DOI: 10.1534/genetics.110.114629

Note

Assessing the Influence of Adjacent Gene Orientation on the Evolution of Gene Upstream Regions in *Arabidopsis thaliana*

Fei He,* Wei-Hua Chen,†,‡ Sinéad Collins,*,[1] Claudia Acquisti,*,[2] Ulrike Goebel,* Sebastian Ramos-Onsins,§ Martin J. Lercher‡ and Juliette de Meaux*,[3]

**Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany, †European Molecular Biology Laboratory, 69117 Heidelberg, Germany, ‡University of Düsseldorf, 40225 Düsseldorf, Germany, and §Centre for Research in Agricultural Genomics, Universitat Autonoma de Barcelona, 08193 Bellaterra, Spain*

Manuscript received January 22, 2010
Accepted for publication March 9, 2010

ABSTRACT

The orientation of flanking genes may influence the evolution of intergenic regions in which *cis*-regulatory elements are likely to be located: divergently transcribed genes share their 5′ regions, resulting either in smaller "private" spaces or in overlapping regulatory elements. Thus, upstream sequences of divergently transcribed genes (bi-directional upstream regions, or URs) may be more constrained than those of uni-directional gene pairs. We investigated this effect by analyzing nucleotide variation segregating within and between Arabidopsis species. Compared to uni-directional URs, bi-directional URs indeed display lower population mutation rate, as well as more low-frequency polymorphisms. Furthermore, we find that bi-directional regions undergo selection for the maintenance of intergenic distance. Altogether, however, we observe considerable variation in evolutionary rates, with putative signatures of selection on two uni-directional upstream regions.

In recent years, the relevance of noncoding DNA for phenotypic evolution has been documented in ever greater detail (Wray 2007). Studies of nucleotide variation within and between species suggest that a substantial number of adaptive fixations can be detected in these regions (Keightley and Gaffney 2003; Andolfatto 2005; Mustonen and Lassig 2007). However, much remains to be learned about the functional and genomic factors that influence the evolution of noncoding DNA.

Several recent studies suggest that the distance and relative orientation of adjacent genes may influence the evolution of gene upstream regions (URs), which host most *cis*-regulatory elements. In the human genome, short URs located between divergently transcribed genes (*i.e.*, oriented head-to-head on the chromosome and <1000 bp) are overrepresented and display a mirror structure indicative of bi-directional *cis*-regulatory activity (Trinklein *et al.* 2004; Engström *et al.* 2006). The preference for closely located head-to-head gene pairs appears to be specific to the mammalian lineage. However, the genes involved are not enriched in mammal-specific genes, suggesting that gene rearrangement per se may provide new regulatory configurations and opportunities for phenotypic novelties (Koyanagi *et al.* 2005). In yeast, adaptive evolution in the relative orientation of adjacent genes could not be proven, despite numerous gene losses following recurrent whole-genome duplications (Byrnes *et al.* 2006).

Cis-regulatory elements are known to impose constraints at least up to ~1 kb upstream of transcription starting sites (Keightley and Gaffney 2003; Zeller *et al.* 2008). As a consequence, when distributed head-to-head, closely located adjacent genes may share a reduced personal "regulatory space" or have common or intermingled *cis*-regulatory elements, both of which may constrain opportunities for evolutionary change. Instead, adjacent genes distributed on the same chromosomal strand each may have their own "regulatory space" and may be more easily decoupled in evolution. DNA located upstream of gene coding regions (in URs) may thus be under distinct constraints, according to the relative orientation and distance of the second gene flanking the UR (Figure 1). In a first approximation,

[1]*Present address:* Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3JT Edinburgh, United Kingdom.

[2]*Present address:* Center for Evolutionary Functional Genomic, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301.

[3]*Corresponding author:* Institute for Evolution and Biodiversity, Hüfferstrasse 1, D48149 Münster, Germany.
E-mail: juliette.de.meaux@uni-muenster.de

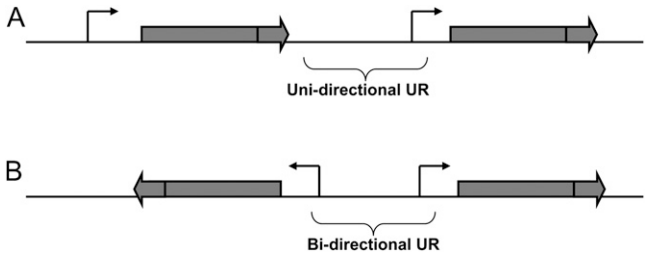

Figure 1.—Schematic of the two classes of upstream regions (URs). Shaded block arrows represent open reading frames along the chromosome and thin solid arrows represent transcription start sites. Uni-directional URs (A) are encoded on the same DNA strand. Bi-directional URs (B) are located between two start codons.

bi-directional URs (*i.e.*, located between head-to-head genes) may have to accommodate up to twice as many constraints for regulation compared to uni-directional URs. This special type of epistasis may constrain the regulatory evolution of divergently transcribed gene pairs, and adaptive novelties may thus more easily evolve in the context of uni-directional URs. Comparative genomics in the Ostreococcus genus of uni-cellular algae has indeed brought evidence for stronger constraints in intergenic regions separating genes in head-to-head orientation (Piganeau *et al.* 2009). However, the same study also revealed that patterns of constraints in the genus Saccharomyces were clearly different.

Compared to animals, the evolution of *cis*-regulation in higher plants has received less attention (Wright and Andolfatto 2008). Nevertheless, the above-mentioned hypothesis is particularly interesting to test in the plant species *Arabidopsis thaliana*. This species has undergone recent genome shrinkage, with a reduction in chromosome number (from 8 to 5) and a 40% reduction in genome size (Johnston *et al.* 2005; Schranz and Mitchell-Olds 2006; Oyama *et al.* 2008). One can therefore hypothesize that constraints on URs may have recently increased in this species because genes on average will be more closely located. We examined nucleotide variation in 29 URs and asked the following questions: (i) Are bi-directional URs more constrained than uni-directional URs? (ii) Is variation in uni-directional URs more likely to be adaptive compared to bi-directional URs? (iii) Does intergenic length influence the evolution of the two types of URs differently?

BI-DIRECTIONAL URs TEND TO DISPLAY LOWER LEVELS OF POLYMORPHISM

We picked 29 light-responsive genes, *i.e.*, nonconstitutively expressed genes with a known or putative function, which were separated by <1000 bp from their closest upstream neighbor. We consider these loci as functionally independent, given that a large proportion of genes are light-responsive in *A. thaliana* (Jiao *et al.* 2005). All upstream neighbors were confirmed by at least a full-length cDNA sequence. Only one pair of neighbor genes were tandem duplicates (locus 760). Sizes of the intergenic regions ranged from 52 to 992 bp. Sixteen of these intergenic regions were bi-directional URs (*i.e.*, located between two start codons). We sequenced these regions in 12 genotypes of *A. thaliana* (Bor4, RRS-10, Shakhdara, Bur-0, Lov-, Ler-1, Ts-1, C24, Tsu-1, Cvi-0, NFA-8, Got-7) and retrieved the Col-0 alleles from the database (http://www.arabidopsis.org). Sequences were submitted to public databases under the accession numbers FN674063–FN674399. Diversity was measured by the average number of pairwise differences, π. On average, levels of variation observed in our set of URs ($\pi = 0.0057$; Table 1) are comparable to levels of synonymous variation reported in *A. thaliana* (Nordborg *et al.* 2005; Schmid *et al.* 2005). Interestingly, variance in levels of diversity across our set of URs ($\sigma = 0.008$) was very similar to variance displayed by the 56 intergenic loci examined in Nordborg *et al.* (2005) ($\sigma = 0.007$). This indicates that our set of URs provides a representative picture of variance in levels of nucleotide variation in noncoding regions. For bi-directional URs, π ranged from 0.0004 to 0.0097, with an average of 0.0029 ($\sigma = 0.002$; Table 1). For uni-directional URs, the average was more elevated (0.0091), and the variance was significantly greater ($\sigma = 0.011$), with values ranging from 0.0003 to 0.0421 (Bartlett's test of homoscedasticity: $\chi^2 = 23.522$, $P < 0.001$). The two distributions were significantly different (Kruskal–Wallis test: $\chi^2 = 4.4308$, d.f. = 1, $P < 0.036$).

We computed maximum-likelihood estimates of population mutation rates (θ) for each locus with MANVa (described in Schmid *et al.* 2005). This rate reflects the rate at which mutations are created at each generation minus those mutations that are immediately removed from the population by negative selection. We found that our data set is best described by three population mutation rates: $\theta_1 = 0.0017$ for 12 loci, $\theta_2 = 0.0057$ for 15 loci, and $\theta_3 = 0.0349$ for 2 loci (likelihood comparisons of model with two rates *vs.* model with three rates, $P = 0.04$). Of 13 uni-directional URs, 3, 8, and 2 loci exhibited θ_1, θ_2, and θ_3 population mutation rates, respectively. By contrast, of 16 bi-directional URs, none exhibited a θ_3 population mutation rate, and 9 and 7 exhibited low (θ_1) and intermediate (θ_2) rates, respectively. We performed 1000 bootstraps over positions for each locus to compare the mean value of θ over all uni-directional and bi-directional URs. On average, θ is significantly lower in bi-directional URs (quantile comparison, $P \ll 0.001$). Our data therefore indicate that bi-directional URs tend to display lower population mutation rates, although levels of diversity are clearly disparate within and between UR types.

The analysis of sequence divergence between species showed the same trend, although differences between the two types of URs were only marginally significant.

TABLE 1

Summary statistics for polymorphism in 29 URs in 13 *A. thaliana* accessions

				Nucleotide diversity								Indel diversity				Neutrality test statistics			
UR type	Locus	Flanking genes		Sample size	No. of sites	No. of segregating sites	No. of sites with outgroup	Ratio alignable region (%)	No. of fixed differences between species	*K*	π	Length of alignment (including indel positions)	Total no. of indels	Average indel length within *A. thaliana*	Indel diversity per site, π(i)	*D*	*H*	*E*	Relative participation to significance of global HKA
Bi-directional URs	770	AT4G35770	AT4G35760	13	420	1	420	100	20	0.049	0.0004	458	0	0.0	0.0000	−1.149	0.318	−0.898	0.93
	570	AT4G02570	AT4G02560	13	597	2	597	100	38	0.067	0.0005	641	1	1.0	0.0002	−1.468	0.422	−1.229	1.38
	220	AT5G64210	AT5G64220	13	815	6	764	94	74	0.104	0.0011	909	2	9.0	0.0007	−1.930	0.604	−1.899	1.05
	670	AT1G19670	AT1G19680	13	778	3	298	38	37	0.136	0.0012	851	7	2.6	0.0030	−0.227	0.520	−0.585	2.33
	400	AT5G66400	AT5G66410	13	718	4	655	91	102	0.176	0.0015	790	3	4.0	0.0009	−0.626	−0.061	−0.352	3.33
	560	AT3G46560	AT3G46550	13	736	6	507	69	36	0.075	0.0016	816	4	2.8	0.0008	−1.440	0.422	−1.229	1.26
	490	AT3G09490	AT3G09480	13	338	3	338	100	36	0.116	0.0017	437	0	0.0	0.0000	−1.233	0.593	−1.290	0.68
	990	AT1G73990	AT1G73980	12	524	4	524	100	45	0.092	0.0019	567	0	0	0.0000	−0.902	0.657	−1.176	0.57
	680	AT2G42680	AT2G42670	13	710	8	452	64	54	0.131	0.0022	849	7	6.1	0.0031	−1.485	0.574	−1.780	0.67
	840	AT2G17840	AT2G17845	10	508	5	508	100	50	0.111	0.0023	1076	12	2.2	0.0039	−1.388	−2.031	1.002	0.34
	500	AT3G24500	AT3G24495	13	526	6	405	77	56	0.158	0.0027	682	2	1.0	0.0007	−0.950	−0.686	−0.022	0.49
	190	AT2G28180	AT2G28190	13	643	9	551	86	58	0.119	0.0034	751	8	11.5	0.0027	−0.944	−0.978	0.492	0.27
	940	AT5G64940	AT5G64930	13	765	12	489	64	31	0.069	0.0036	923	3	6	0.0011	−1.213	0.876	−1.610	2.38
	749	AT1G77490	AT1G77480	13	475	7	475	100	37	0.096	0.0040	576	1	1.0	0.0003	−0.559	0.750	−1.103	0.01
	120	AT5G47110	AT5G47120	13	646	18	627	97	43	0.086	0.0082	873	15	11.7	0.0039	−0.360	−1.978	1.726	4.69
	370	AT5G50370	AT5G50375	11	1040	27	367	35	33	0.096	0.0097	1309	21	6.5	0.0054	0.427	0.362	−0.905	1.78
	Average			12.6	639.9	7.6	498.6	82	46.9	0.105	0.0029	781.8	5.4	4.1	0.0017	−0.965	0.023	−0.679	
	Standard deviation			0.9	174.8	6.7	122.7	22	19.6	0.034	0.0026	225.5	6.1	4.0	0.0017	0.586	0.935	1.020	
Uni-directional URs	430	AT4G23430	AT4G23420	13	499	1	46	9	13	0.355	0.0003	637	1	1.0	0.0008	−1.149	na	na	1.84
	580	AT1G70580	AT1G70570	13	732	3	732	100	76	0.115	0.0011	826	2	2.5	0.0004	−0.478	−2.729	2.104	2.81
	810	AT2G42810	AT2G42820	13	890	5	687	77	63	0.099	0.0019	1062	4	6.8	0.0019	0.179	0.672	0.271	2.85
	667	AT2G26670	AT3G26640	13	1059	11	944	89	134	0.162	0.0026	1210	12	1.8	0.0030	−0.884	−0.800	0.106	1.54
	870	AT5G06870	AT5G06865	13	478	6	478	100	42	0.097	0.0038	559	3	7.3	0.0014	−0.216	0.457	−0.574	0.03
	630	AT1G27630	AT1G27620	13	760	11	638	84	102	0.185	0.0043	897	1	1.0	0.0006	−0.358	0.395	−0.520	0.92
	960	AT2G27960	AT2G27970	13	800	14	665	83	75	0.128	0.0049	1070	0	0	0.0000	−0.541	−0.149	−0.437	0.00
	470	AT5G08470	AT5G08460	13	431	8	398	92	31	0.091	0.0049	614	6	14.0	0.0020	−0.676	−1.574	1.007	0.68
	850	AT4G11850	AT4G11860	12	855	27	717	84	48	0.084	0.0081	1588	7	7.7	0.0013	−1.009	−1.642	1.110	6.94
	650	AT3G26650	AT3G26640	13	845	18	845	100	111	0.155	0.0096	1028	6	11.8	0.0027	1.665	0.193	1.174	0.01
	240	AT3G21240	AT3G21250	12	339	10	339	100	49	0.175	0.0123	383	3	1.0	0.0040	1.081	−0.173	0.981	0.13
	760	AT2G03760	AT2G03770	11	385	33	385	100	21	0.103	0.0220	450	8	4.1	0.0055	−1.154	**-3.085**	2.341	**35.30**
	410	AT3G62410	AT3G62400	13	315	36	315	100	21	0.142	0.0422	406	11	7.4	0.0100	0.641	**-2.573**	3.335	**34.06**
	Average			12.7	645.2	14.1	522.3	81	60.5	0.145	0.0091	825.4	4.9	5.1	0.0026	−0.223	−0.917	0.908	
	Standard deviation			0.6	245.9	11.3	266.6	29	37.6	0.071	0.0115	360.0	3.8	4.5	0.0027	0.883	1.355	1.228	
Average				12.3	627.2	10.3	509.3	81.9	51.7	0.120	0.0055	782.1	5.2	4.5	0.0021	−0.604	−0.323	0.013	
Standard deviation				2.2	215.7	9.2	196.1	24.7	29.0	0.057	0.0081	297.3	4.9	4.0	0.0022	0.817	1.192	1.326	

K is an estimate of per-site divergence between *A. thaliana* and *A. lyrata* (with Jukes–Cantor correction), *D* is Tajima's test statistic (Tajima 1989), *Hn* is the normalized Fay and Wu statistic (Fay and Wu 2000; Zeng *et al.* 2006), and *E* is a composite test statistic (Zeng *et al.* 2006). Sequences were established as described in de Meaux *et al.* (2005) and were deposited under accession nos. FN674063-674399. URs were defined as the UTR and the intergenic regions sensu stricto. ClustalW alignments were optimized by eye. Summary statistics of population variation and neutrality tests were performed using the software MANVa, a software for multilocus analysis of variation (http://www.ub.edu/softevol/manva/). Indel analysis was performed with DNAsp version 5, following model 4 (Librado and Rozas 2009). Summary statistics departing from neutrality are shown in bold.

Orthologous sequences were retrieved from the *Arabidopsis lyrata* genome assembly, kindly provided by J. Fawcett and Y. van de Peer (University of Ghent). Alignments were optimized by eye, and rearranged regions were excluded from the analysis. In addition, fragments of length >30 bp for which >50% of positions did not align with *A. thaliana* were removed because it was not clear whether they were small insertions or highly diverged fragments. The occurrence of such interspersed fragments was variable among loci (Table 1). Levels of divergence were lower for bi-directional loci [average $K = 0.105$, range 0.049–0.176; average $K = 0.145$, range 0.084–0.355 for bi-directional and uni-directional URs, respectively (Table 1); Kruskal–Wallis test: $\chi^2 = 3.23$, d.f. = 1, $P = 0.07$]. Levels of divergence therefore provide a weak confirmation of the patterns found in the polymorphism within species. We find that divergence rates should be interpreted with caution because average levels of divergence do not reflect all the complexities of the interspecific differences in plant noncoding regions. Indeed, the divergence rates reported here are for the alignable noncoding regions only.

COMPARATIVE EVOLUTIONARY DYNAMICS

We nonetheless computed a multilocus Hudson–Kreitman–Aguadé (HKA) to test for homogeneity in the ratio of intraspecific polymorphism to interspecific divergence across all loci (Hudson *et al.* 1987). This test was significant, indicating variable ratios of polymorphism *vs.* divergence across loci ($\chi^2 = 109.3$, d.f. = 28, $P < 0.001$). Partial HKA computed for each locus revealed that two uni-directional loci, 410 and 760, strongly depart from the evolutionary rate shown by most loci (Table 1). These two loci display an excess of polymorphism relative to divergence. It is important to note at this point that positive selection, which is usually associated with high levels of divergence, may be difficult to detect. Indeed, we provide estimates of divergence only for the alignable noncoding regions. The two departing loci displayed significant Fay and Wu's *Hn* statistics, revealing an excess of high-frequency–derived mutations ($Hn = -2.57$, $P = 0.033$ and $Hn = -3.08$, $P < 0.014$, respectively; Table 1) for the two loci. The two loci present relatively low levels of divergence, so that alignment issues did not influence the identification of the derived state at polymorphic sites.

Locus 410 (located between gene AT3G62400 and AT3G62410) displayed two very distinct haplotypes, one being very similar to the *A. lyrata* sequence. Interestingly, the derived haplotype was mutated in the termination codon of AT3G62400, suggesting functional differences between segregating alleles at this locus.

At locus 760 (located between genes AT2G03760 and AT2G03770), both ORFs were intact and both code for sulfotransferases. We screened this region for putative polymorphic binding sites using the plant *cis*-regulatory element database (Higo *et al.* 1999). This UR contained many defense-related motifs consistent with the known expression pattern of AT2G03760, a pathogen-responsive transcript. We identified a sulfur responsive element (GAGAC) gained in the high-frequency allele at position 6 in the 5′ UTR of AT2G03760. It is tempting to speculate that this gain of function has caused the mutation to rise in frequency, but this hypothesis calls for a rigorous experimental validation. Altogether, our data do not bring definitive evidence that uni-directional URs are more likely to be associated with evolutionary novelties in *A. thaliana*, but support the hypothesis that *cis*-regulatory mutations are a potent resource for the exploration of novel and potentially adaptive phenotypes (Clark *et al.* 2006; de Meaux *et al.* 2006; Stern and Orgogozo 2008).

To understand how gene orientation influences UR evolution, we compared the ratio of polymorphism to divergence in bi-directional *vs.* uni-directional URs, using the HKA test (Hudson *et al.* 1987). This test can be implemented to test for differences in evolutionary dynamics between groups of genes because it does not assume equal mutation rates across loci. The test was performed for all positions in the UTR and intergenic regions. To make sure that the two regions apparently subject to selection were not driving differences between UR types, we subtracted the two loci, 410 and 760, from the data set. We found that ratios of polymorphism to fixed differences are comparable between the two types of noncoding regions ($\chi^2 = 0.2$, $P > 0.64$). A Fisher exact test confirmed this finding (odds ratio 1.26, $P = 0.14$). By contrast with the HKA test, this test may be excessively liberal because it assumes identical evolutionary history across loci of both types, an assumption that is met only with a large number of loci (Andolfatto 2005).

Given the elevated variance in levels of diversity and divergence, we did not compute mean statistics for each type of URs but compared distributions for Tajima's *D*, Fay and Wu's *Hn* (normalized *H*, developed by Zeng *et al.* 2006) and Zeng's *E* with Kolmogorov–Smirnov tests (Tajima 1989; Fay and Wu 2000; Zeng *et al.* 2006). Tajima's *D* can detect an excess of low-frequency polymorphism, which is either a signature of polymorphism recovery after a selective sweep or an indication of background selection maintaining slightly deleterious mutations at low frequency. Excess of low-frequency variants can also result from recent demographic events such as bottlenecks. *Hn* measures the excess of high-frequency polymorphism that can hitchhike with selected mutations or be caused by population admixture (Fay and Wu 2000). Zeng's *E* combines both approaches to yield a statistic that is more robust to confounding demographic effects (Zeng *et al.* 2006). For bi-directional URs, the distribution of Zeng's *E* values is shifted toward lower values (Kruskal–Wallis:

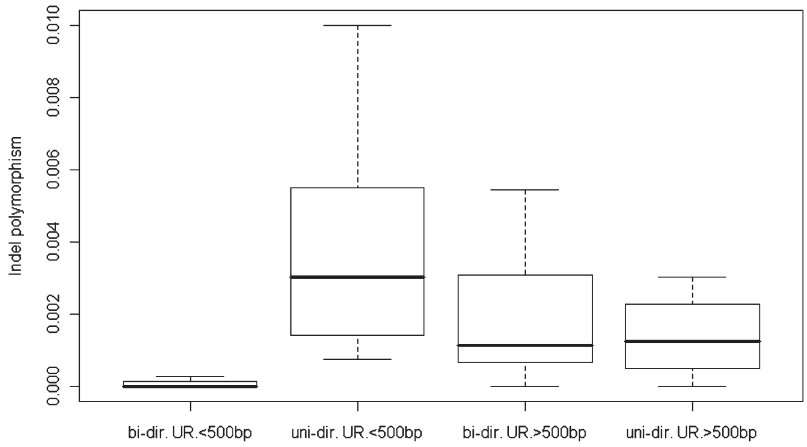

FIGURE 2.—Indel polymorphism and intergenic distance. Short bi-directional URs have a significantly lower level of indel polymorphism (two-way ANOVA: $F_{1,25} = 6.755$, $P < 0.01546$).

$\chi^2 = 8.1$, d.f. $= 1$, $P = 0.004$). For Tajima's *D*, a similar shift is observed ($\chi^2 = 6.4604$, d.f. $= 1$, $P < 0.012$). Both UR types presented similar Fay and Wu's *Hn* (Kruskal–Wallis: $\chi^2 = 2.5$, d.f. $= 1$, $P = 0.11$). In addition, we compared population mutation rates across UR types, after subtracting the two departing loci, and observed the same trend as reported above, but with marginal significance of the difference between uni- and bi-directional URs (quantile comparison, $P = 0.05$). Taken together, these results indicate that the heterogeneity in evolutionary rate alone does not account for the differences that we observed across UR types and confirms that bi-directional URs tend to harbor more low-frequency mutations.

EVOLUTION FOR THE MAINTENANCE OF INTERGENIC DISTANCE IN BI-DIRECTIONAL URs

We eventually examined patterns of insertion/deletion (indel) polymorphisms within *A. thaliana*. In our data set, low levels of polymorphism allow good alignments and an accurate identification of indel boundaries. Individual indel events were defined by the boundaries of gap stretches. Thus, overlapping gaps with different boundaries were considered as two different events. To characterize indel variation, we computed π, the average number of pairwise indel differences, as well as the average indel length for each locus (Table 1). All loci harbored indel polymorphism, with the exception of one bi-directional UR (490) and one uni-directional UR (960). On average, indel length was larger for uni-directional upstream regions than for bi-directional URs (5.531 bp *vs.* 4.359 bp), yet the distributions were not significantly different (Kruskal–Wallis test: $\chi^2 = 0.8239$, d.f. $= 1$, $P > 0.36$). Levels of indel diversity were also comparable between the two types of URs (0.0017 for bi-directional URs and 0.0027 for uni-directional URs; Kruskal–Wallis test: $\chi^2 = 1.0646$, d.f. $= 1$, $P > 0.30$).

More interestingly, however, we observed a relationship between intergenic length and indel polymorphism in bi-directional URs (Spearman's $\rho = 0.581$, $P < 0.018$) but not in uni-directional URs (Spearman's $\rho = -0.406$, $P > 0.16$). Indeed, we observe very low levels of indel polymorphism in bi-directional URs <500 bp ($F_{1,25} = 6.755$, $P < 0.016$, Figure 2). The exclusion of the two loci showing signs of adaptive evolution did not modify this trend, although the effect became only marginally significant ($F_{1,23} = 3.854$, $P = 0.06$). An analogous trend was not observed for the average number of pairwise single nucleotide differences π ($\rho = 0.08$, $P > 0.75$ and $\rho = -0.50$, $P > 0.08$ for bi-directional and uni-directional URs, respectively).

We extended our analysis to a larger set of URs. We used the draft genome of *A. lyrata* and recorded the distance between 10,390 pairs of orthologous flanking coding regions (3331 and 7059 bi- and uni-directional URs, respectively). We then examined the correlation of UR length between species. The correlation between length in *A. thaliana* and length in *A. lyrata* is higher in bi-directional than in uni-directional upstream regions, and the slopes are significantly different ($R = 0.755$ and 0.698, respectively; significant interaction: $t = -4.45$, $P < 1.e\text{-}05$). Bi-directional URs are shorter in *A. thaliana* than in *A. lyrata* but less so than uni-directional URs. This result supports the existence of stronger constraints on intergenic distance in bi-directional URs. The analysis of indel distributions instead did not yield conclusive results (not shown). However, current algorithms are not performing well for generating interspecific sequence alignments in plant noncoding regions. Comparisons of UR length between species provides an interesting example of how differences between species can be quantified without relying on alignments.

CONCLUSION

A few studies have investigated the patterns of coexpression or shared gene ontology categories among gene neighbors in *A. thaliana*, the best annotated plant genome (WILLIAMS and BOWLES 2004; KROM and RAMAKRISHNA 2008; WANG *et al.* 2009). Yet, surprisingly

few studies have investigated the evolutionary dynamics of plant noncoding regions (but see de Meaux *et al.* 2005 and Miyashita 2001). Here, we made the simplifying assumption that the most constrained regulatory elements are located in the 5′ upstream region. It is therefore quite remarkable that we observe consistent differences associated with the orientation of gene neighbors: bi-directional URs display lower polymorphism and a greater number of low-frequency polymorphisms.

Several hypotheses can be invoked to interpret this result. First, polymorphism can be decreased by positive selection on (or around) bi-directional URs or maintained by balancing selection on (or around) uni-directional URs. Any of these hypotheses would indeed explain why both types of URs present similar levels of divergence, although they differ in levels of polymorphism. However, it is not clear why selection patterns should differ between uni- and bi-directional URs in this way. Furthermore, we believe that estimates of interspecific divergence do not reflect adequately the evolutionary dynamics in these regions. Indeed, levels of divergence vary within loci, with conserved fragments flanking highly divergent regions. This has also been reported in other noncoding regions (Miyashita 2001; de Meaux *et al.* 2006). Therefore, levels of divergence give an indication of divergence only for the most conserved fragments.

Several elements indicate that the pattern that we observe could be explained by a second hypothesis: stronger constraints exerted on bi-directional URs. Indeed, we find lower polymorphism levels and a greater number of low-frequency polymorphisms in these regions. Importantly, we also find that indel polymorphisms are less frequent in short bi-directional URs. This relationship was confirmed at the interspecific level because bi-directional URs tended to be less reduced in size than uni-directional URs. Constraints on indel polymorphisms in intergenic regions have been demonstrated in *Drosophila melanogaster* (Ometto *et al.* 2005). In rice, a study of genes displaying *cis*-regulatory differences in a heterotic F_1 cross were shown to be associated with indel variation and not single nucleotide polymorphism (Zhang *et al.* 2008). Indel polymorphism may therefore alter *cis*-regulatory function more strongly than single nucleotide polymorphism. Maintaining a minimal length may be important to avoid collision between regulatory complexes and transcription factors in regions where regulatory elements may be particularly dense. Conversely, if regulatory elements are shared or intermingled, increased intergenic distances may disrupt regulatory interactions. Our study suggests that the distance between binding sites is a predominant functional feature encoded in these regions and shows that patterns of indel polymorphism can reveal features of noncoding DNA that are not apparent from SNP polymorphism data.

This conclusion, however, should not mask that our data also reflect the complexity of noncoding evolution in plant genomes. We actually observe considerable variation in evolutionary dynamics across URs in *A. thaliana*, with great disparity in rates of evolution within loci as well as within and between species (Table 1). The recent progress of high-throughput sequencing technologies will soon allow a comprehensive analysis of the noncoding genome and help confirm the significant differences observed here. Understanding the evolutionary mechanisms that drive noncoding evolution remains a goal of major importance for upcoming research in plants.

We thank Y. van de Peer and J. Fawcett for providing access to the *A. lyrata* assembly produced by Department of Energy–Joint Genome Institute's Community Sequencing Program through a proposal coordinated by Detlef Weigel, Max Planck Institute, Tübingen. This research was funded by the German Funding Agency (SFB680).

LITERATURE CITED

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature **437:** 1149–1152.

Byrnes, J. K., G. P. Morris and W. H. Li, 2006 Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol. Biol. Evol. **23:** 1136–1143.

Clark, R. M., T. N. Wagler, P. Quijada and J. Doebley, 2006 A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. **38:** 594–597.

de Meaux, J., U. Goebel, A. Pop and T. Mitchell-Olds, 2005 Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. Plant Cell **17:** 676–690.

de Meaux, J., A. Pop and T. Mitchell-Olds, 2006 *Cis*-regulatory evolution of chalcone-synthase expression in the genus Arabidopsis. Genetics **174:** 2181–2202.

Engström, P. G., H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa *et al.*, 2006 Complex loci in human and mouse genomes. PLoS Genet. **2:** e47.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Higo, K., Y. Ugawa, M. Iwamoto and T. Korenaga, 1999 Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. **27:** 297–300.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Jiao, Y. L., L. G. Ma, E. Strickland and X. W. Deng, 2005 Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis. Plant Cell **17:** 3239–3256.

Johnston, J. S., A. E. Pepper, A. E. Hall, Z. J. Chen, G. Hodnett *et al.*, 2005 Evolution of genome size in Brassicaceae. Ann. Bot. **95:** 229–235.

Keightley, P. D., and D. J. Gaffney, 2003 Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl. Acad. Sci. USA **100:** 13402–13406.

Koyanagi, K. O., M. Hagiwara, T. Itoh, T. Gojobori and T. Imanishi, 2005 Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. Gene **353:** 169–176.

Krom, N., and W. Ramakrishna, 2008 Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and Populus. Plant Physiol. **147:** 1763–1773.

Librado, P., and J. Rozas, 2009 DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:** 1451–1452.

Miyashita, N. T., 2001 DNA variation in the 5′ upstream region of the Adh locus of the wild plants Arabidopsis thaliana and Arabis gemmifera. Mol. Biol. Evol. **18:** 164–171.

Mustonen, V., and M. Lassig, 2007 Adaptations to fluctuating selection in Drosophila. Proc. Natl. Acad. Sci. USA **104:** 2277–2282.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. Plos Biol. **3:** 1289–1299.

Ometto, L., W. Stephan and D. De Lorenzo, 2005 Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. Genetics **169:** 1521–1527.

Oyama, R. K., M. J. Clauss, N. Formanova, J. Kroymann, K. J. Schmid *et al.*, 2008 The shrunken genome of Arabidopsis thaliana. Plant Syst. Evol. **273:** 257–271.

Piganeau, G., K. Vandepoele, S. Gourbiere, Y. Van de Peer and H. Moreau, 2009 Unravelling cis-regulatory elements in the genome of the smallest photosynthetic eukaryote: phylogenetic footprinting in Ostreococcus. J. Mol. Evol. **69:** 249–259.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

Schranz, M. E., and T. Mitchell-Olds, 2006 Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell **18:** 1152–1165.

Stern, D. L., and V. Orgogozo, 2008 The loci of evolution: How predictable is genetic evolution? Evolution **62:** 2155–2177.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar *et al.*, 2004 An abundance of bidirectional promoters in the human genome. Genome Res. **14:** 62–66.

Wang, Q., L. Wan, D. Li, L. Zhu, M. Qian *et al.*, 2009 Searching for bidirectional promoters in Arabidopsis thaliana. BMC Bioinformatics **10**(Suppl. 1):S29.

Williams, E. J. B., and D. J. Bowles, 2004 Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. **14:** 1060–1067.

Wray, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. **8:** 206–216.

Wright, S. I., and P. Andolfatto, 2008 The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. Annu. Rev. Ecol. Evol. Syst. **39:** 193–213.

Zeller, G., R. M. Clark, K. Schneeberger, A. Bohlen, D. Weigel *et al.*, 2008 Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Res. **18:** 918–929.

Zeng, K., Y. X. Fu, S. Shi and C. I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174:** 1431–1439.

Zhang, H. Y., H. He, L. B. Chen, L. Li, M. Z. Liang *et al.*, 2008 A genome-wide transcription analysis reveals a close correlation of promoter INDEL polymorphism and heterotic gene expression in rice hybrids. Mol. Plant **1:** 720–731.

Communicating editor: O. Savolainen