The Effect of Chromosomes on the Glucose Level of Potatoes

Munkhdul Norovsambuu

Supervisor: Dr. Hong Gu

April 25th, 2023

Abstract:

The goal of this project is to determine which chromosomes and regions of the chromosome to alter to control the glucose level of potatoes so that different potatoes with different levels of glucose could be grown in the future. Random forest and gradient boosted tree method was used to fit the SNPs to the glucose level. Then the R^2 values were calculated and from each method, lists of the SNPs which had the largest effect on the glucose levels were obtained. The SNPs and the genetic position of these SNPs were graphed in a scatter plot and grouped according to chromosome. Then the two lists were compared to the SNPs found be the LOD scores using a Venn diagram. The R^2 value for the random forest was 0.2308673 and for the gradient boosted tree was 0.2580199. Chromosome 1 had the largest effect on the glucose with chromosomes 4,5 and 9 having the second largest effect on the glucose. For most of the chromosomes, regions within 25-100cM had the largest effect on glucose; however, the most influential regions of chromosomes 1 and 7 are within 75-150cM and 0-25cM. The Venn diagram is in figure 2.

Introduction

Glucose is a monosaccharide – a type of carbohydrate -- which the body cannot break down any further. It is one of body's primary fuel sources which is obtained from consuming food (Wright 2022). When a person eats food, their body processes glucose and other carbohydrates into simple sugars that move into their bloodstream. Then the hormone insulin is released into the blood to aid in transporting the sugars into the cells to be used for energy (Mandle, 2020). While glucose is important for the body, people with diabetes have issues with large amounts of glucose due to their bodies being unable to produce insulin correctly which stops glucose from entering their cells causing a buildup in ketones (Wright, 2022). Therefore, foods with differing levels of glucose are needed. This paper focuses on potatoes due to potatoes being the third most important food crop in the world (Consumption, n.d.).

Specifically, this paper identifies the chromosomes and their regions that have the greatest effect on the glucose level of potatoes. Single Nucleotide Polymorphisms (SNP) were used to accomplish this. SNPs are one-nucleotide DNA sequence signatures that can be located in the genome sequence of potato and can therefore be used as DNA markers. This analysis uses two different methods to fit the SNPs to the glucose level, and each method identifies a list of SNPs which are most useful in predicting glucose levels. The two lists are then compared to each other, and the common SNPs are identified and used to determine the chromosomes that have the largest impact on the glucose levels. The two lists are also compared to the SNPs found by using the LOD scores.

Data

The data utilized in this study came from three files taken from a study done by Dr. Helen Tai at Agriculture and Agri-Food Canada. The first file contained the field traits for 300 marked potatoes. The field trait studied for this project, the potatoes' glucose levels, was measured using a strip array. The second data file contained the genotyping calls for 1603 SNPs for most of the marked potatoes found in the first data file. The glucose levels for the marked potatoes were matched to the genotyping calls for the 1603 SNPs and observations without SNP values were removed, which resulted in a data frame containing the glucose level and the SNP values for 264 potatoes. The second data file also contained the SNPs mapped to a chromosome and the genetic distance of the SNP on each of the chromosomes in centimorgan (cM). The third data file contained the LOD scores. LOD scores are test statistics that test the independence of each SNP with the gene expression or trait using a linear mixed model approach with single SNP analysis. LOD scores greater than or equal to 3 are considered statistically significant meaning that the SNP is highly related to the glucose level.

Method

Random forest

Random forest (RF) is a commonly used machine learning algorithm which was created by Leo Breiman (Breiman, 2001). Random forest combines the output of multiple decision trees to reach a single result and can be used for both classification and regression purposes. Random forests have two hyperparameters: the number of trees used and mtry, which is the number of variables chosen randomly at each split (James, 2021, pp.343-344). This algorithm was used to predict the glucose level using the SNPs and to determine which SNPs had the largest effect on the glucose level.

Cross validation

Cross validation is a method used to estimate the test prediction error of a model by training a model on a training set and then calculating the prediction error based on a separate validation set. Cross validation can be used to determine the optimal values for hyper-parameters in a model by choosing values which minimizes the cross-validated error. In this analysis, 10-fold cross validation was used. This is done by splitting the data into 10 equal sized data sets and using 9 of these datasets to train the model while the last dataset is used to validate the model. The cross-validation process is performed 10 times, each time a different dataset is used as the validation set to get 10 prediction errors. Then the 10 prediction errors are averaged to give the cross-validated error Hastie et al., (2009).

Gradient boosted tree

Gradient boosted tree is a boosting method. Boosting methods work by combining weak "learners" into a single strong learner in an iterative manner. Gradient boosting will sequentially grow trees where each individual tree is considered a weak learner. Then each subsequent tree is fit on the negative gradient of the loss function and evaluated over the training data by least squares estimation. Under squared error loss this will result in each tree being fit on the residuals from the preceding tree in the sequence. There are three hyperparameters in the gradient boosted tree model: the shrinkage, and the interaction depth the number of trees. The shrinkage determines the contribution of each tree to the outcome, with smaller values resulting in more accuracy. The interaction depth specifies the maximum depth of each tree (i.e., highest level of variable interactions allowed while training the model). The number of trees is the number of trees grown.

R-Squared

R-squared (R^2) is a method to quantify the extent which a model fits the data. R^2 is the proportion of variance in the response variable explained by the explanatory variables. R^2 is calculated using the equation

$$1 - \frac{RSS}{TSS}$$

RSS (residual sum of squares) =$\sum (y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is the i'th predicted value

TSS (the total sum of squares) =$\sum \left(y_i - y^{-}\right)^2$ where ȳ is the mean of the dependent variable (James, 2013, pp. 69-70).

Analysis

This section will explain how the methods are used to find the SNPs which had the largest effect on the glucose level of potatoes as well as show which chromosomes and regions of the chromosomes are most impactful on the glucose level.

Both random forest and gradient boosted tree are used to assess which SNPs had the most influence on the glucose levels. Random forest is used to fit each of the SNPs to the glucose level. First the data was split into ten equally sized data sets/folds and ten random forests were used to fit the SNPs to the glucose using each fold as a validation set once. When each random forest was being run, a grid search with ten-fold cross validation was used to check the optimal mtry value from the possible values: 345 ,445, 545, 645, 745, 845, 945 and 1045. The incNodePurity was used to rank the SNPs and the top 50 SNPs for each random forest were identified. Then top 50 SNPs for each random forest were combined. A data set containing the unique SNPs and the number of times the unique SNP appeared in the combined list was created.

The incNodePurity for each SNP in the data set was found by summing the incNodePurity of the SNP from each random forest. Then the incNodePurity was used to order the SNPs in the data set from largest to smallest.

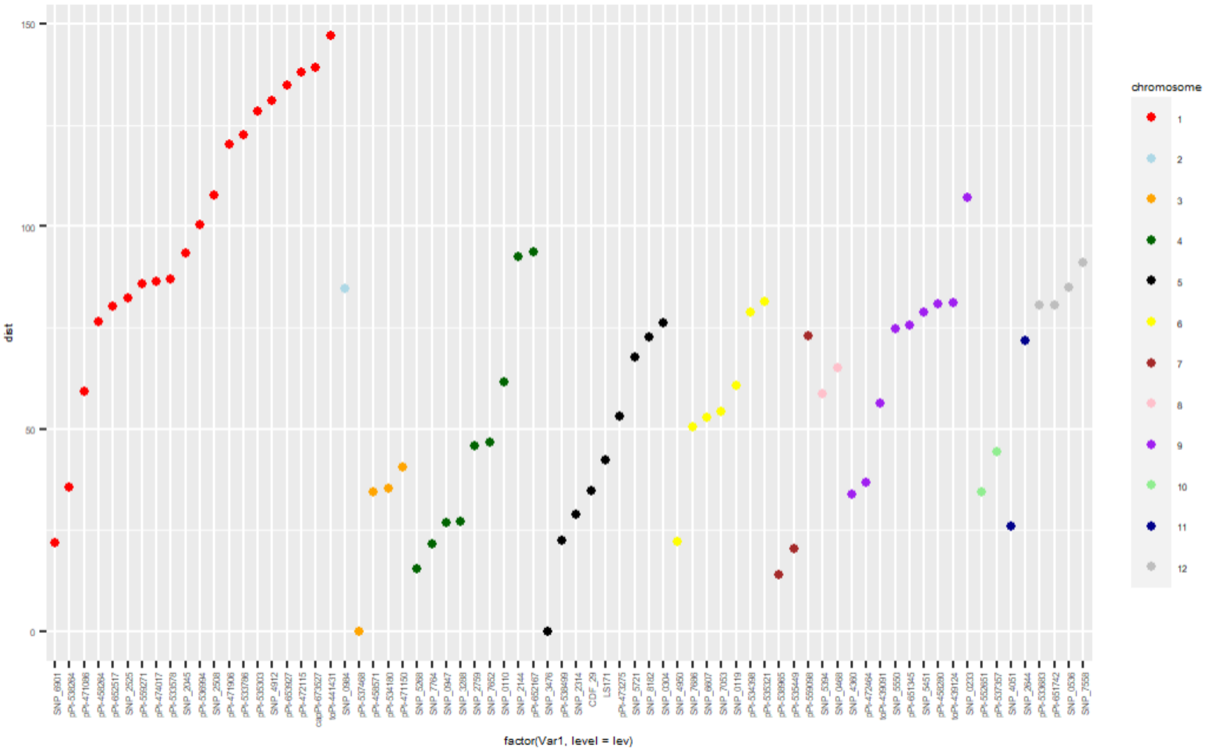
A very similar method was used to get the most important SNPs from the gradient boosted tree. The data was split into ten folds and ten gradient boosted trees were used to fit the SNPs to the glucose using each fold as a validation set once. Ten-fold cross-validation was used to obtain the optimal values for the hyperparameters. Typical values for the shrinkage parameter are small (< 0:1) so the values of shrinkage that were tested were 0.001, 0.01 and 1. Hastie et al. (2009) suggest that if the interaction depth is greater than three then a value of six is often sufficient, so the range of values tested for interaction depth is 1, 2 and 3. The number of trees grown is often a large number so the values checked for n are 200, 400, 600, 800 and 1000. The relative importance measure of the boosted trees was used to rank the SNPs and the top 50 SNPs of each gradient boosted tree was found. Then the same method used for acquiring the data set of the top SNPs for the random forest was used for gradient boosted trees except the relative importance measure was used to rank the SNPs.

For each random forest grown, the validation fold is used to predict the glucose values. All the predicted values for each validation fold were then concatenated together, and this was the $y^{\frown}$ set. All the glucose levels in the validation sets are also concatenated together to get the dependent variable(y) data. These two sets were used to calculate the R^2 value for the random forest. The same method was used to get the R^2 value for the gradient boosted tree except that instead of using each random forest to predict the glucose levels each gradient boosted tree grown was used to predict the glucose levels.

The two data sets containing the most influential SNPs from the random forest and gradient boosted tree were used to create a scatter plot and a Venn diagram. For the scatter plot the SNPs from the two data sets were extracted and combined. Then the SNPs which appeared only once were removed and the chromosomes which each remaining SNP belonged to was added to the data frame and the distance of the SNPs was added to the data frame. This was then graphed using a scatterplot with the distances on the Y axis, the SNPs on the X axis and grouped by chromosomes.

Figure 1

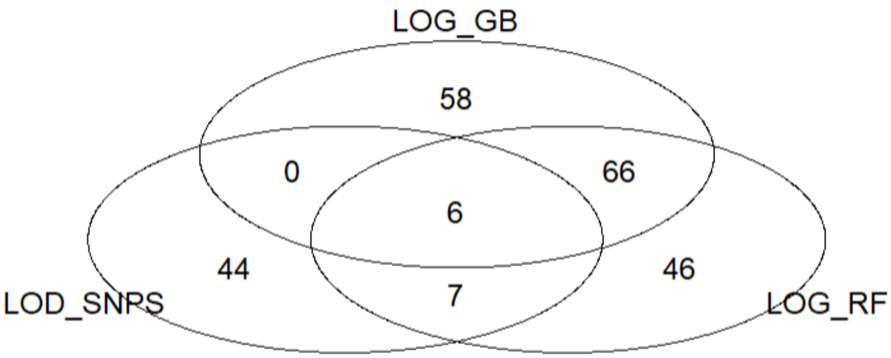
Plot which shows SNPs on x axis, distance on y axis and grouped by chromosome



For the Venn diagram the SNPs which had an LOD score of 3 or greater for glucose, the most important SNPs from the random forest and the important SNPs from the gradient boosted tree were used. LOD_SNPS stands for the SNPs from the LOD table, LOG_GB stands for the SNPs from the gradient boosted tree and LOG_RF stands for the SNPs from the random forest.

Figure 2

Venn diagram of SNPs found from LOD table, gradient boosted tree and random forest



Conclusion

Chromosome 1 has the largest number of SNPs in graph 1 implying that this chromosome has the largest effect on the glucose level. Chromosomes 4, 5 and 9 all have the same the number of SNPs in the plot implying that they all have similar level of effect on the glucose level and since all of them have the second largest number of SNPs after chromosome 1 they would most likely have the second largest effect on the glucose. Therefore these 4 chromosomes would be the chromosomes to alter to control the glucose level of the potatoes.

Most of the chromosomes have most of their SNPs genetic distance within the range of 25-100 cM implying that the regions within this range tend to have the most influence on the glucose level. The exceptions to this are chromosome 1 and chromosome 7. The majority of chromosome 1s SNPs are within the range of 75-150 cM implying that the regions of chromosome 1 which has the largest effect on the glucose are within this range and for chromosome 7 the regions of the chromosome within 0-25 cM have the most impact on glucose levels.

The R squared values (R^2) for the gradient boosted tree and the random forest was calculated to determine how accurate these two methods were. The R squared values were 0.25802 for the gradient boosted tree and 0.23087 for the random forest. Since this study was done using only the SNP values, it is not surprising that the R^2 values are not very large as there are multiple other factors which effect the traits of potatoes other than the SNPs. The R^2 value did tell us that the gradient boosted tree performs better than the random forest due to having a larger R^2 value. This is not surprising since gradient boosted trees generally performs better than random forest though the fact that the difference between the two methods R^2 value is so small implies that the methods are interchangeable.

The Venn diagram is used to compare the SNPs found by the random forest and gradient boosted tree method to the SNPs found to be significant for the glucose from the LOD table. While the gradient boosted tree and the random forest had more than half of the SNPs in common the SNPs from the LOD table barely had any SNPs in common with only 6 SNPs in common with the gradient boosted tree and 13 SNPs in common with the random forest. Since LOD scores are based on single SNP instead of joint on all SNPs, they are less reliable due to the high dimensionality and correlations between SNPs. The Venn diagram seems to represent this as the SNPs found significant using the LOD table has little in common with the SNPs found using random forest and gradient boosted trees.

For further research these methods could be used on the other traits from the first data file to determine which chromosomes had the most effect on the other traits. Other ways to further this research would be to fit different models with larger R^2 values to get more accurate results or to take other factors such as soil or weather into account to determine which factors have the most influence on the glucose.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Consumption. APRE. (n.d.). Retrieved April 25, 2023, from https://apre.org/potato-nutrition/consumption/

Hastie, T., Tibshirani, R., and Freidman, J. (2009). *Elements of Statistical Learning.* Springer, second edition.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). In *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.

Mandl, E. (2020, February 13). *Potatoes and diabetes: Safety, risks, and alternatives*. Healthline. Retrieved April 25, 2023, from https://www.healthline.com/nutrition/potatoes-and-diabetes#effect-on-blood-sugar

Wright, S. A. (2022, August 18). *What is glucose and what does it do?* Healthline. Retrieved April 25, 2023, from https://www.healthline.com/health/glucose#What-is-glucose?