Part 1 - Summary Details

CRDC Project Number: **CSP137C**

Project Title: Development of an unigene set of cotton clones for general microarray analysis of gene expression in cotton plants

Project Commencement Date: 1/10/01 **Project Completion Date:** 30/9/04
Research Program: Plant Breeding and Biotechnology

Part 2 – Contact Details

Administrator: Graham Brill
Organisation: CSIRO Plant Industry
Postal Address: GPO Box 1600 Canberra ACT 2601
Ph: 02-62465431 **Fax:** 02-6246-5000 **E-mail:** Graham.Brill@csiro.au

Principal Researcher: Yingru Wu
Organisation: CSIRO Plant Industry
Postal Address: GPO Box 1600 Canberra ACT 2601
Ph: 02-62464914 **Fax:** 02-6246-5000 **E-mail:** Yingru.Wu@csiro.au

Supervisor: Danny Llewellyn
Organisation: CSIRO Plant Industry
Postal Address: GPO Box 1600 Canberra ACT 2601
Ph: 02-62465470 **Fax:** 02-6246-5000 **E-mail:** Danny.Llewellyn@csiro.au

Researcher 2 Todd Collins (Technical Assistant)
Organisation: CSIRO Plant Industry
Postal Address: P.O. Box 1600 Canberra ACT 2601
Ph: 02-62464922 **Fax:** 02-62465000 **E-mail:** Todd.Collins@csiro.au

Part 3.3 – Final Reports (due 3 months after completion of project)

Background

Cotton provides more than half of the fibre in textile manufacturing globally and is an important contributor to the Australian economy. In spite of the world significance of this plant, we know little about the genes that control the initiation and growth of the cells on the outer surface of the seed that are destined to become the fibre. Recent advances in sequencing of two plant genomes (Arabidopsis and rice) and the advent of DNA microarray technology are profoundly changing and accelerating research in many areas of plant biology. cDNA Microarrays consist of thousands of target cDNAs robotically arrayed on glass slides. Fluorescently labelled cDNA samples, from different tissues or different conditions (representing the range of genes expressed in those tissues) are then hybridised to the arrays. By analysing the fluorescence of the hybridised spots on the microarrays we can assess the gene expression changes of 1000's of genes simultaneously. Microarrays provide a powerful tool for discovery of plant genes involved in important biological processes such as growth, development and defence, to name but a few. At present comprehensive genomics tools are only available for a few plant species, especially Arabidopsis. In order to tap into this new resource and take advantage of the possibilities of this new technology we are developing a general cotton microarray. Our aim is to produce cotton microarrays with a large non-redundant set of up to 10,000 cotton genes chosen from a variety of different cotton tissues and treatments. The different tissues and treatments used for the isolation of clones for the arrays should allow researchers to examine a broad range of cotton research areas, such as fibre development, disease defence and growth. Therefore, the aim of this general cotton microarray or cotton chip is to provide a valuable research tool for cotton scientists. The process of gene discovery will also provide opportunities for generating new Intellectual Property for Australia. A better understanding of the molecular processes in fibre initiation and development should allow us to design new strategies to genetically improve fibre yield and quality in Australian varieties.

Objectives

1. *One of the objectives of this project was to sequence another 7,000 cotton ESTs from a range of cotton cDNA libraries developed from different cotton tissues, including cotton leaf, root, hypocotyl, young ovule and immature embryo etc. at CSIRO Plant Industry. We had already sequenced over 3000 ESTs in an earlier project. It is expected that some genes will be common between different cDNA libraries and so represented more than once in the whole set of genes characterised. Sequence analysis including clustering of the sequences is performed to assemble a non-redundant or unique set of the ESTs (an estimated 4000 to 5000 out of the original 10,000). The non-redundant EST set would be PCR amplified and verified for quality, before printing onto the array.*

The above objective was completed and we have, in fact, been able to sequence over 8000 cDNA clones from cDNA libraries developed from different cotton tissues, including cotton leaf, root, hypocotyl, young ovule and immature embryo. Non-redundant ESTs from 5000 of these clones have been amplified by PCR and ready for printing. The non-redundant ESTs for the remaining 3000 clones have been finalised and will be amplified and printed to the new array in Jan-Feb 2005.

In addition, we have acquired sequences of another 7000 cDNA clones (GH_DE) from our ovule cDNA libraries due to collaborations established with Prof. Jonathan Wendel's group at the University of Iowa, USA. Since these cDNA clones were sequenced from both ends (in contrast to the above ESTs that were sequenced from only one end), the analysis of these sequences requires an additional assembly step. In collaboration with Bayer Crop Science 's Bioinformatics group, the analysis of these set of sequences have just been accomplished and a non-redundant set of ESTs (5500) been finalised. This set of non-redundant ESTs will be amplified in the beginning of next year.

2. *The other objective was to acquire about 9000 fibre ESTs from Dr Ben Burr (Brookhaven Laboratories, USA) and perform the sequence clustering and amplification of the non-redundant clones (6000 to 7000 estimated) as for the other ESTs mentioned above. Together with the proposed in-house ESTs (4000-5000), it would allow us to print a general cotton microarray consisting of at least 10,000 non-redundant cotton genes.*

Unfortunately this set of ESTs never arrived in Australia in spite of an original oral agreement. So we decided to purchase another set of non-redundant fibre ESTs developed at UC Davis by Prof. Wilkins. These 13,000 fibre ESTs include representatives of most of the sequences from the Burr collection and have all been amplified and are ready for printing.

Over all, we have obtained about 24,000 non-redundant cotton cDNA clones of which about 17,000 clones have been PCR amplified, quality verified and readied for printing. In the beginning of this year, we will amplify the rest of the clones and print the arrays. This array, once finished, more than double the number of non-redundant cotton genes that we initially proposed and should provide a powerful tool for cotton researchers in Australia.

Methodology

Microarrays are formed by robotically depositing specific fragments of DNA at indexed locations onto microscope slides. The DNA fragments can originate from a variety of sources including anonymous cDNA clones, EST clones (ie sequenced cDNA clones), anonymous genomic clones, synthesized oligonucleotides, or DNA amplified from open reading frames (ORFs) found in sequenced genomes.

Once produced, the microarrays are hybridised with fluorescently-labelled mRNA-derived probes and the bound probes on the array are then excited by light. The fluorescent signal emitted from each spot is a reflection of the abundance of the corresponding sequence in the original probe and hence tissue from which the original RNA was extracted. Microarray technology is ideally suited for making pair-wise comparisons of samples. Two fluorescent tags – often just refereed to as red and green tags, with different excitation and emission optima, can be used to label two distinct probes (eg. two mRNA populations from physiologically or genetically distinct samples). The two probes are mixed and allowed to hybridise to the same microarray. For each spot on the microarray, the ratio of fluorescence emission at the two wavelengths (red and green channels) reflects the ratio of the abundance of that mRNA species in the two probes.

Microarrays are one of the most powerful tools that have recently been developed to bridge the gap between sequence information and functional genomics. The power lies in its scale, sensitivity and the quantitative nature of the data output. Theoretically, gene expression patterns of a whole plant genome (20,000-30,000 genes) can be studied simultaneously. Because the detection is fluorescence based, the signal output is very sensitive, and individual mRNA species can be detected at a threshold of 1 part in 100 000 to 1 part in 500 000. Because the output is quantitative, subtle changes in gene expression can be detected, in addition to the more dramatic changes observed with such techniques as subtractive hybridisation and differential display. The application of this powerful technology on cotton fibre genomics was only made possible by the installation at CSIRO of a robotic microarrayer and scanners. Operating funds were provided by CRDC for microarray consumables and some additional minor equipment needed to use the microarray techniques in cotton.

Development of anonymous cDNA microarrays (where the DNAs printed on the array have not been sequence characterised) is a more straightforward and less costly procedure than using ESTs, (this was used in the microarrays produced in CSP119C), but has many limitations: inheritant redundancy in cDNA libraries reduces efficiency of gene discovery; lack of sequence information on the genes showing interesting expression pattern unless they are subsequently sequencd. Uni-gene arrays overcome all these limitations because all the genes printed have been sequenced and redundancies have been reduced, so uni-gene arrays are more informative, efficient and also provide wider coverage for a given genome. The process of producing a uni-gene array requires intensive application of Bioinformatics to process sequences including: removing vector sequence and low quality sequence; clustering to assess redundancy; function annotation through database searches etc. The collaborations established with Bayer CropScience's Bioinformatics group in this area has contributed a great deal to this project's progress, although we are now developing the same Bioinfoirmatic expertise (hardware, software and staff) in-hoouse.

Results

The number of sequences generated from CSIRO's cotton cDNA libraries are listed in Table 1. cDNA clones were sequenced from one end (5' end). Sequence quality screening and vector clipping was used to reduce the total number of sequences by 630.

Table 1. EST sequences generated from CSIRO cotton cDNA libraries

Library Name	Source Tissue	Sequences Generated	Total Number of Sequences	After Cleaning	Total Number of Sequences
IE	Immature embryo	1536	8633	1415	8003
CHX	Cycloheximide treated ovule	1048		1037	
ON	Normalized 0 dpa ovule library	1204		1136	
OCF	0 dpa ovule	1101		1091	
CDO	Root and hypocotyl	480		441	
CRH	Root and hypocotyl	1152		1106	
LSL	Later season leaf	2112		1777	

Number of additional sequences generated from our ON and CHX libraries in collaboration with Prof. Wendel are shown in Table 2. All cDNA clones were sequenced from both ends (5’ and 3’ ends). After quality screening and vector clipping, a pair of sequences from one cDNA clone was merged using CAP3 program. This assembly step reduced the total number of sequences by half. Because the availability of the more extensive sequence information, the function annotations of these cDNA clones are more reliable and informative.

Table 2. EST sequences generated from CSIRO cotton ovule libraries in collaboration with Prof. Wendel at Uni. Iowa.

Sequence Name	Source Library	File type provided	Number of sequences	After cleaning And Cap3 assembly	Number of clones
GH_DEa	CHX	6741	11695	3937	6787 clones
GH_DEb	ON	4954		2850	

After the sequences were cleaned, a global assembly using the CAP3 program was performed to assess redundancy levels in each of the EST sets and to generate a uni-gene list. Table 3 shows the redundancy levels of the EST sets with the redundancy level varying from 7.7% to 28.5%.

Table 3. Redundancy analysis of the EST sets

EST Set	Number of Sequences	Number of Contigs	Number of Singletons	Number of Unigenes	Redundancy
GH_DEa	3937	485	2657	3142	20.17%
GH_DEb	2850	371	1993	2364	17.08%
IE	1415	104	922	1026	27.49%
CHX	1037	80	848	928	10.51%
CRH	1106	65	956	1021	7.68%
CDO	441	44	289	333	24.48%
LSL	1777	148	1122	1270	28.53%
OCF	1091	100	847	947	13.19%
ON	1136	58	989	1047	7.83%
TOTAL	14790	2364	7589	9953	32.70%

More detailed analysis on the redundancy distribution from CHX, OCF and ON EST sets is shown in Figure 1. While about 80% sequences of both CHX and OCF are singletons (ie only represented once), over 90% of ON sequences are singletons, suggesting the normalization procedure we applied reduced redundancy levels slightly. Corresponding to the increased number of singleton sequences, the number of ON sequences present in clusters was reduced slightly.

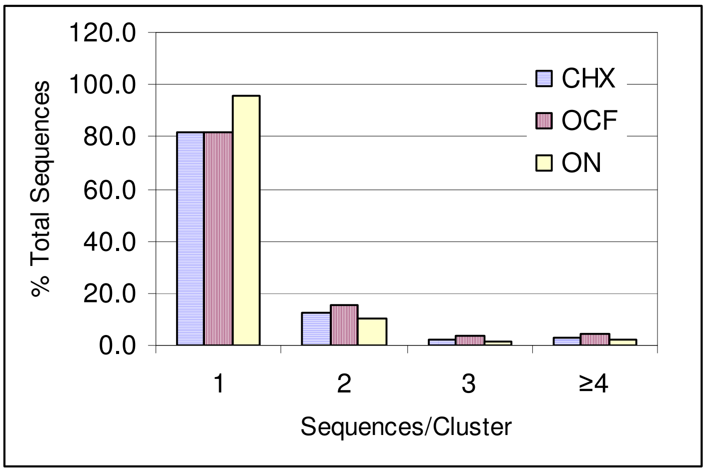

Fig.1. Redundancy distributions of CHX, OCF and ON EST sets

In order to annotate the EST sequences for function, BlastX searches of International protein databases were performed and a summary of this analysis is presented in Table 4. Over all, 74% of sequences have a significant match of known proteins and the rest are probably novel. The detailed results of this analysis has been entered and stored in an Access database.

Table 4. Summary of BlastX searches to functionally annotate EST clones

EST Set	Number of cleaned clones	Significant Protein Hit	No Significant Protein Hit	No Protein Hit
GH_DEa	3937	3494	404	39
GH_DEb	2850	1833	844	173
IE	1415	703	290	422
CHX	1037	686	281	70
ON	1136	679	333	124
OCF	1091	860	222	69
CDO	441	375	48	18
CRH	1106	888	167	51
LSL	1777	1423	242	112
TOTAL	14790	10881 (73.6%)	2831 (19.1%)	1078 (7.3%)

Fig. 2. 20 most abundant and known function groups in OCF EST set

A more detailed function classification for the CHX, OCF and ON EST sets has been carried out and the results are presented in Figure 2 and 3. The ESTs were classified into different functional groups on the basis of strong homology to genes of known function as identified from their BLASTX annotations. Fig. 2 shows the most abundant 20 functional classes of the OCF EST set with the most abundant groups being protein synthesis and DNA transcription, underlining a developmental program of active development and complex differentiation that is occurring in the ovule at and just before fertilisation. The distribution of genes within functional classes in the CHX and ON EST sets were compared to that of the OCF ESTs (Fig. 3) to determine whether there were significant changes in the types of genes represented in the two modified libraries. The most significant differences between the CHX ESTs and the OCF ESTs were an increase of genes of protein phosphorylation and a decrease of protein synthesis genes. Overall, the CHX and ON ESTs displayed different gene expression profiles relative to the OCF ESTs, increasing the potential for greater representation of the cotton genome.

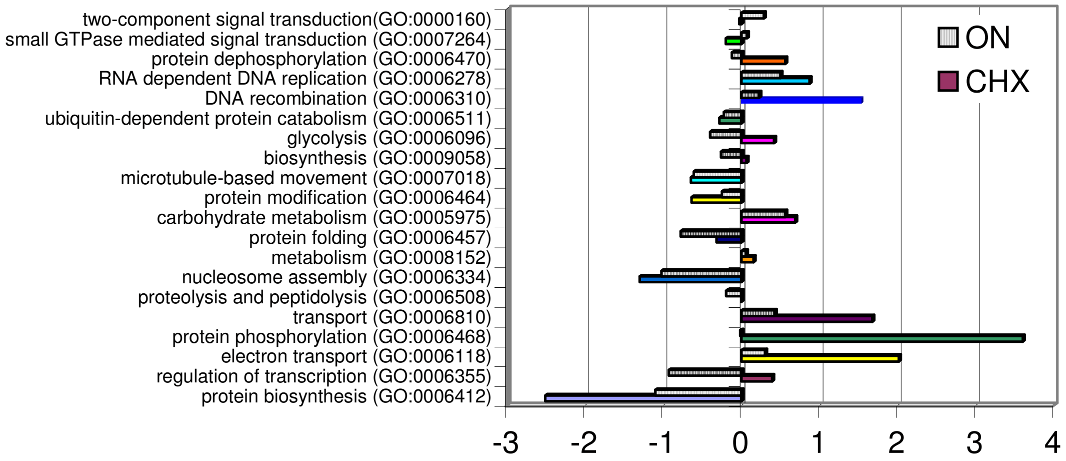

Fig. 3. The change of distributions of CHX and ON ESTs within functional classes as compared to that of the OCF ESTs

Conclusions

We are close to producing probably the biggest cotton uni-gene microarray in the world. This array will be one of the most powerful tools for cotton gene expression studies. Cotton researchers can use this array to research various problems in cotton biology to generate new knowledge and identify molecular markers, as well as key genes important for the Australian cotton Industry.

Outputs

The last few years have seen rapid advances in the analysis of plant genomes that have increased our knowledge of gene structure and function in plants and shown potential for the identification and manipulation of important genes for the agronomic improvement of crop plants through biotechnology. Multi-national companies (such as Monsanto, Dow and Syngenta) are generating genome sequences and libraries of expressed genes from agricultural crops from which they hope to isolate key target genes specifying agronomic traits or protection of plants from pests and diseases. These genome resources will be propriety and protected by patents so will not be available for Australian researchers or Industry except through license arrangements that are generally not to the farmer's advantage. Australian researchers must be active in these areas if we are to participate at an equal level with the multi-nationals over the next decade. CSIRO Plant Industry has initiated genomics approaches to the analysis of genes in cereals such as wheat, barley and rice (through CSIRO and Graingene), but the importance of cotton to the Australian economy and our demonstrated capacity to use biotechnological approaches for cotton improvement warrant a concerted effort in cotton with a focus on key traits of value to the industry, such as fibre quality genes, and disease and pest tolerance genes. The process of gene discovery has generated new Intellectual Property for Australia. Ownership of this IP will enable us to use it in our variety development program and we will be able to trade access under suitable conditions for return access to other proprietary traits of value for Australian cotton. A better understanding of the molecular processes in fibre initiation and development will allow us to design new strategies to genetically improve fibre yield and quality in Australian varieties.

Summary

- *technical advances achieved (eg commercially significant developments, patents applied for or granted licenses, etc.)*

The significant increase in the genes represented in our EST collections as a result of collaborations with other cotton researchers and the purchase of fibre ESTs will increase the coverage of cotton genes by this cotton uni-microarray and renders it a more powerful tool to study cotton fibre development and quality traits as well as many other aspects of cotton biology.

The discovery of a list of genes potentially involved in fibre initiation marks an important first step towards unravelling molecular mechanisms that control fibre initiation and development and has resulted in the filing of a patent by CSIRO adding to our growing portfolio of cotton fibre patents. The Bayer company has been impressed by our progress in this area and has started funding new research into the study of genes determining cotton fibre quality traits.

Provisional Patent: **Yingru Wu, Adriane C. Machado, Danny J. Llewellyn and Elizabeth S. Dennis "**Genes involved in plant fibre development" filed in the end of March 2004

Intellectual Property register required?

The above provisional patent has been lodged by CSIRO.

Further activities

- *to further develop or to exploit the project technology.*

The array will continue to be used for cotton fibre research as part of a new CRDC project (CSP168C) and a project funded by Bayer Crop Sciences through the CSIRO/Bayer Alliance. The unigene array will be available to other researchers on a collaborative basis and with time CSIRO will develop new projects to use the arrays for fusarium and waterlogging research.

- *for the future presentation and dissemination of the project outcomes.*

We plan to announce the development of this uni-gene array in the Australian Cottongrower once it is accomplished.

- *for future research.*

This array is going to be used for the discovery of fibre quality and development genes as detailed in the project “Unravelling the Molecular Basis for Cotton Fibre Quality” CSP168C.

Publications

Yingru Wu, Danny Llewellyn and Elizabeth Dennis (2002) A quick and easy method for isolating good quality RNA from cotton (*Gossypium hirsutum* L.) tissues. Plant Molecular Biology Reporter 20: 213-218.

Yingru Wu, Caitriona Dowd, Emmanuelle Faivre-Nitschke, Danny Llewellyn and Elizabeth Dennis (2002) Cotton Genomics. In: Proceedings of Australian cotton conference, Brisbane.

R.S. Anderssen, **Y. Wu**, R. Dolferus and I. Saunders (2004) An a posteriori Strategy for Enhancing Gene Discovery in Anonymous cDNA Microarray Experiments. Bioinformatics 20: 1721-1727

Yingru Wu, Adriane C. Machado, Rosemary White, Danny J. Llewellyn and Elizabeth S. Dennis. Identification of Early Genes Expressed During Cotton Fibre Initiation Using cDNA Microarrays (Under revision and resubmission)

Yingru Wu, Danny J. Llewellyn and Elizabeth S. Dennis**.** Cycloheximide treatment of cotton ovules alters the abundance of specific classes of mRNAs and provides a method of generating novel ESTs for microarray expression profiling (under revision for resubmission).

Impact on Cotton Industry

It is too early to assess the impact for cotton industry at this stage. However, the process of gene discovery has generated new Intellectual Property for Australia. Ownership of this IP will enable us to use it in our variety development program and we will be able to trade access under suitable conditions for return access to other proprietary traits of value for Australian cotton. A better understanding of the molecular processes in fibre initiation and development will allow us to design new strategies to genetically improve fibre yield and quality in Australian varieties.

Part 4 – Final Report Executive Summary

Genomic technologies are set to revolutionise plant biology and we have already seen significant advances in the rate of gene discovery and new ideas engendered by the sequencing of genomes from two model plants, Arabidopsis and rice. This is pushing the frontiers with new developments in biotechnology as plant scientists start to now try to understand what all the genes in these plants do and how we might manipulate them to improve our crops. Monsanto, for example, is reported to be developing new drought stress tolerance transgenes for crops that has come out of studies of master stress regulator genes first identified in Arabidopsis. Some of the research from these model plants will flow through to cotton but to achieve the maximum benefit, particularly in the area of fibre, as these biological structures are not found in rice and Arabidopsis, we must develop a capacity for genomic studies in cotton. CSIRO has taken the first steps in this direction by developing cDNA microarray technologies for cotton that will help us understand better what genes are important in fibre development as well as in other aspects of cotton biology.

Microarrays are formed by robotically depositing specific fragments of DNA (genes) at indexed locations onto glass microscope slides. The microarrays can then be probed with the genes expressed in particular tissues or under particular conditions that have been chemically tagged with a coloured dye. Scanning the slide can then tell you, by the colour of the spots where the DNA has been printed, how much each genes is expressed in those conditions – thousands of genes at a time, giving a whole picture of the differences in gene expression. Knowing what sequence and hence inferred function of the different genes on the slide can then tell you what genes expression is contribution to the problem under investigation.

This project has focussed on developing a microarray slide that contains unique representatives of large numbers of the genes expressed in cotton (concentrating on those expressed in fibres, but also some other tissues). While it will still be a long way from having all the genes present in cotton (Arabidopsis has at least 30,000 genes) it should have a large enough representation that it will be useful for dissecting many important problems in cotton biology and biotechnology. CSIRO had already printed and used a slide containing over 10,000 cotton genes, but only a small fraction had been sequenced. During this project we added new clones to the set and sequenced about a third of the genes and were able to get the rest sequenced by a collaborator in the US. To add to the genes we already had, we were able to purchase another 13,000 unique genes from the US Cotton Genomics Centre (Thea Wilkins, Director) and have been using bioinformatics programs to search through the gene sequences and weed out those that occur more than once, so that we can produce a so-called uni-gene set (where each different gene is only represented once). This allows us to put more informative genes on the slide as they currently have a limit of about 25,000 spots. The laborious process of amplifying up each gene to produce enough DNA to print hundreds of slides and checking that each gene has been amplified properly has been completed and all that remains now is to assemble the sequences into the unigene set ready for printing. The project has generated an important new tool for cotton researches and the array should contain about 25,000 unique genes - the biggest of its type for cotton available anywhere in the world. We will shortly start to use the slide in a project designed to discover what genes are important in determining fibre quality traits like length, strength and fineness and will hopefully help our breeder's improve the quality of Australian cotton.