The Regulatory Code for Transcriptional Response Diversity and Its Relation to Genome Structural Properties in *A. thaliana*

Dirk Walther[1*], Roman Brunnemann[1,2], Joachim Selbig[1,2]

1 Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany, **2** Institute for Biochemistry and Biology, Potsdam University, Potsdam, Germany

Regulation of gene expression via specific *cis*-regulatory promoter elements has evolved in cellular organisms as a major adaptive mechanism to respond to environmental change. Assuming a simple model of transcriptional regulation, genes that are differentially expressed in response to a large number of different external stimuli should harbor more distinct regulatory elements in their upstream regions than do genes that only respond to few environmental challenges. We tested this hypothesis in *Arabidopsis thaliana* using the compendium of gene expression profiling data available in AtGenExpress and known *cis*-element motifs mapped to upstream gene promoter regions and studied the relation of the observed breadth of differential gene expression response with several fundamental genome architectural properties. We observed highly significant positive correlations between the density of *cis*-elements in upstream regions and the number of conditions in which a gene was differentially regulated. The correlation was most pronounced in regions immediately upstream of the transcription start sites. Multistimuli response genes were observed to be associated with significantly longer upstream intergenic regions, retain more paralogs in the *Arabidopsis* genome, are shorter, have fewer introns, and are more likely to contain TATA-box motifs in their promoters. In abiotic stress time series data, multistimuli response genes were found to be overrepresented among early-responding genes. Genes involved in the regulation of transcription, stress response, and signaling processes were observed to possess the greatest regulatory capacity. Our results suggest that greater gene expression regulatory complexity appears to be encoded by an increased density of *cis*-regulatory elements and provide further evidence for an evolutionary adaptation of the regulatory code at the genomic layout level. Larger intergenic spaces preceding multistimuli response genes may have evolved to allow greater regulatory gene expression potential.

Citation: Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. PLoS Genet 3(2): e11. doi:10.1371/journal.pgen.0030011

Introduction

The regulation of gene expression has evolved in cellular organisms as a major adaptive mechanism to respond to environmental changes [1–5]. How the apparent diversity of responses is encoded in an organism's genome is a central question in understanding transcriptional regulation induced by different environmental and extracellular conditions [6–10]. The induction or repression of particular genes in response to specific environmental challenges is primarily controlled by the recognition and binding of transcriptional regulator proteins (transcription factors) to *cis*-regulatory elements constituted by short DNA sequence motif sites located in the upstream regions of genes [11–13]. Under the simplest scenario of transcriptional regulation, distinct external challenges are matched by specific cognate regulatory sites in upstream regulatory regions of genes that have evolved to respond to the particular perturbation. Genes that are differentially expressed in response to a large number of different external stimuli (multistimuli response genes) are therefore expected to contain more distinct *cis*-regulatory elements in their upstream regions than are genes that respond to only few environmental cues. There are two plausible strategies of how evolution may have shaped the noncoding, regulatory segments of genomes to encode a greater capacity of downstream genes to respond to a wider range of different stimuli by differential gene expression. A broader response spectrum may have evolved via an increased density of regulatory motifs or via an enlarged size of regulatory intergenic regions to accommodate more elements (or both). In analyzing expression patterns of *Caenorhabibditis elegans* and *Drosophila melanogaster* genes in different developmental and tissue differentiation stages, Nelson and coworkers [10] observed that indeed there exists a significant positive correlation between the complexity of a gene's expression, that is, to be expressed in a number of different tissues and developmental stages, and the size of its flanking noncoding, intergenic sequence, suggesting that regulatory requirements may have played a significant role in shaping the architecture of genomes. The association of

Editor: Yoshihide Hayashizaki, RIKEN Genomic Sciences Center, Japan

Received August 21, 2006; **Accepted** December 6, 2006; **Published** February 9, 2007

A previous version of this article appeared as an Early Online Release on December 7, 2006 (doi:10.1371/journal.pgen.0030011.eor).

Abbreviations: FDR, false discovery rate; GO, Gene Ontology; nr, nonredundant; RMA, Robust Microchip Analysis; SEM, standard error of the mean

* To whom correspondence should be addressed. E-mail: walther@mpimp-golm.mpg.de

Author Summary

The induction or repression of specific genes has evolved in living organisms as a mechanism to respond to environmental changes. At the molecular level, this process is mediated via molecular switches, so-called regulatory elements, generally located in the genomic region adjacent to the gene they control, the gene promoter. Upon environmental change, specific proteins bind to such regulatory elements, thereby turning on or off the associated genes. As this molecular response is often specific to the external signal, genes that respond to a large number of different external stimuli should harbor more distinct regulatory elements in their promoter regions than should genes responding only to few environmental challenges. In analyzing data for the plant *Arabidopsis thaliana,* we observed that indeed an increased number of regulatory elements is associated with a broader range of responses. Several other genome structural properties, such as gene size, the occurrence of similar genes in the *Arabidopsis* genome, and the distance between genes, were also observed to be correlated with a broader breadth of response. The results suggest that greater regulatory complexity appears encoded by an increased density of regulatory elements and provide further evidence for an evolutionary adaptation of the regulatory code at the genomic architectural level.

promoters harboring multiple different regulatory sites with differential responses of their downstream genes to varied growth conditions has also been conceptualized by Harbison and coworkers using ChIP-chip yeast data [8]. They distinguished four types of motif arrangements in promoters: single regulators—associated with genes of common functions, repetitive motifs—allowing graded transcriptional responses, multiple regulators—allowing responses in diverse conditions, and co-occurring regulators—for physically interacting regulators.

In this study, we test and quantify the strength of the association of the presence of multiple different regulatory motifs in promoters with the breadth of differential gene expression response to external stimuli in *Arabidopsis thaliana*. For this purpose we use the available compendium of gene expression profiling data in the public AtGenExpress *Arabidopsis thaliana gene* expression repository (http://www.arabidopsis.org/info/expression/ATGenExpress.jsp) alongside previously characterized *cis*-element motifs mapped to upstream gene regions. *Arabidopsis* is an ideal model system for the investigation of the regulatory code in higher eukaryotic organisms due to its complete genome sequence [14], the availability of public domain resources of known *cis*-elements in upstream gene regions such as Athena (http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl) [15], and genome-wide expression profiling data for a large and diverse collection of treatment-control experiments. These experiments encompass a wide range of abiotic and biotic treatments, the application of plant hormones and other chemical treatments performed, and were designed to enable comparative studies by using the same technology platform and reference conditions. We define the property of genes to be differentially regulated in response to many or few conditions as their "breadth of response." It is this quantity and its relation to gene regulatory motifs and other genome structural properties that form the main focus of this study. Rather than internal gene expression regulation during tissue differentiation and organism development for sets of genes with available profiling data in developmental expression series and literature information as used by [10], the available *Arabidopsis* data allow us to assess systematically the differential gene expression response breadth and for nearly all genes more directly by measuring gene expression in response to external stimuli. Furthermore, with the mapping information of previously characterized regulatory motifs, albeit the set may likely not be considered complete, to the *Arabidopsis* genome at hand, we are able to associate gene expression complexity directly to *cis*-elements, their identity, frequency, and spatial distribution and in conjunction with genomic layout properties, such as distances between neighboring genes and gene size. Transcriptional response programs to external stress have been studied in microorganisms, yeast in particular [3]. Studying the association of regulatory capacity and the breadth of transcriptional response in a higher, multicellular and multitissue organism such as *Arabidopsis* will allow comparison and an assessment of the generality of the observed mechanisms.

Our results obtained in *Arabidopsis* lend further support to the notion that larger intergenic regions may have evolved to allow broader differential gene expression capacity [10]. *Arabidopsis* genes showing differential gene expression in response to a greater range of external stimuli are flanked by larger intergenic regions. In addition, increased breadth of gene expression response was observed to correlate with an increased density of motifs in upstream promoter regions, most pronounced in segments immediately upstream of the transcription start site. Among the various *cis*-elements analyzed, the TATA-box motif appears to play a unique role. We observed TATA-box–containing genes to possess a significantly increased breadth of response and to be associated with significantly longer upstream intergenic regions compared to TATA-less genes, thereby possibly allowing for greater regulatory capacity. Identified correlations of several fundamental genome architectural properties with the observed breadth of differential gene expression response are discussed in the context of evolutionary forces shaping the structure of eukaryotic genomes.

Results

Applying a noise-filtered threshold of 2-fold up-regulation or down-regulation in 43 treatment-control experiments, we observed a nearly exponential decrease in the number of genes with increasing cumulative numbers of experiments with differential expression (Figure 1). Most genes were found to be differentially expressed in only few experiments, whereas only a small number of genes were observed to respond to many different external stimuli.

Genes involved in stress response, cell growth, and lipid transport are particularly overrepresented in the set of multistimuli-sensitive genes (Table 1), whereas the housekeeping functions, protein catabolism and synthesis, RNA processing, and DNA repair, as well as uncharacterized genes with as-of-yet-unknown function, are more associated with the group of genes with narrow breadth of differential gene expression response, indicating that they are constitutively expressed. Genes involved in embryonic development were also grouped with narrow response genes. However, this may primarily be explained by the absence of embryonic development samples in the analysis. The gene *AtEXPA8* (At2g40610),

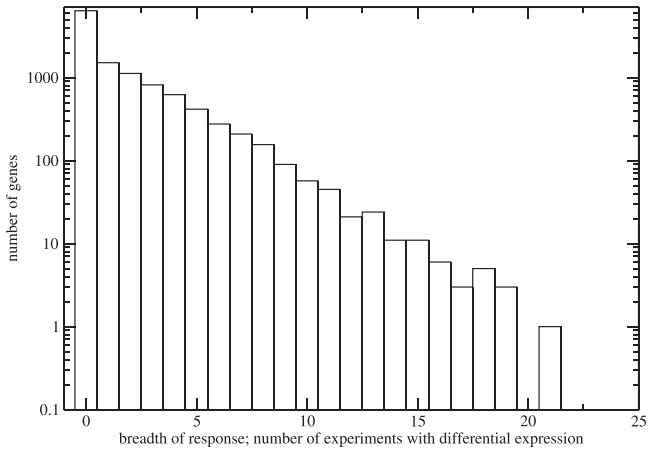

Figure 1. Semilogarithmic Frequency Distribution of the Number of Different Experiments in which Genes Were Found to Be Differentially Expressed Defined Here as the Breadth of Differential Gene Expression Response

doi:10.1371/journal.pgen.0030011.g001

a member of the α-expansin gene family and involved in cell wall modification and loosening, was observed to show the greatest response breadth and was differentially expressed in 22 different experiments.

With the mapping information for 93 previously characterized *cis*-element motifs to all *Arabidopsis* upstream gene regions up to a length of 3,000 nucleotides available from the Athena database [15] (see Methods), it is now possible to correlate the breadth of differential gene expression response to the number of *cis*-regulatory elements for all genes. If a broader differential gene expression response is reflected and even encoded by an increased number of *cis*-elements, we expect a positive correlation. Indeed, very significant positive correlations can be observed for both the number of total and unique elements in promoters of defined lengths (Figure 2), i.e., multistimuli response genes harbor more *cis*-elements in their upstream regions (higher density of elements) than do genes with a narrower scope of responses. This positive correlation was particularly significant and pronounced for an assumed promoter length of 500 nucleotides immediately upstream of the presumed site of transcription initiation. While the significance of the observed correlations was high, absolute motif counts increased only moderately corresponding to an increase of approximately 13% in the 500 nucleotides immediately upstream over the observed range of response breadth. For promoter segments farther upstream, the significance of the observed correlations was weaker, in part explained by fewer observations, albeit detectable, and the relative increase in motif counts smaller with increasing breadth of response.

The number of experiments with differential expression, the breadth of response, includes counts for both up-regulation and down-regulation in treatment-control samples. When only up-regulation or only down-regulation responses were considered differential gene expression events, we observed that the obtained positive correlations of motif counts with breadth of response was primarily caused by up-regulation rather than by down-regulation events for which no significant correlation were obtained to the number of upstream motifs in promoter segments of defined length (Figure S2).

Determining the correct length of gene promoters, i.e., to assess whether a putative *cis*-acting regulatory site will exert an influence on its downstream gene depending upon its distance from the site of transcription initiation, is difficult experimentally, even more so if only *in silico* mapping information is available. Allowing variable promoter segment lengths of up to 3,000 nucleotides, but excluding *cis*-elements that overlap with neighboring upstream genes and only considering *cis*-elements that map to intergenic upstream regions, the magnitude of the correlation of motif counts to breadth of response increased significantly ($r \approx 0.2$ obtained for motif counts in variable-length promoter segments [Figure 3] compared to $r \approx 0.1$ for fixed promoter lengths [Figure 2]). Genes that are differentially regulated in nine different experiments (the greatest observed breadth of response value in the set of experiments analyzed here with more than associated 100 genes) have 50% more *cis*-elements for both overall *cis*-element counts and unique counts versus genes with no detectable differential response in the experiments included in this study.

This result suggests that the length of the upstream intergenic region, too, is positively correlated with the breadth of differential gene expression response. Such a positive correlation has already been reported in analyses of gene expression profiles in different developmental and cell differentiation stages in *C. elegans* and *D. melanogaster* [10]. Indeed, we found a very significant positive correlation of intergenic upstream sequence length and the breadth of response ($r = 0.19$, $p = 1.2E{-}89$; Figure 4A) also in response to external stimuli. On average, highly multistimuli response genes have approximately double the intergenic upstream space compared to genes with no differential response. This trend was observed equally strong for both differential up-regulation or down-regulation events (Figure S3). This finding prompted us to relate other gene properties to the breadth of response. The length of downstream intergenic segments was also found positively correlated with breadth of response, albeit at a less significant level and smaller magnitude ($r = 0.08$, $p = 4.6E{-}18$; Figure 4A) than for the corresponding upstream segment lengths. Downstream intergenic segments are also significantly shorter than upstream segments. We observed that multistimuli response genes are, on average, significantly shorter, with shorter gene length ($r = -0.19$, $p = 1.2E{-}95$ [Figure 4B]) as well as shorter cDNA length ($r = -0.12$, $p = 2.2E{-}42$ [Figure 4E]) and have shorter 5′ ($r = -0.08$, $p = 5.8E{-}16$) and, to a lesser degree, 3′ untranslated regions ($r = -0.05$, $p = 5.1E{-}9$ [Figure 4C]). Commensurate with shorter gene length, the number of introns was also observed to be negatively correlated with breadth of differential expression ($r = -0.2$, $p = 2.3E{-}110$ [Figure 4D]), as was the mean length of intronic segments ($r = -0.2$, $p = 6E{-}114$). In addition to the analyzed gene size–related parameters, multistimuli response genes are also more likely to contain additional paralogs in the *Arabidopsis* genome than are narrow-response genes ($r = 0.17$, $p = 9.4E{-}77$ [Figure 4F] [30% amino acid sequence identity], and $r = 0.09$, $p = 3.1E{-}24$ for the more stringent paralog settings of requiring 70% amino acid sequence identity [see Methods]). Among the various properties analyzed, the length of the upstream region and gene length and the associated number of introns

Table 1. Classification of Genes with High or Low Breadth of Differential Gene Expression Response

Annotation Category	2,000 Genes with Highest Breadth of Response		5,000 Genes with Lowest Breadth of Response	
	FDR *p*-Value	Category	FDR *p*-Value	Category
GO Process	4.47E−08	Response to wounding	9.92E-26	Biological process unknown
	6.98E−08	Response to abscisic acid stimulus	7.14E-08	Ubiquitin-dependent protein catabolism
	2.69E−05	Response to cold	1.64E-06	Embryonic development (sensu Magnoliophyta)
	4.40E−05	Carbohydrate metabolism	1.30E-03	Intracellular protein transport
	5.92E−05	Response to jasmonic acid stimulus	4.13E-03	MRNA processing
	1.43E−04	Cell wall loosening (sensu Magnoliophyta)	1.77E-02	Protein modification
	1.89E−04	Lipid transport	3.07E-02	Protein amino acid glycosylation
	3.55E−04	Toxin catabolism	4.24E-02	ER to Golgi vesicle-mediated transport
	3.56E−04	Response to heat	4.33E-02	Protein transport
	3.74E−04	Response to water deprivation	5.94E-02	DNA repair
	7.98E−04	Response to salt stress	6.05E-02	RNA processing
	8.40E−04	Response to oxidative stress	8.37E-02	Ubiquitin cycle
	1.55E−03	Response to auxin stimulus	—	—
	3.17E−03	Response to gibberellic acid stimulus	—	—
	5.97E−03	Response to salicylic acid stimulus	—	—
	7.34E−03	Cell wall modification	—	—
	7.58E−03	Transmembrane receptor protein tyrosine kinase signaling pathway	—	—
	1.17E−02	Cell wall modification during multidimensional cell growth (sensu Magnoliophyta)	—	—
	1.81E−02	Unidimensional cell growth	—	—
	2.91E−02	Trehalose biosynthesis	—	—
	3.90E−02	Hyperosmotic salinity response	—	—
	4.02E−02	Response to pest, pathogen, or parasite	—	—
	4.09E−02	Response to nematode	—	—
	4.11E−02	Electron transport	—	—
	5.71E−02	Defense response	—	—
	5.80E−02	Response to desiccation	—	—
	5.96E−02	Sucrose catabolism, using β-fructofuranosidase	—	—
	6.14E−02	Photosynthesis, light harvesting	—	—
	6.56E−02	Response to ethylene stimulus	—	—
	6.62E−02	Chromosome organization and biogenesis (sensu Eukaryota)	—	—
	6.74E−02	Multidrug transport	—	—
	7.03E−02	Oligopeptide transport	—	—
	7.08E−02	Abscisic acid mediated signaling	—	—
	7.38E−02	Response to mechanical stimulus	—	—
	7.54E−02	Response to chitin	—	—
	7.71E−02	Glucosinolate biosynthesis	—	—
Gene Family	8.73E−04	Glycoside hydrolase gene families	4.63E−04	Cytoplasmic ribosomal protein gene family
	1.10E−03	Class III peroxidase	1.22E−03	Protein synthesis factors
	1.94E−03	Cytochrome P450	4.44E−02	Primary pumps (ATPases) gene families
	5.64E−03	Expansins	5.92E−02	ABC superfamily
	1.18E−02	GST superfamily	6.17E−02	Core cell cycle genes

Fisher exact test statistics of overrepresented gene categories in the set of the top 2,000 genes sorted by descending breadth of differential gene expression response versus bottom 5,000 genes (left) and vice versa (right) for GO process assignments as well as gene family annotations. Only categories with FDR *p*-value < 0.1 are listed.
doi:10.1371/journal.pgen.0030011.t001

were found most strongly correlated with breadth of differential gene expression response. The sum of the intergenic upstream and downstream distances measuring the size of a gene's intergenic flanking region was not observed to be negatively correlated with the size of the flanked gene ($r = 0.014$, $p = 0.14$); i.e., genes are not distributed evenly such that longer intergenic regions follow from shorter genes. Applying a multiple linear regression approach, we analyzed what level of correlation with breadth of response can be achieved by combining the various and largely independent properties depicted in Figure 4. Performing a stepwise-forward multiple linear regression and using nonunique motif counts in the 500-nucleotide upstream regions (motif density), the three most significant regressors were (1) number of introns (partial correlation coefficient, $\beta = -0.23$), (2) distance to next upstream gene ($\beta = 0.13$), and (3) motif density ($\beta = 0.09$),

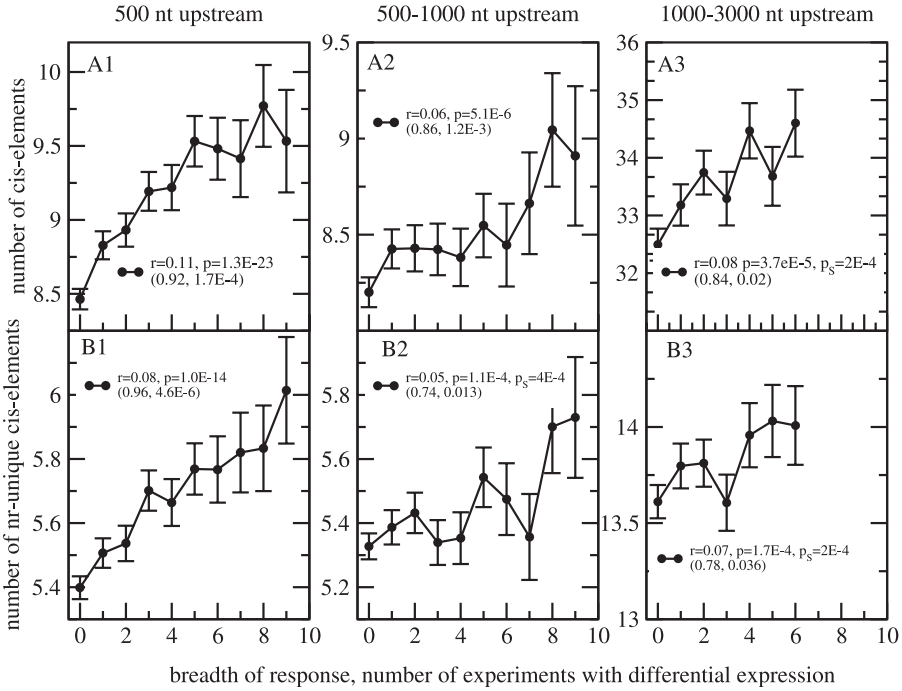

Figure 2. Correlation of the Number of *cis*-Regulatory Elements in Gene Upstream Promoter Regions of Different Length and Position Relative to the Transcription Start Site with the Breadth of Differential Gene Expression Response

The upper row (A1–A3) shows data for all *cis*-elements including counts for multiple occurrences of *cis*-elements. In the lower row (B1–B3), only unique *cis*-element counts (see Methods) have been used. Genes were only included if their associated intergenic upstream region was large enough to fully contain the considered upstream interval, i.e., no overlap of preceding genes and the considered promoter segment was allowed. Shown are the mean values and the SEM. Correlation coefficients, *r*, and associated *p*-values are given for all raw data pairs (individual genes and associated motif counts) and, in parentheses, for the mean values as they are plotted in the graph. If the shuffled *p*-value, p_S, was not zero, it is reported as well (see Methods). Mean *cis*-element counts were only plotted if at least 100 genes were detected at the particular breadth of response. Note: With increasing distance of the considered promoter region from the transcription start site, fewer genes were considered as intergenic upstream regions were frequently not large enough (respective gene counts were A1, B1 = 8,597; A2, B2 = 6,537; A3, B3 = 2,737).

doi:10.1371/journal.pgen.0030011.g002

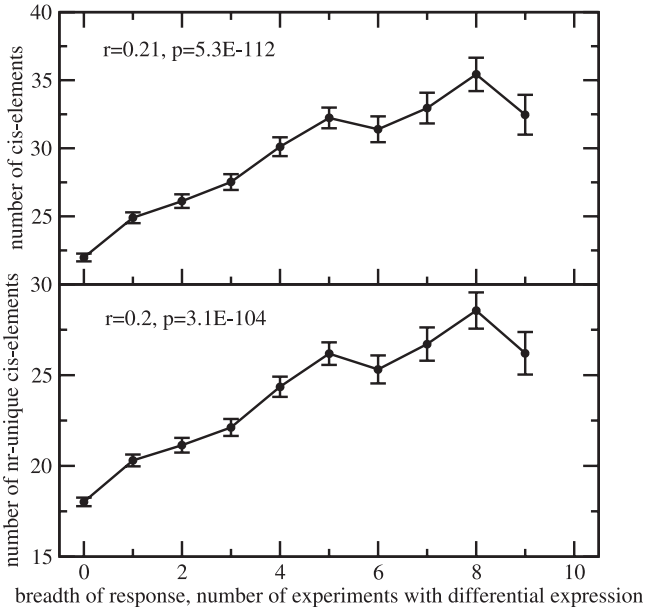

Figure 3. Correlation of the Number of Regulatory *cis*-Elements in Gene Upstream Promoter Regions Up to 3,000 Nucleotides in Length or Truncated at the Next Upstream Gene Boundaries (Variable Promoter Length) with the Breadth of Differential Gene Expression Response

Plotted are the mean values and associated standard errors of the mean. Pearson linear correlation coefficients, *r*, and associated *p*-values are given for all raw data pairs (for all genes considered in the analysis).

doi:10.1371/journal.pgen.0030011.g003

yielding a combined level of correlation of $r = 0.28$, $r^2 = 0.08$ ($p \ll 0.01$, $n = 6{,}976$ genes with complete information and upstream intergenic region longer than 500 nucleotides). Adding more properties (downstream region length, UTR lengths, etc.) did not result in significantly increased correlation levels.

We investigated what gene families and functions are associated with long upstream intergenic segments. By sorting all *Arabidopsis* genes according to their upstream intergenic length and comparing the Gene Ontology (GO) annotations of a subset of genes with longest upstream segments to a set of genes with shortest upstream segment lengths using Fisher exact statistical tests, we found that transcription factors, transcriptional regulatory functions, and genes involved in signaling processes are particularly overrepresented among genes with long intergenic upstream segments, while ribosomal genes involved in protein biosynthesis and genes involved in other housekeeping functions such as glycolysis are relatively overrepresented among genes with short upstream segment lengths, as are genes with currently unidentified function (Table 2). The found association of functional categories with intergenic upstream distances in *A. thaliana* agrees well with very similar observations reported for *C. elegans* and *D. melanogaster* [10].

We analyzed which gene families and biological processes are associated with genes with the greater number of *cis*-elements in their upstream promoter region. When counting regulatory elements in upstream segments of up to 3,000

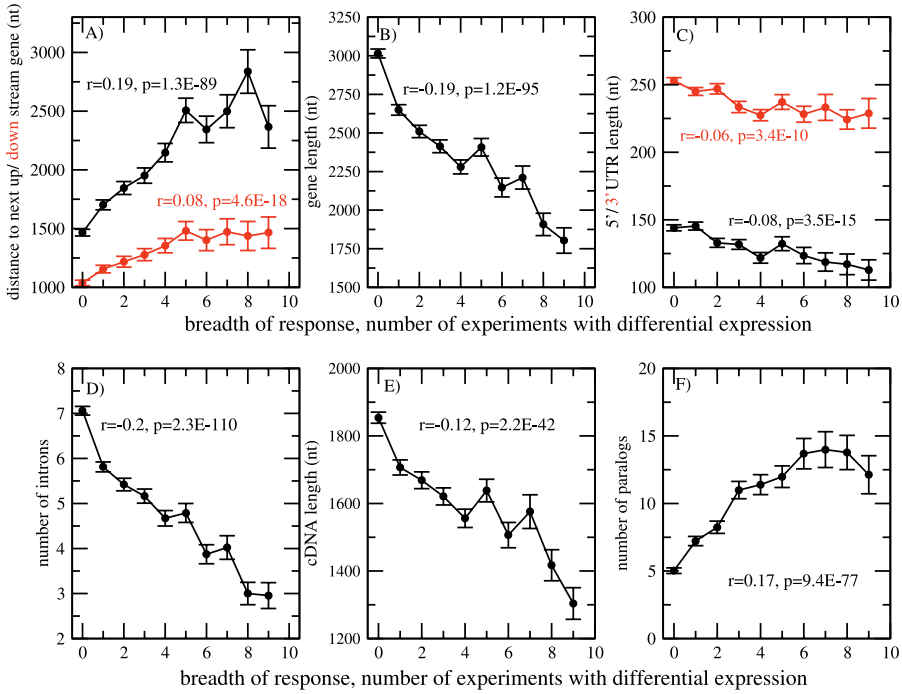

Figure 4. Correlation of the Length of Intergenic Upstream and Downstream Sequence (Distance to the Next 5′-Upstream Gene), Gene Length, 5′/3′ UTR Length, Number of Introns, cDNA Length, and Number of Paralogs with the Breadth of Differential Gene Expression Response

(A) Intergenic upstream and downstream sequence (distance to the next 5′-upstream gene).
(B) Gene length.
(C) 5′/3′ UTR length.
(D) Number of introns.
(E) cDNA length.
(F) Number of paralogs with the breadth of differential gene expression response.
Plotted data points correspond to mean values and associated standard errors of the mean. Pearson linear correlation coefficients, *r*, and associated *p*-values are given for all raw data pairs.
doi:10.1371/journal.pgen.0030011.g004

nucleotides and truncating them at neighboring upstream genes, i.e., variable promoter length, the GO processes and gene families characteristic for genes with many motifs largely coincided with the profile obtained for genes with long upstream segments (Table 2). Likewise, profiles for genes with few motifs matched profiles obtained for genes with short upstream regions. As the motif density was observed to be relatively constant across intergenic regions (Figure S4), this agreement is not surprising because motif counts are positively correlated with intergenic distances. When we confined the promoter count to upstream regions of 500 nucleotides, i.e., analyzing the density of motifs in 500 nucleotides, only one GO process (response to wounding) was borderline significant (false discovery rate [FDR] *p*-value = 0.09). No other GO process category or gene family association was found to be significantly correlated with either high or low motif density (FDR *p*-value > 0.1 for all categories and gene families). However, "response to wounding" and "response to cold" were ranked highest among GO processes associated with high motif density. Several other response-related categories were also contained among the top-ranked processes, such as "response to oxidative stress" (rank 6) and "response to gibberellic acid stimulus" (rank 16) and, as a single-word category, "response" was very significantly associated with high motif density (FDR *p*-value = 5.1E–7), suggesting that higher motif densities are indeed associated with genes that are involved in environmental response pathways.

Well-known stress response elements in plants such as the ABRE-like, ABF, and DRE binding site motifs [16] are among the motifs found most frequently in genes with large breadth of response (Table 3). The occurrence of the TATA-box motif, a commonly found eukaryotic core promoter element involved in transcription initiation and usually located approximately 25 to 30 nucleotides upstream of the transcription start site [17], also appears to be preferentially associated with multistimuli response genes. Genes with a putative TATA-box motif present in the 60 nucleotides upstream of the transcription start site had a significantly greater breadth of response (mean number of experiments with differential expression = 3.4, $n = 2{,}002$) than did genes lacking a TATA-box motif in their 60-nucleotide upstream region (mean number of experiments with differential expression = 1.9, $n = 9{,}795$, $p = 5.0\text{E}{-}113$). Consistent with increased intergenic upstream distances for multistimuli response genes, TATA box–containing genes were also observed to have longer upstream regions (2,370 nucleotides versus 1,737 nucleotides for TATA-less genes, $p = 3.4\text{E}{-}34$). Motifs reported to control housekeeping genes (TELO-box promoter motif [18]; hexamer promoter motif [19]) or implicated to confer tissue-specific expression (LEAFYATAG [20]) were found to be more associated with genes with

Table 2. GO Process and Gene Family Annotations Overrepresented in Genes with Long (Left) and Short (Right) Intergenic Distances to Next Upstream Gene (FDR *p*-Value < 0.1)

Annotation category	2,000 Genes with Longest Intergenic Upstream Regions		5,000 Genes with Shortest Intergenic Upstream Regions	
	FDR *p*-Value	Category	FDR *p*-Value	Category
GO Process	4.73E−19	Regulation of transcription	8.75E−15	Biological process unknown
	3.92E−13	Response to auxin stimulus	1.19E−13	Protein biosynthesis
	4.64E−11	Regulation of transcription, DNA dependent	9.29E−04	Ribosome biogenesis
	4.95E−07	Protein amino acid phosphorylation	7.15E−03	Metabolism
	2.42E−04	Transmembrane receptor protein tyrosine kinase signaling pathway	1.92E−02	Intracellular protein transport
	4.03E−04	Oligopeptide transport	2.50E−02	Biosynthesis
	1.80E−03	Response to gibberellic acid stimulus	4.91E−02	Protein transport
	5.49E−03	Response to jasmonic acid stimulus	6.07E−02	Ubiquitin-dependent protein catabolism
	7.42E−03	Unidimensional cell growth	7.93E−02	TRNA aminoacylation for protein translation
	7.72E−03	Response to abscisic acid stimulus	9.77E−02	Electron transport
	1.53E−02	Flower development	—	—
	1.98E−02	Gibberellic acid–mediated signaling	—	—
	2.34E−02	Response to cytokinin stimulus	—	—
	4.50E−02	Multidrug transport	—	—
	4.75E−02	Trehalose biosynthesis	—	—
	4.96E−02	Response to ethylene stimulus	—	—
	5.86E−02	Response to salicylic acid stimulus	—	—
	8.67E−02	Cell wall loosening (sensu Magnoliophyta)	—	—
	8.92E−02	Response to nematode	—	—
	9.80E−02	Cell wall organization and biogenesis (sensu Magnoliophyta)	—	—
Gene Family	1.02E−06	Transcription factor	1.73E−13	Cytoplasmic ribosomal protein gene family
	2.60E−03	Nodulin-like protein family	2.66E−02	Sulfurtransferasese/rhodanese family
	8.87E−03	GRAS proteins	2.70E−02	Chloroplast and mitochondria gene families
	9.43E−03	MIP family	4.44E−02	Acyl lipid metabolism family
	1.00E−02	MYB	8.27E−02	Protein synthesis factors
	2.91E−02	Organic solute cotransporters	8.62E−02	GST superfamily
	9.37E−02	Receptor kinase–like protein family	—	—
	1.00E−01	C2H2 zinc finger proteins	—	—

doi:10.1371/journal.pgen.0030011.t002

narrow range of differential expression response, i.e., their expression is constitutive or their differential expression response is very specific.

It is conceivable that in transcriptional regulatory signaling cascades, early responders evolved a greater sensory capacity (i.e., respond to many different transcription factors) and then channel the response to common effector genes. In our framework, equating sensory potential to the number of *cis*-elements, we then expect early responders to be associated with a greater number of regulatory elements. The 18 AtGenExpress abiotic stress time series datasets (nine for root and shoot tissue, respectively) allowed testing of this hypothesis. Comparing genes that are differentially expressed during the first 3 h after stimulus to genes that respond at later time points and assuming a fixed promoter length of 500 nucleotides, early-response genes were associated with only marginally or no increased motif count densities with 3% (2%) more (nonredundant [nr]-unique) motifs in 500 nucleotide–upstream regions compared to late genes (7.5 versus 7.3, $p = 5.9E{-}12$; 5.8 versus 5.7, $p = 3.4E{-}6$ for nr-unique motifs, respectively). However, in allowing promoter lengths of up to 3,000 nucleotides and truncating them at neighboring upstream gene sites (variable maximal promoter lengths), early genes were observed to harbor 20% (16%) more (nr-unique) motifs than late genes (34.4 versus 28.5, $p = 2.1E{-}170$; and 13.2 versus 11.4, $p = 1.2E{-}179$ for nr-unique motifs, respectively). As was observed before, an increased number of *cis*-elements for a specific group of genes may originate from two sources: higher motif density and more different motifs and, apparently more significantly, longer upstream regions providing more space for potential regulatory sites. Early-response genes have, on average, 30% longer intergenic upstream segments than do late-response genes (2,668 nucleotides versus 2,064 nucleotides, $p = 2.9E{-}115$) and are more likely to contain TATA-boxes (25% versus 19%, Fisher exact *p*-value = 1.4E–12). Consistent with a transcriptional regulatory signaling cascade, genes involved in the regulation of transcription, signaling, as well as stress response genes are overrepresented among early-responder genes, whereas ribosomal genes and genes generally involved in protein

Table 3. *cis*-Regulatory Elements (Motifs) Associated with Genes with High (Left) versus Low (Right) Breadth of Differential Gene Response and vice versa at Significance Level of FDR *p*-Value < 0.1

2,000 Genes with Highest Breadth of Response		5,000 Genes with Lowest Breadth of Response	
FDR *p*-Value	***cis*-Element**	**FDR *p*-Value**	***cis*-Element**
1.74E−18	ABRE-like binding site motif	8.87E−07	TELO-box promoter motif
7.83E−18	ACGTABREMOTIFA2OSEM	7.18E−06	Hexamer promoter motif
6.71E−09	CACGTGMOTIF	7.76E−06	MYB1AT
7.31E−07	ABFs binding site motif	2.60E−05	CCA1 binding site motif
2.36E−05	AtMYC2 BS in RD22	9.04E−05	MYB4 binding site motif
2.93E−05	TATA-box motif	2.05E−03	BoxII promoter motif
5.42E−05	GBOXLERBCS	4.33E−03	LEAFYATAG
4.90E−04	GBF1/2/3 BS in ADH1	6.76E−03	MYB2 binding site motif
1.96E−03	Octamer promoter motif	6.90E−03	GAREAT
2.09E−03	ABREATRD22	7.43E−03	GCC-box promoter motif
5.19E−03	DREB1A/CBF3	8.61E−03	CDA1ATCAB2
5.41E−03	TGA1 binding site motif	1.63E−02	CCA1 motif1 BS in CAB1
7.70E−03	I-box promoter motif	6.56E−02	MYB binding site promoter
1.63E−02	DRE core motif	7.85E−02	T-box promoter motif
1.93E−02	Z-box promoter motif	7.94E−02	RAV1-B binding site motif
5.55E−02	ABRE binding site motif	8.99E−02	CArG promoter motif
8.29E−02	Evening element promoter motif	9.28E−02	AGATCONSENSUS

Only motifs located in the 500 nucleotides upstream of transcription start sites have been considered, including counts for repeated occurrences of the same motif.
doi:10.1371/journal.pgen.0030011.t003

biosynthesis, metabolism, and other nonsignaling processes are more characteristic of late-response genes (Table 4).

Discussion

The objective of our investigation was to test whether there is a significant relationship between encoded, by way of *cis*-regulatory motifs, and actually observed differential gene expression diversity. Our primary focus has been the relationship of *cis*-regulatory motif counts and differential gene expression using data from *A. thaliana*. In discussing our results, we will first consider several technical aspects of this study and then turn to the biological relevance of our findings.

The currently available data for both data types, *cis*-element motifs and differential gene expression, are by their very nature associated with high levels of noise. The set of *cis*-elements for every gene was obtained by computational mapping of short sequence motifs to upstream gene promoter regions. Without experimental verification, a large number of such predicted sites are expected to be false positives as it is difficult to determine how far upstream the promoter region actually extends, with the exceptions that promoters may not overlap with neighboring gene segments, and where the boundaries of the core-promoters—the site of constitutive transcription initiation with less differential regulatory function—are located. Furthermore, alternative transcription start sites have been reported for up to 50% of the *Arabidopsis* genes [21], associated perhaps with separate regulatory regions. There are also likely to be more, as of yet unknown, motifs than the 93 that were considered here, as there are about 1,500 transcription factors encoded in the *Arabidopsis* genome [22].

On the expression side, detecting true differential expression with only few technical and biological replicate samples is difficult. We applied a simple criterion for calling differential expression as the objective has not been to identify high-confidence gene expression markers for the various conditions but rather to detect global quantitative trends.

However, the fact that we still observed significant correlations between both quantities, despite the high noise levels, suggests that the true correlation between both quantities may, in fact, be even more significant and of greater magnitude than estimated in this study. This is particularly relevant for the investigated correlations between density of motifs and breadth of response (Figure 2).

On the other hand, high noise levels increase the risk of artefacts. One such source of artificial distortion of the data may be the normalization of raw gene expression values. To test the robustness of the observed correlations of motif counts to the breadth of differential gene expression response, we applied two additional normalization techniques (MAS 5.0; Affymetrix proprietary software, http://www.affymetrix.com) and GCRMA (Affymetrix) [23] and compared the results to the data obtained from using Robust Microchip Analysis (RMA; Affymetrix) [24]. Almost identical trends of motif counts in the 500-nucleotide gene upstream regions to the breadth of response have been obtained for all three normalization methods (Figure S5).

In limiting the analysis to genes with expression levels for which no dependence of the breadth of differential response on the expression level was noticeable (Figure S1), we applied additional caution to guard against artefacts. However, the primary observations and conclusions presented here also hold when all genes, including the 10,000 genes with low expression levels in control samples, are used in the analyses.

We have counted *cis*-elements in upstream region as they are presented in the Athena resource regardless of their mapping orientation (forward or reverse strand). Therefore, we compared results from using only motifs that map in the same orientation as the coding strand for a given transcript and obtained qualitatively identical results. Evidently, abso-

Table 4. GO Process and Gene Family Annotations Associated with Genes Differentially Expressed Early (3 h or Earlier [Left]) and Late (After 3 h [Right]) after Abiotic Stress Stimulus in Nine different Abiotic Stress Time Series Experiments and in Root and Shoot Tissue, Respectively (Table S1)

Annotation Category	Early Responding Genes		Late Responding Genes	
	FDR *p*-Value	Category	FDR *p*-Value	Category
GO Process	2.90E−22	Response to abscisic acid stimulus	1.21E−79	Protein biosynthesis
	2.62E−19	Regulation of transcription, DNA dependent	3.24E−30	Ribosome biogenesis
	6.09E−13	Response to cold	1.53E−11	Nucleosome assembly
	1.63E−08	Response to jasmonic acid stimulus	2.63E−11	Chromosome organization and biogenesis (sensu Eukaryota)
	6.14E−08	Response to auxin stimulus	5.03E−09	RNA processing
	6.30E−08	Response to wounding	3.43E−08	Microtubule-based movement
	6.49E−08	Regulation of transcription	1.52E−05	DNA replication
	4.21E−07	Response to water deprivation	2.28E−05	Proteolysis
	4.27E−07	Response to salt stress	3.70E−04	DNA replication initiation
	4.99E−07	Ethylene-mediated signaling pathway	4.61E−04	Glycolysis
	1.21E−06	Defense response	7.88E−04	Regulation of progression through cell cycle
	1.66E−06	Response to gibberellic acid stimulus	1.21E−03	Photosynthesis
	2.85E−06	Trehalose biosynthesis	1.34E−03	Cellular protein metabolism
	6.83E−06	Response to ethylene stimulus	1.80E−03	rRNA processing
	1.02E−05	Response to cytokinin stimulus	2.51E−03	Translational initiation
	1.33E−05	Response to water	3.21E−03	Intracellular protein transport
	1.47E−05	Response to salicylic acid stimulus	3.49E−03	Ubiquitin-dependent protein catabolism
	1.48E−05	Hyperosmotic salinity response	3.95E−03	Ribosome biogenesis and assembly
	1.90E−05	Abscisic acid–mediated signaling	5.72E−03	Zinc ion homeostasis
	3.42E−05	Toxin catabolism	5.72E−03	Metabolism
	3.80E−05	Exocytosis	5.84E−03	Embryonic development (sensu Magnoliophyta)
	5.08E−05	Cold acclimation	6.13E−03	Pentose-phosphate shunt
	3.70E−04	Response to heat	6.25E−03	Translational elongation
	4.54E−04	Response to desiccation	6.57E−03	Electron transport
	5.14E−04	Cytokinin catabolism	6.82E−03	Trichome morphogenesis (sensu Magnoliophyta)
	6.35E−04	Response to chitin	7.36E−03	Photorespiration
	2.34E−03	Negative regulation of abscisic acid–mediated signaling	8.08E−03	Nuclear mRNA splicing, via spliceosome
	2.83E−03	Response to stress	—	—
	3.44E−03	Vesicle docking during exocytosis	—	—
	3.48E−03	L-Phenylalanine biosynthesis	—	—
	3.69E−03	Response to cadmium ion	—	—
	7.24E−03	Protein ubiquitination	—	—
	7.65E−03	Induced systemic resistance, jasmonic acid–mediated signaling pathway	—	—
	9.59E−03	Auxin homeostasis	—	—
Gene Family	8.51E−07	Plant U-box protein (PUB) class III	2.24E−59	Cytoplasmic ribosomal protein gene family
	1.49E−05	WRKY transcription factor superfamily	1.47E−08	Kinesins
	1.08E−04	Trehalose biosynthesis gene families	1.57E−08	Core cell cycle genes
	1.23E−04	Lateral organ boundaries gene family: class II	4.32E−04	Protein synthesis factors
	1.78E−04	PP2C-type phosphatases	1.40E−02	Subtilisin-like serine proteases
	4.42E−04	MYB	1.92E−02	Single gene-encoded CBPs
	1.76E−03	Patatin-like protein family	4.38E−02	Inorganic solute cotransporters
	5.76E−03	GST superfamily	4.59E−02	Other SNAREs
	6.11E−03	AtCIPKs	4.83E−02	Phosphoribosyltransferases (PRT)
	7.88E−03	EF hand–containing proteins: group III	7.19E−02	β-1,3-Glucanase family
	9.97E−03	Transcription factor	9.85E−02	Lipid metabolism gene families
	1.91E−02	Class III peroxidase	—	—
	2.82E−02	Basic region leucine zipper (bZIP) transcription factor	—	—
	4.45E−02	Nodulin-like protein family	—	—
	4.71E−02	Response regulator	—	—
	5.89E−02	Heat shock transcription factors	—	—
	7.12E−02	Receptor kinase–like protein family	—	—
	7.16E−02	HSP70s	—	—
	7.37E−02	CDPKs	—	—
	7.80E−02	EF hand–containing proteins: group IV	—	—
	9.74E−02	Putative glutathione transferase family	—	—

The chosen significance level was FDR *p*-value < 0.01 for GO process and FDR *p*-value < 0.1 for gene family annotations, respectively.
doi:10.1371/journal.pgen.0030011.t004

lute counts were lower. In addition, there is evidence that *cis*-elements can be recognized in both orientations (bidirectional motifs). For example, correlated expression of genes transcribed in opposite direction that are in close proximity to one another suggests that *cis*-elements are shared between the two genes regardless of direction [25].

Regulatory Capacity Is Encoded by Both Motif Density and Absolute Motif Counts Correlated to Intergenic Upstream Space

In this study, we followed the notion that regulatory capacity is associated with information-bearing properties of the intergenic region upstream of the transcription start site of the regulated genes. The simplest and most accessible measure of information content was to correlate *cis*-regulatory motif counts to the breadth of differential gene expression response. We found that increased breadth of response is indeed positively correlated with greater motif density (Figure 2). While the observed correlations were highly significant, their magnitude was generally low. The positive correlation between breadth of response and motif count was relatively strongest for the first 500 nucleotides upstream compared to the other investigated regions, suggesting that this interval may generally be the most relevant promoter segment to control gene expression. Apart from greater density, more motifs and generally greater information content can also arise from more available intergenic space. Analyzing gene expression profiles in different developmental and cell differentiation stages in *C. elegans* and *D. melanogaster,* Nelson and coworkers [10] observed that genes with more complex expression profiles were indeed associated with larger flanking intergenic intervals. Genes with regulatory functions were shown to generally have longer upstream regions, whereas genes involved in housekeeping functions have shorter upstream segments. The results reported here for *Arabidopsis* and analyzing differential expression data for transcriptional response to external stimuli confirm these findings. The requirements for encoding regulatory response complexity to external environmental challenges also appear to have played a role in shaping the layout of the *Arabidopsis* genome. Not conflicting with this result, we saw that motif density also is varied in correlation with different complexity of transcriptional response (Figure 2), suggesting that both principal mechanisms (greater density and/or more space) were employed by evolution to encode complex transcriptional response patterns. While both flanking sequences (upstream and downstream) showed positive correlation of their length to the breadth of response, we found that in *Arabidopsis,* upstream distances were more strongly correlated with response diversity and were generally longer (Figure 4A) than downstream intervals, suggesting that the information content upstream of a gene may be more relevant to the encoding of regulatory properties than downstream segments. Apart from transcription factor binding sites, other mechanisms and motifs, such as histone binding sites, chromatin structural changes, enhancers, silencers, and insulators, as well as sites of epigenetic regulation, may also constitute regulatory elements contained in intergenic segments that can contribute to more complex regulatory properties encoded in longer upstream intervals.

Alternative to our assumption that increased breadth of response correlates with increased regulatory capacity, it might be speculated that because genes that respond to many different stimuli in a similar manner, they are regulated by a common regulatory mechanism and therefore should be expected to be regulated by fewer rather than more *cis*-elements. However, data presented by Gasch et al. [3] on the gene expression response to environmental stress in yeast and the results presented here indicate that multistress response genes appear to be controlled by different and condition-specific regulatory mechanisms.

It has been recognized that often multiple factors are involved in the initiation of transcription, suggesting a combinatorial control of transcription initiation [26,27]. A combinatorial use of the repertoire of *cis*-regulatory elements enlarges the regulatory coding capacity tremendously. Conceptually, the basic premise of this investigation—a positive correlation between encoded and observed differential gene response—also applies in the case of combinatorial control.

Multistimuli Response Genes Are Shorter in *Arabidopsis*

The correlation between expression level and gene size and associated properties such as number and length of introns has been investigated in several studies [28–34]. Surprisingly, while highly expressed genes were found to be shorter in animals and to harbor fewer and shorter introns [30], they appear least compact and relatively largest in plants [33] and *C. elegans* [29]. Several evolutionary scenarios such as selection for economy and genomic design have been discussed to explain the observed trends [31,34]. Using the AtGenExpress dataset analyzed in this study, we also observed a general increase in gene size, intron length, and intron number with expression level in *Arabidopsis*, albeit for very highly expressed genes, the trend was reversed (Figure S7).

The focus of this study has been on the correlation between breadth of differential genes expression response, not its absolute level, and genome architectural properties. It can be assumed that differential gene expression is largely independent of expression level with the obvious boundary effects that lowly expressed genes are more likely to be up-regulated and highly expressed genes have a greater chance of being down-regulated as certain expression levels cannot be exceeded. For medium expression levels, indeed no such correlation was found in our dataset and deviations at low levels likely caused by noise (Figure S1).

Interestingly, we observed that properties related to gene size (gene length, cDNA length, number of introns) are strongly and negatively correlated with breadth of differential gene expression response (Figure 4). While this may reflect a coincidence that informative (involved in regulation or signaling) gene families are generally smaller, the intriguing question arises of whether transcriptional efficiency may have played a role during evolution. Shorter genes with fewer introns may be produced more economically and thereby quicker in response to sudden external stimuli. In fact, the number of introns was the strongest, albeit marginally, of all properties analyzed to correlate with breadth of response (Figure 4 and multiple linear regression results). Informative molecules may also be smaller as they need to diffuse within the cell or tissues to carry their information to the place of perception rather than being constituents of larger and more static cellular machineries. Clearly, comparing our findings in

Arabidopsis with those in warm-blooded animals will shed further light on the generality of our observations.

Special Role of TATA-Boxes

Among the *cis*-regulatory motifs considered in the analysis, the preferential association of the TATA-box motif—a commonly found eukaryotic core promoter element involved in transcription initiation and usually located approximately 25 to 30 nucleotides upstream of the transcription start site [17]—with multistimuli response genes is of particular interest. In the gene set studied here, 17% of all genes contained TATA-box motifs within the first 60 nucleotides upstream of the transcription start. TATA-box genes were found associated with stress response and were shown to be subject to chromatin remodeling factors consistent with their regulation by nucleosomal mechanisms [35]. The TATA-box motif was also described recently to confer increased interspecies gene expression variation of the corresponding genes [36]. Our observation that TATA-box–containing genes have longer intergenic upstream regions is consistent with their chromatin and nucleosomal regulation and also suggests that the expression of TATA-box genes may evolve at higher rates, causing increased variation across species because their upstream regulatory potential is greater and, therefore, more amenable to change and modulation. Interestingly, presence of the TATA-box also strongly correlates with a decreased number of introns in the downstream gene (number of introns for genes with TATA: 3.7, without: 5.9; $p = 1.9E{-}93$). It is presently not clear whether this correlation indicates a direct coupling between transcription and splicing [37] or whether it originates from an indirect correlation via a common principle shaping the genome, such as possibly the breadth of response and the role of both properties in defining it as reported here.

Multistimuli Response Genes Have More Paralogs

Assuming that multistimuli response or "informational hub" genes play more critical roles than genes that have a narrower response scope, it can be speculated that failures of such central genes—by spontaneous disruptive mutations, for example—may be more detrimental to the organism. Therefore, functional backup genes may have evolved to safeguard against such failures. A plausible evolutionary solution would be to retain copies or paralogs of hub gene in the genome from gene or segmental duplication events, i.e., genes with identical or similar function [38,39]. The observed increase of the number of paralogs with increasing breadth of response is consistent with the concept of selection for functional backup (Figure 4F). However, results reported in yeast suggest that rather than redundancy provided by duplicated genes, interactions between unrelated genes appear to be responsible for robustness against mutations [40]. Furthermore, rather than a static robustness provided by "replacement parts," a dynamic reprogramming of the transcriptional regulatory network may be employed during "fail-safe" scenarios [41].

We saw that multistimuli response genes were generally associated with environmental response processes (Table 1). As the number of different external stimuli is large and fine-tuning the perception as well as signal transduction depending upon the stimulus will likely be beneficial, it may have been evolutionarily advantageous to use duplication events to evolve genes (paralogs) with shifted and novel response scope, also explaining a greater number of paralogs for multistimuli response genes [42].

Alternatively, paralogs may allow a greater dynamic range of the response to external stimuli. Corresponding dosage effects and their role in the retention of paralogs have recently been discussed in the literature [43].

Stress Response Genes Have Greater Breadth of Differential Response—A Potential Experiment Bias

We observed that genes implicated in the response to specific stresses (e.g., cold) are also among the genes with a very broad range of differential gene expression in response to various environmental changes (Table 1). However, some of the applied external changes correspond to very similar environmental cues (salt and osmotic stress, for example) and common transcriptional response programs will likely be triggered. Ideally, only very different and unrelated external stimuli would be used in our analyses, but imposing such requirements would reduce the number of experiments to very low numbers. Therefore, it needs to be borne in mind that the similarity between different external conditions and the resulting relative overrepresentation of particular types of external stimuli may cause a bias toward certain gene functional categories (salt stress response, for example). Interestingly, when the abiotic stress series was excluded from the list of experiments (about half of all experiments, Table S1), general stress response GO categories, including abiotic stress response, were still overrepresented among the genes with high breadth of response ("response to wounding," FDR p-value = 3.7E–7; "response to oxidative stress," FDR p-value = 5.5E–4; "response to heat," FDR p-value = 0.04), suggesting that some stress response genes may truly display a large and diverse range of differential response.

Characteristic Sequence Signatures in Longer Upstream Regions of Multistimuli Response Genes?

If, as observed, longer upstream regions are associated with greater gene expression complexity of downstream genes, the question emerges of whether there are characteristic sequence motifs in longer regions that are not found in short intergenic upstream segments. Such motifs may help identify and elucidate new regulatory elements and even mechanisms, DNA structural properties, and their influence on gene regulation, for example. Recent reports that large fractions of the nontranslated segments of genomes are functionally important based on analyses of mutation rates [44] and the greater-than-expected occurrences of specific sequence patterns in noncoding DNA segments associated with genes involved in signaling and transcription regulation processes [45] strongly encourage the pursuit of further research in this direction.

As sessile plants cannot evade changing and adverse environments, they may rely more strongly on elaborate transcriptional response programs and may thus serve as ideal model systems for the study of the regulatory code in the genomes of higher organisms.

Materials and Methods

Gene expression data. Gene expression information was obtained from AtGenExpress. Profiling data based on the ATH1 Affymetrix GeneChip microarray platform [46] from the abiotic stress, biotic,

nutrient, hormone, and chemical treatment-control series have been used. A detailed list of the 43 experiments, included samples, and a brief description of the various conditions, is available in Table S1.

Gene expression data normalization. Raw expression data files were obtained for treatment and associated control samples. All samples associated with a particular treatment-control experiment were preprocessed and normalized together using the Affy-package [47] available in the Bioconductor software suite [48]. Raw expression level data were normalized applying three different methods: RMA [24], GCRMA [23], and MAS 5.0 (Affymetrix). Unless noted otherwise, results are based on RMA normalization as the default normalization method.

Differential expression criteria and probe selection, breadth of response. Genes were considered differentially expressed if $|m_T - m_C| > 1$ and $s > 0.5$ where $s = |m_T - m_C|/(\sigma_T + \sigma_C)$ and m_T, m_C denote the mean logarithm-base-2 transformed expression levels for associated gene probes across the available treatment (t) and control (c) replicate samples, and σ_T, σ_C are the associated standard deviations of the log-2 expression levels. The first condition corresponds to a threshold of minimally 2-fold up-regulation/down-regulation of treatment expression levels relative to the levels in the associated control samples. The second criterion was introduced as a simple statistical significance measure similar to a simplified *t*-test. By normalizing to the standard deviation, rather than the standard error as done in the *t*-test, the risk of biasing the results to experiments with more repeats was reduced. The test metric s was introduced by Golub et al. [49] in their seminal study on cancer classification based on gene expression. To test whether the results are robust with regard to the chosen threshold values, other, more stringent, parameter values were applied. Qualitatively similar results were obtained (Figure S6). All mitochondrial and chloroplastidial gene probes were discarded. Only probes that map uniquely to annotated genes and with available *cis*-regulatory motif information have been considered. To ensure independence of the frequency of detected differential expression events on the absolute expression level and to avoid possible normalization artefacts, we examined for all gene probes the number of experiments in which a gene probe was observed to be differentially expressed as a function of the median rank of this probe across all control samples in the dataset and compared the results for all three normalization methods (Figure S1). For probe ranks greater than 10,000, no significant influence of the absolute expression level on the frequency of differential expression was observed. Therefore, we discarded the 10,000 probes with lowest median rank across all control samples from the analysis. Analyses using all probes have also been conducted and confirm the presented results. Applying the above criteria, 11,797 *Arabidopsis* genes were used in the expression data analysis using their unambiguously mapping array probes, i.e., every probe mapped to only one gene and every gene mapped to only one probe. Mapping information was obtained from The *Arabidopsis* Information Resource (TAIR) [50] and the ATH1-chip annotation file provided by Affymetrix. We define the term "breadth of response" for every gene as the cumulative number of treatment-control experiments in which the gene was found to be differentially expressed.

***cis*-Regulatory elements.** Transcription factor binding site location information was obtained from the Athena promoter annotation resource [15]. In Athena, transcription factor site coordinates are obtained by sequence mapping of consensus motif sequences imported from the Plant *cis*-acting regulatory DNA elements (PLACE) [51] and *Arabidopsis* Gene Regulatory Information Server (AGRIS) [52]. The dataset contained mapping information for 93 different and previously characterized binding site motifs in the 5′ upstream gene segments of up to 3,000 nucleotides in length associated with genes and corresponding probes present on the ATH1 microarray platform.

Construction of sets of an nr-unique *cis*-regulatory elements. Mapping the 93 Athena motifs to gene promoter regions results in several redundancies and motif overlaps that were eliminated in order to construct truly unique sets of motifs and corresponding motif counts associated with every gene. Motifs were sorted alphabetically for every gene and assessed for uniqueness in consecutive order. Motifs whose mapping location either fully or only partially overlaps with already accepted motifs were not included in the unique set of motifs. Thus, redundancy was eliminated based on motif sequence information. The obtained set of unique *cis*-elements comprised a total of 76 different motifs, i.e., 17 motifs were excluded as they always overlapped with other motifs. For example, the motif DREB1A/CBF3 (consensus sequence "RCCGACNT") contains the DRE core motif (consensus sequence "RCCGAC"). Thus, the DREB1A/CBF3 element was never included as a separate motif in the unique set as it always contains the DRE core motif which is sorted alphabetically before it. Multiple occurrences of the same motif at different locations in gene promoters were collapsed to only single counts in the nr-unique *cis*-element set.

Correlation analysis and significance levels. The strength and magnitude of the association of the various investigated gene properties to the number of experiments in which gene probes were observed differentially regulated (breadth of response) was quantified using linear regression and the linear Pearson correlation coefficient, *r*. Two types of correlation statistics have generally been computed: (1) the correlation of all value pairs for all probes and (2) for mean values only, for example, the mean number of *cis*-regulatory elements for genes differentially expressed in *n* different experiments. A minimum of at least 100 probes was required for mean values to be included in the latter, thus excluding mean values of lesser confidence. The significance of the correlation was assessed using standard *p*-value calculation for correlation coefficients based on the *t*-statistic with $t = r \times \mathrm{sqrt}[(N-2)/(1-r^2)]$, and corresponding two-tailed *p*-values were computed from the *t*-distribution where *N* is the number of value pairs to be correlated. In addition to this parametric significance testing, a nonparametric method based on data shuffling was implemented. The pairing of data from two data vectors that were to be tested for correlation was randomly shuffled 10,000 times; i.e., one vector was repeatedly randomly shuffled. The count, C_S, how often the magnitude of the correlation coefficient exceeded the actually observed coefficient for the unshuffled data, divided by the total number of shufflings, N_S, served as a nonparametric *p*-value estimate, p_S, for the correlation coefficients such that $p_S = C_S/N_S$. In almost all cases, the obtained p_S-value was zero, i.e., no shuffled correlation coefficient was obtained with greater correlation than the actually observed one. We therefore only list p_S-values in the few cases with nonzero values (Figure 2). All *p*-values reported for *t*-test comparisons of mean values correspond to the two-tailed value.

GO data and categorical correlations by Fisher exact tests. GO information was obtained from TAIR [50]. Association tests of categorical gene classification data (GO annotations or *cis*-regulatory elements) with two gene sets were performed using the Fisher exact test. The two one-tailed Fisher exact *p*-values corresponding to overrepresentation or underrepresentation of categories in the two sets relative to one another have been calculated based on counts in 2 × 2 contingency tables. Counts n_{11}, n_{12}, n_{21}, and n_{22} in the contingency table refer to n_{11}, number of observations of a particular category in the first gene set; n_{12}, number of other categories in the first gene set; n_{21}, number of observations of category in second gene set; and n_{22}, number of observations of other categories in the second gene set. Listed *p*-values correspond to multiple testing-corrected Fisher exact *p*-values using the FDR method ([53]).

Gene and genomic sequence and gene mapping information. Gene and genomic sequence information and gene mapping information to the *Arabidopsis* genome sequence, including intergenic distances, number of introns, gene, cDNA, and 5′/3′ UTR length information, and all sequence information, were obtained from TAIR database release 6 [50]. For 2,238 of the 11,797 genes, more than one transcript sequence was contained in the TAIR dataset (annotated splice variants). For those genes, always the first instance was taken (".1"-transcript) for the study of transcript-related properties such as UTR length or cDNA length. We verified that similar results were obtained when taking different instances and that the number of splice variants per gene was not correlated with the breadth of response. It should be borne in mind that the ATH1 probes were designed to target the 3′ end of gene transcripts [46] and, therefore, identification of various splice forms of genes is difficult or impossible.

Identification of gene parologs. Duplicated genes within the *Arabidopsis* genome were identified following a method introduced by Gu and coworkers [54,55]. This method is based on an all-against-all sequence comparison of protein sequences that are not splice variants of the same gene and require at least 30% amino acid sequence identity with adjusted higher thresholds for short sequences. A second grouping of genes into paralogous gene families was generated applying a more stringent threshold of 70% amino acid sequence identity.

Multiple linear regression. Multiple linear regression was performed using Statistica 7.1 (StatSoft, http://www.statsoft.com). Forward stepwise regression was applied with F to enter set to 1 and casewise missing value deletion.

Supporting Information

Figure S1. Relationship between Number of Experiments (NE) in Which Genes Were Differentially Regulated as a Function of Absolute

Expression Level Measured by the Median Rank of Gene Expression Levels Across All Control Experiments

Curves correspond to the 500 point-running averages for the *x*-axis sorted data for the three different applied expression data normalization techniques.

Found at doi:10.1371/journal.pgen.0030011.sg001 (847 KB EPS).

Figure S2. Correlation of the Number of *cis*-Regulatory Elements in 500-Nucleotide Upstream Promoter Regions of Varying Length with the Breadth of Differential Gene Expression Response

Red data points refer to counts of *cis*-elements in differentially up-regulated genes, whereas the green data points show data for down-regulated genes in treatment samples versus the respective control samples. Only genes with longer than 500-nucleotide intergenic upstream regions have been considered. Shown are the mean values and the standard errors of the mean (SEM). The *p*-values associated with the observed linear Pearson correlation coefficient are given for all raw data pairs (all gene probes and associated gene properties) and, in parentheses, for the mean values as they are plotted in the graph. Only mean values with more than 100 raw observations (genes) are plotted.

Found at doi:10.1371/journal.pgen.0030011.sg002 (16 KB EPS).

Figure S3. Correlation of the Length of the Intergenic Space in Nucleotides Preceding Genes with Their Breadth of Differential Gene Expression Response

Red data points refer to differentially up-regulated genes, whereas the green data points show data for down-regulated genes in treatment samples versus the respective control samples. Shown are the mean values and the SEM. The *p*-values associated with the observed linear Pearson correlation coefficient are given for all raw data pairs (all gene probes and associated gene properties) and, in parentheses, for the mean values as they are plotted in the graph. Only mean values with more than 100 raw observations (genes) are plotted.

Found at doi:10.1371/journal.pgen.0030011.sg003 (15 KB EPS).

Figure S4. Motif Distribution in Upstream Intergenic Regions without Any Gene Overlaps (Green Histogram), with Possible Preceding Genes Falling into the 3,000-Nucleotide Intergenic Upstream Region

Regulatory motifs are less frequent in *intra*genic regions.

Found at doi:10.1371/journal.pgen.0030011.sg004 (16 KB EPS).

Figure S5. Mean Number of *cis*-Elements in 500-Nucleotide Gene Upstream Regions and Associated Standard Error of the Mean as a Function of Observed Breadth of Differential Gene Expression Response for Three Different Expression Data Normalization Techniques as Indicated in the Figure Legend

The correlation coefficients *r* and associated *p*-values for all raw (mean) data value pairs are given in the figure legend.

Found at doi:10.1371/journal.pgen.0030011.sg005 (20 KB EPS).

Figure S6. Correlation of the Number of *cis*-Regulatory Elements in the 500-Nucleotide Gene Upstream Promoter Regions with the Breadth of Differential Gene Expression Response for Different Thresholds of Differential Expression Parameters (Fold Up-regulation or Down-regulation and *s*, see Methods)

Shown are the mean values and the SEM. Correlation coefficients, *r*, and associated *p*-values are given for all raw data pairs. Note that for more stringent threshold values (e.g., 4-fold versus 2-fold), fewer genes are considered differentially expressed leading to lowered breadth of response values.

Found at doi:10.1371/journal.pgen.0030011.sg006 (17 KB EPS).

Figure S7. Mean Gene Length, Intron Length (All Intronic Segments per Genes Added Up), and Number of Introns of Genes as a Function of Their Expression Levels Measured by the Median Rank of Gene Expression Levels across All Control Experiments and Binned into Bins of Width 1,000

Found at doi:10.1371/journal.pgen.0030011.sg007 (31 KB EPS).

Table S1. Set of 43 AtGenExpress Treatment-Control Gene Expression Profiling Experiments Used in the Analysis

Found at doi:10.1371/journal.pgen.0030011.st001 (25 KB XLS).

Acknowledgments

We are grateful to John Wyrick and team for making the Athena dataset available to us. We thank Matthew Hannah, Judith Gomez, Renate Schmidt, and Alisdair Fernie for helpful discussions and comments on the manuscript. We also thank Sarah Teichmann for her comments on paralogs and differential gene expression.

Author contributions. DW conceived and designed the study, performed the analyses, and wrote the paper. RB processed and analyzed the data. JS suggested analyses and provided cosupervision to RB.

Funding. The authors received no specific funding for this study.

Competing interests. The authors have declared that no competing interests exist.

References

1. Skriver K, Mundy J (1990) Gene expression in response to abscisic acid and osmotic stress. Plant Cell 2: 503–512
2. Shinozaki K, Yamaguchi-Shinozaki K (1997) Gene expression and signal transduction in water-stress response. Plant Physiol. 115: 327–334
3. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257
4. Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 434: 147–151
5. Balázi G, Oltvai ZN (2005) Sensing your surroundings: How transcription-regulatory networks of the cell discern environmental signals. Sci STKE 282: pe20
6. Jenuwein Th, Allis CD (2001) Translating the histone code. Science 293: 1074–1080
7. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet 33: 245–254
8. Harbison ChT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104
9. Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, et al. (2004) A regulatory code for neurogenic gene expression in the *Drosophila* embryo. Development 131: 2387–2394
10. Nelson CE, Hersh BM, Carroll SB (2004) The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5: R25
11. Reznikoff WS, Siegele DA, Cowing DW, Gross CA (1985) The regulation of transcription initiation in bacteria. Annu Rev Genet 19: 355–387
12. Kuhlmeier C, Green JP, Chua NH (1987) Regulation of gene expression in higher plants. Ann Ref Plant Physiol 38: 221–257
13. Orphanides G, Reinberg (2002) A unified theory of gene expression. Cell 108: 439–451
14. Arabidopsis Genome Initiative (2001) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815
15. O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: A resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences. Bioinformatics 21: 4411–4413
16. Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel *cis*-acting element in an *Arabidopsis* gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell* 6: 251–264
17. Roeder RG (1996) The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci 21: 327–335
18. Tremousaygue D, Manevski A, Bardet C, Lescure N, Lesure B (1999) Plant interstitial telomer motifs participate in the control of gene expression in root meristems. Plant J 20: 553–561
19. Chaubet N, Flenet M, Clement B, Brignon P, Gigot C (1996) Identification of *cis*-elements regulating the expression of an *Arabidopsis* histone H4 gene. Plant J 10: 425–435
20. Kamiya N, Nagasaki H, Morikami A, Sato Y, Matsuoka M (2003) Isolation and characterization of a rice WUSHEL-type homeobox gene that is specifically expressed in the central cells of a quiescent center in the root apical meristem. Plant J 35: 429–441
21. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, et al. (2006) Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant Mol Biol 60: 69–85
22. Riechmann JL, Heard J, Martin G, Reuber L, Jiang CZ, et al. (2000) *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. Science 290: 2105–2110
23. Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc 99: 909–917
24. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucl Acids Res 31: e15

25. Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res 14: 1060–1067
26. Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nature 29: 153–159
27. Van Steensel B (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. Nat Genet 37: S18–S24
28. Eyre-Walker A (1995) The distance between *Escherichia coli* genes is related to gene expression levels. J Bacteriol 177: 5368–5369
29. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A 96: 4482–4487
30. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nat Genet 31: 415–418
31. Vinogradov AE (2004) Compactness of human housekeeping genes: Selection for economy or genomic design? Trends Genet 20: 248–253
32. Chiaromonte F, Miller W, Bouhassira EE (2006) Gene length and proximity to neighbors affect genome-wide expression levels. Genome Res 13: 2602–2608
33. Ren XY, Vorst O, Fiers MWEJ, Stiekema WJ, Nap JP (2006) In plants, highly expressed genes are the least compact. Trends Genet 22: 528–532
34. Vinogradov AE (2006) "Genome design" model: Evidence from conserved intronic sequence in human-mouse comparison. Genome Res 16: 347–354
35. Basehoar AD, Zanton SJ, Pugh BF (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell 116: 699–709
36. Tirosh I, Weintberer A, Carmi M, Barkai N (2006) A genetic signature of interspecies variations in gene expression. Nat Genet 38: 830–834
37. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogués G. (2004) Multiple links between transcription and splicing. RNA 10: 1489–1498
38. Sweetlove LJ, Fernie AR (2005) Regulation of metabolic networks: Understanding metabolic complexity in the systems biology era. New Phytologist 168: 9–24
39. Kuepfer L, Sauer U, Blank L (2006) Metabolic functions of duplicate genes in *Saccharamyces cerevisiae*. Genome Res 15: 1421–1430
40. Wagner A. (2000) Robustness against mutations in genetic networks of yeast. Nature 24: 355–361
41. Kafri R, Bar-Even A, Pilpel Y (2005) Transcription control reprogramming in genetic backup circuits. Nat Genet 37: 295–299
42. Papp B, Pal C, Hurst LD (2003) Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet 19: 417–422
43. Veitia RA (2005) Paralogs in polyploids: One for all and all for one? Plant Cell 17: 4–11
44. Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149–1152
45. Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, et al. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Proc Natl Acad Sci U S A 103: 6605–6610
46. Redman JC, Haas BJ, Tanimoto G, Town CD (2004) Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. Plant J 38: 545–561
47. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) Affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315
48. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: Open Software development for computational biology and bioinformatics. Genome Biol 5: R80
49. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: Class discovery by gene expression monitoring. Science 286: 531–537
50. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, et al. (2001) The *Arabidopsis* Information Resource (TAIR): A comprehensive database and Web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29: 101–105
51. Higo K, Ugawa Y, Iwamato M, Korenaga T (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res 27: 297–300
52. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, et al. (2003) AGRIS: *Arabidopsis* Gene Regulatory Information Server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factor. BMC Bioinformatics 4: 25
53. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57: 289–300
54. Gu Z, Cavalacanti A, Chen FC, Bouman P, Li WH (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode and yeast. Mol Biol Evol 19: 256–262
55. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66