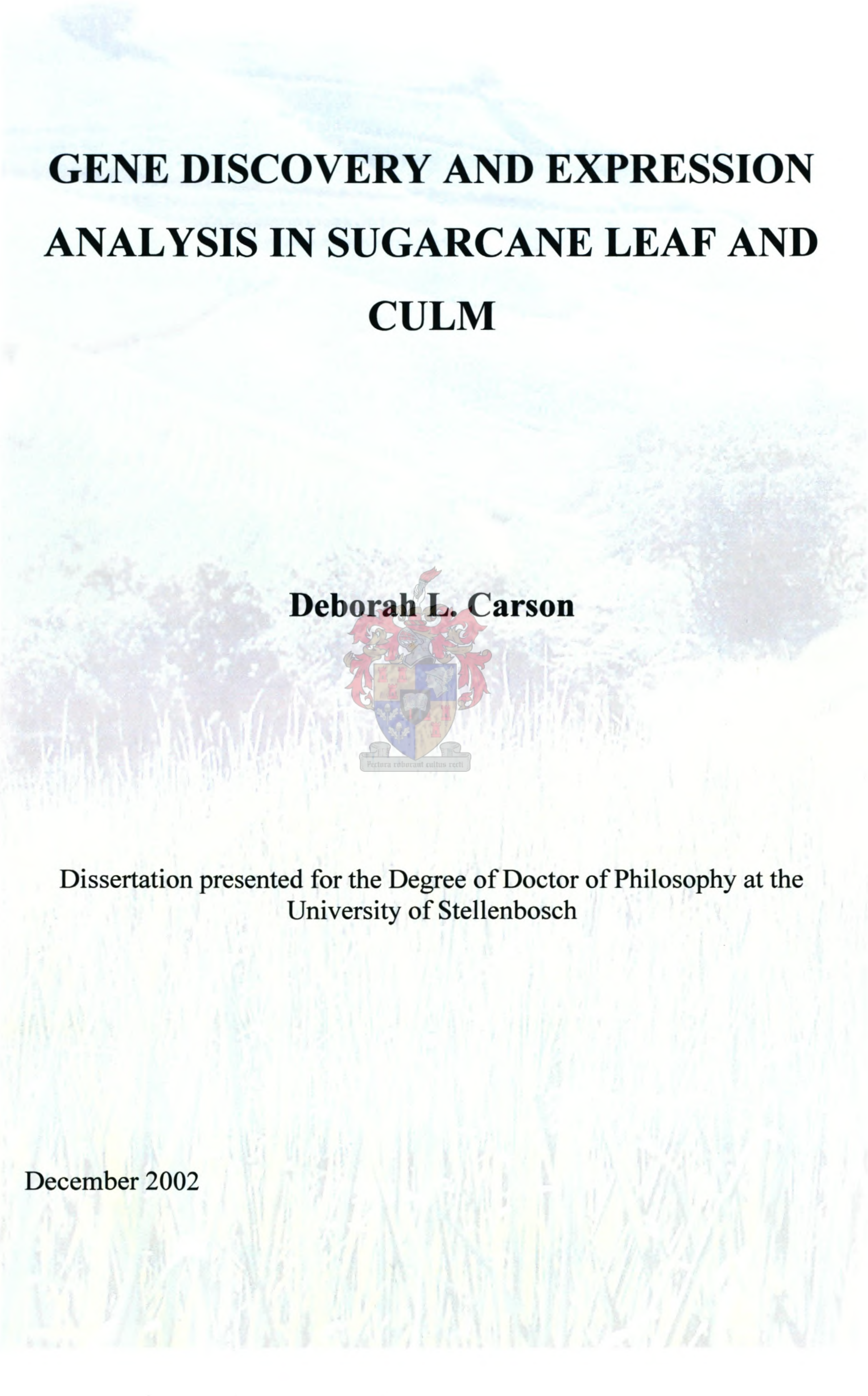


GENE DISCOVERY AND EXPRESSION ANALYSIS IN SUGARCANE LEAF AND CULM

Deborah L. Carson

Dissertation presented for the Degree of Doctor of Philosophy at the
University of Stellenbosch

December 2002

DECLARATION

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and that I have not previously in its entirety or in part submitted it at any other university for a degree.

PREFACE

The research described in this thesis was conducted in the Biotechnology Department at the South African Sugar Association Experiment Station (SASEX), Mount Edgecombe, under the supervision of Professor FC Botha (Institute for Plant Biotechnology, University of Stellenbosch) and Dr BI Huckett (Biotechnology Department, SASEX).

ABSTRACT

Sugarcane (*Saccharum* spp. hybrids) is a commercial crop plant capable of storing up to 20% sucrose on a fresh mass basis in the culm. Knowledge about gene expression during sugarcane growth and maturation is limited. The aim of this study was to assess whether an Expressed Sequence Tag (EST)-based approach towards analysis of sugarcane would reveal new information about gene expression and metabolic processes associated with sugarcane growth and development. The specific objectives were two-fold: firstly, to develop an EST database for sugarcane and secondly, to identify and analyse genes that are expressed in different sugarcane tissue types and developmental stages, with a specific focus on leaf and culm.

An EST database for sugarcane was initiated to obtain information on sugarcane gene sequences. A total cDNA library was constructed from sugarcane immature leaf (leaf roll: meristematic region) tissue and 250 clones randomly selected and subjected to single-pass DNA sequence analysis. Sugarcane ESTs were identified by sequence similarity searches against gene sequences in international databases. Of the 250 leaf roll clones, 26% exhibited similarity to known plant genes, 50% to non-plant genes while 24% represented new gene sequences. Analysis of the identified clones indicated sequence similarity to a broad diversity of genes. A significant proportion of genes identified in the leaf roll were involved in processes related to protein synthesis and protein modification, as would be expected in meristematic tissues. Submission of 495 sugarcane gene sequences to the dbEST database represented the first sugarcane ESTs released into the public domain.

Two subtracted cDNA libraries were constructed by reciprocal subtractive hybridisation between sugarcane immature and maturing internodal tissue. To explore gene expression during sugarcane culm maturation, partial sequence analysis of random clones from maturing culm total and subtracted cDNA libraries was performed. Database comparisons revealed that of the 337 cDNA sequences analysed, 167 showed sequence homology to gene products in the protein databases while 111 matched uncharacterised plant ESTs only. The remaining cDNAs showed no database match and could represent novel genes. The majority of ESTs corresponded to a variety of genes associated with general cellular metabolism. ESTs homologous to various stress response genes were also well represented. Analysis of ESTs from the subtracted library identified genes that may be preferentially expressed during culm maturation.

The expression patterns of sugarcane genes were examined in different tissue sources and developmental stages to identify differentially expressed genes. cDNA arrays containing 1000 random clones from immature leaf and maturing culm cDNA libraries were hybridised with poly $(A)^+$ RNA from immature leaf, mature leaf, immature culm and maturing culm. All cDNAs examined hybridised to all four probes, but differences in signal intensity were observed for individual cDNAs between hybridisation events. No cDNAs displaying tissue- or developmental-stage specific expression were detected. Comparisons between hybridisation patterns identified 61 cDNAs that were more abundantly expressed in immature and mature leaf than the culm. Likewise, 25 cDNAs preferentially expressed in immature and maturing culm were detected. ESTs established for the differentially expressed cDNAs revealed sequence homology to a diverse collection of genes in both the leaf and the culm. These included genes associated with general cellular metabolism, transport, regulation and a variety of stress responses. None of the differentially expressed genes identified in the culm were homologous to genes known to be associated with sucrose accumulation.

To examine differences at the level of gene transcription between low sucrose-accumulating and high sucrose-accumulating tissues, subtracted cDNA libraries were utilised. To isolate cDNAs differentially expressed during culm maturation, cDNA arrays containing 400 random clones (200 from each library) were screened with total cDNA probes prepared from immature and maturing culm poly $(A)^+$ RNA. Results indicated that 36% and 30% of the total number of cDNAs analysed were preferentially expressed in the immature and maturing culm, respectively. Northern analysis of selected clones confirmed culm developmental stage-preferential expression for most of the clones tested. ESTs generated for the 132 differentially expressed clones isolated exhibited homology to genes associated with cell wall metabolism, carbohydrate metabolism, stress responses and regulation, where the specific ESTs identified in the immature and maturing culm were distinct from each other. No developmentally regulated ESTs directly associated with sucrose metabolism were detected.

These results suggest that growth and maturation of the sugarcane culm is associated with the expression of genes for a wide variety of metabolic processes. In addition, genes encoding enzymes directly involved with sucrose accumulation do not appear to be abundantly expressed in the culm.

OPSOMMING

Kommersiële suikerriet variëteite (*Saccharum* spp. hibriede) is in staat om tot 20% sukrose op 'n vars massa basis in die stingel op te berg. Kennis oor geenuitdrukking tydens groei en rypwording is beperk. Die doel van die huidige studie was om vas te stel of 'n grootskaalse karatersisering van die geenvolgordes wat uitgedruk word "Expressed Sequence Tag (EST)-based approach" tot nuwe inligting aangaande die aard en omvang van metabolisme tydens groei en ontwikkeling van suikerriet sal lei. 'n Tweeledige benadering is in hierdie studie gevolg. Eerstens is 'n data basis oor die gene wat uitgedruk word "EST" databasis opgestel. Tweedens is gene geïdentifiseer en gekarakteriseer wat spesifiek op verskillende stadiums van ontwikkeling en in spesifiek weefsel uitgedruk word.

Vir die opstel van die EST-databasis is 250 klone uit 'n totale cDNA biblioteek vanaf RNA uit suikerrietblaarweefsel (blaarrol:meristematiese streek) op 'n lukraak basis gekies en aan 'n enkel eenrigting DNA volgorde analise onderwerp. Suikerrriet EST's is geïdentifiseer deur middel van homologie soektogte teen geenvolgordes in internasionale databasisse. Uit die 250 blaarrol klone het 26% ooreenkomste met bekende plant gene en, 50% met nie-plant gene getoon. Ongeveer 24% het nuwe geenvolgordes verteenwoordig. Analise van die geïdentifeseerde klone het ooreenkomste met 'n breë diversiteit van gene getoon. 'n Betekenisvolle gedeelte van gene wat in die blaarrol geïdentifiseer is, is by proteïensintese en proteïenmodifikasies betrokke. Dit is in ooreenstemming met wat van meristematiese weefsel verwag kan word. Die 495 suikerriet geenvolgordes wat in die internasionale dbEST databasis gestort is, is die eerste sodanige inligting in die publieke domein.

Twee spesifieke cDNA biblioteke (subtraction libraries) wat volgordes spesifiek aan onvolwasse suikerriet en rypwordende internodale weefsel bevat is voorberei. Geenuitdrukking gedurende die rypwordingsproses van die suikerrietstingel is bestudeer deur geenvolgorde analises van onwillekeurige geselekteerde klone van die twee cDNA biblioteke te doen. Van die 337 geenvolgordes wat geanaliseer is het 167 homologie met bekende gene en net 111 ooreenkomste met ongekarakteriseerde plant gene getoon. Die oorblywende geenvolgordes het geen ooreenkomste met bekende gene getoon nie en daar kan dus aanvaar word dat hulle nuwe gene verteenwoordig. Die meerderheid ESTs het ooreenkomste met verskeie gene wat met sellulêre metabolisme geassosieer word getoon. ESTs wat homoloog was aan verskeie spannings geassosieerde gene was ook goed verteenwoordig. Die analise het gene wat by voorkeur tydens stringelrypwording uitgedruk word geidentifiseer.

Die geenuitdrukkingspatrone van suikerriet in weefsels van verskillende oorsprong en ontwikkelingstadia is ondersoek om differensieel uitgedrukte gene te identifiseer. Reekse wat 1000 lukrake cDNA klone van onvolwasse en rypwordende stingel cDNA biblioteke is met poli-(A)-RNA van onvolwasse blaar, volwasse blaar, onvolwasse stingel en volwasse stingel gehibridiseer. Al die cDNA klone wat ondersoek is het met al vier die peilers gehibridiseer. Die intensiteit van die seine het egter grootliks gevarieer. Die analise het gelei tot die identifisering van 61 cDNA klone wat teen hoër vlakke in onvolwasse en volwasse blaar as in die stingel uitgedruk word. Daar is ook 25 cDNA klone wat by voorkeur in onvolwasse en rypwordende stingel uitgedruk word gevind. Gene wat geassosieer word met gewone sel metabolisme, vervoer prosesse, regulering en verskeie spannings-geassosieerde reaksies, is in die twee groepe teenwoordig. Geeneen van die volgordes wat selektief uitgedruk word kan met gene wat direk met sukrose akkumulering verband hou geassosieer word nie.

Ten einde cDNA klone wat differensieel tydens rypwording van die stingel uitgedruk word te isoleer, is 400 cDNA klone (200 van elke biblioteek) lukraak geselekteer en met totale cDNA peilers, wat uit onvolwasse en rypwordende stingel poli-(A)-RNA voorberei is, gesif. Resultate het aangetoon dat 36% en 30% van die totale getal cDNA klonewat geanaliseer is, by voorkeur in die onvolwasse en rypwordende stingel uitgedruk word. RNA kladanalises van geselekteerde klone het getoon dat die meeste ontwikkelingstadium spesifieke uirtdrukkingspatrone het. Daar is gevind dat 132 van die EST klone homologie met gene geassosieerd met selwand- en koolhidraatmetabolisme, spannings geassosieerde- en reguleringsreaksies, toon. Die spesifieke ESTs wat in die onvolwasse en rypwordende stingel geïdentifiseer is het van mekaar verskil. Nie een van die ESTs wat geïdentifiseer is kan direk met sukrose metabolisme geassosieer word nie.

Hierdie werk toon baie duidelik aan dat groei en rypwording van die suikerrietstingel met die uitdrukking van gene geassosieerd is wat by ‘n hele aantal metaboliese prosesse betrokke is. Die resultate toon ook dat die gene wat vir ensieme kodeer wat direk by sukrose akkumulering betrokke is, nie teen hoë vlakke in die stingel uitgedruk word nie.

ACKNOWLEDGEMENTS

A word of thanks are extended to the following:

My supervisor, Prof Frikkie Botha, for first introducing me to the concept of expressed sequence tags. His critical input during manuscript development and thesis preparation is gratefully acknowledged.

My co-supervisor, Dr Barbara Huckett for advice and unwavering support throughout the study.

Dr Barbara Huckett in her capacity as Head of Department (Biotechnology, SASEX) for allowing me to prepare this thesis during working hours. I have always regarded this as a privilege.

Dr Peter Hewitt, Director of SASEX, for his support of Biotechnology research and for encouraging an environment whereby employees are able to further their studies whilst contributing towards a focussed research programme of work.

Avril Harvey and Natalie Williams, for technical assistance with DNA sequence analysis.

Clemson University Genomics Institute, USA, for the production of the cDNA macroarray filters.

Dr Derek Watt for advice and encouragement.

My fellow colleagues in the Biotechnology department, especially for their consideration and patience.

My husband, David, for keeping things going, through both the good times and the bad times.

TABLE OF CONTENTS

Page

DECLARATION .... **ii**

PREFACE .... **iii**

ABSTRACT .... **iv**

OPSOMMING .... **vii**

ACKNOWLEDGEMENTS .... **x**

TABLE OF CONTENTS .... **xi**

CHAPTER 1 : INTRODUCTION **1**

CHAPTER 2 : LITERATURE REVIEW **5**

2.1 A NEW PARADIGM FOR PLANT SCIENCE 5
2.2 IDENTIFYING PLANT GENES 6
2.2.1 Expressed Sequence Tags (ESTs) as a basic tool for gene discovery 6
2.2.2 Access to gene sequence data 8
2.2.3 Advantages and limitations of ESTs for gene discovery 8
2.3 TOWARDS ASSIGNING FUNCTIONS TO PLANT GENES 10
2.3.1 From structural genomics to functional genomics 10
2.3.2 Strategies for gene expression profiling 12
2.3.3 ESTs as a qualitative indicator of plant gene expression 14
2.3.3.1 Analysing diverse plant processes 14
2.3.3.2 Developing global gene expression profiles 16
2.3.4 Transcript profiling and screening for differentially expressed genes 18
2.3.4.1 Organs, tissues and developmental stages 18
2.3.4.2 Metabolic pathways 20
2.3.4.3 Stress-inducible genes 22
2.4 STATUS OF SUGARCANE GENOMICS 24
2.4.1 Metabolism in the sugarcane culm 24
2.4.2 Single-gene studies 25
2.4.3 Current research 26
2.5 APPLICATION FOR CROP IMPROVEMENT 27
2.6 REFERENCES 28

CHAPTER 3 : PRELIMINARY ANALYSIS OF EXPRESSED SEQUENCE TAGS FOR SUGARCANE
[Crop Science 40(6) : 1769-1779 (2000)] **37**

3.1 ABSTRACT 37
3.2 INTRODUCTION 38
3.3 MATERIALS AND METHODS 39
3.3.1 Total and poly (A^+) RNA Isolation 39
3.3.2 Construction of a leaf roll cDNA library 40

3.3.2.1 cDNA synthesis 40
3.3.2.2 Ligation to amplification adaptors 41
3.3.2.3 PCR amplification of cDNA 41
3.3.2.4 Library construction 41
3.3.3 Template preparation 42
3.3.4 Sequencing 42
3.3.5 Sequence data analysis 43
3.4 RESULTS 43
3.4.1 Characteristics of the constructed LR cDNA library 43
3.4.2 Generation of Expressed Sequence Tags 44
3.4.3 Sequencing template 44
3.4.4 Identification of genes 45
3.4.5 Functional identification of sugarcane ESTs 54
3.5 DISCUSSION 56
3.6 REFERENCES 60

CHAPTER 4 : GENES EXPRESSED IN SUGARCANE MATURING INTERNODAL TISSUE
[Plant Cell Reports 20(11) : 1075-1081 (2002)] 63

4.1 ABSTRACT 63
4.2 INTRODUCTION 63
4.3 MATERIALS AND METHODS 65
4.3.1 cDNA library preparation 65
4.3.2 Template preparation 65
4.3.3 DNA sequencing and sequence data analysis 66
4.4 RESULTS AND DISCUSSION 67
4.4.1 Generation of sugarcane ESTs 67
4.4.2 Putatively identified genes in the maturing culm 68
4.4.2.1 Characteristics of the database matches 68
4.4.2.2 Sugarcane ESTs homologous to known gene products 72
4.4.2.3 Sugarcane EST matches to uncharacterised coding sequences 74
4.5 REFERENCES 75

CHAPTER 5 : DIFFERENTIAL GENE EXPRESSION IN SUGARCANE LEAF AND INTERNODAL TISSUES OF VARYING MATURITY
[Manuscript : submitted to South African Journal of Botany] 79

5.1 ABSTRACT 79
5.2 INTRODUCTION 79
5.3 MATERIALS AND METHODS 81
5.3.1 Source of cDNA clones 81
5.3.2 Bacterial clones, production of cDNA arrays and Northern blots 81
5.3.3 RNA extraction and probe synthesis 82
5.3.4 Hybridisation procedures 83
5.3.5 DNA sequencing and sequence data analysis 84
5.4 RESULTS 85
5.4.1 Patterns of transcript abundance 85
5.4.2 Differentially expressed genes in leaf and culm 86
5.4.3 Northern blot analysis 91

5.5 DISCUSSION 92
5.6 REFERENCES 95

CHAPTER 6 : SUGARCANE ESTs DIFFERENTIALLY EXPRESSED IN IMMATURE AND MATURING INTERNODAL TISSUE [Plant Science 162(2) : 289-300 (2002)] 99

6.1 ABSTRACT 99
6.2 INTRODUCTION 100
6.3 MATERIALS AND METHODS 101
6.3.1 Construction of subtracted cDNA libraries 101
6.3.2 Preparation of bacterial clones 102
6.3.3 Production of cDNA macroarrays and Northern blots 102
6.3.4 RNA extraction and probe synthesis 103
6.3.5 Hybridisation procedures 104
6.3.6 DNA sequencing 105
6.3.7 Sequence data analysis 105
6.4 RESULTS 105
6.4.1 Characterisation of the subtracted cDNA libraries 105
6.4.2 Detection of differentially expressed ESTs by cDNA macroarray hybridisation analysis 107
6.4.3 Putative functions of differentially expressed ESTs 113
6.4.4 Confirmation of culm-specific and culm-preferential expression by Northern analysis 114
6.5 DISCUSSION 117
6.6 REFERENCES 120

CHAPTER 7 : CONCLUDING REMARKS 124

PRESENTATIONS AND PUBLICATIONS ARISING FROM THE STUDY .. 127

CURRICULUM VITAE 129

CHAPTER 1

INTRODUCTION

Sugarcane (*Saccharum* spp. hybrids) is a highly productive commercial crop plant capable of storing high concentrations of sucrose in the culm (Moore 1995). Continual improvement of the yield of the crop's primary commodity, sucrose, is the priority of commercial sugarcane industries worldwide. To date, sugarcane improvement has relied solely on traditional breeding that requires no specific knowledge of the physiological or genetic basis of plant performance. However, recent advances in technologies to study higher plants at the genetic level coupled with the ability to genetically modify plants through the introduction of cloned genes have created the opportunity for a more informed breeding selection and the potential to improve important cultivars.

Improving plant performance through genetic manipulation is dependent on an integrated understanding of plant metabolism at a physiological, biochemical and genetic level. Furthermore, the availability of resources such as suitable genes and gene promoter and targeting sequences to direct the expression of cloned genes to specific cells or tissue types is also important. For sugarcane, research directed towards pinpointing important elements associated with the metabolic control of growth, maturation and sucrose accumulation will yield information that can be exploited in genetic engineering efforts to improve crop performance and sucrose yield.

During sugarcane growth, a gradient of maturation and sucrose accumulation is evident down the culm so that high sucrose concentrations are reached in the mature internodes. Since the 1960s, the pathway of sucrose accumulation in sugarcane has been studied extensively at the physiological and biochemical level (for review, see Moore 1995; Grof and Campbell 2001). Although this research has generated a considerable amount of information about sucrose accumulation during culm maturation, the regulation of sucrose metabolism is still not well understood. Studies have indicated that sucrose cycling and compartmentalisation might be important factors for the control of sucrose accumulation.

The mechanisms controlling the partitioning of carbon during sugarcane culm maturation, however, are complex (Whittaker and Botha, 1997). Characterisation of enzymes known to be associated with sucrose metabolism in other plants have identified several key enzymes that synthesise and break down sucrose in sugarcane (Moore 1995). These include invertases, sucrose-phosphate synthase and sucrose synthase. Biochemical studies have examined how the activities of these enzymes correlate with the rate of sucrose accumulation and sucrose content in maturing internodes (Zhu et al., 1997; Lingle et al., 1999). However, the role of these enzymes in regulating sucrose accumulation during culm maturation is still obscure. Furthermore, control at the level of transcription and translation for the genes encoding these enzymes is not well understood for sugarcane. Much more information is therefore required at the gene level about sugarcane metabolism during culm maturation and sucrose accumulation.

Some molecular studies of individual genes encoding enzymes involved in sucrose metabolism have been performed for sugarcane including sucrose phosphate synthase (SPS) (Sugiharto et al., 1997), soluble acid invertase (SAI) (Zhu et al., 2000) and sucrose synthase (SuSy) (Lingle and Dyer, 2001). However, there is a fundamental gap in knowledge about gene expression during sugarcane growth and maturation. The aim of the current study is thus to identify and analyse genes that are expressed in different sugarcane tissue types and developmental states, with a specific focus on leaf and culm.

Recently, new molecular approaches have become available to obtain the entire genome sequence of organisms and to perform genome-wide studies of gene expression. Most of the advances in this regard have been initiated through research on humans and other animals and have been heavily dependent on technological developments for obtaining and analysing large amounts of nucleotide sequence information. The tools developed for genome analysis in non-plant organisms have subsequently been applied to study higher plants with large co-ordinated projects for model systems such as *Arabidopsis* and rice at the forefront of plant genomics research. Complete genome sequencing has been achieved thus far for *Arabidopsis* (Arabidopsis Genome Initiative 2000) with a draft sequence of the rice genome being released recently (Barry 2001). Large amounts of nucleotide sequence data for the expressed portions of plant genomes are now also available for many plant species. Efforts to characterise the functions of these many genes in plant metabolism through analysis of transcript levels are in progress. These studies have adopted the notion that related sets of genes are best studied in parallel (Ewing et al., 1999). In this way, the expression patterns of multiple genes are analysed simultaneously thus allowing comparisons to be made between the expression profiles of genes in different cells, tissue types or developmental states and to identify differentially expressed genes. Although these studies are still in their infancy for plants, they have provided valuable information about the expression behaviour of genes associated with important metabolic processes.

The purpose of the current investigation was fivefold:

Firstly, to critically review progress towards improving the knowledge of higher plant metabolism at the gene level (Chapter 2). Based on the information gained from other plants it is evident that molecular approaches offer great potential to facilitate sugarcane research through the identification and analysis of genes.

Secondly, using the meristematic apex as a model, parameters for identifying sugarcane genes by partial DNA sequence analysis of expressed sequences to generate Expressed Sequence Tags (ESTs) were developed (Chapter 3). As only limited gene sequence information is available in the public domain for sugarcane, a database of sugarcane genes was created.

Thirdly, due to the limited information about gene expression in maturing sugarcane tissue where a high sucrose accumulation rate is evident, ESTs associated with this tissue and type of metabolism were identified. The analysis of these ESTs is presented in Chapter 4. The types of genes abundantly expressed in a sucrose-accumulating tissue are discussed.

Fourthly, to provide information about the expression patterns of sugarcane genes, a comparison between transcript abundance of sugarcane genes in immature and maturing leaf and culm was made. Differentially expressed genes that may be up- or down-regulated during culm development are described in Chapter 5 and new insights into the complexity of sugarcane culm metabolism are provided.

Fifthly, as very little is known about the genetic basis of sugarcane culm maturation, efforts were made to specifically target differences at the level of gene transcription between low sucrose-accumulating tissues and high sucrose-accumulating tissues. Genes differentially expressed in immature and maturing culm tissue are presented in Chapter 6. In this final chapter, candidate developmental-stage preferential genes suitable for further molecular analyses of sugarcane culm development are discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 A NEW PARADIGM FOR PLANT SCIENCE

Physiological and biochemical studies of plant metabolism dominated plant science research in the 20th century and revealed an intricate network of metabolic processes associated with the growth and development of plants and their responses to the external environment. Although the value of this research in improving our understanding of plant performance cannot be overestimated, the overall advances do not describe how plant processes are regulated at the genetic level. Plant growth and development is highly dependent on specific patterns of gene expression. The dynamic changes in the expression of genes during distinct developmental stages or in response to growth conditions are responsible for differences in morphological and other phenotypic traits. Insight into the genetic basis of plant growth and development could have important consequences for efforts to improve plant performance through a better understanding of the relationship between the expression of genes and valuable traits (Miflin 2000).

The development of molecular biology tools to study plant genes has resulted in a large volume of literature describing the spatial and temporal regulation of gene expression. Traditionally, these studies have been conducted by a stepwise analysis of single genes and have been reliant upon prior knowledge of the specific genes or gene products involved in a particular process. Recently, however, technological advances have resulted in the capacity to generate large amounts of gene sequence information and to simultaneously analyse the expression behaviour of multiple genes without prior knowledge of the genome (Bouchez and Höfte 1998). This has lead to a paradigm shift in the way plant metabolism is studied. It is now possible to examine whole complements of genes and to monitor how their expression patterns change during growth and differentiation or in response to various abiotic and biotic stresses. This approach, commonly referred to as "genomics", offers new opportunities for investigating how complex plant metabolic processes are regulated at the gene level and has the potential to supplement the understanding of gene function provided by traditional approaches by allowing a more complete view of biological systems (Lockhart and Winzeler 2000). The full benefit of genomics research will only be realised, however, once genetic and genomic data are integrated with biochemical, physiological and morphological data.

The purpose of this review is to provide an overview of how genomics research has contributed towards the understanding of plant metabolism at the genetic level. The various strategies adopted for the acquisition of gene sequence data and the analysis of plant gene expression are presented with particular reference to the outcomes of recent gene profiling research. Current knowledge of gene expression in sugarcane, an important crop plant, is described and prospects for furthering the understanding of sugarcane metabolism at the genetic level are mentioned. Finally, ways in which the knowledge gained from studying the genetic basis of plant performance could facilitate efforts to improve the productivity of crop plants are portrayed.

2.2 IDENTIFYING PLANT GENES

2.2.1 Expressed Sequence Tags (ESTs) as a basic tool for gene discovery

The development of high-throughput DNA sequencing techniques has made possible the sequencing of entire genomes. The complete genome sequence of an organism is a valuable resource of data providing access to the entire set of genes as well as information on the structure of the genome and the relative order of genes on chromosomes (Bouchez and Höfte 1998). Genomic sequencing projects have been undertaken for several organisms, the most well-known example being the human genome project with the highly publicised release of the complete human genome sequence in 2001 (International Human Genome Sequencing Consortium 2001). In addition, more than 30 other organisms have had their genomes completely sequenced, with approximately another 100 in progress (Lockhart and Winzeler 2000). Plant biologists have capitalised on technological advances made during the sequencing of non-plant genomes with the result that the first complete sequence of the model dicotyledonous plant *Arabidopsis thaliana* was released in 2000 (Arabidopsis Genome Initiative 2000). Sequencing efforts are underway to determine the complete sequence of rice (*Oryza sativa*), the model plant for monocotyledonous species (Yuan et al., 2001). Although the entire genome sequence is the ultimate blueprint of an organism, total sequencing is unlikely for the vast majority of plant and animal species due to factors such as individual genome complexity and practical constraints such as resource availability.

As the coding sequences of genes account for the bulk of the information content of a genome, but only a small proportion of the DNA (Adams et al., 1991), sequencing the expressed portions of the genome presents a more rapid route to gene discovery. In this approach, expressed genes are captured as cDNAs and partial sequences obtained from one or both ends of random anonymous cDNA clones (Bouchez and Höfte 1998). These sequences, usually between 300 and 400 nucleotides in length, are compared to nucleotide or deduced amino acid sequences of genes of known function and on the basis of sequence alignment allow a probable identity to be assigned to the cDNA.

Random cDNA sequencing was first reported in the 1980s (Milner and Sutcliffe 1983) with the term "Expressed Sequence Tags" (ESTs) being popularised in the early 1990s when large-scale partial sequence analysis became possible through the development of high-throughput automated sequenators (Adams et al., 1991; 1992). EST analyses were initially performed for humans (Adams et al., 1991; 1992), mouse (Höög 1991) and the nematode worm *Caenorhabditis elegans* (Waterston et al., 1992), but have since been applied to many other organisms after the ease and rapidity of generating gene sequence data was recognised. EST analysis has been used extensively to identify genes in plants. The earliest published plant EST data sets were from rice (Uchimiya et al., 1992), maize (Keith et al., 1993), *Brassica napus* (Park et al., 1993) and *Arabidopsis* (Höfte et al., 1993). Subsequent to these key papers there have been further reports of the identification of ESTs in those same plants (Newman et al., 1994; Sasaki et al., 1994; van de Loo et al., 1995; Liu et al., 1995; Cooke et al., 1996; Lee et al., 1998; White et al., 2000) and other plant species including, *Lotus japonicus* (Asamizu et al., 2000), *Cryptomeria japonica* (Ujino-Ihara et al., 2000), grape (Ablett et al., 2000) and watermelon (Ok et al., 2000). Presently, however, far more plant EST sequences are available in electronic public databases than are reported in the published literature. These include ESTs from model plants such as *Arabidopsis* as well as important crop plants such as rice, tomato and maize.

2.2.2 Access to gene sequence data

Concurrent growth in information technology and the formation of the World Wide Web (WWW) have automated the search and retrieval of nucleotide sequences, facilitating gene discovery through sequence comparison between anonymous gene sequences and the escalating reservoir of genes with a known identity. Gene sequence information can be accessed through the GenBank genetic sequence database that contains an annotated collection of all publicly available DNA sequences. At present, GenBank lists gene sequences from more than 75 000 organisms. The rapid accumulation of partial cDNA sequences lead to the creation of the public database dbEST, a division of GenBank, that has streamlined access to different EST collections (Boguski et al., 1993). These databases are an immense resource of gene sequence information. A recent release of dbEST (release 112301, November 23, 2001) listed 9 642 172 ESTs from 339 different organisms and included 109 higher plant species. Public access to sequence data from large-scale EST sequencing projects currently in progress for many important crop plants including *Glycine max*, *Hordeum vulgare*, *Lycopersicon esculentum*, *Oryza sativa*, *Sorghum bicolor, Triticum aestivum* and *Zea mays* is also possible through the databases.

The access to nucleotide sequence data through these public databases has made the process of gene identification an effortless task and is a resource available to any laboratory with electronic mailing facilities. It is worth noting, however, that for many plants particularly crops, large numbers of ESTs are available only in private databases. Several proprietary EST databases exist for commercial crops for which gene discovery is being exploited for plant improvement.

2.2.3 Advantages and limitations of ESTs for gene discovery

Using ESTs for gene discovery provide a more efficient alternative to methods such as traditional library screening and polymerase chain reaction (PCR)-based approaches which are often unsuccessful in detecting related genes due to deviations in sequences at the nucleotide level (McCarter et al., 2000). Other strategies for isolating genes such as chromosome walking and gene tagging are usually expensive and time consuming and can only result in limited numbers of genes (Park et al., 1993). By contrast, EST sequencing is simple to perform and with the availability of automated sequenators capable of high-throughput, large amounts of sequence data can be rapidly generated with a limited investment in both time and money. The real value of the EST approach to gene discovery lies therefore in the high frequency with which a putative identity can be assigned to an anonymous cDNA by deduced amino acid sequence comparison with a database of known gene products (Newman et al., 1994). Moreover, ESTs may be generated without prior knowledge of the genome and independently of genome complexity.

A further advantage of EST sequencing is that it connects plant biology to non-plant biology (Newman et al., 1994). Non-plant programmes have been producing sequence data for genes and gene products of known function for longer than plant programmes and in larger numbers. Sequence overlap between plant and non-plant genes frequently allows a putative identity to be assigned to the plant gene. In an early study of ESTs from roots of *Brassica napus,* four ESTs not identified previously in plants were shown to be homologous to genes from a variety of organisms including mouse and bacteria (Park et al., 1993). Similarly, in developing castor seeds, 7% of ESTs analysed could only be identified by homology to non-plant genes (van de Loo et al., 1995). More recently, 9.3% of leaf ESTs from *Brassica napus* (Lee et al., 1998) and 11% of grape leaf and berry ESTs (Ablett et al., 2000) matched genes previously identified in non-plant organisms. EST sequencing presents a useful strategy towards the discovery of genes that have not been identified previously in plants.

Although ESTs are a rich source of tags to genes, certain limitations to their use need to be considered. ESTs are by definition incomplete sequences and are known to contain approximately 3% errors and base ambiguities (Boguski et al., 1993). Although these errors may not significantly influence the probability that the EST can be putatively identified by sequence similarity searches, they can be problematic for further applications such as PCR primer design (Newman et al., 1994). Furthermore, it has been reported that few of the descriptions of genes associated with GenBank public database entries have been validated by wet-lab experiments (Boguski 1999). In some cases, sequences lodged in the databases have been misclassified with respect to function (Newman et al., 1994). The interpretation of the degree of sequence similarity between the EST and the database sequence is also critical as the significance of the alignment is dependent on various factors including the length of the EST, the length of the database sequence and the region of overlap (Newman et al., 1994). As a result, homology searches can only assign a putative identity to an EST that should be validated by other criteria.

2.3 TOWARDS ASSIGNING FUNCTIONS TO PLANT GENES

2.3.1 From structural genomics to functional genomics

The term "structural genomics" pertains to the sequencing of entire genomes, of which sequencing and identifying genes through EST analysis is an integral part. However, the large DNA sequence data sets generated by EST programmes are essentially collections of digital information. These gene sequences alone can provide limited information only about gene function as they are based solely on deductions from electronic sequence homology searches using complex algorithms.

Although putative functions have been assigned to many genes on the basis of homology, it is possible that they may have different roles (Willmann 2001). In *Arabidopsis* for example, seven distinct Cyt P450 sequences were isolated by sequence overlap to previously identified genes from other species but it is likely that each of the corresponding proteins catalyses a different enzymatic reaction in the plant (Newman et al., 1994). In addition, putative functions based on sequence homology alone can be misleading. For instance, it was demonstrated by transposon tagging that clones showing similarity to the chalcone synthase gene were actually a component of the long fatty acid elongase complex (James et al., 1995). Consequently, the wealth of gene sequence information being generated far exceeds that which is understood about the biological roles of genes in cellular metabolism. Assigning functions to genes based on experimental evidence is the fundamental principle behind the functional genomics approach to genome analysis. Hieter and Boguski have defined functional genomics as referring to "the development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics" (Hieter and Boguski 1997). Functional genomics is characterised by a shift from the study of single genes or gene products to the systematic analysis of large numbers of genes in parallel. The fundamental goal is to understand how the genome works through the control and regulation of the expression of genes (Lockhart and Winzeler 2000).

Establishing the function of a gene can be approached in several different ways. Reducing or enhancing the expression of a gene can often lead to an observable phenotypic effect (Maheshwari et al., 2001). Populations of mutant plants can be created using a series of random gene knockouts. These plants may then be screened for any detectable change in phenotype that may be traced back to the particular DNA sequence that was mutated (Miflin 2000). The use of transposable elements to create loss-of-function insertion mutants and subsequent analysis has been applied to several plant species including *Arabidopsis* and maize (Bouchez and Höfte 1998; Martienssen 1998; Somerville and Somerville 1999). However, the success of this approach relies on the production and availability of large collections of plants mutagenised by an insertion element. In addition, it cannot detect when a change in phenotype is linked to the interactions between the mutated gene and other genes (Miflin 2000). For many plant species, especially those with complex polyploid genomes such as cereals and grasses, it is extremely difficult to link a gene function to a particular allele. The function of a gene can also be investigated, however, through the analysis of its expression behaviour. Profiling the abundance of RNA transcripts in specific organs, tissues or cells, or in response to different environmental conditions and stresses provides information about the expression patterns of genes. Determining the expression levels of genes allows deductions to be made about their functions. Furthermore, comparison between expression profiles allows the identification of differentially expressed transcripts that may have specific metabolic or morphogenetic functions (Kuhn 2001). For example, expression profiling has been used to characterise the role of a novel alcohol acyltransferase gene during flavour biogenesis in strawberry (Aharoni et al., 2000). Currently, only about 50% of newly obtained nucleotide sequences obtained through genome and EST sequencing programmes exhibit similarity to previously identified genes of known identity (Bouchez and Höfte 1998). Gene expression profiling will allow predictions of function to be made for those transcripts with no database sequence match as similarities in expression behaviour between sequences of unknown function and known genes may indicate a functional homology (Kuhn 2001).

2.3.2 Strategies for gene expression profiling

The expression behaviour of genes may be monitored either through measuring the expression levels of specific genes, characterising global patterns of expression, or screening for significant differences in mRNA abundance. The development of techniques for genome-wide expression profile analyses are becoming powerful tools for functional genomics research (Breyne and Zabeau 2001). Gene expression profiling examines the expression of genes in a multiparallel fashion and in many cases, is dependent on sophisticated high-throughput or large-scale experimental techniques (Hieter and Boguski 1997).

The following is a brief description of the principles behind the most common strategies currently being used to monitor gene expression and to identify differentially expressed genes in plants. The technical aspects of many of these procedures, as well as their advantages and disadvantages, have been reviewed previously (Schena et al., 1998; Baldwin et al., 1999; Kozian and Kirschbaum 1999; Lockhart and Winzeler 2000; Breyne and Zabeau 2001; Kuhn 2001).

Gene expression profiles can be generated using sequence-based, PCR-based and hybridisation-based methods. The most simplest form of the sequence-based methods is that of EST analysis where the abundance of a specific cDNA in an EST collection is used as a measure of gene expression (Audic and Claverie 1997). In addition, comparative analysis of different EST data sets from particular organs or cell types can be used to identify differentially expressed genes (Kozian and Kirschbaum 1999). The serial analysis of gene expression (SAGE) allows the simultaneous analysis of sequences that derive from different cell or tissue types and is based on counting expressed sequence tags of 14-15 bases from cDNA libraries (Velculescu et al., 1995). The frequency of genes found in SAGE correlate with their abundance in the corresponding cDNA libraries. This method is restricted, however, to organisms for which large EST databases are available and there is only one report in the literature of the application of SAGE to study gene expression in plants (Matsumura et al., 1999). PCR-based methods for transcript profiling include differential display (Liang and Pardee 1992) and amplified restriction fragment length polymorphism (AFLP) of cDNA (Bachem et al., 1996). Differential display uses sets of random primers (with or without anchors) to amplify by PCR portions of cDNA which are then size-fractionated using high resolution polyacrylamide sequencing gels. Comparisons between banding profiles generated from different cDNA sets are used to isolate differentially expressed cDNAs which can then be sequenced and assigned putative identities through gene sequence homology searches. The cDNA-AFLP approach uses restriction enzymes to cut the cDNA population into fragments which are then ligated onto adapters and amplified using selective PCR primers. Although both differential display and cDNA-AFLP have been used successfully to identify differentially expressed genes in plants, the latter is becoming more popular as it is more sensitive and versatile (Breyne and Zabeau 2001).

One of the most traditional methods for isolating differentially expressed genes is the differential screening of cDNA libraries. Hybridisation between complex sets of labeled cDNAs obtained from various experimental conditions (different cells, tissue types or treatments) and cDNA clones in a library is used to identify genes that are differentially regulated (Kuhn 2001). The frequency of identifying differentially expressed genes can be significantly improved if the concentration of differentially expressed transcripts in the cDNA library or complex cDNA population used as a hybridisation probe is enriched through subtractive hybridisation. Although differential screening and subtractive hybridisation are less commonly applied now than previously, they are still robust techniques for detecting differentially expressed genes with a high transcript abundance in plants (Wyrich et al., 1998; Davies and Robinson 2000).

Gene expression profiling based on the hybridisation of transcripts to arrays of DNA molecules immobilised on solid supports is rapidly becoming a powerful tool to identify and analyse genes involved in various plant biological processes (Baldwin et al., 1999). It is based on the principle of a “reverse Northern” where the DNA fragments or oligonucleotides corresponding to different genes (“probes”) are hybridised to complex total mRNA pools converted to cDNA (“targets”) (Bouchez and Höfte 1998). The support bound DNA is in excess so that the amount of target cDNA hybridised to each DNA molecule is a reflection of the abundance of the corresponding transcript in the mRNA population used (Baldwin et al., 1999). Technological advances such as the development of chemical and robotic methods to deposit large numbers of DNA fragments onto glass (microarrays) or membrane filters (macroarrays) allows hundreds or thousands of genes to be analysed simultaneously. Array hybridisation has been shown to be a high-throughput, quantitative and reproducible method for examining the expression profiles of plant genes and for identifying those genes that are differentially expressed (Ruan et al., 1998). The potential of these techniques for genome-wide analysis of gene function in plants have been reviewed extensively (Bouchez and Höfte 1998; Baldwin et al., 1999; Kehoe et al., 1999; Richmond and Somerville 2000; Breyne and Zabeau 2001). Currently however, there are only limited examples of their application to address specific biological processes in plants, being restricted to experimental systems for which large numbers of gene fragments are available such as *Arabidopsis* and rice.

2.3.3 ESTs as a qualitative indicator of plant gene expression

Collections of expressed gene data from a variety of plant species are becoming available as the range of EST programmes continue to expand. These ESTs have been used to obtain information about gene expression patterns in different tissues, cell types and developmental stages from model plants such as *Arabidopsis* (Newman et al., 1994; Cooke et al., 1996; White et al., 2000) and important crop plants including rice (Sasaki et al., 1994; Liu et al., 1995), oilseed rape (Park et al., 1993; Lee et al., 1998), castor bean (van de Loo et al., 1995), grape (Ablett et al., 2000), Lotus (Asamizu et al., 2000) and watermelon (Ok et al., 2000). These studies have identified a broad spectrum of genes corresponding to proteins involved with a myriad of different plant metabolic pathways.

2.3.3.1 Analysing diverse plant processes

The abundance of ESTs for many types of genes vary according to the tissue used for cDNA library production (Höfte et al., 1993, Cooke et al., 1996, Yamamoto and Sasaki 1997). When cDNA libraries are prepared from tissue-types or developmental stages characterised by specific biochemical pathways, the most abundantly expressed genes detected by ESTs correspond to proteins known to exhibit high levels of activity. A typical example is leaf tissue, where the majority of ESTs isolated from leaves of rice (Yamamoto and Sasaki 1997), *Brassica napus* (Lee et al., 1998), grape (Ablett et al., 2000) and watermelon (Ok et al., 2000) are homologous to transcripts involved in photosynthesis. Similarly, in developing endosperm of rice seeds, cDNAs encoding seed storage proteins such as glutelin and prolamin are highly abundant (Liu et al., 1995). In an analysis of more than 10 000 *Arabidopsis* ESTs generated from developing seeds, the number of ESTs in the data set encoding seed metabolic enzymes conformed with conventional biochemical knowledge of seed metabolism (White et al., 2000). When rice callus was cultured in the presence of 2,4-dichlorophenoxyacetic acid, many ESTs encoded ribosomal proteins (Sasaki et al., 1994, Yamamoto and Sasaki 1997). This was expected due to the vigorous growth state of the callus used for cDNA library construction. In addition, ESTs associated with biological processes distinctive to leguminous plants such as nodule development and secondary metabolism are abundant in young plants of *Lotus japonicus* (Asamizu et al., 2000).

The identification of ESTs has provided new insights into metabolic processes that have been studied extensively at the physiological and biochemical level. The transcript composition in ripening grape berries reflected a high degree of specialisation in berry cells (Ablett et al., 2000). These authors showed that 18% of the 2479 ESTs encoded transcripts associated with defense-regulation, detoxification and stress responses while 32% were homologous to genes involved with protein production and processing, and signal transduction. According to Ablett et al. (2000), the high levels of expression for genes associated with defense and maintenance of homeostasis, signal transduction and proteolysis in grape berries suggested that berry tissue was actively engaged in responding to environmental stimuli. Expression analysis of 2231 ESTs derived from the inner bark of *Cryptomeria japonica* revealed that, as expected, genes associated with cell wall formation were well represented but that transcripts similar to a variety of putative stress response genes were abundant (Ujino-Ihara et al., 2000). The apparent abundant expression of genes involved in various drought and wounding signaling pathways indicated that these were major stresses affecting *Cryptomeria japonica*.

Comparisons between ESTs isolated from leaves of rice and *Brassica napus* revealed that the expression of genes varied depending on the developmental state of the leaf (Lee et al., 1998). Fewer photosynthesis-related genes were detected in immature rice leaves than mature *Brassica* leaves, while gene products involved in general metabolic processes were more abundant in rice leaves than *Brassica* leaves (Lee et al., 1998). These authors suggested that immature leaves devote more cellular activities to the biosynthesis of cellular compartments than to photosynthesis.

EST analysis has also identified genes encoding proteins not previously identified in specific plant tissues or developmental stages. For example, in developing rice endosperm, genes encoding a glycine-rich RNA-binding protein and metallothionein-like proteins were reported for the first time in developing seeds (Liu et al., 1995). Similarly, in young *Lotus japonicus* plants, genes related to secondary metabolism and seed development, phenomena not expected to occur in young plants, were detected by EST analysis (Asamizu et al., 2000).

Up to 50% of newly identified ESTs have no match to previously identified gene sequences available in the public databases. For plants processes such as photosynthesis (Wyrich et al., 1998), seed development (White et al., 2000) and plant organ development (Lee et al., 1998; Asamizu et al., 2000), many novel ESTs have been detected that encode proteins with hitherto unknown functions. Furthermore, unidentified ESTs that correspond to genes that are expressed only in specific plant species (Ujino-Ihara et al., 2000) or particular plant organs such as seeds (White et al., 2000) and flowers (Asamizu et al., 2000) have been reported.

Many of these unidentified genes will have critical roles in specific plant molecular processes and emphasise that knowledge of the genetic basis of plant metabolism is still very limited and requires much more study.

2.3.3.2 Developing global gene expression profiles

When large enough data sets are available ESTs can be used to develop global expression profiles for plant tissues and genes. It has been reported that the abundance of a specific cDNA in an EST collection is a measure for gene expression (Audic and Claverie 1997). Quantifying the number of ESTs homologous to a specific gene provides a direct estimate of the level of mRNA expression for individual transcripts. This procedure, known as "electronic or digital northern" has been successfully used to develop expression profiles for various human (Adams et al., 1995) and invertebrate tissues (Santos et al., 1999). As the number of available plant ESTs continues to increase, this technique has also become a mechanism to study gene expression in plants (Ewing et al., 1999, Ablett et al., 2000, White et al., 2000).

When sufficient EST data is available from different plant tissues and developmental stages, comparative analysis of data sets can reveal differences in the expression of individual genes or complements of genes. Comparison between 2479 ESTs from grape berry tissue and 2438 from leaf tissue revealed marked differences in gene expression with only 12% of the ESTs common to both tissues (Ablett et al., 2000). These authors reported elevated transcript levels in the berry for genes associated with disease and defense-response as well as sugar and polysaccharide metabolism, with higher levels in the leaf of transcripts encoding genes involved in photosynthesis, amino acid, lipid and sterol metabolism and cytoskeletal structures.

Statistical analyses of EST data obtained from 10 rice cDNA libraries representing each of the principle tissues in the plant life cycle as well as the same tissues at different developmental stages, revealed correlated patterns of gene expression (Ewing et al., 1999). A mathematical procedure used to organise ESTs into gene clusters indicated that rice tissues could be associated together on the basis of similar patterns of expression and likewise, genes exhibiting tissue-dependent expression patterns were revealed. Identification of gene subsets exhibiting coordinated expression patterns in rice may aid the selection of candidate genes for specific metabolic processes and may provide new information about the interrelationships between different tissues and developmental pathways (Ewing et al., 1999).

The potential of the digital northern approach for analysing gene expression profiles is entirely dependent, however, on the availability of large numbers of ESTs. Consequently, its current application to plant research is limited to a few species. Certain restrictions also apply such as the use of non-normalised cDNA libraries for EST generation that have been prepared in a comparable manner (Ewing et al., 1999). In addition, random sampling of clones from cDNA libraries invariably results in EST collections being enriched with sequences representing highly abundant cDNAs. As cDNA libraries contain DNA sequences representative of the relative abundance of mRNAs in the tissue from which the library was constructed, it is expected that abundantly expressed sequences will be highly-represented, while genes expressed at low levels may not be present in EST data sets (McCarter et al., 2000). To increase the frequency of identifying rare mRNAs, strategies such as the use of subtracted cDNA libraries where common transcripts have been removed, leaving the more rare transcripts in the library, can be employed. Alternatively, ESTs can be sampled from a variety of specialised tissue- and stage-specific libraries, although large numbers of ESTs would be required for this to be effective (Newman et al., 1994). Nevertheless, the detection of redundant sequences in an EST collection does provide useful preliminary information about the types of genes abundantly expressed in the particular tissue or experimental condition being investigated (Cooke et al., 1996). It should be noted, however, that in many cases, gene expression profiles produced using the digital Northern approach have not been validated by conventional laboratory-based procedures for measuring gene expression levels. Efforts in this regard are underway for several plant species including legumes (Asamizu et al., 2000), rice (Delseny et al., 2001) and grapes (Ablett et al., 2000), however no results have yet been published.

2.3.4 Transcript profiling and screening for differentially expressed genes

The systematic analysis of the expression levels for multiple sets of plant genes has provided valuable information about transcript abundance and revealed a wide diversity of differentially expressed genes in a variety of tissue types, developmental stages, specific metabolic pathways and in response to abiotic or biotic stresses. The following commentary highlights some of the new information that has been obtained about important plant processes at the molecular level.

2.3.4.1 Organs, tissues and developmental stages

The transcript profiles of 1443 *Arabidopsis* genes have been analysed in root, leaf and two floral stages using cDNA microarrays (Ruan et al., 1998). In this study, between 1.4% and 5% of cDNAs were found to be highly expressed, 5.3% - 15.9% were moderately expressed while the majority of expressed genes were low abundance (62% - 75%). As expected, most of the highly expressed transcripts were homologous to well-characterised housekeeping or tissue-specific genes. However, several of the highly abundant transcripts were novel, having no homology to previously identified genes. These genes may have important roles in *Arabidopsis* development.

Comparisons between the levels of gene expression in leaf and root, leaf and flower, and flower bud and open flower revealed a large number of differentially expressed genes. Between leaf and root, 34% of the transcripts were significantly regulated while only 16% of genes were differentially expressed between leaf and open flower. Very few sequences (4%) were expressed at significantly different levels between the two floral stages. Database searches indicated homology to a wide range of genes with many of the tissue-specific transcripts displaying expression profiles consistent with previous reports.

To provide more information about genes expressed at both low and high levels during flowering, more than 3000 clones from equalised *Arabidopsis* cDNA libraries prepared from inflorescent apices and flower buds have been analysed by differential screening of high-density filter arrays (Takemura et al., 1999; Hyodo et al., 2000). In these studies, 12% of inflorescence-specific transcripts were expressed at low levels (Takemura et al., 1999) while only 0.8% were highly expressed (Hyodo et al., 2000). Results from these studies indicated that genes associated with the processes of transcription and translation, cell wall biogenesis, and signal transduction were important for the formation of reproductive meristems in *Arabidopsis*. Similar results have also been reported during floral transition in orchids (Yu and Goh 2000).

The biosynthetic pathways responsible for the accumulation of seed storage components have been comprehensively studied due to the economic value of seed products (Ohlrogge and Jaworski 1997; Eastmond and Rawsthorne 2000; White et al., 2000). Characterising gene expression patterns in *Arabidopsis* seeds has revealed new information about genes for which no published data was previously available. In a recent study by Girke et al. (2000), more than 10 000 transcripts from developing *Arabidopsis* seeds were analysed and seed-specific expression patterns for many genes were described for the first time (Girke et al., 2000). Although many genes were detected that are well known to be predominantly expressed in seeds, cDNAs homologous to various transcription factors, kinases, phosphatases, and developmental proteins were highly seed-specific. The tissue-specificity of many of these genes has not been characterised before. In addition, although genes associated with lipid biosynthesis were highly expressed in seeds as expected, they were also shown to be highly expressed in other tissues such as leaf and root, suggesting that lipid biosynthesis could have a "housekeeping" function in plants.

The process of fruit ripening involves considerable biophysical and biochemical changes however limited information is available about the molecular mechanisms controlling these changes, particularly for economically important non-climacteric fruits such as strawberries, blackcurrants, peppers and grapes. When differentially expressed genes were isolated from ripening wild strawberry fruits, putative identities revealed that none of the genes appeared directly related to processes generally associated with ripening such as cell wall metabolism and the accumulation of sugars and pigments (Nam et al., 1999). The ripening-induced transcripts identified in this study were homologous to genes associated with a wide range of processes including lipid metabolism, methionine biosynthesis, stress- and defense-responses, and secondary metabolism. Similarly, in ripening grape berries, more than half the transcripts isolated by differential screening encoded putative stress-response proteins (Davies and Robinson 2000). Transcripts associated with cell wall metabolism also accumulated preferentially during berry ripening (Davies and Robinson 2000). These authors suggested that the dramatic changes in mRNA levels observed are likely to be involved in the physical and metabolic changes occurring during ripening. The expression of cell wall related genes is expected to be linked to rapid cell expansion but may also act as a developmentally regulated defense mechanism against pathogen attack. In addition, some of the stress response proteins may have important roles in protecting the berries from the rapid changes in osmotic pressure and water potential occurring during ripening.

2.3.4.2 Metabolic pathways

Some of the most productive agricultural crops use the C_4 photosynthetic pathway of carbon assimilation. It is known that the division of labour between mesophyll and bundle-sheath cells in C_4 plants is the result of differential gene expression however, the extent of this differential expression is not well understood. Differential screening of a leaf cDNA library from the C_4 grass, *Sorghum bicolor*, identified 58 differentially expressed cDNAs (Wyrich et al., 1998). Of these, 25 were confirmed as mesophyll-specific while eight were bundle-sheath-specific. DNA sequence homology searches identified several new mesophyll cell-specific genes. Furthermore, four transcripts that accumulated preferentially in mesophyll cells had no match to previously identified genes. The expression of these genes was shown to be exclusive to leaves and was regulated by light suggesting that they may have important, but as yet unknown, roles in C_4 photosynthesis (Wyrich et al., 1998).

Further studies of light- and dark-regulated differential gene expression have been conducted in the model C_3 plant, *Arabidopsis thaliana*. Comparison between hybridisation profiles of 432 *Arabidopsis* cDNA fragments arrayed onto nylon filters and screened with mRNA from light-grown and dark-grown seedlings revealed significant differences in transcript abundance for about 15% of the fragments analysed (Desprez et al., 1998). Although some of the differentially expressed genes identified in this study had been reported previously to be light-regulated, for the majority of genes no differential expression between light-grown and dark-grown seedlings had been witnessed before. A functional analysis of *Arabidopsis* photosynthesis-related genes is currently underway to gain more insight into the genetic basis of this important metabolic pathway (Pesaresi et al., 2001).

Plants are dependent on light for growth and development and respond to day/night cycling in a number of physiological ways. Many genes have been identified that are regulated in a diurnal and circadian fashion but have been obtained through the analysis of only a few genes at a time. A large-scale analysis of *Arabidopsis* cDNA microarrays containing over 11 500 ESTs indicated that 11% of the genes exhibited a diurnal expression pattern while 2% cycled with a circadian rhythm (Schaffer et al., 2001). The differentially expressed genes identified in this study encoded a wide range of proteins associated with photosynthesis and carbon and nitrogen metabolism, as well as various transcription factors and protein kinases. Many of the genes detected had not previously been reported to cycle. In addition, a large proportion of differentially expressed genes could not be assigned a putative identity and represent novel genes that are regulated by diurnal and circadian rhythms (Schaffer et al., 2001). Similarly, a preliminary analysis of genes that are modulated during the cell cycle in tobacco also revealed that many of the differentially expressed transcripts either matched genes of unknown function or represented novel sequences (Breyne and Zabeau 2001). These authors used a cDNA-AFLP transcript profiling approach to isolate over 1150 AFLP tags that exhibited a cell cycle modulated profile. For tags that could be assigned a putative identity by DNA sequence analysis, many were homologous to genes known to be differentially regulated during the cell cycle such as histones and tubulins. However, genes exhibiting very low expression levels that encoded transcription factors and other regulatory proteins were also detected (Breyne and Zabeau 2001).

2.3.4.3 Stress-inducible genes

The ability to study coordinated patterns of gene expression using large-scale technologies such as cDNA microarray analysis has also impacted on investigations of plant responses to biotic and abiotic stresses. Plants are exposed to many environmental abiotic stresses such as drought, salinity and high and low temperatures. Efforts to understand the molecular mechanisms of stress tolerance in plants have been confounded by the multigenic responses to different kinds of stresses (Bohnert et al., 2001). Molecular studies of plant stress-responses through large-scale EST and gene expression profile analysis have been recently reported for *Arabidopsis* (Bohnert et al. 2001; Seki et al., 2001), rice and ice plant (*Mesembryanthemum crystallinum*) (Bohnert et al., 2001). When plants were exposed to a salt-stress, a myriad of different genes encoding proteins involved with many diverse processes including general cellular metabolism, signal transduction, cell growth and protein synthesis were both up- and down-regulated in response to the stress (Bohnert et al., 2001). As demonstrated by other similar gene profiling experiments, a large proportion of the differentially regulated genes in this study did not show sequence homology to genes of known function.

Hybridisation analysis of *Arabidopsis* cDNA microarrays identified several new genes that were expressed in response to a drought and cold stress (Seki et al., 2001). When 1300 *Arabidopsis* cDNAs were analysed, 44 genes were isolated that were induced by drought with 30 of these exhibiting DNA sequence homology to genes not previously reported to be drought-inducible (Seki et al., 2001). Likewise, 10 new cold-inducible genes were identified. The roles of these genes in regulating the plant response to the stress are not known. Similarly, when rice seedlings subjected to an anaerobic stress were analysed by SAGE, results revealed the induction and repression of genes not previously known to respond to anaerobic stress (Matsumura et al., 1999). This included elevated expression levels of a gene encoding the seed storage protein prolamin. As the stress response of prolamin gene expression had not previously been described, these authors suggested that this gene might have a novel function in anaerobiosis.

It is evident from the data presented above that a wealth of new information about plant gene expression is rapidly emerging as more research of important plant processes at the gene level is published. Many large-scale EST and gene expression profiling projects are currently underway (Table 2.1) and it is anticipated that the amount of published data will escalate in the near future. One of the challenges for the future will be assigning functions to all the new genes being discovered through these programmes. In addition, managing and integrating the diverse elements of information about the genetic basis of plant metabolism with what is already known at the biochemical and physiological level will be a major task. This will be essential however, for informed decisions to be made about how best to apply this new found knowledge towards efforts to improve plant performance.

Table 2.1 Selected examples of large-scale expression profiling projects currently underway for plants (modified from Richmond and Somerville 2000)

Project title	Principal scientist/Institution
A public soybean EST project	R Shoemaker/ARS, Ames, Iowa, USA
Identification of genes involved in plant response to water stress	J Mullet/Texas A&M University, USA
Genetic engineering of oilseed crops	J Ohlrogge/Michigan State University, USA
The effects of external stimuli on plant responses at the molecular and cellular levels	E Davies/North Carolina State University, USA
Genomic analysis of seed quality traits in corn	B Lemieux/University of Delaware, USA
Regulation of metabolism in developing seeds of *Arabidopsis*	C Benning/Michigan State University, USA
The effect of environment on grain development productivity and quality in wheat	W Hurkman/USDA, Albany, California, USA

2.4 STATUS OF SUGARCANE GENOMICS

2.4.1 Metabolism in the sugarcane culm

Sugarcane (*Saccharum* spp. hybrids) is an important agricultural crop plant that produces approximately 70% of the world's sugar (Grof 2001). Sugarcane belongs to the monocotyledenous Gramineae family that includes other important crop plants such as maize, wheat, sorghum and rice. The physiology and biochemistry of sucrose accumulation has been the subject of intensive study as, economically, it is the most important trait in sugarcane.

However, data from carbon partitioning studies clearly demonstrates that, at best, approximately 70% of the incoming carbon from the leaves to the culm is allocated to the sucrose pool (Whittaker and Botha 1997; Vorster and Botha 1999; Whittaker and Botha 1999). Furthermore, in young internodes, even those with a high sucrose accumulation rate, less than 50% of the incoming carbon is devoted to sucrose accumulation and storage. This illustrates that other metabolic processes are operational in the sugarcane culm, but these have not yet been studied.

Sucrose accumulation during growth and maturation of the sugarcane culm is a complex process. Efforts to understand the regulation of sucrose metabolism have been comprehensively reviewed elsewhere (Moore 1995; Grof and Campbell 2001). Studies have revealed that rapid cycling and turnover of sucrose between the vacuole and metabolic and apoplastic compartments are important factors for sucrose accumulation (Wendler et al., 1990; Moore 1995).

Biochemical research has identified several key enzymes that synthesise and break down sucrose in sugarcane. Characterisation of enzyme activities in the sugarcane culm has been reported for sucrose synthase (Lingle 1999; Black and Botha 2000), sucrose phosphate synthase (Zhu et al., 1997; Black and Botha 2000), pyrophosphate: D-fructose-6-phosphate 1-phosphotransferase (Whittaker and Botha 1999) and various invertases (Zhu et al., 1997; Vorster and Botha 1999; Rose and Botha 2000; Albertson et al., 2001). These studies have assisted in defining the roles for various enzymes in regulating sucrose accumulation. However, the mechanisms by which the processes associated with sucrose accumulation in sugarcane are controlled are still not well understood.

Information about how metabolism in the sugarcane culm is regulated at the gene level is very limited for sugarcane. Knowledge of sugarcane gene expression has been confined to studies of single-genes encoding important enzymes involved in sucrose metabolism. No studies to identify and analyse those genes with important roles in culm maturation have been reported. Recently, however, sugarcane researchers have begun to take advantage of the advances made in genome analysis in other plants to study various aspects of sugarcane culm maturation and sucrose metabolism at the molecular level.

2.4.2 Single-gene studies

A comparative analysis of the expression patterns of two genes encoding sucrose-phosphate synthase (SPS) in sugarcane has been conducted (Sugiharto et al., 1997). Northern hybridisation analysis indicated that the two genes exhibited differential expression patterns in different sugarcane organs and in response to light, suggesting that SPS in sugarcane is encoded by more than one gene. When cDNAs encoding soluble acid invertase (SAI) were isolated from low-sucrose and high-sucrose accumulating sugarcane lines, the nucleotide and deduced peptide sequences of the cDNAs were found to be highly similar but the level of gene expression differed markedly between the two lines (Zhu et al., 2000). Both cDNAs exhibited developmentally regulated gene expression patterns where transcript levels of SAI were high in the meristematic apex and immature internodes and substantially lower in the mature internodes. However, comparative analysis of transcript abundance between the lines revealed that the overall level of expression of SAI mRNA was higher in the low- sucrose accumulating lines. The mRNA expression pattern corresponded favourably with SAI enzyme activity in the low-sucrose and high-sucrose accumulating lines (Zhu et al., 2000).

There are two known isoforms of sucrose synthase (SuSy) in sugarcane that appear to be differentially regulated (Buczynski et al., 1993). Northern analysis of a full-length cDNA corresponding to the SuSy-1 isoform in sugarcane leaf and internodal tissues of varying maturity as well as roots and germinating buds detected transcripts in all tissues examined, although lower levels were evident in immature and mature leaves (Lingle and Dyer 2001). The mRNA expression pattern of SuSy-1 obtained in this study agreed with the distribution of the SuSy-1 protein obtained previously (Buczynski et al., 1993).

2.4.3 Current research

Genomics research is currently underway for sugarcane. The complex polyploid genome of sugarcane presents a challenge to molecular studies but substantial progress has been made into the organisation of the sugarcane genome by comparative genome analyses using data from maize and sorghum (Grivet et al., 1996; Ming et al., 1998; Draye et al. 2001). This work has important consequences for sugarcane research through the identification of markers that could be used in breeding programmes to improve crop performance. Advances made in this aspect of sugarcane genomics are, however, beyond the scope of the current review.

Several EST programmes are underway for sugarcane in various sugarcane-growing countries around the world including Australia, Brazil and South Africa. The goals of these various programmes are essentially similar: to identify and characterise genes expressed during sugarcane growth and maturation in order to pinpoint candidate genes with important roles in processes of agricultural significance such as sucrose accumulation and resistance to pests and diseases.

Some sugarcane ESTs have been made available in the public domain (Carson and Botha 2000) however, due to strategic considerations, most EST data generated by these programmes are currently maintained in private databases. Only two reports of an EST analysis in sugarcane have been published (Carson and Botha 2000; Carson and Botha 2002).

No large-scale gene expression profiling studies are available yet for sugarcane although some results have been published in the Proceedings of the 24th International Society of Sugar Cane Technologists Congress held recently (17-21 September 2001). Here, the analysis of 8504 ESTs, 1085 from young sugarcane stem and 7419 from maturing stem, was presented (Casu et al., 2001). These authors identified a wide range of genes associated with general metabolic processes including cell wall metabolism, chromatin and DNA metabolism, RNA metabolism, protein synthesis and signal transduction. As reported for EST analyses in other plant tissues, approximately 50% of the ESTs generated in this study did not match known gene sequences in the public databases. Comparison between ESTs obtained from the young and maturing stem suggested that while the young internodes were more actively involved in growth, older internodes preferentially expressed genes associated with fibre biosynthesis and degradation as well as a variety of stress-related proteins (Casu et al., 2001). Few sucrose metabolism-related genes were identified. In another report, genes preferentially expressed in maturing internodal tissue were specifically targeted using subtractive hybridisation (Carson et al., 2001). The differentially expressed sequences obtained were homologous to genes predominantly associated with regulatory processes, stress responses, cell wall metabolism and carbohydrate metabolism.

Comparative analyses of transcript abundance for sugarcane genes in meristematic and maturing internodal tissues using cDNA microarrays are currently being conducted, although only preliminary results have been obtained (Casu et al., 2001). Further studies are in progress. In a recent media report (September 2001), a strategic alliance between the Brazilian sugarcane EST genome project, SUCEST, and the private Belgian agbiotech company CropDesign was announced. This alliance plans to perform an in-depth functional genomic evaluation of the SUCEST data to select candidate genes for crop improvement.

2.5 APPLICATION FOR CROP IMPROVEMENT

Genetic research has entered into an exciting transition from being data-poor to data-rich. Exploiting the massive data resources to extract the biological knowledge and apply it towards efforts to improve crop performance is a major challenge for the future.

Genomics has the potential to facilitate crop improvement through the acquisition of knowledge about gene function and how complex biological processes associated with commercially important traits are regulated at the molecular level. Combining this knowledge with the further development of tools for successful genetic manipulation of crop plants will lead to robust genetic modification programmes where rational changes can be formulated from first principles (Somerville and Somerville 1999).

Functional genomics programmes have ensured that there are now large numbers of structural genes available to express specific proteins in plants, including sugarcane. However, their application towards plant improvement is currently limited by the lack of availability of promoters and regulatory elements to target the expression of these genes in the appropriate organs and tissues or during specific developmental stages (Miflin 2000). Further analyses of differentially expressed genes obtained through gene expression profiling experiments may yield new promoters that will be useful to control the expression of economic traits. Presently, the application of genomics research to crop improvement is still in the very early stages, particularly for less widely grown crops such as sugarcane.

2.6 REFERENCES

Ablett E, Seaton G, Scott K, Shelton D, Graham MW, Baverstock P, Lee LS and Henry R (2000) Analysis of grape ESTs: global gene expression patterns in leaf and berry. Plant Sci 159: 87-95.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR and Venter JC (1991) Complemetary DNA sequencing: expressed sequence tags and the human genome project. Science 252:1651-1656.

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C and Venter JC (1992) Sequence identification of 2375 human brain genes. Nature 355: 632-634.

Adams MD, Kerlavage AR, Fleischmann RD, Fauldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O and Venter JC (1995) Initial assessment of human gene diversity and expression patterns based on 83 million nucleotides of cDNA sequence. Nature 377: 3-17.

Aharoni A, Keizer LCP, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas J, van Houwelingen AMML, De Vos RCH, van der Voet H, Jansen RC, Guis M, Mol J, Davis RW, Schena M, van Tunen AJ and O'Connell AP (2000) Identification of the *SAAT* gene involved in strawberry flavor biogenesis by use of DNA microarrays. Plant Cell 12: 647-661.

Albertson PL, Peters KF and Grof CPL (2001) An improved method for the measurement of cell wall invertase activity in sugarcane tissue. Aust J Plant Physiol 28: 323-328.

Asamizu E, Watanabe M and Tabata S (2000) Large scale structural analysis of cDNAs in the model legume, *Lotus japonicus*. J Plant Res 113: 451-455.

Audic S and Claverie J-M (1997) The significance of digital gene expression profiles. Genome Res 7(10): 986-995.

Bachem CWB, van der Hoeven RS, de Bruijn SM, Vreugdenhil D, Zabeau M and Visser GF (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. Plant J 9: 745-753.

Baldwin D, Crane V and Rice D (1999) A comparison of gel-based, nylon filter and microarray techniques to detect differential RNA expression in plants. Curr Opin Plant Biol 2: 96-103.

Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125: 1164-1165.

Black K and Botha FC (2000) Sucrose phosphate synthase and sucrose synthase activity during maturation of internodal tissue in sugarcane. Aust J Plant Physiol 27: 81-85.

Boguski MS, Lowe TMJ and Tolstoshev CM (1993) dbEST – database for "expressed sequence tags". Nature Genetics 4: 332-333.

Boguski MS (1999) Biosequence exegesis. Science 286: 453-455.

Bohnert HJ, Ayoubi P, Borchert C, Bressan RA, Burnap RL, Cushman JC, Cushman MA, Deyholos M, Fischer R, Galbraith DW, Hasegawa PM, Jenks M, Kawasaki S, Koiwa H, Kore-eda S, Lee B-H, Michalowski CB, Misawa E, Nomura M, Ozturk N, Postier B, Prade R, Song C-P, Tanaka Y, Wang H and Zhu J-K (2001) A genomics approach towards salt stress tolerance. Plant Physiol Biochem 39: 295-311.

Bouchez D and Höfte H (1998) Functional genomics in plants. Plant Physiol 118: 725-732.

Breyne P and Zabeau M (2001) Genome-wide expression analysis of plant cell cycle modulated genes. Curr Opin Plant Biol 4: 136-142.

Buczynski SR, Thom M, Chourey P and Maretzki A (1993) Tissue distribution and characterization of sucrose synthase isozymes in sugarcane. J Plant Physiol 142: 641-646.

Carson DL and Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. Crop Sci 40(6): 1769-1779.

Carson DL and Botha FC (2002) Genes expressed in sugarcane maturing internodal tissue. Plant Cell Rep 20(11): 1075-1081.

Carson DL, Huckett BI and Botha FC (2001) Genomics research at SASEX: Perspectives from a small-scale program. Proc Int Soc Sugar Cane Technol 24: 539-541.

Casu R, Dimmock C, Thomas M, Bower N, Knight D, Grof C, McIntyre L, Jackson P, Jordan D, Whan V, Drenth J, Tao Y and Manners J (2001) Genetic and expression profiling in sugarcane. Proc Int Soc Sugar Cane Technol 24: 542-546.

Cooke R, Raynal M, Laudié M, Grellet F, Delseny M, Morris P-C, Guerrier D, Giraudat J, Quigley F, Clabault G, Li Y-F, Mache R, Krivitzky M, Gy IJ-J, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clément B, Philipps G, Hervé C, Bardet C, Tremousaygue D, Lescure B, Lacomme C, Roby D, Jourjon M-F, Chabrier P, Charpenteau J-L, Desprez T, Amselem J, Chiapello H and Höfte H (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. Plant J 9(1): 101-124.

Davies C and Robinson SP (2000) Differential screening indicates a dramatic change in mRNA profiles during grape berry ripening. Cloning and characterisation of cDNAs encoding putative cell wall and stress response proteins. Plant Physiol 122: 803-812.

Delseny M, Salses J, Cooke R, Sallaud C, Regad F, Lagoda P, Guiderdoni E, Ventelon M, Brugidou C and Ghesquière A (2001) Rice genomics: Present and future. Plant Physiol Biochem 39: 323-334.

Desprez T, Amselem J, Caboche M and Höfte H (1998) Differential gene expression in *Arabidopsis* monitored using cDNA arrays. Plant J 14(5): 643-652.

Draye X, Lin Y-R, Qian X-y, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren S-x, Wing RA and Paterson AH (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the Sorghum genome as a foundation. Plant Physiol 125: 1325-1341.

Eastmond PJ and Rawsthorne S (2000) Coordinate changes in carbon partitioning and plastidial metabolism during the development of oilseed rape embryos. Plant Physiol 122: 767-774.

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S and Claverie J-M (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9: 950-959.

Girke T, Todd J, Ruuska S, White J, Benning C and Ohlrogge J (2000) Microarray analysis of developing Arabidopsis seeds. Plant Physiol 124: 1570-1581.

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C and Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organisation in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Grof CPL (2001) Molecular manipulation of sucrose metabolism. Proc Int Soc Sugar Cane Technol 24: 586-587.

Grof CPL and Campbell JA (2001) Sugarcane sucrose metabolism: scope for molecular manipulation. Aust J Plant Physiol 28: 1-12.

Hieter P and Boguski M (1997) Functional Genomics: It's all how you read it. Science 278: 601-602.

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D and Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana.* Plant J 4: 1051-1061.

Höög C (1991) Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. Nucleic Acids Res 19: 6123-6127.

Hyodo H, Takemura M, Yokota A, Ohyama K and Kohchi T (2000) Systematic isolation of highly transcribed genes in inflorescence apices in *Arabidopsis thaliana* from an equalized cDNA library. Biosci Biotechnol Biochem 64(7): 1538-1541.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

James DW, Lim E, Keller J, Plooy I, Ralston E and Dooner HK (1995) Directed tagging of the *Arabidopsis* FATTY ACID ELONGATION (FAE1) gene with the maize transposon *activator*. Plant Cell 7: 309-319.

Kehoe DM, Villand P and Somerville S (1999) DNA microarrays for studies of higher plants and other photosynthetic organisms. Trends Plant Sci 4: 38-41.

Keith CS, Hoang DO, Barrett BM, Feigelman B, Nelson MC, Thai H and Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol 101: 329-332.

Kozian DH and Kirschbaum BJ (1999) Comparative gene-expression analysis. Tibtech 17: 73-77.

Kuhn E (2001) From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in plants. Annals of Bot 87: 139-155.

Lee CM, Lee YJ, Lee MH, Nam HG, Cho TJ, Hahn TR, Cho MJ and Sohn U (1998) Large-scale analysis of expressed genes from the leaf of oilseed rape (*Brassica napus* L.) Plant Cell Rep 17: 930-936.

Liang P and Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257: 967-971.

Lingle SE (1999) Sugar metabolism during growth and development in sugarcane internodes. Crop Sci 39: 480-486.

Lingle SE and Dyer JM (2001) Cloning and expression of sucrose synthase-1 cDNA from sugarcane. J Plant Physiol 158: 129-131.

Liu J, Hara C, Umeda M, Zhao Y, Okita TW and Uchimiya H (1995) Analysis of randomly isolated cDNAs from developing endosperm of rice (*Oryza sativa* L.): evaluation of expressed sequence tags, and expression levels of mRNAs. Plant Mol Biol 29: 685-689.

Lockhart DJ and Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405: 827-836.

Maheshwari SC, Maheshwari N and Sopory SK (2001) Genomics, DNA chips and a revolution in plant biology. Curr Sci 80(2): 252-261.

Martienssen RA (1998) Functional genomics: probing plant gene function and expression with transposons. Proc Natl Acad Sci USA 95: 2021-2026.

Matsumura H, Nirasawa S and Terauchi R (1999) Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE) Plant J 20(6): 719-726.

McCarter J, Abad P, Jones JT and Bird D (2000) Rapid gene discovery in plant parasitic nematodes *via* expressed sequence tags. Nematology 2(7): 719-731.

Miflin B (2000) Crop improvement in the 21st century. J Exp Bot 51: 1-8.

Ming R, Liu S-C, Lin Y-R, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE and Paterson AH (1998) Detailed alignment of Saccharum and Sorghum chromosomes: Comparative organisation of closely related diploid and polyploid genomes. Genetics 150: 1663-1682.

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22: 661-679.

Nam Y-W, Tichit L, Leperlier M, Cuerq B, Marty I and Lelièvre J-M (1999) Isolation and characterisation of mRNAs differentially expressed during ripening of wild strawberry (*Fragaria vesca* L.) fruits. Plant Mol Biol 39: 629-636.

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E and Somerville C (1994) Genes Galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241-1255.

Ohlrogge J and Jaworski J (1997) Regulation of plant fatty acid biosynthesis. Annu Rev Plant Physiol Plant Mol Biol 48: 109-136.

Ok S, Chung YS, Um BY, Park MS, Bae J-M, Lee SJ and Shin JS (2000) Identification of expressed sequence tags of watermelon (*Citrullus lanatus*) leaf at the vegetative stage. Plant Cell Rep 19: 932-937.

Park YS, Kwak JM, Kwon OY, Kim YS, Lee DS, Cho MJ, Lee HH and Nam HG (1993) Generation of expressed sequence tags of random root cDNA clones of *Brassica napus* by single-run partial sequencing. Plant Physiol 103: 359-370.

Pesaresi P, Varotto C, Richly E, Kurth J, Salamini F and Leister D (2001) Functional genomics of *Arabidopsis* photosynthesis. Plant Physiol Biochem 39: 285-294.

Rose S and Botha FC (2000) Distribution patterns of neutral invertase and sugar content in sugarcane internodal tissues. Plant Physiol Biochem 38: 819-824.

Richmond T and Somerville S (2000) Chasing the dream: plant EST microarrays. Curr Opin Plant Biol 3: 108-116.

Ruan Y, Gilmore J and Conner T (1998) Towards *Arabidopsis* genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. Plant J 15(6): 821-833.

Santos TM, Johnston DA, Azevedo V, Ridgers IL, Martinez MF, Marotta GB, Santos RL, Fonseca SF, Ortega JM and Rabelo EML (1999) Analysis of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach. Mol Biochem Parasitol 103: 79-97.

Sasaki T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Murayama-Kayano E, Takiguchi T, Takasuga A, Niki T, Ishimaru K, Ikeda H, Yamamoto Y, Mukai Y, Ohta I, Miyadera N, Havukkala I and Minobe Y (1994) Towards cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J 6: 615-624.

Schaffer R, Landgraf J, Accerbi M, Simon V, Larson M and Wisman E (2001) Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. Plant Cell 13: 113-123.

Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E and Davis RW (1998) Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotech 16: 301-306.

Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y and Shinozaki K (2001) Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. Plant Cell 13: 61-72.

Somerville C and Somerville S (1999) Plant functional genomics. Science 285: 380-383.

Sugiharto B, Sakakibara H, Saumadi and Sugiyama T (1997) Differential expression of two genes for sucrose-phosphate synthase in sugarcane: molecular cloning of the cDNAs and comparative analysis of gene expression. Plant Cell Physiol 38(8): 961-965.

Takemura M, Fujishige K, Hyodo H, Ohashi Y, Kami C, Nishii A, Ohyama K, Kohchi T (1999) Systematic isolation of genes expressed at low levels in inflorescence apices of *Arabidopsis thaliana.* DNA Res 6: 275-282.

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

Uchimiya H, Kidou S, Shimazaki T, Takamatsu S, Hashimoto H, Nishi R, Aotsuka S, Matsubayashi Y, Kidou N, Umeda M and Kato A (1992) Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (*Oryza sativa* L.) Plant J 2: 1005-1009.

Ujino-Ihara T, Yoshimura K, Ugawa Y, Yoshimaru H, Nagasaka K and Tsumura Y (2000) Expression analysis of ESTs derived from the inner bark of *Cryptomeria japonica*. Plant Mol Biol 43: 451-457.

van de Loo FJ, Turner S and Somerville C (1995) Expressed sequence tags from developing castor seeds. Plant Physiol 108: 1141-1150.

Velculescu V, Zhang L, Vogelstein B and Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487.

Vorster DJ and Botha FC (1999) Sugarcane internodal invertases and tissue maturity. J Plant Physiol 155: 470-476.

Waterston R, Martin C, Craxton M, Coulson A, Hillier L, Durbin R, Green P, Showkeen R, Halloran N, Metzstein M, Hawkins T, Wilson R, Berks M, Du Z, Thomas K, Thierry-Mieg J and Sulston J (1992) A survey of expressed genes in *Caenorhabditis elegans*. Nature Genet 1: 114-123.

Wendler R, Veith R, Dancer J, Stitt M and Komor E (1990) Sucrose storage in cell suspension cultures of *Saccharum* sp. (sugarcane) is regulated by a cycle of synthesis and degradation. Planta 183: 31-39.

White JA, Todd J, Newman T, Focks N, Girke T, Martínez de Ilárduya O, Jaworski JG, Ohlrogge JB and Benning C (2000) A new set of Arabidopsis expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol 124: 1582-1594.

Whittaker A and Botha FC (1997) Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. Plant Physiol 115: 1651-1659.

Whittaker A and Botha FC (1999) Pyrophosphate: D-fructose-6-phosphate 1-phosphotransferase activity patterns in relation to sucrose storage across sugarcane varieties. Physiol Plant 107: 379-386.

Willmann MR (2001) *Arabidopsis* enters the post-sequencing era. Trends Plant Sci 6(2): 51.

Wyrich R, Dressen U, Brockmann S, Streubel M, Chang C, Qiang D, Paterson AH and Westhoff P (1998) The molecular basis of C_4 photosynthesis in sorghum: isolation, characterisation and RFLP mapping of mesophyll- and bundle-sheath-specific cDNAs obtained by differential screening. Plant Mol Biol 37: 319-335.

Yamamoto K and Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35: 135-144.

Yu H and Goh CJ (2000) Differential gene expression during floral transition in an orchid hybrid *Dendrobium* Madame Thong-In. Plant Cell Rep 19: 926-931.

Yuan Q, Quackenbush J, Sultana R, Pertea M, Salzberg SL and Buell CB (2001) Rice Bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species. Plant Physiol 125: 1166-1174.

Zhu YJ, Komor E and Moore PH (1997) Sucrose accumulation in the sugarcane stem is regulated by the difference between the activities of soluble acid invertase and sucrose phosphate synthase. Plant Physiol 115: 609-616.

Zhu YJ, Albert HH and Moore PH (2000) Differential expression of soluble acid invertase genes in the shoots of high-sucrose and low-sucrose species of *Saccharum* and their hybrids. Aust J Plant Physiol 27: 193-199.

CHAPTER 3

PRELIMINARY ANALYSIS OF EXPRESSED SEQUENCE TAGS FOR SUGARCANE

[Crop Science 40(6): 1769-1779 (2000)]

3.1 ABSTRACT

Sugarcane, with its complex polyploid genome, is not well understood at the genetic level. Partial sequencing of anonymous cDNA clones is a widely used technique for gene identification. These partial cDNA sequences, or Expressed Sequence Tags (ESTs) have potential application for the identification of important genes for genetic manipulation. This study aimed to initiate the preliminary development of an Expressed Sequence Tag database for sugarcane and thereby gain some potentially useful information on sugarcane gene sequences. A nondirectional cDNA library has been constructed from sugarcane leaf roll (meristematic region) tissue. Two hundred and fifty clones have been randomly selected, subjected to single-pass sequencing from the 5' end of the vector and identified by sequence similarity searches against gene sequences in international databases. Of the 250 leaf roll clones, 26% exhibit similarity to known plant genes, 50% to non-plant genes while 24% represent new gene sequences. Analysis of the identified clones indicated sequence similarity to a broad diversity of genes encoding proteins such as enzymes, structural proteins and regulatory factors. A significant proportion of genes identified in the leaf roll were involved in processes related to protein synthesis and protein modification, as would be expected in meristematic tissues. These results present a successful application of EST analysis in sugarcane and provide a preliminary indication of gene expression in leaf roll tissue.

3.2 INTRODUCTION

Sugarcane (*Saccharum spp*. hybrids) is a perennial monocotyledenous grass belonging to the *Saccharum* genus. It is a crop of substantial economic importance, providing approximately two third's of the world's sugar with an estimated annual worth of about $143 billion (Gallo-Meagher and Irvine 1996). It is not a simple plant on a genetic level, being a very complex polyploid with chromosome numbers ranging between 100 and 130 (Lu et al., 1994). Genetic research into sugarcane has discovered numerous agronomically important phenotypic traits. However, very little information is available about the genes responsible for these traits. Prior to the initiation of this project only 10 sugarcane gene sequences had been identified. Five had been published (Albert et al., 1995; Bugos and Thom 1993a; Bugos and Thom 1993b; Henrik et al., 1992; Tang and Sun 1993) while the remainder had been submitted directly to GenBank database (Alix 1997; Bugos and Thom 1993; Dharmasiri and Harrington 1996; Grof 1995; Sugiharto et al., 1997).

The last decade has seen a rapid proliferation in knowledge about plant and animal genomes through the application of large-scale partial sequencing of anonymous cDNA clones from cDNA libraries and their subsequent identification through homology searches of public databases. This approach, commonly referred to as Expressed Sequence Tag (EST) analysis, has been extensively applied in large-scale cDNA sequencing projects for a variety of both plant and animal species such as humans (Adams et al., 1991; Adams et al., 1992), nematodes (McCombie et al., 1992a; Waterston et al., 1992), *Arabidopsis* (Newman et al., 1994) and rice (Sasaki et al., 1994). These groups have shown that partial cDNA sequences, or ESTs, can be used successfully to identify putative clones for a wide range of gene products. ESTs have been reported both in the literature and public databases for 47 690 rice cDNAs (Uchimaya et al., 1992; Sasaki et al., 1994, dbEST release February 2000), 193 090 Arabidopsis cDNAs (Höfte et al., 1993; Newman et al., 1994, dbEST release February 2000) and 55 466 maize cDNAs (Keith et al., 1993, dbEST release February 2000). However, the availability of plant ESTs in the public databases is substantially less than that available for animal systems. This results in many plant gene identifications being based upon their sequence similarity to animal rather than plant species. There is a need, therefore, to identify and characterise new plant genes in order to increase the availability of plant genes in the international public databases.

Sugarcane biotechnology research world-wide is focussed primarily on two main areas, genetic manipulation and identification of markers. One of the problems associated with genetic manipulation of sugarcane is the lack of homologous gene sequences, especially important for antisense work. Similarly, the lack of known sugarcane genes also has implications for molecular marker programmes. The most recently published sugarcane maps have been constructed using anonymous RFLP and RAPD probes as well as heterologous probes from species such as maize, oats and rice (da Silva et al., 1995; Grivet et al., 1996). The identification of sugarcane genes could thus have significant consequences for sugarcane mapping and genetic manipulation and is therefore of great importance.

As a first step to address this issue we have prepared cDNA libraries from different tissue types in the sugarcane plant. Here we report on the preliminary analysis of 250 anonymous cDNA clones from a library composed of mRNA isolated from the leaf roll (meristematic region) of the commercial sugarcane cultivar NCo376. This work will make a significant contribution towards sugarcane biotechnology.

3.3 MATERIALS AND METHODS

3.3.1 Total and poly (A^+) RNA Isolation

RNA was extracted from the leaf roll (tissue section comprising apical meristem plus approximately 5 cm of etiolated immature leaf whorl) of mature field-grown sugarcane plants (*Saccharum spp.* hybrid, cultivar NCo376) using a modified method of Thompson et al., (1993). Approximately 4g of tissue was used for each extraction. Tissue was ground to a fine powder under liquid nitrogen and transferred to a 50 ml Corning tube on ice. To each sample, 4 ml of RNA extraction buffer (1% (w/v) sodium-dodecyl sulphate, 1 mM aurin tricarboxylic acid (ATA), 4% (w/v) p-aminosalicyclic acid, 10 mM Tris-HCl pH 7.5, 1 mM ethylenedinitrilotetracetic acid and 2% (v/v) 2-mercaptoethanol) and 4 ml phenol:chloroform:isoamylalcohol (50:49:1) was added. Samples were homogenised with an Ultra-Turrax vertical homogeniser for 3-4 minutes and then centrifuged at 4300g for 20 minutes at 4 °C. The aqueous layer was removed, added to 2 M LiCl and 1 mM ATA (final concentration) and allowed to precipitate overnight at 4 °C. Samples were then centrifuged at 4300g for 20 minutes at 4 °C. The pellet was suspended in 1 ml of 50 μM ATA and transferred to a microcentrifuge tube. Samples were centrifuged at 3000g for 2 minutes to remove particulate matter and the supernatant transferred to a fresh tube. RNA was precipitated overnight at 4 °C with 2 M LiCl (final concentration). Samples were then centrifuged at 5000g for 10 minutes at 4 °C, the supernatant discarded and the pellet rinsed with ice-cold 70% (v/v) ethanol. The pellet was resuspended in 250 μl of 50 μM ATA. The RNA was precipitated by the addition of 0.5 volumes 7.5 M ammonium acetate and 3 volumes 95% (v/v) ethanol with incubation for at least 2 hours at –20 °C. After centrifugation at 5000g for 30 minutes at 4 °C the purified RNA was resuspended in 50 μM ATA. mRNA was isolated using Hybond mAP (messenger affinity paper) (Amersham, UK), according to the manufacturer's instructions.

3.3.2 Construction of a leaf roll cDNA library

3.3.2.1 cDNA synthesis

First-strand cDNA synthesis was performed according to a modification of the method described in the Promega Protocols and Applications Guide (1990). Approximately 1 μg of poly (A^+) RNA was used in a first-strand synthesis reaction catalysed by the RNase H^- M-Mulv (Moloney - Murine Leukemia Virus) reverse transcriptase enzyme (Stratagene, La Jolla, CA, USA) and using oligo $d(T)_{18}$ as the primer. Final reaction conditions for first-strand synthesis were: 1 μg mRNA; 0.5 μg/μg mRNA of oligo $d(T)_{18}$; 50 mM Tris-HCl, pH 8.3; 75 mM KCl; 3 mM $MgCl_2$; 10 mM DTT; 1 mM each of dATP, dCTP, dGTP, dTTP; 1.6 u/μl ribonuclease inhibitor; 50 u/μg mRNA of RNase H^- M-Mulv reverse transcriptase. The reaction was incubated at 37 °C for 1 h. Second-strand synthesis was performed directly following first-strand synthesis and proceeded according to the method described in the Promega Protocols and Applications Guide (1990). Components for the second-strand synthesis reaction were added directly to the same tube following first-strand synthesis. Final reaction conditions for second-strand synthesis were: 50 mM Tris-HCl (pH 7.6); 100 mM KCl; 5 mM $MgCl_2$; 5 mM DTT; 0.1 mM NAD; 10 mM $(NH_4)_2SO_4$; 8 u/ml RNase H; 230

u/ml DNA polymerase 1; 5 u/ml *E.coli* DNA ligase; 50 µg/ml BSA; 0.2 mM each of dATP, dCTP, dGTP, dTTP from first-strand reaction. The reaction was incubated at 14 °C for 2 h. After heat inactivation (70 °C, 10 min), second-strand synthesis was completed by the addition of T4 DNA polymerase (2 u/µg mRNA) and incubated for 10 min at 37 °C. The ds cDNA product was phenol:chloroform extracted and purified through a QIAquick Spin column (Qiagen, Germany) according to the manufacturer's instructions. cDNA was ethanol precipitated prior to ligation to amplification adaptors.

3.3.2.2 Ligation to amplification adaptors

cDNA was blunt-end ligated to an annealed amplification adaptor set (Jepson et al., 1991). This adaptor set consisted of the following two oligonucleotides:
Oligonucleotide 1 (29-mer): 5'- ATGCTTAGGAATTCCGATTTAGCCTCATA -3'
Oligonucleotide 2 (12 mer): 5'- TATGAGGCTAAA -3'
Ligation was allowed to proceed overnight at 14 °C. After ligation, cDNA was size fractionated through a Quick-Spin, Linkers 6 column (Roche Molecular Biochemicals, Germany).

3.3.2.3 PCR amplification of cDNA

Ligated, size fractionated cDNA was PCR amplified using oligonucleotide 1 as the primer. The final reaction conditions were as follows: 1X Taq DNA Polymerase buffer (50 mm KCl, 10 mM Tris-HCl (pH 9.0), 0.1% (v/v) Triton X-100); 600 ng oligonucleotide 1; 1.25 mM each dideoxynucleotide triphosphates (dNTPs); 3.5 mM $MgCl_2$; 1 unit Taq DNA Polymerase; 1 µl ds cDNA template. PCR amplification was performed in a Hybaid Omnigene Thermal Cycler under the following conditions: 1 cycle at 73 °C for 1 min, followed by 35 cycles of 94 °C, 0.8 min; 68 °C, 1.1 min; 73 °C, 3.0 min. An aliquot of each amplified cDNA sample was analysed on a 1.5% (w/v) agarose gel in order to confirm that amplification was successful. The remainder was used for cloning.

3.3.2.4 Library construction

All individual PCR amplified cDNA samples were pooled and ethanol precipitated. cDNA was digested with 30 units EcoRI for 2.5 hours and approximately 150 ng -200 ng removed for cloning. cDNA was cloned into the EcoRI site of the Lambda ZAP II cloning vector and packaged according to the manufacturer's instructions (Stratagene, La Jolla, CA, USA).

3.3.3 Template preparation

Aliquots of the constructed leaf roll library were plated out onto solid NZY medium and single plaques randomly picked and stored in SM buffer (100 mM NaCl, 8 mM $MgSO_4.7H_2O$, 20 mM Tris-HCl pH 7.5, 0.01% gelatin) at 4 °C. The insert sizes of individual recombinant phages were examined by specific PCR amplification using the M13 reverse and T7 primers followed by 1.5% (w/v) agarose gel electrophoresis. Templates for the ESTs from the leaf roll library were prepared in two ways. Phagemids (pBluescript SK(-)) plus inserts were excised from individual phages using the ExAssist helper phage system and performed according to the manufacturer's instructions (Stratagene, La Jolla, CA, USA). Individual phagemid clones were plated out onto solid Luria Bertani (LB) medium containing 50 μg/ml ampicillin. For phagemid DNA isolation, a single colony of each clone was removed and inoculated into a 10 ml overnight culture of LB broth containing 50 μg/ml ampicillin. Phagemid DNA was isolated from a 5 ml aliquot of the overnight culture using a Rapid Plasmid Isolation Protocol (Holmes and Quigly 1981) and purified through QIAquick spin columns (Qiagen, Germany). Templates for DNA sequencing were prepared also by specific PCR amplification of cDNA inserts directly from individual phage suspensions in SM buffer, using the M13 reverse and the T7 primers. Amplified inserts were purified using QIAquick spin columns (Qiagen, Germany) prior to sequencing.

3.3.4 Sequencing

Both phagemid and amplified insert cDNA were sequenced by dye terminator cycle sequencing using either the Taq DyeDeoxy Terminator Cycle Sequencing kit (PE Applied Biosystems, Foster City, CA, USA), followed by purification through Centri-Sep Spin columns (Princeton Separations, Adelphia, NJ, USA), or the AmpliTaq DNA polymerase, FS ready reaction kit (PE Applied Biosystems, Foster City, CA, USA). In both cases, all procedures were performed according to the manufacturer's instructions. The M13 Reverse (5') primer was used to generate single-pass partial sequences for all isolated cDNAs. Cycle sequencing was performed in a Hybaid Omnigene Thermal Cycler and sequence analysis was performed using an ABI Prism 310 Genetic Analyser (PE Applied Biosystems, Foster City, CA, USA).

3.3.5 Sequence data analysis

Sequences were edited manually to remove vector and ambiguous sequences. The EST sequences were compared to the nonredundant protein databases by using the BLASTX (Altschul et al., 1990) e-mail server provided by NCBI (blast@ncbi.nlm.nih.gov). Sequences showing a Point Acceptable Mutation (PAM) 120 similarity score of over 80 were considered homologous proteins for the clones (Altschul et al., 1990) while those with scores below 80 were regarded as showing sequence similarity. The EST was identified as the protein showing the highest score among the candidate proteins.

3.4 RESULTS

3.4.1 Characteristics of the constructed LR cDNA library

The titer of the constructed LR cDNA library was 2.96 X 10^5 pfu/ml (unamplified). This titer is comparatively low, relative to the complexity of the polyploid sugarcane genome, but was considered to be sufficiently representative for preliminary analysis of the expressed genes present in sugarcane leaf roll. The titer of the amplified library was 4.2 X 10^9 pfu/ml. Blue/white plaque selection following incubation of an aliquot of the library in the presence of X-gal and IPTG revealed 95% recombinant plaques. The quality of the library was assessed by examining the insert sizes of 468 randomly selected recombinant plaques by specific PCR amplification, using the T7 and M13 reverse primers. Of the 468 selected plaques, 0.09% were found to have no inserts. Insert sizes were found to range between 400 bp and 2500 bp with an average insert size of 600 bp. Sequence analysis of 250 randomly selected clones from the library indicated an absence of contaminating rRNA sequences in the library. In addition, both full-length and near full-length sequences were detected indicating that the leaf roll cDNA library was suitable for the generation of expressed sequence tags.

3.4.2 Generation of Expressed Sequence Tags

For generation of the Expressed Sequence Tags (ESTs), only clones with an insert larger than 400 bp were selected for sequencing. Altogether 250 clones were subjected to single-run partial sequencing, 60 of these using plasmid DNA as sequencing template, and the remaining 190 using DNA obtained by specific PCR amplification of insert DNA from recombinant phages using the T7 and M13 reverse primers. The amount of template DNA used per sequencing reaction differed depending on the source. For plasmid-derived DNA, 1 µg of template was used and for PCR amplified DNA, 100ng-200ng was required. For all sequencing reactions, only the M13 reverse primer (5') was used. As the cDNA library was not a directional library, the orientation of the cDNA inserts was random. This meant that it was not known from which end (5' or 3') the clones had been sequenced. In order to identify individual clones, each of the edited sequences was translated into all six translational reading frames and compared to the nonredundant protein sequences databases in GenBank. Deduced amino acid sequence homology between a sugarcane EST and a known sequence was deemed significant if the BLASTX PAM 120 similarity score was greater than 80 (Altschul et al., 1990). All sugarcane ESTs have been deposited in the GenBank database for ESTs, dbEST.

3.4.3 Sequencing template

A small investigation was conducted to determine whether variation occurred in the amino acid sequence homology results when different forms of template DNA were used for sequence analysis. Conventionally, high quality plasmid DNA is the preferred form of template for sequencing reactions. However, the *in vivo* excision of phagemids from recombinant cDNA clones housed in a λZAP II vector and the subsequent isolation of phagemid DNA is a time-consuming process which can negatively impact on large-scale sequencing efforts. It has been recognised that while direct sequencing of recombinant clones without isolation of plasmid DNA is a favourable alternative, results are often inconsistent. This is because the amount and quality of template DNA generated during PCR amplification of inserts may vary, which in turn can lead to unreliable results. In this study, a comparison was performed between sequencing results obtained using template DNA derived either from recombinant plasmids or PCR-amplified cDNA inserts from recombinant phages. Four different clones were selected arbitrarily. All sequencing reactions and sequence analysis were performed at the same time to minimise experimental error. It is evident that the length of the analysed sequences is similar, regardless of template source (Table 3.1). After editing of sequences to remove the vector component, a final analysed sequence length of approximately 400 bp was obtained for both plasmid and PCR-amplified insert DNA.

Table 3.1 Comparison between the length of analysed DNA sequence and BLASTX PAM120 homology score using two different sources of template DNA

Clone	DNA template	Length of analysed sequence (bases)	PAM120 homology score	Putative identification
B63	PCR-amplified fragment	483	409	SuSy
pB63	plasmid	547	486	SuSy
B81	PCR-amplified fragment	541	136	glutathione S-transferase
pB81	plasmid	410	141	glutathione S-transferase
A73	PCR-amplified fragment	420	160	small nuclear ribonucleoprotein E homolog C29
pA73	plasmid	403	83	small nuclear ribonucleoprotein E homolog C29
B21	PCR-amplified fragment	457	-	none
pB21	plasmid	459	-	none

3.4.4 Identification of genes

Analysis of 250 randomly selected clones revealed that 38% were homologous to peptide sequences present in the NCBI nonredundant protein databases (Tables 3.2 and 3.3). Of the remaining 62% of the ESTs, 49% did not appear to exhibit sequence similarity to any sequence on the databases according to the search criteria used, and thus were interpreted as possibly representing new genes not only in sugarcane but also in all organisms. The other 13% did not show significant homology to previously identified genes in the databases (ie: similarity scores below 80) and thus were putatively identified on the basis of sequence similarity only. Of the 250 clones analysed, 25% showed significant deduced amino acid sequence homology to previously identified plant genes (Table 3.2). Ten clones, although similar to plant genes, did not have PAM 120 scores above 80 and thus could not be considered as homologous. As only 10 previously identified sugarcane genes were registered with GenBank at the time of the database searches (commencing in 1996), all putative clone identities to plant genes came from plants other than sugarcane. Of the 62 identified homologous clones, 31% showed homology to monocotyledenous plant species such as rice, maize and wheat. As expected, these proteins gave high similarity scores. One hundred and thirty seven ESTs (54%) showed sequence similarity to previously identified genes from species other than higher plants, and 20% of these were considered homologous (Table 3.3). The targeted species were widely distributed from bacteria to human.

Table 3.2 Sugarcane ESTs with sequence homology or similarity to known plant genes

The EST no. is the accession number assigned by dbEST. The numbers in the columns designated ID, Similar, and Overlap refer to the number of identical (ID) or similar (Similar) amino acids in a region of a particular length (Overlap). The column designated Organism refers to the source of the protein that exhibits homology or similarity to the sugarcane EST.

EST no.	Putative identification and GenBank accession	ID	Similar	Overlap	Score	Organism
AA080648	60S ribosomal protein L5 [P42796]	69	73	84	329	*Arabidopsis thaliana*
AA080649	60S ribosomal protein L5 [P46287]	43	43	44	202	*Medicago sativa*
AA080650	calcium-dependent protein kinase [P28583]	33	42	49	200	*Glycine max*
AA080655	vacuolar H$^+$- ATPase subunit B [U07052]	55	56	60	280	*Gossypium hirsutum*
AA080670	protein kinase [L27821]	10	19	24	64	*Oryza sativa*
AA080674	3-oxoacyl-[acyl-carrier protein] reductase [S22417]	46	53	61	238	*Brassica napus*
AA080657	unknown [687677]	26	35	61	105	*Arabidopsis thaliana*
AA080659	athila ORF1 [AC007505]	18	25	48	76	*Arabidopsis thaliana*
AA080580	sucrose synthase [S22537]	32	36	52	171	*Oryza sativa*
AA080581	receptor-like protein kinase [Z17991]	19	30	48	90	*Arabidopsis thaliana*
AA080582	ADP-ribosylation factor [S49325]	67	68	69	360	*Zea mays*
AA080583	H2B histone [577825]	17	31	62	75	*Zea mays*
AA080585	translation elongation factor eEF-1 beta-A1 chain [S37103]	28	33	36	151	*Arabidopsis thaliana*
AA080586	enolase [P42895]	97	101	105	511	*Zea mays*
AA080589	casein kinase II, alpha chain [P28523]	74	75	78	404	*Zea mays*

AA080590	acyl-CoA-binding protein [U35015]	50	55	68	260	*Gossypium hirsutum*
AA080599	protein phosphatase 2C [S55457]	51	60	87	228	*Arabidopsis thaliana*
AA080605	60S ribosomal protein L32 [Z17739]	40	43	54	205	*Arabidopsis thaliana*
AA080606	small nuclear ribonucleoprotein E homolog C29 [P24715]	14	17	19	72	*Medicago sativa*
AA080610	sucrose synthase [X81974]	40	44	52	202	*Beta vulgaris*
AA080615	pyruvate kinase, plastid [S44287]	27	27	43	126	*Nicotiana tabacum*
AA080634	sucrose synthase [JT0280]	55	61	68	301	*Triticum aestivum*
AA080636	GTP-binding protein [D12542]	69	77	82	356	*Pisum sativum*
AA080640	mitochondrial processing peptidase [X80236]	45	59	72	237	*Solanum tuberosum*
AA080642	stage III sporulation protein [S39321]	57	70	82	314	*Arabidopsis thaliana*
AA080646	glutathione S-transferase [P46422]	31	43	82	138	*Arabidopsis thaliana*
AA080668	proteasome C2 subunit [D37886]	73	77	78	400	*Oryza sativa*
AA269154	pectin methylesterase [Y08155]	49	63	90	261	*Melandrium album*
AA269161	auxin response factor 1 [U83245]	18	22	29	90	*Arabidopsis thaliana*
AA269164	hypothetical protein (beta-1,3-glucanase) [S31196]	29	42	61	161	*Solanum tuberosum*
AA269165	vacuolar processing enzyme precursor [P49045]	35	44	63	189	*Glycine max*
AA269289	alcohol dehydrogenase [L08591]	72	75	94	384	*Zea mays*
AA269290	disease resistance protein RPM1 [X87851]	18	26	49	86	*Arabidopsis thaliana*
AA269291	chloroplast 30S ribosomal protein S7 [P46292]	32	38	44	159	*Cuscuta europaea*
AI216928	hypothetical polyprotein [S57908]	6	12	16	42	*Oryza sativa*
AI216930	ER lumen protein retaining receptor [P35402]	15	17	25	77	*Arabidopsis thaliana*
AI216931	unknown protein [AC004138]	13	21	24	89	*Arabidopsis thaliana*
AA269292	cathepsin B [X66012]	14	17	25	79	*Triticum aestivum*
AA269294	sucrose synthase [JT0280]	32	39	44	189	*Triticum aestivum*
AA525640	actin depolymerizing factor [X97726]	70	73	75	371	*Zea mays*
AA525645	5' end not determined experimentally [U68408]	21	27	37	119	*Zea mays*
AA525649	glutathione S-transferase [P42761]	39	54	77	199	*Arabidopsis thaliana*
AA525651	3-oxoacyl-[acyl-carrier protein] reductase precursor [P28643]	84	93	97	441	*Cuphea lanceolata*
AA525652	bicolor membrane intrinsic (Mip 1) protein [U87981]	36	37	41	203	*Sorghum bicolor*

AA525655	pyrophosphate-fructose 6-phosphate 1-phosphotransferase (PFP) beta subunit [P21343]	17	17	21	89	*Solanum tuberosum*
AA525658	UDP-glucose dehydrogenase [U53418]	72	79	88	408	*Glycine max*
AA525660	RNA helicase isolog [AC002337]	15	18	21	86	*Arabidopsis thaliana*
AA525661	2-oxoglutarate/malate translocator [D45075]	67	72	89	374	*Panicum miliaceum*
AA525664	acetyl-CoA carboxylase [U10187]	7	11	18	45	*Triticum aestivum*
AA525666	translation initiation factor 5A [Y07920]	89	90	92	480	*Zea mays*
AA525669	Clp protease [AF032123]	15	19	27	95	*Arabidopsis thaliana*
AA525677	ras-related protein RIC1 (GTP binding protein) [S66160]	37	38	41	179	*Oryza sativa*
AA525679	aspartic proteinase precursor [P42211]	75	79	89	390	*Oryza sativa*
AA525680	similar to hypothetical protein from *A. thaliana* [AC002986]	10	17	23	62	*Arabidopsis thaliana*
AA525686	germin-like protein [U75205]	19	22	32	90	*Arabidopsis thaliana*
AA525688	aspartic proteinase precursor [P42211]	58	64	68	327	*Oryza sativa*
AA525692	unknown protein [AC004122]	27	39	55	143	*Arabidopsis thaliana*
AA525697	pectin methylesterase [Y08155]	34	38	59	178	*Melandrium album*
AA577634	farnesyl-diphosphate farnesyltransferase (squalene synthase) [JC5031]	30	48	70	160	*Glycyrrhiza glabra*
AA577635	cellulase homolog OR16pep [S71215]	15	19	23	86	*Arabidopsis thaliana*
AA577636	contains similarity to *S. cerevisiae* hypothetical protein YOR197w [3152597]	18	24	29	105	*Arabidopsis thaliana*
AA577639	lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional enzyme [AF003551]	36	41	42	181	*Zea mays*
AA577641	protein kinase isolog [U90439]	32	40	50	177	*Arabidopsis thaliana*
AA577644	Bowman-Birk protease inhibitor [2123385A]	12	14	26	71	*Pisum sativum*
AA577653	triose-phosphate isomerase, cytosolic [P12863]	60	60	63	302	*Zea mays*
AA577658	GDP-associated inhibitor [Y07961]	71	81	94	376	*Arabidopsis thaliana*
AA577659	chloroplast 50S ribosomal protein L32 [P12197]	10	16	26	44	*Oryza sativa*
AA577663	voltage-dependent anion-selective channel protein (VDAC) [34 kDa outer mitochondrial membrane protein, porin] [P42055]	45	59	77	233	*Solanum tuberosum*
AA577664	hypothetical protein [Z97339]	31	35	39	175	*Arabidopsis thaliana*

AA577666	nucleolar histone deacetylase HD2 [U82815]	22	23	24	111	*Zea mays*
AA577669	unknown protein [U93215]	20	27	38	100	*Arabidopsis thaliana*
AA525685	5-methyltetrahyropteroyl-triglutamate-homocysteine S-methyltransferase [S57636]	19	20	24	100	*madagascar periwinkle*
AI216932	cellulose synthase [U58284]	52	60	64	303	*Gossypium hirsutum*
AA577633	ATP synthase 6 kD subunit, mitochondrial [P80497]	17	17	21	103	*Solanum tuberosum*

Table 3.3 Sugarcane ESTs with sequence homology or similarity to non-plant genes

The EST no. is the accession number assigned by dbEST. The numbers in the columns designated ID, Similar, and Overlap refer to the number of identical (ID) or similar (Similar) amino acids in a region of a particular length (Overlap). The column designated Organism refers to the source of the protein that exhibits homology or similarity to the sugarcane EST.

EST no.	Putative identification and GenBank Accession	ID	Similar	Overlap	Score	Organism
AA080647	yeast hypothetical 16.2 kd protein [P36053]	41	57	87	245	*Saccharomyces cerevisiae*
AI376340	LIM homeobox protein [Z97340]	29	29	29	146	*Caenorhabditis elegans*
AA080651	ribosomal protein L30 [B24028]	34	38	47	189	*Rattus rattus*
AA080653	alpha-fetoprotein enhancer-binding protein [A41948]	8	10	15	51	*Homo sapiens*
AA080669	mouse B-cell receptor CD22-Beta precursor [P35329]	7	8	10	35	*Mus musculus*
AA080671	influenza virus hemagglutinin 5' epitope tag [S71745]	21	22	29	97	*Saccharomyces cerevisiae*
AA080672	erythrocyte membrane protein band 4.2 [U04056]	14	18	29	68	*Mus musculus*
AA080673	ornan sperm protamine P1 [P35307]	12	17	31	44	*Ornithorhyn-chus anatinus*
AA080675	hypothetical protein YJL076W [S56852]	8	16	26	36	*Saccharomyces cerevisiae*
AA080676	proline-rich polypeptide precursor [A42663]	14	17	43	45	*Rattus rattus*
AA080677	putative vitellogenin receptor [U13637]	6	6	7	41	*Drosophila melanogaster*
AA080678	argininosuccinate synthase [P13257]	12	18	31	62	*Methanosarcina barkeri*
AA080656	mucin [M188878]	9	11	19	55	*Homo sapiens*
AA080658	Cii-1=beta-toxin [165350]	7	13	23	40	*Centruroides infamatus*
AA080660	GAG polyprotein [P31622]	13	17	27	54	*sheep pulmonary adenomatosis virus*
AA080661	OATL1 [L08238]	6	7	13	39	*Homo sapiens*

AA080663	human U1 small ribonucleoprotein [P09234]	8	11	17	50	*Homo sapiens*
AA080664	Fin29 [AB007447]	14	19	30	81	*Homo sapiens*
AA080665	60S ribosomal protein L28 [P46779]	13	25	31	76	*Homo sapiens*
AA080667	PAC1 protein [P39946]	12	15	21	69	*Saccharomyces cerevisiae*
AA080587	oligodendrocyte-specific proline-rich protein 2 [C55663]	16	22	44	75	*Homo sapiens*
AA080591	drome broad-complex core-NS-Z3 protein [Q01293]	8	9	19	42	*Drosophila melanogaster*
AA080592	similar to *Saccharomyces cerevisiae* ORF YCR028 [D89224]	10	13	18	70	*Shizosaccharomyces pombe*
AA080595	competence protein S [P80355]	7	14	20	41	*Bacillus subtilis*
AA080598	F55A11.4 [Z72511]	18	30	59	87	*Caenorhabditis elegans*
AA080604	60S ribosomal protein L28 [P46779]	12	23	31	68	*Homo sapiens*
AA080608	gamma-glutamyl transpeptidase 2 [P36268]	15	18	33	71	*Homo sapiens*
AA080609	yeast probable ribosomal protein in VMA3-RIP1 intergenic region [P39990]	38	46	76	163	*Saccharomyces cerevisiae*
AA080611	hypothetical protein L9122.5 [S59413]	8	17	28	42	*Saccharomyces cerevisiae*
AA080612	tenascin-X precursor [A40701]	6	9	16	38	*Homo sapiens*
AA080617	surface antigen [M92048]	15	23	44	68	*Trypanosoma cruzi*
AA080618	mec1p [U31109]	8	15	20	46	*Saccharomyces cerevisiae*
AA080619	similar to calcium channel alpha subunit [U61951]	9	11	21	34	*Caenorhabditis elegans*
AA080621	probable G protein-coupled receptor GPR21 [Q99679]	14	23	37	70	*Homo sapiens*
AA080622	faf=fat facets gene [A49132]	7	7	8	40	*Drosophila melanogaster*
AA080624	putative gtg start codon [X90711]	7	11	17	41	*Bordetella pertussis*
AA080625	51C surface protein [M65164]	6	8	16	34	*Paramecium tetraurelia*
AA080626	hypothetical 92.1 kD protein C24H6.03 in chromosome 1 [Q09760]	17	21	35	90	*Schizosacchar-omyces pombe*
AA080628	phosphoribosylformylglycin-amidine synthase [P35421]	7	8	8	38	*Drosophila melanogaster*
AA080629	ORF 1130 [U20247]	6	8	12	41	*Dichelobacter nodosus*
AA080630	human thrombospondin 1 precursor [P07996]	7	9	12	40	*Homo sapiens*
AA080632	URF5 [1204259B]	16	22	42	69	*Chlamydomonas reinhardtii*
AA080631	ERCC5 [D16305]	6	10	14	37	*Homo sapiens*
AA080635	murine erythroleukemia cardiac calcium channel [U17869]	5	10	13	35	*Mus musculus*

AA080637	ORF gene product [X95373]	7	12	19	40	*Plasmodium falciparum*
AA080638	latent transforming growth factor beta-binding protein 3 precursor [A57293]	8	9	15	42	*Mus musculus*
AA080641	aminopeptidase [U35646]	28	46	82	142	*Mus musculus*
AA080643	TO8A11.2 [Z50875]	25	31	43	137	*Caenorhabditis elegans*
AA080644	peroxisomal membrane protein 47B [U53145]	19	31	46	105	*Candida boidinii*
AA080645	coded for by *C. elegans* cDNA [U50191]	6	10	14	35	*Caenorhabditis elegans*
AA269152	DNA-binding protein [JQ1058]	18	19	20	93	*Mus musculus*
AA269153	F15B9.7 [Z78018]	6	10	12	42	*Caenorhabditis elegans*
AA269155	HIV-EP2 enhancer-binding protein [X65644]	9	9	15	44	*Homo sapiens*
AA269156	60S ribosomal protein L13A [P40429]	27	37	76	123	*Homo sapiens*
AA269157	copper transporting P-type ATPase [U38477]	6	10	13	39	*Mus musculus*
AA269159	unknown [Z69239]	25	29	40	129	*Schizosacchar -omyces pombe*
AA269160	small nuclear ribonucleoprotein [P43330]	23	29	30	124	*Homo sapiens*
AA269163	glycine-rich [U23453]	17	25	42	93	*Caenorhabditis elegans*
AA269166	ZK593.7 [Z69385]	44	62	81	242	*Caenorhabditis elegans*
AA269167	zinc finger protein ZMS1 [P46974]	10	17	21	62	*Candida boidinii*
AA269168	F25F2.2 [Z35599]	8	15	28	43	*Caenorhabditis elegans*
AA269170	ryanodine receptor (skeletal muscle) [P21817]	7	10	12	43	*Homo sapiens*
AA269171	RO7E5.13 [Z32683]	41	53	75	227	*Caenorhabditis elegans*
AA269172	HC1 ORF [X66285]	7	12	23	57	*Mus musculus*
AA269173	BmGATA beta isoform 3 [U16274]	10	11	21	43	*Bombyx mori*
AA269174	pyruvate dehydrogenase E1 component, alpha subunit [D90915]	61	70	93	313	*Synechocystis spp.*
AA269175	glycoprotein GP330, renal-rat (fragments) [A30363]	7	7	9	37	*Rattus rattus*
AA269176	ZC374.2 [Z72518]	7	9	12	50	*Caenorhabditis elegans*
AA269177	guanine nucleotide-binding protein G(O), alpha subunit [P30033]	5	8	8	43	*Rattus rattus*
AI216928	ORF N118 [D84656]	20	37	54	119	*Schizosaccharo-myces pombe*
AA269293	Xenla DG42 protein [P13563]	7	11	17	50	*Xenopus laevis*
AA269295	ran=25 kda ras-related protein [239838]	7	8	8	41	*Homo sapiens*
AA269296	POL1M genome polyprotein [P03299]	5	11	15	37	*human poliovirus 1*
AA269297	f549 [AE000312]	12	17	28	52	*Escherichia coli*

AA269298	RNA-directed RNA polymerase (ORF 1A) [P19751]	8	11	17	42	*murine hepatitis virus*
AA269299	AT-motif binding factor [D26046]	7	10	15	40	*Mus musculus*
AA525639	thioredoxin [P34723]	23	29	46	124	*Penicillium chrysogenum*
AA525641	dTDP-glucose 4-6-dehydratase [D90911]	35	46	53	198	*Synechocystis sp.*
AA525642	hypothetical protein ZC84.1 [S28291]	11	15	26	55	*Caenorhabditis elegans*
AA525643	von Willebrand factor [L76227]	5	12	20	46	*Canis familiaris*
AA525646	envelope protein (human immunodeficiency virus type 1) [U20673]	8	10	26	48	*Homo sapiens*
AA525647	spliceosome associated protein [U41371]	26	37	54	125	*Homo sapiens*
AA525648	lozenge [U47849]	16	19	36	102	*Drosophila melanogaster*
AA525650	pericentrin [P48725]	16	22	51	61	*Mus musculus*
AA525653	human BDNF/NT-3 growth factors receptor precursor [Q16620]	12	15	25	69	*Homo sapiens*
AA525654	metallothionein (MT) [P07216]	9	11	21	61	*Pleuronectes platessa*
AA525659	tropomyosin 1, fusion protein 33 [P49455]	20	28	68	75	*Drosophila melanogaster*
AA525662	chaperone protein SEFB precursor [P33387]	6	7	13	34	*Salmonella enteritidis*
AA525665	adenosine kinase [U33936]	38	45	65	208	*Homo sapiens*
AA525668	hypothetical 24.8 kD protein in FAA3-BET1 intergenic region [P40555]	13	22	35	74	*Saccharomyces cerevisiae*
AA525671	coded for by *C. elegans* cDNA yk89e9.5 [U50199]	25	50	76	133	*Caenorhabditis elegans*
AA525672	human small proline-rich protein 2B [P35325]	6	6	11	41	*Homo sapiens*
AA525673	yeast verprolin [P37370]	8	10	15	41	*Saccharomyces cerevisiae*
AA525674	Ig alpha chain C region [S03297]	5	6	8	29	*Gorilla gorilla*
AA525676	DNA polymerase [D50489]	11	14	26	53	*Hepatitis B virus*
AA525681	human histone H3.1 [P16106]	33	34	41	167	*Homo sapiens*
AA525682	spike protein (porcine respiratory corona virus) [D00658]	7	8	12	43	*Pig*
AA525687	hypothetical protein C23D3.15 in chromosome 1 [AB004534]	12	19	29	68	*Schizosaccharomyces pombe*
AA525689	sequence 3 from Patent WO 8912462 [I11349]	5	5	8	31	-
AA525690	endothelin-2 precursor (ET-2) [P12064]	9	11	32	40	*Canis familiaris*
AA525691	yeast hypothetical 32.8 kd protein in NCE3-HHT2 intergenic region [P53965]	28	43	63	167	*Saccharomyces cerevisiae*
AA525693	rjs [AF061529]	13	22	35	61	*Mus musculus*

AA525694	ORF N150 [D84656]	10	14	23	60	*Schizosaccharo-myces pombe*
AA577629	hypothetical 337.6 kD protein T20G5.3 in chromosome III [P34576]	6	7	10	35	*Caenorhabditis elegans*
AA577630	yeast hypothetical 65.3 kD protein in PRE3-SAG1 intergenic region [P47082]	20	38	71	89	*Saccharomyces cerevisiae*
AA577631	similarity to *C. elegans* retinoic acid receptors [Z92825]	19	28	54	69	*Caenorhabditis elegans*
AA577637	alk 8 [Y14766]	14	20	34	68	*Candida albicans*
AA577640	hypothetical protein in OGT 5' region [P46133]	12	17	28	69	*Escherichia coli*
AA577647	GRK4c [X97568]	20	30	71	79	*Rattus norvegicus*
AA577648	similar to hypothetical proteins [Z99115]	16	20	39	68	*Bacillus subtilis*
AA577649	SNAP-25 interacting protein hrs-2 [U87863]	7	9	12	58	*Rattus norvegicus*
AA577651	membrane protein CD40 [U57745]	9	14	22	41	*Bos taurus*
AA577654	hypothetical protein [D63999]	11	15	33	70	*Synechocystis sp.*
AA577655	C32A3.1 [Z48241]	12	15	25	44	*Caenorhabditis elegans*
AA577660	mataxin [AF059277]	17	28	60	71	*Mus musculus*
AA577661	60S ribosomal protein L10A [P53026]	60	82	112	311	*Mus musculus*
AA577667	probable ubiquitin carboxyl-terminal hydrolase [ubiquitin-specific processing protease] [P34547]	8	11	17	54	*Caenorhabditis elegans*
AA577668	transmembrane glycoprotein [M76753]	7	9	13	43	*human T-cell lymphocyte virus type 1*

During the course of sequencing analysis, several redundant clones were detected. These clones are presented in Table 3.4. The frequency of these clones in the total pool of leaf roll ESTs analysed ranged from 0.8% to 4%. Analysis of the database search results indicated that the DNA sequences of the redundant clones were not identical, that is, they did not simply represent copies of the same clone in the cDNA library (data not shown). In most cases the sequences were homologous to different regions of the gene sequence in the database, although in some cases, small regions of sequence overlap between clones was observed (data not shown).

Table 3.4 Identified leaf roll cDNA clones showing redundancy

Putative genes	Total cDNA clones	Frequency [a]
	#	%
Ribosomal proteins	10	4.0
Protein kinases	4	1.6
Sucrose synthase	4	1.6
Small nuclear ribonucleoprotein	3	1.2
Aspartic proteinase precursor	2	0.8
Glutathione S-transferase	2	0.8
GTP-binding protein	2	0.8
3-oxoacyl-[acyl-carrier protein] reductase	2	0.8
Pectin methylesterase	2	0.8

[a] calculation based on 250 cDNA clones analysed

A slight exception occurred for the ESTs homologous to SuSy. Four clones were identified as being homologous to SuSy. These clones were first sequenced with the M13 Forward primer in order to obtain the full-length sequences of the four individual cDNA clones. Sequence overlap analysis indicated that high sequence homology occurred between clones AA080610 and AA080580, and AA080634 and AA269294 (Table 3.2). Consensus sequences for each of these pairs was generated using Sequence Navigator (PE Applied Biosystems). Overlap alignment of these two consensus sequences generated a total sequence length of 1450 bp with a near-identical overlap of 286 bp. GenBank database searches with the 1450 bp fragment revealed high homology to SuSy isoform I (data not shown). These results suggest that all four SuSy clones represent only one of the known SuSy isoforms. For all of the redundant clones identified, it is suggested that it may be indicative of an increased expression of those genes in the leaf roll.

3.4.5 Functional identification of sugarcane ESTs

All identified ESTs were categorised into general biochemical and metabolic function (Fig. 3.1). The leaf roll cDNA clones exhibited homology to a broad diversity of genes, including enzymes and proteins associated with ubiquitous metabolic pathways, structural proteins and components of transcriptional and translational apparatus. The largest number of clones (35%) was found to encode many proteins as yet uncharacterised. There are several high-throughput gene sequencing programmes currently in progress and many expressed sequences deposited in the GenBank databases by these groups do not yet have an identity. This results in many putative identities to unknown or hypothetical proteins. Of the remaining 65% of clones that were identified, 12.4% were enzymes. Sucrolytic enzymes were the most common, with nine clones representing six different enzymes being identified. These included key regulatory enzymes such as SuSy (AA080580, AA080610, AA080634, AA269294) and triose phosphate isomerase (AA577653). Several other metabolic pathways were represented including the citric acid cycle, fatty acid metabolism, anaerobic metabolism and amino acid biosynthesis. A further 10.8% of ESTs were involved in protein modification and 9.7% in protein synthesis. These included eight different ribosomal proteins, represented by 10 individual clones, and a variety of protein kinases.

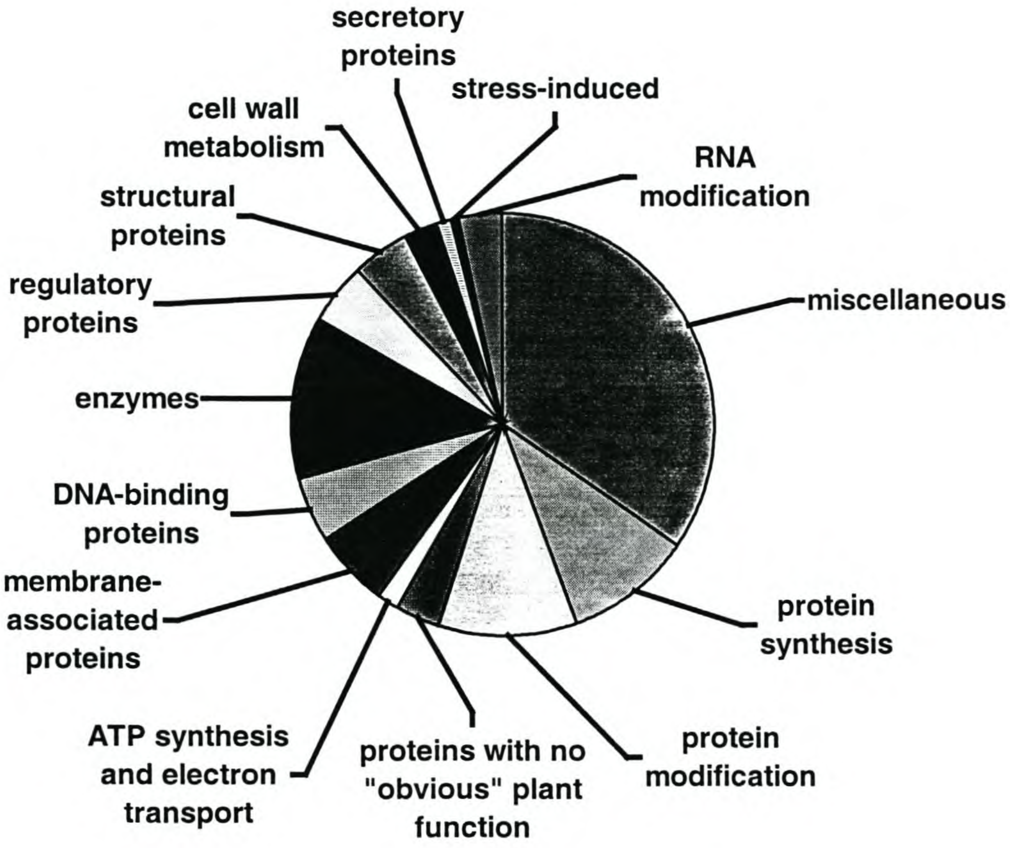


Fig. 3.1 Classification of sugarcane leaf roll ESTs according to biochemical and metabolic function

All cDNA clones for which an identity had been assigned through database searches were included in the analysis, regardless of similarity score.

Membrane-associated proteins contributed a further 5.9% of the total identified clones. Fewer genes were involved in DNA binding (4.9%), regulation (4.9%), structural proteins (4.3%), RNA modification (3.2%), cell wall metabolism (2.7%), secretory proteins (1.1%) and ATP synthesis and electron transport (1.6%). Only one clone was identified as being stress- or defense-related (disease resistance protein RPM1, AA269290). A small percentage of clones (3.2%) were identified as having sequence similarity to proteins involved in functions that are not known to exist in plants. For example, ESTs were putatively identified for *Caenorhabditis elegans* retinoic acid receptors (AA577631), *Canis familiaris* von Willebrand factor (AA525643), human alpha-fetoprotein enhancer-binding protein (AA080653) and several others which cannot be immediately assigned probable functions in plants.

3.5 DISCUSSION

The use of an Expressed Sequence Tag (EST) approach was found to be a very efficient and successful way of identifying genes in sugarcane. Of all the leaf roll cDNA clones identified by database homology searches, 38% had statistically significant similarities to known gene sequences. This value is comparable with that observed for the analysis of clones from maize endosperm and seedling cDNA libraries (39.3%, Shen et al., 1994) but was slightly greater than results from EST projects using cDNA libraries prepared from maize leaf (20%, Keith et al., 1993), various tissues and growth stages in rice (25%, Yamamoto and Sasaki 1997) and equal portions of poly(A^+) RNA from etiolated seedlings, roots, leaves and flowering inflorescences of *Arabidopsis* (32%, Newman et al., 1994). Several reasons have been cited for the apparent differences in results between various EST projects. For example, van de Loo and co-workers have indicated significantly higher values of identification when the tissue used for cDNA library construction was specialised for processes involving well-characterised classes of proteins (van de Loo et al., 1995). In addition, it has been shown that sequencing from the 5' terminus of the mRNA instead of the 3' is more informative, and thus the use of directionally cloned cDNA libraries will result in more significant matches (Shen et al., 1994). In this study, a non-directional cDNA library was prepared so the relatively high percentage of clone identification is probably related to the use of leaf roll tissue for library construction. The leaf roll is the meristematic region of the plant and is metabolically highly active.

It is expected that a high proportion of the genes expressed in the leaf roll are involved in core "housekeeping" metabolic processes, for which DNA sequence information is available on international databases. However, it should be noted that a considerable proportion of clones with significant homology to sequences in the database (20%) have been identified on the basis of homology to non-plant genes. It is possible that these gene sequences have not been well characterised in plants. Due to the rapid growth in numbers of partially sequenced or completely sequenced animal and yeast genes it is likely that there will always be a significant proportion of sugarcane (and other plant) genes identified by homology to non-plant genes. During the course of this study it was also found that routine resubmission of clones with no sequence similarity usually resulted in several more identifications, simply due to new additions to the databases in the interim. It is likely that with the continual rapid escalation of databank submissions from a whole array of organisms, the rate of genes identified will increase simply based on repeated database searches.

During cDNA library construction, it is assumed that all cDNAs present are equally likely to be cloned. The relative frequency of cDNAs in sugarcane leaf roll tissue would therefore reflect the steady-state levels of the mRNA in the leaf roll. Thus the analysis of cDNA abundance may not only identify fundamental housekeeping genes, but also tissue-specific genes. Due to the small sample size of 250 clones in this study, random sequencing resulted primarily in the identification of genes belonging to the superabundant and abundant classes. In order to identify rare genes using this approach it will be necessary to either sequence all the clones in the library, or to prepare a normalized library. However, the high cost both in resources and labour required for large-scale sequencing of total cDNA libraries make it an unpractical option for many small laboratories.

A variety of studies have shown that the composition of clones identified in cDNA libraries reflects the regulation of gene expression related to differentiation, growth condition, or environmental stress. In a recent review of the Rice Genome Project (Yamamoto and Sasaki 1997) results were presented from EST identification of clones from a variety of tissues subjected to different growth conditions. This research has indicated, for example, that many ribosomal proteins and histone genes were found in growth-phase callus while genes encoding globulin and seed storage proteins such as glutelin and prolamine were identified in ripening panicles. Similarly, in developing castor endosperm a significant proportion of identified clones showed homology to storage proteins or components of the protein biosynthetic apparatus (van de Loo et al., 1995). In this study, the distribution of identified genes between the various metabolic pathways indicated that in sugarcane leaf roll genes involved in protein synthesis, protein modification and glycolysis were the most abundant (Fig. 3.1). In addition, there was also a significant proportion of genes coding for structural and cell wall proteins. These results probably reflect the high metabolic rate of the leaf roll. In addition, it was not surprising that only one clone was identified as being stress-induced (disease resistance protein, RPM1). As the leaf roll is protected by several leaf sheaths it is not normally subject to insect or pathogen attack and will therefore not be adversely affected by environmental stresses except under extreme conditions. Some unexpected genes were also detected. Two clones were identified with homologies to a germin-like protein and a stage III sporulation protein, both involved in processes not considered to occur in sugarcane. A similar phenomenon has been observed in maize where proteins involved in nodulation and other processes specifically present in legumes were identified (Shen et al., 1994). These authors suggested that genes with specific functions in some species may have been "borrowed" through evolution to form new genes with different functions, or which simply share some common functional domain.

During the course of the sequencing of the 250 cDNA clones it was found that several types of clones were identified more than once. It is acknowledged that, compared to many other EST projects, a sample size of 250 is very small. It is also assumed that during the construction of the cDNA library, the PCR amplification of the cDNA was proportional and thus the library is representative of the mRNA pool. On this basis, it may be inferred that the occurrence of multiple copies of specific genes may be indicative of their relative frequency and reflect possible trends in level of expression in the leaf roll. Ten of the ESTs showed similarity to eight different ribosomal proteins (Table 3.4). Seven of these were large subunit proteins, one was a small subunit protein and it also included two chloroplast ribosomal proteins. This result was not unexpected due to the vigorous growth state of the leaf roll. Ribosomal proteins are fundamental proteins for living systems and are thought to play a specific regulatory role during development. Many ribosomal genes have been identified in growth-phase callus of rice (Yamamoto and Sasaki 1997) so it seems likely that in sugarcane, ribosomal proteins would be specifically involved in differentiation and growth in the meristematic leaf roll region. Of particular interest in sugarcane is the identification of clones homologous to the SuSy gene. Expression of SuSy in the leaf roll was found to be quite high (1.6% of total genes identified) compared to 0.6% expression in rice endosperm (Liu et al., 1995). Although the reaction catalysed by SuSy is readily reversible, there is evidence that it is primarily involved in the breakdown of sucrose (Kruger 1990). It has been shown that in actively growing tissues where there is high demand for hexose sugars as respiratory substrates, SuSy activity is high (Kruger 1990). The apparent high expression of SuSy in sugarcane leaf roll could therefore be expected to be primarily related to the breakdown of sucrose in order to meet the demand for respiratory metabolites. The homology search results indicate that all the SuSy ESTs might be from the same expressed gene. However, more research is needed to establish whether this is the case. It is interesting to note that the sugarcane cDNA exhibited the highest homologies to the SuSy gene sequences from dicotyledonous species, despite the presence of SuSy gene sequences from other monocotyledonous plants in the database. The reasons for this observation are not immediately apparent. Other clones that were identified more than once could also be related to the active metabolic state of the leaf roll (Table 3.4). For example, expression of pectin methylesterase is related to cell wall biosynthesis during cell division. Likewise, 3-oxoacyl-[acyl-carrier protein] reductase expression is essential for cell membrane biosynthesis. Further work aimed at analysing expression profiles of leaf roll cDNA clones using macroarrays is currently in progress. These results will supplement the trends observed from the random sequencing.

No similar work on the construction of an EST database has yet to be reported for sugarcane.

This research has indicated that genes may be easily identified in sugarcane and has provided information about the metabolic state of the leaf roll, independent of the complexity of the sugarcane genome. It has also provided a resource of gene sequence information for sugarcane that may be applied to sugarcane biotechnology research. Further work is underway to develop an EST database for mature internodal tissue, the region in the plant where sucrose accumulation occurs.

3.6 REFERENCES

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C and Venter JC (1992) Sequence identification of 2375 human brain genes. Nature 355: 632-634.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR and Venter JC (1991) Complemetary DNA sequencing: expressed sequence tags and the human genome project. Science 252:1651-1656.

Albert HH, Carr JB and Moore PH (1995) Nucleotide sequence of sugarcane polyubiquitin cDNA. Plant Physiol 109 (1): 337.

Altschul SF, Gish W, Miller W, Myers EW and Lipman D (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Bugos RC and Thom M (1993a) A cDNA encoding a membrane protein from sugarcane. Plant Physiol 102: 1367.

Bugos RC and Thom M (1993b) Glucose transporter cDNAs from sugarcane. Plant Physiol 103: 1469-1470.

da Silva J, Honeycutt RJ, Burnquist W, Al-Janabi SM, Sorrells ME, Tanksley SD and Sobral BWS (1995) *Saccharum spontaneum* L. 'SES 208' genetic linkage map combining RFLP- and PCR-based markers. Mol Breeding 1: 165-179.

Gallo-Meagher M and Irvine JE (1996) Herbicide resistant sugarcane plant containing the *bar* gene. Crop Sci 36: 1367-1374.

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C and Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): Genome organisation in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Henrik AH, Martin T and Sun SSM (1992) Structure and expression of a sugarcane gene encoding a housekeeping phosphoenolpyruvate carboxylase. Plant Mol Biol 20: 663-671.

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D and Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana.* Plant J 4: 1051-1061.

Holmes DS and Quigly M (1981) A rapid boiling method for the preparation of bacterial plasmids. Anal Biochem 114: 193-197.

Jepson I, Bray J, Jenkins G, Schuch W and Edwards K (1991) A rapid procedure for the construction of PCR cDNA libraries from small amounts of plant tissue. Plant Mol Biol Reporter 9(2): 131-138.

Keith CS, Hoang DO, Barrett BM, Feigelman B, Nelson MC, Thai H and Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol 101: 329-332.

Kruger NJ (1990) Carbohydrate synthesis and degradation. In: Dennis DT and Turpin DH (eds) Plant Physiology, Biochemistry and Molecular Biology. Longman Singapore Publishers (Pte) Ltd., Singapore, pp 59-76.

Liu J, Hara C, Umeda M, Zhao Y, Okita TW and Uchimaya H (1995) Analysis of randomly isolated cDNAs from developing endosperm of rice (*Oryza sativa* L.): evaluation of expressed sequence tags, and expression levels of mRNAs. Plant Mol Biol 29: 685-689.

Lu YH, D'Hont A, Paulet F, Grivet L, Arnaud M and Glaszmann JC (1994) Molecular diversity and genome structure in modern sugarcane varieties. Euphytica 78: 217-226.

McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterback TR, Kahn M, Dubnick M, Kerlavage AR, Venter JC and Fields C (1992) *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. Nature Genet 1: 124-131.

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E and Somerville C (1994) Genes Galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241-1255.

Park YS, Kwak JM, Kwon OY, Kim YS, Lee DS, Cho MJ, Lee HH and Nam HG (1993) Generation of expressed sequence tags of random root cDNA clones of *Brassica napus* by single-run partial sequencing. Plant Physiol 103: 359-370.

Sasaki T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Murayama-Kayano E, Takiguchi T, Takasuga A, Niki T, Ishimaru K, Ikeda H, Yamamoto Y, Mukai Y, Ohta I, Miyadera N, Havukkala I and Minobe Y (1994) Towards cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J 6(4): 615-624.

Shen B, Carneiro N, Torres-Jerez I, Stevenson B, McCreery T, Helentjaris T, Baysdorfer C, Almira E, Ferl RJ, Habben JE and Larkins B (1994) Partial sequencing and mapping of clones from two maize cDNA libraries. Plant Mol Biol 26: 1085-1101.

Tang W and Sun SSM (1993) Sequence of a sugarcane ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit gene. Plant Mol Biol 21: 949-951.

Thompson WF, Everett M, Polans NO, Jorgensen RA and Palmer JD (1993) Phytochrome control of RNA levels in developing pea and mung-bean leaves. Planta 158: 487-500.

Uchimiya H, Kidou S, Shimazaki T, Takamatsu S, Hashimoto H, Nishi R, Aotsuka S, Matsubayashi Y, Kidou N, Umeda M and Kato A (1992) Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (*Oryza sativa* L.). Plant J 2: 1005-1009.

van de Loo FJ, Turner S and Somerville C (1995) Expressed sequence tags from developing castor seeds. Plant Physiol 108: 1141-1150.

Waterston R, Martin C, Craxton M, Coulson A, Hillier L, Durbin R, Green P, Showkeen R, Halloran N, Metzstein M, Hawkins T, Wilson R, Berks M, Du Z, Thomas K, Thierry-Mieg J and Sulston J (1992) A survey of expressed genes in *Caenorhabditis elegans*. Nature Genet 1: 114-123.

Yamamoto K and Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35: 135-144.

CHAPTER 4

GENES EXPRESSED IN SUGARCANE MATURING INTERNODAL TISSUE

[Plant Cell Reports 20(11): 1075-1081 (2002)]

4.1 ABSTRACT

To explore gene expression during sugarcane culm maturation, partial sequence analysis of random clones from maturing culm total and subtracted cDNA libraries has been performed. Database comparisons revealed that of the 337 cDNA sequences analysed, 167 showed sequence homology to gene products in the protein databases while 111 matched uncharacterised plant ESTs only. The remaining cDNAs showed no database match and could represent novel genes. The majority of ESTs corresponded to a variety of genes associated with general cellular metabolism. ESTs homologous to various stress response genes were also well represented. Analysis of ESTs from the subtracted library identified genes that may be preferentially expressed during culm maturation. This research has provided a framework for functional gene analysis in sugarcane sucrose-accumulating tissues.

4.2 INTRODUCTION

Sugarcane is a commercially important crop plant that accounts for approximately 65% of global sugar production. Due to the capacity of sugarcane for storing high concentrations of sucrose in the culm, the processes associated with sucrose accumulation and metabolism during sugarcane growth and maturation have been the subject of intensive study (for review, see Moore 1995). During plant growth, a gradient of maturation and sucrose accumulation develops down the culm such that high sucrose concentrations are reached in the mature internodes. Physiological analyses have shown that sucrose accumulation in sugarcane is a complex, dynamic process. Cycling and turnover of sucrose between the vacuole and metabolic and apoplastic compartments are known to be important factors (Moore 1995) while a complex relationship has been demonstrated to exist in the partitioning of carbon between cellular compartments during sucrose accumulation (Whittaker and Botha 1997). However, the mechanisms regulating these processes are not well understood. Furthermore, despite substantial characterisation of key enzymes involved with sucrose metabolism during culm maturation, their role in regulating sucrose accumulation is still not well defined (Moore 1995; Lingle 1999).

Although the impetus for studying sucrose accumulation in sugarcane is clear, the culm is nonetheless a composite organ associated with a variety of processes and functions. Not only is the culm actively engaged in the metabolism of sucrose but as an integral component of the plants' interface with the external environment, it is also subject to significant stresses. For example, in South Africa sugarcane is particularly susceptible to infestation by the stalk borer, *Eldana saccharina*, which causes severe industry losses.

Molecular analyses of plant growth and development can provide valuable information about metabolic processes at the level of gene transcription. Knowledge of gene expression in sugarcane internodal tissue has been limited thus far to studies of single genes and corresponding proteins directly associated with sucrose metabolism. For example, the expression pattern of sucrose synthase (SuSy) and soluble acid invertase have been reported at the mRNA (Lingle and Dyer 2001; Zhu et al., 2000) and protein level (Buczynski et al., 1993; Zhu et al., 2000). Transcript levels of sucrose-phosphate synthase (SPS) have also been examined (Sugiharto et al., 1997). No information is available however about the types of genes being expressed during culm maturation in sugarcane. Identifying such genes will provide fundamental knowledge about gene expression in sugarcane. Furthermore, this information may facilitate further molecular analyses of metabolic regulation in the maturing sugarcane culm by pinpointing suitable targets that through genetic manipulation could characterise the role of specific genes expressed in sucrose-accumulating tissues.

This paper describes the identification of genes expressed in maturing sugarcane internodal tissue by partial sequence analysis of random cDNAs to generate Expressed Sequence Tags (ESTs). By identifying ESTs in this internodal region, information about those genes most abundantly expressed during culm maturation has been obtained.

4.3 MATERIALS AND METHODS

4.3.1 cDNA library preparation

Two maturing culm cDNA libraries, a total and a subtracted cDNA library, were prepared from mature field-grown sugarcane plants. For this study, maturing culm was defined as internode no. 7, where internode no. 1 is the internode attached to the leaf with the uppermost visible dewlap (van Dillewijn 1952). For both the total and subtracted cDNA libraries, total RNA was extracted from maturing culm tissue as described previously (Carson and Botha 2000). Poly $(A)^+$ RNA was isolated using the Dynabeads mRNA purification kit (Dynal A.S., Oslo, Norway). The total cDNA library was prepared as described previously (Carson and Botha 2000) except that Expand Reverse Transcriptase (Roche Diagnostics GmbH, Mannheim, Germany) was used for first-strand cDNA synthesis. The subtracted cDNA library was constructed by reciprocal subtractive hybridisation between immature (internode no. 2) and maturing culm using the PCR-based Subtractive cDNA Cloning technique (Patel and Sive 1996). Six rounds of subtractive cDNA hybridisation were performed, alternating between three short and three long, and were executed exactly according to the protocol. The protocol was modified for cloning of the subtracted products for library construction. The subtracted cDNA products were first blunt-ended using the Klenow fragment of DNA polymerase I and then ligated to an adapter set that contained an EcoRI restriction site. After restriction digestion with EcoRI, the cDNA products were cloned into the EcoRI site of the Lambda ZAP II phage cloning vector (Stratagene, La Jolla, CA).

4.3.2 Template preparation

Aliquots of the total and subtracted cDNA libraries were plated out onto solid NZY medium and plaques picked at random. Template DNA for sequence analysis was prepared in two ways. Phagemids (pBluescript SK(-)) plus inserts were excised from individual phages (Stratagene, La Jolla, CA) and plated out onto solid Luria Bertani

(LB) medium containing 50µg/ml ampicillin. Phagemid DNA was isolated using a Rapid Plasmid Isolation Protocol (Holmes and Quigly 1981) from a 5ml liquid bacterial culture grown overnight in LB medium containing 50µg/ml ampicillin. Phagemid DNA was further purified through QIAquick spin columns (Qiagen GmbH, Hilden, Germany). Template DNA was prepared also by specific PCR amplification of cDNA inserts directly from individual phages using the universal M13 Forward and Reverse primers. Amplified inserts were purified with QIAquick spin columns prior to sequencing.

4.3.3 DNA sequencing and sequence data analysis

Both phagemid and amplified insert cDNAs were sequenced by dye terminator cycle sequencing using the BigDye™ Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), followed by ethanol precipitation of the extension products. The universal Reverse primer was used to generate single-pass partial sequences for isolated cDNAs. Cycle sequencing was performed in a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) and sequence analysis was performed using an ABI Prism 310 Genetic Analyser (Applied Biosystems, Foster City, CA). Sequences were edited using Sequence Navigator software (Applied Biosystems, Foster City, CA) to remove vector and ambiguous sequences. cDNA sequences were compared with the GenBank non-redundant protein databases and the dbEST database using the BLASTX and BLASTN algorithms, respectively (Altschul et al., 1990). The degree of sequence similarity between the sugarcane cDNA and a known sequence was represented by the *E* value and PAM120 similarity score. Matches were considered significant when *E* values were below 10^{-5} and similarity scores greater than 80 (Newman et al., 1994). The sugarcane EST was putatively identified as either the protein with the lowest *E* value among the candidate proteins generated by the database search or as the dbEST entry with the most significant match. cDNA sequences with matches to the same database accessions were compared to each other using DNASIS® for Windows® Sequence Analysis software (Hitachi Software Engineering Co., Ltd) to identify overlapping clones.

4.4 RESULTS AND DISCUSSION

4.4.1 Generation of sugarcane ESTs

In this study, sequence overlap between random sugarcane cDNA clones and known genes from other organisms was used to identify Expressed Sequence Tags (ESTs) in the maturing sugarcane culm. Random clones were selected from both a total cDNA library and a subtracted cDNA library. The latter had been enriched for transcripts preferentially expressed in the maturing culm and was used to detect genes that may be upregulated during culm maturation. The primary titers of the total and subtracted cDNA library were 2.27 X 10^6 pfu/ml and 1.4 X 10^6 pfu/ml and contained 98% and 94% recombinant clones, respectively. Insert amplification of 250 random clones from the total cDNA library revealed inserts ranging between 350-3000 bp, with an average insert size of 900 bp. For the subtracted cDNA library, insert amplification of 106 clones indicated the inserts were between 230-1200 bp with an average of 400 bp. This insert size was expected according to the particular subtraction scheme used for library construction (Patel and Sive 1996).

Single-pass sequencing of cDNA clones resulted in 225 sequences from the total cDNA library and 112 from the subtracted cDNA library (Table 4.1). To establish ESTs, the cDNAs were first compared with the non-redundant protein databases using the BLASTX algorithm. Search results revealed that 47.6% (107) of total library and 53.6% (60) of subtracted library cDNAs had database matches to known peptide sequences (Table 4.1). Those cDNAs that had no match with sequences in the protein databases were examined at the nucleotide level by comparison with ESTs listed in dbEST database. For both libraries, approximately two-thirds of the cDNAs exhibited sequence similarity to other publicly available ESTs (Table 4.1). The remaining clones with no database match at either the protein or the nucleotide level could represent novel sequences and are useful tags to new sugarcane genes that are expressed in maturing culm tissues.

Table 4.1 Summary of BLASTX and BLASTN (dbEST) search results

	Total cDNA library	Subtracted cDNA library
Number of sequences established	225	112
Significant similarity to known genes (BLASTX $E<10^{-5}$)	107	60
Significant similarity to dbEST entries only (BLASTN $E<10^{-5}$)	78	33
No database match	40	19

4.4.2 Putatively identified genes in the maturing culm

4.4.2.1 Characteristics of the database matches

Sugarcane ESTs with significant sequence similarity to gene products in the non-redundant protein databases are listed in Table 4.2. For ESTs with matches to coding sequences in the dbEST database only, search results are summarised in Table 4.3. Most of the ESTs (83%) described in Table 4.2 matched previously identified plant genes although only 27% were from monocotyledenous species, the most common being maize and rice. Only one EST was homologous to a sugarcane gene (nucleoside diphosphate kinase I). *Arabidopsis* was the most frequently detected dicotyledenous source of gene homologues. These results probably reflect the relative contribution of gene sequences from different plant species to the public databases. The remaining 17% of the ESTs were homologous to genes from a range of non-plant organisms including yeast, bacteria and humans, suggesting that these genes contain highly conserved sequences.

Of the 167 ESTs established from both cDNA libraries, 68 showed overlapping sequences homologous to the same gene (Table 4.2). Of these, approximately half (35) were obtained from the subtracted cDNA library. The most frequently detected genes were an environmental stress-induced protein (11 clones) and an abscisic-acid and environmental stress-inducible gene (5 clones). Other ESTs present in multiple copies included genes such as arabinosidase (4 clones), glycerol-3-phosphate permease homolog (3 clones), callose synthase catalytic subunit-like protein (3 clones), 60S ribosomal protein L7A (3 clones) and several others represented by 2 clones (Table 4.2). The frequency of cDNA clones corresponding to the same gene can provide an estimate of steady-state transcript levels in a particular tissue (Cooke et al., 1996). This suggests that the ESTs present in multiple copies in this study could represent an increase in the expression of the corresponding genes in the maturing sugarcane culm.

Table 4.2 Sugarcane ESTs with significant sequence similarity ($E<10^{-5}$) to known genes in the non-redundant peptide databases

Putative identity	Organism	Accession number	Number of database matches	Library[a]
Auxin biosynthesis				
Putative nitrilase-associated protein	*Arabidopsis thaliana*	AAD20083	1	T
Carbohydrate metabolism				
Pyruvate dehydrogenase E1 alpha subunit	*Arabidopsis thaliana*	U80185	2	T
Alanine aminotransferase 2	*Panicum miliaceum*	P34106	1	S
Glycerol-3-phosphate permease homolog	*Arabidopsis thaliana*	Z97343	3	S
Trehalose-6-phosphate phosphatase	*Arabidopsis thaliana*	AAC39369	2	S
Cell wall metabolism				
UDP-glucose dehydrogenase	*Glycine max*	U53418	1	S
Callose synthase catalytic subunit-like protein	*Arabidopsis thaliana*	CAB88264	3	S
Highly similar to putative callose synthase catalytic subunit	*Arabidopsis thaliana*	AAD30609	1	S
Putative alpha-L-arabinofuranosidase	*Arabidopsis thaliana*	AAF19575	2	S
Arabinosidase	*Bacteroides ovatus*	U15178	4	S
DNA- and RNA- related				
HMGd1	*Zea mays*	X08807	2	T
Replication control protein homolog	*Arabidopsis thaliana*	Z97336	1	T
CND41, chloroplast nucleoid DNA binding protein	*Nicotiana tabacum*	D26015	1	T
S-adenosyl-methionine synthetase 1	*Oryza sativa*	P46611	1	T
Adenosylhomocysteinase	*Catharanthus roseus*	P35007	1	T
Putative small nuclear ribonucleoprotein E homolog C29	*Medicago sativa*	P24715	1	T
Putative small nuclear ribonucleoprotein, Sm D2	*Arabidopsis thaliana*	AAC62848	2	T
Exonuclease 46	*Phage T4*	NCBPX6	1	T
RNA polymerase I, II and III 16.5 kD subunit	*Arabidopsis thaliana*	AAC28252	1	T
Fatty acid metabolism				
Acyl carrier protein II precursor	*Hordeum vulgare*	P08817	1	T
Similar to ATP-citrate-lyase	*Arabidopsis thaliana*	AAB71965	1	T
Putative esterase D	*Arabidopsis thaliana*	AAB84335	1	T
Acyl-CoA-binding protein	*Ricinus communis*	Y08996	2	T

Serine palmitoyltransferase subunit II	*Homo sapiens*	Y08686	1	T
Membrane and transport				
Vacuolar proton pump SFD alpha isoform	*Bos taurus*	AAC02987	1	T
ATP synthase complex subunit 9	*Sorghum sp.*	U61165	1	T
Putative ATPase (ISW2-like)	*Arabidopsis thaliana*	AAF08585	1	T
Plasma membrane MIP protein	*Zea mays*	AAD29676	1	T
Putative protein transport protein SEC61	*Arabidopsis thaliana*	AAC27401	1	T
Metal-binding proteins				
Copper homeostasis factor	*Arabidopsis thaliana*	U88711	1	T
Oligosaccharide synthesis				
Nucleoside diphosphate kinase I	*Saccharum officinarum*	P93554	1	T
Protein modification				
Calcium dependent protein kinase 19	*Arabidopsis thaliana*	S71777	1	T
Kinase associated protein phosphatase	*Zea mays*	U81960	1	T
CTR1 protein kinase isolog	*Arabidopsis thaliana*	AAB64021	1	T
Amidase lytA	*Lactococcus lactis phage US3*	JC1270	1	T
Contains similarity to *Rattus* O-GlcNAc transferase	*Arabidopsis thaliana*	U76557	1	T
Atranbp 1b	*Arabidopsis thaliana*	X97378	1	T
Protein phosphatase 2A	*Oryza sativa*	U49113	1	T
Ubiquitin S6(2)	*Drosophila melanogaster*	M33019	1	T
polyubiquitin	*Pinus sylvestris*	X98063	1	T
Ubiquitin-conjugating enzyme E2	*Homo sapiens*	Q16781	1	T
Ubiquitin-conjugating enzyme	*Lycopersicon esculentum*	S57619	1	T
Putative ubiquitin-conjugating enzyme E2	*Arabidopsis thaliana*	AAD24607	1	S
Ubiquitin-activating enzyme E1-like protein	*Homo sapiens*	AAC69630	2	S
Ubiquitin-conjugating enzyme	*Zea mays*	AAB88617	2	S
Protein synthesis				
60S ribosomal protein L25	*Nicotiana tabacum*	Q07761	2	T
60S ribosomal protein L7A	*Oryza sativa*	P35685	3	T
60S ribosomal protein L37A	*Brassica rapa*	P43209	1	T
60S ribosomal protein L13	*Arabidopsis thaliana*	P41127	1	T
60S ribosomal protein L35	*Rattus norvegicus*	P17078	1	T
60S ribosomal protein L3	*Oryza sativa*	P35684	1	T
60S ribosomal protein L5	*Oryza sativa*	P49625	1	T
40S ribosomal protein S13	*Zea mays*	Q05761	2	T
40S ribosomal protein S3A	*Oryza sativa*	P49397	1	T
40S ribosomal protein S9	*Podospora anserina*	P52810	1	T
Ribosomal protein S3a, cytosolic	*Oryza sativa*	T02874	1	T
Ribosomal protein L25	*Zea mays*	AAC24573	1	T
IF5A-MEDSA initiation factor 5A (eIF-5A)	*Medicago sativa*	P26564	1	T
Initiation factor 5A (EIF-5A)	*Solanum tuberosum*	P56333	2	T
Translation initiation factor 5A	*Zea mays*	Y07920	1	T
Translation initiation factor EIF-5A	*Nicotiana plumbaginifolia*	S21058	1	T
Translation initiation factor, eIF-5A	*Oryza sativa*	CAB96075	1	T

Translation elongation factor eEF-1 α-chain	*Zea mays*	S66338	1	T
Elongation factor 1-α	*Cicer arietinum*	CAA09041	1	T
Protein translation factor SUI1 homolog	*Oryza sativa*	P33278	1	T
Nascent polypeptide associated complex alpha chain	*Nicotiana tabacum*	U74622	1	T
40S ribosomal protein S15A	*Arabidopsis thaliana*	P42798	1	S
Putative ribosomal protein S19 or S24	*Arabidopsis thaliana*	AAD56997	1	S
Regulation/signal transduction				
MIHC [TNFR2-TRAF signalling complex protein, c-IAP2 protein]	*Homo sapiens*	U37546	1	T
GF14-c protein	*Oryza sativa*	U65957	1	T
GTP binding protein Rop1At	*Arabidopsis thaliana*	U49971	1	T
thioredoxin	*Lilium longiflorum*	L18909	1	T
Calmodulin-1	*Nicotiana plumbaginifolia*	e1313841	1	T
Translationally controlled tumor protein homolog (TCTP)	*Oryza sativa*	P35681	1	T
Ripening-associated protein	*Musa acuminata*	AAB82776	1	S
GDP dissociation inhibitor protein OsGDI1	*Oryza sativa*	AAB69870	1	S
Translocon-associated protein, alpha subunit precursor	*Arabidopsis thaliana*	P45434	2	S
Putative signal sequence receptor, alpha subunit	*Arabidopsis thaliana*	AAD29800	1	S
Legumain precursor	*Canavalia ensiformis*	P49046	1	S
Probable protein involved in autophagy yeast apg7 homolog	*Schizosaccharomyces pombe*	T40646	1	S
Cysteine protease	*Zea mays*	X99936	1	S
Tetracycline transporter protein	*Arabidopsis thaliana*	AAC64231	1	S
ran	*Oryza sativa*	BAA81911	1	S
Stress-response				
Blt101 protein	*Hordeum vulgare*	S40406	1	T
Environmental stress-induced protein	*Medicago sativa*	M74191	11	T
Abscisic acid and environmental stress-inducible gene	*Medicago falcata*	Q09134	5	T
Heat shock protein	*Arabidopsis thaliana*	e1250065	1	T
Heat shock protein 82	*Oryza sativa*	P33126	1	T
DNAJ protein	*Solanum tuberosum*	X94301	1	T
PVPR3	*Phaseolus vulgaris*	M75856	1	T
remorin	*Solanum tuberosum*	U72489	1	T
jacalin	*Artocarpus integrifolia*	L03797	1	S
Similar to jacalin	*Arabidopsis thaliana*	AAD55651	1	S
Hypothetical protein wali7	*Triticum aestivum*	T06984	2	S
Osr40g3	*Oryza sativa*	Y08988	1	S
Structural proteins				
Beta-6-tubulin	*Zea mays*	S43327	1	T
Tubulin beta-1 chain	*Oryza sativa*	S52007	1	T
Actin depolymerizing factor	*Zea mays*	X97726	1	T
Actin 1	*Zea mays*	P02582	1	S
Ankyrin-like protein	*Arabidopsis thaliana*	AAD32290	1	S
Urea metabolism				
Arginino-succinate lyase	*Arabidopsis thaliana*	Z97558	1	T

Unclassified				
Unknown protein	*Arabidopsis thaliana*	AAB95291	1	T
Unknown protein	*Arabidopsis thaliana*	AAF63822	1	T
Unknown protein	*Arabidopsis thaliana*	AAB80671	1	T
Unknown	*Arabidopsis thaliana*	U19134	1	T
Hypothetical protein	*Schizosaccharomyces pombe*	Z98598	1	T
Yeast hypothetical protein 16.2 kD protein in PIR3-APE2 intergenic region	*Schizosaccharomyces cerevisiae*	P36053	1	T
KIAA0695 protein	*Homo sapiens*	d1032631	1	T
ORF N118	*Schizosaccharomyces pombe*	D84656	1	T
PfSNF2L	*Plasmodium falciparum*	AAC47719	1	T
R11H6.2	*Caenorhabditis elegans*	Z93386	1	T
WD-40 repeat protein MS14	*Arabidopsis thaliana*	AAD03340	1	T
Unknown protein	*Arabidopsis thaliana*	AAD18124	1	S
Hypothetical protein	*Arabidopsis thaliana*	U89959	1	S
Hypothetical protein	*Schizosaccharomyces pombe*	Z99165	2	S
Hypothetical protein	*Arabidopsis thaliana*	AAF09076	1	S
Hypothetical protein	*Schizosaccharomyces pombe*	e1251101	2	S
Hypothetical protein	*Arabidopsis thaliana*	Z97336	2	S
Hypothetical protein T5L19.190	*Arabidopsis thaliana*	T04010	2	S
Hypothetical protein T21L8.170	*Arabidopsis thaliana*	T12997	1	S
Contains strong similarity to a hypothetical protein and contains three Kelch domains	*Arabidopsis thaliana*	AAF43933	1	S
No definition line found	*Caenorhabditis elegans*	AAC25802	2	S
Putative protein	*Arabidopsis thaliana*	e1248659	3	S
R08F11.1 gene product	*Caenorhabditis elegans*	AAB54243	1	S

[a] T= total cDNA library; S= subtracted cDNA library

4.4.2.2 Sugarcane ESTs homologous to known gene products

ESTs were grouped into general categories according to the proposed function of the primary homolog (Table 4.2). The majority of sugarcane maturing culm ESTs were homologous to genes associated with a variety of general cellular metabolic processes. For the ESTs obtained from the total cDNA library, genes associated with protein synthesis and protein modification were well represented. They included several different ribosomal proteins as well as a variety of translation initiation and elongation factors. More than half of the ESTs related to protein modification were homologous to genes involved in the ubiquitin pathway, several of these being obtained from the subtracted cDNA library. Other ESTs involved in protein modification included several protein kinases and a protein phosphatase. The function of many of the proteins encoded by the genes identified in this study is well known with regard to their role in general cellular metabolism. It has been suggested also that ribosomal proteins, for example, play a specific regulatory role during development in both animals and plants (Reski et al., 1998). Furthermore, the degradation of specific proteins via the ubiquitin pathway has been shown to be associated with metabolic regulation, particularly during senescence and stress-responses (Belknap and Garbarino 1996). Several genes encoding a variety of regulatory proteins were detected in the maturing culm total cDNA library, including well-characterised proteins such as calmodulin and GF14-c. Other ESTs related to general cellular processes such as fatty acid metabolism, structural proteins, membrane-associated proteins as well as ESTs for genes involved with different aspects of DNA and RNA synthesis were also identified in the total cDNA library.

Analysis of the ESTs obtained from the subtracted cDNA library revealed that while some encoded genes associated with general cellular metabolism (protein synthesis, structural proteins, protein modification), the majority of ESTs with a putative identity could be grouped into four functional categories, namely carbohydrate metabolism, cell wall metabolism, regulation/signal transduction and stress-responses (Table 4.2). As was observed for the total cDNA library, many ESTs were obtained that could not be assigned a putative identity (unclassified). These results provide an indication of the types of genes preferentially associated with maturing sugarcane culm tissues. Furthermore, the frequency with which multiple ESTs were detected in the subtracted library corresponding, in particular, to genes associated with carbohydrate metabolism and cell wall metabolism suggests that these genes may have important functions during culm maturation. Of the ESTs homologous to genes associated with carbohydrate metabolism identified in this study, none encoded enzymes and proteins commonly acknowledged as components of the sucrose metabolism pathway (for review, see Moore 1995). This result is surprising considering that internode 7, selected for library construction, has a physiological characteristic of a high sucrose accumulation rate. It has been reported that the composition of clones in cDNA libraries used for EST analyses reflect the expression of genes related to the particular growth state of the tissue examined (Yamamoto and Sasaki 1997). In developing seeds of rice (Liu et al., 1995) and *Arabidopsis* (White et al., 2000), EST analysis revealed that genes coding for seed storage proteins were the most abundant. Similarly, the most abundant transcripts detected in leaves of *Brassica napus* (Lee et al., 1998) and grape

(Ablett et al., 2000) were associated with photosynthesis. Although only limited numbers of ESTs have been analysed in the current study, results suggest that in maturing sugarcane internodal tissue, genes directly associated with sucrose accumulation may not be abundantly expressed. Further research is required to establish the nature of regulation at the level of transcription for these genes during sucrose accumulation in sugarcane.

The identification of ESTs homologous to a diverse selection of stress-response genes in both the total and subtracted cDNA libraries is also noteworthy. The most frequently identified were an environmental stress-induced protein and an abscisic acid and environmental stress-inducible gene. These genes have been shown to be inducible by a multitude of environmental stresses such as low temperature, drought, salt and wounding (Luo et al., 1992). Similarly, other homologs such as PVPR3 are reported to be involved in plant defense (Sharma et al., 1992) while the expression of Os40g3 is induced by salt-stress (Moons et al., 1997). Comparable results have been reported in sugi inner bark (Ujino-Ihara et al., 2000) and grape berries (Ablett et al., 2000) where a significant proportion of ESTs were similar to transcripts of stress-response genes. For sugarcane the results indicate that, during growth and maturation, the culm is actively engaged in responding to a variety of stresses. Additional analysis of the ESTs identified in this study could provide valuable information about the genetic regulation of stress responses in sugarcane.

4.4.2.3 Sugarcane EST matches to uncharacterised coding sequences

Analysis of sugarcane cDNAs that exhibited significant sequence similarity to publicly available ESTs only revealed that all sugarcane ESTs matched uncharacterised ESTs from other monocotyledenous species (Table 4.3A). For both total and subtracted cDNA libraries, the most matches were to ESTs from sorghum. This was not unexpected as sorghum is a close relative of sugarcane. Similarly, ESTs from maize were also frequently detected. This suggests that these ESTs may not only be plant-specific but may also be specific to monocotyledenous plant species. Although the dbEST database matches were not informative with regard to putative gene function it was noted that all sugarcane cDNAs matched ESTs derived from cDNA libraries prepared from various reproductive, vegetative, meristematic and stress-induced tissues (Table 4.3B). This suggests that genes associated with highly active metabolic processes are expressed in the maturing culm.

Table 4.3 Sugarcane cDNAs with significant sequence similarity to ESTs grouped according to most significant plant match (A) and corresponding cDNA library source tissue type (B)

		Total cDNA library	Subtracted cDNA library
A. Plant match			
	Sorghum	52 (67)[a]	22 (67)
	Maize	24 (31)	7 (21)
	Rice	1 (1.2)	-
	Barley	1 (1.2)	1 (3)
	Wheat	-	2 (6)
	Wild wheat sp.	-	1 (3)
B. cDNA library source tissue			
	Reproductive	32 (41)	16 (48)
	Vegetative	28 (36)	13 (39)
	Stress-induced	10 (13)	3 (9)
	Meristematic	8 (10)	1 (0.03)

[a] Values in parentheses are percentage of total

In this study, the identification of ESTs has provided an insight into the complexity of gene expression in the sugarcane culm. Based on transcript composition it is evident that culm maturation is characterised by the expression of a wide variety of genes. This is the first report of a random EST analysis for sugarcane maturing culm and has provided a valuable framework for future functional analysis of gene expression in sucrose-accumulating tissues.

4.5 REFERENCES

Ablett E, Seaton G, Scott K, Shelton D, Graham MW, Baverstock P, Lee LS and Henry R (2000) Analysis of grape ESTs: global gene expression patterns in leaf and berry. Plant Sci 159: 87-95.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Belknap WR and Garbarino JE (1996) The role of ubiquitin in plant senescence and stress responses. Trends in Plant Sci 1(10): 331-335.

Buczynski SR, Thom M, Chourey P and Maretzki A (1993) Tissue distribution and characterization of sucrose synthase isozymes in sugarcane. J Plant Physiol 142: 641-646.

Carson DL and Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. Crop Sci 40(6): 1769-1779.

Cooke R, Raynal M, Laudié M, Grellet F, Delseny M, Morris P-C, Guerrier D, Giraudat J, Quigley F, Clabault G, Li Y-F, Mache R, Krivitzky M, Gy IJ-J, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clément B, Philipps G, Hervé C, Bardet C, Tremousaygue D, Lescure B, Lacomme C, Roby D, Jourjon M-F, Chabrier P, Charpenteau J-L, Desprez T, Amselem J, Chiapello H and Höfte H (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. Plant J 9(1): 101-124.

Holmes DS and Quigly M (1981) A rapid boiling method for the preparation of bacterial plasmids. Anal Biochem 114: 193-197.

Lee CM, Lee YJ, Lee MH, Nam HG, Cho TJ, Hahn TR, Cho MJ and Sohn U (1998) Large-scale analysis of expressed genes from the leaf of oilseed rape (*Brassica napus* L.). Plant Cell Rep 17: 930-936.

Lingle SE (1999) Sugar metabolism during growth and development in sugarcane internodes. Crop Sci 39: 480-486.

Lingle SE and Dyer JM (2001) Cloning and expression of sucrose synthase-1 cDNA from sugarcane. J Plant Physiol 158: 129-131.

Liu J, Hara C, Umeda M, Zhao Y, Okita TW and Uchimiya H (1995) Analysis of randomly isolated cDNAs from developing endosperm of rice (*Oryza sativa* L.): evaluation of expressed sequence tags, and expression levels of mRNAs. Plant Mol Biol 29: 685-689.

Luo M, Liu J-H, Mohapatra S, Hill RD and Mohapatra SS (1992) Characterisation of a gene family encoding abscisic acid- and environmental stress-inducible proteins of alfalfa. J Biol Chem 267(22): 15367-15374.

Moons A, Gielen J, Vandekerckhove J, Van Der Straeten D, Gheysen G and Van Montagu M (1997) An abscisic-acid- and salt-stress-responsive rice cDNA from a novel plant gene family. Planta 202: 443-454.

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22: 661-679.

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E and Somerville C (1994) Genes Galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241-1255.

Ok S, Chung YS, Um BY, Park MS, Bae J-M, Lee SJ and Shin JS (2000) Identification of expressed sequence tags of watermelon (*Citrullus lanatus*) leaf at the vegetative stage. Plant Cell Rep 19: 932-937.

Patel M and Sive H (1996) PCR-based subtractive cDNA cloning. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K (eds) Current Protocols in Molecular Biology, Vol. 1. Greene and Wiley-InterScience, New York, pp 5.9.1-5.9.20.

Reski R, Reynolds S, Wehe M, Kleber-Janke T and Kruse S (1998) Moss (*Physcomitrella patens*) expressed sequence tags include several sequences which are novel for plants. Bot Acta 111: 143-149.

Sharma YK, Hinojos CM and Mehdy MC (1992) cDNA cloning, structure, and expression of a novel pathogenesis-related protein in bean. Mol Plant-Microbe Int 5(1): 89-95.

Sugiharto B, Sakakibara H, Saumadi and Sugiyama T (1997) Differential expression of two genes for sucrose-phosphate synthase in sugarcane: molecular cloning of the cDNAs and comparative analysis of gene expression. Plant Cell Physiol 38(8): 961-965.

Ujino-Ihara T, Yoshimura K, Ugawa Y, Yoshimaru H, Nagasaka K and Tsumura Y (2000) Expression analysis of ESTs derived from the inner bark of *Cryptomeria japonica*. Plant Mol Biol 43: 451-457.

van Dillewijn C (1952) Growth: general, grand period, growth formulae. In: van Dillewijn C (ed) Botany of Sugarcane, Vol. 1. Veenen and Zonen, The Netherlands, pp 97-162.

White JA, Todd J, Newman T, Focks N, Girke T, Martínez de Ilárduya O, Jaworski JG, Ohlrogge JB and Benning C (2000) A new set of Arabidopsis expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol 124: 1582-1594.

Whittaker A and Botha FC (1997) Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. Plant Physiol 115: 1651-1659.

Yamamoto K and Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35: 135-144.

Zhu YJ, Albert HH and Moore PH (2000) Differential expression of soluble acid invertase genes in the shoots of high-sucrose and low-sucrose species of *Saccharum* and their hybrids. Aust J Plant Physiol 27: 193-199.

CHAPTER 5

DIFFERENTIAL GENE EXPRESSION IN SUGARCANE LEAF AND INTERNODAL TISSUES OF VARYING MATURITY

[Manuscript: submitted to South African Journal of Botany]

5.1 ABSTRACT

The expression patterns of sugarcane (*Saccharum* spp. hybrids) genes were examined in different tissue sources and developmental stages to identify differentially expressed genes. cDNA arrays containing 1000 random clones from an immature leaf and maturing culm cDNA library were hybridised with radioactively-labeled poly $(A)^+$ RNA from immature leaf, mature leaf, immature culm and maturing culm. All cDNAs were found to hybridise to all four probes, but differences in signal intensity were observed for individual cDNAs between hybridisation events. No cDNAs displaying tissue- or developmental-stage specific expression were detected. Comparisons between hybridisation patterns identified 61 cDNAs that were more abundantly expressed in immature and mature leaf than the culm. Likewise, 25 cDNAs preferentially expressed in immature and maturing culm were detected. ESTs established for the differentially expressed cDNAs revealed sequence homology to a diverse collection of genes in both the leaf and the culm. These included genes associated with general cellular metabolism, transport, regulation and a variety of stress responses. None of the differentially expressed genes identified in the culm were homologous to genes known to be associated with sucrose accumulation. These results suggest that the genetic regulation of processes related to sugarcane leaf and culm maturation is very complex.

5.2 INTRODUCTION

Sugarcane is a member of the Gramineae family that includes many major monocotyledenous agricultural crop species such as maize, rice, wheat, sorghum and barley. Many of these species have large, complex genomes that present a substantial challenge to molecular studies. Much of what is known at the genetic level for crop plants has been obtained through extensive research on species with relatively small genomes such as sorghum (Draye et al., 2001) and rice (Yuan et al., 2001). The high degree of genome synteny that is known to exist between plants (Schmidt 2000) has facilitated research by enabling the transfer of information and resources from well-studied genomes to related species. In this way, valuable insight into the organisation of the complex polyploid genome of sugarcane has been obtained by comparative genome analyses using data from maize and sorghum (Grivet et al., 1996, Ming et al., 1998). However, while genome structures are being studied intensively, information at the gene level is still very limited for most crop plants.

Gene discovery through the identification of Expressed Sequence Tags (ESTs) has provided access to expressed gene data for many organisms and is a useful tool for large-genome crop plants as ESTs can be generated independent of genome complexity. Gene sequence information in the form of ESTs is now available in the public databases for many crop plants, including rice, maize and sugarcane. Not much is known, however, about the expression patterns of the genes represented by ESTs and their possible roles in growth and development. Most of the research efforts in this area have focussed on rice, the model plant for monocotyledenous species. ESTs from nine different rice cDNA libraries representative of the principal tissues in the plant life cycle have been compared and results revealed similarities and differences in the expression of genes associated with differentiation, specific growth conditions and environmental stress (Yamamoto and Sasaki 1997). However, although many global expression studies are currently being conducted for rice using EST clones, very few reports have been published to date (Delseny et al., 2001).

Research conducted on *Arabidopsis* has suggested that while basic intracellular processes are conserved between organisms, intercellular processes, such as development, may use different proteins (Willmann 2001). Therefore, unlike data on chromosome organisation and genome structure, gene expression data may not be as readily transferable between plant species. Sugarcane has a very distinct physiology where the culm has evolved as a specialised organ capable of storing high concentrations of sucrose. Consequently, to understand the molecular mechanisms controlling culm maturation and sucrose accumulation, it is important to first identify genes that are differentially expressed in different tissue sources and developmental stages.

We have initiated an investigation to establish the expression of sugarcane genes in a variety of leaf and culm growth stages. It has been demonstrated recently that simultaneous expression analysis of multiple genes systematically arrayed onto solid supports using a "reverse Northern" technique is an effective procedure to identify differentially expressed genes in a variety of plant tissues (Desprez et al., 1998, Ruan et al., 1998, Girke et al., 2000). In the present study, we use a reverse Northern analysis to examine the expression of random sugarcane cDNA clones in immature and maturing leaf and culm tissues and report on the identification of genes that are differentially expressed in sugarcane leaf and culm.

5.3 MATERIALS AND METHODS

5.3.1 Source of cDNA clones

Sugarcane cDNA clones were randomly selected from two total cDNA libraries. A leaf roll (immature leaf) cDNA library was prepared as described previously (Carson and Botha 2000). A maturing culm cDNA library was prepared using similar techniques but with some minor modifications. For this library, maturing culm was defined as internode no. 7, where internode no. 1 is the internode attached to the leaf with the uppermost visible dewlap (van Dillewijn et al., 1952). The modifications to the method included the use of the Dynabeads mRNA Purification Kit (Dynal A.S, Oslo, Norway) for isolation of mRNA and Expand Reverse Transcriptase (Roche Diagnostics GmbH, Mannheim, Germany) for first-strand cDNA synthesis. In both cases, the reactions were performed according to the manufacturer's protocol.

5.3.2 Bacterial clones, production of cDNA arrays and Northern blots

Aliquots of the leaf roll and maturing culm library were plated out onto solid NZY medium and plaques picked at random. Phagemids (pBluescript SK(-)) plus inserts were excised from individual phages according to the manufacturer's instructions

(Stratagene, La Jolla, CA, USA) and plated out onto solid Luria Bertani (LB) medium containing 50µg/ml ampicillin. Bacterial glycerol stocks were prepared from phagemid clones by mixing liquid bacterial cultures grown overnight in LB medium containing 50µg/ml ampicillin with sterile glycerol in a 5.7:1 ratio and flash-freezing in liquid nitrogen. Bacterial glycerol stocks were also prepared from *E. coli* SOLR and HB101 cell lines. Phagemid DNA was isolated from a 5 ml overnight liquid bacterial culture using a Rapid Plasmid Isolation Protocol (Holmes and Quigly 1981) and purified through QIAquick spin columns (Qiagen GmbH, Hilden, Germany).

For cDNA macroarray preparation, bacterial clones were spotted onto Hybond™- N+ nylon membranes (Amersham International, Buckinghamshire, United Kingdom) in a 4X4 duplication format using a QBOT (Genetix, Hampshire, United Kingdom). Bacterial cell lines *E. coli* SOLR and HB101 were included as negative controls. Samples were lysed (0.5M NaOH, 1.5M NaCl) for 5 minutes, neutralised (1.5M NaCl, 0.5M Tris-HCl, pH 7.2) and allowed to air-dry. The DNA was then denatured with 0.4M NaOH for 5 minutes and neutralised with 5X SSPE.

For the preparation of Northern blots, total RNA (10µg) was electrophoresed in 1.2% agarose formaldehyde gels and transferred to a Hybond™- N+ membrane (Amersham International, Buckinghamshire, United Kingdom) according to the manufacturer's instructions.

5.3.3 RNA extraction and probe synthesis

Total RNA was extracted as described previously (Carson and Botha 2000). Poly $(A)^+$ RNA was isolated from 75µg total RNA using Dynabeads (Dynal A.S, Oslo, Norway) according to the manufacturer's instructions. For the synthesis of total cDNA probes, a modification of the method described by Sambrook *et al.,* was used (Sambrook et al., 1989). A mixture was prepared containing 12.5µg random hexamer primers (Amersham Pharmacia Biotech Inc, Piscataway, NJ), 20mM each dCTP, dGTP, dTTP and 120µM dATP. Also included in the mixture was 200µM ddCTP, shown previously to significantly improve the efficiency and reproducibility of the reverse transcription reaction (Decraene et al., 1999). This mixture was dried down to complete dryness in a SpeedVac (Savant

Instruments Inc., Holbrook, NY) and resuspended in 5X Expand Reverse Transcriptase buffer (Roche Diagnostics GmbH, Mannheim, Germany), 10mM DTT (final concentration) and DEPC-treated water to a volume of 7.5μl. A 1μg poly $(A)^+$ RNA sample was denatured for 5 minutes at 70°C, cooled on ice and added to the mixture with 20U RNase inhibitor (Roche Diagnostics GmbH, Mannheim, Germany), 100U Expand Reverse Transcriptase (Roche Diagnostics GmbH, Mannheim, Germany) and 50μCi [α-^{33}P]dATP (2500 Ci/mmol) to a final volume of 20μl. The mixture was incubated at 30°C for 10 minutes, followed by 42°C for 45 minutes. The reaction was stopped by the addition of 1.0μl 0.5M EDTA (pH 8.0) and 1.0μl 10% (w/v) SDS. The RNA was hydrolysed by the addition of 3μl of 3N NaOH and incubation for 30 minutes at 68°C. The mixture was allowed to cool to room temperature and then mixed with 10μl 1M Tris-HCl (pH 7.4) and 3μl 2N HCl. The probe was purified by phenol:chloroform extraction and ethanol precipitated to remove unincorporated nucleotides.

Specific cDNA probes for use in Northern blot analysis were labeled with [α-^{32}P]dCTP (3000Ci/mmol) by random primer labeling using the Megaprime™ DNA Labelling system (Amersham International, Buckinghamshire, United Kingdom) according to the manufacturer's protocol. Template DNA was prepared by specific PCR amplification of cDNA inserts from phagemid DNA, using the M13 Forward and Reverse primers. Amplified inserts were purified using QIAquick spin columns (Qiagen GmbH, Hilden, Germany) prior to labelling. Probes were purified to remove unincorporated nucleotides using NucTrap® Probe Purification Columns (Stratagene, La Jolla, CA).

5.3.4 Hybridisation procedures

cDNA macroarray filters were prehybridised overnight at 65°C in a solution of 0.5M sodium phosphate buffer (pH 7.2), 7% (w/v) SDS, 0.9mM EDTA and 10μg/ml denatured salmon sperm DNA (final concentrations). Hybridisation was performed overnight at 65°C with fresh solution minus the denatured salmon sperm DNA but including the appropriate probe. Filters were washed twice with 1X SSC, 0.1% (w/v) SDS for 20 minutes at 65°C, followed by twice with 0.5X SSC, 0.1% (w/v) SDS for 20 minutes at 65°C. Hybridised filters were exposed to a Super Resolution Cyclone

Phosphor Screen (Packard Instrument Company, Meriden, CT) for 4-16 hours and data captured and analysed with OptiQuant™ software (Packard Instrument Company, Meriden, CT). The signal intensity observed for each clone was recorded manually as "high", "medium" or "low". Densitometric quantification of signal intensity using the OptiQuant™ software performed on 48 randomly selected samples confirmed a significant difference in hybridisation signal intensity between the designated three categories.

Northern blot hybridisation was performed exactly according to the protocol supplied with the Hybond™- N+ membrane (Amersham International, Buckinghamshire, United Kingdom). Hybridised membranes were exposed to phosphorscreens as described above. Membranes were also exposed to X-ray film for various times for autoradiography.

5.3.5 DNA sequencing and sequence data analysis

DNA sequencing was performed by dye terminator cycle sequencing using the BigDye™ Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), followed by ethanol precipitation of the extension products. Both procedures were performed according to the manufacturer's instructions. The M13 Forward and Reverse primers were used to generate partial sequences for all isolated cDNAs. Cycle sequencing was performed in a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) and sequence analysis was performed using an ABI Prism 310 Genetic Analyser (Applied Biosystems, Foster City, CA).

Sequences were edited using Sequence Navigator software (Applied Biosystems, Foster City, CA) to remove vector and ambiguous sequences. cDNA sequences were compared with the GenBank non-redundant protein and EST databases using the BLASTX and BLASTN algorithms, respectively (Altschul et al., 1990). The degree of sequence similarity between the sugarcane cDNA clone and a known sequence was represented by the *E* value and PAM120 similarity score. Matches were considered significant when *E* values were below 10^{-5} and similarity scores greater than 80

(Newman et al., 1994). The sugarcane EST was identified as the protein with the lowest *E* value among the candidate proteins generated by the database search.

5.4 RESULTS

5.4.1 Patterns of transcript abundance

Membrane filters containing 1000 random sugarcane cDNA clones, 500 each from immature leaf and maturing culm cDNA libraries, were hybridised separately with probes synthesised from mRNA samples isolated from immature leaf, mature (fully expanded) leaf, immature culm (internode 2) and maturing culm (internode 7). Analysis of hybridisation data revealed that all 1000 cDNAs hybridised to the four mRNA populations indicating that the genes represented by the cDNAs were expressed in all four sugarcane tissue types tested. However, variations in the intensity of hybridisation signal were observed for the cDNA clones both within and between mRNA populations. To provide an assessment of inter-probe variability, the signal intensity of three individual cDNA clones displaying high level constitutive expression in the four tissues was quantified using OptiQuant™ software. Comparison between results indicated that there was approximately 10% variation in signal intensity between separate hybridisation events (data not shown). Manually assigning each cDNA a signal intensity of either "high", "medium" or "low" (as described in Materials and Methods) could therefore provide a rough estimate of the abundance of individual transcripts in each specific tissue type explored (Fig. 5.1). After hybridisation with the immature leaf total cDNA probe, 12% of cDNA clones exhibited a high signal intensity, while 52% and 36% of cDNAs displayed a medium and low signal intensity, respectively (Fig. 5.1). Similarly, hybridisation with a mature leaf probe revealed 13% highly abundant cDNAs, 48% moderate and 39% low abundance cDNAs, respectively. When probed with an immature culm mRNA sample the majority of cDNAs (75%) hybridised with a low signal intensity (Fig. 5.1). For the remaining cDNAs, 11% exhibited high signal intensities and 14% had moderate signals. Similar results were observed for the maturing culm where 76% of cDNA clones hybridised at a low signal intensity with 11% and 13% of clones displaying high and medium signal intensities, respectively. As Fig. 5.1 indicates, the patterns of transcript abundance were alike for the cDNAs when probed with mRNA from the two leaf tissues. This was also the case after hybridisation with the two culm tissues, however, differences in the patterns of transcript abundance between leaf and culm tissues were observed. These results suggested that there were variations in the expression patterns of individual genes between leaf and culm tissues.

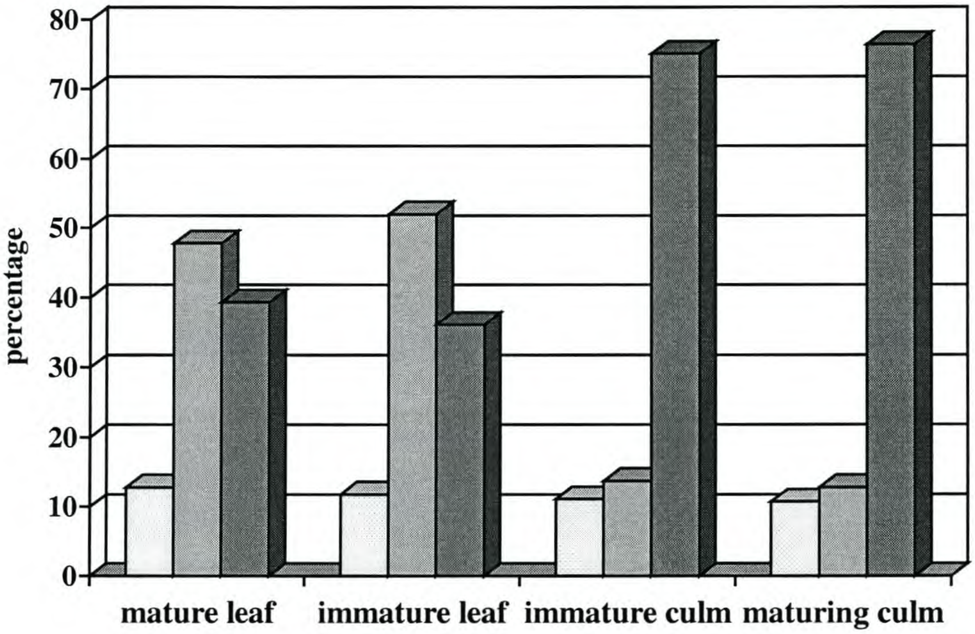


Fig. 5.1 Percentage of cDNA clones exhibiting high (□), medium (▨) and low (■) signal intensities after hybridisation with mRNA from mature leaf, immature leaf, immature culm and maturing culm

5.4.2 Differentially expressed genes in leaf and culm

To isolate transcripts that were differentially expressed in leaf and culm tissues, the hybridisation signal intensities for individual cDNAs were analysed. Signal quantification revealed that hybridisation signals designated "medium" were consistently 2-fold higher than "low", while "high" signals were 2-3-fold higher than "medium". Using these criteria, cDNA clones that exhibited between a 2 to 5-fold higher signal in immature and mature leaf than that recorded after hybridisation with immature and maturing culm were considered to represent transcripts that were more abundant in sugarcane leaf tissues than culm tissues. Transcripts exhibiting higher expression levels in the culm were isolated in a similar manner. cDNAs exhibiting less than a 2-fold difference in signal intensity between the four tissue types tested were not considered to be diagnostic of a differential expression pattern. On this basis, 61 cDNAs were found to encode RNAs that accumulated at higher levels in leaf tissues (both immature and mature) and 25 cDNAs were identified as encoding culm-preferential transcripts (Tables 5.1 and 5.2). Expressed Sequence Tags (ESTs) were established for the differentially expressed cDNAs based on sequence similarities to known proteins. For the cDNAs expressed at high levels in immature and mature leaf, 46% exhibited significant matches to sequences of known gene products in the GenBank database (*E* values $<10^{-5}$) while 34% showed weak similarity only (*E* >0.01) (Table 5.1). No database match using the BLASTX algorithm was obtained for 20% of the cDNAs, suggesting that these transcripts may encode proteins of as yet unknown functions. Of the 25 cDNAs with elevated expression in the culm, 64% displayed significant sequence similarity to other genes, 28% had non-significant similarity and 8% were not similar to any known sequences in the database (Table 5.2). All sugarcane cDNAs showed primary homologies to genes from organisms other than sugarcane. Clones with no database match were used to search the dbEST database to identify possible matches to previously identified ESTs. Results indicated that although all clones did show some sequence similarity to a known EST, in most cases the sequence overlap was poor and not significant.

Table 5.1 Differentially expressed cDNAs more abundant in sugarcane immature and mature leaf than immature and maturing culm

ESTs were established by partial sequence homology searches with known gene sequences in the NCBI GenBank database. Sequence homology/match is the database sequence that the sugarcane cDNA is most similar to. The % sequence identity is at the amino acid level. The *E* value is the statistical indicator of the significance of the match between query and database sequence.

Clone	Sequence homology/match	GenBank Accession	% sequence identity	*E* value
Protein modification				
MB71	*Lycopersicon esculentum* ubiquitin-conjugating enzyme	S57619	58	$9.8X10^{-11}$
MC55	*Homo sapiens* ubiquitin-conjugating-enzyme E2	Q16781	96	$8.8X10^{-31}$
B69	*Pisum sativum* GTP-binding protein	D12542	84	$6.8X10^{-48}$
MC14	*Nicotiana tabacum* CND41, chloroplast nucleoid DNA binding protein	D26015	43	$2.2X10^{-7}$
MI67	*Arabidopsis thaliana* atranbp1b	X97378	58	$2X10^{-19}$
I68	*Arabidopsis thaliana* similar to gene pi010 glucosyltransferase	AAC83030	62	$2X10^{-14}$
MC31	*Oryza sativa* protein phosphatase 2A	U49113	95	$2.5X10^{-39}$
MI43	*Arabidopsis thaliana* putative protein kinase	AAD32292	52	0.001

Protein synthesis				
C63	*Cuscuta europaea* chloroplast 30S ribosomal protein	P46292	72	$8.8X10^{-31}$
J38	*Zea mays* chloroplast 30S ribosomal protein S7	P12339	92	$4X10^{-66}$
MB64	*Zea mays* 40S ribosomal protein S13	Q05761	35	$2.7X10^{-10}$
MI17	*Oryza sativa* ribosomal protein S3a, cytosolic	T02874	86	$8X10^{-48}$
MB47	*Zea mays* translation initiation factor 5A	Y07920	70	$6X10^{-33}$
MB80	*Arabidopsis thaliana* contains similarity to B. subtilis flagellar biosynthesis protein FLHA	AAB61047	42	0.52
Stress response				
C17	*Brassica napus* beta-1,3-glucanase homolog	S31712	45	$5.3X10^{-10}$
B77	*Arabidopsis thaliana* stage III sporulation protein J	S39321	69	$3.2X10^{-38}$
RNA synthesis				
MB53	Uukuniemi virus RNA polymerase (L protein)	P33453	52	0.012
MG76	*Arabidopsis thaliana* RNA polymerase I, II and III 16.5kD subunit	AAC28252	64	$1X10^{-23}$
DNA synthesis				
ME5	Bacteriophage CP-1 DNA polymerase	S51275	34	0.35
Transport				
14	*Hordeum vulgare* vacuolar ATPase B subunit isoform	L11862	84	$1.2X10^{-53}$
J41	*Hordeum vulgare* YLP (vacuolar H^+-ATPase E subunit-1)	U84268	78	$8X10^{-36}$
MI14	*Arabidopsis thaliana* putative ATPase (ISW2-like)	AAF08585	82	$7X10^{-45}$
MI26	*Zea mays* plasma membrane MIP protein	AAD29676	55	$4X10^{-8}$
MD80	*Drosophila melanogaster* neurotransmitter transporter	Y08362	40	0.12
Lipid metabolism				
A50	*Gossypium hirsutum* acyl-CoA-binding protein	U35015	73	$2.9X10^{-39}$
A3	*Brassica napus* 3-oxoacyl-[acyl-carrier protein] reductase	S22417	75	$2.6X10^{-42}$
Sucrose metabolism				
B63	*Hordeum vulgare* sucrose synthase	JT0280	80	$5.5X10^{-41}$
Cell division				
I58	*Zea mays* histone H2B.1	P30755	88	$1X10^{-40}$
Regulation				
H28	*Oryza sativa* thioredoxin H-type (TRX-H) (phloem sap 13 kD protein-1)	Q42443	73	$3X10^{-42}$
MH90	*Homo sapiens* Ca^{2+}-dependent activator protein for secretion	AAC14062	36	6.3
Miscellaneous				
MH80	*Arabidopsis thaliana* unknown protein	NP187332	35	$5X10^{-4}$
MH86	*Arabidopsis thaliana* unknown protein	NP180905	62	$1X10^{-26}$
MI21	*Arabidopsis thaliana* unknown protein	NP187473	68	0.044
F51	*Caenorhabditis elegans* hypothetical 337.6 kD protein T20G5.3 in chromosome III	P34576	60	0.44
H22	*Escherichia coli* hypothetical 31.4 kD protein in MHPT-ADHC intergenic region	P51025	27	4.0
H60	*Chlamydia muridarum* hypothetical protein	NP296498	58	$1X10^{-20}$
J22	*Schizosaccharomyces pombe* hypothetical protein SPBC56F2.01	CAA18880	34	0.92
J46	*Arabidopsis thaliana* hypothetical protein F28J12.260	T04556	40	9.4

MH83	*Rickettsia prowazekii* hypothetical protein RP591	B71664	31	0.96
MI44	*Arabidopsis thaliana* hypothetical protein F14M19.150	T04241	65	0.03
A16	*Homo sapiens* mucin	M57417	47	$2.4X10^{-5}$
A81	*Azorhizobium caulinodans* nodulation protein nodU	S35006	47	0.81
B42	*Bordetella pertussis* putative gtg start codon	X90711	41	0.00046
C34	*Homo sapiens* ryanodine receptor, skeletal muscle	P21817	58	0.002
J43	*Homo sapiens* elastin precursor, long splice form	EAHU	40	0.003
MB33	*Homo sapiens* The KIAA0149 gene product is related to Notch3	D63483	53	0.0015
MB43	*Dictyostelium discoideum* RTOA protein (RATIO-A)	P54681	42	0.89
MC53	*Escherichia coli* ORF4b	D16251	52	0.26
MI27	*Rattus norvegicus* olfactory receptor-like protein I8	P23271	33	1.1
No homology 12 clones				

Table 5.2 Differentially expressed cDNAs more abundant in sugarcane immature and maturing culm than immature and mature leaf

ESTs were established by partial sequence homology searches with known gene sequences in the NCBI GenBank database. Sequence homology/match is the database sequence that the sugarcane cDNA is most similar to. The % sequence identity is at the amino acid level. The *E* value is the statistical indicator of the significance of the match between query and database sequence.

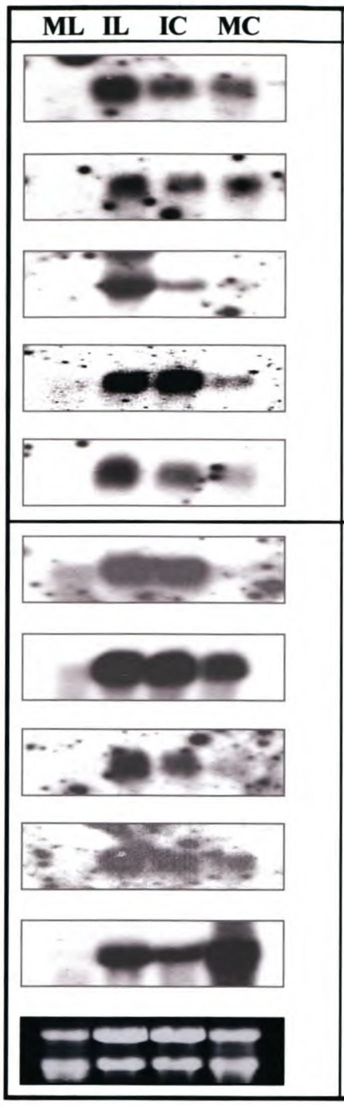
Clone	Sequence homology/match	Genbank Accession	% sequence identity	*E* value
Protein modification				
MA19	*Oryza sativa* ubiquitin-conjugating enzyme	D17786	91	0.00024
C82	*Homo sapiens* Ran=25kDa ras-related protein	239838	87	0.4
Protein synthesis				
MH23	*Oryza sativa* 60S ribosomal protein L5	P49625	79	$3X10^{-53}$
MH28	*Oryza sativa* 60S ribosomal protein L7A	P35685	95	$2X10^{-36}$
MI19	*Oryza sativa* translation initiation factor, eIF-5A	CAB96075	95	$8X10^{-50}$
Stress response				
MB26	*Phaseolus vulgaris* PVPR3	M75856	59	$4.2X10^{-9}$
MB73	*Solanum tuberosum* remorin	U72489	57	$3.7X10^{-15}$
ME42	*Oryza sativa* heat shock protein 82	P33126	94	$1.9X10^{-14}$
Regulation				
MI39	*Arabidopsis thaliana* hypothetical protein T15N24.20 (calcineurin B-like protein 3)	T08923	87	$1X10^{-54}$
MB23	*Oryza sativa* translationally controlled tumor protein homolog (TCTP)	P35681	78	$9.6X10^{-19}$
E70	*Homo sapiens* adenosine kinase	U33936	58	$4.1X10^{-34}$
Structural proteins				
MB42	*Rattus norvegicus* proline-rich protein	M86526	44	0.017
A47	*Homo sapiens* oligodendrocyte-specific proline-rich protein 2	C55663	36	0.15
DNA methylation				
F18	*Catharanthus roseus* 5-methyltetrahydopteroyl-triglutamate-homocysteine S-methyltransferase	S57636	79	$1.6X10^{-17}$

ME46	*Oryza sativa* S-adenosyl-methionine synthetase 1	P46611	96	$6.5X10^{-47}$
Nucleotide sugar biosynthesis				
D51	*Synechocystis* sp. dTDP-glucose 4-6-dehydratase	D90911	66	$2X10^{-26}$
H73	*Arabidopsis thaliana* dTDP-glucose 4-6-dehydratase homolog D18	S58282	56	$1X10^{-25}$
DNA-binding protein				
MA33	*Arabidopsis thaliana* replication control protein homolog	Z97336	77	$3X10^{-9}$
DNA modification				
MA22	*Zea mays* HMGd1	Y08807	85	$5.2X10^{-5}$
Electron transport				
MB75	*Bos taurus* cytochrome b5 reductase	RDB0B5	41	0.012
Miscellaneous				
MH40	*Plasmodium falciparum* rifin PFB1015w	A71601	37	8.9
MA40	*Caenorhabditis elegans* R11H6.2	Z93386	57	$3.2X10^{-10}$
MH1	*Oryza sativa* hypothetical protein	BAA92194	26	6.3
No homology				
2 clones				

All differentially expressed genes were assigned putative identities and grouped into general functional categories according to the proposed function of their homologues in other organisms (Tables 5.1 and 5.2). For the transcripts that accumulated at higher levels in leaf than culm tissues, many were homologous to a wide variety of genes associated with the maintenance and control of cellular metabolism (Table 5.1). These included genes with roles in protein modification (ubiquitin-conjugating enzyme; GTP-binding protein; protein phosphatase 2A), protein synthesis (ribosomal proteins; translation initiation factor 5A), RNA and DNA synthesis (RNA polymerase; DNA polymerase), lipid metabolism (acyl-CoA-binding protein; 3-oxoacyl-[acyl-carrier protein] reductase) and cell division (histone H2B.1). Also identified were various genes associated with transport (vacuolar ATPase; plasma membrane MIP protein), stress responses (beta-1,3-glucanase homolog; stage III sporulation protein J), sucrose metabolism (sucrose synthase) and regulation (thioredoxin H-type). Approximately half of the ESTs could not be assigned a putative identity either due to sequence overlap with uncharacterised proteins (Miscellaneous) or no database match. Significant sequence overlaps to the database homologues were obtained for three of the uncharacterised ESTs (MH86; H60; A16), suggesting that they represent unidentified proteins whose sequence may have been conserved during evolution (Table 5.1). For those genes exhibiting higher expression levels in culm tissues than leaf tissues database searches revealed that, similar to results observed for leaf transcripts, many exhibited sequence similarity to genes associated with general metabolic functions (Table 5.2). This group also comprised genes involved with protein modification and protein synthesis and included those coding for proteins such as a ubiquitin-conjugating enzyme and ribosomal proteins. Other genes encoding structural proteins (proline-rich protein) and those involved with nucleotide sugar biosynthesis (dTDP-glucose 4-6-dehydratase), DNA-binding (replication control protein homolog) and electron transport (cytochrome b5 reductase) were also detected in the culm. Most of the remaining genes could be categorised according to putative roles in stress responses (PVPR3; remorin; heat shock protein), regulation (calcineurin B-like protein 3; TCTP; adenosine kinase), DNA methylation (5-methyltetrahydopteroyl-triglutamate-homocysteine S-methyltransferase; S-adenosyl-methionine synthetase 1) and DNA modification (HMGd1). A small proportion of cDNAs could not be assigned a putative identity due to sequence overlap with poorly characterised genes, with one sugarcane cDNA (MA40) displaying a significant database match.

5.4.3 Northern blot analysis

To evaluate the differential expression results obtained by cDNA macroarray hybridisation, the expression patterns of 10 cDNA clones were analysed by Northern blot hybridisation (Fig. 5.2). Expression levels were evaluated in immature leaf, mature leaf, immature culm and maturing culm. For five clones that macroarray hybridisation indicated as exhibiting higher expression levels in leaf tissues than culm tissues (Table 5.1), Northern analysis indicated very low or undetectable expression of the corresponding transcripts in mature leaves (Fig. 5.2*a*). High transcript levels were observed in immature leaf for all clones tested. Some clones exhibited similar expression levels in immature leaf and immature culm but in general, expression in the culm tissues was lower than that detected in leaf tissues. For five clones recorded as abundantly expressed in culm tissues (Table 5.2), Northern analysis confirmed the results from the macroarray hybridisation except for two clones (MI39; MH28) that showed lower expression levels in maturing culm than immature culm (Fig. 5.2*b*). However, for all clones tested, high expression levels were also detected in the immature leaf, although very low levels were detectable in mature leaf tissue.



Clone	Putative identity	
J41	Vacuolar H$^+$-ATPase E subunit-1	*(a)*
MI67	Atranbp1b	
B63	Sucrose synthase	
H28	Thioredoxin H-type	
MC55	Ubiquitin-conjugating-enzyme E2	
MI39	Calcineurin B-like protein 3	*(b)*
MA19	Ubiquitin-conjugating enzyme	
MH28	60S ribosomal protein L7A	
MB26	PVPR3	
ME46	S-adenosyl-methionine synthetase 1	

Fig. 5.2 Northern blot analysis of selected differentially expressed cDNA clones Expression patterns were examined for individual clones in sugarcane mature leaf (ML), immature leaf (IL), immature culm (IC) and maturing culm (MC). Clones were selected according to results obtained from macroarray screening as either abundantly expressed in immature and mature leaf *(a)*, or immature and maturing culm *(b)*. The bottom panel indicates ethidium-bromide stained rRNA to demonstrate equal sample loading.

5.5 DISCUSSION

In this study, the expression of 1000 random sugarcane gene sequences was investigated in leaf and culm tissues of varying maturation stages. For the first time, information about the expression behaviour of a large number of genes has been obtained for sugarcane, a plant for which little is known regarding the genetic control of growth and development. By randomly selecting clones from immature leaf and maturing culm total cDNA libraries for expression analysis it was anticipated that this would provide some indications, at the level of gene expression, of the morphological and physiological differences between leaf and culm developmental states. Hybridisation signal intensity analyses did suggest that there were differences in the abundance of individual transcripts between the leaf and culm (Fig. 5.1).

Results indicated that the homologous transcripts for the cDNA clones analysed were present in all tissue types tested and no cDNAs displaying tissue-specific expression patterns were detected. Only a small percentage (8.6%) of the cDNAs were found to reflect differential levels of gene expression between leaf and culm tissues. This is in contrast to research on *Arabidopsis* where comparisons between the expression patterns of genes in leaves, roots, flower buds and open flowers revealed large numbers (up to 34%) of differentially expressed genes (Ruan et al., 1998).

The putative identities obtained through gene sequence database searches for those differentially expressed sugarcane genes allows some assessment of gene function in distinct leaf and culm developmental stages. For those genes more highly expressed in immature and mature leaf (Table 5.1), putative identities suggest that many of them are associated with active cell division and growth. Expression analysis of a vacuolar H^+-ATPase (Takanokura et al., 1998) and the GTP-binding protein atranbp1 (Haizel et al., 1997) indicated that these genes exhibited the highest level of expression in meristematic tissues. Similarly, the thioredoxin h gene is preferentially expressed in immature leaves (Ishiwatari et al., 2000) while many genes for ribosomal proteins have been found in growth-phase callus of rice (Yamamoto and Sasaki 1997). The expression of sugarcane cDNAs homologous to these proteins as well as others associated with cellular metabolism such as fatty acid biosynthesis, cell division and RNA synthesis (Table 5.1) is therefore expected to be high in meristematic and growing tissues. For those cDNAs that accumulated preferentially in leaf tissues and whose coding sequences did not match genes with known functions, these could represent novel genes with important roles in sugarcane leaf development. Extensive characterisation is necessary to describe putative functions for these genes in leaf growth and development.

Only one gene directly associated with sucrose metabolism (sucrose synthase) was differentially expressed (Table 5.1). This corresponded to the SS2 sucrose synthase isoform identified previously in sugarcane (Buczynski et al., 1993). The mRNA expression profile detected in this study (Fig. 5.2*a*) corresponded favourably with the SS2 protein expression profile determined previously (Buczynski et al., 1993), with the most abundant expression being detected in immature leaf tissue. During sugarcane culm development, the rate of sucrose accumulation increases with internode expansion and maturation (Whittaker and Botha 1997). Physiological studies of key enzymes associated with sucrose metabolism in sugarcane such as sucrose synthase, sucrose phosphate synthase (SPS) and the various invertases (neutral, soluble acid and cell-wall bound) have established that enzyme activities vary in internodes of differing maturity (for review, see Moore 1995). However, none of the highly expressed transcripts detected in the culm in this study were homologous to genes encoding proteins known to have roles in sucrose accumulation. Most of the genes identified coded for proteins associated with a wide variety of cellular functions (Table 5.2). Attempts to characterise the role of these genes in other plants have revealed responses to a diverse range of stimuli. For example, the expression of S-adenosyl methionine synthetase (SAM) changes significantly in response to treatment with fungal elicitors (Kawalleck et al., 1992), salt stress (Espartero et al., 1994) and sucrose accumulation (Winters et al., 1995). Furthermore, cold, drought and wounding have been shown to increase the mRNA levels for calcineurin B-like proteins in *Arabidopsis* (Kudla et al., 1999). Fungal elicitor treatment and wounding also resulted in an accumulation of transcripts for PVPR3 in *Phaseolus vulgaris* (Sharma et al., 1992). Extensive research is therefore necessary to establish the nature of the role of these genes during sugarcane culm maturation and sucrose accumulation.

Hybridisation data obtained in this study using "reverse Northern" analysis could only provide a general assessment of transcript abundance and included a small proportion of results that did not coincide with conventional Northern hybridisation (Fig. 5.2). It is possible that more sensitive methods will be required to detect developmental stage-specific genes, many of which may be rarely expressed. cDNA macroarray screening does offer, however, an effective preliminary screening procedure that provides valuable new data on the expression patterns of genes and can pinpoint promising candidates for further detailed analyses, although results require independent confirmation by techniques such as Northern analyses (Girke et al., 2000).

The diversity of differentially expressed genes identified in this study provides new insights into the genetic regulation of sugarcane leaf and culm development. In particular, the lack of association between the types of genes identified as preferentially expressed in the culm and available biochemical and physiological data regarding sucrose accumulation in sugarcane illustrate the complexity of sugarcane metabolism at the gene level. In-depth molecular analyses are required to develop an understanding of how sucrose metabolism is regulated at the level of gene transcription and translation.

5.6 REFERENCES

Altschul S, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Buczynski SR, Thom M, Chourey P and Maretzki A (1993) Tissue distribution and characterization of sucrose synthase isozymes in sugarcane. J Plant Physiol 142: 641-646.

Carson DL and Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. Crop Sci 40: 1769-1779.

Decraene C, Reguigne-Arnould I, Auffray C and Piétu G (1999) Reverse transcription in the presence of dideoxynucleotides to increase the sensitivity of expression monitoring with cDNA arrays. BioTechniques 27: 962-966.

Delseny M, Salses J, Cooke R, Sallaud C, Regad F, Lagoda P, Guiderdoni E, Ventelon M, Brugidou C and Ghesquière (2001) Rice genomics: Present and future. Plant Physiol and Biochem 39: 323-334.

Desprez T, Amselem J, Caboche M and Höfte H (1998) Differential gene expression in *Arabidopsis* monitored using cDNA arrays. Plant J 14(5): 643-652.

Draye X, Lin Y-R, Qian X-y, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren S-x, Wing RA and Paterson AH (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the Sorghum genome as a foundation. Plant Physiol 125: 1325-1341.

Espartero J, Pintor-Toro JA and Pardo JM (1994) Differential accumulation of S-adenosylmethionine synthetase transcripts in response to salt stress. Plant Mol Biol 25: 217-227.

Girke T, Todd J, Ruuska S, White J, Benning C and Ohlrogge J (2000) Microarray analysis of developing Arabidopsis seeds. Plant Physiol 124: 1570-1581.

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C and Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organisation in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Haizel T, Merckle T, Pay A, Fejes E and Nagy F (1997) Characterization of proteins that interact with the GTP-bound form of the regulatory GTPase Ran in *Arabidopsis.* Plant J 11(1): 93-103.

Holmes DS and Quigly M (1981) A rapid boiling method for the preparation of bacterial plasmids. Anal Biochem 114: 193-197.

Ishiwatari Y, Nemoto K, Fujiwara T, Chino M and Hayashi H (2000) *In situ* hybridisation study of the rice phloem thioredoxin h mRNA accumulation-possible involvement in the differentiation of vascular tissues. Physiol Plantarum 109: 90-96.

Kawalleck P, Plesch G, Hahlbrock K and Somssich I (1992) Induction by fungal elicitor of S-adenosyl-L-methionine synthetase and S-adenosyl-L-homocysteine hydrolase mRNAs in cultured cells and leaves of *Petroselinum crispum.* Proc Natl Acad Sci USA 89: 4713-4717.

Kudla J, Xu O, Harter K, Gruissem W and Luan S (1999) Genes for calcineurin B-like proteins in *Arabidopsis* are differentially regulated by stress signals. Proc Natl Acad Sci USA 96: 4718-4723.

Ming R, Liu S-C, Lin Y-R, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE and Paterson AH (1998) Detailed alignment of Saccharum and Sorghum chromosomes: Comparative organisation of closely related diploid and polyploid genomes.

Genetics 150: 1663-1682.

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22: 661-679.

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E and Somerville C (1994) Genes Galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241-1255.

Ruan Y, Gilmore J and Conner T (1998) Towards *Arabidopsis* genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. Plant J 15(6): 821-833.

Sambrook J, Fritsch EF and Maniatis T (1989) Molecular cloning, a laboratory manual. (Cold Spring Harbor Press, Cold Spring Harbor, NY).

Schmidt R (2000) Synteny: recent advances and future prospects. Current Opin Plant Biol 3: 97-102.

Sharma YK, Hinojos CM and Mehdy MC (1992) cDNA cloning, structure, and expression of a novel pathogenesis-related protein in bean. Mol Plant-Microbe Int 5(1): 89-95.

Takanokura Y, Komatsu A, Omura M and Akihama T (1998) Cloning and expression analysis of vacuolar H$^+$-ATPase 69-kDa catalytic subunit cDNA in citrus (*Citrus unshiu* Marc.). Biochim et Biophys Acta 1414: 265-272.

van Dillewijn C (1952) Growth: general, grand period, growth formulae. In: van Dillewijn C (ed) Botany of sugarcane, vol 1. Veenen and Zonen, The Netherlands, pp 97-162.

Whittaker A and Botha FC (1997) Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. Plant Physiol 115: 1651-1659.

Willmann MR (2001) *Arabidopsis* enters the post-sequencing era. Trends in Plant Sci 6(2): 51.

Winters AL, Gallagher J, Pollock CJ and Farrar JF (1995) Isolation of a gene expressed during sucrose accumulation in leaves of *Lolium temulentum* L.. J Exp Bot 46: 1345-1350.

Yamamoto K and Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35: 135-144.

Yuan Q, Quackenbush J, Sultana R, Pertea M, Salzberg SL and Buell CR (2001) Rice Bioinformatics: Analysis of rice sequence data and leveraging the data to other plant species. Plant Physiol 125: 1166-1174.

CHAPTER 6

SUGARCANE ESTs DIFFERENTIALLY EXPRESSED IN IMMATURE AND MATURING INTERNODAL TISSUE

[Plant Science 162(2): 289-300 (2002)]

6.1 ABSTRACT

Two subtracted cDNA libraries were constructed by reciprocal subtractive hybridisation between sugarcane immature (low sucrose-accumulating) and maturing (high sucrose-accumulating) internodal tissue. The subtracted libraries contained high, moderate and low abundance transcripts. To isolate cDNAs differentially expressed during culm maturation, 400 random clones (200 from each library) were systematically arrayed onto nylon filters and screened with total cDNA probes prepared from immature and maturing culm poly $(A)^+$ RNA. Results indicated that 36% and 30% of the total number of cDNAs analysed were preferentially expressed in the immature and maturing culm, respectively. Northern analysis of selected clones confirmed culm developmental stage-specific and -preferential expression for most of the clones tested. ESTs generated by partial sequence analysis for all 132 differentially expressed clones indicated 95 unique transcripts. Partial sequence information could assign putative identities to 66% of the differentially expressed ESTs. The majority of ESTs with a putative identity were homologous to genes associated with cell wall metabolism, carbohydrate metabolism, stress responses and regulation, where the specific ESTs identified in the immature and maturing culm were distinct from each other. No developmentally regulated ESTs directly associated with sucrose metabolism were detected. This suggests that growth and maturation of the sugarcane culm is associated with the expression of genes for a variety of processes. This study demonstrates that a combination of cDNA subtraction with macroarray screening is an effective strategy to identify and analyse candidate developmentally regulated genes in sugarcane.

6.2 INTRODUCTION

Plant growth and development is characterised by a multitude of biochemical and physiological changes. Molecular observations provide insight into the genetic regulation of developmental processes, thereby offering the potential to determine when and where specific genes associated with growth and maturation are expressed. In sugarcane, growth and development is characterised by the accumulation of sucrose in developing internodes (Moore 1995). Knowledge about genes expressed during culm maturation is, however, limited. The identification of genic sequences exhibiting culm stage-specific expression will provide some indications of the genetic basis of culm development and, also, will be a source of candidate genes for use in genetic engineering programmes or for genetic marker development.

The rapid accumulation of large amounts of DNA sequence data has made the identification of Expressed Sequence Tags (ESTs) a popular route towards gene detection and expression profile status. There are numerous examples where EST analysis has been applied towards comparative gene expression profiling of specific tissues and growth stages in plants (Cooke et al., 1996; Yamamoto and Sasaki 1997; Ablett et al., 2000). Differential screening is also another widely used technique and has been used successfully to catalogue genes that are developmentally regulated (Wyrich et al., 1998; Nam et al., 1999). However, it is known that both random EST analysis and differential screening are unsuitable for the detection of differentially expressed genes with low transcript abundance (Newman et al., 1994; Wyrich et al., 1998). Studies of developmental gene regulation in ripening grape berries have indicated that stage-specific genes are expressed at high, moderate and low levels (Davies and Robinson 2000). Similarly, developmental changes from gametophyte to sporophyte in *Porphyra purpurea* have also been shown to be associated with high and low abundance phase-specific mRNAs (Liu et al., 1994). Subtractive cDNA cloning is a powerful procedure that offers the advantage of being able to detect both low and high abundance transcripts with differential expression profiles (Liu et al., 1994; Kamakura et al., 1999).

The isolation of differentially expressed clones from subtracted cDNA libraries is based traditionally on a classical differential hybridisation screening. However, this is a time-consuming and material-intensive way to analyse expression patterns due to repeated isolation of cDNAs to saturate the screening (Kozian and Kirschbaum 1999). The recent developments in cDNA array technology offer a potentially more efficient alternative to traditional library screening for detection of differentially expressed clones, as the expression profiles of multiple cDNA fragments are generated simultaneously through a single hybridisation event. Successful application of this approach has been demonstrated in *Arabidopsis* where differential screening of high-density colony filters prepared from an equalised inflorescence cDNA library allowed the systematic isolation of novel genes in the inflorescence (Takemura et al., 1999).

In this study, two subtracted cDNA libraries have been generated by cross-subtraction of mRNA-derived cDNA between immature (low sucrose-accumulating) and maturing (high sucrose-accumulating) internodal tissue. Macroarrays have been used to evaluate the success of the subtractions and to identify the differentially expressed clones. We demonstrate that the enrichment provided by cDNA subtractive cloning, in combination with the efficiency of cDNA macroarray screening, is an effective approach towards identifying developmentally regulated genes in sugarcane.

6.3 MATERIALS AND METHODS

6.3.1 Construction of subtracted cDNA libraries

Two subtracted cDNA libraries were constructed by reciprocal subtractive hybridisation between immature and maturing culm using the PCR-based Subtractive cDNA Cloning technique (Patel and Sive 1996). Immature culm is defined as internode no. 2 and maturing culm as internode no. 7, where internode no.1 is the internode attached to the leaf with the uppermost visible dewlap (van Dillewijn 1952). Total RNA was extracted from internode no. 2 and internode no. 7 according to the method described in (Carson and Botha 2000). Poly $(A)^+$ RNA was isolated using the Dynabeads® mRNA Purification Kit (Dynal A.S, Oslo, Norway), according to the manufacturer's instructions. Double-stranded cDNA was prepared from poly $(A)^+$ RNA as described previously (Carson and Botha 2000) except that Expand Reverse Transcriptase (Roche Diagnostics GmbH, Mannheim, Germany) was used for first-strand cDNA synthesis. The reaction was performed according to the manufacturer's protocol. Six rounds of subtractive cDNA hybridisations were performed, alternating between three short and three long, and were executed exactly according to the protocol. The protocol was modified for cloning of the subtracted products for library construction. The immature culm subtracted cDNA products were restriction digested with EcoRI and cloned into the EcoRI site of the Lambda ZAP II phage cloning vector (Stratagene, La Jolla, CA). For the maturing culm library, the subtracted cDNA products were first blunt-ended using the Klenow fragment of DNA polymerase I and then ligated to an adapter set that contained an EcoRI restriction site. After restriction digestion with EcoRI, the cDNA products were cloned into the EcoRI site of the Lambda ZAP II cloning vector.

6.3.2 Preparation of bacterial clones

Aliquots of the immature and maturing culm subtracted cDNA libraries were plated out onto solid NZY medium and single plaques randomly picked and stored in SM buffer (100mM NaCl, 8mM $MgSO_4.7H_2O$, 20mM Tris-HCl pH 7.5, 0.01% gelatin) at 4 °C. Phagemids (pBluescript SK(-)) plus inserts were excised from individual phages according to the manufacturer's instructions (Stratagene, La Jolla, CA, USA). Individual phagemid clones were plated out onto solid Luria Bertani (LB) medium containing 50μg/ml ampicillin. Bacterial glycerol stocks were prepared from phagemid clones by mixing liquid bacterial cultures grown overnight in LB medium containing 50μg/ml ampicillin with sterile glycerol in a 5.7:1 ratio and flash-freezing in liquid nitrogen. Phagemid DNA was isolated from a 5 ml overnight liquid bacterial culture using a Rapid Plasmid Isolation Protocol (Holmes and Quigly 1981) and purified through QIAquick spin columns (Qiagen GmbH, Hilden, Germany).

6.3.3 Production of cDNA macroarrays and Northern blots

For cDNA macroarray preparation, bacterial clones were spotted onto Hybond™- N+ nylon membranes (Amersham International, Buckinghamshire, United Kingdom) using a QBOT (Genetix, Hampshire, United Kingdom). Samples were lysed with 0.5M NaOH, 1.5M NaCl for 5 minutes, neutralised with 1.5M NaCl, 0.5M Tris-HCl (pH

7.2) and allowed to air-dry. The DNA was then denatured with 0.4M NaOH for 5 minutes and neutralised with 5X SSPE.

For the preparation of Northern blots, total RNA (10μg) was electrophoresed in 1.2% agarose formaldehyde gels and transferred to a Hybond™- N+ membrane (Amersham International, Buckinghamshire, United Kingdom) according to the manufacturer's instructions.

6.3.4 RNA extraction and probe synthesis

Total RNA was extracted as described previously (Carson and Botha 2000). Poly $(A)^+$ RNA was isolated from 75μg total RNA using Dynabeads® (Dynal A.S, Oslo, Norway) according to the manufacturer's instructions. For the synthesis of total cDNA probes, a modification of the method described in (Sambrook et al., 1990) was used. A mixture was prepared containing 12.5μg random hexamer primers (Amersham Pharmacia Biotech Inc, Piscataway, NJ), 20mM each dCTP, dGTP, dTTP, 120μM dATP and 200μM ddCTP (Decraene et al, 1999). This mixture was dried down to complete dryness in a SpeedVac (Savant Instruments Inc., Holbrook, NY) and resuspended in 5X Expand Reverse Transcriptase buffer (Roche Diagnostics GmbH, Mannheim, Germany), 10mM DTT (final concentration) and DEPC-treated water to a volume of 7.5μl. A 1μg poly $(A)^+$ RNA sample was denatured for 5 minutes at 70°C, cooled on ice and added to the mixture with 20U RNase inhibitor (Roche Diagnostics GmbH, Mannheim, Germany), 100U Expand Reverse Transcriptase (Roche Diagnostics GmbH, Mannheim, Germany) and 50μCi $[\alpha\text{-}^{33}P]$dATP (2500 Ci/mmol) to a final volume of 20μl. The mixture was incubated at 30°C for 10 minutes, followed by 42°C for 45 minutes. The reaction was stopped by the addition of 1.0μl 0.5M EDTA (pH 8.0) and 1.0μl 10% (w/v) SDS. The RNA was hydrolysed by the addition of 3μl of 3N NaOH and incubation for 30 minutes at 68°C. The mixture was allowed to cool to room temperature and then mixed with 10μl 1M Tris-HCl (pH 7.4) and 3μl 2N HCl. The probe was purified by phenol:chloroform extraction and ethanol precipitated to remove unincorporated nucleotides.

Specific cDNA probes for use in Northern blot analysis were labelled with [α-^{32}P]dCTP (3000Ci/mmol) by random primer labelling using the Megaprime™ DNA Labelling system (Amersham International, Buckinghamshire, United Kingdom) according to the manufacturer's protocol. Template DNA was prepared by specific PCR amplification of cDNA inserts from phagemid DNA, using the M13 Forward and Reverse primers. Amplified inserts were purified using QIAquick spin columns (Qiagen GmbH, Hilden, Germany) prior to use. Probes were purified to remove unincorporated nucleotides using NucTrap® Probe Purification Columns (Stratagene, La Jolla, CA).

6.3.5 Hybridisation procedures

cDNA macroarray filters were prehybridised overnight at 65°C in a solution of 0.5M sodium phosphate buffer (pH 7.2), 7% (w/v) SDS, 0.9mM EDTA and 10μg/ml denatured salmon sperm DNA (final concentrations). Hybridisation was performed overnight at 65°C with fresh solution minus the denatured salmon sperm DNA. Filters were washed twice with 1X SSC, 0.1% (w/v) SDS for 20 minutes at 65°C, followed by twice with 0.5X SSC, 0.1% (w/v) SDS for 20 minutes at 65°C. Hybridised filters were exposed to a Super Resolution Cyclone Phosphor Screen (Packard Instrument Company, Meriden, CT) for 4-16 hours and data captured and analysed with OptiQuant™ software (Packard Instrument Company, Meriden, CT). The signal intensity for each clone was recorded manually as "high", "medium" or "low". Densitometric analysis using the OptiQuant™ software performed on a random selection of 48 samples established that there was a significant difference in hybridisation signal intensity between the designated three categories.

Northern blot hybridisation was performed exactly according to the protocol supplied with the Hybond™- N+ membrane (Amersham International, Buckinghamshire, United Kingdom). Hybridised membranes were exposed to phosphorscreens as described above. Membranes were also exposed to X-ray film for various times for autoradiography.

6.3.6 DNA sequencing

DNA sequencing was performed by dye terminator cycle sequencing using the BigDye™ Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), followed by ethanol precipitation of the extension products. Both procedures were performed according to the manufacturer's instructions. The M13 Forward and Reverse primers were used to generate partial sequences for all isolated cDNAs. Cycle sequencing was performed in a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) and sequence analysis was performed using an ABI Prism 310 Genetic Analyser (Applied Biosystems, Foster City, CA).

6.3.7 Sequence data analysis

Sequences were edited using Sequence Navigator software (Applied Biosystems, Foster City, CA) to remove vector and ambiguous sequences. Analyses for cDNA sequence similarity to database sequences were conducted by comparison with the nonredundant protein databases and dbEST database using the BLASTX and BLASTN (Altschul et al., 1990) e-mail server, respectively, provided by NCBI (blast@ncbi.nlm.nih.gov). The degree of sequence similarity between the sugarcane cDNA clone and a known sequence was represented by the *E* value. Scores below 10^{-5} for the *E* value were considered as significant and indicated homology between the sugarcane sequence and the database sequence. The EST was identified as the protein with the lowest *E* value among the candidate proteins generated by the database search.

6.4 RESULTS

6.4.1 Characterisation of the subtracted cDNA libraries

Total cDNA was prepared from immature and maturing culm poly $(A)^+$ RNA and digested by restriction endonucleases with 4-bp recognition sequences to obtain short cDNA fragments. PCR amplification of total cDNA resulted in fragments ranging in size from 150bp to 1500bp, with an average size of 300bp. Restricting cDNA prior to performing the subtractive hybridisations prevented preferential amplification of naturally small cDNA molecules during the multiple rounds of PCR amplification, as required by the subtraction scheme used. After six rounds of subtraction, PCR amplification prior to cloning the subtracted products indicated the cDNA size range had altered slightly to between 230bp and 1200bp, with an average size of 350bp for both immature and maturing culm cDNA. After cloning of subtracted cDNA, recombinant plaques were randomly selected from both subtracted cDNA libraries and insert sizes tested by specific PCR amplification using the universal M13 Forward and Reverse primers. Results indicated that of the 136 clones selected from the immature culm library, 94% contained inserts with an average size of 350bp. For the maturing culm library, all 106 clones tested contained inserts with an average size of 400 bp. The titer of the unamplified libraries was $1.2X10^6$ pfu/ml and $1.4X10^6$ pfu/ml, for the immature culm and maturing culm subtracted library, respectively. Blue/white plaque selection following incubation of an aliquot of each library in the presence of X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) and IPTG (isopropyl-β-D-thiogalactoside) indicated 90% and 94% recombinants for the immature and maturing culm library, respectively.

DNA sequence analysis was performed on a random selection of subtracted cDNA clones as a preliminary assessment of library quality. The M13 Reverse primer was used to generate partial cDNA sequences that were translated into all six translational reading frames and compared to the nonredundant protein sequence databases in GenBank. Sequence similarity searches resulted in the putative identification of 60 Expressed Sequence Tags (ESTs) from the immature culm and 72 from the maturing culm subtracted cDNA library (results not shown). Although putative identities of individual clones is insufficient to evaluate the success of the subtraction they do provide some preliminary information. Results indicated that in the random sample selected, the identities of the specific ESTs detected in the two populations were distinct from each other. These pilot results suggested that enrichment for cDNA sequences preferentially expressed in the immature culm and maturing culm had occurred and that the subtracted libraries were suitable for differential expression analysis.

6.4.2 Detection of differentially expressed ESTs by cDNA macroarray hybridisation analysis

To assess the success of the subtractions and to identify differentially expressed sequences, a nylon filter array was constructed using a random selection of clones from the immature and maturing culm subtracted cDNA libraries and screened with total cDNA probes prepared from immature and maturing culm poly $(A)^+$ RNA. The array comprised 400 clones, 200 from each library, approximately one-third of which were ESTs while the remainder were anonymous. The total cDNA probes were prepared from unsubtracted material. Hybridisation signals were recorded as described in Materials and Methods for each clone after screening with both probes. Clones exhibiting a difference in signal intensity of two-fold or more between the two probes were considered as representative of a significant variation in the abundance of corresponding transcripts in the immature and maturing culm. On this basis it was established that 36% of clones selected from the immature culm subtracted library were preferentially expressed in the immature culm. Likewise, 30% of maturing culm clones exhibited tissue-preferential expression. The remaining clones exhibited similar expression levels in both culm developmental stages.

ESTs were established for immature culm and maturing culm differentially expressed clones (Tables 6.1 and 6.2). Of the 72 immature culm-preferential ESTs, 47% could be assigned a putative identity based on strong sequence homology to genes of known function ($E \leq 10^{-5}$). For 53% of the clones, the sequence homology between the sugarcane cDNA and the database match was weak ($E \geq 10^{-5}$) therefore the putative identity could not be considered as significant. For all immature culm ESTs, regardless of the significance of the match, 63% were homologous to known plant genes while 21% matched to non-plant genes. The remaining 16% of clones exhibited no sequence similarity to any published sequences in the GenBank non-redundant peptide database but did display similarities to plant ESTs in the dbEST database (results not shown). Maturing culm-preferential ESTs were analysed in a comparable manner. In this instance, 63% of the 60 preferentially expressed ESTs exhibited sequence homology to known genes, while 37% could not be well characterised due to weak database matches. The percentage of maturing culm ESTs similar to known plant genes was equivalent to those observed for the immature culm (62%), but 33% of the ESTs were similar to non-plant genes. Only 5% of maturing culm ESTs failed to match any sequences in the GenBank peptide database but could all be putatively identified by homology to known plant ESTs (results not shown). None of the differentially expressed ESTs exhibited matches to sugarcane genes.

Table 6.1 Putative identifications of ESTs preferentially expressed in sugarcane immature internodal tissue

Sugarcane ESTs were assigned a putative identity based on partial sequence homology searches with known gene sequences in the NCBI GenBank database. Sequence homology/match is the database sequence that the sugarcane cDNA was most similar to. Accession number represents the number assigned by GenBank for individual entries. The % sequence identity is at the amino acid level and is calculated from sequence alignment between the sugarcane partial cDNA sequence and the GenBank entry by the BLASTX algorithm. The *E* value is the statistical indicator of the significance of the match between query and database sequence.

Clone	Sequence homology/match	Accession number	% sequence identity	*E* value
Cell wall metabolism				
I2-6	*Glycine max* UDP-glucose dehydrogenase	U53418	78	$9.7X10^{-44}$
I2-7	*Glycine max* UDP-glucose dehydrogenase	U53418	92	$4.2X10^{-54}$
I2-13	*Glycine max* UDP-glucose dehydrogenase	U53418	96	$1.4X10^{-9}$
I2-155	*Glycine max* UDP-glucose dehydrogenase	T08818	90	$2X10^{-43}$
I2-160	*Glycine max* UDP-glucose dehydrogenase	Q96558	95	$1X10^{-39}$
I2-190	*Glycine max* UDP-glucose dehydrogenase	Q96558	91	$7X10^{-40}$
I2-241	*Glycine max* UDP-glucose dehydrogenase	Q96558	89	$4X10^{-41}$
I2-261	*Glycine max* UDP glucose dehydrogenase	Q96558	70	$6X10^{-28}$
I2-163	*Arabidopsis thaliana* cellulose synthase catalytic subunit	AF088917	89	$3X10^{-37}$
I2-307	*Arabidopsis thaliana* cellulose synthase catalytic chain	T08583	88	$3X10^{-41}$
I2-184	*Zea mays* cellulose synthase-8	AF200532	93	$3X10^{-38}$
I2-56	*Hordeum vulgare* xyloglucan endotransglycosylase	X93173	56	$5.5X10^{-13}$
Carbohydrate metabolism				
I2-38	*Arabidopsis thaliana* trehalose-6-phosphate synthase homolog	Z97344	64	$1.7X10^{-40}$
I2-245	*Arabidopsis thaliana* trehalose-6-phosphate synthase homolog	T01494	75	$2X10^{-42}$
I2-249	*Arabidopsis thaliana* trehalose-6-phosphate synthase homolog	T01494	75	$4X10^{-43}$
I2-114	*Hordeum vulgare* glyceraldehyde 3-phosphate dehydrogenase, cytosolic	P26517	76	0.0087
I2-147	*Zea mays* glyceraldehyde-3-phosphate dehydrogenase 2	PQ0178	100	0.036
I2-166	*Zea mays* glyceraldehyde-3-phosphate dehydrogenase 2	PQ0178	100	$1X10^{-5}$
Stress responses				
I2-16	*Medicago sativa* environmental stress-induced protein	M74191	39	0.021

I2-35	*Medicago sativa* environmental stress-induced protein	M74191	44	0.074
I2-57	*Medicago sativa* environmental stress-induced protein	M74191	50	0.0059
I2-59	*Medicago sativa* environmental stress-induced protein	M74191	50	0.023
I2-104	*Medicago sativa* environmental stress-induced protein	M74191	53	0.029
I2-28	*Medicago falcata* abscisic acid and environmental stress inducible protein	Q09134	40	2.2X10^{-10}
I2-63	*Arabidopsis thaliana* drought-induced protein Di19	S51478	50	0.00089
I2-74	*Arabidopsis thaliana* drought-induced protein Di19	S51478	45	0.018
I2-238	*Arabidopsis thaliana* drought-induced protein Di19	S51478	44	9X10^{-10}
I2-254	*Arabidopsis thaliana* drought-induced protein Di19	S51478	43	1X10^{-4}
I2-292	Contains similarity to *Arabidopsis thaliana* drought-induced protein Di19	AF075597	57	0.16
I2-311	Contains similarity to *Arabidopsis thaliana* drought-induced protein Di19	AF075597	42	0.81
I2-143	*Arabidopsis thaliana* low temperature and salt responsive protein LT16A	AF104221	72	3X10^{-16}
I2-283	*Oryza sativa* similar to *Arabidopsis thaliana* low temperature and salt responsive protein LTI6B	AP002070	87	4X10^{-22}
I2-107	*Arabidopsis thaliana* TMV resistance protein homolog	Z97336	71	1.6X10^{-15}
I2-11	*Nicotiana tabacum* glutamate decarboxylase	U54774	54	0.4
Regulation				
I2-192	*Arabidopsis thaliana* putative senescence-associated protein 12	T00840	55	4X10^{-33}
I2-296	*Arabidopsis thaliana* putative senescence-associated protein 12	AC003952	61	4X10^{-25}
I2-225	*Hevea brasiliensis* latex-abundant protein	AF098458	57	1X10^{-28}
I2-259	*Hevea brasiliensis* latex-abundant protein	AF098458	51	1X10^{-18}
I2-65	*Lycopersicon esculentum* ubiquitin-conjugating enzyme E2	P35135	92	1.1X10^{-54}
I2-299	*Arabidopsis thaliana* calmodulin-related protein 2, touch-induced	P25070	58	1X10^{-8}
I2-300	*Oryza sativa* r40c1 protein	T03911	77	9X10^{-41}
I2-154	*Saccharomyces cerevisiae* protein kinase 1	M69017	41	9.3
I2-267	*Saccharomyces cerevisiae* ornithine aminotransferase	X06790	49	6X10^{-18}
I2-145	*Callerya reticulata* maturase-like protein	AF142733	46	3.2
I2-159	*Tristaniopsis laurina* maturase K	AF184710	40	3.2
I2-303	*Podospora anserina* adenylate cyclase	Q01513	34	1.3
Photorespiration				
I2-27	*Mesembryanthemum crystallium* glycolate oxidase	U80071	73	7X10^{-32}
Metal-binding proteins				
I2-306	*Arabidopsis thaliana* probable selenium-binding protein	E71401	52	2X10^{-13}
Structural proteins				
I2-308	*Lytechinus pictus* actin, cytoskeletal 3 (LPC3)	Q25379	46	8.4

Miscellaneous				
I2-176	*Mesostigma viride* ChlN subunit of protochlorophyllide reductase	AF166114	51	3.1
I2-216	*Mesostigma viride* ChlN subunit of protochlorophyllide reductase	AF166114	51	4.1
I2-14	*Lymnaea stagnalis* prepro-APGWamide	1811269A	33	0.0061
I2-113	*Caenorhabditis elegans* C27H6.1	Z81042	36	3.6
I2-230	*Drosophila melanogaster* ovo gene product (alt1)	AE003433	25	0.92
I2-263	*Homo sapiens* KIAA0621	AB014521	33	1.7
I2-313	*Drosophila melanogaster* CG10505 gene product	AE003453	33	3.1
I2-223	*Saccharomyces pombe* conserved hypothetical protein SPBC2A9.11c	T40102	29	1.2
I2-208	*Arabidopsis thaliana* unknown protein	AC006300	40	$1X10^{-8}$
I2-227	*Arabidopsis thaliana* hypothetical protein F10M6.80	T05400	37	$4X10^{-6}$
I2-279	*Arabidopsis thaliana* hypothetical protein F22I13.200	T05671	57	0.024
No homology				
12 clones				

Table 6.2 Putative identifications of ESTs preferentially expressed in sugarcane maturing internodal tissue

Sugarcane ESTs were assigned a putative identity based on partial sequence homology searches with known gene sequences in the NCBI GenBank database. Sequence homology/match is the database sequence that the sugarcane cDNA was most similar to. Accession number represents the number assigned by GenBank for individual entries. The % sequence identity is at the amino acid level and is calculated from sequence alignment between the sugarcane partial cDNA sequence and the GenBank entry by the BLASTX algorithm. The *E* value is the statistical indicator of the significance of the match between query and database sequence.

Clone	Sequence homology/match	Accession number	% sequence identity	*E* value
Cell wall metabolism				
I7-143	*Arabidopsis thaliana* callose synthase catalytic subunit-like protein	AL353013	82	$2X10^{-21}$
I7-159	*Arabidopsis thaliana* callose synthase catalytic subunit-like protein	AL353013	73	$2X10^{-19}$
I7-233	*Arabidopsis thaliana* callose synthase catalytic subunit-like protein	AL353013	82	$1X10^{-13}$
I7-204	*Arabidopsis thaliana* highly similar to putative callose synthase catalytic subunit	AC007153	46	$4X10^{-16}$
I7-155	*Arabidopsis thaliana* putative alpha-L-arabinofuranosidase	AC011708	68	$7X10^{-34}$
I7-175	*Arabidopsis thaliana* putative alpha-L-arabinofuranosidase	AC011708	67	$5X10^{-28}$
I7-81	*Cellulomonas fimi* cellulase	B47093	50	2.7
Carbohydrate metabolism				
I7-19	*Arabidopsis thaliana* glycerol-3-phosphate permease homolog	Z97343	65	$6.4X10^{-23}$

I7-83	*Arabidopsis thaliana* glycerol-3-phosphate permease homolog	Z97343	48	$2.9X10^{-34}$
I7-154	*Arabidopsis thaliana* trehalose-6-phosphate phosphatase	AF007778	61	$3X10^{-28}$
I7-186	*Arabidopsis thaliana* trehalose-6-phosphate phosphatase	AF007778	60	$2X10^{-27}$
I7-205	*Zea mays* glyceraldehyde-3-phosphate dehydrogenase	PQ0178	94	0.021
I7-56	*Sparus aurata* fructose-bisphosphate aldolase B	P53447	56	2.4
Stress responses				
I7-43	*Artocarpus integrifolia* jacalin	L03797	41	$5.6X10^{-15}$
I7-132	*Arabidopsis thaliana* similar to jacalin	AC008017	46	$2X10^{-20}$
I7-262	*Triticum aestivum* hypothetical protein wali7	T06984	94	$8X10^{-44}$
I7-266	*Triticum aestivum* hypothetical protein wali7	T06984	94	$3X10^{-44}$
I7-20	*Oryza sativa* osr40g3	Y08988	77	$1.8X10^{-49}$
I7-129	*Oryza sativa* GOS9 protein	P27349	52	0.004
Regulation				
I7-136	*Arabidopsis thaliana* translocon-associated protein, alpha subunit precursor	P45434	60	$2X10^{-34}$
I7-254	*Arabidopsis thaliana* putative signal sequence receptor, alpha subunit	AC006264	59	$3X10^{-36}$
I7-97	*Homo sapiens* ubiquitin activating enzyme E1-like protein	AF094516	44	$2X10^{-5}$
I7-232	*Homo sapiens* ubiquitin activating enzyme E1-like protein	AF094516	45	$2X10^{-9}$
I7-198	*Arabidopsis thaliana* putative ubiquitin-conjugating enzyme E2	AC005825	89	$4X10^{-23}$
I7-257	*Zea mays* ubiquitin conjugating enzyme	AF034946	100	$1X10^{-29}$
I7-33	*Mus musculus* tyrosine kinase	D83002	58	0.47
I7-49	*Saccharomyces cerevisiae* biotin-protein ligase	P48445	30	2.8
I7-75	*Homo sapiens* SPH-binding factor	AF071771	47	3.4
I7-88	Human papillomavirus type 40 replication protein E1	P36727	31	1.3
I7-101	*Streptomyces fradiae* tylactone synthase module 7	U78289	41	0.97
I7-103	*Canavalia ensiformis* legumain precursor	P49046	72	$1.8X10^{-14}$
I7-170	*Schizosaccharomyces pombe* probable protein involved in autophagy yeast apg7 homolog	T40646	43	$6X10^{-7}$
I7-172	*Zea mays* cysteine protease	X99936	84	$3X10^{-26}$
I7-174	*Arabidopsis thaliana* putative tetracycline transporter protein	AC005167	52	$1X10^{-12}$
I7-200	*Oryza sativa* ran	AB015287	100	$5X10^{-54}$
I7-217	*Plasmodium falciparum* erythrocyte membrane binding protein 1	T18378	31	2.6
Protein synthesis				
I7-92	*Arabidopsis thaliana* 40S ribosomal protein S15A	P42798	100	$1.1X10^{-45}$
I7-214	*Arabidopsis thaliana* putative ribosomal protein s19 or s24	AC009465	87	$3X10^{-25}$
Structural proteins				
I7-202	*Arabidopsis thaliana* ankyrin-like protein	AC006533	68	$9X10^{-37}$
Miscellaneous				
I7-244	*Arabidopsis thaliana* hypothetical protein T5L19.190	T04010	50	$1X10^{-24}$
I7-268	*Arabidopsis thaliana* hypothetical protein T5L19.190	T04010	52	$1X10^{-11}$

I7-130	*Arabidopsis thaliana* hypothetical protein T21L8.170	T12997	77	3X10^{-22}
I7-42	*Arabidopsis thaliana* hypothetical protein	AF000657	44	0.009
I7-194	*Arabidopsis thaliana* hypothetical protein	AF000657	38	0.015
I7-164	*Arabidopsis thaliana* hypothetical protein	AC011663	52	2X10^{-25}
I7-39	*Schizosaccharomyces pombe* hypothetical protein	Z99165	52	3.5X10^{-12}
I7-67	*Schizosaccharomyces pombe* hypothetical protein	AL021838	42	1.8X10^{-10}
I7-77	*Arabidopsis thaliana* putative protein	AL021633	58	2.2X10^{-15}
I7-76	*Magnaporthe grisea* putative pol polyprotein	M77661	36	0.27
I7-188	*Arabidopsis thaliana* unknown protein	AC006403	55	2X10^{-34}
I7-65	Tobacco ringspot virus RNA1 polyprotein	U50869	44	0.61
I7-137	*Drosophila melanogaster* CG11451 gene product	AE003592	40	7.9
I7-163	*Drosophila melanogaster* CG9318	AE003667	30	3.6
I7-234	*Drosophila pseudoobscura* decapentaplegic protein precursor	P91699	26	0.52
I7-134	*Homo sapiens* tetratricopeptide repeat protein 3	P53804	31	2.8
I7-247	*Arabidopsis thaliana* contains strong similarity to a hypothetical protein and contains three Kelch domains	AC012188	58	2X10^{-32}
I7-273	*Plasmodium knowlesi* duffy receptor	M68517	27	8.6
No homology 3 clones				

ESTs were analysed to distinguish those cDNA clones with the same database identity and therefore likely to represent the same gene. In the immature culm, nine different types of ESTs were present in multiple copies, with the number of redundancies for individual clones ranging between eight and two (Table 6.3). Similarly, eight individual maturing culm ESTs were present in multiple copies (Table 6.3). Assuming that the poly (A)$^{+}$ RNA populations are adequately reflected in the two subtracted cDNA libraries then the multiple copies of ESTs for individual clones are consistent with an increase in the abundance of the corresponding transcripts. The number of unique ESTs differed markedly between the two libraries. In the immature culm, 51% of the differentially expressed ESTs were single copy, compared with 81% in maturing culm. It must be noted, however, that as only a limited selection of subtracted clones was screened by macroarray hybridisation analysis, these percentages should not be considered as representative.

Table 6.3 Redundancy in immature and maturing culm subtracted libraries

The number of copies of individual preferentially expressed ESTs detected in the analysed sample set is represented.

Putative Identity	Library	Number of Clones
UDP-glucose dehydrogenase	Immature culm	8
Drought-induced protein Di19	Immature culm	7
Environmental stress-induced protein	Immature culm	6
Trehalose-6-phosphate synthase homolog	Immature culm	3
Glyceraldehyde 3-phosphate dehydrogenase, cytosolic	Immature culm	3
Cellulose synthase	Immature culm	3
Putative senescence-associated protein 12	Immature culm	2
Latex-abundant protein	Immature culm	2
Low temperature and salt responsive protein LT16A	Immature culm	2
Callose synthase catalytic subunit-like protein	Maturing culm	4
Glycerol-3-phosphate permease	Maturing culm	2
Jacalin	Maturing culm	2
Translocon-associated protein, alpha subunit precursor	Maturing culm	2
Putative alpha-L-arabinofuranosidase	Maturing culm	2
Trehalose-6-phosphate phosphatase	Maturing culm	2
Hypothetical protein wali7	Maturing culm	2
Hypothetical protein T5L19.190	Maturing culm	2

6.4.3 Putative functions of differentially expressed ESTs

All differentially expressed ESTs identified from the immature culm and maturing culm subtractive cDNA libraries were grouped into general functional categories according to putative identities obtained from the primary BLAST homologues (Tables 6.1 and 6.2). Immature culm ESTs could be divided into nine functional categories (Table 6.1) and maturing culm ESTs into eight (Table 6.2). Seven categories were common to both sets of ESTs. The category designated "Miscellaneous" comprised those ESTs whose putative functions were unclear from the BLAST search results. Clones with no database match were grouped separately. The results shown in Tables 6.1 and 6.2 are summarised in Fig. 6.1 to facilitate the examination of ESTs preferentially expressed in the two distinct culm developmental stages.

The majority of ESTs from both subtracted libraries for which a putative identity could be ascribed were homologous to genes associated with cell wall metabolism, carbohydrate metabolism, stress responses and regulatory processes (Fig. 6.1). The putative identities revealed that the specific differentially expressed ESTs detected in the immature and maturing culm were distinct from each other in all cases except one, suggesting that enrichment for culm developmental-stage specific gene sequences had occurred. EST homologues for glyceraldehyde-3-phosphate dehydrogenase were observed three times in immature culm and once in maturing culm. Two maturing culm clones were homologous to genes associated with protein synthesis with no similar differentially expressed gene sequences being identified in the immature culm. Likewise, single ESTs for glycolate oxidase (photorespiration) and a probable selenium-binding protein (metal-binding proteins) were detected in immature culm only.

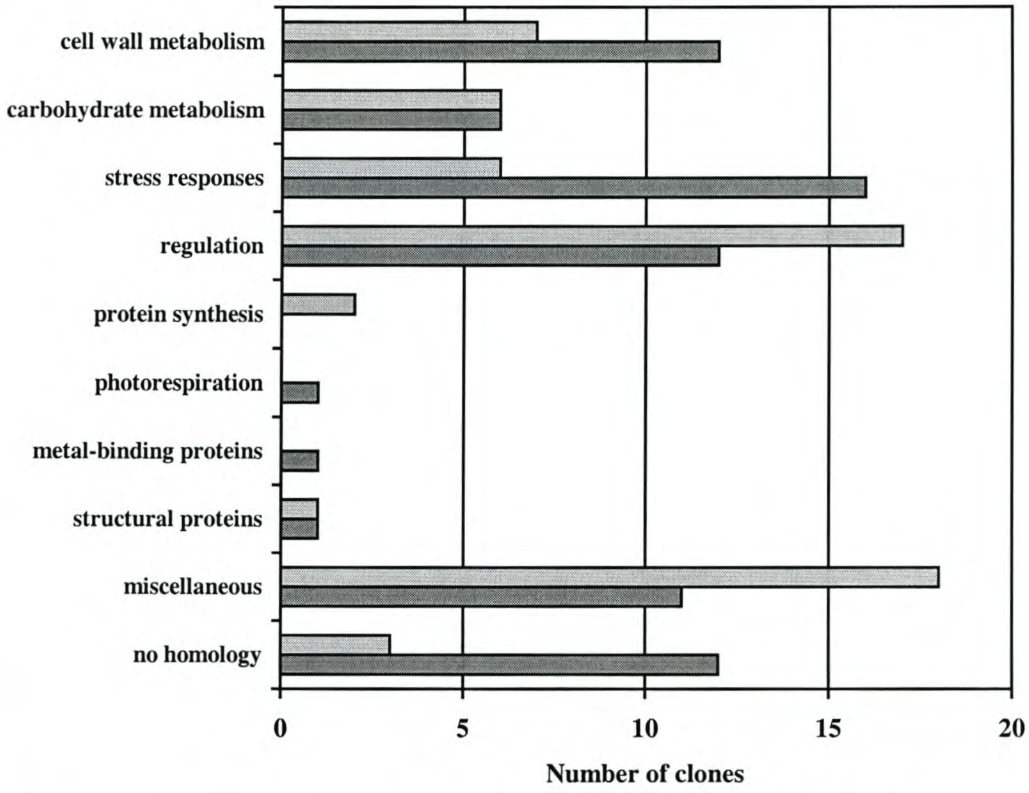


Fig. 6.1 Classification of immature (■) and maturing (□) culm-preferential ESTs according to putative function

All ESTs were included, regardless of *E* value or number of individual copies.

6.4.4 Confirmation of culm-specific and culm-preferential expression by Northern analysis

Northern blot analysis was performed on a selection of differentially expressed ESTs to verify the results obtained from macroarray hybridisation. Sixteen cDNA clones, eight immature and eight maturing culm-preferential were selected and used as probes on Northern blots prepared with total RNA from the leaf, leaf roll, immature culm and maturing culm (Fig. 6.2). In this way, the expression patterns of individual clones could also be examined in non-subtracted tissues (leaf and leaf roll) to confirm the culm-specific and culm-preferential patterns of the subtracted clones. Nine of the clones were selected from the group of redundant ESTs (Table 6.3) while the remainder was randomly selected but included clones whose putative identity was unclear.

For all the clones selected from the immature culm subtracted library, Northern analysis indicated that expression levels of the corresponding transcripts in the immature culm were very high, compared to the maturing culm where transcript levels were either very low or undetectable (Fig. 6.2). This indicates that subtraction between these two tissues had been successful. Six of the clones (I2-299, I2-249, I2-267, I2-160, I2-59, I2-223) were found to exhibit similar expression levels in the leaf roll to that observed in the immature culm, but with very low expression in the leaf. Both the leaf roll (meristematic apex) and immature culm are mitotically highly active therefore it is expected that some similarities in gene expression would occur between these two tissues. For two clones (I2-192, I2-225), expression of the homologous mRNAs were almost exclusive to the immature culm. For the ESTs selected from the maturing culm library, four of the eight clones tested exhibited near exclusive expression in the maturing culm (I7-132, I7-129, I7-163, I7-143). Two clones (I7-163, I7-143) also hybridised to transcripts in the leaf, but at an extremely low level. However, the high levels of mRNA detected in the maturing culm for these four ESTs indicated that subtractive hybridisation had been effective. The expression patterns for the remaining maturing culm clones (I7-155, I7-186, I7-101, I7-214) did not compare favourably with the results from the macroarray screening. Northern analysis indicated that homologous mRNA levels for these clones were highest in the leaf roll and immature culm and for three of the clones, no transcripts could be detected in the maturing culm. This suggests that subtractive hybridisation may not have gone to completion for the maturing culm. Furthermore, the expression results from the cDNA macroarray screening may have included some false positives. Additional hybridisations using four clones from the maturing culm library (I7-19, I7-83, I7-273, I7-244) failed to produce a detectable signal in any of the four RNA samples used for Northern analysis, indicating that the homologous transcripts are expressed at very low levels (results not shown). Results from Northern analysis thereby confirmed that the subtracted libraries contained high, moderate and low abundance tissue-preferential cDNA molecules.

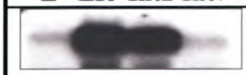
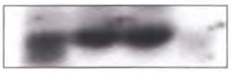
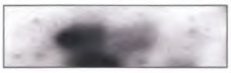
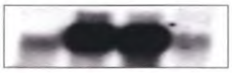
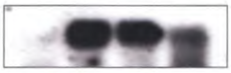
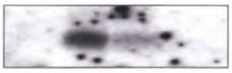
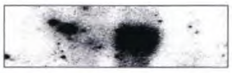
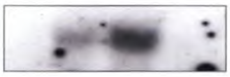
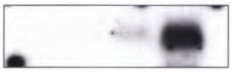
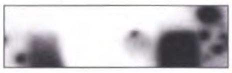
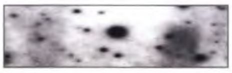
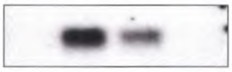
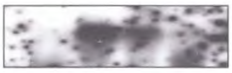
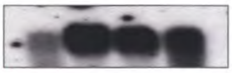
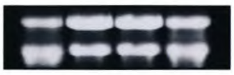
L LR Int2 Int7	Clone	Putative identity	Source
	I2-299	Calmodulin-related protein 2, touch-induced	Int2
	I2-249	Trehalose-6-phosphate synthase homolog	Int2
	I2-267	Ornithine aminotransferase	Int2
	I2-160	UDP-glucose 6-dehydrogenase	Int2
	I2-59	Environmental stress-induced protein	Int2
	I2-223	Conserved hypothetical protein SPBC2A9.11c	Int2
	I2-192	Putative senescence-associated protein	Int2
	I2-225	Latex-abundant protein	Int2
	I7-132	Jacalin	Int7
	I7-129	GOS9 protein	Int7
	I7-163	CG9318	Int7
	I7-143	Callose synthase catalytic subunit-like protein	Int7
	I7-155	Putative alpha-L-arabinofuranosidase	Int7
	I7-186	Trehalose-6-phosphate phosphatase	Int7
	I7-101	Tylactone synthase module 7	Int7
	I7-214	Putative ribosomal protein s19 or s24	Int7
			

Fig. 6.2 Northern blot analysis of 16 selected immature and maturing culm-preferential ESTs

Expression patterns were examined for individual ESTs in sugarcane leaf (L), leaf roll (LR), immature culm (Int2) and maturing culm (Int7). The bottom panel indicates ethidium-bromide stained rRNA to demonstrate equal sample loading.

6.5 DISCUSSION

In this study, cDNA subtractive hybridisation was used to specifically enrich for sugarcane culm differentially expressed cDNA sequences. Screening randomly selected clones from immature and maturing culm subtracted libraries using a serial hybridisation analysis of cDNA macroarrays successfully detected transcripts preferentially expressed in the culm. In this way, 132 cDNAs were identified as candidates for genes that are differentially regulated in the immature and maturing culm. This represents approximately one-third of the cDNA clones tested from the subtracted libraries. Elongation of sugarcane internodal tissue is accompanied by an increasing accumulation of sucrose concentrations (Moore 1995) and consequently, differences in metabolic status between immature and maturing internodes is expected to be characterised by a diversity in gene expression in each of these developmental states. Although the results from this study are based on a subset of clones only and therefore cannot be considered as fully representative, they indicate that there are many transcriptionally regulated genes contributing to the differences between immature and maturing culm.

DNA sequence database searches revealed that many of the differentially expressed sugarcane cDNAs were homologous to the reported sequences of various genes (Tables 6.1 and 6.2). Sequence comparison between sugarcane cDNAs and genes of known function from other organisms is useful in formulating predictions about the functions of the sugarcane genes. The putative identities of the differentially expressed cDNAs allow examination of the abundance of different types of transcripts in the immature and maturing culm. Results indicated that 22% of immature culm and 10% of maturing culm ESTs were stress-related. The homologues for these ESTs included previously characterised genes from other plants such as an environmental stress-induced protein (Luo et al., 1992), os40g3 (Moons et al., 1997), drought-induced protein Di19 (Gosti et al., 1995) and jacalin (Zhang et al., 2000). These authors reported that the expression of these genes were induced by a variety of stimuli such as abscisic acid (environmental stress-induced protein, os40g3), drought (drought-induced protein Di19) and salt stress (os40g3, jacalin). In addition, for the environmental stress-induced protein, os40g3 and the drought-induced protein Di19, expression of these genes was reported to be specific to stem tissues (Luo et al., 1992; Moons et al., 1997; Gosti et al., 1995). Expression of stress-response genes in association with plant development have also been documented for the inner bark of the conifer, *Cryptomeria japonica* (Ujino-Ihara et al., 2000) and during grape berry ripening (Ablett et al., 2000; Davies and Robinson 2000). In the latter case, Davies and Robinson suggested that the considerable changes in osmotic pressure and water potential that occur during ripening may result in the synthesis of proteins involved in stress management. The role of the stress-responsive cDNAs expressed during sugarcane culm development as identified in this study is unknown.

It is not surprising that almost 15% of ESTs preferentially expressed in the immature and maturing culm are homologous to genes associated with cell wall metabolism. The rapid rate of cell expansion in young, developing tissues result in a high demand for cell wall polysaccharide precursors such as cellulose and hemicellulose. UDP-glucose dehydrogenase is a key regulator for the availability of hemicellulose precursors and has been shown to be highly expressed in root, epicotyl and expanding leaves of soybean (Tenhaken and Thulke 1996). Multiple copies of sugarcane cDNAs homologous to UDP-glucose dehydrogenase were detected exclusively in the immature culm during macroarray screening of cDNAs selected from the subtracted libraries (Table 6.3). The high level of expression of the corresponding transcript for these clones in immature internodes and meristematic apex (Fig. 6.2) suggests that this gene plays an important role during sugarcane culm expansion. As the internodes mature and cell expansion slows, cell walls accumulate increasing amounts of complex polysaccharides which is reflected by the preferential expression of ESTs similar to genes for callose production and arabinoxylan metabolism in the maturing culm (Table 6.2). The expression of sugarcane ESTs homologous to genes associated with a variety of regulatory and signal transductory processes in a culm developmental stage-preferential manner implies that these genes may have important roles in mediating culm growth and development. The regulation of expression of a touch-induced calmodulin-related protein in *Arabidopsis thaliana* by a variety of physical stimuli suggested that this gene played a role in enabling the plant to sense and respond to environmental changes (Braam and Davis 1990). The function of a putative senescence-associated protein, expressed in senescing daylily petals, was suggested to be associated with degradative metabolism during senescence and in producing other signal molecules (Panavas et al., 1999). Similarly, the degradation of specific proteins via the ubiquitin pathway is a highly regulated process involving the co-ordinated activities of several enzymes (Belknap and Garbarino 1996). The preferential expression of cDNA homologues for genes actively involved in the ubiquitin pathway in immature and maturing sugarcane culm, as determined in this investigation, demonstrate that these genes have a key function during culm development.

Growth and development of the sugarcane culm is a continuous process which results in a gradient of maturation and sucrose accumulation down the culm (Moore 1995). As the rate of accumulation increases sharply between internodes four and seven, reaching a peak in internode 7 (Whittaker and Botha 1997), enrichment for maturing culm-preferential transcripts by subtractive hybridisation was expected to result in a significant proportion of the expressed sequences detected in this region to be directly associated with sucrose metabolism. It was surprising to observe, therefore, that only 8% and 10% of the analysed clones preferentially expressed in the immature and maturing culm, respectively, were homologous to genes associated with carbohydrate metabolism, none of which were similar to genes known to be directly associated with sucrose metabolism (Tables 6.1 and 6.2). A similar observation was noted in grape berries, known to be active in storing sugars and modified metabolites, where only a small proportion of transcripts were detected for sugar transporters and secondary metabolites (Ablett et al., 2000). Likewise, in ripening wild strawberry fruits, it was reported that none of the ripening-induced cDNAs were homologous to genes directly related to ripening associated processes such as cell wall metabolism and the accumulation of sugars and pigments (Nam et al., 1999). Instead, a wide range of processes was indicated to be upregulated during strawberry fruit ripening (Nam et al., 1999). This too appears to be the case during sugarcane culm maturation. It was intriguing to detect the expression in the culm of ESTs homologous to two enzymes involved in trehalose biosynthesis, trehalose-6-phosphate synthase (TPS) and trehalose-6-phosphate phosphatase (TPP). Not much is understood about the role of trehalose metabolism in the physiology and development of higher plants as evidence for the occurrence of this pathway in plants was established only recently (for review, see Goddijn and van Dun 1999). During the course of this study several differentially expressed cDNAs were isolated whose corresponding transcripts accumulated preferentially in the immature and maturing culm but which could not be assigned a putative identity (Tables 6.1 and 6.2, Fig. 6.1). This was due either to sequence similarity to genes whose function was unclear or due to no match with publicly available sequence data. This forms a resource of potentially novel genes that are developmentally regulated in sugarcane.

In summary, the research results described here demonstrate that combining cDNA subtractive hybridisation with cDNA macroarray expression analysis presents a practical way of isolating sugarcane culm developmental stage-preferential expressed sequences. By using arrayed clones from subtracted libraries, differentially expressed transcripts could be isolated quickly and simply without the need for repeated screening cycles. While it was apparent that results from macroarray screening did not always coincide with a Northern analysis, the potential for inclusion of false positives is a minor disadvantage compared with the efficiency offered by the macroarray system. It can be concluded therefore, that the approach used here is suitable as a preliminary screen to identify candidate genes for further analyses. Work is currently underway to confirm the tissue-specificity of selected cDNAs by examining transcript abundance in more internodes of varying maturity as well as roots. Although this study focussed on a limited selection of clones from the subtracted libraries, insight could be obtained into the transcript complexity of the developing sugarcane culm. The diversity of candidate developmental-stage preferential genes identified form a valuable resource of tools that will facilitate further molecular analyses of culm development. Moreover, these genes may be characterised and evaluated for application in genetic engineering and marker programmes.

6.6 REFERENCES

Ablett E, Seaton G, Scott K, Shelton D, Graham MW, Baverstock P, Lee LS and Henry R (2000) Analysis of grape ESTs: global gene expression patterns in leaf and berry. Plant Sci 159: 87-95.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Belknap WR and Garbarino JE (1996) The role of ubiquitin in plant senescence and stress responses. Trends in Plant Sci 1(10): 331-335.

Braam J and Davis RW (1990) Rain-, wind-, and touch-induced expression of calmodulin and calmodulin-related genes in *Arabidopsis.* Cell 60: 357-364.

Carson DL and Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. Crop Sci 40(6): 1769-1779.

Cooke R, Raynal M, Laudié M, Grellet F, Delseny M, Morris P-C, Guerrier D, Giraudat J, Quigley F, Clabault G, Li Y-F, Mache R, Krivitzky M, Gy IJ-J, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clément B, Philipps G, Hervé C, Bardet C, Tremousaygue D, Lescure B, Lacomme C, Roby D, Jourjon M-F, Chabrier P, Charpenteau J-L, Desprez T, Amselem J, Chiapello H and Höfte H (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. Plant J 9(1): 101-124.

Davies C and Robinson SP (2000) Differential screening indicates a dramatic change in mRNA profiles during grape berry ripening. Cloning and characterisation of cDNAs encoding putative cell wall and stress response proteins. Plant Physiol 122: 803-812.

Decraene C, Reguigne-Arnould I, Auffray C, Piétu G (1999) Reverse transcription in the presence of dideoxynucleotides to increase the sensitivity of expression monitoring with cDNA arrays. BioTechniques 27(5): 962-966.

Goddijn OJM and van Dun K (1999) Trehalose metabolism in plants. Trends in Plant Sci 4(8): 315-319.

Gosti F, Bertauche N, Vartanian N and Giraudat J (1995) Abscisic acid-dependent and –independent regulation of gene expression by progressive drought in *Arabidopsis thaliana.* Mol Gen Genet 246: 10-18.

Holmes DS and Quigly M (1981) A rapid boiling method for the preparation of bacterial plasmids. Anal Biochem 114: 193-197.

Kamakura T, Xiao J-Z, Choi W-B, Kochi T, Yamaguchi S, Teraoka T and Yamaguchi I (1999) cDNA subtractive cloning of genes expressed during early stage of appressorium formation by *Magnaporthe grisea.* Biosci Biotechnol Biochem 63(8): 1407-1413.

Kozian DH and Kirschbaum BJ (1999) Comparative gene-expression analysis. Tibtech 17: 73-77.

Liu QY, van der Meer JP and Reith ME (1994) Isolation and characterisation of phase-specific complementary DNAs from sporophytes and gametophytes of *Porphyra purpurea* (Rhodophyta) using subtracted complementary DNA libraries. J Phycol 30: 513-520.

Luo M, Liu J-H, Mohapatra S, Hill RD and Mohapatra SS (1992) Characterisation of a gene family encoding abscisic acid- and environmental stress-inducible proteins of alfalfa. J Biol Chem 267(22): 15367-15374.

Moons A, Gielen J, Vandekerckhove J, Van Der Straeten D, Gheysen G and Van Montagu M (1997) An abscisic-acid- and salt-stress-responsive rice cDNA from a novel plant gene family. Planta 202: 443-454.

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22: 661-679.

Nam Y-W, Tichit L, Leperlier M, Cuerq B, Marty I and Lelièvre J-M (1999) Isolation and characterisation of mRNAs differentially expressed during ripening of wild strawberry (*Fragaria vesca* L.) fruits. Plant Mol Biol 39: 629-636.

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E and Somerville C (1994) Genes Galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241-1255.

Panavas T, Pikula A, Reid PD, Rubinstein B and Walker EL (1999) Identification of senescence-associated genes from daylily petals. Plant Mol Biol 40: 237-248.

Patel M and Sive H (1996) PCR-based subtractive cDNA cloning. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K (eds) Current Protocols in Molecular Biology, Vol. 1. Greene and Wiley-InterScience, New York, pp 5.9.1-5.9.20.

Sambrook J Fritsch EF and Maniatis T (1990) Molecular Cloning: A laboratory manual. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press.

Takemura M, Fujishige K, Hyodo H, Ohashi Y, Kami C, Nishii A, Ohyama K and Kohchi T (1999) Systematic isolation of genes expressed at low levels in inflorescence apices of *Arabidopsis thaliana.* DNA Res 6: 275-282.

Tenhaken R and Thulke O (1996) Cloning of an enzyme that synthesises a key nucleotide-sugar precursor of hemicellulose biosynthesis from soybean: UDP-glucose dehydrogenase. Plant Physiol.112: 1127-1134.

Ujino-Ihara T, Yoshimura K, Ugawa Y, Yoshimaru H, Nagasaka K and Tsumura Y (2000) Expression analysis of ESTs derived from the inner bark of *Cryptomeria japonica*. Plant Mol Biol 43: 451-457.

van Dillewijn C (1952) Growth: general, grand period, growth formulae. In: van Dillewijn C (ed) Botany of Sugarcane, Vol. 1. Veenen and Zonen, The Netherlands, pp 97-162.

Whittaker A and Botha FC (1997) Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. Plant Physiol 115: 1651-1659.

Wyrich R, Dressen U, Brockmann S, Streubel M, Chang C, Qiang D, Paterson AH and Westhoff P (1998) The molecular basis of C_4 photosynthesis in sorghum: isolation, characterisation and RFLP mapping of mesophyll- and bundle-sheath-specific cDNAs obtained by differential screening. Plant Mol Biol 37 (1998) 319-335.

Yamamoto K and Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35: 135-144.

Zhang W, Peumans WJ, Barre A, Astoul CH, Rovira P, Rougé P, Proost P, Truffa-Bachi P, Jalali AAH and Van Damme EJM (2000) Isolation and characterisation of a jacalin-related mannose-binding lectin from salt-stressed rice (*Oryza sativa*) plants. Planta 210: 970-978.

CHAPTER 7

CONCLUDING REMARKS

This investigation made use of recent developments in genetic analysis to examine processes associated with culm maturation and sucrose accumulation in sugarcane at the molecular level. The results clearly demonstrate the strength of functional genomics for analysing the intricacies of sugarcane metabolism.

The major contributions from this study fall into three main categories. Firstly, it has been demonstrated that EST analysis is a valuable approach towards providing new information about gene expression during sugarcane growth and development. Secondly, the identification and analysis of genes expressed in different tissues and developmental stages has significantly advanced knowledge of gene expression in sugarcane and provided new insights into the complexities of sugarcane metabolism. Thirdly, a collection of genetic resources for application in sugarcane improvement programmes has been developed.

A database of sugarcane gene sequences has been established through the generation of ESTs. This database, containing approximately 1400 sequences from leaf and culm tissues of varying maturity, has facilitated access to a myriad of genes not previously available for sugarcane research. While many of these genes could be putatively identified on the basis of sequence homology with other known genes, approximately 30% of the sugarcane genes detected in this study remain unidentified. As was demonstrated in Chapters 5 and 6, several of these genes exhibit culm-preferential expression patterns and could, therefore, have important roles during culm maturation. Further characterisation is required to establish possible functions for these unidentified genes.

The submission of a large portion of the sugarcane ESTs from the meristematic apex and maturing culm to the dbEST database represented the first sugarcane ESTs to be made available in the public domain. In addition to its value as a molecular tool for sugarcane gene analysis, the EST database generated in this study has also proven to be a useful resource of cDNA clones. This is reflected by the regular requests for clonal material from other research laboratories worldwide, both those working on sugarcane and those focused on other plants.

The sugarcane culm is not simply an organ for sucrose storage but is a complex and dynamic structure. The broad diversity of genes detected through EST analysis (Chapter 4) and the identification of differentially expressed genes (Chapters 5 and 6) indicate that during growth and maturation, the culm is actively engaged in a wide variety of metabolic processes as well as being responsive to environmental conditions and stresses. While many similarities in the expression of genes between leaf tissues and culm tissues were detected, there was evidence for the up-regulation of genes in the culm encoding proteins associated with carbohydrate metabolism, cell wall synthesis, various stress responses and the regulation of cellular metabolism. The involvement of these genes in sugarcane culm metabolism should be explored further.

Genes encoding enzymes directly associated with the pathway of sucrose accumulation do not appear to be abundantly expressed in the maturing sugarcane culm. With the exception of SuSy, no other genes encoding the key enzymes of sucrose metabolism were detected in this study, despite enriching for transcripts preferentially expressed during culm maturation. These findings raise new questions about how sucrose accumulation is regulated at the genetic level in sugarcane. It is possible that genes encoding key enzymes in the sucrose accumulation process are expressed at low levels in the culm. Future research targeted towards characterising the expression behaviour of genes known to be associated with the sucrose metabolic pathway could provide valuable information about the regulation of transcript abundance during culm maturation.

Molecular manipulation of sugarcane is currently limited by the poor availability of genetic resources such as promoters required to drive the tissue or organ-specific expression of introduced transgenes. The tissue-preferential genes identified in this investigation (Chapter 6) provide a useful resource of candidate genes for future studies towards promoter isolation and characterisation.

This research has provided a platform from which future functional genomics strategies for sugarcane can be formulated. Deciphering the roles of the myriad of genes specifically associated with culm maturation will be a major challenge but may identify important targets influencing the process of sucrose accumulation. These efforts, as well as those towards establishing the mechanisms by which the transcript level of genes directly associated with sucrose metabolism are regulated, will add vital additional knowledge to what is currently known about the unique physiology of sugarcane.

PRESENTATIONS AND PUBLICATIONS ARISING FROM THE STUDY

Carson DL, Groenewald JH and Botha FC. Development of an expressed sequence tag (EST) database for sugarcane. South African Genetics Society XV Congress, Stellenbosch, July 1996.

Carson DL and Botha FC. Development of an expressed sequence tag (EST) database for sugarcane. ISSCT Pathology and Molecular Biology Combined Workshop, Umhlanga Rocks, South Africa, May 1997.

Carson DL and Botha FC. The development of an expressed sequence tag (EST) database for sugarcane. 5th International Congress of Plant Molecular Biology, Singapore, September 1997.

Carson DL, Huckett BI and Botha FC (1998) The identification of sugarcane genes by random sequencing of cDNA Clones. Proc SA Sugar Technol Assoc 72: 143-145.

Carson DL, Huckett BI and Botha FC. The identification of sugarcane genes by cDNA sequencing. South African Genetics Society XVI Congress, Bloemfontein, June/July 1998.

Carson DL, Williams NJ, Huckett BI and Botha FC. The identification of expressed sequence tags in sugarcane. Plant and Animal Genome VII, San Diego, USA, January 1999.

Carson DL. Gene identification and expression analysis in sugarcane. Invited talk; Department of Biology, School of Life and Environmental Sciences, University of Natal, Durban, November 1999.

Carson DL and Botha FC (2000) Preliminary analysis of Expressed Sequence Tags for sugarcane. Crop Sci 40: 1769-1779.

Huckett BI, Carson DL, Reddy S and Botha FC. Patterns of gene expression in sugarcane leaf and culm. Sugarcane Genomics Workshop, University of Queensland, Brisbane, Australia, May 2000.

Carson DL, Huckett BI and Botha FC (2000) Patterns of gene expression in sugarcane monitored using cDNA macroarrays. Proc SA Sugar Technol Assoc 74: 181-183.

Carson DL, Huckett BI and Botha FC. The identification of genes preferentially expressed in the sugarcane culm is facilitated through the use of cDNA subtraction. International Consortium of Sugarcane Biotechnology Workshop, Plant and Animal Genome IX, San Diego, USA, January 2001.

Carson DL, Huckett BI and Botha FC. The usefulness of cDNA macroarrays for the identification of differentially expressed genes in sugarcane. Plant and Animal Genome IX, San Diego, USA, January 2001.

Carson DL, Huckett BI and Botha FC (2001) Genomics Research at SASEX: Perspectives from a small-scale programme. Proc Int Soc Sugar Cane Technol 24: 539-541.

Carson DL and Botha FC (2002) Genes expressed in sugarcane maturing internodal tissue. Plant Cell Rep 20(11): 1075-1081.

Carson DL, Huckett BI and Botha FC (2002) Differential gene expression in sugarcane leaf and internodal tissues of varying maturity. Submitted: SA J Bot.

Carson DL, Huckett BI and Botha FC (2002) Sugarcane ESTs differentially expressed in immature and maturing internodal tissue. Plant Sci 162(2): 289-300.

CURRICULUM VITAE

1990: Obtained BSc degree from University of Natal, Durban, South Africa.
Major subjects: Cell Biology; Environmental Biology

1991: Obtained BSc (Hons) degree from University of Natal, Durban, South Africa.
Subject: Cell Biology

1993: Obtained MSc Degree (*cum laude*) from University of Natal, Durban, South Africa.
Project: Effects of nitrogen nutrition on salt stressed *Nicotiana tabacum* var. Samsum *in vitro.*
Supervisors: Prof MP Watt and Dr BI Huckett

1993: Employed in the Biotechnology Department at the South African Sugar Association Experiment Station in the position of Assistant Research Officer.

1998: Registered for PhD degree at the University of Stellenbosch, South Africa.
Supervisors: Prof FC Botha and Dr BI Huckett

1999: Promoted to position of Research Officer in the Biotechnology Department.