Decoding functional regulatory maps via genomic evolutionary footprints in 63 green plants

Feng Tian1,2,3, De-Chang Yang1,2, Yu-Qi Meng1,2, Jinpu Jin1,2* and Ge Gao1,2*

1 State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Center for Bioinformatics, Peking University, Beijing, 100871, China; 2 Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, 100871, China; 3 Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China.

Email:

FT: tianf@mail.cbi.pku.edu.cn
DCY: yangdc@mail.cbi.pku.edu.cn
YQM: mengyq@mail.cbi.pku.edu.cn
JJ: jinjp@mail.cbi.pku.edu.cn
GG: gaog@mail.cbi.pku.edu.cn

*Correspondence: gaog@mail.cbi.pku.edu.cn; jinjp@mail.cbi.pku.edu.cn.
Abstract

Systematic identification of functional transcriptional regulatory interactions is essential for understanding regulatory systems. Here, we firstly established genome-wide conservation landscapes for 63 green plants of seven lineages and then developed an algorithm FunTFBS to screen for functional regulatory elements and interactions by coupling base-varied binding affinities of transcription factors with the evolutionary footprints on their binding sites. Using the FunTFBS and the conservation landscapes, we further identified over two million functional interactions for 21,346 TFs, charting functional regulatory maps of these 63 plants. Our work provides plant community with valuable resources to decode plant transcriptional regulatory system and genome sequences.

Keywords

Conservation landscape, Evolutionary footprint, Functional transcription factor binding site, Transcriptional regulatory network, Green plants
Background

Transcription factors (TFs) control gene expression by binding to specific \textit{cis}-elements, which play essential roles in plant development and stress responses. A systematic identification of the functional transcriptional regulatory interactions would greatly facilitate further mechanism investigation [1, 2]. With the binding motifs of genome-wide TFs determined by experiments in plants [3, 4] and \textit{in silico} mapped into 156 plants [5], directly scanning the TF binding motifs in the promoters of putative target genes becomes a promising option to chart plant transcriptional regulatory maps. As prediction from directly scanning yields a rather high false positive rate, additional data like DNase-seq footprints [6, 7] and conserved elements [8-10] have been incorporated to screen for functional TF binding sites (TFBSs). However, these data are only available in a few model plants [7, 11] and conserved element-based methods are still confounded by evolutionary constraints from other functional elements besides of TF binding [12], hindering this field from systematically charting the transcriptional regulatory maps across the plant kingdom.

A comparison of multiple related genomes with substantial divergence is widely used to detect evolutionary constraints and further identify the functional elements [11, 13, 14]. The availability of over one hundred of plant genomes provides a unique opportunity to calculate genome-wide evolutionary footprints and further to chart functional plant regulatory maps. In this study, we established the first genome-wide conservation landscapes for 63 green plants and developed an algorithm by coupling
base-varied binding affinities of TFs with the evolutionary footprints on their binding
sites to screen functional transcription regulatory elements and to infer interactions,
charting the regulatory maps of these 63 plants.

Results and Discussion

To detect evolutionary constraints, we first chose 63 representative species of seven
lineages (Fig. 1a), with the divergence time from 37 million years ago (MYA) to 106
MYA (Fig. 1b), covering main lineages of green plants. Then, we performed all 322
intragroup pairwise genome alignments and then assembled them into 63 multiple
genome alignments by using each species as a reference genome (see Additional file 1:
Supplementary Methods for more details). Finally, we identified over 67 million
conserved elements and calculated the basewise conservation scores for over 22
billion base pairs, covering approximately 66% of genome sequences and establishing
the first conservation landscapes covering the main lineages of green plants (Fig. 1b;
Additional file 1: Tables S1 and S2).

For these seven groups, approximately 54-87% of genomes are aligned together, and
at least 10-17% of the genomes are under evolutionary constraint (Fig. 1b; Additional
file 1: Tables S1 and S2). Compared with a previous study in A. thaliana [11] which
used nine species in Brassicales to calculate conservation scores, our work, which
embraced more representative species (18 vs 9), aligned 20.49% (18.13 Mb) more
genome sequences and detected 100.96% (4.84 Mb) more conserved non-coding
regions with a higher accuracy (Additional file 1: Figure S1). Compared with other kingdoms, plants present a higher conserved ratio in the noncoding regions than these of humans but less than those organisms with lower proportion of non-coding sequences in genomes, such as fruit flies, worms, and yeast (Fig. 1c). Lineages suffering more rounds of whole genome duplication (e.g., Fabales) or sudden genome expansion (e.g., PACMAD clade) show lower conservation ratios in their genomes, likely due to genomic degeneration after whole genome duplication or a lack of homologs in close species after sudden expansion (e.g., Zea mays [15]) (Fig. 1b, c; Additional file 1: Table S1). Of note, compared to the unannotated nonconserved regions, the unannotated conserved regions in A. thaliana genome present a larger proportion covered by transcriptional signals (27% vs 10%) and a higher expression level (Additional file 1: Figure S2), suggesting that many genes or functional elements still remain to be illustrated even in this most well-annotated plant genome.

Conservation landscapes in humans and fruit flies have been widely used to illustrate their genome sequences [13, 14], thus our conservation landscapes in 63 plants provide the community a unique chance to illustrate plant genomes. For users to conveniently access the conservation data, we have set up a genome browser to visualize each species (Fig. 1d) [16].

The establishment of conservation landscapes in the main lineage of green plants paves the way for a systematic identification of functional TFBSs. However, other functional elements besides of TFBSs (such as noncoding RNAs and stem regions in
RNA structures) may also result in a conservation at the gene promoters [12] (Additional file 1: Figure S3), confounding algorithms that depend on conserved elements to screen for functional interactions. We speculate that the base-varied binding affinity (base frequencies in the binding motifs) of the TF binding motifs could yield a consistent base-varied evolutionary constraint on the functional TFBSs (Fig. 2a). To determine whether this feature could distinguish functional TFBSs from nonfunctional ones, we firstly featured out an evaluation dataset by classifying the TFBSs that were identified from 124 ChIP-seq experiments for 21 TFs [17] into three classes: “Less reliable”, “Highly reliable” and “Functional”. The “Less reliable” and the “Highly reliable” TFBSs represent the TFBSs with low and high consistency among replicates, respectively, and the “Functional” ones are the “Highly reliable” TFBSs that are further supported by expression data (Additional file 1: Supplementary Methods). A method with a higher screening efficiency would result in a lower percentage of TFBSs supported by “Less reliable” TFBSs but a higher percentage of TFBSs supported by “Highly reliable” and “Functional” TFBSs. Compared with the TFBSs whose conservation scores are inconsistent with their binding affinity, the consistent TFBSs are depleted in the “Less reliable” TFBSs and enriched in the “Highly reliable” TFBSs, particularly the “Functional” ones (Fig. 2b), suggesting that this feature does work to screen for functional TFBSs.

Employing this feature, we developed an algorithm called FunTFBS to screen for functional TFBSs and to infer the functional regulatory interactions (Additional file 1:}
Figure S4 and Supplementary Methods). To determine whether our algorithm showed a higher accuracy for functional TFBSs than the other motif-based methods, we firstly compared FunTFBS with the existing DNase-seq footprint-based and conserved element-based methods using the evaluation dataset mentioned above. The result showed that our algorithm presented a 42% and 33% decrease in the percentage of screened TFBSs that were supported by “Less reliable” TFBSs, but it presented a 68% and 67% increase in the percentage of screened TFBSs that were supported by “Functional” TFBSs compared to the DNase-seq footprint-based and conserved element-based methods, respectively, suggesting that our algorithm could more efficiently screen for functional TFBSs (Fig. 2c).

We then assessed the accuracy of our algorithm in inferring transcriptional regulatory interactions based on experimentally validated interactions from *Arabidopsis* transcriptional regulatory map (ATRM) [2]. Our algorithm showed a 95-146% increase in the percentage of edges that were supported by the functional regulatory interactions in ATRM than the DNase-seq footprint-based and conserved element-based methods (Additional file 1: Figure S5), indicating the superiority of FunTFBS in inferring functional regulatory interactions. We further assessed the performance of our algorithm based on another two indexes: the percentage of regulatory pairs that coexist in the same biological process and the percentage of regulatory pairs that are highly correlated in expression [18], where higher numbers in the two indexes represent higher-quality interactions. Our algorithm showed the
highest percentage of TFs and their targets coexisting in the two indexes (20-22% and 20-39% increases than the other two methods in the two indexes, respectively) (Additional file 1: Figures S6 and S7), further confirming the superiority of FunTFBS in screening for functional regulatory interactions.

After confirming the superiority of FunTFBS in screening for functional regulatory interactions, we employed this method with integrated genomic TF binding motifs in these 63 plants [5]. Finally, we identified 21,997,501 functional TFBSs and further inferred 2,196,397 regulatory interactions for 21,346 TFs (Additional file 1: Table S3), charting the functional regulatory maps for the main lineages of green plants.

Our identified functional TFBSs are significantly enriched in expression quantitative trait loci (eQTLs) (Additional file 1: Figure S8), offering a unique chance to unveil the molecular mechanisms that underlie genetic variation and gene expression alternation. For example, according to the 1,203 transcriptomes from the 1001 Arabidopsis genome project [19], one substitution (Chr4:268990 A>T) is associated with lower expression of AT4G00650, a major source of variation in flowering time.

Through browsing our functional TFBSs, we found a TF (AT5G67580) that could bind to that position, and an A to T substitution would definitely weaken its binding (Fig. 2d, e), shedding lights on the putative molecular mechanism. Moreover, the functional regulations also provide informative clues for the function of the TFs. For example, the target genes of the transcription factor AT3G22830 were found to be
enriched in “response to heat” (Fig. 2f), a biological process which corresponds well to its reported heat stress response function [20].

Finally, the conservation landscapes, the FunTFBS tool and the screened functional TFBSs and regulatory interactions were integrated into the PlantRegMap [5, 21] (More details see Additional file 1: Table S4), a popular plant transcriptional regulatory data and analysis platform, to better serve the community.

Conclusions

In this study, we established the first genome-wide, basewise conservation landscapes of 63 green plants in seven lineages. We further developed a novel algorithm by coupling the base-varied binding affinities of the TFs and the evolutionary footprints on their binding sites to screen for functional TFBSs. Using this method, we systematically screened for the functional TFBSs and regulatory interactions in 63 plants, charting functional regulatory maps for the main branches of higher plants. We believe these resources will advance the understanding of plant transcriptional regulatory systems and the further illustration of plant genome sequences.

List of abbreviations

ATRM: Arabidopsis transcriptional regulatory map
eQTLs: Expression quantitative trait loci

MYA: Million year ago
Additional files

Additional file 1: Supplementary Methods, Figures S1-S8, and Tables S1-S4.

Acknowledgments

Part of the analyses were performed on the Computing Platform of the Center for Life Sciences of Peking University, and we thank Dr. Fangjin Chen and Ting Fang for their helps.

Funding

This work was supported by National Natural Science Foundation of China (1470330), International Collaboration Program for Proteome Biological Big Data and Standard System (2014DFB30030), State Key Laboratory of Protein and Plant Gene Research, China Postdoctoral Science Foundation Grant (2014M560017 and 2015T80015 to JJ), Postdoctoral Fellowship at Peking University and Peking-Tsinghua Center for Life Sciences (to JJ), and the National Program for Support of Top-notch Young Professionals (to GG).

Availability of data and materials
The conservation landscapes and the regulatory maps for the 63 plant species are available at PlantRegMap (http://plantregmap.cbi.pku.edu.cn/) [5, 21]. The genome browser for data visualization is available at http://plantregmap.cbi.pku.edu.cn/cis-map.php [16]. The FunTFBS tool and the evaluation dataset are available at http://plantregmap.cbi.pku.edu.cn/funtfbs.php [22]. See Additional file 1: Table S4 for more details about the resource accessibility.

Authors’ Contributions

GG and JJ conceived and designed the research. FT performed all the data analysis and tool development. DCY set up the website for data visualization and user interaction. YQM and DCY set up the genome browser for cis-Map. FT, JJ and GG wrote the manuscript with contribution from all the authors.

Competing Interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable
References

1. Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, et al: An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature 2015, 517:571-575.

2. Jin J, He K, Tang X, Li Z, Lv L, Zhao Y, Luo J, Gao G: An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors. Mol Biol Evol 2015, 32:1767-1773.

3. O’Malley RC, Huang SS, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR: Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 2016, 165:1280-1292.

4. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al: Determination and inference of eukaryotic transcription factor sequence specificity. Cell 2014, 158:1431-1443.

5. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G: PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 2017, 45:D1040-D1045.

6. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al: An expansive human
regulatory lexicon encoded in transcription factor footprints. *Nature* 2012, **489**:83-90.

7. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al: Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep* 2014, **8**:2015-2030.

8. Baxter L, Jironkin A, Hickman R, Moore J, Barrington C, Krusche P, Dyer NP, Buchanan-Wollaston V, Tiskin A, Beynon J, et al: Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell* 2012, **24**:3949-3965.

9. Hupalo D, Kern AD: Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol Biol Evol* 2013, **30**:1729-1744.

10. Van de Velde J, Heyndrickx KS, Vandepoele K: Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. *Plant Cell* 2014, **26**:2729-2745.

11. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al: An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* 2013, **45**:891-898.

12. Wang Y, Fan X, Lin F, He G, Terzaghi W, Zhu D, Deng XW: Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proc Natl Acad Sci USA* 2014, **111**:10359-10364.
13. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.** *Nature* 2007, **450**:219-232.

14. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al: **A high-resolution map of human evolutionary constraint using 29 mammals.** *Nature* 2011, **478**:476-482.

15. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**:1112-1115.

16. cis-element Map of Plant Transcriptional Regulatory Map. [http://plantregmap.cbi.pku.edu.cn/cis-map.php](http://plantregmap.cbi.pku.edu.cn/cis-map.php). Accessed 23 Nov. 2018.

17. Song L, Huang SC, Wise A, Castanon R, Nery JR, Chen H, Watanabe M, Thomas J, Bar-Joseph Z, Ecker JR: **A transcription factor hierarchy defines an environmental stress response network.** *Science* 2016, **354**.

18. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease.** *Nat Biotechnol* 2012, **30**:159-164.

19. 1001 Genomes Consortium: **1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana.** *Cell* 2016, **166**:481-491.

20. Huang YC, Niu CY, Yang CR, Jinn TL: **The Heat Stress Factor HSFA6b**
Connects ABA Signaling and ABA-Mediated Heat Responses. *Plant Physiol* 2016, **172**:1182-1199.

21. Plant Transcriptional Regulatory Map. [http://plantregmap.cbi.pku.edu.cn](http://plantregmap.cbi.pku.edu.cn). Accessed 23 Nov. 2018.

22. Functional Transcription Factor Binding Site Screening tool. [http://plantregmap.cbi.pku.edu.cn/funtfbs.php](http://plantregmap.cbi.pku.edu.cn/funtfbs.php). Accessed 23 Nov. 2018.
**Figure Legends**

**Fig. 1.** Genome-wide conservation landscapes of 63 green plants. **a,** Phylogenetic tree for 63 species of seven lineages, where the branch length within each group is shown as the substitution number per site at four-fold degenerate sites. **b,** The summary of genomes aligned and conserved for each group. Left: The phylogenetic tree shows the evolutionary relationship among groups. The width of the triangle corresponds to the maximum divergence time of involved species, and the height corresponds to the species number (as shown in brackets) in each group. Middle and Right: The box plots show the percentage of genomes aligned (Middle) and conserved (Right) for each group. **c,** The proportion of non-coding regions in the genomes and their relative conservation ratio (proportion of non-coding regions conserved / proportion of coding regions conserved) for plants (63 species), animals (*H. sapiens*, *D. melanogaster* and *C. elegans*) and yeasts (*S. cerevisiae*). Each data point represents a species and its color represents the group. **d,** The screenshot of genome alignment and conservation resource in the genome browser.

**Fig. 2.** Screening for functional regulatory interactions by coupling the base-varied binding affinities of transcription factors (TFs) and consistent evolutionary constraints on their binding sites (TFBSs). **a,** An example illustrates the consistency between the based-varied binding affinities (base frequency in the binding motifs) of the TFs and the evolutionary constraints on a functional TFBS. **b** and **c,** Enrichment analysis of the TFBSs showing consistency or inconsistency between conservation scores and
base frequencies in motifs (b) and the TFBSs screened by DNase-seq footprints (DNase-seq footprint TFBS), conserved elements (Conserved TFBS), and evolutionary footprints (FunTFBS) (c) in “Less reliable”, “Highly reliable”, and “Functional” TFBSs of the evaluation dataset. The percentage of the TFBSs, directly scanned by motifs, supported by each class of evaluation dataset is used as background. (Error bar indicates the standard deviation of the average fold change from 1000 subsampling; p-value was calculated based on the results of the 1000 subsampling, *** p-value < 0.001.) d and e, An eQTL (A to T substitution, highlighted in green) located on a TFBS of AT5G67580 predicted by FunTFBS (d) and the significant difference in the expression level of its target gene AT4G00650 (e) (Wilcoxon rank sum test, *** p-value < 0.001.) f, Enriched GO terms for the target genes of AT3G22830 inferred by FunTFBS.
Fig. 1

(a) Evolutionary tree showing the relationship between different plant species. The tree is color-coded to indicate the number of species in each clade: BOP clade (9), PACMAD clade (9), Lamiales (4), Brassicales (18), Malpighiales (5), Rosales (6), and Fabales (12).

(b) Graph showing the number of substitutions per site over time in million years ago. The x-axis represents time in million years ago, and the y-axis represents the number of substitutions. The graph is divided into two categories: Aligned (%) and Conserved (%).

(c) Scatter plot showing the relative proportion of non-coding regions conserved. The x-axis represents the proportion of non-coding regions in genomes (%), and the y-axis represents the relative proportion of conserved regions (%). The plot includes species such as D. melanogaster, S. cerevisiae, C. elegans, and M. simplifie.

(d) Alignment plot showing the comparison of TFBs across different species. The plot includes Arabidopsis thaliana, Brassica rapa, and Carica papaya. The x-axis represents the sites, and the y-axis represents the TFBs. The plot is color-coded to show the conservation across different species.
Fig. 2

a) TAIR10 Genes
PhyloP
Base frequency
T A C A C G T G T C T

b) Fold enrichment
Inconsistent
Consistent

Less reliable
Highly reliable
Functional

-Log10 (FDR)

ref (A)
alt (T)

response to heat
response to temperature stimulus
protein folding
response to abiotic stimulus
response to high light intensity
response to inorganic substance
response to stress
response to calcium ion
response to light intensity
response to metal ion