PLOS ONE

Nucleotide patterns aiding in prediction of eukaryotic promoters

--Manuscript Draft--

Manuscript Number:	PONE-D-17-24929
Article Type:	Research Article
Full Title:	Nucleotide patterns aiding in prediction of eukaryotic promoters
Short Title:	Nucleotide patterns aiding in prediction of eukaryotic promoters
Corresponding Author:	Tatiana V Tatarinova, PhD University of La Verne La Verne, CA UNITED STATES
Keywords:	deep learning; promoters; classification
Abstract:	Computational analysis of promoters is hindered by their complex architecture. In less studied genomes with complex organization, false positive promoter predictions are common. Accurate identification of TSS and core promoter regions remains an unsolved problem. In this paper, we present a comprehensive analysis of genomic features associated with the promoters and show that probabilistic integrative algorithm models built upon the maps of the distributions of SNPs, RNA sequencing reads on genomic DNA, methylated nucleotides, gene models, TFBS as well as nucleotides and their combinations, allows one achieve accurate DNA sequence classification into "promoter" and "non-promoter", even in absence of the full-length cDNA sequences. Positional clustering of TBFS clearly demonstrates that the cells of Oryza sativa utilize three distinct classes of transcription factors: those that bind preferentially to the [-500,0] region ("promoter-specific"), those that bind preferentially to the [0,500] region ("5′ UTR-specific"), and promiscuous transcription factors, that have weak or no location preference with respect to TSS. For the most informative motifs, their positional preferences were conserved between dicots and monocots.
Order of Authors:	Martin Triska
	Victor Solovyev
	Ancha Baranova
	Alexander Kel
	Tatiana V Tatarinova, PhD
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. **This information is required for submission and will be published with your article, should it be accepted.** A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. **Describe the role** of any sponsors or funders in the study design, data	AK was supported by a grant of the Federal Targeted Program “Research and development on priority directions of science and technology in Russia, 2014-2010”, Contract № 14.604.21.0101, unique identifier of the applied scientific project: RFMEFI60414X0101. AK work was also supported by the following grants of the EU FP7 program: “SYSCOL”, “SysMedIBD”, “RESOLVE” and “MIMOMICS”. TT, AB and MT were supported by the NSF Division of Environmental Biology (1456634) and NSF STTR award 1622840. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

collection and analysis, decision to publish, or preparation of the manuscript. If the funders had **no role** in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*" However, if the study was **unfunded**, please provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work.*" * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	None
Ethics Statement You must provide an ethics statement if your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues.	NA

All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement. **Human Subject Research (involved human participants and/or tissue)** All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript. Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research. **Animal Research (involved vertebrate animals, embryos or tissues)** All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied.	

Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within	All relevant data are within the paper and its Supporting Information files.

the paper and its Supporting Information files." If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: "Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data." "Data are from the XXX study whose authors may be contacted at XXX." * typeset	
Additional data availability information:	Tick here if the URLs/accession numbers/DOIs will be available only after acceptance of the manuscript for publication so that we can ensure their inclusion before publication.

Dear Editors

We have analyzed characteristic features of plant promoter regions, such as drop in DNA methylation and SNP density, peak in TFBS frequency and CG-skew. We analyzed evolutionary conservation of these features and developed a machine learning method for classification of DNA sequence into promoter and non-promoter regions. This careful analysis will be instrumental for development of high confidence TSS prediction method. Accuracy of TSS identification affects the quality of regulatory region analysis. Intelligent integration of multiple types of genomic information (DNA composition, regulatory elements, DNA methylation, RNA-seq coverage data, SNP distribution etc.) can improve annotation of tissue- and developmental stage-specific genes that are often mis-predicted due to atypical sequence composition of stress-related genes in grasses. Integration of multiple noisy features of promoter regions can result in 99% classification accuracy. Features, identified as important by our classification process using deep learning approach, can be used to build a scoring function for promoter prediction.

We hope that our method will be of interest for a wide range of researchers, since the presented approach is easily scalable to other species.

Best regards

Tatiana Tatarinova

Nucleotide patterns aiding in prediction of eukaryotic promoters

Martin Triska

University of Southern California, Los Angeles, CA, USA

martin.triska@usc.edu

Victor Solovyev

Softberry, Inc. Novosibirsk, Russia,

solovictor@softberry.com

Ancha Baranova

George Mason University, USA, and

Research Centre for Medical Genetics, Moscow, Russia

abaranov@gmu.edu

Alexander Kel[#]

geneXplain GmbH, Wolfenbuettel, Germany

Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia,

alexander.kel@genexplain.com

Tatiana V. Tatarinova*[#]

University of La Verne, 1950 3rd St, La Verne, CA 91750,

AA Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia,

ttatarinova@laverne.edu

*#joint senior authors; *corresponding author*

Abstract

Computational analysis of promoters is hindered by their complex architecture. In less studied genomes with complex organization, false positive promoter predictions are common. Accurate identification of TSS and core promoter regions remains an unsolved problem. In this paper, we present a comprehensive analysis of genomic features associated with the promoters and show that probabilistic integrative algorithm models built upon the maps of the distributions of SNPs, RNA sequencing reads on genomic DNA, methylated nucleotides, gene models, TFBS as well as nucleotides and their combinations, allows one achieve accurate DNA sequence classification into “promoter” and “non-promoter”, even in absence of the full-length cDNA sequences.

Positional clustering of TBFS clearly demonstrates that the cells of *Oryza sativa* utilize three distinct classes of transcription factors: those that bind preferentially to the [-500,0] region (“promoter-specific”), those that bind preferentially to the [0,500] region (“5′ UTR-specific”), and promiscuous transcription factors, that have weak or no location preference with respect to TSS. For the most informative motifs, their positional preferences were conserved between dicots and monocots.

Introduction

Promoters are the 5' regions adjacent to the transcriptional start site (TSS) that contain transcription factor binding sites (TFBS) regulating transcription. Each gene contains one or more TSS, and, respectively, one or more promoters, which initiate transcription of a gene. Computational analysis of the eukaryotic promoters is hindered by their complex architecture (Sandelin, Carninci et al. 2007, Solovyev, Shahmuradov et al. 2010, Shahmuradov, Umarov et al. 2017). In 30-60% of eukaryotic genes, depending on species, the proximal part of the promoter, also known as core promotes is contains the TATA motif located approximately 30 nucleotides upstream from the TSS. Most commonly, TATA-containing core promoters are associated with stress-related, tissue-specific and/or highly expressed genes (Troukhan, Tatarinova et al. 2009). Other, TATA-less promoters commonly found in association with broadly expressed genes may have a relatively broad transcription start region (TSR) instead of a pronounced TSS (Tatarinova, Kryshchenko et al. 2014, Shahmuradov, Umarov et al. 2017). To predict the position of the TSR computationally, characteristic

promoter initiation region (Inr) or the downstream promoter element (DPE) may be used (Sandelin, Carninci et al. 2007, Solovyev, Shahmuradov et al. 2010).

Many computational approaches for promoter prediction have been developed. A majority of TSS prediction software tools use sophisticated computational algorithms (e.g. oligonucleotide content-based neural network and linear discriminant approaches), and focus on specific sequence features of promoter region (e.g. TATA-box or CA-motif) (Fickett and Hatzigeorgiou 1997). Genome complexity affects quality of promoter finding methods: for example, presence of several tissue-specific, alternative TSSs negatively affects the prediction accuracy. For the model plant *Arabidopsis thaliana*, modern algorithms identify TATA-containing promoters with sensitivities up to 95% and specificities up to 97% TATA+ promoters (Shahmuradov, Solovyev et al. 2005, Anwar, Baker et al. 2008, Troukhan, Tatarinova et al. 2009, Solovyev, Shahmuradov et al. 2010, Azad, Shahid et al. 2011, Tatarinova, Kryshchenko et al. 2014, Umarov and Solovyev 2017); while for *Homo sapiens* and *Oryza sativa*, prediction accuracy is substantially lower (Umarov and Solovyev 2017). In case of less studied genomes with complex organization (such as the oil palm), false positive as well as false negative error rates can be large, with a false positive promoter prediction occurring once per every 700 - 1000 nucleotides of the genome (Solovyev 2003).

Even the best modern methods of promoter prediction, including genomic sequencing coupled with full-length cDNA capture and ascertainment (Alexandrov, Troukhan et al. 2006, Alexandrov, Brover et al. 2009, Troukhan, Tatarinova et al. 2009), CAGE (Kawaji, Kasukawa et al. 2006, Kawaji, Lizio et al. 2014), 3PEAT(Morton, Petricka et al. 2014), or RAMPAGE (Batut and Gingeras 2013), do not predict TSS positions with the desired 100% accuracy (Tatarinova, Kryshchenko et al. 2014). For example, the mapping of CAGE tags relative to existing human cDNA/mRNA sequences revealed that less than 10% of these tags are located within 10 nucleotides from TSS (Dieterich 2003). When RNA-seq reads covering 12 experimentally validated *O. sativa* promoters were randomly selected from Plant Prom DB (Shahmuradov, Abdulazimova et al. 2012, Shahmuradov, Umarov et al. 2017) and mapped onto the regions TSS+/-1000 nt, the resultant plots showed that the peaks of RNA-Seq coverage and exact positions of known TSS were mismatched, supporting the correct mapping of eukaryotic promoters must be supported by multiple sources of data (Figure 1).

Figure 1: Examples of RNA-seq coverage near 12 randomly selected promoters with experimentally validated transcription start sites.

Two most commonly used techniques to predict eukaryotic promoter by distribution of transcription factors binding sites were proposed in 1995 by Kondrakhin, Kel et al. (1995) and by Prestridge (1995). The method of Kondrakhin, Kel et al. (1995) paired up the detection of TATA boxes with computed weight matrices of transcription factor binding site (TFBS) localizations and achieved much better accuracy in comparison to approaches based on detection of the TATA box alone. Prestridge (1995) combined individual density ratios of all TFBSs into a scoring profile further augmented by a weighted TATA matrix for a TATA box and reported relatively low false positive rate. For both tools, however, applicability remains limited due to insufficiency of reliable for species-specific model training and inability to point at the location of TSS.

In the last decade, several improvements in the promoter prediction were made. Troukhan (2009) combined positional frequency of 5′ EST matches onto genomic DNA with the gene models. This approach, known as TSSer, is, in a nutshell, a deterministic method that predicts one transcription start site per locus. For *Arabidopsis thaliana* promoters, it achieves remarkable accuracy. However, even the most reliable prediction of a single promoter per gene cannot adequately reflect biological complexity underlying its regulation due to common occurrence of alternative promoters, which are often tissue-specific or responsive to the changes in architecture of chromatin (Rye, Sandve et al. 2014). In 2013, the TSSer approach was augmented by non-parametric maximum likelihood approaches to be reborn as NPEST (Tatarinova, Kryshchenko et al. 2014), which allows one to predict the positions of alternative TSSs in the *A. thaliana* genome with better accuracy than the sequences identified in the several "gold standard" databases, such as TAIR (Huala, Dickerman et al. 2001, Berardini, Reiser et al. 2015), Plant Prom DB (Shahmuradov,

Abdulazimova et al. 2012) and Plant Promoter Database (Hieno, Naznin et al. 2014). For example, for the set of 15,875 Arabidopsis promoters derived by both TAIR and NPEST, 11,304 (71%) were predicted within 50 nucleotides of each other and 7,192 (45%) within 10 nucleotides of each other. Thirty percent of TAIR-predicted and 44% of NPEST-predicted promoters identified the "TATA" sequence within the interval [–40, –20] nucleotides upstream from the respective TSS. Nucleotide consensus scores at the TSS (46% of T and 49% of C followed by 65% of A) were stronger for NPEST then for TAIR (43% of T and 35% of C followed by 53% of A). When NPEST predictions were compared to experimentally conformed promoters from other databases, similar data patterns were observed.

Recently, many more types of experimental and computational observations highlighting the TSS positions became available. For example, forty million single nucleotide polymorphisms (SNPs) from the 3,000 Rice Genomes Project (http://snp-seek.irri.org), the largest and highest density SNP collection for any higher plant (Alexandrov, Tai et al. 2015), were shared to facilitate an analysis of genetic variants across the *Oryza sativa* cultivars (Tatarinova, Chekalin et al. 2016). Observed clusters of nucleotide variability were shown to highlight functionally important genomic regions. Interestingly, sharp declines in SNP density were noted about 250 nucleotides upstream of TSS elements; these declines reach their minima exactly at the TSS.

Precise mapping of plant TSSs in the genes with multiple promoters requires incorporation of diverse data types including of tissue/stress specificity of each transcript. Unfortunately, most currently available techniques cannot incorporate diverse data types and simply ignore alternative promoters. Therefore, accurate identification of TSS and core promoter regions remains an unsolved problem. Since evidences for location of TSS are imprecise, the best approach for promoter production should embed probabilistic integrative algorithms. In this paper, we present a comprehensive analysis of genomic features associated with the promoters and show that models built upon the maps of the distributions of SNPs, RNA sequencing reads on genomic DNA, methylated nucleotides, gene models, TFBS as well as nucleotides and their combinations, allows one achieve classification of DNA sequence into "promoter" and "non-promoter", that is accurate even in absence of the full-length cDNA sequences.

RESULTS

Selection of the "gold standard" gene prediction models

Figure 2: Nucleotide consensus around TSS. A) Frequency of CA, B) Frequency of TATA, D) Frequencies of nucleotides A, C, G, T around TSS for Fgenesh, E) Frequencies of A, C, G, T around TSS for MSU

To aid a selection of best available promoter annotation in the rice genome, Fgenesh and MSU mRNA-based gene prediction models were compared. Fgenesh gene prediction set contains 18,389 high quality (5′ full, with mRNA support) gene models, while the MSU gene prediction set contains 20,367 high quality gene models (Kawahara, de la Bastide et al. 2013, Tatarinova, Chekalin et al. 2016). For each gene in each model, we extracted a 1,000 nt long sequence centered at TSS and calculated distributions of genomic features previously associated with the start of transcription: (1) frequency of dinucleotide CA (Shinya and Shimada 1994, Pullen and Friesen 1995, Sandelin, Carninci et al. 2007); (2) frequency of TATA (Kiran, Ansari et al. 2006, Sandelin, Carninci et al. 2007, Troukhan, Tatarinova et al. 2009); (3) nucleotide consensus around TSS (Alexandrov, Troukhan et al. 2006, Alexandrov, Brover et al. 2009, van Heeringen, Akhtar et al. 2011). Figure 2 (C and D) shows that the promoters annotated by Fgenesh shared more pronounced consensus as compared to the promoters annotated by MSU. Fgenesh promoters also had higher frequency of TATA boxes at -30 (B), and more CA dinucleotides at TSS (A). Based on the assumption that these features are reliable quality indicators, Fgenesh annotation was selected for further analysis.

Distribution of transcription factor binding sites

The distributions of transcription factor binding sites (TFBS) in promoters and UTRs of high-confidence rice genes in the regions of -1000 +1000 around TSS around TSS were investigated by utilizing the MATCH software (Kel, Gossling et al. 2003) incorporated in geneXplain platform (www.genexplain.com) and TRANSFAC database (Matys, Kel-Margoulis et al. 2006) comprising 764 plant position weight matrices (PWM) with a strict score threshold of 0.95. MATCH scans the targets promoter sequences with a sliding window equal to the length of the PWM and calculates a score for each of the windows. The maximum value of the score (1.0) corresponds to the sequence that fully fits to the consensus of the PWM. Score threshold of 0.95 allows very little mismatches to the consensus located in the positions with lower degree of conservation

between known binding sites for this TF. The MATCH scores are calculated using an algorithm which considers the nucleotide position-specific entropy measures. In a recent study, MATCH performed with accuracy superior to that of other algorithms (Kondrakhin, Valeev et al. 2016).

In the Fgenesh-predicted rice promoters, MATCH search against the TRASNFAC database revealed 3.2 million potential TF binding sites corresponding to 667 PWMs, while 97 plant PWMs were matchless, possibly due to their exclusive regulatory role in the dicots locations or binding to distal promoters excluded from present study. Interestingly, 487 out of 667 TBFS (73%) were found in proximal promoters of *Oryza sativa* more than 1000 times; the most frequent sites were that for the transcription factors ASR1, DOF56 and PBF. When the frequencies of TFBS found in the proximal promoters were compared with the frequencies for the same PWMs found in randomly shuffled sequences, the most significant promoter-specific enrichments (twice or more) of TFBS were for SPL12, SPL5, GBF1, ABI5, BZIP68, LEC2, and GT1 transcription factors.

Additionally, another set of randomly shuffled sequences was generated using the dinucleotide statistics matching that of the Fgenesh rice promoter regions as described by Stepanova, Tiazhelova et al. (2005). Briefly, the 2000 nt regions [TSS-1000, TSS+1000] were divided onto non-overlapping 100 nt windows, then the dinucleotide statistics were calculated for each window. For each promoter, a 2000 nt long sequence with matching dinucleotide composition was generated, where the positions of TFBS were predicted using the MATCH tool (Kel, Gossling et al. 2003). Kolmogorov-Smirnov tests were applied to find significantly over-represented sequence motifs after selecting only those motifs that occur at least in 100 different rice promoters. Figure 3 shows examples of TBFS that differ and not differ significantly between real and simulated sequences. The most pronounced differences (p-value < 0.002) were detected for the binding sites of TCP15, LIM1, HBP1A, and TCP23. On the other hand, occurrences of binding sites for CMTA2, GATA1, SBF1, and WRKY48 in real and simulated sequences were not different (p-value> 0.99999).

Figure 3: Examples of observed and expected occurrences of TFBS in rice promoters. Different: TCP15, LIM1, HBP1A, TCP23, ARALY493022, AT1G26610, TFIIAL, BZIP910, CBF1, DREB1F, STY1. Observations agree with expectations: CMTA2, GATA1, SBF1, WRKY48.

Positional Specificity of TFBS distribution

A phenomenon of the positional preference in TF binding was previously described by Weirauch, Yang et al. (2014), who showed that, across evolutionary kingdoms, positions of TBFS are not random with respect to the start of transcription. Therefore, we have computed Pearson's correlation coefficients and then clustered the profiles of positional distributions of TFBS across the set of Fgenesh promoters of *Oryza sativa*.

Positional clustering of TBFS clearly demonstrates that the cells of *Oryza sativa* utilize three distict classes of transcription factors: those that bind preferentially to the [-500,0] region ("promoter-specific"), those that bind preferentially to the [0,500] region ("5′ UTR-specific"), and predominantly promiscuous transcription factors with weak or no location preference for respective TSS. Note that some Class 3 TFs cannot be classified as promiscuous (Figure 5), with more regular pattern of positional distribution around the translation start rather than around the transcription start. Examples of the position frequency preference are shown in the Figure 6.

Figure 4: Positional specificity of TFBS distribution

Comparative gene ontology analysis of Class 1, 2 and 3 transcription factors (see

Table 1) showed that Class 1 TFs are enriched in the following GO terms: “sequence-specific DNA binding”, “protein dimerization activity”, “systemic acquired resistance, salicylic acid mediated signaling pathway”, “regulation of transcription from RNA polymerase II promoter”, “response to bacterium”, “jasmonic acid mediated signaling pathway”, “carpel development”, “protein binding”, “negative regulation of defense response”, “protein targeting to membrane”, “regulation of plant-type hypersensitive response”, “plant ovule development”, “response to ozone”. Class 2 TFs were enriched in “DNA binding”, “ethylene-activated signaling pathway”, “response to water deprivation”. Class 3 TFs were enriched in “cellular response to nitrogen levels”.

Figure 5: The distribution pattern for MADSB binding sites highlight ATG codon rather than the respective TSS.

Table 1: GO categories that are significantly different between three TF classes. Total number of genes with GO categories for Class 1, 2 and 3 is 130,164, and 144 correspondingly.

GO	*Class 1*	*Class 2*	*Class 3*	*PVAL*
sequence-specific DNA binding	43	22	61	7.78E-06
protein dimerization activity	21	4	18	0.000438
systemic acquired resistance, salicylic acid mediated signaling pathway	15	2	6	0.000504
regulation of transcription from RNA polymerase II promoter	5	2	16	0.000558
response to bacterium	14	3	4	0.000955
jasmonic acid mediated signaling pathway	15	4	5	0.001897
carpel development	6	1	13	0.002606
protein binding	51	31	35	0.002904
negative regulation of defense response	11	1	5	0.003011
DNA binding	90	146	85	0.007339
protein targeting to membrane	14	4	7	0.010643
regulation of plant-type hypersensitive response	14	4	7	0.010643
ethylene-activated signaling pathway	10	20	4	0.012518
response to water deprivation	48	76	38	0.016138
plant ovule development	12	4	15	0.017373
nucleus	76	126	75	0.017489
response to ozone	9	4	14	0.033868
cellular response to nitrogen levels	13	11	23	0.044478

These three classes of TFs also differ in their expression patterns (Table 2): genes encoding TFs of Class 1 are predominantly expressed in petals, sepals and plant embryos, while mRNAs encoding Class 2 TFs are overrepresented in roots. This observation may explain why TATA boxes were the only cis-element

statistically significantly associated with expression in plant roots (Troukhan, Tatarinova et al. 2009, Triska, Grocutt et al. 2013): possibly, most root-specific transcription factors bind to the 5′ UTR region rather than the region upstream of TSS.

Table 2: Expression specificity of TF from Class 1 and 2

Expression pattern	*Class 1 (99)*	*Class 2 (134)*	*Z-score*
root	66	114	-3.31344
pollen	57	60	1.93163
carpel	72	77	2.39883
seed	70	74	2.40454
leaf lamina base	61	61	2.43144
cauline leaf	64	65	2.4497
collective leaf structure	80	87	2.65976
petal	72	74	2.73045
plant embryo	78	81	2.97259
sepal	76	76	3.17706

Figure 6 shows frequency profiles for TFBSs for ARALY493022_04 (Class 1, left panel), RAP26_03 (Class 2, middle panel), and MYB111_02 (Class 3, right panel). ARALY493022 is basic helix-loop-helix factor, with GGGCCC consensus sequence, its presence in the upstream region is associated with the overall strength of gene expression (Troukhan, Tatarinova et al. 2009, Viola, Uberti Manassero et al. 2011, Weirauch, Yang et al. 2014, Mathelier, Fornes et al. 2016). RAP2.6 is a defense-related, ethylene response transcription factor, recognizing the GCC-box, and showing high binding affinity DNA sequence GCGCCGCCG (Franco-Zorrilla, Lopez-Vidriero et al. 2014). Ali, Abbas et al. (2013) experimentally showed that RAP2.6 works both in tissue-specific and stress-specific manner. Under normal conditions, expression of RAP26 is elevated in roots and stems, while significantly reduced when plant is infected with pathogenic nematodes, such as *H.*

schachtii. To suppress resistance responses of the host plant, the nematodes downregulate expression of RAP2.6. One of the better studied transcription factors with not-position-specific motifs is MYB111, which is involved in the regulation of several genes in flavonoid biosynthesis pathway in cotyledons and leaves (Stracke, Ishihara et al. 2007, Stracke, Jahns et al. 2010) and confers tolerance to UV-B (Stracke, Favory et al. 2010). Binding site MYB111_02 has consensus G[G/T]TAGGT[A/G] (Franco-Zorrilla, Lopez-Vidriero et al. 2014). Such less position-specific motifs often have quite conserved consensus and they could be found in the promoters with relatively low frequency. Such motifs usually provide condition-specific regulation of genes and don't have specific location relative TSS. From this Figure, it is obvious that Class 1 and Class 2 TBFS are useful for delineating potential TSSs, while the mapping of the Class 3 TBFS does not convey positional information that can be used for TSS prediction.

Figure 6: Frequency distribution of TFBS may have different patterns around the start of transcription (position 0 on the horizontal axis). X-axis shows distance from TSS, Y-axis is frequency of motif in each window. Frequencies of ARALY493022_04 TFBS (Class 1) are plotted on the left panel, RAP26_03 TFBS (Class 2) on the middle panel, and MYB111_02 (Class 3) on the right panel.

According to the Kolmogorov-Smirnov test, three classes of TFs differ in the significance of over-representation of their TFBS in promoters as compared to randomly shuffled sequences: in the Class 1 TFs with motifs located predominantly upstream of TSS, 37% of members were significantly overrepresented as pointed by p-values <0.05, while in the Class 2 TFs with TFBS located in 5′ UTRs overrepresentation was confirmed for 20%, and in the Class 3 TFs associated more or less evenly distributed motifs, he over-representation was detected for 15% of class members.

In summary, three classes of transcription factor binding sites differ in their position specificity, significance of detected over-representation, functional classification and patterns of gene expression.

Evolutionary conservation of motif informativeness

We have analyzed evolutionary conservation of the position informativeness (a measure of unevenness of the motif distribution along promoter regions, see Method section) for TFBS motifs in monocots *Oryza sativa* and *Zea mays* and the dicot *Arabidopsis thaliana* (Figure 7). We identified the following correlations

between these measures:

$$I_{corn} = -0.0006 + 0.4422 \times I_{rice}$$

Multiple R^2=0.9786, Adjusted R^2=0.9785, F-statistic: 7732 on 1 and 169 DF, p-value: < 2.2E-16

$$I_{arabidopsis} = 0.001 + 0.6150 \times I_{rice}$$

Multiple R^2=0.6627, Adjusted R^2= 0.6619, F-statistic: 893.8 on 1 and 455 DF, p-value: < 2.2E-16.

Figure 7: Relationship between informativeness of TFBS in rice, corn and Arabidopsis. Each point corresponds to one transcription factor; X axis shows informativeness in rice, Y axis – informativeness in corn and Arabidopsis.

The correlations of the position informativeness for TFBS motifs in two monocots (rice and corn) were higher than that for rice and a dicot plant Arabidopsis. By extracting TFBS with more than 10,000 matches in each of three plant genomes, a list of 171 "common" TFBS was compiled. Each of these TFBS was classified into "promoter-specific" or "5′ UTR-specific". Between rice and corn, 67% of 171 "common" TBFS are consistent in position preference, and 94% of the 90 most informative TBFS in rice had the same positional preference in corn. Between rice and Arabidopsis, 86% of 171 TBFS agreed on their positional preference, while 99% of the 90 most informative TBFS in rice had the same positional preference in Arabidopsis (see Supplemental Data). From these numbers, it appears that rice has better agreement with Arabidopsis than with corn in terms of promoter organization. We hypothesize that this discrepancy is due to less reliable TSS prediction in corn genome as compared to Arabidopsis and rice (Figure 8). Importantly, incorrect prediction of TSS in corn does not affect the informativeness of a TBFS, as defined by deviation from a uniform distribution, but may lead to a systematic "shifting" of the TBFS peaks from upstream of TSS to "5′ UTR specific" and vice versa. In summary, positional preference of the most informative motifs remains conserved between dicots and monocots.

Figure 8: Assessment of promoter prediction quality in Arabidopsis (left) and corn (right). Arabidopsis shows more pronounced consensus at TSS, with higher frequency of TATA at -30 and CA at TSS.

Identification of similar TBFS

Figure 9: An example of five distinct TFBS entries in the TRANSFAC database with very similar position weight matrices (PWMs).

Since TRANSFAC database tends to collect all published motifs, some of its motifs appear to be redundant. For example, several independently published PWMs may be independently built and reported for the same transcription factor (Figure 9). Also, many of the transcription factors that belong to the same protein family may recognize highly similar motifs, which will be reflected by respective PWMs. Figure 9 shows a group of transcription factors with highly similar motifs. Although these transcription factors may differ in their regulatory functions, or participate in differing regulatory networks, for a practical use in promoter prediction, a non-redundant set of these motifs could be made by the clustering based on similarity of their PWMs. By clustering 764 plant PWMs from TRANSFAC, a non-redundant set of 376 sequence motifs was obtained, among which, forty-six were found informative with the scores above 0.0138 (see Supp. Table 1).

SNPs resulting in the TFBS loss and gain

Both the core promoter and the 5′ UTR regions located within 200 bp around the TSS are protected against genomic variations (Figure 10). This protective effect is due to selection constraints on regulatory elements, i.e. TFBS, located near TSS, which prevents their disruption by neutral or near-neutral genomic variants. Analysis of comprehensive collection of plant TBFS (Kondrakhin, Valeev et al. 2016) together with an extensive dataset of the genomic variants in various rice cultivars (Alexandrov, Tai et al. 2015) allowed us to classify regulatory elements of the plants according to their tolerance to the mutations.

Figure 10: Frequency of SNPs located near the TSS in rice.

To achieve that, we considered distribution of SNPs and their effect on TF binding site loss and gain. For each nucleotide change, we have calculated $\Delta = |q - q^*|$ for the TFBS scores before (q) and after (q*) nucleotide change, and compared its value to empirically determined threshold. Calculation of the scores q

and q* was done per the MATCH score formula (see Materials and Methods). If $|\Delta| \geq |\Delta_0|$, the site was considered as "lost" or "gained" depending which score value was larger, q or q*.

Frequencies of site losses and site gains for the promoters and for the random subset of 18,389 intergenic sequences, each 2,000 nt in length, were compared. We hypothesized that functionally important motifs in promoters will have lower amount of variation causing site loss. For each TF, we calculated ratio of the number of site losses in promoters to the number of site losses in intergenic sequences, then the entries in the list were ranked. The binding sites for ABF (CACGTGGC) and CBF4 transcription factors were the most "protected" from the site loss. ABF factors govern osmotic stress response through modulation of the gene expression downstream of SnRK2 kinases in abscisic acid signaling, while CBF4 regulates adaptation to drought. For several important transcription factors, such as MADS8 (involved in the control of flowering time), GT-1 and GATA-1 (response to light), we observed that variation was avoided in positions where nucleotide change can lead to the site gain. Additional data and analysis of SNPs in TFBS is contained in the Table 3, Table 4, and Supplemental Table 5.

The binding sites for AT2G20350 and ARF1 transcription factors were the most "protected" from the site loss. AT2G20350 factors regulate activity of ethylene-activated signaling pathway. The plant hormone ethylene is involved in many aspects of the plant life cycle, including seed germination, root hair development, root nodulation, flower senescence, abscission, and fruit ripening (Johnson and Ecker, 1998). ARF1 is a member of the auxin response factor family, involved in hyperosmotic salinity response. For several important transcription factors, such as WRKY23 (involved in hyperosmotic salinity response and response to auxin), FUS3 (plays a role in embryonic development ending in seed dormancy and response to auxin stimulus), we observed that variation was avoided in positions where nucleotide change can lead to the site gain.

Table 3: Suppression of site loss caused by SNPs in promoters. Frequency intergenic/ Frequency promoters – ratio between frequencies of site loss due to SNPs in intergenic regions and site loss due to SNPs in promoters.

ID	Frequency intergenic/ Frequency promoters	#Promoter Sites	#Intergenic Sites	P-value
P$AT2G20350_01	1.856	1909	3660	6.15E-112
P$ARF1_01	1.814	856	1604	3.30E-47

P$DREBIII4_01	1.677	870	1507	2.78E-35
P$AT2G41690_01	1.648	1072	1825	5.37E-40
P$CBF1_03	1.540	2803	4459	8.89E-74
P$DREB1F_01	1.534	4873	7720	4.52E-124
P$ORA47_01	1.529	4129	6521	8.35E-104
P$RAP210_01	1.527	4519	7128	1.11E-112
P$DEAR3_01	1.526	4525	7134	1.51E-112
P$RAP210_02	1.526	4525	7134	1.51E-112
P$ERF019_01	1.526	4199	6620	1.36E-104
P$RAP21_01	1.526	4236	6675	3.17E-105
P$AT1G71520_01	1.525	4242	6682	3.57E-105
P$DREB1B_01	1.449	2544	3808	1.16E-48
P$HSF3_01	1.423	1262	1855	1.08E-22
P$AT4G16610_01	1.374	8852	12563	7.08E-118
P$AT4G16750_01	1.347	1175	1635	2.62E-15
P$AT2G44940_01	1.338	1231	1701	2.99E-15
P$MADS17_01	1.307	7134	9628	1.37E-66

Table 4: Suppression of site gain caused by SNPs in promoters. Frequency intergenic/ Frequency promoters – ratio between frequencies of site gain due to SNPs in intergenic regions and site gain due to SNPs in promoters.

ID	Frequency intergenic/ Frequency promoters	#Promoter Sites	#Intergenic Sites	P-value
P$WRKY23_01	1.376	1691	2403	2.59E-24
P$FUS3_Q2	1.331	2249	3091	2.07E-25
P$BHLH112_01	1.319	1813	2471	1.14E-19
P$MYB46_02	1.290	1015	1352	4.37E-10
P$WRKY_Q2	1.286	10925	14507	1.56E-88
P$TGA2_Q2	1.275	6140	8084	3.08E-47
P$CDC5_01	1.269	3653	4787	8.44E-28
P$MADS4_01	1.266	1948	2548	1.89E-15

Distribution of RNA-Seq reads

Figure 11: RNA-Seq coverage near the transcription start site

Predictably, an analysis of mapped RNA-seq reads near TSS [-1000; +1000] showed that, on average, coverage peaks are observed immediately downstream of TSS (Figure 11). However, some genes lack a peak of RNS-seq at their TSS. Notably, only 26% of genes display a maximum of the coverage in the range [-50, +250], and only 60% of genes display this maximum in the range [-50, +550].

R-loop forming sequences (RLFS)

Three-stranded nucleic acid R-loop structure is formed between nascent RNA transcript and DNA template (Wongsurawat, Jenjaroenpun et al. 2012). R-loops help prevent methylation of promoters (Ginno, Lott et al. 2012, Ginno, Lim et al. 2013, Sanz, Hartono et al. 2016). Length of the R-loop sequence can be between 150 to 650 nt. R-loop forming sequences are associated with initiation of transcription and other important genic features (Wongsurawat, Jenjaroenpun et al. 2012). R-loops were found to accumulate at the G-rich 5′-UTR regions immediately downstream of the CpG-non-methylated human promoters (Ginno, Lott et al. 2012). We used the QmRRFS tool (Wongsurawat, Jenjaroenpun et al. 2012, Jenjaroenpun, Chew et al. 2015, Jenjaroenpun, Wongsurawat et al. 2015) to map the R-loop forming structures in the area [TSS-1000, TSS+1000]. QmRRFS partitions RLFS three segments, RIZ (DNA region of initiation of R-loops containing at least three contiguous guanines), linker (a spacer up to 50 nt between RIZ and REZ), and REZ (G-rich region supporting extension of R-loop, up to 2000 nt long). According to QmRRFS, 22% of rice genes have at least one RLFS in the area [TSS-1000, TSS+1000], and their distribution is localized to 5′-UTR, in agreement with Ginno, Lott et al. (2012). Distribution of RLFS is unimodal with peak density position around 200 nt downstream from the TSS; over a half of RLFS (52%) are found in the 5′-UTR [TSS, TSS+400] (Supplemental table 6). This position is associated with the polymerase pause region typically occurring after the initiation of transcription (Wongsurawat, Jenjaroenpun et al. 2012, Jenjaroenpun, Chew et al. 2015, Jenjaroenpun, Wongsurawat et al. 2015).

DNA methylation

Figure 12: Methylation around transcription start site in rice in different contexts: red – CG, green - CHG, blue – CHH.

In the intergenic regions and within functional classes of genes and their promoters, the patterns of DNA methylation predictably differ (Tatarinova, Elhaik et al. 2013, Elhaik, Pellegrini et al. 2014). The most pronounced effect was observed for the methylated CpGs (see Figure 12). Intergenic level of CpG methylation was at 0.27, with sharp decline starting around 600 bp upstream of TSS to about 50% of that in intergenic region level at the position of -170, then proceeds to its minimum (0.01) at 8 bp upstream from TSS.

Combining the characteristic features of TSS into promoter classifier

We used 18,389 "promoter" (positives) and 18,389 "non-promoter" (negatives) sequences. To train the model, we used 14,711 positives and negatives each; and for testing 3,678 positives and negatives each. The binary classifier interrogates the candidate sequence and reports whether the sequence is "promoter" or "non-promoter". The best combination of features was: composition of DNA sequence, GC-skew value and presence/absence of the CA-motif in every position. It achieved the best accuracy (0.9995) and has the Matthews correlation coefficient of 0.9989 (see

Table 5). Other features also improve the classification accuracy in comparison with the DNA sequence alone, however, not performing as well as the combination of DNA sequence, GC-skew and CA-motif distribution.

Table 5: Promoter classification accuracy

Features	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	CC
DNA sequence	3424	3030	648	254	0. 8774	0. 9309	0. 8238	0.7591
DNA sequence + CG skew	3635	3653	25	43	0.9907	0.9883	0.9932	0.9832
DNA sequence + CG skew + frequency of CA motif	3674	3678	0	4	0.9994	0.9989	1.0	0.9989
DNA sequence + CG skew + RNA-seq coverage	3658	3666	12	20	0.9956	0.9945	0.9967	0.9913
DNA sequence + CG skew + frequency of TATA motif	3653	3608	70	25	0.9870	0.9932	0.9810	0.9742
DNA sequence + CG skew + DNA methylation	3657	3563	115	21	0.9815	0.9942	0.9687	0.9633
DNA sequence + all TFBS	3241	3386	292	437	0.9009	0.8812	0.9206	0.8024
DNA sequence + all TFBS +CG skew	3619	3668	10	59	0.9906	0.9839	0.9973	0.9813
DNA sequence + selected TFBS+CG skew	3628	3674	4	50	0.9927	0.9864	0.9989	0.9854
DNA sequence + SNP	3430	3138	540	248	0.8929	0.9326	0.8532	0.7882
DNA sequence + SNP+CG skew	3348	3296	382	330	0.9032	0.9103	0.8962	0.8065

DISCUSSION

We have investigated several features of promoter area and identified characteristic patterns of their distribution in this area and assessed their utility for identification of TSS location. Promoter-UTR boundary is marked by the following pronounced trends: (1) drop in SNP density, (2) evolutionary conserved peaks and valleys of positions of regulatory elements, (3) peak of RNA-Seq coverage immediately downstream from the TSS, (4) peak of CG skew, (5) drop in DNA methylation density in CpG, CHH and CHG contexts.

Analysis of SNPs in the context of TBFS in promoter and non-promoter region indicated that TBFS differ by tolerance to polymorphisms. The binding sites for AT2G20350 and ARF1 transcription factors were the most "protected" from the site loss. AT2G20350 factors regulate activity of ethylene-activated signaling pathway. The plant hormone ethylene is involved in many aspects of the plant life cycle, including seed germination, root hair development, root nodulation, flower senescence, abscission, and fruit ripening (Johnson and Ecker 1998). ARF1 is a member of the auxin response factor family, involved in hyperosmotic salinity response. For several important transcription factors, such as WRKY23 (involved in hyperosmotic salinity response and response to auxin), FUS3 (plays a role in embryonic development ending in seed dormancy and response to auxin stimulus), we observed that variation was avoided in positions where nucleotide change can lead to the site gain Binding sites for several important transcription factors, such as AT2G20350 and ARF1 found to be significantly less frequently lost due to polymorphic mutations then it is expected by random chance. We can conclude that sites for such transcription factors are "protected" in evolution from being lost since their high general importance for regulation of plant genes. Indeed, these transcription factors are of utmost importance. They are involved in regulation of many plant developmental processes in response to the plant hormones ethylene and auxin. Both these hormones play a central role in coordination of many growth and behavioral processes in plants and are essential for plant body development. It was interesting to observe that for several transcription factors nucleotide variations were avoided in positions where nucleotide change can lead to the site gain. Among such factors we detect for instance WRKY23 and FUS3. We propose that sporadic creation of new sites for such transcription factors can significantly alter cellular timing, therefore such mutations are avoided in these classes of TBFS. It is

interesting that these factors are also involved in gene regulation in response to the plant hormone auxin. We conclude that TRANSFAC analysis results in interesting observations of the architecture of rice promoters and provide clear avoidance of interplay between SNPs and TF binding sites in rice genome.

Accuracy of TSS identification affects the quality of regulatory region analysis. Intelligent integration of multiple types of genomic information (DNA composition, regulatory elements, DNA methylation, RNA-seq coverage data, SNP distribution etc.) can improve annotation of tissue- and developmental stage-specific genes that are often mis-predicted due to atypical sequence composition of stress-related genes in grasses (Tatarinova 2010, Tatarinova, Elhaik et al. 2013, Elhaik, Pellegrini et al. 2014). Integration of multiple noisy features of promoter regions can result in 99% classification accuracy. Features, identified as important by our classification process using deep learning approach, can be used to build a scoring function for promoter prediction.

MATERIALS AND METHODS

Fgenesh++ Rice Gene Prediction

Fgenesh++ (Find genes using Hidden Markov Models) (Yao, Guo et al. 2005, Bajic, Brent et al. 2006, Solovyev, Kosarev et al. 2006) is a HMM-based *ab initio* gene prediction program (Salamov and Solovyev 2000). We used the rice chromosomes (version MSU 7, (Kawahara, de la Bastide et al. 2013)) to make the initial gene prediction set, applying the Fgenesh gene finder with generic parameters for monocot plants. From this set, we selected a subset of predicted genes that encode highly homologous proteins (using BLAST with E-value cut-off 1.0E-10) to known plant proteins from the NCBI non-redundant (NR) database. Based on this subset, we computed gene-finding parameters, optimized for the rice genome, and executed the Fgenesh++ pipeline to annotate the genes in the genomic scaffolds. The Fgenesh++ pipeline used all available supporting data, such as known transcripts and homologous protein sequences. NR plant and, specifically, rice transcripts were mapped to the rice genomic sequences, therefore identifying a set of potential splice sites. Plant proteins were mapped to the rice genomic contigs, and the high scoring matches were selected to generate protein-supported gene predictions, so that only the highly homologous proteins were used in gene identification.

Amino acid sequences from predicted rice genes were then compared to the protein sequences from plant NR database using the 'bl2seq' routine, and the similarity was significant if it had a blast percent identity ≥ 50, blast score ≥ 100, coverage of predicted protein ≥ 80% and coverage of homologous protein ≥ 80%. BLAST analysis of the predicted sequences was also carried out against the *O. sativa* mRNA dataset, using an identify cutoff of >90%. Predictions that have both NR plant RefSeq and *O. sativa* mRNA support, as well as the 5′ UTR longer than 20 nucleotides and shorter than 1000 were selected for the analysis.

GFF file with Fgenesh++ gene prediction is available as a Supplemental Data file.

MSU Rice Gene Models

The current MSUv7 annotation (http://rice.plantbiology.msu.edu) of rice genome contains 55,986 predicted genes and 66,338 gene models (Kawahara, de la Bastide et al. 2013). Upon exclusion of pseudogenes, transposable elements, and genes with atypical lengths of 5′ UTR (below 20 nt or above 1000 nt long), a high-confidence set contains 20,367 expressed protein-coding rice genes.

Arabidopsis Gene and Promoter Models

Genome annotation files for TAIR 10 version and sequences for 3000 nucleotides upstream from ATG were obtained from The Arabidopsis Information Resource (TAIR) (Lamesch, Berardini et al. 2012, Berardini, Reiser et al. 2015). The upstream sequences were truncated based on the position of the nearest upstream locus. 290,085 EST sequences were obtained from NCBI and TAIR and mapped onto the 27,199 upstream sequences using nucleotide BLAST + (minimum identity percent: 95%; maximum query start of alignment: 5; only plus strand alignments were used). Using the text search, we removed ESTs annotated as 3′ or partial. NPEST (Tatarinova, Kryshchenko et al. 2014) algorithm was used and resulted in prediction of 17,452 transcription start sites for 16,520 protein-coding loci.

Corn Gene and Promoter Models

Genome annotation of maize (B73, 6a) contains 40,602 predicted protein-coding genes (Law, Childs et al. 2015). We excluded genes with atypical lengths of 5′ UTR (below 20 nt or above 1000 nt long), genes

without full-length mRNA support, without valid start and stop codon, or no PFAM annotation. This filtering resulted in 16,180 putative corn TSS.

Positional Informativeness of Transcription Factor Binding Sites

We selected TFBS that occur at last 10,000 times in promoters of a given species. In rice, it amounted to 487, in Arabidopsis -559, and in corn - 171 TBFS. To calculate informativeness of each TFBS, we divided the region around the start of transcription (TSS-1000, TSS+1000) into 100 nt long non-overlapping windows, and calculated the observed frequency of TFBS matches in every window as a ratio of matches within the window to the total number of matches $f_o = \frac{m}{T}$. The expected frequency is calculated as $f_e = \frac{1}{Number\ of\ windows}$. The informativeness defined as $I = \sum_{windows}(f_o - f_e)^2$. The binding sites were ranked from highest to lowest informativeness.

RNA-seq Data

We used following publicly available RNA-seq datasets: SRR034580, SRR034581, SRR034582, SRR034583, SRR034584, SRR034585, SRR034586, SRR034587, SRR034588, SRR034589, SRR034590, SRR034591, SRR034592, SRR034593, SRR034594, SRR034595, SRR034596, SRR034597, SRR034598, SRR034599, SRR042529, SRR074125, SRR074126, SRR074127, SRR074128, SRR074129, SRR074130, SRR074131, SRR074132, SRR074133, SRR074134, SRR074135, SRR074136, SRR074137, SRR074139, SRR074140, SRR074142, SRR074143, SRR074144, SRR074145, SRR074146, SRR074147, SRR074149, SRR074150.

The datasets were processed using the following protocol:

1. Duplicates were removed using tool *clumpify* (http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/clumpify-guide/) allowing for up to two errors per read.
2. Quality trimmed using *trimmomatic* (Bolger, Lohse et al. 2014) with minimum read length=16, minimum quality 28 (sliding window of length 10)
3. Aligned to the MSU 7 rice genome using *Hisat2* (Pertea, Kim et al. 2016) aligner.

Summary statistics is shown in the Table 6.

Table 6: RNA-seq dataset quality.

Experiments	Reads	Read length	Quality	Aligned
SRR034580-SRR034599	~5.5 M	35	Poor	67-72%
SRR042529	8.5M	36	Good	84%
SRR074125-SRR074150	2-5M	26	Good	~1.5% (!)

Identification of Transcription factor binding sites

The prediction of TF binding sites is done using the MATCH tool, which is based on the usage of information vector-based PWM model. This model calculates the *matrix similarity score* (mSS) defined in (Kel, Gossling et al. 2003). This model is a common additive model, which uses a transformed matrix instead of an initial matrix, where each column of the transformed matrix is determined with the help of weighting the corresponding initial column by information content. More specifically, the j^{th} column of the weight matrix is equal (up to the constant $\frac{-MIN}{MAX-MIN}$) to the product of the j^{th} column of the frequency matrix and the value $\frac{I(j)}{MAX-MIN}, i=1,...,l$ where *I(j)*, *MIN*, and *MAX* were defined in Kel, Gossling et al. (2003).

Site loss and gain

We analyzed distribution of SNPs and their effect on TF binding site loss and gain. The effect of a SNP on TF binding sites was computed as the follows. For each SNP and for each PWM model we computed two TFBS scores: q and q^* corresponding to two nucleotides in the SNP – the reference and alternative nucleotides. The score was calculated as: $q=\frac{\sum_{i=1}^{l} I(i) f(b_i,i) - \sum_{i=1}^{l} I(i) f^{\min}(i)}{\sum_{i=1}^{l} I(i) f^{\max}(i)}$, where

$$I(i) = \sum_{b \in \{A,T,G,C\}} f(b,i) \ln(4 f(b,i)) .$$

Next, we calculated $\Delta = |q - q^*|$, and compared its value to empirically determined threshold. If $|\Delta| \geq |\Delta_0|$, the site was considered as "lost" or "gained" depending on sign of the Δ.

We then calculated frequencies of site loss and site gain for all considered SNPs to identify loss or gain of which transcription factor binding sites (TFBS) are significantly enriched by the effect of nucleotide changes in SNPs analyzed. As a background, we considered random nucleotide changes in random genomic positions. We denote study and background sets briefly as "*Yes"* and "*No"* sets (the "*Yes"* set is the set of TFBS sequences overlapping SNPs with either the reference nucleotide or alternative nucleotide; the "*No"* set is the set created by random nucleotide substitutions in random genomic positions). The algorithm for TFBS enrichment analysis, called F-Match, has been described in Kel, Konovalova et al. (2006) and Koschmann, Bhar et al. (2015). Briefly, the procedure finds a critical value (a threshold) for the differences between scores q and q* ($the\ threshold\ \Delta_0$) of each PWM in the library that maximizes the *"Yes/No"* ratio R_{YN} as defined in Equation (1) under the constraint of statistical significance:

$$R_{YN} = \frac{\#Sites_{Yes} / \#Sites_{No}}{\#Seq_{Yes} / \#Seq_{No}} \quad (1)$$

In Equation (1), *#Sites* and *#Seq* are the sites and sequences counted in "*Yes"* and "*No"* sets. A high "*Yes/No"* ratio indicates strong enrichment of binding sites for a given PWM in the "*Yes"* sequences. The statistical significance is computed as follows:

$$\begin{aligned} P(X \geq x) &= \sum_{n=x}^{N} \binom{N}{n} \cdot p^n \cdot (1-p)^{N-n} \\ p &= \#Seq_{Yes} / (\#Seq_{Yes} + \#Seq_{no}) \\ N &= \#Sites_{Yes} + \#Sites_{No} \\ n &= \#Sites_{Yes} \end{aligned} \quad (2)$$

The Yes/No ratio and P-value is computed separately for the site gain and for the site loss. If "*Yes/No"*

ratio >1 and a P-value < 0.01 for a given PWM we consider this as an indication of an enrichment of SNPs by the sites for the given PWM. We can say that sites of this PWM are frequently effected by the SNPs and therefore the gene regulation by the respective TFs is significantly altered by the considered SNPs.

Matrix clustering

Many matrices in the TRANSFAC database are highly similar, up to the point being undistinguishable. To lower the complexity of the training data, we performed hierarchical clustering and used only one matrix from each cluster for promoter classification. The distance between two motifs is calculated as sum of squared differences between all matrix elements. If matrices were not the same size, we slide the shorter matrix over the longer one and take minimal distance. The cut-off for merging clusters was determined empirically by considering the sequence logos of matrices to be merged at each step and deciding which matrices we consider duplicates.

Classification of Promoter Regions

There are many network architectures and the task is to choose a suitable one for a given research problem. We used Convolutional Neural Networks (CNN) architecture for building promoter recognition models developed by Umarov and Solovyev (2017). The software consists of several modules. In the ***learnCNN.py*** modules the CNN model was implemented using *Keras* - a minimalist, highly modular neural networks library, written in Python. It uses the *Theano* library as a backend and utilizes GPU for fast neural network training. *Adam* optimizer was used for training with categorical cross-entropy as a loss function. Our CNN architecture (Figure 13) consists of one convolutional layer with 200 filters of length 21. After the convolutional layer, there is a standard Max-Pooling layer. The output from the Max-Pooling layer is fed into a standard fully connected ReLU layer with 128 neurons. Pooling size was equal to 2. The ReLU layer is connected to the output layer with sigmoid activation, where neurons correspond to promoter and non-promoter classes. The batch size used for training was 16.

Figure 13: Basic CNN architecture that was used in building promoter models implemented in the

learnCNN.py *program (Shahmuradov, Umarov et al. 2017, Umarov and Solovyev 2017).*

Input of the network consisted of nucleotide sequences where each nucleotide is encoded by a four-dimensional vector A (1,0,0,0), T (0,1,0,0), G (0,0,1,0) and C (0,0,0,1) and other dimensions filled by other promoter features such as: GC-skew, DNA methylation, SNP, presence of CA motif, presence of TATA motifs, TFBS. The output is a two-dimensional vector: "promoter" (1, 0) and "non-promoter" (0, 1) prediction. ***learnCNN.py*** learns parameters of the CNN model and outputs the accuracy of promoter prediction for the test set of sequences. It also writes the computed CNN Model into a file, which can be used later in programs for promoter identification in each sequence. We used 70% of these examples for learning, 10% for validation (to find an optimal number of learning epochs) and 20% for testing.

We have extracted 18,389 sequences around transcription start site determined by full-length mRNA. Sequence [TSS-199, TSS+50], containing 200 nucleotides from promoter and 50 nucleotides from 5' UTR, was designated as the "promoter" region, and sequence [TSS+751, TSS+1000], from the coding part of the gene, as "non-promoter".

Quality of prediction was assessed using the following measures: True Positives (TP), True Negatives (TN), False Positive (FP), False Negative (FN), Accuracy, Sensitivity, Specificity, Matthews correlation coefficient (CC):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$CC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Acknowledgments

554 TT and MT were supported by the NSF Division of Environmental Biology (1456634) and NSF STTR award
555 1622840. AB was supported by NSF STTR award 1622840. AK was supported by a grant of the Federal
556 Targeted Program “Research and development on priority directions of science and technology in Russia,
557 2014-2010”, Contract № 14.604.21.0101, unique identifier of the applied scientific project:
558 RFMEFI60414X0101. AK was also supported by the following grants of the EU FP7 program: “SYSCOL”,
559 “SysMedIBD”, “RESOLVE” and “MIMOMICS”.

560

References

Alexandrov, N., S. Tai, W. Wang, L. Mansueto, K. Palis, R. R. Fuentes, V. J. Ulat, D. Chebotarov, G. Zhang, Z. Li, R. Mauleon, R. S. Hamilton and K. L. McNally (2015). "SNP-Seek database of SNPs derived from 3000 rice genomes." Nucleic Acids Res **43**(Database issue): D1023-1027.

Alexandrov, N. N., V. V. Brover, S. Freidin, M. E. Troukhan, T. V. Tatarinova, H. Zhang, T. J. Swaller, Y. P. Lu, J. Bouck, R. B. Flavell and K. A. Feldmann (2009). "Insights into corn genes derived from large-scale cDNA sequencing." Plant Mol Biol **69**(1-2): 179-194.

Alexandrov, N. N., M. E. Troukhan, V. V. Brover, T. Tatarinova, R. B. Flavell and K. A. Feldmann (2006). "Features of Arabidopsis genes and genome discovered using full-length cDNAs." Plant Mol Biol **60**(1): 69-85.

Ali, M. A., A. Abbas, D. P. Kreil and H. Bohlmann (2013). "Overexpression of the transcription factor RAP2.6 leads to enhanced callose deposition in syncytia and enhanced resistance against the beet cyst nematode Heterodera schachtii in Arabidopsis roots." BMC Plant Biol **13**: 47.

Anwar, F., S. M. Baker, T. Jabid, M. Mehedi Hasan, M. Shoyaib, H. Khan and R. Walshe (2008). "Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach." BMC Bioinformatics **9**: 414.

Azad, A. K., S. Shahid, N. Noman and H. Lee (2011). "Prediction of plant promoters based on hexamers and random triplet pair analysis." Algorithms Mol Biol **6**: 19.

Bajic, V. B., M. R. Brent, R. H. Brown, A. Frankish, J. Harrow, U. Ohler, V. V. Solovyev and S. L. Tan (2006). "Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment." Genome Biol **7 Suppl 1**: S3 1-13.

Batut, P. and T. R. Gingeras (2013). "RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs." Curr Protoc Mol Biol **104**: Unit 25B 11.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait and E. Huala (2015). "The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome." Genesis **53**(8): 474-485.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Dieterich, C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003). "CORG: a database for COmparative Regulatory Genomics." Nucleic Acids Res. **31**: 55–57.

Elhaik, E., M. Pellegrini and T. V. Tatarinova (2014). "Gene expression and nucleotide composition are associated with genic methylation level in Oryza sativa." BMC Bioinformatics **15**: 23.

Fickett, J. W. and A. G. Hatzigeorgiou (1997). "Eukaryotic promoter recognition." Genome Res **7**(9): 861-878.

Franco-Zorrilla, J. M., I. Lopez-Vidriero, J. L. Carrasco, M. Godoy, P. Vera and R. Solano (2014). "DNA-binding specificities of plant transcription factors and their potential to define target genes." Proc Natl Acad Sci U S A **111**(6): 2367-2372.

Ginno, P. A., Y. W. Lim, P. L. Lott, I. Korf and F. Chedin (2013). "GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination." Genome Res **23**(10): 1590-1600.

Ginno, P. A., P. L. Lott, H. C. Christensen, I. Korf and F. Chedin (2012). "R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters." Mol Cell **45**(6): 814-825.

Hieno, A., H. A. Naznin, M. Hyakumachi, T. Sakurai, M. Tokizawa, H. Koyama, N. Sato, T. Nishiyama, M. Hasebe, A. D. Zimmer, D. Lang, R. Reski, S. A. Rensing, J. Obokata and Y. Y. Yamamoto (2014). "ppdb: plant promoter database version 3.0." Nucleic Acids Res **42**(Database issue): D1188-1192.

Huala, E., A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, L. A. Mueller, D. Bhattacharyya, D. Bhaya, B. W. Sobral, W. Beavis, D. W. Meinke, C. D. Town, C. Somerville and S. Y. Rhee (2001). "The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant." Nucleic Acids Res **29**(1): 102-105.

Jenjaroenpun, P., C. S. Chew, T. P. Yong, K. Choowongkomon, W. Thammasorn and V. A. Kuznetsov (2015). "The TTSMI database: a catalog of triplex target DNA sites associated with genes and regulatory

elements in the human genome." Nucleic Acids Res **43**(Database issue): D110-116.
Jenjaroenpun, P., T. Wongsurawat, S. P. Yenamandra and V. A. Kuznetsov (2015). "QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences." Nucleic Acids Res **43**(20): 10081.
Johnson, P. R. and J. R. Ecker (1998). "The ethylene gas signal transduction pathway: a molecular perspective." Annu Rev Genet **32**: 227-254.
Kawahara, Y., M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell and T. Matsumoto (2013). "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data." Rice (N Y) **6**(1): 4.
Kawaji, H., T. Kasukawa, S. Fukuda, S. Katayama, C. Kai, J. Kawai, P. Carninci and Y. Hayashizaki (2006). "CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis." Nucleic Acids Res **34**(Database issue): D632-636.
Kawaji, H., M. Lizio, M. Itoh, M. Kanamori-Katayama, A. Kaiho, H. Nishiyori-Sueki, J. W. Shin, M. Kojima-Ishiyama, M. Kawano, M. Murata, N. Ninomiya-Fukuda, S. Ishikawa-Kato, S. Nagao-Sato, S. Noma, Y. Hayashizaki, A. R. Forrest, P. Carninci and F. Consortium (2014). "Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing." Genome Res **24**(4): 708-717.
Kel, A., T. Konovalova, T. Waleev, E. Cheremushkin, O. Kel-Margoulis and E. Wingender (2006). "Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations." Bioinformatics **22**(10): 1190-1197.
Kel, A. E., E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis and E. Wingender (2003). "MATCH: A tool for searching transcription factor binding sites in DNA sequences." Nucleic Acids Res **31**(13): 3576-3579.
Kiran, K., S. A. Ansari, R. Srivastava, N. Lodhi, C. P. Chaturvedi, S. V. Sawant and R. Tuli (2006). "The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants." Plant Physiol **142**(1): 364-376.
Kondrakhin, Y., T. Valeev, R. Sharipov, I. Yevshin, F. Kolpakov and A. Kel (2016). "Prediction of protein-DNA interactions of transcription factors linking proteomics and transcriptomics data." EuPA Open Proteomics **13**: 14–23.
Kondrakhin, Y. V., A. E. Kel, N. A. Kolchanov, A. G. Romashchenko and L. Milanesi (1995). "Eukaryotic promoter recognition by binding sites for transcription factors." Comput Appl Biosci **11**(5): 477-488.
Koschmann, J., A. Bhar, P. Stegmaier, A. E. Kel and E. Wingender (2015). ""Upstream Analysis": An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data." Microarrays (Basel) **4**(2): 270-286.
Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel and E. Huala (2012). "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools." Nucleic Acids Res **40**(Database issue): D1202-1210.
Law, M., K. L. Childs, M. S. Campbell, J. C. Stein, A. J. Olson, C. Holt, N. Panchy, J. Lei, D. Jiao, C. M. Andorf, C. J. Lawrence, D. Ware, S. H. Shiu, Y. Sun, N. Jiang and M. Yandell (2015). "Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes." Plant Physiol **167**(1): 25-39.
Mathelier, A., O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin and W. W. Wasserman (2016). "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles." Nucleic Acids Res **44**(D1): D110-115.
Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender (2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." Nucleic Acids Res **34**(Database issue): D108-110.
Morton, T., J. Petricka, D. L. Corcoran, S. Li, C. M. Winter, A. Carda, P. N. Benfey, U. Ohler and M.

Megraw (2014). "Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures." Plant Cell **26**(7): 2746-2760.
Pertea, M., D. Kim, G. M. Pertea, J. T. Leek and S. L. Salzberg (2016). "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown." Nat Protoc **11**(9): 1650-1667.
Prestridge, D. S. (1995). "Predicting Pol II promoter sequences using transcription factor binding sites." J Mol Biol **249**(5): 923-932.
Pullen, S. S. and P. D. Friesen (1995). "The CAGT motif functions as an initiator element during early transcription of the baculovirus transregulator ie-1." J Virol **69**(6): 3575-3583.
Rye, M., G. K. Sandve, C. O. Daub, H. Kawaji, P. Carninci, A. R. Forrest, F. Drablos and F. consortium (2014). "Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines." BMC Genomics **15**: 120.
Salamov, A. A. and V. V. Solovyev (2000). "Ab initio gene finding in Drosophila genomic DNA." Genome Res **10**(4): 516-522.
Sandelin, A., P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki and D. A. Hume (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-436.
Sanz, L. A., S. R. Hartono, Y. W. Lim, S. Steyaert, A. Rajpurkar, P. A. Ginno, X. Xu and F. Chedin (2016). "Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals." Mol Cell **63**(1): 167-178.
Shahmuradov, I. A., A. Abdulazimova, F. Z. Khan, V. Solovyev, N. Mustafaev, Y. Akbarova, R. Qamar and J. Aliyev (2012). The PlantProm DB: Recent Updates. 2012 International Conference on Biomedical Engineering and Biotechnology (iCBEB). IEEE. Macau, Macao.
Shahmuradov, I. A., V. V. Solovyev and A. J. Gammerman (2005). "Plant promoter prediction with confidence estimation." Nucleic Acids Res **33**(3): 1069-1076.
Shahmuradov, I. A., R. K. Umarov and V. V. Solovyev (2017). "TSSPlant: a new tool for prediction of plant Pol II promoters." Nucleic Acids Res.
Shinya, E. and T. Shimada (1994). "Identification of two initiator elements in the bidirectional promoter of the human dihydrofolate reductase and mismatch repair protein 1 genes." Nucleic Acids Res **22**(11): 2143-2149.
Solovyev, V., P. Kosarev, I. Seledsov and D. Vorobyev (2006). "Automatic annotation of eukaryotic genes, pseudogenes and promoters." Genome Biol **7 Suppl 1**: S10 11-12.
Solovyev, V., Shahmuradov, I. (2003). "PromH: promoters identification using orthologous genomic sequences." Nucleic Acids Research **31**(13).
Solovyev, V. V., I. A. Shahmuradov and A. A. Salamov (2010). "Identification of promoter regions and regulatory sites." Methods Mol Biol **674**: 57-83.
Stepanova, M., T. Tiazhelova, M. Skoblov and A. Baranova (2005). "A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas." Bioinformatics **21**(9): 1789-1796.
Stracke, R., J. J. Favory, H. Gruber, L. Bartelniewoehner, S. Bartels, M. Binkert, M. Funk, B. Weisshaar and R. Ulm (2010). "The Arabidopsis bZIP transcription factor HY5 regulates expression of the PFG1/MYB12 gene in response to light and ultraviolet-B radiation." Plant Cell Environ **33**(1): 88-103.
Stracke, R., H. Ishihara, G. Huep, A. Barsch, F. Mehrtens, K. Niehaus and B. Weisshaar (2007). "Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling." Plant J **50**(4): 660-677.
Stracke, R., O. Jahns, M. Keck, T. Tohge, K. Niehaus, A. R. Fernie and B. Weisshaar (2010). "Analysis of PRODUCTION OF FLAVONOL GLYCOSIDES-dependent flavonol glycoside accumulation in Arabidopsis thaliana plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside accumulation." New Phytol **188**(4): 985-1000.
Tatarinova, T., Alexandrov, N., Bouck, J., Feldmann, K. (2010). "GC3 Biology in Corn, Rice, Sorghum and other grasses." BMC Genomics **11**(308).
Tatarinova, T., E. Elhaik and M. Pellegrini (2013). "Cross-species analysis of genic GC3 content and DNA methylation patterns." Genome Biol Evol **5**(8): 1443-1456.
Tatarinova, T., A. Kryshchenko, M. Triska, M. Hassan, D. Murphy, M. Neely and A. Schumitzky (2014). "NPEST: a nonparametric method and a database for transcription start site prediction." Quant Biol **1**(4):

261-271.
Tatarinova, T. V., E. Chekalin, Y. Nikolsky, S. Bruskin, D. Chebotarov, K. L. McNally and N. Alexandrov (2016). "Nucleotide diversity analysis highlights functionally important genomic regions." Sci Rep **6**: 35730.
Triska, M., D. Grocutt, J. Southern, D. J. Murphy and T. Tatarinova (2013). "cisExpress: motif detection in DNA sequences." Bioinformatics **29**(17): 2203-2205.
Troukhan, M., T. Tatarinova, J. Bouck, R. B. Flavell and N. N. Alexandrov (2009). "Genome-wide discovery of cis-elements in promoter sequences using gene expression." OMICS **13**(2): 139-151.
Troukhan, M., Tatarinova, T. Bouck, J., Flawell, R., Alexandrov, N. (2009). "Genome-wide discovery of cis-elements in promoter sequences using gene expression data." Omics **13**(1).
Umarov, R. K. and V. V. Solovyev (2017). "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks." PLoS One **12**(2): e0171410.
van Heeringen, S. J., W. Akhtar, U. G. Jacobi, R. C. Akkers, Y. Suzuki and G. J. Veenstra (2011). "Nucleotide composition-linked divergence of vertebrate core promoter architecture." Genome Res **21**(3): 410-421.
Viola, I. L., N. G. Uberti Manassero, R. Ripoll and D. H. Gonzalez (2011). "The Arabidopsis class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain." Biochem J **435**(1): 143-155.
Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J. C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. Walhout, F. Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker and T. R. Hughes (2014). "Determination and inference of eukaryotic transcription factor sequence specificity." Cell **158**(6): 1431-1443.
Wongsurawat, T., P. Jenjaroenpun, C. K. Kwoh and V. Kuznetsov (2012). "Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity." Nucleic Acids Res **40**(2): e16.
Yao, H., L. Guo, Y. Fu, L. A. Borsuk, T. J. Wen, D. S. Skibbe, X. Cui, B. E. Scheffler, J. Cao, S. J. Emrich, D. A. Ashlock and P. S. Schnable (2005). "Evaluation of five ab initio gene prediction programs for the discovery of maize genes." Plant Mol Biol **57**(3): 445-460.

Figure 1

Click here to download Figure Figure 1.docx

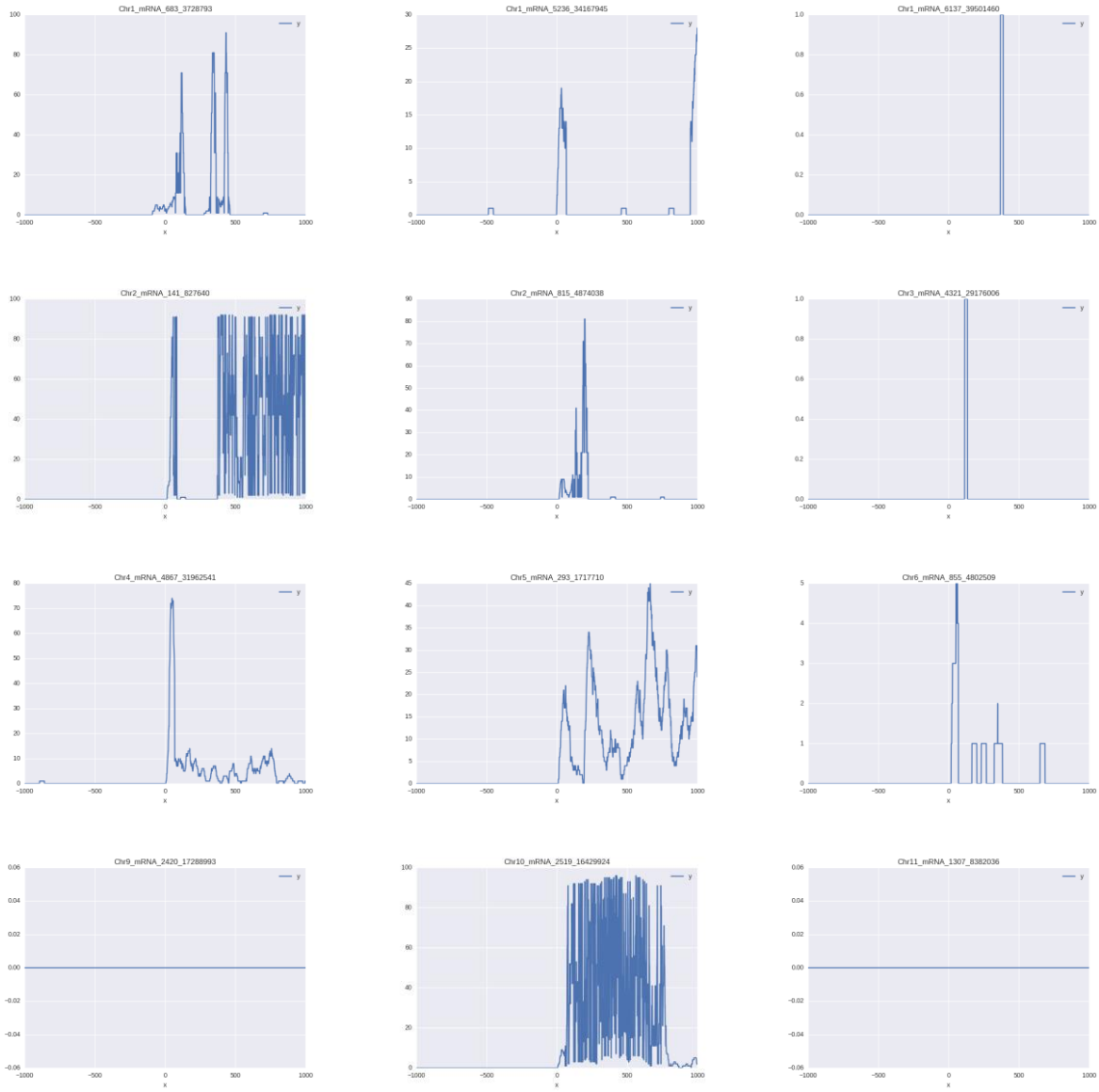


Figure 2

Click here to download Figure Figure 2.docx

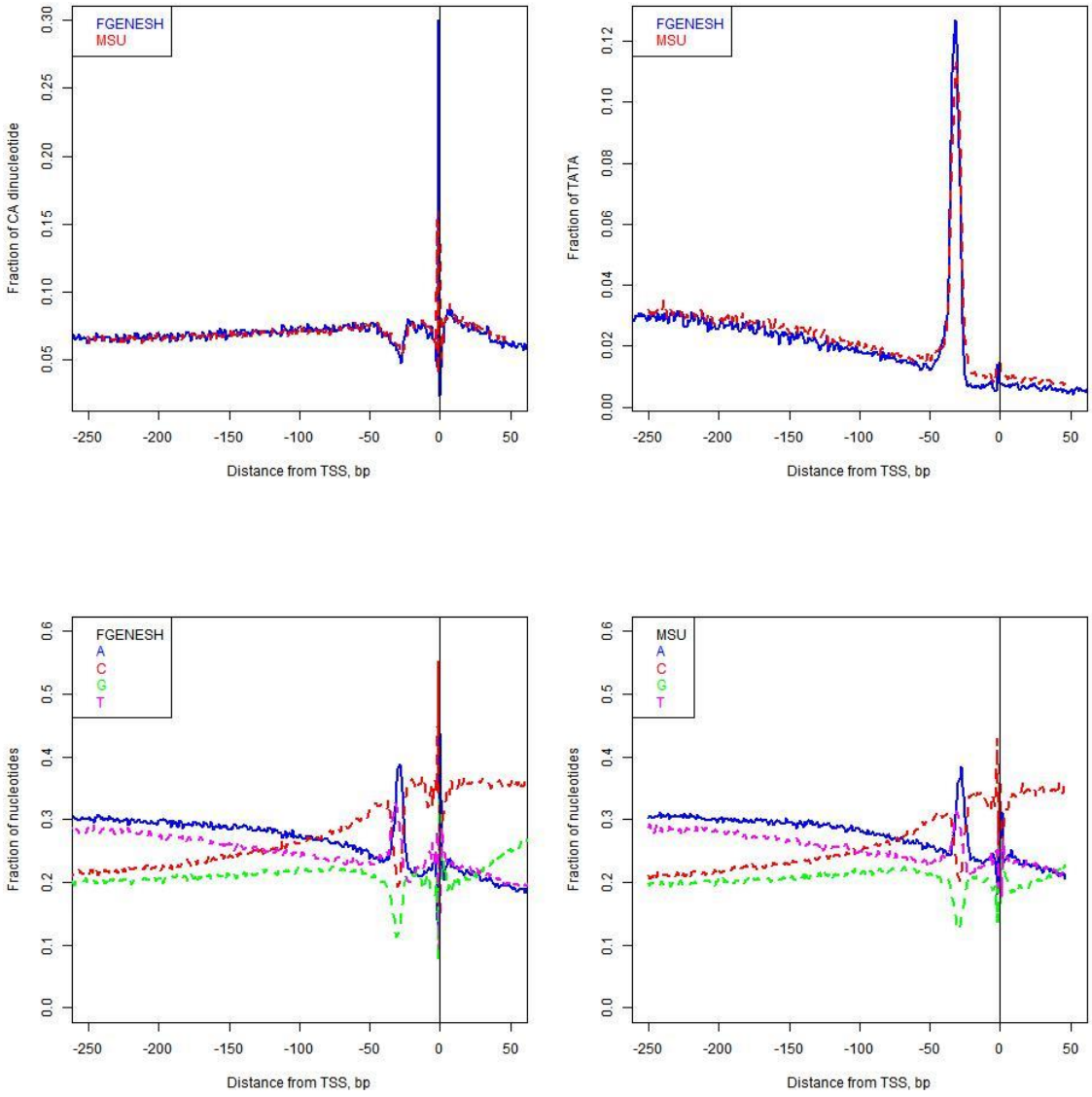
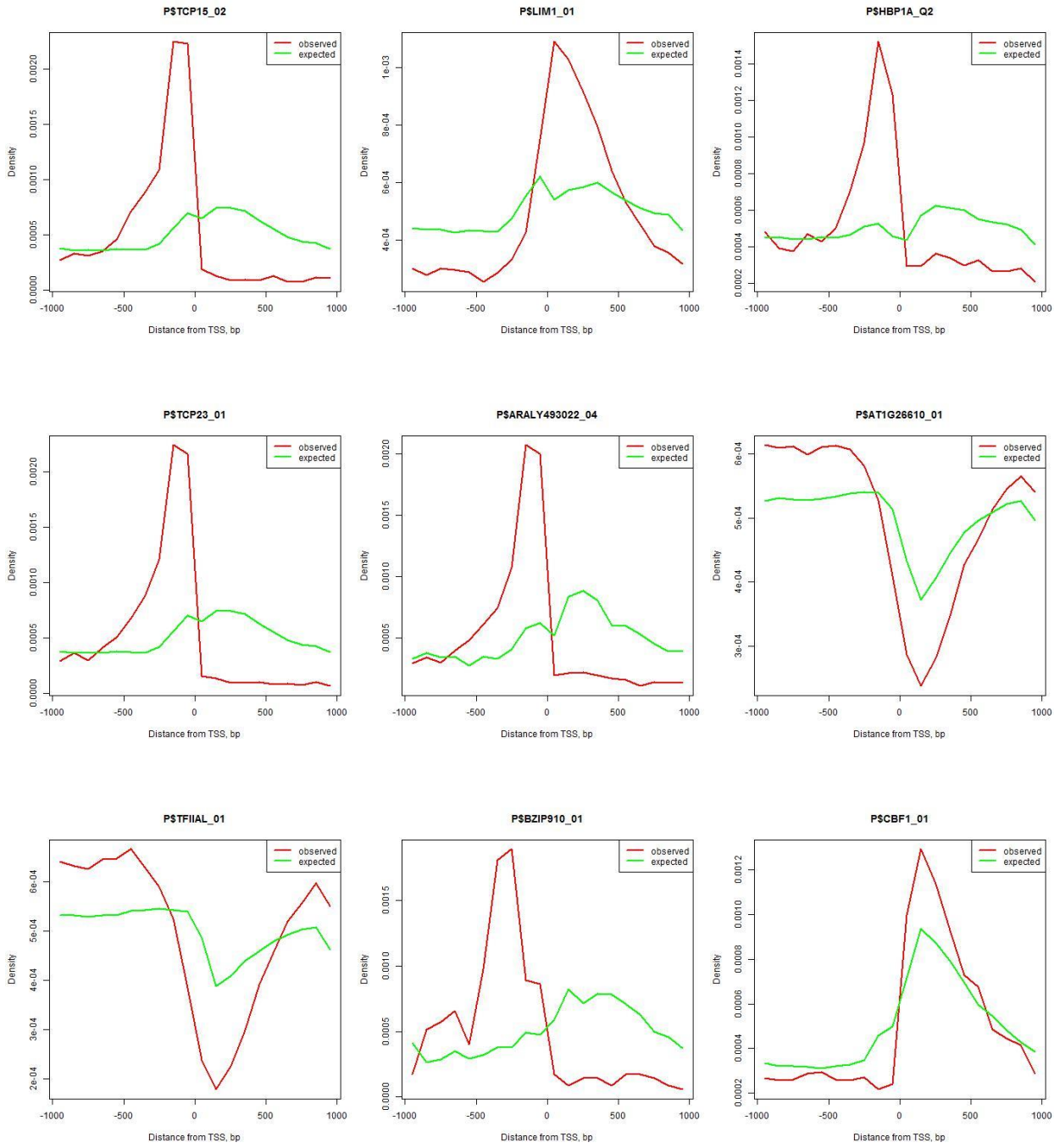
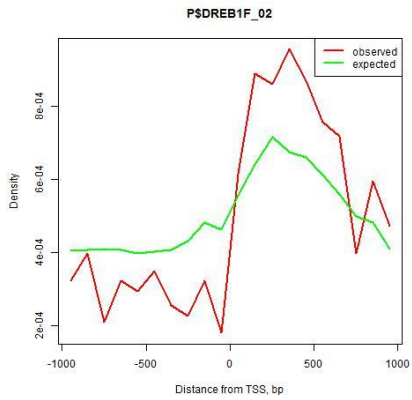
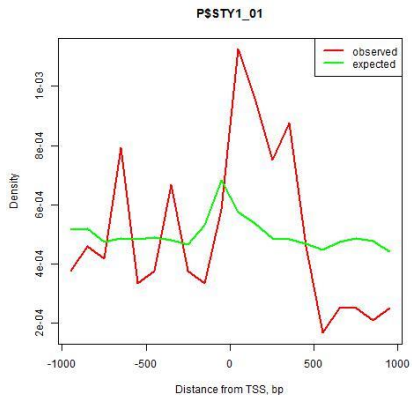


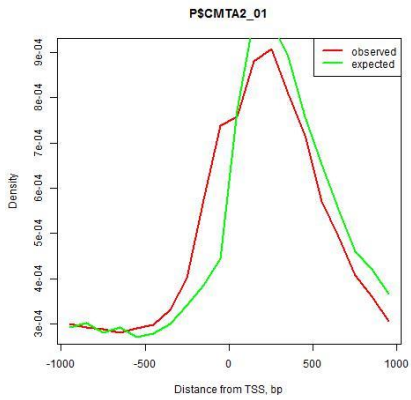
Figure 3

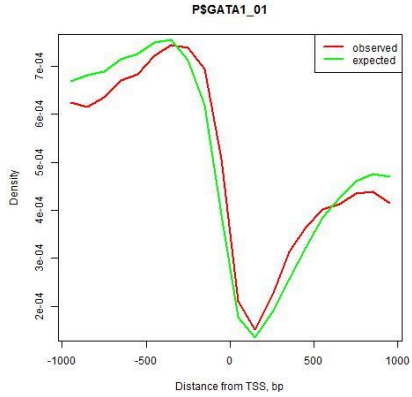
Click here to download Figure Figure 3.docx

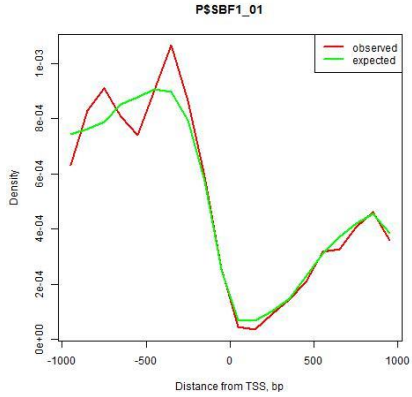



P$DREB1F_02
observed
expected
Density
2e-04
4e-04
6e-04
8e-04
-1000
-500
0
500
1000
Distance from TSS, bp


P$STY1_01
observed
expected
Density
2e-04
4e-04
6e-04
8e-04
1e-03
-1000
-500
0
500
1000
Distance from TSS, bp


P$CMTA2_01
observed
expected
Density
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
9e-04
-1000
-500
0
500
1000
Distance from TSS, bp


P$GATA1_01
observed
expected
Density
2e-04
3e-04
4e-04
5e-04
6e-04
7e-04
-1000
-500
0
500
1000
Distance from TSS, bp


P$SBF1_01
observed
expected
Density
0e+00
2e-04
4e-04
6e-04
8e-04
1e-03
-1000
-500
0
500
1000
Distance from TSS, bp

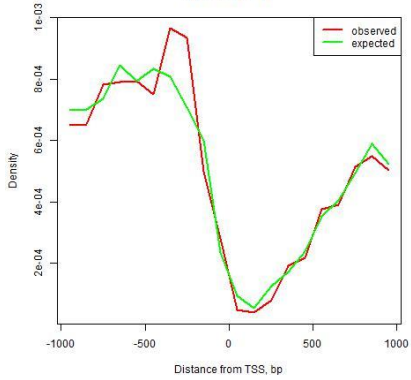

P$WRKY48_01
observed
expected
Density
2e-04
4e-04
6e-04
8e-04
1e-03
-1000
-500
0
500
1000
Distance from TSS, bp

Figure 4

Click here to download Figure Figure 4.docx

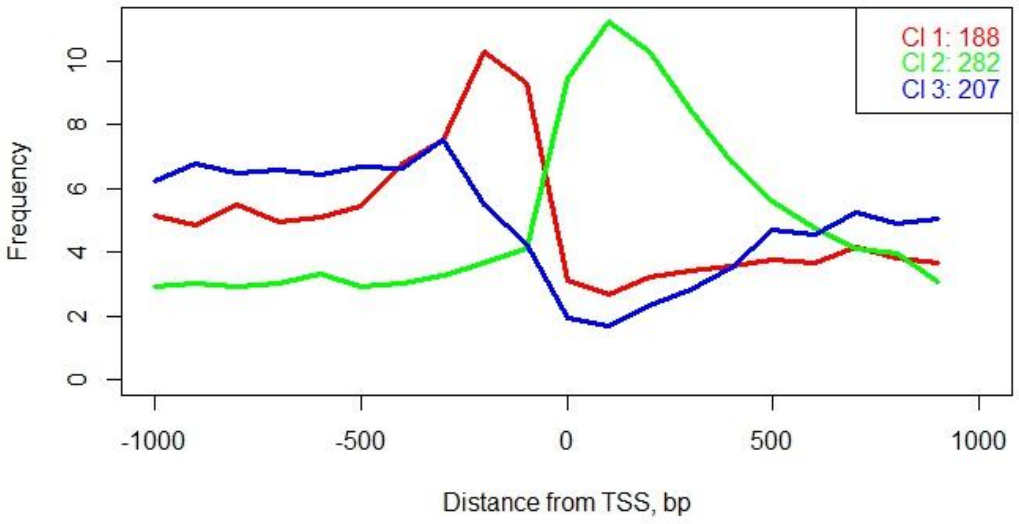


Figure 5

Click here to download Figure Figure 5.docx

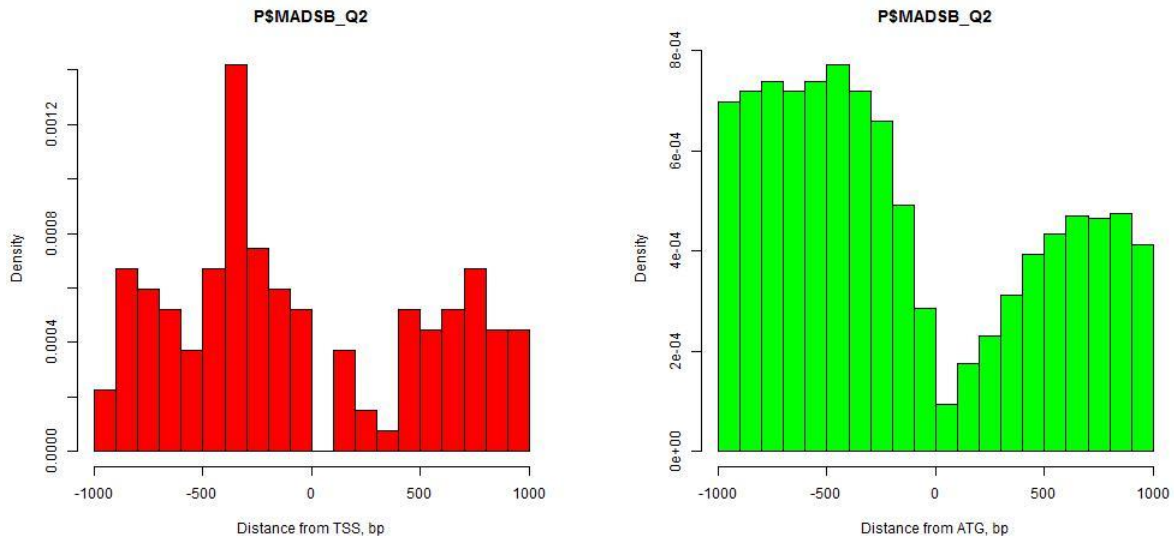


Figure 6

Click here to download Figure Figure 6.docx

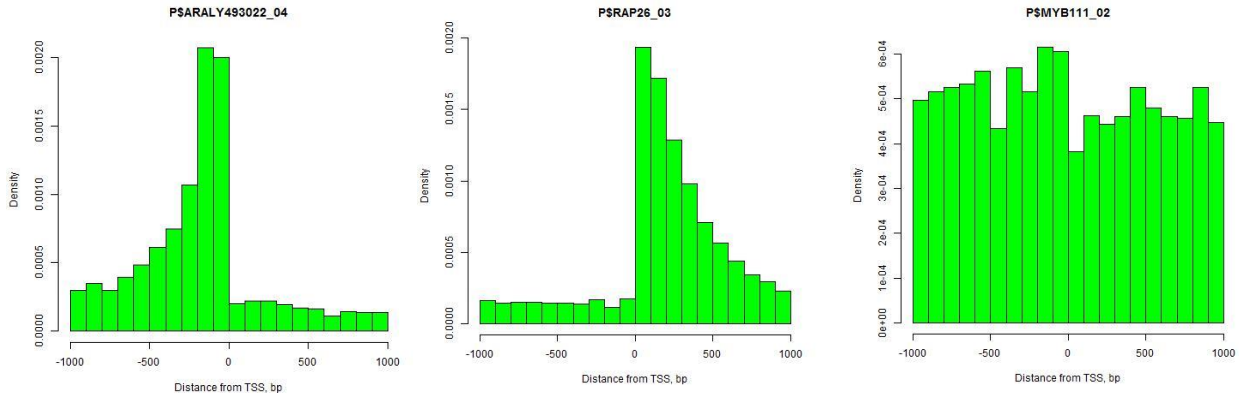


Figure 7

Click here to download Figure Figure 7.docx

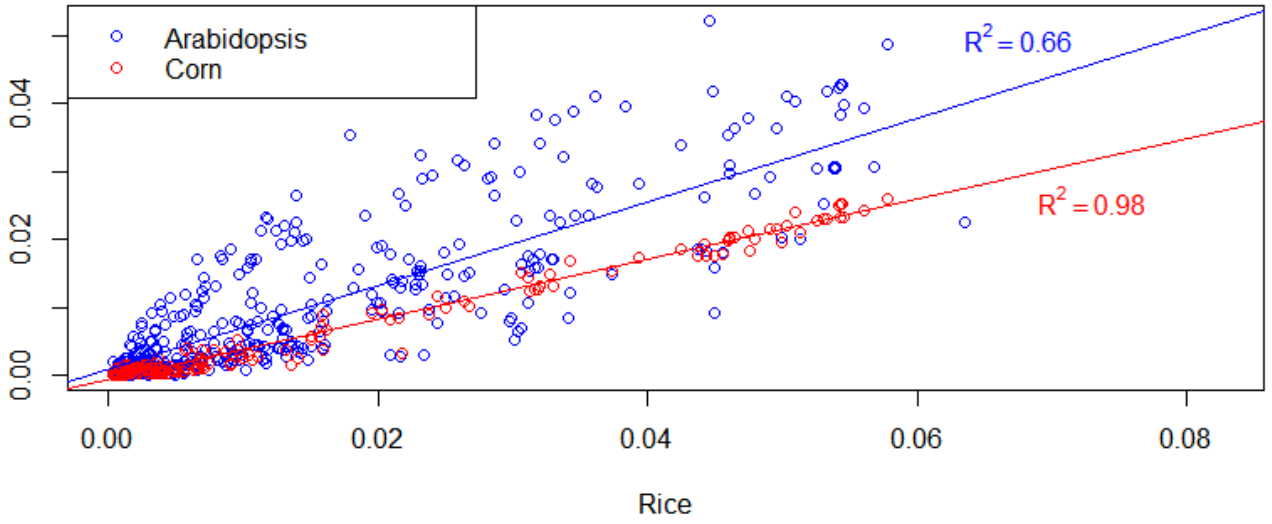


Figure 8

Click here to download Figure Figure 8.docx

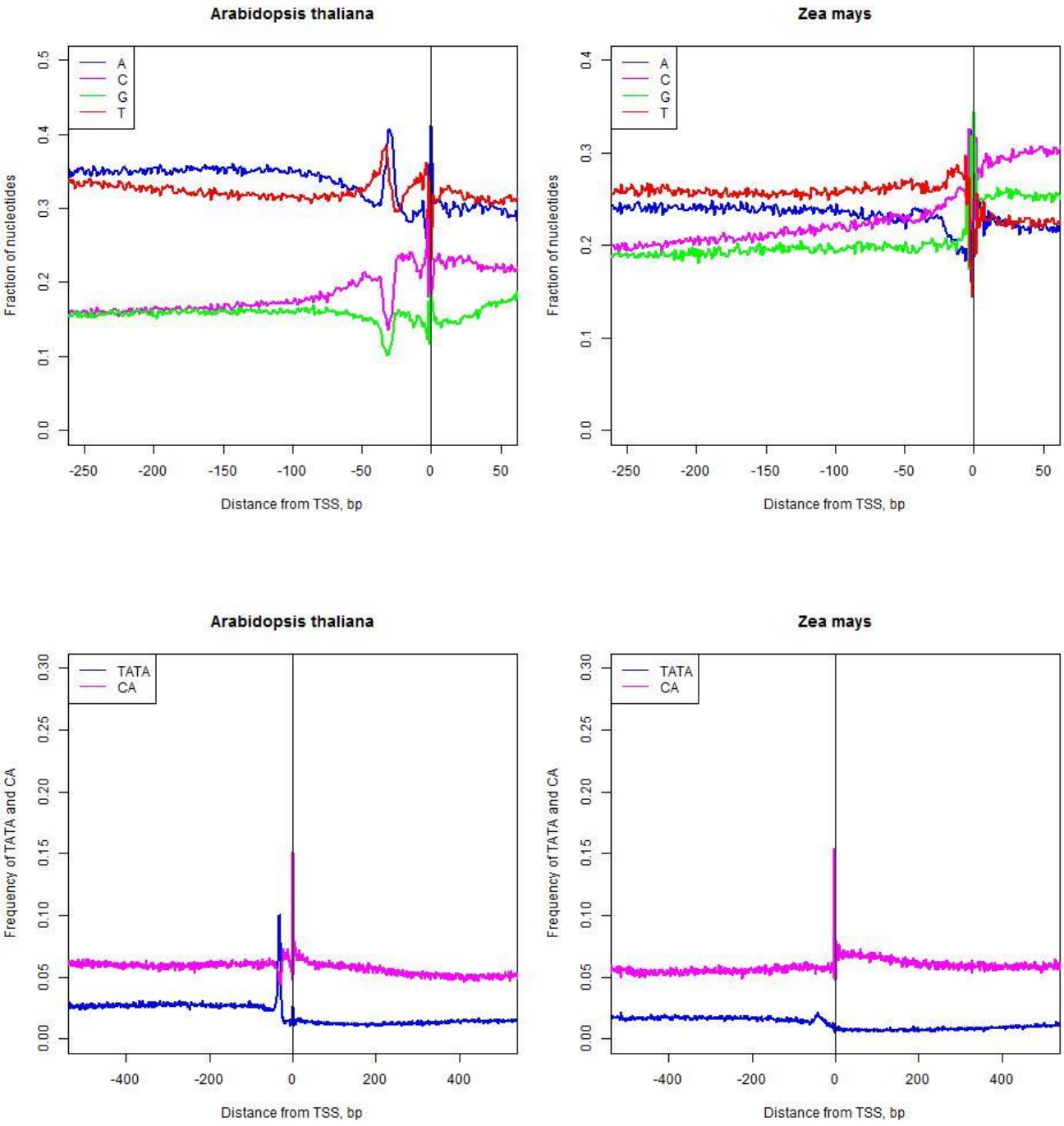


Figure 9

Click here to download Figure Figure 9.docx

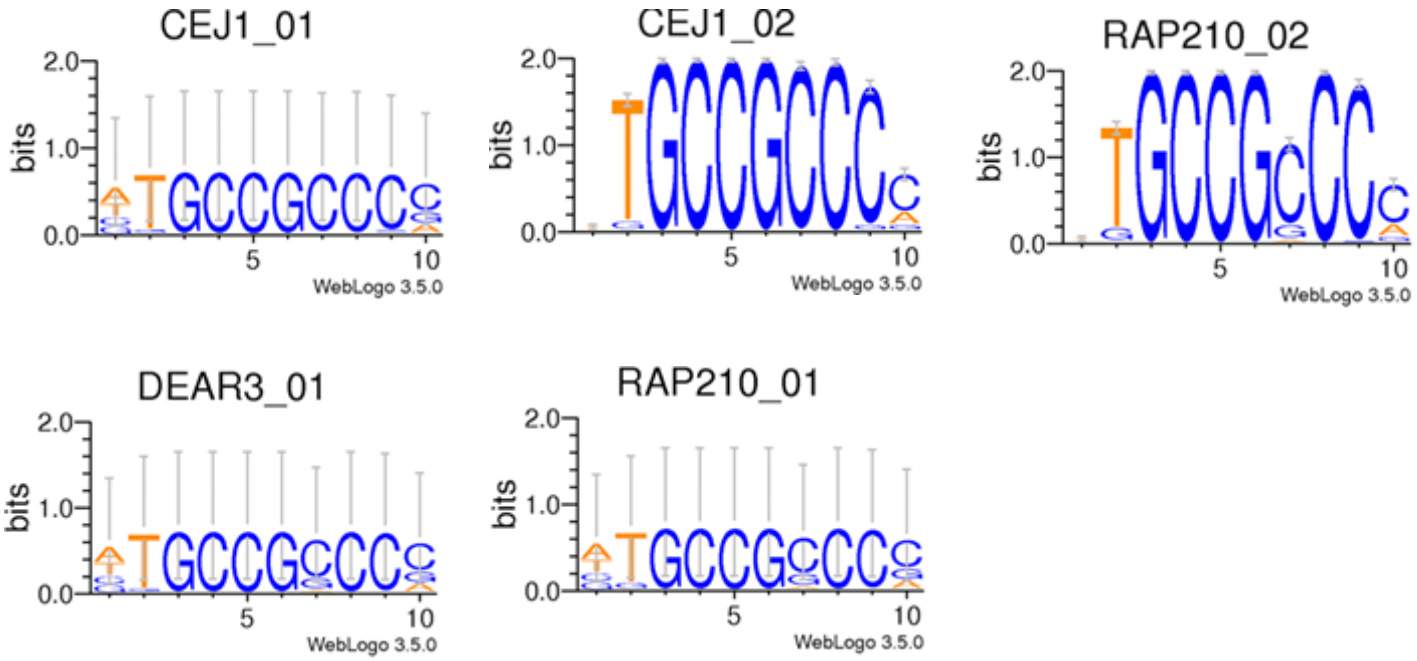


Figure 10

Click here to download Figure Figure 10.docx

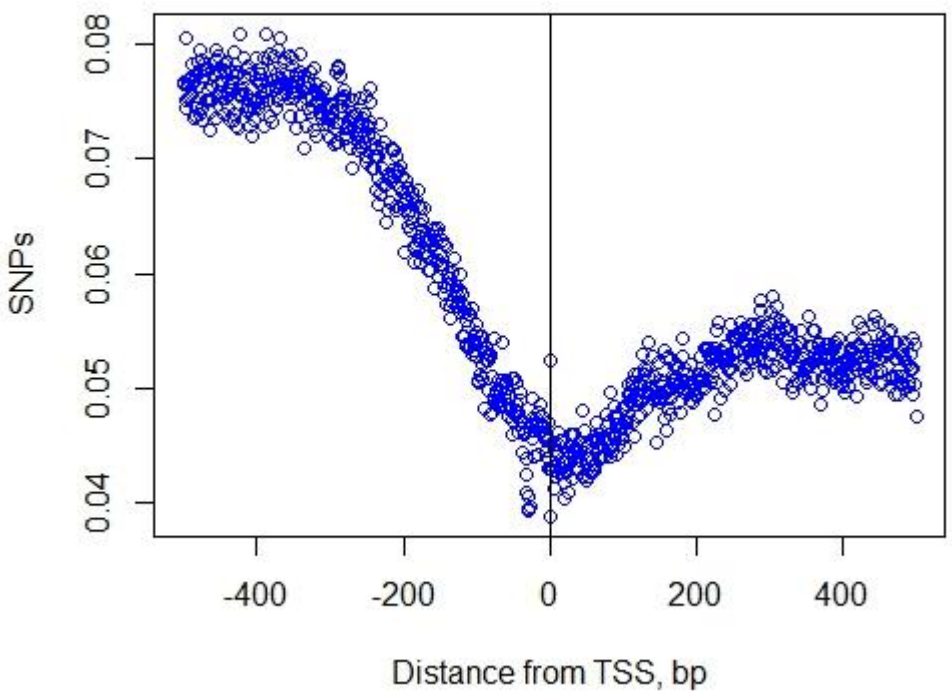


Figure 11

Click here to download Figure Figure 11.docx

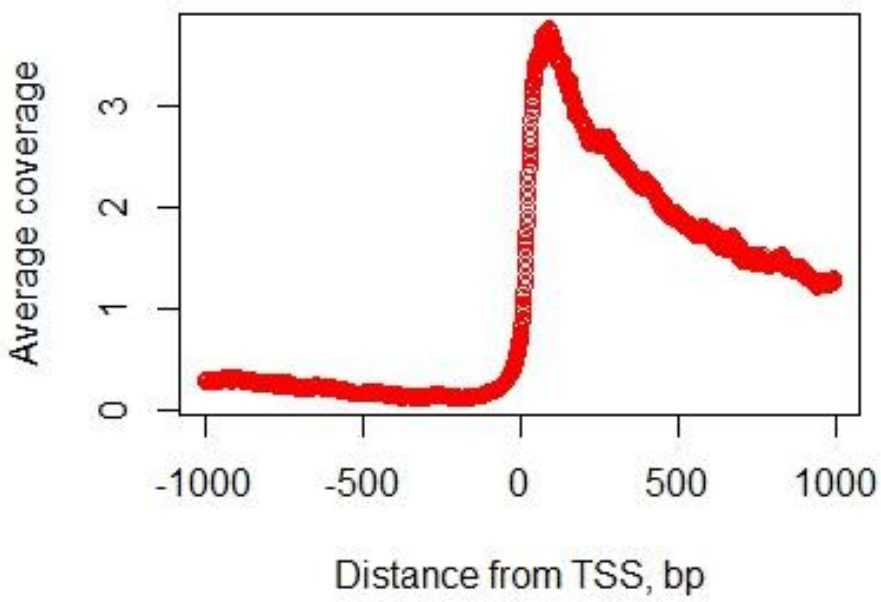


Figure 12

Click here to download Figure Figure 12.docx

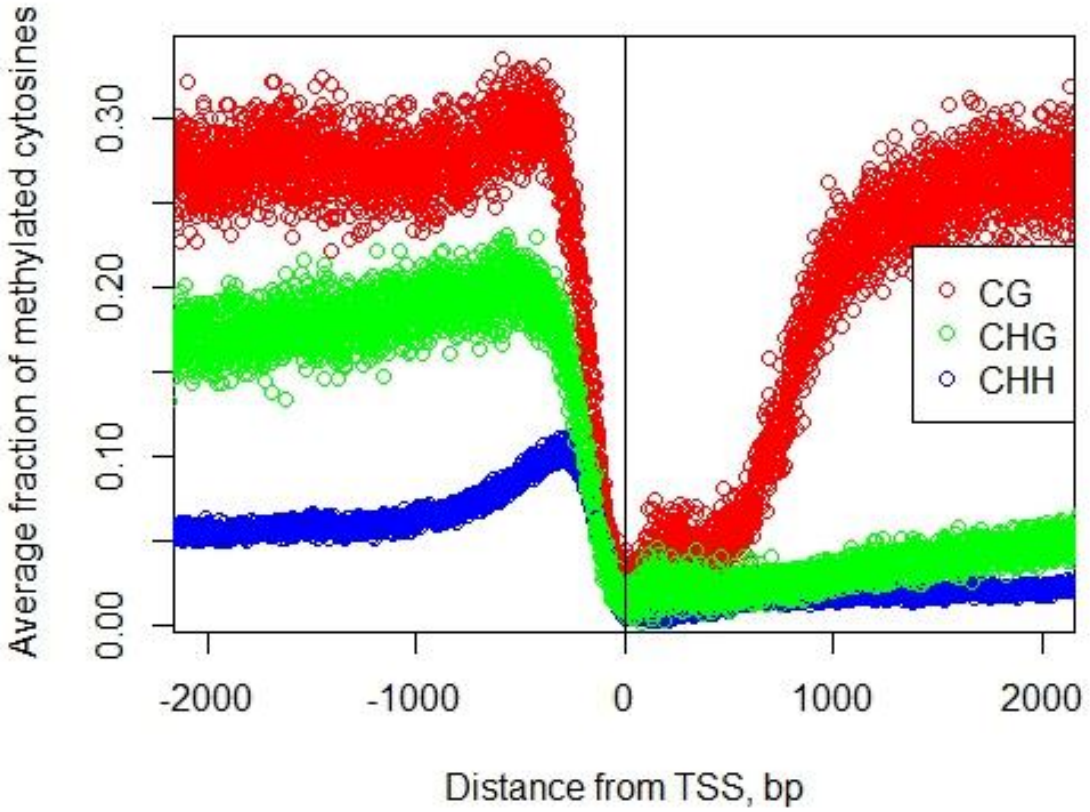
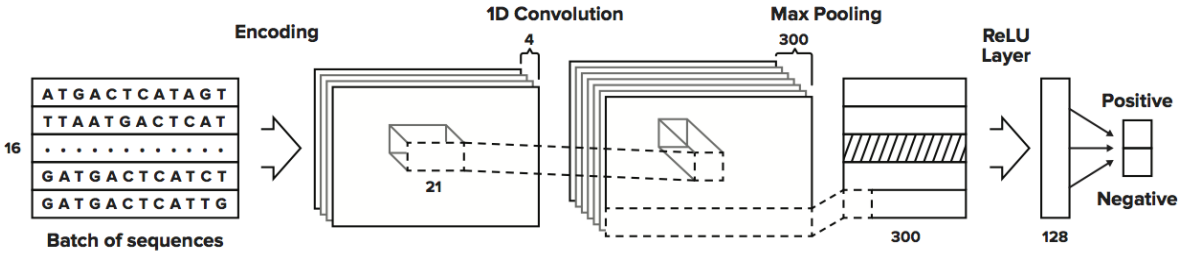


Figure 13

Click here to download Figure Figure 13.docx



Supp data 1

Click here to access/download

Supporting Information

FGEHESH_RICE.gff3

Supp data 2

Click here to access/download

Supporting Information

Supplemental data.xlsx