Mutations in cancer genomes avoid sites recognized by transcription factors

Genome mutations occur in cancer cells much more often than in normal cells. Mutations may unpredictably disrupt various genomic elements and affect important cell functions. Until recently, for the most part only coding regions of cancer genomes have been available for studying from exome sequencing projects. So, the mutations in these regions are much better studied than in the rest of the genome. This has changed with recent advances in the large scale sequencing of complete cancer genomes, which yielded information on mutations in a whole-genome scale, including non-coding regions.
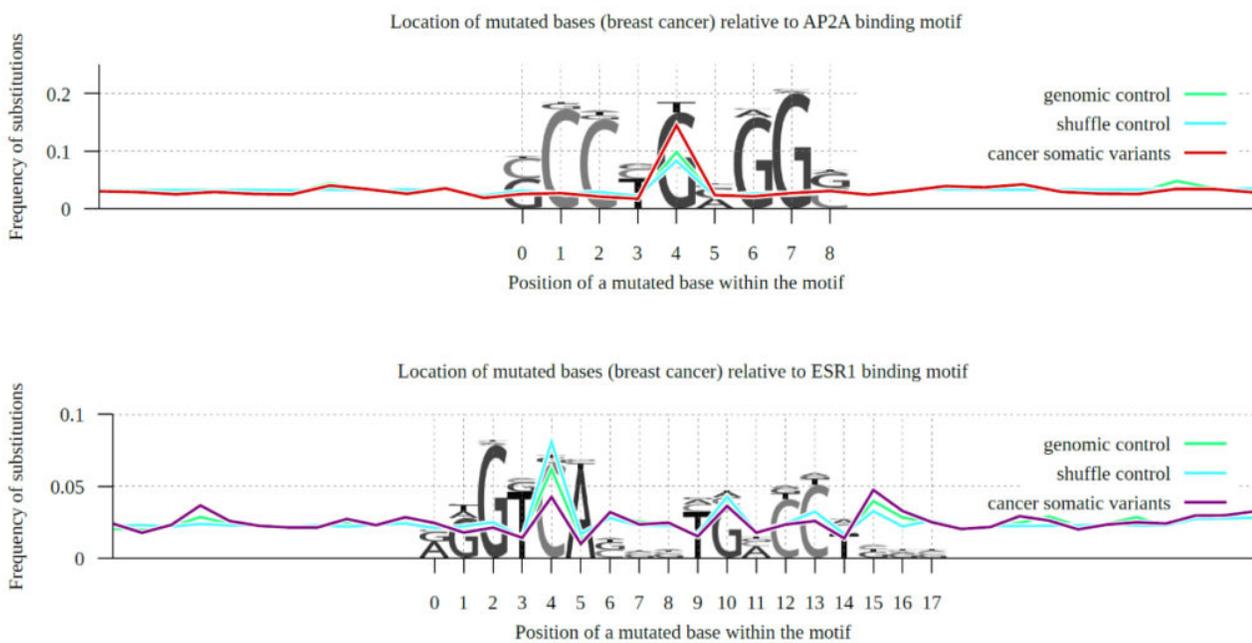

Fig. 1. Relative frequency of breast cancer mutations located within and in the vicinity of transcription factor binding sites of AP2A (top panel) and ESR1 (bottom panel). Shuffled sequences and randomly sampled genomic segments are used as controls. Somatic mutations are enriched at TGA context for AP2A binding patterns and depleted at TCA context for ESR1 binding patterns.

A notable part of the non-coding genome consists of regulatory segments, which control activity of genes in different cell types. Nucleotide sequences of such segments, promoters and enhancers, contain multiple patterns recognized by transcription factors, specific proteins that assist in launching RNA synthesis machinery or regulate its efficiency.

Genome mutations can disrupt DNA sequence patterns recognized by transcription factors, alter DNA-protein binding affinities and, in turn, modify activity of target genes. This can improve or disturb survival of mutated cell lineages. Thus, the number of cells carrying a particular mutation can depend on the mutation location. By studying many cell samples and DNA sites simultaneously, it is possible to observe which DNA patterns would accumulate more mutations than it is expected by chance. This corresponds to the positive selection of mutations in particular sequence context, and this type of selection has been reported in earlier publications. On the other hand, the negative selection, the case of specific sequence context accumulating fewer mutations than it is expected by chance, remained elusive.

We performed computational analysis of millions of non-coding mutations previously identified in samples of different cancer types. We have found that genomes of human cancer cells in some contexts accumulate fewer mutations than it is expected by chance, and these are the contexts in which mutations alter patterns recognized by transcription factors. In particular, this is true for highly mutated cytosine within TCA trinucleotide (TGA on the reverse DNA strand), attacked by APOBEC3 deaminase and converted to TTA in some cancer types, e.g. breast cancer. In some cases these trinucleotides mutate more often than expected if located within larger patterns representing binding sites of selected transcription factors (Fig. 1, top panel). This reflects positive selection. At the same time, for other patterns the mutation frequency is notably decreased (Fig. 1, bottom panel), highlighting negative selection.

Considering all mutations in various contexts we found multiple families of patterns corresponding to binding of various transcription factors being mutated significantly less often than expected. It seems this selection pressure protects cancer cells from rewiring of specific regulatory circuits. It is noteworthy, that similar effects were earlier observed for polymorphic positions of normal genomes.

Further analysis of transcription factors with mutation-proof binding motifs can pinpoint particular regulatory pathways crucial for the survivability and progression of different human cancers.

Ivan Kulakovskiy, Ilya Vorontsov, Vsevolod Makeev
Vavilov Institute of General Genetics, RAS, Moscow, Russia
Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

Publication