Establishing regulatory networks in *Arabidopsis*: Integrating AGRIS with the Identification of direct targets for transcription factors

NSF2010 MCB-0418891

SUMMARY

Senior Personnel

Erich Grotewold, PI – The Ohio State University (OSU)
Ramana V. Davuluri, co-PI – The Ohio State University (OSU)
Betsy A. Read, co-PI – California State University, San Marcos (CSUSM)
Rebecca Lamb, co-PI – The Ohio State University (OSU)

Summary of Project – Hierarchical arrangements (networks) of transcription factors (TFs) provide the information necessary to deploy genes with particular spatial and temporal patterns. The information for this is hardwired in the promoter regions formed by *cis*-regulatory elements that bind specific TFs. Direct target genes for a group of 31 selected TFs involved in regulatory networks associated with flower development and epidermal functions will be identified using chromatin immunoprecipitation followed by analysis of microarrays of promoter sequences (ChIP-CHIP). Identified direct target genes will be validated by *in vitro* DNA-binding experiments, and the TF binding sites will be determined using *in vitro* footprinting. The information obtained will be integrated into the established AGRIS databases, resulting in public resources that integrate TFs, their binding sites and the corresponding regulatory motifs. Results derived from this project will be made available at http://arabidopsis.med.ohio-state.edu/NSF2010Project. Materials developed in this research will be made readily available through the ABRC. The combination of experimental data and powerful computational tools will provide the first steps towards establishing the regulatory networks responsible for the expression of all *Arabidopsis* genes, one of the central objectives of the 2010 project.

Broader Impact – Establishing the regulatory networks that control the expression of all plant genes provides a natural follow-up to the elucidation of the *Arabidopsis* genome and will significantly impact the research in *Arabidopsis* and other plants. This project integrates experimental biology, bioinformatics, statistics and mathematics, offering unique opportunities for interdisciplinary research training. The development of several databases available through the Web will continue to ensure that findings derived from this work will have the broadest possible dissemination. Tools and resources obtained from these studies will be applied for educational purposes in courses at CSUSM and OSU.

PROJECT DESCRIPTION

SPECIFIC AIMS

Our long-term goal is to establish the architecture of the regulatory networks that govern the expression of all plant genes. Over 1,500 transcription factors (TF) are encoded by *Arabidopsis*, but today we only know the cellular processes controlled by a small subset of them. Even in cases of well-studied TFs, the regulatory circuits in which they participate and the target genes that they control remain unknown. Identifying direct target genes for *Arabidopsis* TFs is of central importance in establishing TF function, and provides a first step to start developing maps of regulatory networks that explain the expression patterns of all *Arabidopsis* genes. This proposal benefits from the insightful comments of the reviewers of a prior submission that, although unsuccessful in getting funded, was recommended as "High Priority" by the Panel. Here, we propose to:

Aim 1: Identify direct target genes for *Arabidopsis* TFs - We selected a group of 40 TFs for which we will identify their direct target genes by generating translational fusions with the glucocorticoid receptor hormone-binding domain (GR). These TF-GR fusion proteins will be expressed in mutant plants for the corresponding TF, and used to examine the direct target genes for each of these TFs using genome-wide microarrays. The identified direct target genes will be validated using several complementary approaches. The resulting information will be used to establish the network motifs in which these TFs participate.

Aim 2: Develop At*cis*DB into a genome-wide map of *Arabidopsis cis*-regulatory elements - We have initiated the development of At*cis*DB as part of AGRIS (Arabidopsis Gene Regulatory Information Server), a web resource that integrates information of *Arabidopsis* TFs and their binding sites. We will integrate available experimental data with unique computational tools to build At*cis*DB into a user-friendly and effective resource to predict *cis*-regulatory elements in all *Arabidopsis* genes.

Aim 3: Developing a knowledgebase of regulatory networks - We will model the regulatory networks that govern the expression of all *Arabidopsis* genes, using epidermal cell differentiation and flower development as foundations, by applying a variety of computational tools that include Bayesian networks, combinatorial approaches and linear models on experimental data. We will make these networks available as a knowledgebase, which will be an integral part of AGRIS.

SIGNIFICANCE AND RELEVANCE FOR 2010 PROJECT MISSION

Control of gene expression is central to all cellular processes. TFs function in networks, in which a regulatory protein controls the expression of another, which in turn may modulate the expression of other regulatory proteins, or control genes encoding structural proteins or enzymes. These hierarchical arrangements allow specific signals to be amplified, providing the information necessary for given sets of genes to be deployed with particular space and temporal patterns. Loss of function mutants of regulatory genes may result in phenotypic alterations that represent the integration of all the circuits in which a regulator participates. Because TFs often activate other regulatory genes, mutants alone are not sufficient to identify the direct targets of the corresponding regulators.

Gene function is intimately linked to when and where genes are expressed. This information is hardwired in the promoter regions formed by *cis*-regulatory elements recognized by specific TFs. Thus, establishing the architecture of plant promoters is not just fundamental to understand gene expression, but is directly relevant to gene function. The determination of the regulatory circuits in which TFs participate and the identification of the *cis*-regulatory sequences for all genes have been identified as two of the goals by the Multinational *Arabidopsis* Steering Committee (June 2002). In addition, the 2010 program for this year focuses on understanding gene circuitry. These program objectives will be uniquely met in this proposal by identifying the direct targets for 40 TFs. Most of the selected TFs participate in cellular processes that have been dissected at the genetic level. The identification of direct target genes for these TFs will permit us to determine the network motifs in which these TFs participate. Network motifs [37] can be considered the simplest units of the architecture of the complete *Arabidopsis* regulatory network. Establishing these network motifs will also allow us to validate aspects of the *Arabidopsis cis*-regulatory element map (At*cis*DB) to be developed in parallel. This is both a **tool development** and a **gene function discovery** proposal. From a tool development perspective, it will provide databases and novel computational tools to identify and evaluate the relevance of *cis*-regulatory sequences. From a functional genomics perspective, it will investigate the direct target genes for a selected group of TFs for which limited or no functional data is available. The tools generated and the data produced will be integrated into AGRIS, a web resource that will allow researchers to easily determine networks participating in the regulation of their gene of interest. This proposal also brings together, in a unique way, investigators from very different disciplines, providing a great opportunity for a more comprehensive integration of plant biology, bioinformatics and mathematics. This synergy will continue to enhance the training of students and postdocs in our laboratories in the interface between mathematics and biology.

BACKGROUND

An emerging theme in regulation of gene expression is to identify the regulatory networks in which TFs participate. TFs are classified according to the presence of conserved DNA-binding domains [42]. Members of the same family bind very similar DNA sequences *in vitro*, opening fundamental questions regarding how TFs activate specific sets of target genes *in vivo*, and limiting the value of available *cis*-regulatory element-finding programs. The establishment of regulatory networks in which TFs participate, through the identification of “direct targets” genes, remains a fundamental challenge in biology today.

Regulation of gene expression in eukaryotes – Gene expression can be controlled at multiple levels, including RNA processing, mRNA transport, stability, and translation, and protein modification and stability. However, the process of transcription itself is highly regulated by a group of proteins collectively known as TFs that function in a combinatorial fashion to specify when and how an eukaryotic gene will be expressed. For the purpose of this work, we define TFs as proteins that bind to DNA in a sequence-specific fashion. TFs are recruited to specific *cis*-regulatory elements organized in modules in the promoters of all genes. Each module is responsible for one specific expression output (e.g., temporal or spatial), with the overall expression provided by the combined presence of all the modules, and the corresponding TFs. Indeed, gene regulation may operate through defined logical steps, much in the way in which a computer functions [11; 65]. TFs are often modular in structure, combining the DNA-binding domain with a transcription activation domain that interacts with components of the basal transcriptional machinery to modulate the activity of RNA pol II. Protein-protein interactions and other modifications play a key role in the *in vivo* interaction of TFs with DNA, resulting in the exquisite regulatory specificity that they display *in vivo*. TFs can function as activators or repressors of gene expression, and TF can be an activator of one gene and a repressor of another. Repression can involve one or a combination of several possible mechanisms [5]. TFs frequently function as part of regulatory networks and are hierarchically organized, such that the target for one TF is often the gene encoding another.

Transcription factors in *Arabidopsis* – The *Arabidopsis* genome encodes for at least 1,500 TFs, which can be classified into 40-50 families based on sequence similarity [46]. Less than 200 *Arabidopsis* TFs have been genetically characterized, i.e. mutants resulted in detectable phenotypes. When mutants are available, genetics has been a powerful tool to predict possible networks in which these factors may participate, as shown by the floral development ABC model [22, 24; 25] and for the control of root and leaf hair development [3; 27; 52]. Notably, however, the genes that are directly controlled by the regulators of these processes remain largely unknown. Thus, most genetically characterized TFs continue to be isolated pieces in a complex puzzle, not anchored to any upstream or downstream regulatory process. Expression profile analyses have started to identify many TFs induced or repressed by particular biotic or abiotic conditions [8; 9; 19; 28; 53]. The clustering of gene expression profiles combined with the presence of *cis*-regulatory elements in the promoters of induced or repressed genes provides interesting leads regarding which genes might be directly regulated by particular TFs. Although the confirmation of these TF/target gene associations involves significant additional research, cluster analysis provides a powerful tool to start building regulatory networks, particularly in the absence of mutants for many TF genes. Mutant analyses has taken a giant leap over the past few years, and it is likely that T-DNA insertions in most TF genes will be readily available through the 2010 NSF-funded project (NSF-0115103; PI: Ecker). Indeed, T-DNA insertions or other mutants are available already for almost all the TFs to be investigated in this proposal. However, the availability of a mutant allele often helps little in determining the function of the gene, whether because of genetic redundancy, or because the specific conditions that activate the TF are not met. The limitations associated with reverse genetic approaches to establish the function of TFs are evidenced by the absence of visible phenotypes in more than 35 *R2R3 Myb* genes carrying transposon or T-DNA insertions [35].

The participation of a TF in a given regulatory motif can be inferred from mutant analyses or from gene expression profile clusters. However, determining the ultimate function of a TF depends on identifying which genes it can directly activate. Two main approaches are currently available to identify direct targets of TFs: a) By expressing a fusion of the TF to GR and identifying the mRNAs induced/repressed in the presence of the GR ligand (dexamethasone, DEX), in the presence of an inhibitor of translation (e.g., cycloheximide, CHX); or b) by identifying the DNA sequences that a TF binds *in vivo*, using chromatin immunoprecipitation (ChIP) assays. Both approaches have been successfully used to identify direct targets of a few plant TFs. The power of the analysis of GR fusions to identify direct targets of TFs is exemplified by the identification of NAP as a direct target of the AP3/PI floral homeotic genes (analysis of 35S::AP3-GR plants) [49]; by the identification of PAL as a direct target of AtMYB21 (analysis of 35S::AtMyb21-GR plants) [55]; and by the identification of DFR and the HLH factor JAF13 as direct targets of the Petunia AN1 HLH regulator (35S::AN1-GR) [58]. ChIP has been widely used in animals and yeast to study chromatin-bound factors (e.g., [1; 12]). Indeed, ChIP-CHIP (or genome-wide location analysis) provided a first map of the transcription regulatory networks that govern budding yeast gene expression [30]. ChIP analyses, however, have only recently started to be applied to plant TFs (e.g., [26]). Both of these methods have strengths and limitations, and it is probably a combination of both, linked with cluster analysis of RNA profiling experiments, that will provide the most accurate image of the direct targets of a particular TF.

Identification of *cis*-regulatory elements – Regulatory elements are often identified by generating deletions and mutations in the promoter of interest. Sometimes, *cis*-regulatory elements have been confirmed by *in vivo* footprint experiments [32; 43; 44] or by the availability of *in vivo* mutations [45]. More often, if the TF(s) that regulate the gene are known, then *cis*-regulatory elements are predicted based on the *in vitro* DNA-binding properties of the TFs. TFs from the same family often recognize similar DNA sequences. For example, G-boxes (CACGTG) are bound by the GBF subfamily of bZIP proteins, E-boxes conform to the typical binding site of bHLH factors [39], and WRKY factors bind the W-box consensus $(^{T}/_{N}{}^{T}/_{N}TGAC^{C}/_{T})$ [14]. Several computational tools and public databases are available that help predict *cis*-regulatory elements, starting from a query DNA sequence. TRANSFAC (http://transfac.gbf.de) is one of the most widely used databases of TF binding sites and TFs, yet contains few plant accessions. EPD (http://www.epd.isb-sib.ch) holds information on previously characterized promoter sequences, yet contains only 18 *Arabidopsis* accessions (Dec.'03). This database is confined to sequences that are found immediately upstream of the transcription start site (TSS) and does not contain any annotation about regulatory elements. PlantCARE (http://oberon.rug.ac.be:8080/PlantCARE/index.html) contains information on about 435 different plant TF binding sites in 159 promoters. From these, 281 correspond to dicots, just a fraction of the total *cis*-regulatory regions present. A serious limitation of PlantCARE and similar resources, such as the recently developed AthaMap [40], is the very similar DNA-binding specificity of factors that belong to the same TF family. The output of PlantCARE with a query sequence will result in the prediction of possible sites recognized by any of the 100+ bHLH or AP2/EREBP present in *Arabidopsis*. While PlantCARE has recently made available computational programs to identify new regulatory elements from the analysis of transcriptome data [31; 62], their output has not been integrated with the *cis*-element prediction.

EXPERIMENTAL DESIGN AND METHODS

Specific Aim 1: Identify direct target genes for *Arabidopsis* TFs

Rationale and overview – To determine the functions of the selected TF, and as a first step in establishing the regulatory networks in which these *Arabidopsis* TF participate, we will identify direct targets for (at least) the 40 selected TFs using GR fusion constructs (TF-GR) in transgenic *Arabidopsis* plants and genome-wide microarrays. For most of the TFs selected, mutants are available that results in a visual phenotype (See Justification of the TF Genes Selected). In those cases, the TF-GR proteins will be expressed in the corresponding mutant backgrounds to ensure that the TF-GR fusion protein is functional, complementing the mutant phenotype in the presence of DEX. In a few other cases, phenotypes associated with the available mutants have not yet been reported. In those cases, the TF-GR fusion will be expressed in wild type (Columbia) plants.

a) Generation of transgenic plants expressing TF-GR fusions – TF-GR translational fusions will be obtained by PCR amplification of the TF coding regions with primers that permit their cloning into a Gateway™ pENTR TOPO® plasmid. The TFs will then be transferred into modified binary Gateway™ destination plasmids containing the GR region either 5' of the *att*R1 or 3' of the *att*R2 to obtain N- or C-terminal (in frame) GR fusions (TF-GR or GR-TF) respectively. Initially, we will test the TF-GR fusions, but if they fail to complement the mutant phenotype, we will test the GR-TF construct. By using the Gateway™ system, we minimize the time involved in cloning. The technology allows us to easily mobilize the initial TF genes into multiple vectors (e.g., binary, for *E. coli* expression, for GFP fusion, etc.). The 35S::TF-GR constructs will be introduced into *Arabidopsis* plants mutant for each TF (when mutants are available) by *Agrobacterium*-mediated transformation. Single T-DNA insertion T1 lines will be identified to increase transgene stability, and checked for mutant complementation after DEX treatment. Expression of the transgene will be verified by RT-PCR. Plants will be carefully monitored for the appearance of new phenotypes or for complementation of the mutant phenotype that may suggest activity of the TF-GR in the absence of DEX. Five independently transformed single T-DNA lines expressing each transgene will be maintained. In those cases in which the mutant (homozygous) prevents propagation (e.g., no flowers), the mutants will be propagated as heterozygotes and the transgenes will be introduced by crossing. The treatments of the plants with DEX will be carried out at one of two stages of development, depending on whether the specific TF participates in an epidermal process or in flower development. For the analysis of genes with a function in the epidermis, we will carry out the analyses in two-week old seedlings. For the analysis of TFs involved in flower development, we will treat inflorescences with flowers at various stages of development after bolting. TFs with unknown function will be investigated in both conditions, unless the expression pattern of the TF is known.

Two-week old seedlings - Seedlings corresponding to five independent transformed lines carrying each 35S::TF-GR transgene will be grown in MS+1% sucrose agar plates for 15 days at 22°C in continuous light. Hormone induction experiments will be carried out by flooding the plants for 2 hours with MS media containing 10 μM DEX (+DEX), 100 μM CHX (+CHX), 100 μM CHX and 10 μM DEX (+CHX +DEX), or MS alone (Mock). Plants transformed with an empty vector will be grown as control in each set of experiments and treated in identical conditions. After the 2 hours treatment, seedlings will be collected and RNA will be extracted. These conditions have been previously successfully used, for example, to identify direct target for *AtMyb21* [55].

Flowers – The hormone induction experiments in the flowers will be conducted as previously described for AP3/PI [49]. Flowers and siliques at different developmental stages will be soaked in a solution containing 10 μM DEX (+DEX), 100 μM CHX (+CHX), 100 μM CHX and 10 μM DEX (+CHX +DEX), or water alone (Mock). Control plants transformed with an empty vector will be treated in identical conditions. After the 2 hours treatment, flowers and siliques will be collected and RNA will be extracted. The generation, analysis and treatment of *Arabidopsis* transgenic plants will be carried out in the Grotewold and Lamb labs. Undergraduate students will carry out the generation and analysis of the transformed *Arabidopsis* plants, and the isolation of tissues and RNA extraction experiments. Senior personnel more familiar with *Arabidopsis* will perform the visual analyses.

***b) Identification of direct target candidate genes by genome wide expression analysis* –** The direct targets of our select TFs will be identified by using high-density oligonucleotide probe microarrays (ATH1 GeneChips) containing 24,000 *Arabidopsis* gene sequences [21]. Total RNA will be extracted from seedlings or floral tissues expressing the TF-GR fusions following hormone induction experiments under four different conditions: 1) +DEX, 2) +CHX, 3) +CHX +DEX, and 4) Mock. The preparation of cRNA from total RNA and subsequent steps leading to the hyrbridization and scanning of the ATH1 Array will be performed following standard Affymetrix procedures. cRNAs will be purified and randomly fragmented prior to hybridization. Arrays will be hybridized, washed and stained with streptavidin-phycoerythrin. Chips will be scanned with the GeneArray Scanner. The GeneChip Suite 3.2 software will assist with background subtraction and normalization of the data that will then be used for scaling, statistical analysis, and data mining. The average intensity of each array will be scaled to 100, to compare the hybridization intensity across arrays. To control for biological variation, RNA samples will be pooled from at least five individual plants receiving the same treatment. The detected gene expression will be the average response of the biological replicates. To further ensure data quality and comparability, for each TF-GR fusion, tested biological replicates will be performed across treatments. Reproducibility of the independent replicates under each of the defined conditions will be assessed using an ANOVA and individual correlation coefficients for each pair. Stringent criteria will be used to select differentially expressed genes. A putative direct target of a TF will be identified as a gene whose expression is greater than or equal to the threshold expression level, and whose expression is:

[1] Increased or decreased 2 fold or more in +DEX +CHX versus +CHX
[2] Increased or decreased 2 fold or more in +DEX versus Mock.
[3] Not significantly different in +DEX versus +DEX +CHX

Although direct target candidates will derived from the comparison of +DEX +CHX/+CHX treatments (keeping in mind that gene expression can be affected by the CHX treatment, hence the +CHX control), downstream or secondary targets will be identified by comparing the expression of genes under the +DEX and Mock treatment conditions. Similar experiments will be performed using RNA extracted from transgenic plants harboring empty vectors to rule out alterations in gene expression induced by DEX alone. The GeneSpring (or comparable) microarray software will compile the various data sets according to MIAME standards and MAGE data exchange formats. The microarray analysis will be carried out at CSUSM in the Read lab. A minority graduate RISE (Research Initiative for Scientific Enhancement) or MARC (Minority Access to Research Careers) scholar, working under the direction of the co-PI, will be responsible for the cRNA synthesis and microarray hybridizations. The actual hybridizations, staining, and scanning will be performed by the co-PI and students at the Scripps Research Institute. Data analysis will be performed by the graduate student and the co-PI, each working together with a undergraduate RISE or MARC scholar. The data generated from these experiments will be centrally located and made available to all participants immediately (see Management Plan Appendix A-2).

Expected outcomes and solutions to possible pitfalls - The experiments described here are devised to ***narrow down*** the space of genes that may correspond to direct targets for each of the selected TFs. The complementation of the mutant phenotypes (or recapitulation of reported over-expression phenotypes) by the 35S::TF-GR constructs (after DEX treatment) ensures that the TF-GR fusion is controlling the corresponding target genes. If complementation of the mutant phenotype is not observed using the 35S promoter with either TF-GR or GR-TF, then we will utilize the native promoter to express the corresponding TF-GR fusion. We expect that the 35S::TF-GR construct will activate or repress genes, in addition to the direct targets, that are not normal targets for the corresponding TF (false positives), either because of the higher level of expression of the TF obtained from the 35S promoter or because the 35S promoter is expressed in cell types in which the TF is normally not expressed. For the TFs involved in flower development, obtaining RNA from floral tissues will minimize the problem. Obtaining epidermal tissue is not that easy. However, if in our initial studies we find out that this is a serious problem, we will investigate whether enriching for epidermal cells by FACS analysis of young leaf protoplasts obtained from AtML1::GFP plants (AtML1 is an available L1-layer specific promoter [54]) before RNA extraction decreases the number of false positives and negatives. As part of a synergistic collaboration with Dr. Alan Lloyd (see letter), his lab will contribute to generate TF-GR transgenic plants for several factors involved in epidermal cell differentiation. The studies to be conducted in Dr. Lloyd's lab will further serve to validate the microarray experiments carried out as part of this proposal. We expect to be able to carry out preliminary experiments with already available 35S::EGL3-GR plants even before a funding decision for this proposal is reached.

Microarray analyses are vulnerable to methodological problems associated with variables such as differences in the absolute amount of labeled test and experimental cRNA hybridized to the slides and the lack of linearity between the quantified signal and the expression level of the corresponding genes. All these problems can also result in false positives. Many of these false positives will be filtered in the validation process described below. It is also possible that the 2-fold difference in expression used as criteria for the analysis of the microarray data could make us miss genes, which are indeed direct targets for the TF (false negatives). *A priori*, we have no idea of how many direct target candidate genes we will identify for each TF. Previous studies in plants [50] and yeast [17] found few (<10), but obviously this remains an open question for the specific set of TFs to be investigated here. It would be a mistake to expect that the analyses described here will provide a dynamic image of all the target genes that a TF can control. Rather, our approach is intended to provide one snapshot of the possible genes that a TF can directly regulate. Co-PI Read is well experienced with microarray work. In 1999, as a Visiting Scientist at the Novartis Agricultural Institute, Read helped establish their microarray facility. Dr. Read has also taught four different intensive one-week microarray technology workshops, attended by scientists from around the world (http://www.csusm.edu/bread/Microarraymain.htm). Currently, Dr. Read and Dr. Grotewold labs are collaborating in genome-wide microarray expression experiments aimed at identifying *Arabidopsis* genes induced by flavonoids.

c) Validation of the identified direct target genes – Several complementary criteria will be employed to validate the identified direct target genes:

1. Significantly different expression in +DEX +CHX vs. +CHX and Mock controls TF-GR plants
2. Significantly different expression in wild type compared to mutant plants (when mutants available)
3. Binding of the TF to the direct target gene *in vitro* and *in vivo*

Real Time RT-PCR will be used to establish whether a candidate direct target gene meets the first two criteria. A combination of electromobility shift assays (EMSA) and ChIP will be used to determine whether a putative direct target genes meets the third criteria and to find out the specific DNA elements recognized by the corresponding TF.

Expression analysis of putative direct target genes – Real time RT-PCR will be performed using cDNA obtained from RNA extracted from all five independent transformed lines for altered accumulation in +CHX +DEX, vs. +CHX and Mock controls. Only genes that display a significant difference (increase or decrease depending on whether the TF is acting as an activator or repressor) after DEX treatment in all five lines (compared to actin and GAPDH controls) will be considered putative candidate direct target genes for each TF. Similarly, the expression of each of the putative target genes will be tested in wild type or mutant plants for the corresponding TF gene. It is expected that in general, if a gene is a direct target for a given

TF, its expression will be different in mutant versus wild type tissues. Briefly, SYBR Green RT-PCR amplifications of genes identified as direct targets for one particular TF will be performed. cDNA synthesis, primer design, and SYBR green RT-PCR will be performed as described [48]. All assays will be performed in triplicate and include: a standard curve of four serial dilution points for the actin and GAPDH controls (50 ng to 50 pg), a no-template control, and the test cDNA. At the end of each real-time RT-PCR run, data analysis will be performed using the iCycler iQ analysis software. The cycle threshold values will be exported into a Microsoft Excel sheet and/or MiniTab for subsequent data analyses. A one-way ANOVA will be used to compare the gene expression across treatment conditions with fold increase in mRNA relative to actin and GAPDH, as the dependent variable. The real time PCR experiments will be carried out by the Read lab (CSUSM). A graduate RISE or MARC scholar and the co-PI each working with a minority undergraduate student (see Broader Impacts Section) will validate the direct targets of ~5-10 TFs each year.

Analysis of protein-DNA interactions – Once the number of putative direct target genes has been narrowed down by the expression analyses described above, we will assess whether the promoters of the target genes have binding sites for the respective TFs. This will be achieved by ChIP [41; 56], using the protocol recently adapted for *Arabidopsis* [26]. This method will permit us to confirm the binding of the TF to the promoter of the putative direct target in plant cells. In parallel, we will investigate the binding of the TF *in vitro* to the putative target genes using EMSA. The binding site in the direct target gene will be further identified using footprinting methods. In order to carry out these experiments, recombinant TFs will need to be obtained first (for EMSA & *in vitro* footprint analyses), as well as to produce antibodies (for ChIP analyses) in those cases when antibodies or epitope-tagged proteins are not available.

Expression of recombinant TFs and generation of antibodies – To express the TFs in *E. coli*, we will move the entire ORF of the TFs from the constructs in the Entry vectors into vectors that allow expression in *E. coli* as N-terminal poly-histidine fusions (N_6His-TF; pDEST™ vectors). Expression and purification from *E. coli* extracts will be carried out using affinity chromatography on Ni-NTA columns, as done before [18; 51; 63]. Pure N_6His-TF will be used as antigens into rabbits using available commercial providers. We have successfully used before Cocalico Biologicals Inc., who have agreed on a special low rate if we provide them with 30-40 antigens over a two years time period. The specificity of the Abs for the TF of interest will be verified by Western analysis of wild type and mutant plant protein extracts. If the Abs need to be further purified to increase their specificity, then fragments of the proteins that are most divergent between members of a gene family and between recently duplicated genes will be expressed as GST fusions. Clones for the recombinant proteins as well as Abs will be made available to the entire scientific community through the ABRC (see letter). The expression of proteins in *E. coli*, their purification and the analysis of Abs will be carried out in the Grotewold lab. Undergraduate students will help, particularly with the expression and purification of recombinant TFs in *E. coli*, techniques that are well established in the lab and that are routinely carried out by students.

ChIP analyses – ChIP will be carried out essentially as described [26], by comparing results using tissues obtained from wild type and mutant plants. In those TF cases for which mutant plants are not available, the ChIP experiments will be carried out in 35S::TF-GR tissues, with and without DEX. Briefly, *Arabidopsis* tissues (flowers or leaves, depending on the TF to be tested) will be isolated and incubated with 1% formaldehyde at 30°C for 15 minutes. The cross-linking reaction will be terminated by the addition of glycine. After nuclei have been isolated from N_2(l)-ground tissue, chromatin will be extracted as described [26]. After IP of the protein-DNA complexes in RIPA buffer, RNA and proteins are eliminated by RNase/protease treatment followed by reversal of the cross-links by incubation at 65°C and DNA isolation. Primers specific to the promoters and other parts of the genes (first introns often serve as binding sites for TFs) identified as putative direct targets will be used to determine by PCR whether they have been recovered in the IP. As negative controls, we will include DNA obtained from plant tissues without the formaldehyde treatment, PCR from a gene not regulated by the TF, and controls in which an unrelated antibody has been added instead of the TF-specific antibodies. The ChIP experiments will be carried out in the Grotewold lab, where the technique is currently being standardized.

In vitro DNA-binding experiments – We will also determine whether the putative direct targets can be bound by the corresponding TF *in vitro*. Briefly, 300–500 bp regions corresponding to the promoters of the putative direct targets will be amplified by PCR using one of the primers radioactively labeled with ^{32}P, as done before [18]. The PAGE-purified labeled PCR fragment will be used as probe in EMSA with the recombinant pure TFs described above. If no binding is observed, we will test additional regions in the promoter and first intron, aided by the computer prediction of where the possible binding sites might be. When a specific fragment was identified in the ChIP experiments as containing a candidate-binding site, then that fragment will be used for the *in vitro* binding experiments. If a shifted complex is observed in EMSA, experiments with specific and non-specific DNA competitors should further establish the specificity of the observed DNA-binding activity. The binding site for the TF will ultimately be determined using DNAase I footprinting experiments, as done previously [18]. This information will be used to further advance At*cis*DB (Objective 2). In addition, knowledge on the *cis*-regulatory elements recognized by TFs will be used to determine the contribution of these *cis*-regulatory elements to the overall expression of the direct target gene. This will be accomplished by expressing promoter-GFP fusions in transgenic *Arabidopsis* plants and comparing the GFP expression patterns driven by wild type and mutant promoters. Briefly, promoters for the direct target genes will be amplified by PCR from genomic DNA or available BAC and cloned in Entry vectors. After site directed mutagenesis, mutant and wild type promoters are transferred to the Gateway pK7WGF2 plasmid containing an in-frame GFP, which is used for transformation into *Agrobacterium* and subsequent transformation of *Arabidopsis* plants, wild type and mutant for the corresponding TF. Promoter-GFP fusions will be made available to the community through the ABRC. While we recognize the importance of this *in vivo* analysis of the promoters of the identified direct target genes, it is unlikely that during the time period of the proposal we will be able to analyze in this way every direct target gene identified.

Absence of a shifted protein-DNA complex in the EMSA experiments could be indicative of i) not having met the appropriate conditions for binding, ii) need for a partner or post-translational modification for binding, or iii) the fragment being used as a probe not containing the binding site. Distinguishing between these different possibilities should be greatly facilitated by the results obtained from the ChIP experiments and by the knowledge available on the specific TF (e.g., bHLH factors tend to bind DNA as homo- or heterodimers, thus absence of DNA-binding could indicate that we are missing the partner).

<u>Expected outcomes and solutions to possible pitfalls</u> – Although the ChIP experiments are fairly straight forward, Dr. Jonathan Arias (University of Maryland) offered assistance, if such was needed. We considered carrying out ChIP-CHIP analysis, instead of using TF-GR fusions, to identify the direct targets for the TFs selected here. However, the ChIP-CHIP technique is far more challenging and microarrays representing the entire *Arabidopsis* genome for these experiments are not yet available. Because this is likely to change within the next 1-2 years, we may then consider comparing the results obtained with the TF-GR fusion with those of ChIP-CHIP experiments for a few TFs to determine which of the two methods better describes the direct targets of TFs. The Grotewold lab is well experienced in all the other approaches described in this objective, thus unsolvable technical problems are not expected.

Modified[1] Specific Aim 1: Identify direct target genes for *Arabidopsis* TFs

[1] This original aim 1 was modified to this version based on the comments of the panel

To determine the functions of the selected TF, and as a first step in establishing the regulatory networks in which these *Arabidopsis* TF participate, we will identify direct targets for 31 selected TFs (see revised list of TFs at the end of this document) using the ChIP-CHIP technology. Originally, we intended to identify direct targets by making TF-GR fusions. The ChIP-CHIP technology will now be used instead, but we thought that it would be of great interest for the community to have a careful comparison of the results obtained by these two approaches. Thus, we will take advantage of our collaboration with Dr. Alan Lloyd (U. Texas, collaboration letter attached in original proposal) to compare by ChIP-CHIP and TF-GR fusion (already generated in his lab) the direct targets for GL3, EGL3, TT8, GL1 and PAP1. Only for these five genes, microarrays will be made as described below in Aim 1b. The validation of all the putative direct targets will be carried out as outlined in the original proposal, by using EMSA and *in vitro* footprinting analyses. In contrast to the original proposal, in this revised objective we will already have the pure recombinant proteins available to carry out these validation analyses.

a) ChIP-CHIP analyses **–** As a first step for the ChIP-CHIP analysis, antibodies will be obtained for all 31 TFs. Briefly, to express the TFs in *E. coli*, we will move the entire ORF of the TFs from the constructs into vectors that allow expression in *E. coli* as N-terminal poly-histidine fusions (N_6His-TF; pDEST™ vectors). Expression and purification from *E. coli* extracts will be carried out using affinity chromatography on Ni-NTA columns, as we have extensively done in the past. Pure N_6His-TF will be used as antigens into rabbits using available commercial providers. The specificity of the Abs for the TF of interest will be verified by Western analysis of wild type and mutant plant protein extracts. If the Abs need to be further purified to increase their specificity, then fragments of the proteins that are most divergent between members of a gene family and between recently duplicated genes will be expressed as GST fusions. Clones for the recombinant proteins as well as Abs will be made available to the entire scientific community through the ABRC (see letter in original proposal). The expression of proteins in *E. coli*, their purification and the analysis of Abs will be carried out between the Read, Lamb and Grotewold labs. Undergraduate students will help, particularly with the expression and purification of recombinant TFs in *E. coli*.

ChIP analyses **–** ChIP will be carried out essentially as described previously [30]. We have enlisted the help of Dr. Chris Town (TIGR). Briefly, *Arabidopsis* tissues (flowers or leaves, depending on the TF to be tested) will be isolated and incubated with 1% formaldehyde at 30°C for 15 minutes. The cross-linking reaction will be terminated by the addition of glycine. After nuclei have been isolated from N_2(l)-ground tissue, chromatin will be extracted. After IP of the protein-DNA complexes in RIPA buffer, RNA and proteins are eliminated by RNase/protease treatment, the cross-links reversed by incubation at 65°C, followed by DNA isolation. The DNA is broken down into ~700 bp fragments, and after labeling, used to hybridize microarrays (CHIP). For each TF, two DNA samples are generated, one corresponding to enriched chromatin (i.e., precipitated with the TF specific Ab) and another corresponding to control (i.e., using a pre-immune, non specific sera). For each TF, all the experiments will be carried out in duplicate, thus for each TF four hybridizations will be necessary. The CHIPs to be used are in the process of being developed by NimleGen (http://www.nimblegen.com/). Currently, NimbelGen has one array that contains 195,000 features corresponding to 7 oligos (55-65 nt long) per promoter represented in a single CHIP. They are working on another that will have oligos spanning the entire genome, 100 bp apart from each other (3 arrays total). We expect the first set of array to become available for these studies by late summer. The use of those arrays essentially represents our original objective to identify regulatory elements in promoters. It is clear however that in many cases regulatory elements might also be present in the introns or 3' sequences. It would therefore be convenient to compare with at least a small subset of genes whether the results with the promoter CHIP and the genome wide CHIP give similar or very different results. NimbleGen is a service providing company, thus they will carry out the hybridizations and will provide us with the raw data to analyze. Co-PI Davuluri is already working with NimbleGen in a project aimed at identifying target genes for animal E2F factors. Thus. expertise to analyze the results from these ChIP-CHIP experiments is already in house. The bulk of the analysis of the data will be carried out at CSUSM, using the expertise developed by co-PI Read for the analysis of microarray data.

***b) Identification of direct target candidate genes by genome wide expression analysis* –** This will be done only with the five genes listed before, and the results of the TF-GR and ChIP-CHIP will be compared. Seedlings corresponding to five independent transformed lines carrying each 35S::TF-GR transgene will be grown in MS+1% sucrose agar plates for 15 days at 22°C in continuous light. Hormone induction experiments will be carried out by flooding the plants for 2 hours with MS media containing 10 μM DEX (+DEX), 100 μM CHX (+CHX), 100 μM CHX and 10 μM DEX (+CHX +DEX), or MS alone (Mock). Plants transformed with an empty vector will be grown as control in each set of experiments and treated in identical conditions. After the 2 hours treatment, seedlings will be collected and RNA will be extracted. These conditions have been previously successfully used, for example, to identify direct target for *AtMyb21* [55]. The direct targets of our select TFs will be identified by using high-density oligonucleotide probe microarrays (ATH1 GeneChips) containing 24,000 *Arabidopsis* gene sequence. Total RNA will be extracted from seedlings expressing the TF-GR fusions following hormone induction experiments under four different conditions: 1) +DEX, 2) +CHX, 3) +CHX +DEX, and 4) Mock. The preparation of cRNA from total RNA and subsequent steps leading to the hyrbridization and scanning of the ATH1 Array will be performed following standard Affymetrix procedures. cRNAs will be purified and randomly fragmented prior to hybridization. Arrays will be hybridized, washed and stained with streptavidin-phycoerythrin. Chips will be scanned with the GeneArray Scanner. The GeneChip Suite 3.2 software will assist with background subtraction and normalization of the data that will then be used for scaling, statistical analysis, and data mining. The average intensity of each array will be scaled to 100, to compare the hybridization intensity across arrays. To control for biological variation, RNA samples will be pooled from at least five individual plants receiving the same treatment. The detected gene expression will be the average response of the biological replicates. To further ensure data quality and comparability, for each TF-GR fusion, tested biological replicates will be performed across treatments. Reproducibility of the independent replicates under each of the defined conditions will be assessed using an ANOVA and individual correlation coefficients for each pair. Stringent criteria will be used to select differentially expressed genes. A putative direct target of a TF will be identified as a gene whose expression is greater than or equal to the threshold expression level, and whose expression is:

[4] Increased or decreased 2 fold or more in +DEX +CHX versus +CHX

[5] Increased or decreased 2 fold or more in +DEX versus Mock.

[6] Not significantly different in +DEX versus +DEX +CHX

Although direct target candidates will derived from the comparison of +DEX +CHX/+CHX treatments (keeping in mind that gene expression can be affected by the CHX treatment, hence the +CHX control), downstream or secondary targets will be identified by comparing the expression of genes under the +DEX and Mock treatment conditions. Similar experiments will be performed using RNA extracted from transgenic plants harboring empty vectors to rule out alterations in gene expression induced by DEX alone. The GeneSpring (or comparable) microarray software will compile the various data sets according to MIAME standards and MAGE data exchange formats. The microarray analysis will be carried out at CSUSM in the Read lab. A minority graduate RISE (Research Initiative for Scientific Enhancement) or MARC (Minority Access to Research Careers) scholar, working under the direction of the co-PI, will be responsible for the cRNA synthesis and microarray hybridizations. The actual hybridizations, staining, and scanning will be performed by the co-PI and students at the Scripps Research

Institute. Data analysis will be performed by the graduate student and the co-PI working together with a undergraduate RISE or MARC scholar. The data generated from these experiments will be centrally located and made available to all participants immediately (see Management Plan).

Specific Aim 2: Develop At*cis*DB into a genome-wide map of *Arabidopsis cis*-regulatory elements.

Rationale and overview – The dissection of promoters and the identification of the *cis*-regulatory elements responsible for the temporal and spatial expression of a gene is a central focus of many research programs today. We will continue to evolve At*cis*DB into a map of computationally predicted *cis*-regulatory motifs in all *Arabidopsis* genes. We will combine mathematical algorithms in novel ways, for example to identify motifs over-represented in genes belonging to a metabolic pathway (likely to be coordinately regulated) or co-regulated in expression genome-wide profiling experiments. This map will be further complemented by links that will indicate which of these sites are confirmed from experimental data, such as *in vitro* DNA-binding studies or mutational analysis.

Experimental design - The generation of the At*cisDB* consists of the following steps:

Step 1: We already developed a database of upstream sequences for all annotated *Arabidopsis* genes (see Preliminary Results section). As described, we are in the process of improving this database by integrating data from full-length cDNAs that allow us to determine the TSS and the first intron. Similarly, analyses of core promoters for genes with well-defined TSS will be incorporated.

Step 2. AGRIS already contains a draft annotation of At*cis*DB 1.1. We will continue to improve the annotation of putative *cis*-regulatory elements based on the following criteria and approaches:

1. *Experimentally proven TF binding sites:* About 49 binding sites for specific TFs have been experimentally demonstrated. AGRIS provides links of the TFs to their respective binding sites in the corresponding target genes. In At*cis*DB, these experimentally validated binding sites are clearly distinguished from other computationally predicted motifs. Position Weight Matrix Methods (PWMs, [60]) are currently being applied to those *Arabidopsis* TFs for which at least 5-10 experimentally known binding sites exist. PWMs have been used to estimate the likelihood that a given sequence binds to a specific TF [15]. Using PWMs, we will scan the promoters of At*cis*DB for putative binding sites.

b) TF consensus binding sites: AtTFDB contains the TFs and their consensus binding sites, whenever they have been determined. We have developed Perl scripts to scan the promoters for consensus binding site occurrence, and this information is already available in AGRIS. Currently, we have consensus binding sites for 24 families/sub-families of TFs.

c) Conserved motifs: A number of DNA motifs present in genes that respond to specific biotic or abiotic stress conditions have been identified (e.g., ABRE, DRE, CArG; a comprehensive table can be obtained from AGRIS). In addition, new conserved motifs have been uncovered by analyzing co-regulated genes from genome-wide gene expression analyses (e.g., [8; 28]). In most cases, the specific TFs that recognize these motifs have not yet been identified. Today, AGRIS lists a total of 32 such motifs.

We are making available a link at AGRIS where users can report back any of the validated annotations of At*cis*DB 1.1 from their experiments. As part of this proposal, we will continue to improve the annotation of plant promoters for the presence of binding sites or conserved motifs. Although we have received substantial feedback from the community regarding novel motifs, much of the work involved in this step will continue to require a careful and systematic analysis of the literature.

Step 3. We are currently developing and will continue to develop new mathematical algorithms and tools for the identification of new conserved motifs. Two of these approaches are described below:

a) Common motifs in co-regulated genes: We are taking advantage of the increasing amount of microarray expression data to identify motifs in co-expressed genes. Co-expression may reflect co-regulation, which may be deduced by an integrated analysis of genomic sequence and genome-wide expression profile data. This strategy has been used extensively for organisms with smaller genomes, such as yeast [23; 57; 61]. Our analyses will be complemented by extensive searches of publicly available microarray expression data to include all such information about TFs and their regulatory motifs present in their target gene promoters. In addition, we will make available a link in AGRIS that will allow investigators to search for conserved motifs or TF binding sites in specific clusters of co-regulated genes. Briefly, the Id of the co-regulated genes will be submitted in batch through AGRIS. We will automate the analysis of the corresponding promoters using a local version of MEME (version 3.0.4; available from ftp://ftp.sdsc.edu/pub/sdsc/biology/meme), which will provide the model motifs as a PWM, using an Expectation Maximization algorithm to fit a two-component finite mixture model to the sequence data. The output will be parsed to filter only the most significant motifs and within these motifs a search for *cis*-acting elements will be carried against our motif and TF binding site database. A statistical analysis of the resulting motifs will be carried based on its log likelihood ratio, its width and number of occurrences, a higher order Markov model of the background letter distribution (given as an input file in the command line summary), and the size of the training set to rapidly determine whether the identified motifs are over-represented in the query data set, compared to the entire genome. The statistical analyses described here and in other parts of this proposal will be greatly enhanced by the collaboration with Dr. Ralf Bundschuh (Dept. of Physics, OSU, see letter). As a test data set, we are working together with Dr. JC Jang (OSU, Dept. of Hort. & Crop Sciences) to analyze motifs present in genes whose expression is affected by glucose (see letter). We will also test the recently developed Network Component Analysis method [33] to data generated from microarray experiments. The results will be formatted and sent by e-mail to the scientist. An agreement will be reached with investigators submitting data for these analysis that the motifs that we identify could be made available (within 6 months of submission) through AGRIS.

b) Common motifs in metabolic pathway genes: Genes encoding enzymes that participate in a common metabolic pathway are often co-regulated by a few TFs (e.g., [6; 36] and references therein). AraCyc [38] has recently been made available through TAIR as a resource that integrates *Arabidopsis* genes with metabolic pathways. Using AraCyc, we will search for DNA motifs over-represented in genes corresponding to a given metabolic pathway. The analysis of the promoters of these genes will be carried out similarly as described above, yet the smaller number of genes that participate in a pathway (3-15) makes it possible to develop new algorithms to also identify combinations of motifs.

Step 4. Each of the annotated binding sites will be updated with information derived from experimental data available in the literature and from our own experiments (Aim 1). The database with annotations from experimental support will be integrated into At*cis*DB 2.0, a dynamic database that will continue to be updated as additional experimental data accumulates. In addition, we will include:

Comparative analyses of regulatory sequences from related plants – The comparison of genomic sequences as a tool to identify important regulatory sequences is gaining momentum [59]. Indeed, co-PI Davuluri is taking advantage of the available genomic sequences of human and mouse to identify *cis*-regulatory elements (MPromDB: http://bioinformatics.med.ohio-state.edu) recognized by TF with relevance in disease [64]. Recently, a pilot study was carried out to find conserved non-coding sequences (CNS) between the regulatory regions of *Arabidopsis* and cauliflower [10]. Genome sequence of *Arabidopsis* relatives is expected to expand as part of a new 2010 proposal submitted concurrently to this one (*The Genomics of Natural Variation: The Arabidopsis Relatives Sequencing Project; PI/co-PIs: Benfey, Dangl, Mitchell-Olds & Wessler*). As DNA sequence information from Arabidopsis relatives accumulates, derived from this and similar projects, we will perform comparative sequence analysis of orthologous gene promoters using VISTA [13; 34] and other algorithms developed in the Davuluri lab for the comparison of mice and human promoters [64]. We expect that this comparative analysis will also underscore other conserved elements that might be recognized by as yet unidentified factors.

Motif Discovery for potential new binding sites - Promoter regions of potential target genes of each TF will be analyzed by identifying over-represented sequence motifs using MEME [2]. We will also use other motif finding programs in DNA-sequences such as AlignACE [47], Motif Sampler [62], and cis/TF [4]. A confidence value will be assigned to each motif based on an E-value calculated by MEME and a specificity score. AlignACE identifies motifs that are over-represented in the promoter regions of co-regulated genes. Motif Sampler is a retrained GibbsSampler [29], with a reliable background model based on a set of carefully selected intergenic sequences of *Arabidopsis*. In contrast to other methods to search for binding sites by clustering genes that have similar expression patterns, cis/TF considers that B is a likely binding site for TF T if the expression of T correlates with the composite expression patterns of all genes containing B, not just those that are expressed similarly as B [4]. Cis/TF has been shown to predict the correct sites in experimentally supported TF-binding interactions in *S. cerevisiae* [4]. The identification of putative binding sites by TFs not yet identified may provide a powerful complement to other approaches to establish gene function, as some of the genes with no homology may encode TFs with novel DNA-binding domains. The identified conserved sequences with no obvious similarity to the binding of known TFs could be used as probes in protein-DNA binding experiments to identify factors that bind to them. While such an analysis highlights the tremendous power of the computational studies described here, they are outside the scope of this proposal. The availability of At*cis*DB will also permit us to carry out a statistical analysis to determine how often binding sites for two or more TFs appear together. The combinatorial nature of transcriptional regulation will be further explored by making use of different statistical models such as CART [7], MART [16] and logistic regression [30][16] methods. At*cis*DB 2.0 will also contain binding sites, identified as described above, in the introns of the genes.

<u>Expected outcomes and solutions to possible pitfalls</u> – A publicly available promoter database with annotation of *cis*-regulatory elements (At*cis*DB) is a major outcome of this aim. It will contain the annotation of both experimentally known and computationally annotated potential TF binding sites. At present, we have not identified any potential software problems regarding the development of At*cis*DB. The databases of AGRIS are optimized to handle large data sets for easy and faster access. When necessary, these databases will be moved to a more powerful server, preferably a larger Linux cluster of multiple nodes. GDVTK is optimized for large datasets, and successfully implemented in many other similar promoter annotation projects. Nevertheless, Davuluri's lab will continue to expand GDVTK by incorporating more Java classes to handle new data structures. Similarly as we continue to update At*cis*DB periodically, we will also integrate other *cis* element databases as they become available.

Aim 3: Developing a knowledge base of regulatory networks

Rationale and overview – Regulatory networks direct the patterns of gene expression. The number of TFs and the target genes that participate in a cellular process would generate enormous number of potential regulatory network structures. The availability of genome sequences and high-throughput technologies are helping the development of computational approaches to systematically dissect transcriptional networks. We will apply statistically rigorous computational tools, including Bayesian networks and linear models, to search for the network structure that is most consistent with the experimental data. In parallel, we will do extensive literature searches to construct the known components of regulatory networks that participate in epidermal cell differentiation and flower development. We will make these networks available as a knowledgebase and part of AGRIS. We will make use of GDVTK tools to present the data in user-friendly graphical form.

Experimental design - At*cis*DB 2.0 will contain information about *Arabidopsis* TFs and their target genes. Using that information, we will prepare a data matrix (**M**) consisting of binary entries M_{ij}, where M_{ij},=1 indicates gene *i* is a target gene of TF *j* and M_{ij},=0 indicates gene *i* is not a target gene of TF *j*. For each TF *j* we'll construct its target matrix (**T**), which is a sub-matrix of **M,** containing only the rows corresponding to target genes of *j* and all the columns. Using matrix **T,** we will build various networks of transcriptional regulation similar to the ones described [30; 37]. These networks could be:

a) *autoregulatory loop* - TF*i* acting on its own promoter, or *i* is its own target gene
b) *feedforward loop* - TFi_1 regulates another TFi_2 and TFi_1 and TFi_2 together regulate a target gene
c) *multi-component loop* - a closed regulatory network, in which TFi_1 regulates TFi_2 and viceversa
d) *single-input modules* - a single TF uniquely binds to two or more target gene promoters
e) *multi-input modules* - a set of TFs bind together to a set of target gene promoters
f) *regulator chain* - a chain of three or more TFs, in which the first one binds to the promoter of second TF, the second one binds to the third, etc., until the last target gene in the chain is not a TF.

We will also explore statistical tools (Bayesian networks [20] and graph theory [20]) to model the networks described above. The networks hence constructed will be part of AGRIS, and will be available in graphical form as a searchable database. All the TFs and their target genes involved in each network will be inter-linked with the corresponding entries in At*cis*DB and AtTFDB. GDVTK will be used for building and web-presentation of the network database. All the computational studies described in this aim will be carried out in the Davuluri and Grotewold labs, using the facilities available (see Management Plan Appendix A-2).

Expected outcomes and solutions to possible pitfalls – A publicly available knowledge base of regulatory networks linking flower development and epidermal cell differentiation with other cellular processes through TF/target gene connections is a major outcome of this Aim. The establishment of these networks will involve the integration of already available data with the experimental and computational analyses described here. We expect that the deduced networks will serve as powerful hypothesis generating engines to drive the design of the next generation of experiments in *Arabidopsis* and other plants. While the matrix will initially be binary, we expect that, as we start accumulating information, we will be able to integrate "time" as an additional dimension. This will permit us to incorporate cases like a TF activating different sets of genes in different tissues, or TFs expressed over narrow time windows. Data on physical interactions between TFs (primarily obtained from the literature) will also be integrated into these analyses.

Summary – The studies proposed here are aimed at starting to establish the regulatory networks that govern the expression of *Arabidopsis* genes. By combining powerful computational approaches with the experimental identification of direct target genes for a subset of TFs involved in well characterized genetic processes and representing major TF families present in *Arabidopsis*, we will establish an interactive and user-friendly database of *Arabidopsis cis*-regulatory elements (At*cis*DB). Combining this with the growing knowledge on co-regulated genes will allow us to start building a map of regulatory networks, using two well described cellular processes as centers. We expect the network to expand as direct targets for additional TFs are identified.

REFERENCES CITED

1.**Andrau, J.C., Van Oevelen, C.J., Van Teeffelen, H.A., Weil, P.A., Holstege, F.C., and Timmers, H.T.** (2002). Mot1p is essential for TBP recruitment to selected promoters during in vivo gene activation. EMBO J **21,** 5173-5183.

2.**Bailey, T.L., and Elkan, C.** (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2,** 28-36.

3.**Bernhardt, C., Lee, M.M., Gonzalez, A., Zhang, F., Lloyd, A., and Schiefelbein, J.** (2003). The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root. Development **130,** 6431-6439.

4.**Birnbaum, K., Benfey, P.N., and Shasha, D.E.** (2001). cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships. Genome Res **11,** 1567-1573.

5.**Blackwell, T.K., and Walker, A.K.** (2002). Getting the right dose of repression. Genes & Dev. **16,** 769-772.

6.**Braun, E.L., Matulnik, T., Dias, A., and Grotewold, E.** (2001). Transcription factors and metabolic engineering: Novel applications for ancient tools. Recent. Adv. Phytochem. **35,** 79-109.

7.**Breiman, L.** (1996). Bagging predictors. Machine Learn. **24,** 123-140.

8.**Chen, W., Provart, N.J., Glazebrook, J., Katagiri, F., Chang, H.S., 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.

9.**Cheong, Y.H., Chang, H.S., Gupta, R., Wang, X., Zhu, T., and Luan, S.** (2002). Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis. Plant Physiol. **129,** 661-677.

10.**Colinas, J., Birnbaum, K., and Benfey, P.N.** (2002). Using cauliflower to find conserved non-coding regions in Arabidopsis. Plant Physiol. **129,** 451-454.

11.**Davidson, E.H.** (2001). Gene regulatory functions in development. In Genomic Regulatory Systems: Development and Evolution (San Diego: Academic Press), pp. 11-23.

12.**Decary, S., Decesse, J.T., Ogryzko, V., Reed, J.C., Naguibneva, I., Harel-Bellan, A., and Cremisi, C.E.** (2002). The retinoblastoma protein binds the promoter of the survival gene bcl- 2 and regulates its transcription in epithelial cells through transcription factor AP-2. Mol Cell Biol **22,** 7877-7888.

13.**Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., and Frazer, K.A.** (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res **10,** 1304-1306.

14.**Eulgem, T., Rushton, P.J., Robatzek, S., and Somssich, I.E.** (2000). The WRKY superfamily of plant transcription factors. Trends Plant Sci **5,** 199-206.

15. **Frech, K., Herrmann, G., and Werner, T.** (1993). Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. Nucleic Acids Res **21,** 1655-1664.
16. **Friedman, J.H.** (1999). Stochastic gradient boosting (Stanford, CA: Dept. of Statistics, Stanford University).
17. **Gross, C., Kelleher, M., Iyer, V.R., Brown, P.O., and Winge, D.R.** (2000). Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. J. Biol. Chem. **275,** 32310-32316.
18. **Grotewold, E., Drummond, B.J., Bowen, B., and Peterson, T.** (1994). The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell **76,** 543-553.
19. **Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J.A., and Kay, S.A.** (2000). Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. Science **290,** 2110-2113.
20. **Heckerman, D.** (1997). A tutorial on learning with Bayesian networks. Data Mining Knowl. Disc. **1,** 79-119.
21. **Hennig, L., Menges, M., Murray, J.A.H., and Gruissem, W.** (2004). *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. Plant Mol. Biol. **0,** 1-9.
22. **Honma, T., and Goto, K.** (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature **409,** 525-529.
23. **Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M.** (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol **296,** 1205-1214.
24. **Jack, T.** (2001). Relearning our ABCs: new twists on an old model. Trends Plant Sci **6,** 310-316.
25. **Jack, T.** (2002). New members of the floral organ identity AGAMOUS pathway. Trends Plant Sci **7,** 286-287.
26. **Johnson, C., Boden, E., Desai, M., Pascuzzi, P., and Arias, J.** (2001). In vivo target promoter-binding activities of a xenobiotic stress- activated TGA factor. Plant J **28,** 237-243.
27. **Kellogg, E.A.** (2001). Root hairs, trichomes and the evolution of duplicate genes. Trends Plant Sci **6,** 550-552.
28. **Klok, E.J., Wilson, I.W., Wilson, D., Chapman, S.C., Ewing, R.M., Somerville, S.C., Peacock, W.J., Dolferus, R., and Dennis, E.S.** (2002). Expression profile analysis of the low-oxygen response in *Arabidopsis* root cultures. Plant Cell **14,** 2481-2494.
29. **Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C.** (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262,** 208-214.
30. **Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A.** (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298,** 799-804.
31. **Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P., and Rombauts, S.** (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res **30,** 325-327.
32. **Li, G., and Hall, T.C.** (1999). Footprinting *in vivo* reveals changing profiles of multiple factor interactions. Plant J. **18,** 633-641.

33.**Liao, J.C., Boscolo, R., Yang, Y.-L., Tran, L.M., Sabatti, C., and Roychowdhury, V.P.** (2003). Network component analysis: reconstruction of regulatory signals in biological systems. PNAS **100,** 15522-15527.

34.**Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I.** (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics **16,** 1046-1047.

35.**Meissner, R.C., Jin, H., Cominelli, E., Denekamp, M., Fuertes, A., Greco, R., Kranz, H.D., Penfield, S., Petroni, K., Urzainqui, A., Martin, C., Paz-Ares, J., Smeekens, S., Tonelli, C., Weisshaar, B., Baumann, E., Klimyuk, V., Marillonnet, S., Patel, K., Speulman, E., Tissier, A.F., Bouchez, D., Jones, J.J., Pereira, A., Wisman, E., and et al.** (1999). Function search in a large transcription factor gene family in Arabidopsis: assessing the potential of reverse genetics to identify insertional mutations in R2R3 MYB genes. Plant Cell **11,** 1827-1840.

36.**Memelink, J., Menke, F.L.W., van der Fits, L., and Kijne, J.W.** (2000). Transcriptional regulators to modify secondary metabolism. In Metabolic engineering of plant secondary metabolism, R. Verpoorte and A.W. Alfermann, eds (Dordrecht: Kluwer Academic Publishers), pp. 111-125.

37.**Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.** (2002). Network motifs: simple building blocks of complex networks. Science **298,** 824-827.

38.**Mueller, L.A., Zhang, P., and Rhee, S.Y.** (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol. **132,** 453-460.

39.**Murre, C., McCaw, P.S., and Baltimore, D.** (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell **56,** 777-783.

40.**Ole Steffens, N., Galuschka, C., Schindler, M., Bulow, L., and Hehl, R.** (2004). AtaMap: an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. Nucl. Acids Res. **32,** D368-D372.

41.**Orlando, V., and Paro, R.** (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. Cell **75,** 1187-1198.

42.**Pabo, C.O., and Sauer, R.T.** (1992). Transcription factors: Structural families and principles of DNA recognition. Annu. Rev. Biochem. **61,** 1053-1095.

43.**Paul, A.-L., and Ferl, R.J.** (1991). In Vivo footprinting reveals unique *cis*-elements and different modes of hypoxic induction in maize *Adh1* and *Adh2*. Plant Cell **3,** 159-168.

44.**Paul, A.-L., and Ferl, R.J.** (1994). *In vivo* footprinting identifies and activating element of the maize *Adh2* promoter specific for root and vascular tissues. Plant J. **5,** 523-533.

45.**Pooma, W., Gersos, C., and Grotewold, E.** (2002). Transposon insertions in the promoter of the *Zea mays a1* gene differentially affect transcription by the Myb factors P and C1. Genetics **161,** 793-801.

46.**Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., and Yu, G.** (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290,** 2105-2110.

47.**Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M.** (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol **16,** 939-945.

48.**Roux, M.M., Pain, A., Klimpel, K.R., and Dhar, A.K.** (2002). The lipopolysaccharide and beta-1,3-glucan binding protein gene is upregulated in white spot virus-infected shrimp (Penaeus stylirostris). J Virol **76,** 7140-7149.

49.**Sablowski, R.W.M., and Meyerowitz, E.M.** (1998). A Homolog of *NO APICAL MERISTEM* is an immediate target of the floral homeotic genes *APETALA3/PISTILLATA*. Cell **92,** 93-103.

50.**Sablowski, R.W.M., Moyano, E., Culianez-Macia, F.A., Schuch, W., Martin, C., and Bevan, M.** (1994). A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. EMBO J. **13,** 128-137.

51.**Sainz, M.B., Grotewold, E., and Chandler, V.L.** (1997). Evidence for direct activation of an anthocyanin promoter by the maize C1 protein and comparison of DNA binding by related Myb domain proteins. Plant Cell **9,** 611-625.

52.**Schiefelbein, J.W.** (2000). Constructing a plant cell. The genetic control of root hair development. Plant Physiol. **124,** 1525-1531.

53.**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,** 279-292.

54.**Sessions, A., Weigel, D., and Yanofsky, M.F.** (1999). The *Arabidopsis thaliana MERISTEM LAYER 1* promoter specifies epidermal expression in meristems and young primordia. Plant J. **20,** 259-263.

55.**Shin, B., Choi, G., Yi, H., Yang, S., Cho, I., Kim, J., Lee, S., Paek, N.C., Kim, J.H., and Song, P.S.** (2002). AtMYB21, a gene encoding a flower-specific transcription factor, is regulated by COP1. Plant J **30,** 23-32.

56.**Solomon, M.J., Larsen, P.L., and Varshavsky, A.** (1988). Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell **53,** 937-947.

57.**Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B.** (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell **9,** 3273-3297.

58.**Spelt, C., Quattrocchio, F., Mol, J.N., and Koes, R.** (2000). Anthocyanin1 of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural anthocyanin genes. Plant Cell **12,** 1619-1632.

59.**Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R.** (1999). Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. Nucleic Acids Res **27,** 3899-3910.

60.**Stormo, G.D.** (2000). DNA binding sites: representation and discovery. Bioinformatics **16,** 16-23.

61.**Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M.** (1999). Systematic determination of genetic network architecture. Nat Genet **22,** 281-285.

62.**Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y.** (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics **17,** 1113-1122.

63.**Williams, C.E., and Grotewold, E.** (1997). Differences between plant and animal Myb domains are fundamental for DNA-binding, and chimeric Myb domains have novel DNA-binding specificities. J. Biol. Chem. **272,** 563-571.

64.**Yoon, H., Liyanarachchi, S., Wright, F.A., Davuluri, R., Lockman, J.C., De La Chapelle, A., and Pellegata, N.S.** (2002). Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and identifies CSPG2 as a direct target of p53. Proc Natl Acad Sci U S A **99,** 15632-15637.

65.**Yuh, C.-H., Bolouri, H., and Davidson, E.H.** (1998). Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. Science **279,** 1896-1902.