RSAT::Plants: Motif Discovery within Clusters of Upstream Sequences in Plant Genomes

Bruno Contreras-Moreira (1, 2), Jaime Castro-Mondragón (3,4), Claire Rioualen (3,4), Carlos P. Cantalapiedra (1) and Jacques van Helden (3,4)

1. Estación Experimental de Aula Dei-CSIC, Av. Montañana 1.005, 50059 Zaragoza, Spain.
2. Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.
3. INSERM, U1090 TAGC, Marseille, F-13288, France.
4. AixMarseilleUniversity, U1090 TAGC, Marseille, F-13288, France.

Corresponding author: Bruno Contreras Moreira bcontreras@eead.csic.es

Running head: Discovering regulatory sequences in co-expressed plant promoters

Summary/Abstract

The plant-dedicated mirror of the Regulatory Sequence Analysis Tools (RSAT, http://plants.rsat.eu) offers specialized options for researchers dealing with plant transcriptional regulation. The web site contains whole-sequenced genomes from species regularly updated from Ensembl Plants and other sources (currently 40), and supports an array of tasks frequently required for the analysis of regulatory sequences, such as retrieving upstream sequences, motif discovery, motif comparison, pattern matching. RSAT::Plants also integrates the footprintDB collection of DNA motifs. This protocol explains step-by-step how to discover DNA motifs in regulatory regions of clusters of co-expressed genes in plants. It also explains how to empirically control the significance of the result, and how to annotate the discovered motifs with putative binding factors.

Key Words: co-expression, DNA motif, position-weight matrix, upstream sequence, cluster

1. Introduction

Transcriptome data (microarrays, RNA-seq) have been extensively used as a proxy for genetic regulation in many organisms, as the analysis of genome-wide profiles of gene transcription under different treatments uncovers clusters of genes with correlated behaviors, which may result from direct or indirect co-regulation. A classical application of this approach was done by Beer and co-workers ***(1)*** with yeast microarray data sets obtained in a variety of experimental conditions. In that experiment, expression data-mining was demonstrated to be an effective strategy for finding regulons, groups of genes that share regulatory mechanisms and functional annotations.

Other studies have unveiled that the outcome of these approaches largely depends on the genomic background of the species under study. For instance, Sand and others ***(2)*** reported that the significance of DNA motifs discovered in *Saccharomyces cerevisiae* promoters is much higher for regulons than for random gene sets of the same sizes, but for human promoters the signal-to-noise ratio is almost null, because random gene sets give highly significant motifs due to heterogeneities in promoter compositions and biases due to repetitive elements. For metazoans, it is thus a real challenge to distinguish *bona fide* motifs from noise ***(2)***. These observations suggest that motif discovery on sequence clusters faces intrinsic properties of the genomes under study, regardless of the software used for the task.

Among plants, these strategies have so far been tested on the model *Arabidopsis thaliana*, and they have been successfully applied to the identification of novel *cis*-regulatory elements validated with synthetic promoters ***(3)***. Yet, with the exception of this model, these sorts of experiments have not been possible in plants until recently. In spite of this, the growing list of available plant genomes encourages these analyses in combination with expression profiles obtained from either microarray or RNA-seq data sets, as in the recent work of Yu and collaborators ***(4)***, provided that these factors are considered:

- Plant genomes are rich in repetitive elements (RE) distributed along the genome ***(5)***, which pose particular problems for motif discovery statistics (violation of the independence assumption).

- Current genome assemblies range from 119.7Mb (*A.thaliana*) to 6.48Gb (*Triticumaestivum*). *Brachypodium distachyon*, a model species for grasses, is 271.9Mb. The quality of these assemblies and their RE content is also quite variable, as shown in **Fig. 1** and **Table 1**.

- Upstream regions, defined by annotated gene coordinates, are also of variable length, going from 1,123b on average in *A.thaliana* to 1,856b in *Aegilopstauschii* (see **Table 1**).

This chapter presents a step-by-step protocol for the task of discovering and annotating DNA motifs in clusters of upstream sequences for species supported by RSAT::Plants, which have been obtained mostly from Ensembl Plants (http://plants.ensembl.org)***(6)***, but also include data from the JGI Genome Portal (http://genome.jgi.doe.gov) ***(7)***, and the National Institute of Agrobiological Sciences in Japan (http://barleyflc.dna.affrc.go.jp/bexdb)***(8)***. In addition, RSAT::Plants integrates footprintDB (http://floresta.eead.csic.es/footprintdb)***(9)***, a collection of position-specific scoring matrices (PSSM) representing transcription factor binding motifs (TFBM), as well as their cognate binding proteins, which can be used to annotate discovered motifs and to predict potentially binding transcription factors, as illustrated in the chapter by Contreras-Moreira and Sebastiánin this book.

Discovering regulatory elements within natural genomic sequences is certainly an important scientific goal on its own, but can also be part of the design and validation of synthetic promoters. We envisage at least two applications in this context:

1. The characterization of promoters of genes with known expression properties, which can then be used to engineer the expression of genes of interest.

2. The validation of engineered promoters in order to make sure that they contain the expected regulatory elements which might be natural or engineered depending on the application.

2. Materials

This protocol requires disposing of:

1. A computer with any Web browser installed.

2. A set of gene clusters from any of the species currently supported at RSAT::Plants (http://plants.rsat.eu, *see***Note 1**). Here we will use three example clusters of co-expressed maize genes, shown in **Table 2** (s*ee***Note 2**). More generally, expression data can be obtained from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) ***(10)***and used to produce gene clusters of plant genes (see**Table 3**).

3. Methods

The following protocol enumerates the steps required to discover DNA motifs, based on the over-representation of k-mers (oligonucleotides) and dyads (spaced pairs of oligonucleotides), in clusters of upstream sequences. The protocol comprises two stages, analyzing first co-expressed genes and then random clusters as a negative control (*see***Note 3**). Only after both stages have been completed it is possible to objectively estimate the relevance of the results.

The time required for carrying out the following steps is approximately one hour.

3.1. Collecting the Full Set of Promoters for the Genome of Interest

Before the proper analysis of the gene cluster, we will retrieve the promoter sequences of all the genes of the organism of interest, which will serve below to estimate the background model.

1. Open a connection to the RSAT::Plants server. It can be reached athttp://plants.rsat.euand also at http://floresta.eead.csic.es/rsat. On the left-side menu, select 'Sequence tools -> retrieve sequence'.

2. Choose 'Single organism -> Zea_mays.AGPv3.29' for the examples of this protocol (*see***Note 1**). At the time of publication this corresponds to Ensembl Plants release 29, but that might change over time.

3. Choose 'Genes -> all'; this will retrieve all upstream sequences of the maize genome.

4. Set appropriate upstream bounds. Default values are -2000,-1. To replicate the work of Yu et al ***(4)*** these should be set to 'From' -1000 'To' +200, with position 0 corresponding to transcriptional start sites (TSS). Beware that TSS positions in plant genomes often correspond to start codons, probably due to incomplete annotations.

5. We recommend to tick the option 'Mask repeats', as plant genomes are frequently repeat-rich (see **Fig. 1** and **Table 1**; maize genome contains 78% of REs). This option should not be used if you suspect the transcription factors of interest bind to repeated sequences.

6. Press 'GO' and wait until the retrieve-seq result page is displayed (*see***Note 4**). The results include the executed command and a URL to the 'sequences' file, which must be saved. We will refer to this URL as '**all.fasta.URL**'. This FASTA-format file can also be stored as a local file on your computer, but note it can be rather large (52Mb in this example).

3.2. Analyzing Upstream Sequences of Co-Expressed Genes

We will now retrieve the upstream sequences of a cluster of co-expressed genes, and use *peak-motifs* to discover exceptional motifs in their promoters. The tool *peak-motifs*was initially conceived to discover motifs in ChIP-seq peaks, but it can also be used to analyze other sequence types, as illustrated here.

1. Choose cluster E2F from **Table 2**, copy the corresponding gene IDs (last column) and paste them in a new text file that you will store on your computer. Insert newline characters between genes (*see***Note5**).

2. In the left menu of the RSAT server, click on 'retrieve sequence' to get a fresh form. Make sure that the option 'Genes -> selection' is activated and that the right organism, in this case ‘Zea_mays.AGPv3.29’, is selected. Tick 'Mask repeats', and set the same size limits as for the whole collection of promoters: from -1000 to +200. Paste the list of IDs of your gene cluster (one gene ID per row).

3. Press 'GO' and wait a few seconds until the result page is displayed. Inspection of these sequences might reveal N-masked sequence stretches, which correspond to annotated repeats. Save both

'query genes' and 'sequences' files to local files on your computer, we will refer to them as '**cluster.genes**' and '**cluster.fasta**' later on this protocol.

4. Press the 'peak-motifs' button. The **peak sequences**section is automatically filled with a link to the selected cluster sequences.

5. Add a title for this job, such as 'E2F cluster'.

6. On the right side of ‘Peak sequences’, under **Control sequences**, paste the '**all.fasta.URL**' on the 'URL of a sequence file available on a Web server' entry.

7. Click on 'Reduce peak sequences' and leave both fieldsblank (‘number of top sequences to retain’ and ‘cut peak sequences’)to avoid having the sequences clipped.

8. Click on 'Motif discovery parameters'. Select two algorithms: 'Discover over-represented words' (**oligo-analysis**) and 'Discover over-represented spaced word pairs' (**dyad-analysis**). Uncheck the program **position-analysis** (*see***Note 6**).

9. Click on 'Compare discovered motifs with databases' and select appropriate databases which will be used to annotate any found motifs. For plant promoters, we recommend to check '*footprintDB-plants*', but you can also check other databases such as '*Athamap*', '*ArabidopsisPBM*' and '*JASPAR plants*' (*see***Note 7**). You can also upload your own collection of DNA motifs in TRANSFAC format.

10. Click on ‘Reporting Options’. Set 'Origin' to 'end' and 'Offset' to -200 (*see***Note 8**).

11. Select outputtype (display or email)and press 'GO'.

12. After few seconds the server should have uploaded the sequences and display a page with the URL of the future result page. You can already click on this link: the result page will be

periodically updated to show the progress of the analysis. At the end of the processing, a box will appear at the top of the result page, with a short summary of the discovered motifs, and links to different sections of the results. Once the job is complete click on the link **[Download all results (peak-motifs_archive.zip)]** to **save the results** on your computer. You will later be able to uncompress this archive in order to check the result after its removal from the server (results are only available on the server for 7 days after job completion). We also recommend downloading the full set of discovered motifs, by clicking on the link **[Download all matrices (transfac format)]** and saving a local file named '**cluster.motifs.tf**'. This file contains all motifs in the form of position-weight matrices (PWMs) in TRANSFAC format.

On the result page, the section entitled *'Discovered motifs (with motif comparison)'* lists the discovered motifs, displays their sequence logos and their distribution along clustered sequences, in addition to top matches with the motif databases selected on **step 9**. The top motifs found by *oligo-analysis* and *dyad-analysis*are reported in **Table 4**.

3.3. Negative Control: Random Groups of Genes

In this section, we propose a procedure to obtain an empirical estimation of the rate of false positives, by discovering motifs in the promoters of genes picked up at random.

1. On the left-side menu of RSAT::Plants select 'Build control sets -> random gene selection'.

2. Choose 'Organism -> Zea_mays.AGPv3.29' for the examples of this protocol.

3. Set 'Number of genes' to the size of one of the sample clusters on **Table 2**. For instance, the size of the negative control sets would be 18 for cluster E2F, 16 for cluster ABI4, and 56 for cluster WRI1. For convenience, in this tutorial only one random group is generated (the default), but this utility can generate several random groups in one go (*see***Note 9**).

4. Press 'GO' and click the 'Next step' button 'retrieve sequences' at the bottom of the result page. In the retrieve-seq form, set the other parameters as above: from -1000 to +200, check the 'Mask repeats' option and press 'GO'.

5. Save 'query genes' and 'sequences' files to local '**random.genes**' and '**random.fasta**' filesand repeat steps 4-11 of **section 3.2**. The top motifs found by oligo-analysis and dyad-analysis on such a random cluster are reported on **Table 4**.

3.4. Validating Motifs by Scanning Promoter Sequences

This part of the protocol is devoted to validating sequence motifs discovered by their over-representation, which are scanned against the original sequences from which they were discovered, plus, optionally, orthologous sequences from a related species (*see***Note 10**). The first goal of this section is to check whether the discovered motifs show patterns of occurrence along promoter sequences, and to see how many cluster sequences actually harbor them. This can be done empirically by comparing the results of expression-based motifs with those of shuffled motifs, with columns permuted, which play the role of negative controls. A second goal is to investigate whether these regulatory motifs are conserved on orthologous promoters of a related plant, *Sorghum bicolor* in this case study.

1. On the left-side menu select 'Comparative genomics -> get orthologs-compara'.

2. Choose 'Reference organism -> Sorghum bicolor' for the maize example.

3. Upload file '**cluster.genes**' generated in **step 3** of **section 3.2**. Press 'GO' and finally press 'retrieve sequences' on the next screen.

4. Repeat steps 4-6 of **section 3.1**but now select *Sorghum bicolor* as organism. Save ‘sequences' to local file '**cluster_orths.fasta**'.

5. On the left-side menu select 'Build control sets -> permute-matrix'.

6. Upload '**cluster.motifs.tf**' (obtained in **step 12** of **section 3.2**) and press 'GO'. Save the results file as '**cluster.motifs.perm1.tf**' (*see***Note 11**).

7. Select 'Pattern matching -> matrix scan (full options)'.

8. In the sequence box paste the contents of '**cluster.fasta**' and, optionally, '**cluster_orths.fasta**', if you wish to assess motif conservation. Alternatively, **steps 7-12** can be performed separately with maize and *S.bicolor* sequences.

9. Upload file '**cluster.motifs.tf**' and select 'TRANSFAC' format.

10. In the 'Background model' section select Markov order 2 and choose 'Organism-specific -> Zea_mays.AGPv3.29'. Press 'GO'.

11. Save the 'Scan result' file as '**cluster.scan.ft**' and press the 'feature map' button to draw a map of the matched motif instances.

12. Repeat **steps 6-11** using the set of permuted PWMs '**cluster.motifs.perm1.tf**' and save the results as '**cluster.perm1.scan.tf**'.

3.5. Interpretation of Results

The last stage of the protocol is the interpretation of results, which requires having at hand results of both clusters of co-expressed genes and random clusters, which play the role of negative controls. **Fig.2** summarizes the results of clusters in **Table 2** compared to 50 random clusters of the same size. There are three types of evidence to look at, which will be discussed with the examples on this figure.

- The **distributions of motif significance** yielded by *oligo-analysis* (A,E,I) and *dyad-analysis* (B,F,J). Motifs discovered in random clusters (grey bars) typically have significances below 4 . The motifs found in ABI4 and WRI1 clusters (black bars) are not more significant than those of random gene sets of the same sizes. . The reason for having significant motifs in the random gene sets may result from the occasional presence of low complexity motifs, which should not be considered as reliable predictions. In contrast, the most significant oligomer found within E2F upstream sequences clearly supersedes those of random clusters, and a very similar motif is reported by *dyad-analysis*, with a lower but still strong significance. For this reasons, E2F motifs can be considered as promising predictions.

  Panels A, E and I also show the comparisons between some motifs returned by *peak-motifs* and those reported by the authors of the reference experimental study (Yu et al ***(4)***;they used MEME as motif discovery tool). For E2F and WRI1 the different motif discovery tools return similar motifs (logos) with some differences in the matrix width and in the conservation at some positions. Note that this protocol did not produce any motifs matching the binding sequence reported by Yu et al ***(4)***.

- The **distributions of scanning scores** (C,G,K) show to which extent motif matches in upstream sequences of both maize genes and their *S.bicolor* orthologues (dark boxes) depart from matches of permuted matrices (lighter boxes, *see***Note 11**), used here as negative controls. On these boxplots, the horizontal bars indicate the median score of all the predicted sites in a given set of promoter sequences, and the shaded rectangles show the interquartile range, i.e. the extent between the 25% and 75% percentiles. In the example, the results for E2F motifs confirm their relevance (Fig. 2G): the interquartile range of the E2F cluster (dark rectangle) is clearly separated from the corresponding rectangle of the random selections (gray box). For the ABI4 cluster (Fig2C), there is a noticeable overlap between the interquartile boxes of the cluster and the random gene selections. Besides, the random selections show several “outliers” (circles) indicating sites predicted with high matching scores. Even though the mean scores are clearly higher for the actual cluster, the results may thus not be considered very significant. WRI1 results show a somewhat intermediate situation, where the interquartile boxes show a moderate overlap, but the random gene selections frequently bear relatively high-scoring sites (circles) for the discovered motifs.

- The **distributions of scores in footprintDB** (D,H,L) describe how similar the discovered motifs are when compared to motifs (PWMs) annotated in footprintDB. Similarities are measured by the normalized correlation score (*Ncor, see***Note 12**). In each example 50 random sets of promoterswere analyzed with *peak-motifs*, and the discovered motifs compared to footprintDB. The black bar indicates the best matching score for the original, expression-based gene clusters, and the corresponding logo is overlaid on the histogram. For E2F and WIR1, the best matching motifs correspond to the motifs experimentally confirmed by Yu et al *(4)*. However, in both cases motifs discovered from random gene selections present even better matching scores with some motif database. This result indicates that the matching score between a discovered motif and a repository, while essential for annotation purposes (identifying putative factors for a given gene cluster), is not particularly helpful in order to distinguish relevant expression-supported motifs from PWMs constructed from random sequence clusters. For ABI4, the best-scoring matches correspond to phytochrome interacting factors. These proteins belong to the bHLH family of transcription factors and there are many annotated motifs for them in databases such as footprintDB.

In summary, motifs discovered in promoters of co-expressed genes should always be evaluated based on a combination of complementary criteria:

1. The primary key of interpretation is the significance reported by the motif discovery algorithms. This significance has to be interpreted by comparison with the results obtained in random promoter sets of the same size as the gene cluster of interest (negative controls).
2. Sequence scanning permits to predict putative binding sites, but the matching scores should be evaluated relative to randomized motifs (column-permuted).
3. Comparison between discovered motifs and databases of known TF-binding motifs suggests candidate transcription factors which could intervene in the co-regulation of the co-expressed cluster.

4. Notes

1. As gene models can change from one assembly to another it is important to use the right assembly version, which is indicated for each genome on **Table 1**. If the assembly of interest it not available on RSAT::Plant server, please contact the first author.

2.Twelve clusters of maize genes, found to be co-expressed in 22 transcriptomes and enriched on Gene Ontology terms (http://geneontology.org)***(11)***, were analyzed in detail by Yu et al ***(4)***. First, they discovered potential regulatory motifs within their upstream sequences, and then they performed electrophoretic mobility shift assays (EMSA) to confirm them. Table 2 shows three of those clusters which are used in this protocol. For each cluster a list of gene identifiers is given next to the EMSA-confirmed motifs. The remaining nine clusters were left out for being too small, as the statistical approaches in this protocol require at least ~10-15 genes. Cluster MYB59 was left out due to space restrictions but its results can be browsed at http://plants.rsat.eu/data/chapter_expression_clusters/

3. A crucial parameter to evaluate the results of motif discovery is to estimate the rate of false positives (FP). RSAT programs compute a significance score, which is the minus log of the expected number of false positives (e-value = $10^{-\text{signif}}$). For example, a motif associated with a significance of 1 should be considered as poorly significant, since on average we would expect 10^{-1} = 0.1 false positives, i.e. one FP every 10 random trials. In contrast, a significance of e.g. 16 is very promising, since on average such a result would be expected every 10^{-16} random trials. However, the theoretical significance relies on the correctness of the background model (computed here as k-mer and dyad frequencies in the whole set of promoters). In some cases, sets of plant promoters can discard from the theoretical model, due to heterogeneity of the input (e.g. inclusion of repetitive sequences). The negative control consists in measuring the significance obtained by submitting a random selection of promoters from the

organism of interest (maize in the example). Although each of these genes is likely to be regulated by one or more transcription factors (and its promoter should contain corresponding binding sites), in principle the random set as a whole should not be co-regulated, so that the elements would differ from gene to gene, and there should thus be no over-represented motif in their promoters.

4. Should the connection to the server interrupt it might be safer to go back and choose 'email' as delivery option. The mail message provides a link to the data, which is actually stored at the server.

5. It is crucial to have one gene ID per row for submitting queries to retrieve-seq, because only the first word of each row is considered as a query.

6. This program is generally relevant when analyzing sets containing a large number of sequences such as ChIP-seq peaks or genome-wide promoter sets.

7. Plant transcription databases are unfortunately still very fragmentary, so one might be tempted to check more complete collections such as *footprintDB* or *JASPAR core all*. However, the results should be interpreted with caution, because there is no conservation of *cis*-regulation between plants and other kingdoms of the tree of life.

8. The option *'Origin'* indicates the reference position relative to each sequence (start, center or end). When this option is set to 'end', the coordinates are computed relative to the end of the sequence, with negative values indicating upstream location. The option *'Offset'* enables to shift the reference point by a given number. For the current example, setting the offset to -200 will give coordinates from -1000 to +200, the 0 corresponding to the TSS.

9. Clearly, more than one random cluster should be evaluated, as suggested in **Fig. 2**, where the results of up to 50 random groups are displayed next to the clusters of *(4)*.

10. Orthologues reported are annotated in Ensembl Compara, generated by a pipeline where maximum likelihood phylogenetic gene trees play a central role. These gene trees, reconciled with their species tree, have their internal nodes annotated to distinguish duplication or speciation events, and thus support the annotation of orthologous and paralogous genes, which can be part of complex one-to-many and many-to-many relations. Adapted from: http://www.ensembl.org/info/genome/compara/homology_method.html.

11. This will permute the columns of input PWMs producing matrices with different consensus.Column-permuted matrices are used as negative controls because they conserve the information content and nucleotide frequencies of the original motifs, but at the same time alter the sequence of nucleotides captured by the original motif, which is not recognized anymore.

12. 'Ncor' is the relative width-normalized Pearson correlation of two PWMs aligned with *matrix-scan*. This normalized score prevents spurious matches that would cover only a subset of the aligned matrices (e.g. matches between the last column of the query matrix and the first column of the reference matrix, or matches of a very small motif against a large one).

Acknowledgements

This work was funded in part by Fundación ARAID and by the Enseignants-Chercheursinvités program of Aix-Marseille Université (to BCM).CR was supported by the France Génomique National infrastructure, funded as part of the Investissementsd'Avenir, program managed by the AgenceNationale pour la Recherche (contract ANR-10-INBS-09).

References

1. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117:185–198

2. Sand O, Turatsinze TJ, van Helden J (2008) Evaluating the prediction of cis-acting regulatory elements in genome sequences, In: Frishman, D. and Valencia, A. (eds.) Modern genome annotation, pp. 55–89 Springer

3. Koschmann J, MachensF, BeckerM, NiemeyerJ, SchulzeJ, BülowL, StahlDJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in Arabidopsis. Plant Physiol 160:178–191

4. YuCP, ChenSC, ChangYM, LiuWY, LinHH, LinJJ, ChenHJ, LuYJ, WuYH, LuMY, LuCH, ShihAC, KuMS, ShiuSH, WuSH, LiWH (2015) Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. Proc Natl AcadSci USA 112:E2477–2486

5. SchmidtT, Heslop-HarrisonJ (1998) Genomes, genes and junk: The large-scale organization of plant chromosomes. Trends Plant Sci 3:195–199

6. Kersey PJ, Allen JE, Armean I, et al. (2016) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 44:D574–580

7. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates.Nucleic Acids Res. 42(1):D26-31.

8. Tanaka T, Sakai H, Fujii N, Kobayashi F, Itoh T, Matsumoto T, Wu J. (2013) bexdb: Bioinformatics workbench for comprehensive analysis of barley-expressed genes.Breeding Science 63:430-434.

9. SebastianA, Contreras-MoreiraB (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics 30:258–265

10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. (2013) NCBI GEO: archive for functional genomics data sets--update.Nucleic Acids Res. 41:D991-5.

11. The Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–D1056.

12. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40:D1202-10.

Figure Captions

Figure 1. Genome size of some plant species annotated in RSAT::Plants, showing the fraction of Nsand repeat-masked segments.“Ns” are stretches of uncharacterized nucleotides which often connect assembled sequence contigs. “repeat-masked” segments are sequences with significant similarity to plant repetitive DNA sequences, which are masked in order to calculate background oligonucleotide frequencies. The full dataset is available at http://plants.rsat.eu/data/stats. Most genomes have been downloaded from Ensembl Plants ***(6)***. The yeast genome (*S.cerevisiae*) is plotted as a reference model organism.

Figure 2. Summary of motif discovery results with three clusters of maize genes (ABI4, top; E2F, middle; WRI1, bottom) used along the protocol, see Table 2. Dark bars correspond to clusters of co-expressed genes, grey bars to 50 random clusters of genes drawn from the maize genome. Maximum significance of *oligo-analysis* (A, E, I) and *dyad-analysis* (B, F, J) motifs. The sequence logo of motifs reported by each algorithm is shown on top, indicating the number of sites used to compute it and the Ncor score of the comparison to the expected motif (bottom) (see **Note 11**). Note that the *oligo-analysis* sequence logo of E2F was trimmed to fit in the panel, the original has width=20. Panels C,G,K show the scores of discovered motifs when scanned back to the original maize upstream sequences and sequences from orthologous genes in *Sorghum bicolor*. Here dark bars are the reported PWMs, while the grey bars correspond to permuted PWMs. Panels D,H,L show the Ncor scores of discovered motifs when compared to annotated PWMs in footprintDB. A full report including cluster MYB59 can be browsed at http://plants.rsat.eu/data/chapter_expression_clusters .

Table captions

Table 1. Features of some plant genomes in RSAT::Plants, taken from http://plants.rsat.eu/data/stats. Each ID concatenates the organism, the assembly version and the source. Most genome IDs add to the end the Ensembl Plants release number. For instance, *Arabidopsis_thaliana.TAIR10.29*, corresponds to *A.thaliana* assembly 10 from TAIR (https://www.arabidopsis.org)***(12)***, annotated in release 29 of Ensembl Plants. The yeast genome (*S.cerevisiae*) is listed as a reference."%Ns" are stretches of uncharacterized nucleotides which often connect assembled sequence contigs. "%repeat-masked" segments are sequences with significant similarity to plant repetitive DNA sequences, which are masked.

Table 2. Clusters of maize (*Zea mays*) genes used along the protocol, extracted from the published work of Yu et al *(**4**)*. Experimentally verified regulatory motifs of these clusters are shown.

Table 3. Number of high throughput sequencing expression data sets available at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) as of January, 2016.

Table 4. Top hexamers and dyads enriched on the E2F cluster of maize usptream sequences and a random cluster of the same size. Abbreviations: exp_freq=expected relative frequency, occ=observed occurrences, exp_occ=expected occurrences, occ_P=occurrence probability (binomial), occ_E=E-value for occurrences, occ_sig=occurrence significance.

Tables

Table 1

Organism / assembly ID	Genome size (Mb)	Contigs	%Ns	% repeat-masked	Gene models	Mean upstream length
Aegilops_tauschii.ASM34733v1.29	3,314	429,892	18.8	10.2	37,035	1,856
Amborella_trichopoda.AMTR1.0.29	706	5,745	5.4	12.0	28,721	1,832
Arabidopsis_lyrata.v.1.0.29	207	695	11.1	21.8	32,667	1,411
Arabidopsis_thaliana.TAIR10.29	120	7	0.2	19.3	33,602	1,123
Brachypodium_distachyon.v1.0.29	272	83	0.4	20.1	26,552	1,723
Brassica_oleracea.v2.1.29	489	32,928	8.8	11.0	59,225	1,628
Brassica_rapa.IVFCAASv1.29	284	40,367	3.8	13.9	42,846	1,622
Chlamydomonas_reinhardtii.v3.1.29	120	1,558	12.5	11.1	14,487	1,148
Cyanidioschyzon_merolae.ASM9120v1.29	17	22	0.0	2.2	5,106	804
Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.29	5	1	0.0	0.6	4,497	129
Glycine_max.Wm82.a2.v1.JGI	978	1,190	2.4	43.1	56,044	1,806
Hordeum_vulgare.082214v1.29	4,045	19,705	66.8	9.0	26,066	1,769
Hordeum_vulgare.HarunaNijo.20151026.NIAS	2,006	1,712,261	11.3	50.7	51,249	804
Leersia_perrieri.Lperr_V1.4.29	267	12	0.4	31.3	30,615	1,629
Medicago_truncatula.MedtrA17_4.0.29	413	2,186	5.5	25.3	54,073	1,678
Musa_acuminata.MA1.29	473	12	17.4	9.6	37,579	1,469
Oryza_indica.ASM465v1.29	427	10,490	3.8	7.5	88,438	1,512
Oryza_longistaminata.O_longistaminata_v1.0.30	326	60,198	9.8	24.1	31,686	1,566
Oryza_sativa.IRGSP-1.0.29	374	61	0.0	40.5	91,080	1,444
Ostreococcus_lucimarinus.ASM9206v1.29	13	21	0.0	14.7	7,640	510

Physcomitrella_patens.ASM242v1.29	480	2,106	5.4	47.1	32,273	1,607
Populus_trichocarpa.JGI2.0.29	417	2,518	3.2	32.5	41,377	1,794
Prunus_persica.Prupe1_0.29	227	202	1.2	16.1	29,499	1,635
Saccharomyces_cerevisiae.R64-1-1.29	12	17	0.0	6.4	7,126	423
Selaginella_moellendorffii.v1.0.29	213	759	1.9	36.9	34,888	1,168
Setaria_italica.JGIv2.0.29	406	336	1.2	4.8	35,471	1,673
Solanum_lycopersicum.SL2.50.29	824	3,144	10.4	20.1	38,735	1,724
Solanum_tuberosum.SolTub_3.0.29	811	13	15.8	39.1	42,974	1,763
Sorghum_bicolor.Sorbi1.29	738	3,304	5.5	63.2	34,567	1,773
Theobroma_cacao.Theobroma_cacao_20110822.29	346	711	4.4	20.9	29,188	1,253
Triticum_aestivum.IWGSC1.0+popseq.29	6,483	317,977	3.3	35.6	112,496	1,391
Triticum_urartu.ASM34745v1.29	3,747	499,222	19.7	4.4	37,604	1,806
Vitis_vinifera.IGGP_12x.29	486	33	3.3	39.9	29,971	1,728
Zea_mays.AGPv3.29	2,068	523	0.6	78.2	39,625	1,829

Table 2

Cluster name	Confirmed motif	Number of sequences	Gene IDs
ABI4	GCGCRSGCGGSC	16	GRMZM2G025062 GRMZM2G053503 GRMZM2G069082 GRMZM2G069126 GRMZM2G069146 GRMZM2G076896 GRMZM2G081892 GRMZM2G124011 GRMZM2G129674 GRMZM2G142179 GRMZM2G169654 GRMZM2G172936 GRMZM2G173771 GRMZM2G174347 GRMZM2G175525 GRMZM2G421033
E2F	TTCCCGCCA	18	AC197146.3_FG001 GRMZM2G017081 GRMZM2G021069 GRMZM2G037700 GRMZM2G057571 GRMZM2G062333 GRMZM2G065205 GRMZM2G066101 GRMZM2G075978 GRMZM2G100639 GRMZM2G112074 GRMZM2G117238 GRMZM2G130351 GRMZM2G139894 GRMZM2G154267 GRMZM2G162445 GRMZM2G327032 GRMZM2G450055
WRI1	CGGCGGCGS	56	AC210013.4_FG019 GRMZM2G008430 GRMZM2G009968 GRMZM2G010435 GRMZM2G010599 GRMZM2G014444 GRMZM2G015097 GRMZM2G017966 GRMZM2G022019 GRMZM2G027232 GRMZM2G028110 GRMZM2G035017 GRMZM2G041238 GRMZM2G045818 GRMZM2G047727 GRMZM2G048703 GRMZM2G064807 GRMZM2G068745 GRMZM2G074300 GRMZM2G076435 GRMZM2G078779 GRMZM2G078985 GRMZM2G080608 GRMZM2G092663 GRMZM2G096165 GRMZM2G098957 GRMZM2G107336

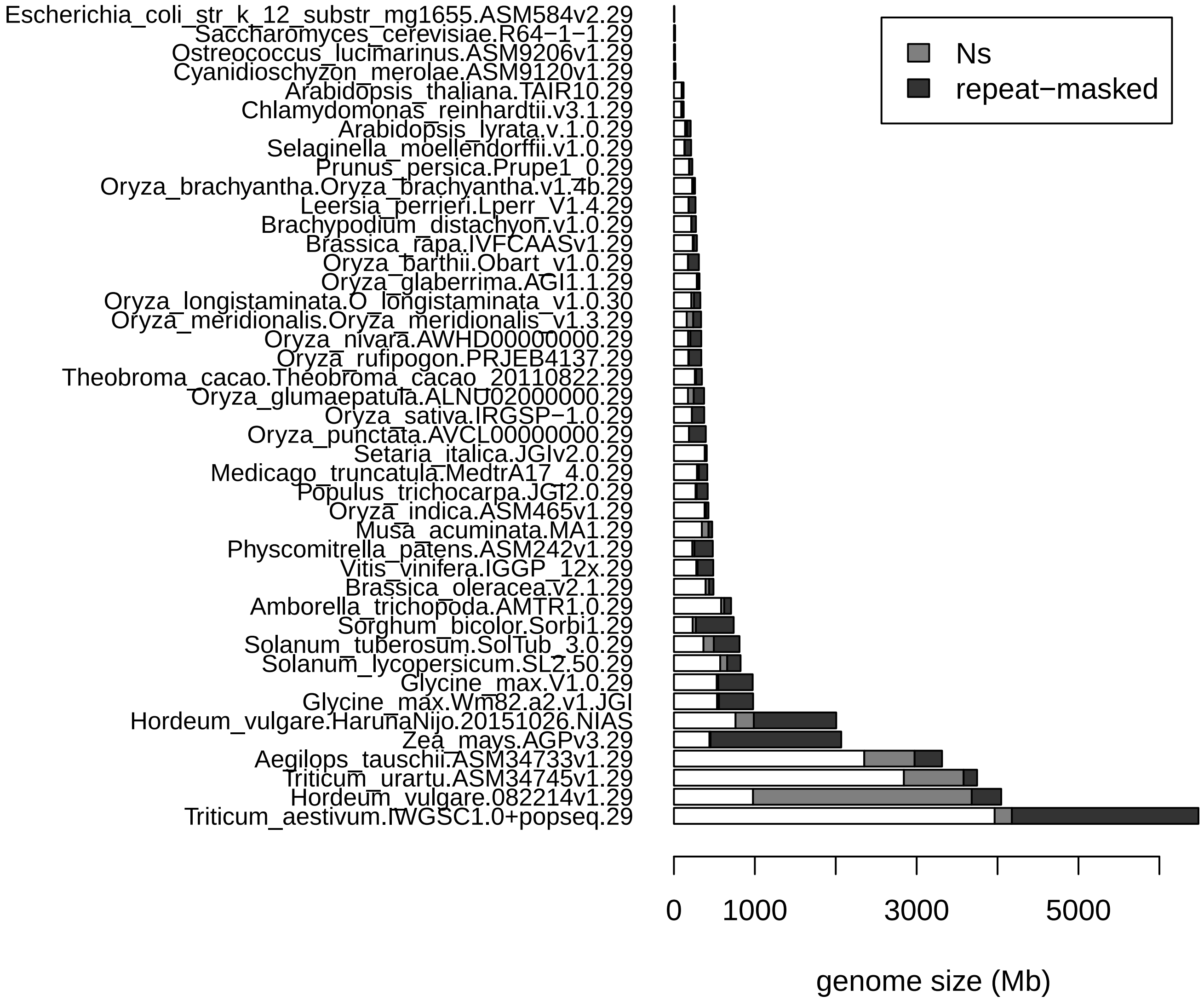
GRMZM2G108348 GRMZM2G111987 GRMZM2G115265 GRMZM2G119865 GRMZM2G122871 GRMZM2G126603 GRMZM2G126928 GRMZM2G132095 GRMZM2G140799 GRMZM2G148744 GRMZM2G150434 GRMZM2G151252 GRMZM2G152599 GRMZM2G170262 GRMZM2G181336 GRMZM2G311914 GRMZM2G312521 GRMZM2G322413 GRMZM2G325606 GRMZM2G343543 GRMZM2G353785 GRMZM2G409407 GRMZM2G439201 GRMZM5G823135 GRMZM5G827266 GRMZM5G831142 GRMZM5G835323 GRMZM5G870606 GRMZM5G882378

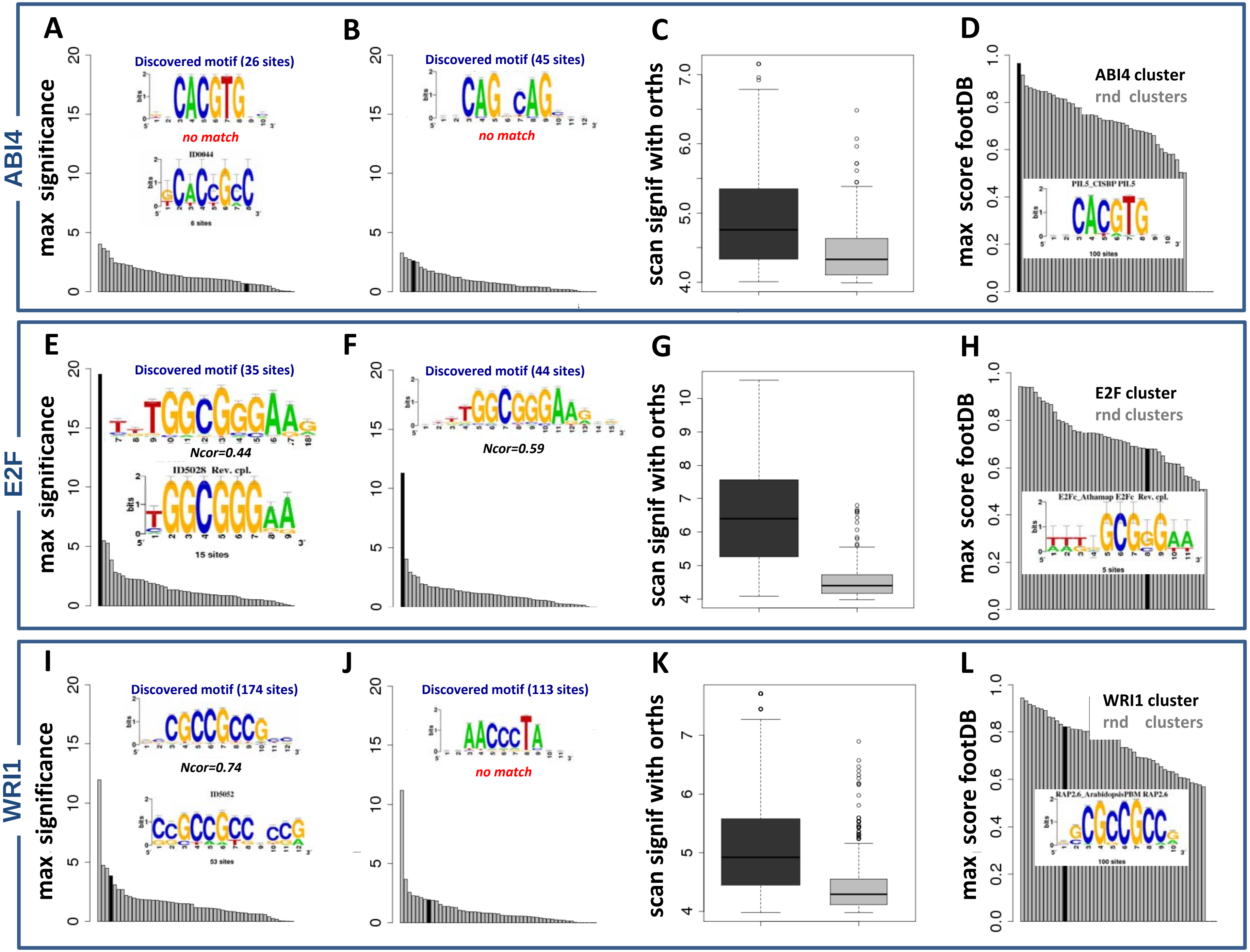
Table 3

Taxon	GEO RNA-seq series
Metazoa	4869
Homo sapiens	1911
Fungi	398
Saccharomyces cerevisiae	167
Viridiplantae	649
Arabidopsis thaliana	235
Zea mays	62
Oryza sativa	51
Bacteria	415
Archaea	12
Total	6378

Table 4

cluster	type	motif	exp_freq	occ	exp_occ	occ_P	occ_E	occ_sig
E2F	hexamer	gcggga	0.00046	37	6.65	3.1e-16	6.5e-13	12.19
E2F	hexamer	cgggaa	0.00031	28	4.55	1.1e-13	2.2e-10	9.66
E2F	hexamer	cccgcc	0.00072	36	10.49	5.7e-10	1.2e-06	5.93
random	hexamer	cttcga	0.00032	15	4.78	0.00014	2.9e-01	0.53
random	hexamer	ccaaaa	0.00083	27	12.16	0.00016	3.4e-01	0.47
random	hexamer	aacacc	0.00046	18	6.78	0.00025	5.2e-01	0.28
E2F	dyad	gcgn{1}gaa	0.00036	31	5.21	1.3e-14	2.6e-10	9.58
E2F	dyad	ggcn{1}gga	0.00062	40	8.79	1.3e-14	2.7e-10	9.57
E2F	dyad	ggcn{2}gaa	0.00042	27	6.00	2.9e-10	6.1e-06	5.22
random	dyad	accn{8}aaa	0.00055	23	7.66	5.7e-06	1.2e-01	0.91
random	dyad	aatn{3}aaa	0.00126	39	17.95	1.1e-05	2.4e-01	0.62
random	dyad	cttn{2}gac	0.00027	15	3.87	1.4e-05	2.9e-01	0.53


Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.29
Saccharomyces_cerevisiae.R64−1−1.29
Ostreococcus_lucimarinus.ASM9206v1.29
Cyanidioschyzon_merolae.ASM9120v1.29
Arabidopsis_thaliana.TAIR10.29
Chlamydomonas_reinhardtii.v3.1.29
Arabidopsis_lyrata.v.1.0.29
Selaginella_moellendorffii.v1.0.29
Prunus_persica.Prupe1_0.29
Oryza_brachyantha.Oryza_brachyantha.v1.4b.29
Leersia_perrieri.Lperr_V1.4.29
Brachypodium_distachyon.v1.0.29
Brassica_rapa.IVFCAASv1.29
Oryza_barthii.Obart_v1.0.29
Oryza_glaberrima.AGI1.1.29
Oryza_longistaminata.O_longistaminata_v1.0.30
Oryza_meridionalis.Oryza_meridionalis_v1.3.29
Oryza_nivara.AWHD00000000.29
Oryza_rufipogon.PRJEB4137.29
Theobroma_cacao.Theobroma_cacao_20110822.29
Oryza_glumaepatula.ALNU02000000.29
Oryza_sativa.IRGSP−1.0.29
Oryza_punctata.AVCL00000000.29
Setaria_italica.JGIv2.0.29
Medicago_truncatula.MedtrA17_4.0.29
Populus_trichocarpa.JGI2.0.29
Oryza_indica.ASM465v1.29
Musa_acuminata.MA1.29
Physcomitrella_patens.ASM242v1.29
Vitis_vinifera.IGGP_12x.29
Brassica_oleracea.v2.1.29
Amborella_trichopoda.AMTR1.0.29
Sorghum_bicolor.Sorbi1.29
Solanum_tuberosum.SolTub_3.0.29
Solanum_lycopersicum.SL2.50.29
Glycine_max.V1.0.29
Glycine_max.Wm82.a2.v1.JGI
Hordeum_vulgare.HarunaNijo.20151026.NIAS
Zea_mays.AGPv3.29
Aegilops_tauschii.ASM34733v1.29
Triticum_urartu.ASM34745v1.29
Hordeum_vulgare.082214v1.29
Triticum_aestivum.IWGSC1.0+popseq.29
Ns
repeat−masked
0
1000
3000
5000
genome size (Mb)

ABI4
A
max significance
0
5
10
15
20
Discovered motif (26 sites)
CACGTG
no match
ID0044
CACCGCC
6 sites
B
Discovered motif (45 sites)
CAG CAG
no match
C
scan signif with orths
4.0
5.0
6.0
7.0
D
max score footDB
0.0
0.2
0.4
0.6
0.8
1.0
ABI4 cluster
rnd clusters
PIL5_CISBP PIL5
CACGTG
100 sites
E2F
E
Discovered motif (35 sites)
TGGCGGGAAG
Ncor=0.44
ID5028 Rev. cpl.
TGGCGGGAA
15 sites
F
Discovered motif (44 sites)
TGGCGGGAAG
Ncor=0.59
G
scan signif with orths
4
5
6
7
8
9
10
H
max score footDB
E2F cluster
rnd clusters
E2Fc_Athamap E2Fc Rev. cpl.
TTTGCGGGAA
5 sites
WRI1
I
Discovered motif (174 sites)
CGCCGCCG
Ncor=0.74
ID5052
CCGCCGCC CCG
53 sites
J
Discovered motif (113 sites)
AACCCTA
no match
K
scan signif with orths
4
5
6
7
L
max score footDB
WRI1 cluster
rnd clusters
RAP2.6_ArabidopsisPBM RAP2.6
GCGCCGCCG
100 sites
bits