An EST-Based Genomics Project in Potato, *Solanum tuberosum*

Gregory Cormack

A Thesis
in
The Department
of
Biology

Presented in Partial Fulfilment
of the Requirements for the Degree
of Master of Science at
Concordia University
Montreal, Canada

September 2004

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-94655-X
Our file Notre référence
ISBN: 0-612-94655-X

ABSTRACT

An EST-Based Genomics Project
in Potato, *Solanum tuberosum*

Gregory Cormack

Expressed sequence tags (ESTs) are partial DNA sequences generated from either the 5′ or the 3′ end of cDNA clones. Many large scale cDNA sequencing projects generating thousands of ESTs have been performed. Such EST collections can reflect a substantial proportion of the expressed genes of a species under a given set of conditions. Two potato cDNA libraries derived from pathogen challenged tissue were enriched by virtual subtraction and single pass sequencing of selected clones resulted in 4,795 EST sequences. 4,184 unigenes (3,305 singletons + 879 contigs) were discovered from grouping the ESTs into clusters and contigs. The virtual subtraction enrichment before sequencing of the cDNA libraries was found to be highly effective at reducing EST redundancy and enriching for ESTs of genes expressed at low levels, namely transcription factor genes. As much as 30 fold decreases in the numbers of ESTs representing certain highly expressed genes were observed while some transcription factor ESTs were up to 10 times more abundant than in public EST sets developed from randomly selected clones.

In addition, the EST collection was analyzed for percent full length cDNA composition corresponding to genes of different lengths. A potato microarray was constructed from the enriched set of cDNA clones. The *B2* gene of potato, a gene believed to be involved in its disease response, was cloned into the gene overexpression vector pRD526 and into the RNAi gene silencing vector pDARTHVECTOR. The constructs were each transformed into *Agrobacterium tumefaciens.* Future studies conducted using transgenic plants made from these constructs and the microarray will enable the elucidation of the function of *B2* through the discovery of interactions between *B2* and other disease response genes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

EST Sequencing

Expressed Sequence Tags (ESTs) are partial DNA sequences generated from either the 5' or the 3' end of cDNA clones. The ESTs resulting from the large scale sequencing of cDNA libraries provide a cost-effective way to gain insight into the genome of a species (Ohlrogge & Benning, 2000). EST sequencing has a number of advantages over whole genome sequencing. It immediately provides information on the genes being transcribed in the species in question under the given conditions. Genomic DNA sequence data, on the other hand, must first be extensively analyzed to find possible genes, revealing little information regarding the conditions under which they are expressed. Repetitive elements present in untranscribed genomic DNA make assembling genomic sequence a very difficult task. In sequencing the *Arabidopsis thaliana* and rice genomes, contig maps, which represent the optimal selection of BAC clones to cover whole chromosomes, were used. Contig map construction is confounded in polyploid genomes by the presence of highly repetitive DNA (Mayer & Mewes, 2001). EST sequencing hastens gene discovery and characterization by the clues provided using ESTs to search public databases for alignment matches to genes of known function (Clarke *et al.*, 2002). Compared to genomic DNA sequencing data, ESTs provide a quick route to gene expression information by representing the transcriptome of an organism at a given time and set of conditions.

EST sequencing can function as a first step in gene discovery and characterization. Similarity comparisons of a set of EST sequences against a database of DNA or protein sequences of known function can provide putative annotations for the ESTs. The putative functional assignments can facilitate the process of experimentally determining the functions of the genes represented by the ESTs (Clarke *et al.*, 2002). The program BLASTx was used to compare 2,137 unique wheat EST sequences for alignment similarity against the public *Arabidopsis thaliana* protein sequence database at the Institute for Genomic Research (TIGR) (Clarke *et al.*, 2002). Using a cut-off E value of 10^{-6} to determine a match, 40% (853) of the EST sequences could be assigned a function. Asamizu *et al.* (2000) also used the program BLASTx to make sequence comparisons between translated EST sequences and the sequences in the protein database at the National Centre for Biotechnology Information (NCBI). ESTs were assigned the most statistically significant functional annotation above an E value threshold of 10^{-14}. Forty percent (4,816) of the 12,028 groups of non-redundant 3′-end ESTs used in the study could be assigned a function through similarity to genes of known function. In addition, annotated genes are often grouped into broader functional categories. Crookshanks *et al.* (2001) classified ESTs into twelve functional groups, such as protein destination, development, and protein synthesis by comparing the translated ESTs to the functional categories at the Munich Information Center for Protein Sequences (MIPS) (http://mips.gsf.de). Gene discovery and characterization can begin with the assignment to ESTs of putative functions based on sequence homology to genes or gene products of known function.

Since EST sequences are usually obtained from randomly chosen clones from cDNA libraries, abundantly expressed genes are often sequenced multiple times (Rounsley *et al.*, 1996). This redundancy of sequences can be reduced by using normalized libraries in which the frequency of highly expressed genes is decreased. Numerous subtractive cDNA hybridization methods have been developed with the purpose of reducing the representation of highly expressed genes and increasing the relative abundance of rare transcript cDNAs. Essentially, these involve the hybridization of a set of "tester" cDNAs containing the target cDNAs of interest against a set of "driver" cDNAs. The unhybridized fraction representing the target cDNAs is then separated from the hybridized common cDNAs and used to construct the cDNA library. These methods have been successful at identifying many genes which are highly induced, e.g. *PR*-genes, but have been found to be inefficient for obtaining low abundance transcripts (Duguin & Dinauer, 1990; Hara *et al.*, 1991; Hendrick *et al.*, 1984). Employed as a part of cDNA library construction, subtractive hybridization (Schweinfest *et al.*, 1990), suppression subtractive hybridization (Diatchenko *et al.*, 1996), and the cap-trapper method (Carninci *et al.*, 2000) are three examples of techniques which have been able to overcome this problem. Other techniques have been developed which normalize existing cDNA libraries based on reassociation kinetics (Soares *et al.*, 1994; Bonaldo *et al.*, 1996). A bias favouring the enrichment for short cDNAs has, however, been associated with these techniques. In addition, 3′ end truncated clones have been shown to result frequently. But, subsequent refinements to these normalization methods have partially overcome these problems (Bonaldo *et al.*, 1996).

Virtual subtraction is a normalization method developed by Li and Thomas (1998). It takes a cDNA library and enriches for the cDNAs of rare transcripts by decreasing the relative representation of the cDNAs of abundant transcripts. cDNA libraries are arrayed at low density on nylon membranes so that individual cDNA recombinants can be easily distinguished. These are screened with radiolabeled probes which are made from cDNAs usually derived from a mixture of mRNA populations representing different tissues and/or treatments. cDNA clones with strong hybridization signals represent clones from highly expressed genes. Weak signals represent clones from genes with low abundance mRNAs. Only clones with low signals are chosen for sequencing. Li and Thomas successfully used this methodology to enrich an *Arabidopsis thaliana* cDNA library for clones of genes expressed in developing embryos. A developing seed cDNA library was constructed and arrayed on nylon membranes. Radiolabeled probes made from leaf, root, silique, and abundant seed cDNAs were hybridized to the array. Five hundred clones with low signal were sequenced and BLAST searched against the Genbank database. Only rare cDNAs and no moderately to abundantly transcribed genes were found (Li & Thomas, 1998). Thus, a set of ESTs can be generated with less sequence redundancy by employing various subtraction techniques.

In spite of the enrichment for clones derived from low abundance mRNAs, a collection of ESTs is likely to contain redundancy. This redundancy can be managed by grouping sequences together based on sequence similarity. A first step involves the clustering of EST sequences. Sequences are grouped together based on a minimum percentage of sequence similarity over a minimum number of base pairs (Burke *et al.*, 1999). Clusters are similar to gene family groupings but are less flexible than the criteria that are traditionally used to define a gene family which incorporate probable or proven function in consideration of family and superfamily groups. Within clusters, contigs are assembled. Contig assembly combines sequences from the same gene into a single contiguous sequence. A tentative consensus (TC) sequence representing the ESTs of a given contig is constructed. It is designated "tentative" because of uncertainty due to questionable sequence quality, low redundancy of sequences, allelic variation (due to the fact that sequences in the public databases are usually derived from different genotypes), the presence of sequences from homeologous loci within polyploid species' genomes, and mRNA splice variants. Tentative contigs are assembled from several overlapping ESTs to form a consensus sequence which represents the sum of sequence information contained in the individual ESTs and the possible resolution of sequence ambiguities that might be present in a small subset of the contributing sequences. Consensus sequences are usually longer than any of the individual ESTs from which they are made. This increases the chances of finding statistically significant sequence alignments in public gene sequence databases, which may provide a functional annotation for the EST (Rounsley *et al.*, 1996). Grouping ESTs into clusters and contigs facilitates the management and best use of the redundancy present in a collection of EST sequences.

A number of large scale cDNA sequencing projects have been realized. One of the first sets of potato ESTs was developed by Crookshanks *et al.* (2001) who generated 6,077 EST sequences in the single pass sequencing of a mature potato tuber cDNA library. Analysis of the frequency of different genes in the EST set was used to determine the relative expression profiles of genes from tuber of potato. The 6,077 clones were assembled into 828 clusters and 1,533 singletons. The clones were also classified according to the *Arabidopsis thaliana* functional classification found at the Munich Information Center for Protein Sequences (MIPS).

Genes involved in protein synthesis, protein targeting, and cell defense were found to predominate in tuber. Asamizu *et al.* (2000) constructed cDNA libraries from the aboveground organs, flower buds, roots, liquid-cultured seedlings, and green siliques of *Arabidopsis thaliana*. Randomly chosen cDNA clones from each library were sequenced yielding a total of 14,026 5'-end ESTs and 39,207 3'-end ESTs. The 3'-EST sequences were clustered in order to identify the number of independent unique genes. The 3'-EST sequences were clustered into 12,028 non-redundant groups. BLASTx similarity searches against the NCBI protein database were performed on the non-redundant 3'-EST groups. Among these, 4,816 groups were found to be similar to genes of known function. 1,864 groups were found to be similar to genes classified as hypothetical. The remaining 5,348 groups were classified as novel sequences since they had no high-scoring sequence matches in existing databases. Results of EST sequencing programs for rice (Yamamoto & Sasaki, 1997), potato (Ronning *et al.*, 2003), sugarcane (Ma et al., 2003), and poplar (Sterky *et al.*, 1998) have also been reported. The results of other large scale EST sequencing programs which have taken place in recent years are also evident by the large EST collections for 13 plant species that can be seen among the gene indexes posted on the internet by TIGR.

EST sequence collections can provide information on gene expression based on the rational that the frequency of EST sequences in a data set is proportional to the frequency of the corresponding mRNA in the tissue from which the EST was derived (Audic & Claverie, 1997). Since any given collection of EST sequences represents a small sample of the overall population of expressed genes, the statistics of sampling small numbers from a large population must be considered. A rigorous statistical test developed by Audic and Claverie (1997) addresses this issue. Another limitation of the digital Northern approach is the possible under-representation of certain mRNA species due to obstacles to the reverse transcription process - mRNA secondary structure may cause reverse transcriptase to ineffectively produce cDNA (Ohlrogge & Benning, 2000). Ewing *et al.* (1999) conducted an extensive gene expression analysis of the rice ESTs available in Genbank's dbEST database. An expression profile was computed for each gene represented by at least 5 ESTs in each of 10 different cDNA libraries. Ten expression measurements were therefore used to derive each gene's expression profile. Correlated patterns of gene expression between different tissues were discovered. In addition, when gene expression profiles were clustered, genes with similar functions were found grouped together. The digital Northern approach can make use of certain EST collections to provide gene expression information.

Microarrays and the Study of the Plant Disease Response

DNA microarray technology is a miniaturized hybridization based method enabling the examination of the expression levels of thousands of genes simultaneously. DNA representing each gene is immobilized on a specially coated glass slide at a density of up to 10,000 spots / 3.24 cm^2 (Kehoe *et al.*, 1999). The DNA of a microarray is termed the probe and the fluorescently labelled soluble cDNA generated from the reverse transcription of mRNA and used for hybridization is termed the target (Aharoni & Vorst, 2001). Fluorescent dyes with different excitation and emission optima are used to label two different target cDNA samples. The two differently labelled samples are simultaneously hybridized to a single microarray. The two sets of labelled targets are from mRNA samples being compared, such as two different tissue types of the same plant, from pathogen challenged plant tissue versus uninfected tissue, or from plant tissue at two different developmental stages (Richmond & Somerville, 2000). For each gene represented on the microarray, the strength of each of the two fluorescence emissions represents the amount of the gene's specific mRNA that is represented by the given target label that has hybridized to the array. The signals are quantified enabling the calculation of expression ratios for each gene (Aharoni & Vorst, 2001). The result is often displayed as a false-color image. After hybridization and scanning, the spots on the microarray might be shown as 1 of 3 different colors. For example, yellow might be used to represent spots with a signal ratio of 1, where both signals are similar in intensity; red might be used to show that the "red" probe signal was stronger; and green might be used to show that the "green" probe signal was stronger (Kehoe *et al.*, 1999). Large scale gene expression studies have been carried out using microarrays. For example, a microarray containing approximately 7,000 full length *Arabidopsis thaliana* cDNA sequences was constructed by Seki *et al.* (2002) to examine the expression profiles of genes under conditions of drought, cold, and high salinity stress. The transcripts of 53, 277, and 194 genes were found to increase more than five fold after cold, drought, and high salinity treatments, respectively. Among these induced genes, 22 responded to all three stresses. Microarray technology enables the simultaneous analysis of the expression patterns of thousands of genes. It is well suited to the study of the gene expression patterns comprising the response of potato to challenge by the oomycete pathogen, *Phytophthora infestans*.

Microarray analysis of the response of plants to pathogen challenge has been instrumental in the identification of new pathogenesis-related genes, the identification of co-regulated genes and the associated regulatory systems, and the uncovering of interactions between different signaling pathways (Wan *et al.*, 2002). A large number of potential defense-related genes in maize were identified by Baldwin *et al.* (1999) using a 1,500 gene microarray. They reported 117 genes which consistently had altered mRNA levels 6 hours after treatments with the fungal pathogen *Cochliobolus carbonum*. Using a 2,375 gene *Arabidopsis thaliana* microarray, Schenk *et al.* (2000) observed substantial changes in the abundance of 705 mRNAs in response to inoculation with the fungal pathogen *Alternaria brassicicola* or treatment with the defense activating signaling molecules salicylic acid, methyl jasmonate, or ethylene. Among the 705 genes with altered expression, 106 genes had not been previously ascribed function.

Cluster analysis of microarray data to group genes with similar patterns of expression has enabled the identification of co-regulated genes. In a microarray experiment monitoring the expression pattern of 1,300 *Arabidopsis thaliana* genes under drought and cold stresses, in transgenic plants overexpressing the transcription factor DREB1A, 12 stress inducible genes were found to be co-regulated by DREB1A. Eleven of these 12 genes were found to contain the dehydration-responsive element (DRE) or DRE-related CCGAC core motif in their promoter regions (Seki *et al.*, 2001). Another study identified a cluster of 41 *Arabidopsis thaliana* genes that showed the same expression pattern in response to bacterial, fungal, oomycete, or viral infection (Chen *et al.*, 2002). A novel motif was found to be statistically over represented in the promoters of the genes in this cluster, lending further support for their co-regulation.

Since the expression patterns of thousands of genes can be studied simultaneously in a single experiment, interactions between signal transduction pathways can be observed. Cross-talk has been documented between different defense response pathways and between defense response pathways and other plant response pathways (Wan *et al.*, 2002). In the comparison of gene regulation by salicylic acid, methyl jasmonate, ethylene, and pathogen challenge, 169 mRNAs co-regulated by multiple treatments or defense pathways were identified (Schenk *et al.*, 2000). Genes regulated by salicylic acid and methyl jasmonate represented the largest number of genes that were coinduced or corepressed. It was also found that half of the genes induced by ethylene treatment were also induced by methyl jasmonate treatment. There is a substantial network of regulatory interactions and coordination occurring during plant defense among the different defense signaling pathways (Schenk *et al.*, 2000). Cross-talk between defense response pathways and other plant response pathways has been elucidated using microarray technology. For example, the study by Chen *et al.* (2002) found five transcription factors activated by both abiotic stress and bacterial infection.

The *B2* Gene of Potato

The *B2* gene of potato is thought to be involved in plant defense through its interaction with the suppressor element binding factor (SEBF) which binds the promoter of the pathogenesis-related gene *PR-10a*. *PR-10a* has been found to be transcriptionally activated in response to pathogen infection or elicitor treatment (Marineau *et al.*, 1987; Matton & Brisson, 1989; Constabel & Brisson, 1992). SEBF is a regulatory protein which has been found to bind in a sequence specific manner to the promoter region of *PR-10a* of potato and to repress transcription (Boyle & Brisson, 2001). The authors also speculate that SEBF may act as a transcriptional repressor for a number of other pathogenesis-related genes via the suppressor element binding sequence in their promoters. A study of protein-protein interactions with SEBF using the yeast two hybrid system uncovered a strong interaction between SEBF and the B2 protein (N. Brisson lab, University of Montreal, unpublished results). The interaction between SEBF and the B2 protein points to a possible role for B2 as a suppressor of SEBF function.

In the present work, a large number of potato ESTs were generated from cDNA libraries made from leaf and tuber tissues challenged by the oomycete pathogen, *Phytophthora infestans*. Virtual subtractions were performed to reduce the redundancy of cDNAs chosen for sequencing. The success of the virtual subtractions at reducing the abundance of clones from highly transcribed genes and increasing the proportional representation of clones from genes expressed at low levels was gauged by comparing the frequencies of clones of transcription factors and highly expressed genes between this study's EST set and a set of unenriched ESTs. Sequences from the virtually subtracted cDNA libraries were grouped into clusters and contigs and annotated via alignment similarity searches to the Genbank public database. Estimates of the percent full length cDNA composition of the sequenced clones were also performed.

A potato microarray was constructed from PCR amplicons of the cDNAs sequenced as part of the virtual subtractions. The elucidation by microarray analysis of the gene expression patterns comprising the response of potato to challenge by *Phytophthora infestans* is anticipated but is not included in the present work. A gene thought to be involved in the disease response of potato, *B2* has been cloned into a gene overexpression vector and a gene silencing vector and each construct transformed into *Agrobacterium tumefaciens*. These will be used in the future to generate *B2* transgenic plants which will be used for microarray analysis to better understand the effect of *B2* expression on the expression of other plant genes.

MATERIALS AND METHODS

Virtual Subtraction

Virtual subtraction was performed to enrich cDNA sets for clones from genes with low abundance mRNAs. Two virtual subtractions were done using cDNA libraries constructed from potato (*Solanum tuberosum* cv Kennebec) infected with the incompatible *Phytophthora infestans* Race Zero pathovar. The SV7 virtual subtraction was carried out with a cDNA library made from infected leaf tissue and the SV8 virtual subtraction used a cDNA library made from infected tuber tissue.

Each of the two cDNA libraries was made with equal mixtures of mRNAs from each of seven time points after infection, 0.5, 2, 6, 10, 20, 28, and 50 hours. The cDNA libraries were made in the Lambda Zap II vector and converted to a pBluescript II SK$^+$ plasmid library in SOLR *E. coli* by massive plasmid rescue using the Stratagene ZAP-cDNA Synthesis Kit according to the manufacturer's protocol.

The Virtual Subtraction Procedure

Each cDNA library in SOLR *E. coli* was plated on LB-agar+amp+kan plates with ampicilin at a concentration of 100 µg/mL and kanamycin at a concentration of 50 µg/mL. For each virtual subtraction, approximately 10,000 colonies were randomly picked and cultured in 175 µL of LB-broth+amp+kan in 384-well microtiter plates. After inoculation, each plate was covered with a lid, sealed with parafilm around its edges, placed in a plastic bag to prevent evaporation, and incubated with agitation (250 RPM) for 16 hours at 37°C. The plates were then stored at 4°C until replica plating could take place.

The 384-well plates were replica spotted twice and the replicas were inspected for growth. Using a sterilized 384-pin inoculator, culture from each well of the 384-well plate was transferred to two precut (115 X 75 mm) nylon membranes (Hybond-N+, Amersham Pharmacia Biotech, Cat. #RPN303B) which were placed, culture sides up, in large (150 X 15 mm) Petrie dishes which contained LB-agar+amp+kan. Plates were incubated for 48 hours at 37°C. The membranes were then inspected for amount of growth of bacterial spots. Those bacterial spots showing minimal growth and those membrane areas showing no bacterial growth where growth would be expected were noted.

The membranes were immersed in a 1.5 M NaCl, 0.5 M NaOH lysis solution for 5 minutes and then transferred to a 1.5 M NaCl, 0.5 M Tris-HCl neutralization solution for 5 minutes. Membranes were transferred to fresh neutralization solution and the remains of the lysed bacteria were rubbed off with gloved hands. The membranes were transferred to a 2X SSC (0.3 M NaCl, 0.03 M Na_3 citrate•$2H_2O$) solution for 30 seconds. Membranes were air dried and the nucleic acid was cross linked to the membranes using a UV cross linker (Hoefer UVC 500).

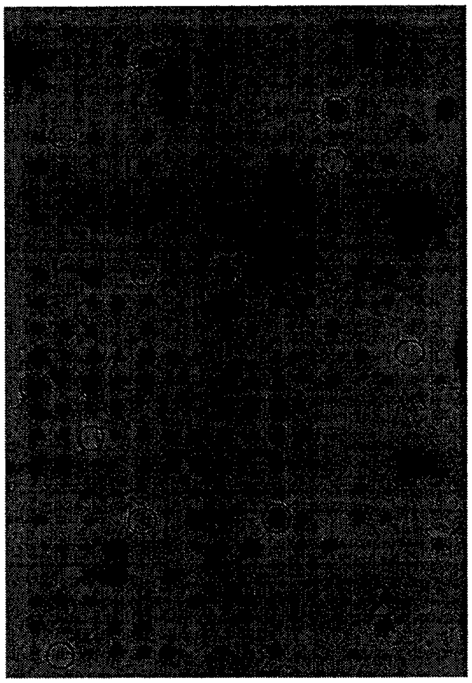
Three membranes were placed side by side on a piece of wetted vinyl mesh (Nitex nylon mesh, 112 µm, Sefar, Cat. #HC3-112), rolled up, and placed in a hybridization bottle with 25 mL of prehybridization solution composed of 0.25 M Na_2HPO_4, 1% BSA, 7% SDS, and 1 mM EDTA. The membranes were incubated with rotation (4 RPM) at 65°C for between 3 and 16 hours.

For each virtual subtraction, probes were made from a sample of double stranded cDNAs representative of each time point which had been kept aside during the construction of the original cDNA library. The protocol of Boehringer Mannheim's High Prime DNA Labelling Kit was followed in order to generate the radioactively labelled probes. The protocol involves the annealing of random hexamer primers to the denatured linear cDNAs and primer extension using the Klenow fragment of DNA polymerase I and dGTP, dTTP, and dCTP combined with dATP^{32} (3,000 Ci/mmol). Unincorporated nucleotides and hexamer primers were then removed by passage through a Sephadex G-50 Quickspin column (Boehringer Mannheim). The radioactivity of the probe was measured by scintillation counter and some labeled time-zero cDNA probe equal to one seventh of the total radioactivity was added to enhance the hybridization signal of genes not induced in the pathogen response.

The prehybridization solution was replaced with approximately 20 mL of hybridization solution (0.25 M Na_2HPO_4, 1% BSA, 7% SDS, and 1 mM EDTA). The probe was denatured for 5 minutes at 100°C and added to the hybridization solution. The membranes were incubated overnight at 65°C with rotation at 4 RPM. The hybridization solution was removed and the membranes were washed with 25 mL of 2X SSC, 0.1% SDS for 30 minutes at 25°C and 30 minutes at 35°C. Membranes were washed twice more with 25 mL of 1X SSC, 0.1% SDS for 30 minutes at 45°C and 55°C, respectively. The final wash was with 25 mL of 0.1X SSC, 0.1% SDS for 10 minutes at 55°C. Membranes were then dried on sheets of Whatman paper. They were then exposed to X-ray film (Kodak Biomax) for 29 hours and subsequently developed in an automatic film developer (Agfa Curix 60).

The films were marked for the no growth areas and low growth bacterial spots made note of previously. No sequencing was carried out for these cultures. Those clones corresponding to locations on the films showing no hybridization signal or a low hybridization signal were chosen for sequencing (See figure 1 below.). Approximately 25% of the colonies used for virtual subtraction screening were chosen for sequencing. Five μL of the culture of selected clones was transferred from its original 384-well plate location to the well of a 96-well plate containing 200 μL of LB-broth+amp+kan. Each 96-well plate was incubated for 24 hours at 37°C. Forty five μL of 100% glycerol was added to each of these cultures and stirred thoroughly by up and down by pipetting. Glycerol stocks were stored at -80°C until used to inoculate new cultures for plasmid purification.

Figure 1: Autoradiogram Showing Virtual Subtraction Hybridization Results



Examples of clones that would be chosen for sequencing are represented by open circles. Examples of clones of strongly expressed genes that would not be sequenced are also shown. These are represented by crossed circles.

O = chosen for sequencing
Ø = excluded from sequencing

The fresh cultures grown in 2 mL of TB-broth+amp+kan inoculated with 5 µL of the glycerol stock were incubated for 24 hours at 37°C. Plasmid purification was carried out with a Qiagen Qiaprep 96 miniprep kit according to the manufacturer's protocol. The final elution of the plasmids was with 100 µL of 10 mM Tris-Cl pH 8.5. The Montreal Genome Centre sequenced the cDNA inserts from the 5' end using the pBluescript T3 primer. The sequencing results were returned from the centre as FASTA and trace files along with the results of BLASTx searches of the GenBank protein database.

Cluster / Contig Analysis of the SV7 and SV8 EST Sequences

The EST sequences of the SV7 and SV8 virtual subtractions were grouped into clusters using the program d2_cluster of the stackPACK package of analysis programs. Sequences with a minimum of 96% identity for at least 50 bp were included in the same cluster. Clustering was done within the SV7 and SV8 sets and with the combined set of the two databases.

Contig assembly was done within the clustered sequences with Phrap. Contig assembly was more stringent than clustering since contigs are constructed by anchoring together only the highest quality parts in common between reads. For each contig, a consensus sequence representing each EST in that contig was generated. The cluster / contig groupings are kept at the Functional Genomics of Abiotic Stress (FGAS) web site at http://bioinfo.usask.ca/cgi-bin/abiotic/login.cgi.

Functional Annotation of the SV7 and SV8 EST Sequences

The SV7 and SV8 EST sequences that were returned from the Montreal Genome Centre included the results of BLASTx searches of the NCBI nr protein database. The BLASTx results were used to assign a putative function to each EST. Sequence alignments with E values equal to or smaller than 10^{-10} were considered statistically significant and the annotation of the match with the lowest E value was used as the annotation of the given EST.

Evaluation of the SV7/SV8 Virtual Subtractions for Transcription Factor Enrichment

The SV7 and SV8 EST sequences derived from virtual subtraction were analyzed for the frequency of each of the 11 major transcription factor gene families of *Arabidopsis thaliana* (Riechmann & Ratcliffe, 2000). The same analysis was also carried out on a data set of random ESTs taken from a cDNA library developed from potato challenged with *Phytophtora infestans* downloaded from NCBI. By searching NCBI's nucleotide database with Entrez using the search term "Ronning Baker Buell P. infestans-challenged potato leaf, incompatible reaction", the entire set of 5,434 ESTs is displayed. The options to display the results as FASTA files and send the sequences to file enable the complete set of ESTs to be downloaded from NCBI.

An exemplary amino acid sequence for a member of each transcription factor gene family was taken from Genbank (NCBI) using "Entrez" to search Genbank's protein database. A query describing the gene family in question along with the term "solan*" (e.g. "WRKY solan*") was used to direct the searches to species from the family Solanacea, including potato, tomato, green pepper, and petunia. Among the search results, a full length amino acid sequence was chosen to represent each gene family. Priority was given to sequences from potato or other Solanacea species. *Arabidopsis thaliana* sequences were chosen when no matches with a Solanacea species were obtained. The exemplary full length amino acid sequence representative of the gene family was used as the query to BLASTp search the Genbank *Arabidopsis thaliana* database and three additional representative sequences were taken for each gene family. In order to reflect the range of sequence divergence within the family, sequences with a representative range of E values for the sequence alignments were chosen. Each amino acid sequence chosen had to have the given gene family name present in its description. Each of the three amino acid sequences was used for a tBLASTn search of the combined SV7

/ SV8 EST database at the Functional Genomics of Abiotic Stress (FGAS) website. A non-redundant list of matches with E values equal to or below 10^{-10} along with annotations, scores, and E values was recorded in a spreadsheet.

A database of 5,434 random ESTs from an unenriched potato cDNA library was similarly searched for representatives of each of the 11 transcription factor gene families. The library was made using potato leaf tissue from plants challenged with an incompatible race (US-1) of *Phytophtora infestans* (Ronning *et al.*,2003). The 5,434 ESTs were downloaded from Genbank, uploaded to the FGAS website, and searched using the web-based BLAST programs there. The three amino acid sequences used previously to represent each of the 11 gene families were used to search the 5,434 unenriched EST data set using tBLASTn. The number of matches equal to or below an E value of 10^{-10} and the percentage of the 5,434 unenriched ESTs that this number of matches represents were recorded.

A Chi-squared goodness of fit test was used to test the statistical significance of differences in frequencies of ESTs representing transcription factors in the virtually subtracted versus randomly selected EST data sets. This was tested with the null hypothesis that there were no differences in the frequencies of transcription factor ESTs between the two data sets. The "observed" values were the numbers of matches to transcription factors in the virtual subtraction data set and the "expected" values were the numbers of matches in the randomly selected data set normalized for the number of sequences in the virtual subtraction data set. The calculated Chi-squared value was compared to the critical Chi-squared value obtained at a significance level of 0.005 with 10 degrees of freedom.

The Effect of Virtual Subtraction on EST Redundancy

The combined SV7 and SV8 EST set was compared to the unenriched EST set for the abundance of ESTs derived from 11 highly expressed genes. Ronning *et al.* (2003) constructed 48 different consensus sequences from ESTs common to all nine of the potato cDNA libraries used in the study. The top 11 of the consensus sequences containing the most ESTs and each representing a full length gene were used as queries in tBLASTx searches of the combined SV7 / SV8 database at FGAS. A list of matches with E values equal to or below 10^{-10} along with annotations, scores, and E values was recorded in a spreadsheet. Similarly, the 5,434 unenriched potato EST data set was searched using the

11 full length gene sequences. The number of matches equal to or below an E value of 10^{-10} and the percentage of the 5,434 unenriched ESTs that this number of matches represents were recorded.

A Chi-squared goodness of fit test was used to test the statistical significance of differences in frequencies of ESTs representing highly expressed genes in the virtually subtracted versus randomly selected EST data sets. The null hypothesis is that there were no differences in the frequencies of ESTs for the 11 selected highly expressed genes between the two data sets. The "observed" values were the numbers of matches to the 11 query sequences in the virtual subtraction database and the "expected" values were the numbers of matches in the randomly selected data set normalized for the number of sequences in the virtual subtraction data set. The calculated Chi-squared value was compared to the critical Chi-squared value obtained at a significance level of 0.005 with 10 degrees of freedom.

Percent Full Length cDNAs among the SV7/SV8 Clones

The SV7 / SV8 EST database was searched with full length DNA sequences in order to determine percent full length cDNA composition values corresponding to different query sequence length groups. A full length DNA sequence is functionally defined in the present work as a cDNA clone that includes the start codon. Thirty full length query sequences of different lengths were used to search the EST database.

Full length TC (tentative consensus) query sequences for searching the SV7 / SV8 EST database were found in three ways. The consensus sequences for five of the combined SV7 / SV8 EST database's most redundant contigs listed on the FGAS website were used to BLASTn search the nucleotide database at TIGR (http://www.tigr.org). The searches were limited to *Arabidopsis thaliana* and the available Solanacea species including *Lycopersicum esculentum, Nicotiana tobaccum, Nicotiana benthamiana,* and *Capsicum annum.* Once the start codon was located in the most statistically significant TC sequence match of each search, the sequence was trimmed of its DNA sequence upstream of the ATG. Finding the start codon in each search's best TC match was accomplished by homology comparison carried out via BLASTx searches of the Genbank protein database using the TC sequence as the query. Additional full length TC query sequences were found by searching the combined SV7 / SV8 EST project information spreadsheet which included, among other information, the annotations for each of the SV7 / SV8 EST sequences. The search word "constitutive" resulted in matches to many EST annotations. The names of the proteins associated in the annotations with the word "constitutive" were then used to key word search the annotations of the potato nucleotide database at TIGR. The start codon was located in the TC matches as described previously. Each TC was then trimmed to remove the 5′ UTR region of its DNA sequence, upstream of the ATG. Direct key word searches of the annotations of the TIGR potato nucleotide database were performed to find additional full length TC query sequences. The search terms "ribosome OR ribosomal", "ubiquitin", "polyubiquitin", and "UDP-Glucose:protein transglucosylase OR uptg2" were used. The ATGs in the resulting TCs were located and sequence upstream of them trimmed as before. Each of the TCs found using the different methods was checked to ensure the inclusion of a 3′ UTR. The presence of at least one component EST sequenced in the 3′ to 5′ direction with a (trimmed) poly T header followed by 300 nucleotides or more or at least one 5′ to 3′ EST with a (trimmed) poly A tail preceded by 300 nucleotides or more was taken as evidence of the inclusion of a 3′ UTR in the TC sequence.

Additional full length DNA query sequences for searching the SV7 / SV8 EST database were found in the form of complete potato genes. The nucleotide database at Genbank was searched using Entrez and search terms specifying DNA sequence length and organism. For example, the query "1500:1600[SLEN] AND Solanum tuberosum[Organism]" searched the database for potato DNA sequences between 1500 bp and 1600 bp in length. Potato sequences between the sequence length ranges of 3000 - 4000 bp, 2500 - 2600 bp, 2000 - 2100 bp, 1500 - 1600 bp, and 1000 - 1100 bp were searched for. Among the search results, only complete potato genes were chosen; genomic DNA sequence and genes with introns were excluded. Sequence upstream of the start codons of the complete potato genes was removed before the subsequent searches of the SV7 / SV8 EST database.

The full length DNA sequences were used to BLASTn search the combined SV7 / SV8 EST database at FGAS. The resulting matches for each full length query sequence were listed in a spreadsheet along with their scores and E values. The alignment of each SV7 / SV8 EST match to its full length query sequence was indicated in terms of the number of the nucleotide position where the alignment began in the query and the number of the nucleotide position where the alignment began in the subject. Based on this alignment information, each match was assigned a “not full length”, “full length”, or “likely full length” evaluation.

The full length DNA sequences used to search the SV7 / SV8 EST database were grouped into query sequence groups based on their lengths. Sequences were grouped together such that the difference between the length of the longest and the shortest sequence in a group was no more than about 100 bp. Each query sequence group can be identified by its length number which represents the average length of the full length query sequences in it. The percent full length cDNA composition was calculated for each query sequence group.

Microarray Construction

cDNA inserts of 4,416 clones were amplified by PCR and purified for microarray construction. One hundred µL PCR reactions to amplify the cDNA inserts were done in 96-well PCR reaction strips (Ultident 96-well flexible PCR microplate, catalogue # 17-T323-96N). Each reaction contained 4 µL of 25 fold diluted plasmid from a Qiagen miniprep, 0.5 µM of each of two custom designed pBluescript vector primers - "M13-forward-long" (GTTTTCCCAGTCACGACGTTG) and "M13-reverse-long" (TGAGCGGATAACAATTTCACACAG), 0.3 mM of each dNTP, reaction buffer at 1X (10 µL of MBI/Fermentas *Taq* DNA polymerase reaction buffer), and 1.9 units of *Taq* DNA polymerase (MBI/Fermentas). Plates were sealed with a silicone sealing mat (Axygen Scientific Axymat, catalogue # AM-96-PCR-RD). The cycling conditions were: denaturation at 95°C for 2 minutes followed by 36 amplification cycles of 1 minute denaturation at 95°C, 30 seconds annealing at 59°C, 3 minutes extension at 72°C, and a final extension at 72°C for 7.5 minutes.

The PCR reactions were run on 1% agarose gels in order to ensure the presence of PCR products and to check for single or multiple bands. For each 96-well plate of PCR reactions, two 1% agarose gels were run; each gel had 48 samples. Eight µL of each PCR reaction was electrophoresed at 100 volts for 52 minutes. Gels were stained for 25 minutes in an ethidium bromide solution at a concentration of 0.5 µg/mL, destained for 20 minutes in dH_2O, and photographed under UV light. PCR products with multiple bands are not included in microarray analyses.

Unincorporated nucleotides and primers were removed from the PCR products using the vacuum filtration Millipore MultiScreen-PCR 96-Well Filtration System according to the manufacturer's protocol. In the final step of the purification, the purified DNA sample was dissolved in 60 µL of dH_2O and transferred to a 96-well plate. Samples were lyophilized for 2 hours at 35°C using a Savant SpeedVac SPD111V equipped with a rotor capable of holding two 96-well plates and redissolved in 10 µL of dH_2O.

Five µL of each purified PCR product was mixed with 5 µL of 2X printing buffer (90% DMSO / 100 mM KCl / 40 mM Tris-Cl pH 6.5) and the samples were transferred to 384-well plates (Whatman catalogue # 7701-5101). The twelve 384-well plates were labelled A to F for the SV7 clones and U to Z for the SV8 clones. Control clones, listed in table 1, were prepared in the same manner as the microarray's experimental samples and were added to the unoccupied wells available in plate F. The plates were centrifuged and the samples were printed on Corning Amino-Silane coated UltraGAPS (catalogue #s 40015/40016) and Corning Amino-Silane coated GAPS II microarray slides (catalogue # 40006) using the Virtec/BioRad robot microarray slide printer at Concordia's Centre for Structural and Functional Genomics. Three spots of each clone were printed on each slide.

Table 1: Potato Microarray Controls

	Genes encoding proteins...	Clone source
Constitutively expressed potato genes (Ronning *et al.*, 2003)	GAPDH (glyceraldehyde-3-phosphate dehydrogenase)	SV5-54-C3
	DNAJ protein	SV5-53-C3
	EF-1-ALPHA (elongation factor 1-alpha)	SV5-53-H8
	SAMDC (S-adenosylmethionine decarboxylase)	SV5-37-F4
	alpha tubulin	SV8-05-A5
	TCTP; P23 (translationally controlled tumer protein)	SV5-47-D10
Potato genes upregulated by pathogen challenge	vetispiradiene synthase (Yoshioka *et al.*, 1999)	SV8-21-E2
	Gst-1 (glutathione s-transferase) (Collinge & Boller, 2001; Strittmatter *et al.*, 1996)	SV7-03-H5
	PR-10a (A.K.A. *STH-21*) (defense-related gene)	Brisson Lab (U of M)
Potato genes downregulated by pathogen challenge	Rubisco (ribulose bisphosphate carboxylase)	SV7-19-B11
	chlorophyll a/b binding protein	SV8-07-C1
Negative control	empty pBluescriptSK$^-$ amplified using M-13-forward-long and M-13-reverse-long	lab stock

B2 Cloning

Figure 2: The Complete *B2* Gene Sequence

The *B2* gene was used for making constructs for its 35S promoter driven overexpression and for the RNAi silencing of its expression. The overexpression vector pRD526 and the gene silencing vector pDARTHVECTOR were used for this purpose.

```
   1 TTAACCCTCA CTAAAGGGAA CAAAAGCTGG AGCTCCACCG CGGTGCGGCC
  51 GCTCTAGAAC TAGTGGATCC CCCGGGCTGC AGGAATTCGG CTCGAGGGAA
 101 AAGTGCAAAA AAAAGCAGAA AAATTCATTC TTTTCACAGT AAAGCTTGAA
 151 TCTTACACAA TTTCCTCACT TGGGTATTCA AGAAAACCGC AATAAAAAGC
 201 TCATGGAGAT CAACAACAAC AACAATCAAT CATCTTTCTG GCAGTTCAGT
 251 GACCAGCTTC GTCTGCAGAA CAACAACTTA GCAAATCTCT CTTTGAATGA
 301 TTCAATCTGG AGCAGTAACT ATGGCTCTAA AAGGCCTGAA GAAAGAAGAA
 351 ATTTTGATAT CAGGGTAGGT GGTGACTTCA ACTCTACTGC TAATACTTCT
 401 TCAAACAAGT CAAATTACAA TCTTTTTAGC AACGATGGCT GGAAAATTGC
 451 TGACCCATCT GCTCTCACGG CGGCGAACGG CGGTGGTGCT GCCGGAAAAG
 501 GGGTACTTGG GGTTGGTTTA AATGGTGGAT TCAACAAAGG GGTTTACTCA
 551 AATCAAGCTT TGAACTTCAG TTATAGTAAG GGTACTAATA ATGTTGCAGT
 601 AGGTACCAAA GGGATAAACA AGAAATTTGG TAAAGGGTTT TTTGAAGATG
 651 AGCATAAAAG TGTGAAGAAG AATAACAAGA GTGTTAAAGA GAGTAACAAG
 701 GATGTTAATA GTGAGAAACA GAATGGTGTT GATAAAAGGT TTAAGACTTT
 751 GCCACCAGCA GAATCTTTGC CAAGAAATGA GACGGTTGGT GGATATATTT
 801 TTGTTTGCAA CAATGATACT ATGGCTGAGA ATCTCAAAAG GGAGCTCTTT
 851 GGCTTGCCCC CACGTTACAG GGACTCAGTT AGGCAAATAA CACCTGGATT
 901 GCCTCTTTTT CTGTACAACT ACTCGACCCA TCAGCTTCAC GGAGTATTTG
 951 AGGCTGCANC TTTTGGTGGG TCAAATATTG ATCCATCGGC CTGGGAGGAC
1001 AAGAAGAACC CTGGTGAATC TCGCTTTCCT GCTCAGGTTC GTGTCGTGAC
1051 AAGGAAAGTC TGTGAACCAC TTGAAGAGGA TTCATTCAGG CCAATCCTTC
1101 ACCACTACGA CGGCCCTAAA TTCCGCCTCG AGCTAAACGT TCCAGAGGCT
1151 ATTTCTCTTC TCGACATTTT TGAAGAGAAC AAGAACTAAA TGAATGTTCT
1201 TGTATTACAA GCAGAGAATG GACAATATAC CATTATAAAA AAAAAAAAAA
1251 AAAACTCGAG GGGGG
```

boxed text = start and stop codons of the gene

grey highlighted regions = annealing sites of primers used for cloning *B2* into pRD526. See table 2 for primer sequences.

bold italicized regions = annealing sites of primers used for cloning *B2* into site 1 of pDARTHVECTOR. See table 4 for primer sequences.

underlined regions = annealing sites of primers used for cloning *B2* into site 2 of pDARTHVECTOR. See table 4 for primer sequences.

The pRD526 vector

The plant gene expression vector pRD526 has a strong constitutive CaMV double 35S promoter and a multiple cloning site for insertion of a gene of choice. The plasmid is used for *Agrobacterium tumefaciens* mediated transformation in compatible plant species which include *Solanum tuberosum*. The pRD526 sequence is not available in Genbank. The origin of the plasmid is briefly described here.

pRD526 originates from the pBin19 plasmid. pBin19's kanamycin resistance gene was changed to the wild type kanamycin resistance gene, *NPT-II*, resulting in pRD400. A small 50 bp HindIII-EcoRI fragment of DNA was replaced with a cassette of 934 bp composed of a CaMV double 35S promoter (620 bp), followed by a *cis*-active "translation activator" element (a 44 bp sequence called "AMV" - Datla *et al.*, 1993), followed by the multiple cloning site (NcoI, XbaI, and BamHI, 20 bp), followed by a Nos transcriptional terminator (250 bp). The resulting pRD256 is 12,661 bp in size.

Cloning *B2* into pRD526

PCR primers incorporating unique restriction sites were designed to amplify the complete *B2* gene coding region for insertion into the pRD526 vector.

Table 2: PCR Primers for Cloning *B2* into pRD526

Primer name	Primer direction	Primer sequence	Unique restriction site present
B2Cormack-Full-For2	forward	TGtctagaATGGAGATCAACAACAACAA	XbaI (T/CTAGA)
B2Cormack-Full-Rev2	reverse	CAggatccATTTAGTTCTTGTTCTCTTCAA	BamHI (G/GATCC)

The *B2* gene annealing sites for these primers are indicated in figure 2.

The forward primer includes *B2*'s ATG start codon which was in frame with the ATG present in the vector's multiple cloning site. Thus, translation products beginning at the AUG from the multiple cloning site would yield a fusion including the complete B2 protein sequence.

The 100 µL PCR reaction contained 300 ng of pBluescriptSK$^-$ plasmid DNA containing the full length *B2* gene, 0.3 µM of each primer, 0.2 mM of each dNTP, 1X reaction buffer (MBI/Fermentas *Pfu* DNA polymerase reaction buffer), and 2.5 units of *Pfu* DNA polymerase. The cycling conditions were denaturation at 95°C for 1.5 minutes followed by 35 amplification cycles of 45 seconds denaturation at 95°C, 45 seconds annealing at 58°C, 3 minutes extension at 72°C, and a final extension at 72°C for 5 minutes.

The PCR product was purified using Qiaquick, Qiagen's PCR product purification kit, according to the manufacturer's protocol except the final elution of cleaned DNA was in 100 µL of dH_2O. The eluted DNA was lyophilized and redissolved in 17 µL of dH_2O. The purified fragment was digested with XbaI in a 20 µL reaction: 17 µL purified PCR product, 1X reaction buffer (MBI/Fermentas' Y+/Tango buffer), and 10 units of XbaI (MBI/Fermentas). The mixture was incubated at 37°C for 1 hour and 45 minutes.

The reaction was cleaned using a Qiaquick column according to the manufacturer's protocol except for elution in 100 µL of dH_2O. The DNA was lyophilized and redissolved in 17 µL of dH_2O. The DNA fragment was BamHI digested in a 20 µL reaction: 17 µL purified DNA, 1X reaction buffer (MBI/Fermentas' G+ buffer), and 10 units of BamHI (MBI/Fermentas). The mixture was incubated at 37°C for 1 hour and 45 minutes. Again the digestion was cleaned using Qiaquick with a 100 µL elution in dH_2O. The DNA was quantitated by electrophoresis of an aliquot in a 1% agarose gel. The PCR product was 1 kb as expected.

Approximately 3 µg of the pRD526 plasmid vector was digested with XbaI and purified using Qiaquick according to the manufacturer's protocol except the final elution was with 100 µL of dH_2O. The DNA was lyophilized and redissolved in 17 µL of dH_2O. The plasmid DNA was digested with BamHI, purified in the same way, and quantitated by ethidium bromide staining after electrophoresis in a 0.7% agarose gel.

The digested vector and insert dissolved in dH_2O were mixed together in a roughly 1:3 molar ratio of vector to insert: 2.6 µg of vector + 0.7 µg of PCR product. The mixture was lyophilized and redissolved in 4 µL of dH_2O. The 5 µL ligation reaction included DNA, 1X reaction buffer (0.5 µL of MBI/Fermentas T4 DNA ligase buffer), and 2.5 Weiss units of T4 DNA ligase (0.5 µL of MBI/Fermentas T4 DNA ligase). The mixture was incubated overnight at 16°C.

The ligations were transformed into chemically competent *E. coli*, strain XL1-Blue, heat shocked at 37°C in the presence of the ligation products, and plated on LB-kanamycin plates. Numerous transformants resulted. Forty eight transformants were screened for the presence of the *B2* gene insert. The plasmid was isolated with Qiagen minipreps according to the manufacturer's protocol regarding the isolation of large plasmids. PCR screening was done in 50 µL reactions which contained 4 µL of plasmid, 0.5 µM of each of the primers B2Cormack-Full-For2 and B2Cormack-Full-Rev2, 0.2 mM of each dNTP, 1X reaction buffer (*Taq* DNA polymerase reaction buffer), 1.5 mM $MgCl_2$, and 2.5 units of *Taq* DNA polymerase. The cycling conditions were denaturation at 94°C for 4 minutes followed by 34 amplification cycles of 45 seconds denaturation at 94°C, 30 seconds annealing at 57.5°C, 1 minute extension at 72°C, and a final extension at 72°C for 4 minutes. Five samples had a clear band of the expected 1 kb size. The transformant represented by the brightest of these bands was confirmed by sequencing to contain the pRD526 vector with the appropriate insertion of the complete *B2* gene.

Sequencing the pRD526+*B2* Construct

More PCR product was produced using the plasmid sample corresponding to the best 1 of 5 transformants which showed a clear band of size 1 kb. Five identical 100 µL reactions were performed. Each reaction contained 5 µL of plasmid, 0.5 µM of each of the original *B2* subcloning primers, 0.2 mM of each dNTP, reaction buffer at 1X, 1.5 mM $MgCl_2$, and 5 units of *Taq* DNA polymerase (MBI/Fermentas). The cycling conditions were denaturation at 94°C for 4 minutes followed by 34 amplification cycles of 45 seconds denaturation at 94°C, 30 seconds annealing at 57.5°C, 1 minute extension at 72°C, and a final extension at 72°C for 4 minutes. The PCR products were purified using Qiaquick following the manufacturer's protocol except for elutions in 100 µL volumes of dH_2O. The eluates were pooled and concentrated by lyophilization followed by redissolution in 25 µL of dH_2O. The PCR product was sequenced at the Centre for Structural and Functional Genomics at Concordia University. Two sequencing reactions were performed, each using one of the original *B2* subcloning primers. The internal sequence of the gene was confirmed.

Additional plasmid from the same transformant was prepared for further sequencing using a Qiagen miniprep kit following the manufacturer's protocol for the isolation of large plasmids. In order to have enough DNA and in a concentrated form, 12 separate miniprep columns were used to isolate plasmid DNA from a common 20 mL LB-kanamycin culture of the clone. The elutions, which were in 100 µL volumes of dH_20, were then pooled, lyophilized, and redissolved in a smaller volume of dH_20. *B2* gene specific primers were designed to anneal to a region within the gene and sequence outwards, toward the *B2* gene/pRD526 vector junctions.

Table 3: *B2* Gene Specific Sequencing Primers

Primer name	Primer direction	Primer sequence	Vector/gene junction sequenced
B2-sequencing-rev	reverse	GCCGTGAGAGCAGATGGGTCAGCA	5′
B2-sequencing-forw	forward	GATCCATCGGCCTGGGAGGACAAG	3′

Two sequencing reactions, one per primer, were done at the Sheldon Biotechnology Centre (Montreal, Canada). The 5′ and 3′ junction points of the pRD526+*B2* construct were elucidated by the two resulting sequences.

A complete contig was constructed from these two DNA sequences and from the two sequences that resulted from the PCR product sequencing reactions. The correct insertion of the complete *B2* gene into the pRD526 vector was confirmed.

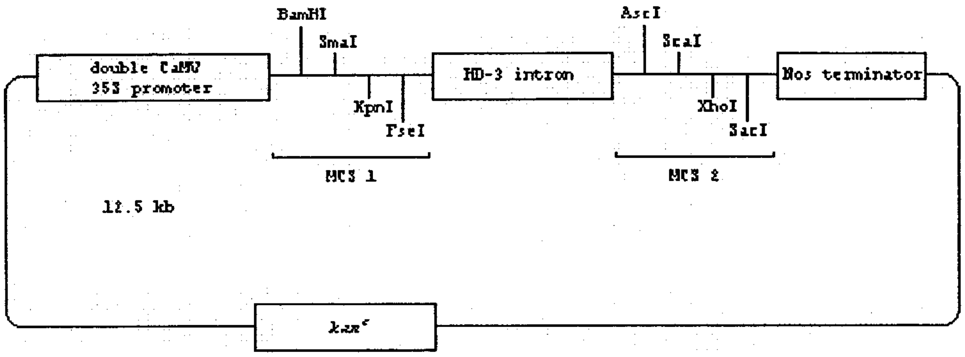
The pDARTHVECTOR vector

The plant gene silencing vector, pDARTHVECTOR, produces a self-annealing hairpin RNA in transformed plants. Two copies of a fragment of the gene are cloned in opposite orientation in the vector's two multiple cloning sites which are separated by an intron region. The sense and antisense sequences within the same RNA molecule lead to a hairpin-like structure with a region of double stranded RNA. Such a structure is effective at gene silencing (Wesley *et al.*, 2001).

pDARTHVECTOR is used for *Agrobacterium tumefaciens* mediated plant transfection and the hairpin-like RNA is expressed under the control of the strong CaMV double 35S promoter.

pDARTHVECTOR originates from the kanamycin resistant pBin19 plasmid. pBin19's CaMV 35S promoter was doubled and an insert consisting of two multiple cloning sites separated by an intron was added to its multiple cloning site. Thus, in the upstream to downstream direction, pDARTHVECTOR's hairpin-RNA forming region consists of a double CaMV 35S promoter, multiple cloning site 1, an intron region, multiple cloning site 2, and a Nos terminator. The multiple cloning site 1 contains the restriction enzyme sites BamHI, SmaI, KpnI, and FseI and site 2, the restriction enzyme sites AscI, ScaI, XhoI, and SacI.

Map of pDARTHVECTOR



Cloning *B2* into pDARTHVECTOR

A forward and an inverted copy of a fragment of the potato *B2* gene were cloned into pDARTHVECTOR's two cloning sites. Two sets of PCR primers incorporating unique restriction sites were designed to PCR amplify the desired *B2* gene section for insertion into pDARTHVECTOR's two multiple cloning sites.

Table 4: PCR Primers for Cloning a Fragment of *B2* into Multiple Cloning Sites 1 and 2 of pDARTHVECTOR

Multiple cloning site	Primer name	Primer direction	Primer sequence	Unique restriction site present	For cloning *B2* gene fragment[1]
1	B2-forw-a	forward	CATCggatccGCAGTTCAGTGACCAGCTT	BamHI (G/GATCC)	241 - 589
	B2-rev-b	reverse	CTGCggtaccATTAGTACCCTTACTATAAC	KpnI (GGTAC/C)	
2	B2-forw-y-new	forward	AATCAActcgagTTCTGGCAGTTCAGTGACCA	XhoI (C/TCGAG)	236 - 591
	B2-rev-z-new	reverse	GGTACCTAggcgcgccTTATTAGTACCCTTAC	AscI (GG/CGCGCC)	

[1] The numbers refer to the nucleotide numbers of the complete *B2* gene sequence (figure 2). The annealing sites for these primers are indicated on the *B2* gene sequence.

The 100 µL PCR reactions contained 300 ng of pBluescriptSK$^-$ plasmid DNA containing the full length *B2* gene, 0.3 µM of each primer of the given primer pair, 0.2 mM of each dNTP, 1X reaction buffer (MBI/Fermentas *Pfu* DNA polymerase reaction buffer), and 2.5 units of *Pfu* DNA polymerase. The cycling conditions were denaturation at 95°C for 1.5 minutes followed by 35 amplification cycles of 45 seconds denaturation at 95°C, 45 seconds annealing at 61.1°C, 3 minutes extension at 72°C, and a final extension at 72°C for 5 minutes.

Each of the two PCR products was purified using Qiaquick according to the manufacturer's protocol except the final elutions of cleaned DNA were in 100 µL volumes of dH_2O. The eluted DNA samples were lyophilized and redissolved in 17 µL volumes of dH_2O. The purified DNA was then digested in 20 µL reactions. The site 1 amplicon was digested with KpnI: 17 µL of PCR product, 1X reaction buffer (MBI/Fermentas' KpnI+ reaction buffer), and 10 units of KpnI (MBI/Fermentas). The mixture was incubated at 37°C for 1 hour and 45 minutes. The reaction was cleaned and the eluted DNA lyophilized and redissolved in dH_2O, as was done for the PCR products. The 17 µL DNA sample was then digested in another 20 µL reaction, as described previously, using BamHI (MBI/Fermentas) with the G+

reaction buffer at 1X. Again the digestion was cleaned and eluted in 100 µL of dH_2O. The amplicon for pDARTHVECTOR's site 2 was similarly prepared by XhoI and AscI digestion. The DNA samples were quantitated by ethidium bromide staining after electrophoresis in a 1% agarose gel. Bands of approximately 0.37 kb were observed as expected.

pDARTHVECTOR's site 1 was digested with the KpnI and BamHI restriction enzymes and purified. Approximately 3 µg of pDARTHVECTOR was digested with KpnI, Qiaquick purified, and the eluted DNA lyophilized and redissolved in 17 µL of dH_2O. The plasmid DNA was then digested with BamHI, purified in the same way, and quantitated by ethidium bromide staining after electrophoresis in a 0.7% agarose gel.

The site 1 digested vector and site 1 insert dissolved in dH_2O were mixed together in a roughly 1:3 molar ratio of vector to insert - 0.51 µg of vector + 0.047 µg of insert. The mixture was lyophilized and redissolved in 4 µL of dH_2O. The 5 µL ligation included DNA, 1X reaction buffer (MBI/Fermentas T4 DNA ligase reaction buffer), and 2.5 Weiss units of T4 DNA ligase (0.5 µL of MBI/Fermentas T4 DNA ligase). The mixture was incubated overnight at 16°C.

The ligations were transformed into chemically competent *E. coli*, strain XL1-Blue, heat shocked at 37°C in the presence of the ligation products, and plated on LB-kanamycin plates. Thirty seven transformants resulted. Eight transformants were screened for the presence of the *B2* gene fragment insert. The plasmid was isolated with a Qiagen miniprep kit according to the manufacturer's protocol regarding the isolation of large plasmids. PCR screening was done in 100 µL reactions which contained 1 µL of plasmid, 0.5 µM of a custom designed pDARTHVECTOR forward primer which anneals to a unique segment of the vector's double 35S promoter (AGACCCTTCCTCTATATAAGGAAGTTC, "pDV-35Spromoter-forw-S1"), 0.5 µM of "B2-rev-b", 0.2 mM of each dNTP, *Taq* DNA polymerase reaction buffer at 1X, 1.5 mM $MgCl_2$, and 5 units of *Taq* DNA polymerase. The cycling conditions were denaturation at 94°C for 4 minutes followed by 34 amplification cycles of 45 seconds denaturation at 94°C, 30 seconds annealing at 62°C, 30 seconds extension at 72°C, and a final extension at 72°C for 4 minutes. Three samples had a clear band of the expected 0.4 kb size.

One of these three transformants testing positive for a site 1 insertion became the starting material for a site 2 insertion. Five Qiagen plasmid preps were pooled, lyophilized, and redissolved in 17 µL of dH_2O.

To insert the inverted amplicon into site 2, the vector was digested with XhoI, purified and redissolved in 17 µL of dH_2O, digested with AscI, purified, and quantitated by ethidium bromide staining after electrophoresis in an agarose gel. The digested vector and insert dissolved in dH_2O were mixed together in a roughly 1:3 molar ratio of vector to insert: 0.083 µg of vector + 0.082 µg of PCR product. The mixture was lyophilized and redissolved in 4 µL of dH_2O for a 5 µL ligation reaction. After ligation and transformation into *E. coli*, strain XL1-Blue, numerous transformants resulted.

Twelve transformants were screened for the presence of the site 2 *B2* gene fragment insert. The plasmid was isolated with a Qiagen miniprep kit according to the manufacturer's protocol regarding the isolation of large plasmids. PCR screening was done in 100 µL reactions which contained 2 µL plasmid, 0.5 µM of a custom designed pDARTHVECTOR forward primer which anneals to the vector's intron region which is upstream of site 2 (GCTCCCTTTTGTGGTTGATTTAGATGG, "pDV-HD-3-forw-S2"), 0.5 µM of a custom designed pDARTHVECTOR reverse primer which anneals to the vector's NosTerminator region which is downstream of site 2 (CGCAAGACCGGCAACAGGATTCAATCT, "pDV-NosTer-rev-S2"), 0.2 mM of each dNTP, *Taq* DNA polymerase reaction buffer at 1X, 1.5 mM $MgCl_2$, and 5 units of *Taq* DNA polymerase. The cycling conditions were the same as those for the site 1 screening except, in this case, the annealing temperature was 64.4°C. Five samples had a clear band of the expected 0.5 kb size (384 bp *B2* fragment + 150 bp vector sequence).

Sequencing the pDARTHVECTOR+2times*B2* Construct

Sequencing further confirmed the presence of proper inserts in sites 1 and 2, for 1 of these 5 transformants. Two sets of PCR reactions were performed generating DNA fragments covering sites 1 and 2, respectively. Each of the PCR products was sequenced. The primers used for the PCR and sequencing reactions are summarized in table 5 below.

Table 5: Primers Used for the PCR Amplification and Sequencing of pDARTHVECTOR+2times*B2*

PCR amplified region of pDARTHVECTOR+2times*B2* construct	PCR primer name	PCR primer direction	PCR primer sequence	Primers used for sequencing
MCS 1[a]	pDV-35Spromoter-forw-S1	forward	AGACCCTTCCTCTATATAAGGAAGTTC	pDV-HD-3-rev-S1
	pDV-HD-3-rev-S1	reverse	GGGCCAGACCACAAGGGTCTATTAGTC	
MCS 2[b]	pDV-HD-3-sequencing-forw-S2	forward	GCTGGTACTGGTGATCAAATGGCTCAA	pDV-HD-3-sequencing-forw-S2
	pDV-POSTNosTer-rev-S2	reverse	TGTTTGATGGTGGTTCCGAAATCGGCA	

[a] The PCR fragment generated extended from the vector's CaMV double 35S promoter region, through its multiple cloning site 1 (plus insert), to its intron region.

[b] The PCR fragment generated extended from the vector's intron region, through its multiple cloning site 2 (plus insert), to a region of the vector downstream of its Nos terminator.

Twelve identical PCR reactions were performed for each of the two amplifications of multiple cloning site plus insert. The PCR products were Qiaquick purified with elutions in 100 µL volumes of dH_2O. Each site's purified PCR products were pooled, lyophilized, and redissolved in smaller volumes of dH_2O. Each amplicon, along with the appropriate primer, was sent to be sequenced at Sheldon Biotechnology Centre. The sequencing results confirmed the correct insertions of the *B2* gene fragments into the pDARTHVECTOR plasmid.

Transfection of the constructs into *Agrobacterium tumefaciens*

The constructs pRD526+*B2* and pDARTHVECTOR+2times*B2* were each transformed into electroporation competent *Agrobacterium tumefaciens*, strain LBA4404 using the protocol described in Dashek (1997).

RESULTS

SV7 and SV8 Sequence Quality and Cluster / Contig Information

Two cDNA libraries from pathogen challenged potato tissue were enriched by virtual subtraction and used for EST sequencing. The SV7 cDNA set was selected from a cDNA library made from *Phytophthora infestans* infected Kennebec potato leaf tissue and the SV8 cDNA set was selected from a library made from *Phytophthora infestans* infected Kennebec potato tuber tissue. After virtual subtraction of the SV7 library, a total of 2,496 EST sequences were generated. Of these, 2,282 (91.4%) are considered to be of high quality, 35 (1.4%) of marginal quality, and 177 (7.1%) of unacceptable quality. After virtual subtraction of the SV8 library, a total of 2,299 EST sequences were generated with 2,169 (94.3%) considered to be of high quality, 18 (0.8%) of marginal quality, and 112 (4.9%) of unacceptable quality. Sequence quality was determined by the program Phred which reads DNA sequencer trace data, calls bases, and assigns quality values to the individual bases. The highest of 5 consecutive averages take over a 20 bp window is used to represent the overall quality value for a given sequence.

Relative sequence identity was used to cluster and contig EST sequences with the SV7 and SV8 EST sets and within the combined set of ESTs. The results are summarized in table 6 below.

Table 6: Cluster/Contig Breakdown of SV7/SV8 EST Sequences

	in SV7	in SV8	in SV7 & SV8 combined
total number of sequences	2,496	2,299	4,795
number of sequences in clusters	477	616	1,490
number of singletons	2,019	1,683	3,305
number of clusters	210	248	604
number of contigs	310	332	879

Sequences that cluster together have at least 96% nucleotide sequence identity over a 50 bp window and likely represent closely related members of gene families. Sequences within the same contig have a greater similarity to one another because only the highest quality regions in common between reads are anchored together to assemble a contig. Contig assembly is conservative and may split sequences from the same gene into separate contigs if sequences have errors. For several contigs, alignment comparisons between the sequences within the contig were made using the BLAST 2 sequences program available at NCBI. The alignments ranged from 96-100% identity and normally over regions much longer than the 50 bp window of sequence similarity required for cluster formation.

Evaluation of the SV7/SV8 Virtual Subtractions for Transcription Factor Enrichment

The combined collection of SV7 and SV8 ESTs was analyzed for the presence of the 11 major transcription factor gene families of *Arabidopsis thaliana* (Riechmann & Ratcliffe, 2000). The results are summarized in table 7 below.

Table 7: SV7/SV8 Potato EST Sequences with High Similarity to Known Transcription Factors

Gene family	Match number	EST identifier	Annotation	Score	E value
WRKY	1	SV8-16-F6	WRKY DNA binding protein [Solanum tuberosum]	106	7.0E-25
	2	SV8-6-E5	WRKY transcription factor 71 [Arabidopsis thaliana]	105	2.0E-24
	3	SV7-10-E7	thermal hysteresis protein STHP-64 [Solanum dulcamara] ; DNA-binding protein 1 [Nicotiana tabacum] ; WRKY transcription factor 20 [Arabidopsis thaliana]	98	2.0E-22
	4	SV7-15-H5	WRKY transcription factor 35 [Arabidopsis thaliana]	73	1.0E-14
	5	SV8-2-D4	thermal hysteresis protein STHP-64 [Solanum dulcamara]; DNA-binding protein 1 [Nicotiana tabacum]; ZAP1 [Arabidopsis thaliana]	72	1.0E-14
	6	SV7-3-F8	WRKY transcription factor 69 [Arabidopsis thaliana] ; AR411 [Arabidopsis thaliana] ; somatic embryogenesis related protein [Dactylis glomerata]	70	1.0E-13
	7	SV7-12-C11	DNA-binding protein 4 [Nicotiana tabacum]	63	7.0E-12
	8	SV7-7-B9	putative WRKY-type DNA binding protein [Arabidopsis thaliana]	61	3.0E-11
	9	SV7-17-D5	WRKY DNA-binding protein 53 [Arabidopsis thaliana]	57	4.0E-10
	0.19% of the combined SV7/SV8 4,795 EST database				
MYB	1	SV8-20-A1	myb family transcription factor [Arabidopsis thaliana]	154	5.0E-39
	2	SV8-23-D4	myb-related transcription factor [Lycopersicon esculentum]	152	1.0E-38
	3	SV7-20-D7	myb-related transcription activator, putative [Arabidopsis thaliana]	147	1.0E-36
	4	SV8-10-E6	MYB-like transcription factor DIVARICATA [Antirrhinum majus] ; syringolide-induced protein 1-3-1B [Glycine max] ; I-box binding factor putative [Arabidopsis thaliana]	115	6.0E-27
	5	SV7-25-F10	I-box binding factor, Myb-related transcription activator [Lycopersicon esculentum]	114	1.0E-26
	6	SV7-2-G10	I-box binding factor [Lycopersicon esculentum]	114	1.0E-26
	7	SV8-20-A3	tuber-specific and sucrose-responsive element binding factor, myb-related protein transcription factor [Solanum tuberosum]	110	9.0E-26
	8	SV7-2-A4	MYB-family transcription factor, putative [Arabidopsis; I-box binding factor - like protein [Arabidopsis	108	4.0E-25
	9	SV7-23-E4	myb-related protein, MYB-family transcription factor, I-box	60	1.0E-18

			binding factor - like [Arabidopsis thaliana]		
	0.19% of the combined SV7/SV8 4,795 EST database				
AP2/ EREBP	1	SV8-10-E7	AP2 domain containing protein transcription factor, transcription factor TINY [Arabidopsis thaliana]	167	5.0E-43
	2	SV7-3-A8	apetala2 domain TINY like protein [Arabidopsis	91	6.0E-20
	3	SV8-5-A12	transcription factor TINY [Arabidopsis thaliana]	90	7.0E-20
	4	SV7-2-E8	TINY-like AP2 domain transcription factor [Arabidopsis	89	2.0E-19
	5	SV8-17-G9	ethylene responsive element binding protein [Fagus sylvatica]	82	2.0E-17
	6	SV7-10-E10	Avr9/Cf-9 rapidly elicited protein 111B [Nicotiana tabacum] ; DRE/CRT-binding protein DREB1A, putative ethylene responsive element binding factor [Arabidopsis thaliana]	82	3.0E-17
	7	SV8-9-A5	transcription factor JERF1 [Lycopersicon esculentum]	79	1.0E-16
	8	SV7-3-H2	Avr9/Cf-9 rapidly elicited protein 111B [Nicotiana tabacum] ; transcriptional activator CBF1 DRE/CRT-binding protein DREB1A [Arabidopsis thaliana]	78	4.0E-16
	9	SV8-8-G1	ethylene responsive element binding protein [Fagus sylvatica]; putative Ckc2 [Arabidopsis thaliana]	76	1.0E-15
	10	SV7-23-E9	DNA binding protein EREBP-3, ethylene-responsive element binding [Nicotiana tabacum] ; AP2 domain containing protein RAP2.5 [Arabidopsis thaliana]	73	9.0E-15
	11	SV8-12-C6	AP2 domain-containing transcription factor, DNA binding protein, ethylene responsive element binding protein [Nicotiana tabacum]	71	5.0E-14
	12	SV8-21-A1	DNA binding protein EREBP-4; ethylene-responsive element binding factor [Nicotiana sylvestris]	69	4.0E-13
	13	SV8-7-F9	U5 snRNP-specific 40 kDa protein, splicing factor [Homo sapiens]	60	6.0E-11
	14	SV8-22-B11	PATHOGENESIS-RELATED GENES TRANSCRIPTIONAL ACTIVATOR PTI6[L.esculentum] ; AP2 domain containing protein, ethylene responsive element, DNA-binding protein [P.armeniaca]	58	3.0E-10
	15	SV7-25-F7	ethylene responsive element binding protein, AP2 domain containing protein [Prunus armeniaca]	47	3.0E-10
	0.31% of the combined SV7/SV8 4,795 EST database				
NAC	1	SV7-20-H2	putative NAC domain protein [Solanum tuberosum]	270	1.0E-73
	2	SV7-2-E6	NAC domain protein NAC2 [Phaseolus vulgaris]	223	1.0E-59
	3	SV7-10-B7	NAC2, No apical meristem protein[Arabidopsis thaliana]	177	2.0E-45
	4	SV7-17-A1	hypothetical protein SENU5, senescence up-regulated, NAM-like protein [Lycopersicon esculentum]	171	5.0E-44
	5	SV8-17-D5	unknown protein [Arabidopsis thaliana]; NAC, putative [Arabidopsis thaliana]	171	1.0E-43
	6	SV7-5-B7	NAM (no apical meristem) (CUC2) (NAC) [Petunia x hybrida]	103	3.0E-23
	7	SV7-22-B1	hypothetical protein SENU5, senescence up-regulated [Lycopersicon esculentum]; NAM-like protein [Arabidopsis thaliana]	101	5.0E-23
	8	SV8-23-B12	nam-like protein 1 [Petunia x hybrida]	72	1.0E-13
	9	SV7-7-G5	jasmonic acid 2 [Lycopersicon esculentum] ; NAM-like protein, NAC domain protein[Arabidopsis thaliana]	64	1.0E-11
	0.19% of the combined SV7/SV8 4,795 EST database				
bHLH/ MYC	1	SV8-15-D8	phaseolin G-box binding protein PG2 [Phaseolus vulgaris]; bHLH protein - like [Arabidopsis thaliana]	241	1.0E-64
	2	SV8-10-D6	THY5 protein, TGACG-motif binding protein STF, transcription factor [Lycopersicon esculentum]	152	9.0E-39

	3	SV7-24-C3	putative protein [Arabidopsis thaliana]; putative DNA-binding protein [Arabidopsis thaliana]	100	1.0E-22
	4	SV7-25-B9	transcription factor bHLH protein, G-box binding protein [Arabidopsis thaliana]	75	1.0E-14
	5	SV8-21-F2	putative protein [Arabidopsis thaliana]; putative bHLH transcription factor [Arabidopsis thaliana]	66	2.0E-12
	6	SV8-21-H8	putative bHLH transcription factor [Arabidopsis thaliana]	62	1.0E-10
	0.13% of the combined SV7/SV8 4,795 EST database				
bZIP	1	SV7-12-H3	bZIP transcription factor [Arabidopsis thaliana]	194	6.0E-51
	2	SV7-23-G6	bZIP transcription factor [Nicotiana tabacum] ; NPR1-interactor protein 1 [Lycopersicon esculentum]	175	5.0E-45
	3	SV7-11-G2	mas-binding factor MBF2 [Solanum tuberosum]	174	8.0E-45
	4	SV8-10-D6	THY5 protein, TGACG-motif binding protein STF, transcription factor [Lycopersicon esculentum]	152	9.0E-39
	0.08% of the combined SV7/SV8 4,795 EST database				
HB	1	SV7-15-E12	Homeobox-leucine zipper protein HAT22 (HD-ZIP protein 22) [Arabidopsis thaliana]	187	7.0E-49
	2	SV7-21-A9	homeobox protein HAT22, leucine zipper protein, homeodomain transcription factor [Arabidopsis thaliana]	180	1.0E-46
	0.04% of the combined SV7/SV8 4,795 EST database				
Z-C2H2	1	SV7-24-E2	zinc finger protein [Arabidopsis thaliana]	308	8.0E-85
	2	SV8-16-A10	zinc finger protein, DNA/RNA binding protein [Arabidopsis thaliana]	230	2.0E-61
	3	SV7-14-C7	zinc finger and C2 domain protein [Arabidopsis thaliana]	223	2.0E-59
	4	SV8-1-C7	zinc finger protein [Arabidopsis thaliana]	115	5.0E-41
	5	SV8-6-F2	zinc finger protein Glo3-like [Arabidopsis thaliana]	122	5.0E-29
	6	SV8-20-H12	zinc finger protein [Arabidopsis thaliana]	102	7.0E-23
	0.13% of the combined SV7/SV8 4,795 EST database				
MADS	1	SV7-26-A8	MADS transcriptional factor [Solanum tuberosum]	194	5.0E-51
	0.02% of the combined SV7/SV8 4,795 EST database				
ARF-Aux/IAA	1	SV7-14-C7	ADP-ribosylation factor 1-directed GTPase activating protein [Rat] ; zinc finger and C2 domain protein [A thaliana]...	223	2.0E-59
	2	SV8-6-F2	zinc finger protein Glo3-like [Arabidopsis thaliana]; ADP-ribosylation factor GTPase activating protein 1 [Homo sapiens]	122	5.0E-29
	3	SV7-16-H3	glycine-rich protein – Arabidopsis (5e-06)	76	4.0E-15
	0.06% of the combined SV7/SV8 4,795 EST database				
Dof	1	SV7-19-D6	Dof zinc finger protein, elicitor-responsive [Oryza sativa]	116	3.0E-27
	2	SV7-14-D1	ascorbate oxidase promoter-binding protein [Cucurbita maxima] ; H-protein promoter binding factor-2a, zinc finger protein OBP4 [Arabidopsis thaliana]	73	2.0E-17
	3	SV7-13-E8	Dof zinc finger protein [Oryza sativa]	64	1.0E-11

	4	SV7-13-F1	Dof zinc finger protein [Oryza sativa]	64	1.0E-11
	0.08% of the combined SV7/SV8 4,795 EST database				

Three amino acid sequences representing each gene family were used to tBLASTn search the combined SV7 / SV8 EST database and a non-redundant list of matches was created.

The database of 5,434 unsubtracted pathogen challenged potato cDNA library derived ESTs was searched in the same way for representatives of each of the 11 transcription factor gene families. For each transcription factor gene family, the number of matches and the percentage of ESTs that this number of matches represents are recorded in table 8 below.

Table 8: EST Frequencies for Transcription Factors in an Unenriched Potato EST Set Derived from Pathogen Challenged Potato Tissue

Transcription factor gene family	Number of EST matches	Percentage of the database represented
WRKY	3	0.06
MYB	3	0.06
AP2/EREBP	16	0.29
NAC	1	0.02
bHLH/MYC	2	0.04
bZIP	0	0
HB	2	0.04
Z-C2H2	4	0.07
MADS	3	0.06
ARF-Aux/IAA	2	0.04
Dof	0	0

A Chi-squared test was used to determine if the frequency of ESTs for the 11 transcription factors in the

SV7 / SV8 data set was significantly different from that in the unenriched potato EST data set. In most cases, the frequency of transcription factors in the SV7 / SV8 database was higher than the frequency in the unenriched EST database. The Chi-squared test tested the statistical significance of these higher frequencies:

H_0: The frequency of ESTs representing transcription factors is no different between the virtually subtracted EST data set and the unenriched EST data set.

H_1: The frequency of ESTs representing transcription factors is different between the virtually subtracted EST data set and the unenriched EST data set.

The Chi-squared test statistic was calculated using the observed numbers of matches in the SV7 / SV8 database for each transcription factor gene family and the expected values which were calculated from the frequencies of transcription factors in the unenriched EST data set. The results are summarized in table 9 below.

Table 9: Chi-squared Calculations for the Transcription Factor Gene Families

Transcription factor gene family	Observed frequency[a] (O)	Expected frequency[b] (E)	$(O - E)^2/E$
WRKY	9	2.647	15.245
MYB	9	2.647	15.245
AP2/EREBP	15	14.119	0.055
NAC	9	0.882	74.677
bHLH/MYC	6	1.765	10.164
bZIP	4	(1)[c]	9.000
HB	2	1.765	0.031
Z-C2H2	6	3.530	1.729
MADS	1	2.647	1.025
ARF-Aux/IAA	3	1.765	0.865
Dof	4	(1)[c]	9.000

$$X^2 = \sum [(O - E)^2/E]$$

$$= 137.036$$

[a]the number of tBLASTn matches in the SV7 / SV8 database

[b]the number of tBLASTn matches expected in the SV7 / SV8 database based on the frequency of transcription factors in the unenriched database

[c]In cases where no ESTs with significant matches were observed in the unenriched database, the value of 1 was used.

The critical Chi-squared value for a 0.005 level of significance with 10 degrees of freedom is 25.188. Since the calculated Chi-squared statistic of 137.036 is larger than the critical value in this one-tailed test, the null hypothesis was rejected. Thus, the higher frequencies of transcription factors observed for the SV7 / SV8 subtracted database are significantly different from those in the unenriched database. The overall frequency of ESTs with significant similarity to transcription factors in the SV7 / SV8 data set was 68 out of 4,795 ESTs (14 per 1,000 ESTs) whereas in the comparable unenriched data set, it was 36 out of 5,434 ESTs (7 per 1,000 ESTs). The greatest increases in the virtually subtracted set were for WRKY, MYB, NAC, bHLH/MYC, bZIP, and Dof-like transcription factors. Strangely, MADS family transcription factor members were at a lower frequency in the virtually subtracted data set.

The Effect of Virtual Subtraction on EST Redundancy

The SV7 / SV8 EST database was analyzed for the presence of sequences from 11 highly expressed potato genes. Consensus sequences representing highly expressed genes (Ronning *et al.*, 2003) were assembled from ESTs to constitute full length gene sequences. The frequency of ESTs derived from these genes in the SV7 / SV8 database was determined by tBLASTx searches and is summarized in table 10 below.

Table 10: Potato ESTs from the SV7/SV8 Data Set with High Similarity to Highly Expressed Genes

Abundantly transcribed genes	Match number	EST identifier	Annotation	Score	E value
Heat shock cognate protein 80	1	SV8-12-A02	heat shock protein hsp80 [Lycopersicon esculentum]	314	3.0E-86
	0.02% of the combined SV7/SV8 4,795 EST database				

Catalase (CAT1)	none				
Elongation factor 1-alpha (EF-1-α)	1	SV8-16-G10	elongation factor 1 alpha [Oryzias latipes]	314	1.0E-94
	0.02% of the combined SV7/SV8 4,795 EST database				
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	1	SV8-22-A03	Glyceraldehyde 3-phosphate dehydrogenase [Dictyostelium discoideum]	124	1.0E-57
	2	SV7-07-F07	glyceraldehyde 3-phosphate dehydrogenase [Arabidopsis thaliana]	111	2.0E-25
	0.04% of the combined SV7/SV8 4,795 EST database				
Elongation factor 1-alpha (EF-1-α)	1	SV8-16-G10	elongation factor 1 alpha [Oryzias latipes]	340	2.0E-94
	0.02% of the combined SV7/SV8 4,795 EST database				
L3 ribosomal protein	none				
DNA J protein	1	SV8-11-B06	DnaJ protein [Solanum tuberosum]	157	3.0E-43
	2	SV8-19-D11	DnaJ-like protein [Arabidopsis thaliana]	51	8.0E-12
	3	SV7-06-B03	dnaJ-like protein [Arabidopsis thaliana]	47	3.0E-10
	0.06% of the combined SV7/SV8 4,795 EST database				
S-adenosylmethionine decarboxylase **(SAMDC)**	1	SV7-05-C06	adenosylmethionine decarboxylase proenzyme[Solanum tuberosum]	365	1.0E-102
	2	SV7-08-A01	S-adenosylmethionine decarboxylase [Arabidopsis thaliana]	225	1.0E-69
	3	SV8-11-E11	S-adenosylmethionine decarboxylase [Arabidopsis thaliana]	225	5.0E-67
	4	SV7-08-B01	S-adenosylmethionine decarboxylase [Arabidopsis thaliana]	225	5.0E-67
	5	SV8-11-A07	hypothetical protein 1 [Catharanthus roseus]; S-adenosylmethionine decarboxylase [Oryza sativa]	197	2.0E-51
	0.10% of the combined SV7/SV8 4,795 EST database				
Alpha-tubulin	1	SV8-05-A05	alpha-tubulin [Nicotiana tabacum]	501	1.0E-154
	2	SV8-16-E09	beta-8 tubulin [Zea mays]	74	3.0E-29
	0.04% of the combined SV7/SV8 4,795 EST database				
Translationally controlled tumor protein (TCTP; P23)	none				

Chaperonin-60 beta chain precursor	1	SV7-06-E03	mitochondrial chaperonin hsp60 [Arabidopsis thaliana]	162	5.0E-40
	2	SV8-10-G02	mitochondrial chaperonin 60 [Cucurbita sp.]	111	3.0E-25
	0.04% of the combined SV7/SV8 4,795 EST database				

The frequency of the same 11 highly expressed genes in the unenriched 5,434 EST data set derived from pathogen challenged tissue was also determined by tBLASTx searches and is summarized in table 11 below.

Table 11: EST Frequencies for Highly Expressed Genes in an Unenriched Potato EST Set Derived from Pathogen Challenged Potato Tissue

Highly expressed genes	Number of EST matches	Percentage of the database represented
Heat shock cognate protein 80	19	0.35
Catalase (CAT1)	43	0.79
Elongation factor 1-alpha (EF-1-α)	38	0.70
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	26	0.48
Elongation factor 1-alpha (EF-1-α)	39	0.72
L3 ribosomal protein	6	0.11
DNA J protein	8	0.15
S-adenosylmethionine decarboxylase (SAMDC)	14	0.26
Alpha-tubulin	7	0.13
Translationally controlled tumor protein (TCTP; P23)	3	0.06
Chaperonin-60 beta chain precursor	11	0.20

A Chi-squared test was performed to determine if the frequency of ESTs representing the 11 highly expressed genes in the SV7 / SV8 data set was significantly different from that in the unenriched potato EST data set. For the SV7 / SV8 database, the frequencies of all of the highly expressed gene ESTs were lower as compared to those for the unenriched EST database. The Chi-squared test tested the statistical significance of these lower frequencies:

H_0: The frequency of ESTs representing highly expressed genes is no different between the virtually subtracted EST data set and the unenriched EST data set.

H_1: The frequency of ESTs representing highly expressed genes is different between the virtually subtracted EST data set and the unenriched EST data set.

The calculation of the Chi-squared test statistic used the observed numbers of matches in the SV7 / SV8 database for each highly expressed gene and the expected values which were calculated from the frequencies of highly expressed genes in the unenriched EST data set. The results are summarized in table 12 below.

Table 12: Chi-squared Calculations for the Highly Expressed Genes

Abundantly transcribed genes	Observed frequency[a] (O)	Expected frequency[b] (E)	$(O - E)^2/E$
Heat shock cognate protein 80	1	16.766	14.825
Catalase (CAT1)	0	37.944	37.944
Elongation factor 1-alpha (EF-1-α)	1	33.531	31.561
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	2	22.943	19.117
Elongation factor 1-alpha (EF-1-α)	1	34.414	32.443
L3 ribosomal protein	0	5.294	5.294
DNA J protein	3	7.059	2.334
S-adenosylmethionine decarboxylase (SAMDC)	5	12.354	4.377
Alpha-tubulin	2	6.177	2.824
Translationally controlled tumor protein (TCTP; P23)	0	2.647	2.647
Chaperonin-60 beta chain precursor	2	9.706	6.119

$$X^2 = \sum [(O - E)^2/E]$$

$$= 159.486$$

[a] the number of tBLASTx matches in the SV7 / SV8 database

[b] the number of tBLASTx matches expected in the SV7 / SV8 database based on the frequency of highly expressed genes in the unenriched database

The critical Chi-squared value for a 0.005 level of significance with 10 degrees of freedom is 25.188. Since, in this one-tailed test, the calculated Chi-squared statistic of 159.486 is larger than the critical value, the null hypothesis was rejected. Thus, the lower frequencies of ESTs representing highly expressed genes observed for the SV7 / SV8 subtracted database are significantly different from the frequencies in the unenriched database. In the SV7 / SV8 EST set, the overall frequency of ESTs with significant similarity to highly expressed genes was 17 out of 4,795 ESTs (4 per 1,000 ESTs) whereas, in the unenriched EST set, it was 214 out of 5,434 ESTs (39 per 1,000 ESTs). The most pronounced decreases in the virtually subtracted EST set were in the numbers of ESTs representing heat shock cognate protein 80, CAT1, EF-1-alpha, and GAPDH. Virtual subtraction led to a very dramatic reduction of redundancy for ESTs derived from highly expressed genes.

Percent Full Length cDNAs among the SV7/SV8 Clones

The combined SV7 / SV8 EST database at FGAS was searched using 30 full length DNA query sequences. For each search performed, the full length status of each of the cDNAs represented by a match was determined based on the alignment to the query sequence. For each group of full length DNA query sequences similar in length, an estimate of the percent full length cDNA composition of the SV7 / SV8 clones representing genes of that length was calculated using the full length cDNA statuses for all of the matches resulting from all of the searches carried out for that query sequence group. The results are summarized in table 13 below.

Table 13: Percent Full Length cDNAs among the SV7/SV8 Clones

Query sequence group[a]	Full length nucleotide sequence used to search the combined SV7/SV8 EST database[b]	Matches to the combined SV7/SV8 EST database[c]			Nucleotide where alignment begins in query	Nucleotide where alignment begins in subject	cDNA length status[d]	Percent full length cDNA composition for query sequence group[e]
		EST identifier	Score	E value				
3091 bp	Complete potato gene Genbank GI# 435002 (found by Genbank Entrez nucleotide search with "3000:4000[SLEN] AND Solanum tuberosum[Organism]") • encodes "PHA1" • query length = 3136 bp	SV7-21-D02	396	1.0E-110	2533	56	nfl	0
	Complete potato gene Genbank GI# 313348 (found by Genbank Entrez nucleotide search with "3000:4000[SLEN] AND Solanum tuberosum[Organism]") • encodes "leaf type L starch phosphorylase" • query length = 3085 bp	SV7-22-A01	589	1.0E-168	1834	17	nfl	
	Complete potato gene Genbank GI# 435000 (found by Genbank Entrez nucleotide search with "3000:4000[SLEN] AND Solanum tuberosum[Organism]") • encodes "PHA2" • query length = 3083 bp	SV8-09-B04	254	3.0E-67	2893	18	nfl	
	Complete potato gene Genbank GI# 3702676 (found by Genbank Entrez nucleotide search with "3000:4000[SLEN] AND Solanum tuberosum[Organism]") • encodes "alpha-glucan phosphorylase precursor" • query length = 3058 bp	SV7-24-C08	698	0.0	639	16	nfl	
2251 bp	TIGR TC94072, potato (found by key word searches beginning with "constitutive") • TC's ATG located via homology with complete tobacco gene Genbank GI# 19812, "luminal binding protein (blp5)" • query length = 2251 bp	SV8-04-E02	1348	0.0	1385	15	nfl	0
		SV7-10-H09	1338	0.0	1492	15	nfl	
		SV8-09-H02	825	0.0	1720	12	nfl	
		SV8-11-D01	159	1.0E-38	1740	10	nfl	

1988 bp	Complete potato gene Genbank GI# 3097270 (found by Genbank Entrez nucleotide search with "2000:2100[SLEN] AND Solanum tuberosum[Organism]") • encodes "ferrochelatase" • query length = 2031 bp	SV7-04-A09	1455	0.0	304	11	nfl	25
	Complete potato gene Genbank GI# 169537 (found by Genbank Entrez nucleotide search with "2000:2100[SLEN] AND Solanum tuberosum[Organism]") • encodes "pyrophosphate-fructose 6-phosphate 1-phosphotransferase (PFP) alpha-subunit" • query length = 1988 bp	SV7-23-B02	133	5.0E-31	1595	51	nfl	
	Complete potato gene Genbank GI# 5734586 (found by Genbank Entrez nucleotide search with "2000:2100[SLEN] AND Solanum tuberosum[Organism]") • encodes "external rotenone-insensitive NADPH dehydrogenase (ndb1)" • query length = 1985 bp	SV7-02-A12	1518	0.0	727	10	nfl	
	Complete potato gene Genbank GI# 248336 (found by direct key word search of TIGR nucleotide database annotations with "polyubiquitin") • encodes "polyubiquitin" • query length = 1949 bp	SV8-1-F11	436	1.0E-122	1	38	fl	
1482 bp	Complete potato gene Genbank GI# 3550662 (found by Genbank Entrez nucleotide search with "1500:1600[SLEN] AND Solanum tuberosum[Organism]") • encodes "w-3 desaturase" • query length = 1546 bp	SV7-04-D03	1489	0.0	286	10	nfl	50
		SV7-13-H06	1396	0.0	252	30	nfl	
		SV7-18-G04	1223	0.0	1	55	fl	
	Complete potato gene Genbank GI# 20160363 (found by Genbank Entrez nucleotide search with "1500:1600[SLEN] AND Solanum tuberosum[Organism]") • encodes "allene oxide synthase (aos2)" • query length = 1530 bp	SV8-17-C12	1203	0.0	1	81	fl	
		SV8-03-A12	80	5.0E-15	1240	32	nfl	
	TIGR TC93270, potato (found by key word searches beginning with "constitutive") • TC's ATG located via homology with complete tomato gene Genbank GI# 3668353, "ornithine decarboxylase" • query length = 1506 bp	SV7-09-G11	1110	0.0	1	100	fl	

	Complete potato gene Genbank GI# 14627127 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “fatty acid hydroperoxide lyase (hpl)” • query length = 1480 bp	SV7-16-E03	1382	0.0	615	11	nfl	
	Complete potato gene Genbank GI# 473168 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “E1 alpha subunit of pyruvate dehydrogenase (lipoamide)” • query length = 1467 bp	SV8-11-E01	1247	0.0	571	17	nfl	
		SV8-03-A04	1152	0.0	1	61	fl	
	Complete potato gene Genbank GI# 21406 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “ADP/ATP translocator” • query length = 1466 bp	SV7-17-F07	202	5.0E-52	478	36	nfl	
	Complete potato gene Genbank GI# 609269 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “beta-tubulin” • query length = 1455 bp	SV8-16-E09	335	5.0E-92	832	21	nfl	
	Complete potato gene Genbank GI# 15778631 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “G protein beta subunit 2 (GB2)” • query length = 1445 bp	SV8-24-B05	1334	0.0	1	127	fl	
		SV7-08-F08	700	0.0	1	138	fl	
	Complete potato gene Genbank GI# 21165526 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “mitogen-activated protein kinase” • query length = 1443 bp	SV7-13-B07	1402	0.0	1	94	fl	
1316 bp	Complete potato gene Genbank GI# 21484 (found by Genbank Entrez nucleotide search with “1500:1600[SLEN] AND Solanum tuberosum[Organism]”) • encodes “induced stolon tip protein” • query length = 1334 bp	SV7-05-C06	870	0.0	566	19	nfl	**67**

	TIGR TC94489, potato (found by key word searches beginning with "constitutive") • TC's ATG located via homology with complete tomato gene Genbank GI# 12002864, "JAB" • query length = 1288 bp	SV8-16-B10	1459	0.0	1	27	fl	
	Complete potato gene Genbank GI# 19913102 (found by direct key word search of TIGR nucleotide database annotations with "UDP-Glucose:protein transglucosylase OR uptg2") • encodes "UDP-Glucose:protein transglucosylase (uptg2)" • query length = 1269 bp	SV7-22-C01	739	0.0	1	6	fl	
1045 bp	TIGR TC94433, potato (found by using consensus sequence of most redundant contig #168 to search TIGR nucleotide database) • TC's ATG located via homology with complete *Arabidopsis thaliana* gene Genbank GI# 42568504, "expressed protein" • query length = 1093 bp	SV7-21-G07	1356	0.0	44	16	nfl	**78**
		SV8-05-B11	1185	0.0	1	180	fl	
		SV8-15-D11	1162	0.0	1	182	fl	
		SV8-14-H01	1049	0.0	1	171	fl	
		SV7-18-A05	1003	0.0	1	199	fl	
		SV8-13-D02	906	0.0	182	8	nfl	
		SV8-17-H06	74	2.0E-13	107	106	lfl	
	Complete potato gene Genbank GI# 1030067 (found by Genbank Entrez nucleotide search with "1000:1100[SLEN] AND Solanum tuberosum[Organism]") • encodes "isoflavone reductase homologue" • query length = 1033 bp	SV8-14-D08	1314	0.0	1	39	fl	
	Complete potato gene Genbank GI# 7141301 (found by Genbank Entrez nucleotide search with "1000:1100[SLEN] AND Solanum tuberosum[Organism]") • encodes "soluble NSF attachment protein" • query length = 1011 bp	SV7-17-E07	353	2.0E-97	43	49	lfl	
743 bp	TIGR TC92911, potato (found by direct key word search of TIGR nucleotide database annotations with "ribosome OR ribosomal") • TC's ATG located via homology with complete *Solanum brevidens* gene Genbank GI# 37625522, "60S ribosomal protein L13 (Ci-1)" • query length = 795 bp	SV7-01-F10	1088	0.0	1	49	fl	**85**
		SV8-18-A02	200	1.0E-51	418	16	nfl	

	TIGR TC96152, potato (found by direct key word search of TIGR nucleotide database annotations with "ubiquitin") • TC's ATG located via homology with complete *Arabidopsis thaliana* gene Genbank GI# 42565865, "ubiquitin-conjugating enzyme family protein" • query length = 727 bp	SV8-19-C09	573	1.0E-164	1	126	fl	
		SV8-12-A06	565	1.0E-161	1	119	fl	
		SV8-23-C03	509	1.0E-145	1	123	fl	
	TIGR TC104895, potato (found by using consensus sequence of most redundant contig #67 to search TIGR nucleotide database) • TC's ATG located via homology with complete *Arabidopsis thaliana* gene Genbank GI# 18408977, "60S ribosomal protein L26 (RPL26A)" • query length = 705 bp	SV8-03-H10	1118	0.0	1	54	fl	
		SV8-21-G12	1118	0.0	1	62	fl	
		SV8-11-B11	1110	0.0	1	32	fl	
		SV8-17-B04	1108	0.0	1	62	fl	
		SV8-01-A01	472	1.0E-133	326	22	nfl	
		SV7-02-H04	250	1.0E-66	1	48	fl	
		SV7-15-F02	212	3.0E-55	1	31	fl	
		SV8-06-A03	98	1.0E-20	1	54	fl	
551 bp	Complete potato gene Genbank GI# 633682 (...found by using consensus sequence of most redundant contig #205 to search TIGR nucleotide database. This complete potato gene was a highly statistically significant match.) • encodes "cytochrome-c reductase" • query length = 572 bp	SV8-02-G09	959	0.0	1	57	fl	100
		SV8-16-D07	944	0.0	1	81	fl	
		SV8-22-C06	908	0.0	1	48	fl	
		SV8-14-B07	908	0.0	1	46	fl	
		SV8-03-H04	904	0.0	1	64	fl	
		SV8-19-H06	904	0.0	1	72	fl	
		SV8-09-A04	896	0.0	1	76	fl	
		SV8-18-C07	894	0.0	8	19	1fl	
	TIGR TC103965, potato (found by using consensus sequence of most redundant contig #153 to search TIGR nucleotide database) • TC's ATG located via homology with complete tomato gene Genbank GI# 3850777, "glutaredoxin" • query length = 556 bp	SV8-19-D01	999	0.0	1	28	fl	
		SV8-23-A08	989	0.0	1	51	fl	
		SV8-21-D04	912	0.0	1	29	fl	
		SV8-12-H08	848	0.0	1	41	fl	
		SV8-07-A12	767	0.0	1	42	fl	
		SV7-14-E05	416	1.0E-117	1	21	fl	
	TIGR TC94064, potato (found by using consensus sequence of most redundant contig #214 to search TIGR nucleotide database) • TC's ATG located via homology with complete cotton gene Genbank GI# 1553128, "Ribosomal protein L44 isoform a" • query length = 524 bp	SV8-09-C10	952	0.0	1	60	fl	
		SV7-09-C12	948	0.0	1	54	fl	
		SV7-09-D07	932	0.0	1	55	fl	
		SV8-14-C06	874	0.0	1	69	fl	
		SV7-07-F05	841	0.0	1	48	fl	
		SV7-09-E06	694	0.0	1	58	fl	
		SV7-09-C05	545	1.0E-155	1	60	fl	
		SV8-16-B05	402	1.0E-112	1	55	fl	
		SV8-15-F03	248	3.0E-66	1	119	fl	

[a] The full length DNA sequences used to BLASTn search the SV7 / SV8 EST database were grouped into query sequence groups based on their lengths. In any group, there is no more than an approximately 100 bp difference between the longest and the shortest sequence.

[b] The full length DNA query sequences were found and used to BLASTn search the SV7 / SV8 EST database. A full length DNA sequence is defined here as extending from the ATG to the end of the 3′ UTR. This is the query length.

[c] The full length DNA sequences were used as queries in BLASTn searches of the SV7 / SV8 EST database.

[d] nfl = "not full length"
fl = "full length"
lfl = "likely full length"

[e] For each query sequence group, the percent full length cDNA composition was calculated. A "likely full length" cDNA was considered to be full length for calculation purposes. For example, the percent full length cDNA composition for the last query sequence group is 100% (23/23).

Microarray Construction

The SV7 and SV8 virtual subtractions provided the clones with which to build the potato microarray. The inserts of the purified plasmids were PCR amplified; the amplicons were purified and concentrated and then printed on microarray slides, ready for hybridization. A total of 4,416 clones were processed including 22 96-well plates from SV7 and 24 96-well plates from SV8.

Cloning *B2* into pRD526

The complete *B2* gene of potato was cloned into the pRD526 vector. The gene was PCR amplified from a previously identified cDNA clone, ligated into pRD526, and transformed into *E. coli*. *E. coli* transformants were screened by PCR and 5 out of 48 transformants showed a clear band at the expected 1 kb location. One of these was confirmed by DNA sequencing to contain the pRD526 vector with the appropriate insertion of the complete *B2* gene.

The PCR product generated by the original subcloning primers for *B2* was sequenced. Two sequencing reactions were performed, each using one of the *B2* subcloning primers. The internal sequence of the gene was confirmed. The pRD526+*B2* construct plasmid was also sequenced. Gene specific sequencing primers were used to perform two sequencing reactions, each using a primer designed to anneal to a region within the *B2* gene and sequence toward a *B2* gene / pRD526 vector junction. The 5′ and 3′ junction points of the construct were elucidated by the two resulting sequences.

The four sequences for the pRD526+*B2* construct were assembled into a complete contig. The DNA sequence showed clearly that the gene's start codon is in frame with the translation initiation ATG present in the vector, that the two restriction enzymes used to clone the gene (XbaI and BamHI) cut in the proper places in both the vector and insert, and that the cut insert was ligated correctly. Pairwise BLAST analysis against the known *B2* gene revealed that no mutations were introduced in the cloning process and that the entire open reading frame of the gene was intact. The complete sequence of the *B2* insert is presented in figure 3 below.

Figure 3: *B2* gene insertion into the pRD526 vector

GGAAACCTCCTCGGATTCCATTGCCCAGCTATCTGTCACTGTTATTGTGAAGATAGTGG
AAAAGGAAGGTGGCTCCTACAAATGCCATCATTGCGATAAAGGAAAGGCCATCGTTGAA
GATGCCTCTGCCGACAGTGGTCCCAAAGATGGACCCCCACCCACGAGGAGCATCGTGGA
AAAAGAAGACGTTCCAACCACGTCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTG
ACGTAAGGGATGACGCACAATCCCACTATCCTTCGCAAGACCCTTCCTCTATATAAGGA
AGTTCATTTCATTTGGAGAGGAGATCTTTTTATTTTTAATTTTCTTTCAAATACTTCCA
CCatgGCCt/ctaga**ATG**GAGATCAACAACAACAACAACTCAACTCCATCTTTCTGGCA
GTTCAGTGACCAGCTTCGTCTGCAGAACAACTAACTTAGCAAATCTCTCTTTGAATGAT
TCAATCTGGAGCAGTAACTATGGCTCTAAAAGGCCTGAAGAAAGAAGAAATTTTGATAT
CCAGGGTAGGTGGTGACTTCAACTCTACTGCTAATACTCTTCTAAACAAGTCAAATTAC
AATCTTTTTAGCAACGATGGCTGGAAAATTGCTGACCCATCTGCTCTCACGGCGGCGNA
ACGGCGGTGGTGCTGCCGGAAAAGGGGTACTTGGGGTTGGTTTAAATGGTGGATTCAAC
AAAGGGGTTTACTNCAAATCAAGCTTATGAACTTCAGTTATAGNTAAGGGTACTAATAA
TGTTGCATTAGGTACCAAAGGGATAAACAANNAAATTTGGATAAAGGGTTTTTTGAAGA
TGAGCATAAAAGTGTGAAGAAGAATAACAAGAGTGTTAAAGAGAGTAACAAGGATGTTA
ATAGTGAGAAACAGTATGGTGTTGATAAAAGGTTTAAGNACTTTGCCACCAGCAGAATC
TTTGCCAAGAAATGAGACGGTTGGTGGATATATTTTTGTTTGCAACANATGATACTATG
GCTGAGAATCTCAAAAGGGAGCTCTTTGGCTTGCCCCCACGTTACAGGGACTCAGTTAG
GCAAATAACACCTGGATTGCCTCTTTTTCTGTACAACTACTCGACCCATCAGCTTCACG
GAGTATTTGAGGCTGCAAGCTTTGGTGGGTCAAATATTGATCCATCGGCCTGGGAGGAC
AAGAAGAACCCTGGTGAATCTCGCTTTCTGCTCAGGTTCGTGTCGTGACAGGAAAGTCT
GTGAACCACTTGAAGAGGATTCATTCAGGCCAATCCTTCACCACTACGACGGCCCTAAA
TTCCGCCTCGAGCTAAACGTTCCAGAGGCTATTTCTCTTCTCGACATTTTTGAAGAGAA
CAAGAAC**TAA**ATg/gatccCCGATCGTTCAAACATTTGGCAATAAAGTTTCTTAAGATT
GAATCCTGTTGCCGGTCTTGCGATGATTATCATATAATTTCTGTTGAATTACGTTAAGC
ATGTAATAATTAACATGTAATGCATGACGTTATTTATGAGATGGGTTTTTATGATTAGA
GTCCCGCAATTATACATTTAATACGCGATAGAAAACAAAATATAGCGCGCAAACTAGGA
TAAATTATCGCGCGCGGTGTCATCTCATGTTACTAGATCGGGAATTCACTGGCCGTCGT
TTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCGAACTTAATCGCCTTGCAGCA
CATCCCCCTTTCGC

The sequence shows the insertion of the complete coding region of the *B2* gene into the pRD526 vector. Flanking vector sequences of 363 nucleotides and 354 nucleotides can be seen.

bold letters =-start and stop codons of the *B2* gene lower case atg = translation initiation atg present in the vector

t/ctaga = XbaI restriction enzyme target site

g/gatcc = BamHI restriction enzyme target site

forward slashes (/) = show were the *B2* gene was cut and ligated into the pRD526 vector

first underlined region = partial sequence of the B2Cormack-Full-For2 subcloning primer visible in the cloned insert

second underlined region = partial sequence of the reverse compliment of the B2Cormack-Full-Rev2 subcloning primer visible in the cloned insert

The pRD526+*B2* construct was transformed into *Agrobacterium tumefaciens*.

Cloning *B2* into pDARTHVECTOR

Parts of the *B2* gene were cloned into the pDARTHVECTOR plasmid. Each section of the *B2* gene to be cloned was PCR amplified out of a previously identified cDNA clone, digested, and ligated into a separate multiple cloning section of the pDARTHVECTOR plasmid. On two separate occasions, PCR screening was performed on minipreps of transformants in order to find transformants with an insert. The presence of an appropriately sized site 1 insert and an appropriately sized site 2 insert was confirmed by the first and second screenings, respectively.

PCR reactions were performed using one of these transformants which tested positive for *B2* gene fragment insertions into sites 1 and 2. The sequences of two PCR products provided confirmation of the correct insertions of the *B2* gene fragments into the pDARTHVECTOR plasmid. See figures 4 and 5 below.

Figure 4: The DNA Sequence for the pDARTHVECTOR Site 1 Insert

TCTATGATAAGGAAGTTCATTTCATTTGGAGAg/^{i}gatccGCAGTTCAGTGACCAGCTTC
GTCTGCACGAACAACAACTTAGCAAATGCTCTCTTNTGAATGATTCAATCTGGAGCAGT
CAACTATGGCTCTAAAAGGCCTGAAGAAAGAAGNAATTTTGATATCAGGGTAGGTGGTG
ACTTCAACTCTACTGCTAATACTTCTGTCAAACAAGTCAAATTACAATCTTTTTAGCAA
CGATCGGCTGGAAAATTAGCTGACCCATCATGCTCTCACGGCGGCGAACGGCGGTGGTG
CTGCCGGAAAAGGGGTACTTGGGGTTGGCTTAAATGGTGGATTCAACAAAGGGNTTTAC
TCCAATCAAGCTTTGAACTTCAGTTATAGTAAGGGTACTAATggtac/^{j}cGGCCGGCCAA
AAAACTGGTTTGTTCTGCCCTTCTATGNTTTTTGTTACTTCTTCTGTCAACTAGTGTNT
CAATGCACTAAGAGTNT

The sequence shows the insertion of the appropriate *B2* gene fragment into site 1 of pDARTHVECTOR. The flanking vector sequence can be seen.

/^{i} = double 35S promoter region / *B2* gene fragment junction

/^{j} = *B2* gene fragment / intron region junction

first underlined region = partial sequence of the B2-forw-a subcloning primer visible in the cloned insert

second underlined region = partial sequence of the reverse compliment of the B2-rev-b subcloning primer visible in the cloned insert

g/gatcc = BamHI restriction enzyme target site

ggtac/c = KpnI restriction enzyme target site

Figure 5: The DNA Sequence for the pDARTHVECTOR Site 2 Insert

```
TTCGGATTCGGATCACAAAACCAATATAGAGAGTATAATCTTTTTGGTAATTGCTCCCT
TTTGTGGTTGATTTAGATGGCAAAAAAgg/ᵖcgcgccTTATNNAGTACCCTTACTATAAC
TGAAGTTCAAAGCTTGATTTGAGTAAACCCCTTTGTTGAATCCACCGATTTAAACCGAA
GCCCCAAGTACCCCTTTTCCGAGCAGCACCAGCGCGCCGTTCGCCGCCGTGAGAGCAGA
TGGGTCACGCAATATATTCCAGCCATCGTGGCTAGGAAAGATTGTCANATTNGACTTGT
GTGAAGAAAGTATTAAGCAGTAAGAGTNTGAAGTCACCACCTACCCTGATATCAAGAGA
TTTCTTCTTTCTTCAGGCCTTNGAGAGCCGATAGTATACGNGCGTCCAGATTGAGATCA
TTCAAAGAGAGATTCTGCTCAAGTTGTTGTTCTAGCAGNACGAAGCTGGTACACTGAAC
TGCCAGAAc/ᑫtcgagAGCTCGAATTTCCCCGATCGTTCAAACATTNGGCAATAAAGTTT
CTTAAGATTGAATCCTGTGCCGGTCTTGCGANGATTATCATATAATTCTGTTGAATTAC
GTTAGCATGTAATAATTAAGCATGTAATGCATGAGCGTTNTTTATGAGANAGGGTATTT
ATGATTAGAGTCCCGGGAATTATACGATTTGATACGCGATTGAAAACACAAATATAGGC
GCGCAAAAGCTAGGGATACACATTCATCGCTGCCGCGGGTGCTCAGTACTAATGTGTCA
CTAAGATCCGGGGAATTCCATCTTGGGTCCTGTCCGGAATTAACAAACCGNTTCCGTTG
GACCTGGGGCAACNACCTGTGGCAGTTTAACCCATGTTGAATCCGGCCTTGNTGCAAGG
GCTATTTCCCGCGTTTTCGGGCCAAAG
```

The sequence shows the insertion of the appropriate *B2* gene fragment into site 2 of pDARTHVECTOR. The flanking vector sequence can be seen.

/^{p} = intron region / *B2* gene fragment junction

/^{q} = *B2* gene fragment / Nos terminator region junction

first underlined region = partial sequence of the B2-rev-z-new subcloning primer visible in the cloned insert

second underlined region = partial sequence of the reverse compliment of the B2-forw-y-new subcloning primer visible in the cloned insert

gg/cgcgcc = AscI restriction enzyme target site

c/tcgag = XhoI restriction enzyme target site

After excluding vector sequence and restriction enzyme target sites from both sequences, the sequences were BLASTed against each other. The *B2* gene fragments were found to be inserted into the vector with correct orientation in terms of one another. Thus, the mRNA that will eventually be produced in the transformed plant should anneal to itself, forming the hairpin RNA structure. Effective gene silencing should result since the sense and antisense arms of the hairpin would be of adequate length (Wesley *et al.*, 2001).

The pDARTHVECTOR+2times*B2* construct was transformed into *Agrobacterium tumefaciens*.

DISCUSSION

Employed using an existing cDNA library, virtual subtraction is a normalization technique capable of decreasing the representation of highly expressed genes among the clones of a cDNA library chosen for sequencing. The choice of which cDNAs of highly expressed genes to be excluded from sequencing depends on the selection of the cDNAs used to make the probes for hybridization against the randomly picked and arrayed library cDNAs. In the present work, the probes for each virtual subtraction were made from cDNAs originating from both infected and uninfected tissue. As such, following hybridization, cDNAs representing genes highly transcribed following pathogen challenge, e.g. the well characterized *PR*-genes, would not have been chosen to be sequenced and cDNAs representing genes normally transcribed at high levels in uninfected tissue, such as housekeeping genes, would have been similarly excluded. Consequently, the cDNA libraries would have been enriched for genes expressed at low levels, a set of genes containing possible candidates for involvement in potato disease response signalling. Unlike early subtractive cDNA hybridization-based methods which had limited success at enriching for low abundance transcripts

(Duguin & Dinauer, 1990; Hara *et al.*, 1991; Hendrick *et al.*, 1984), the virtual subtractions of the present work were successful at doing so.

The virtual subtraction enrichment of the pathogen challenged potato leaf and tuber cDNA libraries was successful at enriching for genes expressed at low levels based on the frequency of clones with high similarity to the transcription factors of the 11 *Arabidopsis thaliana* transcription factor gene families. The frequencies of representatives of these gene families among the 4,795 subtracted SV7 / SV8 ESTs were nearly twice as great as a comparable set of unenriched ESTs. In addition, the frequency of clones from genes with abundant transcripts was substantially reduced. There were significantly lower numbers of ESTs representing highly expressed genes among the subtracted ESTs as compared to an unenriched EST set.

The estimates of the percent full length cDNA clones for the different sized genes represented by the query sequence groups indicate that for the SV7 / SV8 cDNA set, virtually all clones identified for genes of 551 bp and smaller were full length cDNAs. As expected, for larger genes, the percentage of full length cDNAs represented in the clone set diminishes. Hence, the percentages of full length clones for the different genes represented are 100%

(551 bp), 85% (743 bp), 78% (1045 bp), 67% (1316 bp), 50% (1482 bp), 25% (1988 bp), 0% (2251 bp), and 0% (3091 bp).

A microarray was constructed from the cDNA clones of two virtually subtracted *Phytophthora infestans* challenged potato tissue cDNA libraries. The microarray can be used in future analyses of the gene expression patterns of the disease response of potato and for the elucidation of the function of unknown genes such as *B2*. Through microarray based studies involving transgenic plants overexpressing *B2* and transgenic plants silencing *B2*'s expression, interactions between *B2* and other disease response genes could be discovered.

REFERENCES

Aharoni, A. and Vorst, O. (2001). DNA microarrays for functional plant genomics. Plant Molecular Biology 48: 99-118.

Asamizu, E., Nakamura, Y., Sato, S., Tabata, S. (2000). A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. DNA research 7:175-180.

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

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. Current Opinion in Plant Biology 2: 96-103.

Bonaldo, M.F., Lennon, G., and Soares, M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Research, 6: 791-806.

Boyle, B. and Brisson, N. (2001). Repression of the defense gene *PR-10a* by the single stranded DNA bonding protein SEBF. The Plant Cell 13: 2525-2537.

Burke, J., Davison, D., and Hide, W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res. 9: 1135-1142.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M. , Konno, H., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. (2000). Normalization and Subtraction of Cap-Trapper-Selected cDNAs to Prepare Full-Length cDNA Libraries for Rapid Discovery of New Genes. Genome Research 10: 1617-1630.

Chen, W., Provart, N.J., Glazebrook, J., Katagiri, F., Chang, H., Eulgem, T., Mauch, F., Luan, S., Zou, G., Whitham, S.A., Budworth, P.R., Tao, Y., Xie, Z., Chen, X., Lam, S., Kreps, J.A., Harper, J.F., Si-Ammour, A., Mauch-Mani, B., Heinlein, M., Kobayashi, K., Hohn, T., Dangl, J.L., Wang, X., and Zhu, T. (2002). Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses. Plant Cell 14: 559-574.

Clarke, B., Lambrecht, M., and Rhee, S.Y. (Epub 2002). *Arabidopsis* genomic information for interpreting wheat EST sequences. Funct Integr Genomics. 2003, 3(1-2):33-38.

Collinge, M. and Boller, T. (2001). Differential induction of two potato genes, *Stprx2* and *StNAC*, in response to infection by Phytophthora infestans and to wounding. Plant Molecular Biology 46(5): 521-529.

Constabel, C.P. and Brisson, N. (1992). The defense-related *STH-2* gene product of potato shows race-specific accumulation after inoculation with low concentrations of *Phytophthora infestans* zoospores. Planta 188: 289-295.

Crookshanks, M., Emmersen, J., Welinder, K.G., and Nielsen, K.L. (2001). The potato tuber transcriptome: analysis of 6077 expressed sequence tags. FEBS Letters 506: 123-126.

Dashek, W.V. (1997). Methods in Plant Biochemistry and Molecular Biology. Boca Raton, Florida: CRC Press.

Datla, R.S.S., Bekkaoui, F., Hammerlindl, J.K., Pilate, G., Dunstan, D.I., and Crosby, W.L. (1993). Improved high-level constitutive foreign gene expression in plants using an AMV RNA4 untranslated leader sequence. Plant Science 94: 139-149.

Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D., Siebert, P.D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. U S A. 93: 6025-6030.

Duguid, J.R. and Dinauer, M.C. (1990). Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. Nucleic Acids Research. 18: 2789-2792.

Ewing, R., Poirot, O., and Claverie, J.M. (1999-2000). Comparative analysis of the *Arabidopsis* and rice expressed sequence tag (EST) sets. In Silico Biol. 1(4): 197-213.

Hara,E., Kato, T., Nakada, S., Sekiya, S., and Oda, K. (1991). Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. Nucleic Acids Research. 19: 7097-7104.

Hendrick, S.M., Cohen, D.I., Nielsen, E.A., and Davis, M.M. (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature 308: 149-53.

Kehoe, D.M., Villand, P., and Somerville, S. (1999). DNA microarrays for studies of higher plants and other photosynthetic organisms. Trends in Plant Science. 4(1): 38-41.

Li, Z. and Thomas, T.L. (1998). *PEI1*, an Embryo-Specific Zinc Finger Protein Gene Required for Heart-Stage Embryo Formation in Arabidopsis. Plant Cell. 10: 383-398.

Ma, H.M., Schulze, S., Lee, S., Yang, M., Mirkov, E., Irvine, J., Moore, P., and Paterson, A. (2004). An EST survey of the sugarcane transcriptome. Theor. Appl. Genet. 108: 851-863.

Marineau, C., Matton, D.P., and Brisson, N. (1987). Differential accumulation of potato mRNAs during the hypersensitive response induced by arachidonic acid elicitor. Plant Mol. Biol. 9: 335-342.

Matton, D.P. and Brisson, N. (1989). Cloning, expression, and sequence conservation of pathogenesis-related gene transcripts of potato. Mol. Plant-Microbe Interact. 2: 325-331.

Mayer, K. and Mewes, H. (2001). How can we deliver the large plant genomes? Strategies and perspectives. Current Opinion in Plant Biology 5: 173-177.

Ohlrogge, J. and Benning, C. (2000). Unravelling plant metabolism by EST analysis. Current Opinion in Plant Biology 3: 224-228.

Richmond, T. and Somerville, S. (2000). Chasing the dream: plant EST microarrays. Current Opinion in Plant Biology 3: 108-116.

Riechmann, J.L. and Ratcliffe, O.J. (2000). A genomic perspective on plant transcription factors. Current Opinion in Plant Biology 3: 423-434.

Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., Bougri, O., Hart, A.L., Utterbach, T.R., Vanaken, S.E., Riedmuller, S.B., White, J.A., Cho, J., Pertea, G.M., Lee, Y., Karamycheva, S., Sultana, R., Tsai, J., Quackenbush, J., Griffiths, H.M., Restrepo, S., Smart, C.D., Fry, W.E., Van der Hoeven, R., Tanksley, S., Zhang, P., Jin, H., Yamamoto, M.L., Baker, B.J., and Buell, C.R. (2003). Comparative analyses of potato expressed sequence tag libraries. Plant Physiology 131: 419-429.

Rounsley, S.D., Glodek, A., Sutton, G., Adams, M.D., Somerville, C.R., Venter, J.C., and Kerlavage, A.R. (1996). The construction of Arabidopsis expressed sequence tag assemblies. A new resource to facilitate gene identification. Plant Physiology 112: 1177-1183.

Schenk, P.M., Kazan, K., Wilson, I., Anderson, J.P., Richmond, T., Somerville, S.C., and Manners, J.M. (2000). Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc. Natl. Acad. Sci. USA 97: 11655-11660.

Schweinfest, C.W., Henderson, K.W., Gu, J., Kottaridis, S.D., Besbeas, S., Panotopoulou, E., and Papas, T.S. (1990). Subtractive Hybridization cDNA Libraries From Colon Carcinoma and Hepatic Cancer. Genet. Annal. Techn. Appl. 7: 64-70.

Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Caminci, 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.

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Taji, T., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y., and Shinozaki, K. (2002). Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J. 31(3): 279-292.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91(20): 9228-9232.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T., Boerjan, W., Gustafsson, P., Uhlen, M., Sundberg, B., and Lundeberg, J. (1998). Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. USA 95: 13330-13335.

Strittmatter, G., Gheysen G., Gianinazzi-Pearson V., Hahn K., Niebel A., Rohde W.,and Tacke E. (1996). Infections with various types of organisms stimulate transcription from a short promoter fragment of the potato *gst1* gene. Mol Plant-Microbe Interact. 9(1): 68-73.

Wan, J., Dunning, F.M., and Bent, A.F. (2002). Probing plant-pathogen interactions and downstream defense signaling using DNA microarrays. Funct. Integr. Genomics 2: 259-273.

Wesley, S.V., Helliwell, C.A., Smith, N.A., Wang, M., Rouse, D.T., Liu, Q., Gooding, P.S., Singh, S.P., Abbott, D., Stoutjesdijk, P.A., Robinson, S.P., Gleave, A.P., Green, A.G., and Waterhouse, P.M. (2001). Construct design for efficient, effective, and high-throughput gene silencing in plants. The Plant Journal 27(6): 581-590.

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

Yoshioka, H., Yamada, N., and Doke, N. (1999). cDNA cloning of sesquiterpene cyclase and squalene synthase, and expression of the genes in potato tuber infected with *Phytophthora infestans*. Plant Cell Physiology 40(9): 993-998.