Stephanie Sjoberg
CROP_SCI 545 HW2

Objective: To exam the impact of 1) missing rate; 2) sample size; and 3) method on imputation accuracy evaluated as correlation coefficient, match proportion across genotypes, and match proportions on major and minor allele homozygous separately.

Question 1: Choose a dataset (please specify the dataset number) from the recommended list (http://zzlab.net/StaGen/2018/Data/PublicData.pdf), or a dataset outside the list (please specify source, 5 extra points), or your own data that can be released to public (please specify source, 5 extra points). You can sample partial of individuals or markers; however, the final data set must contain over 100 individuals and 5,000 markers with known chromosome and base pair positions. Display marker locations on chromosomes, distribution of missing rate (both maker wise and individual wise), and minor allele frequency (20 points).

Statement: The dataset used in this assignment is marker data from my own research containing 15,937 SNP markers on 469 winter wheat breeding lines. Genotyping was performed by the USDA genotyping lab in Fargo, ND using the Wheat 90K Illumina SNP chip. This is a subset of the 90K SNPs that are most informative for these breeding lines. Although this data has been published on by other graduate students, it should not be released to the public.

Methods: To display marker locations on chromosomes and the missing data rate, "rQTL" package was used. The input data file was set up in the format of A, B, and H and included the marker names and locations. The minor allele frequency was calculated using a looping function. This required a genotype matrix input of numeric values 0, 1, and 2.

Results and Interpretation: The map of 15,937 SNP markers, spanning across the 21 chromosomes (1-7) of wheat's 3 genomes (A, B, and D) is shown in Figure 1. A total of 2,699 markers do not fall on any chromosome, according to the most recent wheat genome assembly. These markers can still be important when performing statistical analyses such as a GWAS, so they have been designated as part of the "UN" chromosome. Missing data for both individuals and markers are shown in Figure 2. Across the entire dataset, only 97.8% of data is missing. Finally, the distribution of minor allele frequency across all markers is shown in Figure 3. The minor allele frequency ranges from 0.01 to 0.50, with a mean of 0.27.

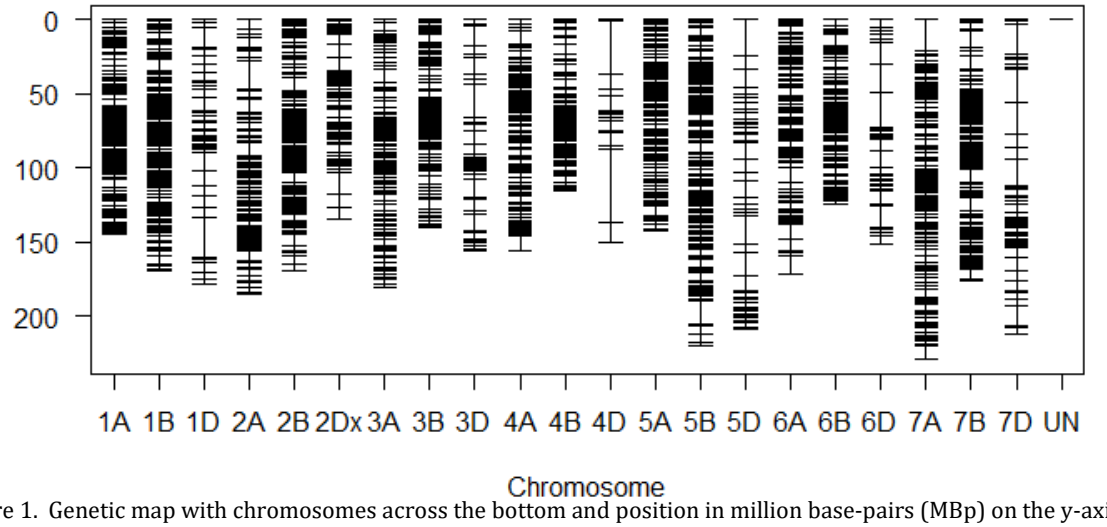

Figure 1. Genetic map with chromosomes across the bottom and position in million base-pairs (MBp) on the y-axis.

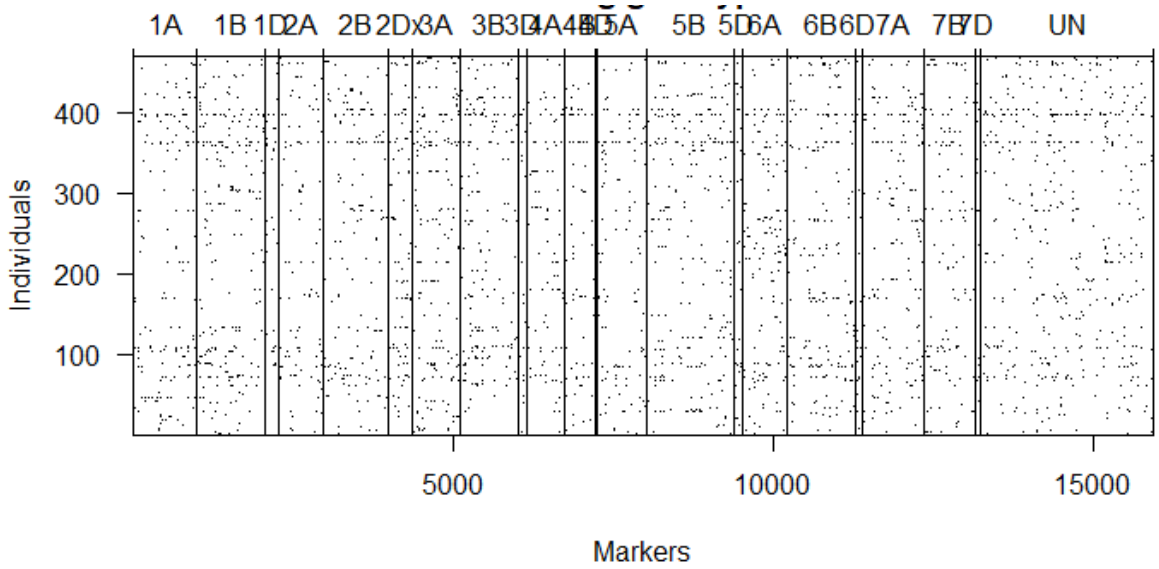

Figure 2. Missing data shown by black points for individuals across the y-axis and markers along the x-axis.

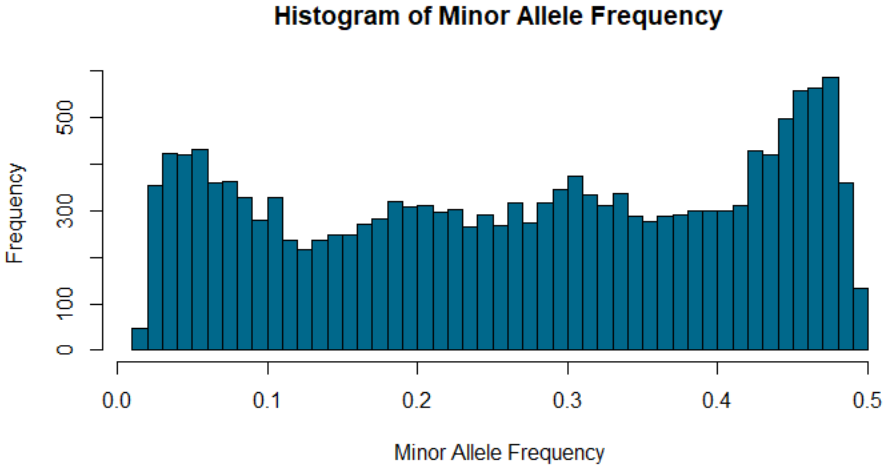

Figure 3. Histogram of minor allele frequency for all markers.

Question 2: Randomly select 5%, 25% and 50% of data points and set them as missing values. Impute these missing values with the stochastic imputation method. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 30 times, report average, standard deviation and number of replicates. Describe the relationship between the missing rate and imputation accuracy.

Hypothesis: Stochastic imputation adds a random component to predicting missing values. It imputes values for each case by drawing at random from the distribution of the missing values given the observed values. This method only uses the observed data, so regardless of the amount of missing data the accuracy should remain the similar.

Methods: The stochastic imputation formula function was used to impute data for each dataset with percent missing data set to 5%, 25%, an 50%.

Results and Interpretation: The accuracy of each imputation was repeated 30 times for each data set and the average and standard deviation are shown in Table 1. The variation in accuracy between the missing rates is small, 0.005. This supports the hypothesis that stochastic imputation

accuracy would remain the same across the missing rates because it is using a random value to fill the missing data. As the missing rate increases, the standard deviation decreases.

Stochastic Imputation			
Set missing	# Reps	Accuracy avg.	Accuracy SD
5%	30	0.1900	0.0017
25%	30	0.1896	0.0006
50%	30	0.1895	0.0005

Table 1. Accuracy results from imputation using the Stochastic method for three sets of data.

Question 3: Redo (2) by replacing stochastic method with KNN method. Describe the differences of results from (2).

Hypothesis: The nearest KNN imputation method uses a set number of nearest neighbors with observed data to fill missing data. If there is more missing data, we expect that the imputation accuracy will decrease because there are fewer neighbors for the method to find and use to fill the missing data.

Methods: The impute.knn function contained the 'impute' package in R was used to impute data for each dataset with percent missing data set to 5%, 25%, an 50%.

Results and Interpretation: The accuracy of each imputation was repeated 30 times for each data set and the average and standard deviation are shown in Table 2. The variation in accuracy between the 5% and 25% missing rate is small, 0.070, although when the missing rate is 50% the accuracy decreases by nearly 50%, by 0.406. This supports the hypothesis that KNN imputation accuracy would decrease as the missing data rate increases because there would be fewer neighbors to reference. As the missing rate increases, the standard deviation decreases.

KNN Imputation			
Set missing	# Reps	Accuracy avg.	Accuracy SD
5%	30	0.8916	0.0010
25%	30	0.8220	0.0001
50%	30	0.4164	0.0001

Table 2. Accuracy results from imputation using the KNN method for three sets of data.

Question 4: The neighbors in KNN refer to individuals and attributes refer to genetic markers for imputation of missing genotypes. Redo (3) by switching neighbors to genetic markers and attribute to individuals. Describe the differences.

Hypothesis: If the KNN imputation uses markers instead of individuals, accuracy should not be affected.

Methods: The same methods for KNN imputation from Question 3 was used for this question, with one exception. The dataset needed to be transposed so that the genotypes were oriented in rows and the markers in columns. The same missing data rates were used for this analysis, 5%, 25%, and 50%.

Results and Interpretation: The accuracy of each imputation was repeated 30 times for each data set and the average and standard deviation are shown in Table 3. The variation in accuracy between the 5% and 25% missing rate is small, 0.006, although when the missing rate is 50% the accuracy decreases more, by 0.195. The difference between KNN using individuals as neighbors and markers as neighbors between datasets with 5%, 25%, and 50% is 0.12, 0.06, and 0.15, respectively. Interestingly, the accuracy for datasets with missing rates of 5% and 25% is higher when KNN uses individuals as neighbors but for the dataset with a missing rate of 50% the accuracy is higher when KNN uses markers as neighbors. This could be explained by the fact that even when 50% data is missing there are still almost 8,000 markers left to use as comparisons for filling the missing data. As the missing rate increases, the standard deviation decreases.

KNN Imputation with marker as neighbor			
Set missing	# Reps	Accuracy avg.	Accuracy SD
5%	30	0.7716	0.0003
25%	30	0.7651	0.0000
50%	30	0.5699	0.0000

Table 3. Accuracy results from imputation using the KNN method with marker as neighbor for three sets of data.

Question 5: Fix the missing rate at 25% and perform imputation with BEAGLE. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 10 times, report average, standard deviation and number of replicates. Describe the advantage over KNN.

Hypothesis: The BEAGLE imputation method should be the most accurate, compared to KNN and Stochastic imputation.

Methods: The BEAGLE package in R, run via Java, was used to impute data for the dataset with percent missing data set to 25%.

Results and Interpretation: The accuracy of each imputation was repeated 10 times for each data set and the average and standard deviation are shown in Table 4. Contrary to the hypothesis, the BEAGLE imputation accuracy is less than the KNN imputation method accuracy but is still much higher than the Stochastic imputation method. The expectation that BEAGLE would impute more accurately was based on the method of using a linear interpolation algorithm. The finding that KNN provides a more accurate method of imputation for this data suggests that the individuals in this dataset are very similar, so neighboring information is the most helpful in imputing missing data.

BEAGLE Imputation			
Set missing	# Reps	Accuracy avg.	Accuracy SD
25%	10	0.7573	0.0017

Table 4. Accuracy results from imputation using the BEAGLE method on one set of data.

Question 7: Find another method and demonstrate that it has better imputation accuracy than both KNN and BEAGLE (20 points, report is limited to one extra page).

Hypothesis: A random forest algorithm can be used for non-parametric imputation will increase the accuracy of imputation over BEAGLE and KNN.

Methods: The random forest algorithm is a non-parametric method that does not make explicit assumption about an arbitrary function. Instead, it estimates the function so that it can be as close to the data points as is realistic. For each variable, a random forest model is built and then the model is used to predict the missing values in the variable based on the observed values. The 'missForest' package can be used to impute missing data from the simulated 25% missing data set that was used for Question 5, BEAGLE imputation.

Results and Interpretation: The accuracy of each imputation was repeated 30 times for each data set and the average and standard deviation are shown in Table 5. The accuracy exceeds both KNN and BEAGLE when there is 25% missing data. This indicates that our data fits a non-parametric model better than a normal linear model.

missForest Imputation		
Set missing	**# Reps**	**Accuracy avg.**
25%	10	0.8975

Table 5. Accuracy results from imputation using the missForest method on one set of data.