Wheat PHG v2

PHG Creation

The wheat PHG v2 was created in Eduard Akhunov's lab by Katie Jordan using 472 US wheat cultivars. The accessions were exome sequenced to give 2.89M markers. The FASTQ files were converted to gVCF then loaded into the PHG.

The wheat PHG v1 containd 65 UW wheat cultivars. When the Akhunov lab tested imputation accuracy they calculated 95% accuracy when using the same genotype protocol. The imputation accuracy when using a reduced coverage or GBS protocol gave an imputation accuracy of 85-90%. https://doi.org/10.1093/g3journal/jkab390

PHG Imputation

The data in T3/Wheat was imputed using the VCF files for each genotype project. The procedure is described in step 3 of the PHG Wiki User instructions.

The PHG was build using PHG version 0.31. For imputation the PHG was updated to version 0.40 using the liquibase plugin. The configuration settings used minReads=0 paramater which specifies that variants calls were imputed for all positions in the Wheat PHG database. During imputation the VCF files were aligned to the CONSENSUS haplotypes stored in the PHG database. A unique readMethod and pathMethod were used for each VCF file (genotype project). The PHG imputation pipeline build a new index for each VCF file (genotype project). When running the ImputePipelinePlugin we used the "-imputeTarget pathToVCF" so that the output is a VCF of the imputed data.

WheatCAP Goals

1. Convert low density genotype data into a single high-density protocol
2. Convert historic data, including other protocols, into a single high-density protocol

Accuracy testing

For accuracy testing I compared the genotype results between the VCF files for all common accessions. For example if both files had 0/0 or both files had 1/1 then the genotypes match. If one file had 0/0 and another had 1/1 then the genotypes do not match. Missing data is not counted. If all the genotypes match for common accessions then the comparison would give 100% accuracy.

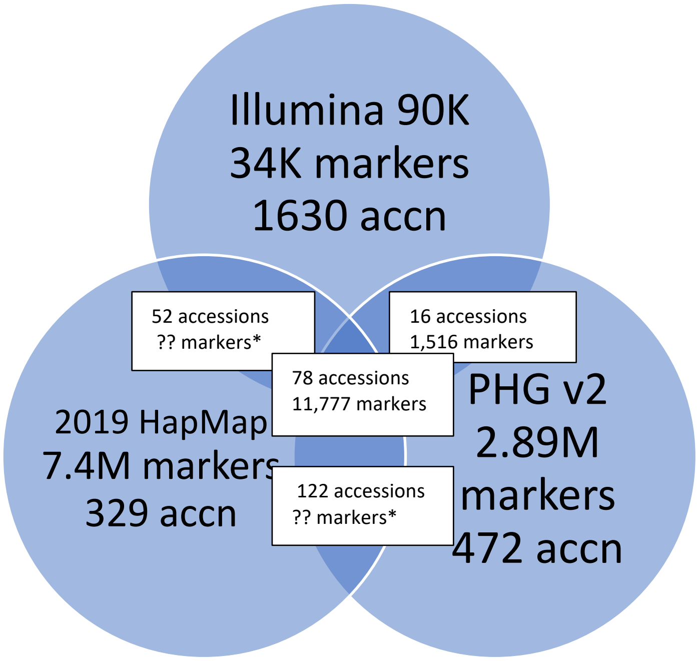
The steps of the imputation pipeline are

1. IndexHaplotypesBySNPPlugin - index the PHG for each SNP specified in the VCF file, creates an index file
2. SNPToReadMappingPlugin - compare a VCF file with the index and create readMappings
3. HaplotypeGraphBuilderPlugin – create path through PHG using method: CONSENSUS
4. BestHaplotypePathPlugin - a set of read mappings to infer the best (most likely) path through the graph given the read mappings. Read mappings are a list of reads with a set of haplotypes to which that read aligned
5. HaplotypeGraphBuilderPlugin – creates path through PHG using method: CONSENSUS and PathMethod

6. PathsToVCFPlugin – exports paths to VCF file

Illumina 90K genotype protocol (34K markers, 1630 accessions)
Exome capture 2019 HapMap genotype protocol (7.4M markers, 329 accessions)
Accession List:

1. 472 accessions in PHG
2. 101 accessions in common between PHG and 90K
3. 85 accessions in common between PHG, 90K, and 2019_HapMap
4. 329 accession in 2019 HapMap
5. 130 accession in 2019 HapMap, and 90K, but not in PHG



* most of the markers match between the 2019 HapMap and PHG. The exact number of matches is unknown because the 2019 HapMap has not be aligned to RefSeq v2.

1. **Accuracy for different protocols (102 accessions in PHG)**

Compared genotype calls for common markers between 90K and 2019 HapMap. The markers indexed is a count of the markers that match by position and reference allele between input VCF and PHG.

Imputation accuracy compared to 2019_HapMap
these don't match venn diagram?

Protocol	comon markers all sets	common markers 2019_HapMap and PHG	chromosome 1A
90K	94.3%	93.8%	86.9%
9K	93.6%	92.3%	85.5%

show graph by marker
show graph by MAF

2. **2.** Accuracy for accessions in PHG vs not in PHG
For this test the 90K protocol was divided into 2 subsets; 23 accessions in PHG and 108 accessions not in PHG.

Protocol	Accuracy between 90K imputed and 2019_HapMap 9,378 markers
90K (accessions in PHG)	92%
90K (accessions not in PHG)	79%

3. **3.** Accuracy for down sampled datasets
This test used the Exome Capture protocol for 13HWW cultivars not in PHG at 8-10x. It was down sampled using a custom script. For a down sampe of 10, keep every tenth sample and remove the 9 others.

Downsample	Markers	Accuracy
10	76,147	94%
30	25,608	94%
100	7,618	94%
300	2,865	93%

This test used the 90K protocol for 79 accessions in the PHG. It was down sampled using a custom script. For a down sample factor of 10, keep every tenth sample and remove the 9 others. For a down sample factor of 100 keep every one hundredth sample and removes the 99 others.

Downsampled	Markers	Accuracy
1	21,814	93%
10	2,486	93%
30	1054	93%
100	553	87%