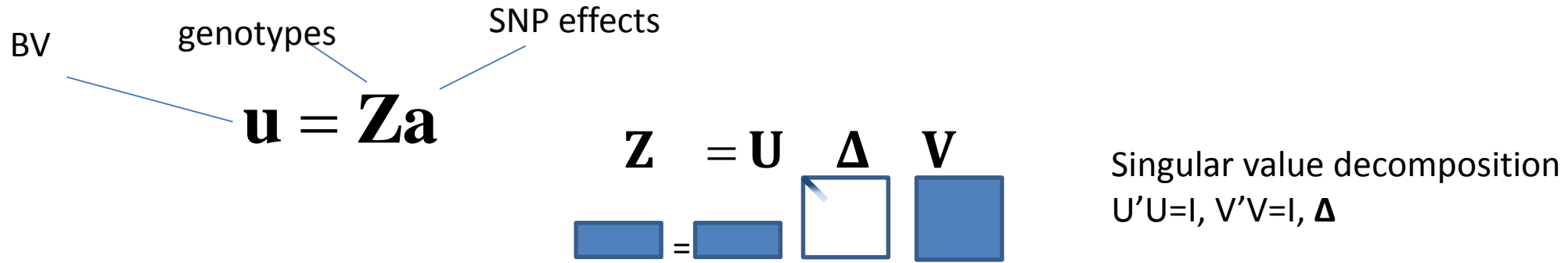


Possible implications of limited dimensionality of genomic information

**Ignacy Misztal, Ivan Pocrnic and Daniela
Lourenco**

University of Georgia

Dimensionality of genomic information



$$\mathbf{G} = \mathbf{U}\mathbf{\Delta}\mathbf{\Delta}\mathbf{U}' = \mathbf{U}\mathbf{D}\mathbf{U}'$$

Genomic relationship matrix
 $\text{Rank}(\mathbf{G}) \leq \min(\#\text{SNP}, \#\text{anim})$

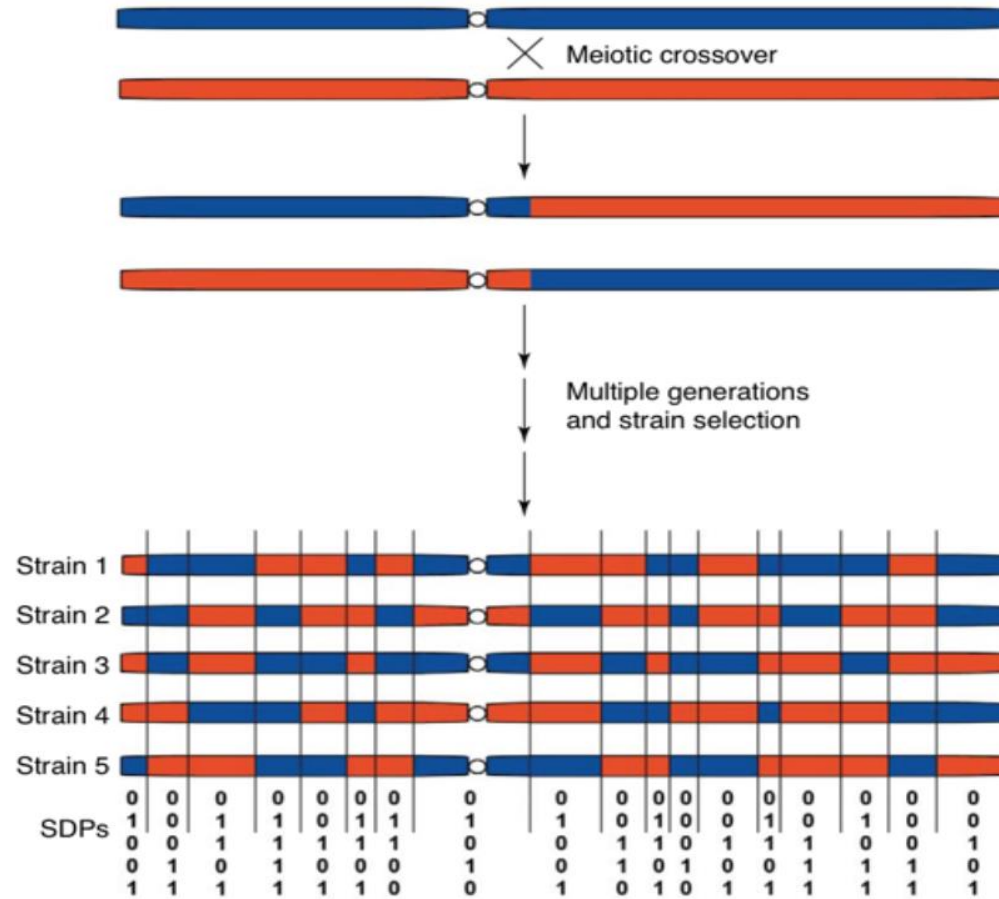
$$\mathbf{Z}'\mathbf{Z} = \mathbf{V}'\mathbf{\Delta}\mathbf{\Delta}\mathbf{V}$$

SNP BLUP design matrix
 $\text{Rank}(\mathbf{Z}'\mathbf{Z}) \leq \min(\#\text{SNP}, \#\text{anim})$

Same dimensionality for genotypes, GRM and SNP BLUP

Dimensionality around 5-15k (VanRaden, 2008; Maciotta et al., 2013)

Origin of Haplotype blocks



Cuppen, 2005

Chromosome segments



- Theory of junctions (Fisher, 1949):
 - Heterogenetic and homogenic tracts in genome
- For randomly mating population of constant size the number of tracts:

$$E(Me) = 4 \text{ Effective population size (Ne)} * \text{ Genomic size (L)} \text{ (Stam, 1980)}$$

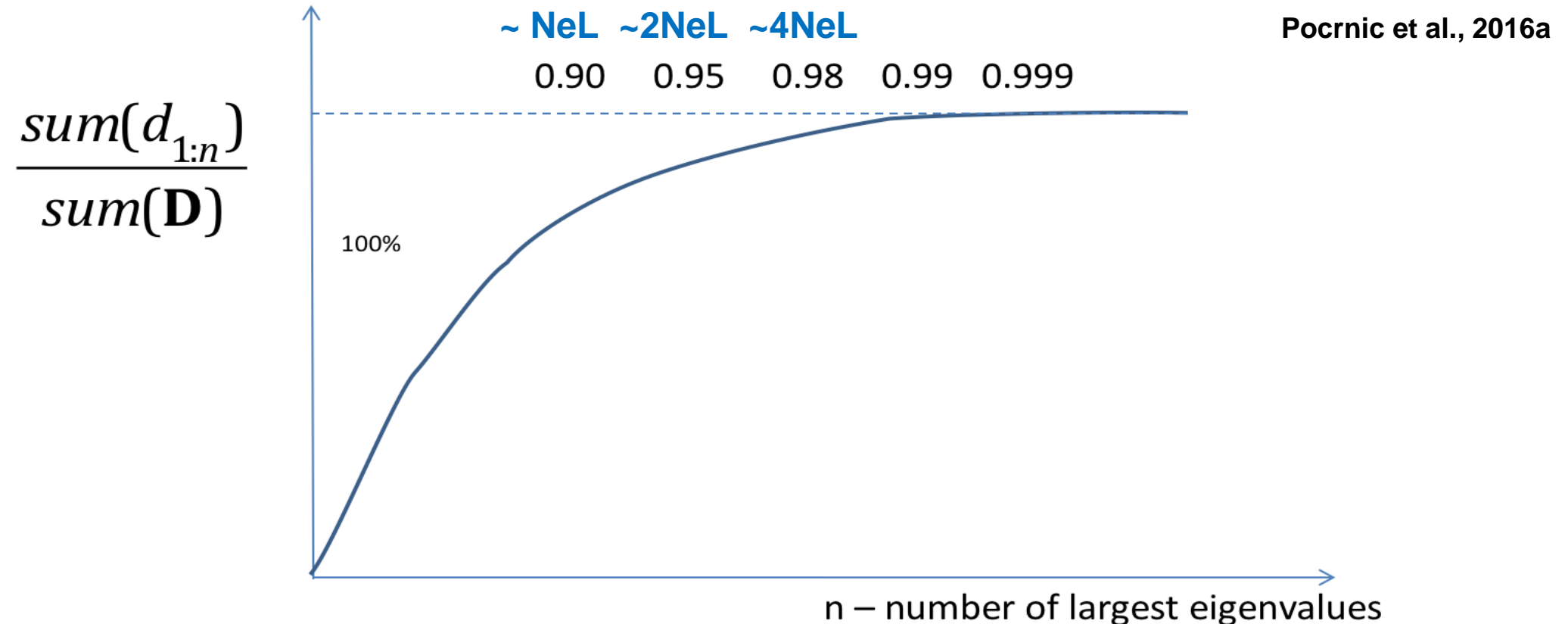
- Independent chromosome segments Me (Goddard, 2009; Daetwyler et al., 2010)
- Need $12 Me$ SNPs to detect 90% of junctions (MacLeod et al., 2005)

Number of junctions/chromosome segments/haplotype blocks

- $\sim 4N_eL$ Stam (1980) 12,000 for Holsteins
- $2N_eL$ Hayes et al. (2009) 6,000
- $2N_eL/[\log(N_eL)]$ Goddard et al. (2011) ~ 500

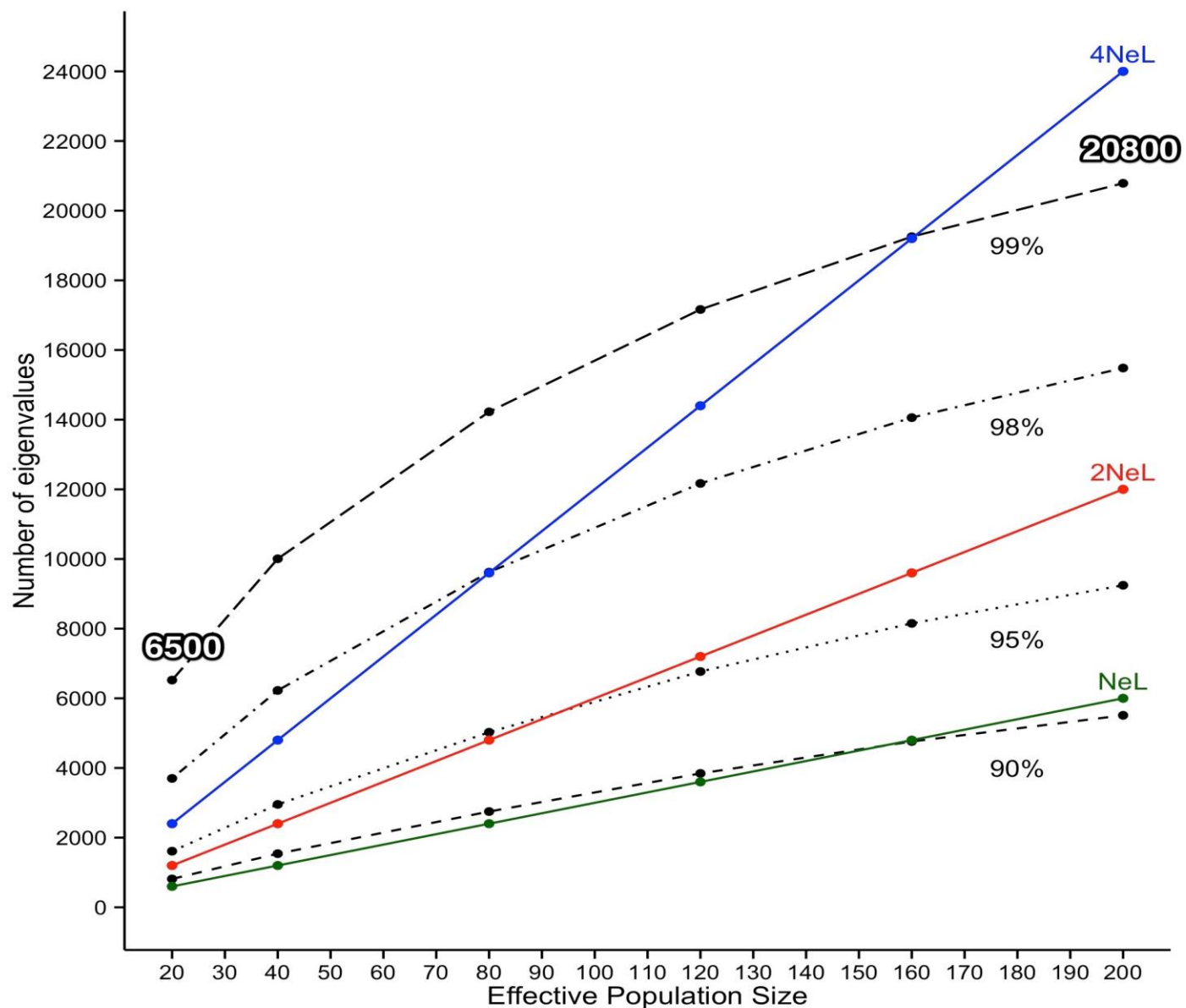
Fraction of **G** variance explained

$$\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}'$$



What fraction of variance in **G** is information and what is noise?

Number of largest eigenvalues to account for a given variance



How to determine dimensionality in practice - APY inversion of GRM

Breeding values \mathbf{u} N chromosome segments \mathbf{s}
 $\mathbf{u} = \mathbf{T}\mathbf{s}$

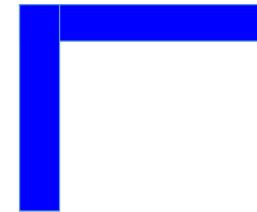
Choose any N animals called “core”: \mathbf{u}_c

$\mathbf{s} = \mathbf{Q}\mathbf{u}_c + \boldsymbol{\varepsilon}_c$ Segments linear function of core N animals

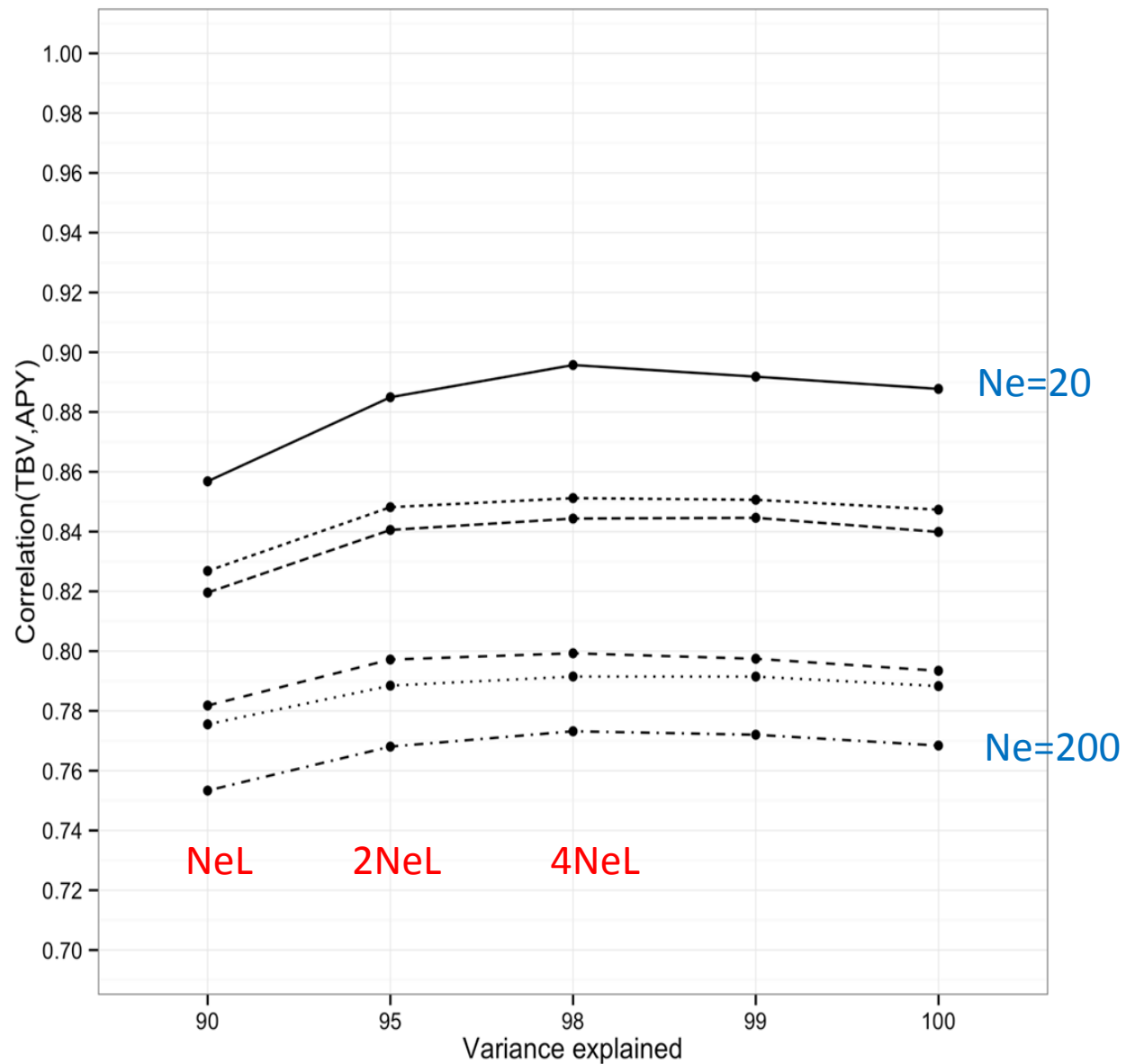
$\mathbf{u}_n = \mathbf{T}_n\mathbf{s} = \mathbf{P}_{nc}\mathbf{u}_c + \boldsymbol{\varepsilon}_n$ Noncore animals linear functions of core animals

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{P}_{nc} & \mathbf{I} \end{bmatrix}$$

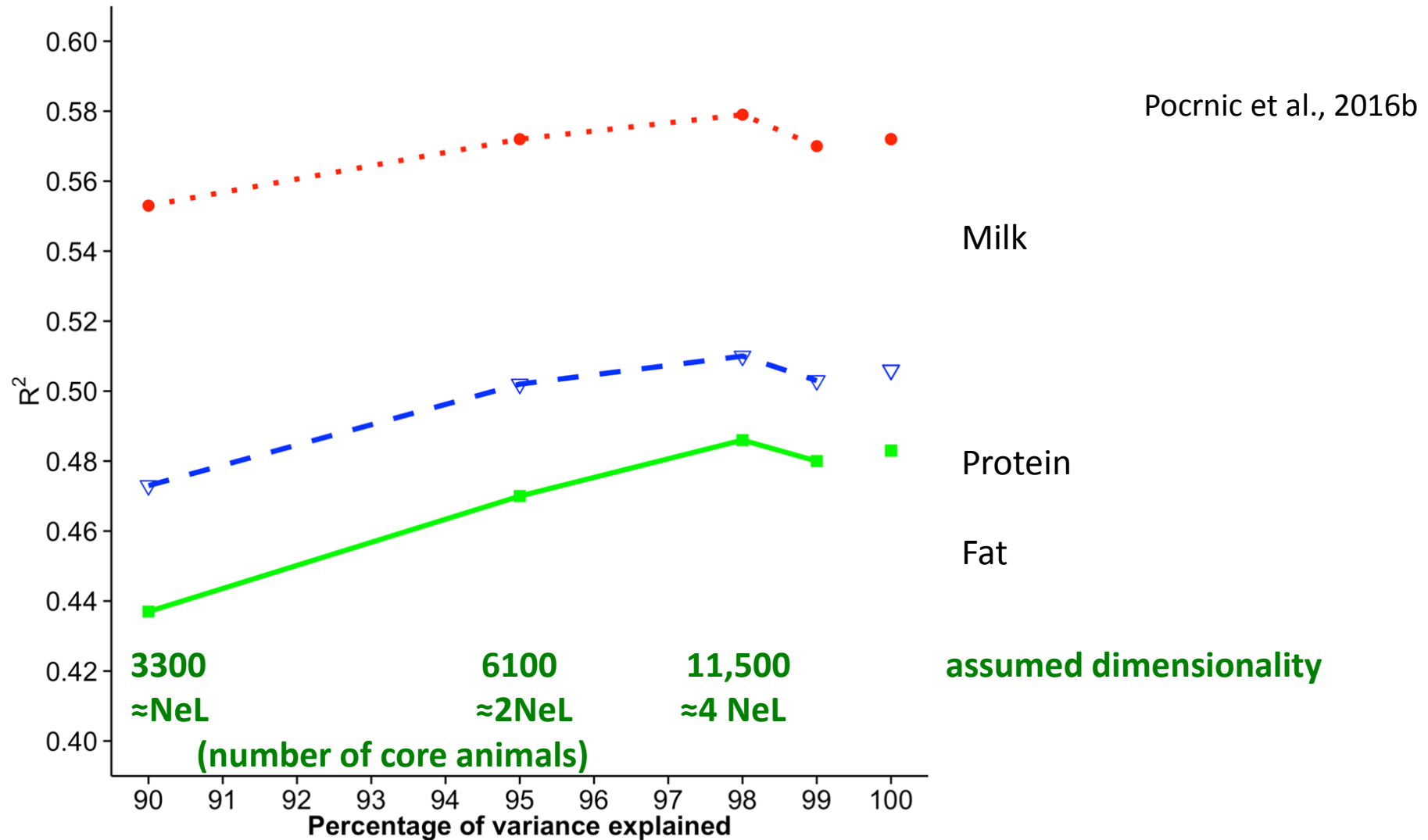
Sparse inverse



True accuracies as function of number of eigenvalues



Reliabilities – Jerseys (75k animals)



100% = full inverse \rightarrow lower accuracy

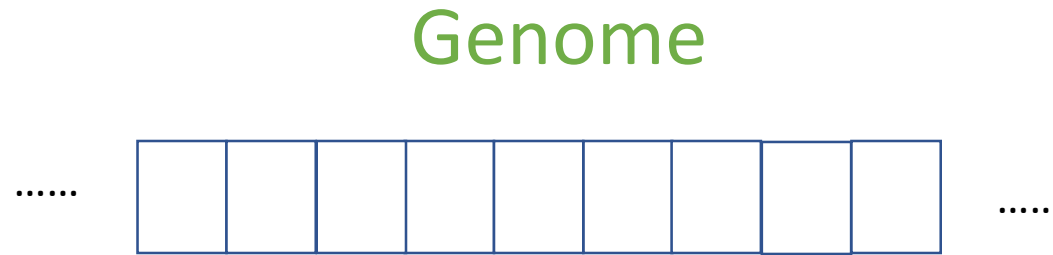
Estimated effective population size and the number of segments

Specie	Effective population size	Me
Holsteins	149	18k
Jerseys	101	12k
Angus	113	13k
Pigs	43	4k
Chicken	44	4k

Impact of reduced dimensionality

- Accuracy with SNP selection
- Theoretical accuracies
- Persistency of GEBV
- GWAS

Understanding of limited dimensionality (I)



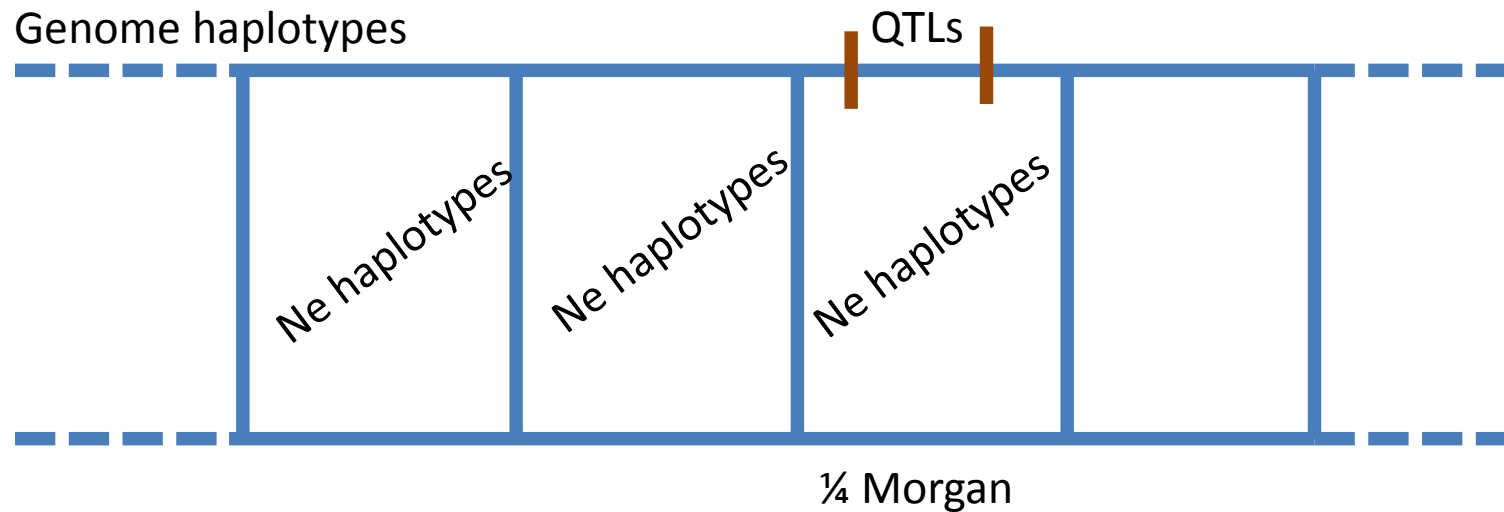
$\approx 4 N_e L$ segments

Average size $L/(4N_eL)$

With $N_e=100$, $L=30$, Genome size 3 Gb \Rightarrow 1 segment \approx 250 kb

Understanding of limited dimensionality (II)

Number of haplotypes: $4 N_e L$
Ne within each $\frac{1}{4}$ Morgan segment



Dimensionality of $\frac{1}{4}$ Morgan case: N_e

→ Reduced dimensionality with weighted GRM

ssGBLUP accuracies using SNP60K and 100 QTNs – simulation study



Fragomeni et al. (2017)

Data: 60k genotyped animals
60k SNP + 100 QTN

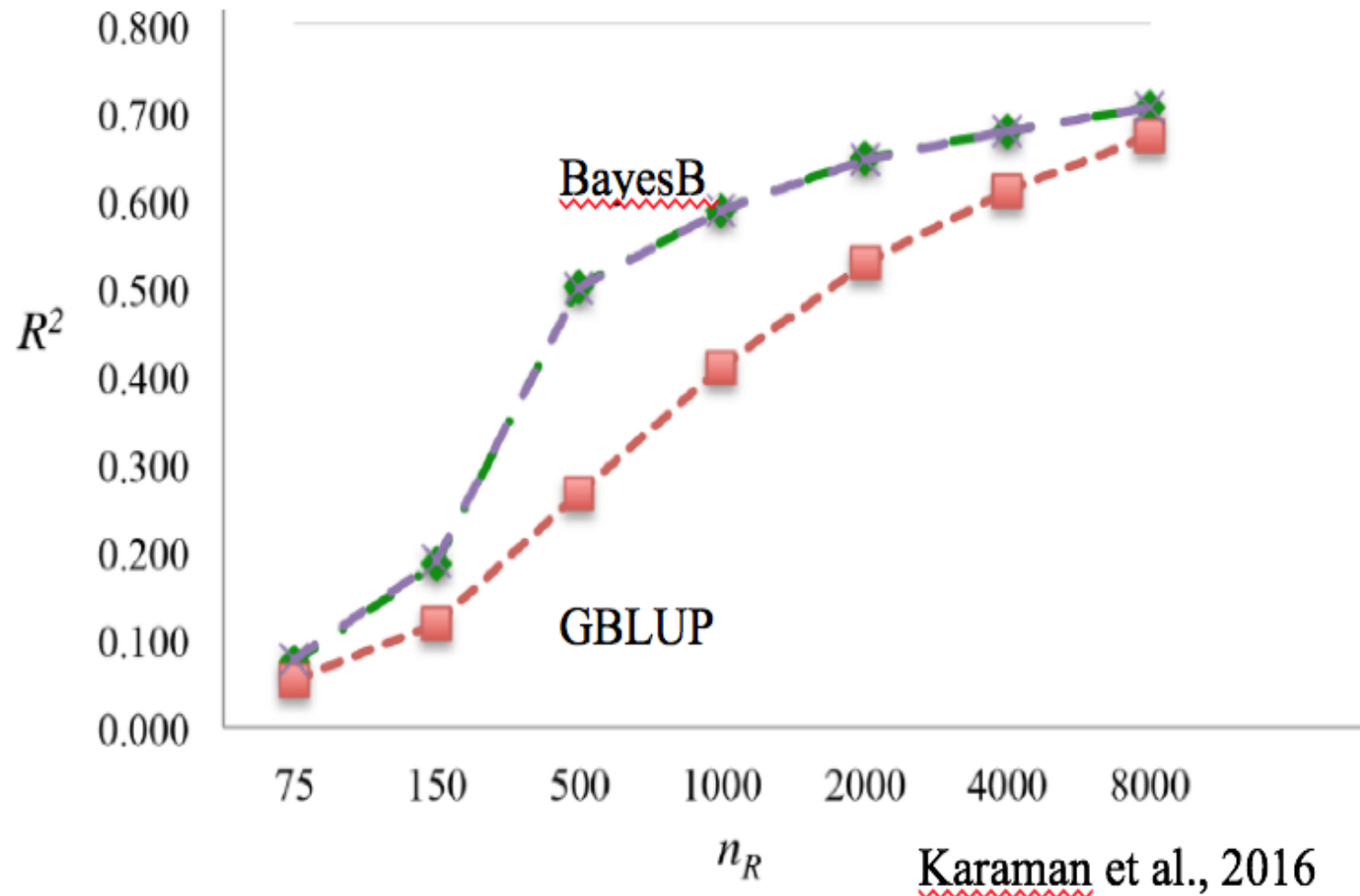
98% Dimensionality:

19k – unweighted G

5k - weighted G

98 - only QTN

Advantage of SNP selection and size of data



Accuracy as generation of core animals

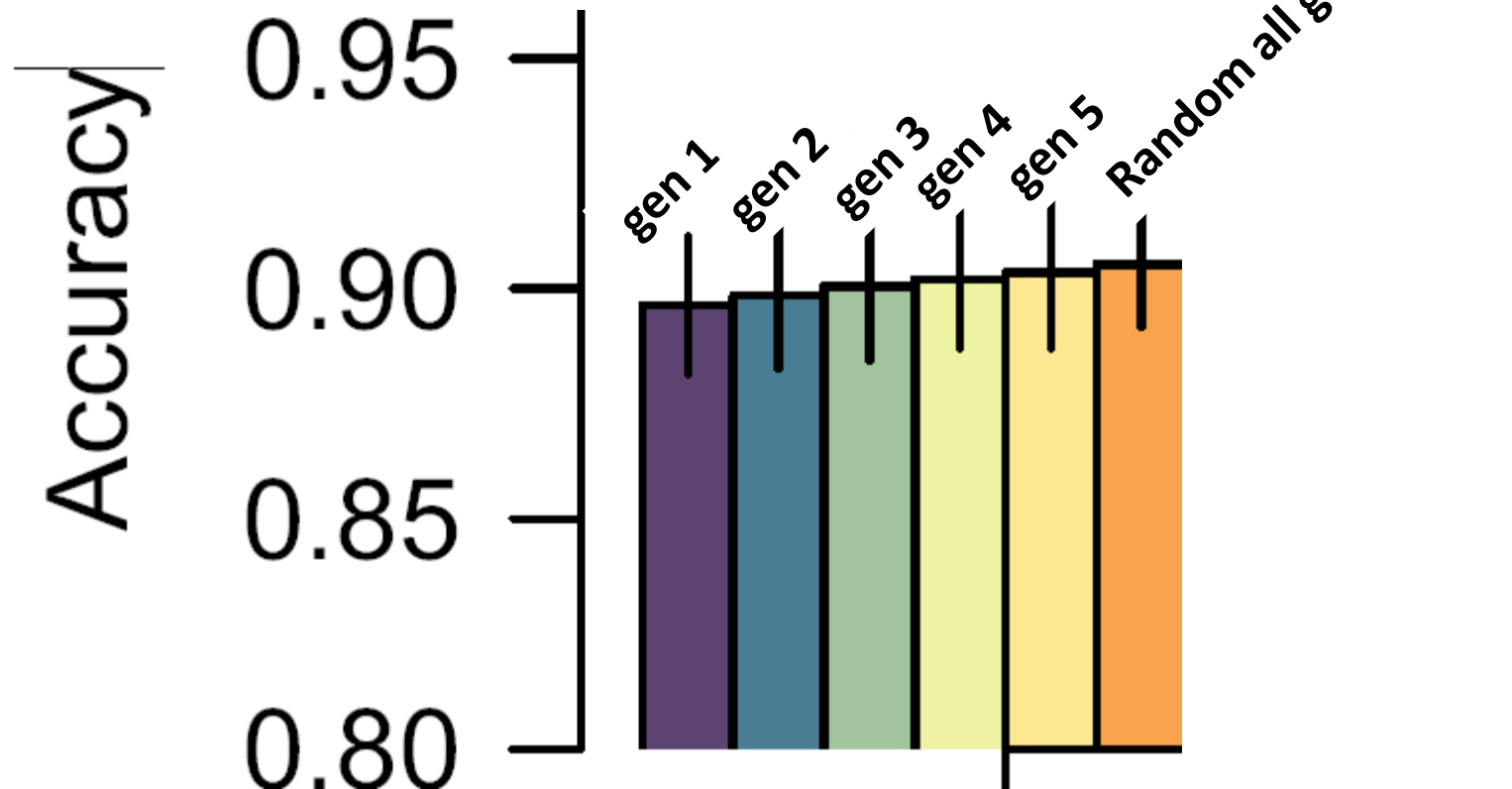


Bradford et al. (2017)

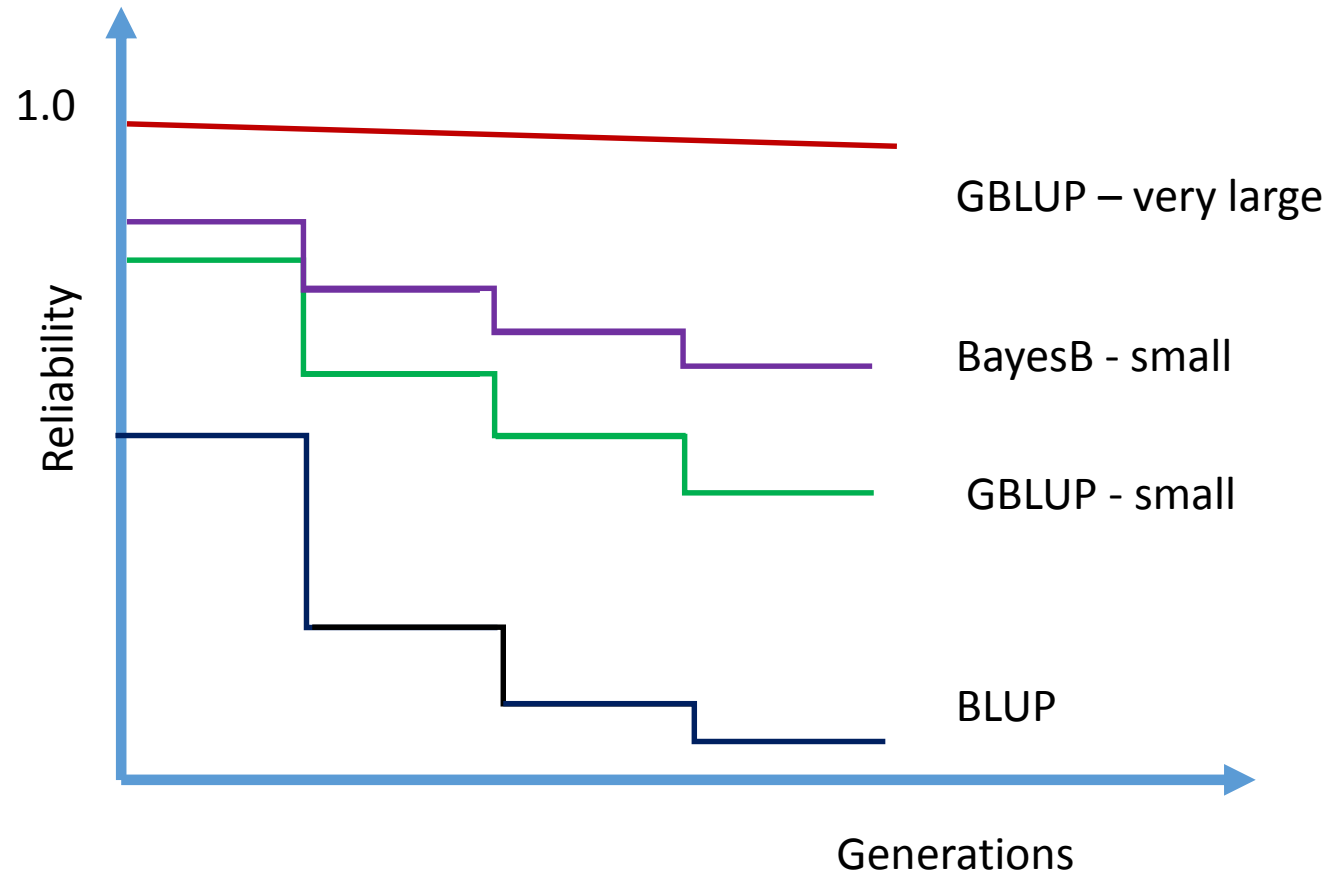
Noncore animals linear
functions of core animals

$$\mathbf{u}_n = \mathbf{P}_{nc} \mathbf{u}_c + \boldsymbol{\varepsilon}_n$$

Selection does not change
segments (additive model
only)



Persistence over generations with different sizes of reference populations



Very large – equivalent to 4NeL animals with 99% accuracy

Are SNP effects from Holstein national populations converging?

Accuracy approximations

- Based on equal sized segments (Daetwyler et al., 2008)

$$r = \sqrt{\frac{Nh^2}{Nh^2 + M_e}}$$

N – number of animals

M_e – number of segments

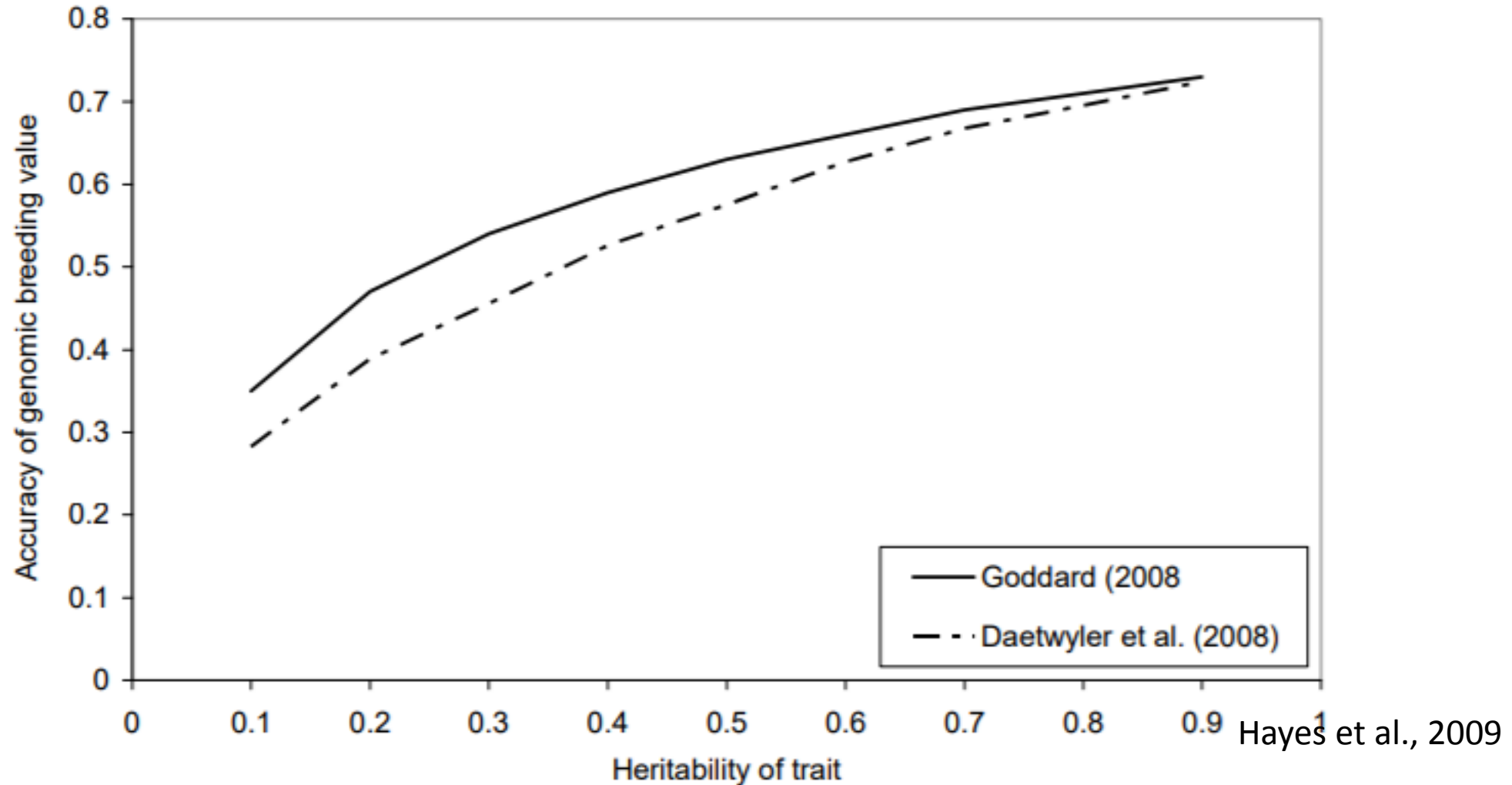
- Based on segments modified by QTL frequencies (Goddard, 2009)

$$r = \sqrt{1 - \frac{\lambda}{2N\sqrt{\alpha}} \ln \left(\frac{1 + \alpha + 2\sqrt{\alpha}}{1 + \alpha - 2\sqrt{\alpha}} \right)}$$

$M_e / (h^2 \ln(2N_e))$

$$1 + 2(M_e / Nh^2 \ln(2N_e)).$$

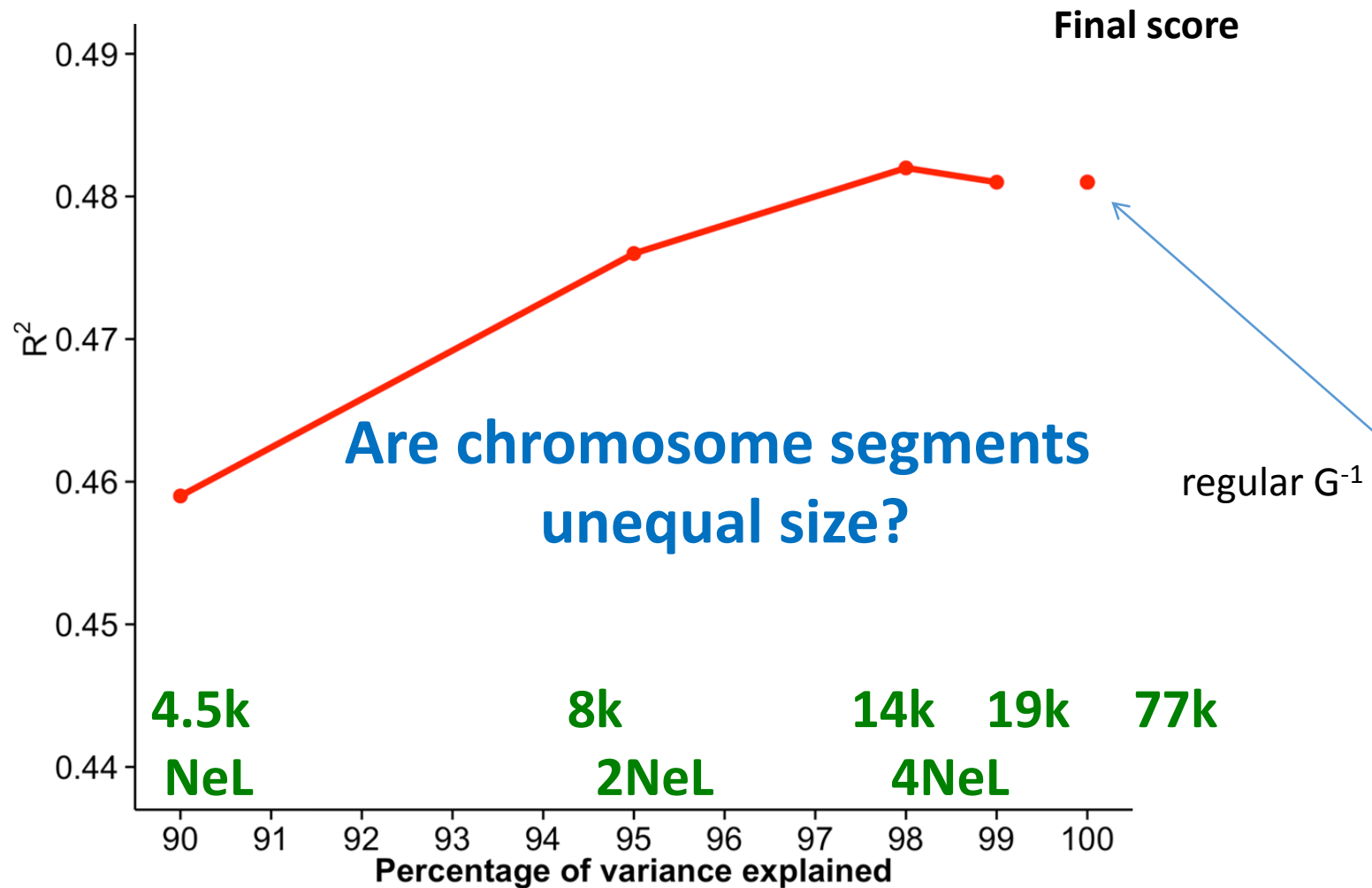
Figure 1. Accuracy of genomic breeding values with 5000 phenotypic records, effective population size of 100 and increasing heritability, predicted by the deterministic formula of Goddard (2008) or Daetwyler et al. (2008).



Theory and practice

- Theoretical formulas not useful (Brard and Ricard, 2015)
 - Effect of selection?
 - Wrong numbers?
 - Segments not equal?

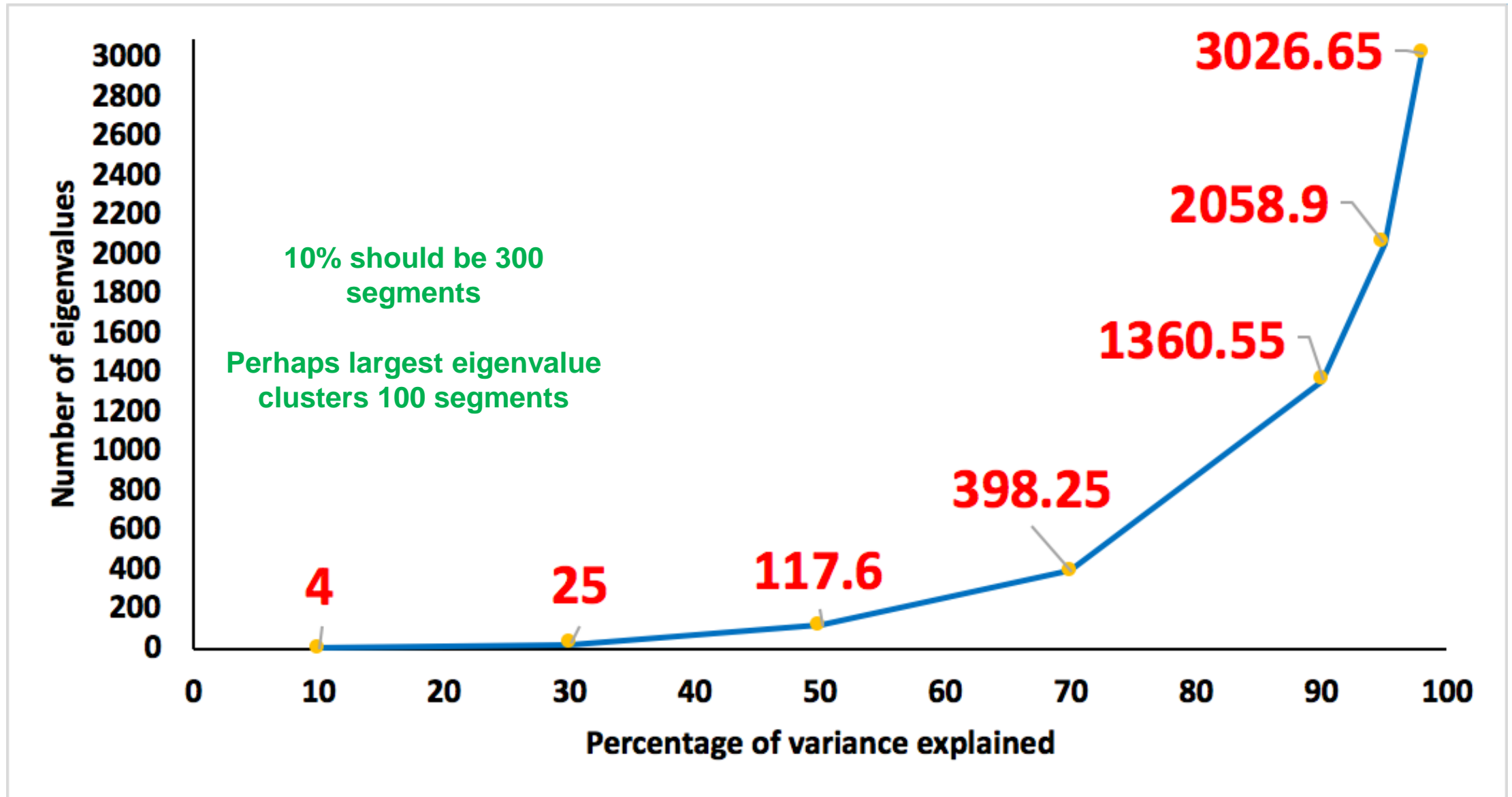
Reliabilities assuming different dimensionality with APY inverse – Holsteins



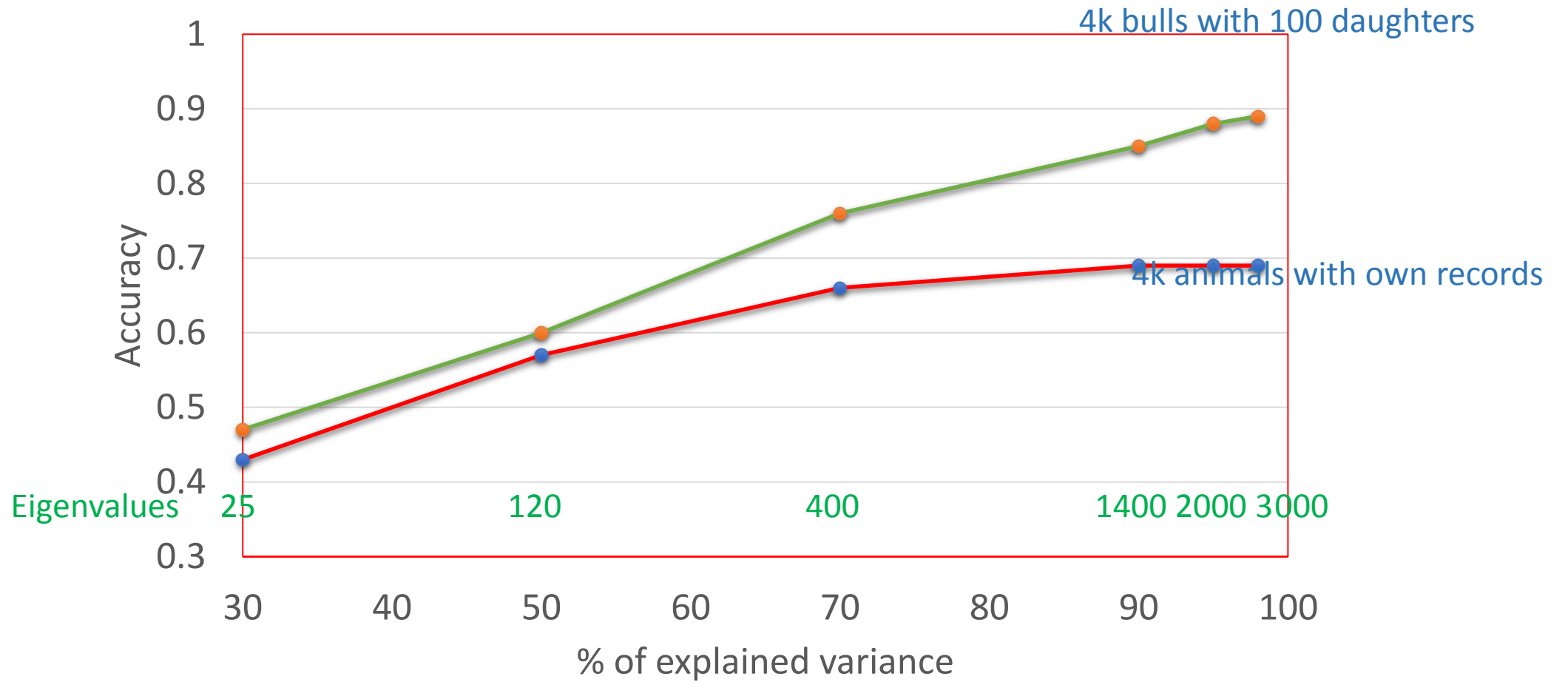
Is genomic selection on chromosome segments or chromosome clusters ?

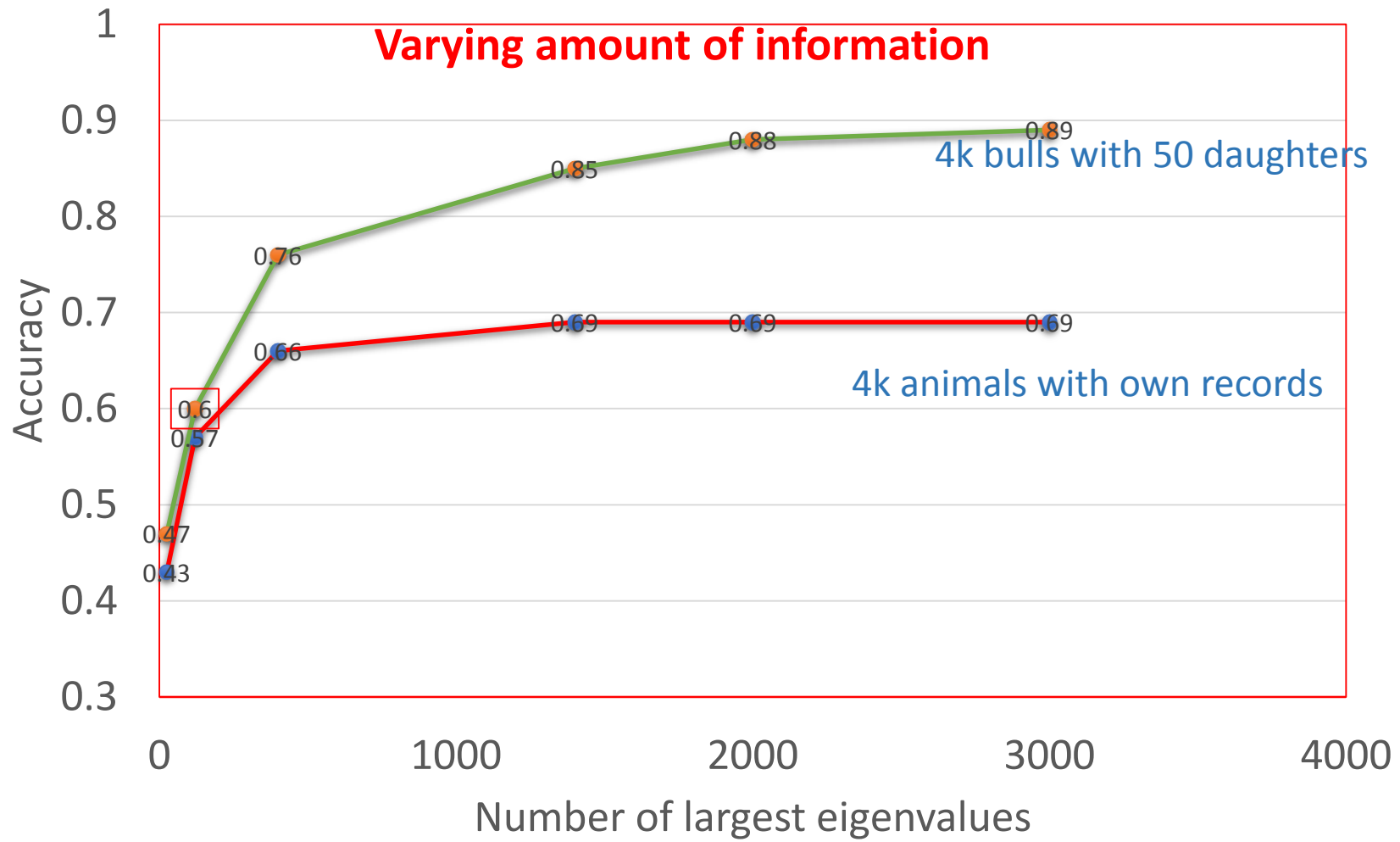
- Simulation
 - 6k animals with 50 k SNP
 - $N_e \approx 50$, $L = 10M$
- GBLUP
 - Use GRM with limited number of eigenvalues (corresponding to 10 to 99% variation)
 - 4k animals in reference population, 2k in validation

Eigenvalue profile of GRM

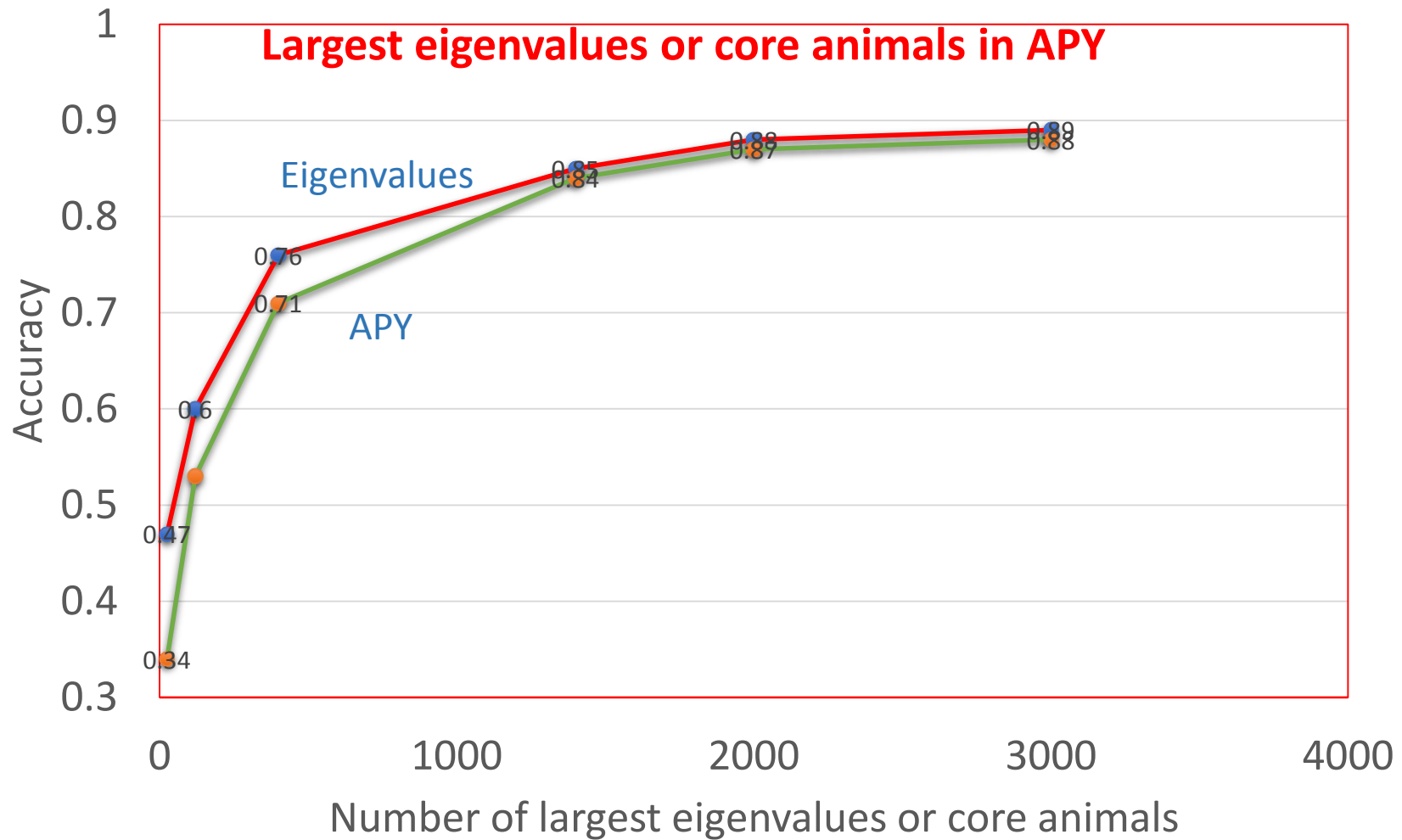


Accuracies of GBLUP using GRM with largest eigenvalues only





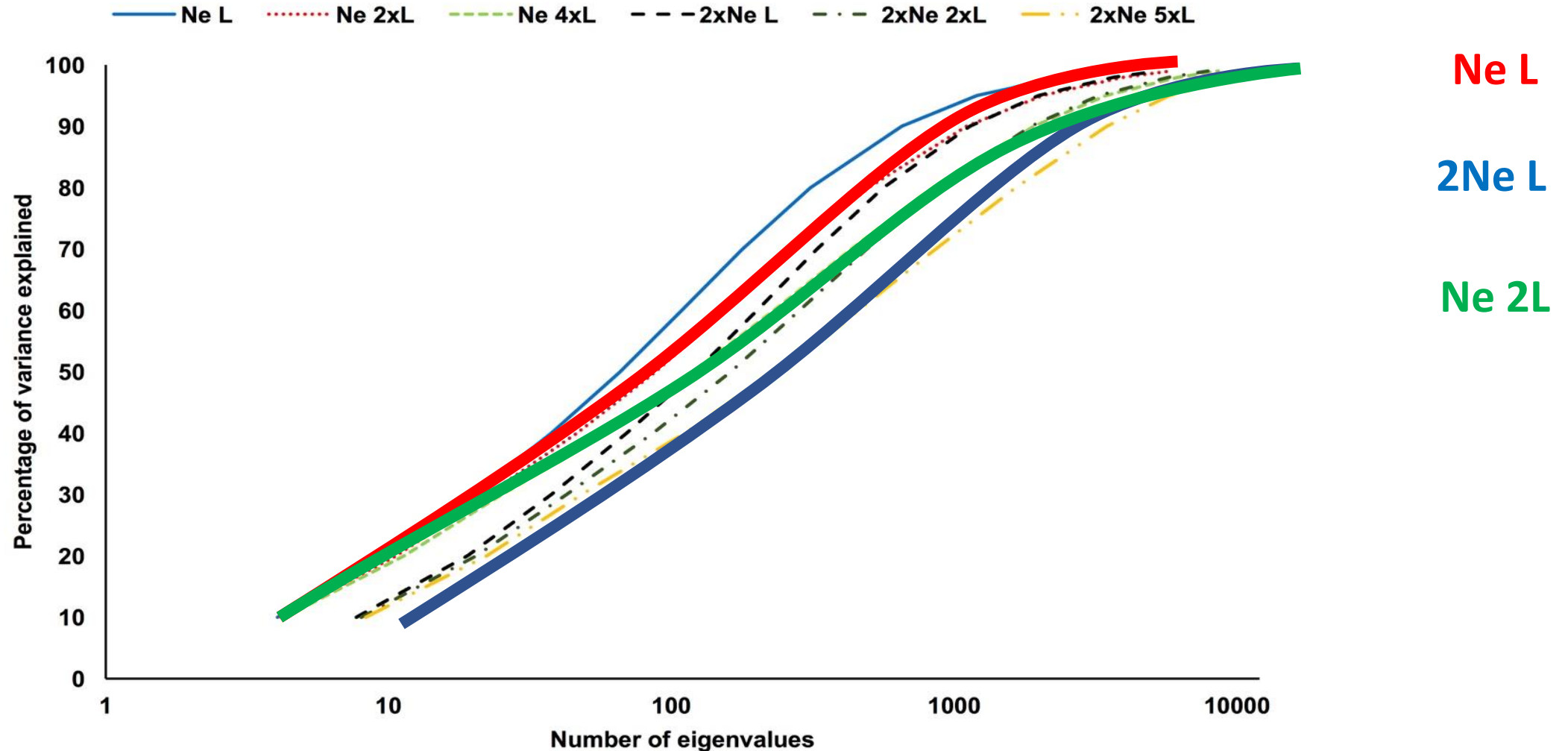
Does APY algorithm for inversion of GRM work on segments or eigenvalues



Selection on largest eigenvalues – important ancestors – reduced N_e
If largest eigenvalues excluded- increased diversity?

How are eigenvalues influenced by effective population size and genome length?

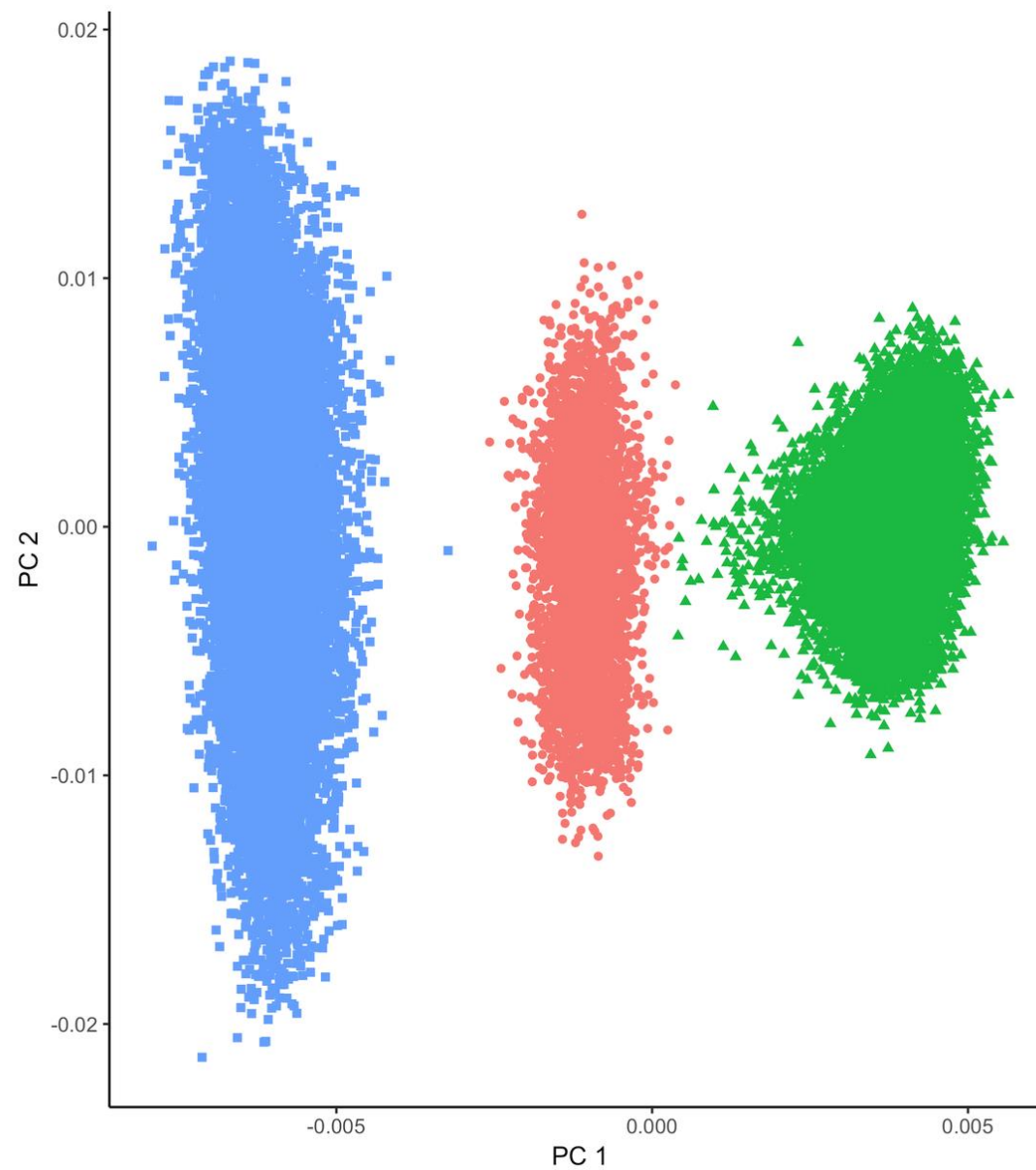
Graph of dimensionality as % for different Ne and L



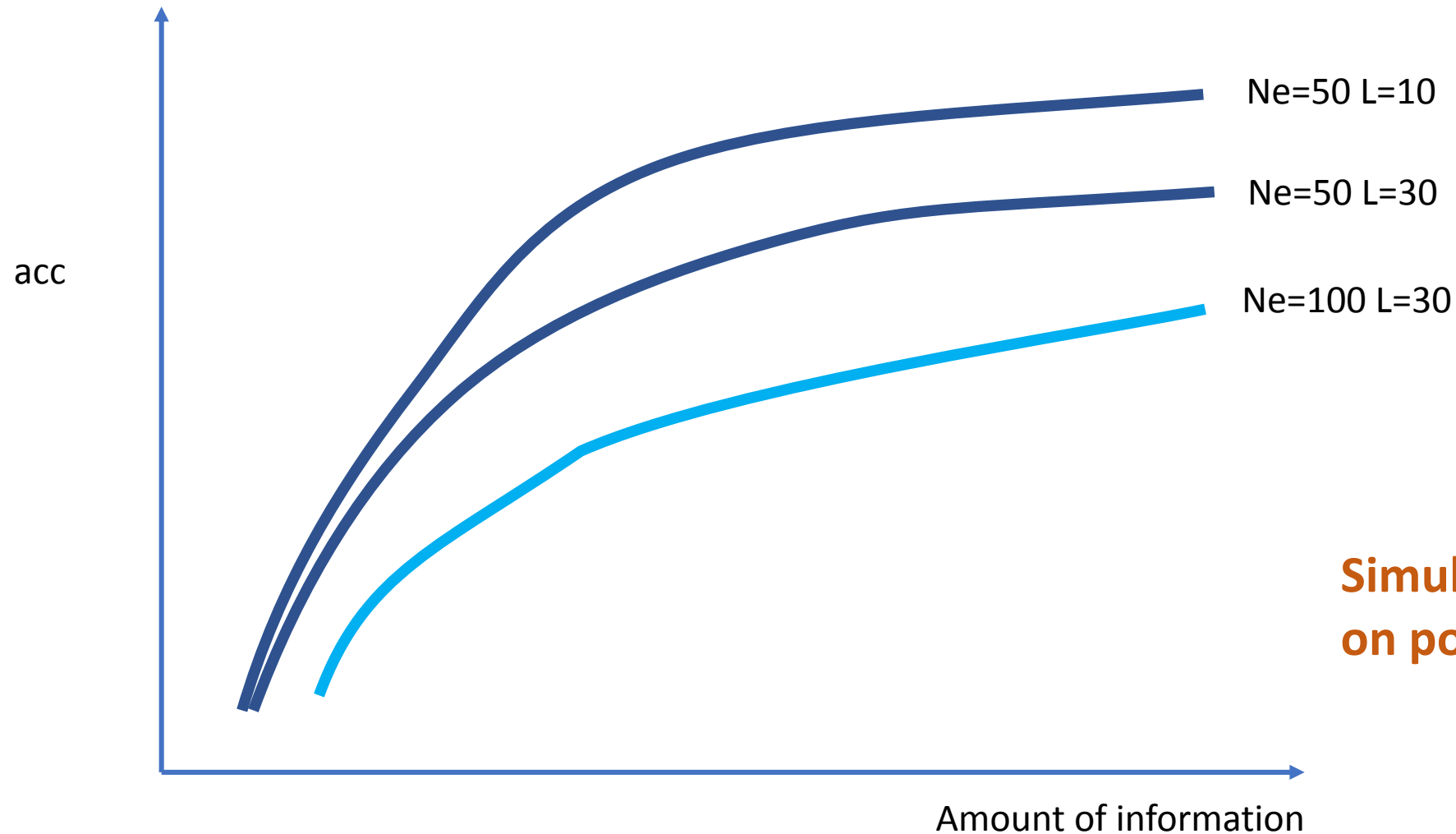
Largest eigenvalues do not depend on genome size - cluster haplotypes across all genome

PCA Plot

PC1 and PC2 pool segments across genome



Hypothetical accuracies as function of Ne and genome length

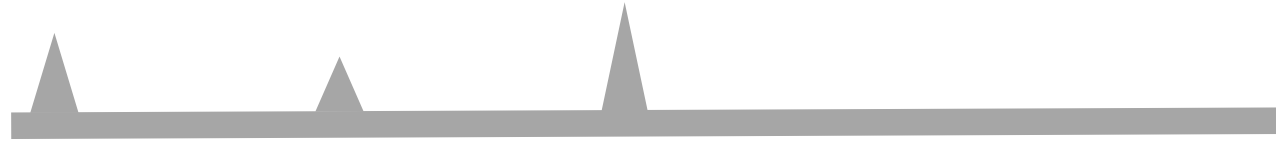
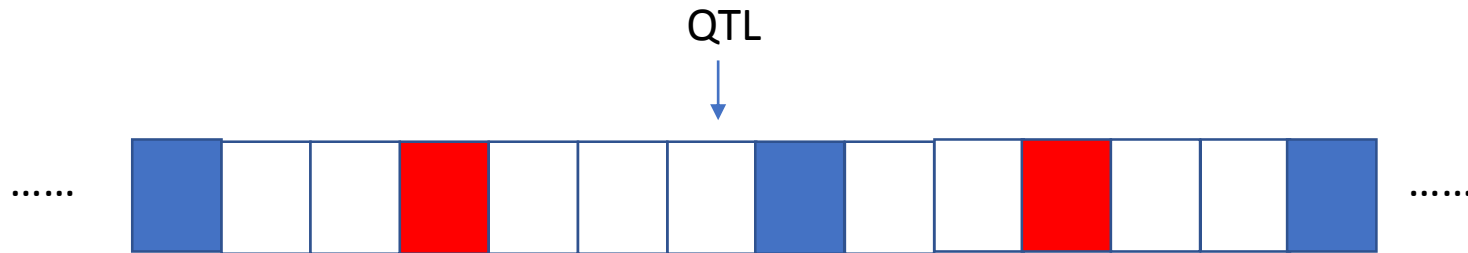


Simulation results depend on population parameters

Some hypothesis on GWAS

First cluster

Second cluster



BayesB1



BayesB2



GBLUP



Classical with pop.
stratification



Conclusions

- Large impact of limited dimensionality of genomic info
 - Accuracies
 - Persistence
 - GWAS
 - ...
- Little data required for medium accuracy, large data for high accuracy
- Many hypotheses - potential studies with real data sets
- Collaborators welcome, funding available



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



Group and sponsors



Shogo
Tsuruta



Ignacio
Aguilar



Breno
Fragomeni



Ivan
Pocrnic



Daniela
Lourenco



Yutaka
Masuda



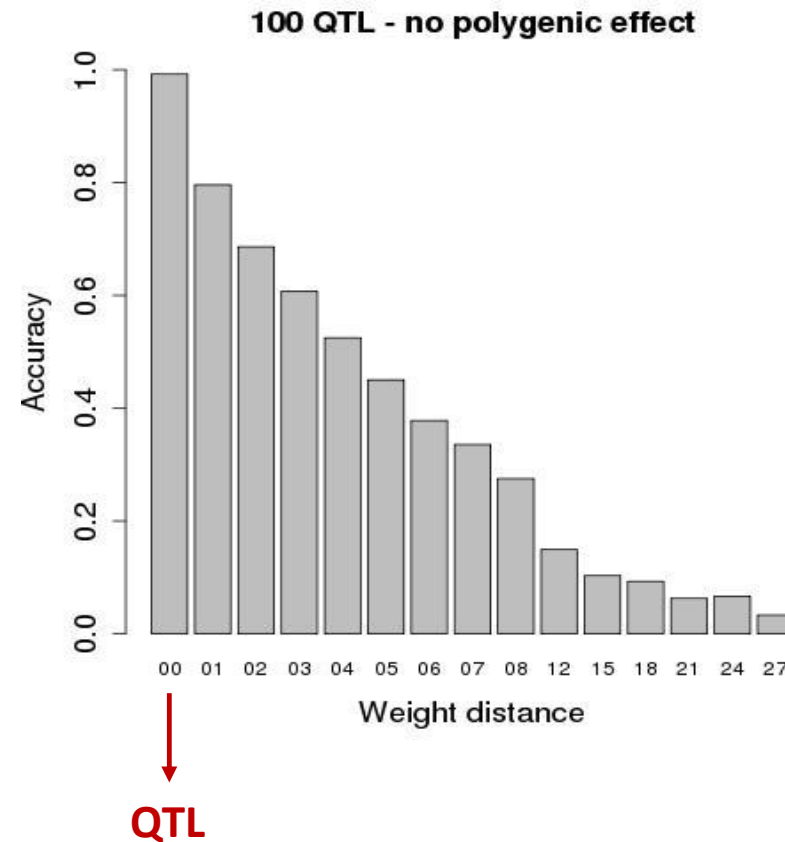
Andres
Legarra



Heather
Bradford

Accuracy and distance from markers to QTL

Fragomeni et al. (2017)



Questions - summary

- Is accuracy of GBLUP proportional to explained variance in \mathbf{G} ?
- Can accuracy of GBLUP be expressed in the terms of variance explained by N largest segments, i.e. eigenvalues of \mathbf{G}
 - e.g. 10% variance = 10% accuracy; 50%=50%; ... ?
- Is dimensionality of \mathbf{G} related to number of core animals?
- What accuracies with n core animals that have perfect BV?
- Do accuracies of GBLUP reach 0.99 with many animals?
- Are APY and SVD/EIGEN methods related?