**Materials and Methods**

**Datasets**

**Illumina 9K and 90K arrays**

The T3/Wheat database contains a large amount of historical data using Illumina 9K and 90K arrays. This historical data was imputed using the PHG.

**Illumina 9K and 90K data processing**

The Illumina array data is stored in the T3/Wheat database as Ref = A\_allele and Alt = B\_allele. In order to use the data with the PHG it was aligned to the reference genome by strand and orientation. See [GitHub/align2Genome](https://github.com/TriticeaeToolbox/PHGv2/tree/main/align2Genome). For each SNP the Ref allele was retrieved from the FASTA reference using samtools faidx. The A\_allele and B\_allele was corrected to match the Ref allele and the genotype data complemented as necessary.

**2019 HapMap**

The 2019 HapMap data uses the same Exome Capture protocol as the data used to create the PHG. This data is used to check the imputation accuracy for accession not used to create the PHG database.

**2019 HapMap data processing**

The dataset was converted from RefSeq\_v1 to RefSeq\_v2 using liftover software: see [GitHub/liftover](https://github.com/TriticeaeToolbox/PHGv2/tree/main/liftover). The dataset was then filtered to remove markers where NCC > 30 (Number of no-called accessions). Remove marker that do not match PHGv2 by location and Ref allele. These markers are either not correctly identified from the SNP calling or misidentified during the liftover to RefSeq\_v2. Remove markers where genotypes do not match with PHGv2 greater than 20%. For additional details see [GitHub/2019 HapMap](https://github.com/TriticeaeToolbox/PHGv2/tree/main/2019_hapmap). The remaining dataset has 470K markers.

**PHGv2**

The PHGv2 data consist of 472 USA WheatCAP germplasm used to create the PHG. The data was aligned to RefSeq\_v2 and SNP calls using GATK pipeline.

**PHGv2 data processing**

The raw dataset has 20M markers. The dataset after filtering by MAF has by 5M markers.

**HQ\_EC**

These are 13 high quality accessions that are not used to create the PHG

|  |  |  |  |
| --- | --- | --- | --- |
| protocol | trials | accessions | markers |
| Illumina 9K array | 4 | 6613 |  |
| Illumina 90K array | 17 | 1630 |  |
| 2019 HapMap | 1 | 329 | 7.5M |
| 2019 HapMap filtered | 1 | 329 | 471K |
| PHGv2 | 1 | 472 | 20M |
| PHGv2 filtered | 1 | 472 | 5M |
| HQ\_EC | 1 | 13 | 7.5M |

**Population Analysis by clustering**

Good imputation accuracy from the PHG requires that the PHG is created with accessions that are similar to the accessions to be imputed. Cluster analysis is used to examine if the accessions in the datasets are closely related. The results are shown in [Github/cluster-snprelate3](https://github.com/TriticeaeToolbox/PHGv2/tree/main/cluster-snprelate3). Clustering HQ\_EC and PHGv2 show that HQ\_EC is a closely related to the PHG so I would expect high accurate imputation. Clustering 2019\_HapMap and PHGv2 shows most of the accessions are closely related but there are about a dozen accession that can be seen clearly in the cmdscale(ibs) plot that are outside the accessions in the PHGv2.

**Imputation Accuracy**

Imputation of HQ\_EC dataset – This dataset uses the same genotyping protocol as the data used to create the PHG and the accessions are similar to those in PHG so it is expected that theses’ accessions will have high imputation accuracy. The HQ\_EC data was down sampled using every 30th sample then imputed. The Imputed output was compared to the original and the imputation accuracy was 94%.

Imputation of 2019\_HapMap dataset – This dataset also uses the same genotyping protocol as the data used to create the PHG but the accessions are more diverse and the genotype data was originally mapped to RefSeq\_v1 instead of RefSeq\_v2. The 2019\_HapMap data was down sampled using every 30th sample then imputed. The Imputed output was compared to the original and the imputation accuracy ranged from 87% to 95%. The accessions with the lowest imputation accuracy were from Aegilops tauschii, Cultivated diploid, and wild types. The results are shown in [GitHub/accuracy\_2019hapmap](https://github.com/TriticeaeToolbox/PHGv2/tree/main/accuracy_2019hapmap).

Imputation of Illumina 9K and 90K arrays – For this accuracy analysis I created two subsets:

inPHG – 90K data whose accessions are in 2019\_HapMap and the PHG

notinPHG – 90K data whose accession are in 2019\_HapMap but not in PHG

I used the 2019\_HapMap to check the imputation accuracy of the imputed data

The results are shown in [GitHub/accuracy\_90K](https://github.com/TriticeaeToolbox/PHGv2/tree/main/accuracy_90K).

A second analysis of the 90K array data looked possible causes for lower imputation accuracy. The unfiltered 90K data had an average imputation accuracy of 92% with a range of 86% to 94%. The process of converting the A\_allele/B\_allele to Ref allele/Alt allele may not give the correct results so I looked at complimenting the genotypes of for each marker to maximize the agreement with the 2019\_HapMap. This slightly improved the accuracy by at most 1%. I looked at removing accessions from the 90K data where the genotypes did not match the 2019\_HapMap data. This also slightly improved the accuracy by at most 1%. For the third test I looked at removing markers from the 90K data where the genotypes did not match the 2019\_HapMap data. This method had the greatest effect on accuracy by increasing the accuracy up 99%. I expect that these markers have problems because they either map to multiple locations or the procedure of mapping them to RefSeq\_v2 did not work correctly. The results are shown in [GitHub/accuracy 90K\_filtered](https://github.com/TriticeaeToolbox/PHGv2/tree/main/accuracy_90K_filtered).