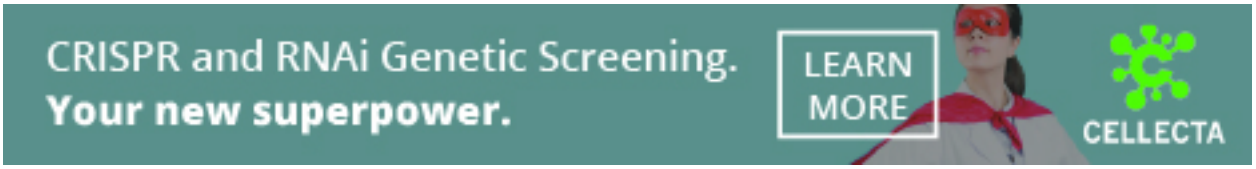
Conserved Noncoding Sequences in the Grasses4

Dan Choffnes Inada, Ali Bashir, Chunghau Lee, et al.

Genome Res. 2003 13: 2030-2041
Access the most recent version at doi:10.1101/gr.1280703



Letter

Conserved Noncoding Sequences in the Grasses

Dan Choffnes Inada,[1] Ali Bashir,[1] Chunghau Lee,[1] Brian C. Thomas,[2] Cynthia Ko,[3] Stephen A. Goff,[3] and Michael Freeling[1,5]

[1]*Department of Plant and Microbial Biology,* [2]*College of Natural Resources, University of California, Berkeley, Berkeley, California 94720, USA;* [3]*Torrey Mesa Research Institute, Syngenta Corporation, San Diego, California, USA*

As orthologous genes from related species diverge over time, some sequences are conserved in noncoding regions. In mammals, large phylogenetic footprints, or conserved noncoding sequences (CNSs), are known to be common features of genes. Here we present the first large-scale analysis of plant genes for CNSs. We used maize and rice, maximally diverged members of the grass family of monocots. Using a local sequence alignment set to deliver only significant alignments, we found one or more CNSs in the noncoding regions of the majority of genes studied. Grass genes have dramatically fewer and much smaller CNSs than mammalian genes. Twenty-seven percent of grass gene comparisons revealed no CNSs. Genes functioning in upstream regulatory roles, such as transcription factors, are greatly enriched for CNSs relative to genes encoding enzymes or structural proteins. Further, we show that a CNS cluster in an intron of the *knotted1* homeobox gene serves as a site of negative regulation. We show that CNSs in the *adh1* gene do not correlate with known *cis*-acting sites. We discuss the potential meanings of CNSs and their value as analytical tools and evolutionary characters. We advance the idea that many CNSs function to lock-in gene regulatory decisions.

[Supplemental material is available online at www.genome.org.]

Regions of DNA coding for protein are expected to exhibit sequence conservation between related species over evolutionary time due to the functional constraints of protein structure. It has become apparent during the past five years that noncoding sequence also exhibits functional constraints. Conservation outside of coding exons can be detected by cross-species orthologous gene comparisons. Such regions of noncoding DNA displaying strong sequence homology among distantly related organisms have been described as "phylogenetic footprints" (Gumucio et al. 1996), and can be as small as 6 bp when several diverged species are used. Conserved noncoding sequences (CNSs) in animals (Hardison 1997, 2000; Dubchak et al. 2000) and plants (Kaplinsky et al. 2002; Gau and Moose 2003) are phylogenetic footprints large enough to be detected when but two species are compared under conditions set by the investigator to reject all chance alignments. Significance levels are set so that CNSs represent sequences conserved due to selection while surrounding, nonconserved sequence has undergone the randomizing effects of mutations over time, as computed and discussed by Kaplinsky and coworkers (2002). Recently diverged noncoding sequences are expected to align in cross-species comparisons. For example, cauliflower (*Brassica*) and *Arabidopsis* are closely related crucifers that share about 20% of their noncoding promoter sequence (Colinas et al. 2002), as expected by neutral carryover from the ancestor. Such carryover sequences do not result from selection alone and are not considered CNSs.

Various algorithms and stringency schemes are currently employed to identify CNSs in mammals. Dubchak and coworkers (2000) and Loots and coworkers (2002) developed a global alignment (Avid) and scoring package called *Vista* (or *rVista*, which overlays potential protein binding sites). *Vista* identifies a CNS as any aligned sequence exhibiting 70% or greater sequence identity over at least a 100-bp interval. (The parameters can be altered by the user for customized results.) Another program, called SynPlot (Göttgens et al. 2001) uses an alternative alignment strategy and plots a graph of sequence identity along a horizontal axis to allow the user to discern visually conserved regions. Kaplinsky and coworkers (2002) used BLAST local alignment (Altschul et al. 1990) to find short, near-exact homologies between maize and rice orthologous genes, and graphed the results manually.

Individual CNSs in mouse–human gene comparisons have been shown to function based on transgenic CNS knock-out experiments (Loots et al. 2000) and by a tight concordance between the position of CNSs and DNAse I-hypersensitive sites (Göttgens et al. 2001). Furthermore, mammalian CNSs were found to be enriched for known transcription factor binding sites 10 bp or larger (Levy et al. 2001), and for clusters of known sites (Loots et al. 2002). It is generally accepted that CNSs are beacons to gene regulatory elements (Hardison 2000; Blanchette and Tompa 2002).

In mammals, using a human versus mouse comparison, Jareborg and coworkers (1999) calculated that large phylogenetic footprints (>60% identity over 100 bp, which is less stringent than the 70% identity requirement for mammalian CNSs used by other research groups) constitute a large portion of noncoding DNA: 36% of promoter sequence, 50% of 5′ UTR, 23% of introns, and 56% of 3′ UTR sequence. Indeed, regardless of identification strategy, a typical mammalian gene pair contains many often large regions of homology. No orthologous mammalian gene pair published to date is devoid of CNSs. Kaplinsky and coworkers (2002) found short CNSs in five of six grass genes examined, and compared these results to the mammalian literature and to six mammalian genes analyzed using grass CNS parameters (Kaplinsky et al. 2002.) They found that mammalian genes had many more, and much larger CNSs. Kaplinsky and coworkers hypothesized that the relative noncoding simplicity of grass genes reflects a relative simplicity of regulatory networks.

The grass family (Poaceae) is a monophyletic taxon of approximately 10,000 species, including all the major grain crops, such as *Oryza sativa* (rice) and *Zea mays* (maize). All but the most

[4]Present address: Syngenta Biotechnology, Research Triangle Park, NC 27713, USA.
[5]Corresponding author.
E-MAIL freeling@nature.berkeley.edu; FAX (510) 642-4995.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1280703.

13:2030–2041 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

basal grasses are thought to have originated from a common ancestor approximately 50 million years ago (Mya; Kellogg 2001). Kaplinsky and coworkers (2002) showed that maize–rice CNSs could be used successfully in all grasses as PCR primer binding sites. CNSs are now being used to extract promoter and other noncoding sequence from wild grasses, and to design pan-grass gene-anchored mapping tools. Given the importance of grass plants to humans and the availability of large amounts of orthologous sequence in the two distant grasses, rice and maize, these species are well suited for CNS analysis.

RESULTS

Development of Bioinformatic Tools

In order to identify regions of high conservation between gene pairs, we developed software to identify and display CNSs in a graph format. Although most mammalian studies identified conserved regions based on a global sequence alignment followed by an analysis of percent nucleotide identity within a window of a defined number of nucleotides (e.g., *Vista*; Dubchak et al. 2000), we employed a technique using the BLAST algorithm (Altschul et al. 1990; Kaplinsky et al. 2002). BLAST is an algorithm that, when applied to two input sequences, generates local alignments between the two sequences (in either the plus or the minus strand) nucleated by a minimum number of identical bases (the word size) and extends outward in both directions, allowing a certain number of base mismatches and gaps as determined by the mismatch allowance and gap penalties options. The results are provided in decreasing order of calculated statistical significance.

There are two elements to our software package. The first element, the "CNS Blaster" uses a Perl-based script to read a GenBank text file with CDS (coding sequence) annotation and a second text file containing the orthologous genomic comparison sequence, both chosen by the user. Coding exons are masked from comparison in the annotated sequence. The software then employs BLAST 2 SEQUENCES (bl2seq) to generate files containing the CNS analysis data, with parameters set exactly as by Kaplinsky and coworkers (2002). Using bl2seq to identify conserved sequences allows the detection of conserved regions on opposite strands of the comparison sequences, regions that are not positionally conserved in the two orthologs, and regions that are reiterated in one species, but not the other. To our knowledge, no global alignment algorithms are able to identify these classes of conserved elements. The second element of our package is the "Gene Annotation Viewer," which is a Java applet that reads the sequence annotation and CNS information from the first stage, aligns the orthologs at the beginning (5′ATG) of the coding sequence, and produces a simple, interactive display of the compared genes. We present an example of the graphic readout as Figure 1. Double-clicking on a CNS line on either gene generates a pop-up box with the CNS sequence displayed as text (upper window of Fig. 1). Dragging the mouse over any region of gene brings up the underlying sequence as text in another window (lower window of Fig. 1), with annotation in italics (this sequence shows the 5′UTR-exon1 junction, with the exon sequence in italics). Being able to copy sequence and paste into

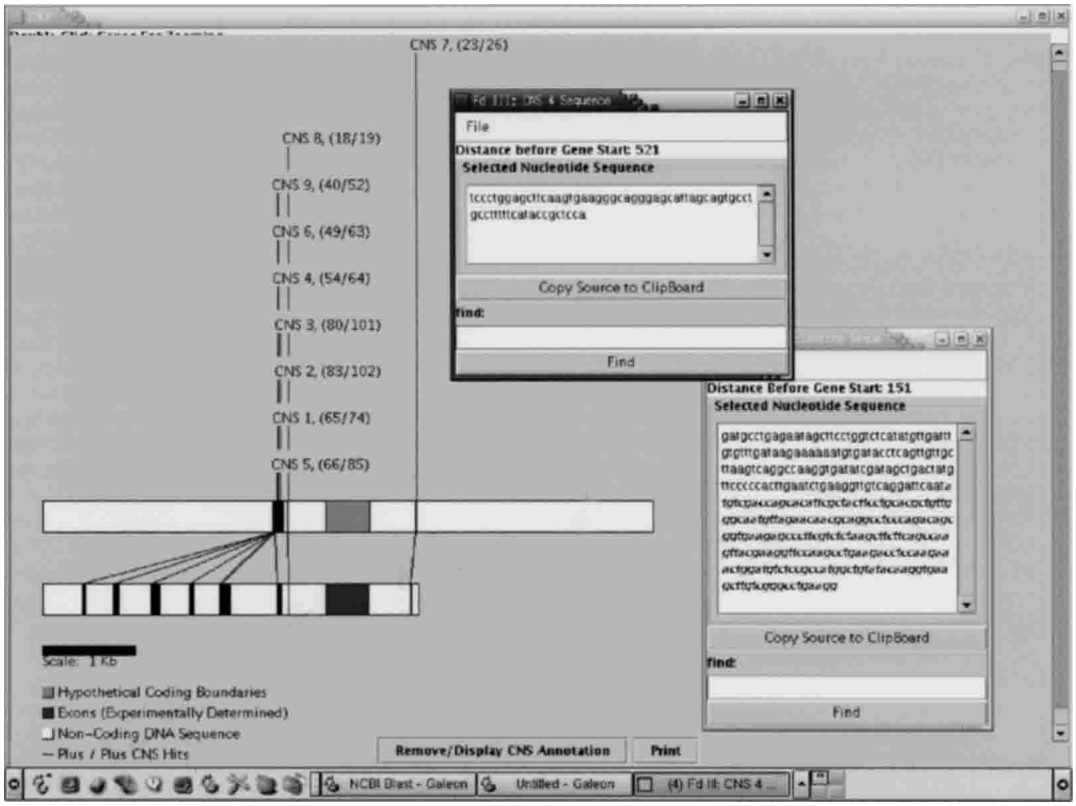


Figure 1 The CNS Viewer software displays the bl2seq results of an orthologous pair of *ferredoxin3* genes aligned on their start codon. The CNS repeats in the 5′ region of maize have been shown to be in a 5′ UTR intron (Nakano et al. 1997). The text boxes permit quick input to other applications, such as BLASTX or ORF predictors. The button "Remove/Display CNS Annotation" brings up the CNS list with the option to toggle "off." Once off, a CNS does not appear in the graphic, but the data remain. The dark blue box denotes the experimentally verified exon in maize. The light blue box denotes orthologous space between the rice 5′ATG and the stop.

Table 1. The 52 Grass Genes Used for CNS Analysis

Gene locus name	Name of gene product	Maize GenBank accession	Locus name in maize	Rice GenBank/TMRI accession
abi4	AP2 domain transcription factor	AY125490	ABA insensitive 4	CLC22051
adh1	alcohol dehydrogenase	AF123535	alcohol dehydrogenase1	AF172282
alt1	alanine aminotransferase	AF055898	alanine aminotransferase1	AAAA01000312.1
amy2	beta amylase	AF068119.1	beta amylase2	HTC0096333-A01.R.4.4
bt2	ADP glucose pyrophosphorylase	AF334959	brittle endosperm2	AAAA01001202.1
bx8	UDP-glucosyltransferase	AF331854	benzoxazinoid resistant 8	AL606615.2
cat3	catalase	L05934	catalase3	AAAA01004939
ckx1	cytokinin oxidase	AF044603	cytokinin oxidase	AAAA01000072
dek1	calpain-like membrane protein	AY061804	defective kernel1	CLC1853
dhy1	dehydrin-like protein	AF314251	dehydrin-like1	AAAA01000476.1
fad7	fatty acid desaturase	D63954	fatty acid desaturase7	AAAA01001919
fd3	ferredoxin III	AB001387	ferredoxin3	CL003093.85
fd6	ferredoxin VI	AB001386	ferredoxin6	CL006554.354
g2	putative transcription factor	AF298118	golden2	AP002855
gcn5	histone acetyltransferase	AJ428542	maize GCN5 (yeast) homolog	AF229187
gl8	beta ketoacyl CoA reductase	AF302098	glossy8	AAAA01000778.1
gn1	Knotted1-like homeobox transcription factor	AY312168	gnarley1	CL033612.103
gpa1	glycerald. 3-phosph. Dehydrog. A subun.	X15408	glyc.-3-Ph. dH, A sub.	AAAA01001985
gpc1	glyceraldehyde 3 phosphate dehydrogenase, cytosolic, C subunit	X15596	glyc.-3-Ph. dH, C sub.	CL027672.49
hsbp1	heat shock binding protein (empty pericarp2)	AF494285	heat shock protein binding protein1	CL034628.160
incw3	cell wall invertase	AF043346	cell wall invertase 3	CL032573.44
incw4	cell wall invertase	AF043347	cell wall invertase 4	AF155121
kn1	Knotted1 homeobox transcription factor	AY312169	knotted1	CL020544.98
lrs1	liguleless2-related transcription factor	AY180107	lg2-related sequence1	AP003794
lkrsdh	lysine ketoglutarate reductase/saccharopine dehydrogenase	AF271636	lysine ketoglutarate reductase/saccharopine dehydrogenase	AAAA01003034.1
mis1	aldehyde dehydrogenase, disease upregulated	AF467541	maize homolog of fis1	AAAA01000475.1
mlip15	cold-upregulated bZIP transcription factor	D63956	low temperature induced protein	CLB2912.4
mpcbp	calmodulin binding protein	AF250191	calmodulin binding protein	CL006707.158
noi	nitrate induced protein	AF030385	nitrate induced	HTC032618-A0.F.66.6
nr1	nitrate reductase	AF153448	nitrate reductase1	CLB8272.5
o2	O2	X15544	opaque endosperm2	AAAA01000962
ole16	oleosin (16 kD)	U13701	oleosin, 16 kD	CL024588.71.8
ole18	oleosin (18 kD)	U13702	oleosin, 18 kD	AF019212
pdc2	pyruvate decarboxylase	AF370004	pyruvate decarboxylase2	AAAA01000307.1
pdc3	pyruvate decarboxylase	AF370006	pyruvate decarboxylase3	CL039322.220
pep1	phosphoenolpyruvate carboxylase	E17154	PEPCase, cytosol. C4	AAAA01003016
phyt2	acidic phytase II	AJ223471	phytase 2	AL662942.2
psei1	cystein proteinase inhibitor	D63342	cystatin1	HTC027330-A01.122
rr1	cytokinin-inducible response regulator 1	AB031011	response regulator1	AAAA01005417.1
sbe1	starch branching enzyme I	AF072724	branching enzyme1	HTC116571-A01.R.5.5
sh1	sucrose synthase-1	X02382	shrunken1	X64770
sh2	ADP glucose pyrophosphorylase, 60kD	M81603	shrunken2	CL034645.143.17
su1	starch debranching enzyme I	AF030882	sugary1 (isoamylase)	AAAA01000810
sus1	sucrose synthase-2	L33244	sucrose synthase	X59046
tb1	teosintebranched1 transcription factor	AF464738	teosintebranched1	AAAA01001682.1
te1	TE1	AF348319	terminal ear1	AAAA01003386
trpA	tryptophan synthase, alpha	X76713	trytophan synthase a	CL004278.147
tse	endoxylanase, tapetum-specific	AF149016	tapetum specific endoxylanase1	AAAA01015527
y1	phytoene synthase	U3263.1	yellow1	AAAA01004470.1
zag2	Zea Agamous 2 transcrition factor	X80206	Zea Agamous2	AAAA01000742.1
zmm1	MADS-box transcription factor	X81199	Zea Mays MADS-box protein 1	AAAA01000742.1
zog1	cis-zeatin-O-glycosyltransferase	AF466203	cis-zeatin-O-glycosyltransferse1	CL010916.37.36

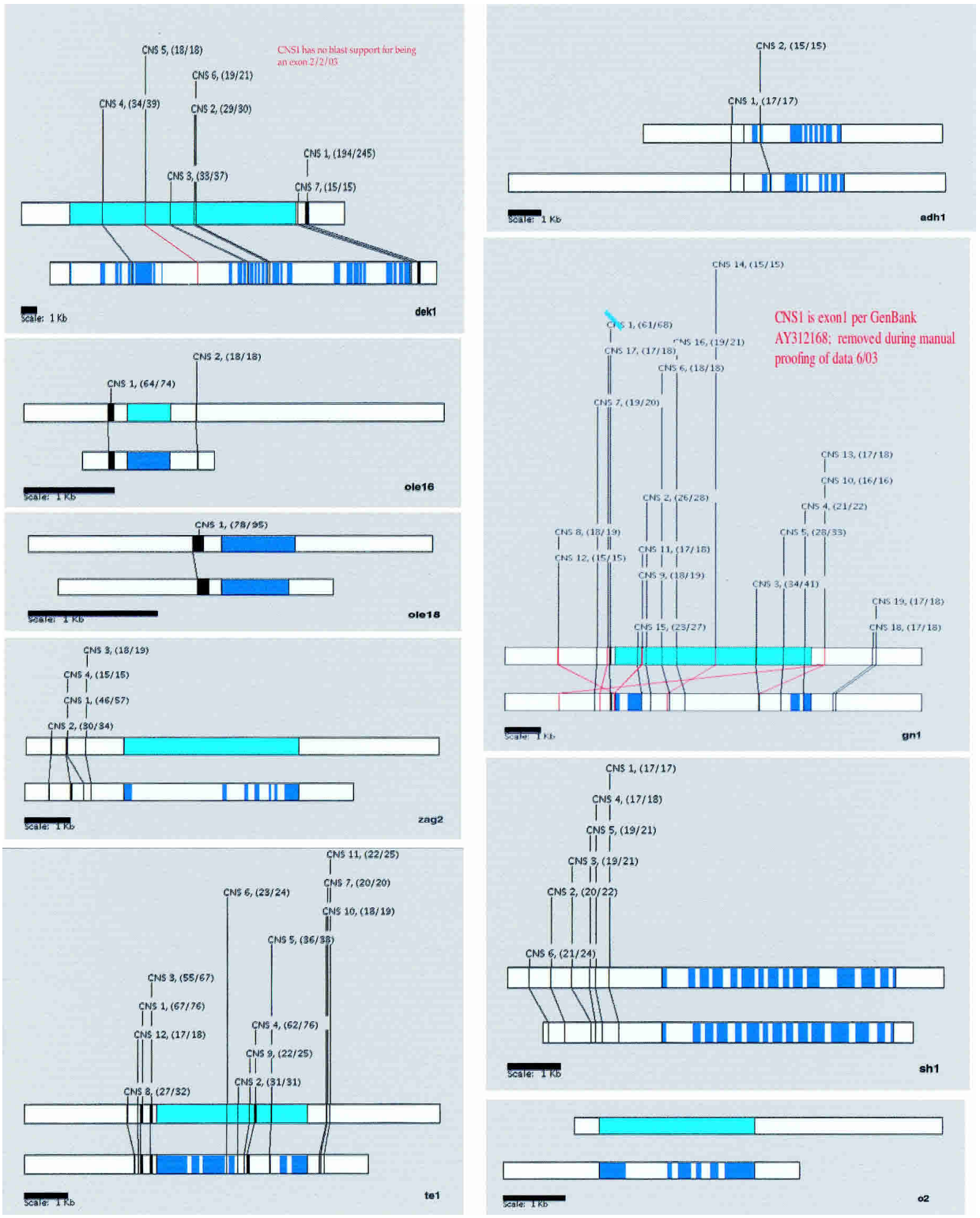


Figure 2 Examples of CNS graphic readouts from our 52 genes. From the upper right corner and moving clockwise, *alcohol dehydrogenase*1-F, homeobox gene *gnarley1*, *shrunken1* encoding an enzyme, downstream bZip gene *opaque-2*, RNA binding motif gene *terminal ear1*, MADS-box *zea agamous2*. *oleosin16* and *18*, and *defective kernel1* encoding a membrane-bound calpain. Red lines indicate change of strand. A blue line through a CNS denotes removal during manual proofing of data.

other applications facilitates manual annotation. For example, some sequences, although deposited in GenBank as annotated noncoding sequence, returned CNSs with homology to plant coding sequences. Although not the focus of the present study, such sequences are of interest for the study of gene structure and cross-species genetic colinearity.

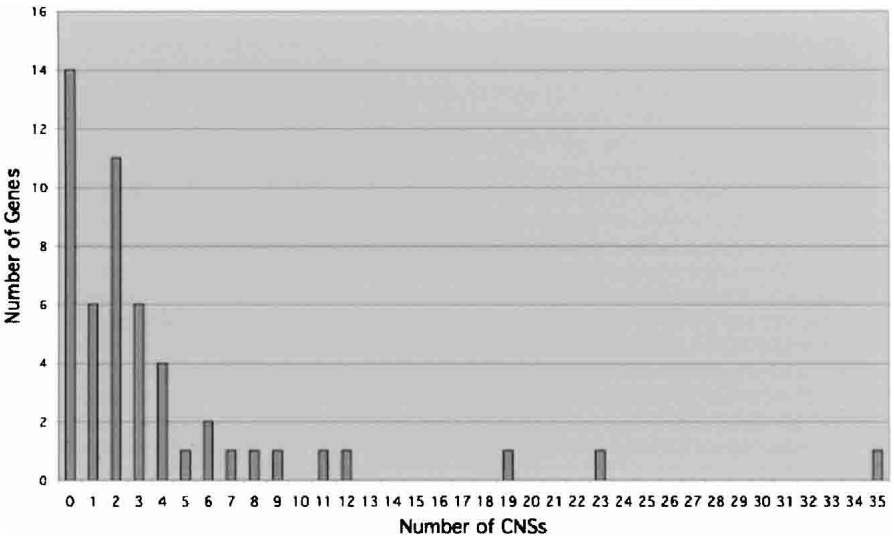


Figure 3 Our 52 two genes distributed by total number of CNSs. The most CNS-rich genes are upstream, developmental regulatory genes. The six CNS-richest, in descending order, are *lrs1* (bZip-TGA-1a family), *kn1* and *gn1* (Class I homeobox), *te1* (RNA-binding product; apex domain identity), *sus1* (encodes an enzyme), and *tb1* (transcription factor).

For this study, we focused on sequences having a bl2seq statistical score of at least the significance of a 15-nucleotide (nt) identical match between the compared sequences. Kaplinsky and coworkers (2002) calculated that BLAST hits between maize and rice at this significance did not occur by chance carryover.

Fifty-Two Maize–Rice Orthologous Gene CNS Descriptions

We chose 52 annotated maize genes from GenBank and compared them with their rice orthologs. To compare larger sets of noncoding sequence and to be more confident of exon structure, we limited our comparison set of maize genes to those having approximately 1 kb or more of annotated promoter sequence as well as the entirety of exons and introns, and had been annotated experimentally using cDNA sequence. Using these criteria, we gathered a set of maize genes diverse in terms of size, structural complexity, and function (Table 1). We then identified the rice ortholog by searching the draft *Oryza sativa* ssp. *japonica* (Torrey Mesa Research Institute [TMRI] database http://www.tmri.org; Goff et al. 2002), *O. sativa ssp. indica* (Yu et al. 2002), and partial japonica (http://www.rgp.dna.affrc.go.jp; Rice Genome Project 2002) genomic sequences for the most homologous sequence to the annotated maize gene. Eight of our maize gene sequences carried one or more adjacent genes. These linked genes were also found together in rice, and these orthologs were also the best homologs. We also checked that the rice and maize coding regions showed approximately 85% (75%–90%) sequence identity, which is typical for rice/maize orthologs diverged from the common grass ancestor. Table 1 gives the GenBank or TMRI accession numbers for all 52 gene pairs. Supplemental Data (Table 4) identify the length of gene space used for each of these genes. Graphic representations of our BLAST results are exemplified in the nine gene pairs in Figure 2. All 52 graphs are available as Supplemental Data via the *Genome Research* Web site (www.genome.org) or our own Web site (http://genomics.cnr.Berkeley.edu/cns/ User = reviewer; Password = super).

Grass CNS Statistics

Figure 3 shows the number of CNSs per gene in our data set. The majority of genes compared have at least one CNS, and nearly all genes have four or fewer. A smaller number of genes have from five to 35 conserved noncoding elements. The largest single class of genes (27%) have no CNSs detectable using these parameters. Figure 4 distributes CNSs by length. Most CNSs are small (<20bp), but CNSs of larger size occur frequently. We detected nine CNSs greater than 80 bp in length.

CNS-Rich Genes Tend to Be Upstream Regulatory Genes

CNS analysis reveals enormous disparity in genes of different function (Tables 2 and 3A,B). Because the genes are heterogeneous in length of their constituent coding and noncoding regions, they do not lend themselves readily to a full statistical analysis, but some general trends are apparent. Overall, genes encoding structural proteins or enzymes have few or no detectable conserved regions (e.g., *adh1*, *amy2*), even if their regulation is known to be developmentally complex (e.g., *adh1*). Genes encoding upstream, developmental transcription factors (e.g., *lrs1*, *gn1*, *kn1*, *abi4*, *tb1*) or proteins thought to act in a complex regulatory manner (e.g., *te1*) are, in general, enriched for CNSs. Exceptions to this prevailing trend exist. For example, *dek1* encodes an essential, membrane-associated calpain-like protein, and has a

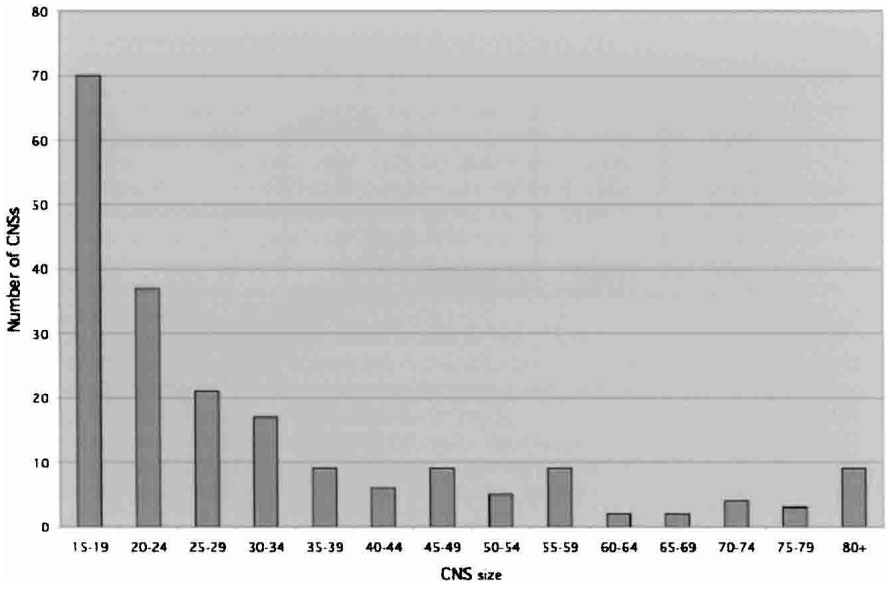


Figure 4 CNSs in 52 grass genes distributed by size.

Table 2. Average Number of CNSs per Genes Organized by Functional Class

Class of gene (percent of total genes)	Average number of CNSs
Enzymes (56%)	2.4
Structural (19%)	2.1
Gene regulation (25%)	9.0

relatively high number of CNSs at seven, including a very large 245-bp sequence in the 3′UTR. On the other hand, the TGA-1a-type bZip transcription factor *o2*, regulating balance of storage proteins in the maize endosperm (a relatively "downstream" function), has no significant CNSs compared to its rice ortholog. This is especially interesting because another bZip-encoding gene in the same subfamily, *lrs1*, is the most CNS-rich gene in this data set, with 35 CNSs.

We calculated CNS density in grass genes by dividing the total number of bases occupied by CNS by the total number of noncoding bases studied per gene (Supplemental Table 4). Table 3B presents these CNS density data organized by type of gene. The results show that genes encoding regulatory proteins have a higher CNS density than genes of other functions.

Grass Versus Mammalian CNS Frequencies, Sizes, and Densities

Kaplinsky and coworkers (2002) found that an orthologous human–mouse sequence comparison containing six genes had an average of 17.7 CNSs per gene using the grass CNS parameters, and this measure was in agreement with other studies in mammals (especially Jareborg et al. 1999). To determine whether mammalian genes in general have more CNSs than plant genes, using our parameters, we compared a set of seven functionally diverse human gene sequences with their mouse orthologs. (Genes studied: ODC1, MPO, PROC, PIM1, MYCN, IL6, ADCYAP1.) Every gene pair revealed many more, larger CNSs than found between grass genes (data not shown). For example, *Pim-1* kinase (human M27903; mouse M13995) has 31 CNSs, nine of which are greater than 100 bp in length, and two in excess of 300 bp.

Since grass genes encoding housekeeping enzyme functions have, in general, few and small CNSs, but all mouse–human gene comparisons revealed numerous, large CNSs, we expected that a comparison of mammalian housekeeping orthologs would reveal comparatively abundant CNSs. To test this hypothesis, we chose three human–mouse enzyme-encoding genes whose functions are also represented on our grass list (Table 1). The grass gene *glyceraldehyde-3-phosphate dehydrogenase* (*gpa1*) has only two CNSs, and a density of 2.49%, however, the homologous mammalian gene GAPDH (human J04038; mouse NW_000265) has 15 CNSs, five of them over 70 bp in length, at a density of 12.3%. Whereas the grass comparison for *alcohol dehydrogenase* (*adh1*) produced two CNSs, both under 20 bp in length for a density of 0.27%, a homologous gene from mammals (ADH1; human NT_016354; mouse NT_039242) revealed 15 CNSs, four of them over 50 bp, and a noncoding space that is 3.6% CNS. Finally, we compared two pairs of genes encoding enzymes involved in sugar homeostasis in both grasses—*sucrose synthase* (*sus1*)—and mammals—*glucose-6-phosphatase* (G6PC: human NT_010755 ; mouse NT_03952). The grass gene comparison produced six CNSs, none longer than 24 bp in length, and CNSs constituted 2.37% of the gene's noncoding sequence. The mammalian gene pair produced over 50 CNSs with one 403 bp and another 174 bp in length at a density slightly higher than 10%.

A very low proportion of grass noncoding sequence is conserved between rice and maize: Approximately 2% of the noncoding nucleotides used in this study are conserved, as identified by our parameters (Suppl. Table 4). Such a low percentage of conserved noncoding sequence content in plants is in sharp contrast to the situation in the mammalian genome (first observed by Kaplinsky et al. 2002). Our analysis of many mammalian genes of diverse function revealed that between 5- and 20-fold more noncoding human–mouse gene space is occupied by CNSs than in grasses (see above; other data not shown). Although Jareborg and coworkers (1999) used different parameters for identifying conserved regions in a large set of human–mouse orthologs, they estimated that 26%–56% of mammalian noncoding space is conserved.

A Cluster of CNSs in a *knotted1* Transcription Factor Gene Intron Corresponds to a Region Where Transposon Insertions Result in Ectopic Expression

To address the possibility that CNSs can serve as gene regulatory components, we sought to identify mutations in and around CNSs. A report from the Hake laboratory (Greene et al. 1994) studied nine transposon insertions into the *kn1* Class I homeobox gene, each of which resulted in a strong leaf phenotype conferred due to the ectopic synthesis of the KN1 product in leaf initial cells, primordia, and leaves. These workers located the insertion points of these nine dominant *Knotted* mutants to eight different positions, all within approximately 300 bp in the 5′ half of intron 3. Figure 5 maps the locations of the Greene et al. (1994) insertion sites onto the complete *kn1* gene sequence. As demonstrated by Figure 5, 13 of the *kn1* CNSs, including seven over 50 bp long and two over 100 bp long, are clustered in the 5′ half of intron 3, an area of high maize–rice homology and the location of insertions resulting in ectopic gene expression. The concordance of insertion position with this CNS-rich region strongly implicates the underlying sequences in the negative regulation of *kn1* in the leaves.

To further investigate the potential regulatory nature of the *kn1* third intron CNSs, we sought to address the question: Is the 5′ portion of the intron, which has numerous CNSs, enriched for known plant transcription factor (TF) binding motifs relative to the 3′ portion, which has very few CNSs? We submitted the maize *kn1* intron sequence to the PlantCARE Web application

Table 3A. Number of Genes by Functional Class Sorted by Number of CNSs

Class of gene	O CNSs	1–10 CNSs	11–20 CNSs	21 or more CNSs
Enzymes	9 (31%)	19 (66%)	1 (3%)	0
Structural	4 (40%)	6 (60%)	0	0
Gene Regulation	1 (8%)	8 (62%)	2 (15%)	2 (15%)

Percentage of genes per class is shown in parentheses.

Table 3B. CNS Density Sorted by Gene Functional Class

Class of gene	Percentage noncoding sequence that is CNS
Enzymes	1.29%
Structural	2.41%
Gene regulation	4.28%

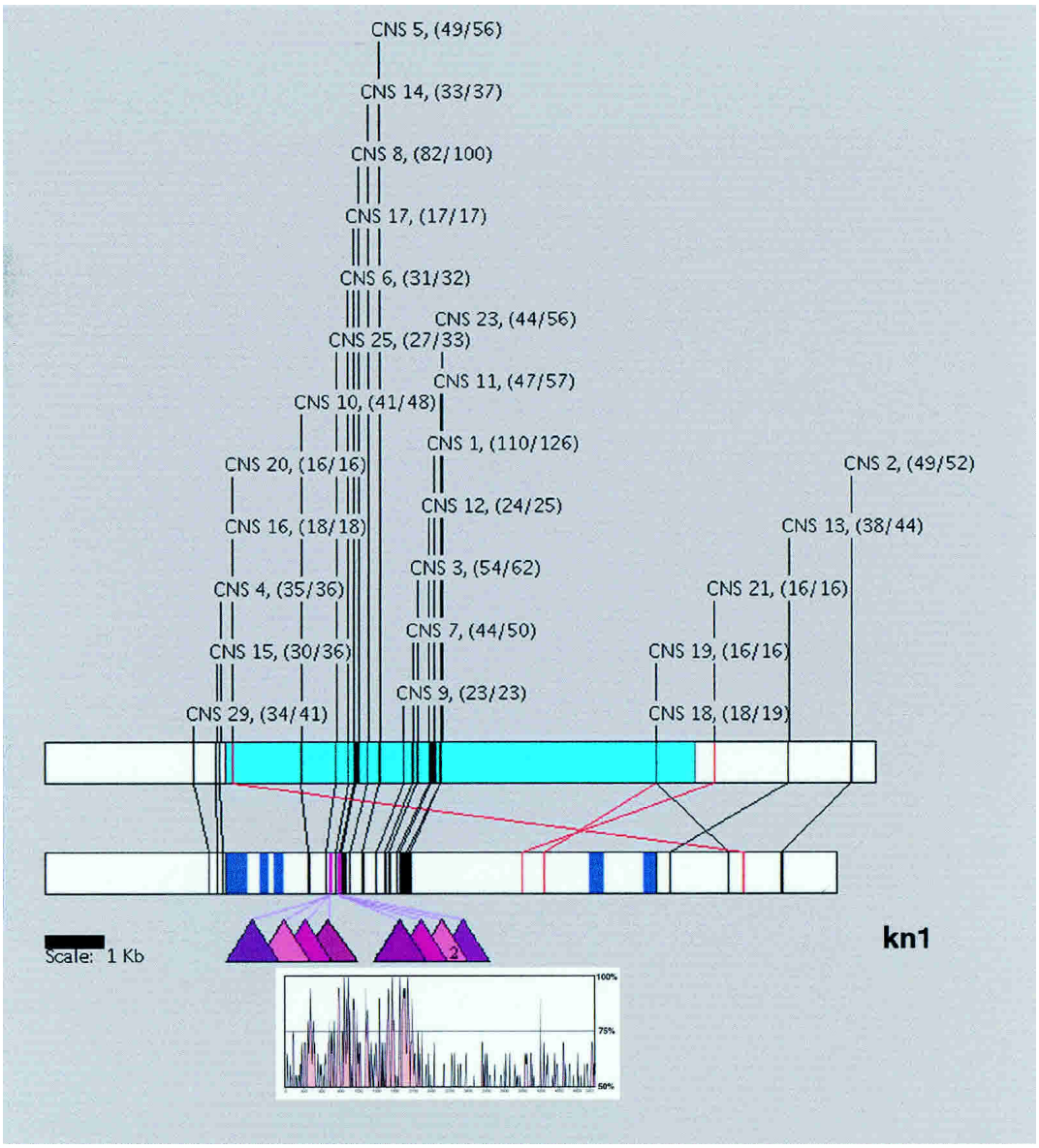


Figure 5 CNSs may bind a negative regulatory factor(s). Greene and coworkers (1994) positioned nine transposon insertions (colored triangles) within the third intron of the *knotted1* Class I homeobox gene in maize. These insertions were found as dominant mutants conferring ectopic gene expression and phenotype. Note how they cluster in a region rich in CNSs. The bottom-most plot is a global alignment of the third intron (*Vista*; Loots et al. 2002), showing bp identity in a 20-bp sliding window.

(http://oberon.rug.ac.be:8080/PlantCARE/index.html; Lescot et al. 2002), an online tool that searches for 435 experimentally determined higher plant TF binding site boxes and motifs (as of November 2002; most of these sequences overlap the binding site of another, so this is not a "unibinding site" number). In the PlantCARE database, a typical box is 4 bp. In this experiment, we used the 5′ half of the third intron as a CNS-rich sequence, and the 3′ half as a CNS-poor control sequence. The 5′ half contained 480 sites representing 55 binding motifs. The 3′ half contained 612 sites and 59 motifs. Visual inspection of the patterns of motifs found no significant differences in abundance or distribution relative to CNSs. Therefore, although a regulatory role for the *kn1* third intron CNSs is demonstrated by the ectopic gene expression resulting from disruptive insertions, we could not correlate CNS richness with known TF-binding-site richness.

Conserved Elements in the Promoter of *adh1* Do Not Include Known *cis*-Acting Binding Sites Nor a Potential Scaffold Attachment Region

To further address the potential regulatory nature of CNSs, we identified CNSs in a gene whose promoter DNA regulatory elements have been mapped relative to its regulation. *Adh1*, a well studied gene in grasses, is necessary for surviving seed and seedling flooding. It has two small CNSs, and only CNS1 (17/17) is in the promoter, at about 350 nt upstream of the 5′ATG (−274 to −290, inclusive).

ADH1 is expressed constitutively in certain vascular cells, root caps, scutellum, pollen mother cells, and the generative cell of the male gametophyte, is anaerobically and auxin-inducible in root and mesocotyl parenchyma, but is not expressed in leaves (for review, see Freeling and Bennett 1985; M. Freeling, unpubl.).

Additionally, it is known that different naturally occurring alleles have different, reciprocating quantitative balances of tissue-specific expression (Woodman and Freeling 1981). In addition to TATA and CAAT boxes identified by most annotators of *adh1* upstream sequence, biochemical experiments on the *Adh1-F* allele have found promoter G-box-like anaerobic regulatory elements (AREs; Walker et al. 1987) positioned between −140 and −99, which are bound by proteins. DMSO in vivo footprints also mark the ARE general region (Paul and Ferl 1991). GenBank annotation (M17134; Ferl et al. 1987) identifies a 61-bp S1 hypersensitive region with a 3′ (proximal) end at −296, which is close but does not overlap the 5′ end of CNS1. Two DNAseI-hypersensitive regions (Paul et al. 1987) have been located with respect to the regulation of *Adh1-F* at −35 to −150 (anaerobic-sensitive) and −160 to −700 (constitutive). Osmium tetroxide footprinting identified a putative AT-rich scaffold attachment region far upstream at −780 to −500 (Paul and Ferl 1993).

Because the TATA, CAAT, and ARE regions, for which protein binding is certain, coincide with no CNSs, we lowered our CNS statistical significance cutoff to the level of a 13/13 exact match. We did not find hits in the general area within 200 bp from the start of transcription where these sites exist. The DNAseI-hypersensitive sites include almost the entire promoter, and so lack specific diagnostic power. The putative scaffold attachment region is not a CNS. In general, areas of the *Adh1-F* promoter known to bind proteins are not CNSs, and those protected by bound molecules in footprint experiments show a poor correlation with CNS1 or other, less significant phylogenetic footprints. Although it is possible that the S1 hypersensitive region is related to CNS1, which is 5 bp proximal, CNSs are not obvious beacons to the sites where proteins bind the *adh1* promoter.

CNS Discovery Made Unexpected Contributions to Grass Gene Annotation

Our 52 CNS analyses, using available maize sequence surrounding single genes, discovered four close-by candidate gene coding regions by virtue of strong sequence conservation (two exons downstream of *gl8*, three exons downstream of *gpa1*, two exons upstream of *hsbp1*, and four exons nearby and upstream of *tse*). The neighbor of *hsbp1* encodes no known transcript present in the plant databases. However, it has homology (e-value 1e-60) to the hypothetical gene At1g52640 in *Arabidopsis*. Upstream of the annotated maize *tse* gene is a patch of exon-like maize/rice homologies. These regions are probably exons in maize, because they are nearly identical to the rice predicted *tse*-like genes and products of similar function in *Arabidopsis* and papaya (data not shown).

DISCUSSION

In this analysis of 52 grass gene pairs for conserved noncoding sequences, we found that, in general, those genes with the largest number, size, and density of CNSs are transcription factors or other regulatory genes. Those genes with few or no CNSs are generally enzyme-encoding or structural protein-encoding. Although most genes had from zero to four conserved elements, 11 genes had from five to 35 CNSs. We were unable to identify any CNSs using our parameters in 27% of the genes studied. The CNSs identified were usually small segments between 15 and 19 bp in length, although some CNSs were much larger, on the order of 50–100 or more bp. CNSs were approximately as likely to occur in upstream regions as in introns or downstream regions. The short size and infrequency of grass CNSs is consistent with the promoter comparisons of Gau and Moose (2003).

We know of no study in any taxon that identifies a subpopulation of genes as "richer" in noncoding components. It is not apparent from previous work that upstream regulatory genes (such as developmental transcription factors that are only "on" at specific times and places) should be thought of as more complex than worker genes (such as enzymes) that are used at many developmental times and places. On the contrary, genes that are available for transcriptional regulation at many different times and places in an organism are known to have complex promoters with many, often reiterated, binding sites for the many different specific transcription factors used to exert control (Quinn 1996; Bolouri and Davidson 2002). Our data imply that "boss" genes, or many of them, have more noncoding space under higher sequence selection than "worker" genes, as if upstream regulatory genes were themselves under either more or special control.

Our most CNS-complex genes (*lrs1*, *te1*, *gn1*, *kn1*), along with the previously reported *lg1* (Kaplinsky et al. 2002) encode developmental regulatory factors; these are identified in the legend of Figure 2. The *lrs1* gene, encoding a basic leucine-zipper protein, is the genomic duplicate (paralog) of maize *lg2*, required for precise transitions along the proximo-distal axis of the leaf and apical-basal axis of the shoot (Walsh et al. 1998), *te1*, with a probable RNA-binding motif, specifies the identity of a particular region of shoot apex (Veit et al. 1998), and *gn1* and *kn1* are Class I homeobox genes required for proper apical-basal identities along the shoot (Freeling 1992). Many of the zero-CNS class are enzymes. Even so, there are exceptions to the trend. Transcription factors *o2* and *g2* have no or few CNSs. Some enzyme-encoding genes have several, including *sucrose synthase1(sh1)* with six CNSs in the 5′ region. When more maize genome sequence is available, it will be possible to better understand the relationship between CNS number and gene function.

In One Case, a Cluster of CNSs May Bind Negative Regulatory Factors

We showed that a CNS-rich region of the *kn1* third intron has undergone nine transposon insertions, each involving ectopic expression of KN1 in the leaf. These data suggest that the region has been conserved in sequence because it may bind some regulatory factor or factors that turn *kn1* "off" in the leaf. The transposon *Mu* insertions themselves do not cause the maximum phenotype; this phenotype occurs when an insertion mutant is in a *Mu*-active genetic background. Greene and coworkers (1994) interpreted this result as indicating that the negative regulatory function was only fully interrupted when *Mu* ends were bound with their transposases and other related molecules, preventing the binding of factors that normally would be responsible for *kn1* down-regulation.

Investigating the potential molecular function of *kn1* CNSs, we found that the 13 CNSs in maize–rice *kn1-intron3* are also present in the *intron3* of the *kn1* ortholog in barley, a grass in yet another subfamily of grass, as was expected from the previous work of Kaplinsky and coworkers (2002). However, none of these 13 CNSs is present elsewhere in GenBank sequence (data not shown, BLAST search performed 12/02) . Detailed examination of the plant TF binding motifs and the distribution of their specific sites in this CNS-rich region of *kn1-intron3*, using the 3′ CNS-poor half of the same intron as a control, found nothing emerging from the noisy background.

Comparisons of global and local alignments in this 13-CNS region, and local alignments done at a variety of stringencies, found that about half of these CNSs are discreet, in that they are surrounded by gaps, whereas about half are surrounded by sequence that can be aligned, but is less well conserved. Thus, examination of the structure of this CNS-rich region of *kn1-intron3* (Fig. 5 and associated text) does not clearly answer the question: Are these conserved CNSs 13 independent modules, or

one big module, or some combination? Future studies could look for factors binding the *kn1* third intron CNSs and may elucidate any potential role in the tissue-specific regulation of gene expression.

Are CNSs Transcription Factor Binding Sites?

Although we suspect that a CNS-rich region may serve a regulatory function, it is apparent from the large fraction of grass genes having few or no CNSs that many plant genes have not conserved, and thus, do not require, elements of the size identified by our parameters. Because our approach identifies conservation above a 15/15 identity threshold, we are likely unable to resolve conserved footprints resulting from selection at binding sites of single regulatory proteins. Individual transcription factors bind relatively small lengths of sequence, generally 4–10 bp (e.g., Bucher 1999). However, transcription factors do not bind as single elements, but function in large complexes of many independent factors over a stretch of DNA (Bolouri and Davidson 2002). A few mammalian CNS studies (Levy et al. 2001, on TF binding sites >10 bp long; Blanchette and Tompa 2002; Loots et al. 2002) suggest that mammalian CNSs, although large, are augmented for transcription factor binding sites.

In contrast to mammalian studies, plant orthologous gene comparisons return a significant fraction having no detectable conserved segments above the threshold level. Because all of the sequences compared presumably represent functional alleles, and because genes require *cis*-acting binding sites for expression, any conserved regulatory sequences must fall below the 15/15 significance threshold of our parameters. Interestingly, the known regions of protein binding in the *adh1* promoter did not coincide with a promoter CNS, indicating that the promoter regions responsible for tissue-specific, anaerobic response, hormone response, and constitutive expression do not produce a large enough footprint to have been conserved between maize and rice since their common ancestor.

It is possible that we were unable to detect enrichment for transcription factor binding motifs in CNSs due to the rudimentary nature of current knowledge of such sites in plants. Indeed, the majority of sites listed in the PlantCARE database are 4 bp in length, and only constitute a few hundred known sites. Further, very small sequences are extremely common in coding and noncoding DNA, and the presence of functional sites is often difficult to distinguish due to the high number of sequence matches for nonfunctional sites. Since the "noise" at any element length below approximately 10 bp is expected to be high, our results were expected. Therefore, future attempts to correlate CNSs with DNA-binding motifs should utilize a larger database of sites and focus primarily on the larger, less frequent motifs, or clusters of such motifs.

Mammalian and Grass Genes Differ in CNS Content

In order to properly compare maize–rice and human–mouse CNS data, it is important to show that a neutral base pair has approximately the same chance to undergo a base substitution in either lineage, and that this frequency is adequate to have randomized functionless DNA. Multiplying published average mutation rates in grasses and mammals by their estimated divergence times, Kaplinsky and coworkers (2002) calculated substitutions/neutral bp/yr to be 0.35 for maize–rice and 0.32 for human–mouse. These products are nearly the same. Using these products, Kaplinsky et al. further calculated the chance that 15 bp of positionally conserved sequence could be carried over from the grass or mammalian ancestor without selection; the chances were very low, at 4 in a million. These calculations argue that the comparisons of grass and mammalian CNS frequency and density data are proper.

It is intriguing that mammalian CNS studies find vastly greater tracts of conserved noncoding sequences among genes of all functional classes. We showed that for the homologous *glyceraldehyde-3-phosphate dehydrogenase*, *alcohol dehydrogenase,* and sugar homeostasis "worker" genes in grasses and mammals, the maize–rice comparisons revealed fewer and shorter CNSs than the human–mouse comparisons. Overall, we found that the average grass gene has about three CNSs whereas the typical mammalian gene has between 15 and 20. Whereas plant genes generally have short (20-bp) CNSs that range to approximately 100 bp in length, mammalian genes routinely have CNSs over 100 bp.

If mammalian genes' noncoding sequences have been subject to selection over more numerous and much larger regions than grass genes, then the functional constraints on noncoding regions are greatly weaker in grasses than mammals. Perhaps grass genomes have undergone a particularly rapid evolutionary process during periods of high transposon activity and genome restructuring. Such a process might be expected to contribute to the evolution of chromosome-level structure, but seems to have little effect on the relatively small stretches of DNA surrounding and encoding genes. In all eight cases where more than one gene was found on our maize sequence, both genes were found together in rice, and relative exons-intron sizes were similar as well.

Are CNSs Signatures of Developmental Complexity?

Since the similar mutation rates calculated using nonsynonomous/synonomous base substitution frequencies in coding sequence between human–mouse and maize–rice allow for a comparison of genomes similarly evolved at the nucleotide level, we believe the most likely explanation for relative mammalian CNS richness lies in one or more of the profound biological differences between mammals and grasses. (It is tempting to extrapolate what we find in maize–rice comparisons to other flowering plants, just as it is tempting to extrapolate from human–mouse to all other vertebrates or metazoans. Since there is not yet enough genomic sequence data to test these extrapolations, we now confine ourselves to grasses and mammals.)

A simple definition of complexity involves the product of the number of parts and the number of connections between those parts. Because mammals have so many more stem-cell populations, organs, and organ systems than do grasses, and because plant cells are cemented inside walls, which prohibits making new connections by rotating or migrating, mammals must surely be the more complex organisms. Mammalian cells follow strict developmental pathways and differentiate terminally, events undoubtedly necessitating the permanent modification of chromatin state, often associated with the binding of regulatory proteins to DNA (Li 2002). Plant cells, in contrast, are distinguished by their developmental "plasticity" (Walbot 1985, 1996). In general, plant form and behavior are more accommodating to the environment than those of animals. For example, when organs in a meristic series change form in plants, organ identities often appear to overlap. With rare exceptions, differentiated plant cells quickly dedifferentiate in culture, and express totipotency. While overall plasticity is expressed in most every plant structure and behavior, the flexibility of plant reproductive strategies, including ingenious asexual schemes, is noteworthy. Mammalian cells, in contrast, are more determined, and developmental boundaries are more exact.

It is possible that mammalian RNA transcripts are richer in information content compared to grasses, allowing for the production of numerous specialized proteins from a single gene. In such a scenario, mammalian introns and UTRs would be rich in

conserved instructions for the proper function of the splicing machinery. Alternative generation and splicing of RNA transcripts has been shown to be common in mammals generating different transcripts depending on developmental or other cues, whereas alternative splicing in plants is relatively uncommon. Perhaps binding at CNSs facilitates this capacity for which mammalian genes are more complex.

Perhaps CNSs are involved in transcriptional regulation, but do not serve as traditional transcription factor binding sites. Some may perform roles in binding or producing regulatory RNAs (microRNAs). Alternatively, some CNSs may serve not to catalyze the stepwise binding of individual proteins, but in the assembly of large multiprotein complexes to perform complex regulatory functions. In such a case, a large CNS might not be a "binding site," in the classical sense but rather a "structural template" for the formation of chromatin-level control factors, as proposed by Kaplinsky et al. (2002).

One Particularly Robust Idea for the Function of the CNS Difference Between Grasses and Mammals

Among the several previously mentioned functions that may be served by large CNSs and CNS-rich regions, one function may be particularly useful as a guide for further experiments. There may be a greater negative consequence if mammalian genes, not just upstream regulatory genes but most every gene, are not turned "off" and kept off in the many, determined stem cell populations. The difference between grass and mammalian CNS frequencies, size, and densities may reflect a fundamental difference in the amount of regulation expended to limit gene availability in stem-like cell populations, and to lock-in such deterministic decisions.

Limitations

It is possible that our selection of 52 maize genes, although diverse in function, is not a representative sample of grass genes, since we chose many of the individual genes most intensively studied in maize over the years. Although a few dozen sequenced maize BACs exist, we chose to focus on genes for which an experimental exon annotation exists. We wanted these first 52 genes to have known coding regions and intron-exon boundaries involving one functional transcript. As a result, we are reasonably sure that the CNSs identified are not coding sequences. However, it is possible that among these genes, a conserved region—annotated as noncoding—may be present in a translated, alternative transcript which has not yet been captured as an expressed sequence.

Currently, the high-quality rice genome (the public Rice Genome Project) is incomplete, and maize, sorghum, and other grasses have very limited stretches of genomic sequence completed. All cross-species comparisons are by necessity limited by the length and quality of the available genomic sequences. We are not currently able to assess the possibility that plant genes may be regulated by distant, conserved elements, such as is the case with the mammalian α-globin locus, which possesses an enhancer 40 kb away from its coding region (Vyas et al. 1995). Likewise we have not yet determined whether individual CNSs display specificity, affecting the expression of certain genes in a region, but not others (Loots et al. 2000).

Because our maize genomic sequences were generally limited to the space surrounding a single gene, we were only able to use synteny to certify orthology for eight of our 52 genes. For these genes the true ortholog was also the best hit in rice. Analysis of the other 47 genes required that the maize gene be compared to its best homolog in all of the rice genomic sequence available, which included two whole genome contig databases, plus the collection of contigs produced by the Rice Genome Project. We also checked to see whether average exon nucleotide sequence identities were near 85% (75%–90%), which is what we generally find between syntenic maize and rice genes. We are confident that the comparisons presented are orthologs, and derive from a common ancestor about 50 Mya. It is possible that some gene pairs may share sequence similarity because they are members of a diverged gene family, but trace back to an ancestor more ancient than the first grass. This situation could result, for example, from rapid evolution of an ortholog or from a deletion removing an ortholog, leaving only distant homologs for comparison. As more long genomic sequences become available from rice and other grasses, synteny will be increasingly valuable to identify unambiguous orthologs.

At the bioinformatics level, the number of CNSs detected is limited by the BLAST parameters (the stringency of the local alignment). Although we define CNS size, position, and sequence based on our standard settings, it is always possible to obtain more or fewer "hits" by altering the settings (Kaplinsky et al. 2002), by allowing lower statistical significance, or by changing to a global alignment tool such as *Vista* (see Fig. 5 for a comparison of *Vista* and bl2seq results). Our program is not set to mask repetitive sequences, such as simple sequence repeats (SSRs) or the ends of transposable elements automatically, although it is possible to do so manually. We felt it best not to prejudge what sort of sequence might confer selective advantage.

Practical Considerations

In addition to the identification of pan-grass PCR binding sites (Kaplinsky et al. 2002), maize–rice genomic alignments should be valuable in annotating genes with hypothetical support or no support at all. The history of identifying human genes computationally using one species' sequence alone (Hogenesch et al. 2001) is not promising. In mammalian studies, CNS investigations led to the discovery of a gene implicated in human health in a region of DNA considered by genome annotators to be noncoding (Pennacchio et al. 2001). Even though we selected genes with small gene space, four of them included an unannotated gene or coding region. In all four cases, both maize and rice carried the expected new gene in the expected location based on the assumption of maize/rice microsynteny (Bennetzen 2000).

CNSs as Tools for the Study of Gene Evolution and the Origin of Novelty

Lewis (1951) proposed that, following gene or genome duplication, two previously identical genes occasionally might diverge in function, as one copy becomes free from the constraints of selection. Another explanation for the retention of duplicates suggests that duplicate genes each lose a different, specific *cis*-acting function such that the entire function is now encoded by two rather than one gene (Force et al. 1999; Lynch and Force 2000).

Maize is an excellent model system with which to test hypotheses about the consequences of duplication, with a well documented history of tetraploidization (Ahn and Tanksley 1993; Devos and Gale 2000) occurring about 11 Mya (Gaut and Doebley 1997). Rice has not undergone recent genome duplications. In maize, the MADS-box transcription factor genes *zag2* and *zmm1* are duplicates, as are *sh1* and *sus*, and both pairs of genes share most but not all of their CNSs. Such retained duplicates in maize, and especially those rich in CNSs, present opportunities to study the evolution of gene regulation, be it the origin of novelty via the Lewis scheme, or the fractionation of one ancestral complement of *cis*-acting sites into two subfunctionalized genes (Force et al. 1999). Although it be will in-

teresting to learn the molecular functions of CNSs in plants, much research on the consequences of gene duplication and the origin of novelty can proceed without knowing what CNSs bind or to what end.

METHODS

All source code for the computational analysis presented here is available upon request. Our blaster and viewer work well but presently require the use of the Unix command line.

CNS BLAST and Gene Annotation Viewer Software

We developed a two-part software package in order to identify and graphically display CNSs. This package offers an advantage relative to previous work (Kaplinsky et al. 2002) in allowing completely automated viewing of results, without the tedious step of manually drawing conserved regions onto a cartoon gene layout. It is important to keep in mind that this program is not designed for heuristically discovering genes, or discerning splice sites. The programming aspect of this goal is divided into two distinct parts, a Perl Script and a Java Applet.

Before running either program, one must first start with an annotated sequence (typically a maize sequence from GenBank). A cutoff of 3000 bases upstream and downstream of the coding region is used to define the noncoding sequence associated with a given maize gene. This annotated sequence is used to probe other available databases, here rice, via BLASTN, and to manually select the sequence, or sequences, considered strongest from the BLAST results.

At this point, user involvement ends and a Perl-based program takes these two sequences and begins analysis. Before beginning analysis, the program must identify the strand orientation of the compared genes. In all cases for this study, the maize sequence is in the plus orientation. The program analyzes the plus/plus and plus/minus matches in a bl2seq output. If the plus/minus matches carry more weight than the plus/plus matches, the program generates the reverse complement of the rice sequence, and reruns BLAST.

The next step is to establish a rough alignment of the two sequences beginning with the start of translation of the gene product. If both sequences are annotated, the program records the exon indices for both maize and rice and proceeds. However, if the rice sequence is not annotated, then the maize annotation must be used to predict the coding regions in the rice ortholog. Using the previously generated BLAST result, which should have created sequences in the same orientation, the start and end indices of the maize coding region are used to predict the rice boundaries. If the corresponding rice indices are found, then the program records the indices and continues. If only one index is returned, then the missing rice index is estimated by adding or subtracting the length of the coding region of maize.

If the genes are not similar enough to include matches that cover the start and end indices of the maize gene, then a simpler, less accurate secondary procedure is executed. This simpler alignment takes the first BLAST result hit representing a coding region in the maize gene, and discerns the distance of the maize sequence from the start and end indices of the maize gene in question. Using those distances, the program guesses where the rice coding region starts and stops. Although the last two techniques are much less exact, they provide a rough outline that helps in understanding the relative position of the retrieved CNSs.

The last step for the Perl program is the retrieval of CNSs. The program begins by substituting an 'n' for every coding nucleotide in the annotated sequence, thereby masking the coding regions from the subsequent bl2seq search. The bl2seq is rerun, returning all hits in noncoding regions. The program keeps only matches greater than 15 nucleotides in length in order to enrich the results for the most significant regions of similarity.

The second aspect of this software package is the viewer, which is designed to take the information generated above and create an easy to understand interface, as described earlier. The viewer is programmed in Java and utilizes the Swing package, available free as part of the Java™ Foundation Classes (JFC). The applet itself takes in a flat file and parses it; each gene is then represented as an Object containing Objects of coding regions, noncoding regions, and CNSs. Information is then retrieved from each Object by the user, primarily through mouse interaction.

Java was chosen for several reasons. It allows for interoperability across most platforms running the Java 2 Runtime Environment. Second, it easily and quickly permits user interaction. Finally, one can convert the applet into an application while conserving most of the original code, for use on auxiliary or private databases.

The initial window of the Gene Annotation Viewer presents a thumbnail view of information regarding the accession numbers of the compared sequences and provides an overview of the CNSs identified, their sequences, their orientation, and their cross-species nucleotide conservation. Most CNSs are indicated by a black line connecting orthologs, these sequences being in the +/+ orientation; red lines connect +/− alignments. A toggle switch allows the user to choose which CNSs will be displayed on the graphical view. The graphical viewer itself has a zoom function, and upon dragging across the gene schematic, displays the nucleotide sequence text desired. Exons, introns, and CNSs can be selected individually for text-based web applications such as BLASTN, BLASTX, and open reading frame finders.

In the Annotation Viewer, an annotated sequence is colored dark blue, and unannotated sequence is marked with a single light blue rectangle to indicate the predicted position of the corresponding gene's coding region, including introns. The software does not annotate the comparison sequence; therefore, it does not display exon structure of an unannotated input sequence.

Manual Adjustment of CNSs

Because our software returns all conserved regions above the parameter thresholds, we manually removed some conserved regions that clouded the larger data analysis. Conserved regions that we identified as probable coding exons by BLASTN or BLASTX (as described above) were removed from consideration as CNSs. Additionally, we removed conserved segments that were primarily dinucleotide simple sequences in order to remove noise from a gene graphic. Only occasionally (approximately 5% of genes) did genes have conserved simple sequence dinucleotides.

ACKNOWLEDGMENTS

We thank members of the Freeling Laboratory for contributions in the early stages of this project. Karen Osmont, Noriko Inada, Damon Lisch, Keith Slotkin, Maggie Woodhouse, George Theodoris, Stanley Lee, and Randall Tyers provided sequences or analyses toward our data set. We thank Virginia Walbot for comments on plant plasticity, Randall Tyers for critical review of the manuscript, and Nancy Nelson for her valuable assistance. We are grateful for the thoughtful comments of two anonymous reviewers. D.C.I. was supported by a National Science Foundation Graduate Research Fellowship. This work was supported by the Plant and Microbial Biology-Syngenta Collaborative Research Agreement, UC-Berkeley.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ahn, S. and Tanksley, S.D. 1993. Comparative linkage maps of rice and maize genomes. *Proc. Natl. Acad. Sci.* **90:** 7980–7984.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tools. *J. Mol. Biol.* **215:** 403–410.

Bennetzen, J.L. 2000. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* **12:** 1021–1029.

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Bolouri, H. and Davidson, E.H. 2002. Modeling DNA sequence-based *cis*-regulatory gene networks. *Dev. Biol.* **246:** 2–13.
Bucher, P. 1999. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9:** 400–407.
Colinas, J., Birnbaum, K., and Benfey, P. 2002. Using cauliflower to find conserved noncoding regions in *Arabidopsis*. *Plant Physiol.* **129:** 451–454.
Devos, K.M. and Gale, M.D. 2000. Genome relationships: The grass model in current research. *Plant Cell* **12:** 637–646.
Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., and Frazer, K.A. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10:** 1304–1306.
Ferl, R.J., Nick, H.S., and Laughner, B.H. 1987. Architecture of a plant promoter: S1 nuclease hypersensitive features of maize *Adh1*. *Plant Mol. Biol.* **8:** 299–307.
Force, A.M., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerate mutations. *Genetics* **151:** 1531–1545.
Freeling, M. 1992. A conceptual framework for maize leaf development. *Dev. Biol.* **153:** 44–58.
Freeling, M. and Bennett, D.C. 1985. Maize *Adh1*. *Annu. Rev. Genet.* **19:** 297–323.
Gau, H. and Moose, S.P. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15:** 1143–1158.
Gaut, B.S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci.* **94:** 6809–6814.
Goff, S.A., Ricke, D., Lan, T-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296:** 92–114.
Göttgens, B., Gilbert, J.G.R., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. 2001. Long-range comparison of human and mouse SCL loci: Localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res.* **11:** 87–97.
Greene, B., Walko, R., and Hake, S. 1994. Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. *Genetics* **138:** 1275–1285.
Gumucio, D.L., Shelton, D.A., Zhu, W., Millinoff, D., Gray, T., Bock, J.H., Slightom, J.L., and Goodman, M. 1996. Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. *Mol. Phylogenet. Evol.* **5:** 18–32.
Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.
Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.
Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.
Kaplinsky, N.J., Braun, D.M., Penterman, J., Goff, S.A., and Freeling, M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. *Proc. Nat. Acad. Sci.* **99:** 6147–6151.
Kellogg, E.A. 2001. Evolutionary history of the grasses. *Plant Physiol.* **125:** 1198–1205.
Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., and Rombauts, S. 2002. PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* **30:** 325–327.
Levy, S., Hannenhalli, S., and Workman, C. 2001. Enrichment of regulatory signals in conserved noncoding sequence. *Bioinformatics* **17:** 871–877.
Lewis, E.B. 1951. Pseudoallelism and gene evolution. *Cold Spring Harbor Symp. Quant. Biol.* **16:** 159–174.
Li, E. 2002. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat. Rev. Genet.* **3:** 662–673.
Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulatory of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.
Loots, G.G., Ovcharenki, I., Pachter, L., Dubchak, I., and Rubin, E. 2002. rVista for comparative sequence-based discovery of functional transcription factors binding sites. *Genome Res.* **12:** 832–839.
Lynch, M. and Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.
Nakano, R., Matsumura, T., Sakakibara, H., Sugiyama, T., and Hase, T. 1997. Cloning of maize ferredoxin III gene: Presence of a unique repetitive nucleotide sequence within an intron found in the 5′-untranslated region. *Plant Cell Physiol.* **38:** 1167–1170.
Paul, A.-L. and Ferl, R.J. 1991. In vivo footprinting reveals unique *cis*-elements and different modes of hypoxic induction in maize *Adh1* and *Adh2*. *The Plant Cell* **3:** 159–168.
Paul, A.-L. and Ferl, R.J. 1993. Osmium tetroxide footprinting of a scaffold attachment region in the maize *Adh1* promoter. *Plant Mol. Biol.* **22:** 1145–1151.
Paul, A.-L., Vasil, V., Vasil, I.K., and Ferl, R.J. 1987. Constitutive and anaerobically induced Dnase-I-hypersensitive sites in the 5′ region of the maize *Adh1* gene. *Proc. Nat. Acad. Sci.* **84:** 799–803.
Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J-C., Krauss, R.M., and Rubin, E.M. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169–173.
Quinn, J.P. 1996. Neuronal-specific gene expression—The interaction of both positive and negative transcriptional regulators. *Prog. Neurobiol.* **50:** 363–379.
Veit, B., Briggs, S.P., Schmidt, R.J., Yanofsky, M.F., and Hake, S. 1998. Regulation of leaf initiation by the terminal ear 1 gene of maize. *Nature* **393:** 166–168.
Vyas, P., Vickers, M.A., Picketts, D.J., and Higgs, D.R. 1995. Conservation of position and sequence of a novel, widely expressed gene containing the major human α-globin regulatory element. *Genomics* **29:** 679–689.
Walbot, V. 1985. On the life strategies of plants and animals. *Trends Genet.* **1:** 165–169.
Walbot, V. 1996. Sources and consequences of phenotypic and genotypic plasticity in plants. *Trends Plant Sci.* **1:** 27–32.
Walker, J., Howard, E., Dernnis, E., and Peacock, W. 1987. DNA sequences required for anaerobic expression of the maize *alcohol dehydrogenase 1* gene. *Proc. Nat. Acad. Sci.* **84:** 6624–6628.
Walsh, J., Waters, C.A., and Freeling, M. 1998. The maize gene liguleless2 encodes a basic leucine zipper protein involved in the establishment of the blade-sheath boundary. *Genes & Dev.* **12:** 208–218.
Woodman, J.C. and Freeling, M. 1981. Identification of a genetic element which controls the organ-specific expression of *adh*1 in maize. *Genetics* **98:** 357–378.
Yu, J., Hu, S., Wang, J., Wong, G-K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296:** 79–92.

WEB SITE REFERENCES

http://ncbi.nlm.nih.gov; GenBank
http://www.tmri.org; Torrey Mesa Research Institute, Syngenta Inc. rice genome portal.
http://oberon.rug.ac.be:8080/PlantCARE/index.html; PlantCARE (cis-acting regulatory elements; Lescot et al. 2002).
http://www.rgp.dna.affrc.go.jp; Rice Genome Project.
http://genomics.cnr.Berkeley.edu/cns/; User = reviewer; password = super (our 52 genes' graphics).

Received February 18, 2003; accepted in revised form June 16, 2003.