Table of Contents

[Materials and Methods 1](#_Toc145331319)

[Practical Haplotype Graph 1](#_Toc145331320)

[Datasets 1](#_Toc145331321)

[Illumina 9K and 90K 1](#_Toc145331322)

[2019 HapMap 2](#_Toc145331323)

[HQ\_EC 2](#_Toc145331324)

[Analysis 2](#_Toc145331325)

[Population clustering 2](#_Toc145331326)

[Data curation 3](#_Toc145331327)

[Imputation Accuracy 3](#_Toc145331328)

[Imputation Accuracy by marker position 3](#_Toc145331329)

[Imputation Accuracy by Market Class 4](#_Toc145331330)

[Imputation Accuracy by Accession 4](#_Toc145331331)

# Materials and Methods

Practical Haplotype Graph

The Practical Haplotype Graph (PHG) software was developed and is maintained by the Buckler Lab. The wheat PHG (PHGv2) was created by Katherine Jordan. The PHG software to create the database was version 0.0.35. The PHGv2 data consist of 472 USA WheatCAP germplasm. The data was aligned to RefSeq\_v2 and SNP calls using GATK pipeline. The raw dataset has 20M markers. The dataset after filtering by MAF has by 5M markers. A list of accessions and the matching T3/Wheat names are found in [GitHub/imputation/T3matches](https://github.com/TriticeaeToolbox/PHGv2/blob/main/imputation/PHG472_T3Matches.xlsx)

The PHG database was converted to version 0.0.40 and the imputation was done using version 0.0.40. The input for PHG imputation were VCF files from T3/wheat (e.g., from Illumina 9K or 90K genotyping protocols). The imputed output is a VCF file with 2.9M markers. The steps to create the PHG database and impute VCF files is described in more detail in [GitHub/imputation](https://github.com/TriticeaeToolbox/PHGv2/tree/main/imputation).

## Datasets

Illumina 9K and 90K

The T3/Wheat database contains a large number of accessions genotyped using Illumina 9K and 90K arrays. This historical data was used as input for imputation using the PHG. 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 were corrected to match the Ref and Alt alleles. The genotype data is complemented if Ref = B\_allele or Ref = complement(B\_allele).

### 2019 HapMap

The 2019 HapMap data uses the same exome capture protocol as the data used to create the PHG. This data was aligned to RefSeq\_v1 so it was converted to RefSeq\_v2 coordinates before use. This data was used to check the imputation accuracy for accessions not used to create the PHG database.

The dataset was converted from RefSeq\_v1 to RefSeq\_v2 using liftover software: see [GitHub/liftover](https://github.com/TriticeaeToolbox/PHGv2/tree/main/liftover). To correct for errors in the liftover procedure, markers were removed that do not match PHGv2 by location and Ref allele. The dataset was filtered to remove markers where that had more than 10% missing data. The dataset was filtered to remove markers that have more than 20% disagreement with the PHGv2 dataset. For additional details see [GitHub/2019 HapMap](https://github.com/TriticeaeToolbox/PHGv2/tree/main/2019_hapmap). The remaining dataset has 470K markers.

### HQ\_EC

These are 13 HWW cultivars that were not used to create the PHG. The variants were produced using HiSat2 to the RefSeq\_v2 genome and GATK. These accessions should be closely similar to those used to create the PHGv2.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

# Analysis

## Population 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. Each dataset was merged with the PHGv2 VCF data where only markers in common across datasets were retained. 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 accessions that can be seen clearly in the cmdscale(ibs) plot that are outside the accessions in the PHGv2. The clustering of the 90K and PHGv2 shows significant differences in the genotype data. These differences may be due to differences in the genotype protocol and processing the 90K data. Converting the A\_allele/B\_allele to Ref/Alt and converting from RefSeq\_v1 to RefSeq\_v2 may introduce differences in the genotypes.

## Data curation

Please describe here the process, for each genotyping protocol, of comparing allele scores to identify either bad accessions or bad markers.

## Imputation Accuracy

Limitations of PHG imputation - The PHG software imputes genotypes by finding the best match between the input SNP and the haplotypes stored in the PHG. If the input SNP is not contained in the PHG then the imputed output will be different, then the input. The best imputation accuracy requires that the input SNP is located within the reference ranges defined when the PHG is created. For our Wheat PHG the reference ranges are the gene regions. For our imputation we used the parameter “minRead = 0” which tells the program to also impute nearby SNPs.

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 retaining every 30th variant then imputed. The Imputed output compared to the original and the imputation accuracy was 92.1% (minRead = 0).

### Imputation Accuracy by marker position

For the HQ\_EC dataset I looked at how the imputation accuracy depends on marker location. The imputation accuracy using only the markers within the reference ranges was slightly higher 92.2% (minRead = 1). The imputation accuracy using markers outside the reference ranges was lower, 91.2%. The PHG software is designed to give the best imputation accuracy within the reference ranges using a Hidden Markov Model to choose the best path through the graph.

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).

### Cluster analysis by market class

Market class categories wheat cultivars by hardness, color, and growing season. The common classes are: Hard Red Winter (HRW), Hard Red Spring (HRS), Soft Red Winter (SRW), Soft White (SW), Hard White (HW), Durum

1. Does the PHGv2 contain a good representation of each market class? Cluster analysis shows distinct cluster for HRW, HRS, and SRW.

2. Can market class predict the imputation accuracy?

### Cluster analysis by accession

Cluster analysis can be used to determine if the input population (90K, 9K) is similar to the population used to create the PHGv2. If there is overlap between the two populations the I expect that the imputation accuracy to be higher. The datasets for this analysis were created by combining 90K and PHGv2 genotype data. Only markers in common between the two datasets were included. The clustering was done using simple Principle component analysis (PCA) to show PC1 and PC2. The clustering was also done using cmdscale(ibs). Classical multidimensional scaling Identity-By-State.

|  |  |  |  |
| --- | --- | --- | --- |
| **Protocol** | **Input** | **Accessions** | **Imputation accuracy inPHG-notinPHG** |
| HQ\_EC | Down sampled markers | 13 | 95% |
| 2019\_HapMap | Down sampled markers | 329 | 87%-95% |
| 90K |  | 80 in PHG | 95% |
| 90K |  | 54 not in PHG | 88% |
| 9K |  | 64 in PHG | 94% |
| 9K |  | 48 not in PHG | 88% |

Another analysis of the 90K array data looked for possible causes for lower imputation accuracy. The unfiltered 90K data had an average imputation accuracy of 92% with a range of 88% to 95%. The process of converting the A\_allele/B\_allele to Ref allele/Alt allele may not give the correct results so I looked at complementing the genotypes of for each marker to maximize the agreement with the 2019\_HapMap. In this process I counted the matching genotypes for each marker between the 90K and the 2019\_HapMap. I repeated this step for the complement of the 90K data. If the complemented 90K data gave a larger number of matches then I used that data as an input to the imputation process. 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 analysis was done by simply removing accessions from the average accuracy where the imputation accuracy was < 90% for that accession. 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 analysis was done by removing markers from the average accuracy where the imputation accuracy was < 90% for that marker. This method had the greatest effect on accuracy by increasing the accuracy up to 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).

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Markers** | **Accessions** | **Imputation accuracy inPHG-notinPHG** |
| Uncorrected | 14484 | 125 | 86%-94% |
| Aligned to 2019\_hapmap | 14484 | 125 | 87%-95% |
| Remove bad accessions | 14484 | 104 | 87%-96% |
| Remove bad markers | 9839 | 125 | 95%-98% |