Title
Phasing Variants in Poplar Trees using a Hybrid of Short & Long Read Technologies

Permalink
https://escholarship.org/uc/item/50q137sm

Authors
Schackwitz, Wendy
Martin, Joel
Lipzen, Anna
et al.

Publication Date
2013-03-26
Phasing Variants in Poplar Trees using a Hybrid of Short & Long Read Technologies

Wendy Schackwitz\textsuperscript{1}, Joel Martin\textsuperscript{1}, Anna Lipzen\textsuperscript{1}, Len Pennacchio\textsuperscript{1}, Gerald Tuskan\textsuperscript{2}

\textsuperscript{1}Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA
Oak Ridge National Lab // ORNL – Oak Ridge, TN

\textsuperscript{a}To whom correspondence may be addressed. E-mail: wsschackwitz@lbl.gov

March 25, 2013

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

\textsuperscript{[LBNL]} This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.
Goal: Phasing SNPs in Poplar

Phasing SNPs assists in correlating genotype & phenotype.

Strategy: A Hybrid Approach

Can Short Reads be Utilized to Call SNPs & Long Reads to Phase Them?

Proposed Work Flow

Confirm Work Flow with Test Data

A 100Kb region was chosen to be manually analyzed so the true phase of heterozygous SNPs could be determined. Once the true phase had been ascertained, the region was used to identify the correct parameters for the scripts.

1) Phase Test Data via Manual Analysis

It is not possible to phase the read of variants directly from long reads because high error rate leads to inconsistent phasing. We manually determined the true phase by summing up all reads to generate a consensus phase.

2) Tune Parameters Using Manual Analysis

We ran GATEK-ReadBackedPhasing on Short Read only and Hybrid test data altering the parameters to determine their effect on phased block size.

Conclusion: By choosing the correct parameters, manual analysis & GATEK-ReadBackedPhasing completely agreed over the 100Kb region.

Results: Phasing of Two Poplar Genomes

Hybrid Approach Results in Dramatically Longer Phased Blocks & Phases a Much Greater Percentage of the Genome

Future Phasing

Do low read base qualities of Long Reads affect ability to phase?

Observation: For phasing programs to effectively use data from long error prone reads the minimum base quality threshold must be lowered to below default levels. This seems to have a minimal impact on the quality of the phasing, possibly because only high quality SNPs were analyzed. However, I would expect that allowing poor quality data into the analysis could result in mistakenly broken phase blocks.

Solution: Update GATEK-ReadBackedPhasing VariantRecords class to apply pipeline filters by read group and qualities.

Can reference bias be prevented when aligning reads?

Observation: The Long Read sequence used has a specific error profile weighted toward indels. Thus when a read has a base that does not match the reference, the preferred way to align the read is to “correct” it via a gap or insertion. This leads to an under-representation of reads containing the alternate allele.

Solution: Is it possible to feed the aligner a reference with ambiguity bases at the known SNP locations, so both versions of the reads are treated equally? BLASR has added a scoring matrix option and this is currently being attempted for the long reads.

Can even longer haplotype blocks be generated?

Observation: GATEK-ReadBackedPhasing only reports serially phased SNPs. If a SNP cannot be phased, it “breaks” the phase block & begins a new one. Broken blocks can stem from false positives, regions of poor depth, or reference bias.

Solution: 1) Remove un-phase-able SNPs by tossing low quality calls. 2) Adjust phasing algorithm to allow “Leap Flopping” around un-phase-able SNPs.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.