Course roadmap

AlphaSimR

Day 1: Simulation of breeding programmes **BASICS**

Day 2: … Quantitative genetics

Day 3: … Estimation with linear mixed models

Day 4: … Spatial variation & GxE interactions

Day 5: … Ancestral recombination graphs

Day 5 agenda – Ancestral Recombination Graphs

- 09:00-10:30 Spatial modelling, ARG
- 10:30-11:00 Refreshments break
- 11:00-12:30 ARG
- 12:30-13:30 Lunch break
- 13:30-15:00 ARG / ARG Practicals
- 15:00-15:30 Refreshments break
- 15:30-17:00 ARG Practicals / Open-end

Spatial modelling improves genetic evaluation in smallholder breeding programs

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-09

Learning objectives

Separating genetic and environmental effects is a critical component of any quantitative genetics model

- Showcase challenge and solution for modelling data from smallholder settings
- Aside: APY approximation

Modelling environmental effects

- Naïve model

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}^*$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$$
$$\boldsymbol{e}^* \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_{e^*}^2)$$

- Model contemporary group effect, say, herd, herd-season, …

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{Wh} + \boldsymbol{e}$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$$
$$\boldsymbol{h} \sim N(\boldsymbol{0}, \boldsymbol{H}\sigma_h^2)$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

Extensive literature on “fixed” vs. “random” treatment due to data structure or views/opinions:

- unbalanced designs & bias
- ability to estimate effects
- …

Challenge with small contemporary groups

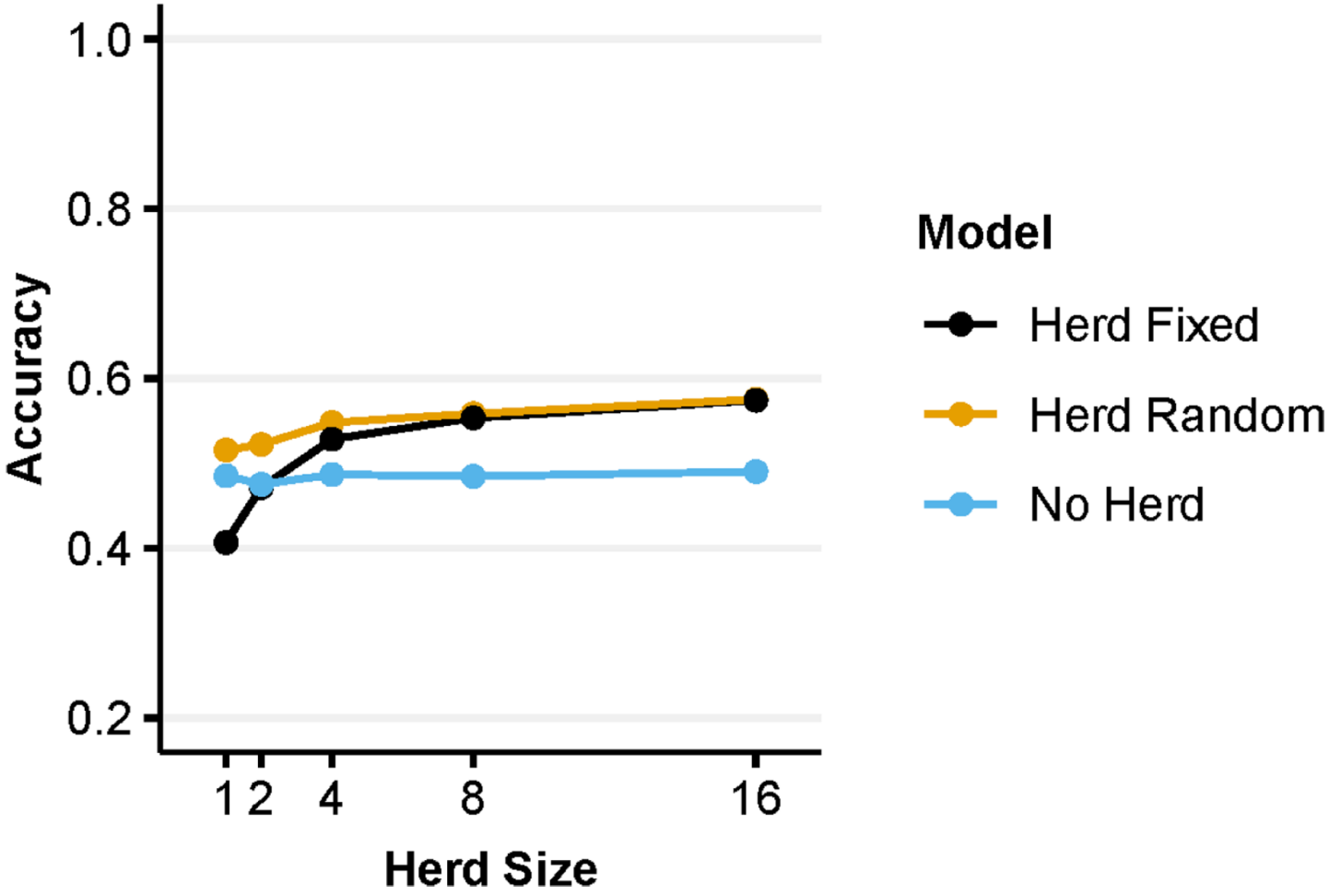
JDS Communications®
2021; 2:366–370

https://doi.org/10.3168/jdsc.2021-0092
Short Communication
Genetics

Genomic evaluations using data recorded on smallholder dairy farms in low- to middle-income countries

Owen Powell,[1]* Raphael Mrode,[2,3] R. Chris Gaynor,[1] Martin Johnsson,[1,4] Gregor Gorjanc,[1] and John M. Hickey[1]

Challenge with small contemporary groups



Environmental/Spatial modelling

- A solution?
  - borrow information from neighbours (spatial model) and/or
  - measure key environmental indicators (location covariates)

Selle *et al. Genet Sel Evol* (2020) 52:69
https://doi.org/10.1186/s12711-020-00588-w

GSE Genetics Selection Evolution

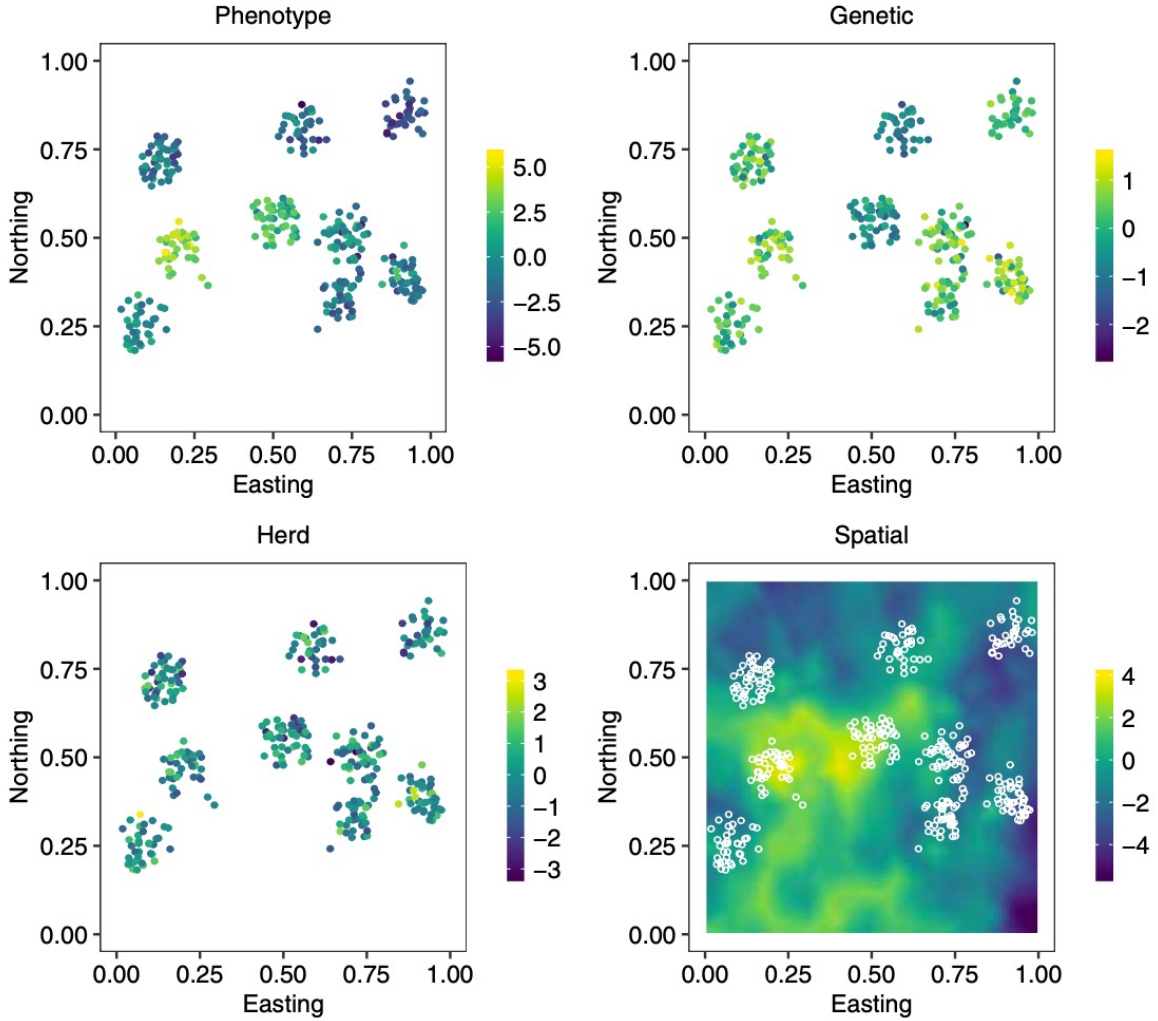
RESEARCH ARTICLE Open Access

Check for updates

Spatial modelling improves genetic evaluation in smallholder breeding programs

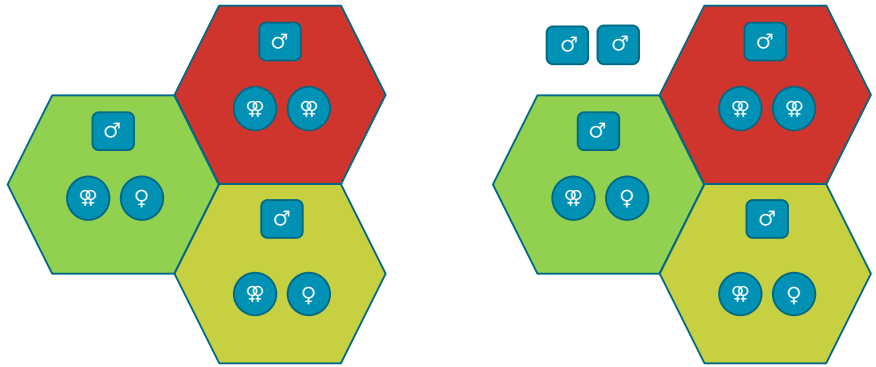
Maria L. Selle[1*], Ingelin Steinsland[1], Owen Powell[2], John M. Hickey[2] and Gregor Gorjanc[2]

Spatial context



Simulation

- Smallholder breeding programme
- Connectedness scenarios



- Phenotype = Location + Herd + Genetics + Noise
  0.40 0.25 0.10 0.25

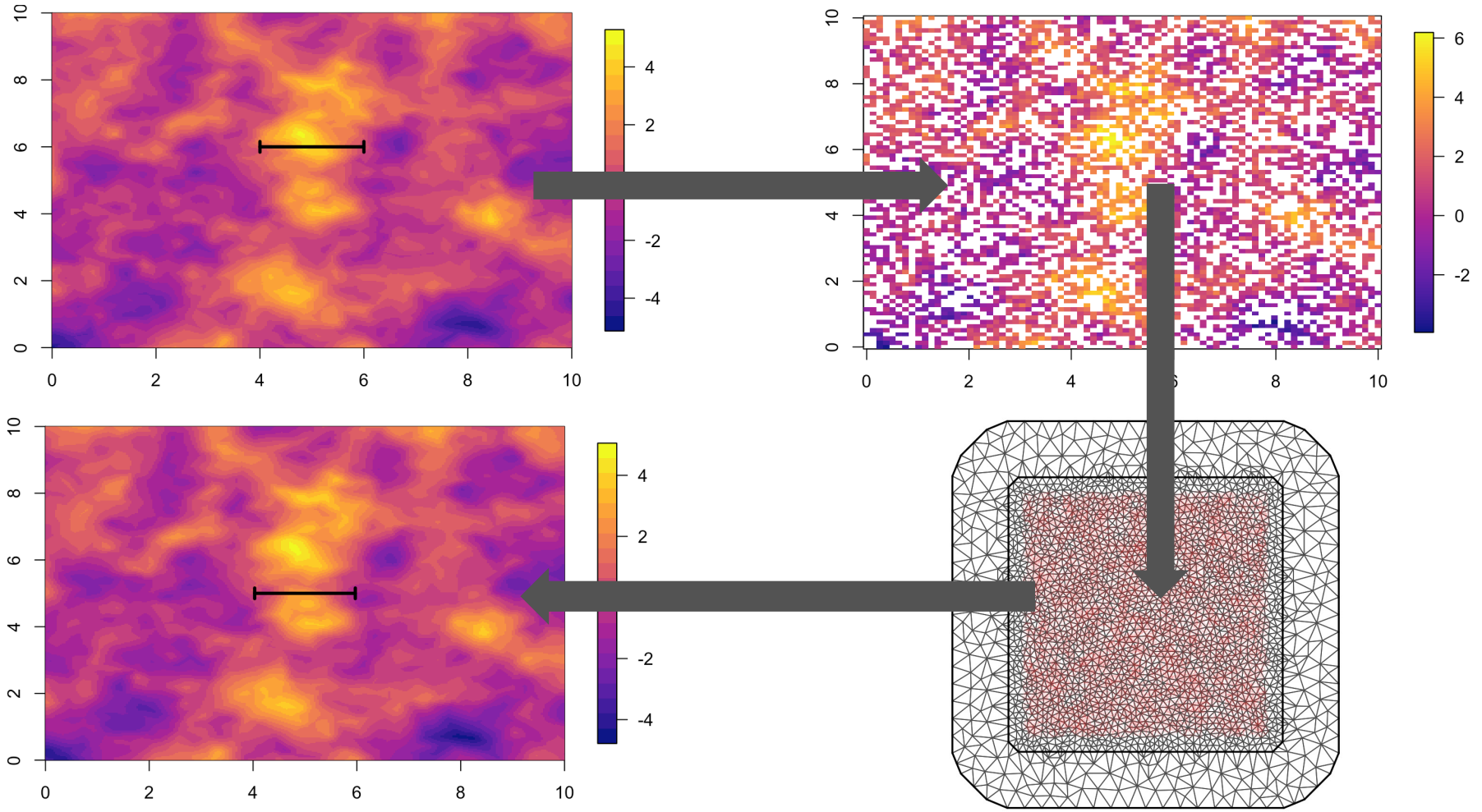
- Location = sum of 10 spatially varying covariates (rainfall, temp, …)
- Observe 5 covariates with noise and 2 as binary indicators

Spatial modelling

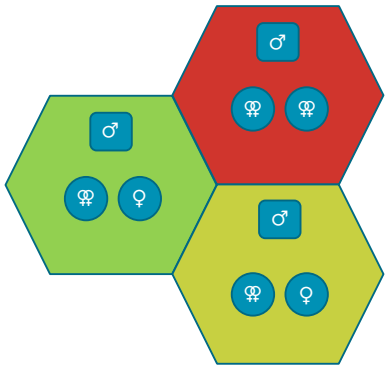
$$\text{Phenotype} = \text{Location} + \text{Herd} + \text{Genetics} + \text{Noise}$$

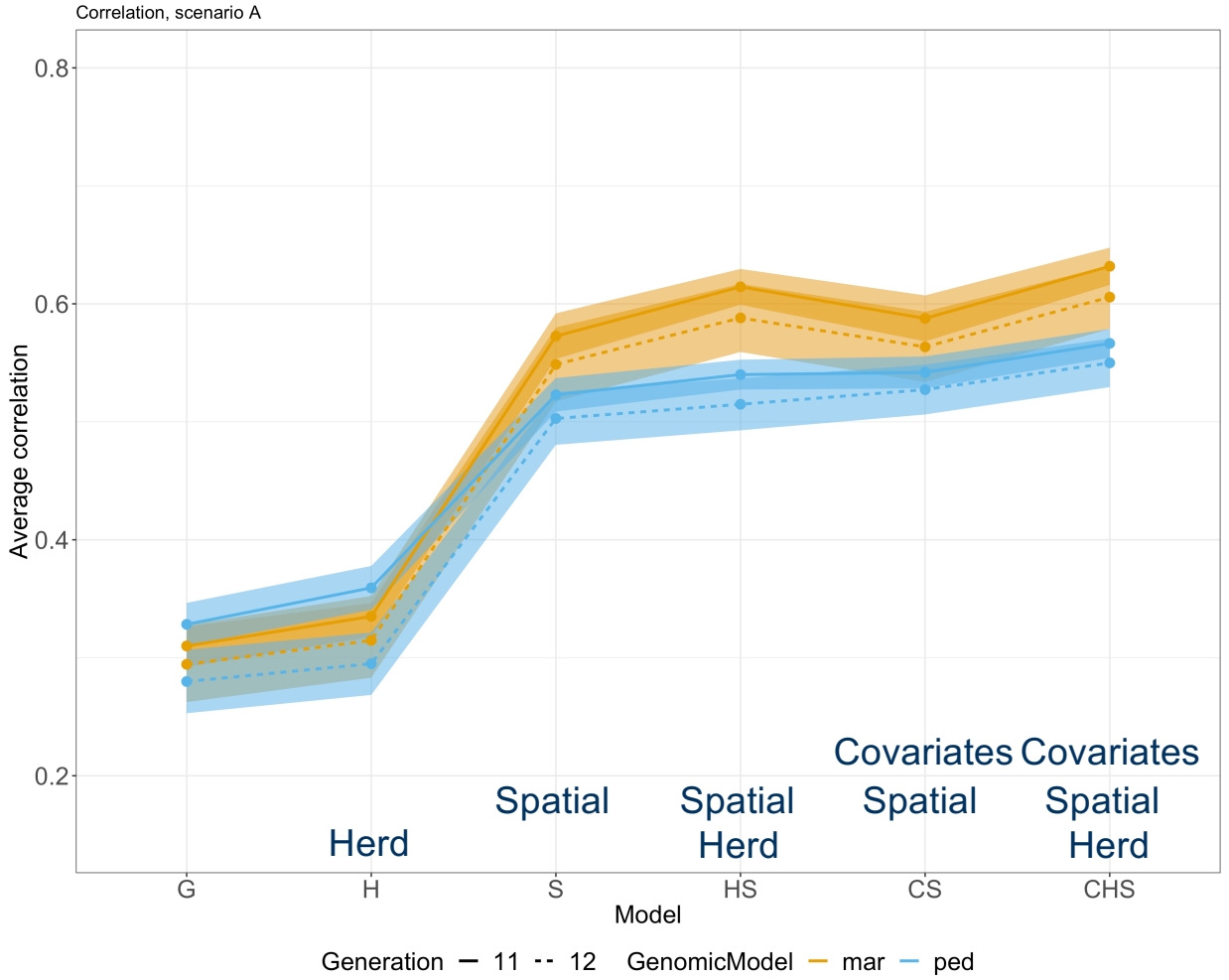
- Location
  - Spatial $\mathbf{l} \sim N(\mathbf{0}, \text{Matern}(\mathbf{E}, \text{Kappa}, \text{Var}_L))$
  - Spatial + Covariates
- Herd $h \sim N(\mathbf{0}, \mathbf{I}\text{Var}_H)$
- Genetics
  - Pedigree-based $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}_{ped}\text{Var}_A)$
  - Genome-based $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}_{mar}\text{Var}_A)$

SPDE approach (similar idea to APY)

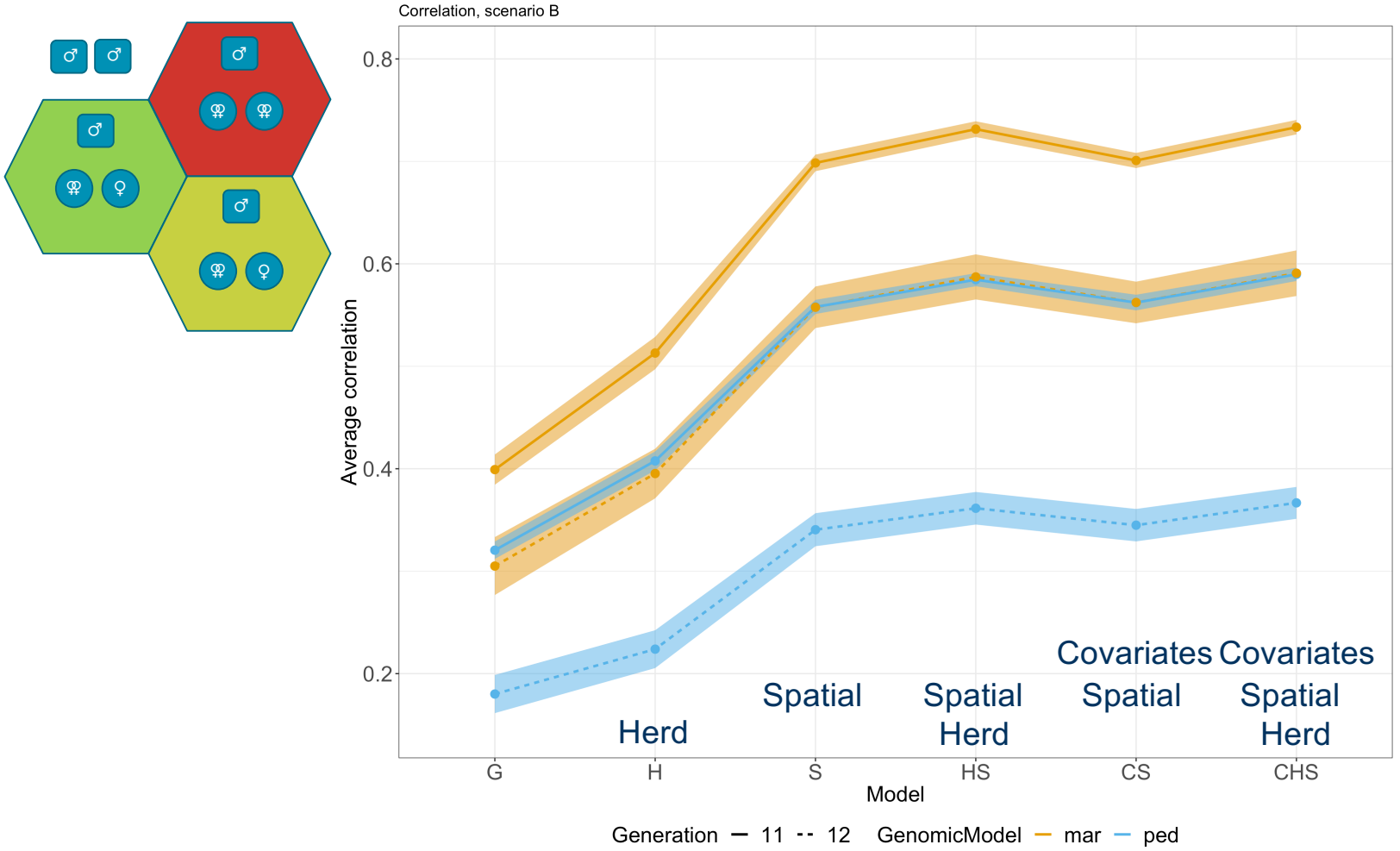


Accuracy of evaluation & prediction (simulation)





Accuracy of evaluation & prediction (simulation)



Real application

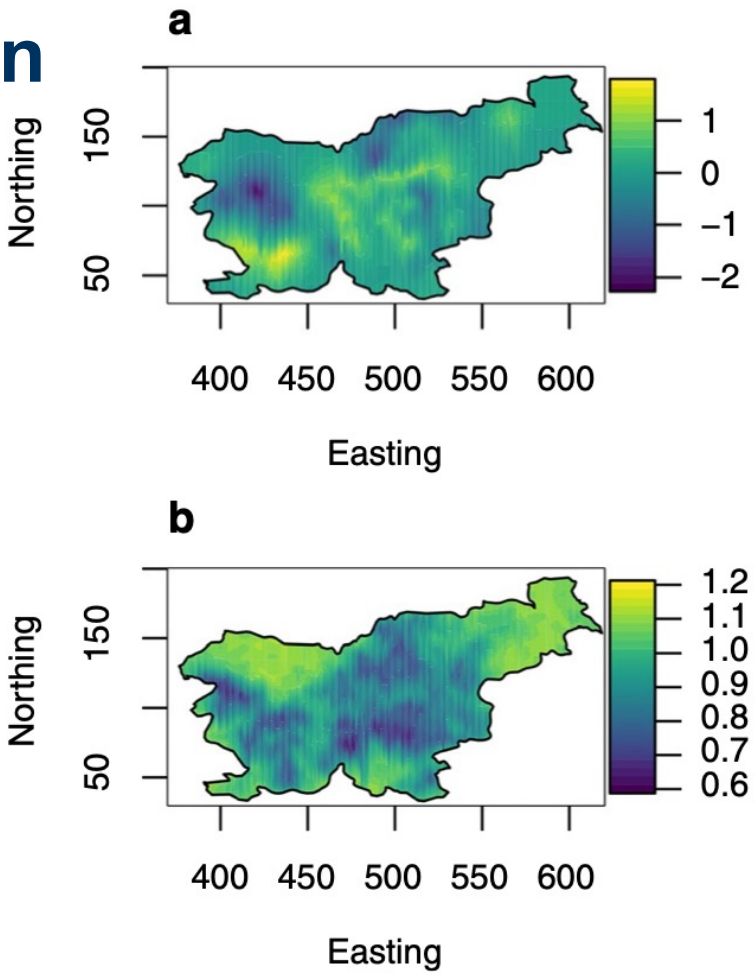
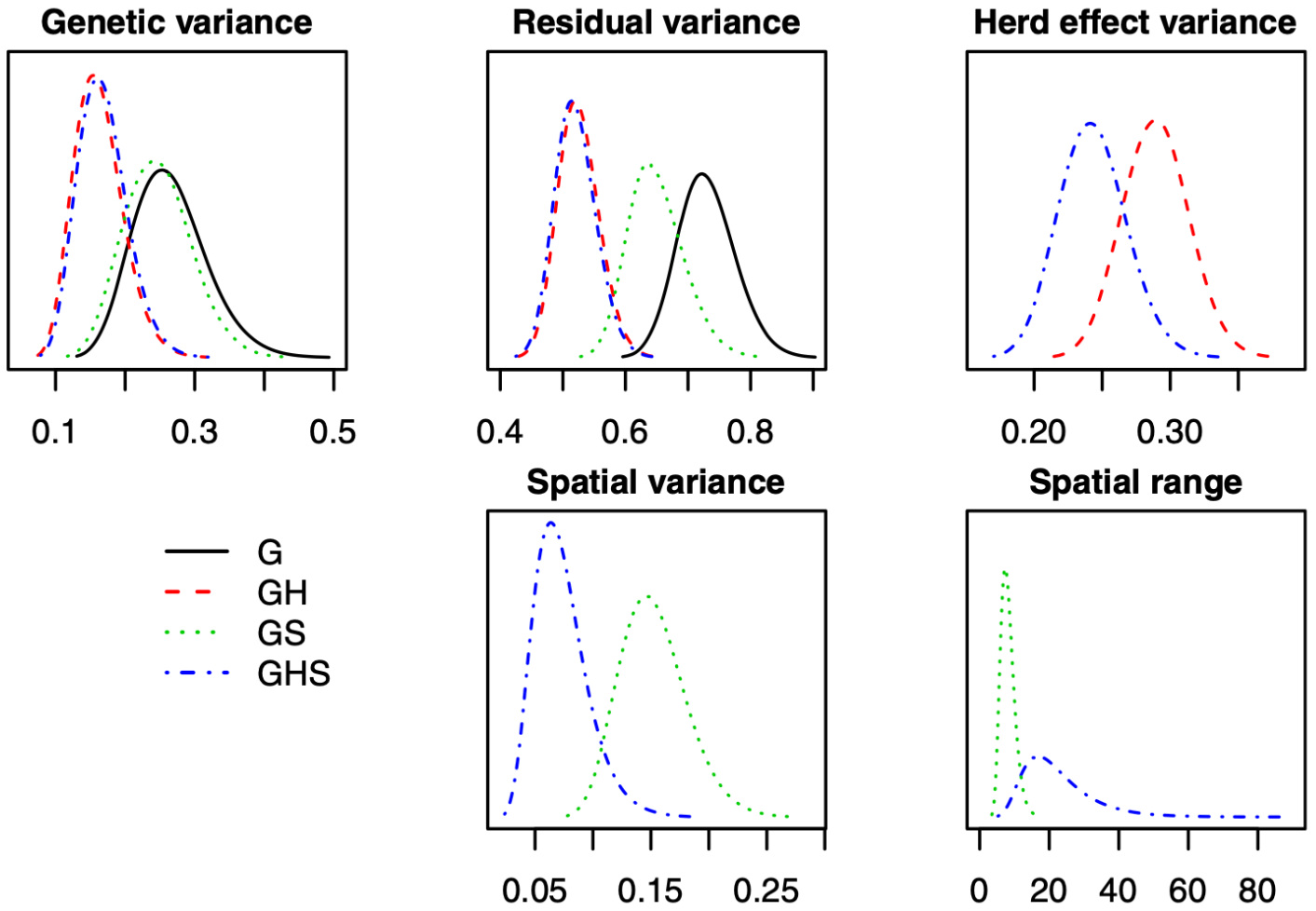


Fig. 5 Posterior mean (**a**) and standard deviation (**b**) of the estimated spatial effect (in units of posterior spatial standard deviation) from model GHS fitted to the real data—the axis units are in km

Estimated variance components



Does it matter? Yes!

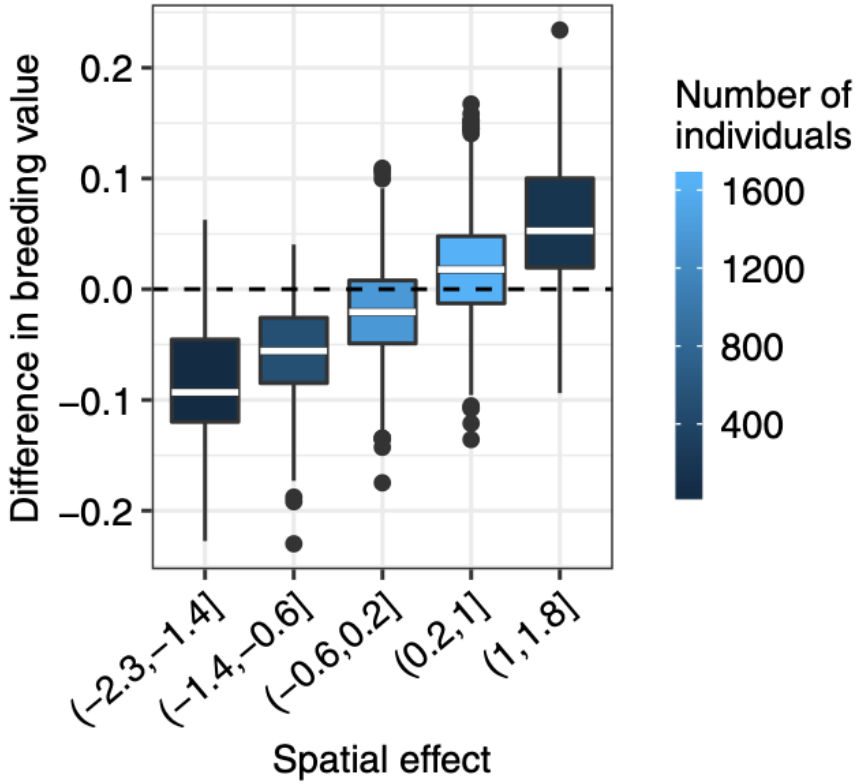


Fig. 6 The difference in estimated breeding values (in units of posterior genetic standard deviation) between models GH and GHS by the estimated spatial effect (in units of posterior spatial standard deviation) from model GHS fitted to the real data

Questions?!

APY approximation

- Number of genotyped individuals is growing rapidly!!!!

- Fitting genome-based models is becoming a challenge
  - Marker model scales with number of markers
  - Individual model scales with number of individuals

- Many solutions proposed to manage the scale
- APY (Algorithm for Proven and Young) is one of them
- Showcasing it due to connection to Day 3 and spatial models

Idea

- Pedigree-based model

$$Var(\boldsymbol{a}|\boldsymbol{T}) = \boldsymbol{TDT^T}\sigma_a^2 = \boldsymbol{A}\sigma_a^2$$

  – if we use MME, we need inverse of **A**, sparse with pedigree data

  – sparse because of pedigree structure (conditioning on parents, Mendelian sampling terms are independent of parents)

- Genome-based model

$$Var(\boldsymbol{a}|\boldsymbol{W}) = \boldsymbol{WW}^T\sigma_\alpha^2 = \boldsymbol{G}\sigma_\alpha^2$$

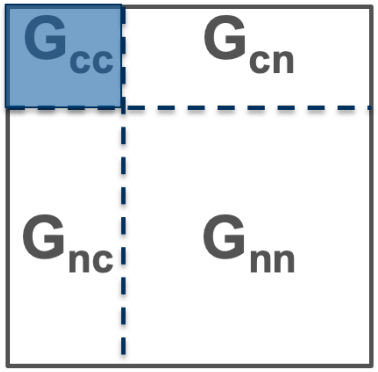
  – if we use MME, we need inverse of **G**, dense with genomic data

  – dense because genomic data informs what we share or don't share (no independence)

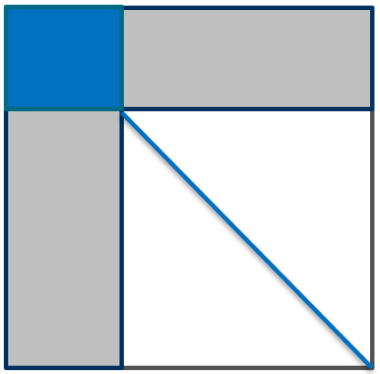
Idea

- With limited number of markers and large number of ind., we soon incur linear dependencies → we can explain genome of some ind. with genome of other ind. & what is not explained is independent
- APY
  - Pick some "core" individuals (=locations/knots in spatial community) & "non-core" individuals
  - Split **G** between these two
  - Inverse **G** can be made sparse

$\boldsymbol{G}$



$\boldsymbol{G^{-1}}$



(Misztal, 2016)

How?

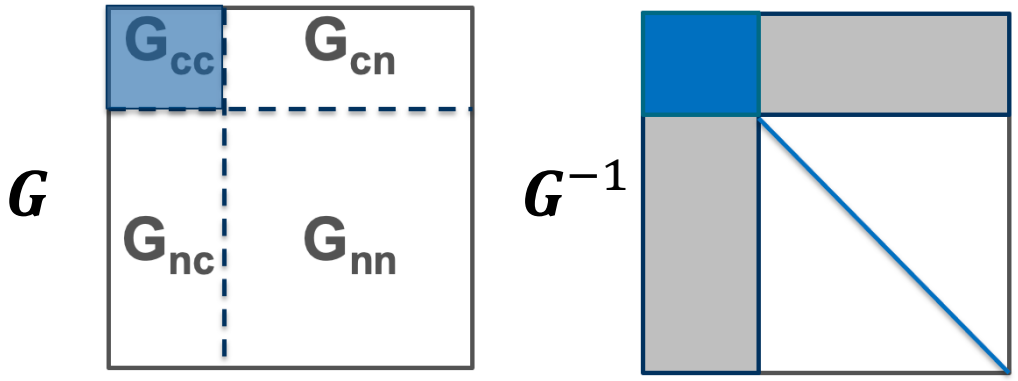
$$Var(\boldsymbol{a}|\boldsymbol{W}) = \boldsymbol{W}\boldsymbol{W}^T\sigma_\alpha^2 = \boldsymbol{G}\sigma_\alpha^2$$
$$\boldsymbol{a}_c = \boldsymbol{W}_c\boldsymbol{\alpha} + \boldsymbol{r}_c$$
$$\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$$
$$\boldsymbol{r}_c \sim N(\boldsymbol{0}, \boldsymbol{I}m\sigma_\alpha^2)$$
$$\begin{aligned}\boldsymbol{a}_n &= \boldsymbol{W}_n\boldsymbol{\alpha} + \boldsymbol{r}_n \\ &= f(\boldsymbol{a}_c) + \boldsymbol{r}_n\end{aligned}$$

$$E(\boldsymbol{a}_n|\boldsymbol{a}_c) = E(\boldsymbol{a}_n) + Cov(\boldsymbol{a}_n, \boldsymbol{a}_c)Var(\boldsymbol{a}_c)^{-1}\left(\boldsymbol{a}_c - E(\boldsymbol{a}_c)\right)$$
$$Var(\boldsymbol{a}_n|\boldsymbol{a}_c) = Var(\boldsymbol{a}_n) - Cov(\boldsymbol{a}_n, \boldsymbol{a}_c)Var(\boldsymbol{a}_c)^{-1}Cov(\boldsymbol{a}_c, \boldsymbol{a}_n)$$

How?

$$E(\boldsymbol{a}_n|\boldsymbol{a}_c) = Cov(\boldsymbol{a}_n, \boldsymbol{a}_c)Var(\boldsymbol{a}_c)^{-1}\boldsymbol{a}_c$$
$$= \boldsymbol{G}_{n,c}\boldsymbol{G}_{c,c}^{-1}\boldsymbol{a}_c$$

$$Var(\boldsymbol{a}_n|\boldsymbol{a}_c) = Var(\boldsymbol{a}_n) - Cov(\boldsymbol{a}_n, \boldsymbol{a}_c)Var(\boldsymbol{a}_c)^{-1}Cov(\boldsymbol{a}_c, \boldsymbol{a}_n)$$
$$= \boldsymbol{G}_{n,n} - \boldsymbol{G}_{n,c}\boldsymbol{G}_{c,c}^{-1}\boldsymbol{G}_{c,n}$$



(Misztal, 2016)

Core optimisation

Pocrnic *et al. Genetics Selection Evolution* (2022) 54:76
https://doi.org/10.1186/s12711-022-00767-x

RESEARCH ARTICLE Open Access

Optimisation of the core subset for the APY approximation of genomic relationships

Ivan Pocrnic[1*], Finn Lindgren[2], Daniel Tolhurst[1], William O. Herring[3] and Gregor Gorjanc[1]

Core optimisation - algorithms

Algorithm 1 Core subset optimisation using conditional covariance matrix $\mathbf{C}$

Require: n_c, $\mathbf{k}$, and $\mathbf{C}_0 = \mathbf{W}\mathbf{W}^\top$ ▷ Core subset size, Vector for core animals, and Covariance matrix

1: **for** i in 1 to n_c **do** ▷ Loop over the core subset

2: $k_i \leftarrow argmax\big(diag(\mathbf{C}_{i-1})\big)$ ▷ Find the i-th core animal

3: $\mathbf{e} \leftarrow 0$ ▷ Update the "selector" vector

4: $\mathbf{e}_{k_i} \leftarrow 1$

5: $\mathbf{C}_i \leftarrow \mathbf{C}_{i-1} - \mathbf{C}_{i-1}\mathbf{e}_{k_i}\left(\mathbf{e}_{k_i}^\top \mathbf{C}_{i-1}\mathbf{e}_{k_i}\right)^{-1}\mathbf{e}_{k_i}^\top \mathbf{C}_{i-1}$ ▷ Update the covariance matrix

6: **end for**

Core optimisation - algorithms

Algorithm 1 Core subset optimisation using conditional covariance matrix $\mathbf{C}$

Require: n_c, $\mathbf{k}$, and $\mathbf{C}_0 = \mathbf{W}\mathbf{W}^\top$ ▷ Core subset size, Vector for core animals, and Covariance matrix

1: **for** i in 1 to n_c **do** ▷ Loop over the core subset

2: $k_i \leftarrow argmax\big(diag(\mathbf{C}_{i-1})\big)$ ▷ Find the i-th core animal

3: $\mathbf{e} \leftarrow 0$ ▷ Update the "selector" vector

4: $\mathbf{e}_{k_i} \leftarrow 1$

5: $\mathbf{C}_i \leftarrow \mathbf{C}_{i-1} - \mathbf{C}_{i-1}\mathbf{e}_{k_i}\left(\mathbf{e}_{k_i}^\top \mathbf{C}_{i-1}\mathbf{e}_{k_i}\right)^{-1}\mathbf{e}_{k_i}^\top \mathbf{C}_{i-1}$ ▷ Update the covariance matrix

6: **end for**

Algorithm 2 Core subset optimisation using conditional SNP genotype matrix $\mathbf{W}$

Require: n_c, $\mathbf{k}$, and $\mathbf{W}_0 = \mathbf{W}$ ▷ Core subset size, Vector for core animals, and SNP genotype matrix

1: **for** i in 1 to n_c **do** ▷ Loop over the core subset

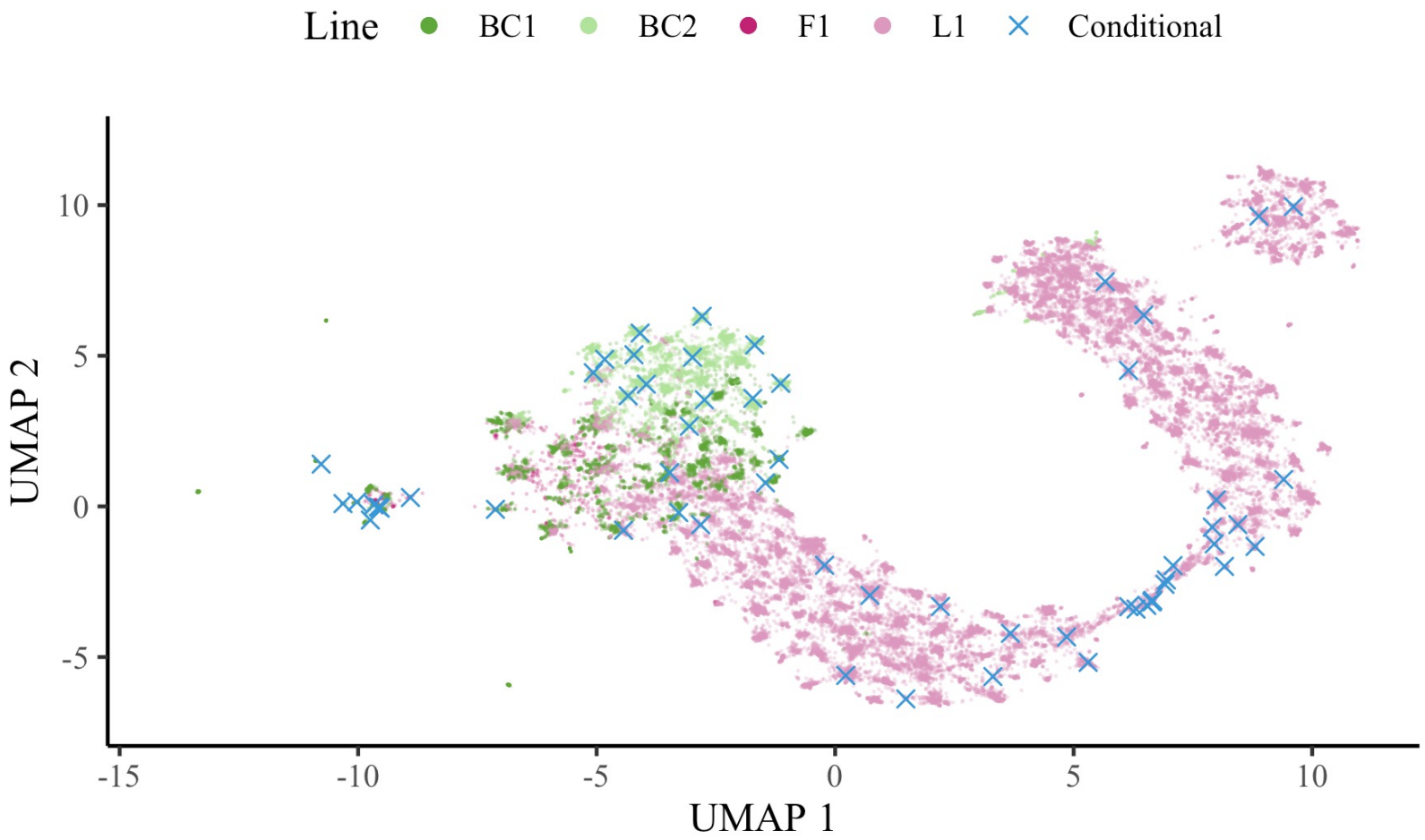
2: $k_i \leftarrow argmax\big(diag(\mathbf{W}_{i-1}\mathbf{W}_{i-1}^\top)\big)$ ▷ Find the i-th core animal

3: $\mathbf{w}_{k_i} \leftarrow \mathbf{W}_{i-1}[k_i,]$ ▷ Conditional SNP genotypes of the animal i

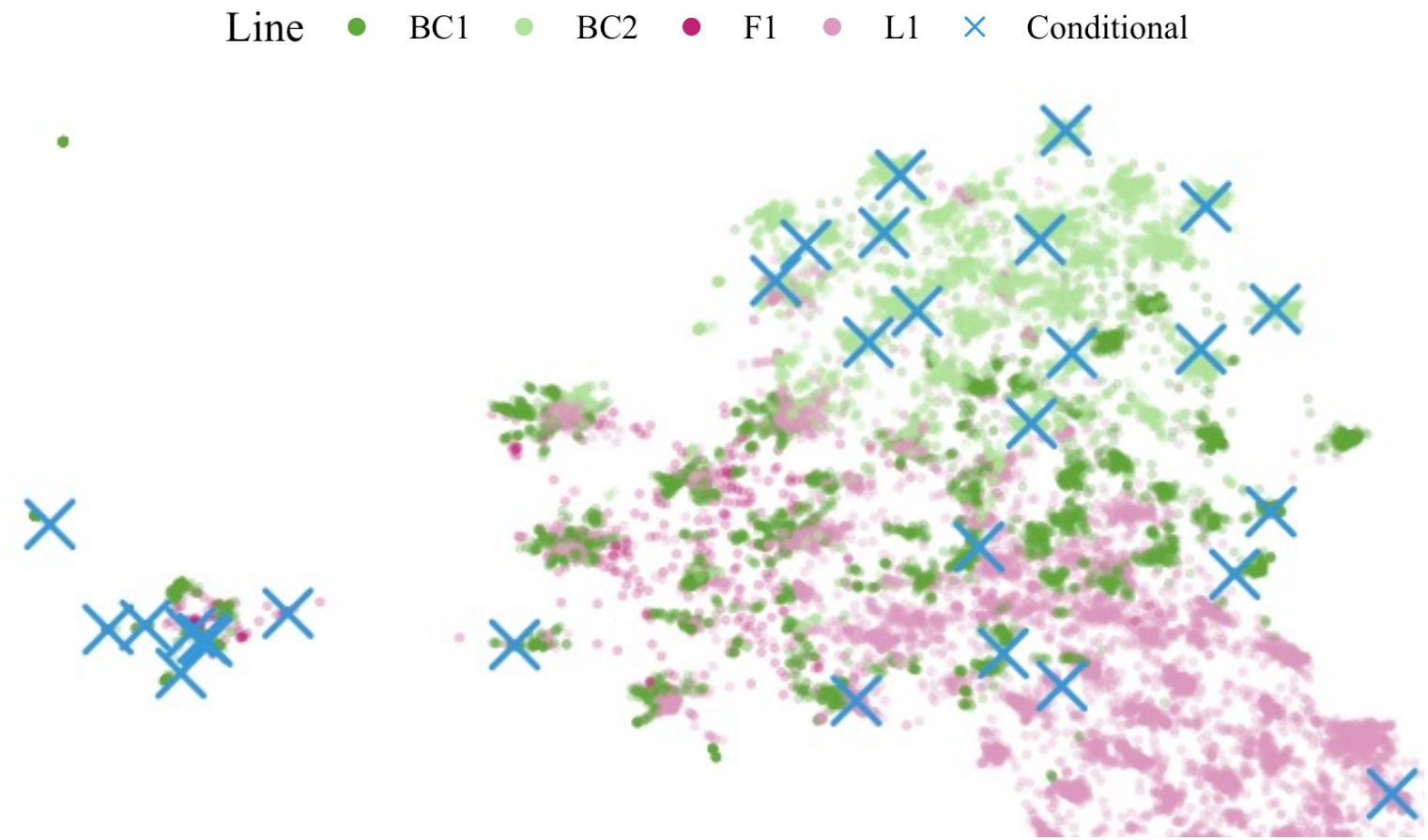
4: $\mathbf{W}_i \leftarrow \mathbf{W}_{i-1} - \mathbf{W}_{i-1}\mathbf{w}_{k_i}^\top \mathbf{w}_{k_i} / \left(\mathbf{w}_{k_i}\mathbf{w}_{k_i}^\top\right)$ ▷ Update the SNP genotype matrix

5: **end for**

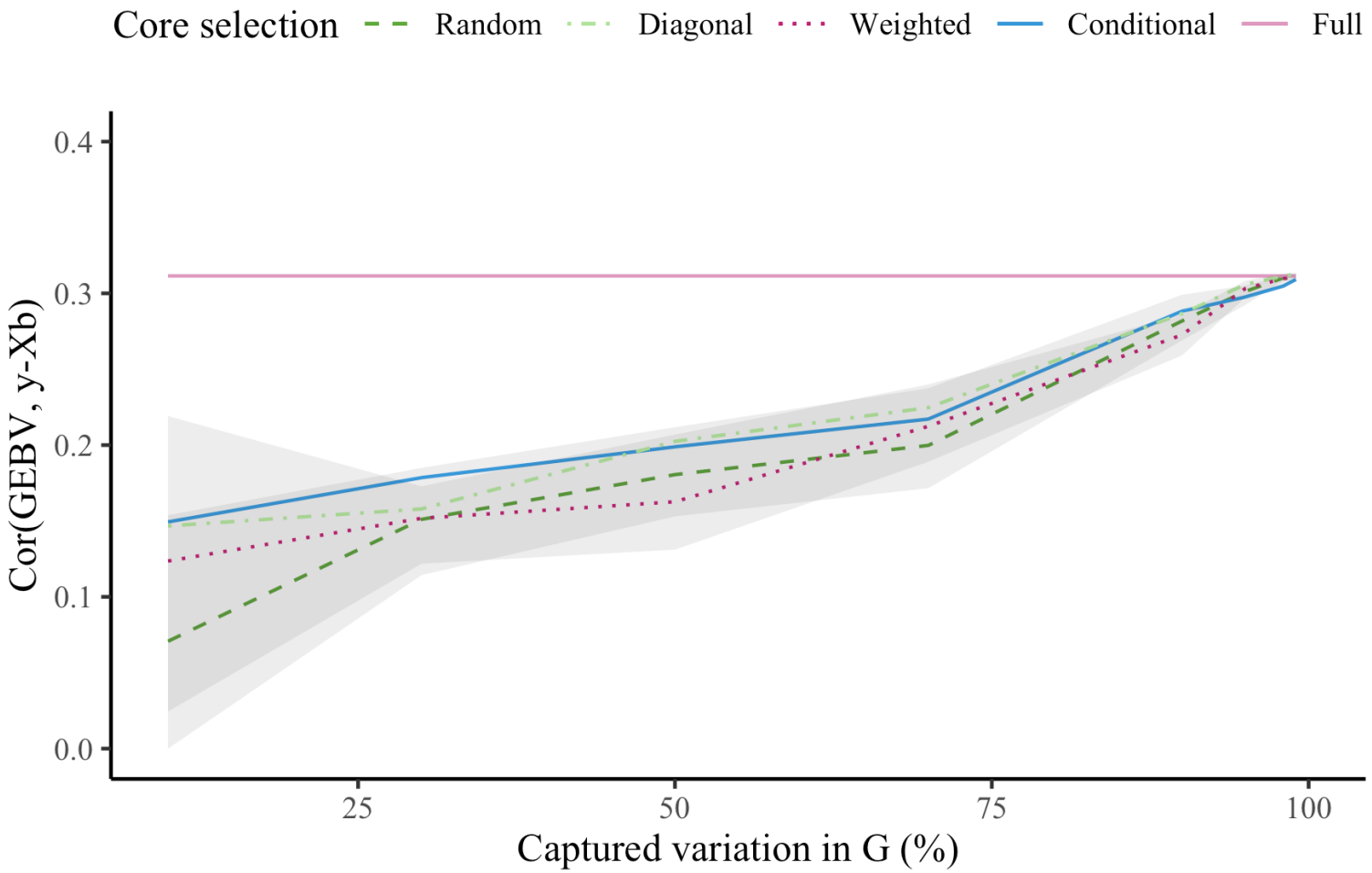
Core optimisation – UMAP visualisation



Core optimisation – UMAP visualisation



Impact on accuracy of prediction



Learning objectives

Separating genetic and environmental effects is a critical component of any quantitative genetics model

- Showcase challenge and solution for modelling data from smallholder settings
- Aside: APY approximation

Questions?!

Spatial modelling improves genetic evaluation in smallholder breeding programs

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-09