Gene tree- & tree sequence-based linear mixed models

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-09

Warning disclaimer

Active area of learning, exploration, & research in our lab!!!

- Ideas & work in progress (pre-publication stage!)
- Building experience with applications at this stage

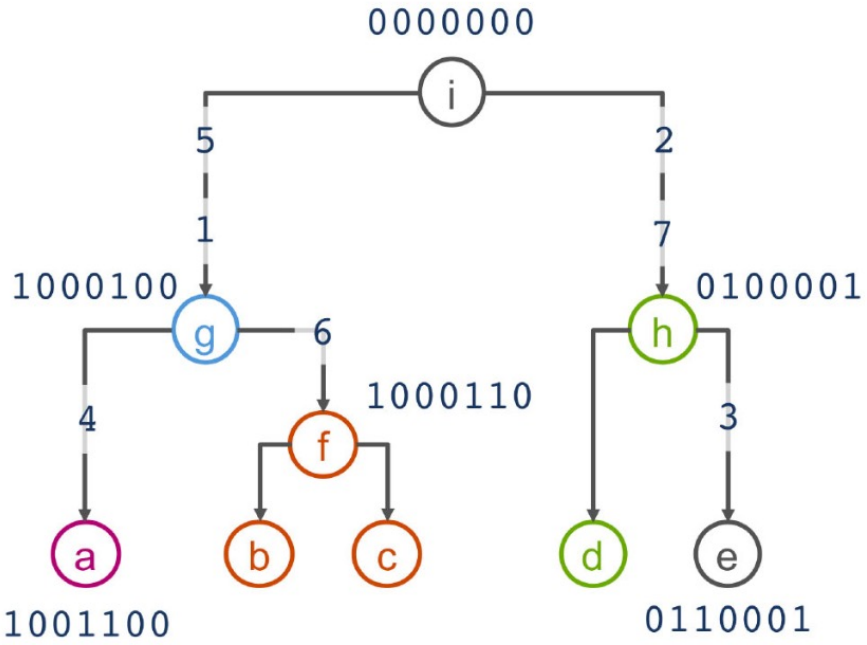
Learning objectives

- Showcase two approaches to modelling haplotype effects for a non-recombining region (real & simulation)

- Showcase tree-sequence results for a rice dataset

Hierarchical Modelling of Haplotype Effects on a Phylogeny

Maria Lie Selle[1], Ingelin Steinsland[1], Finn Lindgren[2], Vladimir Brajkovic[3], Vlatka Cubric-Curik[3] and Gregor Gorjanc[4]*



$$
\begin{aligned}
h_i &\sim N\left(0, \sigma^2_{h_m}\right) \\
h_{g'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_g|h_{g'} &\sim N(\rho h_{g'}, \sigma^2_{h_c}) \\
h_a|h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_f, h_b, h_c, |h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_{h'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_h, h_d|h_{h'} &\sim N\left(\rho h_{h'}, \sigma^2_{h_c}\right) \\
h_e|h_h &\sim N\left(\rho h_h, \sigma^2_{h_c}\right)
\end{aligned}
$$

Model & Results

$$h_1 \sim \mathcal{N}(0, \sigma^2_{h_m}),$$

$$h_j | h_{p(j)} \sim \mathcal{N}(\rho h_{p(j)}, \sigma^2_{h_c}).$$

$$\boldsymbol{h} = \mathbf{T}(\rho)\,\varepsilon,$$

$$\mathbf{T}(\rho)^{-1}\boldsymbol{h} = \boldsymbol{\varepsilon},$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{h}) &= \mathrm{Var}(\mathbf{T}(\rho)\,\boldsymbol{\varepsilon}),\\ &= \mathbf{T}(\rho)\,\mathrm{Var}(\boldsymbol{\varepsilon})\mathbf{T}(\rho)^T = \mathbf{T}(\rho)\,\mathbf{D}(\rho)\,\mathbf{T}(\rho)^T\sigma^2_{h_c}\\ &= \mathbf{H}(\rho)\,\sigma^2_{h_c} = \mathbf{V}_h\left(\rho, \sigma^2_{h_c}\right),\end{aligned}$$

$$\boldsymbol{h} | \rho, \sigma^2_{h_c} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_h\left(\rho, \sigma^2_{h_c}\right)\right),$$

$$\mathbf{H}(\rho)^{-1} = \frac{1}{\sigma^2_{h_c}}\mathbf{T}(\rho)^{-1T}\mathbf{D}(\rho)^{-1}\mathbf{T}(\rho)^{-1}.$$

Application to cattle mtDNA (collaborators provided the tree)

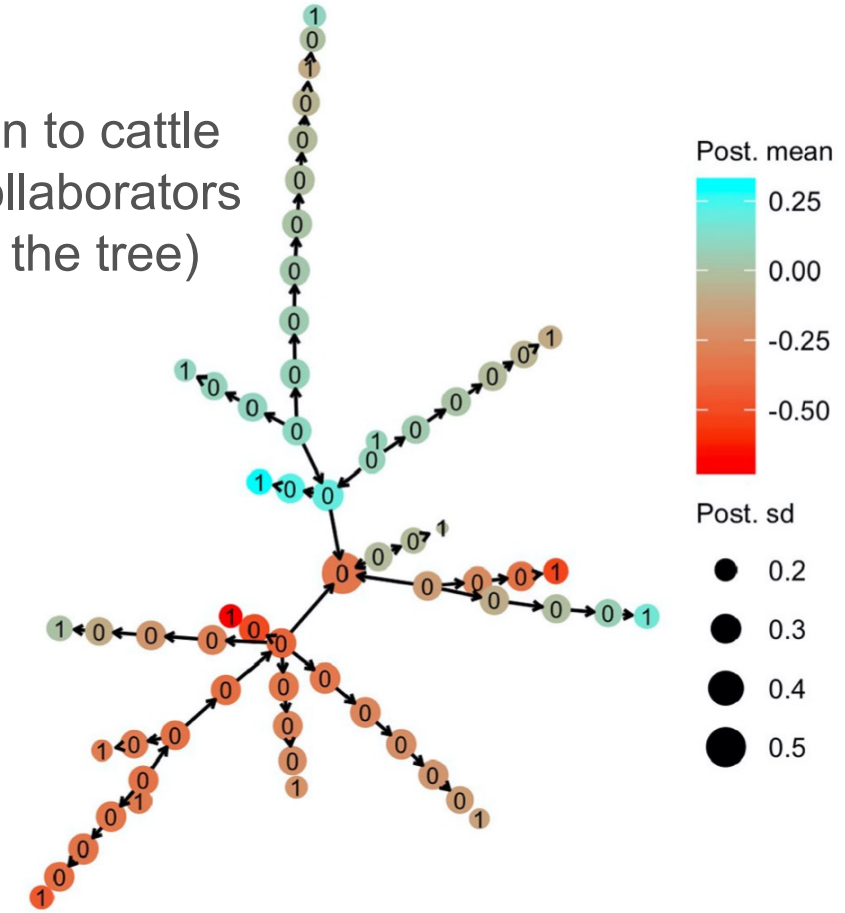
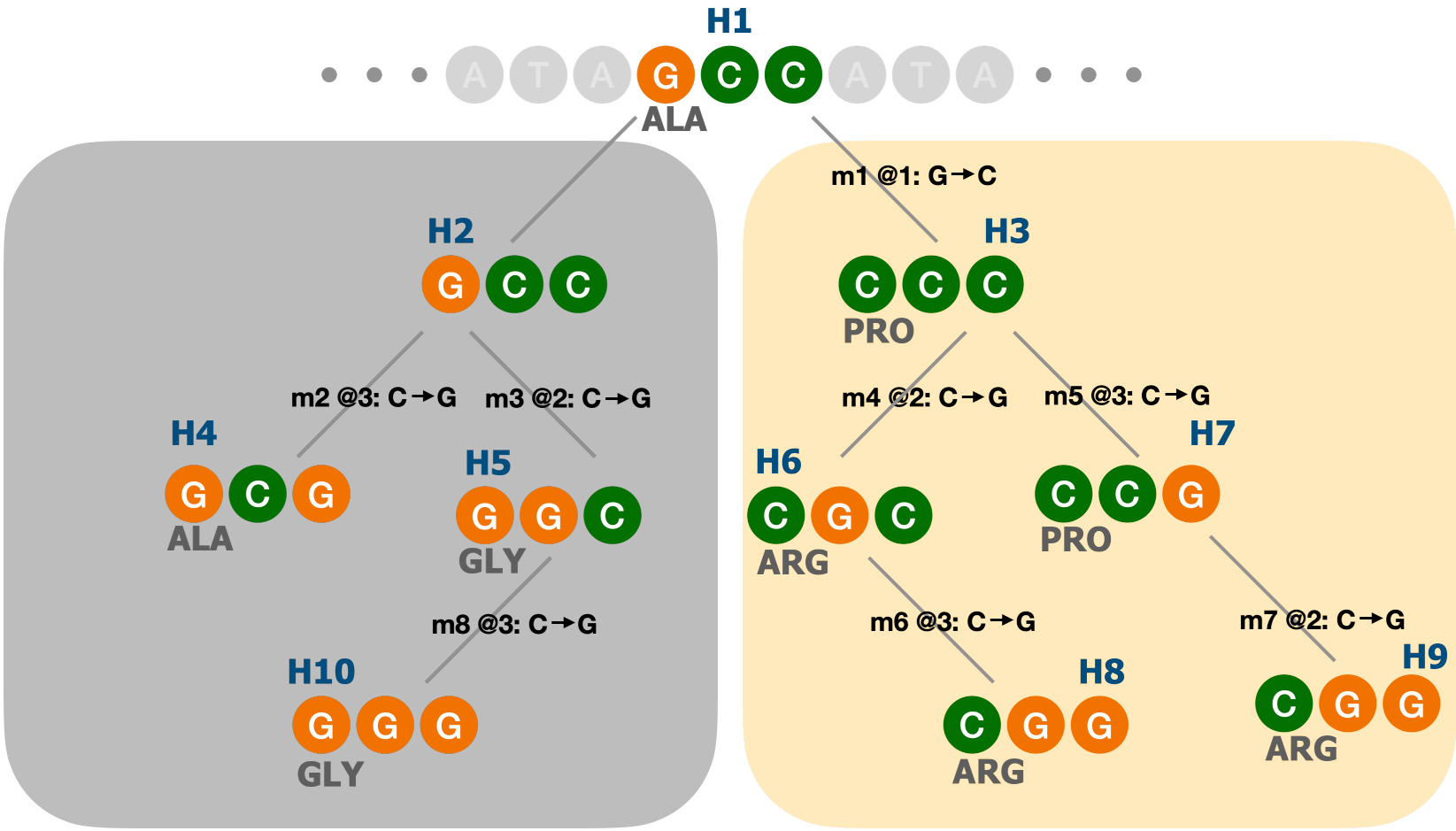


FIGURE 5 | Posterior mean and standard deviation for mitochondrial haplotype effects on milk yield in cattle. Posterior means are denoted with node color, while posterior deviations are denoted by the node size. The numbers on each haplotype node indicate if the haplotype had a direct link to the observed phenotype (1) or not (0).

Gabriela Mafra Fortuna

Revisiting gene tree modelling approach

Imagine this (possibly unlikely) gene tree



Haplotypes description

Hid	Pid	A1	A2	A3	Mutated	Mutation site	Hap. Value	Mut. effect
1	NA	0	0	0	0	NA	0	NA
2	1	0	0	0	0	NA	0	NA
3	1	1	0	0	1	1	1	1
4	2	0	0	1	1	3	0	0
5	2	0	1	0	1	2	3	3
6	3	1	1	0	1	2	2	1
7	3	1	0	1	1	3	1	0
8	6	1	1	1	1	3	2	0
9	7	1	1	1	1	2	2	1
10	5	0	1	1	1	3	3	0

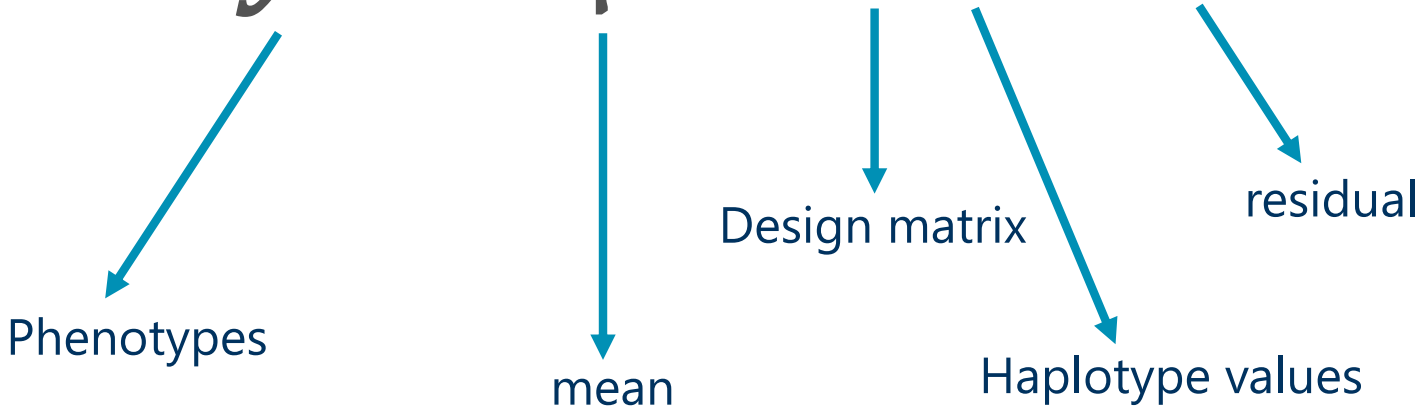
Simulated data

- Using a small balanced example where we pull nine haplotypes at random

IndID	HapID	Hap. Value	Phenotype
1	H3	1	y_1
2	H2	0	y_2
3	H2	0	y_3
4	H5	3	y_5
5	H10	3	y_6
6	H9	2	y_7
7	H8	2	y_8
8	H6	2	y_9
9	H7	1	y_{10}

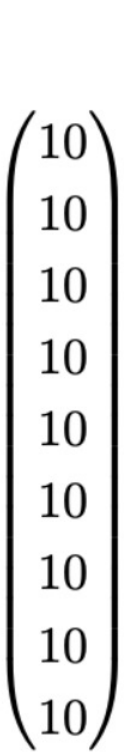
Simulated data

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zh} + \boldsymbol{e}$$



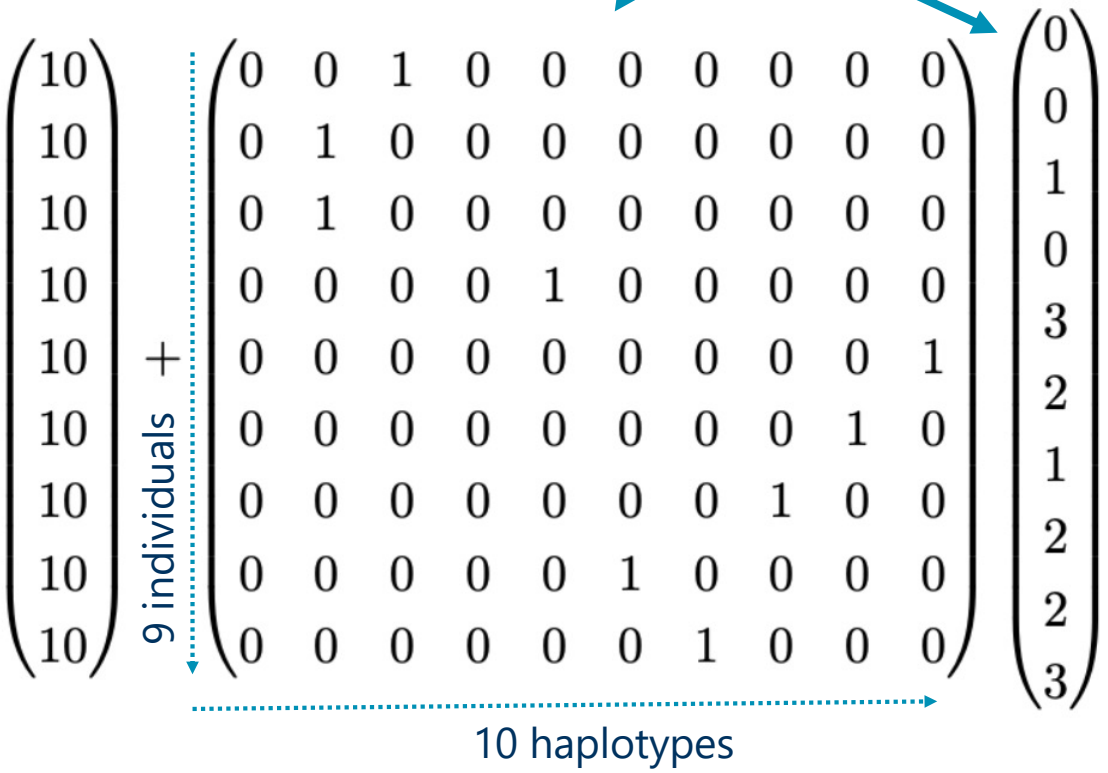
Simulated data

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zh} + \boldsymbol{e}$$



Simulated data

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zh} + \boldsymbol{e}$$



Simulated data

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zh} + \boldsymbol{e}$$

$e \sim N(0,1)$

$$\begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 3 \\ 2 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 0.7588282 \\ -2.2183234 \\ -0.2927689 \\ -0.8998502 \\ 1.2981962 \\ 0.4850921 \\ -0.5006985 \\ -0.1756993 \\ -1.4570039 \end{pmatrix}$$

Simulated data

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zh} + \boldsymbol{e}$$

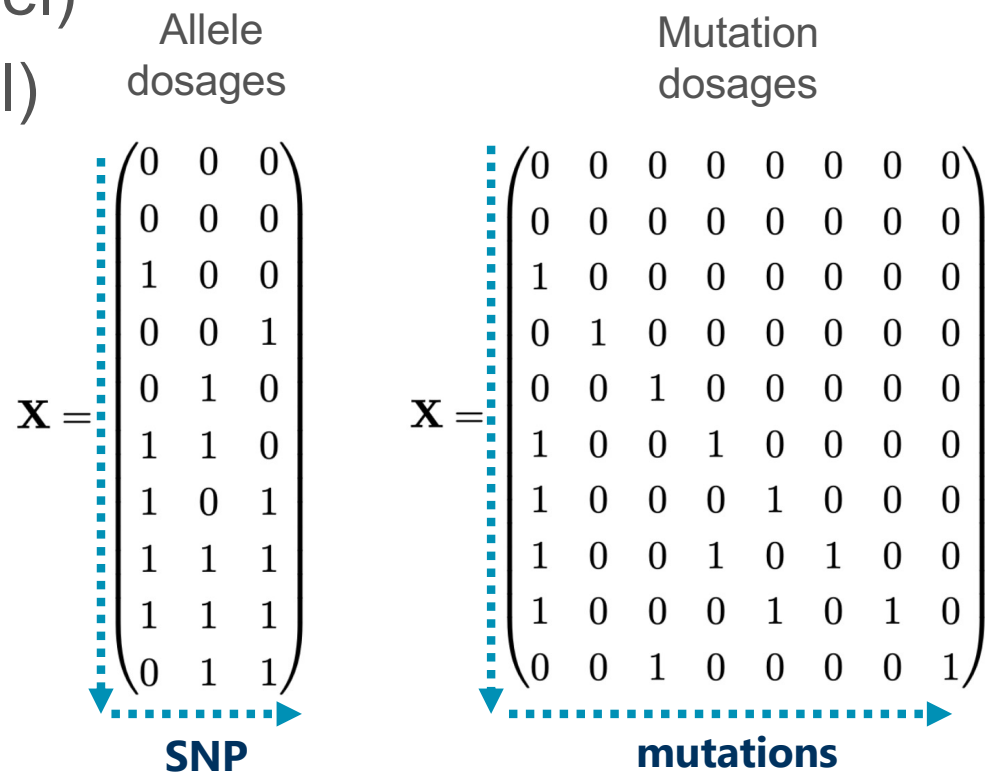
$$\begin{pmatrix} 11.758828 \\ 7.781677 \\ 9.707231 \\ 12.100150 \\ 14.298196 \\ 12.485092 \\ 11.499301 \\ 11.824301 \\ 9.542996 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 3 \\ 2 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 0.7588282 \\ -2.2183234 \\ -0.2927689 \\ -0.8998502 \\ 1.2981962 \\ 0.4850921 \\ -0.5006985 \\ -0.1756993 \\ -1.4570039 \end{pmatrix}$$

Estimating haplotype values

a) SNP-BLUP (marker model)
b) GBLUP (individual model)

Based on

1. Allele dosages
2. Mutation dosages



Estimating haplotype values

a) **SNP-BLUP**

b) GBLUP

Based on

1. **Allele dosages**
2. Mutation dosages

Allele dosages

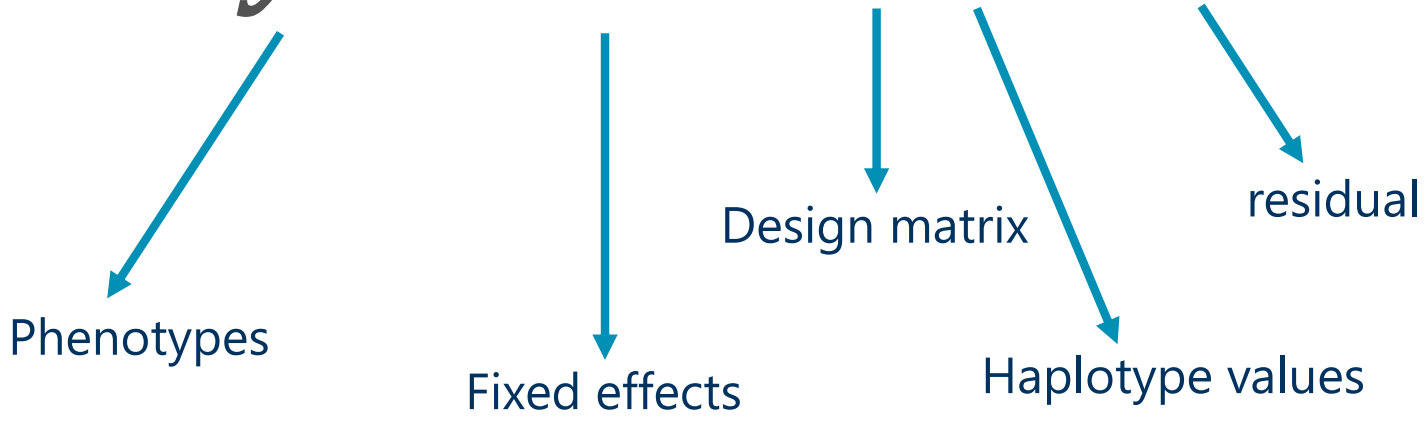
$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Mutation dosages

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{Zh} + \boldsymbol{e}$$



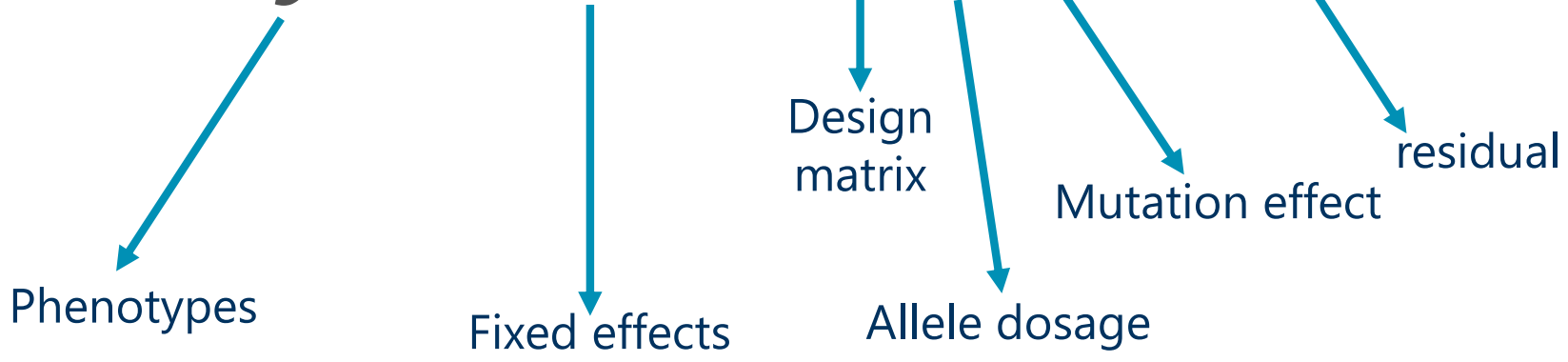
Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{Z}\boldsymbol{h} + \boldsymbol{e}$$

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{\alpha}$$

Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{ZX\alpha} + \boldsymbol{e}$$



Estimating allele substitution effect: SNP-BLUP

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{ZX\alpha} + \boldsymbol{e}$$

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_{\alpha}^{2}) \qquad \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_{e}^{2})$$

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{X}^T\mathbf{Z}^T\mathbf{Z}\mathbf{X} + \mathbf{I}{\sigma_e}^2/{\sigma_\alpha}^2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

Estimating allele substitution effect: SNP-BLUP

a) When **X** is a matrix of allele dosages

$$\begin{pmatrix} 9 & 5 & 5 & 4 \\ 5 & 6 & 3 & 3 \\ 5 & 3 & 6 & 3 \\ 4 & 3 & 3 & 5 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 101.00 \\ 57.11 \\ 62.21 \\ 47.83 \end{pmatrix}$$

Estimating allele substitution effect: SNP-BLUP

a) When **X** is a matrix of allele dosages

$$\begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 9.99 \\ 0.08 \\ 1.77 \\ 0.46 \end{pmatrix}$$

Return allele substitution effect

Estimating haplotype values: SNP-BLUP

Haplotype values are given by $\boldsymbol{h} = \boldsymbol{X}\boldsymbol{\alpha}$

$$\begin{pmatrix} 0.00 \\ 0.00 \\ 0.08 \\ 0.46 \\ 1.77 \\ 1.85 \\ 0.54 \\ 2.32 \\ 2.32 \\ 2.24 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.08 \\ 1.77 \\ 0.46 \end{pmatrix}$$

Estimating haplotype values

a) **SNP-BLUP**

b) GBLUP

Based on

1. Allele dosages
2. **Mutation dosages**

Allele dosages

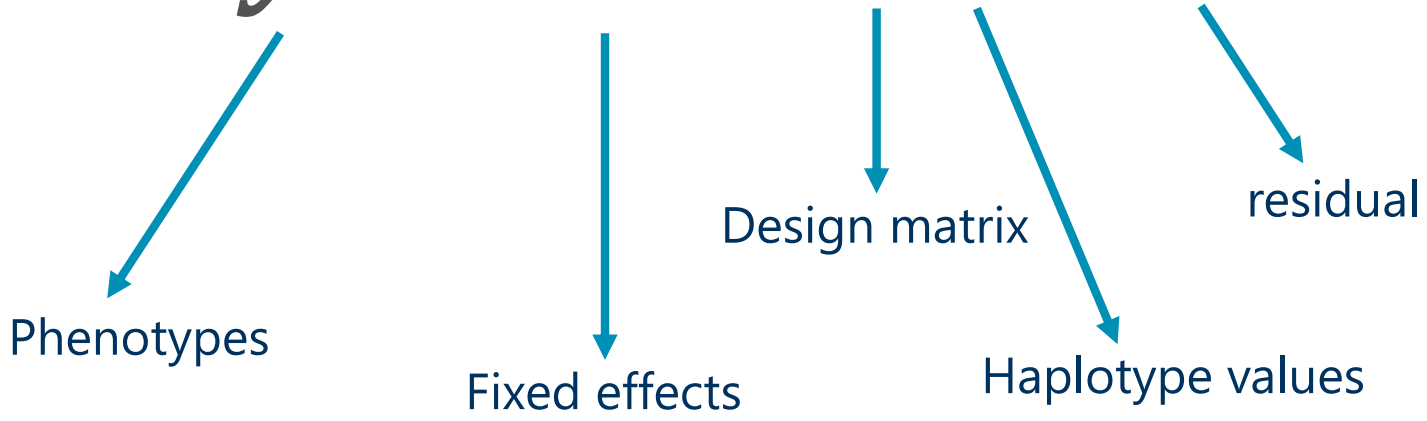
$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Mutation dosages

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{Zh} + \boldsymbol{e}$$



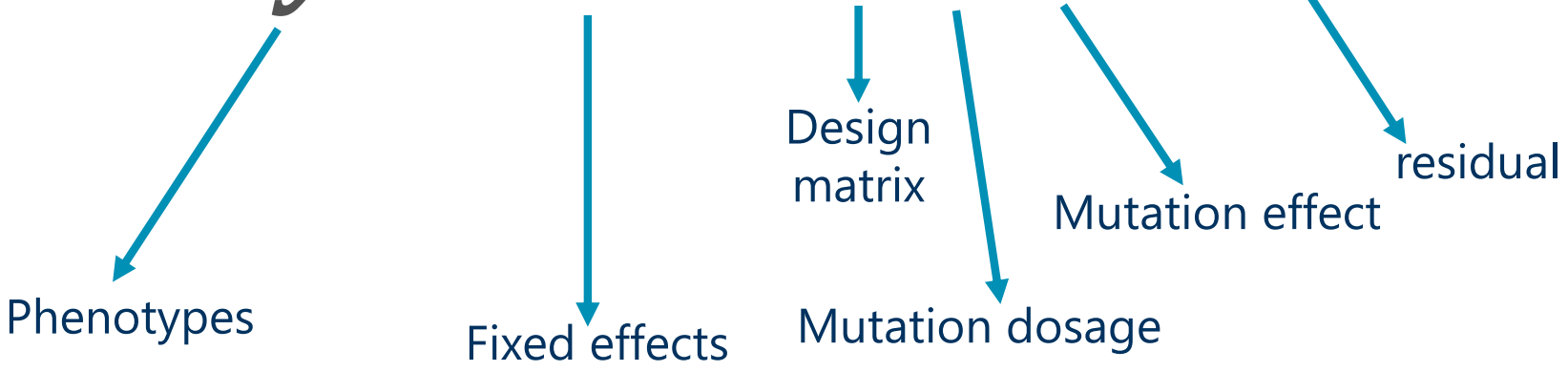
Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{Z}\boldsymbol{h} + \boldsymbol{e}$$

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{\alpha}$$

Estimating haplotype values

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{ZX\alpha} + \boldsymbol{e}$$



Estimating haplotype values: SNP-BLUP

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{ZX\alpha} + \boldsymbol{e}$$

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_{\alpha}^{2}) \qquad \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_{e}^{2})$$

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{X}^T\mathbf{Z}^T\mathbf{Z}\mathbf{X} + \mathbf{I}{\sigma_e}^2/{\sigma_\alpha}^2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

Estimating haplotype values: SNP-BLUP

b) When **X** is a matrix of mutation dosages

$$\begin{pmatrix} 9 & 5 & 0 & 2 & 2 & 2 & 1 & 1 & 1 \\ 5 & 5.0001 & 0e+00 & 0.00000 & 2.00000 & 2.00000 & 1.00000 & 1.00000 & 0.00000 \\ 0 & 0.0000 & 1.e-05 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 2 & 0.0000 & 0e+00 & 2.00001 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00000 \\ 2 & 2.0000 & 0e+00 & 0.00000 & 2.00001 & 0.00000 & 1.00000 & 0.00000 & 0.00000 \\ 2 & 2.0000 & 0e+00 & 0.00000 & 0.00000 & 2.00001 & 0.00000 & 1.00000 & 0.00000 \\ 1 & 1.0000 & 0e+00 & 0.00000 & 1.00000 & 0.00000 & 1.00001 & 0.00000 & 0.00000 \\ 1 & 1.0000 & 0e+00 & 0.00000 & 0.00000 & 1.00000 & 0.00000 & 1.00001 & 0.00000 \\ 1 & 0.0000 & 0e+00 & 1.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00001 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 105.6274 \\ 58.35461 \\ 0.00000 \\ 27.92479 \\ 25.22575 \\ 22.75531 \\ 12.48743 \\ 11.17953 \\ 13.32951 \end{pmatrix}$$

Estimating haplotype values: SNP-BLUP

$$\begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 9.6740235 \\ 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 2.3647471 \\ 1.2022096 \\ -0.2508792 \\ -0.3962383 \\ -1.2657303 \end{pmatrix}$$

Estimating haplotype values: SNP-BLUP

$$\begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 9.6740235 \\ 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 2.3647471 \\ 1.2022096 \\ -0.2508792 \\ -0.3962383 \\ -1.2657303 \end{pmatrix}$$

Mutation effects

Estimating haplotype values: SNP-BLUP

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{\alpha}$$

$$\begin{pmatrix} 0.0000000 \\ 0.0000000 \\ 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 3.0642863 \\ 1.9017487 \\ 2.8134070 \\ 1.5055104 \\ 3.6554916 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 2.3647471 \\ 1.2022096 \\ -0.2508792 \\ -0.3962383 \\ -1.2657303 \end{pmatrix}$$

Estimating haplotype values: SNP-BLUP

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{\alpha}$$

Haplotype values

$$\begin{pmatrix} 0.0000000 \\ 0.0000000 \\ 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 3.0642863 \\ 1.9017487 \\ 2.8134070 \\ 1.5055104 \\ 3.6554916 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.6995392 \\ 0.0000000 \\ 4.9212219 \\ 2.3647471 \\ 1.2022096 \\ -0.2508792 \\ -0.3962383 \\ -1.2657303 \end{pmatrix}$$

Compared

Allele dosage	Mutation dosage	HPV
0.00	0.0000000	0
0.00	0.0000000	0
0.08	0.6995392	1
0.46	0.0000000	0
1.77	4.9212219	3
1.85	3.0642863	2
0.54	1.9017487	1
2.32	2.8134070	2
2.32	1.5055104	2
2.24	3.6554916	3

Estimating haplotype values: GBLUP

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{h} + \boldsymbol{e}$$

$$\boldsymbol{h} \sim N(\mathbf{0}, \boldsymbol{X}\boldsymbol{X}^T\sigma_\alpha^2) \qquad \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$$

Inverse might not exist

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + (\mathbf{X}\mathbf{X}^T + \mathbf{I}\gamma)^{-1}\sigma_e{}^2/\sigma_\alpha{}^2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

Add tiny value

Estimating haplotype values: GBLUP

$$\begin{pmatrix} 9 & 0 & 2 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 57.271 & 0.000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0.000 & 59.271 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.000 & 0.000 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.000 & 0.000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0.000 & 0.000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 105.62742 \\ 0.00000 \\ 19.34801 \\ 10.37355 \\ 0.00000 \\ 14.59528 \\ 12.73832 \\ 11.57578 \\ 12.48743 \\ 11.17953 \\ 13.32951 \end{pmatrix}$$

Estimating haplotype values: GBLUP

$$\begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 9.6740240 \\ 0.0000000 \\ -0.0000006 \\ 0.6995386 \\ 0.0000000 \\ 4.9212214 \\ 3.0642857 \\ 1.9017482 \\ 2.8134065 \\ 1.5055099 \\ 3.6554910 \end{pmatrix}$$

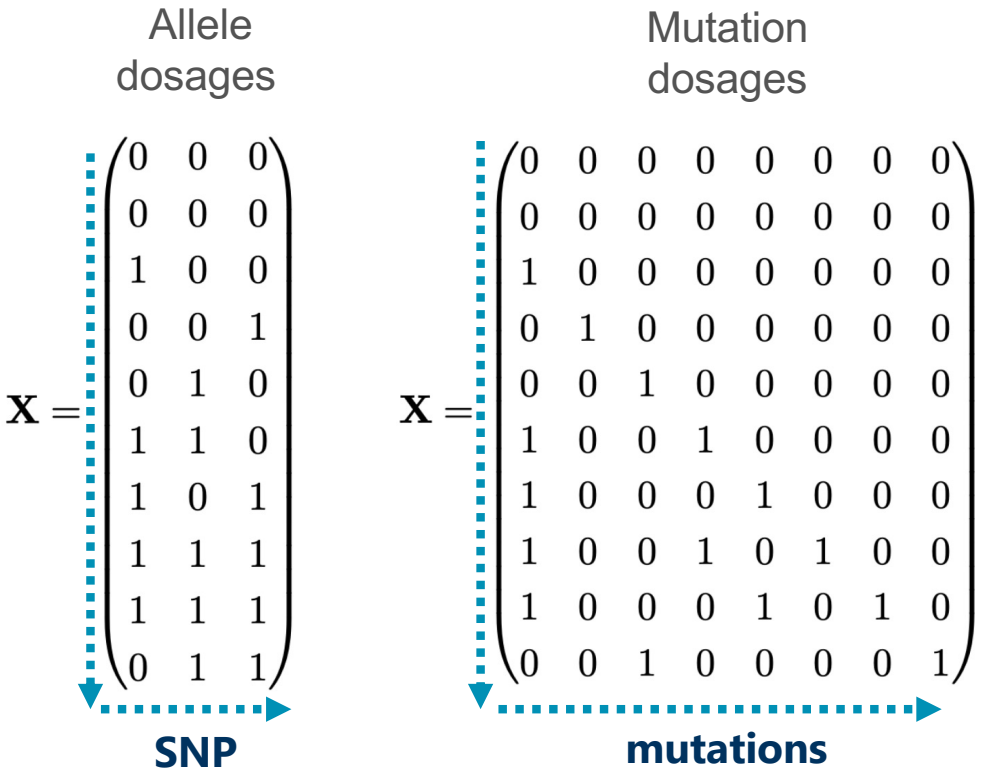
Haplotype values

Compared

SNP-BLUP	GBLUP
0.0000000	0.0000000
0.0000000	−0.0000006
0.6995392	0.6995386
0.0000000	0.0000000
4.9212219	4.9212214
3.0642863	3.0642857
1.9017487	1.9017482
2.8134070	2.8134065
1.5055104	1.5055099
3.6554916	3.6554910

Estimating haplotype values

a) SNP-BLUP
b) GBLUP



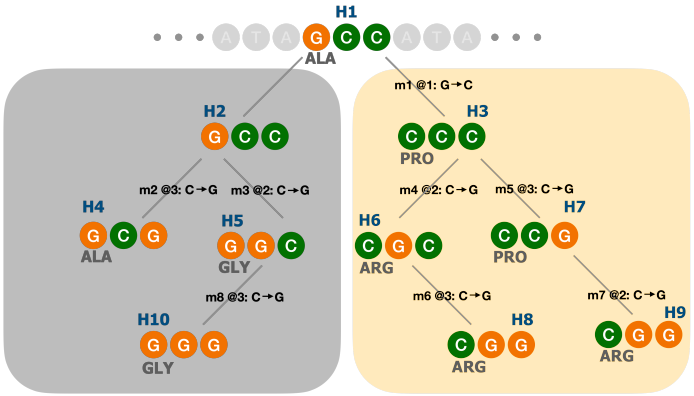
Estimating haplotype values using a tree

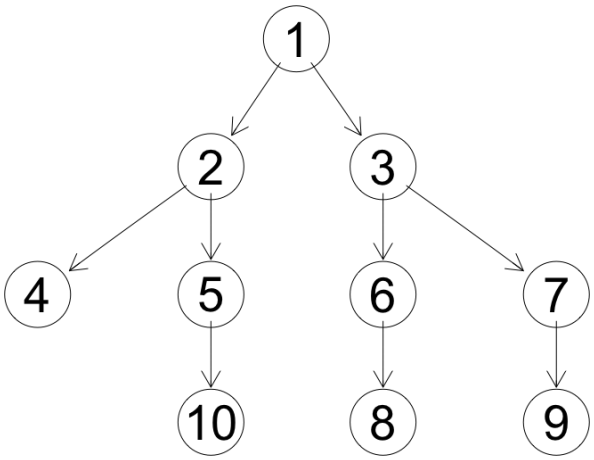
a) SNP-BLUP

b) GBLUP

c) Tree structure

Based on relationships among haplotypes on a tree (TBLUP)





Estimating haplotype values on a gene tree

- Each haplotype value can be explained by its parent plus an "innovation" term (mutation, untyped variants, epistasis, …)

$$h_1 = r_1$$

$$h_1 \sim N\left(0, \sigma_h^2\right)$$

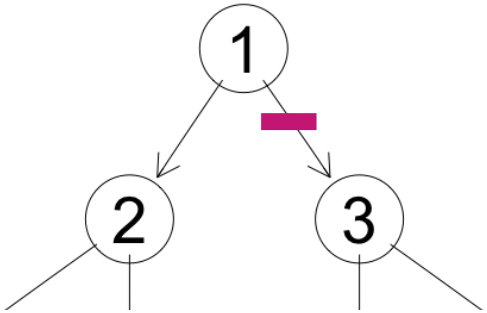
$$h_2 = h_1 + r_2$$

$$h_2 \sim N\left(0, \sigma_h^2\right) \rightarrow h_2 | h_1 \sim N(h_1, \sigma_r^2)$$

$$h_3 = h_1 + m_{1\rightarrow 3} + r_3$$

$$h_3 \sim N\left(0, \sigma_h^2\right) \rightarrow h_3 | h_1 \sim N(h_1, \sigma_m^2 + \sigma_r^2)$$

…



Estimating haplotype values on a gene tree

- Because of the structure this can be solved as P(G)BLUP:

$$\boldsymbol{y} = \mathbf{1}b + \boldsymbol{Zh} + \boldsymbol{e}$$

$$\boldsymbol{h} \sim N(\mathbf{0}, \boldsymbol{H}\sigma_h^2)$$

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + \mathbf{H}^{-1}{\sigma_e}^2/{\sigma_m}^2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

*** σ_m^2 is the mutation variance and we are assuming it the same as σ_α^2**

Estimating haplotype values on a gene tree

- As it is the case for **A**, building **H** can be expensive; but $\boldsymbol{H}^{-1}$ is easy using the generalised Cholesky decomposition:

$$\boldsymbol{H}^{-1} = \mathbf{T}^{T^{-1}}\mathbf{R}^{-1}\mathbf{T}^{-1}$$

Haplotype relationship matrix

Upper triangular

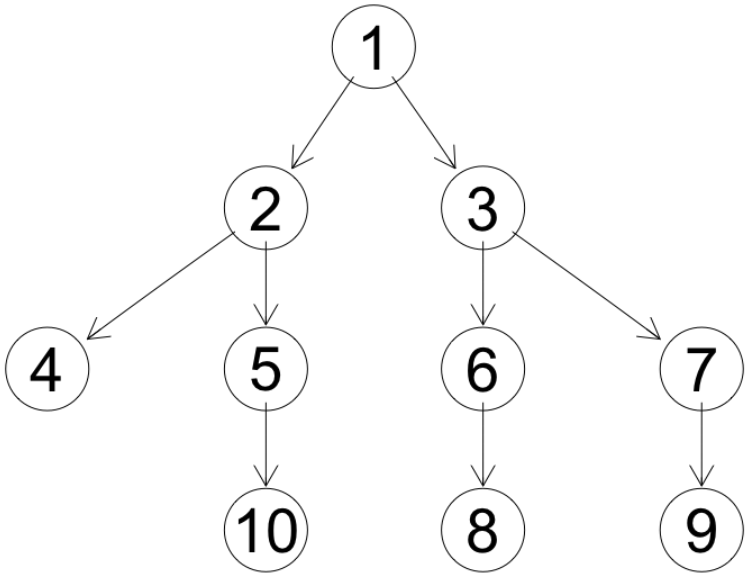
Diagonal Innovation variances (Mendelian sampling in A)

Lower triangular

Estimating haplotype values on a gene tree

- **T** gives the connections between haplotypes

$$\mathbf{T} = \begin{pmatrix}
1 & . & . & . & . & . & . & . & . & . \\
1 & 1 & . & . & . & . & . & . & . & . \\
1 & 0 & 1 & . & . & . & . & . & . & . \\
1 & 1 & 0 & 1 & . & . & . & . & . & . \\
1 & 1 & 0 & 0 & 1 & . & . & . & . & . \\
1 & 0 & 1 & 0 & 0 & 1 & . & . & . & . \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & . & . & . \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & . & . \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & . \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$



Estimating haplotype values on a gene tree

- **R** holds innovation variances (conditional haplotype variance) obtained as:

If h_1 :

$$\boldsymbol{R}_{1,1} = \boldsymbol{x}_1 \boldsymbol{x}_1^T + \gamma/\sigma_m^2$$

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \rightarrow x_1$$

$$\gamma = 1e^{-07}$$

$$\sigma_m^2 = 1$$

Estimating haplotype values on a gene tree

- **R** holds innovation variances (conditional haplotype variance) obtained as:

If h_1 :

$$\boldsymbol{R}_{1,1} = \boldsymbol{x}_1 \boldsymbol{x}_1^T + \gamma/\sigma_m^2$$

For any other h_i :

$$\boldsymbol{R}_{i,i} = n + \gamma/\sigma_m^2$$

Number of mutations
separating h_i from its parent

Estimating haplotype values on a gene tree

$$\mathbf{T} = \begin{pmatrix}
1 & . & . & . & . & . & . & . & . & . \\
1 & 1 & . & . & . & . & . & . & . & . \\
1 & 0 & 1 & . & . & . & . & . & . & . \\
1 & 1 & 0 & 1 & . & . & . & . & . & . \\
1 & 1 & 0 & 0 & 1 & . & . & . & . & . \\
1 & 0 & 1 & 0 & 0 & 1 & . & . & . & . \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & . & . & . \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & . & . \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & . \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix}
1e-07 & . & . & . & . & . & . & . & . & . \\
. & 1e-07 & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & . & . & . \\
. & . & . & . & 1 & . & . & . & . & . \\
. & . & . & . & . & 1 & . & . & . & . \\
. & . & . & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & 1
\end{pmatrix}$$

Estimating haplotype values on a gene tree

$$\boldsymbol{H}^{-1} = \mathbf{T}^{T\,-1}\mathbf{R}^{-1}\mathbf{T}^{-1}$$

$$\mathbf{H}^{-1} = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ . & . & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 \\ . & . & . & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & . & 1 & 0 & 0 & 0 & 0 & -1 \\ . & . & . & . & . & 1 & 0 & -1 & 0 & 0 \\ . & . & . & . & . & . & 1 & 0 & -1 & 0 \\ . & . & . & . & . & . & . & 1 & 0 & 0 \\ . & . & . & . & . & . & . & . & 1 & 0 \\ . & . & . & . & . & . & . & . & . & 1 \end{pmatrix} \begin{pmatrix} 1e+07 & . & . & . & . & . & . & . & . & . \\ . & 1e+07 & . & . & . & . & . & . & . & . \\ . & . & 1 & . & . & . & . & . & . & . \\ . & . & . & 1 & . & . & . & . & . & . \\ . & . & . & . & 1 & . & . & . & . & . \\ . & . & . & . & . & 1 & . & . & . & . \\ . & . & . & . & . & . & 1 & . & . & . \\ . & . & . & . & . & . & . & 1 & . & . \\ . & . & . & . & . & . & . & . & 1 & . \\ . & . & . & . & . & . & . & . & . & 1 \end{pmatrix} \begin{pmatrix} 1 & . & . & . & . & . & . & . & . & . \\ -1 & 1 & . & . & . & . & . & . & . & . \\ -1 & 0 & 1 & . & . & . & . & . & . & . \\ 0 & -1 & 0 & 1 & . & . & . & . & . & . \\ 0 & -1 & 0 & 0 & 1 & . & . & . & . & . \\ 0 & 0 & -1 & 0 & 0 & 1 & . & . & . & . \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & . & . & . \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & . & . \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & . \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Estimating haplotype value on a gene tree

- when solving the system of equations shown before:

$$\begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 9.6740235 \\ 0.0000000 \\ -0.0000001 \\ 0.6995391 \\ -0.0000001 \\ 4.9212218 \\ 3.0642862 \\ 1.9017487 \\ 2.8134070 \\ 1.5055104 \\ 3.6554915 \end{pmatrix}$$

Compared

Allele	Mutation			
SNP-BLUP	SNP-BLUP	GBLUP	TBLUP	HPV
0.00	0.0000000	0.0000000	0.0000000	0
0.00	0.0000000	−0.0000006	−0.0000001	0
0.08	0.6995392	0.6995386	0.6995391	1
0.46	0.0000000	0.0000000	−0.0000001	0
1.77	4.9212219	4.9212214	4.9212218	3
1.85	3.0642863	3.0642857	3.0642862	2
0.54	1.9017487	1.9017482	1.9017487	1
2.32	2.8134070	2.8134065	2.8134070	2
2.32	1.5055104	1.5055099	1.5055104	2
2.24	3.6554916	3.6554910	3.6554915	3

Compared

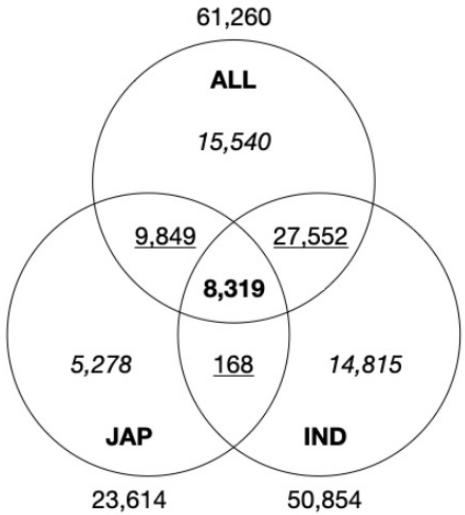
Data Model	Estimation Model	correlation	b_0	b_1	mse
Balanced	**Covariate_allele**	0.898	0.938	0.377	0.096
Balanced	**Covariate_mutation**	0.997	0.979	-0.031	0.004
Balanced	**Animal_allele**	0.898	0.938	0.377	0.096
Balanced	**Animal_mutation**	0.997	0.979	-0.031	0.004
Balanced	**ARG_mutation**	0.997	0.979	-0.031	0.004
Unbalanced	**Covariate_allele**	0.879	1.222	0.264	0.221
Unbalanced	**Covariate_mutation**	0.998	1.043	-0.005	0.003
Unbalanced	**Animal_allele**	0.879	1.222	0.264	0.221
Unbalanced	**Animal_mutation**	0.998	1.043	-0.005	0.003
Unbalanced	**ARG_mutation**	0.998	1.043	-0.005	0.003

Questions?!

Rice & tree sequence exploration

Ines Rebollo

- Rice breeding dataset from Uruguay
- 936 lines (381 Indica & 555 Japonica)
- 22,741 yield records from 828 trials
- 61,260 GBS markers

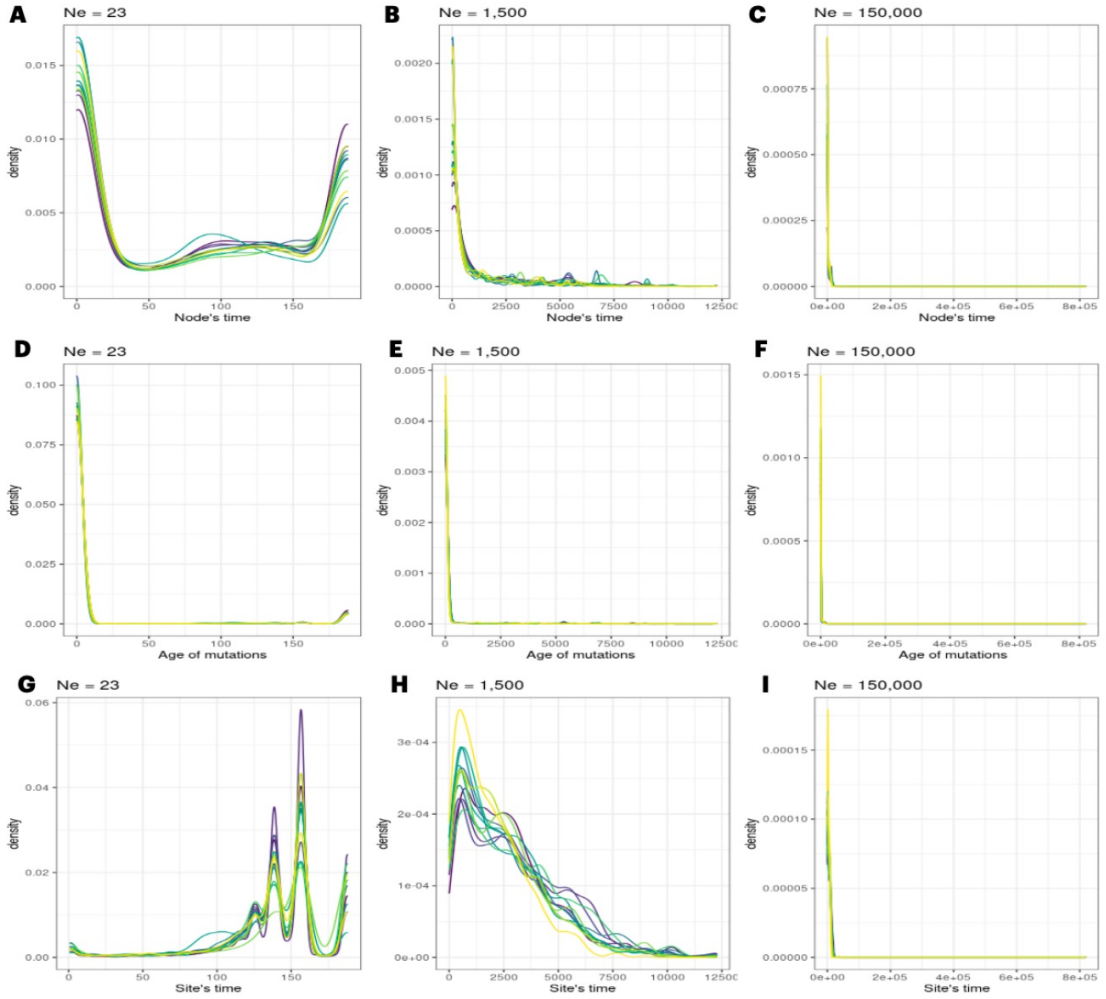


Rice – ancestral alleles & tree sequence

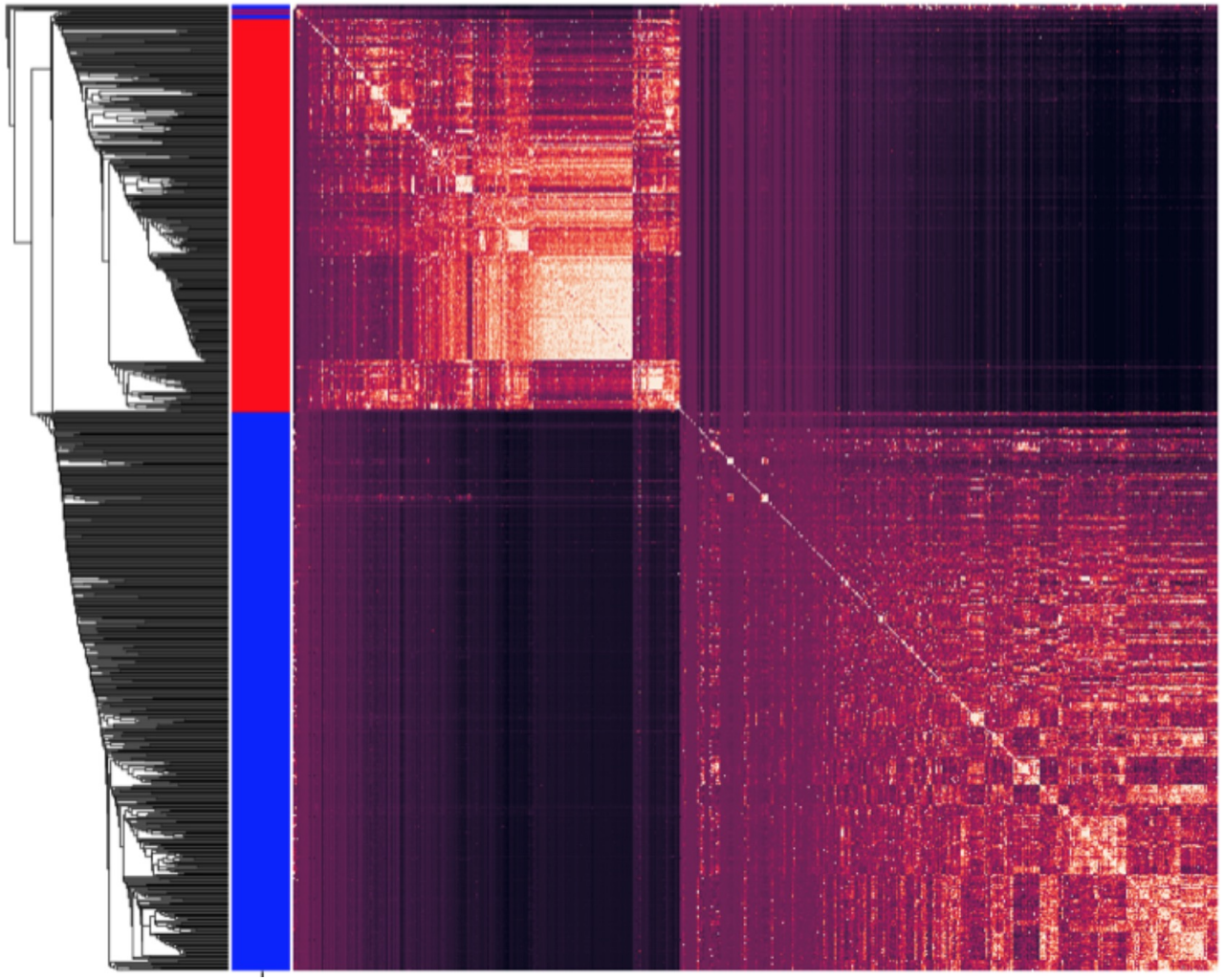
Genomic dataset	Total Sites	Sites inferred	Ancestral = Major	Ancestral = Minor	Ancestral ≠ Minor ≠ Major
ALL	61,260	49,518 (80.8%)	33,485 (67.6%)	15,802 (31.9%)	231 (0.5%)
IND	50,854	40,891 (80.4%)	21,253 (52.0%)	19,363 (47.3%)	275 (0.7%)
JAP	23,614	18,519 (78.4%)	10,905 (58.9%)	7,511 (40.6%)	94 (0.5%)

Chromosome	Nodes	Edges	Trees	Sites	Mutations
TOTAL	95,612	693,524	31,925	61,260	1,031,672
1	11,195	73,825	4,429	8,245	119,709
2	9,584	67,879	3,543	6,549	106,885

Rice - tree sequence



Rice - GNN



Rice – two interesting loci

	Locus			
	DST		NSB & NSBB	
	IND	JAP	IND	JAP
ANC-ANC	30 (4%)	1084 (98%)	208 (27%)	889 (80%)
ANC-ALT	145 (19%)	3 (0%)	542 (71%)	8 (1%)
ALT-ANC	585 (77%)	23 (2%)	8 (1%)	213 (19%)
ALT-ALT	2 (0%)	0 (0%)	4 (1%)	0 (0%)

IND - *indica*; JAP - *japonica*.

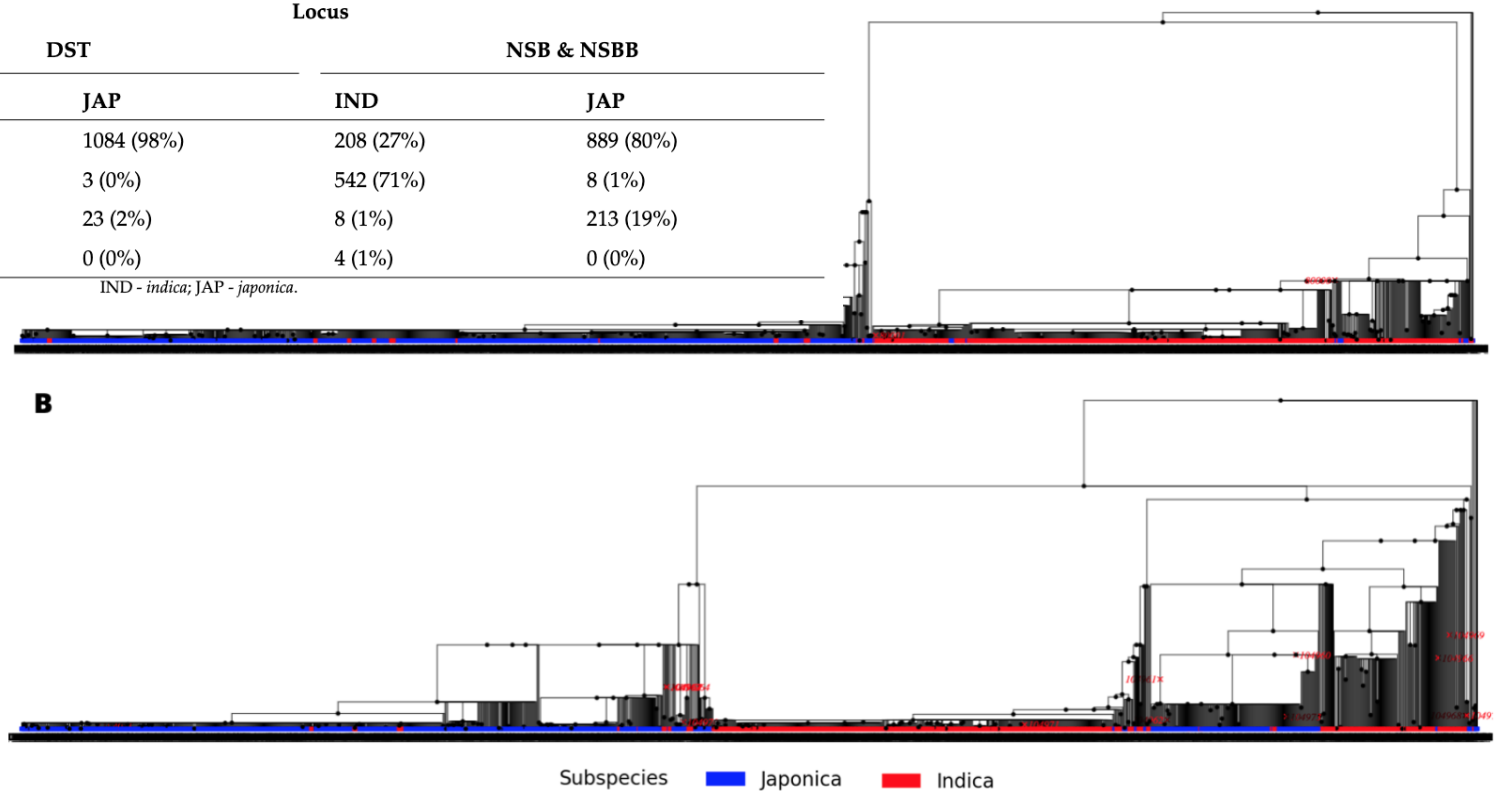
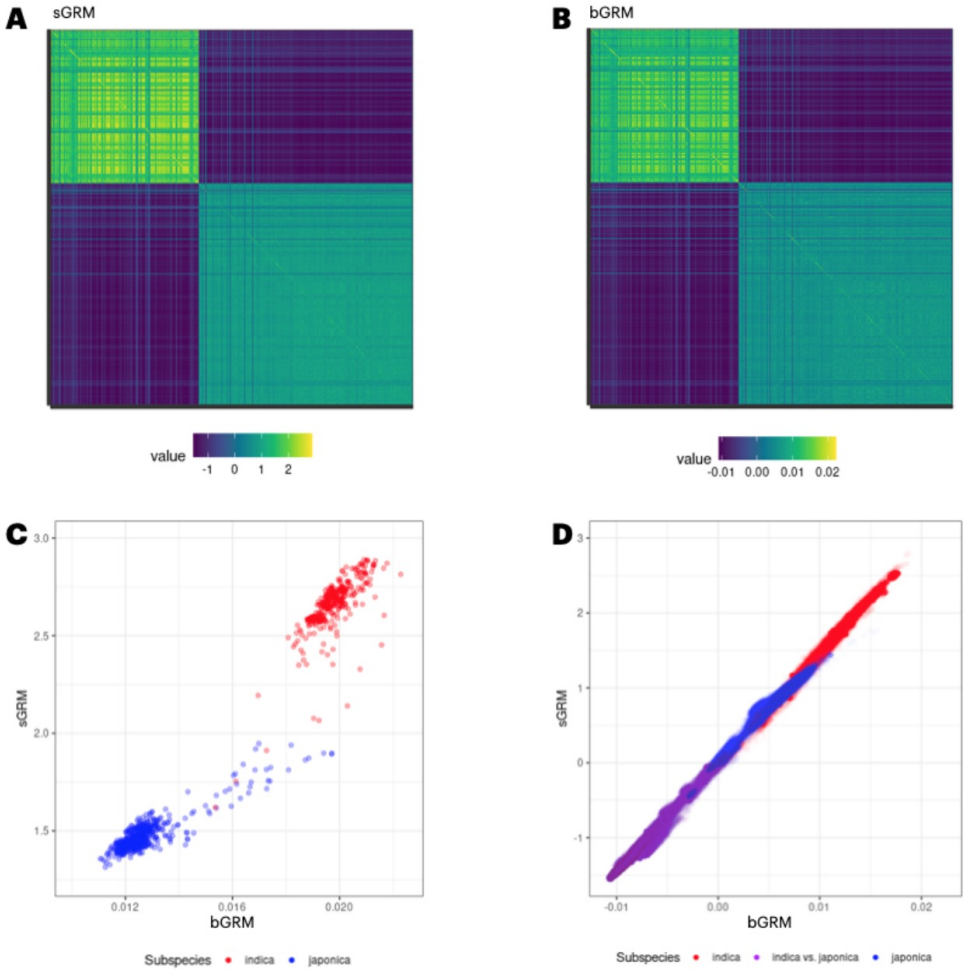


Figure 2 Differential local tree structure at the genome positions in (A) the Drought and Salt Tolerance gene (*DST*), associated with panicle length only in *japonica* and (B) a locus associated with number of panicle secondary branches and number of spikelets per panicle secondary branch in both *indica* and *japonica*. The tree in (A) shows a very clear and deep separation between *indica* and *japonica* while the tree in (B) shows segregation in both *indica* and *japonica*.

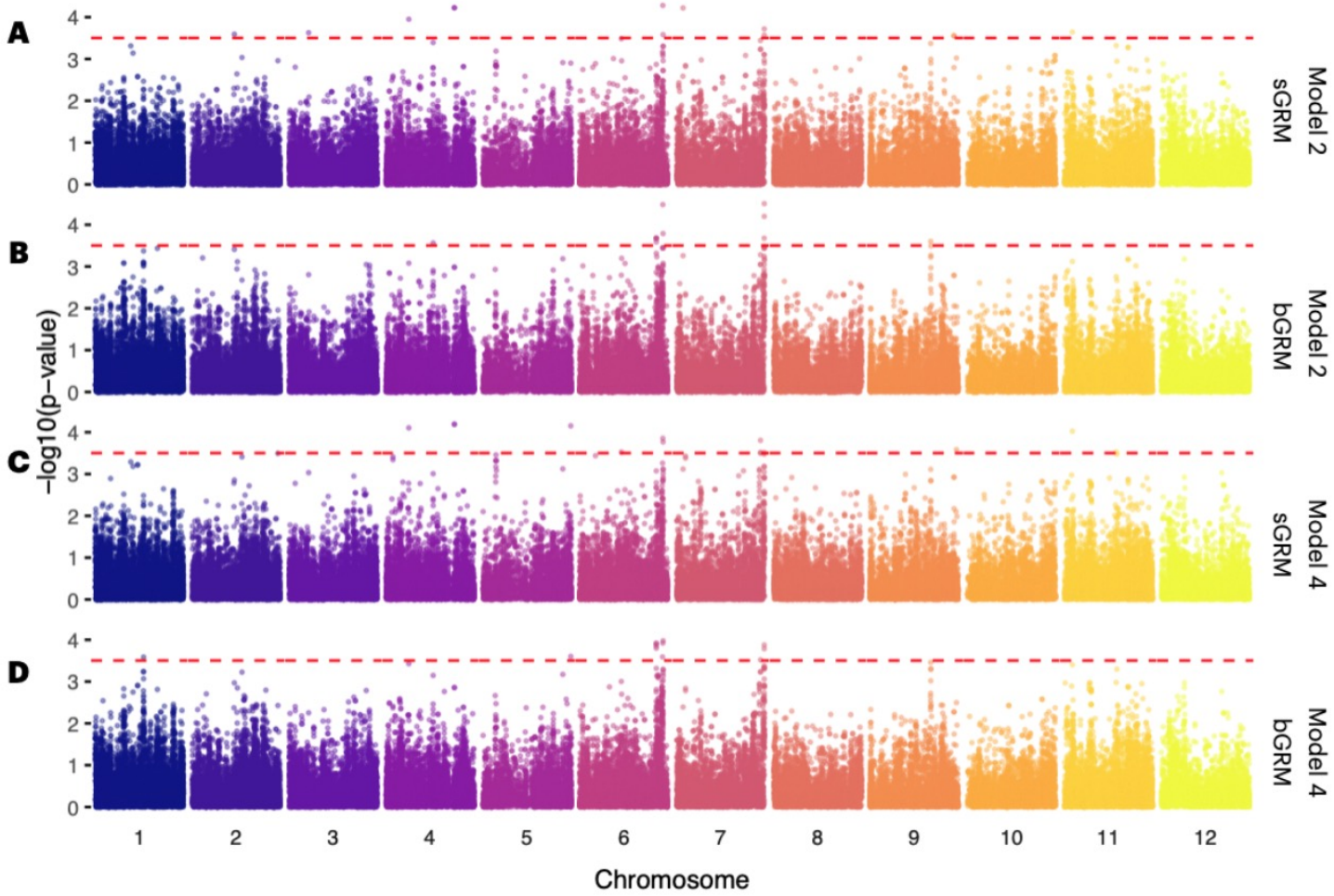
ts.relatedness() → site- & branch-based NRM



Cross-validation accuracy (meh)

Scenario	Relationship matrix	Number of individuals				Model 2	
		Training set		Prediction set		Main effects	
		IND	JAP	IND	JAP	IND	JAP
$CV_{IJ \rightarrow IJ}$	PRM	262	264	65	66	**0.70 (0.03)**	0.40 (0.02)
	SRM					0.68 (0.01)	0.47 (0.01)
	BRM					**0.70 (0.01)**	**0.48 (0.01)**
$CV_{I \rightarrow J}$	PRM	262			66		**0.15 (0.03)**
	SRM						0.11 (0.03)
	BRM						0.10 (0.02)
$CV_{J \rightarrow I}$	PRM		264	65		-0.33 (0.03)	
	SRM					**0.43 (0.03)**	
	BRM					0.20 (0.04)	
$CV_{I \rightarrow I}$	PRM	262		65		0.70 (0.01)	
	SRM					0.69 (0.01)	
	BRM					**0.71 (0.01)**	
$CV_{J \rightarrow J}$	PRM		264		66		0.42 (0.02)
	SRM						**0.49 (0.02)**
	BRM						**0.49 (0.02)**

Rice - GWAS


A
B
C
D
Model 2 sGRM
Model 2 bGRM
Model 4 sGRM
Model 4 bGRM
−log10(p−value)
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome

Rice – GWAS peak trees and haplotypes

Local tree	SNP	z marker effect
tree 1	S6_29476724	4.15
	S6_29476748	0.06
	S6_29476763	3.31
	S6_29476787	-3.21
tree 2	S6_29480408	-2.94
tree 3	S6_29480471	2.68
	S6_29480530	2.63
tree 4	S6_29516992	-0.32
tree 5	S6_29517004	2.59
tree 6	S6_29531546	3.03
tree 7	S6_29557666	3.29
	S6_29557756	0.82
	S6_29557803	2.80
	S6_29561680	3.20
tree 8	S6_29561694	3.77

Local tree	Haplotype	IND	JAP	z-haplotype effect
tree 1 chr6:29,476,724-29,476,787	0000	55 (7%)	645 (58%)	0.0
	0001	2 (0%)	0 (0%)	-3.21
	0010	3 (1%)	1 (0%)	3.31
	0011	519 (68%)	32 (3%)	0.33
	0100	173 (23%)	66 (6%)	0.06
	0101	2 (0%)	0 (0%)	-1.88
	1010	0 (0%)	2 (0%)	4.11
	1011	8 (1%)	364 (33%)	4.01
tree 2 chr6:29,480,408	0	227 (30%)	710 (64%)	0.0
	1	535 (70%)	400 (36%)	-2.94

Rice – GWAS peak trees and haplotypes

Local tree	SNP	z marker effect
tree 1	S6_29476724	4.15
	S6_29476748	0.06
	S6_29476763	3.31
	S6_29476787	-3.21
tree 2	S6_29480408	-2.94
tree 3	S6_29480471	2.68
	S6_29480530	2.63
tree 4	S6_29516992	-0.32
tree 5	S6_29517004	2.59
tree 6	S6_29531546	3.03
tree 7	S6_29557666	3.29
	S6_29557756	0.82
	S6_29557803	2.80
	S6_29561680	3.20
tree 8	S6_29561694	3.77

Haplotype	IND	JAP	z-haplotype effect
000000000000000	6 (0.8%)	622 (56%)	0.00
001111110111011	499 (65.5%)	23 (2.1%)	3.25
101111101110111	4 (0.5%)	335 (30.2%)	3.96
010000000111011	167 (21.9%)	9 (0.8%)	3.02
000000000111011	49 (6.4%)	2 (0.2%)	3.34
010000000111010	0 (0.0%)	46 (4.1%)	2.57
001111100111011	4 (0.5%)	2 (0.2%)	3.55
001111110110111	6 (0.8%)	0 (0.0%)	3.49
101111100110111	2 (0.3%)	4 (0.4 %)	4.01
010000000000000	0 (0.0%)	5 (0.5 %)	0.06
Other 51	25 (3.3%)	62 (5.6%)	-

Learning objectives

- Showcase two approaches to modelling haplotype effects for a non-recombining region (real & simulation)
  → interesting implications (waiting on real data analysis)

- Showcase tree-sequence results for a rice dataset
  → the dataset is too small (genomic models similar results to pedigree model)
  → issues with tree sequence inference? (too many mutations!?)

Questions?!

Gene tree- & tree sequence-based linear mixed models

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-09