

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Patterns of computed conserved noncoding sequence loss following the paleopolyploidies in the Maize and Brassica lineages and their functional consequences

Permalink

<https://escholarship.org/uc/item/6z92q4gq>

Author

Subramaniam, Shabari

Publication Date

2013

Peer reviewed|Thesis/dissertation

Patterns of computed conserved noncoding sequence loss following the paleopolyploidies in the maize and *Brassica* lineages and their functional consequences

By

Sabarinath Subramaniam

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Plant Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Freeling, Chair

Professor Richard Harland

Adjunct Associate Professor Frank Harmon

Fall 2013

Patterns of computed conserved noncoding sequence loss following
the paleopolyploidies in the maize and *Brassica* lineages and their
functional consequences

Copyright 2013

by

Sabarinath Subramaniam

Abstract

Patterns of computed conserved noncoding sequence loss following the paleopolyploidies in the maize and *Brassica* lineages and their functional consequences

by

Sabarinath Subramaniam

Doctor of Philosophy in Plant Biology

University of California, Berkeley

Professor Michael Freeling, Chair

Following polyploidy, duplicate genes are often deleted, and if they are not, then duplicate regulatory regions are sometimes lost. What is the mechanism for this loss, and what is the chance that such a loss removes function? To explore these questions, we followed individual *Arabidopsis thaliana*-*Arabidopsis thaliana* conserved noncoding sequences (CNSs) into the *Brassica* ancestor, through a paleohexaploidy and into *Brassica rapa*. Thus, a single Brassicaceae CNS has six potential orthologous positions in *Brassica rapa*; a single *Arabidopsis* CNS has three potential homeologous positions. We reasoned that a CNS, if present on a singlet *Brassica* gene, would be unlikely to lose function compared to a more redundant CNS, and this is the case. Redundant CNSs often become nondetectable. Using this logic, each mechanism of CNS loss was assigned a metric of functionality. By definition, proven deletions do not function as sequence. Our results indicated that CNSs that become nondetectable by base substitution or large insertion are almost certainly still functional (redundancy does not

matter much to their detectability frequency), while, those lost by inferred deletion or indels are about 75% likely to be nonfunctional. Overall, an average nondetectable, once-redundant CNS > 30 bps in length has a 72% chance of being nonfunctional, and that makes sense because 97% of them sort to a molecular mechanism with “deletion” in its description, but base substitutions do cause loss. Similarly, proved-functional G-boxes become undetectable by deletion 82% of the time. Fractionation mutagenesis is a procedure that uses polyploidy as a mutagenic agent to genetically alter RNA expression profiles, and then to construct testable hypotheses as to the function of the lost regulatory site. We show fractionation mutagenesis to be a “deletion machine” in the *Brassica* lineage.

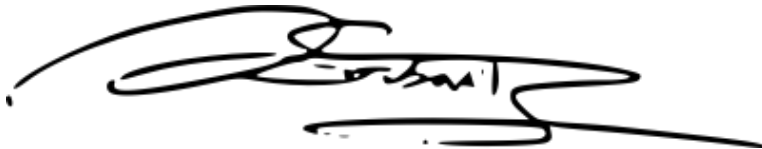
Dedication & Acknowledgements:

Foremost, I would like to express my sincere gratitude to my advisor Prof. Michael R Freeling for his support and guidance through my PhD study and research, as well as for his patience, motivation, enthusiasm, and immense knowledge. I cannot imagine a better advisor and mentor.

I would like to thank my wife, Sree Subha Madhusudanan, for her love, kindness and unwavering support over the past 15 years. Your unrelenting optimism has been a beacon of hope for me in my darkest hour.

I would like to thank Dr. Eric Lyons, whom I consider a true friend, for being always willing to help and give his best suggestions. It would have been a lonely lab without him.

I would like to dedicate this thesis to my parents, Shivakumar MS and Rajkumari Shivakumar and to my brother, Sakthidhar Subramaniam, the most courageous soul I know. Your courage and strength have always been my inspiration.

A handwritten signature in black ink, appearing to read 'Sabarinath Subramaniam', with a long horizontal flourish extending to the right.

-Sabarinath Subramaniam, Dec 2013

Table of Contents

Chapter 1: Introduction to conserved noncoding sequences (CNSs) in higher plants 1

Introduction and definitions	1
Detection of CNSs in plant genomes	4
General Characteristics of Plant CNSs Suggest a Predominant Regulatory Function	5
Plant versus vertebrate CNSs	6
Experiments that address the possibility that some CNSs are artifacts, mutation coldspots, gene conversions, misannotations and the like	7
CNS-rich genes, bigfoot genes, are particularly enigmatic	9
Fresh data on how CNS richness influences the ability of a gene to be retained post-tetraploidy.	9
Conclusions.....	11
Figures	12
Tables	17

Chapter 2: The fate of *Arabidopsis thaliana* homeologous CNSs and their motifs in the paleohexaploid *Brassica rapa* 19

Introduction	19
Results	23
<i>At-At</i> α CNSs updated to Version 2	23
General features of detectability of <i>At-At</i> CNSs in <i>Br</i>	23
α CNS length vs. detectability	24
The functionality metric: Deletion to loss-of-function is the primary mechanism for removal of α CNSs >30 bp long in <i>Br</i>	25
Detectability of CNS-enriched transcription factor binding site (TFBS) motifs, especially the G-box, in <i>Br</i> orthologous positions.....	26
Detectability studies for G-boxes in α CNSs, CNSs that are particularly likely to function.....	28
Discussion	29
Purifying selection in <i>Brassica rapa</i> resulted in many gene regulatory regions that have lost cis-acting binding sites, and 75% of the time, these sites were deleted and therefore have no chance to function in the ancestral manner.....	29
Even fully fractionated <i>Brassica rapa</i> genes (singlets) are likely to be functionally redundant at least 10% of the time.....	31
5' and 3' α CNSs generally confer function	33
The G-box and motif detectability.....	34
Fractionation: nature's "deletion machine"	35
Methods.....	37
Rationale for confining this paper to CNSs defined in one specific way.....	37
<i>Arabidopsis</i> α CNSs.....	37
Locating orthologous coordinates for α CNSs within <i>Br</i>	38
Identifying the molecular mechanisms that caused the lack of detectability of α CNSs in <i>Br</i> doublets and triplets.....	39
mRNA levels for <i>Br</i> genes in seedling root and shoot.....	42
Revised CNS-enriched TFBS motif list	42
α CNS that are reinforced by overlap with published Pil3-like5 protein (PIL5) binding sites and their G-boxes.....	43
Tables	44

Figures	45
Supplemental Data	50
Chapter 3: Computational pipeline to track the evolutionary fates of arabidopsis conserved noncoding sequences through a paleohexaploidy and into the genomes of its close relative, <i>Brassica rapa</i>: PL3 Genespace Tweaker + Mapping Over	51
Introduction	51
Results	54
<i>CNSs between Aethionema arabicum (Aa) and Arabidopsis thaliana (At)</i>	54
<i>Detectability of Aa-At CNSs in Br</i>	54
<i>Cartoons of CNS detectability</i>	55
Methods.....	55
<i>Genespace Tweaker</i>	55
<i>Mapping Over of At tweaked CNSs to expected syntenic chromosomal region in Br</i>	56
<i>CNS detectability in Br</i>	56
<i>Genespace Tweaker and MO code</i>	57
Acknowledgements.....	57
Figures	58
Supplemental Tables	66
Chapter 4: Subfunctionalization, nonfunctionalization and cis-acting modules: arabidopsis conserved noncoding sequences used as markers to quantify fractionation patterns following the hexaploidy in the <i>Brassica rapa</i> lineage.....	67
Introduction	67
Results	70
<i>Many At-Aa CNSs do not map over to any Br ortholog</i>	70
<i>The alternative classes of MO CNS fractionation cartoons of Brassica rapa v. Chiifu doublets.</i>	70
<i>Subfunctionalization-nonfunctionalization data for Br doublets with 2 CNSs: Even the simplest case gives complex results</i>	71
<i>Doublets subfunctionalize increasingly as their CNS/gene numbers increase</i>	72
<i>Genome dominance and nonfunctionalization; the ancestral (intact) gene of a nonfunctionalized doublet tends to be on the dominant subgenome of the doublet, as expected</i>	72
<i>The excess of deletion runs for linked CNS is universal over the entire dataset</i>	73
<i>Cis-acting modules often contain duplicated TF-binding motifs</i>	73
<i>CNSs occurring in runs present an opportunity to find motif patterns specific to cooperative binding principles that may direct the assembly of transcriptomes or their components</i>	74
Discussion	75
List of Supplemental Documents	77
Methods.....	77
<i>Preparing our automated CNS discovery pipeline output for Map Over (MO) from At to Br</i> ..	77
<i>The Mapping Over (MO) Pipeline</i>	78
Figures	79
Tables	83

Chapter 5: MotifView (in CoGe): Visualization tool for comparative analysis of motifs and ENCODE-like features.....	85
Preface	85
Introduction	86
Results	87
<i>Output example with many motifs.....</i>	<i>87</i>
<i>Studying Mutations via unexpected patterns of fractionation</i>	<i>88</i>
<i>Tab-delimited output.....</i>	<i>89</i>
Conclusion.....	89
Methods.....	90
Acknowledgements.....	92
Tables.....	93
Figures.....	94
Chapter 6: During the comparison of genomes, orphaned conserved noncoding sequences may signify regulatory novelty.....	98
Introduction	98
Preliminary Results	100
<i>Validation of orphan CNSs with GEvo.....</i>	<i>100</i>
<i>Detectability of orphan CNSs in Br subgenomes</i>	<i>100</i>
<i>Frequency of TFBS motifs within orphan CNSs.....</i>	<i>101</i>
Discussion	101
Methods.....	103
Tables.....	104
Figures.....	105
Supplemental tables	111
Addendum I: Annotated bibliography of latest CNS research	112
Preface	112
Bibliography.....	112
Addendum II: The comparison of genome-wide, orthologous conserved noncoding sequences (CNSs) in arabidopsis obtained using two very different methods: pairwise blast versus multiple global alignments..	116
Introduction	116
VEGI and MF-lab CNSs overlap.	118
Conclusion.....	119
Methods.....	119
Figures.....	121
Supplemental Data	127
Bibliography	128

Chapter 1: Introduction to conserved noncoding sequences (CNSs) in higher plants

The following chapter has been published as a peer-reviewed article in *Current Opinions in Plant Biology*. GIVEN THE DUAL AUTHORSHIP, I THINK IT IS IMPORTANT FOR YOU TO MAKE CLEAR YOUR CONTRIBUTIONS

Freeling M, Subramaniam S. (2009) "Conserved noncoding sequences in higher plants" *Current opinions in Plant Biology*, 12(2):126-32.

Introduction and definitions

Plant genomes carry a great diversity of all sorts of sequences and many have functions. Some of these sequences specify functions important to the plant by encoding RNA sequences, some of which (genes or coding domain sequences) encode proteins. Some sequences encode binding functions important to the plant, such as the DNA sites near genes that bind regulatory proteins (motifs), while others may function to block the binding or movement of chromosomal proteins (e.g. insulators). One of the goals of molecular biology studies is to discover the exact functions specified by the genome. However, this is not a simple task. A typical plant has about 30,000 genes, and this does not include genes that function largely selfishly such as most transposons. All of these about 30,000 genes encode one or more messenger RNAs (mRNAs) and many of these genes contain different parts: introns, exons, RNA binding sites, DNA binding sites and similar. Somewhere near the transcriptional units of a gene comprised of its coding regions are the chromosomal regulatory regions that enable the gene to be a part of one or more biological pathways or networks, transcription factor binding sites, enhancer sites, insulator sites and so forth. Added up, there are "millions" of specific DNA sequences that carry specific coding, binding or blocking functions important to gene function sequences with a chromosomal start, stop and strand. We know something functional, however vague, about several thousand of these sequences and almost nothing about the meaning of their combinations.

There is a way to infer that particular sequences functioned over evolutionary time, and presumably still function, even though the function itself remains unknown. This is because sequences that carry specific function will resist deletions, insertions, and base substitutions

expected to alter function; these sequences are said to be "conserved." This conservation derives from the ongoing process of purifying selection, where most mutations in functional nucleotides confer some lowering of fitness and tend to not pass on to subsequent generations. Thus, there is a tendency for the DNA that encoded ancestral functions to be conserved over time. These blocks of conserved sequence are called "phylogenetic footprints" (Table 1). Exons, with their open reading frames, comprise a good portion of the longer phylogenetic footprints in plants, as expected.

As described already, some sequences with specific functions for the plant do not encode proteins, and these are sometimes resolved as phylogenetic footprints. Some encode those parts of the transcripts that are not translated, and some RNAs fold into functional gene products themselves. A specific category of phylogenetic footprints represent sequences that function by binding or blocking some molecule, so called conserved noncoding sequences, or CNSs; they have precise definitions that fit the biological realities of plants (Table 1). Since functional DNA that does not encode a product must function as a CNS, the annotation of protein-coding sequence and RNA product-coding sequence is a necessary part of CNS discovery. Hence every CNS requires validation following the release of each new version of a genome's annotations.

In order to visualize the optimal number of plant CNSs, much care has been given to the divergence times between the chromosomal regions being footprinted (compared). Among divisions of eukaryotes, flowering plants have the maximum frequency of successful polyploidy over the last 150 million years or so. Figure 1 shows a heavily pruned phylogenetic tree of plants with those genomes at the tips that have been sequenced and released. For comparison, eutheria (a vertebrate clade about [160 million years old](#)?) have had no successful tetraploidies. Each ancient plant tetraploidy is a small starburst. Each tetraploidy has a doubled genome for only a short time, and then the chromosomes rearrange and one or the other of the two duplicate (homeologous) genes tend to be lost; this post-polyploidy process is called "fractionation." Plant species too distantly related, like any dicot compared to any monocot (Figure 1), present difficulties simply lining up the chromosomes into orthologous pairs. CNS discovery begins by comparing genes within chromosomes that are

obviously matched, but diverged enough to ensure that functionless DNA has randomized. However, the divergence time separating the two chromosomal regions cannot be too recent either, because then sequences would be similar just by virtue of recent divergence (Fig. 1m Tab. 1). Sometimes there are no two sequenced species that are usefully diverged. However, sometimes the most recent tetraploidy happened a useful number of years ago, as with the most recent *Arabidopsis thaliana* tetraploidy, called "alpha" (notice the starburst in the window of Figure 1). In such a case, CNSs can be discovered by comparing retained segments of syntenic, homologous (homeologous, Table 1 definitions) chromosomes within the same genome.

We use a local alignment algorithm, blastn, to find CNSs in usefully diverged, homologous DNA sequences. The settings and e-value cutoffs are specifically defined for plant CNS discovery (Table 1). Local alignment algorithms are preferred because noise is easily detected: noise includes alignments-- called "blast hits"-- that are not syntenous, and tend to be palindromic or overly simple. The CNS significance cutoff is set just above noise. The CNSs in Figure 2, colored orange, exemplify CNS discovery between the grasses sorghum and rice. Note that the divergence between sorghum-like grasses and rice-like grasses happened within the window of useful divergence (Fig. 1). CNSs are sorted to their nearest gene pair. The overall result for two comparable genomes is a long list of CNSs. Each is a pair of homologous sequences, either orthologous or homeologous (Table 1), and each has been sorted to one gene in the genome. CNS discovery has been automated for plants (Fig. 2: purple rectangles on the rice model annotation line). Following genes by their CNS-richness, or by the DNA-binding motifs within their CNSs, is beginning to generate useful data on how CNSs-- or clusters or families of CNSs-- specify function.

Specific function is known for only a few plant CNSs. One is the cluster of CNSs in the longest intron of the grass homeobox gene *knotted1*. When this gene is ectopically expressed in the leaves of grass plants, the leaf cells do not mature properly and do not stop proliferating. Figure 3 shows the region of a maize leaf between sheath and blade covered by grotesque finger-like projections, called knots; this is a dramatic version of the dominant *knotted1* phenotype.

The laboratory of Sarah Hake, Plant Gene Expression Center, ARS, Albany California (cited in Figure 3 legend), found that 9 different transposon insertions happened in this large intron, and just in the CNSs, and that these could cause ectopic expression and knots. We hypothesize that the transposon insertions prevent the binding of some product to the intron CNSs, so *knotted1* turns on in leaves. All monocot *knotted1* genes tested so far have these intron CNSs.

Detection of CNSs in plant genomes

The original paper that found CNSs between maize BACs and the rice genome defined and defended the blastn settings and cutoffs of Table 1 (KAPLINSKY et al. 2002). Eric Lyons, the project lead and programmer of the CoGe suite of software tools (<http://coge.iplantcollaborative.org/CoGe/index.pl>), compared and contrasted various DNA sequence comparison algorithms for CNS resolution, purity and discovery speed. He concluded that blastn (or bl2seq) using our exact settings was an excellent compromise (LYONS et al. 2008). Of course, any algorithm that dips into nonsyntenic, repetitive noise will find some additional syntenic footprints. Figures 4A and B show GEvo panels of blastn output from comparisons of homeologous (Table 1) regions of *Arabidopsis thaliana* settings defined in Table 1, (Figure 4A) and similar settings with a slightly lowered e-value cutoff (Figure 4B and legend). The choice among 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 (LYONS et al. 2008).

CNS discovery has been automated. The CNS Discovery Pipeline 1.0 (WOODHOUSE et al. 2010) entails repeat masking, tandem identification, orthologous pair finding, CNSs discovery, and CNS sorting to the nearest orthologous gene pair.. This pipeline utilizes a new algorithm for finding bonafide orthologous pairs when comparing genomes with more than one possible syntenic partner, as is the case for all plant comparisons seeking syntenic runs of gene pairs (TANG et al. 2010). The rice-sorghum orthologous CNSs have been cataloged and form the starting point for hypothesis testing (SCHNABLE et al. 2011). The rice genome model line in Figure 2 is annotated with more than the exons, introns and transcriptional unit. The purple bars on this model line are the locations of the Version 1.0 (v1) CNS discovery pipeline automated sorghum-rice CNS calls. Also note

the orange-sorghum-rice blast hits above the rice model line, and note how there is excellent correspondence between the v1 pipeline calls and the actual data, that data being the blast hits themselves (rectangles colored Orange in Fig. 4). Thus, Figure 2 exemplifies how the v1 CNS discovery pipeline was proofed for accuracy. Please visit this experiment at <http://genomeevolution.org/r/a51s>; the researcher may change settings, algorithms, dimensions and more and then rerun the experiment on-the-fly.

General Characteristics of Plant CNSs Suggest a Predominant Regulatory Function

The general conclusions from the original maize-rice orthologous CNS studies (INADA et al. 2003, KAPLINSKY et al. 2002) have been replicated—in general—in the two large scale homeologous CNS studies in *Arabidopsis* (FREELING et al. 2007, THOMAS et al. 2007) and rice (LIN et al. 2008). These α CNS (defined in Table1) data demonstrate that, in plants, α CNSs average from 20-30 bp in length. α CNSs are predominantly found in close proximity to one and only one gene pair.

In an effort to associate CNSs with function, genes, gene names or their Gene Ontology (GO) terms have been quantified for CNS-richness. To date, all studies have concluded that “regulatory genes” are CNS-rich, with genes encoding transcription factors generally being more CNS-rich than genes encoding protein kinases, and much richer than genes encoding ancient structural functions like ribosomal subunits, motors, or energy metabolic pathways. In one study of arabidopsis α CNSs (FREELING et al. 2007), -- those homeologous CNSs retained from the most recent tetraploidy-- 246 gene pairs occupied an exceptionally long stretch of chromosome, where the regions full of CNSs were conspicuous in being exon voids, sometimes called "gene deserts" in the mammalian literature. For Arabidopsis $> 4\text{kb } 5' \text{ plus } 3'$ of CNS-defined genespace "exceptionally long," but this definition does not apply to all genomes. These genes were called “Bigfoot” and tended to be annotated with “response to...” GO biological function categories, these often being “transcription factor activity” as well. The distance between Bigfoot gene 5' CNSs and the start of exon1 averaged 3.1 kb. This same study found a statistically significant over-abundance in CNS sequence of the most famous

transcription factor binding site in plants, the G box palindrome (CACGTG). Several other known motifs, plus previously unknown 7mers, were significantly over-represented, but none to the extent of the G-box and Gbox-like motifs. CNSs in both Arabidopsis and rice occur 5' and 3' to genes and in introns, and are skewed toward the 5' end (LIN et al. 2008). Some intron CNSs, like those found inside the grass *knotted1* gene, are noncoding (INADA et al. 2003), while some are certainly un-annotated exons expressed in alternative transcripts. In general, intron CNSs are suspect, especially those that are touching an annotated exon; they are likely un-annotated exons. The data for the 5' and 3' plant CNSs supports their role-- in general-- as binding sites for gene regulatory products.

Most CNSs contain known transcription factor binding sites, and CNSs generally are enriched in several such motifs. Most plant genes function in developmentally complicated ways in the absence of CNSs. It must be understood that known transcription factor binding motifs tend to be short, far too short to be detected by themselves as sequence, with the minimum CNS being a 15/15 exact match. CNSs may well only detect clusters of sites, reiterated binding sites or other macrostructure associated with particular sorts of genes, about which we know little.

Plant versus vertebrate CNSs

Mammalian CNS research, now called CNE research, began years earlier than comparable research in plants (HARDISON et al. 2000). Among the early conclusions derived from 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 (INADA et al. 2003, KAPLINSKY et al. 2002). This clustering of CNSs around gene pairs supported the feasibility of sorting a particular CNS to a particular gene using proximity alone, an activity not possible to do in man-mouse comparisons. Such assignments are arbitrary, and some of these assignments are expected to be proved wrong, since some CNSs are expected to encode functions involving more than one gene in a chromosomal domain. Additionally, the most extreme animal CNSs are more conserved than are the most conserved plant CNSs. Several hundred ultraconserved elements (UCEs), deeply conserved vertebrate CNSs, have been found to be identical or near-identical for 200-500 bp in human, mouse, and

rat, with a majority of them showing on the order of 96% identity with birds (chicken) that diverged from mammals approximately 310 MYA (SAKURABA Y et al. 2008, BEJENARO et al. 2004). Plant CNSs are short, and also may mutate relatively quickly over evolutionary time, although the binding functions themselves may well be conserved. This problem of detecting divergent functional homologous binding sites by sequence only is called "binding site turnover" in animals (MOSES et al. 2006). It is not yet clear whether or not binding site turnover mechanisms in animals apply to plants.

Some known, characterized cis-acting binding sites are CNSs, cases like that of the grass *knotted1* big intron CNS cluster already discussed (Fig. 2). These have been reviewed (FREELING et al. 2009, Table 2).

Experiments that address the possibility that some CNSs are artifacts, mutation coldspots, gene conversions, misannotations and the like.

Because CNSs cluster around genes that are "response to" transcription factors -- these are very often bigfoot genes -- and because CNSs are enriched in known DNA binding motifs the indication is that CNSs as defined are positively correlated with cis-acting, functional sequences that bind regulatory molecules. However, that does not mean that each CNS is real or even that the average CNS is real. Artifacts are possible.

There is evidence against CNSs being mutational coldspots in animals (KIM et al. 2007, DRAKE et al. 2006), but there is no such evidence for plants. However, it seems unlikely that replication or repair processes should be error-free. If one sequence copies over to another syntenic, homologous sequence -- that is, suffers a gene conversion event over a short stretch of paired DNAs -- then an artifactual CNS could be created essentially at any time in evolutionary history. Recent studies found many regions of gene conversion within the rice genome (WANG et al. 2007, XU et al. 2008), including approximately 6 mb of near-identical sequence at the ends of homeologous chromosomes 11 and 12. Since such gene conversions are not understood, they should not be discounted as possible sources of "CNSs" that have no function. Another source of artifactual CNSs are unannotated RNA genes (like *MIR*, *Ta-si-RNA* or targets for siRNAs, Lee et al. this volume); analyses of current data indicted that 1.5% of CNSs can be explained in this way (THOMAS et al. 2007). Finally, CNSs could be undiscovered, small genes or exons that are used differentially. New

genes are being discovered continuously, and any of these could conceivably be erroneously interpreted as a CNS. Proteogenomic studies in *Arabidopsis* resulted in the expansion of exons for 2446 TAIR7 models and called 838 new, usually short, genes (CASTELLANA et al. 2008). We back-translated each of these unexpected peptides using tblastn and compared all sequences to each published *Arabidopsis* α CNS. None of the homeologous CNSs identified within *Arabidopsis thaliana* could be translated into any sequence on this new protein list. Even so, CNSs, and especially intron CNSs adjacent to an exon, not only could but also should turn out to be protein coding. One of us (Freeling, unpublished data) has data from blastn analyses of intron CNSs to *distantly related* plants: some CNSs are conserved in such a way as to be best explained as alternatively spliced exons that are not yet annotated as such in *Arabidopsis*. Because of the purely computational nature of the average CNS's identity, there is always a chance that it will be artifactual. However, that the *bulk* of CNSs are functional binding sites is supported by the fact that CNSs are located near regulatory genes and especially "responds to" transcription factors, and that CNSs are enriched for known DNA-binding motifs. Those two results are not expected of any of the artifactual or trivial explanation.

CNSs can exist for any duplicate DNA segment.

It is possible to find CNS-like sequences among the noncoding regions of some tandem duplicates (unpublished). Unfortunately, there is no perfect way to know when the tandem array occurred since gene conversions are well-known in duplicate arrays (GAO et al. 2004) and selection within tandem genes is expected to be reduced. For orthologous genes-- between rice and sorghum for example-- every gene diverged from its ortholog at the same time. Using similar logic, all alpha homeologous pairs in *Arabidopsis* happened contemporaneously. For a orthologous gene-pairs dataset, each gene's divergence frequency for any endpoint is comparable to the frequencies of other contemporaneous pairs. There is no contemporaneous control for tandem genes or segmental duplicates. Even so, it might well be possible to control for CNS discovery among these duplicates.

CNS-rich genes, bigfoot genes, are particularly enigmatic

Genes high up in the regulatory cascade, especially those that respond to external stimuli, tend to be CNS-rich, while genes that encode housekeeping 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 regulatory genes are also the most regulated*. When faced with enigma, metaphors are useful to some of us. Here is a metaphor from the market. From the company's point of view, the CEO is at the highest control level and the worker is near the lowest. However, ask any CEO about her job and hear about how changes in unions, markets, investor sentiment, government regulations, tax laws, unexpected lawsuits and a myriad of additional connections to "the system" limit and often over-ride any control the CEO might hope to exert on behalf of company success. Perhaps the genes at the highest control level, like CEOs, are also those under the most control themselves. Bigfoot genes may be these enigmatic CEO genes.

Fresh data on how CNS richness influences the ability of a gene to be retained post-tetraploidy.

As seen in Figure 1, there is a complex phylogenetic relationship between the pre-grass tetraploidy, the radiation of grasses, and the tetraploidy that happened just after the maize and sorghum lineages split. The rice, sorghum and maize genomes are sequenced. For rice, it was possible to obtain a list of orthologous pairs with sorghum, each with a CNS collection obtained from the CNS Discovery Pipeline v1.0, and the rice genome can also be organized into two homeologous, highly fractionated genomes anchored by their homeologous pairs. Thus, it was possible for Schnable and coworkers (SCHNABLE et al. 2011) to ask whether CNS richness of rice orthologous pairs correlated with having been retained after the pre-grass tetraploidy. Figure 5 (reprinted here) shows these data. As sorghum-rice orthologous CNS-richness increases, the chances of having been retained as a pair from the last tetraploidy increases. When the ">15 CNSs/gene" bin of Figure 5 is subdivided, retention levels go up. Six of the 15 rice-sorghum gene pairs with >28 CNSs possess a retained homeolog (40% retention) and 25 of the 56 gene pairs with 22-28 CNSs possess a retained homeolog (45% retention).

This result was not expected on the basis of prevailing theory on why genes are retained post-tetraploidy. Based on the robust gene balance hypothesis ((BIRCHLER et al. 2010)) generally applicable reason that genes are retained as pairs post-tetraploidy is that , were one of the duplicates removed, dosage imbalance would ensue and fitness would go down. Thus, for those gene networks where dosage balance is functionally preferred, purifying selection maintains the status quo by preventing fractionation of the pair. Most genes lose one or the other duplicate; that is, most genes fractionate. The consequences of gene dosage theory on gene and genome duplication have been reviewed at length (EDGER et al. 2009, FREELING et. 2009, SEMON et al. 2007). In general, selection for gene dosage balance is usually envisioned at the level of gene product balance, as in the concentration of protomeric proteins assembling into larger complexes. The most advanced theoretical treatment of gene dosage sensitivity (haploinsufficiency) is at the protein assembly level (VEITIA, 2010), and this may explain why genes encoding components of ribosomes, motors, and proteosomal cores are over-retained post-tetraploidy. Since transcription factors may participate in complicated aggregates, enhancosomes (LEVINE, 2010), the dose-sensitivity of genes encoding transcription factors might well be explained at the protein-protein level as well. However, the data of Figure 5 suggest an alternative explanation. Some genes have binding sites in their genespace, and for some of these genes the concentration of binding sites may be just as important as the concentration of protein that binds them.: A bigfoot transcription factor gene, for example, might be dose-sensitive not because of any optimal protein-protein stoichiometries involving the gene's product, but because of DNA-protein stoichiometries being selectively important on its own flanks.

The newly sequenced maize genome (SCHNABLE et al. 2009) permitted an experiment that begins to utilize the analytical power of phylogenetics. Rice homeologous pairs of bigfoot genes were analyzed for orthologous sorghum-rice CNS content. There were hundreds of cases where one of the homeologs had more CNSs than the other. There was a significant tendency for the homeolog with the greater CNS count to be retained in the maize lineage tetraploidy. Clearly, the differential CNS loss must have occurred in the branch shared by all grasses. These data (SCHNABLE et al. 2011) make sense if CNS-richness is directly related to gene dosage sensitivity, but this

explanation is not proved; CNS-loss *could* switch the expression profile per organ, region or time leading to an avoidance of dose-sensitivity in the pregrass branch of the phylogenetic tree. In the absence of dose-sensitivity, loss is expected. More research is necessary on the molecular biology of gene dose sensitivity.

Conclusions

Plant CNSs often function as cis-acting regulatory binding sites, but published CNSs lists undoubtedly contain artifacts of a number of types. Genes that are particularly CNS-rich and take-up a lot of chromosomal space -- bigfoot genes -- tend to be in a "responds to..." GO category and also tend to be transcription factors. These long genes are of particular interest. They are retained preferentially following ancient tetraploidies, and are retained less often if they lose CNSs over evolutionary time. This conclusion is from an experiment that marks the beginning of CNS-driven phylogenetic research. The relationship between CNS-richness and gene dosage sensitivity is complex, and begs for continued research.

The first paper on plant CNSs (KAPLINSKY et al. 2002) suggested that CNSs may be particularly close arrays of transcription factor binding motifs that function as do enzymes, to mechanically facilitate the binding of proteins into enhancer (or insulator or transcription-factor) complexes. That idea, in the absence of relevant data, seems about as useful today as it was over a decade ago.

Deciphering the meaning of CNSs is an important part of understanding the language of gene regulation.

Figures

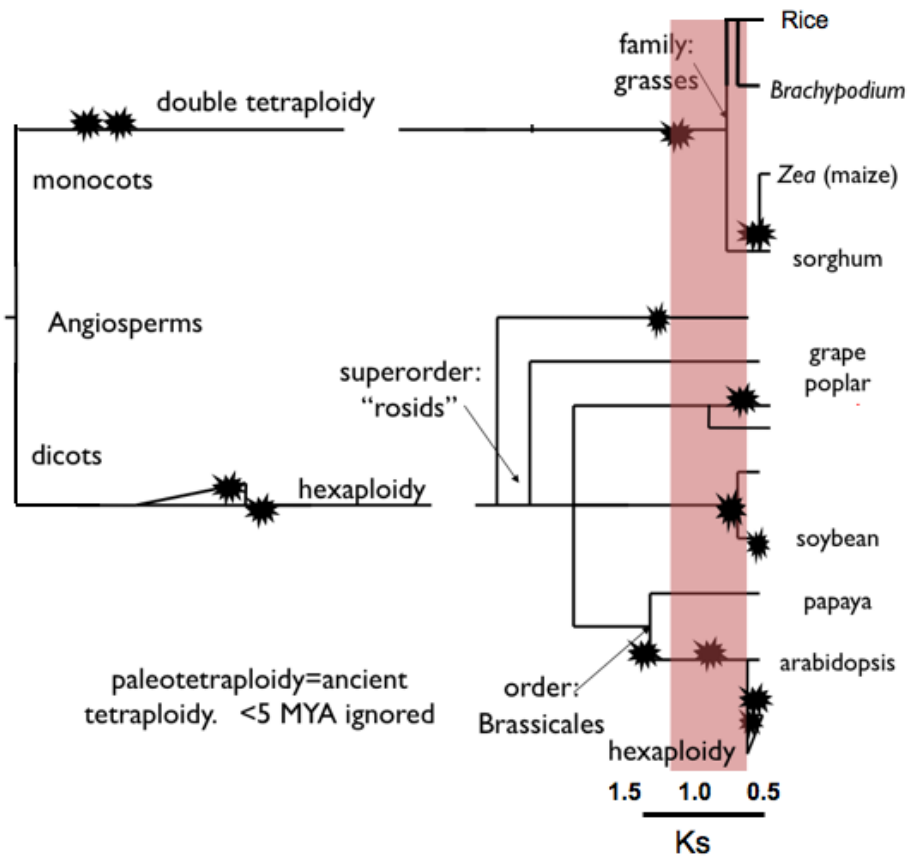


Figure 1. Heavily pruned phylogenetic tree of plants with those genomes sequenced and released at the tips. The colored bar refers to the window of useful divergence for CNS discovery.



Figure 2. GEvo panel comparison of orthologous regions of *Sorghum bicolor* (top panel) and *Oryza sativa* (bottom panel). Red bars indicate High scoring segment pairs (HSPs) between the two regions, some of them shown connected with green lines. HSPs called as CNSs are annotated in purple bars drawn below the HSP graphic in the *Sorghum* panel. This result can be regenerated at <http://genomevolution.org/r/a51s>.



Figure 3. One extreme example of the mutant phenotype conferred by dominant *knotted1* (*kn1*) mutants. This plant has grotesque finger-like projections of more-or-less normal leaf coming off its outer surface. These failures to stop dividing result from the ectopic expression of KN1 protein in leaves, an organ where this gene is normally not expressed. The Hake laboratory showed that phenotypes of this general type were made by multiple transposon insertions into specific regions of the *kn1* big intron [GREENE et al. 1994] containing number of clustered CNSs [INADA et al. 2003].

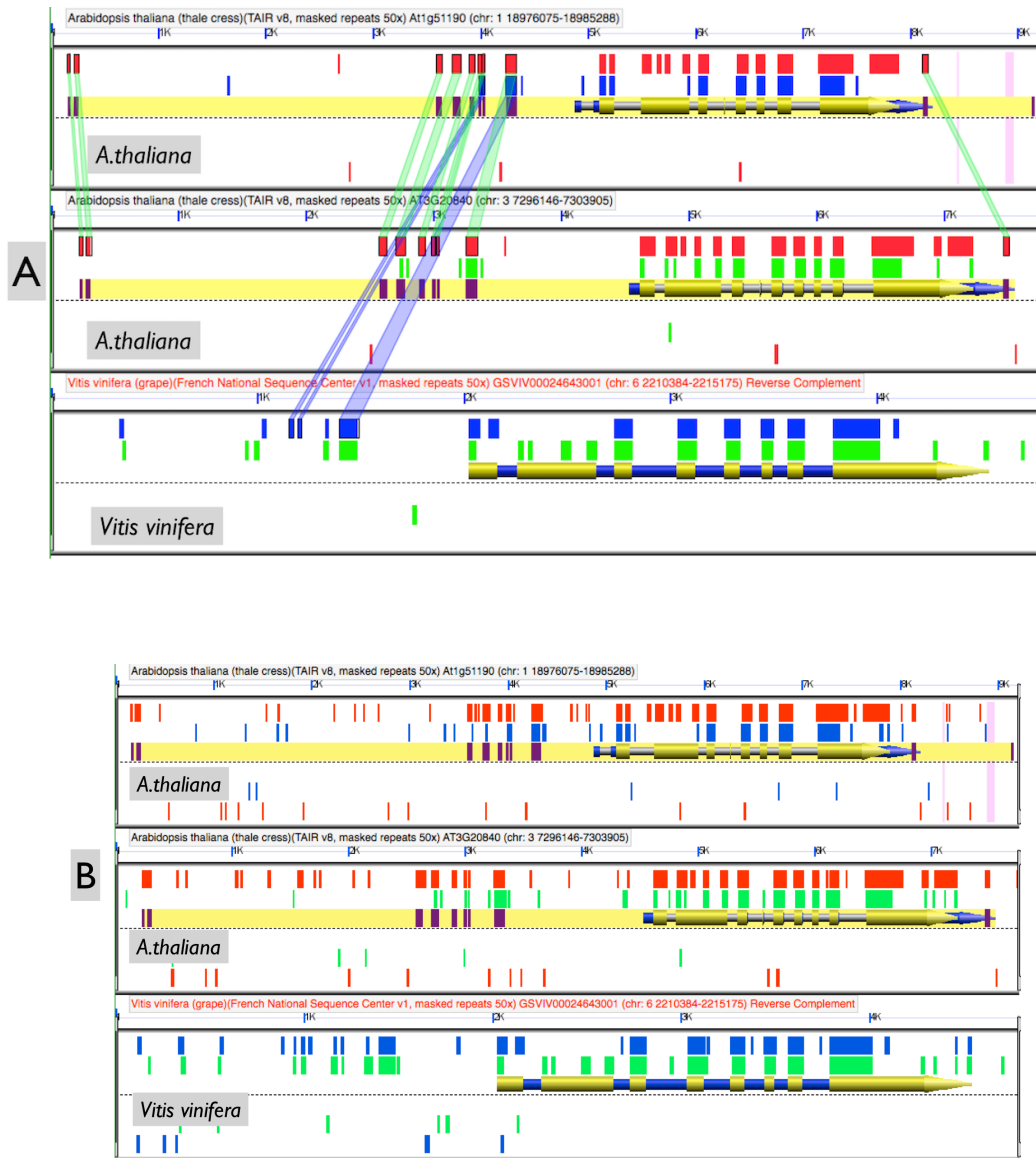


Figure 4. GEvo panel comparison of homeologous regions of *Arabidopsis thaliana* (top two panels) and corresponding orthologous region in *Vitis Vinifera* (bottom panel), A. Using blastn at optimal CNS discovery settings (Table 1) B. similar settings with a slightly lowered e-value cutoff. Red bars indicate High scoring segment pairs (HSPs) for each pairwise comparison, some of them shown connected with purple lines. HSPs called as CNSs are annotated in purple bars drawn below the HSP graphic in the *A. thaliana* panels.

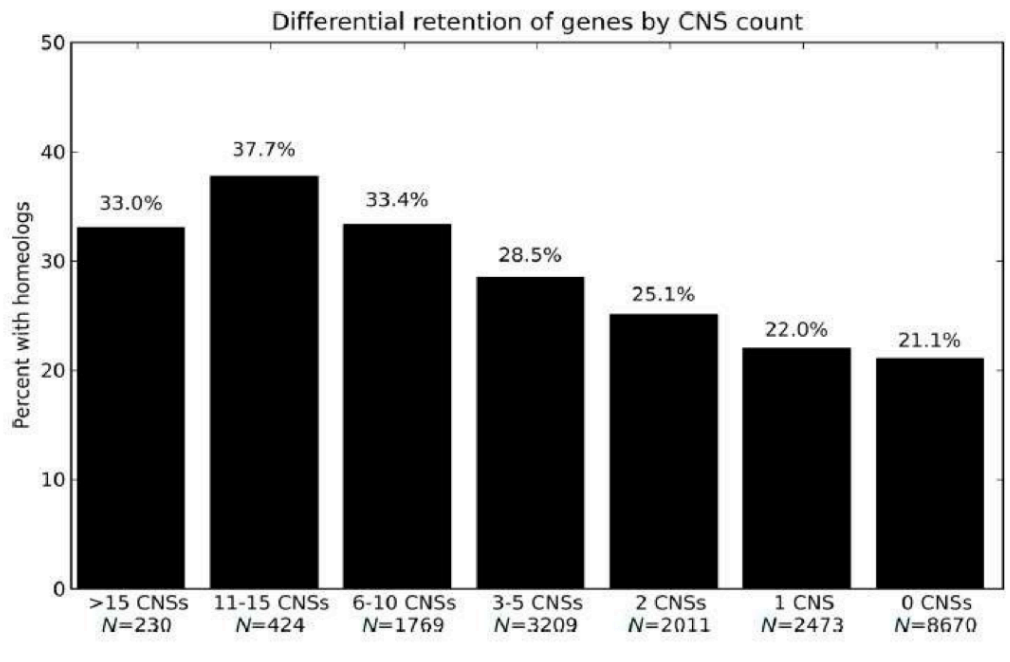


Figure 5. Relationship between CNS richness and retention.

Tables

Phylogenetic footprint	The most inclusive term for sequences that tend to stay the same over evolutionary time, while sequences within the genome known to be without specific function are becoming randomized. Thus, a phylogenetic footprint often marks a sequence with function.
Plant CNSs (conserved noncoding sequences)	Plant CNSs, a subset of phylogenetic footprints, have a set definition that permits quantitative comparisons between "usefully diverged" homologous genes in all sorts of plants. A bl2seq (blastn-2-sequences with default settings; TATUSOVA and MADDEN, 1999) hit between the nonprotein-coding sequences near <i>usefully diverged</i> , syntenic (either orthologous or homeologous) genes. A CNS pair must have an e-value equal to or more significant than a 15/15 exact nucleotide match (KAPLINSKY et al. 2002; INADA et al. 2003) and must be syntenic (FREELING and SUBRAMANIAM, 2009). "Usefully diverged" is of paramount importance, and will be independently defined in this glossary.
Plant α CNS and orthologous CNSs	Orthologous CNSs, or just CNSs, may usefully be distinguished from paralogous CNSs, such as those generated by polyploidy. In <i>Arabidopsis</i> , these are rare (Thomas et al. 2007).
"usefully diverged"	In order to predict function, a pair of bonafide CNSs needs to have diverged for a long enough period of time to let sequences randomize or delete if not under selection. Otherwise, conservation is nothing but neutral carryover. For biological reasons, if divergence continues for too long, detection by sequence similarity becomes difficult. This "usefully diverged" window in plants is between approximately Ks 40% and 90% where Ks is the rate of synonymous base pair substitution in codons, used here as a measure of divergence time only.

Table 1. Definitions involving CNSs and their identification in plants

Type	Function associated with plant CNSs	Citation
G	Grass regulatory genes are rich in orthologous CNSs	(INADA et al. 2003, GUO et al. 2003)
G	<i>Arabidopsis</i> genes that are induced by stimuli and/or encode transcription factors are rich in homeologous CNSs, and are often “Bigfoot” genes. <i>Japonica</i> rice homeologous CNSs are similar.	(THOMAS et al. 2007, LIN et al. 2008)
G	Grass regulatory genes are rich in orthologous CNSs	(INADA et al. 2003, GUO et al. 2003)
G	<i>Arabidopsis</i> genes that are induced by stimuli and/or encode transcription factors are rich in homeologous CNSs, and are often “Bigfoot” genes. <i>Japonica</i> rice homeologous CNSs are similar.	(THOMAS et al. 2007, LIN et al. 2008)
G	<i>Arabidopsis</i> homeologous CNSs are significantly enriched for several known transcription factor binding motifs, especially the G-box	(FREELING et al. 2007)
S	Intron CNSs in a Class I homeobox gene (<i>kn1</i>) bind a negative <i>cis</i> regulator, a binding that is disrupted by <i>Mu</i> transposons, but only in <i>Mu</i> -active lines.	GREENE et al. 1994, INADA et al. 2003)
S	5' CNSs contain conserved, known transcription factor binding motifs and motif-patterns: <i>RAB16/17</i> in grasses, <i>rbc alb</i> in dicots, and **proximal promoters in dicots.	(BUCHANAN et al. 2004, WEEKES et al. 2007, VANDEPOELE et al. 2006)
S	Two 5' CNSs of the shoot meristemless (STM) homeobox gene in dicots and monocots, and their binding motifs, <i>cis</i> -regulate repression/re-establishment of leaf expression.	(UCHIDA et al. 2007)
S	A QTL for flowering time in maize mapped to a <i>cis</i> -acting, regulatory grass CNS 70 bp upstream of an <i>Ap2</i> -like gene.	(SALVI et al. 2007)
S	Some 5'UTR grass CNSs are uORFs, one mechanism to down-regulate translation	(TRAN et al. 2008)

Table 2. CNSs characterized as *cis*-acting binding sites, from Freeling and Subramaniam 2009. G=General. S=Specific ** Vandepoele and coworkers 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.

Chapter 2: The fate of *Arabidopsis thaliana* homeologous CNSs and their motifs in the paleohexaploid *Brassica rapa*

The following chapter has been published as a peer reviewed article in Genome Biology and Evolution. Again, outline your specific contributions

Subramaniam S, Wang X, Freeling M, Pires JC. (2013) “The fate of *Arabidopsis thaliana* homeologous CNSs and their motifs in the paleohexaploid *Brassica rapa*” Genome Biology and Evolution, 5: 646-60. PMID: 23493633.

Introduction

A perplexing and long-standing problem in classical genetics is to know when a recessive mutant specifies a complete knockout of function. Even the sequence of mutants with recessive phenotypes compared to the wild type progenitor may not answer the question of functionality. If the mutation happened during evolution, inferred from comparisons of mutant with a more ancestral outgroup, it is even more difficult to predict functionality. Because of the history of paleopolyploidy in all plant lineages (Van de Peer, 2011), updated at http://coge.iplantcollaborative.org/wiki/index.php/Sequenced_plant_genomes, and the consequent potential for functional redundancy, duplicate genes or regulatory site sequences mutate into nondetectability. This post-polyploidy gene loss, called fractionation, is widespread and frequent. It's important to know if such “loss” results in loss of function. One way to show that a loss of sequence detectability is a loss-of-function is to show that the loss is by deletion of sequence, since a deleted sequence cannot function.

When a genome doubles or triples, as with paleotetraploids or hexaploids, each chromosome with each gene is initially duplicated. What follows is a process of chromosomal evolution called “diploidization,” during which the polyploid becomes rearranged and altered to act as a meiotic diploid (WOLFE, 2001). The newly diploidized polyploid tends to lose one or the other of its duplicate genes (fractionation), usually much of the time, as expected in theory (LYNCH and FORCE, 2000), and realized in practice (SANKOFF et al. 2010). The fractionation mechanism is a sort of intrachromosomal recombination inferred from short repeats flanking progenitor deleted

sequences (PETROV et al. 1996, DEVOS et al. 2002), and is known for post-paleotetraploid maize (WOODHOUSE et al. 2011) and post-paleohexaploid *Brassica rapa* (*Br*) (TANG et al. 2012). Even if a gene pair survives polyploidy, perhaps because of subfunctionalization (LYNCH and FORCE, 2000) or tendency to maintain product dosage balance (FREELING, 2009), that does not mean that all parts of the gene will remain duplicated. This study follows individual conserved noncoding sequences (CNSs) known to exist around many arabidopsis genes as they now exist in *Brassica rapa* (*Br*), a hexaploid. Figure 1 follows one ancestral Brassicaceae gene as it gets duplicated during the alpha paleotetraploidy, and then follows as each alpha homeolog splits into the lineage that will be arabidopsis (*At*) or *Brassica rapa* (*Br*), and then through the *Brassica* lineage, on through the paleohexaploidy, and finally follows the genes into the six potential chromosomal positions on the three *Br* subgenomes. Sometimes a *Br* gene is fractionated and takes all of its *At*-orthologous CNSs with it, but sometimes the duplicate transcriptional unit and its *cis* sequences persist. In such cases, sometimes the *At* CNS being followed goes undetectable but the gene remains and is transcriptionally active. This has been shown previously in grasses (SCHNABLE et al. 2011). The red arrow on Figure 1 denotes such a CNS loss. The small squares decorating the gene models of Figure 1 are CNSs.

The mechanism of CNS fractionation in plants has not been studied previously, although it is known that plant CNSs lose detectability as divergence time increases (REINEKE et al. 2011). This mechanism is important because several CNSs have been shown to function as cis-acting regulators and are enriched in known DNA-binding motifs (FREELING and SUBRAMANIAM, 2009; RAATZ et al. 2011), they are associated with DNaseI open chromatin (ZHANG et al. 2012) and with the suppression of gene expression (SPANGLER et al. 2012). Thus, CNS loss of detectability could predict loss of a specific regulatory function, but only in the case that the CNS loss marks loss of CNS function.

Crucifer CNSs in arabidopsis have a history. Previous work (THOMAS et al. 2007) found 14,944 conserved noncoding sequences (alphaCNSs, α CNSs, *At-At* CNSs) retained following the most recent tetraploidy in the *Arabidopsis thaliana* (arabidopsis, *At*) lineage.

Genes retained as pairs following this tetraploidy, called homeologs (or homoeologs, Ohnologs, syntenic paralogs), have diverged a modal 0.76 in synonymous base substitution rate (Ks), and this was shown to be an adequate evolutionary divergence proxy to ensure that associated CNSs avoided purifying selection because CNSs on average, functioned. When divergence times become too great (>0.9 modal Ks), detection of CNSs becomes difficult, and when there is too little divergence, or when the detection algorithm settings are set without regard to noise levels (KAPLINSKY et al. 2002; Thomas et al. 2007; LYONS and Freeling 2008), CNSs no longer indicate putative conserved function.

We know enough about the genome of *Br* to make some predictions. The three ancestral genomes of the new *Br* hexaploid do not remain intact for long. Fractionation soon removed most of the redundant duplicated genes (WANG et al. 2011) and is predicted to have removed some duplicate CNSs as well. Thus, each CNS in *Br* is retained as a singlet, a doublet or a triplet depending on whether or not its gene is retained, and if its gene is retained, depending on whether or not the CNS itself remains detectable. Some go undetectable, as with the CNS position at the tip of the red arrow of Figure 1. The background "neutral" base substitution rate between arabidopsis and *Br* orthologs (0.38, legend Figure 1) will tend, in theory, to substitute nucleotides in CNSs that contribute little or nothing to CNS function. The α CNSs of arabidopsis contain sequences that come with varying blastn e-values and lengths down to 15 bp. Since plant CNSs contain DNA-binding motifs (FREELING and SUBRAMANIAM, 2009) as they do in mammals (VON ROHR et al. 2007; PENNACCHIO et al. 2007), motifs known to be short and inexact (7-12 bp with alternatives), some of the *At* α CNSs should not be detected in *Br* even though they might contain functional motifs because the motifs are shorter than the minimal length of detectable CNSs. In short, if base substitution were the prevailing mechanism of going non-detectable, then CNSs could drift into non-detectability and still conserve typical, functional DNA-binding motifs. However, no matter what the mechanism of non-detectability, selection for functional loss should be greater for a CNS on a singleton gene, the CNS being more unique-sequence, as compared to CNS on each of a doublet or triplet *Br* gene, the more redundant situation. This inference is the basis of our

essential strategy; see the second footnote of Table 1 where our strategy is applied to real data.

If a CNS is undetectable using our standard `blastn` criteria, we use computational methods to deduce the preponderant mechanism of each CNS's mutation from the ancestral sequence, detailed in the methods section. Since much is known about particular G-boxes within CNSs (FREELING et al. 2008), based on previous work on this motif and transcription factors that bind some G-boxes, we study how G-boxes become nondetectable as well. Our categories of loss: (1) base substitutions (the pseudogene pathway), (2) proved deletions (removal of one or both flanking markers as well as the CNS), (3) computationally inferred deletions, (4) indels, or (5) large insertions. Indels have been recently identified as a significant mutational endpoint in plants (HOLLISTER et al. 2010).

Understanding how plant CNSs go undetectable is important for several reasons. In animals, explanations have been proposed for how undetectable enhancer sequences sometimes retain function, including ‘binding site turnover’ (HANCOCK et al. 1999; LUDWIG et al. 2000; DERMITZAKIS and CLARK, 2002) and “dormant TF binding sites” (JUNION et al. 2012). These mechanisms require nonfunctional sequences drifting along the pseudogene pathway before they mutate back to function. Such mechanisms become less likely as nonfunctional DNA is deleted more quickly in plants. Additionally, knowing the mechanism of CNS fractionation is particularly important in light of a genetic-type method we have proposed called “fractionation mutagenesis” (FREELING et al. 2012). This method quantifies the RNA levels of duplicate genes in a polyploid and also compares each homeolog's CNS-loss pattern. A mutant expression pattern is then associated with a lost CNS or a cluster of CNSs, and these previously mysterious sequences acquire a testable hypothesis as to ancestral function. We will show that this method of fractionation mutagenesis comprises a natural “deletion machine” 84% of the time in the post-hexaploidy *Brassica* lineage, and deletion mutations are certainly loss-of-function.

Results

At-At α CNSs updated to Version 2

The updated version 2 α CNSs list (Supplement Table 1) now contains 11,448 sequences, or 5,724 α pairs (α pairs are homeologous pairs derived from the most recent whole genome duplication event in the lineage of *Arabidopsis thaliana*). These α CNSs were used to search for retention within *Br* at orthologous loci. As a control for our CNSs discovered through manual comparison of homeologous regions in *At* (v2), we ran our automated CNS Discovery Pipeline 3.0 (https://github.com/gturco/find_cns/tree/master/pipeline) over our homeologous gene pairs to generate an automated *At-At* α CNS dataset (Supplement Table 2). There is 80% concordance between the automated and manually generated CNS datasets; the v2 data set was used in this study. The α CNSs, (both v2 and the pipeline 3.0, for comparison) have been added to the gene models of TAIR8 in CoGe as genome dataset ID=39598. Our Supplemental Table 1 includes links to GEvo in CoGe using these customized genespace models, thus facilitating reproduction and proofing of our results. Figure 2 (<http://genomeevolution.org/r/4db1>) shows GEvo blastn output graphic, where the query is arabidopsis *At1G75520*, a bigfoot gene encoding a RING zinc finger protein of unknown function, displayed with its corresponding α -homeolog. Both manually curated (v2, color-coded purple) and automated pipeline 3.0-generated α CNSs (color-coded green) are annotated on this graphic along with the blastn HSPs (high-scoring segment pairs; color-coded orange) corresponding to regions of high sequence similarity between the homeologs.

General features of detectability of *At-At* CNSs in *Br*

The paleohexaploidy in the *Br* lineage generated three subgenomes, with one of them (subgenome III) having almost twice as many genes as either of the other two. Genome dominance and purifying selection explain this phenomenon, using the exact same argument that was proven valid in maize. In terms of CNS detectability, we expect the dominant subgenome (III) to carry most of the genes that are singlets, and subgenomes I and II to have endured the most gene and CNS loss.

For each of the 6330 *At* genes with a retained *At* α -homeolog, each used for CNS discovery, we used our synteny screening blocks technique (TANG et al. 2012) to identify all possible orthologous

regions in the *Br* genome. Given that the recent hexaploidy in *Br* occurred following divergence from the arabidopsis lineage (Figure 1), we expected to find up to three orthologous copies for each *At* gene. Figure 3 is a GEvo graphic (<http://genomeevolution.org/r/4db6>) of the same bigfoot gene shown in Figure 2, this time showing blastn hits to the three detected orthologous regions within *Br*. The top panel shows the aCNS-rich *At* gene (*AT1G75520*), a member of SHI ring Zn-finger gene family, and the three panels below show its detectable orthologs in *Br*. aCNSs (v2, purple bars) and the gene space (yellow background) are annotated on the *At* gene panel. HSPs between the *At* gene and each *Br* subgenome ortholog is annotated as *Br* I, *Br* II and *Br* III (Figure 3). Analysis of the overlap of aCNS positions with corresponding HSPs to each of the *Br* orthologous positions in Figure 3 gives insight into the detectability of aCNSs in each of the three *Br* regions. One of the *At-At* CNSs (315; 5;CNS_s677) shown in Figure 3-- highlighted in grey-- has corresponding HSPs in *Br* II and III subgenomes, but has an undetectable ortholog in *Br* subgenome I. Another CNS, highlighted in pink (315; 8;CNS_s680) has corresponding HSPs only in *Br* II subgenome and has undetectable orthologs in *Br* subgenomes I and III.

Of the 16330 *At* genes used in *At-At* CNS discovery, we identified at least one *Br* ortholog for 6245 *At* genes, with 2391 *At* genes having a single detectable orthologous copy in *Br* (singlets), 1723 *At* genes with two orthologous copies in *Br* (doublets) and 654 *At* genes with three orthologous copies in *Br* (triplets). We expected to find, in the absence of mutation, an α CNS whenever its gene was present. So, each gene in a doublet or triplet would have an expected CNS. Based on *At-Br* orthologies, we expected to find 9179 CNSs within the expected orthologous positions, 3882 as singlets, 3678 as doublets and 1619 as triplets. Our detectability results are in Supplemental Table 2. Many mutations to nondetectability occurred.

α CNS length vs. detectability

The version 2 CNS collection includes CNSs as short as 15 bps and as long as 283, and each has an e-value more significant than that of a 15/15 exact nucleotide match. Even one base substitution would render some of these sequences undetectable using our blastn settings, so we expected that detectability would increase with length, and it did.

α CNS length versus detectability was plotted for all version 2 α CNSs. Figure 4-A shows these data for *Br* ortholog singlets, doublets and triplets. In general, detectability is greater in singlets than doublets than triplets, as expected from our previous results and our general understanding of purifying selection and CNS redundancy. For singlets, detectability increases from 40% for 15-19bp to 96% for >76 bp, with the 31-40 bp bin being 85% detectable. For the bin 51-75 bp, detectability was 62, 72 and 91% for triplets, doublets and singlets, respectively. We chose those 2509 α CNSs that are >30 bps in length to analyze further as to the molecular mechanism of their loss of detectability.

Pooling all α CNSs that are 31 bases or longer, we compared the degree of detectability in *Br* as a measure of the number of expected orthologous copies; we compared singlets with doublets with triplets. Figure 4-B: each α CNS is localized to one of the nine “categories” of *Br* genome: singlet subgenome I, singlet II, singlet III, doublet I, doublet II, doublet III, triplet I, triplet II and triplet III. Figure 4-B includes numbers of genes in each category, and probabilities that particularly interesting differences are significantly different. Detectability for CNSs on singlet genes is generally greater than that for doublet or triplet; that is expected because it should be more difficult to remove a singlet CNS without removing function. Detectability of CNSs on singlet genes of subgenome III is nearly 100%, and is significantly more than detectability of singlets on subgenomes II and I. There is certainly subgenome bias in the detectability of CNSs. This interesting result is not easy to explain, is probably important, and will be discussed.

The functionality metric: Deletion to loss-of-function is the primary mechanism for removal of α CNSs >30 bp long in *Br*

Having located the stretch of chromosome in *Br* where the missing CNS could be, we devised a global alignment algorithm, global-npe, to identify the predominant mechanism of removal of CNSs. Table 1, Column 1 lists these predominant mutational causes for the failure to detect a CNS. For each mechanism, impact on detectability *within singlet genes* was used as a control, and recorded as data in Column A of Table 1. We then sorted the 498 redundant (doublets and triplets), undetectable CNSs into mutational mechanism categories, and

recorded these data in Column B. B/A is the functionality metric, with a fully functional CNS category scoring 1, by definition. B/A ranged from 5.3 (nonfunctional deletions) to 1 (fully functional). The functionality metric is useful. Since deleted DNA cannot be functional, we now know that computed deletions have a 64% chance of being nonfunctional. Base substitutions, however, have a 91% chance of still being functional (but there are not many CNSs that have gone undetectable for this reason). Insertions may destroy detectability, while function is almost always maintained. Overall, the average nondetectable, once-redundant CNS has a 72% chance of being nonfunctional, and that makes sense because 97% of them sort to a molecular mechanism with “deletion” in its description. Deletion is the predominant mutational mechanism for the lack of detectability of α CNSs in *Br*, but other mechanisms operate as well.

Although our functionality metric differences imply that genes with undetectable CNSs generally function, it is more rigorous to test directly to see if the loss of CNSs is somehow correlated with the loss of gene function. Cheng and coworkers (2012) published RPKM values in leaves, stems and roots of seedling *Brassica rapa* Chiifu; our subgenomes I, II and III are their subgenomes MF1, MF2 and LF. Using two different cutoffs for potential gene death, there was no correlation between loss of CNSs and potential gene death (Supplemental Table 3). There was a slight tendency for subgenome I to have more dead genes than other subgenomes, and, as expected, the more stringent cutoff found fewer (ca. 7%) potentially dead genes as compared more potentially dead genes (ca. 17%) for the permissive threshold. Conclusions from this control experiment: CNS nondetectability or even total CNS loss is not correlated with gene death, so there is no need to modify the predictions of the functionality metric of Table 1. Note that three organs in one environment do not monitor all of the possible expression endpoints, so the frequencies of genes that are actually dead are definitely below these potential death values (Supplemental Table 3).

Detectability of CNS-enriched transcription factor binding site (TFBS) motifs, especially the G-box, in *Br* orthologous positions

TFBS motifs contained within arabidopsis α CNSs, and enriched >2-fold within CNSs, were studied without regard to whether or not their

CNSs were detectable in *Br*. They were detected as an exact match in the expected genespace region. Were any such motif in arabidopsis lacking function in the *Brassica* lineage, base substitution alone (*At-Br* Ks=0.38) would likely lead to nondetectability: a 5mer would become undetectable 91% of the time, and a 6mer, like the G box, would go undetectable 95% of the time if base substitution were the only mutational mechanism operating (which is certainly not the case). Supplemental Table 4 updates α CNS enrichment data from 2007 (FREELING et al. 2007)-- using Version 2 CNSs. For our TFBS detectability and enrichment studies, we included all 11,448 *At* α CNSs, not just those >30 bp long. We studied 12 motifs, and 8 of them contained the 5'ACGTG core (colored red in Suppl. Table 4); this core is part of the G-box. Of these, each motif was counted as complement plus reverse-complement. The most enriched motif was the G-box (CACGTG palindrome) at 12.9 fold. 12.9 times more G-boxes are in α CNSs than in nontransposon, noncoding, nonCNS control space, normalized by position relative to the gene. The 6 base pair G-box derivatives plus core were all significantly enriched, but at values as low as 6-fold. The 4 not-G-box motifs were significantly enriched at between 2.4 and 8.1-fold. We found a poor correlation between motif enrichment and detectability in *Br* for these 12 motifs, although the G-box itself—most enriched—was third highest in detectability at 63%. Higher than the G-box in detectability was the 5'CCGTCC “meristem” box at 65%, with an enrichment of 8.1-fold. The jasmonic acid box, 5'GCCGCC, enriched to a paltry 2.4-fold, was relatively highly detectable at 50%, and highly enriched G-box-core 8-mer derivative 5'CACGTGGC was detectable in *Br* only 26% of the time. Detectability is certainly giving us clues as to what sequence is essential for any generalized function, and what sequences may be superfluous, as will be discussed. For example, the 5'ACGTG G-box core was the most detectable motif of all, at 67.5%. In the absence of additional information, it seems obvious that some motifs may function in *Br* even though mutated while others have more absolute requirements for continuing function. The G-box itself is a CNS-enriched motif that seems to have a requirement for near-perfect sequence conservation to preserve function, and is especially intolerant to changes in the 5'ACGTG core. Therefore, the G box is a known motif that should be useful to study in order to independently determine the mutational mechanism that causes nondetectability.

Detectability studies for G-boxes in α CNSs, CNSs that are particularly likely to function

Since we already provided evidence that 75% lost CNSs were deleted and thus, were mutated to no function, our premise is that G-boxes are primarily lost by deletion. Given our ability to pull-out and analyze expected orthologous regions within *Br* for comparison to the corresponding conserved noncoding space in arabidopsis, we looked at all CNS-enclosed G-boxes and, more importantly, a subset of these that were experimentally shown to function in light regulation. PIL5 is an arabidopsis transcription factor of the basic-helix-loop-helix type that is known to bind sequence containing a G-box. Oh and coworkers (OH et al. 2009) used microarray data and ChIP-chip (fragments from chromatin immunoprecipitation were localized by hybridization on microarrays) data to infer that, of the 748 arabidopsis genomic binding sites occupied by PIL5, 166 genes were up-regulated in light directly by PIL5. Each PIL5 site represents a strong argument for a *functional* G-Box. We compared these PIL5 sites for overlap with our aCNS dataset and identified 27 aCNSs containing a PIL5-informed G-Box. Each CNS was traced in all expected orthologous positions in *Br*. For each undetectable G-Box, we used our global-npe alignment to identify the predominant evolutionary mechanism that mutated the motif to undetectability in the *Br* lineage. Again, G-boxes were followed for detectability independent of whether the CNS expected to carry them was detected in *Br* or not. Figure 5 gives these results for all aCNS G-boxes and for the 27 G-boxes comprising the “most likely to function” subset, side-by-side. Our overall result: deletions—not point mutations, indels or insertions-- removed the detectability of the majority of G-boxes: 73% for CNS-contained motifs and 82% for PIL5 informed G-boxes. Base substitutions account for a smaller but significant portion, approximately 15%, of G-box mutations to nondetectability.

Discussion

Purifying selection in *Brassica rapa* resulted in many gene regulatory regions that have lost cis-acting binding sites, and 75% of the time, these sites were deleted and therefore have no chance to function in the ancestral manner

Table 1 summarizes the complete CNS detectability data of Supplemental Table 1. Based on the length versus detectability data of Figure 4A, we demanded that CNSs be >30 bp long for our focal analysis of Table 1. We reasoned that CNSs localized to genes that were fractionated down to one (singlets) would lose their genes rarely; from Table 1, this “loss” frequency was 10%, and this became our least-redundant control pool. (This frequency of 10% is not negligible, and is discussed in the next section). Those CNSs that existed near doublet and triplet genes are expected to be relatively more redundant and more liable to loss-of-function mutation, so these CNSs became our experimental pool. We expected that more-redundant CNSs should go undetectable by whatever mutational mechanisms operated in the *Br* lineage to a greater frequency than they go undetectable in the singlet controls. This was indeed the case (Figures. 4A and B, Supplemental Table 1). Overall, an *At* α CNS >30bp long, either 5’ or 3’ of its gene, mutates to undetectability in *Br* 33% of the time (Table 1, last row, column B). The functionality metric for those CNSs that go undetectable by proved chromosomal deletions was 5.3, becoming our maximum not functional value; the CNS must be nonfunctional because the original DNA is not there. A functionality metric of 1 indicates complete functionality since redundancy makes no difference; nondetectability by large insertion had a negligible effect on functionality. The functionality metric for those CNSs going undetectable because of base substitutions was 1.8, meaning that only 34% of CNSs in this category lost function, 66% of them still functioned even though they were undetectable. However, only 1% (16/1543) more-redundant CNSs (column B) CNSs went undetectable for this reason. Considering all 498 cases where a more-redundant CNS went undetectable in the *Br* lineage (Table 1, last row), 72% of these went nonfunctional, as expected because they were largely placed in categories characterized by the word “deletion.” There was no correlation between CNS loss and potential gene death (Supplemental Table 3).

Different branches on the plant phylogenetic tree have differed greatly in transposon blooms and polyploidies. Although there is no experimental evidence, it is possible that the rate of deletion and/or the size of the average deletion differs greatly among plant lineages, so extrapolating from our “mostly deletions” conclusion in *Brassica* to other plant lineages is not warranted. Interestingly, researchers in the detectability of ultraconserved noncoding sequences in vertebrates noticed that post-paleotetraploid teleost fish lost CNS detectability much faster than sister vertebrate lineages not undergoing polyploidy (LEE et al. 2011). Again, an “induction” relationship is possible, but not proved.

The deletion mechanism we envision is the intrachromosomal recombination mechanism discovered for transposons in drosophila (PETROV et al. 1996), described for transposons and genes, respectively, in maize (DEVOS et al. 2002; WOODHOUSE et al. 2010), evidenced in *Br* as rare exons carrying deletions (Tang et al. 2012), in rice (TIAN et al. 2009) and inferred here to be the prevailing mutation mechanism in *Br*. The importance of short direct repeats flanking deleted DNA was first shown as a RecA-responsive process in bacteria (ALBERTINI et al.1982). Not all deletions need to be caused by the same mechanism. Some deletions may be mediated by flanking transposons and/or mis-repair of gaps caused in the movement process (WICKER et al. 2010). Similarly, strand slippage in the replication fork could generate short intrachromosomal recombination deletions (PETROV, 2002). Whatever the mechanism, the fact that we often see kilobase stretches of *Br* (and in maize: WOODHOUSE et al. 2010) removed when a gene and all of its CNS are fractionated does not mean that deletions in plants are long. It seems obvious that, once an initial deletion renders the gene functionless, then some combination of [rate of deletion] and/or [length of deletion] will incrementally remove the entire cis-acting unit.

There has been enough work on rates and lengths of CNS deletion in animals to permit a gross comparison of our *Brassica* lineage CNS deletion process and that operating in any animal to be studied. There are no examples of an ordinary gene being lost in the human lineage by deletion; all are still present *in situ* as pseudogenes (SCHRIDER et al. 2009); the human-chimpanzee is about as diverged as are the *Br*

subgenomes! When genes were lost from the pheromone network in old world apes, the genes remain as obvious pseudogenes (LIMAN and INNAN, 2003); they were not deleted. Petrov (2002), in a theoretical essay on how the C-value paradox is best solved by a balance of deletion and insertion, reviewed the data of others on average rate of deletion per bp substitution and the average size of these deletions. He reviewed data in drosophila, *C. elegans*, crickets, primates+rodents and grasshoppers. Average rates of deletions per base pair substitution ranged from a low of 5% in the mammals to a high of 8.7% in drosophila. The average size of a deletion was more variable, from a low of 1.6 bp in grasshoppers thru 3.2 base pairs for the mammals to a maximum of 48 bps for *C. elegans*.

For the purpose of illustration, imagine a 10 kb stretch of DNA that used to contain an entire gene and cis-acting elements. In the *Brassica* lineage, a $K_s=15\%$ (the *Br-Br* K_s , Figure 1) is enough divergence time to remove the entire genespace—exons plus all CNSs-- without a trace; this occurred routinely during gene fractionation after the paleohexaploidy (WANG et al. 2011). Using Petrov's (PETROV et al. 1996) maximum animal deletion rate/bp (0.13 for drosophila) and drosophila's deletion length of 38 bps, This 10 kb of functionless DNA would suffer $0.15 \times 0.13 = 2\%$ of its bps, or 200 deletion events, each averaging 35 bp long, giving 7 kb of deletion. Using the mammalian rate and length, and the crude arithmetic method above, primates and rodents would delete only 240 bps of the 10kb. The *Brassica* lineage uses a combination of deletion rate and length to more rapidly delete its functionless DNA than animals with tiny genomes and vast population sizes, and far more rapidly than do primates and rodents. The most obvious difference between plants and animals are the hundreds of millions of pollen shed per plant, each grain being a haploid gametophyte. Somehow, the plant's biology accommodates the "genetic load" commensurate with its relatively strong pressure of purifying selection.

Even fully fractionated *Brassica rapa* genes (singlets) are likely to be functionally redundant at least 10% of the time

All of the CNSs used in this study were from homeologous gene pairs retained from the most recent tetraploidy in the arabidopsis lineage (modal K_s 76%, Figure 1, called alpha). Thus, each of these CNSs is redundant or nearly so in arabidopsis. When this gene is triplicated as

part of the paleohexaploidy in the *Br* lineage, there are originally three copies of the progenitor gene and its associated CNSs. If all but one of these genes fractionates, one might guess that the remaining one gene, and each of its CNSs, would confer some nonredundant, unique function. However, 10% of the CNSs expected to be with these singletons are undetectable. In fact, the assessment of “nonredundant” is probably not the whole truth. For every singleton *Br* gene used in this study, there is a possible out-paralog (KOONIN, 2005)--the descendants of the arabidopsis alpha pair--gene family in *Br* that can be a singleton, doublet or triplet. This out-paralog lineage is included in Figure 1 as greyed-out. In some cases, it may not matter which of these genes is active as long as a “correct” contingent of them are retained to specify optimal product levels. As a general test of this reasoning, we analyzed Supplemental Table 2 and asked, “do singleton genes in *Br* have more first cousin genes (descendants of this gene’s α -pair) with retained doublet and triplets?” In other words, did the genome compensate for the loss of an arabidopsis gene by amplifying retention of the “out-paralog” gene, (KOONIN, 2005) such that we should really consider all six potential *Br* orthologs of the alpha pair when we are studying dosage relationships. The data (from Supplemental Table 2): When an *At* gene is retained as a singlet ortholog in *Br*, there is at least one detectable out-paralog 87% of the time. However, if the *At* is retained as a doublet, then at-least-one out-paralog retention drops to 75%, and if retained as a triplet, the out-paralog retention frequency drops further to 72%. Clearly, there is some overlap of function among the six immediate descendants of an alpha pair. Consider further: Before alpha in the Brassicales phylogenetic tree, the beta duplication generated another potential of six “out, out paralogs” that we have not yet included in our conceivably dose-sensitive networks (Figure 1). For these reasons, we are careful to note that a “singleton” gene in *Br* is not “nonredundant,” but “less redundant”.

The sort of reasoning used above is complicated because there is an overall expectation that – for gene functions requiring a fixed stoichiometry of product level-- many genes will be selected for maintaining the *status quo* of product balance (BIRCHLER and VEITIA, 2010). However, there have been multiple polyploidies in the lineage of all plants (Figure 1) that certainly led to functional

redundancy. Add to this complexity the fact that genes on over-fractionated subgenomes are expected to—on average—express to lower RNA levels than do genes on the less fractionated homeolog (SCHNABLE et al. 2011). For *Br*, subgenome III is the dominant subgenome (CHENG et al. 2012).

5' and 3' α CNSs generally confer function

Plant CNS function is supported by conservation itself (REINEKE et al. 2011; LOCKTON, 2005), the association of CNS-richness with particular genes and motifs (FREELING and SUBRAMANIAM, 2009), the positive association of CNSs with open chromatin (ZHANG et al. 2012) and by expression association studies (SPANGLER et al. 2011). This current study approaches the function question by comparing less-redundant CNS loss—those on singlet *Br* genes—to the loss of more-redundant CNSs, as they are expected to exit on doublet and triplet *Br* genes. Figure 4A shows the relationship between detectability and CNS length: with one exceptional data point, CNSs expected to be on singlet genes are more detectable than CNSs expected on doublets are more detectable than on triplets. This makes sense if purifying selection is strongest when there is only one copy, moderately strong when there are two copies and weak for triplets. For the bin carrying the shortest CNSs, detectability is 2.9 fold higher for a singlet than for a triplet. For the bin carrying the median-lengthed CNS (31-30 bp), a singlet is 1.8 fold more detectable as a singlet than as a triplet, with the doublet in the middle. This result implicates selection—and α CNS function-- unless mutation rates are somehow correlated with the redundancy of cis-acting regulatory units motifs; that is not reasonable.

The results of Figure 4B involving redundancy versus detectability are, in general, expected, but the differences in detectability of singlet CNSs depending on the subgenome (I vs. II vs. III) is disturbing. That subgenome III always has more detectable CNSs than do the other two subgenomes cannot be ignored. As with the data of Figure 4A, purifying selection seems to act most strongly on CNSs that are less redundant. Removing a unique CNS from a gene could well remove an essential or selectable function. However, why should it matter on what subgenome the singlet α CNS is located? It does. Singlets on I: II: III are detectable at 82, 87 and 96% with differences significant,

$p < 0.05$. While none of these differences is much smaller than 100%, there still must be an explanation. The most obvious is that subgenome III is less mutagenetic; it deletes at a lower rate than the other subgenomes for some structural reason. This “mutationist” hypothesis was very much in contention as an explanation of biased fractionation, where one subgenome’s genes gets deleted significantly more often than the other subgenome (FREELING et al. 2012, SANKOFF et al. 2010). However, this mutationist alternative was considered carefully and disproved unequivocally in the case of the maize paleotetraploid (SCHNABLE et al. 2011). It was shown that both subgenomes of maize suffer mutations--deletions via intrachromosomal recombination (WOODHOUSE et al. 2010)—at the same rate but that one subgenome expresses its genes to higher levels than the other on average, so pairs of genes tended to fractionate the homeolog that expresses least. These workers (SCHNABLE et al. 2011) didn't just demonstrate that genome dominance predicts biased fractionation; they actually tested the rate of deletion of functionless transposon and intron DNA between subgenomes and found the rates to be the same. For these reasons, the “selectionist, not mutationist” explanation was adopted for the maize lineage tetraploidy, and predicted to apply to the paleohexaploidy in the *Brassica* lineage as well. For *Brassica rapa*: fractionation is biased with subgenome III being the least deleted (WANG et al. 2011), the mechanism of exon loss is deletion (TANG et al. 2012), and subgenome III dominates its RNA levels over subgenomes I and II (CHENG et al. 2012), just as is the case of maize. Perhaps singlets on subgenome III carry genes that are particularly and continuously important for growth and development, and singlets on the not-dominant subgenomes just don't matter quite as much.

The G-box and motif detectability

Figure 5 graphs detectability in *Br* (as an exact motif sequence) of (1) G-boxes within all *At* α CNSs and (2) a subset of these G-boxes that are also experimentally validated (by ChIP-chip) PIL5 helix-loop-helix transcription factor binding sites. As expected from the CNS detectability results (Table 1), G-boxes that lose exact sequence are almost always deleted, not lost by base substitution.

Supplemental Table 4 presents our update of CNS enrichment values

given our version 2 of the *At* α CNS list, and slightly updated methods. This Table also presents detectability data for all motifs enriched significantly in CNSs by >2-fold. For those motifs, it's important to know that, in every case, complement and reverse complement were enriched to an equal degree (by Chi-square). The 5-mer within the G-box, 5'ACGTG is more detectable than the G-box itself, and is more detectable than any of the 11 G-box derivatives. For example, the most CNS-enriched G-box derivative, 5'CACGTGGC and its reverse complement, was among the least detectable at 26%; we conclude that this 8-mer motif contains alternative bp substitution sequences that still function even when mutated for the majority of similar sequences, and we draw a similar conclusion for most of the G-box derivative motifs. While these particular G-like-boxes may (or may not) be the optimum DNA-binding partner for one or a few protein-DNA interactions, this sequence is really not a motif. Rather, they are each a specific sequence that contains a motif. We suggest that this G-box situation is typical of the generally overlapping, multiple motif data that comprise our current plant motif lists. For example, one such list that attempts to be exhaustive—the 426 regular expression motifs gathered together in the MotifView application in CoGe-- fall into many sets of overlapping sequences, each supported by a unique experimental datum. Detectability measures over evolutionary time may help consolidate binding sequences into actual motifs. By this reasoning, the “G-box” is not an actual motif, but a derivative. The actual motif by this reasoning, the core shared by all or most related sites, could be 5'ACGTG because it is the most detectable of CNS-enriched boxes in the “G-box” family.

The general aim of bringing together CNSs and motifs or clusters of motifs—and especially ChIP-seq sites (much needed data)—is not even well formulated for plants. We know next to nothing about what proteins actually bind CNSs, how many different binding sites generally occupy CNSs, or if the spacing of sites within or among CNSs is important.

Fractionation: nature's "deletion machine"

Knowing that the predominant reason arabidopsis CNSs go undetectable is deletion, leading to loss-of-function, is crucial for the intelligent application of a new strategy for enhancer-like site analysis:

fractionation mutagenesis. For example, the *Brassica rapa* paleohexaploid informs intelligent, hypothesis-driven enhancer experiments in arabidopsis. Fractionation mutagenesis is exemplified in the GEvo blastn output graphic of Figure 3, where the query is arabidopsis *At1G75520*, a particularly CNS-extensive gene encoding a RING zinc finger protein of SHI-type. Its 17 CNSs, covering 7.5 kb of chromosome in addition to the 1.9 kb transcriptional unit, have been largely retained in triplicate in *Br*. However, fractionation has rendered undetectable—*probably deleted*—a few longer orthologs of α CNSs: those circled in Figure 3 are clearly present in *At* and BrII. The arrows indicate individual sequences in *At*. If there were a particular RNA-level pattern that was missing or aberrant in BrI and BrIII, but ancestral in BrII, the CNS denoted by the rightmost arrow would become a candidate sequence with a hypothesis as to its meaning. Looking further to the right in Figure 3, we find that CNSs on subgenome III (*BrIII*) are ancestral, but some CNSs are missing from orthologs on subgenomes I and II. Since subgenome III is the dominant subgenome (CHENG et al. 2012), this bias for loss is expected.

There will soon be many more *orthologous At* CNSs when usefully diverged *Brassicaceae* genomes are sequenced and aligned, and when CNSs obtained from multiple alignment data are merged with our pairwise CNS list. Those fractionated in *Br* should predominately lose function. It is valid to think of the fractionations following polyploidy in the Brassicas (and probably following other plant polyploidies as well) as deletion machines ideally suited to be used in the procedure of fractionation mutagenesis.

Methods

Rationale for confining this paper to CNSs defined in one specific way

Our *At-At* CNSs reflect one definition of a CNS: a syntenic noncoding conservation detected by blastn with significance at or better than a 15/15 exact match, and between genomes or subgenomes diverged to a modal Ks of 0.9-0.5 (FREELING and SUBRAMANIAM, 2009). The CNS dataset produced has the advantages of having been the object of some study, and because this data set depends on local alignments generating an even-handed sampling of conserved noncoding regions no matter how far they may exist from any conserved coding sequence. It is certainly more sensitive to anchor on a coding part of the gene and extend alignments, but this sensitivity only applies close to the anchor. For example, multiple global alignments anchored on the start of transcription and moving up to 1 kb 5' have provided an excellent CNS data set, and they do overlap with ours (BAXTER et al. 2012), but this data set goes deficient in those CNSs moving from 500bp to 15 kb away from the nearest exon; transposon insertions disrupt global alignments. No one method of obtaining CNSs is best. Fortunately, complete coverage is not important for us to see how arabidopsis CNSs are lost in the post-hexaploidy *Br* lineage (Figure 1), so we use homeologous arabidopsis CNSs from our *At-At* v2 data set (Supplemental Table 1) described below.

Arabidopsis α CNSs

In 2006, there were no sequenced Brassicaceae genomes within the window of Ks 0.5-0.9. However, the two alpha *Arabidopsis thaliana* subgenomes descended from its most recent paleotetraploidy were nicely diverged for CNS discovery, so the TAIR4 version of the Arabidopsis genome was compared to itself (THOMAS et al. 2007). Of the original 14,944 individual *At-At* CNS sequences of version 1, 3,635 CNSs were removed: 82 were found to be out of synteny, 22 erroneous CNS calls, 169 CNSs were reassigned to neighboring genes, 1831 CNSs were invalidated due to wrong direction, and 1531 CNSs were found to overlap CDSs or RNA genes, called as annotation of plant genomes became more complete. Version 2 CNSs, identified in relation to TAIR8 annotations, are syntenous in relation to other homeologous features. Column A of Supplemental Table 1 is a notation for each of these Version 2 CNSs that includes the *At* gene

name to which each sorts; the actual sequence of this sequence is displayed later in the row. Our CNS calls may be proofed easily with the GEvo links of Supplemental Table 1. GEvo is the sequence comparison tool in the CoGe toolbox (<http://genomeevolution.org>) of comparative genomics applications (LYONS and FREELING, 2008). GEvo provides a graphical comparison of multiple genomic regions indicating high-scoring segment pairs (HSPs) for a variety of sequence similarity testing algorithms (Selected under the ‘Algorithm’ tab of GEvo), between the defined genomic regions. GEvo was used extensively during our version 1 to version 2 update. The 11,302 version 2 *At-At* CNSs have been “burnt” onto a TAIR8 genome on the model line—this genome is identified as id 39598 in CoGe. These CNSs can be visualized within GEvo by selecting “Yes” for “Show pre-annotated CNSs” under the ‘Results Visualization Options’ tab in GEvo. Using GEvo, our precalled CNS positions can be readily compared through HSPs (high-scoring segment pairs) generated by blastn (default blastn settings with a spike of 15 nucleotides). Column B of Supplemental Table 1 is easy to parse for CNS length; we focus only on those 2,509 longer CNSs for our focal experiment in which we define a functionality metric for each category of α CNS loss (Table 1). However, all Version 2 CNSs are used for other experiments and all motif experiments.

Locating orthologous coordinates for α CNSs within *Br*

For each of the 6330 *At* genes with a retained *At* α -homeolog (Supplemental Table 1) we used the synteny screening blocks technique (TANG et al. 2012), to identify all possible orthologous regions in the *Brassica rapa* (*Br*) Chiifu (Chinese cabbage) genome. Given the recent hexaploidy in the *Br* lineage (Figure 1), we expected to find up to three orthologous copies for each *At* gene. Of the 6330 *At* genes used for version 2 CNS discovery, we (. 2012), identified at least one *Br* ortholog for 6245 of them, with 2391 *At* genes having a single detectable orthologous copy in *Br* (singlets), 1723 *At* genes with two orthologous copies in *Br* (doublets) and 654 *At* genes with three orthologous copies in *Br* (triplets).

In the absence of CNS fractionation, and assuming that our CNSs were sorted to their correct gene, we expected to find the CNS

whenever we found the gene in *Br*. Based on *At-Br* orthologies, we expected to find 9179 CNSs within expected orthologous positions, 3882 as singlets, 3678 as doublets and 1619 as triplets. We attempted to detect each of these 9179 α CNSs within each expected orthologous *Br* region. Nucleotide sequence of the gene space (expanded genomic region around and including coding regions and spanning from farthest upstream to downstream CNS) of each *At* gene containing one or more of these 9179 CNSs was masked for very repeated sequences (50X copies across entire *At* genome). The corresponding gene space of each detected orthologous region (singlet, doublet or triplet ortholog) in *Br* was also repeat-masked. Each *At* and *Br* orthologous genespace pair was compared using the same blast settings used for *At-At* CNS discovery (THOMAS et al. 2007). Every blast HSP hit to the *Br* orthologous genespace was then screened for synteny, using a perl script, to filter out probable noise, while simple sequences were filtered out using the DUST filter option of BLAST.

α CNSs that do *not* show a hit using the above blastn settings are valuable data. In the following (next) section, we describe analysis of such sequences using a global alignment algorithm to determine the nature of evolutionary modifications that may have contributed to the lack of detectability of these CNSs. The general idea is this: The CNSs "lost" in a singlet are assumed to still function, but to have drifted in functionless sequence, or to have suffered "binding site turnover" (MOSES et al. 2006). Our Discussion section argues that this assumption is not the whole story, but we did make this assumption. Any frequency of nondetectability above the baseline of loss in singletons was interpreted as being caused by actual functional loss either by base substitution, deletion, small indels, or a large insertion.

Identifying the molecular mechanisms that caused the lack of detectability of α CNSs in *Br* doublets and triplets

Locating orthologous coordinates within Br for

α CNSs undetectable using blastn

Earlier, we described the use of our previously published CNS discovery blastn settings to measure detectability of α CNSs within expected orthologous gene spaces within each of the three *Br* subgenomes. Each α CNS that was not detected in the expected orthologous region of *Br* using our standard settings was retested to

determine the predominant mechanisms that could potentially contribute to the lack of detectability. We started by identifying and extracting the nucleotide sequence for the expected orthologous regions for each α CNS that was not detectable within *Br*. As in the case of the blastn analysis used for measuring CNS detectability, we used the *Br* orthologous gene spaces (coding region+40kb on both sides of coding region) as the subject sequence. The query sequences were the coding regions of the *At* gene to which the α gene to which the paces (coding region+40kb on both sides of coding region) as the subject sequence. The query sequences were the coding regions of the *At* gene to which the α CNS was assigned in Version2 (Supplemental Table 1) to which was added the nucleotide sequence extending out to and including the farthest upstream and downstream CNS. In Supplemental Figure 1, the GEvo link points to a graphic where this *At* gene space is highlighted yellow; to see this requires selecting “see genespace” in the GEvo options panel.

Each pair of *At* and *Br* gene spaces were compared using blastz with default settings. The position of each α gene spaces were compared using blastz with default settings. The position of each α CNS was studied for overlap with any blastz HSP (high scoring segment pair) between the *At* and *Br* genespaces. Those CNSs that overlapped with blastz HSPs were assigned the location of the HSP (start and stop positions) as its expected location in *Br*. CNSs that did not overlap a blastz HSP, but were found flanked by blastz HSPs, were assigned an expected position between the flanking HSPs. In cases where flanking HSPs were not present, depending on the position of the CNS relative to the gene, the expected location was defined from either the start of the orthologous genespace to the start position of the gene, or the stop position of the *Br* gene to the end of the *Br* genespace (TANG et al. 2012). An example of this procedure follows. Supplemental Figure 1 (<http://genomeevolution.org/r/4dc3>) shows an annotated view of the same GEvo panel described earlier (Figure 3), but now displays blastz HSPs between the *At* gene and its *Br* orthologs. In this figure, α CNS 315; 2;CNS_s680 (Pink highlight in Figure 3) overlaps with a blastz HSP in *Br* II and *Br* III but falls between 2 flanking HSPs in *Br*I. The search sequences used for studying the mechanisms acting on this CNS in *Br*II and *Br*III are the blastz HSPs labeled *Br*II and *Br*III (Supplemental Figure 1). The orthologous region in *Br*I falls between

these two blastz HSPs, indicated in pink highlight in Supplemental Figure 1.

*Identifying the molecular mechanisms that caused the lack of detectability of α CNSs in *Br* doublets and triplets*

There are several possible reasons for a CNSs to go undetectable: deletions of an entire chromosomal segment resulting in the removal of one or more CNSs, or relatively smaller-scale changes including insertions, smaller deletions, a combination of both (indels) and base substitutions making individual CNSs undetectable. For smaller CNSs, even one base substitution would either destroy the minimum exact match blast word size or drop the CNS below the e-value cutoff, that equal to a 15/15 exact match. We wrote a simple perl script to use a global alignment algorithm (NEEDLEMAN and WUNSCH, 1970) with cost-free ends (BLOSUM 62) to align the nucleotide sequences of each α CNS without a detectable ortholog in *Br* with the expected orthologous regions within *Br*; these regions were found as described previously.

Using the genomic positions for the expected location of each CNS in *Br* to inform our search, we generated an alignment between each CNS sequence and the nucleotide sequence (repeat masked) corresponding to the expected location in *Br*. The *Br* and *At* sequences were aligned using a global alignment algorithm (not a blast family algorithm) with no end gap penalties (NEEDLEMAN and WUNSCH, 1970). We did this because the CNS sequence length is much shorter than the *Br* subject sequence. A score value was generated for each alignment and a p-value statistic was used to measure quality of alignment. As control for the alignment for each α CNS, we used a perl script to generate 10,000 'scrambled' random sequences, each representing a "random" permutation of the nucleotides that make up each CNS. The alignment score for each CNS was compared to those of the 10,000 random sequences to generate the p-value of significance. We define any alignment with p-value less than or equal to 0.05 as being "above noise." Each high-quality, optimal alignment generated by the global-npe algorithm was analyzed using perl scripts for deletions, base substitutions, insertions and exact matches at each position. Alignments with gaps only on the *Br* sequence were classified (using a perl script) as deletions, those with gaps only on the *At* sequence were

classified as insertions and when gaps occurred on both *Br* and *At* sequences, such alignments were classified as indels. We then used these “gaps” data to infer the predominant mechanism contributing to lack of detectability of *At-At* α -CNSs in *Br*.

mRNA levels for Br genes in seedling root and shoot

Since our *Br* sequence is from Chiifu, a Chinese cabbage variety, it is important to know that our RNA expression data is from this same genotype. RPKM (reads per kilobase per million mapped reads) data for genes expressed in seedling stem, leaves and roots has been analyzed and presented in Supplemental Table 3 (CHENG et al. 2012) as a control experiment for potential gene death.

Revised CNS-enriched TFBS motif list

Using Version 1 of the *At* α CNS list, previous work (FREELING et al. 2007) identified a few known transcription factor-binding sites (TFBS), as regular expression motifs, that were significantly enriched in aCNS sequence as compared to noncoding, nonconserved sequence. While the G-box, a “strictly conserved” palindromic hexamer, was by far the most significantly enriched, other “strictly conserved” motifs were significantly enriched over 2-fold as well. By “strictly conserved” we mean that at least 5 nucleotides within the consensus sequence for the motif must be conserved in the same order; e.g. For the G-box the consensus motif is CACGTG, the core of the consensus motif ‘ACGTG’ should be conserved to be considered “strictly conserved”. Because this work begins with a revised CNS list, Version 2, and because we wanted to refine how we controlled for nonfunctional motifs (we didn't mask transposons in our previous work, for example), we updated our enriched motif list (Supplemental Table 4). We did not use all of the often overlapping motifs available, and in the literature, but concentrated only on 12 motifs picked that were, like the G-box, more strictly conserved and enriched by >2X in CNSs: CACGTG (the G-Box), 5' ACGTGGC (in the ACGT category), GCCGCC (jasmonic acid box), 5' AAACCCTA, and 5' CCGTCC (Freeling et al. 2007) to which we added [CT]ACGTGGC, CACGTGGC, ACGTGGCA, ACGTGTC, AAACCCTAA, TGTCTC, CCACGTGG. Several of these motifs can be seen (underline) to be similar. This strictly conserved criterion was used so we could more easily infer whether or not they were intact following mutation to

nondetectability. Specific references for each motif sequence are in Supplemental Table 4 and in a beta-test application in CoGe: MotifView, (<http://genomevolution.org/CoGe/MotifView.pl>). Noncoding, nonconserved, and nontransposon regions from *within the same gene space* as each α CNS were used as the control for each of these motif enrichment studies.

α CNS that are reinforced by overlap with published Pil3-like5 protein (PIL5) binding sites and their G-boxes

Oh and coworkers (OH et al. 2009) used ChIP-chip (chromatin immuno-precipitation with microarray sequence recognition methods) data to infer that 748 arabidopsis genomic binding sites were occupied by basic helix-loop-helix transcription factor PIL5 and 166 nearby genes were up-regulated directly by PIL5. Since PIL5 has been shown to bind CACGTG, each G-box within a PIL5 "peak" represents a strong argument for a functional G-Box. We compared these PIL5 sites for overlap with our α CNSs. These 32 G-boxes were assumed to be particularly likely to be functional. 5 of these CNSs did not have an ortholog in *Brassica rapa*; understanding these is outside of our topic. The remaining 27 were studied at all orthologous positions in *Br*.

The relatively low number of α CNS-PIL5 peak overlaps was expected. α CNSs (not being orthologous CNSs) can only include those cis-acting sites that were retained after the most recent tetraploidy in the arabidopsis lineage. Further, our unanchored blastn pairwise CNS discovery tool, while necessary to find CNSs that are far from coding sequence syntenic anchors, is known to miss many if not most of the cis-acting sites that are close to the transcription unit (THOMAS et al. 2007).

Tables

Predominant cause of <i>At-At</i> CNS loss in <i>Br</i> determined by computation	B/A	Total # of undetectable 5'+3' CNSs> 30 bases* / total (%)		
	Functionality metric**	A= # (%) for CNSs in singlet <i>Br</i> genes (control)	B=# (%) for CNSs in doublet or triplet <i>Br</i> genes	A+B
	1= functional	--	--	--
1. Segmental deletions , flanking feature(s) lost. Functionless.	5.3	16/966 (~1.6%)	129/1543 (8.4%)	145/2509
2. Computed deletions , but with no flanking feature also deleted	3.4	43/966 (4.45%)	226/1543 (15.2%)	269/2509
3. Indels , both insertions and deletions of any size.	2.0	22/966 (2.2%)	90/1543 (4.4%; 2.3% in doublets, 8.9% in triplets)	112/2509
4. Insertions , of any length	1.1	14/966 (1.4%)	37/1543 (1.5%)	51/2509
5. Base Pair substitutions	1.8	11/966 (1.13%)	16/1543 (2%, 1.5% in doublets; 4.14% in triplets)	27/2509
All <i>At</i> CNSs undetectable in <i>Br</i>	3.8	106/966 (10%)	498/1543 (33%)	604/2509
All <i>At</i> CNSs detectable in <i>Br</i>	nonsense	860 (89%)	1045 (67%)	1905/2509

Table 1. Gene redundancy was used to infer whether or not an undetectable *At* CNS is undetectable in *Br* because it has been mutated into nonfunction and categorization by “predominant cause” was done computationally.

*Almost all CNSs in arabidopsis that are >30 bp long cannot be rendered blastn-undetectable by the *At-Br* modal Ks of 38% (Figure 1).

Our logic: we assume that a *Br* gene that has been fractionated down to a singlet will tend to require the function of all of its ancestral CNSs, but that CNSs on doublet or triplet genes will be subject to occasional loss of function due to redundancy. Thus, any predominant cause category that tends to NOT reduce function should have B/A ratio nearer to 1. Thus **B/A, the functionality metric, varies from 1 (functional) to 5.3 (functionless).

Figures

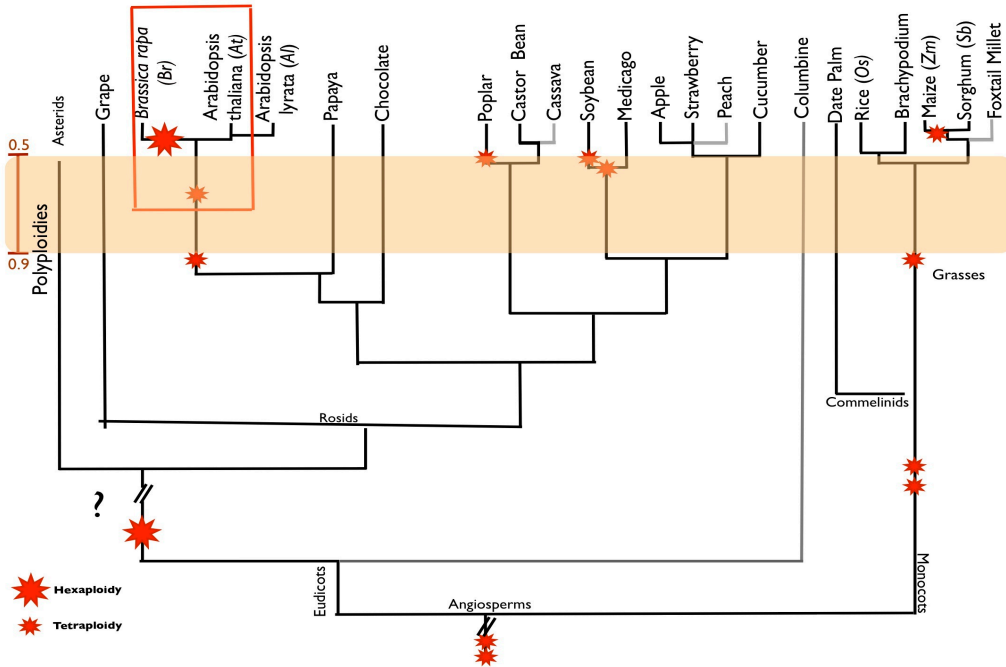


Fig. 1. The gene tree of a pre-alpha tetraploidy Brassicaceae gene with a protein-coding sequence (black arrow) and five conserved noncoding sequences (boxes on the model line) as it duplicates at the alpha and, in the *Brassica* lineage, undergoes an additional paleohexaploidy before it was sequenced in *Brassica rapa* (*Br*). The modal K_s values, for each of these three events is shown, as downloaded from the SynMap application in the CoGe toolbox. The red CNS exemplifies the sort of CNS we follow. It is detected as a conserved sequence between the two homeologous genomes of arabidopsis (*Arabidopsis thaliana*, *At*), but is fractionated (red arrow) from one of the *Br* homeologs in this doublet. The grey lineage is of the “out-paralog” Brassica lineage, in this case represented in *Br* as a singleton gene. Note that a CNS was lost just 5’ of coding sequence, and is not present in the out-paralog lineage. Even though this CNS did exist in the test lineage, we did not test for it because we began with homeologous *At-At* CNSs.

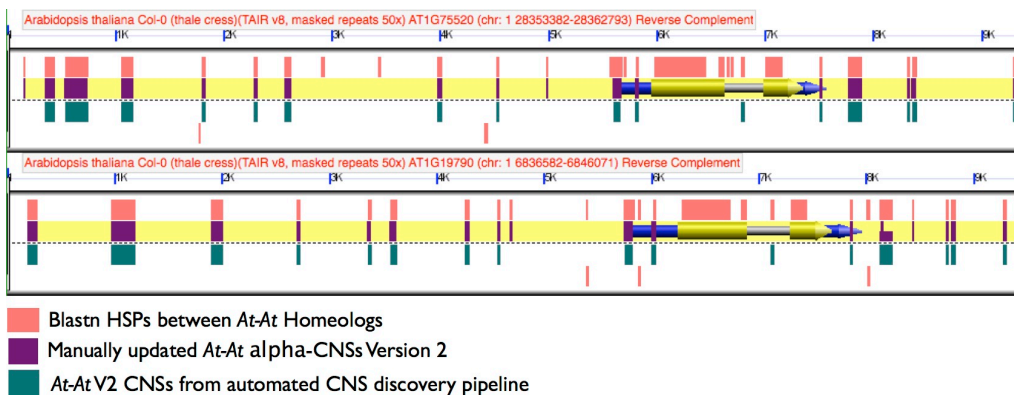


Fig. 2. A CNS-rich pair of genes in arabidopsis, retained from the most recent (α) paleotetraploidy, compared as sequence using “find CNS” blastn settings and displayed in the GEvo viewer. Panels of genomic regions (which can be regenerated at <http://genomevolution.org/r/4db1>) annotated using the GEvo application in the CoGe suit of tools (<http://coge.iplantcollaborative.org>). The figure compares an *At* gene (*AT1G75520*), a member of SHI transcription factor gene family and its homeolog. Blastn HSPs between the two genes (orange rectangles), manually updated α CNSs (purple blocks on upper model line; V2, Supplement Table 1) and CNSs detected using automated CNS pipeline (green blocks on lower model line; Supplement Table 3) are annotated in this figure. Note the similarity of the two CNS annotations, and how the HSP data in this experiment generally supports our CNS calls.

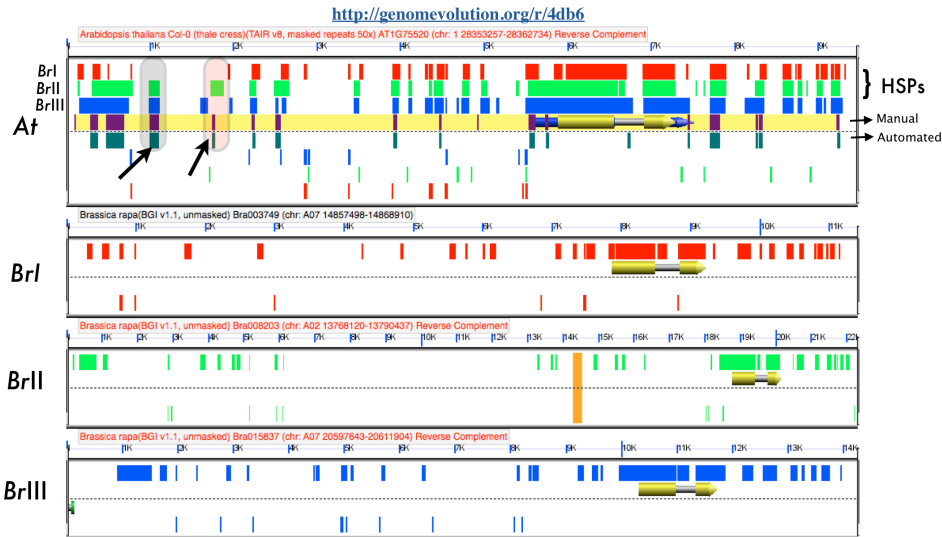


Fig. 3. The fates of different arabidopsis CNS sequences from Figure 2 in the three subgenomes of *Brassica rapa* (*Br*) visualized in GEvo blastn comparison. Regenerate this experiment at <http://genomeevolution.org/r/4db6>); the GEvo application (<http://coge.iplantcollaborative.org>). The top panel shows the α CNS-rich *At* gene (*AT1G75520*) of Figure 2, a member of SHI gene family, and the three panels below show its detectable orthologous genespaces in *Br* (*Br* I, II and III). α CNSs (purple bars) and the gene space (yellow background) are annotated on the *At* gene panel. High-scoring segment pairs (HSPs) corresponding to pairwise blastn comparison between the *At* gene and each of the three panels are indicated on the *At* gene panel as red bars for *Br*I, green bars for *Br*II and blue bars for *Br*III); the default color scheme in GEvo differs. Grey area highlighted follows the detectability of one α CNS across all three *Br* orthologous regions. Orthologous copies of this CNS are detectable in *Br*II and *Br*III subgenomes, but undetectable in *Br*I.

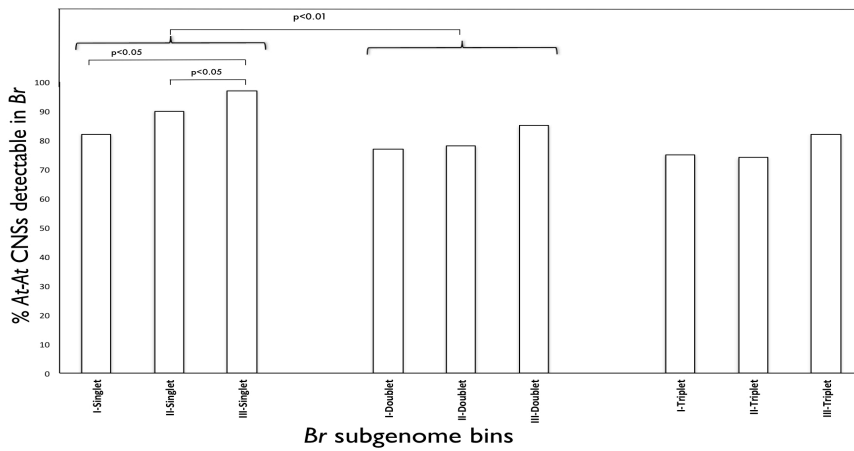
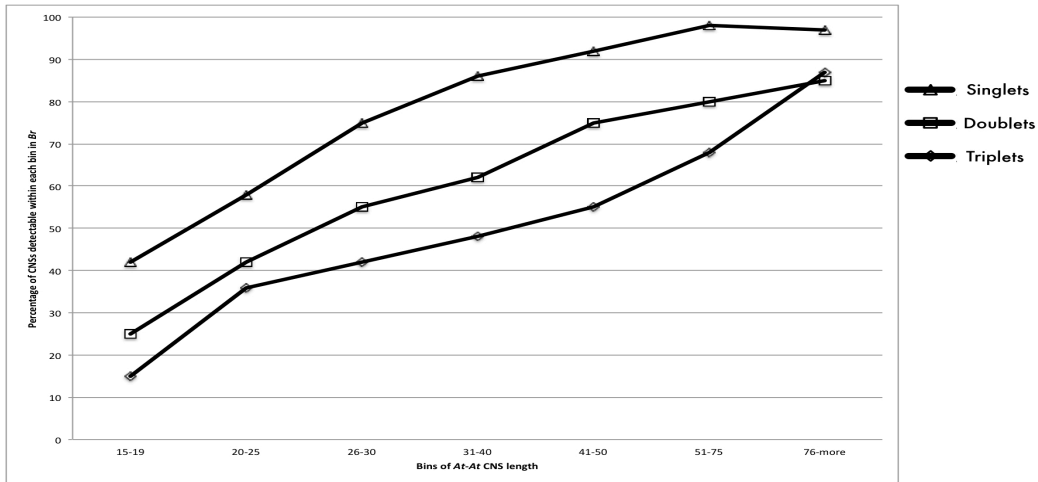


Fig. 4 **A.** CNS length versus detectability. Percent detectability of α CNSs in *Br* over bins containing CNSs of a given length (bps). Expected number of copies for each α CNS is based on number of detectable syntenous orthologs for each *At* gene in *Br* genome. **B.** Comparison of detectability of all expected copies (singlet, doublet or triplet) of *At-At* CNSs (31 bases or longer) between the three subgenomes within *Br*. Expected number of copies for each *At-At* CNS is based on number of detectable syntenous orthologs for each *At* gene in *Br* genome.

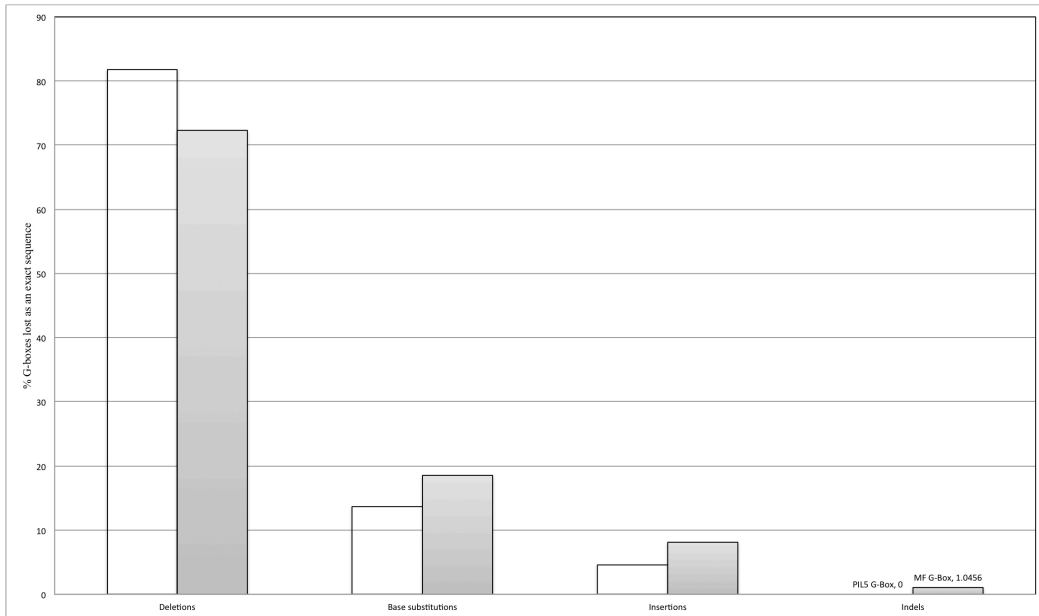


Fig. 5. Predominant mechanism causing lack of detection of the G-Box (CACGTG) within α CNSs within the expected orthologous segments of *Brassica rapa*. Data for G-Boxes detected using a regular expression are labeled MF-G-Boxes; data for PIL-5 defined G-Boxes detected within α CNSs are labeled as PIL5-G-Box. Pil5-G-boxes are a subset.

Supplemental Data

Supplemental data is available for download at

http://figshare.com/articles/Chapter_2_Supplemental_Data/856772

Supplemental Figure 1. GEvo panel example showing regions of *Br* pulled out from each *Br* subgenome for each asubge

Supplemental Table 1. Version 2 *At* α CNS list.

Supplemental Table 2. Detectability of α CNSs in expected positions in *Br*

Supplemental Table 3. mRNA levels in Chiifu: RPKM

Supplemental Table 4. Motifs known to be enriched in CNSs: Update and detectability in *Br*.

Chapter 3: Computational pipeline to track the evolutionary fates of arabidopsis conserved noncoding sequences through a paleohexaploidy and into the genomes of its close relative, *Brassica rapa*: PL3 Genespace Tweaker + Mapping OverWOULD IT BE USEFUL TO ITALIMIZE MAPPING OVER AND OTHER ALGORITHMS TO DISTINGUISH THEM FROM „BACKGROUND“

Introduction

Conserved non-coding sequences (CNSs) are islands of non-coding sequence that show an unexpectedly low level of divergence. In plants, these sequences have been identified by a variety of multiple alignment protocols (FELTUS et al. 2006). We use a pairwise alignment method similar to that first used in plants (KAPLINSKY et al. 2002; LYONS and FREELING, 2008) because then we can capitalize on the even-handed way the blastn (ALTSCHUL et al. 1990) algorithm finds CNSs 15 kb away from an associated gene with the same efficiency as it does a CNS very close to its gene, as expected of a local alignment algorithm.

Blastn-based local alignment of paired homologous genomes-- those diverged between median Ks 0.5-0.9—facilitates CNS discovery, and has now been automated. (TURCO et al, 2013).

In this pipeline, nicknamed PL3.0, the CDS of each official and newly annotated gene in the query and subject genomes are compared using LASTZ (HARRIS, 2007) run with default parameters. For each syntenic gene pair, as identified by the QUOTA-ALIGN algorithm (TANG et al. 2011), regions of sequence starting 12 kb upstream of the annotated start site of each gene and extending 12 kb past the end of transcription were extracted from the 50× masked genomic sequence files. In addition to the 50× repetitive sequence masking, all annotated protein coding regions (CDSs) were also masked. BL2Seq was used to compare the two regions using the following parameters: word size 7bp (-W 7), gap penalties extension 2(-E 2), nucleotide mismatch penalty 2 (-q 2), nucleotide match reward 1 (-r 1), cost to open a gap 5 (-G 5), and DUST filtered turned “on”(- FT). Hits with a bit score less than 29.5 (equivalent to a perfect match of 15 base pairs) (KAPLINSKY et al. 2002) were discarded, based on estimates of how background neutral mutation might be expected to “carry over” a short near-identical sequence by chance alone; These standard “find CNS” settings have been justified (LYONS and FREELING, 2008).

The ideal window to identify CNSs most likely to have biological function is to only compare genomic regions that have experienced between 0.5 and 0.9 synonymous substitutions per site (FREELING and SUBRAMANIAM, 2009). For less diverged homologous genomic regions, some functionless sequences will still retain detectable sequence similarity, while in more diverged genomic regions many functionally constrained sequences will have diverged too much from each other to be identified as homologous, with only the largest, most conserved CNSs remaining detectable. For that reason, PL3.0 is only recommended for genomes diverged within this “window of useful divergence.”

PL3 was proofed by using a “gold standard” pair of genomes, the two genomes within arabidopsis retained from its most recent tetraploidy; median Ks is 0.65. These 14944 CNSs were hand annotated using alignment viewers and a great deal of time (THOMAS et al. 2007). The CNSs called manually, the “gold standard”, were used to decorate the TAIR arabidopsis genome that was most recent at the time (THOMAS et al. 2007). This decorated genome was put into the formats required to upload into the CoGe toolbox of comparative genomics applications (now hosted by iPlant, NSF’s plant cyberinfrastructure project). Within CoGe is a particularly useful sequence comparison alignment tool called GEvo. Figure 1, illustrates how this “gold standard” arabidopsis genome looked in a GEvo panel when focused on a segment of arabidopsis chromosome; the green rectangles on the model line are these manually annotated CNSs. PL3 generated CNSs—identified as chromosome, start, stop, strand (in relation to its nearest paired gene)—were also used to decorate the same arabidopsis genome. These are also decorating Figure 1, but are color-coded purple. Simply comparing the patterns of the gold standard with the experimental, the PL3 output, generates a proof. Here, the patterns are very similar. A more rigorous proof was to repeat the entire PL3 process at the level of a single chromosomal region. Figure 1 also includes the homeologous arabidopsis sequence that was used for the original CNS calls, so the computed and the experimental elements can be compared “on the fly”. GEvo was set on “find CNSs” blastn parameters and the resulting blast hits (HSPs or high-scoring segment pairs colored orange in Figure 1) were compared to the gold standard and pre-computed CNSs. The result here is a good

match, similar to the results published in the PL3 release paper (TURCO et al, 2013).

There are now several useful whole genome sequences in the family Brassica, or crucifers. Figure 2 is a small phylogenetic tree. *Aethionema arabicum* (*Aa*) whose sequence was recently published (HAUDRY et al. 2013), falling within the window of useful CNS discovery, should be useful to capture a complete dataset of orthologous arabidopsis (*At*)-*Aa* CNSs. Additionally, the *Brassica* crops are becoming sequenced (Fig. 2). In particular, we wanted to follow the fate of arabidopsis CNSs into the three subgenomes of *Brassica rapa* (v. Chiifu, Chinese cabbage) (WANG et al. 2011). Unfortunately, *Brassica rapa* (*Br*) and *At* are too closely related (Ks 0.15); they are not in our window of useful CNS discovery. So, we embarked on a plan to 1) use PL3 to find all arabidopsis orthologous CNSs and 2) “map them over” to the three potential, expected syntenic positions in the *Br* genome to see if they are present (+) or absent (-), and if the gene to which they associate is present or absent (fractionated). From previous work on small-scale mapping over homeologous *At* aCNSs to *Br*, we knew that an undetectable CNS is a deleted CNS (and not a base-substituted pseudoCNS) about 80% of the time (SUBRAMANIAM et al, 2013 and Chapter 2).

The desired outcome of Mapping Over from arabidopsis to *Brassica rapa* would be to generate reliable, detailed cartoons of, for example, an *At* gene that has 10 CNSs, and what became of those CNSs in a doublet of *Brassica rapa*. For an example of subfunctionalization, for a gene with 3 5' CNSs: 5' + - + / 5' + + - . Mapping over draws every cartoon for every *At* CNS in every orthologous gene in *Br*, where + means the CNS exists in the expected position and – means it is a probable deletion.

This Chapter documents how these two aims were accomplished. These aims turned out to be far more ambitious than expected. The reason for this was, fundamentally, that PL3 -- proofed using homeologous CNSs derived from two genomes from which many genes had been fractionated and thereby “spaced-out” into islands of CNSs+gene-- did not assign CNSs to genes accurately enough when two compact genomes were compared (*At* and *Aa*). The PL3 output CNS list was not accurate enough to map over directly. For that

reason, the Map-Over pipeline (MO) described in this paper has two components: 1) genespace tweaker and 2) mapping over (MO). The purpose of the former is to “fix” PL3 output.

Figure 3 illustrates the concept behind MO 2.0. The outgroup (such as sorghum to maize and arabidopsis to Brassica) is too phylogenetically close to the derived taxa to use PL3 directly, so the orthologous CNSs of the outgroup—derived by comparing it to rice or *Aa*, respectively—are mapped over (MO, the blue arrows) to the close relative genomes.

Results

CNSs between *Aethionema arabicum* (*Aa*) and *Arabidopsis thaliana* (*At*)

CNSs were obtained The *Aa* genome was sequenced and annotated by the VEGI group. (HAUDRY et al. 2013), but we did not use their published CNS list (See Addendum II for a comparison between PL3 CNSs and the CNSs published by the VEGI group). The *Aa* genome was compared to *At* TAIR10 using the PL3 protocol defined in TURCO et al. 2013. CNSs were then proofed and validated through manual proofing in GEvo with customized CNS-decorated genomes, as explained in the Introduction; all in the Freeling lab (see Acknowledgements) pitched-in to find mis-assigned CNSs or other PL3 problem, thus informing some recoding. After several rounds of coding and proofing, the Genespace Tweaker (Supplemental Table 1) application was finalized. 65456 validated, revised (PL3-tweaked) CNSs were identified between *Aa-At* orthologous regions, 16885 on the 3-prime side of genes, 28965 classified as 5-prime CNSs and 19606 identified within intronic regions. Of the 5' and 3' CNSs, 22553 or 49% were detected over 500 base pairs from the gene and classified as distal CNSs and 23297 were proximal CNSs. The *Aa-At* CNSs can be viewed using the GEvo graphics module of the CoGe comparative genomics platform using the GEvo links in the CNS spreadsheet (Supplemental Table 1). See methods for the details of how the customized GEvo proofing genome, dsgid 19870 (decorated with PL3-tweaked CNSs), was constructed.

Detectability of *Aa-At* CNSs in *Br*

Aa-At PL3-tweaked CNSs ranged in length from 15 to 1986 bps, each (by definition) with a bit score ≥ 29.5 (Supplemental Table 2). Based on our previous findings (SUBRAMANIAM et al. 2013), we expected

that detectability would increase with CNS length, and it did: Figure 4 shows these data for *Br* ortholog singlets, doublets and triplets. Detectability of *Aa-At* CNSs is greater in singlets than in doublets and triplets, a result that is consistent with our previous results with *At-At* CNSs—that duplication relaxes selection-- and will not be the object of further study here.

Cartoons of CNS detectability

Supplemental Table 3 shows for each *At* gene, that has two orthologs (doublet) in *Br*, a cartoon indicating the detectability of *Aa-At* CNSs within each of the *Br* orthologs. Detectable CNSs are indicated with a '+' and undetectable CNSs are marked as '-'. Start of transcription and end of the gene body is indicated with '\$'. 'l's indicate exons. These cartoons are useful to determine the frequencies of subfunctionalization and non-functionalization of doublet genes following polyploidies, for example. Figure 5 is an example of such a cartoon.

Methods

Genespace Tweaker

CNSs identified by the CNS discovery pipeline go through a 'manual' validation process, during which a tester checks to make sure the CNS fits the earlier described definition, and has been 'assigned' to the correct gene. This is particularly true in the case of Mapping Over, where it is vital to know which CNS goes with which gene. Genespace Tweaker features a series of simple perl scripts, each designed and developed to address an "error" in the PL3 code as identified by proofing the PL3 output.. The end result is a 'tweaked' CNS list. This CNS list is not perfect, but the cartoons generated are "greater than 80% accurate" and often perfect, and were judged useful as the CNSs of the outgroup to be mapped-over to close relatives.

Our proofers and myself found the following PL3 issues (errors). 1) Problems with co-annotation. As part of the CNS discovery pipeline, the search and query genomes are co-annotated with each other's annotations to achieve reciprocal consistency in terms of the syntenous regions being identified. Depending on the quality of these annotations, sometimes, genomic regions tend to get mis-labeled as genes, which in turn confuses the CNS assignment.

2) Overlapping genespaces. A genespace is defined as the genomic region stretching between the farthest CNSs assigned to a gene (5' to 3'). In cases where the gene does not have a CNS on one side, the gene space boundary is defined by the edge of the transcription unit boundary. Our rule for genespaces is that each gene has its own gene space, which should *never* overlap with another gene space. Rules were devised to eliminate these overlaps. 3) In rare cases, miscalled CNSs (Figure 6) caused genespaces between two neighboring genes or genes on opposite strands to overlap. Once understood, these overlaps were eliminated. 3) Reassigning CNSs initially tagged as “RNA” in PL3. Part of the CNS Discovery pipeline excludes all CNSs annotated as RNAs. Our testing has shown discrepancies in several RNA calls (Figure 7), which have now been addressed. CNSs that were improperly removed were added back into the CNS list and assigned to the most proximal gene. 4) We decided to remove uORFs from the CNS list. 5) We assigned each CNS to the closest syntenic gene in arabidopsis, without exception. That is, we ignored the direction (strand) of the associated genes or their ortholog.

Mapping Over of *At* tweaked CNSs to expected syntenic chromosomal region in *Br*

For each *At* gene with a syntenic ortholog in *Aa*, we used the synteny screening blocks technique, nicknamed “quota-align” (TANG et al. 2011), to identify all possible orthologous regions in the *Brassica rapa* (*Br*) genome. Given the recent hexaploidy in the *Br* lineage (TANG et al. 2012) following divergence from the arabidopsis lineage (Figure 1), we expected to find up to three orthologous copies for each *At* gene. Of the *At* genes used for PL3, we identified at least one *Br* ortholog for 11592 of them, with 4963 *At* genes having a single detectable orthologous copy in *Br* (singlets), 4278 *At* genes with two orthologous copies in *Br* (doublets) and 1540 *At* genes with three orthologous copies in *Br* (triplets).

CNS detectability in *Br*

We expected to find the same number of orthologous copies for every α CNS within *Br*, as for the *At* gene's exons. Based on *At-Br* orthologies, we expected to find CNSs within expected orthologous positions, 25162 as singlets, 24608 as doublets and 9995 as triplets. We attempted to detect each of these CNSs within each expected

orthologous Br region. Nucleotide sequence of the gene space (expanded genomic region around and including coding regions and spanning from farthest upstream to downstream CNS) of each At gene containing one or more of these CNSs was masked for very repeated sequences (50X copies across entire At genome). The corresponding gene space of each detected orthologous region (singlet, doublet or triplet ortholog) in Br was also repeat-masked. Each At and Br orthologous gene space pair was compared using the same blast settings used for At-At CNS discovery (Thomas et al. 2007). Every blast HSP hit to the Br orthologous gene space was then screened for synteny to filter out probable noise, but we did not filter out simple sequence.

Genespace Tweaker and MO code

All code has been deposited <https://github.com/sshabari/Mapping-Over-Scripts/>. This code does not have instructions for independent installation. Rather, we include instructions on how to link this code to the CNS Discovery Pipeline (PL3)

Acknowledgements.

I would like to acknowledge the tireless testing efforts of members of the Freeling lab including my professor, M. Freeling and lab researchers D. Burgess and M. Woodhouse. Their incredible support allowed me to identify ‘bugs’ within the CNS datasets and to develop potential solutions that were, again, returned to them to proof again.

Figures

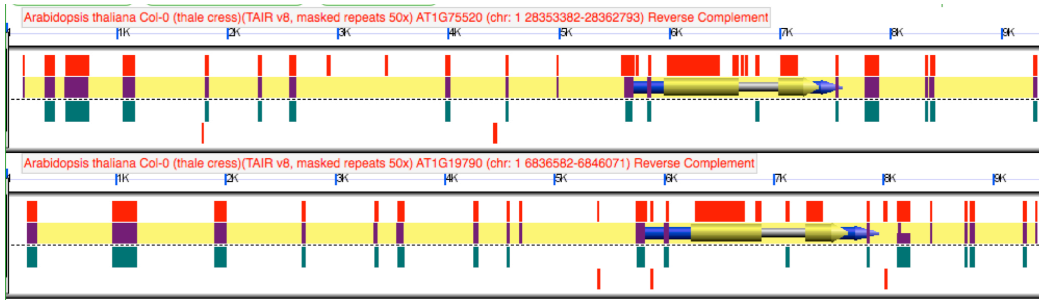


Figure 1. A CNS-rich pair of genes in Arabidopsis, retained from the most recent (a) paleotetraploidy, compared as sequence using “find CNS” blastn settings and displayed in the GEvo viewer. Panels of genomic regions (which can be regenerated at <http://genomevolution.org/r/4db1>) annotated using the GEvo application in the CoGe suite of tools (<http://coge.iplantcollaborative.org>). The figure compares an *At* gene (*AT1G75520*), a member of SHI transcription factor gene family and its homeolog. Blastn HSPs between the two genes (orange rectangles), manually updated aCNSs (the “gold standard” purple blocks on upper model line; V2, Supplement Table 1) and CNSs detected using automated PL3 CNS pipeline (green blocks on lower model line; Supplement Table 3) are annotated in this figure. Note the similarity of the two independent CNS annotations, and how the HSP data (orange rectangles) in this experiment generally supports CNS calls

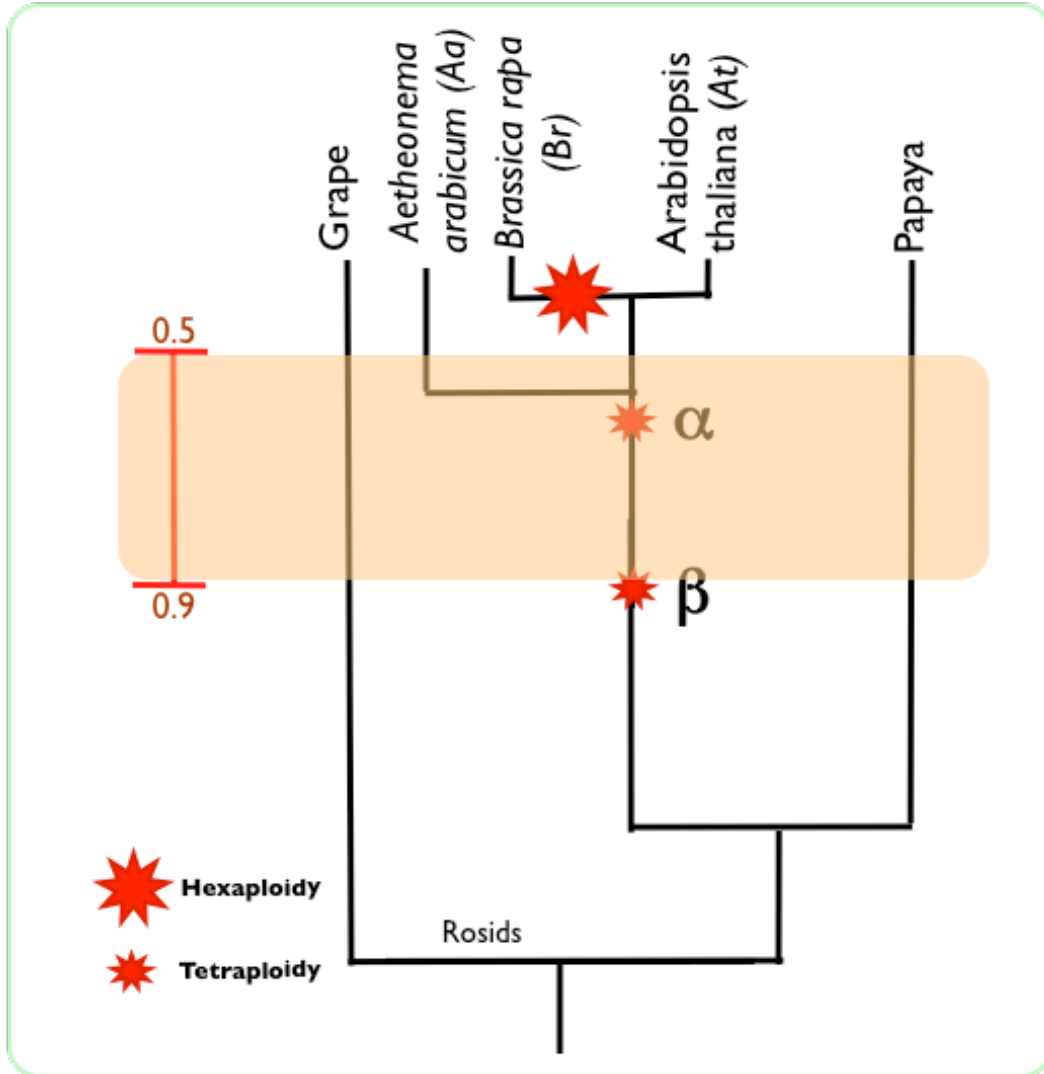


Figure 2. Cladogram showing approximate phylogenetic relationships between *Aethionema arabicum*, *Arabidopsis thaliana* and the paleohexaploid *Brassica rapa*. The Ks 0.5-0.9 window for useful CNS discovery is colored orange. Notice that the paleohexaploidy is outside of this window.

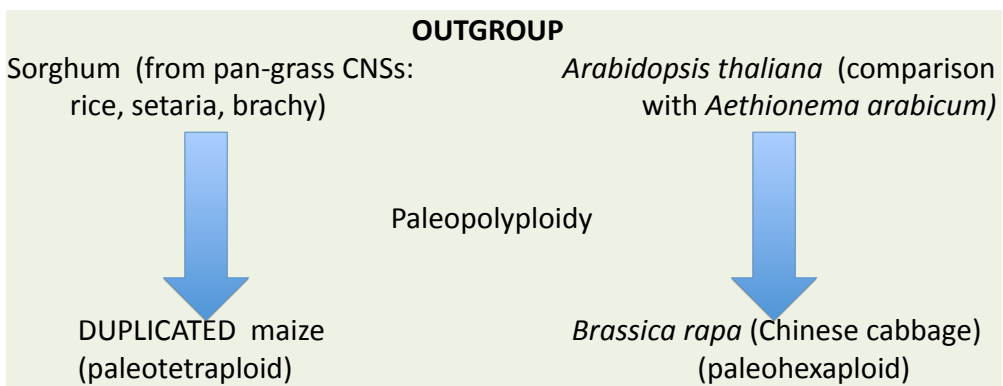


Figure 3. Strategy of the Mapping Over pipeline. Both maize and *Brassica rapa* are not diverged enough from their outgroup to permit direct CNS discovery by PL3. Therefore, CNSs in the outgroup are mapped over (blue arrow) to the polyploid relative.

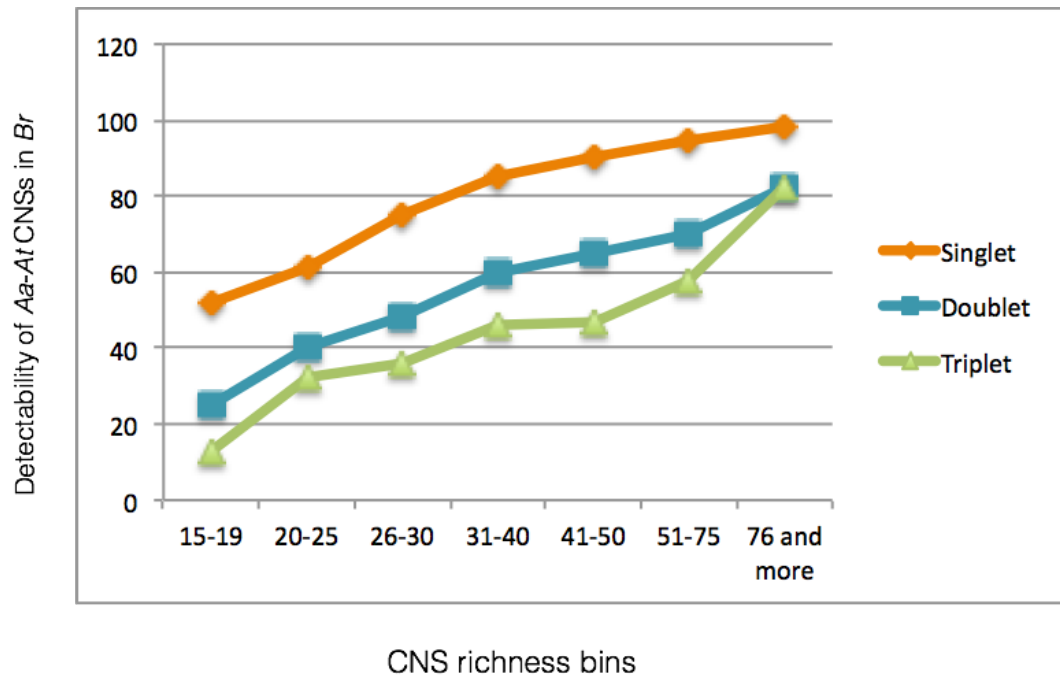
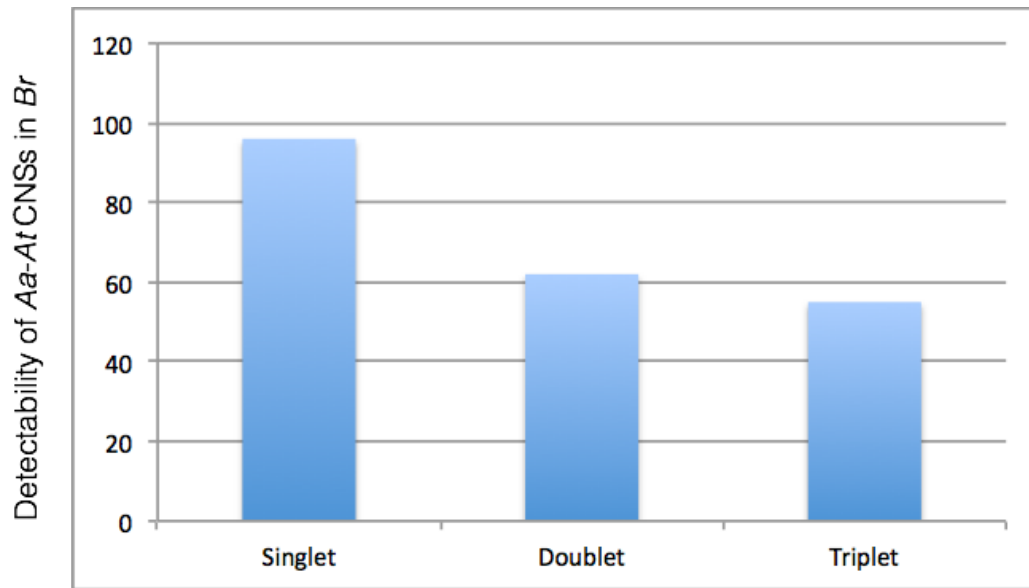


Figure 4. Comparison of detectability of *Aa-At* CNSs in *Br* over bins of CNS richness.



Expected number of copies of *Aa-At* CNSs in *Br*

Figure 5. Comparison of detectability of all expected copies of *Aa-At* CNSs (over 30 bases long) in *Br*. Duplication certainly relaxes selection on CNSs.

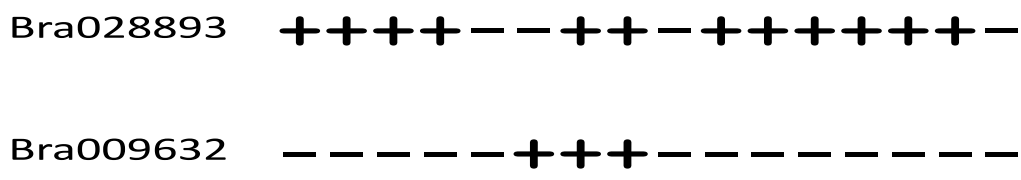
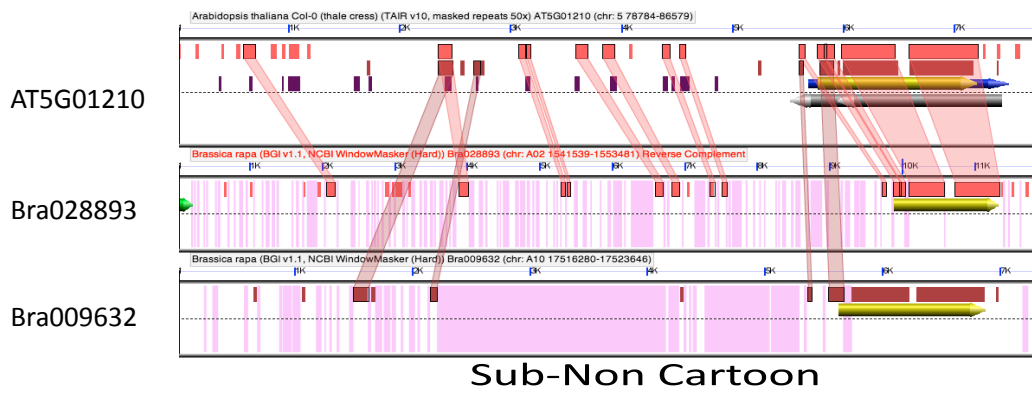
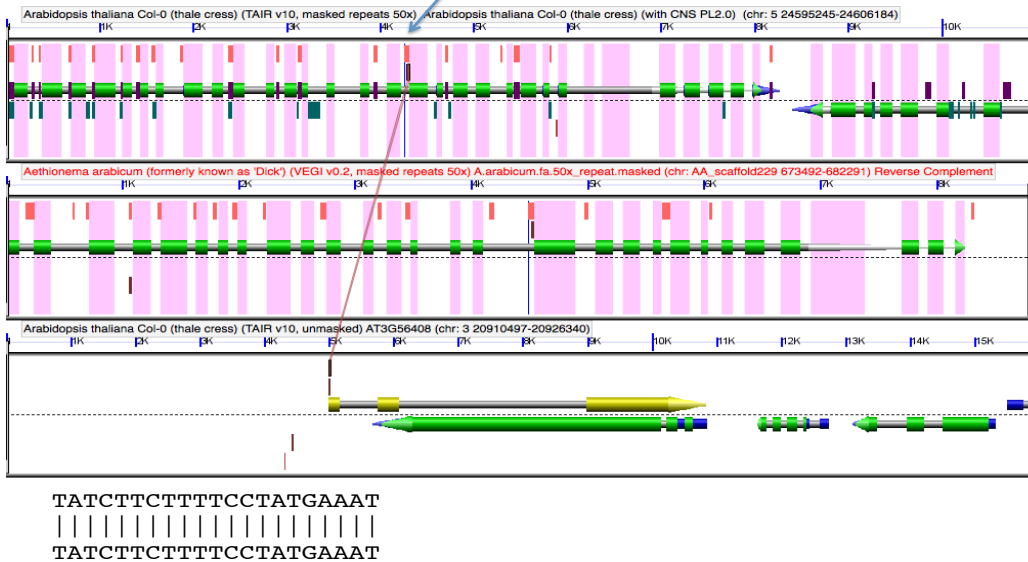


Figure 6. Example of GEvo panel of detectability with cartoon at the bottom.



Figure 7. <http://genomeevolution.org/r/8zvq>. Arrow points to the bogus CNS that, because it was called causes the gene spaces to overlap. Errors like this were corrected.

invalidated as *cns_rna* (single hit to overlapping antisense of an unrelated gene)



A cluster of new *cns_rnas* called due to *hsp*s with antisense RNA overlapping homeologous gene

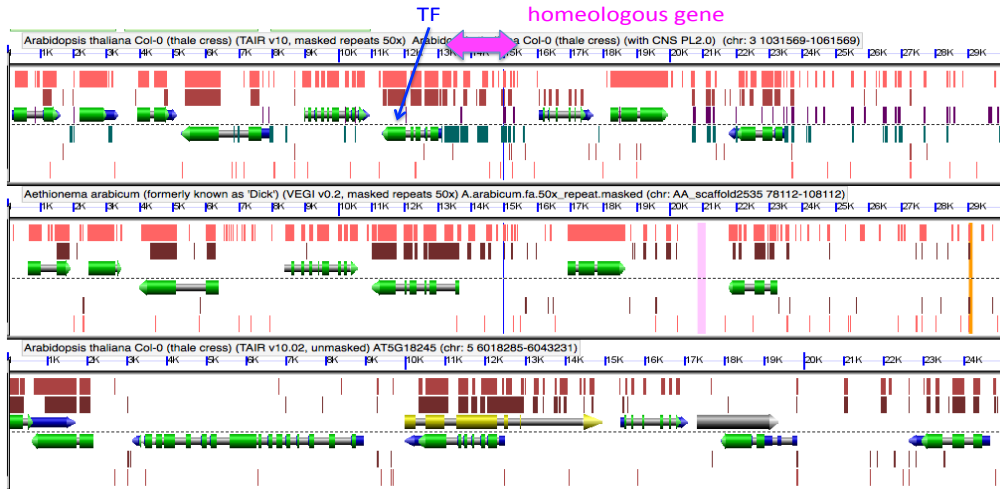


Figure 8. <http://genomevolution.org/r/91v1>,
<http://genomevolution.org/r/9365>

Supplemental Tables

Supplemental data available for download at:

http://figshare.com/articles/Computational_pipeline_to_track_the_evolutionary_fates_of_arabidopsis_conserved_noncoding_sequences_through_a_paleohexaploidy_and_into_the_genomes_of_its_close_relative_Brassica_rapa_PL3_Genespace_Tweaker_Mapping_Over_Supplemental_Data/856778

- S1. Output of reassigned Aa-At CNSs with GEvo panels
- S2. Output of MO pipeline I: CNS detectability
- S3. Output of MO pipeline II: Cartoons with GEvo links.

Chapter 4: Subfunctionalization, nonfunctionalization and cis-acting modules: arabidopsis conserved noncoding sequences used as markers to quantify fractionation patterns following the hexaploidy in the *Brassica rapa* lineage.

Introduction

Gene duplication of any sort results in a relaxation of selection on retained duplicates. This essential consequence of gene duplication was used by E.B. Lewis to explain, in theory, how new genes, “neomorphs”, might arise from preexisting genes (LEWIS 1951): duplication followed by relaxation of selection followed by divergence. This scheme set the rules for understanding the evolution of the first *HOX* gene clusters and segment identity in vertebrates (GARCIA-FERNANDEZ 2005). While it was known for decades that duplicate genes could acquire complementary functions, as with some duplicate genes involved with anthocyanine pigmentation in the maize leaves-culm versus aleurone (STADLER, 1951), it was not until the concept of subfunctionalization was introduced (FORCE et al. 1999) that the potential power of gene duplication became undeniable. Figure 1 is a simple diagram of a gene with two independent cis-regulatory elements E1 and E2, functioning to express product in cell-type 1 and 2, respectively. Figure 1 defines graphically the terms subfunctionalization and nonfunctionalization as they apply to alternate mutation patterns exhibited by a retained duplicate.

Subfunctionalization was originally proposed as an explanation for the over-retention of duplicates following whole genome duplications. The idea is that, once a gene pair is subfunctionalized, it is made permanent (LYNCH and FORCE 2000). Subsequent work has shown that subfunctionalization is not the only or even the general explanation for changes in gene content following whole genome duplications (FREELING 2009)—the preservation of gene product balance is more general— but subfunctionalization certainly should preserve duplicate genes when it occurs, subfunctionalization just happens as a part of neutral drift, and subfunctionalization has been shown to be important to understand gene diversification and potential for fine-tuning (ROTH et al. 2007). The *HOX1* subfunctionalized genes in the mouse have been dissected and an ancestral reconstruction has taken their places seemingly without ill effects (TVRDIK and CAPECHI 2006). Subfunctionalization has been proposed to be a step

toward the evolution of novelty (neofunctionalization) (HE and ZHANG 2005; RASTOGI and LIBERLES 2005).

Brassica rapa v. Chiifu (*Br*) has three subgenomes as compared to the arabidopsis genome because of a hexaploidy occurring in its lineage (WANG et al. 2011): The LF subgenome (III) has more genes than the other two (MF subgenomes, or I and II) and those genes located on subgenome III tend to express to higher message levels as compared to their retained homeologs (CHENG et al. 2012), thus expressing a phenomenon common among ancient polyploids called “genome dominance” (SCHNABLE et al. 2011b; GARSMEUR 2013). In a study that presages this paper, we (SUBRAMANIAM et al. 2013) used the published conserved noncoding sequences (CNSs)—derived from comparing the alpha genomes within arabidopsis (*Arabidopsis thaliana*, v. Columbia or col= *At*)— and asked “what becomes of these *At* CNSs in *Br*?”, where they are expected to be present whenever their associated gene is present. It was shown that duplication of genes and their CNSs clearly relaxed selection on any one CNSs, and that, when a CNS was not detected, that was because the sequence was deleted (84% of the time) rather than going undetectable by disintegration by base substitutions or indels. The idea of beginning with CNSs associated with expected genes in *At* and then following them to expected syntenic positions in *Br* inspired this study. However, here we begin with many more CNSs because we compared the arabidopsis genome with that of *Aethionema arabicum* (HAUDRY et al. 2013) an early diverging family member and separated from arabidopsis by a *Ks* of 0.65 which is excellent for CNS discovery (SUBRAMANIAM and FREELING 2012) using our Pipeline 3.0 (TURCO et al. 2013). We then take each CNS, adjust its relationship to exons and other CNSs, and then “map over” each CNS to the expected one, two or three orthologous positions in *Br*; when present, the *Br* gene gets a “+” and when undetectable, a “-“. The CNS cartoons we generate for all gene (homeolog) doublets in *Br* comprise the primary graphic results of this study. The patterns of the +’s and -’s permit sorting these doublets into categories: subfunctionalization (s), nonfunctionalization (n), afunctionalization (a=ancestral), and “loss of one CNS only”(a-1). Neofunctionalization, or novelty, is not visible in this experimental design because we do not here attempt to reconstruct the *Br* ancestral CNS patterns. Although CNSs are, by definition, functional, here

CNSs are largely used as markers for cis regulatory space, so it should not be over-troubling that we rarely know what any CNS actually does.

Based on past results and common sense, we had some expectations.

1) The frequency of subfunctionalization should increase as the number of CNSs per gene increase. There are simply more possibilities to subfunctionalize at the CNS + or – level. Genes that encode response-to transcription factors, “bigfoot genes”, tend to be CNS-rich (Freeling et al. 2007). 2) The subgenomes of *Br* are different from one-another; genes on III (LF) are more dominant than genes on II (MF2) than genes on I (MF1). 2) We expect that nonfunctionalization of CNSs will tend to happen on the recessive genome of any doublet. Sometimes nonfunctionalization is dramatic, where many CNSs are removed from one gene while the homeolog remains unmutated (ancestral); these cases should display an exaggerated proclivity to be on recessive subgenomes. 3) Most importantly, we should be able to identify cis-acting modules composed of multiple, adjacent CNSs. The first report of CNSs in plants contained the suggestion that CNSs might work together to bring into proximity proteins assembling into components of the transcriptosome (KAPLINSKY et al. 2002), components that would otherwise not assemble in the nuclear soup. Our experiments rest on the following assumption: if a stretch of CNSs perform one and only one function, and one CNS is lost, the others will soon be deleted as well, leading to a run of CNS loss. This is no different from the expectation that, if a gene’s exons are impaired, then the rest of the gene and all of its CNSs will soon be deleted. The mechanism of duplicate DNA loss, the fractionation mechanism, is known to be aberrant intrachromosomal recombination removing several bps of sequence between short tandem repeats, but rarely if ever removing gene-sized pieces of DNA (WOODHOUSE et al. 2010; TANG et al. 2012).

Following mapping over of *At* CNSs onto *Br*, our aim is to compute a cartoon (E.g. +++--+-+++++ 5’ RNA ++++ RNA 3’ +++-+) for each gene, and for each doublet. These cartoons become the pre-existing mutant collection used to make sense of community RNAseq data as it accumulates and gains in specificity. This general aim to correlate CNS loss with changes in RNAseq expression profile is called “fractionation mutagenesis” (FREELING et al. 2012).

Results

Many *At-Aa* CNSs do not map over to any *Br* ortholog.

26,208 *At-Aa* CNSs have been sorted to those 4354 arabidopsis genes that have *Br* doublets. Therefore, there are 52416 positions on these *Br* doublets for a CNS to be detectable as a sequence (+) or not detectable as a sequence(-). In reality, there are 15,718 *Br* CNSs—representing 7859 arabidopsis CNSs, that are present at neither *Br* position in the doublet (0/0 CNSs); 30% of arabidopsis CNSs do not exist at all anywhere in *Br* (,including at the expected orthologous positions of either *Br* homeolog).

Supplemental Table 1 shows the results of mapping over of *Aa-At* CNSs to *Br*. Over 78% of CNSs that are not detectable in either *Br* homeolog have a bit score (original bit score from PL3) between 29.5 (the lower threshold of CNS detection) and 35. Also, over 75% of these CNSs are shorter than 30 bases. It's possible that a majority of the 0/0 CNSs are not detectable in either homeolog because they were on the borderline of detection to start with. We also find a strong correlation between genes with 0/0s and genes with out-paralogs (closely related paralogs). So the reason the 0/0 genes and their CNSs might be mutating into nondetectability is that they are just above our significance threshold combined with consequences of the paleohexaploidy and *the preceding polyploidies in the lineage* generating redundancies via gene families that cover essential gene functions.

The alternative classes of MO CNS fractionation cartoons of *Brassica rapa* v. *Chiifu* doublets.

Supplemental Table 2 is our Master Spreadsheet organized around 8708 *Br* genes present as doublets, each doublet aligned with one of 4354 arabidopsis orthologs. One of the two genes in the doublet is from a more dominant subgenome, given the dominance order as III is more dominant than is II than is I. Each arabidopsis ortholog has 0-63 CNSs distributed around the noncoding genespace. The methods section describes how the MO pipeline begins with the two outputs of our CNS Discovery Pipeline (PL3) and makes some adjustments to permit the precision needed to print proofable, realistic fractionation cartoons. Figure 2A is a screenshot arranged from Supplemental Data 1 showing (A) A list of doublets with fractionation pattern calls on

them, and (B) a “typical” doublet of *Br* genes with 6 arabidopsis CNSs (*At1g22250*) that have fractionated by balanced subfunctionalization (*se/se*); the GEvo panel (Fig 2B; re-generated at <http://genomeevolution.org/r/9ukz>) provides blastn evidence for these 6 CNSs and how one is lost from each gene. This proofing is facilitated by a special CoGe edition of TAIR10 with CNSs decorating the model line. Figure 2C is a qTeller-Brassica screenshot of a two homeolog plot of FPKMs derived from all *Br* RNAseq data including these two *Br* genes; regenerate these data http://qteller.com/brassica2/scatter_plot.php?name1=Bra012317&name2=Bra031377. Chiifu data (from the Wang lab, Beijing) is identified by citation and color-coded green in qTeller-Brassica. Most of the data on this plot are from the B3 inbred line of *Br*, and may or may not reflect our Chiifu inbred line cartoons. We used the tools GEvo and qteller-Brassica, using links embedded in Supplemental Table 2—our Master Spreadsheet-- to proof Map Over itself, to proof our cartoons, and to check on gene expression, especially if a dead gene is suspected. There’s no need to trust our code; each cartoon is an automated estimate of what any researcher can evaluate using the GEvo link and/or qTeller link provided.

Each of the 4354 *Br* doublets exhibits one of six sorts of fractionation patterns (Examples in Figure 1A): a/a (ancestral; no fractionation), a/a-1 (one homeolog has one CNS deleted), 0/0 (no arabidopsis CNSs to fractionate), s-e/s-e (even # of subfunctionalized CNSs), subfunctionalized-dominant/subfunctionalized (s-d/s: subfunctionalized but one homeolog has more CNSs than the other), nonfunctionalized-dominant/nonfunctionalized (n-d/n: one homeolog is ancestral, and the other has more than one CNS removed). Supplemental Table 1, if sorted by Column A, will generate every doublet with all fractionation categories calls recorded with links to proofing tools. Supplemental Figure 1 is a pie chart distributing all doublets into these categories. The most populated categories are ancestral (a/a) and where one gene of the doublet has one CNS deletion (a/a-1) categories.

Subfunctionalization-nonfunctionalization data for *Br* doublets with 2 CNSs: Even the simplest case gives complex results.

Doublets with 2 CNSs represent the simplest test collection of doublets with which to test the fundamental rules governing fractionation after a

paleopolyploidy. Figure 2 is such a two-CNS case where both are 5'. Table 1 shows example cartoons of subfunctionalization and nonfunctionalization in 2 CNS cases where the CNSs are linked (l) — either both 5' or both 3' — or split, in cases where an exon separates them. Cases where both CNSs were located in intron were invalidated and ignored for this 2CNS experiment). Linked subfunctionalizations are significantly below expectations and linked nonfunctionalizations are significantly above random expectation compared to the control at 1.1 linked/split for all cases of ancestral 2CNS doublets (1.1 is the expected ratio in the absence of any bias). We will go on to test our hypothetical explanation: the paucity of linked subfunctionalizations is because, in some cases, the two CNSs are components of a cis-acting regulatory module (CRM) carrying out a singular function.

Doublets subfunctionalize increasingly as their CNS/gene numbers increase.

The % of doublets that subfunctionalize goes up dramatically as CNS counts go up. For 2-CNS doublet genes, subfunctionalization is at 9.2% of total doublets (122/1224), which rises to 53.7% (378/648) for genes with 8-10 CNSs, to 72% (298/412) for genes with 11-19 CNSs, to 84.1% (148/176) for genes with 15-20 CNSs and to a maximum of 87.0% (40/46) for those 23 doublets that have 21-61 CNSs. Since only one CNS subfunctionalization is enough to subfunctionalize a doublet that was otherwise ancestral or nonfunctionalized, our result proves that our reasonable expectation is indeed true. This is not a trivial hypothesis since subfunctionalized doublets should render a retained pair permanent (FORCE et al. 1999; LYNCH and FORCE 2000), and transcription factor genes are notoriously rich in CNSs and are also over-retained post paleopolyploidy (as will be discussed).

Genome dominance and nonfunctionalization; the ancestral (intact) gene of a nonfunctionalized doublet tends to be on the dominant subgenome of the doublet, as expected.

Each nonfunctionalized doublet is classified in Supplemental Data 1 as a n-d/n doublet; the d denotes “dominant” in the sense of resisting fractionation. When n-d is actually on the expected (dominant) subgenome, then that gene is noted “e” for expected on the Master Spreadsheet. Expectations for subgenome are met 1.94 to 1, or twice as often as expected by chance alone. If the nonfunctionalization is dramatic — where n-d has 4 or more CNSs excess over the mutated

homeolog (n), then the expectations are met 5.6 times more often than they would be if subgenomes were randomly distributed. Genes on recessive subgenomes tend to lose CNSs

A significant subgenome effect on nonfunctionalization is seen in the 242 simplest sort of nonfunctionalized genes: the 141 n-d genes from the 2 CNS doublets of Table 1. The chances of an n-d gene being on the expected dominant subgenome (III dominates II dominates I) is 1:1 by chance alone and 76:45 (1.7:1) for n-d genes.

Even the deletion of one and only one CNS from a doublet (a-1 genes) tends to happen more often than not from a recessive subgenome, as compared to ancestral genes (a from a/a or a/a-1 doublets)(Supplemental Table 1). However, this subgenome effect is relatively small compared to those exhibited in nonfunctionalized doublets.

The excess of deletion runs for linked CNS is universal over the entire dataset.

We began by counting all of the expected CNS positions in *Br* doublets (after removing the 0/0 CNSs positions) for the entire doublet dataset. This was used as the denominator. For the numerator, we counted all the –'s (minuses) in all the cartoons. This ratio allowed us to calculate the single CNS deletion frequency (0.17). On the basis of this frequency we then calculated the expected distribution of runs of 1,2,3,4 and 5. This is all without correcting for the reality that gene noncoding space is not infinitely long—so our actual runs data is expected to be shorter than expectations because “runs are naturally cut off if they are near an end”. This won't matter because our runs, known to be under-estimates, are very significantly longer and occur at a higher frequency than expected by chance alone, as shown in Table 2.

Cis-acting modules often contain duplicated TF-binding motifs.

We looked at the enrichment of known TFBS motifs within *At-Aa* CNSs that are deleted in runs in *Br*. Of all *Aa-At* CNSs within our dataset, 120 CNSs were deleted one-at-a-time when doublets were evaluated. We have only considered doublets so there would not be complications involving different degrees of relaxation of selection based on redundancy. Within the “runs” CNSs, we searched for all

instances of 221 known plant transcription factor motifs. (The entire motif list with a reference as to source publication can be accessed at: <http://coge.iplantcollaborative.org/CoGe/tmp/MotifView/Motiflist.txt>). We found at least one instance of 80 of the 220 motifs within *At* CNSs where one was lost in the doublet. 34 motifs, each at least 5 bases long and without ambiguity in sequence were selected for further analysis (Supplemental Table 3, Column A). These motifs had only one base represented in each position, no choice between any two bases was permitted, and no n's indicating a choice of any of the four bases. So, using regular expressions is justified without the need for weighted matrices. We used all *At* CNSs where one was lost in *Br*, *but in a run of one*, as control (Supplemental Table 3, Column C) for deriving the expectations for each motif within CNSs deleted in runs of two and three. Within the *Aa-At* CNS dataset, there were 42 pairs of CNSs found undetectable in runs of two and 22 sets of CNSs lost three-in-a-row.

We asked whether or not these sequential CNSs tended to carry reiterated motifs as compared to CNSs within motifs lost one-at-a time. Columns D and E show the ratio of the observed number of each motif in two-in-a-row and three-in-a-row undetectable CNSs. We found that 29/34 motifs at least 5 bases long, with at least one representation in a single loss CNS, were represented between 1-16 times higher than expectations based on CNSs lost one at a time (Column D). 23/34 motifs were found to be overrepresented 2-21 times in CNSs lost three-at-a-time compared to control (Column E). 18/34 motifs were overrepresented in both experimental CNS sets. For example, the G-Box, the most highly enriched of motifs in CNSs was found to be overrepresented 2-fold over expectations within runs of two deleted CNSs and 7-fold over expectations within runs of three deleted CNSs (Supplemental Table 3, Row 17). All ratios are significantly different ($p < 0.05$) from a 1:1 expected by Chi-square.

CNSs occurring in runs present an opportunity to find motif patterns specific to cooperative binding principles that may direct the assembly of transcriptomes or their components.

As will be discussed, the over-abundance of CNS loss in runs likely means that particular patches of CNSs are performing a single function. That runs are enriched in duplicated motifs reinforces the importance of these runs; when motifs are reiterated—much like when

stop codons are reiterated, the importance of that motif becomes undeniable. Thus, a comparison of CNS runs sequence and motif patterns with control sequences and patterns presents an opportunity to discover new cis-regulatory binding functions, new CRMs. Comparing patterns of motifs is not as straightforward as comparing sequences. A collaboration with others in the Freeling lab has begun toward this end in order to follow the results of this chapter to the sort of exciting conclusions that could be the impact of its eventual publication.

Discussion

Three expectations were introduced previously. Each has been realized.

Our first expectation was that subfunctionalization frequency should increase as number of CNSs/gene increases. Indeed, this is the case, from 9.2% for genes with 2 CNSs to a maximum of 87% for genes with over 26 CNSs. Genes encoding transcription factors, and especially those with “responds to...” GO terms, are known to be particularly rich in CNSs (FREELING et al. 2007), and are also retained post whole genome duplications, as reviewed (FREELING 2009). Many genes that have no CNSs at all, such as genes encoding ribosomal subunits, are also highly retained post polyploidy. The reciprocal relationship between being retained post tandem duplication versus post tetraploidy (FREELING, 2009) proved that the general mechanism driving retention was not subfunctionalization, but the preservation of gene product balances. That gene encoding transcription factors tended to be retained could well be that the transcriptosome is a complex, subunit-interactive machine with potentially important subunit:subunit stoichiometries. Alternatively, the stoichiometry of transcription factor-DNA binding site could also constitute a selected balance (SCHNABLE et al. 2011a). However, the original function proposed for subfunctionalization—gene pair preservation post-polyploidy (LYNCH and FORCE 2000)—while not the general mechanism for retention, is the preferred mechanism of retention for genes that are subfunctionalizable. This subfunctionalizability, if the cause of retention, should operate for tandemly duplicated genes as well as gene duplicated following tetraploidy, assuming that all of the gene is tandemly duplicated! Having many cis-acting sites, if they are actually present, certainly confers upon a gene a proclivity to subfunctionalize. There is no

reason some categories of genes might be highly retained post-polyploidy both because they subfunctionalize their cis-regulatory sites and because their products participate in dose-sensitive interactions.

Our second expectation was that nonfunctionalization should have its deletions focused on genes on the recessive subgenomes. Specifically, degree of CNS deletion should decrease in the order of decreasing subgenome dominance: III is dominant over II is dominant over I. That is the case. There is a small tendency for the *Br* lineage to remove any single CNS from recessive subgenomes rather than subgenome III (Figure 2) and there is a huge 5.56-fold chance that nonfunctionalizations involving ≥ 4 CNSs will be from the expected, recessive subgenome. We conclude that there is a natural tendency in those ancient polyploid lineages that derive from allotetraploidy, but not autotetraploidy (GARSMEUR 2013), to favor nonfunctionalization over subfunctionalization simply because subgenomes do not express their genes to the same levels. Therefore, homeologous genes are simply not equally “important” to purifying selection.

Our third expectation is the most important, and best directs us in new research directions. The Introduction documents seminal studies on motifs within mammalian CNSs and evidence for combinatorial gene cis-regulation; a recent paper on this subject does double-duty as an excellent review (GUTURU et al. 2013). Our expectation is, if adjacent CNSs really contain motifs that work together as a cluster to serve as a cis-acting regulatory module (CRM) of singular function, loss of any one CNS should rapidly expand to a loss of all CNSs in the CRM. The experimental endpoint is expected to be an over-representation of runs of adjacent CNS deletion. This tendency was dramatic. Even the simplest case of genes with two CNSs show that adjacent CNSs were significantly more likely to be lost together than when two CNS are located with an exon in between (Table 1). Using our entire dataset of *Br* genes in doublets, runs of CNS deletion occur vastly more often than predicted by the chance accumulation of CNSs lost one at a time (Table 2). We interpret these “runs” of CNS deletion as defining cis-acting modules of function. Since we know that the fractionation mechanism in maize is the accumulation of small (5-15 bps) deletions—with no evidence of multiple gene deletions-- (WOODHOUSE et al. 2010), there is no basis for thinking that an obligatory very long deletion mechanism somehow results in these

runs of adjacent CNS loss. It is encouraging that these runs of CNS loss are particularly prone to duplicate within their CNSs particular, known DNA-binding motifs. Most reiterated, specifically in *At* CNSs deleted in runs, are CACGTG (G-box), TGACG, and TATAAT (Supplemental Table 3).

It is reassuring to have a future direction of these results so clearly indicated. The laboratory of G. Bejerano (GUTURU et al. 2013) is combining motifs as estimated by positional weight matrices with structural annotations of actual protein complexes known to bind these motifs. Mammalian systems are far ahead of plants in knowledge of actual protein-promoter structure. This approach is called “3D Positional Weight Matrices.” We are aiming to do the same, with the addition of using fractionation mutagenesis involving runs of co-deleted CNSs to identify particular clusters of motifs within CNSs as candidate cis-acting modules (CRMs). Of course, using forward genetics to identify CRMs assures the singularity of function crucial to making sense of the molecular details of protein-protein and protein-DNA binding that constitute the reality of cis transcriptional regulation. We expect plant systems, with their frequent polyploid outgroups, will be particularly useful in understanding the rules of combinatorial cis-acting gene regulation.

List of Supplemental Documents

All supplemental data available for download at:

http://figshare.com/articles/Chapter_4/868859

Supplemental Table 1. The problem of arabidopsis CNSs that do not map over to Brassica

Supplemental Table 2. Master Spreadsheet of cartoons

Supplemental Table 3. Motif enrichments in runs of deleted CNSs.

Methods

Preparing our automated CNS discovery pipeline output for Map Over (MO) from *At* to *Br*

Conserved noncoding sequences (CNSs) in *Arabidopsis thaliana* v. Columbia (Genome version TAIR10) were obtained for each gene that had a syntenic ortholog in the sequenced genome of *Aethionema arabicum* (*Aa*) using the CNS Discovery Pipeline PL3.0 (Turco et al.

2013). The crucifer genera *Arabidopsis* and *Aethionema* are about 50 MY diverged with a modal K_s of about 65%. This falls within the ideal divergence window for CNS Discovery using the CNS Discovery Pipeline PL3.0.

The CNS Discovery pipeline generates two lists as its primary output, one a pairs list indicating the *At-Aa* gene pairs as identified using the QUOTA-ALIGN algorithm and a CNS list for those *At-Aa* orthologous pairs. We coded Map Over (MO) to take these two lists, adjust their content for complete pairing, precision of CNS positions within genespace, and then asked, for each *Aa-At* CNS: Is the CNS detectable in any of the expected (up to three) syntenic, orthologous positions in the paleohexaploid *Brassica rapa*-Chiifu genome? If detectable, the presence of the CNS is indicated with a “+” sign and if not detectable, by a “-“ sign. MO is a pipeline we developed expressly for this study, but is generally applicable, as described in Chapter 3.

Comparing *Br* directly with *At* is problematic because our set CNS blastn cutoff of 29.5 bitscore (15/15 exact match and syntenic) finds too many similar sequences that were simply carried-over from the ancestor without being selected. If we tighten our CNS criteria, we lose comparability. If we compare *Aa* with *Br* we lose the excellent annotation of arabidopsis. So, we obtain our CNSs first using our best methodologies and then map them over onto relatives. As more Brassica and related genomes are sequenced, these same *At* CNSs will be Mapped Over using the identical pipeline (MO), and the output will then be comparable.

The Mapping Over (MO) Pipeline

The Mapping Over Pipeline is available for download from: http://figshare.com/articles/Computational_pipeline_to_track_the_evolutionary_fates_of_arabidopsis_conserved_noncoding_sequences_through_a_paleohexaploidy_and_into_the_genomes_of_its_close_relative_Brassica_rapa_PL3_Genespace_Tweaker_Mapping_Over_Supplemental_Data/856778

Figures

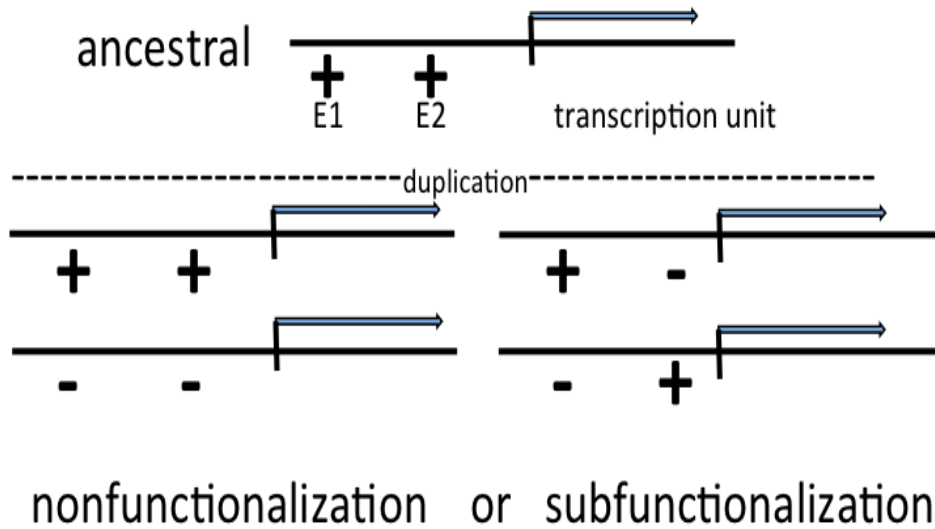


Figure 1. Graphic definition of subfunctionalization and nonfunctionalization. E1 and E2 are “enhancers” driving expression in different cells in the organism. Post duplication gene pairs suffer two + to - mutations to no function. The patterns of these two mutations are definitive.

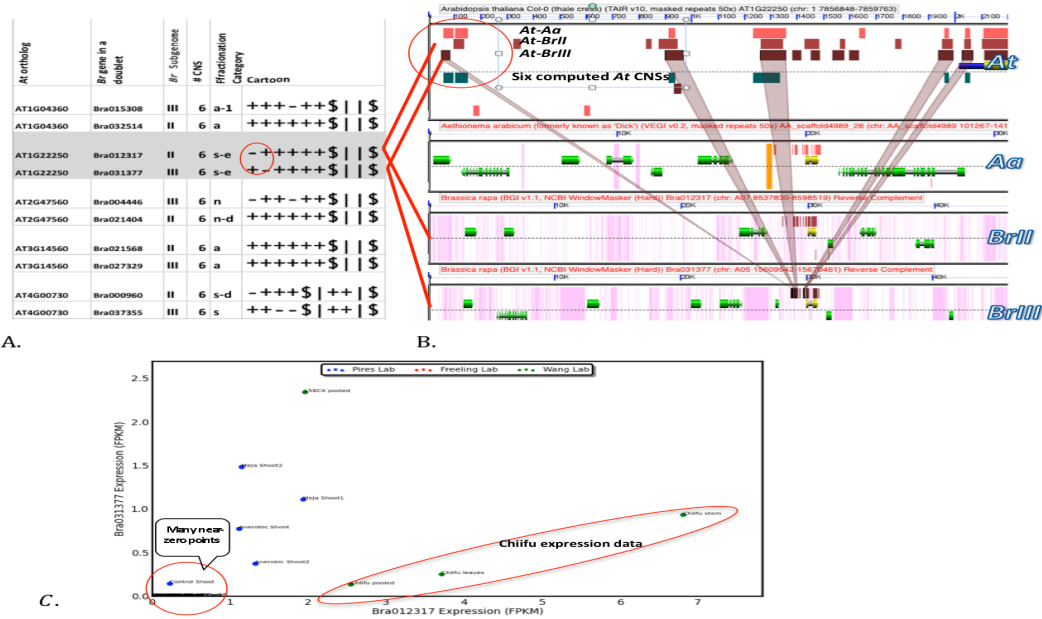


Figure 2. The alternative ways a *Br* doublet can fractionate, and how cartoons were proofed and gene expression evaluated using links embedded in the Master Spreadsheet (Supplemental Data 1). A. A screenshot of selected columns from the Master Spreadsheet, with one *Br* doublet—the orthologs of *At1g22250*—highlighted grey. “\$|” designates the most 5’ exon and |\$ the most 3’; CNSs between these symbols are in introns. For the highlighted doublet, “s-e” means that these are subfunctionalized evenly, as diagrammed in the cartoon. B. This subfunctionalization is seen in this screenshot (circle) of a GEvo blastn comparison of *At1g22250* in arabidopsis (*At*), its *Aethionema* (*Aa*) ortholog, and its two *Br* orthologs (identified by subgenome), with hits (HSPs) displayed onto the arabidopsis gene, with computed *At/Aa* CNSs decorating the model line as six green rectangles. C. The x axis contains FPKM data for RNA levels rendered in the qTeller-*Brassica* pipeline (instance 10-7-2013) for *Bra012317* on BrII. The Y axis is homeologous *Br* gene *Bra031377* expression reflecting the BrIII panel of B. and the +-++++\$| cartoon in A. See text for Fractionation categories and the GEvo and qTeller links that regenerate the graphic data of B and C, respectively.

Distribution of 4354 *Br* doublets into fractionation categories

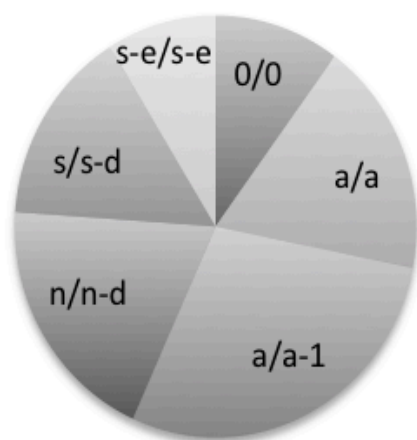


Figure 3. The distribution of categories of doublets from Supplemental Table 1, our Master List. Pie-chart distribution of 4354 *br* doublets into fractionation categories.

Deletion of one CNS only per doublet is biased to recessive subgenomes

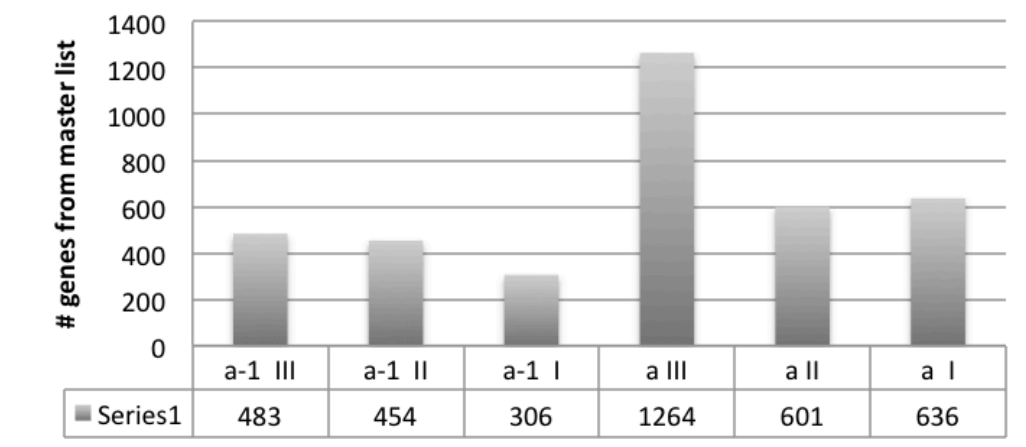


Figure 4. Genes with one CNS deleted from a/a-1 doublets-- a-1 genes-- are located preferentially on the recessive subgenomes II and III of *Br*. The control genes for this experiment are ancestral (a; unfractionated) from both a/a and a/a-1 doublets. III is the dominant subgenome. The deletion of one and only one CNS from a doublet tends to be significantly biased to a recessive subgenome.

Tables

Mode of Fractionation	# Linked	*# Split	Linked/Split Expect 1.1	Linked + Split
*Sub	51	71	0.7	122
*Non	150	93	1.6	242
**Ancestral (a)=control	322	288	1.1	610

Table 1. Number of genes among doublets with 2 CNSs that are either subfunctionalized or nonfunctionalized, and whether these two CNSs are linked or split.

*Sub=subfunctionalization (+ -/ - + or - +/+ -); non-nonfunctionalization (+ +/- - or - -/+ +) and “split” means that at least one exon divides the CNSs in arabidopsis.

**Ancestral genes (from a/a and a/a-1 doublets) serve as a control for the unfractionated linked:split ratio.

X: CNS deletion runs of length X	#doublet cartoons in <i>Br</i>	*Expected number of cartoons with at least one CNS deletion of X length	Observed number of cartoons with CNS deletion runs of X length	**Obs/Expected
1	8710	3825	4208	1.1
2	6106	179	865	4.8
3	4740	24	223	9.3
4	3592	3	93	29
5	2774	<1	51	123

Table 2. Observed versus expected number of doublet cartoons with 'X' deletion runs.

*Expectations are based on the single CNS deletion frequency of 0.17, assuming that each CNS is deleted independently of any other.

** All ratios except for single run deletions are significantly different ($p < 0.05$) from a 1:1 expected by Chi-square.

Chapter 5: MotifView (in CoGe): Visualization tool for comparative analysis of motifs and ENCODE-like features.

Preface

This chapter describes MotifView (<http://genomevolution.org/CoGe/MotifView.pl>), the latest in the suite of tools within the CoGe Comparative Genomics platform. When I joined the Freeling lab in 2009, I was faced with the challenge of extending the power of CoGe's superior data comparison and visualization to include viewing and comparing annotations as small as a motif. A motif can be as short as five bases. This is not a simple problem. A typical region visualized in GeVo, CoGe's Genome Visualization module, is several kilobases (kb) in length and the shortest possible window, for meaningful comparative analysis, is more than 250 bases wide. The graphics module within CoGe had to be adjusted so that motif graphics could be scaled based on window size.

There were several additional challenges with this coding project. All known plant motifs had to be collected, catalogued and stored in an easily accessible data format. There is currently no single complete resource of all plant transcription factor-binding site (TFBS) motifs, so scripts had to be written to collect data from multiple sources and to maintain the data in a format easy to use and update. From a user-experience point of view, it was important to keep the same look and feel of CoGe and maintain the same workflow with which CoGe users are familiar.

Using MotifView, users can enter their own motifs or select from a list of all known motifs grouped into various categories, as will be described in detail. MotifView output allows users to view the instances of each motif on the genomic regions selected, and to check and compare multiple genomic regions, aligned by sequence similarity, in one snapshot in a secure, user-friendly manner with easy to visualize graphics. MotifView lays the foundation for an enhanced system to be used for visualizing and comparing all ENCODE-like data, all features with a version, chromosome, start, stop.

MotifView is growing constantly with new and diverse datasets being added to the back end, including pre-annotated datasets, some of which

will be discussed here in detail. [A “preannotated dataset is a CoGe whole genome annotation, TAIR10 for example, to which is added user-generated features using the same visualization format as used for exons, but using a different color.] The power of MotifView is that its way more than just a motif-finding tool. As any computational tool, collaboration allows for growth. By collaborating with groups such as the Soudhamini group at NCBS, Bangalore, India, MotifView will soon be featuring pre-annotated datasets involving motifs involved in various stress responses found through independent verification.

Future versions of the tool will include options with which users can search an entire genome for a set of motifs and get enrichment statistics for each motif. MotifView, in its nascent stages, has been published in the following:

- Mathews G, Subramaniam S, Freeling M. 2013. Computational identification of conserved root hair elements in maize. Poster #25, 55th Annual Maize Genetics Conference, St Charles, Il.
- Chettoor A, Nelsom W, Subramaniam S, Evans M. 2013. Molecular Genetic Dissection of Auxin in Maize Embryo Sac Development. Poster #205, 55th Annual Maize Genetics Conference, St Charles, Il.
- Subramaniam S, de Leon-Horton, Kieran, Lyons E, Freeling M. 2012. MotifView: A comparative genomics tool for analyzing motifs and their genes, Poster #268, 54th Annual Maize Genetics Conference, Portland, Oregon.

Introduction

It's important to know where transcription factors (TFs) bind in the neighborhood of a gene to begin to understand what affects that gene's regulation. Finding known TF binding sites (TFBSs) can be done using a multitude of computational techniques (BRILLI et al. 2007; Doi et al. 2008; BUSKE et al. 2010; BROWN et al. 2013; GAO et al. 2013).

Several algorithms exist that facilitate a variety of ways to detect TFBSs within DNA sequence (CHENG et al. 2008; CLAEYS et al. 2012; LAJOIE et al. 2012; KIM and Yi 2013; LISERON-MONFILS et al. 2013). There are also several databases that have TFBSs for plants including PLACE (Higo et al. 1998), AthaMap (STEFFENS et al. 2004), AtcisDB/AGRIS (DAVULURI et al. 2003), STIFDB

(SHAMEER et al. 2009). None of these databases support useful search applications, and none aim to display the results of such searches as a part of a comparative sequence alignment experiment.

There are online tools that let you find motifs in any DNA sequence you enter as long as the sequence is short (BAILEY et al. 2009, THIJS et al. 2002, POLULIAKH et al. 2003, CARLSON et al. 2007, OKUMURA et al. 2007). There is nothing out there, October, 2013, supporting searches for TFBSs in entire pre-loaded genomes, and none that support visual comparison of motif patterns among preloaded homologous genes. The output for most online tools for TFBS finding are textual with a tab delimited sheet for each nucleotide sequence. The user must copy the results of one analysis to another web input box to compare results for multiple genes. Also different databases have different types of TFBS motifs in the backend, probably because there is no one repository for all experimentally discovered plant TFBS motifs. MotifView has an exhaustive database of plant TFBS motifs in the backend, together with any or all of CoGe's whole genome sequences (numbering 21,025 genomes on November 1, 2013), and these are also available at the backend of MotifView because it is an application within CoGe.

Results

Output example with many motifs

Figure 2 shows an example results graphic from MotifView. The two panels shown in the figure are for the arabidopsis gene *At4g19230*; this gene encodes a protein involved in ABA catabolism, This arabidopsis gene has an ortholog (syntenic) in the recently sequenced (HAUDRY et al. 2013) crucifer, *Aethionema arabicum*. All motifs within the "Drought" and "Heat" stress categories were selected for visualization in this result graphic. Motifs are drawn using a simple graphic in the shape of a diamond filled with the color corresponding to each motif. Clicking on each motif open a popup box with information on the motif annotation including the motif name, the motif sequence as well as the literature reference associated with the TFBS motif. Homology between the two regions is clearly evident by the relatively large number of HSPs (blastn hits or "High-scoring Segment Pairs") between these two regions, both between coding as well as noncoding space. In fact a majority of the HSPs within noncoding space in this graphic have been confirmed as conserved noncoding sequences and are shown

as precomputed CNSs [from our automated pipeline PL3, (TURCO et al. 2013)] embedded as vertical dark green rectangles on the model line on CoGe custom arabidopsis TAIR10 genome dsid 19870. The annotation label for one such CNS is shown in Figure 2 in a popup window. Although there are a lot of motifs in Figure 2, making a preliminary analysis is easy. For example, it's evident that homology extends to TFBS motifs shared between the two regions since several conserved motifs overlap with CNSs. Differences in motif distribution between the two regions could correspond to either different (species-specific) functions for the two homologs or to different modes of regulation for the two genes, or—in some cases—to noise or differential detectability issues.

Studying Mutations via unexpected patterns of fractionation

Post-polyploidy, originally duplicate genomes fractionate: they often lose one or the other of genes or CNSs, but generally not both. The fractionation mechanism in maize and *Brassica rapa* is deletion and not base substitution (WOODHOUSE et al. 2010; SUBRAMANIAM et al. 2013). When genes are retained as duplicates, fractionation mutants—recognized as losses of one or more CNSs—sometimes show incomplete expression patterns (like “off in pollen”; FREELING et al, 2012) compared to the dominant less-fractionated or unfractionated gene duplicate. Given that CNSs are rich in TFBS motifs (Freeling et al. 2007), it stands to reason that fractionation could impact gene expression by removing one or a combination of TFBS motifs within CNSs surrounding the duplicated gene; this has been shown clearly for G-box fractionations in doublets (duplicates)(SUBRAMANIAM et al. 2013) Using MotifView, the impact of post-polyploidy fractionation can be evidenced through visual comparison between multiple genomic regions. Figure 3 shows four syntenic genomic regions, in order from top to bottom: *Arabidopsis thaliana* (*At*), *Aethionema arabicum* (*Aa*) and the doublet pair of syntelogs in *Brassica rapa* (*Br*). Conserved non-coding sequences identified between *At* and *Aa* are indicated as dark green rectangles. The four genomic, orthologous regions have been interrogated with the ABRE-drought response motif and the results were filtered for overlap with CNSs. Every instance of the ABRE motif is shown as a navy blue color-filled diamond that-- in the *At* panel-- can be clearly seen to overlap two *Aa-At* CNSs. The *Aa* panel shows these motifs in the corresponding orthologous positions in *Aa*.

The bottom *Br* panel shows the motif retained within the expected orthologous positions relative to the gene. However in the case of the other *Br* gene, the homeolog, only one of the ABRE motifs was retained in the window. The other motif is clearly absent, gone by fractionation following the most recent *Brassica* lineage hexaploidy (WANG et al. 2011). The gene missing the ABRE-motif is a potential fractionation mutant, and a candidate for being a drought-stress-nonresponsive mutant.

Tab-delimited output

Results from MotifView can be downloaded in tab-delimited format. Table 1 is a sample output file showing motifs within one of the selected genomic regions. This file can be exported and opened in any spreadsheet software for further analysis, including Open Office and Excel.

Conclusion

CoGe is a comparative genomics platform that is being used by, on average, 4000 scientists a month. CoGe is powered by iPlant, the community-based plant cyberinfrastructure project funded by NSF. MotifView has been fully functional for months, and is now available on the CoGe Home Page menu along with the other applications like GEvo and SynMap. MotifView provides greater value for CoGe researchers interested in cis-regulatory function.

MotifView, in its current avatar, is a complete tool for finding and visualizing plant TFBS motifs in any genomic region, and is uniquely useful for creating complex comparative motif/CNS,HSP graphics. Most importantly, this tool allows bioinformatics specialists who have listed motifs and associated them with genes or regions can visualize this output and thereby proof the code of their pipelines. [Comparative genomics pipelines without mandatory graphical proofing opportunities during development had better have other, clever ways to avoid those bugs that live within the illogical complexities of evolutionary history.] Every list associating a motif with a gene should have a MotifView link.

MotifView is a flagship application. We envision a fleet of similar applications organized around various ENCODE-like features that can be assigned a genome location (version, chromosome, start, stop).

The code behind MotifView can be extended to include any type of annotation. This will let users look for overlaps between TFBS motifs based on sequence and CHIP-Seq sites, methylated Cs, smallRNA targets, Hi-C regions and the like. In addition, the underlying graphics module is currently being expanded to view RNA-Seq data as histograms popping-up nearby genes on the MotifView graphic. Please contact the first author for suggestions for continued enhancements.

Methods

MotifView was built using the API and graphics modules of CoGe (LYONS, 2009) written primarily in PERL and Python. The code for MotifView is available to view in the CoGe repository at <http://coge.iplantcollaborative.org/CoGe/tmp/MotifView/MotifView.pl>. However, since this is an CoGe add-on and not a stand-alone application, there are no instructions for installation beyond comments annotating the code itself. MotifView has been designed to maintain the same look and feel as the rest of the CoGe platform. This is done so that it fits seamlessly into CoGe and can be used by CoGe users with relative ease. A tutorial is available on the wiki supporting CoGe <http://coge.iplantcollaborative.org/wiki/index.php/MotifView>. Users start by choosing two or more genomic regions they want to compare for TFBS motifs (Fig 1A). The CoGe database contains >20 thousand genomes plus versions, each with annotations in the gff format. Information relevant to the regions selected by the user will be automatically filled in to the form as the user types the gene id.

There are four tabs in the web page of MotifView. The form (Figure 1A) has the same options as the one for GEvo (CoGePedia page on GEvo). The second tab ‘Algorithms’ allows users to select blast options to use in the comparison (Figure 1 B). Currently only two options, blastn (local) and Lagan (global) have been enabled. Why? Because it makes sense to visually compare small regions of high homology when motif finding is involved. As will be explained later in this document, blastn HSPs (with a 29.5 bitscore cutoff) are used for detecting exons encoding protein and conserved noncoding sequences (CNSs) between different regions. Since TFBS motifs are super-enriched within CNSs, and CNSs are, by definition, functional, confining analyses to TFBSs within CNSs is often a good idea. Additionally, knowing where CNSs are within the MotifView panel

may be important for fractionation analyses since CNSs are excellent markers for chromosomal positions, and may be used in ways similar to using SNPs.

The third tab ‘Results Parameters’ (Figure 1C) facilitates the same options as with GEvo, with a couple of very important additions. One of the key aspects of MotifView is that it allows the user to graphically render not just TFBS motifs within the selected regions, but also to visually examine these motifs in relation to other annotations such as CNSs, thus facilitating visual pattern recognition. Users can choose which preloaded annotations to be viewed along with the TFBS motifs. Another useful option allows users to view only those motifs that overlap with selected annotations. For example, a user can choose a dataset within CoGe with CNSs and choose to see only TFBS motifs that are localized within CNSs. An example analysis and its output using this scenario will be shown later.

The fourth and final tab presents options unique to MotifView, providing menus of individual (like “G box”) or functional group (like “responds to light”) motifs (Fig 1 D). There are four options for the user to enter/select/choose:

- 1) Manual entry: Clicking the “Search for User-Defined Motifs” header, drops down a text box into which users can manually enter the regular expression corresponding to the TFBS motif. Multiple motifs can be entered in the text box separated by newlines. Users have the option to either define a color of choice in the format (<motif>:<colorname>) or let MotifView choose a randomly assigned color for each motif. Select from the Comprehensive List of motifs: MotifView has in the backend a database of plant TFBS motifs collected from a variety of data sources, with each source accurately identified in the literature and on other lists if present. (MotiveView’s Motif List is available for download in tab-delimited format from the MotifView homepage, and at <http://coge.iplantcollaborative.org/CoGe/tmp/MotifView/Motiflist.txt>. Clicking on the header of Motif List drops down two list boxes, the left one showing every motif within the MotifView database, the right one to be populated with the user-chosen motifs. Users can select each motif, either by double-clicking on the motif name or by single clicking the

- motif and clicking on the “Add” button to add the motif to the user selected motif list.
- 2) **Select Motifs from Stress categories:** Motifs with experimental evidence—cited clearly on our lists-- for involvement in stress responses have been categorized on the basis of the stress response type(s). Clicking the main header drops down multiple headers, each corresponding to a stress type. Each stress type header opens a drop down menu showing all motifs within each stress type in the format: the name of the motif/motif regular expression in IUPAC format/Color specific to each motif. Users can select one, multiple or use the ‘Select All’ button to choose all motifs associated with each stress category. Colors for the motifs within each stress type are pre-assigned because choosing colors manually is time-consuming process especially if you want motifs of the same stress category to be variations of a similar color. An example of this automated color-chooser will be shown later.
 - 3) **Select Motifs from Transcription Factor Family categories;** motifs have been classified on the basis of the family of TFs shown to bind each motif. Motif families follow the IUPAC convention, these being based on conserved protein domains and not on functional network. Similar in organization to the stress categories list, the TF family category header opens to reveal subheadings each representing a TF family. Users can choose from the drop down menu one, multiple or all motifs in each family, and also choose to view motifs belonging to multiple TF families on the same results graphic. As in the case of the stress categories, the color scheme for TF families has been pre-assigned so that TFBS motifs belonging to the same TF family have similar color.

Acknowledgements

Funding provided by a grant from the NSF to MF: IOS-1248106. We thank Eric Lyons for all his support with the underlying code of the CoGE platform.

Tables

Motif Name	Start	Stop	Motif sequence	Lit Ref	Title	Color	Strand
ABRE-like binding	5406	5413	[CGT]ACGTG	Curr. Opin. P	Molecular re	tomato4	+
ABRE-like binding	5665	5672	[CGT]ACGTG	Curr. Opin. P	Molecular re	tomato4	+
ABRE-like binding	5404	5411	[CGT]ACGTG	Curr. Opin. P	Molecular re	tomato4	-
RAV	7095	7107	CAACA.*CAC	Kagaya, Nuc	NA	DeepSkyBlue	+
RAV	6153	6165	CAACA.*CAC	Kagaya, Nuc	NA	DeepSkyBlue	-
EREBP-DERB	6759	6764	[AG]CCGAC	Stockinger, I	NA	cyan	-
EREBP-DERB	6063	6068	[AG]CCGAC	Stockinger, I	NA	cyan	-
EREBP-DERB	3384	3389	[AG]CCGAC	Stockinger, I	NA	cyan	-
EREBP-DERB	1053	1058	[AG]CCGAC	Stockinger, I	NA	cyan	-
EREBP-DERB	51	56	[AG]CCGAC	Stockinger, I	NA	cyan	-
G-box promoter	5406	5411	CACGTG	Proc Natl Aca	Isolation and	VioletRed1	+
G-box promoter	5406	5411	CACGTG	Proc Natl Aca	Isolation and	VioletRed1	-
ATB2/AtbZIP	11062	11067	ACTCAT	Plant Cell Ph	A Novel Subg	DarkMagenta	+
ATB2/AtbZIP	11832	11837	ACTCAT	Plant Cell Ph	A Novel Subg	DarkMagenta	+
ATB2/AtbZIP	10569	10574	ACTCAT	Plant Cell Ph	A Novel Subg	DarkMagenta	-
ATB2/AtbZIP	6670	6675	ACTCAT	Plant Cell Ph	A Novel Subg	DarkMagenta	-
ATB2/AtbZIP	47	52	ACTCAT	Plant Cell Ph	A Novel Subg	DarkMagenta	-
ABA; ABF; bZIP	5406	5413	[TC]ACGTGG	J Biol Chem.	ABFs, a fami	DarkOrchid1	+
ABA; response	5405	5414	[GA][TC]ACG	Mol Gen Gen	Identification	OrangeRed3	+
MYB1 binding	8802	8809	[AC]TCC[AT]A	Proc Natl Aca	Isolation and	DarkSeaGree	+
MYB4 binding	2298	2304	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	5558	5564	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	5689	5695	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	5956	5962	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	7768	7774	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	9692	9698	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	9846	9852	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	9867	9873	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	+
MYB4 binding	11630	11636	A[AC]C[AT]A	Plant Cell. 20	Expression p	LightGreen	-

Table 1: Motif results output list

Figures

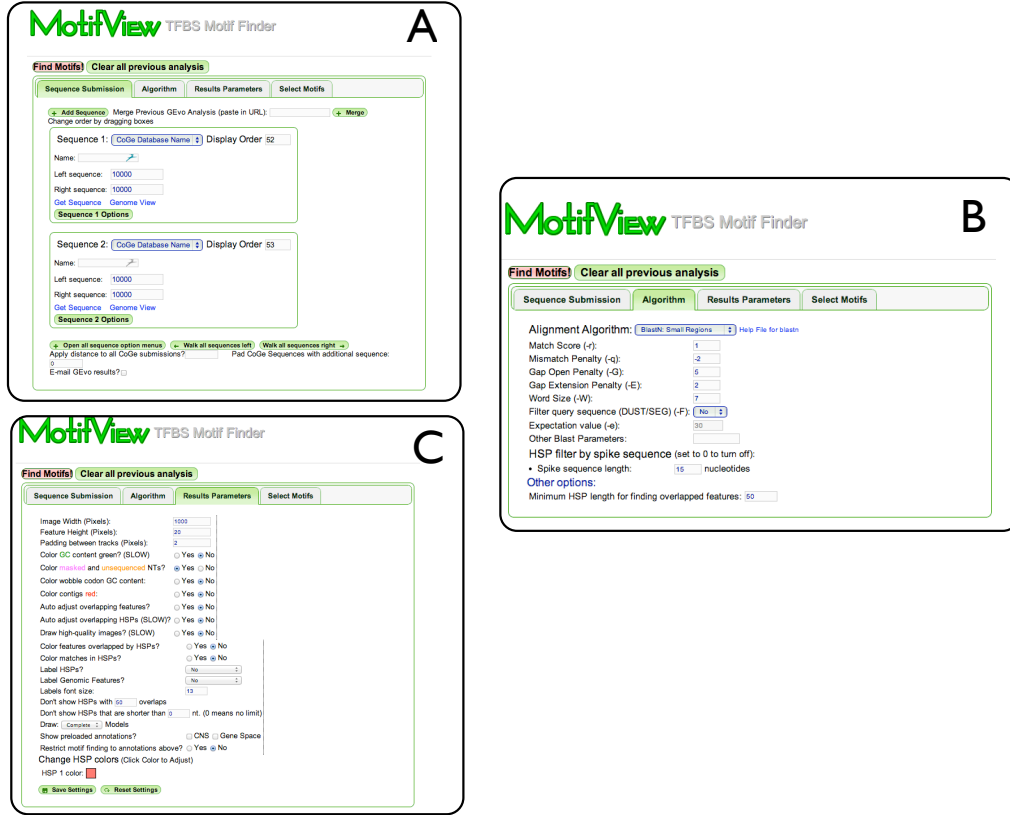


Figure 1 A,B,C. Screenshots of MotifView showing input tabs. A) Select genomic regions. B) Choose algorithm for pairwise comparison of genomic regions. C) Results parameters tab showing options that user can select. Regenerate this page at <http://coge.iplantcollaborative.org/CoGe/MotifView.pl>.

D

User-defined motif(s)

User-selected list of motifs

Categories of TF family

Categories of Stress

Figure 1D. Screenshots of MotifView showing input tabs. D) Motif to choose/select/enter. Describe each one. Regenerate this page at <http://coge.iplantcollaborative.org/CoGe/MotifView.pl>



Figure 2. Drought/Heat Stress motifs,At-Aa.

Figure 2. MotifView output of drought and heat stress motifs within orthologous region between *Arabidopsis thaliana* (At) and *Aethionema arabicum*(Aa). Motifs are drawn as diamond-shaped graphics filled with motif-specific color. Regenerate this MotifView experiment and test-drive our application at <http://coge.iplantcollaborative.org/CoGe/MotifView.pl>



Figure 3. MotifView results panels showing impact of fractionation on upstream ABRE drought response motifs within *Brassica rapa*. Regenerate this experiment and test-drive our application at <http://coge.iplantcollaborative.org/CoGe/MotifView.pl>

Chapter 6: During the comparison of genomes, orphaned conserved noncoding sequences may signify regulatory novelty.

(largely the work of my summer intern Valentine Verzat, visiting from Ecole Normale Supérieure, France. (This Chapter is not written as a draft paper))

Introduction

Conserved Noncoding Sequences (CNSs) and the genes they are associated are subject to fractionation—duplicate DNA loss—following whole genome duplication. In previous chapters we explored the evolutionary fate of homeologous CNSs following the alpha Whole Genome Duplication (WGD) event in *Arabidopsis thaliana* (*At*), as well as orthologous CNSs between *Aethionema arabicum* (*Aa*) and *At* as they are fractionated after the pre-Brassica paleohexaploidy. We used “Find CNS” blastn settings (within our CNS Discovery Pipeline PL3. Turco et al. 2013) to determine the detectability of each CNS within all three expected orthologous regions in *Br*.

We assign CNSs to genes based on the simple rule of synteny and proximity; the nearest syntenic gene gets the CNS. This mode of assignment makes a lot of sense as we find particular families of genes to be CNS rich, e.g.: genes encoding transcription factors. As newer genomes get sequenced, we have the ability-- using our Map-Over pipeline (Chapter 3)-- to be able to detect CNSs within these new genomes. The success of this Map-Over strategy depends on the existence of one or more orthologous copies of the ancestral gene to which the CNS has been assigned.

We noticed-- during the course of our mapping *At-At* CNSs onto *Br* (Chapter 2)-- that there are several cases where CNSs that were thought to act in *cis* to particular genes still exist even though the “mother” gene was deleted. We call these exceptional sequences “orphans”.

There are several possible explanations for a CNS becoming orphaned, and some are trivial. 1) The CNSs was assigned to the wrong gene in the first place (trivial). 2) The CNSs was co-opted to perform a new function following the loss of its original *cis*-gene (not trivial). 3) The CNSs could be carrying motifs that make them part of a larger network

of regulation modulating the function of multiple downstream genes (not trivial); while such cis chromosomal domains have not been reported in plants, an early mammalian CNS paper described just such a situation (LOOTS et al., 2004). 4) The CNS could be an unannotated, independent gene that avoided annotation (not trivial).

The general objective of this experimental plan was to find all CNSs orphaned or potentially orphaned in *Brassica rapa* v. Chiifu. We intended to validate orphan CNSs detected by the Map Over pipeline and to analyze invalidated orphans to improve the computational pipeline for orphan detection as well as to analyze validated orphans to understand their evolutionary history and potential new function, if any. Each orphan CNS has the potential to be an evolutionary placeholder for post-polyploidy regulatory novelty.

Preliminary Results

Validation of orphan CNSs with GEvo

From the Mapped-Over *Br* CNSs analyzed, we identified 1322 orphans (21%) (Fig. 1). Each of these CNSs was then validated using the GEvo program within the CoGe comparative genomics platform. GEvo panels showing orphan CNSs were examined for two possible errors (Fig. 2) in the orphan detection script:

1. In 35% of invalidated cases, an orthologous copy of the ancestral gene to which the CNS is associated, was found to be actually present in a syntenic position. This type of error could be because of heavy fractionation on the gene which throws off the QUOTA-ALIGN algorithm that finds syntenic pairs.
2. Erroneous CNS calls. 65% of invalidated CNSs were invalidated due to lack of sufficient evidence to be called a valid CNS. This could have been because of problems with the alignment algorithm or because of the poor quality of the CNS e.g. Simple sequence. Of the 1322 putative orphans, 1090 were invalidated by this method.

Furthermore, all orphan CNSs overlapping with small open reading frames (sORFs) identified by Hanada *et al.* 2007 were invalidated. Of the 7159 sORFs identified in the *At* genome, none overlapped with our orphan dataset.

We ended up with a list of 232 potentially orphan CNSs: each exists although its putative mother gene has been deleted. It is important to note that the link between CNS and gene assumes that genes do not have overlapping components. The cis-acting unit= gene is assumed to be unique. This assumption is just a hypothesis. Our CNS-discovery script *assumes* that the closest gene is the one and only one on which the CNS works in cis (Freeling and Subramaniam, 2009). If a nearby gene to the CNS in *A. arabicum* is also present in *B. rapa*'s subgenome with this CNS, we deduced that the orphan was a false positive. Removing such false positives, we narrowed the list to 106 improved candidate orphans.

Detectability of orphan CNSs in *Br* subgenomes

Previous studies have shown biased fractionation between subgenomes I, II and III of *Br*. (WANG *et al.* 2011; CHENG *et al.*, 2012). We noticed that the distribution of the 106 candidate orphan CNS

detectability in *Br* was not equal between the three subgenomes (Table 1). Detectability follows the conservation of the subgenomes: subgenome III has more orphans compared to II and I.

Frequency of TFBS motifs within orphan CNSs

Plant CNSs are enriched in transcription factor or other cis-acting binding sites (Freeling and Subramaniam, 2009). We compared the distribution and frequency of motifs within orphan and non-orphan CNSs. We used a list of known plant transcription factor binding motifs (List is available for download at <http://genomevolution.org/CoGe/MotifView.pl>). These motifs are comprised of simple motifs (with redundancy in multiple positions) and complex motifs (no redundancy, e.g.: CACGTG). A simple motif was found at least once in every orphan CNSs while complex motifs were much less common as compared to non-orphan CNSs (Fig. 3).

We observed that the correlation between orphan CNS length and number of motifs decreases with the increase in motif complexity (Figure A-C). For non-orphan CNSs this correlation decreases from 0.57 (strong) to 0.21 (weak) and for orphan CNSs from 0.64 (strong) to 0.41 (weak). This observation seems to validate the relevance of our motif sorting: the more complex the motif the less their presence is correlated with the size of sequence and thus the motif probably occurs at random. We also noted that number of orphan CNSs have more motifs per CNS but this difference was not significant (Figure 5).

Discussion

Examination of orphans using GEvo graphics has allowed us to study regions of the genome and relationships between CNSs and genes. Using the validation step, false positives were eliminated to end with a list of 106 possible orphans. It is possible that orphan CNSs have evolved differently from other CNSs. Orphan CNSs are less numerous per gene than non-orphan CNSs. Moreover we have found that the distribution of CNSs in the three subgenomes of *B. rapa* follows differences in gene retention in these subgenomes after diploidization.

Orphan CNSs represent a part of a cluster which has evolved separately. This is why we hypothesize that CNSs could act on several

genes or even on a genomic region (like animals CNS). We have found a list of 25 motifs that appear in orphan and non orphan CNSs. They all appear more frequently in orphans than in non orphans. It seems to indicate that orphan CNSs have an important role in cis-regulation and are really interesting sequences with a potential of regulatory novelty. However we have found a list of 53 motifs significantly enriched in non-orphan CNSs compared to orphans (data not shown).

An experiment for the future. It could be interesting to use FPKM data (RNAseq data) to determine whether a new gene is being regulated by the orphan CNS. I have compared subgenomes to determine the gene associated with the CNS, but it could be more relevant to know the expression of the putative new gene. Furthermore as far as motifs are concerned, we obtained a list of 25 relevant motifs but without real information about their roles and functions. It could be interesting to know what exactly these motifs do in other plants. Orphans could have global roles in modulating chromatin structure as insulators (blocking distant enhancers) or as proverbial matrix attachments sites.

Methods

We started with two spreadsheets. Supplemental Table 1 with *Aa-At* CNSs with syntenic genes identified in *Br* subgenome I and II (orphans in subgenome III) and Supplemental Table 2 of *Aa-At* CNSs without syntenic genes identified in *Br* subgenomes (orphans in subgenomes I, II and III).

For all analyzes of data the programming language used is Perl. The statistic comparisons have been done with the function `wilcoxon()` (WILCOXON, 1945) of the module "stats" of Scipy (OLIPHANT, 2007). The Wilcoxon test is a non-parametric statistical hypothesis test used when comparing two related samples. It can be used as an alternative to the Student's t-test when the population cannot be assumed to be normally distributed. Coefficient of correlation has been calculated with the function `corrcoeff()` of Numpy (OLIPHANT, 2007). Analysis of orphan CNSs detected by the script was done using the GEvo module within the CoGe comparative genomics platform. Figure 6 shows the strategy for mapping over orphan CNSs over to *Br*.

Tables

Subgenome	detectability in orphan CNSs (occurrence)	detectability in orphan CNSs (%)	retention of <i>A. thaliana's</i> orthologous genes (%)
I	36	34	36
II	32	30	46
III	54	51	70

Table 1. Distribution of detectability of orphan CNSs in *B. rapa's* subgenomes.

Figures

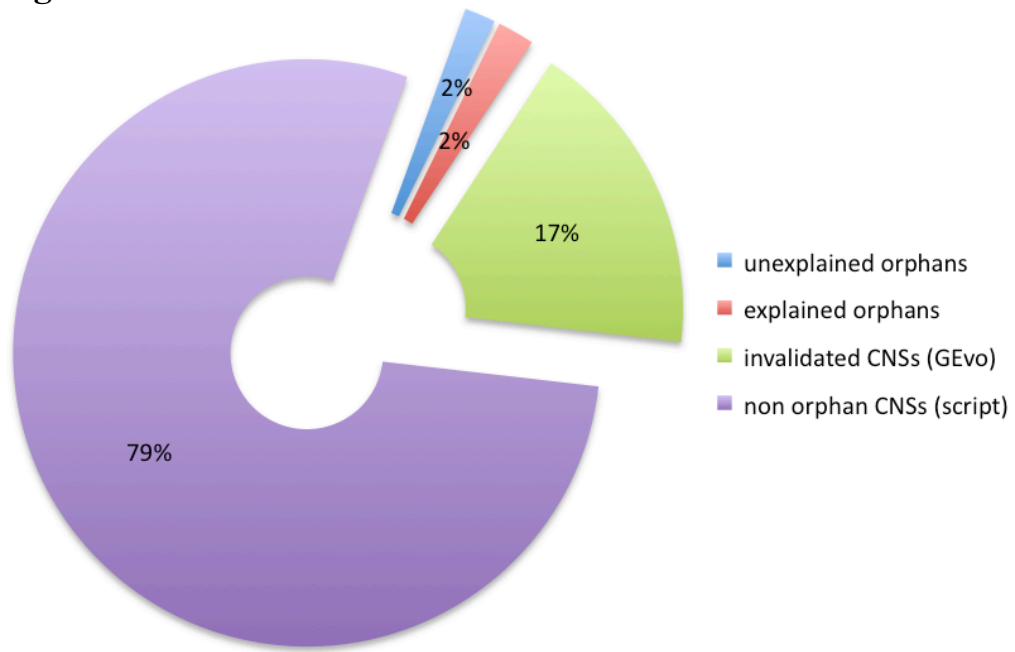


Figure 1. Quantitative results of orphan analysis (n = 6216)

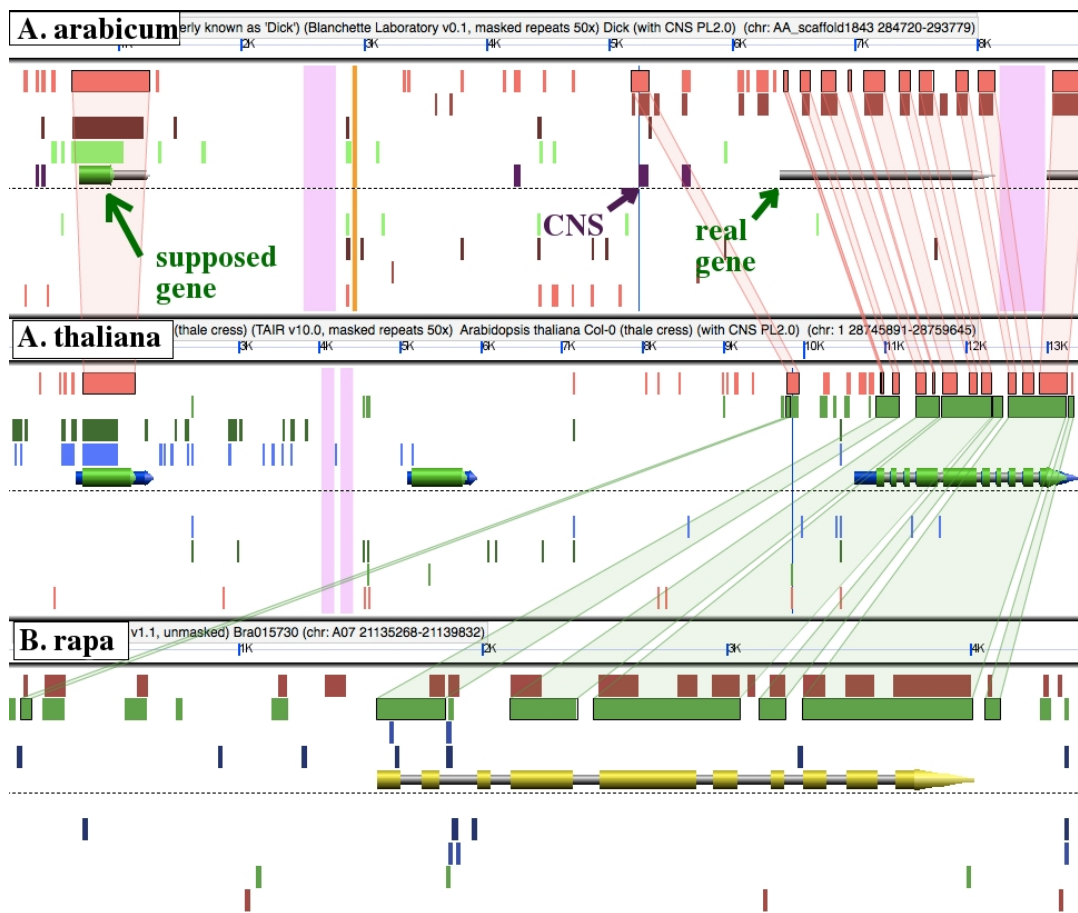


Figure 2. Example orphan CNSs. First panel represents *Aa* genome, second line *At* genome and third line *Br* subgenome I. The CNS was simply sorted to the wrong gene, and was not orphaned at all.

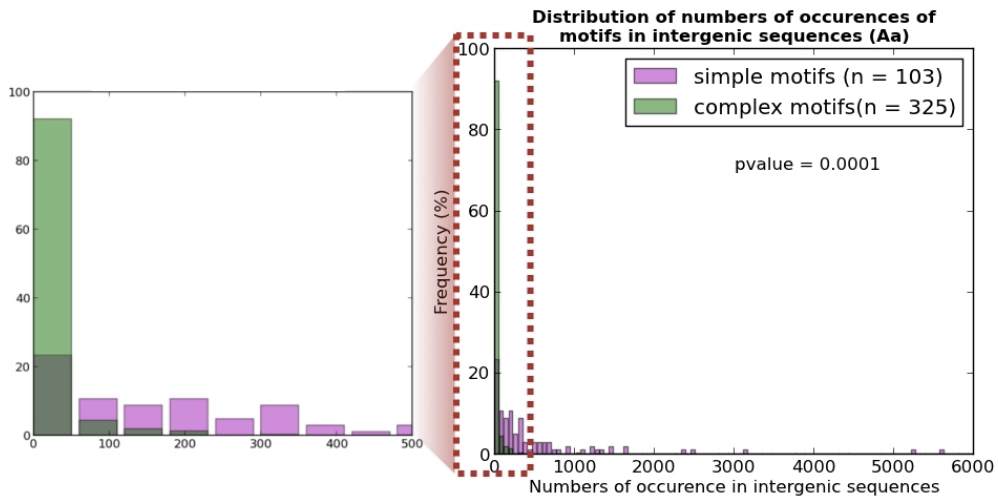


Figure 3: Histogram of occurrence in CNSs of complex and simple motifs.

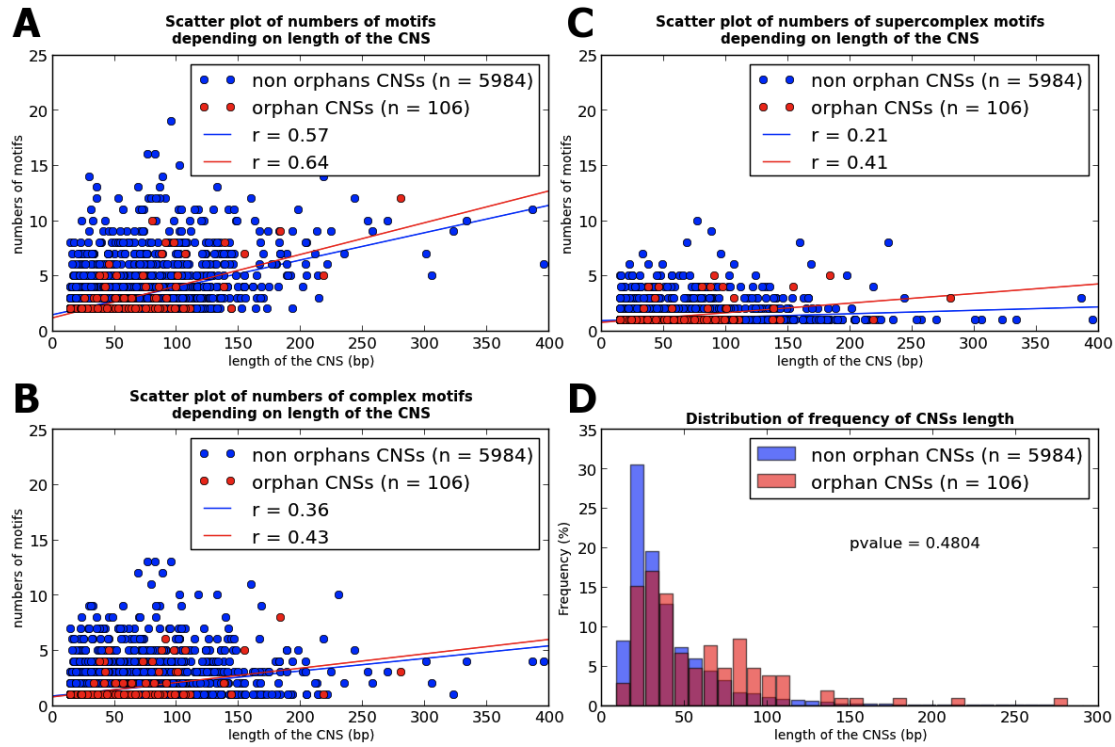


Figure 4. (A-C) Scatter plot of numbers of motifs per *Aa-At* CNS (D) Histogram of length of CNSs compared between orphan and non-orphan CNSs.

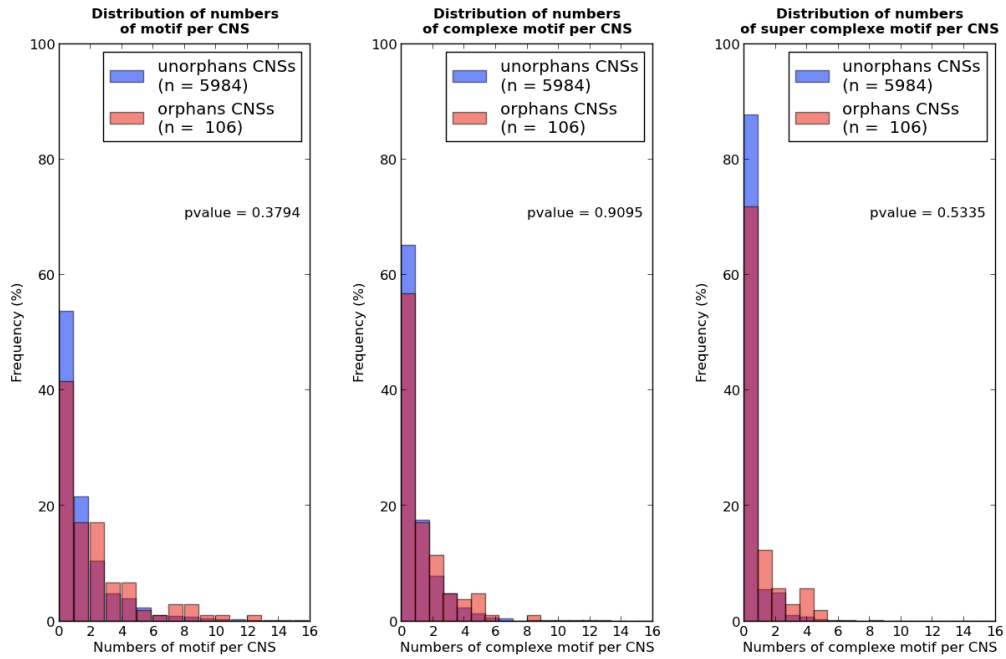


Figure 5. Histogram of frequency of motifs in orphan and non-orphan Aa-At CNSs.

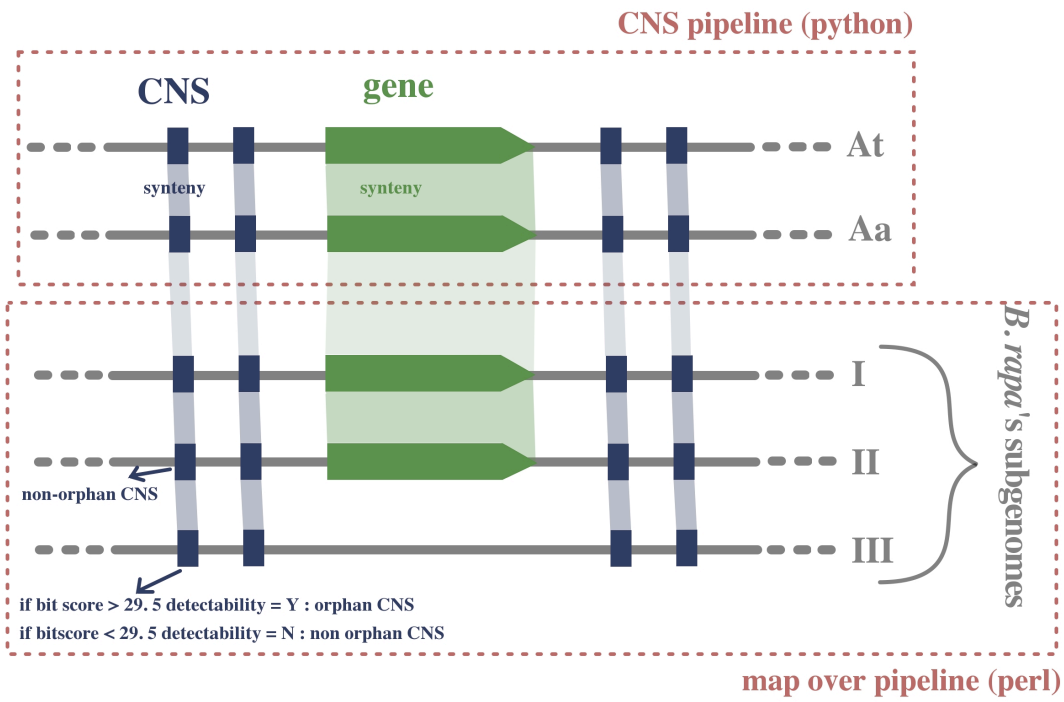


Figure 6. Method of detection of orphan CNSs in *Br*

Supplemental tables

Supplemental data is available for download at
http://figshare.com/articles/During_the_comparison_of_genomes_orphaned_conserved_noncoding_sequences_may_signify_regulatory_novelty/857597

Addendum I: Annotated bibliography of latest CNS research

Preface

The following is a list of research publications on the topic of conserved noncoding published since our review on CNSs (SUBRAMANIAM and FREELING, 2009). Papers of particular importance are annotated with “***”.

Bibliography

- Amiel, Jeanne, Sabina Benko, Christopher T Gordon, and Stanislas Lyonnet. 2010. Disruption of Long-Distance Highly Conserved Noncoding Elements in Neurocristopathies. *Annals of the New York Academy of Sciences* 1214: 34–46.
- Babarinde, Isaac Adeyemi, and Naruya Saitou. 2013. Heterogeneous Tempo and Mode of Conserved Noncoding Sequence Evolution among Four Mammalian Orders. *Genome Biology and Evolution* 5: doi: 10.1093/gbe/evt177 .
- Babbitt, Courtney C, Olivier Fedrigo, Adam D Pfefferle, et al. 2010. Both Noncoding and Protein-Coding RNAs Contribute to Gene Expression Evolution in the Primate Brain. *Genome Biology and Evolution* 2: 67–79.
- Cantone, Carmela, Maria Sica, Luciano Gaudio, and Serena Aceto 2009 The OrcPI Locus: Genomic Organization, Expression Pattern, and Noncoding Regions Variability in *Orchis Italica* (Orchidaceae) and Related Species. *Gene* 434(1-2): 9–15.
- Godoy M, Franco-Zorrilla JM, Perez-Perez J, Oliveros JC, Lorenzo O, Solano R. 2011. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *Plant J* **66**(4): 700-711.
- Halligan, Daniel L, Fiona Oliver, Jack Guthrie, et al. 2011 Positive and Negative Selection in Murine Ultraconserved Noncoding Elements. *Molecular Biology and Evolution* 28(9): 2651–2660. ***
- Harmston N, Lenhard B. 2013. Chromatin and epigenetic features of long-range gene regulation. *Nucleic acids research* **41**(15): 7185-7199.
- Hufton, Andrew L, Susanne Mathia, Helene Braun, et al. 2009 Deeply Conserved Chordate Noncoding Sequences Preserve

- Genome Synteny but Do Not Drive Gene Duplicate Retention. *Genome Research* 19(11): 2036–2051. ***
- Janes, D E, C Chapus, Y Gondo, et al. 2011 Reptiles and Mammals Have Differentially Retained Long Conserved Noncoding Sequences from the Amniote Ancestor. *Genome Biology and Evolution* 3: 102–113.
 - Jeong, Yongsu, and Sangtaek Oh 2010 Identification of an Evolutionarily Conserved, Functional Noncoding Element Regulated by Six1 Homeoprotein. *Genes & Genetic Systems* 85(3): 233–240.
 - Kang, Seong Wook, Sang Hyun Kim, Naeun Lee, et al. 2012 1,25-Dihydroxyvitamin D3 Promotes FOXP3 Expression via Binding to Vitamin D Response Elements in Its Conserved Noncoding Sequence Region. *Journal of Immunology* (Baltimore, Md.: 1950) 188(11): 5276–5282.
 - Kritsas K, Wuest SE, Hupalo D, Kern AD, Wicker T, Grossniklaus U. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Res* 22(12): 2455-2466.
 - Lebenthal, Ilana, and Ron Unger 2010 Computational Evidence for Functionality of Noncoding Mouse Transcripts. *Genomics* 96(1): 10–16.
 - Licastro D, Gennarino VA, Petrera F, Sanges R, Banfi S, Stupka E. 2010. Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics* 11:151 doi:10.1186/1471-2164-11-151.
 - Liska, Frantisek, Pavel Snajdr, Lucie Sedová, et al. 2009 Deletion of a Conserved Noncoding Sequence in Plzf Intron Leads to Plzf down-Regulation in Limb Bud and Polydactyly in the Rat. *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 238(3): 673–684.
 - Lomelin, David, Eric Jorgenson, and Neil Risch 2010 Human Genetic Variation Recognizes Functional Elements in Noncoding Sequence. *Genome Research* 20(3): 311–319.
 - Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. *Science* 333(6045): 1019-1024.

- Matsunami, Masatoshi, and Naruya Saitou 2013 Vertebrate Paralogous Conserved Noncoding Sequences May Be Related to Gene Expressions in Brain. *Genome Biology and Evolution* 5(1): 140–150.
- Moghadam, Hooman K, Moira M Ferguson, and Roy G Danzmann 2009 Comparative Genomics and Evolution of Conserved Noncoding Elements (CNE) in Rainbow Trout. *BMC Genomics* 10: 278, doi:10.1186/1471-2164-10-278.
- Parker, Stephen C J, Loren Hansen, Hatice Ozel Abaan, Thomas D Tullius, and Elliott H Margulies 2009 Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome. *Science (New York, N.Y.)* 324(5925): 389–392.
- Picot E, Krusche P, Tiskin A, Carré I, Ott S. 2010. Evolutionary analysis of regulatory sequences (EARS) in plants. *Plant J* 64(1): 165-176.
- Plunkett, Jevon, Scott Doniger, Thomas Morgan, et al. 2010 Primate-Specific Evolution of Noncoding Element Insertion into PLA2G4C and Human Preterm Birth. *BMC Medical Genomics* 3: 62, doi:10.1186/1755-8794-3-62.
- Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu CR, Korkin D. 2012. Long identical multispecies elements in plant and animal genomes. *Proc Natl Acad Sci U S A* 109(19): E1183-1191.
- Salmon, Armel, Joshua A Udall, Jeffrey A Jeddloh, and Jonathan Wendel 2012 Targeted Capture of Homoeologous Coding and Noncoding Sequence in Polyploid Cotton. *G3 (Bethesda, Md.)* 2(8): 921–930. ***
- Takahashi, Mahoko, and Naruya Saitou 2012 Identification and Characterization of Lineage-Specific Highly Conserved Noncoding Sequences in Mammalian Genomes. *Genome Biology and Evolution* 4(5): 641–657. ***
- Vavouri, Tanya, and Ben Lehner 2009 Conserved Noncoding Elements and the Evolution of Animal Body Plans. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 31(7): 727–735. ***
- Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present

in the jawed vertebrate ancestor. *Mol Biol Evol* **26**(3): 487-490.

- Wang, Xiaohu, Yibing Zhang, Xuexian O Yang, et al. 2012 Transcription of Il17 and Il17f Is Controlled by Conserved Noncoding Sequence 2. *Immunity* 36(1): 23–31.
- Yao X, Wang H, Li H, Yuan Z, Li F, Yang L, Huang H. 2009. Two types of cis-acting elements control the abaxial epidermis-specific transcription of the MIR165a and MIR166a genes. *FEBS letters* **583**(22): 3711-3717.
- Zhen, Ying, and Peter Andolfatto 2012 Methods to Detect Selection on Noncoding DNA. *Methods in Molecular Biology* (Clifton, N.J.) 856: 141–159. ***

Addendum II: The comparison of genome-wide, orthologous conserved noncoding sequences (CNSs) in arabidopsis obtained using two very different methods: pairwise blast versus multiple global alignments.

To: Mathieu Blanchette
blanchem@cs.mcgill.ca
<http://www.mcb.mcgill.ca/~blanchem>
Trottier 3107, McGill University, Canada

Re: Our potential collaboration on the comparison of the VEGI Project's arabidopsis conserved noncoding sequences (CNSs) obtained by multiple orthologous global alignments among several crucifer genomes and the Freeling Lab's arabidopsis CNS list obtained through employing a pipeline using pairwise blasts between orthologous arabidopsis and the genome of crucifer *Aethionema arabicum*. Our object is to develop a consolidated arabidopsis CNS list that is superior to either individual list.

This is the Freeling lab's preliminary report.

Sabarinath Subramaniam and Michael Freeling, Freeling lab,
November 25, 2013, UCB, CA, USA

Introduction

In 2013, an international consortium, lead by M. Blanchette at McGill University, released the genome sequences of three new species in the crucifers (family Brassicaceae) and, using arabidopsis as a particularly important comparator, plus these three new genomes and five previously released crucifer sequences, published an atlas of over 90,000 crucifer conserved noncoding sequences ([HAUDRY et al. 2013](#)).

For the purposes of comparison, the arabidopsis edition of these globally, multiply aligned CNSs are used here, nicknamed "VEGI CNSs" after the McGill University project of the same name (<http://biology.mcgill.ca/vegi/>). An accession of *Aethionema arabicum* was sequenced and annotated as a part of the VEGI project, and this genome is diverged from the arabidopsis—with a modal Ks of 0.65--to an extent that falls within the window that is useful for pairwise

arabidopsis- *Aethionema* (*At-Aa*) CNS discovery using blastn (ALTSCHUL et al. 1990) parameters set just above noise (LYONS and FREELING 2008) as implemented in the CNS Discovery Pipeline 3.0 (TURCO et al. 2013). The arabidopsis edition of these “local alignment” CNSs are called MF-Lab CNSs after their lab of origin. Since the methodologies are so different, it seemed useful to compare these two CNS collections using their arabidopsis editions.

We had some expectations.

1. Since VEGI CNSs collect arabidopsis sequences conserved with not just *Aethionema*, but with any of eight additional crucifer species, we expected that Freeling Lab CNSs would not include all of the VEGI CNSs. That is, if *Aethionema* has lost arabidopsis orthologs with CNSs present in another crucifer then the MF-lab CNS list should be deficient. Figure 1 is the crucifer phylogenetic tree as taken from Haudry and coworkers (2013).
2. As long as the window of examination is large enough to include all CNSs, blastn finds one or multiple local alignments with equal sensitivity no matter how far away from the gene a CNS might be or how divergent the sequence between CNS patches. Global alignments, on the other hand, while very sensitive close to anchor exons, or anchor CNSs, are less tolerant of sequence divergences or reiterations as they accumulate as distances from genes increases. So, MF-Lab CNSs are expected to be less accurate close to genes and more accurate further away, as compared to VEGI CNSs.

With the help of visualization software, CNSs may be manually seen and proofed. Before we compared VEGI and MF-Lab CNSs, we developed a customized CoGe “proofing” genome (dsgid19870) of arabidopsis TAIR 10 with the 89927 VEGI and 65430 MF-lab CNSs (Supplemental Table 1; Master VEGI-MF-lab CNS list) --color coded purple and green, respectively-- decorating the model lines. Figure 2 is a typical stretch of our proofing arabidopsis genome (upper panel) as visualized in GEvo, the sequence comparison viewer of the CoGe comparative genomics toolbox. Notice that VEGI and MF-CNSs obviously overlap, at least in this typical region. Figure 2 also includes a re-creation of the blastn results between *At* and *Aa* as a sanity check. Every MF-lab CNS (Supplemental Table 1; Master VEGI-MF-lab

CNS list) comes with a link to an on-the-fly CoGe experiment as exemplified in Figure 2. Thus any conclusion about any CNS can be proofed by any researcher at the CoGe toolbox website without need for the command line.

VEGI and MF-lab CNSs overlap.

As expected, there are more VEGI than MF-lab CNSs; specifically, about 37% more. Figure 3 are pie charts showing the nature of the overlap, where “no” is the single most important metric, where there is no overlap.

When all of the VEGI CNSs were arranged in order down the chromosome-- using chromosome, start, descending-- it was clear that long (>19 CNSs) linked stretches of “No Overlap” were frequent. We manually removed all of these “no” regions from *At* chromosome 1, extrapolated the deleted genome over its entirety, and redrew the VEGI to MF-Lab pie chart; “no overlap” dropped to 50% (data not shown). This was generally expected because some genes in *At*, with their CNSs don't exist in *Aa*, and, for that reason, cannot find an overlap with a MF-lab CNS. In gross terms, CNSs between VEGI and MF-Lab overlap about 50% of the time.

Blastn CNSs come with a measure of significance, the bitscore(ALTSCHUL et al. 1990). We used MF-CNSs sorted into bitscore bins as queries to the VEGI CNS list. The “just above noise” cutoff for CNS discovery is bitscore 29.5, and this equals the e value of a 15/15 exact match. The bitscore is independent of window size, so is a more useful measure of CNS significance than is an e-value, for example. Figure 4 shows that the more significant (higher) the bitscore, the more likely that this CNS will be shared by both VEGI and MF-lab lists. It is clear that a bitscore 29.5 MF-Lab CNS—an *At-Aa* syntenous hit as short as 15 bp-- has a low chance (28%) to be recognized by VEGI methods.

Blastn should exhibit no sensitivity bias among CNSs because of location within a gene. We categorized MF-Lab CNSs to one of five gene regions (Fig. 5). MF-Lab CNSs in the intron and, to a lesser extent, 3' proximal region (<500 bps past the last exon) are preferentially not present in the VEGI dataset. Therefore, in part,

sensitivity bias in the VEGI methods explains those non-overlapped CNSs.

As explained in the Introduction, we expected a local alignment algorithm, like blast, to be superior to a global algorithm when comparing sequences far away from homologous anchors, like exons or other CNSs. Therefore, we expected that MF-Lab CNSs increasingly far away from their genes should be increasingly not shared by the VEGI dataset. The data of Figure 6 does show this for 5' CNSs greater than 750 bp 5' from the closest exon, but the trend is not pronounced until distances become greater than 6kb. Additionally, CNSs 5' and within 1kb of exon show a distinct tendency to not overlap as well as the more distant ones. We have not explored the reason for this.

Conclusion

The MF-lab CNSs most likely to overlap VEGI CNSs are those with higher bitscores, those located 0.75 to 6.0 kb from the nearest exon, and those not in the intron or 3' proximal region. VEGI arabidopsis CNSs not present in *Aa* because there is no orthologous gene in *Aa* are naturally unique to the VEGI dataset. We conclude that a combined CNS list—the “Combined VEGI plus MF-Lab” is an improvement over either CNS dataset alone. This Combined Arabidopsis CNS List is Supplemental Table 1, where—in cases of overlap-- the largest possible length, the combined length, is used to delineate the start and stop, and the formal CNS designation becomes Combined Arabidopsis (Tair10) CNS chr. start, stop.

Methods

MF-Lab CNS methods begin with the two lists that are output from the CNS Discovery Pipeline PL 3.0 (TURCO et al. 2013) run for arabidopsis (TAIR10 , CoGe dsid 19870) and *Aethionema* (v0.2, CoGe dsid 80498) available in CoGe. The exact position of a CNS in relation to individual exons has become particularly important for some projects. PL3.0 pipeline was coded by proofing against a manually annotated gold standard, and that comparison was arabidopsis-arabidopsis: comparing the two alpha subgenomes (Thomas et al. 2007). *At-At* CNSs tend to be “spaced out” and not too compact because of post-tetraploidy fractionation. In contrast, the orthologous pairs of *At* and *Aa* are jammed together. To obtain the

intragenic CNS accuracy described here, PL3 output was “tweaked” with an additional pipeline called Genespace Tweaker, as described in Chapter 3 and available for download and deployment at [http://figshare.com/articles/ Computational pipeline to track the evolutionary fates of arabidopsis conserved noncoding sequences through a paleohexaploidy and into the genomes of its close relative Brassica rapa PL3 Genespace Tweaker Mapping Over Supplemental Data/856778](http://figshare.com/articles/Computational_pipeline_to_track_the_evolutionary_fates_of_arabidopsis_conserved_noncoding_sequences_through_a_paleohexaploidy_and_into_the_genomes_of_its_close_relative_Brassica_rapa_PL3_Genespace_Tweaker_Mapping_Over_Supplemental_Data/856778).

Figures

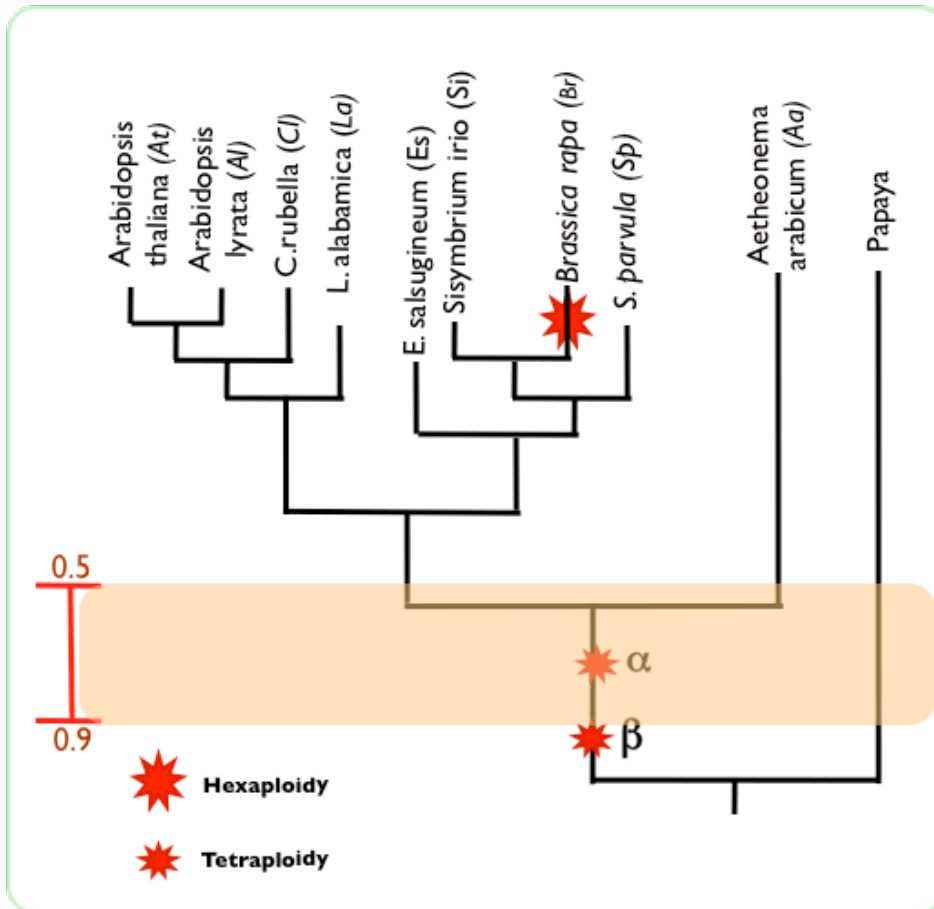
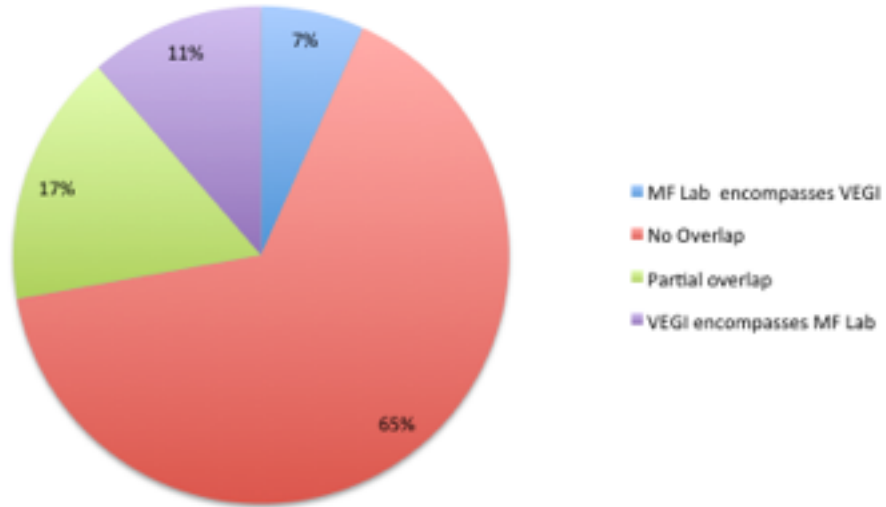


Figure 1. The phylogenetic tree of sequenced crucifers used to obtain the VEGI project CNS (HAUDRY et al, 2013), including *Arabidopsis* and *Aethionema*, used to obtain the MF-lab CNSs. *At* and *Aa*, and the alpha subgenomes themselves, are the only pairs that are within the K_s window of 0.5-0.9, the window of useful CNS discovery using MF-lab methods. VEGI methods utilize all of these genomes.



Figure 2. A GEvo screenshot of an stretch of arabidopsis chromosome 1 (around *Atlg30360*; upper panel) aligned with an orthologous stretch of chromosome of *Aethionema arabicum*, with our proofing *At* genome in the upper panel , the *Aa* genome in the lower panel , and with the orange rectangles being *At-Aa* blastn “hits”, see text for definitions of the colored rectangles on the proofing *At* genome. Regenerate this experiment: <http://genomeevolution.org/r/a24t> .

Arabidopsis CNSs from VEGI project compared to CNSs from MF lab, of 89927



Arabidopsis CNSs from MF Lab compared to VEGI project CNSs, of 65430

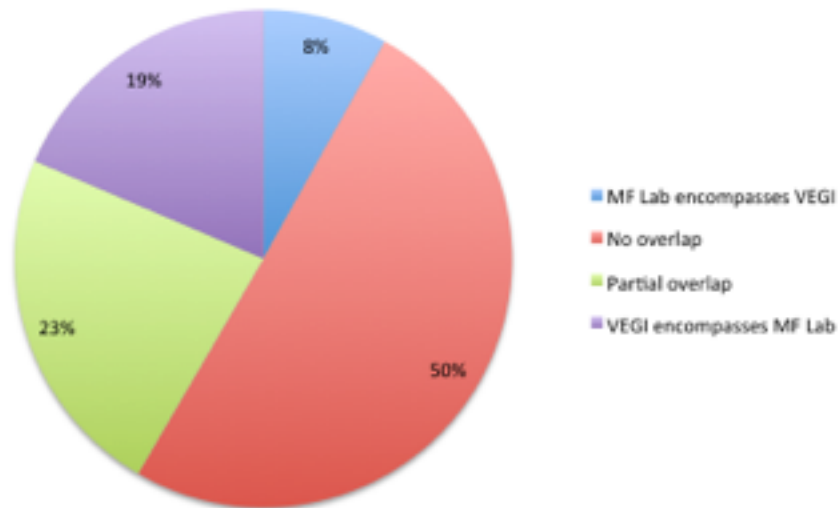
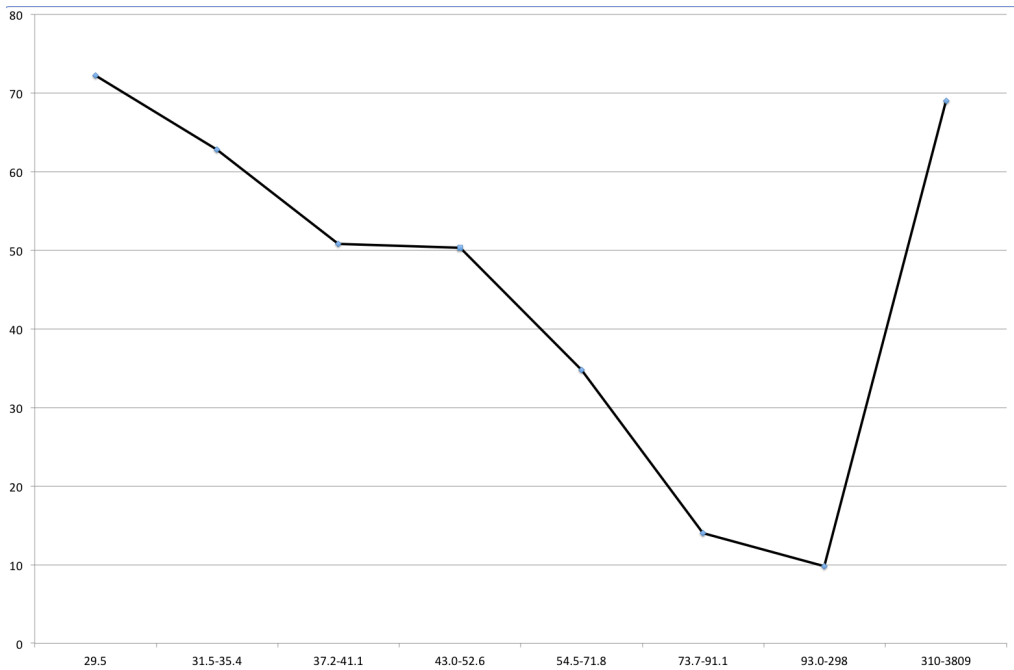


Figure 3. Reciprocal overlap experiments comparing the VEGI and MF-Lab CNSs from Supplemental Information 1.



X axis= bins of CNSs based on bitscore
 Y axis= % no-overlap between MF-Lab and VEGI CNSs

Figure 4. Except for those few MF-Lab CNSs with exceptionally high bitscores (probably artifacts), the higher the bitscore, the more likely that the CNS will also be in the VEGI dataset. CNSs with bitscores between 54.5 and 300 have a 87% chance of being in both datasets.

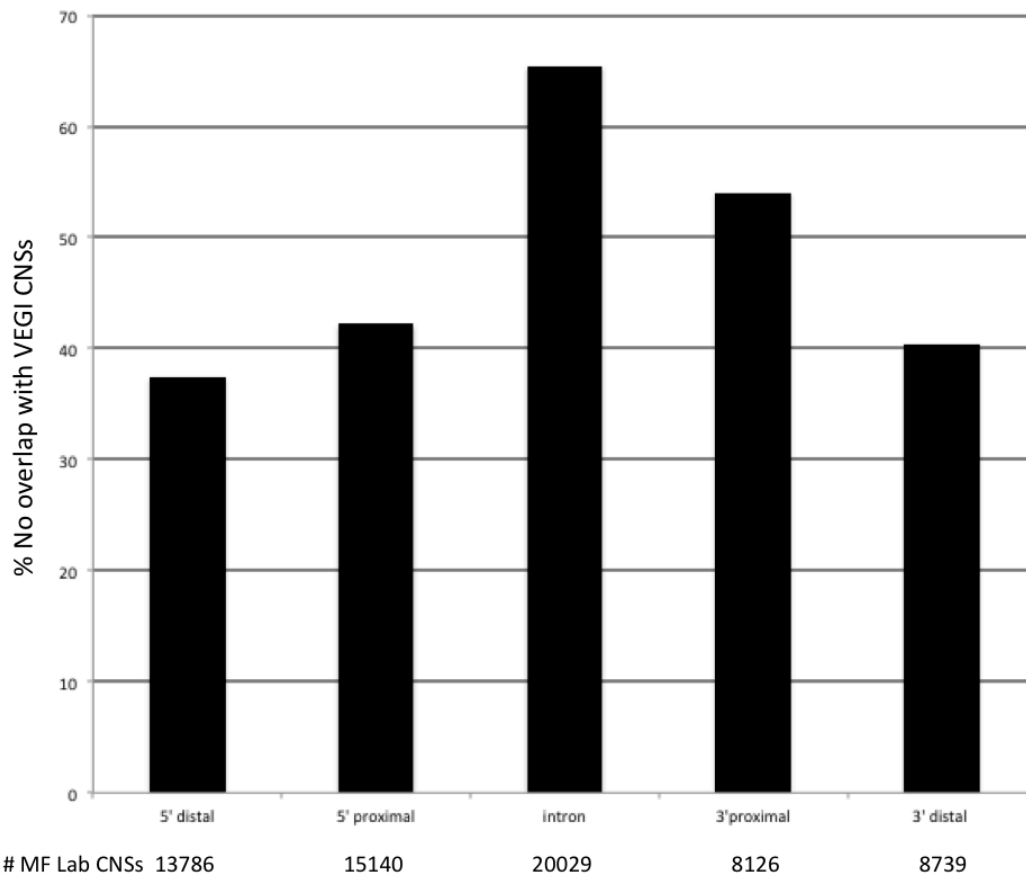


Figure 5. MF-lab CNSs in some gene regions are less likely to be shared by VEGI than others. The Y-axis is % no overlap. The bins on the X axis are CNSs sorted to 5' distal (> 500 bp upstream of exon 1), 5' proximal (\leq 500bps 5' of exon 1), intron, 3' proximal (\leq 500bps downstream of the last exon), and 3' distal >500bp downstream of the last exon.

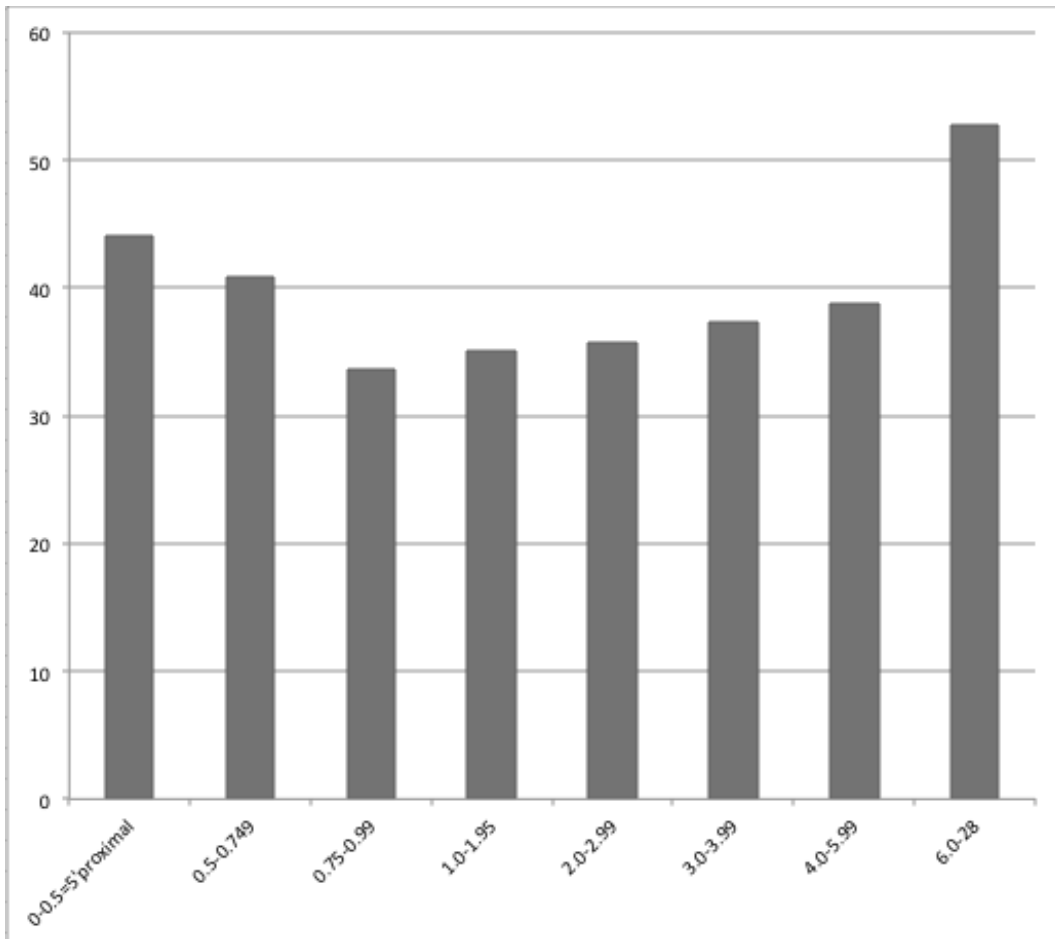


Figure 6. The relationships between the distance, in bps, between a 5' MF-lab CNS and the first exon and its likelihood to also be on the VEGI CNS list. X axis are bins of CNS by 5' distance from exon 1 and the Y axis is % no-overlap with any VEGI CNS.

Supplemental Data

Supplemental data is available for download at

http://figshare.com/articles/Addendum_II_Supplemental_Table_1/868841

Supplemental Table 1: Combined Master CNS List combined CNS datasets of VEGI and MF-blastn CNSs.

Bibliography

- AHITUV N., ZHU Y., VISEL A., HOLT A., AFZAL V., PENNACCHIO L. A., RUBIN E. M., 2007 Deletion of ultraconserved elements yields viable mice. *PLoS Biology*. **5**: e234.
- ALBERTINI A. M., HOFER M., CALOS M. P., MILLER J. H., 1982 On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell*. **29**: 319-328.
- ALTSCHUL S. F., GISH W., MILLER W., MYERS E. W., LIPMAN D. J., 1990 Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**: 403–410.
- BAILEY T. L., BODÉN M., BUSKE F. A., FRITH M., GRANT C. E., CLEMENTI L., REN J., LI W. W., NOBLE W. S., 2009 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*. **37**:W202-W208.
- BAXTER L., JIRONKIN A., HICKMAN R., MOORE J., BARRINGTON C., KRUSCHE P., DYER N. P., BUCHANAN-WOLLASTON V., TISKIN A., BEYNON J., DENBY K., OTT S., 2012 Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell* **10**: 3949-3965.
- BIRCHLER J. A., VEITIA R. A., 2010 The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytology*. **186**: 54-62.
- BIRCHLER J. A., VEITIA R. A., 2007 The gene balance hypothesis : from classical genetics to modern genomics. *The Plant Cell*. **19**: 395-402.
- BRILLI M., FANI R., LIO P., 2007 MotifScorer: using a compendium of microarrays to identify regulatory motifs. *Bioinformatics* **23**: 493-495.
- BROWN P., BAXTER L., HICKMAN R., BEYNON J., MOORE J.

- D., OTT S., 2013 MEME-LaB: motif analysis in clusters. *Bioinformatics* **29**: 1696-1697.
- BUCHANAN C., KLEIN P., MULLET J., 2004 Phylogenetic analysis of 5'-noncoding regions from the ABA-d responsive rab16/17 gene family of sorghum, maize and rice provides insight into the composition, organization and function of cis-regulatory modules. *Genetics* **168**: 1639-1654.
- BULGER M., GROUDINE M., 2010 Enhancers: the abundance and function of regulatory sequences beyond promoters. *Developmental Biology* **339**: 250-257.
- BUSKE F. A., BODEN M., BAUER D. C., BAILEY T. L., 2010 Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* **26**: 860-866.
- CARLSON J. M., CHAKRAVARTY A., DEZIEL C. E., GROSS R. H., 2007 SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Research* **35**: W259-64.
- CASTELLANA N. E., PAYNE S. H., SHEN Z., STANKE M., BAFNA V., BRIGGS S. P., 2008 Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci USA* **105**: 21034-21038.
- CHENG F., WU J., FANG L., SUN S., LIU B., LIN K., BONNEMA G., WANG X., 2012 Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PloS One* **7**: e36442.
- CHENG H., KIM B. H., GRISHIN N. V., 2008 MALISAM: a database of structurally analogous motifs in proteins. *Nucleic acids research* **36**: D211-217.
- CLAEYS M., STORMS V., SUN H., MICHOEL T., MARCHAL K., 2012 MotifSuite: workflow for probabilistic motif detection and assessment. *Bioinformatics* **28**: 1931-1932.
- DAVULURI R. V., SUN H., PALANISWAMY S. K., MATTHEWS

- N., MOLINA C., KURTZ M., GROTEWOLD E., 2003. AGRIS: Arabidopsis gene regulatory information server., an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC bioinformatics* **4**: 25.
- DERMITZAKIS E. T., CLARK A. G., 2002 Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**: 1114-1121.
- DEVOS K. M., BROWN J. K., BENNETZEN J. L., 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. *Genome Research*. **127**:1075-1079.
- DOI K., HOSAKA A., NAGATA T., SATOH K., SUZUKI K., MAULEON R., MENDOZA M. J., BRUSKIEWICH R., KIKUCHI S., 2008 Development of a novel data mining tool to find cis-elements in rice gene promoter regions. *BMC plant biology* **8**: 20-30.
- DRAKE J., BIRD C., NEMESH J., THOMAS D., NEWTON-CHEH C., RAYMOND A., EXCOFFLER L., ATTAR H., ANTONARAKIS S., DERMITZAKIS E et al., 2006 Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics* **38**:223-227.
- EDGER P. P., PIRES J. C., 2009 Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res* **175**:699-717.
- FELTUS F. A., SINGH H. P., LOHITHASWA H. C., et al 2006 A Comparative Genomics Strategy for Targeted Discovery of Single Nucleotide Polymorphisms and Conserved-Noncoding Sequences in Orphan Crops. *Plant Physiology* **1404**: 1183–1191.
- FORCE A., LYNCH M., PICKETT F. B., AMORES A., YAN Y. L., POSTLETHWAIT J., 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **1514**: 1531-1545.

- FREELING M., WOODHOUSE M. R., SUBRAMANIAM S.,
TURCO G., LISCH D., SCHNABLE J. C., 2012. Fractionation
mutagenesis and similar consequences of mechanisms removing
dispensable or less-expressed DNA in plants. *Current opinion in
plant biology* **152**: 131-139.
- FREELING M., RAPAKA L., LYONS E., PEDERSEN B., THOMAS
B. C., 2007 G-boxes, bigfoot genes, and environmental
response: characterization of intragenomic conserved noncoding
sequences in *Arabidopsis*. *Plant Cell* **195**:1441-1457.
- FREELING M., SUBRAMANIAM S., 2009 Conserved noncoding
sequences CNSs in higher plants. *Current Opinion in Plant
Biology* **122**:126-132.
- FREELING M., 2009 Bias in plant gene content following different
sorts of duplication: tandem, whole-genome, segmental, or by
transposition. *Annual Review of Plant Biology* **60**: 433-453.
- FREELING M., THOMAS B. C., 2006 Gene-balanced duplications,
like tetraploidy, provide predictable drive to increase
morphological complexity. *Genome Research* **16**: 805-814.
- GAO L. Z., INNAN H., 2004 Very low gene duplication rate in the
yeast genome. *Science* **306**:1367-1370.
- GAO Z., ZHAO R., RUAN J., 2013 A genome-wide cis-regulatory
element discovery method based on promoter sequences and
gene co-expression networks. *BMC genomics* **14**: S1-4.
- GARCIA-FERNANDEZ J., 2005 Hox, ParaHox, ProtoHox: facts and
guesses. *Heredity* **942**: 145-152.
- GARSMEUR O., SCHNABLE J. C., ALMEIDA A., JOURDA C., D'HONT
A., FREELING M., 2013 Two evolutionarily distinct classes of
paleopolyploidy. *Mol Biol Evol* doi:10.1093/molbev/mst230.
- GREENE B., WALKO R., HAKE S., 1994 Mutator insertions in an
intron of the maize knotted1 gene result in dominant
suppressible mutations. *Genetics* **1384**:1275-1285.

- GUO H., MOOSE S. P., 2003 Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **155**:1143-1158.
- GUTURU H., DOXEY A. C., WENGER A. M., BEJERANO G., 2013 Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philosophical transactions of the Royal Society of London Series B., Biological sciences* **368**: doi:10.1098/rstb.2013.0029.
- HAEUSSLER M., 2011 When needles look like hay: how to find tissue-specific enhancers in model organism genomes. *Developmental Biology*. **350**: 239-254.
- HANADA K., ZHANG Z., BOREVITZ J. O., LI W. H., SHIU S. H., 2007 A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Research*. **17**: 632-640.
- HANCOCK J. M., SHAW P. J., BONNETON F., DOVER G. A., 1999 High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. *Mol Biol Evol*. **16**:253-265.
- HARDISON R. C., 2000 Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **169**:369-372.
- HARRIS R. S., 2007 Improved Pairwise Alignment of Genomic DNA. Ph.D. thesis, The Pennsylvania State University.
- HAUDRY A., PLATTS A. E., VELLO E., HOEN D. R., LECLERCQ M., WILLIAMSON R. J., FORCZEK E., JOLY-LOPEZ Z., STEFFEN J. G., HAZZOURI K. M ET AL., 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature genetics* **458**: 891-898.
- HE Q., BARDET A. F., PATTON B., PURVIS J., JOHNSTON J.,

- PAULSON A., GOGOL M., STARK A., ZEITLINGER J., 2011 High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nature Genetics*. **43**: 414-420.
- HE X., ZHANG J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **1692**: 1157-1164.
- HIGO K., UGAWA Y., IWAMOTO M., HIGO H., 1998 PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic acids research* **261**: 358-359.
- HOLLISTER J. D., ROSS-IBARRA J., GAUT B. S., 2010 Indel-associated mutation rate varies with mating system in flowering plants. *Mol Biol Evol.* **27**:409-416.
- HONG J. W., HENDRIX D. A., LEVINE M. S., 2008 Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.
- INADA D. C., BASHIR A., LEE C., THOMAS B. C., KO C., GOFF S. A., FREELING M., 2003 Conserved noncoding sequences in the grasses. *Genome Res* **139**:2030-2041.
- JUNION G., SPIVAKOV M., GIRARDOT C., BRAUN M., GUSTAFSON E. H., BIRNEY E., FURLONG E. E., 2012 A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**:473-486.
- KADAUKE S., BLOBEL G. A., 2009 Chromatin loops in gene regulation. *Biochim Biophys Acta* **1789**:17-25.
- KAPLINSKY N. J., BRAUN D. M., PENTERMAN J., GOFF S. A., FREELING M., 2002 Utility and distribution of conserved noncoding sequences in the grasses. *Proceedings of the National Academy of Sciences of the United States of America*. **99**:6147-6151.
- KIM J., YI G. S., 2013 RMOD: a tool for regulatory motif detection in

- signaling network. PLoS one **87**: DOI: 10.1371/journal.pone.0068407.
- KIM S., PRITCHARD J., 2007 Adaptive evolution of conserved noncoding elements in mammals. PLoS Genetics **39**:10.137journal.pgen.0030147
- KOONIN E. V., 2005 Orthologs, paralogs and evolutionary genomics. Annual Review of Genetics **39**:309-338.
- LAJOIE M., GASCUEL O., LEFORT V., BREHELIN L., 2012 Computational discovery of regulatory elements in a continuous expression space. Genome biology **13**: R109.
- LEE A. P., KERK S. Y., TAN Y. Y., BRENNER S., VENKATESH B., 2011 Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Molecular Biology and Evolution **283**:1205-1215.
- LEVINE M., 2010 Transcriptional enhancers in animal development and evolution. Curr Biol **2017**:R754-763.
- LEWIS E. B., 1951 Pseudoallelism and gene evolution. Cold Spring Harb Symp on Quant Biol **16**: 159-174.
- LIMAN E. R., INNAN H. 2003 Relaxed selective pressure on an essential component of pheromone transduction in primate evolution. Proceedings of the National Academy of Sciences of the United States of America **100**:3328-3332.
- LIN H., OUYANG S., EGAN A., NOBUTA K., HAAS B. J., ZHU W., GU X., SILVA J. C., MEYERS B. C., BUELL C. R., 2008 Characterization of paralogous protein families in rice. BMC Plant Biol **8**:18-32.
- LISERON-MONFILS C., LEWIS T., ASHLOCK D., MCNICHOLAS P. D., FAUTEUX F., STROMVIK M., RAIZADA M. N., 2013 Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. BMC plant biology **13**: 42-59.

- LOOTS G. G., LOCKSLEY R. M., BLANKESPOOR C. M., WANG Z. E., MILLER W., RUBIN E. M., FRAZER K. A., 2000 Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136-140.
- LUDWIG M. Z., BERGMAN C., PATEL N. H., KREITMAN M., 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564-567.
- LYNCH M., FORCE A., 2000 The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**:459-473.
- LYONS E., FREELING M., 2008 How to usefully compare homologous plant genes and chromosomes as DNA sequence. *The Plant Journal* **53**:661-673.
- MOSES A. M., POLLARD D. A., NIX D. A., IYER V. N., LI X. Y., BIGGIN M. D., EISEN M. B., 2006 Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput Biol* **210**:1219-1231.
- NEEDLEMAN S. B., WUNSCH C. D., 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* **48**:443-53.
- NELSON A. C., WARDLE F. C., 2013 Conserved non-coding elements and cis regulation : actions speak louder than words. *Development.* **140**: 1385-1395.
- OH E., KANG H., YAMAGUCHI S., PARK J., LEE D., KAMIYA Y., CHOI G., 2009 Genome-wide analysis of genes targeted by Phytochrome interacting factor 3-like5 during seed germination in *Arabidopsis*. *Plant Cell.* **21**:403-19.
- OKUMURA T., MAKIGUCHI H., MAKITA Y., YAMASHITA R.,

- NAKAI K., 2007 Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions Nucl. Acids Res. **35**: 227-231.
- OSSOWSKI S., SCHNEEBERGER K., LUCAS-LIEDO J. I., WARTHMAN R. M., SHAW R. G., WEIGEL D., LYNCH M., 2010 The rate and molecular spectrum of spontaneous mutations in arabidopsis thaliana. Science **327**: 92-94.
- PENNACCHIO L. A., LOOTS G. G., NOBREGA M. A., OVCHARENKO A., 2007 Predicting tissue-specific enhancers in the human genome. Genome Res. **17**:201-11.
- PETROV D. A., LOZOVSKAYA E. R., HARTL D. L., 1996 High intrinsic rate of DNA loss in Drosophila. Nature. **384**:346-9.
- PETROV D. A., 2002. Mutational equilibrium model of genome size evolution. Theor Popul Biol. **61**:531-544.
- POLULIAKH N., TAKAGI T., NAKAI K., MELINA: motif extraction from promoter regions of potentially co-regulated genes. Bioinformatics **193**: 423-424.
- RAATZ B., EICKER A., SCHMITZ G., FUSS E., MULLER D., ROSSMAN S., THERES K., 2011 Specific expression of LATERAL SUPPRESSOR is controlled by an evolutionarily conserved 3' enhancer. Plant J. **68**:400-412.
- RASTOGI S., LIBERLES D. A., 2005 Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC evolutionary biology **51**: 28-35.
- REINEKE A. R., BORNBERG-BAUER E., GU J., 2011 Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic acids research. **39**: 6029-6043.
- RITTER D. I., LI Q., KOSTKA D., POLLARD K. S., GUO S.,

- CHUANG J. H., 2012 The importance of being cis : evolution of orthologous fish and mammalian enhancer activity. *Molecular Biology Evolution*. **27**: 2322-2332.
- ROTH C., RASTOGI S., ARVESTAD L., DITTMAR K., LIGHT S., EKMAN D., LIBERLES D. A., 2007 Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol* **3081**: 58-73.
- SALVI S., SPONZA G., MORGANTE M., TOMES D., NIU X., FENGLER K. A., MEELEY R., ANANIEV E. V., SVITASHEV S., BRUGGEMANN E ET AL., 2007 Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci USA* **104**:11376-11381.
- SANKOFF D., ZHEN C., ZHU Q., 2010 The collapse of gene complement following whole genome duplication. *BMC Genomics*. **11**:313-324.
- SCHNABLE J. C., PEDERSEN B. S., SUBRAMANIAM S., FREELING M., 2011a. Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses. *Frontiers in Plant Genetics and Genomics*. **2**:1-7.
- SCHNABLE J. C., SPRINGER N. M., FREELING M., 2011b. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 4069-4074.
- SCHNABLE P. S., WARE D., FULTON R. S., STEIN J. C., WEI F., PASTERNAK S., LIANG C., ZHANG J., FULTON L., GRAVES T. A et al., 2009 The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**:1112-1115.
- SCHRIDER D. R., COSTELLO J. C., HAHN M. W., 2009 All human-

- specific gene losses are present in the genome as pseudogenes. *J Comput Biol.* **16**:1419-1427.
- SÉMON M., WOLFE K. H., 2007 Consequences of genome duplication. *Current Opinion in Genetics & Development.* **17**: 505-512.
- SHAMEER K., AMBIKA S., VARGHESE S. M., KARABA N., UDAYAKUMAR M., SOWDHAMINI R., 2009 STIFDB-Arabidopsis Stress Responsive Transcription Factor DataBase. *International journal of plant genomics* **583429**:1-9.
- SPANGLER J. B., SUBRAMANIAM S., FREELING M., FELTUS F. A., 2011 Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. *New Phytol.* **193**:241-252.
- SPANGLER J. B., FICKLIN S. P., LUO F., FREELING M., FELTUS A., 2012 Conserved non-coding regulatory signatures in *Arabidopsis* co-expressed gene modules. *PLoS One.* **7**: e45041, 1-7.
- STADLER L., 1951 Spontaneous mutation in maize. *Cold Spring Harbor Symp in Quant Biol* **16**: 49-63.
- STEFFENS N. O., GALUSCHKA C., SCHINDLER M., BULOW L., HEHL R., 2004 AthaMap: an online resource for in silico transcription factor binding sites in the *Arabidopsis thaliana* genome. *Nucleic acids research* **32**: D368-372.
- SUBRAMANIAM S., FREELING M., 2012 Conserved noncoding sequences in plant genomes. In *Plant Genome Diversity*, ISBN 978-3-7091-1129-1. Volume **1**: Plant genomes., their residents, and their evolutionary dynamics ed. JF Wendel. Springer.
- SUBRAMANIAM S., WANG X., FREELING M., PIRES J. C., 2013 The fate of *Arabidopsis thaliana* homeologous CNSs and their motifs in the Paleohexaploid *Brassica rapa*. *Genome biology and evolution* **54**: 646-660.
- TANG H., LYONS E., PEDERSEN B., SCHNABLE J. C.,

- PATERSON A. H., FREELING M., 2011 Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics*. **12**:102-113.
- TANG H., LYONS E., 2012 Unleashing the genome of *Brassica rapa*. *Frontiers in Plant Science*. **3**: 1-12.
- TANG H., WOODHOUSE M. R., CHENG F., SCHNABLE J. C., PEDERSEN B. S., CONANT G., WANG X., FREELING M., PIRES J. C., 2012 Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* **190**: 1563-1574.
- THIJS G., MARCHAL K., LESCOT M., ROMBAUTS S., DE MOOR B., ROUZÉ P., MOREAU Y., 2002 A Gibbs Sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *Journal of Computational Biology special issue Recomb'2001*. **9**: 447-464.
- THOMAS B. C., RAPAKA L., LYONS E., PEDERSEN B., FREELING M., 2007 Arabidopsis intragenomic conserved noncoding sequence. *Proceedings of the National Academy of Sciences of the United States of America*. **104**:3348-3353.
- TIAN Z., RIZZON C., DU J., ZHU L., BENNETZEN J. L., JACKSON S. A., GAUT B. S., MA J., 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res*. **19**:2221-2230.
- TRAN M., SCHULTZ C., BAUMANN U., 2008 Conserved upstream open reading frames in higher plants. *BMC genomics* **9**: 361-378.
- TURCO G., SCHNABLE J. C., PEDERSEN B., FREELING M., 2013 Automated Conserved Non-Coding Sequence CNS Discovery Reveals Differences in Gene Content and Promoter Evolution among Grasses. *Frontiers in Plant Science* **4**: 170-185.
- UCHIDA N., TOWNSLEY B., CHUNG K. H., SINHA N., 2007

- Regulation of SHOOT MERISTEMLESS genes via an upstream-conserved noncoding sequence coordinates leaf development. *Proc Natl Acad Sci USA* **104**:15953-15958.
- VAN DE PEER Y., 2011 A mystery unveiled. *Genome Biology*. **12**:113.
- VANDEPOELE K., CASNEUF T., VAN DE PEER Y., 2006 Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biology* **7**:R103.
- VEITIA RA., 2010 A generalized model of gene dosage and dominant negative effects in macromolecular complexes. *FASEB J* **24**:994-1002.
- VON ROHR P., FRIBERG., M. T., KADARMIDEEN H. N., 2007 Prediction of transcription factor binding sites using genetical genomics methods. *J. Bioinform Comput Biol.* **5**:773-793.
- WANG X., et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature genetics*. **43**:1035-1039.
- WANG X., TANG H., BOWERS J. E., FELTUS F. A., PATERSON A. H., 2007 Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**:1753-1763.
- WEEKS K. E., CHUZHANOVA N. A., DONNISON I. S., SCOTT I. M., 2007 Evolutionary hierarchies of conserved blocks in 5'-noncoding sequences of dicot *rbcS* genes. *BMC Evol Biol* **7**:51.
- WICKER T., BUCHMANN J. P., KELLER B., 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.* **20**:1229-1237.
- WILCOXON F., 1945 Individual comparisons by ranking methods. *Biometrics Bulletin* **1** **6**: 80–83.
- WOLFE K. H., 2001 Yesterday's polyploids and the mystery of

- diploidization. *Nature reviews Genetics*. **2**:333-341.
- WOODHOUSE M. R., SCHNABLE J. C., PEDERSEN B. S., LYONS E., LISCH D., SUBRAMANIAM S., FREELING M., 2010 Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS biology* **8**: e1000409.
- XU S., CLARK T., ZHENG H., VANG S., LI R., WONG G. K-S., WANG J., ZHENG X., 2008 Gene conversion in the rice genome. *BMC Genomics* **9**:1-8.
- ZHANG W., WU Y., SCHNABLE J. C., ZENG Z., FREELING M., CRAWFORD G. E., JIANG J., 2012. High-resolution mapping of open chromatin in the rice genome. *Genome Res.* **22**:151-162.