Evaluation of the positional correlations between whole genome annotations: novel statistical approaches development, advancement of the GenometriCorr methodologies

E.V.Zhuravleva[1], Y.A.Medvedeva[2], E.D.Stavrovskaya[1] , L.M. Cope[3], A.A. Mironov[1], V.J. Makeev[4,5], S.J. Wheelan[3]

[1]*Bioinformatics Group of Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Moscow, GSP-1, Leninskiye Gory, MSU, 1-73, Moscow, Russia,* e-mail:zhuravlka@mail.ru

A.V. Favorov [4,5]

[2]*Institute of Predictive and Personalized Medicine of Cancer, Barcelona,*

[3]*Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine , 550 North Broadway, Baltimore, Maryland, United States of America,*

[4]*Research Institute of Genetics and Selection of Industrial Microorganisms, 1-st Dorozhniy pr., 1, 117545 Moscow, Russia,*

[5]*Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333, Moscow, Gubkina str., Moscow, Russia*

e-mail:favorov@sensi.org

The result of each massive genome-scale biological experiment is a whole-genome annotation of the specific feature distribution across genome. The idea to consider correlations between genome annotations as a measure of relation and, possibly, of interaction of two features becomes generally accepted. The field continuously expands due to substantial increase of genome-scale data amounts. Nevertheless, the set of solutions that are available to the society is far from being completed.

GenometriCorr R package [1] was one of the first for the analysis of correlation relations between genome annotations. The software performs variety of different statistical analyses on each input, so that a variety of biologically significant relationships are queried. This includes looking for proximity, looking for uniform spacing, looking for increased or decreased overlaps of intervals or points. The input annotations are represented as sets of intervals.

In the presented work, we continue the GenometriCorr software development. In particular, we formulate and implement specific criteria for comparison between interval annotation and

coverage-like annotation. The second represents itself a set of intervals with a score ascribed to each interval. The score could represent, e.g. coverage of sequenced reads for this genomic region or a quality score. The approach is based on comparison of medians (Wilcoxon - Mann - Whitney test) of two distributions, which are that of the scores distributions of the reference intervals that are inside the query intervals (interval annotation) and that of the intervals that are outside. Features are independent of the two medians do not differ significantly.

We also develop new structure of GenometriCorr package, which allows using software as a library of statistical approaches for finding position correlation between annotations. Currently, all the the available statistical tests are called altogether. The current version of GenometriCorr is 1.1.14; a set of improvements is already released.

The work was supported by Russian Foundation for Basic Research (grant 14-04-01872).

1. A.V.Favorov et al. (2012) Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Comput Biol.