Arabidopsis intragenomic conserved noncoding sequence

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

PNAS published online Feb 14, 2007;
doi:10.1073/pnas.0611574104

Arabidopsis intragenomic conserved noncoding sequence

Brian C. Thomas*, Lakshmi Rapaka†, Eric Lyons†, Brent Pedersen*, and Michael Freeling†‡

*College of Natural Resources and †Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720

Contributed by Michael Freeling, December 28, 2006 (sent for review October 11, 2006)

After the most recent tetraploidy in the *Arabidopsis* lineage, most gene pairs lost one, but not both, of their duplicates. We manually inspected the 3,179 retained gene pairs and their surrounding gene space still present in the genome using a custom-made viewer application. The display of these pairs allowed us to define intragenic conserved noncoding sequences (CNSs), identify exon annotation errors, and discover potentially new genes. Using a strict algorithm to sort high-scoring pair sequences from the bl2seq data, we created a database of 14,944 intragenomic *Arabidopsis* CNSs. The mean CNS length is 31 bp, ranging from 15 to 285 bp. There are ≈1.7 CNSs associated with a typical gene, and *Arabidopsis* CNSs are found in all areas around exons, most frequently in the 5′ upstream region. Gene ontology classifications related to transcription, regulation, or "response to . . ." external or endogenous stimuli, especially hormones, tend to be significantly overrepresented among genes containing a large number of CNSs, whereas protein localization, transport, and metabolism are common among genes with no CNSs. There is a 1.5% overlap between these CNSs and the 218,982 putative RNAs in the *Arabidopsis* Small RNA Project database, allowing for two mismatches. These CNSs provide a unique set of noncoding sequences enriched for function. CNS function is implied by evolutionary conservation and independently supported because CNS-richness predicts regulatory gene ontology categories.

gene regulation | small RNA | transcription factor

Conserved noncoding sequences (CNSs) can offer insight into the evolution of gene regulation. CNSs are pairwise phylogenetic footprints in noncoding gene space and are useful when divergence is enough to ensure that conservation implies function, but not so much as to impair the detection of homology. Candidates for CNS function include matrix attachment regions (1, 2), transcription factor (TF) binding sites, and multiple TF binding sites (1, 3–14), chromosome-level regulatory regions (7), DNase I hypersensitive sites (15), and enhancers (such as *sonic hedgehog*; e.g., ref. 16). In the one case in which CNS function has been addressed in plants (homeobox gene *kn1* in grasses), intron CNSs bind a repressor that prevents ectopic expression (17).

Our interest in CNS is with plants and specifically *Arabidopsis thaliana* because the *Arabidopsis* genome is the most accurately annotated genome in plants. *Arabidopsis* had its most recent tetraploid ancestor sometime between 23 and 70 million years ago, and this duplication event has been analyzed by using several distinct methods (18–20). We chose to study the intragenomic footprints present in *Arabidopsis* (21). Presently, no other finished plant genome is diverged from *Arabidopsis* to an extent useful for CNS discovery; poplar is too distant and *Brassica* is too close. There have been multiple large segmental or whole-genome duplications in the *Arabidopsis* lineage (19, 20, 22–27).

Identifying a CNS begins by comparing two syntenic sequences (orthologs, homeologs, or other paralogs). Terms other than "CNS" are used for footprints where more than two syntenic sequences are compared, as is now common in vertebrates especially when ultra-conserved regulatory elements are being studied (11, 16, 28, 29). Once two sequences are aligned and evaluated for annotation errors, exons are masked, and the resulting alignments are in "noncoding" regions of sequence similarity. Accurate CNS identification is a visual process requiring a viewer to graphically display alignment results, to facilitate research on alignments, and to store CNS data.

When the 30,039 protein-coding *A. thaliana* genes in GenBank are minimized (by removing transposons and condensing local duplicates to one gene), 80% of the resulting 25,220-gene genome (30) is represented in syntenous chromosomal regions [ref. 19, refined in ref. 30; supporting information (SI) Table 1]. We show that comparisons of DNA sequence between these syntenic regions generate useful data. We used a special software tool to aid our genome investigation and graphically represent the syntenic stretches of the *Arabidopsis* genome, called the *Arabidopsis* bl2seq Viewer. A typical image generated from our viewer is seen in Fig. 1.

Technically speaking, we are measuring "alpha" CNSs [retained from the α tetraploidy (19)] between homeologs and not CNSs between orthologs. Duplicate genes within the same genome are under different selective pressures compared with orthologous genes in different genomes (31); subfunctionalized CNSs are expected between homeologs but not between orthologs. The database of the 14,944 *Arabidopsis* CNSs developed in this investigation is available in SI Table 2.

Results and Conclusions

Manual Inspection of Gene Pairs. Using our viewer, we manually annotated every identifiable gene pair retained from the *Arabidopsis* tetraploidy. We chose to include all local duplicates and any associated High Scoring Pair (HSP) in a single syntenous gene space. The typical case of local duplication is in tandem, with the duplicates being adjacent. However, our local regions also included reverse tandems and duplicates with one or two intervening genes, as indicated in notes frozen with our gene spaces. As seen in Fig. 1*A*, the CNS content of the query gene is sometimes duplicated and present syntenously near both of the subject tandem repeats. It is interesting to note that in this gene space diagram, the four CNSs have subfunctionalized, being present near one or the other of the duplicate genes on the subject (lower) gene.

Several of the viewer screenshots in Fig. 1 depict query genes pointed left-to-right in different 5′ to 3′ orientations than subject genes. This represents the order of the genes as they appear on their respective genomes, and by extension, the orientation of the Blast hit as either +/+ or +/−. Because it is important to us that all of our results may be readily replicated by using our online viewer application, we think it best to display the gene spaces in this manner, rather than to flip one or the other gene space.

Author contributions: B.C.T. and M.F. designed research; B.C.T., L.R., B.P., and M.F. performed research; B.C.T., E.L., and B.P. contributed new reagents/analytic tools; B.C.T. and M.F. analyzed data; and B.C.T. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: CNS, conserved noncoding sequence; GO, gene ontology; HSP, High Scoring Pair; NGCS, nongenic conserved sequence; smRNA, small RNA; TAIR, The *Arabidopsis* Information Resource; TF, transcription factor.

‡To whom correspondence should be addressed. E-mail: freeling@nature.berkeley.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0611574104/DC1.

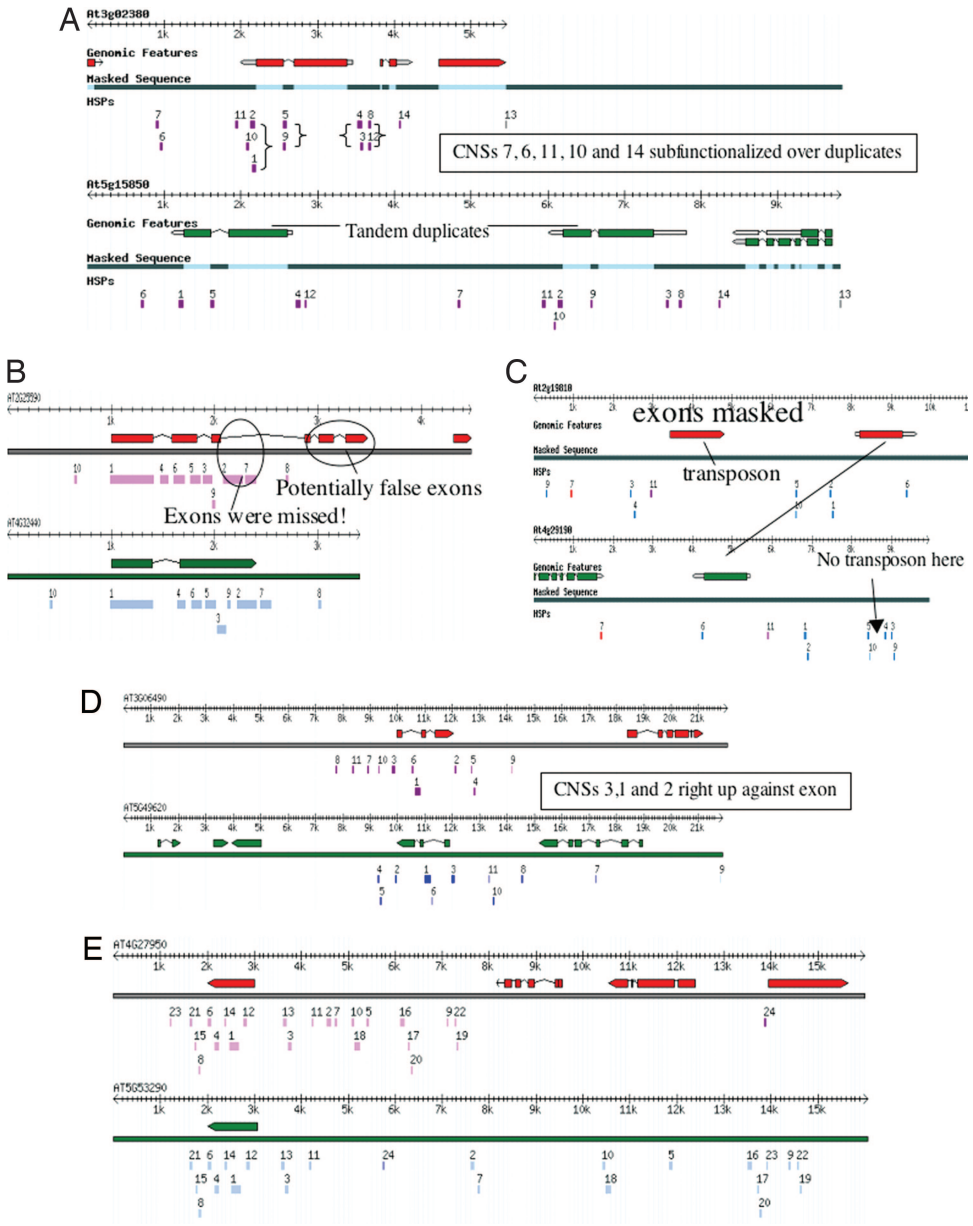


Fig. 1. Screenshots of gene space images generated from the *Arabidopsis* Bl2seq Viewer. In all images, red boxes represent the query gene, and green boxes are the subjects. Query genes are always drawn on top of the image. Purple and light-blue boxes represent CNSs, which are numbered as found in the bl2seq high-scoring pair report. (*A*) *At3g02380* vs. *At5g15850*: These two genes represent a local duplication (seen on the lower strand, *At5g15850*) where the CNS have subfunctionalized across the duplicated genes. (*B*) Gene model where an exon was missed in the query gene's annotation. (*C*) A known transposon in the promoter region of the query strand that is absent in the subject strand. (*D*) ''Appressed'' is a label applied to a CNS very close to a known exon. Here, three CNSs (on query: 3, 1, 2) lie very close to an exon. (*E*) A ''Bigfoot'' gene pair, showing CNSs spanning 5,000 bp (query) and 15,000 bp (subject) upstream of exon 1.

Annotation Issues. Approximately 10% of the gene pairs generated an HSP pattern implicating a probable annotation error in one or both homeologs. We attempted to correct these problems by temporarily changing our bl2seq settings from their default setting of −2 mismatch penalty to −1. This change results in larger HSPs, but with more mismatches, and helps to merge two HSPs split over a single exon. Other types of possible annotation errors included an exon present in the TAIR (The *Arabidopsis* Information Resource) annotation that lacked bl2seq support (Fig. 1*B*). The counterpart to this type of error, bl2seq support for an exon without any annotation, was noted as well (Fig. 1*B*). We observed instances where a string of HSPs from the bl2seq report between two syntenic regions matched a gene model perfectly but were not called as genes at all. Similarly, there were examples of HSP patterns that were obviously homeologs yet were called as two separate genes or as only a part of a hypothetical model. We note also that even though our investigation was done with assembly Version 5 of the TAIR annotation resource, we have checked our findings in the Version 6 annotation and these issues remain.

Transposons commonly insert between CNSs and the gene to which they associate; transposons were ignored in our analysis. Because many CNSs exist distal to such insertions, we conclude that the addition of a few kilobases of extra space, and whatever sequence lies within this space, need not remove CNS function. One such case involving insertion of a retrotransposon is shown in Fig. 1*C*, inserted between promoter CNSs 10 and 4.

Hypothetical Genes. There are 3,008 "hypothetical genes" in our proofed *Arabidopsis* genome (assembly Version 5.0; SI Table 1, or

GENETICS

search in the viewer by "hypothetical"). Seventy-seven hypothetical genes are retained: 0.025 retention frequency. Compare this with the 0.22 retention frequency for genes with an "average" GO classification of "molecular function unknown." One explanation for this large difference is that only 11.4% of hypothetical genes are real. Another explanation is that hypothetical genes are special or originated after the tetraploidy event. Changes from assembly Version 5 to Version 6 upgraded rather than removed hypothetical genes.

***Arabidopsis* CNS Database.** We used a hierarchical set of rules to correctly assign each bl2seg HSP to one gene (see *Methods*). The primary rule was based on proximity. Applying these rules resulted in a database containing 14,944 CNSs as 7,472 pairs (SI Table 2). The mean CNS is 30.7 bp in length, with a median of 24 and a range from 15–285. The mean number of CNSs per gene is 1.7; the mode is 0. Histograms displaying these data are shown in SI Fig. 3. None of the larger CNSs resulted in a significant ($e < 1.0$) Blastx score when searched against the entire *Viridiplantae* GenBank dataset (see *Methods*). Nevertheless, the larger CNSs make excellent candidates for unannotated exons or exons of unannotated genes.

CNS Characteristics and Gene Association. CNSs from *Arabidopsis* defined in the CNS database have a mean %AT composition of 65.25 ± 12.7. This percentage is similar to the mean for intergenic regions of 67.1% (Genome Indices 8/04: http://gi.kuicr.kyoto-u.ac.jp). GC content, CpG content, and CpNpG content are all similar to known values for similar gene regions in *Arabidopsis* (SI Table 2).

We searched each CNS for an overrepresentation of simple sequence repeats. Simple sequence repeat motifs are not found in the majority of CNSs in the database; typically <1% for any given simple sequence repeat (data not shown).

Some categories of genes have larger or smaller numbers of CNSs. We grouped all genes by their CNS count and then compared the gene ontology (GO) terms associated within each group. SI Table 3 shows that the group of genes with 0 CNSs is dominated by terms related to "ribosome," "protein metabolism," "localization," and "protein transport": the general theme inferring housekeeping and basal metabolic processes. Fig. 2 displays the gene groups with one or more CNSs. The red bars indicate significant overrepresentation of a GO term in genes with increasing numbers of CNSs. The legend explains the numbers embedded in this figure.

GO terms related to "nucleotide binding," "kinase activity," "chromatin," and "nucleosome" appear with genes with at least one CNS. At the high end of the list, GO terms associated with genes containing 14 CNSs are associated with "response" events, either to environmental stress ("endogenous stimulus," "osmotic stress," "salt stress") or to metabolic/pathogenic stress ("jasmonic acid," "salicylic acid," "endogenous stimulus"). The highest CNS count with a GO term significantly overrepresented at the $P \leq 0.001$ level is 18 CNSs: "response to auxin stimulus." Genes with modest levels of CNS-richness are annotated with GO terms involving signal transduction (Fig. 2).

Note the group of genes containing 4–14 CNSs in Fig. 2. These genes share a set of GO terms heavily biased toward "transcription" and "regulation." For comparison, we analyzed the CNS-richness of 44 MIR genes within 18 gene spaces for CNS-richness. The average number of CNSs/MIR gene space is 4.6, which is similar to the mean 4.5 CNSs per gene associated with GO: "transcription factor activity."

The biological process "response to . . ." terms are of unique significance. Investigating our 588 most CNS-rich genes (CNS count per gene), we obtained a list of 39 genes with the GO term "response to biotic stimuli" (GO:0009628). We found that 62% of these genes are also annotated as TF genes (GO:0003700).

Among the 39 "response to . . ." genes, all 5 growth hormones (29 genes with GO:0009725) but cytokinin were represented as specific stimuli: 16 genes for auxin, 10 for ethylene, 7 for ABA, and 6 for GA. GO:0009605, "response to external stress," carried 11 genes, and among these included 9–11 genes each representing response to the specific agents wounding, salt, pathogens, salacylic acid, and jasmonic acid.

CNS Distribution Around *Arabidopsis* Genes. We identified 4,208 (omitting local duplicates and genes in more than one space) genes containing UTR annotation for both ends of the gene. This set of genes contained a total of 9,778 CNSs. Having detailed annotation for these genes allowed us to sort the CNSs into five non-protein-coding regions: 5′, 5′ UTR, intron (within CDS regions), 3′ UTR, and 3′ (SI Table 2). 237 CNSs spanned the boundary between 5′ and 5′ UTR, and 29 CNSs spanned 3′ UTR and 3′; these were divided equally between the two contending regions for the count. The summary of the distribution of CNSs around an *Arabidopsis* gene is 5′ to intron to 3′ is 2.3:0.7:1. It is apparent that CNSs exist in the 5′ region of a gene 2.3 times more often than in the 3′.

Occasionally (in 9.5% of our pairs), we found an HSP much larger than a nearby exon in the gene space. If the HSP remains after masking out exons and rerunning the bl2seq comparison, we annotated the HSP as "appressed." If the HSP scores high in a Blastx search against all plant proteins, then we classify it as an exon and remove it from the CNS database. Some mammalian genes with splice variants have CNSs conserved next to alternatively spliced segments (32), so our list of appressed CNSs could prove useful for further study.

We found 126 gene pairs that had CNSs spread over a much larger region of the genome than an average gene pair. These big footprint ("Bigfoot") genes were labeled as such if they spanned at least 4 kb of chromosome 5′ plus 3′ of exon (e.g., Fig. 1*E*).

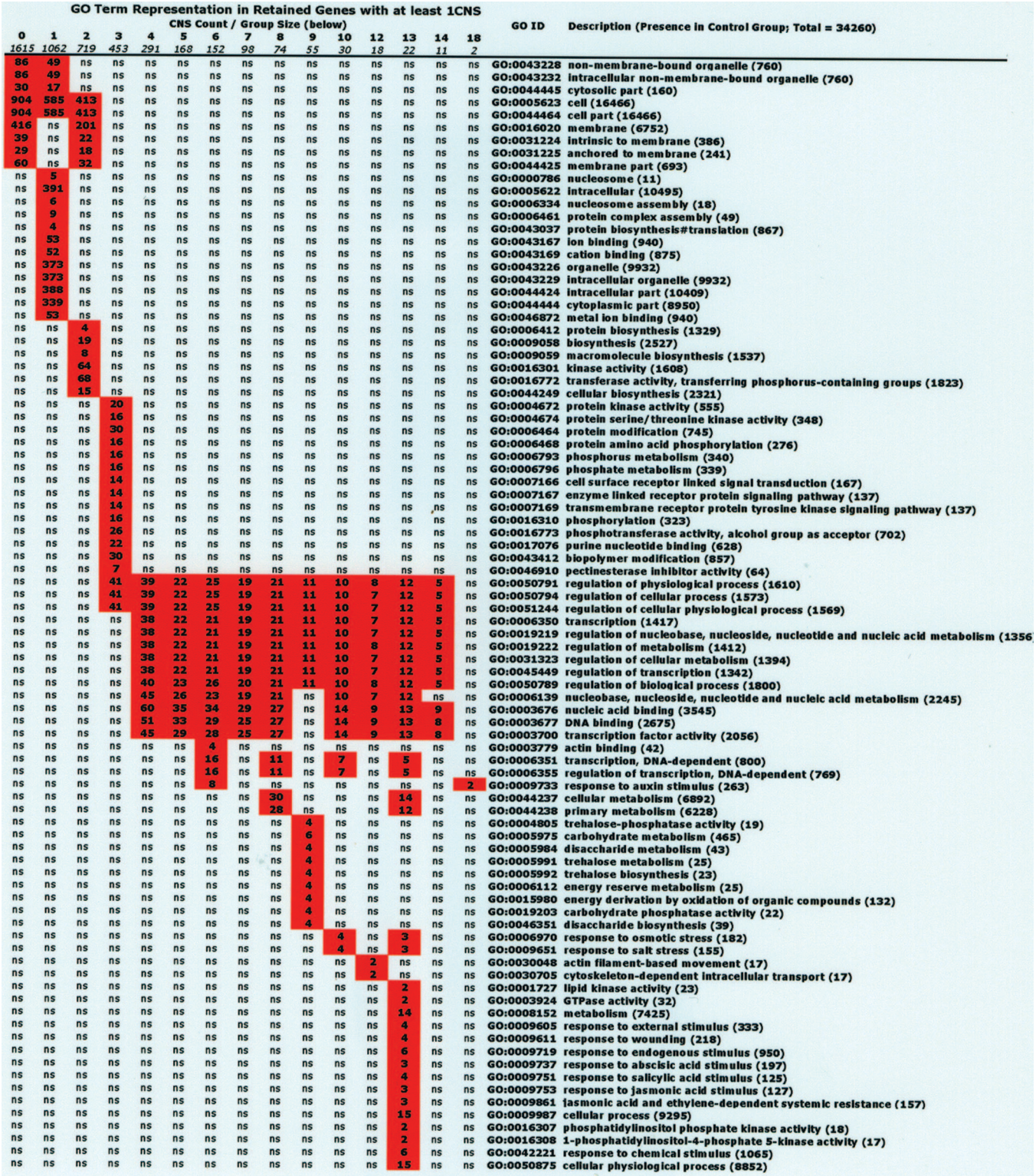
Very occasionally, we found sequences that are paired, syntenous, seem unlikely to code for a protein, and also do not seem to be associated with any gene in cis. Often, such sequences have an over-simple structure, and queries using Blastn (under conditions favoring distant homologous hits) find hundreds of such hits in *Arabidopsis* at over 80% nucleotide identity and coverage. These are annotated with the keyword "NGCS" (nongenic conserved sequence) to make them easy to recognize, and these HSPs were generously included in the CNS database.

Comparing the CNSs to *Arabidopsis* Small RNAs (smRNAs). We searched the database of 218,982 (206,077 unique) smRNA sequences from the *Arabidopsis thaliana* Small RNA Project (September 2006; http://asrp.cgrb.oregonstate.edu) against a partial CNS database composed of the 10,826 CNS ≥ 19 bp. We allowed up to two mismatches or gaps. Each of the 198 hits was proofed manually, and 146 were validated. Those removed were unannotated, repetitive sequence (NGCS, uniformly hit by many smRNAs many times), and also known transposons and RNA genes populating our CNS database in error. These invalidated hits are listed in SI Table 2 with an explanatory note. We found that of these CNS ≥ 19 bp, only 1.3% matched (zero to two mismatches) a smRNA sequence. Using CNS ≥ 21 bp increased the percentage to 1.5%. We conclude that, with caveats, the typical CNS function is unlikely to involve either the encoding or the binding of RNAs.

The 146 CNSs that do match a smRNA had the following 5′ to intron to 3′ ratio: 39:23:56 or 0.7:0.4:1. This 3′ bias is different from the 2.3:0.7:1 distribution of all CNSs. This 3′ skew is so striking that we conclude that "many" of these 146 potential regulatory smRNA binding sites actually function. Nevertheless, smRNA involvement in CNS function is rare.

Discussion

Approximately 25% of the genes in an *Arabidopsis* genome (after minimizing as described in *Methods*) have a pair retained following the most recent (α) tetraploidy. Therefore, we do not capture all or even the majority of CNSs in *Arabidopsis* in a way that would be



GO Term Representation in Retained Genes with at least 1CNS

CNS Count / Group Size (below)															GO ID	Description (Presence in Control Group; Total = 34260)
0	1	2	3	4	5	6	7	8	9	10	12	13	14	18		
1615	1062	719	453	291	168	152	98	74	55	30	18	22	11	2		
86	49	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043228	non-membrane-bound organelle (760)
86	49	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043232	intracellular non-membrane-bound organelle (760)
30	17	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044445	cytosolic part (160)
904	585	413	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0005623	cell (16466)
904	585	413	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044464	cell part (16466)
416	ns	201	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0016020	membrane (6752)
39	ns	22	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0031224	intrinsic to membrane (386)
29	ns	18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0031225	anchored to membrane (241)
60	ns	32	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044425	membrane part (693)
ns	5	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0000786	nucleosome (11)
ns	391	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0005622	intracellular (10495)
ns	6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006334	nucleosome assembly (18)
ns	9	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006461	protein complex assembly (49)
ns	4	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043037	protein biosynthesis#translation (867)
ns	53	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043167	ion binding (940)
ns	52	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043169	cation binding (875)
ns	373	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043226	organelle (9932)
ns	373	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043229	intracellular organelle (9932)
ns	388	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044424	intracellular part (10409)
ns	339	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044444	cytoplasmic part (8950)
ns	53	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0046872	metal ion binding (940)
ns	ns	4	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006412	protein biosynthesis (1329)
ns	ns	19	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0009058	biosynthesis (2527)
ns	ns	8	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0009059	macromolecule biosynthesis (1537)
ns	ns	64	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0016301	kinase activity (1608)
ns	ns	68	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0016772	transferase activity, transferring phosphorus-containing groups (1823)
ns	ns	15	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0044249	cellular biosynthesis (2321)
ns	ns	ns	20	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0004672	protein kinase activity (555)
ns	ns	ns	16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0004674	protein serine/threonine kinase activity (348)
ns	ns	ns	30	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006464	protein modification (745)
ns	ns	ns	16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006468	protein amino acid phosphorylation (276)
ns	ns	ns	16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006793	phosphorus metabolism (340)
ns	ns	ns	16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0006796	phosphate metabolism (339)
ns	ns	ns	14	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0007166	cell surface receptor linked signal transduction (167)
ns	ns	ns	14	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0007167	enzyme linked receptor protein signaling pathway (137)
ns	ns	ns	14	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway (137)
ns	ns	ns	16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0016310	phosphorylation (323)
ns	ns	ns	26	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0016773	phosphotransferase activity, alcohol group as acceptor (702)
ns	ns	ns	22	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0017076	purine nucleotide binding (628)
ns	ns	ns	30	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0043412	biopolymer modification (857)
ns	ns	ns	7	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	GO:0046910	pectinesterase inhibitor activity (64)
ns	ns	ns	41	39	22	25	19	21	11	10	8	12	5	ns	GO:0050791	regulation of physiological process (1610)
ns	ns	ns	41	39	22	25	19	21	11	10	7	12	5	ns	GO:0050794	regulation of cellular process (1573)
ns	ns	ns	41	39	22	25	19	21	11	10	7	12	5	ns	GO:0051244	regulation of cellular physiological process (1569)
ns	ns	ns	ns	38	22	21	19	21	11	10	7	12	5	ns	GO:0006350	transcription (1417)
ns	ns	ns	ns	38	22	21	19	21	11	10	7	12	5	ns	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism (1356
ns	ns	ns	ns	38	22	21	19	21	11	10	8	12	5	ns	GO:0019222	regulation of metabolism (1412)
ns	ns	ns	ns	38	22	21	19	21	11	10	7	12	5	ns	GO:0031323	regulation of cellular metabolism (1394)
ns	ns	ns	ns	38	22	21	19	21	11	10	7	12	5	ns	GO:0045449	regulation of transcription (1342)
ns	ns	ns	ns	40	23	26	20	21	11	10	8	12	5	ns	GO:0050789	regulation of biological process (1800)
ns	ns	ns	ns	45	26	23	19	21	ns	10	7	12	ns	ns	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism (2245)
ns	ns	ns	ns	60	35	34	29	27	ns	14	9	13	9	ns	GO:0003676	nucleic acid binding (3545)
ns	ns	ns	ns	51	33	29	25	27	ns	14	9	13	8	ns	GO:0003677	DNA binding (2675)
ns	ns	ns	ns	45	29	28	25	27	ns	14	9	13	8	ns	GO:0003700	transcription factor activity (2056)
ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	ns	ns	ns	GO:0003779	actin binding (42)
ns	ns	ns	ns	ns	ns	16	ns	11	ns	7	ns	5	ns	ns	GO:0006351	transcription, DNA-dependent (800)
ns	ns	ns	ns	ns	ns	16	ns	11	ns	7	ns	5	ns	ns	GO:0006355	regulation of transcription, DNA-dependent (769)
ns	ns	ns	ns	ns	ns	8	ns	ns	ns	ns	ns	ns	ns	2	GO:0009733	response to auxin stimulus (263)
ns	ns	ns	ns	ns	ns	ns	ns	30	ns	ns	ns	14	ns	ns	GO:0044237	cellular metabolism (6892)
ns	ns	ns	ns	ns	ns	ns	ns	28	ns	ns	ns	12	ns	ns	GO:0044238	primary metabolism (6228)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0004805	trehalose-phosphatase activity (19)
ns	ns	ns	ns	ns	ns	ns	ns	ns	6	ns	ns	ns	ns	ns	GO:0005975	carbohydrate metabolism (465)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0005984	disaccharide metabolism (43)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0005991	trehalose metabolism (25)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0005992	trehalose biosynthesis (23)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0006112	energy reserve metabolism (25)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0015980	energy derivation by oxidation of organic compounds (132)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0019203	carbohydrate phosphatase activity (22)
ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	ns	ns	ns	GO:0046351	disaccharide biosynthesis (39)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	3	ns	ns	GO:0006970	response to osmotic stress (182)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	3	ns	ns	GO:0009651	response to salt stress (155)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	ns	GO:0030048	actin filament-based movement (17)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	ns	GO:0030705	cytoskeleton-dependent intracellular transport (17)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	GO:0001727	lipid kinase activity (23)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	GO:0003924	GTPase activity (32)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	14	ns	ns	GO:0008152	metabolism (7425)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	GO:0009605	response to external stimulus (333)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	GO:0009611	response to wounding (218)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	6	ns	ns	GO:0009719	response to endogenous stimulus (950)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	3	ns	ns	GO:0009737	response to abscisic acid stimulus (197)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	4	ns	ns	GO:0009751	response to salicylic acid stimulus (125)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	3	ns	ns	GO:0009753	response to jasmonic acid stimulus (127)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	3	ns	ns	GO:0009861	jasmonic acid and ethylene-dependent systemic resistance (157)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	15	ns	ns	GO:0009987	cellular process (9295)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	GO:0016307	phosphatidylinositol phosphate kinase activity (18)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	2	ns	ns	GO:0016308	1-phosphatidylinositol-4-phosphate 5-kinase activity (17)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	6	ns	ns	GO:0042221	response to chemical stimulus (1065)
ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	15	ns	ns	GO:0050875	cellular physiological process (8852)

Fig. 2. GO term representation in retained genes with ≥1 CNS. ''CNS Count'' refers to the set of genes each with the indicated number of CNSs. ''Group Size'' indicates how many of the genes in the set had any GO annotation and is used in calculating significance (see the GOstat web site, http://gostat.wehi.edu.au). The control for every gene group was the TAIR database of 34,620 genes. Numbers in parentheses after the GO description are total genes so annotated per genome. The numbers in the cells of the figure itself represent the number of genes representing each term. Red is used to highlight the trend in the data. ''ns'' is used when the listed GO ID may have been present in the group of genes but was not more significant than our cutoff ($P = 0.001$).

possible were the comparison between orthologous genes of *Arabidopsis* and a usefully diverged relative (the *Brassicaceae* equivalent of man–mouse or maize–rice). Because the genes retained following tetraploidy in *Arabidopsis* are not expected to be a random sampling of ancestral genes (20, 33–35), the CNSs in the database also cannot be a random sampling. Additionally, our CNS database

is incomplete because genes that are duplicated can subfunctionalize cis-acting regulatory sequences (36). Subfunctionalized CNSs are not present in this analysis because a useful out-group is needed to resolve them. The generalized result for all eukaryotes is that duplicates diverge, sometimes rapidly, although it is usually difficult to clearly differentiate subfunctionalization from gain-of-function (21, 37–47).

The *Arabidopsis* bl2seq Viewer facilitates the use of synteny in improving the model annotation of those genes retained as pairs, as well as the comparison of any region of any length with any other stretch of chromosome. Most dramatically, if a gene of interest is poorly annotated but its pair is well annotated, the gene of interest's annotation is thus increased. Hundreds of paired genes have markedly different models and/or inexplicably different GO annotations, and most may be corrected by applying the annotations of the better-understood gene onto the lesser-understood gene. There are dozens of examples where known TF genes are paired with genes not annotated as TF genes or to an anonymous sequence.

The most important result of these studies is that CNS-richness predicts genes that contain the GO term "transcription factor activity" and, as CNS-richness increases even more, "response to . . ." GO terms. We show that genes annotated with a "response to . . ." GO term are simultaneously annotated as a TF gene 62% of the time. GO terms associated with signal transduction populated the middle regions of CNS-richness. Genes with zero CNSs tended to be household and/or metabolic genes (Fig. 2). It is of particular interest that those genes highest in the regulatory cascade, "response to . . ." or first-responder genes, are themselves covered with CNSs (a CNS presumably being a site where exogenous regulatory molecules bind the gene space). In other words, the highest-level regulatory genes tend to be, themselves, most highly regulated. This "enigma" does make sense in a scheme where the targets of transcriptional regulation feed back to the regulators via a systemic regulatory pathway.

Inada and coworkers (17), studying maize–rice CNSs, noticed that genes with upstream regulatory functions (mostly TF genes) had an average of 9 CNSs per gene, whereas the average gene had only 2.4 CNSs per gene. In vertebrates, there are ≈1,400 noncoding sequences conserved in all vertebrates from fish to man, these being among the most conserved of man–mouse CNSs and marking particularly CNS-rich genes. Most or all of these are enhancers of developmental regulatory genes (29). Thus, our result that CNS-richness is positively correlated with transcription factor activity (and even more so with GO terms involving "response to . . ." stimuli of all sorts, these describing genes that are annotated TF genes 62% of the time) fits a general rule that may apply to plants and animals alike.

Recently, there has been a burst of new information on the importance of smRNAs [micro RNAs (miRNAs) and, in specific cases, siRNAs] in developmental gene regulation, in addition to the better understood involvement of siRNA in silencing of repetitive elements (48–51). There are 146 CNSs that could possibly bind smRNAs, and these are distributed far more 3′ in the gene space than the norm. These few reflect only the 1.5% of CNSs that were hit with zero to two mismatches/gaps by one or more smRNA in the massive *Arabidopsis thaliana* Small RNA Project database. Our data do not support the hypothesis that CNSs are smRNA targets or that CNSs mark new RNA-encoding genes.

For maize–rice, the modal gene had 0 CNSs and on average, a gene had 2.4 CNSs (17). The modal *Arabidopsis* gene also has 0 intragenic CNSs, and there is an average of 1.7 intragenic CNSs per gene. As mentioned in the Introduction, CNSs and intragenic (α) CNSs measured here are not identical. That said, the mean number of CNSs per gene, 2.7 and 1.7, are in the same broad range. Either of these frequencies are far smaller than man–mouse CNS content where almost all genes have some CNSs, and most have so many that are so long (covering approximately half of the noncoding gene space) that individual gene spaces overlap into a continuum of conservation (52, 53). *Arabidopsis–Arabidopsis*, man–mouse, and maize–rice all have exons that have diverged to approximately the same extent.

The CNS database is not a comprehensive sampling. A few very large, very CNS-rich gene spaces dominate the CNS list as a whole. We noticed the extremes of these genes in the viewer, and they are typically TF genes surrounded by a low-exon-density void, a void often filled with several CNSs. Fig. 1*E* shows such a gene. If the gene space extended 4 kb beyond the exons either 5′ or 3′, we noted it as "Bigfoot," to denote the large footprint defined by this gene space. These 252 Bigfoot genes (see the column labeled "BF" in SI Table 1) are a unique contribution to *Arabidopsis* gene annotation and deserve further study.

The *Arabidopsis* CNS database described here provides a unique set of noncoding sequences enriched for function. Because smRNA involvement is rare, CNSs probably bind protein. CNS function is implied by evolutionary conservation and is supported by significant correlation of CNS-richness of a gene and its associated GO category annotations.

Methods

The *Arabidopsis* bl2seq Viewer. The *Arabidopsis* bl2seq Viewer (http://synteny.cnr.berkeley.edu/AtCNS) (hereafter "the viewer") is a web application whose primary function is to visualize the output from bl2seq (54). Source code is available.

Retained Pairs List and Defining Syntenic Regions. We manually inspected each of the 3,179 gene pairs as described (30) and 40–200 kb around the pair in our viewer. We arbitrarily set the gene space boundaries to include all exons, introns, and CNS. Locally duplicated arrays of genes were included in one gene space if present. SI Table 1 is our gene list, which includes the additional retained sequence pairs we discovered during manual inspection of gene space in the *Arabidopsis* genome and also the known *MIR* genes from Rfam (http://microrna.sanger.ac.uk). During annotation of every gene pair, entries were made in our database to indicate particularly common or interesting gene space configurations. The terms are as follows: "DUPLICATE GENES IN SPACE" indicates locally duplicated (usually tandem) genes; "ANNOTATION ISSUE" indicates one or both genes of the pair have an annotation inconsistency; "DUPLICATE EXON/HSP BITS IN SPACE" indicates regions where sequence has been duplicated (HSP refers to "High Scoring Pair" from the bl2seq report); "APRESSED CNS" indicates that a putative CNS is very close to an exon; "NGCS" denotes a nongenic conserved sequence, as explained in *Results*. Each NGCS is given a fake gene location number followed by an "*_oa*" for "our additional" (e.g., *At5g45614_oa*) and are listed along with typical genes in SI Table 1.

Defining CNS in a Gene Space and the *Arabidopsis* CNS Database. HSPs (High Scoring Pairs and, at this stage, putative CNSs) were assigned to a gene space by using the following hierarchical rule set:

1. HSPs are assigned to the closest gene on the homeolog.
2. HSPs separated from a retained gene by more than 2 genes, not including hypothetical genes, must belong to another retained gene or else become candidates for a rare NGCS.
3. An HSP approximately (±2 kb) midway between two retained genes on both homeologs is assigned to the retained gene with the most undisputed HSPs already assigned (and we add a note because this situation is rare). If this sorting rule cannot be applied (no current HSPs assigned), then proceed to rule 4.
4. An HSP in the 5′ region of a gene is preferred over one in the 3′ region.

At this point, the gene space is "frozen" and the remaining HSPs are added to the list of "Putative CNSs." An HSP becomes a CNS after a second round of manual and automated inspection. During

this round, HSPs of any length on the incorrect strand or not syntenic were invalidated and removed from the list if including them would increase the length of the gene space. Simplicity did not invalidate HSPs. Any HSP >24 bp was further proofed and, if found to be located close to an exon or over 50 bp in length, compared with the *Viridiplantae* protein database (www.ncbi.nlm.nih.gov/BLAST) using Blastx. Any hit with an e-value more significant than 1.0 was inspected to determine whether a small gene or exon was possible. Additionally, each HSP suspected of being exon or RNA-coding was used as a Blastn query to *Arabidopsis* sequence at the European *Arabidopsis* Stock Center's BLASTView (an Ensemble project at http://atensembl.arabidopsis.info/Multi/blastview?species=Arabidopsis_thaliana) as well as the *Arabidopsis* Tiling Array Transcriptome Express Tool (ref. 55; http://signal.salk.edu/cgi-bin/atta). Any high-scoring result from these comparisons was noted in our database. An invalidated HSP does not show up on our CNS list (SI Table 2), with the exception of those HSPs invalidated in the process of this research, such as repetitive sequences hit by smRNAs; these are invalidated by turning them red in our viewer and adding a note.

When the associated gene had sufficient annotation, we classified CNS locations into 5′, 5′ UTR, intron, 3′ UTR, and 3′, and recorded the data in SI Table 2. The term "appressed" was used to indicate a CNS immediately juxtaposed to an exon (usually the 5′ or the 3′ terminal exon). There were 4,208 genes that contained UTR annotation and so the more exact locations could only be assessed on 9,778 CNSs from the database.

CNS with Genes by GO Category. Genes were categorized by GO terms from the GenBank annotation file (TAIR 6–05). Except for MIR genes, genes encoding RNA were not counted in this study, although all GenBank genes appeared on our viewer as an aid to CNS annotation. As explained, we sometimes found a gene that was lacking annotation or was vaguely annotated ("hypothetical" or "expressed protein"). We did not duplicate the GO annotation for a gene in a retained pair lacking GO annotation using information from the partner. Our analysis did not find new miRNA-encoding genes except as additional duplicates in gene spaces (i.e., no new MIR gene spaces were identified).

We grouped genes by their total number of CNSs and created a histogram using the R statistical analysis software package (www.r-project.org). We used these bin sizes to create a list of TIGR gene identifiers, which were then submitted to the application GOstat (56) to determine whether any GO terms associated with the gene list were significantly overrepresented. Each group of genes was compared against the control GOstat database TAIR, which represents the entire *Arabidopsis* genome (34,260 genes). We filtered this result using a significance cutoff of $P \leq 0.001$, and did not select to cluster the results (Cluster = −1). We corrected for multiple testing using the false discovery method (Benjamini). Each bin of genes corresponding to CNS count for the group was submitted separately to GOstat, and the results were collated to produce Fig. 2 and SI Table 3. GO terms were sorted by their appearance in a bin. We also used GOstat output to address questions of GO term gene overlap, again without clustering the results.

Nucleic Acid Secondary Structure. To determine whether CNS entries in the database could encode an RNA, or fold as a single-stranded DNA, with a significant secondary structure, we submitted each CNS to the M-Fold (57). We used settings appropriate for folding DNA sequence (NA = DNA). The calculated negative minimum free energy for each CNS is listed in SI Table 2 next to each CNS.

NGCS. Occasionally, a larger HSP or a cluster of HSPs exists between homeologs and is present in strict synteny in relation to adjacent genes. However, the sequence of these NGCS is clearly simpler than that found in exons and usually found in many copies throughout the genome. These NGCS are included as CNSs (see above), although some are likely to be transposons positioned syntenously by chance alone, as evidenced by being highly repetitive and the targets of siRNAs (SI Table 2).

We thank Damon Lisch for discussions. The College of Natural Resources, University of California, Berkeley, partially subsidized the Statistics and Bioinformatics Consulting Service. This work was supported by National Science Foundation Grant DBI-034937 (to M.F.).

1. Avramova Z, Tikhonov A, Chen M, Bennetzen JL (1998) *Nucleic Acids Res* 26:761–767.
2. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA (2003) *Trends Genet* 19:119–124.
3. Loots GG, Ovcharenki I, Pachter L, Rubin E (2002) *Genome Res* 12:832–839.
4. Dubchak I, Frazer K (2003) *Genome Biol* 4:122.
5. Hardison RC (2003) *PLoS Biol* 1:E58.
6. Hardison RC (2000) *Trends Genet* 16:369–372.
7. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) *Science* 288:136–140.
8. Loots GG, Ovcharenko I (2004) *Nucleic Acids Res* 32:W217–W221.
9. Levy S, Hannenhalli S, Workman C (2001) *Bioinformatics* 17:871–877.
10. Bejerano G, Siepel AC, Kent WJ, Haussler D (2005) *Nat Methods* 2:535–545.
11. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, *et al.* (2005) *Genome Res* 15:1034–1050.
12. Siepel A, Haussler D (2004) *J Comput Biol* 11:413–428.
13. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al.* (2003) *Nature* 424:788–793.
14. Sobral BW, Mangalam H, Siepel A, Mendes P, Pecherer R, McLaren G (2001) *Novartis Found Symp* 236:59–81, discussion 81–84.
15. Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J, Bentley DR, Green AR (2001) *Genome Res* 11:87–97.
16. Goode DK, Snell P, Smith SF, Cooke JE, Elgar G (2005) *Genomics* 86:172–181.
17. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, Freeling M (2003) *Genome Res* 13:2030–2041.
18. Guyer D, Tuttle A, Rouse S, Volrath S, Johnson M, Potter S, Gorlach J, Goff S, Crossland L, Ward E (1998) *Genetics* 149:633–639.
19. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) *Nature* 422:433–438.
20. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) *Proc Natl Acad Sci USA* 102:5454–5459.
21. Haberer G, Hindemitt T, Meyers BC, Mayer KF (2004) *Plant Physiol* 136:3009–3022.
22. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) *Plant Cell* 12:1093–1101.
23. Blanc G, Hokamp K, Wolfe KH (2003) *Genome Res* 13:137–144.
24. Blanc G, Wolfe KH (2004) *Plant Cell* 16:1667–1678.
25. Vision TJ, Brown DG, Tanksley SD (2000) *Science* 290:2114–2117.
26. Kowalski S, Lan TH, Feldmn K, Paterson A (1994) *Genetics* 138:499–510.
27. Patterson A, Lan T-H, Reischmann K, Chang C, Lin Y, Liu S, Burow M, Kowalski S, Kastar C, DelMonte T, *et al.* (1996) *Nat Genet* 14:380–382.
28. Hughes JR, Cheng J-F, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR (2005) *Proc Natl Acad Sci USA* 102:9830–9835.
29. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, *et al.* (2005) *PLoS Biol* 3:e7.
30. Thomas BC, Pederson B, Freeling M (2006) *Genome Res* 16:934–946.
31. Koonin EV (2005) *Annu Rev Genet* 39:309–338.
32. Sorek R, Ast G (2003) *Genome Res* 13:1631–1637.
33. Seoighe C, Gehring C (2004) *Trends Genet* 20:461–464.
34. Blanc G, Wolfe KH (2004) *Plant Cell* 16:1679–1691.
35. Freeling M, Thomas BC (2006) *Genome Res* 16:805–814.
36. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) *Genetics* 151:1531–1545.
37. Gu Z, Nicolae D, Lu HH, Li WH (2002) *Trends Genet* 18:609–613.
38. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH (2002) *Mol Biol Evol* 19:256–262.
39. Makova KD, Li W-H (2003) *Genome Res* 13:1638–1645.
40. Raes J, Van de Peer Y (2003) *Appl Bioinf* 2:91–101.
41. Wagner A (2002) *Mol Biol Evol* 19:1760–1768.
42. Causier B, Castillo R, Zhou J, Ingram R, Xue Y, Schwarz-Sommer Z, Davies B (2005) *Curr Biol* 15:1508–1512.
43. Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) *Mol Biol Evol* 23:469–478.
44. Li WH, Yang J, Gu Z (2005) *Trends Genet* 21:1–6.
45. Rastogi S, Liberles DA (2005) *BMC Evol Biol* 5:28.
46. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA (2007) *J Exp Zool B Mol Dev Evol* 308:58–73.
47. Gu Z, Rifkin SA, White KP, Li WH (2004) *Nat Genet* 36:577–579.
48. Allen E, Xie Z, Gustafson AM, Carrington JC (2005) *Cell* 121:207–221.
49. Axtell MJ, Bartel DP (2005) *Plant Cell* 17:1658–1673.
50. Bartel B (2005) *Nat Struct Mol Biol* 12:569–571.
51. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) *Genes Dev* 18:2368–2379.
52. Jareborg N, Birney E, Durbin R (1999) *Genome Res* 9:815–824.
53. Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M (2002) *Proc Natl Acad Sci USA* 99:6147–6151.
54. Tatusova TA, Madden TL (1999) *FEMS Microbiol Lett* 174:247–250.
55. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HW, Kim C, Nguyen M, *et al.* (2003) *Science* 302:842–846.
56. Beissbarth T, Speed T (2004) *Bioinformatics* 1:1–2.
57. Zuker M (2003) *Nucleic Acids Res* 31:3406–3415.