



UNIVERSITY OF
GEORGIA

College of Agricultural &
Environmental Sciences



Dimensionality of genomic information and its impact on GWA and variant selection

Sungbong Jang

S. Tsuruta, N.G. Leite, I. Misztal, D. Lourenco

University of Georgia

GWA in AB&G

- **GWA searches for major genes for traits of interest**
- **Sequence data is becoming available (> 30 M SNPs)**
 - GWA narrows down to the most important sequence variants
- **How many samples do we need?**
 - Explore the limited dimensionality of genomic information
 - Chromosome segments segregating independently in the population (M_e)

Ex) $\downarrow N_e \rightarrow \uparrow$ relationship among individuals and \uparrow LD $\rightarrow \downarrow M_e$



Dimensionality of genomic information

- **Independent chromosome segments (M_e)**

- N_e and M_e of livestock animals are small
- $N_e = 30 \sim 150$ and $M_e = 4000 \sim 15000$ (chicken, pig, dairy and beef cattle)
- $M_e = 4N_eL$ (Stam et al., 1980)

Pocrnic et al., (2016)

- **If limited M_e contains all additive information**

- 4000 ~ 15000 of M_e represents the dimensionality of genomic information
- Dimensionality of genomic information can be approximated by Eig of \mathbf{G}
 - Number of largest eigenvalues explaining certain % of \mathbf{G}

Misztal et al., (2016)



Objective

- **Explore the limited dimensionality of genomic information for GWA**
 - Establish the relationship between the number of animals representing the number of independent chromosome segments and the ability to detect causative variants



Data: Simulation

- QMSim (Sargolzaei et al., 2009)
- Cattle population
- Random mating
 - $N_e = 20$: 5 males / 15,000 females
 - $N_e = 200$: 50 males / 15,000 females
- 20 generations – used from 11-20
 - 75k genotyped animals from gen. 16-20



Step1: Genotype data scenarios

- **29 Chromosomes (total 23 Morgans)**
- **500k SNPs + QTNs (sequence data)**
- **Different numbers of QTNs: QTN = 200 and QTN = 2000**
- **Different heritabilities to mimic different amounts of data:**
 - $h^2 = 0.3, 0.9, 0.99$: low to very high reliability of EBVs



Step2: Construction of discovery sets

- Case of $N_e = 20$ and $N_e = 200$

	Ne = 20 QTN = 2000 $h^2 = 0.3$	Ne = 200 QTN = 2000 $h^2 = 0.3$
Eig50	80	530
Eig60	140	920
Eig70	220	1540
Eig80	400	2650
Eig90	890	5300
Eig95	1800	8800
Eig98	4100	15200
Eig99	7100	22000
All	30000	30000

- $\downarrow N_e \rightarrow \downarrow M_e$



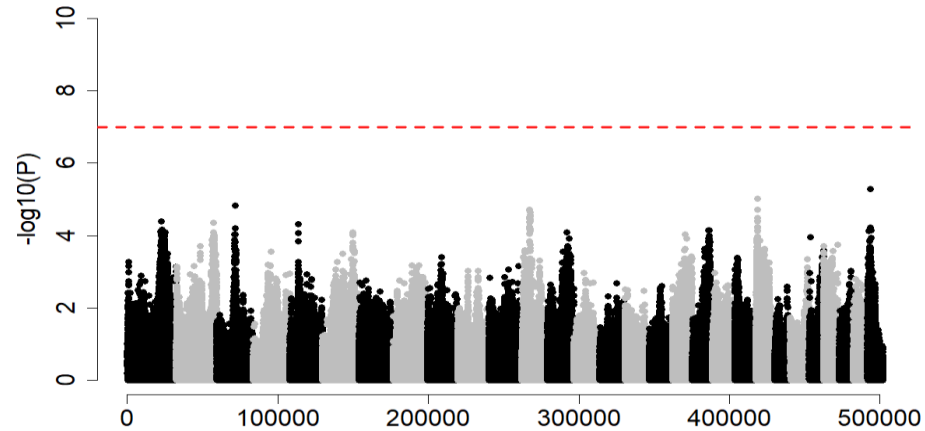
Step3: GWA

- **GEMMA (Zhou et al., 2012)**
- **Population structure was accounted for by G**
- **Total variance explained by significant QTNs**
 - $\%Var = 2pq\alpha^2 / \sigma_a^2$

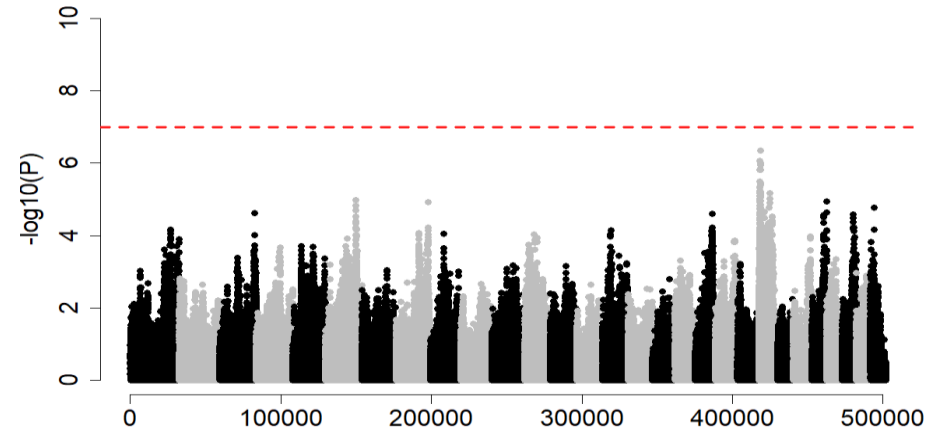


GWA results: $N_e = 20$ QTN = 2000 $h^2 = 0.3$

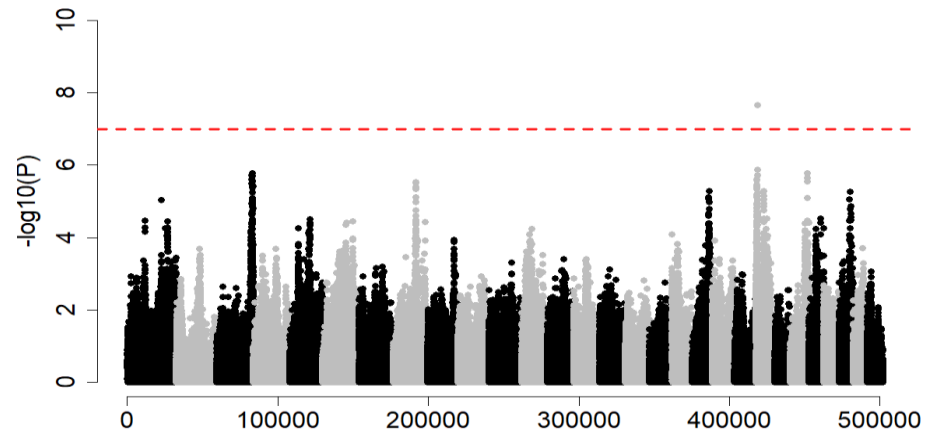
Eig95



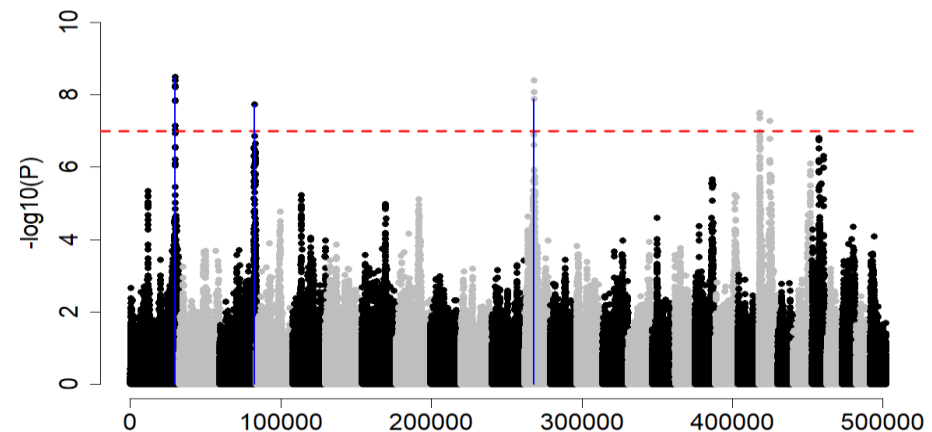
Eig98



Eig99

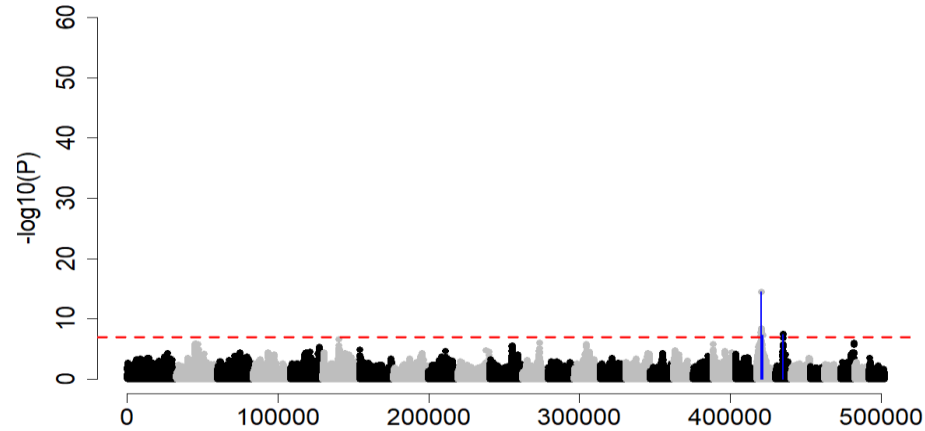


All

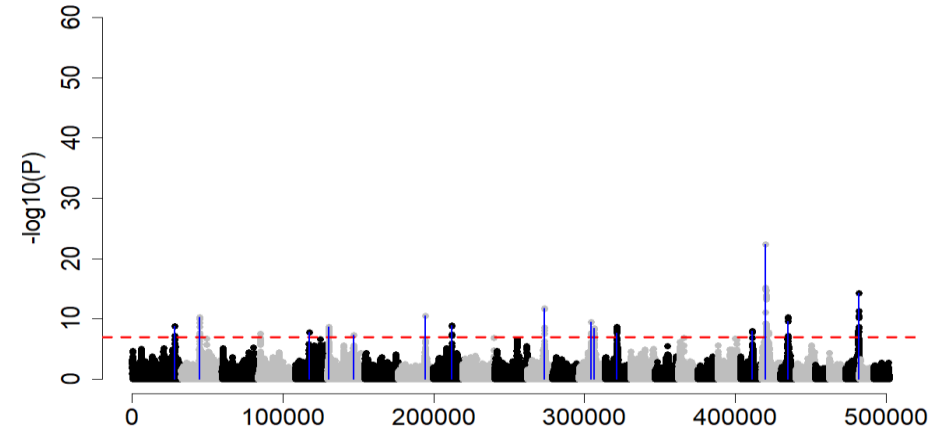


GWA results: $N_e = 20$ QTN = 2000 $h^2 = 0.99$

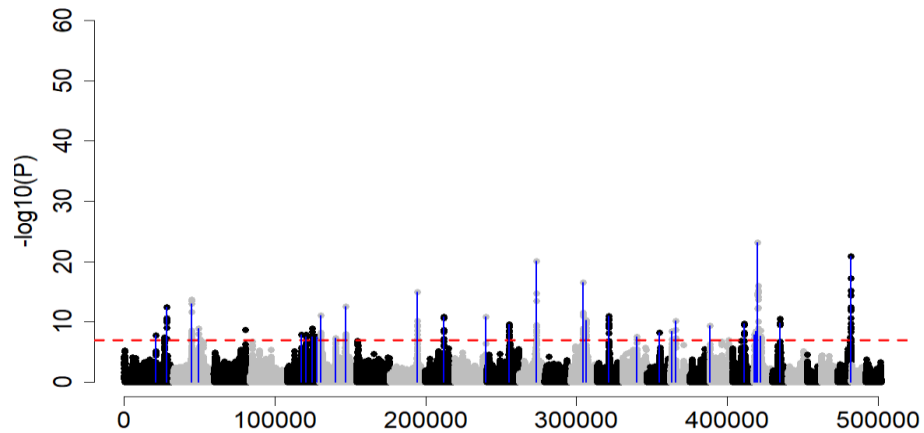
Eig95



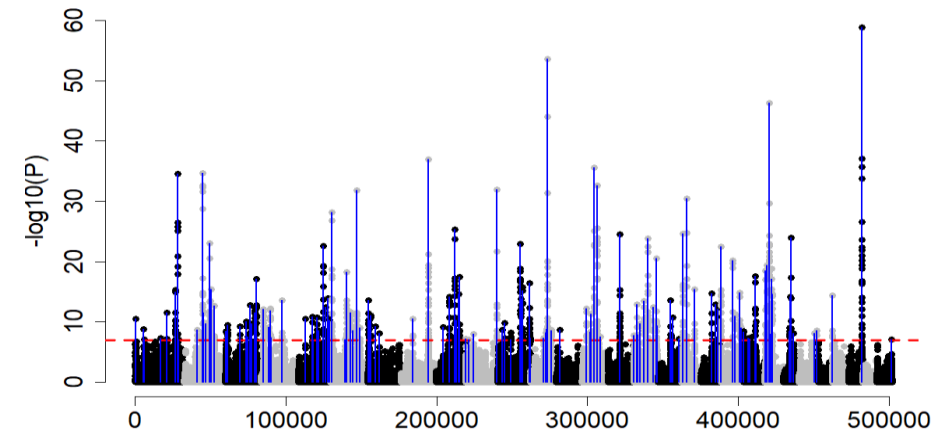
Eig98



Eig99

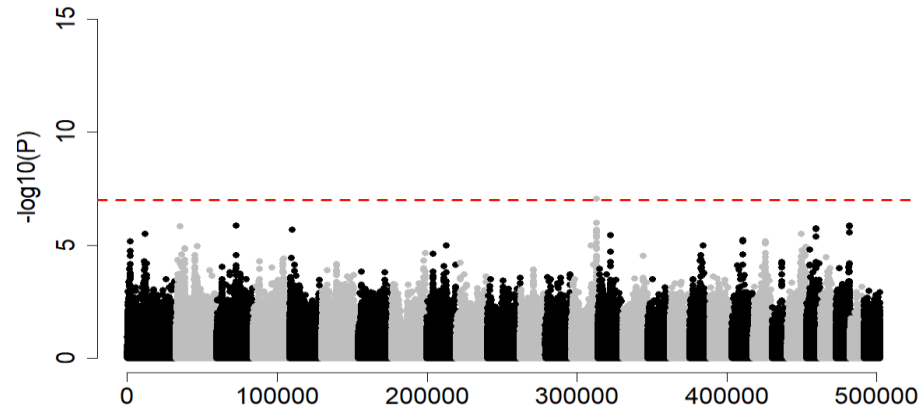


All

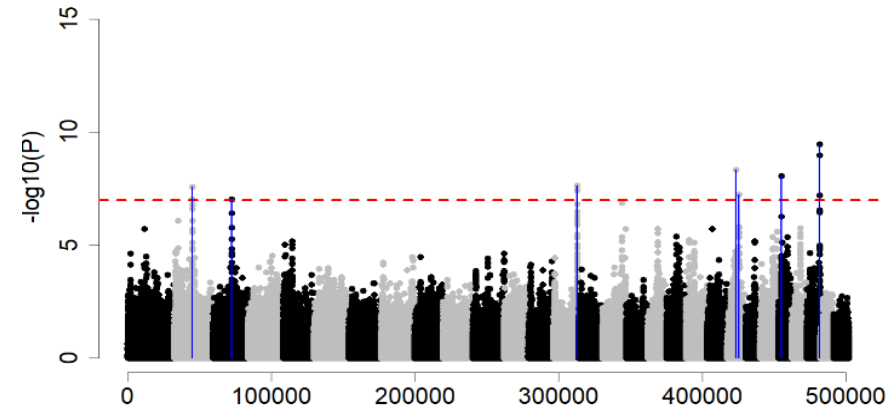


GWA results: $N_e = 200$ QTN = 2000 $h^2 = 0.3$

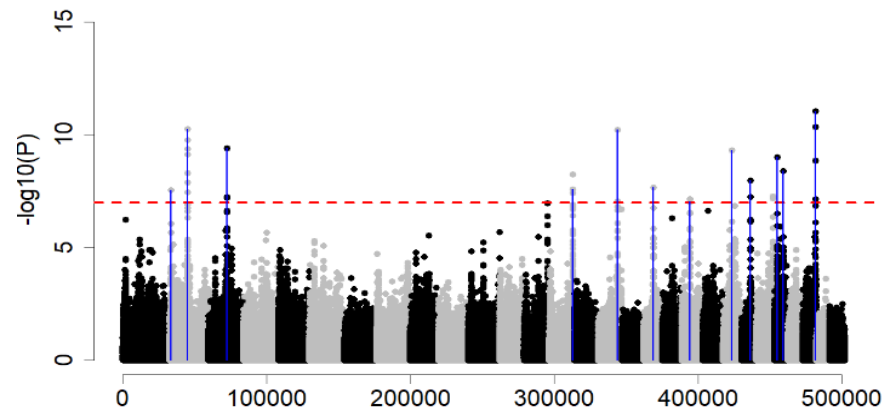
Eig95



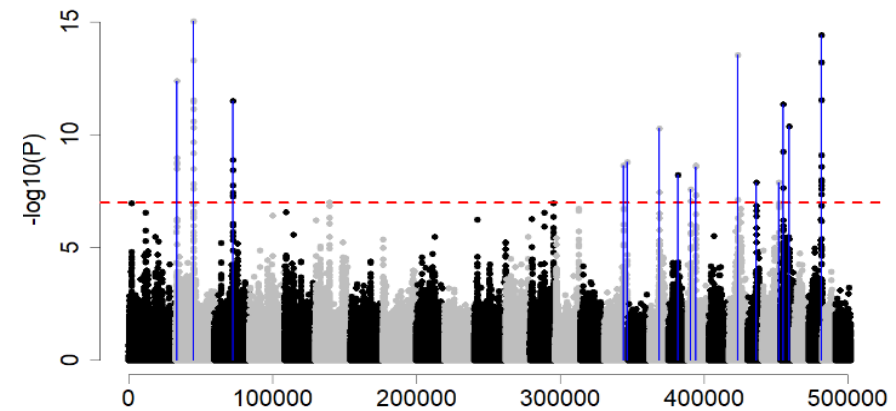
Eig98



Eig99

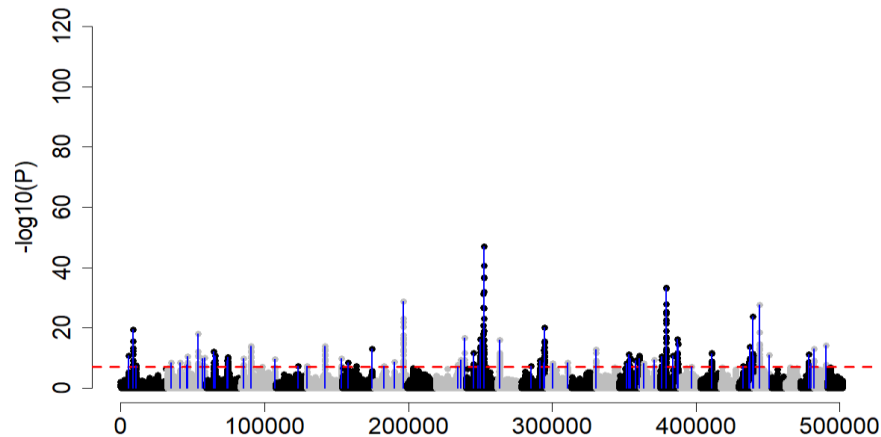


All

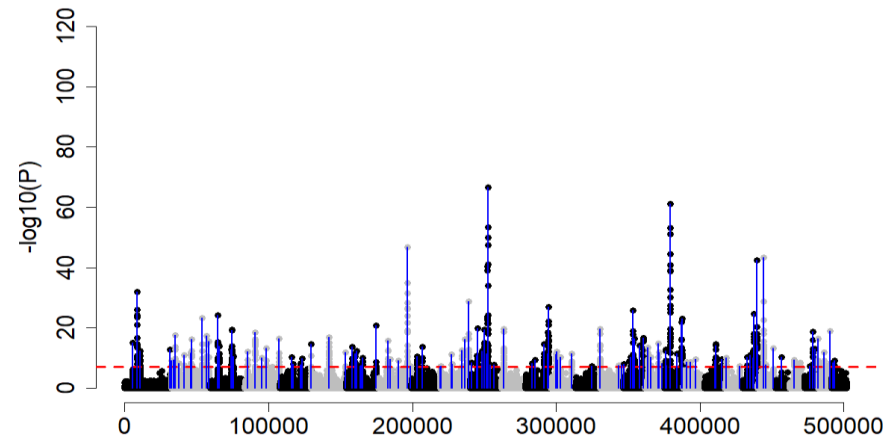


GWA results: $N_e = 200$ QTN = 2000 $h^2 = 0.99$

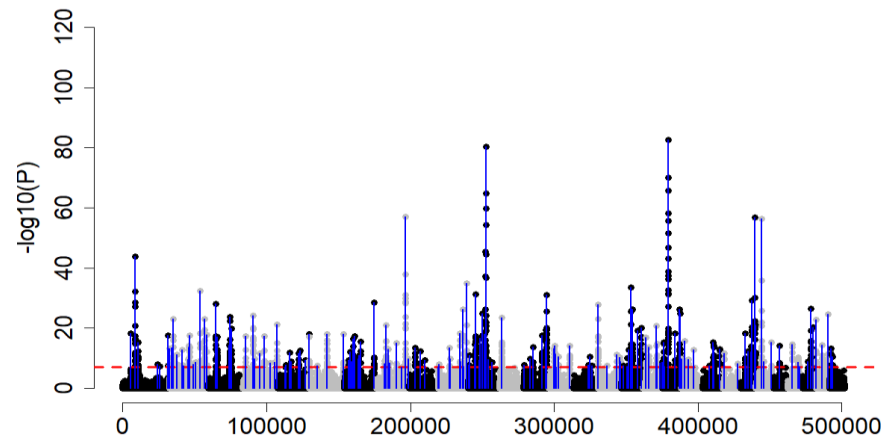
Eig95



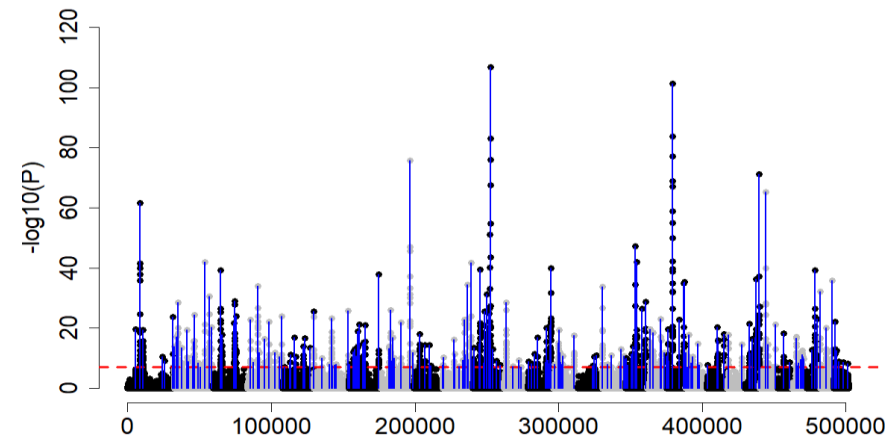
Eig98



Eig99



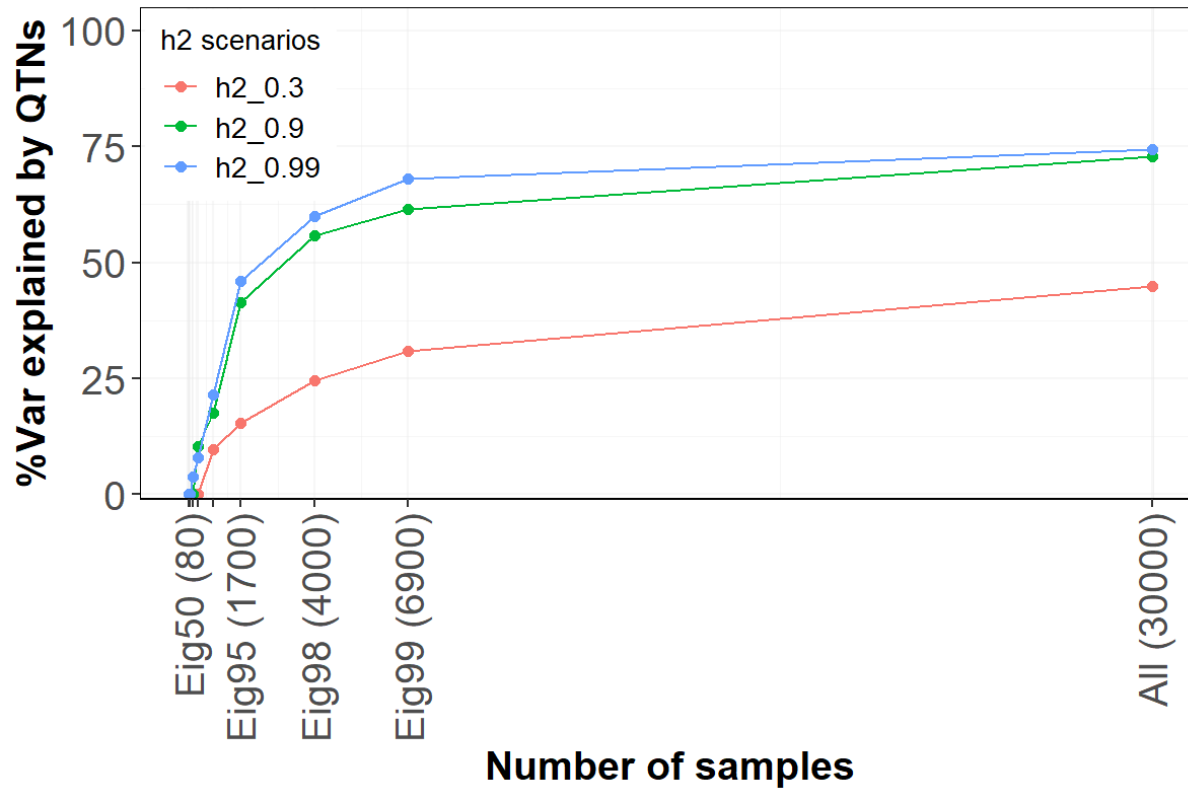
All



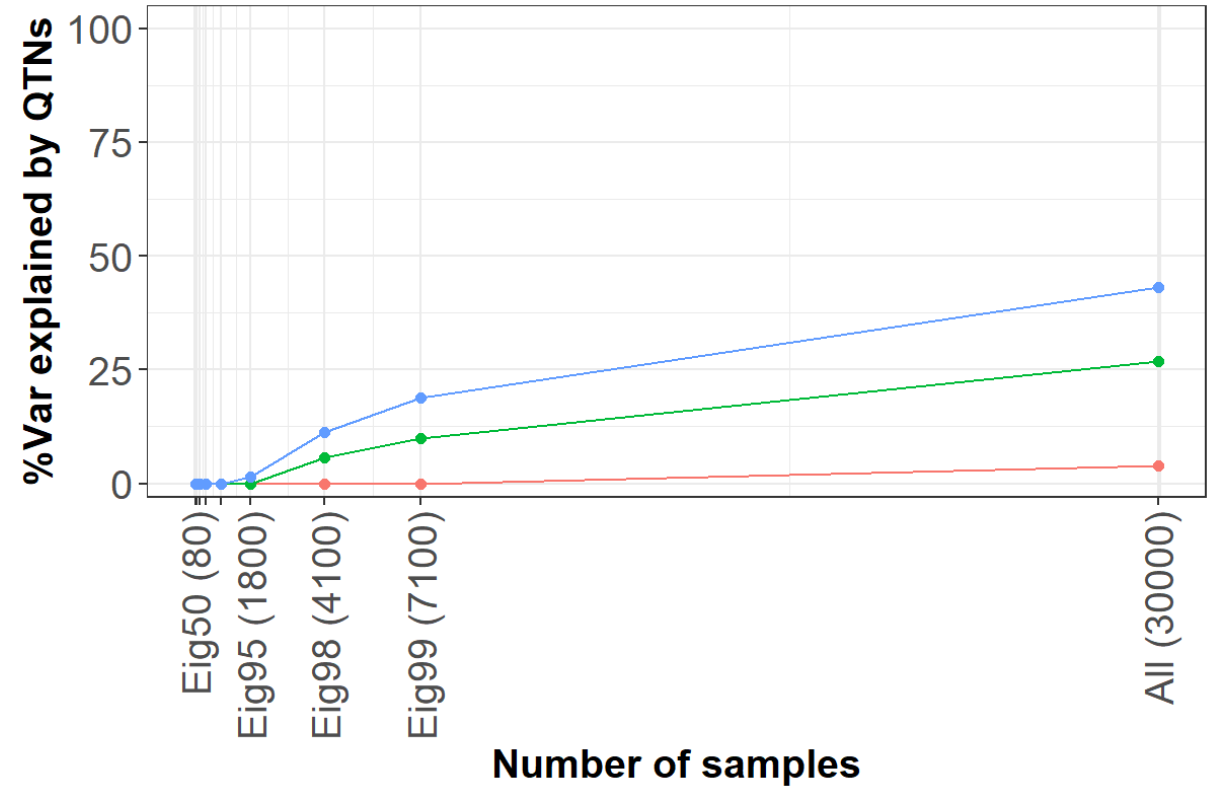
Total variance explained by QTNs

- $N_e = 20$

QTN = 200



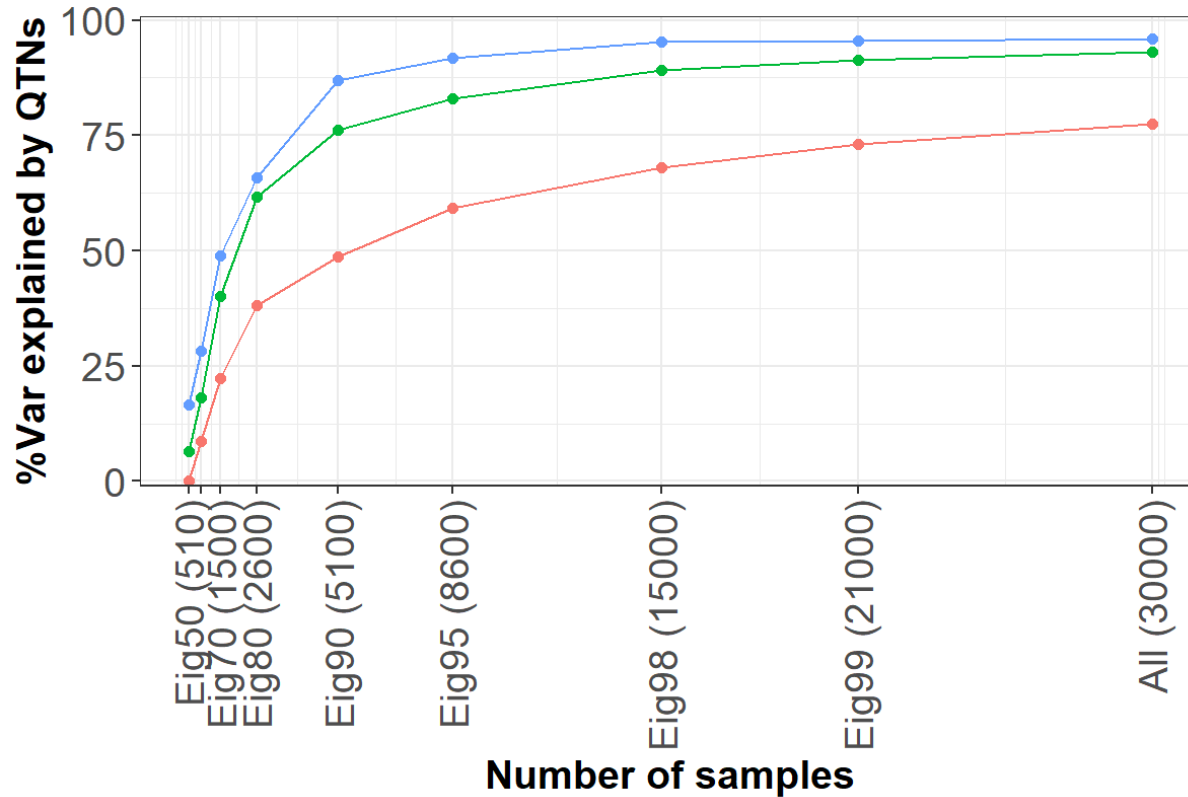
QTN = 2000



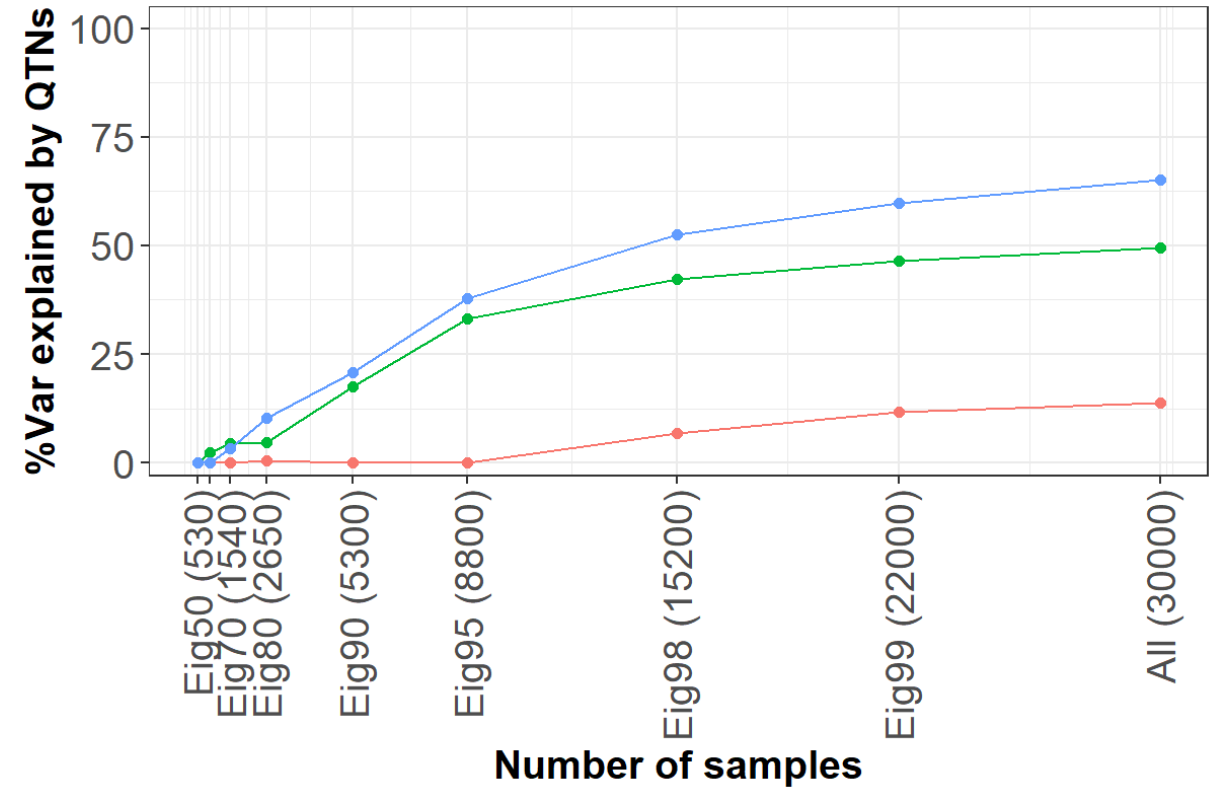
Total variance explained by QTNs

- $N_e = 200$

QTN = 200



QTN = 2000



Conclusions

- **The suitable sample size depends on the N_e and the number of QTNs**
 - Smaller populations require more data to capture causative variants
 - Larger populations: Eigen98 – Eigen99 could capture most informative QTNs
 - More polygenic trait requires more data to identify causative variants
- **Dimensionality of genomic information allows to approximate the suitable sample size for GWA**
- **In progress:**
 - Deriving equations to relate sample size and amount of data with N_e and M_e
 - Testing the impact of incorporating selected variants for GP



Thanks 😊

jsbng@uga.edu