TRANSFAC-ing the rice genome

Alexander Kel

Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia,
alexander.kel@genexplain.com

Tatiana V. Tatarinova

University of Southern California, Los Angeles, CA, USA and

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

tatarino@usc.edu

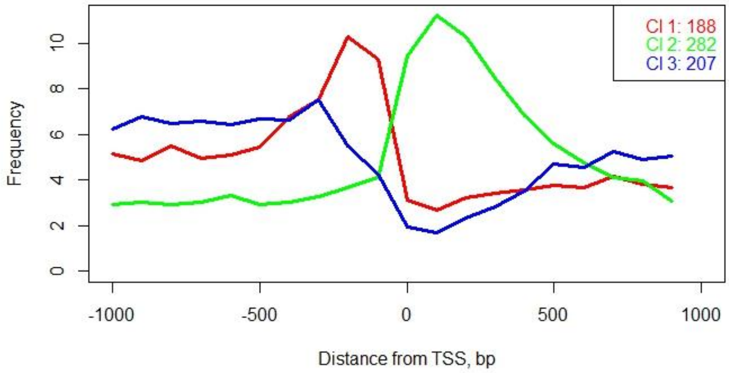
Recent publication of 40 million single nucleotide polymorphisms (SNPs) dataset from the 3,000 Rice Genomes Project (http://snp-seek.irri.org), the largest and highest density SNP collection for any higher plant, facilitated analysis of functionality and distribution of genetic variants across the complete *Oryza sativa* genome [1, 2]. The observed profound patterns of nucleotide variability reveal functionally important genomic regions. As expected, nucleotide diversity is much higher in intergenic regions than within gene bodies (regions spanning gene models), and protein-coding sequences are more conserved than untranslated gene regions. We have observed a sharp decline in nucleotide diversity that begins at about 250 nucleotides upstream of the transcription start and reaches minimal diversity exactly at the transcription start. We found the transcription termination sites to have remarkably symmetrical patterns of SNP density, implying presence of functional sites near transcription termination. Also, nucleotide diversity was significantly lower near 3′ UTRs, the area rich with regulatory regions.

Functional classes of genes differ in conservation. We show that the DNA-binding transcription factors (TFs) are the most conserved group of genes, whereas kinases and membrane-localized transporters are the most variable ones. TFs may be conserved because they belong to some of the most connected regulatory hubs that modulate transcription of vast downstream gene networks, whereas signaling kinases and transporters need to adapt rapidly to changing environmental conditions. To investigate the distributions of SNPs at a finer scale, we have a subset of 20,367 “high-confidence” protein coding rice genes that have sufficient experimental support, and have a reliably predicted transcription start site.

At the first stage of our project we have investigated distribution of TFBS around TSS. We searched for the potential TF-binding sites in promoters and UTRs of high-confidence rice genes in the regions of -1000 +1000 around TSS. For the site search we used the MATCH program [3] incorporated in geneXplain software platform (www.genexplain.com). It applies TRANSFAC database [4] with all 764 plant position weight matrices (PWM) with a strict score threshold 0.95. MATCH program scans the targets promoter sequences with a sliding window of the length of the PWM and calculates a score for each such window. The maximum value of the score of 1.0 corresponds to the sequence that fully fits to the consensus of the PWM. Score threshold of 0.95 allows few mismatches to the consensus in the positions that are not well conserved between known sites for this TF. The score in the MATCH program is calculated using a specific algorithm that considers the entropy measure for each position of PWM. In a recent study, we demonstrate the higher accuracy of this algorithm in comparison to the other algorithms [5].

MATCH search revealed 3.2 million potential TF binding sites corresponding to 667 PWMs. No sites were found for 97 PWMs. We think, the corresponding transcription factors have got their sites outside of the proximal promoter arias that we consider in this study. The most frequent sites were for the transcription factors ASR1, DOF56 and PBF. We compared the frequencies of found TF sites in the proximal promoters with the frequencies of found hits for the same PWMs in randomly shuffled sequences. We found the most significant enrichment (more than 2 times) of TF binding sites in promoters in comparison to the randomly shuffled sequences for the following transcription factors: SPL12, SPL5, GBF1, ABI5, BZIP68_01, LEC2_01 and GT1.

Clustering of TBFS shows that there are three classes of transcription factors: those that bind preferentially to the [-500,0] (promoter) region, those that bind preferentially to the [0,500] (UTR) region, and the agnostic transcription factors, that have weak location



preference.

Comparative gene ontology analysis of Class 1 and 2 transcription factors showed that Class 1 is enriched in the following GO terms: “gibberellic acid mediated signaling pathway,” “regulation of hydrogen peroxide metabolic process,” “systemic acquired resistance,” “salicylic acid mediated signaling pathway,” “plant ovule development,” “cellular response to phosphate starvation,” “endoplasmic reticulum unfolded protein response,” and “response to xenobiotic stimulus.” Class 2 is enriched in “response to ethylene,” “cellular response to cold,” “response to auxin,” and “response to water deprivation.” Also, the two classes of TF have different expression patterns: Class 1 genes are expressed more in petals, sepals and plant embryo, while Class 2 genes are more expressed in roots. This observation may explain why TATA box was the only cis-element in promoter that was statistically significantly associated with expression in plant roots [6, 7]: possibly root-specific transcription factors bind to the UTR region.

Next, we have analyzed distribution of SNPs and their effect on TF binding site loss and gain. For each predicted TBFS we have calculated $\Delta = |q - q^*|$, and compared its value to empirically determined threshold. $q = \dfrac{\sum_{i=1}^{l} I(i) f(b_i, i) - \sum_{i=1}^{l} I(i) f^{\min}(i)}{\sum_{i=1}^{l} I(i) f^{\max}(i)}$, where $I(i) = \sum_{b \in \{A,T,G,C\}} f(b,i) \ln(4 f(b,i))$. If $\Delta \geq \Delta_0$, the site was “lost” or “gained”. We compared frequency of site loss and gain in promoters and in random subset of 20,367 intergenic sequences, each 2000 nt long. Our hypothesis was that that functionally important motifs in promoters will have less SNPs that cause site loss. For each TF, we calculated ratio of the number of site losses in promoters to the number of site losses in intergenic and ranked the list.

Most “protected” from the site loss are binding sites for ABF transcription factors (CACGTGGC) that function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signaling in response to osmotic stress. They are followed by binding sites for CBF4 (regulator of drought adaptation). We also observed that SNPs are avoided in positions where nucleotide change can lead to the site gain of several important transcription factors, such as MADS8 (involved in the control of flowering time), GT-1 and GATA-1 (response to light). We propose that sporadic creation of new sites for such transcription factors can significantly alter cellular timing, therefore such mutations are avoided. 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.

Acknowledgments

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”, grant number: 14.604.21.0101 to the Institute of Chemical Biology and Fundamental Medicine, SBRAS. This work was also supported by the following grants of the EU FP7 program: “SYSCOL”, “SysMedIBD”, “RESOLVE” and “MIMOMICS”. TT was supported by the NSF Division of Environmental Biology (1456634) and NSF STTR award 1622840.

References

1. Tatarinova, T.V., et al., *Nucleotide diversity analysis highlights functionally important genomic regions.* Sci Rep, 2016. **6**: p. 35730.
2. Alexandrov, N., et al., *SNP-Seek database of SNPs derived from 3000 rice genomes.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1023-7.
3. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences.* Nucleic Acids Res, 2003. **31**(13): p. 3576-9.
4. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
5. Kondrakhina, Y., et al., *Prediction of protein-DNA interactions of transcription factors linking proteomics and transcriptomics data.* EuPA Open Proteomics, 2016. **13**: p. 14–23.
6. Troukhan, M., et al., *Genome-wide discovery of cis-elements in promoter sequences using gene expression.* OMICS, 2009. **13**(2): p. 139-51.
7. Triska, M., et al., *cisExpress: motif detection in DNA sequences.* Bioinformatics, 2013. **29**(17): p. 2203-5.