RESEARCH ARTICLES

G-Boxes, Bigfoot Genes, and Environmental Response: Characterization of Intragenomic Conserved Noncoding Sequences in Arabidopsis

Michael Freeling, a Lakshmi Rapaka, a Eric Lyons, a Brent Pedersen, b and Brian C. Thomas b

a Department of Plant and Microbial Biology, University of California, Berkeley, California 94720
b College of Natural Resources, University of California, Berkeley, California 94720

A tetraploidy left Arabidopsis thaliana with 6358 pairs of homoeologs that, when aligned, generated 14,944 intragenomic conserved noncoding sequences (CNSs). Our previous work assembled these phylogenetic footprints into a database. We show that known transcription factor (TF) binding motifs, including the G-box, are overrepresented in these CNSs. A total of 254 genes spanning long lengths of CNS-rich chromosomes (Bigfoot) dominate this database. Therefore, we made subdatabases: one containing Bigfoot genes and the other containing genes with three to five CNSs (Smallfoot). Bigfoot genes are generally TFs that respond to signals, with their modal CNS positioned 3.1 kb 5’ from the ATG. Smallfoot genes encode components of signal transduction machinery, the cytoskeleton, or involve transcription. We queried each subdatabase with each possible 7-nucleotide sequence. Among hundreds of hits, most were purified from CNSs, and almost all of those significantly enriched in CNSs had no experimental history. The 7-mers in CNSs are not 5’- to 3’-oriented in Bigfoot genes but are often oriented in Smallfoot genes. CNSs with one G-box tend to have two G-boxes. CNSs were shared with the homoeolog only and with no other gene, suggesting that binding site turnover impedes detection. Bigfoot genes may function in adaptation to environmental change.

INTRODUCTION

Functional DNA sequence changes at a lower rate over evolutionary time than sequence without function. Exon sequence tends to be conserved, whereas functionless sequence is randomized by substitution, lost by conversion, or deleted entirely. Therefore, if two genes or chromosomal regions have diverged from a common ancestor, be they in different species (orthologs) or duplications within the same genome (paralogs), those few noncoding regions that retain a high degree of sequence similarity provide a measure of noncoding DNA function, where function is inferred from conservation (Hardison, 2000, 2003). In flowering plants, the alignment algorithm BLAST-2-sequences (Tatusova and Madden, 1999) has been used successfully to detect conserved noncoding sequences (CNSs) in comparisons of maize (Zea mays) and rice (Oryza sativa) orthologous genes (Kaplinsky et al., 2002; Inada et al., 2003) and to find intragenomic CNSs in Arabidopsis thaliana (Thomas et al., 2007) by comparing the syntenic duplicates (syntenic paralogs and homoeologs) retained following its most recent tetraploidy (called the α-event). The two genomes from the α-event (Simillion et al., 2002; Bowers et al., 2003; Maere et al., 2005) diverged approximately the same extent as have maize-rice and man-mouse (Kaplinsky et al., 2002); all of these have diverged sufficiently such that sequence conservations are not due to neutral carryover. Table 1 defines important terms we use: CNS, αCNS, gene space, and phylogenetic footprint. It should be understood that function does not assure that noncoding sequence will be conserved. Some sequences evolve quickly; others defy detection. CNSs are a subset of functional noncoding sequences.

Haberer et al. (2004) demonstrated that intragenomic (α) phylogenetic footprints in proximal promoters of Arabidopsis syntenic gene pairs could be found with an anchored alignment tool (Brudno et al., 2004), anchoring on the 5′ ATG. However, these workers did not explore sequence outside of the 500-bp proximal promoter. Guo and Moose (2003) used LAGAN (Brudno et al., 2003) with a Vista display (Mayor et al., 2000) to explore CNSs within the grasses, especially between maize and rice, concluding that transcription factor binding sites are within some CNSs and that CNSs were sometimes shared among different grasses.

This report explores the functions of 14,944 intragenomic CNSs (αCNSs) retained from a set of 3179 gene space pairs retained from the α-event in the Arabidopsis lineage (Thomas et al., 2006, 2007). These data can be accessed by a custom Web application called the Arabidopsis BIL2seq Viewer available at http://synteny.cnr.berkeley.edu/AtCNS. According to Thomas et al. (2007), these αCNSs exist in gene spaces with a modal frequency of zero and a mean of 1.7. The median αCNS length is 25 bp, and these sites are distributed in all regions of a gene with a general preference to be 5’ of a gene. The ratio of CNSs within the gene space, going from 5′ to 3′, 5′:3′:UTR:intron:3′:UTR:3’ is.

1 To whom correspondence should be addressed. E-mail freeling@nature.berkeley.edu; fax 510-642-4995. The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Michael Freeling (freeling@nature.berkeley.edu).

°Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.107.050419
encode or bind small RNAs. Thus, the likely general function of these CNSs are not simple sequences and not likely to either expression or phylogenetic relatedness (Prakash et al., 2004; Van Hellemont et al., 2005) without the addition of expert knowledge such as coexpression or phylogenetic relatedness (Prakash et al., 2004; Van Hellemont et al., 2005). Using phylogenetic relatedness, ~1400 enhancers are deeply (>300 million years) conserved in vertebrates (Wooife et al., 2005).

One cis-acting binding motif is known to be vastly overrepresented in the plant genome, a microsatellite, and thus was truly expected to be enriched in CNSs: GAGA. GAGA (CTCT) motifs are known to bind particular proteins, and these are known to interact with chromatin remodeling complexes to alter animal gene expression (Lehman, 2004; Meister et al., 2004; Kooiker et al., 2005). Simple repeated sequences are also known to mark regions of the human X chromosome that avoid silencing (McNeil et al., 2006). To not exclude GAGA or any similar regulatory sequence, we included simple sequences in our CNS list. In so doing, we accepted the increased noise level expected when vastly overrepresented sequences locate in α-syntenous gene space by chance alone. As it turned out, few αCNSs are simple sequence repeats (Thomas et al., 2007).

Genes retained after either a local or whole-genome duplication are a biased subset of ancestral gene content (Freeling and Thomas, 2006). Therefore, the αCNSs may be a biased sample of functional sites simply because the genes available for CNS analysis are themselves biased. In the Arabidopsis lineage, genes in the large molecular function GO categories “transcription factor activity” and “protein kinase,” along with other genes whose products interact in complexes, tend to be retained following tetraploidies at frequencies nearly double expectations (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005). So far, the combined data of gene retention following either local or whole-genome duplications in eukaryotes fits with the predictions of the gene balance hypothesis (Freeling and Thomas, 2006), but vertebrate tetraploid data are more difficult to acquire. The gene balance hypothesis (Veitia, 2002; Birchler et al., 2003; Papp et al., 2003; Birchler et al., 2005) predicts that
genes whose products function in subunit–subunit interactions or participate at the top of regulatory cascades will tend to be more susceptible to gene dosage change and therefore will be overretained following tetraploidy. For this reason, connected genes tend to be retained as pairs following tetraploidy. Since CNSs can only be detected in α-gene pairs, our CNS database is obviously biased toward connected genes.

RESULTS

Gene Retainability and CNS Richness Are Weakly or Not Correlated

Our CNS database is necessarily confined to the analysis of those 25% of genes that have α-pairs. To extrapolate from our data on functional noncoding sequence to the entire genome, we needed to access the relationship, if any, between retention post-tetraploidy and CNS richness. When genes are sorted by GO category and each was compared for both retention and CNS richness, no correlation was found (data not shown; see Supplemental Table 1 online). Because GO categories have overlapping gene contents, we attempted the same correlation among genes in different, nonoverlapping TF families (DATF families in Supplemental Table 1 online; Guo et al., 2005 at http://datf.cbi.pku.edu.cn/, July, 2005). There was no obvious correlation (graph not shown; see data in Supplemental Table 1 online). Had every gene in Arabidopsis been retained following the tetraploidy, and thus been available for analysis, we would have found approximately fourfold the number of α-CNSs.

GAGA Motifs and PLACE TF Binding Motifs in α-CNSs

GAGA

On the basis of published work (see Introduction), we expected that CNSs would preferentially contain GAGA repeats. We expected significantly more GAGA sequences in CNS sequence than in an equal amount of control noncoding sequence (see Methods); this we call “enrichment.” For GA6 with less than two mismatches (≤10/12), CNS enrichment was 0.8, meaning that there were actually more GA6-type sequences in control noncoding sequence than in CNS: GA6 = 0.6; GA5 = 0.9; GA4 = 1.16; and [GA] GA[GA][AG][GA][GA] (Kooiker et al., 2005) = 1.0. We expected a GAGA CNS enrichment proportion to be significantly >1. This expectation was not met.

PLACE cis-Acting Binding Motifs

Data on CNS function suggests that some CNSs might be clusters of known TF binding sites (see Introduction). The PLACE Angiosperm cis-acting sequence database (Higo et al., 1999) had accumulated 431 such TF binding motifs on August 1, 2005. Each motif and its reverse complement, in IUPAC (Prosite-type) format, were found in both α-CNS sequences and the control noncoding sequences within the same gene spaces. By requiring that at least 50 CNSs were hit by any motif reduced the list to 136 TF binding motifs. Enrichments ranged from 6.2-fold to 0.46-fold with a median at 0.9-fold. Fourteen motifs gave CNS enrichments more than twofold. Table 2 shows these data ranked descending for enrichment in CNSs; the top 14 and a few additional PLACE motifs are included here. The top-ranked eight motifs and 10 of the top 14 carry a G-box: CACGTG (Williams et al., 1992; Menkens et al., 1995; Gao et al., 2004). The specific PLACE sequence CACGTGGC, the most CNS-enriched PLACE motif (6.2-fold), is part of a Type I-AA_G-ABRE element, CCACGTGGC, known to operate in genes responsive to both abscisic acid (ABA) and external stress (Choi et al., 2000). TF binding sites for jasmonic acid response (pathogenic stress; Brown et al., 2003) and three additional known boxes (color-coded in Figure 1 and cited at PLACE) were enriched more than twofold in our α-CNSs. Boxes DRE-CRT and ARE (thought to confer response to auxin) are enriched to a lesser extent. We conclude that the most CNS-enriched motif, the G-box, lies in a noncoding gene space near particular gene pairs. The 268 CNSs with a perfect G-box are noted in the CNS list (see Supplemental Table 2 online). The G-box CNSs are usually positioned 5’ of the ATG (229 are 5’, eight are in intron, 29 are 3’, and one is near a gene we called ourselves, designated -oa) at a mean distance of 1588 bp and a range from 40 to 8958 bps upstream of the A of the 5’ ATG. Note that this distance is well 5’ of that considered to be the normal promoter region of a gene.

Bigfoot and Smallfoot Genes

The Problem Caused by a Few Hundred Very CNS-Rich Genes

Thomas et al. (2007) showed that some genes are particularly CNS rich, and these genes tend to fall into GO categories that are TF or MIR, each with a mean of approximately five CNSs/gene. The most CNS-enriched genes (10 to 20 CNSs/gene space) populate GO categories characterized by “response to...” plant hormones, internal or exogenous stress, light, desiccation, jasmonic acid, and so forth. This previous work engendered the hypothesis that CNS richness characterizes, in particular, genes that respond first to signals of all sorts: the first responder hypothesis. However, the average paired gene in Arabidopsis has 1.7 CNSs, and the modal gene has 0. These more CNS average genes tend to encode enzymes or structural proteins. The result is that a few CNS-rich genes dominate our CNS database with their many and potentially special CNSs. To address this problem, we created two subdatabases, one from the longest gene spaces in Arabidopsis (Bigfoot genes) and the other from genes with a modest number, three to five, of CNSs/gene (called Smallfoot genes). Table 3, row 1, quantifies the genes and sequences that define these two subdatabases, and Table 3 compares them. Twenty-four genes were shared.

Bigfoot Genes

Figure 2 shows a screenshot of a Bigfoot gene pair in the Arabidopsis BI2seq Viewer (Thomas et al., 2006, 2007). Note that it is transcribed from right to left. More often than not, the 126 pairs of Bigfoot genes exist in a region of chromosome devoid of other genes, with CNSs populating the void (Figure 1). Comparisons of vertebrate genomes identified gene deserts that are
Table 2. DATF Families and Counts in July, 2005, Showing Nonrandom Distribution of Both CNS Richness and Frequency of Genes That Are Bigfoot; Bigfoot Genes Are Possible among Retained Pairs Only

<table>
<thead>
<tr>
<th>TF Families</th>
<th>Total Genes</th>
<th>Genes Retained</th>
<th>Retention Frequency</th>
<th>Average No. CNS/Gene</th>
<th>Average Length</th>
<th>Fraction Bigfoot</th>
<th>Bigfoot Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF:LUG</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:NZZ</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:SAP</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:CCAAT-DR1</td>
<td>2</td>
<td>2</td>
<td>1.00</td>
<td>2.00</td>
<td>75.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:ULT</td>
<td>2</td>
<td>2</td>
<td>1.00</td>
<td>3.00</td>
<td>66.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:S1Fa-like</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:VOZ</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:GIF</td>
<td>3</td>
<td>2</td>
<td>0.67</td>
<td>0.00</td>
<td>0 of 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:HRT-like</td>
<td>3</td>
<td>2</td>
<td>0.67</td>
<td>0.00</td>
<td>0 of 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:MBF1</td>
<td>3</td>
<td>2</td>
<td>0.67</td>
<td>0.00</td>
<td>0 of 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:PBF-2-like (Whirly)</td>
<td>3</td>
<td>2</td>
<td>0.67</td>
<td>1.00</td>
<td>21.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:LFY</td>
<td>3</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:PcG</td>
<td>4</td>
<td>2</td>
<td>0.50</td>
<td>0.00</td>
<td>0 of 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:C2C2-YABBY</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:FHA</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:EIL</td>
<td>6</td>
<td>2</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:LI1</td>
<td>6</td>
<td>4</td>
<td>0.67</td>
<td>0.50</td>
<td>10.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:BES1</td>
<td>6</td>
<td>2</td>
<td>0.33</td>
<td>5.00</td>
<td>140.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:ALFIN</td>
<td>7</td>
<td>6</td>
<td>0.86</td>
<td>1.00</td>
<td>19.33</td>
<td>0 of 6</td>
<td></td>
</tr>
<tr>
<td>TF:CAMTA</td>
<td>7</td>
<td>6</td>
<td>0.86</td>
<td>1.00</td>
<td>32.67</td>
<td>0 of 6</td>
<td></td>
</tr>
<tr>
<td>TF:ARID</td>
<td>7</td>
<td>4</td>
<td>0.57</td>
<td>3.00</td>
<td>132.50</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:E2F/DP</td>
<td>8</td>
<td>2</td>
<td>0.25</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF:CPP</td>
<td>8</td>
<td>4</td>
<td>0.50</td>
<td>8.00</td>
<td>239.50</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:PLATZ</td>
<td>9</td>
<td>4</td>
<td>0.44</td>
<td>5.00</td>
<td>140.00</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:HMG</td>
<td>10</td>
<td>4</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:CCAAT-HAP2</td>
<td>10</td>
<td>6</td>
<td>0.60</td>
<td>1.67</td>
<td>38.67</td>
<td>2 of 6</td>
<td>0.33</td>
</tr>
<tr>
<td>TF:TAZ</td>
<td>10</td>
<td>6</td>
<td>0.60</td>
<td>3.33</td>
<td>117.67</td>
<td>0 of 6</td>
<td>0.00</td>
</tr>
<tr>
<td>TF:GRF</td>
<td>10</td>
<td>4</td>
<td>0.40</td>
<td>6.00</td>
<td>286.50</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:SRS</td>
<td>10</td>
<td>6</td>
<td>0.60</td>
<td>11.00</td>
<td>514.67</td>
<td>2 of 6</td>
<td>0.33</td>
</tr>
<tr>
<td>TF:TUB</td>
<td>11</td>
<td>2</td>
<td>0.18</td>
<td>2.00</td>
<td>37.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:CCAAT-HAP3</td>
<td>11</td>
<td>2</td>
<td>0.18</td>
<td>7.00</td>
<td>131.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:PHD</td>
<td>12</td>
<td>2</td>
<td>0.17</td>
<td>3.00</td>
<td>68.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:JUMONJI</td>
<td>13</td>
<td>2</td>
<td>0.15</td>
<td>1.00</td>
<td>28.00</td>
<td>0 of 2</td>
<td></td>
</tr>
<tr>
<td>TF:CCAAT-HAP5</td>
<td>13</td>
<td>4</td>
<td>0.31</td>
<td>2.00</td>
<td>73.00</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:GARP-ARR-B</td>
<td>13</td>
<td>6</td>
<td>0.46</td>
<td>2.67</td>
<td>100.33</td>
<td>2 of 6</td>
<td>0.33</td>
</tr>
<tr>
<td>TF:AB3/VP1</td>
<td>13</td>
<td>6</td>
<td>0.46</td>
<td>13.67</td>
<td>528.33</td>
<td>4 of 6</td>
<td>0.67</td>
</tr>
<tr>
<td>TF:Nin-like</td>
<td>14</td>
<td>8</td>
<td>0.57</td>
<td>3.00</td>
<td>82.25</td>
<td>0 f 8</td>
<td>0.00</td>
</tr>
<tr>
<td>TF:ZIM</td>
<td>15</td>
<td>10</td>
<td>0.67</td>
<td>1.40</td>
<td>28.00</td>
<td>0 of 10</td>
<td>0.00</td>
</tr>
<tr>
<td>TF:ZF-HD</td>
<td>15</td>
<td>4</td>
<td>0.27</td>
<td>13.50</td>
<td>450.00</td>
<td>4 of 4</td>
<td>1.00</td>
</tr>
<tr>
<td>TF:GeBP</td>
<td>16</td>
<td>4</td>
<td>0.25</td>
<td>0.00</td>
<td>0.00</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:SBP</td>
<td>17</td>
<td>6</td>
<td>0.35</td>
<td>2.67</td>
<td>101.67</td>
<td>2 of 6</td>
<td>0.33</td>
</tr>
<tr>
<td>TF:ARF</td>
<td>23</td>
<td>2</td>
<td>0.09</td>
<td>6.00</td>
<td>211.00</td>
<td>2 of 2</td>
<td>1.00</td>
</tr>
<tr>
<td>TF:HSF</td>
<td>24</td>
<td>4</td>
<td>0.17</td>
<td>2.50</td>
<td>79.50</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:TCP</td>
<td>24</td>
<td>8</td>
<td>0.33</td>
<td>5.13</td>
<td>179.00</td>
<td>2 of 8</td>
<td>0.25</td>
</tr>
<tr>
<td>TF:Trihelix</td>
<td>29</td>
<td>12</td>
<td>0.41</td>
<td>2.33</td>
<td>68.67</td>
<td>2 of 12</td>
<td>0.17</td>
</tr>
<tr>
<td>TF:C2C2-GATA</td>
<td>29</td>
<td>16</td>
<td>0.55</td>
<td>5.38</td>
<td>210.75</td>
<td>4 of 16</td>
<td>0.25</td>
</tr>
<tr>
<td>TF:AUX/IAA</td>
<td>29</td>
<td>16</td>
<td>0.55</td>
<td>5.63</td>
<td>211.75</td>
<td>2 of 16</td>
<td>0.13</td>
</tr>
<tr>
<td>TF:C2C2-co-like</td>
<td>31</td>
<td>16</td>
<td>0.52</td>
<td>3.00</td>
<td>118.25</td>
<td>0 of 16</td>
<td>0.00</td>
</tr>
<tr>
<td>TF:GRAS</td>
<td>33</td>
<td>15</td>
<td>0.45</td>
<td>3.27</td>
<td>85.67</td>
<td>6 of 15</td>
<td>0.40</td>
</tr>
<tr>
<td>TF:C3H</td>
<td>35</td>
<td>16</td>
<td>0.46</td>
<td>2.06</td>
<td>48.13</td>
<td>2 of 16</td>
<td>0.13</td>
</tr>
<tr>
<td>TF:C2C2-DOF</td>
<td>36</td>
<td>16</td>
<td>0.44</td>
<td>6.19</td>
<td>213.88</td>
<td>6 of 16</td>
<td>0.38</td>
</tr>
<tr>
<td>TF:B3</td>
<td>39</td>
<td>4</td>
<td>0.10</td>
<td>6.50</td>
<td>116.00</td>
<td>0 of 4</td>
<td></td>
</tr>
<tr>
<td>TF:AS2</td>
<td>42</td>
<td>8</td>
<td>0.19</td>
<td>4.25</td>
<td>118.75</td>
<td>2 of 8</td>
<td>0.25</td>
</tr>
<tr>
<td>TF:GARP-G2-like</td>
<td>43</td>
<td>16</td>
<td>0.37</td>
<td>7.00</td>
<td>239.13</td>
<td>2 of 16</td>
<td>0.13</td>
</tr>
<tr>
<td>TF:WRKY</td>
<td>73</td>
<td>24</td>
<td>0.33</td>
<td>3.75</td>
<td>109.00</td>
<td>4 of 24</td>
<td>0.17</td>
</tr>
<tr>
<td>TF:zZIP</td>
<td>75</td>
<td>30</td>
<td>0.40</td>
<td>3.87</td>
<td>122.40</td>
<td>2 of 16</td>
<td>0.13</td>
</tr>
</tbody>
</table>

(Continued)
adjacent to TF and developmental genes and are full of deeply conserved noncoding sequences (Ovcharenko et al., 2005); Bigfoot gene space seems to be a plant convergence on the gene desert phenomenon. The 252 Bigfoot genes are noted in Supplemental Table 1 online. Using GOStat (see Methods), there were 140 to 142 genes each in four overrepresented (P = 0.00) GO categories: 0005488, binding; 0003677, DNA binding; 0003700, TF activity; and 0003676, nucleic acid binding. Other significantly overrepresented GO categories (P < 3 \times 10^{-8}) were several with the phrase “regulation of...” and GO:0008755 hormone-mediated signaling. Bigfoot genes are predominantly (~66%) TF genes. Other very broad categories of Bigfoot gene pairs are 16 unknown protein pairs, six protein kinase pairs, 18 enzyme pairs, one RNA binding pair and, interestingly, two DVL gene pairs (A12N061 and A08N202). DVL genes (genes that were not in The Arabidopsis Information Resource [TAIR]; May, 2005), are a group of genes encoding small polypeptides that, when overexpressed, confer developmental phenotypes (Wen et al., 2004). The GOstat results for Bigfoot genes (Table 4) expand upon the “first responder” list of GO terms that are most CNS rich of all \alpha-pairs in Arabidopsis (Thomas et al., 2007). It was not a surprise that particularly long gene spaces (Bigfoot) tended to be CNS rich.

We investigated whether or not Bigfoot genes occur at random in the phylogenetic trees of gene families. To answer this question, we looked more closely at the 56 TF families recognized by DATF (Table 2). Twenty-four of 44 \alpha-paired homeobox (HB) genes are Bigfoot (55%), but 0 of 20 MADS box \alpha-paired genes are Bigfoot. The range of percentage of Bigfoot among all TF families is greater yet (even though some families have so few
Table 3. Comparison of Features of Bigfoot and Smallfoot Genes, Where Bigfoot Genes Have >4 bp of Syntenous, CNS-Rich Space 3' + 5' of the CDS, Smallfoot Genes Have Three to Five CNSs, and 24 Genes Are Both Bigfoot and Smallfoot

<table>
<thead>
<tr>
<th>Feature</th>
<th>Bigfoot Genes</th>
<th>Smallfoot Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes (CNSs; total bp CNS; total bp control noncoding)</td>
<td>252 genes (2897; 459,548 bp; 26,141,400 bp)</td>
<td>1197 genes (7532; 252,934 bp; 5,426,926 bp)</td>
</tr>
<tr>
<td>GO categories</td>
<td>Narrow spectrum, 78% TFs; many respond to environmental or hormonal signals (Table 4)</td>
<td>Broad spectrum. Signal transduction, then transcription, metabolism, and cytoskeleton (Table 5)</td>
</tr>
<tr>
<td>Does the direction of transcription influence enrichment/purification of a random 7-mer?</td>
<td>None significantly preferred complement versus reverse complement, and only one pair had a nominal P &lt; 0.001 (see Supplemental Table 3 online)</td>
<td>Many 7-mers were polar. Of the eight significantly enriched, two were significantly polar and four were polar at P nominal &lt;0.001. The next most significantly enriched were 13% polar (see Supplemental Table 4 online)</td>
</tr>
<tr>
<td>How important is purifying selection in determining the overall sequence of gene space?</td>
<td>Very. The difference between slope 0.45 and the expected 1.0 of the hits CNS versus hits control of the trend line of Figure 3A implies massive purification of CNSs.</td>
<td>Very. The difference between the slope 0.51 and 1.0 on the similar plot as discussed for Bigfoot genes attests to importance of purification (data not shown).</td>
</tr>
<tr>
<td>Which 7-mer sequences tend to be purified from CNSs (removed)?</td>
<td>Most purified are long runs of A or T, and purified are almost any sequence with more than three As or Ts in a row. In general, [A+T] is purified. More research is needed.</td>
<td>Similar to Bigfoot genes. A general definition of CNS should include sequences purified and sequences under positive selection.</td>
</tr>
<tr>
<td>Which 7-mers are enriched? Are they TF binding sites?</td>
<td>The most significant carry CACGTG (G-box binds bZIP or HLH TFs), plus a few known TF binding sites, but most are unknown (see Supplemental Table 3 online).</td>
<td>One of the eight significant 7-mers could bind a MYC TF (see Supplemental Table 4 online). As with Bigfoot genes, significant and “worth interest” 7-mers are mostly unknown (e.g., 5’-CTTCTTC).</td>
</tr>
</tbody>
</table>

α-paired genes that they cannot be studied at this time). Breaking the 24 HB Bigfoot genes into subfamilies, the HD-ZIPI and HD-ZIPIII subfamilies are almost all Bigfoot (90%), while those genes in the sister HD-ZIPIII and -IV subfamilies are more like the average TF. The Bigfoot notation measures a CNS property that is positively correlated with CNS richness but is a unique feature of gene space. As with CNS richness, Bigfoot genes characterize some gene lineages but not others.

**Smallfoot Genes**

We then subjected a list of 1197 Smallfoot genes to the same analyses described for Bigfoot genes. Table 5 shows the GOStat results: 45 terms were enriched at P < 0.001, and 20 of them involve signal transduction; the six highest ranked terms identify protein kinase genes, including GO:004713, protein Tyr kinase activity (rank 1, P = 9E-12), and GO:004674. GO:004544, regulation of transcription (rank 8, P = 1.9E-9), and GO:003700, TF activity (rank 24, P = E-7), are also represented. Lower-ranked terms include cytoskeleton and RNA helicase. Note that no term involving signal transduction was significantly enriched in Bigfoot genes (Table 4), so Bigfoot and Smallfoot genes function quite differently.

**Enriched and Purified Motifs in Bigfoot αCNSs**

**Bigfoot Genes**

As presented previously, most Bigfoot genes encode TFs, most are CNS rich, and all use an exceptional amount of chromosomes for conserved function (Figure 2, Table 2). Even so, Bigfoot CNSs comprise only 4.3% of the average Bigfoot gene space. The Bigfoot subdatabase contains 2538 αCNS sequences (see Supplemental Table 2 online) whose total sequence lengths sum to 91,970 bps of nonoverlapping, aligned (5’-3’) αCNS sequence. These CNSs are distributed around Bigfoot genes with a greater overall 5’:3’ bias (3.1:1) than the similar statistic for the average gene (1.8:1). The frequency and distribution of these CNSs will, of course, influence the frequency and distribution of motifs located within CNSs; that’s why these CNS distribution data are reviewed here.

**Random 7-bp Motif Enrichment and Purification**

Supplemental Table 3 online lists all 16,384 (47) possible 7-bp DNA sequence motifs and data on their properties, hit numbers, and hit locations in the aligned Bigfoot subdatabase. Data on each motif in the database include number and frequency of nonoverlapping hits to αCNSs (for both complement and reverse complement), hits to gene space control sequences, the ratio of αCNS hits to control hits (called “fold enrichment” if αCNSs are preferred and “fold purification” if noncoding nonCNS gene space is preferred), and a χ² evaluation of significance (see Methods) for those 7-mers with hits totaling 10 or more; nominal P values are listed and evaluated in a special column based on Bonferroni corrections for multiple tests (see Methods). Figure 3A plots the number of occurrences of each of the 16,384 motifs plus its reverse complement in αCNSs (y axis) versus an equal amount of gene space control sequences (x axis; see Methods).
Given normal distributions, and no difference between αCNS sequence and nearby noncoding sequence, a trend line in Figure 3A should emerge with a slope of 1.0; this slope, the null hypothesis, reflects a 1:1 motif hit ratio between CNS and control noncoding sequence. Unexpectedly, the slope of the trend line in Figure 3A is 0.46. Many of the 7-bp motifs that are most overrepresented in noncoding space (enclosed in the largest oval in Figure 3A) have been purified from αCNS sequences, so much so that the slope of the trend line is far below 1.0. Those motifs most overrepresented in noncoding sequence in general (high A+T sequence and especially in runs; Figures 3B and 3C) tend not to be in CNSs. There are 150 significantly purified 7-mers and 282 additional 7-mers that have low enough P values to be worth interest (see Methods). Supplemental Tables 3 and 4 online have these 7-mer rows color-coded in pink and blue, respectively.

We found 14 motifs, seven pairs, that are significantly enriched (overrepresented, the opposite of purified) in Bigfoot αCNSs; those furthest from and above the trend line are circled in Figure 3. Most of these motifs have a core ACGT. The additional significantly enriched 7-mer motifs were experimentally unprecedented: a CATG core and a CCAC/GTGG core. The fold enrichment for these 14 most significant 7-bp sequences varies from 3.21 (TCACATG and reverse complement: 61 hits) to 9.89 (CACGTGC and reverse complement: 40 hits). Among all 8-mers with hits in Bigfoot gene αCNSs >9, GCACGTGC was most enriched at 23-fold (n = 10). This maximum-enriched 8-mer is clearly related to the most enriched PLACE motif (stress and ABA responsiveness) discussed previously: CACGTGGC. Among all 8 bp, this PLACE motif also ranks among the most enriched in the Bigfoot αCNSs at 9.7-fold (n = 22).

There are 28 7-mers, those that hit Bigfoot CNSs in the “worth interest” category of significance; these usually share core motifs CATG, ACGT (G-box core), or CCAC/GTGG that are known to be associated with significantly enriched 7-mers and sometimes group into additional orderly patterns.

GAGA motifs are not significantly enriched. The GAGAGAG/AGAGAGA 7-mers (and reverse complements) were 2.1-fold enriched with an insignificant P value, one below our “worth interest” category. The most significantly enriched GAGA-like sequence is GGAGAAG and its reverse complement at 2.75-fold enrichment (nominal P = 0.003), with 30 hits in the Bigfoot αCNS; this 7-mer is also not “worth interest.”

Sequences That Are Missing (Purified) from the αCNSs of Bigfoot Genes

We use the term “purified” to be the opposite of “enriched” to connote the selective process that presumably removed unfit
sequence from functional regions of gene space. A total of 150 7-bp sequences are significantly purified from BF CNSs, and an additional 282 7-mers are purified at “worth interest” levels of significance. Many of the significantly purified 7-mers are highly abundant in noncoding space, which accounts for the 0.46 slope of the trend line in Figure 3A. Figures 3B and 3C plot all purified 7-mers that hit a Bigfoot CNS at least 10 times in the order of their purification significance, with the most significant on the left (arrows in Figure 3C at P nominal < 0.00001, P nominal < 0.001, and P nominal < 0.05). The x axes are percentage of GC of the 7-mer (Figure 3B) and 7-mers with runs of 4 nucleotides (Figure 3C). It can be seen from Figures 3B and 3C that significant purification involves runs of A and T and generally high percentage of AT, the very sequences that characterize the bulk of noncoding space.

### Distribution of Significantly Enriched or Purified 7-bp Sequences within Bigfoot Gene Space and Strand Preference

The 5′:3′ ratio of the 2538 CNSs around Bigfoot genes, the genes used for this 7-mer analysis, is 3:1. The 5′:3′ ratio is skewed to the right, with the enrichment in the 3′ direction. This indicates that the enrichment is not random and is likely related to the structure of the genome or the function of the genes. The enrichment of 7-mers in the 3′ direction is consistent with the enrichment of 6-mers in the 3′ direction observed in previous studies. This suggests that the enrichment of 7-mers is not random and is likely related to the structure of the genome or the function of the genes. This enrichment could be due to the fact that 7-mers are more likely to occur in the 3′ direction due to the way the genome is structured.
significantly enriched 7-mers is 3.15 kb upstream of the 5′ATG; the mean 3′ position is 1.9 kb downstream from the stop codon upstream. Data for the 28 “worth interest” 7-mers are similar. These average locations within Bigfoot gene space are well outside the transcription unit and outside the 0 to −500-bp proximal promoter that is generally studied experimentally.

None of the 7-mers to Bigfoot CNSs showed preference for orientation with regard to the direction of transcription; there were neither significant nor “worth interest” motifs (see Supplemental Table 3 online). This is in contrast with similar data for Smallfoot genes, as will be shown.

**Smallfoot Gene Data**

We prepared an aligned subdatabase of αCNSs with genes labeled as Smallfoot genes as described and subjected this gene list to the same analyses as was the Bigfoot gene list (Supplemental Table 5 online is the 7-mer to Smallfoot genes data sheet).
Tables 4 and 5 show that the gene GO annotations of Bigfoot and Smallfoot genes are very different. In general, the 7-mers enriched were very different as well, as summarized in Table 3 and Supplemental Table 4 online. Those eight sequences (complement only) enriched significantly and 53 sequences enriched at a "worth interest" significance level were, with one possible exception (a MYC gene binding site; see Supplemental Table 4 online), without experimental precedent and not the same as were hit in the Bigfoot database. The most significantly enriched 7-mer is 5\textsuperscript{9} CTTCCTTC, and there are several more reasonably similar sequences ranking among the 61 most significant. There are several sequences that are 5\textsuperscript{9} CACG-like, all of unknown function. In general, 7-mers that significantly hit Smallfoot CNSs are approximately five times more numerous but about half as enriched as are the comparable Bigfoot 7-mers (Table 3). If we had twice as many 7-mer hits, as would be the case with a twice-as-large database, there would be a manifold increase in the number of significantly enriched 7-mers.
For 7-mers to Bigfoot genes, it didn’t matter whether the complement or reverse complement was used. Smallfoot gene CNSs hit significantly by 7-mers are often biased (Table 3; see Supplemental Tables 3 and 4 online) to one strand or the other. This directionality fits the compact nature of Smallfoot genes and tends to validate the 7-mer motif’s functionality in relation to transcription or translation of Smallfoot genes. Twenty-five percent of significantly enriched and 5% of “worth interest”-enriched 7-mers are biased significantly (Bonferroni-corrected; see Methods). These percentages rise to 50 and 13%, respectively, if strand bias is judged at the \( P \) nominal < 0.001 level. Even purified sequences to Smallfoot genes show strand bias at approximately one-third the values of enriched 7-mers. Under each of these conditions, every 7-mer to Bigfoot gene CNS showed zero strand bias; Bigfoot CNSs provided an exceptionally useful control.

**G-Boxes in CNSs Are Not Significantly Associated with Conserved Neighbor 7-Mers or Motifs, Except Other G-Boxes**

The G-box, CACGTTG, has been studied in detail for two decades. Some of these studies have found associated motifs that, with the G-box, comprise a cis-acting array. However, the G-boxes have been in core promoters (not as far upstream as our CNS G-boxes). Studies on ribulose-1-5-bisphosphate carboxylase small subunit genes, for example, have discovered a conserved modular array consisting of a G-box and an adjacent I-box (Arguello-Astorga and Herrera-Estralla, 1998), and a GCC motif (jasmonic acid, GGGCCG) was found near a G-box in *pmt1a* gene of tobacco (*Nicotiana tabacum*; Xu and Timko, 2004), and both were necessary for jasmonic acid induction. Weaker associations between G-boxes and adjacent sequences have been reported (Giuliano et al., 1988). The *Arabidopsis* G-box and an I-box–like motif (ATATCCGA) were associated with photosynthesis GO terms in *Arabidopsis* (Vanderpoole et al., 2006). None of these G-box–linked motifs or any PLACE motif served near G-boxes; no motifs were found other than the G-box itself. We compiled a table of all CNSs and all significantly enriched Bigfoot 7-mers and looked for any co-occurrence of 7-mer pairs to CNSs; we found nothing significant.

**Evaluation of Shared CNSs**

Some sorts of negative data are important. With trivial exceptions, there were zero *Arabidopsis* \( \alpha \)-CNSs, not even our most significant CNSs, detectable by bi2seq in other *Arabidopsis* gene space except the duplicate from the \( \alpha \)-tetraploidy. Supplemental Table 5 online contains these data. We examined 23 Bigfoot gene pairs with an out-group gene from the same lineage, a gene categorized by Bowers et al. (2003) to have diverged from \( \alpha \)-duplicate genes as part of the \( \beta \)-tetraploidy. (The \( \beta \)-tetraploidy, more ancient than the \( \alpha \)-tetraploidy, occurred after the monocot-dicot split [Bowers et al., 2003], but it remains unclear whether or not the \( \beta \)-tetraploidy is in poplar [Tuskan et al., 2006].) Not one CNS was conserved between either \( \alpha \)-gene and any \( \beta \)-out-group gene. These three-gene comparisons were at the tips of each of the 23 gene trees. The \( \beta \)-tetraploidy event has been estimated to have happened \( \sim \)100 million years ago (Bowers et al., 2003; Maere et al., 2005). Therefore, unlike the case in vertebrates, where some enhancer CNSs are conserved for >350 million years (see Introduction), the maximum level of CNS conservation in higher plants is relatively shallow. That does not mean that the function specified by these CNSs is not conserved; it means that we cannot detect such conservation at the sequence level.

**DISCUSSION**

**Limitations of Intragenomic Footprints**

We can only analyze 25% of the gene content of *Arabidopsis*, which is the gene content that was retained in pairs following \( \alpha \)-tetraploidy. We found that those genes retained, and studied, are not expected to have more or fewer CNSs than the remaining 75% of the minimized genome would have had they all retained post-tetraploidy or if there were a brassicoid genome diverged to the same extent as are \( \alpha \)-pairs. To our knowledge, no *Arabidopsis* relative useful for pairwise (CNS) research has been sequenced to date.

Without an out-group branching off the *Arabidopsis* lineage before the \( \alpha \)-tetraploidy, we cannot measure subfunctionalized or neofunctionalized (gain-of-function) sequences, although post-tetraploidy duplicate genes certainly diverge (Gu et al., 2002a, 2002b; Makova and Li, 2003; Raes and Van de Peer, 2003; Gu et al., 2004; Haberer et al., 2004; Li et al., 2005; Rastogi and Liberles, 2005; Roth et al., 2006). However, sometimes divergence is slower than expected (Koonin, 2005; Chapman et al., 2006), or divergence may vary by gene functional category (Ha et al., 2007). What we measure are sequences shared by homoeologs for the entirety of the time since tetraploidy. These 7470 noncoding sequence pairs in our \( \alpha \)-CNS database are likely to specify the same function today as they provided in the ancestral gene before tetraploidy. Cotton (*Gossypium hirsutum*), a candidate out-group, is simply too ancient to facilitate *Arabidopsis*–cotton CNS discovery (data not shown), and poplar is even more distant. Papaya (*Carica papaya*; in the family Caricaceae, in the same order as *Arabidopsis*) could constitute a useful out-group when the sequence is finished (Lai et al., 2006). Lack of a properly positioned out-group or usefully diverged genomic
sequences fundamentally limits this and all current work using the \( \alpha \)-tetraploidy of *Arabidopsis*.

### Results That Validate the Functionality of CNSs

Table 6 summarizes our major results. Three results imply CNS functionality independent of evolutionary conservation. (1) CNS richness varies with gene GO category (Thomas et al., 2007). Genes annotated GO “respond to...” are most CNS-rich, then TF genes in general, then signal transduction genes, then metabolic genes, and then, with zero CNSs, genes encoding ribosomal subunits and components of the mitochondria. Our GO term analyses of Bigfoot (Table 4) and Smallfoot (Table 5) genes substantiates this result. (2) The most significantly CNS-enriched motifs in all CNSs and in the Bigfoot CNS list are G-box motifs (Figures 2 and 3A). G-boxes are proven to bind known families of TFs (Menkens et al., 1995; Toledo-Ortiz et al., 2003), lending credibility to our database. (3) The significantly and worth interest CNS-enriched motifs in the Smallfoot but not (0%) Bigfoot CNSs are often aligned in relation to the direction of transcription. Twenty-five percent of the significantly enriched and 6% of those worth interest are polar, and these values increase to 50 and 13%, respectively, if we reduce the significance level to \( P < 0.001 \). Bigfoot 7-mers show zero polarity. The CNS enrichment significance of complement versus reverse complement is given in Supplemental Tables 3 and 4 online. CNS polarity, based on the direction of transcription, implies a function physically linked to the transcription process, which fits Smallfoot gene spaces with their CNSs being close to or within the transcriptional units. We could think of no alternative to function that could account for these three results, but experimental verification of CNS function remains outstanding. Presumably, this function is to bind or affect the movement (e.g., boundary elements) of regulatory factors. Small RNAs binding gene transcripts are unlikely candidates for binding CNS sequences (Thomas et al., 2007), although such binding is possible for 1.3% of the CNSs. We are left with the hypothesis that CNSs generally bind proteins or other macromolecules. CNSs of the sort that are aligned near Smallfoot gene exons behave like typical TF binding sites. On the other hand, CNSs far upstream of Bigfoot exons, even though they may carry a G-box, are more like enhancers.

### Many CNSs May Be Explained by Selection for or against Particular Sequences, but Almost All of These Sequence Motifs Were Unknown Previously

As seen from the slope of the regression line in Figure 3A (0.46 and not the expected 1.0) and in the description of the similar line comparing Bigfoot and Smallfoot genes in Table 3 (0.51 and not the expected 1.0), it is clear that CNS regions of noncoding gene space have been significantly purified of many particular 7-mer sequences. CNSs are defined largely by what sequences they do not contain. Apparently, when CNSs are interrupted by many sorts of sequences, function is impeded. Thus, a CNS must tend to be a module of function rather than a chance closeness of independent binding sequences. In other words, a CNS generally does not have spacer DNA that could be any sequence. We know that runs of five or six of any nucleotide and runs of four of A or T are particularly removed from CNS space. AAAAAGG in Smallfoot CNSs is purified at 0.13 in CNS versus in noncoding control space (\( P = 0 \)).

With exceptions, Bigfoot and Smallfoot CNSs have had very similar sequences removed but are characterized by very different enriched 7-mers (see Supplemental Tables 3 and 4 online). Almost all of our significant or worth interest 7-mers have no experimental history. For example, there are nine 7-mers of the 50 most enriched in Smallfoot genes marked “High GC” (e.g., 5\(^‘\)-CGTGCC, 5\(^‘\)-CGGCCGC, or 5\(^‘\)-GAGCCGT of Supplemental Table 4 online). The core CGGCTCC may be similar to a few of these; this is called Box A (Logeman et al., 1995) in the PLACE database (Higo et al., 1999). Our results indicate that the linear sequence of conserved noncoding DNA sometimes explains function. We need much more CNS data (10-fold) in plants if we are to generate enough hits to fully evaluate linear DNA sequence as an explanation for CNS function. The sorghum sequence as a rice comparison should be perfectly suited to this aim (Inada et al., 2003). It remains possible that a conserved CNS sequence is but one of several or many that maintains identical functional conformations.

### Table 6. Four Major Conclusions as Clues to CNS Functions

| 1. | Some CNS-enriched motifs, like the G-box and jasmonic acid (jerr or GCC box in Figure 1 and MYCATERD1 box in Figure 3A), have a rich experimental history, but the vast majority of CNS-enriched motifs have yet to be studied. The G-box in very 5′ CNSs may be functionally distinct from the well-described G-box in proximal promoters of light-sensitive genes. Adding the Bigfoot and Smallfoot genes together, CNSs are significantly enriched in 22 7-mer sequences, and another 81 7-mers are enriched (nominal \( P < 0.001 \)) to an extent “worth interest.” Similar counts of sequences, function is impeded. Thus, a CNS must tend to be a module of function rather than a chance closeness of independent binding sequences. In other words, a CNS generally does not have spacer DNA that could be any sequence. We know that runs of five or six of any nucleotide and runs of four of A or T are particularly removed from CNS space. AAAAAGG in Smallfoot CNSs is purified at 0.13 in CNS versus in noncoding control space (\( P = 0 \)). |
| 2. | Those genes with the most CNSs tend to be “first responder” genes, these being TF genes 64% of the time (Thomas et al., 2007), induced by environment or hormonal signals (Table 4). Those genes that take up an extraordinary length of chromosome, Bigfoot genes, tend to be CNS rich and are especially represented among “first responder” genes, these being TF genes of particular subfamilies or classes (i.e., HD-ZIPI). Genes with two to five CNSs tend to be involved in signal transduction and not transcription (Table 5). The modal gene is not regulatory and has 0 CNSs (Thomas et al., 2007).
| 3. | Unlike the situation in vertebrates, we found no evidence for highly conserved (100 to 400 million years) noncoding sequences within the gene trees of *Arabidopsis*. Even so, the most conserved sequences in plants seem to resemble the most conserved noncoding sequences in animals in terms of regulatory/developmental function and location in gene voids.
| 4. | There is no one functional explanation for CNS richness. While Bigfoot genes, G-box CNS genes, “first responder” genes, and TF genes all tend to be CNS rich, analyses of the detailed gene contents of these categories finds more uniqueness than overlap. |
**CNS Function in Arabidopsis**

### Genes with Zero CNSs

It is curious why the modal gene carries zero CNSs and is also expected to contain many TF binding sites. We think there is an explanation: the expected TF binding sites are either conserved as motifs but not as recognizable linear sequence or that there has been enough binding site turnover among redundant sequences in a cluster to obscure the conserved function, as has been demonstrated in *Drosophila* (Ludwig et al., 1998, 2005; Moses et al., 2004, 2006) and mammals (Frith et al., 2006). When *Arabidopsis* α-gene pairs are anchored on the 5′-A of their ATG and compared using a modified local alignment program, such as DIALIGN (Brudno et al., 2003; used by Haberer et al., 2004), convincing footprints were detected in the 500-bp upstream region. These less significant intragenomic footprints include potentially the TF binding sites that are missing in our modal pair of homoeologs. Motif-finding algorithms coupled with phylogenetic footprinting between poplar and *Arabidopsis* have been successful in finding convincing arrays of TF binding sites in the promoters of dicot MADS box TF genes (De Bodt et al., 2006). Additionally, Vandepoele et al. (2006) found motifs in *Arabidopsis* genes using three input data sets: genome sequence of *Arabidopsis*, genome-wide expression data for *Arabidopsis*, and the complete genome of the related out-group dicot, poplar. Confining themselves to 500 to 1000 bp upstream of exon 1, these workers used alignments to find *Arabidopsis*-poplar orthologs and used a nonalignment-based MotifSampler (Thijs et al., 2001, 2002) to find overrepresented, conserved (functional) short proximal promoter cis-acting motifs among coregulated genes. There is adequate evidence that a TF binding site in the proximal promoters of *Arabidopsis* genes can sometimes be identified if the correct tools are used. Presumably, future research, perhaps using binding site reconstruction tools such as MONKEY (Moses et al., 2004, 2006), will find that CNSs, and especially the larger CNSs, contain more than one binding motif, or that anchoring on CNSs might permit finding conserved motifs adjacent to CNSs and far from exons. The fact remains that we did not find any motif associations within CNSs with the exception that G-boxes tend to occur multiple times in CNSs.

We suggest that plant researchers interested in cis-regulatory modules include all noncoding gene space as we define it (Table 1) and not just a bit of 5′ sequence.

### CNS Function and Gene Expression

The most significant short sequence enriched in CNSs is the G-box; this is a clue as to the role(s) of CNSs in gene expression. However, because the historically documented G-box is the most significantly CNS-enriched TF binding motif does not mean that the G-box–containing CNSs should exhibit historical properties. It is possible that the G-box has many functions and those CNS-associated have not been studied. The placement of the average G-box is in a CNS 1.5 kb upstream of exon 1, and the GO annotation of the G-box–containing gene spaces did not include response to light signals. This far 5′ location is not far by Bigfoot gene standards with their mean CNS 3.1 kb 5′. Perhaps the G-boxes within CNSs are as yet unstudied enhancers. On the other hand, research on G-box–like motifs has linked them to several different functions. The ACGT core embedded in a number of related sequences is known to bind TGA-type bZIP TFs (Izawa et al., 1993), although this core is not always necessary for binding (de Pater et al., 1994). The phenotypic consequences of such binding can involve both positive and negative regulation of transcription in response to developmental cues, ABA, auxin, pathogenic stress, salicylic acid, and ethylene (Tucker et al., 2002). However, the GOStat results on our genes with G-box CNSs did not yield any significant term in the “response to...” gene except “response to ethylene.” Further research on these distant CNS G-boxes and their genes is needed.

All three of the original studies of CNSs in plants (Kaplinsky et al., 2002; Guo and Moose, 2003; Inada et al., 2003) noted that, compared with man-mouse, there were far fewer CNSs in plants and they were far shorter, so much so that the definition used by mammalian workers to define a CNS would find no CNSs in plants at all. This observation was predicated on the comparability of data from man-mouse and maize-rice, both with exons diverged to ~85% nucleotide sequence identity over long high-scoring pairs from BLAST-2-sequences. Our work on αCNSs in *Arabidopsis* is consistent with the observation that plants have vastly fewer and shorter CNSs than mammals. Most compelling is the difference in distribution of CNSs along the grass or *Arabidopsis* chromosome versus the man-mouse chromosome. In plants, CNSs are almost always clustered near exons so that sorting them to gene was feasible; mammalian intergenic regions are often covered with CNSs, and raising bitseq mismatch penalties does not help. Sorting them to gene without experimental data would generate ambiguities. Inada et al. (2003) suggested that CNSs largely keep stem cell/meristem genes from being ectopically expressed, where they might cause disease/cancer, and that many more genes use this regulatory mode in mammals compared with plants due to several differences in biology. This hypothesis was based on one case study and some arguments. Our results support this useful but untested hypothesis.

### CNS Richness, Bigfoot Genes, and Evolvability

When faced with stress, such as must occur with environmental change, animals move and plants endure on penalty of extinction. Major trends in animal form and development might be seen primarily as adaptations involving organismal and cellular movement, with trends involving synapses becoming paramount over evolutionary time in the metazoan lineage. Perhaps the comparable major (most complex, perhaps) trends during higher plant evolution might be in those gene lineages that, when expressed as functional modules, modify abilities to endure. Over the last 500 million years of plant evolution, the maximums of adaptation to environmental extremes have probably gone up and not down over time; maximums of plant morphological complexity have certainly gone up (Freeling and Thomas, 2006). Table 4 details the many “response to...” GO categories that are significantly overrepresented in the Bigfoot database. We suggest that Bigfoot genes and other CNS-rich genes include those special endurance genes, genes we have lumped together in the general category “first responder” to environmental signals. As an illustration, the CNS-rich TF family HB, the even more CNS-rich...
subfamily, HD-ZIP (homeodomain-basic leucine zipper), and especially the Bigfoot HD-ZIPI class are known to be induced by environmental stress (Henriksson et al., 2005). The relationship between CNSs and signal response invites further research. Recent research from the Z.J. Chen laboratory (Ha et al., 2007) found that α-duplicates that were induced by environmental stress had more divergent expression patterns than α-duplicates expressed programmatically in the course of development.

Most Arabidopsis genes are positioned close to one another on the chromosome and have zero αCNSs. Bigfoot genes are not only exceptional in size, but, as first responders, function high in the regulatory cascade. If CNS richness implies a high level of gene regulation, then Bigfoot genes are, at the same time, high-level regulatory genes and also the most highly regulated (Inada et al., 2003; Thomas et al., 2007). This enigma suggests a systems sort of regulatory model. During the course of higher plant evolution, mutations (subfunctionalizations and neofunctionalizations) in particular genes, important genes, may help explain adaptations to changing environments. Bigfoot genes present exceptionally large targets for regulatory mutation. Since Bigfoot genes are exceptionally enriched in “response to...” genes (Table 4), the report that “response to...” homoeologs may diverge in expression particularly rapidly (Ha et al., 2007) supports these conclusions. All else being equal, it seems wise to look at alleles, derivatives and cooptions involving these large, CNS-rich, first responder genes to better understand the genetic basis of plant endurance.

METHODS

α-Pairs, Gene Spaces, and the αCNS Database

We used the Arabidopsis thaliana intragenic CNS database as by Thomas et al. (2007) and its supplemental data. When pairs, gene spaces, or CNSs gain information due to this study, that information is added to our data tables as Supplemental Tables 1 and 2 online.

Gene Categories and Statistics

Genes were categorized by GO term (per Thomas et al., 2007), as MIR genes, and those genes that were also TFs were subcategorized into DATF’s 56 families (Database of Arabidopsis Transcription Factors, July, 2005; http://datf.cbi.pku.edu.cn/) (Guo et al., 2005). Except for MIR genes, genes encoding RNA were not counted in this study. When a gene in a category list had an uncalled, unannotated, or vaguely annotated homoeologous partner, we did not add the poorly annotated gene to the GO annotation lists. However, we did update the DATF lists, thereby increasing the number of genes above that on the original list (documented in Supplemental Table 1 online, columns B and C). Genes we called ourselves in the process of validating α-gene pairs were not used when GO terms were needed. Our analysis did not find new MIR genes except as additional duplicates in known gene spaces; MIR genes were from “RNA Families Database of Alignments and CMs” or Rfam (Thomas et al., 2007). Gene lists derived from the annotation keyword Bigfoot and Smallfoot and gene lists from motif enrichment analyses, like "contains a G-box CNS" (lists from Supplemental Table 1 online) were evaluated for GO category representation using GOstat (http://gostat.wehi.edu.au) (Beissbarth and Speed, 2004), using TAIR GO annotations, with a P value cutoff ≤0.001.

Analysis of CNS Sequence for Over- and Underrepresentation of Particular DNA cis-Acting Motifs or 7-Mer Sequences

We used the χ² test to evaluate the significance of a PLACE (cis-acting binding sequence) motif or random 7-mer hit to CNSs versus hits to control noncoding gene space. We limited data to having 10 hits minimum to facilitate this test. The results are the nominal P values displayed in Supplemental Tables 3 and 4 online. In instances where many, usually thousands, of motifs/7-mers were assessed in the same experiment, nominal P values were multiplied by this repeat number, generating a Bonferroni-corrected P value. Values are considered significant if their corrected P value is <0.05 and color-coded pink in our spreadsheets. Based on observations to be presented, motifs/sequences giving values with P values just below Bonferroni significance (nominal P < 0.001) are identified as “worth interest,” but are not called “significant”; these are color-coded blue.

The 7-mers were chosen because they were the longest random sequences that gave us adequate numbers of hits to CNSs >9 and because Guo and Moose (2003) used them previously.

CNS sequences were from Thomas et al. (2007); we include these in Supplemental Table 2 online. The control nongenic sequences were prepared by extracting the total gene space from our manual annotations and then by partitioning sequence fragments into categories representing αCNSs, noncoding nonCNS controls, and exons. We also prepared CNS and control noncoding sequence fragments from a subdatabase composed only of gene spaces from 126 pairs of genes labeled as Bigfoot pairs and 1197 pairs called Smallfoot genes. CNS and control fragments were from the gene pairs actually used to prepare the database. There were always far more noncoding nonCNS control sequences than CNS sequences. Therefore, we normalized the hits to our control sequences so we could compare our αCNS hits to control hits, assuming a 1:1 correlation would be expected, or that there would be no difference in 7-mer content between αCNS and control sequence. When there were ≥10 CNS hits, significance of difference from the 1:1 expectation was estimated by χ² and corrected for multiple tests (see above). Most ratios were not significantly different than 1.0. Those ratios >1.0 were marked as “O” in Supplemental Tables 3 and 4 online, and those ≤1 were marked as “U” whether or not the difference from 1.0 was significant or “worth interest.”

Evaluation of Our Results and Conclusions with TIGR Assembly Version 5 in Light of Version 6

After we froze our αCNS database, TAIR released a newer version of their annotations, TIGR assembly 6. Using the lists of changes available from the TAIR website (ftp://ftp.arabidopsis.org/Genes/TAIR6_annotation_release/), we estimated the effect these changes might have on our data. Of the 6358 genes we analyzed in pairs, 551 genes were revised. Of these, 228 changed gene models, 195 changed protein structures only, and 353 new splice variants were added. No genes used in our analysis were deleted or had splice variants removed. No “new” exons are represented as αCNSs in our database, presumably because this would require misannotation of both homoeologs, not just one. The αCNS database prepared using version 5 gene annotation did not require correction in light of version 6 annotation (Thomas et al., 2007).

Websites Cited

Websites used in our study include the following: Public Arabidopsis Synteny Viewer 1.0 (http://synteny.cnru.berkeley.edu/AtCNS/), Arabidopsis Small RNA Project (http://asrp.cgrb.oregonstate.edu/), BLASTView (ensemble) set on Arabidopsis TIGR assembly (http://atesensembl.arabidopsis.info/Multi/blastview?species=arabidopsis_thaliana), Database of Arabidopsis Transcription Factors (http://datf.cbi.pku.edu.cn/),
Acknowledgments

We thank Damon Lisch for discussions, our College of Natural Resources for partial subsidy of the Statistics and Bioinformatics Consulting Service, and especially the National Science Foundation (DBI-034937 to M.F.).

Received January 18, 2007; revised March 10, 2007; accepted April 19, 2007; published May 11, 2007.

References


