UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Dimensionality of genomic information and its impact on GWA and variant selection: A simulation study

Sungbong Jang

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

University of Georgia

2021 ASAS-CSAS-SSASAS
07/16/2021

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 (Me)

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

Dimensionality of genomic information

- **Independent chromosome segments ($\mathbf{M_e}$)**
  - N_e and M_e of livestock animals are small (Pocrnic et al., (2016))
  - N_e = 30 ~ 150 and M_e = 4000 ~ 15000 (chicken, pig, dairy and beef cattle)
  - $M_e = 4N_eL$ (Stam et al., 1980)

- **If limited $\mathbf{M_e}$ contains all additive information** (Misztal et al., (2016))
  - 4000 ~ 15000 of M_e represents the dimensionality of genomic information
  - How to approximate the dimensionality of genomic information?
  - SVD of **Z** => $\mathbf{Z = U\Delta V}$ ($\mathbf{U'U = I}$ and $\mathbf{V'V = I}$), where $\mathbf{\Delta}$ = diagonal of singular values
  - Eigen of **G** => var(u) = var($\mathbf{U\Delta V}a$) ~ $\mathbf{U\Delta\Delta' U'}$, where $\mathbf{\Delta\Delta'}$ = diagonal of eigenvalues
  - Rank of **G** $\leq$ min(N_{snp}, N_{ind}, M_e), limited dimensionality

Objectives

- **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

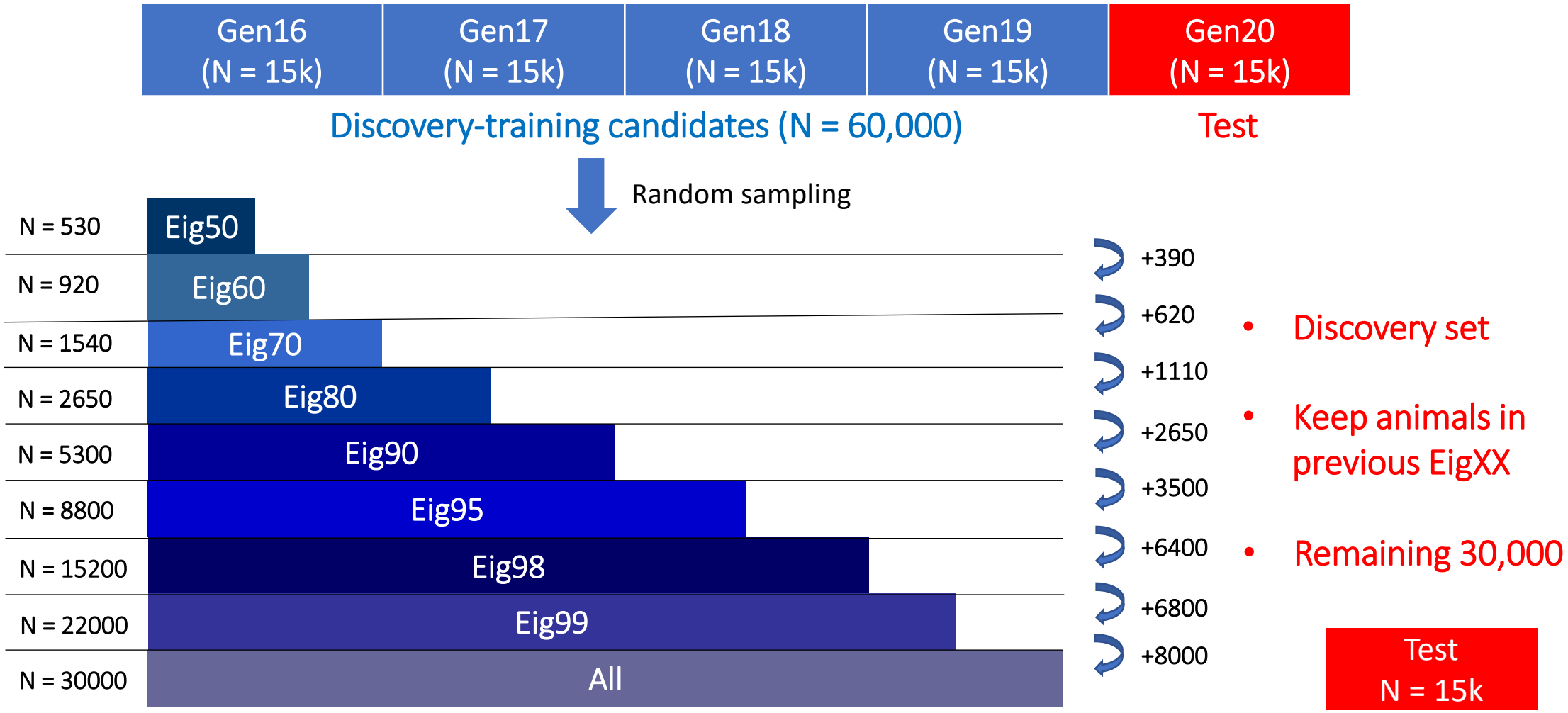
- QMSim (Sargolzaei et al., 2009)
- Cattle population
- Random mating
  - Ne = 20: 5 males / 15,000 females
  - Ne = 200: 50 males / 15,000 females
- 20 generations – used from 11-20
  - 75k genotyped animals from gen. 16-20

- 29 Chromosomes (Total 23 Morgans)
- 500k SNPs
- 200, 2000 QTNs
- Heritability = 0.3, 0.9, 0.99

Step1: Genotype data scenarios

- **withQTN: 500k SNPs + QTNs (sequence data)**
- **Different Ne: Ne = 20 and Ne = 200**
- **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 accuracy of EBVs

Step2: Number of animals for discovery/training



Step2: Number of animals for discovery/training

- **Case of Ne = 20 and Ne = 200**

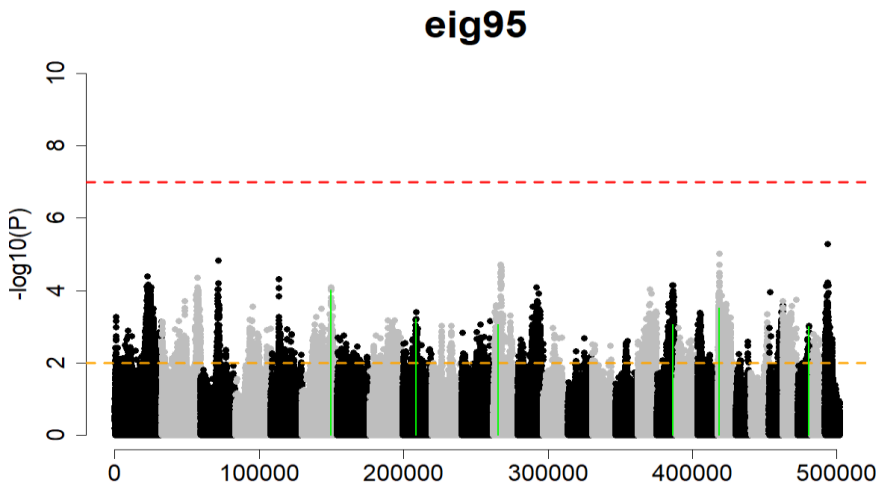
	Ne = 20 QTN = 2000 h^2 = 0.3	Ne = 200 QTN = 2000 h^2 = 0.3
Eig50	80	530
Eig60	140	920
Eig70	220	1540
Eig80	400	2650
Eig90	890	5300
Eig95	1800	8800
Eig98	4100	15200
Eig99	7100	22000
All	30000	30000

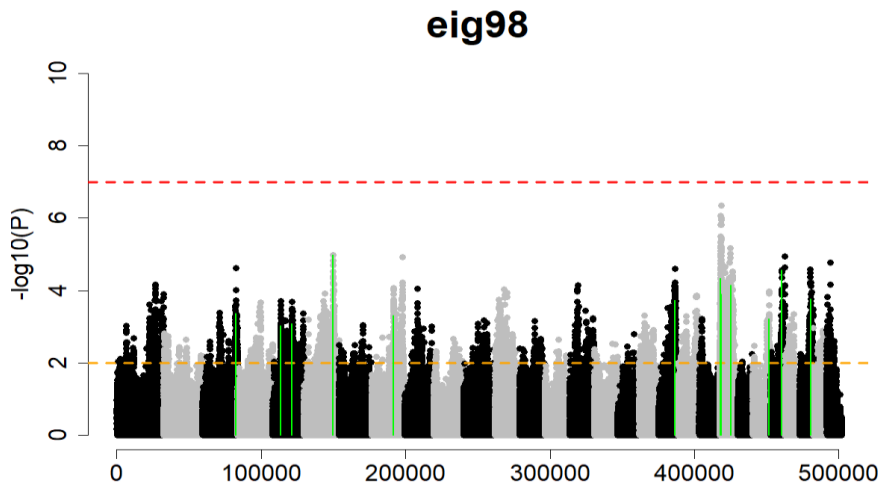
- $\downarrow \mathbf{N_e} \rightarrow \downarrow \mathbf{M_e}$

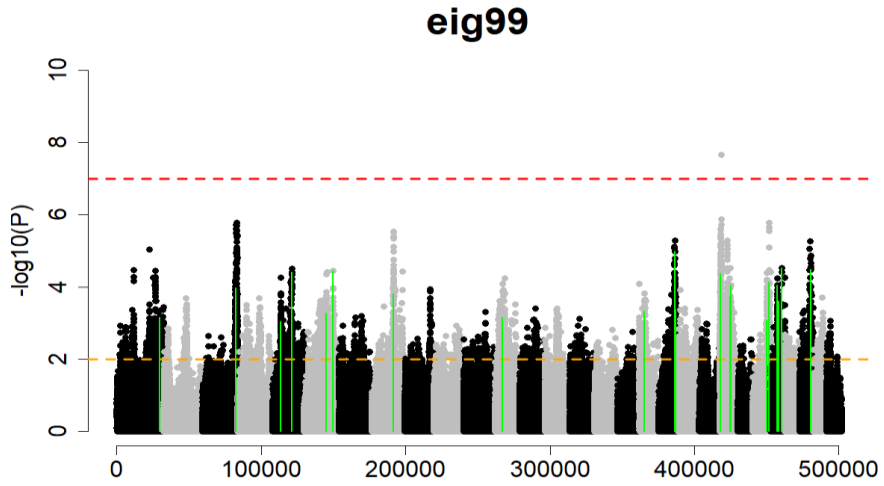
Step3: GWA

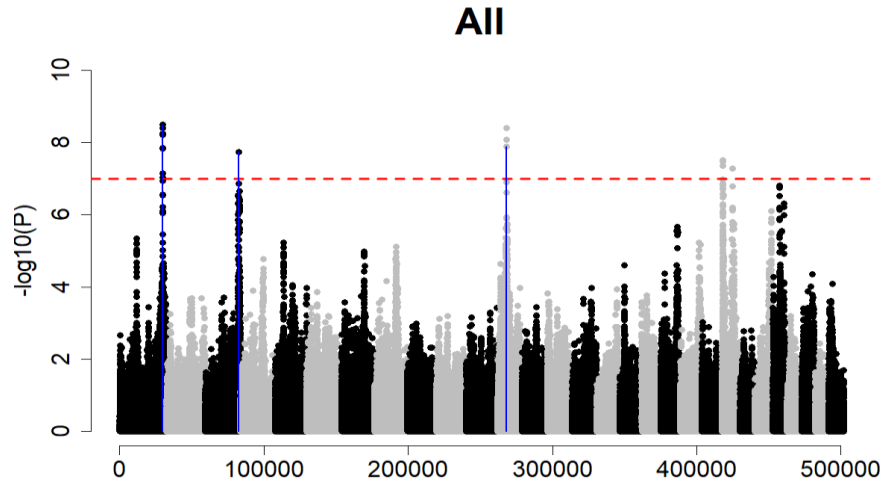
- **GEMMA (Zhou et al., 2012)**
- **Population structure was accounted for by G**
- **Total variance explained by significant QTNs**
  - $\%\mathrm{Var} = 2\mathrm{pq}(\beta)^2/\sigma_a^2$

GWA results: Ne = 20 QTN = 2000 h^2 = 0.3

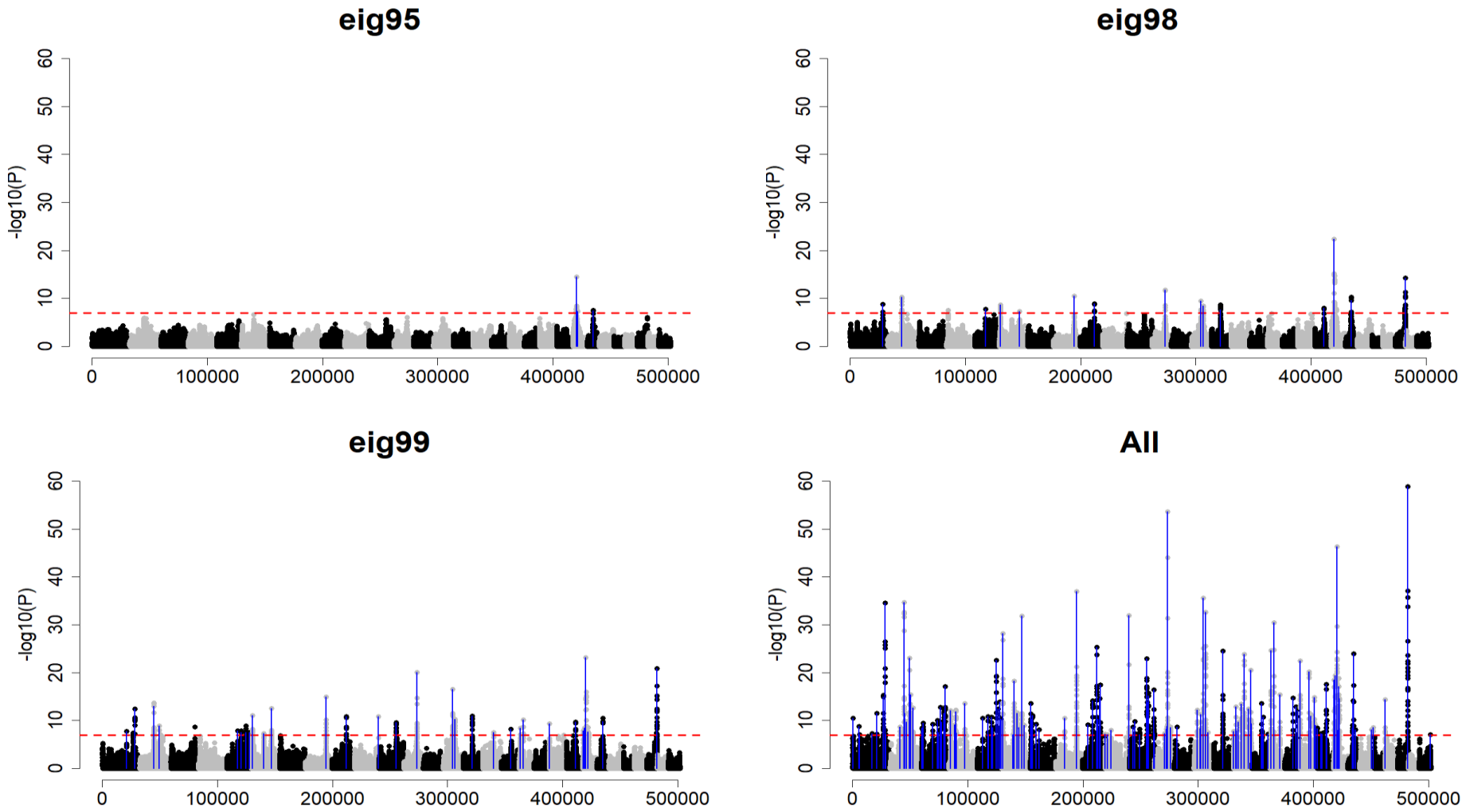




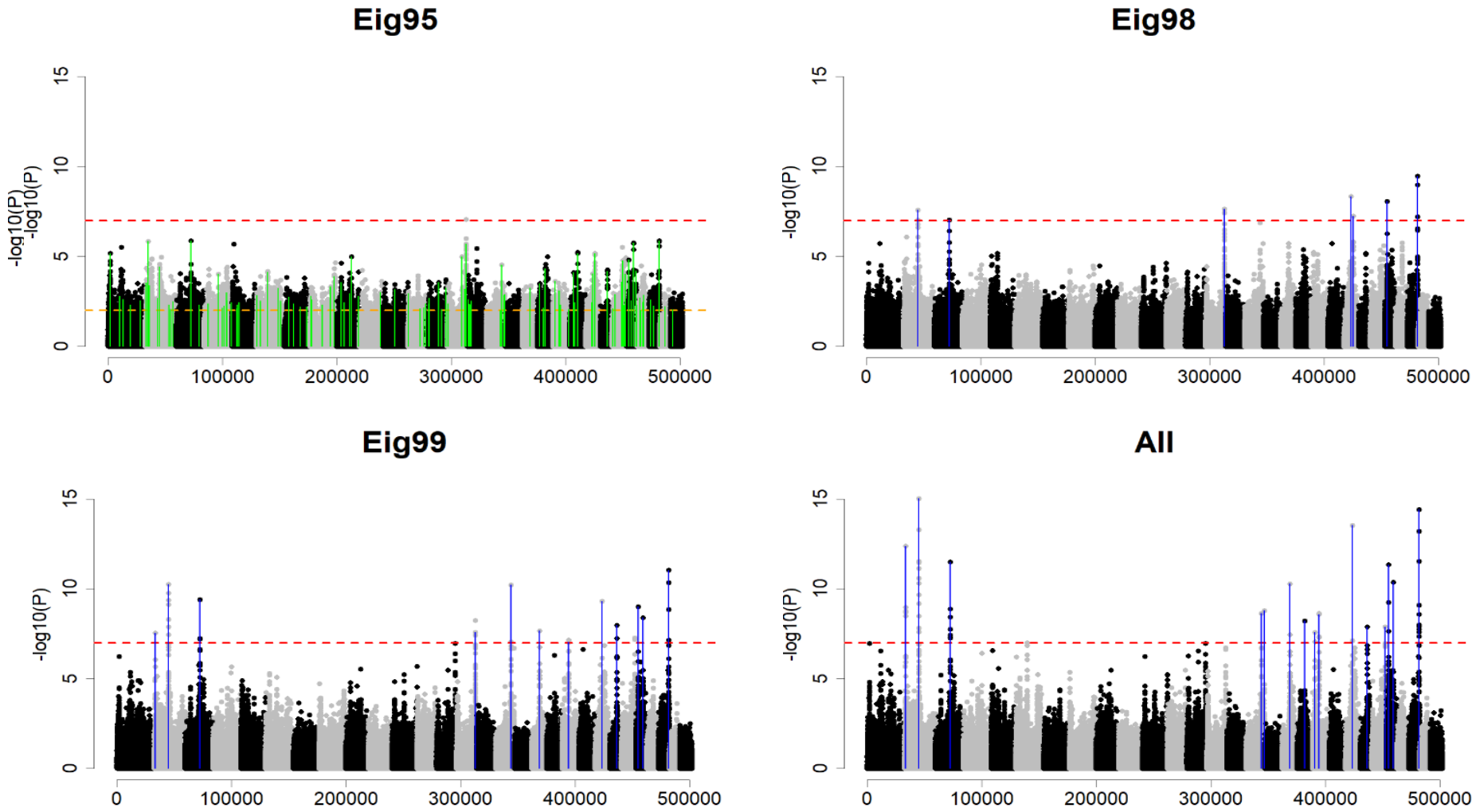




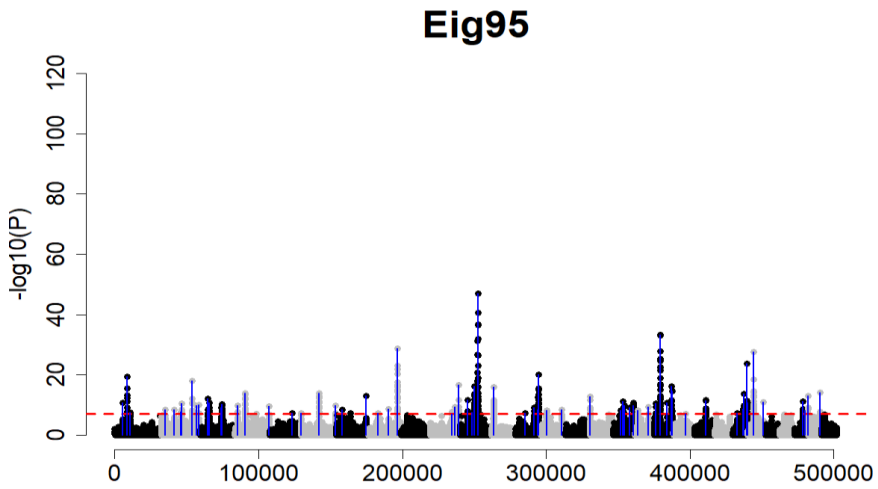
GWA results: Ne = 20 QTN = 2000 h^2 = 0.99

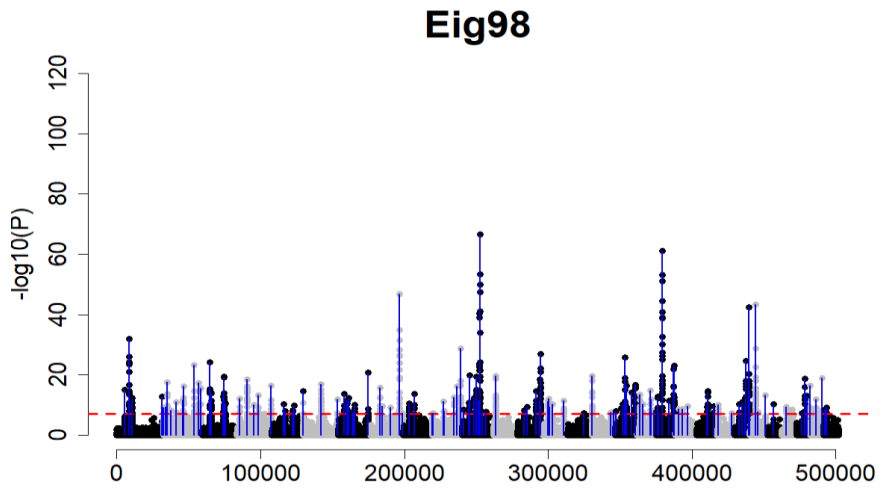


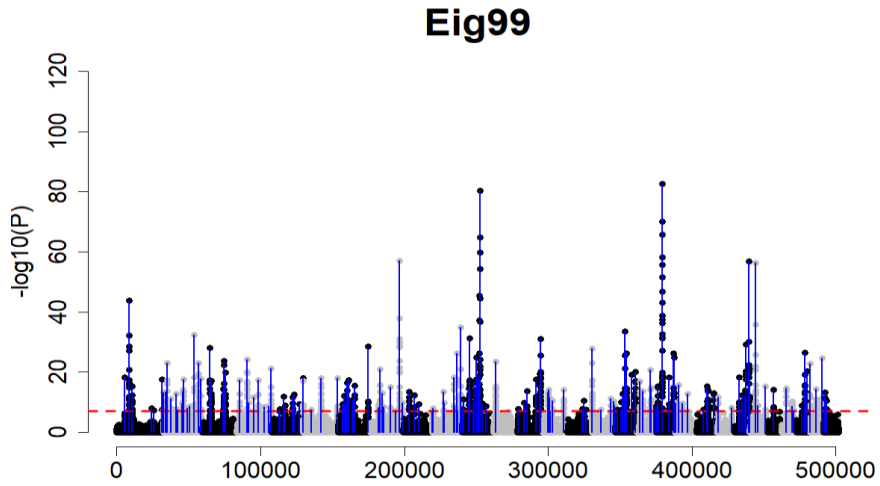
GWA results: Ne = 200 QTN = 2000 h^2 = 0.3

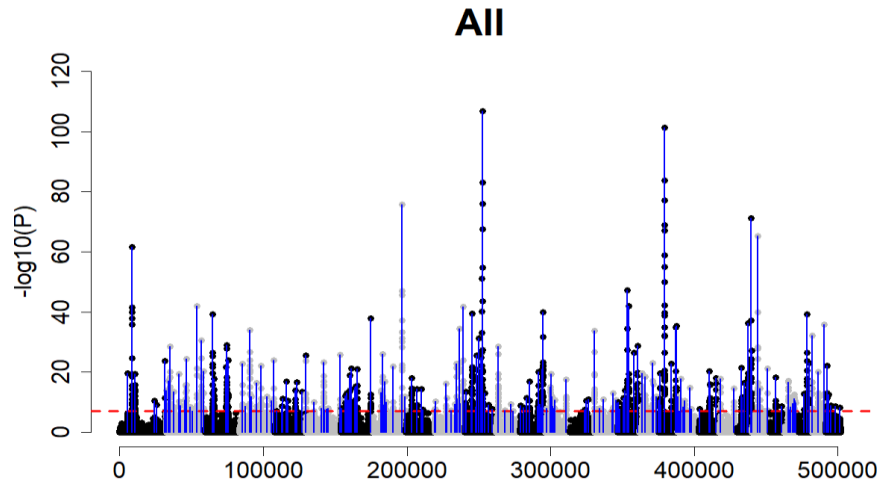


GWA results: Ne = 200 QTN = 2000 h^2 = 0.99





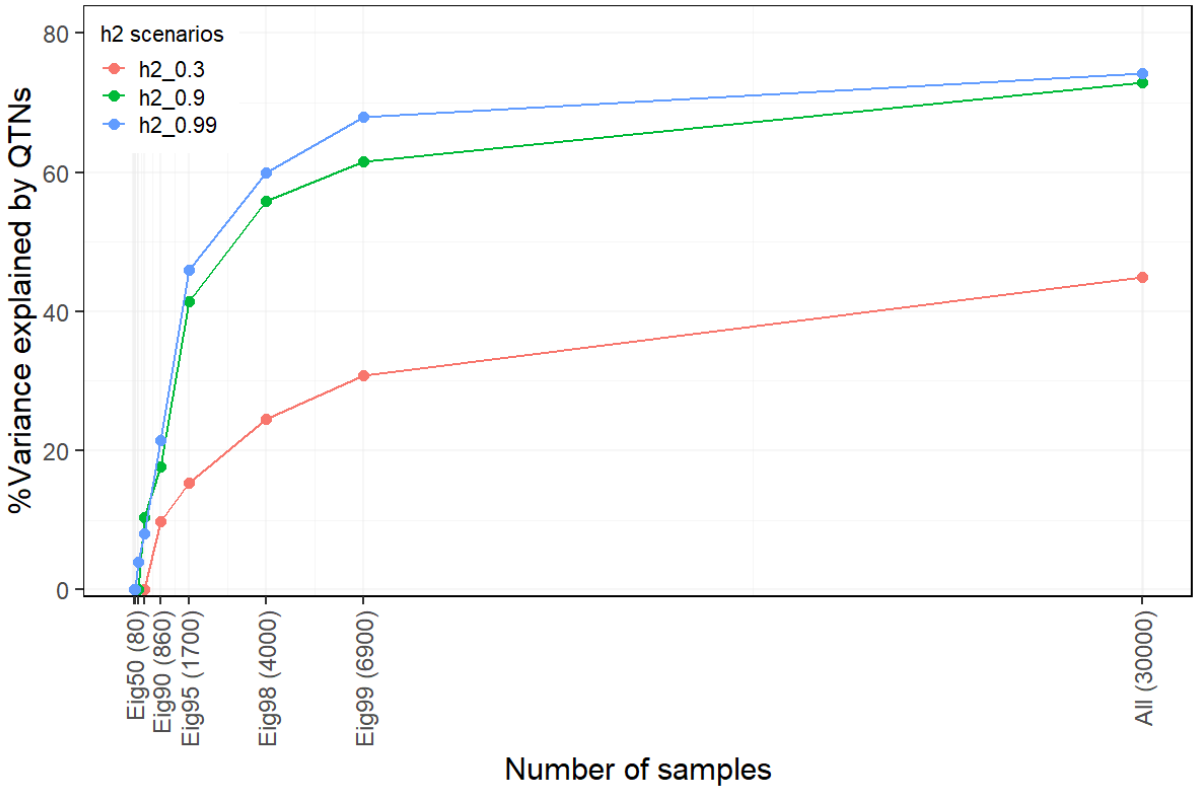




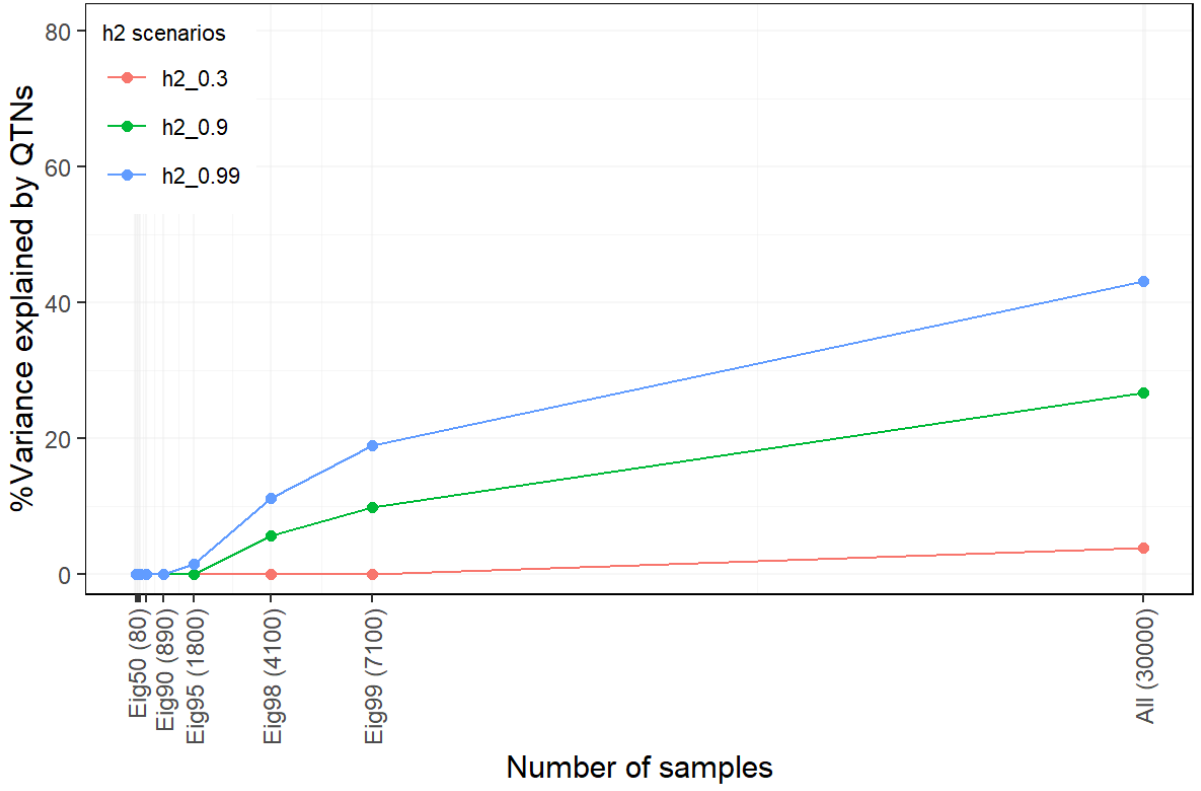
Total variance explained by QTNs

- **Ne = 20**

QTN = 200



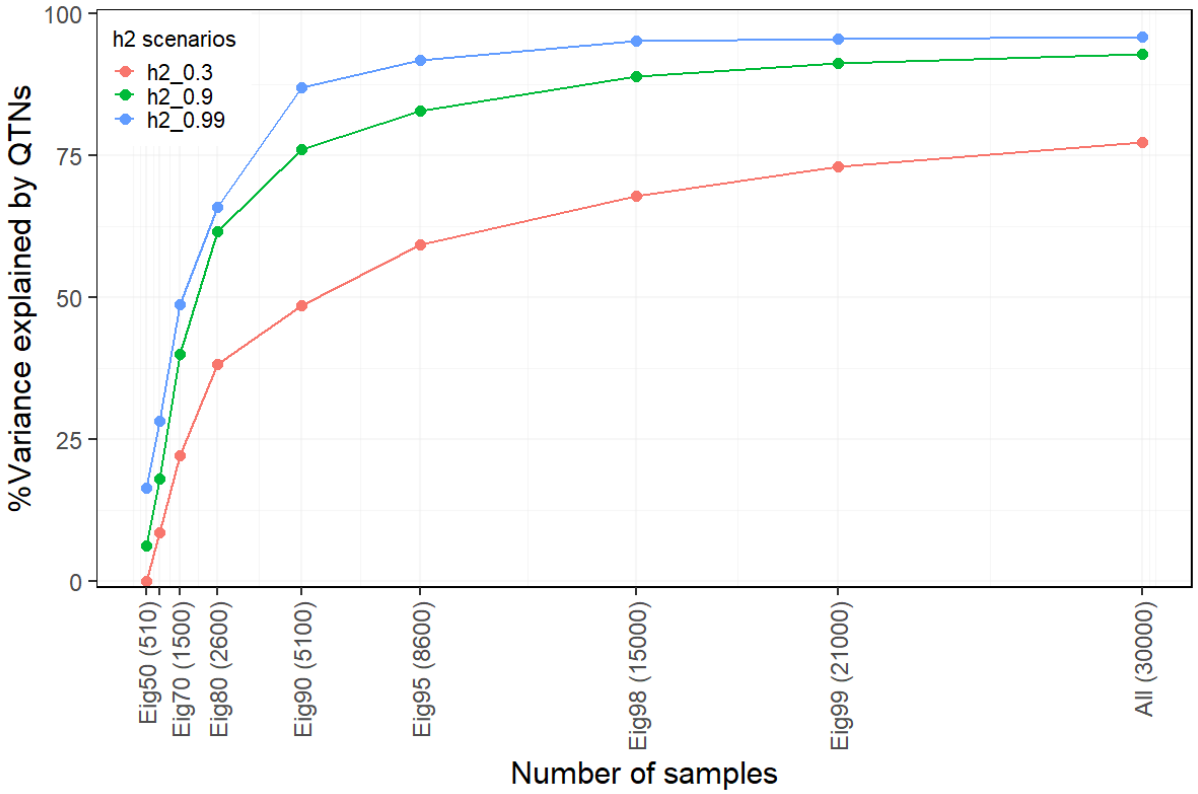
QTN = 2000



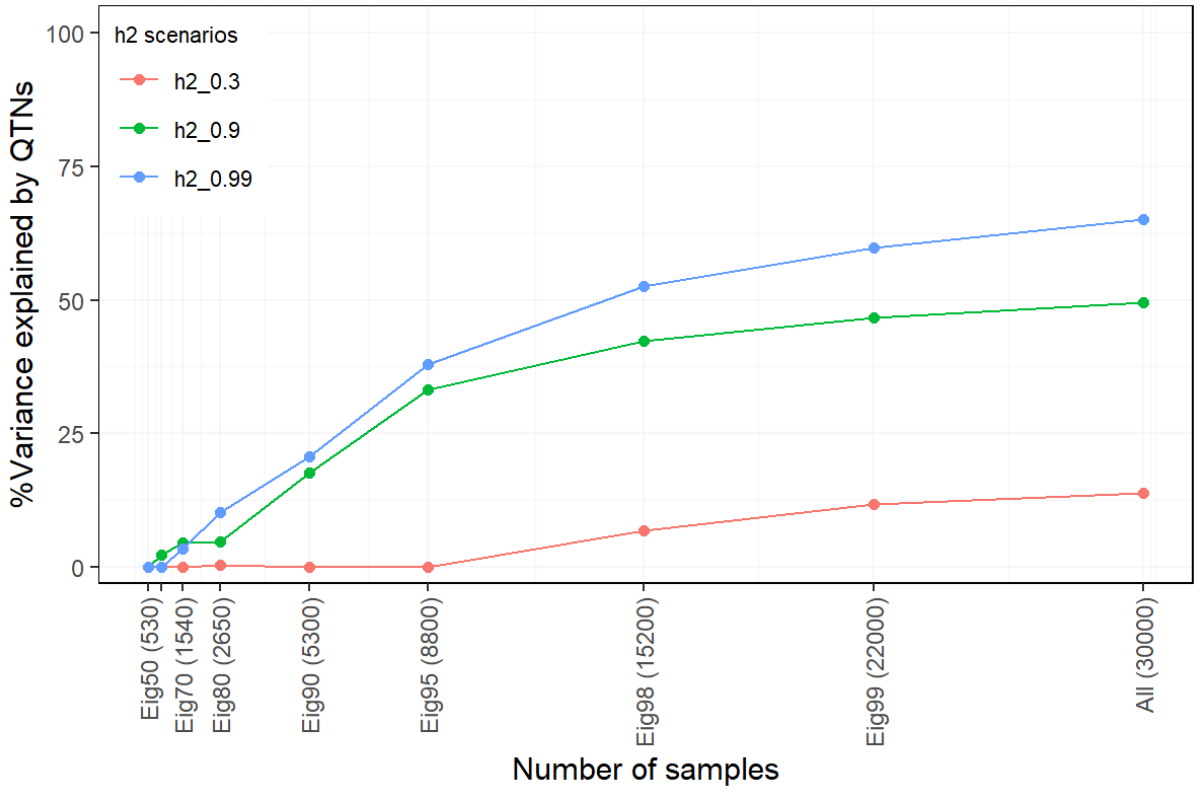
Total variance explained by QTNs

- **Ne = 200**

QTN = 200



QTN = 2000



Conclusions

- **The suitable sample size depends on the Ne and the number of QTNs**
  - Smaller populations require more data to capture causative variants
  - Larger populations: Eigen98 – Eigen99 to capture most informative
  - 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 Ne and Me
  - Testing the impact of incorporating selected variants for GP

Thanks ☺

jsbng@uga.edu