Limited dimensionality of genomic information and implications on genomic selection

Ivan Pocrnić[1,2],
Daniela Lourenco[1] ,Gregor Gorjanc[2], Ignacy Misztal[1]

[1]University of Georgia, USA [2]University of Edinburgh, UK

@IvanPocrnic @HighlanderLab

Limited dimensionality of genomic information and implications for genomic selection

AIMS & CONCLUSIONS

- Dimensionality of genomic information is limited and is a function of effective population size (Ne)
- It can be estimated by a matrix decomposition of a genomic relationship matrix or linkage-disequilibrium matrix
- Genomic selection utilizes genomic dimensions (~clusters of independent chromosome segments)
  - For moderate to high accuracy we need few genomic dimensions
  - For very high accuracy we need many genomic dimensions
- Limited dimensionality enables computationally efficient algorithms

INTRODUCTION

How to utilize genomic information?

- Genome Wide Association Studies; GWAS
- Genomic Selection; GS (e.g., GBLUP, SNP-BLUP, …)

Recent trends

- Massive SNP array datasets a new norm
- Whole-genome sequencing and causative variants

Big data

- Cattle genome: 3 billion bp and 30 million SNP
- US Holstein: around 3 million genotyped individuals

METHODS & MATERIAL

How to measure genomic dimensionality?

- Inheritance is chunkular (in blocks)

- Estimating junctions/segments/blocks (Me) is cryptic:
  - Theory of Junctions (Fisher, 1949)
  - E(Me) = 4NeL (Stam, 1980) L-genome length in Morgan
  - Many formulae exist (Review by Brard and Ricard, 2015)

- Alternative:

  Z – matrix of gene content

  

  Singular value decomposition: **Z = U D V'**

- Applied to:

  Genomic relationship matrix (GBLUP): **G = (ZZ'/k) = UDDU'**

  Linkage-disequilibrium matrix (SNP-BLUP): **Z'Z = V'DDV**

- Genomic information (**Z**, **ZZ'** or **Z'Z**) has limited dimensionality
- Use only a limited number of genomic dimensions by retaining dimensions with the largest eigenvalues

Data:

- US Holstein, US Jersey, US Angus, Broiler Chicken, Commercial Pig lines
- Validation based on the accuracy of GS

RESULTS

Species	Genotyped (k)	SNP (k)	Dimensionality (k)	Ne
Pig	23	37	4.1	48
Chicken	16	39	4.2	44
Jersey	75	61	11.5	101
Angus	81	38	10.6	113
Holstein	77	61	14.0	149

- Capturing 98% of variance in **G** seems to be a good measure of limited dimensionality
  - Point that maximizes the accuracy of GS
  - 1-2% of variance in **G** is noise
  - 98% of variance in **G** corresponds to 4NeL segments
- Allows application of efficient algorithms for big data
  - The Algorithm for Proven and Young; APY (Misztal, 2016)
  - GS possible for millions of genotyped individuals
  - No. of eigenvalues that explain certain %Var in **G** is used as the no. of core animals in the APY
- Possibility to calculate optimal SNP array size
  - 12x no. of segments (MacLeod et al., 2005)

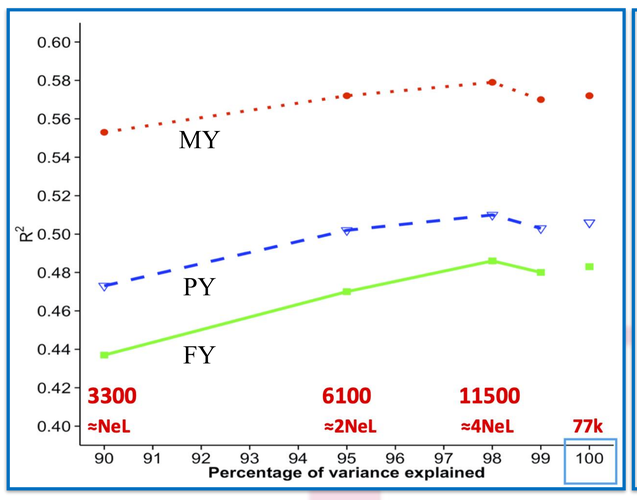

Figure 1. Reliability of GS for milk (MY), protein (PY), and fat yield (FY) in US Jersey.

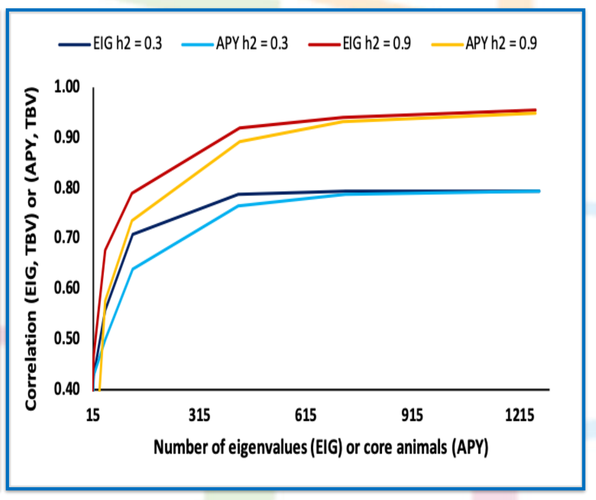

Figure 2. Accuracy of simulated GS based on APY or eigenvalue based **G.** Note the steep but diminishing returns in accuracy with the number of genomic dimensions.

FUTURE STUDIES

- Associate genomic dimensions and underlying genome segments to actual haplotypes
- Study the impact of genomic dimensions on GWAS results
- Gain better understanding of mechanisms driving the accuracy of GS