**Dante - genotyping of complex and expanded short tandem repetitions**

Budiš J.1,2,3, Kucharík M.4, Ďuriš F.2,3, Gazdarica J.2,5, Zrubcová M.5, Ficek A.5, Szemes T.2,5,6, Brejová B.1, Radvanszky J.2,5,7

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

2 Geneton Ltd., Bratislava, Slovakia

3 Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia

4 Medirex a.s., Bratislava, Slovakia

5 Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

6 Comenius University Science Park, Bratislava, Slovakia

7 Biomedical Research Centre, Slovak Academy of Sciences Bratislava, Slovakia

Short tandem repeats (STRs) are stretches of repetitive DNA in which short sequences, typically made of 2-6 nucleotides, are repeated several times. Since STRs have many important biological roles and also belong to the most polymorphic parts of the human genome, they became utilized in several molecular-genetic applications. Precise genotyping of STR alleles, therefore, was of high relevance during the last decades. Despite this, massively parallel sequencing (MPS) still lacks the analysis methods to fully utilize the information value of STRs in genome scale assays.

We propose an alignment-free algorithm for genotyping and characterizing STR alleles based on sequence reads originating from STR loci of interest called Dante. The method accounts for natural deviations from the expected sequence, such as variation in the repeat count, sequencing errors, ambiguous bases, and complex loci containing several different motifs.

In addition, we implemented a correction for copy number defects caused by the polymerase induced stutter effect as well as a prediction of STR expansions that, according to the conventional view, cannot be fully captured by inherently short MPS reads.

We tested Dante on simulated data sets and on data sets obtained by targeted sequencing of protein coding parts of thousands of selected clinically relevant genes. In both these data sets, Dante outperformed HipSTR and GATK genotyping tools. Furthermore, Dante was able to predict allele expansions in all tested clinical cases.

The presented work was supported by the “REVOGENE – Research centre for molecular genetics” project (ITMS 26240220067) supported by the Operational Programme Research and Development funded by the ERDF.