DOI 10.18699/SBB-2022-038

Interpreting non-coding genome variation with DNA sequence motifs

Kulakovskiy I.V.[1, 2]*, Makeev V.J.[2, 3]

[1] *Institute of Protein Research, RAS, Pushchino, Russia*

[2] *Vavilov Institute of General Genetics, RAS, Moscow, Russia*

[3] *Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

* *ivan.kulakovskiy@gmail.com*

Key words: non-coding genome variation, regulatory SNP, transcription factor binding, sequence motifs

Nowadays, it is impossible to imagine the future of human healthcare without personalized biomedicine and pharmacogenomics. While genome sequencing is becoming cheaper and more reliable, overall interpretability and thus the practical utility of the individual genome data remain limited. On the one hand, genome-wide association studies1 reveal millions of statistical associations between genome variants and various traits, including predisposition to complex diseases, such as immune disorders and cancer. On the other hand, the detected associations do not immediately reveal the genuine causal variants, particularly, because of linkage disequilibrium. Prioritization of the causal variants requires statistical fine-mapping or functional annotation in a wet lab or *in silico*. Interpretation and prioritization are especially challenging for the non-coding variants in gene regulatory regions, which do not alter protein sequence and function directly. Instead, they disturb the cellular state by altering gene expression and protein abundance. For many pathologic phenotypes, such as immune disorders, the majority of causal variants are located in the transcription regulatory regions. Annotating individual regulatory elements of gene regulatory regions facilitates functional annotation of non-coding variants, and reveals underlying molecular mechanisms.

We present a versatile bioinformatics framework for the analysis of high-throughput sequencing data in the context of gene regulation and transcription factor binding, including *de novo* DNA motif discovery (ChIPMunk [1]), pattern matching (SARUS [2] and MoLoTool [3]), motif comparison (MACRO-APE [4]), and basic annotation of noncoding variants (PERFECTOS-APE [5]). With this toolkit, we assembled the well-known HOCOMOCO [3] collection of sequence motifs recognized by human TFs and used it to annotate single-nucleotide variants linked with autoimmune diseases. On top of that, we show how advanced statistics backed up with motif analysis allows for mapping causative allele-specific regulatory elements (ADASTRA [6] and ANANASTRA [7]), and discuss surprising pitfalls and peculiarities of machine learning applications in the analysis of regulatory variants [8].

Acknowledgements: Russian Science Foundation (20-74-10075 to I.V.K.).

References

1. Kulakovskiy I.V. et al. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*. 2010;26(20):2622-3. doi: 10.1093/bioinformatics/btq488.
2. Kulakovskiy I.V. et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* 2013;41(D1):D195-D202. doi: 10.1093/nar/gks1089.
3. Kulakovskiy I.V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018;46(D1):D252-D259. doi: 10.1093/nar/gkx1106.

4. Vorontsov I.E. et al. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol Biol*. 2013;8(1):23. doi: 10.1186/1748-7188-8-23.
5. Vorontsov I.E. et al. PERFECTOS-APE: Predicting regulatory functional effect of SNPs by approximate P-value estimation. In: Proceedings of the 6th International Conference on Bioinformatics Models, Methods and Algorithms, BIOINFORMATICS 2015. Lisbon, 2015;102-108. doi: 10.5220/0005189301020108.
6. Abramov S. et al. Landscape of allele-specific transcription factor binding in the human genome. 2015. *Nat Commun*. 2021;12:2751. doi:10.1038/s41467-021-23007-0.
7. Boytsov A. et al. ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs. *Nucleic Acids Res.* 2022;gkac262. doi:10.1093/nar/gkac262.
8. Penzar D.D. et al. What do neighbors tell about you: the local context of cis-regulatory modules complicates prediction of regulatory variants. *Front Genet*. 2019;10:1078. doi: 10.3389/fgene.2019.01078.