A Genome Wide Association Study to Identify SNPs Associated with Human Height and Training a Support Vector Machine to Predict Height with P-value Prioritized SNPs

Demarcus Briers[1*], Samuel Moijueh[1*]

Abstract: Human height is a complex polygenic trait influenced by potentially thousands of interacting loci. Additionally height has shown to be about 80% genetic while environmental factors such as nutrition account for approximately 20% of human height. As a result of these genetic and environment influences, height is a difficult trait to predict. Previous studies to locate height associated SNPs and predict height used far fewer SNPs accounting for human height and were limited to case-control classifications of height in ethnically homogeneous population. We conducted a height association study on 313 individuals. We further developed a genetic height Support Vector Machine using p-value prioritized SNPs that achieved a maximal AUC of 0.86 in case-control classification and a predictive accuracy of 72% in multi-class classification. The genome study was a relatively low statistical power for identifying SNPs associated to height but this was project was mainly a proof of concept since applying a GWAS to an SVM hasn't be performed before.

INTRODUCTION

Human height is a complex polygenic trait that has been difficult to predict due to the small interacting effects exerted by thousands of genetic loci **[Wood et al 2014]**. In addition to this conundrum, past research has shown that approximately 80-90% of a person's growth before skeletal maturity is genetic and the other 10-20% is attributed to environmental effects such as nutrition **[Aulchenko et al]**. As a result, there are currently no genetically rigorous models for predicting a person’s height outside of extreme binary classifications in ethnically homogenous populations. As far as physiological estimates of height, one common metric used by physicians is to use a child’s height at age two as a predictive estimate. This a decent estimate for final height because by age two children have reached a percentile on the CDC height growth chart that they will stay in as they develop into adults. Another estimate physicians use for predicting height within 2 to 3 inches (8.5cm) is a function of the parents' height. This prediction estimate is called the mid-parental, or median height of the parents, which explained 40% of the sex and age adjusted height variance **[Aulchenko et al]**. However, generally it is unpredictable to know which genes relating to height the child will inherit from their parents, and it is uncertain how the genes themselves interacted to determine adult height. A disadvantage common to both estimations is that they are not defined by a person’s genetic profile, but solely rely on the physiological measurement. Elucidating a strategy to accurately determine the causal mutations in polygenic phenotypes such as human height of an individual based on genetic markers promises to have wide application in the fields of pediatric endocrinology, embryonic screening, crime scene forensics, and the classification of polygenic diseases.

The first approach to developing a predictive model was to perform a Genome Wide Association Study (GWAS) on a large array of single nucleotide polymorphisms (SNPs) from a population of individuals. A GWAS is a statistical multiple test comparison used to identify genetic loci that are significantly associated with a particular trait. GWAS is commonly used in genetic epidemiology to detect and potentially prevent or treat genetic variations that contribute to common to complex diseases such as asthma, cancer, diabetes, heart diseases, among others. In this project, a GWAS was performed in order to identify causal SNPs statistically associated with human height.

Using a p-value prioritized list of SNPs which has been shown to explain more of height variability than genome-significant SNPS alone **[Wood et al**

1 Bioinformatics Department, Boston University
* Authors Contributed Equally

2014], the combination of major and minor alleles of each SNPs were mapped into a Euclidean space in order to train a model that predict the height of individuals. Each Euclidean mapping of all height associated SNPs represent a unique genetic signature that can be used to classify and predict adult height with a Support Vector Machine (SVM). It was necessary that the predictive model be lenient in the classification of human height since variation due to environment accounts for approximately 20% of height. From the 4 height stratifications, 3 binary stratifications of height and 1 multi-class stratification of height, we were able to classify height with accuracies between 71-96%. The best performing binary classification model achieved a mean cross-validation AUC of .85. Previous binary classification models in the literature only achieved an AUC of 0.75 using 184 SNPs and AUC 0.65 using 54 SNPs **[Liu et al 2013]**.

METHODS

Genetic Association

The GWAS was performed in R with the aid of several R packages. The genome wide study consisted of 313 individuals who published their 23andme and FamilyTree SNP array data to opensnp.org. 23andme and FamilyTree DNA both use Illumina Omni Express genotyping chips to identify autosomal, sex, and mitochondrial loci. The raw SNP array data files were curated to only include SNPs common to all individuals. This filtering protocol narrowed the SNP data from an average of over 900,000 SNPs to approximately 177,000 ubiquitously called SNPs. This effectively took care of the Person-Wide missing rate, or individuals in the dataset missing more than 5% of the data, and took care of the SNP-wide missing rate, SNPs that were missing alleles in more than 5% of the population. A Perl script was written to parse each raw SNP array data file and create a single CSV genotype file that contained the SNP data for all individuals where each line represented an individual and every pair of columns represented a SNP. The genotype file was imported into R as a matrix. The genotype file and the phenotype files were both indexed by user. The GWAS was performed using a linear regression model such as the one shown below:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

The linear regression model assumes that the error is normally distribution, and the data is approximately normal. To ensure this, each individual height was converted to centimeters, and the distribution was plotted in a histogram. A better visualization of the normality of the data was done using a QQ-plot. A t-test was also performed on the phenotypes to statistically quantify normality.

Once the SNP data was imported in R, a computationally intensive genome filtering protocol was done to ensure overall SNPs genomic quality. This was done by performing multiple (177616 choose 2) Hardy Weinberg Equilibrium Chi Square tests. A function from the '*genetics*' R package was used to perform the HWE chi square test. The null hypothesis is that the SNP in the population is in Hardy Weinberg Equilibrium. The alternative hypothesis is the SNP in the population is not in Hardy Weinberg Equilibrium. In order to maximize runtime speed and performance especially considering the large enumeration, the important implementation of Vectorization was used to significantly reduce the number of computations. Instead of calling functions multiple times on vectors of length 1, it is better practice to call a function once on vectors of multiple length. This significantly reduced execution time. This part of the implementation was computationally intensive, and so was run in the background on the Boston University Shared Computing Cluster (SCC).

After the SNPs were filtered by genotype, the alleles were converted from nucleotides (A,G,C,T) into numeric values based on the major and minor allele frequency; this step was necessary for the linear regression. Let A_1 and A_2 be the minor and major allele respectively. An additive genotype matrix (X_a) and dominant genotype matrix (X_d) were created using the following scoring matrix:

$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

The X_a and X_d matrices had N rows of individuals and M columns of SNPs.

During this part of the R Implementation, a genotype scoring matrix required by EIGENSTRAT was created and exported as a text file. The EIGENSTRAT manual reads:

The genotype file contains 1 line per SNP. Each line contains 1 character per individual: 0 means zero copies of reference allele. 1 means one copy of reference allele. 2 means two copies of reference allele. 9 means missing data.

The most frequent allele (the major allele) was chosen as the reference allele. EIGENSTRAT was used to do the Principle Component Analysis which corrects for population stratification in genome-wide association studies. The first ten principal components were obtained and used to fit the linear model. This part of the R implementation was run on the Boston University SCC.

After the dummy variable coding, the linear regression model was fit using the principal components. The 5th principle component was found to account for most of the variation of the model therefore it was the only Principal Component included in the linear models. The null hypothesis of the GWAS is that there are no causal loci statistically associated with the phenotypic trait. The alternative hypothesis is there are causal loci statistically associated with the phenotypic trait. The fitted linear regression was run and an ANOVA likelihood test was applied to obtain the p-values for each SNP. The obtained p-values were used to create the QQ-plot and the Manhattan Plot. An R package, '*qqman*', was used to generate the QQ-plot and the Manhattan Plot. The genomic inflation factor, Lambda, was calculated to check for p-value inflation.

Mapping Alleles to Euclidian Space

After a list of SNPs were obtained from the GWAS, the p-value prioritization **[Frayling, T. M. 2014]** and mapping of allelic combinations into a Euclidean space **[Brinza et al 2010]**. A unique set of 500 features or scores were computed for every individuals as the training features for the SVM model. With the list of the top 500 p-value prioritized SNPs, the allelic combinations were mapped to a Euclidean space using the following scoring method:

$$\boldsymbol{\delta_i = \{w/w = 1; w/m = 2;\ m/m = 3\}}$$

The delta represents the score of a single SNP while W and M represent if the allele is the wild type or mutant allele respectively in the training population of 313 individuals from OpenSNP. Any missed allele calls from the genotyping were imputed to be heterozygous using the OpenSNP individuals as the reference population. After a score of each SNP was computed for all prioritized SNPs in all 313 individuals, the values were scaled and normalized to have a mean and variance of 0 and 1 respectively to maintain compatibility with SciLearn-kit algorithms for classification. [**Pedregosa et al 2011**].

Height Class Stratification

Using the unsupervised learning algorithm, k-means clustering with k being 2 or 3, users were given labels for the height class they belong most closely clustered to **[Figure X]**. In the 3 class classification system Class 1 (short) consists of individuals between 150cm-170.2cm, Class2 (average) includes individuals between 170.2cm-180cm, and Class 3 (tall) includes individuals between 180-193cm. Each class has a standard deviations of 5.13, 3.36, and 3.73 respectively which shows that k-means clustering stratified the height classes relatively equally. We also created several binary height classes. The top 500 prioritized SNPs were evaluated for their predictive capabilities using the previously described methodology on five different multiclass models to evaluate the accuracy of 4 different multiclass SVM models and a simple Naive Bayes classification model in Python SciLearn-Kit. The 5 multiclass SVM models had adjusted parameters to evaluate over fitting of the model to the OpenSNP dataset. C parameters of 0.1, 0.5, 1.0 and the kernel function Radial Basis Function (rbf) were evaluated. 1 multiclass SVM with C parameter of 1.0 and a linear kernel function. The last model was a simple Naive Bayes Classifier. The factor C in SVM models “is a parameter that allows one to trade off training error vs. model complexity. A small value for C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM." **[Joachims 2002, page 40]**. The performance of all models were evaluated with 9-fold cross-validation due to the low number of individuals in the training and test datasets.

Training a Support Vector Machine

Machine Learning algorithms such as Support Vector Machine are used widely in Bioinformatics. Support Vector Machine attempt to maximize the margin between features of labeled training data and predict new labels for unseen data. The 2 class system of

Support Vector Machines

Results

GWAS

The mean height of the individuals in the GWAS is 174.6 cm. The variance of the individual's heights in the GWAS was 102.5 cm. The standard deviation of height was 10.12 cm [**SEE TABLE 1**]. The distribution of the heights is approximately normal. A normal density curve and a kernel density curve laid over the distribution relatively well **(Figure 1)**. A Theoretical Density vs Empirical Kernel Density curve shows that distribution is approximately normal. A QQ-plot of the phenotype shows that the normality assumption of the GWAS is satisfied **(Figure 2)**. A two sample t-test between a randomly normal distribution generated from the mean and standard deviation of the heights, and the actual distribution of the height yielded a p-value of approximately 0.7. The null hypothesis is that the two samples have both been drawn from the same population. The alternative hypothesis is that the two sample are not drawn from the same population. The p-value from the two sample t-test is 0.7 therefore we fail to reject the null hypothesis and conclude that the two samples are drawn from the same population.

The linear regression model used for generating the association test statistic (p-value) was taking from the Analysis of Variance of two models and storing the p-value:

$$(model_1)\, height = X.mx + PC5$$

$$(model_2)\, height = PC5$$

$$anova(model_2, model_1)$$

The fifth PC was determined to account for a significant amount of the model's variance. **(Figure 3)**. The QQ plot is a great visualization tool identifying confounding factors in a GWAS such as population stratification. If these confounding factors are unaccounted for in the linear model of the GWAS then the p-values will appear inflated. This is quantified by the genomic inflation factor. The genomic inflation factor λ is defined as the ratio of the median of the observed distribution of the test statistic to the expected median. Values of lambda greater than 1 indicate that the association test statistics (p-values) are inflated and an excess of false positives. The value of lambda in this GWAS of $\lambda = 0.9985$ show that there are no evidence of p-value inflation in the model.

QQ plot shows the observed distribution of p-values across the SNPs in the GWAS (y-axis) compared to the expected distribution p-values under the null hypothesis (x-axis). A GWAS under the null hypothesis where there are no causal polymorphisms, the QQ plot would be a straight line across the diagonal. In this case, the observed p-values should adhere to a uniform distribution. In the ideal GWAS case where there are causal polymorphisms, the QQ plot will form a straight line across the diagonal matching X=Y until it slightly deviates towards the end. This 'tail' at the end of the QQ plot represents a small number of causal polymorphisms among the tens of thousands of unassociated SNPs.

The QQ plot in this GWAS showed no sign of p-value inflation ($\lambda = 0.9985$). The lack of a distinct 'tail' at the end of the QQ plot suggests that there may not enough power in the GWAS study to adequately detect significant associations. (Figure 4) In genome wide studies, researchers usually require a sample of at least 200,000 individuals in order to detect significant SNPs; the sample size of 313 individuals undermined the power of this GWAS.

The Manhattan Plot of the genetic loci confirmed the suspicion of a low power study. (Figure 5) There was only one SNP with p-value above the suggestion line. There were no SNPs with p-values above the genome wide significance threshold.

SVM Classification Performance

After model training and 9 fold cross-validation with different supervised learning models it was determined that a multi class SVM with a C parameter of 1.0, RBF kernel function, and a p-value prioritized list of the top 500 SNPs provided the highest accuracy during cross-validation. The SVM(c=0.1, kernel=rbf) achieved an accuracy of 47%, SVM(c=0.5, kernel=rbf) achieved 61%, SVM(c=1.0, kernel=rbf) achieved 71%, SVM(c=1.0, kernel=linear) achieved 57%, and the Naive Bayes classifier achieved 53% accuracy **[Figure Y]**. The performance of the binary classification models were evaluated by taking the average AUC from 9-fold cross validation. One-vs-all SVM classifications determined if an individuals was extreme tall or short versus not. The 95th percentile, 90th percentile, 5th percentile, and 10th percentile of our reference population were shown to have a mean AUC of 0.85, 0.84, 0.76, and 0.70 of with our SVM classification.

Conclusion and Discussion

A major challenge in genome wide association studies is setting an appropriate threshold to correct for the multiple hypothesis testing problem. The probability of falsely identifying a significant SNP merely due to chance even if the particular outcome were low, increases as more hypotheses tests are performed. In a multiple test case, we would therefore make lots of false positive associations. This is essentially the multiple testing problem: When statistical tests are used repeatedly, such as in the multiple comparisons in GWAS, there is a greater potential of making a Type I error. The multiple testing problem is the case of rejecting the null hypothesis incorrectly more frequently under a single hypothesis test.

The Bonferroni correction can be applied to reduce the probability obtaining false-positive (Type I error) in multiple hypothesis testing on a single set of data. However, the Bonferroni correction becomes too conservative as the number of hypothesis tests increases. As a result, this reduces the power of the GWAS and increases the risk of generating false negatives (type II error). Type I and Type II errors are inversely related. The False Discovery Rate (FDR) is also procedure for controlling Type I error but is not as conservative as Bonferroni correction and has allows greater statistical power in GWAS studies.

The Bonferroni correction threshold can be calculated by dividing the specified Type I error (alpha) by the number of SNPs in the GWAS:

$$Bonferroni\,correction = \alpha / N$$

The False Discovery Rate can be calculated as shown below:

$FDR = (N \times \alpha)/R$, where R is the number of cases where the null hypothesis is rejected. Intuitively, the FDR is the proportion of cases where one rejects the null hypothesis that are false positive.

Removing genotypes with low minor allele frequencies (MAF) less than 6% was an important part of the filtering protocol as it ensures more accurate results in the analysis. The power of the GWAS, which is the probability that the null hypothesis will be rejected when the null hypothesis is false, tends to increase as the MAF increases. As the power increases, the chances of Type II error, the failure to reject a false null hypothesis (a false negative), occurring decreases. Type II error cannot be controlled because it depends on the true value of the parameter for probability model which is unknown. In this GWAS, the sample size was not large enough for sufficient statistical power. The GWAS showed that there were no significant SNPs in the study.

Linkage Disequilibrium (LD) is the non-random association of alleles at one or more loci. In the presence of LD, the hypothesis test of the GWAS are not independent. SNPs that involve that in LD is that GWAS would have to involve haplotype analysis. A haplotype is a collection of specific alleles in a cluster of tightly-linked genes on a chromosome that are likely to be inherited together. Haplotype analysis requires a different ANOVA scoring and not much is known about it. As a result, most researchers favor single point GWAS.

Previous attempts to classify human height focus on binary height classification of extreme height classes with relatively few SNPs that could reach genome-wide significant in relatively low powered GWAS studies. Although this height GWAS was low powered **Wood, T. M. 2014** has shown that decreasing the p-value threshold from genome-wide significance does not drastically reduce the predictive capabilities of a height model. By decreasing the significance threshold more SNPs can be used as indicators to help explain more of the variance and heritability of human height. Although the genetic function these SNPs may be involved with has not been explored the goal is to find genetic "signatures" for polygenic traits such as adult human height.

Using the top 500 significant SNPs of each individual, map the combination of major and minor alleles for each nucleotide into a Euclidean space, train a model, and predict the height of users. Each Euclidean mapping of all height associated SNPs represent a unique genetic signature that could be used to classify and predict adult height with a Support Vector Machine (SVM). It was necessary that the predictive model be lenient in the classification of human height since variation due to the environment accounts for 20% of height.

Stratified 5-fold cross-validation was performed on the dataset using the leave-one-out method, and also ensuring there was a balance of height classes in each of the 5 training folds. By repeating stratified 5-fold cross-validation on a monotonically increasing collection of SNPs, from 40 - 500 SNPs, it was shown that utilizing increasing numbers of prioritized SNPs did improve the classification accuracy during cross-validation **(Table 2)**.

We also recreated the upper and lower 5th quantiles to compare the accuracy and AUC of our model against previous predictive model. W

Future Direction

A higher power GWAS with more individuals will enable the researchers to obtain more significant list of SNPs. This will invariably improve the accuracy of the SVM because these genetic loci will be more statistically associated with height and offer better estimate and statistical power. Performing a GWAS on whole exome data is a possibly worth exploring.

For future study, the list of significant SNPs will be examined for evidence of LD. In the presence of LD, the GWAS will be modified to account for haplotype genetic linkage. Literature shows that accounting for haplotype analysis in the presence of LD should improve evidence for candidate genes or SNPs [**Barendse 201**1]. The scientific paper mentions that GWAS almost invariably use single point analysis to detect casual loci however, increased levels of information be achieved by the analysis of haplotype analysis. Since many genetic loci are inherited as haplotype blocks we would gain additional information by performing imputation based on the ancestry of the individual.

Another genetic phenomena that may be occurring in height genetic association is epistasis. A paper from 2006 found that epistasis between loci on chromosomes 2 and 6 influences human height **[Liu et al 2006]**. A different scoring system would need to be developed for these epistatic interactions. As the case with haplotype, account for these epistatic interactions has potential to improve the estimation power of causal SNPs.

Our classification models showed high accuracy for binary classification on the upper and lower percentiles but, we believe that biological and environmental factors make it harder to determine the height of individuals in our height model with 3 height groups. We believe this is because many individuals who are on the border of being in a short, normal, or tall height class may by their reported height my genetically belong to another height group. To address this mislabeling we will experiment with creating overlapping height groups to allow for 20% of height. This means the short height group we be composed mainly of phenotypically short individuals and a few normal height individuals that are close to being classified as short. We will select individuals to represent height groups using sampling with replacement.

Comparison of Original and Revised Proposal

In our original proposal we proposed to create a genetic model that would look at SNPs of interest from 23andme and other personal genetics companies such as Ancestry and FamilyTree DNA to predict the adult human height of an individual. To build a rigorous genetic model we would perform a GWAS study to identify novel SNPs, locate height related SNPs in the literature, and train a Hidden Markov Model. We also considered the effects of environment on adult height and proposed a strategy to address the environmental effects. However, our primary aim was to determine how tall a person was given a set of alleles for height associated SNPs and their ancestry. After receiving feedback from our reviewers it was recommend that we use more formal terminology and formatting in our research proposal. Also, the reviewers recognized that we should explicitly state the societal importance of predicting human height, and discuss its positive and negative implications. Although the idea to study human height might appeal to human curiosity, the biological and ethical implications of such advancements will allow us to foresee tangible goals and applications from out project. We also realized we should have results to show, even if we did not achieve our goal of classifying height more accurately than other published methods. So our revised proposal included biological implication of a genetic model, our approach for performing a GWAS in R, classifying human height with a SVM. As a secondary goal we proposed to calculate the necessary power of our statistical test to determine the causal

SNPs associated with height. Our revised proposal more thoroughly explain our motivation, aims, and gave a timeline to when we expected to complete our project.

Commentary of the Experience

Working in a 2 person group was a good experience but we encountered many challenges trying to locate and download publicly available genotype data that included the person's height. Since metadata such as height can be used to personally identify individuals, special permission must be granted by the NIH or similar regulatory bodies to access the height and genotype data in many publications. So we explored the availability of personal genomics datasets voluntarily released by individuals and located OpenSNP and the Harvard Personal Genome Project. We found that many personal genomics databases provide incomplete meta-data about the individual's height, and multiple formats of genotype files. We had to write web crawlers and file parsing scripts to search for each user's height, age, gender, ethnicity, and standardize the data into a compatible format. We enjoyed the challenge of taking the statistical, algorithmic, and computational tools we developed this year and applying them to real data that could potentially impact the field.

Commentary of the Peer Review Process

The peer review process was invaluable in shaping the project design. In the original proposal, team members had planned to classify height using an HMM. Team members planned to use the genetic ancestry and the distribution of allele at specific loci as the parameters for a Hidden Markov model. Team members were going to calculate the posterior probability of a height given the alleles and genetic ancestry. The ancestry of individuals were to be calculated using the ADMIXTURE software and the HapMap2/1000 Genomes populations. Both peer reviewers suggested that an HMM would be a challenge task in of itself and that perhaps we should consider another classifier or develop a detailed timeline of goals and solutions to potential problems. Upon reading further literature, team members decided to develop an SVM classifier to predict a person's height class based on their genetic profile.

Division of Labor

It was very much of a collaborative effort and an equal division of labor. Both team members brought invaluable computational skill and biology acumen to the project. Samuel proposed the idea of doing a GWAS in R and Demarcus proposed the idea of predicting height using a Support Vector Machine. Demarcus wrote the web crawlers to obtain the phenotypes and genetic data from the OpenSNP and Harvard Personal Genome Project websites, and curated the data using a series of Bash commands. Samuel and Demarcus worked on the programming together. Samuel wrote the R code for the GWAS and developed the linear regression model to account for p-value inflation and population stratification. Samuel was essentially project manager; he ensured that progress on the project was moving along at a steady pace. Demarcus headed up the research and development aspect of the project. His responsibility included keeping up to date on current literature on human height and experimenting with models for the SVM. Samuel and Demarcus both contributed to connecting the genetic result from the GWAS to the SVM. Team members contributed equally to the power point presentation and the scientific final report.

References

1. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. 52, 399–433 (1918)

2. Silventoinen, K. et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 6, 399–408 (2003).

3. Visscher, P.M. et al. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet. 2, e41 (2006).

4. Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics, 46(11).doi:10.1038/ng.3097

5. Aulchenko, Y. S., Struchalin, M. V, Belonogova, N. M., Axenovich, T. I., Weedon, M. N., Hofman, A., Borodin, P. M. (2009). Predicting human height by Victorian and genomic methods. European Journal of Human Genetics: EJHG, 17(8), 1070–5.doi:10.1038/ejhg.2009.5

6. Pedregosa *et al.* (2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

7. Brinza, D., Schultz, M., Tesler, G., & Bafna, V. (2010). RAPID detection of gene–gene interactions in genome-wide association studies. Bioinformatics, 26(22), 2856–2862. doi:10.1093/bioinformatics/btq529

8. Barendse W (2011) Haplotype Analysis Improved Evidence for Candidate Genes for Intramuscular Fat Percentage from a Genome Wide Association Study of Cattle. PLoS ONE 6(12): e29601. doi: 10.1371/journal.pone.0029601

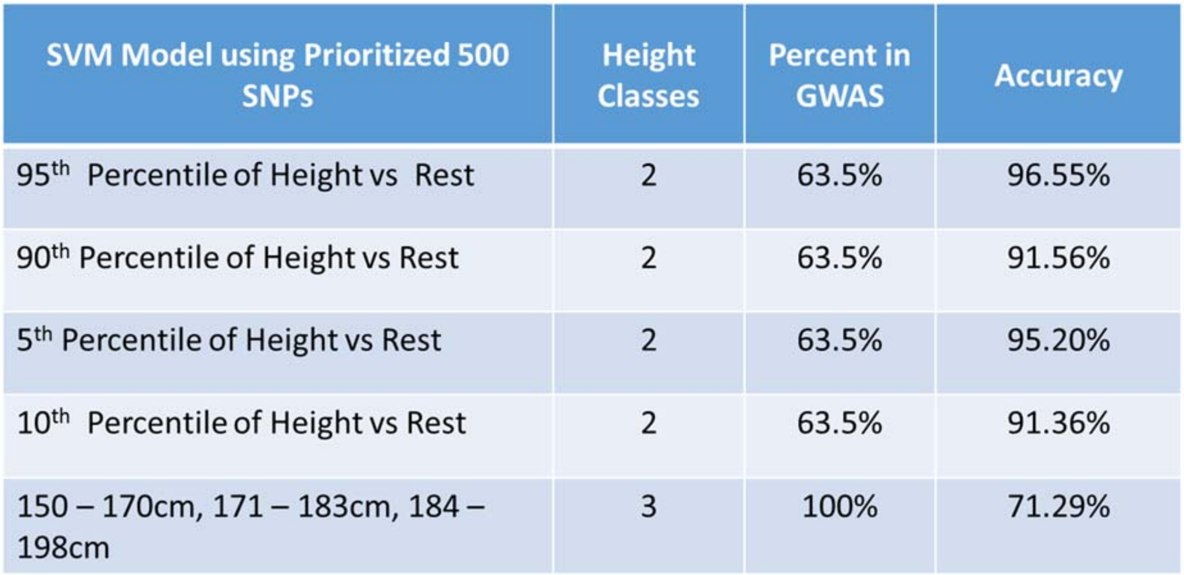
9. Liu, Y.-Z., Guo, Y.-F., Xiao, P., Xiong, D.-H., Zhao, L.-J., Shen, H., … Deng, H.-W. (2006). Epistasis between loci on chromosomes 2 and 6 influences human height. The Journal of Clinical Endocrinology and Metabolism, 91(10), 3821–5. doi:10.1210/jc.2006-0348

Appendix

Table 1: Descriptive Statistics of GWAS Population

Mean Height (cm)	Variance (cm)	Standard Deviation (cm)
174.6	102.5	10.123

Table 2: Accuracies of the SVM Classification Models



SVM Model using Prioritized 500 SNPs	Height Classes	Percent in GWAS	Accuracy
95th Percentile of Height vs Rest	2	63.5%	96.55%
90th Percentile of Height vs Rest	2	63.5%	91.56%
5th Percentile of Height vs Rest	2	63.5%	95.20%
10th Percentile of Height vs Rest	2	63.5%	91.36%
150 – 170cm, 171 – 183cm, 184 – 198cm	3	100%	71.29%

Figure 1: Distribution of Heights for the GWAS Study

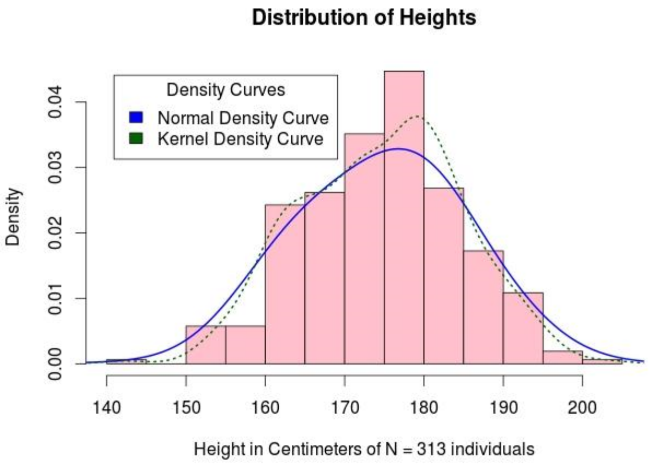
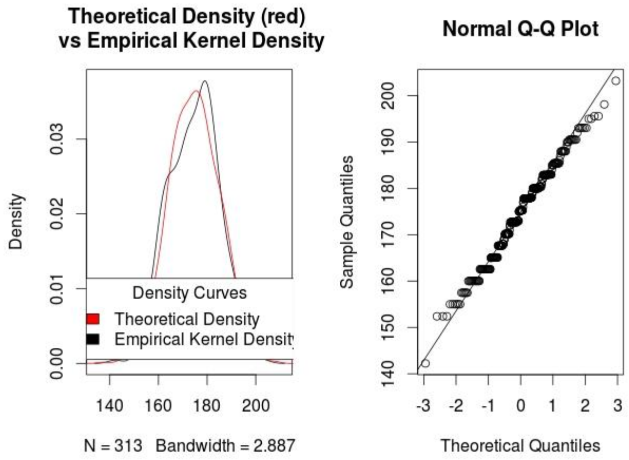
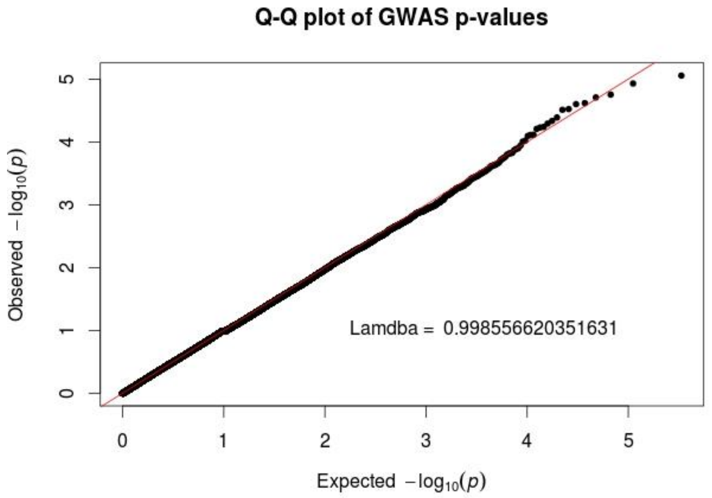
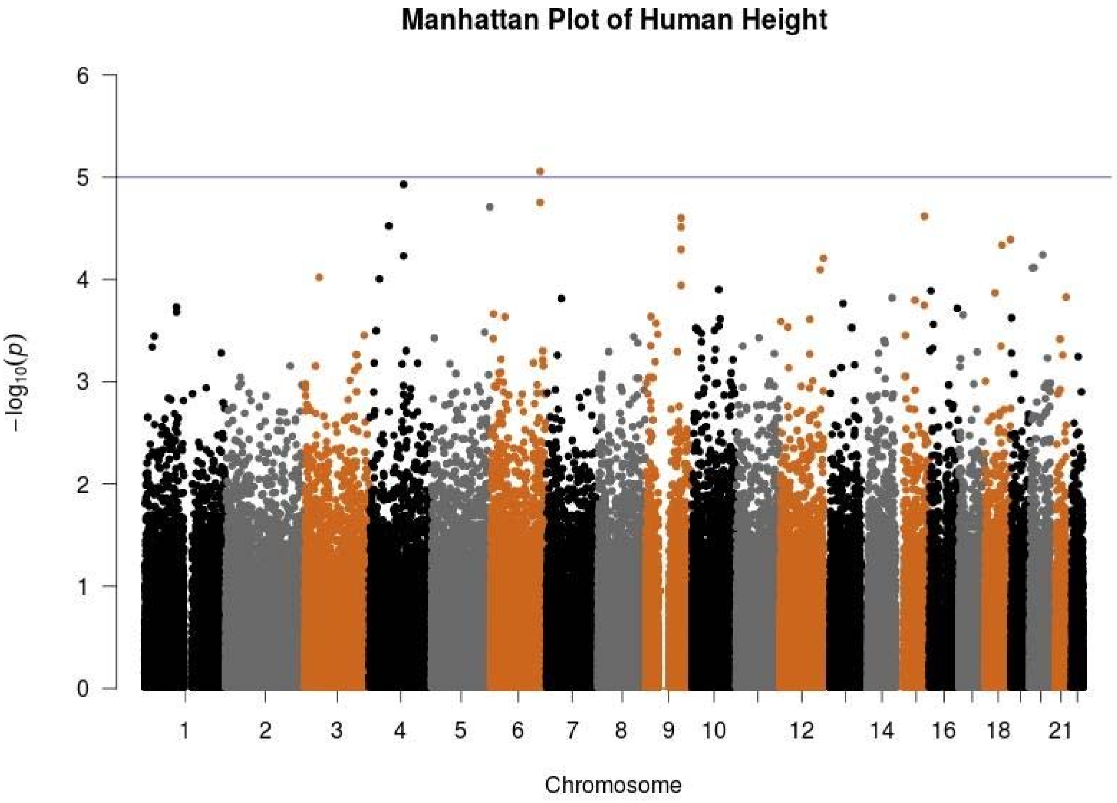
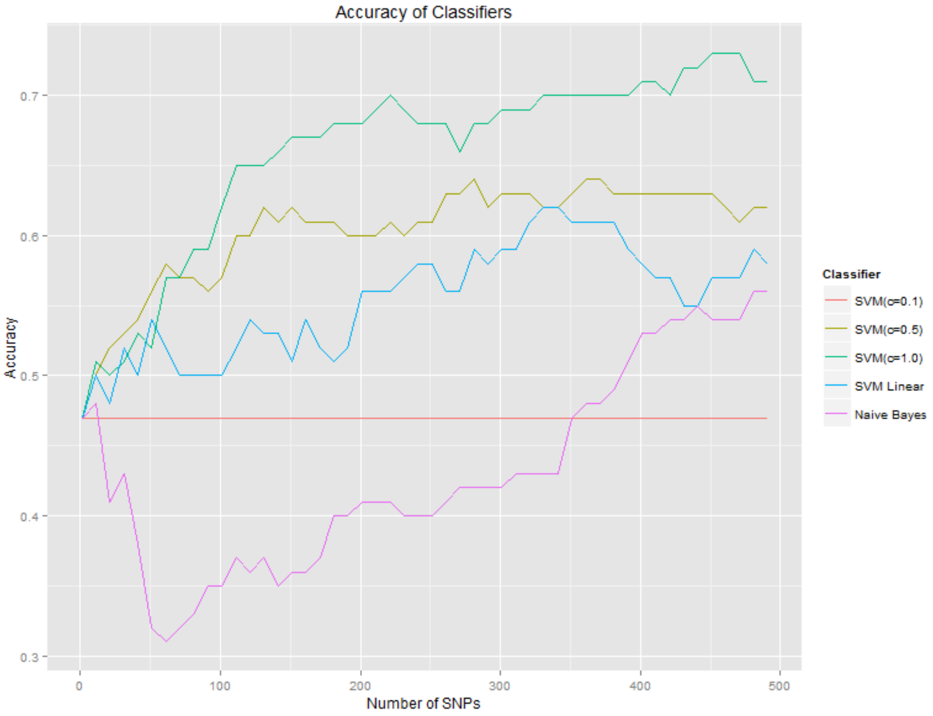


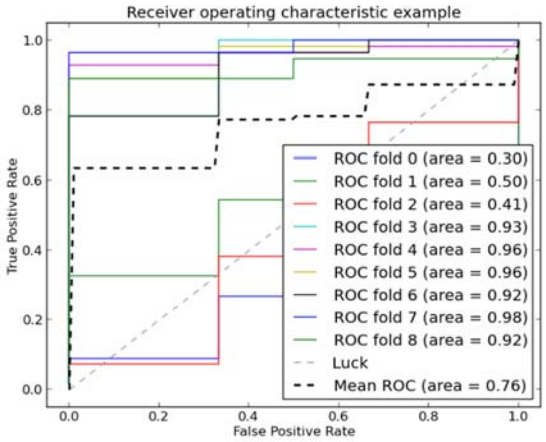
Figure 2: Normal Distribution of Heights and QQ Plot

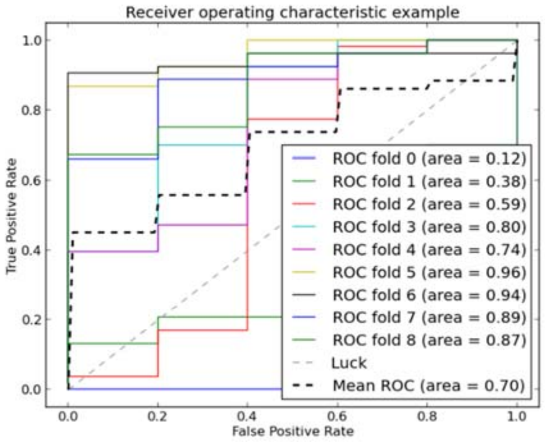


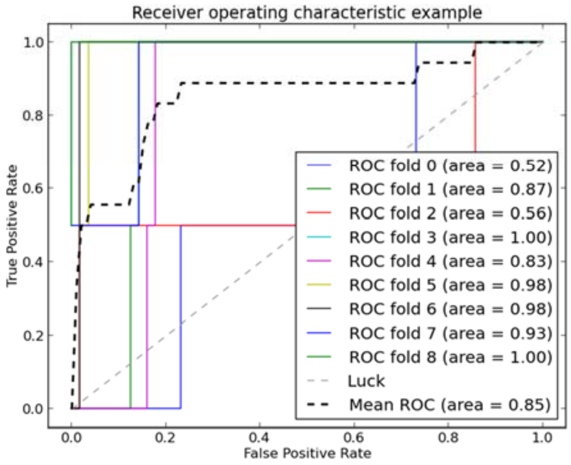


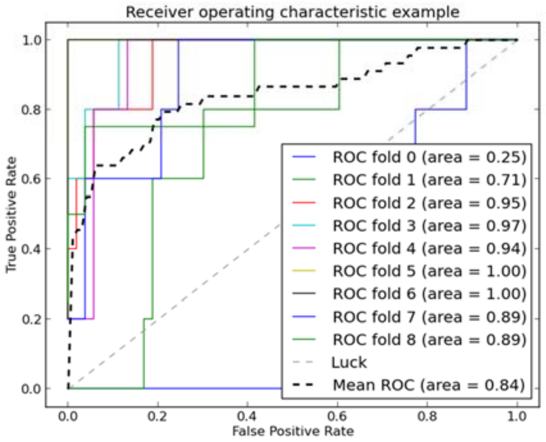

Manhattan Plot of Human Height
$-\log_{10}(p)$
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
Chromosome

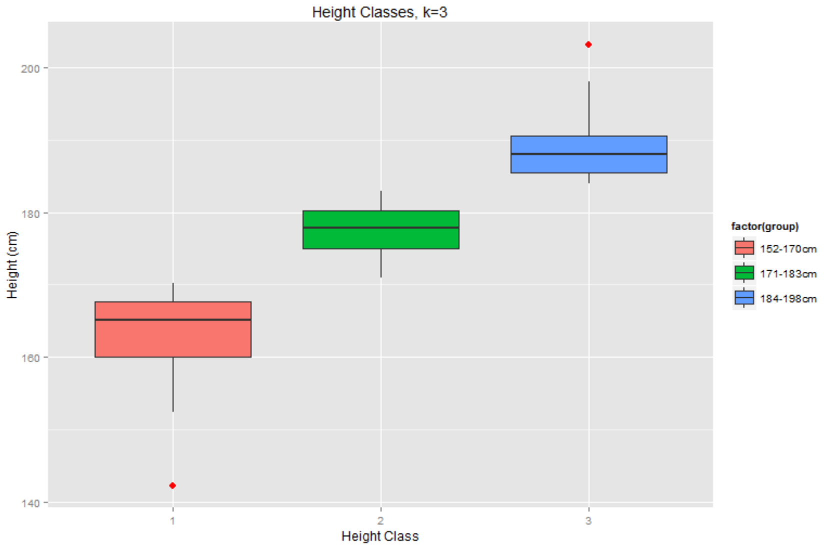

Accuracy of Classifiers
Accuracy
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500
Number of SNPs
Classifier
SVM(c=0.1)
SVM(c=0.5)
SVM(c=1.0)
SVM Linear
Naive Bayes


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (area = 0.30)
ROC fold 1 (area = 0.50)
ROC fold 2 (area = 0.41)
ROC fold 3 (area = 0.93)
ROC fold 4 (area = 0.96)
ROC fold 5 (area = 0.96)
ROC fold 6 (area = 0.92)
ROC fold 7 (area = 0.98)
ROC fold 8 (area = 0.92)
Luck
Mean ROC (area = 0.76)


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (area = 0.12)
ROC fold 1 (area = 0.38)
ROC fold 2 (area = 0.59)
ROC fold 3 (area = 0.80)
ROC fold 4 (area = 0.74)
ROC fold 5 (area = 0.96)
ROC fold 6 (area = 0.94)
ROC fold 7 (area = 0.89)
ROC fold 8 (area = 0.87)
Luck
Mean ROC (area = 0.70)


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (area = 0.52)
ROC fold 1 (area = 0.87)
ROC fold 2 (area = 0.56)
ROC fold 3 (area = 1.00)
ROC fold 4 (area = 0.83)
ROC fold 5 (area = 0.98)
ROC fold 6 (area = 0.98)
ROC fold 7 (area = 0.93)
ROC fold 8 (area = 1.00)
Luck
Mean ROC (area = 0.85)


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (area = 0.25)
ROC fold 1 (area = 0.71)
ROC fold 2 (area = 0.95)
ROC fold 3 (area = 0.97)
ROC fold 4 (area = 0.94)
ROC fold 5 (area = 1.00)
ROC fold 6 (area = 1.00)
ROC fold 7 (area = 0.89)
ROC fold 8 (area = 0.89)
Luck
Mean ROC (area = 0.84)


Height Classes, k=3
Height (cm)
200
180
160
140
1
2
3
Height Class
factor(group)
152-170cm
171-183cm
184-198cm