Deriving dimensionality of genomic information from limited SNP information

Ivan Pocrnić, D.A.L. Lourenco & I. Misztal

UNIVERSITY OF GEORGIA

Department of Animal and Dairy Science, Athens GA, USA

EAAP, Dubrovnik HR, August 2018

Background

- Recently great interest in whole-genome sequences and causative variants
- Cattle: 3 billion base pairs; around 30 million SNP markers
- Problems:
  - This is very big data
  - Do we need to use all of that?
- Actually; genomic information redundant and limited

How to calculate dimensionality ?

- **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

- Rank of **G** or **Z'Z** ≤ min($\#_{SNP}$, $\#_{IND}$, $\#_{Me}$)

Dimensionality in livestock species

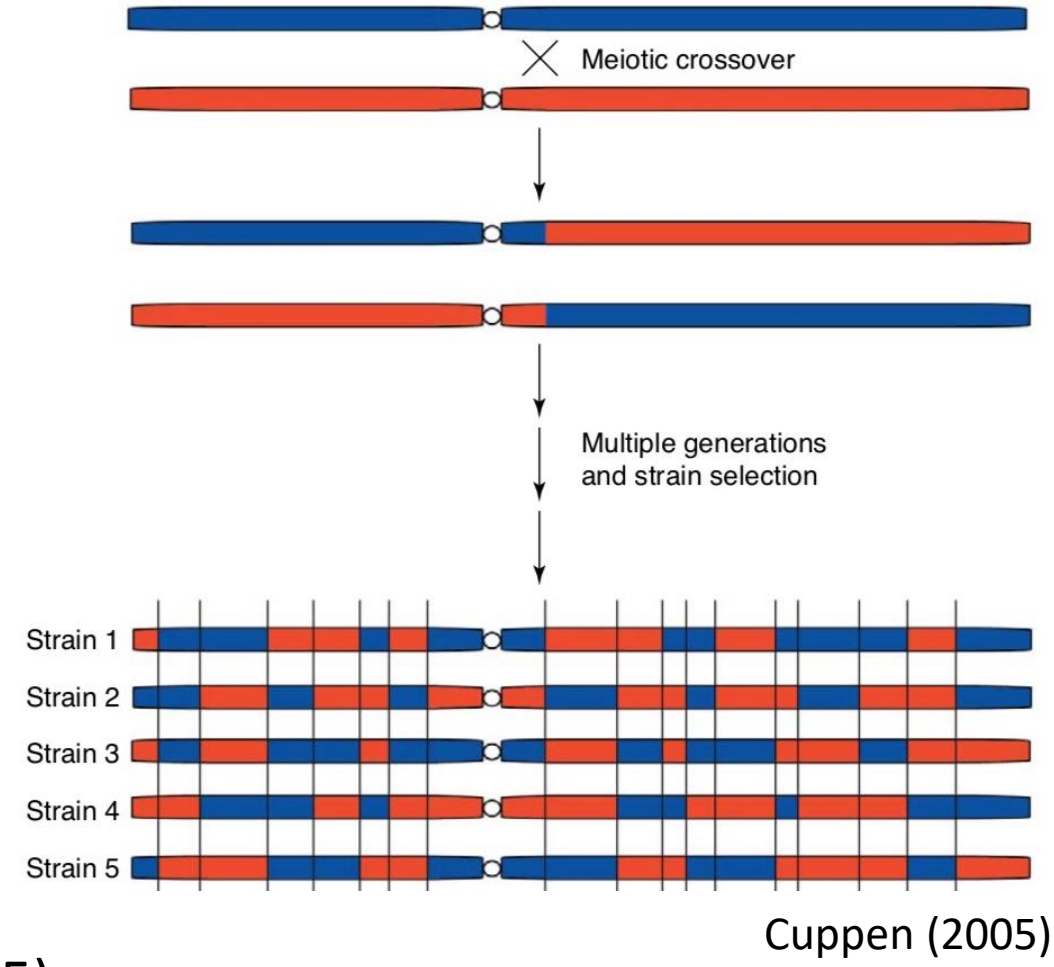
Ne	Dimensionality at 98%	Genotyped	SNP
48 (Pig)	4.1 k	23 k	37 k
44 (Chicken)	4.2 k	16 k	39 k
101 (Jersey)	11.5 k	75 k	61 k
113 (Angus)	10.6 k	81 k	38 k
149 (Holstein)	14 k	77 k	61 k

- Dimensionality is **5-15k** – not **30M**
- If we estimate 5-15k segments exactly = perfect prediction of GEBV
- Using more than 98% can slightly decrease accuracy (e.g. in APY)

Independent chromosome segments

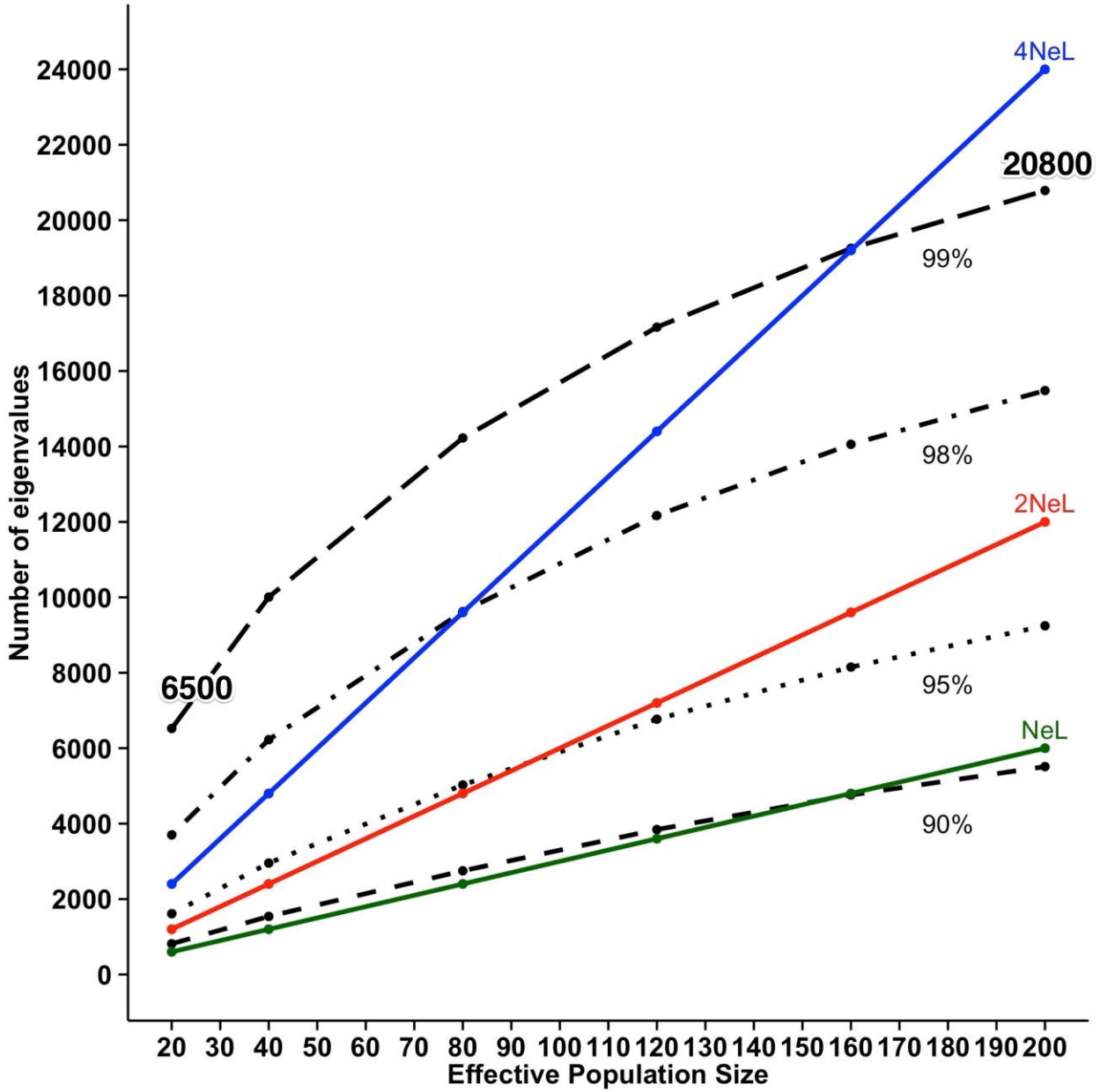
- E(Me) = $4N_eL$ Stam (1980)
  - Me – Independent chromosome segments
  - N_e – Effective population size
  - L – Length of genome in Morgans

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



Cuppen (2005)

Number of largest eigenvalues to account for a given variance



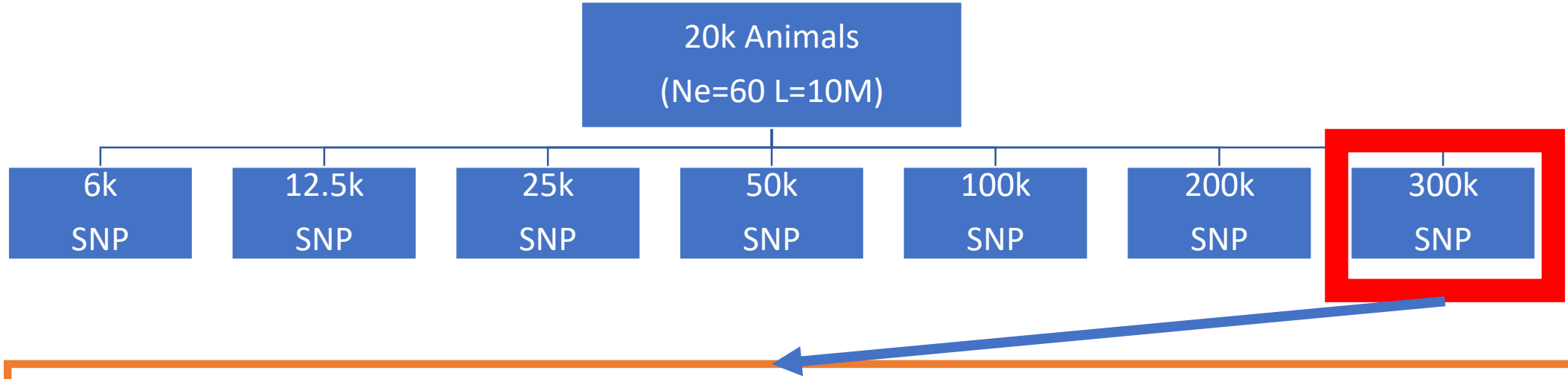
Pocrnic *et al.* (2016)

Objectives

- To find dimensionality for large population that has small SNP chip and/or small number of genotyped animals
- To derive approximate formulae for Ne and L

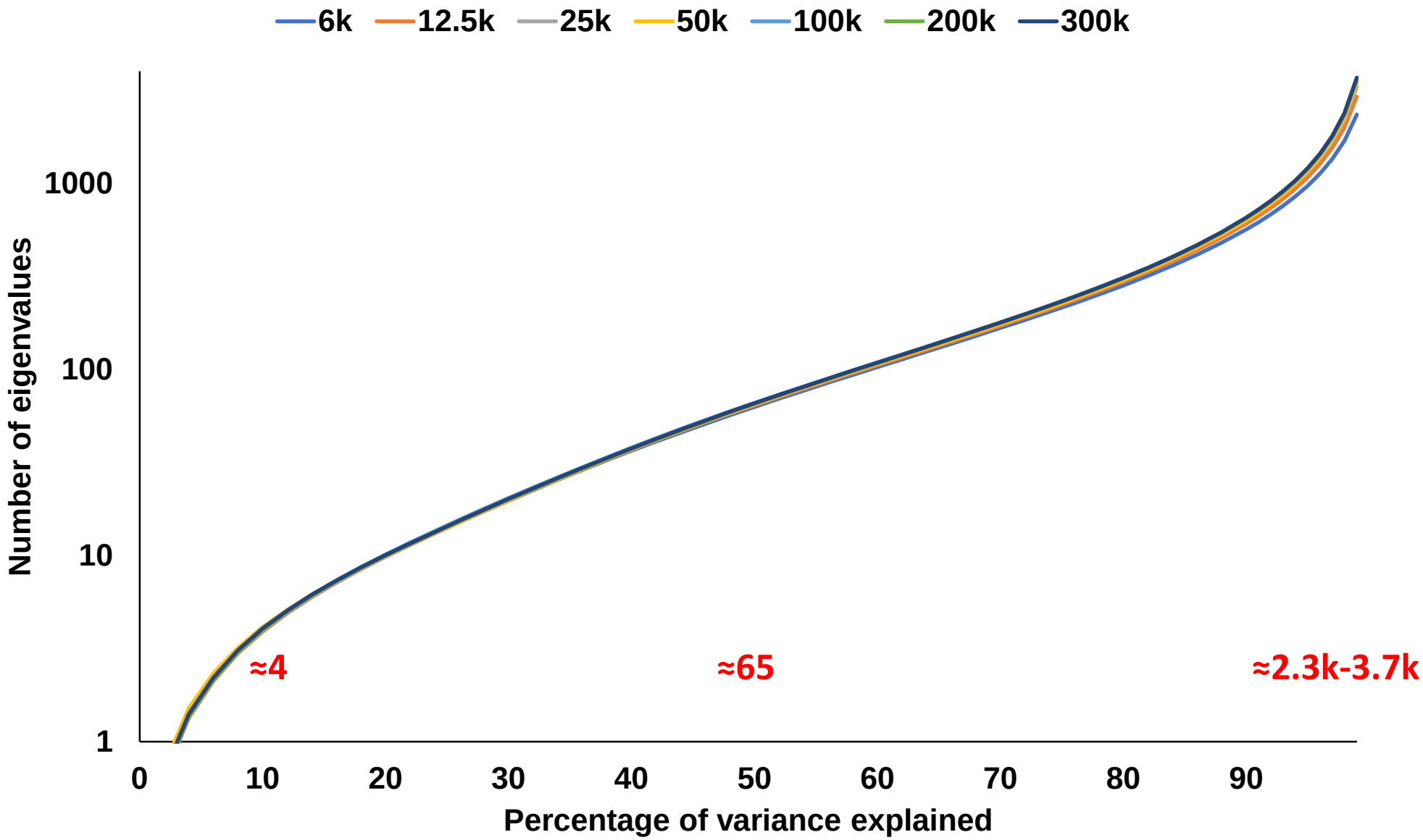
Data: Simulation

- 10 Chromosomes
- 3000 QTL
- Historical population + 10 recent generations (9 and 10 genotyped)
- 5 replications
- QMSim software (Sargolzaei and Schenkel, 2009)
- Additionally tested with huge maize and pig datasets (Not shown)

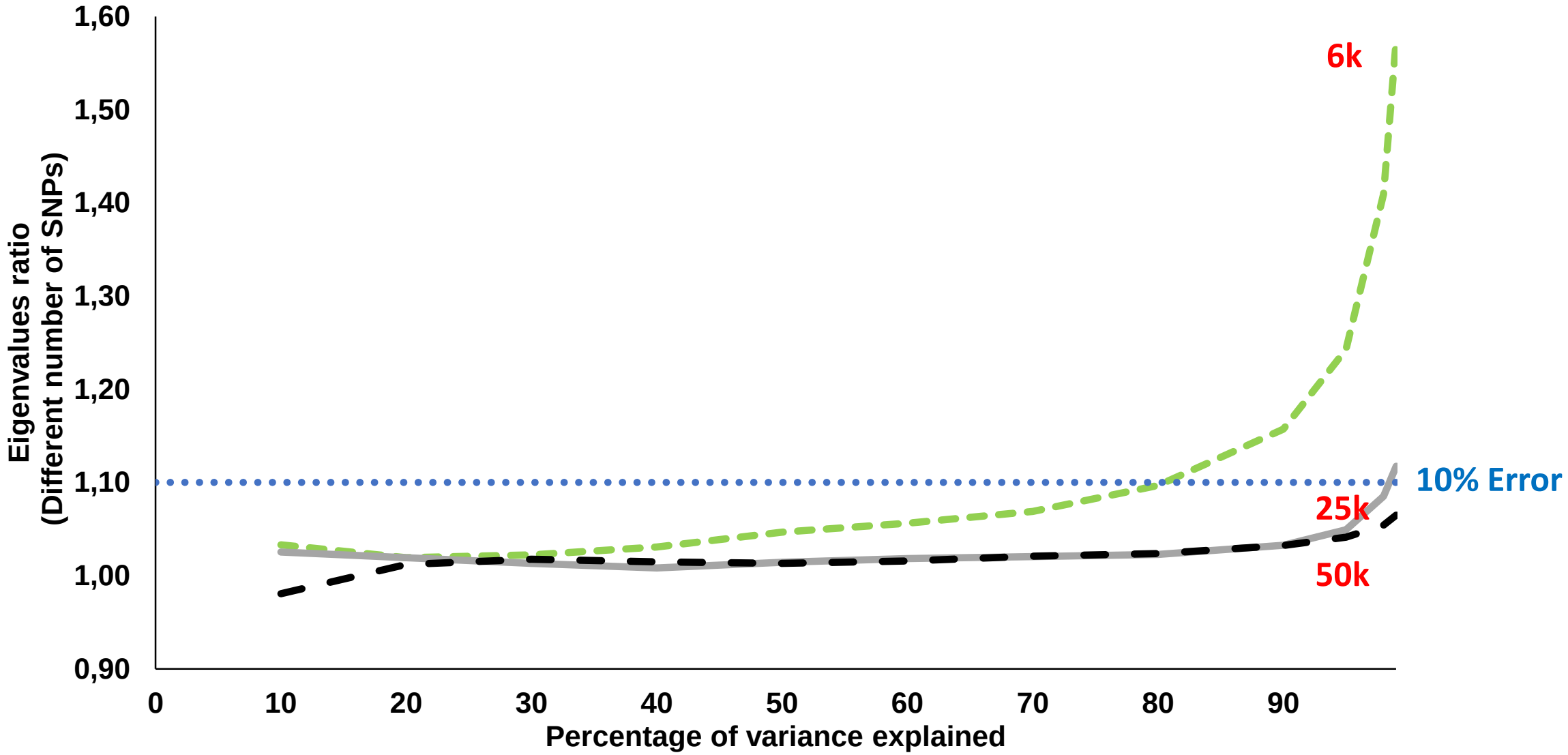


- Change number of genotyped animals: 10k, 5k, 2k, 1k
- Change genome length: 20M and 40M
- Change Ne and genome length: Ne=120 L=10; Ne=120 L=20; Ne=120 L=50

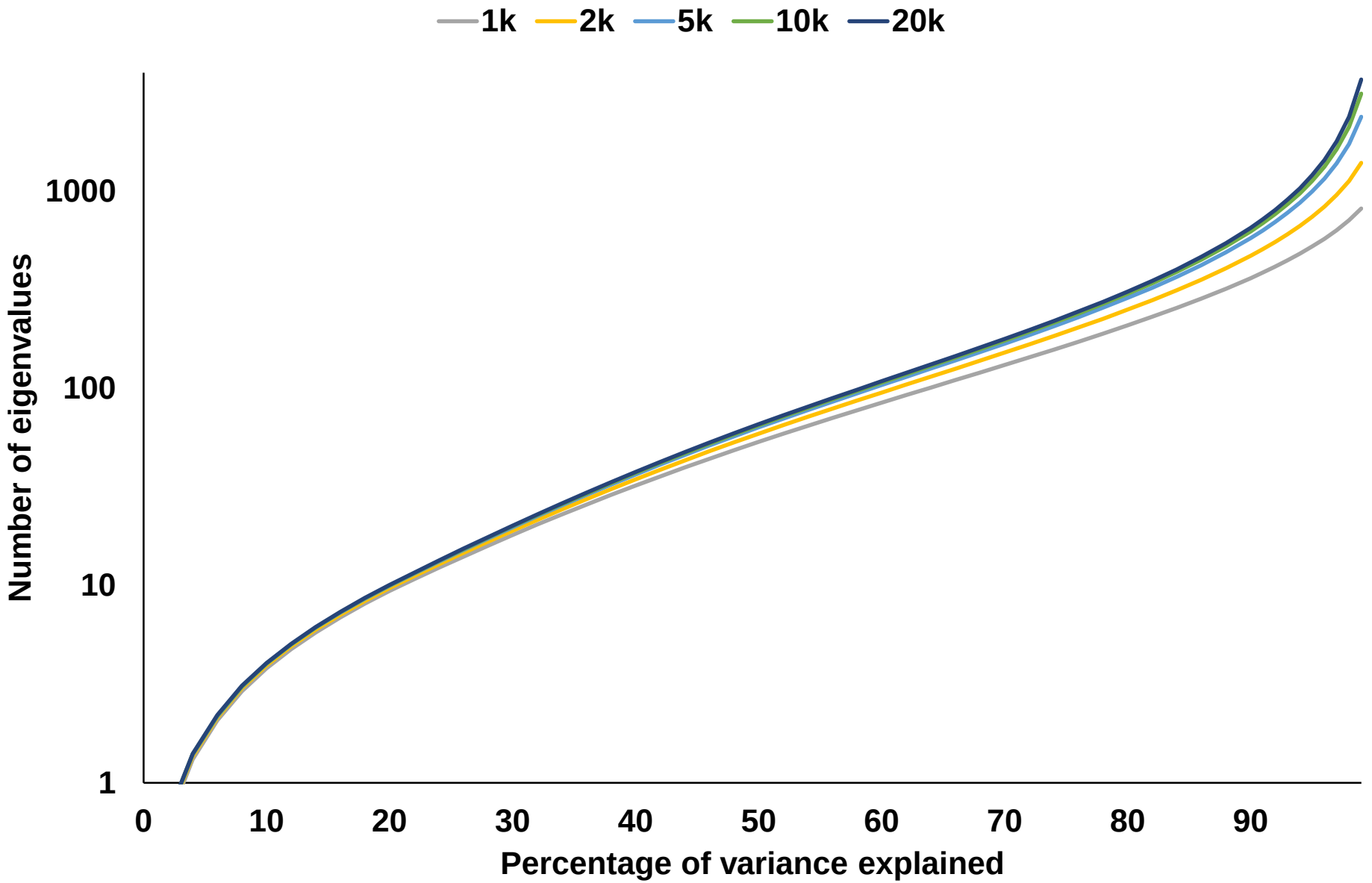
Profile plot: Variation of #SNP



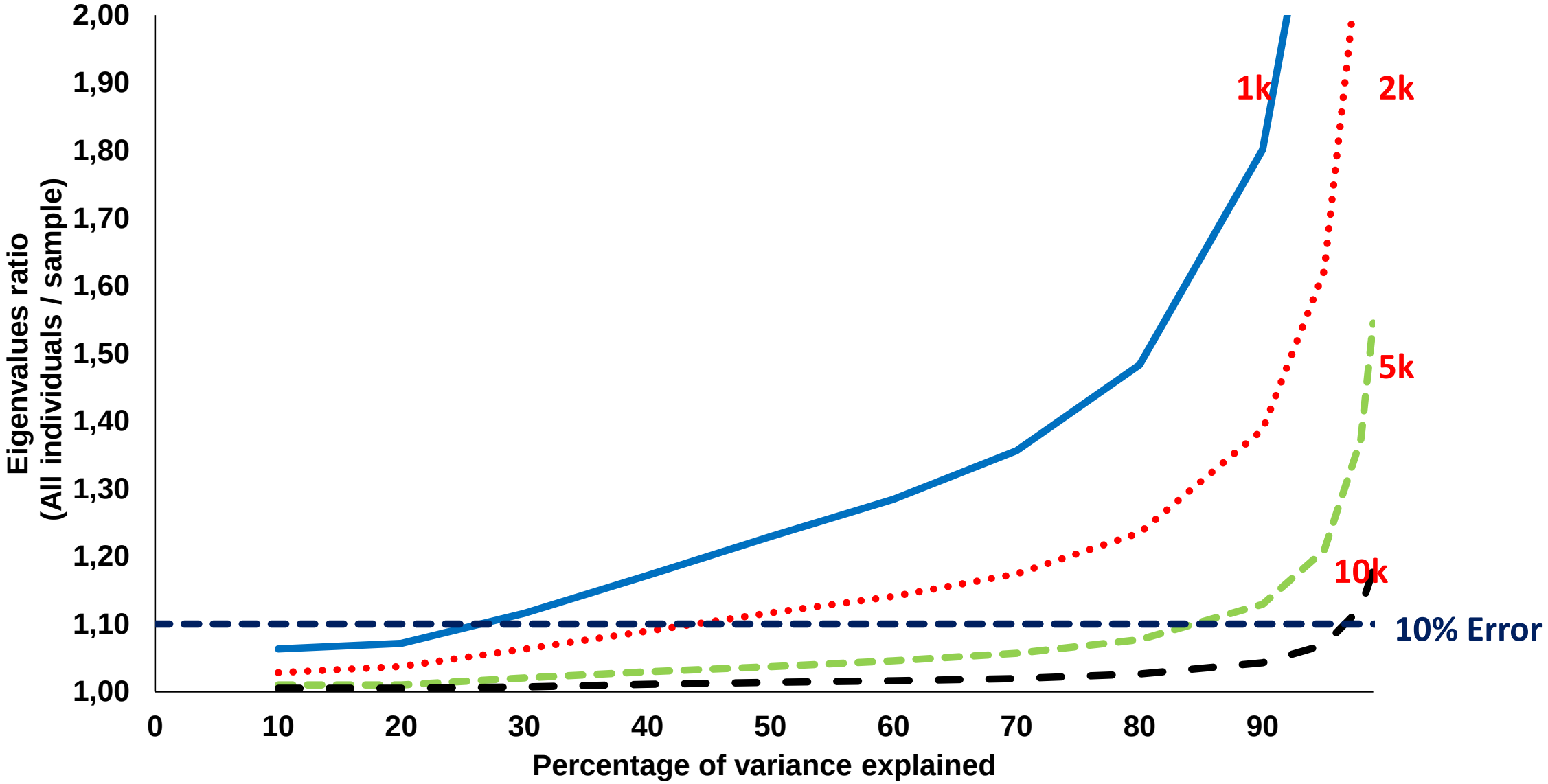
Variation of #SNP relative to 300k SNP



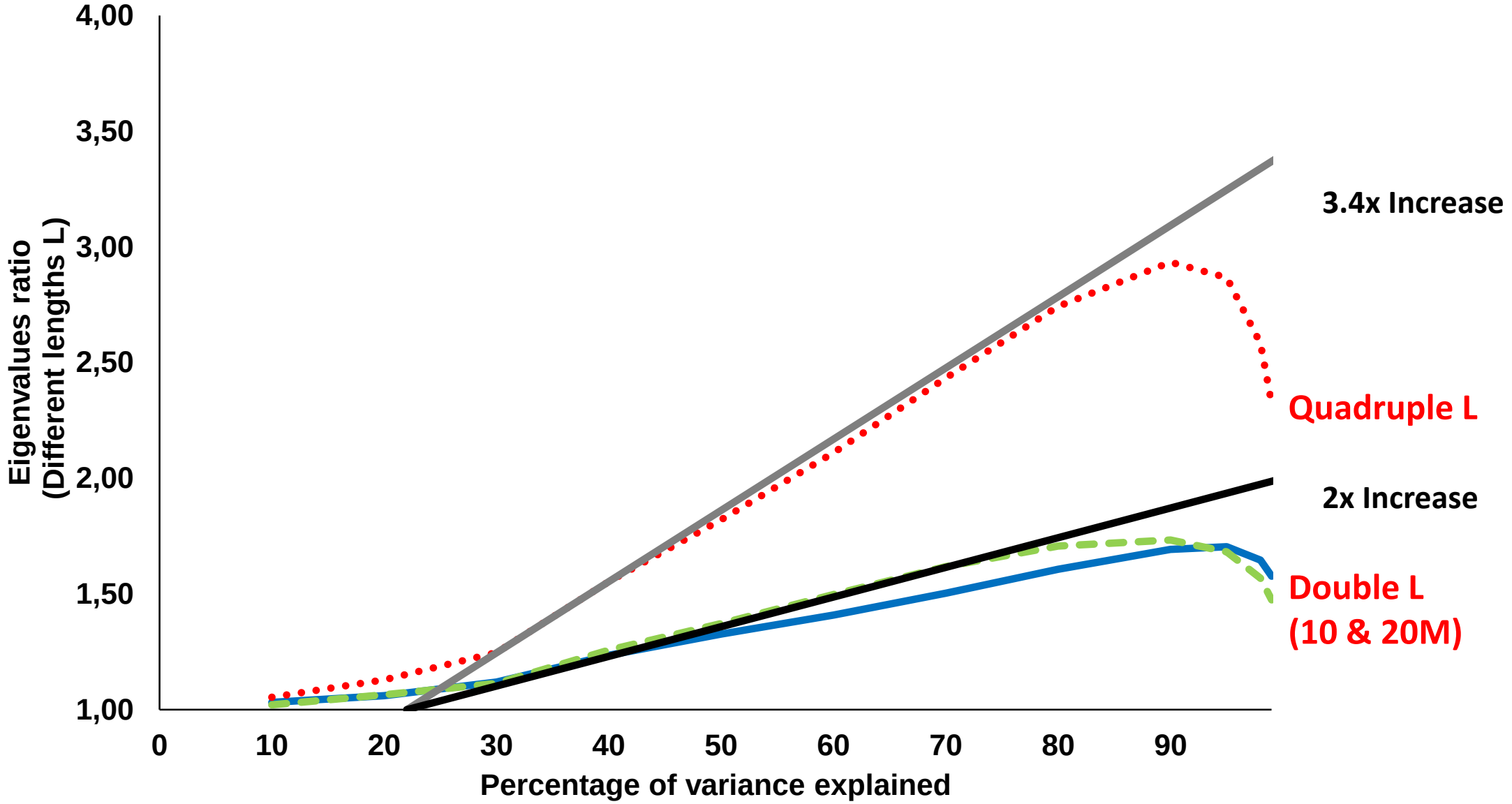
Profile plot: Variation of #Genotyped animals



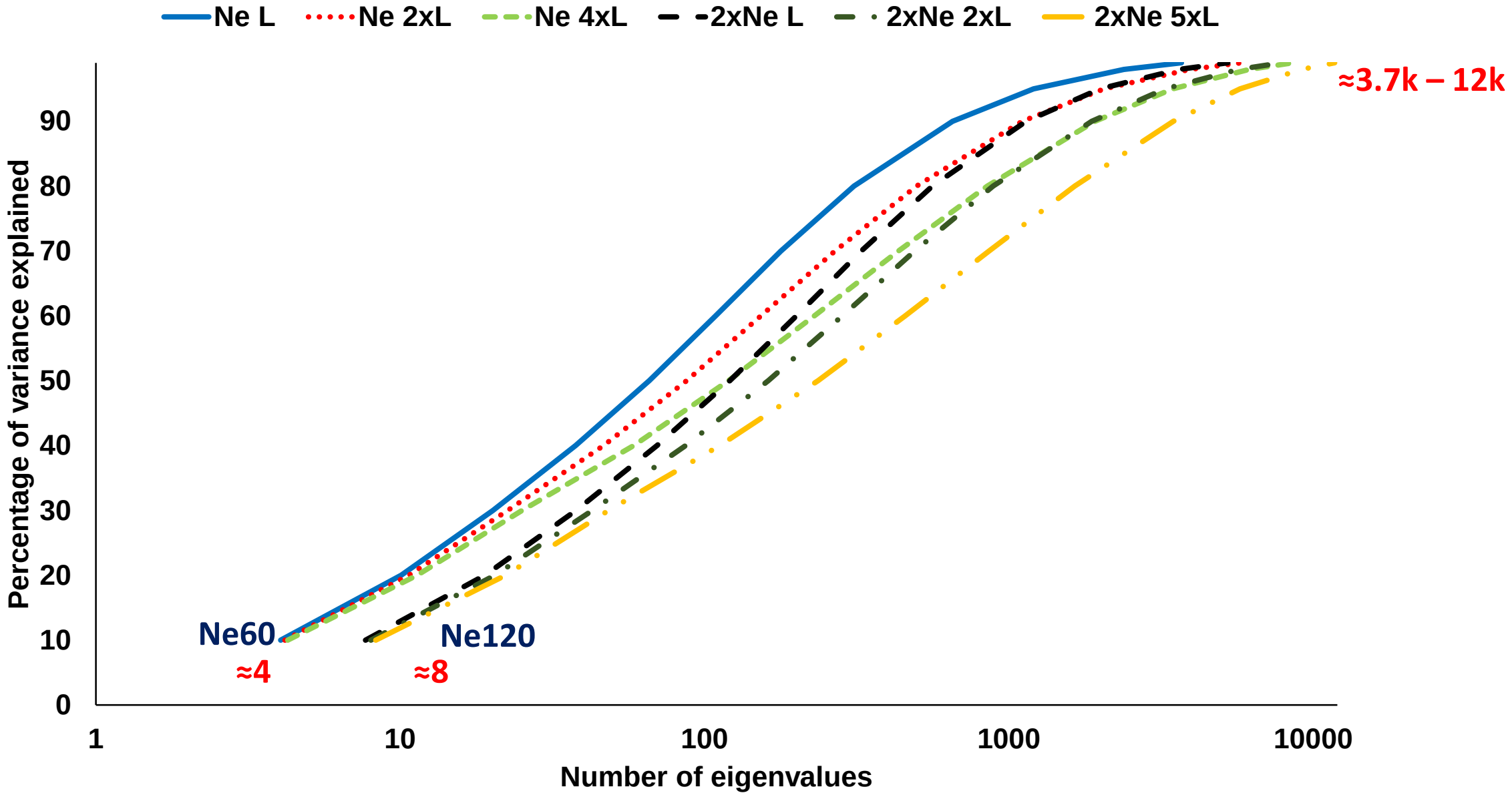
Variation of #Genotyped relative to 20k animals



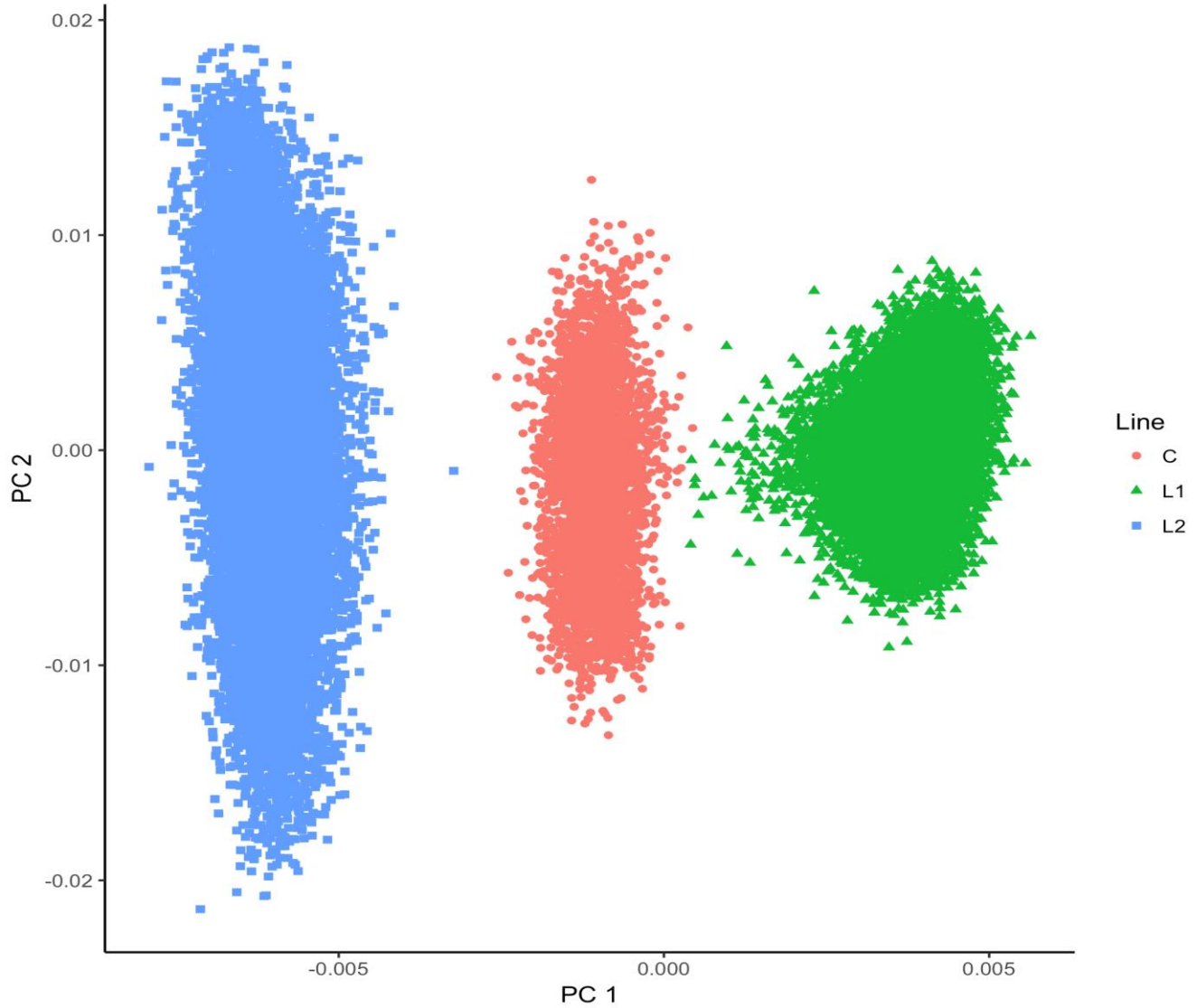
Variation of genome length relative to 10M



Profile plot: Variation of #NeL



First two components show population structure



Approximate formulae

- $Ne \approx 15*(Eig10\%)$
- $L \approx 10*[(Eig50\%/Eig10\%)/16]^{\log(2)/\log(1.35)}$
- $L \approx 10*[(Eig80\%/Eig10\%)/77]^{\log(2)/\log(1.65)}$
- $4NeL \approx 600*(Eig10\%)*[(Eig50\%/Eig10\%)/16]^{\log(2)/\log(1.35)}$
- $4NeL \approx 600*(Eig10\%)*[(Eig80\%/Eig10\%)/77]^{\log(2)/\log(1.35)}$

- Empirical formulae - population similar to this simulation
- Deviation from reality (e.g. selection, genotyping errors, etc.)

Conclusions

- Few largest eigenvalues account for many segments across genome
- Smallest eigenvalues account for individual segments
- We can predict dimensionality using small datasets
- At least 1k genotyped animals and (15xEIG%) SNP to accurately estimate segments
- Approximate formulae for population parameters can be derived

Thank you !!!