Parents vs Genetic Power
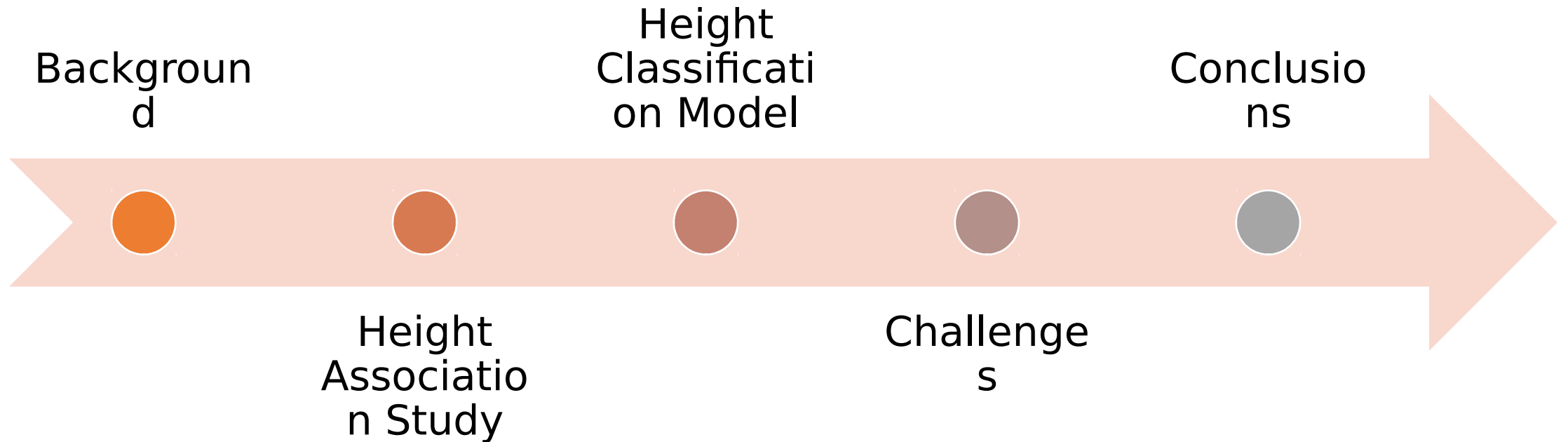
Predicting Adult Human Height

Samuel Moijueh

Demarcus Briers

Presentation Outline

Background

Height Association n Study

Height Classificati on Model

Challenge s

Conclusio ns

How much of human height is genetic?

Human height is a complex polygenic trait, difficult to predict

There are currently no genetically rigorous models used to predict a person's height

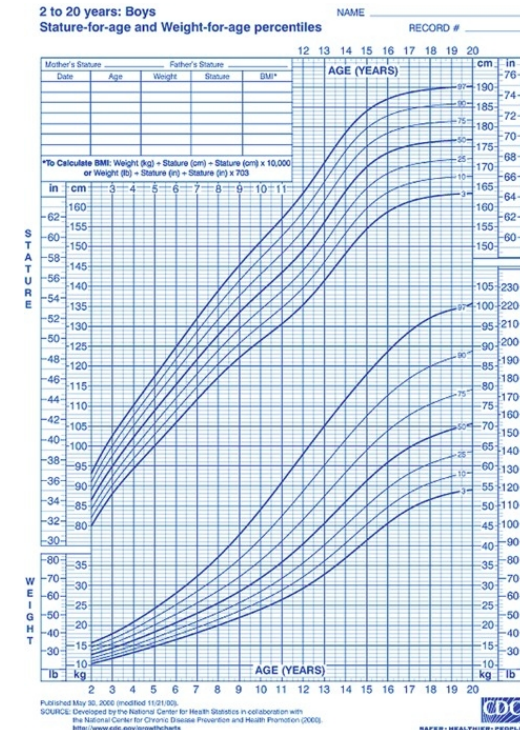Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

Current Methods to predict human height?

Mid-Parental Estimate

CDC Height-Weight Chart

Example of Mid-parental Estimate

Formula
- Males: (Fa + Mo + 5in) / 2
- Females: (Fa + Mo -5in) / 2

Example(Demarcus)

Predicted: (68in + 59in + 5in) /2 = **66in (+/- 3.3in)**

Actual: **67in**
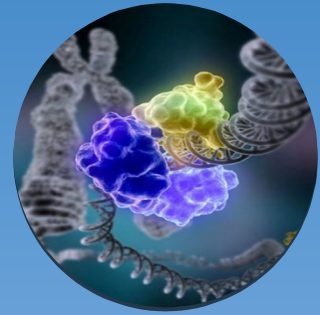
Motivation for Genetically Predicting Human Height

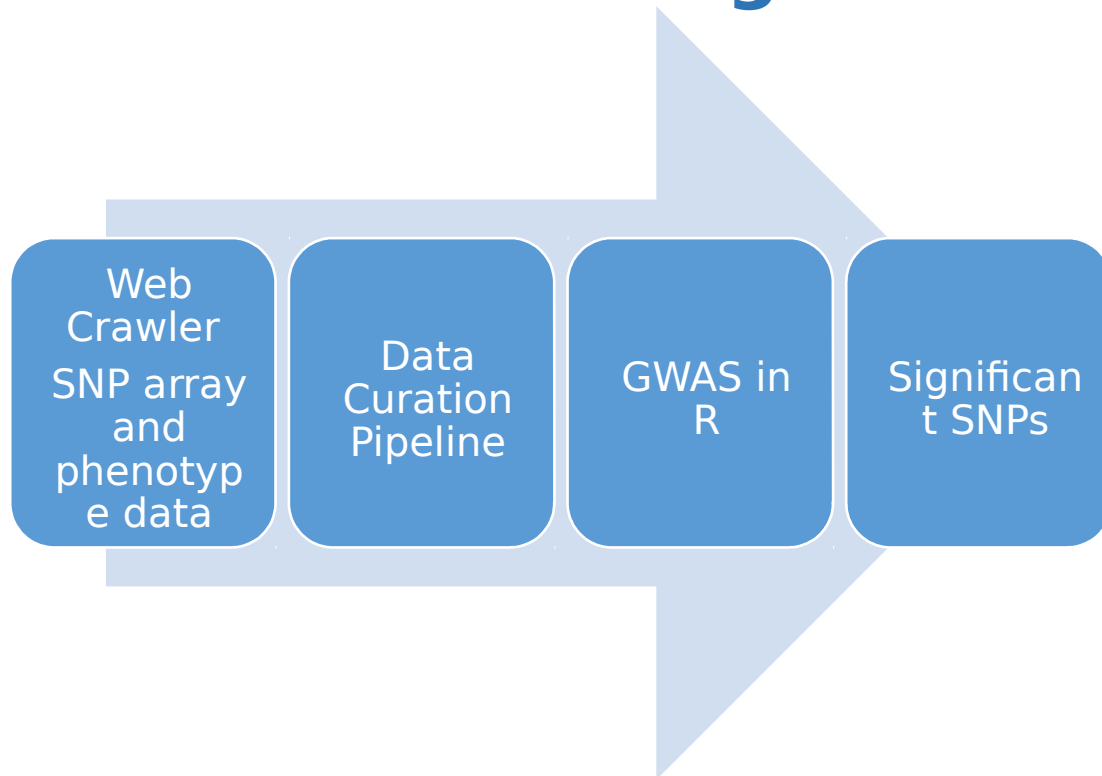Crime Scene Investigations Genetic Privacy

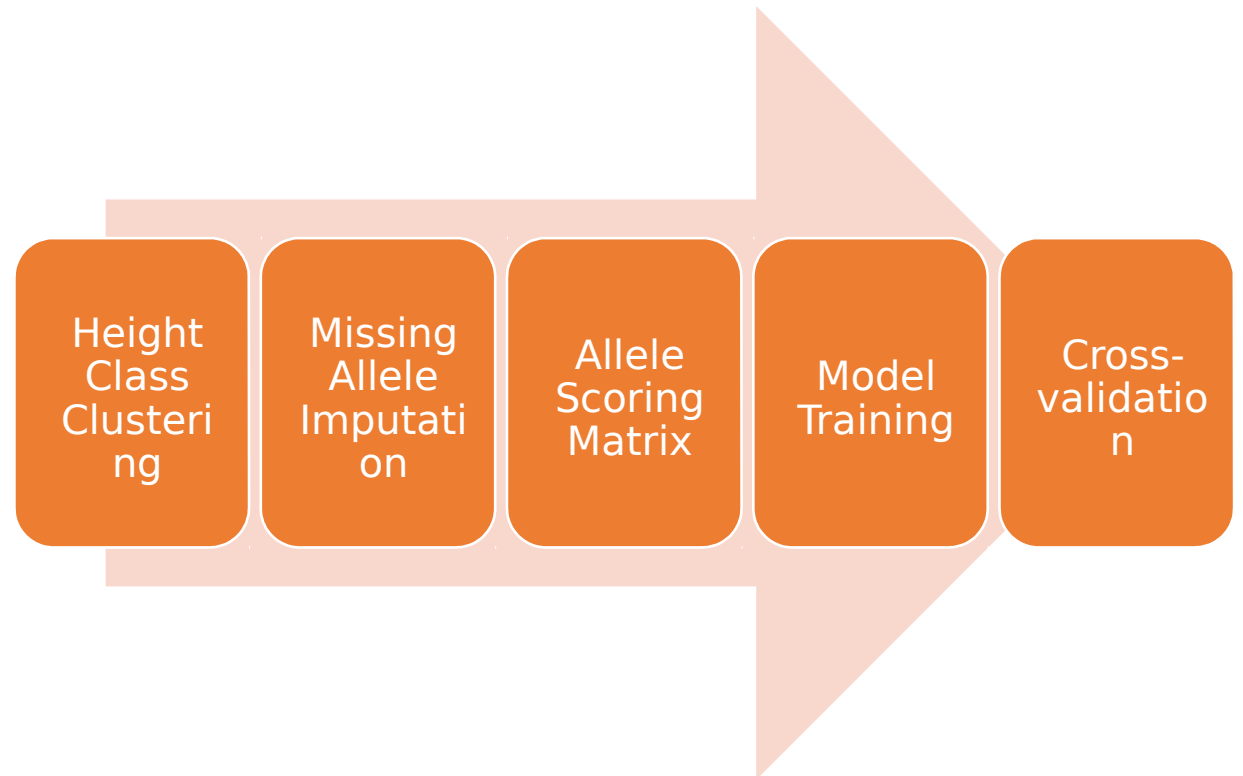Embryonic Screenings

Pediatric Endocrinology & Hormone Therapy

Reveal nature of polygenic traits (skin, eye color, complex disease)

Objectives of Adult Height Study
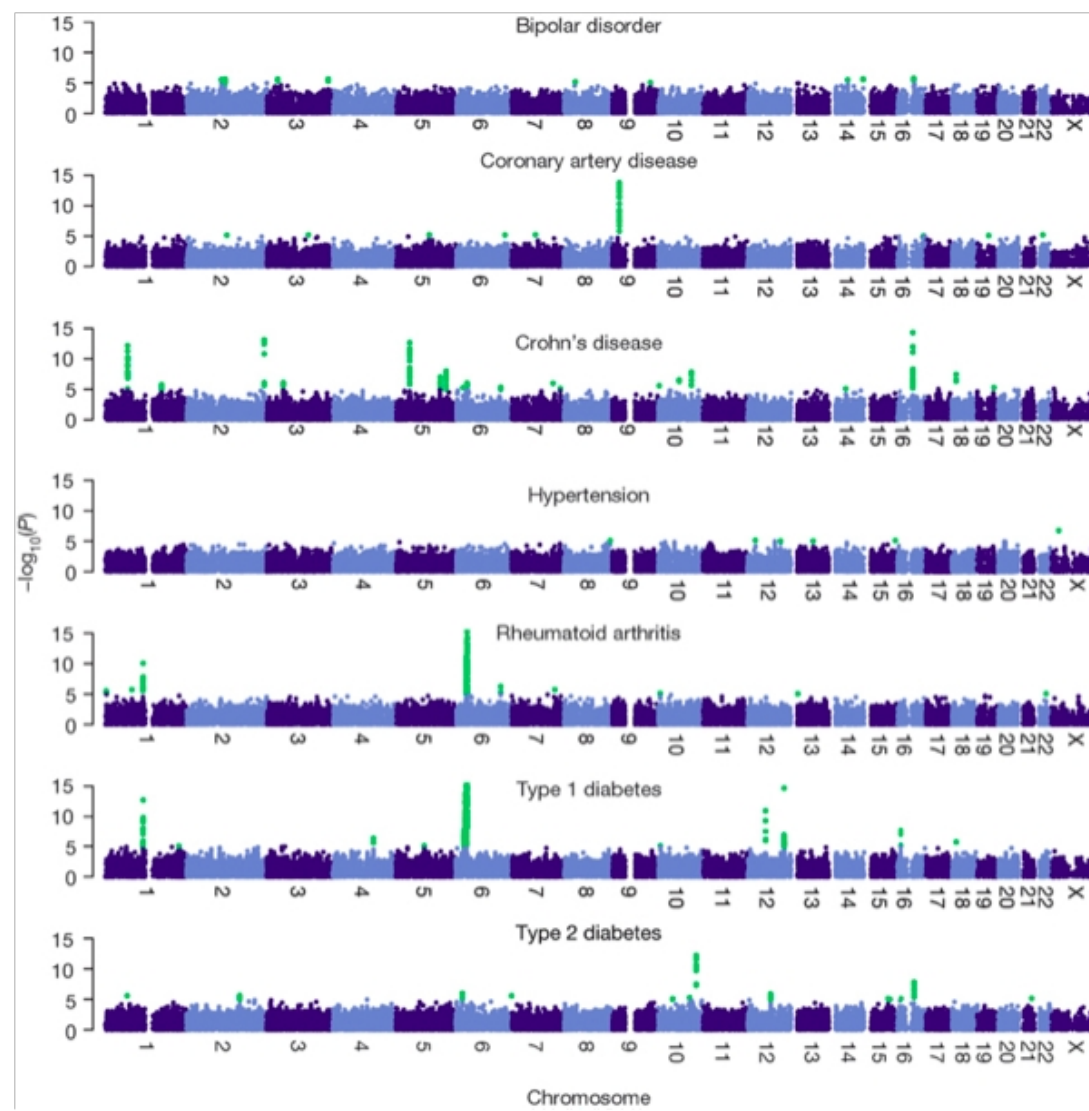
Identify SNPs Associated with Height

Train an SVM to Classify Height

Genome Wide Association Study

Multiple Hypothesis tests performed to identify causal single nucleotide polymorphisms (SNPs) that are statistically associated with human height
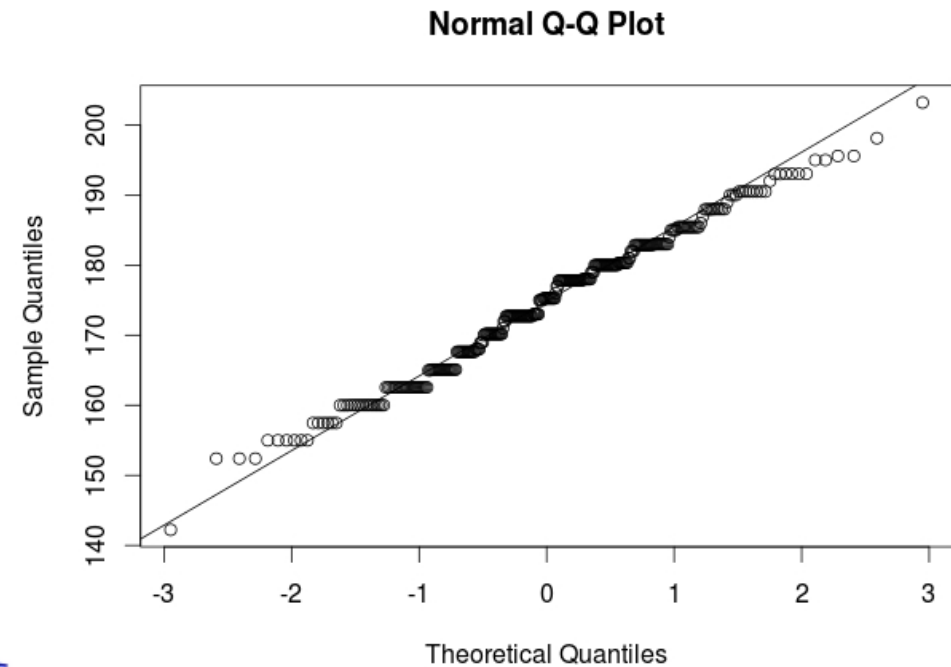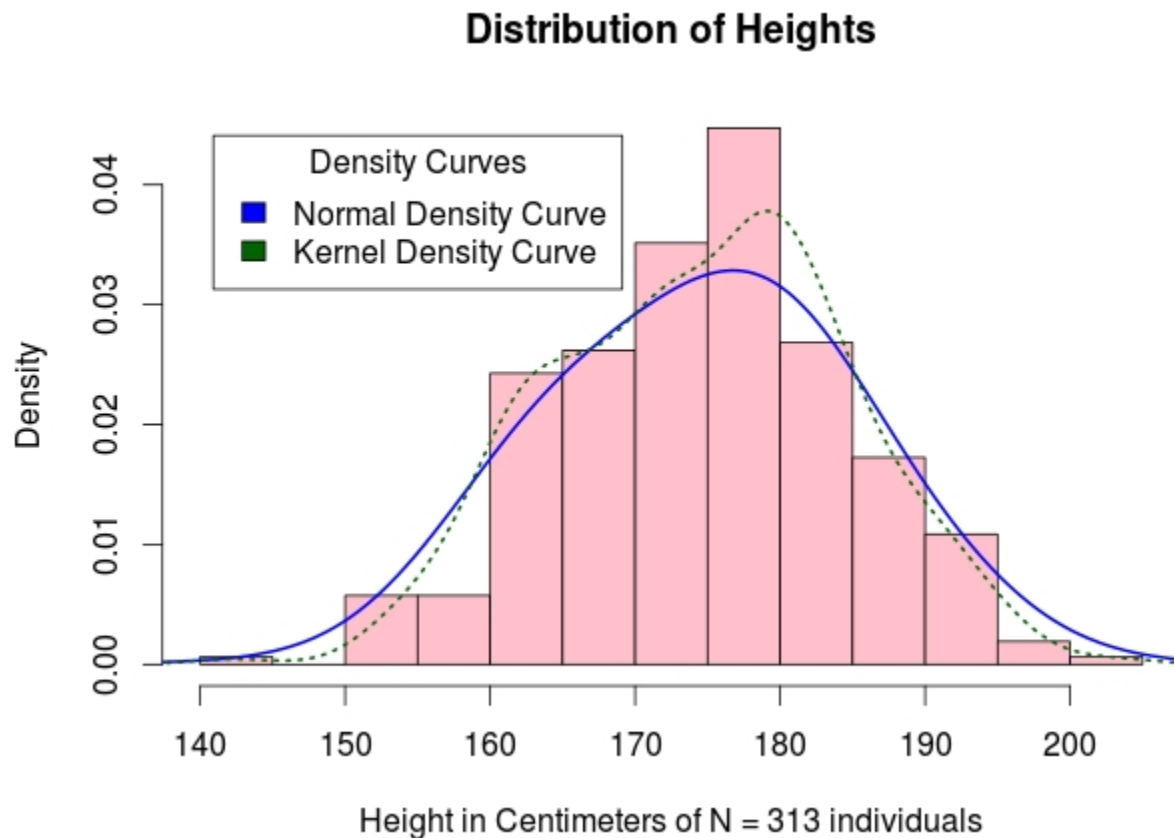
$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

Genome Wide Association Normality Assumption

313 individuals, SNP array data obtained from 23andme and FamilyTree

Distribution of Heights

Normal Q-Q Plot

Developing the Linear Regression Model to Perform GWAS

Genome Filtering Protocol

- Removed SNPs with MAF < 6%, poor genotyping, or not in HWE (performed HWE Chi-test)

EIGENSTRAT: Principle Component Analysis to correct for Population Stratification

Dummy Variable Coding based on the MAF

- - Necessary for converting alleles into numbers for the Regression model

Linear Regression Model
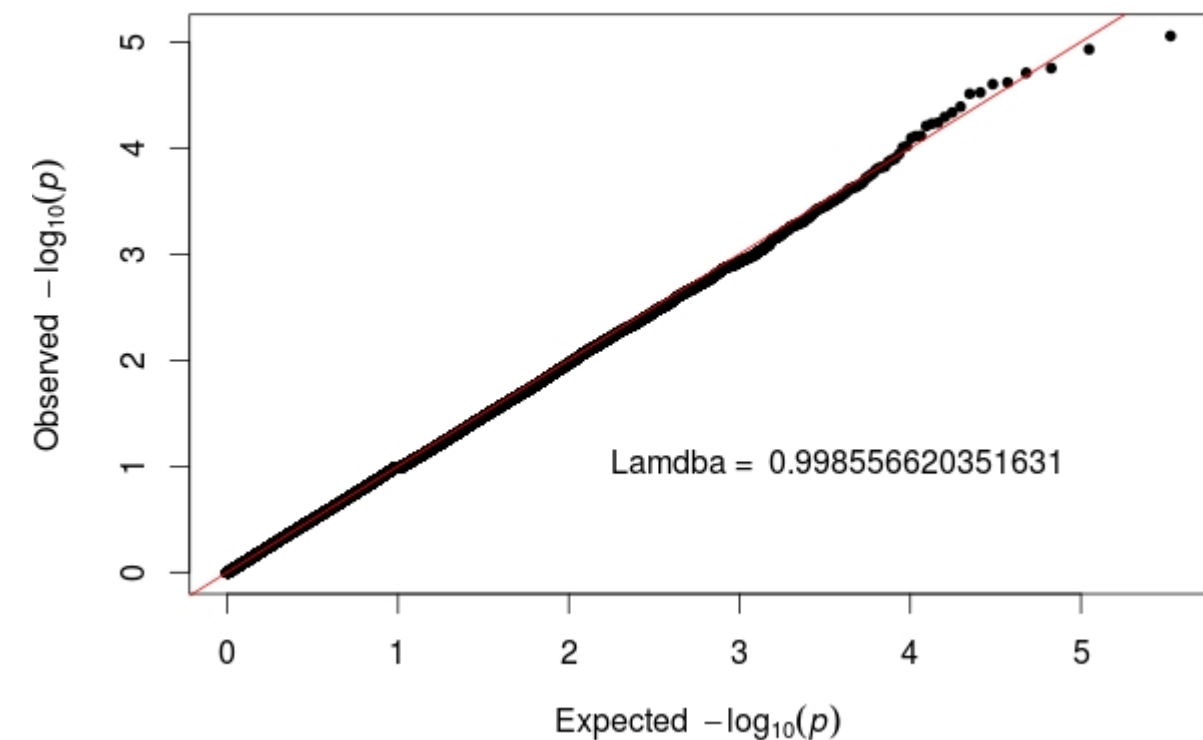
$$Y1 = B0 + B1*PC5$$
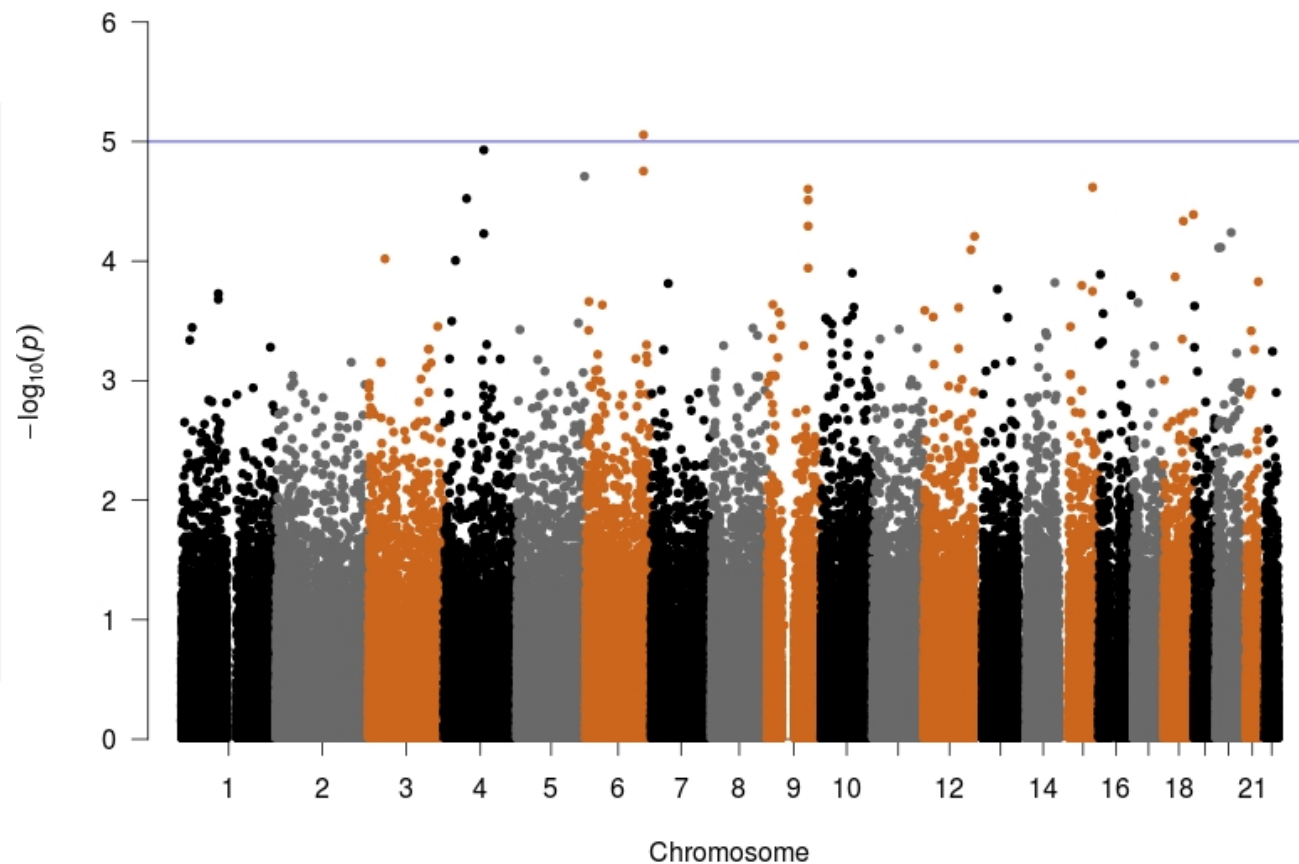$$Y2 = B0 + B1*Xa + B2*Xd + B3*PC5$$
*ANOVA likelihood ratio test to obtain p-value for each SNP.
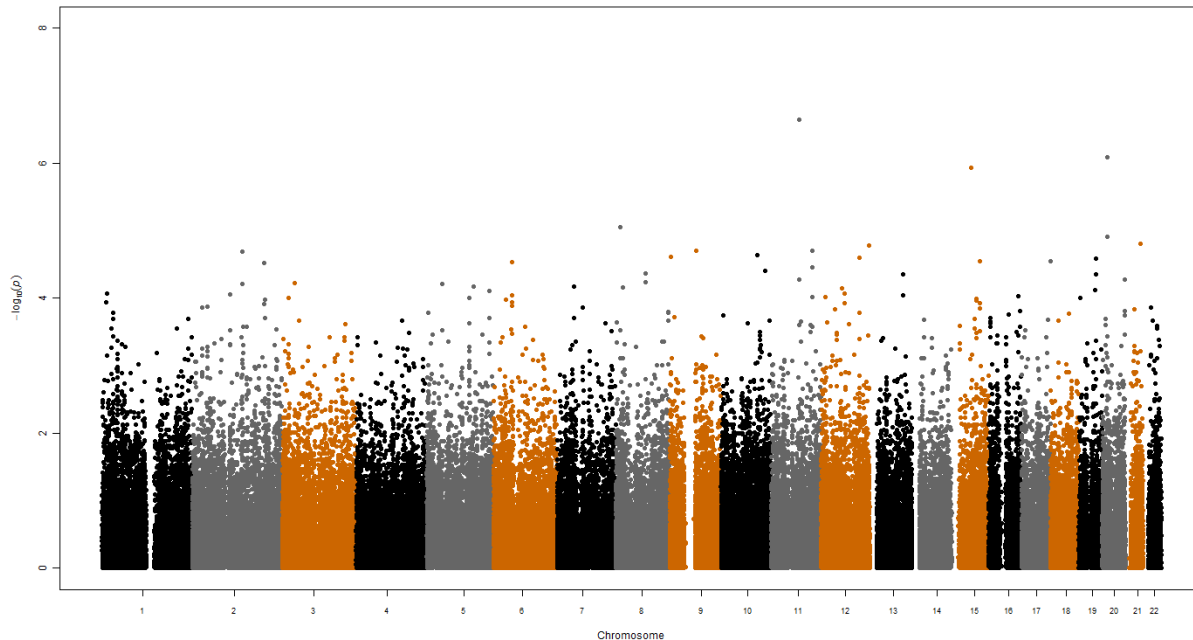
GWAS Results

Q-Q plot of GWAS p-values

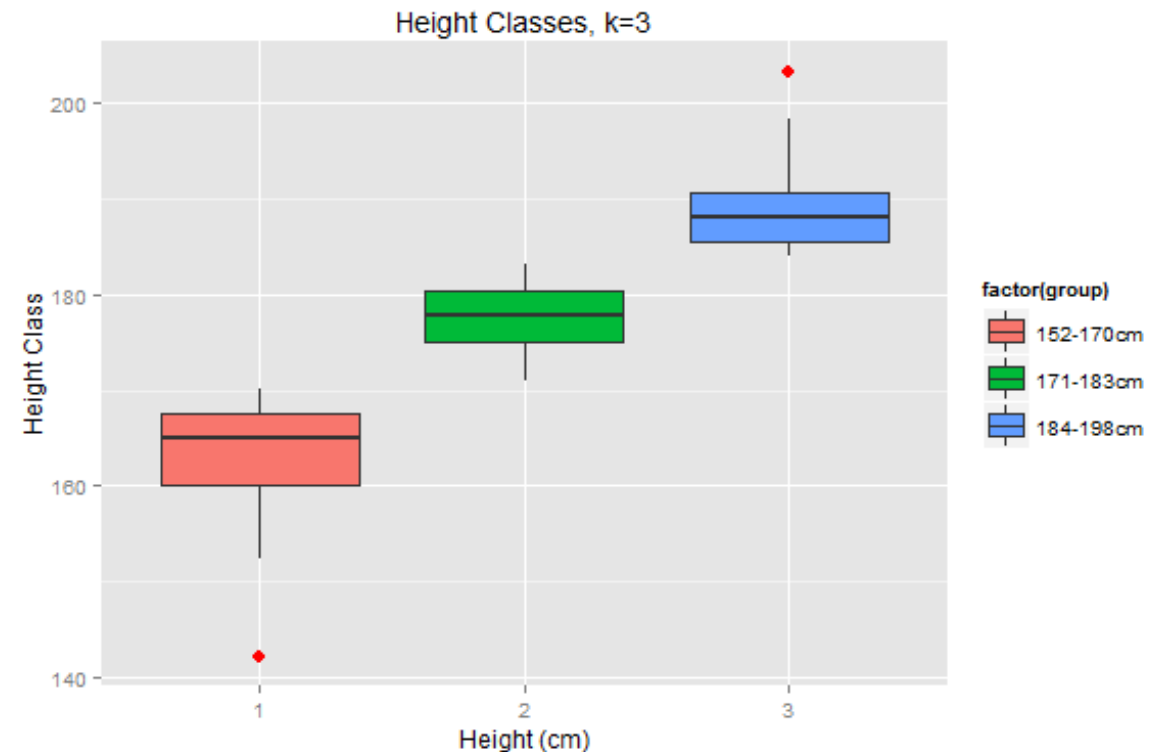Lamdba = 0.998556620351631

Manhattan Plot of Human Height

Objective 2: Learning the Signature of Adult Height
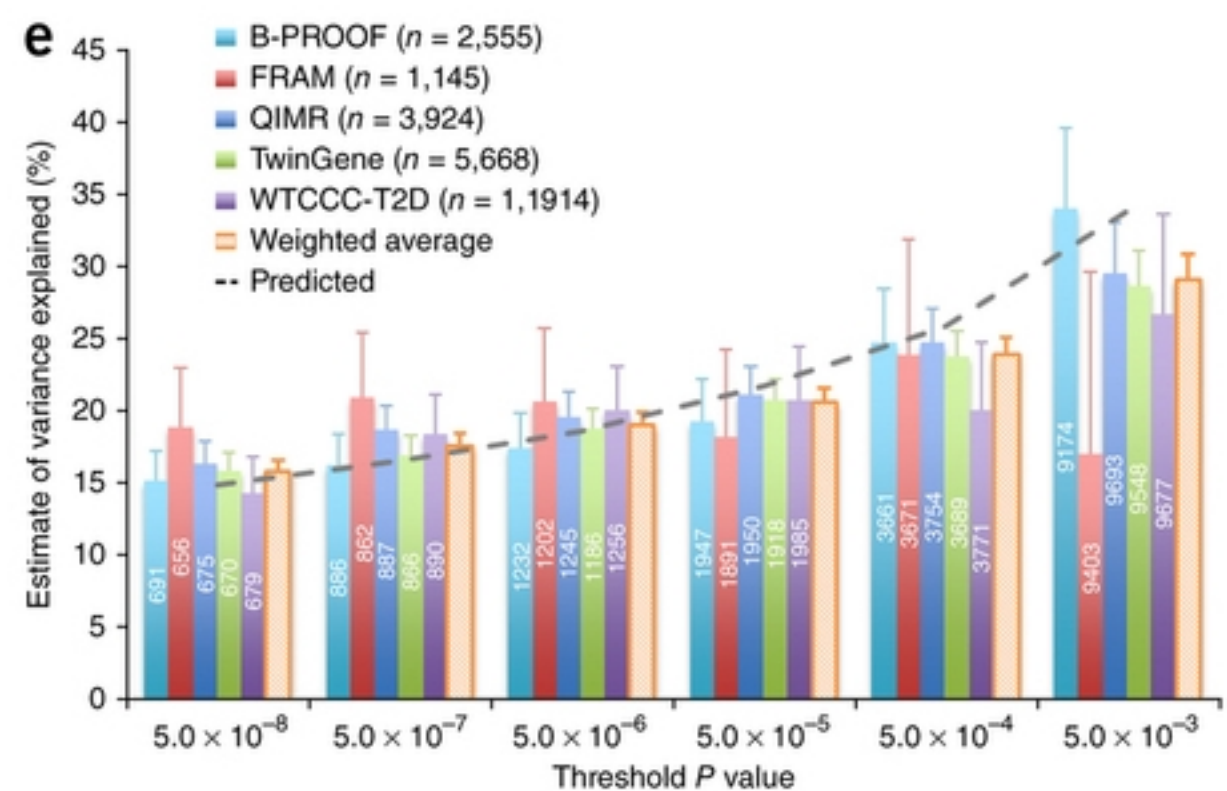
List of genome-significant SNPs

List of reported heights grouped into k classes

Height Classes, k=3

Height Class

factor(group)
152-170cm
171-183cm
184-198cm

Height (cm)

−log₁₀(p)
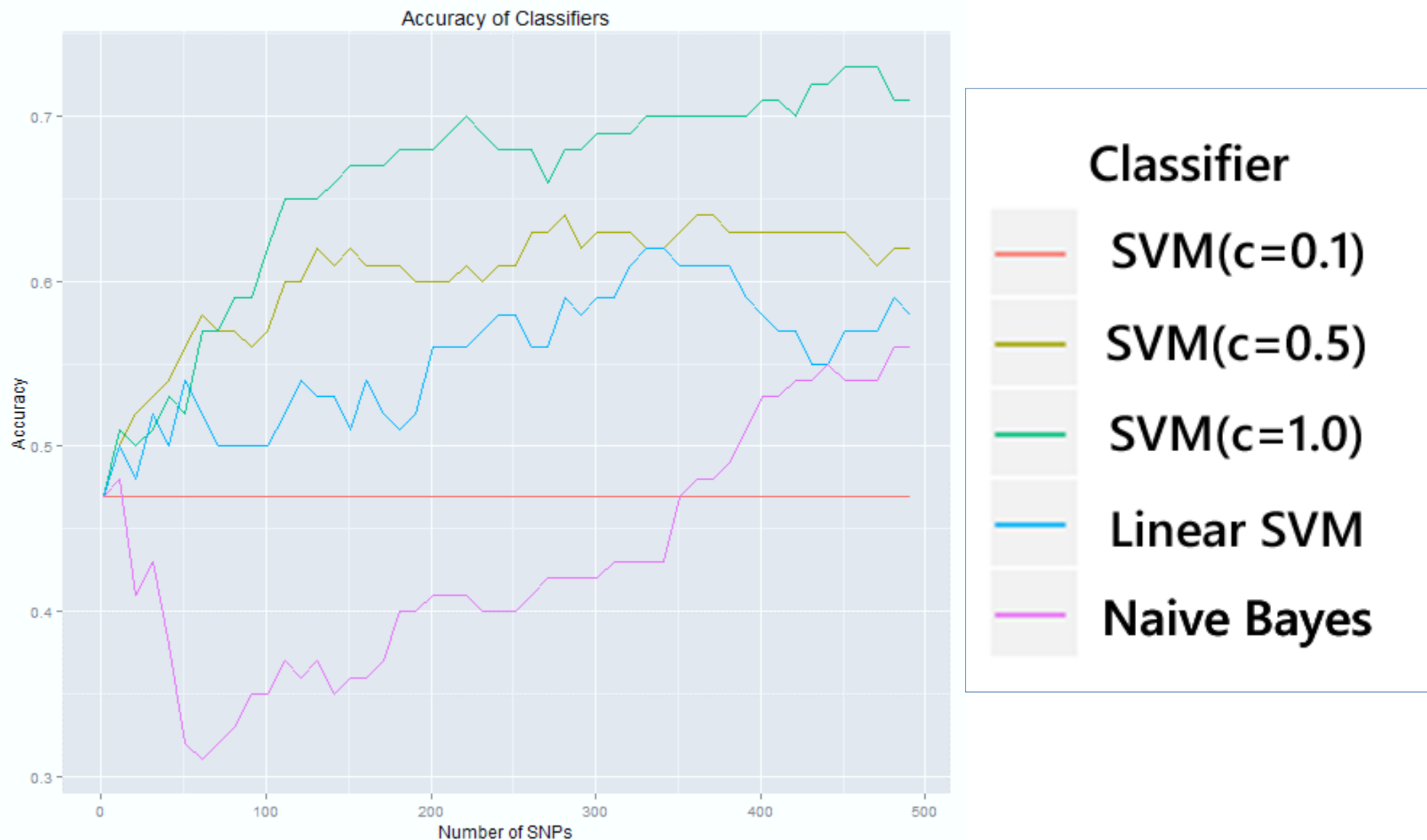
Chromosome

There is a better method for SNP selection... p-value prioritization

- Common Phenotype / Common Alleles
- Genome wide threshold too restrictive

(Wood et al. Nature 2014)
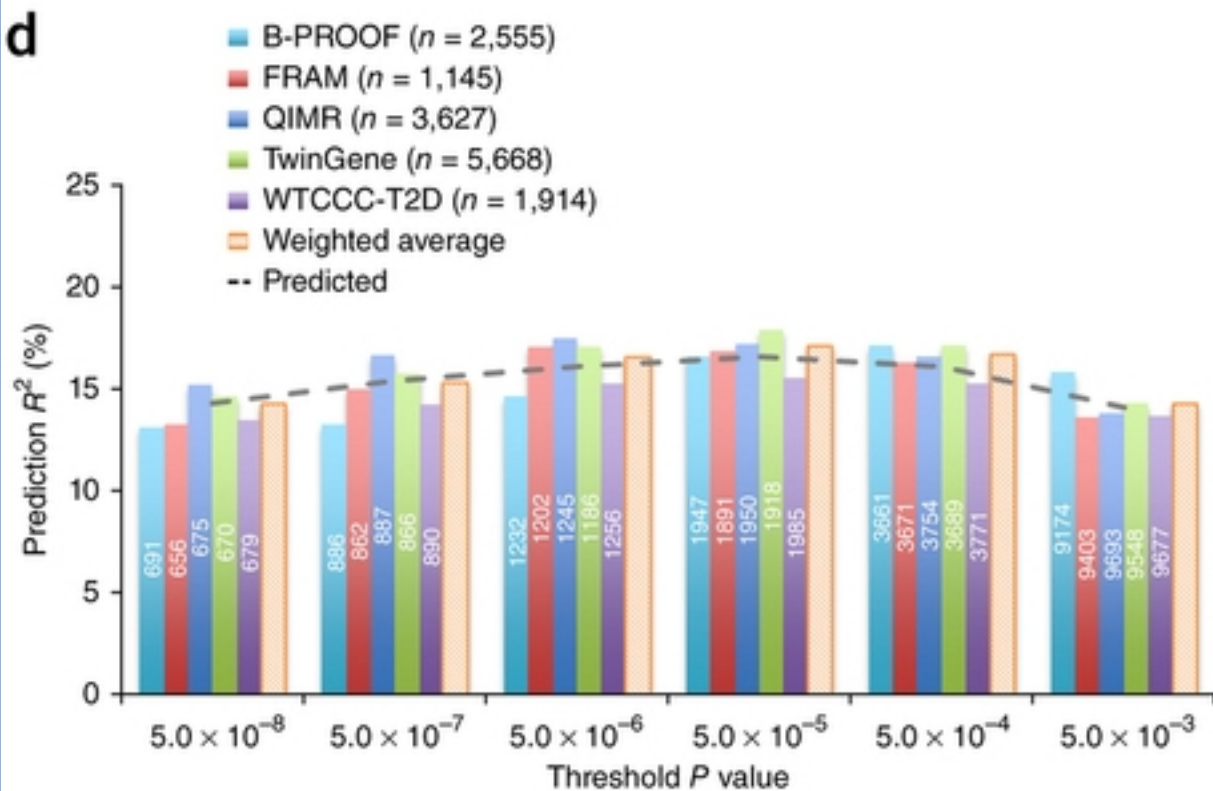
5-fold Cross-Validation SVM

SVM Results

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%

95th percentile = 191.2cm

90th Percentile = 187.5cm

10th Percentile = 161.5cm

5th Percentile = 157.9cm

Conclusions and Future Work

- Increasing the sample size of our GWAS will improve the power, significance of our list of prioritized SNPs, and the accuracy of our SVM classification.

- Since environmental effects such as nutrition account for 10-20% of height variability, we will create overlapping height classes populated with resampling of users.

Acknowledgements

- *OpenSNP*
- *Harvard Personal Genome Project*
- *[first page image] Credit: © jinga80 / Fotolia*
- *[height chart]*
  http://www.pediatricsconsultant360.com/content/two-girls-short-stature

Questions