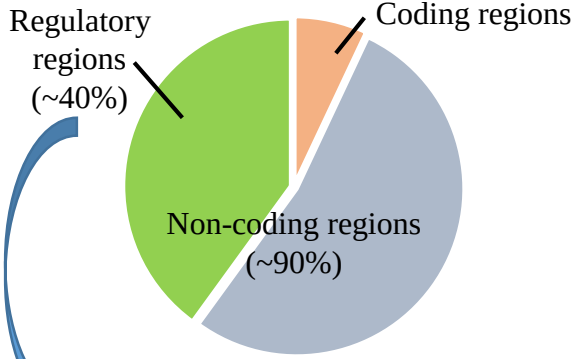
Software pipeline for the analysis of the functional role of nucleotide substitutions in regulatory regions of genes and its testing on polymorphisms associated with obesity

Ekaterina A. Matrosova[1], Vadim M. Efimov[1,2], Elena V. Ignatieva[2]

[1] Novosibirsk State University, Novosibirsk, Russia;

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

Distribution of the nucleotide substitutions associated with diseases in the genome (according to GWAS)



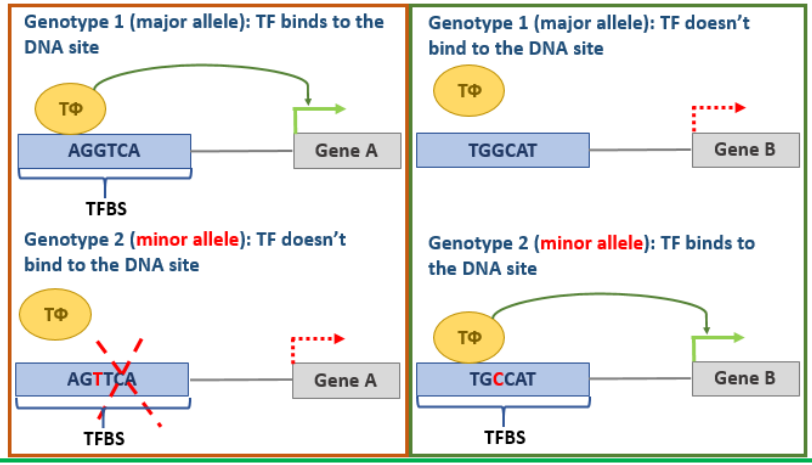
[Science. 2012; 337(6099):1190–1195]

However, the molecular genetic mechanisms of the influence of the revealed polymorphisms on the development of pathologies have been studied insufficiently.

Genetic variants in regulatory regions

- **causative** (affect the functional activity of TFBSs)
- may mark specific haplotypes containing **causal SNPs**

Possible mechanisms of the influence of regulatory polymorphisms on the functions of TFBSs: replacing a major allele with a minor one can disrupt TFBS (left side of the picture) or lead to its *de novo* appearance

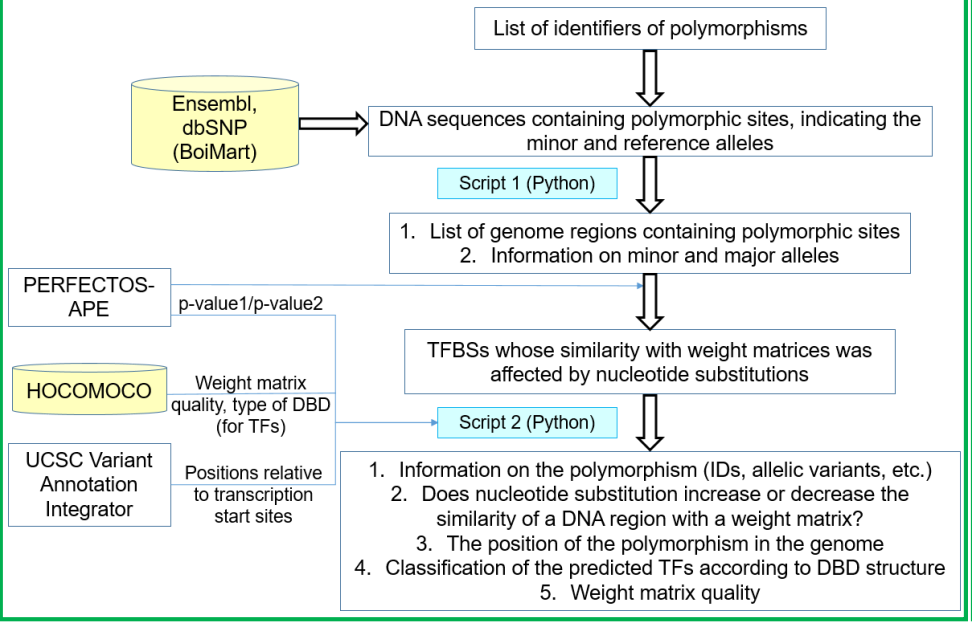


The aim was to create a software pipeline to evaluate the potential effects of nucleotide substitutions on TFBSs. Functionality of the software pipeline was tested on nucleotide variants in regulatory regions of the *BDNF* gene associated with the development of monogenic forms of obesity.

RESULTS

1 We have developed a software pipeline that includes two original Python scripts that integrate and analyze information obtained from the UCSC Variant Annotation Integrator and PERFECTOS-APE programs, as well as from the dbSNP, Ensembl, and HOCOMOCO databases.

The scheme of the software pipeline, which allows to analyze the effects of nucleotide substitutions on TFBSs



2 The software pipeline was tested on 5 sets of polymorphisms.

The analyzed sets of polymorphisms with marked localization relative to the *BDNF* gene

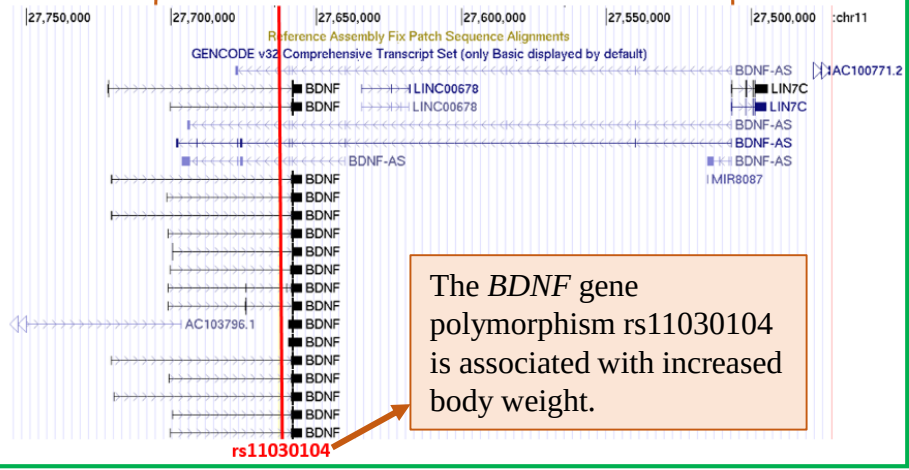
№5: *BDNF* gene body, MAF>0.01 (185 SNPs)

№4: The promoter regions of the *BDNF* gene -2000/-1, MAF<0.01 (301 SNPs)

№2: Control – in the vicinity of rs11030104, MAF>0.01 (206 SNPs)

№3: Control – in the vicinity of rs11030104, MAF<0.01 (220 SNPs)

№1: Case – in linkage disequilibrium, $R^2 \geq 0.6$ (87 SNPs)



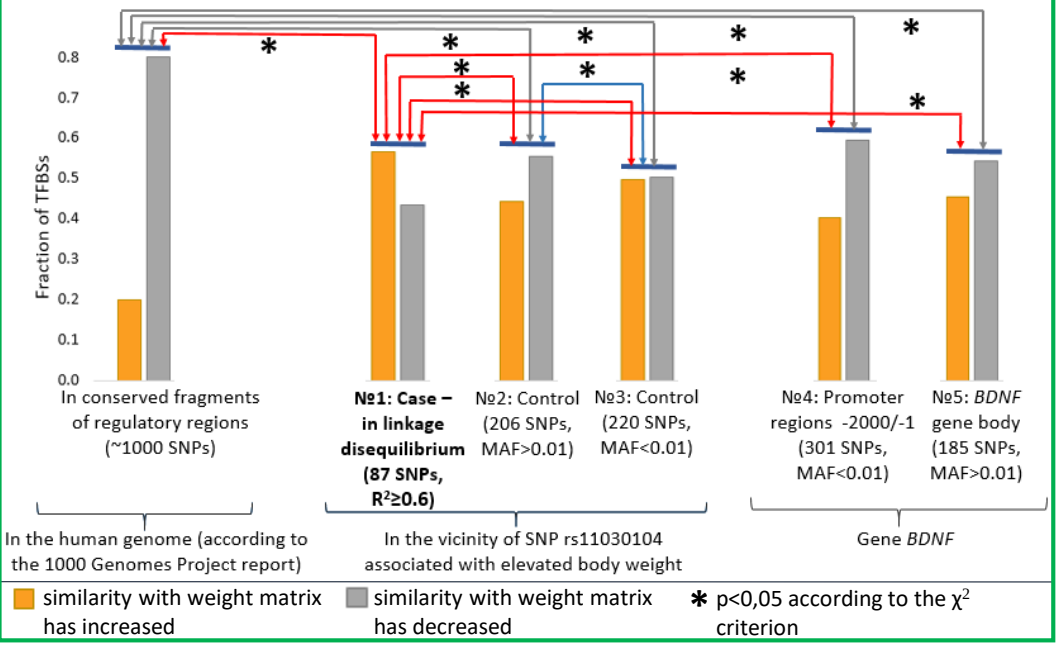
3 According to the 1000 Genomes Project report, for SNPs located in conserved fragments of regulatory regions of the human genome the ratio between the number of nucleotide substitutions leading to occurrence of TFBSs and the number of substitutions damaging the TFBSs is 1:4.

All 5 SNPs sets showed a significant difference in the proportions between positive and negative effects of nucleotide substitutions on TFBSs in comparison with this ratio. For set №3, the ratio is close to 1:1. For sets №2, 4, 5, the ratio is only qualitatively similar to 1:4. For set №1, positive effects of SNPs are more often than negative effects.

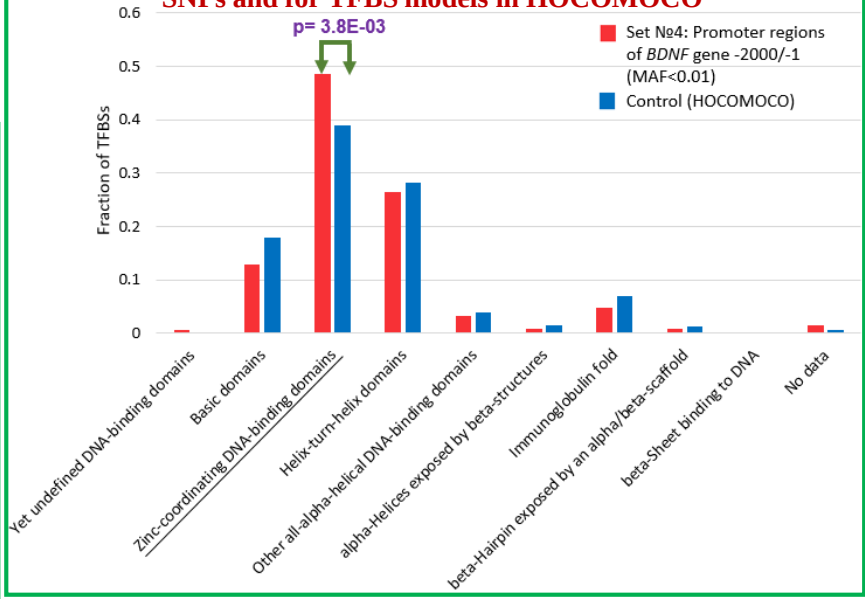
4 When comparing the distributions of TF sets with the distribution of TFBS models in HOCOMOCO by the classes of their DNA-binding domains (DBDs), significant differences were identified using Kullback criterion for the set №4 ($p=10^{-2}$).

When comparing the proportion of TFs of a particular DBD class, significant differences were found between sets of TFs for the class of Zinc-coordinating DBDs.

The effect of nucleotide substitutions on the similarity of sites with weight matrices



Fractions of TFs with the DBD of certain type for the analyzed set of SNPs and for TFBS models in HOCOMOCO



Conclusion: We have developed a software pipeline that takes a list of SNP identifiers as an input and allows us to predict the potential effects of nucleotide substitutions on potential TFBSs. The efficiency of the pipeline was tested on 5 sets of polymorphisms in the regulatory regions of the *BDNF* gene associated with obesity.

Acknowledgment: Supported by the State Budgeted Project 0324-2019-0040-C-01.