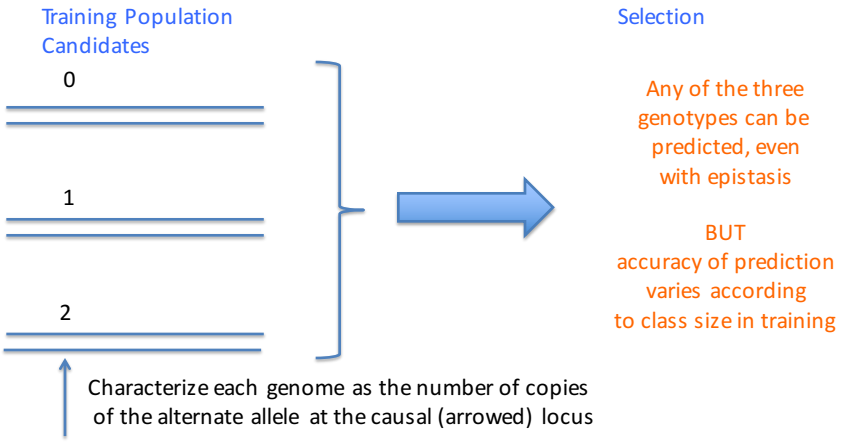
Characterizing the Upper Limits of Accuracy of Genomic Prediction for Individual Selection Candidates

Rohan Fernando
Hao Cheng, Emre Karaman, Xiaochen Sun
Dorian Garrick

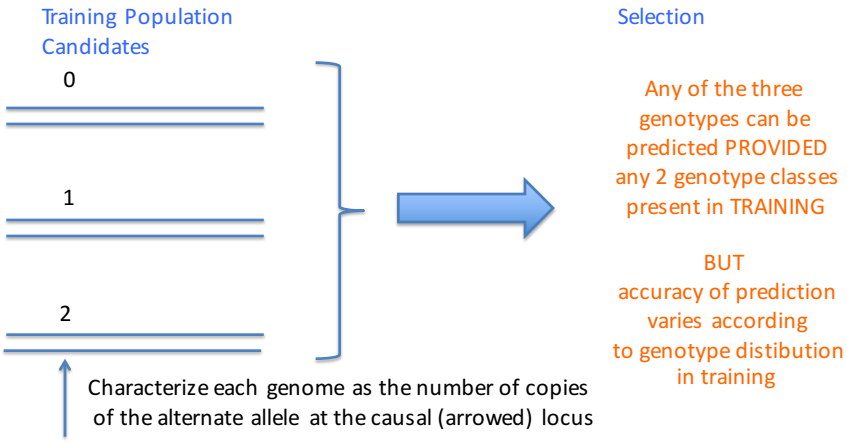
One QTL as Fixed Class Effect

Consider a single QTL/QTN affecting a complex trait



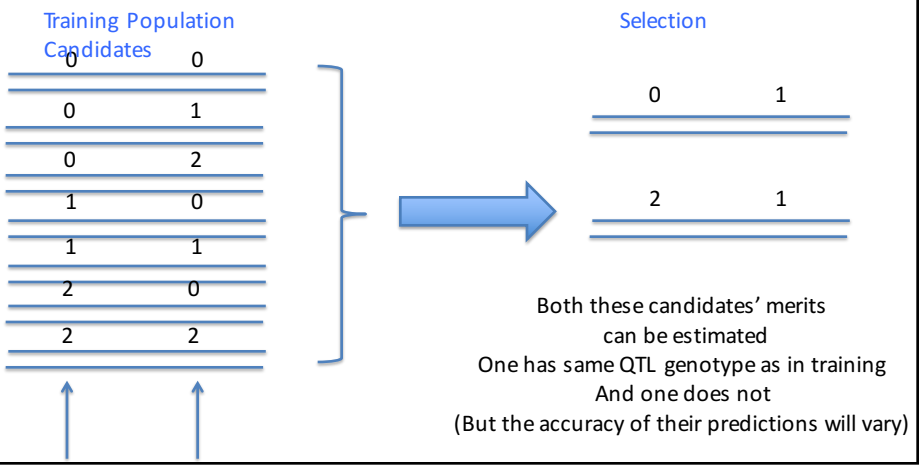
One QTL as Additive Fixed Effect

Consider a single QTL/QTN affecting a complex trait



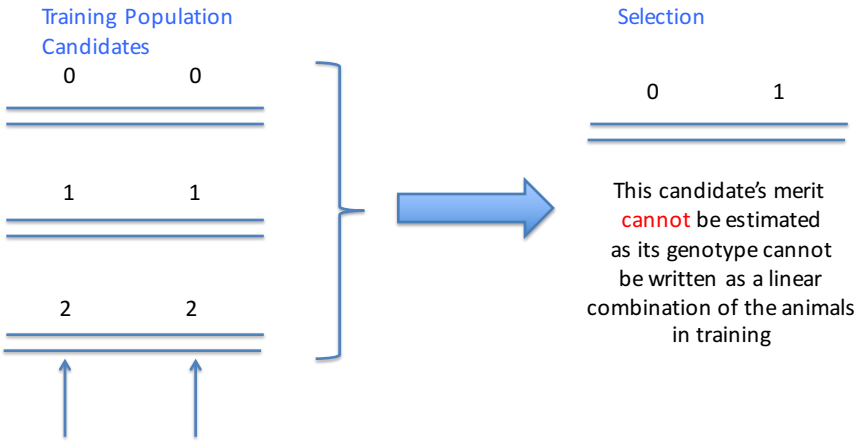
Two QTL as Additive Fixed Effects

Suppose there are 2 QTL/QTN affecting a complex trait



Two QTL as Additive Fixed Effects

Suppose there are 2 QTL/QTN affecting a complex trait



QTL as Additive Fixed Effects

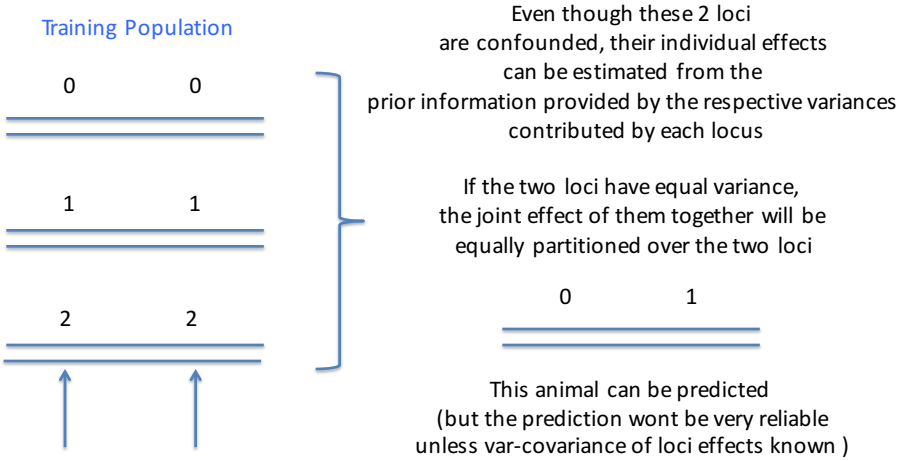
The same problem can occur in different ways……..

Causes of Linear Dependence

- First, because there are more gene variants (ie loci) than animals in training
- Second, because the genome consists of chromosomes and nearby loci will be linked
  - Relatively few haplotypes
- Third because our datasets comprise families so we may have a lot of data that simply represents "replicates" of similar data
- The best data would have all QTL x-classified

QTL as Additive Random Effects

Suppose there are 2 QTL/QTN affecting a complex trait



More Generally

- Suppose there are many QTL/QTN
  - Provided the covariate for any one locus CAN be written as a linear function of all the other covariates, its effect can be estimated as a fixed effect
  - If there are *p* loci and *n* animals with *p>n* then there cannot be more than *n* linearly independent loci (there may be less)
    - Even though all the loci will get estimated effects when treated as random effects
      - But those estimated effects depend upon variance components which we don't know so simply approximate
        - » eg all equal variance as in GBLUP

More Generally (contd)

- The apparent effects of a locus may represent the combined effect of more than one QTL
  - Any animal in validation whose genotypes represent a new combination of QTL not seen in training will not be well estimated
    - Even though they will get a prediction from a model that treats them as random
- We need to be able to characterize animals in validation in terms of the extent their genomes are estimable…..

Estimability

- Being estimable
  - Does not mean that the animals merit will be WELL estimated
    - Some or all of the locus effects my still be poorly estimated (ie have large prediction error variances)
- But being inestimable does indicate that the estimated merit will be spurious
- In real life, some fraction of the validation animals genome will be estimable and some fraction will not…we can characterize this…

Linear Regression

- Given $y = Xb + e$
- We predict the observation vector *y*, by partitioning it into
  - one part that can be explained by the "explanatory" variables (ie *X*)
  - and another (orthogonal) part that contains that part that *cannot be explained* by the explanatory variables

Linear Regression

$$y = Xb + e$$

$$\hat{b} = [X'X]^{-1}X'y$$

$$\hat{y} = X\hat{b} = X[X'X]^{-1}X'y$$

$$\hat{e} = y - \hat{y} = [I - X[X'X]^{-1}X']y$$

$$y = \hat{y} + \hat{e}$$

With these vectors being orthogonal
(so neither contains information about the other)

We now apply exactly this concept to the training and validation genotypes.....

Application to Genotypes

$$k = M'b + e$$

k is a vector of validation animal genotypes

$$\hat{b} = [MM']^{-1}Mk$$

M is a matrix of training population genotypes

$$\hat{k} = M'\hat{b} = M'[MM']^{-1}Mk$$

$$\hat{e} = k - \hat{k} = [I - M'[MM']^{-1}M]k$$

$$k = \hat{k} + \hat{e}$$

Genotypes in Vallidation that
CAN be explained from TRAINING

Genotypes in Validation that
CANNOT be explained from TRAINING

Ideal Prediction Candidate

$$k = \hat{k}$$

- This means that the validation animals whole genome genotype is a linear combination of the genotype combinations of animals in training
- If the QTL effects are well estimated and the QTN are genotyped, we should be able to get good accuracy of prediction

Worst-case Prediction Candidate

$$k = \hat{e}$$

- This means that the none of the validation animals whole genome genotype is a linear combination of the genotype combinations of animals in training
- We cannot predict this animal regardless of knowledge of the QTN effects
  - We will still get apparent predictions that may suggest some animals are good or even very good

More Common Outcome

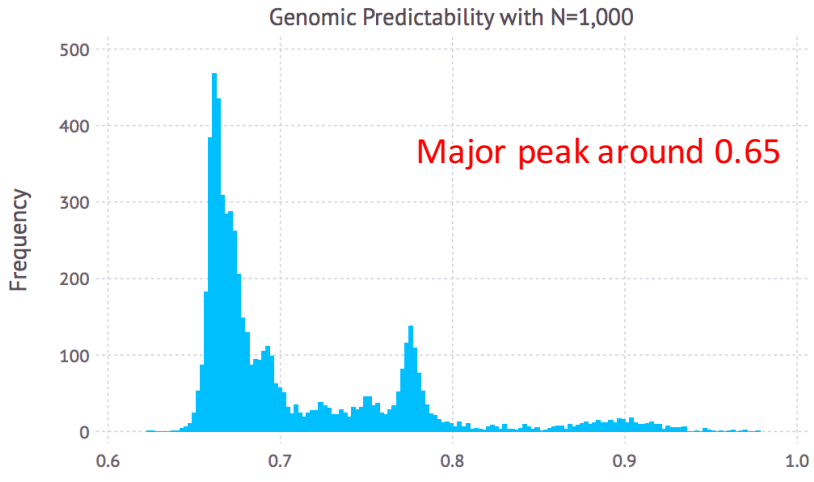
- Unless we have a sufficiently large training population, or the candidate has close relatives in training

$$k = \hat{k} + \hat{e}$$

- And the relative "sizes" of these vectors place a ceiling on predictive ability if QTN are distributed across the entire genome
- So we can quantify the upper limit for predictive ability for every selection candidate

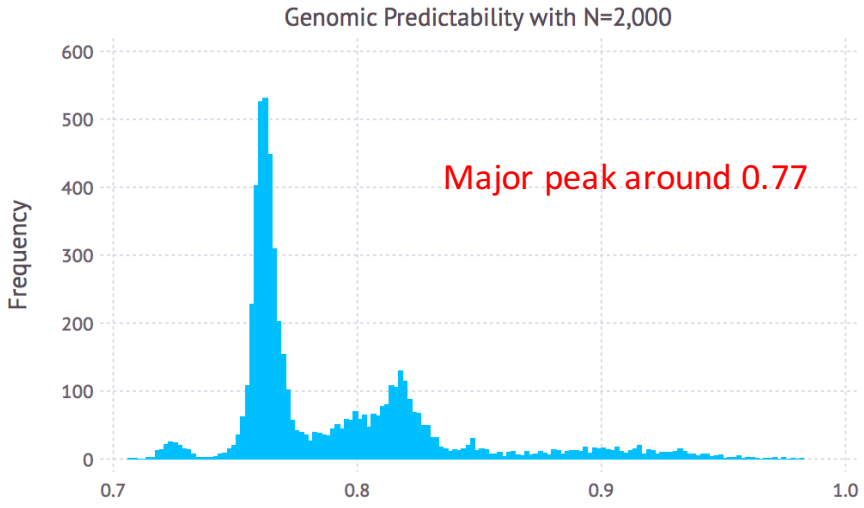
$$U = \hat{k}'\hat{k}/k'k$$

Application to Bovine 50k



Gelbvieh Upper bounds for genomic prediction from 1,000 animals with 50k with perfect LD etc

Application to Bovine 50k



Gelbvieh Upper bounds for genomic prediction from 2,000 animals with 50k with perfect LD etc

Additional Aspects

- We can further modify our approach to account for LD between QTL and marker loci
- We validated that approach using simulated phenotypes and the human 1,000 genomes project

Application to Human Genome Sequence

- Use actual 1,092 phased WGS data as founders
- Dropped down for 100 generations with 10,000 individuals per generation and a mutation rate of $1x10^{-8}$
- Only data from the last generation analysed
- Discarded loci with MAF<0.005
- Only used 0.1M of each of HSA1-HSA5
  - Whole genome was therefore 0.5M
    - Need to scale training population size by 60 to represent a 30M genome
- Only used 84 loci/cM and 1 in 60 was a QTN

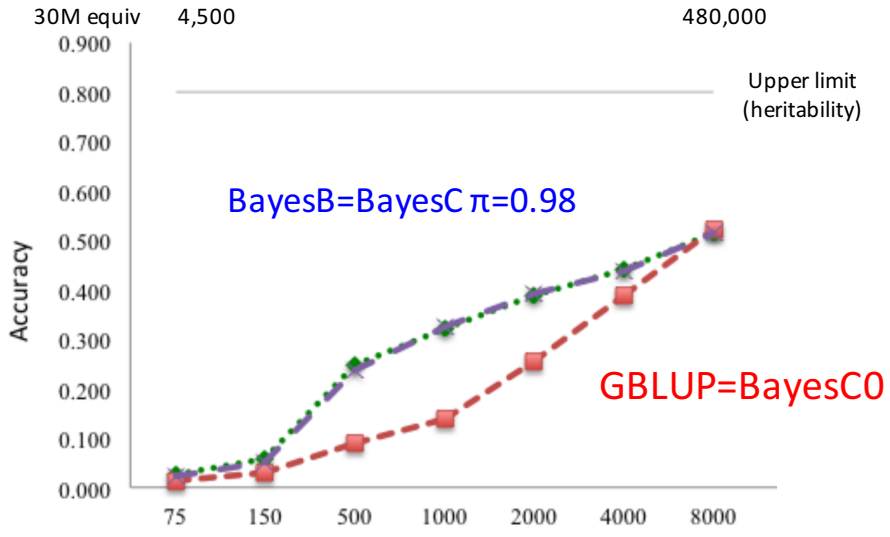
Simulated data from Human WGS

- Three scenarios for generating simulated phenotypes from an additive model and for choosing marker loci for genomic prediction
- S_Hi-Lo: Markers high MAF – QTL low MAF
- S_Hi-Rnd: Markers high MAF – QTL random
- S_R-R: Marker and QTL selected at random
- Heritability 0.8 (like human height)
- Every scenario replicated 10 times

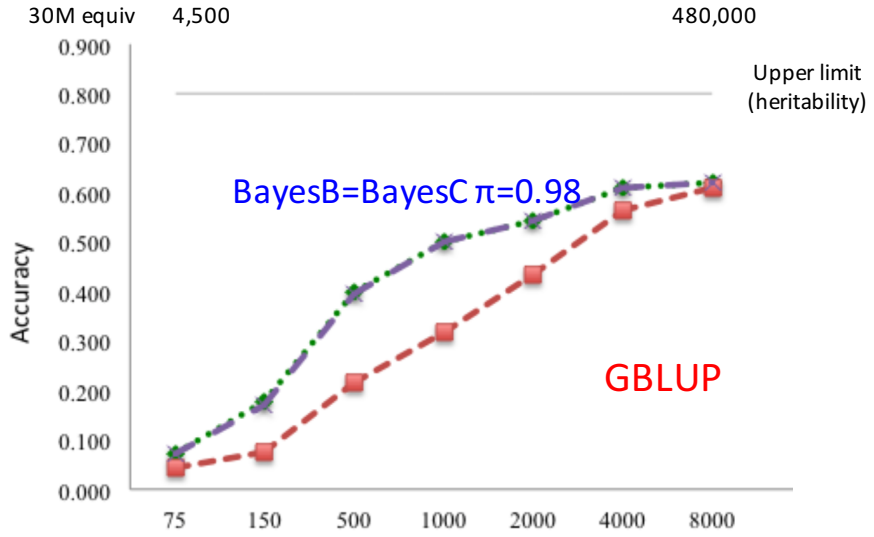
Results from Cross-Validation

- Among 10,000 individuals in the last generation
- Randomly chose 2,000 for validation
  - Validation is correlation with phenotype
- Randomly chose individuals for varying sizes of training data
  - Used 75 150 500 1,000 2,000 4,000 and 8,000
- For 30M genome these correspond to
  - 4,500 9,000 30,000 60,000 120,000 240,000 480,000

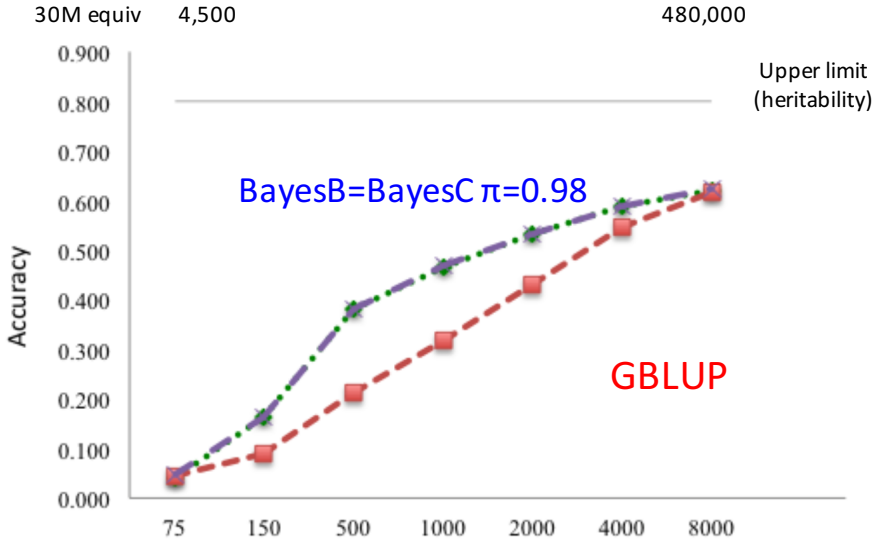
High MAF Markers Low MAF QTL



High MAF Markers Random QTL



Random Markers & QTL



Summary

- Likely Predictive Ability for a complex additive polygenic trait can be determined based on characteristics of the genomes of the training and validation populations
- Predictive Ability is (potentially) variable for selection candidates unless the training population is extremely large

Summary

- There is little difference between methods of prediction in small training populations (like 10,000 individuals for N_e=10,000 with h^2=0.8)
- There is little difference between methods of prediction in very large training populations like ½ million or more humans
- At intermediate sized training populations, mixture methods give a significant increase in predictive ability