Practical Haplotype Graph (PHG) to call genotypes from skim sequences to aid in genomic selection

Lynn Johnson[1], Dan C. Ilut[1], Zack Miller[1], Terry M. Casstevens[1], Peter J. Bradbury[1,2], Punna Ramu[1], Cinta M. Romay[1], Edward S. Buckler[1,2,*]

[1]Institute of Genomic Diversity, Cornell University, Ithaca, NY, USA. [2]US Department of Agriculture – Agriculture Research Service (USDA-ARS).

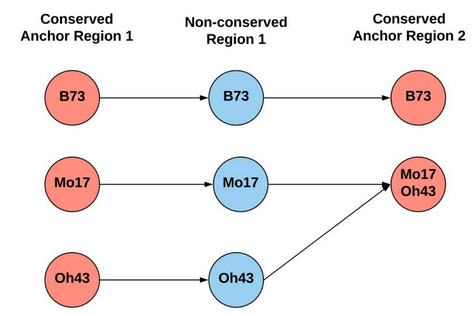
* Correspondence should be addressed to E.S.B. (esb33@cornell.edu)

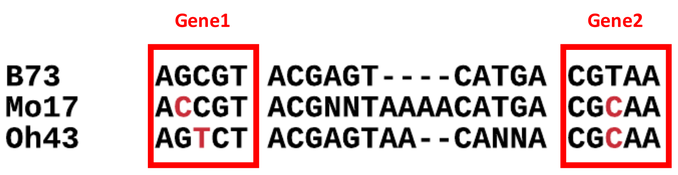
Introduction: Why a graph?

Biology Produces a consistent pattern of a genome

- Conserved genes (and other elements)
- Non-conserved intergenic regions of tremendous variation
- Architecture similar across many species

Maize Example:





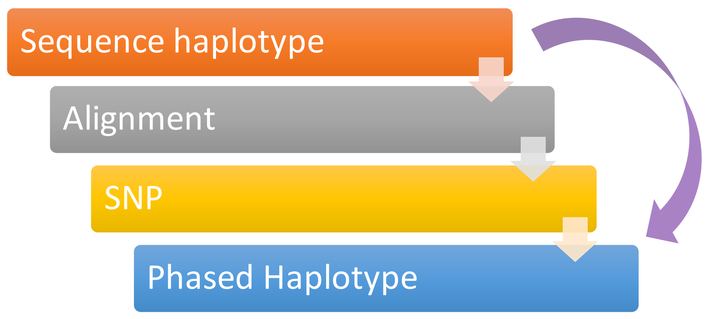
- **Representing genomic diversity and complexity is challenging**
- **Graphs can compactly represent sequence from multiple genomes**
- **Aligning to a pan-genome is better than aligning to a single reference**

Why the PHG?

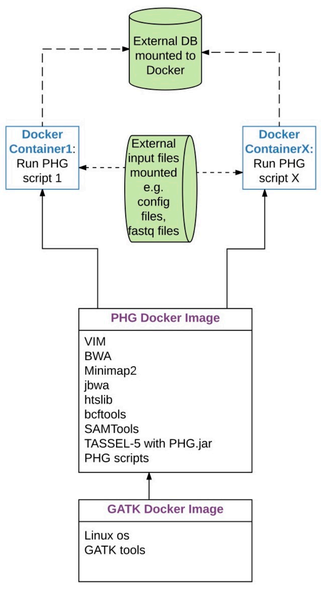
- PHG accepts sequence from multiple technologies (rAmpSeq, Nextera, Nanopore, etc.)
- By pulling taxa into consensus haplotypes, the haplotype graph can be built with low coverage input genomes
- We don't have good assemblies, intergenic regions are horrible
- PHG creates a useful graph even when the data isn't perfect

PHG Goals

- ❖ Identify the haplotype from low depth sequences
- ❖ Create custom genomes for alignment
- ❖ Call rare haplotypes
- ❖ Compress data



PHG docker instance



Load docker Image from .tar

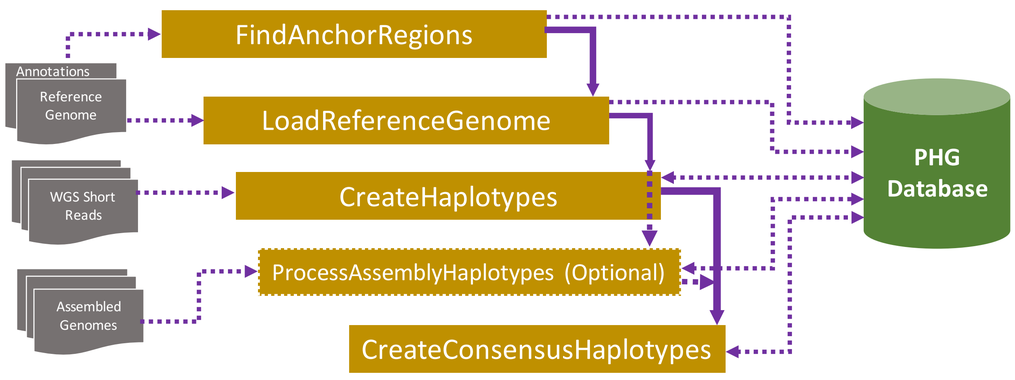
docker load –input phgdocker.tar

PHG scripts in docker root dir

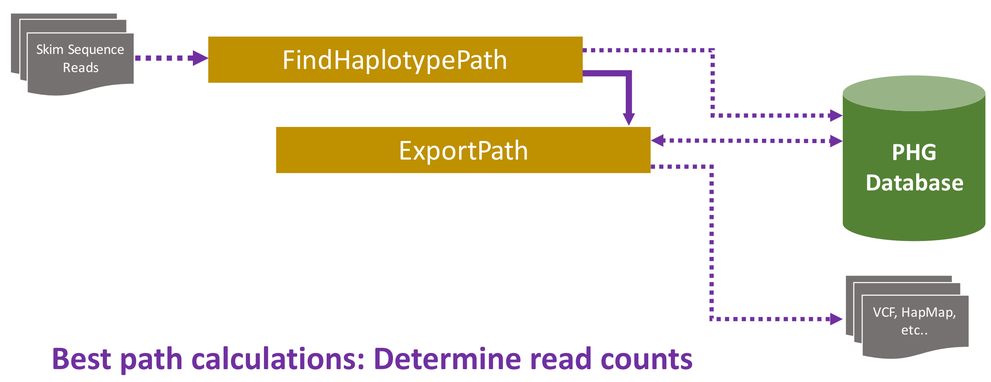
- **LoadGenomeIntervals.sh (1)**
- **LoadAssemblyAnchors.sh (1a)**
- **CreateHaplotypes.sh (2)**
- **CreateConsensi.sh (3)**
- **FindPath.sh (4)**
- **ExportPath.sh (5)**

Calling genotypes from skim sequences using the PHG

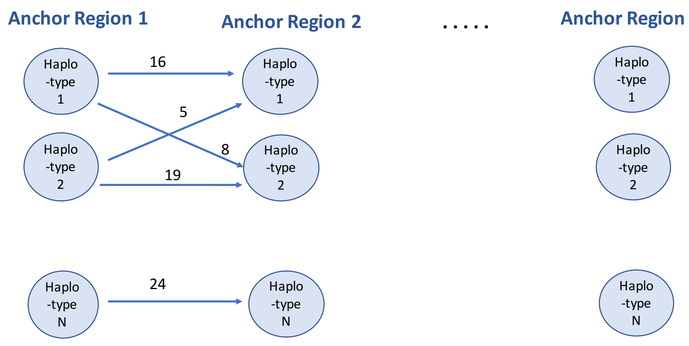
1. Populate the database with haplotypes



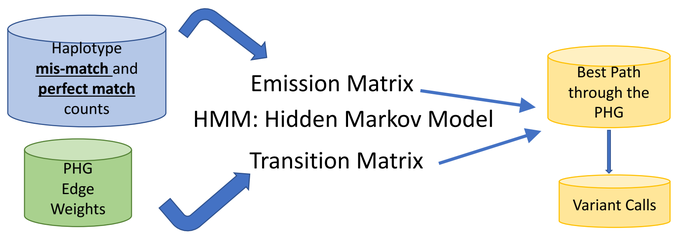
2. Infer genotypes from skim sequences



Best path calculations: Determine read counts and edge weights for each haplotype

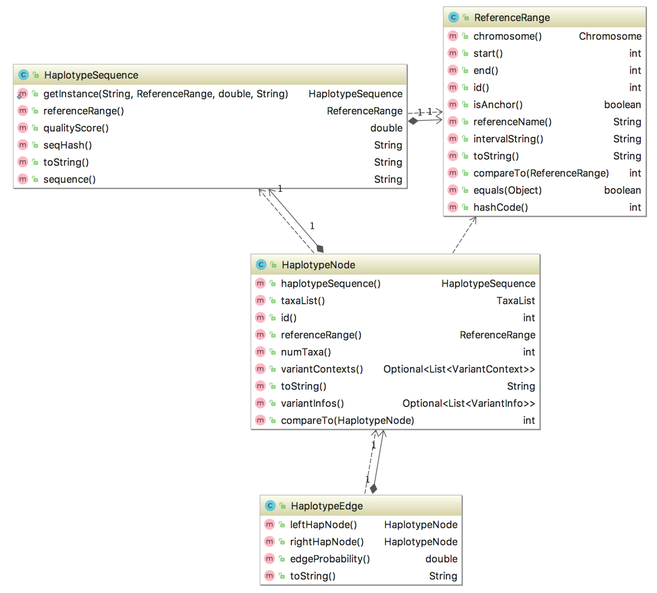


Use HMM to find the best path through PHG

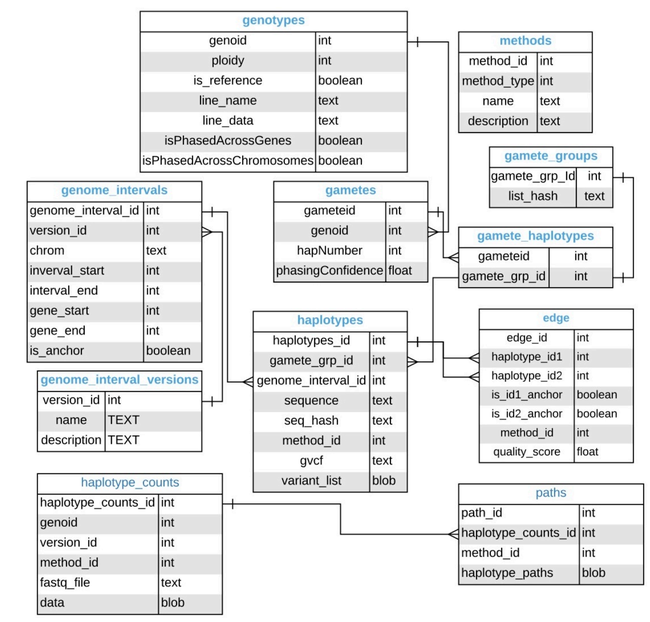


- **Perfect match count**: Number of sequence reads that match the haplotype sequence exactly.
- **Mis-match count**: Number of reads that map to a haplotype but do not match exactly.

PHG API



PHG Database schema



Current deployment:

- Maize -308 taxa
- Sorghum -140 taxa (under development)
- Cassava -348 taxa (under development)

In maize:

- Tested using W22 GBS sequence
- Pathway: 85% of nodes called correctly
- Error rate calling SNPs = 2% (compared to Axiom Array)

Common pipeline to call genotypes from technology independent skim sequences (GBS, rAmpSeq, Nextera, Nanopore, RNA-seq, DArT-seq, etc.)

PHG will be available as a docker image

Running PHG scripts in a docker container

(1) docker run –name load_phg_container \
-v localMachine:/pathToOutputDir:/tmpDir/outputDir \
-v localMachine:/pathToReference:/tmpDir/refDir \
-t phgdocker:latest \
/**LoadGenomeIntervals:sh** config.txt ref.fa anchors.bed data.txt

(2) docker run –name haplotypes_phg_container \
-v localMachine/InputFiles/:/tmpDir/data/ \
-v localMachine/InputFiles/Reference/:/tmpDir/data/reference/ \
-v localMachine/InputFiles/WGSBams/:/tmpDir/data/bam/DedupBAMS/ \
-v localMachine/DockerOutput/phgMaize.db:/tmpDir/output/phgMaizeDB.db \
-t phgdocker:latest \
/**CreateHaplotypes.sh** /tmpDir/data/config.txt taxon single anchors.bed

(3) docker run –name consensus_phg_container \
-v localMachine/InputFiles/Reference/ref.fa:/tmpDir/data/reference/ref.fa \
-v localMachine/InputFiles/config.txt:/tmpDr/data/config.txt \
-v localMachine/DockerOutput/phgMaize.db:/tmpDir/outputDir/phgMaize.db \
-v localMachine/DockerOutput/gvcfsToLoad/:/tmpDir/data/outputs/gvcfs/ \
-v localMachine/DockerOutput/fastasToLoad/:/tmpDir/data/fastas/ \
-t phgdocker:latest \
/**CreateConsensi.sh** /tmpDir/data/config.txt ref.fa ref_version

(4) docker run –name findPath_phg_container \
-v localMachine/InputFiles/Reference/:/tmpDir/data/reference/ \
-v localMachine/InputFiles/GBSFastq/:/tmpDir/data/fastq/ \
-v localMachine/inputFiles/config.txt:/tmpDir/data/config.txt \
-v localMachine/DockerOutput/FindPathDir/:/tmpFileDir/outputdir/ \
-t phgdocker:latest \
/**FindPath.sh** myTaxa config.txt CONSENSUS ref.fa HAP_METHOD ref_version PATH_METHOD

For more details: https://bitbucket.org/bucklerlab/practicalhaplotypegraph/wiki/Home

Acknowledgements

This work was generously supported by the USDA-ARS, the Bill & Melinda Gates Foundation (OPP1159867), and the NSF Plant Genome Research Project (IOS#1238014).