Journal of Genetics and Genomics xxx (2018) 1–7

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original research

Systematic identification and annotation of multiple-variant compound effects at transcription factor binding sites in human genome

Si-Jin Cheng [a, b], Shuai Jiang [a, b], Fang-Yuan Shi [a, b], Yang Ding [a, b], Ge Gao [a, b, *]

[a] State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Center for Bioinformatics, Peking University, Beijing 100871, China
[b] Beijing Advanced Innovation Center for Genomics, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 31 December 2017
Received in revised form
3 May 2018
Accepted 25 May 2018
Available online xxx

Keywords:
Compound effect
Transcription factor binding site
Variant annotation
Bioinformatics
Genetic variants

ABSTRACT

Understanding the functional effects of genetic variants is crucial in modern genomics and genetics. Transcription factor binding sites (TFBSs) are one of the most important *cis*-regulatory elements. While multiple tools have been developed to assess functional effects of genetic variants at TFBSs, they usually assume that each variant works in isolation and neglect the potential "interference" among multiple variants within the same TFBS. In this study, we presented COPE-TFBS (Context-Oriented Predictor for variant Effect on Transcription Factor Binding Site), a novel method that considers sequence context to accurately predict variant effects on TFBSs. We systematically re-analyzed the sequencing data from both the 1000 Genomes Project and the Genotype-Tissue Expression (GTEx) Project via COPE-TFBS, and identified numbers of novel TFBSs, transformed TFBSs and discordantly annotated TFBSs resulting from multiple variants, further highlighting the necessity of sequence context in accurately annotating genetic variants. COPE-TFBS is freely available for academic use at http://cope.cbi.pku.edu.cn/.

1. Introduction

Transcription factors (TFs) bind to *cis*-regulatory elements (transcription factor binding sites, TFBSs) to regulate the transcription of downstream genes (Latchman, 1997). Variants within TFBSs can impact the binding strength of TFs (Maurano et al., 2015) and participate in the biogenesis of human diseases, including cancers (Huang et al., 2014; Liu et al., 2017).

With the rapidly decreasing cost of deep sequencing, numerous genetic variants have been identified across the whole human genome, of which more than 7.8 million lie within TFBSs (Sherry et al., 2001). Many tools have been used to evaluate the functional effects of these TFBS variants (Boyle et al., 2012; Fu et al., 2014; Coetzee et al., 2015; Zuo et al., 2015; Ward and Kellis, 2016; Kumar et al., 2017). Notably, these tools generally handle each variant independently, and neglect the potential "interference" resulting from multiple variants within the same TFBS.

Here, we presented the tool COPE-TFBS (Context-Oriented Predictor for variant Effect on Transcription Factor Binding Site) to annotate the TFBS variant effects in a context-sensitive way as an integrative module of our COPE variant effect annotation framework (Cheng et al., 2017). Different from existing tools, COPE-TFBS reconstructs genomic sequences from phased genotype to take multiple variants into consideration. By re-analyzing the sequencing data from both the 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015) and the Genotype-Tissue Expression (GTEx) Project (The GTEx Consortium, 2013), we first systematically investigated the distribution of TFBS variants in human genome, and found that multiple variants within the same TFBS may produce different functional effects in combination than individually. The web server and the standalone package of COPE-TFBS are freely available at http://cope.cbi.pku.edu.cn/.

2. Results

2.1. Overview of COPE-TFBS

COPE-TFBS employed Position Weight Matrices (PWMs) to quantitatively measure the functional effects of variants at TFBSs.

* Corresponding author. Beijing Advanced Innovation Center for Genomics, Peking University, Beijing 100871, China.
E-mail address: gaog@mail.cbi.pku.edu.cn (G. Gao).

https://doi.org/10.1016/j.jgg.2018.05.005

All PWMs were downloaded from Fu et al. (2014), including PWMs from ENCODE (The ENCODE Project Consortium, 2012), TRANSFAC (Wingender et al., 1996) and JASPAR (Mathelier et al., 2016) databases. All known TFBSs validated by high-throughput experiments were downloaded from Fu et al. (Fu et al., 2014; The ENCODE Project Consortium, 2012).

Two independent modules were included in COPE-TFBS: TFBS gain module and TFBS breaking module (Fig. 1A). To evaluate the complex effects resulting from multiple variants, genomic sequences reconstructed from phased genotype were used to predict variant effects on promoter. For TFBS gain module, variants within promoter regions (defined as −2.5 kb from the transcription start site) were identified first. Then, reconstructed promoter sequences were extracted based on the phased promoter region variants. Finally, COPE-TFBS scanned all sub-sequences that contained variants, and then evaluated their statistical significance by TFM-Pvalue (Touzet and Varre, 2007), a tool calculating the probability that the background model can achieve a score larger than or equal to the observed PWM score. Only novel TFBSs with *p*-value for mutant sequence ≤ 4e-8 and *p*-value for reference sequence > 4e-8 were reported as putative novel TFBSs. The cut-off was adopted from Touzet and Varre (2007), which was widely used to evaluate the significance of PWM scores, including FunSeq2 (Fu et al., 2014) and RegulomeDB (Boyle et al., 2012). For TFBS breaking module, variants overlapped with known TFBSs were identified first. Then, TFBS sequences were reconstructed based on the phased variants. Finally, COPE-TFBS compared the PWM scores of both reference and mutant sequences and reported the PWM score alteration.

2.2. Complex compound effects resulting from multiple variants within the same TFBS

By applying COPE-TFBS to the sequencing data from both the 1000 Genomes Project and the GTEx Project, we found that multiple variants within the same TFBS may interfere with each other and result in complex compound effects (Figs. 1B and S1) that differ from individual effects: 1) putative novel TFBS: it is created by multiple variants and none of the variants could result in the novel TFBS independently; 2) transformed TFBS: it means that the novel TFBS is transformed from a known TFBS by multiple variants, and

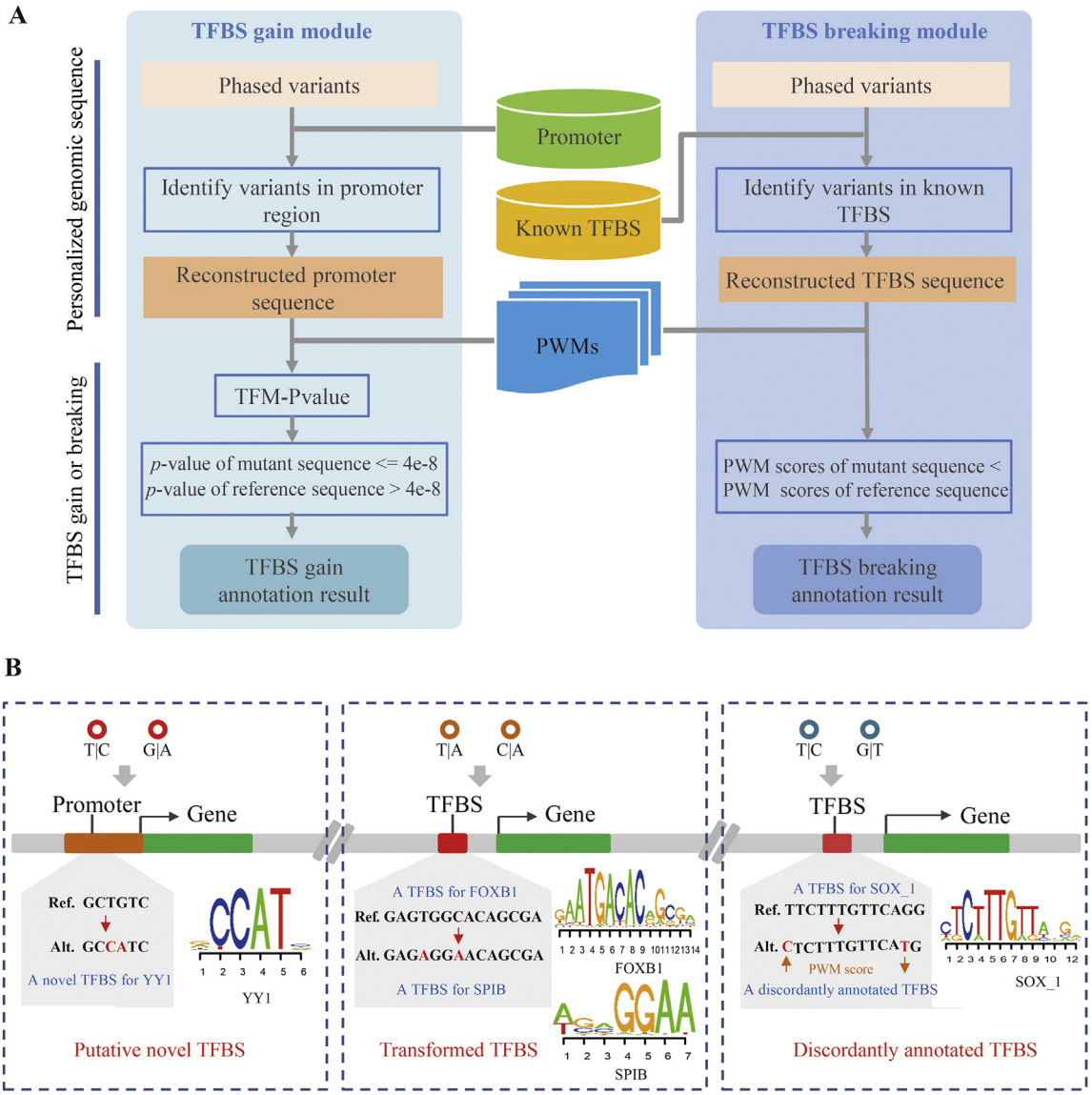

Fig. 1. The workflow of COPE-TFBS and the method for identification of complex compound effects. **A**: The workflow of COPE-TFBS. **B**: Schematic diagram of complex compound effects resulting from multiple variants within a TFBS.

Please cite this article in press as: Cheng, S.-J., et al., Systematic identification and annotation of multiple-variant compound effects at transcription factor binding sites in human genome, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.05.005