Estimating haplotype and mutation effects in the context of genome sequence via ancestral recombination graphs

Gabriela Mafra Fortuna, Jana Obsteter, Ajda Moskric, Gregor Gorjanc

2024 EAAP, Florence

@GregorGorjanc @HighlanderLab

Ancestral Recombination Graph (ARG)

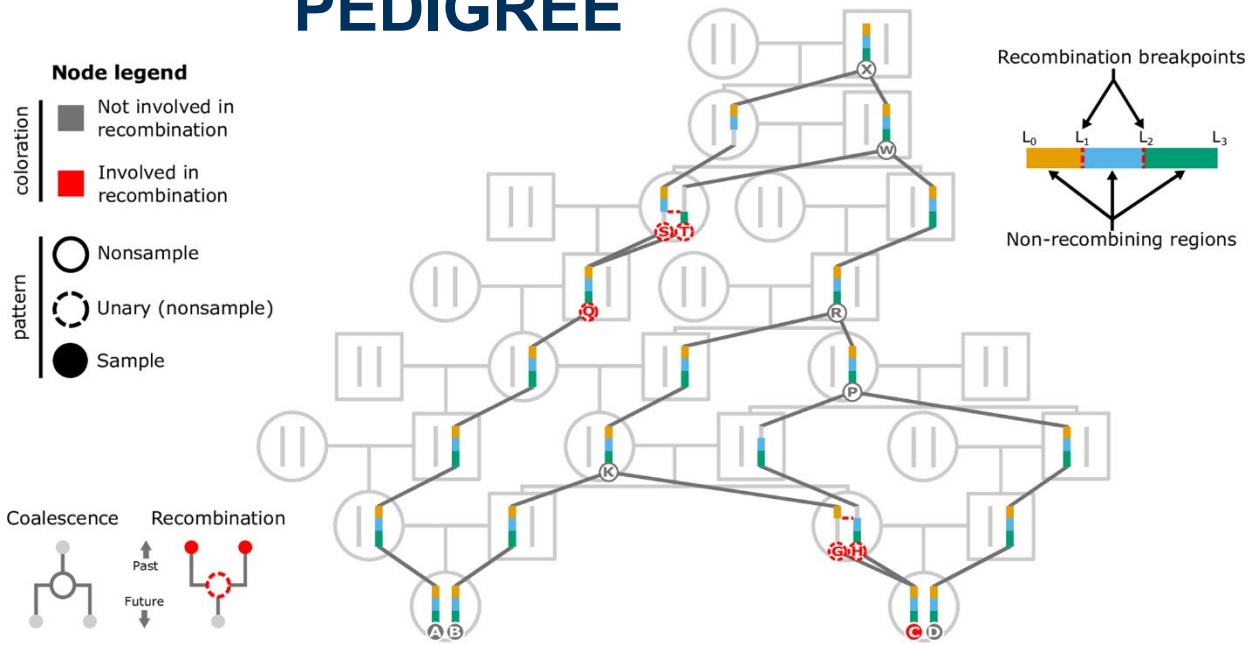
- **Representation of DNA variation by focusing on DNA events**
  - *branch/coalescent events* (where DNA molecule split/merge),
  - *mutation events* (where DNA molecule changes), and
  - *recombination events* (where DNA molecules exchange)

- **Insight into evolutionary history**
  (time and place of DNA events, demography, divergence, introgression, …)

- **Succinct data representation!**
  (can handle large scale WGS datasets)

How do ARGs look like?

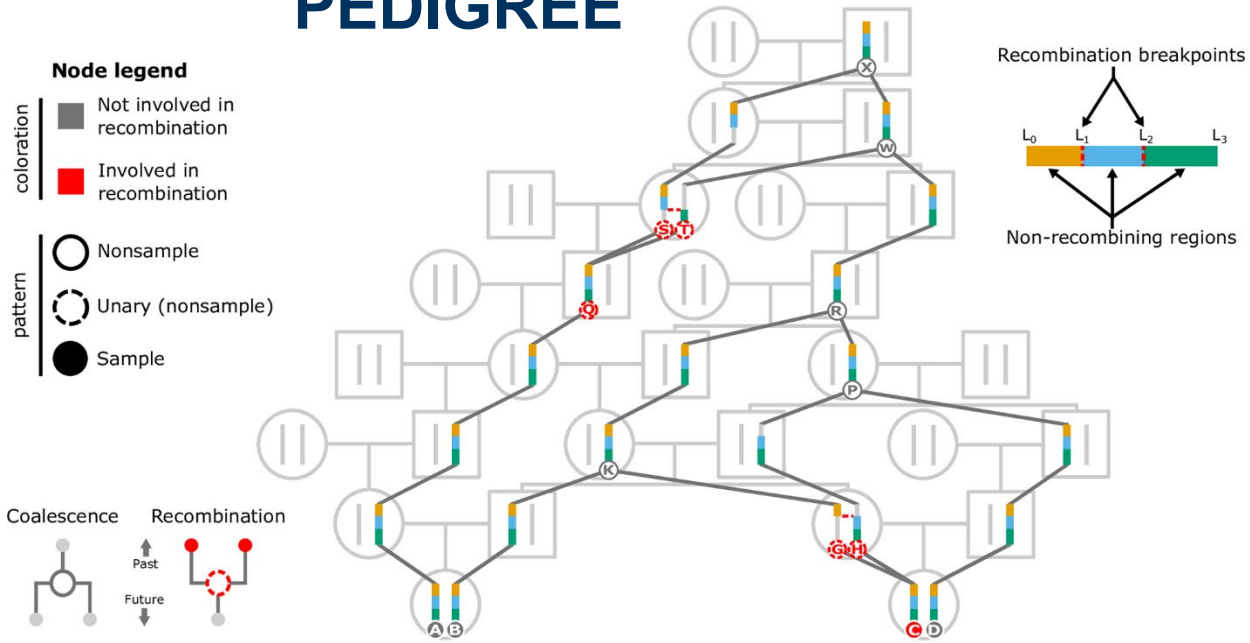
PEDIGREE



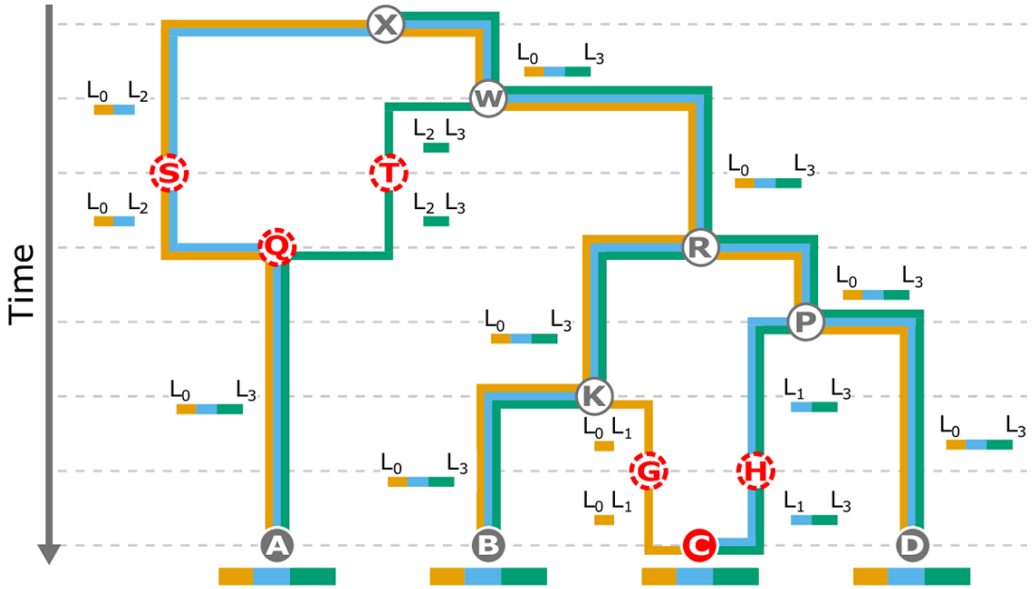
Lewanski et al. (2024)

How do ARGs look like?

PEDIGREE



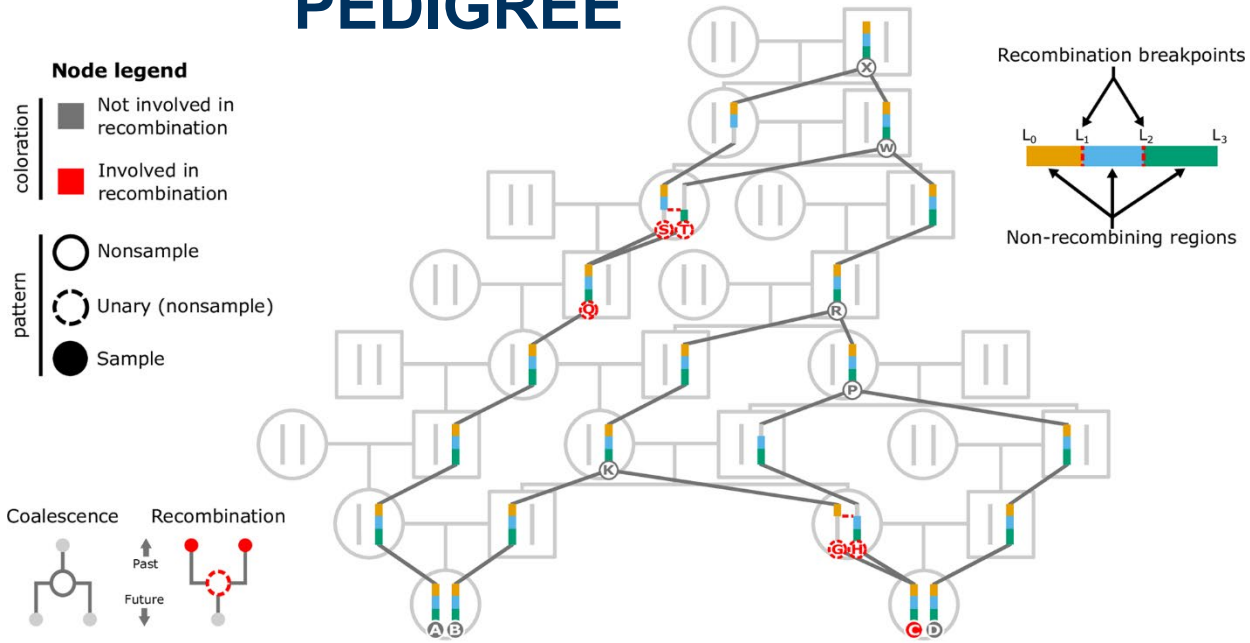
ARG



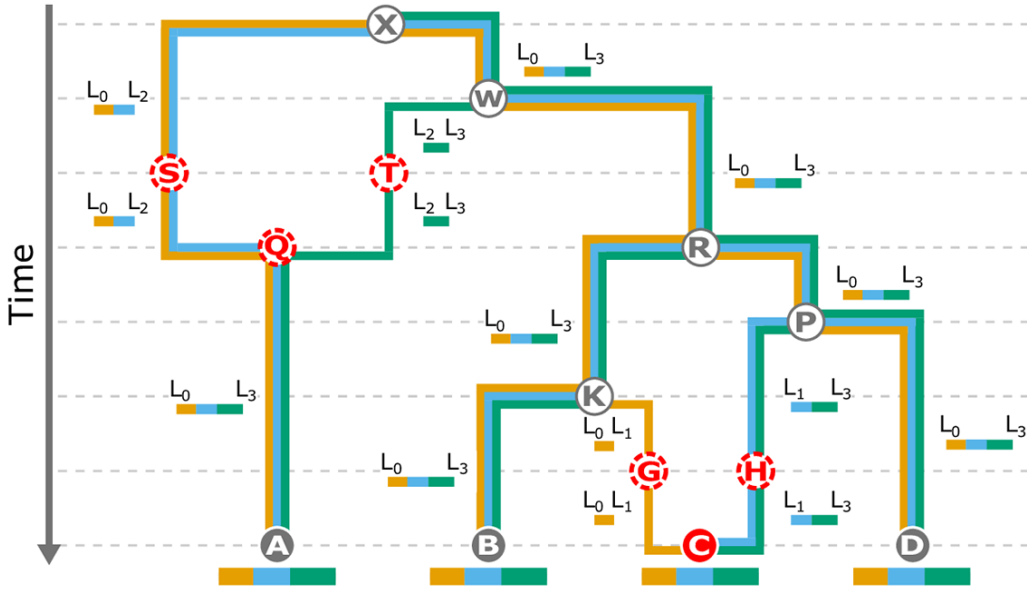
Lewanski et al. (2024)

How do ARGs look like?

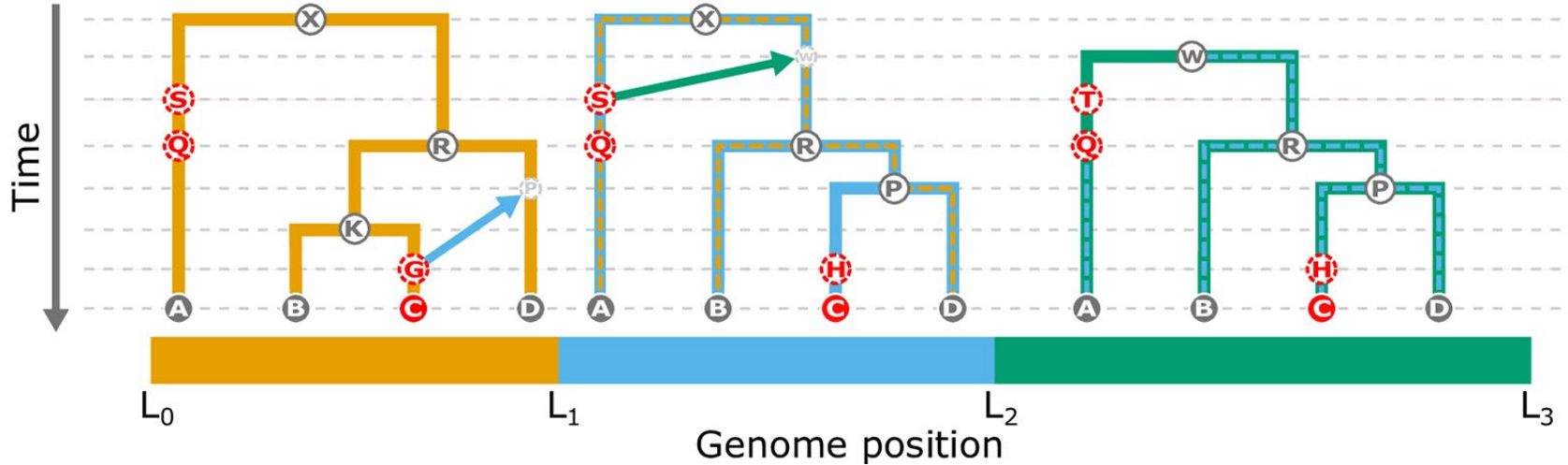
PEDIGREE



ARG



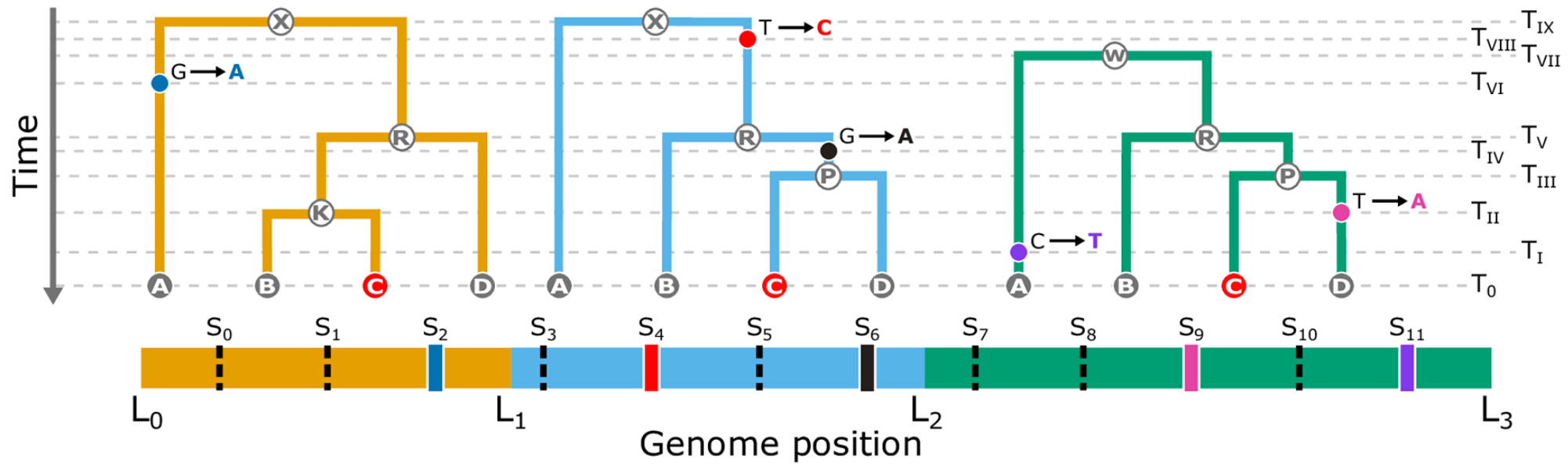
LOCAL TREES



Lewanski et al. (2024)

Genotypes → Haplotypes → ARG

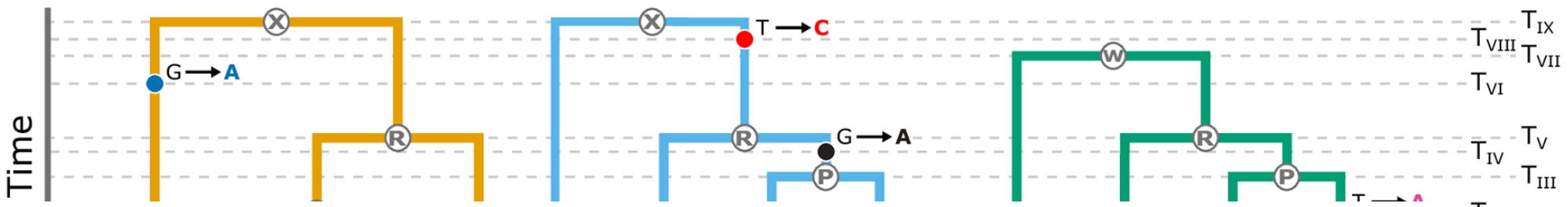
Sample	Site ID 0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	T	A	C



Lewanski et al. (2024)

Genotypes → Haplotypes → ARG

Sample	Site ID 0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	T	A	C



Nodes

Node ID	Age
A	T_0
B	T_0
C	T_0
D	T_0
K	T_{II}
P	T_{III}
R	T_V
W	T_{VII}
X	T_{IX}

Edges

Genome start	Genome end	Parent	Child
L_0	L_2	X	A
L_0	L_1	K	B
L_0	L_1	K	C
L_0	L_2	X	R
L_0	L_1	R	D
L_0	L_1	R	K
L_1	L_3	R	B
L_1	L_3	R	P
L_1	L_3	P	C
L_1	L_3	P	D
L_2	L_3	W	A
L_2	L_3	W	R

Sites

Site ID	Position	Ancestral
0	S_0	G
1	S_1	T
2	S_2	G
3	S_3	G
4	S_4	T
5	S_5	C
6	S_6	G
7	S_7	T
8	S_8	A
9	S_9	T
10	S_{10}	A
11	S_{11}	C

Mutations

Mutation ID	Site	Node	Derived	Age
0	2	A	A	T_{VI}
1	4	R	C	T_{VIII}
2	6	P	A	T_{IV}
3	9	D	A	T_{II}
4	11	A	T	T_I

Lewanski et al. (2024)

Ancestral Recombination Graph (ARG)

- **Representation of DNA variation by focusing on DNA events**
  - *branch/coalescent events* (where DNA molecule split/merge),
  - *mutation events* (where DNA molecule changes), and
  - *recombination events* (where DNA molecules exchange)

- **Insight into evolutionary history**
  (time and place of DNA events, demography, divergence, introgression, …)

- **Succinct data representation!**
  (can handle large scale WGS datasets)

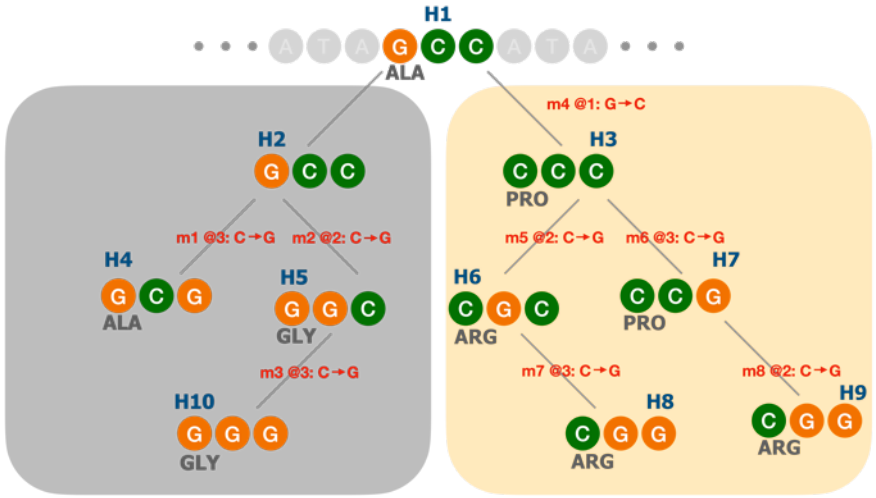
- **AIM: Quantitative genetic modelling based on ARGs!?**

TODAY: QG model for non-recombining DNA

Model options:

a) SNP-BLUP

b) G-BLUP



Based on

1. Allele dosages

Allele dosages

$$\mathbf{X} = \begin{array}{c} \\ \text{Individual} \end{array} \overset{\text{SNP}}{\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}$$

TODAY: QG model for non-recombining DNA

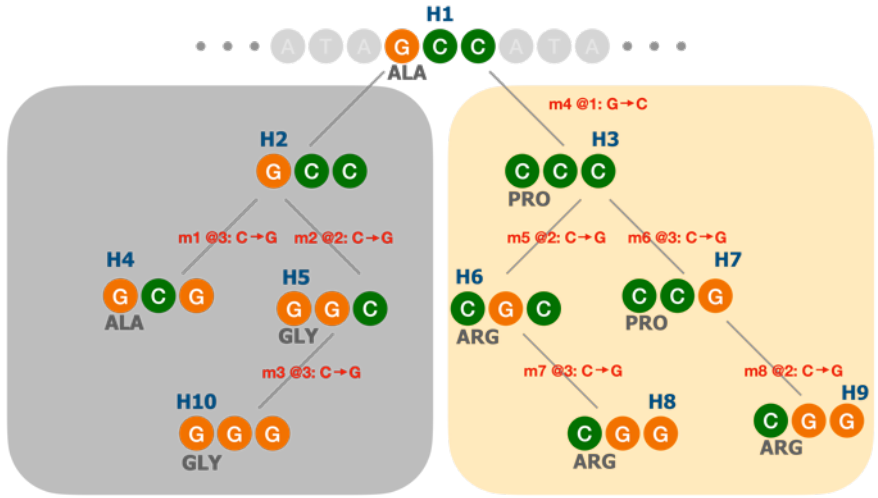
Model options:

a) SNP-BLUP

b) G-BLUP

Based on

1. Allele dosages
2. Mutation dosages



Allele dosages

$$\mathbf{X} = \begin{matrix} \\ \text{Individual} \end{matrix} \begin{matrix} \text{SNP} \\ \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$\mathbf{X} = \begin{matrix} \\ \text{Individual} \end{matrix} \begin{matrix} \text{Mutation} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

Mutation dosages

TODAY: QG model for non-recombining DNA

Model options:

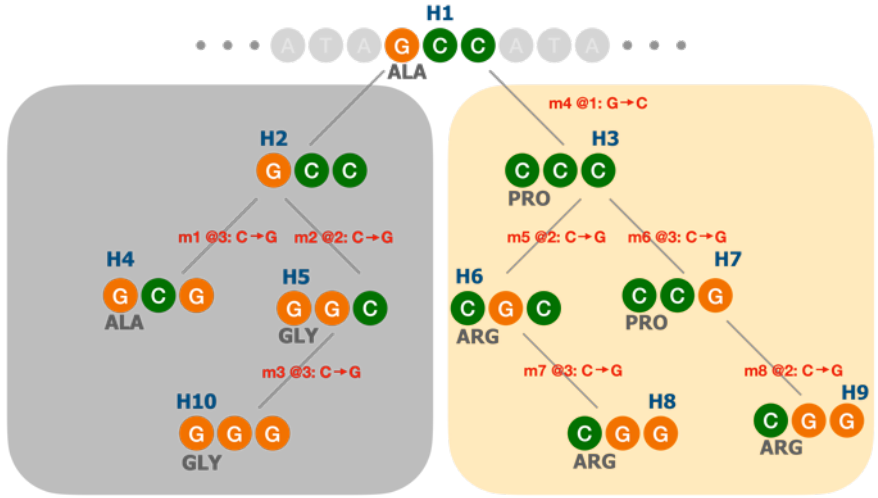
a) SNP-BLUP

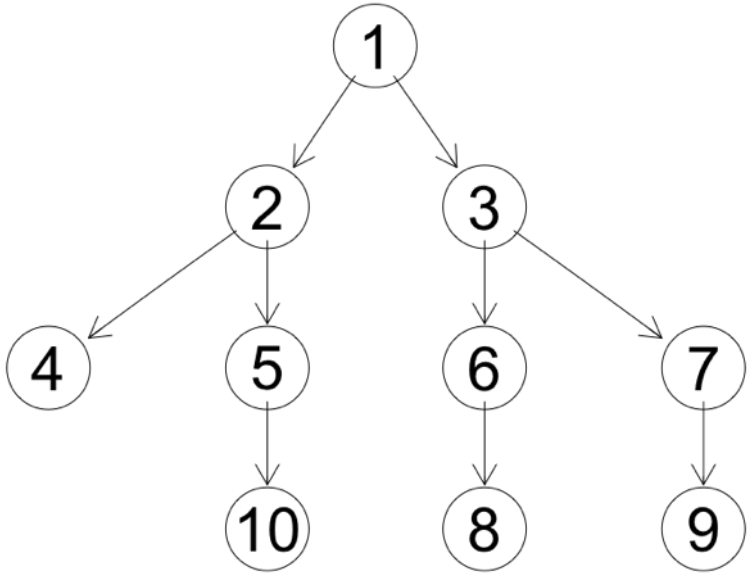
b) G-BLUP

c) Tree-BLUP / ARG-BLUP

Based on

1. Allele dosages
2. Mutation dosages
3. Tree / ARG structure

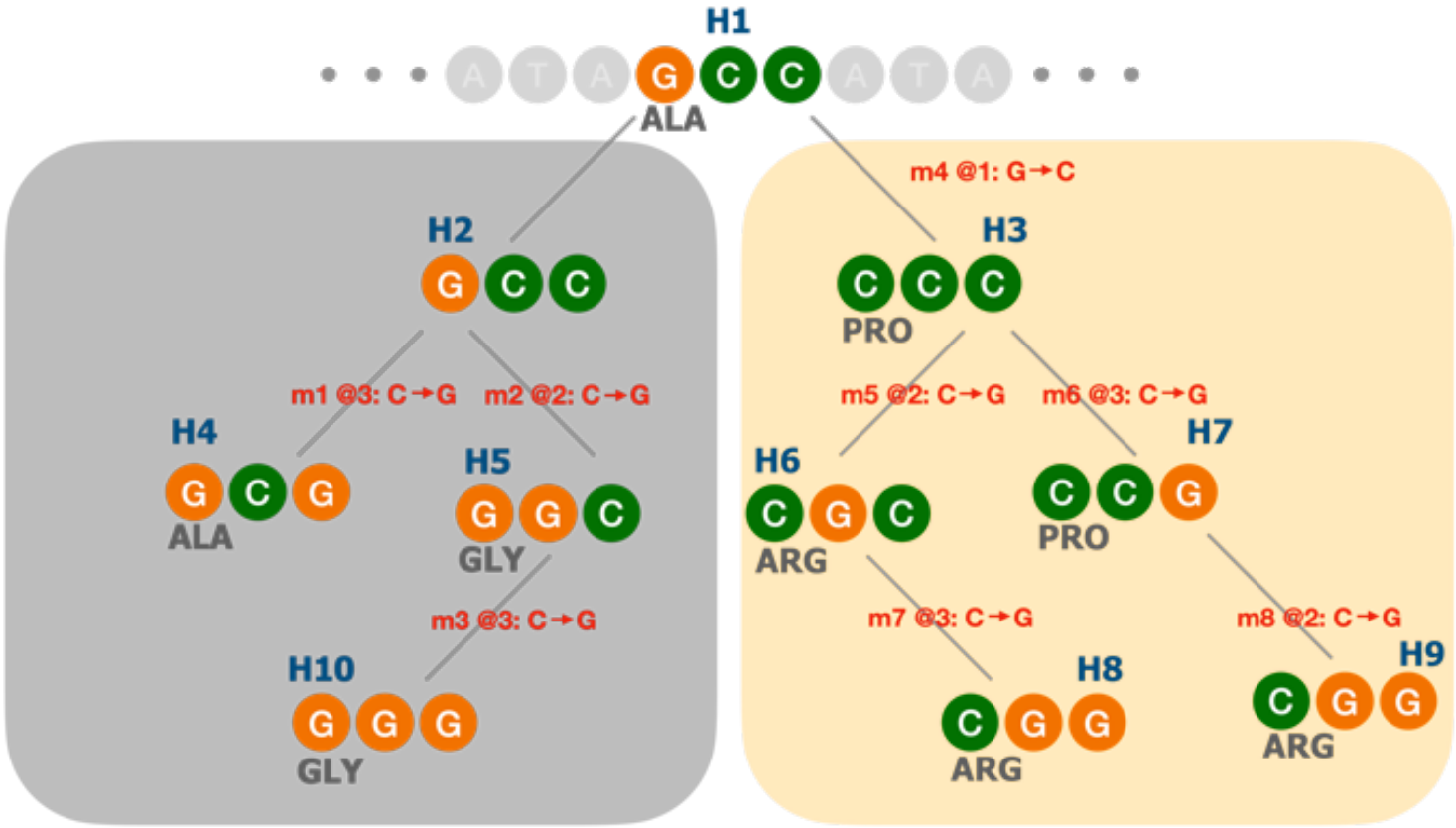




Haplotype tree model

Haplotype's value can be modelled as the value of it's parental haplotype plus an "innovation" term

(mutation, structural variants, unobserved biology, epistasis, …)



$$h_1 \sim N(0, \sigma_h^2)$$
$$h_2 \sim N(0, \sigma_h^2)$$
$$h_2 | h_1 \sim N(h_1, \sigma_m^2)$$

…

$$h_2 = h_1 + e_2$$
$$h_3 = h_1 + m_{1\to3} + e_3$$
$$h_4 = h_2 + (m_{2\to4} + e_4)$$

Tree-BLUP

- Phenotype model

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

- Haplotype tree model

$$\boldsymbol{h} \sim N(\mathbf{0}, \boldsymbol{H}(\sigma_m^2 + \sigma_\epsilon^2))$$

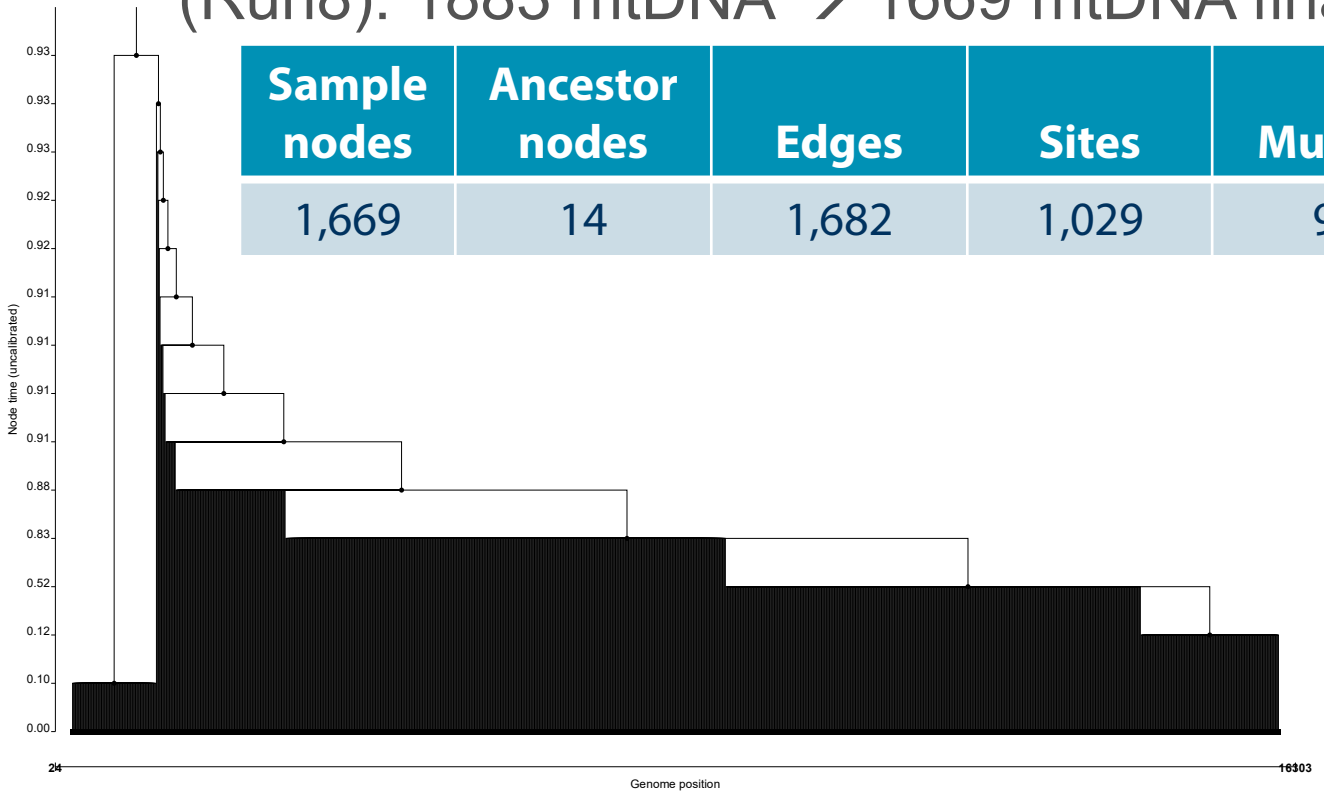
- Covariance coef. matrix between haplotypes **H** has a simple inverse (precision matrix) due to DNA tree structure

- Can fit with most linear mixed model software (ASReml, blupf90, INLA, Mix99, MixBLUP, …)

Proof of concept – real data + simulations

- **Real data**
  - Cattle mtDNA
  - Inferred mtDNA ancestral alleles
  - Inferred mtDNA tree for 1000 Bull Genomes (Run8): 1883 mtDNA → 1669 mtDNA final

Sample nodes	Ancestor nodes	Edges	Sites	Mutations
1,669	14	1,682	1,029	9,598



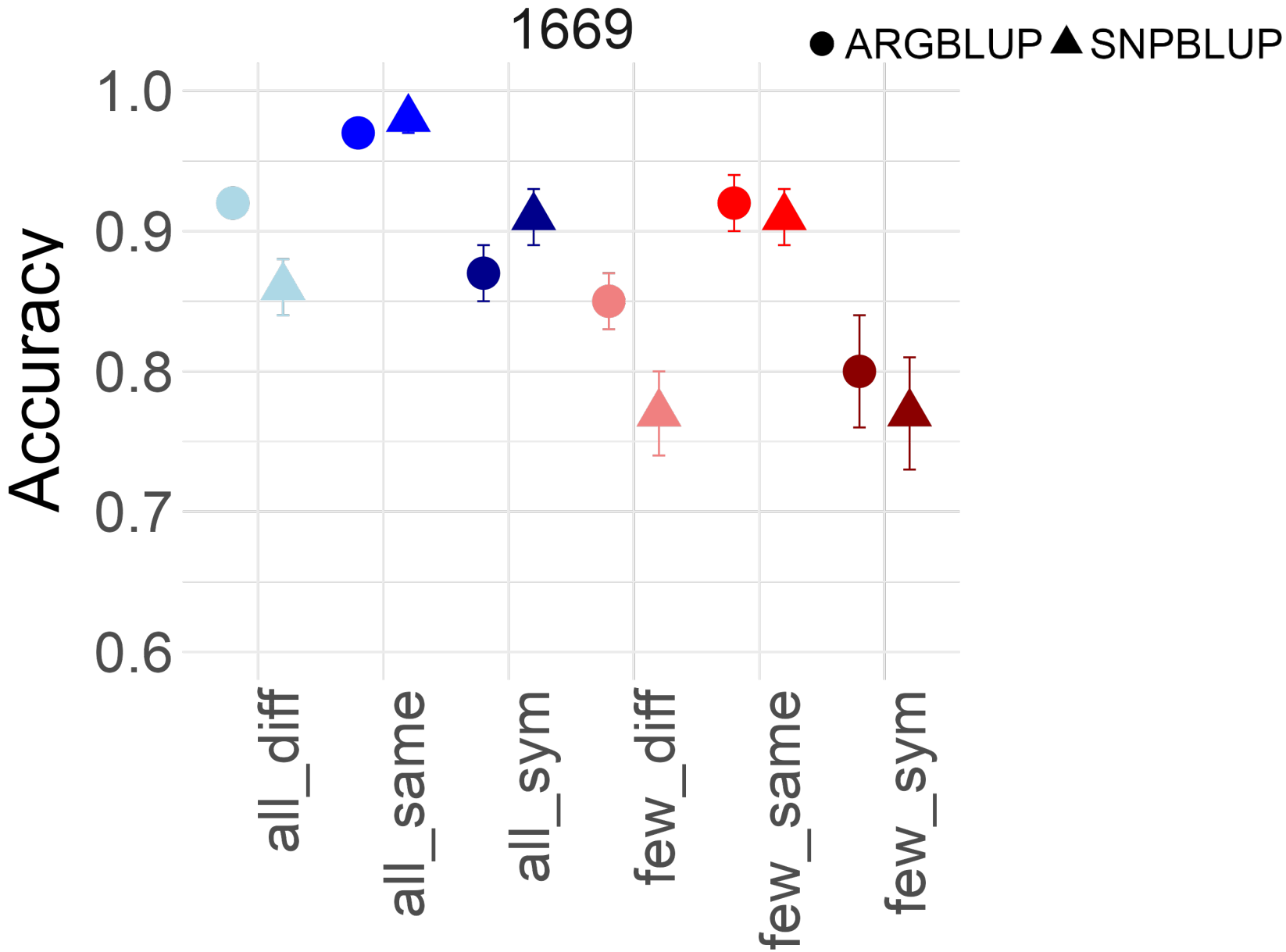
Simulations

ALL all mutations are causative
FEW 10% of mutations are causative

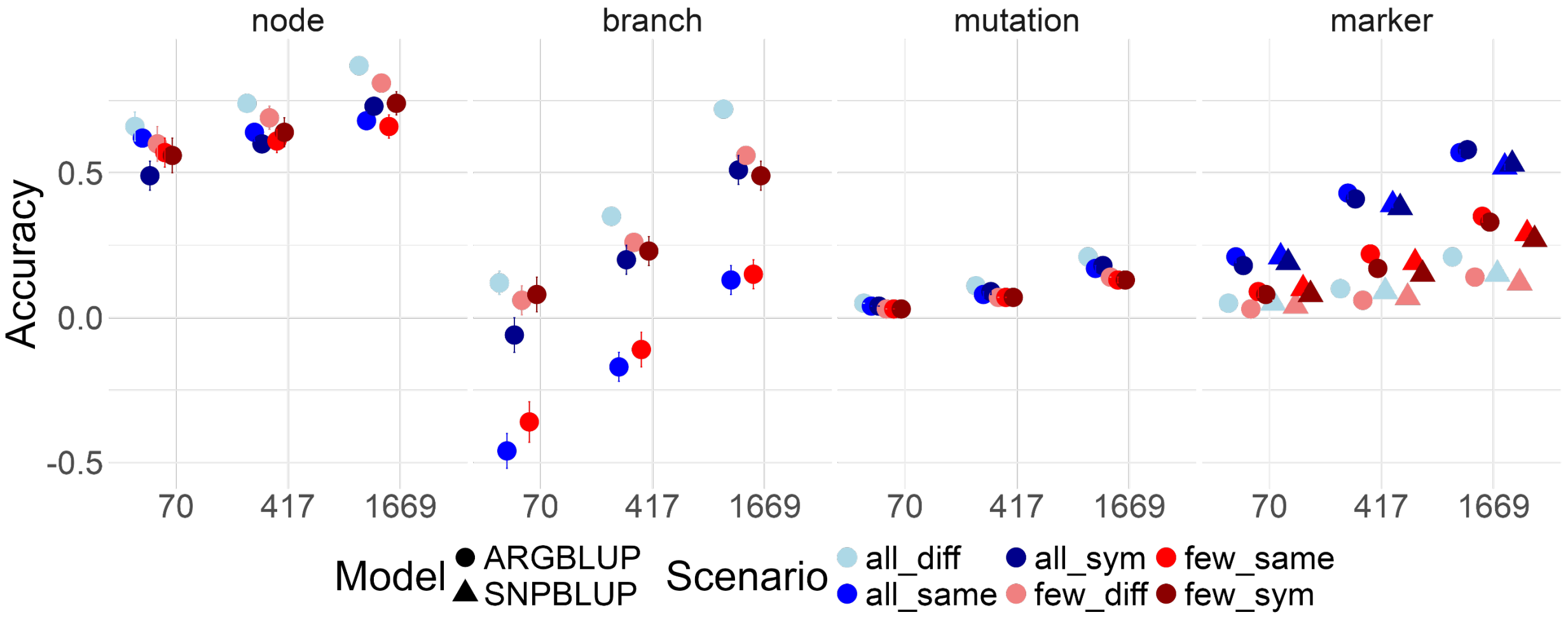
DIFF all mutation effects are different
SAME mutation type defines effect
SYM symmetric mutation effects

1669 phenotypes
417 phenotypes
70 phenotypes

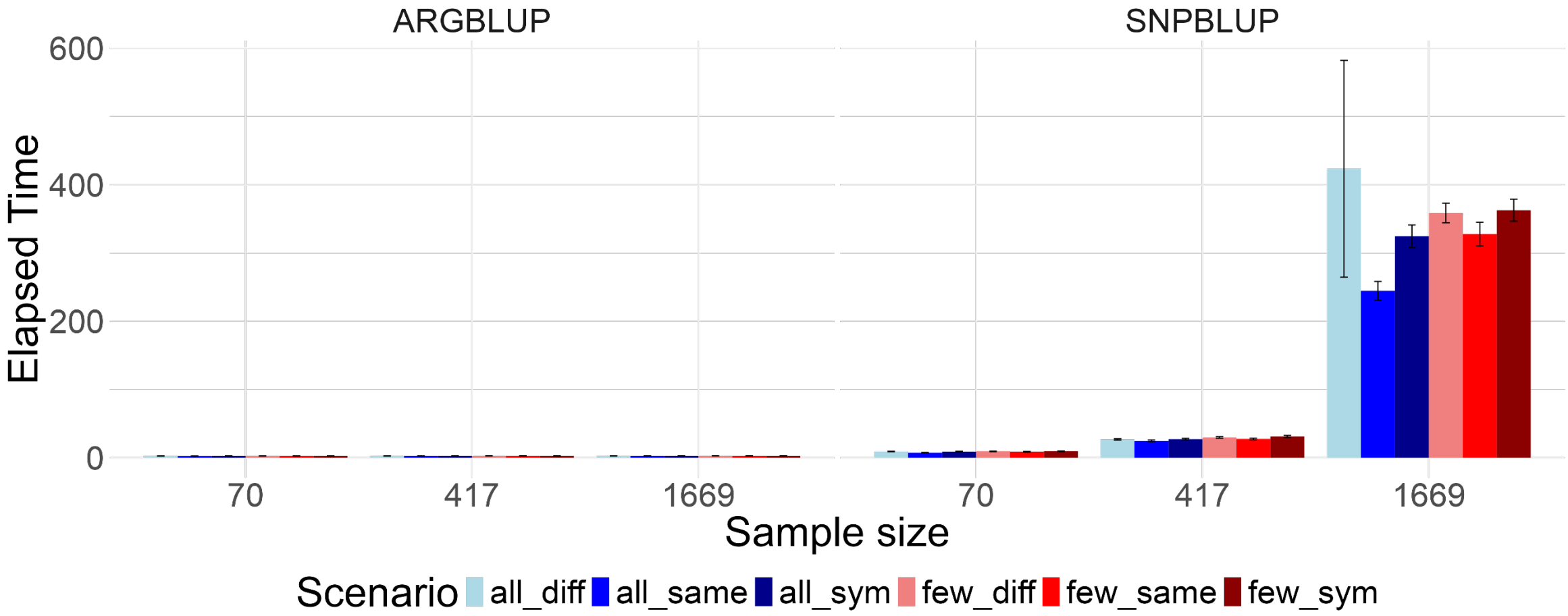
Accuracy of estimated haplotype effects (samples/tips)



Accuracy of various estimated effects



Execution time with INLA (seconds)



Conclusion

- Ancestral Recombination Graphs (ARGs) are the ultimate object in population genetics/genomics → now feasible & popular
- We are interested in quantitative genetic modelling with ARGs
- Here we presented a Tree-BLUP to work with a single non-recombining DNA region (mtDNA real data & simulation)
- QG modelling with the full ARG involves recombination events (haplotypes of different lengths) → we have a good working model for that in the pipeline

Estimating haplotype and mutation effects in the context of genome sequence via ancestral recombination graphs

Gabriela Mafra Fortuna, Jana Obsteter, Ajda Moskric, **Gregor Gorjanc**

2024 EAAP, Florence

@GregorGorjanc @HighlanderLab