Published online 5 April 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 14* ***6029–6043***
doi:10.1093/nar/gkr179

Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes

Anna R. Reineke, Erich Bornberg-Bauer and Jenny Gu*

Institute for Evolution and Biodiversity, University of Münster, Hüfferstrasse 1, 48149, Münster, Germany

Received June 25, 2010; Revised February 22, 2011; Accepted March 15, 2011

ABSTRACT

The discovery of regulatory motifs embedded in upstream regions of plants is a particularly challenging bioinformatics task. Previous studies have shown that motifs in plants are short compared with those found in vertebrates. Furthermore, plant genomes have undergone several diversification mechanisms such as genome duplication events which impact the evolution of regulatory motifs. In this article, a systematic phylogenomic comparison of upstream regions is conducted to further identify features of the plant regulatory genomes, the component of genomes regulating gene expression, to enable future *de novo* discoveries. The findings highlight differences in upstream region properties between major plant groups and the effects of divergence times and duplication events. First, clear differences in upstream region evolution can be detected between monocots and dicots, thus suggesting that a separation of these groups should be made when searching for novel regulatory motifs, particularly since universal motifs such as the TATA box are rare. Second, investigating the decay rate of significantly aligned regions suggests that a divergence time of ~100 mya sets a limit for reliable conserved non-coding sequence (CNS) detection. Insights presented here will set a framework to help identify embedded motifs of functional relevance by understanding the limits of bioinformatics detection for CNSs.

INTRODUCTION

A major focus in the post-genomic era is to understand the temporal expression of genes defining developmental stages (1), physiological states (2), stress responses (3) and adaptive fitness (4). Gene expression is controlled by regulatory elements embedded in non-coding regions of the genome, the discovery of which remains elusive (5). Many genomic surveys of conserved non-coding sequences (CNSs), which are used as a proxy to identify embedded putative gene expression regulators, are proving to be particularly challenging (6–13). While the conservation of examples such as the TATA-box *cis*-regulatory motif has been identified by Berendzen *et al.* (14) across all species, the number of such universal highly conserved motifs are few and limited.

Unlike *cis*-elements of vertebrates that are frequently ≥100 bp in length (15), experimentally identified plant *cis*-elements are infrequently over 30 bp with a median observed length of 8 bp (8,16,17). Bioinformatic scans of genomes estimate plant CNSs to have a predicted median length of 25 bp (18); however, the functional relevance of these CNSs remains to be confirmed. While CNS does not necessarily imply a regulatory role, they are often used as proxies to identify possibly embedded *cis*-regulatory motifs. Second, while nuclear genes of plants and animals show similar substitution rates (9,19,20), non-coding regions in plants appear to have a higher degeneracy (16,21). This may be a consequence of differences in genomic evolutionary mechanisms observed between animals and plants. Genomes of plants have been found to be more diverse than vertebrate genomes due to increased duplication events (22), polyploidy (23,24), increased recombination (25), transposable elements (TEs) (26) and gene silencing (21,27,28), for example. These evolutionary processes compound to the challenges of detecting CNSs when conservation becomes difficult to distinguish against the degenerate background frequency (13). Therefore, the bioinformatics discovery of possibly embedded regulatory motifs for further characterization and understanding of gene regulation is also complicated by the decay of strong functionally important sequence signatures in upstream regions.

Several bioinformatics strategies have been employed to identify putative CNSs and embedded regulatory motifs, among which is to leverage conservation through sequence comparisons that suggest possible functional importance.

*To whom correspondence should be addressed. Tel: 0049-251 - 83-21086; Fax: 0049-251 - 83-24668; Email: j.gu@uni-muenster.de

Early approaches use probability-based algorithms employing a variation of strategies such as Expectation Maximization and Gibbs sampling to find overrepresentation of motifs. These algorithms include MEME (29), MotifSampler (30) and AlignACE (31). Improvements in recent algorithms are often achieved through flexible search parameters (32,33), suffix-trees (34), development of mixed-models (35,36), aided analysis with supplementary high-throughput experimental data (30,37–39), prior knowledge of transcription factor binding site features (40–42), implementation of graph-based methods (43) and the incorporation of phylogenetic relationships (7,44). Finally, consensus interpretation of results from multiple different algorithms has been proposed to improve the discovery of conserved motifs (34,45). The reported success of these strategies, however, has mostly been applied and often remains limited to metazoan CNS (13,46–48). More recently, a new method has been developed to identify CNS in plants by determining the statistical significance of aligned segments (49). Nevertheless, the success of any algorithm improves with a stronger understanding of the data from which the desired feature is to be extracted.

As genomes become increasingly available through the advancement of high throughput sequencing, comparative genomics through orthologs and paralogs are becoming a popular strategy yielding some success of CNS identification (6,9,12,50–53). However, the validity of such comparisons needs to be explored, as there are a number of evolutionary mechanisms leading to rapid divergence of sequences, thus rendering them difficult to detect significant similarities. The useful divergence of sequences for comparison is a recognized issue to be considered in comparative genomic-based analysis (54,55). This article addresses the limits of which comparative genomics can be employed specifically in plants to identify putative CNSs by investigating the decay rate of identified significantly aligned sequences in upstream regions with respect to the age of divergence. Most critical to the success of any algorithm is having the proper null model such that statistical significance of CNSs can be properly assigned. This entails understanding the effects of underlying biological processes. Furthermore, defining the divergence limit for CNS discovery also highlights which comparisons will yield interpretable results. Comparisons between upstream regions spanning 5 kb for monocots and dicots were made as well as regions downstream from the coding sequence. The fundamental importance of these findings provides practical guidelines and considerations for future successful research efforts in understanding the regulatory genome of plants.

MATERIALS AND METHODS

Data sets used for phylogenomic comparisons

Comparative analysis of upstream regions between the following plant genomes were used: *Arabidopsis thaliana* v9.0 (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release/) (56), *Arabidopsis lyrata* v1.0 (http://www.jgi.doe.gov/, http://genome.jgi-psf.org/Araly1/Araly1.download.ftp.html), *Carica papaya* v1.0 (http://www.life.illinois.edu/ming/, ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Cpapaya/) (57), *Glycine max* v1.01 (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Gmax/annotation/) (58), *Medicago truncatula* v3.0 (http://www.medicago.org/genome/downloads/Mt3/) (59,60), *Populus trichocarpa* v2.0 (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Ptrichocarpa/annotation/) (61), *Ricinus communis* v0.1 (http://www.phytozome.net/ricinus, http://castorbean.jcvi.org/castorbean_downloads.shtml), *Manihot esculenta* v1.1 (http://www.phytozome.net/cassava, ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Mesculenta/), *Prunus persica* v1 (http://www.phytozome.org/peach), *Cucumus sativa* v1 (http://www.phytozome.net/cucumber.php) (62), *Zea mays* v4a.53 (http://ftp.maizesequence.org/current/filtered-set/) (63), *Sorghum bicolor* v1.0 (ftp://ftp.jgi-psf.org/pub/JGI_data/Sorghum_bicolor/v1.0/Sbi/annotation/Sbi1.4/) (64), *Oryza sativa* v6.0 (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.1/all.dir/) (65), *Bryachypodium distachyon* v1.0 (http://files.brachypodium.org/Annotation/) (66), *Ostreococcus lucimarinus* v2.0 (http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.download.ftp.html) (67), *Ostreococcus tauri* v2.0 (http://genome.jgi-psf.org/Ostta4/Ostta4.download.ftp.html) (67) and *Micromonas pusilla* v2.0 (http://genome.jgi-psf.org/MicpuC2/MicpuC2.download.ftp.html) (68).

Estimating time of divergence between compared genomes

MUSCLE (69) was used to align rbcL, matK and atpb genes, markers that have been used for previous phylogenetic tree constructions (http://www.ncbi.nlm.nih.gov/) (70–73). BEAST (74) was used to construct a phylogenetic tree with these alignments to estimate divergence times between the species used for this investigation. BEAUTI (74) was used to define the following settings for BEAST: the WAG substitution model, a relaxed clock model, randomly generated starting tree, and MCMC chain length of 10 000 000. Taxons were set based on the APG III classification (72) with the following priors: *A. thaliana–A. lyrata* ~5.15 mya (75), *P. trichocarpa–R.communis* ~81 mya (76), *A. thaliana–C. papaya* ~72 mya (76) and *Z. mays–S. bicolor* ~16 mya (77). *Glycine max* and *Medicago truncatula* is estimated to diverge ~50–54 mya (personal communication with Nevin D. Young). The first 200 trees were burned with TreeAnnotator (74), BEAST results were analyzed with Tracer (74) and the resulting phylogenetic trees were visualized with FigTree (74). The divergence times of the three gene trees generated from rbcL, matK and atpb genes (Supplementary Data S1) were averaged.

Ortholog detection and comparative analysis

Inparanoid (78) was used to identify corresponding orthologous genes between the genomes using default parameters. The longest splice variants were used as the representative for orthologs, and identified clusters were separated into two sets; a set containing only one-to-one relationships between two orthologous genes (singleton orthologs) and a second set with clusters containing multiple genes (multiple orthologs). The alignment similarities for multiple orthologs clusters were calculated using DIALIGN for all possible combinatorial orthologous pairs. The best and average similarity scores for each multiple orthologs clusters were used for comparison in the analysis.

Retrieval of upstream and downstream regions

Upstream regions were extracted based on the annotated transcription start site [(TSS) corresponding to the mRNA start position noted in the gff3-file] and translation start site (ATG–the start codon position found noted in gff3-file) for each respective genome data sets. The retrieved upstream regions were then partitioned into 10 × 500 bp segments from 0 to +5 kb. Downstream regions were extracted with the length of 500 bp from the end of the transcript representing the longest splice variant (−500 bp).

Tool performance comparisons to identify CNS

Several tools were used and compared to identify aligned upstream regions with a minimum length of 8 bp which we designate in this investigation as putative CNS: BLASTN (79), CHAOS (80), DIALIGN (81) and LAGAN (82). Comparisons were made using the calculated sequence similarities for +500 bp upstream of the TSS between all plant genome pairings, and the results were then used as benchmark. Upstream regions were aligned when each sequence in the comparison contained no more than 5% missing nucleotides and have a length of 500 bp. Default settings were used for DIALIGN and LAGAN for alignment. For these tools, CNSs were defined by an aligned region with a minimum of 8 consecutive base pairs in length. Alignments were concatenated when separated by a maximum of five unaligned bases.

Default parameter settings and a word length of 8 bp were used for BLASTN and CHAOS to detect sequence similarity. Alignments were also concatenated using the following strategy to calculate similarity scores for CNS: (i) identified overlapping alignments were resolved first by taking the longer of two alignments (ii) for overlapping alignments of equal length, the one with a lower *e*-value was selected.

The similarity scores S of alignments using these tools were calculated by the sum of non-overlapping CNS lengths located within the respective 500 bp segment, where *i* is the identified putative CNS:

$$S_{\text{score}} = \frac{\sum_{i=1}^{n} \text{length}(\text{CNS}_i)}{500\text{bp}}$$

This metric was used as a measure to detect the reducing size of CNS search space with respect to evolutionary decay and proximity from either the transcription or translation start site. Similarity score distributions of 2000 randomly chosen upstream regions for each compared plant genomes were used to construct cumulative null models for dicots and monocots, respectively, used in the calculation of statistical significance.

Identifying effects of indels and TEs

Effects of insertions, deletions, and TEs were investigated using DIALIGN (81) and BLASTN (79) to find CNS disruptions within the range of +5 kb upstream regions of TSS and ATG. Cross comparison alignments were made between every 500 bp segments up to 5 kb in the upstream regions of compared species. Two 500 bp regions are loosely defined as significantly similar with similarity scores above the threshold of 1.5 from the mean of the null model. Corresponding 500 bp sequences with direct positional best scores are assumed to have no disruptions in upstream regions resulting from indels.

TEs were identified with RepeatMasker (http://www.repeatmasker.org) in the 5 kb upstream regions with a div-setting of 20% against the angiosperm library. The distribution of TE in sequences was calculated by counting all masked positions of the sequences in the RepeatMasker output file, where all interspersed repeats in the sequence are masked.

Statistical analysis

Statistical tests were performed on the alignment similarity score distributions of dicots, monocots, and algae using the Mann–Whitney U test with R. Statistical analyses between all plant pair comparisons were performed on similarity scores calculated with the singleton ortholog data set using the Kruskal–Wallis test, a non-parametric ANOVA-like multivariance test from the pgirmess package of R. The Kruskal–Wallis test was also applied for the cross-comparison analysis between all singleton orthologs and the comparison between downstream and upstream regions. Both the Mann–Whitney U-test and Kruskal–Wallis test were applied to identify differences in CNS amount and length.

RESULTS

Workflow for CNS divergence and tool comparisons

A workflow including ortholog detection, comprehensive cross-comparison of upstream regions up to 5 kb, tool comparisons, and TE identification was implemented to understand the decay rate of CNS in plant regulatory genomes (Figure 1). CNSs are defined in this investigation as aligned upstream regions of similarity with a length of >8 bp. First, technical considerations were addressed through tool comparisons, data quality and divergence time between compared genomes using only the first +500 bp of upstream regions from different starting reference points. The search space for CNSs was then expanded to 5 kb in order to discern positional effects and those that may be introduced by putative TEs.

The first 500 bp upstream regions using both the ATG and TSS reference points were extracted for all orthologous genes (Figure 1, left path) to address

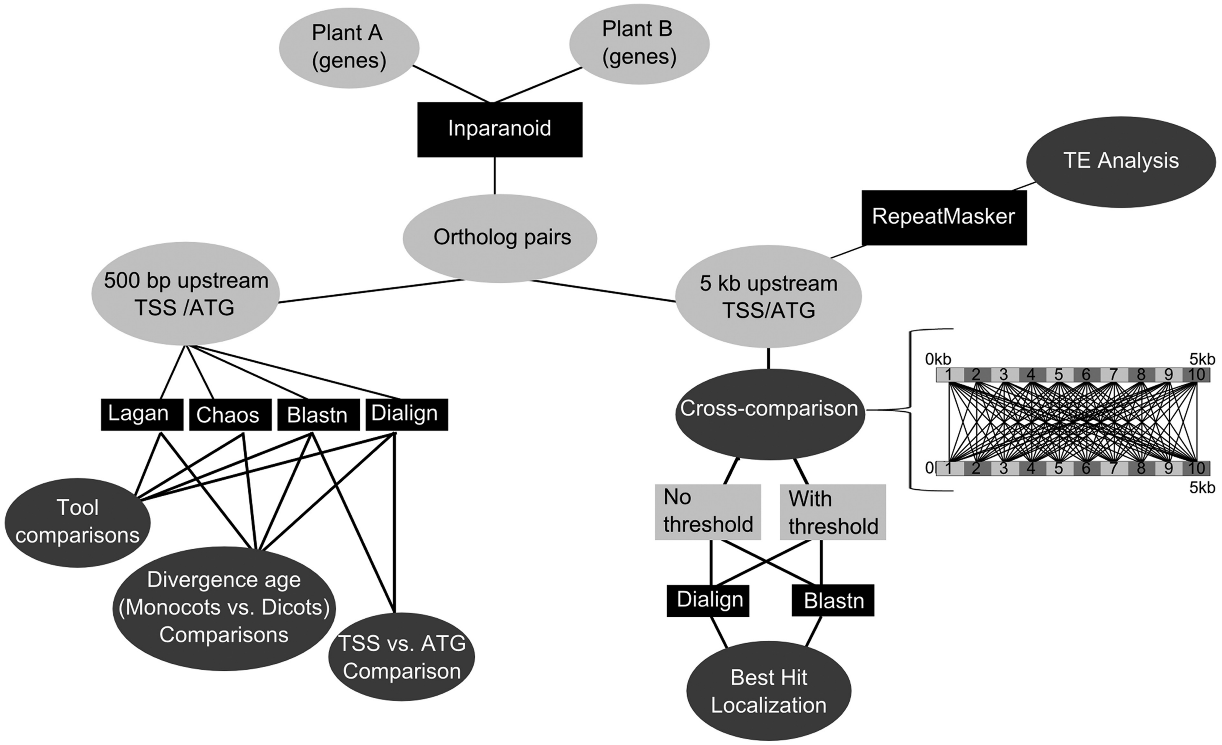


Figure 1. Workflow for phylogenomic CNS identification. A workflow was implemented to address the effects of tool performance, data quality, divergence age and localization for bioinformatic identification of putative CNSs when using comparative genomics in plants. Orthologous gene pairs were found with Inparanoid and upstream regions of genes were extracted using both the TSS and ATG reference points for a length of 500 bp and 5 kb. Tool comparisons were performed using the first 500 bp of orthologously paired upstream regions as benchmark. Comparisons of upstream region similarity with respect to divergence time, proximity to TSS and ATG and between monocots and dicots were conducted. Cross-comparisons between 500 bp segments covering a search space of 5 kb were also conducted to localize regions with the highest alignment similarity. RepeatMasker was used to identify putative TEs.

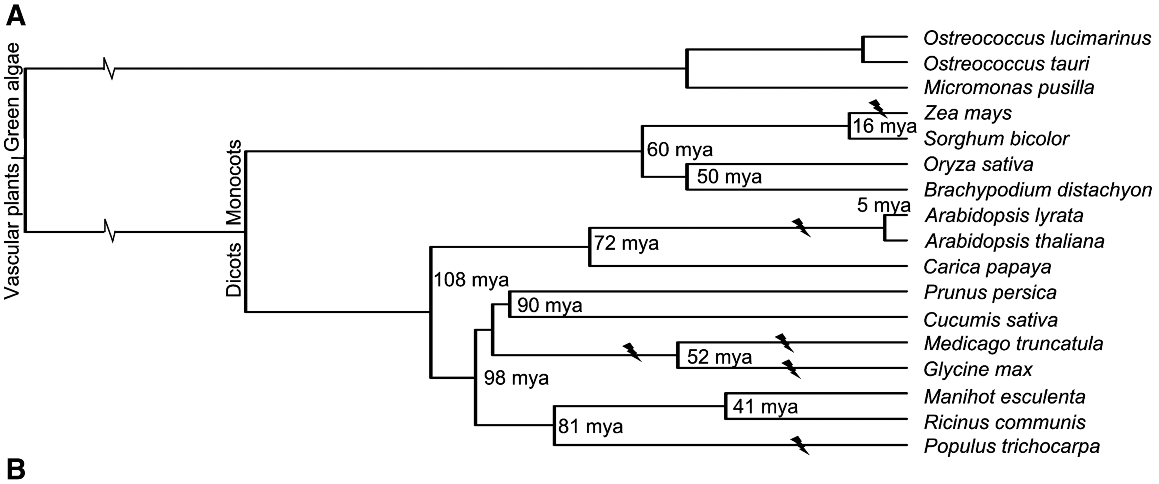
technical considerations in searching for CNSs. Orthologs between two genomes were found using Inparanoid (78). For this study, singleton orthologs were used as the benchmark to ensure our analysis addresses effects of evolutionary decay rather than alternative interfering mechanisms that may be associated with duplication and multiple orthologs. Singleton orthologs, genes with only direct one-to-one relationships, have been shown to have an increased possibility of sharing gene function compared with multiple orthologs where neo- or subfunctionalization often occurs after gene duplication in terms of gene function (83).

Pairwise upstream region similarities were then compared to evaluate tool performances (Supplementary Figure S3). The tools evaluated were BLASTN (79), CHAOS (80), DIALIGN (81) and LAGAN (82). Comparisons showed BLASTN and DIALIGN to perform best, and therefore these tools were used to estimate the decay of upstream regions with respect to divergence time. Furthermore, comparison of similarity scores and CNS properties between monocots and dicots, using both the ATG/TSS reference points, were performed for the first 500 bp upstream region.

After benchmarking with the first 500 bp, the search space for putative CNSs was then expanded to include 5 kb upstream regions using both the ATG and TSS (Figure 1, right path) as reference points. A cross-comparison of all ten 500 bp segments of the extracted 5-kb was performed to localize regions with the highest aligned similarity. Interspersed repetitive elements that may be putative TEs in the 5-kb upstream region were also analyzed with RepeatMasker.

Estimation of divergence time between plant species

Since a phylogenomic strategy was employed in this investigation, the phylogenetic relationship and age of divergence must be estimated between all compared species. The majority of the genome data sets used in this investigation have previously published divergence times, which made it possible to estimate the split between the remaining species using BEAST (Figure 2) (74). The result suggests that *R. communis* and *M. esculenta* diverged more recently at ~41 mya compared with the divergence age of *G. max* and *M. truncatula* estimated at 50–54 mya (N. D. Young, personal communication). Since the published divergence times are based on different methods, a comparison of these results may not be appropriate. Nevertheless, the published data and results of the BEAST analysis are good estimates for our research aims and the sequence of divergence events is more important than the exact divergence times for this investigation. Due to possible substitution rate variations between the different plants pairs (84), *O. sativa* was used as the fixed species for comparisons in monocots, and *A. thaliana*



B

Cross comparison chart

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Ostreococcus lucimarinus																	
2. Ostreococcus tauri	X																
3. Micromonas pusilla	X																
4. Zea mays					X	X											
5. Sorghum bicolor				X		X											
6. Oryza sativa				X	X		X										
7. Brachypodium distachyon						X											
8. Arabidopsis lyrata									X								
9. Arabidopsis thaliana								X		X						X	
10. Carica papaya									X								
11. Prunus persica												X					
12. Cucumis sativa											X						
13. Medicago truncatula														X		X	
14. Glycine max													X				
15. Manihot esculenta																X	
16. Ricinus communis															X		
17. Populus trichocarpa									X				X			X	

Figure 2. Phylogenetic tree with estimated divergence times used for comparative analysis. (**A**) Phylogenetic tree and estimated divergence times were calculated using BEAST (74). The following previously published divergence times were used: *C. papaya–A. thaliana,* 72 mya (76); *A. thaliana–P.trichocarpa,* 108 mya (76); *P. tichocarpa–M. truncatula,* 98 mya (76); *P. persica–C. sativa,* 90 mya (76); *R. communis–P. trichocarpa,* 81 mya (76); *Z. mays–S. color,* 16 mya (77); *O. sativa–B. distachyon,* 50 mya (64) and *O. sativa–S. bicolor,* 60 mya (64). Divergence time of *R. manihot–M. esculenta* and *G. max–M. truncatula* were estimated using BEAST with rbcL, atpb and matK genes. The divergence time of the Euphorbiaceae pair (*R. communis–M. esculenta*) was more recent compared with the divergence time of the Fabaceae pair (*G. max–M. truncatula*). WGD events (lightning) are also marked (17). (**B**) Table of comparisons made between plants. Fixed plant species, depending on the divergence age being compared in the analysis, are circled.

and *P. trichocarpa* for dicot comparisons, depending on the possible comparisons for the given divergence age (Figure 2B).

Comparison of alignment tool performance in identifying CNS

Alignment performance between different tools was conducted through comparing similarity scores resulting from BLASTN(79), CHAOS (80), DIALIGN (81) and LAGAN (82), and comparisons revealed DIALIGN and BLASTN to perform best (Figure 3, Supplementary Figures S3 and S4). The results of DIALIGN is comparable with those of BLASTN (79) and better than that of LAGAN (82) and CHAOS (80). For DIALIGN and BLASTN, alignments could be made for nearly all sequence pairs to calculate similarity, which is necessary for estimating the decay rate, whereas LAGAN and CHAOS did not successfully identify CNSs in many sequence comparisons and therefore no similarity score could be calculated (Supplementary Figure S3). BLASTN implements a local alignment strategy whereas DIALIGN combines local alignment to seed subsequent global alignment. Both of these tools have previously been used in other CNS investigations (6,8,85,86). A case study using a known heat shock element (74,75) shows that both DIALIGN and BLASTN were successful in identifying the motif TTCnnGAA, with DIALIGN being more sensitive (Figure 3). Results from using DIALIGN are reported for the remaining of this investigation, however complementary BLASTN results can also be found in Supplementary Data.

Selection of reference point: transcription versus translation start sites

Due to variable genome qualities, a comparison of alignments starting from different reference points, the transcription (TSS) and translational start sites (ATG), were compared to determine possible effects on alignment performance. Using the ATG as the reference point showed slight improvements in the similarity scores of

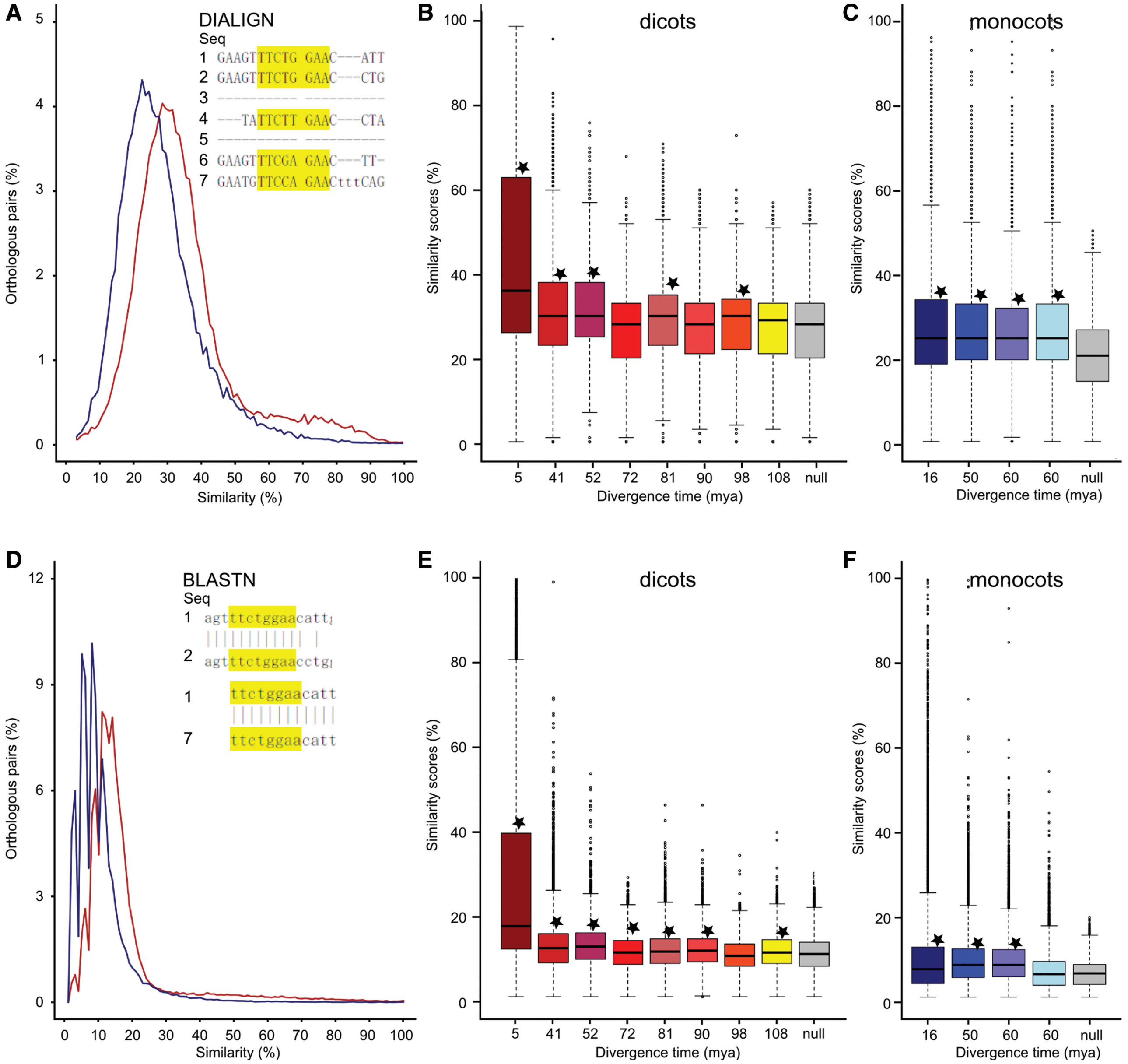


Figure 3. Upstream region conservation in dicots, monocots and algae. (**A**) The distribution of the similarity levels for orthologous upstream regions in all studied dicot plant pairs (red) and monocot plant pairs (blue) are shown for DIALIGN. Similarity scores of monocots has a lower distribution compared with that of dicots (Mann–Whitney U-test, $P = 2.2e^{-16}$). (**B**) Distribution of similarity scores using DIALIGN with respect to divergence time and null model (grey) for dicots and (**C**) monocots. Median values are shown (center black bars). (**D**) Results from using BLASTN for dicots (red) and monocots (blue) are significantly different (Mann–Whitney U-test, $P = 2.2e^{-16}$). (**E**) Similarity scores with respect to divergence time and null model (grey) for dicots and (**F**) monocots analyzed with BLASTN. Significantly different distributions with respect to null-models constructed based on randomly paired upstream regions are marked (asterisk, Kruskal–Wallis test, $P = 0.01$). Case study results of tool performance using well characterized heat shock elements are shown, respectively. DIALIGN identified heat shock element in orthologous genes from five out of seven plants [(i) *A. thaliana*, (ii) *A. lyrata*, (iii) *C. papaya*, (iv) *P. trichocarpa*, (v) *M. truncatula*, (vi) *G. max* and (vii) *R. communis*] whereas BLASTN only identified the motif in three out of seven plants [(i) *A. thaliana*, (ii) *A. lyrata*, (iii) *C. papaya*, (iv) *P. trichocarpa*, (v) *M. truncatula*, (vi) *G. max* and (vii) *R. communis*].

+500 bp upstream in monocots (Figure 3C and Supplementary Figure S4.1.C). The majority of distances between TSS and ATG were calculated to be <500 bp (Supplementary Figure S6). For dicots however, TSS versus ATG have no effects on the results. The distribution of ATG–TSS distances between monocots and dicots are not significantly different (Supplementary Figure S6A), but there seems to be an effect on the distance between the annotated sites due to genome quality. The *Z. mays* genome shows a large fraction of distances between TSS and ATG to be >500 bp. Incidentally the *Z. mays* genome also contain the highest amount of masked regions indicative of poor sequence quality. These observations suggest upstream regions in *Z. mays* to be of poor quality compared with the other plant species used in this study. Alternatively, large distances between TSS and ATG may be caused by incorrect annotation of the first exon or alternative TSSs (87). For example, some genes contain multiple TSSs and the median distance between the two TSSs was observed to be 184 and 149 bp for *A. thaliana* and *O. sativa,* respectively (87). Finally, genome quality continues to be an issue as the annotation of these genomes improves. A substantial portion of the distances could not be calculated between the TSS and ATG sites due to the lack of positional information and, subsequently, annotation (Supplementary Figure S6). For example, previous investigations have shown that ~66% of the TSSs in *A. thaliana* are annotated with available 5′-UTR positional information (88). Results from using both the TSS and ATG reference points are included and are reported (Supplementary Data).

Monocot CNS decay faster than those found for dicots

Significantly lower similarity scores were observed in monocots compared with dicots using the Mann–Whitney U-test (Figure 3AD, $P = 2.2e^{-16}$). As an outgroup family, three green algae species from the Mamiellaceae family, *Ostreococcus lucimarinus*, *Ostreococcus tauri* and *Micromonas pusilla*, were also included for alignment score comparisons using the ATG as calculated with DIALIGN and BLASTN (Supplementary Figure S4.1A and S4.2A). The TSS information for algae was not available and therefore was not conducted. The distributions of upstream region similarity scores for the three different plant groups were found to differ significantly (Mann–Whitney U-test, $P = 2.2e^{-16}$). Furthermore, while similarity values between orthologs observed for dicots differ from those observed for green algae, the distribution curves both have a similarity peak around ~35% compared with the monocot distribution which peaked at ~25% (Supplementary Figure S4.1A and S4.2A).

Variation in selection pressures in the non-coding regions immediately flanking the transcripts was also investigated. A comparison of the downstream region (−500 bp) with the first two upstream 500 bp segments in monocots and dicots show a difference in the ranking of regions with the highest distribution of similarity scores. The first +500 bp in the upstream region was ranked to have the highest similarity values, followed by the downstream region and second 500 bp segment (+0.5–1 kb upstream). In contrast, monocots had better alignment scores in downstream regions followed by the first and then second 500 bp segments upstream of the TSS (Supplementary Figure S8, Kruskal–Wallis test, $p = 0.001$). The findings further suggest differences in regulatory genome evolution of dicots and monocots.

Decay of upstream regions with respect to divergence time

Comparative analysis of singleton orthologous upstream regions was conducted between plant genomes paired based on the estimated divergence time calculated with BEAST (see ‘Materials and Methods’ section). The results show that similarity scores between orthologous upstream regions of different genome pairings decrease with increasing divergence age (Figure 3). The significance of similarity values becomes difficult to distinguish from the null model of randomly paired genes as genomes approach a divergence time around 100 mya ± 10 mya, and therefore comparative genomic identification of CNSs is challenged at this divergence limit. The observation holds true for both monocots and dicots.

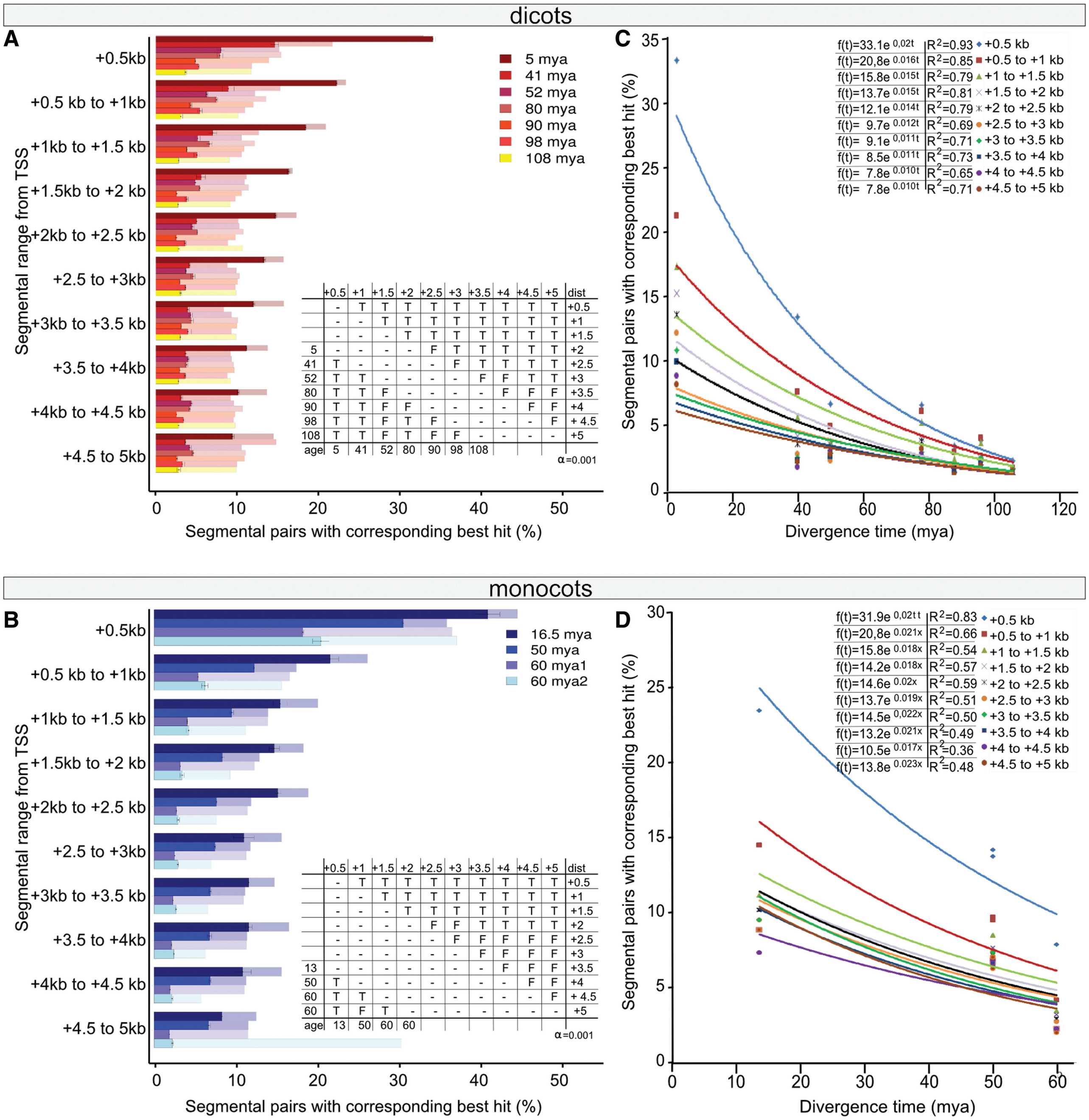
Gene duplication events leading to sub- or neofunctionalization may also effect the conservation of upstream regions, and therefore potential impacts were investigated for gene families containing multiple paralogs. Upstream regions of singleton orthologs were compared with gene families with multiple orthologs (Supplementary Figure S2). Two measures were used to discern if differences in the similarity scores could be observed between these two groups of orthologs. First, the average calculated similarity score for alignments between all multiple orthologs pairings was compared with the similarity values observed for singleton orthologs. The second measure used the best calculated similarity scores for each compared pairings within multiple orthologs. Results show that while the distribution of best hit similarity scores is significantly higher than that of singleton orthologs, the averages of pairwise comparisons between multiple orthologs are lower (Wilcoxon test, $P = 2.2e^{-16}$). The lower variance and average similarity values observed for multiple orthologs, compared with the best calculated similarity values within the set and singletons, suggests that upstream regions for multiple orthologs may be subjected to less selection pressure to maintain conservation (Supplementary Figure S2).

Decay rate with respect to distance from the TSS

Since TEs can make up nearly half the genome of many species and reshape genomic structure (89), the mobility of these elements can present problems in detecting CNS. Therefore upstream regions were aligned with respect to the identified orthologous gene coding regions to account for possible displacement events. Furthermore, each 500 bp segments of the upstream regions were aligned across 5 kb of the orthologous upstream region to account for possible indel events resulting in shifted frames. The analysis was done reciprocally between all compared species with the exception of *C. papaya* due to the lack of available upstream data beyond 500 bp. This cross-combinatorial design also addresses potential issues with respect to errors in TSS annotation which can significantly affect alignment results (Figure 1).

Scans of 500 bp segments across 5 kb show possible effects of TEs in repositioning putative CNSs. Direct conservation, wherein significantly aligned regions corresponded to the orthologous 500 bp segment without relocation, appeared mostly in the +500 bp region for all plant pairs with detected occurrences from ~35% of *A. thaliana*–*A. lyrata* comparisons decreasing with divergence time to ~8% in *A. thaliana*–*P. trichocarpa* for dicots (Figure 4A). Monocots also show the same results starting from ~40% in *Z. mays*–*S. bicolour* to ~20% for *O. sativa*–*S. bicolour* (Figure 4B). Additionally, the frequencies of direct conservation also decrease with distance from the TSS (Figure 4AB). Similar results can be observed irrespective of whether the ATG or BLASTN were used instead (Supplementary Figure S5). The decay appears to plateau at ~3 kb upstream for dicots (Figure 4A) and ~2.5 kb for monocots (Figure 4B). As expected, the highest instances of significantly aligned regions can be detected with shorter divergence time and in closer proximity to the TSS (Kruskal–Wallis test, $P < 0.001$, Figure 4A and B).

The decay rates of dicot upstream regions can be fitted with an exponential function (Figure 4C) and are observed to be highest in near proximity to the TSS. The first +500 bp segment in the upstream region have the highest decay rate of $d(t) = 33.1e^{0.02t}$ followed by the rate for the second segment (+0.51.0 kb) with $d(t) = 20.8e^{0.016t}$.



	+0.5	+1	+1.5	+2	+2.5	+3	+3.5	+4	+4.5	+5	dist
	-	T	T	T	T	T	T	T	T	T	+0.5
	-	-	T	T	T	T	T	T	T	T	+1
	-	-	-	T	T	T	T	T	T	T	+1.5
5	-	-	-	-	F	T	T	T	T	T	+2
41	T	-	-	-	-	F	T	T	T	T	+2.5
52	T	T	-	-	-	-	F	F	T	T	+3
80	T	T	F	-	-	-	-	F	F	F	+3.5
90	T	T	F	F	-	-	-	-	F	F	+4
98	T	T	F	T	F	-	-	-	-	F	+ 4.5
108	T	T	F	T	F	F	-	-	-	-	+5
age	5	41	52	80	90	98	108				

f(t)=33.1e ^{0,02t}	R^2=0.93	◆	+0.5 kb
f(t)=20,8e $^{0.016t}$	R^2=0.85	■	+0.5 to +1 kb
f(t)=15.8e $^{0.015t}$	R^2=0.79	▲	+1 to +1.5 kb
f(t)=13.7e $^{0.015t}$	R^2=0.81	×	+1.5 to +2 kb
f(t)=12.1e $^{0.014t}$	R^2=0.79	✳	+2 to +2.5 kb
f(t)= 9.7e $^{0.012t}$	R^2=0.69	●	+2.5 to +3 kb
f(t)= 9.1e ^{0,011t}	R^2=0.71	◆	+3 to +3.5 kb
f(t)= 8.5e $^{0.011t}$	R^2=0.73	■	+3.5 to +4 kb
f(t)= 7.8e $^{0.010t}$	R^2=0.65	●	+4 to +4.5 kb
f(t)= 7.8e $^{0.010t}$	R^2=0.71	●	+4.5 to +5 kb

	+0.5	+1	+1.5	+2	+2.5	+3	+3.5	+4	+4.5	+5	dist
	-	T	T	T	T	T	T	T	T	T	+0.5
	-	-	T	T	T	T	T	T	T	T	+1
	-	-	-	T	T	T	T	T	T	T	+1.5
	-	-	-	-	F	F	T	T	T	T	+2
	-	-	-	-	-	F	F	F	F	F	+2.5
	-	-	-	-	-	-	F	F	F	F	+3
13	-	-	-	-	-	-	-	F	F	F	+3.5
50	T	-	-	-	-	-	-	-	F	F	+4
60	T	T	-	-	-	-	-	-	-	F	+ 4.5
60	T	F	T	-	-	-	-	-	-	-	+5
age	13	50	60	60							

f(t)=31.9e ^{0,02tt}	R^2=0.83	◆	+0.5 kb
f(t)=20,8e $^{0.021x}$	R^2=0.66	■	+0.5 to +1 kb
f(t)=15.8e $^{0.018x}$	R^2=0.54	▲	+1 to +1.5 kb
f(t)=14.2e $^{0.018x}$	R^2=0.57	×	+1.5 to +2 kb
f(t)=14.6e $^{0.02x}$	R^2=0.59	✳	+2 to +2.5 kb
f(t)=13.7e $^{0.019x}$	R^2=0.51	●	+2.5 to +3 kb
f(t)=14.5e ^{0,022x}	R^2=0.50	◆	+3 to +3.5 kb
f(t)=13.2e $^{0.021x}$	R^2=0.49	■	+3.5 to +4 kb
f(t)=10.5e $^{0.017x}$	R^2=0.36	●	+4 to +4.5 kb
f(t)=13.8e $^{0.023x}$	R^2=0.48	●	+4.5 to +5 kb

Figure 4. Upstream region decay with respect to divergence time. Detected decay rate of upstream regions based on DIALIGN results. (**A**) Percentage of direct regional conservation is shown to decrease with divergence time and distance from the TSS for dicots and (**B**) monocots (Kruskal–Wallis test $P = 0.001$, transparent bars). A correction was applied to identify alignments with similarity levels >1.5σ from the mean of the null model (solid bars). Statistical difference between comparisons is noted in the inserted table (T = true and F = false). (**C**) The decay rate of identified putative CNSs was fitted with an exponential function for an estimation of the decay rate in dicots and (**D**) monocots. The exponential decay rate and the corresponding correlation coefficient (R^2) are shown in the inserted table.

Decay rates in monocots were also estimated (Figure 4D), although there is insufficient data for a proper fitting at this time. As soon as more genomes become available for monocot comparisons, this topic will have to be revisited.

Distribution of putative TEs

The amount of sequences with detected interspersed repeats that may be putative TEs varied significantly between the different plant species. In some cases, nearly no interspersed repeats were detected in species such as *P. persica* and *A. lyrata* while ~70% of *O. sativa* upstream regions are detected to contain TE (Figure 5A). Statistically significant differences could not be detected in the amount of TE between monocots and dicots. Interestingly, less interspersed repeats were detected in the first 500 bp upstream of the ATG compared with regions further upstream for all plants (Figure 5B). This result explains the strong differences in the decay rate between the first +500 bp compared with segments further upstream (Figure 4C and D). This analysis was conducted using the upstream region of the ATG instead of the TSS to also identify putative TE between these two reference points. The findings support a potentially higher success rate of identifying embedded motifs in the first +500 bp segment for initial bioinformatics analysis. For *Z. mays,* large portion of sequences were already masked due to sequence quality issues and therefore these masked regions in combination with detected interspersed repeats resulted in the higher observances of putative TE of *Z. mays* compared with the other plants.

Other identified CNS features

Some features of CNS were identified in this comparative study of plant regulatory genomes. First, the density of identified CNSs in +500 bp differs significantly between monocots and dicots (Supplementary Figure S7A).

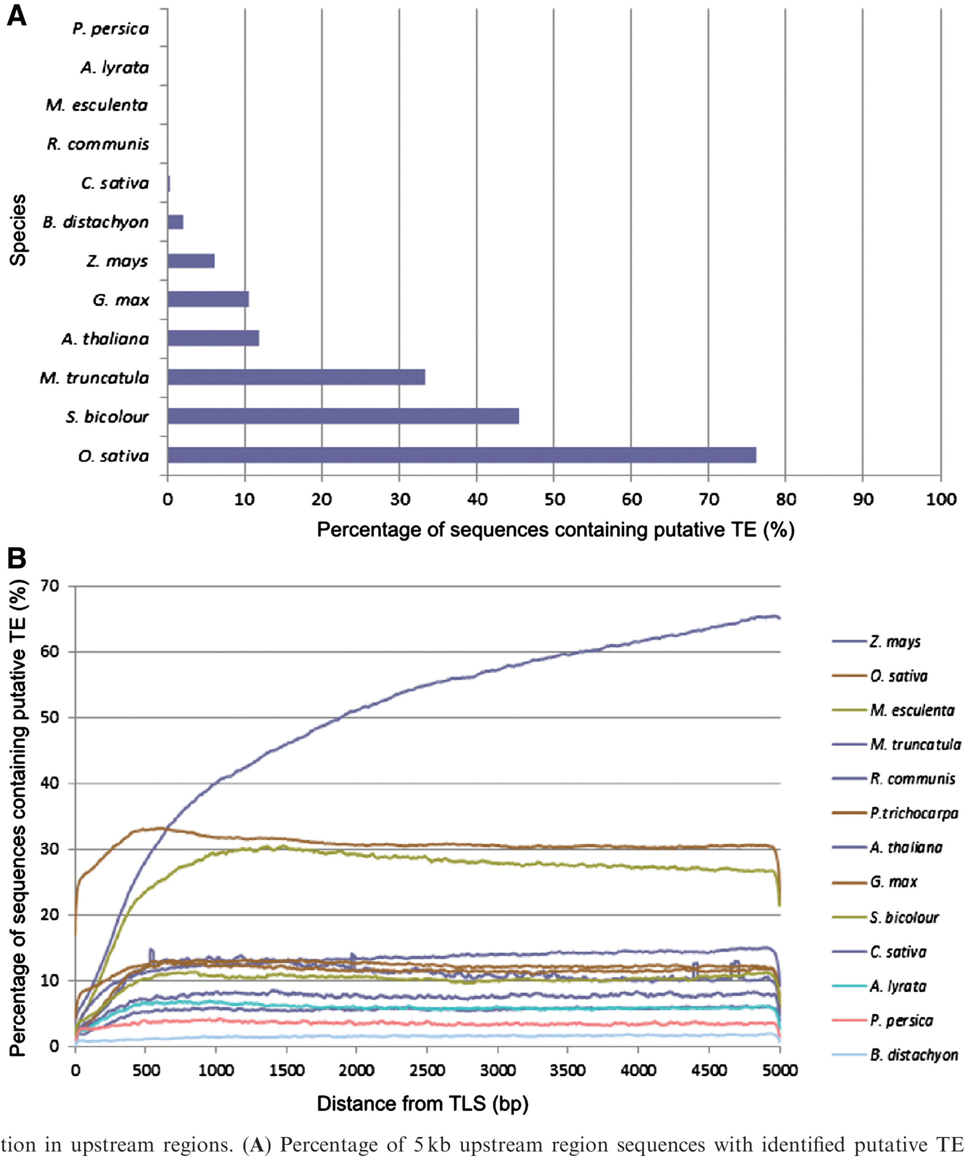


Figure 5. TE distribution in upstream regions. (**A**) Percentage of 5 kb upstream region sequences with identified putative TE in each plant species. (**B**) Percentage of upstream regions with putative TE is plotted with respect to upstream localization from the ATG.

Monocots are found to contain, on average,10 CNSs in the +500 bp upstream region of TSS while dicots show a slightly higher average of 12 CNS (Mann–Whitney U-test, $P = 2.2e^{-16}$, Supplementary Figure S7A). However, using the ATG instead, an average density of 12 CNSs was calculated for both monocots and dicots, with algae showing an average of 14 CNSs (Supplementary Figure S7). Second, the length of CNSs was also investigated (Figure S7B). Monocot CNSs are significantly shorter than those of dicots and algae (Kruskal–Wallis test, $P = 0.001$). The finding holds true for CNSs found in the +500 bp upstream regions of both ATG and TSS. The length of CNSs is observed to increase with decreasing divergence time, as expected (Supplementary Figure S7B). Finally, the nucleotide composition of the first 500 bp segments were also investigated and show significant differences in distribution between dicots, monocots and algae (ANOVA, $P = 2.2e^{-16}$). A higher AT content was observed in dicots (66%) compared with algae containing 32%. A more balanced nucleotide composition is observed for monocots at 51% AT. It should be noted that the two groups, dicots and algae, with a skewed AT-GC content were also identified to have higher detected similarity levels in aligned CNS.

DISCUSSION

The increasing availability of plant genomes allows us to conduct comparative analysis between species to identify conserved features of regulatory genomes. However, previous research efforts have shown that the identification of CNSs and, more importantly, regulatory motifs that form binding sites for transcription factors is difficult due to properties of plant CNSs. Identified motifs are often short, averaging from 20 to 30 bp in length (10), and evolutionary mechanisms such as gene duplication (90,91) and mobile TEs (27,89,92–94) resulting in high diversification are more frequently observed (21). As such, exploring the limits of when comparative genomics is useful for bioinformatics investigations is needed to advance future efforts to understand the regulatory genome.

First, the evolutionary split between the analyzed plant species were successfully reconstructed to estimate the divergence ages necessary for further phylogenomic comparison of upstream regions. Using this constructed time tree (Figure 1), systematic pairwise comparisons between two plant species based on the divergence time were conducted to align orthologous upstream regions. The limits of bioinformatics detection for putative CNSs were explored, although the functional validity of these identified CNSs needs to be further experimentally substantiated. Current probabilistic-based algorithms depend on the fundamental basis that orthologous sequences must share similarity to detect signals of putative CNSs which are used as proxies to subsequently identify possibly embedded *cis*-regulatory motifs.

Both BLASTN and DIALIGN showed similar results (Figure 3), thus also indicating the reliability of the presented results despite having different alignment algorithms. Sensitivity to weakly conserved sequences was important for this study to estimate the decay rate of CNSs, which was missed by both CHAOS and LAGAN (Supplementary Figure S4), although comparable performance with BLASTN and DIALIGN can be achieved in regions of high similarity. Therefore, the use of either BLASTN or DIALIGN is encouraged for investigations requiring alignments of non-coding regions between more distantly related species.

Common among all plant groups, however, is that the similarity of orthologous upstream regions is significantly different from randomly paired upstream regions up to ~100 mya (Figure 3). After which, detection of significant similarity between aligned CNSs is difficult to distinguish from the null model. The findings suggest that the divergence age between compared genomes should be considered to increase the success of finding conserved motifs. The boundary of 100 mya may be the useful divergence limit to identify statistically significant CNSs when using probabilistic-based mining strategies. Other null models or strategies should be considered for more distant comparisons and findings encourage further bioinformatics development to consider the biological features identified in this investigation. Conversely, the divergence time between sister species containing non-saturated substitution patterns, such as between *A. thaliana* and *A. lyrata*, should also not be neglected (95). Identified CNS as proxies for *cis*-regulatory elements between recently diverged species may not reflect true functional conservation. However, when including multiple and more distantly related sequences, the impact of non-saturated substitutions between more closely related species is reduced.

The effects of duplication events on shared similarity levels between orthologous upstream regions were also detected. For example, although the divergence age between *O. sativa*–*Z. mays* is similar to that of *O. sativa*–*S. bicolor*, the findings show orthologous upstream regions from *O. sativa*–*Z. mays* to have lower scores. This may be a consequence of an independent whole genome duplication (WGD) occurring at 11.4 mya followed by gene deletion events in the lineage of *Z. mays* (63). Similarly, both *G. max*–*M. truncatula* share a common WGD followed by additional independent WGDs (23). The divergence age is estimated to be similar to that of *R. communis*–*M. esculenta*, however orthologous upstream regions show lower alignment scores compared with the two Euphorbiaceae. WGDs are frequently observed in plant evolution (21,23,90,91) and must be considered in phylogenomic comparisons.

The distances calculated between TSS and ATG using the current available annotations are mostly <500 bp (Supplementary Figure S6). Monocots show slight improvements in alignments when using the ATG as the reference point compared with TSS (Figure 3 and Supplementary Figure S4). This may likely be caused by the large distances between TSS and ATG in *Z. mays* due to the presence of alternative open reading frames. For example, an average of two TSS with a median distance of 149 bp in *O. sativa* has been observed for each locus (87). Unlike transcription regulating motifs that are mostly found upstream of TSS to activate gene expression, motifs embedded in the 5′-UTR region between the TSS and ATG have higher impact on modulating the abundance of transcripts that are expressed (96,97). Since this study is dependent on the quality of genome annotation, it should be noted that the boundaries of the 5′-UTR and 3′-UTR, and therefore TSS, is not always known (88). The analyses were performed for upstream sequences using both positional reference points, TSS and ATG, to minimize the effects of genome quality on the interpretation of our analysis. Both analysis show the same conclusions and are available for review (Supplementary Data).

The highest region of conservation was identified to be located in the +0.5 kb upstream of the TSS as expected from previous findings, suggesting that the immediate proximity has higher propensity to host modules of regulatory elements (Figure 4AB) (14,98,99). Moreover, the frequency of TEs is lower in this region compared with analyzed segments further upstream which has huge implications for effects of indels on upstream region conservation over the analyzed 5 kb (Figure 5B). The amount of conserved regions is observed to decrease with increasing proximity from the TSS and plateau between +2.5 and 3 kb upstream.

CNS conservation appears to be disrupted by indels such as TEs which are found, for example, in over 70% of upstream regions in rice. The amount of identified TE is found to vary strongly between the different species and agrees with previous observations (100–102). TEs are partly responsible for size variation in closely related species and TE amplification, as well as gene duplication, have been shown to contribute to high *c*-values (101). Nevertheless, cautious interpretation should be taken due to bias that may be introduced by the TE library used in this study. While all species were compared with the same curated angiosperm library implemented in RepeatMasker, the library contains more sequences for the well-studied model plant *O. sativa* compared with recently sequenced plants such as *P. persica* and *R. communis*.

The estimated decay rate using an exponential fitted function shows higher decay rates for +500 bp compared with upstream regions located more distant from the TSS. The fitting suggests most of the observed degeneration occurred immediately following the divergence since <10% of the sequences retained good alignment scores for more distant comparisons. Furthermore, the data also suggest that the largest degeneration can be found closest to the TSS. This may be a consequence of a larger number of identified CNSs compared with other regions (14,103), and therefore resulting in a higher probability of observing a phenotypic change if possibly embedded regulatory motifs are disrupted. Furthermore, as mentioned earlier, recently diverged sister species are not yet saturated with substitutions and therefore a high decay rate of putative CNSs that is yet to be subjected to strong functional selection pressure is not unexpected.

Differences can be clearly observed between monocots and dicots with monocots showing lower alignment scores in the upstream region (Figure 3). Previous studies have shown that monocot chloroplast genomes have a higher evolutionary rate in the coding regions compared with dicots (71). The effects are also evident in the analysis presented here for non-coding regions of the nuclear genome. This suggests that selection of genomes for comparisons between different plant groups should be considered. However, it should be noted that currently available monocot genomes are only from those of grasses and may introduce a bias for the conclusions. In earlier studies, a higher evolutionary rate in grasses compared with other monocots have been shown (73). Subsequent investigation for differences in CNSs associated with non-grass monocots will be needed when these genomes become available for analysis.

Monocots additionally show higher alignment scores in the downstream region (−500 bp), followed by the first and second upstream segments from the TSS. Whereas analysis rank the first 500 bp segment in dicots to have the highest distribution of similarity scores followed by the downstream region having similar alignment score distribution with the second 500 bp segment upstream. This may be indicative of three possible scenarios. First, the selection pressure to conserve CNS in the 0.5 kb upstream and downstream regions may be similar in monocots. Second, the selection pressure in monocots on the +500 upstream region may not be as strong as the corresponding region in dicots. Alternatively, third, CNSs in monocots have a higher decay rate, and therefore significant alignments between orthologous CNS are not retained over a long period of time.

Additional plant group specific features have also been observed with respect to the density, length and composition of nucleotides in CNS. Dicots contain an overall higher density of detected CNSs per 500 bp when compared with monocots (Figure 4). Majority of identified CNS have a length between 8 and 40 bp, suggesting CNS regulatory motif length may be static. Monocot CNSs appear to be significantly shorter compared with dicot and algae CNSs, however the findings may be an artifact resulting from the lower CNS coverage in the 500 bp promoter of monocots (Supplementary Figure S7). Interestingly, this distribution does not change significantly with divergence age. These features provide additional support suggesting different regulatory strategies are employed by respective plant groups. Furthermore, previous studies have shown that some genes corresponding to the GO of stress response have a particularly high amount of CNSs that can be located up to several hundreds base pairs, suggesting gene function should be considered in understanding CNS features (11). In Arabidopsis there are 252 genes identified to be highly enriched in CNS called bigfoot genes (18). Considering that 18 000 orthologous pairs were conducted in this investigation between *A. thaliana* and *A. lyrata,* these bigfoot genes comprise only a minor fraction and therefore have a low impact on this results of this investigation.

Finally, the nucleotide composition between plant groups shows significant differences with preferences for the AT content in dicots and GC content in algae. A higher GC content in rice compared with *A. thaliana* has been found previously. Interestingly, although different nucleotide compositions in both species are observed, an AT enrichment is nevertheless observed in the first +500 bp upstream region of the TSS for both species (87). This may be a result of sequence signature effects associated with the TATA-box. A previous comparison between *A. thaliana* and *A. lyrata* found that mutation rates vary among genomic regions as a function of base composition and is largely dependent on the GC-content (95). GC rich regions have an increased transition:transversion ratio, which may be one contributing factor to the higher observed decay rate (104). The low amount of detected CNSs in monocots compared with dicots could be a result of differences in the decay rate associated with regions of higher GC content which was observed to be 49% in monocots and 33% in dicots. However, this interpretation cannot be applied to clearly explain the observed differences in algae where a higher GC content is also observed but a significant difference in the distribution of identified CNS from dicots was not detected.

CONCLUSIONS

The results of this systematic phylogenetic approach to understand properties of plant CNSs highlighted specific differences and important considerations to guide future research efforts to understand the regulatory component of genomes. First, the selection of plant group and species used for comparative genomics to identify novel regulatory motifs must be considered. We recommend making comparisons only with plants specific to monocots and dicots separately since different evolutionary rates have been observed for each group. Second, we estimate ~100 mya to be the divergence limit to which plant upstream regions can be compared. After which time, significant similarities between putative orthologous CNSs cannot be distinguished from alignments between randomly selected upstream regions. The findings do not imply that CNSs cannot be detected beyond a divergence age of 100 mya since non-coding regions showing no conservation in sequence are nonetheless conserved in function (8). Instead the analysis suggests that alternative considerations of null models and strategies that incorporate additional information for a large scale global search will be needed.

Current algorithms combining multiple features such as incorporating comparative genomics with gene expression data (105,106) and evolutionary models (53,107) to distinguish functional CNSs holds promise. Finally, although we are not the first to investigate the limits of comparative genomics to study CNSs (51), our investigation has successfully addressed these fundamental issues in a systematic, generalized manner with respect to divergence time, plant groups, proximity to the coding region, and other various features that will contribute significantly to our basic fundamental knowledge of CNS identification, guiding future research efforts and algorithm development to understand the transcriptional regulation of genes in systems biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Nevin Dale Young for personal communications with confirmation of the divergence date between *G. max* and *M. truncatula,* Rebecca Freiburg for support with BEAST and Philine Feulner, Andreas Schüler and Juliette de Meaux for the critical reading of this manuscript.

FUNDING

Human Frontier Science Program (R6P0033/2006-C); Alexander von Humboldt Foundation. Funding for open access charge: Alexander von Humboldt Foundation, Deutsche Forschungsgemeinschaft, and the Open Access Publication fund of University of Muenster.

Conflict of interest statement. None declared.

REFERENCES

1. Dinneny,J.R., Long,T.A., Wang,J.Y., Jung,J.W., Mace,D., Pointer,S., Barron,C., Brady,S.M., Schiefelbein,J. and Benfey,P.N. (2008) Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science*, **320**, 942–945.
2. Covington,M.F., Maloof,J.N., Straume,M., Kay,S.A. and Harmer,S.L. (2008) Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.*, **9**, R130.
3. Kilian,J., Whitehead,D., Horak,J., Wanke,D., Weinl,S., Batistic,O., Angelo,C.D. and Harter,K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J.*, **50**, 347–363.
4. Whitehead,A. and Crawford,D.L. (2006) Neutral and adaptive variation in gene expression. *Proc. Natl Acad. Sci. USA*, **103**, 5425–5430.
5. Mattick,J.S., Amaral,P.P., Dinger,M.E., Mercer,T.R. and Mehler,M.F. (2009) RNA regulation of epigenetic processes. *BioEssays*, **31**, 51–59.
6. Inada,D.C., Bashir,A., Lee,C., Thomas,B.C., Ko,C., Goff,S.A. and Freeling,M. (2003) Conserved Noncoding Sequences in the Grasses. *Genome Res.*, **13**, 2030–2041.
7. Van Hellemont,R., Monsieurs,P., Thijs,G., De Moor,B., Van de Peer,Y. and Marchal,K. (2005) A novel approach to identifying regulatory motifs in distantly related genomes. *Genome Biol.*, **6**, R113.
8. Haberer,G., Mader,M.T., Kosarev,P., Spannagl,M., Yang,L. and Mayer,K.F.X. (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. *Plant Physiol.*, **142**, 1589–1602.
9. Guo,H. and Moose,S.P. (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell*, **15**, 1143–1158.
10. Freeling,M. and Subramaniam,S. (2009) Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol.*, **12**, 126–132.
11. Thomas,B.C., Rapaka,L., Lyons,E., Pedersen,B. and Freeling,M. (2007) Arabidopsis intragenomic conserved noncoding sequence. *Proc. Natl Acad. Sci. USA*, **2006**, 3348–3353.
12. Dubchak,I., Brudno,M., Loots,G.G., Pachter,L., Mayor,C., Rubin,E.M. and Frazer,K.A. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.*, **10**, 1304–1306.

13. Priest,H.D., Filichkin,S.A. and Mockler,T.C. (2009) Cis-regulatory elements in plant cell signaling. *Curr. Opin. Plant Biol.*, **12**, 643–649.
14. Berendzen,K.W., Stüber,K., Harter,K. and Wanke,D. (2006) Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, **7**, 522–541.
15. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
16. Higo,K., Ugawa,Y., Iwamoto,M. and Higo,H. (1998) PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res.*, **26**, 358–359.
17. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and trancription factors. *BMC Bioinformatics*, **4**, 25.
18. Freeling,M., Rapaka,L., Lyons,E., Pedersen,B. and Thomas,B.C. (2007) G-boxes, bigfoot genes, and environmental response: Characterization of intragenomic conserved noncoding sequences in Arabidopsis. *Plant Cell*, **19**, 1441–1457.
19. Li,W.H. and Sadler,L.A. (1991) Low Nucleotide Diversity in man. *Genetics*, **129**, 513–523.
20. Chen,J.Q., Wu,Y., Yang,H., Bergelson,J., Kreitman,M. and Tian,D. (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.*, **26**, 1523–1531.
21. Lockton,S. and Gaut,B.S. (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.*, **21**, 60–65.
22. De Bodt,S., Maere,S. and Van de Peer,Y. (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.*, **20**, 591–597.
23. Blanc,G. and Wolfe,K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667–1678.
24. Shoemaker,R.C., Schlueter,J. and Doyle,J.J. (2006) Paleopolyploidy and gene duplication in soybean and other legumes. *Curr. Opin. Plant Biol.*, **9**, 104–109.
25. Gaut,B.S., Wright,S.I., Rizzon,C., Dvorak,J. and Anderson,L.K. (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.*, **8**, 77–84.
26. Freeling,M., Lyons,E., Pedersen,B., Alam,M., Ming,R. and Lisch,D. (2008) Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res.*, **18**, 1924–1937.
27. Woodhouse,M.R., Pedersen,B. and Freeling,M. (2010) Transposed genes in Arabidopsis are often associated with flanking repeats. *PLoS Genet.*, **6**, e1000949.
28. Kejnovsky,E., Leitch,I.J. and Leitch,A.R. (2009) Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol. Evol.*, **24**, 572–582.
29. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2008) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, 369–373.
30. Thijs,G., Moreau,Y., Smet,F.D., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., Moor,B.D. and Marchal,K. (2002) Inclusive: Integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, **18**, 331–332.
31. Roth,F.P., Hughes,J.D., Estep,P.W. and Chrurch,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
32. Wilczynski,B., Dojer,N., Patelak,M. and Tiuryn,J. (2009) Finding evolutionary conserved cis-regulatory modules with a universal set of motifs. *BMC Bioinformatics*, **10**, 82.
33. Thompson,W.A., Newberg,L.A., Conlan,S., McCue,L.A. and Lawrence,C.E. (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res.*, **35**, 232–237.
34. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
35. Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.
36. Zhang,L., Zuo,K., Zhang,F., Cao,Y., Wang,J., Zhang,Y., Sun,X. and Tang,K. (2006) Conservation of noncoding microsatellites in plants: implication for gene regulation. *BMC Genomics*, **7**, 323–337.
37. Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**, 165–171.
38. Kreiman,G. (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.*, **32**, 2889–2900.
39. Chang,L.-W., Nagarajan,R., Magee,J.A., Milbrandt,J. and Stormo,G.D. (2006) A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome Res.*, **16**, 405–413.
40. Wang,Z., Wei,G.H., Liu,D.P. and Liang,C.C. (2007) Unravelling the world of cis-regulatory elements. *Med. Biol. Eng. Comput.*, **45**, 709–718.
41. Klepper,K., Sandve,G.K., Abul,O., Johansen,J. and Drablos,F. (2008) Assessment of composite motif discovery methods. *BMC Bioinformatics*, **9**, 123–139.
42. Chang,W.C., Lee,T.Y., Huang,H.D., Huang,H.Y. and Pan,R.L. (2008) PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics*, **9**, 561–575.
43. Hu,J., Hu,H. and Li,X. (2008) MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs. *Nucleic Acids Res.*, **36**, 4488–4497.
44. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.
45. Phan,V. and Furlotte,N.A. (2008) Motif Tool Manager a web-based framework for motif discovery. *Bioinformatics*, **24**, 2930–2931.
46. Doi,K., Hosaka,A., Nagata,T., Satoh,K., Suzuki,K., Mauleon,R., Mendoza,M.J., Bruskiewich,R. and Kikuchi,S. (2008) Development of a novel data mining tool to find cis-elements in rice gene promoter regions. *BMC Plant Biol.*, **8**, 20.
47. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y.T., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
48. Li,N. and Tompa,M. (2006) Analysis of computational approaches for motif discovery. *Algorithms Mol. Biol.*, **1**, 8.
49. Picot,E., Krusche,P., Tiskin,A., Carre,I. and Ott,S. (2010) Evolutionary analysis of regulatory sequences (EARS) in plants. *Plant J.*, **64**, 165–176.
50. Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.
51. Lyons,E., Pedersen,B., Kane,J., Alam,M., Ming,R., Tang,H., Wang,X., Bowers,J., Paterson,A., Lisch,D. *et al.* (2008) Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, **148**, 1772–1781.
52. Kaplinsky,N.J., Braun,D.M., Penterman,J., Goff,S.A. and Freeling,M. (2002) Utility and distribution of conserved noncoding sequences in the grasses. *Proc. Natl Acad. Sci. USA*, **99**, 6147–6151.
53. He,X., Ling,X. and Sinha,S. (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. *PLoS Comput. Biol.*, **5**, e1000299.
54. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
55. Creux,N.M., Ranik,M., Berger,D.K. and Myburg,A.A. (2006) Comparative analysis of orthologous cellulose synthase promoters

from Arabidopsis, Populus and Eucalyptus : evidence of conserved regulatory elements in angiosperms. *New Phytol.*, **179**, 722–737.

56. Kaul,S., Koo,H.L., Jenkins,J., Rizzo,M., Rooney,T., Tallon,L.J., Feldblyum,T., Nierman,W., Benito,M.I., Lin,X.Y. *et al.* (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
57. Ming,R., Hou,S.B., Feng,Y., Yu,Q.Y., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L.T. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991–997.
58. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
59. Young,N.D., Cannon,S.B., Sato,S., Kim,D., Cook,D.R., Town,C.D., Roe,B.A. and Tabata,S. (2005) Sequencing the genespaces of Medicago truncatula and Lotus japonicus. *Plant Physiol.*, **137**, 1174–1181.
60. Mudge,J., Vasdewani,J., Schiex,T., Spannagl,M., Monaghan,E., Nicholson,C., Oldroyd,G.E., Humphray,S.J., Schoof,H., Mayer,K.F.X. *et al.* (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc. Natl Acad. Sci. USA*, **103**, 14959–14964.
61. Tuskan,G.A., DiFazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2010) The genome of black cottonwood, Populus trichcarpa(Torr & Grey). *Science*, **313**, 1596–1604.
62. Huang,S.W., Li,R.Q., Zhang,Z.H., Li,L., Gu,X.F., Fan,W., Lucas,W.J., Wang,X.W., Xie,B.Y., Ni,P.X. *et al.* (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet.*, **41**, 1275–1281.
63. 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.
64. Dubchak,I., Grimwood,J., Gundlach,H., Paterson,A.H., Bowers,J.E., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A., Schmutz,J. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
65. Goff,S.A., Goff,S.A., Ricke,D., Lan,T.H., Glazebrook,J., Sessions,A., Oeller,P., Varma,H., Lange,B.M., Moughamer,T. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp). *Science*, **296**, 92–100.
66. Vogel,J.P., Garvin,D.F., Mockler,T.C., Schmutz,J., Rokhsar,D., Bevan,M.W., Barry,K., Lucas,S., Harmon-Smith,M., Lail,K. *et al.* (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
67. Palenik,B., Grimwood,J., Aerts,A., Rouze,P., Salamov,A., Putnam,N., Dupont,C., Jorgensen,R., Derelle,E., Rombauts,S. *et al.* (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA*, **104**, 7705–7710.
68. Worden,A.Z., Lee,J.-h., Mock,T., Rouzé,P., Simmons,M.P., Aerts,A.L., Allen,A.E., Cuvelier,M.L., Derelle,E., Everett,M.V. *et al.* (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science*, **324**, 268–272.
69. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
70. Wikström,N., Savolainen,V. and Chase,M.W. (2001) Evolution of the angiosperms: calibrating the family tree *Proc. Biol. Sci.*, **268**, 2211–2220.
71. Chaw,S.M., Chang,C.C., Chen,H.L. and Li,W.H. (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J. Mol. Evol.*, **58**, 424–441.
72. Bremer,B., Bremer,K., Chase,M.W., Fay,M.F., Reveal,J.L., Soltis,D.E., Soltis,P.S., Stevens,P.F., Anderberg,A.A., Moore,M.J. *et al.* (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.*, **161**, 105–121.
73. Sanderson,M.J., Thorne,J.L., Winkström,N. and Bremer,K. (2004) Molecular evidence on plant divergence times. *Am. J. Bot.*, **91**, 1656–1665.
74. Drummond,A.J. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214–222.
75. Koch,M.A., Haubold,B. and Mitchell-Olds,T. (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (Brassicaceae). *Mol. Biol. Evol.*, **17**, 1483–1498.
76. Forest,F. and Chase,M.W. (2009) Eurosid I. In Hedges,S.B. and Kumar,S. (eds), *The timetree of life*. Oxford University Press, New York, pp. 188–196.
77. Gaut,B.S. and Doebley,J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA*, **94**, 6809–6814.
78. Brien,K.P.O., Remm,M. and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, 476–480.
79. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.
80. Brudno,M., Chapman,M., Gottgens,B., Batzoglou,S. and Morgenstern,B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
81. Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998) DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
82. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A., Batzoglou,S. and Progra,N.C.S. (2003) LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
83. Ganko,E.W., Meyers,B.C. and Vision,T.J. (2007) Divergence in expression between duplicated genes in Arabidopsis. *Mol. Biol. Evol.*, **24**, 2298–2309.
84. Tang,H., Wang,X., Bowers,J.E., Ming,R., Alam,M. and Paterson,A.H. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.*, **18**, 1944–1954.
85. Kohn,M.H. (2008) Rapid sequence divergence rates in the 5 prime regulatory regions of young Drosophila melanogaster duplicate gene pairs. *Gene Expression*, **584**, 575–584.
86. Guo,X., Wang,Y., Keightley,P.D. and Fan,L. (2007) Patterns of selective constraints in noncoding DNA of rice. *BMC Evol. Biol.*, **7**, 208.
87. Tanaka,T., Koyanagi,K.O. and Itoh,T. (2009) Highly diversified molecular evolution of downstream transcription start sites in rice and arabidopsis. *Plant Physiol.*, **149**, 1316–1324.
88. Chung,B.Y., Simons,C., Firth,A.E., Brown,C.M. and Hellens,R.P. (2006) Effect of 5′UTR introns on gene expression in Arabidopsis thaliana. *BMC Genomics*, **7**, 120.
89. Deragon,J.M., Casacuberta,J.M. and Panaud,O. (2008) Plant transposable elements. *Genome Dyn.*, **4**, 69–82.
90. Van de Peer,Y., Maere,S. and Meyer,A. (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, **10**, 725–732.
91. Freeling,M., Thomas,B.C., Freeling,M. and Thomas,B.C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.*, **16**, 805–814.
92. Tenaillon,M.I., Hollister,J.D. and Gaut,B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends Plant Sci.*, **15**, 471–478.
93. Lockton,S. and Gaut,B.S. (2009) The contribution of transposable elements to expressed coding sequence in Arabidopsis thaliana. *J. Mol. Evol.*, **68**, 80–89.
94. Hollister,J.D. and Gaut,B.S. (2009) Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, **19**, 1419–1428.
95. Derose-Wilson,L.J. and Gaut,B.S. (2007) Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of Arabidopsis thaliana and Arabidopsis lyrata. *BMC Evol. Biol.*, **7**, 66–78.

96. Liu,W.X., Liu,H.L., Chai,Z.J., Xu,X.P., Song,Y.R. and Qu,L.Q. (2010) Evaluation of seed storage-protein gene 5′ untranslated regions in enhancing gene expression in transgenic rice seed. *Theor. Appl. Genet.*, **121**, 1267–1274.
97. Wang,C.T. and Xu,Y.N. (2010) The 5′ untranslated region of the FAD3 mRNA is required for its translational enhancement at low temperature in Arabidopsis roots. *Plant Sci.*, **179**, 234–240.
98. Pan,D. and Zhang,L. (2008) A holistic view of evolutionary rates in paralogous and orthologous genes. *Lect. Notes Comp. Sci.*, **5227**, 967–974.
99. Lichtenberg,J., Yilmaz,A., Welch,J.D., Kurz,K., Liang,X., Drews,F., Ecker,K., Lee,S.S., Geisler,M., Grotewold,E. *et al.* (2009) The word landscape of the non-coding segments of the Arabidopsis thaliana genome. *BMC Genomics*, **10**, 463–394.
100. Lockton,S. and Gaut,B.S. (2010) The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. *BMC Evol. Biol.*, **10**, 10.
101. Hawkins,J.S., Kim,H., Nason,J.D., Wing,R.A. and Wendel,J.F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.*, **16**, 1252–1261.
102. Piegu,B., Guyot,R., Picault,N., Roulin,A., Saniyal,A., Kim,H., Collura,K., Brar,D.S., Jackson,S., Wing,R.A. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res.*, **16**, 1262–1269.
103. Molina,C. and Grotewold,E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.
104. Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**, 6.
105. Wang,X., Haberer,G. and Mayer,K.F.X. (2009) Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics*, **10**, 284.
106. Baele,G., Van de Peer,Y. and Vansteelandt,S. (2009) Efficient context-dependent model building based on clustering posterior distributions for non-coding sequences. *BMC Evol. Biol.*, **9**, 87–110.
107. Linder,H.P. and Rudall,P.J. (2005) Evoltionary history of poales. *Ann. Rev. Ecol. Evol. Syst.*, **36**, 107–124.