Selection pressure on breast cancer somatic mutations revealed by bioinformatics sequence analysis

I.E. Vorontsov[1], I.V. Kulakovskiy[2,1], G.N. Khimulya[1], D.D. Nikolaeva[3], V.J. Makeev[1,2,4]

[1]*Vavilov Institute of General Genetics, Gubkina str. 3, Moscow, Russia,* ilya.vorontsov@vigg.ru

[2]*Engelhardt Institute of Molecular Biology, Vavilova str. 32, Moscow, Russia*

[3]*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[4]*Department of Medical and Biological Physics, Moscow Institute of Physics and Technology, Moscow Region, Russia*

Among different variations of the human genome single nucleotide variants, SNVs, are the most common. SNVs located in coding regions may directly affect function of a particular protein through alterations of the protein sequence and, consequently, the structure.

Nucleotide substitutions occurring in regulatory regions do not alter the protein but may change expression of the corresponding genes. In particular, SNVs in promoters and enhancers may alter transcription factor (TF) DNA binding and thus affect efficiency of transcription initiation. With hundreds of human TFs binding patterns known it is finally possible to predict regulatory effects of mutations purely by sequence analysis *in silico*. In the past, SNVs were primarily studied in a population context as single-nucleotide polymorphisms, SNPs. The high-throughput sequencing gave birth to principally new data on somatic mutations, in particular, those emerging in cancer.

Here we discuss a new version of PERFECTOS-APE (Vorontsov *et al.*, 2015), the software to PrEdict Regulatory Functional Effect of SNVs by Approximate P-value Estimation. We applied PERFECTOS-APE to analyze somatic mutations detected in 21 breast cancer samples by Nik-Zainal *et al.*, 2012. Using HOCOMOCO (Kulakovskiy *et al.*, 2013) collection of transcription factor binding patterns we identified TFs whose binding sites were affected by somatic substitutions in breast cancer cells.

Binding sites of several transcription factors were damaged by mutations significantly more often than expected by chance. At the same time, for dozens of transcription factors binding sites were protected from mutations, i.e. were affected by them significantly less often than

expected by chance. We believe this is the evidence for positive and negative selection of cancer somatic mutations in regulatory regions.

1. I. Vorontsov *et al.* (2015) PERFECTOS-APE: Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation, *Proceedings of the 6th International Conference on Bioinformatics Models, Methods and Algorithms*, Lisbon, Portugal, p.102–108.

2. S. Nik-Zainal *et al.* (2012) Mutational Processes Molding the Genomes of 21 Breast Cancers, *Cell,* 149(5)**:**979–993.

3. Kulakovskiy *et al.* (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, *Nucleic Acids Res.*, 41(Database issue):D195–202.