J Mol Evol (2009) 69:249–259
DOI 10.1007/s00239-009-9271-0

Unravelling *cis*-Regulatory Elements in the Genome of the Smallest Photosynthetic Eukaryote: Phylogenetic Footprinting in *Ostreococcus*

Gwenael Piganeau · Klaas Vandepoele · Sébastien Gourbière · Yves Van de Peer · Hervé Moreau

Received: 8 May 2009 / Accepted: 27 July 2009 / Published online: 20 August 2009

Abstract We used a phylogenetic footprinting approach, adapted to high levels of divergence, to estimate the level of constraint in intergenic regions of the extremely gene dense *Ostreococcus* algae genomes (Chlorophyta, Prasinophyceae). We first benchmarked our method against the *Saccharomyces* sensu stricto genome data and found that the proportion of conserved non-coding sites was consistent with those obtained with methods using calibration by the neutral substitution rate. We then applied our method to the complete genomes of *Ostreococcus tauri* and *O. lucimarinus*, which are the most divergent species from the same genus sequenced so far. We found that 77% of intergenic regions in *Ostreococcus* still contain some phylogenetic footprints, as compared to 88% for *Saccharomyces*, corresponding to an average rate of constraint on intergenic region of 17% and 30%, respectively. A comparison with some known functional *cis*-regulatory elements enabled us to investigate whether some transcriptional regulatory pathways were conserved throughout the green lineage. Strikingly, the size of the phylogenetic footprints depends on gene orientation of neighboring genes, and appears to be genus-specific. In *Ostreococcus*, 5′ intergenic regions contain four times more conserved sites than 3′ intergenic regions, whereas in yeast a higher frequency of constrained sites in intergenic regions between genes on the same DNA strand suggests a higher frequency of bidirectional regulatory elements. The phylogenetic footprinting approach can be used despite high levels of divergence in the ultrasmall *Ostreococcus* algae, to decipher structure of constrained regulatory motifs, and identify putative regulatory pathways conserved within the green lineage.

Keywords Phylogenetic footprinting · Non-coding DNA · *cis*-regulatory elements · *Saccharomyces* · *Ostreococcus*

Gwenael Piganeau and Klaas Vandepoele equally contributed to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9271-0) contains supplementary material, which is available to authorized users.

G. Piganeau · H. Moreau
UPMC Univ Paris 06, UMR 7621, Laboratoire Arago, 66651 Banyuls/mer, France

G. Piganeau (✉) · H. Moreau
CNRS, UMR7621, LOB, Observatoire Océanologique, 66651 Banyuls/mer, France
e-mail: gwenael.piganeau@obs-banyuls.fr

K. Vandepoele · Y. Van de Peer
Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), 9052 Gent, Belgium

K. Vandepoele · Y. Van de Peer
Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium

S. Gourbière
Laboratoire de Biologie et d'Ecologie Tropicale et Méditerranéenne, Université de Perpignan Via Domitia, UMR 5244, 52 Avenue Paul Alduy, 66860 Perpignan, France

Introduction

Discriminating between functional and junk sequences in the non-coding fraction of a genome has become one of the major challenges of functional and evolutionary genomics (Bird et al. 2006). Indeed, functional non-coding DNA is involved in the regulation of gene expression and thus in the evolution of novelties and adaptation between species (Castillo-Davis 2005). Functional non-coding sequences

fall into two main categories: protein binding sites such as transcription factor binding sites (TFBSs), enhancers, and silencers, which are involved in the control of gene expression, and sequences that control chromatin organization such as insulators and matrix attachment regions (Cooper and Sidow 2003).

At least two different in silico approaches have been developed and may be combined to extract functional non-coding elements from the bare genome sequence data. Word count approaches search for over or under-represented motifs, but the main shortcoming of these methods is the high rate of false positives they generate (Hampson et al. 2002). The comparative approach, or "phylogenetic footprinting" as defined by Tagle et al. (1988), relies on homologous sequence data from at least two species combined with evolutionary theory, which states that substitutions accumulate much faster at non-functional DNA bases than at functionally constrained base positions. The comparative approach, using pairwise and multiple genome comparisons, has been successfully applied to identify conserved elements in mammals (Dermitzakis et al. 2004; Xie et al. 2005), vertebrates (Bejerano et al. 2004; Siepel et al. 2005), *Drosophila* (Bergman and Kreitman 2001; Halligan et al. 2004; Siepel et al. 2005), *Caenorhabditis* (Shabalina and Kondrashov 1999; Siepel et al. 2005), and the *Saccharomyces* sensu stricto group (Chin et al. 2005; Cliften et al. 2001; Kellis et al. 2003). In these groups of organisms, species divergence is such that pairwise aligned segments still retain "false positives," that is, sequence identity because of shared ancestry, not because of selective constraint on a DNA sequence. As a consequence, the proportion of constrained non-coding sites has to be calibrated by the rate of neutral substitution, to correct for conserved but neutrally evolving sequences in *Saccharomyces sensu stricto* (Chin et al. 2005) or human–chimpanzee comparisons (Keightley et al. 2005).

This dense genome data coverage for evolutionary close groups of eukaryotic model organisms is not available for species belonging to the other four eukaryotic supergroups of the eukaryotic tree of life (Keeling et al. 2005), once the phylum of Unikonts (Fungi and Metazoans) has been removed. These organisms account for most of the eukaryotic diversity and whole genome data are still scattered along highly divergent branches, so that when two genomes from the same phylum are sequenced (as in Chlorophyta, Ciliates or Apicomplexan), the evolutionary distance reaches saturation on neutrally evolving sites. As a consequence, the methodological problem of discriminating between the phylogenetic footprints generated by selective constraints, and the footprints generated by shared ancestry, shifts the problem to that of discriminating the footprints generated by selective constraints and the footprints generated by the alignment algorithm itself.

The genus *Ostreococcus* belongs to the prasinophytes, an ecological important group dominating marine photosynthetic picoeukaryotes (Vaulot et al. 2008). They are the smallest eukaryotic free-living photosynthetic organisms identified to date, with a size of 1 μm, and are found worldwide in the marine environment (Rodriguez et al. 2005) and in the Sargasso Sea shotgun metagenome sequence data (Piganeau et al. 2008). *Ostreococcus tauri* and *Ostreococcus lucimarinus* cells are morphologically similar, even at electron microscopy level, and are characterized by a single chloroplast, a single mitochondrion and a cytoplasm bounded by a membrane lacking any detectable cell wall or scales. These species show specific adaptations to different environments as depth and/or light intensity (Rodriguez et al. 2005). The genome sequences of *O. tauri* (Derelle et al. 2006) and *O. lucimarinus* (Palenik et al. 2007) have recently been completed and revealed very short intergenic regions, raising the issue of the structure of its regulatory elements. The analysis of their protein coding genes revealed high levels of divergence, as measured by synonymous, non-synonymous and intronic rates of molecular evolution (Jancek et al. 2008; Piganeau and Moreau 2007), raising in turn a methodological problem to detect conserved intergenic regions.

From the comparison of these two genomes, we investigated (i) how we could estimate the degree of sequence conservation in intergenic sequences, (ii) whether some of these footprints are conserved through the green lineage by comparing the footprints with functional footprints previously identified in *Arabidopsis* and (iii) whether gene orientation of flanking genes influenced footprint size.

Methods

Datasets

Whole genome sequences and gene annotations for *O. tauri* and *O. lucimarinus* were downloaded from http://bioinformatics.psb.ugent.be/genomes/ and http://genome.jgi-psf.org/euk_home.html. When several gene annotations were available, we chose the Eugene annotation method that predicted shorter intergenic regions (to reduce positives due to unannotated coding sequences). *Saccharomyces bayanus* and *Saccharomyces cerevisiae* sequences were downloaded from the *Saccharomyces* genome database (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/) and *Candida glabrata* sequences from the Genolevure database (http://cbi.labri.fr/Genolevures/download/GL2_index.php).

CDS could be mapped on chromosomes by BLAST (Altschul et al. 1990): positions of each gene on chromosomes were extracted using codes implemented in C language.

Orthologous intergenic regions (OIR) were defined as intergenic sequences between two genes having orthologs (defined as reciprocal best blast hits) in the same order and orientation in the two genomes compared. These OIRs were then extracted from the whole genome sequence data to be aligned with different alignment tools.

Alignment Software and Processing

We propose a permutation based validation scheme to estimate the significance of detected conserved sequences. Briefly, for each pair of orthologous intergenic sequences, we apply an alignment method on the real dataset and on 100 randomized datasets. As such, it is possible to compare the observed degree of conservation to an empirically determined distribution. This is then used to correct the observed identity between two sequences, to get an estimate of the proportion of constrained sites in each intergenic sequence.

This approach has been first applied to the *S. cerevisiae* and *S. bayanus* genomes to test and validate our method, because both our approach and the calibration by the neutral substitution rate approach can be used in this genus, as a consequence of the wealth of genomes available. We then applied our method to the two unicellular green algae *O. tauri* and *O. lucimarinus*.

ACANA (Huang et al. 2006) is a pairwise sequence alignment algorithm that uses a Smith–Waterman-like dynamic programming algorithm. ACANA has been shown to be highly accurate for divergent sequences, as compared to BLASTZ (Schwartz et al. 2003), CHAOS (Brudno et al. 2003a) and DIALIGN (Pohler et al. 2005) for local alignments (Huang et al. 2006). After benchmarking additional alignment algorithms that have been especially designed to align non-coding DNA sequences, Lagan 1.1 (Brudno et al. 2003b) and Mavid 2.0 build 4 (Bray and Pachter 2004), we found that ACANA 1.10 used in local mode (shortly ACANAL) with parameters -C 5 -T 1 best discriminated real from random sequences, as most OIRs with at least one significant footprint were obtained with this algorithm (results not shown). For both *Ostreococcus* and *Saccharomyces* OIR alignment software comparison led us to use ACANAL for all further analysis to study the nature and the level of constraint in OIR of *Ostreococcus* and *Saccharomyces*. Local footprints reported by ACANA have a minimal length of 9 bp.

Screening Footprints for Functional *cis*-Regulatory Elements

In order to investigate the putative function of our footprints, we assigned a significance level to each local alignment, hereafter "footprint." The problem of assessing the significance of an alignment is complex (Altschul et al. 1994). To estimate the significance of footprints obtained for ACANAL, we ran the alignment method to the real and to 100 re-shuffled datasets, keeping the mononucleotide frequencies equal (i.e., for each re-shuffled dataset, all positions of both sequences were randomly permutated). We then calculated a score, defined as the number of nucleotide matches multiplied by the percent identity of the match, for each footprint, and estimated a *P*-value by counting how many times a bigger score was observed in the re-shuffled dataset. Note that stretches of 'N' (sequence gaps in final genome assembly) were maintained as fixed blocks during the shuffling process. This part was computer intensive because of running each alignment program over 400,000 times (number of intergenic regions times 100 times) and took approximately 3 weeks per whole genome comparison on a 2.8 GHz, Intel Computer with 1.5 GB of RAM.

We first investigated whether some of the significant footprints ($P < 0.05$) could be due to incomplete or mis-annotations. We therefore screened the footprints identified by ACANAL for open reading frames (blastx against GenBank, *E*-value $< 1\times 10^{-3}$) and found that putative ORFs are negligible in our data (*Saccharomyces* $\sim 1.02\%$ and *Ostreococcus* $\sim 1.66\%$). Next, we screened the footprints for the presence of RNA genes (using the 116 RNA annotations of *O. tauri* and 132 for *Saccharomyces*) and found no OIR containing RNA genes in *Ostreococcus* and 36 in *Saccharomyces* out of 6 and 52 RNA loci present in the complete OIR dataset, respectively. These RNAs were removed for further analysis.

We compiled a list of 589 yeast regulatory elements described in the literature (referred to as reference motifs): 160 from Kellis et al. (2003), 50 from SCPD (Zhu and Zhang 1999), and 379 from Elemento and Tavazoie (2005) (Elemento and Tavazoie 2005). Similarly, we collected all plant motif instances from AGRIS (Davuluri et al. 2003) and PLACE (Higo et al. 1999), yielding 986 (partially redundant) reference motifs. We retrieved all footprints and compared these sequences with the reference motifs using DNA-pattern from rsa-tools (http://rsat.ulb.ac.be/rsat/) allowing zero substitutions. The fold enrichment for reference motifs was calculated by taking the ratio of the observed over the expected frequency of motifs located in footprints, where the latter was computed by counting motif instances on mono-nucleotide reshuffled sequences (i.e., maintaining base composition).

Estimating the Proportion of Nucleotide Sites Under Constraint

For each alignment with i segments containing Id_i identical nucleotides per segment, we defined a conservation score, *S*, that gives the fraction of conserved nucleotides in the alignment.

$$S = \sum_i \frac{\mathrm{Id}_i}{\mathrm{length\_of\_alignment}}$$

The proportion of conserved nucleotides in an alignment j, S_j, can be expressed as the sum of the proportion of nucleotides conserved as a result of constraint, F_j, that share 100% identity, and the proportion of nucleotide conserved by chance, that equals $1 - F_j$, by the average identity observed in the re-shuffled random sequences, $S_{\mathrm{random},j}$:

$$S_j = F_j \times 1 + (1 - F_j) \times \overline{S}_{\mathrm{random},j}$$

Thus for each intergenic alignment, the proportion of nucleotides under constraint, F_j can be estimated as:

$$F_j = \frac{S_j - \bar{S}_{\mathrm{random},j}}{1 - \bar{S}_{\mathrm{random},j}}$$

Footprints for each OIR in *Ostreococcus* and *Saccharomyces* are available as Supplementary material.

Results

Estimation of the Level of Constraint on Intergenic Regions

We first benchmarked our method on *Saccharomyces* genome data, in order to compare our estimate of the proportion of phylogenetic footprints with other estimates based on multiple genomes alignments containing species showing less divergence. We focused on the two most distant *Saccharomyces* sensu stricto, *S. cerevisiae* and *S. bayanus*, showing saturated substitution patterns (Ks $\sim$ 1.2), before performing a whole genome comparison of the two marine unicellular algae *O. tauri* and *O. lucimarinus*.

We extracted 2,758 OIRs from *Ostreococcus* and 2,203 from *Saccharomyces.* An OIR is defined as a pair of intergenic regions flanked by the same orthologous genes (showing similar relative transcriptional orientations) in both species. Note that this definition is more stringent compared to the frequently applied definition of orthologous promoters, which does not consider the conservation of both flanking genes. Sequences features of OIRs together with features of the compared genomes are described in Table 1. Since the *S. bayanus* genome sequence assembly is distributed over 1,098 contigs, it is not meaningful to compare the overall percent of OIRs (number of OIRs divided by total number of orthologous genes) between *Ostreococcus* and *Saccharomyces*. Strikingly, *Ostreococcus* intergenic regions are, on average, 25% shorter than *Saccharomyces* intergenic regions, suggesting greater compaction of regulatory elements in *Ostreococcus*, and/or smaller regulatory elements in OIRs.

For each OIR, we estimated the proportion of sites under constraint, F, as the excess of identity in the observed alignment as compared to the expected alignment for the shuffled intergenic region (see Methods section). Our estimate of the average F in yeast is 30.4%, consistent with previous estimates of 30% based on conservation calibration with the local neutral substitution rate (Chin et al. 2005). However, F is not normally distributed, suggesting that the average level of constraint is not a good descriptor of the genome wide level of constraint on intergenic regions (Fig. 1). It appears that 12% (*Saccharomyces*) to 23% (*Ostreococcus*) of intergenic regions contain no detectable footprint at all ($F = 0$), and that there is a large

Table 1 General features of the *Saccharomyces* and *Ostreococcus* genomes and sequence features of orthologous gene set and OIRs used in analysis

	O. tauri	*O. lucimarinus*	*S. cerevisiae*	*S. bayanus*
Genome size (Mb)	12.6	13.2	12.1	nd
Chromosomes	20	21	16	16
Gene number	7892	7651	6563	nd
% of genes with intron(s)	20	25	5	nd
Length of OIRs (bp)	233	343	463	449
GC frequency in OIR	0.62	0.66	0.34	0.36
Gene orthologs	6243		4395	
*aaId** (%)	70		82	
Ks	>2		1.2	
nb of OIR	2758		2203	
Footprints (nb of OIR)	8612 (2540)		12749 (2183)	
Footprints $P < 0.05$ (nb of OIR)	2350 (1539)		4710 (1769)	

OIR orthologous intergenic region, *aaId* average amino acid identity between orthologs, *GC* average GC content

* From Palenik et al. (2007)

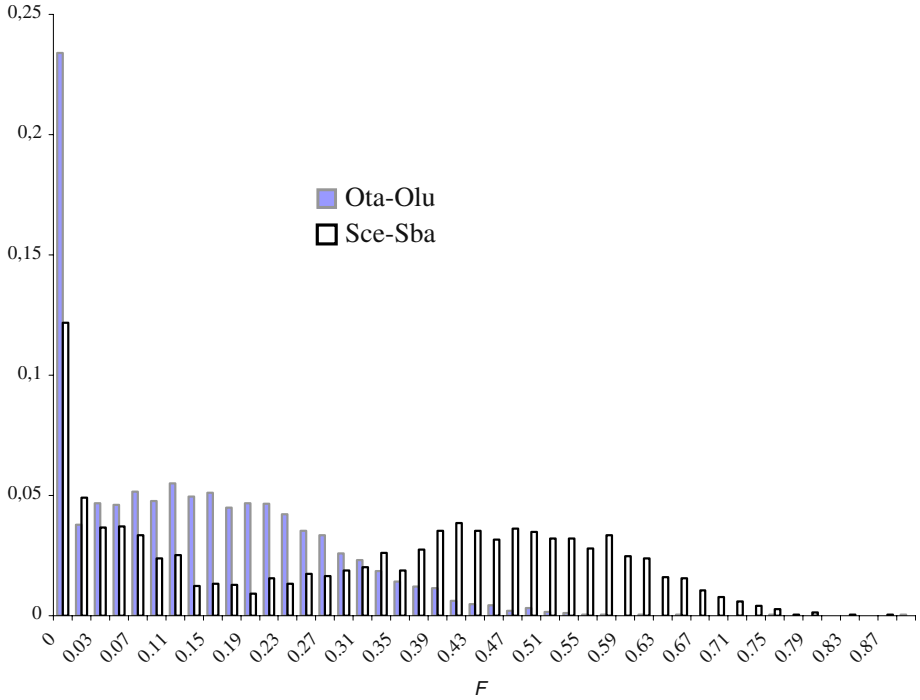

Fig. 1 *F* distribution in *Ostreococcus* (Ota-Olu) and in *Saccharomyces* (Sce-Sba)

variation in the level of constraint for the remaining OIR. Excluding the intergenic regions with $F = 0$, the average level of constraint in intergenic regions raises from 30.4 to 34.7% in yeast and from 13 to 17% in *Ostreococcus*.

Moreover, *F* is strikingly different with regard to gene orientation of neighboring genes (Table 2). Head-to-head OIR (divergent gene pairs) are on average longer in both species, followed by head-to-tail OIR and tail-to-tail OIR (converged gene pairs), hereafter HH, HT, and TT. However, if following (Chin et al. 2005), we define conserved sequences as regulatory elements, the regulatory element structure seems to be different between *Saccharomyces* and *Ostreococcus*. In yeast, the longer regulatory elements are in HT OIR (200 bp) whereas in *Ostreococcus*, the longer regulatory elements are in HH OIR (53 bp) (Table 2). In addition, the low *F* value for *Ostreococcus* TT compared to HH OIR (0.08 and 0.18, respectively) suggests that 3′ regulatory elements are rare. This is in contrast with *Saccharomyces*, where the average *F* value for HH and TT is identical (0.17).

Table 2 Average *F* differs between gene orientation

	HH	HT	TT
Saccharomyces			
Number	538	1037	608
Average length (bp)	585	487	304
Average *F*	0.17	0.45	0.17
Average length of conserved sites[a]	97	200	44
Number of OIR with $F = 0$	115	1	150
Ostreococcus			
Number	864	839	837
Average length	437	293	155
Average *F*	0,18	0,13	0,08
Average number of conserved sites[a]	53	30	11
Number of OIR with *F*=0	129	134	331

HH head-to-head orientation for the two neighboring genes, two start codons limit the OIR, *HT* head-to-tail orientation for the two neighboring genes, one start and one stop codon limit the OIR, *TT* tail-to-tail orientation for the two neighboring genes, two stop codons limit the OIR

[a] The length of conserved sites in each OIR *i*, is the product of F_i by the length of the smallest OIR

From these results, we can test null models of regulatory element structure for the three different types of OIR, as defined in Hermsen et al. (2008) (Fig. 2). First, let us define a simple model of regulatory element structure, the additive regulatory element model. In this model, a HH region contains two 5′-regulatory element sequences of total length l_{HH}, a TT region contains two 3′-regulatory element sequences of total length l_{TT}, and a HT region contains one

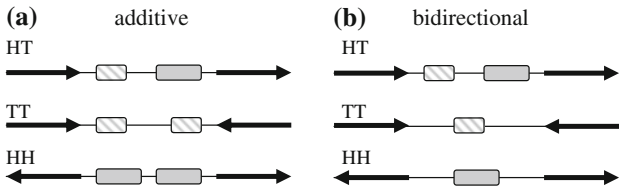

Fig. 2 Regulatory element structures. *Black arrows* give the 5′->3′ orientation of neighboring genes, *shaded box*: 3′ regulatory elements, of total length l_{TT}, *gray box*: 5′ regulatory elements, of total length l_{HH}

of each type of regulatory element sequences, of total length $l_{HT} = l_{TT} + l_{HH}$. We have three distributions to estimate the distribution of two variables, l_{TT} and l_{HH}. We can thus use the remaining distribution to test whether the expected distribution equals the observed distribution. We estimate l_{TT}, l_{HT}, and l_{HH} from the total number of constrained nucleotides in each OIR, excluding OIR with no conserved site ($F = 0$). Note that this is different from the Hermsen et al. (2008) study that is based on the total length of the intergenic regions.

Second, we defined the bidirectional regulatory element structure, where a HH region contains one 5′-regulatory element sequences of length l_{HH}, a TT region contains one 3′-regulatory element sequences of length of length l_{TT}, and a HT region contains one of each type of regulatory element sequences.

The additive regulatory element structure model cannot be rejected from the *Ostreococcus* data (Table 3), whereas this model does not fit the *Saccharomyces* data ($P < 10^{-5}$), because of the greater than expected length of regulatory elements in HT regions. Strikingly, even the bidirectional regulatory element structure model, predicting an average l_{HT} of 154 bp cannot account for the observed length of regulatory elements in HT regions (Table 3).

Benchmarking Method on *Saccharomyces* Data: Comparison with Known Motifs and Gene Orientation Effect

To evaluate the power of our phylogenetic footprinting approach to identify *cis*-regulatory elements, we compared all significant footprints (as defined in "Methods" section) against a reference set of regulatory motifs (589 for fungi and 936 for green plants; see "Methods" section). For 97% of the *Saccharomyces* footprints (4559/4710) and 99% of the *Ostreococcus* footprints (2329/2350), there was a perfect match with a reference motif. Although the degenerate nature of several of the reference regulatory elements, in general, hinders the identification of biologically functional motif instances (Vavouri and Elgar 2005), it is interesting to note that for both species a large number of reference motifs occur much more frequently in our footprints than expected by chance. The enrichment for reference motifs was calculated by taking the ratio of the observed over the expected frequency of motifs located in footprints, where the latter was computed by counting motif instances on reshuffled sequences. Considering the 50 motifs from the Promoter Database of *S. cerevisiae* (SCPD), 74% are twofold or more enriched in the significant footprints. Examples in yeast are the ESR1 binding site (GATGAG Kellis_g036), the SCPD_GCN4 motif regulating biosynthetic genes in response to amino acid starvation, the PAC binding site (CTCATCGCA Elemento_21; involved in rRNA transcription) and the Met4 binding site (AACTGTGGC Kellis_g057; involved in amino acid metabolism) all showing a greater than threefold enrichment.

To verify if some of the reference motifs could resemble 3′ regulatory signals captured in our footprints, we analyzed the frequencies of all reference motifs over the HH, HT, and TT OIRs. Whereas most motifs are most frequent in the HT class and nearly absent in the TT class—indicative of 5′ promoter regulatory motifs, we did find some motifs that occur with high frequencies in TT OIRs. One example in yeast is the TATATA upstream efficiency element (nataTATATAyATATATAnn, 4% HH, 41% HT, and 56% TT; $n = 27$), an mRNA 3′-end processing element appearing upstream of the poly(A) cleavage site (Graber et al. 1999). Another motif WTATWTACADG described by Kellis and co-workers is also depleted in HH OIRs (4% HH, 69% HT, and 27% TT; $n = 48$) and resembles a down-stream element identified in a set of co-expressed genes whose product localizes to the cytosolic translational machinery, the mitochondrial DNA translational machinery or the mitochondrial outer membrane (Kellis et al. 2003). Also in *Ostreococcus*, we found some motifs, including the plant poly(A) signal AATAAT (4% HH, 16% HT, and 80% TT; $n = 25$), enriched in TT and depleted in HH OIRs (Fig. 3). These results demonstrate that our footprints also capture 3′ regulatory elements most probably playing a regulatory role in mRNA splicing, localization, or stability.

Table 3 Estimates of the average lengths of the conserved regulatory elements under the additive and the bidirectional models (see text and Fig. 2)

	l_{HH} estimate	l_{TT} estimate	l_{HT} estimate	predicted l_{HT}	P value
Ostreococcus					
Additive model	26	5	30	32	0.07
Bidirectional model	53	11	30	64	$<10^{-5}$
Saccharomyces					
Additive model	48	22	200	77	$<10^{-5}$
Bidirectional model	97	44	200	154	$<10^{-5}$

The test of each model is done by comparison of the estimated total length of HT regulatory elements, l_{HT}, with the distribution of the predicted distribution under each model, obtained from 100,000 random sampling of $l_{HH} + l_{TT}$

Searching for Function of Footprints in *Ostreococcus*: Investigating Motif Conservation for Four *cis*-Regulatory Elements

For *Ostreococcus*, the verification of known regulatory elements is hampered by our limited knowledge about

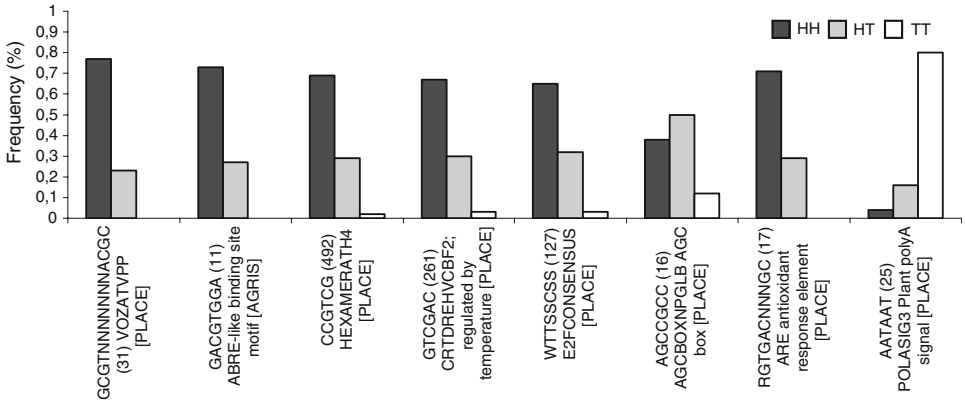

Fig. 3 Gene orientation frequencies for known plant motifs enriched in conserved Ostreococcus footprints. The series refer to the frequency of motifs in conserved footprints counted per OIR class (*HH* head-to-head, *HT* head-to-tail, and *TT* tail-to-tail). Numbers in parenthesis indicate the number of significant footprints matching a known motif used to calculate the frequencies. The data for all motifs are provided as Supplementary material

transcriptional control in these green algae. Therefore, we selected four well studied *cis*-regulatory elements (E2F, TELO, I-box, and CRE/DRE motif) from land plants (i.e., *Arabidopsis thaliana*) and investigated whether these motifs are conserved in green algae and, if so, whether they are present in our *Ostreococcus* footprints. For the E2F box, involved in the regulation of DNA replication genes during S-phase, we initially selected 38 *Arabidopsis* genes containing a consensus WTTSSCSS motif in their promoter sequence (Vandepoele et al. 2005) and annotated as involved in DNA replication. For the corresponding orthologous *Ostreococcus* genes present in our data set, we found that half (7/14) of the OIRs contained a (W)TTSSCSS motif, and in each of these seven cases, the motif was located in a footprint. Examples of genes with a conserved E2F site are two DNA polymerase subunits (alpha and beta), a MCM subunit and an ORC subunit (with $P < 0.01$, 0.01, 0.51, and 0.86, respectively hypergeometric distribution; Pilpel et al. 2001), well-described E2F target genes in plants and animals. Complementary, mapping the location of the E2F footprints in relationship to the gene orientation reveals that 97% of all motifs occur in HH or HT OIRs, confirming its role as a promoter *cis*-regulatory element (Fig. 3). The TELO box (AAACCCTA) is frequently found in the promoter of cytosolic ribosomal proteins in *Arabidopsis* (Tremousaygue et al. 1999). We selected 26 *Ostreococcus* genes orthologous to known TELO target genes from *Arabidopsis* that were annotated as involved in ribosome biogenesis and assembly. However, only one *Ostreococcus lucimarinus* gene in this dataset contains the (A)AAACCCT(A) motif, which was not located in the OIR of the *O. tauri* gene. The I-box (CTTATC) is a promoter element frequently found in light-responsive genes. Starting from a set of 297 *Arabidopsis* I-box genes (Vandepoele et al. 2006), we found that 33 orthologous *Ostreococcus* genes contain the (C)TTATC motif, of which only three motif instances are located in footprints. Although this result might suggest that our identified footprints only partially detect these *cis*-regulatory elements, an alternative scenario is that these boxes—as they are defined in *Arabidopsis*—are not functional in green algae. To investigate this possibility, we first identified the set of *Ostreococcus* genes (orthologous to the *Arabidopsis* target genes) that contain the motif (discarding if it is located in a footprint or not) and then assessed whether the functional annotation linked to the *Arabidopsis* reference genes is conserved in *Ostreococcus*. Whereas for the E2F box a strong (conserved) GO enrichment toward DNA replication was found (hypergeometric distribution $P < 0.01$), the GO enrichment for photosynthesis ($P < 0.01$) observed in *Arabidopsis* I-box target genes is not observed in *Ostreococcus*. Consistent with this, the presence of an I-box in the promoter of *Ostreococcus* rbcS, a known I-box target gene in higher plants (Altschul et al. 1990), could not be confirmed.

The (G/a)(T/c)CGAC CRE/DRE transcription factor-binding site has been shown to be involved in the Low Temperature response in *Hordum* (barley) and *Arabidopsis* (Xue 2003; Sharma et al. 2005 for review). The GTCGAC motif is overrepresented in our significant footprints and frequently found in HH OIR in *Ostreococcus* (67% HH vs. 30% HT and 3% TT; Fig. 3).

The low temperature response involves transcriptional activation and repression of several pathways, and activated pathways include those involved in the accumulation of osmoprotectants like sugars, amines, and compatibles solutes (Sharma et al. 2005). We investigated whether we could identify regions of similarity between the 66 upregulated *A. thaliana* genes under cold response containing this motif in their 1 kb upstream sequence (Vogel et al.

2005) and the *Ostreococcus* genes also having this motif in a significant footprint. To assess whether the number of genes with regions of similarity was higher than expected by chance, we compared it with the number of genes with regions of similarity having a significant footprint and no GTCGAC motif. We found a significant excess of homologues with the *Ostreococcus* genes also having a GTCGAC motif as compared with the *Ostreococcus* genes not having this motif (Fisher exact test, $P = 0.01$).

We also investigated that *Ostreococcus* contains a gene homologue to the transcription factors associated to this motif in *A. thaliana*, the CRT/DRE binding factors. These transcription factors belong to the large AP2 multigene family of DNA-binding proteins (Riechmann and Meyerowitz 1998). There are only two homologues to these genes in both *Ostreococcus* genomes, annotated as transcription factors as they contain the AP2 like DNA-binding domain (JGI protein ID 23938 and 9237 in *O. lucimarinus*; 23659 and 30470 in *O. tauri*).

Discussion

We investigated the level of constraint in non-coding regions from pairs of divergent streamlined genomes using a phylogenetic footprinting approach. This methodology is specifically adapted to three genomic features that seem to characterize recent unicellular eukaryotic genome projects. First, these genomes are gene dense and thus contain short intergenic regions that can be analyzed over their full length. Second, pairwise genome comparison shows a well-preserved short scale synteny, enabling a stringent definition of OIRs by shared adjacent orthologous genes. Third, genomes available for comparison share a high level of divergence, as measured by complete saturation at neutral evolving sites, making constraint estimates by calibration with neutral substitution rates practically unsuitable. The method applied in this study is expected to produce a very low fraction of false positives. First, because it has been developed for highly divergent species, where all neutral sites have been overwritten and there is no sequence similarity due to shared ancestry alone. Second, the observed alignments were benchmarked against alignments obtained from randomized sequence dataset to quantify the amount of conservation expected by chance and to determine significance values for the different footprints.

On the other hand, our method has three main limitations. The first one is common to all phylogenetic footprinting methods, that is that they are only able to detect the highly constrained functional sequences, which is only a fraction of the actual functional non-coding sequences of an organism (Li et al. 2007; Samanta et al. 2006; Wittkopp 2006). The second one is that our method will not detect constrained sequences of low complexity, as mono-nucleotide repeats. This is because the randomized sequences (based on the real dataset) will be too much like the real dataset to allow discrimination. The third one is that we will not detect any conserved sequence shorter than 9 bp, the minimum size of segments reported by the local alignment algorithm ACANA applied in this study.

Using the *S. cerevisiae—S. bayanus* genome comparison as a benchmark, we show that our method effectively detects regulatory elements previously identified by other methods relying on multiple species comparison. Our estimation of the proportion of constrained sites in yeast non-coding regions, 30%, is consistent with previous estimates relying on multispecies comparison and calibration by the neutral substitution rate (Chin et al. 2005). We then applied our method to the two available genomes of *Ostreococcus*, the smallest free-living eukaryotic photoautotrophic cells, to investigate regulatory element structure. There are approximately half as many constrained sites in intergenic regions in *Ostreococcus* as compared to *Saccharomyces*, with an average proportion of constrained sites, *F*, of 13%. This is consistent with the higher divergence of the two *Ostreococcus* genomes (Table 1), and reflects the degradation of some evolutionary information with time. On the other hand, this degradation enables to come closer to the sequence backbone of gene expression control, that remain conserved between even more distantly related species, as exemplified by the conserved non-coding sequences of master control genes in development, conserved between fly and chicken (Blanco et al. 2005).

We also found that the level of constraint depends on the type of intergenic region considered, given by the 5′ or 3′ orientation of the neighboring genes. This has not been reported for multicellular genomes as such, because OIRs are not defined for adjacent orthologous gene pairs, but are rather defined by taking a region upstream and downstream one gene. However, 3′ UTR sequences have been shown to be more constrained than 5′ UTR in vertebrates and to a lesser extent in Drosophila, possibly reflecting widespread post-transcriptional regulation by microRNAs (Siepel et al. 2005). Since microRNA interference is not believed to occur in to the two organisms we analyzed (Cerruti and Casas-Mollano 2006 for review), the different trend we observe is not surprising.

Previous studies have already pinpointed some qualitative inter-kingdom differences between genome structure, as the positive correlation between first intron length and expression in *Arabidoposis* and rice as opposed to animals (Ren et al. 2006). We observed that the regulatory element structure estimated from the *Ostreococcus* genomes is markedly different from the yeast regulatory element structure. Indeed, the regulatory element structure in *Ostreococcus* is consistent with an additive regulatory element

structure model, where the total length of constrained sites in HT intergenic regions corresponds to the sum of constrained sites in a 3′ regulatory element and in a 5′ regulatory element. This also may suggest some kind of optimization of intergenic sequence space in *Ostreococcus*, because the length of each type of intergenic region is proportional to the level of constraint. On the other hand, the regulatory element structure in yeast is biased toward a higher level of constraint in HT regions. There are at least three possible explanations for this trend. First, cases of bidirectional regulatory element structure in HH intergenic regions have been reported experimentally in yeast (Ishida et al. 2006), and a bidirectional regulatory element structure suggests a greater compaction of regulatory elements in HH non-coding regions. However, if we assume that all HH and TT intergenic regions contain bidirectional regulatory elements, we show that there is still an average excess of regulatory element sequences in HT region (46 bp, Table 3), so that this cannot be the sole factor responsible for this trend, in addition to the fact that it is unlikely that all HH regions contain bidirectional regulatory elements. Recently, a study on the distribution of TT, HT, and HH total intergenic sequence lengths in *S. cerevisiae*, ignoring sequence conservation, suggested that about 30% of HH regions contain a bi-directional transcriptional regulatory regions (Hermsen et al. 2008). Second, yeast HT intergenic regions could evolve at a slower pace than HH and TT intergenic regions. This could be a consequence of a recent partial loss of an interleaving orthologous gene leading to a present HT OIR. Indeed, it has been shown that before the whole genome duplication (WGD), gene order orientation in *Saccharomyces* was more biased toward HH and TT occurrences, and that single gene deletions in pairs of HH and TT genes are responsible for the present distribution of adjacent gene orientation (Byrnes et al. 2006). To test this scenario, we investigated whether HT OIR prior to WGD are shorter and have a lower F than more recent HT OIR, using the available genome data of *A. gossypii*, a pre-WGD species. We re-estimated F in the 245 adjacent gene pairs of the *S. cerevisiae—S. bayanus* that were already adjacent and in same orientation between Sc and Ag. We found a higher average F in these pre-WGD OIR than in the more recent OIR (preWGD:postWGD OIR:HH: $F = 18.6{:}16.8$, TT: $F = 47.8{:}44.7$, and HT: $F = 15.2{:}17.2$). However, the higher level of constraint in HT regions remains in the preWGD gene pairs and there is thus no detectable effect of single gene deletion on the pattern we observe in yeast.

Third, transcriptional interference, the perturbation of one transcription unit by another, could exercise different selective pressures on different gene orientations. There is experimental evidence for differential strengths of transcriptional interference as a consequence of gene orientation in mammalian cells, with more interference in tandems of HT genes (Eszterhas et al. 2002), but to our knowledge, this has not been investigated in yeast.

Estimating the proportion of sites under constraint in non-coding regions is often independent from a functional analysis non-coding regulatory elements, and this may lead to apparent conflicting interpretations (Bush and Lahn 2005; Keightley et al. 2006). Comparing our footprints against a reference set of yeast and plant motifs indicates that a high proportion of conserved footprints identified in this study captures known *cis*-regulatory elements. A detailed analysis of the E2F pathway targeting DNA replication genes in animals and higher plants reveals that several *Ostreococcus* genes are orthologous to *bona fide* replication genes containing an E2F binding site in their promoter. This finding indicates that this pathway is also evolutionary conserved in green algae. Interestingly, all E2F binding sites present in *Ostreococcus* DNA replication that were conserved between species were detected using our pairwise footprinting approach. We have also established a three level homology based correspondence between (i) a *cis*-regulatory motif overrepresented in our footprints (the cold response CRE/DRE element), (ii) the transcription factors binding this motif, and (iii) the genes up regulated by this motif in *Arabidopsis*. Experimental analysis is now required to demonstrate that the *cis*-regulatory motif proposed is a binding site upregulating gene transcription under low temperature, to conclude about the conservation of the low temperature regulation pathway between *Ostreococcus* and *Arabidopsis*.

Altogether, these findings indicate that the E2F pathway is conserved in green algae and that there may be conservation of the Low Temperature pathway involving the CRE/DRE transcription binding site. In contrast, the *cis*-regulatory control of cytosolic ribosomal proteins and light-regulated genes in green algae might be driven by other promoter elements than the TELO or the I-box, respectively. These results indicate that, like observed between different yeast species (Tanay et al. 2005), the underlying *cis*-regulatory network has evolved substantially within the green plant lineage, even for highly conserved processes like photosynthesis or ribosome biogenesis. However, these results show that the footprints we describe provide a starting point to characterize other *cis*-regulatory elements in green algae and may thus enable to unravel very ancient regulatory pathways. Deciphering the genomic information in free-living unicellular eukaryotes will enable us to picture the toolbox of functional non-coding sequences in ancestor eukaryotic cells. On the other hand, it will enable to pinpoint *cis*-regulatory divergence and major regulatory novelties implied in the evolution of multicellularity in plants and animals.

Acknowledgments We would like to thank an anonymous referee for constructive comments on a previous version and Eric Bonnet for help with alignment software and sequence shuffling. We would also like to thank Severine Jancek, Nigel Grimsley, Stéphane Rombauts, Pierre Rouzé, David Waxman, and Jan Wuyts for critical comments and stimulating discussions. This collaboration was founded by Tournesol. G.P. was granted an EMBO short-term fellowship and a "Marine Genomics Europe" GAP fellowship (European Network of Excellence 2004–2008 GOCE-CT-2004-505403). K.V. is a postdoctoral fellow of the Fund for Scientific Research, Flanders. This work was supported by the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6(2):119–129

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

Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335–1345

Bird CP, Stranger BE, Dermitzakis ET (2006) Functional variation and evolution of non-coding DNA. Curr Opin Genet Dev 16:559–564

Blanco J, Girard F, Kamachi Y, Kondoh H, Gehring WJ (2005) Functional analysis of the chicken delta1-crystallin enhancer activity in Drosophila reveals remarkable evolutionary conservation between chicken and fly. Development 132:1895–1905

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699

Brudno M, Chapman M, Göttgens B, Batzoglou S, Morgenstern B (2003a) Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinform 4:66

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S (2003b) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Bush EC, Lahn BT (2005) Selective constraint on noncoding regions of hominid genomes. PLoS Comput Biol 1:e73

Byrnes JK, Morris GP, Li WH (2006) Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol Biol Evol 23:1136–1143

Castillo-Davis CI (2005) The evolution of noncoding DNA: how much junk, how much func? Trends Genet 21:533–536

Cerruti H, Casas-Mollano JA (2006) On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet 50:81–99

Chin CS, Chuang JH, Li H (2005) Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res 15:205–213

Cliften P, Hillier L, Fulton L, Graves T, Miner T, Gish W, Waterston R, Johnston M (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 11:1175–1186

Cooper GM, Sidow A (2003) Genomic regulatory regions: insights from comparative sequence analysis. Curr Opin Genet Dev 13:604–610

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of Arabidopsis *cis*-regulatory elements and transcription factors. BMC Bioinform 4:25

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA 103:11647–11652

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

Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol 6:R18

Eszterhas S, Bouhassira E, Martin D, Fiering S (2002) Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol Cell Biol 22:469–479

Graber JH, Cantor CR, Mohr SC, Smith TF (1999) Genomic detection of new yeast pre-mRNA 3′-end-processing signals. Nucleic Acids Res 27:888–894

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD (2004) Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. Genome Res 14:273–279

Hampson S, Kibler D, Baldi P (2002) Distribution patterns of over-represented k-mers in non-coding yeast DNA. Bioinformatics 18:513–528

Hermsen R, ten Wolde PR, Teichmann S (2008) Chance and necessity in chromosomal gene distributions. Trends Genet 24:216–219

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

Huang W, Umbach DM, Li L (2006) Accurate anchoring alignment of divergent sequences. Bioinformatics 22:29–34

Ishida C, Aranda C, Valenzuela L, Riego L, DeLuna A, Recillas-Targa F, Filetici P, López-Revilla R, González A (2006) The UGA3-GLT1 intergenic region constitutes a promoter whose bidirectional nature is determined by chromatin organization in *Saccharomyces cerevisiae*. Mol Microbiol 59:1790–1806

Jancek S, Gourbiere S, Moreau H, Piganeau G (2008) Clues about the genetic basis of adaptation emerge from comparing the proteomes of two *Ostreococcus* ecotypes (Chlorophyta, Prasinophyceae). Mol Biol Evol 25:2293–2300

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW (2005) The tree of eukaryotes. Trends Ecol Evol 20:670–676

Keightley PD, Lercher MJ, Eyre-Walker A (2005) Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 3:e42

Keightley PD, Lercher MJ, Eyre-Walker A (2006) Understanding the degradation of hominid gene control. PLoS Comput Biol 2:e19 author reply e26

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Li L, Zhu Q, He X, Sinha S, Halfon MS (2007) Large-scale analysis of transcriptional *cis*-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8:R101

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA 104:7705–7710

Piganeau G, Moreau H (2007) Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta). Gene 406:184–190

Piganeau G, Desdevises Y, Derelle E, Moreau H (2008) Picoeukaryotic sequences in the Sargasso sea metagenome. Genome Biol 9:R5

Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 29:153–159

Pohler D, Werner N, Steinkamp R, Morgenstern B (2005) Multiple alignment of genomic sequences using CHAOS, DIALIGN and ABC. Nucleic Acids Res 33:W532–W534

Ren XY, Vorst O, Fiers MW, Stiekema WJ, Nap JP (2006) In plants, highly expressed genes are the least compact. Trends Genet 22:528–532

Riechmann JL, Meyerowitz EM (1998) The AP2/EREBP family of plant transcription factors. Biol Chem 379:633–646

Rodriguez F, Derelle E, Guillou L, Le Gall F, Vaulot D, Moreau H (2005) Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). Environ Microbiol 7:853–859

Samanta M, Tongprasit W, Sethi H, Chin C, Stolc V (2006) Global identification of noncoding RNAs in *Saccharomyces cerevisiae* by modulating an essential RNA processing pathway. Proc Natl Acad Sci 103:4192–4197

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–107

Shabalina SA, Kondrashov AS (1999) Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet Res 74:23–30

Sharma P, Sharma N, Deswal R (2005) The molecular biology of the low-temperature response in plants. Bioessays 27:1048–1059

Siepel A, Bejerano G, Pedersen J, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier L, Richards S, Weinstock G, Wilson R, Richard A, Gibbs R, Kent W, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439–455

Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. Proc Natl Acad Sci 102:7203–7208

Tremousaygue D, Manevski A, Bardet C, Lescure N, Lescure B (1999) Plant interstitial telomere motifs participate in the control of gene expression in root meristems. Plant J 20:553–561

Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GT, Gruissem W, Van de Peer Y, Inze D, De Veylder L (2005) Genome-wide identification of potential plant E2F target genes. Plant Physiol 139:316–328

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

Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton (< or =3 μm) in marine ecosystems. FEMS Microbiol Rev 32:795–820

Vavouri T, Elgar G (2005) Prediction of *cis*-regulatory elements using binding site matrices—the successes the failures and the reasons for both. Curr Opin Genet Dev 15:395–402

Vogel JT, Zarka DG, Van Buskirk HA, Fowler SG, Thomashow MF (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. Plant J 41:185–211

Wittkopp P (2006) Evolution of *cis*-regulatory sequence and function in Diptera. Heredity 97:139–147

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434:338–345

Xue GP (2003) The DNA-binding activity of an AP2 transcriptional activator HvCBF2 involved in regulation of low-temperature responsive genes in barley is modulated by temperature. Plant J 33:373–383

Zhu J, Zhang MQ (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. Bioinformatics 15:607–611