Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes

Konstantinos Kritsas, Samuel E. Wuest, Daniel Hupalo, et al.

Genome Res. 2012 22: 2455-2466 originally published online September 17, 2012
Access the most recent version at doi:10.1101/gr.129346.111

Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes

Konstantinos Kritsas,[1] Samuel E. Wuest,[2,5] Daniel Hupalo,[3] Andrew D. Kern,[4] Thomas Wicker,[1,6] and Ueli Grossniklaus[1,6]

[1]*Institute of Plant Biology & Zürich-Basel Plant Science Center, University Zürich, CH-8008 Zürich, Switzerland;* [2]*Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland;* [3]*Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA;* [4]*Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA*

Ultraconserved elements (UCEs), stretches of DNA that are identical between distantly related species, are enigmatic genomic features whose function is not well understood. First identified and characterized in mammals, UCEs have been proposed to play important roles in gene regulation, RNA processing, and maintaining genome integrity. However, because all of these functions can tolerate some sequence variation, their ultraconserved and ultraselected nature is not explained. We investigated whether there are highly conserved DNA elements without genic function in distantly related plant genomes. We compared the genomes of *Arabidopsis thaliana* and *Vitis vinifera*; species that diverged ~115 million years ago (Mya). We identified 36 highly conserved elements with at least 85% similarity that are longer than 55 bp. Interestingly, these elements exhibit properties similar to mammalian UCEs, such that we named them UCE-like elements (ULEs). ULEs are located in intergenic or intronic regions and are depleted from segmental duplications. Like UCEs, ULEs are under strong purifying selection, suggesting a functional role for these elements. As their mammalian counterparts, ULEs show a sharp drop of A+T content at their borders and are enriched close to genes encoding transcription factors and genes involved in development, the latter showing preferential expression in undifferentiated tissues. By comparing the genomes of *Brachypodium distachyon* and *Oryza sativa*, species that diverged ~50 Mya, we identified a different set of ULEs with similar properties in monocots. The identification of ULEs in plant genomes offers new opportunities to study their possible roles in genome function, integrity, and regulation.

[Supplemental material is available for this article.]

An increasing number of studies indicate that although the larger part of eukaryotic genomes consists of non-protein-coding DNA, this is far from being nonfunctional. Conserved noncoding sequences (CNSs) are found in large numbers in all animal genomes (Dermitzakis et al. 2002, 2004). CNSs are still conserved between humans and pufferfish, which diverged 450 million years ago (Mya) (Woolfe et al. 2005). Their average sequence identity varies depending on the genomes compared.

There are varying degrees of conservation of CNSs, with noncoding ultraconserved elements (ncUCEs) forming the extreme end of the distribution. UCEs were first identified as DNA stretches that are 100% identical between the mouse, rat, and human genomes over at least 200 bp (Bejerano et al. 2004). NcUCEs were mainly described among eutherian genomes, such as human, mouse, rat, dog, and cow (Bejerano et al. 2004; Stephen et al. 2008; Elgar 2009). Although most ncUCEs only appeared during tetrapod evolution (Stephen et al. 2008), many were already present in the jawed vertebrate ancestor, spanning ~530 Mya of evolutionary time; however, their conservation falls off to ~80% (Wang et al. 2009). Because we currently do not know any biological process that would not tolerate at least some sequence variation, the function of these ultraconserved and ultraselected elements is enigmatic.

The majority of the ncUCEs and CNSs seem to be under purifying selection, indicating that they are not mutation cold spots but are strongly constrained functional elements (Drake et al. 2006; Chen et al. 2007; Katzman et al. 2007). In insects, ncUCEs occur much less frequently and are smaller in size than the mammalian ones, thus being more similar to CNSs, which are often shorter and less conserved (Glazov et al. 2005).

In animals, ncUCEs and CNSs are enriched near specific functional groups of genes, e.g., encoding transcription factors and developmental regulators (Bejerano et al. 2004; Glazov et al. 2005; Vavouri et al. 2007). It was demonstrated that ncUCEs and CNSs can function as enhancers controlling tissue-specific gene expression (Woolfe et al. 2005; Pennacchio et al. 2006; Paparidis et al. 2007; Visel et al. 2008; McEwen et al. 2009). Nevertheless, their role as enhancers is not sufficient to explain their high conservation, because all protein–DNA, DNA–DNA, or DNA–RNA interactions known to date tolerate significant sequence divergence without affecting their functions (Ludwig et al. 2000, 2005; Romano and Wray 2003; Poulin et al. 2005; Rastegar et al. 2008). Therefore, ncUCEs and CNSs are likely to serve additional—so far unknown—functions that constrain their sequence.

Because ncUCEs are often single-copy sequences and strongly depleted from segmental duplications and human copy number

[5]Present address: Institute of Evolutionary Biology and Environmental Studies & Institute of Plant Biology, Zürich-Basel Plant Science Center, University Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland.
[6]Corresponding authors
E-mail wicker@botinst.uzh.ch
E-mail grossnik@botinst.uzh.ch
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.129346.111. Freely available online through the *Genome Research* Open Access option.

variants (Derti et al. 2006; Chiang et al. 2008), it was suggested that they could serve as genome integrity retention agents that act in a copy counting mechanism for chromosomes (Derti et al. 2006). In other words, ncUCEs in diploid cells should be present in exactly two copies to ensure genome integrity. To accurately assess their number, ncUCEs would have to be identical in sequence to avoid interactions with duplicated genomic regions. However, sequence retention and extreme conservation do not mean that they are essential for viability. In fact, deletion of four ncUCEs in the mouse did not cause obvious phenotypic abnormalities (Ahituv et al. 2007). Nonetheless, mutations in ncUCEs are deleterious over evolutionary time as evidenced by the fact that ncUCEs are under stronger selection than protein-coding regions (Katzman et al. 2007).

Until now little is known about the occurrence of CNSs in plant genomes. Most plant CNSs described to date are relatively small and reside close to genes. In monocots, apart from three exceptions (Bossolini et al. 2007; Wicker et al. 2008), most CNSs are short (average 20 bp), flanking a small number of orthologous genes (Kaplinsky et al. 2002; Guo and Moose 2003; Inada et al. 2003). A recent study describes the existence of long identical sequences (over 100 bp) between plant genomes; however, the reported sequences are part of regions of known function or origin, such as repeats, exons, or organellar DNA (Reneker et al. 2012).

Here, we focus on the identification of large UCE-like elements (ULEs) in dicot and monocot genomes. Special care was taken to ensure that ULEs are not part of any genic sequence with known function. By comparing the genome sequences of *Arabidopsis thaliana* (mouse-ear cress) and *Vitis vinifera* (grapevine), we identified 36 large and highly conserved ULEs, which are >55 bp long and share at least 85% sequence identity. The divergence time between the two species is estimated to be 115 Mya (Fawcett et al. 2009), allowing significant changes in DNA sequence to occur. Monocots have their own set of ULEs, and many are shared by the more closely related genomes of *Brachypodium distachyon* (purple false brome), *Oryza sativa* (rice), *Sorghum bicolor* (sorghum), and *Zea mays* (maize). Strikingly, despite a complete lack of sequence similarity between plant ULEs and animal ncUCEs, they share common properties, indicating that the evolutionary conservation of ULEs and ncUCEs may result from similar functional constraints and selective pressures in plants and animals.

Results

Identification of plant UCE-like elements (ULEs) between the *A. thaliana* and *V. vinifera* genomes

To identify ULEs in plants, whole-genome comparisons of the two dicot species *A. thaliana* and *V. vinifera* were performed. Among dicots with sequenced genomes *Vitis* is the most distantly related to *Arabidopsis*. The genome of *Arabidopsis* was used as an anchor for the ULE search against *Vitis*. We define a ULE as a noncoding DNA sequence sharing at least 85% identity. To exclude that these sequences serve as potential transcription factor–binding sites, we searched the *Arabidopsis* Gene Regulatory Information Server (AGRIS), a database for transcription factor binding sites (Palaniswamy et al. 2006). The average size of the 763,000 predicted *cis*-regulatory elements is 6.4 bp. Often, such transcription factor–binding sites are clustered, leading to larger conserved stretches (Davidson 2001). Using AGRIS, we found 28 large putative transcription factor–binding sites or clusters (>25 bp) with the biggest being 50 bp. Thus, we searched for ULEs that were longer than 55 bp.

The *Arabidopsis* genome was split into fragments of 1200 bp with a 600-bp sliding window and 600 bp overlap. These fragments were used in BLASTN searches against the *Vitis* genome. All conserved sequences >55 bp long with ≥85% similarity were investigated further, using a set of stringent criteria for the identification of ULEs (Table1): To exclude gene sequence motifs that may still have been present in this data set, candidate sequences were used in BLASTN searches against all *Arabidopsis* coding sequences. BLASTN searches were also carried out against collections of *Arabidopsis* tRNAs, ribosomal genes, and known ncRNAs. The remaining sequences were used in BLASTN searches against mitochondrial and chloroplast DNA. Transposable elements were excluded from our data set. The remaining candidates were used in BlastX searches against the nonredundant NCBI protein database to identify and eliminate any further protein-coding sequences that might not have been annotated in *Arabidopsis*. Finally, we removed conserved sequences overlapping intron–exon junctions because they might be part of alternative splicing products or wrongly annotated exons. To ensure that only ULEs of low copy number remained in our data set, candidates with >5 copies were removed.

In total, 36 candidate ULEs between the *Arabidopsis* and *Vitis* genomes met our criteria (Supplemental Table S1). The resulting ULEs reside in intergenic or intronic regions and all occur as single copies in the genome. We identified two paralogous elements, ULE27 and ULE28, which are found in tandem on chromosome 2. These ULEs are within 300 bp of each other. ULE27 is 2 bp longer than ULE28 but otherwise 100% identical. In *Vitis* ULE25 is found in two tandem copies within 250 bp on chromosome 4. One of the *Vitis* copies is 12 bp longer than the other, but the shared sequences are 100% identical.

The ULEs comprise a total of 2396 bp. The longest one is 105 bp, and sequence identity ranges from 85% to 98%, with an average of 87.7%. Twenty-two were found in intergenic regions and 14 in introns. All ULEs were screened against *Arabidopsis* ESTs and novel transcripts detected after exosome depletion (Chekanova et al. 2007). For 28/36 ULEs there was no evidence of transcription, while the remaining eight were at least partially covered by transcripts. The distribution of ULEs along the five *Arabidopsis* chromosomes is shown in Figure 1.

ULEs are conserved among dicot but not more distantly related genomes

We tested whether the identified ULEs are present in other eudicot genomes, namely, *Populus trichocarpa* (poplar), *Carica papaya* (papaya), *Cucumis sativus* (cucumber), and *Arabidopsis lyrata* (lyre-leaved rock-cress) (Supplemental Table S2; Tuskan et al. 2006; Ming et al. 2008; Huang et al. 2009). The phylogenetic relationships of these species are shown in Figure 2. Twenty-two ULEs (22/36) were also present in the poplar genome, with similarities ranging from 83% to 98%, and a similar average identity as between *Arabidopsis*

Table 1. Criteria for ULE identification

ULEs are	ULEs are not
1. >55 bp long	1. Coding sequences
2. ≥85% identity	2. tRNA, rRNA, ncRNA
3. Low copy number (≤5)	3. mtDNA, chlDNA
	4. Transposable elements
	5. *E. coli* contamination
	6. Encoding a protein motif
	7. In intron–exon junctions

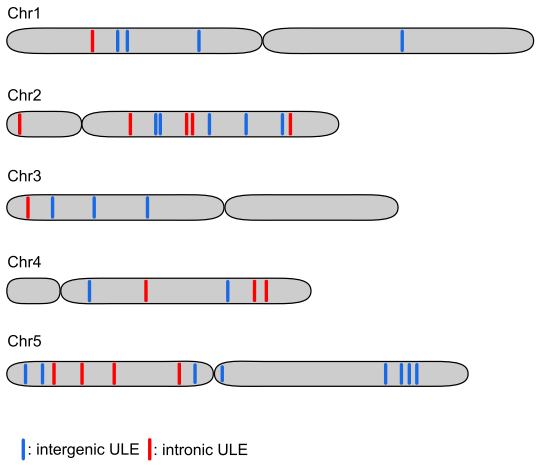

Figure 1. Distribution of ULEs along *Arabidopsis* chromosomes. (Blue lines) Intergenic ULEs; (red lines) intronic ULEs. ULEs of both types are found on all chromosomes: On chromosome 1, ULEs are found on average every 6 Mb, whereas on chromosomes 2, 3, 4, and 5, ULEs are found on average every 1.9 Mb, 5.8 Mb, 3.8 Mb, and 2.2 Mb, respectively.

and *Vitis*. High levels of conservation were also found within the less complete genome of papaya, where 20 ULEs have identities ranging from 84% to 100%.

Only nine of 36 ULEs were found in the cucumber genome with identities ranging from 85% to 98%. All but one of these corresponded to intronic ULEs, which suggests that scaffold data from cucumber are good enough for comparisons of genes but intergenic regions are not. Also, the genomes of poplar and papaya are less complete than the *Arabidopsis* genome, which may explain why not all ULEs were found. To test this, we investigated whether genes neighboring the ULEs that are not present in poplar or/and papaya are also absent from those genomes. Indeed, the closest gene to these ULEs was not found or only partially present (less than a third of the corresponding sequence) in either the poplar or papaya genome (Supplemental Table S3). Finally, we looked for ULEs in the sequenced genome of another member of the mustard family, *A. lyrata* (Hu et al. 2011), where all but one ULE were conserved with identities between 93% and 100%.

We also searched for the 36 ULEs in the genomes of rice but found only one (ULE3) with 89% identity. ULE3 was partially conserved in two other monocot genomes, *Brachypodium* and maize (Supplemental Table S2). ULE3 is located upstream of gene *At2g33440*, which encodes an RNA-binding domain and is expressed at different developmental stages, but is functionally uncharacterized. None of the remaining ULEs, except for ULE19 in *Brachypodium,* were conserved. The identified ULEs were also searched against the genomes of the moss *Physcomitrella patens* and the green alga *Chlamydomonas reinhardtii*, but no ULEs were conserved.

ULEs are mostly found in conserved collinear positions

To examine genes and other genic features that neighbor ULEs in *Arabidopsis* and *Vitis*, a genomic region spanning 3 kb from the 5′ and 3′ ends of each ULE was analyzed. These 6-kb windows were used in BLASTN searches against the coding sequences of the two genomes. Among the 22 intergenic ULEs, 15 were located upstream of genes, three downstream from genes, and four in genomic regions where the nearest gene is >2 kb away.

To further assess ULE organization, we used the same 6-kb window and compared it by dotplot with an equivalent window in *Vitis*. To study whether ULEs are located in collinear regions, we classified flanking regions of ULEs as collinear when at least one of the neighboring genes was homologous. We found that 29/36 ULEs were found in collinear regions (see example in Fig. 3A). Interestingly, seven ULEs were found in noncollinear regions, where exclusively the ULE was conserved in the 6-kb segment, indicating that ULEs can be independent elements not necessarily associated with nearby genes (see example in Fig. 3B). For these seven noncollinear ULEs, we examined a larger region (50 kb) between *Arabidopsis* and *Vitis*: in five cases only the ULE was conserved (ULE2, ULE5, ULE9, ULE35, ULE36). It is intriguing that some of the ULEs are not found in collinear regions relative to *Vitis*, since in animals, UCEs remain in collinear positions. One possible explanation for the noncollinear ULEs is that transposable element (TE) activity can lead to movement of genes and other sequences, thereby eroding collinearity (Wicker et al. 2010). Indeed, transposed genes in *Arabidopsis* are often associated with flanking repeats (Woodhouse et al. 2010), and this is also the case for three of the noncollinear ULEs (ULE2, ULE5, ULE35), which contain repeats within 3 kb of their borders (http://epigara.biologie.ens.fr/cgi-bin/gbrowse/a2e). Whether these repeats were associated with the movement of the ULEs or inserted afterward cannot easily be distinguished.

ULEs are flanked by a sharp drop of the A+T content

To investigate whether ULEs have specific sequence characteristics, we compared the base composition at the boundaries of the ULEs, which are not conserved, with the one inside the ULEs (Fig. 4), as it was done for highly conserved noncoding sequences in vertebrates (Walter et al. 2005). We analyzed ULEs and their flanking regions in three blocks of sequences: 400 bp of flanking sequence plus 10 bp of the corresponding end of each ULE at the 5′ and 3′ borders, and 30 bp from the middle of each ULE. We calculated the A+T content for each of these three blocks and observed a sharp drop in

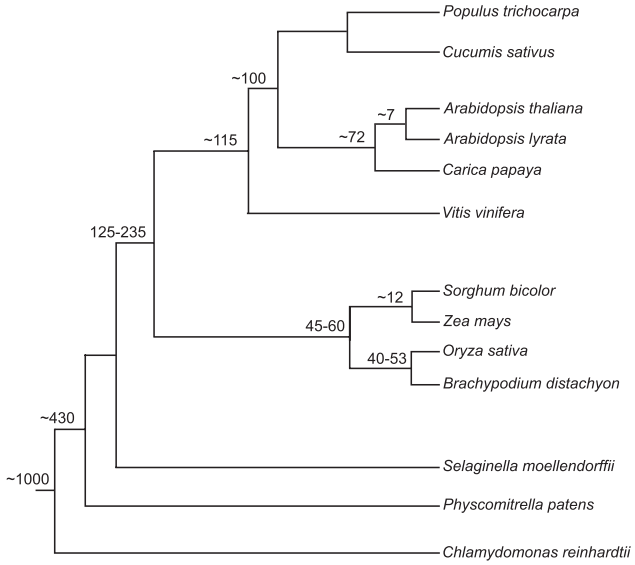

Figure 2. Phylogenetic relationships between major sequenced plant genomes. The phylogenetic tree is adapted from phytozome.org. Divergence distances in million years ago (Mya) are indicated beside the nodes and are taken from Stewart and Rothwell (1993), Yang et al. (1999), Davies et al. (2004), Kuittinen et al. (2004), Swigoňová et al. (2004), Yoon et al. (2004), Tuskan at al. (2006), Fawcett et al. (2009), and International Brachypodium Initiative (2010).

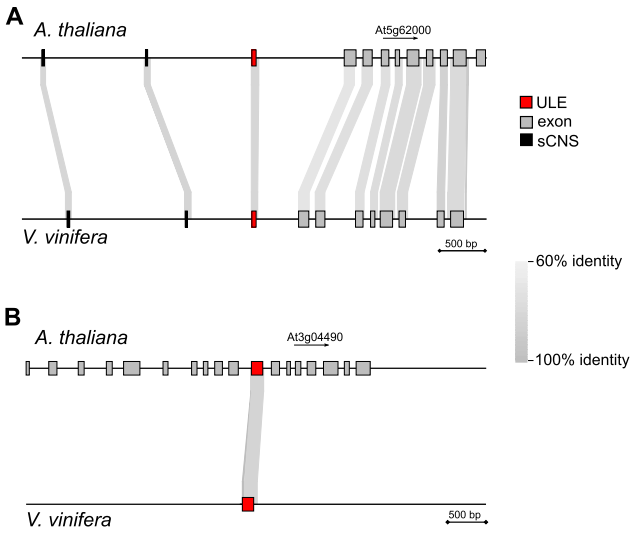

Figure 3. Comparison of a 6-kb region surrounding two selected ULEs in *Arabidopsis* and *Vitis*. Conserved regions are indicated by shaded areas. (Red) ULEs; (black) small conserved noncoding sequences (sCNSs) below 30 bp; (gray) exons. (*A*) Comparison in collinear regions between *Arabidopsis* and *Vitis*. (*B*) Comparison between noncollinear regions between *Arabidopsis* and *Vitis*.

A+T frequency starting just before the borders of the ULEs. Within the ULEs, the A+T content was lower than in flanking regions (Fig. 4A). The same was observed when we analyzed the A+T frequency of each ULE individually (data not shown). We calculated the average A+T content in the *Arabidopsis* genome to be 63%, which is the same as the average A+T content in the regions flanking the ULEs (63%). In contrast, the average A+T content of the ULEs is 57% (Fig. 4A) and differs significantly from the A+T content of the sequences flanking the ULEs (0.57 vs. 0.63, $P = 0.00104$ by paired Wilcoxon signed-rank test). Thus, there is a sharp drop in A+T content at the borders of the *Arabidopsis* ULEs.

In *Vitis* we also observed a sharp drop of the A+T content at the ULE borders (Supplemental Fig. S1). The average A+T content of the *Vitis* genome is 65%, while the ULEs have an average A+T content of 57%. As in *Arabidopsis*, the A+T content of *Vitis* ULEs is significantly lower than that of the flanking sequences, which is 61% (0.57 vs. 0.61, $P = 0.0103$ by paired Wilcoxon signed-rank test).

ULEs are associated with specific functional categories of genes

We investigated whether ULEs are clustered near genes of distinct biological or molecular function. We examined the Gene Ontology (GO) annotations of genes flanking intergenic ULEs. For intronic ULEs, we considered only those genes in which they were located. ULE-flanking genes showed a significant enrichment for genes involved in development ($P \leq 2.2 \times 10^{-16}$). They also showed significant functional enrichment for genes associated with transcription factor activity ($P = 1.99 \times 10^{-5}$) and nucleic acid binding activity ($P = 3.5 \times 10^{-7}$) (Fig. 5).

ULE-associated genes exhibit expression peaks in undifferentiated tissues

Our GO analysis indicated that genes associated with ULEs are involved in development and are, in turn, likely to be developmentally regulated, too. To test this hypothesis, we estimated gene expression signals from a large collection of Affymetrix ATH1-array data querying a total of 103 different tissue and cell types of *Arabidopsis* (for details, see Supplemental Table S4; Supplemental Methods). From 54 ULE-associated genes, 41 are targeted by probe sets present on the ATH1 array. We visualized the average expression of ULE-associated genes across different developmental stages, tissues, and cell types. As shown in Figure 6A, several genes exhibit elevated expression levels in gametophytes, embryos, and meristems. Figure 6B summarizes the number of expression peaks found in distinct tissues of this developmental atlas. Because of small sample numbers, however, we could not test whether this increase is statistically significant. Furthermore, the data set consists of a heterogeneous pool of data from different laboratories, tissue origins, and preparation protocols. Therefore, we classified tissues and cell types into four categories according to their differentiation state from (1) mainly undifferentiated cell populations to (4) mostly fully differentiated cell populations (Supplemental Table S4) and found that arrays from cell populations consisting of mainly undifferentiated cells (i.e., gametes and cells from the shoot meristem, early embryo, and endosperm stages, as well as the root quiescent center) showed a significantly increased number of expression peaks (see Fig. 6C; observed: 20, expected 12.4, *P*-value from randomly resampling 100,000 gene sets: $P = 0.00939$). From these results, we estimate that ~50% of ULE-associated genes show highest expression in an undifferentiated cell type. However, low expression of ULE-associated genes in other cell types does not necessarily exclude the importance of gene activity in these tissues. Overall, these results suggest that ULE-associated genes are developmentally regulated in plants and are often highly expressed in reproductive tissues.

ULEs are depleted from segmental duplications

The fact that ULEs are single copy in the genome may indicate that multiple copies may be deleterious, possibly because that would interfere with the proposed copy counting mechanism (Derti et al. 2006). We searched whether ULEs are depleted from segmental duplications (SDs). During evolution, *Arabidopsis* has undergone multiple whole-genome and large-scale duplication events. We took into account SDs identified in *Arabidopsis* by Blanc et al. (2003), i.e., chromosome regions that share similar genes in the same order, excluding genes duplicated in tandem and transposable elements. In this survey, 108 blocks of SDs sharing six or more duplicated genes were identified, which cover 71% of the *Arabidopsis* genome (80 Mb). The more recent duplications are estimated to have occurred 24–40 Mya (Blanc et al. 2003). Since these SDs refer to coding regions, we considered ULEs to be in segmental duplications when the closest genes to intergenic ULEs or genes containing intronic ULEs were within segmental duplications.

All intronic ULEs and, with the exception of one (ULE8 flanking *At2g15510*), all genes neighboring intergenic ULEs were outside SDs. To investigate the statistical significance of the identified trend for ULEs, a permutation test was applied in which 1000 randomized data sets were sampled. Our test shows that the absence of ULEs from SDs is clearly nonrandom ($P < 0.00036$). The depletion of ULEs from SDs indicates that they are dosage sensitive and that there are selective constraints to keep them single copy.

ULEs are under purifying selection and not mutational cold spots

The high sequence conservation of ULEs between *Arabidopsis* and *Vitis* indicates that ULEs are selectively constrained sequences.

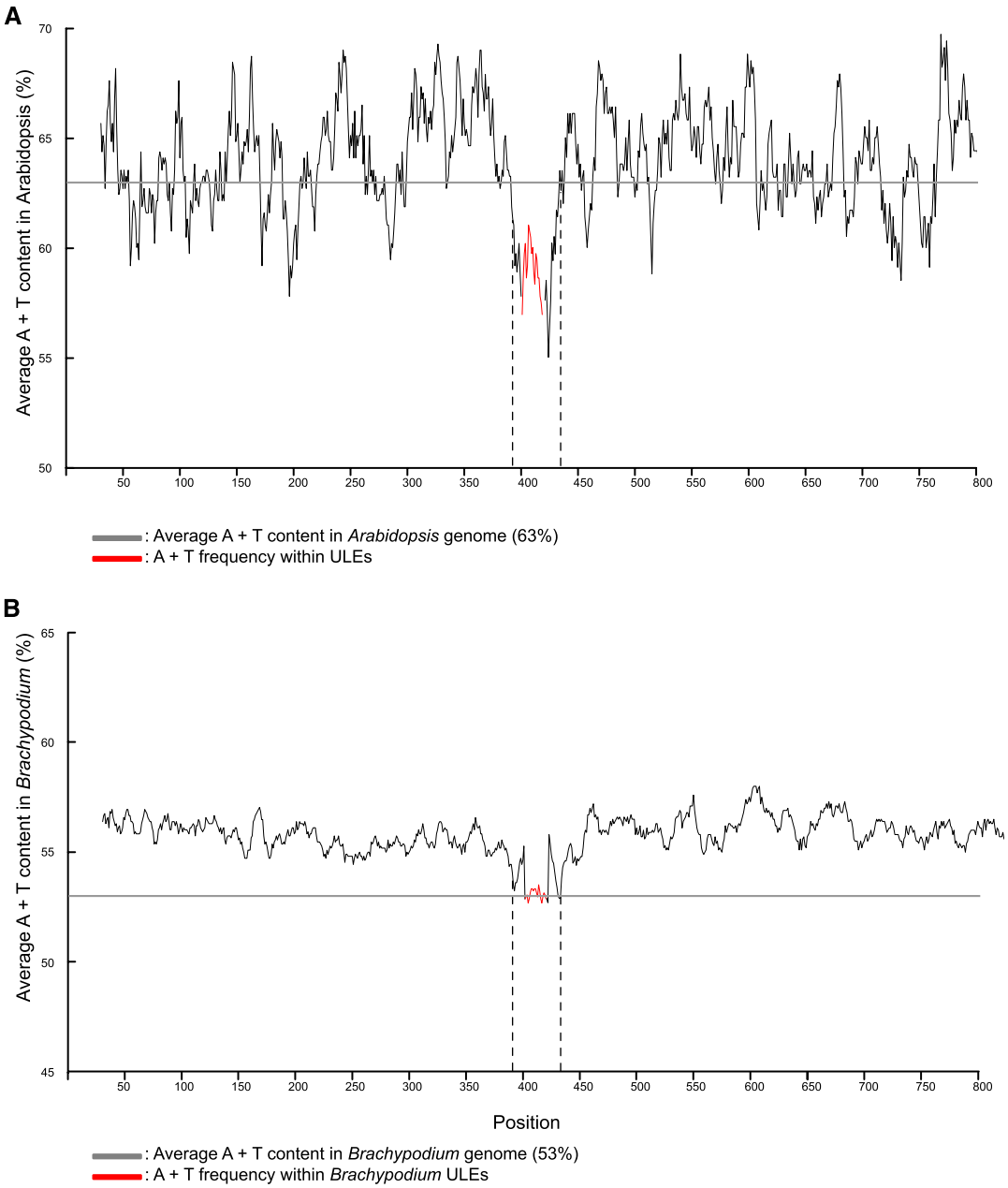

Figure 4. A+T content distribution within ULEs and their flanking regions. (Red) A+T frequency within ULEs; (black) the frequency of flanking regions; (gray line) the average A+T content of the respective genome; (dashed vertical lines) the last nucleotide of the neighbor regions before ULEs. (*A*) A+T frequency in *Arabidopsis* ULEs (34/36). (*B*) A+T frequency in *Brachypodium* ULEs (869/870) present in the genomes of rice, sorghum, and maize, and their flanking regions.

Alternatively, their conservation could be due to the fact that they lie in regions with low mutation rates. To address this question, we estimated the distribution of selection coefficients from polymorphism data on the ULEs in 83 resequenced *Arabidopsis* accessions (Fig. 7; Supplemental Tables S5, S6). The strength of selection acting on ULEs was compared relative to protein-coding regions and ULE-flanking regions (500 bp from the borders), respectively. Using the derived allele frequency (DAF) spectrum, we fit a Bayesian hierarchical model to estimate mean selection coefficients for each class of site. The hierarchical model was fit using a Markov chain Monte Carlo (MCMC) while controlling for the effect of ascertainment on the ULE sites (Katzman et al. 2007; Kern 2009). The potency of removal of deleterious alleles increases as the selection coefficient decreases. Posterior estimates of mean selection coefficients between classes of sites indicate that ULEs may be under slightly stronger purifying selection than ULE flanking sites or exons; however, because the credible sets overlap, such a difference is not statistically significant, only consistent with the hypothesis that ULEs might be under stronger purifying selection. This demonstrates that purifying selection rather than reduced mutation rates preserve ULEs at the DNA level. Thus, ULEs are under evolutionary pressure, which suggests that they are, indeed, functional elements.

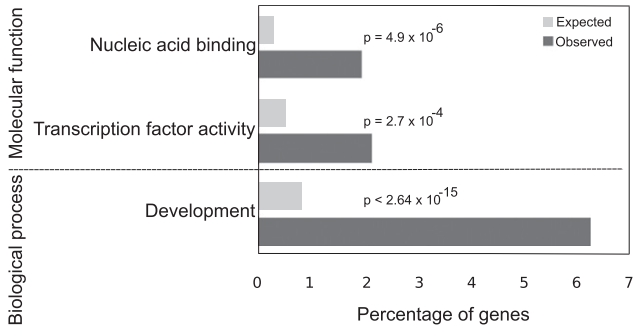

Figure 5. Expected versus observed percentage of genes in Gene Ontology annotation under the Molecular Function and Biological Process categories, corrected for multiple testing (Bonferroni correction).

ULEs are not associated with recombination hot spots or origins of replication, nor are they modified by DNA methylation

The observation that A+T content drops at the borders of the ULEs is intriguing because it implies a structural basis of these elements. Various cellular processes may be influenced by the A+T content, including recombination and replication. Indeed, it was shown that sequences with many ATs and TAs have lower recombination rates than those containing AGs, TCs, CAs, and TGs (Guo et al. 2009). Thus, we explored the possibility that ULEs are enriched at recombination hot spots (RHSs). RHSs are DNA regions with a higher rate of meiotic crossing-over than the surrounding DNA. In *Arabidopsis*, studies in dense SNP regions from a sample of 19 accessions revealed around 260 RHSs, which tend to occur in intergenic regions and are 1–2 kb long (Kim et al. 2007). None of the ULEs overlapped these RHSs. However, permutation test of 1000 randomized data sets showed that the absence of ULEs from RHSs is not significant.

Furthermore, we investigated whether ULEs are part of origins of DNA replication. Recently, ~1500 putative origins of replication were mapped in *Arabidopsis* at a genome-wide scale (Costas et al. 2011). In this study, next-generation sequencing was used to map newly synthesized DNA at the G_1/S transition using synchronized cells. Only three ULEs are located within mapped origins of replication, the two tandem ULEs on chromosome 2 (ULE27, ULE28) and ULE33 on chromosome 1. However, this depletion of ULEs from origins of replication is not significant after applying a permutation test.

DNA methylation of cytosines is involved in epigenetic regulation. This epigenetic mark is heritably transmitted to following generations and affects various processes such as gene expression, genomic imprinting, transposon silencing, and timing of replication (for review, see Vanyushin and Ashapkin 2011). In *Arabidopsis*, the methylome at a single-base-pair resolution has been assessed in DNA of 5-wk-old plants and flower buds, respectively (Cokus et al. 2008; Lister et al. 2008). The methylome of flower buds identified more than 2 million methylated cytosines accounting for 5.26% of genomic cytosines (Lister et al. 2008). We used this single-base-pair resolution DNA methylation map to investigate the methylation pattern of ULEs. The majority of the ULEs do not have any detectable methylation marks. Only seven ULEs are methylated in either the CG, CHH, or CHG context (Supplemental Table S7). However, after applying a permutation test, the lack of ULE methylation is not statistically significant.

A distinct set of ULEs is shared between monocot genomes

Surprisingly, *Arabidopsis* ULEs were not present in genomes that are more distantly related than those of dicots. Thus, we asked whether there is another set of ULEs found explicitly in monocot genomes. We compared the genome of *Brachypodium* against that of *Oryza sativa* sb *japonica*. Divergence time between the two species is estimated at 40–53 Mya (International Brachypodium Initiative 2010), which is less than between *Arabidopsis* and *Vitis* (~115 Mya). We applied the same criteria as before (Table 1) and found 4572 *Brachypodium* ULEs that are at least 85% identical to rice and >55 bp long. The median size and identity of these sequences are 69 bp and 87%, respectively, similar to the ones found in dicots. Like the *Arabidopsis* ULEs, the majority of *Brachypodium* ULEs are single copy in the genome (4491 out of 4572). Interestingly, 870 sequences are also shared in the maize and sorghum genomes, which reflects conservation over 50 Mya (Fig. 2; Supplemental Table S8; International Brachypodium Initiative 2010).

We tested whether, apart from being single copy, the *Brachypodium* ULEs share other properties with *Arabidopsis* ULEs. Similarly, we calculated the A+T composition of these sequences relative to their flanking regions, which show no conservation (Fig. 4B). The average A+T content of the 870 *Brachypodium* ULEs shared with other monocots is 53%, which is identical to the average A+T content of the *Brachypodium* genome (53%), but differs significantly from that of the sequences flanking the ULEs, which is 55% (0.53 vs. 0.55, $P = 0.000351$ by paired Wilcoxon signed-rank test). Surprisingly, a similar drop of A+T composition at the borders of ULEs is present in both dicots and monocots. It is clear that *Brachypodium* and *Arabidopsis* ULEs, although distinct in sequence, share common characteristics.

Human UCEs are more abundant than *Arabidopsis* ULEs even when filtered under stricter criteria

Our results indicate that in plants ULEs are less common than UCEs are in mammalian genomes. However, in our study, we used filter criteria that were more stringent compared with those used in mammalian studies. Thus, there might be fewer mammalian UCEs had they been analyzed under our criteria. To address this question, we reanalyzed the 481 UCEs identified by Bejerano et al. (2004). In our analysis, we excluded UCEs within protein-coding sequences and functional ncRNAs and removed mitochondrial or *E. coli* sequences. In total, 390 elements, of 100% identity and length ≥200 bp, meet the criteria we used, indicating that, even under these stringent criteria, mammalian ncUCEs are more abundant than plant ULEs.

Discussion

In this study, we sought to identify and characterize highly conserved, noncoding elements in plant genomes. Synteny studies in flowering plant genomes revealed that the *Arabidopsis* genome is the most reshuffled, whereas the grapevine and papaya genomes have a better conserved ancestral genome structure (Huang et al. 2009). Thus, any conserved sequence between *Vitis* and *Arabidopsis* suggests a functional role. We focused on long stretches of conserved DNA (>55 bp), not necessarily associated with genes, in an unbiased search. In addition, our study used particularly stringent criteria in order to avoid any overlap with known genic sequences. Moreover, we were only interested in ULEs found at low copy number in the genome, thus targeting elements with a possible dosage effect that is prohibitive to accumulating high copy numbers.

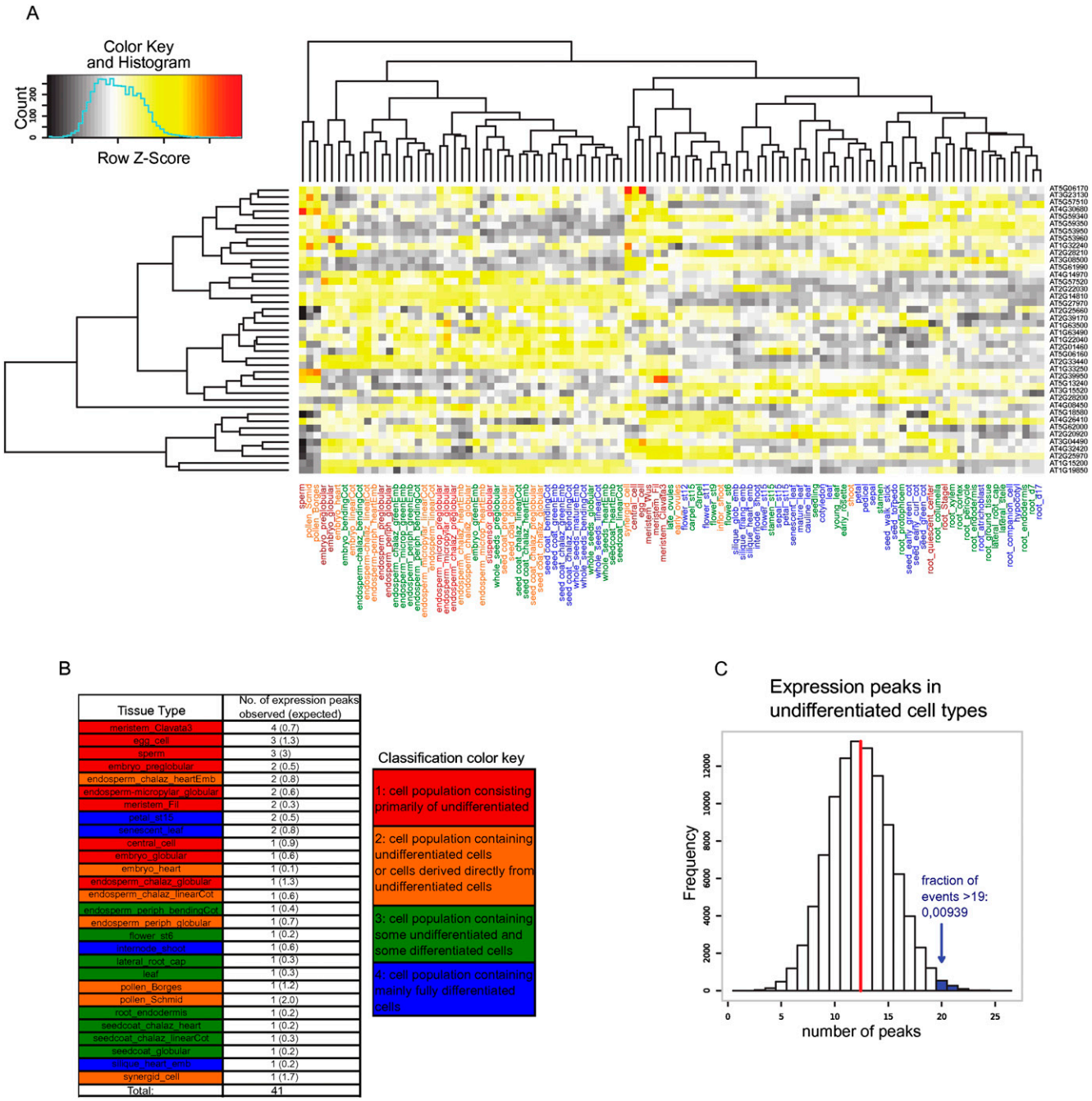

Figure 6. ULE-associated genes are developmentally regulated. (*A*) Heatmap representing color-coded relative expression among a large collection of *Arabidopsis* tissues/cell types. (Dark colors) Low expression; (bright colors) high expression. Expression values were scaled per row (i.e., per gene) to visualize expression peaks of a transcript across developmental stages. Per-gene as well per-tissue clustering was applied to visualize patterns in the expression profiles. Sample descriptions are color-coded as described in *B*. Only ULE-associated transcripts represented on the ATH1 array are shown (41/56). (*B*) Table of tissues in which expression peaks of ULE-associated genes occur. The color code indicates the differentiation state of the respective tissue/cell type. (*C*) ULE-associated gene expression peaks are significantly enriched in undifferentiated cells. The number of events where the maximal mean expression signal for one of 41 ULE-associated genes was found in undifferentiated cell types (i.e., gametes, shoot meristem cells, root quiescent center, and early embryo/endosperm) is significantly higher than expected by chance. The histogram depicts the frequencies of expression peaks occurring in undifferentiated cell types among groups of 41 genes randomly sampled from the whole array. Resampling of random groups indicated that the same or higher number of expression peaks in undifferentiated cell types occurs only in 939 out of 100,000 instances ($P = 0.00939$).

Plant ULEs are fewer and less conserved than mammalian UCEs

One striking result from our comparative studies is the relatively low number of ULEs found in dicot genomes compared with sets of UCEs reported in mammals. If only the 390 mammalian ncUCEs that passed our filtering criteria are considered, the frequencies of these elements lie in the same range, i.e., one ncUCE/ULE per 8.0 Mb, 3.3 Mb, and 13.5 Mb in the human, *Arabidopsis*, and *Vitis*

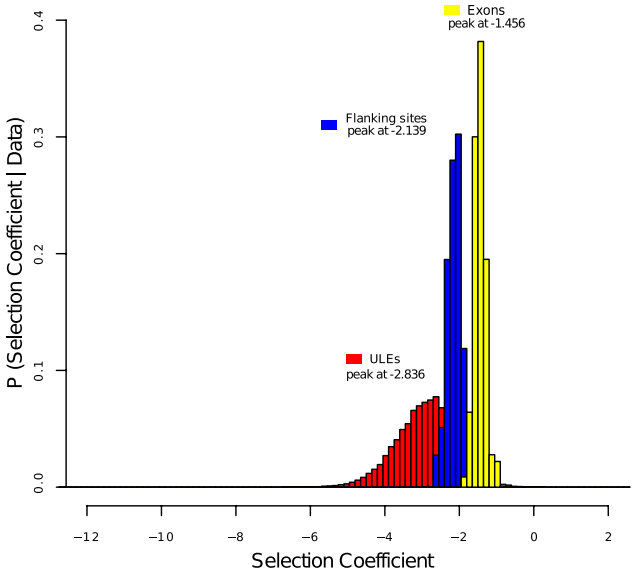

Figure 7. Selection coefficients for genomic regions. Shown are the posterior distributions of mean selection coefficients across classes of sites in the *Arabidopsis* genome. The values shown are α, the mean population scaled selection coefficient (2Nes). The values given are the maximum a posteriori (MAP) estimates from our MCMC (Supplemental Figs. S2, S3).

genomes, respectively. However, the ncUCEs identified by Bejerano et al. (2004) represent only the tip of the iceberg of the total number of highly conserved sequences. There are several thousands of UCEs (13,736) at least 100 bp long that are shared between human and placental mammals (Stephen et al. 2008). In addition, there is a large number of conserved, noncoding elements that are slightly less than 100% identical (Dermitzakis et al. 2002; Woolfe et al. 2005). Thus, it appears that conserved noncoding sequences are more abundant in animals than in plants, perhaps because in animal genomes, gene order is retained over millions of years (Li et al. 2010). More ULEs are lying in plant genomes when the genome comparison is made among less evolutionary distant plant species, such as the 870 we found shared by monocot genomes, with frequencies of one ULE per 0.4 Mb, 0.5 Mb, 0.9 Mb, and 2.6 Mb in *Brachypodium*, rice, sorghum, and maize, respectively. In fact, in contrast to dicot plants, monocots show a substantial conservation of gene order (International Brachypodium Initiative 2010).

The vast majority of the identified *Arabidopsis* ULEs arose after the divergence of dicots and monocots. However, *Arabidopsis* ULEs are well conserved in other dicot genomes, such as those of poplar, papaya, cucumber, and *A. lyrata*, but between monocots and dicots only one ULE was retained. This is in sharp contrast to mammalian UCEs, where a major proportion covers an evolutionary time of ~530 Mya (Wang et al. 2009).

Why do plant genomes appear to contain fewer ULEs? One reason could be that plants and vertebrates have molecular clocks running at different speeds. It has been suggested that *Arabidopsis* has a faster molecular clock relative to other angiosperms (Paterson et al. 2010), whereas amniote evolution was accompanied by a slowdown in the molecular clock (Stephen et al. 2008). Thus, it is possible that plant ULEs evolved at a higher rate due to a faster molecular clock. This could also explain why ULEs are not conserved between monocots and dicots, because they might have diverged beyond recognition. Alternatively, our set of ULEs may represent distinct dicot and monocot innovations.

Plant genomes have the tendency to reorganize frequently, for example, by undergoing whole-genome duplications (Masterson 1994). This could also contribute to the smaller number of ULEs because genome duplication events might have relaxed the selective constraints on ULEs, allowing them to evolve faster.

ULEs from plants and animals have similar characteristics

While CNSs have been characterized in various plant genomes, only a few studies focused on ultraconserved sequences, which represent an extreme case of conservation. A bioinformatics comparison between the *Arabidopsis* and rice genome identified 25 ultraconserved sequences that were longer than 100 bp, the longest being 1491 bp (Zheng and Zhang 2008). A more detailed examination of these sequences, however, showed that most were identical to segments of mitochondrial DNA, such that horizontal gene transfer—be it artifactual or biological—provides a likely explanation for their unltraconservation (Freeling and Subramaniam 2009). A very recent study also identified a large number of highly conserved elements between sequenced plant and animal genomes but came to the conclusion that there are no sequences similar to mammalian UCEs in plants (Reneker et al. 2012). However, they did not filter out certain sequence classes, such as organellar DNA, rDNA, and *E. coli* contamination, as we did in our search for ULEs. Furthermore, their criteria were quite different from ours, and thus they could not identify the ULEs we report here. Although plant ULEs and mammalian ncUCEs are distinct sets of conserved sequences, they share a surprising number of common properties. Dicot ULEs and mammalian ncUCEs (Katzman et al. 2007) are under strong purifying selection. New alleles arising within ULEs may therefore be deleterious, making it unlikely that they become fixed in a population; hence, their astounding sequence conservation.

We found that the A+T frequency is low at the borders of plant ULEs. The same feature is also shared among vertebrate and nematode conserved sequences (Walter et al. 2005; Vavouri et al. 2007; Chiang et al. 2008). The fact that the drop in A+T content at the borders of ULEs and ncUCEs is a conserved feature between animal and plant genomes indicates that their function may have a structural basis. The A+T content can affect DNA topology, nucleosome positioning, and higher-order chromatin organization (Segal et al. 2006; Hughes and Rando 2009), and influence DNA replication, repair, and recombination. However, ULEs do not appear to correlate with functional elements related to the structural features we tested and are not enriched in RHSs, origins of replication, or regions of DNA methylation. Like in ncUCEs from vertebrates and insects, the majority of dicot ULEs described in this study are found in the vicinity of genes involved in development and near genes whose molecular function is assigned to transcription factor activity. In addition, the majority of genes neighboring ULEs show strong expression in undifferentiated cells.

Based on the common properties between ULEs and mammalian ncUCEs, it is tempting to speculate that both sets of conserved sequences represent convergent evolutionary products that may be involved in the regulation of developmental genes. This is further supported by functional assays of ncUCEs showing that they act as enhancers during early embryo development in lamprey and mouse (Pennacchio et al. 2006; Visel et al. 2008). But why then are they so highly conserved? Enhancers usually do not require a high degree of sequence conservation (Stormo 2000), nor

are they unusually large, even if clustered. Recent findings suggest that ncUCEs might have dual or even more functions, since part of the human ncUCEs are both transcribed and act as enhancers (Licastro et al. 2010). Except for enhancers, ULEs could potentially represent part of conserved *cis*-regulatory modules (CRMs), where one or more transcription factors bind to regulate the expression of neighbor genes. About 18,500 CRMs located upstream of genes are shared by *Arabidopsis* and poplar (Ding et al. 2012). Merely one (ULE6) out of 13 intergenic ULEs tested is part of a such a CRM. In addition, in vertebrates, a proportion of conserved noncoding elements (ncUCEs and CNSs) do not share common target genes in all six genomes tested (Sun et al. 2008). This finding suggests that mere *cis*-regulatory activity is unlikely to be the only explanation for the existence and high conservation of these elements.

Strikingly, ULEs are depleted from SDs in *Arabidopsis*, similarly to what was reported for mammalian ncUCEs (Derti et al. 2006). However, the existence of ULEs predates the existence of the segmental duplications we investigated in our analyses. This advocates that an evolutionary force kept the ULEs as single copies even though segmental duplications cover >70% of the *Arabidopsis* genome (Blanc et al. 2003). These observations suggest that either ULEs per se, or the genomic regions that contain them, are dosage sensitive and that a deviation from single copy could have an impact on the plant's fitness. These results also support the idea that ULEs function as agents involved in a chromosome copy-counting mechanism (Derti et al. 2006). Here the maternal and paternal copies of ULEs/ncUCEs may recognize each other, perhaps through pairing, in order to determine the exact copy number of chromosomes, which in a diploid cell should be exactly two. Deviation from ULE/ncUCE copy number or sequence could trigger events that are deleterious to a cell with an abnormal number of chromosomes, but deleterious effects could also occur at the organismal or population level.

Despite the recent efforts to elucidate the function of conserved noncoding sequences, their role still remains elusive. ULEs have distinct characteristics and our data suggest that, in addition to sequence constraints, they are functional elements that are under purifying selection. Future studies are needed to shed light onto the purpose of their existence and their function.

Methods

Sequence analyses

All analyses were performed on LINUX systems. For the identification of ULEs, we developed software designed in Perl; all scripts are available upon request. Stand-alone BLAST software was obtained from NCBI (ncbi.nlm.nih.gov). For genome comparison studies, local BLAST databases were created. The *A. thaliana* genome sequence was downloaded from The Institute of Genomic Research (TIGR), now available from TAIR (arabidopsis.org). The genome of grapevine was obtained from Genoscope version1 (genoscope.cns.fr/externe/GenomeBrowser/Vitis), poplar version 1.1, *P. patens* version 1.1, and *C. reinhardtii* version 4.0 from the Joint Genome Institute (JGI) (jgi.doe.gov), *Oryza sativa* sb. *japonica* (rice) version 6 (rice.plantbiology.msu.edu/), *B. distachyon* from brachypodium.org, maize version 2 from plantGDB.org, papaya version 4 from phytozome.net/papaya.php, and cucumber scaffold data from cucumber.genomics.org.cn/page/cucumber/index.jsp. The sequence data from *A. lyrata* were produced by JGI in collaboration with the user community.

Coding, mitochondrial, and chloroplast sequences of *A. thaliana* were obtained from arabidopsis.org, TAIR9, ncRNA sequences from NCBI (ncbi.nlm.nih.gov, 08/1/2010), and PMRD: Plant microRNA database (Zhang et al. 2010). Transposable elements from the TREP database (wheat.pw.usda.gov/ITMI/Repeats), as well as repeats from the Plant Repeat Databases (plantrepeats.plantbiology.msu.edu/index.html). To identify possible *E. coli* contaminations, candidates were used in BLASTN searches against the *E. coli* genome version NC 000313. The number of conserved sequences that were culled after applying the above filters is shown on Supplemental Table S9A.

Similar filters were applied for the identification of monocot ULEs (Supplemental Table S9B). *Brachypodium dystachion* (version 1.0) was used in BLASTN searches against the *Oryza sativa* sb *japonica* genome (version 6.0). Databases from brachypodium.org, phytozome.org, rice.plantbiology.msu.edu, and plantgdb.org were used to discard candidates showing similarity to coding sequences, proteins, chloroplast, and mitochondrial DNA. For repetitive elements, PTREP and Plant Repeat Database were used. For small RNAs, PMRD, Cereal small RNAs database (http://sundarlab.ucdavis.edu/smrnas/), plant snoRNA database (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna), and NCBI were used.

For annotation of the *Vitis* genes surrounding ULEs, the two regions were aligned using DOTTER (Sonnhammer and Durbin 1995) to determine positions of introns, exons, and start and stop codons.

Characterization of ULEs

The A+T composition was calculated in a 10-bp window with a 1-bp sliding step width in each of the three sequence blocks. Two ULEs found in two closely spaced tandem copies were excluded from this analysis. For a comparison of A+T content of ULEs and flanking regions, ULEs were compared with sequences composed of one-half of the length of the ULEs flanking their 3′ and 5′ borders, respectively. A paired Wilcoxon signed-rank test was applied to assess significance.

For functional categorization, 54 genes located in flanking regions of intergenic ULEs or genes enclosing intronic ULEs were selected. Genes were grouped into different functional categories by using the TAIR9 Gene Ontology (GO) annotation. The data were compared with the functional categories assigned for all TAIR9 *Arabidopsis* genes. Fisher's exact tests were performed to determine over-representation of gene categories. *P*-values were corrected for multiple testing with the Bonferroni correction.

Expression analysis of ULE-associated genes

Original ATH1-array data from different *Arabidopsis* tissues were used as described (Wuest et al. 2010). Additional data sets were downloaded from public repositories (Supplemental Table S4). Data from the root quiescent center (Nawy et al. 2005), discrete seed compartments (Le et al. 2010), and cell types of the shoot apical meristem (Yadav et al. 2009) were added to the tissue atlas. The tissue data totally includes a set of 103 tissue types of gametophytic, sporophytic, and embryonic origin. Gene expression signals were calculated by dChIP (Version 2010) using invariant-set normalization and a PM-only model. Probe-set definitions according a newer *Arabidopsis* genome release (TAIR9) were downloaded from brainarray.mbni.med.umich.edu; ATH1-version 10, based on TAIR9 genomic sequences (Dai et al. 2005) and probes mapping to multiple probe sets were removed from the analysis. For this, duplicated probe sequences in the probe-set definitions were identified in R (Version 2.8.1) and a new chip description file generated using the Bioconductor package affxparser (Bengtsson et al. 2010) (bioconductor.org). The mappings contain a total 21,253 probes mapping to unique gene identifiers (AGIs). From 56 ULE-

associated genes, 41 were contained within the updated mappings. Log_2-transformed dChip expression values were imported into R Version 2.11.1, where all subsequent analyses were performed. To simplify analyses, replicated array signals were averaged. Heatmaps were generated using functionality provided by the R-package gplots (Version 2.8.0) (Warnes et al. 2010).

Purifying selection of ULEs

The coordinates of the conserved elements were used to extract flanking regions that spanned 500 bp upstream and downstream from each ULE. Genome annotation information from TAIR9 (arabidopsis.org) was used to randomly select a group of 50 coding sequences from the collection of all exons across the five *A. thaliana* chromosomes. Eighty-three genomes, obtained from the ongoing 1001 *Arabidopsis* Genomes project (1001genomes.org) (Cao et al. 2011), supplied variation data for the sequences in each group. Separately, sequences from all three groups of *A. thaliana* sequence were aligned to their *A. lyrata* and *V. vinifera* counterparts using BLAST. The sequence at the node of the *A. lyrata/A. thaliana* phylogenetic precursor was ancestrally reconstructed using maximum likelihood as implemented in the PAML v4.3 software suite (Yang 2007) under a HKY85 nucleotide substitution model (Hasegawa et al. 1985). The ancestral sequence, aligned to the *A. thaliana* population data, provided a reference to determine whether the variations seen in the alignment were ancestral or derived. By parsing the collection of *A. thaliana* individuals and comparing variation to the ancestral sequence, we were able to unfold a derived allele frequency (DAF) spectrum.

To estimate the strength of selection on each group of sequences, we took a hierarchical Bayesian approach. To fit our Bayesian hierarchical model, we used the Markov Chain Monte Carlo (MCMC) algorithm described in Katzman et al. (2007), which uses the Metropolis–Hastings algorithm for updates. Briefly, this model aims to estimate the mean and standard deviation of an unknown normal distribution representing the selective effect of new alleles in each of a series of "classes" of DNA (ULEs, exons, flanking sites). Individual alleles are each assumed to have their own selection coefficients, drawn as independent, identically distributed random variables from this distribution. Furthermore, selection coefficient estimates were corrected to account for divergence-based ascertainment biases present in the ULE sequences (Kern 2009).

To evaluate the elements, flanking regions, and exonic regions, we ran six independent chains of 500,000 samples for the group of ULEs and for the flanking and exonic regions, respectively. To assess whether the chains converged, we plotted Gelman's potential scale reduction factor (Brooks and Gelman 1998) as implemented in the Coda R package (Plummer et al. 2006). After a reliable convergence, we discarded the first 25,000 iterations as burn in, and we used the remaining samples to estimate a selection coefficient distribution as plotted in Supplemental Figure S3.

Mammalian UCE analysis

Mammalian UCEs, the human genome (February 2009, hg19), and the mitochondrial genome were obtained from the University of California Santa Cruz (genome.ucsc.edu). UCEs were used in BLAST searches against human cDNA sequences (enseble.org), eukaryotic tRNAs (gtrnadb.ucsc.edu), and ncRNAs from the Noncoding RNA database (biobases.ibch.poznan.pl/ncRNA). All mammalian UCEs that did not have matches in these data sets were used in BlastX searches against the nonredundant NCBI database to search for protein similarities. Subsequently, UCEs were used in BLASTN searches against the *E. coli* genome.

Data access

The sequence data reported in this manuscript have been submitted to the EMBL Nucleotide Sequence Database (EMBL-Bank) (http://www.ebi.ac.uk/embl/). The accession numbers for the *A. thaliana* ULEs are HE963851–HE963886. The accession numbers for the monocot ULEs are HE963887–HE964756.

Acknowledgments

We are indebted to Detlef Weigel, Jun Cao, Korbinian Schneeberger, and Stephan Ossowski (Max Planck Institute for Developmental Biology, Tübingen) for providing access to SNP data for the ULEs in the *Arabidopsis* accessions they resequenced; and to Sharon Kessler for comments on the manuscript. This work was supported by the University of Zürich and a Syngenta PhD-Fellowship of the Zürich-Basel Plant Science Center. D.H. and A.D.K. are supported by Dartmouth College, the Neukom Institute, and NSF grant MCB-1052148.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5:** e234. doi: 10.1371/journal.pbio.0050234.

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

Bengtsson H, Bullard J, Gentleman R, Hansen KD, Morgan M. 2010. affxparser: Affymetrix file parsing. SDK. R package version 1.22.0. http://bioconductor.wustl.edu/bioc/html/affxparser.html.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* **13:** 137–144.

Bossolini E, Wicker T, Knobel PA, Keller B. 2007. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: Implications for wheat genomics and grass genome annotation. *Plant J* **49:** 704–717.

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *J Comput Graph Statist* **7:** 434–455.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koening D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43:** 956–963.

Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, Hooker T, Yazaki J, Li P, Skiba N, Peng Q, et al. 2007. Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. *Cell* **131:** 1340–1353.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80:** 692–704.

Chiang CW, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu CT. 2008. Ultraconserved elements: Analyses of dosage sensitivity, motifs and boundaries. *Genetics* **180:** 2277–2293.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452:** 215–219.

Costas C, Sanchez MP, Stroud H, Yu Y, Oliveros JC, Feng S, Benguria A, Lopez-Vidriero I, Zhang Z, Solano R, et al. 2011. Genome-wide mapping of *Arabidopsis thaliana* origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol* **18:** 395–400.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33:** e175. doi: 10.1093/nar/gni179.

Davidson EH. 2001. *Genomic regulatory systems*. Academic Press, San Diego.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc Natl Acad Sci* **101:** 1904–1909.

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578–582.

Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, Antonarakis SE. 2004. Comparison of human chromosome 21 conserved nongenic

sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* **14:** 852–859.
Derti A, Roth FP, Church GM, Wu CT. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38:** 1216–1220.
Ding J, Hu H, Li X. 2012. Thousands of *cis*-regulatory sequence combinations are shared by *Arabidopsis* and poplar. *Plant Physiol* **158:** 145–155.
Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al. 2006. Conserved non-coding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38:** 223–227.
Elgar G. 2009. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Brief Funct Genomics Proteomics* **8:** 256–265.
Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc Natl Acad Sci* **106:** 5737–5742.
Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. *Curr Opin Plant Biol* **12:** 126–132.
Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of *homothorax* mRNA splicing. *Genome Res* **15:** 800–808.
Guo H, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15:** 1143–1158.
Guo WJ, Ling J, Li P. 2009. Consensus features of microsatellite distribution: Microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. *Genomics* **93:** 323–331.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22:** 160–174.
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* **43:** 476–481.
Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* **41:** 1275–1281.
Hughes A, Rando OJ. 2009. Chromatin 'programming' by sequence—is there more to the nucleosome code than %GC? *J Biol* **8:** 96.
Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, Freeling M. 2003. Conserved noncoding sequences in the grasses. *Genome Res* **13:** 2030–2041.
International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463:** 763–768.
Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved non-coding sequences in the grasses. *Proc Natl Acad Sci* **99:** 6147–6151.
Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317:** 915.
Kern AD. 2009. Correcting the site frequency spectrum for divergence-based ascertainment. *PLoS ONE* **4:** e5152. doi: 10.1371/journal.pone.0005152.
Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **39:** 1151–1155.
Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O. 2004. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168:** 1575–1584.
Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, et al. 2010. Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci* **107:** 8063–8070.
Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and de novo assembly of the Giant Panda genome. *Nature* **463:** 311–317.
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.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.
Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567.
Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol* **3:** e93. doi: 10.1371/journal.pbio.0030093.
Masterson J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* **264:** 421–424.
McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet* **5:** e1000762. doi: 10.1371/journal.pgen.1000762.
Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452:** 991–996.
Nawy T, Lee JY, Colinas J, Wang JY, Thongrod SC, Malamy JE, Birnbaum K, Benfey PN. 2005. Transcriptional profile of the *Arabidopsis* root quiescent center. *Plant Cell* **17:** 1908–1925.
Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. 2006. AGRIS and AtRegNet. a platform to link *cis*-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* **140:** 818–829.
Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, Elgar G, deGraaff E, Lopez-Rios J, Zeller R, Grzeschik KH. 2007. Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. *Dev Growth Differ* **49:** 543–553.
Paterson AH, Freeling M, Tang H, Wang X. 2010. Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* **61:** 349–372.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.
Plummer M, Best N, Cowles K, Vines K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6:** 7–11.
Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85:** 774–781.
Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, Hadzhiev Y, Thies WG, Scherer G, Strähle U. 2008. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* **318:** 366–377.
Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu CR, Korkin D. 2012. Long identical multispecies elements in plants and animal genomes. *Proc Natl Acad Sci* **109:** 1183–1191.
Romano LA, Wray GA. 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both *cis* and *trans*-acting components of transcriptional regulation. *Devlopment* **130:** 4187–4199.
Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JP, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772–778.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** 1–10.
Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25:** 402–408.
Stewart W, Rothwell GW. 1993. *Paleobotany and the evolution of plants*, 2nd ed. Cambridge University Press, Cambridge, UK.
Stormo GD. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Sun H, Skogerbø G, Wang Z, Liu W, Li Y. 2008. Structural relationships between highly conserved elements and genes in vertebrate genomes. *PLoS ONE* **3:** e3727. doi: 10.1371/journal.pone.0003727.
Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of maize and sorghum genome progenitors. *Genome Res* **14:** 1916–1923.
Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596–1604.
Vanyushin BF, Ashapkin VV. 2011. DNA methylation in higher plants: Past, present and future. *Biochim Biophys Acta* **1809:** 360–368.
Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol* **8:** R15. doi: 10.1186/gb-2007-8-2-r15.
Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40:** 158–160.

Walter K, Abnizova I, Elgar G, Gilks WR. 2005. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet* **21:** 436–440.

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:** 487–490.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S et al. 2010. gplots: Various R programming tools for plotting data. R package version 2.8.0. http://cran.r-project.org/web/packages/gplots/index.html.

Wicker T, Narechania A, Sabot F, Stein J, Vu GT, Graner A, Ware D, Stein N. 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9:** 518. doi: 10.1186/1471-2164-9-518.

Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* **20:** 1229–1237.

Woodhouse MR, Pedersen B, Freeling M. 2010. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet* **6:** e1000949. doi: 10.1371/journal.pgen.1000949.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3:** e7. doi: 10.1371/journal.pbio.0030007.

Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, Wellmer F, Rahnenführer J, von Mering C, Grossniklaus U. 2010. *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. *Curr Biol* **20:** 506–512.

Yadav RK, Girke T, Pasala S, Xie M, Reddy GV. 2009. Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche. *Proc Natl Acad Sci* **106:** 4941–4946.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J Mol Evol* **48:** 597–604.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* **21:** 809–818.

Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z. 2010. PMRD: Plant microRNA database. *Nucleic Acids Res* **38:** 806–813.

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

Received July 22, 2011; accepted in revised form August 8, 2012.