New quality control workflow developed for next-generation sequencing
November 11, 2018.

Next-generation sequencing (NGS) is a new method for determining the nucleotide order of genomes or parts of genomes. Currently, NGS quality control relies on setting thresholds for quality parameters. However, the disadvantage of thresholding is that the quality controller may not notice when the quality decreases, if the quality stays within the thresholds.

In her master's thesis, Carin Aschan, who majors in Complex Systems, develops a new workflow for NGS quality control. The workflow utilises statistical process control (SPC), which is a technique that uses statistics to account for the natural variance in NGS and helps determining if the process is in or out of control. In addition, for finding the transitions from in to out of control, Aschan uses a changepoint detection method.

In her thesis, Aschan compares five multivariate SPC methods and two changepoint detection methods. The performance of the methods are compared using simulated data, that is based on three real datasets. One of the quality datasets is based on exome sequencing and another of the datasets is based on gene panel sequencing. Interestingly, a result of plotting the distributions of the quality parameters is that the gene panel quality data is more skewed and the exome quality data is more symmetrical. Although the original quality data comprises 50 quality parameters, ten parameters are selected, in order to save calculation time and storage. The parameters are selected based on correlation values and NGS workflow knowledge.

As a result of the comparison, a multivariate exponentially weighted moving average control chart in combination with a Bayesian online changepoint detection method are chosen for the quality control workflow. The combination finds the transition from in to out of control after the fewest out of control samples. Furthermore, the distance from the estimated changepoint to the correct changepoint is among the shortest. Nevertheless, a disadvantage of the Bayesian changepoint detection is that it needs prior information. The prior needs to not only be optimised separately for each laboratory, but also for each time a laboratory changes its protocol.

Aschan hypothesizes that in the future some sort of statistical process control will be part of the daily routine at any NGS laboratory. She believes that the workflow she has developed is an important first step, but that the developed workflow still needs perfecting. For example are paired-end sequencing related parameters not part of the developed workflow.

More information:
Carin Aschan
carin.aschan@aalto.fi