Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Plant Biology

Conserved noncoding sequences (CNSs) in higher plants

Michael Freeling and Shabarinath Subramaniam

Plant conserved noncoding sequences (CNSs)—a specific category of phylogenetic footprint—have been shown experimentally to function. No plant CNS is conserved to the extent that ultraconserved noncoding sequences are conserved in vertebrates. Plant CNSs are enriched in known transcription factor or other *cis*-acting binding sites, and are usually clustered around genes. Genes that encode transcription factors and/or those that respond to stimuli are particularly CNS-rich. Only rarely could this function involve small RNA binding. Some transcribed CNSs encode short translation products as a form of negative control. Approximately 4% of *Arabidopsis* gene content is estimated to be both CNS-rich and occupies a relatively long stretch of chromosome: Bigfoot genes (long phylogenetic footprints). We discuss a 'DNA-templated protein assembly' idea that might help explain Bigfoot gene CNSs.

Address
Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Corresponding author: Freeling, Michael (freeling@nature.berkeley.edu)

Current Opinion in Plant Biology 2009, **12**:1–7

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Masahiro Yano and Roberto Tuberosa

1369-5266/$ – see front matter

DOI 10.1016/j.pbi.2009.01.005

Introduction

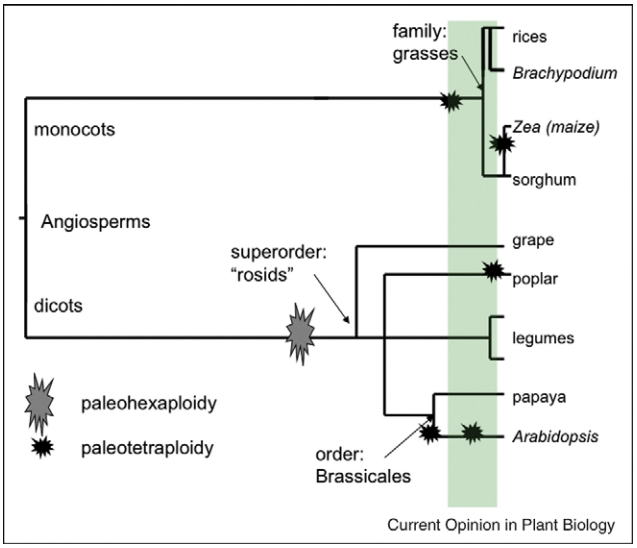
Conserved noncoding sequences (CNSs) are a subclass of phylogenetic footprints. Plant CNSs have been defined exclusively in order to maximize the chance that each CNS sequence exists because it was not removed by purifying selection and genetic drift. When properly defined, a CNS's existence implies that the DNA sequence functioned. Borrowing from vertebrate CNS research results [1–3], functions are expected to include DNA-binding of transcription or chromatin factors. Of course, some CNSs originally thought to be noncoding are, in light of ongoing experimental results, expected to actually code for a protein or an end-product RNA. While CNSs in animals and plants are known to function, the specific function is usually not yet known.

CNSs in plants are syntenic DNA sequence alignments just above noise that are shared between *usefully divergent* regions of homologous chromosome. The concept of 'usefully divergent' will be explained. CNSs are shared between or among orthologous chromosomes; homeologous CNSs are shared between chromosomes that diverged within a single nucleus, as occurs following segmental duplication or following tetraploidy. In theory, CNSs can also be detected between paralogous genes in tandem arrays, but the divergence times of such gene pairs are difficult to measure, especially when considering the consequences of gene conversion. Blastn, using NCBI default settings, finds CNSs above noise in pairwise comparisons when the e-value is less than or equal to a 15/15 exact base pair match, a definition first devised and defended by Kaplinsky *et al*. [4]. This noise cutoff was defined using the BLAST algorithm blastn but alternative alignment alternatives like LAGAN and CHAOS may be adjusted to be just under noise as well. This cutoff is approximately LAGAN at length 21 bp and 70% identity [5] or CHAOS at wordsize 8, score 25, rescore 250; see algorithm citations at CoGe: http://synteny.cnr.berkeley.edu/CoGe/. CoGe is an online suite of databases, interfaces, and applications for comparative genomics research. CoGe supports LAGEN, CHAOS as well as various BLAST tools. The choice among local alignment algorithms is less important than carefully setting the noise cutoff and far less important than picking alignments that are within the window of useful divergence [6], as will be discussed. Global alignments can be more accurate than local alignments but must be anchored and then extend no more than 500 bp along a typical plant chromosome. Five hundred bp upstream from exon 1 may be adequate to footprint a proximal promoter [7], but does not cover CNS-rich gene space [8] even in *Arabidopsis*, a plant with a small genome with correspondingly small gene spaces.

Figure 1 is a heavily pruned phylogenetic tree with paleopolyploidies denoted as starbursts; all citations have been published previously [9]. The shaded rectangle in Figure 1 is the aforementioned window of *useful divergence*. If chromosomes diverged too recently, then patches of conserved noncoding sequence will happen naturally by neutral carry over from the ancestor. If the chromosomes are too diverged, CNSs become difficult to detect by sequence alone, probably as a natural consequence of how binding sites evolve. Because maximally diverged grasses are within the CNS discovery window of Figure 1, and because the complete genomes of two rice subspecies (grasses) are soon to be joined by the grass genomes sorghum, *Brachypodium*, and maize (each

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

Figure 1



The window of useful CNS divergence—the vertical band—applied to a heavily pruned phylogenetic tree relating those plant species represented by fully sequenced genomes, and decorated with starbursts representing paleopolyploidy events. Phylogeny and citations: Missouri Botanical Garden, 'trees' section of http://www.mobot.org/MOBOT/research/APweb/welcome.html in March 2008. Recent polyploidy as is the case in the legume soybean is not noted. Different grass subfamilies and probably tribes originated within this window. The most recent tetraploidy (α) in the *Arabidopsis* lineage is within this window. Although the pre-grass tetraploidy and the split among different orders of rosids is outside this window, these ancient homeologous or orthologous CNSs certainly exist; they are few but informative. The maize tetraploidy, and probably the poplar tetraploidy, on the contrary, are positioned on the too-recent side of the window. Homeologous 'CNSs' derived from these too-recent pairs are often neutral carryovers from the ancestor. To use these footprints as an indication of functional sequence would demand a more stringent definition of 'CNS' than that given here.

available online), with *Setaria* on the horizon (Joint Genomes Institute, http://www.jgi.doe.gov/genome-projects/) CNS research has just begun.

Figure 2 illustrates some plant CNSs as they exist near genes. These images and the syntenic alignment data they represent are drawn by GEvo, the alignment viewer in CoGe [5]. Each image can be regenerated on-the-fly by engaging its tinyurl (hosted by tinyURL.com) presented beneath each graph, after which settings may be adjusted and research may continue. These putative CNSs cannot be annotated as noncoding until each has been compared to the most up-to-date protein, peptide, and RNA-gene databases. For example, it became clear only recently that circular miniproteins are encoded in the typical plant genome [10]. CNSs must be periodically reassigned to be genes in light of improved data implicating authentic gene products.

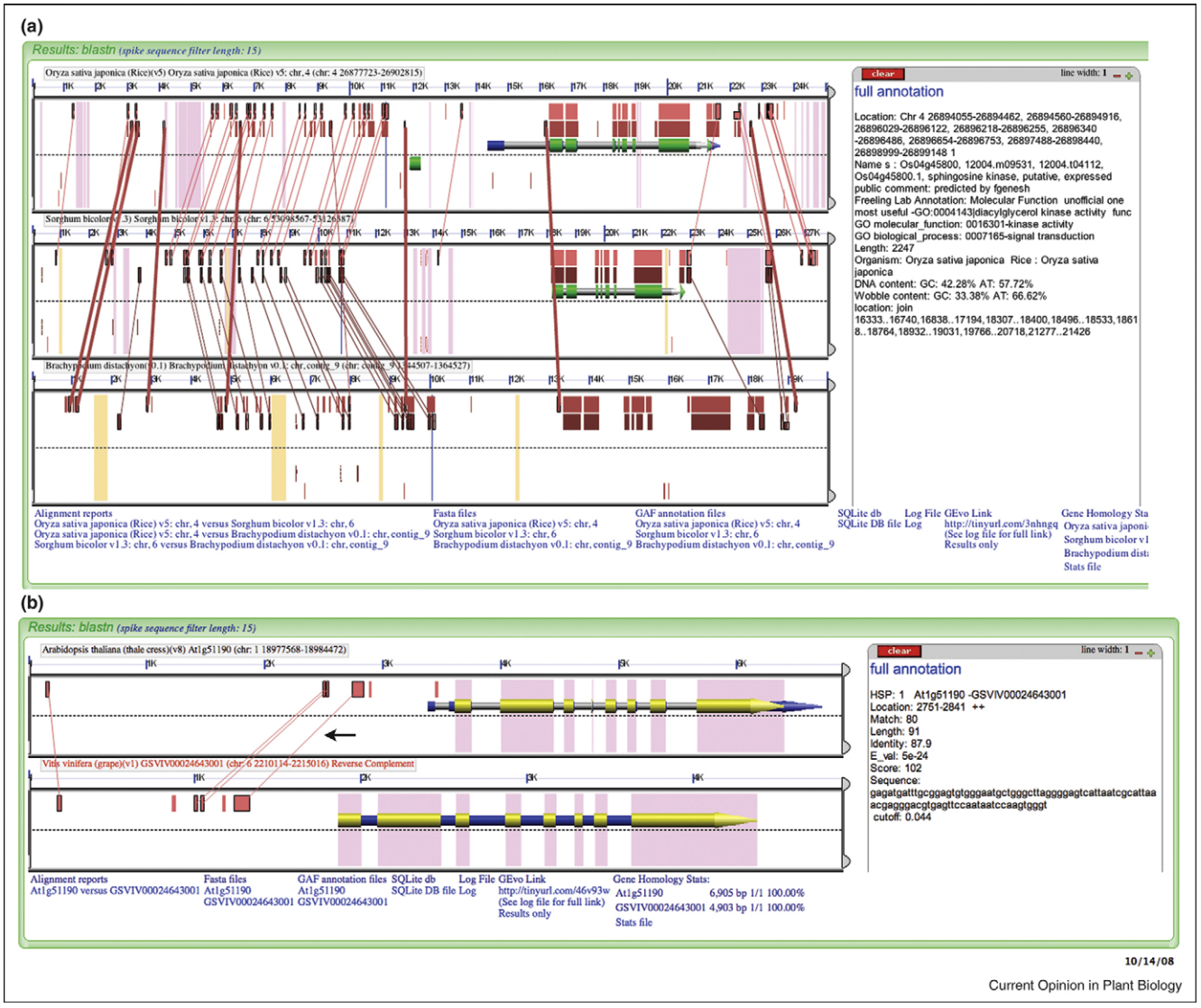
Figure 2a depicts orthologous blastn alignments of a *sphingosine kinase* gene from the grasses rice, sorghum, and *Brachypodium distachyon*. Sphingosine is a lipid component of membranes containing 18 carbons and sphingosine-1-phosphate is known to act as a signal regulating guard cell turgor [11]. Note the 12 kb of upstream chromosome and 2 kb of downstream chromosome that are covered by apparent CNSs; this is an extreme Bigfoot gene, as first studied in *Arabidopsis* [17]. As is typical, the region of proximal promoter has no CNSs at all (although there is certainly protein binding happening here). While most Bigfoot genes encode transcription factors, a variety of other functions are represented, although the significance of this breadth is not yet understood. Sorghum–rice and sorghum–*Brachypodium* high scoring pairs are connected by lines (Figure 2a) when syntenous; rice–*Brachypodium* syntenic CNSs are not drawn to avoid clutter, but may be inferred. Note how the spacing between orthologous CNSs can change significantly, a common feature of CNS pattern.

Plant ultraconserved sequences: a cautionary tale

A recent comparison of the sequences of *Arabidopsis* and rice [12] found many very similar sequences meeting the vertebrate definition of 'ultraconserved': >100 bp stretches of 100% identity. These identical blocks were often linked to each other forming clusters of 'ultraconserved' sequences on chromosomes. Since rice is a monocot and *Arabidopsis* is a dicot, such sequence identity is certainly remarkable. This result was interpreted as evidence for ultraconserved noncoding sequences comparable to those discovered in mammals [13]; mammalian CNSs can go back to the origin of vertebrates. Published evidence suggested that such sequences do not occur in plants [8], but it is difficult to prove a negative assertion. We began proofing the most significant of Zheng and Zhang's 'ultraconserved' noncoding sequences [12] and found that Sequence#4, a sequence within a chromosomal cluster, hit *Arabidopsis* and *japonica* rice at 100% identity, and also hit *indica* rice and sorghum as well, each in a few locations and always at very high or perfect sequence identity, which is consistent with the results of Zheng and Zhang. Because we performed our multiple blasts using the blast application within CoGe—following the tutorial called 'CoGe with rosids' [14]—each hit chromosomal segment could be quickly aligned with all others and visualized graphically. It was apparent that none of these hits were in syntenous positions. Vertebrate ultraconserved sequences are, by definition, orthologous (syntenous). We found that none of the other *Arabidopsis*–rice 100% identical sequences on the ultraconserved list were syntenous, so the term 'conserved' was not used properly. Using some trial-and-error, it turned out that Sequence#4 also exists in a known exon of mitochondrial DNA of all species mentioned except *japonica* rice. We chose as BLAST query a 7 kb segment of *Arabidopsis* mitochondrial DNA with Sequence#4 in the middle and used this against subject whole plant genomes. Large clusters of the original 'ultraconserved' sequences were accounted

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

Figure 2



GEvo graphic representations of plant CNSs, represented as blastn hits (colored rectangles), using settings as defined in text. Lines connect blast hits that are syntenous. Color code: masked sequence is purple and sequence gaps are yellow. Regenerate these data from CoGe on-the-fly using http://tinyurl.com/5dr95h and http://tinyurl.com/46v93w, respectively. **(a)** This grass *sphingosine kinase* (textbox to the right) has an exceptionally long and complicated phylogenetic footprint (Bigfoot). Aligning these three orthologs in the three pairwise combinations uncovers many CNSs (sorghum–rice and sorghum–*Brachypodium* syntenic pairs are connected by lines, as are some of the rice–*Brachypodium* syntenic pairs). Dark lines identify CNSs that are potentially specific to a grass lineage. **(b)** This rosid CNS pair is the most conserved CNS in plants discovered to date. The arrow marks an αCNS already known in *Arabidopsis* aligned with its best hit in grape, a fellow member of the superorder rosid but about as distant a rosid relative as is possible. The box lists characteristics of this special CNS. The associated gene encodes a transcription factor putatively functioning in meristem identity.

for in this way; all were 'ultraconserved' bits of a larger segment of mitochondrial DNA that was, itself, inserted intact and highly 'conserved.' We conclude that horizontal transfer is a likely explanation. There are many verified cases of horizontal transfer involving mitochondrial DNA in plants [15]. Further study will be required to decide whether or not this horizontal transfer is artifactual or biological. The above cautionary tale illustrates the obligation of researchers to proof the datasets delivered up in tables by computational biologists and bioinformatics specialists. Individual datapoints—case studies—comprising such lists must make sense.

The oldest published plant CNSs are between papaya and *Arabidopsis*, and these are neither long nor near identical [16]. Figure 2b is a GEvo graphic of blastn output for the single most significant (expect value = 5e − 24) hit between all *Arabidopsis* homeologous (α) CNSs [17] and the French grape genome [18]. This 80/91 nucleotide blastn hit is approximately 600 bp

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

upstream of an orthologous *Ainegumenta-like* transcription factor gene thought to be involved in meristem maintenance. This sequence, characterized in the box to the right of the graphic, is the most significant *Arabidopsis*–grape CNS, but it is not near animal standards [13,19] for being called 'ultraconserved.'

Plant versus vertebrate CNSs

Two of the most experimentally useful contributions of early plant CNS research was that plant CNSs are considerably smaller and less numerous than those in mammals, and that plant CNSs do not generally 'run together' on the chromosome [4,5,20]. With exceptions, individual plant genes may be assigned CNSs, and this CNS-richness metric engendered biological insight. Additionally, the most extreme animal CNSs are much more conserved than are the most conserved plant CNSs. The meanings of these plant–animal CNS data differences are not yet clear. Plant CNSs degrade or permute (mutate) relatively quickly over evolutionary time, although it seems likely that the binding functions themselves are conserved. This problem of alignment detectability is called 'binding site turnover' [21].

Experiments that address the possibility that some CNSs are 'artifacts'

The reason we define CNSs specifically in terms of synteny, window of divergence time, and alignment settings is to maximize the chance that any CNS functioned, and was therefore conserved. Even though it is difficult to imagine how the replication/correction machine could operate without error, it is formally possible that CNSs are mutational coldspots. There is evidence against this hypothesis in animals [22,23]. The ultimate coldspot is a patch of recent gene conversion; this trivial explanation could well explain paralogous (in the same nucleus) CNSs, such as the homeologous CNSs that have been retained after the most recent tetraploidy in the *Arabidopsis* lineage [17] or the pre-grass tetraploidy in the rice lineage [24]. Recent studies found many regions of gene conversion within the rice genome [25,26], including the approximately 6 Mb of near-identical sequence at the low-numbered ends of chromosomes 11 and 12. Three sorts of data argue against the conversion hypothesis. First, CNS-rich genes are not average genes, but regulatory and/or 'response to …' genes. Second, known transcription factor-binding motifs characterize the *Arabidopsis* homeologous CNS dataset, as will be reviewed. It seems improbable that *ABI13/VP1* transcription factor CNSs or G-box CNSs, for example, would be preferentially converted in their noncoding sequences. Third, none of the homeologous, *Arabidopsis* CNSs over 24 bp are identical to each other, indicating that gene conversions in noncoding DNA has not been significant, at least over the past several million years. Even so, any single homeologous CNS could be the result of a patch of gene conversion, and there is no reason to expect a CNS that originated by gene conversion to function. Because of the size of CNSs, it originally seemed possible that CNSs might be *MIR* or tasiRNA-encoding, or binding sites for small RNAs; data indicates that this RNA-world alternative could be correct for only 1.5% of *Arabidopsis* homeologous CNSs [17]. There are some sequences in the *Arabidopsis* αCNS Version 1 database [17] that are actually RNA genes, and are errors. These errors will be corrected in the upcoming Version 2 database (to be published online in CoGe downloads).

New genes are being discovered continuously, and any of these could conceivably have been included as an erroneous CNS. For example, recent proteogenomic studies in *Arabidopsis* resulted in the expansion of exons for 2446 TAIR7 models and called 838 new, usually short, genes [27]. We back-translated each of these unexpected peptides using tblastn and compared all sequences to each published [17] αCNS. No exact matches were found.

What about finding CNSs associated with those many duplicate genes that are not syntenic?

It is experimentally powerful to know that all CNSs in a dataset are contemporaneous. For orthologous CNSs, species divergence is the start point for the evolutionary clock. For homeologous CNSs, divergence began with the tetraploidy or segmental duplication. However, not all duplicate genes in a plant genome are syntenous, and these might well contain CNSs even though the origins and relative timing of the origins of these duplications are obscure. Local (tandem) duplicates comprise 5–30% of a genome depending on real differences, the algorithm used and on the definition of 'gene' [28]. Between one-fourth and three-fourths of the genes in *Arabidopsis* have been reported to have moved from their ancestral position, if they had one, or were generated *de novo* since the split between papaya and *Arabidopsis*, and these include the sorts of genes known to be CNS-rich, like *MADS-box* and *B3-box* transcription factor genes [29]. In theory, exon divergence could be used as an internal measure of 'the window of useful divergence'—and perhaps as a measure of regional gene conversion—and properly defined CNSs could then be collected for nonsyntenic duplicates as well as for syntenic duplicates.

General characteristics of plant CNSs

The general conclusions from the original maize-rice orthologous CNS studies [4,5,20] have been replicated—in general—in the two large-scale homeologous CNS studies in *Arabidopsis* [8,17] and also in rice [24]. Plant CNSs average from 20–30 bp in length. The CNSs tend to be positioned close to one gene, with exceptions, so individual pairs of genes or their Gene Ontology (GO) terms were quantified for CNS-richness. All studies conclude that 'regulatory genes' are CNS-rich, with genes encoding transcription factors being generally more CNS-

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

rich than genes encoding protein kinases. Still, the average (median) homeolog in plants has 0–2 CNS. In one *Arabidopsis* homeologous CNS study [8], 246 genes occupied an exceptionally long stretch of chromosome where the regions full of CNSs were conspicuous in being exon voids; these genes were called 'Bigfoot' and were annotated with 'response to' GO biological function categories, these often being 'transcription factor activity' as well. The distance between Bigfoot gene 5′CNSs and exon 1 averaged 3.1 kb. This same study found a CNS over-abundance of the most famous transcription factor binding site in plants, the G-box (CACGTG palindrome); several other known motifs—plus previously unknown 7mers—were significantly over-represented, but none near the extent of the G-box. CNSs occur 5′, 3′, and in introns, but the rice homeologous CNS distributions [24] were less skewed to 5′ than were similar distributions around *Arabidopsis* genes.

While it seems reasonable that some CNSs contain active transcription factor binding sites, it is important to emphasize that most plant genes function in developmentally complicated ways in the absence of CNSs.

Case studies implying specific CNS function

Box 1 summarizes data, with citations, from which general or specific CNS function was inferred. From the totality of animal data, CNSs are expected to include *cis*-acting regulatory sites [1–3], and they do (Box 1). There is one early report of a CNS involving nuclear matrix attachment [30]. Box 1 also includes regulatory CNSs that function to slow down translation because they are themselves translated [31]; while these sequences are formally 'coding,' since their function is not in the product peptide *per se*, then they seem properly included together with CNSs. It seems likely that ongoing proteomic work in plants will find sequences for short peptides that will turn out to be CNSs, but that is not yet the case.

Uchida *et al*. [32] completed perhaps the most developmentally complete study of two CNSs in the proximal promoter of the Class I *KNOX* flagship gene, a gene named *Knotted1* in grasses and relatives (monocots) and *STM* in *Arabidopsis* and relatives (dicots). One of these CNSs contains a known *cis*-acting binding motif called the K-box and the other the RB-box. The binding functions of each box apparently accounts for conservation. These workers went on to perform *in vitro* mutation experiments, often using reporter genes, to work out the details of these regulatory functions and to relate them to the compound nature plant leaves. In fact, the dicot CNSs are not conserved well enough in the monocots to be termed 'CNSs' using dicot–monocot pairwise blastn comparisons, but multiple alignments anchored on the 5′ATG demonstrate clearly the underlying homology of all *STM* promoters studied, at least at their core boxes.

The most inspirational result of CNS research to date [33] is the mapping and functional confirmation of an important maize quantitative trait locus (QTL), *Vgt1* (a flowering time locus named *Vegetative to generative transition1*). *Vgt1* maps to a 24 bp maize–sorghum–rice CNS 70 kb upstream from an *Apetala2-like* maize gene. *Vgt1* acts in *cis* on the expression levels of the gene encoding this transcription factor in *cis*, so the word 'enhancer' is probably applicable. Interestingly, while the early-flowering

Box 1

[a]Type	Function associated with plant CNSs	Ref
G	Grass regulatory genes are rich in orthologous CNSs	[5,20]
G	*Arabidopsis* genes that are induced by stimuli and/or encode transcription factors are rich in homeologus CNSs, and are often 'Bigfoot' genes. *japonica* rice homeologous CNSs are similar.	[17,24]
G	*Arabidopsis* homeologous CNSs are significantly enriched for several known transcription factor binding motifs, especially the G-box	[8]
S	Intron CNSs in a Class I homeobox gene (*knotted1*) bind a negative *cis* regulator, a binding that is disrupted by *Mu* transposons, but only in *Mu*-active lines.	[20,38]
S	5′CNSs contain conserved, known transcription factor binding motifs and motif patterns: *RAB16/17* in grasses, *rbc a/b* in dicots, and [b]proximal promoters in dicots.	[7,39,40]
S	Two 5′CNSs of the *SHOOT MERISTEMLESS*/knotted1 homeobox gene in dicots/monocots, and their binding motifs, *cis*-regulate repression/re-establishment of leaf expression.	[32]
S	A *cis*-acting, regulatory QTL for flowering time in maize, positioned 70 kb upstream of an *Apetala2*-like gene, contained an intact CNS in the early flowering parental line, but the late-flowering parental had this CNS disrupted by a 144 bp MITE insertion.	[33]
S	Some 5′UTR grass CNSs are uORFs, one mechanism to downregulate translation.	[31]

[a] G = general, S = specific.

[b] Vandepoele *et al*. [7] used phylogenetic footprinting of homologous proximal promoters of dicot genes, along with TF binding motif over-representation and transcript co-expression, to infer functional regulatory modules composed of two or more transcription factor binding sites in close proximity. Being confined to noncoding space close to the start of transcription, this work addresses the transcription factor binding potential of a few CNSs, not CNSs in any general way, and specifically not Bigfoot gene CNSs.

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

Vgt1 allele had an intact CNS, a 144 bp MITE transposable element insertion disrupted the CNS of a late-flowering allele, and this allele also expressed the *Apetala2* transcription factor mRNA to a lower level. Flowering time is a plant trait that is particularly variable among closely related taxa, and a trait that responds to breeding, implying polymorphism within the breeding population. While it is never logically sound to extrapolate from one point, the work of Salvi *et al*. [33] should inspire others to use CNSs as mapping markers, especially for mapping those sorts of phenotypes that account for the *differences* among plants that make them taxonomically distinct and account for the *improvements* on which agriculture is based. QTL research, because it maps complex and poorly understood phenotypes existing in the wild to the genome, thereby providing a way to reduce phenotype to sequence, is perhaps the most pioneering of any genetic approach toward better understanding life. It is the obligation of genomic research to provide infrastructure for QTL studies.

There have been no advances in understanding a fundamental 'enigma' concerning plant CNS distribution. Those genes high up the regulatory cascade—especially those that respond to stimuli—tend to be CNS-rich. However, genes known to function on cellular building blocks—those encoding house-keeping enzymes, for example—tend to be CNS-poor. CNS-richness seems a reasonable quantitative metric for at least one sort of gene regulation. Using this metric, *the most important regulatory genes are also the most regulated*. Consider a corporation metaphor. It is not silly to think that the CEO (chief executive officer) of a corporation might also be the most regulated. CEOs must comply by law with rules and regulations of little concern to most workers. Only when considering individual networks of gene regulation within a larger regulatory system can this enigma begin to make sense. CNS-rich genes are perhaps under the control of a level of 'systems' coordination. We might guess that this system coordinates a plant's or population's ability to endure fluctuating stresses, but we actually know very little.

One particularly intriguing hypothesis explaining why certain plant genes are Bigfoot

Knowing little, some ideas are more 'beautiful' than other ideas only because they stimulate the imagination. One such idea for Bigfoot CNS-regions (e.g. Figure 2a) may be the DNA-templated protein complex idea [4]. This idea posits that particular protein–protein or protein–RNA complexes that eventually assemble as a part of chromatin do so under the direction of a DNA template and do not do so by self-assembly alone. A pattern of CNSs would then *catalyze* this assembly. This idea might be particularly applicable to changing chromatin in epigenetically heritable ways. We predict that chromosomal regions around Bigfoot genes are prone to this sort of epigenetic marking. All ideas and data pertaining to structure begetting subsequent structure—structural inheritance—have been of special interest ever since studies, largely from the Sonneborn lab, on the inheritance of the directions of unit territories in the cortex of ciliates [34].

At this point, it would be good to know the current state of plant research on chromosomal units of function, loops perhaps, on the boundary sequences that define these blocks of chromatin, and on how chromatin states might be altered and then inherited in a cell lineage. Somewhere in these research results might be candidate DNA–protein complexes associated with the CNSs of plant Bigfoot genes. We found no plant research to review. If plants and animals share the essential mechanisms delimiting chromosomal domains and units of epigenetic function, as might be the case [35], then mechanisms thought to apply to animal epigenetic domains might inform plant research. Animal insulators, like vertebrate CTCF [36], are conserved proteins that bind to conserved DNA sequences. Such binding can block a promoter from a distant enhancer, alter chromosomal looping architecture and may be involved in changes in silencing or epigenetic state, as reviewed [37]. In any case, a purely hypothetical templating role for clusters of plant CNSs—those 5′ average 3.1 kb away from the Bigfoot gene exon1 [8]—predicts that the pattern of these CNSs should influence chromatin structure and function.

Conclusion

These are early and exciting times for plant CNS research. Mysteries abound.

Acknowledgements

Funded by NSF DBI0337083 to MF. Supported by the CoGe system, a comparative genomics software product of the Freeling laboratory, developed by Eric Lyons (lead), Brent Pedersen, and Josh Kane.

References

1. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome Res* 2005, **15**:137-145.

2. Hardison RC: **Comparative genomics**. *PLoS Biol* 2003, **1**:E58.

3. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome**. *Genome Res* 2007, **17**:201-211.

4. Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M: **Utility and distribution of conserved noncoding sequences in the grasses**. *Proc Natl Acad Sci U S A* 2002, **99**:6147-6151.

5. Guo H, Moose SP: **Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution**. *Plant Cell* 2003, **15**:1143-1158.

6. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequence**. *Plant J* 2008, **53**:661-673.

7. Vandepoele K, Casneuf T, Van de Peer Y: **Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics**. *Genome Biol* 2006, **7**:R103.

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005

8. Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC: **G-boxes, Bigfoot genes and environmental response: characterization of the intragenomic conserved noncoding sequences of Arabidopsis**. *Plant Cell* 2007, **19**:1441-1457.

9. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH: **Synteny and collinearity in plant genomes**. *Science* 2008, **320**:486-488.

10. Gruber CW, Elliott AG, Ireland DC, Delprete PG, Dessein S, Goransson U, Trabi M, Wang CK, Kinghorn AB, Robbrecht E *et al.*: **Distribution and evolution of circular miniproteins in flowering plants**. *Plant Cell* 2008, **20**:2471-2483.

11. Coursol S, Stunff J, Lynch D, Gilroy S, Assmann S, Spiegel S: **Arabidopsis sphingosine kinase and the effects of phytosphingosine-1-phosphate on stomatal aperture**. *Plant Phys* 2005, **137**:724-737.

12. Zheng WX, Zhang CT: **Ultraconserved elements between the genomes of the plants *Arabidopsis thaliana* and rice**. *J Biomol Struct Dyn* 2008, **26**:1-8.

13. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**:1321-1325.

14. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang XW, Bowers J, Paterson A, Lisch D, Freeling M: **Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids**. *Plant Phys* 2008, **148**:1772-1781.

15. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution**. *Nat Rev Genet* 2008, **9**:605-618.

16. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL *et al.*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**:991-996.

17. Thomas BC, Rapaka L, Lyons E, Pedersen B, Freeling M: **Intragenomic conserved noncoding sequences in Arabidopsis**. *Proc Natl Acad Sci U S A* 2007, **104**:3348-3353.

18. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al.*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**:463-467.

19. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes**. *BMC Genomics* 2004, **5**:99.

20. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, Freeling M: **Conserved noncoding sequences in the grasses**. *Genome Res* 2003, **13**:2030-2041.

21. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB: **Large-scale turnover of functional transcription factor binding sites in Drosophila**. *PLoS Comput Biol* 2006, **2**:e130.

22. Kim S, Pritchard J: **Adaptive evolution of conserved noncoding elements in mammals**. *PLoS Genetics* 2007, **3**:e147.

23. Drake J, Bird C, Nemesh J, Thomas D, Newton-Cheh C, Raymond A, Excoffler L, Attar H, Antonarakis S, Dermitzakis E *et al.*: **Conserved noncoding sequences are selectively constrained and not mutation cold spots**. *Nature Genetics* 2006, **38**:223-227.

24. Li X, Tan L, Wang L, Hu S, Sun C: **Isolation and characterization of conserved noncoding sequences among rice (*Oryza sativa* L.) paralogous regions**. *Mol Genet Genomics* 2008 doi: 10.1007/s00438-008-0388-4.

25. Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH: **Extensive concerted evolution of rice paralogs and the road to regaining independence**. *Genetics* 2007, **177**:1753-1763.

26. Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK-S, Wang J, Zheng X: **Gene conversion in the rice genome**. *BMC Genomics* 2008, **9**:1-8.

27. Castellana N, Payne S, Shen Z, Stanke M, Bafna V, Briggs S: **Proteogenomic discovery, correction and confirmation of Arabidopsis gene models**. *Proc Natl Acad Sci U S A* 2008 doi: 10.1073/pnas.0811066106.

28. Rizzon C, Ponger L, Gaut BS: **Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice**. *PLoS Comput Biol* 2006:2.

29. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D: **Many or most genes in Arabidopsis transposed after the origin of the order Brasicales**. *Genome Res* 2008, **18**:1924-1937.

30. Avramova Z, Tikhonov A, Chen M, Bennetzen JL: **Matrix attachment regions and structural colinearity in the genomes of two grass species**. *Nucleic Acids Res* 1998, **26**:761-767.

31. Tran M, Schultz C, Baumann U: **Conserved upstream open reading frames in higher plants**. *BMC Genomics* 2008, **9**:361.

32. Uchida N, Townsley B, Chung KH, Sinha N: **Regulation of SHOOT MERISTEMLESS genes via an upstream-conserved noncoding sequence coordinates leaf development**. *Proc Natl Acad Sci U S A* 2007, **104**:15953-15958.

33. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E *et al.*: **Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize**. *Proc Natl Acad Sci U S A* 2007, **104**:11376-11381.

34. Beissen J, Sonneborn T: **Cytoplasmic inheritance of the organization of the cell cortex of *Paramecium aurelia***. *Proc Natl Acad Sci U S A* 1965, **53**:275-282.

35. van Driel R, Fransz P: **Nuclear architecture and genome functioning in plants and animals: what can we learn from both?** *Exp Cell Res* 2004, **296**:86-90.

36. Zhang R, Burke L, Rasko J, Lobanenkov V, Renkawitz R: **Dynamic association of mammalian insulator protein CTCF with centrosomes and the midbody**. *Exp Cell Res* 2004, **294**:86-93.

37. Bushey AM, Dorman ER, Corces VG: **Chromatin insulators: regulatory mechanisms and epigenetic inheritance**. *Mol Cell* 2008, **32**:1-9.

38. Greene B, Walko R, Hake S: **Mutator insertions in an intron of the maize *knotted1* gene result in dominant suppressible mutations**. *Genetics* 1994, **138**:1275-1285.

39. Buchanan C, Klein P, Mullet J: **Phylogenetic analysis of 5′-noncoding regions from the ABA-responsive *rab16/17* gene family of sorghum, maize and rice provides insight into the composition, organization and function of cis-regulatory modules**. *Genetics* 2004, **168**:1639-1654.

40. Weeks KE, Chuzhanova NA, Donnison IS, Scott IM: **Evolutionary hierarchies of conserved blocks in 5′-noncoding sequences of dicot *rbcS* genes**. *BMC Evol Biol* 2007, **7**:51.

Please cite this article in press as: Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants, Curr Opin Plant Biol (2009), doi:10.1016/j.pbi.2009.01.005