

**GENOMIC AND PROTEOMIC ANALYSIS OF DROUGHT TOLERANCE IN SORGHUM**

*(SORGHUM BICOLOR (L.) MOENCH)*



**A dugna Abdi Woldesemayat**

UNIVERSITY *of the*  
WESTERN CAPE

A thesis submitted in partial fulfilment of the requirements for the degree of  
Doctor Philosophiae at the South African National Bioinformatics Institute,  
Department of Biotechnology, Faculty of Natural Science, University of the Western  
Cape

Supervisors: **Professor Alan Christoffels and Professor Bongani K. Ndimba**

November 2014

**GENOMIC AND PROTEOMIC ANALYSIS OF DROUGHT TOLERANCE IN SORGHUM**

*(SORGHUM BICOLOR (L.) MOENCH)*

**Aadugna Abdi Woldeesemayat**

**KEYWORDS**

*In silico* candidate gene identification

Drought tolerance

Functional genomics

Gene-trait-association

Novel gene prediction

Differential expression

MALDI-TOF-TOF/Mass-spectrometry

Protein identification

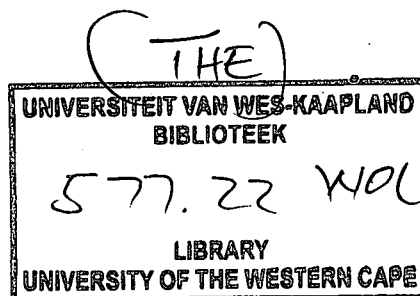
Proteomics

*Sorghum bicolor* (L.) Moench





UNIVERSITY *of the*  
WESTERN CAPE



## ABSTRACT

### **Genomic and Proteomic Analysis of Drought Tolerance in Sorghum (*Sorghum bicolor* (L.) Moench)**

**Adugna Abdi Woldesemayat**

**PhD Thesis, South African National Bioinformatics Institute, Department of Biotechnology,  
University of the Western Cape**

Drought is the most complex phenomenon that remained to be a potential and historic challenge to human welfare. It affects plant productivity by eliciting perturbations related to a pathway that controls a normal, functionally intact biological process of the plant. Sorghum (*Sorghum bicolor* (L.) Moench), a drought adapted model cereal grass is a potential target in the modern agricultural research towards understanding the molecular and cellular basis of drought tolerance. This study reports on the genomic and proteomic findings of drought tolerance in sorghum combining the results from *in silico* and experimental analysis. Pipeline that includes mapping expression data from 92 normalized cDNAs to genomic loci were used to identify drought tolerant genes. Integrative analysis was carried out using sequence similarity search, metabolic pathway, gene expression profiling and orthology relation to investigate genes of interest. Gene structure prediction was conducted using combination of *ab initio* and extrinsic evidence-driven information employing multi-criteria sources to improve accuracy. Gene ontology was used to cross-validate and to functionally assign and enrich genes.

An integrated approach that subtly combines functional ontology based semantic data with expression profiling and biological networks was employed to analyse gene association with plant phenotypes and to identify and genetically dissect complex drought tolerance in sorghum. The gramene database was used to identify genes with direct or indirect association to drought related ontology terms in sorghum. Where direct association for sorghum genes were not available, genes were captured using Ensemble Biomart by transitive association based on the putative functions of sorghum orthologs in closely related species. Ontology mapping represented a direct or transitive association of genes to multiple drought related ontology terms based on sorghum specific genes or orthologs in related species. Correlation of genes to enriched gene ontology (GO)-terms ( $p$ -value  $< 0.05$ ) related to the whole-plant structure was used to determine the extent of gene-phenotype

association across-species and environmental stresses.

Seeds of sorghum varieties obtained from the International Crop Research Institute for Semi-Arid Tropics (ICRISAT), India, were used in to identify drought responsive proteins. Plant samples were grown in greenhouse under differential conditions and post-flowering drought stress was induced on the onset of flowering using  $30\% \pm 5$  water field capacity (FC). Harvested samples were stored at  $-80^{\circ}\text{C}$  until use for protein identification. TCA/acetone method was used to extract proteins from leaf tissue (Btx642 sorghum variety) for this analysis. After conducting protein separation using two-dimensional gel electrophoresis (2DE), Coomassie Brilliant Blue (CBB) stained gels were scanned using Molecular Imager PharoFX Plus System (BIO-RAD). Spot detection and analysis of differential expression pattern was performed using PDQuest™ software (Bio-Rad) version 8.0.1 build 055. Sixteen spots selected based on abundance and resolution were used for protein identification using MALDI-TOF MS/MS and MASCOT search engine and database.

A total of 10619 UniGenes (75.5%) were mapped to reference genome using pair-wise sequence similarity search and were retained as high scoring segment pairs (HSPs) out of which 9763 were mapped to the existing gene loci. We report 123 (1.3%) putative uncharacterised proteins that matched UniGenes as drought responsive genes (DRGs) which were not previously ascribed. The remaining 856 (6.1%) HSPs that mapped to intergenic region were further screened for analysis in gene structure prediction. Interproscan (IPS) analysis of the 123 HSPs revealed 60.5% known signature domains. The number of genes with IPS and those enriched by GO analysis (P-value < 0.05) were comparable. Pathway analysis revealed 14 metabolic pathways related to drought tolerance and 32 genes responsible to encode protein enzyme that catalyse substrate conversions in the respective pathways. Expression profiling showed 12 genes significantly expressed under drought stress. Result from analysis of orthology group showed 265 non-redundant genes with response to drought stress. Identification of two merged genes and 3 corresponding transcripts, one novel mRNA, 64 novel exons, 74 five and 3595 three prime novel UTRs account for update of more than 4400 single gene models (~12.6%) of the existing annotation. This study reports 241 novel genes of which 69% represent drought responsive (DR), 6% complete gene model structure and 47.7% single exonic. On the basis of gene-gene and gene-phenotype association study, we report 169 sorghum genes identified for drought tolerance across species and environmental stresses. Out of this, 56% have shown multiple stress tolerance in sorghum, 90% exhibited drought tolerance with other species and 10% remain sorghum specific. Mapping biological ontologies validates the results and provides role for identified genes association. The resource enables us to

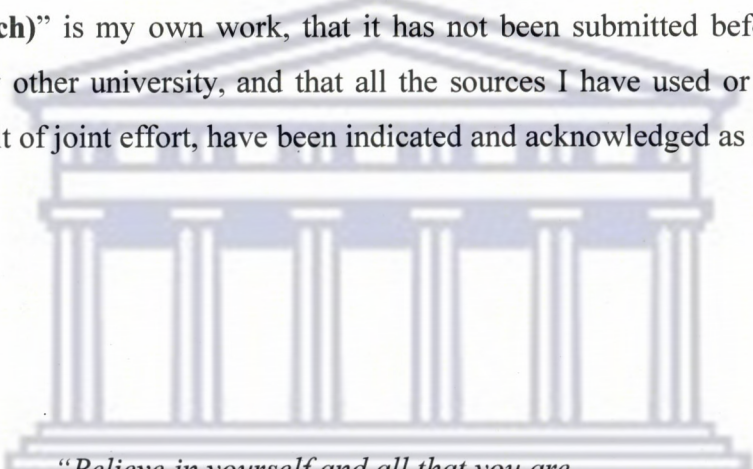
perform cross-species queries for genes that are likely to be associated with stress tolerance, as a means to identify unique opportunities to create stress resistance in sorghum and other crop species.

This study identified nine protein enzymes with novel function in sorghum from seven spots that resulted significant score out of sixteen selected for mass spectrometry analysis. Two spots were shown to have mixture of two different proteins each. Four classes of known protein functional groups were identified namely: proton (H<sup>+</sup>) transporting related (11.1%); carbohydrate metabolism (22%); carbon assimilation (33%); 4) Stress tolerance, defence and immunity (11.1%) and RNA binding proteins (11.1%) and an unknown protein. Most of the proteins were identified to be chloroplastic showing functions related to photosynthesis under drought stress. Of all identified proteins, 78% were shown to be significantly up-regulated suggesting their role in drought tolerance. However, this study also shows a typical mechanism where plants induce signal transduction alarm to bypass stress condition by down-regulating a rate limiting enzyme. This study reflects functional correlation of some key protein enzymes experimentally identified with some other enzymes identified *in silico* thus serving as validation tool bridging the gaps between genomic and proteomic research.

Our integrated *in silico* approach proves to be unique tool for detecting biologically plausible candidate genes. This study has successfully identified significant array of prioritized candidate known and novel genes that are critical to response to drought and related stresses. The pathways identified in this study signify the interplays of biochemical reactions that make up the metabolic network constituting fundamental interface for the crop to build defensive mechanism against drought stress. Multiple informants in the gene prediction prove to be reliable and dependable. This result entails yet untapped natural genetic variation in sorghum suggesting its key position in agricultural productivity and comparative proteogenomics as a model for grass family. The resource in this study represents a useful reference for future research in sorghum and related cereals.

## DECLARATION

I declare that “**Genomic and Proteomic Analysis of Drought Tolerance in Sorghum (*Sorghum bicolor* (L.) Moench)**” is my own work, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged as complete references.



*“Believe in yourself and all that you are.*

*Know that there is something inside you that is greater than any challenge.”*

*– Christian D. Larson*

**Adugna Abdi Woldesemayat**

**November 2014**

**Signed**

DEDICATION



*Dedicated to my late brother Teferi Abdi Woldesemayat*

UNIVERSITY *of the*  
&  
WESTERN CAPE  
*to my wife Genet Girma Lema*



## Acknowledgements

My sincere gratitude goes to my supervisors Prof. Alan Christoffels and Prof. Bongani K. Ndimba for their support and guidance throughout the course of my PhD. Without their constructive comments and suggestion this thesis would not have been possible. Prof. Alan thank you for accepting me into your research group. Your input towards my skill in bioinformatics and genomics was immense. Prof. Bongani thank you for your willingness to accept me into your lab and proteomics research group and for your efforts towards my practical know-how in proteomics. It's my pleasure to thank you both for enriching and taking my career steps forward.

My sincere thanks also go to Peter van Heusden for his relentless assistance during my study. Your contribution towards my computational bioinformatics part is invaluable. My in-depth gratitude also goes to Dr. Junaid Gamiieldien for his advice and help in building semantic query component and in reading the thesis chapter. I have special thanks for Mario Jones who assisted me in so many ways during my study. My sincere gratitude also goes to the following SANBI faculty and staff members who contributed to a success of my study in different ways: Dr. Nicki Tiffin, Dr. Gordon Harkins, Prof. Simon Travers, Ferial Mullins, Maryam Salie, Fungiwe Mpithi and Samantha Alexander. And I will not forget to thank Junita Williams and Dale Gibbs for their assistance during my early study. A thousand of thanks go to my friends and other people at SANBI: Dr Uliana Hasee; Dr. Mark Wamalwa; Dr. Sumir Panji, Dr Ashley Pretorius, Dr. Musa Nur Gabere, Dr. Mushal Ali, Dr Samson Miyanga, Dr. Samuel Kojo Kwofie, Dr. Siaka. Their contribution in one way or another have special places in my thesis. My special thanks go to Dr. Jean-Baka Domelevo Entfellner for reading of a chapter of my thesis and constrictive comments. Your advice and generous assistance have surely contributed to a success of my study. Emil, Darlington, Ebrahim, Dr. Sara, Dr. Ruben, Dr. Monique, and all others whom I forget to mention have contributed to my achievements.

I appreciate the assistance rendered to me in so many ways by the Proteomics Research Group: Dr. Rudo Ngara, Dr Omodele, Fabian and Lizex thank you for your contribution towards my proteomics knowledge. Roya, Anati, Didi, Rendani, Dr Tekalani, Xolisa, Dr. Putuma and others who I can't mention the name both from proteomics research group at UWC and ARC, your presence and support was like a family. I would also like to extend my special thanks to Ms Frances Starkey and Ms Melvine Pretorius for administrative assistance in the Department of Biotechnology during my study.

I am indebted to the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa for financial support. I am also grateful to the National Research Foundation of South Africa (NRF) and The National Agricultural Proteomics Research and Service Unit (NAPRSU) for the support rendered and to the International Crop Research Institute for Semi-Arid Tropics (ICRISAT) for the provision of research materials.

A profound appreciation goes to my friend Dr Kebede Ketero (Pretoria, SA) for his assistance in so many ways during my study. Many thanks Kebe. Dr. Mesfine Taffese (Ethiopia, Addis), Kebe Baleme (Ethiopia, Addis), Dr Mulatu Geleta (Sweden, SLU, Alnard), Dr Namera Shargie (Johannesburg, SA), members of Biftu PLC. (Ethiopia, Addis) deserve special thanks for their advice, cheering and assistance during my study. Sincere appreciation also goes to Ato Gemechu Keneni, a PhD student, from the Ethiopian Institute of Agricultural Research (EIAR) for giving me one sorghum variety.

Special thanks go to my dad Abdi Woldesemayat, my siter Almaz Abdi, Gete Abdi, my brothers Yohannes Abdi, Terecha Abdi, my cousin Sisay Girma, my mother-in-law Birtukuan Assefa and my uncle Bayssa Woldemichael and extended family for your unwavering support during the years of my study. My younger brother Taferi Abdi who died of Hepatitis B Virus and my step-mother were who I lost during the course, may God keep your soul in peace.

Finally, my little daughter Mary who tolerated the pain of missing me during the last two years of my study deserves many thanks. I love you baby. My beloved wife Genet Girma, equally contributed to a success of my PhD. You have been the pillar of strength, words are never sufficed to thank you. You are so wonderful, I love you.

*Thank you God for helping me to be strong and to overcome all the challenges during the course.*

## Table of contents

Contents	Pages
KEYWORDS.....	ii
ABSTRACT.....	iii
Acknowledgements.....	viii
Table of contents.....	x
List of abbreviations.....	xvii
List of tables.....	xix
List of figures.....	xxi
Chapter 1: General Introduction and Literature Review.....	1
Abstract.....	1
1.1 Introduction.....	2
1.2 Drought.....	3
1.2.1 What is drought?.....	3
1.2.2 Drought: “A primarily African problem”.....	3
1.2.3 Mode of responses and mechanisms of drought tolerance.....	4
1.3 Signal transduction and Transcriptional regulation network.....	5
1.3.1 Signal transduction cascade.....	5
1.3.2 Role of ABA in dehydration tolerance response.....	7
1.3.3 Transcriptional Regulatory Networks.....	8
1.4 Sorghum: “A drought-hardy and agriculturally important crop”.....	12
1.4.1 Sorghum Metabolic pathway: C4 crop.....	12
1.4.1.1 C4-pathway: A secret behind drought-hardy sorghum.....	13
1.4.1.2 Sorghum Metabolic pathway: SorghumCyc.....	13
1.4.2 Sorghum genome sequence: opportunity.....	14
1.4.2.1 Sorghum consensus gene prediction.....	14
1.4.2.2 Sorghum Genomics.....	16
1.4.2.3 Sorghum proteomics.....	16
1.4.2.4. Proteogenomics: Integrative genomics and proteomics.....	17
1.5 Homology and syntenic relationships.....	19
1.6 Gene-Trait Association.....	19
1.7 Candidate gene identification strategies.....	20
1.7.1 Integrated In silico Candidate Gene Approach (InsCGA).....	22

1.7.1.1. Analytical Strategies.....	22
1.7.1.1.1 Primary process (Core issues): a priori analytical process based knowledge.....	23
1.7.1.1.1.1 Core issue 1: Genes of interest (knowledge base).....	23
1.7.1.1.1.2 Core issue 2: Gene repositories and tools.....	24
1.7.1.1.1.3 Core issue 3: Identification of candidate genes.....	24
1.7.1.1.2 Central Process: underlying process for identification of candidate genes.....	24
1.7.1.1.3 Specialized process: a posteriori analytical process based knowledge.....	25
1.7.1.2 Genomic resources.....	27
1.7.1.2.1 Expressed Sequence Tag (EST) mapping.....	27
1.7.1.2.1.1 What is the advantage of ESTs?.....	28
1.7.1.2.1.2 ESTs: Tool for gene discovery.....	29
1.7.1.2.1.3 ESTs: as a source of data on gene expression and regulation.....	29
1.7.1.2.1.4 ESTs: tool for gene mapping (genome Landmarks).....	30
1.7.1.2.1.5 Challenges and limitations of ESTs.....	30
1.7.1.3 UniGene.....	31
1.7.1.3.1 UniGene clusters.....	31
1.7.1.3.2 UniGene build procedure: steps for inclusion of ESTs into clusters.....	32
1.7.1.4 Functional Genomic Annotation.....	33
1.7.1.5 Functional Ontology annotation.....	34
1.7.2 Experimental approach for gene identification.....	35
1.7.2.1 Drought phenotyping.....	35
1.7.2.2 Source of stay-green genes.....	36
1.7.2.3 Protein identification and purification.....	37
1.7.2.3.1 Protein Optimization and Quantification.....	37
1.7.2.4. Differential expression.....	38
1.8 Rationale of the thesis.....	39
1.9 Aims and objectives.....	39
1.10 Overview of the thesis.....	40
Chapter 2: In silico identification of candidate genes for drought tolerance in Sorghum ( <i>Sorghum bicolor</i> (L.) Moench).....	41
Abstract.....	41
2.1 Introduction.....	43
2.2 Materials and methods.....	48
2.2.1 Data sources.....	48

2.2.1.1 Genome Data.....	48
2.2.1.2 UniGene Data.....	48
2.2.1.3 NCBI EST.....	48
2.2.1.4 TIGR plant transcripts.....	1
2.2.2 pre-processing (quality filtering process).....	1
2.2.3 Mapping experimental data to reference genome.....	1
2.2.3.1 Building gene models in the intergenic regions.....	4
2.2.3.2 Annotation Comparison.....	6
2.2.3.3 Prediction of gene structure models using AUGUSTUS.....	7
2.2.3.4 Consistency in gene predictions.....	8
2.2.3.5 Filtering the gene model structures.....	8
2.2.4 Metabolic pathway analysis.....	9
2.2.5 Gene Ontology functional enrichment analysis.....	9
2.2.5.1 GO functional enrichment analysis using BLAST2GO.....	9
2.2.5.2 GO functional enrichment analysis using AGRIGO.....	10
2.2.6 Use of expression profiling for candidate gene identification.....	11
2.2.6.1 Statistical analysis of gene expression.....	12
2.2.7 Analysis of orthologous groups.....	12
2.3. Results.....	13
2.3.1 Mapping experimental data to reference genome.....	13
2.3.1.1 BLAST Sequence Similarity Search: Identification of candidate drought responsive genes.....	13
2.3.1.2 Reannotation of sorghum genome.....	13
2.3.1.3 Annotation comparison and an update.....	14
2.3.1.4 Structural and positional modification of the candidate genes.....	14
2.3.1.5 Novel gene structure model prediction.....	16
2.3.1.5.1 Genomic distribution of the novel genes.....	18
2.3.1.5.2 Alternative Splicing (AS): Intron retention and exon skipping.....	18
2.3.1.5.3 Distribution of Exon and Intron structure.....	21
2.3.1.5.4 Intronless (single exonic) genes.....	22
2.3.1.5.5 Prediction of pseudogenes.....	23
2.3.1.5.6 Identification of nearest intergenic distances.....	23
2.3.2 Metabolic pathways analysis.....	24
2.3.2.1 Functional GO-enrichment analysis of the pathway.....	26

2.3.2.2	Pattern of sequence distribution and GO annotation.....	27
2.3.2.3	Interpro Domain Analysis.....	28
2.3.3	Analysis of gene-expression profiling.....	29
2.3.3.1	Functional GO-enrichment analysis of gene-expression.....	30
2.3.4	Analysis of orthology groups.....	33
2.3.4.1	GO enrichment analysis of genes through orthology groups.....	33
2.4	Discussion.....	36
2.4.1	Identification of candidate genes by mapping experimental data to reference genome.....	36
2.4.2	Annotation comparison and update.....	37
2.4.2.1	Genome annotation modification.....	37
2.4.2.2	Novel gene structure model prediction.....	38
2.4.2.2.1	Complete and partial gene structure models.....	38
2.4.2.2.2	Single exonic and intronless genes.....	39
2.4.2.2.3	Correlation between intronless and pseudogenes.....	39
2.4.2.2.4	Splisomes.....	40
2.4.3	Metabolic pathways.....	40
2.4.4	Functional GO enrichment and Interpro domain analysis.....	42
2.4.5	Differential gene expression profiling.....	42
2.4.6	Analysis of orthology relationship.....	43
2.5	Conclusion.....	44
Chapter 3: Gene-gene and gene-phenotype association: a novel integrated approach to dissect complex drought tolerance in sorghum ( <i>Sorghum bicolor</i> (L.) Moench).....		45
Abstract.....		45
3.1	Introduction.....	47
3.2	Materials and Method.....	50
3.2.1	Data source and data mining.....	50
3.2.2	Identification of gene association using functional ontology based semantic query building. .	52
3.2.3	Cross-species comparative analysis: correlating gene-trait association across species.....	52
3.2.4	Multiple responses of genes across environmental stresses.....	53
3.2.4.1	Metabolic pathway and phylogenetic relationship.....	53
3.2.5	Integration of gene trait association with gene differential expression.....	53
3.2.6	Functional-annotation and GO Enrichment.....	54
3.3	Result.....	55
3.3.1	Gene association across-environmental stresses: Functional-cross-talk.....	55

3.3.1.1 Characteristic feature of 'SORBI_03g026070' and 'SORBI_09g030600': Implication in stress signal transduction pathway.....	56
3.3.1.2 Phylogenetic relationship.....	57
3.3.2 Comparative gene association across-species: Functional overlapping and specificity.....	58
3.3.3 Semantic integration of existing data based on functional ontology.....	59
3.3.4 Integration of differential expression data set.....	60
3.3.5 Functional-annotation and GO Enrichment.....	62
3.4 Discussion.....	66
3.4.1 Association to multi-environmental-stress tolerance.....	66
3.4.2 Resistance from whole-plant to individual level components.....	67
3.4.3 Understanding gene-phenotype-association through integration of functional ontologies.....	68
3.4.4. Cross-species functional crosstalk.....	68
3.4.5 Deciphering drought stress tolerance through integration of semantic knowledge.....	69
3.4.6 Association of expression profiling with drought phenotypes.....	70
3.5 Conclusion.....	71
Chapter 4: Identification of Drought Responsive Proteins in Sorghum ( <i>Sorghum bicolor</i> (L.) Moench) using Differential Expression Profiling and MALDI-TOF-TOF MS.....	72
Abstract.....	72
4.1 Introduction.....	74
4.2 Materials and Methods.....	76
4.2.1 Plant material.....	76
4.2.2 Experiment and growth conditions.....	77
4.2.3 Protein Optimization and Extraction.....	78
4.2.4 Quantification of Protein.....	82
4.2.5 Separation of protein.....	83
4.2.5.1 Electrophoretic separation of proteins: one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (1D-SDS-PAGE).....	83
4.2.5.1.1 Preparation of resolving gel.....	83
4.2.5.1.2 Preparation of stacking gel.....	84
4.2.5.1.3 Running Electrophoresis for 1D SDS PAGE.....	85
4.2.5.1.4 Staining gels and analysing proteins for 1D SDS PAGE.....	86
4.2.5.2 Two-Dimensional (2D) Sodium Dodecyl Electrophoresis (SDS-PAGE) Sulfate-Polyacrylamide Gel.....	86
4.2.5.2.1 Sample preparation for (protein loading) 2D Gels.....	86

4.2.5.2.2. Selection of appropriate technology for Isoelectric Focusing (IEF; First Dimension)	86
4.2.5.2.3 Rehydration of IPG Strips	87
4.2.5.2.4 First Dimension (IEF of IPG Strips)	87
4.2.5.2.4.1 IPG Strips Equilibration	88
4.2.5.2.5 Second Dimension SDS-PAGE	89
4.2.5.2.6 Detection of proteins by staining gels for 2D SDS PAGE	90
4.2.6 Image digitizing and analysis of protein spots	90
4.2.6.1 Spot imaging by Molecular Imager FX Pro Plus Multi-imager System	90
4.2.6.2 PDQuest analysis of 2D SDS-PAGE	90
4.2.7. Mass Spectrometry	91
4.2.7.1 Protein Identification using MALDI-TOF MS	91
4.2.7.1.1 Excision of Coomassie stained protein spots	91
4.2.7.1.2 Proteins in-gel digestion	91
4.2.7.1.3 Peptides extraction: Zip-tip procedure	92
4.2.7.1.4. Spot analysis using MALDI-TOF-MS	94
4.2.7.1.5 Searching for known protein sequences from databases	95
4.3 Results	96
4.3.1 Protein separation	96
4.3.1.1 One-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (1D-SDS PAGE)	96
4.3.1.2 Two-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (2D-SDS-PAGE)	96
4.3.2 Spot detection and protein differential expression analysis (PDQuest analysis)	96
4.3.3. Protein identification using MALDI-TOF-TOF-MS/MS	100
4.3.3.1 Protein spot analysis and spectral acquisition	100
4.3.4 Functional and subcellular classification of proteins	103
4.4 Discussion	108
4.4.1 Functional categories of the proteins identified using MALDI-TOF-MS/MS	108
4.4.1.1 Energy generation: proton (H <sup>+</sup> ) transporting protein	108
4.4.1.2 Glycolysis and other carbohydrate metabolism associated proteins	110
4.4.1.3 Regulation of photosystem (carbon assimilation)	113
4.4.1.4 RNA-Binding protein	115
4.4.1.5 Stress tolerance, defence and immunity related proteins	115



4.4.1.6 Unknown (Hypothetical proteins).....	116
4.5 Conclusion.....	117
Chapter 5: General discussion and conclusions.....	118
5.1 Summary.....	118
5.2. In silico Identification of candidate genes for drought tolerance.....	118
5.3 Gene-gene and gene-phenotype association in sorghum drought tolerance.....	120
5.4 Identification of drought responsive proteins using DE profiling and MALDI-TOF-TOF MS/MS .....	121
5.5 Conclusive remarks and future research plan.....	122
References.....	124
Appendices.....	193
Appendix 1.....	193
Appendix 2.....	195
Appendix 3.....	209
Appendix 4.....	211



UNIVERSITY *of the*  
WESTERN CAPE

## List of abbreviations

AAO:	ABA-aldehyde oxidase
ABA:	Abscisic Acid
ABRE:	ABA-dependent cis-acting element or ABA-Responsive Element Repeat
AREBs:	ABA-responsive element-binding proteins
BCAT:	Branched-chain amino acid transaminase
BCORFs:	Best Candidate Open Reading Frames
BCUCs:	Best Candidate UniGene Clusters
CBF:	C-Repeat binding factor
CDR:	Candidate Drought Responsive
CDRG:	Candidate Drought Responsive Genes
CGs:	Cyanogenic Glucosides
CGA:	Candidate Gene Approach
DigiCGA:	Digital Candidate Gene Approach
DR:	Drought Responsive
DREBs:	DRE-binding proteins
DRE/C-RT:	Dehydration Responsive Element/C-Repeat
DRESTs:	Drought Responsive ESTs
DRG:	Drought Responsive Genes
ESTs:	Expressed Sequence Tags
FDR:	False Discovery Rate
GEO:	Gene Expression Omnibus
GTA:	Gene Trait Association
GWA:	Genome Wide Association
HLC:	High Level Confidence
ICGBs:	Initial Comprehensive Gene Builds
InsCGA:	In Silico Candidate Gene Approach
MAPKs:	Mitogen-activated protein kinases
NAPRSU	National Agricultural Proteomics Research and Service Unit.
ICRISAT	International Crop Research Institute for Semi-Arid Tropics
MCSU:	Molybdenum cofactor sulphurase
NCED:	9-cis-epoxycarotenoid dioxygenase
ODRG:	Overlapping Drought Responsive Genes

PASA:	Program to Assemble Spliced Alignments
PMN:	Plant Metabolic Network
TDRESTs:	TIGR transcripts DRESTs
TFs:	Transcriptional factors
TIGR	The Institute for Genomic Research
Two DGE:	Two dimensional Gel Electrophoresis
VLIB:	Valine, Leucine, and Isoleucine biosynthesis
VLID:	Valine, Leucine, and Isoleucine degradation
ZEP:	Zeaxanthin oxidase



UNIVERSITY *of the*  
WESTERN CAPE

## List of tables

Contents	Pages
Table 2.1 Summary of sorghum transcript and genomic data.....	48
Table 2.2 Parameters used for the annotation comparison in the PASA pipeline.....	7
Table 2.3 Comparison and update of annotation between existing and current prediction.....	14
Table 2.4 Chromosomal distribution of the modified existing genes models.....	16
Table 2.5 Functional distribution of the novel gene structures model.....	17
Table 2.6 Distribution of novel genes based on the method of prediction.....	17
Table 2.7 Chromosomal distribution of the novel gene model.....	18
Table 2.8 Genomic distribution of spliced and retained intron based on PASA analysis.....	19
Table 2.9 Genomic distribution of skipped and retained exons based on PASA analysis.....	20
Table 2.10 PASA based identification of alternative splicing (AS) for the novel genes.....	20
Table 2.11 Exons and introns distributions for the novel genes in the sorghum genome.....	22
Table 2.12 Patterns of exonic and intronic features in the novel gene models.....	23
Table 2.13 Functional description of sorghum drought related metabolic pathways.....	26
Table 2.14 Functional GO enrichment of the genes involved in the metabolic pathways.....	27
Table 2.15 Description of the top ten interpro domains in decreasing order of frequency.....	28
Table 2.16 Description of the 45 significantly expressed maize genes under drought condition and the corresponding sorghum orthologs.....	32
Table 2.17 Sorghum orthologs and the corresponding genes from closely related species.....	33
Table 3.1 Gene-phenotype association based GO enrichment analysis.....	62
Table 3.2 Functional association of genes with different ontologies terms.....	64
Table 4.1 Solutions and buffers systems including storage condition and other description.....	79
Table 4.2 Bradford assay for protein quantification.....	82
Table 4.3 Preparation of resolving and stacking gels for SDS-PAGE.....	85
Table 4.4 Buffer system for equilibration of IPG Strips.....	88
Table 4.5 List of proteins identified by MALDI-TOF-TOF-MS/MS analysis.....	103
Table 4.6 Functional description of protein identified.....	104
Table 4.7 Sequence description of the identified peptide in the corresponding protein.....	106
Table S1.1 Justifiable parameters for staged clustering UniGenes.....	194
Table S2.1 Overview of UniGene libraries (build # 29).....	195
Table S2.2 Chromosomal distribution of UniGene clusters mapped to genome.....	196

Table S2.3 Comparison and update of annotation.....197  
Table S2.4 Databases containing potential candidate drought responsive genes.....197  
Table S2.5 Relevant tools for identification of the candidate gene.....198



UNIVERSITY *of the*  
WESTERN CAPE

## List of figures

Contents	Pages
Figure 1.1: Signal transduction, transcriptional regulation and ABA biosynthesis pathways networks in response to drought stress.....	7
Figure 1.2: The Core ABA signalling pathway (Source: adapted from Hubbard et al., 2012).....	8
Figure 1.3: Generic flow chart for cell signalling and gene regulation network.....	11
Figure 1.4: Pipeline for sorghum consensus gene prediction based on existing gene prediction data (Paterson et al., 2009).....	15
Figure 1.5: Conceptual strategic work flow for the candidate gene analysis and prioritization.....	26
Figure 1.6: An overview of the process of mRNA translation.....	27
Figure 2.1: Pipeline for mapping experimental data to reference genome and annotation comparison .....	2
Figure 2.2: Pipeline for building gene structure models.....	5
Figure 2.3: Schematic gene structure model for annotation comparison showing modified and novel genes.....	15
Figure 2.4: Pattern of exon and intron number and the average length.....	21
Figure 2.5: Intergenic distances between the novel and nearest existing gene structure model.....	24
Figure 2.6: Oxidative phosphorylation metabolic pathway.....	25
Figure 2.7: Heat map showing up and down-regulated sorghum orthologs based on maize RNA-seq expression data.....	29
Figure 2.8: Volcano plot showing differential expression of genes.....	30
Figure 2.9: Venn diagram showing distribution of significantly expressed genes.....	31
Figure 2.10: Sorghum orthologs correlating among species and drought related GO terms.....	34
Figure 3.1: Work-flow for gene-phenotype association across-species and stresses.....	51
Figure 3.2: Functional correlation and specificity of the drought tolerance with other stresses.....	55
Figure 3.3: Phylogenetic relationship of the sorghum genes with orthologs.....	58
Figure 3.4: Species specific and common drought responsive genes in a closely and distantly related species.....	59
Figure 3.5: Sorghum genes transitive association to multiple drought related terms.....	60
Figure 3.6: Hierarchical clustering of gene expression showing heatmap.....	61
Figure 3.7: Volcano plots showing gene expression.....	61
Figure 3.8: Scatter plot for semantic similarities in enriched GO-terms.....	63
Figure 3.9: Summarized description of drought related gene-trait associations.....	65

Figure 4.1: Sorghum sample control and stressed plants grown in green house conditions.....	78
Figure 4.2: summary of procedure for in-gel tryptic digestion of protein spots.....	93
Figure 4.3: PDQuest analysis of sorghum Btx642 leaf proteins for the 3 biological replicates.....	97
Figure 4.4: Representative 2DE showing spots selected for analysis of protein identification.....	98
Figure 4.5: Schematic representation of 2DGE differential expression of protein spot.....	99
Figure 4.6: Visualization of superimposed spots using multichannel viewer.....	100
Figure 4.7: Functional category of the proteins identified.....	105
Figure 4.8: Category of subcellular localization of the proteins identified.....	105
Figure 4.9: MASCOT probability distribution based on peptides (A) and protein score (B).....	105
Figure S1.1: Flow chart for sorghum staged clustering of UniGene-build procedure.....	193
Figure S2.1: Predicted overlapping gene annotation and characterization status.....	200
Figure S2.2: GO annotation based on blasting and mapping to non-redundant databases.....	200
Figure S2.3: The 13 metabolic pathways among 14 identified (Figures S2.4A – G).....	207
Figure S2.4: Description of interpro domain analysis: List of protein signatures identified.....	207
Figure S2.5: Sorghum % GO-terms assigned genes identified from maize orthologs.....	208
Figure S2.6: Mapping of GO terms related to responses to stress based on biological process.....	208
Figure S3.1: Graphical views of significantly enriched GO-terms.....	209
Figure S3.2: GO-annotation Vs % gene enriched in the particular GO-domain.....	209
Figure S3.3: Interactive association of genes for enriched GO-terms.....	210
Figure S4.1: MALDI-TOF-TOF-MS/MS spectrum of protein spot in-gel tryptic digest.....	211



**Chapter 1:** General Introduction and Literature Review

**Abstract**

Drought stress is the most complex phenomena that disrupts a normal biological functioning of the plant. Because of the complexity in genetic and physiological conditioning and the environmental influences, the progress towards drought tolerance has been very slow. Thus advances in the study for drought tolerance requires an identification and detailed analysis of many and possibly all components of the complex biological processes. Data sets consisting but not limited to genomics and proteomics provide basis for systems biology the integration and handling of which through modeling will be the best way to arrive at a predictive level for improving drought tolerance. This chapter attempts to indicate the challenges that drought causes to human welfare and describes sorghum (*Sorghum bicolor* (L.) Moench) as a model crop in assisting the combat against this challenge as a target material in genomic and proteomic research. This chapter provides general review on the main topics that will be incorporated in the thesis, outlining the main components and giving brief description on each. Because genomic and proteomic analysis of drought tolerance is the heart of the thesis, this chapter gives review with due attention on candidate gene identification, highlighting procedures for *in silico* and experimental analysis. It highlights computational pipeline for candidate gene identification and prioritization. It gives details of gene structure prediction, role of ESTs as tool for gene discovery and UniGene build procedure. It provides description on signal transduction and transcriptional regulation of drought associated genes; provides brief review on identification of proteins differentially expressed under drought stress conditions.



## 1.1 Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is the fifth most important cereal crop worldwide. It is grown in rain fed lowland and semi-arid tropics with remarkable tolerance to adverse conditions. It holds a key position in the North-East and Southern Africa where it is grown on an average area of more than seven million hectares per year (Rao *et al.*, 2011). In Southern Africa sorghum has played an important role in the development of food security where it is largely grown by small scale farmers. In South Africa, though it is expected to find unreported landraces cultivated by non-commercial farmers, sorghum is mainly grown on a commercial level and utilized on a wider scale than other countries in the Southern African Development Community (SADC) (Rao *et al.*, 1989; Bernstein, 2013).

Worldwide production of sorghum is over 63 million metric tones of grains annually from over 44 million hectares of land with an average yield of 1.4 metric tones per hectare (Sasaki and Antonio , 2009). Drought contributes heavily to the constant food insecurity and rampant poverty characteristics of sub-Saharan Africa. Drought coupled with global climatic trend is becoming a bottle-neck for crop yields worldwide. This challenge on top of the exponential growth of world population aggravated the demand for limited fresh water supplies for crop production (Sorghum Genomics Planning Workshop Participants; SGPW, 2005). This suggests that dry land and water-use efficient crops such as sorghum will be of increasing importance to enhance food production and livelihoods of poor farmers (Bennetzen, 1997; and Nguyen, 2000). Sorghum is grown under sever moisture stress and incredibly adapted to marginal condition compared to most other crops grown in low rainfall areas of the world (Doggett, 1988; Doggett, 1991; Rosenow *et al.*, 1996). This uniqueness is an attribute of the presence of important genes and useful alleles that enable the crop to survive in arid environments (Dogget, 1988). Despite its importance, relatively few studies have been undertaken to develop drought resistant varieties of sorghum. It is imperative to develop strategies to harness existing and emerging sciences to exploit the potential of this crop to reduce or limit drought stresses and to promote food and economic security. Genomics and proteomics provide new approaches that may allow relatively rapid progress in producing crops with improved drought tolerance (Bennetzen, 1997). Thus, improving drought tolerance in sorghum would increase and stabilize grain and food production in drought affected areas of South Africa and other parts of the world.

Therefore, identification of drought tolerant genes and gene products in sorghum is of significant importance for further development of the potential germplasm and their utilization in the commercial sectors and breeding programs. Hence, this project aims at investigating genes and gene products that respond to drought stress employing integrated genomic and proteomic approaches. Application of bioinformatics to develop a comprehensive dataset of gene function that will serve as a powerful reference of protein properties and functions towards the molecular understanding of drought tolerance in sorghum is also a target.

## **1.2 Drought**

### **1.2.1 What is drought?**

Though difficult to define, drought is a prolonged and abnormal moisture deficiency. It is an insidious hazard of nature resulting when there is less precipitation than normal over an extended period, usually a season or more (Brewer and Heim, 2011; Fu and Tang, 2013). Drought is among environmental factors but the most likely disastrous abiotic stress (Slettebak, 2012). According to Wilhite and Glantz (1985), and latter reviewed by Mishra (2010), drought is defined as a conceptual phenomenon which represents prolonged period of less precipitation than normal resulting in extensive damage to crops and as an operational thought that speculates the extent of drought in a given geographical location where the beginning, severity, and end of droughts would be set. Based on the differences in regions, needs and disciplinary approaches, the definitions reflect operationally four basic approaches for identifying and measuring drought, namely: meteorological (long-term region-specific drought characteristic), agricultural (moisture deficient soil fails to meet needs of crop at a time (McKee *et al.*, 1993), hydrological (truncation level of water reserve causing deficiencies in surface and subsurface water supplies (Tallaksen *et al.*, 2004; Shiao *et al.*, 2007) and socio-economic (effects of water shortfall rippling through socio-economic systems). All the other approaches deal with measuring drought as a physical phenomenon, whereas a socio-economic measure drought in terms of supply and demand (Yuan and Zhou, 2012).

### **1.2.2 Drought: “A primarily African problem”**

Drought is a major environmental stress that would adversely affects crop yield and quality. The harmful effect of drought stress on crop productivity is increasingly implicated by a global rise in temperature level. This is exacerbated by human activities such as over-utilization of water resources, over-irrigation, improper drainage, besides natural causes. Among other stresses, cold,

salinity and drought are major stresses which adversely affect plant growth and productivity. Drought and salinity have osmotic, ionic and nutritional constraint effects on plants (Tuteja and Mahaja, 2007; Mahajan *et al.*, 2008). Water is the main limiting factor to plant productivity in arid and semi-arid zones where approximately four-tenths of the world's agricultural land lies and which still a lot of variables contributing to the drought conditions (Fensholt *et al.*, 2012). Erratic rainfall, degradation of land, loss of biodiversity, shift in a vegetation cover to agricultural lands, and global warming are among those variables eliciting drought. This contributed tremendously to the virtual decline of world food production from time to time compared to the ever increasing rate of world population and the upcoming alert from increased temperature (global warming) with inescapable consequences. This urgently necessitates to look out of box to seek the long term solution. Developing and enhancing stress tolerant crops using modern approach is one of the mitigating action among others.

Interaction between Plant drought stress response and resistance are complex biological processes that need to be analysed at a systems level using functional genomics and proteomics in addition to the traditional physiological approaches to dissect experimental models that address drought stresses encountered by crops in the field.

Drought combined with current global climatic changes is unavoidable challenge for the modern agriculture particularly in Africa. Genomic approaches in combination with proteomics will likely be needed to significantly improve drought tolerance in crops and to provide understanding and subsequent utilization of stress response and acclimation networks (Mittler and Blumwald, 2010). With due care to the experimental conditions, our work was designed to investigate sorghum drought tolerance that include all the above variables in addition to *in silico* identifications.

### **1.2.3 Mode of responses and mechanisms of drought tolerance**

Plant responses to most stresses comprises physiological or morphological changes. The effect of stress is usually manifested with retarded photosynthetic and growth rate at the whole plant level. This is usually perceived in association with alteration in carbon and nitrogen metabolism (Cornic and Massacci, 1996; Deeba *et al.*, 2012).

According to Arraudeau, 1989, Physiological mechanisms of drought tolerance could be grouped into three major categories. These are i) accumulation and translocation of assimilates, (ii) osmotic

adjustment, and (iii) maintenance of cell wall elasticity. However, in addition to this, mechanisms such as biochemical, genetic and molecular and all the basis behind these mechanisms including Reductive Oxygen Species (ROS), Antioxidants and Detoxification genes and Cell Membrane Stability (CMS) as a measure of drought (Pareek, 2010) are all important interplay that have prominent roles affecting plant survival. Sorghum as a plant mainly grown in drought affected regions have evolved a wider basis of adaptive mechanisms that accounts for either resistance to or escape from drought (Price *et al.*, 2002). Drought-escaping is the early flowering of the plant before drought occurrence while drought resistance could be manifested in various ways as identified by price *et al.*, 2002 such as drought avoidance by mechanism of maintaining tissues water potential and drought tolerance which could further be identified as dehydration avoidance and dehydration tolerance. However, for a holistic perception of plant resistance, understanding the way of their interactions and responses to drought with respect to the above mentioned mechanisms particularly biochemical and molecular responses is essential.

The process of responses and mechanisms of drought tolerance requires initial signal with sufficient amount to generate signalling molecules that allow signal transduction cascade to effect transcriptional factors. This on one hand, mediates the Abscisic Acid (ABA) biosynthesis and on the other alters growth and development at all plant level regulating the expression level of target genes. The mechanistic process flow of the plant drought response according to Reddy *et al.*, 2004 include: Sub-or-supra-optimal environmental stimuli → Sensing mechanisms → Local and Global (long distance) signals → Signal transduction → Genomic and proteomic (post-genomic) responses → Alterations in cytoplasmic and apoplastic metabolism → Altered growth and development at level of cell, organ and organism → Acclimation (Adaptation/ Adaptive adjustment) or stress-induced death (see detail in section 1.4; Figures 1.1 and 1.).

### **1.3 Signal transduction and Transcriptional regulation network**

#### **1.3.1 Signal transduction cascade**

Stress related activation of signalling cascade of molecular networks is the basis for plant adaptation to environmental influences. Plant adaptation involves a chain of processes that include stress signal perception, signal transduction, activation of transcription factors and ABA biosynthesis, the expression of specific stress-induced genes and gene products and other related metabolites (Figure 1.1). Because of the complexity in signal transduction pathway (Atkinson *et al.*, 2012), yet this process has not been fully demonstrated from the perception of the stress to the adaptive response

manifested by gene expression in plant cells where enhanced access is achieved to plant survival. Most abiotic stresses are mutually exclusively related with an exception of ABA that commonly regulates the main pathway that effects the response to these stresses (Pareek, 2010). Drought is among environmental factors but the most likely disastrous abiotic stresses (Slettebak, 2012). The generic concept, however is that there is the common regulatory system and cross-talk among abiotic stresses, with particular emphasis on the Mitogene Activated Protein Kinase (MAPK) cascades and the cross-talk between ABA and abiotic signalling (Haung *et al.*, 2012). Survival of plants therefore depends on the extent of their adaptation to respond to these stresses at the molecular, cellular, physiological and biochemical level given the deferential rate of expression patterns of dehydration inducible genes (Nakashima *et al.*, 2009). Plants have developed adaptive responses both to multiple or specific environmental stress (Huang *et al.*, 2012). It is also practical for plants to respond to stresses as individual cells and synergistically as a whole organism (Tuteja, 2007; Bansal *et al.*, 2011; Chawla *et al.*, 2011).

The generic signal transduction pathway, as reviewed in Xiong *et al.* 2002; Mahajan and Tuteja, 2005; Hubbard *et al.*, 2012 and Haung *et al.*, 2012, is described briefly as follows. Signal transduction pathway begins with extracellular stress signal perception by the membrane receptors. This activates downstream intracellular signal cascade allowing the generation of the second messenger (secondary signalling molecules) such as calcium, Inositol phosphate (IP) and Reactive Oxygen Species (ROS; Figure 1.1). This further modulates and triggers up-regulation of cytoplasmic calcium level which in turn sensed by calcium binding proteins ( $\text{Ca}^{+2}$  sensors) that eventually change their conformational structures in the presence of calcium ion as they lack any enzymatic activity. Interacting with their down stream signalling components (eg. kinases and/or phosphatases), these sensors then initiate a respective transduction (eg.phosphorylation) cascade to target the major stress responsive genes or the transcription factors regulating these genes. Then responsive expression of these genes result in important gene products that involve in plant adaptation and survival. Chemicals such as ABA, salicylic acid, ethylene are also produced as a consequence of stress induced change in gene expression. These molecules may involve in second round signalling cascade not necessarily following the same initial pathway (Larkindale *et al.*, 2002). Other molecules like protein modifiers though involve in the modification and assembly of signalling components for effecting myristoylation, glycosylation, methylation and ubiquitination may not directly participate in signalling (Xiong *et al.*, 2002).

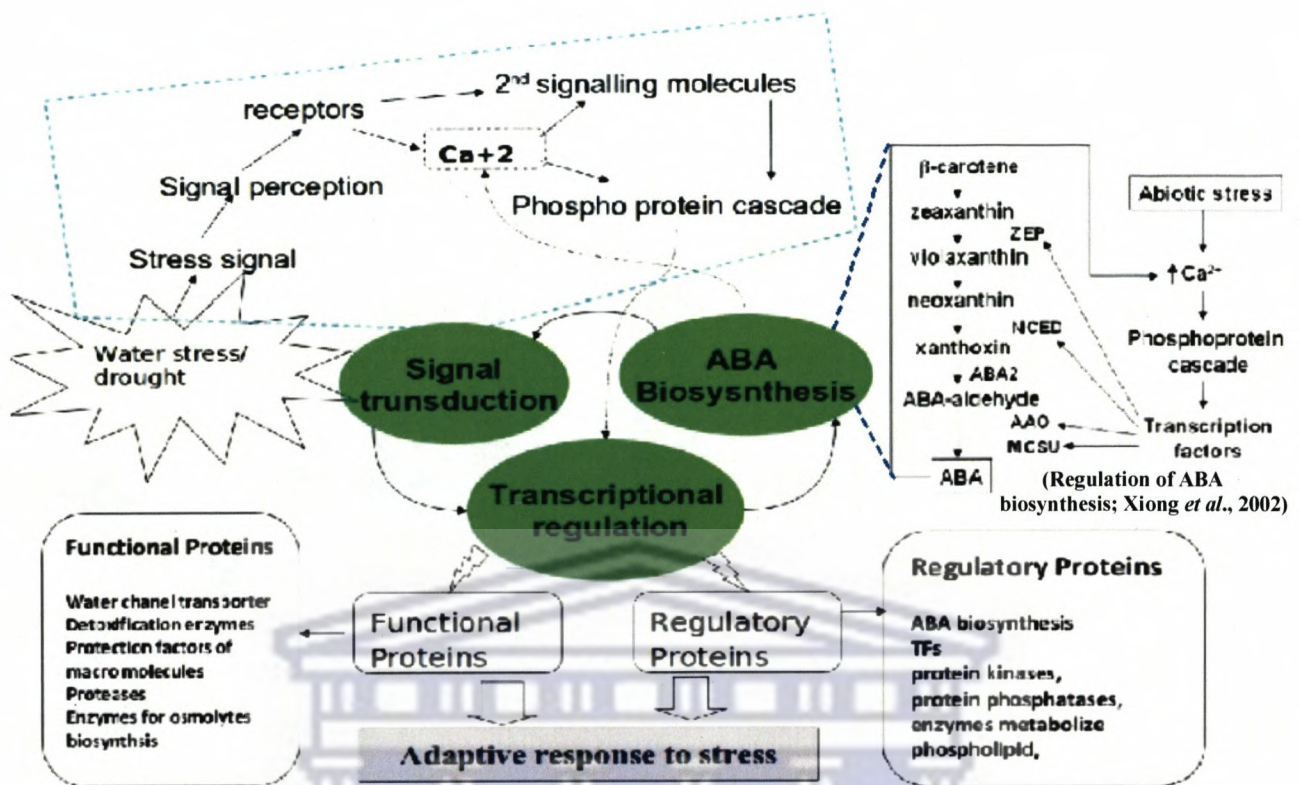


Figure 1.1: Signal transduction, transcriptional regulation and ABA biosynthesis pathways networks in response to drought stress.

Stress inducible genes are categorized based on the duration of changes in expression since they first received stress-signal into early (those induced instantly and often expressed transiently) and late (slow inducible genes and often exhibit a prolonged and sustained expression level (Cramer *et al.*, 2007). It was previously known that early genes encode for the transcription factors that activate the major stress responsive genes (delayed genes) the expression of which result in the production of various osmolytes, antioxidants, molecular chaperones and LEA-like proteins, which function in stress tolerance (Cheong *et al.*, 2002).

### 1.3.2 Role of ABA in dehydration tolerance response

The phytohormone ABA plays major role in plant responses to stress. While ABA is rapidly synthesised at time of stress, it is equally rapidly dissociated at time of relive (Huang *et al.*, 2012) as the later may lead to inhibition of normal growth of plants if not otherwise. It has been shown that ABA is produced under water-deficit conditions causing stomatal closure and induction of expression of genes which play important role in the tolerance response of plants to abiotic stresses (Figure 1.2; Yamaguchi-Shinozaki and Shinozaki, 2006). It has also been shown that exogenous

application of ABA induces most of the genes that respond to drought, salt, and cold stress (Shinozaki *et al.*, 2003; Zhu 2002). However, studies have indicated that not all genes which respond to stresses are induced by exogenous application of ABA (Zhu, 2002; Shinozaki *et al.*, 2003; Yamaguchi-Shinozaki and Shinozaki, 2005).

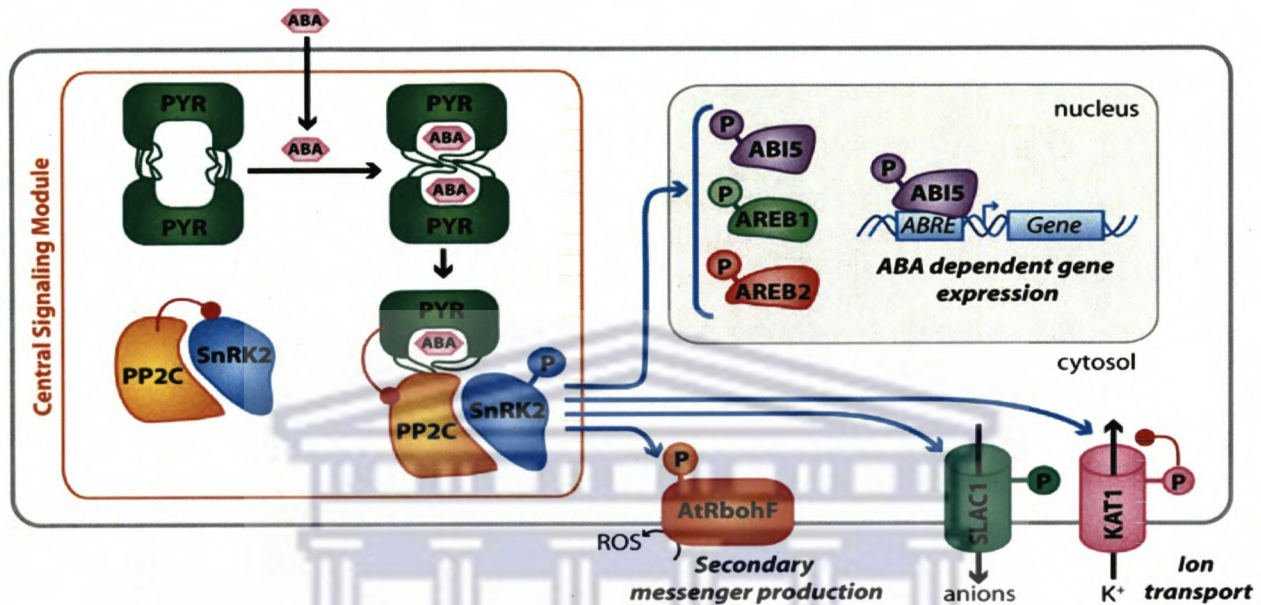


Figure 1.2: The Core ABA signalling pathway (Source: adapted from Hubbard *et al.*, 2012)

This suggests the existence of both ABA-independent and ABA-dependent signal transduction cascades that lie between signal perception and the expression of specific genes. Some genes, for example RD29A, are reported to be regulated in both ABA-dependent and ABA-independent pathways (Figure 1.3; Yamaguchi-Shinozaki and Shinozaki, 1993; Yamaguchi-Shinozaki and Shinozaki, 1994, 2005; Zhang *et al.*, 2012).

### 1.3.3 Transcriptional Regulatory Networks

Understanding the regulatory interactions and interconnections between genomics, transcriptomics, proteomics and metabolomics is not only important in the discovery of the genes and gene products and assigning functions to these genes and their products. It is also crucial in understanding the behaviour of complex biological systems and processes in terms of their molecular constituents (Kirschner, 2005). Molecular mechanisms governing gene expression patterns have been studied on model crops such as *Arabidopsis* (*Arabidopsis thaliana*; Yamaguchi-Shinozaki and Shinozaki, 2006) and rice (*Oryza sativa*; Qin *et al.* 2007 and Khan, 2012) in response to dehydration and cold stresses. Molecular basis of stress tolerance mainly rely on active control of transcription factors either collectively or independently, and/or constitutively over expressing the target genes through

binding to the cis-acting element in the promoter region of the same target genes (Lee *et al.*, 2002; Nakahima *et al.*, 2009). This suggests that cis-acting elements (transcription factor binding sites, or DNA-binding element) are the major evidence for transcriptional regulatory networks and for the cross-talk among abiotic stresses (Knight and Knight, 2001).

Based on the transactivation assays, plant-specific TFs function as transcriptional activators (Lu *et al.*, 2012) and these belong to families defined by their characteristic DNA-binding domains (DBDs) such as AP2/ERF, B3, NAC, SBP, and WRKY (Yamasaki *et al.*, 2012). Transcriptional factors regulate gene expression by inducing activators or repressors which is the activity of the RNA polymerase (Ciarmiello *et al.*, 2011). Major plant-specific TFs that are active in response to abiotic stress have been identified (Nakashima *et al.*, 2009) in many cereal crops. For example AP2/ERF, a large group of plant-specific TFs that includes four major subfamilies: the AP2, RAV, ERF and dehydration-responsive element-binding protein (DREB) subfamilies (Sakuma *et al.*, 2002), first identified in Arabidopsis homeotic gene APETALA 2 (Jofuku *et al.*, 1994), and a similar domain ethylene-responsive element binding proteins (EREB) was found in tobacco (*Nicotiana tabacum*; Ohme-Takagi and Shinshi, 1995). Recently, a maize stress-responsive NAC transcription factor, ZmSNAC1 was proven to confer enhanced tolerance to dehydration in transgenic Arabidopsis (Niu and Bate, 2010; Lu *et al.*, 2012). Though detailed description of its role in stress-responsive was not indicated, 15 distinct clusters of NAC subfamilies have been phylogenetically identified by genome-wide survey and characterization of green-bug induced NAC transcription factors in sorghum (Zhang, 2013). The roles of B3, NAC, SBP, and WRKY TFs in stress signal transduction and transcriptional regulation have been well reviewed in Corrêa *et al.* (2008); Agrwal and Jha (2010); Mizoi *et al.* (2012) and Lu *et al.* (2012).

According to Riechmann and Ratcliffe (2000); Wingender *et al.* (2001) and Warren (2002), classification of TFs into different protein families is based on the following two criteria: their primary and/or three-dimensional structure and similarities in the DNA-binding and multimerization domains. The extent of binding of transcription factors (TFs) to these DNA-binding sites switches or effects the expression pattern of dehydration inducible genes such as RD29 (Mahdevar *et al.*, 2012). This means that gene expression and its regulation are dependent, to a great extent, on the binding efficiency of TFs on to the TFBS (Mahdevar *et al.*, 2012).

There are four independent transcription regulatory system for gene expression which involve two



major cis acting elements, ABA-independent cis-acting element (Dehydration Responsive Element (DRE/C-RT or C-Repeat) and ABA-dependent cis-acting element (ABRE or ABA-Responsive Element Repeat; Jia *et al.*, 2012). Figure 1.3 shows the generic map of the transcriptional regulation network and signal transduction cascade and the gene regulation under drought condition for ABA-dependent and independent pathways. On the other hand, following the two major groups of cis acting elements, dehydration inducible TFs are likely divided into two main groups namely AREBs (ABA-responsive element-binding proteins) and DREBs (DRE-binding proteins) such that each group further divided into subgroups based on the type of signalling pathways they involve in (Yamaguchi--Shinozaki and Shinozaki, 1993; Shinozaki *et al.*, 2003; Shinozaki and Yamaguchi-Shinozaki, 2007; Nakashima *et al.*, 2009; Mizoi *et al.*, 2012).

ABA-responsive element-binding proteins is activated under ABA-dependent signal transduction pathway whereas DREBs (DRE-binding proteins) is activated under ABA-independent signal transduction pathway. DREB proteins further sub-grouped in to DREB1/C-repeat binding factor (CBF) and DREB2 (Nakashima *et al.*, 2009). Both of these TFs belong to the ERF/AP2 family and actively bind to their common conserved DNA-binding motif, A/GCCGAC (Shinozaki and Yamaguchi-Shinozaki, 2007). The fact that the DREB1/CBF is rapidly and transiently induced by cold stress resulting in the regulation of expression of target cold stress-inducible genes mainly distinguishes between the two (Shinozaki and Yamaguchi-Shinozaki, 2006; Shinozaki and Yamaguchi-Shinozaki, 2007; Nakashim *et al.*, 2009; Mizoi *et al.*, 2012). Their function in the development of cold-stress tolerance doesn't necessarily require post-translational modification as proven through transgenic plant experiment (Liu *et al.*, 1998). It was shown that most of the CBF/DREB1 regulated target genes contain a conserved DRE motif with (A/G) CCGACNT sequence (Shinozaki and Yamaguchi-Shinozaki, 2007) in their promoter regions. On the contrary, DREB2 genes are stimulated by drought stress and are responsible for the activation and controlling of expression of target drought stress-inducible genes (Liu *et al.*, 1998).

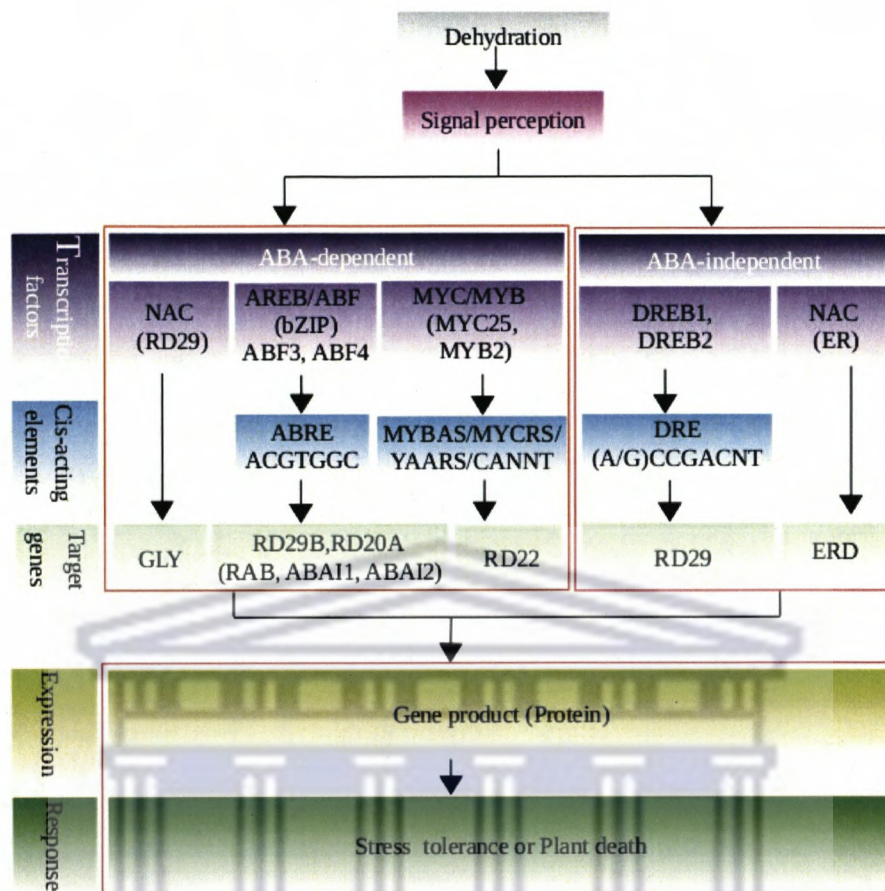


Figure 1.3: Generic flow chart for cell signalling and gene regulation network

This figure shows a generalized flow chart for cell signalling and gene regulation network under dehydration condition based on ABA-dependent and independent pathways.

Unlike CBF/DREB1 genes, the developmental function of DREB2 genes in drought stress tolerance involves post-translational activation (Liu *et al.*, 1998). However, DREB1 (also called CBF) and DREB2 involve in cross-talk between cold and drought signalling since they share the same *cis*-acting element, DRE element (also known as *LTI78*) in the promoters of genes such as *RD29A* (Knight and Knight, 2001). This suggests that DRE element is a common site for cross-talk between cold and drought signal transduction pathways. DREB2 genes include but not limited DREB2A and DREB2B which are thought to be the major TFs that function under drought and high-salinity stress conditions (Sakuma *et al.*, 2006). DREB2A protein was known to be activated by post-translational modification such as phosphorylation (Liu *et al.*, 1998). A constitutively active form of DREB2A was demonstrated in Arabidopsis that transactivates the target drought inducible genes (Sakuma *et al.*, 2006).

Similarly, a common target gene, RD29, is known to be regulated by the two different TFs, AREBs

and DREBs through binding to two separate binding elements called ABRE and DRE respectively (Agarwal and Jha, 2010 and Zhang *et al.*, 2011). On the other hand, there seems a phenomenon where ABA-independent transduction pathway regulates the ABA-dependent pathway by way of a transactivational integration of the four major genes that involve in the ABA biosynthesis, namely: Zeaxanthin oxidase (ZEP), 9-cis-epoxycarotenoid dioxygenase (NCED), ABA-aldehyde oxidase (AAO) and molybdenum cofactor sulphurase (MCSU; Figure 1.3; Xiong *et al.*, 2002; Tuteja, 2007).

It has been shown that ABA is involved in regulating many aspects of plant developmental patterns such as seed germination and maturation, desiccation tolerance and seed dormancy and hence plays an integral role in the plant's response to drought stress. This is probably associated with the fact that plenty of the drought-inducible genes identified and studied to date are also induced by ABA (Wang *et al.*, 2003). Identification of the genes involved and understanding their roles during stress perception and physiological regulation has become an important and exciting research field in recent years (Zhang *et al.*, 2006).

#### **1.4 Sorghum: “A drought-hardy and agriculturally important crop”**

Sorghum is the most drought-hardy crop that suitably thrives under rain-fed environments (Nagaraj and Rao, 2011; Rao *et al.*, 2011). Hardy crops thrive in adversely drought affected regions with minimal production risky. Early analysis on physico-chemical nature of drought resistance in crop plants by Newton and Martin (1930) and later likely supported by Ben-Hammouda *et al.* (1995) and Yu *et al.* (2003) shows that sorghum contain relatively higher drought-hardy varieties. Comparative evaluations also show that sorghum is drought-hardy than most drought grown crops (eg. rice) and produces well with relatively little moisture (McClain, 1997). Furthermore, sorghum has stay-green genes that potentially contribute to drought hardiness and yield productivity than a close relative grass, maize (Subedi and Ma, 2005). This suggests that sorghum is an agriculturally important crop a choice for which substitutes are so limited in sub-Saharan Africa and most arid and semiarid regions thus continue feeding millions of resource poor.

##### **1.4.1 Sorghum Metabolic pathway: C4 crop**

Sorghum is one of the only two major food crops (including maize) that evolved C4 photosynthetic pathway (Long 1998; Brutnell *et al.*, 2010). Mainly because of the advantage of eliminating energy loss in photorespiration that C3 pathway fail to exhibit, C4 plants are the world most productive

(Long *et al.*, 2006) compared to the C3 plants. A comparative analysis of the key photosynthetic enzyme genes in sorghum, maize (C4) and rice (C3) signified the contribution of duplication both at the whole-genome and at the individual gene level for the evolution of C4 pathway (Wyrich *et al.*, 1998; Wang *et al.* 2009). In total, there has been multiples of independent evolution of C4 pathway at the time of angiosperm evolution (Edwards *et al.*, 2001; Surrige, 2002) with multiple origins (Gaut, *et al.*, 1997; Swigonova *et al.*, 2004) implying genetic predisposition in some C3 plants to C4 evolution (Wang *et al.*, 2009; Paterson *et al.*, 2010). This suggests that C4 pathway may be the only adaptive feature to use among plants for avoiding high energy loss in photorespiration (Long *et al.*, 2006).

#### **1.4.1.1 C4-pathway: A secret behind drought-hardy sorghum**

The secret behind exceptional hardiness of sorghum and its thriving to adverse agro-ecological situations lies on C4-pathway (Britton, 2003; Gowik and Westhoff, 2011). However, other drought and crop related factors may also contribute to this fact (Manavalan *et al.*, 2012). Sorghum has characteristic xeromorphic features (Britton, 2003) such as dense root system (for efficient water absorption), thick waxy cuticle (for efficient reduction of water loss), rolling leaf in dry condition (for trapping moisture and reduce transpiration), low number of stomatal sunken in leaf surface (keeps the transpiration rate low but allowing gas exchange).

#### **1.4.1.2 Sorghum Metabolic pathway: SorghumCyc**

Sorghum-specific metabolic pathway database was constructed as the first released (SorghumCyc) at the pathways section in the Gramene metabolic pathway databases. From here users can search for genes, proteins, enzymes, reactions, metabolites and can upload and analyse high-throughput expression data and generate cellular overview of pathways using Omics-viewer tool (Naithani, 2013). SorghumCyc is one of the eight species metabolic pathway databases that Gramene hosts and it provides 328 pathways (Youens-Clark *et al.*, 2011) which are suited for web-based browsing as well as for bulk downloading in several options including the BioPax (Demir *et al.*, 2010) and Systems Biology Markup Language (SBML) (Gauges *et al.*, 2006) formats for advanced users. The annotated pathways are used as external references in the sorghum genome browser (Youens-Clark *et al.*, 2011). However, many of the pathways might be incomplete or may contain errors based on the fact that functions of many of the sorghum genes are either provided by homology and/or HMM based predictions (Youens-Clark *et al.*, 2011; Naithani, 2013).

However, according to Plant Metabolic Network (PMN), it has been shown that *Sorghum bicolor* has 428 pathways, 2802 enzymatic reactions of which 41 are transport reactions, 392 pathway holes (reactions within metabolic pathways of *Sorghum bicolor* for which no corresponding enzyme has been identified in the genome) were identified. Similarly, 8615 enzymes of which 316 are transporters, 2155 compounds and 1 complex have also been identified with no transcription units and tRNAs (Chae *et al.*, 2013).

#### **1.4.2 Sorghum genome sequence: opportunity**

Sorghum is only the first among the few major food crops that evolved C4 photosynthetic pathway and is only the second after rice among grass family whose genome has completely been sequenced. It's approximately 730 Mb genome size (Bennett and Leitch, 2005; Paterson *et al.*, 2009) which is much larger than the genome of its distant relative, rice from which sorghum diverged 42 mya (Paterson *et al.*, 2004; Paterson, 2008; Bolot *et al.*, 2009). However, it is by far a small, diploid genome compared to its closest relatives maize (*Zea mays*) and sugarcane (*Saccharum officinarum*), both of which have much larger polyploid genomes (Arumuganathan and Earle, 1991). The completion of sorghum genome sequencing is undoubtedly an opportunity through which further characterization of the genomes other than Saccharine cereals may shed light on mechanisms, levels, and patterns of evolution of genome size and structure. This will lay the foundation for further studies on sugar-cane and other economically important members of the group (Paterson, 2008; Paterson *et al.*, 2009). With the genome having been sequenced, sorghum consensus gene predictions were built around several evidence sources (Paterson *et al.*, 2009; see also prediction pipeline, Figure 1.4).

##### **1.4.2.1 Sorghum consensus gene prediction**

The principal products of any genome project involve at least three main components including the genomic sequence itself, the genes and the map integrating the genes (Zhuo *et al.*, 2001). This three basic products are applied to and included in the sorghum genome projects though the extent of enrichment of each product varying from genome to genome project. Here we briefly describe the consensus sorghum gene prediction procedure providing the pipeline (Figure 1.4) that illustrate the prediction procedure based on the original information in Paterson *et al.* (2009) and on the overview of the consensus gene prediction.

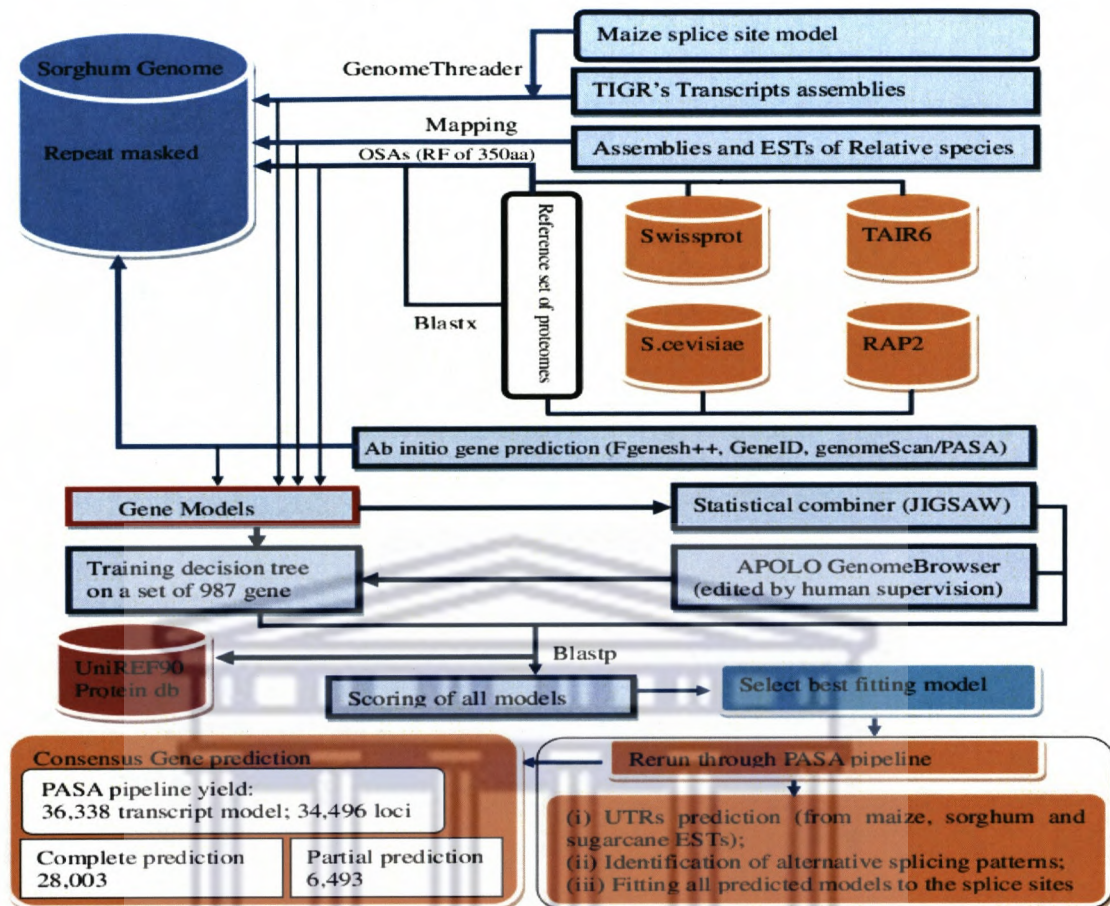


Figure 1.4: Pipeline for sorghum consensus gene prediction based on existing gene prediction data (Paterson *et al.*, 2009).

Sorghum gene prediction has passed through a series of process that include three major steps as follows (Figure 1.4):

First, evidence sources for the consensus gene prediction were organized;

Second, gene models were identified and sorted. Result from this pipeline yielded 36,338 transcript models at 34,496 loci out of which 28,003 were complete gene model (complete gene structure) whereas 6493 were partial, candidate genes lacking a start and/or stop codon included in annotation if they were not overlapping with complete predictions (Paterson *et al.*, 2009). Partial gene models may result from several, not mutually exclusive reasons: (i) sequencing or assembly errors (Venter *et al.*, 2001); (ii) transposon activity; Brennecke *et al.*, 2007) (iii) insufficient evidences from *ab initio* predictions or expressed sequence tag (EST) matches (Curwen *et al.*, 2004).

Thirdly, UTRs were identified by utilizing the gene prediction set using PASA with all available sorghum ESTs to run producing 842 alternatively spliced alignments that increased UTR to 17,744 transcripts (Paterson *et al.*, 2009).

### 1.4.2.2 Sorghum Genomics

Genetic and physical mapping as well as genome sequencing and the future perspective for the enhanced progress in genome characterization at the post genomic era has been well addressed in Paterson *et al.* (2008). Well-developed physical and genetic maps and large bacterial artificial chromosome (BAC) libraries are already available to facilitate the progress in sorghum genomics (Klein *et al.*, 2000; Menz *et al.*, 2002; Paterson, 2008; Paterson *et al.*, 2009). The availability of complete sorghum genome sequence has laid foundation for progress in functional genomics (Paterson, 2008; Paterson *et al.*, 2009) and this feature combined with the relatively small genome size provides basis for understanding of the structure, function and evolution of the grass genomes. Sorghum genome remains unduplicated at the whole-genome level compared to its cereal closest relative maize which has undergone one whole-genome duplication (Swigonova *et al.*, 2004) since the two crops were diverged 12 mya (Gaut, *et al.*, 1997; Swigonova *et al.*, 2004) and to its saccharine member *Saccharum* (sugar-cane) that has undergone genome duplication at least twice (Ming *et al.*, 1998).

### 1.4.2.3 Sorghum proteomics

Proteomic analysis seemingly is a current focus particularly on tagging genes for complex traits such as drought tolerance mainly based on the fact that the fate of gene expression is determined at least partially by post-translation modification to which most proteins are subjected (Bevilacqua *et al.*, 2003) and that it is difficult to monitor regulation of gene expression at transcription level. Several studies have been conducted to analyse changes in proteomes in response to salinity and drought stress such as Peng *et al.*, 2009 in Bread Wheat; Merracini *et al.*, 2012 in coffee and Ndimba *et al.*, 2010; Ngara and Ndimba, 2011; Ngara *et al.*, 2012 and Ndimba and Ngara, 2013 in sorghum. However, compared to genomics, proteomic studies conducted on plants in general and on sorghum in particular is limited in reference to drought tolerance. Recently, studies that target salinity and drought stress have been done using proteomic techniques and Mass Spectrum (MS) on sorghum as a focus organism (Ndimba *et al.*, 2010; Ngara and Ndimba, 2011; Ngara *et al.*, 2012 and Ndimba and Ngara, 2013). Studies to dissect the complex traits using proteome analysis may help target candidate proteins dissecting the genetic foundation of the variation of quantitative traits, and essentially using for validation purposes.

While drought primarily generates osmotic stress, salinity induces osmotic stress through effects on the ionic homeostasis within the plant cell (Zhu, 2002) where in both cases identification of stress-responsive proteins are made possible as a result of physiological and biochemical defence actions

(Salekdeh *et al.*, 2002; Ndimba *et al.*, 2005). Physiological and biochemical mechanisms are among strategies by which plants respond to various stress signals. As to the latter one, drought stress induces the accumulation of important biochemical components (proteins) which are produced as a result of the expression of drought-inducible genes (Seki *et al.*, 2003). These genes are functionally classified into two groups (Shinozaki *et al.*, 2003). The first group includes those proteins that probably directly involve in protecting plants from dehydration and are termed as functional proteins such as the detoxification enzymes (enzymes required for the biosynthesis of various osmoprotectants such as proline, sugar and sugar alcohol), proline, sugar transporters, water channels, protection factors of macromolecules such as late embryogenesis abundant proteins (LEA), chaperones, and enzymes that conducts proteolysis (proteases/peptidase/proteinases) (Seki *et al.*, 2003). The other group of proteins which provide regulatory functions in the signal transduction and the expression of drought inducible genes are known as regulatory proteins (Seki *et al.*, 2003). These include various transcription factors, protein kinases, protein phosphatases, enzymes involved in phospholipid metabolism, and other signalling molecules such as calmodulin-binding protein, mRNA binding (Seki *et al.*, 2003; Shinozaki and Yamaguchi-Shinozaki, 2006). The combinatorial effect of the two groups of gene products build adaptive responses in plant to stresses such as drought. This probably suggests the possibility of selecting drought tolerance biomarkers as a source of potential adaptive traits. Drought tolerance biomarkers represent biochemical compounds found in the plant biological tissues which are directly or indirectly associated with the presence and progression of a drought stress (Jorrín-Novo *et al.*, 2009). Thus, a comprehensive approach that include genomic, transcriptomic and proteomic analysis provides basis for a complete biomarker development from RNA expression (Figure 1.6) to validated quantitative RT-PCR assays; from MS-based protein identification to validated, multiplex immunoassays (Ong and Mann, 2005; Whiteaker *et al.*, 2011). Identification of such protein markers is empirically and imperatively crucial to improve drought tolerance in agriculturally useful crops such as sorghum and as a strategy in the proteomics research.

#### **1.4.2.4. Proteogenomics: Integrative genomics and proteomics**

A notion that integrative approach investigates 'complex' systems, which cannot probably be understood by investigation of individual components in separate, brings better understanding of how these complex systems are based on the development of computational models so that the response of the biological systems to any sort of perturbation (e.g. drought perturbation), can be



predicted (Aggarwal and Lee, 2003). Proteogenomics is a recently emerging term representing an incredibly vital tool for integrating protein-level information into the genome annotation process to attain genome annotation quality (Gupta *et al.*, 2007; Payne *et al.*, 2010). Proteogenomics involving both biological experiments and *in silico* analysis interplay a pivotal role between proteomics and genomics to utilize information from expressed proteins, often derived from mass spectrometry, to improve genome annotations (Gupta *et al.*, 2007; Ansong *et al.*, 2008). This shows the most likely interdependence between approaches in dissecting the complex trait such as drought. While genomics depends on and benefited from a highly complementary information offered by proteomics which transmit the most biological functions through protein into new insight of biological traits (Nilsson *et al.*, 2010), proteomics would not be possible without the previous achievements of genomics which provided the 'blueprint' of possible gene products, the focal point of proteomics studies (Tyers and Mann, 2003). The difference between the two partly lies on the factors that determine the expression of what they are supposed to measure. Genomics measures the genotype of an organism confounded by a rich and long history of genetics, whereas proteomics measures the phenotype shaped by both the genotype and the environment of the organism built upon an equally long history of biochemistry (Cox and Mann, 2011) and structural biology.

So, the improvement of the drought tolerance, as a complex trait, basically depends on the integrative complementation of genomics and proteomics approaches. Genomics, on one hand provides access to agronomically desirable alleles present at quantitative trait loci (QTLs) which affect such responses (Cushman and Bohnert, 2000). By and large, genomics-based approaches, aim at emphasising on the integrated analysis of stress-dependent behaviour relying on the physiological and biochemical observations that need to be linked together by a functional genomics with all information on gene complement, transcription and transcript regulation, the behaviour of proteins, protein complexes and pathways (Bohnert *et al.*, 2006). On the other hand, proteomics provides strategy that complements other functional genomics approaches, including microarray-based expression profiles (Shoemaker and Linsley, 2002) and systematic phenotypic profiles at the cell/tissue and organism level (Giaever *et al.*, 2002). Therefore, it is imperative to integrate genomics and proteomics data sets through application of bioinformatics to develop a comprehensive database of gene function that will serve as a powerful reference of protein properties and functions, which will also be useful both in building and testing hypotheses towards drought tolerance (Tyers & Mann, 2003).

### 1.5 Homology and syntenic relationships

Sorghum, a morphologically distinct species from other cereals but also highly diverse in its genetic bases with a wider range of adaptive traits and growth regional patterns has collinearity of orthologous regions with rice (a distantly related) and maize (a closely related cereals) conserved in tandem duplication (Li and Gill, 2002). It has earlier been shown that sorghum and maize had largely conserved gene content and gene number with extended regions of map collinearity (Bennetzen and Freeling, 1997). However, these two species did not exhibit a cross-hybridization of interspersed repetitive DNAs (Hulbert *et al.* 1990) as this part of the DNAs do not agreeably cross-hybridize between members of different plant genera. Collinearity in gene density in larger genome sized (750 Mbp) sorghum euchromatin is similar to that in smaller genome sized (400 Mbp) rice euchromatin. The difference in the sizes of the genomes of the two species may reflect the greater amounts of repetitive DNA in sorghum's pericentromeric heterochromatin (Mullet *et al.*, 2002). This suggestion has been supplemented in the recent sorghum genome analysis that identified more than 60% repetitive elements of sorghum constitute the genome (Paterson *et al.*, 2009). Similarly, sorghum genes are known to be about 24% and 7% grass and sorghum-specific respectively (Paterson *et al.*, 2009). Some drought tolerant genes have been suggested to occur due to recent gene and microRNA duplications. Based on the resulted gene duplication, in the rice, maize and sorghum genomes, it is suggested that there might be high functional gene redundancy, rapid gene silencing and/or loss from the duplicated genomes (Xu and Messing, 2008a and 2008b). Thus, of all identified sorghum genes, 34, 496 loci, during genome sequencing, only about 24, 580 (71.3%) are protein coding genes (Paterson *et al.*, 2009).

### 1.6 Gene-Trait Association

Several studies have reflected the challenges behind associating genes with phenotypes that detect genetic covariance across biological scale (Chesler *et al.*, 2005; Pierlé *et al.*, 2012). A typical phenotypic variation in quantitative genetics is caused by the change in gene expression which is under investigation. If such an alteration in gene expression is due to functional mutations, then the normal biological function is stored in the phenotypic variants with respect to the particular trait. This notion suggests that candidate genes which are defined by the known biological function directly or indirectly involve in the developmental processes of the investigated traits, which in turn could be proven by evaluating the effects of the causative gene variants in gene-trait association (Zhu and Zhao, 2007). Based on candidate gene approach that assumes some understanding of the

genetics of the trait, it is possible to test the hypothetical correlation between DNA polymorphisms in a particular gene and the trait of interest. This could be exemplified by examining a large collection of sorghum germplasm for a correlation between DNA sequence alleles of functional stay-green and the extent of chlorophyll in leaves (Park *et al.*, 2007; Hörtensteiner, 2009) and the increase in plant productivity (Harris *et al.*, 2007). However, genome scan as opposed to candidate gene approach, involves in testing for association of the segments of the genome by genotyping densely distributed genetic marker loci covering all the chromosomes. On the other hand, genome wide association studies (GWAS), takes care of one (or more) of the genetic loci being considered in association with the trait of interest as either causal for the trait or in linkage disequilibrium with the causal locus in a segmental genome (Rafalski, 2010). But, just for the reason of coverage, candidate gene association, could be considered a subset of a more general genome scan approach. Based on this, Rafalski (2010) has described some principles behind association analysis: when a genetically rich and diverse germplasm is genotyped and grouped into a densely spaced loci sharing SNP haplotypes (or alleles) that distributed along the genome, the respective distribution of phenotypic values for each haplotype are generated and can be compared and statistically evaluated.

Association studies in relation to a semantically built query components for known traits lays a platform for the identification of candidate genes pertaining to sorghum and its relatives contributing to drought and associated stress tolerance.

### **1.7 Candidate gene identification strategies**

One of the products of any genome project are the predicted genes along with the enriched information in terms of their location and functionality (Zhuo *et al.*, 2001). Genome-Wide Approach (GWA) and Candidate Gene Approach (CGA) are the two widely accepted approaches for identification and prediction of genes associated with complex quantitative traits like drought tolerance (Zhu and Zhao, 2007). While the former only locates glancing chromosomal segment of quantitative trait loci (QTLs) basing on the genetic distance which usually may harbour multiple candidate genes, the latter approach has been shown to be extremely powerful and most promising for studying the genetic architecture of complex traits. This shows that CGA is more effective and economical method for direct gene discovery (Byrne and McMullen, 1996; Harris *et al.*, 2007; Zhu and Zhao, 2007). Here we provide a comprehensive list/description of strategies used for the CGA. Candidate gene approach is extremely useful in determining the genetic foundation of complex drought tolerance when coupled with drought-regulated ESTs (Byrne and McMullen, 1996; Perez-

Torres *et al.*, 2009) and unique genetic materials such as near-isogenic lines (NILs; Nguyen, 2000; Sanchez *et al.*, 2002; Harris *et al.*, 2007). However, CGA relies on the *a priori* knowledge of possible candidates (Zhu and Zhao, 2007) for which reason digital candidate gene approach (DigiCGA) has recently emerged though is still in its infancy (Zhu *et al.*, 2010).

The gene-disease research, genetic association studies, biomarker and drug target selection in many organisms (Tabor *et al.*, 2002) including plants have gained popularity in CGA. Strategy for CGA that include rapid discovery of genes primarily employs cataloguing and categorizing genetically complex trait (Cushman and Bohnert, 2000). However, this strategy starts with selection of some target genes based on biological pathways or genome location relative to the landmarks identified such as known QTL for the target trait. One or some or all of the following studies based on the complexity of the trait are required (Faris *et al.*, 1999):

- 1) identifying biochemical pathways involve the gene of interest such as stay green or any trait (QTL) of interest;
- 2) examining a drought-related EST database, microarray and RNA-seq data and the mutagenesis approach for identification of valuable candidates and for verification of their association with the drought tolerance traits;
- 3) determining by intelligently guessing the sequence of at least some genes that are involved in the pathways, probably from homology studies (sequence homology);
- 4) text mining for orthologs in existing literature and databases for information on drought and related abiotic stress genes;
- 5) mapping genes to QTL to tag-genes as putative candidate. This step is likely based on the following points: i) mapping ESTs very near the trait QTLs for further candidate gene analysis (Nguyen, 2000). However, fine mapping of such genes that generate a QTL will only be possible in NIL populations, in which a single QTL provides all the population variation for drought tolerance; ii) mapping ESTs to genome annotation to uncover/tag-genes as putatively predicted/unpredicted candidate. Since most of the ESTs can be located on the physical map, one will be able to target a subset of candidate genes for co-segregation analysis in different populations;
- 6) conducting extensive biochemical and genetic studies to confirm the phenotypic effect;
- 7) hypothesizing a model regarding how drought tolerance is manifested based on several candidate genes with different functions in different QTL locations that could be verified;
- 8) using gene knock-out population by transposable elements, the existence of the known

- transposon sequence within a gene will normally knock out the function of that gene;
- 9) Producing F<sub>1</sub> plants containing a transposon within a candidate gene that can be identified by the polymerase chain reaction (PCR) using primers specifically designed for the gene of interest and the appropriate transposon.
  - 10) Performing a segregation analysis on the F<sub>2</sub> seed from the identified plant to confirm the putative function of the gene. Initial mapping activities can generally map targeted loci to 1-5 cM regions of the sorghum genetic map (Klein *et al.*, 2000).

Analysis of large segregating populations (~1,000 plants) is usually required to provide sufficient genetic resolution for efficient map-based cloning. Fine-mapping can then reduce the target locus in euchromatic regions to less than 100 kbp, a size that can be readily sequenced using standard BAC-based shotgun sequencing approaches. Interestingly, ~100 kbp of sorghum DNA, on average, will encode protein (Paterson *et al.*, 2009): First, if the targeted region is less than 500 kbp, shotgun sequencing of BAC DNAs spanning the region followed by BLASTX analysis can be used to identify sorghum genes that are related to other known protein coding genes. Second, the sorghum sequence can be compared to the sorghum EST database to identify the transcribed portions of the BAC sequence and third, other genes encoded by the sorghum BAC sequence can be identified by aligning the sorghum sequence with orthologous rice or maize sequences. Finally, gene prediction programs such as FGENESH (Salamov and Solovyev, 2000); riceGAAS (Sakata *et al.*, 2002) and AUGUSTUS (Stanke *et al.*, 2006a,b&c) can be used to identify regions of the sorghum genome that may encode genes. Genome sequence, genomic tools including microsatellite, cDNA, EST and cosmid libraries (Colbourne *et al.*, 2005) now allow access to link systematically the drought response information compiled over decades with the genomic data available for sorghum (Schwarzenberger *et al.*, 2009).

### **1.7.1 Integrated *In silico* Candidate Gene Approach (InsCGA)**

#### **1.7.1.1. Analytical Strategies**

In the current study, we used an integrated *in silico* candidate gene approach (InsCGA) by modifying a traditional CGA by integrating multi-analytical processes. Conceptual strategy for InsCGA was designed to include three main interrelated components which are not functionally mutually exclusive. These are: A) Primary process (Core issues): *a priori* analytical process based knowledge; B) Central process: Underlying analytical process of candidate gene identification; C) specialized process: *a posteriori* analytical process based knowledge for candidate gene

identification and prioritization (Figure 1.5). Each of these processes is further divided into sub components. While the primary process represents three core issues which are predominantly based on the predetermined knowledge with regard to the gene of interest, its cumulative effect add on to the central process as a prerequisite for the chain of coherence and as the basis for possible analytical process. This also indirectly links to the specialized process. The central process consists of five underlying analytical processes (UP) of candidate gene identification the combined effect of which would provide for generation of a list of potential candidate genes that would be catalogued in the following process. Last but not lest, the specialized *a posteriori* analytical process consisting of further evaluative steps directly dependent on output synthesized from the analytical underlying processes. The outcome of this would be processed in a hierarchical procedure representing identified list of genes, ranked, prioritized and validated promising elements.

#### **1.7.1.1.1 Primary process (Core issues): *a priori* analytical process based knowledge**

This process represents the baseline, the minimum predetermined knowledge, required for the identification of candidate genes.

##### **1.7.1.1.1.1 Core issue 1: Genes of interest (knowledge base)**

Though this issue is difficult to deal with, possible facts and evidences, concepts and ideas were put together to develop a consensus. This question considers defining a proposed type and near true nature of candidate genes to be under investigation looking for molecular, genetic, physiological and biochemical speculation and background features of the genes of interest. Candidate genes were proposed from the conceptual understanding and knowledge point of view and related evidences from reviewed and revised works based on the existing literature and from the inheritance, epistasis and epigenetic stand point. Since these are drought stress related adaptive genes, this theme considers the polygenic nature of drought tolerance in its quantitative and complex chain of continuous variation. This allows an interplay of several adaptive genes acting, reacting and interacting with one or more of their own partners and co-partners and with the niche or broadly ecophysiological environment they are hosted and existent in towards gaining response to develop adaptive traits. Such complex quantitative traits or formally QTLs shall follow in additive effects a polygenic inheritance unlike Mendelian genetics which predominantly govern a single genetic effect. Based on this, literature was gathered and reviewed to gain insight and understanding on the basis of proposed candidate genes. Typical characteristics of drought tolerance as the knowledge base were reviewed such as polygenic and complex quantitative nature (Ribaut *et al.*, 2002; Ravi *et al.*, 2011), genomic based approaches to discerning the regulatory mechanisms of abiotic stress

tolerance in plants (Sreenivasulu *et al.*, 2007), species specific advancement of drought tolerance (Campos *et al.*, 2004), systems biology-based approaches toward understanding drought tolerance in food crops (Jogaiah *et al.*, 2013).

#### **1.7.1.1.2 Core issue 2: Gene repositories and tools**

Databases that contain potential candidate drought responsive genes were identified (Table S2.4). Similarly, relevant tools for identifying candidate drought responsive genes were identified and selected (Table S2.5).

#### **1.7.1.1.3 Core issue 3: Identification of candidate genes**

Five major analytical components (Figure 1.5) based on the principle behind integrated InsCGA are selected for identification. The results from each were treated as specialized process, a *posteriori* analytical process based knowledge.

#### **1.7.1.1.2 Central Process: underlying process for identification of candidate genes**

Designing appropriate methodology for the analytical processes is vital to reveal the functional processes underlying strategy for candidate gene analysis. This is well suited to model the interplay over the entire processes between each analytical aspects of every functional process concordant with functional genomics though not feasible all the time to look at the whole (Westerhoff and Kell 2007). Central process represents and plays analytically predictive role based on the five major processes for identification and prediction of candidate genes. Sequence similarity search (SSS) (Altschul *et al.*, 1990; Conesa *et al.*, 2005), analysis of differential gene expression profiling (DE) (Vanderschuren *et al.*, 2013), analysis of metabolic pathways using Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Deokar *et al.*, 2011), analysis of orthology groups (Yu and Buckler., 2006), and analysis of gene ontology (GO) terms and functional GO enrichment (Zhou. *et al.*, 2007) were used as the key instrumental processes for candidate gene identification (see Figure 1.5, B).

Output for the candidate genes from SSS and all other analytical processes were analysed based on the functional genomics approach. Based on the evaluation, known and novel candidate genes were identified for drought responses. Known genes from existing annotation (Paterson *et al.*, 2009) which were not reported as drought tolerant were identified as potential candidate drought responsive, examined for their functional annotation and classified as hypothetical, putatively uncharacterised and unknown proteins. Novel genes identified were filtered, prioritized and tested for accuracy of prediction (see Figure 1.5C). Based on the criteria set for quality control on top of statistical significance mentioned above, namely: genomic coordinates, prediction score, gene

coverage, percent evidence prediction support, intergenic distance from the nearest neighbouring genes were used.

#### **1.7.1.1.3 Specialized process: *a posteriori* analytical process based knowledge**

The analytically processed results were integrated by pooling respective outcomes together (see Figure 1.5, C) for cataloguing and listing candidate genes. This was further employed to identify the top ranking candidate genes based on statistical models used for scoring similarities (Chen *et al.*, 2009), fuzzy measures for clustering categorical data (Gan *et al.*, 2009) and Pearson's correlation for a non-parametric measure of statistical dependence between variables representing candidate genes (Schafleitner *et al.*, 2007).

The candidate genes prioritization procedure (see Figure 1.5, graphical representation for hierarchical analysis output) is meant to provide a guiding procedure for evaluating hierarchical achievements of candidate gene analysis. The experimental validation work for known and novel candidate drought responsive genes was supposed to be conducted using RT-qPCR technology. The molecular validation of large size of candidate genes is both technically and economically challenging (Yin *et al.*, 2014). Thus, identified candidate drought responsive genes were validated using functional GO gene enrichment based on the fact that semantic similarity measures using GO enrichment is widely used in validation and trait gene prioritization (Pesquita *et al.*, 2009). Furthermore, identification of drought responsive proteins using differential expression profiling and MALDI-TOF-TOF MS/MS (chapter 4) was evaluated to reveal proteins encoded by genes in similar functional category with those identified *in silico*, an implicated functional validation.



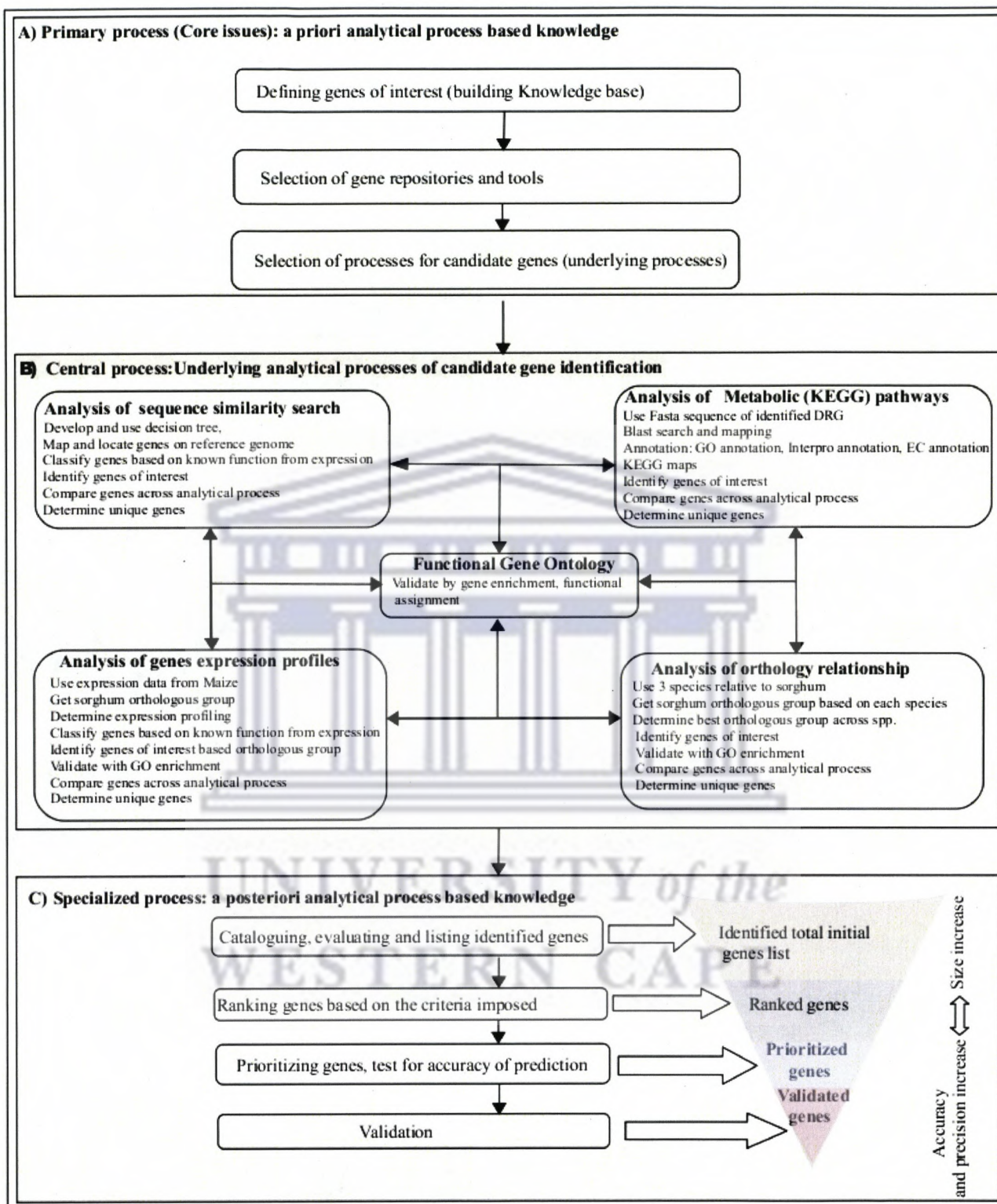


Figure 1.5: Conceptual strategic work flow for the candidate gene analysis and prioritization.

The conceptual work-flow consists of three component parts which are functionally interrelated: A) **Primary process (Core issues)**: representing *a priori* analytical process based knowledge to consist three core issues. These include defining the characteristic features of gene of interest to build knowledge base, selecting gene repositories and appropriate tools and identifying relevant analytical process for the candidate gene identification. This component represent a reference knowledge that provides basis for analytical process in identification of candidate genes. B) **Central process**:

Underlying Processes (UP) of candidate gene identification in this manuscript represent those analytical aspects used to help identify the candidate gene analysis based on systems approach to functional genomics. UP representing the heart of the conceptual work flow is the collective form of five interrelated aspects of analytical processes used in this research. These include analysis of sequence similarity search, metabolic (KEGG) pathways, gene expression profiling, orthology relation and functional gene ontology as a central for all the analysis. The combined effect of all the analysis allow generation of a list of potential candidate genes. However, it is important to note that challenges may not be avoided in this analysis because drought responsive candidate genes (DRCG) may vary in attributes based on the complexity of trait. C) **Specialized process:** *a posteriori* analytical process based knowledge represent the post analytical process out-put that include cataloguing and evaluation processes to rank and prioritize genes reducing to a promising number based on the criteria used. This also include graphical representation of the prioritization steps denoting the hierarchical out comes of a nest of gene sets coming up with validated promising elements. Starting with the largest size of genes increase in accuracy and precision with decreasing in size of genes to the minimum possible promisingly validated for use.

### 1.7.1.2 Genomic resources

#### 1.7.1.2.1 Expressed Sequence Tag (EST) mapping

Expressed sequence tags being produced in a large batch, denote a snapshot of spatio-temporally expressed genes that represent tags of a particular cDNA library featured in a short single-pass sequence read (Parkinson and Blater, 1995). ESTs or mRNA fragments, however, depending on the sequencing technology mostly varies in length from 200 to 500 nucleotides. Significant figure of ESTs have been generated with an advancement of high throughput technologies and the automated sequencing projects playing role in the discovery of genes and the assignment the functions (Nagaraj, *et al.*, 2007).

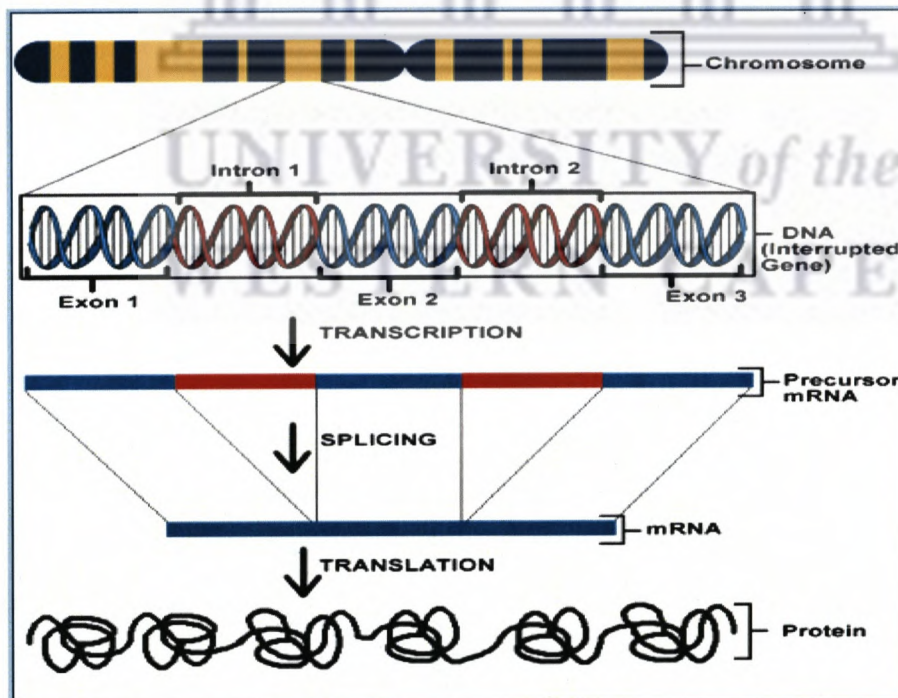


Figure 1.6: An overview of the process of mRNA translation

Adapted from NCBI: <http://www.ncbi.nlm.nih.gov/About/primer/est.html>

A global comprehensive EST database (dbEST), a division of GenBank, National Centre for Biotechnology Institute (NCBI) contains single-pass cDNA sequences data and other information from a number of organisms including sorghum (Boguski *et al.*, 1993). Sorghum dbEST constitute a collection of 209,835 ESTs in the collection release '130101' (Sakharkar *et al.*, 2009). Obviously, finding a gene that codes for a protein, or proteins, is not an easy task. Thus, gene identification in sorghum is relatively difficult for the fact that sorghum genome is composed of 60 - 75% repetitive elements and other features like introns interspersed with a relatively few-DNA coding sequences (Paterson *et al.*, 2009).

Although there are limited EST collections for most species apart from human and mouse, the available resources are very useful in gene identification if the resource are employed from related taxonomic groups (Brendel *et al.*, 2004). Sorghum consensus gene prediction has employed EST's from 15 relative species for mapping and computing gene structure models. This was relied on a similarity-based approach by adding external homologs from maize for splice site models (Paterson *et al.*, 2009; see also section 1.1.3.4.1 and Figure 1.5). In addition, it is important to realise that the collection of ESTs reflecting the level and complexity of gene expression (Zhang *et al.*, 2010), to-date the majority of EST entries constitute mainly from grass family (Sakharkar *et al.*, 2009).

#### **1.7.1.2.1.1 What is the advantage of ESTs?**

Ever since the launching of the first EST sequencing project for human in 1991 (Lander *et al.*, 2001) and the establishment of dbEST as a division of Genbank in 1992 (Boguski *et al.*, 1993), many similar projects have been completed or are under way for many plant species. All of these projects have a common agenda such as for: providing useful tools within and between genomic comparisons (Fulton *et al.*, 2002; Schlueter *et al.*, 2004), gene discovery (Ewing *et al.*, 1999; Ronning *et al.*, 2003; Hughes and Friedman, 2005), molecular marker identification (Michalek *et al.*, 2002), microarray development (Alba *et al.*, 2004; Arpat *et al.*, 2004; and Close *et al.*, 2004), polyploid species genomic resource development (Udall *et al.*, 2006), sequence data source and quality filtering; highlight transcript sequence diversity and splicing (Wolfberg and Landsman 1997). This entails that ESTs provide ample opportunities in providing routes for gene discovery, gene expression and regulation, and genome mapping and landmarking.

#### **1.7.1.2.1.2 ESTs: Tool for gene discovery**

Expressed Sequence Tags provide research communities with a quick and cheaper route for discovering new genes, for obtaining data on gene expression and regulation, and for constructing genome maps (Duggan *et al.*, 1999; Alba *et al.*, 2004; Manickavelu *et al.*, 2012). As ESTs represent mRNAs (transcripts) that were copied from coding region of the genome, they are experimentally proven to be a potential tool in the discovery of genes (Duggan *et al.*, 1999). Since mRNA in a cell do not contain sequences from the intergenic regions, nor from the non-coding introns that are present within many genes, isolating mRNA appears impressive to finding expressed genes in the genome. Some typical examples are: 1) representation of the 60 % of *Arabidopsis thaliana* genes by 105, 000 ESTs (Arabidopsis, 2000); 2) identification of wheat genes by EST and full-length cDNA (Manickavelu *et al.*, 2012) in complementary to the ongoing genome sequencing ([www.wheatgenome.org](http://www.wheatgenome.org)), and 3) the coverage of about 90% of the switchgrass gene space (Wang *et al.*, 2012). Consequently, ESTs as a tag for expressed genes have gained practical advantage in spite of only a single sequencing experiment required at each cycle of cDNA generation (Christoffels *et al.*, 1999; Christoffels *et al.*, 2001). Although no basic steps as such is required for error checking, it would not basically affect the identification of the genes from which the EST were derived (Duggan *et al.*, 1999). Hence, steps in detecting ESTs include: Observation of phenotypic syndromes; examination of the DNA of the stressed tissue for mutations (alteration) and isolation of genes involved in specific trait (eg. drought responsive genes).

#### **1.7.1.2.1.3 ESTs: as a source of data on gene expression and regulation**

While identification of the gene and its genomic location is important, it is not the ultimate goal of genome sequencing *per se*. However, importantly, one wants to gain understanding of the spatio-temporal patterns of gene expression and the process associated with it. This eventually elucidate understanding of the gene expression patterns with respect to time and space in an altered condition for which identification of protein coded for the genes is informative. Huge progress has been realized in terms of characterization and analysis of expression data underpinning the gene expression profiles (Cramer *et al.*, 2011; Jamil *et al.*, 2011). The regulation of alterations in gene expression (Baena-González, 2010) is a consequence of the fast growing technologies in this field. ESTs, microarray, RNA-seq, large-scale gene expression (transcriptome) profiling are among many of these technologies (Alba *et al.*, 2004). EST data has proven to be important for mining UTRs as both 5' and 3' ESTs contain significant sections of the UTRs along with protein coding regions

(Nagaraj *et al.*, 2007). Some functionalities of ESTs are intended to facilitate the candidate gene discovery through: (i) access to specialised subsets of the data base; (ii) identification of isoforms based on physical or developmental expression states; or (iii) locating entries based on physical location within the genome (Christoffels *et al.*, 2001).

#### **1.7.1.2.1.4 ESTs: tool for gene mapping (genome Landmarks)**

ESTs as a tool for genome mapping provides potential landmarks for navigating through genomic region in search for specific segment of the DNA sequence that code for a particular protein. Thus, EST as genomic landmark provides a clue to identify and discover new genes (Rapp and Wheeler, 2005). Among other potential mapping techniques, Sequence Tagged Site (STS), is the most reliable landmarks owing to the fact that it is a DNA sequence easily recognized and appearing only once in a genome or chromosome (Olson *et al.* 1989). A common source of STSs that most likely represent unique identity of a specific species directly related to an expressed gene is 3' ESTs in the NCBI repository. Mapping of ESTs to specific chromosomal locations can be generated using physical mapping techniques, such as radiation hybrid mapping (Christoffels *et al.*, 2001), Happy mapping, or FISH (Thangavelu *et al.*, 2003; Cossins and Crawford, 2005), and *in silico* alignment mapping (Rheadet *al.*, 2010) to the genome that originated the EST specially for organisms having complete genome sequence.

#### **1.7.1.2.1.5 Challenges and limitations of ESTs**

ESTs may be represented in databases as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand. Since genes are frequently expressed as mRNA splice variants, so many overlapping ESTs ultimately redundantly originated. As a solution to this EST limitation, UniGene database has been deliberately established to automatically partition EST sequences into a non-redundant set of gene-oriented clusters (Stanton *et al.*, 2003). There are several other limitations in association with the EST approaches despite the fact that it is widely recognized as an efficient strategy to identify genes. The difficulties in isolation of mRNA from some tissues and cell types create the scarcity of expression data related to certain genes that may only reside in these tissues or cell types (Chawla *et al.*, 2011). Second is that important gene regulatory sequences may be found within an intron. Because ESTs are small segments of cDNA, generated from a mRNA in which the introns have been removed, much valuable information may

be lost by focusing only on cDNA sequencing. Despite these limitations, ESTs continue to be invaluable in characterizing the human genome, as well as the genomes of other organisms. They have enabled the mapping of many genes to chromosomal sites and have also assisted in the discovery of many new genes. EST sequences are grouped into a cluster if they share a minimum of 95% identity over a 40 nucleotide or longer with fewer than 20 bases of mismatch sequence at their end (Christoffels *et al.*, 2001).

Aiming at reducing the number and the redundancy in ESTs for the downstream analysis, EST contigs have been produced by assembling ESTs the best resources of which include TIGR gene indices (Lee *et al.*, 2005), STACK (Christoffels *et al.*, 2001), and UniGene (Stanton *et al.*, 2003). On this review we are briefly reviewing UniGene by taking greater attention to the building procedure and the stages and the criteria behind the procedure.

### **1.7.1.3 UniGene**

UniGenes are largely automated analytically or experimentally unified system for producing a partitionally organized GenBank sequences into a non-redundant set of gene-oriented transcripts (Pontius *et al.*, 2003). It is a well defined data set containing clustered members of groups of ESTs (Miller *et al.*, 1999; Dai *et al.*, 2005). Unlike what its nomenclature primarily denotes, a database for genes, UniGene represent an NCBI database of the transcriptome (Pontius *et al.*, 2003). Sets of transcript sequences in a database of each UniGene entry are probably originated from the same gene or expressed pseudo-gene (Isokpehi *et al.*, 2009), along with all informations on amino acid similarities, gene expression, cDNA clone reagents, and chromosomal location (Mewes *et al.*, 2002; Wheeler *et al.*, 2005). This indicates that set of transcripts involved in each UniGene appears to represent isoforms of the same gene. Hence, the motivation for establishment of the UniGene database was primarily aimed at the resolution of the difficulty created by the high level of redundancy of transcribed sequences of which effective use is very unlikely (Wheeler *et al.*, 2005). UniGene databases are updated weekly with new EST sequences and bimonthly with newly characterized sequences (Wheeler *et al.*, 2007).

#### **1.7.1.3.1 UniGene clusters**

UniGene clusters primarily consists of the EST, mRNA, and the mixture of ESTs and mRNA/cDNA sequences, with all the coding sequences (CDSs) annotated on genomic DNA, into subsets of

related sequences (Pontius *et al.*, 2003; Wheeler *et al.*, 2005). Based on this formulation, sorghum UniGene was built around sequences from the known genes obtained from GenBank and the ESTs from dbEST (Matsumoto *et al.*, 2011). A transcript based building method was employed taking into considerations the alignments between all transcript sequences to generate clusters of sequences originating from the same gene. In this method, the clustered members of transcript set include mRNA, ESTs, and high-throughput cDNA (HTC) sequences available from GenBank that contain a final number of clusters (sets) equal to 14 057 based on the UniGene Sbi build 29 and 13 733 UniGene Sbi build 30 (Albert *et al.*, 2005).

#### **1.7.1.3.2 UniGene build procedure: steps for inclusion of ESTs into clusters**

UniGene Build is a staged procedure that includes several important steps evaluating members of the clusters that have to be included (Table S1.1). Each stage is adding less reliable data to the results of the preceding stage referring in decreasing reliability order to pairwise identity, annotation, shared clone and information (Hide *et al.*, 1999). Here we present an updated complete description of standard check as selective criteria for UniGene building procedure and inclusion of sequences into the UniGene clusters (for a complete pipeline or flow chart, see Figure S1.1). Unlike the TIGR gene index, UniGene follow a combination of supervised and unsupervised methods with variable levels of stringency in clustering and with no consensus sequences being produced (Pontius *et al.*, 2003).

The primary step in the UniGene clustering is the quality filtering process where the sequence quality will be checked for repeats, low information content and vectors by running repeatmasker against known genomic repeat (Table S1.1; Smit *et al.*, 2004; Tarailo-Graovac *et al.*, 2009), DUST for identifying a run of a single pyrimidine or purine in DNA and low-complexity DNA sequences (Morgulis *et al.*, 2006), TRF, a tandem repeat finder for finding the tandem repeats (Benson, 1999) and Cross-matching for contaminants, fragments of vector, mitochondrial and ribosomal sequences against database (Heller *et al.*, 2010).

UniGene rule underlines that a sequence to be included in UniGene cluster must at least be 100 base pairs in the clone insert with high sequence quality and not repetitive (Table S1.1). Hence, sequences are discarded if the minimum length falls below 100 informative base pairs (Hide *et al.* 1999). The third important step is the sequence record where ESTs and or other sequences are

required to satisfy the minimum set of data necessary for a gene record that include a unique identifier, or GeneID, assigned by NCBI; and other information (Benson *et al.*, 1999; Pontius *et al.*, 2003). It is important to bear in mind that gene records are only created for genomes whose annotation and assembly are completely represented by Whole Genome Sequencing (WGS). In this case UniGene builds also include genome basis. For instance, sorghum genome completely represented by WGS assemblies and record containing annotated genes (Paterson *et al.*, 2009), has gene records and RefSeq created for all high quality loci defined as per the criteria for Entrez genes (Maglott *et al.*, 2007).

As the next step, set of ESTs is compared with set of initial clusters using megaBLAST to add high similarity pairs to the clusters discarding links that would join the initial clusters. Likewise, EST to EST links are also created to increase the dimension of initial clusters and to create purely EST clusters, however discarding the unmatched ones. Importantly, clone-based edges are added to avoid overlapping 5' and 3' ESTs for assigning to the same cluster merging two clusters with clone IDs that link at least two 5' ends and two 3' ends coming from distinct cluster (Girma, 2012). If a sequence with no polyadenylation signal or tail is found in a cluster, it is discarded. These are called anchored clusters, because their 3' ends are presumed to be known (Kim *et al.*, 2005). Lower level of stringency would operate for ESTs that do not belong to an anchored cluster (Hide *et al.*, 1999; Mukhopadhyay *et al.*, 2002) and those meeting this evaluation are added to the cluster as the guest members. Clusters with just a size merge with the one of most similar sequence based on a comparison against the rest. The resulting clusters are compared with the immediate preceding UniGene build to provide succession for the following build bearing in mind that Genbank accessions are safer alternative than cluster IDs as reference at time of merging clusters. The summary of all the parameters required in UniGene staged clustering is given in Table S1.1.

#### 1.7.1.4 Functional Genomic Annotation

The result of any genome-project is an ever-widening gap between drafted and fully annotated that only promises to continue (Chain *et al.*, 2009). The value of the genome depends on the extent of annotation because the latter fills the gap from the sequence to the biology of the organism (Stein, 2001). This entails that genome annotation is the process of attaching biological information to sequences in terms of identifying repeat-poor, gene-rich euchromatin (Schmutz *et al.*, 2010) thereby identifying protein coding genes (gene prediction) and repeat-rich heterochromatin regions (Lippman *et al.*, 2004).



Searching for sequence homology is the basis for annotation on which genome annotation is based (Pevsner, 2009), though, manual annotation is still imperative to enrich the annotation platform (Flicek *et al.*, 2013). Based on the information they assign to the genome, two interrelated types of genomic annotation exist. Structural annotation is concerned in the identification of genomic elements as collective process of identifying genes (Open Reading Frames (ORFs) and their localisation, genes and gene structure, coding regions, promoters, and regulatory elements (motifs) and their location). On the other hand, functional annotation using GO as a reference guide (Ashburner *et al.*, 2000) is assigning biochemical and biological functions, involved regulation and interactions and expression profiles to each structural element or gene (Ansong *et al.*, 2008; Bright *et al.*, 2009). With this regard, among the total genes predicted in sorghum, about 30% are thought to have no functional annotation (Paterson *et al.*, 2009). Proteogenomics, however, may play a major role in making use of translated information to upgrade genomic annotations.

#### **1.7.1.5 Functional Ontology annotation**

Gene Ontology (GO) is the structured and controlled vocabulary of terms for the description of three important non overlapping biological domains (Harris *et al.*, 2004) representing and processing of information about gene products and functions (Smith *et al.*, 2003). The GO project was initiated in 1998 based on three model organisms with targeted goals to develop a set of ontologies, to describe key domains of molecular biology and to apply GO terms in the annotation of sequences, genes or gene products for centralized public resource that allow universal access to the ontologies and annotation data sets (Harris *et al.*, 2004). Other than using it as a *de facto* reference-guide for functional annotation, GO helps standardize the way evidence codes are used for curating the various databases in plants (Ashburner *et al.*, 2000; Jaiswal *et al.*, 2005). The gramene database has provided a wide range of ontologies for extensive use of controlled vocabularies to describe specific biological attributes in comparative genomics across taxonomic groups (Ware *et al.*, 2002; Liang *et al.*, 2008).

The gramene ontology database is a source of an integrated information based on structured controlled vocabularies for the knowledge domains such as: 1) Plant Ontology (PO) for description of plant anatomy and the stages of plant development (Jaiswal *et al.*, 2005; Avraham *et al.*, 2008); 2) Trait Ontology (TO) for specifying of phenotypic traits related to mutants and QTLs (Jaiswal *et al.*, 2002; Ware *et al.*, 2002; Bard and Rhee, 2004 and Gaulton *et al.*, 2007); 3) Plant Gene

Ontology (GO) for describing genes and their products from the perspectives of the three important biological domains vis-à-vis molecular function, biological process and cellular component (Harris *et al.*, 2004); 4) Environment Ontology (EO) for enumerating the concepts and relations of different simulation environments and ontology of landforms (Smith and Mark, 2003); 5) Gramene's taxonomy ontology (GR\_tax) for displaying the taxonomy tree of the plant species in the ontology format (Ware *et al.*, 2002; Jaiswal *et al.*, 2006 and Liang *et al.*, 2008); and 6) Plant Growth Ontology (PGO) for description of plant growth and developmental stages (Pujar *et al.*, 2006; Liang *et al.*, 2008). In addition, Crop Ontologies (CO), have been established for 12 Global Challenge Programme (GCP) mandated crops including sorghum for developing and describing crop-specific ontologies primarily to address the concept of anatomy and plant agronomic traits (Shrestha *et al.*, 2010) and define how many other domains and publicly available ontologies were required to fully reflect the concept. So, plant ontology can be successfully integrated with functional annotation and mapping for detecting QTLs that control complex traits such as drought tolerance and for testing the interplay between gene action and development (He *et al.*, 2010).

## **1.7.2 Experimental approach for gene identification**

### **1.7.2.1 Drought phenotyping**

Plants inherent diversity and reactions to varying level of environmental factors such as moisture stresses and complexity of drought tolerance contribute drought phenotyping to real challenge. Thus, sorghum drought phenotyping would require meticulous selection of drought testing environments such as hydroponics (Mir *et al.*, 2012), growth chambers (Wall *et al.*, 2001; Gholipour *et al.*, 2010), Greenhouses (Kebede *et al.*, 2001), Rain-Out Zones (Harris *et al.*, 2007) and Fields (Salekdeh *et al.*, 2009; Van Oosterom *et al.*, 2011; Hammer *et al.*, 2012). These depend on the question in target to properly address the actual biological variability in the test population. Appropriate experimental design (Salekdeh *et al.*, 2009), robustness of drought phenotyping parameters (El Soda *et al.*, 2010; Brito *et al.*, 2011), accurate, precise and timely measuring and sampling (Cobb *et al.*, 2013), automated data processing and evaluation (Burke *et al.*, 2010; Cabrera-Bosquet *et al.*, 2012) are all essential for successful drought phenotyping in particular reference to sorghum. Detail and complete description of drought phenotyping have been given in articles and reviews (Hervé and Seraji, 2009; Monneveux and Riboult, 2011; Masuka *et al.*, 2012).

### 1.7.2.2 Source of stay-green genes

Stay-green is a character that regulate chlorophyll and chlorophyll-binding protein degradation during senescence (Park *et al.*, 2007; Hörtensteiner, 2009). Stay-green mutants are delayed in leaf senescence compared to functional stay-greens which have the potential to increase plant productivity (Harris *et al.*, 2007). A stay-green phenotype could arise in one of four fundamentally distinct ways (Thomas and Smart, 1993) and could occur in one of the five ways (Thomas *et al.*, 2000). Stay green in sorghum (*Sorghum bicolor* (L.) Moench) is a vital trait related with post-flowering drought tolerance. Four major stay-green, *Stg1*, *Stg2*, *Stg3*, and *Stg4* (Xu *et al.*, 2000) and many minor QTL that can modulate expression of the stay-green traits had been earlier identified with distinct genomic location controlling the functional basis of senescence or 'stay-green' in the source plant (Borrell *et al.*, 2014). However, recently, based on the analysis using 55 simple sequence repeats (SSR) markers with genome coverage, eight distinct sources of stay-green genes have been identified in the sorghum germplasm of which most originating from Sudan and Ethiopian collections (Harris *et al.*, 2007; Kassahun *et al.*, 2011). Several sorghum genotypes have been identified that exhibit the stay-green trait some of the most commonly used highlighted in this section are BTx 423, BTx623, Btx642, Tx7000 and E 36-1.

BTx 423 (also called PI 659985 MAP), from which most genomic resources for sorghum has been developed gained popularity in breeding and genomic research programmes (Kresovich *et al.*, 2005). BTx623 beyond its attribution in agricultural sector, its values in genomic research has gained a central position owing to its published fully sequenced sorghum genome (Paterson *et al.*, 2009). Another genotype likely to be focused on is Btx642, formerly known as B35. Btx642 is a back cross 1 derivative of IS12555, a dura landrace cultivar from Ethiopia (Walulu *et al.*, 1994). This has been an especially useful source of stay-green genotype for research (Tuinstra *et al.*, 1998; Xu *et al.*, 2000), for the development of commercial hybrids (Kumar *et al.*, 2011) and for distinct responses to drought during the pre-flowering and post-flowering stages compared to Tx7000. It is susceptible to pre-flowering drought and highly resistant to post-flowering drought (stay-green) with a relatively low yield potential. Tx7000, an elite open-pollinated variety Caprock, high-yielding with nonstay-green, pre-flowering drought tolerant was released from a 'Kafir' × 'Milo' cross and distributed in the late 1940s but later a hybrid male parent of Texas 660 was released in the late 1950s (Pratt *et al.*, 2005). Tx7000, exhibits two lines: A-line (male sterile or the seed parent) is the cytoplasmic sterile counterparts with the B-line which restore fertility in A1 cytoplasm. B-line contain the restorer gene

for male-fertility counter-parting with A-line to maintain the male-sterility (Hunt *et al.*, 2011). E 36-1, Ethiopian released variety is striga susceptible stay-green drought tolerant (Van Oosterom *et al.*, 1996) and has been widely used in the drought-breeding program at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

### 1.7.2.3 Protein identification and purification

Proteome analysis has become an increasingly crucial task involving wider research area that target dissection of complex traits, biomarker discovery, cancer prevention, novel candidate gene identification and discovery, food safety, protein interaction studies, medicine treatment, disease screening and many more. The advancement of proteomics in the postgenomic era has proven to be the global domain in understanding the application of proteomes across the board and in recognizing peptide fingerprints as the sole bottom line evidence for identification and validation of gene markers. Strategies have been developed for identification and purification of proteins that involve solubilization of proteins with detergents, separation of proteins by sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE) and digestion of the gel-trapped proteins ('in-gel' digestion) (Shevchenko *et al.*, 1996) and without detergent, comprising protein extraction with strong chaotropic reagents such as urea and thiourea, protein precipitation and digestion under denaturing conditions ('in-solution' digestion) which may usually be followed by two-dimensional peptide separation (Washburn *et al.*, 2001). Both strategies convert proteins extracted from tissues efficiently to peptides suited for advanced proteome analysis. Though, less automated and the peptides may not be easily recovered, in-gel digestion is advantages in cleaning impurities which otherwise interfere with digestion, whereas in-solution digestion, in spite of its readiness to be automated, the proteome may be incompletely solubilized and digestion gets impeded by interfering substances (Wiśniewski *et al.*, 2009). As strategies for protein identification, in-gel or in-solution based approaches can further be divided as follows, detail work flow for both strategies are provided in (Capelo *et al.*, 2009): (i) protein dissolution or denaturation (Park and Russel, 2000), (ii) protein reduction (Capelo *et al.*, 2009) (iii) protein alkylation (Sechi and Chait, 1998) and (iv) protein digestion (Nesvizhskii, 2010; Stauber *et al.*, 2010).

#### 1.7.2.3.1 Protein Optimization and Quantification

One of the most efficient way to optimize the quantitative detection of protein is to minimize complexity in sample by increasing the chromatographic gradient times or through biochemical

fractionation prior to Liquid chromatography–mass spectrometry (LC-MS/MS) analysis (Lee *et al.*, 2010; Schulze and Usadel, 2010). However, for gel-free quantitative proteomic analysis, phenol protein extraction method has proven to be the most efficient and reproducible, and hence used to optimize the in-solution digestion method (Lee *et al.*, 2010 ).

Strategies for differential proteomics have been designed to measure the quantitative variation in protein abundance between stressed/disturbed and normal/reference samples. However, since a particular quantitation strategy is influenced by factors such as the technical reproducibility, biological variation and sample amount, it is important to have enrichment and purification steps of sub-proteomes before actual comparison of protein abundances (Falick *et al.*, 2011). Thus, based on the physical and chemical properties of the different tryptic peptides, differential proteomics is categorized into gel-based and mass-spectrometry-based quantitative (Schulze and Usadel, 2010).

#### **1.7.2.4. Differential expression**

Analysis of expression profiling can provide insight into the protein function revealing regulatory pathways to indicate potential effects of stresses and serving as a diagnostic syndrome (Baginsky *et al.*, 2010). Gene expression studies using cDNA tag sequencing, two-dimensional gel proteome analysis and microarray and RNA-seq technologies have improved our understanding as the approaches providing relevant experimental information that can be viewed as validative steps for *in silico* identification (Claverie and Notredame, 2011; Matas *et al.*, 2011; Amiour *et al.*, 2012). Genome wide expression data have provided opportunities to investigate thousands of genes at the same time and to understand effects of stress on tissues under consideration. Stress-associated traits and genes are complex, multigenic and networked in nature. Thus, their identification through analysis of regulatory systems and pathways they are involved in provides understanding on the basics behind stress response related gene expression and regulatory network (de la Fuente, 2010).

Differential protein expression is the analytical measure of changes in protein expression under changing moisture conditions from ambient to the drought stress for different tissue types or developmental stages (Rabello, 2008). Subtracted cDNA using RT- PCR (Jiang *et al.*, 2004) and Liquid-Chromatography/Mass-Spectrometry (Pandey and Mann, 2000; Zhang *et al.*, 2006) are tools currently applicable for identifying, visualizing and analysing differential expression of proteins.

## 1.8 Rationale of the thesis

Drought stress is a major cause of agricultural failure ultimately hampering crop productivity (Tester and Langridge, 2010; Meena *et al.*, 2012). World crop damage due to drought stress exceed \$10 billion annually (Mutava, 2009). Understanding the relationship between drought phenotype of the plant during drought stress condition and the typical drought tolerance pathway is a novel approach to improve crop productivity. A completely sequenced and annotated genomic resource and associated proteome provide wide opportunities to identify promising drought responsive genes and biomarkers which can be targeted in breeding programs to enhance sorghum productivity. Studies using sorghum and other model crops have demonstrated that drought tolerance is a function of complex polygenic traits mediated by complex metabolic networks that uses a wide range of QTLs, protein and gene families and metabolic pathways to respond to drought stress (Cattivelli *et al.*, 2008).

Sorghum genome annotation and progressive update revealed sorghum-specific genes with novel drought-related functions (Paerson *et al.*, 2009 and Moris *et al.* 2013). However, in contrast to the large number of studies assessing candidate genes for drought impacts under a single environmental stress, relatively little attention has been devoted to understanding how genes function in multiple pathways under several environmental stresses. Addressing this issue, integrally and comparatively evaluates how metabolism and gene expression are altered under water stress conditions in relation to other stresses.

The vast natural variability in sorghum arising from the major and hybrid races (Harlan and de Wet, 1972) across the wide geographical area and integrated “omics” datasets, offer opportunity for harnessing a comprehensive catalogues of drought responsive genes in order to unravel the unique association of genes to drought tolerance. A total of approximately 22000 unique transcripts and of about 15000 UniGene clusters (UniGene build # 29; Pontius *et al.*, 2003) exist that are organized into about 90 diverse libraries from several sorghum genotypes (Kresovich *et al.*, 2005). This resource together with a collection of 200,000 sorghum ESTs which originated under drought condition from experimental tissues offers opportunities to discover novel sorghum specific drought responsive genes and proteins. The result of this work would provide basis for understanding metabolic pathways that modulate drought tolerance. Furthermore, comprehensive catalogue of drought responsive genes resulted from this thesis would represent unique value in sorghum productivity.

## 1.9 Aims and objectives

1. Identify candidate genes associated with drought tolerance using integrated *in silico* approach;
2. Predict novel drought responsive genes structure models using a combinatorial approach;
3. Identify proteins differentially expressed under drought stress conditions using MALDI-TOF mass spectrometry analysis;

4. Identify gene-trait and gene-phenotype association using integrated ontologies and query building approach;
5. Validate identified candidate drought responsive genes and proteins using functional GO gene enrichment.

### 1.10 Overview of the thesis

This thesis is organised into the following chapters:

**Chapter one** outlines the main components of the thesis and gives a general review on each. The detailed purpose of this chapter has been described in the abstract section.

**Chapter two** provides description on the methodology, the findings and the arguments there in with regard to the candidate gene analysis illustrating the pipelines for mapping experimental data to reference genome and for building gene models. It also provides description on the functional annotation of genes putatively uncharacterised in the previous annotation. Furthermore, it describes the gene ontology terms associated with genes and biochemical pathways that these genes involve in. This chapter also provides a detailed description of novel gene structure prediction.

**Chapter three** gives a detailed description of the integrated approach in determining gene-trait and gene-phenotype association. This chapter also details the role of functional ontology in determining gene association with plant phenotypes.

**Chapter four** provides a detailed description of the drought responsive proteins identified using MALDI-TOF-TOF MS/MS. It also provides description on protein quantification and separation; spot analysis, qualitative and quantitative differential expression patterns and lastly protein spot identification.

**Chapter five** highlights general description of each chapter outlining only the main points. Integrates and compares results across the different chapters. Lastly, it gives conclusive remarks, applications and future research plans.

## CHAPTER 2

### Chapter 2: *In silico* identification of candidate genes for drought tolerance in Sorghum (*Sorghum bicolor* (L.) Moench)

#### Abstract

**Background:** Genetic dissection and understanding of the biological functions of drought regulated traits are complex in nature that necessitate integrative approaches. Identification and characterization of candidate genes for drought tolerance using a model organism and high-throughput technologies is the best strategy to enhance productivity. Here we present an integrated approach for the identification and analysis of candidate genes for drought tolerance using sorghum, a model cereal, well adapted to drought-affected regions that sustain millions of resource-poor lives.

**Methodology:** Integrated *in silico* identification of candidate gene analysis (InsCGA) is a powerful and holistic approach that takes into consideration functional features of genes. Here we demonstrate a novel approach for InsCGA using a conceptual work flow that consists of three interrelated components to integrate multi-functional analytical processes and to prioritize candidate genes. We have developed a pipeline and decision tree that includes mapping expression data from 92 normalized cDNAs to the sorghum genome to identify and characterize drought tolerant genes. The approach integrates a sequence similarity search, metabolic pathways, gene expression profiling and orthology relation to investigate genes of interest. Genome functional reannotation enabled modification and update of the existing sorghum annotation and the discovery of novel gene features where the latter employed the combination of *ab initio* and extrinsic evidence-driven information and multiple sources of criteria to improve accuracy in gene prediction. Gene ontology was used to validate and to functionally assign and enrich genes across-analytical processes.

**Result:** A total of 10619 UniGene clusters derived from drought expression libraries were mapped to the genome based on pair-wise sequence similarity search. Out of this, 9763 (91.9%) UniGenes mapped to existing gene loci. The remaining 856 (8.1%) didn't match with any existing gene annotation and were considered as putative novel gene loci. One hundred and twenty-three of the 9763 genes (1.3%) used the drought response expression information from the UniGene libraries. Identification of two merged genes and 3 corresponding transcripts and 64, 74 and 3595 novel exons, 3' and 5' UTRs respectively enabled a potential update of more than 4400 (~12.6%) existing



single gene models. In this study, 241 genes were identified to be novel and interestingly, nearly 50% of these represent single exon genes. Approximately 69% represent DR and 6% complete gene structure model with 3' and 5' UTRs. Five classes of interprodomains namely protein domains (205, 33%), protein families (179, 28%), motives or binding sites (6, 1%), repeats (12, 2%) and null (228, 36%) were identified based on the drought responsive UniGenes that mapped to known genes. Interproscan revealed approximately 49% of the drought responsive genes possess interpro ids. This is almost comparable with the number of corresponding known genes (69, 57%) enriched using GO enrichment ( $P$ -value  $< 0.05$ ). Analysis of biochemical metabolism revealed 14 metabolic pathways related to drought tolerance and 32 genes that encode protein enzymes that catalyse substrate conversions in the respective pathways. This indicates a biological network among categories of genes involved with some playing the rate limiting role in all pathways. Expression profiling showed 12 genes significantly expressed under drought stress conditions. Similarly, results from analysis of ortholog groups showed 265 non-redundant genes responsive to drought stress which were verified by gene ontology enrichment for water deprivation (118, 44.5%), desiccation (21, 8%), heat (91, 34.3%), ABA stimulus (109, 41.1%) and ABA mediated signalling pathways (37, 14%).

**Conclusion:** Our results have shown that the consistency of our method proven to be a powerful approach for identifying candidate drought responsive genes (CDRGs). This study has successfully identified significant array of prioritized candidate known and novel genes that are critical to respond to drought and related stresses. In line with its C4 photosynthetic evolution, the pathways identified in this study signify the interplays of biochemical reactions that make up the metabolic network constituting fundamental interface for sorghum to build defensive mechanism against drought stress. Multiple informants that we used in the gene prediction method prove to be reliable and dependable. This result entailed that 12.6% of the existing annotation has been modified and 1% novel gene models have been incorporated to the sorghum genome suggesting untapped natural genic and genetic variation in sorghum and its key position in agricultural economy and comparative genomics as a model for grass family.

## 2.1 Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is one of the few crops that uniquely sustain life and productivity under environments of intermittent or perpetual stresses from drought or water shortfalls. Several studies indicated that the unique adaptation of sorghum to such arid regions of the world may probably be credited to its recent C4 photosynthetic pathway evolution (Ghannoum *et al.*, 2002 and 2003; Miyao, 2003; Ripley *et al.*, 2007; Ghannoum, 2009), anatomical structure (Nguyen *et al.*, 2004; Harris *et al.*, 2007; Xin and Wang, 2011) and physio-biochemical processes (Lay and Anderson, 2005; Pagariya *et al.*, 2011 and 2012).

Previous studies have demonstrated an increase in sorghum productivity by improving resistance to pathogen infection and notorious weeds such as striga (*Striga hermonthica*) (Ejeta and Gressel, 2007). Others have demonstrated various aspects of sorghum functional information related to drought. These studies can be classified into four general categories:

(I) Research using traditional practices and indigenous knowledge identified largely morphological traits of diverse varieties based on local practices and knowledge (Teshome *et al.*, 1999; Abdi and Asfaw, 2005; Altieri, 2004 and 2009). Although this method uses various indigenous knowledge based selection criteria towards developing and maintaining genetic variability in traits of interest at the phenotype levels (Teshome *et al.*, 1999; Abdi and Asfaw, 2005), easily identifiable and heritable genetic variation could not be developed in drought tolerance as it lacks significant levels of interaction of genes with drought stresses (Edmeades, 2013).

(II) Conventional breeding that include diversity assessment and resource location (Teshome *et al.*, 1997 and 1999; Brush *et al.*, 2000; Subudhi *et al.*, 2000; Abdi *et al.*, 2002; Dillon *et al.*, 2007 and Ejeta, 2007) mainly focused on diversity studies and selection of the best trait aiming for higher yield under uniform and high input conditions. These studies, however, ignore varietal diversity, and the traits are less well adapted to changing environment and may be unable to cope with stresses such as drought (Teshome *et al.*, 1997; Abdi *et al.*, 2002 and Ceccarelli and Grando, 2007).

(III) Molecular breeding and QTL mapping (Tanksley and McCouch, 1997; Tuinstra *et al.*, 1997; Tao *et al.*, 2000; Xu *et al.*, 2000; Kebede *et al.*, 2001; Sanchez *et al.*, 2002; Eathington *et al.*, 2007) target QTLs responsible for a particular trait on a genomic region by linkage analysis and

association mapping. Others focus on the use of DNA markers that are closely-linked to target loci used as a substitute for phenotypic screening (Harris *et al.*, 2007). However, these studies mostly didn't engage various "omics" data sources generated under stress conditions that can potentially identify candidate genes responsible for specific or multiple stresses tolerance (Collard *et al.*, 2008).

(IV) Recent studies which largely include whole-genome sequencing, genome scanning, comparative genomics and transcriptomics applied high throughput data sets to describe sorghum related information. While genome sequencing determine the sequence of chemical base pairs (Childs *et al.*, 2001; Bedel *et al.*, 2005; Pratt *et al.*, 2005; Paterson *et al.*, 2009; The International Brachypodium Initiative, 2010; Mace and Jordan, 2011; Morris *et al.*, 2013), genome scanning identifies DNA markers linked to an inherited trait allowing genotyping and co-segregation of the markers (Zhang *et al.*, 2001; Harris *et al.*, 2007; Morris *et al.*, 2013). Moreover, comparative genomics, compares complex traits based on single or several tissues from different species to understand the functional basis of these traits (Draye *et al.*, 2001; Tuberosa and Salvi, 2006; Paterson *et al.*, 2009). Recently relatively few studies in sorghum such as Dugas *et al.* (2011) have shown efforts that described functional annotation of sorghum transcriptome in response to osmotic stress and identified more than 50 differentially expressed orthologs and (Mace *et al.*, 2013) that identified racial variation and domestication events in sorghum by analysing genomic sequences of wider geographic origin of several accessions.

All these data, however, suggest that relatively limited studies have reported on drought candidate gene identification in sorghum compared to other grass species like Arabidopsis (Xiong *et al.*, 2006; Atkinson *et al.*, 2013), barley (Ramalingam *et al.*, 2002; Tondelli *et al.*, 2006; Cseri *et al.*, 2011), maize (Ramalingam *et al.*, 2003; Xu *et al.*, 2014), pear millet (Sehgal *et al.*, 2012; Parvathaneni *et al.*, 2013), rice (Ramalingam *et al.*, 2003; Nguyen *et al.*, 2004; Yue *et al.*, 2006), sugar cane (Gupta *et al.*, 2010; Chandrakant, 2012) and wheat (Webster *et al.*, 2012; Diab *et al.*, 2013). Sorghum is known to have high genetic variability, however, the genes that play rate limiting roles in pathways controlling drought tolerance are not known. Approximately 50% of the 35845 existing protein coding genes lack experimentally validated information. For example, 10278 predicted genes (28.7%) are largely similar to or weakly similar to putatively uncharacterised proteins. Another 4324 genes (12.1%) represent predicted proteins and 3351 genes (9.38%) are similar to or weakly similar to expressed or putatively expressed proteins (Paterson *et al.*, 2009). In addition, the sorghum transcriptome (sorghum\_79\_annotation) was used to identify 27,608 transcripts of which

3984 transcripts represent unknown protein function. Thus, in the post-genomic era, assigning drought tolerance phenotype to any of these genes by gene knock-down experiment will be important for plant transformation and map-based cloning to improve sorghum drought tolerance and produce yield stability in drought affected area.

Traditionally, the candidate gene approach aims at dissecting a single drought gene in a pathway that contributes to a drought-response cascade and to measure its contribution to tolerance (Vinocur and Altman, 2005). Real progress in the study of drought tolerance at the gene level will require the identification and detailed analysis of many and possibly all components of the complex biological processes (Tyers & Mann, 2003).

Integrated *in silico* candidate gene approach (InsCGA) is the most promising method that allow for mapping expression data to metabolic pathways, gene expression profiling and orthology relation to investigate genes of interest. InsCGA considers functional features of the traits under study complemented by multi-analytical processes unlike genome wide association (GWA) study which usually overlooks such special features (Zhu and Zhou, 2007). Moreover, InsCGA employ efficient and reproducible study design that can provide a generic scheme for maximizing identification of promising candidate genes unlike traditional CGA which is mostly criticised for inefficient study design or suboptimal analytical approaches (Jorgensen *et al.*, 2009; Thomas, 2010).

During the past decade, unique genetic material such as near-isogenic lines (NILs) (Byrne and McMullen, 1996; Sanchez *et al.*, 2002; Harris *et al.*, 2007) were used to identify complex quantitative traits. Today, next generation sequencing technology has accelerated the identification of genes. For instance, Dugas *et al.* (2011) used RNA-seq data to identify genes responsive to osmotic stress and abscisic acid. Shakoor *et al.* (2014) used microarray data to identify and functionally characterize genotype-specific tissue expression in sorghum. However, genomic data sets such as a normalized library of drought-regulated expressed sequence tags (ESTs) also provide a well-defined view of the transcriptome (Pontius *et al.*, 2003), so called UniGenes (putative unique genes). Presently the sorghum gene space is represented by approximately 200,000 sorghum ESTs which has been clustered into approximately 22000 unique transcripts. These transcripts are grouped into about 15000 UniGene clusters representing more than 90 diverse libraries from several genotypes (Kresovich *et al.*, 2005). Each UniGene cluster, in addition to representing a unique gene, also includes information such as map location and the tissue types where these genes have been

expressed (Hoeven *et al.*, 2002). Therefore, the UniGene transcripts expressed under drought conditions together with its genomic location represent a collection of candidate genes (Irizarry *et al.*, 2005). This substantiates the importance of UniGene in InsCGA for identification and analysis of genes based on tissue, developmental stage and stress condition of the plant.

Furthermore, *in silico* candidate gene identification relies on an updated genome annotation. Genome annotation is updated dynamically as additional information on molecular and genome biology is obtained (Haas *et al.*, 2005; Haas *et al.*, 2011). For instance, the rice genome was reannotated at least four times (Ouyang *et al.*, 2007) and continue to be refined (Kawahara *et al.*, 2013), while the Arabidopsis genome has been annotated at least five times (Haas *et al.*, 2003 and 2005). However, between 2009 and 2013, the sorghum genome has undergone two annotation updates (Wang and Paterson, 2013). To our knowledge few sorghum studies have reported functional annotation using RNA-seq technology (Dugas *et al.*, 2011) or whole-genome sequencing to resequence the sorghum genome (Mace *et al.*, 2013).

There are two approaches routinely used to discover novel genes: first, *ab initio*, an intrinsic method is the most straight forward approach with no external input (Bonneau *et al.*, 2001; Korf, 2004; Lomsadze *et al.*, 2005; Stanke *et al.*, 2006a) that basically rely on a target genomic sequence. The prediction accuracy is limited and dependent on the user defined training set derived parameters for the underlying probabilistic model, often a Generalized Hidden Markov Model (GHMM) (Stanke *et al.* 2006c). The second approach employs extrinsic evidence-driven information to generate hint for finding genes based on similarity search (Stanke *et al.* 2006b). The current method used in our gene prediction pipeline is a combination of both intrinsic and extrinsic approaches that includes a validation protocol (Harrow *et al.*, 2009).

In this project, we embarked on a genomic approach to identify, characterize and prioritize sorghum candidate genes for drought tolerance using an integrated *in silico* candidate gene approach. The InsCGA executed multi-analytical processes that include sequence similarity search, metabolic pathways, differential gene expression profiling, identification of orthologs and functional gene ontology based gene enrichment. We set out to identify drought tolerant genes in the current sorghum annotation by mapping UniGene data obtained from drought resistant libraries. These genes were then functionally annotated. The sorghum genome was reannotated using the Program to Assemble Spliced Alignments (PASA) and publicly available experimental data. Gene models

were generated for mRNA reads that mapped to intergenic regions. This study presents unique approach and resources that complement existing efforts in sorghum research and potentially contribute to further understanding sorghum genomics and comparative studies.



UNIVERSITY *of the*  
WESTERN CAPE

## 2.2 Materials and methods

### 2.2.1 Data sources

Sorghum genome sequence, UniGene data and ESTs were used to identify drought responsive (DR) genes (Table 2.1).

Table 2.1 Summary of sorghum transcript and genomic data

Sequence origin	Sequence type	Number of sequences	GC %	URL
Genome	Chromosomes	10	41.6	<a href="http://www.phytozome.net/sorghum">http://www.phytozome.net/sorghum</a>
	Super scaffold	3394	40.9	<a href="http://www.phytozome.net/sorghum">http://www.phytozome.net/sorghum</a>
UniGene	Non-Clustered UniGene	199087	52.9	<a href="http://www.ncbi.nlm.nih.gov/UniGene/">http://www.ncbi.nlm.nih.gov/UniGene/</a>
	Clustered UniGene	14057	52.0	<a href="http://www.ncbi.nlm.nih.gov/UniGene/">http://www.ncbi.nlm.nih.gov/UniGene/</a>
ESTs	TIGR transcripts	209835	53.3	<a href="http://www.ncbi.nlm.nih.gov/dbEST/">http://www.ncbi.nlm.nih.gov/dbEST/</a>
	NCBI dbEST	20199	50.7	<a href="http://www.ncbi.nlm.nih.gov/dbEST/">http://www.ncbi.nlm.nih.gov/dbEST/</a>

#### 2.2.1.1 Genome Data

Genome assembly (sbi1, fasta format) and annotation data (sbi1.4, GFF file) were downloaded from the phytozome database (Goodstein *et al.*, 2012) (Table 2.1). The genome assembly was downloaded with 10 chromosomes and 3394 super-scaffolds which are small unmapped pieces of genome, that may or may not contain annotated genes and coordinates (Paterson *et al.*, 2009). The genome is represented as 697,578,683 base pairs arranged in  $2n=20$  chromosomes, 34,496 loci containing protein-coding transcripts and 36,338 protein-coding transcripts (Paterson *et al.*, 2009).

#### 2.2.1.2 UniGene Data

The UniGene database represent a collection of non-redundant stage-wise clustered and unified view of transcriptome that comprise ESTs derived from differentially expressed cDNA libraries (Pontius *et al.*, 2003; Rudd *et al.*, 2003). A total of 199087 UniGene sequences (build#29) were retrieved from NCBI UniGene (Table 2.1) of which 14057 sequences were unique and represented the longest sequence for each cluster. These included 41 clusters comprising ESTs derived exclusively from drought resistant libraries and 11353 clusters without any ESTs from drought resistant libraries.

#### 2.2.1.3 NCBI EST

A total of 20199 drought related ESTs were downloaded from the EST database (dbESTs) (Boguski

*et al.*, 1993) (Table 2.1). Based on EST data generated from drought stress experiments under differential expression conditions, 36 libraries were treated with water-stressed conditions at the pre-flowering developmental stage. The remaining 56 were treated under drought conditions at the post-flowering developmental stages targeted for stay-green traits. Sequences of a mixture of poly(A)+ RNA were organized in a total of 92 normalized cDNA libraries made of 48 body sites and 44 developmental stages of plant tissues grown under differential conditions (Table S2.1).

#### **2.2.1.4 TIGR plant transcripts**

A total of 209835 transcripts, which are all expressed sequence tags responsive to drought stress (DRESTs) were obtained from the TIGR plant transcript assembly database (Childs *et al.*, 2007) and were cross-checked for redundancies with dbEST from NCBI (Table 2.1; Boguski *et al.*, 1993).

#### **2.2.2 pre-processing (quality filtering process)**

Genome and EST sequences were screened for repeats, low complexity and vectors using RepeatMasker v. 3.0 (Smit *et al.*, 2012). A run of single pyrimidine or purines were identified using the DUST program (Morgulis *et al.*, 2006). Drought response phenotype information was obtained from the EST library description field to label ESTs within a UniGene cluster as a drought responsive EST (DREST). UniGene clusters were defined as follows: (i) DREST-only – all ESTs in the cluster were DREST, (ii) non-DREST clusters – non of the ESTs in the cluster were DREST and (iii) a mix of DREST and non-DREST.

#### **2.2.3 Mapping experimental data to reference genome**

The sorghum genome file was partitioned into its respective chromosomes (1-10) and more than 3300 super scaffolds using a python script. The partitions were used to minimize the size into each chromosome when mapping experimental sequences to the genome. The UniGene dataset and the TIGR ESTs were mapped to the sorghum genome in a two step approach namely: (i) UniGene dataset containing drought ESTs were mapped to the sorghum genome using EXONERATE and BLAT (Figure 2.1). Coordinates of sequences that mapped to intergenic regions were used as HINTs for AUGUSTUS (Figure 2.1). (ii) UniGene dataset and the TIGR ESTs were mapped to the sorghum genome using BLAT and then valid alignments were assembled by PASA to improve the existing gene annotations (Figure 2.1).



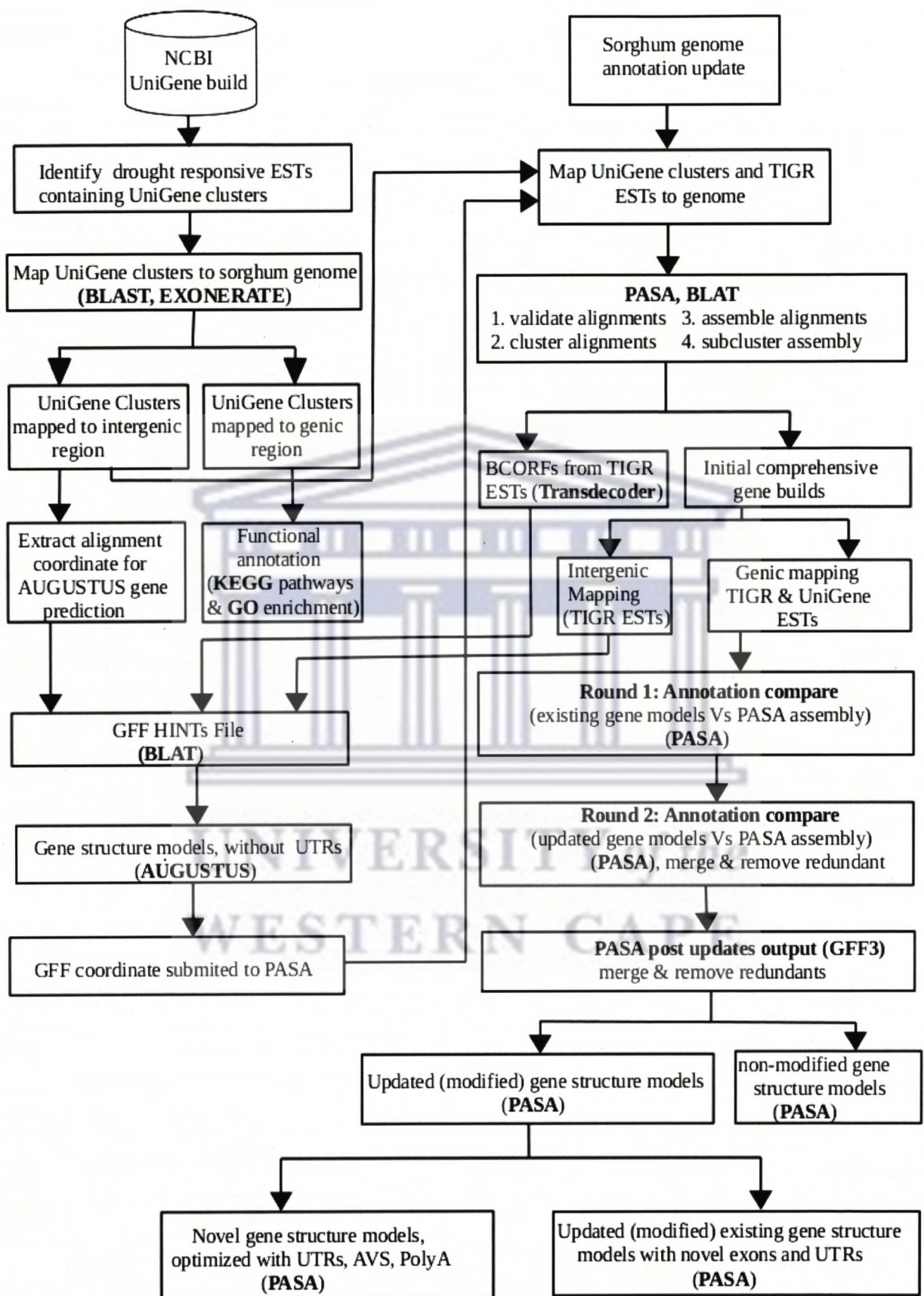


Figure 2.1: Pipeline for mapping experimental data to reference genome and annotation comparison

Keys to legend: BCUCs = best candidate UniGene clusters; BCORFs = best candidate open reading frames. A summarized description of an implementation of the pipeline for mapping genes and annotation comparison shown in

Figure 2.1 is given as follows: mapping UniGenes to genome using BLAST provides the best drought responsive hits (HSPs) that mapped to intergenic regions and the known sorghum genes. Almost all the known genes identified were putatively uncharacterised and they were selected for further functional annotation. The HSPs that were originated from intergenic mapping were consolidated along 2000nt spanning. The associated genomic regions were extracted for mapping back to the corresponding UniGenes by EXONERATE (Figure 2.2). The resulting best candidate UniGenes clusters were used by BLAT to generate HINTs which were in turn used by AUGUSTUS as an extrinsic experimental evidence in gene structure model prediction. Novel genes produced by AUGUSTUS were further optimized by PASA. The UniGenes that mapped to intergenic region by BLAST algorithm were also used as transcript input for annotation comparison by PASA to find out annotation updates. The initial step in the PASA pipeline was the cleaning up of any existing output in the MYSQL database by using utility codes existing as part of the program. This step was essential to enable the pipeline starting afresh. The process for annotation comparison was then started by running alignment assembly and by employing the minimum criteria for overlapping transcript alignments and for sub clustering into gene structure (Table 2.2). Comprehensive initial gene build was established by mapping valid alignment assemblies to genome. The gene builds mapped to the intergenic region that come from the TIGR transcripts were used by BLAT as the second input to generate HINTs for novel gene structure prediction by AUGUSTUS whereas the gene builds mapped to the genic region (existing gene annotation) were used for further annotation comparison. A two round approach was implemented by PASA for processing a complete annotation comparisons: first round, compared pre-existing gene structure annotations with alignment assemblies and second round, run it again, using the output of updated genes from the first round to capture a few more updates or to verify the initial updates if there is no further updates from the second round. The annotation comparison included analysis of alternative spliced alignments and identification of the best candidate ORFs (BCORFs) in PASA transcript assemblies using TRANSDECODER, a program built-in PASA. The BCORFs originated from TIGR ESTs were the third input for BLAT to generate HINTs.

Existing sorghum gene annotations were functionally characterised as hypothetical, putatively uncharacterised, or unknown proteins. The identification of drought responsive transcripts that overlap these existing annotated genes will add drought information and provide additional annotation coordinates that can potentially rectify sorghum gene annotations against existing gene models (Figure 2.3 and Table 2.4).

A total of 209835 TIGR transcripts (DRESTs) or TDRESTs and 10619 UniGene clusters (Table 2.1) were cleaned by a program called SeqClean (section 2.2.3.5) and then aligned to the sorghum genome using the PASA pipeline. Specifically, the first step in the PASA pipeline uses BLAT (Kent, 2002), a pre-installed program required by PASA (PASA2 v. PASA2-r20130425beta; Haas *et al.*, 208) to align transcripts to the genome. An in-built assembly function within PASA was triggered after the transcripts were aligned to the genome and resulted in 5970 assemblies out of 16835 validated TDREST alignments. Similarly, 749 PASA assemblies from 756 validated UniGene cluster alignments were obtained. The PASA assembly was undertaken once after clustering the

alignments into groups and reassigning them using the validated coordinates of the alignments. Transcripts that aligned to the genome were retained if they met the following threshold: greater than 95% identity and 90 % alignment coverage. PASA output included initial comprehensive gene builds (ICGBs; GFF format) that were mapped to intergenic region. These gene builds were used as input to BLAT to generate a “HINT” file for AUGUSTUS. The other PASA output used to generate a “HINT” file was the product of TRANSDECODER, the BCORFs. Transcripts that do not map to existing sorghum gene annotation, but to intergenic regions were extracted and analysed according to the methods outlined in section 2.2.3.1.

One of the modifications in the existing annotation is the change in structural and positional categories of the existing gene models. The candidate gene models with structural and positional modifications are described as follows: i) *Bidirectionally extended overlapping genes*: Transcripts were identified as having 3' and 5' ends extended in both direction and overlapping with the existing genes; ii) *Unidirectionally extended overlapping genes*: these overlapped the existing annotation and unidirectionally extended just on one end of the gene structure generating either 3' or 5' end extended gene but not both; iii) *perfect overlapping genes*: the current gene model were identified as exactly matching the coordinates of the existing genes; iv) *partial overlapping genes at the 5' end*: the current gene models share start coordinate with the existing genes. v) *partial overlapping genes at the 3' end*: the new gene models structure share the 3' end with existing gene model; vi) *Inner overlapping genes*: the new gene modes fall exclusively within the range of the existing gene models (Figure 2.3a).

### **2.2.3.1 Building gene models in the intergenic regions**

UniGene sequences were blast searched against the genome, to identify sequences that map to intergenic regions using the following parameters: an e-value cutoff  $1e-10$ ; high scoring alignment pairs (HSPs) or hits with at least 80% identity over the entire length of query. HSPs corresponding to the same query were retained if they span a maximum of 2000bp. Raw blast output was parsed using in house perl script (Blast\_parser.pl) to identify UniGene sequences that overlapped existing gene annotations and those sequences that mapped to intergenic regions. UniGene sequences that mapped to intergenic regions were retained even if these sequences did not correspond to DRESTs.

Pipeline for identification of candidate genes through mapping UniGene to genome

Codes, commandlines and programs for running the pipeline

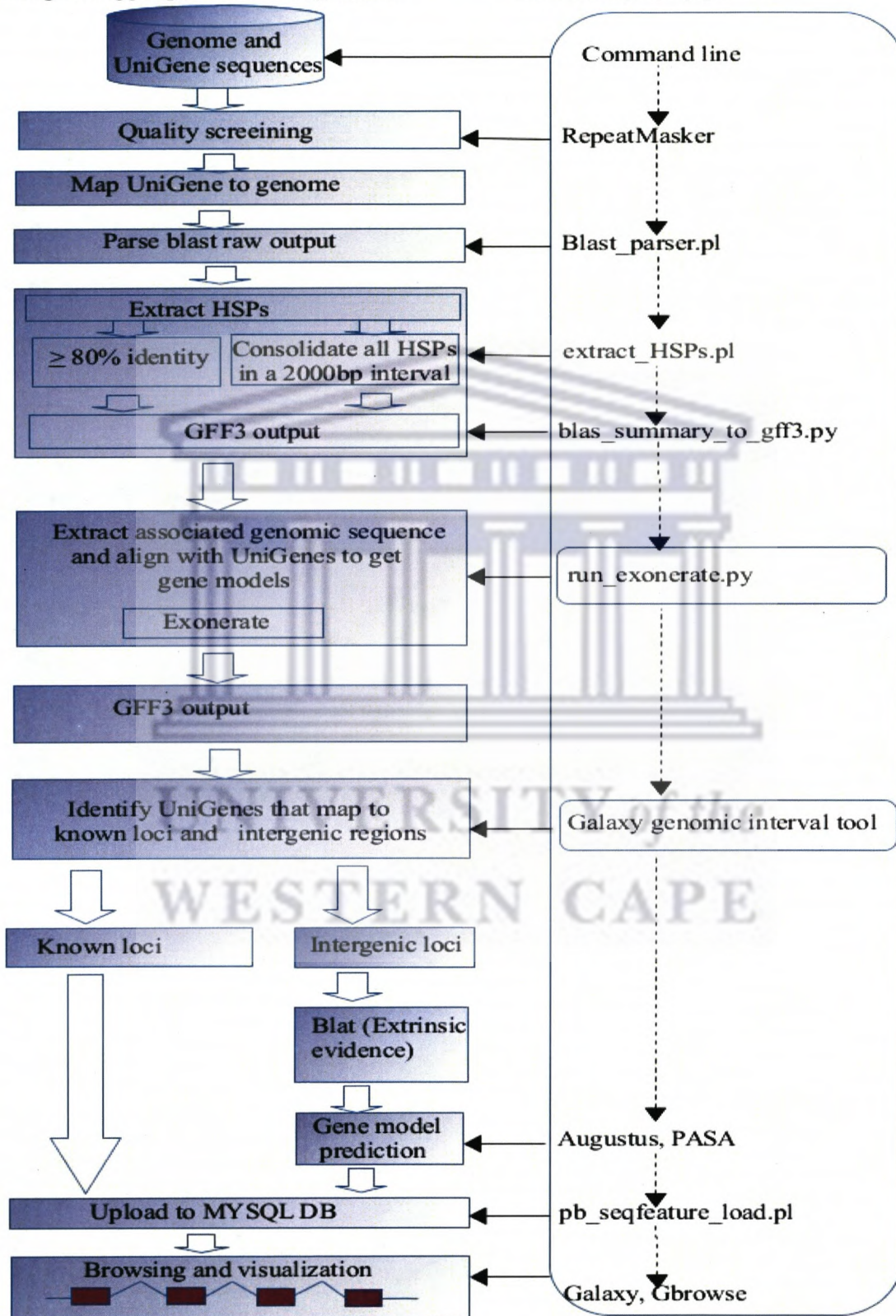


Figure 2.2: Pipeline for building gene structure models.

UniGene EST sequences were used to map drought responsive genes to sorghum genome (section 2.2.3). Genome and UniGene sequence including DRESTs were downloaded as described in section 2.2.1. The sequences were screened for quality using RepeatMasker (see also section 2.2.2). Repeat masked UniGene sequences were blasted against genome using e-value cutoff  $1e-10$ . The raw output was parsed and HSPs were extracted using different perl script. Percent identity with  $\geq 80\%$  was used to select the HSPs. The HSPs originated from same query sequences were consolidated (see sections 2.2.3 and 2.2.3.3) along the genomic length of 2000bp and were converted into GFF3 formats to extract associated genomic region. These were aligned to the corresponding UniGene sequences using Exonerate. Galaxy genomic interval tool was used to get intersected and subtracted data sets which were classified as known and intergenic region (novel loci) respectively. Genes were visualized by loading the GFF3 formatted files of the candidates and the genome annotation onto the MySQL DB using a perl script (pb\_loadfeature.pl). All scripts can be downloaded from <ftp://adugnasorghum:data88notweed@ftp.sanbi.ac.za>.

The genomic coordinates of these HSPs were extracted and converted to GFF3 format using a perl script ('extract\_HSP.pl'). The output was summarized by a python script called 'blast\_summary\_to\_gff3.py' and was used as an input by EXONERATE (Slater and Birney, 2005; 'run\_exonerate.py') to extract the genomic segments from the masked genome and to align with the corresponding UniGene sequences. The resulting genomic coordinates were converted to GFF3 formatted files (Figure 2.2). GFF3 formatted UniGene file and the sorghum genome annotation GFF3 files were loaded to the galaxy genomic suite (Goecks *et al.*, 2010) using the "Get Data" option. The UniGene data was compared with the genome annotation to find the known genes that correspond to the mapped UniGene sequences using the "Compare two Datasets" option. Intergenic (novel) loci were identified using the "Subtract Whole Dataset" in the galaxy genomic suite.

The GFF3 formatted output of the BCUCs were used as input to BLAT (Kent, 2002) to generate an alignment file which was in turn used as a hint by AUGUSTUS (Stanke *et al.*, 2006b). This was used along with sorghum genome and expression data, maize parameter and meta parameter as extrinsic evidence in a sequence homology search and combined with intrinsic evidence (ab initio method, Stanke *et al.*, 2006a) in gene prediction (Figures 2.1 and 2.2).

### **2.2.3.2 Annotation Comparison**

The PASA pipeline was used to compare the existing sorghum genome annotation with the new genome mapping coordinates derived from the DRESTs. Of the available tools, PASA can be used to report differences between existing and newly created annotations (Yandell and Ence, 2012). Table 2.2 shows the parameters set in the PASA pipeline for the annotation comparison and minimum full length ORF size. Based on the parameters, all the valid single gene model updates

that retained PASA assembly reference id were computed and compared with the existing gene structure models. We used the term “update” to explain annotation modification (improvements or addition of new features) of anyone or multiple genes, transcripts, exons, CDS and UTRs of the existing annotation as structural attainment based on spatio-temporal and differential expression data. The annotation update in this thesis included: (1) modification of the existing annotation, and (2) discovery of novel loci.

Table 2.2 Parameters used for the annotation comparison in the PASA pipeline

No.	Annotation comparison		Minimum full length ORF size	
	Parameters	Minimum %	Parameters	Minimum value
1	Genomic overlap	50	Annotation version	2
2	Protein coding	40	Maximum utr exons	2
3	Length for non-fulllength compare	70	Compare ID	2
4	Length for full-length compare	70	Trust full length status	0
5	Predicted protein compare	70	stomp	0
6	Alignment length	70	Minimum % overlap	80

The PASA pipeline uses built-in dependency alignment tools such as BLAT, GMAP and BLAT-GMAP as default aligners. In this prediction, BLAT was used because of the reasons outlined in section 2.2.3.4. The default values used for the thread number of the pipeline, the number of top scoring spliced alignments and the minimum % overlap of the transcripts to be clustered were equivalent to 'two', 'one' and 30 respectively.

### 2.2.3.3 Prediction of gene structure models using AUGUSTUS

BLAT was used as an AUGUSTUS dependency alignment tool because it is more accurate and much faster than existing tools. It uses '-ooc=11.ooc' option that tells the program to load over-occurring 11-mers from external file which basically increases the speed by a factor of 40 (Kent, 2002). For mRNA/DNA alignments, it allows extension of all perfect hits, stitches homologs into single larger alignment unsplicing mRNA on to the genome that uses each base of the mRNA only once which correctly positions splice sites (Kent, 2002; Li and Durbin, 2010). The three types of initial gene set were used by BLAT to generate hints. The initial gene sets are namely: best candidate UniGene clusters (BCUCs) generated by EXONERATE from UniGene clusters mapped to intergenic regions, BCORFs generated from the TDRESTs alignment assemblies predicted by TRANSDECODER in PASA pipeline and ICGBs predicted by PASA based on the alignment assemblies mapped to intergenic regions in genome annotation comparison. BLAT initially

produced “\*.psl” formatted files at a DNA sequences homology with  $\geq 95\%$  identity and the default coverage of 80%. This was then sorted by using pslSort program and command line “sort -K 10,10” and then used pslReps to select the best alignments which was finally subjected to UCSC standard tool pslCDnaFilter to filter again the alignments and report only the top HSPs for each UniGene EST sequence before the last run of BLAT to create hints. The setting of parameters for pslCDnaFilter was based on the EST/mRNA of the UniGene track construction protocol in BLAT software. The parameters are: – minId = 0.95 – minCover = 0.25 – localNearBest = 0.001 – minQSize = 20 – minNonRepSize = 16 – ignoreNs – bestOverlap – polyASizes = polyAFile, where polyAFile was generated by UCSC program faPolyASizes. Hints, evidence driven files originated from the best filtered alignments based on expression data, were then produced by BLAT using AUGUSTUS utility script, blat2hints.pl, for use by AUGUSTUS program for gene structure prediction.

The following parameters were used for running AUGUSTUS stand-alone software: AUGUSTUS – species=species –hintsfile=hints.E.gff –extrinsicCfgFile=extrinsic.ME.cfg genome.fa. Species and genome were set to represent sorghum according to the options given in the program. Hints were separately used by AUGUSTUS to predict the gene model structure and the out put were pooled together. AUGUSTUS either accept or ignore a hint depending on the level of compatibility and reliability of the hint to predict a gene structure (Stanke *et al.*, 2006c) whereby the gene predicted was assigned to *ab initio* for hints which were not compatible. A combination of *ab initio* and homology based prediction (Bonneau *et al.*, 2001; Zhang, 2008; Walsh *et al.*, 2009) were used to identify potential novel candidate genes.

#### **2.2.3.4 Consistency in gene predictions**

The consistency in gene prediction by the AUGUSTUS was checked using multiple data sources selected based on sequences mapped to the intergenic regions. The results in the bitscore in AUGUSTUS prediction from each datasets were compared and the evidence support were used to show consistency in gene prediction. These were used to evaluate the novelty of genes structure model prediction in combination with the parameters used for screening gene models (see section 2.2.3.5).

#### **2.2.3.5 Filtering the gene model structures**

The following parameters were used to filter the novel gene structures. I) genomic coordinates of

the novel genes in relation to the intergenic distance between existing nearest neighbouring genes and the predicted genes; II) length of the predicted genes; III) score of the predicted genes; IV) percentage evidence support where prediction was homology; V) Strand orientation of the predicted genes in relation to the existing genes or the currently predicted genes if they were neighbourhood and VI) the gene prediction confidence agreement (based on the criteria for gene confidence used by Broad Institute, <http://www.broadinstitute.org/>). The parameters are not necessarily in order of their weight. Each of the screening criteria contributes to the novelty of the gene structure models. However, we used curiously the genomic coordinates as the primary screening parameter so as no any novel gene has overlapping coordinate with existing gene models. Coordinates for all known sorghum genes were obtained from phytozome (release v2.1, Sbi1.4, the latest release) to compare with the genomic coordinates of the AUGUSTUS gene models. This was done only after the novel gene models were optimized by PASA. Genes satisfied any of the four listed criteria were considered valid. Manual curation and post PASA update structural annotation of the novel structure models were conducted.

#### **2.2.4 Metabolic pathway analysis**

Biochemical pathway analysis was performed using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa *et al.*, 2000) which is supported by BLAST2GO database and software (Conesa *et al.*, 2005). A total 123 UniGene sequences that mapped to the sorghum genome and overlapped known genes were searched against the BLAST2GO databases using the BLASTX search algorithm with the e-value cut-off  $1e-10$  as a default parameter. The number of hits and the HSPs length cut off value per query sequence were set to 50 each. EC-weight was set at 1 or 0 depending on whether or not the influence of the evidence codes on the GO annotations is required or ignored (eg. IEAs) respectively (Conesa and Götz, 2009). A list of Enzyme Code (EC), KEGG pathway maps, interpro annotation and statistics, GO annotation and combined graphs for GO-domains: BP, CC, MF were identified by BLAST2GO. Gene enrichment analysis for genes mapped to metabolic pathways and Interpro domains was carried out as described in the BLAST2GO based GO enrichment protocol, section, 2.2.5.1.

#### **2.2.5 Gene Ontology functional enrichment analysis**

##### **2.2.5.1 GO functional enrichment analysis using BLAST2GO**

GO enrichment annotation was configured to e-value cut-off  $<1.0e-6$ . In addition, default values



were considered for the annotation parameters such as annotation cut-off = 55, a GO-weight=5 and HSP-hit coverage = zero. Based on the BLAST2GO software package guide-line, HSP-hit coverage greater than zero may create chances of missing any best hit from the HSP spans (Conesa *et al.*, 2005). Once setting the parameters, BLAST2GO employed a BlastX program, to search for matching nucleotides against NCBI non-redundant database. Each UniGene/EST sequence was assigned a GO term and an InterPro domain identifier. The occurrence of GO terms assigned to each UniGene were compared to occurrence of the background set of GO-annotated transcripts in the entire database using a hypergeometric distribution. Gene ontology domains namely biological process, cellular components and molecular function based tree-type combined-graph was configured using default values provided by BLAST2GO for all enriched GO terms (adjusted p-value < 0.05). Mapping was performed to associate the blast HSP-hits to functionally enrich information from GO DB. Because, BLAST2GO basically relies on resources stored in GO DB which is linked to functional information from NCBI, PIR and GO, and all query protein Ids for mapping are linked to repositories of millions of functionally annotated gene products of several hundreds of species (Harris *et al.*, 2004). All annotations were associated to an evidence code which provides information about the quality of this functional assignment (Camon *et al.*, 2004). Default parameters were used to assign InterPro domain and GO term to identified gene models. Sorghum peptides were selected for the occurrence of functional motifs and protein signature for which statistical significance of over-representations of each GO term exist. Enrichment status or over-representation of GO terms were checked using Fisher's exact test in comparison to the background set based on p-values. The gene set with the lowest p-value represent a significance level of enrichment. Terms representing all the GO domains were used in annotation for the enriched ones with adjusted p-value (False Discovery Rate (FDR),  $p < 0.05$ ).

#### **2.2.5.2 GO functional enrichment analysis using AGRIGO**

GO enrichment analysis for candidate known genes identified by BLAST sequence similarity search based on mapping UniGene clusters to sorghum genome was performed using AGRIGO (Du *et al.*, 2010), a web-based tool and database for the gene ontology analysis. This was compared with the result performed using BLAST2GO. Query sequences of a total 123 known genes that matched the same total (123 UniGene clusters) were used as an input for AGRIGO to evaluate the genes to which the enriched GO terms were assigned. The number of genes associated to enriched GO terms were then compared to the total number of genes obtained the Interpro information from BLAST2GO analysis (section 2.2.5.1).

Similarly, GO enrichment analysis for the genes identified by other two underlying processes (analysis of expression profiling, section 2.2.6 and analysis of orthologous groups, section 2.2.7) were performed using AgriGO (Du *et al.*, 2010) separately each after the candidate genes were identified. Singular Enrichment Analysis (SEA) (Huang *et al.*, 2009), a version of Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) was performed based on enrichment of the GO terms obtained after comparing the statistical test with pre-calculated background set. GO term enrichment and the number of genes mapped to the enriched terms were determined by Parametric Analysis of Gene Set Enrichment (PAGE) using a Z-score value which eventually was converted to the p-value for correction (Benjamini and Huchberg, 1995; Benjamini and Yekutieli, 2001) inferring the statistical significance of the GO term enrichment. AGRIGO allows checking for enrichment status of GO terms using Fisher's exact test as a default against the background set (Du *et al.*, 2010) based on p-values. Adjusted p-value, FDR,  $p < 0.05$  was used to determine the significance level of enrichment. The gene set returned with p-value lower than 0.05 were retained.

The final set of CDRGs associated with all GO-terms with direct or indirect correlation with drought stress responses were selected based on the Biological process, Cellular components and Molecular functions. The GO term descriptors that related to drought tolerance included but not limited to “response to drought stress”, “response to desiccation tolerance”, “response to water deprivation”, “cellular response to drought stress”, “cellular response to desiccation” and “cellular response to water deprivation”. Mapping of the GO-terms related to responses to stress based on biological processes was configured by tree traversing mode.

### **2.2.6 Use of expression profiling for candidate gene identification**

To investigate potential candidate genes that respond to drought stress conditions, we analysed expression data for sorghum orthologs in maize, a closely related cereal crop. Maize RNA-seq expression data under drought stress was downloaded from Gene Expression Omnibus (GEO) (Kakumanu *et al.*, 2012) employing spatio-temporal analysis to determine tissue and stage-specific expression and drought stress condition. A software package, TIGR MultivariateExperiment Viewer (MeV) (MeV\_4\_8\_1) expression-data analysis tools (Saeed *et al.*, 2003) was used to analyse the differentially expressed genes and visualize the results as heat maps. Maize genes that were over-expressed ( $\geq 2$ -fold RNA-seq) under drought stress were used to identify orthologs in sorghum based on the ortholog pairs recorded in the ENSEMBL Biomart database (Smedley *et al.*, 2009). Sorghum orthologs corresponding to maize over-expressed genes were captured and used as input to AGRIGO (Du *et al.*, 2010) to identify the functions of these orthologs based on GO term

enrichment (FDR,  $p < 0.05$ ; section 2.2.5.2).

### 2.2.6.1 Statistical analysis of gene expression

In order to identify sorghum genes that drought responsive, we used maize as a test case because maize is the closest ancestral cereal crop to sorghum. Specifically we identified maize genes expressed in leaves and ovary tissues that were drought responsive (published data) using multivariate analysis of variance. Statistically significant over-represented genes were identified using parametric and non-parametric analyses. These genes were used to find orthologs in sorghum and validated using GO enrichment analysis.

Significant differences in the gene expression level was evaluated by employing an unpaired t-Test (Baldi and Long, 2001; Huang *et al.*, 2008) to estimate between subject variance. Non-parametric Fisher's exact test (Agresti, 2007; Bullard *et al.*, 2010) was used to evaluate the effect of treatments on the gene expression outcome, and a FDR calculation (Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2003) for genes identified at  $p < 0.05$  were performed. Rank products, a non-parametric statistical method (Breitling *et al.*, 2004; Hong *et al.*, 2006) was employed to minimize the discrepancy between the actual and false discovery of differentially expressed genes. Tissue and treatment based groupings of the samples were employed to determine the effect of these parameters on the gene expression. The treatments used in this analysis represent drought stress and well-watered condition while tissue types are fertilized ovary and basal leaf meristem (Kakumanu *et al.*, 2012).

### 2.2.7 Analysis of orthologous groups

A total of 9693 sorghum UniGene clusters out of a total of 14057 UniGene clusters that contain one or more drought responsive ESTs (see Table 2.1 for the data source) was used for orthology analysis. Sorghum drought responsive orthologs were identified in the distantly and closely related cereal crops namely Arabidopsis, rice and maize (placed in the order of increasing evolutionary proximity to sorghum). Sorghum orthologs in these three species were retrieved from the ENSEMBL compara database using ENSEMBL BioMart (Smedley *et al.*, 2009).

Percent identity and orthology confidence levels were used as parameters to retrieve matching orthologs. All available homology types (one2one, one2many and many2many) that have more than 50% identity and high level orthology confidence (1) as a threshold value cut off were considered for selecting the best quality orthologs. These were then used as input for GO enrichment analysis using AGIRGO (see section 2.2.5.2 for a description of the GO annotation protocol).

## **2.3. Results**

This chapter is aimed at identifying novel genes and characterising existing genes as drought tolerant using experimental expression data in sorghum. All drought responsive genes were functionally annotated. Sorghum genome annotation was improved by PASA.

### **2.3.1 Mapping experimental data to reference genome**

#### **2.3.1.1 BLAST Sequence Similarity Search: Identification of candidate drought responsive genes**

Of the 14057 UniGene clusters used as the query sequences, 10619 were mapped to the genome by exonerate whereas 3378 did not due to reasons such as mapping error, low complexity, contamination, etc. Of the UniGene clusters that mapped to the genome at a threshold level of  $\geq 80\%$  identity, 9763 overlapped with the existing gene models. Among those that overlapped with the existing gene annotation, 123 UniGene clusters (1.2%) represent purely drought responsive. The remaining 9640 UniGenes clusters that mapped to existing annotations were non-drought responsive sequences except for 258 relatively short ESTs that were identified as drought responsive interspersed. On the other hand, 856 UniGene clusters did not overlap with existing annotated gene models and were considered to be novel hits. These set of gene loci included 128 (1.3%) clusters that represent drought responsive sequences (Table S2.2).

#### **2.3.1.2 Reannotation of sorghum genome**

PASA updates were prediction of novel structures of the known or existing gene models. Based on annotation comparison, 4 separate genes Sb03g045450, Sb03g045460, Sb04g008510 and Sb04g008530 from two chromosomes merged into two new genes namely Sb03g045450\_Sb03g045460 on chromosome 3: 72720937 – 72725839 and Sb04g008510\_Sb04g008530 on chromosome 4: 9869026 – 9888743. These genes transcribed into corresponding 2 merging transcripts: Sb03g045450.1\_Sb03g045460.1 chromosome 3:72720937–72725839 and Sb04g008510.1\_Sb04g008530.1 chromosome 4: 9869026–9888743. Furthermore, a novel transcript, Sb04g007110.2.1 on chromosome 4: 7175432–7182182 was identified as an additional isoform of the gene Sb04g007110 in the current annotation that was not present in the existing annotation. While the two other transcripts of the same gene such as Sb04g007110.2 and Sb04g007110.3 were still identified to be valid single gene model updates, a transcript known as 'Sb04g007110.1' remain without PASA-modified. In addition, the current annotation update

includes 64 novel exons, 74 five prime UTRs and 3595 three prime UTRs (Table 2.3; Table S2.3).

### 2.3.1.3 Annotation comparison and an update

A comparison between the current and existing sorghum genome annotation has resulted in a total of PASA improved 4349 genes and 4447 mRNAs (Table S2.3 and Figure 2.3). This makes a 12% PASA updates out of the total non-redundant 36337 mRNAs leaving the other 31890 existing transcripts unupdated (Table S2.3). On the other hand, among 441 genes initially predicted as novel, a total of 241 gene models were filtered and optimized by PASA (section 2.2.3.6 and Figure 2.1). This result was based on evidences used from three initial gene sets that mapped to intergenic region namely 856 best hit UniGene clusters, 500 BCORFs selected from the top best long ORFs and 520 initial comprehensive gene builds. The last two evidences were obtained from the 20199 TDRESTs analysed by PASA (Table 2.6). Merging genes, transcripts and different isoforms and new exonic and UTR features were identified and contributed to annotation update (see section 2.3.1.2; Table 2.3).

Table 2.3 Comparison and update of annotation between existing and current prediction

Source data	Input Sequence (seq no.)	3' UTR	5' UTR	Exon	Transcript		Gene (merged)
					merged	novel	
TIGR transcript	20199	3503	39	28	2	1	2
UniGene cluster	10619	92	35	36	2	1	2
Total	30818	3595	74	64	2 (unique)	1(unique)	2 (unique)

### 2.3.1.4 Structural and positional modification of the candidate genes

In this study, 95.4% of the modified genes represent perfect overlapping with the existing gene model sharing start position @ 5' and stop at 3' ends (Figure 2.3Transcript D), 85 genes (1.9%) represent extended overlapping @ 5' ends sharing the end position @ 3' (Figure 2.3Transcript B). Again, 96 genes (2.1%) represent an extended overlapping @ 3' ends sharing the start position @ 5' (Figure 2.3Transcript C). Interestingly, 21 existing genes (0.5%) were modified to be extended bidirectionally both at the 5' and 3' ends (Figure 2.3Transcript A). Still, two genes (Figure 2.3Transcript G and H) were noted each representing partial overlapping with the existing gene models at 3' and extended at 5' and the reverse structural patterns respectively. Only a gene in two different positions each was modified with structural pattern exhibiting partial overlapping @ 5' end and extended at 3' (Figure 2.3 Transcript F) and the inner overlapping position (Figure 2.3 Transcript I). While the majority of modified genes represent perfect overlapping, no gene was

detected flanking @ the 5' end and partially overlapping @ the 3' (Figure 2.3 Transcript E). Four separated existing protein coding genes with the transcribed mRNAs merged into two new genes and the corresponding mRNAs (Figure 2.3B; Gene B, C and D; Table 2.3). In this study, a novel mRNA 'Sb04g007110.2.1.1' was also identified from a parent gene 'Sb04g007110' previously identified to have similar function with Zinc finger transcription factor (Table 2.3). Of all the total 6719 extrinsic input from the three initial gene sets that mapped to the intergenic regions, 241 (3.6%) represent novel genes for drought tolerance (Figure 2.3C; Gene F). For detail description of novel gene prediction, refer to section 2.2.3.4 and 2.3.1.5.

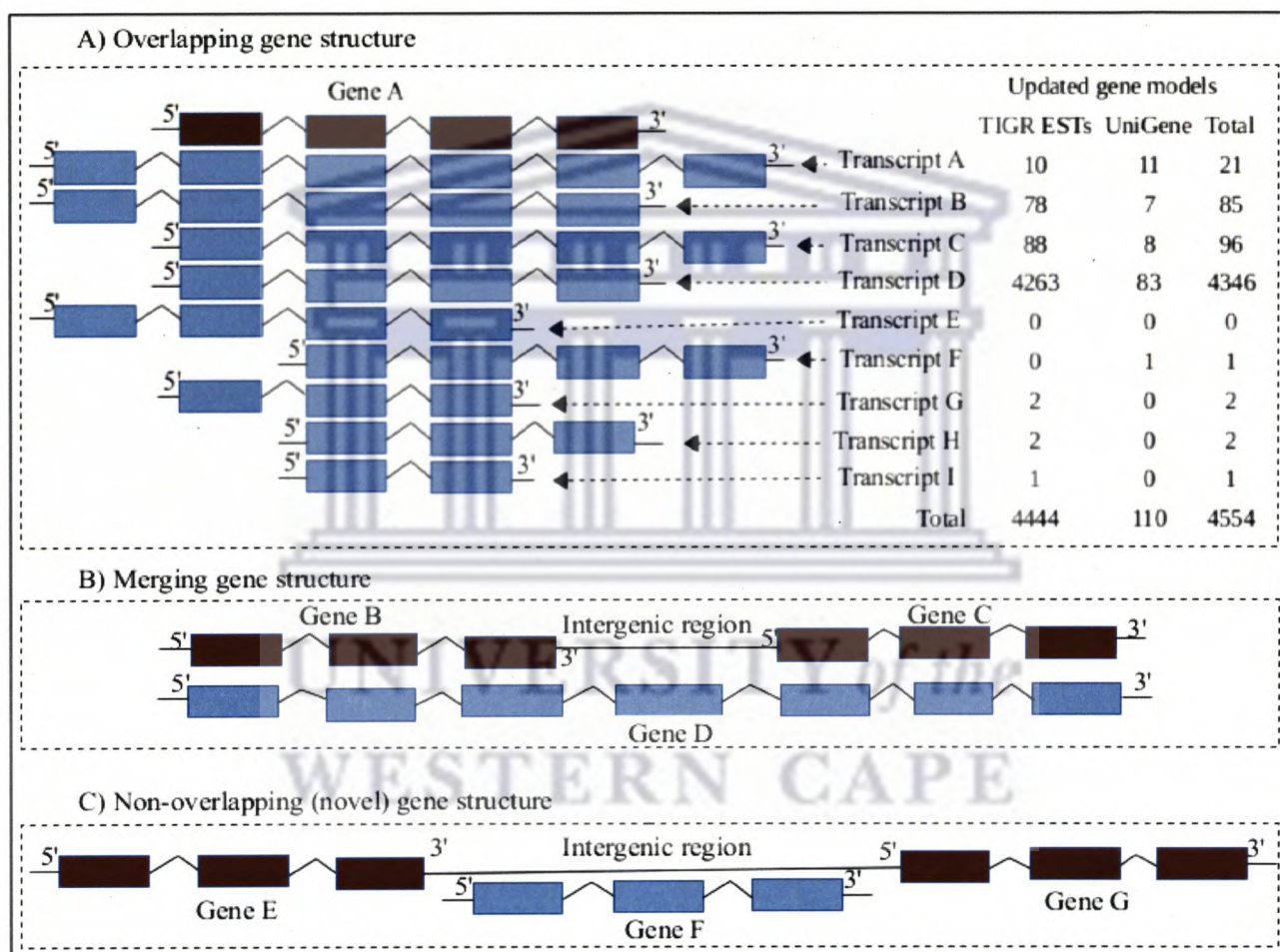


Figure 2.3: Schematic gene structure model for annotation comparison showing modified and novel genes. In this figure A) represent the current modified gene structure models overlapping existing genes: Gene A = represent an existing hypothetical gene structure model against which all overlapping genes that currently identified were assumed to be mapped. Transcript A represents an extended overlapping gene @ both 5' and 3' edges; Transcript B represents an extended overlapping gene @ 5' edge sharing start position @ 3' edge; Transcript C denotes an extended overlapping gene @ 3' edge sharing start position @ 5' edge; Transcript D represents perfect overlapping gene that conform or share start position @ 5' and stop @ 3' edges; Transcript E represents a partial overlapping @ 3' and extended overlapping @ 5' edge; Transcript F represents a partial overlapping @ 5' and extended @ 3' edge; Transcript G represents a partial overlapping gene that conform or share start position @ 5' edge; Transcript H represents a partial overlapping gene that

conform or share start position @ 3' edge; and Transcript I represents an inner overlapping gene. The values given corresponding to each overlapping transcript in Figure 2.3A describe the actual number of existing genes modified in our prediction based on TDERSTs and UniGene datasets. Figure 2.3B represents merging gene structure models where 'Gene B' and 'Gene C' were assumed to be merged into 'Gene D'. Figure 2.3C represents non-overlapping (novel) gene structure model that mapped to intergenic region: 'Gene E' represents the left nearest neighbouring existing gene model to the novel 'Gene F'; 'Gene F' represents a non-overlapping gene that mapped to the intergenic region between 'Gene E' and 'Gene G'. 'Gene G' represents the right nearest neighbouring existing gene model to the novel 'Gene F'. The gene names denote arbitrary example. Each bar represent exon structure and the inverted 'v' shaped structure positioned between any two adjacent bars represent intron splicing. The gene model structure with red bars denote existing gene models and those with blue are assumed to represent the currently identified genes that match existing models (transcript A-I), merging gene ('Gene D') and novel gene model ('Gene F'). This schematic gene structure model assumes both strand orientations based on the patterns of loci observed overlapping with the exiting annotation in our result.

Table 2.4 Chromosomal distribution of the modified existing genes models

Scaffolds	Chromosomal distribution of the existing modified gene models									Total
	A <sup>1</sup>	B <sup>2</sup>	C <sup>3</sup>	D <sup>4</sup>	E <sup>5</sup>	F <sup>6</sup>	G <sup>7</sup>	H <sup>8</sup>	I <sup>9</sup>	
Ch1	3	16	24	868	0	0	1	0	0	912
Ch2	6	11	7	535	0	0	0	1	0	560
Ch3	2	19	14	645	0	0	0	1	0	681
Ch4	2	13	11	525	0	0	1	0	0	552
Ch5	2	3	7	157	0	0	0	0	0	169
Ch6	2	8	5	367	0	0	0	0	0	382
Ch7	0	5	10	287	0	0	0	0	0	302
Ch8	0	4	4	206	0	0	0	0	0	214
Ch9	3	2	3	371	0	1	0	0	0	380
Ch10	1	4	11	375	0	0	0	0	1	392
Super	0	0	0	10	0	0	0	0	0	10
<b>Total</b>	<b>21</b>	<b>85</b>	<b>96</b>	<b>4346</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>2</b>	<b>1</b>	<b>4554</b>

Key to legend:

A<sup>1</sup> Transcript A; B<sup>2</sup> Transcript B; C<sup>3</sup> Transcript C; D<sup>4</sup> Transcript D; E<sup>5</sup> Transcript E;

F<sup>6</sup> Transcript F; G<sup>7</sup> Transcript G; H<sup>8</sup> Transcript H and I<sup>9</sup> Transcript I (Figure 2.3)

### 2.3.1.5 Novel gene structure model prediction

We initially identified 414 novel genes using AUGUSTUS gene prediction program and PASA pipeline. We then subject these to series screening procedure and filtered out 241 (Table 2.5). Among these, 14 genes were complete gene model structure possessing both 3' UTR and 5' UTR

edges and 18 genes with only 3' UTR and 2 genes with 5'prime. This entails that 34 genes have got at-least one end complete or semi-complete genes structure ether with only 3' UTR edge or 5' UTR or both. The remaining 207 were partial genes without any UTR segment but with the start and stop codons. Of all the total novel genes identified in this study, 115 were found to be single exonic and were contributed to the identification of 47.7% intronless genes. Other genes included those having 2 exons (51 genes, 21.2%), 3 exons (40 genes, 16.6%), 4 exons (17 genes, 7.1%), 5 exons (10 genes, 4.2%), 6 and 7 exons (3 genes each, 1.3%) and 2 nine exonic genes (1%) with no eight exonic (Table 2.12). Table 2.5 provides the size of novel genes functionally related to drought responses and also gives statistics for the complete and partial gene structure models.

Table 2.5 Functional distribution of the novel gene structures model

Known drought-related function	Complete genes		Partial genes			Total
	3' and 5' UTRs attained	Only 3' UTR attained	Only 5' UTR attained	at-least end attained	one truncated at both ends	
DR <sup>1</sup>	14	18	2	34	144	168
NDR <sup>2</sup>	0	0	0	0	37	73
Total	14	18	2	34	171	241

Key to legend: <sup>1</sup> Drought responsive genes; <sup>2</sup> Non-drought responsive genes

Genes were scanned for pseudogene behaviour or false coding sequence based on the criteria used by Ensembl (Ensembl Gene Set). A transcript is considered to be a pseudogene if 1 of these 4 criteria are met, however not limited just to these criteria: 1) It is a single exon transcript and it matches a multi-exon transcript elsewhere in the genome. (2) The transcript is completely masked out by repeat masker. 3) The transcript contains no introns and multiple frame shifts and 4) The transcript contains frame shifts and all the introns are > 80% covered by repeats.

Table 2.6 Distribution of novel genes based on the method of prediction

Source data Type of initial datasets	Prediction method				
	Extrinsic inputs for AUGUSTUS		Homology	<i>ab-initio</i>	Total
	Hints	Total			
UniGene Clusters	BCUCs	856	74	2	76
TIGR transcripts	BCORFs	500	32	105	137
	ICGBs	520	24	4	28
Total		1870	130	111	241

Of the predicted 241 novel genes, 168 genes (69.7%) represent DR and 73 genes (30.3%) were



found to be NDR (Table 2.5). Similarly, 130 genes (54%) had extrinsic evidence support for which data for percent evidence support was recorded thus determined to be sequence homology based prediction. The remaining 46% genes didn't have any evidence support and hence represent *ab initio* prediction (see Table 2.6).

### 2.3.1.5.1 Genomic distribution of the novel genes

Determination of genomic distribution of novel gene loci along sorghum genome is important to facilitate exploration of their possible role in trait specific function. UniGene EST mapping in the plant genomes should be a powerful tool for association studies because they allow locus assignment based on sequence homology mapping. Here we demonstrated the chromosomal distribution of the novel gene model based on the drought responsive function. The highest number of genes (32, 13%) reside on chromosome 1 of which 53% are drought responsive. Because the total figure of complete gene structure is only 6%, their distribution along the genome is not more than 2% in each chromosome. Table 2.7 describes structure based distribution of novel gene that are functionally related to the drought response.

**Table 2.7 Chromosomal distribution of the novel gene model**

Scaffolds	Complete gene structure		Partial gene structure				Total
	DR	ND-DR	One end retained		truncated at both ends		
			DR	ND-DR	DR	Non-DR	
chr1	3	0	1	0	14	14	32
chr2	4	0	2	0	10	8	24
chr3	1	0	2	0	17	4	23
chr4	1	0	4	0	11	9	25
chr5	0	0	2	0	10	7	19
chr6	0	0	0	0	14	5	19
chr7	1	0	5	0	10	8	25
chr8	2	0	0	0	16	7	25
chr9	0	0	2	0	18	2	22
chr10	2	0	2	0	12	9	25
super	0	0	0	0	1	1	2
Total	14	0	20	0	133	74	241

### 2.3.1.5.2 Alternative Splicing (AS): Intron retention and exon skipping

Alternative splicing (AS) is the major source of transcriptome and proteome variation (Kim *et al.*, 2007). The combination of various transcript splice junctions result in transcripts with splice events such as shuffled (alternate) exons, alternative 5' or 3' splicing sites, alternative donor and acceptor, retained introns, skipped exon and different transcript termini (Ner-Gaon *et al.*, 2004). Table 2.8 and

Table 2.9 show patterns of alternative splicing for intron retention and exon skipping respectively and Table 2.10 shows AS in the *S. bicolor* genome detected by PASA based on the predicted novel genes. While retained intron is responsible for 20% of the AS, skipped exon is for 7% of the total AS.

Table 2.8 Genomic distribution of spliced and retained intron based on PASA analysis

Scaf <sup>1</sup>	Intron							
	coordinate	Spliced			Retained			
		orie <sup>2</sup>	asm <sup>3</sup>	MEP <sup>4</sup>	coordinate	orie <sup>2</sup>	asm <sup>3</sup>	MEP <sup>4</sup>
chr1	1819870-1820018	+	64	1	1819870-1820018	+	65	1
chr1	2245082-2245182	-	83	5	2245082-2245182		84	2
chr1	2630315-2630410	-	101	1	2630315-2630410	-	102	2
chr1	8195826-8195903	-	306	1	8195826-8195903	-	307	1
chr1	47862385-4786251	+	638	6	47862385-4786251	+	639	6
chr1	56966890-56966985	-	779	2	56966890-56966985	-	780	1
chr1	58887490-58887568	-	819	3	58887490-58887568	-	820	1
chr1	58887601-58887781	-	819	1	58887601-58887781	-	820	1
chr1	61110177-61110310	-	896	9	61110177-61110310	-	897	2
chr1	67980410-67980506	+	1125	6	67980410-67980506	+	1126	2
chr2	69162027-69162097	-	2358	2	69162027-69162097	-	2357	2
chr3	62293574-62293685	-	3153	3	62293574-62293685	-	3154	1
chr4	65061928-65062017	+	4139	2	65061928-65062017	+	4140	1
chr6	45323794-45323870	-	4626	18	45323794-45323870	-	4627	1
chr6	60237462-60237612	+	4942	2	60237462-60237612	+	4943	1
chr7	8345323-8345537	+	5138	5	8345323-8345537	+	5139	4
chr8	6642051-6642134	+	5536	7	6642051-6642134	+	5537	1
chr8	6641856-6641938	+	5536	4	6641856-6641938	+	5537	1
chr8	44687799-44687915	-	5622	1	44687799-44687915	-	5621	1
chr9	2436568-2436638	-	5784	4	2436568-2436638	-	5785	1
chr9	52085543-52085653	-	6061	2	52085543-52085653	-	6060	4
chr9	46038980-46039077	+	5954	32	46038980-46039077	+	5955	15
chr9	46038375-46038506	+	5953,	178	46038375-46038506	+	5955	4
			5954					
chr9	54606686-54606830	+	6117	2	54606686-54606830	+	6118	2
chr9	57490227-57490310	-	6208	5	57490227-57490310	-	6209	1
chr10	58496008-58496090	-	1717	2	58496008-58496090	-	1718	1

Key to legend: <sup>1</sup> Chromosomes # 1-10; <sup>2</sup> Orientation of the strand; <sup>3</sup> Assembly # that the transcripts belong to and <sup>4</sup> Maximum evidence support

Table 2.9 Genomic distribution of skipped and retained exons based on PASA analysis

Scaf <sup>f</sup>	Exon							
	Skipped				Retained			
	coordinate	orie <sup>2</sup>	asm <sup>3</sup>	MEP <sup>4</sup>	coordinate	orie <sup>2</sup>	asm <sup>3</sup>	MEP <sup>4</sup>
chr2	15949877-15955756	+	1983	1	15951004-15951091, 15951181-15951253	+	1984	1
chr2	75790551-75791220	+	2540	1	75790984-75791097	+	2539	7
chr3	71528810-71529369	-	3403	2	71529086-71529156	-	3404	1
chr4	4717042-4717529	+	3609	1	4717411-4717488	+	3608	11
chr6	1517417-1519399	-	4521	1	1518211-1518383	-	4520	2
chr6	60298761-60299726	-	4947	1	60299069-60299251	-	4946	9
chr9	54972512-54974544	-	6127	2	54972660-54973133	-	6126	10
chr10	2956478-2957725	-	1340	1	2956662-2956707	-	1339	1
chr10	59988260-59988901	+	1779	3	59988341-59988410	+	1780	2

Key to legend: refer to Table 2.8

Table 2.10 PASA based identification of alternative splicing (AS) for the novel genes

Types of AS	AS event	Non-AS events
Alternative acceptor	42	-
Alternative donor	14	-
Alternative 5' site	-	-
Alternative 3' site	-	-
Alternative exon	16	-
Ends in intron	10	-
Retained exon	-	9
Retained Intron	27	-
Skipped exon	9	-
Spliced intron	-	27
Starts in intron	18	-
Total	136	36

The maximum splice junction was created by alternate acceptor with 31% and alternate donor with 10% splice events respectively. An alternate exon mainly related to an increase in coding diversity (Boyd *et al.*, 1993) contributed to 12% splice events, whereas ends and starts in the intron resulted in 7 and 13% splice events respectively (see Table 2.10).

### 2.3.1.5.3 Distribution of Exon and Intron structure

A total of 512 exons were identified and characterized, thus the type of exons vary with number, length and position. The number varies from 1-9 and the length from 3 (initial exon) to 3614 (terminal exon) bases with the position known as initial, inner and terminal. The trend in average size of exon length decreases with increase in their number per gene (Figure 2.4; Table 2.11).

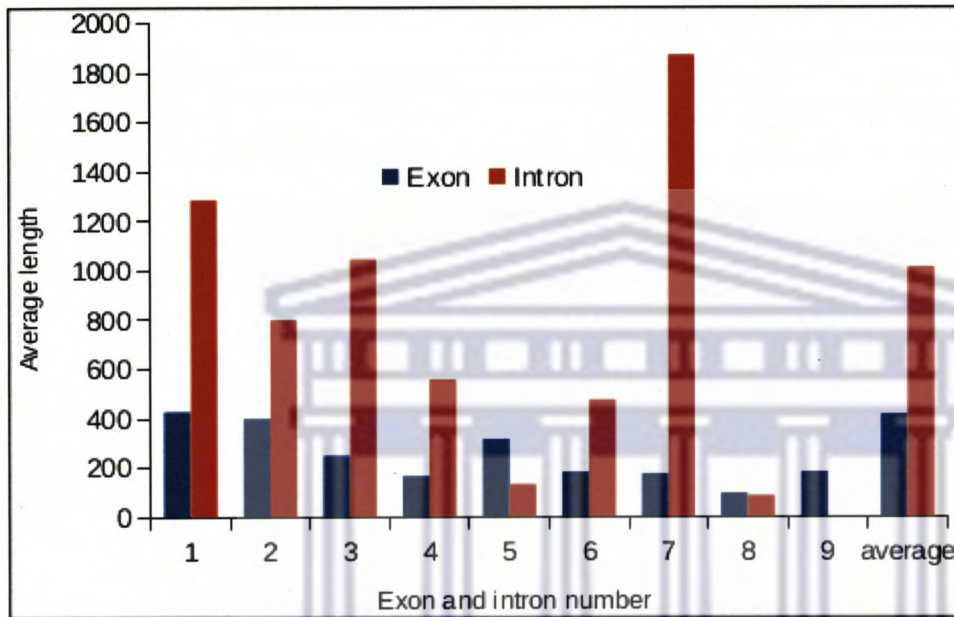


Figure 2.4: Pattern of exon and intron number and the average length

Key to legend:- 1=exon1, intron1; 2=exon2, intron2; 3=exon3, intron3; 4=exon4, intron4; 5=exon5, intron5; 6=exon6, intron6; 7=exon7, intron7; 8=exon8, intron8; 9=exon9, intron9 and average = exon total average; intron total average.

A total number of introns identified for 241 novel genes were found to be 275 and were varying in distribution from two to nine exonic genes. One hundred and fifteen single exonic genes observed in this study were devoid of any intron implicating the presence of intronless genes. While the number of introns and the length vary from 0 to 8 and from 69 to 21019 bases respectively, their location is restricted to inner position.

**Table 2.11 Exons and introns distributions for the novel genes in the sorghum genome**

Scaffo lds	Total features per scaffold		Exons per gene					Length (bp) of features											Chromosome size
								shortest		longest			total		average		Standard deviation		
			G	E	I	M	A	E	I	G	E	I	G	E	I	E	I	E	
Chr1	32	79	47	9	2.5	13	74	209	2065	3866	8778	27793	32497	351.8	691.2	369.5	960.6	2164-67845075	
Chr2	24	53	29	7	2.2	6	69	254	1910	4440	5290	17327	11248	558.1	385.9	1183.8	848.2	135092-64307455	
Chr3	23	40	17	4	2.4	3	89	236	2297	3392	3689	18730	14207	468.3	835.7	591.5	931.7	25612-73118483	
Chr4	25	63	38	9	2.5	7	66	224	2732	10968	13188	2130	37302	338.3	981.6	455	2089	10238-67290539	
Chr5	19	44	25	5	2.3	19	71	245	2018	6695	8795	12688	25038	288.4	1001.5	393.7	1711.3	43486-33757957	
Chr6	19	38	19	5	2	31	75	236	3518	3899	6502	19387	14540	510.2	765.5	765.3	903.7	11607-52987788	
Chr7	25	50	25	4	2	6	79	224	3156	3332	7905	21258	16599	425.2	664	525.1	930.2	41803-56863519	
Chr8	25	45	20	5	1.8	6	71	218	2858	2680	4372	25730	15539	571.8	777	695.1	794.6	10323-34152034	
Chr9	22	47	25	6	2.1	44	69	233	3614	15338	16668	25195	27435	536.1	1097.4	813.6	3066.6	17759-55481927	
Chr10	25	51	26	5	2.04	20	72	212	1855	6768	13475	21609	23136	423.7	889.9	420	1485.5	13087-53827720	
Super	2	2	0	2	1	739	0	739	887	0	889	5695	0	798.2	0	1626	0	4-8720612	
Ava	11	47	27	6	2.1	81	67	275	2446	5580	8141	17958	19777	479.1	735.4	712.6	1247.4	28289-51668465	

Key to legend: G=Genes; E = Exons; I = Introns; M = Maximum; A = Average

Distribution of exons and intron for novel genes throughout sorghum genome is shown in (Table 2.11). This data depicts that the least # of genes (2, 1%) were predicted from super scaffold probably owing to its relative smaller size and lower gene density (Paterson *et al.*, 2009) and that highest prediction was from chromosome 1 with 32 genes showing its biggest size.

#### 2.3.1.5.4 Intronless (single exonic) genes

A recently known prokaryotic characteristics of certain eukaryotic genes are thought to play role in our understanding of the evolutionary patterns of related genes and complex genomes. One such prokaryotic nature of the eukaryotic cells is the existence of intronless genes in their genomes as reported over the past few decades (Tine *et al.*, 2011; Zou *et al.*, 2011). For example, 901 predicted human genes including G protein-coupled receptor genes (Gentles and Karlin, 1999), single-copy primate-specific human single exonic genes (Tay *et al.*, 2009) and that of species-specific intronless enriched genes in Arabidopsis, Oryza, and Populus (Yang *et al.*, 2009) are few among others. Thus, species specificity is one of the distinctive feature of intronless genes. In this study, 2 single exonic intronless genes were found to be complete gene structure, 6 were partially complete in which five prime and three prime UTRs were retained for one and 5 genes respectively. The remaining 107 intronless genes were truncated completely (Table 2.12). Since there is some correlation between intron loss and processed pseudo gene (Zhu and Niu, 2013) and truncation as a feature in common between the events (Terai *et al.*, 2010; Arisue *et al.*, 2011), we speculate some of the identified single exonic genes to be pseudogene.

Table 2.12 Patterns of exonic and intronic features in the novel gene models

Number of features per gene	Feature-less	Single	Double	Triple	Quadruple	Multiple features					Total
						5	6	7	8	9	
Total Exonic	0	241	126	75	35	18	8	5	2	2	512
Total intronic	0	126	75	35	18	8	5	2	2	0	271
Gene per exon <sup>a</sup>	0	115 <sup>b</sup>	51	40	17	10	3	3	0	2	241
Gene per intron <sup>c</sup>	115 <sup>d</sup>	51	40	17	10	3	3	0	2	0	241

Key to legend: <sup>a</sup>Gene number per exonic feature; <sup>b</sup>Single exonic genes; <sup>c</sup>Gene number per intronic feature and <sup>d</sup>Intronless genes.

Although intron loss in evolution has been described, the mechanism involved is still unclear. Three models have been proposed (Rogozin *et al.*, 2003) : the reverse transcriptase (RT) model, genomic deletion model and double-strand-break repair model. The RT model, also termed mRNA-mediated intron loss, suggests that cDNA molecules reverse transcribed from spliced mRNA recombines with genomic DNA causing intron loss. Many studies have attempted to test this model based on its predictions, such as simultaneous loss of adjacent introns, 3'-side bias of intron loss, and germline expression of intron-lost genes (Roy and Gilbert, 2006). Evidence either supporting or opposing the model has been reported. The mechanism of intron loss proposed in the RT model shares the process of reverse transcription with the formation of processed pseudogenes. If the RT model is correct, genes that have produced more processed pseudogenes are more likely to undergo intron loss.

#### 2.3.1.5.5 Prediction of pseudogenes

Based on what Ensembl already stipulated as characteristic features of pseudogenes (Curwen *et al.*, 2004; Flicek *et al.*, 2013 and 2014) and then applied by Goodstadt and Ponting, 2006, the finding of single exonic intronless genes triggers our suspect for the presence of pseudogenes in this result. However, this requires further investigation to substantiate the finding.

#### 2.3.1.5.6 Identification of nearest intergenic distances

Figure 2.5 illustrates the intergenic distances between the novel and the nearest neighbouring existing gene structure model. Identification of intergenic distance is not only important in estimating the extent of neighbourhood of the genes but also to speculate the proximity of functionally important regulatory domains. However, the large span of this region poses complexity in scrutinizing the extent its importance. Because this region is less divergent in sequence structure than the rest part of the genome, it represents a

more likely conserved region between species (Thornton *et al.*, 2002) thus plays role in phylogenetic analysis and gene regulatory activities.

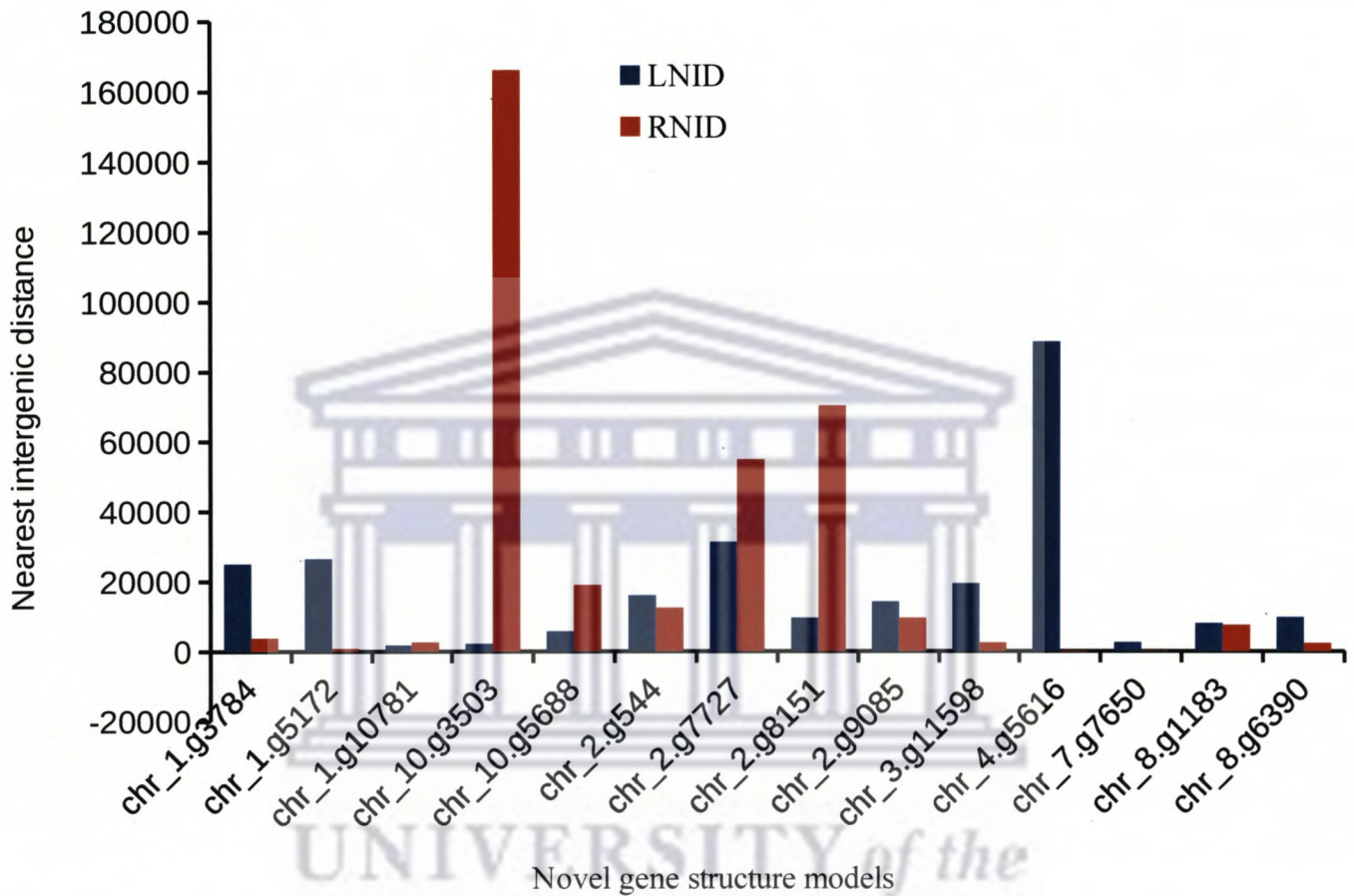


Figure 2.5: Intergenic distances between the novel and nearest existing gene structure model.

**Legend:** LNID represents left nearest neighbouring intergenic distance; RNID denotes right nearest neighbouring intergenic distance. The intergenic distances between the novel gene structure models and the nearest existing neighbouring genes were described based on the fourteen novel genes structure models known to have the complete gene structure. The right (red bar) nearest intergenic distance between the novel gene model and the existing gene represents the longest distance in most cases (eg. chr\_10.g3503) than the left (blue) nearest intergenic distance.

### 2.3.2 Metabolic pathways analysis

We identified twelve known metabolic pathways which displayed strong correlation with sorghum drought tolerance. Two other pathways namely cholinesterase (EC:3.1.1.1) and an adenylypyrophosphatase (EC: 3.6.1.3) that catalyse drug metabolism-other enzymes and purine metabolism respectively were identified for which we did not find any known gene responsible for encoding the enzyme as yet for sorghum (Table 2.14). We arbitrarily picked five metabolic

pathways to discuss the result in detail (Figure S2.3). However, Table 2.13, gives description of all the pathways and the genes involved. Figure S2.3 shows all the KEGG pathways identified except oxidative phosphorylation (Figure 2.6) which is shown in the body.

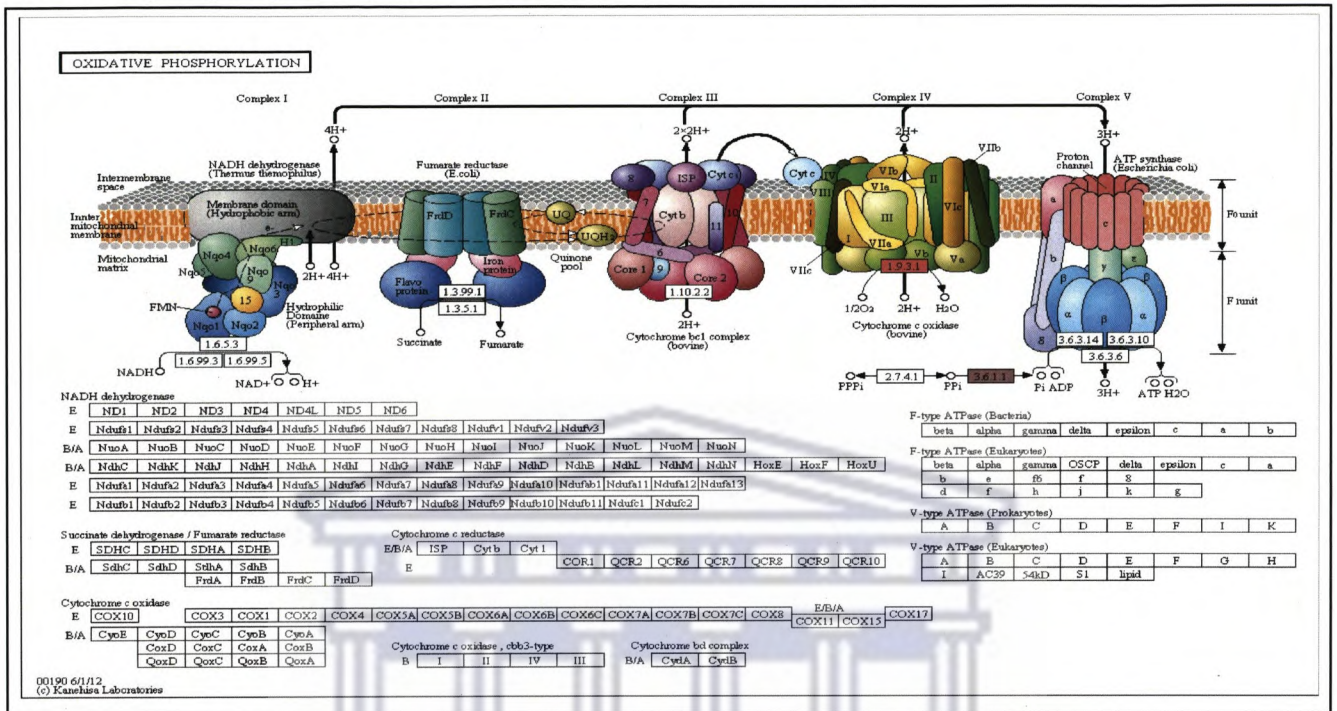


Figure 2.6: Oxidative phosphorylation metabolic pathway.

This pathway is one of the 14 metabolic pathways identified in this analysis and is a pathway for the production of respiratory energy (Atkin and Macherel, 2009) in mitochondria, power house of the cell. Cytochrome c oxidase subunit 1 (EC: 1.9.3.1; Table 2.13), the enzyme encoded by sorghum gene *cox1* was identified to be involved in the catalytic reaction of the final step of protein complex (complex IV) in the electron transport chain (Calhoun *et al.*, 1994). Inorganic diphosphatase (EC: 3.6.1.1; Table 2.13) is the another enzyme identified to be involved in the electron transport system by catalyzing the conversion of diphosphate into monophosphate thus control amount of inorganic phosphate (Pi) that should be coupled with adenosine dinucleotide phosphate (ADP) in the last step of oxidative phosphorylation (Affourtit *et al.*, 2012).

Glucosinolate biosynthesis catalysed by CYP79A1 [EC:1.14.13.41] is identified with gene CYP79A1 to be involved in synthesizing Cyanogenic Glucosides (CGs), a secondary metabolites in most plants. The three pathways which share certain characteristics in common are Pantothenate biosynthesis (EC: 2.6.1.42), Valine, Leucine, and Isoleucine biosynthesis (VLIB) (EC 2.6.1.42), Valine, Leucine, and Isoleucine degradation metabolic pathways (VLID) (EC 2.6.1.42) (Figure 2.6). These pathways are responsible for the amination of 4-methyl-2-oxopentanoate attributed to the presence of a branched-chain amino acid transaminase (BCAT) (EC 2.6.1.42) activity. Three genes namely Sb04g010240, Sb06g025140 and Sb09g008180 are responsible commonly for the biochemical reaction in association with the amination of 4-methyl-2-oxopentanoate entailing their



shared peculiarity and networking across pathways.

**Table 2.13 Functional description of sorghum drought related metabolic pathways**

Pathway	Enzyme	Enzyme ID	Pathway map ID	Pathway ID	gene (id) involved
Aminoacyl-tRNA biosynthesis	ligase	EC:6.1.1.16	map00970	KO:K01883	*genes
Cysteine and methionine metabolism	Dioxygenase (iron(II)-requiring)	EC:1.13.11.54	map00270	KO:K08967	Sb01g046360
Drug metabolism - other enzymes	ali-esterase	EC:3.1.1.1	map00983	*KO	novel
Glucosinolate biosynthesis	CYP79A1, tyrosine N-mono-oxygenase	EC:1.14.13.41	Map00966 (sbi00966)	KO:K13027	Sb01g001200
Glycerophospholipid metabolism	kinase (ATP dependent)	EC:2.7.1.107	map00564	KO:K00901	Sb01g032250
Glycerolipid metabolism	kinase (ATP dependent)	EC:2.7.1.107	map00561	KO:K00901	Sb01g032250
Phosphatidylinositol signalling system	kinase (ATP dependent)	EC:2.7.1.107	map04070	KO:K00901	Sb01g032250
Nicotinate and nicotinamide metabolism	kinase	EC:2.7.1.23	map00760	KO:K00858	Sb09g019130
Oxidative phosphorylation	Cytochrome c oxidase subunit 1 (Oxidase); Inorganic diphosphatase	EC:1.9.3.1; EC:3.6.1.1	Map00190 map00190	KO:K02256 KO:K02256	Sb09g022400, COX1; **genes
Pyrimidine metabolism	RNA polymerase	EC:2.7.7.6	map00240	**KO	***genes
Purine metabolism	RNA polymerase; adenylypyrophosphatase	EC:2.7.7.6; EC:3.6.1.3	Map00230 map00230	**KO KO:K01509	***genes novel
Pantothenate biosynthesis	BCAT	EC:2.6.1.42	sbi00770	KO:K00826	****genes
VLIB	BCAT	EC:2.6.1.42	sbi00290	KO:K00826	****genes
VLID	BCAT	EC:2.6.1.42	sbi00280	KO:K00826	****genes

Key to legend:  
 \*KO K01044 carboxylesterase 1; K03927 carboxylesterase 2; K03928 carboxylesterase; K15743 carboxylesterase 3/5  
 \*\*KO KO:K03006; KO:K02999; KO:K03002; KO:K03006; KO:K03018; KO:K03021; KO:K03043; KO:K03046  
 \*genes Sb01g047380; Sb02g032450  
 \*\*genes Sb09g001530; Sb10g009880; Sb09g004450; Sb09g021610; Sb01g022340; Sb04g036230; Sb04g005710; Sb04g034340; Sb03g013530; Sb03g040910  
 \*\*\* genes Sb05g019520; Sb03g017630; Sb03g020184; Sb04g001790; Sb04g009491; Sb05g019520; Sb06g021120; Sb07g003680; Sb09g027223; Sb09g027230; Sb10g006995  
 \*\*\*\*genes Sb04g010240; Sb06g025140; and Sb09g008180

In this study 28 genes were identified to be involved in the three pathways among which the three genes mentioned are responsible for transamination (conversion of 4-methyl-2-oxopentanoate). Oxidative phosphorylation pathway is involved in the production of energy by maintaining mitochondrial respiration at times of water-stress conditions (Atkin and Macherel, 2009). Two genes, COX1 and Sb09g022400, responsible for the process of electron transport in oxidative phosphorylation were identified to encode for cytochrome c oxidase 1 and diphosphatase respectively (Figure 2.6). A detailed description of the fourteen metabolic pathways is given in Table 2.14 where 32 functionally enriched genes are indicated.

### 2.3.2.1 Functional GO-enrichment analysis of the pathway

A total of 477 sorghum genes in all the pathways were identified to which 583 significantly enriched GO-terms were assigned (P-value and FDR < 0.01). However, analysis revealed 32 genes responsible to encode protein enzyme that catalyse substrate conversions in the respective pathways (Table 2.13).

Table 2.14 Functional GO enrichment of the genes involved in the metabolic pathways

Gene identifier	GO-Term	Attribute <sup>a</sup>	test gene set frequency	Background set frequency	P-value	FDR
GO:0009081	branched chain family amino acid metabolic process	P	22/56 (39.3%)	54/26245 (0.2%)	9.50E-043	8.80E-041
GO:0009108	coenzyme biosynthetic process	P	20/56 (35.7%)	145/26245 (0.6%)	1.80E-030	1.60E-028
GO:0006732	coenzyme metabolic process	P	23/56 (41.1%)	316/26245 (1.2%)	3.50E-029	3.20E-027
GO:0006752	group transfer coenzyme metabolic process	P	16/56 (28.6%)	67/26245 (0.3%)	7.80E-028	7.20E-026
GO:0051186	cofactor metabolic process	P	24/56 (42.9%)	452/26245 (1.7%)	2.20E-027	2.00E-025
GO:0043436	oxoacid metabolic process	P	32/56 (57.1%)	1301/26245 (5%)	3.40E-027	3.10E-025
GO:0019752	carboxylic acid metabolic process	P	32/56 (57.1%)	1301/26245 (5%)	3.40E-027	3.10E-025
GO:0006082	organic acid metabolic process	P	32/56 (57.1%)	1302/26245 (5%)	3.40E-027	3.20E-025
GO:0034641	cellular nitrogen compound metabolic process	P	27/56 (48.2%)	718/26245 (2.7%)	3.50E-027	3.20E-025
GO:0042180	cellular ketone metabolic process	P	32/56 (57.1%)	1318/26245 (5%)	5.00E-027	4.60E-025
GO:0006790	response to osmotic stress	P	7/56 (12.5%)	631/26245 (2.4%)	0.0004	0.036
GO:0009651	catalytic activity	F	56/56 (100%)	13636/26245 (52%)	1.30E-016	5.00E-015
GO:0006725	cofactor binding	F	16/56 (28.6%)	792/26245 (3%)	7.20E-012	2.90E-010
GO:0006970	coenzyme binding	F	14/56 (25%)	589/26245 (2.2%)	2.30E-011	9.10E-010
GO:0003824	lyase activity	F	13/56 (23.2%)	570/26245 (2.2%)	2.20E-010	8.60E-009
GO:0048037	3-chloroallyl aldehyde dehydrogenase activity	F	5/56 (8.9%)	24/26245 (0.1%)	4.20E-009	1.70E-007
GO:0050662	magnesium ion binding	F	9/56 (16.1%)	316/26245 (1.2%)	2.70E-008	1.10E-006
GO:0016829	oxidoreductase activity	F	18/56 (32.1%)	2349/26245 (9%)	1.10E-006	4.20E-005
GO:0004028	carboxylic acid binding	F	6/56 (10.7%)	166/26245 (0.6%)	1.80E-006	7.10E-005
GO:0000287	oxidoreductase activity...	F	7/56 (12.5%)	271/26245 (1%)	2.00E-006	8.10E-005
GO:0016491	carbon-carbon lyase activity	F	6/56 (10.7%)	179/26245 (0.7%)	2.70E-006	0.00011
GO:0016835	cytoplasm	C	51/56 (91.1%)	9051/26245 (34.5%)	1.40E-018	3.50E-017
GO:0016740	cytoplasmic part	C	45/56 (80.4%)	7660/26245 (29.2%)	3.40E-015	8.60E-014
GO:0016746	mitochondrion	C	22/56 (39.3%)	1853/26245 (7.1%)	1.00E-011	2.60E-010
GO:0005737	intracellular part	C	51/56 (91.1%)	12750/26245 (48.6%)	1.60E-011	4.00E-010
GO:0044444	intracellular	C	51/56 (91.1%)	13212/26245 (50.3%)	8.30E-011	2.10E-009
GO:0005739	cytosol	C	17/56 (30.4%)	1740/26245 (6.6%)	7.80E-008	1.90E-006
GO:0044424	mitochondrial lumen	C	7/56 (12.5%)	166/26245 (0.6%)	8.30E-008	2.10E-006
GO:0005622	mitochondrial matrix	C	7/56 (12.5%)	166/26245 (0.6%)	8.30E-008	2.10E-006
GO:0005829	mitochondrial part	C	9/56 (16.1%)	546/26245 (2.1%)	2.40E-006	6.10E-005
GO:0031980	plastid	C	16/56 (28.6%)	2109/26245 (8%)	5.70E-006	0.00014

<sup>a</sup>: P= Biologicl Process, F = Molecular Function and C = Cellular Components. This description includes only the first ten top enriched GO-terms in decreasing order of their p-values for all the GO-domains and with the GO-terms indicated at the top of each category of domain representing the highest enriched.

### 2.3.2.2 Pattern of sequence distribution and GO annotation

Sequence distribution based on blast hits which were associated to the GO-terms for the biological process is shown in Figure S2.2a. This distribution revealed that the highest proportion of sequences were mapped to biological, metabolic and cellular process with the order of 50, 38.2 and 34.2 percent. A fairly high proportion in sequence matching the GO-terms for response to stimuli and to

stress was demonstrated with 20.3 and 18.7% respectively. This results may suggest that the length of the blast pairwise matching alignment regions that are associated with the particular GO-term and the percent identity with the matching sequence are indicative to shared function (Lomax, 2005). Furthermore, there seems a positive correlation between the presence of sequence matching the region associated with the GO-terms and expression of genes in response to a given tissue-specific biological process a GO-term stands for. On the other hand, mapping result for species specific blast hit shows highest score for maize, whereas highest blast top hit for sorghum (Figure S2.2b and c). The GO-level distribution of annotation (Figure S2.2d) shows that biological process takes 49.7% share, cellular component 30.5% and molecular function 25.3%.

### 2.3.2.3 Interpro Domain Analysis

Protein domains represent 33% of the main categories of interpro domains identified. Interpro-domains with known signature, represent 381 (60.5%) of a 630 total figure (Figure S2.4).

Table 2.15 Description of the top ten interpro domains in decreasing order of frequency

Ser. <sup>1</sup>	Interpro domain	Accession <sup>2</sup>	F, P <sup>3</sup>	Functional description	References
1	DnaJ domain	IPR001623	22, 6.4	Act as protein chaperon; cooperation of Hsp40 with Hsp70 and endosomal trafficking <sup>4</sup>	<sup>10, 11</sup>
2	Gamma thionin	IPR008176	18, 5.3	plant defensins induced in response to drought	<sup>12</sup>
3	Ribosomal protein L29e	IPR002673	17, 5	forms part of the 60S ribosomal subunit, structural constituent of ribosome <sup>5</sup>	<sup>13</sup>
4	Zinc finger, CCHC-type	IPR001878	17, 5	Drought stress response in plants	<sup>14</sup>
5	DUF4281 <sup>6</sup>	IPR025461	16, 4.7	Protein domain functionally uncharacterized, found both in prokaryotes and eukaryotes	<sup>15</sup>
6	RNA recognition <sup>7</sup>	IPR000504	16, 4.7	Expression of EgRBP42 transcript under drought stress	<sup>16, 17</sup>
7	Cytochrome c oxidase, subunit VIIa	IPR003177	14, 4.1	catalyzes the reduction of oxygen to water in the inner mitochondrial membrane forming the functional core of the enzyme complex <sup>8</sup>	<sup>18</sup>
8	oligopeptide transporter family <sup>9</sup>	IPR000109	14, 4.1	showing an enhanced response in 35S:ABF3 plants that may contributing to drought-tolerance	<sup>19</sup>
9	Peptidase S10, serine carboxypeptidase	IPR018202	13, 3.8	protein recognition and binding, serine carboxypeptidase-like gene OsBISCPL1 in rice is involved in regulation of defense responses against biotic and oxidative stress	<sup>20</sup>
10	CBS domain	IPR000644	12, 3.5	transcript levels of CBS domain containing proteins are altered in response to drought	<sup>21</sup>

Key to legend:

<sup>1</sup> Serial number; <sup>2</sup> FinterPro Accession; <sup>3</sup> Frequency of occurrence, %; <sup>4</sup> Intracellular; <sup>5</sup> involve in translation and ribosome biogenesis; <sup>6</sup> Protein length range between 147 and 232 amino acids with known two functionally important conserved residues (W and P); <sup>7</sup> motif domain; <sup>8</sup> transferring the electrons from cytochrome c via its binuclear copper A center to the bimetallic center of the catalytic subunit 1; <sup>9</sup> Proton-dependent; <sup>10</sup> Greene *et al.*, 1998; <sup>11</sup> Girard *et al.*, 2005; <sup>12</sup> Lay and Anderson, 2005; <sup>13</sup> Kuwano *et al.*, 1991; <sup>14</sup> Vij and Tyagi, 2008; <sup>15</sup> Marchler-Bauer *et al.*, 2013; <sup>16</sup> Zhou *et al.*, 2014; <sup>17</sup> Yeap *et al.*, 2012; <sup>18</sup> Wang and Vanlerberghe, 2013; <sup>19</sup> Abdeen *et al.*, 2010; <sup>20</sup> Liu *et al.*, 2008; <sup>21</sup> Kushwaha *et al.*, 2009.

Interpro-domain analysis clearly shows that the frequency of protein domains in the sequences varies greatly. Table 2.15 shows the description of the ten top interpro-domains in decreasing order of their occurrence.

### 2.3.3 Analysis of gene-expression profiling

Pattern of gene expression was analysed using both parametric (unpaired t-Test,  $p < 0.01$ ) and non-parametric tests (rank product,  $P < 0.01$  and Fisher's exact test,  $p < 0.05$ ). Using unpaired parametric t-Test, 49 and 879 statistically significantly expressed gene were identified based on treatment (drought stress and well-watered) and tissue (fertilized ovary and basal leaf meristem) based grouping respectively. On the other hand, with the rank product using treatment based grouping, 75 and 34 up-regulated and down-regulated genes were respectively identified when treated under drought condition. Using tissue based grouping, 52 up-regulated and 41 down-regulated genes were respectively identified under same condition. Similarly, based on the Fishers's exact test, 55 treatment based (25 up-regulated and 30 down-regulated) and 824 tissue based grouping (226 up-regulated and 598 down-regulated) genes were identified (Figures 2.7 and 2.8). Figure 2.9 shows the distribution of significantly expressed genes under drought stress using venn diagram (Oliveros, 2007) based on different statistical models.

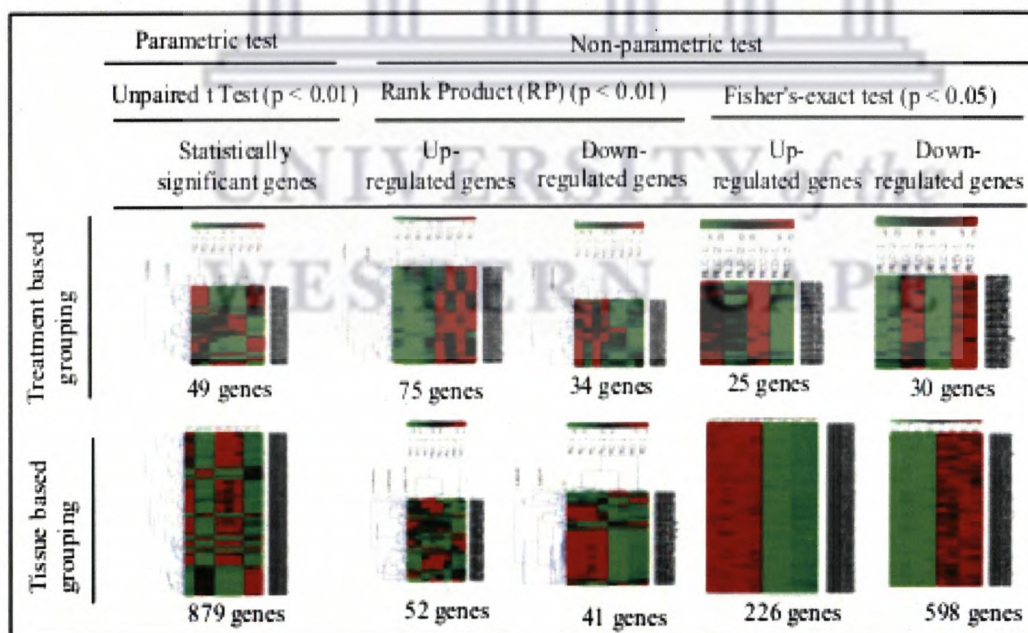


Figure 2.7: Heat map showing up and down-regulated sorghum orthologs based on maize RNA-seq expression data. The comparison based on parametric (unpaired t-Test or between subject comparison,  $p < 0.01$ ) and non-parametric test (Rank Product, RP,  $p < 0.01$ ) and Fisher's Exact test ( $p < 0.05$ ) have shown the up and down-regulated genes across treatment and tissue based grouping. Evaluation by treatment based grouping is determined to see significance difference in gene expression due to effect of differential condition under which the samples were tested while tissue based grouping is tempted to detect the effect of differences in tissues on the gene expression indicating the type of

tissues contributed for more significant expression. All data showing significant expression, either up or down regulation of genes in both groupings represent results obtained under drought conditions for ovary and leaf meristem tissues.

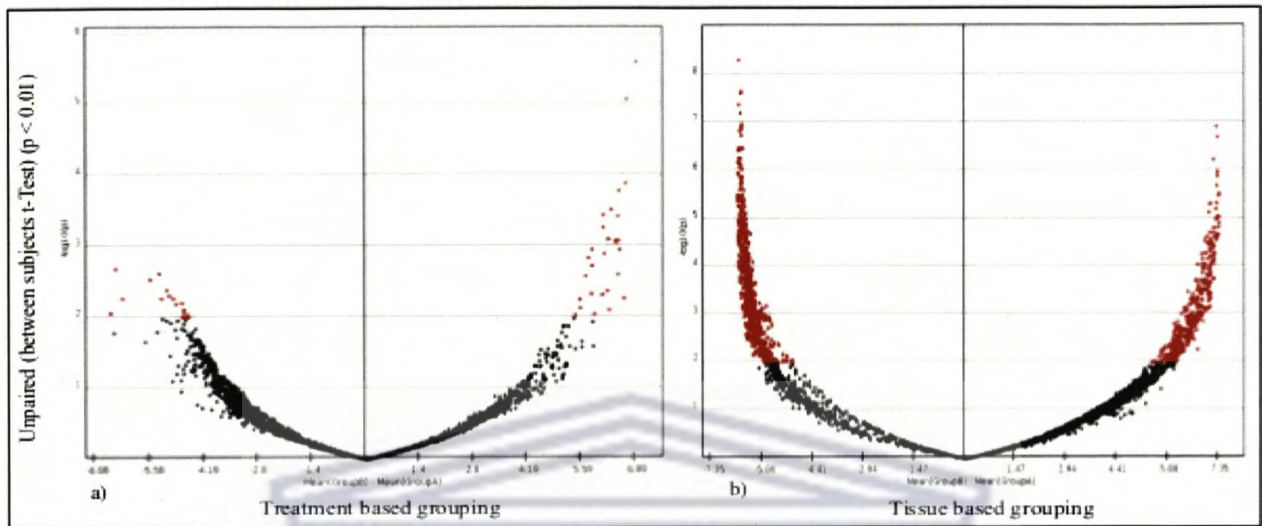


Figure 2.8: Volcano plot showing differential expression of genes

This figure describes volcano plot showing differential expression of genes with the most significant genes at the top of the plot. The red dots indicate genes-of-interest that display both large-magnitude of fold-changes (the change in mean values of the A group and B group, in our case, x-axis) and a fairly high value of statistical significance ( $-\log_{10}$  of p-value, y-axis). The upper line across the plot shows where  $p = 0.01$  (i.e. where the fold-change is equal to two ( $\log_2 = 1$ ) above which lie all genes having  $p < 0.01$  and below which having  $p > 0.01$ ). Volcano plots a) represent unpaired t-test based on treatment grouping, b) depicts the gene expression pattern based on between subject variance with tissue based grouping of samples. This plot shows higher number of genes expressed under drought condition in tissue based grouping (Figure 2.8b) with more down-regulated genes of a specific tissue than in treatment based grouping (Figure 2.8a) which shows not only relatively fewer number of genes expressed in total under same stress condition but also relatively less down-regulated genes.

### 2.3.3.1 Functional GO-enrichment analysis of gene-expression

The combined result of all the statistical tests gave 1079 significant non-redundant genes with 45 significantly expressed genes supported by all models (Table 2.16). We used the 45 maize genes to query sorghum orthologs using ENSEMBL BIOMART and retrieved 41 high level identity ( $> 90\%$ ) and high confidence (one) for further analysis using functional ontology (Figure S2.6) where we obtained 32 non-redundant genes to which enriched GO-terms were assigned (Table 2.16). This denoting that sorghum genes identified from maize, closest relative species, show conserved functional similarity in drought stress response notably in activities related to reproduction, photosynthetic cellular metabolic process and ion and chlorophyll binding typically involving both photosyntheses I and II. Figure S2.5 shows the patterns of functional GO-terms assignment of

sorghum genes identified from their maize ortholog based on gene expression profiling in reproductive and leaf meristem tissue under drought condition as revealed from RNA-seq data (Kakumanu *et al.*, 2012).

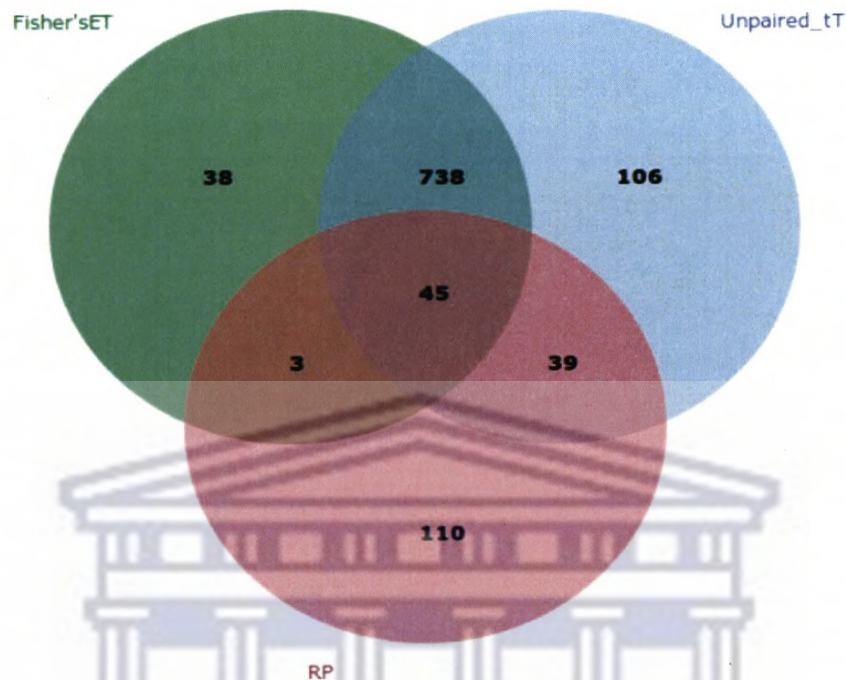


Figure 2.9: Venn diagram showing distribution of significantly expressed genes

Key to legend: Fisher'sET = Fishers Exact Test; Unpaired\_tT = Unpaired t test and RP = Rank product. The figure illustrates the number of statistically significantly expressed genes identified by each statistical model which is equivalent to the sum of values in each respective circle. The overlapping shows the number of significantly expressed genes supported by multiple models. The number of expressed genes supported by all statistical models is represented by 45.

Based on the GO enrichment analysis, cellular and metabolic processes including responses to stimulus represent the major task of significantly enriched genes involved in biological process. On the other hand, organelle, cell and cell parts serve as the integral component for the majority of genes assigned to cellular component where ion binding is a sole activities representing the main function of the majority of genes with limited genes taking part transcriptional regulatory and structural molecular activities (Figure S2.6). From the GO enrichment analysis, it is possible to realise that our result gives information related to a response of ovary tissue to drought stress where it was more likely affected as also shown in Figure 2.8b. Expression profiling result showed 12 genes significantly expressed under drought condition in agreement with the published result (Kakumanu *et al.*, 2012).

Table 2.16 Description of the 45 significantly expressed maize genes under drought condition and the corresponding sorghum orthologs

Genes name		Unpaired t-Test (p < 0.01)				Rank Product (p < 0.01)			Fisher's Exact-Test (p < 0.05)	
Maize gene	Sorghum ortholog	Raw p value	Adj p value	FDR	p-Values (Down)	RP-Values (Down)	p-Values (Up)	RP-Values (Up)	Right Tail p-value	2-Tail p-value
AC205122.4_FG003	N/A	0.0026207152	0.0026207152	0.009536492	0.6074809	1066.7379	0.0070398645	88.33089	0.014285714	0.028571429
AC207722.2_FG009	Sb02g032040	3.23E-007	3.23E-007	4.77E-005	0.99507207	1995.2388	5.51E-005	36.512115	0.014285714	0.028571429
AC216353.2_FG005	Sb04g034340	9.36E-006	9.36E-006	2.69E-004	0.52217984	966.8189	0.0042196778	150.68915	0.014285714	0.028571429
GRMZM2G001653	Sb06g016090	3.90E-005	3.90E-005	5.32E-004	0.97915143	1570.506	0.009499576	193.44215	0.014285714	0.028571429
GRMZM2G013342	Sb01g004330	6.64E-005	6.64E-005	7.49E-004	0.9913062	1936.3949	0.005182358	187.29814	0.014285714	0.028571429
GRMZM2G015419	Sb04g006480	0.0022894286	0.0022894286	0.008569004	0.17839696	565.9524	9.71E-004	97.553955	0.014285714	0.028571429
GRMZM2G016066	Sb02g010190	1.41E-005	1.41E-005	3.21E-004	0.98470736	1866.7698	0.008630196	119.37978	0.014285714	0.028571429
GRMZM2G017290	Sb01g006370	3.99E-004	3.99E-004	0.0024433285	0.9976845	2061.1926	0.0018744699	168.86763	0.014285714	0.028571429
GRMZM2G018627	Sb01g015400	1.74E-006	1.74E-006	1.11E-004	0.9330365	1616.922	0.003914334	147.49261	0.014285714	0.028571429
GRMZM2G022958	N/A	1.49E-006	1.49E-006	9.77E-005	0.98819333	1899.9612	1.74E-004	56.964867	0.014285714	0.028571429
GRMZM2G026015	Sb03g004560	6.18E-005	6.18E-005	7.14E-004	0.97915184	1823.742	0.0043977946	152.34447	0.014285714	0.028571429
GRMZM2G033885	Sb02g036260	3.95E-006	3.95E-006	1.66E-004	0.9952757	1998.7881	0.0014334182	109.00573	0.014285714	0.028571429
GRMZM2G038519	N/A	1.98E-005	1.98E-005	3.66E-004	0.9717091	1777.1985	0.0041221376	149.75601	0.014285714	0.028571429
GRMZM2G042118	Sb03g043760	0.0057001295	0.0057001295	0.017320754	0.0013782866	107.56193	0.038240034	68.723564	0.014285714	0.028571429
GRMZM2G046284	Sb08g004500	0.0040300074	0.0040300074	0.013198274	0.8609881	1441.6995	9.50E-004	96.456825	0.014285714	0.028571429
GRMZM2G052869	N/A	0.004447227	0.004447227	0.014209432	3.52E-004	70.26057	0.18936387	580.7853	0.014285714	0.028571429
GRMZM2G057075	N/A	0.003801683	0.003801683	0.012733478	2.33E-004	63.383923	0.99999577	866.5349	0.014285714	0.028571429
GRMZM2G062610	Sb01g003250	6.50E-004	6.50E-004	0.003436323	0.9992536	2137.4414	0.008859202	1123.0922	0.014285714	0.028571429
GRMZM2G063162	Sb04g009670	1.75E-004	1.75E-004	0.0014064441	0.9926675	1954.9122	0.0069296015	1633.9083	0.014285714	0.028571429
GRMZM2G071450	Sb10g000230	1.31E-005	1.31E-005	3.03E-004	0.987799	1895.428	0.00937659	192.54845	0.014285714	0.028571429
GRMZM2G072280	Sb02g037410	1.86E-004	1.86E-004	0.0014676332	0.9297752	1606.6776	0.0034266326	160.0729	0.014285714	0.028571429
GRMZM2G073934	Sb01g017010	3.39E-004	3.39E-004	0.0022259252	0.0027226463	132.78181	0.77147156	2291.8972	0.014285714	0.028571429
GRMZM2G080107	Sb08g005300	2.81E-005	2.81E-005	4.30E-004	0.9931468	1961.6893	0.0031128076	138.47742	0.014285714	0.028571429
GRMZM2G080603	Sb08g022740	0.0037057353	0.0037057353	0.012500893	0.046547923	331.15103	7.17E-004	88.76978	0.014285714	0.028571429
GRMZM2G085646	Sb01g006370	3.61E-004	3.61E-004	0.002296463	0.9425488	1649.0367	0.0060517387	170.18427	0.014285714	0.028571429
GRMZM2G092311	Sb04g004770	1.80E-005	1.80E-005	3.62E-004	0.94385076	1653.9498	0.0016497031	115.0895	0.014285714	0.028571429
GRMZM2G098520	Sb05g003480	5.39E-004	5.39E-004	0.003020165	0.96303225	1731.8862	0.0077650547	142.271	0.014285714	0.028571429
GRMZM2G099454	Sb01g048140	3.19E-004	3.19E-004	0.002132841	4.83E-004	77.36669	0.8880407	307.5208	0.014285714	0.028571429
GRMZM2G100754	Sb06g000820	6.78E-004	6.78E-004	0.0035212275	0.0015012722	110.53399	0.93812984	185.9457	0.014285714	0.028571429
GRMZM2G113033	Sb05g003480	2.28E-005	2.28E-005	3.90E-004	0.97665393	1807.3234	0.00613655667.17	175.34497	0.014285714	0.028571429
GRMZM2G122937	Sb06g023630	0.0093331365	0.0093331365	0.025354303	0.0012553012	105.054855	E-004	88.33089	0.014285714	0.028571429
GRMZM2G126772	Sb04g003110	0.0031622113	0.0031622113	0.011079486	0.00399067	148.20943	0.9902799	1498.0651	0.014285714	0.028571429
GRMZM2G130173	N/A	1.18E-004	1.18E-004	0.0010597043	4.28E-004	74.34362	0.6535199	134.9499	0.014285714	0.028571429
GRMZM2G153184	Sb07g021260	7.10E-005	7.10E-005	7.79E-004	0.9764376	1805.9309	0.003273961	140.4451	0.014285714	0.028571429
GRMZM2G155216	Sb09g028720	1.51E-004	1.51E-004	0.001263552	0.9841137	1861.1422	0.0010347753	100.13144	0.014285714	0.028571429
GRMZM2G160268	Sb07g021260	1.22E-005	1.22E-005	2.98E-004	0.99112386	1934.2758	0.0030831213	137.90283	0.014285714	0.028571429
GRMZM2G166944	N/A	7.37E-004	7.37E-004	0.0037214751	2.04E-004	60.823204	0.43264207	188.9694	0.014285714	0.028571429
GRMZM2G168651	N/A	7.26E-006	7.26E-006	2.31E-004	0.612799	1073.1229	0.0014249363	108.82252	0.014285714	0.028571429
GRMZM2G174984	Sb07g005660	9.58E-006	9.58E-006	2.66E-004	0.99849874	2093.0405	0.002832909	168.0419	0.014285714	0.028571429
GRMZM2G306345	Sb09g019930	6.67E-004	6.67E-004	0.0034861472	0.9872561	1889.6614	7.21E-004	89.26275	0.014285714	0.028571429
GRMZM2G351977	Sb03g027030	2.34E-005	2.34E-005	3.85E-004	0.9444444	1656.0074	0.0013994911	107.910805	0.014285714	0.028571429
GRMZM2G414192	Sb01g015400	1.75E-006	1.75E-006	1.09E-004	0.9331764	1617.3859	0.0038507208	146.75131	0.014285714	0.028571429
GRMZM2G447785	N/A	0.009865722	0.009865722	0.026495868	0.37666243	804.86896	3.44E-004	175.94577	0.014285714	0.028571429
GRMZM2G451224	Sb09g028260	1.99E-005	1.99E-005	3.64E-004	0.85424936	1428.4792	0.008418151	181.18259	0.014285714	0.028571429
GRMZM5G845611	Sb01g048470	7.07E-005	7.07E-005	7.83E-004	0.99077183	1930.4094	0.0062892283	580.7853	0.014285714	0.028571429

This table shows statistical description of the 45 maize genes significantly up-regulated under drought condition and the corresponding 35 sorghum orthologs. The statistical significance of these genes was supported by all statistical models used in the analysis.

### 2.3.4 Analysis of orthology groups

The value of identification of orthologous groups is not only noted for genome annotation, but is also spectacular in evolutionary findings of genes and gene products, comparative genomic studies, and the identification of candidate genes (Koonin *et al.*, 2004, Mitchell *et al.*, 2007).

Out of the 18,815 (6915 non redundant) initially identified orthologs from the three sorghum relative species, a total of 13,801 (6492 (93%) non redundant) were screened whose % identity is > 50 and confidence level is only 1 (see Table 2.17). This represents 42.7%, 37.6% and 19.7% contribution of orthologs from maize, rice and Arabidopsis respectively suggesting greater number of shared genes with proximity to ancestral species. Before performing ontology enrichment using the combination of all the orthologs recovered, we determined to see the extent of species representation and subtotal genes commonly identified by more than one species (section 2.3.4.1, Table 2.17 and Figure 2.10 (Venn-diagram). To this end, 2098 genes were identified to represent sorghum orthologs contributed in common by all species as indicated in the Venn-diagram (see Figure 2.10). Table 2.17 shows the patterns of sorghum orthologs with respect to the corresponding sorghum relative species based on 9693 sorghum UniGene clusters as an entry.

#### 2.3.4.1 GO enrichment analysis of genes through orthology groups

Considering only the 2098 genes into GO enrichment analysis, did not result significantly high gene enrichment suggesting the necessity to rather consider all the genes partly because non-common orthologs which potentially contribute to drought tolerance seem to remain unrepresented in GO enrichment and partly because the common genes represent only 30% of the initial total figure which doesn't seem to be representative.

Table 2.17 Sorghum orthologs and the corresponding genes from closely related species

Sorghum relative spp.	Genes <sup>b</sup>	Genes <sup>c</sup>	Homology type (%)			% Identity (> 50)	Confidence level			sorghum orthologs <sup>g</sup>
			A <sup>d</sup>	B <sup>e</sup>	C <sup>f</sup>		Low	High	N/A	
<i>Z.maize</i>	686	8835	3918	4291	626	6009	377	8458	0	5889
<i>O.sativa japonica</i> <sup>d</sup>	509	7183	4677	1540	966	5512	666	6508	9	5186
<i>A.thaliana</i>	1194	8334	1659	2532	4143	3461	1344	6993	2	2723
Total										6492 <sup>h</sup>

Key to legend: <sup>a</sup> Rice; <sup>b</sup> genes without sorghum orthologs; <sup>c</sup> Total sorghum orthologs; <sup>d</sup> one2one orthology type; <sup>e</sup> one2many orthology type; <sup>f</sup> many2many orthology type; <sup>g</sup> selected sorghum orthologs above 50% identity and high confidence level; <sup>h</sup> non-redundant total sorghum orthologs.

Then we decided to use the 6492 genes as query input for GO functional enrichment among which



6321 enriched annotation and 239 significant GO-terms were identified under p-value, FDR < 0.05. We reduced the number to our most genes of interest with 1102 highly enriched drought responsive genes by selecting a response to stress as a key GO-term. Consequently, huge over-representation of genes seems to promising in drought stress tolerance, for instance genes involving in responses to water deprivation (118), dessication (21), heat (91), ABA stimulus (109) and ABA mediated signalling pathways (37) are among few identified GO functional terms (see Table 2.18 and Figure S2.6). Table 2.18 shows the brief description of GO functional enrichment of DR sorghum genes identified based on orthology groups and Figure S2.6 represents mapping of the GO-terms related to responses to stress based on biological processes. Surprisingly enough significant over-representation of high number of genes which were validated by gene ontology functional enrichment were identified from sorghum orthologs in evolutionarily related grass species.

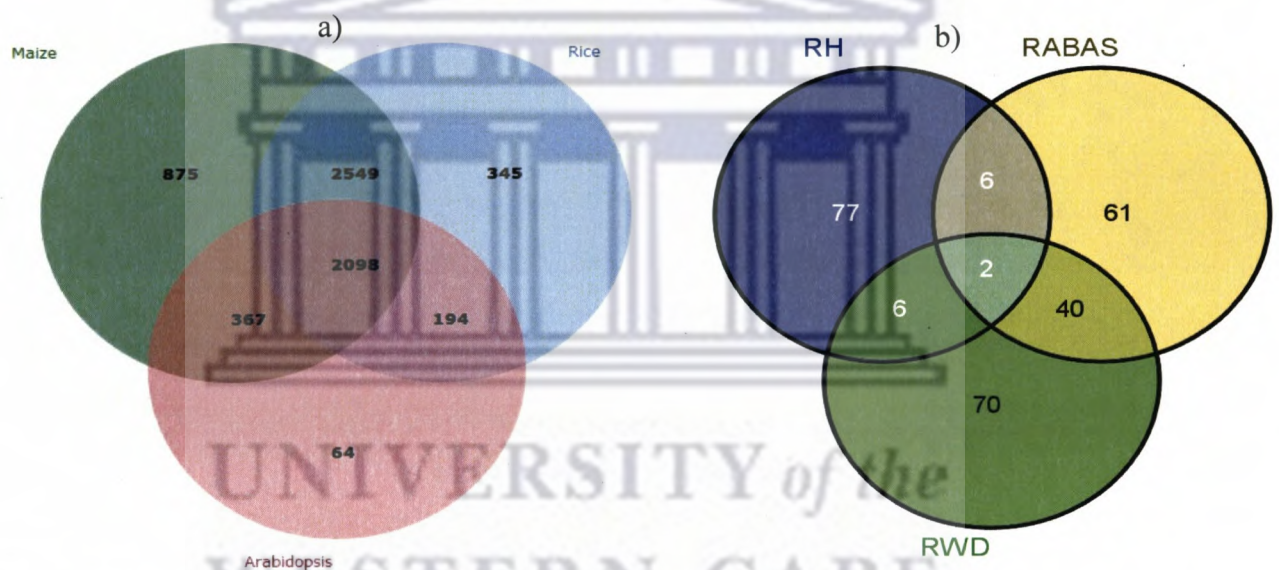


Figure 2.10: Sorghum orthologs correlating among species and drought related GO terms

Key to legend: RWD = response to water deprivation; RH = response to heat and RABAs = response to ABAs.

This Venn-diagram showing patterns of overlapping of sorghum orthologous genes among its relative species and among GO terms related to drought stress across species defined promising drought responses based on the orthologous group: a) shows patterns of sorghum orthologs evolutionary and functional crosstalk with genes in other species. This may give some clue though needs to be further proven on the extent of shared conserved synteny among species related to sorghum such that closely related species such as maize and rice share higher conserved (2549, 39%) orthologs than relatively distantly related species to sorghum eg. maize and Arabidopsis that share 367 (6%) sorghum orthologs and rice and Arabidopsis share only 3% orthologs. Surprisingly, 2098 sorghum orthologs shared among all the species seemingly represent ancestral genes. All the genes in the diagram represent sorghum orthologs in the respective species. The non-shared ones indicate the unique sorghum orthologs found only in the corresponding species. b) patterns of genes involved in key selected GO-terms representing stress response. Functional overlapping is indicated as a clue for gene network among categories involve in complex stress responses with some genes playing the rate limiting role (eg.

Sb09g026860.1 and Sb07g014940.1) acting in all pathways. Pathway controlling response to water deprivation (RWD) shares 40 overlapping genes with response to ABA stimulus (RABAS) and six genes with response to heat (RH) (see Table 2.18). In total 265 unique genes were identified from orthology groups to play active role in drought related responses.

Table 2.18 GO functional enrichment of DR sorghum genes based on orthology groups.

GO-D <sup>A</sup>	GO ID	GO-term	E/T <sup>B</sup>	GO/Bg <sup>C</sup>	P <sup>D</sup>
BP	GO:0006950	Response to stress	1102/1102	3705/26245	0.00
BP	GO:0006950	Response to water deprivation	118/1102	374/26245	4.98e-55
BP	GO:0009269	Response to dessication	21/1102	71/26245	5.4e-09
BP	GO:0009408	Response to heat	91/1102	324/26245	1.35e-38
BP	GO:0009737	Response to ABA stimulus	109/1102	664/26245	2.87e-28
BP	GO:0009738	ABA mediated signalling pathway	37/1102	192/26245	4.87e-11
CC	GO:0044444	Cytoplasmic part	477/1102	7660/26245	2.07e-20
CC	GO:0009536	Plastid	161/1102	2109/26245	5.95e-11
CC	GO:0009507	Chloroplast	144/1102	183/26245	1.45e-10
CC	GO:0043227	Membrane-bounded organelle	628/1102	10408/26245	1.24e-27
CC	GO:0005886	Plasma membrane	136/1102	1557/26245	6.93e-13
MF	GO:0051082	Unfolding protein binding	36/1102	184/26245	1.83e-10
MF	GO:0048037	Cofactor binding	81/1102	792/26245	5.83e-10
MF	GO:0005506	Ion binding	91/1102	890/26245	3.17e-11
MF	GO:0016209	Antioxidant activity	72/1102	276/26245	3.88e-28
MF	GO:0009055	Electron transfer activity	86/1102	984/26245	1.59e-07

**Key to legend:** <sup>A</sup> GO-D denotes GO domain; <sup>B</sup> E/T denote enriched genes Vs test set: the numerator and denominator represent the number of enriched genes and test set respectively; <sup>C</sup> GO/Bg denote genes with GO-terms Vs back ground set: the numerator and denominator represent the number of genes for which corresponding GO-terms were assigned and the background set respectively; <sup>D</sup> P denote P-value (FDR). The first top GO-term in each GO-domain, eg. BP: response to stress (GO:0006950), CC: Cytoplasmic part (GO:0044444) and MF: Unfolding protein binding (GO:0051082), represent the most likely term associated to the corresponding number of enriched genes. As such, 1102, 477, and 36 genes respectively were assigned to these GO-terms.

## 2.4 Discussion

Detection of complex trait related genetic determinants on an *in silico* basis probably is the best approach to identify functional candidate genes. An integrated *in silico* approach that we applied in this study generated a wide array of candidate drought responsive genes in sorghum.

### 2.4.1 Identification of candidate genes by mapping experimental data to reference genome

Mapping data to reference genome is not only important for molecular characterization of genome structure and evolution in the grass family (Rubin *et al.* 2000; Feuillet and Keller, 2002; Keller *et al.*, 2011), but is also vital for comparative genomics in aspects including but not limited to predicting and verifying gene models, identifying and characterizing putatively known candidate genes, improving genome annotation, and identifying homologs between genomes of related species in the eukaryotes (Schnable and Lyons, 2012; Varshney *et al.* 2011). Sequence similarity search now for more than two decades since introduction of BLAST (Altschul *et al.*, 1990) has been the focus in DNA or protein query search for sequence similarities against known databases (Kisman *et al.* (2005) with likelihood of matched sequences on similarity measure returning a set of high-scoring pairs (HSPs) and reflecting evolutionary relationship. The result of the UniGene mapping to sorghum genome in the current study captured 123 DRG (1% out of 14057 UniGene clusters and 1.3% out of the 9258 DRESTs, part of UniGene clusters) strongly supported by sorghum existing gene models originally not ascribed as drought tolerance. Based on the status of their functional annotation of these genes, we able to classify as hypothetical (61%), as putatively uncharacterised (1%) and as unknown proteins, 38% (Figure S2.1). Analysis of a collection of the 9258 DR single-pass ESTs selected out of 20199 initially derived from 92 different sorghum cDNA libraries that were incorporated into a set of 14057 UniGene clusters (Pontius *et al.*, 2003) revealed significant number of identifiable CDRG.

Forty-one UniGene clusters purely DRESTs (0.3% out of 14057) that mapped to the reference genome might give information on the function of the genes towards drought tolerance and these could be tagged as potential target for further investigation. However, 82.4% DREST free UniGene clusters that mapped to genome represent drought stress susceptibility suggesting that drought tolerance loci are not frequently and uniformly distributed along the genome owing to

developmental stage or tissue-specificity of spatio-temporal patterns of gene expression under stress condition.

## **2.4.2 Annotation comparison and update**

### **2.4.2.1 Genome annotation modification**

Locating protein coding genes using *in silico* tracing is probably the most difficult task of genome annotation. Identification of new structure of the existing gene model which we refer to as annotation update and the novel gene structure model are the two major achievements in the current study. The need for annotation comparison is not just restricted to different versions of annotation of the same genome but also of different sources derived from distinct gene prediction pipelines (Standage and Brendel, 2012). Since gene structure prediction is not just a one time complete endeavour that exhaustively describe all possible gene sets in the genome, a long term dynamic and additive process of a variety of efforts requires for progressive update of genome annotation.

Comparative genomics has provided opportunities to investigate not only genome structures but additional phenomena such as alternative splicing, exonic variances and untranslated parts by tracing homology based similarities and differences between organisms (Modrek and Lee, 2003; Singh *et al.*, 2008). Alternative splicing in several model organisms eg. human, mouse and rat is associated with an increased frequency of exon creation and/or loss (Modrek and Lee, 2003; Singh *et al.*, 2008) though there is low level of alternatively spliced genes in plants probably for reasons related to plant evolution (Barbazuk *et al.*, 2008, Keren *et al.*, 2010).

Incorporation of additional expression data sets is the main source of annotation update. Genome annotation is already subject to change in different organisms via the use of new data repositories and tools publicly available (Koonin *et al.*, 2003). In this study, improvements made as a consequence of structural and positional modification of one or multiple gene features in the existing genome annotation signifies annotation update. Variability in the genomic features that may be associated with diversification of tissue-specific expression patterns of protein coding genes and the resulting diversification in protein function may be a biological implication of such modifications.

Analysis of the present genome annotation compare revealed 12.5% modification that includes

single gene model updates of which about 5% represent both structural and positional rearrangements. The rest 95% attained only structural changes being chromosomal assignment remained unchanged. Merging genes based on multiple overlapping transcripts and novel exonic and UTR contributed to the improvement of annotation. The PASA improved 4557 mRNAs originated from updated 4455 genes based on the datasets obtained from UniGene clusters and drought responsive TIGR transcript is an indication of changes in gene structure and probable shift in the start and stop codons. Addition of 64 novel exons, 74 five prime UTR, 3595 three prime UTR entail the generation of different isoforms from a single gene model leading to variations between protein coding genes and the corresponding products (Modrek and Lee, 2002). This additional genetic variation may indispensably be implicated in the enhancement of drought tolerance and yield stability in sorghum.

#### **2.4.2.2 Novel gene structure model prediction**

The major outcome of annotation update in this study is the findings of 241 novel sorghum genes out of which 69% were found to be DR. This result describes 34% among the drought responsive genes were originated from extrinsic evidence based prediction using AUGUSTUS gene prediction program (Stanke *et al.*, 2006b) and PASA pipeline (Haas *et al.*, 2008). The gene building and mapping pipelines we designed for this purpose have been instrumental in filtering out the 441 initially identified genes by subjecting to series of screening procedures. In contrary to its genome size the number of identified genes in Sb1.4 (Reseales v2.1, Sbi1.4, latest release) (Paterson *et al.*, 2009) is relatively low in sorghum as a model species for cereals compared to other model species. Rice whose genome size is 75% smaller than that of sorghum has gone through several reannotation and refinement steps (Ouyang *et al.*, 2007) with a total of 55,986 identified Non-TE and TE loci and total of 66, 343 gene models (RGAP 7 Summary; Kawahara *et al.*, 2013). This entails the relatively slow annotation updates in sorghum genome from 2009 where it was initially annotated. This study, contributed to the improvement of the existing annotation by adding novel putative gene structure models, thus enhancing the quality of sorghum genome annotation and furthering our understanding of sorghum genomics.

##### **2.4.2.2.1 Complete and partial gene structure models**

Interestingly, 14 novel genes (6%) were found to be complete in structure with the presence of both 3 prime and 5 prime UTR edges. The rest 8% were shown to have only 3 prime UTR and 1% with

only 5 prime UTR. This result reveals that 34 genes (14%) were at least semi-complete gene structured i.e. only 3 prime UTR edge or 5 prime UTR or both) and the remaining 207 were shown to be only partial with no any UTR segment of the gene but with the start and stop codons. This truncation could arise due to an in-frame stop codon (Liu *et al.*, 2009) or however often exhibited in the nature of our test dataset that most ESTs are shorter and they are sensitive to errors in predicting whether a gene is truncated at one or both ends (Klassen and Currie, 2012). The finding of the complete gene structure is implicated in to a more potentially featured new functional elements to the sorghum genome annotation.

#### **2.4.2.2.2 Single exonic and intronless genes**

Further analysis on our result depicts 115 genes (47.7%) to be single exonic which were all intronless and the remaining genes (52.3%) included those having 2 exons (51, 21.2%), 3 exonic (40, 16.6%), 4 exonic (17, 7.1%), 5 exonic (10, 4.2%), 6 exonic (3, 1.3%) with no seven exonic gene and 2 eight exonic genes (1%). The finding of single exonic intronless genes is very interesting as several reports shown that intron loss play role in drought tolerance. For instance, the ORF of CBF4 (C-repeat binding factor /dehydration-responsive elements binding (CBF/DREB1) proteins), a drought stress inducible gene was shown to be associated with an intronless expressed gene (Haake *et al.*, 2002). Besides, sequence analysis of DREB1 genes (dehydration-responsive-element-binding protein), which play an important role in increasing stress tolerance in plants, showed that they are intronless (Akhtar *et al.*, 2012).

#### **2.4.2.2.3 Correlation between intronless and pseudogenes**

A more interesting in connection with intron loss is its correlation in frequency with processed pseudogene abundance which would be seen as a novel strategy to test the reverse transcriptase model of intron loss (Zhu and Niu, 2013). Pseudogenes are defined as functional anomalies of the previously intact protein coding genomic loci having sequence homology with the functional genes which often referred to as parent paralogs. With such defunct due to frame shifts mutation, interrupted stop codon and gaps within conserved regions (Balakirev and Ayala, 2003 and Niimura, 2013), they are grouped into three known classes based on their origin: (a) retrotransposition of mRNA from functional protein-coding loci back into the genome, referred to as processed pseudogene (Balakirev and Ayala, 2003); (b) duplication of functional genes referred to as

duplicated (also unprocessed) pseudogenes (Li *et al.*, 2013) and (c) *in situ* mutations in previously functional protein-coding genes referred to as unitary pseudogenes (Zhang *et al.*, 2010). A consensus for the gene to be categorized into pseudogene based on the combined criteria set by Ensembl (Curwen *et al.*, 2004; Flicek *et al.*, 2013 and 2014) and those used by Goodstadt and Ponting, 2006 pointed out these features as pseudo-genes: 1) a characteristic short introns with less than 10 bp, 2) frame shift (in-frame stop codon disruptions); 3) the lack of conserved syntenic gene order in dispersed genes (syntenic distance of 20 genes); 4) any single or multiple disrupted interspersed gene, identified syntenic but with more than one disruption and dispersed gene with single exon. Even though we suspect for the presence of pseudogenes from this result in correlation with the finding of single exonic intronless genes, it requires further investigation.

#### **2.4.2.2.4 Splisomes**

Alternative splicing is a major regulatory mechanism in eukaryote gene expression and it has evolutionary implications in diversification of structures in gene products and their functions (Keren *et al.*, 2010). This allows generation of multiple mRNA species and proteins from a single gene with all potential informational content of eukaryotic genomes (Ner-Gaon *et al.*, 2004). The identified 136 AS in our result suggests importance of splice event in regulation of the levels and tissue specificity of gene expression in sorghum crop. In most cases it may cause unprecedented disorder without the occurrence of such phenomena (Tazi *et al.*, 2009). Alternate exon, a 12% splice event in our finding is related to an increase in coding diversity within genes coding for extracellular matrix proteins (Boyd *et al.*, 1993). Our analysis shows nine alternative splice junctions of which 80% may significantly involve in the variability of transcripts.

#### **2.4.3 Metabolic pathways**

With the complete sequencing and annotation of eukaryotic genomes, it's becoming easier a task to assign the coding regions where the majority of genes encode products with known metabolic and biochemical functions (Ouzounis *et al.*, 1996; Schilling *et al.*, 1999; Brown *et al.*, 2014). In this study, we identified fourteen metabolic pathways related to drought tolerance and the total of 32 genes for which enriched drought associated GO-terms were assigned out of 477 involved. Sorghum has the ability to synthesize dhurrin, a Cyanogenic Glucosides (CGs) and store in the tissues without any effect of the toxic cyanide unlike most plants (Niang, 2008). In sorghum an enzyme, CYP79A1 [EC:1.14.13.41] (tyrosine N-monooxygenase also called tyrosine N-

hydroxylase) which is grouped in a class of Oxidoreductases (Halkier *et al.*, 1995; Bak *et al.* 2000) is involved in biosynthesis of the cyanogenic glucoside dhurrin, along with some other enzymes. As study shows, dhurrin synthesis in sorghum depends on developmental stage and growth condition and is largely determined by transcriptional regulation of the biosynthetic enzymes CYP79A1 and CYP71E1 (Busk and Moller, 2002). A gene known with Sb01g001200 name being classified under the protein family Pfam p450 is responsible to encode CYP79A1 and CYP71E1 (Bak *et al.*, 2000). Though clear understanding is not in place how dhurrin involve in drought tolerance in sorghum, the alternative pathway for the degradation of dhurrin is hypothesized in the latter developmental stage stimulating sorghum for endogenous turnover pathway allowing the plant to recycle the nitrogen bound in dhurrin without the risk of toxic effects from hydrogen cyanide released inside the cells (Bach, 2012). However, based on functional ontology assignment, the gene, Sb01g001200 (CYP79A1), a putatively uncharacterised hypothetical protein, has been identified with the GO-accessions GO:009414 and GO:0009269 with the corresponding GO-terms for the biological process, response to both water deprivation and desiccation respectively suggesting the direct involvement of this particular gene in function related to drought tolerance.

On the other hand, the three metabolic pathways namely PcoAB, VLIB and VLID with an enzyme EC # [EC:2.6.1.42] in common among others involve in the amination of 4-methyl-2-oxopentanoate attributed to the presence of a branched-chain amino acid transaminase (BCAT) activity. The enzyme is classified under transferases, transferring nitrogenous groups, and transaminases based on the particular reaction it catalyses. The three genes namely Sb04g010240, Sb06g025140 and Sb09g008180 were identified to involve commonly in these pathways and were associated to GO-terms such as response to stress, water deprivation and response to desiccation implicating that the genes are actively involved in drought stress tolerance.

From this analysis, it is possible to suggest that a set of genes seem to involve in the pathways in two specific approaches: 1) across pathways playing a multiplex metabolic role and 2) within pathway(s) playing pathway specific metabolic role. The three genes mentioned above involve in all the three metabolic pathways in common whereas *cox1* involve only in Oxidative phosphorylation though it may interact with more than one gene within pathway as it shares the pathway with others. However, not clear why three genes act together in playing on the same enzymatic activity may rise



a question about their functional duplication. Analysis of patterns of expression profiling of such genes may give clue to distinguish between them, however, still important to know that the timing of enzymatic role may not necessarily correlate with the mRNA abundance (Glanemann *et al.*, 2003).

#### **2.4.4 Functional GO enrichment and Interpro domain analysis**

In total, this analysis revealed significant number of genes (477) networking in all the pathways for which 583 GO-terms were significantly enriched under P-value and FDR < 0.01) of which 32 potential genes noted to be responsible for encoding key enzymes. Analysis of GO annotation, in terms of the GO-domain representation, revealed 32% branched chain family amino acid metabolic process, 29% and 30% cofactor and oxidoreductase binding protein respectively and 91% with 29% cytoplasm and plastid cellular component respectively.

Interpro domain analysis revealed high frequency of protein domains related to drought tolerance (Isokpehi *et al.*, 2011) such as zinc finger domain representing common elements in drought stress response in plants (Vij and Tyagi, 2008) and Chaperon DnaJ domain protein playing functional role in the cooperation of Hsp40 with Hsp70 (Greene *et al.*, 1998) and in intracellular or endosomal trafficking (Girard *et al.*, 2005).

#### **2.4.5 Differential gene expression profiling**

Analysis of gene expression is vital means of interpreting gained information regarding gene expression or transcriptional profiling to discover and develop defensive process in complex trait controlled systems. It discloses polygenic and pleiotropic networks that modulate systems functioning (Chesler *et al.*, 2005). This suggests that gene expression analysis plays pivotal role in candidate gene analysis from distinctly observed expressed set of genes or gene features or subtypes that provide clue of a particular biological state. Expression profiling can be used to prioritize a candidate gene list that would otherwise have been a difficult task of using reverse genetics to assign functionality to genes (Kreps *et al.*, 2002). It can serve as a proper tool to more accurately classify gene features (Goldstein *et al.*, 2013). Analysis of gene expression based on maize orthologs revealed 32 significantly expressed sorghum genes (a maize orthologs) in association with drought tolerance, majority being from ovary tissues in line with the published work (Kakumanu *et al.* 2012). This denotes conserved functional similarity in drought stress responses between the two

crops based on the fact that expression context (co-expression of genes with others having counterparts in other genome) is largely conserved between orthologs (Dutilh *et al.*, 2006). However, sequence similarity based extrapolation of gene expression profiles of a species to the ortholog of its closest relative will only be applied if the similarity holds true for functional conservation across species (Sánchez *et al.*, 2000). GO annotation notably revealed maximum %age of photosynthetic cellular metabolic process, ion and chlorophyll binding typically involving in both photosyntheses I and II significantly correlating with drought stress responses.

#### **2.4.6 Analysis of orthology relationship**

Identification of orthologous groups has been instrumental in wide array of research areas. Some of these include but not limited to evolutionary findings of genes and or proteins (Devos and Gale, 2000, Glazier *et al.*, 2002; Wu *et al.*, 2006), genome annotation (Itoh *et al.*, 2007) and comparative genomic studies (Koonin *et al.*, 2004, Proost *et al.*, 2009) and identification of genes using CGA (Mitchell *et al.*, 2007). The present analysis provided huge over-representation of genes promising in drought stress tolerance with the total prioritized added up to 5.1% functionally enriched orthologs based on the 7223 total identified initial list. For instance in responses to water deprivation (118), desiccation (21), heat (91), ABA stimulus (109) and ABA mediated signalling pathways (37) were among few identified. There is hint for functional overlapping across pathways indicating the presence of gene network among categories involve in complex stress responses. Maize which diverged from sorghum 12 mya (Swigonova *et al.*, 2004) contributed the largest orthologs than rice and Arabidopsis from which sorghum diverged 42 mya (Paterson *et al.*, 2004; Tang *et al.*, 2008) and ~150mya (Bancroft, 2001) receptively. This further asserts that more functional gene conservation is plausible between sorghum and maize given their close relationship and the largely conserved gene content and number with extended regions of map collinearity as an evidence (Song *et al.*, 2002). In other words, the relative evolutionary distances between sorghum and other species created the extent of variation in the use of orthology in predicting candidate genes.

## 2.5 Conclusion

Our approach proves to be a well designed tool for detecting biologically plausible candidate genes. Reliability and validity of our data contributed to identification of significant array of prioritized candidate genes that are critical to response to drought and related stress genes. Because drought tolerance is a complex poligenetic traits, detection and genetic dissection of candidate genes requires the use of multi-analytical processes. Mapping experimental data to reference genome, pathway analysis, expression profiling, analysis of orthologous group and genome annotation and identification of novel genes contributed to 620 (~2%) non-redundant functionally enriched drought tolerant genes which were not ascribed in previous annotation. Structural and positional modification of gene annotation is implicated in the genomic structural variation and its consequences on a probable functional variation. Identification of 41 purely DR UniGene clusters underscore the importance of sorghum UniGenes in candidate gene discovery.

Expression profiling and orthologous group identification show high gene conservation along evolutionary related lineage, however the closer the lineage the greater the shared functional features would be. All the metabolic and biochemical pathways identified in this study suggest sorghum's C4 photosynthetic peculiarity. As a basic factional unit of metabolic system, these pathways interplay to create biochemical reactions that make up the metabolic network, constitute a fundamental interface to build a defensive mechanism against drought stress.

The pipeline designated for novel gene prediction is most dependable and reliable that employed multiple informants and standard quality control. This result has modified 12.6% of the existing annotation and incorporated almost 1% of the novel gene models. Yet untapped genetic variation in sorghum was witnessed in this study entailing the need for future research target. The result in the present study is of interest to further research in molecular breeding in sorghum towards enhancing drought tolerance and yield stability.

**Chapter 3: Gene-gene and gene-phenotype association: a novel integrated approach to dissect complex drought tolerance in sorghum (*Sorghum bicolor* (L.) Moench)**

**Abstract**

**Background:** Identification of genes associated with complex traits is a common challenge in eukaryotic genomes. Dissecting genetic determinants for normal biological function in plants under drought stress is difficult due to complexity in drought factors and the polygenicity of the trait. However, association studies has long been a way-forward in genetic dissection of complex traits such as drought resistance.

**Methodology:** An integrated approach that combine functional ontology based semantic data with expression profiling and biological networks was employed to analyse gene association with plant phenotypes and to identify and genetically dissect complex drought tolerance in sorghum (*Sorghum bicolor* (L.) Moench) and related species. The gramene database was used to identify genes with direct or indirect association to drought related ontology terms in sorghum. Where direct association for sorghum genes were not available, genes were captured using Ensemble Biomart by transitive association based on the putative functions of sorghum orthologs in closely related species. Semantic query building components were used to determine associations to all ontology terms. Ontology mapping represented the direct or transitive association of genes to multiple drought related ontology terms based on sorghum specific genes or orthologs in related species. Based on this, trait functional specificity and overlapping across species was determined. Non-redundant multi-ontology supported genes were further enriched using gene ontology (GO) enrichment analysis. Metabolic pathways, functional biological interaction and phylogenetic distances were identified for selected gene associations. Comparative GO associated drought responsive genes were determined between sorghum and related species. Correlation of genes to the enriched GO-terms related to the whole-plant structure was used to determine extent of gene-phenotype association across-species and environmental stresses. Genes enriched for functions related to drought resistant ( $p < 0.05$ ) were used in analysis of gene expression profiling.

**Result:** We demonstrated the effectiveness of our approach by cross-examining the association of

169 sorghum genes identified for drought tolerance across species and environmental stresses. While 56% of these have shown multiple stress tolerance in sorghum, 90% exhibited drought tolerance in multiple species and 10% identified to be sorghum specific. We integrated gene-to-phenotype associations and relevant public expression datasets from related cereal crops and model organisms. Based on the biological processes they were involved in, we identified 1117 sorghum candidate genes which potentially respond to five different abiotic stresses such as drought (169), salt (352), cold (222), heat (92) and oxidative stress (282). Based on expression profiling, 88 of these genes were associated with drought response. A total of 2224 non-redundant genes exhibited strong association with GO (6%), trait ontology (TO, 13%), plant ontology (PO, 4.4%), plant growth ontology (GRO, 32%) and plant environment ontology (EO, 75%) of which 30% is shared among all ontologies ( $p$ -value  $< 0.05$ ). Mapping further ontology validated the results and provided biological functions for identified genes.

**Conclusion:** Our approach allows intermarriage of gene association with functionally interrelated, but not overlapping ontologies terms to identify and genetically dissect complex drought tolerance in sorghum (*Sorghum bicolor* (L.) Moench). The resource enables us to perform cross-species and stresses queries for genes that are likely to be associated with multiple stress tolerances, as a means to identify novel targets for engineering stress resistance in sorghum and other crop species.

### 3.1 Introduction

One of the most daunting tasks in plant genetics is to identify the genetic determinants related to complex traits and to decipher the molecular basis of these traits. The use of technologies such as association studies and expression arrays have provided opportunities for the major trade-offs (Brunner *et al.*, 2004). Association studies usually correlate complex trait-related genetic backgrounds to the corresponding chromosomal regions (Atwel *et al.*, 2010), whereas expression profiling allows researchers to obtain a list of differentially regulated genes in an experimental sample with respect to the reference (Alba *et al.*, 2004; Wu *et al.*, 2026; Carrera *et al.*, 2007). Association studies can also be viewed in comparison with linkage analysis as both of them have long been instrumental in genetic analysis of quantitative drought tolerance (Champoux *et al.*, 1995; Tuberosa *et al.*, 2002; Eeuwijk *et al.*, 2010; Varshney *et al.*, 2012). However, the former tend to be more effective than the latter for analysing complex traits since it applies more statistical power to identify large number of genes with minimal effect (Risch *et al.*, 1996). In all of these technologies, a common characteristic challenge has been a resulting large number of genes corresponding to the analysis (Panagiotou *et al.*, 2013; Paux *et al.*, 2012; Wei *et al.*, 2012).

Integrated functional ontology based gene association using gene set enrichment tools for identification of complex traits is the most promising current approach to obtain relevant and concise number of genes (Jaiswal *et al.*, 2002; Cline *et al.*, 2007; Horan *et al.*, 2008; Ficklin *et al.*, 2010 and Lee *et al.*, 2011). This approach uses gene ontology as the basis for query building and semantic integration of data (Song *et al.*, 2013). It allows identification of genes regulating complex traits by using expression data profiling (Carrera *et al.*, 2007), biological networks (Cline *et al.*, 2007) and data mining from known biological information (Jaiswal *et al.*, 2002). Previous studies on sorghum gene-phenotype modelling at crop level were concerned about the challenge of gene-gene and gene-environment interactions with respect to modern breeding approaches (Capman *et al.*, 2002; Chapman *et al.*, 2003). Integrated functional ontology approach employs multiple means for identification of key physiological and developmental traits (Tonsor *et al.*, 2005) that relate to gene-phenotype association.

The ontology-based identification of complex traits using association analysis includes a wider spectrum of interrelated components. This approach uses primarily five ontologies namely Gene

Ontology (GO); Trait Ontology (TO); Plant Structure Ontology (PO); Plant Growth Ontology (GRO) and Environment Ontology (EO) to integrate association studies such as gene-gene and gene-trait associations. Ontology has been represented and used since the time of Parmenides, an ancient Greek philosopher (Cordero, 2004) to describe the fundamental characteristics of reality based on the differences and similarities with all the relationships manifested within the realities. Ontology as the philosophical thinking of the nature of *being* deals with the concept of what makes the biological entity itself and what not. It represents well-structured and controlled vocabularies with well-defined relationships (Grube *et al.*, 1993; Smith *et al.* 2007) of which the structure denotes the current representation of biological knowledge (Ashburner *et al.*, 2000).

Among other ontologies, the GO is the first that our approach uses to dissect a genetically complex trait. GO is a well-defined and structured shared knowledge in three interrelated but non-overlapping domains of molecular biology such as biological process (BP), molecular function (MF) and cellular component (CC) which are all attributes of genes, gene products or gene-product groups (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2001; Bard and Rhee, 2004 and Rhee *et al.*, 2008). The three GO domains mentioned above represent a biological objective to which the gene or gene product contributes, a biochemical activity of a gene product and the place in the cell where a gene product is active respectively. GO is very important because it makes possible the annotation of homologous gene and protein sequences across organisms based on shared biology and the association of genes to the respective nodes within an ontology (Ashburner *et al.*, 2000). It deals with gene-centered information such as gene-gene relationship, association and interaction as well as protein-protein interaction (Yue *et al.*, 2006; Pattin and Moore, 2008 and 2009; Moore *et al.*, 2010; Califano *et al.*, 2012) and mapping of genes to known GO-terms based on biological functions from all GO-domains.

Trait Ontology (INGER. 1996; Jaiswal *et al.*, 2002; Youens-Clark *et al.* 2010 and Arnaud *et al.*, 2012) is a structured vocabulary of terms denoting a phenotypic traits in plants notably plant height, chlorophyll content and stay green are few among many. These are classified into genetic, agronomic, biochemical, physiological, developmental traits based on the categories they represent which are familiar in nature but not distinct and are often complementary (Jaiswal *et al.*, 2002). Such non-distinctiveness is solved by TO allowing 'one to many' relationships (Bard and Rhee,

2004; Jaiswal *et al.*, 2006; Balhoff *et al.*, 2010 and Kattge *et al.*, 2011). TO deals with gene trait association (Plant Ontology Consortium, 2002; Bard and Rhee, 2004; Dixon *et al.*, 2007 and Youens-Clark *et al.*, 2010).

Third, PO is the first generic ontological representation of anatomical and morphological structure of all plants (Yamazaki *et al.*, 2005; Ilic *et al.*, 2007; Avraham *et al.*, 2008). Like TO, PO addresses the same problem arising due to inconsistencies in terminologies used to describe plant structure in publications and genomic databases. Plant Ontology allows description of gene association to plant morphological and anatomical structures (Da Cruz and de Macedo Vieira, 2010; Harnsomburana *et al.*, 2011 and Cooper *et al.*, 2013).

Fourthly, GRO describes distinct growth and developmental stage contained within plant biology (Pujar *et al.*, 2006) dealing with gene association with such distinct plant physical growth and age based on differences in tissues groups (Pujar *et al.*, 2006). On the other hand, EO represents description of a well-defined regimen of a plant (Youens-Clark *et al.*, 2010). Environment Ontology models the association and interaction of genes to different environment regimen and factors.

All these ontologies provide distinct descriptions of attributes for association to respective drought terms. However, it may not be convenient and convincing if the identification and analysis of gene association is limited to only one or few complex traits which are obviously lacking complementary information. Integrative approach that considers all the ontologies combined that can examine wider content of complex traits with expression data and biological networks is the most promising method to address this issue.

Therefore, this study investigates gene-gene and gene-phenotype association by using an integrated approach to genetically dissect complex drought tolerance in sorghum (*Sorghum bicolor* (L.) Moench).



## 3.2 Materials and Method

### 3.2.1 Data source and data mining

Plant related functional ontologies were identified using the gramene database. These include; Gene Ontology (GO), Trait Ontology (TO), Environment Ontology (EO), Plant structure Ontology (PO) and Growth Ontology (GRO). These were used to retrieve sorghum genes directly or indirectly associated with drought tolerance. To determine direct association, drought related ontology terms were first identified for each specified ontology including the number of genes that they represent for sorghum (Figure 3.1). Where direct association of sorghum gene-trait in question was not available from the respective ontologies, potential drought tolerant sorghum genes were captured using Ensemble Biomart by transitive association based on the putative functions of sorghum gene orthologs in other closely related crop species. We used ontology mapping (see Figure 3.2) to represent direct or transitive association of sorghum genes to multiple drought related ontology terms based on orthology functional relationship in maize, a closely related cereal crop as well as the distantly related rice and Arabidopsis.

Once sorghum drought associated genes for all ontologies were identified and retrieved, those gene associations supported by all ontology terms in each ontology group were retained and merged to capture only unique entries. Further, genes supported by all ontology groups were used as an input for functional GO-enrichment,  $P < 0.01$  using Agrigo (Du *et al.*, 2010).

Species specific and common genes were identified based on gene functional association across-species and multi-environmental stresses. VENNY (see Figures 3.2 and 3.5), an interactive tool for comparing lists of genes with Venn Diagrams (Oliveros, 2007) was used to display and visualize unique and common gene groups based on the attributes they are involved in across species and multiple drought related stresses. KEGG pathway (Kanehisa and Goto., 2000), biological networks (Supek *et al.*, 2011) and (Mihara *et al.*, 2010) were used to show gene association in terms of their metabolic role, functional biological interaction and phylogenetic distances respectively (see section 3.2.4 and 3.3.1.1 for detail).

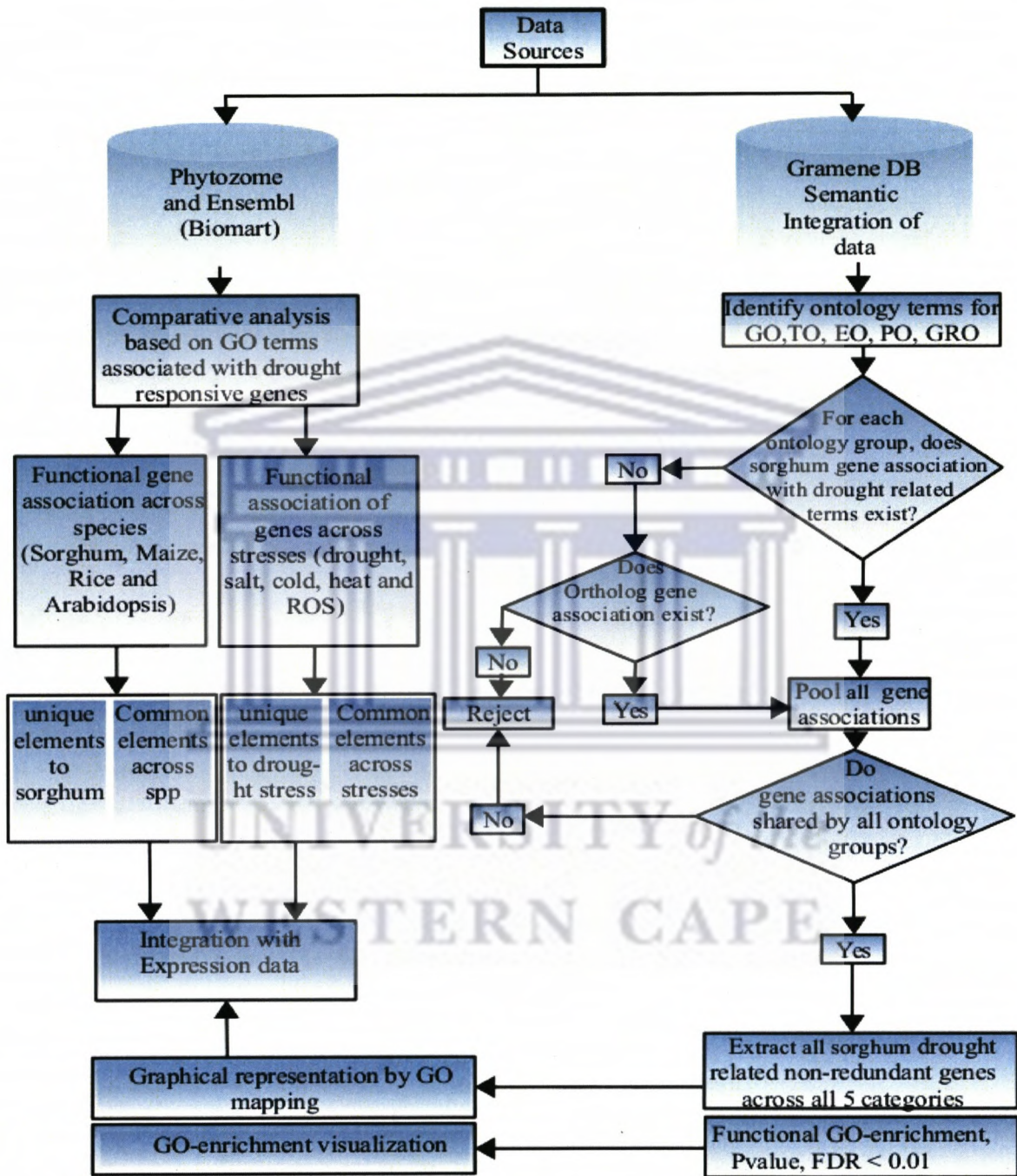


Figure 3.1: Work-flow for gene-phenotype association across-species and stresses

This figure demonstrates the work flow for the gene-phenotype association in sorghum drought tolerance across species by comparing sorghum drought responsive genes with orthologs in related species and across stresses by comparing these genes against other stress such as salt, cold, heat and ROS. Gramene database was used in identification of sorghum genes with drought phenotype association using semantic integration of data based on the known drought related ontology terms for each identified plant ontologies such as GO, TO, EO, PO and GRO. Ensemble Biomart was

used to get sorghum orthologs that have transitive association with known drought regulated functions from related species such as rice and maize. Sorghum specific genes and orthologs having association with drought responses were then integrated with expression data to evaluate expression profiling.

### **3.2.2 Identification of gene association using functional ontology based semantic query building**

Gramene (Ware *et al.*, 2002), Ensembl (Ensembl plant) (Flicek *et al.*, 2013) and Ontology (LePendu *et al.*, 2008) databases were used to identify sorghum genes functionally associated with plant phenotypes. The resultant extracted knowledge was semantically integrated based on drought related terms from different ontologies. Semantic queries pertain to knowledge or data expressed on the basis of a common vocabulary that leverage semantic information stored in ontologies used to filter and retrieve the data from relational tables (Yu, 2011). Investigation of gene-phenotype association was based on the correlation of genes to the enriched GO-terms related to the whole-plant structure, namely: plant phenotypic traits, plant anatomical and morphological structure and growth and developmental stages. Genes supported by all ontology terms were analysed for GO-enrichment ( $p\text{-value} < 0.05$ ) and the result obtained was integrated with expression data for comprehensive analysis.

### **3.2.3. Cross-species comparative analysis: correlating gene-trait association across species**

Comparative analysis were determined based on GO associated drought responsive genes for all GO-domains across species. Functional conservation was speculated for genes from sorghum and other three species. Ensembl Biomart (Smedley *et al.*, 2009) was used to trace sorghum orthologs in maize, rice and Arabidopsis based on non-redundant sorghum genes identified for GO with direct or indirect association to sorghum drought tolerance. Sorghum specific genes and those sharing attributes with other species were identified by determining cross-species gene functional association using Venny (Oliveros, 2007).

A list of sorghum orthologs were compared against each other for specificity and commonality in drought tolerance based on orthology relationship of all species. Genes functionally conserved were detected by investigating the attributes of orthologs in the respective species and then correlated with the genes identified for drought tolerance in sorghum.

### **3.2.4 Multiple responses of genes across environmental stresses**

Using the same initial input as in section 3.2.3 above, functional correlation of drought responsive genes were compared with genes responsive to other stresses that include salt, cold, heat and Reactive Oxygen Species (ROS). Sorghum drought tolerance specific gene association and multiple stress responses of genes were identified using same procedure described in section 3.2.3. (see Figure 3.1). Genes were selected based on the extent of their association to each environmental stress under particular ontology terms and then filtered based on their enrichment significance level (P-value, FDR < 0.05). Where data was lacking for sorghum, closely related orthologs were used to retrieve gene association. Sorghum-rice orthologs were almost entirely used because gramene data source is exhaustive (Ware *et al.*, 2002) for rice gene association.

#### **3.2.4.1 Metabolic pathway and phylogenetic relationship**

A metabolic pathway for two genes, Sb03g026070 and Sb09g030600, universally expressed along all stress environments was identified using Kyoto Encyclopaedia of Genes and Genomes (KEGG) map (see section 3.3.1.1 for the detail). All the genes co-occurring with the two genes and involved in the same pathway were identified both within the protein domains that the two genes belong and between different protein domains. A dendrogram showing a phylogenetic relationship of the two sorghum genes and their functional classification with orthologs from ancestrally related species was generated using SALAD, a tool of systematic comparison of proteome data (Mihara *et al.*, 2010).

#### **3.2.5 Integration of gene trait association with gene differential expression**

Sorghum expression data related to drought stress was obtained from NCBI Expression Omnibus (GEO) database (Barrett *et al.*, 2007). This was based on experimental data on sorghum transcriptome analysis (RNA-Seq) on 9 days seedlings in response to osmotic and abscisic acid stresses (Dugas *et al.*, 2011). This was done to integrate the patterns of gene trait association with tissues on which expression profiling showed drought phenotypes. Gene expression profiling was shown using heat map (Figure 3.6) and up and down regulated genes were visualized using volcano plot (Figure 3.7a&b). Statistical significance was determined using parametric t-test (P-value < 0.01).

### 3.2.6. Functional-annotation and GO Enrichment

Gene association was determined based on the enrichment level of GO terms based on the cut-off threshold value (p-value, FDR < 0.05). GO-terms with p-value less than 0.05 were considered significantly enriched for all the three domain namely BP, CC and MF. Similarly, enriched genes (FDR < 0.05) which exhibited strong association with their respective plant attribute from TO, PO, GRO and EO were also determined. To visualise drought related GO-term associated genes, a configured combined-graph tree (Figure S3.3); interactive biological networks (see Figure 3.9) and scatter plots (Figure 3.8) for multidimensional scaling of semantic similarities (Supek *et al.*, 2011) were generated using default values.



### 3.3 Result

#### 3.3.1 Gene association across-environmental stresses: Functional-cross-talk

Based on the biological processes in which they were involved, 1117 sorghum candidate genes were identified which potentially respond to five different abiotic stresses such as drought (169), salt (352), cold (222), heat (92) and oxidative stress (282) (Figure 3.2).

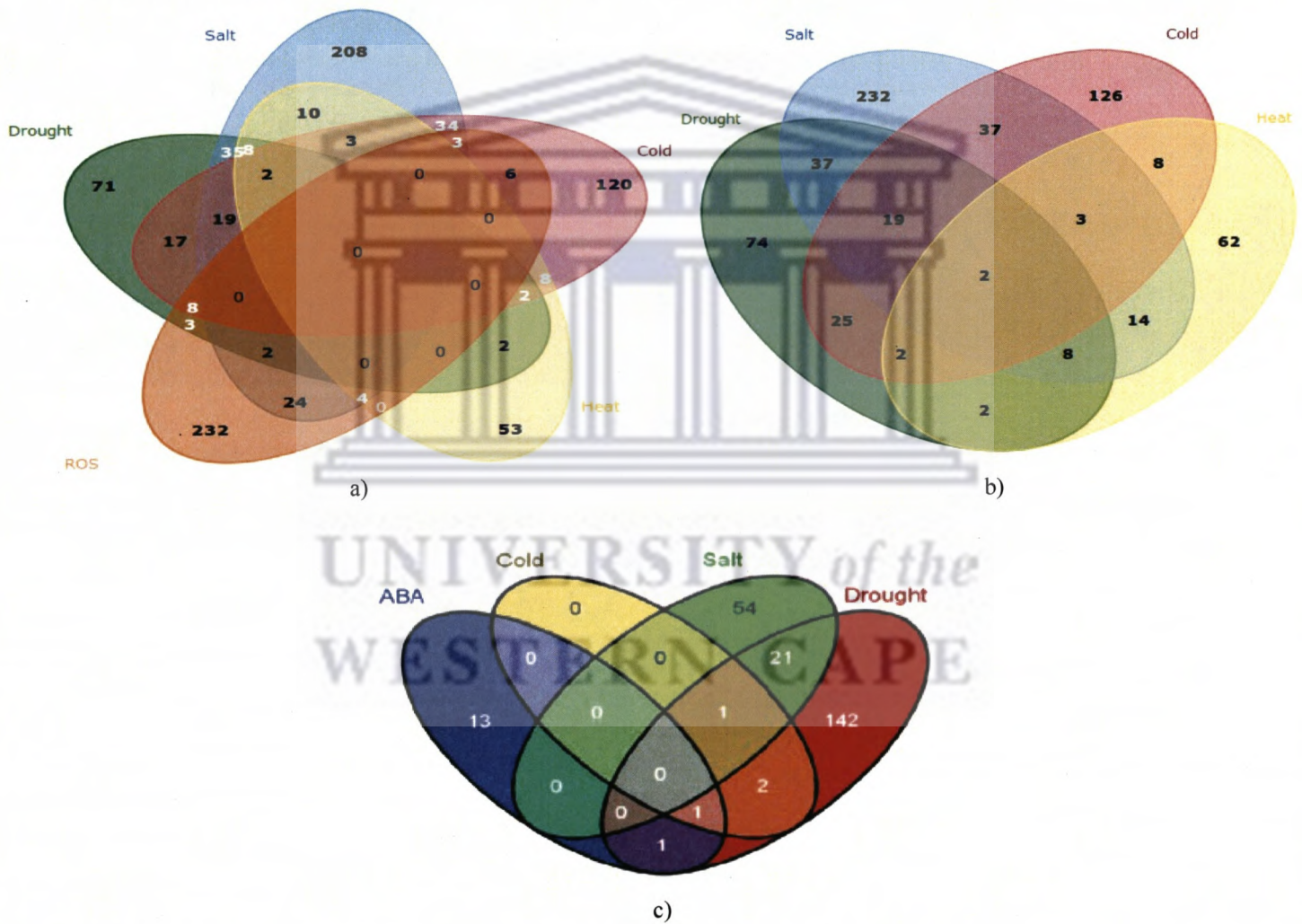


Figure 3.2: Functional correlation and specificity of the drought tolerance with other stresses.

Genes associated with five environmental stresses based on data from gramene database are indicated in Figure 3.2a; genes associated with four environmental stresses based on data from gramene database are indicated in Figure 2.3b and similarly, genes associated with four environmental stresses based on sequence similarity search are indicated in figure 2.3c. The venn diagram was constructed by using an integrative, flexible tool for comparing lists of genes or proteins (Oliveros, 2007).

A diverse functional array of sorghum gene association is characterized by the over-expression of some specific genes for multiple traits. For instance, two peculiar genes (Sb03g026070 and Sb09g030600) were identified to be simultaneously expressed in all the four stresses namely drought, salt, cold and heat. Furthermore, many other genes have been shown to have common expression in two or more environmental stresses. For example, 2 genes (Sb01g037090 and Sb02g043450) for drought, cold and heat, other 2 genes (Sb03g039820 and Sb09g022290) for drought, salt and ROS, still other 2 genes (Sb01g003880 and Sb10g023010) for drought and heat and 3 genes (Sb0010s007790, Sb01g031520 and Sb10g022780) for drought and ROS were found to be simultaneously expressed.

Similar results were also observed for the large number of genes interacting across environmental stresses. For example, 8 genes were shown to act commonly in three stresses: i) drought, cold and ROS and ii) drought, heat and salt each (Figure 3.2). Seventeen genes in drought and cold, 19 genes in drought, salt and cold and 35 other genes in drought and salt were commonly responsive (see Figure 3.2). On the other hand, stress specific genes were identified for all the five stresses such as 71 genes for drought, 232 for ROS, 208 for salt, 120 for cold and 53 genes for heat were shown to be unique elements.

### **3.3.1.1 Characteristic feature of 'SORBI\_03g026070' and 'SORBI\_09g030600': Implication in stress signal transduction pathway**

The two functionally cross-talking genes SORBI\_03g026070 or Sb03g026070 share several common features. As member of the 10 sorghum protein serine/theonine phosphatase catalytic subunit gene family that encode enzyme Protein Phosphatase 2C (PP2C), they involve in the regulation of the plant hormone (ABA) signal transduction pathway (Zeevaar *et al.*, 2005). Protein Phosphatase 2C (K14497, EC:3.1.3.16) catalyse the cleavage of phosphate group from phosphorylated serine/theonine protein phosphatases (Arino *et al.*, 2011) and play an important regulatory role in stress signalling (Fuchs *et al.*, 2013). A PP2C inhibitory action regulates a downstream process of the carotenoid biosynthesis, one of the biochemical pathways included in the ABA signal transduction pathway, effects stomatal closure or seed dormancy as a consequences of gene expression.

The two genes were shown ubiquitously expressed across all the environmental stresses investigated in this study. They are grouped under the same class “Environmental Information Processing; Plant hormone signal transduction” and under protein family (Pfam) “PP2C and PP2C\_2”. They take part in the same signalling pathway with other eight sorghum genes: SORBI\_01g039890; SORBI\_02g022090; SORBI\_03g029890; SORBI\_03g032740; SORBI\_03g039630; SORBI\_06g001720; SORBI\_09g026860 and SORBI\_09g029080 and with nine common orthologs out of which protein phosphatase 2C represent 89%. In both cases, the first hits obtained using blast search among homologous proteins was originated from maize signifying it's closest ancestral relation to sorghum. However, Sb09g030600 is uniquely characterized by position on chromosome 9: 59,156,548 – 59,158,857, reverse stranded with shorter protein (400aa) and nucleotide (1203nt) sequences, two more (least-significant hit Pfams such as SpoIIE and DUF1378. On the other hand, Sb03g026070 is located on chromosome 3: 52,451,500 - 52,456,092 forward stranded with protein and nucleotide length of 482aa and 1449nt respectively. Both are hypothetical proteins not previously reported as drought responsive.

### 3.3.1.2 Phylogenetic relationship

A dendrogram showing phylogenetic relationship of the two sorghum genes 'Sb03g030600' and 'Sb09g026070' with homologs based on species with common ancestor is shown in Figure 3.3. These genes share a common ancestral position with Os05g0592800, a rice gene, involving in PP2C inhibited negative regulation of signal transduction pathway to control ABA signalling.



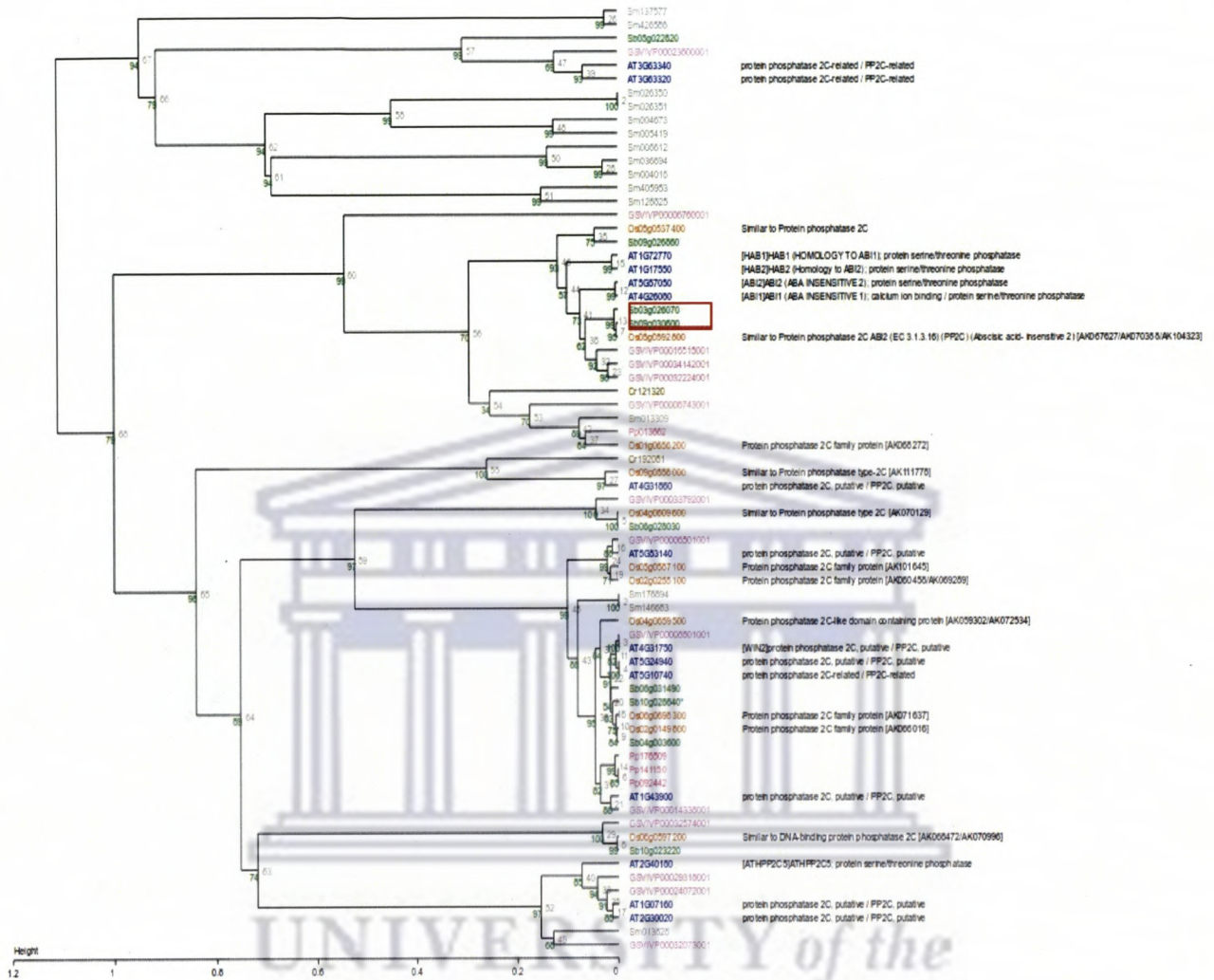


Figure 3.3: Phylogenetic relationship of the sorghum genes with orthologs.

This dendrogram was generated using SALAD, a tool of systematic comparison of proteome data among the species SALAD databases (Mihara *et al.*, 2010).

### 3.3.2 Comparative gene association across-species: Functional overlapping and specificity

Figure 3.4 shows species specific and shared gene locus and probably functional conservation in a closely and distantly related species of grass families. The total number of genes represented in sorghum, maize, rice, and Arabidopsis were 169, 138, 213 and 613 respectively. The representation of these genes in each species in this data was based on the relevant drought related terms in the EO, TO, PO, GRO and the GO. Potential drought tolerant genes in sorghum having shared functionality with closely related species were identified based on the putative functions of their orthologs in all related species under consideration.

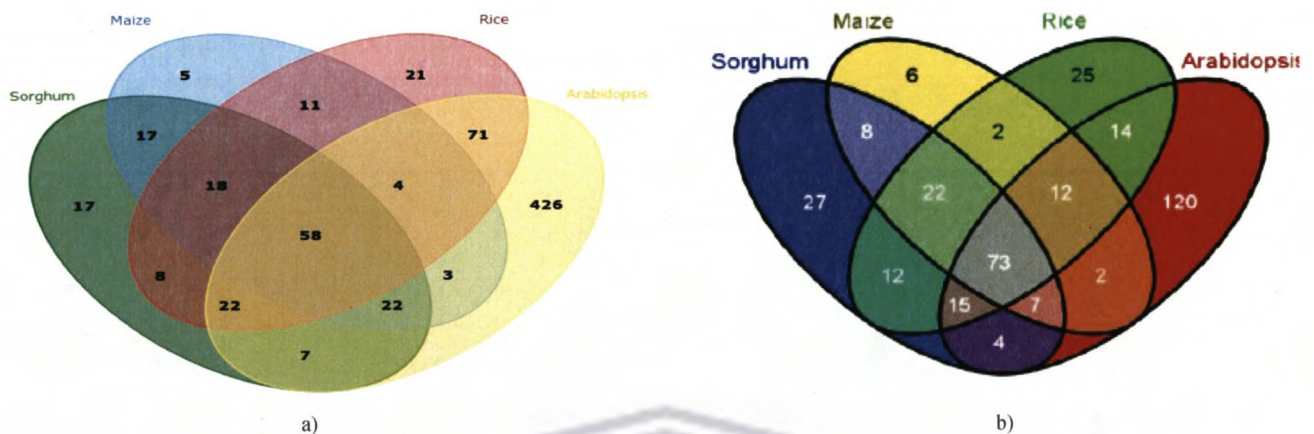


Figure 3.4: Species specific and common drought responsive genes in a closely and distantly related species.

This venn diagrama in Figure 2.4a shows a distribution of sorghum orthologous genes in the other three related species in association with drought related ontology terms based on data from gramene db, and in Figure 2.4b shows the distribution of sorghum orthologs in the other three species associated with drought related ontology terms based on sequence similarity search. The numbers displayed in venn diagram of this figure correspond to the number of genes. Superimposed region of all circles shows the number of genes shared in all four species. Overlapping regions between any three species indicate gene locus and functional conservation between the three of the four species and similarly the shared regions between any two species meant gene locus and functional conservation in the two of the four species. Parts that don't overlap between circles show unique drought responsive genes of each species.

### 3.3.3 Semantic integration of existing data based on functional ontology

Figure 3.5 shows the ontology mapping using semantic integration of existing sorghum perturbation related information where lists of potential candidate drought responsive genes were identified from their transitive association rather than directly from sorghum gene-trait association. Where our query for relevant terms in the different ontologies yields no existing information for sorghum, we opted to use transitive gene association to multiple traits through rice orthologs. Ontology mapping was used in functional validation of the 168 known genes which were previously putatively uncharacterised.

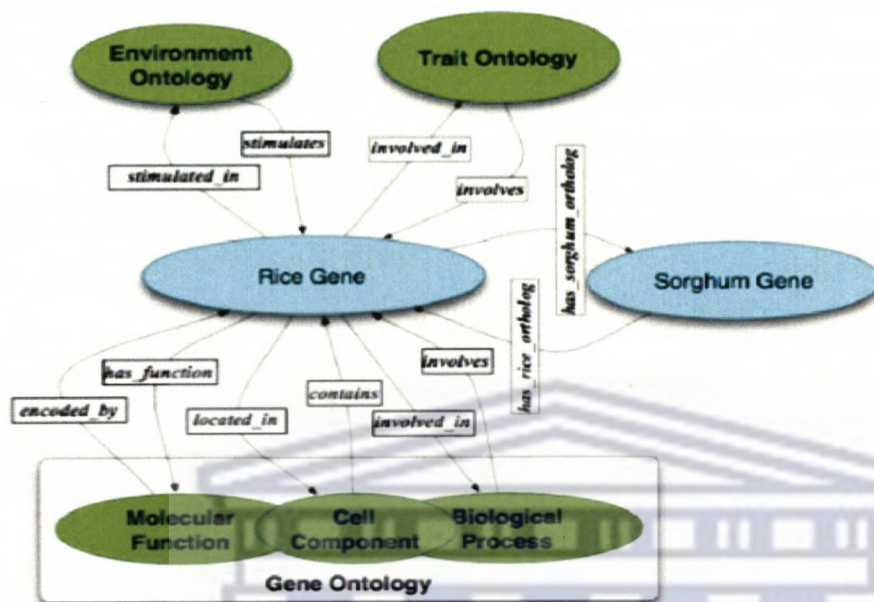


Figure 3.5: Sorghum genes transitive association to multiple drought related terms

This figure shows sorghum genes transitive association to multiple drought related terms based on functional ontology. Based on plant environment ontology, trait ontology and gene ontology information, sorghum orthologs in rice are associated with drought response. This shows that sorghum may share functionally similar genes with distantly related crop species, rice.

### 3.3.4 Integration of differential expression data set

Expression data was integrated with information from functional ontology and successful association drought responsive genes with phenotypes was demonstrated. Based on evaluation of tissue type contributing to gene expression, 46 significantly up-regulated genes were shown to have strong correlation with drought tolerance. On the other hand, evaluation of treatment effect revealed 42 significantly up-regulated genes under drought condition for which strong association from all plant attributes had been determined. This result shows the higher percentage of genes representation in tissue-specific expression under stress condition than with drought stimulation irrespective of tissue type. This result is concurrent with the published work on sorghum transcriptome analysis on 9 days seedlings in response to osmotic and abscisic acid stresses (Dugas *et al.*, 2011).

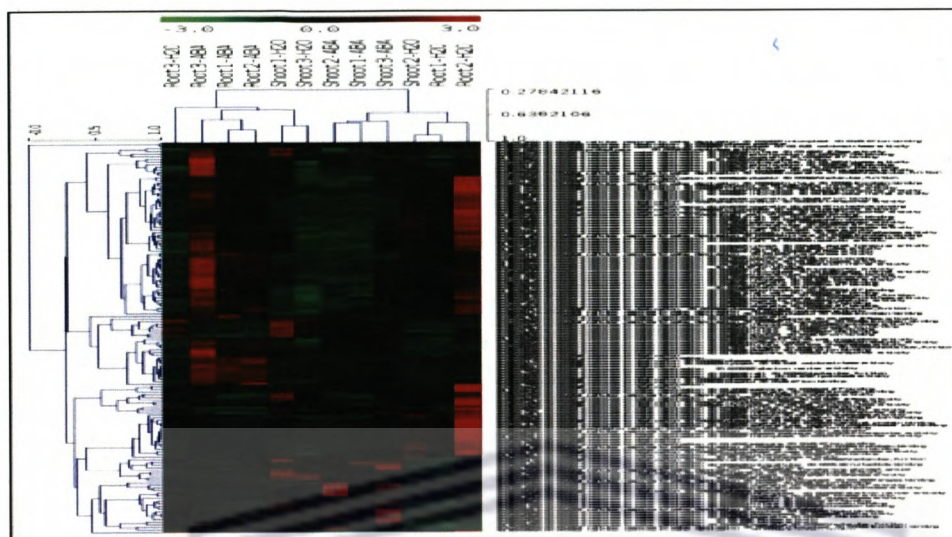


Figure 3.6: Hierarchical clustering of gene expression showing heatmap

The hierarchical clustering of gene expression profiling in Figure 3.6 is based on the information derived from the sorghum drought related ontology terms and the GEO database. Figure 3.6 shows heat map depicting up and down-regulated sorghum genes under drought condition based on data from sorghum RNA-seq in response to osmotic and abscisic acid stresses (Dugas *et al.*, 2011).

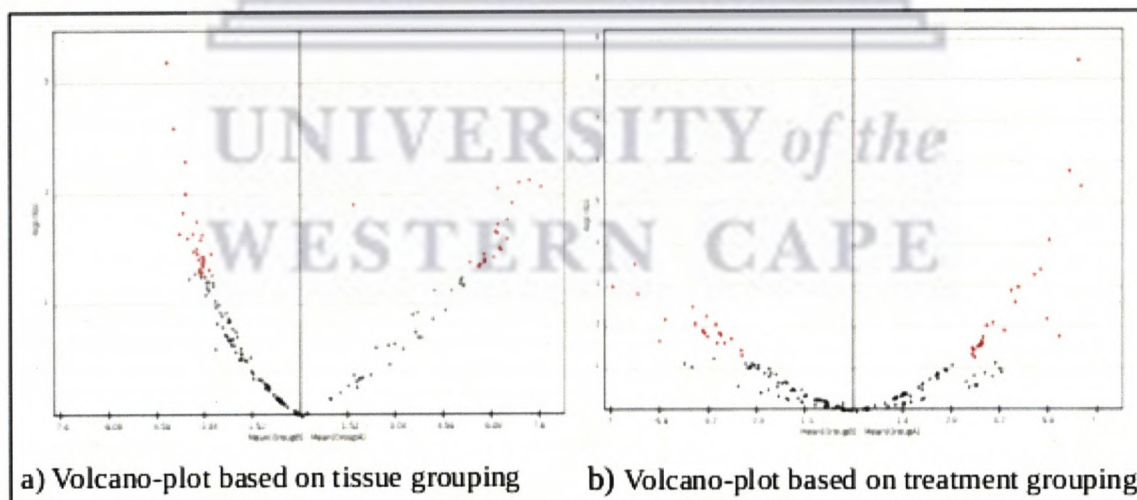


Figure 3.7: Volcano plots showing gene expression

This figure shows volcano plots with differential expression of genes (most significant genes at the top of the plot). Volcano plots (a) represent unpaired t-test based on evaluation of tissue type contributing to gene expression and (b) based on evaluation of treatment effect on experimental samples. The red dots indicate genes with statistically significant value for up and down-regulated genes and fold-changes above 2. Above this value all genes have  $p < 0.01$  and below it  $p > 0.01$ .

### 3.3.5 Functional-annotation and GO Enrichment

Based on the association of genes with drought related GO terms, 167, 148, 133 significantly enriched genes (Table 3.1; p-value, FDR < 0.05) were identified for all the three domain namely BP, CC and MF respectively. This was further screened to 126 non-redundant genes supported by all GO-domains. Similarly, using the same method 296, 1681, 98 and 712 enriched genes ( p-value, FDR < 0.05) which were shown more likely to have strong association to the plant attribute were filtered from TO, EO, PO and GRO respectively (see Table 3.2). This makes a total filtered non-redundant gene for association to be 2224 which were further screened down to 2118 enriched transcripts or 1820 genes.

Table 3.1 Gene-phenotype association based GO enrichment analysis

GOD <sup>1</sup>	GO-term	GO-ID	Genes	P-value	FDR	Trait association
BP	response to water deprivation	GO:0009414	138	1.80E-039	8.90E-036	Drought stress tolerance
BP	response to desiccation	GO:0009269	25	5.90E-008	3.70E-006	Drought stress tolerance
BP	response to osmotic stress	GO:0006970	65	1.50E-012	2.50E-010	Osmotic stress tolerance
BP	response to salt stress	GO:0009651	93	1.80E-009	2.10E-007	Salt stress tolerance
BP	response to heat	GO:0009408	56	3.30E-006	0.00019	Heat tolerance
BP	response to cold	GO:0009409	138	7.40E-021	3.70E-018	Cold tolerance
BP	response to oxidative stress	GO:0006979	95	9.00E-008	7.00E-006	Oxidative stress tolerance
BP	response to ROS	GO:0000302	43	5.00E-007	40E-005	Oxidative stress tolerance
BP	oxidation reduction	GO:0055114	62	5.80E-007	3.80E-005	Drought tolerance
CC	plastid	GO:0009536	294	1.60E-015	9.20E-014	Drought stress tolerance
CC	chloroplast	GO:0009507	257	6.50E-014	3.40E-012	Drought stress tolerance
CC	chloroplast thylakoid	GO:0009534	96	2.90E-013	1.50E-011	Drought stress tolerance
CC	thylakoid	GO:0009579	103	2.80E-012	1.20E-010	Drought stress tolerance
CC	chloroplast stroma	GO:0009570	37	3.00E-008	1.20E-006	Drought stress tolerance
MF	oxidoreductase activity	GO:0016491	285	3.80E-009	8.30E-007	Drought stress tolerance
MF	protein binding	GO:0005515	676	3.90E-007	3.30E-005	Drought stress tolerance
MF	water channel activity	GO:0015250	14	9.40E-005	0.0049	Drought stress tolerance

Key to legend: <sup>1</sup>GO-domain. This table gives description of the GO enrichment for drought responsive genes identified in this analysis (P-value, FDR < 0.05).

We have shown a graphical representation of significantly enriched GO-terms assigned to the identified genes that demonstrated strong association with drought-responses based on biological processes, cellular components and molecular functions (Figure S3.1). This representation was demonstrated using enrichment graphical views (Figure S3.1), GO annotation (Figure S3.2), biological network (Figure S3.3) and scatter plots (Figure 3.8).

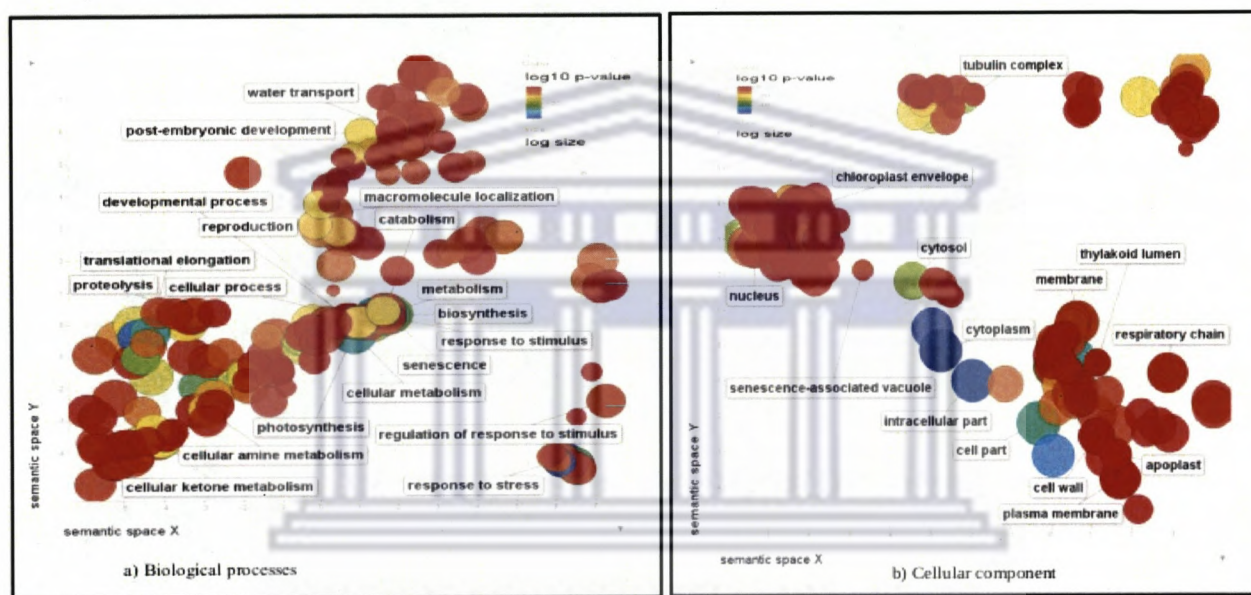


Figure 3.8: Scatter plot for semantic similarities in enriched GO-terms.

This scatter plot that was generated using multidimensional scaling shows semantic similarities in the enriched and non-redundant drought-related GO-terms association to a set of genes. As multidimensional scaling provides an option of using an eigenvalue of the GO-terms' pairwise distance matrix, the coordinate position of the GO-terms' semantic similarities in the enriched genes will be displayed to the two-dimensional spaces (Figure 3.8; Supek *et al.*, 2011; Li *et al.*, 2013).

Biological networks of gene association for which enriched GO-terms exist can be shown by using interactive biological networks (Martin *et al.*, 2004; Eden *et al.*, 2009; Supek *et al.*, 2011) based on all deterministic factors attributed to the three GO-domains (Daraselia *et al.*, 2007; Gruber, 2009).

Table 3.2 Functional association of genes with different ontologies terms

No.	Ontologies	Ontology Terms	Ontology accessions	Number of identified		Screened Genes universally supported
				Genes	QTLs	
1	Gene (GO)	Biological process (BP)	GO:0009414	167	-	126
		Cellular_component (CC)	GO:0005575	148	-	
		Molecular function (MF)	GO:0003674	133	-	
2	Trait Ontology (TO)	drought tolerance (BP)	GO:0009414,GO:0009819	150	-	296
		drought tolerance (TO)	TO:0000276	-	25	
		drought susceptibility index	TO:0000155	-	4	
		total biomass yield	TO:0000457	-	141	
		leaf rolling time	TO:0000503	-	28	
		leaf rolling tolerance	TO:0002662	-	5	
		deep root dry weight	TO:0000081	-	25	
		plant dry weight	TO:0000352-	-	6	
		chlorophyll content	TO:0000495	12	74	
		stay green trait	TO:0002712	2	-	
		biochemical trait	TO:0000277	2	-	
		leaf senescence (BP)	TO:0000249, GO:0010150	132	-	
		growth & development trait	TO:0000357	2	-	
		3	Environment Ontology (EO)	drought environment	EO:0007404	
sodium chloride regimen	EO:0007048			1193	-	
salt regimen	EO:0007185			398	-	
watering regimen	EO:0007383			2406	-	
Cold temperature regimen	EO:0007174			1372	-	
4	Plant structure ontology (PO)	Inflorescence	PO:0009049	10232	-	98
		tassel inflorescence	PO:0020126	-	20	
5	Growth (GRO)	reproductive stage	GRO:0007140	2803	-	712
		seedling stage	GRO:0007047	9088	-	
		Booting stage	GRO:0007148	286	-	
		early-booting stage	GRO:0007149	1949	-	
		Late-booting stage	GRO:0007150	1	-	
		Flowering Stage	GRO:0007151	6497	-	
		Heading stage	GRO:0007044	6454	-	
Total	5	27	27	11987	328	2224
unique						

This table shows the number of genes associated with drought related ontology terms identified at different stages based on step-wise screening procedure. Note: in cases of TO and PO the results indicate both genes and QTLs in correlation with the ontology terms based on transitive association of sorghum genes to their gramene species particularly of maize, rice and Arabidopsis genes. However, we didn't include QTLs in this particular analysis for consistency reason.

Figure 3.9 shows summarized result description of drought related gene-trait associations based on functional ontology enrichment analysis.

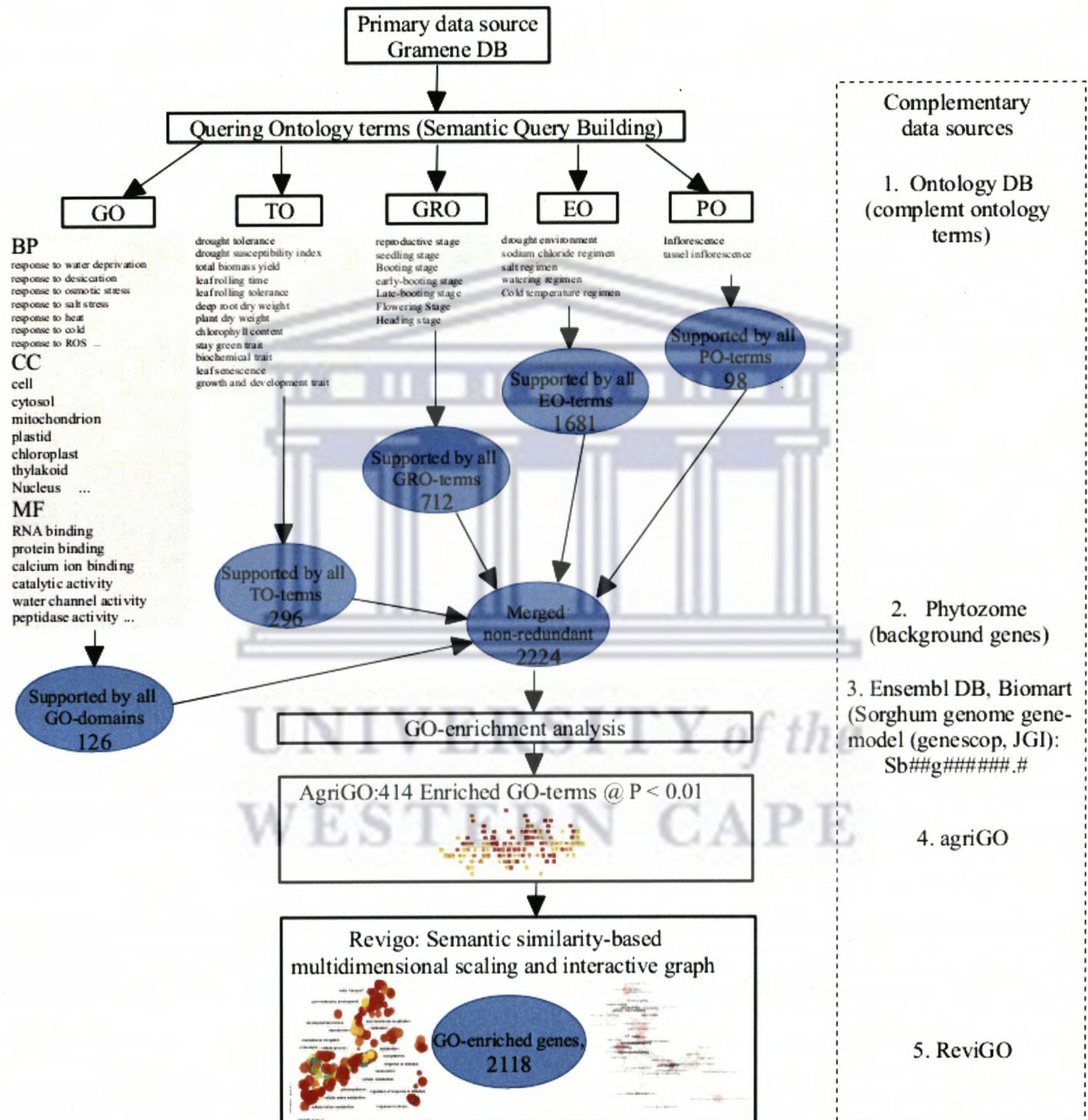


Figure 3.9: Summarized description of drought related gene-trait associations

This figure summarizes the drought related gene-trait associations including other basic functional ontology terms in sorghum plant and GO-enrichment visualization.



### 3.4 Discussion

Understanding the genetic basis of complex traits has long been a challenge because of the complexity in stage and development specific physio-biochemical processes at cellular and whole-plant level (Tripathy *et al.* 2000; Rampino *et al.*, 2006; Farooq *et al.*, 2009). However recent advances in studies have shown that this challenge is tractable and within reach of functional genomics and association studies (Fleury *et al.*, 2010; Langridge and Fleury, 2011; Varshney *et al.*, 2011; Mir *et al.*, 2012; Swamy and Kumar, 2013; and Varshney *et al.* 2014). Identification of genes associated with multiple environmental stresses and their functional correlation across species as a bigger interest in the current study was successfully demonstrated generating genes associated with drought tolerance using integrated and the most efficient and promising approach. In this study, results have shown that genes with more functionally ubiquitous across species and stresses have been evidenced.

#### 3.4.1 Association to multi-environmental-stress tolerance

Multiple responses of genes across environmental stresses is the genetic foundation of plant adaptation to environmental heterogeneity. The initially identified 169 sorghum drought responsive genes has been shown to be up regulated under several stress conditions. About 34% of these genes, were found to be responsible for the defence and tolerance responses in multiple environmental stresses. Further more, about 11% of the genes were detected to be salt and cold responsive. Interestingly, 2 genes (Sb03g026070 and Sb09g030600) were responsive to all stresses except to ROS. These two functionally ubiquitous genes may be considered as the rate limiting factors in sorghum drought tolerance based on the fact that their involvement in the “Plant hormone signal transduction” pathways plays significant inhibitory regulation (Leung and Giraudat, 1999). The basis of involvement in ABA pathway regulation of these genes is mainly associated with their function to encode for protein phosphatase 2C (PP2C; EC:3.1.3.16; K14497), an enzyme involved in the negative regulation of the ABA signalling transduction pathway and mediate stomatal closure or seed dormancy (Park *et al.*, 2009; Umezawa *et al.*, 2009; Ma *et al.*, 2009 and Nishimura *et al.*, 2010). The involvement of genes in ABA signalling modulation observed in this study agrees with a study shown to enhance drought tolerance over multiple stresses using transformed maize ABP9 gene (Zhang *et al.*, 2011) in Arabidopsis. These genes and more others were shown to be over-expressed across stresses and were noted as vital elements in advancing tolerance to multiple

environmental stresses in sorghum. When genes happen to be expressed under multiple environmental stresses, they are regulated and acted upon complex biological processes. Such genes would on one hand be involved in several pathways that network and interact with many other genes and on the other represent quantitative expression dynamics under varying experimental stress conditions (Breitling *et al.*, 2005). This gives an impression of how genes might be controlled in the plant pathways during simultaneous exposure to different stresses (Kilian *et al.*, 2007).

These phenomena build a ground for physio-biochemical and molecular mechanistic function of a gene across environmental heterogeneity where respective stresses are prevailing. If the plant varieties with such a set of genes are predominant in the population, then the resultant effect is equivalent to ubiquitous adaptation and productivity. This represent a fundamental element both in agricultural and evolutionary biology. This study reports a multi-environmental stress tolerant genes in sorghum which were previously ascribed only as hypothetical proteins.

#### **3.4.2 Resistance from whole-plant to individual level components**

This approach, not only combines functional ontology based semantic data analysis with association study, expression profiling and biological networks, but also strives to resolve the whole-plant resistance into individual components through identification of functionally enriched drought expressed genes which were associated to predetermined key drought relevant ontology terms. Based on other findings, drought resistance (DR) can be broken down into several component parts (Yue, 2006). First, drought tolerance (DT) which is the focus of this study, is primarily manifested via osmotic adjustment (OA), antioxidant capacity and desiccation tolerance. Here 65 genes were identified to be responsive to OA, 282 genes were responsive to ROS, and 22 genes were responsive to desiccation tolerance. In addition to morphological, physiological and biochemical responses (Anjum *et al.*, 2011), accumulation and translational of assimilates (Ji *et al.*, 2012) and maintenance of cell wall elasticity (Bartlett *et al.*, 2012a and 2012b; Ajithkumar and Panneerselvam, 2012; Sanchez *et al.*, 2012 and Scholz *et al.*, 2012) also define DT. Secondly, Drought Avoidance (DA) is characterized by enhanced water uptake and reduced water loss. Nineteen genes with cellular responses to water deprivation and 126 genes with physical response to water deprivation were identified. The third and the last DR component is drought escape (DE) which is characterized by a short life cycle or developmental plasticity of the plant. This study

identified 1949 genes responsible for DE which were filtered to a few genes associated with early booting and only one gene “Sb03g003110”, a rice ortholog (BGIOSGA002217) with late booting (Yaqoob *et al.*, 2012; Ribot *et al.*, 2012 and Lopes, *et al.*, 2014). Based on this, it is apparent that our results concur with other findings such as Yue, 2006. Most of the genes identified satisfy drought resistance criteria. The results also suggest that our approach in dissecting the complex polygenic traits into particular elements of plant DR is holistic in nature and promising.

### **3.4.3 Understanding gene-phenotype-association through integration of functional ontologies**

A high genetic diversity and rich functionality of sorghum that engage gene association with important and complex traits gave foundation for its well adaptation to adverse environments. In this study, it was very interesting to have initially identified 1681 genes (75.5%) based on drought stress related environmental regimens that were commonly enriched by all EO terms. This was an indication that sorghum is a potential source of drought tolerance. An investigation of such a huge DRGs using a novel approach that accommodates different plant species closely related to sorghum implicates towards improving crops for drought tolerance. Our analysis shows that this approach is effective in examining an interoperability of the ontologies which were not functionally overlapping. It also valued an interrelationship of the plant traits with all other plant attributes including the environment where the plant is normally adapted. This concept is basically linked to an important aspect of plant life. The extent of adaptability, survival and yield production is an empirical association to the genetic make up of the plant itself and the optimal values of all the plant attributes. These include traits such as chlorophyll content, stomatal closure, morphological and anatomical structural fitness of the plant, early or late maturing. All the biological functioning that include the fundamental biological processes, molecular functions and the cellular components are all fundamental parts of plant life. This suggests, on the one hand, the need for realizing the plant tendency for such a balanced biological functioning to depend critically on specific environment (Jones, 2007) and on the other hand, the need for understanding physiological and molecular basis of complex trait such as drought tolerance.

### **3.4.4. Cross-species functional crosstalk**

Our analysis based on 169 identified sorghum DRGs across species indicated 90% that exhibit drought tolerance in multiple species without being sorghum specific. Analysis of sorghum genes functional correlation with its orthologs in other species showed that 11% were shared only with

maize, nearly 5% with rice and Arabidopsis each. Still, 12% of sorghum genes were shared with maize and rice in common and 15% with rice and Arabidopsis. Moreover, 34% of the total sorghum genes were commonly shared by all three species. Cross-species gene association suggests existence of homologous groups, that descends from a common ancestral gene (Fitch, 1970 and 2000; Putnam *et al.*, 2007). This entails an evolutionary proximity of sorghum to these species and the conservation of specific genomic regions across species with some extent of similarity in functional association to drought tolerance. Orthologs unlike paralogs which evolve to functional diversification (Lynch and Katju, 2004) typically occupy the same functional niche in different organisms (Makarova *et al.*, ). Orthology is related to conserved structural elements or conserved neighbourhood (Arnesano *et al.*, 2005; Bandyopadhyay *et al.*, 2006) in ancestrally related species. However, based on evolutionary definition (Snel *et al.*, 2002), one orthologous group often contain different functions. Based on KEGG definition (Kanehisa and Goto, 2000), unless all constituent members such as a conserved sub-pathway or a molecular complex are fulfilled, sequence similarity only may not represent a functional group. This suggests further investigation to determine the functional association of orthologous group and to define conserved gene order across species. The presence of 10% sorghum specific genes implicates a uniqueness of sorghum compared to all other species suggesting its distinct position in phylogenetic order. This may further suggest the presence of unique genes encoded by the sorghum genome which have been selectively structurally and functionally evolved and have developed sorghum-specific plasticity in response to changes in environmental conditions and more specifically to drought and related environmental stresses.

#### **3.4.5 Deciphering drought stress tolerance through integration of semantic knowledge**

Functional ontology has long been an instrumental for genetic deciphering of complex traits such as drought stress tolerance through semantic knowledge. The ontology mapping produced in this study is based on semantic integration of existing sorghum perturbation related information from the functional ontology points of view. The ontology mapping was an implication of potential candidate genes resulted in association with drought stress tolerance. Added to its validative role, ontology mapping gained advantage from transitive association based on orthologs. Transitive association of sorghum orthologs with drought related ontology terms complements data from sorghum genes to make sufficient association with multiple drought-related terms in several ontologies. In the current

analysis, of the total genes expressed association across all ontology terms, at least 50% had transitive association (Li *et al.*, 2003; Vinayagam *et al.*, 2004 and Mungall *et al.*, 2010). This suggests that associations among the sorghum genes and drought terms for their orthologous counterparts is important for discerning genetic dissection of complex drought tolerance.

#### **3.4.6 Association of expression profiling with drought phenotypes**

A number of other studies have been approached using expression data in combination with text information from several areas but not limited to quantitative genetics (Narain, 2010); Molecular breeding (Spence *et al.*, 2005; Cattivelli *et al.*, 2008) and biomedical research (Bailey and Ulch, 2004; Tiffin *et al.*, 2005 and Pillitteri *et al.*, 2011). Integration of expression data with information from functional ontology related to complex drought tolerance successfully identified association of relevant drought responsive genes with phenotypes based on their expression status under drought environments. Among 169 genes tested for significant expression, 52% showed strong association to drought tolerance. In this category, 48% exhibited strong association under differential conditions irrespective of tissues specificity while the remaining 52%, showed association with different specific tissues. This shows the value of expression profiling in segregating genes based on their attributed association and in complementing other strategies in drought tolerance research.

UNIVERSITY of the  
WESTERN CAPE

### 3.5 Conclusion

Analysis of gene association with drought phenotype using multiple environmental variants revealed promising results in drought tolerance. In this study we were able to show genetic dissection of complex traits by comparing gene association across-environmental stresses and species and by integrating ontology based semantic data with expression profiling. Evaluation of the results using multivariate analysis provided significant array of genes associated with drought tolerance particularly 46 significantly up-regulated tissue specific genes and 42 significantly up-regulated genes irrespective of tissue specificity under drought condition. Ontology mapping played a validating role for all identified putatively uncharacterised 1820 genes. Our approach adds on to the existing efforts by providing researchers with unique integrative data analysis systems towards genetic dissection of complex poligenic traits. The results may have profound implications in comparative study of major cereal crops and in breeding programs towards improving drought tolerance in sorghum.



## CHAPTER 4

### **Chapter 4: Identification of Drought Responsive Proteins in Sorghum (*Sorghum bicolor* (L.) Moench) using Differential Expression Profiling and MALDI-TOF-TOF MS/MS**

#### **Abstract**

**Background:** Drought is a major threat to the world food production affecting plant growth and productivity by causing plant metabolic and photosynthetic impairments. Understanding how drought stress alters the normal physiological and biochemical functions provides a means for enhancing drought tolerance and crop productivity. Sorghum (*Sorghum bicolor* (L.) Moench), a dry-land adapted cereal crop is beginning to be useful model for genomic and proteomic research towards developing drought tolerance varieties. Here we report results on a proteomic analysis of drought response in sorghum.

**Methodology:** Sorghum seed varieties obtained from the International Crop Research Institute for Semi-Arid Tropics (ICRISAT), India, Delhi, were used in this research to identify drought responsive proteins. Control and experimental plants were grown in a greenhouse (ARC, Stellenbosch) under normal and drought conditions. Post-flowering drought stress was induced on the onset of flowering and samples were harvested at the  $30\% \pm 5$  FC. The samples were immediately stored at  $-80^{\circ}\text{C}$  @ the University of the Western Cape (UWC), Biotechnology Department, NAPRSU until use for protein identification. Trichloroacetic acid (TCA)/acetone and Bradford assay methods were used to extract and quantify proteins respectively from sorghum leaf tissue (Btx642 variety) for this analysis. After conducting protein separation using two-dimensional gel electrophoresis (2DE), Coomassie Brilliant Blue (CBB) stained gels were scanned using Molecular Imager PharoFX Plus System (BIO-RAD). Spot detection and matching and analysis of differential expression pattern was performed using PDQuest™ software (Bio-Rad) version 8.0.1. Sixteen spots of interest based the intensity or abundance and resolution were selected for protein identification using MALDI-TOF-TOF MS/MS and searching database using MASCOT.

**Result:** This study identified nine protein enzymes from seven spots resulted significant score out of sixteen selected for mass spectrometry analysis. Our result shows two spots (12 and 14) which were mixture of two different proteins each. Five functional categories of proteins were identified.

These are 1) energy generating or proton (H<sup>+</sup>) transporting related protein (11.1%); 2) Glycolysis and gluconeogenesis and other carbohydrate metabolism associated proteins (22%); 3) Photosystem regulation (carbon assimilation, 33%); 4) Stress tolerance, defence and immunity related proteins (11.1%); 5) RNA binding proteins (11.1%); and 6) Unknown (11.1%). In addition, three different classes of subcellular localization were identified where 78% of the proteins positioned in chloroplasts suggesting the photosynthetic role under drought stress. Of the identified protein 77.7% were found to be significantly expressed (up-regulated). This suggests the role of proteins in drought tolerance. However, this study also shows one typical mechanism where plants induce signal transduction alarm to bypass stress condition by down regulating a rate limiting enzyme.

**Conclusion:** This result demonstrates novel functions of the proteins in sorghum describing their central role in maintaining a normal functioning of metabolic and photosynthetic pathways under drought stress. A functional correlation that was depicted between some of the key protein enzymes experimentally identified and the others *in silico* generated proves novelty and validity bridging the gap between genomic and proteomic research. The data presented in this study forms a useful resource as a reference for future research.



UNIVERSITY of the  
WESTERN CAPE



## 4.1 Introduction

Drought stress is a major constraint to the world production and is tremendously hampered sorghum [*Sorghum bicolor* (L.) Moench] productivity. With a recurrent production decline over time throughout arid and semiarid regions (Hatfield *et al.*, 2011), a yearly yield loss caused by drought was estimated between 9.3 and 15.5% (Sultan *et al.*, 2013). Drought is a regular phenomenon in most African climate due to a shortage of precipitation over an extended period (Herrmann *et al.*, 2005 and Bola *et al.*, 2014). However, sorghum (*Sorghum bicolor* (L.) Moench) among rare hardy crops is the most striving to drought affected African and global regions (Nagaraj and Rao, 2011; Rao *et al.*, 2011). Several studies have investigated the mechanisms that underling drought resistance in sorghum. These include but not limited to C4 photosynthetic pathway evolution (Ripley *et al.*, 2007 and Ghannoum, 2009), physio-biochemical processes (Lay and Anderson, 2005; Pagariya *et al.*, 2011) and anatomical structure (Nguyen *et al.*, 2004; Harris *et al.*, 2007; Xin and Wang, 2011) that contributed to sorghum's unique adaptation to drier and hotter regions.

Sorghum is the model for comparative genomics standing as fifth most important cereal crop worldwide based on production scale grown in rain fed lowland and semi-arid tropics with remarkable tolerance to adverse conditions (Subodhi, 2011). Sorghum, on top of its achievements in the possible enhancements through the use of pan-grass tools and information, it has the potential to become a model system for understanding C<sub>4</sub> plants and other members of the tribe Andropogonae (Wang *et al.*, 2009). With a relatively small genome of the haploid size ca. 760 Mbp ranking second only to rice (420Mbp) among the major crops in the Poaceae family, and expected smaller distance between individual genes than its large genome relative, maize (~2922-6171Mbp), sorghum can be a model for plant genetics and genetic engineering (Bennetzen, 2000; Paterson, 2008). It has been demonstrated that sorghum shares fundamental drought tolerance pathway with maize. However, because of its greater adaptation to drought-affected areas, sorghum evolved superior genes in that pathway which may also function to provide tolerance in maize as well (Bennetzen, 2000). This entails that sorghum is used as a key plant species in comparative analysis for grass genomes, and as a source of beneficial genes for agriculture.

Drought-stress is manifested in sorghum at both pre-flowering and post-flowering stages resulting in a drastic reduction in grain yield. The former occurs when plants are under significant moisture

stress particularly from panicle differentiation until flowering and the later during the grain filling where lodging further results in total loss of crop yield (Kebede et al., 2001). In contrary, stay-green forms resistance mechanism determining sorghum responses to terminal drought stress. Several quantitative trait loci (QTL) for stay-green and terminal drought tolerance have been identified in improved sorghum varieties. Stay-green is a drought tolerance trait in sorghum that gives plants resistance to premature senescence under severe soil moisture stress during the post-flowering stage (Xu *et al.*, 2000; Reddy *et al.*, 2014).

One of the genotypes identified to offer a post-flowering drought resistance in sorghum with stay-green gene is Btx642, formerly known as B35, a back cross derivative of a dura landrace cultivar (Walulu et al., 1994). QTLs from this and other genotypes have been successfully used to improve lodging resistance with positive association with grain yield and reduce post-flowering drought-induced leaf senescence to provide post flowering drought resistance under water limited environments (Borrell et al., 2000; Harris et al., 2007; Reddy *et al.*, 2014). In addition, sorghum has a relatively high capacity for osmotic adjustment (OA) under drought stress with especial characteristics to display diversity for OA (Babita et al., 2010; Liu et al., 2014). Different morphological and physiological mechanisms associated with drought tolerance including stomatal regulation, variation in leaf cuticle thickness, root morphology and many others have been investigated in sorghum (Pathan *et al.*, 2004). Such unique attributions in sorghum entails the presence of important genes and useful alleles enabling the crop to survive in arid environments (Dogget, 1988; Holden and Peacock, 1993 and Collins et al., 2008).

There is a significant genetic variation for drought tolerance in sorghum (Ayana and Bekele, 1998; Abdi *et al.*, 2002; Paterson *et al.*, 2009; Jordan *et al.*, 2012). However, attempts to exploit this resource through conventional plant breeding methods have been slow and arduous. Conventional breeding method only allow inferences to plant performance such as yield or secondary traits associated with yield (e.g., anthesis-silking interval in maize (Ribaut *et al.*, 1996) or stay green in sorghum (Crasta *et al.*, 1999; Jordan *et al.*, 2012).

Several studies have been conducted to enhance sorghum productivity based on genomic and transcriptomic analysis (Tuberosa and Salvi, 2006; Manners and Casu, 2011) and to identify drought tolerance genes. However, such efforts are relatively quite minimal with respect to proteomic analysis particularly with reference to complex drought tolerance in sorghum. Recently,

studies have been initiated that target drought stress and salinity using proteomic techniques such as MALDI-TOF-TOF and Mass Spectrum (MS) using sorghum as a target organism (Ndimba *et al.*, 2010; Ngara and Ndimba, 2011; Ngara *et al.*, 2012; Ndimba and Ngara, 2013). Proteome analysis is most powerful in targeting candidate proteins and dissecting the genetic foundation of drought tolerance and are likely to be considered validative (Sharma *et al.*, 2013). One important evidence for that is the determination of gene expression by post-translation modification (Jorrín-Novo and Maldonado, 2009 and Pang *et al.*, 2010).

It is imperative to develop strategies to harness existing and emerging sciences to exploit the inherent potential of the crop to reduce or limit drought stresses and to promote food and economic security. Proteomics provide new approaches that may allow relatively rapid progress in producing crops with improved drought tolerance (Timperio *et al.*, 2008). Advances in proteomics with the wider application of bioinformatic tools widen research horizons in agriculture. This creates opportunities to investigate major proteins that involve in drought tolerance (Paterson, 2008) and in developing crop varieties with traits of interest (Bennetzen, 2000; Ndimba and Ngara, 2013). It provides strategy that complements functional genomics including microarray-based expression profiles (Shoemaker and Linsley, 2002; Sharma *et al.*, 2013) and systematic phenotypic profiles at the cell/tissue and organism level (Giaever *et al.*, 2002; Ngara and Ndimba, 2011). Identification of drought responsive proteins suggests high significance in the development of potential biomarker. Therefore, proteomics data sets serve as a powerful reference of protein properties and functions, which will also be useful both in building and testing hypotheses towards drought tolerance (Tyers & Mann, 2003). Hence, this study investigates proteins that are differentially expressed in sorghum inbred lines under drought stresses by implementing the software PDQuest and MALDI-TOF MS/MS for protein spots identification. The result will be cross-referenced with non-redundant protein databases to identify significant scoring match.

## **4.2 Materials and Methods**

### **4.2.1 Plant material**

This experiment was conducted in the greenhouse of the Agricultural Research Council (ARC), Stellenbosch, Cape town, South Africa in 2012 and the wet lab experiment in the National Agricultural Proteomic Research and Service Unit (NAPRSU), Biotechnology Department,

University of the Western Cape (UWC) in the academic year of 2013. Three sorghum seeds (Btx642, Tx7000A, Tx7000B) were obtained from the International Crop Research Institute for Semi-Arid Tropics (ICRISAT), India, Delhi and one (E36-1) from the Ethiopian Institute of Agricultural Research (EIAR, through a PhD student, Ato Gemechu Keneni). Sorghum lines listed above were initially used to rate the pre-and post-flowering drought tolerance however here we only report the analysis of protein identification based on Btx642 post-flowering drought tolerance. Btx642 also formerly called B35, is a derivative of IS12555, a dura landrace cultivar from Ethiopia (Walulu *et al.*, 1994) and is post-flowering drought resistant expressing stay-green phenotype under moisture stress condition at the grain filling stage, it however is susceptible to pre-flowering drought stress (Rosenow *et al.*, 1996; Tuinstra *et al.*, 1998).

#### **4.2.2 Experiment and growth conditions**

This experiment was conducted to evaluate and identify drought responsive proteins differentially expressed at the post flowering stage. Sorghum seeds were disinfected using sodium hypochlorite 5% solution, washed thoroughly with distilled water prior to sowing. Sorghum seeds were sown and germinated on a plastic tray using sandy soil for 25 days until the vigorous seedlings were selectively transferred to the plastic pot. Mixed with water, nutrient solution was given for seedlings until they were transferred. Once they were transferred, plants were grown for treatments in a 25cm diameter and 19cm depth plastic pots (Figure 4.1) with a  $(4.75 \pm 0.25)$  kg compost soil, without being additionally fertilized.

The experiment consisted of two treatments: (1) well watered, a 99% of maximum water Field Capacity (FC), the amount of water content retained in the soil after excess has been drained away and no more water movement downward is taking place (Veihmeyer and Hendrickson, 1931 and Klute, 2003) both at pre-flowering and post-flowering stages; (2) well-watered at the pre-flowering and drought-stressed ( $30\% \pm 5$ ; FC) at the post-flowering. Three replicate plants were grown in the greenhouse and leaf tissues from Btx642 were sampled both from control and experimental plants.

The greenhouse experiment was conducted over approximately three to four months ( $114.5 \pm 18.5$ d). Each pot contained one plant (Figure 4.1). The plants were watered with a tap water (from Stellenbosch municipality, Cape town, South Africa) and replenished with compost nutrients.

Watering was administered at daily basis twice during relatively hottest time of a day until the start of stressing by withholding water. Green house was facilitated with controlled temperature at an average of  $(28^{\circ}\text{C}\pm 3)$  during the time of plant growth. Drought stressing was initiated before anthesis on the beginning of flowering. Plant leave samples in treatment 1 and 2 were collected when the FC for plants in treatment 2 attained about 30% ( $30\pm 5$ ) and ready for sampling.

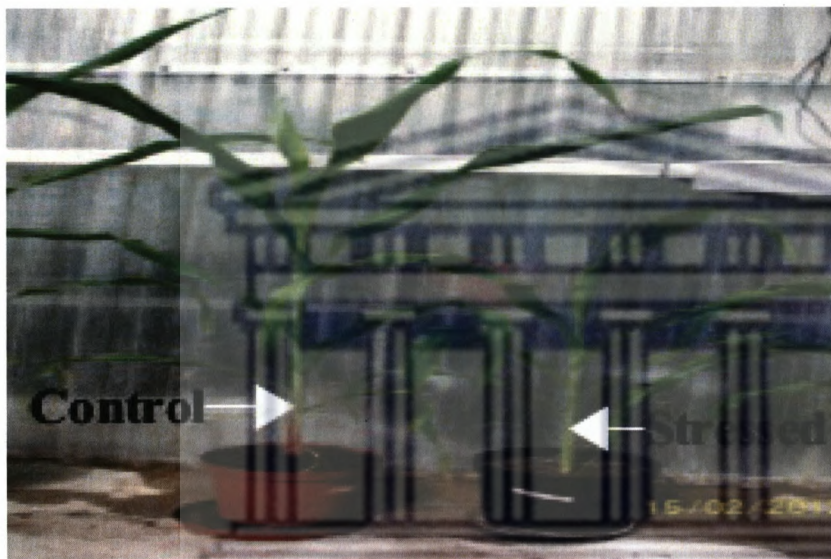


Figure 4.1: Sorghum sample control and stressed plants grown in green house conditions

This shows sorghum sample control and stressed plants treated in the green house condition both with the same sowing date. Except for the difference in the water level in the soil within the pot, all the other conditions ideally remained the same. However, certain unanticipated differences from external factors that we could not avoid might still be expected to exist.

#### 4.2.3 Protein Optimization and Extraction

The protein precipitation and extraction was performed by testing the precipitation of protein sample with 10% TCA alone and with 10% TCA in acetone separately and the later was found to be effective. The protein optimization and extraction for sorghum Btx642 leaf tissue was performed as stated below. Table 4.1 shows the list of all the chemical solutions and buffers systems used throughout this study stating in accordance to their alphabetical order and indicating the respective storage condition with other description.

Table 4.1 Solutions and buffers systems including storage condition and other description

Chemicals	% solution, composition and used volume	Type of solution	Description	storage temperature
Acetone	80% (v/v). i.e. 80ml of acetone in 20ml of ddH <sub>2</sub> O	Rinsing or washing solution	in deionized (distilled) water (place in -20°C atleast an hour before applying if stored under 4°C)	4°C / -20°C
Acetonitrile	100µl 100% acetonitrile	Washing solution	Cover the gel piece completely to allow shrinking of the gel piece until turns white	4 °C
Acetonitrile	10µl 100% ACN	wetting solution	to wet the Zip-tip Pipettor tip column and to remove organic impurities	4 °C
Agarose	0.5% (w/v)	sealing solution	prepared in 1X SDS-PAGE running buffer with a tiny bromophenol blue for sealing the IPG strips for SDS-PAGE analysis. Used in gel electrophoresis to separate protein by charge (pI); MW and size.	4°C
Ammonium bicarbonate	500µl of 50 mM Ammonium bicarbonate	Washing solution	removes any impurities from the gel pice of the protein spot	4 °C
Ammonium bicarbonate in acetonitrile	500µl of 50 % 50 mM Ammonium bicarbonate/50 % acetonitrile	destaining solution	removes staining color from the gel piece of the protein spot	4 °C
Ammonium Persulfate (APS)	10% (w/v)	oxidizing agent (fresh)	in distilled water (used fresh)	
Bradford reagent: BIO-RAD Protein Assay dye	1:4	stock (reagent):	diluted with deionized water;	4 °C
Bovine Serum Albumin (BSA) Standards	5 mg/ml	stock solution reagent	in urea buffer	4 °C
Coomassie Brilliant Blue (CBB)	1.25% (w/v)	stock solution	CBB R-250 in deionized water	4 °C
CBB stain I	50 ml of 1.25% (w/v) CBB stock solution, 10% (v/v) glacial acetic acid and 25% (v/v) propan-2-ol	staining solution I	in distilled water.	Room temperature
CBB stain II	6.25 ml of 1.25% (w/v) CBB stock solution, 10% (v/v) glacial acetic acid and 10% (v/v) propan-2-ol	staining solution II	in distilled water.	Room temperature
CBB stain III	6.25 ml of 1.25% (w/v) CBB stock solution and 10% (v/v) glacial acetic acid	staining solution III	in distilled water.	Room temperature
Destain	10% (v/v) acetic acid and 1% (v/v) glycerol	Destaining solution	N/A	Room temperature
DTT	50% (w/v)			4 °C,
Equilibration buffer	6 M urea, 2% SDS, 0.05 M Tris-HCl, pH 8.8 and 20% (v/v)	equilibration buffer	in distilled water	-20 °C

Ethanol	glycerol 70% (v/v)	Fixing solution	in distilled water.	Room temperature
Fixing solution	40% (v/v) ethanol and 10% (v/v) acetic acid		in distilled water.	Room temperature
0.1 M HCl (of 500ml)	Required solution: $V_i = 500 \text{ mL} * (1 \text{ L}) / (1000 \text{ mL}) = 0.5 \text{ L}$ $C_i = 0.1 \text{ mol/L}$ Source solutions: weight = 36.5 % = 36.5 * 1/100 = 0.365 density = 1181 g/L	Aqueous solution used as reaction component	in distilled water -determine the vol of conc HCl, with a density and weight of HCl in a total volume of a 0.1 mol/L HCL	4 °C
<p>Concentration of the source (stock) soln:  <math>C_f = \text{density} * \text{weight} / M(\text{HCl})</math>  <math>= 1181 \text{ g/L} * 0.365 / 36.4609 \text{ g/mol}</math>  <math>= 11.8227 \text{ mol/L}</math></p> <p>Amount of solute before and after dilution remain same:  <math>V_f = V_i * C_i / C_f</math>  <math>= 0.5 \text{ L} * 0.1 \text{ mol/L} / 11.8227 \text{ mol/L}</math>  <math>= 0.00422917 \text{ L}</math>  <math>= 0.00422917 \text{ L} * (1000 \text{ mL}) / (1 \text{ L})</math>  <math>= 4.22917 \text{ mL}</math></p>				
2X SDS sample loading buffer	60 mM Tris pH 6.8, 2% (w/v) SDS, 10% (v/v) glycerol, 200 mM DTT, 0.025% (w/v) bromophenol blue	loading buffer (reducing buffer)		Room temperature
1X SDS-PAGE	25 mM Tris, 192 mM glycine containing 0.1% (w/v) SDS	running buffer		Room temperature
Tris-buffered saline (TBS)	50 mM Tris and 150 mM NaCl, pH 7.5.			4 °C
TCA10%	10% (w/v)	Precipitation solution	TCA in acetone (freshly Prepared) precipitates the protein extract	-20 °C
Tris-Hcl	0.375 M Tris-HCl, pH 8.8, 50% (v/v) glycerol with a tint of bromophenol blue	Displacing	in urea buffer	4 °C
0.5 M Tris-HCl, pH 6.8	0.5 M Tris in distilled water adjusted to pH 6.8 with concentrated Hcl.	Tris-HCl	in agaros gel	4 °C
1.5 M Tris- HCl, pH 8.8	1.5 M Tris in distilled water adjusted to pH 8.8 with concentrated HCl	Tris-HCl	in agaros gel	4 °C
Trypsin digest	- 10µl trypsin buffer (20ng/µl in 50mM acetic-acid) - 50µl 100mM ABC in 50µl stock frozen aliquots to final conc.10ng/µl, - incubate on ice for 15 minutes, - incubate overnight (15-20hrs) @ 37°C	Trypsin buffered solution	Digest (degrade) protein into peptides	-20 °C
0.1%TFA	10µl 0.1%TFA in Milli-Q grade water	Washing, equilibration	- to equilibrate the column and to remove the wetting solution (wash1),	Room temperature

		solution	salt and loosely bound impurities.	
			- to clean tryptic peptides in acidic 0.1% TFA(aq) from some hydrophobic contaminants	
0.1% TFA in 33% ACN	5µl 0.1% TFA in 33% ACN	Elution solution	- to elute protein peptides	Room temperature
Urea buffer	9 M urea, 2 M thiourea and 4% 3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS).	Extraction buffer (solubilization solution)		-20 °C

Plant tissues stored under  $-80^{\circ}\text{C}$  were targeted in protein extraction. Tissues were ground by grinding in a liquid nitrogen using autoclaved pestle and mortar into fine powder. Then the powder was transferred into falcon 50mL conical centrifuge tubes and kept under  $-80^{\circ}\text{C}$  until the time of extraction. The tissue samples were weighed immediately upon removal from  $-80^{\circ}\text{C}$  using autoclaved spatula that was kept at same temperature for adding the ground matter to 2ml eppendrofs kept at  $-80^{\circ}\text{C}$ . This was precipitated with 10% (w/v) TCA in acetone which was freshly prepared and stored at  $-20^{\circ}\text{C}$  until use (Méchin *et al.*, 2007).

The reason for using TCA and acetone lies on their ability to denature and precipitate protein so as the solution inactivates the phenol oxidases and oxidases, blocking phenol oxidation into quinones, which would result in protein binding into insoluble complexes (Carpentier *et al.*, 2005, Mechin *et al.*, 2007). Further, they block proteases activity (Johnson *et al.*, 2011) and phenol extraction (Granier, 1988). The precipitated tissues were briefly vortexed to homogenize the solution which were then centrifuged at room temperature for 10 minutes at 13,400rpm to separate the unlysed tissue debris in the form of pellet at the bottom from the undesired lysate supernatant at the top of solution. By discarding the supernatant, the pellet was washed with 1.5ml of ice-cold 80% (v/v) acetone three times and after each wash was centrifuged at 13,400rpm for 10 minutes. Acetone solubilizes the pigments, lipids, and terpenoids present in the tissue very easy (Mechin *et al.*, 2007) and 2-mercaptoethanol (2ME) prevents the formation of disulfide bonds during precipitation (Gallina *et al.*, 2002). The pellet (that comes from 500mg of ground tissue) was air dried for 5 minutes at room temperature and then re-suspended in a pre-optimized volume (0.5 ml) of urea buffer (see Table 4.1) by vortexing vigorously overnight at room temperature. To optimize the volume of urea and amount of tissue, we did serial experiments by varying volumes of urea buffer against specific amount of ground tissue and found a ~1:1 ration of urea buffer to weighed ground



powder tissue.

Homogenised tissues were centrifuged at 15,700rpm for 10 minutes and air dried at room temperature. The soluble protein in the supernatant was collected in a fresh 2ml eppendorf and kept @ -20°C until further processes for quantification and electrophoretic step.

#### 4.2.4 Quantification of Protein

Quantification of protein concentration for all extracts was determined using a modified Bradford Assay (Bradford, 1976) as previously described in Ndimba *et al.* (2003). Two copies of each standard Bovine serum albumin (BSA) was prepared by mixing up BSA stock solution (5mg of BSA in one ml of urea buffer), extraction buffer, 0.1 M HCl and deionized water in 2 ml plastic cuvettes. Table 4.2 shows the Bradford assay chemical composition for protein quantification. Five BSA standards with final concentration ( $\mu\text{g}/\mu\text{l}$ ) 0 or blank, 5, 10, 20, 40 and 50 were used. In addition, duplicate protein extracts were prepared in a separate plastic cuvettes and 5 $\mu\text{l}$  of unknown protein sample was mixed with 5 $\mu\text{l}$  of urea buffer, 10  $\mu\text{l}$  of 0.1 M HCl including 80  $\mu\text{l}$  of deionized water.

Table 4.2 Bradford assay for protein quantification

BSA final <sup>1</sup>	0 (Blank)	5	10	20	40	50	Protein
BSA stock <sup>2</sup>	0	1	2	4	8	10	sample <sup>3</sup>
Extraction Buffer ( $\mu\text{l}$ )	10	9	8	6	2	0	5
HCL ( $\mu\text{l}$ )	10	10	10	10	10	10	10
Deionized water ( $\mu\text{l}$ )	80	80	80	80	80	80	80
Bio-RAD dye reagent	900	900	900	900	900	900	900
Total	1000	1000	1000	1000	1000	1000	1000

Key to legend: <sup>1</sup> BSA final Concentration ( $\mu\text{g}/\mu\text{l}$ ); <sup>2</sup> BSA stock solution (5mg/ml) ( $\mu\text{l}$ ); <sup>3</sup> Protein sample 5 $\mu\text{l}$  (unknown)

A volume of 900 $\mu\text{l}$  of the diluted Bradford reagent which was prepared by mixing a concentrated Bradford reagent (Bio-RAD) with deionized water in a 4:1 ratio was added to all the component mixture of BSA standards and the protein extracts respectively to make 1ml of the respective final

volume. After mixing and incubating each at room temperature for 5 minutes, absorbance values were measured by setting spectrophotometer (an instrument to measure intensity of light absorbed as wavelength) at 595nm. The standard curve was determined from the absorbance values of the BSA standards and this was used as a basis for determining the concentration of all unknown proteins.

#### **4.2.5 Separation of protein**

##### **4.2.5.1 Electrophoretic separation of proteins: one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (1D-SDS-PAGE)**

###### **4.2.5.1.1 Preparation of resolving gel**

Electrophoretic separation of protein extracts using 1D SDS-PAGE was carried out based on the description given in (Ludevid *et al.*, 1984 and Brini *et al.*, 2007). Acrylamide/Bis stock solution with a ratio 37.5:1 (2.6% C, BIO-RAD) was used for making two types of gels (12% resolving gel and 5% stacking gel) in the system. As indicated in Table 4.3a, resolving gel was prepared by mixing up 4.3ml of deionized water with 3ml of a 40% Acrylamide/Bis stock solution with a ratio 37.5:1 (2.6% C, BIO-RAD), 2.5ml of a 1.5 M Tris-HCl gel buffer, pH 8., 0.1ml of both 10% SDS (w/v) and APS (w/v) each together for a single gel cast. Of this total volume, 5ml was used to cast the resolving gel. A standard gel electrophoretic casting system (Mini-PROTEAN® 3, BIO-RAD) was used to cast 1D gels on a 1mm thickness mounted glass plates (BIO-RAD, spacer) with the width and height of 10.1cm and 8.3 cm respectively. A 1.5cm space at the top of resolving gel was left for the 5% (v/v) stacking gel solution (Table 4.3b), which was filled by overlaying with 1ml of a 100% isopropanol to remove any air bubbles. This was kept @ room temperature for 20-30 minutes to allow polymerization soon after adding TEMED and mixing the solution.

As TEMED and ammonium persulfate are known to be initiators of polymerization (Maurer, 1978), these were only added at the latter stage to block early polymerization. Because a reproducible polyacrylamide gels electrophoresis requires polymerization, the presence of Catalyst-redox system such as, APS-TEMED complex in the reaction mixture is important to provide free radicals (Issa *et al.*, 2011; Thakur *et al.*, 2011 and Posch *et al.*, 2013). Therefore, a tiny concentration of tertiary aliphatic amines (such as TEMED) accelerate the polymerization (Maurer, 1978; Parthasarathy *et al.*, 2011; Gunavadhi *et al.*, 2012). Other than being the initiator of polymerization,

it's high purity when prepared, relative stability in 0°C and low tendency to generate molecular oxygen, makes APS a good candidate for suitable catalytic system in polymerization of PAGE without affecting the selected buffer condition and electrical conductivity (Maurer, 1978; Lobos *et al.*, 1991; Kiatkamjornwong *et al.*, 1999; Sahiner *et al.*, 2007). However, recently, APS has shown a varied conductivity with no considerable effect on morphology (Choi *et al.*, 2004) and viscosity of the gels (Nelson *et al.*, 2005). The APS-TEMED-complex, thus, initiate chemical polymerization by the free monomer radical which are dependent on the base catalysed formation of free oxygen radicals (McJury *et al.*, 2014). Both the free oxygen radicals and free base are supplied by sulphate and TEMED respectively. Thus, as avoiding retarded polymerization is an issue of reproducibility of gels, due cautious were given to prevent inhibitors of chemical polymerization such as molecular oxygen (Decker and Jenkins, 1985), cooling and metal impurities (Fevotte *et al.*, 2013) and by selecting standardized condition such as relatively higher pH so that the solutions will gelate within not more than an hour (Liu *et al.*, 2013). The isopropanol was rinsed off with deionized water and the gel surfaces were made dry by blotting with filter paper.

The relatively larger pores size in the stacking gel provides greater chances for proteins to be concentrated into thinner areas which assist protein mobility on the basis of net charge compared to the smaller pore sizes of the resolving gel in this area which allow the separation of protein on the basis of MW (Westmeier, 2006). Such an increased acrylamide concentration retards the protein mobility in resolving gel than in stacking.

#### **4.2.5.1.2 Preparation of stacking gel**

Stacking gel (5%, v/v) was similarly prepared from the combination of all but with different volumes of the above ingredients used for resolving gel except 0.5 M Tris-HCl gel buffer, pH 6.8 used in place of 0.5 M Tris-HCl gel buffer, pH 6.8 as indicated in Table 4.3b. A mixture of 3.64ml of deionized water with 0.63ml of a 40% Acrylamide/Bis stock solution (37.5:1, 2.6% C, BIO-RAD), 0.63ml of a 1.5 M Tris-HCl gel buffer, pH 6.8, 0.05ml of both 10% SDS (w/v) and APS (w/v) each was made for a single gel cast. TEMED was similarly added only at the latest stage for the reason already described. Once the 5% (v/v) stacking gel was ready prepared, 1ml of the solution was added to overlay the dry resolving gel. After this step, a 1mm thick BIO-RAD well combs with 10 - 15 well combs based on the size of the gel and on the number of protein samples to be loaded were inserted in the stacking gel soon after pouring to form wells for loading samples.

This was left to polymerization for another 30 minutes at same temperature.

#### 4.2.5.1.3 Running Electrophoresis for 1D SDS PAGE

Electrophoresis for 1D SDS PAGE was carried out immediately following protein sample preparation. Protein samples were mixed with 2X SDS sample loading buffer, also called reducing buffer, [60 mM Tris pH 6.8, 2% (w/v) SDS, 10% (v/v) glycerol, 200 mM DTT (a product to break covalent disulfide bonds to allow full denaturation of proteins before loading, Rabilloud, 1996), 0.025% (w/v) bromophenol blue or also called a tracking dye used to control the protein samples while electrophoresis is running] at a ratio of 1:1. The mixture was briefly vortexed, and heated to denaturate at 95 °C for 5 minutes and pulse centrifuged. Complete denaturation of all proteins following immediate heating limits degradation through the combination of heat, SDS, and reductant. However, prolonged excessive heating was prevented because it may cause selective aggregation and band smearing by breaking peptide bonds (Gallagher, 2006).

Table 4.3 Preparation of resolving and stacking gels for SDS-PAGE

Composition	(a) 12% Resolving gel buffer	(b) 5% Stacking gel buffer
Deionized water	4.3 ml	3.64ml
40% Acrylamide/Bis stock solution 37:5:1 (2.6% C)	3ml	0.63ml
1.5 M Tris-HCl, pH 8.8	2.5 ml	-
0.5 M Tris-HCl, pH 6.8	-	0.25 ml
10% SDS	0.1 ml	0.02 ml
10% APS	0.1 ml	0.02 ml
TEMED	0.008ml	0.005ml
<b>Total</b>	<b>18.008ml</b>	<b>4.565ml</b>

A 20 µg total protein quantities in each well (lane) was loaded onto 10-welled gels once the stacking gel was solidified and all the buffer tank apparatus, a vertical slab gel electrophoresis apparatus (Bangalore Genei) were in position. Electrophoresis was run in 1X SDS-PAGE running buffer (25 mM Tris, 192 mM glycine and 0.1% (w/v) SDS) on a Bio-RAD Electrophoretic Cell assembly system (Mini-PROTEAN® 3) as previously described by (Simpson, 2006). The loaded gel wells were totally immersed in the running buffer solution and the 3 µl of the PageRuler™ unstained protein ladder (Fermentas) was used as a marker. Electrophoresis was conducted at 100V for the

first 15 minutes and then maintained at 120V until the tracking dye (bromophenol blue) migration approached the extreme lower part of the plates.

#### **4.2.5.1.4 Staining gels and analysing proteins for 1D SDS PAGE**

After removing the gels from the gels plates, protein bands were detected by staining with coomassie brilliant blue (CBB) R250 for 2 to 3 hours depending on the clarity of the background as previously described in (Neuhoff *et al.*, 1988). The protein bands were analysed by using a Gel Documentation System (GDS) linked to a computerized Image Analyser (BioRad PharosFX™ plus molecular imager). The reproducibility of the result was checked by repeating the experiment three times.

#### **4.2.5.2 Two-Dimensional (2D) Sodium Dodecyl Electrophoresis (SDS-PAGE) Sulfate-Polyacrylamide Gel**

##### **4.2.5.2.1 Sample preparation for (protein loading) 2D Gels**

Protein sample for loading on 2D gels was meticulously prepared as previously described in (Grabski and Novagen, 2001). Sample preparation is the key for a reliable result in the 2D-gel electrophoresis which is determined by several factors such as solubility, size, charge, and isoelectric point (pI) of the protein of interest (Jiang *et al.*, 2004). This creates the variation in sample complexity of the protein extracts used into the 2D gels and the length and pH range of immobilised pH gradient (IPG) strips used for IEF. Thus, it minimizes the complexity in protein extracts with least possible ionic strength of the denatured buffer maintaining proteins original charges and its solubility. A volume of 150-200µg protein extracts (as calculated from Bradford assay) were added to the rest of reaction mixture or also called rehydration solution (2µl of 50% DTT (v/v), 1.25µl of ampholytes (BIO-RAD), a tint of bromophenol blue) making a final volume of 125µl by adding urea buffer. The sample solution was vortexed to mix for 3-5 minutes at room temperature and then pulse centrifuged.

##### **4.2.5.2.2. Selection of appropriate technology for Isoelectric Focusing (IEF; First Dimension)**

IEF techniques are categorized in two forms such as classical (conventional) isoelectric focusing (Righetti *et al.*, 1986; Garfin 2000) and modern techniques (Righetti *et al.*, 1986; Righetti and Bossi, 1998 and Fredman *et al.*, 2009). The former is characterized by a carrier ampholyte generated pH gradient and is older tube gel method (Görg, 1988 and 1991) compared to the later

which is a more advanced technique and is of a typical nature of Immobilized pH gradient (IPG). In this experiment, the modern IPG based IEF technique was used to resolve protein extracts in the first dimension of 2-DE as it is currently most applicable for its characteristic reproducibility, ease of use and throughput (Garfin, 2001) used in most proteomics labs. The length and pH gradient ranges of interest of the IPG strips were identified and selected before directly embarking to series of steps for first dimension of 2D SDS PAGE.

#### **4.2.5.2.3 Rehydration of IPG Strips**

After making the solution ready as stated above, rehydration step of IPG Strips was followed by loading the sample mixture. Immobiline™ Dry Strip Reswelling Tray (GE Healthcare, Amersham, UK) was used to load the sample in each channels (1-12) depending on the number of protein samples to be separated. IPG strips (a Linear and ReadyStrip™ IPG strips) with varying length and pH, however, for this experiment, 7 cm of length, pH range of 4-7 were used and cautiously positioned on top of the protein sample but with the gel side facing downward. The IPG strips were carefully handled to avoid tearing of any part of strips that otherwise would result in the absence of focused protein in that region. Before leaving the IPG strips passively rehydrate for overnight or at least 20 hrs at room temperature to regain its normal gel size of 0.5 mm thickness, three important events were double checked. First, careful avoidance of any air bubbles in the process that would otherwise be trapped in the agarose that joins the strip to the top of the second-dimension gel creating blank stripes in the vertical dimension. Second, sufficient rehydration of a region of the IPG strip so that no part of the strips let the protein focused; Thirdly, covering the strips by overlaying mineral oil (PlusOne DryStrip Cover Fluid; GE Healthcare) to avoid sample loss as vapour while rehydrating.

#### **4.2.5.2.4 First Dimension (IEF of IPG Strips)**

Having successfully completed the rehydration step, the IPG strips were briefly rinsed with deionized water to make free of any sample of protein remain not absorbed and remove any excess liquid from the strip by cautious blotting with wet filter paper. Then after, the strips were transferred to the IEF focusing tray (platform) (Ettan™ IPGphor II™, GE Healthcare) placing this time the gel side up. Immediately following this, with the purpose to collect excess salts and any contaminants from the sample during focusing wetted but not soaked wicks blotted using filter paper were placed at the two extreme ends (anode, +ve and cathode, -ve ends) of IPG strips. Mineral oils were used to

cover the strips before starting the focusing which was aimed to prevent loss of sample due to evaporation and CO<sub>2</sub> intake while running IEF.

After closing the lid of the GE Healthcare platform, IEF was set in a three part stepwise programme @ 20°C for 7 cm IPG strips. These were namely: 250V for 15minutes, 4 000V for 1hr and lastly 4000V for 12000Vhr. These figures representing the total volt-hours for the used IPG strips, suggest the running (focusing) conditions of the IPG strips on the GE Healthcare platform. As was intended from the start of sample preparation (see also section 4.3.1), the same type of sample, buffer, and IPG strip pH range together were used to better control IEF running conditions of the IPG strips.

#### 4.2.5.2.4.1 IPG Strips Equilibration

On completion of the IEF run, two intermediate but important steps were taken place before Second Dimension SDS-PAGE starts. Equilibration of the focused IPG strips in SDS reducing buffers, and sealing of the strips by embedding on the top of the second-dimension gel. Resolved IPG strips were equilibrated in SDS-containing buffers to make the proteins easily soluble and to provide conditions for SDS binding. The resolved IPG strips were placed in reswelling tray channels with gel up side and covered by 2ml equilibration buffer.

As indicated in Table 4.4, two types of equilibration buffers were used to equilibrate the focused strips based on the types of chemical agents added to the equilibration bases buffer (6 M urea, 2% (w/v) SDS, 50 mM Tris/HCl, pH 8.8 and 20% (v/v) glycerol). Dithiothreitol (DTT) and Iodoacetamide (IAA) were used as reducing and alkalining agents respectively.

Table 4.4 Buffer system for equilibration of IPG Strips

Equilibration bases buffer		DTT Equilibration Buffer	IAA Equilibration Buffer
Reagents	amount (final concentration)		
Urea	36 g urea (6M)	<b>Reagent:</b> DDT	<b>Reagent:</b> IAA
20% SDS	10 ml (2%)	<b>Action:</b> reduces sulfhydryl groups	<b>Action:</b> alkylates sulfhydryl groups
1.5 M Tris/HCl, pH 8.8 gel buffer	3.3 ml (0.05 M)	<b>Composition:</b> add 200mg of DTT to 10ml of equilibration base buffer to make 2% DTT Equilibration Buffer	<b>Composition:</b> add 250 mg dry IAA to equilibration buffer to make 2.5% IAA equilibration
50% Glycerol	40 ml (20%)		
Water	Adjust to 100 ml		

While DTT equilibration buffer reduces sulfhydryl groups (Cleland, 1964; Herbert *et al.*, 2013), the IAA equilibration buffer alkylates the same sulfhydryl groups (Sondej *et al.*, 2011). Focused IPG strips were first incubated in the 2.5ml DTT equilibration buffer, (an equilibration bases buffer containing 2% (w/v) DTT) agitating for 10 minutes and then decanting the buffer. The strips were incubated again by 2.5% (w/v) IAA for additional 10 minutes with gentle agitation at room temperature.

#### **4.2.5.2.5 Second Dimension SDS-PAGE**

In order to run second-dimension gels resolving gel solution and sealing solution were first prepared as described in Table 4.1. A Bio-Rad multi-cast Mini-PROTEAN<sup>®</sup>3 with 12 gels per run (Garfin, 2001) was used to cast Mini format 2D SDS-PAGE gels on spacer glass plates (BIO-RAD) as previously described in section 4.2.5.1.1 for 1D gel. Similar to the 1D SDS PAGE, the resolving gel solution (see Table 4.1) was used to make the 2D gels by pouring into the cast plates and overlaying each gel with 1 ml of 100% isopropanol for the same reason previously mentioned in section 4.2.5.1.1. The procedure for making resolving gel is as described in section 4.2.5.2. After rinsing off the overlaid isopropanol from a polymerised gel, gel surfaces were blot-dried (see section 4.2.5.1.1). Then after, an equilibrated IPG strips, 7 cm (section 4.2.5.2.4.1) which were carefully rinsed with 1X SDS-PAGE running buffer (Table 4.1) were placed on top of the solidified 12% resolving gels with the plastic backing against the spacer plate with the gel side of IPG facing outside to the shorter glass plater. Unstained protein ladder (PageRuler<sup>™</sup>, Fermentas Life Sciences, Ontario, Ca) was used as protein marker by applying 3 µl on a piece of filter paper. After letting it air-dried, this was positioned at the positive side of the IPG strips.

A sealing solution (see Table 4.1) which was prepared by mixing 1 ml of 0.5% (w/v) molten agarose with 1X SDS-PAGE running buffer containing a small amount of bromophenol blue (migration tracking dye during electrophoresis) was used to seal the IPG strips with the resolving gels. Lastly, to run electrophoresis for second Dimension SDS-PAGE, all the gel plates containing gels in the apparatus multi-cast Mini-PROTEAN<sup>®</sup>3 (Bio-Rad) were carefully assembled and then the power supply was set to 200V constant for 35 minutes using Dodeca cell Bio-Rad (Mini-PROTEAN<sup>®</sup>3).



#### **4.2.5.2.6 Detection of proteins by staining gels for 2D SDS PAGE**

A modified CBB R-250 staining protocol (Wang *et al.*, 2007) for SDS-PAGE gels were used to detect proteins resolved in the first and second dimension SDS PAGE. Application of 3 but consecutive staining procedures each for the minimum of 30min was carried out. Its broadest range of proteins stain, makes CBB R-250 the most common stain for protein detection in polyacrylamide gels (Smejkal, 2004). The gels were first immersed in CBB I (see Table 4.1) after dismounting all gel plates from electrophoretic the apparatus. This was then heated for a minute in a microwave at full power and placed for 30 - 45 minutes on a shaker @ room temperature. After discarding the used CBB I, similar procedures were applied for CBB II and CBB III. For the chemical constituent in the solution of CBB II and III please see Table 4.1. After CBB III stain, the gels were destained using destaining solution (see Table 4.1) by agitating at room temperature. This continued until clear protein spot was observed.

#### **4.2.6 Image digitizing and analysis of protein spots**

##### **4.2.6.1 Spot imaging by Molecular Imager FX Pro Plus Multi-imager System**

Gels were first digitized or imaged using an instrument named Molecular Imager PharoFX Plus Multi-imager System (Bio-RAD) which is flexible and expandable machine. Then after, the gels were analysed with an image evaluation system using computerized Bio-Rad imaging systems integrated with PDQuest™ software which are robust to export and import images (Voordijk *et al.*, 2003). Coomassie stained gels were imaged with a multi-imaging system.

##### **4.2.6.2 PDQuest analysis of 2D SDS-PAGE**

PDQuest™ software (Bio-Rad) version 8.0.1 build 055 (Garrels, 1989) was used to analyse the digitized gel images. PDQuest analysis were presented in terms of spot detection and quantification, gel comparison and statistical analysis. PDQuest software provides nine progressive analysis steps to generate an accurately detected and matched protein spots to make ready for MALDI-TOF-MS/MS downstream analysis. First, new experiment work flow was established for the 2D-gel electrophoresis. To create consistency and reproducibility of the work flow, the parameters were set for the analysis and the default setting and steps were followed to compare the biological effects of the treated samples against the untreated ones based on the 2D-gels that consist the protein spots of interest. A master gel, a virtual combination of all the spots of interest from all the six gel images,

was automatically created and was used as a reference for analysis of protein spots.

Spots were labelled using PDQuest advanced annotation feature with text. Spot detection and quantification of the 2D-gel were carried out automatically after correcting the raw image data and subtracting the gel background. PDQuest uses 3-D Gaussian distributions and models to automatically detect and resolve merged spots. From this spot intensity were determined. Normalization of gels helped to balance the differences in spot quantities and group consensus based manual editing of spots provided expression consensus for all biological replicate. Based on this, differential expression of proteins were qualitatively and quantitatively and statistically (Students t-Test) detected using dummy variable (indicators) such as 1 or 0 and the values for the 2-fold expression change respectively. The p-value < 0.05 (95% significance level) was used as the cut off value for statistical significance of confidence level. For analysis and identification of protein using MALDI-TOF mass spectrometry and other related tools, qualified spots were manually picked by pipette tips, however, spots could automatically be cut using the ExQuest™ spot cutter (Bio-RAD).

#### **4.2.7. Mass Spectrometry**

##### **4.2.7.1 Protein Identification using MALDI-TOF MS**

###### **4.2.7.1.1 Excision of Coomassie stained protein spots**

Before proceeding with excision of protein spots, Coomassie Brilliant Blue (CBB) stained gels were scanned using Molecular Imager PharoSFX Plus System (BIO-RAD) as stated in section 4.2.6.1. The intact CBB stained gel was then rinsed with 70% Ethanol and with dH<sub>2</sub>O and then incubated on the shaker twice for 20 mins each. In order to identify proteins in each selected spot, the CBB stained gel was excised as close to the spot as possible, with no excess around the spot. Manually, protein spots were picked using yellow pipette tips by cutting the tips at its second node and transferred into sterile labelled micro-centrifuge tubes.

###### **4.2.7.1.2 Proteins in-gel digestion**

Excised gel protein spots were in gel digested with trypsin buffer (according to the revised unpublished protocol by Prof. B. Ndimba, Proteomics laboratory, University of the Western Cape

(UWC). A 50kDa marker band was used as a “reference” piece of gel and was processed in parallel with the test gel pieces. Gel pieces were then washed twice using the washing solution (500µl 50mM ammonium bicarbonate (ABC) and acetonitrile) by occasional vortexing for 5 minutes each. The gel pieces were washed again using the same solution and volume for 30 minutes by occasional vortexing. The gel pieces were then destained using 500µl 50mM ABC in 50% acetonitrile (ACN) for 30 minutes, occasionally vortexed. This step was repeated as necessary to remove any remnant CBB colour. The supernatant was removed at every wash. Reduction and alkylation steps were skipped (see Figure 4.2) as the 2D gel had already been undergone reduction and alkylation to equilibrate IPG strips in the first dimension of the 2D PAGE. Reduction and alkylation help to maintain the normal state of the cystine-containing peptides which otherwise have been in-gel digested ( Liu *et al.*, 2013 and Gonzalez-Fernandez *et al.*, 2014) until protein spots are exposed to trypsin for further in-gel-digestion. These are also of help to minimize the appearance of unknown masses in MS analysis for disulfide bond formation and side chain modification (Wang *et al.*, 2011). Dehydration of the gel was carried out using 100µl of 100% acetonitrile by leaving at room temp for 10 minutes until shrink, followed by Speed Vac SC100 (ThermoSavant, Waltham, MA, USA) to completely desiccate the gel-piece, until the colour turns white.

Proteins were in-gel digested with 10µl sequencing grade modified trypsin (Promega, Madison, WI, USA) dissolved in 50µl of 50 mM ABC for at least 15 hrs at 37°C. Peptide bonds on the carboxyl side of Lysine and Arginine residues were cleaved as a result of action of trypsin, which is a Serine protease. However, cleavage can be prevented or slowed by proximal acidic, aromatic or proline residues, as proline having the most significant effect. Peptide fragments with one missed cut are common and should be taken into consideration during mass analysis (Gasteiger *et al.*, 2005). The digested proteins were then stored at -20°C until further analysis. The pH level of the tryptic digest was checked using paper strip and made to remain below 4 by quenching with 30µl of 2% trifluoroacetic acid (TFA).

#### **4.2.7.1.3 Peptides extraction: Zip-tip procedure**

Peptides were extracted based on the three steps of Zip-tip procedure using the ZipTip® Pipette Tips for sample preparation (Millipore, Billerica, MA, USA; PR02358, Rev. A, 02/07, © 2007 Millipore Corporation; <https://www.millipore.com/userguides/tech1/p36404>). These steps were: 1)

Wetting the column and equilibrating the zip-tip pipette tips, 2) Binding and washing peptides and 3) Eluting the peptides into clean vials. In order to equilibrate, a 100% ACN (wetting solution) was used to wet the Zip-tip Pipettor column with the maximum volume of 10µl setting and to remove organic impurities. The wetting solution was aspirated into the pipettor tips by depressing a pipettor plunger to a dead stop, and dispensed to waste. This was repeated twice for maximum equilibration of the zip-tip pipette tip (Figure 4.2).

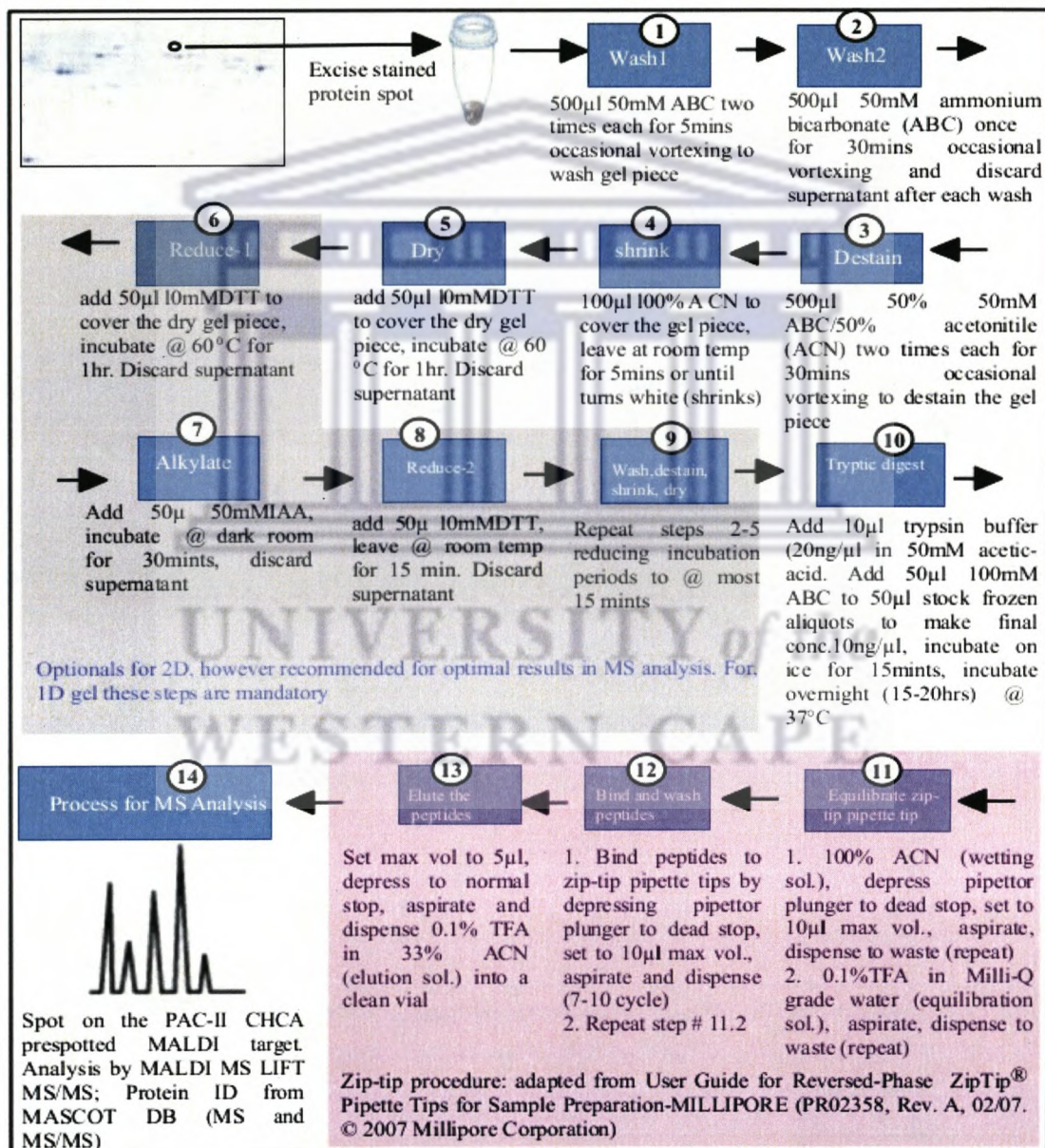


Figure 4.2: summary of procedure for in-gel tryptic digestion of protein spots.

This diagram shows the procedure for in-gel tryptic digestion of protein spots from sorghum leaf tissues. In some protocols the steps 6 to 9 (grey) of this protocol are considered to be optional for 2D gel though highly recommended

for optimal results in mass spectrometry applications (Shevchenko *et al.*, 1996b).

An equilibration solution (0.1%TFA in Milli-Q grade water) was aspirated into the pipettor to equilibrate the column and to remove the wetting solution (wash1), salt and loosely bound impurities. These were dispensed to waste and the process was repeated at-least twice. Figure 4.2 summarises the procedure for in-gel tryptic digestion of protein spots from sorghum leaf tissues according to a protocol revised and the zip-tip procedure based on the user guide for reversed-phase ZipTip<sup>®</sup>; Pipette Tips for sample preparation-MILLIPORE. Billerica, MA, USA.

To bind the peptides to zip-tip pipette tips (a conditioned column in acidic TFA in H<sub>2</sub>O), by depressing pipettor plunger to dead stop, the samples were aspirated and dispensed (7-10 cycle) using similar volume setting as before for equilibrating. The tryptic peptides in acidic 0.1% TFA(aq) was washed with 0.1% TFA in water to remove some hydrophobic contaminants. This step was repeated at-least twice. Peptides were eluted using a selective organic/aqueous mixture 0.1% TFA in 50% CH<sub>3</sub>CN/ H<sub>2</sub>O, maximum volume was set to 5µl. By depressing pipettor plunger to standard stop, the elution solution was aspirated and dispensed into a clean vial carefully for at least 7-10 times without letting air into the sample.

#### **4.2.7.1.4. Spot analysis using MALDI-TOF-MS**

Samples were spotted on the PAC-II CHCA pre-spotted MALDI target for analysis by MALDI MS and were lifted by MS/MS. Each tryptic in-gel-digested protein sample of which the peptides were already eluted using zip-tip procedure with 5µl elution solution was used to spot on the MALDI target plate by setting the already calibrated P10 (0.5-10µl) pipette (Bio-RAD) to 0.5µl volume. Each peptide sample was spotted at four positions to acquire reproducibility and representativity of the technical replicates. Samples spotted onto a MALDI target plate were let well air dried for at-least an hour to load on to the MALDI TOF/TOF Mass Spectrometer (Bruker Daltonics Ultraflex<sup>™</sup>, Germany). Spot analysis and identification using MALDI-TOF-TOF MS/MS was initiated by training the algorithm with four best spots at the extreme corners of the spotted samples and then followed by calibration with a 200 - 600 shots to give inertia for changing the status and to obtain the reference Mass. Autoanalysis was allowed to run to generate peptide mass fingerprints (PMFs) for each spot. The absolute masses of the unknown protein of interest which was first cleaved into

smaller peptides was accurately measured with a mass spectrometer (MALDI-TOF; Clauser *et al.*, 1999) and then compared to the known proteins in the databases. The proteins which are encoded by the genome sequences in the databases of interest are cut into peptides *in silico* for which absolute masses are calculated by software program and compared to the masses of the unknown proteins. The best match is found based on the statistical analysis for the unknown peptides, however, most PMF algorithms assume that the peptides come from a single protein (Shevchenko *et al.*, 1996a).

#### **4.2.7.1.5 Searching for known protein sequences from databases**

Search for known protein was performed against pertinent protein databases such as SWISSPROT (Boeckmann *et al.*, 2003) and UNIPROT (UniProt Consortium, 2008) databases, the National Centre for Biotechnology Information (NCBI; Wheeler *et al.*, 2007) and Mass spectrometry protein sequence database (MSDB; Choudhary *et al.*, 2001) where MASCOT (Perkins *et al.*, 1999; Koenig *et al.*, 2008) was used to browse the MSDB. The absolute masses obtained for the unknown protein peptides based on MALDI-TOF result was used as a fingerprint to search known proteins in the databases. In all the searches, the hits were ranked according to the scores such that the more the proteolytic peptides were contained in a candidate protein, the higher was the score ranked which can match measured masses. This means that molecular weight search (MOWSE; Pappin *et al.*, 1993) scores higher than 66 ( $p < 0.05$ ) matches for a protein candidate were considered as positive identifications. A minimum of 10% protein sequence coverage is also mandatory to assign proteins. However, to avoid ambiguity where multiple proteins were meant to satisfy the minimum sequence coverage, a protein with the highest MOWSE score was considered to be a probable positive candidate.

## **4.3 Results**

### **4.3.1 Protein separation**

#### **4.3.1.1 One-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (1D-SDS PAGE)**

In order to determine the quality of protein extract in two-dimensional (2D) gel electrophoresis, we first run a one-dimensional (1D) gel electrophoresis (as described in section 4.2.5.1). This is probably because proteins in one dimension are separated so that all the proteins will lie clearly along a lane largely improving the depth of proteomic coverage though sample fractionation is associated with a moderate decrease of quantitative measurement repeatability (Gautier *et al.*, 2012). One-D SDS- PAGE can give adequate separation when the number of proteins is low, however, for a very complex mixture not effective which is sometimes difficult to perform analysis using MALDI TOF MS (Marvin *et al.*, 2000).

#### **4.3.1.2 Two-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (2D-SDS-PAGE)**

Two-dimensional gel electrophoresis analysis of the sorghum Btx642 leaf tissue proteins revealed quantitative changes as a result of responses to drought stress. The extent of expression of drought responsive proteins (DRPs) may impact the level and differences in drought tolerance indicating that the result has also shown substantial qualitative changes. This may entail that there is inherent difference in the expression level of DRPs to wards drought stresses which may be related to tissue developmental stage specificity. It has been already reported in grain sorghum for the heat shock level that there is natural genotypic differences in the extent of Heat Shock Proteins (HSPs) expression based on the spatio-temporal variation (O'Farrell, 1975; Ngara *et al.*, 2012).

### **4.3.2 Spot detection and protein differential expression analysis (PDQuest analysis)**

The main objective of this chapter is to identify the DRPs by analysing three biological replicates of sorghum leaf tissue from the post-flowering drought stressed Btx642 variety. Differential expression analysis is a common and powerful method for screening and characterizing genes or gene products based on specificity and patterns of spatio-temporal expression. Specificity in plant tissues and development stages in the evaluation of differential expression of proteins for detection

of drought tolerance is the key approach for analysis and identification of tissue or age specific biomarkers.

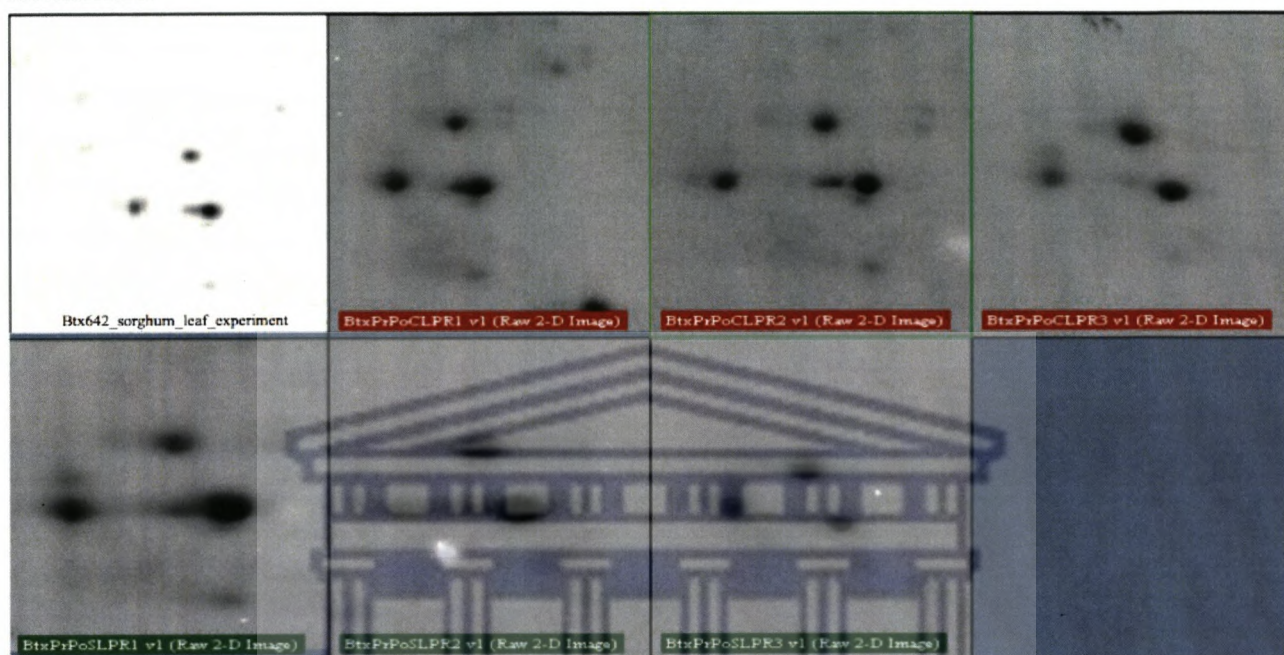


Figure 4.3: PDQuest analysis of sorghum Btx642 leaf proteins for the 3 biological replicates

This diagram shows the PDQuest analysis for sorghum Btx642 leaf proteins where gel image of the 3 biological replicates for both control (normal watering condition) and treatment (drought stressed condition,  $30\pm 5\%$  FC) groups are shown with an automatically generated master gel at the left corner.

PDQuest analysis shown in part in Figure 4.3 for sorghum Btx642 leaf tissues for the three biological replicate groups was based on differential conditions that include normal watering condition (control) and drought stressed condition at  $30\pm 5\%$  FC (treatment). Protein spots in the PDQuest analysis describe observed qualitative and quantitative variations as a consequence of differential expression, as demonstrated in Figure 4.4. Here, analysis of separation of protein extracts from sorghum leave tissue that was subjected to 2D SDS-PAGE using the IPG strips with length of 7cm and pH range of 4-7 and differential expression profiling using PDQuest are indicated. This analysis result in protein spot identification and differential expression represents a sole reflection of inherent compositional contents and complex mixture of the protein sample with differences in protein abundance between control and the treatment groups. On the other hand, Figure 4.5 illustrates schematic representation of the 2DGE showing differential expression of protein spots visualization, expression patterns and spot 3D views of sorghum Btx642 leaf proteome



contrasting the relative abundance of proteins in control and treatment samples based on the PDQuest analysis. Figure 4.6 shows the visualization of superimposed spots using multichannel viewer.

The sixteen protein spots selected for MALDI-TOF-MS/MS analysis were classified based on their spot density, protein (spots) separation based on Molecular Weight (MW) and their isoelectric point (pI) and differential expression profile. Based on their molecular mass, spot # 5, 6, and 12 represent proteins of higher molecular mass  $\geq 70$ kDa. On the other hand, the protein with spot number 11 and 14 are almost equivalent to 50 kDa. However, the majority of protein spots are in the lower molecular mass weight range  $< 40$ kDa such as spot 1, 7 and 16 are representing  $\sim 10$ kDa (Figure 4.4).

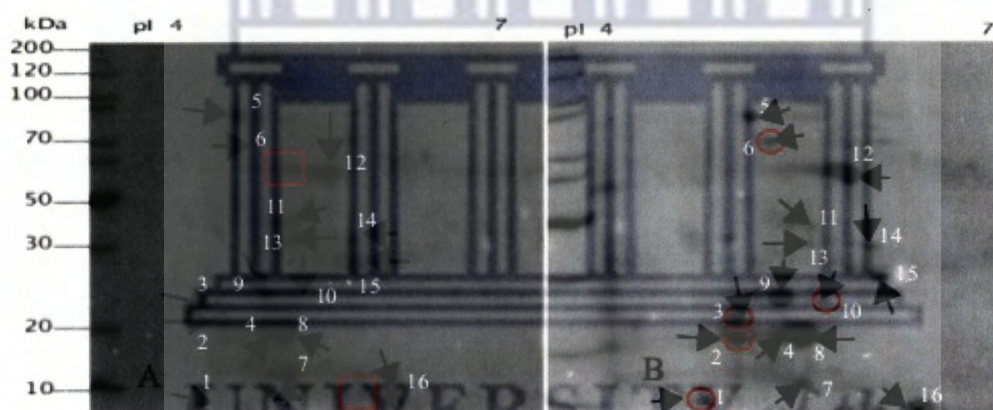


Figure 4.4: Representative 2DE showing spots selected for analysis of protein identification

Two-Dimensional (2D) gel electrophoresis of sorghum (Btx642) leaf tissues under: (A) control (normal conditions) and (B) treatment (drought stress condition). Drought responsive proteins in the circles are unique to drought treatment and in the square unique to control plants. The gel was visualized by CBB. The arrows indicate protein spots identified for MALDI-TOF TOF MS/MS analysis.

Protein spots identified with spot number 2, 8, 9, 12, and 16 were grouped in large sized spots based on 3D view of the protein spot visualization. Spots identified as spot 1, 4, 5, 7, and 14 were small sized and those spots # 3, 6, 10, 11, 13 and 15 were considered to be faint spots. Thus 75% of all the spot based on their molecular weight, were located below 50kDa of which one-third constituted 80% of the large sized spots. In terms of the pI position, almost 88% of the spots were located between 5–6 pH range except the two spots (spot 1 and spot 16) which were positioned at about 4.5 and 6.5pI values respectively. Based on the differential expression patterns, almost 63% (spots 1, 2,

3, 6, 8, 9, 10, 12, 13 and 16) were up-regulated, 12% (spots 11, 14) down-regulated and 25% (spots 4, 5, 7 and 15) remain unresponsive in stressed samples compared to the control ones.

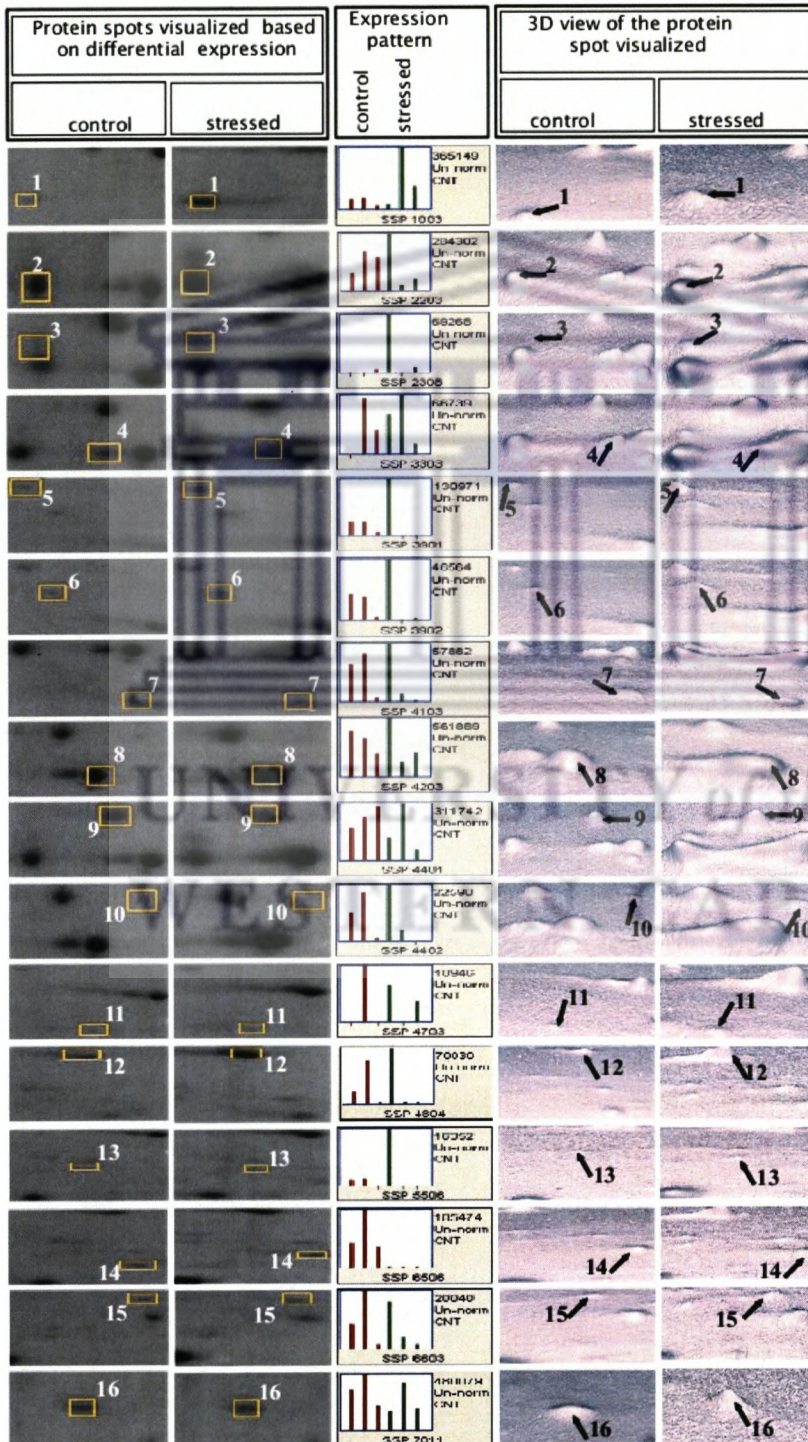


Figure 4.5: Schematic representation of 2DGE differential expression of protein spot

This figure describes 2DGE differential expression of protein spot visualization, expression patterns and spot 3D view of sorghum Btx642 leaf proteome for the control and treatment samples. Each particular spot for which corresponding differential expression pattern and the visualization of 3D view has been shown is indicated in the rectangular box respective to the number given to each spot per control and stressed samples. The difference in colour of the numbers assigned to the spot in spot density visualization and the 3D view (white and black respectively) is simply to contrast the background. Nine proteins were identified from spot number 1, 8, 9, 12, 13, 14, and 16. Two spot (# 12 and 14) resulted in to two proteins each (for the description, see section 4.3.3).

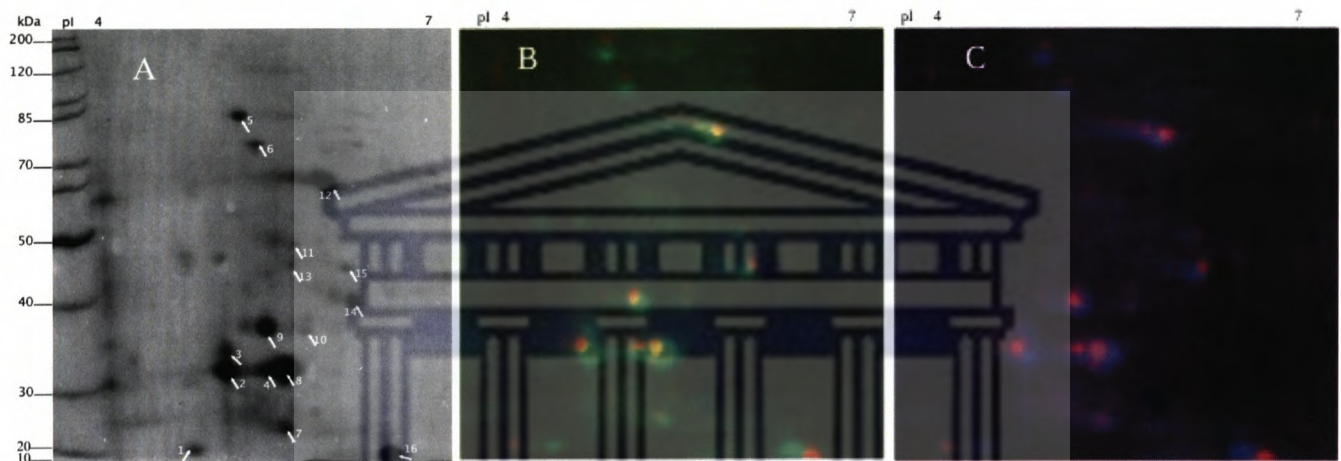


Figure 4.6: Visualization of superimposed spots using multichannel viewer

Separation of sorghum proteins using 2DGE from green leaf tissues treated under drought stress condition. Spots labelled from 1 to 16 are selected for MALDI-TOF-MS/MS analysis. PDQuest software provides an option to superimpose the corresponding spots from the control and stressed gel replicates whereby one can be moved around the corresponding spots. This option is helpful when automatic spot matching is difficult. In this Figure, A, represent a black and white gel image of a representative replicate before superimposition of the control and stressed spot using the multichannel viewer, B represent a gel image of a representative replicate of superimposed control (red-colored) and treatment (green-colored), as automatically assigned and C represent a gel image of a representative replicate of superimposed control (red-colored) and treatment (blue-colored) spot using the multichannel viewer. Green is more contrastive in B than red-colored spots whereas red is more contrastive in C than blue-colored spots (i.e. red is more contrastive in C than in B).

### 4.3.3. Protein identification using MALDI-TOF-TOF-MS/MS

#### 4.3.3.1 Protein spot analysis and spectral acquisition

A total of 86 protein spots were separated by 2D gel electrophoresis, of which 16 spots matched across all gels (Figure 4.4) and hence were selected for protein spot identification using MALDI-TOF-TOF MS/MS analysis. A relatively low number of protein spots identified in this study may probably be related to protocols (outlined in this section) that were not applied in the study because

of the time frame and resource issues. Protein identification using MALDI-TOF-TOF based on a sample desalting or concentrating with ZipTip<sub>C18</sub> method and optimization will generally result in high protein resolution (Lim *et al.*, 2003). However, protein separation and identification from the 2D gel electrophoresis could be affected by some other factors that directly or indirectly contribute to a relatively low number of protein spot identification. These factors include such as length and pH range of the IPG strips, an extent of sensitivity of staining dye, the colorimetric protein assays, electrophoretic technical and mechanical factors and the biological nature of the plant materials. Other studies show that a relative amount of loaded polyacrylamide may affect the protein recoveries of biologically-active and complex proteins (Reguera and Leschine, 2009). Most protein samples are complex solutions as they are processed containing many non-protein, interfering agents which partly or entirely affect the efficiency of protein assay. Therefore, the extent of the accuracy and sensitivity of protein assays may determine the quality of protein estimation. For instance, the Bradford Protein Assay we commonly used is considered incompatible with the common buffer and is sensitive to many contaminants, such as non-ionic detergents present in biological samples in concentrations greater than 0.1% (Pihlasalo *et al.* 2012). Studies recommend instead using 2D Quant-Kit GE Healthcare to quantify protein with such a reagent concentration in the sample buffer (Tannu and Hemby 2006). On the other hand, a fluorescence-based assay has been used for detailed measurements of expression levels and protein mapping enabling specific protein identification and further characterization (Tannu and Hemby, 2006; Mackintosh *et al.*, 2005). The use of membrane based protein separation, ultrafiltration, is also considered to give high protein purification through separating out dissolved smaller impurities and retaining a relatively larger sized proteins (Deutscher, 1990; Saxena *et al.*, 2009).

After plating the tryptic digest on the Pre-spotted AnchorChip (PAC)-II CHCA MALDI target (Bruker Daltonics systems), spectra were acquired for each sample. Peptide mass fingerprints were generated by using a MALDI-TOF-MS. MALDI-TOF-MS mostly demonstrate a relative failure to identify proteins with low molecular mass than the reverse resulting in the fragmentation of protein due to the low traceable peptides (Thiede *et al.*, 2005). However, the limitation of MALDI-TOF-MS may be resolved by complementing PMF using *de novo* sequencing MS/MS methods (Thiede *et al.*, 2005).

Unknown protein peptides isolated from sorghum leaf tissue recovered from spot selections were

microsequenced using MALDI-TOF-MS/MS with the high pattern of spectral acquisition. Non-redundant protein databases search using MASCOT software putatively identified 9 potentially significant proteins each with at least one or two peptides (Table 4.5). Two spots (12 and 14) each representing a combined spot were found to contain a mixture of proteins. ATP synthase subunit beta and enolase 1; and fructose-bisphosphate aldolase and double-stranded RNA-binding protein 5 were identified from spot 12 and 14 respectively. Thiede *et al.*, 2013 obtained similar result whereby 665 spots out of 816 revealed one to five proteins each with one 2-DE spot identified up to 23 proteins. According to the author's demonstration, it is possible to identify multiple proteins within a single 2-DE spot, regardless of the high protein resolution power of the 2-D. It is suggested that the extent of sensitivity and resolution power of MALDI-TOF could probably determine the degree of tryptic peptide detection. Similarly, Lim *et al.* (2003) has identified 9% of the spots containing multiple proteins in the study conducted to compare a MALDI/TOF peptide mass mapping with  $\mu$ LC-ESI tandem MS.

Based on MOWSE score and theoretical pI values and the number of peptide, it happened to a double-stranded RNA-binding protein 5 to seemingly be unlikely hit and may represent false positive. All the different protein spots experimentally identified using MALDI-TOF-TOF MS/MS were functionally categorized based on the extent of sequence homology with the theoretical proteins from the non-redundant protein databases. Three data bases namely the National Centre of Biotechnology Institute (NCBI), UniProtKB and SwissProt were cross-referenced both for green plants and all based on protein search. The best and significant matching proteins were selected based on primarily but not limited to MOWSE score value corresponding to the resultant combination of search result. The list of all proteins identified is indicated in Table 4.5 with all the necessary information. These include the identified and selected protein spots for analysis using MALDI-TOF-TOF MS/MS, MOWSE score for the best matches, the species to which the best match was adhered, the accession numbers as identifiers of the identified proteins and the observed experimental and predicted theoretical molecular weight (MW) along with isoelectric points (pI). Figure S4.1 shows MALDI-TOF-TOF-MS/MS spectrum of an in-gel tryptic digest of putatively identified proteins for the three top scoring proteins (Fructose-Bisphosphate Aldolase, ATP synthase subunit beta and Ribulose Bisphosphate Carboxylase large chain, RuBisCO). Table 4.6 gives brief functional description of each of the identified protein and the source of function information. Figure 4.7 shows the functional categories of the putatively identified proteins whereas Figure 4.8

illustrates the category of subcellular localization of the identified proteins. Of the sixteen spots selected for mass spectrometry analysis, seven protein spots gave significant score for the nine proteins that were identified.

#### 4.3.4 Functional and subcellular classification of proteins

Five different functional categories of proteins were putatively identified based on the analysis of MALDI-TOF-TOF-MS/MS and the protein database search using MASCOT.

**Table 4.5** List of proteins identified by MALDI-TOF-TOF-MS/MS analysis

Code <sup>1</sup> Spot SSP # <sup>2</sup>	Accession <sup>3</sup>	Protein <sup>4</sup>	Organism (Species) <sup>5</sup>	peptides <sup>6</sup>	SC (%) <sup>7</sup>	MOWSE Score	Observed MW/pI	Predicted MW/pI	Cellular location	Expression
14_6506	ALFC_ORYSJ	Fructose-bisphosphate aldolase	Oryza sativa subsp. japonica	5	15.7	209.4	43.1/5.6	42/6.5	chloroplast	-
12_4804	ATPB_SACHY	ATP synthase subunit beta	Saccharum hybrid	5	15.9	204.8	60.9/5.2	53.9/5.2	Chloroplast	+
13_5506	RBL_LACSA	RuBisCO, large chain	Lactuca sativa	4	10.3	161	45.9/5.9	52.9/5.9	Chloroplast	+
8_4203	PSBO_FRIAG	Oxygen-evolving enhancer protein 1	Fritillaria agrestis	2	7.6	74.1	34.8/5	34.8/6.3	chloroplast	+
12_4804	ENO1_MAIZE	Enolase 1	Zea mays	2	5.2	51.2	60.9/5.1	48/5.1	Chloroplast	+
1_1003	gi 242038635	hypothetical protein SORBIDRAFT_01.g012710	Sorghum bicolor	2	10.0	50.31	23.90/4.25	23.9/4.3	Chloroplast	+
16_7011	PSBP_WHEAT	Oxygen-evolving enhancer protein 2	Triticum aestivum	1	3.1	35.31	23.9/6.4	27.3/9.5	Chloroplast	+
9_4401	CHIT_PETHY	Acidic endochitinase	Petunia hybrida	2	7.5	39	39.7/5.6	27.6/5.6	Secreted, Extracellular region	+
14_6506	DRB5_ORYSJ	Double-stranded RNA-binding protein 5	Oryza sativa subsp. japonica	1	2.7	31.3	43.1/5.6	43.1/12.4	HNRNP <sup>8</sup>	-

#### Key to legend:

- <sup>1</sup> represents the code number arbitrarily assigned to the spot that were selected for the MALDI-TOF-MS/MS analysis;
- <sup>2</sup> represents the sample spot protein number assigned by the PD-Quest software to the differentially; expressed spots correspondent to the code number of the spot selected for MALDI-TOF-MS/MS analysis;
- <sup>3</sup> represents the identifier recovered for the known protein identified from the non-redundant protein databases
- <sup>4</sup> represents the predicted known protein identified from the database matching any one of the expressed protein experimentally
- <sup>5</sup> represents the species to which the best and significant match for protein was obtained
- <sup>6</sup> represents number of peptides for each protein identified
- <sup>7</sup> represents the percent sequence coverage
- <sup>8</sup> represents heterogeneous nuclear ribonucleoprotein particles

The categories include 1) Energy generation or proton (H<sup>+</sup>) transporting protein, 2) Glycolysis and

gluconeogenesis and other carbohydrate metabolism associated proteins 3) Photosystem regulation (carbon assimilation) 4) Stress tolerance, defence and immunity related proteins; 5) RNA binding proteins and 6) Unknown (Figure 4.7 and Table 4.5 and Table 4.6). On the other hand three different classes of subcellular localization (Figure 4.8) was identified based on cellular component specifically adhered to the identified proteins. This is because protein sub-cellular localization primarily determines its function (Andersen *et al.*, 2003; Wang *et al.*, 2003 and Smith *et al.*, 2004). Table 4.7 shows the sequence description for the identified peptide in the corresponding protein. Figure 4.7 shows the Functional categories of the putatively identified proteins and Figure 4.8 illustrate the category of subcellular localization of the proteins identified. MASCOT probability distribution based on peptides and proteins scores is shown in Figure 4.19, labelled A and B respectively.

Table 4.6 Functional description of protein identified

Proteins	Sequences	Functional description
Fructose-bisphosphate aldolase	K.GLVPLAGSNNESWCQGLDGLASRE	Glycolysis, catalysis the conversion of D-fructose 1,6-bisphosphate into glyceraldehyde 3-phosphate and D-glyceraldehyde 3-phosphate
ATP synthase subunit beta	R.IFNVLGEPIDNLGPVDTSATFPIHR.S	ATP generation from ADP using proton gradient across the membrane
RuBisCO, large chain	R.EITLGFVDLLR.D	Carbon fixation. RuBisCO catalyzes two reactions: the carboxylation of D-ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate in the photorespiration process. Both reactions occur simultaneously and in competition at the same active site
OEE 1 <sup>1</sup>	K.DGIDYAAVTQLPGER.V	Stabilizes the manganese cluster which is the primary site of water splitting
Enolase 1	K.FRAPVEPY.-	catalysis of the conversion of 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP)
SORBIDRAFT_01g012710 <sup>2</sup>	R.TGCSFDGSGNGQCQTGDCCGVLR.C	Response to salt and defence to fungal
Acidic endochitinase	R.VPGYGVITNIINGGIECGK.G	Defense against chitin containing fungal pathogens
OEE 2 <sup>3</sup>	R.EFPGQVLR.Y	May be involved in the regulation of photosystem II
DS-RNA-binding <sup>4</sup>	R.RNAAADAVLLR.A	Binds double-stranded RNA

Legend: <sup>1</sup> Oxygen-evolving enhancer protein 1; <sup>2</sup> Unknown, hypothetical protein; <sup>3</sup> Oxygen-evolving enhancer protein 2; <sup>4</sup> Double-stranded RNA-binding protein 5. This figure shows the functional description of the proteins identified and the sequences of the proteins. Detail description has been given in the Table 4.5. The sequence of the proteins are also found in Table 4.7.

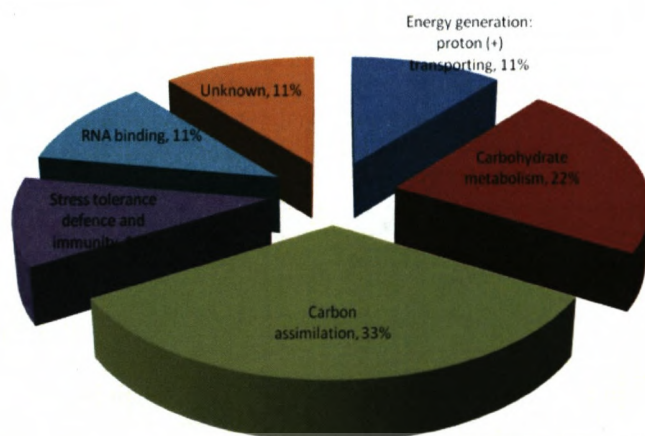


Figure 4.7: Functional category of the proteins identified

This figure shows that photosystem regulation and or carbon assimilation involve 33% of the proteins identified such as Rubisco, OEE1 and OEE2 whereas carbohydrate metabolism involving 22% include proteins such as Fructose 1, 6, Bis Phosphatase and Aldolase. The rest of the proteins such as ATP synthase subunit beta, double stranded RNA-binding protein 5 and acidic endochitinase each involve in energy generation, RNA-binding and defence and immunity categories respectively. The hypothetical SORBIDRAFT\_01g012710 is putative protein.



Figure 4.8: Category of subcellular localization of the proteins identified.

This figure shows that the maximum number of proteins were identified and localized in the chloroplast indicating that the role of these proteins to be related with photosynthetic activity. This on the other hand, suggests that most of the proteins are fairly responsible for sorghum drought tolerance.

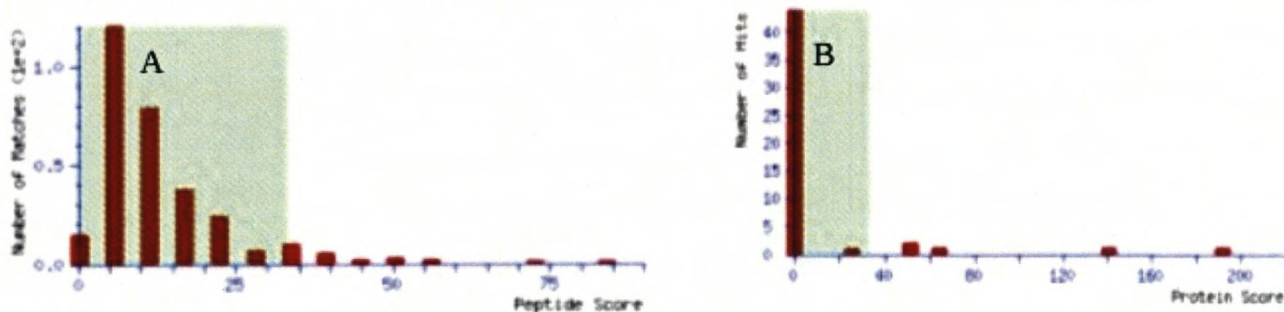


Figure 4.9: MASCOT probability distribution based on peptides (A) and protein score (B).



The protein and peptides scores greater than 30 and 35 respectively were significant ( $P < 0.05$ ). The top scoring match for protein was 195 whereas for peptide was 85. All significant scoring match were obtained with a random hit less than 5% chance, based on Mascot reports.

**Table 4.7 Sequence description of the identified peptide in the corresponding protein**

A) hypothetical protein SORBIDRA FT_01g012710												
10	20	30	40	50	60	70	80	90	100	110	120	
MASVLLLEFL	VVVVFLAA	AADAATFVY	KEQYTVGA	AVPQQQLD	HQQTWVNP	AGTTGGRVA	<b>RTGCSFDGSS</b>	<b>WQCCQTQDQ</b>	<b>GVLR</b> CTGTGQ	PFPTLAEFGL	HCYQGLDFID	
130	140	150	160	170	180	190	200	210	220	230	240	
ISLVGGHVP	MEFLPADGDS	GCPKGGPCD	ADVTGQCAA	LKATGQGNP	CTVFKDTEY	CTGSAANTG	PTDYSKFFK	LCPGAYSPK	IGATSTYCF	GQTHVWVFC	P	
B) RuBisCO, large chain												
10	20	30	40	50	60	70	80	90	100	110	120	
MEFQTEKAS	VGFRAGVQY	KLTYTFEYE	TKDIDLAA	RVTFQGVFP	EEGAAVAAE	SBTGTMTVM	TDGLTELRY	KGRCYGIKPV	FGKRNQYIAY	VAYPLDLFER	GSVTRMPTSI	
130	140	150	160	170	180	190	200	210	220	230	240	
VGVYFKRAL	RALRLELDRI	PTAYVKTQG	PFHGIQVER	KLKCYORPL	OCTIKPKLGL	SAGQYORAVY	ECLR <b>GOLDPT</b>	<b>KLDENVWQSP</b>	<b>FMWVDRFLF</b>	CAEAIKQQA	ETQEKQHYL	
250	260	270	280	290	300	310	320	330	340	350	360	
<b>NATAQTEKH</b>	<b>MGRALFARKL</b>	<b>GVPIWQDYL</b>	<b>TGQFTARTEL</b>	<b>AMYCRMSLL</b>	<b>LNIHRMAGAV</b>	<b>IDRQSNQIH</b>	<b>FRVLAKALRM</b>	<b>SGGDHINSQ</b>	<b>VVQKLEQERE</b>	<b>ETLQFVLLR</b>	<b>DDFIEKDBR</b>	
370	380	390	400	410	420	430	440	450	460	470	480	
GIYFTQWVS	LQVLPVAGS	GIHWMSFAL	TEIFQDQSVL	QFQDGLGHP	WQNAFGAVAN	EVVALRACVGA	RNEGRDLATE	<b>QREIIRKATK</b>	WPELAAACE	WWEIKVFPQ	AMOTLDQ	
C) Fructose-bisphosphate aldolase												
10	20	30	40	50	60	70	80	90	100	110	120	
MASATLKEK	FLPKSEKMA	TRGAAAFQV	TVSHVYRGA	YDDELVKAK	<b>TIASRQDIL</b>	AMEEERATQ	<b>KRLASTGLR</b>	<b>TEARRGAVK</b>	LLVTAPLQK	YISGAILFE	TLVQSTVDR	
130	140	150	160	170	180	190	200	210	220	230	240	
KIVDILTEK	IVPGIKVDR	<b>LVLQAGHNS</b>	<b>WQGLQKSLA</b>	<b>SRRAAYTQS</b>	<b>ARFARQIVV</b>	SIPHPBELA	<b>VKRAAMELR</b>	YAAIQDML	VFIVEPILL	DQEHIDATF	EVAGVQVST	
250	260	270	280	290	300	310	320	330	340	350	360	
FFYMAENIVM	FEGILLKPM	VTPGAECKR	ATPEQVSDT	LKLIHRRIFF	AVPQDPLSG	GGSEYKATQ	LRAHQGQNP	WVPSFYARA	LQFTCLKTW	GGPENVKAA	DALLRKAAR	
370	380	390	400	410	420	430	440	450	460	470	480	
SLAQLGKYS	DOEAAEAKQ	MFVQIVY										
D) ATP synthase subunit beta												
10	20	30	40	50	60	70	80	90	100	110	120	
MBTNPTEK	GVSTIEKESV	QRIDQIGPV	LDITFPQGL	PNHYALIVK	SRDTAKQIN	<b>WVCEVQGLL</b>	NHRVRAVMS	ATDGLRQRE	VIDTGTPLV	PVGGATLGR	<b>FWVLGEPID</b>	
130	140	150	160	170	180	190	200	210	220	230	240	
<b>LQPVTRATF</b>	<b>FIMR</b> RAPFI	ELDTKLSPE	<b>TKIKVYDIA</b>	<b>PTFRGRIQL</b>	FGAGVQKTV	LIDELINKIA	KARQGVVFG	GVQERTGSR	DLYEMRQK	VIREKIEES	<b>KVALVYQGN</b>	
250	260	270	280	290	300	310	320	330	340	350	360	
<b>EPFGARRVQ</b>	<b>LVALTMAEYF</b>	<b>RDVIRQIVLL</b>	<b>FIDRIFRFG</b>	<b>ADREVALLS</b>	<b>EMPRAVVQP</b>	TLSTHMSLQ	ERITSTKRS	ITSIGAVYF	ADDLTPAPA	TFFAHLDAT	VLSRGLASK	
370	380	390	400	410	420	430	440	450	460	470	480	
IYPAVDPLDS	TSTMLQPRV	QREHYETAQ	VKETLQRYK	LQDIIAILGL	DELREEDRT	VARAKIERF	LQPFVAVY	FTGSPKTVG	LAETIRGQL	ILSHELDGLF	EQAFYLVNI	
490	500											
DEASTKALM	EESEKLEK											
E) Oxygen-evolving enhancer protein 1												
10	20	30	40	50	60	70	80	90	100	110	120	
MASELQAAT	LIPARVGAFA	RTHLRENSH	SKAFQDSEF	AGRLTCSIR	DLRDIAGRT	DAAKLAGFAL	ATSALVISA	SAEGVPRKL	FDEIQRTYM	EVKGGTANG	CPTIEGQTE	
130	140	150	160	170	180	190	200	210	220	230	240	
FVYKTKYTL	KQLCLEPTEF	TVKAEIGIN	APPEPQTKL	MRLTYTLEK	IEGPFVAPD	GTVKTEERD	<b>IDYAAVTQL</b>	<b>FGSERVPLF</b>	<b>TVKQLVATK</b>	PEKFSQSYLV	PSYRGGPFL	
250	260	270	280	290	300	310	320	330	340	350	360	
PRGRGSAQY	DNAVALPAQ	RDEEELVKE	KIRVYSSYQ	KITLQVTRK	PETQEVQVF	ESIQPQDIL	GGKAPQVKI	QQIWTAKL				
F) Enolase 1												
10	20	30	40	50	60	70	80	90	100	110	120	
MAVITWYKA	RQIFDSRGNP	TVEVDVGLSD	GSYARGVPS	GASTGIYKAL	ELKGGGSDYL	GEVYLRVSN	VHNIIGPAIV	GRCPTEQVKI	DHPVQGLDG	TIRHWKMKQ	KLGAHAILAY	
130	140	150	160	170	180	190	200	210	220	230	240	
ELAVKAGAM	VKIKFLYQHI	ARLAGKILV	LPVFAFVNI	GQSHAGKLA	MDFHILPTQ	ASFPKAMQ	GVVYNSLKS	IIRKQYQDA	TVVDEGGFA	PNIQENRGL	ELLKAAIEKA	
250	260	270	280	290	300	310	320	330	340	350	360	
GYTKVQVIM	DVAASEPFG	KDKTYDLRF	KENRGGSKI	SGDELKDLK	SPVSEYFES	IEDPFDQDM	STYAKLDEI	GQKQVIVSD	LLVTHPTVA	KAINERTMA	LLLVKVIQK	
370	380	390	400	410	420	430	440	450				
<b>PTESIEAWQ</b>	<b>SIKAGQWVA</b>	<b>SHRSGETED</b>	<b>FIADLSVGLS</b>	<b>TQIKTGAPC</b>	<b>RBERLAKYND</b>	<b>LLRIEELQD</b>	<b>AAVYAGAKR</b>	<b>APVEPT</b>				
G) Acidic endochitinase												
10	20	30	40	50	60	70	80	90	100	110	120	
MEKFMGLAL	SPVYFLFTG	TLAQVGSIV	TSDFDQMK	HNRDARCFV	KPFTYDAFIA	AANSFPQPT	TGDDTARQK	IAAFPQDTE	ETTGGLSPD	GPFAGGYCFL	REGNQKNGY	
130	140	150	160	170	180	190	200	210	220	230	240	
YURGFILQT	QSTILAGRA	IEQDLVHMD	LVATDAVSE	KTALWYMET	QENKPFCEV	IYRWTFSA	DTSANKVPI	<b>GVITRIIRKQ</b>	<b>IEQGRQGNK</b>	VKLRIGTYR	<b>HVSI</b> IMVAPC	
250	260											
ENLDCYQRE	FAEV											
H) Oxygen-evolving enhancer protein 2												
10	20	30	40	50	60	70	80	90	100	110	120	
MASTSCFLQ	STAKLAASR	PAPAVRGTG	FVYKAQNDK	AAEDAAYVS	KNAALSLAG	AAAIKVVSP	AAAAYGAAN	VFGARQDND	FVAYSQEGK	LMIKAMNPS	<b>KERSFPQVLI</b>	
130	140	150	160	170	180	190	200	210	220	230	240	
RYEDNFDAT	<b>HLSEVIMPTT</b>	<b>KKTITDQSF</b>	<b>EEPLQGVFL</b>	<b>LQGVYQSGT</b>	<b>DREGGPESA</b>	VATANVLESS	APVYQGVVY	SITVLTAD	<b>QDEQGGQLI</b>	<b>TAVVADGKLY</b>	<b>VCKAQKQNR</b>	
250	260											
FRGAKIVSE	AAQSFVA											
I) Double-stranded RNA-binding protein 5												
10	20	30	40	50	60	70	80	90	100	110	120	
MYRQLQLLA	QRSCFSLPY	VCTREGPDA	FRFKATYFR	GETFDGPRK	<b>ETLRQAEHAA</b>	AEVALARLKL	RGFESLTA	VLEDGTGTR	LLQETARAG	LKLPYTYTR	SQPSISPVF	
130	140	150	160	170	180	190	200	210	220	230	240	
STVELAGMF	AGDFAKTGH	AEKRAAMAAW	ESLQKSHRT	TVSFLVFDLV	VIYCHGGVF	VYVMSAQA	QGARRRRRR	AGACRRQQA	RRAEAGRLR	RNRWNRQGS	GVSTTEASR	
250	260	270	280	290	300	310	320	330	340	350	360	
KVLFVIRVAY	QTSMSAYAA	AAGRTQDTA	SPAPAAAAAS	GGVLFRRR	RAGARAERK	RAGAAARRH	<b>AARQGRRNA</b>	<b>AADAVLLRAY</b>	LPRRRRRR	EALRRRRRV	RAAGKRRPT	
370	380	390	400	410								
DPQLKRAVA	AAAAAQGGRS	SDLIQGVVV	VKTSIQSLAQ	PAPI								

This table shows the description of sequences of the identified proteins (A-I) where the peptide sequences are indicated in red colour.

The distribution of the peptide sequences in the respective protein is equivalent to the number of peptides identified for each protein (Table 4.5). For instance, Fructose-Bisphosphate Aldolase and ATP synthase subunit beta are represented each with five peptide sequences each, RuBisCO with four peptide sequences and all the rest contain two sequences each, except OEE2 and double-stranded RNA-binding protein 5 which contain only one peptide sequence each.



## 4.4 Discussion

### 4.4.1 Functional categories of the proteins identified using MALDI-TOF-MS/MS

#### 4.4.1.1 Energy generation: proton (H<sup>+</sup>) transporting protein

ATP synthase subunit beta (EC=3.6.3.14; *Saccharum hybrid* (Sugarcane); ATPB; spot 12) is a protein encoded on plastid, chloroplast. ATP synthase subunit beta, which is also known with the name ATP synthase F1 sector subunit beta or F-ATPase subunit beta was upregulated under drought stress with observed and theoretical MW/pI 60.9/5.2 and 53.9/5.2 respectively and with MOWSE score 204. Based on the gene ontology (GO) functional information, as inferred from electronic annotation (IEA), this enzyme is associated to biological process both in ATP hydrolysis and plasma membrane ATP synthesis coupled proton transport. ATP synthase generate ATP from ADP using proton gradient across the membrane with the beta subunits serving as catalytic sites (UniProt Consortium, 2013).

It has been previously identified that the binding of proteins to their interacting proteins may cause functional or positional alteration and probably leading to the formation of protein complex (van Hemert *et al.* 2001). However, activities of plasma membrane H<sup>+</sup> ATPase (Borch *et al.* 2002) and chloroplast and mitochondrial ATP synthase (Bunney *et al.* 2001) are regulated by regulatory proteins such as 14-3-3 protein (Yan *et al.*, 2004). ATP synthase beta subunit is involved in the biological process through chloroplast thylakoid membrane and with cellular component in forming proton-transporting ATP synthase complex.

Of the 70% chloroplastic proteins identified in the sorghum leaf proteome map, ATP synthase subunit beta 2 was observed to be upregulated as a response to water stress, the extent of dehydration at 30% FC where water availability become the major constraint falling below the needs of cellular functions of the plant. The increased abundance in the ATP synthase subunit beta in the treatment samples may represent its drought stress responsiveness being expressed at 60 kDa and pI 5.2 (Figure 4.5; Figure 4.6; Table 4.5 and Table 4.6). This result go in agreement with other reports on drought stress in wheat (Jiang *et al.*, 2012), in Arabidopsis (Ndimba *et al.*, 2005), and in sunflower leaf tissue (Atkin and Macherel, 2009). Ngara *et al.* (2012) has also identified the

upregulated ATP synthase subunits in sorghum in response to salt stress. The overexpression of ATP synthase beta subunit, shown in this work might be an indicative evidence for the requirement of an increase in energy level to bypass the window of stress.

Following the extent of drought stress, ATP synthase expression produces chlorplastic ATP to generate energy. Excitation of chloroplast by the light absorption tend to produce reducing substances to alter the redox status of the cell and to maintain the cellular energy balance. However, these substances are transported by ATP synthase subunit by forming the ATP synthase complex to mitochondria where ATP is either phosphorylated or produced as heat to boost energy for the cell (Atkin and Macherel, 2009). Reoxidization of chloroplast produced reducing substances require two important optional processes to take place, first in plastid itself exemplified by antioxidant systems and second in the mitochondria by the photorespiratory cycle (Kromer, 1995) where the involvement of Calvin cycle is dependable (Atkin and Macherel, 2009). The notion of the two processes is agreeably supported by the chemiosmotic coupling, the mechanism of coupling electron transport to ATP generation, where by energy stored in the form of proton gradients across biological membranes is transferred to ATP rather than a direct transfer of high-energy containing chemicals (Mitchell, 1973). The electron transport coupling to ATP generation is applicable not only in mitochondria but also in chloroplast where proton gradient allow ATP synthesis across the plasma membrane (Boyer, 1997; Cooper and Hausman, 2000). This phenomenon signifies the integrative operation of subcellular structure in mechanism of ATP and oxidative phosphorylation and the biochemical pathways associated to the processes.

The production of ATP caused as a consequence of the freely transferred energy along the chain of electron transport system (ETS) and oxidative phosphorelation is the functional relay of several protein-enzyme-substrate-complexes working together. The two important final steps of the physically separated, however functionally intact ETS and oxidative phosphorelation are catalysed by cytochrome c oxidase subunits (complex IV) and by proton translocating ATP synthase subunits (Complex V); oxidative phosphorylation metabolic pathway, refer to chapter 2 section 2.3.2, Figure 2.6 and Table 2.3 of this thesis). Interestingly, these key protein enzymes which are responsible for the control of ETS and oxidative phosphorelation were identified to be drought responsive in this study employing *in silico* functional genomics and experimental proteomic approach respectively.

In other words, the ATP synthase beta subunit identified in this study is playing the rate limiting step in the oxidative phosphorylation pathway which is tightly coupled with the ETS (Hejl *et al.*, 1993).

One of the two important drought responsive protein enzymes identified *in silico*, cytochrome c oxidase subunit 1 (COX1) also referred to as cyt c-oxidase or oxidase (EC:1.9.3.1; COX1; SbioMp23; chapter 2 section 2.3.2, Figure 2.6 and Table 2.3), is known to catalyse the final reaction of the ETS which is essentially irreversible, thus acting as a rate limiting (control point) for electron transport (Cadenas and Davies, 2000; Namslauer *et al.*, 2003). The other protein enzyme, inorganic diphosphatase (EC:3.6.1.1; encoded by a group of genes ; chapter 2 section 2.3.2, Figure 2.6 and Table 2.3) catalyses the conversion of diphosphate into inorganic phosphate, an important substrate for the chemical reaction of the ATP phosphorylation (Mitchell, 1966 and Beard, 2005). Inorganic diphosphatase regulates the rate limiting reaction played by ATP synthase subunit beta by way of limiting the amount of inorganic phosphate (Pi) that should be coupled with Adenosine Dinucleotide Phosphate (ADP) in the last step of oxidative phosphorylation. However, these phosphorylative substrates are stored inside the mitochondrial matrix (Affourtit *et al.*, 2012; Traba *et al.*, 2011) while the site of ATP uptake is fairly outside (Heldt and Flügge, 2013). The impermeability of the mitochondrial membrane to ADP and Pi poses another challenge (Nicholls and Ferguson, 2013; Cohen and Venkatachalam, 2014) whereby the need for shuttle system to transport them across the mitochondrial membrane is crucial. This is where ATP synthase subunits come into action as the ultimate rate limiting, final step of oxidative phosphorylation (Osellame *et al.*, 2012).

#### **4.4.1.2 Glycolysis and other carbohydrate metabolism associated proteins**

Fructose-Bisphosphate Aldolase (EC: 4.1.2.13; spot 14; *Oryza sativa* subsp. japonica) and Enolase 1 also referred to as phosphopyruvate hydratase (spot 12; *Zea mays*) are the two identified and functionally grouped under the category of proteins associated to glycolysis and other carbohydrate metabolism. Fructose-Bisphosphate Aldolase is responsible for the metabolic process of glycolysis and gluconeogenesis catalysing the conversion of D-fructose 1,6-bisphosphate into glycerone phosphate and D-glyceraldehyde 3-phosphate (Cooper *et al.*, 1996; Sáez and Slebe, 2000). On the other hand, Enolase 1 (EC:4.2.1.11; Sb01g040040 and Sb02g023480) also involving in the carbon metabolism is responsible for the conversion of 2-phosphoglycerate (2-PG) to phosphoenolpyruvate

(PEP) in the processes of glycolysis and gluconeogenesis (Pagliaro *et al.*, 1989; Lakshmanan *et al.*, 2013).

A chloroplastic Fructose-bisphosphate aldolase (Fructose 1,6-bisphosphate aldolase) also referred to as aldolase or fructose 1,6-bisphosphate D-glyceraldehyde 3-phosphate lyase is one of the proteins putatively identified in this study with the highest MOWSE score 209.4 at the MW 42kDa and pI 6.5 and with 5 peptides. Plants cells localize fructose 1,6-bisphosphate aldolase in plastid (chloroplast) endosymbiotically (Henze *et al.*, 1995), in addition to its commonly known cytosolic (cytoplasmic) subcellular localization (Yamazaki *et al.*, 2004; Lao *et al.*, 2013). Fructose-bisphosphate aldolase (Figure 4.6 and Table 4.5) in this experiment was found to be down-regulated in sorghum Btx642 leaf samples treated under post-flowering drought stress ( $30\pm 5\%$ FC). The decreased abundances of the fructose 1, 6-bisphosphate presumably suggest the suppression of photosynthetic apparatus whereby chloroplast and mitochondria might have experienced higher levels of oxidative damage under drought stress. This down regulations of the enzyme might indicate the scarcity of secured concentration of CO<sub>2</sub> for assimilation. Studies have earlier shown that plants have developed precautionary mechanisms to drought stress in first phases of dehydration mostly by decreasing the photosynthetic activities and starch content using stomatal closure and by altering the level of sugars and starch (Haake *et al.*, 1998) however gradually regaining as dehydration continued by passive reopening (Schwab *et al.*, 1989). The down regulation of Fructose-bisphosphate aldolase in this study seems to selectively slow or shut down a metabolic pathway which results stress worsening. Fructose-bisphosphate aldolase is the enzyme that catalyse a reversible reaction to allow gluconeogenesis which normally utilize high energy input. Thus, at time of stress, drought-tolerant plants use the mechanism of slowing down the expression and abundance of this protein enzyme to block gluconeogenesis and conserve the energy for the cell. Similar result has been reported for fructose-bisphosphatases aldolase in the study conducted to examine the expression profiles of Arabidopsis genes and gene products under drought stress (Seki *et al.*, 2002; Kilian *et al.*, 2007); under salt and osmotic stresses (Ndimba *et al.*, 2005) and in the transcriptional analysis of drought-responsive genes based on signal transcription and biochemical pathways in tomato (Gong *et al.*, 2010). As sorghum is a C<sub>4</sub> plant, this result might also reflect the eventual impact of C<sub>4</sub> photosynthetic pathway under sever drought stress and hot temperature.

However, as opposed to aldolase (Spot 14), Enolase 1 (spot 12) has shown significant increase in subunit protein abundance when sorghum leaf tissues were treated under drought condition at the post-flowering stage in contrast to the control group (Figure 5.6). As a glycolytic enzyme, the up-regulation of the Enolase1 under drought stress may seem to regain or maintain the active functioning of glycolysis by counter acting the increased effect of aldolase on glycolytic flux which otherwise may lead to the complete drop of the energy level that could cause of programmed cell death (Apoptosis; Colussi *et al.*, 2000). Early studies have indicated the role of enolase in the formation of sugars from three-carbon molecules in photosynthesising leaves of C4-plants via glycolytic conversion reaction (Karpilov *et al.*, 1977). Increased activity of enolase has been noted in NADP-ME and NAD-ME subtypes, to provide the necessary PEP by converting PGA for C4 photosynthetic pathway (Monson, 2003; Sage, 2004). Thus, the upregulation of Enolase1 in this study presumably implicated in the C4 metabolic pathway.

However, several findings have also shown that Enolase is a multi-functional enzyme other than its actual glycolytic functions (Pancholi, 2001; Entelis *et al.*, 2006). In *Saccharomyces cerevisiae*, Enolase acts as an RNA chaperone associating cytosolic tRNA with mitochondrial surface and as Hsp48p signifying functional correlation with the cell wall (Entelis *et al.*, 2006). It has also been revealed that in plants a bi-functional Enolase take part in the control of cold (Lee *et al.*, 2002) and drought-responsive (Knight and Knight, 2001) gene transcription (CBF/DREB1). We also described previously in the chapter one of this thesis (section 1.3.3; Figure 1.3) that DREB1 and DREB2 represent decisive components in the cross-talk between cold and drought signalling because they use DRE element as the common transcription factors binding site in the promoters of gene *RD29A* (Knight and Knight, 2001). Several previous studies have identified that many of glycolytic enzymes including Enolase have acquired more non-glycolytic functions in signalling transduction and transcriptional regulation (Kima and Dang, 2005). The up-regulation of enolase 1 in this experiment, therefore, presumably suggest a nonglycolytic actions of this multifunctional enzyme in addition to its innate activity, hence we speculate its involvement either in the coordination of cross-talk between cold and drought or separately in each signal transduction pathways.

In *Arabidopsis thaliana*, enzymes of glycolysis are present on the surface of mitochondria and free in the cytosol. The functional significance of this dual localization has now been established by

demonstrating that the extent of mitochondrial association is dependent on respiration rate in both *Arabidopsis* cells and potato (*Solanum tuberosum*) tubers (Grahama *et al.*, 2007).

Fructose-bisphosphate aldolase (EC:4.1.2.13; Sb03g008050) takes part in rate limiting (Wong and Whitsides, 1994) steps both in gluconeogenesis and glycolysis (Cooper *et al.*, 1996). The synthesis of new glucose necessitates the reduction of phosphoenolpyruvate (PEP) to fructose 1,6-bisphosphate which is catalysed by aldolase as the last reaction of gluconeogenesis (Horecker *et al.*, 1972; Sáez and Slebe, 2000). In case of glycolysis, the last product of gluconeogenesis, fructose 1,6-bisphosphate is used as a substrate to be oxidized back to PEP in which case aldolase play the first reaction catalytic role in glycolysis (Cooper *et al.*, 1996). The aldolase that is involved in the gluconeogenesis and glycolysis occurs both as chloroplastic and cytoplasmic protein. Two type aldolase classes are known to exist (Cooper *et al.*, 1996) with further division in to sub-aldolase-classes (Walther *et al.*, 1998). This suggests that different sorts of aldolase isomeric enzymes are necessary to control the catalytic reactions of the pathways in glycolysis and gluconeogenesis (Sáez and Slebe, 2000).

#### **4.4.1.3 Regulation of photosystem (carbon assimilation)**

Ribulose Bisphosphate Carboxylase (RuBisCo, EC:4.1.1.39; Spot 13; Sb03g020182; Sb05g003480; Sb05g025125; Sb05g025130; Sb08g001646) is among three identified protein enzymes in the functional category of photosystem regulation (carbon assimilation; pathway ID:ec00710). RuBisCo (*Lactuca sativa*) was upregulated in this experiment in the sorghum leaf tissue treated under drought condition at the post-flowering stage (Figure 4.5) and is responsible for the carbon fixation (Berg *et al.*, 2010; Peretó *et al.*, 2010), a special ingredient in dicarboxylate metabolism (pathway ID:ec00630) with a bipartite relationship with glycolysis (Heymans and Singh, 2003). That RuBisCo large chain displayed upregulated expression profiles in this study goes in agreement with a previously demonstrated work (Spreitzer and Salvucci 2002). RuBisCo is treated under this functional category for its outstanding performance in carbon assimilation by catalysing carbon fixation in one of the major metabolic process in the photosynthetic organism (Bar-Even *et al.*, 2010; Bauwe *et al.*, 2010; Hohmann-Marriott and Blankenship, 2011). RuBisCo has been previously shown to be involved in a glucose metabolism (glycolysis) following a pathway that bypasses extra-glycolytic reactions in generation of pyruvate from D-fructose-6-phosphate and supplementing CO<sub>2</sub> at the minimal cost of energy expenditure from hydrolyzed ATP (Karpinets *et*



*al.*, 2014). This way, it looks that RuBisCo plays stress tolerance role by minimizing energy uptake of a cell at time of stresses when metabolising carbohydrate contributing to the amount of energy reserved for use. This mechanism definitely assists the plant staying alive, probably in association with stay-green genes (Thomas *et al.*, 2000). This function of RuBisCo seems to correlate with the one that Fructose-Bisphosphate Aldolase does in conserving energy pool in drought-stressed plants by inhibiting gluconeogenesis that otherwise consume high energy (Gong *et al.*, 2010) leading the plant cell to oxidative damage and apoptosis (programmed death; Apel and Hirt, 2004; Ott *et al.*, 2007).

Oxygen-evolving enhancer protein 1 (OEE1, *Fritilaria agrestis*) is among other identified proteins functionally categorized in the carbon assimilation (pathway ID:ec00710). Oxygen-evolving enhancer protein 1 (spot 8) was expressed experimentally and theoretically at the MW/pI level 34.8/5 and 34.8/6.3 respectively with MOWSE score 74.1 and was observed to be upregulated in the sorghum Btx642 leaf tissue under drought stress in the treatment group than in control group. The implicated over-expression of OEE1 following post-flowering drought stress in sorghum leaf suggests its role in the photosystem II complex having strong association with photosynthesis. Other studies have also shown similar result when evaluating the leaf tissue from the Chinese spring and needles of maritime pine (Bahrman *et al.*, 2004) in drought stress.

Oxygen-evolving enhancer protein 2 (OEE2; Photosystem II (P680 chlorophyll a; *Petunia hybrida*) also referred to as photosystem II oxygen-evolving enhancer protein 2 (Pathway ID: sbi00195; biochemical pathway Photosynthesis; Sb01g049040; Sb02g002690) is the third protein categorized into the functional group of carbon assimilation (photosystem regulation). A 23.9/6.4 kDa/pI Spot 16 was increasingly expressed to have matched with OEE2 (*Triticum aestivum*) with a MOWSE score 35.31 and was up-regulated with functional correlation with OEE1. The increased abundance of OEE2 suggests the establishment of photosynthetic activity through maintenance of the evolution of oxygen from the system in drought stressed leaves. Other study has previously demonstrated that OEE1 take a key position in oxygen evolution and stability of PSII (Sugihara *et al.*, 2000).

The relative abundance of the two subunits was observed to be comparable with seemingly slightly ubiquitous nature of subunit 1. The two proteins were previously identified to be encoded by

nuclear-chloroplast bound to be photosystem II (PSII) peripheral location on the thylakoid membrane of luminal region (Sugihara *et al.*, 2000). The simultaneous increased expression of the protein spots of the two subunits OEE1 and OEE2 entails their integrative and coordinated functioning of the photosystem II (PSII) in accelerating carbon assimilation in the drought stressed sorghum leaf tissue. Results from the experiment of salinity stress showed the concurrent increase of the two proteins with increase in the quantity of PSII centre (Sugihara *et al.*, 2000; Kim *et al.*, 2005) though reduction in gross photosynthesis was exhibited following increased salinity (Takemura *et al.*, 2000). This implicate that the OEE1 and 2 protein subunits actively respond to multiple stresses to maintain functioning of photosynthesis.

#### **4.4.1.4 RNA-Binding protein**

Double-stranded RNA-binding protein 5 (EC:2.7.7.49; spot 14; *Oriza sativa* subsp.japonoca, Figure 4.7; Table 4.5) is the heterogeneous nuclear-riboneocleoprotein particle functionally known to bind double-stranded RNA (Fierro-Monti and Mathews, 2000; Buratti and Baralle, 2001). A 43.1kDa protein spot which was migrated at the pI 5.6 with a relatively lower MOWSE score was identified to be down-regulated in sorghum Btx6642 leaf sample under differential drought stress compared to the control plant samples. Some studies have indicated that this protein has significant role in detecting and controlling the viral genome (Yoneyama *et al.*, 2004) and in microRNA-mediated gene regulation (Han *et al.*, 2004). Study conducted using *Arabidopsis thaliana* proves glycine-rich cold-inducible zinc finger RNA-binding protein to contribute to the enhancement of freezing tolerance (Kim *et al.*, 2005). However, as to our knowledge no report has been produced to indicate the functional role of double-stranded RNA-binding protein 5 in association with post-flowering drought responses in sorghum leaf tissues. A significantly decreased abundance of the double-stranded RNA-binding protein 5 shown in this result suggests that this protein does not play significant role in responding to drought stress in sorghum Btx642 leaf tissues under post-flowering drought stress. Alternatively, the extent of dehydration ( $30\pm 5\%$ FC) imposed to induce drought stress might not have been sufficient to induce the expression of this protein.

#### **4.4.1.5 Stress tolerance, defence and immunity related proteins**

Acidic endochitinase (EC=3.2.1.14; Sb07g027310; spot 9; Figure 4.7 and Table 4.5) that was identified in this experiment is the only protein classified as having functional correlation with defence against chitin containing fungal pathogens (Grover, 2012). Based on the gene ontology

information, this protein involve in chitin and polysaccharide degradation in addition to plant defence activity (Grover, 2012; Hartl *et al.*, 2012). Acidic endochitinase has been identified from an increasingly abundant spot 9 experimentally at the MW/pI 39.7/5.6 by sequence homology search against non-redundant protein database. The search result identified the highest match with *Petunja hybrida* protein with MOWSE score 39. Acid endochitinase has been observed to be up-regulated in sorghum leaf tissue under post-flowering drought stress entailing strong functional correlation of this protein with drought response. Other studies have reported similar results by identifying Abscisic Acid and osmotic stress-responsive genes encoding acid endochitinase from *Lycopersicon* and by functionally characterizing the protein in the traps of the carnivorous pitcher plant genus depicting positive association with drought stress (Chen *et al.*, 1994; Rottloff *et al.*, 2011).

#### **4.4.1.6 Unknown (Hypothetical proteins)**

Hypothetical protein SORBIDRAFT\_01g012710 (gi|242038635; spot 1) was identified in this study as putatively uncharacterised protein with an observed MW/pI 23.9/4.3, 2 protein peptides and MOWSE score 50.3. As the only one with unknown functional category, much is not known about the role of this protein albeit sorghum gene prediction annotation (Paterson *et al.*, 2009) serve as preliminary information. Database search for the sequence similarity of the microsequence of the protein spot 1 as a query showed the top match with a sorghum protein, the hypothetical protein SORBIDRAFT\_01g012710, encoded by the gene 'Sb01g012710'. This putatively uncharacterised protein was found to be upregulated in sorghum leaf tissue of the Btx642 variety when treated under post-flowering drought compared to the well watered plant samples. The increased abundance of this protein implicate the involvement of this protein in response to drought stress. Ontological and gramene/ensembl plants based information suggest defence response to fungus and incompatible interaction and response to salt stress to be hypothetical function of this protein (UniProt Consortium, 2013). All this information suggest that the protein responds to multi-stresses including drought tolerance as the current study implicate.

#### 4.5 Conclusion

This study represents experimental investigation of proteomes responsive to drought stress by employing expressional proteomics. The differential expressional profiles exhibited in this research led to the identification of key protein enzymes that play central roles in metabolic pathways that determine the drought phenotype of the stressed plants thus leading to functional proteomics. A long standing application of 2DE for separation and visualization of proteomes and the Bruker ultraflex based MALDI-TOF-TOF-MS/MS gave the complete hightech proteomic platform for the analysis of this research and establishment of sorghum leaf proteome map. Many experimentally identified proteins were shown to be correlated with *in silico* identified, for example ATP synthase subunit beta (complex V) which is the rate limiting enzyme for oxidative phosphorelation pathway correlated with the cytochrome c oxidase subunit 1 (cox1, complex IV), a rate limiting enzyme for the ETS pathway. This study serves as a validative work for *in silico* genomic identification of gene and gene products signifying the importance and implication of protogenomics in agricultural research. The result of this study can therefore be used as a model for all genomic research combining with proteomics. The identified drought-induced genes and gene products can be used in breeding programmes to enhance crop productivity. In total, the current result can be used as a stepping stone towards agricultural achievements through proteomic research using sorghum as a model crop.

UNIVERSITY of the  
WESTERN CAPE

## CHAPTER 5

### Chapter 5: General discussion and conclusions

#### 5.1 Summary

Water deficit is the most important constraint to sorghum productivity worldwide. A broad genetic basis and natural variability added to the C4 photosynthetic pathway provides for physiological and biochemical plasticity in sorghum to adapt to drought affected regions. Plant adaptation and specifically sorghum, is a complex phenomenon that involves a chain of processes from the stress specific activation of signalling cascades to the expression of specific stress-induced genes and gene products. Several studies have indicated that extracellular signal perception activates the downstream intracellular signal cascade generating the second messenger (Haung *et al.*, 2012 and Hubbard *et al.*, 2012). This triggers up-regulation of cytoplasmic calcium levels facilitating the interaction of calcium binding proteins (Ca<sup>2+</sup> sensors) with the down stream signalling components. This initiate a transduction (phosphorylation) cascade to target the major stress responsive genes or the transcription factors that regulate these genes then produce the gene products that are involved in plant adaptation and survival. The study described in this thesis focused on genomic and proteomic analyses of drought tolerance to identify potential drought responsive genes and gene-products, and discover novel drought related functions encoded by the sorghum genome. We aimed at providing a better understanding of drought-related genes that are switched on in response to drought stress.

The points of discussion in this chapter are divided into four sections. The first section deals with the *in silico* identification of candidate genes for drought tolerance in sorghum and the discovery of novel genes using *ab initio* and an extrinsic evidence approach. The second section explores the gene-gene and gene-phenotype association for drought related determinants in sorghum, and the third section deals with identification of proteins using differential expression profiling and MALDI-TOF-TOF MS/MS analysis. Lastly, the concluding chapter combines these three chapters and discuss the implications.

#### 5.2. *In silico* Identification of candidate genes for drought tolerance

The *In silico* candidate gene approach (InsCGA) was successful in identification and prioritization

of drought responsive genes. Identification of candidate genes by mapping experimental data to a reference genome is a promising approach because it allows for identification of genomic regions associated with complex drought tolerance. For example, the tissue attributes of the UniGene dataset (9258 DRESTs) were exploited to identify 123 drought responsive candidate genes.

With the complete sequencing and annotation of the sorghum genome, it was possible to assign the coding regions of the majority of sorghum genes to metabolic functions (Paterson *et al.*, 2009). The fourteen metabolic pathways identified in this study represent the biochemical basis of functional networks of the 477 sorghum genes involved in these pathways in general and the 32 drought specific genes that are responsible for encoding enzymes that catalyse substrate conversion in particular. Analysis of the metabolic pathways revealed a functional set of genes involving a multiplex metabolic role across the pathways and within a pathway. For example, Mitogen-Activated Protein Kinases (MAPKs) regulate similar multiplex roles in plant signalling pathways mediating the functioning of cellular responses to drought and salt stresses (Munnik and Meijer, 2001). By and large, pathway analysis revealed a wide array of potential candidate genes interacting in response to complex drought stress.

GO annotation and Interpro domain analysis demonstrated a high frequency of protein domains related to drought tolerance that exist as common elements in plants. Analysis of gene expression reflected conserved functional similarity in drought stress responses between species in agreement with the previous finding that co-expression of genes is largely conserved between orthologs (Dutilh *et al.*, 2006). At least 50% of all sorghum genes associated with drought phenotypes were identified in maize, rice and Arabidopsis suggesting an evolutionarily conserved mechanism to external stresses. Yet, 10% of sorghum drought responsive genes were not identified in other cereals.

The results of the current study reveals the identification of novel gene structures and single gene model updates from existing annotation that include novel exons, five and three prime UTRs. These gene updates will be submitted to sorghum database as GFF formatted files. The value of an annotation comparison could be realized both in annotation of different versions of the same genome and different sources of distinct gene structure prediction pipelines (Standage and Brendel,

2012). These provides opportunities for comparative genomics to use orthologous groups in different genomes to find similarities and differences between organisms by comparing genome sequences and exonic variants (Singh *et al.*, 2008). On the other hand, in this thesis, novel sorghum genes were predicted using the combination of both *ab initio* and extrinsic evidence (Stanke *et al.*, 2006 and Haas *et al.*, 2008) of which 69% were shown to be DR. Recently, the use of a combination of sequence similarity search and *ab initio* approaches is the most concern in functional genomics in predicting the complete gene structure models with sets of exons that can be spliced (Leroy *et al.*, 2012; Kornblihtt *et al.*, 2013; Castellana *et al.*, 2014). The present study has also identified novel genes with single exonic (intron less) feature (47.7%) with high probability to function in drought tolerance based on the fact that most intronless genes were shown to be drought inducible (Haake *et al.*, 2002 and Akhtar *et al.*, 2012).

### **5.3 Gene-gene and gene-phenotype association in sorghum drought tolerance**

Cellular and whole-plant level complexity in tissues and developmental stage-specific physio-biochemical processes have been the most challenging in understanding the genetic foundation of drought tolerance. This challenge is best solved by understanding the association between drought phenotype of the plants under drought stress condition and the typical drought tolerance pathway. Recent advances in functional genomics has provided opportunity to bridge the gaps (Farooq *et al.*, 2009 and Pagariya *et al.*, 2011). Identification of the genes with functional association across environmental stresses and among species cross-talk using an integrated approach was the main objective of this chapter. This study identified 34% of the 169 sorghum drought responsive genes with functional conservation in other species (maize, rice and Arabidopsis) and functional roles in multi-stress (cold, heat, salt and ROS) tolerance. This confirms that sorghum has a potential value as the model in comparative genomics and in agricultural research. In line with other studies (Yue *et al.*, 2006), the current analysis partitions the whole-plant resistance into separate components of functionally enriched drought expressed genes namely drought tolerance (369), drought avoidance (147) and drought escape (1950) genes. Functional ontology mapping revealed more than 50% genes with transitive association (Vinayagam *et al.*, 2004 and Mungall *et al.*, 2010) suggesting the importance of orthology groups in discerning genetic dissection of complex drought tolerance. Integrating gene expression profiling and ontology related data revealed 48% tissue specific association of drought responses implicating the power of expression data in determining gene-

phenotype association. This has also been demonstrated by several previous works (Pillitteri *et al.*, 2011).

#### **5.4 Identification of drought responsive proteins using DE profiling and MALDI-TOF-TOF MS/MS**

Sorghum Btx642 is a known post-flowering drought tolerant variety and a source of stay-green genes (Jordan *et al.*, 2012). This study aimed to identify drought responsive proteins using Btx642, a terminal drought tolerant sorghum variety. The result of this data has contributed to nine key protein enzymes (78%-upregulated) that play a central role in the biochemical processes and metabolic pathways that determine drought phenotypes. Identification of key up-regulated proteins suggest their role in post flowering drought tolerance entailing functions related to stay green genes that regulate post-flowering terminal drought stress (Harris *et al.*, 2007).

Of the protein enzymes identified in this study, ATP synthase subunit beta (spot 12) and Fructose-Bisphosphate Aldolase (spot 14); Enolase 1 (spot 12); Ribulose Bisphosphate Carboxylase (RuBisCo; spot 13); Oxygen-evolving enhancer protein 1 (spot 8); Oxygen-evolving enhancer protein 2 (spot 16) are among others involved in photosynthetic metabolic activities in various ways. ATP synthase subunit beta is a relatively large protein enzyme acting as a rotary mechanical motor (Junge *et al.*, 2009) to generate ATP, a rate limiting role of Oxidative phosphorylation (or OXPHOS) (Hüttemann *et al.*, 2008; Verdin *et al.*, 2010). Contrary to OXPHOS vitality in plant metabolism, studies have indicated the production of Reactive Oxygen Species (ROS, eg. superoxide and hydrogen peroxide) that can cause the spread of free-radicals, probably leading to apoptosis through damage of cells contributing to senescence in plants (Fleury *et al.*, 2002). However, enzymes responsible for the OXPHOS pathway are the potential target of drugs and poisons that inhibit their activities in attempt to maximize antioxidant role (Wallac *et al.*, 2000; Roy *et al.*, 2008). This data is similar to the result we identified *in silico*, 'Drug Metabolism-Other Enzymes', catalysed by enzyme Ali-esterase (EC:3.1.1.1; chapter 2), probably a novel discovery in sorghum metabolism, the expression of which may be used as a BIOSENSOR to evaluate hazardous compounds induced due to insecticidal contamination or other environmental pollution in plants (Carvalho *et al.*, 2003).



The different classes of sub-cellular localization of these proteins indicates the type of function they are associated with (Smith *et al.*, 2004) suggesting chloroplastic enzymes in this research have got to do with photosynthetic metabolism in conferring drought tolerance in plants. Moreover, this study showed the functional correlation of the key enzymes identified in this experiment with many of the enzymes identified *in silico*, that can be taken as a validation for *in silico* genomic approach.

### **5.5 Conclusive remarks and future research plan**

The current work has identified comprehensive catalogues of significantly enriched candidate multi-stress responsive genes. Since the analysis in this thesis includes distinctly different studies on genes and proteins related to drought responses, it is worth integrating and comparing the results across the different studies. Based on InsCGA (chapter 2), a total of 612 known and novel drought responsive genes were identified using mapping expression data to reference genome (306 genes), pathway analysis (28 genes), expression profiling (11 genes) and analysis of orthology relation (260 genes). Seven (1.1%) genes were commonly identified in all analytical approaches used in InsCGA. On the other hand, among a total of 1831 genes identified using the gene-gene and gene-phenotype association study (chapter 3), 1743 functionally enriched drought responsive genes were identified by integrating five plant ontologies using semantic query components and 88 DRGs using expression profiling with 77 genes commonly identified by all approaches used in the gene association study. Comparing the results from the different analytical approaches used to identify drought responsive genes in the two chapters, 89 genes were found to have overlapping biological functions. Further more, nine drought responsive proteins which are functionally novel in sorghum were identified in the 4<sup>th</sup> chapter using differential expression analysis and MALDI-TOF-TOF-MS/MS. Most of the genes and proteins identified were shown to have multiple stress tolerance. For example, based on InsCGA approach, 32 genes (~4%) which were identified in pathway analysis and 265 genes (~ 43%) in orthology groups were all associated with multiple responses across environmental stresses. Likewise, 34% of the total DRGs identified using gene-gene and gene-phenotype association have demonstrated multiple stress resistance that suggests a wide range of biological functions and prevalence of adaptive traits across environmental heterogeneity. This study has shown apparent concordance between the *in silico* and the experimental results that account for valid functional components and the quality of our data. Specifically, the results in this study signify the empirical value of integrative approach to identify drought responsive genes that

would not otherwise be captured by using just a single or few analytical approaches. Overall, this study identified a total of 2113 known DRGs, 241 novel genes and 9 functionally novel proteins in sorghum. All the novel genes identified in this study and the existing gene models modified will be submitted to the sorghum annotation database.

A future plan includes the validation of some representative known and novel genes using RT-qPCR. However, transformation of the promising sorghum candidate multiple stress responsive genes into an open pollinated farmers varieties should be the main future target to improve sorghum productivity. To do this, we propose a very concise note. Construction of promoters/reporter gene will be used to manipulate and clone DNA. Promoter GUS reporter gene fusions will be used by constructing chimerical promoters. All promoter fragments will be assembled, confirmed by sequencing, excised and transferred in front of the GUS reporter gene. The promoter reporter gene constructs will be cloned and all DNA fragments to be generated by PCR will be confirmed by DNA sequencing. Electroporation technique will be used to introduce the promoter/GUS constructs into the appropriate experimental organism (eg. *Agrobacterium tumefaciens* strain). Plants from sorghum preferred farmer's varieties will be transformed using appropriate protocol. Integration of the chimerical genes into the sorghum genome will be examined by DNA gel blot analysis or by PCR using appropriate protocol (Stockhaus *et al.*, 1997; Le *et al.*, 2012; Atkinson *et al.*, 2013).

UNIVERSITY of the  
WESTERN CAPE

## References

- Abdeen, Ashraf, Jaimie Schnell<sup>1</sup>, Brian Miki. "Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor ABF3." *BMC Genomics*, 11(2010):69.
- Abdi A, E Bekele, Z Asfaw, A Teshome. "Patterns of Morphological Variation of Sorghum (*Sorghum Bicolor* (L.) Moench) Landraces in Qualitative Characters in North Shewa and South Welo, Ethiopia," *Hereditas* 137, no. 3 (2002): 161–172.
- Abdi, A. and Z Asfaw. "In-situ (on-farm) conservation dynamics and the patterns of uses of sorghum (*Sorghum bicolor* (L.) Moench) Landraces in North Shewa and South Welo, Central Highlands of Ethiopia." *Ethiopian Journal of Biological Sciences* 4, no. 2(2005): 161-184.
- Acland, Abigail, Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A. Benson, Colleen Bollin, Evan Bolton, Stephen H. Bryant, Kathi Canese, and Deanna M. Church. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 41, no. D1 (2013): D8–D20. <http://europepmc.org/articles/PMC3531099>.
- Affourtit, Charles, Casey L. Quinlan, and Martin D. Brand. "Measurement of Proton Leak and Electron Leak in Isolated Mitochondria." In *Mitochondrial Bioenergetics*, 165–82. Springer, 2012. [http://link.springer.com/10.1007/978-1-61779-382-0\\_11](http://link.springer.com/10.1007/978-1-61779-382-0_11).
- Agarwal, P. K., and B. Jha. "Transcription Factors in Plants and ABA Dependent and Independent Abiotic Stress Signalling." *Biologia Plantarum* 54, no. 2 (2010): 201–212.
- Aggarwal, K., and H. K. Lee. "Functional Genomics and Proteomics as a Foundation for Systems Biology." *Briefings in Functional Genomics & Proteomics* 2, no. 3 (2003): 175–184.
- Agresti, Alan. *An Introduction to Categorical Data Analysis*. Vol. 423. Wiley-Interscience, 2007. [http://books.google.co.za/books?hl=en&lr=&id=gCskkCZWjyIC&oi=fnd&pg=PR5&dq=Exact+Inference+for+Categorical+Data&ots=5LWwE9Puca&sig=nRP\\_qgyc7hGItdbDvoWCP1\\_99Mw](http://books.google.co.za/books?hl=en&lr=&id=gCskkCZWjyIC&oi=fnd&pg=PR5&dq=Exact+Inference+for+Categorical+Data&ots=5LWwE9Puca&sig=nRP_qgyc7hGItdbDvoWCP1_99Mw).
- Ajithkumar, I. Paul, and R. Panneerselvam. "ROS Scavenging System, Osmotic Maintenance, Pigment and Growth Status of *Panicum Sumatrense* Roth. Under Drought Stress." *Cell Biochemistry and Biophysics* (2013): 1–9.
- Akhtar, M., A. Jaiswal, G. Taj, J. P. Jaiswal, M. I. Qureshi, and N. K. Singh. "DREB1/CBF Transcription Factors: Their Structure, Function and Role in Abiotic Stress Tolerance in

- Plants." *Journal of Genetics* 91, no. 3 (2012): 385–395.
- Alba, R., Z. Fei, P. Payton, Y. Liu, S. L. Moore, P. Debbie, J. Cohn, M. D'Ascenzo, J. S. Gordon, and J. K. C. Rose. "ESTs, cDNA Microarrays, and Gene Expression Profiling: Tools for Dissecting Plant Physiology and Development." *The Plant Journal* 39, no. 5 (2004): 697–714.
- Albert, V. A., D. E. Soltis, J. E. Carlson, W. G. Farmerie, P. K. Wall, D. C. Ilut, T. M. Solow, L. A. Mueller, L. L. Landherr, and Y. Hu. "Floral Gene Resources from Basal Angiosperms for Comparative Genomics Research." *BMC Plant Biology* 5, no. 1 (2005): 5.
- Altieri, Miguel A. "Agroecology, Small Farms, and Food Sovereignty." *Monthly Review* 61, no. 3 (2009): 102–13.
- . "Linking Ecologists and Traditional Farmers in the Search for Sustainable Agriculture." *Frontiers in Ecology and the Environment* 2, no. 1 (2004): 35–42.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215, no. 3 (1990): 403–410.
- Amiour, Nardjis, Sandrine Imbaud, Gilles Clément, Nicolas Agier, Michel Zivy, Benoît Valot, Thierry Balliau, Patrick Armengaud, Isabelle Quilleré, and Rafael Cañas. "The Use of Metabolomics Integrated with Transcriptomic and Proteomic Studies for Identifying Key Steps Involved in the Control of Nitrogen Metabolism in Crops Such as Maize." *Journal of Experimental Botany* 63, no. 14 (2012): 5017–5033.
- Andersen, Jens S., Christopher J. Wilkinson, Thibault Mayor, Peter Mortensen, Erich A. Nigg, and Matthias Mann. "Proteomic Characterization of the Human Centrosome by Protein Correlation Profiling." *Nature* 426, no. 6966 (2003): 570–74.
- Anjum, Shakeel Ahmad, Xiao-yu Xie, L. C. Wang, Muhammad Farrukh Saleem, Chen Man, and Wang Lei. "Morphological, Physiological and Biochemical Responses of Plants to Drought Stress." *African Journal of Agricultural Research* 6, no. 9 (2011): 2026–2032.
- Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S., Smith, R.D. "Proteogenomics: needs and roles to be filled by proteomics in genome annotation." *Brief. Funct. Genomic. Proteomic.* 7, (2008):50–62.
- Apel, Klaus, and Heribert Hirt. "Reactive Oxygen Species: Metabolism, Oxidative Stress, and Signal Transduction." *Annu. Rev. Plant Biol.* 55 (2004): 373–99.
- Arabidopsis, G. I. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana." *Nature* 408, no. 6814 (2000): 796.

- Ariño, Joaquín, Antonio Casamayor, and Asier González. "Type 2C Protein Phosphatases in Fungi." *Eukaryotic Cell* 10, no. 1 (January 2011): 21–33. doi:10.1128/EC.00249-10.
- Arisue, Nobuko, Nirianne MQ Palacpac, Kazuyuki Tanabe, and Toshihiro Horii. "Clues to Evolution of the SERA Multigene Family in the Genus Plasmodium." *Gene Duplication. InTech; ISBN* (2011): 978–953.
- Arnaud, Elizabeth, Laurel Cooper, Rosemary Shrestha, Naama Menda, Rex T. Nelson, Luca Matteis, Milko Skofic, Ruth Bastow, Pankaj Jaiswal, and Lukas A. Mueller. "Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes." In *KEOD*, 220–225, 2012.  
[http://wiki.plantontology.org/images/6/6e/Ref\\_TO\\_KEOD\\_2012.pdf](http://wiki.plantontology.org/images/6/6e/Ref_TO_KEOD_2012.pdf).
- Arnesano, Fabio, Lucia Banci, Ivano Bertini, and Manuele Martinelli. "Ortholog Search of Proteins Involved in Copper Delivery to Cytochrome c Oxidase and Functional Analysis of Paralogs and Gene Neighbors by Genomic Context." *Journal of Proteome Research* 4, no. 1 (2005): 63–70.
- Arpat, A., M. Waugh, J. P. Sullivan, M. Gonzales, D. Frisch, D. Main, T. Wood, A. Leslie, R. Wing, and T. Wilkins. "Functional Genomics of Cell Elongation in Developing Cotton Fibers." *Plant Molecular Biology* 54, no. 6 (2004): 911–929.
- Arrau deau MA. "Breeding strategies for drought resistance" (1989) – [agris.fao.org](http://agris.fao.org).
- Arumuganathan, K., and E. D. Earle. "Nuclear DNA Content of Some Important Plant Species." *Plant Molecular Biology Reporter* 9, no. 3 (1991): 208–218.
- Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, and Janan T. Eppig. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25, no. 1 (2000): 25–29.
- Atkin, Owen K., and David Macherel. "The Crucial Role of Plant Mitochondria in Orchestrating Drought Tolerance." *Annals of Botany* 103, no. 4 (2009): 581–97.
- Atkinson, N. J., and P. E. Urwin. "The Interaction of Plant Biotic and Abiotic Stresses: From Genes to the Field." *Journal of Experimental Botany* 63, no. 10 (2012): 3523–3543.
- Atkinson, Nicky J., Catherine J. Lilley, and Peter E. Urwin. "Identification of Genes Involved in the Response of Arabidopsis to Simultaneous Biotic and Abiotic Stresses." *Plant Physiology* 162, no. 4 (2013): 2028–41.
- Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li,

- Dazhe Meng, Alexander Platt, Aaron M. Tarone, and Tina T. Hu. "Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines." *Nature* 465, no. 7298 (2010): 627–631.
- Avraham, Shulamit, Chih-Wei Tung, Katica Ilic, Pankaj Jaiswal, Elizabeth A. Kellogg, Susan McCouch, Anuradha Pujar, Leonore Reiser, Seung Y. Rhee, and Martin M. Sachs. "The Plant Ontology Database: a Community Resource for Plant Structure and Developmental Stages Controlled Vocabulary and Annotations." *Nucleic Acids Research* 36, no. suppl 1 (2008): D449–D454.
- Ayana A and E Bekele. "Geographical Patterns of Morphological Variation in Sorghum (*Sorghum Bicolor* (L.) Moench) Germplasm from Ethiopia and Eritrea: Qualitative Characters." *Hereditas* 129, no. 3 (1998): 195–205.
- Babita, M., M. Maheswari, L. M. Rao, Arun K. Shanker, and D. Gangadhar Rao. "Osmotic Adjustment, Drought Tolerance and Yield in Castor (< i> Ricinus Communis</i> L.) Hybrids." *Environmental and Experimental Botany* 69, no. 3 (2010): 243–49.
- Bach, Inga. "Sorghum-Metabolism of Dhurrin" (2012).  
[http://plen.ku.dk/english/research/plant\\_biochemistry/natural\\_products/sorghum/](http://plen.ku.dk/english/research/plant_biochemistry/natural_products/sorghum/).
- Baena-González, E. "Energy signaling in the regulation of gene expression during stress." *Mol. Plant* 3, (2010): 300–313.
- Baginsky, S., Hennig, L., Zimmermann, P., Gruissem, W. "Gene expression analysis, proteomics, and network discovery." *Plant Physiol.* 152, (2010): 402–410.
- Bahrman, Nasser, Jacques Le Gouis, Luc Negróni, Laurence Amilhat, Philippe Leroy, Anne-Lyse Lainé, and Odile Jaminon. "Differential Protein Expression Assessed by Two-Dimensional Gel Electrophoresis for Two Wheat Varieties Grown at Four Nitrogen Levels." *PROTEOMICS* 4, no. 3 (March 1, 2004): 709–19. doi:10.1002/pmic.200300571.
- Bailey, Wendy J., and Roger Ulrich. "Molecular Profiling Approaches for Identifying Novel Biomarkers." *Expert Opinion on Drug Safety* 3, no. 2 (2004): 137–151.
- Bak, Søren, Carl Erik Olsen, Barbara Ann Halkier, and Birger Lindberg Møller. "Transgenic Tobacco and Arabidopsis Plants Expressing the Two Multifunctional Sorghum Cytochrome P450 Enzymes, CYP79A1 and CYP71E1, Are Cyanogenic and Accumulate Metabolites Derived from Intermediates in Dhurrin Biosynthesis." *Plant Physiology* 123, no. 4 (2000): 1437–48.
- Balakirev ES and Ayala FJ. "PSEUDOGENES: Are They "Junk" or Functional DNA?" *Annual*

- Review of Genetics* 37, no. 1 (2003): 123-151.
- Baldi, Pierre, and Anthony D. Long. "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized T-Test and Statistical Inferences of Gene Changes." *Bioinformatics* 17, no. 6 (2001): 509-519.
- Balhoff, James P., Wasila M. Dahdul, Cartik R. Kothari, Hilmar Lapp, John G. Lundberg, Paula Mabee, Peter E. Midford, Monte Westerfield, and Todd J. Vision. "Phenex: Ontological Annotation of Phenotypic Diversity." *PLoS One* 5, no. 5 (2010): e10500.
- Bancroft, Ian. "Duplicate and Diverge: The Evolution of Plant Genome Microstructure." *TRENDS in Genetics* 17, no. 2 (2001): 89-93.
- Bandyopadhyay, Sourav, Roded Sharan, and Trey Ideker. "Systematic Identification of Functional Orthologs Based on Protein Network Comparison." *Genome Research* 16, no. 3 (2006): 428-435.
- Bansal, K. C., S. K. Lenka, and N. Tuteja. "Abscisic Acid in Abiotic Stress Tolerance: An 'omics' Approach." *Omics and Plant Abiotic Stress Tolerance. Bentham Science, Sharjah* (2011): 143-150.
- Bar-Even, Arren, Elad Noor, Nathan E. Lewis, and Ron Milo. "Design and Analysis of Synthetic Carbon Fixation Pathways." *Proceedings of the National Academy of Sciences* 107, no. 19 (2010): 8889-94.
- Barbazuk, W. Brad, Yan Fu, and Karen M. McGinnis. "Genome-Wide Analyses of Alternative Splicing in Plants: Opportunities and Challenges." *Genome Research* 18, no. 9 (2008): 1381-1392.
- Bard, JBL, and SY. Rhee. "Ontologies in Biology: Design, Applications and Future Challenges." *Nature Reviews Genetics* 5, no. 3 (2004): 213-222.
- Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. "NCBI GEO: Mining Tens of Millions of Expression Profiles—database and Tools Update." *Nucleic Acids Research* 35, no. suppl 1 (2007): D760-D765.  
[http://nar.oxfordjournals.org/content/35/suppl\\_1/D760.short](http://nar.oxfordjournals.org/content/35/suppl_1/D760.short).
- Bartlett, Megan K., Christine Scoffoni, and Lawren Sack. "The Determinants of Leaf Turgor Loss Point and Prediction of Drought Tolerance of Species and Biomes: A Global Meta-Analysis." *Ecology Letters* 15, no. 5 (2012): 393-405.

- Bartlett, Megan K., Christine Scoffoni, Rico Ardy, Ya Zhang, Shanwen Sun, Kunfang Cao, and Lawren Sack. "Rapid Determination of Comparative Drought Tolerance Traits: Using an Osmometer to Predict Turgor Loss Point." *Methods in Ecology and Evolution* 3, no. 5 (2012): 880–888.
- Bauwe, Hermann, Martin Hagemann, and Alisdair R. Fernie. "Photorespiration: Players, Partners and Origin." *Trends in Plant Science* 15, no. 6 (2010): 330–36.
- Beard, Daniel A. "A Biophysical Model of the Mitochondrial Respiratory System and Oxidative Phosphorylation." *PLoS Computational Biology* 1, no. 4 (2005): e36.
- Bedell, Joseph A., Muhammad A. Budiman, Andrew Nunberg, Robert W. Citek, Dan Robbins, Joshua Jones, Elizabeth Flick, Theresa Rohlfing, Jason Fries, and Kourtney Bradford. "Sorghum Genome Sequencing by Methylation Filtration." *PLoS Biology* 3, no. 1 (2005): e13.
- Ben-Hammouda, M., R. J. Kremer, H. C. Minor, and M. Sarwar. "A Chemical Basis for Differential Allelopathic Potential of Sorghum Hybrids on Wheat." *Journal of Chemical Ecology* 21, no. 6 (1995): 775–786.
- Benjamini, Yoav, and Daniel Yekutieli. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics*, 2001, 1165–88. <http://www.jstor.org/stable/10.2307/2674075>.
- Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* (1995): 289–300. <http://www.jstor.org/stable/10.2307/2346101>.
- Bennett, M. D., and I. J. Leitch. "Plant DNA C-values Database." *Royal Botanic Gardens, Kew* (2005).
- Bennetzen JL, M Freeling. "The unified grass genome: synergy in synteny". *Genome Research* 7, (1997): 301-306
- Bennetzen JL. "Comparative genomics approaches to the study of drought tolerance." In: *Molecular approaches in genetic improvements for stable production in water-limited environments* (2000): 41-44; Jean-Marcel Ribuat and David Poland (eds.); a strategic planning workshop held at CYMMIT, B, BAtan, Mexico, 21-25, June 1999.
- Bennetzen, J.L. "The Potential of Biotechnology for the Improvement of Sorghum and Pearl millet." In: *IntsorMil and ICRISAT* (eds), *Proceedings of the International Conference on Genetic*



- Improvement of Sorghum and Pearl Millet, sep. 23-27, 1996, Lubbock, Texas. (1997): pp. 13-20.
- Benson, D. A., M. S. Boguski, D. J. Lipman, J. Ostell, B. F. F. Ouellette, B. A. Rapp, and D. L. Wheeler. "GenBank." *Nucleic Acids Research* 27, no. 1 (1999): 12–17.
- Benson, G. "Tandem Repeats Finder: a Program to Analyze DNA Sequences." *Nucleic Acids Research* 27, no. 2 (1999): 573–580.
- Berg, Ivan A., Daniel Kockelkorn, H. W. Ramos-Vera, Rafael Say, Jan Zarzycki, and Georg Fuchs. "Autotrophic Carbon Fixation in Biology: Pathways, Rules, and Speculations." *Carbon Dioxide as Chemical Feedstock*, 2010, 33–53.
- Bernstein, H. "Commercial agriculture in South Africa since 1994: natural, simply capitalism." *J. Agrar. Change* 13, (2013): 23–46.
- Bevilacqua A., M. Cristina, S. Capaccioli, and A. Nicolin. "Post-Transcriptional Regulation of Gene Expression by Degradation of Messenger RNAs"(2003). *Journal of cellular physiology* 195:356–372.
- Birney, Ewan, and Richard Durbin. "Using GeneWise in the Drosophila Annotation Experiment." *Genome Research* 10, no. 4 (2000): 547–548.
- Boeckmann B, A. Bairoch, R. Apweiler<sup>1</sup>, Marie-CI Blatter, A Estreicher, E Gasteiger, M J. Martin<sup>1</sup>, Ke Michoud, C O'Donovan<sup>1</sup>, I Phan, S Pilbout and M Schneider. "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003". *Nucleic Acids Research* 31, 1(2003): 365–370.
- Boguski, M S, T M Lowe, and C M Tolstoshev. "dbEST--Database for 'Expressed Sequence Tags'." *Nature Genetics* 4, no. 4 (1993): 332–333. doi:10.1038/ng0893-332.
- Bohnert, H. J., Q. Gong, P. Li, and S. Ma. "Unraveling Abiotic Stress Tolerance Mechanisms—getting Genomics Going." *Current Opinion in Plant Biology* 9, no. 2 (2006): 180–188.
- Bola, G., C. Mabiza, J. Goldin, K. Kujinga, I. Nhapi, H. Makurira, and D. Mashauri. "Coping with Droughts and Floods: A Case Study of Kanyemba, Mbire District, Zimbabwe." *Physics and Chemistry of the Earth, Parts A/B/C* 67 (2014): 180–86.
- Bolot, S., M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing, C. Feuillet, and J. Salse. "The 'inner Circle' of the Cereal Genomes." *Current Opinion in Plant Biology* 12, no. 2 (2009): 119–125.
- Bonneau, Richard, and David Baker. "Ab Initio Protein Structure Prediction: Progress and

- Prospects." *Annual Review of Biophysics and Biomolecular Structure* 30, no. 1 (2001): 173–189.
- Bordat, Amandine, Vincent Savoie, Marie Nicolas, Jérôme Salse, Aurélie Chauveau, Michael Bourgeois, Jean Potier, et al. "Translational Genomics in Legumes Allowed Placing in Silico 5460 Unigenes on the Pea Functional Map and Identified Candidate Genes in *Pisum Sativum* L." *G3: Genes, Genomes, Genetics* 1, no. 2 (2011): 93–103.
- Borrell AK., Graeme LH and Robert GH. "Does Maintaining Green Leaf Area in Sorghum Improve Yield under Drought? II. Dry Matter Production and Yield." *Crop Science* 40, 4(2000):1037-1048.
- Borrell, A.K., Oosterom, E.J., Mullet, J.E., George-Jaeggli, B., Jordan, D.R., Klein, P.E., Hammer, G.L. "Stay-green alleles individually enhance grain yield in sorghum under drought by modifying canopy development and water uptake patterns." *New Phytol.* (2014).
- Boyd, C D, R A Pierce, J E Schwarzbauer, K Doege, and L J Sandell. "Alternate Exon Usage Is a Commonly Used Mechanism for Increasing Coding Diversity within Genes Coding for Extracellular Matrix Proteins." *Matrix (Stuttgart, Germany)* 13, no. 6 (November 1993): 457–469.
- Boyer PD. "The ATP synthase—a splendid molecular machine." *Annual Review of Biochemistry* 66 (1997): 717-749
- Bradford MM. "A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding," *Analytical Biochemistry* 72, no. 1 (1976): 248–254.
- Breitling, Rainer, Patrick Armengaud, and Anna Amtmann. "Vector Analysis as a Fast and Easy Method to Compare Gene Expression Responses between Different Experimental Backgrounds." *BMC Bioinformatics* 6, no. 1 (2005): 181.
- Breitling, Rainer, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. "Rank Products: A Simple, yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments." *FEBS Letters* 573, no. 1 (2004): 83–92. <http://www.sciencedirect.com/science/article/pii/S0014579304009354>.
- Brendel, V., L. Xing, and W. Zhu. "Gene Structure Prediction from Consensus Spliced Alignment of Multiple ESTs Matching the Same Genomic Locus." *Bioinformatics* 20, no. 7 (2004): 1157–1169.

- Brennecke, Julius, Alexei A. Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J. Hannon. "Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*." *Cell* 128, no. 6 (2007): 1089–1103.
- Brewer MJ and RR Heim Jr. "International Drought Workshop Series." *Bulletin of the American Meteorological Society* 92, no. 7 (2011): E29–E31.
- Bright, Lauren, Shane Burgess, Bhanu Chowdhary, Cyprianna Swiderski, and Fiona McCarthy. "Structural and Functional-annotation of an Equine Whole Genome Oligoarray." *BMC Bioinformatics* 10, no. Suppl 11 (2009): S8.
- Brito, Giovani Greigh de, Valdinei Sofiatti, Marleide Magalhães de Andrade Lima, Luiz Paulo de Carvalho, and João Luiz da Silva Filho. "Physiological Traits for Drought Phenotyping in Cotton." *Acta Scientiarum. Agronomy* 33, no. 1 (2011): 117–125.
- Britton, N. F. "Essential Mathematical Biology." Springer Verlag, 2003.
- Brown, Garth R., Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, et al. "Gene: A Gene-Centered Information Resource at NCBI." *Nucleic Acids Research*, 2014, gku1055.
- Brunner, Amy M., Victor B. Busov, and Steven H. Strauss. "Poplar Genome Sequence: Functional Genomics in an Ecologically Dominant Plant Species." *Trends in Plant Science* 9, no. 1 (2004): 49–56.
- Brush, Stephen B. "Genes in the Field. On-Farm Conservation of Crop Diversity." *IPGRI, IDRC, Lewis Publishers*, (2000): 3–26.
- Brutnell, T.P., Wang, L., Swartwood, K., Goldschmidt, A., Jackson, D., Zhu, X.-G., Kellogg, E., Van Eck, J. "Setaria viridis: a model for C4 photosynthesis." *Plant Cell Online* 22, (2010): 2537–2544.
- Bullard JHE, Hansen PKD, and Dudoit S. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." *BMC Bioinformatics* 11, no. 1 (2010): 94.
- Bunney TD, HS van Walraven, and AH. de Boer. "14-3-3 protein is a regulator of the mitochondrial and chloroplast ATP synthase." *PNAS* 98, no. 727 (2001): 4249–4254.
- Buratti, Emanuele, and Francisco E. Baralle. "Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator of CFTR Exon 9." *Journal of Biological Chemistry* 276, no. 39 (2001): 36337–36343.

- Burke, J. J., C. D. Franks, Gloria Burow, and Zhanguo Xin. "Selection System for the Stay-green Drought Tolerance Trait in Sorghum Germplasm." *Agronomy Journal* 102, no. 4 (2010): 1118–1122.
- Busk, Peter Kamp, and Birger Lindberg Møller. "Dhurrin Synthesis in Sorghum Is Regulated at the Transcriptional Level and Induced by Nitrogen Fertilization in Older Plants." *Plant Physiology* 129, no. 3 (January 7, 2002): 1222–1231. doi:10.1104/pp.000687.
- Byrne, P. F., and M. D. McMullen. "Defining Genes for Agricultural Traits: QTL Analysis and the Candidate Gene Approach." *Probe* 7, no. 1 (1996): 24–27.
- Cabrera-Bosquet, Llorenç, José Crossa, Jarislav von Zitzewitz, María Dolors Serret, and José Luis Araus. "High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding ConvergeF." *Journal of Integrative Plant Biology* 54, no. 5 (2012): 312–320.
- Cadenas, Enrique, and Kelvin JA Davies. "Mitochondrial Free Radical Generation, Oxidative Stress, and Aging." *Free Radical Biology and Medicine* 29, no. 3 (2000): 222–230.
- Calhoun M, Thomas J, Gennis R. "The cytochrome oxidase superfamily of redox-driven proton pumps". *Trends Biochem Sci* 19, no. 8 (1994): 325–230. doi:10.1016/0968-0004(94)90071-X
- Califano, Andrea, Atul J. Butte, Stephen Friend, Trey Ideker, and Eric Schadt. "Leveraging Models of Cell Regulation and GWAS Data in Integrative Network-Based Association Studies." *Nature Genetics* 44, no. 8 (2012): 841–847.
- Camon E, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Res.*, vol. 32, no. suppl 1, pp. D262–D266, 2004.
- Campos, H., M. Cooper, J. E. Habben, G. O. Edmeades, and J. R. Schussler. "Improving Drought Tolerance in Maize: A View from Industry." *Field Crops Research* 90, no. 1 (2004): 19–34.
- Capelo, J. L., R. Carreira, M. Diniz, L. Fernandes, M. Galesio, C. Lodeiro, H. M. Santos, and G. Vale. "Overview on Modern Approaches to Speed up Protein Identification Workflows Relying on Enzymatic Cleavage and Mass Spectrometry-based Techniques." *Analytica Chimica Acta* 650, no. 2 (2009): 151–159.
- Carpentier SC, E Witters, K Laukens, P Deckers, R Swennen and B Panis. "Preparation of Protein Extracts from Recalcitrant Plant Tissues: An Evaluation of Different Methods for Two-

- Dimensional Gel Electrophoresis Analysis,” *Proteomics* 5, no. 10 (2005): 2497–2507.
- Carrera, Esther, Tara Holman, Anne Medhurst, Wendy Peer, Heike Schmuths, Steven Footitt, Frederica L. Theodoulou, and Michael J. Holdsworth. “Gene Expression Profiling Reveals Defined Functions of the ATP-Binding Cassette Transporter COMATOSE Late in Phase II of Germination.” *Plant Physiology* 143, no. 4 (2007): 1669–1679.
- Castellana, N.E., Shen, Z., He, Y., Walley, J.W., Briggs, S.P., Bafna, V., others. “An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*.” *Mol. Cell. Proteomics* 13, (2014): 157–167.
- Cattivelli, L. Fulvia Rizza, Franz-W. Badeck, Elisabetta Mazzucotelli, Anna M. Mastrangelo, Enrico Francia, Caterina Mare, Alessandro Tondelli, A. Michele Stanca. Drought tolerance improvement in crop plants: An integrated view from breeding to genomics (2008). *Field Crops Research* 105: 1–14.
- Ceccarelli, Salvatore, and Stefania Grando. “Decentralized-Participatory Plant Breeding: An Example of Demand Driven Research.” *Euphytica* 155, no. 3 (2007): 349–360. doi:10.1007/s10681-006-9336-8.
- Chae L, KA Dreher, Ricardo Nilo-Poyanco, Chuan Wang, Taehyong Kim, Seung Yon Rhee, Peifen Zhang. 2013. Carnegie Institution for Science. Plant Metabolic Network (PMN), <http://pmn.plantcyc.org/SORGHUMBICOLOR/organism-summary?object=SORGHUMBICOLOR> on [www.plantcyc.org](http://www.plantcyc.org), June 05, 2013).
- Chain, P. S. G., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, and C. Buhay. “Genome Project Standards in a New Era of Sequencing.” *Science* 326, no. 5950 (2009): 236–237.
- Champoux, M. C., G. Wang, S. Sarkarung, D. J. Mackill, J. C. O’Toole, N. Huang, and S. R. McCouch. “Locating Genes Associated with Root Morphology and Drought Avoidance in Rice via Linkage to Molecular Markers.” *Theoretical and Applied Genetics* 90, no. 7–8 (1995): 969–981.
- Chandrakant, PM. “Candidate Genes as Molecular Markers for Evaluating and Validating Sugarcane Germplasm for Salinity Stress,” 2012. HYPERLINK "<http://ir.inflibnet.ac.in:8080/jspui/handle/10603/4336>"<http://ir.inflibnet.ac.in:8080/jspui/handle/10603/4336>.
- Chapman, Scott, Mark Cooper, Dean Podlich, and Graeme Hammer. “Evaluating Plant Breeding

- Strategies by Simulating Gene Action and Dryland Environment Effects.” *Agronomy Journal* 95, no. 1 (2003): 99–113.
- Chawla, K. B. P., M. Kuiper, and A. M. Bones. “Systems Biology: a Promising Tool to Study Abiotic Stress Responses.” *Omics and Plant Abiotic Stress Tolerance. Betham eBooks: International Centre for Genetic Engineering and Biotechnology, New Delhi, India* (2011): 163–172.
- Chen, Ri-Dong, Long-Xi Yu, Ann Francine Greer, Hassan Cheriti, and Zohreh Tabaeizadeh. “Isolation of an Osmotic Stress-and Abscisic Acid-Induced Gene Encoding an Acidic Endochitinase from *Lycopersicon Chilense*.” *Molecular and General Genetics MGG* 245, no. 2 (1994): 195–202.
- Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G. “ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.” *Nucleic Acids Res.* 37, (2009): W305–W311.
- Cheong, Y.H., Chang, H.-S., Gupta, R., Wang, X., Zhu, T., Luan, S. “Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*.” *Plant Physiol.* 129, (2002): 661–677.
- Chesler, Elissa J., Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D. Mountz, Nicole E. Baldwin, and Michael A. Langston. “Complex Trait Analysis of Gene Expression Uncovers Polygenic and Pleiotropic Networks That Modulate Nervous System Function.” *Nature Genetics* 37, no. 3 (2005): 233–242.
- Childs, Kevin L., John P. Hamilton, Wei Zhu, Eugene Ly, Foo Cheung, Hank Wu, Pablo D. Rabinowicz, Chris D. Town, C. Robin Buell, and Agnes P. Chan. “The TIGR Plant Transcript Assemblies Database.” *Nucleic Acids Research* 35, no. suppl 1 (2007): D846–D851.
- Childs, Kevin L., Robert R. Klein, Patricia E. Klein, Daryl T. Morishige, and John E. Mullet. “Mapping Genes on an Integrated Sorghum Genetic and Physical Map Using cDNA Selection Technology.” *The Plant Journal* 27, no. 3 (2001): 243–55.
- Choi JW, MG Han, SY Kim, SG Oh, SS Im. “Poly (3, 4-Ethylenedioxythiophene) Nanoparticles Prepared in Aqueous DBSA Solutions,” *Synthetic Metals* 141, no. 3 (2004): 293–299.
- Christoffels A, Miller R, Hide W. “STACK\_PACK and STACK (Sequence Tag Alignment and Consensus Knowledgebase): A novel, comprehensive, hierarchical EST clustering and consensus generation and analysis system providing unique insight into the human genome.”

- Am J Hum Genet. 65, 4 (1999): 477.
- Christoffels, A., A. van Gelder, G. Greyling, R. Miller, T. Hide and W. Hide (2001). STACK: Sequence Tag Alignment and Consensus Knowledgebase . *Nucleic Acids Research*, 29, 234–238.
- Ciarmiello, L. F., P. Woodrow, A. Fuggi, G. Pontecorvo, and P. Carillo. “Plant Genes for Abiotic Stress” (2011).
- Clouser KR, Baker P, Burlingame AL. "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." *Anal. Chem.* 71 14 (1999): 2871–82.
- Claverie, Jean-Michel, and Cedric Notredame. “Bioinformatics for Dummies.” John Wiley & Sons, (2011).
- Cleland WW. “Dithiothreitol, a New Protective Reagent for SH Groups.” *Biochemistry* 3, no. 4 (1964): 480–482.
- Cline, Melissa S., Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, and Benjamin Gross. “Integration of Biological Networks and Gene Expression Data Using Cytoscape.” *Nature Protocols* 2, no. 10 (2007): 2366–2382.
- Close, T. J., S. I. Wanamaker, R. A. Caldo, S. M. Turner, D. A. Ashlock, J. A. Dickerson, R. A. Wing, G. J. Muehlbauer, A. Kleinhofs, and R. P. Wise. “A New Resource for Cereal Genomics: 22K Barley GeneChip Comes of Age.” *Plant Physiology* 134, no. 3 (2004): 960–968.
- Cobb, Joshua N., Genevieve DeClerck, Anthony Greenberg, Randy Clark, and Susan McCouch. “Next-generation Phenotyping: Requirements and Strategies for Enhancing Our Understanding of Genotype–phenotype Relationships and Its Relevance to Crop Improvement.” *Theoretical and Applied Genetics* (2013): 1–21.
- Cohen, Adam E., and Veena Venkatachalam. “Bringing Bioelectricity to Light.” *Annual Review of Biophysics*, no. 0 (2014): <http://www.annualreviews.org/doi/abs/10.1146/annurev-biophys-051013-022717>.
- Colbourne JK, Singan VR, Gilbert DG, 2005. wFleaBase: the Daphnia genome database. *BMC Bioinformatics* 6: 45.
- Collard, Bert C. Y., Casiana M. Vera Cruz, Kenneth L. McNally, Parminder S. Virk, and David J. Mackill. “Rice Molecular Breeding Laboratories in the Genomics Era: Current Status and

- Future Considerations.” *International Journal of Plant Genomics* 2008 (2008). doi:10.1155/2008/524847.
- Collins NC, F Tardieu, R Tuberosa - Plant Physiology. “Quantitative trait loci and crop performance under abiotic stress: where do we stand?” *Plant Physiology* 147 (2008): 469–486.
- Colussi, C., M. C. Albertini, S. Coppola, S. Rovidati, F. Galli, and L. Ghibelli. “H<sub>2</sub>O<sub>2</sub>-Induced Block of Glycolysis as an Active ADP-Ribosylation Reaction Protecting Cells from Apoptosis.” *The FASEB Journal* 14, no. 14 (2000): 2266–76. doi:10.1096/fj.00-0074com.
- Conesa, A, S Gotz. “Blast2GO Tutorial.” Bioinformatics and Genomics Department, Prince Felipe Research Center, Valencia SPAIN (2009): 1-30.
- Conesa, A, S Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. “Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research.” *Bioinformatics* 21, no. 18 (2005): 3674–3676. <http://bioinformatics.oxfordjournals.org/content/21/18/3674.short>.
- Cooper GM, Hausman RE. "The Cell: A molecular approach fifth edition." (2000) – [tocs.ulb.tu-darmstadt.de](http://tocs.ulb.tu-darmstadt.de)
- Cooper SJ, GA Leonard, SM McSweeney, AW Thompson, JH Naismith, S Qamar, A Plater, A Berry and WN Hunter. “The crystal structure of a class II fructose-1,6-bisphosphate.” *Structure*, Volume 4, Issue 11, (1996): 1303–1315.
- Cooper, Laurel, Ramona L. Walls, Justin Elser, Maria A. Gandolfo, Dennis W. Stevenson, Barry Smith, Justin Preece, Balaji Athreya, Christopher J. Mungall, and Stefan Rensing. “The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses.” *Plant and Cell Physiology* 54, no. 2 (2013): e1–e1.
- Cordero, Nestor-Luis. “By Being, It Is: The Thesis of Parmenides” (2004). <http://philpapers.org/rec/CORBBI>.
- Cornic, G., Massacci, A. “Leaf photosynthesis under drought stress, in: Photosynthesis and the Environment.” *Springer*, (1996): 347–366.
- Corrêa, L. G. G., D. M. Riaño-Pachón, C. G. Schrago, R. V. Dos Santos, B. Mueller-Roeber, and M. Vincentz. “The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes.” *PLoS One* 3, no. 8 (2008): e2944.
- Cossins, Andrew R., and Douglas L. Crawford. “Fish as Models for Environmental Genomics.” *Nature Reviews Genetics* 6, no. 4 (2005): 324–333.



- Cox, J., and M. Mann. "Quantitative, High-resolution Proteomics for Data-driven Systems Biology." *Annual Review of Biochemistry* 80 (2011): 273–299.
- Cramer, Grant R., Ali Ergül, Jerome Grimplet, Richard L. Tillett, Elizabeth AR Tattersall, Marlene C. Bohlman, Delphine Vincent, Justin Sonderegger, Jason Evans, and Craig Osborne. "Water and Salinity Stress in Grapevines: Early and Late Changes in Transcript and Metabolite Profiles." *Functional & Integrative Genomics* 7, no. 2 (2007): 111–134.
- Cramer, Grant R., Kaoru Urano, Serge Delrot, Mario Pezzotti, and Kazuo Shinozaki. "Effects of Abiotic Stress on Plants: a Systems Biology Perspective." *BMC Plant Biology* 11, no. 1 (2011): 163.
- Crasta OR, WW Xu, DT Rosenow, J Mullet, HT Nguyen. "Mapping of post-flowering drought resistance traits in grain sorghum: association between QTLs influencing premature senescence and maturity." *Molecular and General Genetics* 262, no. 3 (1999): 579-588.
- Cseri, András, Mátyás Cserhádi, Maria Von Korff, Bettina Nagy, Gábor V. Horváth, András Palágyi, János Pauk, Dénes Dudits, and Ottó Törjék. "Allele Mining and Haplotype Discovery in Barley Candidate Genes for Drought Tolerance." *Euphytica* 181, no. 3 (2011): 341–56.
- Curwen, Val, Eduardo Eyras, T. Daniel Andrews, Laura Clarke, Emmanuel Mongin, Steven MJ Searle, and Michele Clamp. "The Ensembl Automatic Gene Annotation System." *Genome Research* 14, no. 5 (2004): 942–950.
- Cushman, J. C., and H. J. Bohnert. "Genomic Approaches to Plant Stress Tolerance." *Current Opinion in Plant Biology* 3, no. 2 (2000): 117–124.
- Da Cruz, S.M.S., de Macedo Vieira, A.C. "Managing Semantic Annotations on Medicinal Plants." In *Information Society (i-Society), 2010 International Conference on*, 554–559. IEEE, 2010. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6018776](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6018776).
- Dai, M., P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, and H. Akil. "Evolving Gene/transcript Definitions Significantly Alter the Interpretation of GeneChip Data." *Nucleic Acids Research* 33, no. 20 (2005): e175–e175.
- Daraselia, Nikolai, Anton Yuryev, Sergei Egorov, Ilya Mazo, and Iaroslav Ispolatov. "Automatic Extraction of Gene Ontology Annotation and Its Correlation with Clusters in Protein Networks." *BMC Bioinformatics* 8, no. 1 (2007): 243.
- Deeba, F., Pandey, A.K., Ranjan, S., Mishra, A., Singh, R., Sharma, Y.K., Shirke, P.A., Pandey, V. "Physiological and proteomic responses of cotton (*Gossypium herbaceum* L.) to drought

- stress." *Plant Physiol. Biochem.* (2012): 53, 6–18.
- De la Fuente, Alberto. "From 'differential Expression' to 'differential Networking'—identification of Dysfunctional Regulatory Networks in Diseases." *Trends in Genetics* 26, no. 7 (2010): 326–333.
- Decker C and AD Jenkins. "Kinetic Approach of Oxygen Inhibition in Ultraviolet-and Laser-Induced Polymerizations," *Macromolecules* 18, no. 6 (1985): 1241–1244.
- Demir, Emek, Michael P. Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, and Joanne Luciano. "The BioPAX Community Standard for Pathway Data Sharing." *Nature Biotechnology* 28, no. 9 (2010): 935–942.
- Deokar, A.A., Kondawar, V., Jain, P.K., Karuppayil, S.M., Raju, N.L., Vadez, V., Varshney, R.K., Srinivasan, R. "Comparative analysis of expressed sequence tags (ESTs) between drought-tolerant and-susceptible genotypes of chickpea under terminal drought stress." *BMC Plant Biol.* 11, (2011): 70.
- Devos, Katrien M., and Mike D. Gale. "Genome Relationships: The Grass Model in Current Research." *The Plant Cell Online* 12, no. 5 (2000): 637–46.
- Deutscher, M. P. (ed.). "*Guide to Protein Purification.*" Vol. 182. Academic Press, San Diego, CA.1990.
- Diab, Ayman A., R. V. Kantety, N. Z. Ozturk, D. Benscher, M. M. Nachit, and M. E. Sorrells. "Drought-Inducible Genes and Differentially Expressed Sequence Tags Associated with Components of Drought Tolerance in Durum Wheat," 2013. <http://dspacetest.cgiar.org/handle/10883/2643>.
- Dillon, Sally L., Frances M. Shapter, Robert J. Henry, Giovanni Cordeiro, Liz Izquierdo, and L. Slade Lee. "Domestication to Crop Improvement: Genetic Resources for Sorghum and Saccharum (Andropogoneae)." *Annals of Botany* 100, no. 5 (2007): 975–89.
- Dixon, Anna L., Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny CC Wong, Jenny Taylor, Edward Burnett, Ivo Gut, and Martin Farrall. "A Genome-Wide Association Study of Global Gene Expression." *Nature Genetics* 39, no. 10 (2007): 1202–1207.
- Doggett, H. "Sorghum history in relation to Ethiopia." In: *Plant Genetic Resources of Ethiopia*, (Engels, J. M. M., Hawks, J. G., and Worede, M., eds.). Cambridge University press, Cambridge. (1991): 140 – 159.

- Doggett, H. "Sorghum." John Wiley & Sons, Inc., New York. 1988.
- Draye, Xavier, Yann-Rong Lin, Xiao-yin Qian, John E. Bowers, Gloria B. Burow, Peter L. Morrell, Daniel G. Peterson, Gernot G. Presting, Shu-xin Ren, and Rod A. Wing. "Toward Integration of Comparative Genetic, Physical, Diversity, and Cytomolecular Maps for Grasses and Grains, Using the Sorghum Genome as a Foundation." *Plant Physiology* 125, no. 3 (2001): 1325–41.
- Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. "agriGO: A GO Analysis Toolkit for the Agricultural Community." *Nucleic Acids Research* 38, no. suppl 2 (2010): W64–W70.
- Dugas, Diana, Marcela Monaco, Andrew Olson, Robert Klein, Sunita Kumari, Doreen Ware, and Patricia Klein. "Functional Annotation of the Transcriptome of Sorghum Bicolor in Response to Osmotic Stress and Abscisic Acid." *BMC Genomics* 12, no. 1 (2011): 514.
- Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. "Expression Profiling Using cDNA Microarrays." *Nature Genetics* 21, no. 1 Suppl (1999): 10–14.
- Dutilh Bas E., Martijn A. Huynen, and Berend Snel. "A Global Definition of Expression Context Is Conserved between Orthologs, but Does Not Correlate with Sequence Conservation." *BMC Genomics*. (2006). 7(1): 1–10.
- Eathington, Sam R., Theodore M. Crosbie, Marlin D. Edwards, Robert S. Reiter, and Jason K. Bull. "Molecular Markers in a Commercial Breeding Program." *Crop Science* 47, no. Supplement\_3 (2007): S-154.
- Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. "GORILLA: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists." *BMC Bioinformatics* 10, no. 1 (2009): 48.
- Edmeades Greg O. Progress in Achieving and Delivering Drought Tolerance in Maize- An Update (2013): 1-39.
- Edwards, Gerald E., Robert T. Furbank, Marshall D. Hatch, and C. Barry Osmond. "What Does It Take to Be C4? Lessons from the Evolution of C4 Photosynthesis." *Plant Physiology* 125, no. 1 (2001): 46–49.
- Ejeta, G. "Breeding for Resistance in Sorghum: Exploitation of an Intricate Host–Parasite Biology." *Crop Science* 47, no. 3 (2007): S216–S227.
- Ejeta, G. and J. Gressel. "Integrating new technologies for striga control: Towards ending the Witch-hunt." Published by: World Scientific Publishing Co. Pte., Ltd., 5 Toh Tuck Link,

Singapore, 596224, (2007).

- El Soda, Mohamed, Satya Swathi Nadakuduti, Klaus Pillen, and Ralf Uptmoor. "Stability Parameter and Genotype Mean Estimates for Drought Stress Effects on Root and Shoot Growth of Wild Barley Pre-introgression Lines." *Molecular Breeding* 26, no. 4 (2010): 583–593.
- Entelis, Nina, Irina Brandina, Piotr Kamenski, Igor A. Krasheninnikov, Robert P. Martin, and Ivan Tarassov. "A Glycolytic Enzyme, Enolase, Is Recruited as a Cofactor of tRNA Targeting toward Mitochondria in *Saccharomyces Cerevisiae*." *Genes & Development* 20, no. 12 (2006): 1609–20.
- Ersoz, Elhan S., Mark H. Wright, Jasmyn L. Pangilinan, Moira J. Sheehan, Christian Tobias, Michael D. Casler, Edward S. Buckler, and Denise E. Costich. "SNP Discovery with EST and NextGen Sequencing in Switchgrass (*Panicum Virgatum* L.)." *PLoS One* 7, no. 9 (2012): e44112.
- Ewing, R. M., A. B. Kahla, O. Poirot, F. Lopez, S. Audic, and J. M. Claverie. "Large-scale Statistical Analyses of Rice ESTs Reveal Correlated Patterns of Gene Expression." *Genome Research* 9, no. 10 (1999): 950–959.
- Falick, Arnold M., William S. Lane, Kathryn S. Lilley, Michael J. MacCoss, Brett S. Phinney, Nicholas E. Sherman, Susan T. Weintraub, H. Ewa Witkowska, and Nathan A. Yates. "ABRF-PRG07: Advanced Quantitative Proteomics Study." *Journal of Biomolecular Techniques: JBT* 22, no. 1 (2011): 21.
- Faris, J. D., W. L. Li, D. J. Liu, P. D. Chen, and B. S. Gill. "Candidate Gene Analysis of Quantitative Disease Resistance in Wheat." *TAG Theoretical and Applied Genetics* 98, no. 2 (1999): 219–225.
- Farooq M., A. Wahid, Lee D-J, O Ito & K H. M. Siddique. "Advances in Drought Resistance of Rice". *Critical Reviews in Plant Sciences*. (2009). **28(4)**: 199–217.
- Fensholt, R., Langanke, T., Rasmussen, K., Reenberg, A., Prince, S.D., Tucker, C., Scholes, R.J., Le, Q.B., Bondeau, A., Eastman, R., others. "Greenness in semi-arid areas across the globe 1981–2007—An Earth Observing Satellite based analysis of trends and drivers." *Remote Sens. Environ.* 121, (2012): 144–158.
- Feuillet, C, and B. Keller. "Comparative Genomics in the Grass Family: Molecular Characterization of Grass Genome Structure and Evolution." *Annals of Botany* 89, no. 1 (2002): 3–10.
- Fevotte G, N Gherras, J Moutte. "Batch cooling solution crystallization of ammonium oxalate in the

- presence of impurities: study of solubility, supersaturation, and steady-state Inhibition." *Crystal Growth & Design* 13, no.7 (2013): 2737–2748.
- Ficklin, Stephen P., Feng Luo, and F. Alex Feltus. "The Association of Multiple Interacting Genes with Specific Phenotypes in Rice Using Gene Coexpression Networks." *Plant Physiology* 154, no. 1 (2010): 13–23.
- Fierro-Monti, Ivo, and Michael B. Mathews. "Proteins Binding to Duplexed RNA: One Motif, Multiple Functions." *Trends in Biochemical Sciences* 25, no. 5 (2000): 241–46.
- Fitch, Walter M. "Distinguishing Homologous from Analogous Proteins." *Systematic Biology* 19, no. 2 (1970): 99–113.
- . "Homology: A Personal View on Some of the Problems." *Trends in Genetics* 16, no. 5 (2000): 227–231.
- Fleury, Christophe, Bernard Mignotte, and Jean-Luc Vayssière. "Mitochondrial Reactive Oxygen Species in Cell Death Signaling." *Biochimie* 84, no. 2 (2002): 131–41.
- Fleury, Delphine, Stephen Jefferies, Haydn Kuchel, and Peter Langridge. "Genetic and Genomic Tools to Improve Drought Tolerance in Wheat." *Journal of Experimental Botany* 61, no. 12 (2010): 3211–3222.
- Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, et al. "Ensembl 2014." *Nucleic Acids Research* 42 (2014): D749–D755.
- Flicek, Paul, Ikhlaq Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, et al. "Ensembl 2013." *Nucleic Acids Research* 41 (2013): D48–D55.
- Fu X and Z Tang. "Planning for Drought-resilient Communities: An Evaluation of Local Comprehensive Plans in the Fastest Growing Counties in the US." *Cities* 32 (2013): 60–69.
- Fuchs, Stefan, Erwin Grill, Irute Meskiene, and Alois Schweighofer. "Type 2C Protein Phosphatases in Plants." *The FEBS Journal* 280, no. 2 (2013): 681–93. doi:10.1111/j.1742-4658.2012.08670.x.
- Fulton, T. M., R. Van der Hoeven, N. T. Eannetta, and S. D. Tanksley. "Identification, Analysis, and Utilization of Conserved Ortholog Set Markers for Comparative Genomics in Higher Plants." *The Plant Cell Online* 14, no. 7 (2002): 1457–1467.
- Gallagher SR. "One-Dimensional SDS Gel Electrophoresis of Proteins." *Curr. Protoc. Mol. Biol.* 97 (2006):10.2A.1-10.2A.44.

- Gallina A, TM Hanley, R Mandel, M Trahey, CC Broder, GA Viglianti and HJ-P Ryser. "Inhibitors of Protein-Disulfide Isomerase Prevent Cleavage of Disulfide Bonds in Receptor-Bound Glycoprotein 120 and Prevent HIV-1 Entry," *Journal of Biological Chemistry* 277, no. 52 (2002): 50579–50588.
- Gan, G., Wu, J., Yang, Z. "A genetic fuzzy k-Modes algorithm for clustering categorical data." *Expert Syst. Appl.* 36, (2009): 1615–1620.
- Garfin DE. "Isoelectric Focusing," *Separation Science and Technology* 2 (2000): 263–298.
- Garfin DE. 2-D Electrophoresis for Proteomics: A Methods and Product Manual, Bio-Rad, (2001).
- Garrels JI. "The QUEST System for Quantitative Analysis of Two-dimensional Gels". *The journal of biological chemistry* 264, 9 (1989): 5269-5282.
- Gasteiger E., C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch. "Protein Identification and Analysis Tools on the ExPASy Server". *The Proteomics Protocols Handbook* (2005): 571-607.
- Gauges, Ralph, Ursula Rost, Sven Sahle, and Katja Wegner. "A Model Diagram Layout Extension for SBML." *Bioinformatics* 22, no. 15 (2006): 1879–1885.
- Gaulton, Kyle J., Karen L. Mohlke, and Todd J. Vision. "A Computational System to Select Candidate Genes for Complex Human Traits." *Bioinformatics* 23, no. 9 (2007): 1132–1140.
- Gaut, B. S., L. G. Clark, J. F. Wendel, and S. V. Muse. "Comparisons of the Molecular Evolutionary Process at rbcL and ndhF in the Grass Family (Poaceae)." *Molecular Biology and Evolution* 14, no. 7 (1997): 769–777.
- Gautier V., E. Mouton-Barbosa, D Bouyssié, N. Delcourt, M Beau, J Girard, C Cayrol, O Burllet-Schiltz, B Monsarrat, and A. G. de Peredo. "Label-free Quantification and Shotgun Analysis of Complex Proteomes by One-dimensional SDS-PAGE/NanoLC-MS" . *Molecular & Cellular Proteomics* 11, (2012): 527–539.
- Gene Ontology Consortium,. "Creating the Gene Ontology Resource: Design and Implementation." *Genome Research* 11, no. 8 (2001): 1425–33.
- Gentles, Andrew J., and Samuel Karlin. "Why Are Human G-Protein-Coupled Receptors Predominantly Intronless?" *Trends in Genetics* 15, no. 2 (1999): 47–49.
- Ghannoum O. "C4 Photosynthesis and Water Stress," *Annals of Botany* 103, no. 4 (2009): 635–44.
- Ghannoum, Oula, Jann P. Conroy, Simon P. Driscoll, Matthew J. Paul, Christine H. Foyer, and David W. Lawlor. "Nonstomatal Limitations Are Responsible for Drought-Induced

- Photosynthetic Inhibition in Four C4 Grasses." *New Phytologist* 159, no. 3 (2003): 599–608.
- Ghannoum, Oula, Susanne von Caemmerer, and Jann P. Conroy. "The Effect of Drought on Plant Water Use Efficiency of Nine NAD–ME and Nine NADP–ME Australian C4 Grasses." *Functional Plant Biology* 29, no. 11 (2002): 1337–48.
- Ghannoum, Oula. "C4 Photosynthesis and Water Stress." *Annals of Botany* 103, no. 4 (2009): 635–44.
- Gholipour, Manoochehr, P. V. Prasad, Raymond N. Mutava, and Thomas R. Sinclair. "Genetic Variability of Transpiration Response to Vapor Pressure Deficit Among Sorghum Genotypes." *Field Crops Research* 119, no. 1 (2010): 85–90.
- Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, and B. Andre. "Functional Profiling of the *Saccharomyces Cerevisiae* Genome." *Nature* 418, no. 6896 (2002): 387–391.
- Girard, Martine, Viviane Poupon, Francois Blondeau, and Peter S. McPherson. "The DnaJ-Domain Protein RME-8 Functions in Endosomal Trafficking." *Journal of Biological Chemistry* 280, no. 48 (2005): 40135–43.
- Girma, Yemane. "Mining genomic resources for snp and snp-caps markers and divergence for drought tolerance in sorghum [*Sorghum Bicolor* (L.) Moench]." University of agricultural sciences, 2009.
- Glanemann C, A. Loos, N. Gorret, L. B. Willis, X. M. O'brien, P. A. Lessard, and A. J. Sinskey, "Disparity between changes in mRNA abundance and enzyme activity in *Corynebacterium glutamicum*: implications for DNA microarray analysis." *Appl. Microbiol. Biotechnol.*, vol. 61, no. 1 (2003): 61–68.
- Glazier, Anne M., Joseph H. Nadeau, and Timothy J. Aitman. "Finding Genes That Underlie Complex Traits." *Science* 298, no. 5602 (2002): 2345–2349.
- Goecks, Jeremy, Anton Nekrutenko, James Taylor, and others. "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences." *Genome Biol* 11, no. 8 (2010): R86.
- Goldstein, Theodore C., Evan O. Paull, Matthew J. Ellis, and Joshua M. Stuart. "Molecular Pathways: Extracting Medical Knowledge from High-Throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114–3120.
- Gong, Pengjuan, Junhong Zhang, Hanxia Li, Changxian Yang, Chanjuan Zhang, Xiaohui Zhang, Ziaf Khurram, *et al.* "Transcriptional Profiles of Drought-Responsive Genes in Modulating

- Transcription Signal Transduction, and Biochemical Pathways in Tomato.” *Journal of Experimental Botany*. erq167, (2010). doi:10.1093/jxb/erq167.
- Gonzalez-Fernandez F., D. Sung, K. M. Haswell, A. Tsin, D. Ghosh. “Thiol-dependent antioxidant activity of interphotoreceptor retinoid-binding protein”. *Experimental Eye Research* 120 (2014):167–174.
- Goodstadt, L, and CP Ponting. “Phylogenetic Reconstruction of Orthology, Paralogy, and Conserved Synteny for Dog and Human.” *PLoS Comput Biol* 2, no. 9 (2006): e133. doi:10.1371/journal.pcbi.0020133.
- Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, *et al.* “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40, no. D1 (2012): D1178–D1186. doi:10.1093/nar/gkr944.
- Görg A, W Postel, and S Günther. “Two-Dimensional Electrophoresis. The Current State of Two-Dimensional Electrophoresis with Immobilized pH Gradients,” *Electrophoresis* 9, no. 9 (1988): 531–546.
- Gorg A. “Two-Dimensional Electrophoresis.” *Nature* 349 (1991): 545–546.
- Gowik U, P Westhoff. “The Path from C3 to C4 Photosynthesis”. *Plant Physiology* (2011): 56-63.
- Grabski AC and RRB Novagen. “Preparation of Protein Samples for SDS-Polyacrylamide Gel Electrophoresis: Procedures and Tips.” 2001, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.378.613>.
- Graham, James WA, Thomas CR Williams, Megan Morgan, Alisdair R. Fernie, R. George Ratcliffe, and Lee J. Sweetlove. “Glycolytic Enzymes Associate Dynamically with Mitochondria in Response to Respiratory Demand and Support Substrate Channeling.” *The Plant Cell Online* 19, no. 11 (2007): 3723–38.
- Granier F. “Extraction of Plant Proteins for Two-Dimensional Electrophoresis,” *Electrophoresis* 9, no. 11 (1988): 712–718.
- Greene, Michael K., Karol Maskos, and Samuel J. Landry. “Role of the J-Domain in the Cooperation of Hsp40 with Hsp70.” *Proceedings of the National Academy of Sciences* 95, no. 11 (1998): 6108–6113.
- Grover, Anita. “Plant Chitinases: Genetic Diversity and Physiological Roles.” *Critical Reviews in Plant Sciences* 31, no. 1 (2012): 57–73.
- Gruber, Tom. “Ontology.” *Encyclopedia of Database Systems* (2009): 1963–1965.



- Gunavathi M, LAA Maria, VN Chamundeswari and M Parthasarathy. "Nanotube-Grafted Polyacrylamide Hydrogels for Electrophoretic Protein Separation," *Electrophoresis* 33, no. 8 (2012): 1271–1275.
- Gupta, Nitin, Stephen Tanner, Navdeep Jaitly, Joshua N. Adkins, Mary Lipton, Robert Edwards, Margaret Romine, Andrei Osterman, Vineet Bafna, and Richard D. Smith. "Whole Proteome Analysis of Post-translational Modifications: Applications of Mass-spectrometry for Proteogenomic Annotation." *Genome Research* 17, no. 9 (2007): 1362–1377.
- Gupta, Vikrant, Saurabh Raghuvanshi, Ambika Gupta, Navin Saini, Anupama Gaur, M. S. Khan, R. S. Gupta, J. Singh, S. K. Duttamajumder, and S. Srivastava. "The Water-Deficit Stress-and Red-Rot-Related Genes in Sugarcane." *Functional & Integrative Genomics* 10, no. 2 (2010): 207–14.
- Haake V, R Zrenner, U Sonnewald, M Stitt. "A moderate decrease of plastid aldolase activity inhibits photosynthesis, alters the levels of sugars and starch, and inhibits growth of potato plants." *The Plant Journal* 14, 2 (1998): 147–157.
- Haake V., D. Cook, J. Riechmann, O., Pineda, M.F. Thomashow and J.Z. Zhang. "Transcription Factor CBF4 Is a Regulator of Drought Adaptation in Arabidopsis," *Plant Physiology* 130, no. 2 (2002): 639–648.
- Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK Jr, Maiti R, Chan AP, Yu C, Farzad M, Wu D, White O, Town CD. "Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release". *BMC Biol.* (2005). 22(3): 1323–1337.
- Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. "Automated Eukaryotic Gene Structure Annotation Using EVIDENCEModeler and the Program to Assemble Spliced Alignments." *Genome Biology* 9, no. 1 (2008): R7. doi:10.1186/gb-2008-9-1-r7.
- Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith Jr, Linda I. Hannick, Rama Maiti, Catherine M. Ronning, Douglas B. Rusch, and Christopher D. Town. "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies." *Nucleic Acids Research* 31, no. 19 (2003): 5654–5666.
- Haas, Brian J., Qiandong Zeng, Matthew D. Pearson, Christina A. Cuomo, and Jennifer R. Wortman. "Approaches to Fungal Genome Annotation." *Mycology* 2, no. 3 (2011): 118–141.
- Halkier, Barbara Ann, Hanne Linde Nielsen, Birgit Koch, and Birger Lindberg Moller. "Purification

- and Characterization of Recombinant Cytochrome P450TYR Expressed at High Levels in Escherichia Coli.” *Archives of Biochemistry and Biophysics* 322, no. 2 (1995): 369–77.
- Hammer, Graeme L. “Drought Adaptation in Sorghum: Associations with Nodal Root Angle and Root System Architecture.” In *Plant and Animal Genome XX Conference (January 14-18, 2012)*. Plant and Animal Genome, 2012.
- Han, Meng-Hsuan, Saiprasad Goud, Liang Song, and Nina Fedoroff. “The Arabidopsis Double-Stranded RNA-Binding Protein HYL1 Plays a Role in microRNA-Mediated Gene Regulation.” *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 4 (2004): 1093–98.
- Harlan, J. R., and J. M. J. De Wet. “A Simplified Classification of Cultivated Sorghum.” *Crop Science* 12, no. 2 (1972): 172–176.
- Harnsomburana, Jaturon, Jason M. Green, Adrian S. Barb, Mary Schaeffer, Leszek Vincent, and Chi-Ren Shyu. “Computable Visually Observed Phenotype Ontological Framework for Plants.” *BMC Bioinformatics* 12, no. 1 (2011): 260.
- Harris K, PK Subudhi, A Borrell, D Jordan, D Rosenow, H Nguyen, P Klein<sup>6</sup>, R Klein and J Mullet. “Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence.” *J. Exp. Bot.* (2007) 58 (2): 327-338. doi: 10.1093/jxb/erl225
- Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, and C. Mungall. “The Gene Ontology (GO) Database and Informatics Resource.” *Nucleic Acids Research* 32, no. Database issue (2004): D258–61.
- Harrow J, A. Nagy, A. Reymond, T. Alioto, L. Patthy, S. E. Antonarakis, R. Guigó, and others. “Identifying protein-coding genes in genomic sequences.” *Genome Biol* 10, no. 1 (2009): 201.
- Hartl, Lukas, Simone Zach, and Verena Seidl-Seiboth. “Fungal Chitinases: Diversity, Mechanistic Properties and Biotechnological Potential.” *Applied Microbiology and Biotechnology* 93, no. 2 (2012): 533–43.
- Hatfield, Jerry L., Kenneth J. Boote, B. A. Kimball, L. H. Ziska, Roberto C. Izaurralde, Don Ort, Allison M. Thomson, and D. Wolfe. “Climate Impacts on Agriculture: Implications for Crop Production.” *Agronomy Journal* 103, no. 2 (2011): 351–70.
- He, Qiuling, Arthur Berg, Yao Li, C. Eduardo Vallejos, and Rongling Wu. “Mapping Genes for Plant Structure, Development and Evolution: Functional Mapping Meets Ontology.” *Trends*

- in *Genetics* 26, no. 1 (2010): 39–46.
- Hejl AAM, FA Einhellig, JA Rasmussen. "Effects of juglone on growth, photosynthesis, and respiration." *Journal of Chemical Ecology* 19, 31993 (1993): 559-568.
- Heldt, H., and U. Flügge. "Subcellular Transport of Metabolites in Plant Cells." *The Biochemistry of Plants* 12 (2013): 49–85.
- Heller, R., S. T. Jensen, P. R. Rosenbaum, and D. S. Small. "Sensitivity Analysis for the Cross-Match Test, With Applications in Genomics." *Journal of the American Statistical Association* 105, no. 491 (2010): 1005–1013.
- Hemert van , M.J., Steensma, H.Y. and van Heusden, G.P. "14-3-3 proteins:key regulators of cell division, signalling and apoptosis." *Bioessays* (2001) 23:936–946
- Henze K. A Badrt, M. Wetirent, R. Cerff, and W. Martin. "A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution." *Proc. Natl. Acad. Sci.* Vol. 92 (1995): 9122-9126.
- Herbert, Ben, Andrew Arthur Gooley, and Keith Leslie Williams. *Electrophoresis Separation Methods*. Google Patents, 2013. <http://www.google.com/patents/US20140114053>.
- Herrmann, Stefanie M., Assaf Anyamba, and Compton J. Tucker. "Recent Trends in Vegetation Dynamics in the African Sahel and Their Relationship to Climate." *Global Environmental Change* 15, no. 4 (2005): 394–404.
- Hervé P, and R Serraj. "Gene Technology and Drought: A Simple Solution for a Complex Trait?" *African Journal of Biotechnology* 8, no. 9 (2009).
- Heymans, Maureen, and Ambuj K. Singh. "Deriving Phylogenetic Trees from the Similarity Analysis of Metabolic Pathways." *Bioinformatics* 19, no. suppl 1 (2003): i138–i146.
- Hide, W., R. Miller, A. Ptitsyn, J. Kelso, C. Gopallakrishnan, and A. Christoffels. "EST Clustering Tutorial." *ISMB in Heidelberg, Germany* 6 (1999). [http://www.lausanne.isb-sib.ch/~galisson/masterCS\\_EPFL/EST\\_tutorial.pdf](http://www.lausanne.isb-sib.ch/~galisson/masterCS_EPFL/EST_tutorial.pdf).
- Hoeven Vd , Rutger, Catherine Ronning, James Giovannoni, Gregory Martin, and Steven Tanksley. "Deductions about the Number, Organization, and Evolution of Genes in the Tomato Genome Based on Analysis of a Large Expressed Sequence Tag Collection and Selective Genomic Sequencing." *The Plant Cell Online* 14, no. 7 (2002): 1441–1456.
- Hohmann-Marriott, Martin F., and Robert E. Blankenship. "Evolution of Photosynthesis." *Annual Review of Plant Biology* 62 (2011): 515–48.

- Holden J, WJ Peacock. "Genes\_crops and the environment" - books.google.com)[BOOK] Cambridge University Press (1993): PP. 1-163.
- Hong, Fangxin, Rainer Breitling, Connor W. McEntee, Ben S. Wittner, Jennifer L. Nemhauser, and Joanne Chory. "RankProd: A Bioconductor Package for Detecting Differentially Expressed Genes in Meta-Analysis." *Bioinformatics* 22, no. 22 (2006): 2825–2827.
- Horan, Kevin, Charles Jang, Julia Bailey-Serres, Ron Mittler, Christian Shelton, Jeff F. Harper, Jian-Kang Zhu, John C. Cushman, Martin Gollery, and Thomas Girke. "Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis." *Plant Physiology* 147, no. 1 (2008): 41–57.
- Horecker, B.L., Tsolas, O. & Lai, C.Y. Aldolases. In *The Enzymes*. (Boyer, P.D., ed), Academic Press, New York, USA, (1972): 213–258,
- Hörtensteiner, S. "Stay-green Regulates Chlorophyll and Chlorophyll-binding Protein Degradation During Senescence." *Trends in Plant Science* 14, no. 3 (2009): 155–162.
- Huang da W, Sherman BT and Lempicki RA. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4, no. 1 (2008): 44–57.
- Huang da W, Sherman BT and Lempicki RA. "Lempicki. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." *Nucleic Acids Research* 37, no. 1 (2009): 1–13.
- Huang, G.-T., Ma, S.-L., Bai, L.-P., Zhang, L., Ma, H., Jia, P., Liu, J., Zhong, M., Guo, Z.-F. "Signal transduction during cold, salt, and drought stresses in plants." *Mol. Biol. Rep.* 39, (2012): 969–987.
- Hubbard, K.E., Siegel, R.S., Valerio, G., Brandt, B., and Schroeder, J.I. "Abscisic acid and CO<sub>2</sub> signalling via calcium sensitivity priming in guard cells, new CDPK mutant phenotypes and a method for improved resolution of stomatal stimulus–response analyses". *Annals of Botany* 109 (2012):5-17.
- Hughes, A. L., and R. Friedman. "Expression Patterns of Duplicate Genes in the Developing Root in *Arabidopsis Thaliana*." *Journal of Molecular Evolution* 60, no. 2 (2005): 247–256.
- Hunt, Colleen, David Butler, and Brian Cullis. "Analysis of Sorghum Breeding Trials Using Pedigree Information." In *Australasian Applied Statistics Conference (Genstat and ASReml) Palm Cove, Australia*, (2011).
- Hulbert, S.H., Richter, T.E., Axtell, J.D., Bennetzen, J.L. "Genetic mapping and characterization of

- sorghum and related crops by means of maize DNA probes." *Proc. Natl. Acad. Sci.* 87, (1990): 4251–4255.
- Hüttemann, Maik, Icksoo Lee, Alena Pecinova, Petr Pecina, Karin Przyklenk, and Jeffrey W. Doan. "Regulation of Oxidative Phosphorylation, the Mitochondrial Membrane Potential, and Their Role in Human Disease." *Journal of Bioenergetics and Biomembranes* 40, no. 5 (2008): 445–56.
- Ilic, Katica, Elizabeth A. Kellogg, Pankaj Jaiswal, Felipe Zapata, Peter F. Stevens, Leszek P. Vincent, Shulamit Avraham, Leonore Reiser, Anuradha Pujar, and Martin M. Sachs. "The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant." *Plant Physiology* 143, no. 2 (2007): 587–599.
- Irizarry, Rafael A., Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal, Bryan C. Frank, Edward Gabrielson, Joe GN Garcia, Joel Geoghegan, and Gregory Germino. "Multiple-Laboratory Comparison of Microarray Platforms." *Nature Methods* 2, no. 5 (2005): 345–50.
- Isokpehi, R.D., Rajnarayanan, R.V., Jeffries, C.D., Oyeleye, T.O., Cohly, H.H. "Integrative sequence and tissue expression profiling of chicken and mammalian aquaporins." *BMC Genomics* 10,(2009): S7.
- Isokpehi, Raphael D., Shaneka S. Simmons, Hari HP Cohly, Stephen IN Ekunwe, Gregorio B. Begonia, and Wellington K. Ayensu. "Identification of Drought-Responsive Universal Stress Proteins in Viridiplantae." *Bioinformatics and Biology Insights* 5 (2011): 41.
- Issa SMA, BL Schulz, NH Packer and NG Karlsson. "Analysis of Mucosal Mucins Separated by SDS-Urea Agarose Polyacrylamide Composite Gel Electrophoresis," *Electrophoresis* 32, no. 24 (2011): 3554–3563.
- Itoh, Takeshi, Tsuyoshi Tanaka, Roberto A. Barrero, Chisato Yamasaki, Yasuyuki Fujii, Phillip B. Hilton, Baltazar A. Antonio, Hideo Aono, Rolf Apweiler, and Richard Bruskiewich. "Curated Genome Annotation of *Oryza Sativa* Ssp. Japonica and Comparative Genome Analysis with *Arabidopsis Thaliana*." *Genome Research* 17, no. 2 (2007): 175–83.
- Jaiswal, Pankaj, Doreen Ware, Junjian Ni, Kuan Chang, Wei Zhao, Steven Schmidt, Xiaokang Pan, Kenneth Clark, Leonid Teytelman, and Samuel Cartinhour. "Gramene: Development and Integration of Trait and Gene Ontologies for Rice." *Comparative and Functional Genomics* 3, no. 2 (2002): 132–136.
- Jaiswal, Pankaj, Shulamit Avraham, Katica Ilic, Elizabeth A. Kellogg, Susan McCouch, Anuradha

- Pujar, Leonore Reiser, Seung Y. Rhee, Martin M. Sachs, and Mary Schaeffer. "Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages." *Comparative and Functional Genomics* 6, no. 7-8 (2005): 388–397.
- Jaiswal, Pankaj, Junjian Ni, Immanuel Yap, Doreen Ware, William Spooner, Ken Youens-Clark, Liya Ren, Chengzhi Liang, Wei Zhao, and Kiran Ratnapu. "Gramene: a Bird's Eye View of Cereal Genomes." *Nucleic Acids Research* 34, no. suppl 1 (2006): D717–D723.
- Jamil, A., S. Riaz, M. Ashraf, and M. R. Foolad. "Gene Expression Profiling of Plants Under Salt Stress." *Critical Reviews in Plant Sciences* 30, no. 5 (2011): 435–458.
- Ji, Kuixian, Yangyang Wang, Weining Sun, Qiaojun Lou, Hanwei Mei, Shihua Shen, and Hui Chen. "Drought-Responsive Mechanisms in Rice Genotypes with Contrasting Drought Tolerance during Reproductive Stage." *Journal of Plant Physiology* 169, no. 4 (2012): 336–343.
- Jia, H., S. Zhang, M. Ruan, Y. Wang, and C. Wang. "Analysis and Application of RD29 Genes in Abiotic Stress Response." *Acta Physiologiae Plantarum* (2012): 1–12.
- Jiang L, L He, M Fountoulakis. "Comparison of protein precipitation methods for sample preparation prior to proteomic analysis." *Journal of Chromatography A*, 1023 (2004): 317–320.
- Jiang, Shan-Shan, Xiao-Na Liang, Xin Li, Shun-Li Wang, Dong-Wen Lv, Chao-Ying Ma, Xiao-Hui Li, Wu-Jun Ma, and Yue-Ming Yan. "Wheat Drought-Responsive Grain Proteome Analysis by Linear and Nonlinear 2-DE and MALDI-TOF Mass Spectrometry." *International Journal of Molecular Sciences* 13, no. 12 (2012): 16065–83.
- Jofuku, K. D., B. G. Den Boer, M. Van Montagu, and J. K. Okamoto. "Control of Arabidopsis Flower and Seed Development by the Homeotic Gene APETALA2." *The Plant Cell Online* 6, no. 9 (1994): 1211–1225.
- Jogaiah, Sudisha, Sharathchandra Ramsandra Govind, and Lam-Son Phan Tran. "Systems Biology-Based Approaches toward Understanding Drought Tolerance in Food Crops." *Critical Reviews in Biotechnology* 33, no. 1 (2013): 23–39.
- Johnson CJ, JP Bennett, SM Biro, JC Duque-Velasquez, CM. Rodriguez, RA. Bessen, TE. Roche. "Degradation of the Disease-Associated Prion Protein by a Serine Protease from Lichens," *PLoS ONE* 6, no. 5 (May 11, 2011): e19836, doi:10.1371/journal.pone.0019836.
- Jones, Hamlyn G. "Monitoring Plant and Soil Water Status: Established and Novel Methods Revisited and Their Relevance to Studies of Drought Tolerance." *Journal of Experimental Botany* 58, no. 2 (2007): 119–130. doi:10.1093/jxb/erl118.

- Jordan, D. R., C. H. Hunt, A. W. Cruickshank, A. K. Borrell, and R. G. Henzell. "The Relationship between the Stay-Green Trait and Grain Yield in Elite Sorghum Hybrids Grown in a Range of Environments." *Crop Science* 52, no. 3 (2012): 1153–61.
- Jorgensen TJ, I. Ruczinski, B. Kessing, M. W. Smith, Y. Y. Shugart, and A. J. Alberg. "Hypothesis-Driven Candidate Gene Association Studies: Practical Design and Analytical Considerations." *Am. J. Epidemiol.* 170, no. 8 (2009): 986–993.
- Jorrin-Novo JV., and M. Maldonado, S. E. Zome, L. Valledor, M. A. Castillejo, M Curto, J. Valero, B. Sghaier, G. Donoso, I. Redondo. "Plant proteomics update (2007–2008): second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards". *Journal of proteomics* 72 (2009): 285–314.
- Junge, Wolfgang, Hendrik Sielaff, and Siegfried Engelbrecht. "Torque Generation and Elastic Power Transmission in the Rotary FOF1-ATPase." *Nature* 459, no. 7245 (2009): 364–70.
- Kakumanu, Akshay, Madana MR Ambavaram, Curtis Klumas, Arjun Krishnan, Utlwang Batlang, Elijah Myers, Ruth Grene, and Andy Pereira. "Effects of Drought on Gene Expression in Maize Reproductive and Leaf Meristem Tissue Revealed by RNA-Seq." *Plant Physiology* 160, no. 2 (2012): 846–67.
- Kanehisa, Minoru, and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28, no. 1 (2000): 27–30.
- Karpilov, IuS, I. L. Novitskaia, A. N. Kuz'min, A. I. Maslov, and E. I. Popova. "Reversibility of Glycolysis in Leaves of C4-Plants." *Biokhimiia (Moscow, Russia)* 42, no. 10 (1977): 1810–16.
- Karpinets, Tatiana V., Byung H. Park, Mustafa H. Syed, Martin G. Klotz, and Edward C. Uberbacher. "Metabolic Environments and Genomic Features Associated with Pathogenic and Mutualistic Interactions Between Bacteria and Plants." *Molecular Plant-Microbe Interactions* 27, no. 7 (2014): 664–77.
- Kassahun, B., Bidinger, F.R., Hash, C.T., Kuruvinashetti, M.S. "Stay-green expression in early generation sorghum [*Sorghum bicolor* (L.) Moench] QTL introgression lines." *Euphytica* 172, (2011): 351–362.
- Kattge, Jens, Kiona Ogle, Gerhard Bönisch, Sandra Díaz, Sandra Lavorel, Joshua Madin, Karin Nadrowski, Stephanie Nöllert, Karla Sartor, and Christian Wirth. "A Generic Structure for Plant Trait Databases." *Methods in Ecology and Evolution* 2, no. 2 (2011): 202–213.
- Kawahara, Y., de la Bastide, M., Hamilton J. P., Kanamori, H., McCombie, W. R., Ouyang, S.,

- Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., Matsumoto, T. "Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data." *Rice* 6, (2013):4.
- Kebede, H., P. K. Subudhi, D. T. Rosenow, and H. T. Nguyen. "Quantitative Trait Loci Influencing Drought Tolerance in Grain Sorghum (*Sorghum Bicolor* L. Moench)." *TAG Theoretical and Applied Genetics* 103, no. 2 (2001): 266–76.
- Keller, Oliver, Martin Kollmar, Mario Stanke, and Stephan Waack. "A Novel Hybrid Gene Prediction Method Employing Protein Multiple Sequence Alignments." *Bioinformatics* 27, no. 6 (2011): 757–763.
- Kent, W. J. "BLAT-the BLAST-Like Alignment Tool." *Genome Research* 12, no. 4 (2002): 656–664. doi:10.1101/gr.229202.
- Keren, Hadas, Galit Lev-Maor, and Gil Ast. "Alternative Splicing and Evolution: Diversification, Exon Definition and Function." *Nature Reviews Genetics* 11, no. 5 (2010): 345–355.
- Khan M. A. "Current Status of Genomic Based Approaches to Enhance Drought Tolerance in Rice (*Oryza Sativa* L.), an Over View." *Molecular Plant Breeding* 2 (2012): 00477.
- Kiatkamjornwong S, N Siwarungson, and A Nganbunsri. "In Situ Immobilization of Alkaline Protease during Inverse Suspension Polymerization of Polyacrylamide and Poly (acrylamide-Co-Methacrylic Acid) Hydrogel Beads," *Journal of Applied Polymer Science* 73, no. 11 (1999): 2273–2291.
- Kilian, Joachim, Dion Whitehead, Jakub Horak, Dierk Wanke, Stefan Weinl, Oliver Batistic, Cecilia D'Angelo, Erich Bornberg-Bauer, Jörg Kudla, and Klaus Harter. "The AtGenExpress Global Stress Expression Data Set: Protocols, Evaluation and Model Data Analysis of UV-B Light, Drought and Cold Stress Responses." *The Plant Journal* 50, no. 2 (2007): 347–363.
- Kim, Dea-Wook, Randeep Rakwal, Ganesh Kumar Agrawal, Young-Ho Jung, Junko Shibato, Nam-Soo Jwa, Yumiko Iwahashi, *et al.* "A Hydroponic Rice Seedling Culture Model System for Investigating Proteome of Salt Stress in Rice Leaf." *Electrophoresis* 26, no. 23 (2005): 4521–39.
- Kim, Eddo, Alon Magen, and Gil Ast. "Different Levels of Alternative Splicing among Eukaryotes." *Nucleic Acids Research* 35, no. 1 (2007): 125–131. doi:10.1093/nar/gkl924.
- Kim, Jung-whan, and Chi V. Dang. "Multifaceted Roles of Glycolytic Enzymes." *Trends in*



- Biochemical Sciences* 30, no. 3 (March 2005): 142–50. doi:10.1016/j.tibs.2005.01.005.
- Kim, Namshin, Seokmin Shin, and Sanghyuk Lee. “ECgene: Genome-based EST Clustering and Gene Modeling for Alternative Splicing.” *Genome Research* 15, no. 4 (2005): 566–576.
- Kim, Yeon-Ok, Jin Sun Kim, and Hunseung Kang. “Cold-Inducible Zinc Finger-Containing Glycine-Rich RNA-Binding Protein Contributes to the Enhancement of Freezing Tolerance in *Arabidopsis thaliana*.” *The Plant Journal* 42, no. 6 (2005): 890–900.
- Kirschner, M. W. “Department of Systems Biology Harvard Medical School Boston, Massachusetts 02115.” *Cell* 121 (2005): 503–504.
- Kisman, Derek, Ming Li, Bin Ma, and Li Wang. “tPatternHunter: Gapped, Fast and Sensitive Translated Homology Search.” *Bioinformatics* 21, no. 4 (2005): 542–544.
- Klassen, JL and CR Currie. “Gene Fragmentation in Bacterial Draft Genomes: Extent, Consequences and Mitigation.” *BMC Genomics* 13, no. 1 (2012): 14. doi:10.1186/1471-2164-13-14.
- Klein, P. E., R. R. Klein, S. W. Cartinhour, P. E. Ulanich, J. Dong, J. A. Obert, D. T. Morishige, S. D. Schlueter, K. L. Childs, and M. Ale. “A High-throughput AFLP-based Method for Constructing Integrated Genetic and Physical Maps: Progress Toward a Sorghum Genome Map.” *Genome Research* 10, no. 6 (2000): 789–807.
- Klute A. “Field Capacity and Available Water Capacity,” 2003, <https://dl.sciencesocieties.org/publications/books/articles/sssabookseries/methodsofsoilan1/901?show-t-%20%20f=tables&wrapper=no>.
- Knight H, MR Knight. "Abiotic stress signalling pathways: specificity and cross-talk." *Trends in plant science* 6, no.6 (2001): 262–267.
- Koenig T , B H. Menze<sup>1</sup>, M. Kirchner, F. Monigatti, K. C. Parker, T. Patterson, J.J. Steen, F. A. Hamprecht, H. Steen."Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics." *J. Proteome Res.* **7**, no. 9 (2008): 3708–17.
- Koonin, Eugene V., and Michael Y. Galperin. “Genome Annotation and Analysis,” 2003. <http://www.ncbi.nlm.nih.gov/books/NBK20253/>.
- Koonin, Eugene V., Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Dmitri M. Krylov, Kira S. Makarova, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskaya, and B. Sridhar Rao. “A Comprehensive Evolutionary Classification of Proteins Encoded in Complete Eukaryotic Genomes.” *Genome Biology* 5, no. 2 (2004): R7.

- Korf, Ian. "Gene Finding in Novel Genomes." *BMC Bioinformatics* 5, no. 1 (2004): 59.
- Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., Muñoz, M.J. "Alternative splicing: a pivotal step between eukaryotic transcription and translation." *Nat. Rev. Mol. Cell Biol.* 14, (2013): 153–165.
- Kreps, Joel A., Yajun Wu, Hur-Song Chang, Tong Zhu, Xun Wang, and Jeff F. Harper. "Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress." *Plant Physiology* 130, no. 4 (2002): 2129–41.
- Kresovich, S., B. Barbazuk, J. A. Bedell, A. Borrell, C. R. Buell, J. Burke, S. Clifton, M. M. Cordonnier-Pratt, S. Cox, and J. Dahlberg. "Toward Sequencing the Sorghum Genome: a US National Science Foundation-sponsored Workshop Report." *Plant Physiology* 138, no. 4 (2005): 1898.
- Kromer S. "Respiration during photosynthesis." *Annual review of plant biology* 46 (1995): 45-70.
- Kumar S, A., S. I. Alam, N. Sengupta, and R. Sarin. "Differential Proteomic Analysis of Salt Stress Response in Sorghum Bicolor Leaves." *Environmental and Experimental Botany* 71, no. 2 (2011): 321–328.
- Kushwaha HR, AK Singh, S K Sopory, SL Singla-Pareek and A Pareek. Genome wide expression analysis of CBS domain containing proteins in *Arabidopsis thaliana* (L.) Heynh and *Oryza sativa* L. reveals their developmental and stress regulation. *BMC Genomics*, 10 (2009): doi:10.1186/1471-2164-10-200.
- Kuwano Y, Olvera J, Wool IG . "The primary structure of rat ribosomal protein L38". *Biochem. Biophys. Res. Commun.* 175, 2(1991): 551–5.
- Lakshmanan, Meiyappan, Zhaoyang Zhang, Bijayalaxmi Mohanty, Jun-Young Kwon, Hong-Yeol Choi, Hyung-Jin Nam, Dong-Il Kim, and Dong-Yup Lee. "Elucidating Rice Cell Metabolism under Flooding and Drought Stresses Using Flux-Based Modeling and Analysis." *Plant Physiology* 162, no. 4 (2013): 2140–50.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, and W. FitzHugh. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409, no. 6822 (2001): 860–921.
- Langridge, Peter, and Delphine Fleury. "Making the Most of 'omics' for Crop Breeding." *Trends in Biotechnology* 29, no. 1 (2011): 33–40.
- Lao X, J-ichi Azuma and M Sakamoto. "Two cytosolic aldolases show different expression patterns

- during shoot elongation in Moso bamboo, *Phyllostachys pubescens* Mazel.” *Physiologia Plantarum*, 149 (3), (2013): 422–431.
- Larkindale, Jane, and Marc R. Knight. “Protection Against Heat Stress-induced Oxidative Damage in Arabidopsis Involves Calcium, Abscisic Acid, Ethylene, and Salicylic Acid.” *Plant Physiology* 128, no. 2 (2002): 682–695.
- Lay F.T. and Anderson M.A. “Defensins – Components of the Innate Immune System in Plants.” *Current Protein and Peptide Science* 6 (2005): 85-101.
- Le, D.T., Aldrich, D.L., Valliyodan, B., Watanabe, Y., Van Ha, C., Nishiyama, R., Guttikonda, S.K., Quach, T.N., Gutierrez-Gonzalez, J.J., Tran, L.-S.P., others. “Evaluation of candidate reference genes for normalization of quantitative RT-PCR in soybean tissues under various abiotic stress conditions.” *PLoS One* 7, (2012): e46487.
- Lee, Dong-Gi, Norma L. Houston, Severin E. Stevenson, Gregory S. Ladics, Scott McClain, Laura Privalle, and Jay J. Thelen. “Mass Spectrometry Analysis of Soybean Seed Proteins: Optimization of Gel-free Quantitative Workflow.” *Analytical Methods* 2, no. 10 (2010): 1577–1583.
- Lee, Hojong, Yan Guo, Masaru Ohta, Liming Xiong, Becky Stevenson, and Jian-Kang Zhu. “LOS2, a Genetic Locus Required for Cold-Responsive Gene Transcription Encodes a Bi-Functional Enolase.” *The EMBO Journal* 21, no. 11 (2002): 2692–2702.
- Lee, Insuk, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. “Prioritizing Candidate Disease Genes by Network-Based Boosting of Genome-Wide Association Data.” *Genome Research* 21, no. 7 (2011): 1109–1121.
- Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, and I. Simon. “Transcriptional Regulatory Networks in *Saccharomyces Cerevisiae*.” *Science Signalling* 298, no. 5594 (2002): 799.
- Lee, Y., J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung, and J. Quackenbush. “The TIGR Gene Indices: Clustering and Assembling EST and Known Genes and Integration with Eukaryotic Genomes.” *Nucleic Acids Research* 33, no. suppl 1 (2005): D71–D74.
- Leroy, P., Guilhot, N., Sakai, H., Bernard, A., Choulet, F., Theil, S., Reboux, S., Amano, N., Flutre, T., Pelegriin, C., others. “TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes.” *Front. Plant Sci.* 3, (2012).

- Leung J. and Giraudat J., 1999. "ABSCISIC ACID SIGNAL TRANSDUCTION." *Annual Review of Plant Physiology and Plant Molecular Biology* 49 (1999): 199-222.
- Li H, and R Durbin. "Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 26, no. 5 (2010): 589–595. doi:10.1093/bioinformatics/btp698.
- Li W., W Yang, XJ Wang. "Pseudogenes: Pseudo or Real Functional Elements?" *Journal of Genetics and Genomics* 40, 4 (2013): 171–177.
- Li, Li, Christian J. Stoeckert, and David S. Roos. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13, no. 9 (2003): 2178–2189.
- Li, Wanlong, and Bikram S. Gill. "The Colinearity of the Sh2/A1 Orthologous Region in Rice, Sorghum and Maize Is Interrupted and Accompanied by Genome Expansion in the Triticeae." *Genetics* 160, no. 3 (2002): 1153–1162.
- Li, Yong-Fang, Yixing Wang, Yuhong Tang, Vijaya Gopal Kakani, and Ramamurthy Mahalingam. "Transcriptome Analysis of Heat Stress Response in Switchgrass (*Panicum Virgatum* L.)." *BMC Plant Biology* 13, no. 1 (2013): 153.
- Liang, Chengzhi, Pankaj Jaiswal, Claire Hebbard, Shuly Avraham, Edward S. Buckler, Terry Casstevens, Bonnie Hurwitz, Susan McCouch, Junjian Ni, and Anuradha Pujar. "Gramene: a Growing Plant Comparative Genomics Resource." *Nucleic Acids Research* 36, no. suppl 1 (2008): D947–D953.
- Lim, H., Eng, J., Yates III, J.R., Tollaksen, S.L., Giometti, C.S., Holden, J.F., Adams, M.W.W., Reich, C.I., Olsen, G.J., Hays, L.G. "Identification of 2D-gel proteins: A comparison of MALDI/TOF peptide mass mapping to  $\mu$  LC-ESI tandem mass spectrometry." *J. Am. Soc. Mass Spectrom* 14, (2003): 957–970. doi:10.1016/S1044-0305(03)00144-2.
- Lippman, Zachary, Anne-Valérie Gendrel, Michael Black, Matthew W. Vaughn, Neilay Dedhia, W. Richard McCombie, Kimberly Lavine, Vivek Mittal, Bruce May, and Kristin D. Kasschau. "Role of Transposable Elements in Heterochromatin and Epigenetic Control." *Nature* 430, no. 6998 (2004): 471–476.
- Liu F, M Ye, C Wang, Z Hu, Y Zhang, H Qin, K Cheng, and H Zou. "Polyacrylamide Gel with Switchable Trypsin Activity for Analysis of Proteins," *Analytical Chemistry* 85, no. 15 (2013): 7024–7028.
- Liu M., Z. Zhang, T. Zang, C. Spahr, J. Cheetham, D. Ren, and Z. S. Zhou. "Discovery of Undefined Protein Cross-Linking Chemistry: A Comprehensive Methodology Utilizing  $^{18}\text{O}$ -

- Labeling and Mass Spectrometry". *Anal. Chem.* 85 (2013): 5900–5908.
- Liu, Huizhi, Wang, Xiaoe, Zhang, Huijuan, Yang, Yayun, Ge, Xiuchun, Song, Fengming. "A rice serine carboxypeptidase-like gene *OsBISCP1* is involved in regulation of defense responses against biotic and oxidative stress." *Gene* 420, no. 1 (2008): 57–65.
- Liu, Peng, Lina Yin, Xiping Deng, Shiwen Wang, Kiyoshi Tanaka, and Suiqi Zhang. "Aquaporin-Mediated Increase in Root Hydraulic Conductance Is Involved in Silicon-Induced Improved Root Water Uptake under Osmotic Stress in Sorghum Bicolor L." *Journal of Experimental Botany*, 2014, eru220.
- Liu, Q., M. Kasuga, Y. Sakuma, H. Abe, S. Miura, K. Yamaguchi-Shinozaki, and K. Shinozaki. "Two Transcription Factors, DREB1 and DREB2, with an EREBP/AP2 DNA Binding Domain Separate Two Cellular Signal Transduction Pathways in Drought-and Low-temperature-responsive Gene Expression, Respectively, in Arabidopsis." *The Plant Cell Online* 10, no. 8 (1998): 1391–1406.
- Lobos SR and GC Mora. "Alteration in the Electrophoretic Mobility of OmpC due to Variations in the Ammonium Persulfate Concentration in Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis," *Electrophoresis* 12, no. 6 (1991): 448–450.
- Lomax, Jane. "Get Ready to GO! A Biologist's Guide to the Gene Ontology." *Briefings in Bioinformatics* 6, no. 3 (2005): 298–304.
- Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. "Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm." *Nucleic Acids Research* 33, no. 20 (2005): 6494–6506.
- Long SP, Xin-Guang Zhu, SL Naidu and DR Ort. "Can improvement in photosynthesis increase crop yields?" *Plant, Cell* 29, no. 3 (2006): 315–330.
- Lopes, M. S., H. D. Saglam, M. Ozdogan, and M. Reynolds. "Traits Associated with Winter Wheat Grain Yield in Central and West Asia." *Journal of Integrative Plant Biology* (2014).
- Lu, Min, Sheng Ying, Deng-Feng Zhang, Yun-Su Shi, Yan-Chun Song, Tian-Yu Wang, and Yu Li. "A Maize Stress-responsive NAC Transcription Factor, *ZmSNAC1*, Confers Enhanced Tolerance to Dehydration in Transgenic Arabidopsis." *Plant Cell Reports* 31, no. 9 (2012): 1701–1711.
- Ludevid MD, M Torrent, JA Martinez-Izquierdo, P Puigdomenech, J Palau. "Subcellular Localization of Glutelin-2 in Maize (*Zea Mays* L.) Endosperm," *Plant Molecular Biology* 3, no.

- 4 (July 1, 1984): 227–234, doi:10.1007/BF00029658.
- Lynch, Michael, and Vaishali Katju. “The Altered Evolutionary Trajectories of Gene Duplicates.” *TRENDS in Genetics* 20, no. 11 (2004): 544–549.
- Ma, Yue, Izabela Szostkiewicz, Arthur Korte, Danièle Moes, Yi Yang, Alexander Christmann, and Erwin Grill. “Regulators of PP2C Phosphatase Activity Function as Abscisic Acid Sensors.” *Science* 324, no. 5930 (2009): 1064–1068.
- Mace, Emma S., Shuaishuai Tai, Edward K. Gilding, Yanhong Li, Peter J. Prentis, Lianle Bian, Bradley C. Campbell, Wushu Hu, David J. Innes, and Xuelian Han. “Whole-Genome Sequencing Reveals Untapped Genetic Potential in Africa’s Indigenous Cereal Crop Sorghum.” *Nature Communications* 4 (2013).  
<http://www.nature.com/ncomms/2013/130827/ncomms3320/full/ncomms3320.html>.
- Mackintosh, J.A., Veal, D.A., Karuso, P. “Fluoroprofile, a fluorescence-based assay for rapid and sensitive quantitation of proteins in solution.” *Proteomics* 5, (2005): 4673–4677. doi:10.1002/pmic.200500095.
- Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T. “Entrez Gene: gene-centered information at NCBI.” *Nucleic Acids Res.* 35,(2007): D26–D31.
- Mahajan, S., and N. Tuteja. “Cold, Salinity and Drought Stresses: An Overview.” *Archives of Biochemistry and Biophysics* 444, no. 2 (2005): 139–158.
- Mahajan, S., Pandey, G.K., Tuteja, N. “Calcium-and salt-stress signaling in plants: shedding light on SOS pathway.” *Arch. Biochem. Biophys.* 471, (2008): 146–158.
- Mahdevar, G., M. Sadeghi, and A. Nowzari-Dalini. “Transcription Factor Binding Sites Detection by Using Alignment-based Approach.” *Journal of Theoretical Biology* (2012).
- Makarova, Kira S., Alexander V. Sorokin, Pavel S. Novichkov, Yuri I. Wolf, and Eugene V. Koonin. “Clusters of Orthologous Genes for 41 Archaeal Genomes and Implications for Evolutionary Genomics of Archaea.” *Biol Direct* 2 (2007): 33.
- Manavalan, L. P., H. T. Nguyen, and S. Shabala. “1 Drought Tolerance in Crops: Physiology to Genomics.” *Plant Stress Physiology* (2012): 1.
- Manickavelu, A., K. Kawaura, K. Oishi, T. Shin, Y. Kohara, N. Yahiaoui, B. Keller, R. Abe, A. Suzuki, and T. Nagayama. “Comprehensive Functional Analyses of Expressed Sequence Tags in Common Wheat (*Triticum Aestivum*).” *DNA Research* 19, no. 2 (2012): 165–177.
- Manners JM , RE Casu. “Transcriptome analysis and functional genomics of sugarcane”. *Tropical*

*Plant Biology* 4 (2011): 9-21.

Marraccini, P., Vinecky, F., Alves, G.S., Ramos, H.J., Elbelt, S., Vieira, N.G., Carneiro, F.A., Sujii, P.S., Alekcevetch, J.C., Silva, V.A., others. "Differentially expressed genes and proteins upon drought acclimation in tolerant and sensitive genotypes of *Coffea canephora*." *J. Exp. Bot.* (2012): ers103.

Marchler-Bauer A, C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, and others. "CDD: conserved domains and protein three-dimensional structure." *Nucleic Acids Res.* 41 (2013): D348–D352.

Martin, David, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. "GOToolBox: Functional Analysis of Gene Datasets Based on Gene Ontology." *Genome Biology* 5, no. 12 (2004): R101.

Marvin L., A Millar, V. Saulot, N. Machour, R. Charlionet, F Tron and C. Lange . "Identification of proteins from one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis using electrospray quadrupole-time-of-flight tandem mass spectrometry." *MASS SPECTROMETRY Rapid Commun. Mass Spectrom* 14 (2000): 1287–1292.

Masuka, Benhilda, Jose Luis Araus, Biswanath Das, Kai Sonder, and Jill E. Cairns. "Phenotyping for Abiotic Stress Tolerance in MaizeF." *Journal of Integrative Plant Biology* 54, no. 4 (2012): 238–249.

Matas, Antonio J., Trevor H. Yeats, Gregory J. Buda, Yi Zheng, Subhasish Chatterjee, Takayuki Tohge, Lalit Ponnala, Avital Adato, Asaph Aharoni, and Ruth Stark. "Tissue-and Cell-type Specific Transcriptome Profiling of Expanding Tomato Fruit Provides Insights into Metabolic and Regulatory Specialization and Cuticle Formation." *The Plant Cell Online* 23, no. 11 (2011): 3893–3910.

Matsumoto, Takashi, Tsuyoshi Tanaka, Hiroaki Sakai, Naoki Amano, Hiroyuki Kanamori, Kanako Kurita, Ari Kikuta, Kozue Kamiya, Mayu Yamamoto, and Hiroshi Ikawa. "Comprehensive Sequence Analysis of 24,783 Barley Full-length cDNAs Derived from 12 Clone Libraries." *Plant Physiology* 156, no. 1 (2011): 20–28.

Maurer HR. "Disc Electrophoresis and Related Techniques of Polyacrylamide Gel Electrophoresis." Walter de Gruyter, (1978).

Mcclain, W. R. "Comparative Evaluations of Rice and a Sorghum–sudangrass Hybrid as Crawfish Forage Crops." *The Progressive Fish-culturist* 59, no. 3 (1997): 206–212.

- McJury M, M Oldham, VP Cosgrove, PS Murphy, S Doran, MO Leach, and S Webb. "Radiation Dosimetry Using Polymer Gels: Methods and Applications.," 2014, <http://www.birpublications.org/doi/abs/10.1259/bjr.73.873.11064643>.
- McKee, T. B., N. J. Doesken, and J. Kleist. "The Relationship of Drought Frequency and Duration to Time Scales." In *Proceedings of the 8th Conference on Applied Climatology*, 17:179–183. American Meteorological Society Boston, MA, 1993.
- Méchin V, C Damerval, and M Zivy. "Total Protein Extraction with TCA-Acetone," in *Plant Proteomics* (2007): 1–8, <http://link.springer.com/protocol/10.1385/1-59745-227-0:1>.
- Meena, R. K., S. B. Verulkar, and G. Chandel. "Nutrient Characters Analysis in Rice Genotypes Under Different Environmental Conditions." *Bull. Environ. Pharmacol. Life Sci.; Volume 1* (2012): 61–64.
- Menz, M. A., R. R. Klein, J. E. Mullet, J. A. Obert, N. C. Unruh, and P. E. Klein. "A High-density Genetic Map of Sorghum Bicolor (L.) Moench Based on 2926 AFLP®, RFLP and SSR Markers." *Plant Molecular Biology* 48, no. 5 (2002): 483–499.
- Mewes, H. W., D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil. "MIPS: a Database for Genomes and Protein Sequences." *Nucleic Acids Research* 30, no. 1 (2002): 31–34.
- Michalek, W., W. Weschke, K. P. Pleissner, and A. Graner. "EST Analysis in Barley Defines a UniGene Set Comprising 4,000 Genes." *TAG Theoretical and Applied Genetics* 104, no. 1 (2002): 97–103.
- Mihara, Motohiro, Takeshi Itoh, and Takeshi Izawa. "SALAD Database: A Motif-Based Database of Protein Annotations for Plant Comparative Genomics." *Nucleic Acids Research* 38, no. suppl 1 (2010): D835–D842.
- Miller, R. T., A. G. Christoffels, C. Gopalakrishnan, J. Burke, A. A. Ptitsyn, T. R. Broveak, and W. A. Hide. "A Comprehensive Approach to Clustering of Expressed Human Gene Sequence: The Sequence Tag Alignment and Consensus Knowledge Base." *Genome Research* 9, no. 11 (1999): 1143–1155.
- Ming, R., S. C. Liu, Y. R. Lin, J. Da Silva, W. Wilson, D. Braga, A. Van Deynze, T. F. Wenslaff, K. K. Wu, and P. H. Moore. "Detailed Alignment of Saccharum and Sorghum Chromosomes: Comparative Organization of Closely Related Diploid and Polyploid Genomes." *Genetics* 150, no. 4 (1998): 1663–1682.



- Mir, Reyazul Rouf, Mainassara Zaman-Allah, Nese Sreenivasulu, Richard Trethowan, and Rajeev K. Varshney. "Integrated Genomics, Physiology and Breeding Approaches for Improving Drought Tolerance in Crops." *Theoretical and Applied Genetics* 125, no. 4 (2012): 625–645.
- Mishra, Ashok K., and Vijay P. Singh. "A Review of Drought Concepts." *Journal of Hydrology* 391, no. 1 (2010): 202–216.
- Mitchell, Rowan A. C., Paul Dupree, and Peter R. Shewry. "A Novel Bioinformatics Approach Identifies Candidate Genes for the Synthesis and Feruloylation of Arabinoxylan." *Plant Physiology* 144, no. 1 (2007): 43–53. doi:10.1104/pp.106.094995.
- Mitchell, Peter. "Chemiosmotic Coupling in Energy Transduction: A Logical Development of Biochemical Knowledge." In *Membrane Structure and Mechanisms of Biological Energy Transduction*, (1973): 5–24. [http://link.springer.com/chapter/10.1007/978-1-4684-2016-6\\_2](http://link.springer.com/chapter/10.1007/978-1-4684-2016-6_2).
- . "Chemiosmotic Coupling in Oxidative and Photosynthetic Phosphorylation." *Biological Reviews* 41, no. 3 (1966): 445–501.
- Mittler, Ron, and Eduardo Blumwald. "Genetic Engineering for Modern Agriculture: Challenges and Perspectives." *Annual Review of Plant Biology* 61 (2010): 443–462.
- Miyao M. "Molecular Evolution and Genetic Engineering of C4 Photosynthetic Enzymes." *Journal of Experimental Botany* 54, no. 381 (2003): 179–89.
- Mizoi, J., K. Shinozaki, and K. Yamaguchi-Shinozaki. "AP2/ERF Family Transcription Factors in Plant Abiotic Stress Responses." *Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819, no. 2 (2012): 86–96.
- Modrek, Barmak, and Christopher J. Lee. "Alternative Splicing in the Human, Mouse and Rat Genomes Is Associated with an Increased Frequency of Exon Creation And/or Loss." *Nature Genetics* 34, no. 2 (2003): 177–180.
- Modrek, Barmak, and Christopher Lee. "A Genomic View of Alternative Splicing." *Nature Genetics* 30, no. 1 (2002): 13–19.
- Monneveux P and JM Ribaut. "Drought Phenotyping in Crops: From Theory to Practice." *Available at Generation Challenge Program Website Wwww. Generationcp. Org* (2011).
- Monson RK. "Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis." *International Journal of Plant Sciences* 164, no. S3 (2003): S43-S54.
- Moore, Jason H., Folkert W. Asselbergs, and Scott M. Williams. "Bioinformatics Challenges for Genome-Wide Association Studies." *Bioinformatics* 26, no. 4 (2010): 445–455.

- Morgulis, A., E. M. Gertz, A. A. Schäffer, and R. Agarwala. "A Fast and Symmetric DUST Implementation to Mask Low-complexity DNA Sequences." *Journal of Computational Biology* 13, no. 5 (2006): 1028–1040.
- Morris, Geoffrey P., Punna Ramu, Santosh P. Deshpande, C. Thomas Hash, Trushar Shah, Hari D. Upadhyaya, Oscar Riera-Lizarazu, Patrick J. Brown, Charlotte B. Acharya, and Sharon E. Mitchell. "Population Genomic and Genome-Wide Association Studies of Agroclimatic Traits in Sorghum." *Proceedings of the National Academy of Sciences* 110, no. 2 (2013): 453–58.
- Mott, Richard. "EST\_GENOME: A Program to Align Spliced DNA Sequences to Unspliced Genomic DNA." *Computer Applications in the Biosciences: CABIOS* 13, no. 4 (1997): 477–78. <http://bioinformatics.oxfordjournals.org/content/13/4/477.short>.
- Mukhopadhyay, Snehasis, Changhong Tang, Jeffery Huang, Mulong Yu, and M. Palakal. "A Comparative Study of Genetic Sequence Classification Algorithms." In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop On*, 57–66. IEEE, 2002.
- Mullet, J. E., R. R. Klein, and P. E. Klein. "< I> Sorghum Bicolor</i>—an Important Species for Comparative Grass Genomics and a Source of Beneficial Genes for Agriculture." *Current Opinion in Plant Biology* 5, no. 2 (2002): 118–121.
- Mungall, Christopher J., Georgios V. Gkoutos, Cynthia L. Smith, Melissa A. Haendel, Suzanna E. Lewis, and Michael Ashburner. "Integrating Phenotype Ontologies across Multiple Species." *Genome Biology* 11, no. 1 (2010): R2.
- Munnik, T. and Meijer, H. J. "Osmotic stress activates distinct lipid and MAPK signalling pathways in plants". *FEBS Lett.* (2001). **498**: 172-178.
- Mutava, R. N. "Characterization of Grain Sorghum for Physiological and Yield Traits Associated with Drought Tolerance." Kansas State University, 2009.
- Nagaraj, N., G. Basavaraj, and P. P. Rao. "Policy Brief on Future Outlook and Options for Target Crops: The Sorghum and Pearl Millet Economy of India". International Crops Research Institute for the Semi-Arid Tropics, Patancheru, India. [www.icrisat.org](http://www.icrisat.org) (2011) Pp: 1-16.
- Nagaraj, S. H., R. B. Gasser, and S. Ranganathan. "A Hitchhiker's Guide to Expressed Sequence Tag (EST) Analysis." *Briefings in Bioinformatics* 8, no. 1 (2007): 6–21.
- Naithani, S. "Plant Pathways & Gene Expression Analysis in Gramene." In *Plant and Animal Genome XXI Conference*. Plant and Animal Genome, 2013. <http://www.gramene.org/pathway/sorghumcyc.html>.

- Nakashima, K., Y. Ito, and K. Yamaguchi-Shinozaki. "Transcriptional Regulatory Networks in Response to Abiotic Stresses in Arabidopsis and Grasses." *Plant Physiology* 149, no. 1 (2009): 88–95.
- Namslauer, Andreas, Anna Aagaard, Andromachi Katsonouri, and Peter Brzezinski. "Intramolecular Proton-Transfer Reactions in a Membrane-Bound Proton Pump: The Effect of pH on the Peroxy to Ferryl Transition in Cytochrome c Oxidase,□." *Biochemistry* 42, no. 6 (2003): 1488–98.
- Narain, Prem. "Quantitative Genetics: Past and Present." *Molecular Breeding* 26, no. 2 (2010): 135–143.
- Ndimba BK, S Chivasa, JM. Hamilton, WJ Simon and AR Slabas. "Proteomic Analysis of Changes in the Extracellular Matrix of Arabidopsis Cell Suspension Cultures Induced by Fungal Elicitors," *Proteomics* 3, no. 6 (2003): 1047–1059.
- Ndimba BK, R Ngara. "Sorghum and sugarcane proteomics." *Genomics of the Saccharinae* 11 (2013): 141-168.
- Ndimba BK, Thomas LA and Ngara R. "Sorghum 2-Dimensional Proteome Profiles and Analysis of HSP70 Expression Under Salinity Stress." *Kasetsart J. (Nat. Sci.)* 44 (2010): 768-775.
- Ndimba, Bongani K., Stephen Chivasa, William J. Simon, and Antoni R. Slabas. "Identification of Arabidopsis Salt and Osmotic Stress Responsive Proteins Using Two-Dimensional Difference Gel Electrophoresis and Mass Spectrometry." *Proteomics* 5, no. 16 (2005): 4185–96.
- Nelson EB, B Lungwitz, K Dismuke, M Samuel, G Salamat, T Hughes, J Lee, P Fletcher, D Fu, R Hutchins, M Parris, GJ Tustin. "Viscosity Reduction of Viscoelastic Surfactant Based Fluids." (2005), <http://www.google.com/patents/US6881709>.
- Ner-Gaon, Hadas, Ronit Halachmi, Sigal Savaldi-Goldstein, Eitan Rubin, Ron Ophir, and Robert Fluhr. "Intron Retention Is a Major Phenomenon in Alternative Splicing in Arabidopsis." *The Plant Journal* 39, no. 6 (2004): 877–885. doi:10.1111/j.1365-313X.2004.02172.x.
- Nesvizhskii, AI. "A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics." *Journal of Proteomics* 73, no. 11 (2010): 2092–2123.
- Neuhoff V, N Arold, D Taube, W Ehrhardt. "Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using

- Coomassie Brilliant Blue G-250 and R-250." *Electrophoresis* 9, 6 (1988): 255-62.
- Newton, R., and W. M. Martin. "Physico-chemical Studies on the Nature of Drought Resistance in Crop Plants." *Canadian Journal of Research* 3, no. 4 (1930): 336–383.
- Ngara R and BK Ndimba. "Mapping and characterisation of the sorghum cell suspension culture secretome." *African Journal of Biotechnology* 10, 2 (2011): 253-266.
- Ngara R, R Ndimba, J Borch-Jensen, ON Jensen. "Identification and profiling of salinity stress-responsive proteins in Sorghum bicolor seedlings." *Journal of Proteomics* 75, no.13 (2012): 4139–4150.
- Nguyen TTT, N Klueva, V Chamareck, A Aarti, G Magpantay, ACM Millena, MS Pathan. "Saturation Mapping of QTL Regions and Identification of Putative Candidate Genes for Drought Tolerance in Rice." *Molecular Genetics and Genomics* 272, no. 1 (2004): 35–46.
- Nguyen, H.T. "Molecular Dissection of Drought Resistance in Crop Plants: from Traits to Genes, Plant Molecular Genetics Laboratory, Department of Plant and Soil Science." Texas Tech University, Lubbock, Texas 79409, USA. (2000).
- Niang AS. "Effect of Abiotic Stresses and Cyanide Treatment on the Cyanide Assimilatory Pathway in Arabidopsis Thaliana." ProQuest, 2008.
- Nicholls DG, and S Ferguson. *Bioenergetics*. Academic Press, (2013).  
<http://books.google.co.za/books?hl=en&lr=&id=b3fTWHBTHAAC&oi=fnd&pg=PP1&dq=+mitochondrial+membrane+is+impermeable+to+ADP+and+Pi&ots=vrqV4hrm59&sig=JCSRHQtuyn1-nAjfw0prxDpeJuc>.
- Niimura Y. "Identification of Chemosensory Receptor Genes from Vertebrate Genomes." *Methods in Molecular Biology* 1068, (2013): 95-105.
- Nilsson, T., M. Mann, R. Aebersold, J. R. Yates, A. Bairoch, and J. J. M. Bergeron. "Mass Spectrometry in High-throughput Proteomics: Ready for the Big Time." *Nature Methods* 7, no. 9 (2010): 681.
- Nishimura, Noriyuki, Ali Sarkeshik, Kazumasa Nito, Sang-Youl Park, Angela Wang, Paulo C. Carvalho, Stephen Lee, Daniel F. Caddell, Sean R. Cutler, and Joanne Chory. "PYR/PYL/RCAR Family Members Are Major in-Vivo ABI1 Protein Phosphatase 2C-Interacting Proteins in Arabidopsis." *The Plant Journal* 61, no. 2 (2010): 290–299.
- Niu, Xiping, and Nicholas J. Bate. *Maize Stress-Responsive NAC Transcription Factors and Promoter and Methods of Use*. Google Patents, 2010.

- O'Farrell PH. "High resolution two-dimensional electrophoresis of proteins". *J. Biol. Chem.* **250** (1975): 4007-4021.
- Ohme-Takagi, M., and H. Shinshi. "Ethylene-inducible DNA Binding Proteins That Interact with an Ethylene-responsive Element." *The Plant Cell Online* 7, no. 2 (1995): 173–182.
- Oliveros, J. C. "VENNY. An Interactive Tool for Comparing Lists with Venn Diagrams." *BioinfoGP, CNB-CSIC* (2007).
- Olson, M., L. Hood, C. Cantor, and D. Botstein. "A Common Language for Physical Mapping of the Human Genome." *Science* 245, no. 4925 (1989): 1434–1435.
- Ong, Shao-En, and Matthias Mann. "Mass Spectrometry–based Proteomics Turns Quantitative." *Nature Chemical Biology* 1, no. 5 (2005): 252–262.
- Osellame, Laura D., Thomas S. Blacker, and Michael R. Duchen. "Cellular and Molecular Mechanisms of Mitochondrial Function." *Best Practice & Research Clinical Endocrinology & Metabolism* 26, no. 6 (2012): 711–23.
- Ott, Martin, Vladimir Gogvadze, Sten Orrenius, and Boris Zhivotovsky. "Mitochondria, Oxidative Stress and Cell Death." *Apoptosis* 12, no. 5 (2007): 913–22.
- Ouyang S , W Zhu, J Hamilton, H Lin, M Campbell, K Childs, F Thibaud-Nissen, R L. Malek, Y Lee, L Zheng, J Orvis, B Haas, J Wortman and C. Robin Buell. "The TIGR Rice Genome Annotation Resource: improvements and new features". *Nucl. Acids Res.* **35**, (2007): D883-D887
- Ouzounis, Christos, Georg Casari, Chris Sander, Javier Tamames, and Alfonso Valencia. "Computational Comparisons of Model Genomes." *Trends in Biotechnology* 14, no. 8 (1996): 280–85.
- Pagariya M C, M. Harikrishnan, PA Kulkarni, RM Devarumath, PG Kawar. "Physio-biochemical analysis and transcript profiling of *Saccharum officinarum* L. submitted to salt stress." *Acta Physiologiae Plantarum* 33, no. 4 (2011): 1411-1424
- Pagariya, Madhuri Chandrakant, Rachayya Mallikarjun Devarumath, and Prashant Govindrao Kawar. "Biochemical Characterization and Identification of Differentially Expressed Candidate Genes in Salt Stressed Sugarcane." *Plant Science* 184 (2012): 1–13.
- Pagliari, L., KAREN Kerr, and D. LANSING Taylor. "Enolase Exists in the Fluid Phase of Cytoplasm in 3T3 Cells." *Journal of Cell Science* 94, no. 2 (1989): 333–42.
- Panagiotou, Orestis A., Cristen J. Willer, Joel N. Hirschhorn, and John PA Ioannidis. "The Power of

- Meta-Analysis in Genome-Wide Association Studies.” *Annual Review of Genomics and Human Genetics* 14 (2013): 441–465.
- Pancholi, V. “Multifunctional A-Enolase: Its Role in Diseases.” *Cellular and Molecular Life Sciences CMLS* 58, no. 7 (2001): 902–20. doi:10.1007/PL00000910.
- Pandey, A., Mann, M. “Proteomics to study genes and genomes.” *Nature* 405, (2000): 837–846.
- Pang Q., S. Chen, S. Dai, Y. Chen, Y. Wang and X. Yan. “Comparative proteomics of salt tolerance in *Arabidopsis thaliana* and *Thellungiella halophila*”. *Journal of proteome research* 9 (2010): 2584-2599.
- Pappin DJ, Hojrup P, Bleasby AJ. “Rapid identification of proteins by peptide-mass fingerprinting”. *Curr Biol.* 3, no. 6 (1993): 327-32.
- Pareek, A. *Abiotic Stress Adaptation in Plants*. Springer, 2010.
- Park, S. Y., J. W. Yu, J. S. Park, J. Li, S. C. Yoo, N. Y. Lee, S. K. Lee, S. W. Jeong, H. S. Seo, and H. J. Koh. “The Senescence-induced Staygreen Protein Regulates Chlorophyll Degradation.” *The Plant Cell Online* 19, no. 5 (2007): 1649–1664.
- Park, Sang-Youl, Pauline Fung, Noriyuki Nishimura, Davin R. Jensen, Hiroaki Fujii, Yang Zhao, Shelley Lumba, Julia Santiago, Americo Rodrigues, and F. Chow Tsz-fung. “Abscisic Acid Inhibits Type 2C Protein Phosphatases via the PYR/PYL Family of START Proteins.” *Science* 324, no. 5930 (2009): 1068–1071.
- Park, Zee-Yong, and David H.R. “Thermal Denaturation: a Useful Technique in Peptide Mass Mapping.” *Analytical Chemistry* 72, no. 11 (2000): 2667–2670.
- Parkinson, J., and M. Blaxter. “Expressed Sequence Tags: An Overview.” *Methods Mol Biol* 533 (2009): 1–12.
- Parvathaneni, Rajiv K., Vinod Jakkula, Francis K. Padi, Sebastien Faure, Nethra Nagarajappa, Ana C. Pontaroli, Xiaomei Wu, Jeffrey L. Bennetzen, and Katrien M. Devos. “Fine-Mapping and Identification of a Candidate Gene Underlying the d2 Dwarfing Phenotype in Pearl Millet, *Cenchrus Americanus* (L.) Morrone.” *G3: Genes| Genomes| Genetics* 3, no. 3 (2013): 563–72.
- Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M. “Semantic similarity in biomedical ontologies.” *PLoS Comput. Biol.* 5, (2009): e1000443.
- Paterson AH, JE Bowers, and BA Chapman. “Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics.” *PNAS* 101, 26 (2004): 9903–9908.

- Paterson AH. "Genomics of sorghum." *International journal of plant genomics* 2008 (2008): Article ID 362451, 6 pages <http://dx.doi.org/10.1155/2008/3624512008>.
- Paterson AH., JE. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, and A. Poliakov. "The Sorghum Bicolor Genome and the Diversification of Grasses." *Nature* 457, no. 7229 (2009): 551–556.
- Paterson, Andrew H., Michael Freeling, Haibao Tang, and Xiyin Wang. "Insights from the Comparison of Plant Genome Sequences." *Annual Review of Plant Biology* 61 (2010): 349–372.
- Pathan MS, B Courtois , HT Nguyen , and PK Subudhi. "Molecular Dissection of Abiotic Stress Tolerance in Sorghum and Rice." In: *Physiology and Biotechnology Integration for Plant Breeding* (2004), Nguyen HT and Abraham Blum (eds); Print ISBN: 978-0-8247-4802-9, eBook ISBN: 978-0-203-02203-0, DOI: 10.1201/9780203022030.ch14.
- Pattin, Kristine A., and Jason H. Moore. "Exploiting the Proteome to Improve the Genome-Wide Genetic Analysis of Epistasis in Common Human Diseases." *Human Genetics* 124, no. 1 (2008): 19–29.
- . "Role for Protein-Protein Interaction Databases in Human Genetics" (2009). <http://informahealthcare.com/doi/abs/10.1586/epr.09.86>.
- Paux, Etienne, Pierre Sourdille, Ian Mackay, and Catherine Feuillet. "Sequence-Based Marker Development in Wheat: Advances and Applications to Breeding." *Biotechnology Advances* 30, no. 5 (2012): 1071–1088.
- Payne, Samuel H., Shih-Ting Huang, and Rembert Pieper. "A Proteogenomic Update to Yersinia: Enhancing Genome Annotation." *BMC Genomics* 11, no. 1 (2010): 460.
- Peng, Z., M. Wang, F. Li, H. Lv, C. Li, and G. Xia. "A Proteomic Study of the Response to Salinity and Drought Stress in an Introgression Strain of Bread Wheat." *Molecular & Cellular Proteomics* 8, no. 12 (2009): 2676–2686.
- Peretó, Juli G., Ana María Velasco, Arturo Becerra, and Antonio Lazcano. "Comparative Biochemistry of CO<sub>2</sub> Fixation and the Evolution of Autotrophy." *International Microbiology* 2, no. 1 (2010): 3–10.
- Pérez-Torres, E., Paredes, M., Polanco, V., Becerra, V., others. "Gene expression analysis: a way to study tolerance to abiotic stresses in crops species." *Chil. J Agric Res* 69, (2009): 260–269.
- Perkins DN, Pappin DJ, Creasy DM, Cottrell JS . "Probability-based protein identification by

- searching sequence databases using mass spectrometry data." *Electrophoresis* 20, 18 (1999): 3551–67.
- Pevsner, J. *Bioinformatics and Functional Genomics*. John Wiley & Sons, 2009.
- Pierlé, Sebastián A., Michael J. Dark, Dani Dahmen, Guy H. Palmer, and Kelly A. Brayton. "Comparative Genomics and Transcriptomics of Trait-gene Association." *BMC Genomics* 13, no. 1 (2012): 669.
- Pihlasalo, S., Kulmala, A., Rozwandowicz-Jansen, A., Hänninen, P., Härmä, H. "Sensitive Luminometric Method for Protein Quantification in Bacterial Cell Lysate Based on Particle Adsorption and Dissociation of Chelated Europium." *Anal. Chem.* 84, (2012): 1386–1393.
- Pillitteri, Lynn Jo, Kylee M. Peterson, Robin J. Horst, and Keiko U. Torii. "Molecular Profiling of Stomatal Meristemoids Reveals New Component of Asymmetric Cell Division and Commonalities among Stem Cell Populations in Arabidopsis." *The Plant Cell Online* 23, no. 9 (2011): 3260–3275.
- Plant Ontology Consortium. "The Plant Ontology™ Consortium and Plant Ontologies." *International Journal of Genomics* 3, no. 2 (2002): 137–142.
- Pontius, Joan U., Lukas Wagner, and Gregory D. Schuler. "21. UniGene: A Unified View of the Transcriptome." *The NCBI Handbook*. Bethesda, MD: National Library of Medicine (US), NCBI, 2003. <http://sites.google.com/site/mitchela/ch21.pdf>.
- Posch A, T Franz, S Hartwig, B Knebel, H Al-Hasani, W Passlack, N Kunz, Y Hinze, X Li, J Kotzka, and S Lehr. "2D-ToGo Workflow: Increasing Feasibility and Reproducibility of 2-Dimensional Gel Electrophoresis," *Archives of Physiology and Biochemistry* 119, no. 3 (2013): 108–113.
- Pratt, Lee H., Chun Liang, Manish Shah, Feng Sun, Haiming Wang, Alan R. Gingle, Andrew H. Paterson, Rod Wing, Ralph Dean, and Robert Klein. "Sorghum Expressed Sequence Tags Identify Signature Genes for Drought, Pathogenesis, and Skotomorphogenesis from a Milestone Set of 16,801 Unique Transcripts." *Plant Physiology* 139, no. 2 (2005): 869–884.
- Price, Adam H., Jill E. Cairns, Peter Horton, Hamlyn G. Jones, and Howard Griffiths. "Linking Drought-resistance Mechanisms to Drought Avoidance in Upland Rice Using a QTL Approach: Progress and New Opportunities to Integrate Stomatal and Mesophyll Responses." *Journal of Experimental Botany* 53, no. 371 (2002): 989–1004.
- Proost, Sebastian, Michiel Van Bel, Lieven Sterck, Kenny Billiau, Thomas Van Parys, Yves Van de



- Peer, and Klaas Vandepoele. "PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants." *The Plant Cell Online* 21, no. 12 (2009): 3718–3731.
- Pujar, Anuradha, Pankaj Jaiswal, Elizabeth A. Kellogg, Katica Ilic, Leszek Vincent, Shulamit Avraham, Peter Stevens, Felipe Zapata, Leonore Reiser, and Seung Y. Rhee. "Whole-plant Growth Stage Ontology for Angiosperms and Its Application in Plant Biology." *Plant Physiology* 142, no. 2 (2006): 414–428.
- Putnam, Nicholas H., Mansi Srivastava, Uffe Hellsten, Bill Dirks, Jarrod Chapman, Asaf Salamov, Astrid Terry, Harris Shapiro, Erika Lindquist, and Vladimir V. Kapitonov. "Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization." *Science* 317, no. 5834 (2007): 86–93.
- Qin, F., M. Kakimoto, Y. Sakuma, K. Maruyama, Y. Osakabe, L. S. P. Tran, K. Shinozaki, and K. Yamaguchi-Shinozaki. "Regulation and Functional Analysis of ZmDREB2A in Response to Drought and Heat Stresses in Zea Mays L." *The Plant Journal* 50, no. 1 (2007): 54–69.
- Rabello, A.R., Guimarães, C.M., Rangel, P.H., da Silva, F.R., Seixas, D., de Souza, E., Brasileiro, A.C., Spehar, C.R., Ferreira, M.E., Mehta, Â. "Identification of drought-responsive genes in roots of upland rice (*Oryza sativa* L)." *BMC Genomics* 9, (2008): 485.
- Rabilloud T. "Solubilization of proteins for electrophoretic analyses." *Electrophoresis*, 17 (5), (1996): 813–829
- Rafalski, J.A. "Association Genetics in Crop Improvement." *Current Opinion in Plant Biology* 13, no. 2 (2010): 174–180.
- Ramalingam, J., C. M. Vera Cruz, K. Kukreja, J. M. Chittoor, J.-L. Wu, S. W. Lee, M. Baraoidan, M. L. George, M. B. Cohen, and S. H. Hulbert. "Candidate Defense Genes from Rice, Barley, and Maize and Their Association with Qualitative and Quantitative Resistance in Rice." *Molecular Plant-Microbe Interactions* 16, no. 1 (2003): 14–24.
- Rampino, Patrizia, Stefano Pataleo, Carmela Gerardi, Giovanni Mita, and Carla Perrotta. "Drought Stress Response in Wheat: Physiological and Molecular Analysis of Resistant and Sensitive Genotypes." *Plant, Cell & Environment* 29, no. 12 (2006): 2143–2152.
- Rao AVRK, SP Wani, and P Singh. "Use of Agroclimatic Datasets for Improved Planning of Watersheds." International Crops Research Institute for the Semi-Arid Tropics, Patancheru, India. (2011):145-155.
- Rao, A.S., House, L.R., and Gupta, S.C. Review of Sorghum, Pearl Millet, and Finger Millet

- Improvement in SADCC Countries. ICRISAT, Patancheru, India (1989).
- Rapp, Barbara A., and David L. Wheeler. "Bioinformatics Resources from the National Center for Biotechnology Information: An Integrated Foundation for Discovery." *Journal of the American Society for Information Science and Technology* 56, no. 5 (2005): 538–550.
- Ravi, K., V. Vadez, S. Isobe, R. R. Mir, Y. Guo, S. N. Nigam, M. V. C. Gowda, T. Radhakrishnan, D. J. Bertioli, and S. J. Knapp. "Identification of Several Small Main-Effect QTLs and a Large Number of Epistatic QTLs for Drought Tolerance Related Traits in Groundnut (*Arachis Hypogaea* L.)." *Theoretical and Applied Genetics* 122, no. 6 (2011): 1119–1132.
- Reddy, Attipalli Ramachandra, Kolluru Viswanatha Chaitanya, and Munusamy Vivekanandan. "Drought-induced Responses of Photosynthesis and Antioxidant Metabolism in Higher Plants." *Journal of Plant Physiology* 161, no. 11 (2004): 1189–1202.
- Reddy, Nagaraja Reddy Rama, Madhusudhana Ragimasalawada, Murali Mohan Sabbavarapu, Seetharama Nadoor, and Jagannatha Vishnu Patil. "Detection and Validation of Stay-Green QTL in Post-Rainy Sorghum Involving Widely Adapted Cultivar, M35-1 and a Popular Stay-Green Genotype B35." *BMC Genomics* 15, no. 1 (2014): 909.
- Reguera, G., Leschine, S. "Fast and efficient elution of proteins from polyacrylamide gels using nanosep® centrifugal devices." Dep. Microbiol. Univ. Mass.-Amherst Amherst. 2009.
- Rhead, Brooke, Donna Karolchik, Robert M. Kuhn, Angie S. Hinrichs, Ann S. Zweig, Pauline A. Fujita, Mark Diekhans, Kayla E. Smith, Kate R. Rosenbloom, and Brian J. Raney. "The UCSC Genome Browser Database: Update 2010." *Nucleic Acids Research* 38, no. suppl 1 (2010): D613–D619.
- Rhee, Seung Yon, Valerie Wood, Kara Dolinski, and Sorin Draghici. "Use and Misuse of the Gene Ontology Annotations." *Nature Reviews Genetics* 9, no. 7 (2008): 509–515.
- Ribaut JM, DA Hoisington, JA Deutsch, C Jiang, D Gonzalez-de-Leon. "Identification of quantitative trait loci under drought conditions in tropical maize. 1. Flowering parameters and the anthesis-silking interval." *Theoretical and Applied Genetics* 92, no.7 (1996): 905-914.
- Ribaut, J. M., M. Banziger, J. Betran, C. Jiang, G. O. Edmeades, K. Dreher, and D. Hoisington. "Use of Molecular Markers in Plant Breeding: Drought Tolerance Improvement in Tropical Maize." *Quantitative Genetics, Genomics, and Plant Breeding*, (2002): 85–99.
- Ribot, Gerlitt González, Paola Silva, and Edmundo Acevedo. "Morphological and Physiological Traits of Assistance in the Selection of High Yielding Varieties of Durum Wheat (*Triticum*

- Turgidum L. Spp. Durum) for the Rainfed Mediterranean Environments of Central Chile.” *American Journal of Plant Sciences* 3 (2012): 1809.
- Riechmann, J. L., and O. J. Ratcliffe. “A Genomic Perspective on Plant Transcription Factors.” *Current Opinion in Plant Biology* 3, no. 5 (2000): 423–434.
- Righetti PG and A Bossi. “Isoelectric Focusing of Proteins and Peptides in Gel Slabs and in Capillaries.” *Analytica Chimica Acta* 372, no. 1 (1998): 1–19.
- Righetti PG *et al.*, “Conventional Isoelectric Focusing and Immobilized pH Gradients for Hemoglobin Separation and Identification.” *The Hemoglobinopathies* 15 (1986): 47–59.
- Ripley BS, ME Gilbert, DG Ibrahim, CP Osborne. “Drought Constraints on C4 Photosynthesis: Stomatal and Metabolic Limitations in C3 and C4 Subspecies of *Alloteropsis Semialata*.” *Journal of Experimental Botany* 58, no. 6 (2007): 1351–63.
- Risch, Neil, and Kathleen Merikangas. “The Future of Genetic Studies of Complex Human Diseases.” *Science-AAAS-Weekly Paper Edition* 273, no. 5281 (1996): 1516–1517.
- Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., Koonin, E.V. “Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.” *Curr. Biol.* 13, (2003): 1512–1517.
- Ronning, C. M., S. S. Stegalkina, R. A. Ascenzi, O. Bougri, A. L. Hart, T. R. Utterbach, S. E. Vanaken, S. B. Riedmuller, J. A. White, and J. Cho. “Comparative Analyses of Potato Expressed Sequence Tag Libraries.” *Plant Physiology* 131, no. 2 (2003): 419–429.
- Rosenow, D.T., Ejeta, G., Clark, E.L., Gilbert, M.L., Henzell, R.G., Borrell, A.K., and Muchow, R.C., 1996. Breeding for pre-and post-flowering Drought resistance in Sorghum. In IntsorMil and ICRISAT (eds) proceedings of the international conference on genetic Improvement of sorghum and pearl millet, Sep. 23-27, 1996, Lubbock, Texas.
- Rottloff, Sandy, Regina Stieber, Heiko Maischak, Florian G. Turini, Günther Heubl, and Axel Mithöfer. “Functional Characterization of a Class III Acid Endochitinase from the Traps of the Carnivorous Pitcher Plant Genus, *Nepenthes*.” *Journal of Experimental Botany*, 2011, err173.
- Roy, Amit, Agneyo Ganguly, Somdeb BoseDasgupta, Benu Brata Das, Churala Pal, Parasuraman Jaisankar, and Hemanta K. Majumder. “Mitochondria-Dependent Reactive Oxygen Species-Mediated Programmed Cell Death Induced by 3, 3'-Diindolylmethane through Inhibition of F0F1-ATP Synthase in Unicellular Protozoan Parasite *Leishmania Donovanii*.” *Molecular*

*Pharmacology* 74, no. 5 (2008): 1292–1307.

- Roy, S.W., Gilbert, W. “The evolution of spliceosomal introns: patterns, puzzles and progress.” *Nat. Rev. Genet.* 7, (2006): 211–221.
- Rubin, Gerald M., Mark D. Yandell, Jennifer R. Wortman, George L. Gabor, Catherine R. Nelson, Iswar K. Hariharan, Mark E. Fortini, Peter W. Li, Rolf Apweiler, and Wolfgang Fleischmann. “Comparative Genomics of the Eukaryotes.” *Science* 287, no. 5461 (2000): 2204–15.
- Rudd, Stephen, Hans-Werner Mewes, and Klaus FX Mayer. “Sputnik: A Database Platform for Comparative Plant Genomics.” *Nucleic Acids Research* 31, no. 1 (2003): 128–32. <http://nar.oxfordjournals.org/content/31/1/128.short>.
- Saeed, A. I., Vasily Sharov, Joe White, Jerry Li, Wei Liang, Nirmal Bhagabati, J. Braisted, M. Klapa, T. Currier, and M. Thiagarajan. “TM4: A Free, Open-Source System for Microarray Data Management and Analysis.” *Biotechniques* 34, no. 2 (2003): 374. <http://www.ncbi.nlm.nih.gov/pubmed/12613259>.
- Sáez DE and JC Slebe. “Subcellular localization of aldolase B.” *Journal of Cellular Biochemistry* Volume 78, Issue 1 (2000): 62–72. DOI: 10.1002/(SICI)1097-4644(20000701)78:1<62::AID-JCB6>3.0.CO;2-W.
- Sage RF. “The evolution of C4 photosynthesis.” *New Phytologist* 161, no. 2 (2004): 341–370.
- Sahiner N and M Singh. “In Situ Micro/nano-Hydrogel Synthesis from Acrylamide Derivates with Lecithin Organogel System,” *Polymer* 48, no. 10 (2007): 2827–2834.
- Sakata, Katsumi, Yoshiaki Nagamura, Hisataka Numa, Baltazar A. Antonio, Hideki Nagasaki, Atsuko Itonuma, Wakako Watanabe, et al. “RiceGAAS: An Automated Annotation System and Database for Rice Genome Sequence.” *Nucleic Acids Research* 30, no. 1 (2002): 98–102.
- Sakharkar, Meena, Pandjassarame Kanguene, and Venkatarajan S. Mathura. “Biological Sequence Databases.” In *Bioinformatics: A Concept-Based Introduction*, 39–46. Springer, 2009.
- Sakuma, Y., K. Maruyama, Y. Osakabe, F. Qin, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki. “Functional Analysis of an Arabidopsis Transcription Factor, DREB2A, Involved in Drought-responsive Gene Expression.” *The Plant Cell Online* 18, no. 5 (2006): 1292–1309.
- Sakuma, Y., Q. Liu, J. G. Dubouzet, H. Abe, K. Shinozaki, and K. Yamaguchi-Shinozaki. “DNA-Binding Specificity of the ERF/AP2 Domain of Arabidopsis DREBs, Transcription Factors Involved in Dehydration-and Cold-Inducible Gene Expression.” *Biochemical and Biophysical Research Communications* 290, no. 3 (2002): 998–1009.

- Salamov, Asaf A., and Victor V. Solovyev. "Ab Initio Gene Finding in Drosophila Genomic DNA." *Genome Research* 10, no. 4 (2000): 516–22.
- Salekdeh, Gh H., J. Siopongco, L. J. Wade, B. Ghareyazie, and J. Bennett. "A Proteomic Approach to Analyzing Drought-and Salt-responsiveness in Rice." *Field Crops Research* 76, no. 2 (2002): 199–219.
- Salekdeh, Ghasem Hosseini, Matthew Reynolds, John Bennett, and John Boyer. "Conceptual Framework for Drought Phenotyping During Molecular Breeding." *Trends in Plant Science* 14, no. 9 (2009): 488–496.
- Sanchez AC., P. K. Subudhi, D. T. Rosenow, and H. T. Nguyen. "Mapping QTLs Associated with Drought Resistance in Sorghum (*Sorghum Bicolor* L. Moench)." *Plant Molecular Biology* 48, no. 5–6 (2002): 713–26.
- Sánchez R, Ursula Pieper, Francisco Melo, Narayanan Eswar, Marc A. Martí-Renom, M. S. Madhusudhan, Nebojša Mirković, and Andrej Šali. "Protein Structure Modeling for Structural Genomics." *Nature Structural & Molecular Biology* 7 (2000): 986–90.
- Sanchez, Diego H., Franziska Schwabe, Alexander Erban, Michael K. Udvardi, and Joachim Kopka. "Comparative Metabolomics of Drought Acclimation in Model and Forage Legumes." *Plant, Cell & Environment* 35, no. 1 (2012): 136–149.
- Sasaki, T. and Antonio, B. A. 2009. Plant genomics: *Sorghum in sequence*. *Nature:News and Views* Vol 457: 547-548
- Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, and S. Federhen. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 39, no. suppl 1 (2011): D38–D51.
- Saxena, A., Tripathi, B.P., Kumar, M., Shahi, V.K. "Membrane-based techniques for the separation and purification of proteins: an overview." *Adv. Colloid Interface Sci.* 145, (2009): 1–22.
- Schafleitner, R., Gutierrez, R., Espino, R., Gaudin, A., Pérez, J., Martínez, M., Domínguez, A., Tincopa, L., Alvarado, C., Numberto, G., others. "Field screening for variation of drought tolerance in *Solanum tuberosum* L. by agronomical, physiological and genetic analysis." *Potato Res.* 50, (2007): 71–85.
- Schilling, Christophe H., Stefan Schuster, Bernhard O. Palsson, and Reinhart Heinrich. "Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-Genomic Era." *Biotechnology Progress* 15, no. 3 (1999): 296–303.

- Schlueter, J. A., P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker. "Mining EST Databases to Resolve Evolutionary Events in Major Crop Species." *Genome* 47, no. 5 (2004): 868–876.
- Schmutz, Jeremy, Steven B. Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L. Hyten, Qijian Song, Jay J. Thelen, and Jianlin Cheng. "Genome Sequence of the Palaeopolyploid Soybean." *Nature* 463, no. 7278 (2010): 178–183.
- Schnable, James C., and Eric Lyons. "Comparative Genomics with Maize and Other Grasses: From Genes to Genomes!" *Maydica* 56, no. 2 (2012). <http://cra-journals.cineca.it/index.php/maydica/article/view/690>.
- Scholz, Fabian G., Sandra J. Bucci, Nadia Arias, Frederick C. Meinzer, and Guillermo Goldstein. "Osmotic and Elastic Adjustments in Cold Desert Shrubs Differing in Rooting Depth: Coping with Drought and Subzero Temperatures." *Oecologia* 170, no. 4 (2012): 885–897.
- Schulze, Waltraud X., and Björn Usadel. "Quantitation in Mass-spectrometry-based Proteomics." *Annual Review of Plant Biology* 61 (2010): 491–516.
- Schwab, K. B., U. Schreiber, and U. Heber. "Response of Photosynthesis and Respiration of Resurrection Plants to Desiccation and Rehydration." *Planta* 177, no. 2 (1989): 217–27.
- Schwarzenberger, A., C. Courts, and E. Von Elert. "Target Gene Approaches: Gene Expression in *Daphnia Magna* Exposed to Predator-borne Kairomones or to Microcystin-producing and Microcystin-free *Microcystis Aeruginosa*." *BMC Genomics* 10, no. 1 (2009): 527.
- Sechi, Salvatore, and Brian T. Chait. "Modification of Cysteine Residues by Alkylation. A Tool in Peptide Mapping and Protein Identification." *Analytical Chemistry* 70, no. 24 (1998): 5150–5158.
- Sehgal, Deepmala, Vengaldas Rajaram, Ian P. Armstead, Vincent Vadez, Yash P. Yadav, Charles T. Hash, and Rattan S. Yadav. "Integration of Gene-Based Markers in a Pearl Millet Genetic Map for Identification of Candidate Genes Underlying Drought Tolerance Quantitative Trait Loci." *BMC Plant Biology* 12, no. 1 (2012): 9.
- Seki, M., A. Kamei, K. Yamaguchi-Shinozaki, and K. Shinozaki. "Molecular Responses to Drought, Salinity and Frost: Common and Different Paths for Plant Protection." *Current Opinion in Biotechnology* 14, no. 2 (2003): 194–199.
- Seki, Motoaki, Mari Narusaka, Junko Ishida, Tokihiko Nanjo, Miki Fujita, Youko Oono, Asako

- Kamiya, *et al.* "Monitoring the Expression Profiles of 7000 Arabidopsis Genes under Drought, Cold and High-Salinity Stresses Using a Full-Length cDNA Microarray." *The Plant Journal* 31, no. 3 (2002): 279–92.
- Shakoor Nadia, R. Nair, O. Crasta, G. Morris, A. Feltus and S. Kresovich. "A Sorghum bicolor expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums." *BMC Plant Biology* 14 (2014): 35
- Sharma R, D. De Vleeschauwer, MK. Sharma, and PC. Ronald. "Recent Advances in Dissecting Stress-Regulatory Crosstalk in Rice". *Molecular Plant* 6 (2013): 250-260
- Shevchenko A., M.Wilm, O. Vorm, and M. Mann. "Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels". European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. *Analytical Chemistry* 68 (1996a): 850-858.
- Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M. "Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels." *Proc. Natl. Acad. Sci. U.S.A.* 93, 25 (1996b): 14440–5.
- Shiau, J. T., S. Feng, and S. Nadarajah. "Assessment of Hydrological Droughts for the Yellow River, China, Using Copulas." *Hydrological Processes* 21, no. 16 (2007): 2157–2163.
- Shinozaki K and K Yamaguchi-Shinozaki. "Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression". *The Plant Cell* 18, no.5 (2006): 1292–1309.
- Shinozaki, K., K. Yamaguchi-Shinozaki, and M. Seki. "Regulatory Network of Gene Expression in the Drought and Cold Stress Responses." *Current Opinion in Plant Biology* 6, no. 5 (2003): 410–417.
- Shinozaki, Kazuo, and Kazuko Yamaguchi-Shinozaki. "Gene Networks Involved in Drought Stress Response and Tolerance." *Journal of Experimental Botany* 58, no. 2 (2007): 221–227.
- Shoemaker DD, PS Linsley. "Recent developments in DNA microarrays." *Current opinion in microbiology* 5, no. 3(2002): 334–337.
- Shrestha, Rosemary, Elizabeth Arnaud, Ramil Mauleon, Martin Senger, Guy F. Davenport, David Hancock, Norman Morrison, Richard Bruskiwich, and Graham McLaren. "Multifunctional Crop Trait Ontology for Breeders' Data: Field Book, Annotation, Data Discovery and Semantic Enrichment of the Literature." *AoB Plants* 2010 (2010).

- Singh, Rohit, Jinbo Xu, and Bonnie Berger. "Global Alignment of Multiple Protein Interaction Networks with Application to Functional Orthology Detection." *Proceedings of the National Academy of Sciences* 105, no. 35 (2008): 12763–12768.
- Slater, Guy SC, and Ewan Birney. "Automated Generation of Heuristics for Biological Sequence Comparison." *BMC Bioinformatics* 6, no. 1 (2005): 31. doi:10.1186/1471-2105-6-31.
- Slettebak, R. T. "Don't Blame the Weather! Climate-related Natural Disasters and Civil Conflict." *Journal of Peace Research* 49, no. 1 (2012): 163–176.
- Smedley, Damian, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. "BioMart - Biological Queries Made Easy." *BMC Genomics* 10 (2009): 22. doi:10.1186/1471-2164-10-22.
- Smejkal GB. "The Coomassie Chronicles: Past, Present and Future Perspectives in Polyacrylamide Gel Staining," 2004, <http://informahealthcare.com/doi/pdf/10.1586/14789450.1.4.381>.
- Smit, A. F. A., R. Hubley, and P. Green. 1996–2004. *RepeatMasker Open-3.0*, 2006.
- Smit, A. F. A., R. Hubley, and P. Green. *RepeatMasker Open-3.0 (Institute for Systems Biology, Seattle, WA)*, 2012.
- Smith, Barry, and David M. Mark. "Do Mountains Exist? Towards an Ontology of Landforms." *Environment and Planning B* 30, no. 3 (2003): 411–428.
- Smith, Barry, Jennifer Williams, and Schulze-Kremer Steffen. "The Ontology of the Gene Ontology." In *AMIA Annual Symposium Proceedings*, 2003:609. American Medical Informatics Association, 2003.
- Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, and Christopher J. Mungall. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration." *Nature Biotechnology* 25, no. 11 (2007): 1251–1255.
- Smith, Wendell A., Brandon T. Schurter, Flossie Wong-Staal, and Michael David. "Arginine Methylation of RNA Helicase a Determines Its Subcellular Localization." *Journal of Biological Chemistry* 279, no. 22 (2004): 22795–98.
- Snel, Berend, Peer Bork, and Martijn A. Huynen. "The Identification of Functional Modules from the Genomic Association of Genes." *Proceedings of the National Academy of Sciences* 99, no. 9 (2002): 5890–5895.
- Sondej MA, P Doran, JA Loo, I-B Wanner. "Sample Preparation of Primary Astrocyte Cellular and



- Released Proteins for 2-D Gel Electrophoresis and Protein Identification by Mass Spectrometry." *Sample Preparation in Biological Mass Spectrometry* (2011): 829–849, [http://link.springer.com/chapter/10.1007/978-94-007-0828-0\\_39](http://link.springer.com/chapter/10.1007/978-94-007-0828-0_39).
- Song, Fuqi, Gregory Zacharewicz, and David Chen. "An Ontology-Driven Framework towards Building Enterprise Semantic Information Layer." *Advanced Engineering Informatics* 27, no. 1 (2013): 38–50.
- Song, Rentao, Victor Llaca, and Joachim Messing. "Mosaic Organization of Orthologous Sequences in Grass Genomes." *Genome Research* 12, no. 10 (2002): 1549–55. doi:10.1101/gr.268302.
- Sorghum Genomics Planning Workshop Participants (SGPW), 2005. Toward Sequencing the Sorghum Genome: A U.S. National Science Foundation-Sponsored Workshop Report, American Society of Plant Biologist. *Plant Physiology* 138: 1898-1902
- Spreitzer, Robert J., and Michael E. Salvucci. "Rubisco: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme." *Annual Review of Plant Biology* 53, no. 1 (2002): 449–75.
- Sreenivasulu, N., S. K. Sopory, and P. B. Kavi Kishor. "Deciphering the Regulatory Mechanisms of Abiotic Stress Tolerance in Plants by Genomic Approaches." *Gene* 388, no. 1 (2007): 1–13.
- Standage DS and VP Brendel. "ParsEval: Parallel Comparison and Analysis of Gene Structure Annotations." *BMC Bioinformatics* 13 (August 1, 2012): 187. doi:10.1186/1471-2105-13-187.
- Stanke M, O Keller, I Gunduz, A Hayes, S Waack and B Morgenstern. "AUGUSTUS: *ab initio* prediction of alternative transcripts." *Nucl. Acids Res.* 34, suppl 2 (2006a) : W435-W439. doi: 10.1093/nar/gkl200
- Stanke M, A Tzvetkova and B Morgenstern."AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome." *Genome Biology* 7, Suppl 1 (2006b): S11
- Stanke M, O. Schöffmann, B. Morgenstern and S. Waack. "Gene prediction in eukaryotes with a generalized hidden markov model that uses hints from external sources." *BMC Bioinformatics* 7, (2006c):62.
- Stanton, J. A. L., A. B. Macgregor, and D. P. L. Green. "Identifying Tissue-enriched Gene Expression in Mouse Tissues Using the NIH UniGene Database." *Applied Bioinformatics* 2 (2003): 65–74.

- Stauber, Jonathan, Luke MacAleese, Julien Franck, Emmanuelle Claude, Marten Snel, Basak Kükreer Kaletas, Ingrid MVD Wiel, Maxence Wisztorski, Isabelle Fournier, and Ron Heeren. "On-tissue Protein Identification and Imaging by MALDI-ion Mobility Mass Spectrometry." *Journal of the American Society for Mass Spectrometry* 21, no. 3 (2010): 338–347.
- Stein, Lincoln. "Genome Annotation: From Sequence to Biology." *Nature Reviews Genetics* 2, no. 7 (2001): 493–503.
- Stockhaus, J., Schlue, U., Koczor, M., Chitty, J.A., Taylor, W.C., and Westhoff, P. The promoter of the gene encoding the C4 form of phosphoenolpyruvate carboxylase directs mesophyll specific expression in transgenic C4 Flaveria spp. *Plant Cell* (1997). **9**: 479–489
- Storey JD and Tibshirani R. "Statistical significance for genomewide studies." *Proc. Natl. Acad. Sci* 100, no. 16 (2003): 9440–9445.
- Subedi, K. D., and B. L. Ma. "Nitrogen Uptake and Partitioning in Stay-green and Leafy Maize Hybrids." *Crop Science* 45, no. 2 (2005): 740–747.
- Subodhi P. "Omics Approaches for Abiotic Stress Tolerance in Plants". In: Omics and Plant abiotic Stress Tolerance (2011): 10-38.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, *et al.* "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 43 (2005): 15545–50. doi:10.1073/pnas.0506580102.
- Subudhi, P. K., D. T. Rosenow, and H. T. Nguyen. "Quantitative Trait Loci for the Stay Green Trait in Sorghum (*Sorghum Bicolor* L. Moench): Consistency across Genetic Backgrounds and Environments." *Theoretical and Applied Genetics* 101, no. 5–6 (2000): 733–41.
- Sugihara K, Hanagata N, Dubinsky Z, Sigeyuki B and Karube I. "Molecular characterization of cDNA encoding oxygen evolving enhancer protein 1 increased by salt treatment in the mangrove *Bruguiera gymnorhiza*." *Plant and Cell Physiology* 41, no.11 (2000): 1279–1285.
- Sultan, Benjamin, P. Roudier, P. Quirion, A. Alhassane, B. Muller, Michael Dingkuhn, P. Ciaï, M. Guimberteau, S. Traore, and C. Baron. "Assessing Climate Change Impacts on Sorghum and Millet Yields in the Sudanian and Sahelian Savannas of West Africa." *Environmental Research Letters* 8, no. 1 (2013): 014040.
- Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." *PLoS One* 6, no. 7 (2011): e21800.

- SurrIDGE, C. "Agricultural Biotech: The Rice Squad." *Nature* 416, no. 6881 (2002): 576–578.
- Swamy, B. P., and Arvind Kumar. "Genomics-Based Precision Breeding Approaches to Improve Drought Tolerance in Rice." *Biotechnology Advances* 31, no. 8 (2013): 1308–1318.
- Swigonova Z, J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J. L. Bennetzen, and J. Messing, "Close split of sorghum and maize genome progenitors," *Genome Res.*, vol. 14, no. 10a (2004): 1916–1923.
- Tabor, H. K., N. J. Risch, and R. M. Myers. "Candidate-gene Approaches for Studying Complex Genetic Traits: Practical Considerations." *Nature Reviews Genetics* 3, no. 5 (2002): 391–396.
- Takemura T, N Hanagata, K Sugihara, S Baba, I Karube, Z Dubinsky. "Physiological and biochemical responses to salt stress in the mangrove, *Bruguiera gymnorrhiza*." *Aquatic Botany* 68, no. 1 (2000): 15–28.
- Tallaksen, L. M., and H. A. J. Van Lanen. *Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater*. Vol. 48. Elsevier Science, 2004.
- Tang H., JE. Bowers, X. Wang, R. Ming, M. Alam, AH. Paterson. "Synteny and Collinearity in Plant Genomes." *science* 320 no. 5875 (2008):486-488 DOI: 10.1126/science.1153917
- Tanksley, Steven D., and Susan R. McCouch. "Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild." *Science* 277, no. 5329 (1997): 1063–66.
- Tannu, N.S., Hemby, S.E. "Two-dimensional fluorescence difference gel electrophoresis for comparative proteomics profiling." *Nat. Protoc.* 1, (2006): 1732–1742.
- Tao, Y. Z., R. G. Henzell, D. R. Jordan, D. G. Butler, A. M. Kelly, and C. L. McIntyre. "Identification of Genomic Regions Associated with Stay Green in Sorghum by Testing RILs in Multiple Environments." *TAG Theoretical and Applied Genetics* 100, no. 8 (2000): 1225–32.
- Tarailo-Graovac, M., and N. Chen. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics* (2009): 4.10. 1–4.10. 14.
- Tay, Sen-Kwan, Jason Blythe, and Leonard Lipovich. "Global Discovery of Primate-Specific Genes in the Human Genome." *Proceedings of the National Academy of Sciences* 106, no. 29 (2009): 12019–12024.
- Tazi, Jamal, Nadia Bakkour, and Stefan Stamm. "Alternative Splicing and Disease." *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1792, no. 1 (2009): 14–26.
- Terai, Goro, Aya Yoshizawa, Hiroaki Okida, Kiyoshi Asai, and Toutai Mituyama. "Discovery of

- Short Pseudogenes Derived from Messenger RNAs." *Nucleic Acids Research* 38, no. 4 (2010): 1163–1171.
- Teshome, Awegechew, B. R. Baum, L. Fahrig, J. K. Torrance, T. J. Arnason, and J. D. Lambert. "Sorghum [*Sorghum Bicolor* (L.) Moench] Landrace Variation and Classification in North Shewa and South Welo, Ethiopia." *Euphytica* 97, no. 3 (1997): 255–63.
- Teshome, Awegechew, Lenore Fahrig, J. Kenneth Torrance, J. D. Lambert, T. J. Arnason, and B. R. Baum. "Maintenance of Sorghum (*Sorghum Bicolor*, Poaceae) Landrace Diversity by Farmers' Selection in Ethiopia." *Economic Botany* 53, no. 1 (1999): 79–88.
- Tester, M., and P. Langridge. "Breeding Technologies to Increase Crop Production in a Changing World." *Science* 327, no. 5967 (2010): 818–822.
- Thangavelu, Madan, Allan B. James, Alan Bankier, Glenn J. Bryan, Paul H. Dear, and Robbie Waugh. "HAPPY Mapping in a Plant Genome: Reconstruction and Analysis of a High-resolution Physical Map of a 1.9 Mbp Region of *Arabidopsis Thaliana* Chromosome 4." *Plant Biotechnology Journal* 1, no. 1 (2003): 23–31.
- The International Brachypodium Initiative. "Genome Sequencing and Analysis of the Model Grass *Brachypodium Distachyon*." *Nature* 463, no. 7282 (2010): 763–68.
- Thiede, Bernd, Wolfgang Höhenwarter, Alexander Krahl, Jens Mattow, Monika Schmid, Frank Schmidt, and Peter R. Jungblut. "Peptide Mass Fingerprinting." *Methods* 35, no. 3 (2005): 237–47.
- Thiede B, Christian JK, Margarita S Achim T, Robert S, Ursula ZA, Monika S and Peter RJ. "High Resolution Quantitative Proteomics of HeLa Cells Protein Species Using Stable Isotope Labeling with Amino Acids in Cell Culture(SILAC), Two-Dimensional Gel Electrophoresis(2DE) and Nano-Liquid Chromatography Coupled to an LTQ-Orbitrap Mass Spectrometer." *Molecular & Cellular Proteomics* 12 (2013): 529-538.
- Thomas D. "Gene–environment-wide association studies: emerging approaches." *Nature Reviews Genetics* 11, (2010): 259–272.
- Thomas, H., and C. J. Howarth. "Five Ways to Stay Green." *Journal of Experimental Botany* 51, no. suppl 1 (2000): 329–337.
- Thomas, H., and C. M. Smart. "Crops That Stay Green1." *Annals of Applied Biology* 123, no. 1 (1993): 193–219.
- Thornton, M.A., Zhang, C., Kowalska, M.A., Poncz, M. "Identification of distal regulatory regions

- in the human  $\alpha$ IIb gene locus necessary for consistent, high-level megakaryocyte expression.” *Blood* 100, (2002): 3588–3596. doi:10.1182/blood-2002-05-1307
- Tiffin, Nicki, Janet F. Kelso, Alan R. Powell, Hong Pan, Vladimir B. Bajic, and Winston A. Hide. “Integration of Text-and Data-Mining Using Ontologies Successfully Selects Disease Gene Candidates.” *Nucleic Acids Research* 33, no. 5 (2005): 1544–1552.
- Timperio AM, MG Egidi, L Zolla. “Proteomics applied on plant abiotic stresses: role of heat shock proteins (HSP)”. *Journal of Proteomics* 71 (2008): 391–411
- Tine, Mbaye, Heiner Kuhl, Alfred Beck, Luca Bargelloni, and Richard Reinhardt. “Comparative Analysis of Intronless Genes in Teleost Fish Genomes: Insights into Their Evolution and Molecular Function.” *Marine Genomics* 4, no. 2 (2011): 109–19.
- Tondelli, A., E. Francia, D. Barabaschi, A. Aprile, J. S. Skinner, E. J. Stockinger, A. M. Stanca, and N. Pecchioni. “Mapping Regulatory Genes as Candidates for Cold and Drought Stress Tolerance in Barley.” *Theoretical and Applied Genetics* 112, no. 3 (2006): 445–54.
- Tonsor, S. J., C. Alonso-Blanco, and M. Koornneef. “Gene Function beyond the Single Trait: Natural Variation, Gene Effects, and Evolutionary Ecology in *Arabidopsis Thaliana*.” *Plant, Cell & Environment* 28, no. 1 (2005): 2–20.
- Traba, J., A. Del Arco, M. R. Duchon, G. Szabadkai, and J. Satrústegui. “SCaMC-1 Promotes Cancer Cell Survival by Desensitizing Mitochondrial Permeability Transition via ATP/ADP-Mediated Matrix Ca<sup>2+</sup> Buffering.” *Cell Death & Differentiation* 19, no. 4 (2011): 650–60.
- Tripathy, J. N., Jingxian Zhang, S. Robin, Thuy T. Nguyen, and H. T. Nguyen. “QTLs for Cell-Membrane Stability Mapped in Rice (*Oryza Sativa* L.) under Drought Stress.” *Theoretical and Applied Genetics* 100, no. 8 (2000): 1197–1202.
- Tuberosa, Roberto, and Silvio Salvi. “Genomics-Based Approaches to Improve Drought Tolerance of Crops.” *Trends in Plant Science* 11, no. 8 (2006): 405–12.
- Tuberosa, R, S Salvi, M C Sanguineti, P Landi, M Maccaferri, and S Conti. “Mapping QTLs Regulating Morpho-Physiological Traits and Yield: Case Studies, Shortcomings and Perspectives in Drought-Stressed Maize.” *Annals of Botany* 89, no. 7 (2002): 941–963.
- Tuinstra MR, G Ejeta, and P Goldsbrough. “Evaluation of near-Isogenic Sorghum Lines Contrasting for QTL Markers Associated with Drought Tolerance,” *Crop Science* 38, no. 3 (1998): 835–42.
- Tuinstra, Mitchell R., Edwin M. Grote, Peter B. Goldsbrough, and Gebisa Ejeta. “Genetic Analysis

- of Post-Flowering Drought Tolerance and Components of Grain Development in Sorghum Bicolor (L.) Moench." *Molecular Breeding* 3, no. 6 (1997): 439–48.
- Tuteja, N., Mahajan, S. "Calcium signaling network in plants: an overview." *Plant Signal. Behav.* 2, (2007):79–85.
- Tuteja, N. "Abscisic Acid and Abiotic Stress signalling." *Plant signalling & Behavior* 2, no. 3 (2007): 135–138.
- Tyers, M., and M. Mann. "From Genomics to Proteomics." *Nature* 422, no. 6928 (2003): 193-197.
- Udall, J.A., Wendel, J.F. "Polyploidy and crop improvement." *Crop Sci.* 46, (2006): S–3.
- Umezawa, Taishi, Naoyuki Sugiyama, Masahide Mizoguchi, Shimpei Hayashi, Fumiyoshi Myouga, Kazuko Yamaguchi-Shinozaki, Yasushi Ishihama, Takashi Hirayama, and Kazuo Shinozaki. "Type 2C Protein Phosphatases Directly Regulate Abscisic Acid-Activated Protein Kinases in Arabidopsis." *Proceedings of the National Academy of Sciences* 106, no. 41 (2009): 17588–17593.
- UniProt Consortium. "Update on activities at the Universal Protein Resource (UniProt) in 2013." *Nucleic acids research* 41, D1 (2013): D43-D47. doi: 10.1093/nar/gks1068.
- UniProt Consortium. "The universal protein resource (UniProt)". *Nucleic acids research* 36 (2008): 190-195.
- van Hemert MJ, HY Steensma, GP van Heusden. "14-3-3 proteins: key regulators of cell division, signalling and apoptosis." *BioEssays*, 10 (2001): 936–946.
- Van Oosterom E.J., R. Jayachandran, and F. R. Bidinger, "Diallel Analysis of the Stay-Green Trait and Its Components in Sorghum," *Crop Science* 36, no. 3 (1996): 549–555.
- Van Oosterom, E.J., A. K. Borrell, K. S. Deifel, and G. L. Hammer. "Does Increased Leaf Appearance Rate Enhance Adaptation to Postanthesis Drought Stress in Sorghum?" *Crop Science* 51, no. 6 (2011): 2728–2740.
- Vanderschuren, H., Lentz, E., Zainuddin, I., Gruissem, W. "Proteomics of model and crop plant species: Status, current limitations and strategic advances for crop improvement." *J. Proteomics* 93, (2013): 5–19.
- Varshney, R. K., M. J. Paulo, S. Grando, F. A. van Eeuwijk, L. C. P. Keizer, P. Guo, S. Ceccarelli, A. Kilian, M. Baum, and A. Graner. "Genome Wide Association Analyses for Drought Tolerance Related Traits in Barley (< I> Hordeum Vulgare</i> L< I>.</i>)." *Field Crops Research* 126 (2012): 171–180.
- Varshney, Rajeev K., Lekha Pazhamala, Junichi Kashiwagi, Pooran M. Gaur, L. Krishnamurthy, and

- Dave Hoisington. "Genomics and Physiological Approaches for Root Trait Breeding to Improve Drought Tolerance in Chickpea (*Cicer Arietinum* L.)." In *Root Genomics*, 233–250. Springer, 2011. [http://link.springer.com/10.1007/978-3-540-85546-0\\_10](http://link.springer.com/10.1007/978-3-540-85546-0_10).
- Varshney, Rajeev K., Mahendar Thudi, Spurthi N. Nayak, Pooran M. Gaur, Junichi Kashiwagi, Lakshmanan Krishnamurthy, Deepa Jaganathan, Jahnavi Koppolu, Abhishek Bohra, and Shailesh Tripathi. "Genetic Dissection of Drought Tolerance in Chickpea (*Cicer Arietinum* L.)." *Theoretical and Applied Genetics* 127, no. 2 (2014): 445–462.
- Varshney, Rajeev K., Wenbin Chen, Yupeng Li, Arvind K. Bharti, Rachit K. Saxena, Jessica A. Schlueter, Mark TA Donoghue, Sarwar Azam, Guangyi Fan, and Adam M. Whaley. "Draft Genome Sequence of Pigeonpea (*Cajanus Cajan*), an Orphan Legume Crop of Resource-Poor Farmers." *Nature Biotechnology* 30, no. 1 (2011): 83–89.
- Veihmeyer FJ and AH Hendrickson. "The Moisture Equivalent as a Measure of the Field Capacity of Soils," *Soil Science* 32, no. 3 (1931): 181–194.
- Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, and Robert A. Holt. "The Sequence of the Human Genome." *Science* 291, no. 5507 (2001): 1304–1351.
- Verdin, Eric, Matthew D. Hirschey, Lydia WS Finley, and Marcia C. Haigis. "Sirtuin Regulation of Mitochondria: Energy Production, Apoptosis, and Signaling." *Trends in Biochemical Sciences* 35, no. 12 (2010): 669–75.
- Vij, Shubha, and Akhilesh K. Tyagi. "A20/AN1 Zinc-Finger Domain-Containing Proteins in Plants and Animals Represent Common Elements in Stress Response." *Functional & Integrative Genomics* 8, no. 3 (2008): 301–307.
- Vinayagam, Arunachalam, Rainer König, Jutta Moormann, Falk Schubert, Roland Eils, Karl-Heinz Glatting, and Sándor Suhai. "Applying Support Vector Machines for Gene Ontology Based Gene Function Prediction." *BMC Bioinformatics* 5, no. 1 (2004): 116.
- Vinocur B, Altman A. "Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations". *Curr Opin Biotechnol* 16, no. 2 (2005):123-132.
- Voordijk S, D Walther, G Bouchet, RD Appel. "Image Analysis Tools in Proteomics," eLS, 2003, <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0006216/full>.
- Wall, G. W., T. J. Brooks, N. R. Adam, A. B. Cousins, B. A. Kimball, P. J. Pinter, R. L. LaMorte, J. Triggs, M. J. Ottman, and S. W. Leavitt. "Elevated Atmospheric CO<sub>2</sub> Improved Sorghum

- Plant Water Status by Ameliorating the Adverse Effects of Drought.” *New Phytologist* 152, no. 2 (2001): 231–248.
- Walsh, Ian, Alberto JM Martin, Catherine Mooney, Enrico Rubagotti, Alessandro Vullo, and Gianluca Pollastri. “Ab Initio and Homology Based Prediction of Protein Domains by Recursive Neural Networks.” *BMC Bioinformatics* 10, no. 1 (2009): 195. <http://www.biomedcentral.com/1471-2105/10/195>.
- Walther EU, Dichgans M, Maricich SM, Romito RR, Yang F, Dziennis S, Zackson S, Hawkes R, Herrup K. "Genomic sequences of aldolase C (Zebrin II) direct lacZ expression exclusively in non-neuronal cells of transgenic mice." *Proc. Natl. Acad. Sci. U.S.A.* 95, 5(1998): 2615–20. doi:10.1073/pnas.95.5.2615. PMC 19434. PMID 948293
- Walulu RS, DT Rosenow, DB Wester and HT Nguyen. “Inheritance of the Stay Green Trait in Sorghum,” *Crop Science* 34, no. 4 (1994): 970–972.
- Wang J and GC Vanlerberghe. “A lack of mitochondrial alternative oxidase compromises capacity to recover from severe drought stress.” *Physiol Plant*, (2013):DOI: 10.1111/ppl.12059.
- Wang X, U Gowik, H Tang, JE Bowers, P Westhoff. “Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses”. *Genome* 10, no. 6 (2009): R68
- Wang X, X Li, and Y Li. “A Modified Coomassie Brilliant Blue Staining Method at Nanogram Sensitivity Compatible with Proteomic Analysis,” *Biotechnology Letters* 29, no. 10 (2007): 1599–1603.
- Wang Xi-Yin , Paterson AH. “Genome Sequencing and Comparative Genomics in Cereals.” *Cereal Genomics II* (2013): 101-126. [http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org\\_Sbicolor](http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor).
- Wang Y., Q. Lu, S. Wu, B. L. Karger, and W.S. Hancock. “Characterization and Comparison of Disulfide Linkages and Scrambling Patterns in Therapeutic Monoclonal Antibodies - Using LC-MS with Electron Transfer Dissociation”. *Anal Chem.* 83, 8 (2011): 3133–3140.
- Wang, Ping, Yalan Wu, Xin Ge, Lan Ma, and Gang Pei. “Subcellular Localization of B-Arrestins Is Determined by Their Intact N Domain and the Nuclear Export Signal at the C Terminus.” *Journal of Biological Chemistry* 278, no. 13 (2003): 11648–53.
- Wang, W., B. Vinocur, and A. Altman. “Plant Responses to Drought, Salinity and Extreme Temperatures: Towards Genetic Engineering for Stress Tolerance.” *Planta* 218, no. 1 (2003): 1–14.



- Wang, Y., X. Zeng, N. J. Iyer, D. W. Bryant, T. C. Mockler, and R. Mahalingam. "Exploring the Switchgrass Transcriptome Using Second-Generation Sequencing Technology." *PloS One* 7, no. 3 (2012): e34225.
- Ware, Doreen H., Pankaj Jaiswal, Junjian Ni, Immanuel V. Yap, Xioakang Pan, Ken Y. Clark, Leonid Teytelman, Steven C. Schmidt, Wei Zhao, and Kuan Chang. "Gramene, a Tool for Grass Genomics." *Plant Physiology* 130, no. 4 (2002): 1606–1613.
- Warren, A. J. "Eukaryotic Transcription Factors." *Current Opinion in Structural Biology* 12, no. 1 (2002): 107–114.
- Washburn, Michael P., Dirk Wolters, and John R. Yates. "Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology." *Nature Biotechnology* 19, no. 3 (2001): 242–247.
- Webster, Hollie, Gabriel Keeble, Bernard Dell, John Fosu-Nyarko, Y. Mukai, Paula Moolhuijzen, Matthew Bellgard, Jizeng Jia, Xiuying Kong, and Catherine Feuillet. "Genome-Level Identification of Cell Wall Invertase Genes in Wheat for the Study of Drought Tolerance." *Functional Plant Biology* 39, no. 7 (2012): 569–79.
- Wei, Kai-Fa, Juan Chen, Yan-Feng Chen, Ling-Juan Wu, and Dao-Xin Xie. "Molecular Phylogenetic and Expression Analysis of the Complete WRKY Transcription Factor Family in Maize." *DNA Research* 19, no. 2 (2012): 153–163.
- Westerhoff, H.V., Kell, D.B. "The methodologies of systems biology." *Syst. Biol. Philos. Found.* (2007): 23–70.
- Westermeier R. "Electrophoresis in Practice." (John Wiley & Sons, 2006), [http://books.google.co.za/bookshl=en&lr=&id=jFfaDsnnzQ04C&oi=fnd&pg=PR5&dq=Electrophoresis+in+practice+&ots=KJ8Q3Smm-F&sig=ILDWdcl2eCKTPibaIUR\\_hTLXwg](http://books.google.co.za/bookshl=en&lr=&id=jFfaDsnnzQ04C&oi=fnd&pg=PR5&dq=Electrophoresis+in+practice+&ots=KJ8Q3Smm-F&sig=ILDWdcl2eCKTPibaIUR_hTLXwg).
- Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, and W. Helmborg. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 33, no. suppl 1 (2005): D39–D45.
- Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, and S. Federhen. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 35, no. suppl 1 (2007): D5–D12.
- Whiteaker, Jeffrey R., Chenwei Lin, Jacob Kennedy, Liming Hou, Mary Trute, Izabela Sokal, Ping Yan, Regine M. Schoenherr, Lei Zhao, and Uliana J. Voytovich. "A Targeted Proteomics-

- based Pipeline for Verification of Biomarkers in Plasma.” *Nature Biotechnology* 29, no. 7 (2011): 625–634.
- Wilhite, D. A., and M. H. Glantz. “Understanding: The Drought Phenomenon: The Role of Definitions.” *Water International* 10, no. 3 (1985): 111–120.
- Wingender, E., X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, and R. Ohnhäuser. “The TRANSFAC System on Gene Expression Regulation.” *Nucleic Acids Research* 29, no. 1 (2001): 281–283.
- Wiśniewski JR, A Zougman, N Nagaraj. “Universal Sample Preparation Method for Proteome Analysis.” *Nature Methods* 6, no. 5 (2009): 359–362.
- Wolfsberg, T.G., Landsman, D. “A comparison of expressed sequence tags (ESTs) to human genomic sequences.” *Nucleic Acids Res.* 25, (1997): 1626–1632.
- Wong CH , Whitesides GM. "Enzymes in synthetic organic chemistry." Tetrahedron organic chemistry series volume 12 (1994): QD262 W65 1994 547. 7'0459-dc2094-2329.
- Wu, F., L. A. Mueller, D. Crouzillat, V. Pétiard, and S. D. Tanksley. “Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade.” *Genetics* 174, no. 3 (2006): 1407–1420.
- Wu, Yingru, Adriane C. Machado, Rosemary G. White, Danny J. Llewellyn, and Elizabeth S. Dennis. “Expression Profiling Identifies Genes Expressed Early during Lint Fibre Initiation in Cotton.” *Plant and Cell Physiology* 47, no. 1 (2006): 107–127.
- Wyrich, R., U. Dressen, S. Brockmann, M. Streubel, C. Chang, D. Qiang, A. H. Paterson, and P. Westhoff. “The Molecular Basis of C 4 Photosynthesis in Sorghum: Isolation, Characterization and RFLP Mapping of Mesophyll-and Bundle-sheath-specific cDNAs Obtained by Differential Screening.” *Plant Molecular Biology* 37, no. 2 (1998): 319–335.
- Xin Z and ML Wang. “Sorghum as a Versatile Feedstock for Bioenergy Production.” *Biofuels* 2, no. 5 (2011): 577–88.
- Xiong, L., K. S. Schumaker, and J. K. Zhu. “Cell signalling During Cold, Drought, and Salt Stress.” *The Plant Cell Online* 14, no. suppl 1 (2002): S165–S183.
- Xiong, Liming, Rui-Gang Wang, Guohong Mao, and Jessica M. Koczan. “Identification of Drought Tolerance Determinants by Genetic Analysis of Root Response to Drought Stress and Abscisic Acid.” *Plant Physiology* 142, no. 3 (2006): 1065–74. doi:10.1104/pp.106.084632.

- Xu, J. H., and J. Messing. "Diverged Copies of the Seed Regulatory Opaque-2 Gene by a Segmental Duplication in the Progenitor Genome of Rice, Sorghum, and Maize." *Molecular Plant* 1, no. 5 (2008a): 760–769.
- Xu, J.-H., Messing, J. "Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species." *Proc. Natl. Acad. Sci.* 105, (2008b): 14330–14335.
- Xu W, PK Subudhi, OR Crasta, DT Rosenow, JE Mullet, HT Nguyen. "Molecular mapping of QTLs conferring stay-green in grain sorghum (*Sorghum bicolor* L. Moench)." *Genome* 43, no.3 (2000): 461-469.
- Xu, Jie, Yibing Yuan, Yunbi Xu, Gengyun Zhang, Xiaosen Guo, Fengkai Wu, Qi Wang, et al. "Identification of Candidate Genes for Drought Tolerance by Whole-Genome Resequencing in Maize." *BMC Plant Biology* 14, no. 1 (2014): 83.
- Xu, W., D. T. Rosenow, and H. T. Nguyen. "Stay Green Trait in Grain Sorghum: Relationship Between Visual Rating and Leaf Chlorophyll Concentration." *Plant Breeding* 119, no. 4 (2000): 365–367.
- Xu, W., P. K. Subudhi, O. R. Crasta, D. T. Rosenow, J. E. Mullet, and H. T. Nguyen. "Molecular Mapping of QTLs Conferring Stay-green in Grain Sorghum (*Sorghum Bicolor* L. Moench)." *Genome* 43, no. 3 (2000): 461–469.
- Xu, Wenwei, Prasanta K. Subudhi, Oswald R. Crasta, Darrell T. Rosenow, John E. Mullet, and Henry T. Nguyen. "Molecular Mapping of QTLs Conferring Stay-Green in Grain Sorghum (*Sorghum Bicolor* L. Moench)." *Genome* 43, no. 3 (2000): 461–69.
- Yamaguchi-Shinozaki, K., and K. Shinozaki. "A Novel Cis-acting Element in an Arabidopsis Gene Is Involved in Responsiveness to Drought, Low-temperature, or High-salt Stress." *The Plant Cell Online* 6, no. 2 (1994): 251–264.
- . "Characterization of the Expression of a Desiccation-responsive Rd29 Gene of Arabidopsis Thaliana and Analysis of Its Promoter in Transgenic Plants." *Molecular and General Genetics MGG* 236, no. 2 (1993): 331–340.
- . "Organization of Cis-acting Regulatory Elements in Osmotic-and Cold-stress-responsive Promoters." *Trends in Plant Science* 10, no. 2 (2005): 88–94.
- . "Transcriptional Regulatory Networks in Cellular Responses and Tolerance to Dehydration and Cold Stresses." *Annu. Rev. Plant Biol.* 57 (2006): 781–803.

- Yamasaki, K., T. Kigawa, M. Seki, K. Shinozaki, and S. Yokoyama. "DNA-binding Domains of Plant-specific Transcription Factors: Structure, Function, and Evolution." *Trends in Plant Science* (2012).
- Yamazaki D, K Motohashi, T Kasama, Y Hara and T Hisabori. "Target proteins of the cytosolic thioredoxins in *Arabidopsis thaliana*." *Plant Cell Physiol* 45, no. 1 (2004): 18-27.
- Yamazaki, Yukiko, and Pankaj Jaiswal. "Biological Ontologies in Rice Databases. An Introduction to the Activities in Gramene and Oryzabase." *Plant and Cell Physiology* 46, no. 1 (2005): 63–68.
- Yan J, C He, J Wang, Z Mao, SA Holaday, RD Allen and H Zhang. "Overexpression of the *Arabidopsis* 14-3-3 Protein GF14 $\lambda$  in Cotton Leads to a "Stay-Green" Phenotype and Improves Stress Tolerance under Moderate Drought Conditions." *Plant Cell Physiol.* 45, no.8 (2004): 1007–1014.
- Yandell M and Ence D: A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet* 2012, **13(5)**: 329-342.
- Yang, Xiaohan, Sara Jawdy, Timothy J. Tschaplinski, and Gerald A. Tuskan. "Genome-Wide Identification of Lineage-Specific Genes In *Arabidopsis*, *Oryza* And *Populus*." *Genomics* 93, no. 5 (2009): 473–480.
- Yaqoob, Muhammad, Nazir Hussain, and Abdul Rashid. "ASSESSMENT OF GENETIC VARIABILITY IN RICE (*ORYZA SATIVA* L.) GENOTYPES UNDER RAINFED CONDITIONS." *Journal of Agricultural Research* 50, no. 3 (2012).
- Yeap WC, Ooi T, Namasivayam P, Kulaveerasingam H, Ho C-L. "EgRBP42 encoding an hnRNP-like RNA-binding protein from *Elaeis guineensis* Jacq. is responsive to abiotic stresses. *Plant Cell Rep stresses* 31, no.10(2012):18231-1843.
- Yin, H., Xue, W., Chen, S., Bogorad, R.L., Benedetti, E., Grompe, M., Kotliansky, V., Sharp, P.A., Jacks, T., Anderson, D.G. "Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype." *Nat. Biotechnol.* (2014).
- Yoneyama, Mitsutoshi, Mika Kikuchi, Takashi Natsukawa, Noriaki Shinobu, Tadaatsu Imaizumi, Makoto Miyagishi, Kazunari Taira, Shizuo Akira, and Takashi Fujita. "The RNA Helicase RIG-I Has an Essential Function in Double-Stranded RNA-Induced Innate Antiviral Responses." *Nature Immunology* 5, no. 7 (2004): 730–37.
- Youens-Clark, Ken, Ed Buckler, Terry Casstevens, Charles Chen, Genevieve DeClerck, Paul

- Derwent, Palitha Dharmawardhana, Pankaj Jaiswal, Paul Kersey, and A. S. Karthikeyan. "Gramene Database in 2010: Updates and Extensions." *Nucleic Acids Research* 39, no. suppl 1 (2011): D1085–D1093.
- Yu, J. Q., S. F. Ye, M. F. Zhang, and W. H. Hu. "Effects of Root Exudates and Aqueous Root Extracts of Cucumber (< I> Cucumis Sativus</i>) and Allelochemicals, on Photosynthesis and Antioxidant Enzymes in Cucumber." *Biochemical Systematics and Ecology* 31, no. 2 (2003): 129–139.
- Yu, Jianming, and Edward S. Buckler. "Genetic Association Mapping and Genome Organization of Maize." *Current Opinion in Biotechnology* 17, no. 2 (2006): 155–160.
- Yu, Liyang. "A Developer's Guide to the Semantic Web." Springer, 2011.  
<http://link.springer.com/content/pdf/10.1007/978-3-642-15970-1.pdf>.
- Yuan, W., and G. Zhou. "Theoretical study and research prospect on drought indices [J]." *Advance in Earth Sciences* 6 (2012): 014
- Yue, Bing, Weiya Xue, Lizhong Xiong, Xinqiao Yu, Lijun Luo, Kehui Cui, Deming Jin, Yongzhong Xing, and Qifa Zhang. "Genetic Basis of Drought Resistance at Reproductive Stage in Rice: Separation of Drought Tolerance From Drought Avoidance." *Genetics* 172, no. 2 (February 2006): 1213–28. doi:10.1534/genetics.105.045062.
- Yue, Bing, Weiya Xue, Lizhong Xiong, Xinqiao Yu, Lijun Luo, Kehui Cui, Deming Jin, Yongzhong Xing, and Qifa Zhang. "Genetic Basis of Drought Resistance at Reproductive Stage in Rice: Separation of Drought Tolerance From Drought Avoidance." *Genetics* 172, no. 2 (January 2, 2006): 1213–1228. doi:10.1534/genetics.105.045062.
- Zhang ZD, A Frankish, T Hunt, J Harrow, M Gerstein. "Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates." *Genome Biol* 11, no. 3 (2010): R26.
- Zhang, Guojie, Guangwu Guo, Xueda Hu, Yong Zhang, Qiye Li, Ruiqiang Li, Ruhong Zhuang, Zhike Lu, Zengquan He, and Xiaodong Fang. "Deep RNA Sequencing at Single Base-pair Resolution Reveals High Complexity of the Rice Transcriptome." *Genome Research* 20, no. 5 (2010): 646–654.
- Zhang, Hengyou. "Genome-Wide Survey and Characterization of Greenbug Induced NAC Transcription Factors in Sorghum [*Sorghum bicolor*(L.) Moench]." In *Plant and Animal Genome XXI Conference*. Plant and Animal Genome, 2013.

- Zhang, J., H. G. Zheng, A. Aarti, G. Pantuwan, T. T. Nguyen, J. N. Tripathy, A. K. Sarial, S. Robin, R. C. Babu, and Bay D. Nguyen. "Locating Genomic Regions Associated with Components of Drought Resistance in Rice: Comparative Mapping within and across Species." *Theoretical and Applied Genetics* 103, no. 1 (2001): 19–29.
- Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., Samatova, N.F. "Detecting differential and correlated protein expression in label-free shotgun proteomics." *J. Proteome Res.* 5, (2006): 2909–2918.
- Zhang, J., W. Jia, J. Yang, and A. M. Ismail. "Role of ABA in Integrating Plant Responses to Drought and Salt Stresses." *Field Crops Research* 97, no. 1 (2006): 111–119.
- Zhang, Xia, Lei Wang, Hui Meng, Hongtao Wen, Yunliu Fan, and Jun Zhao. "Maize ABP9 Enhances Tolerance to Multiple Stresses in Transgenic Arabidopsis by Modulating ABA Signaling and Cellular Levels of Reactive Oxygen Species." *Plant Molecular Biology* 75, no. 4–5 (2011): 365–378.
- Zhang, Yang. "Progress and Challenges in Protein Structure Prediction." *Current Opinion in Structural Biology* 18, no. 3 (2008): 342–348.
- Zhou C, RJ Chen, XL Gao, LH Li, ZJ Xu. "Heterologous expression of a rice RNA-recognition motif gene OsCBP20 in Escherichia coli confers abiotic stress tolerance." *Plant Omics* 7, No. 1(2014): 28-34.
- Zhu, J. K. "Salt and Drought Stress Signal Transduction in Plants." *Annual Review of Plant Biology* 53 (2002): 247.
- Zhu, M, and S Zhao. "Candidate Gene Identification Approach: Progress and Challenges." *International Journal of Biological Sciences* 3, no. 7 (2007): 420–427.
- Zhu, M.-J., Li, X., Zhao, S.-H. "Digital candidate gene approach (DigiCGA) for identification of cancer genes, in: Cancer Susceptibility." *Springer*, (2010): 105–129.
- Zhu, Tao, and Deng-Ke Niu. "Frequency of Intron Loss Correlates with Processed Pseudogene Abundance: A Novel Strategy to Test the Reverse Transcriptase Model of Intron Loss." *BMC Biology* 11, no. 1 (2013): 23.
- Zhuo, D., W. D. Zhao, F. A. Wright, H. Y. Yang, J. P. Wang, R. Sears, T. Baer, D. H. Kwon, D. Gordon, and S. Gibbs. "Assembly, Annotation, and Integration of UniGene Clusters into the Human Genome Draft." *Genome Research* 11, no. 5 (2001): 904–918.

Zhou, X., Su, Z. "EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species." *BMC Genomics* 8, (2007): 246.

Zou, Ming, Baocheng Guo, and Shunping He. "The Roles and Evolutionary Patterns of Intronless Genes in Deuterostomes." *International Journal of Genomics* 2011 (August 11, 2011): e680673.  
Doi:10.1155/2011/680673.



UNIVERSITY *of the*  
WESTERN CAPE

## Appendices

### Appendix 1

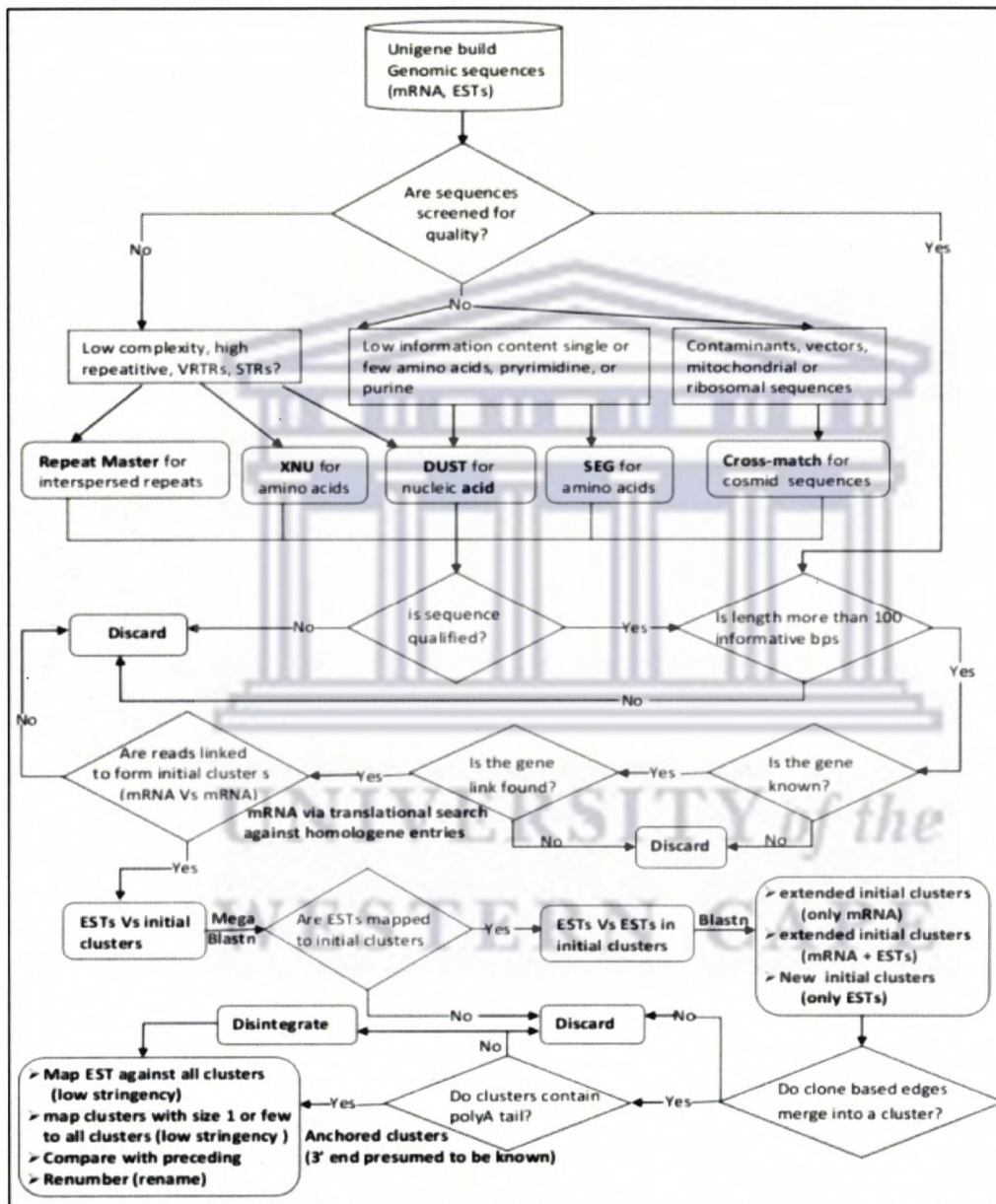


Figure S1.1: Flow chart for sorghum staged clustering of UniGene-build procedure.

This figure serves to justify the exclusion of ESTs from the UniGene clusters. At time of transcripts clustering into UniGene cluster, several parameters were considered. Sequences were rejected from the clusters if one or more criteria were not satisfied. Repeats, sequence length of less than 100 productive bases, enterenze gene record, gene link or links to homologue for fully sequenced organism are common factors. In addition, the extent of sequence similarity, clone-based ends (presence or absence of polyA-tail), sequences mapping to clusters and whether less-sized clusters mapping to other UniGene clusters are included into the parameters.



Table S1.1 Justifiable parameters for staged clustering UniGenes

no.	Parameters	Threshold (cut off values) or required quality	References
1	Repeat masking	High GC content, free of low complexity, mitochondrial or ribosomal sequence, etc (see Figure S1.1)	Smit <i>et al.</i> 2012
2	sequence length	$\geq 100$ informative base pairs	Hide <i>et al.</i> , 1999
3	gene link	Homologene	Sayers <i>et al.</i> , 2011
4	gene record	Entrenze gene	Sayers <i>et al.</i> , 2011
5	sequence homology	$\geq 80\%$ identity	ERSOZ <i>et al.</i> , 2012
6	clone-based ends (presence or absence of polyA-tail)	Presence of PolyA-tail	Kim <i>et al.</i> , 2005
7	extent of sequences mapping to clusters	Presence of 3prime end	Hide <i>et al.</i> , 1999; Bordat <i>et al.</i> , 2011
8	extent of less-sized clusters	mapping of less-sized clusters to other UniGene clusters	Sayers <i>et al.</i> , 2011 and Acland <i>et al.</i> , 2013



UNIVERSITY of the  
WESTERN CAPE

## Appendix 2

Table S2.1 Overview of UniGene libraries (build # 29)

Main parts of plants	Tissues	Number of libraries	Library ID	Name of Library	Sequence number	
Body site	Callus	1	Lib.13735	Callus culture/cell suspension	10449	
	embryo	2	Lib.5437	Embryo 1 (EM1)	9843	
	leaf		12	Lib.15546	Wounded leaves	11221
				Lib.10086	Pathogen-infected compatible 1 (PIC1)	9092
				Lib.14497	Drought-stressed after flowering	6295
				Lib.14500	Drought-stressed before flowering	3597
				8 not shown	Not shown	each <1000
				ovary	3	Lib.5439
				Lib.7266	Ovary 2 (OV2)	4983
				1 not shown	Not shown	<1000
				panicle	2	Lib.9519
				1 not shown	Not shown	<1000
				pollen	1	Lib.14372
	chloroplast	1	not shown	Not shown	each <1000	
	root		3	Lib.15544	Acid- and alkaline-treated roots	7744
				Lib.16897	Anaerobic roots	6113
				Lib.20762	Sorghum bicolor BTx623 Root hair	5468
	shoot		2	Lib.16898	GA- or brassinolide-treated seedlings	11134
				1 not shown	not shown	<1000
	Whole	4	All not shown	Not shown	each <1000	
	Unspecified tissue	1	not shown	Not shown	<1000	
	mixed		13	Lib.4037	Dark Grown 1 (DG1)	11099
				Lib.13713	Heat-shocked seedlings	10558
				Lib.15545	Oxidatively-stressed leaves and roots	10086
				Lib.4038	Water-stressed 1 (WS1)	10039
Lib.13770				Ethylene-treated seedlings	6942	
Lib.13736				Salt-stressed seedlings	6737	
Lib.13769				Salicylic acid-treated seedlings	5801	
Lib.12996				Abscisic acid-treated seedlings	4907	
Lib.14109				Iron-deficient seedlings	3984	
Lib.14297				Nitrogen-deficient seedlings	3849	
Lib.14296				Phosphorous-deficient seedlings	3723	
2 Not shown					each < 1000	
not yet classified					4	Lib.5441
	Lib.2801	Light Grown 1 (LG1)	9451			
	2 Not shown	Not shown	each < 1000			
Developmental stage	Germinating seed	2	Not shown	Not shown	Each < 1000	
	seedling	25	Not shown	Not shown	each < 1000	

Main parts of plants	Tissues	Number of libraries	Library ID	Name of Library	Sequence number
	vegetative	9	Not shown	Not shown	each < 1000
	flowering	1	Not shown	Not shown	< 1000
	ripening	1	Not shown	Not shown	< 1000
	unknown	1	Not shown	Not shown	< 1000
	developmental stage	3	Not shown	Not shown	each < 1000
	not yet classified	2	Not shown	Not shown	each < 1000

Table S2.2 Chromosomal distribution of UniGene clusters mapped to genome

Chromosomes	UniGene clusters that overlapped Known genes			UniGene clusters that mapped to intergenic region			Grand Total
	DR <sup>1</sup>	Non-DR(DR) <sup>2</sup>	Total	DR	Non-DR	Total	
Chr1	20	1733 (37)	1753	22	91	113	1866
Chr2	14	1181 (33)	1195	20	69	89	1284
Chr3	16	1322 (39)	1338	15	73	88	1427
Chr4	13	1150 (35)	1163	12	64	76	1239
Chr5	4	432 (13)	436	7	49	56	493
Chr6	19	876 (21)	895	10	64	74	966
Chr7	9	670 (23)	679	9	61	70	749
Chr8	8	544 (13)	552	7	63	70	622
Chr9	14	883 (27)	897	12	66	78	976
Chr10	6	750 (15)	756	7	54	61	817
Super	0	99 (2)	99	7	74	81	180
Total	123	9640 (258)	9763	128	728	856	10619

Key to legend:

<sup>1</sup> UniGene clusters that represent drought response

<sup>2</sup> UniGene clusters that do not represent drought response, however, contain shorter ESTs as member of the clusters which were expressed under drought conditions.

**Table S2.3** Comparison and update of annotation

Description	Existing annotation		Current annotation			
	gene models	mRNA	TIGR transcripts		UniGene clusters	
			gene models	mRNA	gene models	mRNA
Original	34496	36338	30145	31890	34374	36216
Updated	-	-	3724	4447	106	122*
Total	34496	36338	34494	36337	34496	36338

\* Out of 122 mRNAs obtained using UniGene clusters, 12 mRNAs including the merging mRNAs were redundant with mRNAs obtained using TIGR transcript. This supplementary table shows the comparison and update of annotation between existing and current prediction of sorghum genome based on PASA alignment evidence

Table S2.4 Databases containing potential candidate drought responsive genes

No.	Databases	Link or reference
1	Genome database: Phytozome	<a href="http://www.phytozome.net/sorghum">http://www.phytozome.net/sorghum</a> or <a href="ftp://ftp.jgi-psf.org/pub/comp/gen/phytozome/v9.0/Sbicolor_v1.4/">ftp://ftp.jgi-psf.org/pub/comp/gen/phytozome/v9.0/Sbicolor_v1.4/</a> ; Goodstein <i>et al.</i> , 2012 <a href="http://www.gramene.org/">http://www.gramene.org/</a> ; Ware <i>et al.</i> , 2002
	Gramene	<a href="http://www.ensembl.org/index.html">http://www.ensembl.org/index.html</a> ; Flicek <i>et al.</i> , 2013 and 2014
	Ensembl	
2	sequences databases: RefSeq	<a href="http://www.ncbi.nlm.nih.gov/refseq/">http://www.ncbi.nlm.nih.gov/refseq/</a>
	NCBI	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
	UniGene	<a href="http://www.ncbi.nlm.nih.gov/UniGene/">http://www.ncbi.nlm.nih.gov/UniGene/</a> ; Rudd <i>et al.</i> , 2003
	dbESTs	<a href="http://www.ncbi.nlm.nih.gov/dbEST/">http://www.ncbi.nlm.nih.gov/dbEST/</a> ; Boguski

		<i>et al.</i> , 1993
	Uniprot	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>
3	Protein domain databases	
	Uniref	<a href="http://www.uniprot.org/help/uniref">www.uniprot.org/help/uniref</a>
	Uniprot	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>
	Swissprot	<a href="http://www.expasy.ch/sprot/">http://www.expasy.ch/sprot/</a>
		<a href="http://www.ebi.ac.uk/swissprot/">http://www.ebi.ac.uk/swissprot/</a>
	Interoroscan	<a href="https://www.ebi.ac.uk/interpro/">https://www.ebi.ac.uk/interpro/</a>
	Signal peptides	<a href="http://www.signalpeptide.com/">http://www.signalpeptide.com/</a>
4	expression databases:	
	Gene Expression Omnibus (GEO):	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>
	experimental genome microarray data,	
	RNA-seq data	

Table S2.5 Relevant tools for identification of the candidate gene

Ser. #	Tools	Links or references
1	alignment tools:	
	Blast,	<a href="http://blast.ncbi.nlm.nih.gov/Blast.cgi">http://blast.ncbi.nlm.nih.gov/Blast.cgi</a> ; (Altschul <i>et al.</i> , 1990)
	Blat,	<a href="http://genome.ucsc.edu/FAQ/FAQblat.html">http://genome.ucsc.edu/FAQ/FAQblat.html</a>
	est2genome	<a href="http://bioinfo.nhri.org.tw/cgi-bin/emboss/est2genome">http://bioinfo.nhri.org.tw/cgi-bin/emboss/est2genome</a> , (Mott, 1997)
	exonerate	<a href="http://www.csc.fi/english/research/sciences/bioscience/programs/exonerate/index_html">http://www.csc.fi/english/research/sciences/bioscience/programs/exonerate/index_html</a>
	ClustalW	<a href="http://www.genome.jp/tools/clustalw/">http://www.genome.jp/tools/clustalw/</a>
	ClustalX	<a href="http://www.clustal.org/clustal2/">http://www.clustal.org/clustal2/</a>
2	Gene structure prediction tools:	
	Augustus,	<a href="http://augustus.gobics.de/binaries/">http://augustus.gobics.de/binaries/</a>
	est2genome	<a href="ftp://ftp.hgc.jp/pub/mirror/ebi/software/exonerate/">ftp://ftp.hgc.jp/pub/mirror/ebi/software/exonerate/</a>
	exonerate, and	<a href="http://bioinfo.nhri.org.tw/cgi-bin/emboss/est2genome">http://bioinfo.nhri.org.tw/cgi-bin/emboss/est2genome</a>
	PASA	<a href="http://www.evidencemodul">http://www.evidencemodul</a>
	Evidence Moduler (EVM),	<a href="http://evidencemodeler.sourceforge.net/">http://evidencemodeler.sourceforge.net/</a>

3	Gene enrichment analysis tools:	
	AgriGO	<a href="http://www.agriGO.org/">http://www.agriGO.org/</a> ; Du et al., 2010
	Blast2go	<a href="http://www.blast2go.de/">http://www.blast2go.de/</a> ; Conesa et al., 2005
	MeV,	<a href="http://www.tm4.org/mev.html">http://www.tm4.org/mev.html</a>
	Genevestigator,	<a href="https://www.genevestigator.ethz.ch/">https://www.genevestigator.ethz.ch/</a>
4	Genome viewer tools:	
	Galaxy,	<a href="http://usegalaxy.org/">http://usegalaxy.org/</a> ; <a href="https://main.g2.bx.psu.edu/">https://main.g2.bx.psu.edu/</a> ; Goecks et al., 2010
	Gbrowse	<a href="http://gmod.org/wiki/GBrowse/tool_data">http://gmod.org/wiki/GBrowse/tool_data</a> ;
	USCS-Gbrowse	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a> ;
	Biomart-ensemble,	<a href="http://www.ensembl.org/info/data/biomart.html">http://www.ensembl.org/info/data/biomart.html</a> ;
	Biomart-Gramene	<a href="http://archive.gramene.org/biomart/martview/334378ea281f8c0071af7d45f4ab01f9">http://archive.gramene.org/biomart/martview/334378ea281f8c0071af7d45f4ab01f9</a> ;
	Biomart Phytosome	<a href="http://www.phytosome.net/biomart/martview/6ae06d47637222933cd94d5beb507c71">http://www.phytosome.net/biomart/martview/6ae06d47637222933cd94d5beb507c71</a>
5	Expression profiling analysis tools:	
	Multivariate Experiment Viewer (MeV)	<a href="http://www.tm4.org/mev.html">http://www.tm4.org/mev.html</a> ; Saeed et al., 2003
6	Sequence quality filtering tools:	
	RepeatMasker	<a href="http://www.repeatmasker.org/">http://www.repeatmasker.org/</a> ; Smit et al. 2012
	Statistical analysis tools:	
	R programming,	<a href="http://www.r-project.org/">http://www.r-project.org/</a>
	MeV	<a href="http://www.tm4.org/mev.html">http://www.tm4.org/mev.html</a>
7	<b>Programming tools:</b>	
	linux OS	<a href="http://www.gnu.org/">http://www.gnu.org/</a>
	Perl,	<a href="http://www.perlmonks.org/">http://www.perlmonks.org/</a>
	Python,	<a href="http://www.tutorialspoint.com/python/python_tools_utilities.htm">http://www.tutorialspoint.com/python/python_tools_utilities.htm</a>
	R, programming modules,	<a href="http://www.r-project.org/">http://www.r-project.org/</a>
	Bioperl	<a href="http://www.bioperl.org/">www.bioperl.org/</a>
	Biopython	<a href="http://www.biopython.org/">www.biopython.org/</a> ; <a href="http://biopython.org/wiki/Main_Page">http://biopython.org/wiki/Main_Page</a>

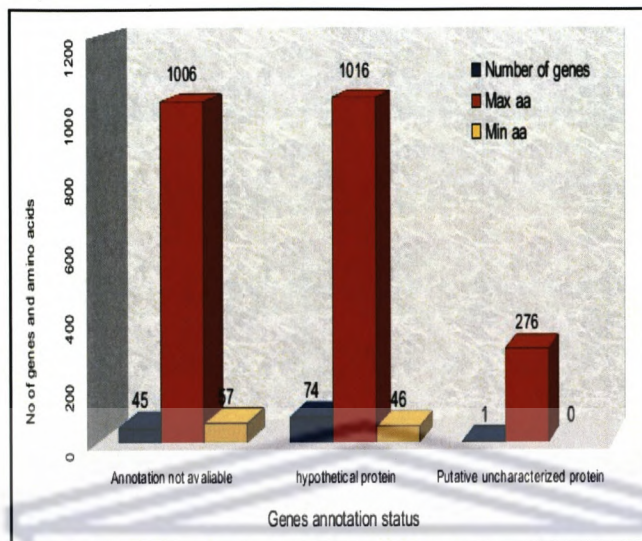


Figure S2.1: Predicted overlapping gene annotation and characterization status

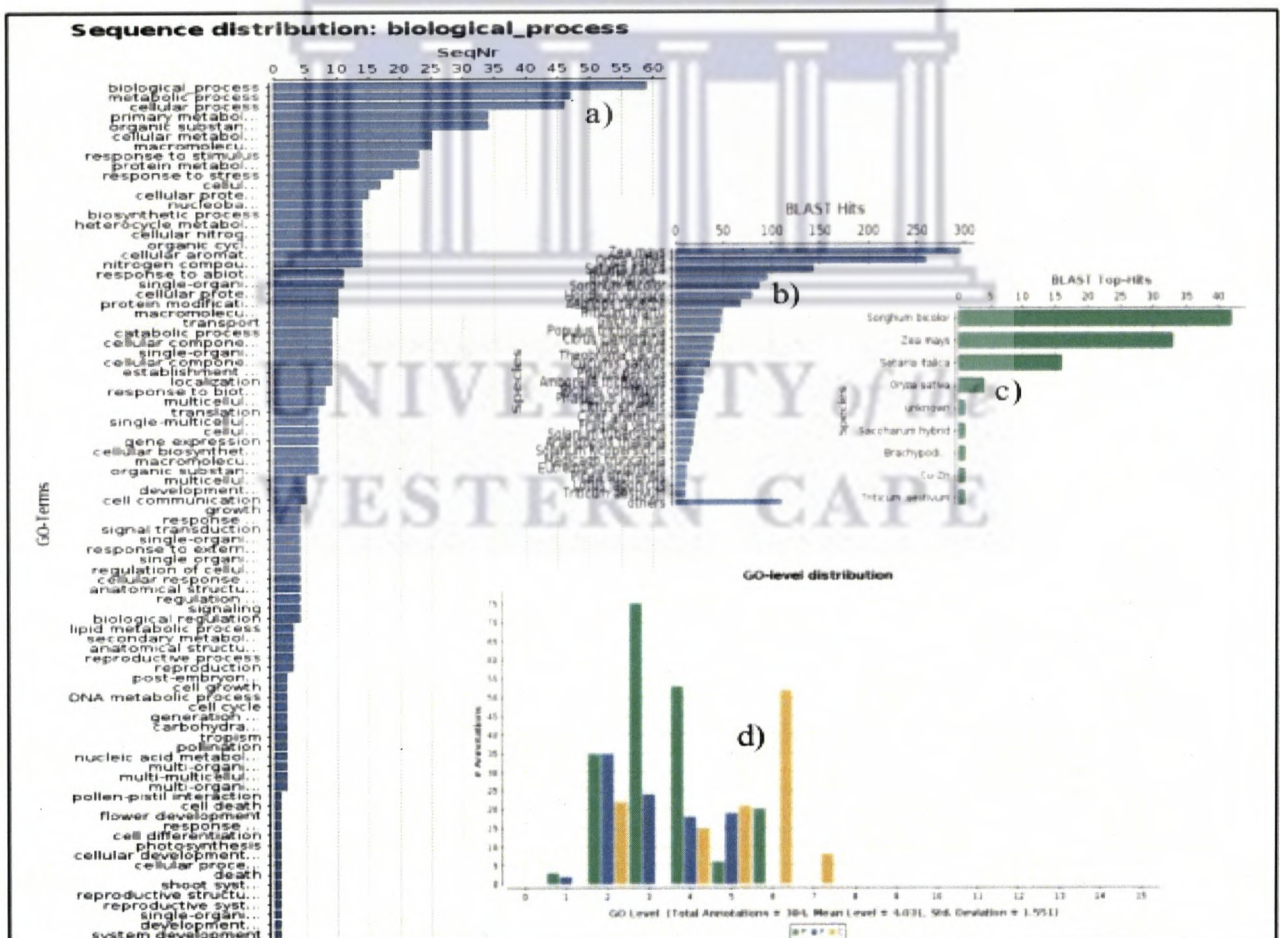


Figure S2.2: GO annotation based on blasting and mapping to non-redundant databases

GO annotation based on blasting and mapping sequences to non-redundant databases based on blast2go: a) patterns of sequences distribution based on blast hits associated to the GO-terms for the biological process; b) Species specific score for the total blast-hit; c) species specific blast top hits and d) GO-level distribution of annotations





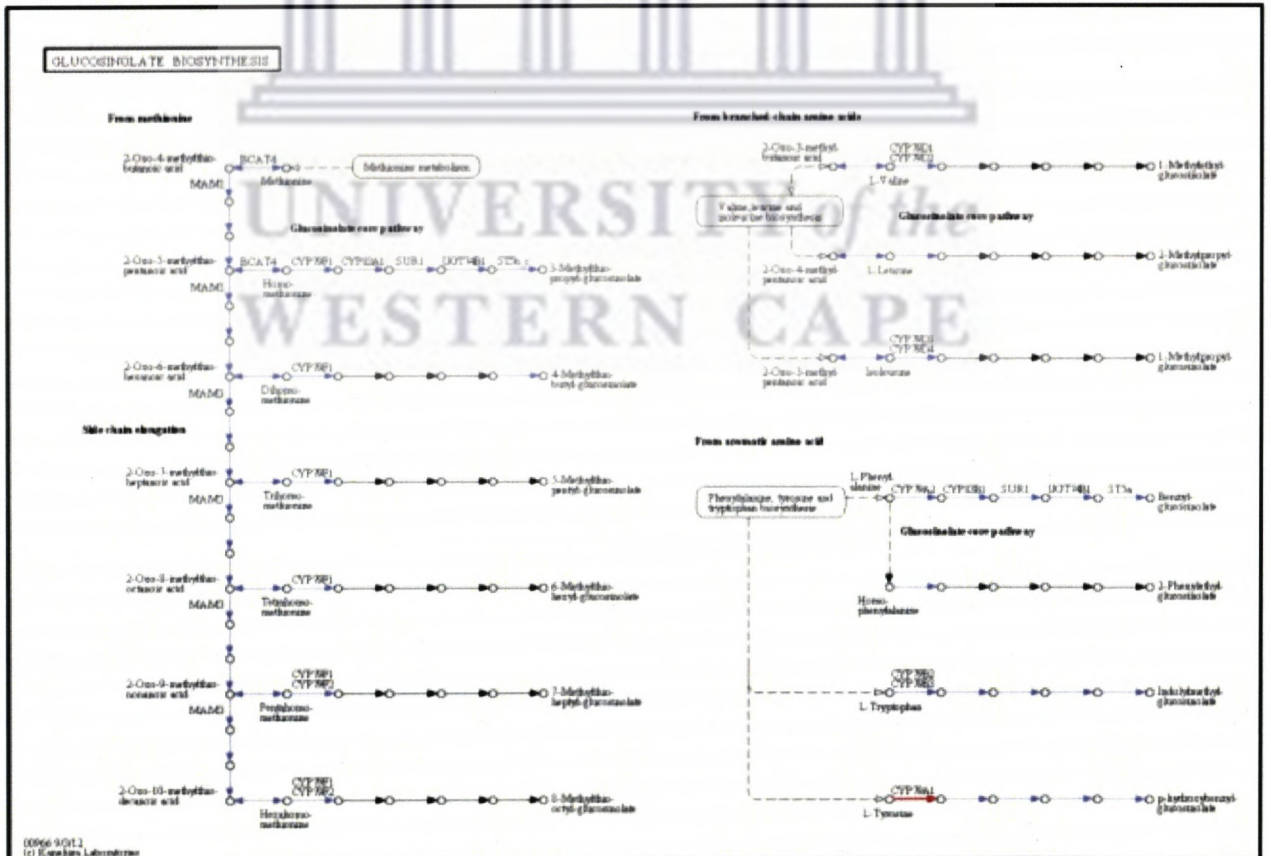
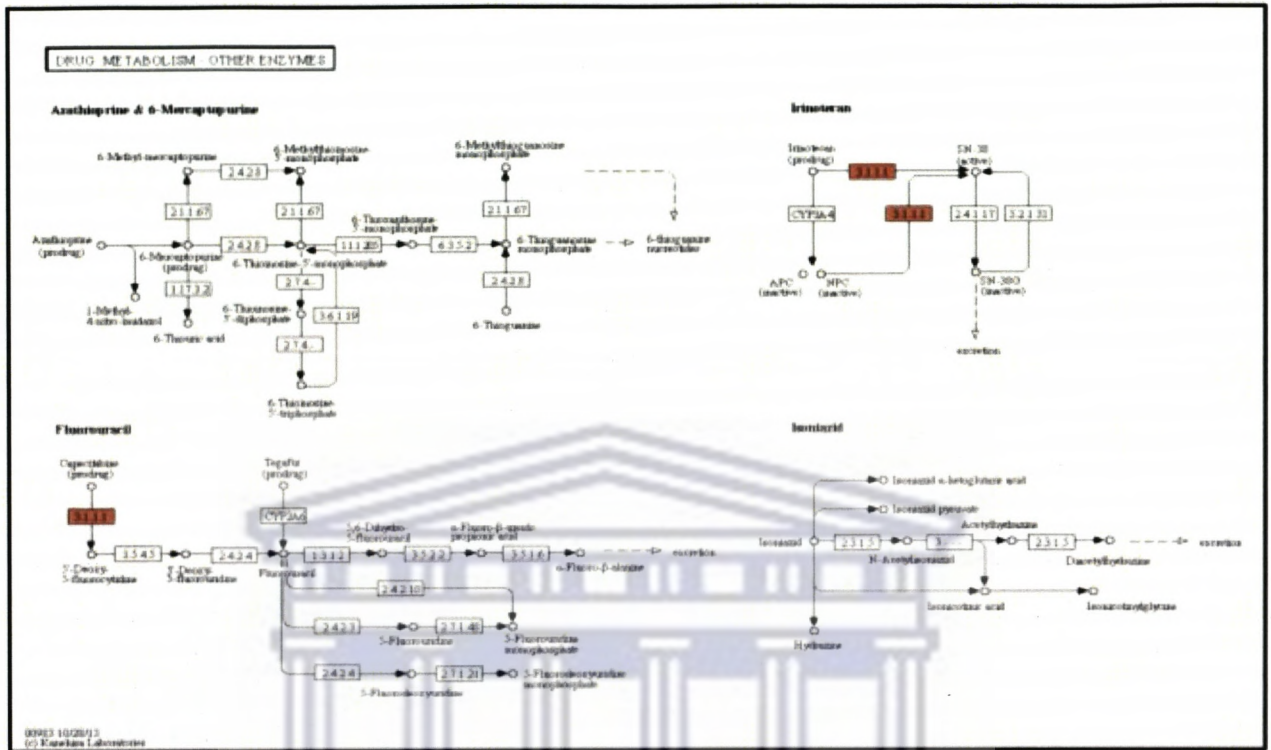


Figure S2.3B: Diagram of the two metabolic pathways (Drug metabolism - other enzymes and Glucosinolate biosynthesis), continued.

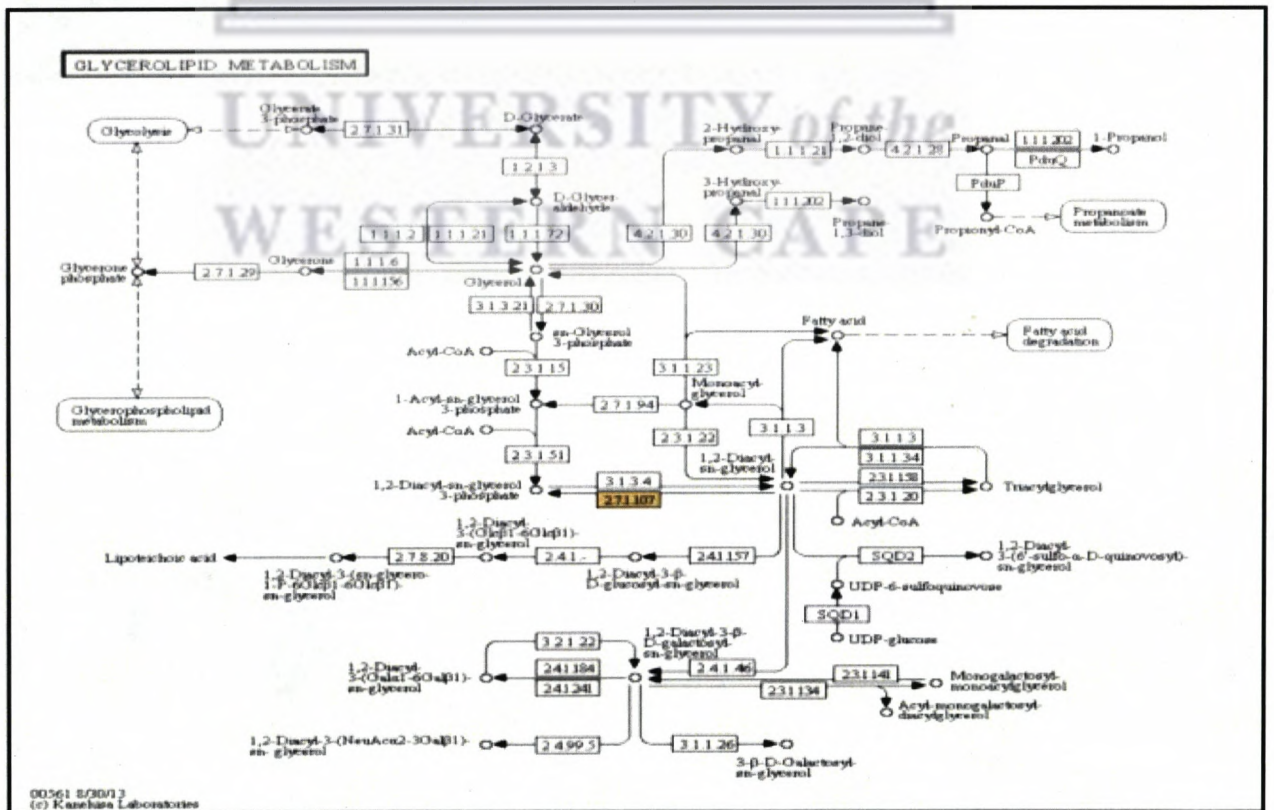
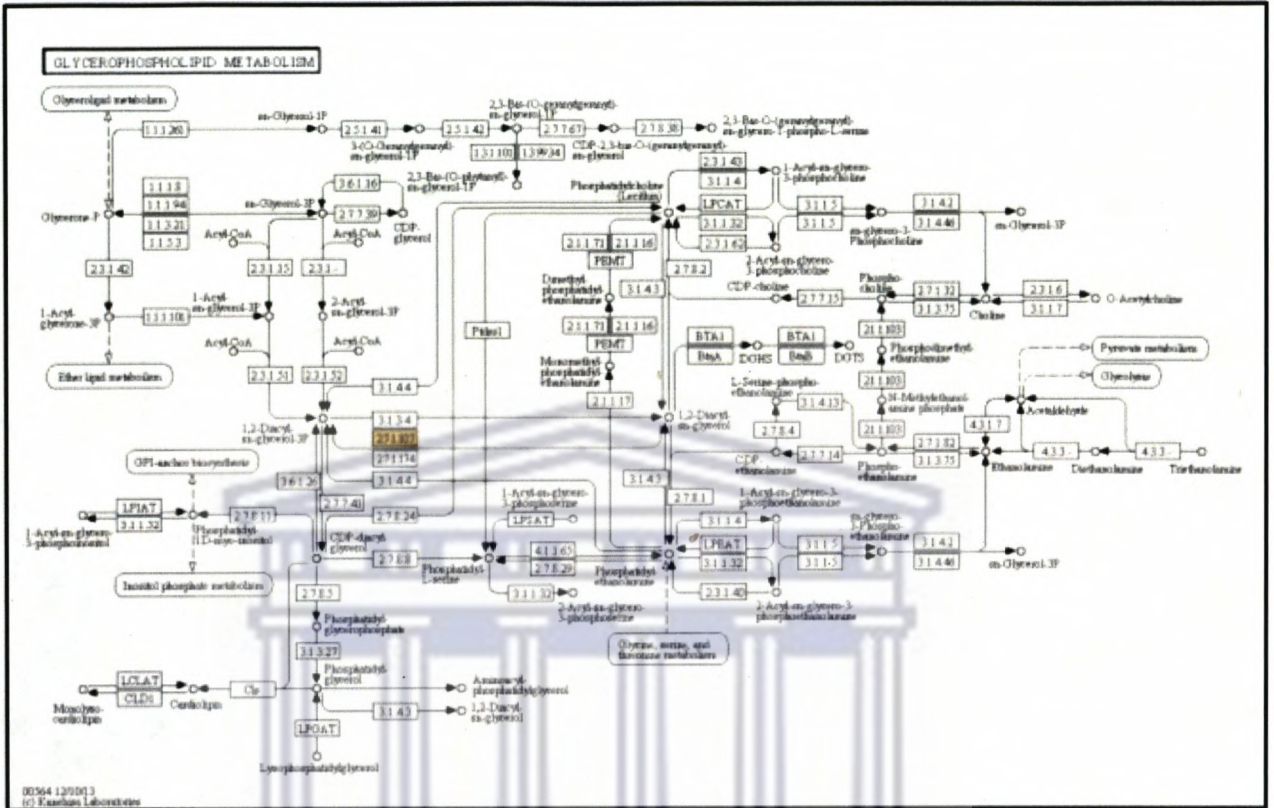


Figure S2.3C: Diagram of the two metabolic pathways (Glycerophospholipid metabolism and Glycerolipid metabolism), continued.







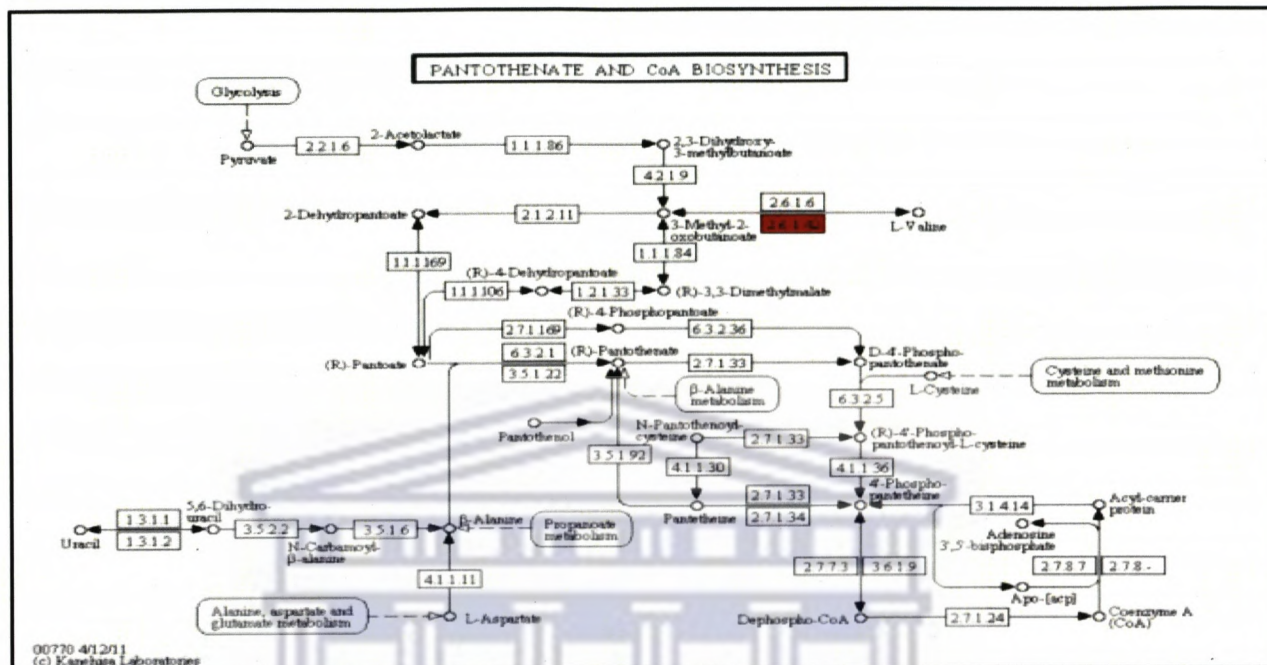


Figure S2.3G: Diagram of the two metabolic pathways (Pantothenate and CoA biosynthesis), continued.

Figure S2.3: The 13 metabolic pathways among 14 identified (Figures S2.4A – G).

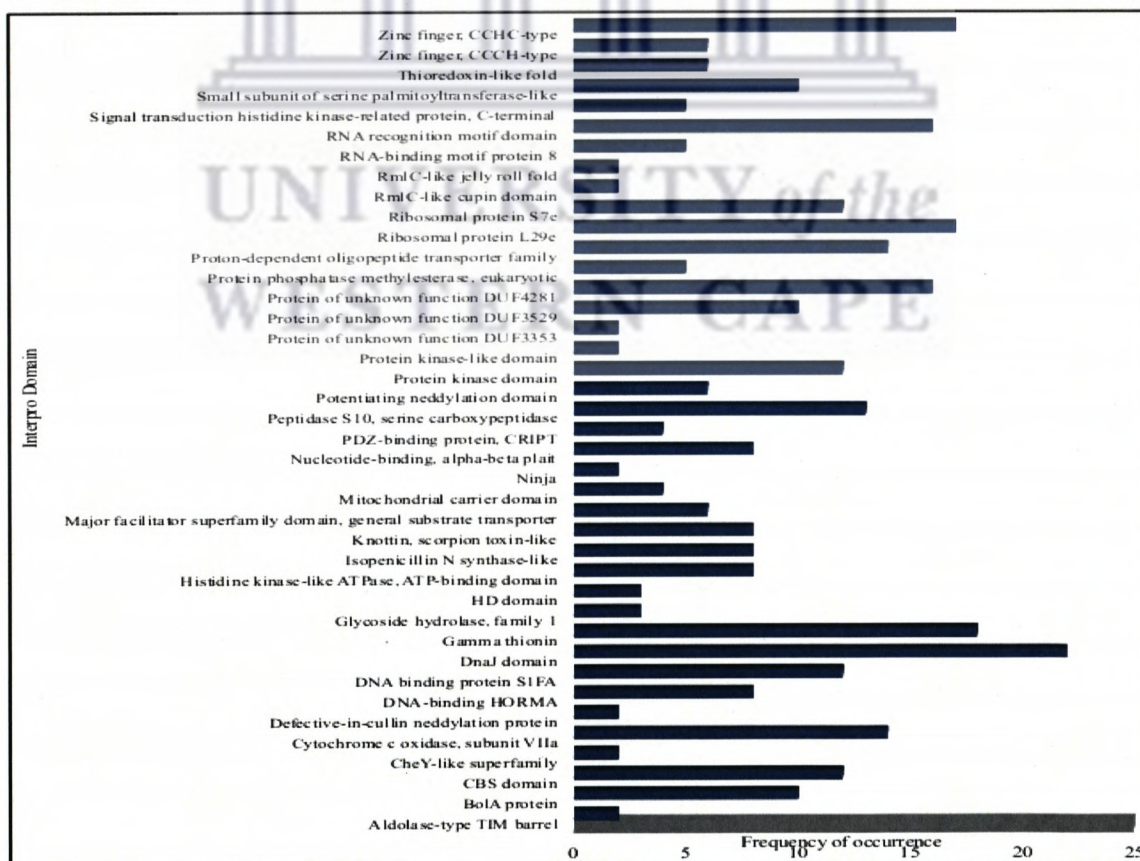
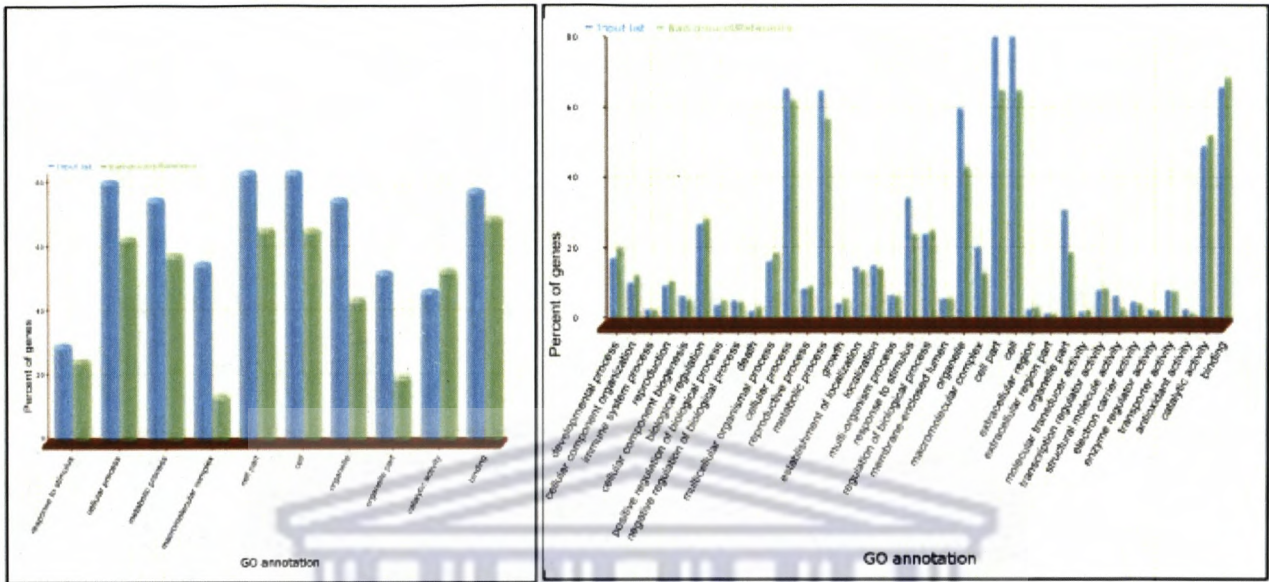


Figure S2.4: Description of interpro domain analysis: List of protein signatures identified.



a) chart for significantly expressed 45 genes supported by all statistical models based on analysis of gene-expression (see also section 2.4.3.1.1) b) chart for significantly expressed total 1079 genes supported by individual statistical models based on analysis of gene-expression (see also section 2.4.3.1.1)

Figure S2.5: Sorghum % GO-terms assigned genes identified from maize orthologs.

This figure shows the percentage of sorghum genes identified from maize orthologs that are assigned to different GO-terms associated to drought effects on gene expression in maize reproductive and Leaf meristem tissue (Kakumanu et al., 2012). a) chart for significantly expressed genes supported by all statistical models and b) chart for significantly expressed total 1079 genes supported individual statistical models.

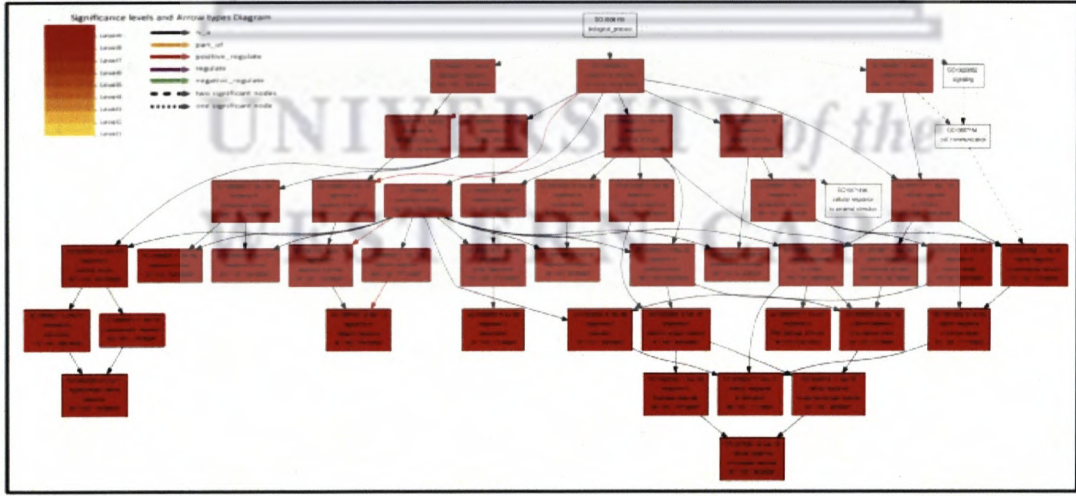


Figure S2.6: Mapping of GO terms related to responses to stress based on biological process.

This map was generated using agriGO software package (Du et al., 2010 ) by browsing in tree traversing mode selecting only significantly overrepresented GO-terms most related to drought tolerance. Drought stress regulated genes were represented as an expression of mRNAs with more than two-fold changes under differential conditions. The number inside the coloured box represent GO accessions, p-values, # of enriched genes in the test set per total number of genes involved in the test set and those in the sorghum database associated with GO-term per total number of genes in the database involved in the background set of the GO-term (see legend for further description).

### Appendix 3

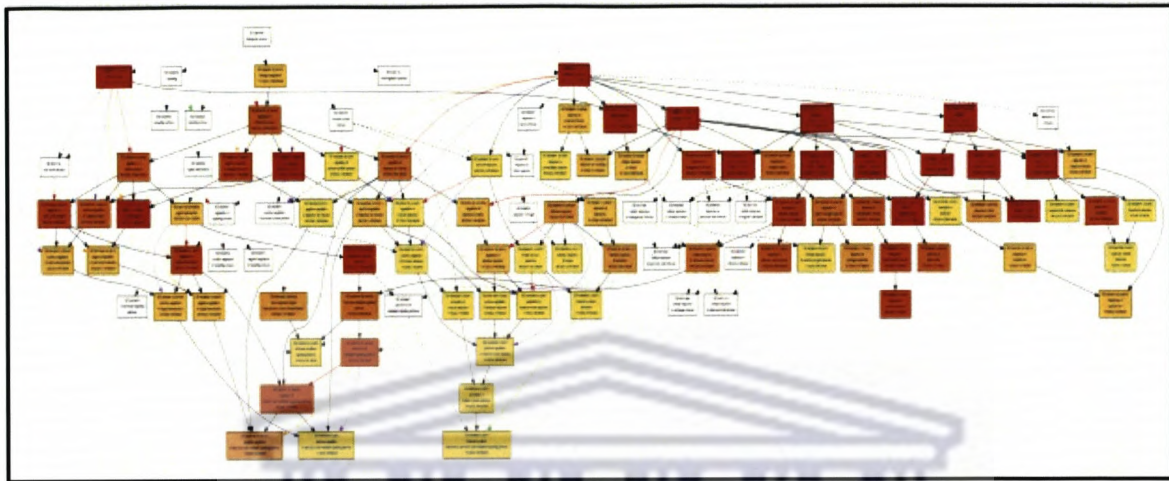


Figure S3.1: Graphical views of significantly enriched GO-terms

The graphical views of significantly enriched GO-terms for the identified genes shows strong association with drought-related terms based on biological processes. Analysis of GO of sorghum drought-stress responsive genes using agriGO (Du et al., 2010). Drought stress regulated genes were represented as an expression of mRNAs with more than two-fold under differential conditions. Information in each box shows the GO ids, the p-value and GO term, number of enriched genes in the test set per total number of genes involved in the test set and those in the sorghum database associated with GO-term per total number of genes in the database involved in the background set of the GO-term (see legend for further description).

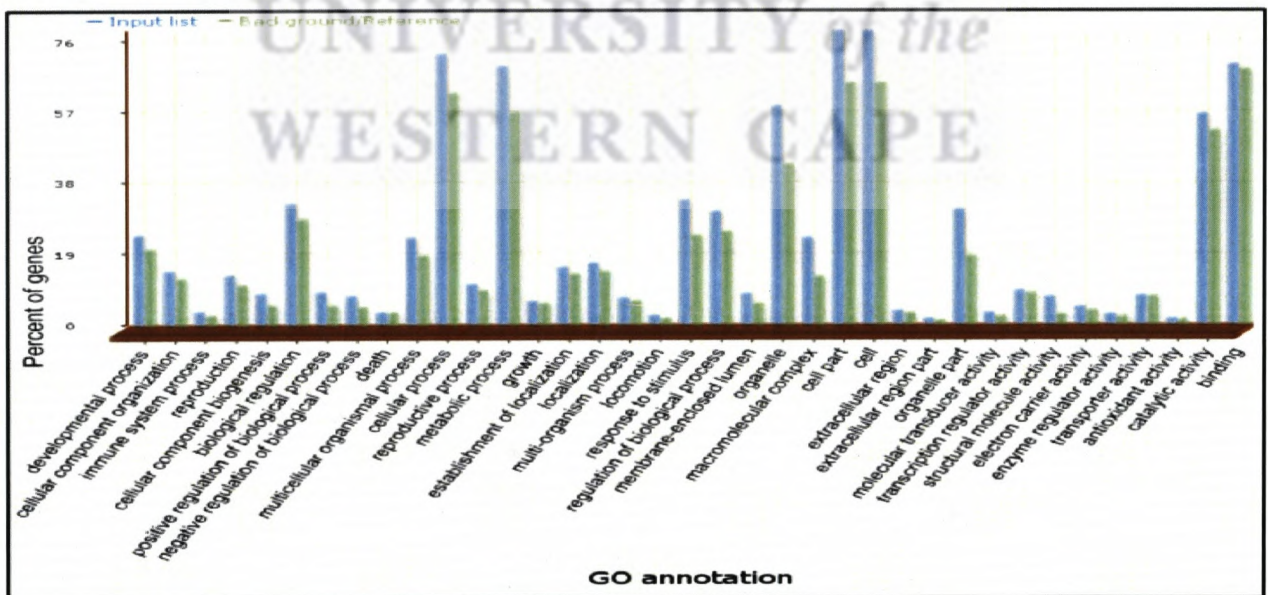


Figure S3.2: GO-annotation Vs % gene enriched in the particular GO-domain.

The y-axis shows the number of genes in percentage that was significantly enriched and the x-axis represent the GO-annotation depicting all the GO-terms that are associated to specific genes for all the GO-domains.



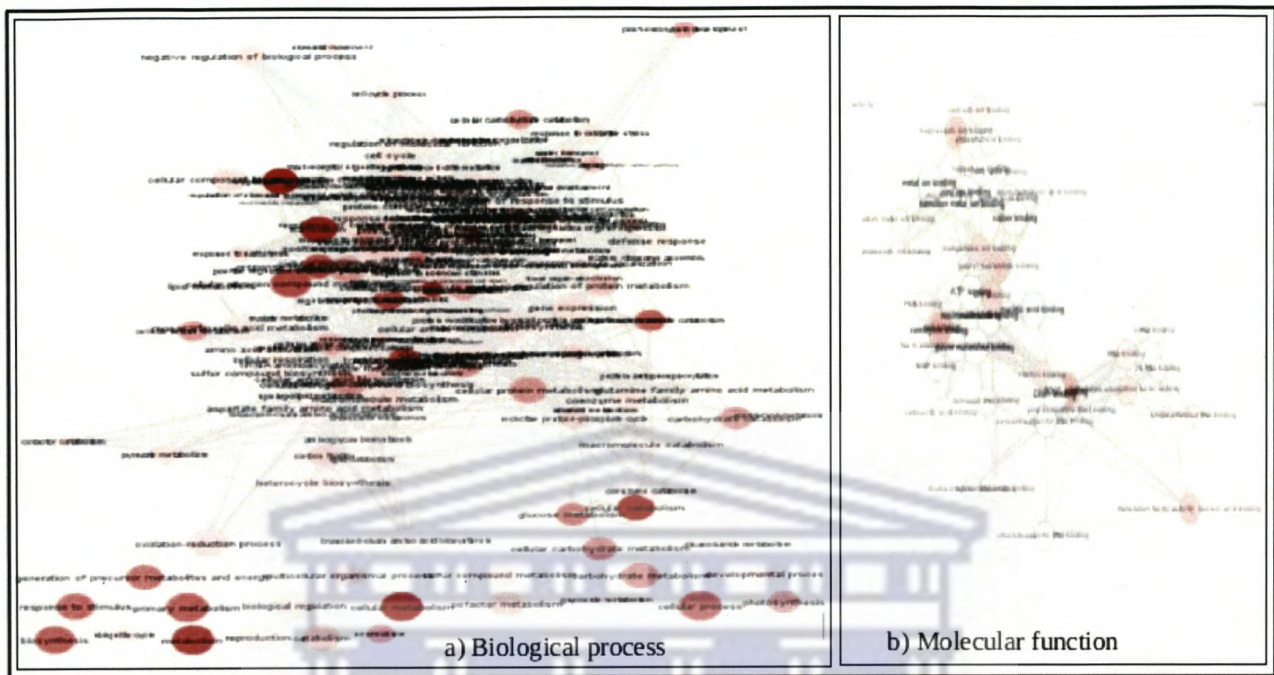
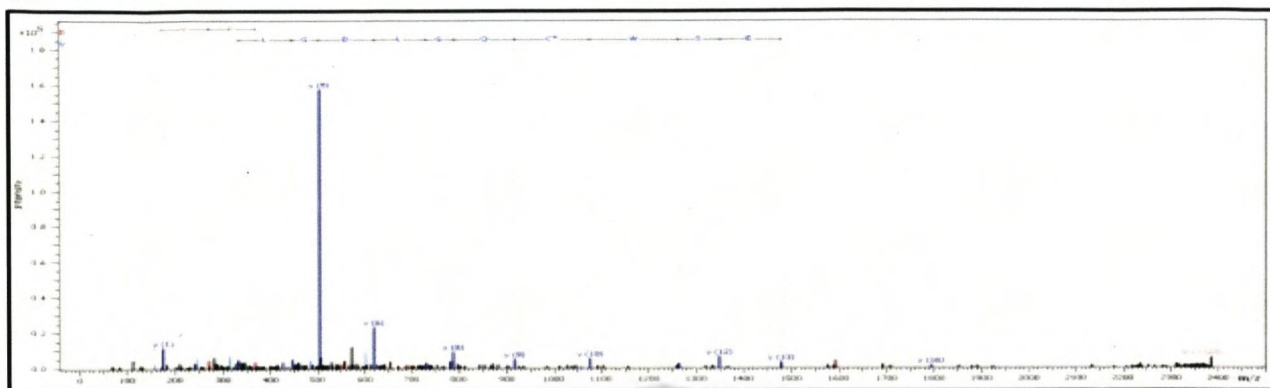


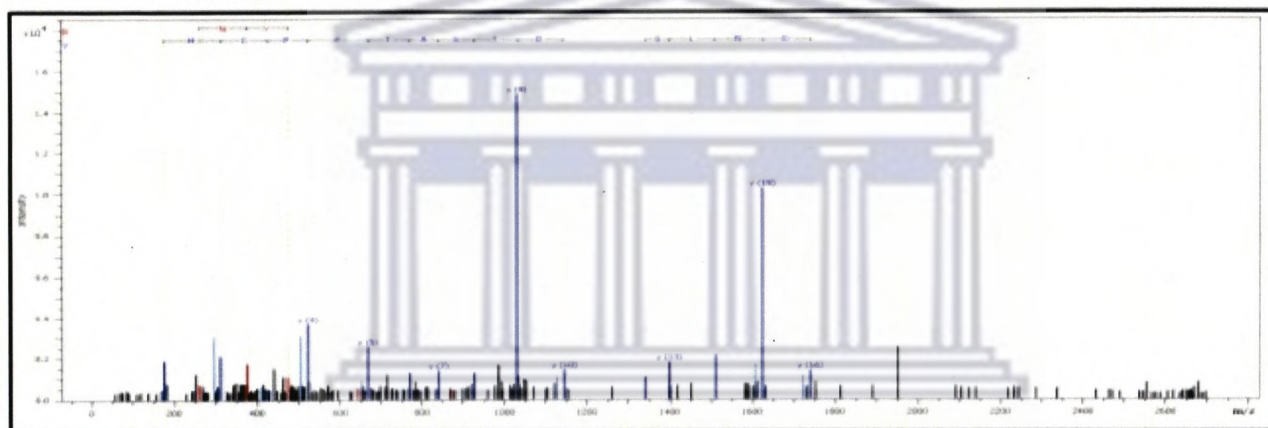
Figure S3.3: Interactive association of genes for enriched GO-terms. This graphical representation of interactive association of genes represent significantly enriched GO-terms with all deterministic factors: a) Interactive association of genes significantly involve in a) biological processes and b) molecular functions.

UNIVERSITY of the  
WESTERN CAPE

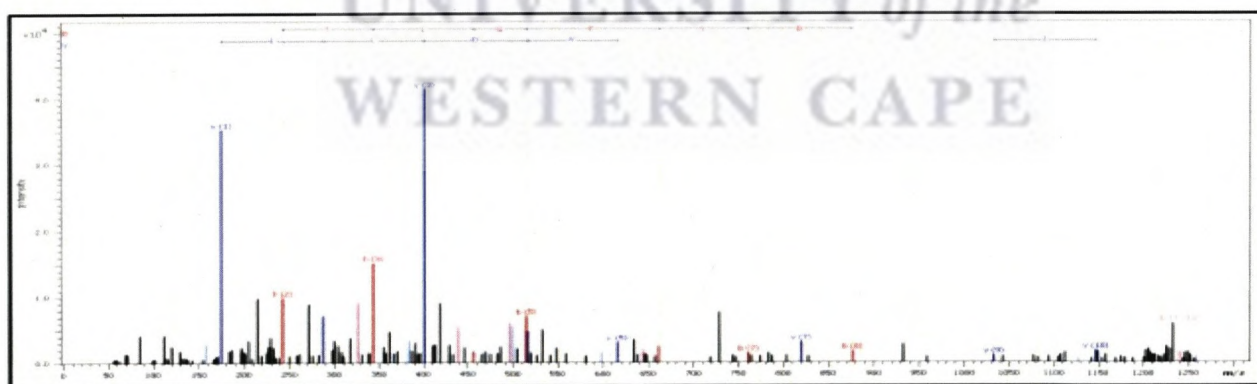
## Appendix 4



Fructose-bisphosphate aldolase, chloroplastic OS=*Oryza sativa* subsp. *japonica* GN=Os11g0171300 PE=1 SV=2



ATP synthase subunit beta, chloroplastic OS=*Saccharum* hybrid GN=atpB PE=3 SV=1



Ribulose bisphosphate carboxylase large chain OS=*Lactuca sativa* GN=rbcl PE=3 SV=2 RuBisCO

Figure S4.1: MALDI-TOF-TOF-MS/MS spectrum of protein spot in-gel tryptic digest.

This figure shows MALDI-TOF-TOF-MS/MS spectrum of an in-gel tryptic digest of putatively identified proteins: Fructose-bisphosphate aldolase, ATP synthase subunit beta and Ribulose bisphosphate carboxylase large chain, RuBisCO. The spectra were produced by calibrating internally using trypsin autohydrolysis peaks for the  $m/z$  values. These three spectra are presented here to represent the peptides from the top three proteins identified in this analysis based on MOWSE score.