Accurate detection of clinically relevant Copy Number Variants (CNVs) is essential in the diagnosis of genetic diseases since CNVs are responsible for a large fraction of Mendelian conditions. While the bioinformatics pipelines specializing in the detection of Single-Nucleotide Variants and short indels using Whole Exome Sequencing (WES) data are mature and provide satisfactory performance (https://precision.fda.gov/challenges/consistency), the identification of larger deletions and duplications still remains a challenge. Although a plethora of tools have been developed to call CNVs from WES data, most of these algorithms are characterized by limited resolution, insufficient performance, and unsatisfactory classification metrics. Although fine-tuning of CNV calling parameters may substantially improve the overall algorithm performance, there are still no well-established guidelines that would help optimize the detection rate of CNV-calling pipeline.

Presently computational analysis step, especially CNVs calling, is the major bottleneck in sequencing research projects. There are over 25 tools dedicated for the detection of Copy Number Variants (CNVs) using Whole Exome Sequencing (WES) data based on read depth analysis. The tools reported consist of several steps, including: (i) calculation of read depth for each sequencing target, (ii) normalization, (iii) segmentation and (iv) actual CNV calling. The essential aspect of the entire process is the normalization stage, in which systematic errors and biases are removed and the reference sample set is used to increase the signal-to-noise ratio.

Although some CNV calling tools use dedicated algorithms to obtain the optimal reference sample set, most of the advanced CNV callers do not include this feature. Reference sample set is caalculated based on the correlation between samples calculated in regions divided into groups according to their chromosome membership.

In the project, we want to divide the regions into groups not according to the belonging to the chromosome (as it was done so far) but according to the features of the targets, e.g. median depth of coverage across samples, GC content etc.

To our knowledge it is the first attempt to design an effective algorithm to divide regions into groups. We expect, that the new division into groups (not according to the belonging to the chromosome) will significantly change the group of reference samples for a given sample, which will significantly affect the quality of CNV detection.