TITLE

Pangenome read alignment and variant calling

SHORT DESCRIPTION

Pan-genome references, also known as graph reference genomes, can represent multiple strains in a single reference structure. To utilize them fully, we need tools to employ them as we currently use single references. In this project, we aim to tackle two important sequence analysis tasks and transform them to graph-based references: read-alignment and variant calling.

BACKGROUND

While DNA is sequenced at staggering speed, the resulting mountain of data does not necessarily lead to new insights because scientists struggle to visualize, analyze and interpret the new information. This is particularly true in plant genomics, because plant genomes are more complex due to repeats and ploidy.

Recently, pan-genomes have become popular because they provide an intuitive way to encode variation in a graph. This graph representation, encodes the information of many DNA sequences, i.e. multiple reference genomes. Because they provide a broader sampling of known sequences, pan-genomes promise to offer more accurate and comprehensive read alignment and variant calling analyses - two of the most fundamental bioinformatics methods to analyze genomes - compared to a linear reference. This benefit should be particularly pronounced in plant genomes, which are more diverse and complex, compared to human genomes - the focus of current efforts.

The objective of this proposal is to implement and benchmark novel methods use graph references for read alignment and variant calling. Read-alignment and variant calling will be a necessity for the foreseeable future to genotype large collections or to integrated large bodies of existing short-read sequencing data.

OBJECTIVES

This proposal has a novel development component and benchmarking component that includes state-of-the-art methods.

1. Develop algorithm for (short) read alignment on pan-genome graphs.
2. Comparative read-alignment benchmark in the context of polyploid organisms based on simulated data. This benchmark will include the method from Objective (1) as well as relevant state-of-the-art.
3. Develop variant caller for small variants from graph read alignments, including SNPs and indels (Based on Pilon[1])
4. Benchmark variant in the context of polyploid organisms, with attention to heterozygous variants

[1] https://github.com/broadinstitute/pilon/

Student: Msc student in LST, KT, NB, CS or bioinformatics

Contact: t.abeel@tudelft.nl

Partners: Virtual Laboratory for Plant Breeding (http://vlpb.nl/)