Evaluation of VARIATHON 2013 - Bacterial and Yeast simulated data sets

Eric Rivals

Deep genomic sequencing is now the primary way to predict genetic mutations on a genome wide scale. Once the genomic reads have been sequenced, the analysis consists basically in two main steps : first, mapping the reads against a reference genome sequence, second, inferring putative mutations from the differences observed between the reads and the genome. The mutations one wishes to detect are either small scale : substitutions, plus insertions and deletions (collectively termed indels), or structural variants. As many tools have been and are still developed for predicting mutations, it is necessary to generate benchmarks for assessing their precision and sensitivity. This is one goal of the Variathon challenge (http ://bioinf.dimi.uniud.it/variathon). This year edition focuses on small scale variations.

Here, we report the assessment results regarding the synthetic bacterial and yeast data sets. Both data sets were generated by mutating a real genome, randomly extracting reads, and then generating random sequence errors. Each step of the prediction can be evaluated. Hence, we will comment on both the mapping accuracy and the variant calling quality of submissions and of another pipeline (which we performed ourselves for the sake of comparison). We detail the results separately for substitutions, insertions and deletions and try to investigate the relative impact of each step.

This is a joint work with V. Maillol within the collaborative framework of the Variathon.