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 1×10^{-8} using Xsim
- 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
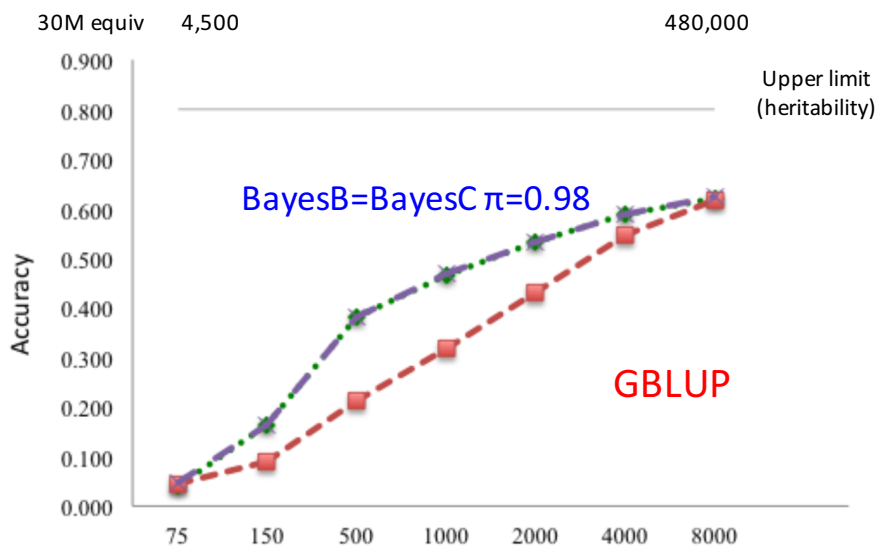   - 4,500 9,000 30,000 60,000 120,000 240,000 480,000

High MAF Markers Low MAF QTL
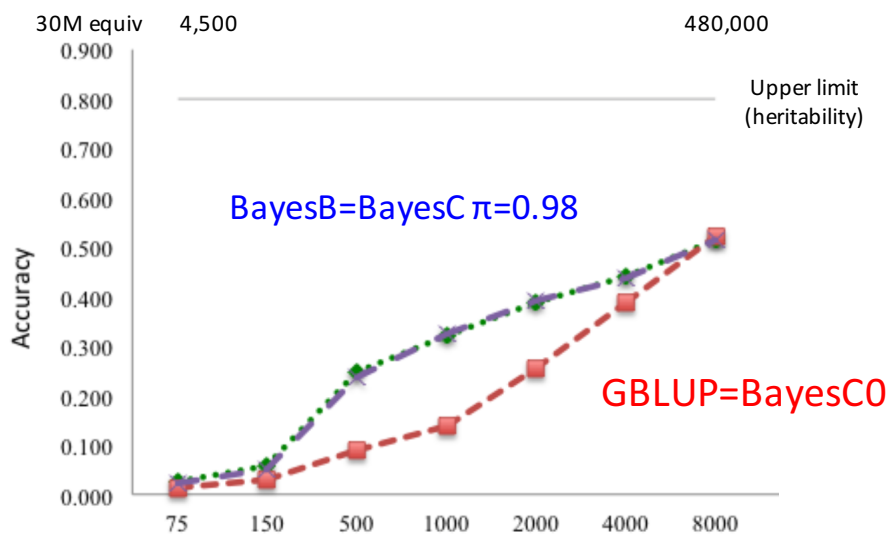
High MAF Markers Random QTL
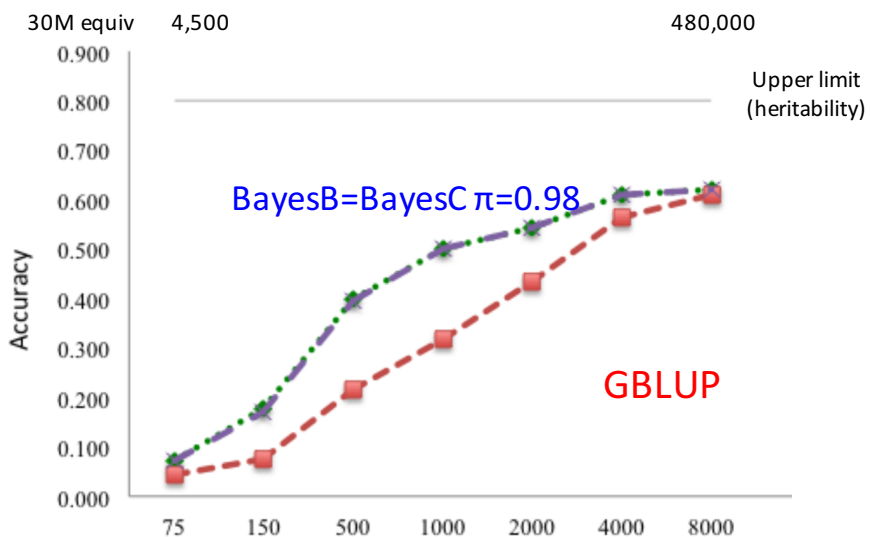
Random Markers & QTL

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