Crossbred evaluations using ssGBLUP and algorithm for proven and young with distinct sources of data

Ivan Pocrnić,
D.A.L. Lourenco, C.Y. Chen, W.O. Herring, I. Misztal

UNIVERSITY OF GEORGIA

PIC®

EAAP, Ghent BE, August 2019

Problem statement

- Pig breeding very specific structure
- Selection in nucleus
- Combining several lines is common

1. Is there predictivity across and within the lines and their crosses

2. Can we use Algorithm for Proven and Young (APY) for multiple lines / crossbreed datasets

3. How to quantify overlapping chromosome segments in these lines

Data

- PIC (Genus) purebred lines (**L1** and **L2**) and F1 cross (**C**)
- 2 Traits ($h^2 \approx 0.3$)
- Pedigree 727.3k
- 43.5k SNP markers and 46.5k genotyped animals

	Line 1	Line 2	Cross
Trait 1	180k	25.3k	5.6k
Trait 2	178.8k	25k	5.4k
Genotypes	26.5k	15.9k	3.9k

Statistical models

- **M1 – All lines joint**

$$\mathbf{y}_t = \mathbf{X}_t\mathbf{b}_t + \mathbf{Z}_t\mathbf{u}_t + \mathbf{W}_t\mathbf{c}_t + \mathbf{e}_t$$

$$\mathrm{Var}(\mathbf{u}) = \begin{bmatrix} \sigma^2_{uT1} & \sigma_{uT1,uT2} \\ \sigma_{uT2,uT1} & \sigma^2_{uT2} \end{bmatrix} \otimes \mathbf{H},$$

$$\mathrm{Var}(\mathbf{c}) = \begin{bmatrix} \sigma^2_{pT1} & 0 \\ 0 & \sigma^2_{pT2} \end{bmatrix} \otimes \mathbf{I},$$

$$\mathrm{Var}(\mathbf{e}) = \begin{bmatrix} \sigma^2_{eT1} & \sigma_{eT1,eT2} \\ \sigma_{eT2,eT1} & \sigma^2_{eT2} \end{bmatrix} \otimes \mathbf{I}.$$

- **M2 – Each line as different trait**

$$\mathbf{y}_l = \mathbf{X}_l\mathbf{b}_l + \mathbf{Z}_l\mathbf{u}_l + \mathbf{W}_l\mathbf{c}_l + \mathbf{e}_l$$

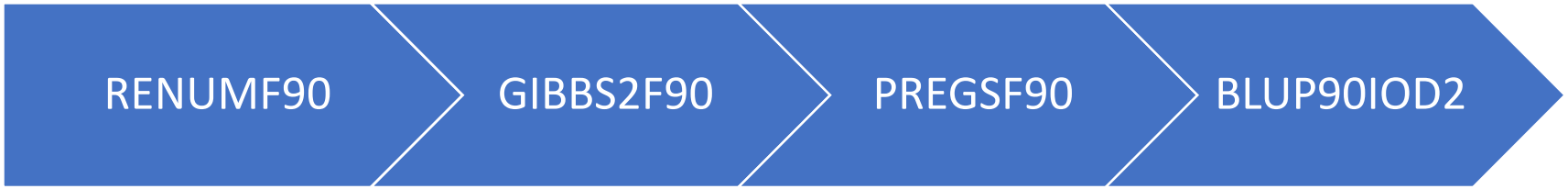
$$\mathrm{Var}(\mathbf{u}) = \begin{bmatrix} \sigma^2_{uL1} & \sigma_{uL1,uL2} & \sigma_{uL1,uC} \\ \sigma_{uL2,uL1} & \sigma^2_{uL2} & \sigma_{uL2,uC} \\ \sigma_{uC,uL1} & \sigma_{uC,uL2} & \sigma^2_{uC} \end{bmatrix} \otimes \mathbf{H},$$

$$\mathrm{Var}(\mathbf{c}) = \begin{bmatrix} \sigma^2_{pL1} & 0 & 0 \\ 0 & \sigma^2_{pL2} & 0 \\ 0 & 0 & \sigma^2_{pC} \end{bmatrix} \otimes \mathbf{I},$$

$$\mathrm{Var}(\mathbf{e}) = \begin{bmatrix} \sigma^2_{eL1} & 0 & 0 \\ 0 & \sigma^2_{eL2} & 0 \\ 0 & 0 & \sigma^2_{eC} \end{bmatrix} \otimes \mathbf{I}.$$

Genomic setup and computational details

BLUPF90 Software Family



- **Genomic relationship matrices:**

$$\mathbf{G}_0 = \mathbf{MM}' / 2\Sigma p_j(1 - p_j)$$

VanRaden (2008)

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Aguliar *et al.* (2010)

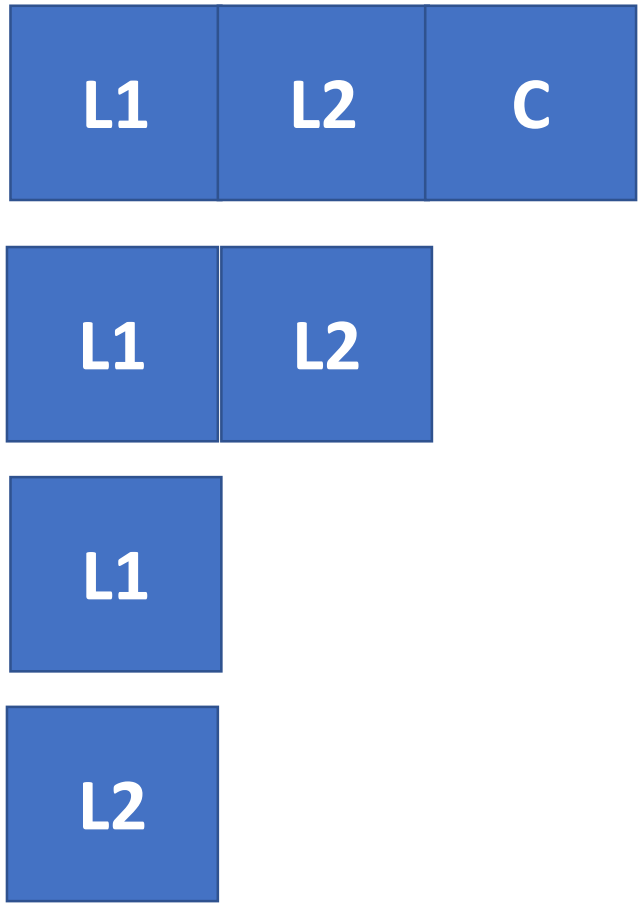
$$\mathbf{G} = 0.95\mathbf{G}_0 + 0.05\mathbf{A}_{22}$$

- **Direct or APY inverse of GRM**

APY; Misztal *et al.* (2014)

Scenarios

- **Available phenotypes:**



- **Available genotypes:**

- **Core animals APY:**

- Eigenvalues - 90, 98, or 99 % variance of **G**
- Randomly selected
  - From L1
  - From L2
  - From L1, L2 and C

Validation

- Genotyped animals born in 2017 with phenotypes removed
  - 2770 - L1
  - 2623 - L2
  - 2557 - C

- Accuracy
  - corr (y-Xb, GEBV)
  - GEBV based on either direct or APY inverse

- Inflation
  - $(y-Xb) = b_0 + b_1 GEBV + e$

- Corr (GEBV_apy, GEBV_direct)

Predictive ability – Trait 1

Phenotypes

	MODEL 1			MODEL 2		
	L1	**L2**	**C**	**L1**	**L2**	**C**
L1 + L2 + C	0.33	0.34	0.26	0.33	0.35	0.24
L1 + L2	0.33	0.34	0.26			
L1	0.33	0.15	0.19	0.33	0.15	0.19
L2	0.18	0.35	0.21	0.18	0.35	0.20

Validation subset

Predictive ability – Trait 2

Phenotypes

	MODEL 1			MODEL 2		
	L1	**L2**	**C**	**L1**	**L2**	**C**
L1 + L2 + C	0.24	0.36	0.25	0.25	0.38	0.22
L1 + L2	0.24	0.36	0.25			
L1	0.25	0.14	0.19	0.25	0.15	0.20
L2	0.11	0.36	0.18	0.12	0.38	0.19

Validation subset

Model alternatives?

- Expand model with (exact) Unknow Parent Groups (Misztal *et al.* 2013)
- Use metafounders instead UPG (Legarra *et al.* 2015)
- Use alleles breed of origin and breed-specific relationship matrices (Ibanez-Escriche *et al.* 2009; Christensen *et al.* 2014)

Predictive ability APY – Model 1

	TRAIT 1			TRAIT 2		
CORE*	**L1**	**L2**	**C**	**L1**	**L2**	**C**
L1 + L2 + C	0.33	0.33	0.25	0.24	0.36	0.24
L1	0.33	0.29	0.24	0.23	0.31	0.23
L2	0.32	0.34	0.25	0.23	0.36	0.24

***CORE** = Number of eigenvalues that explain 98% variance in **G**

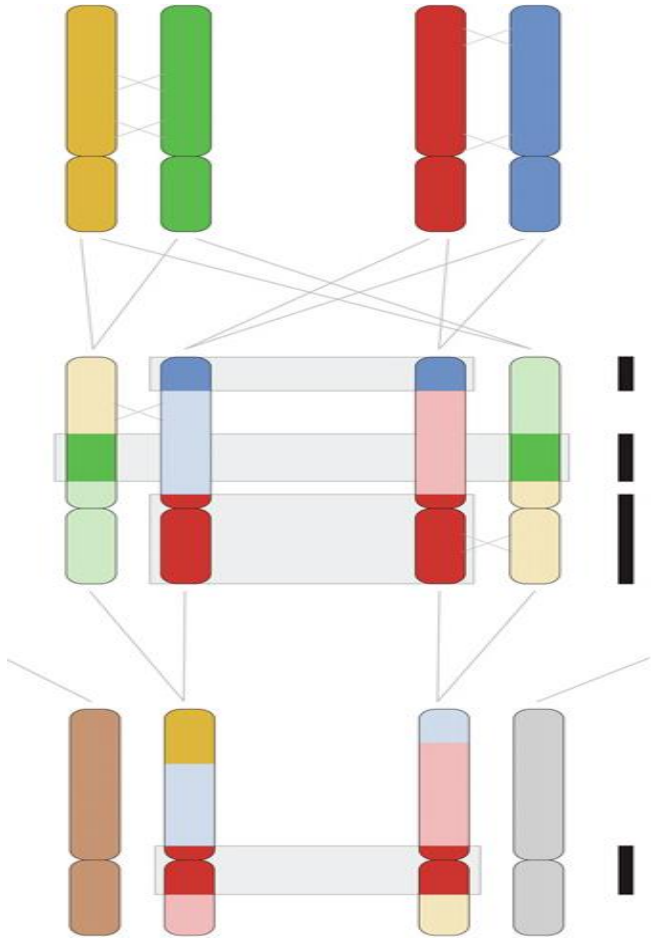
Corr (GEBV_apy, GEBV_direct) > 0.99

Estimating junctions/segments/blocks is cryptic

- Theory of junctions — Fisher (1949)
- $E(Me) = 4N_eL$ — Stam (1980)

  - Me – Independent chromosome segments
  - N_e – Effective population size
  - L – Length of genome in Morgans
  - Idealistic population structure

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



Huff *et al.* (2011)

Finding the number of segments

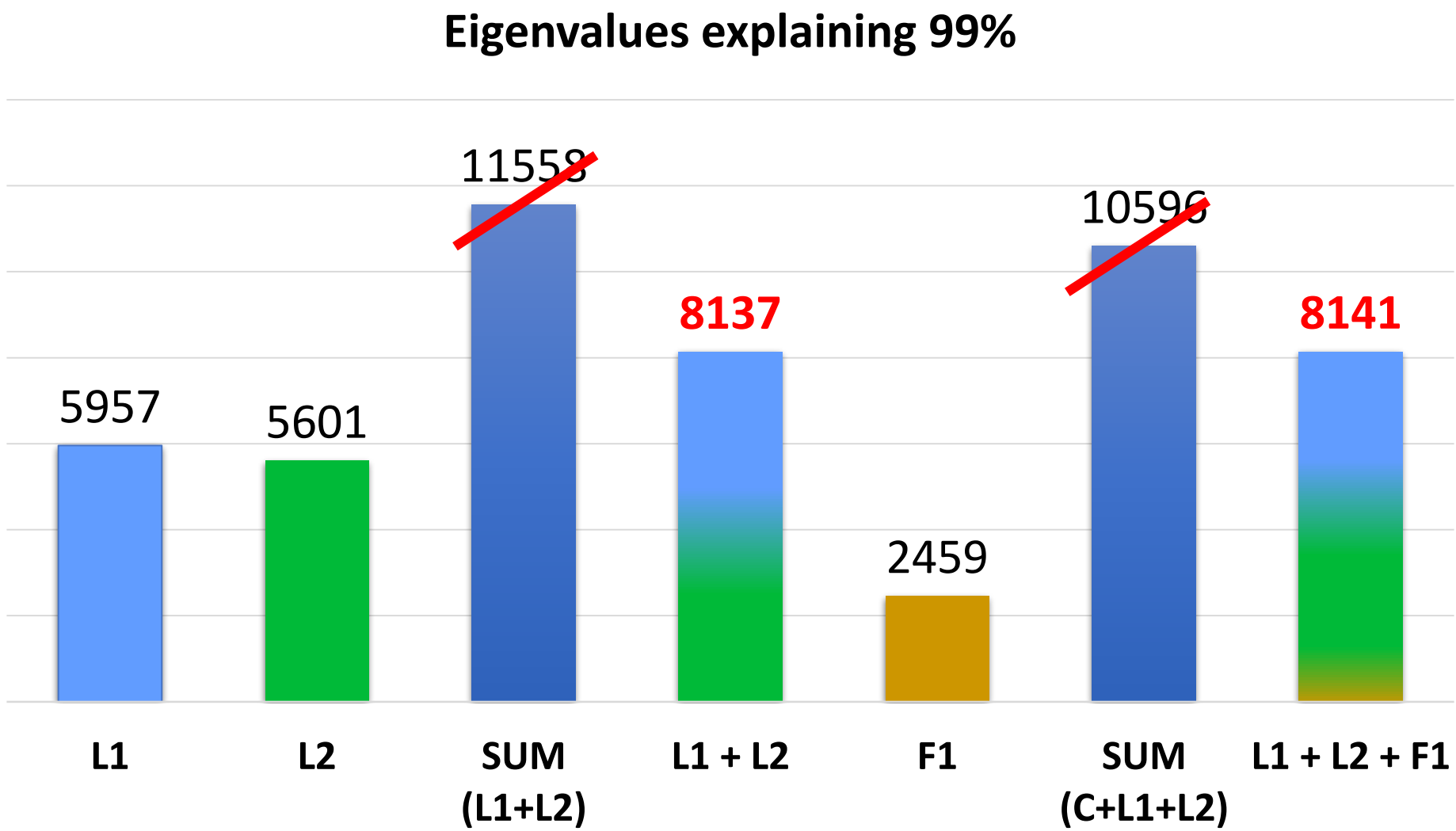
- **Z** – matrix of gene content
- Singular value decomposition: **Z = U D V'** **(U'U=I, V'V=I)**

- 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($\#_{SNP}$, $\#_{IND}$, $\#_{Me}$)

Number of shared segments



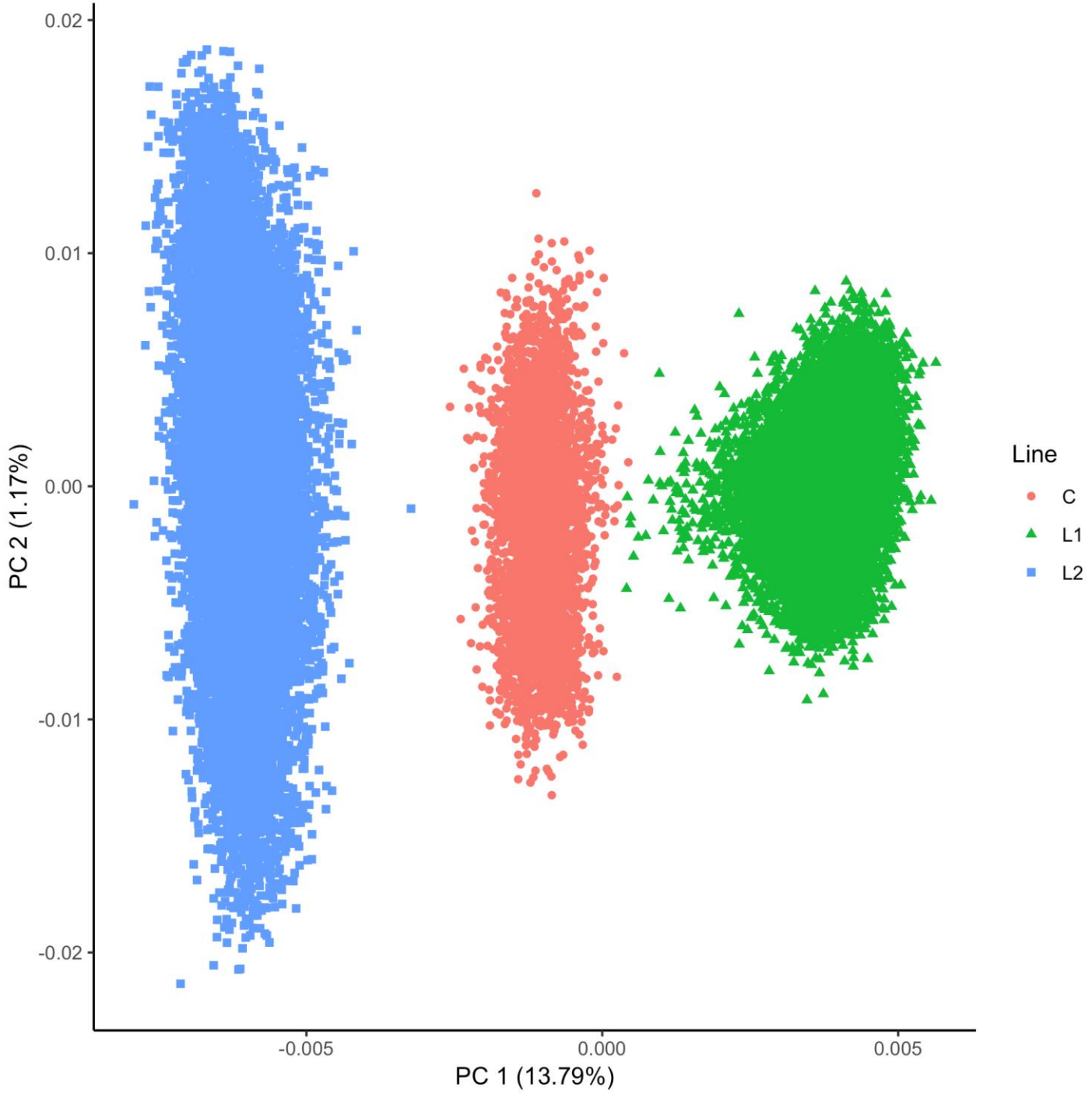
Number of shared segments – summary

- If no segments shared → lines completely independent
- Otherwise → some historical genetic connectedness between L1 and L2
- Wright's F_{ST} (L1, L2) ≈ 0.15

- How many generations ago did these lines separated before they merged again?

- Genomic info from crossbreeds already present in L1 + L2

PCA plot



Conclusions

- APY + ssGBLUP is appropriate tool for multiple lines / crossbreed datasets
- Core animals should consider all available lines

- Predictivity across the lines is possible due to the shared segments between them

- Number of shared segments can be obtained from the eigenvalue analysis of genomic information

Thank you !!!