Evaluation of VARIATHON 2013 - simulated data sets from Human Genome.

Veli Mäkinen

May 13, 2013

Abstract

Variation calling is the process of detecting how an individual differs from the consensus genome of the species. The standard scenario to establish the task is to sequence a sample of randomly positioned DNA fragments from an individual with high expected coverage, align the sequence *reads* to consensus genome, and check for differences called by many read alignments. The details vary depending on which kind of variation one is after. Substitutions and short insertions and deletions (indels) are easiest to detect by voting, but longer indels and structural variant require more sophisticated methods.

For the variation calling challenge we generated three simulated ground truth data sets from Human Genome. Two of the data sets were generated by choosing a random subset of already published common variants to make up the simulated genome. For the third data set we generated artificial long deletions. Reads were generated from the generated artificial diploid genomes and given out for the contestants. We also shared the superset of variations from which the random subset was chosen, so that contestant could use those to improve the accuracy and to normalize the prediction. This was to simulate the realistic scenario of variation calling, as frequent mutations common in human population have been quite well identified and the catalogue is likely to be even more complete in the future. The latter data set was aimed at a really challenging scenario, to identify deletions overlapping in diploid genome alleles. Such overlaps and nearby long deletions can occur locally, but for the sake of challenge, we extrapolated the situation to obtain a large and challenging data set. For this data set we expect that only methods that can exploit the superset of known variations can obtain good accuracy.

Evaluations on predictions were conducted to measure alignment accuracy and variation calling accuracy. Details of the evaluations and results will be given in the talk. The talk also discusses some ongoing efforts on more fair evaluation measures that take into account different kind of invariances and approximately corrects answers in predictions.

This is joint work with Krista Longi and the VARIATHON 2013 team.