oi 10.18699/bgrs2024-1.4-29

Integration of genomic and clinical data: assessment of the influence of SNPs in the TATA-box region of human genes on changes in the expression of these genes and their associations with clinical and phenotypic observations

Podkolodnyy N.L.[1, 2]*, Filonov S.V.[1], Podkolodnaya O.A.[1], Tverdokhleb N.N.[1], Ponomarenko P.M.[1], Rasskazov D.A.[1], Bogomolov A.G.[1], Ponomarenko M.P.[1]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia*

** pnl@bionet.nsc.ru*

Key words: TATA box; affinity; TBP; single nucleotide polymorphism; database; genome-wide analysis; Human_SNP_TATAdb database; ClinVar database; candidate SNP markers of diseases

Motivation and Aim: Genome-wide association studies (GWAS) have shown that most SNPs that are significantly associated with disease susceptibility are located in non-coding regions [1–3], and more than 90 % of them are located in regulatory elements [4]. Now, one of the most studied regulatory regions is the TATA box region in the promoter, the sequence of which determines the affinity of the TBP protein (TATA binding protein), which is a key transcription initiation factor. Mutations in this region can affect the binding of the TBP protein to the promoter and, consequently, gene expression [5, 6]. Previously, the Institute of Cytology and Genetics SB RAS (ICG) has developed a method for predicting the affinity of TBP to gene promoters based on a three-step binding mechanism, including sliding of TBP along DNA, stopping of TBP at the binding site, and fixing the TBP-promoter complex due to bending of the DNA helix [7]. The method showed a high correlation of theoretical predictions with measured values during repeated experimental testing by independent groups of researchers. Based on this model, the ICG has developed the SNP_TATA_Z-tester and SNP_TATA_Comparator web services, which allow calculating a statistical estimation of the SNP-induced change in the affinity of TBP binding to the human gene promoter and predicting changes in expression that may be associated with a genetic predisposition to diseases or phenotypic features of the organism. Using this web service, we previously identified candidate SNP markers for autoimmune diseases [8], behavioral disorders [9], chronopathologies [10] and other diseases. In this work, we integrated into a database information about SNPs in human gene promoters, obtained by automatic extraction from various heterogeneous data sources, as well as the results of assessing the affinity of TBP to the promoter and the specificity of the TBP binding site using a three-step binding model and assessing their effect on gene expression for the reference genome promoters and promoters with SNPs as well as clinical and phenotypic observations associated with SNPs presented in the ClinVar database. This makes it possible to identify candidate genetic markers of various clinical and phenotypic manifestations among SNPs located in the TATA box region.

Methods and Algorithms: Data on genes and their attributes, transcription starts and transcripts were obtained from the Ensembl web service. To access the services and

databases used in the work, the Bioconductor library of the R language was used, with the following packages:
1. biomaRt is a package that provides an interface to the ENSEMBL collection of databases, allowing large volumes of data to be retrieved in a unified way and used in data analysis in Bioconductor.
2. BSgenome.Hsapiens.NCBI.GRCh38 is package that provides access to the Homo sapiens (Human) genome sequence provided by NCBI (GRCh38.p13).
3. SNPlocs.Hsapiens.dbSNP155.GRCh38 is a dbSNP 155 access package including information on 949,021,448 SNPs in chromosomes 1-22, X, Y and MT.
4. Information on the associations of snps with phenotypic and clinical observations presented in the ClinVar database (Apr, 22, 2024).
For each promoter, SNPs located within [–90,–1] from the start of transcription were identified. The affinity of TBP for DNA was calculated using a three-step binding model previously developed at the Institute of Cytology and Genetics SB RAS [7] and a multi-threaded high-performance version of the SNP_TATA_Z-tester program also implemented by us. This program also allows you to evaluate the statistical significance of changes in the affinity of the TBP protein for the promoter due to point nucleotide substitutions (SNPs) in the promoter using a z-test.
Results: A software package was developed to extract data from available information resources about all human genes and their transcripts, promoter sequences and SNPs localization in them. A multi-threaded version of the SNP_TATA_Comparator program was developed to assess changes in the affinity of TBP for DNA and a genome-wide analysis of the influence of known SNPs in the promoters of all human genes on changes in the expression of these genes was carried out. Integration of genomic data with clinical and phenotypic observations presented in the clinVar database was carried out. As a result, the database Human_SNP_TATAdb contains the following information [11]:
● 62,603 genes, of which 19,314 encode proteins.
● 117,414 transcripts, of which 63,141 encode proteins.
● 5,305,816 SNP variants in gene promoters in the [–90,–1] interval from the start of transcription, of which 3,199,285 are in the promoters of protein-coding genes.
● For 445,875 SNP variants in the promoter of a protein-coding gene, we predicted that they statistically significantly (p-value < 0.05) change the level of TBP affinity for this promoter. Among them, for 3,847 SNP variants, there are clinical and phenotypic observations presented in the ClinVar database.
The results of genome-wide data analysis are presented, including the features of the distribution of genes by the number of transcripts, the distribution of SNPs affecting the affinity of TBP to DNA by positions within promoters, as well as patterns linking the affinity of TBP to the promoter, the specificity of the TBP binding site to the promoter and other characteristics of promoters. The results of genome-wide analysis showed that the affinity of TBP to the promoter and the specificity of its binding site are statistically related to other characteristics of promoters important for the functional classification of promoters and the study of the features of differential gene expression.
Conclusion: This work presents the integrated database Human_SNP_TATAdb, which includes information on single nucleotide polymorphisms in human gene promoters, the results of assessing the affinity of TBP to the promoter using a three-step binding model, and assessing their impact on gene expression for wild-type promoters and promoters with single nucleotide polymorphism as well as clinical and phenotypic observations associated with snps presented in the clinvar database. The results of genome-wide

analysis showed that the affinity of TBP for the promoter and the specificity of its binding site are statistically associated with other characteristics of promoters that are important for the functional classification of promoters and the study of differential gene expression patterns.

We have shown that the affinity of the TBP protein to the promoter, the specificity of the TBP binding site to the promoter, and assessments of changes in these characteristics with single nucleotide polymorphisms presented in the Human_SNP_TATAdb database may be important for identifying candidate markers of genetic susceptibility to diseases, identifying and functional interpretation of classes of promoters that are similar in the mechanism of regulation of the early stage of transcription initiation, etc.

Funding: The study is supported by the Government Budget Projects (No. FWNR-2022-0020 and No. 0251-2022-0005).

References

1. Hindorff L.A., Sethupathy P., Junkins H.A., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106(23):9362-9367. doi 10.1073/pnas.0903103106
2. French J.D., Edwards S.L. The role of noncoding variants in heritable disease. *Trends Genet*. 2020(36):880-891. doi 10.1016/j.tig.2020.07.004
3. Chandra V., Bhattacharyya S., Schmiedel B.J. et al. Promoter interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat Genet.* 2021;(53):110-119. doi 10.1038/s41588-020-00745-3
4. Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. doi 10.1126/science.1222794
5. Ravarani C., Chalancon G., Breker M. et al. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nat Commun*. 2016;7:10417
6. Ponomarenko M., Rasskazov D., Arkova O. et al. How to Use SNP_TATA_Comparator to Find a Significant Change in Gene Expression Caused by the Regulatory SNP of This Gene's Promoter via a Change in Affinity of the TATA-Binding Protein for This Promoter. *Biomed Res Int*. 2015;2015:359835
7. Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A Step-by-step Model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl Biochem Biophys*. 2008;419:88-92. doi 10.1134/S1607672908020117
8. Ponomarenko M.P. et al. Candidate SNP markers of genderbiased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front Immunol*. 2016a;7:130. doi 10.3389/fimmu.2016.00130
9. Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A. et al. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics*. 2016;17(Suppl. 14):995. doi 10.1186/s12864-016-3353-3
10. Ponomarenko P., Rasskazov D., Suslov V., Sharypova E., Savinkova L., Podkolodnaya O., Podkolodny N.L., Tverdokhleb N.N., Chadaeva I., Ponomarenko M., Kolchanov N. Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed Res Int*. 2016b;2016:8642703. doi 10.1155/2016/8642703
11. Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilov J Genet Breed*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85