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


**Wendy Schackwitz[1], Joel Martin[1], Anna Lipzen[1], Len Pennacchio[1], Gerald Tuskan[2]**

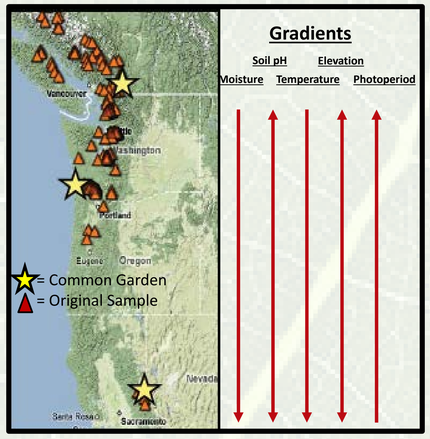
Affiliations:
[1]Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA
[2]Oak Ridge National Lab, P.O. Box 2008, Oak Ridge, TN, 37831, USA


DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

## Goal: Phasing SNPs in Poplar

### Phasing SNPs Assists in Correlating Genotype & Phenotype

Poplar Sample Area

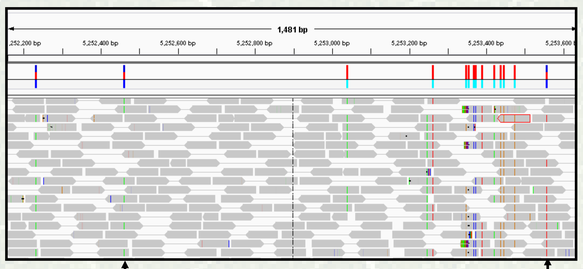



Common Garden - Astoria, OR

Poplar grow throughout the West coast & are adapted to extremely variable conditions. To examine what allows for this wide range of growth conditions, Jerry Tuskan's team has collected 1000 different individuals from British Columbia to California. In 2009, three "Common Gardens" were established where each individual was cloned in triplicate. Nearly all of these trees have been sequenced using short read technology, revealing a huge degree of variation in genotype. Correlating this genomic variation to phenotype would be greatly be strengthened if the variants could be phased into long haplotype blocks.

## Problem: Short Reads Can't Phase Distant SNPs

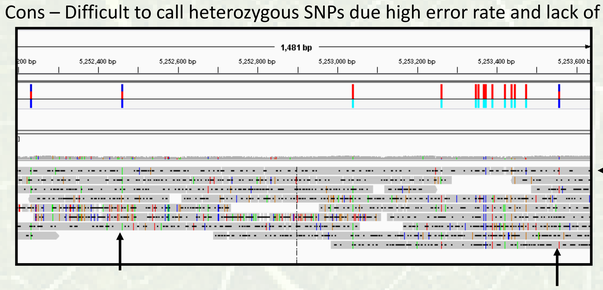
Pros – High throughput, lots of depth, low error rate
Cons – Too short to phase distant SNPs




These two SNPs are too far apart to be phased using short reads

## Problem: Long Reads Can't Call SNPs

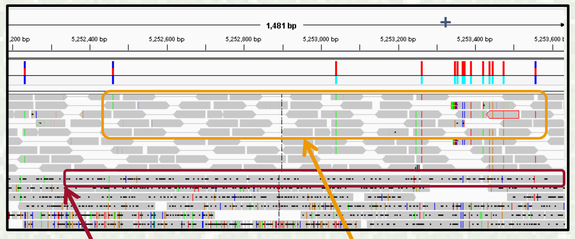
Pros – Can tie distant SNPs
Cons – Difficult to call heterozygous SNPs due high error rate and lack of depth




The top read spans both heterozygous SNPs, allowing them to be phased, but finding the SNPs among the errors is difficult

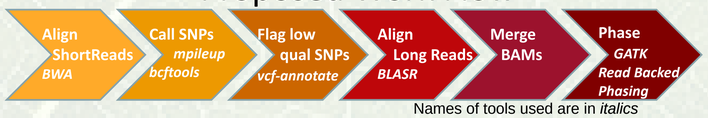
## Strategy: A Hybrid Approach

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




High quality SNPs called from Short Reads
+ Long Read that can tie distant SNPs together
= Large Phased Blocks?

### Proposed Work Flow




Names of tools used are in *italics*

## 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

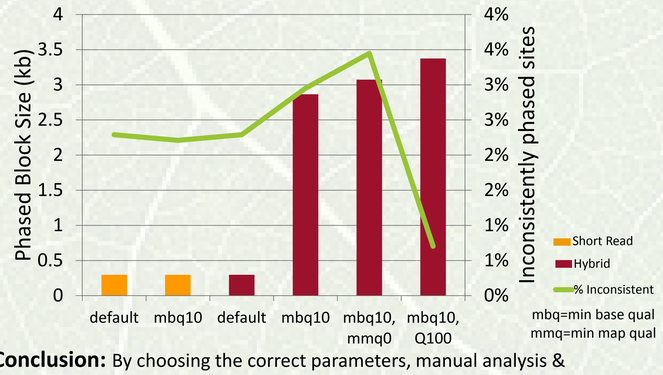
It is not possible read the phase of variants directly from long reads because high error rate leads to inconsistent phasing. We manually determined the true phase by summing up all reads to generate a consensus phase.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ref | A | T | T | G | T | A | C | A | C | G | T | T | G | C | T |
| alt | G | A | A | A | A | G | T | G | T | A | C | G | A | T | A |
| Read1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0/1 | 0/1 | x | 0 | 0/1 |
| Read2 | 1 | 1 | 0 | 0 | 0 | | | | | | | | | | |
| Read3 | | 0 | 0 | 1 | 1 | 1 | 0 | x | | | | | | | |
| Read4 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | |
| Read5 | | 0 | 1 | 1 | 1 | 1 | x/1 | x | | | | | | | |
| Read6 | | | 1 | 1 | 1 | x | 1 | 1 | | | | | | | |
| Read7 | | | 1 | 1 | x | 1 | 1 | 1 | 0 | | | | | | |
| Read8 | | | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | |
| Read9 | | | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | |
| Read10 | | | | | x | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Read11 | | | | | 1 | 1 | 1 | 1 | 0 | x | 0 | 0 | 0 | 1 | 0 |
| Read12 | | | | | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Consensus** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| **Phase** | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

0=ref allele
1=alt allele
x=deletion
/=ambiguous

### 2) Tune Parameters Using Manual Analysis

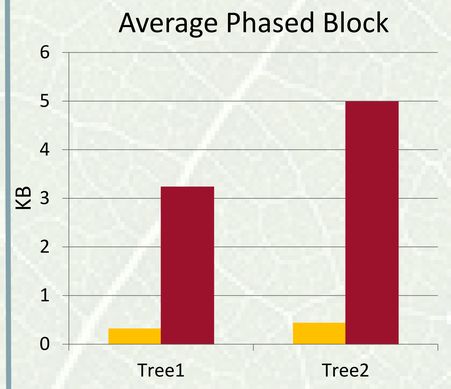
We ran GATK-ReadBackedPhasing on Short Read only and Hybrid test data altering the parameters to determine their effect on phased block size.

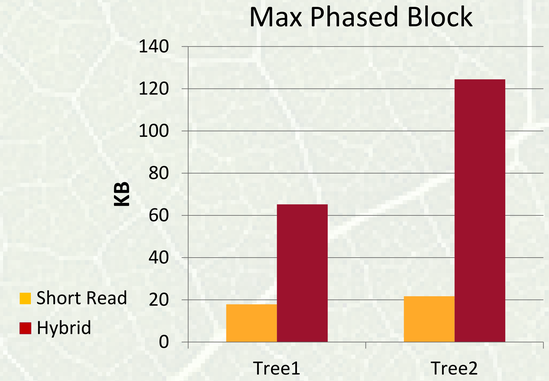


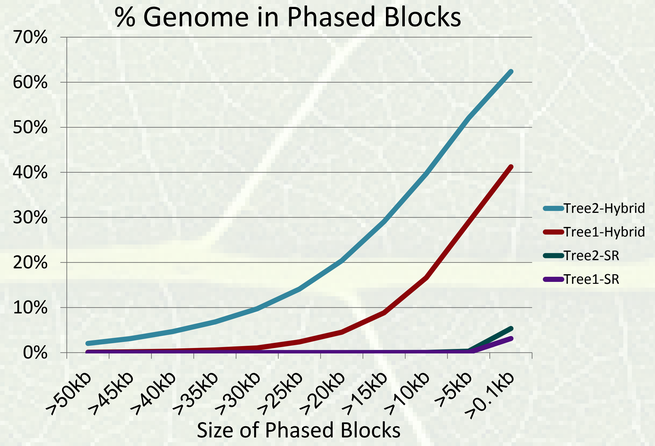

**Conclusion:** By choosing the correct parameters, manual analysis & GATK-ReadBackedPhasing completely agreed over the 100kb region.

## Results: Phasing of Two Poplar Genomes

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



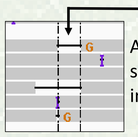

## Future Phasing

### 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 GATK-ReadBackedPhasing VariantReads class to apply pileup 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.




Alternate allele is a "G" here – based on the phase results of other reads, this read should be a G at this location. Because the aligner was not "SNP aware" it chose a del instead of allowing the "G" to align.

**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:** GATK-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 Frogging" 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.