# Efficient Analysis of Genome Annotation Colocalization

Askar Gafurov1,2, Tomáš Vinař1, Paul Medvedev3, Broňa Brejová1

1 Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

2 LIRRM, Montpellier, France  
3 The Pennsylvania State University, USA

An annotation is a set of genomic intervals sharing a particular function or property. Examples include genes, conserved elements, and epigenetic modifications. A common task is to compare two annotations to determine if one is enriched or depleted in the regions covered by the other, suggesting some underlying biological mechanism. We study the problem of assigning statistical significance to such a comparison based on a null model representing two random unrelated annotations. To make the problem trackable, we use a Markov chain as a null model. We developed an exact quadratic-time algorithm based on dynamic programming as well as a linear-time algoritm based on a normal approximation. We also incorporate background information into our analyses by differentiating among several genomic contexts. These contexts can capture various confounding factors, such as GC content or assembly gaps. Currently, we are extending the tool to also separately analyze fixed-sized windows of the genome to pinpoint areas contributing to the enrichment the most. We demonstrate the efficiency and accuracy of our algorithms on synthetic and real data sets, including the recent human telomere-to-telomere assembly. The use of genomic contexts to correct for GC-bias resulted in the reversal of some previously published findings. In one striking example, the set of all human exons appears enriched for overlap with copy number losses but enrichment turns into depletion after taking into account assembly gaps and GC content. Our software, called MCDP2, is freely available at https://github.com/fmfi-compbio/mcdp2 under the MIT licence.

This work was partially supported by a grant from the European Union Horizon 2020 research and innovation program No. [872539] (PANGAIA); and grants from the Slovak Research and Development Agency [APVV-22-0144] and the Scientific Grant Agency VEGA [1/0140/25, 1/0538/22].