waxes, and non-steroidal triterpenoids (Rasbery *et al.*, 2007). The various SQE genes (SQE1 - SQE6) may be differently responsive to environmental stimuli, and certain SQE isozymes may be produced in conjunction with other triterpenoid biosynthetic enzymes to make specific products in response to biotic or abiotic challenges. In *Medicago truncatula*, Suzuki *et al.*, 2002, showed that one SQE is up-regulated upon methyl-jasmonate treatment, whereas a second SQE is unaffected. In this study, SQE3 is 2-fold up-regulated and is associated with GO-terms *response to wounding* (GO:0009611) and *response to jasmonic acid stimulus* (GO:0009753).

Major latex-like protein (MLP) (Table 4.1, group 8) is associated with fruit and flower

development in addition to plant pathogenesis responses (Ruperti *et al.*, 2002). MLP is linked to the GO-term *response to biotic stimulus* (GO:0009607).

Chloroplast lipoxygenase (AtLOX2) is required for wound-induced jasmonic acid accumulation in *Arabidopsis* (Table 4.1, group 16). Plant lipoxygenases are involved in the biosynthesis of lipid-derived signaling molecules. According to Bell *et al.*, 1995, LOX2 is required for the wound-induced synthesis of the plant growth regulator JA in leaves. AtLOX2 is annotated with GO-terms *response to wounding* (GO:0009611) and *response to jasmonic acid stimulus* (GO:0009753).

Allene oxide cyclase 2 (AOC2) (Table 4.2, group 40) is an enzyme that catalyzes one of the reactions of the JA biosynthetic pathway during leaf senescence in *Arabidopsis* (He *et al.*, 2002). AOC2 is linked to the GO-term *jasmonic acid biosynthetic process* (GO:0009695).

TCH4 is a cell wall-modifying enzyme, found to be rapidly up-regulated in response to environmental stimuli (Table 4.2, group 37). TCH4 is a member of the gene family encoding xyloglucan endotransglycosylase (XET)-related proteins (Antosiewicz *et al.*, 1997). This gene mapped to the GO-term *cellulose and pectin-containing cell wall organization and biogenesis* (GO:0009664).

Genes with positive ER3 values in the reverse library are down-regulated in response to the treatment with *R. solanacearum*. These genes are also of importance and the enriched *biological processes* GO-terms, gives an indication of which biological processes were down-scaled in order to switch on other defense-related processes. Many of these candidate genes are involved in the generation of ROS (reactive oxygen species) from reserves situated in the mitochondria and chloroplast (Table 4.3 on page 135). A few selected down-regulated genes are described below.

Brassinosteroid insensitive 1 (BRI1) (Table 4.3, group 27), is a steroidal plant hormone involved in numerous plant processes such as the promotion of cell expansion and cell elongation. Homeostasis of brassinosteroids is essential for normal growth and development in higher plants (Tanaka *et al.*, 2005). BRI1 is annotated under more with the 3 GO-terms *detection of brassinosteroid stimulus* (GO:0009729), *response to UV_B* (GO:0010224), *brassinosteroid homeostasis* (GO:0010268).

Lhcb1.1, a component of the LHCIIb light harvesting complex associated with photosystem II (Table 4.3, group 9), Lhcb2.2, the light-harvesting chlorophyll a/b-binding (LHC) proteins that constitute the antenna system of the photosynthetic apparatus (Table 4.3, group 24), a mitochondrial ribosomal protein similar to ATP synthase delta chain (Table 4.3, group 28), and CP29.1, a chlorophyll a/b binding protein similar to light-harvesting complex PSII (LHCB4.2) (Table 4.3, group 15) are the main genes involved in photosynthesis, mapping to the GO-term *photosynthesis* (GO:0015979). Examples where photosynthesis

is down scaled upon attack by insects and pathogens are given below. Truman *et al.*, 2006, examined the transcriptional dynamics of basal defense responses between *A. thaliana* and *Pseudomonas syringae*, a gram-negative bacterium with polar flagella. According to this study, down-regulated genes encoding photosynthetic function were highly over-represented (with a p-value of 1.17×10^{-23} from TIGR GO), indicating bacterial infection impacted strongly on photosynthesis. Zhu-Salzman *et al.*, 2004, aimed at understanding the transcriptional response of sorghum to infestation by the greenbug aphid (*Schizaphis graminum*). Greenbug-responsive transcript profiles were also compared with those after treatments by MeJA and SA. Photosynthesis-related genes were suppressed strongly by MeJA, and to a lesser extent by SA and aphids. Zhu-Salzman *et al.*, 2004, suggests that down-regulation of these genes allow energy reallocation to defense responses, with suppression of less important functions such as photosynthesis.

O-Acetylserine Sulphydrylase (OAS) (Table 4.3, group 20) catalyzes the final step of cysteine biosynthesis in plants. It occurs as several isoforms found in the cytosol, the plastids and the mitochondria (Jost *et al.*, 2002). OAS linked to the GO-term *cysteine biosynthetic process from serine* (GO:0006535).

WAK1, a cell wall-associated kinase (Table 4.3, group 36), is a member of the WAK family that links the plasma membrane to the extracellular matrix. Park *et al.*, 2001 suggests that the interaction of WAK1 with AtGRP3, a glycine-rich extracellular protein, occurs in a pathogenesis-related process in plants. Their study also showed co-expression of WAK1 and AtGRP3, and co-induction by SA treatment. Blanco *et al.*, 2005 emphasized the crucial role that SA plays in stress resistance in plants. They identified two groups of early SA-regulated genes of *Arabidopsis*. Group1 was classified as genes involved in cell protection (i.e. glycosyltransferases and glutathione S-transferases) and group 2 as genes involved in signal transduction (i.e. protein kinases and transcription factors). WAK1 was identified as a gene involved in signal transduction, part of group 2. As expected, WAK1 mapped under more the GO-terms *cell surface receptor linked signal transduction* (GO:0007166) and *response to salicylic acid stimulus* (GO:0009751). Since WAK1 is a positive regulator of the SA signaling pathway and WAK1 is down-regulated in response to treatment with *R. solanacearum*, one can speculate that resistance to *R. solanacearum* leads to the SA signaling pathway not being activated. As *R. solanacearum* is a necrotrophic pathogen, it is expected that the most prominent pathway in resistance against it would be the JA or ET signaling pathway. These pathways are known to have an antagonistic effect on the SA signaling pathway (Dong, 1998). Therefore it makes sense to find genes playing a role in the activation of the JA biosynthetic pathway up-regulated in response to treatment with

R. solanacearum: AtLOX2 (Table 4.1, group 16), AOC2 (Table 4.2, group 40) and Squalene epoxidase 3 (SQE3) (see discussion on page 139).

AHA1, a plasma membrane proton ATPase (Table 4.3, group 11) is an enzyme that catalyzes the decomposition of ATP into ADP and a free phosphate ion. This dephosphorylation reaction releases energy, which the enzyme harnesses to drive other chemical reactions. AHA1 mapped under more the enriched GO-term *proton transport* (GO:0015992). HPR, a hydroxypyruvate reductase (Table 4.3, group 32) mapped the GO-term *photorespiration* (GO:0006352). Glyceraldehyde-3-phosphate dehydrogenase B subunit (GAPB) (Table 4.3, group 2) is a chemical compound that occurs as an intermediate in several central metabolic pathways such as glycolysis and gluconeogenesis. It mapped the GO-term *glycolysis* (GO:0006096). The reason that these energy production genes are down-regulated in the resistant interaction between *A. thaliana* and *R. solanacearum* may be the down scaling of other biological processes so that energy can be reallocated to defense-related processes.

Chapter 5

Concluding discussion

Genetic modification, as opposed to breeding which is a slow and limited approach, is a powerful approach that can be used to improve stress resistance or drought tolerance in plants for example (Zhang *et al.*, 2009). However, using this approach, information about the genes involved in drought stress or a specific resistant interaction, is required in advance. For this reason, gene discovery of key stress response genes is very important. The first step in the process of finding novel genes, is usually to construct a cDNA or EST library. In order to identify and clone the relevant subsets of differentially expressed genes of interest, an enrichment technique, such as SSH or normalization can be used during library construction.

Norelli *et al.*, 2009 and Jayashree *et al.*, 2005, both used SSH to create EST libraries for gene discovery. All clones in these libraries were sequenced and after base calling and vector screening, sequences were clustered into contigs. Resulting sequences were putatively annotated after comparison with sequences in Genbank using BLASTN, BLASTX and TBLASTX searches. Sequences were also assigned to functional categories using BLAST2GO and manual annotation. Jayashree *et al.*, 2005, developed the Chickpea Root EST Database, which is a relational database system designed to provide on-line data mining of the sequence data and bioinformatic analysis results for the chickpea EST library. This is an illustration of the need for database management systems to store and automate sequence handling in projects like these. SSHSuite (Weckx *et al.*, 2004) is an integrated software package for analysis of large-scale SSH data from EST libraries. SSHHandler, which is the main part of the SSHSuite program, performs base calling, vector clipping, assembly, repeat masking, BLAST searches and parsing of the BLAST output files. SSHSuite runs only on a Linux workstation and requires various external software packages to be installed.

Following a slightly different strategy, Zhang *et al.*, 2009 and Morissette *et al.*, 2008, created SSH cDNA libraries and used quantitative screening to evaluate the quality of their cDNA libraries in order to select truly differentially expressed transcripts for sequencing. EST cluster analysis and the CAP3 program were respectively used to identify the 68% and 36% unique genes in these libraries. The unique sequences were searched against the NCBI

database with BLASTN and BLASTX for significant homologs. Sequences were placed in putative functional classes based on BLAST similarity and in the case of no significant BLAST similarity match, were identified as potential novel genes. Quantitative RT-PCR was used to verify the microarray results.

In this study, the SSHscreen-SSHdb pipeline was developed to automate most of the necessary steps mentioned in the previously described projects, including the screening, managing and annotation of clones from SSH cDNA libraries. The pipeline addresses the problem with SSH libraries that there is a lot of redundancy (i.e. the same gene fragment cloned multiple times), and for this reason it is not cost effective to sequence the whole library. SSHscreen allows clones to be prioritized for sequencing, whereafter a few clones are sequenced and sorted into redundant partner groups using SSHdb. Based on availability of funds etc., the next batch of clones can then be sequenced.

The aim of this pipeline differs from that of SSHSuite in that SSHSuite deals with large-scale sequence data from EST libraries and together with sequence identification, it also processes and manages sequence trace files. The SSHscreen-SSHdb pipeline on the other hand focuses on handling and annotating smaller subsets of cDNA sequences at a time. Base calling or assembly of EST reads is carried out prior to upload of sequences to SSHdb using standard software such as CLCBio. It provides an interactive user interface where groups of users can upload, view and store sequence data and microarray results. SSHscreen performs microarray data analyses and the resulting top tables can be uploaded and stored in SSHdb, which is a web-based tool with no other software requirements than a web-browser. After selecting and sequencing subsets of clones from the library, these sequences can also be uploaded to the database. SSHdb automatically performs vector clipping, clusters clones with similar sequences in redundant partner groups, performs BLASTN and BLASTX similarity searches against the Genbank database and finally parses the BLAST results, which is in XML format, so that it can be uploaded to the database. SSHdb links user selected BLAST annotations, from the top 10 BLAST hits stored in the database, to the corresponding limma/SSHscreen top table entries.

The SSHscreen-SSHdb pipeline was successfully developed and used for the screening of SSH cDNA libraries to identify differentially expressed genes potentially playing a role in stress response in cowpea, pearl millet and *Arabidopsis* ecotype Kil-0. The aim with the cowpea drought expression SSH library was to identify and isolate genes contributing to drought tolerance. With the pearl millet SSH library, the aim was to identify genes activated and deactivated against biotic stress. The aim with the *Arabidopsis* SSH library was to identify candidate genes up- and down-regulated in response to treatment with the bacterial wilt pathogen *Ralstonia solanacearum*.

For each library, a small number of microarray slides were sufficient for screening. SSHscreen was used to analyze spot intensity data, thereby quantitatively screening the clones in each SSH library. SSH libraries were developed over time by the MPPI lab, and the early SSH library construction and screening of pearl millet and *Arabidopsis* libraries were done without some of the data and control spots. For the pearl millet reverse library, only Enrichment Ratio 3 (ER3) slides indicating up/down-regulation were available and accordingly no rarity/abundance (ER2) conclusions could be inferred, and for the *Arabidopsis* library, no spiked-in controls were available for normalization. However, this dissertation demonstrates the flexibility of the SSHscreen-SSHdb pipeline by extracting useful data and conclusions from these libraries. The cowpea library is the most recent SSH library, including all necessary controls and hybridizations, so a full analysis was possible.

Clones from the forward and reverse subtraction libraries for the cowpea and *Arabidopsis* libraries were spotted on the same slides, and analyzed in one run, specifying the SSHscreen parameter library="both". For the pearl millet library, SSHscreen could only be used to analyze the forward library since reverse library probes were only printed on the UD vs. UT slides. For this reason, limma was used for the data analysis, and hence no rarity/abundance information is available for the pearl millet reverse library. The best normalization strategy, when screening SSH libraries, is to give full weight to a set of non-differentially expressed spiked-in control spots and no weight to the cDNA spots. This control spot based normalization method, called "up-weighting print-tip loess", was used for normalizing the cowpea and pearl millet libraries. Since there was no satisfactory set of control spots on the *Arabidopsis* microarray slides, print-tip loess normalization was used (without assigning weights). Although the *Arabidopsis* data was over-normalized, statistically significant genes could still be identified. Top tables of the statistically significant differentially expressed genes were generated. In the cowpea libraries, 58% and 28% of the clones were significantly up- and down-regulated (having an adjusted p-value < 0.05) in the forward and reverse libraries respectively. In the pearl millet libraries, these figures were 58% and 30% and in the *Arabidopsis* libraries, 18% for both. Top regulated genes were selected, sequenced and uploaded to SSHdb: 118 cowpea sequences, 174 pearl millet sequences and 262 *Arabidopsis* sequences. Of these clones, 33%, 37% and 55% were unique and the rest were redundant clones in the cowpea, pearl millet and *Arabidopsis* libraries respectively.

This pipeline facilitated the selection of six cowpea genes for further study. GST, THAU and LEA were statistically significantly up-regulated, and CHL and LTP were statistically significantly down-regulated as expected from the clones being in their respective libraries. Each of these clones had an SSHscreen ER3 value greater than 0.8 or less than -0.8, indicating more than 1.5 fold regulation with an adjusted p-value of less than 0.05, which was confirmed

by qPCR results. 26S escaped subtraction in the construction of the reverse subtraction library, as confirmed by the SSHscreen ER3 values as well as its qPCR result.

For the pearl millet forward library, defense-related genes were expected to be up-regulated and 24% of the redundant partner groups were well characterized defense response genes. For the pearl millet reverse library, almost 1/5 of the selected non-redundant clones were down-regulated photosynthesis genes, suggesting the down-scaling of less important processes such as photosynthesis, so that energy can be reallocated to defense-related processes.

Interesting candidate genes were identified in the *Arabidopsis* libraries. WAK1 is a positive regulator of the SA signaling pathway and since WAK1 is down-regulated in response to treatment with *R. solanacearum*, one can speculate that resistance to *R. solanacearum* leads to the SA signaling pathway not being activated. As *R. solanacearum* is a necrotrophic pathogen, it is expected that the most prominent pathway in resistance against it would be the JA or ET signaling pathways, which are known to have an antagonistic effect on the SA signaling pathway (Dong, 1998). Therefore it makes sense to find genes playing a role in the activation of the JA biosynthetic pathway up-regulated in response to treatment with *R. solanacearum*: (i) a lipoxygenase which is required for wound-induced JA accumulation in *Arabidopsis* (AtLOX2), (ii) Allene oxide cyclase 2, an enzyme that catalyzes one of the reactions of the JA biosynthetic pathway during leaf senescence (AOC2) and (iii) Squalene epoxidase 3, an oxidoreductase (SQE3) (see discussion on page 139). According to de Torres Zabala *et al.*, 2009, pathogen-modulated ABA signaling also rapidly antagonizes SA-mediated defenses. To complete the picture, Phospholipase D alpha 1 (PLDalpha1), which is up-regulated in this resistant reaction, is a positive regulator of abscisic acid (ABA) mediated stomatal movements.

There are several advantages to using the SSHscreen-SSHdb pipeline. SSH increases the probability of obtaining low-abundance differentially expressed cDNA fragments, which might provide important information in a gene discovery project. SSHscreen provides a quantitative method for screening SSH libraries to identify truly differentially expressed genes, since the SSH technique does not always yield only differentially expressed transcripts. Groups of researchers working together on an SSH project can register to view the same data in SSHdb. SSHdb users can be sure that their data is secure since users need to login and can only view projects that they are registered for.

The pipeline can be improved by writing a GUI version of SSHscreen in R, taking the user through a step-by-step analysis of the microarray data. This will ensure a more user-friendly version of the R package SSHscreen. A better option might be to pull SSHscreen into being part of the web-based tool, such as WebArray which is an online platform for microarray data analysis (Xia *et al.*, 2005), so that the user can select argument options from SSHscreen

using drop-down lists for example. A further objective is to add SSHscreen as an R “library”, as part of the BioConductor project (Gentleman *et al.*, 2004). The functionality of SSHdb can be expanded by also performing BLAST searches against other databases such as TAIR when working with *Arabidopsis* data, as well as the GO database to further annotate each clone in order to do functional classification of the clones in the library.

It is anticipated that this pipeline will make gene discovery projects easier for researchers, so that less time is spent performing individual BLAST searches and managing gene-lists in MS Excel, and more time deriving biological conclusions and getting insight while seeking to answer the biological question under study. The sequence database part of SSHdb can also be used in itself to manage and annotate clones from other transcriptome sequencing projects. The software is open source and easily accessible. SSHscreen can be downloaded from <http://microarray.up.ac.za/SSHscreen/>. SSHdb is available at <http://sshdb.bi.up.ac.za>.

Summary

A pipeline was developed for the quantitative screening and sequence management of clones from suppression subtractive hybridization (SSH) cDNA libraries. The pipeline is particularly useful for gene discovery in non-sequenced organisms, and was illustrated with SSH library data from pearl millet (*Pennisetum glaucum*) and cowpea (*Vigna unguiculata*) and *Arabidopsis* (*Arabidopsis thaliana*) ecotype Kil-0. The objective of each library was to identify stress-response genes.

cDNA microarrays provide a high-throughput screening method. Accordingly, these SSH libraries were amplified by PCR and spotted onto glass microarray slides. Subtracted and un-subtracted cDNA samples, that were used to construct the SSH libraries were prepared as Cy3- and Cy5-labeled targets and hybridized to the microarrays. The R package SSHscreen version 2.0.0, available from <http://microarray.up.ac.za/SSHscreen/>, was developed to analyze the resulting microarray data using limma (linear models for microarray data) functions. Commonly, loess normalization is used for within-slide normalization, however this is based on the assumption that most of the genes on the array are not differentially expressed. This is legitimate for most whole genome microarray experiments, however it is not appropriate when the array is constructed from an SSH library which is enriched for differentially expressed genes. Therefore, control spot-based normalization was used in the SSHscreen analysis. Empirical Bayes methods were employed to calculate the moderated t-statistic using functions from the limma package. This procedure in effect borrows information from the ensemble of genes to aid with inference about individual genes, taking advantage of the parallel structure whereby the same model is fitted to the data for each gene. In the *Arabidopsis*, pearl millet and cowpea forward libraries, 18%, 58% and 58% of the clones were identified as significantly up-regulated (adjusted p-value < 0.05) and in the reverse libraries, 18%, 30% and 28% significantly down-regulated, respectively.

SSHscreen analysis was used to assist in selection of clones for sequencing. The SSHscreen data output (ranked gene lists in terms of differential expression), as well as the selected sequences in FASTA format were uploaded to SSHdb. For the *Arabidopsis* library, 114 out of the 262 sequenced clones (55%) were identified as unique/non-redundant; and for the pearl millet and cowpea libraries respectively, 37% and 33% of the sequenced clones were

unique. SSHdb was developed as a web-based tool for sequence management and annotation of clones in SSH libraries and can freely be accessed at <http://sshdb.bi.up.ac.za>. BLAST analysis that was carried out when sequences were uploaded to SSHdb was used to combine clones with the same sequence into redundant partner groups, as well as identify putative annotations for each group.

Individual clones from the abovementioned SSH libraries were selected and an independent technique, quantitative PCR, was used to validate the microarray/SSHscreen results. The pipeline was applied successfully to *Arabidopsis*, pearl millet and cowpea SSH cDNA libraries. Interesting genes in each case were identified for further study.

Bibliography

- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. and Moreno, R. F. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 5013, 1651–1656.
- Agbicodo, E., Fatokun, C., Muranaka, S., Visser, R. and Linden van der, C. (2009) Breeding drought tolerant cowpea: constraints, accomplishments, and future prospects. *Euphytica* **167**, 3, 353–370.
- Alba, R., Fei, Z., Payton, P., Liu, Y., Moore, S. L., Debbie, P., Cohn, J., D’Ascenzo, M., Gordon, J. S., Rose, J. K. C., Martin, G., Tanksley, S. D., Bouzayen, M., Jahn, M. M. and Giovannoni, J. (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant Journal* **39**, 5, 697–714.
- Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (1997) *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. Garland Science/Taylor and Francis Group.
- Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.
- Antosiewicz, D., Purugganan, M., Polisensky, D. and Braam, J. S. (1997) Cellular localization of *Arabidopsis* xyloglucan endotransglycosylase-related proteins during development and after wind stimulation. *Plant Physiology* **115**, 4, 1319–1328.
- Arondel, V., Vergnolle, C., Cantrel, C. and Kader, J. (2000) Lipid transfer proteins are encoded by a small multigene family in *Arabidopsis thaliana*. *Plant Science* **157**, 1, 1–12.
- Bachem, C. W., Oomen, R. J. and Visser, R. G. (1998) Transcript Imaging with cDNA-AFLP: A Step-by-Step Protocol. *Plant Molecular Biology Reporter* **16**, 157–173.
- Bell, E., Creelman, R. and Mullet, J. (1995) A chloroplast lipoxygenase is required for wound-induced jasmonic acid accumulation in *Arabidopsis*. *The Proceedings of the National Academy of Sciences Online (US)* **92**, 19, 8675–8679.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**, 1, 289–300.
- Berger, D., Crampton, B., Hein, I. and Vos, W. (2007) *Microarrays: Methods and Protocols*.

- Methods in Molecular Biology (Series Editor J.M. Walker). Screening of cDNA Libraries on Glass Slide Microarrays. Humana Press, Totowa, New Jersey, USA. 2nd edition.
- Bevan, M., Mayerb, K., Whitec, O., Eisenc, J. A., Preussd, D., Bureaue, T., Salzbergc, S. L. and Mewesb, H.-W. (2001) Sequence and analysis of the Arabidopsis genome. *Current Opinion in Plant Biology* **4**, 2, 105–110.
- Birch, P., Avrova, A., Duncan, J., Lyon, G. and Toth, R. (1999) Isolation of potato genes that are induced during an early stage of the hypersensitive response to *Phytophthora infestans*. *Molecular Plant-Microbe Interactions* **12**, 365–361.
- Blanco, F., Garretón, V., Frey, N., Dominguez, C., Pérez-Acle, T., der Straeten, D. V., Jordana, X. and Holuigue, L. (2005) Identification of NPR1-dependent and independent genes early induced by salicylic acid treatment in Arabidopsis. *Plant Molecular Biology* **59**, 6, 927–944.
- Bloom, J., Khan, Z., Kruglyak, L., Singh, M. and Caudy, A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**, 211.
- Bolwell, G. (1999) Role of active oxygen species and NO in plant defence responses. *Current Opinion in Plant Biology* **2**, 4, 287–294.
- Botha, A., Lacock, L., van Niekerk, C., Matsioloko, M., du Preez, F., Loots, S., Venter, E., Kunert, K. and Cullis, C. (2006) Is photosynthetic transcriptional regulation in *Triticum aestivum* L. cv. ‘TugelaDN’ a contributing factor for tolerance to *Diuraphis noxia* (Homoptera: Aphididae)? *Plant Cell Reports* **25**, 1, 41–54.
- Bruland, T., Anderssen, E., Doseb, B., Bergum, H., Beisvag, V. and Lægred, A. (2007) Optimization of cDNA microarrays procedures using criteria that do not rely on external standards. *BMC Genomics* **8**, 377.
- Büttner, M. and Singh, K. B. (1997) *Arabidopsis thaliana* ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein. *Proceedings of the National Academy of Sciences* **94**, 11, 5961–5966.
- Cheung, F., Haas, B., Goldberg, S., May, G., Xiao, Y. and Town, C. (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* **7**, 1.
- Collett, H., Shen, A., Gardner, M., Farrant, J., Denby, K. and Illing, N. (2004) Towards transcript profiling of desiccation tolerance in *Xerophyta humilis*: construction of a normalized 11 k X. humilis cDNA set and microarray expression analysis of 424 cDNAs in response to dehydration. *X. humilis* cDNA set and microarray expression analysis of 424 cDNAs in response to dehydration. *Physiologia Plantarum* **122**, 1, 39–53.

- Copeland, R. (2008) *Essential sqlalchemy*. O'Reilly.
- Crampton, B., Hein, I. and Berger, D. (2009) Salicylic acid confers resistance to a biotrophic rust pathogen, *Puccinia striata*, in pearl millet (*Pennisetum glaucum*). *Molecular plant pathology* **10**, 2, 291–304.
- Cui, X. and Churchill, G. A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.
- Dalgaard, P. (2002) *Introductory Statistics with R (Statistics and Computing)*. Springer.
- de Bianchi, S., Dall'Osto, L., Tognon, G., Morosinotto, T. and Bassi, R. (2008) Minor Antenna Proteins CP24 and CP26 Affect the Interactions between Photosystem II Subunits and the Electron Transport Rate in Grana Membranes of Arabidopsis. *Plant Cell* **20**, 4, 1012–1028.
- de Torres Zabala, M., Bennett, M., Truman, W. and Grant, M. (2009) Antagonism between salicylic and abscisic acid reflects early host-pathogen conflict and moulds plant defence responses. *Plant Journal* **59**, 3, 375–386.
- Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences* **93**, 12, 6025–6030.
- Diatchenko, L., Lukyanov, S., Lau, Y. F. and Siebert, P. D. (1999) Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. *Methods Enzymol* **303**, 349–380.
- Dingkuhn, M., Singh, B., Clerget, B., Chantreau, J. and Sultan, B. (2006) Past, present and future criteria to breed crops for water-limited environments in West Africa. *Agricultural Water Management* **80**, 241–261.
- Dixon, D., Laphorn, A. and Edwards, R. (2002) Plant glutathione transferases. *Genome Biology* **3**, 3.
- Dong, X. (1998) SA, JA, ethylene, and disease resistance in plants. *Current Opinion in Plant Biology* **1**, 4, 316–323.
- Dowd, C., Wilson, I. and McFadden, H. (2004) Gene expression profile changes in cotton root and hypocotyl tissues in response to infection with *Fusarium oxysporum f. sp. vasinfectum*. *Molecular Plant-Microbe Interactions* **17**, 6, 654–667.
- Eckelkamp, C., Ehmann, B. and Schopfer, P. (1993) Wound-induced systemic accumulation of a transcript coding for a Bowman-Birk trypsin inhibitor-related protein in maize (*Zea mays* L.) seedlings. *FEBS letters* **323**, 1-2, 73–76.
- Feise, R. J. (2002) Do multiple outcome measures require p-value adjustment? *BMC Medical*

Research Methodology **2**, 8.

- Freestone, P., Nyström, T., Trinei, M. and Norris, V. (1997) The universal stress protein, UspA, of *Escherichia coli* is phosphorylated in response to stasis. *Journal of Molecular Biology* **274**, 3, 318–324.
- Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Current Opinion in Plant Biology* **9**, 4, 436–444.
- Galau, G., Wang, H. and Hughes, D. (1993) Cotton Lea5 and Lea14 Encode Atypical Late Embryogenesis-Abundant Proteins. *Plant Physiology* **101**, 2, 695–696.
- Ge, Y., Dudoit, S. and Speed, T. (2003) Resampling-based multiple testing for microarray data analysis, with discussion. *TEST* **12**, 1–78.
- Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, 10, R80.
- Goldman, J., Hanna, W., Fleming, G. and Ozias-Akins, P. (2003) Fertile transgenic pearl millet [*Pennisetum glaucum* (L.) R. Br.] plants recovered through microprojectile bombardment and phosphinothricin selection of apical meristem-, inflorescence-, and immature embryo-derived embryogenic tissues. *Plant Cell Reports* **21**, 10, 999–1009.
- Hardiman, G. (2004) Microarray platforms - comparisons and contrasts. *Pharmacogenomics* **5**, 5, 487–502.
- He, Y., Fukushige, H., Hildebrand, D. and Gan, S. (2002) Evidence Supporting a Role of Jasmonic Acid in Arabidopsis Leaf Senescence. *Plant Physiology* **128**, 3, 876–884.
- Heath, M. (2000) Hypersensitive response-related death. *Plant Mol Biol* **44**, 3, 321–334.
- Hein, I., Campbell, E., Woodhead, M., Hedley, P., Young, V., Morris, W., Ramsay, L., Stockhaus, J., Lyon, G., Newton, A. and Birch, P. J. (2004) Characterisation of early transcriptional changes involving multiple signalling pathways in the Mla13 barley interaction with powdery mildew (*Blumeria graminis* f. sp. *hordei*). *Planta* **218**, 5, 803–813.
- Herrero, J., Al-Shahrour, F., Díaz-Uriarte, R., Mateos, Á., Vaquerizas, J., Santoyo, J. and Dopazo, J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research* **31**, 13, 3461–3467.
- Hikichi, Y., Yoshimochi, T., Tsujimoto, S., Shinohara, R., Nakaho, K., Kanda, A., Kiba, A. and Ohnishi, K. (2007) Global regulation of pathogenicity genes at early stages of the infection process of *Ralstonia solanacearum*. *Plant Biotechnology* **24**, 149–154.

- Hillmann, A., Dunne, E. and Kenny, D. (2009) cDNA Amplification by SMART-PCR and Suppression Subtractive Hybridization (SSH)-PCR. *DNA and RNA Profiling in Human Blood* 223–243.
- Iuchi, S., Yamaguchi-Shinozaki, K., Urao, T., Terao, T. and Shinozaki, K. (1996) Novel Drought-Inducible Genes in the Highly Drought-Tolerant Cowpea: Cloning of cDNAs and Analysis of the Expression of the Corresponding Genes. *Plant Cell Physiology* **37**, 8, 1073–1082.
- Jayashree, B., Buhariwalla, H. K., Shinde, S. and Crouch, J. H. (2005) A legume genomics resource: The Chickpea Root Expressed Sequence Tag Database. *Electronic Journal of Biotechnology* **8**, 2.
- Jones, J. and Dangl, J. (2006) The plant immune system. *Nature* **444**, 323–329.
- Jost, R., Berkowitz, O., Wirtz, M., Hopkins, L., Hawkesford, M. and Hell, R. (2002) Genomic and functional characterization of the oas gene family encoding O-acetylserine (thiol) lyases, enzymes catalyzing the final step in cysteine biosynthesis in *Arabidopsis thaliana*. *Gene* **253**, 2, 237–247.
- Kanzaki, H., Saitoh, H., Ito, A., Fujisawa, S., Kamoun, S., Katou, S., Yoshioka, H. and Terauchi, R. (2003) Blackwell Publishing Ltd. Cytosolic HSP90 and HSP70 are essential components of INF1-mediated hypersensitive response and non-host resistance to *Pseudomonas cichorii* in *Nicotiana benthamiana*. *Molecular Plant Pathology* **4**, 5, 383–391.
- Kasukawa, T., Furuno, M., Nikaido, I., Bono, H., Hume, D., and D.P. Hill, C. B., Baldarelli, R., Gough, J., Kanapin, A., Matsuda, H., Schriml, L., Hayashizaki, Y., Okazaki, Y. and Quackenbush, J. (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Research* **13**, 6B, 1542–1551.
- Kavar, T., Maras, M., Kidric, M., Sustar-Vozlic, J. and Meglic, V. (2008) Identification of genes involved in the response of leaves of *Phaseolus vulgaris* to drought stress. *Molecular Breeding* **21**, 2, 159–172.
- Kerr, M. and Churchill, G. (2001) Experimental Design of DNA Microarray Experiments. *Biostatistics* **2**, 183–201.
- Kwon, Y., Kim, S., Jung, M., Kim, M., Oh, J., Ju, H., Kim, K., Vierling, E., Lee, H. and Hong, S. (2006) *Arabidopsis* hot2 encodes an endochitinase-like protein that is essential for tolerance to heat, salt and drought stresses. *The Plant Journal* **49**, 2, 184–193.
- Law, P., Claudel-Renard, C., Joubert, F., Louw, A. and Berger, D. (2008) MADIBA: A web server toolkit for biological interpretation of Plasmodium and plant gene clusters. *BMC Genomics* **9**, 105, 1471–2164.
- Lawton, K., Friedrich, L., Hunt, M., Weymann, K., Delaney, T., Kessmann, H., Staub, T. and Ryals, J. (1996) Benzothiadiazole induces disease resistance in *Arabidopsis* by

- activation of the systemic acquired resistance signal transduction pathway. *The Plant Journal* **10**, 1, 71–82.
- Lazzari, B., Caprera, A., Vecchietti, A., Stella, A., Milanese, L. and Pozzi, C. (2005) ESTree db: a tool for peach functional genomics. *BMC Bioinformatics* **1**, 6 Supplement 4, S16.
- Liang, P. and Pardee, A. B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 5072, 967–971.
- Liu, C., Zhang, Y., Cao, D., He, Y., Kuang, T. and Yang, C. (2008) Structural and functional analysis of the antiparallel strands in the lumenal loop of the major light-harvesting chlorophyll a/b complex of photosystem II (LHCIIb) by site-directed mutagenesis. *Journal of Biological Chemistry* **283**, 1, 487–495.
- Lonnstedt, I. and Speed, T. (2002) Replicated Microarray Data. *Statistica Sinica* , 12, 31–46.
- Luo, M., Liang, X., Dang, P., Holbrook, C., Bausher, M., Lee, R. and Guo, B. (2005) Microarray-based screening of differentially expressed genes in peanut in response to *Aspergillus parasiticus* infection and drought stress. *Plant Science* **169**, 4, 695–703.
- Mahalingam, R., Gomez-Buitrago, A., Eckardt, N., Shah, N., Guevara-Garcia, A., Day, P., Raina, R. and Fedoroff, N. V. (2003) Characterizing the stress/defense transcriptome of Arabidopsis. *Genome Biology* **4**, 3, R20.
- Mishra, G., Zhang, W., Deng, F., Zhao, J. and Wang, X. (2006) A Bifurcating Pathway Directs Abscisic Acid Effects on Stomatal Closure and Opening in Arabidopsis. *Science* **312**, 5771, 264–266.
- Morissette, D., Dauch, A., Beech, R., Masson, L., Brousseau, R. and Jabaji-Hare, S. (2008) Isolation of mycoparasitic-related transcripts by SSH during interaction of the mycoparasite *Stachybotrys elegans* with its host *Rhizoctonia solani*. *Current Genetics* **53**, 67–80.
- Muchero, W., Ehlers, J., Close, T. and Roberts, P. (2009) Mapping QTL for drought stress-induced premature senescence and maturity in cowpea [*Vigna unguiculata* (L.) Walp.]. *TAG Theoretical and Applied Genetics* **118**, 5, 849–863.
- Nawrath, C., Heck, S., Parinthewong, N. and Métraux, J.-P. (2002) EDS5, an Essential Component of Salicylic Acid-Dependent Signaling for Disease Resistance in Arabidopsis, Is a Member of the MATE Transporter Family. *The Plant Cell* **14**, 275–286.
- Norelli, J., Farrell, R., Bassett, C., Baldo, A., Lalli, D., Aldwinckle, H. and Wisniewski, M. (2009) Rapid transcriptional response of apple to fire blight disease revealed by cDNA suppression subtractive hybridization analysis. *Tree Genetics and Genomes* **5**, 27–40.
- O'Neill, M. J. and Sinclair, A. H. (1997) Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research* **25**, 13, 2681–2682.
- Park, A., Cho, S., Yun, U., Jin, M., Lee, S., Sabetto-Martins, G. and Park, O. (2001) Interaction of the Arabidopsis Receptor Protein Kinase Wak1 with a Glycine Rich Protein

- AtGRP-3. *Journal of Biological Chemistry* **276**, 28, 26688–26693.
- Park, T., Yi, S., Lee, S., Lee, S. Y., Yoo, D., Ahn, J. and Lee, Y. (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* **19**, 6, 694–703.
- Paschall, J., Oleksiak, M., VanWye, J., Roach, J., Whitehead, J., Wyckoff, G., Kolell, K. and Crawford, D. (2004) FunnyBase: a systems level functional annotation of Fundulus ESTs for the analysis of gene expression. *BMC Genomics* **5**, 96.
- Quass, C. (1995) Guidelines for the Production of Cowpeas. *National Department of Agriculture, Pretoria, South Africa*.
- R Development Core Team (2009) *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria.
- Ralph, D., McClelland, M. and Welsh, J. (1993) RNA fingerprinting using arbitrarily primed PCR identifies differentially regulated RNAs in mink lung (Mv1Lu) cells growth arrested by transforming growth factor beta 1. *Proceedings of the National Academy of Sciences* **90**, 22, 10710–10714.
- Ralph, S., Chun, H. J., Cooper, D., Kirkpatrick, R., Kolosova, N., Gunter, L., Tuskan, G., Douglas, C., Holt, R., Jones, S., Marra, M. and Bohlmann, J. (2008) Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics* **9**, 1.
- Ramm, M., Dangoor, K. and Sayfan, G. (2006) *Rapid Web Applications with TurboGears: Using Python to Create Ajax-Powered Sites*. Prentice Hall Open Source Software Development Series. Prentice Hall PTR Upper Saddle River, NJ, USA.
- Rasbery, J. M., Shan, H., LeClair, R. J., Norman, M., Matsuda, S. P. T. and Bartel, B. (2007) Arabidopsis thaliana Squalene Epoxidase 1 Is Essential for Root and Seed Development. *The Journal of Biological Chemistry* **282**, 23, 17002–17013.
- 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 Research* **31**, 1224–228.
- Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23**, 20, 2700–2707.
- Rogers, E. E. and Ausubel, F. M. (1997) Arabidopsis Enhanced Disease Susceptibility Mutants Exhibit Enhanced Susceptibility to Several Bacterial Pathogens and Alterations in

- PR-1 Gene Expression. *Plant Cell* **9**, 3, 305–316.
- Rohrmeier, T. and Lehle, L. (1993) WIP1, a wound-inducible gene from maize with homology to Bowman-Birk proteinase inhibitors. *Plant Molecular Biology* **22**, 5, 783–792.
- Rupert, B., Bonghi, C., Ziliotto, F., Pagni, S., Rasori, A., Varotto, S., Tunutti, P., Giovannoni, J. and Ramina, A. (2002) Characterization of a Major Latex Protein (Mlp) Gene Down-Regulated by Ethylene During Peach Fruitlet Abscission. *Plant Science* **163**, 265–272.
- Ryals, J., Neuenschwander, U., Willits, M., Molina, A., Steiner, H. and Hunt, M. (1996) Systemic Acquired Resistance. *The Plant Cell* **8**, 10, 1809–1819.
- Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Klapa, J. B. M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V. and Quackenbush, J. (2003) TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **34**, 2, 374–378.
- Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Taji, T., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y. and Shinozaki, K. (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant Journal* **31**, 3, 279–292.
- Singh, B., Ajeigbe, H., Tarawali, S., Fernandez-Rivera, S. and Musa, A. (2003) Improving the production and utilization of cowpea as food and fodder. *Field Crops Research* **84**, 169–177.
- Smith, R., Buchser, W., Lemmon, M., Pardin, J., Bixby, J. and Lemmon, V. (2008) EST Express: PHP/MySQL based automated annotation of ESTs from expression libraries. *BMC Bioinformatics* **9**, 186.
- Smyth, G. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 3.
- Smyth, G. (2005) *Limma: linear models for microarray data*. Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health). Springer Science+Business Media, Incorporated.
- Smyth, G., Michaud, J. and Scott, H. (2005) The use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 9, 2067–2075.
- Smyth, G., Ritchie, M., Thorne, N. and Wettenhall, J. (2008) *limma: Linear Models for Microarray Data User's Guide*. *The Walter and Eliza Hall Institute of Medical Research*.
- Smyth, G. and Speed, T. (2003) Normalization of cDNA microarray data. *Methods* **31**, 4,

265–273.

- Smyth, G., Yang, Y. and Speed, T. (2003) Statistical Issues in cDNA Microarray Data Analysis. *Methods in Molecular Biology* **224**, 111–136.
- Soltis, D., Soltis, P., Albert, V., Oppenheimer, D., dePamphilis, C., Ma, H. and Theissen, M. F. G. (2002) Missing links: the genetic architecture of flowers [correction of flower] and floral diversification. *Trends in Plant Science* **7**, 1, 22–31.
- Spreeth, M., Slabbert, M., de Ronde, J., van den Heever, E. and Ndou, A. (2004) Screening of Cowpea, Bambara Groundnut and Amaranthus Germplasm for Drought Tolerance and Testing of the Selected Plant Material in Participation with Targeted Communities. *WRC Report No 944/1/04*, WRC, Pretoria.
- Steyn, A., Smit, C., du Toit, S. and Strasheim, C. (1994) *Modern statistics in practice*. J.L. van Schaik Publishers.
- Student (William Sealy Gosset) (1908) The probable error of a mean *Biometrika* **6**, 1, 1–25.
- Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T. and Dixon, R. A. (2002) A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Journal* **32**, 6, 1033–1048.
- Tanaka, K., Asami, T., Yoshida, S., Nakamura, Y., Matsuo, T. and Okamoto, S. (2005) Brassinosteroid Homeostasis in Arabidopsis Is Ensured by Feedback Expressions of Multiple Genes Involved in Its Metabolism. *Plant Physiology* **138**, 1117–1125.
- Thatcher, L., Anderson, J. and Singh, K. (2005) Plant defence responses: what have we learnt from Arabidopsis? *Functional Plant Biology* **32**, 1, 1–19.
- Thomma, B., Eggermont, K., Tierens, K.-J. and Broekaert, W. (1999) Requirement of Functional Ethylene-Insensitive 2 Gene for Efficient Resistance of Arabidopsis to Infection by *Botrytis cinerea*. *Plant Physiology* **121**, 4, 1093–1101.
- Thompson, J., Higgins, D. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Computational Biology* **22**, 22, 4673–4680.
- Timko, M., Rushton, P., Laudeman, T., Bokowiec, M., Chipumuro, E., Cheung, F., Town, C. and Chen, X. (2008) Sequencing and analysis of the gene-rich space of cowpea. *BMC Genomics* **9**, 1.
- Toegel, S., Huang, W., Piana, C., Unger, F., Wirth, M., Goldring, M., Gabor, F. and Viernstein, H. (2007) Selection of reliable reference genes for qPCR studies on chondroprotective action. *BMC Molecular Biology* **8**, 1.
- Treviño, M. B. and O’Connell, M. A. (1998) Three Drought-Responsive Members of the Nonspecific Lipid-Transfer Protein Gene Family in *Lycopersicon pennellii* Show Different Developmental Patterns of Expression. *Plant Physiology* **116**, 1461–1468.

- Truman, W., de Zabala, Torres, M. and Grant, M. (2006) Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance. *The Plant Journal* **46**, 1, 14–33.
- Tunnacliffe, A. and Wise, M. (2007) The continuing conundrum of the LEA proteins. *Naturwissenschaften* **94**, 791–812.
- Tusher, V., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 9, 5116–5121.
- van den Berg, N., Berger, D., Hein, I., Birch, P., Wingfield, M. and Viljoen, A. (2007) Tolerance in banana to Fusarium wilt is associated with early up-regulation of cell wall-strengthening genes in the roots. *Molecular Plant Pathology* **8**, 3, 333–341.
- van den Berg, N., Crampton, B., Birch, P., Hein, I. and Berger, D. (2004) High throughput screening of SSH cDNA libraries using DNA microarray analysis. *BioTechniques* **37**, 5, 818–824.
- Varshney, R. K., Close, T. J., Singh, N. K., Hoisington, D. A. and Cook, D. R. (2009) Orphan legume crops enter the genomics era! *Current Opinion in Plant Biology* **12**, 2, 202–210.
- Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 5235, 484–487.
- Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.
- Weckx, S., Rijk, P. D., Broeckhoven, C. V. and Del-Favero, J. (2004) SSHSuite: an integrated software package for analysis of large-scale suppression subtractive hybridization data. *BioTechniques* **36**, 6, 1043–1045.
- Wei, C., Li, J. and Bumgarner, R. (2004) Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics* **5**, 1.
- Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D. and McClelland, M. (1992) Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research* **20**, 19, 4965–4970.
- Wettenhall, J. M. and Smyth, G. K. (2004) limmaGUI: A graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**, 18, 3705–3706.
- Wi, S. J., Kim, W. T. and Park, K. Y. (2006) Overexpression of carnation S-adenosylmethionine decarboxylase gene generates a broad-spectrum tolerance to abiotic stresses in transgenic tobacco plants. *Plant Cell Reports* **25**, 10, 1111–1121.
- Wit, E. and McClure, J. (2004) *Statistics for Microarrays: Design, Analysis and Inference* John Wiley & Sons, Ltd.
- Xia, X., McClelland, M. and Wang, Y. (2005) WebArray: an online platform for microarray

- data analysis. *BMC Bioinformatics* **6**, 1.
- Yang, G., Ross, D., Kuang, W., Brown, P. and Weigel, R. (1999) Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acids Research* **27**, 6, 1517–1523.
- Yang, Y., Dudoit, S., Luu, P. and Speed, T. (2001) Normalization for cDNA microarray data. *Microarrays: Optical Technologies and Informatics* **4266**, 141–152.
- Yang, Y. and Thorne, N. (2003) *Normalization for two-color cDNA microarray data*. Science and Statistics: A Festschrift for Terry Speed. Institute of Mathematical Statistics, Lecture Notes - Monograph Series.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, 4.
- Zegzouti, H., Marty, C., Jones, B., Bouquin, T., Latche, A., Pech, J. C. and Bouzayen, M. (1997) Improved Screening of cDNAs Generated by mRNA Differential Display Enables the Selection of True Positives and the Isolation of Weakly Expressed Messages. *Plant Molecular Biology Reporter* **15**, 3, 236–245.
- Zhang, L., Li, F.-G., Liu, C.-L., Zhang, C.-J. and Zhang, X.-Y. (2009) Construction and analysis of cotton (*Gossypium arboreum* L.) drought-related cDNA library. *BMC Research Notes* **1**, 120.
- Zhong, R., Kays, S., Schroeder, B. and Ye, Z. (2002) Mutation of a chitinase-like gene causes ectopic deposition of lignin, aberrant cell shapes, and overproduction of ethylene. *Plant Cell* **14**, 1, 165–179.
- Zhou, J., Wang, X., Jiao, Y., Qin, Y., Liu, X., He, K., Chen, C., Ma, L., Wang, J., Xiong, L., Zhang, Q., Fan, L. and Deng, X. (2007) Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle. *Plant Molecular Biology* **63**, 5, 591–608.
- Zhou, X. and Su, Z. (2007) EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics* **8**, 1.
- Zhu-Salzman, K., Salzman, R. A., Ahn, J. and Koiwa, H. (2004) Transcriptional Regulation of Sorghum Defense Determinants against a Phloem-Feeding Aphid. *Plant Physiology* **134**, 1, 420–431.

Appendix

SSHscreen R Documentation

High-throughput screening of SSH cDNA libraries using DNA microarray data

Description

This package implements the calculations outlined in van den Berg *et al.* (2004). It also implements a revised and extended approach to the one presented there, to include more sophisticated normalization and statistical analysis steps (using the limma package) as described in Berger *et al.* (2007). Further improvements in the functionality of SSHscreen lead to the latest version 2.0.0. We recommend the use of \geq R-2.2.0 for this package.

Usage

```
SSHscreen(source = NULL, path = NULL, bc.method = "normexp", offset = 50, wa.method = "printtiploess", ba.method = "Aquantile", ndups = 1, spacing = 1, irregular = FALSE, method = "ER3", toplist = 100, adjust = "none", spot.ave = TRUE, mfrow = c(2,2), norm.plot = FALSE, weights = FALSE, negflags = 1, legend = TRUE, sort = "B", cutoff = "none", library = "F", proportion = 0.75, ...)
```

Arguments

- source** Character string specifying the image analysis program, which produced the output files. Choices are "agilent", "arrayvision", "genepix", "genepix.median", "bluefuse", "imagine", "quantarray", "smd.old", "smd", "spot" or "spot.close.open". If source is "other" the data files should be tab-delimited text files and should contain columns named 'SpotLabel', 'GeneList', 'Rf', 'Gf', 'Rb' and 'Gb' which contain the names of the spot labels, gene names, as well as the red and green foreground and background. Note that for this option only global loess within-array normalization is possible, as no layout parameters are specified.
- path** Character string giving the directory containing the files. This directory should contain the image analysis files, Targets file and SpotTypes file. The last two files are specified according to the conventions of the limma package. See limma

documentation for more details. In the case of “genepix” or “spot” data, this directory should also contain the GAL file.

- bc.method Character string specifying background correct method. Possible values are “none”, “subtract”, “half”, “minimum”, “movingmin”, “edwards” or “normexp”.
- offset Numeric value to add to intensities before log-transforming. This may eliminate the usual ‘fanning’ of log-ratios at low intensities associated with local background subtraction.
- wa.method Character string specifying the within-array normalization method. Choices are “none”, “median”, “loess”, “printtiploess”, “composite”, “control” and “robust-spline”. When “control” or “composite” is used, all spots where the spot name (‘Name’ column in the image analysis output files) contains the word ‘control’ are used. The weights argument can be specified together with “loess” or “printtiploess” within-array normalization (this is called up-weighting print-tip loess normalization), when a suitable set of spiked-in control spots (non-differentially expressed and spanning the whole intensity range) are available on each slide.
- weights If TRUE the up-weighting within-array normalization method for spiked-in control spots will be used. That is print-tip loess (or global loess) normalization, assigning zero weight to the cDNAs and double weight to the control spots. Note that ‘wa.method’ must be “printtiploess” (or “loess”).
- ba.method Character string specifying the between-array normalization method to be used. Choices are “none”, “scale”, “quantile”, “Aquantile”, “Gquantile”, “Rquantile”, “Tquantile” or “vsn”.
- ndups Positive integer giving the number of times each gene is printed on an array.
- spacing The spacing between the rows of the expression matrix corresponding to duplicate spots, ‘spacing=1’ for consecutive spots.
- irregular If each gene is spotted the same number of time, but the spacing between replicate spots are irregular, then irregular=TRUE should be specified in the function call. This will sort the gene list by gene ID. Spacing should then be specified as 1 along with the value of ndups.
- method Two options are available:
1. “ER1” for a ER1 versus ER2 comparison. Two groups of arrays are required: one hybridized with un-subtracted driver (UD) and subtracted tester (ST) and another hybridized with un-subtracted tester (UT) and subtracted tester (ST). Dye-swaps may be included.
 2. “ER3” for a ER3 versus -ER2 (invER2) comparison. Two groups of arrays are required: one hybridized with un-subtracted tester (UT) and sub-

- tracted tester (ST) and another hybridized with un-subtracted driver (UD) and un-subtracted tester (UT). Dye-swaps may be included.
- toplist The number of 'top' genes provided in the top table (output). These genes will also be indicated with a cross on the SSHscreen ER plots. This will only be valid if the cut-off argument is "none" (by default).
- adjust Method to use to adjust the p-values for multiple testing e.g., "holm" or "fdr". See 'p.adjust' for the available options. If "none" then the p-values are not adjusted.
- spot.ave If TRUE duplicate spots will be averaged before fitting a linear model. If FALSE duplicate spots on each array will be analyzed separately using the limma 'duplicateCorrelation' function.
- mfrow Specifies the plot layout for the MA plots. A vector specifying the number of rows (nr) and number of columns (nc), i.e. c(nr,nc). nr * nc should be equal to the number of microarray slides (in order to view all MA-plots in one window).
- norm.plot If TRUE, MA plots will be produced of arrays before and after normalization and saved to the working directory (specified by the path argument) in pdf format.
- negflags A value between 0 and 1 (can also be 0 or 1) to change the spot quality weights of all spots which receive a negative flag from the image analysis program. These spot quality weights are used during normalization.
- legend If TRUE, a legend of plotting symbols and colours will be included in MA plots.
- sort Select a criteria by which the genes will be sorted. Choices are "B" for the B-statistic, "t" for the t-statistic or "p" for the p-value. If sort is "none", a fixed number of genes specified by 'toplist' will be selected. If sort is not "none", a cutoff value can also be specified.
- cutoff Select a cutoff value from which the number of genes in the top table will automatically be determined. This value should correspond to the sort argument, for example 0 when sort="B" or 0.05 when sort="p".
- library A character string specifying the library to analyze. "F" for a forward library analysis only, "R" for a reverse library analysis only and "both" when the forward and reverse libraries are spotted on the same microarray slides. Note that when "both" are used, the cDNAs should be named cDNA_F for a clone from the forward library and cDNA_R for a clone from the reverse library.
- proportion Numeric value between 0 and 1, indicating the assumed proportion of genes which are differentially expressed.

Details

This function makes use of functionality provided in the *limma* package to import, normalize and analyze DNA microarray data for SSH screens. For an 'ER1' analysis, the function calculates ER1 and ER2 ratios for each gene, and constructs a ER1 versus ER2 plot. For an 'ER3' analysis, the function calculates ER3 and -ER2 (invER2) ratios for each gene, and constructs a ER3 versus invER2 plot.

- ER1 (Enrichment Ratio 1) = $\log_2(ST/UD)$
- ER2 (Enrichment Ratio 2) = $\log_2(ST/UT)$
- ER3 (Enrichment Ratio 3) = $\log_2(UT/UD)$

This package is for simple microarray designs corresponding to the dye-swap design described in van den Berg *et al.* (2004). For the 'ER1' analysis option it assumes that data is available for two groups of arrays: one hybridized with ST and UD and another hybridized with ST and UD. Each of these groups can include any number of dye-swap replicates. The design should be specified in the Targets file. For the 'ER3' analysis option it assumes that data is available for two groups of arrays: one hybridized with ST and UT and another hybridized with UT and UD. Each of these groups can include any number of dye-swap replicates. The design should be specified in the Targets file.

For the 'ER1' analysis a linear model is fitted to the data for each clone using *limma*, in order to identify clones for which $ER1 > ER2$ (significantly). It is assumed that the ST sample can be treated as a common reference so that the contrast $UD - UT$ can be tested. This is equivalent to testing for which clones ER1 is significantly larger than ER2. The top number of clones specified by the 'toplist' argument are returned in the top table output if 'cutoff' is not set. If 'cutoff' is not equal to "none", the number of genes are automatically determined using information from the 'sort' and 'cutoff' arguments.

For the 'ER3' analysis, two linear models are fitted to separate parts of the data. A linear model fit is applied to test $ER3 \neq 0$, to determine significantly up- and down-regulated clones in a -ER2 (invER2) versus ER3 screen, and another linear model fit is applied to test $-ER2 \neq 0$, to identify rare and abundant clones in the treated sample.

Value

If `method="ER1"` and `library="F"` or "R", the value is a matrix of identified clones of length 'toplist' for which the relation $ER1 > ER2$ is most significant, as determined by the linear model.

If `method="ER3"` and `library="F"` or "R", the value is a list object with components:

- tt.ud Matrix of identified clones of length 'toplist' for which $ER3 \neq 0$ (i.e. up or down regulated). These are the most significant clones determined by the linear model.
- tt.ar Matrix of identified clones of length 'toplist' for which $-ER2 \neq 0$ (i.e. rare or abundant). These are the most significant clones determined by the linear model.

If method="ER1" and library="both", the value is a list object with components:

- tt.F Matrix of identified clones of length 'toplist' for which the relation $ER1 > ER2$ is most significant in the forward library, as determined by the linear model.
- tt.R Matrix of identified clones of length 'toplist' for which the relation $ER1 > ER2$ is most significant in the reverse library, as determined by the linear model.

If method="ER3" and library="both", the value is a list object with components:

- tt.ud.F Matrix of identified clones of length 'toplist' for which $ER3 \neq 0$ (i.e. up or down regulated) in the forward library. These are the most significant clones determined by the linear model.
- tt.ar.F Matrix of identified clones of length 'toplist' for which $-ER2 \neq 0$ (i.e. rare or abundant) in the forward library. These are the most significant clones determined by the linear model.
- tt.ud.R Matrix of identified clones of length 'toplist' for which $ER3 \neq 0$ (i.e. up or down regulated) in the reverse library. These are the most significant clones determined by the linear model.
- tt.ar.R Matrix of identified clones of length 'toplist' for which $-ER2 \neq 0$ (i.e. rare or abundant) in the reverse library. These are the most significant clones determined by the linear model.

If method="ER1", a ER1 versus ER2 plot will be produced and saved to the working directory (specified by the 'path' argument) in pdf format (for the forward and reverse libraries separately). Significantly up- and down-regulated clones will be marked. The number of clones marked on the plots, is as specified by the 'toplist' or 'cutoff' arguments.

If method="ER3", two ER3 versus -ER2 (invER2) plots will be produced and saved to the working directory (specified by the 'path' argument) in pdf format (for the forward and reverse libraries separately). A plot where the significantly up- and down-regulated clones are marked and a plot where the significantly rare and abundant clones are marked. The number of clones marked on the plots, is as specified by the 'toplist' or 'cutoff' arguments.

Note

The latest version of SSHscreen, as well as demonstration data with an R script giving an example of the implementation of SSHscreen, can be downloaded at: <http://microarray.up.ac.za/SSHscreen/>.

Author(s)

Nanette Coetzer <nanette.coetzer@gmail.com>, Dave Berger <dave.berger@fabi.up.ac.za>, Wiesner Vos.

References

1. Van den Berg, N., Crampton, B., Birch, P., Hein, I. and Berger, D. (2004) High throughput screening of SSH cDNA libraries of SSH cDNA libraries using DNA microarrays. *BioTechniques* **37** (5), pp. 818-824.
2. Berger, D.K., Crampton, B., Hein, I. and Vos, W. (2007) Screening of cDNA Libraries on Glass Slide Microarrays. *In: Microarrays: Methods and Protocols (2nd Ed) (Editor J Brampal), Series: Methods in Molecular Biology (Series Editor JM Walker), Humana Press, Totowa, New Jersey, USA.*