AN INTEGRATED BIOINFORMATICS APPROACH TO THE DISCOVERY OF *CIS*-REGULATORY ELEMENTS INVOLVED IN PLANT GRAVITROPIC SIGNAL TRANSDUCTION

X. Liang,* K. Shen,**,*** J. Lichtenberg,* S.E. Wyatt,**,*** and L.R. Welch*,***,****

Abstract

Gravity is a common stimulus affecting plant growth and development, from seed germination to positioning of flowers for pollination and seeds for dispersal. Classic models of plant gravitropism have revolved around biophysical perception of the gravity stimulus and the effects of plant growth regulators on the growth response. Transcriptional regulation of the gravitropic mechanism has been largely ignored. The aim of this experiment is to identify putative regulatory functional elements, including transcription factor binding sites and *cis*-regulatory modules involved in gravitropic signal transduction.

In this article, we detailed a strategy to identify putative *cis*-regulatory elements by analyzing gene expression data from microarray experiments. Genes involved in the gravitropic perception–response pathway were identified based on their changes in expression level after gravity stimulation. Genes were clustered according to their expression patterns (transcriptional regulation profiles), and gene promoter were analyzed using genomics regulatory analysis software to identify candidate *cis*-regulatory elements and *cis*-regulatory modules.

Analysis of the microarray data indicated that 154 genes were involved in the gravitropic response. The genes were grouped into 9 clusters based on expression profile similarities. An analysis of the promoters of the 154 genes resulted in the identification of 32 putative regulatory elements and 55 putative regulatory modules. Some of the elements are associated with individual clusters and other elements are associated with multiple clusters, potentially indicating elements involved in specific and in general gravitropic response processes, respectively.

Key Words

Gravitropism, motif discovery, regulatory genomics, module discovery, gene expression

* School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio, USA

** Department of Environmental and Plant Biology, Ohio University, Athens, Ohio, USA

*** Molecular and Cellular Biology Program, Ohio University, Athens, Ohio, USA

**** Biomedical Engineering Program, Ohio University, Athens, Ohio, USA; e-mail: {xl187007,ks280007,lichtenj,wyatts,welch}@ohio.edu

(paper no. 210-1013)

1. Introduction

Gravitropism refers to the plant growth response in the Earth's gravity field. Plants use gravity to control everything from the direction of growth of an emerging seedling, to the positioning of plant organs (stems, branches, primary and lateral roots, flowers and seed pods). Simplistically, gravitropism has been broken down into three steps: perception, signal transduction and growth response [1]. Commonly, the gravitational stimulus is believed to be sensed by dense organelles: statoliths. Statoliths, often amyloplasts full of starch, are located in specific cell layers (the columella cells of the root cap and in the starch sheath surrounding the vascular tissue in the shoots) that are known to control the gravitropic response (reviewed in [2]). When a plant is re-positioned relative to the gravity field, the statoliths "settle" on the new physical "bottom" of the cell, initiating a signal transduction cascade. Although in existence for over 100 years, the Starch–Statolith hypothesis does not tell the full story. Mutants lacking starch-filled statoliths also respond to gravity, indicating that additional mechanisms must be involved [3, 4]. The signal transduction phase of the gravitropic pathway has been dominated by the movement and redistribution of the plant growth regulator auxin. However, how statolith sedimentation directs redistribution of auxin is still largely unknown. More recently, research has identified other potential aspects of the signal transduction phase (for review see [5]). Potential roles for cytoplasmic pH [6], cytoskeletal rearrangements (for review see [7]), inositol 1, 4, 5-triphosphate [8, 9], and reactive oxygen species [10, 11] have all been proposed, but how these fit into the pathway is not clear. A plant's response to the gravitropic stimulus is a growth response that results in organ curvature to maintain the organ's original position to the gravity field. This curvature is the result of the different elongation of the cells within the elongation zone. The response leads to cell expansion, cell wall synthesis, and other physiologic

events (for review see [12]).

Most gravitropic experiments have focused on the biophysical movement of statoliths, imaging of proton and calcium gradients, or the physiological interactions and responses involved with auxin and the growth response. Physiological experiments and mutant analyses have provided the majority of the data. Little has been done on a genome scale to identify components of the mechanism that might be under transcriptional control.

The analysis presented here is designed to identify potential *cis*-regulatory elements that are controlled by events of gravitropic signal transduction. A bioinformatics approach to identify *cis*-regulatory elements from gravitropic microarray data is employed. The approach employed a pipeline for mining microarray and genome sequence data to identify regulatory features by searching for functional elements that are significantly over- and under-represented in the DNA regulatory regions of genes clustered based on their transcription profiles. The raw data (the expression values from a gene expression microarray experiment), analyzed here, were obtained from Kimbrough *et al.* [13]. Briefly, 7-day-old *Arabidopsis thaliana* seedlings were either rotated 135° to provide a gravity stimulus or oscillated gently for 5 s to control for the mechanical movement of rotation. The RNA was extracted from the root tips from each group at six time points after treatment: 0, 2, 5, 15, 30, and 60 min. The RNA was amplified and hybridized to *Arabidopsis* ATH1 GeneChips®. The resulting microarray data were analyzed, and the discovered genes were clustered based on their expression profiles. The clusters were then subjected to a motif discovery analysis pipeline to determine the interesting motifs and modules shared among the promoter regions of the genes involved in these clusters.

The remainder of the article is organized as follows: In Secton 2, the results of the microarray experiments are presented together with the analysis of the expression-based clusters derived from said analysis. In Section 3, the methods employed during the analysis presented here are discussed in detail. Finally, in Section 4, conclusions, derived from the results and their discussion, are presented.

2. Results

Analysis of the microarray data (obtained from Kimbrough *et al.* [13]) identified 154 genes significantly up- or down-regulated in the gravitropic response. The genes were grouped into 9 clusters based on similarities in their expression profiles across the time course. Promoters for all but 4 of the genes were available in a regulatory database, so a total of 150 genes, spread across the different clusters, were analyzed. For each cluster, the number of genes, and groupings of the clustered genes with similar GO Terms are shown in Table 1(a)–(i). The same process was conducted for the entire gene list (Table 1(j)). Reviewing the result, several clusters (Cluster 3, 5, and 7) share common functionality focused on the response to outside stresses. Others (Cluster 2, 6, and 8) are annotated with membrane localization. Cluster 1 is annotated as being related to signal transduction and cell communication, a feature to be expected during gravitropic response. Finally, Cluster 9 is associated with protein binding in general and heat shock protein binding in particular, which is of interest in general stress response and to be expected during gravitropic response as well. A strong relation to stress responses was strongly suggested based on the GO clusters generated for the complete gene list (Table 1(j)).

A detailed analysis of the promoter regions of the clustered genes identified *regulatory genomic signatures* [14, 15], i.e., putative *cis*-regulatory elements and modules associated with gravitropic control of transcription.

Genes which have similar expression patterns typically share the same regulatory element (word) in their promoter regions. In eukaryotes, regulatory elements usually consist of 6–10 nucleotides. Thus, for each cluster, the top five statistically over-represented 6-mers are presented in Table 2. The words, which represent putative regulatory elements, were sorted in descending order by the $S\ln(SEs)$ score (S is the number of sequences in which a word occurred and Es is the number of sequences in which the word was expected to occur). The table also shows, for each word, the number of occurrences (O), the reverse complement, the rank of the reverse compliment in the sorted word list, whether it is a palindrome, and the p-value detailing the significance of the word based on the computed number of occurrences.

Because a regulatory element may vary, a *motif* is often used to represent its variations. Interesting words in each gene cluster were selected for *word-based clustering*, wherein a motif was constructed from all words that were similar to (i.e., are within a hamming distance of 1 of) the *top* words. Motif logos for the top two words from each cluster are presented in Table 3.

Often, gene regulation is controlled by multiple regulatory elements (called a *cis*-regulatory module) that work in conjunction. To identify putative *cis*-regulatory modules, a module discovery algorithm was applied to the top 25 statistically over-represented words. Table 4 shows the

Table 1(a)
GO Analysis for Cluster 1

Cluster 1						
Number of genes	Group	GO	Genes	Group count	Total count	Description
14	1	GO:0007165	at1g02340.1	2	996	Signal transduction
		GO:0007154	at3g07890.1	2	1098	Cell communication

Table 1(b)
GO Analysis for Cluster 2

Cluster 2						
Number of genes	Group	GO	Genes	Group count	Total count	Description
18	1	GO:0005507	at5g37990.1	3	118	Copper ion binding
		GO:0016020	at2g07695.1	10	7444	Membrane
		GO:0012505	at4g39830.1	7	4603	Endomembrane system
		GO:0031225	at3g60270.1	2	240	Anchored to membrane
			at4g11190.1			
			at2g33050.1			
			at3g16530.1			
			at1g78460.1			
			at1g79680.1			
			at1g70990.1			

Table 1(c)
GO Analysis for Cluster 3

Cluster 3						
Number of genes	Group	GO	Genes	Group count	Total count	Description
16	1	GO:0004601	at5g15180.1	2	120	Peroxidase activity
		GO:0016684	at3g03670.1	2	120	Oxidoreductase activity, acting on peroxide as acceptor
		GO:0006979		2	218	Response to oxidative stress

Table 1(d)
GO Analysis for Cluster 4

Cluster 4 (using the p value of 0.2)						
Number of genes	Group	GO	Genes	Group count	Total count	Description
28	1	GO:0008289	at4g33550.1	2	163	Lipid binding
			at5g59320.1			
	2	GO:0022402	at2g32590.1	2	185	Cell cycle process
		GO:0007049	at3g11520.1	2	208	Cell cycle

putative regulatory modules, which consist of statistically over-represented word pairs.

The selection process for the most interesting putative regulatory words (Table 5) and modules (Table 7) produced short lists. The list of significant words as putative regulatory elements and the putative elements making up the predicted modules are checked against the currently known transcription factor binding sites contained in the Arabidopsis Gene Regulatory Information Server (AGRIS) database [16, 17] (Tables 6 and 8), and TRANSFAC [18] and JASPAR [19] databases (Table 9). Among the significant words (32 in total), five of them are known and reported in AGRIS, and 14 can be found in the TRANSFAC and JASPAR databases.

Interestingly, several words are similar, i.e., they have the same core which may play an important regulatory role. Specifically, TAAGCC and TCTAAG have the same core of TAAG. TAACTC, TCTAAC, and TGTAAC not

Table 1(e)
GO Analysis for Cluster 5

Cluster 5						
Number of genes	Group	GO	Genes	Group count	Total count	Description
21	1	GO:0006869	at5g23400.1	3	117	Lipid transport
		GO:0008289	at5g25610.1	3	163	Lipid binding
		GO:0012505	at5g53870.1	9	4603	Endomembrane system
		GO:0016020	at4g12470.1	10	7444	Membrane
		GO:0006952	at3g20820.1	3	683	Defense response
			at1g62510.1			
			at3g22120.1			
			at2g30540.1			
			at5g39110.1			
			at3g24510.1			

Table 1(f)
GO Analysis for Cluster 6

Cluster 6						
Number of genes	Group	GO	Genes	Group count	Total count	Description
17	1	GO:0006118	at4g37370.1	3	681	Transport
		GO:0006091	at4g31970.1	3	829	Generation of precursor metabolites and energy
		GO:0019825	at1g26380.1	2	248	Oxygen binding
	2	GO:0005783	at1g09080.1	2		Endoplasmic reticulum
			at4g37370.1			
	3	GO:0006869	at4g36670.1	2	117	Lipid transport
		GO:0012505	at2g16005.1	7	4603	Endomembrane system
		GO:0008289	at4g31970.1	2	163	Lipid binding
		GO:0016020	at4g29020.1	8	7444	Membrane
		GO:0006810	at1g26380.1	4	1952	Transport
		GO:0051234	at5g01870.1	4	1971	Establishment of localization
		GO:0051179	at1g12090.1	4	1981	Localization
			at3g15980.1			

only share the common core TAAC, but TAACTC and TCTAAC are also associated with Myb2 binding site motif and Myb Recognition Element (MRE) motif in Chalcone Synthase (CHS) respectively. Myb2 and MRE are known to interact [20], while Myb2 is furthermore known to be involved in the regulation of salt tolerance in *Arabidopsis thaliana* [21]. CCTTTC and ACCTTC share CCTT, with CCTTTC being associated with the CArG2 motif in APETALA3 (AP3). GGATAC and CAATAC share ATAC and are associated with the GATA box as well as AGATAG and AGATCA, which share AGAT, also an integral part of the GATA family. Furthermore, three pairs of words have a hamming distance of 1: CGAACC and CCAACC; TCTAAC and TGTAAC; GTATCC and GTATCT.

Table 1(g)
GO Analysis of Cluster 7

Cluster 7						
Number of genes	Group	GO	Genes	Group count	Total count	Description
7	1	GO:0004601	at5g39580.1	3	120	Peroxidase activity
		GO:0016684	at1g20620.1	3	120	Oxidoreductase activity, acting on peroxide as acceptor
		GO:0050832	at2g37130.1	2	85	Defense response to fungus
		GO:0009620	at4g21850.1	2	134	Response to fungus
		GO:0016491		4	1507	Oxidoreductase activity
		GO:0006952		2	683	Defense response
	2	GO:0008289	at5g39580.1	3	163	Lipid binding
		GO:0006869	at4g12550.1	2	117	Lipid transport
		GO:0016020	at5g57220.1	7	7444	Membrane
			at2g37130.1			
			at4g12510.1			
			at2g38530.1			
			at2g05540.1			
	3	GO:0051707	at5g39580.1	3	482	Response to other organism
		GO:0009607	at2g37130.1	3	525	Response to biotic stimulus
		GO:0051704	at4g39950.1	3	547	Multi-organism process
	4	GO:0006091	at5g57220.1	3	829	Generation of precursor metabolites and energy
		GO:0006118	at2g07698.1	2	681	
			at2g46750.1			
	5	GO:0019825	at5g57220.1	2	248	Oxygen binding
			at4g39950.1			
	6	GO:0012505	at5g39580.1	6	4603	Endomembrane system
			at4g12550.1			
			at5g57220.1			
			at4g12510.1			
			at2g37130.1			
			at2g05540.1			

3. Methods

3.1 Microarray Analysis

The raw microarray data were analyzed using Bioconductor [22], an R package suite for microarray analysis. First, the data were normalized using an un-scaled standard error (NUSE) plot for quality assessment:

$$\mathrm{NUSE}(\theta_{\mathrm{gi}}) = \frac{\mathrm{SE}(\theta_{\mathrm{gi}})}{\mathrm{med_i}(\mathrm{SE}(\theta_{\mathrm{gi}}))} \qquad (1)$$

Expression values are corrected for background noise using GCRMA [23] (Fig. 1). Rank Product, a non-parametric method [24], was used to identify the differentially expressed genes in the data set. Differentially expressed genes were then selected based on the false positive prediction. A p-value of 0.15 was chosen and resulted in a list of 154 genes.

Once the differentially expressed genes were identified, they were clustered based on their transcriptional expression pattern. Two criteria were used to cluster the genes:

Table 1(h)
GO Analysis of Cluster 8

Cluster 8						
Number of genes	Group	GO	Genes	Group count	Total count	Description
16	1	GO:0016020	at1g70710.1	12	7444	Membrane
		GO:0031224	at5g23840.1	4	779	Intrinsic to membrane
		GO:0044425	at3g43720.1	4	1212	Membrane part
		GO:0012505	at3g28550.1	7	4603	Endomembrane system
		GO:0005623	at1g06120.1	13	15514	Cell
		GO:0044464	at3g06460.1	13	15514	Cell part
		GO:0031225	at3g20570.1	2	240	Anchored to membrane
			at2g39510.1			
			at5g12940.1			
			at5g49770.1			
			at3g04320.1			
			at3g05020.1			
			at1g47600.1			

Table 1(i)
GO Analysis of Cluster 9

Cluster 9						
Number of genes	Group	GO	Genes	Group count	Total count	Description
5	1	GO:0031072	at3g30450.1	2	148	Heat shock protein binding
		GO:0005515	at2g17060.1	3	2275	Protein binding
			at2g14140.1			

within cluster similarity and *between cluster dissimilarity*. The *Point Accepted Mutation* matrix ...[25] was used since the objective was to partition genes into several groups rather than to find the hierarchical structure. Initially, genes were clustered into 10, 15, and 20 groups. Clustering in 15 or 20 groups resulted in less than 10 genes per cluster, which is not ideal for further analysis because it does not provide large enough data sets to detect statistically interesting words. Thus, 10 clusters proved to be the most appropriate. For 10 clusters the algorithm produced an empty cluster, which is subsequently discarded.

3.2 GO Analysis

To provide insight into the GO terms associated with the genes of the nine clusters as well as the complete set of genes subjected to the motif discovery analysis, a GOstat [26] analysis was conducted. GOstat supported the clustering of extracted GO terms and the associated genes, allowing functional similarity assessments of the genes.

The GOstat analysis was executed for the 9 gene clusters as well as for the whole gene list containing 150 elements, since 4 of them cannot be associated with actual gene products. A p-value, which evaluates the matched level of genes and corresponding GO items, was set to 0.1 as a threshold. As no output can be generated for cluster 4 with the p-value set to 0.1, it was to be adjusted to 0.2.

3.3 Statistically Over-Represented Words

For each cluster of genes, the promoter regions were retrieved from AGRIS67. All words of the specified length 6, which were present in the promoters, were enumerated. The expected number of occurrences for each word was computed using an order 4 Markov model (the method is described in [14]). Equations (2) and (3) show, respectively, the equations that were used to compute the expected number of occurrences and the expected number of sequences hit, for each word w, with p_w being the probability of the word w, l_i being the length of sequence i (out of a total of m sequences) and v being the length of w.

$$Eo(w) = \sum_{i=1}^{m} (l_i - v + 1)p_w \qquad (2)$$

Table 1(j)
GO Analysis of the Entire Set of 150 Genes

Group	GO	Genes			Group count	Total count	Description
1	GO:0012505	at1g74500.1	at1g75780.1	at2g33050.1	49	4603	Endomembrane system
	GO:0016020	at3g08970.1	at3g60270.1	at3g15980.1	65	7444	Membrane
	GO:0006869	at4g33550.1	at3g07890.1	at1g26380.1	9	117	Lipid transport
	GO:0044464	at2g07696.1	at5g06720.1	at5g25610.1	81	15514	Cell part
	GO:0005623	at2g38530.1	at1g06120.1	at1g77210.1	81	15514	Cell
	GO:0005507	at1g18830.1	at5g12940.1	at3g16530.1	5	118	Copper ion binding
	GO:0031225	at5g15180.1	at3g04320.1	at5g39110.1	6	240	Anchored to membrane
	GO:0043170	at2g32590.1	at3g05020.1	at1g09080.1	10	6920	Macromolecule metabolic process
	GO:0005622	at5g47990.1	at1g47600.1	at4g36670.1	18	9003	Intracellular
	GO:0050832	at5g01870.1	at3g47380.1	at3g61890.1	3	85	Defense response to fungus
	GO:0044424	at4g12550.1	at2g16005.1	at3g03670.1	17	8514	Intracellular part
	GO:0043283	at5g65600.1	at5g26260.1	at5g23840.1	6	4744	Biopolymer metabolic process
	GO:0043227	at1g78460.1	at5g57220.1	at3g43720.1	13	7166	Membrane-bounded organelle
	GO:0043231	at3g06460.1	at5g11210.1	at2g39510.1	13	7164	Intracellular membrane-bounded organelle
		at5g49770.1	at2g05540.1	at3g11520.1			
		at4g29020.1	at1g20620.1	at5g04160.1			
		at2g04070.1	at5g23400.1	at4g31970.1			
		at3g24510.1	at5g39580.1	at3g22120.1			
		at3g29970.1	at4g39830.1	at1g70990.1			
		at3g20570.1	at2g07698.1	at1g79680.1			
		at4g12510.1	at2g17060.1	at4g11190.1			
		at5g37990.1	at1g12090.1	at4g12470.1			
		at2g07695.1	at4g17785.1	at1g70710.1			
		at3g20820.1	at4g30270.1	at1g61500.1			
		at5g64510.1	at5g53870.1	at3g28550.1			
		at4g37370.1	at1g62510.1	at4g28100.1			
		at4g28710.1	at2g37130.1	at2g30540.1			
2	GO:0008289	at4g33550.1	at1g12090.1	at3g43720.1	11	163	Lipid binding
		at3g22120.1	at5g01870.1	at4g12550.1			
		at4g12510.1	at4g12470.1	at1g62510.1			
		at2g38530.1	at5g59320.1				
3	GO:0006118	at3g60270.1	at2g07695.1	at5g53870.1	13	681	
	GO:0006091	at2g45550.1	at1g26410.1	at2g46750.1	14	829	Generation of precursor metabolites and energy

(*Continued*)

Table 1(j)
(*Continued*)

Group	GO	Genes			Group count	Total count	Description
		at4g37370.1	at5g47990.1	at2g07698.1			
		at4g31970.1	at5g57220.1	at3g20570.1			
		at2g30540.1	at1g26380.1				
4	GO:0004601	at1g20620.1	at2g07695.1	at1g64590.1	6	120	Peroxidase activity
	GO:0016684	at5g39580.1	at1g66800.1	at5g06720.1	6	120	Oxidoreductase activity, acting on peroxide as acceptor
	GO:0016491	at1g26410.1	at5g15180.1	at1g06120.1	16	1507	Oxidoreductase activity
		at4g39830.1	at2g37130.1	at4g21850.1			
		at3g03670.1	at2g30540.1	at1g26380.1			
		at2g37540.1					
5	GO:0019825	at5g57220.1	at4g37370.1	at4g31970.1	6	248	Oxygen binding
		at2g45550.1	at5g47990.1	at4g39950.1			
6	GO:0006952	at5g23400.1	at2g17060.1	at2g33050.1	9	683	Defense response
		at5g39580.1	at4g12470.1	at2g37130.1			
		at4g11190.1	at3g20820.1	at4g23670.1			
7	GO:0044237	at1g66800.1	at1g79680.1	at4g39950.1	18	9054	Cellular metabolic process
		at1g06120.1	at1g09080.1	at1g01480.1			
		at5g38020.1	at4g11190.1	at4g17785.1			
		at5g65600.1	at2g07698.1	at1g68530.1			
		at5g49770.1	at3g61890.1	at1g61500.1			
		at1g74500.1	at3g08970.1	at2g07696.1			
8	GO:0044238	at5g49770.1	at1g66800.1	at2g07698.1	19	9160	Primary metabolic process
		at1g79680.1	at5g24210.1	at3g61890.1			
		at1g09080.1	at1g06120.1	at1g01480.1			
		at1g47600.1	at5g38020.1	at4g17785.1			
		at4g11190.1	at5g65600.1	at1g68530.1			
		at1g61500.1	at3g08970.1	at2g07696.1			
		at1g74500.1					

$$Es(w) = \sum_{j=1}^{m} (1 - (1 - p_w)^{l_j - v + 1}) \quad (3)$$

Based on these two expected values, for each word, multiple scores were computed: SlnSE score ($S \ln\left(\frac{S}{Es}\right)$ [21]) and P-value (4),

$$1 - \sum_{j-1}^{|s|} \sum_{i=0}^{l_j - v + 1} \binom{l_j - v + 1}{i} p_w^i (1 - p_w)^{l_j - v + 1 - i} \quad (4)$$

A p-value threshold (0.05) was set for choosing significant words. Among the top five words, the top two words and the words with a p-value smaller than 0.05 were selected as significant.

3.4 Word-Based Cluster

To construct motifs, top scoring words were chosen as seeds. All enumerated words that exhibited a hamming distance of 1 from the seed word were identified and used to construct motif logos based on position weight matrices using the TFBS Perl module by Lenhard and Wassermann [27].

Table 2
Top 5 Statistically Over-Represented Words of the 9 Clusters
The words were sorted in descending order based on SlnSEs score

Cluster 1									
Word	S	Es	O	Eo	SlnSEs	RevComp	Position RevComp	Palindrome	Pval
TCCCAT	8	3.4851	8	4.14894	6.64755	ATGGGA	2223	No	0.060484
TGATAC	7	2.96985	7	3.43636	6.00179	GTATCA	1502	No	0.06048
GGAACA	8	3.83552	8	4.65882	5.8811	TGTTCC	3393	No	0.100222
CGAACC	5	1.5724	5	1.69412	5.78417	GGTTCG	417	No	0.029237
TAAGCC	6	2.31415	6	2.58696	5.7163	GGCTTA	1805	No	0.048075
Cluster 2									
Word	S	Es	O	E	SlnSES	RevComp	Position RevComp	Palindrome	Pval
TAACTC	9	3.82579	10	4.64035	7.69913	GAGTTA	3383	No	0.020566
CCAACC	9	3.92801	10	4.79208	7.46182	GGTTGG	468	No	0.024886
GGCTTA	8	3.50605	9	4.17722	6.59961	TAAGCC	821	No	0.027108
GATGTA	8	3.52945	8	4.21053	6.5464	TACATC	3595	No	0.064651
TCTAAG	6	2.03414	6	2.2439	6.49013	CTTAGA	2799	No	0.027051
Cluster 3									
Word	S	Es	O	E	SlnSES	RevComp	Position RevComp	Palindrome	Pval
GCTCTA	7	2.42874	7	2.69725	7.40976	TAGAGC	1221	No	0.020465
AGATAG	10	5.36604	10	6.93243	6.22494	CTATCT	2421	No	0.162696
AGTGTT	11	6.33525	13	8.68421	6.06942	AACACT	941	No	0.102439
ACCTCT	6	2.21741	6	2.43902	5.97253	AGAGGT	291	No	0.038065
GCCATA	10	5.63704	11	7.4	5.73226	TATGGC	2398	No	0.129297
Cluster 4									
Word	S	Es	O	Eo	SlnSEs	RevComp	Position RevComp	Palindrome	Pval
AGATCA	21	14.3345	25	22.0914	8.01895	TGATCT	1025	No	0.295071
TCTAAC	17	10.8665	21	14.6437	7.60798	GTTAGA	1510	No	0.068925
GTAAGT	15	9.16075	18	11.6651	7.39683	ACTTAC	2486	No	0.050994
GTATCC	12	6.50209	13	7.64925	7.3534	GGATAC	879	No	0.04836
CCATTA	18	11.9867	21	16.8164	7.31827	TAATGG	3294	No	0.181999
Cluster 5									
Word	S	Es	O	Eo	SlnSES	RevComp	Position RevComp	Palindrome	Pval
CTCATG	14	7.89675	20	10.3143	8.01649	CATGAG	1631	No	0.004801
GTATCT	14	8.33286	16	11.0885	7.26391	AGATAC	3374	No	0.097377
AGAATC	17	11.3066	26	17.4024	6.93299	GATTCT	3819	No	0.032026
GGATAC	8	3.61972	9	4.0393	6.34436	GTATCC	3447	No	0.022551
ACTGAG	8	3.65912	8	4.0885	6.25775	CTCAGT	2739	No	0.056571

(*Continued*)

Table 2
(*Continued*)

Cluster 6									
Word	S	Es	O	Eo	SlnSES	RevComp	Position RevComp	Palindrome	Pval
CTCTCC	10	4.95692	10	5.99394	7.018	GGAGAG	3762	No	0.083491
CAATAC	12	6.98193	14	9.30677	6.49897	GTATTG	2312	No	0.09034
GCATCG	6	2.04804	6	2.2029	6.44926	CGATGC	NA	No	0.025044
GTACGT	10	5.33541	11	6.5625	6.2822	ACGTAC	959	No	0.070367
TGTAAC	13	8.21936	19	11.7345	5.95994	GTTACA	3777	No	0.031052
Cluster 7									
Word	S	Es	O	Eo	SlnSES	RevComp	Position RevComp	Palindrome	Pval
CTTTCG	9	4.32507	10	5.1831	6.59517	CGAAAG	1257	No	0.038953
ATCTGA	11	6.14726	15	8.10227	6.40078	TCAGAT	1397	No	0.019043
CCATCC	6	2.40417	6	2.64348	5.48733	GGATGG	2495	No	0.052292
GTGAAG	9	4.92507	10	6.07981	5.42597	CTTCAC	1231	No	0.08951
TAGCTT	13	8.59185	15	13.2762	5.38375	AAGCTA	3684	No	0.353323
Cluster 8									
Word	S	Es	O	E	SlnSES	RevComp	Position RevComp	Palindrome	Pval
TCATTC	13	7.76988	17	11.4459	6.69103	GAATGA	128	No	0.073941
CTTAAC	11	6.16998	13	8.21053	6.36019	GTTAAG	210	No	0.074444
GTGAAT	13	8.03022	20	12.0482	6.26258	ATTCAC	1937	No	0.022049
ATTAAC	14	8.99597	17	14.5263	6.19192	GTTAAT	2935	No	0.291188
TTACAC	12	7.28713	16	10.3911	5.98557	GTGTAA	2149	No	0.06365
Cluster 9									
Word	S	Es	O	Eo	SlnSES	RevComp	Position RevComp	Palindrome	Pval
GAGTAT	4	0.825029	4	0.913043	6.31452	ATACTC	1093	No	0.014108
GGAAGC	4	0.932334	4	1.04651	5.82544	GCTTCC	2006	No	0.021967
CATCTT	5	1.6255	6	2.01667	5.61811	AAGATG	2344	No	0.017162
CCTTTC	4	0.982098	4	1.10976	5.61743	GAAAGG	NA	No	0.026461
ACCTTC	4	0.983157	4	1.11111	5.61312	GAAGGT	892	No	0.026563

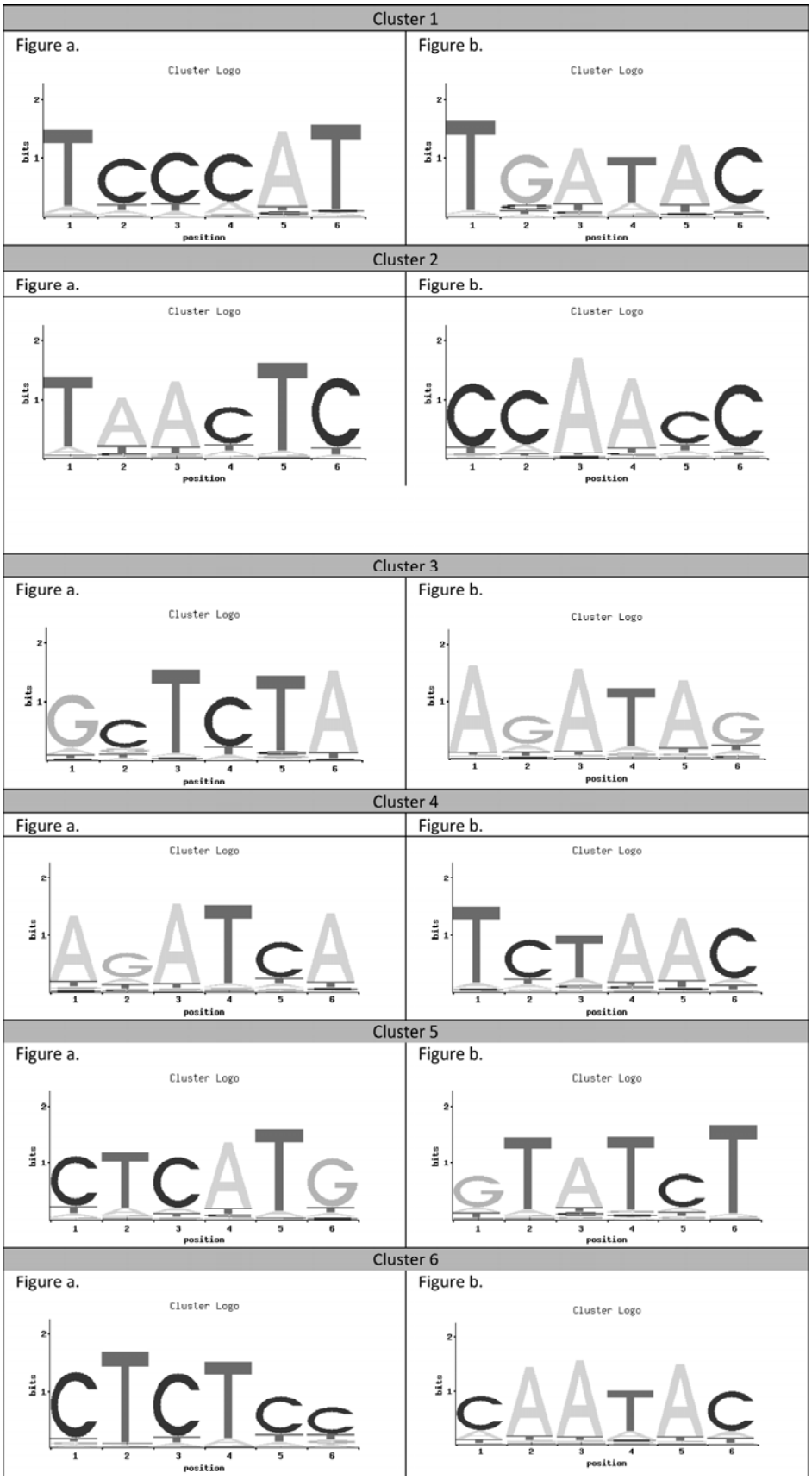
3.5 Module Discovery

To identify putative binding modules, all combinations of word pairs, and their frequencies, were enumerated. The primary statistical value that was used in module discovery was the number of sequences in which a word pair was expected to occur. This statistic is generated from the expected value of sequence hits for a single word, based on the assumption that, for each position, the probability of occurrence of each nucleotide is independent. Let Z_j be a binary random variable, defined as follows:

$$Z_j(W_k) = \begin{cases} 1, & \text{if } W_k \text{ occurs in sequence } j; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

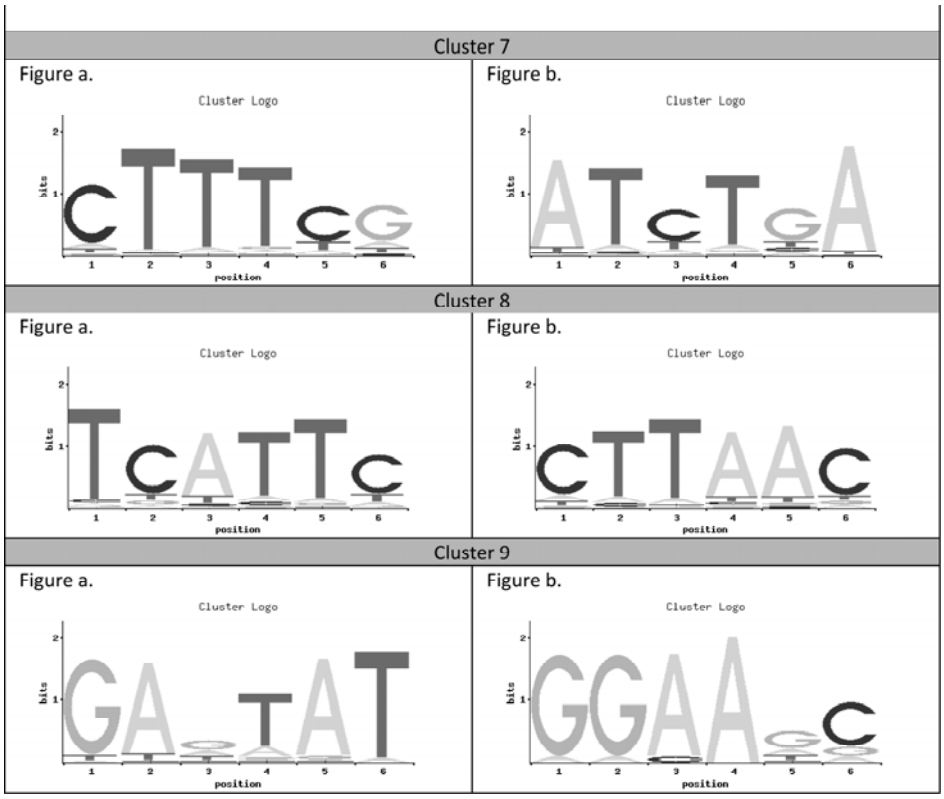
Suppose that there are m sequences, and that the length of sequence j is l_j. W_k represents a word, $|W_k|$ is the length of such word, and p_{wk} is the probability of the specific word. Let $W = \{W_1, \ldots, W_n\}$ be the set of all words enumerated from the sequences. The number of sequences in which of a set (pair) of words is expected to occur, is computed according to (6).

Table 3
Motif Logos
The table presents motif logos for top two over-represented words of each gene cluster. Subfigure *a* corresponds to the highest ranked word, while Subfigure *b* corresponds to the second highest ranked word



(*Continued*)

Table 3
(*Continued*)



$$E\left(\sum_{j=1}^{m} Z_j(W)\right) = \sum_{j=1}^{m}\prod_{k=1}^{t}\left(1-(1-p_{w_k})^{l_j-|W_k|+1}\right) \quad (6)$$

Besides the statistical scores, this project also computed gap and density for each word pair. The density distribution was evaluated and generated. Density for word pair is defined as $|W_i|+|W_j|/span$, where *span* is the total number of nucleotides covered by the word pair and the gap between the words of the pair. Note that density is in the range of 0–100%.

In this case, the top 25 statistically over-represented words of each cluster were chosen for the module discovery. All modules containing at least one word from the significant word list (Table 5) were considered as interesting modules (Table 7).

3.6 Comparison

Predicted *cis*-regulatory elements were compared with the currently available knowledge of transcription factor binding sites. The words that belong to either significant words or interesting modules were compared to the TFBS list in AGRIS. The reference information of matched motifs, including the matched binding sites and the reported publication, was reported for corresponding *cis*-regulatory elements.

In addition to the AGRIS-based lookup, a comparison of the word and module elements against the established transcription factor binding site knowledge compiled in the TRANSFAC8 and JASPAR9 databases was conducted using a modification of the approach by Jacox and Elnitski [28]. The sequences of each cluster were marked up with annotations regarding the TRANSFAC binding sites and subsequently analyzed for overlap with the discovered words. The matches between words and binding sites were used as the foundation to assess if a transcription factor would bind to a word by computing the ratio of actual matches between word and transcription factor binding sites and the total occurrences of the word. A threshold of 0.75 was applied to limit the results to a set of significant transcription factor matches.

4. Conclusions

This paper identifies regulatory genomic signatures for sets of related genes. Starting with microarray data, genomic analysis software was employed to identify putative regulatory elements and modules. Thirty two words (Table 5) were selected from the top ten over-represented words of each cluster, and are considered as putative *cis*-regulatory elements due to their statistical over-representation.

In addition to analyzing single words, this research project also identified significant pairs of over-represented words, which constitute putative regulatory modules. After comparing statistically over-represented modules and selected words, 55 modules, (Table 7), were chosen as putative *cis*-regulatory modules with the highest potential biological interest. All the modules contain a pair of words, in which at least one of them was selected from the significant word list (Table 5). Out of the 55 modules, six modules' components are both from significant words. Note that several words are shared by more than three modules. Furthermore, the average density and distance

Table 4
Top 10 Modules
The top 10 modules for each cluster, sorted in descending order based on their Sln(S/Es) score.

Cluster 1			
Module	S	Es	Sln(S/Es)
ACCAAT_GAAAAC	10	4.91741	7.09803
GGTTTG_ACCAAT	8	3.4573	6.71162
GTCAAT_GAAAAC	10	5.21406	6.51227
GACAAT_GAAAAC	10	5.21406	6.51227
GTCATT_ACCAAT	7	2.90888	6.14698
GTCAAT_ACCAAT	8	3.91875	5.70935
ATTCCG_TGATAC	5	1.77686	5.17295
ATTCCG_GTCATT	5	1.77686	5.17295
TTGTGT_GAAAAC	10	5.96994	5.15849
GTCAAT_GGAACA	7	3.3852	5.08548
Cluster 2			
Module	S	Es	Sln(S/Es)
GAACTA_CCAACC	7	1.30133	11.7777
ATCTTA_TTACTC	11	3.9052	11.3915
TAGAAT_CCAACC	8	1.95733	11.2629
CTCTAT_ACTCTA	7	1.45388	11.0017
CAAATC_AATAAC	10	3.48142	10.5514
TTACTC_TCTAAG	6	1.03571	10.54
ATGAAG_CCAACC	8	2.1875	10.3734
TTTGCA_TAACTC	8	2.32562	9.88363
TTTGCA_CCAACC	8	2.38558	9.68001
TAGAAT_GATGTA	7	1.76565	9.64176
Cluster 3			
Module	S	Es	Sln(S/Es)
ACTGAG_GCATTG	6	0.682823	13.0397
ATTAGC_GCTCTA	6	0.925382	11.2158
CCCTCC_GCTCTA	4	0.242921	11.2053
GCCATA_GCTCTA	6	1.03879	10.5222
ACTGAG_GATAGG	5	0.643251	10.2533
AGGATA_GCTCTA	6	1.1675	9.82138
AAACGC_ACCTCT	5	0.711836	9.74673
TGAGTT_GCTCTA	6	1.18413	9.73652
ACCTAC_GCATTG	5	0.718395	9.70087
CCTATA_ACTGAG	6	1.20509	9.63123

(*Continued*)

Table 4
(*Continued*)

Cluster 4			
Module	S	Es	Sln(S/Es)
GCTATA_GTATCC	10	2.61313	13.4204
GTAGAA_GTAAGT	12	3.97595	13.2557
GTATTA_TCTTAT	19	9.72642	12.7223
GTAGAA_GTATCC	10	2.84172	12.5818
TGAGTC_GTAAGT	11	3.55579	12.4225
GCAATG_GTATCC	9	2.27587	12.3738
GAAACT_TCTTAT	19	10.005	12.1857
AAGCCT_GTAAGT	11	3.63476	12.1809
CCACAA_TCTAAC	13	5.14182	12.058
TCTTAT_AGATCA	18	9.2187	12.0445
Cluster 5			
Module	S	Es	Sln(S/Es)
GTATCT_CTCATG	12	3.48888	14.8239
GCTTAT_CTCATG	12	3.52175	14.7114
GCTTAT_GTATCT	12	3.71202	14.08
GTTCAC_GGATAC	8	1.43408	13.7514
GGATAC_CTCATG	8	1.53173	13.2243
TACAAG_CTCATG	12	4.2618	12.4226
AGATGT_CTCATG	12	4.31969	12.2607
GTTTTC_CTCATG	14	5.84617	12.2258
CAAGTG_GCTTAT	11	3.78712	11.7292
GTTCAC_CTCATG	10	3.0987	11.716
Cluster 6			
Module	S	Es	Sln(S/Es)
AGTGAC_CACTCT	9	2.25362	12.4622
CGTGTC_CACAGT	6	0.826081	11.8969
CACTCT_CTCTCC	8	1.84574	11.7325
AGTGAC_CTCTCC	8	1.9313	11.37
TCATAG_CAATAC	9	2.55861	11.3198
GATCGA_GTACGT	7	1.39758	11.2782
GGTGAT_TCATAG	9	2.57787	11.2523
TGTAAC_CTCTCC	9	2.60036	11.1742
AACGAT_TGTAAC	11	4.06626	10.9469
CATTTC_TGTAAC	13	5.83075	10.4234

(*Continued*)

Table 4
(*Continued*)

Cluster 7			
Module	S	Es	Sln(S/Es)
ATCACA_CTTTCG	9	2.48752	11.5734
CGGGTA_TGCCAG	5	0.495852	11.5546
TAGCTT_CTTTCG	9	2.61704	11.1166
CAACGT_ATCTGA	9	2.77125	10.6013
CTTAAG_ATCTGA	9	2.78466	10.5579
ATCACA_ATCTGA	10	3.52527	10.4263
GTTCTC_GTCTAA	6	1.07328	10.3263
GTAACC_ATCTGA	7	1.60119	10.3261
GTCTAA_ATCTGA	8	2.23224	10.2115
Cluster 8			
Module	S	Es	Sln(S/Es)
GTGAAT_TCATTC	12	4.38729	12.0743
ATTAAC_CTTAAC	11	3.90788	11.3839
TTGCTG_ATTAAC	12	4.69791	11.2535
TCGAAC_GTCAAG	8	1.96711	11.223
ACATTG_TTACAC	11	4.0788	10.913
CCCATG_ATAGTG	7	1.49161	10.8224
TTACAC_GTGAAT	11	4.12317	10.794
ACATTG_CTTAAC	10	3.4685	10.5886
TCGAAC_TTGCTG	9	2.80362	10.4968
GTGAAT_CTTAAC	10	3.50615	10.4807
Cluster 9			
Module	S	Es	Sln(S/Es)
ACCTTC_GAGTAT	4	0.178573	12.4362
CACCGA_GGAAGC	4	0.215494	11.6845
CAACTC_CCTTTC	4	0.246191	11.1518
CCTCAC_GAGTAT	4	0.280597	10.6285
CATCTT_GAGTAT	4	0.292942	10.4563
CTGACA_GAGTAT	4	0.313277	10.1879
CATCTT_GGAAGC	4	0.330686	9.97153
CTATGT_GGAAGC	4	0.33226	9.95253
CCTCAC_ACCTTC	4	0.333838	9.93358
CGAATC_CCTTTC	4	0.335205	9.91723

Table 5
Significant Words
Selected over-represented words from each cluster's top five over-represented words. For every cluster, the top two over-represented words (as determined by the SlnSEs score) were chosen, in addition to words with p-values less than 0.05.

Cluster 1		
Word	SlnSEs	p-Value
TCCCAT	6.64755	0.060484
TGATAC	6.00179	0.06048
CGAACC	5.078417	0.029237
TAAGCC	5.07163	0.048075
Cluster 2		
Word	SlnSEs	p-Value
TAACTC	7.69913	0.020566
CCAACC	7.46182	0.024886
GGCTTA	6.59961	0.027108
TCTAAG	6.049013	0.027051
Cluster 3		
Word	SlnSEs	p-Value
GCTCTA	7.40976	0.020465
AGATAG	6.22494	0.162696
ACCTCT	5.97253	0.038065
Cluster 4		
Word	SlnSEs	p-Value
AGATCA	8.01895	0.295071
TCTAAC	7.60798	0.068925
GTATCC	7.3534	0.04836
Cluster 5		
Word	SlnSEs	p-Value
CTCATG	8.01649	0.004801
GTATCT	7.26391	0.097377
AGAATC	6.93299	0.032026
GGATAC	6.34436	0.022551
Cluster 6		
Word	SlnSEs	p-Value
CTCTCC	7.018	0.083491
CAATAC	6.49897	0.09034
GCATCG	6.44926	0.025044
TGTAAC	5.95994	0.031052
Cluster 7		
Word	SlnSEs	p-Value
CTTTCG	6.59517	0.038953
ATCTGA	6.40078	0.019043

(*Continued*)

Table 5
(*Continued*)

Cluster 8		
Word	SlnSEs	P-Value
TCATTC	6.69103	0.073941
CTTAAC	6.36019	0.074444
GTGAAT	6.36019	0.022049
Cluster 9		
Word	SlnSEs	P-Value
GAGTAT	6.31452	0.014108
GGAAGC	5.82544	0.021967
CATCTT	5.61811	0.017162
CCTTTC	5.61743	0.026461
ACCTTC	5.61312	0.026563

are associated with each module (Table 7). Note that, while most modules have the density lower than 10%, module CCTCAC_GAGTAT has a density of 39.2106%, with an average distance of 179 bps between its elements.

5. Acknowledgements

We would like to acknowledge Frank Drews, Matt Wiley, Rami Al-Ouran, Lee Nau, and Kyle Kurz for their support in the development of the software applied in this research. We acknowledge the support of the Ohio University Stocker Endowment, Ohio University's Graduate Research and Education Board (GERB), the Ohio Supercomputer Center, and the Choose Ohio First Initiative of the University System of Ohio. Additionally, salaries and research support were provided by state funds appropriated to the Ohio Plant Biotechnology Consortium through The Ohio State University, Ohio Agricultural Research and Development Center.

Table 6
AGRIS Look-up of Significant Words
This table reports the words for the specific clusters that are contained in the AGRIS database

Cluster 1					
–					
Cluster 2					
Word	SlnSEs	*P*-Value	Name	Consensus Motif	Reference
TAACTC	7.69913	0.020566	MYB2 binding site motif	TAACT(G/C)GTT	29
Cluster 3					
Word	SlnSEs	*P*-Value	Name	Consensus Motif	Reference
AGATAG	6.22494	0.162696	GATA promoter motif	[AT]GATA[GA]	30
Cluster 4					
Word	SlnSEs	*P*-Value	Name	Consensus Motif	Reference
TCTAAC	7.60798	0.068925	MRE motif in CHS	TCTAACCTACCA	31
Cluster 5					
Word	SlnSEs	*P*-Value	Name	Consensus Motif	Reference
GTATCT	7.26391	0.097377	EIL1 BS in ERF1; EIL2 BS in ERF1; EIL3 BS in ERF1	TTCAAGGGGGCATGTATCTTGAA	32
			EIN3 BS in ERF1	GGATTCAAGGGGGCATGTATCTTGAATCC	
Cluster 6					
–					
Cluster 7					
–					
Cluster 8					
–					
Cluster 9					
Word	SlnSEs	*p*-Value	Name	Consensus Motif	Reference
CCTTTC	5.61743	0.026461	CArG2 motif in AP3	CTTACCTTTCATGGATTA	33

Table 7
Significant Modules
This table presents interesting modules selected from top 10 over-represented modules of each cluster. For each module, the statistical value, SlnSE score, and several features, including density, and distance are shown in this table

Cluster 1				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
ATTCCG_TGATAC	5.17295	TGATAC	2.22%	403
Cluster 2				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
GAACTA_CCAACC	11.7777	CCAACC	2.04%	895
TAGAAT_CCAACC	11.2629	CCAACC	5.59%	422
TTACTC_TCTAAG	10.54	TCTAAG	22.72%	549
ATGAAG_CCAACC	10.3734	CCAACC	6.11%	316
TTTGCA_TAACTC	9.88363	TAACTC	7.27%	631
TTTGCA_CCAACC	9.68001	CCAACC	2.62%	753
Cluster 3				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
ATTAGC_GCTCTA	11.2158	GCTCTA	8.70%	518
CCCTCC_GCTCTA	11.2053	GCTCTA	15.40%	893
GCCATA_GCTCTA	10.5222	GCTCTA	2.56%	552
AGGATA_GCTCTA	9.82138	GCTCTA	7.72%	909
AAACGC_ACCTCT	9.74673	ACCTCT	7.16%	891
TGAGTT_GCTCTA	9.73652	GCTCTA	4.91%	673
Cluster 4				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
GCTATA_GTATCC	13.4204	GTATCC	6.10%	715
GTAGAA_GTATCC	12.5818	GTATCC	2.18%	1047
GCAATG_GTATCC	12.3738	GTATCC	2.13%	916
CCACAA_TCTAAC	12.058	TCTAAC	10.55%	463
TCTTAT_AGATCA	12.0445	GTATCT; AGATCA	2.58%	739
Cluster 5				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
GTATCT_CTCATG	14.7114	CTCATG	21.51%	543
GCTTAT_GTATCT	14.08	GTATCT	6.84%	670
GTTCAC_GGATAC	13.7514	GGATAC	2.41%	913
GGATAC_CTCATG	13.2243	CTCATG	6.46%	933
TACAAG_CTCATG	12.4226	CTCATG	3.15%	905
AGATGT_CTCATG	12.2607	CTCATG	2.90%	728
GTTTTC_CTCATG	12.2258	CTCATG	3.96%	704
GTTCAC_CTCATG	11.716	CTCATG	3.13%	727

(*Continued*)

Table 7
(*Continued*)

Cluster 6				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
CACTCT_CTCTCC	11.7325	CTCTCC	17.47%	807
AGTGAC_CTCTCC	11.37	CTCTCC	3.84%	981
TCATAG_CAATAC	11.3198	CAATAC	1.87%	1045
TGTAAC_CTCTCC	11.1742	TGTAAC; CTCTCC	1.92%	926
AACGAT_TGTAAC	10.9469	TGTAAC	1.43%	1091
CATTTC_TGTAAC	10.4234	TGTAAC	3.16%	862
Cluster 7				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
ATCACA_CTTTCG	11.5734	CTTTCG	3.70%	960
TAGCTT_CTTTCG	11.1166	CTTTCG	16.84%	1118
CAACGT_ATCTGA	10.6013	ATCTGA	2.80%	738
CTTAAG_ATCTGA	10.5579	ATCTGA	6.86%	559
ATCACA_ATCTGA	10.4263	ATCTGA	8.24%	545
GTAACC_ATCTGA	10.3261	ATCTGA	1.94%	1060
GTCTAA_ATCTGA	10.2115	ATCTGA	3.06%	788
Cluster 8				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
GTGAAT_TCATTC	12.0743	GTGAAT; TCATTC	2.93%	752
ATTAAC_CTTAAC	11.3839	CTTAAC	3.14%	780
TTACAC_GTGAAT	10.794	GTGAAT	19.27%	403
ACATTG_CTTAAC	10.5886	CTTAAC	3.78%	507
GTGAAT_CTTAAC	10.4807	GTGAAT; CTTAAC	6.08%	702
Cluster 9				
Module	SlnSEs	Significant Word Contained	Average Density	Average Distance
ACCTTC_GAGTAT	12.4362	GAGTAT	1.81%	682
CACCGA_GGAAGC	11.6845	GGAAGC	1.59%	806
CAACTC_CCTTTC	11.1518	CCTTTC	1.85%	828
CCTCAC_GAGTAT	10.6285	GAGTAT	39.21%	179
CATCTT_GAGTAT	10.4563	CATCTT; GAGTAT	2.84%	655
CTGACA_GAGTAT	10.1879	GAGTAT	1.53%	835
CATCTT_GGAAGC	9.97153	CATCTT; GGAAGC	2.60%	621
CTATGT_GGAAGC	9.95253	GGAAGC	0.70%	1758
CCTCAC_ACCTTC	9.93358	ACCTTC	2.24%	977
CGAATC_CCTTTC	9.91723	CCTTTC	2.19%	613

Table 8
AGRIS Look-up of Interesting Modules
The information from the AGRIS database 67 for the words in module's component is provided

Cluster 1					
–					
Cluster 2					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
TAACTC	7.69913	0.020566	MYB2 binding site motif	TAACT(G/C)GTT	29
Cluster 3					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
AGATAG	6.22494	0.162696	GATA promoter motif	[AT]GATA[GA]	30
Cluster 4					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
TCTAAC	7.60798	0.068925	MRE motif in CHS	TCTAACCTACCA	31
Cluster 5					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
GTATCT	7.26391	0.097377	EIL1 BS in ERF1; EIL2 BS in ERF1; EIL3 BS in ERF1	TTCAAGGGGGCATGTATCTTGAA	32
			EIN3 BS in ERF1	GGATTCAAGGGGGCATGTATCTTGAATCC	
Cluster 6					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
TCATAG	11.3198	–	LS5 promoter motif	ACGTCATAGA	34
Cluster 7					
–					
Cluster 8					
–					
Cluster 9					
Word	SlnSEs	P-Value	Name	Consensus Motif	Reference
CCTTTC	5.61743	0.026461	CArG2 motif in AP3	CTTACCTTTCATGGATTA	33

Table 9
TRANSFAC Analysis
For each significant word, this table shows transcription factors known to be bind to the word, as published in TRANSFAC8 and JASPAR9 database. Any known binding sites that hit significant words more than 3/4 of their total occurrences are shown

Cluster 1							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
TGATAC	7	7	MA0035	Gata1	90.93 - 90.93	NGATNN	14\14
TGATAC	7	7	MA0037	Gata3	86.31 - 86.31	HGATWR	14\14
Cluster 2							
–							
Cluster 3							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
AGATAG	10	10	MA0035	Gata1	94.35 - 94.35	NGATNN	18\18
AGATAG	10	10	MA0037	Gata3	100.00 - 100.00	HGATWR	18\18
AGATAG	10	10	V$GATA3_01	Gata3	85.26 - 95.56	NNGATWDNN	16\18
AGATAG	10	10	V$GATA6_01	Gata6	85.51 - 92.33	NNHGATWNNN	18\18
AGATAG	10	10	V$GATA_Q6	Gata	91.43 - 98.09	WGATARN	18\18
Cluster 4							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
AGATCA	21	25	V$HNF4_Q6_02	HNF4	88.87 - 88.87	AGKYCA	48\48
AGATCA	21	25	V$HNF4_Q6_03	HNF4	90.67 - 90.67	NGDBCA	48\48
GTATCC	12	13	MA0035	Gata1	93.07 - 93.07	NGATNN	19\19
Cluster 5							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
CTCATG	14	20	MA0089	TCF11-MafG	87.05 - 87.05	NATGAC	29\29
GTATCT	14	16	MA0035	Gata1	92.07 - 92.07	NGATNN	26\26
GTATCT	14	16	MA0037	Gata3	87.55 - 87.55	HGATWR	26\26
AGAATC	17	26	V$STAT5A_04	STAT5A	85.03 - 88.96	NNNTTCYN	32\38
GGATAC	8	9	MA0035	Gata1	93.07 - 93.07	NGATNN	11\11
Cluster 6							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
GCATCG	6	6	MA0035	Gata1	97.24 - 97.24	NGATNN	6\6
Cluster 7							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
CTTTCG	9	10	MA0080	SPI1	85.47 - 85.47	VGGAAS	21\21
ATCTGA	11	15	V$CAP_01	CAP	85.57 - 94.08	NCABHNNN	25\25

(*Continued*)

Table 9
(*Continued*)

Cluster 8							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
TCATTC	13	17	V$CAP_01	CAP	89.34 - 97.86	NCABHNNN	31\31
TCATTC	13	17	V$GEN_INI2_B	GEN INI2	85.83 - 100.00	BBNCANTB	24\31
TCATTC	13	17	V$GEN_INI_B	GEN INI	86.69 - 100.00	NBNCANTB	24\31
Cluster 9							
Word	S	O	TF_ID	TF_Name	Score_Range	TFBS	Matches
CATCTT	5	6	V$CAP_01	CAP	86.97 - 93.14	NCABHNNN	10\12
GGAAGC	4	4	V$CETS168_Q6	CETS168	85.92 - 100.00	CMGGAAGY	6\7
GGAAGC	4	4	V$PEA3_Q6	PEA3	89.16 - 91.24	ACWTCCK	6\7
GGAAGC	4	4	V$STAT3_02	STAT3	91.22 - 98.25	NNNTTCCN	7\7

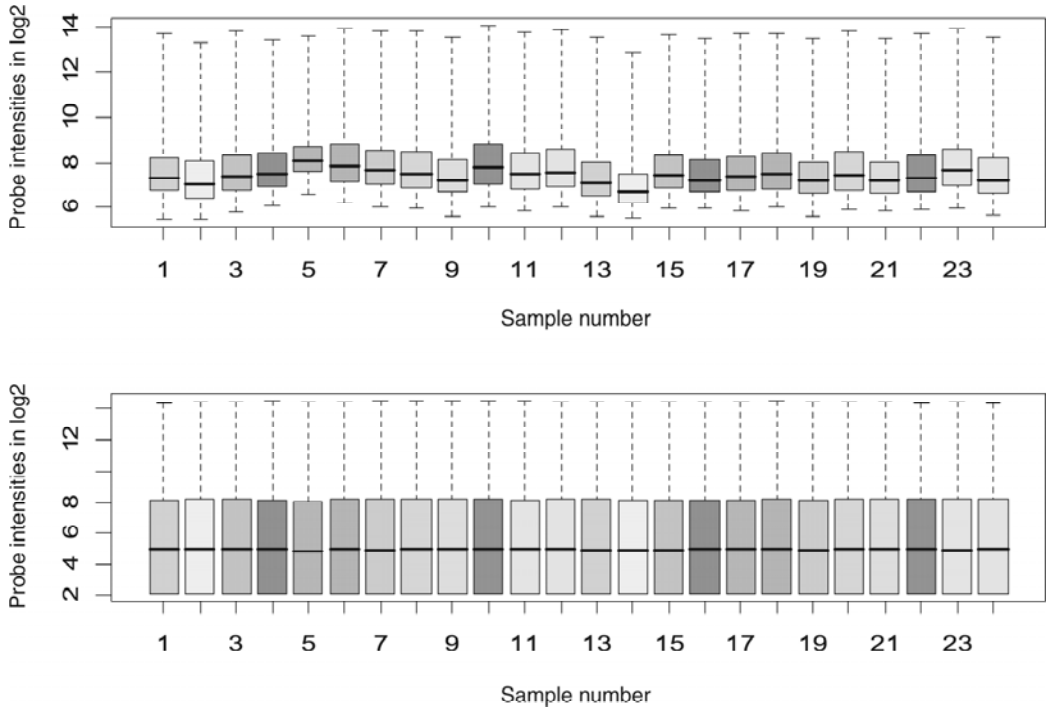


Figure 1. The normalization of samples. The raw data (top) before normalization and the data after GCRMA normalization (bottom) are compared.

References

[1] F.D. Sack, Plant gravity sensing *International Review of Cytology-a Survey of Cell Biology*, *127*, 1991, 193–252.

[2] J.Z. Kiss, Mechanisms of the early phases of plant gravitropism, *CRC Critical Review Plant Science*, *19*(6), 2000, 551–573.

[3] T. Caspar & B.G. Pickard, Gravitropism in a starchless mutant of Arabidopsis – Implications for the Starch-Statolith Theory of Gravity Sensing, *Planta*, *177*(2), 1989, 185–197.

[4] F.D. Sack & J.Z. Kiss, Rootcap structure in wild-type and in a starchless Mutant of Arabidopsis, *American Journal of Botany*, *76*(2), 1989, 454–464.

[5] B.R. Harrison & P.H. Masson, ARL2, ARG1 and PIN3 define a gravity signal transduction pathway in root statocytes, *Plant Journal*, *53*(2), 2008, 380–392.

[6] A.C. Scott & N.S. Allen, Changes in cytosolic pH within Arabidopsis root columella cells play a key role in the early signaling pathway for root gravitropism, *Plant Physiol*, *121*(4), 1999, 1291–1298.

[7] E.B. Blancaflor & P.H. Masson, Plant gravitropism: Unraveling the ups and downs of a complex process, *Plant Physiology*, *133*(4), 2003, 1677–1690.

[8] I.Y. Perera, I. Heilmann, & W.F. Boss, Transient and sustained increases in inositol 1,4,5-trisphosphate precede the differential growth response in gravistimulated maize pulvini. *Proceedings National Academy Sciences United States of America*, *96*(10), 1999, 5838–5843.

[9] W.F. Boss, I.Y. Perera, J. Love, & I. Heilmann, Altering phosphoinositide metabolism by expressing human type I inos-

itol polyphosphate 5' phosphatise in tobacco cells. *Molecular Biology of the Cell*, 12: 2001, 820.

[10] J.H. Joo, Y.S. Bae, & J.S. Lee, Role of auxin-induced reactive oxygen species in root gravitropism, *Plant Physiology*, *126*(3), 2001, 1055–1060.

[11] A.M. Clore, S.M. Doore, & S.M.N. Tinnirello, Increased levels of reactive oxygen species and expression of a cytoplasmic aconitase/iron regulatory protein 1 homolog during the early response of maize pulvini to gravistimulation, *Plant Cell and Environment*, *31*(1), 2008, 144–158.

[12] P.B. Kaufman, L.-L. Wu, T.G. Brock, & D. Kim, Hormones and the orientation of growth, in P.J. Davies (Ed.), *Plant Hormones*, Second Edition, Dordrecht, The Netherlands: Kluwer Academic Publishers, 547–571.

[13] J.M. Kimbrough, R. Salinas-Mondragon, W.F. Boss, C.S. Brown, & H.W. Sederoff, The fast and transient transcriptional network of gravity and mechanical stimulation in the *Arabidopsis* root apex, *Plant Physiology*, *136*(1), 2004, 2790–2805.

[14] J. Lichtenberg, E. Jacox, J.D. Welch, K. Kurz, X. Liang, M.Q. Yang, F. Drews, K. Ecker, S.S. Lee, L. Elnitski, & L.R. Welch, Word-based characterization of promoters involved in human DNA repair pathways, *BMC Genomics*, *10*(Suppl 1), 2009, S18.

[15] J. Lichtenberg, A. Yilmaz, J.D. Welch, K. Kurz, X. Liang, F. Drews, K. Ecker, S.S. Lee, M. Geisler, E. Grotewold, & R.W. Lonnie, The word landscape of the non-coding segments of the *Arabidopsis thaliana* genome, *BMC Genomics*, *10*, 2009, 463.

[16] S.K. Palaniswamy, S. James, H. Sun, R.S. Lamb, R.V. Davuluri, & E. Grotewold, AGRIS and AtRegNet: A platform to link *cis*-regulatory elements and transcription factors into regulatory networks, *Plant Physiology*, *140*(3), 2006, 818–829.

[17] R.V. Davuluri, H. Sun, S.K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, & E. Grotewold, AGRIS: *Arabidopsis* Gene Regulatory Information Server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors, *BMC Bioinformatics*, *4*(1), 2003, 25.

[18] E. Wingender, X. Chen, R. Hehl, H. Karas, & I. Liebich, TRANSFAC: An integrated system for gene expression regulation, *Nucleic Acids Research*, *28*, 2000, 316–319.

[19] J.C. Bryne, E. Valen, M.H. Tang, T. Marstrand, & O. Winther, JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update, *Nucleic Acids Research*, *36*, 2008, D102–106.

[20] H.M. Hsu, S.J. Ong, M.C. Lee, & J.H. Tai, Transcriptional regulation of an iron-inducible gene by differential and alternate promoter entries of multiple myb proteins in the protozoan parasite *Trichomonas vaginalis*, *Eukaryotic Cell*, *8*(3), 2009, 362–372.

[21] J.H. Yoo, C.Y. Park, J.C. Kim, W.D. Heo, M.S. Cheong, H.C. Park, M.C. Kim, B.C. Moon, M.S. Choi, Y.H. Kang, J.H. Lee, H.S. Kim, S.M. Lee, H.W. Yoon, C.O. Lim, D.J. Yun, S.Y. Lee, W.S. Chung, & M.J. Cho, Direct interaction of a divergent CaM isoform and the transcription factor, MYB2, enhances salt tolerance in arabidopsis. *The Journal of Biological Chemistry*, *280*(5), 2005, 3697–3707.

[22] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y.C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Lacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, & J.H. Zhang, Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, *5*(10), 2004, R80.

[23] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, & F. Spencer, A model-based background adjustment for oligonucleotide expression arrays, *Journal of the American Statistical Association*, *99*(468), 2004, 909–917.

[24] R. Breitling, P. Armengaud, A. Amtmann, & P. Herzyk, Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS Letter*, *573*(1–3), 2004, 83–92.

[25] L. Kaufman & P.J. Rousseeuw, Finding groups in data (John Wiley & Sons, Inc. 1990), 190, 191, 212, 217.

[26] T. Beissbarth & T.P. Speed, GOstat: Find statistically over-represented gene ontologies within a group of genes, *Bioinformatics*, *20*(9), 2004, 1464–1465.

[27] B. Lenhard & W.W. Wasserman, TFBS: Computational framework for transcription factor binding site analysis, *Bioinformatics*, *18*, 2002, 1135–1136.

[28] E. Jacox & L. Elnitski, Finding occurrences of relevant functional elements in genomic signatures, *International Journal of Computational Science*, *2*(5), 2008, 599–606.

[29] C. Martin & J. Paz-Ares, MYB transcription factors in plants, *Trends in Genet*, *13*, 1997, 67–73.

[30] G.R. Teakle, I.W. Manfield, J.F. Graham, & P.M. Gilmartin, *Arabidopsis Thaliana* GATA factors: Organisation, expression and DNA-binding characteristics, *Plant Molecular Biology*, *50*(1), 2002, 43–57.

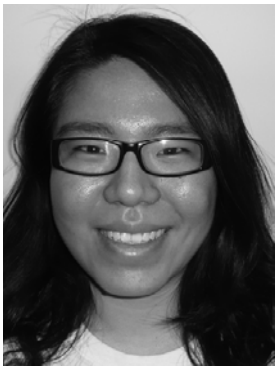
[31] U. Hartmann, W.J. Valentine, J.M. Christie, J. Hays, G.I. Jenkins, & B. Weisshaar, Identification of UV/blue light-response elements in the *Arabidopsis thaliana* chalcone synthase promoter using a homologous protoplast transient expression system, *Plant Molecular Biology*, *36*, 1998 741–754.

[32] R. Solano, A. Stepanova, Q. Chao, & J.R. Ecker, Nuclear events in Ethylene signaling: A transcriptional cascade mediated by ethylene-insensitive3 and ethylene-response-factor1, *Genes of Development*, *12*, 1998, 3703–3714.

[33] J.J. Tilly, D.W. Allen, & T. Jack, The CArG boxes in the promoter of the *Arabidopsis* floral organ identity gene APETALA3 mediate diverse regulatory effects, *Development*, *125*(9), 1998, 1647–1657.

[34] C. Despres, C. Delong, S. Glaze, E. Liu, & P.R. Fovert, The *Arabidopsis* NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors, *Plant Cell*, *12*, 2000, 279–290.

Biographies



Xiaoyu Liang received her B.S. degree in management information system from University of Shanghai for Science and Technology, Shanghai, China in 2006. She is currently a master student in Bioinformatics at Ohio University, USA. Her research interest is *cis*-regulatory module discovery.

Kaiyu Shen received his B.S. degree in molecular and cellular biology from the University of Science and Technology of China, Hefei, China in 2006. He is currently a Ph.D. candidate in Molecular and Cellular Biology at Ohio University. His research interests are gravitropism and microarray data analysis.

Jens Lichtenberg received his B.Sc. and M.Sc. degrees in Business Informatics from the Clausthal University of Technology, Germany in 2002 and 2004, respectively. He is currently a Ph.D. candidate in Bioinformatics at Ohio University, USA, and the President of the Regional Student Group Ohio for the Student Council of the International Society of Computational Biology. His research interests include regulatory genomics and proteomics.

Sarah Wyatt is an associate professor of Plant Developmental Biology at Ohio University. She received her Ph.D. degree in Plant Physiology and Molecular Biology at Purdue University. Her main research interests are plant gravitropism, signal transduction, and bioinformatic analysis of genomic data. She is also faculty advisor for the DNA Analysis/ genomics facility at Ohio University.

Lonnie Welch received his Ph.D. degree in Computer and Information Science from the Ohio State University. He is the Stuckey Professor of Electrical Engineering and Computer Science at Ohio University, and is a member of the Graduate Faculties of the Biomedical Engineering Program and of the Molecular and Cellular Biology Program. He directs the Bioinformatics Laboratory, where he performs research in the area of computational genomics. His research has been sponsored by the Defense Advanced Research Projects Agency, the Navy, NASA, the National Science Foundation, the Army, and the Ohio Board of Regents.