Limited dimensionality of genomic information and effective population size

Ivan Pocrnić [1], D.A.L. Lourenco [1], Y. Masuda [1], A. Legarra [2] & I. Misztal [1]

[1]University of Georgia, USA

[2]INRA, France

WCGALP, Auckland NZ, Feb 2018

Dimensionality of genomic information (i)

- **Z** – matrix of gene content
- Singular value decomposition: **Z = U D V'** **(U'U=I, V'V=I)**

- Eigenvalues: Genomic relationship matrix **G** = (**ZZ'**/k) = **UDDU'**
- Eigenvalues: SNP-BLUP design matrix **Z'Z = V'DDV**
- Genomic information (**Z**, **ZZ'** or **Z'Z**) has the same limited dimensionality

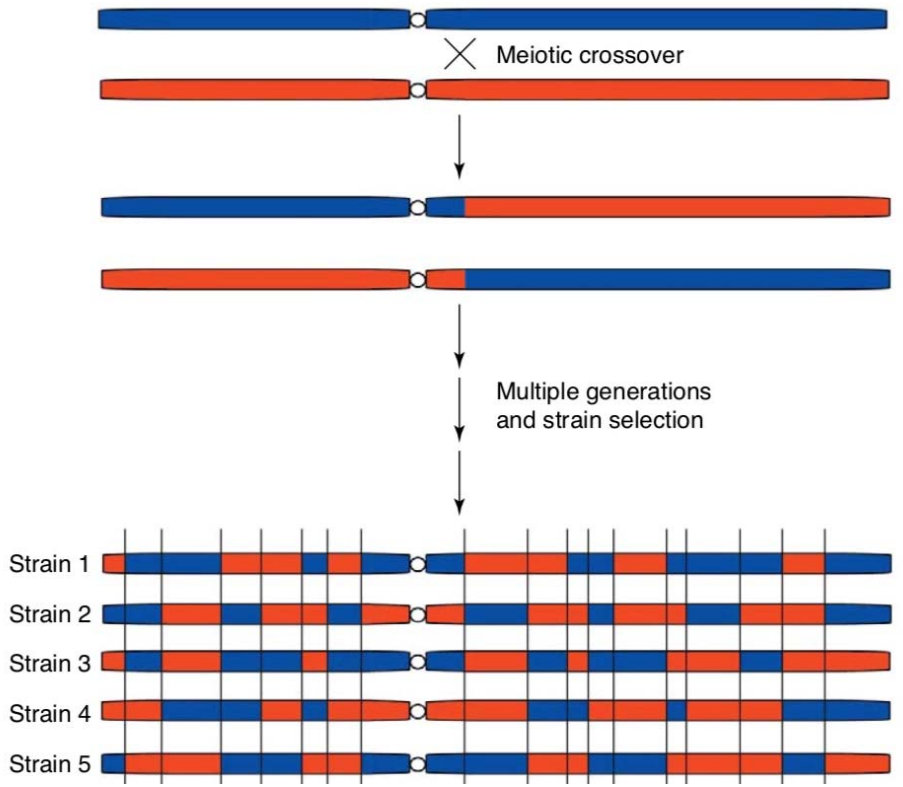
Dimensionality of genomic information (ii)

- Rank of **G** or **Z'Z** ≤ min(N_{SNP}, N_{IND}, Me)
  - N_{SNP} – number of SNPs
  - N_{IND} – number of genotyped animals
  - Hypothesis: Me – number of independent chromosome segments

Independent chromosome segments

- $E(Me) = 4N_eL$ Stam (1980)
  - Me – Independent chromosome segments
  - N_e – Effective population size
  - L – Length of genome in Morgans

- Me
  - $2N_eL$ Hayes *et al.* (2009)
  - $2N_eL/[\log(N_eL)]$ Goddard *et al.* (2011)
  - Many more Brard and Ricard (2015)



Cuppen (2005)

Objectives

- Test the limited dimensionality of genomic information:
  - Simulated idealized populations with variation in Ne
  - Livestock populations (different species and breeds)

- Establish connection of dimensionality with Ne and accuracy of GS

Data: Simulation

- 6 populations (Ne=20,40,80,120,160,200)
- 10 discrete generations taken from historical population
- Mating random / no selection
- Generations 8-10 genotyped (75,000 individuals)
- Validation: Gen. 10 (no phenotypes)
- 30 Chromosomes (100 cM each)
- 49,980 SNPs
- 4,980 QTLs
- Heritability = 0.3
- 5 replications
  - QMSim software (Sargolzaei and Schenkel, 2009)

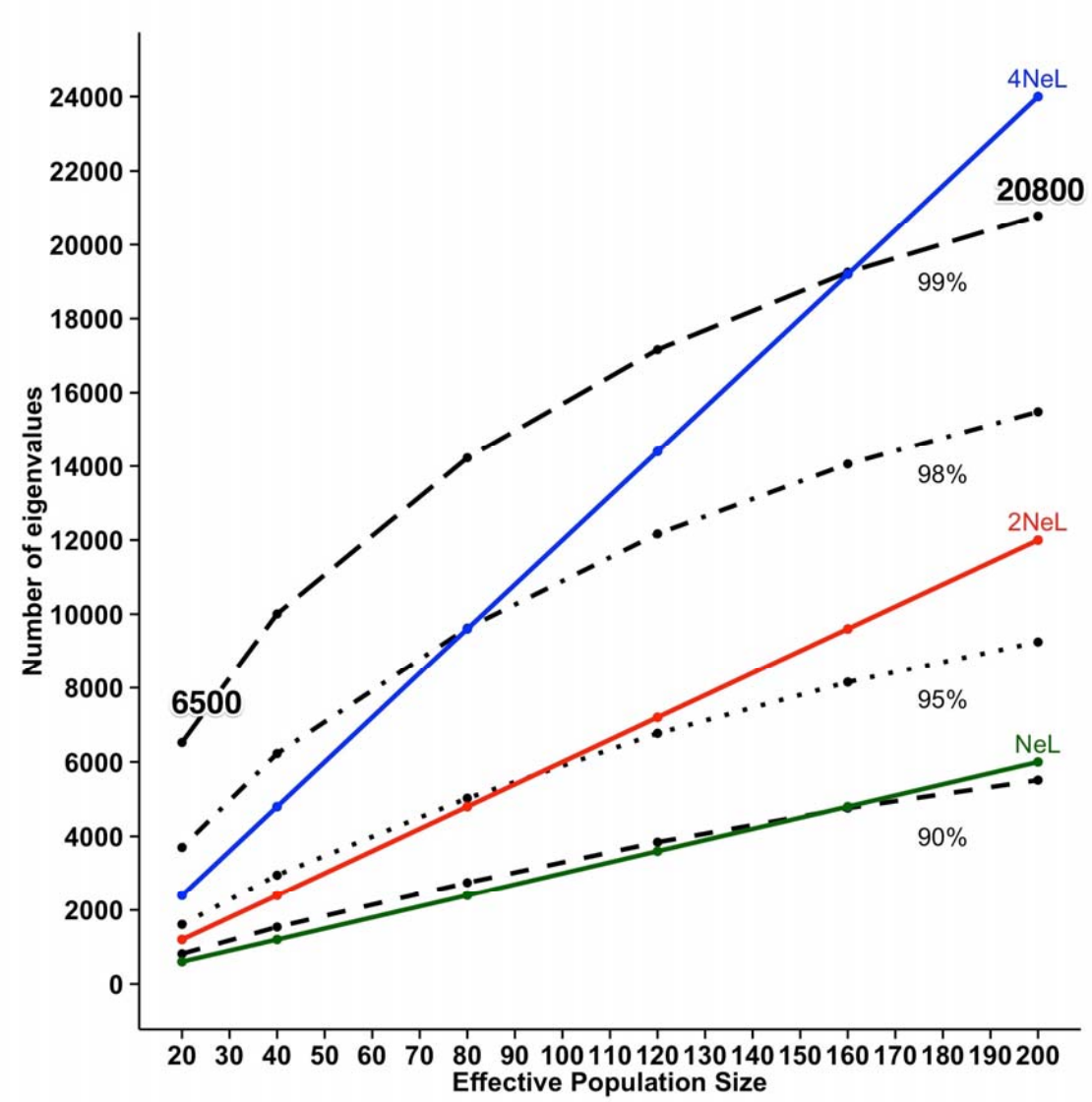
Data: Livestock populations

Species / Breed	Traits	Animals	Genotyped	SNP
Holstein	FS	10.7 M	77 k	61 k
Jersey	MY, FY, PY	2.4 M	75 k	61 k
Angus	BW, WW, PWG	8.2 M	81 k	38 k
Pig	LS, SB	2.4 M	23 k	37 k
Chicken	BWG, RFI, BMP, WGT	199 k	16 k	39 k

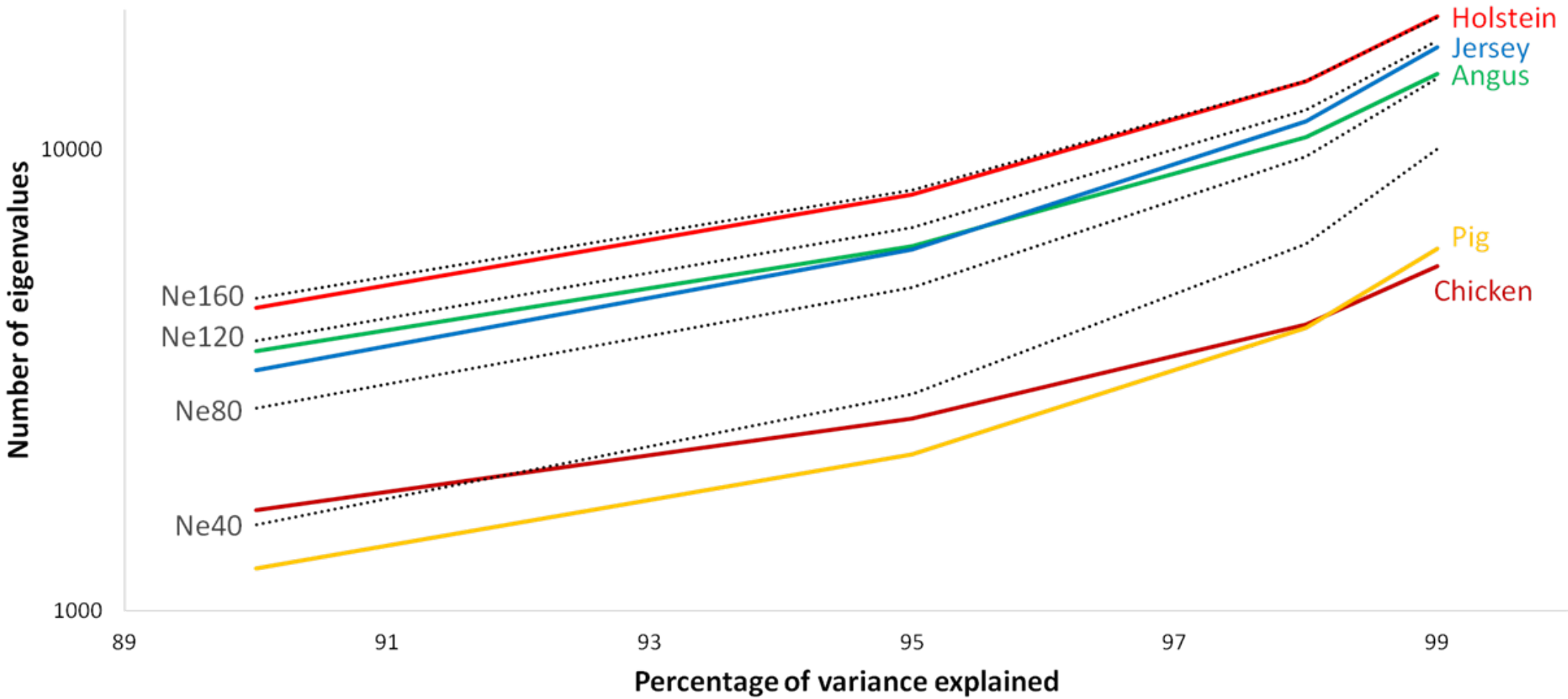
Models and calculations

- Construct $\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\Sigma p_j(1-p_j)}$ (VanRaden, 2008)
- Find number of the largest eigenvalues in $\mathbf{G_0}$ that explained 90, 95, 98 and 99 percent of variance
- Calculate GEBV via ssGBLUP:
  - Regular inverse of GRM
  - Algorithm for Proven and Young (APY) inverse of the GRM (Misztal et al., 2014; Masuda et al., 2016)
- Core animals for APY were randomly selected based on eigenvalues
- Accuracies calculated as cor(TBV,GEBV) in simulations, forward prediction or predictability for livestock populations

Number of largest eigenvalues to account for a given variance



Numbers of largest eigenvalues that explain a given percentage of variance

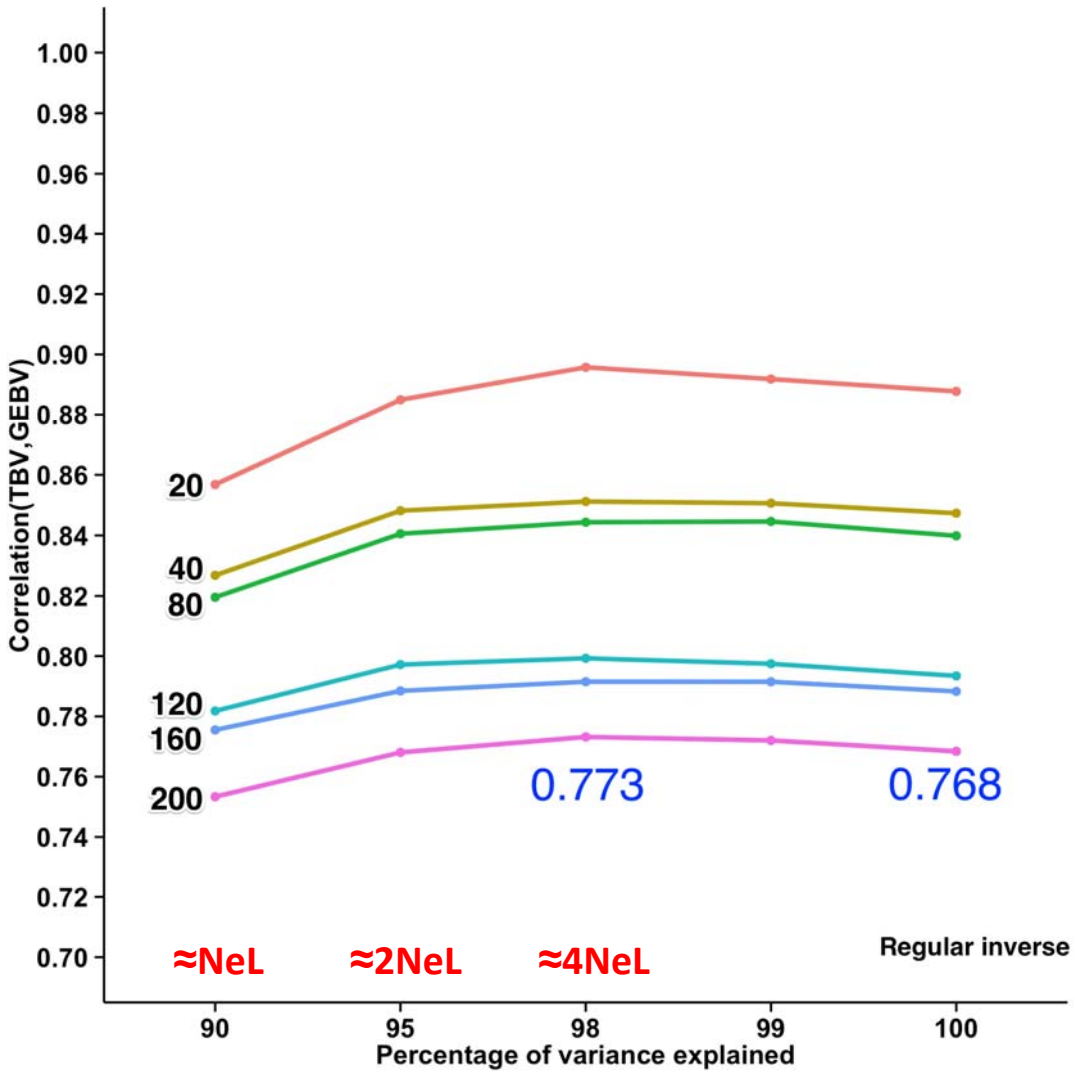


Optimal SNP Chip Size and estimated Ne

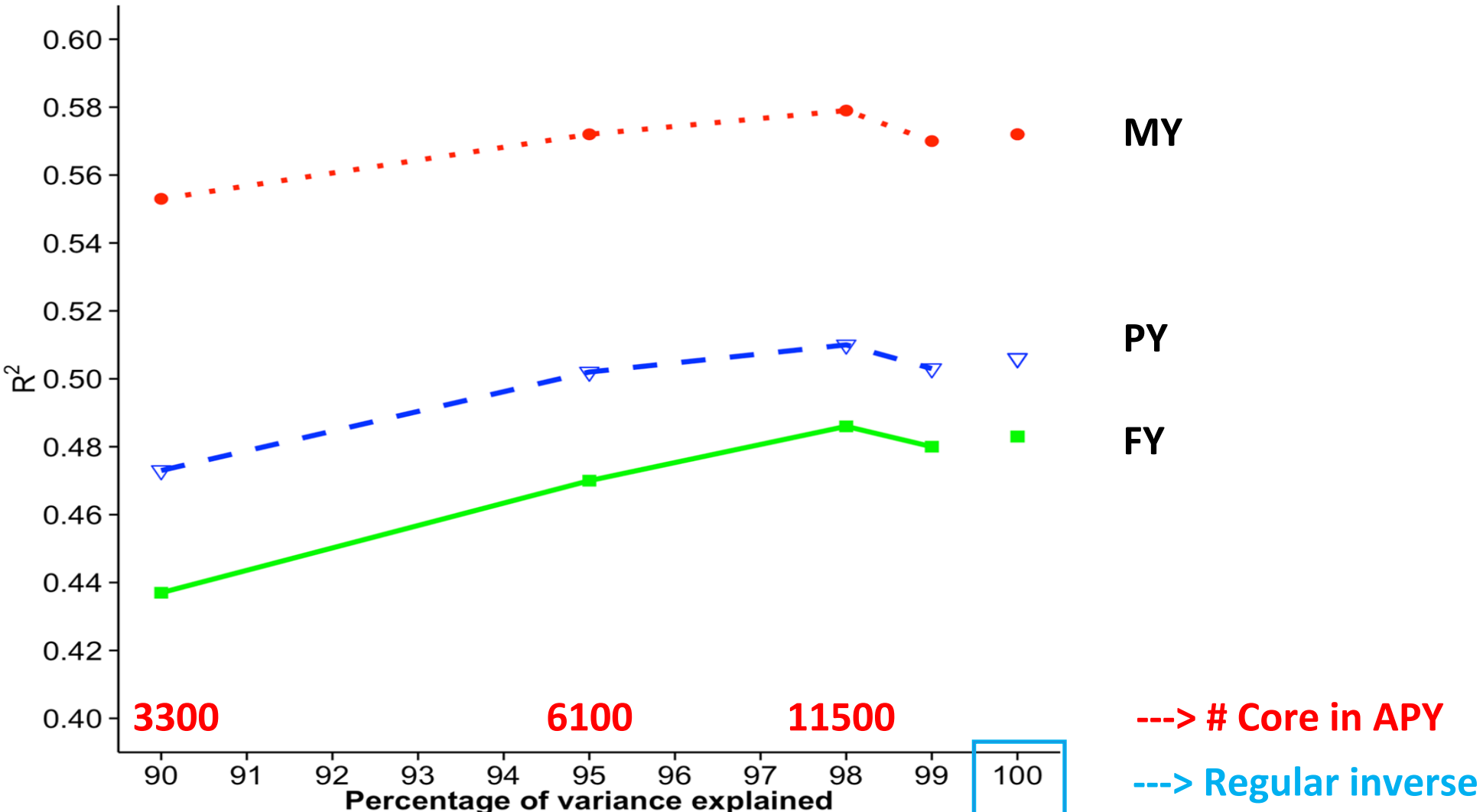
Ne	Dimensionality at 98%	Optimal chip density	Actual chip used
48 (Pig)	4.1 k	49 k	37 k
44 (Chicken)	4.2 k	50 k	39 k
101 (Jersey)	11.5 k	138 k	61 k
113 (Angus)	10.6 k	127 k	38 k
149 (Holstein)	14 k	168 k	61 k

12 times number of segments
(MacLeod et al., 2005)

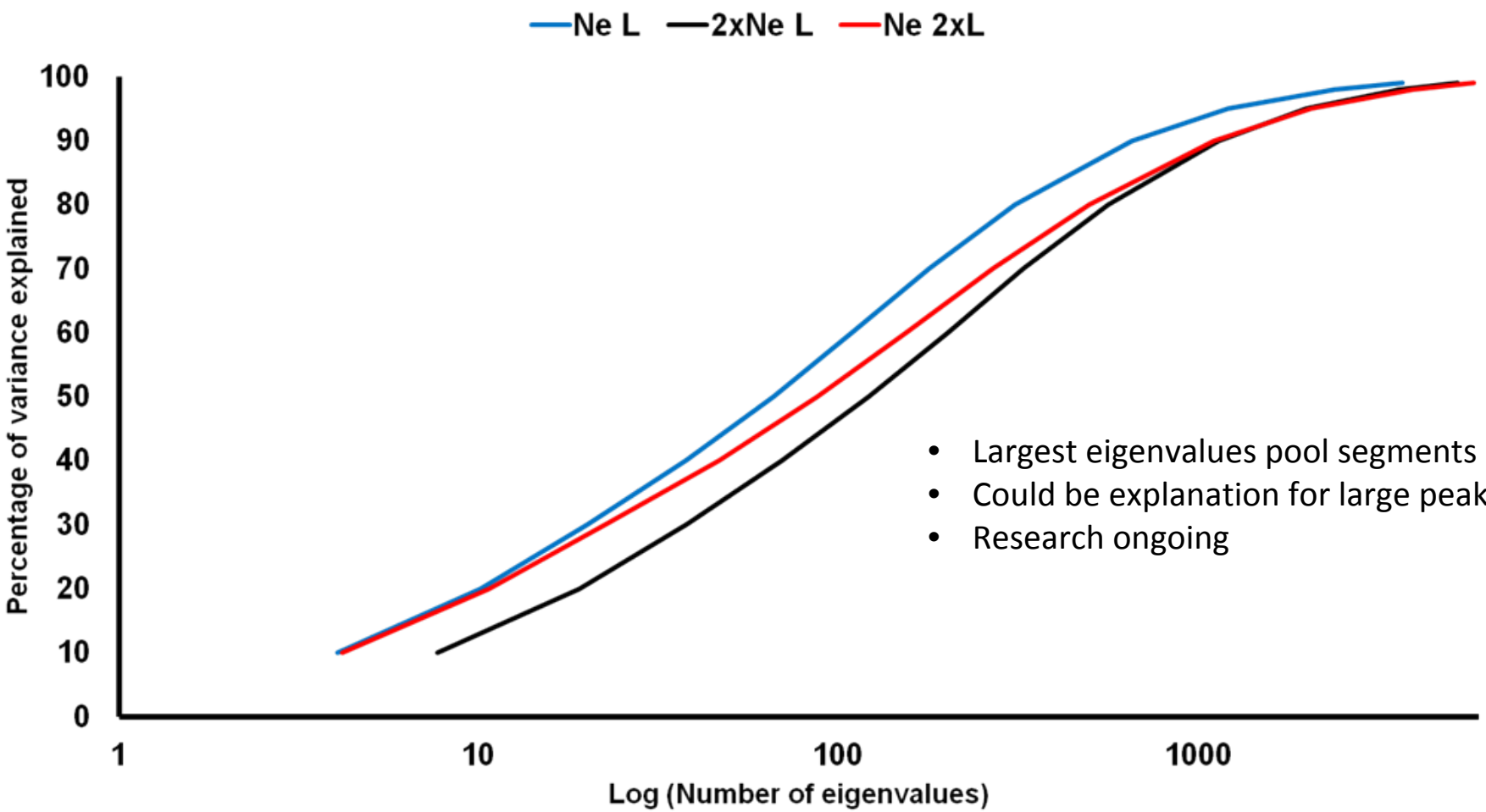
Accuracies as function of number of core animals



Coefficients of determination(R^2) for Jersey



Eigenvalue decomposition for different Ne and L



- Largest eigenvalues pool segments across genome
- Could be explanation for large peaks with BayesB
- Research ongoing

Conclusions

- Dimensionality of genomic information is limited, and function of Ne
- Eigenvalues at 98% correspond to approximately 4NeL segments
- Enables computation of optimal SNP chip size and Ne
- Maximizes accuracy in APY (low-cost GS possible for millions animals)
- Future research: impact on GWAS

Thank you !!!