DOI 10.18699/SBB-2022-227

Annotation of uORFs in the OMIM genes allows to identify disease-causing variants in 5`UTRs

Filatova A.[1]*, Reveguk I.[2], Piatkova M.[3, 4], Bessonova D.[5], Kuziakova O.[6], Demakova V.[5], Romanishin A.[6, 7], Fishman V.[8, 9], Imanmalik Y.[8], Chekanov N.[8], Skitchenko R.[10], Barbitoff Y.[11, 12, 13], Kardymon O.[8], Skoblov M.[1]

[1] *Research Centre for Medical Genetics, Moscow, Russia*
[2] *Laboratoire de Biologie Structurale de la Cellule, École Polytechnique, Paris, France*
[3] *Institute of Chemistry, FEB RAS, Vladivostok, Russia*
[4] *Institute of High Technologies and Advanced Materials, Far Eastern Federal University, Vladivostok, Russia*
[5] *Medical Center, Far Eastern Federal University, Vladivostok, Russia*
[6] *Institute of Life Sciences and Biomedicine, Far Eastern Federal University, Vladivostok, Russia*
[7] *Institute of Life Sciences, Immanuel Kant Baltic Federal University, Kaliningrad, Russia*
[8] *Bioinformatic group, AIRI, Moscow, Russia*
[9] *Molecular Mechanisms of Ontogenesis, Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[10] *Computer Technologies Laboratory, ITMO University, St. Petersburg, Russia*
[11] *Bioinformatics Institute, St. Petersburg, Russia*
[12] *Dpt. of Genomic Medicine, D.O. Ott Research Institute of Obstetrics, Gynaecology, and Reproductology, St. Petersburg, Russia*
[13] *Dpt. of Genetics and Biotechnology, St. Petersburg State University, St. Petersburg, Russia*
* *maacc@yandex.ru*

Key words: uORF, 5`UTR, translation initiation start, Mendelian disorders, hereditary diseases

Motivation and Aim: To date, the identification of disease-causing variants generally focuses on the coding part of the genes, although an increasing number of studies highlight the role of non-coding variants in diseases development. It was previously shown that variants in the 5`UTR could lead to hereditary diseases through disruption of upstream open reading frames (uORF). However, the identification of such variants in routine clinical genetic testing still remains a challenge since there is no precise annotation of uORFs in human disease-associated genes.
Methods and Algorithms: We used different types of publicly available data (Ribo-seq, mRNA-seq, CAGE data previously predicted by uORF) and self-generated data (Kozak scores) to manually annotate upstream translation initiation sites (TIS) in ~3,600 OMIM genes. Based on obtained data, we created a list of uORFs and investigated their various features. Experimental validation was conducted using dual luciferase assay. Two software tools were written in Python and are publicly available on the Github repository.
Results: We annotated about 5 thousand upstream TISs in 1835 OMIM genes. The major group of these TISs was related to uORFs. Comparison of our annotation with the previous studies produced a list of "high confidence" uORFs. In addition, we created a tool to assess the effects of nucleotide variants located in uORFs. This allowed us to identify variants from HGMD/ClinVar that disrupt uORFs and thereby could lead to hereditary diseases. We noticed that the distribution of uORFs-affecting variants differs between pathogenic variants and population variants from gnomAD. In addition, for

several genes, we performed luciferase experiments to test the translation of the predicted uORFs and to confirm the pathogenicity of the mutations identified in these uORFs. Finally, drawing on manually curated data, we developed a machine-learning algorithm that allows us to predict the TISs and uORFs in other human genes.
Conclusion: Thus, we attempted to obtain qualitative annotation of uORFs in the OMIM genes, which may be useful in interpreting the patient-derived variants in 5`UTRs.
Acknowledgements: This study was performed using the infrastructure of AIRI, Artificial Intelligence Research Institute (Moscow, Russia). Preliminary analysis of Ribo-seq data was carried out within the state assignment of the Ministry of Science and Higher Education of the Russian Federation for RCMG.