DOCTORAL CANDIDATE: Ksenia Khelik
DEGREE: Philosophiae Doctor
FACULTY: Mathematics and Natural Sciences
DEPARTMENT: Informatics
AREA OF EXPERTISE: Bioinformatics
SUPERVISORS: Torbjørn Rognes, Alexander Johan Nederbragt, Karin Lagesen, Geir Kjetil Sandve

DATE OF DISPUTATION: 20th of August 2019

DISSERTATION TITLE: Improved computational approaches for genome assembly accuracy assessment and error correction

Novel approaches for genome assembly accuracy assessment and improvement

This research has proposed three new tools called NucDiff, NucBreak, and NucMerge for genome assembly accuracy assessment and error correction. NucDiff can be used when assembly errors should be found through comparison of an assembly with an available high quality genome of the same or closely related species. It localizes and annotates all differences between the assembly and the genome. The other two tools can be utilized when paired-end Illumina reads are provided together with a genome assembly. NucBreak localizes assembly errors with the help of reads mapped back to the assembly. NucMerge corrects errors in the assembly by using information provided by an alternative assembly and alignments of remapped reads. According to the benchmarking results provided in the research, the first tool performs a more comprehensive comparison of the genomes while the other two tools enable detection and correction of errors with less number of false positive results compared to the existing alternatives.

Genome assembly accuracy is a crucial factor in many genomic and comparative genomic analyses. Any errors in an assembly may compromise the analysis results and conclusions based upon them. Therefore, it is very important to have an opportunity to locate and, if possible, to correct errors prior to analysis. The proposed tools can be used as one of the possible alternatives to reach this goal.