TITLE

De novo haplotype reconstruction of polyploid plants: part 2

SHORT DESCRIPTION

Implement, improve and evaluate haplotype aware assembly algorithms on long-read data from (simulated) potato

BACKGROUND

Chromosome polyploidy is largely unique to plant genomes and provides a wealth of additional genetic diversity that can be leveraged in plant breeding. Drought and disease resistance is highly desired traits in plants to ensure agriculture can continue to feed the global population.

To more effectively breed plants it is critical to know where variants are located in the genome. While this is relatively straight-forward for haploid organisms, it is largely impossible to do this on a systematic scale for polyploid plants. The main challenges are to accurately associate heterozygous variants to the correct chromosome copy and establish which variants are sitting on the same copy of a chromosome. This is particularly relevant for crops that resist haploidization such as the common potato, Solanum tuberosum.

De novo haplotype reconstruction, i.e. reconstructing the individual copies of a chromosome that occurs in multiple copies is a computationally complex problem. While methods exist to solve this problem for one or two copies, there are no of the shelf solutions available for ploidies greater than two. Ongoing research in our, show promise that 3rd generation sequencing may make haplotype resolution for higher ploidy plants a reality.

Developments in other projects with long read sequencing technology have enabled generation of large read data sets for S. tuberosum. In this proposal we would like to evaluate the innovative algorithms developed at the Delft Bioinformatics Lab can be used to completely reconstruct separate chromosome copies. Specifically lessons learned from the earlier haplotype assembly small project will be incorporated, implemented and benchmarked.

Related work: This work builds on the Msc thesis work of Lucas van Dijk that built a prototype haplotype assembler. Lucas now works in my group as a PhD student and will be involved in supervising the Msc graduate that will execute this project.

Data

Long read data from *S. tuberosum* (Chr4 of DM is being made available in the context of the GreenHapMap project).

Simulated long-read data from synthetic in-silico polyploid genomes

OBJECTIVES

The objectives in the proposal are two-fold. First we want to implement a number of changes to PHASM[1] (our *de novo* haplotype assembler) to make it work as intended and second we want to expand the scope of the benchmarks and explore alternative approaches.

Implement modifications in the PHASM haplotype reconstruction algorithms based on the lessons learned from the previous project:

1. Fix overlap detection issue through two actions: (i) less graph cleaning and (ii) machine learning to detect repeat induced overlaps.
2. Individual benchmarks for each step in the Overlap-Layout-Haplotype algorithms, with particular attention for the first step.

Expand benchmark and explore alternate approach

1. Head-to-head comparison of PHASM and Falcon-UNZIP on simulated diploid and tetraploid genomes.
2. Explore options to use PHASM in the context of guided haplotype assembly, rather than pure *de novo.* In this case, a consensus assembly is used to guide the initial Overlap-Layout steps.

The benchmarks of both algorithms will consist of:

- Evaluation on simulated error-free
- Evaluation on simulated error-containing
- Evaluation on *S. tuberosum* data set that were previously generated

The benchmark of the algorithm's correctness is primarily done on the simulated data because, the ground-truth of the input genome structures is known. Genome simulations are done with with *aneusim*[2]

References

[1]https://github.com/AbeelLab/phasm

[2] https://github.com/AbeelLab/aneusim

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

Contact: t.abeel@tudelft.nl

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