Modelling complex traits with ancestral recombination graphs

Hanbin Lee

hblee@umich.edu

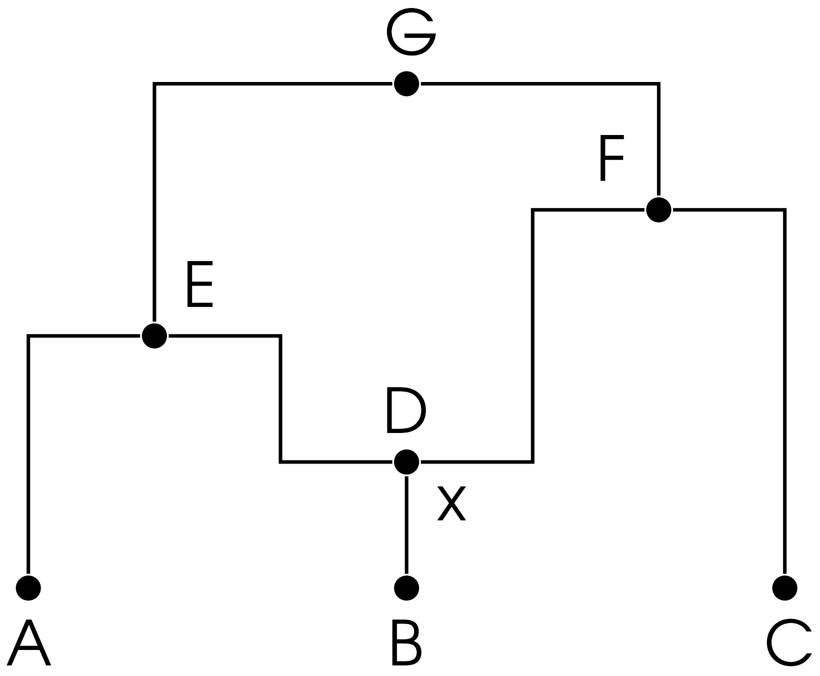
University of Michigan, Ann Arbor

Mar 7, 2025

Overview

Overview

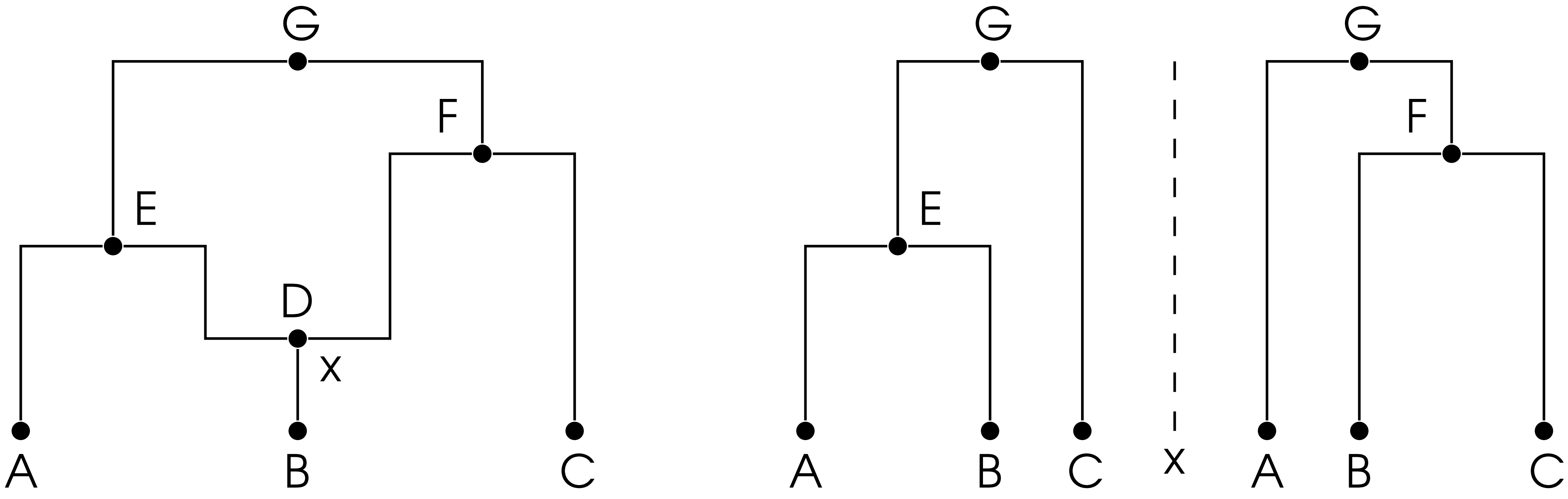
The ancestral recombination graph (ARG) describes the evolutionary relationship between genetic materials in the presence of recombination and drift



From (Wong et al. 2024)

Overview

The ancestral recombination graph (ARG) describes the evolutionary relationship between genetic materials in the presence of recombination and drift

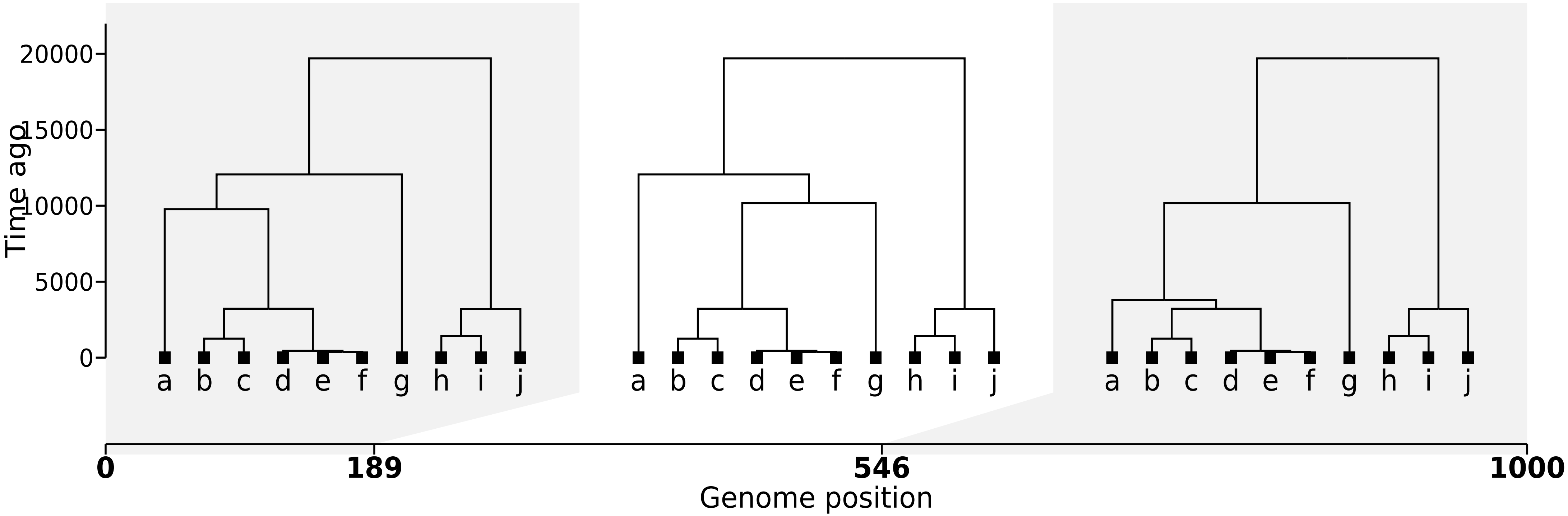


From (Wong et al. 2024)

Overview

The full probabilistic process is complicated

In this work, we condition on the realized ARG, resulting a sequence of local trees



From tskit docs

Overview

What is the conditional distribution of a trait given the trees?

Since the genealogy is fixed, the only randomness that remains is mutation

$$\text{Trait} \mid \text{Local trees} \quad \sim \quad ?$$

Linear mixed model

Linear mixed model

Linear mixed models are popular in quantitative genetics

$$\mathbf{y} = \underbrace{\mathbf{Z}\mathbf{u}}_{\text{random effects}} + \underbrace{\mathbf{X}\mathbf{b}}_{\text{fixed effects}} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ includes genotyped variants and $\mathbf{X}$ is the covariate matrix

In particular, the SNP effects $\mathbf{u} \sim p(\cdot)$ is *random*

Linear mixed model

Linear mixed models are popular in quantitative genetics

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{random effects}} + \underbrace{\mathbf{Xb}}_{\text{fixed effects}} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ includes genotyped variants and $\mathbf{X}$ is the covariate matrix

In particular, the SNP effects $\mathbf{u} \sim p(\cdot)$ is *random*

Some questions …

Linear mixed model

Linear mixed models are popular in quantitative genetics

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{random effects}} + \underbrace{\mathbf{Xb}}_{\text{fixed effects}} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ includes genotyped variants and $\mathbf{X}$ is the covariate matrix

In particular, the SNP effects $\mathbf{u} \sim p(\cdot)$ is *random*

Some questions …

- What's the source of $\mathbf{u}$'s randomness?

Linear mixed model

Linear mixed models are popular in quantitative genetics

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{random effects}} + \underbrace{\mathbf{Xb}}_{\text{fixed effects}} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ includes genotyped variants and $\mathbf{X}$ is the covariate matrix

In particular, the SNP effects $\mathbf{u} \sim p(\cdot)$ is *random*

Some questions …

- What's the source of $\mathbf{u}$'s randomness?
- Why are $\mathbf{u}$'s (vector of random effects) entries independent?

Linear mixed model

Linear mixed models are popular in quantitative genetics

$$\mathbf{y} = \underbrace{\mathbf{Z}\mathbf{u}}_{\text{random effects}} + \underbrace{\mathbf{X}\mathbf{b}}_{\text{fixed effects}} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ includes genotyped variants and $\mathbf{X}$ is the covariate matrix

In particular, the SNP effects $\mathbf{u} \sim p(\cdot)$ is *random*

Some questions …

- What's the source of $\mathbf{u}$'s randomness?
- Why are $\mathbf{u}$'s (vector of random effects) entries independent?

We answer these questions from a genealogical perspective

Setup and derivation

Setup and derivation

The trait $\mathbf{y}$ is a linear function of the genotype $\mathbf{G}$

$$\mathbf{y} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

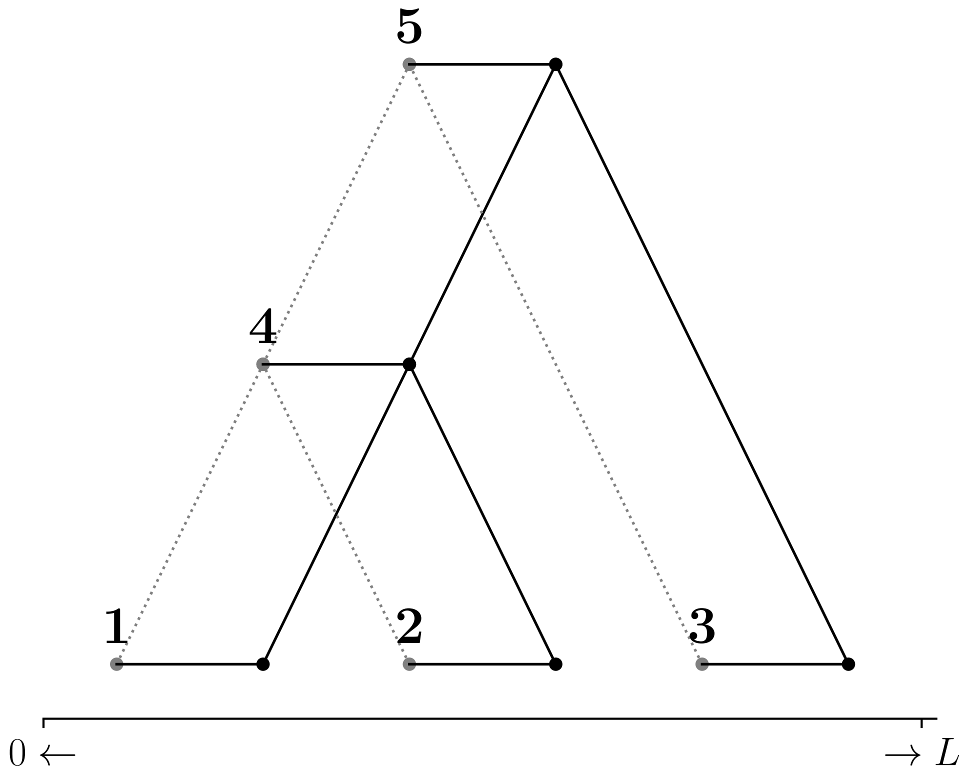
$\mathbf{y} \in \mathbb{R}^N, \mathbf{G} \in \mathbb{R}^{N \times P}, \boldsymbol{\beta} \in \mathbb{R}^P$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^N$

$\mathbf{G}$ contains *all* positions the genome including genotyped ones

N: number of samples, P: length of the genome

How do we get traits?

How do we get traits?

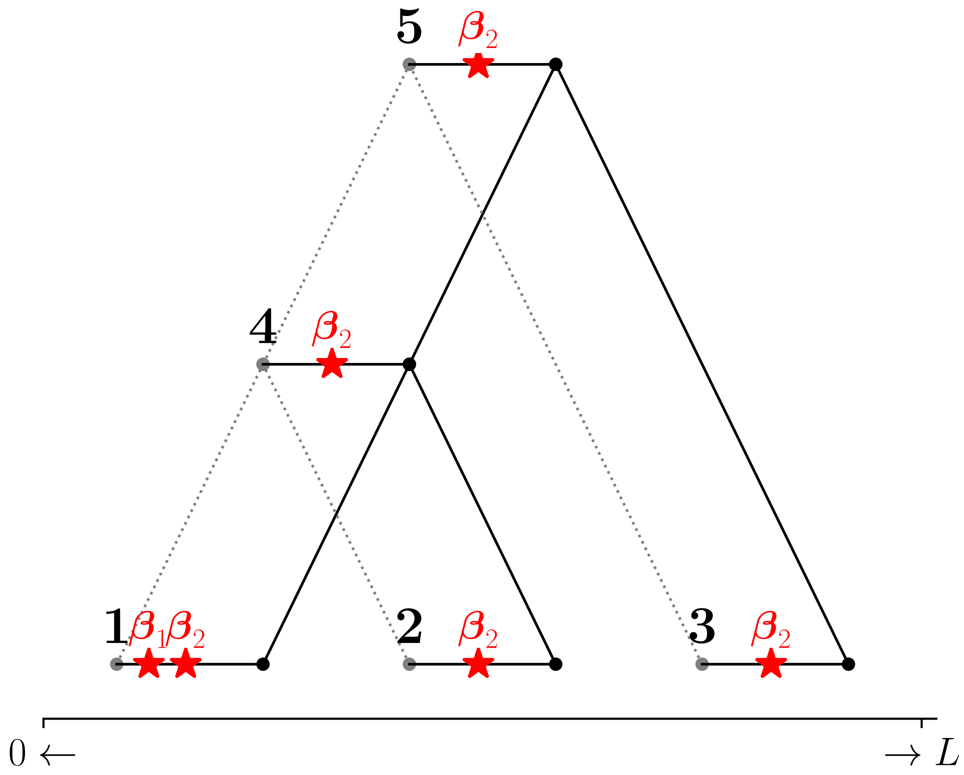


Consider a local tree that spans over a region

We get trait values by adding up effect sizes (β)

How do we get traits?

$$\mathbf{y}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2,\ \mathbf{y}_2 = \boldsymbol{\beta}_2,\ \mathbf{y}_3 = \boldsymbol{\beta}_2$$



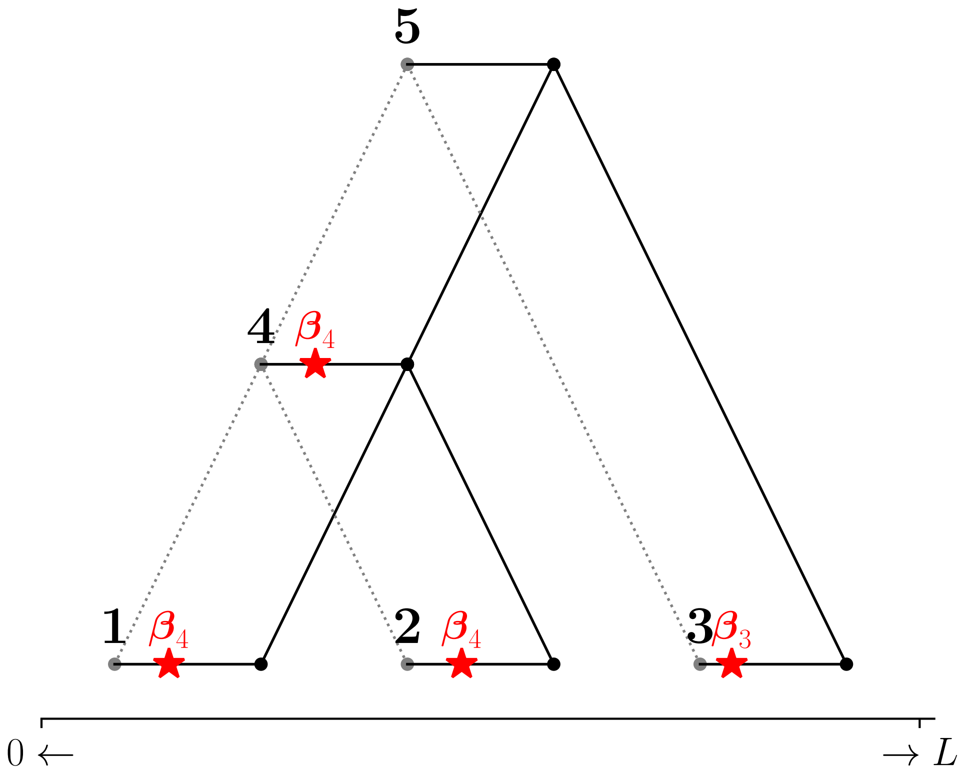
Consider a local tree that spans over a region

We get trait values by adding up effect sizes ($\boldsymbol{\beta}$)

- $\mathbf{y}_n = \mathbf{G}_{n1}\boldsymbol{\beta}_1 + \mathbf{G}_{n2}\boldsymbol{\beta}_2$

How do we get traits?

$$\mathbf{y}_1 = \boldsymbol{\beta}_4,\ \mathbf{y}_2 = \boldsymbol{\beta}_4,\ \mathbf{y}_3 = \boldsymbol{\beta}_3$$



Consider a local tree that spans over a region

We get trait values by adding up effect sizes ($\boldsymbol{\beta}$)

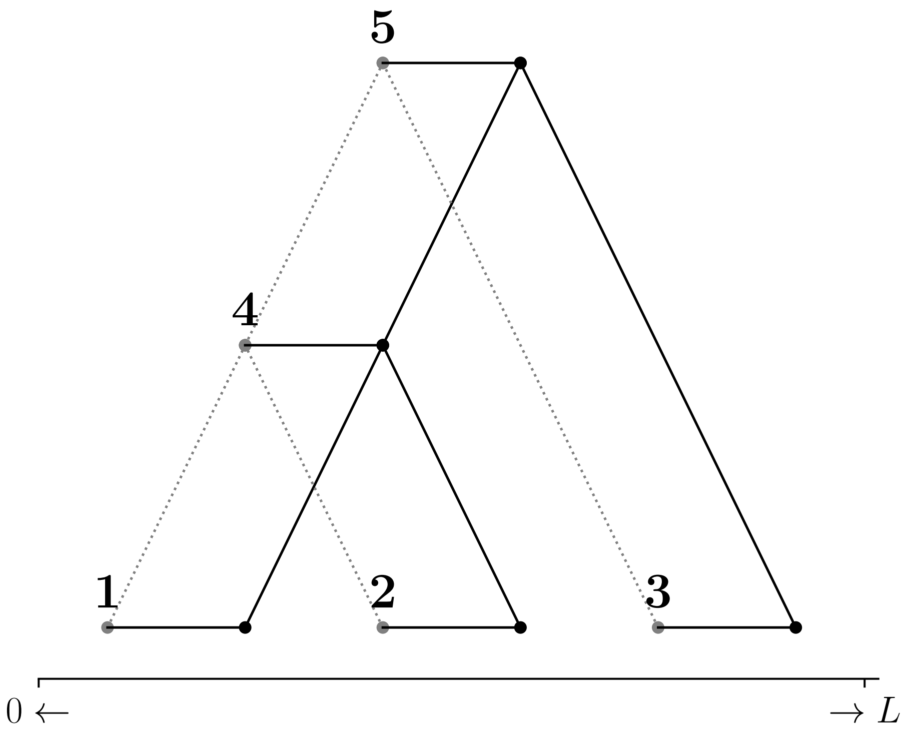
- $\mathbf{y}_n = \mathbf{G}_{n1}\boldsymbol{\beta}_1 + \mathbf{G}_{n2}\boldsymbol{\beta}_2$
- $\mathbf{y}_n = \mathbf{G}_{n3}\boldsymbol{\beta}_3 + \mathbf{G}_{n4}\boldsymbol{\beta}_4$

Branch-centric view of trait transmission

Branch-centric view of trait transmission

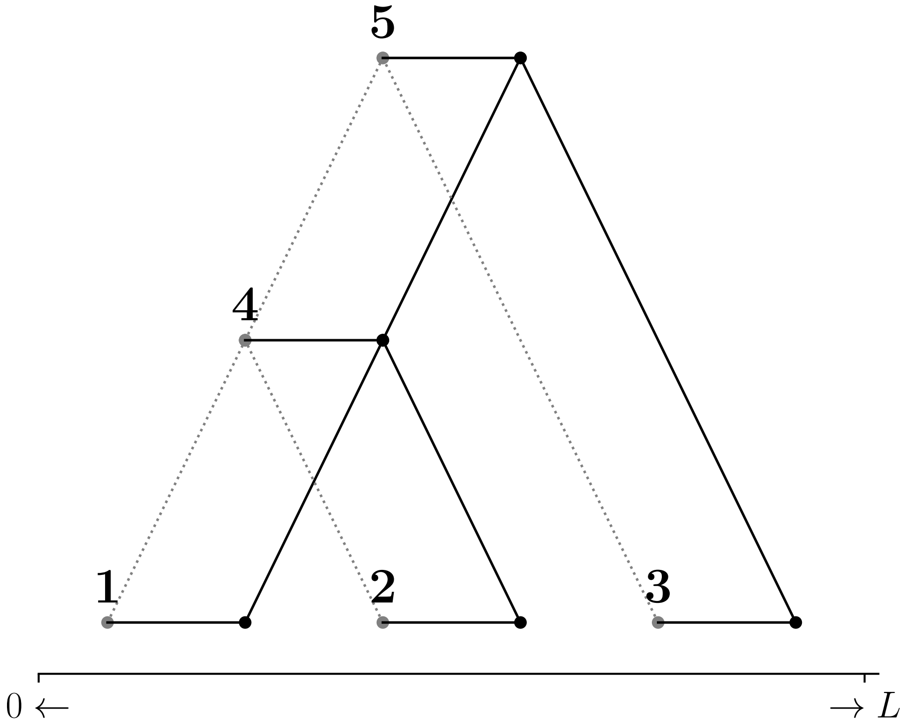
Inherit a branch first, then a mutation

Branch-centric view of trait transmission



Inherit a branch first, then a mutation

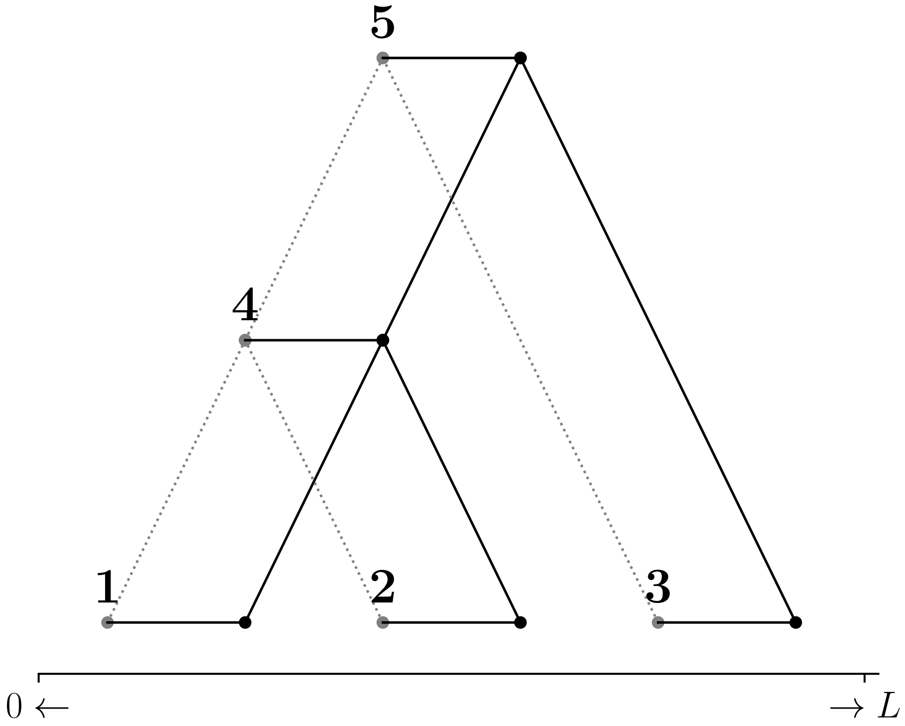
Branch-centric view of trait transmission



Inherit a branch first, then a mutation

- Sample 1 inherits edges $1 - 4$ and $4 - 5$

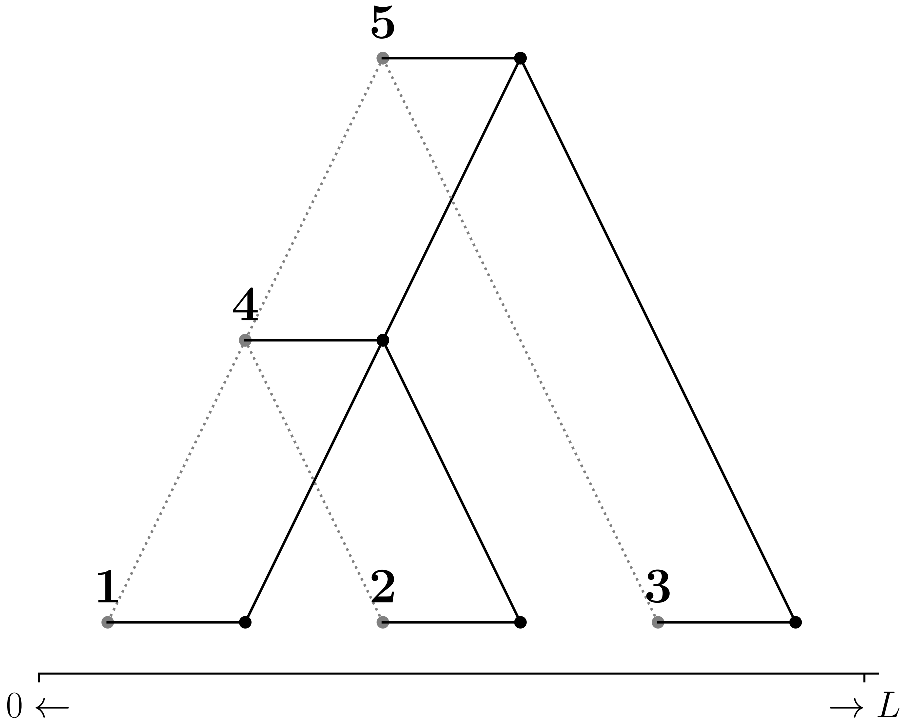
Branch-centric view of trait transmission



Inherit a branch first, then a mutation

- Sample 1 inherits edges $1 - 4$ and $4 - 5$
- Sample 2 inherits edges $2 - 4$ and $4 - 5$

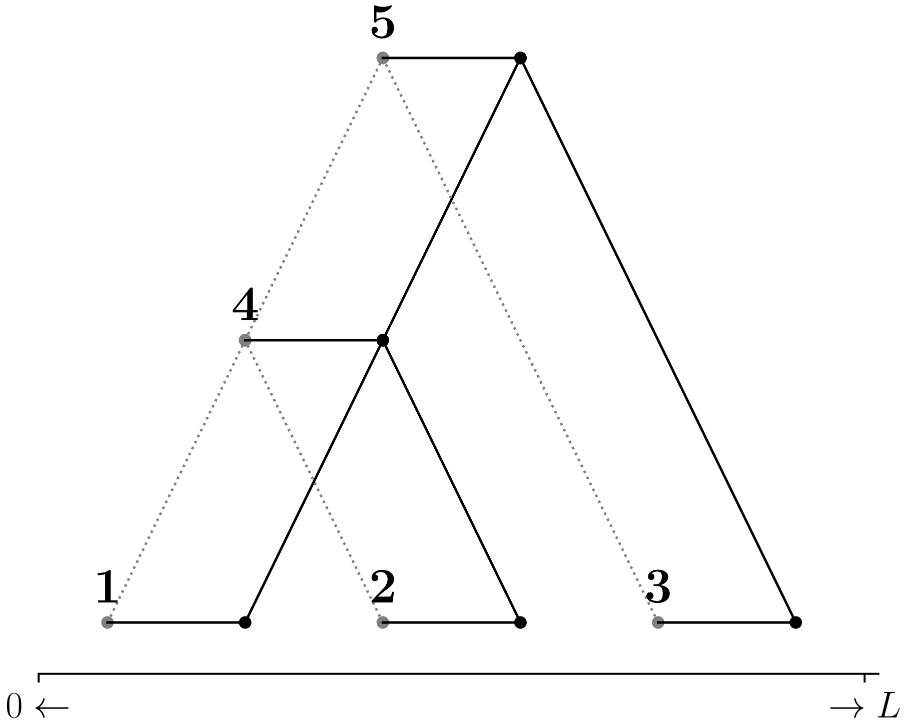
Branch-centric view of trait transmission



Inherit a branch first, then a mutation

- Sample 1 inherits edges $1 - 4$ and $4 - 5$
- Sample 2 inherits edges $2 - 4$ and $4 - 5$
- Sample 3 inherits edge $3 - 5$

Branch-centric view of trait transmission

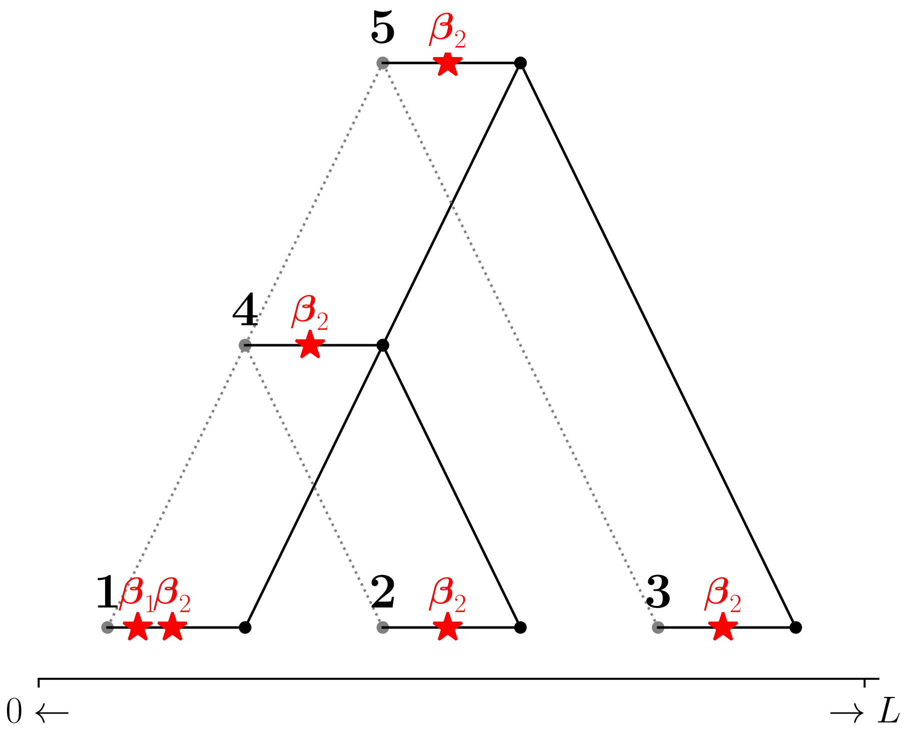


Inherit a branch first, then a mutation

- Sample 1 inherits edges $1 - 4$ and $4 - 5$
- Sample 2 inherits edges $2 - 4$ and $4 - 5$
- Sample 3 inherits edge $3 - 5$

Branch’s effect $=$ Sum of mutations’ effect

Branch-centric view of trait transmission



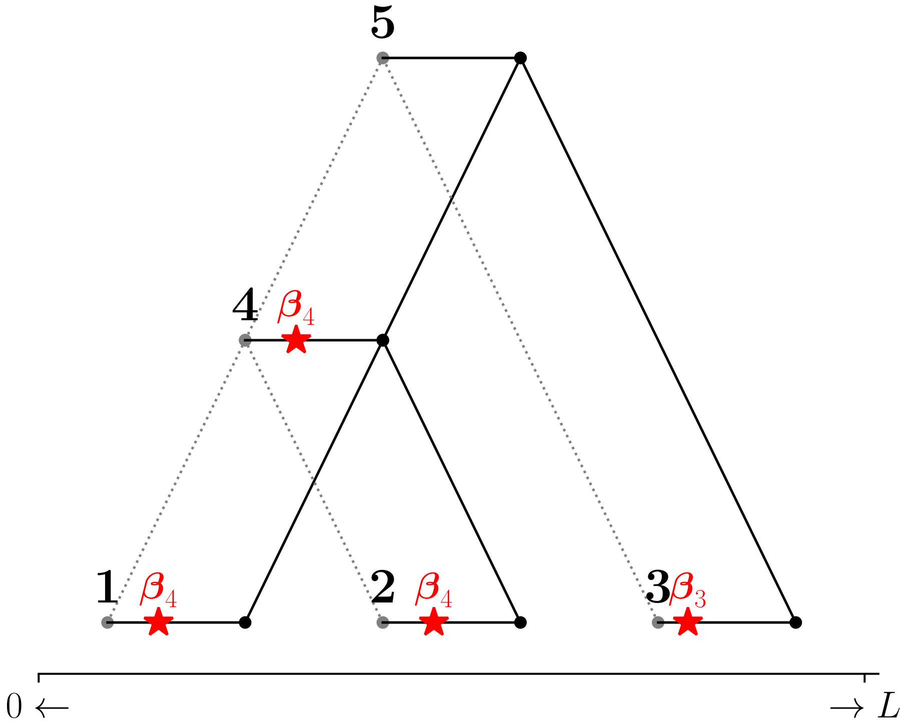
Inherit a branch first, then a mutation

- Sample 1 inherits edges $1 - 4$ and $4 - 5$
- Sample 2 inherits edges $2 - 4$ and $4 - 5$
- Sample 3 inherits edge $3 - 5$

Branch’s effect $=$ Sum of mutations’ effect

- Effect of $4 - 5 = 0$ (1st realization)

Branch-centric view of trait transmission



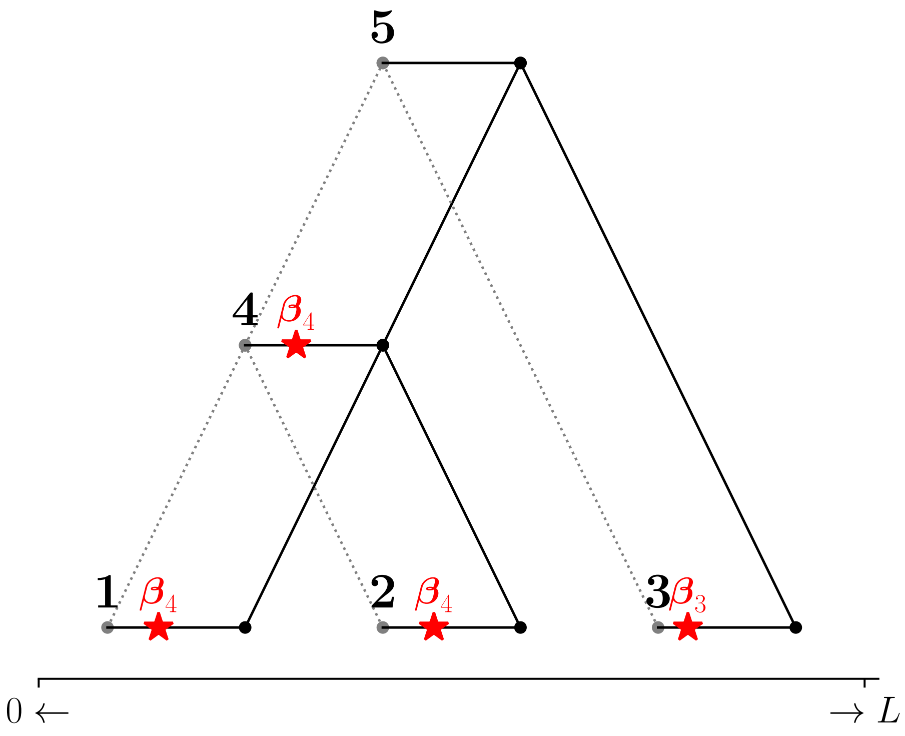
Inherit a branch first, then a mutation

- Sample 1 inherits edges $1-4$ and $4-5$
- Sample 2 inherits edges $2-4$ and $4-5$
- Sample 3 inherits edge $3-5$

Branch's effect $=$ Sum of mutations' effect

- Effect of $4-5=0$ (1st realization)
- Effect of $4-5=\beta_4$ (2nd realization)

Branch-centric view of trait transmission



Inherit a branch first, then a mutation

- Sample 1 inherits edges $1-4$ and $4-5$
- Sample 2 inherits edges $2-4$ and $4-5$
- Sample 3 inherits edge $3-5$

Branch's effect $=$ Sum of mutations' effect

- Effect of $4-5=0$ (1st realization)
- Effect of $4-5=\beta_4$ (2nd realization)

Branch effect is a random variable!

From variants to branches

From variants to branches

$$\text{Trait} = \sum_p \text{Variant}_p \text{ effect size} \quad \Rightarrow \quad \text{Trait} = \sum_e \text{Branch}_e \text{ effect size}$$

From variants to branches

$$\text{Trait} = \sum_{p} \text{Variant}_p \text{ effect size} \quad \Rightarrow \quad \text{Trait} = \sum_{e} \text{Branch}_e \text{ effect size}$$

$$\boldsymbol{v}_e = \text{Branch}_e \text{ effect size} = \sum_{p} \text{Variant}_p \text{ on Branch}_e$$

From variants to branches

$$\text{Trait} = \sum_p \text{Variant}_p \text{ effect size} \quad \Rightarrow \quad \text{Trait} = \sum_e \text{Branch}_e \text{ effect size}$$

$$\boldsymbol{\upsilon}_e = \text{Branch}_e \text{ effect size} = \sum_p \text{Variant}_p \text{ on Branch}_e$$

$$\mathbf{y} = \sum_p \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathbf{y} = \sum_e \mathbf{Z}_e \boldsymbol{\upsilon}_e + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}_{ne} =$ the number of haplotypes of n that inherit e

Ancestral recombination graph linear mixed model (ARG-LMM)

Ancestral recombination graph linear mixed model (ARG-LMM)

Split $\boldsymbol{v}$ to $\mathbf{u} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v}$ and $\mathbf{f} = \mathrm{E}\boldsymbol{v}$

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{Random effects}} + \underbrace{\mathbf{Zf}}_{\text{Fixed effects}} + \boldsymbol{\varepsilon}$$

Ancestral recombination graph linear mixed model (ARG-LMM)

Split $\boldsymbol{v}$ to $\mathbf{u} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v}$ and $\mathbf{f} = \mathrm{E}\boldsymbol{v}$

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{Random effects}} + \underbrace{\mathbf{Zf}}_{\text{Fixed effects}} + \boldsymbol{\varepsilon}$$

This is the ancestral recombination graph linear mixed model (ARG-LMM) and $\mathbf{Z}\mathrm{Cov}(\mathbf{u})\mathbf{Z}^T$ is the expected genetic relatedness matrix (eGRM) (Fan, Mancuso, and Chiang 2022; Zhang et al. 2023)

Ancestral recombination graph linear mixed model (ARG-LMM)

Split $\boldsymbol{v}$ to $\mathbf{u} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v}$ and $\mathbf{f} = \mathrm{E}\boldsymbol{v}$

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{Random effects}} + \underbrace{\mathbf{Zf}}_{\text{Fixed effects}} + \boldsymbol{\varepsilon}$$

This is the ancestral recombination graph linear mixed model (ARG-LMM) and $\mathbf{Z}\mathrm{Cov}(\mathbf{u})\mathbf{Z}^T$ is the expected genetic relatedness matrix (eGRM) (Fan, Mancuso, and Chiang 2022; Zhang et al. 2023)

- The random effects are tied to a physical process - Mutations!

Ancestral recombination graph linear mixed model (ARG-LMM)

Split $\boldsymbol{v}$ to $\mathbf{u} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v}$ and $\mathbf{f} = \mathrm{E}\boldsymbol{v}$

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{Random effects}} + \underbrace{\mathbf{Zf}}_{\text{Fixed effects}} + \boldsymbol{\varepsilon}$$

This is the ancestral recombination graph linear mixed model (ARG-LMM) and $\mathbf{Z}\mathrm{Cov}(\mathbf{u})\mathbf{Z}^T$ is the expected genetic relatedness matrix (eGRM) (Fan, Mancuso, and Chiang 2022; Zhang et al. 2023)

- The random effects are tied to a physical process - Mutations!
- We start from more lower-level evolutionary statements to recover mixed model assumptions

Ancestral recombination graph linear mixed model (ARG-LMM)

Split $\boldsymbol{v}$ to $\mathbf{u} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v}$ and $\mathbf{f} = \mathrm{E}\boldsymbol{v}$

$$\mathbf{y} = \underbrace{\mathbf{Zu}}_{\text{Random effects}} + \underbrace{\mathbf{Zf}}_{\text{Fixed effects}} + \boldsymbol{\varepsilon}$$

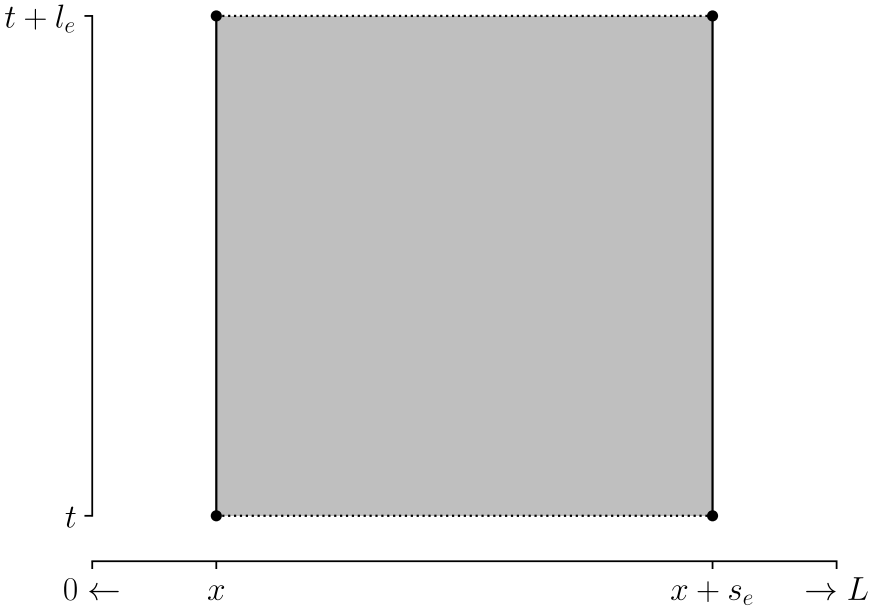
This is the ancestral recombination graph linear mixed model (ARG-LMM) and $\mathbf{Z}\mathrm{Cov}(\mathbf{u})\mathbf{Z}^T$ is the expected genetic relatedness matrix (eGRM) (Fan, Mancuso, and Chiang 2022; Zhang et al. 2023)

- The random effects are tied to a physical process - Mutations!
- We start from more lower-level evolutionary statements to recover mixed model assumptions
- Independent random effects, random effect weights, normality, . . .

How do we weigh branches of the ARG?

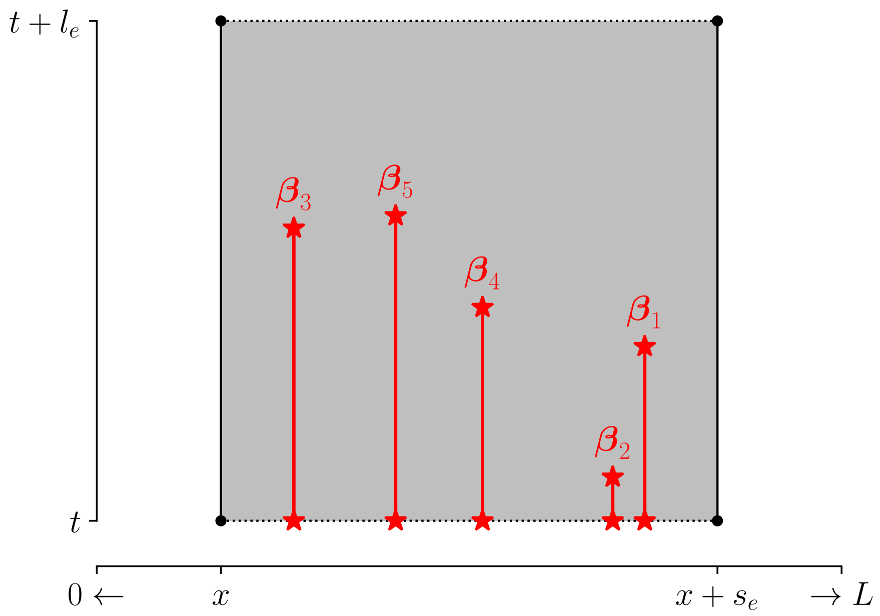
How do we weigh branches of the ARG?

l_e : length in time $\quad$ s_e : span in base pairs



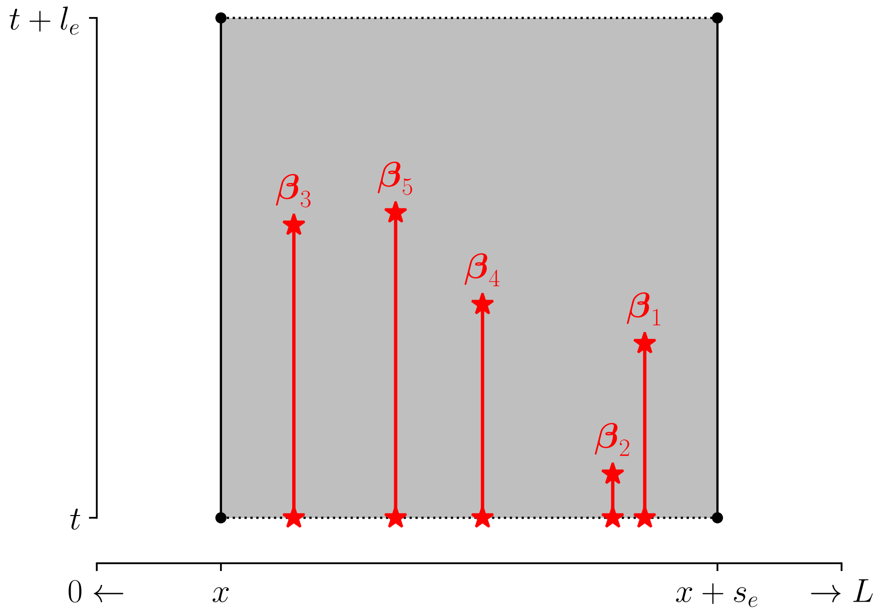
How do we weigh branches of the ARG?

l_e : length in time s_e : span in base pairs



How do we weigh branches of the ARG?

l_e : length in time $\quad s_e$: span in base pairs



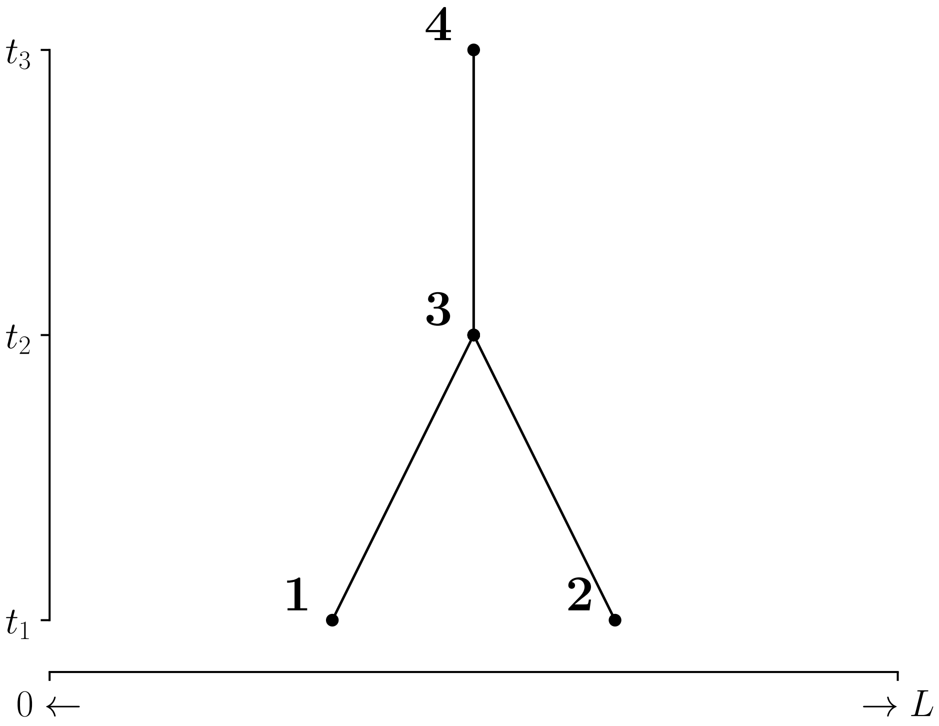
$$\mathrm{Var}(\mathbf{u}_e) \quad \propto \quad \text{Number of mutations} \quad \propto \quad \text{Area} = l_e s_e$$

Complex traits through the lens of ARG-LMM

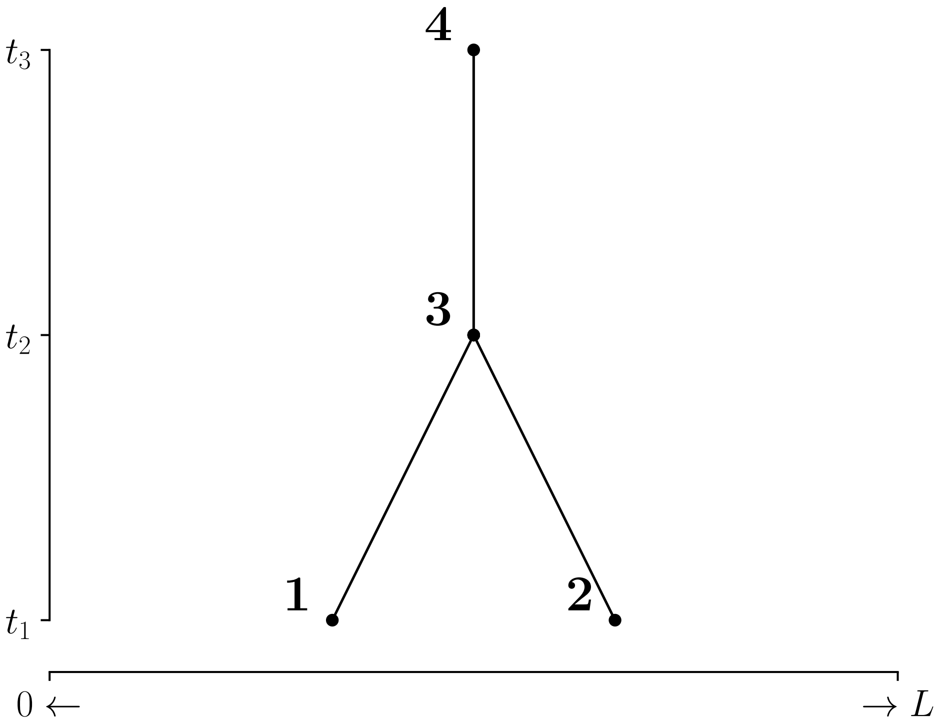
What does ARG-LMM tell us about complex trait analysis?

Genetic value covariance

Genetic value covariance

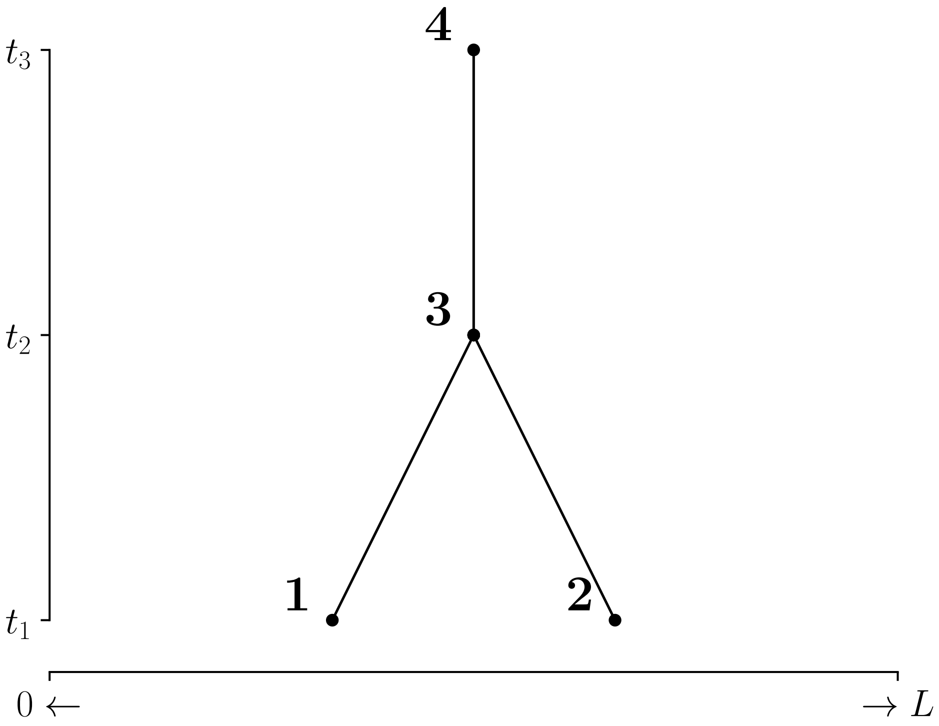


Genetic value covariance



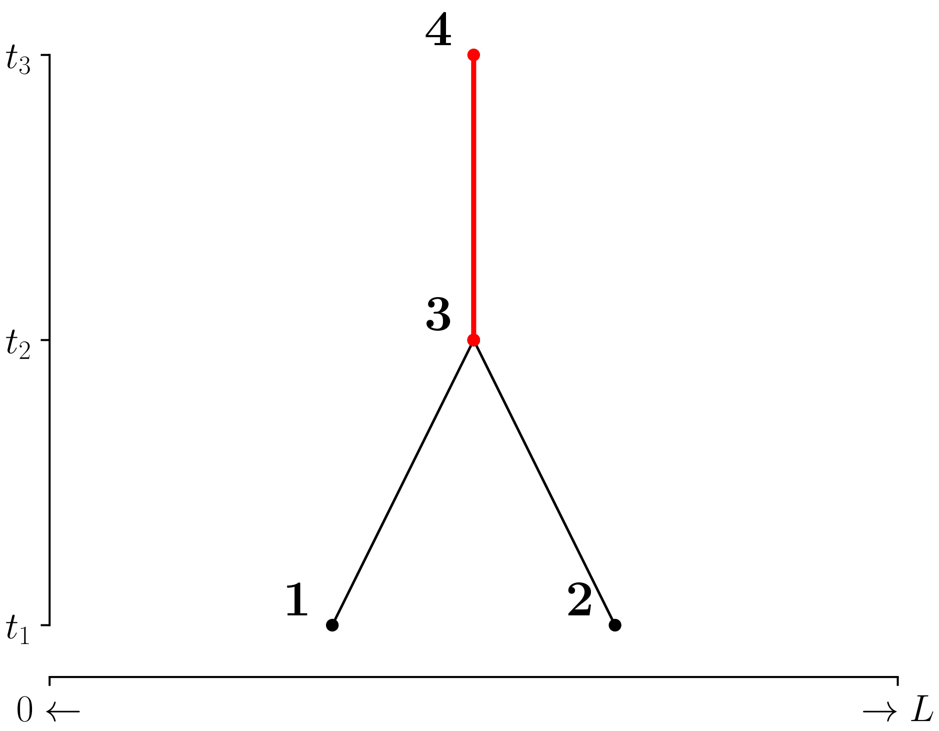
$$\mathbf{y}_1 = \mathbf{u}_{13} + \mathbf{u}_{34} \quad \text{and} \quad \mathbf{y}_2 = \mathbf{u}_{23} + \mathbf{u}_{34}$$

Genetic value covariance



$$\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathrm{Cov}(\mathbf{u}_{13} + \mathbf{u}_{34}, \mathbf{u}_{23} + \mathbf{u}_{34}) = \mathrm{Cov}(\mathbf{u}_{34}, \mathbf{u}_{34}) = \mathrm{Var}(\mathbf{u}_{34}) \propto t_3 - t_2$$

Genetic value covariance



$$\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathrm{Cov}(\mathbf{u}_{13} + \mathbf{u}_{34}, \mathbf{u}_{23} + \mathbf{u}_{34}) = \mathrm{Cov}(\mathbf{u}_{34}, \mathbf{u}_{34}) = \mathrm{Var}(\mathbf{u}_{34}) \propto t_3 - t_2$$

Heritability is *ill*-defined in ARG-LMM

Heritability is *ill*-defined in ARG-LMM

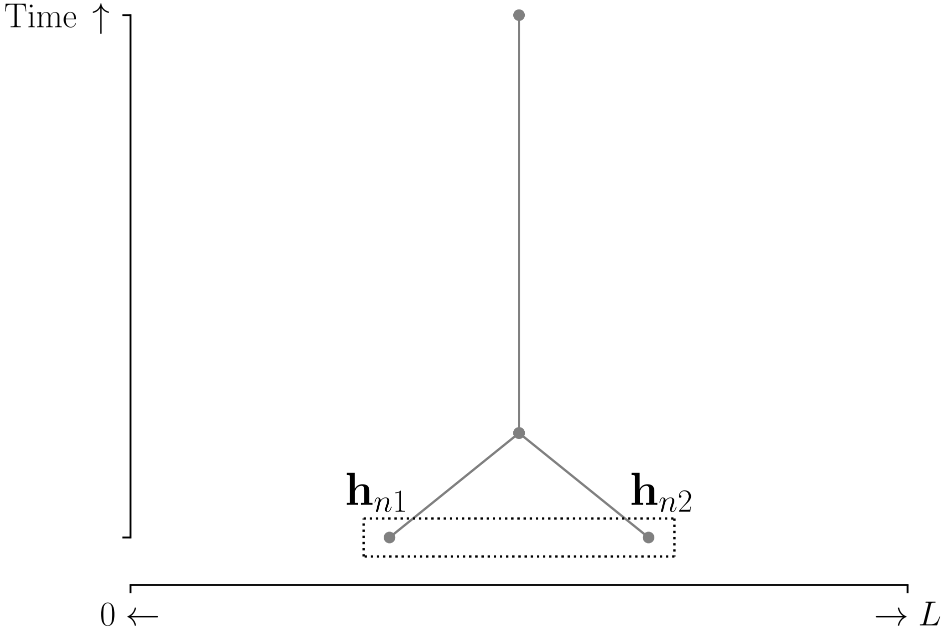
$$\text{Heritability: } h_g^2 = \frac{\mathrm{Var}(\mathbf{g}_n)}{\mathrm{Var}(\mathbf{y}_n)} = \frac{\mathrm{Var}(\mathbf{g}_n)}{\mathrm{Var}(\mathbf{g}_n) + \mathrm{Var}(\boldsymbol{\varepsilon}_n)}$$

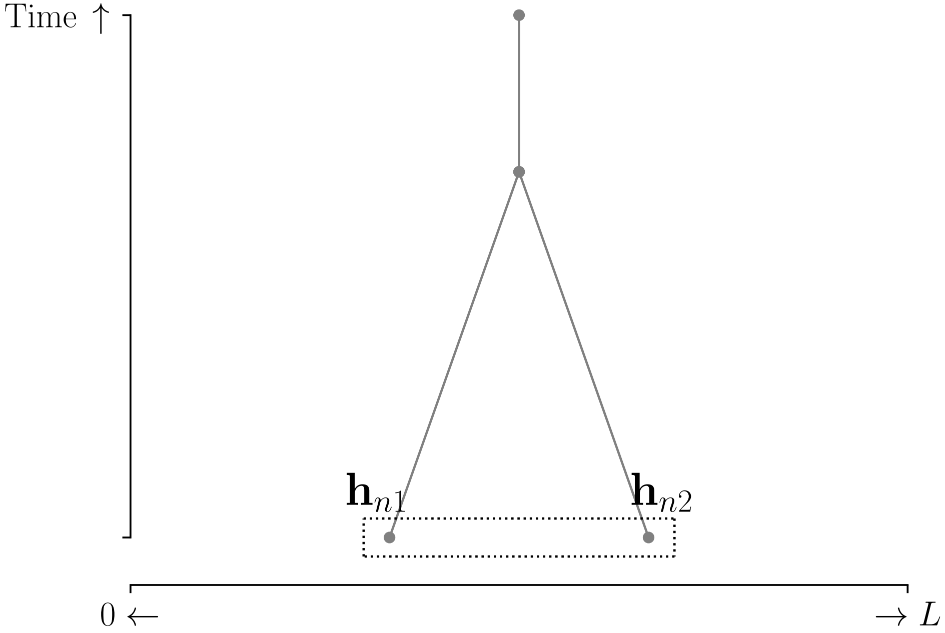
This applies to all individuals $n \in \{1, \ldots, N\}$

Heritability is *ill*-defined in ARG-LMM

However, all individuals have a different amount of genetic variance (except haploids)

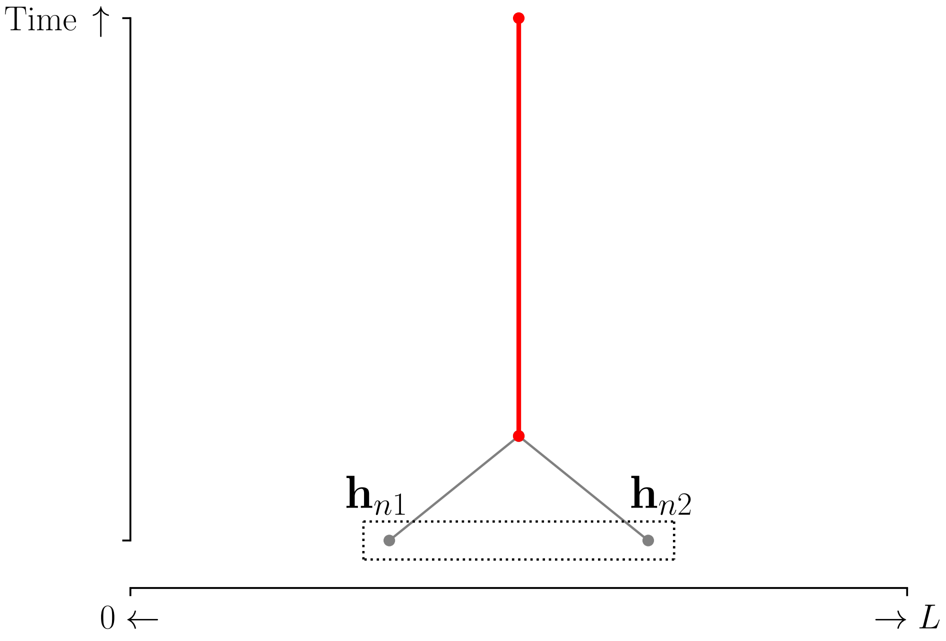
$$\mathrm{Var}(\mathbf{g}_n) = \mathrm{Var}(\mathbf{h}_{n1} + \mathbf{h}_{n2}) = \mathrm{Var}(\mathbf{h}_{n1}) + \mathrm{Var}(\mathbf{h}_{n2}) + 2\mathrm{Cov}(\mathbf{h}_{n1}, \mathbf{h}_{n2})$$

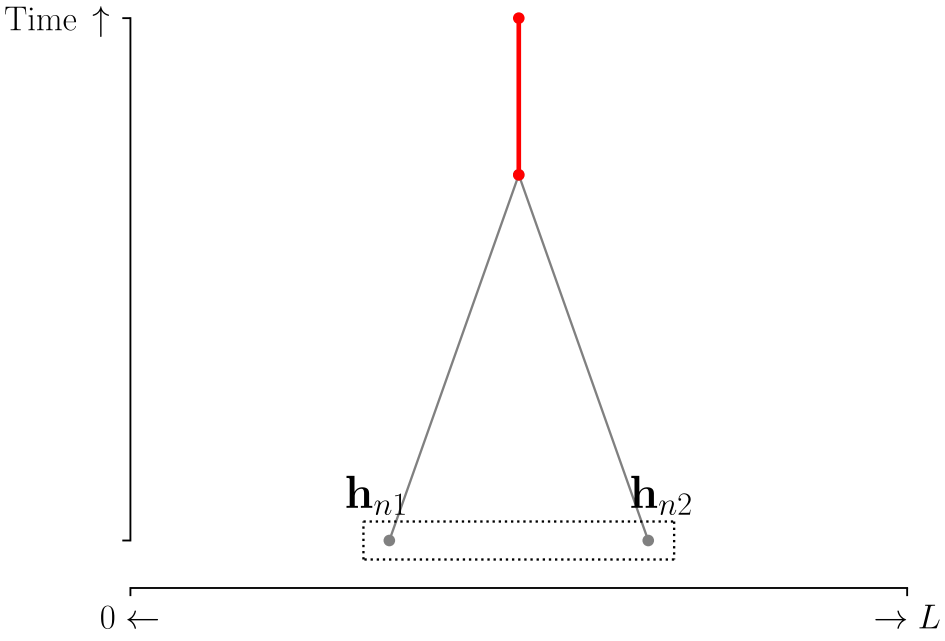




Heritability is *ill*-defined in ARG-LMM

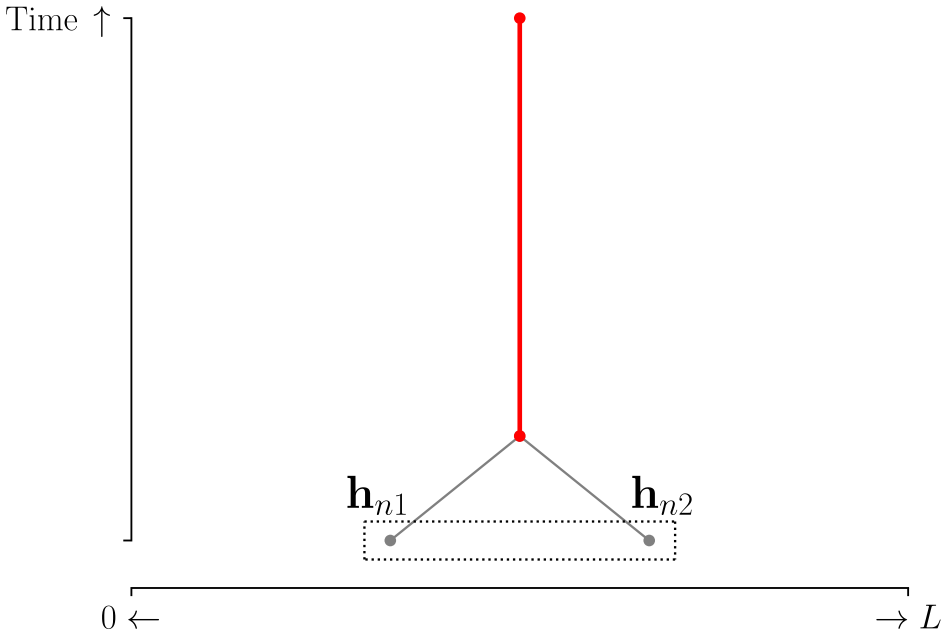
$$\mathrm{Var}(\mathbf{g}_n) = \mathrm{Var}(\mathbf{h}_{n1} + \mathbf{h}_{n2}) = \mathrm{Var}(\mathbf{h}_{n1}) + \mathrm{Var}(\mathbf{h}_{n2}) + 2\underbrace{\mathrm{Cov}(\mathbf{h}_{n1}, \mathbf{h}_{n2})}_{\text{Self-relatedness}}$$

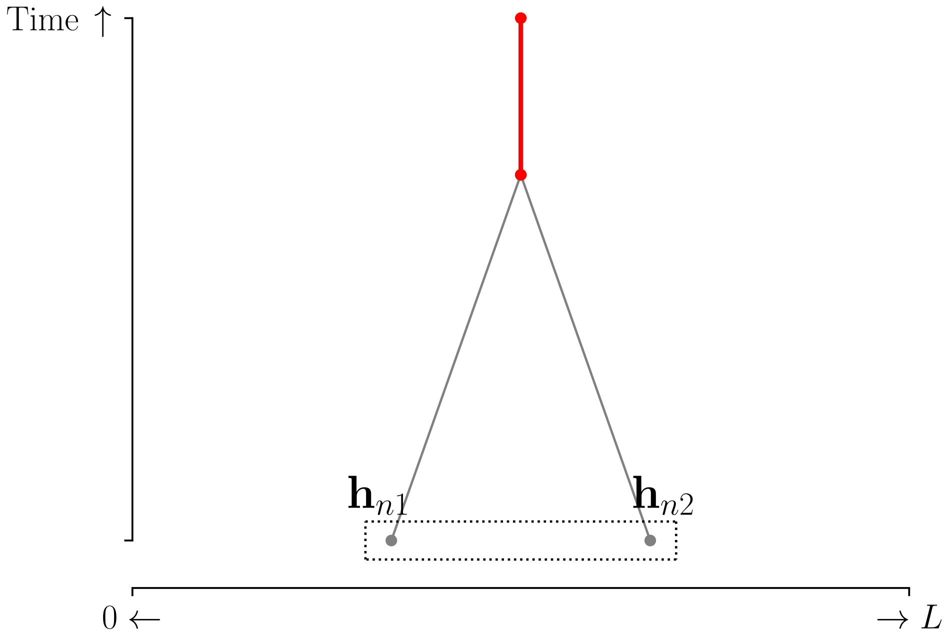




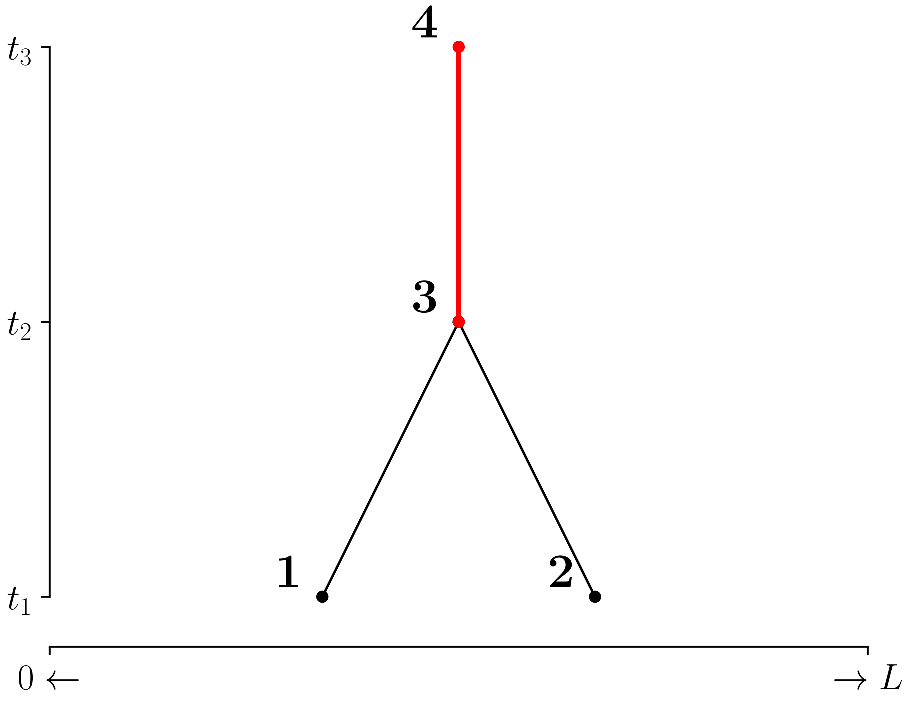
Heritability is *ill*-defined in ARG-LMM

We can’t define a single quantitity $h_g^2 = \frac{\mathrm{Var}(\mathbf{g}_n)}{\mathrm{Var}(\mathbf{g}_n)+\mathrm{Var}(\boldsymbol{\varepsilon}_n)}$ for everyone



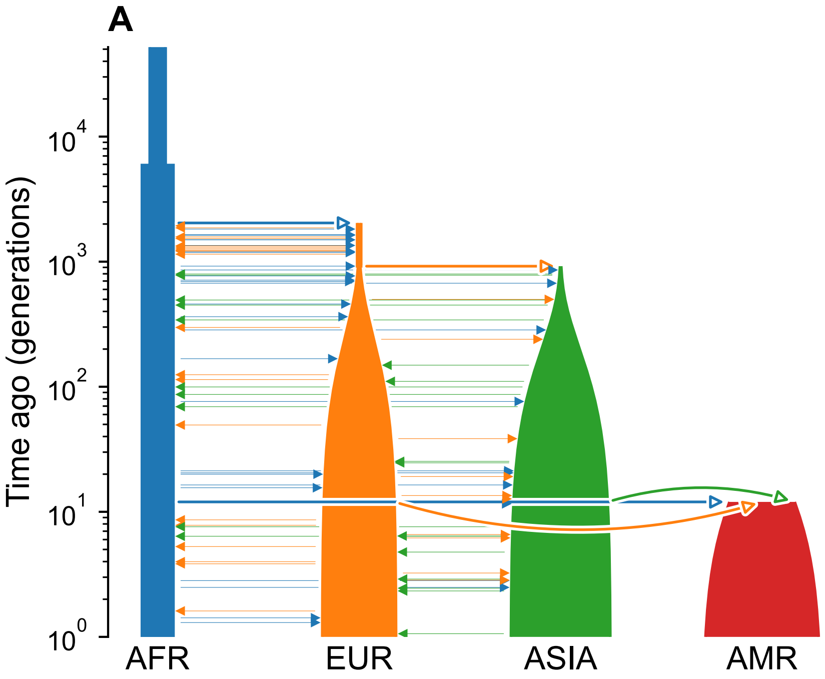


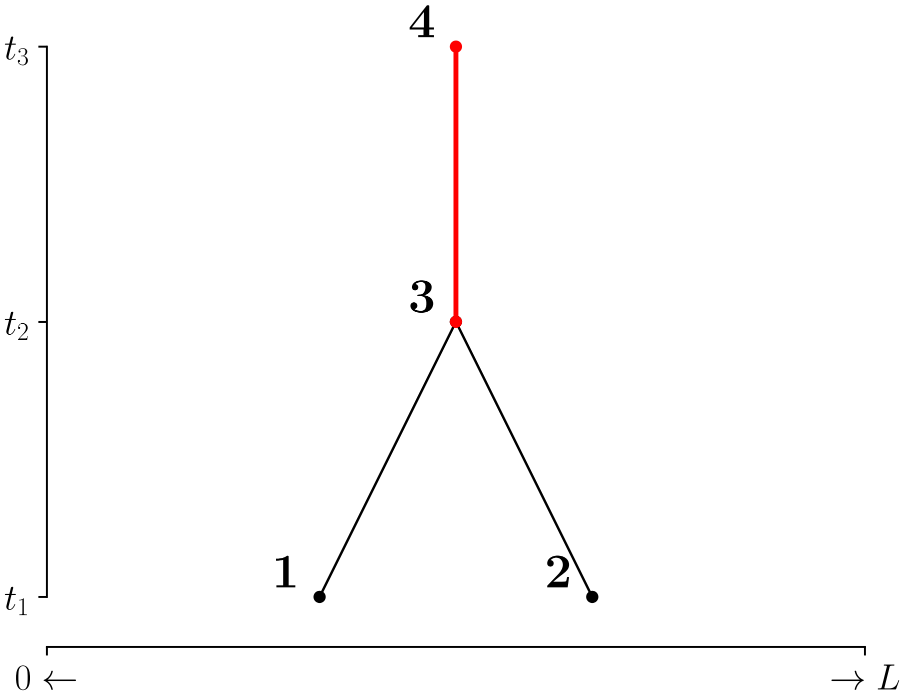
Polygenic prediction is constrained by demography



Demographic model from (Browning et al. 2018)

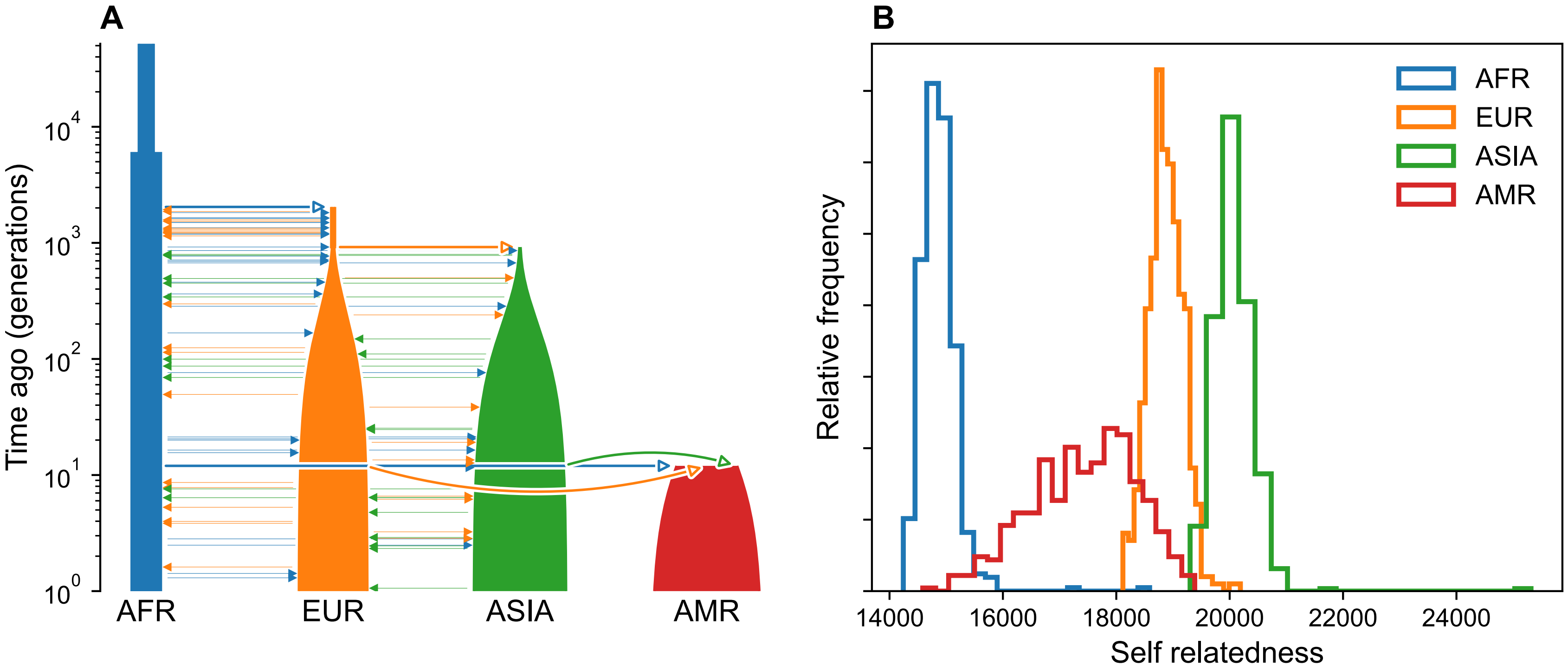
Polygenic prediction is constrained by demography





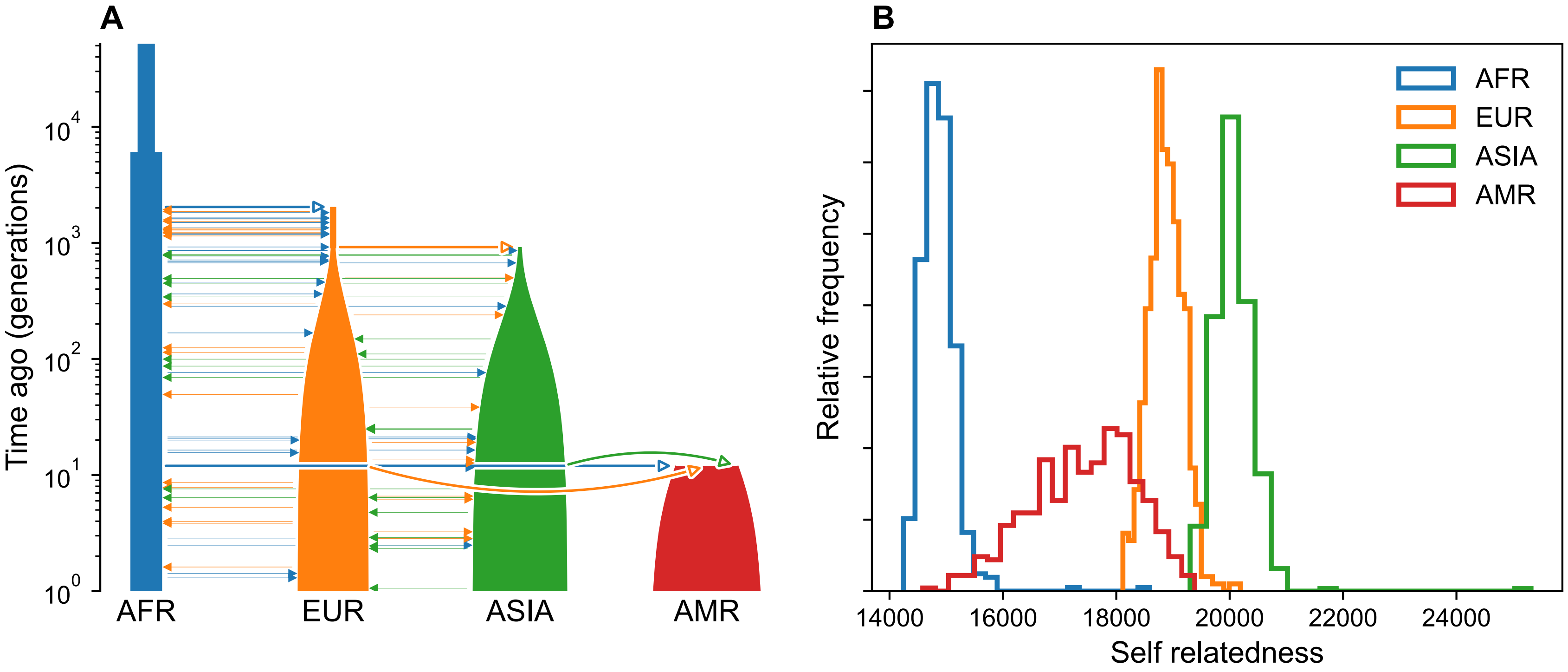
Demographic model from (Browning et al. 2018)

Polygenic prediction is constrained by demography



Demographic model from (Browning et al. 2018)

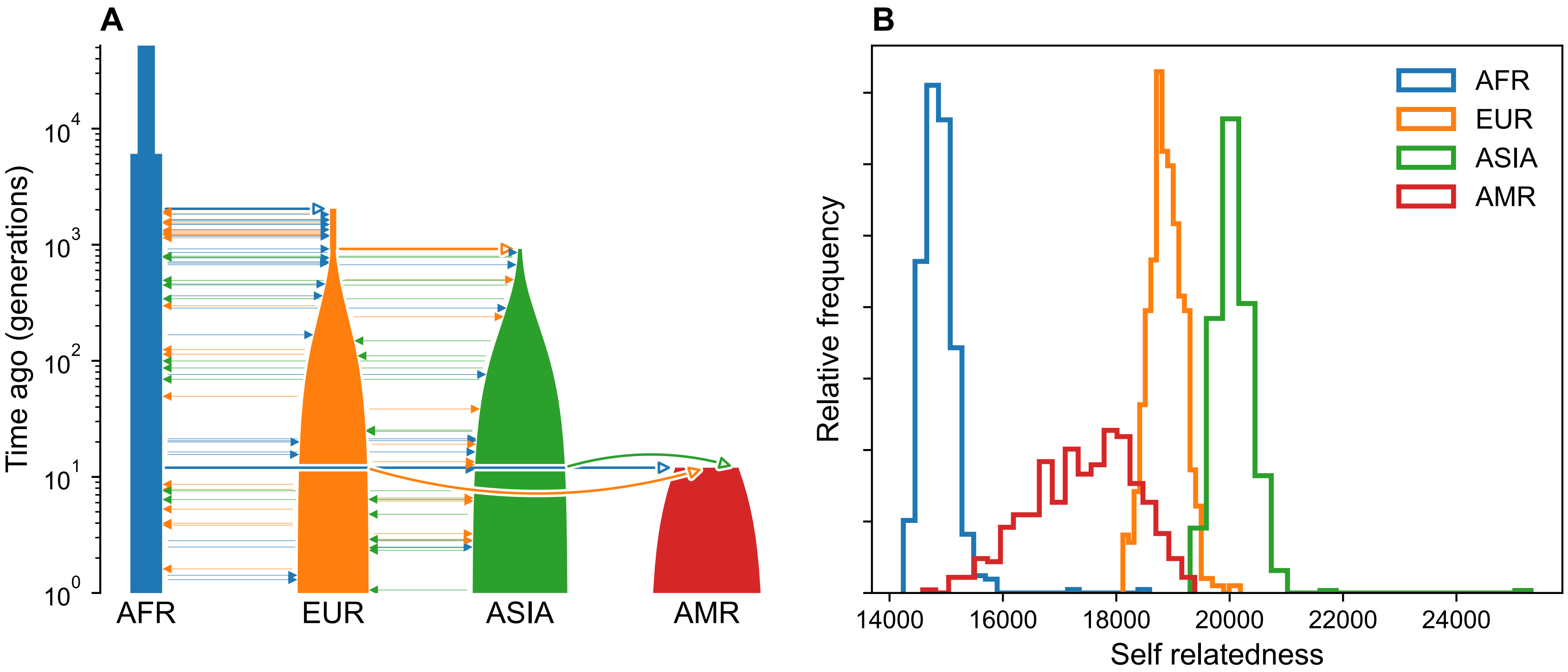
Polygenic prediction is constrained by demography



Some people are less genetically variable than others

Demographic model from (Browning et al. 2018)

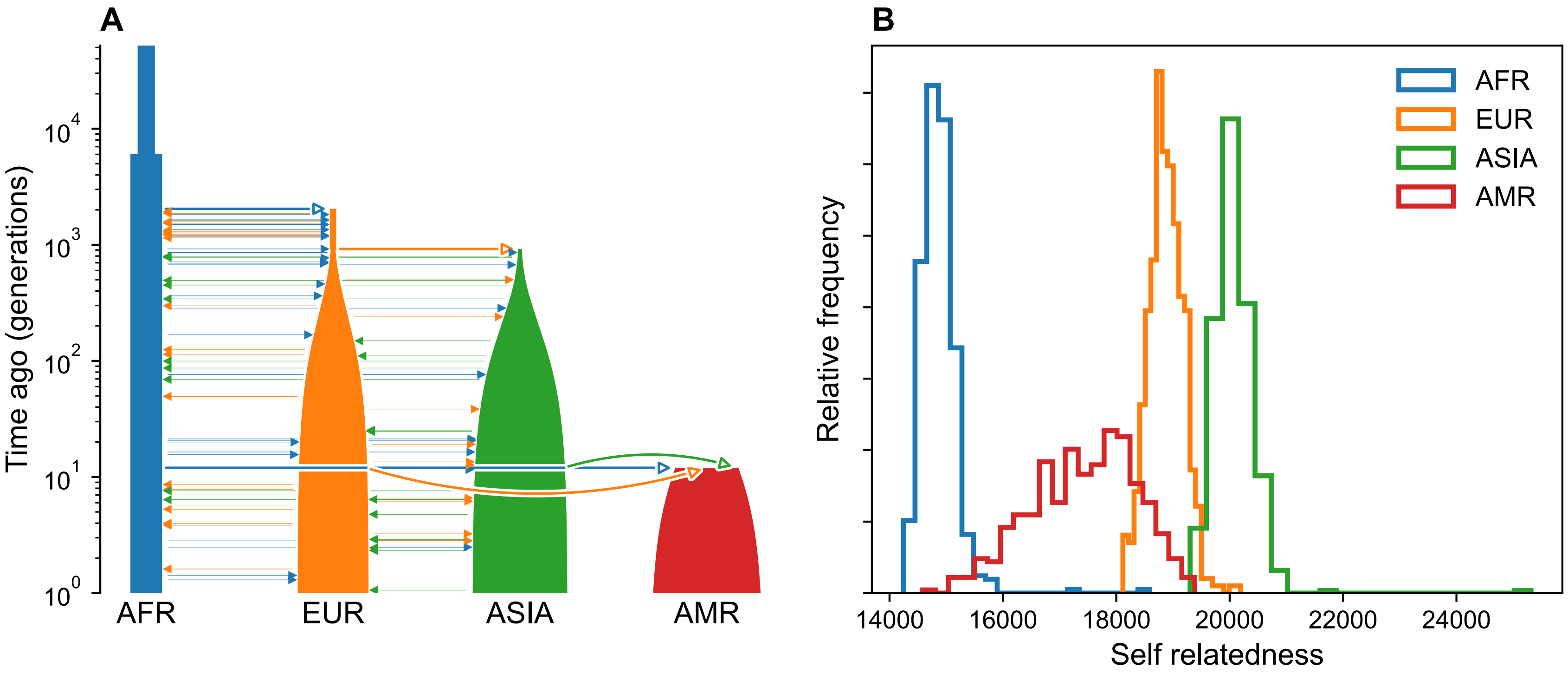
Polygenic prediction is constrained by demography



Some people are harder to predict genetically than others

Demographic model from (Browning et al. 2018)

Polygenic prediction is constrained by demography



Some populations are inherently harder to predict!

Demographic model from (Browning et al. 2018)

tslmm, fitting ARG-LMM to tree sequences

tslmm, fitting ARG-LMM to tree sequences

tslmm utilizes an efficient *genetic relatedness matrix - vector product* to fit the restricted maximum likelihood (REML) objective

tslmm, fitting ARG-LMM to tree sequences

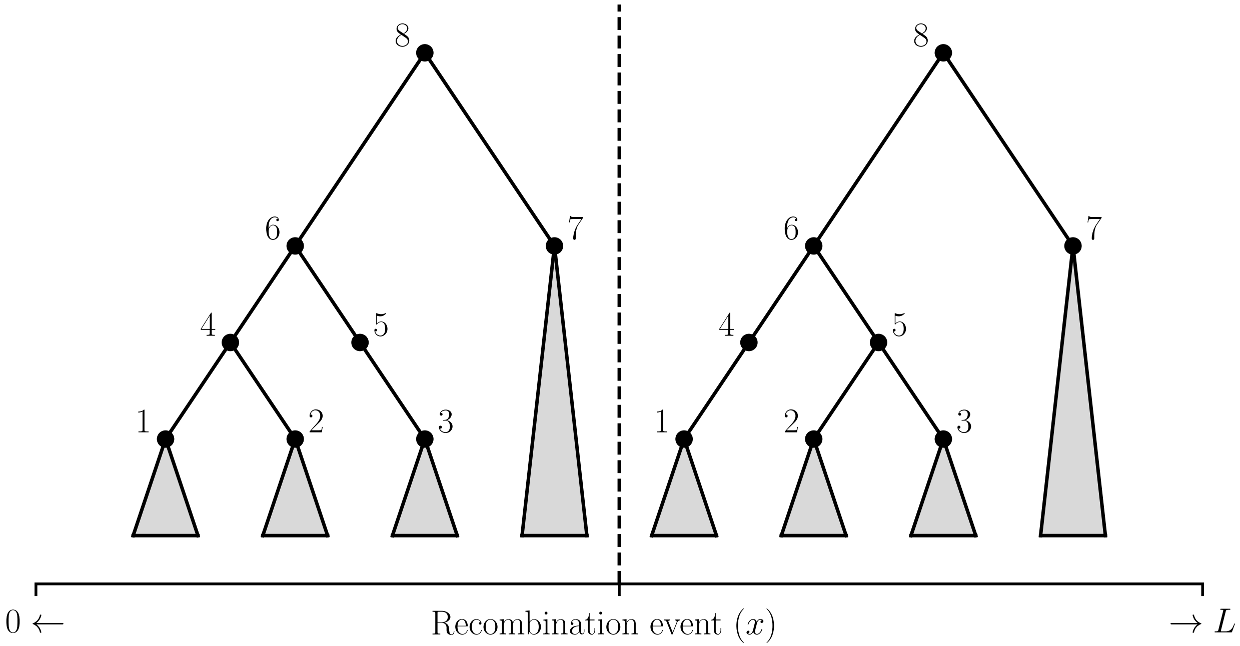
tslmm utilizes an efficient *genetic relatedness matrix - vector product* to fit the restricted maximum likelihood (REML) objective

It can estimate variance components and compute polygenic scores by best linear unbiased prediction (BLUP)

The matrix-vector product algorithm

Modified figures by Nathaniel Pope (Oregon)

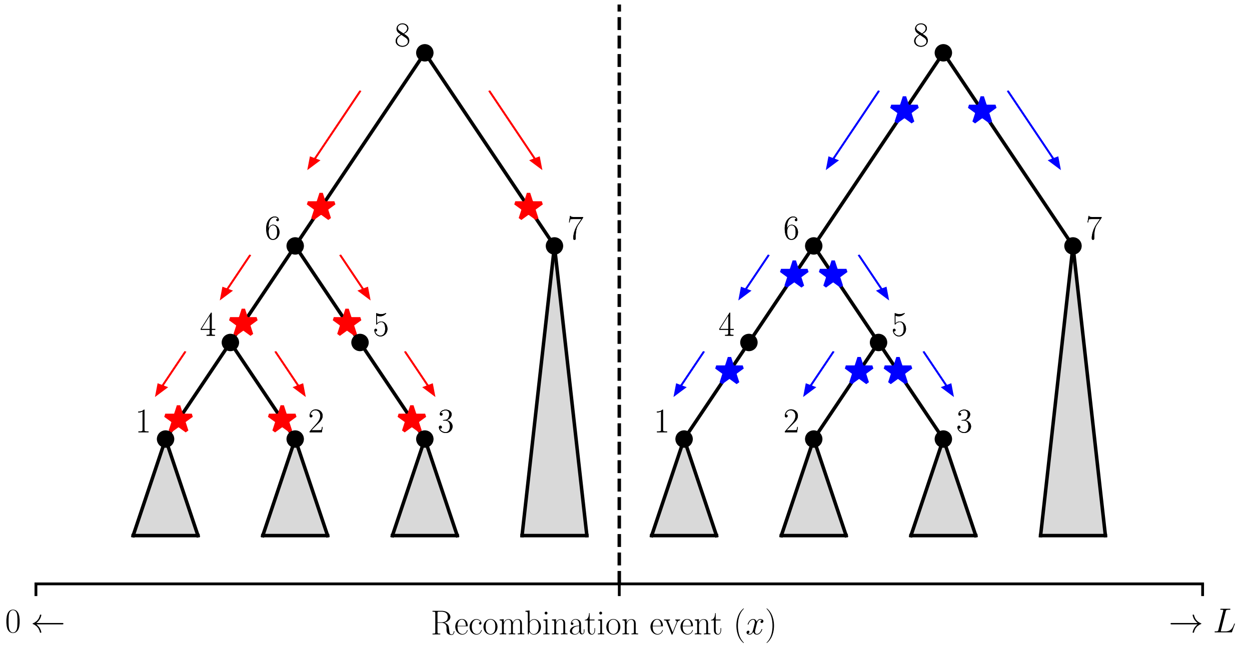
The matrix-vector product algorithm



The algorithm needs to pass mutations to the correct samples

Modified figures by Nathaniel Pope (Oregon)

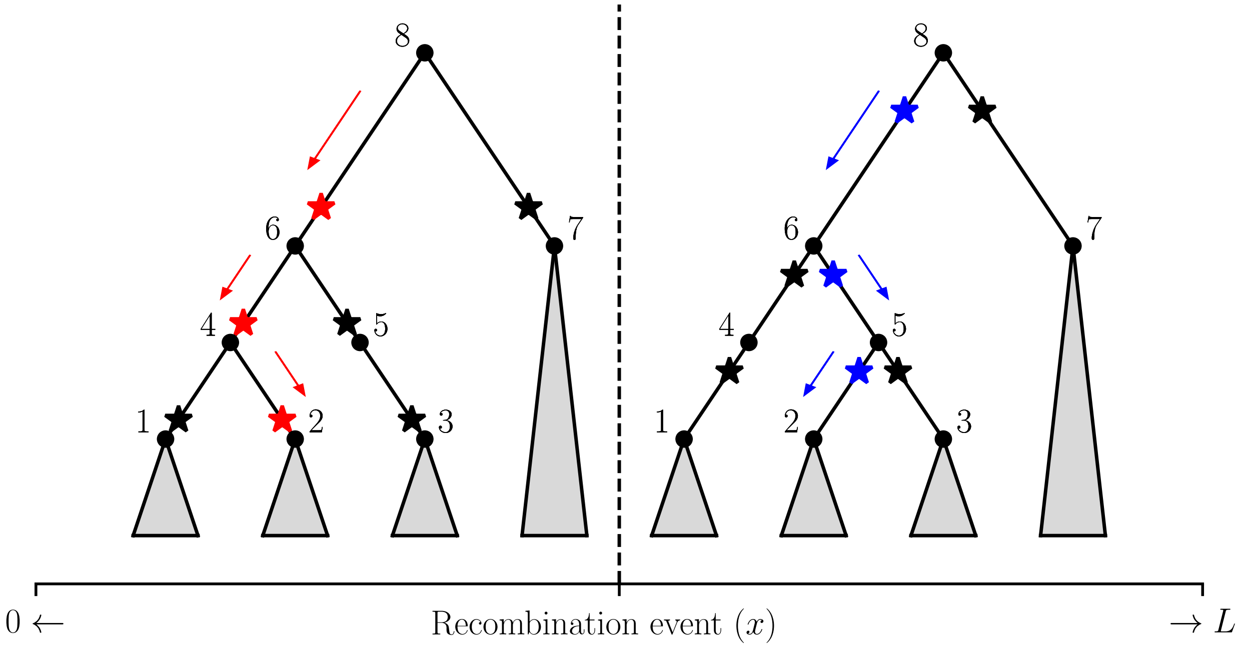
The matrix-vector product algorithm



A naive approach is to push the mutations down to the leaves every time

Modified figures by Nathaniel Pope (Oregon)

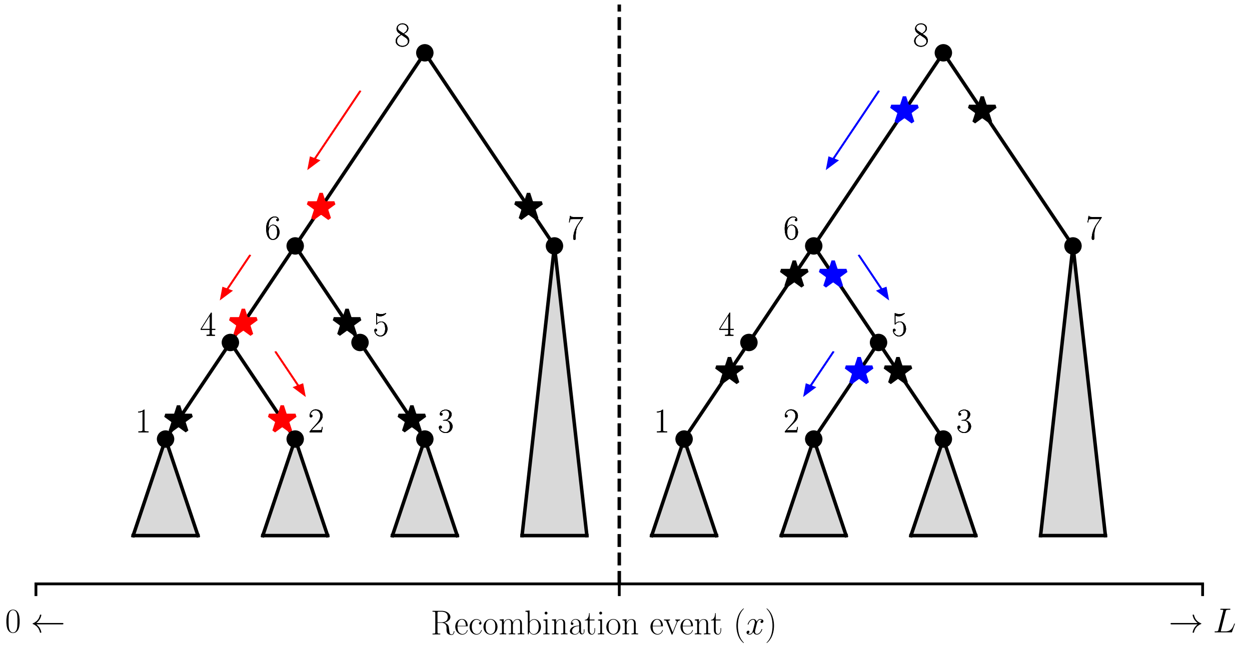
The matrix-vector product algorithm



Wait until the subtree's topology changes due to edge insertion/deletion

Modified figures by Nathaniel Pope (Oregon)

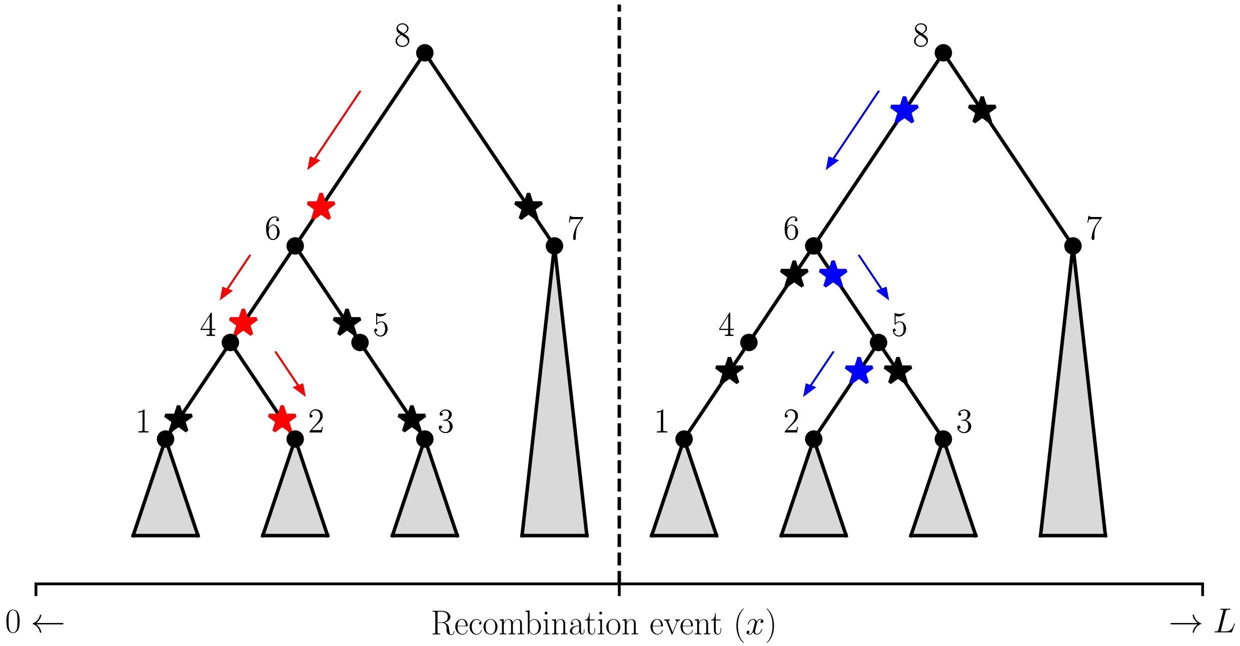
The matrix-vector product algorithm



The wrong recipient will receive the mutations if we procrastinate further

Modified figures by Nathaniel Pope (Oregon)

The matrix-vector product algorithm



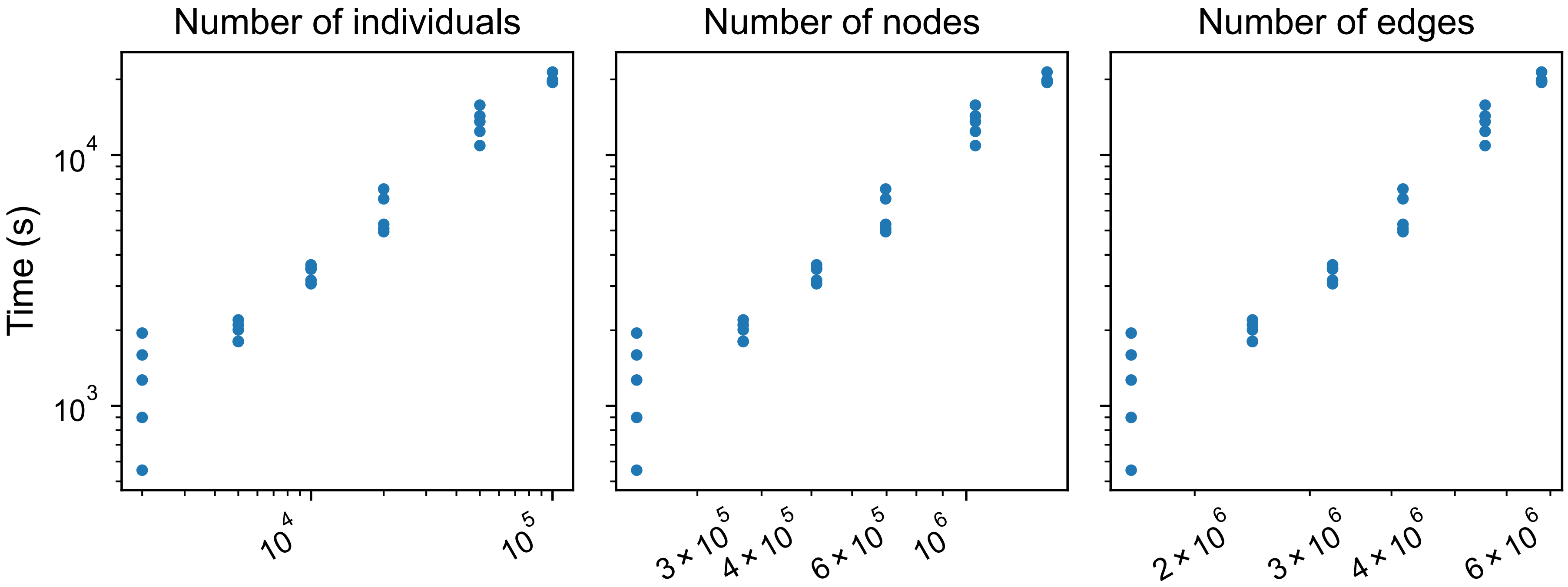
Fitting REML $\mathcal{O}(n_s^3) \;\Rightarrow\; \mathcal{O}(n_s + n_t \log n_s)$

n_s: number of samples, n_t: number of trees

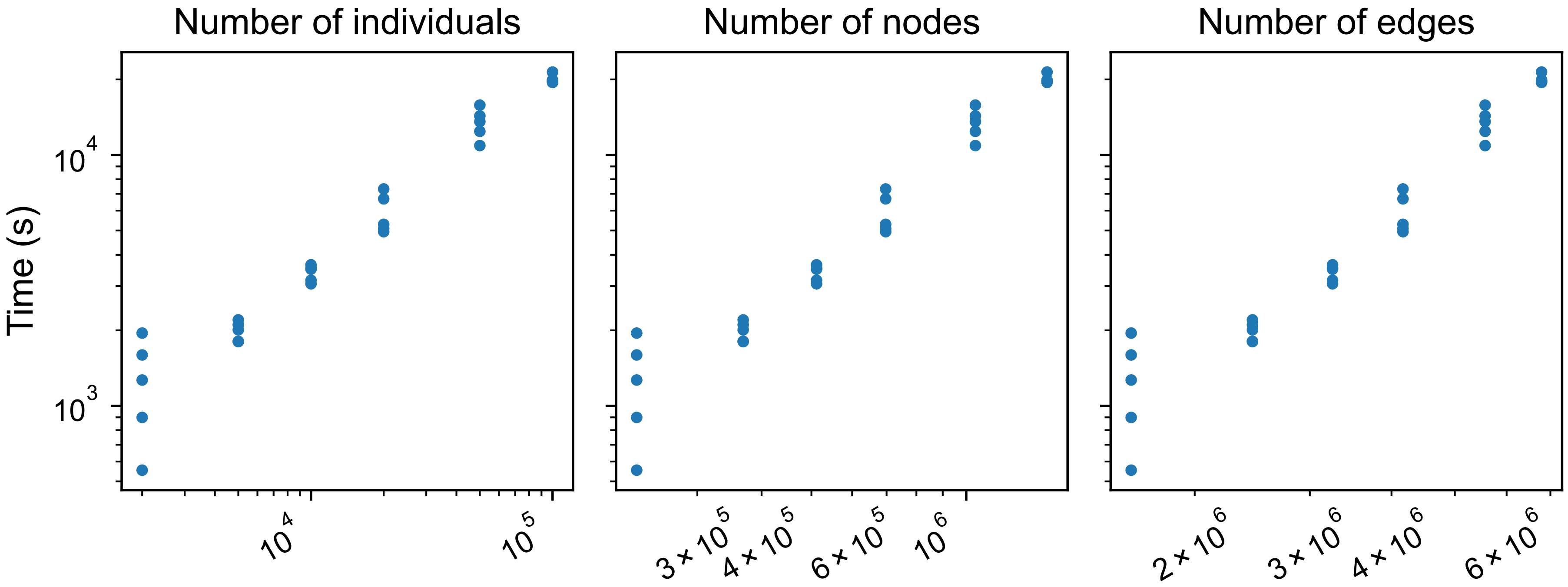
Modified figures by Nathaniel Pope (Oregon)

Runtime for variance component estimation

Runtime for variance component estimation



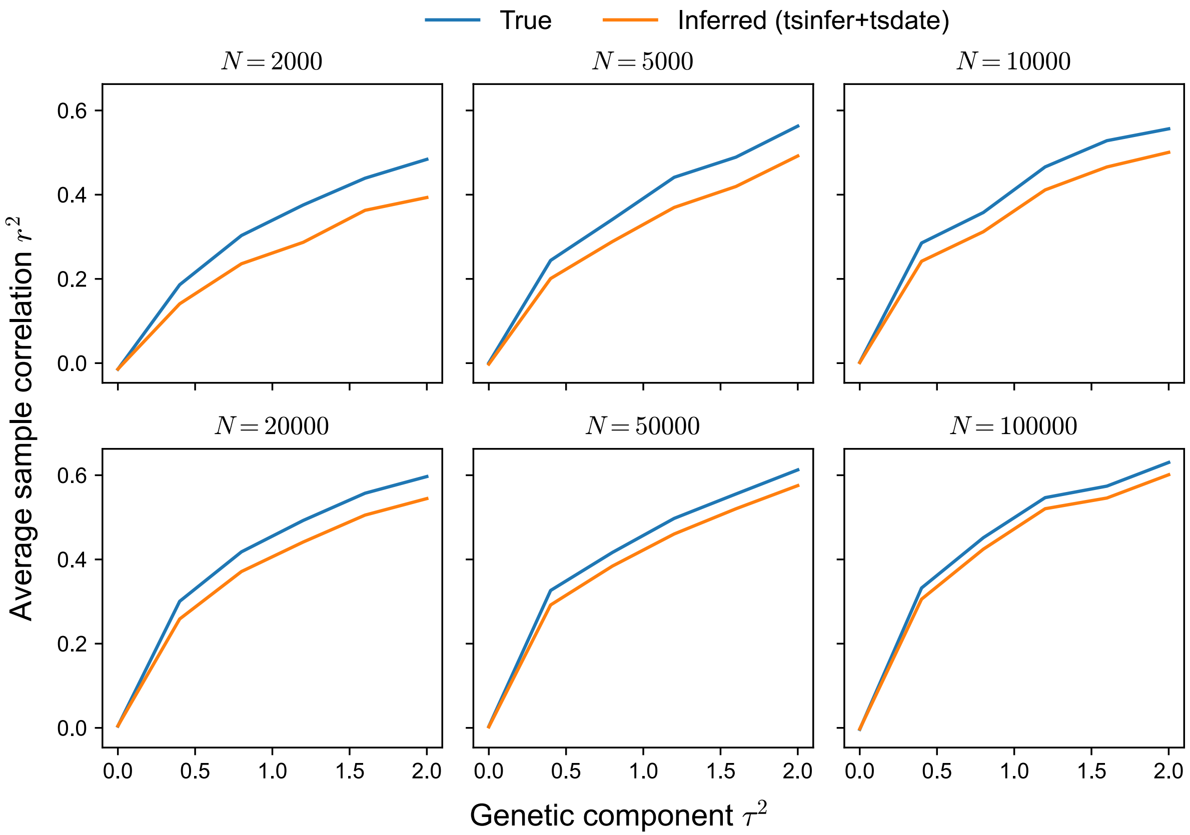
Runtime for variance component estimation



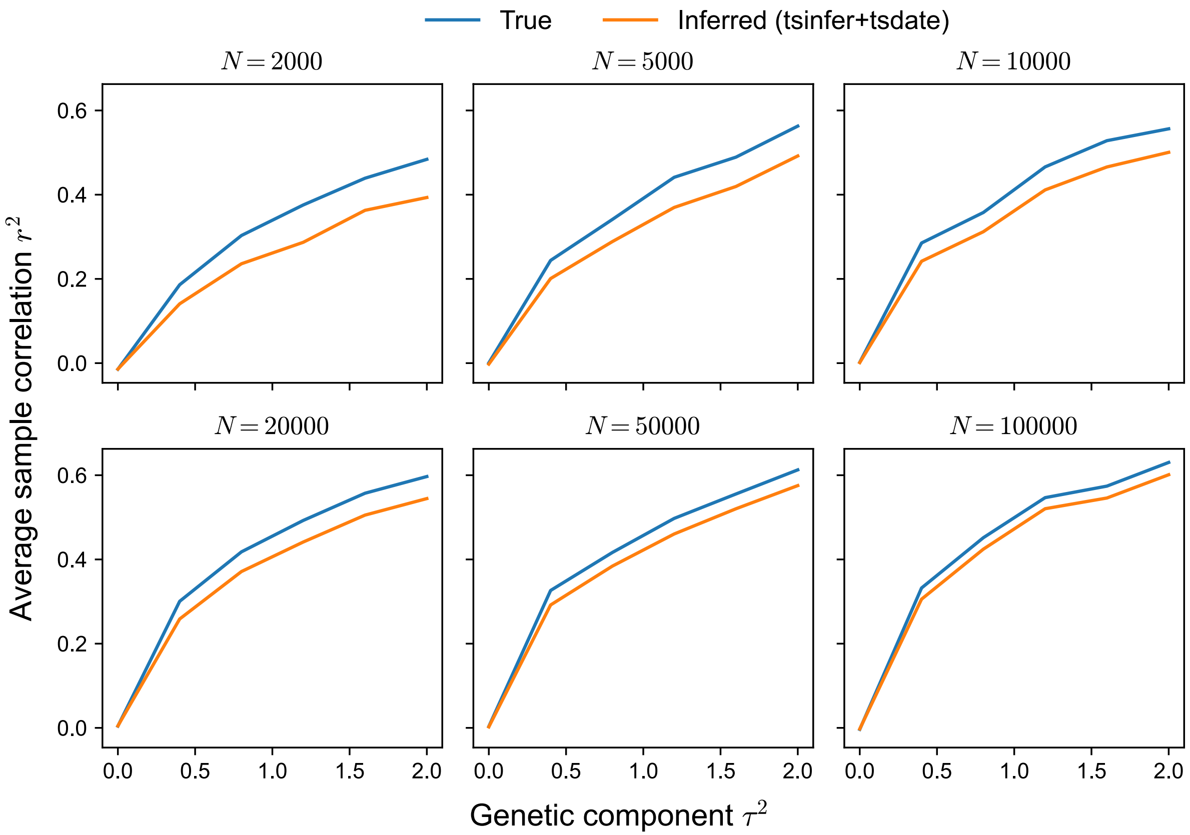
The runtime scales linearly with respect to the number of individuals (genome length = 10^8)

Best linear unbiased prediction (BLUP)

Best linear unbiased prediction (BLUP)

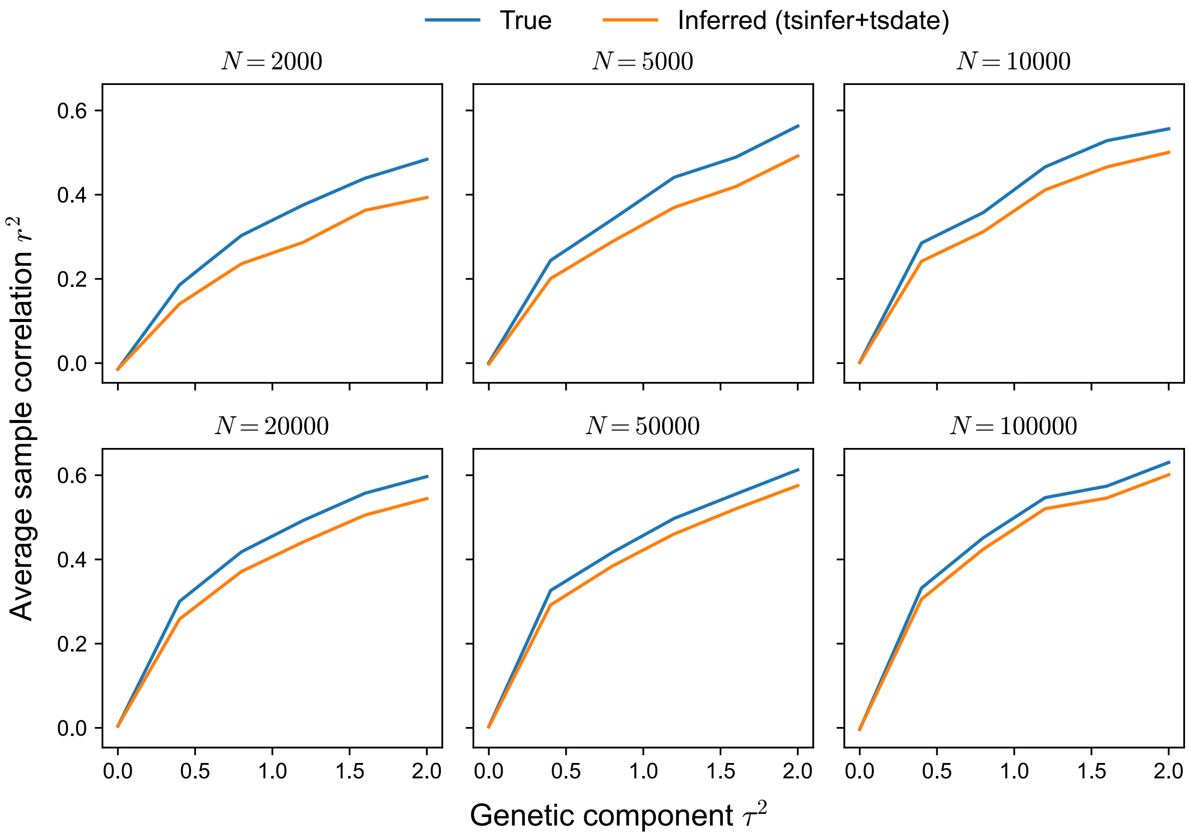


Best linear unbiased prediction (BLUP)



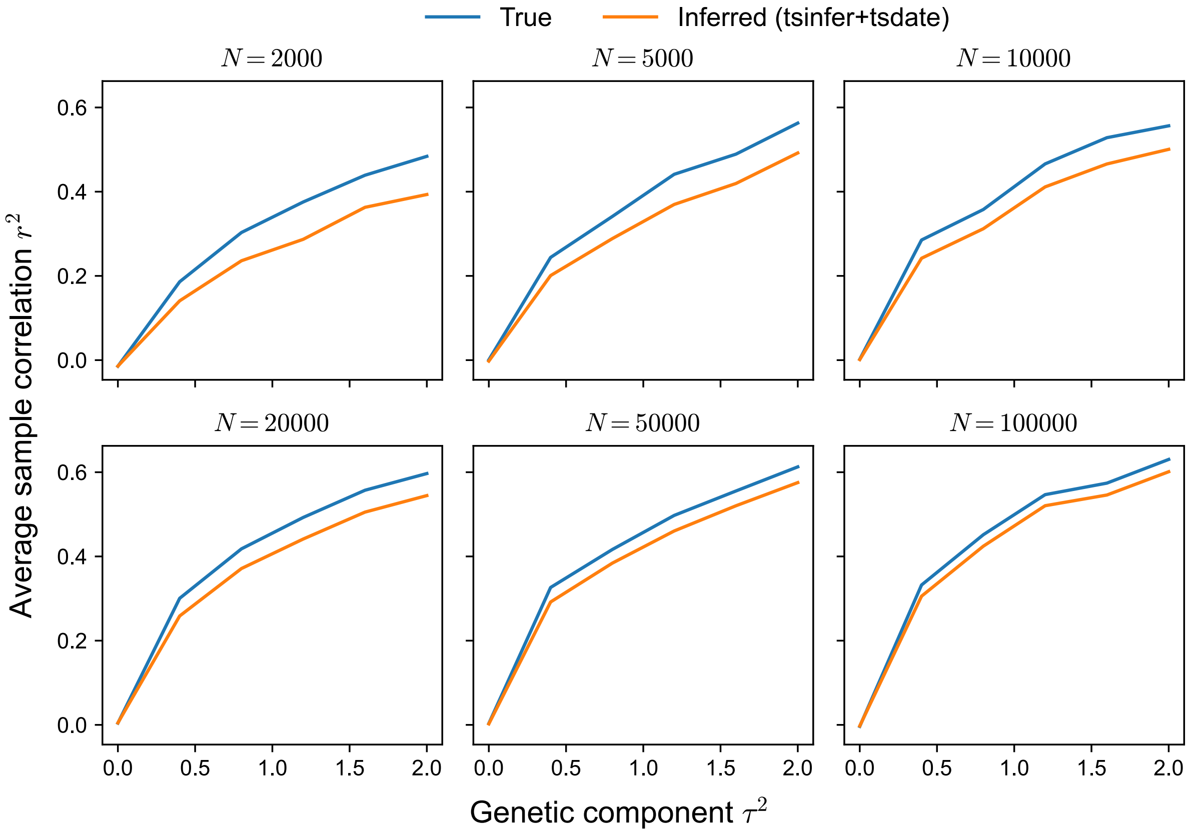
We measured the accuracy of polygenic scores computed from **tslmm**

Best linear unbiased prediction (BLUP)



Training and testing on two non-overlapping groups embedded in the same tree sequence

Best linear unbiased prediction (BLUP)



True trees are better, but inferred trees are not too behind!

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Missing heritability, Mutations vs Mendelian segregation

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Missing heritability, Mutations vs Mendelian segregation

A powerful trait simulator based on ARG-LMM (Cranmer, Brehmer, and Louppe 2020)

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Missing heritability, Mutations vs Mendelian segregation

A powerful trait simulator based on ARG-LMM (Cranmer, Brehmer, and Louppe 2020)

Super interesting technical details and proofs (10+ backup slides prepared)

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Missing heritability, Mutations vs Mendelian segregation

A powerful trait simulator based on ARG-LMM (Cranmer, Brehmer, and Louppe 2020)

Super interesting technical details and proofs (10+ backup slides prepared)

Predicting polygenic scores of internal nodes (Edge and Coop 2018; Peng, Mulder, and Edge 2024)

Summary & Future directions

ARG-LMM lays an explicit connection between population and quantitative genetics

Pseudoreplication due to shared ancestry (Rosenberg and VanLiere 2009)

Missing heritability, Mutations vs Mendelian segregation

A powerful trait simulator based on ARG-LMM (Cranmer, Brehmer, and Louppe 2020)

Super interesting technical details and proofs (10+ backup slides prepared)

Predicting polygenic scores of internal nodes (Edge and Coop 2018; Peng, Mulder, and Edge 2024)

Time conditioned analysis (random vs fixed effects) (Fan, Mancuso, and Chiang 2022)

Thank you for listening

Link to (Lehmann et al. 2025), **tslmm** preprint coming soon

Collaborators: Nathaniel Pope (Oregon), Jerome Kelleher (Oxford), Gregor Gorjanc (Edinburgh), and Peter Ralph (Oregon)

References

Browning, Sharon R., Brian L. Browning, Martha L. Daviglus, Ramon A. Durazo-Arvizu, Neil Schneiderman, Robert C. Kaplan, and Cathy C. Laurie. 2018. “Ancestry-Specific Recent Effective Population Size in the Americas.” Edited by Kirk E. Lohmueller. *PLOS Genetics* 14 (5): e1007385. https://doi.org/10.1371/journal.pgen.1007385.

Cranmer, Kyle, Johann Brehmer, and Gilles Louppe. 2020. “The Frontier of Simulation-Based Inference.” *Proceedings of the National Academy of Sciences* 117 (48): 30055–62. https://doi.org/10.1073/pnas.1912789117.

Edge, Michael D, and Graham Coop. 2018. “Reconstructing the History of Polygenic Scores Using Coalescent Trees.” *Genetics* 211 (1): 235–62. https://doi.org/10.1534/genetics.118.301687.

Fan, Caoqi, Nicholas Mancuso, and Charleston W. K. Chiang. 2022. “A Genealogical Estimate of Genetic Relationships.” *The American Journal of Human Genetics* 109 (5): 812–24. https://doi.org/10.1016/j.ajhg.2022.03.016.

Lehmann, Brieuc, Hanbin Lee, Luke Anderson-Trocme, Jerome Kelleher, Gregor Gorjanc, and Peter L. Ralph. 2025. “On ARGs, Pedigrees, and Genetic Relatedness Matrices,” March. https://doi.org/10.1101/2025.03.03.641310.

Peng, Dandan, Obadiah J. Mulder, and Michael D. Edge. 2024. “Evaluating ARG-Estimation Methods in the Context of Estimating Population-Mean Polygenic Score Histories,” May. https://doi.org/10.1101/2024.05.24.595829.

Rosenberg, Noah A., and Jenna M. VanLiere. 2009. “Replication of Genetic Associations as Pseudoreplication Due to Shared Genealogy.” *Genetic Epidemiology* 33 (6): 479–87. https://doi.org/10.1002/gepi.20400.

Salehi Nowbandegani, Pouria, Anthony Wilder Wohns, Jenna L. Ballard, Eric S. Lander, Alex Bloemendal, Benjamin M. Neale, and Luke J. O’Connor. 2023. “Extremely Sparse Models of Linkage Disequilibrium in Ancestrally Diverse Association Studies.” *Nature Genetics* 55 (9): 1494–1502. https://doi.org/10.1038/s41588-023-01487-8.

Wakeley, John. 2008. *Coalescent Theory*. Greenwood Village, CO: Roberts & Company.

Wong, Yan, Anastasia Ignatieva, Jere Koskela, Gregor Gorjanc, Anthony W Wohns, and Jerome Kelleher. 2024. “A General and Efficient Representation of Ancestral Recombination Graphs.” Edited by G Coop. *GENETICS* 228 (1). https://doi.org/10.1093/genetics/iyae100.

Technical Notes

Edge splitting

Edge splitting

- Nodes and edges are reused across multiple trees in a tree sequence

Edge splitting

- Nodes and edges are reused across multiple trees in a tree sequence
- Edges, in particular, may not have a unique set of samples along their span

Edge splitting

- Nodes and edges are reused across multiple trees in a tree sequence
- Edges, in particular, may not have a unique set of samples along their span
- Salehi Nowbandegani and colleagues *bricked* the edges to divide them (Salehi Nowbandegani et al. 2023)

Edge splitting

- Nodes and edges are reused across multiple trees in a tree sequence
- Edges, in particular, may not have a unique set of samples along their span
- Salehi Nowbandegani and colleagues *bricked* the edges to divide them (Salehi Nowbandegani et al. 2023)
- Henceforth, we assume that edges are splitted to have a unique subtopology

$$\mathbf{Z}_{ne} = \text{The number of haplotypes of individual } n \text{ that inherit } e$$

The overall matrix $\mathbf{Z}$ is an individual-edge design matrix.

Collapsing variants to edges

Collapsing variants to edges

- When does an individual possess a derived allele? $(\mathrm{ancestral} = 0, \mathrm{derived} = 1)$

Collapsing variants to edges

- When does an individual possess a derived allele? ($\mathrm{ancestral} = 0, \mathrm{derived} = 1$)
- Let $\mathbf{1}_{ep}$ be the indicator *random* variable of a mutation at edge e and position p

Collapsing variants to edges

- When does an individual possess a derived allele? ($\mathrm{ancestral} = 0, \mathbf{derived} = 1$)
- Let $\mathbf{1}_{ep}$ be the indicator *random* variable of a mutation at edge e and position p

An individual should be a descendant of an edge ($\mathbf{Z}_{ne} = 1$)

Collapsing variants to edges

- When does an individual possess a derived allele? ($\mathrm{ancestral} = 0, \mathrm{derived} = 1$)
- Let $\mathbf{1}_{ep}$ be the indicator *random* variable of a mutation at edge e and position p

An individual should be a descendant of an edge ($\mathbf{Z}_{ne} = 1$)

$$+$$

That edge should have mutation ($\mathbf{1}_{ep} = 1$)

Collapsing variants to edges

- When does an individual possess a derived allele? ($\text{ancestral} = 0, \textbf{derived} = 1$)
- Let $\mathbf{1}_{ep}$ be the indicator *random* variable of a mutation at edge e and position p

An individual should be a descendant of an edge ($\mathbf{Z}_{ne} = 1$)

$$+$$

That edge should have mutation ($\mathbf{1}_{ep} = 1$)

$$\mathbf{G}_{np} = \sum_{e:p \in e} \mathbf{Z}_{ne} \mathbf{1}_{ep} \quad \Leftrightarrow \quad \mathbf{G}_{p} = \sum_{e:p \in e} \mathbf{Z}_{e} \mathbf{1}_{ep}$$

Collapsing variants to edges

- When does an individual possess a derived allele? ($\text{ancestral} = 0, \textbf{derived} = 1$)
- Let $\mathbf{1}_{ep}$ be the indicator *random* variable of a mutation at edge e and position p

An individual should be a descendant of an edge ($\mathbf{Z}_{ne} = 1$)

$$+$$

That edge should have mutation ($\mathbf{1}_{ep} = 1$)

$$\mathbf{G}_{np} = \sum_{e:p \in e} \mathbf{Z}_{ne} \mathbf{1}_{ep} \quad \Leftrightarrow \quad \mathbf{G}_{p} = \sum_{e:p \in e} \mathbf{Z}_{e} \mathbf{1}_{ep}$$

- Assumes that there are no parent-child mutation pairs, but allows *some* recurrent mutations

Exchange the summations

Exchange the summations

Recall $\mathbf{G}_p = \sum_{e:p\in e} \mathbf{Z}_e \mathbf{1}_{ep}$ and $\mathbf{y} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$

Exchange the summations

Recall $\mathbf{G}_p = \sum_{e:p\in e} \mathbf{Z}_e \mathbf{1}_{ep}$ and $\mathbf{y} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$

Substitute $\mathbf{G}_p$

$$\color{red}{\sum_{p=1}^{P} \sum_{e:p\in e}} \mathbf{Z}_e \boldsymbol{\beta}_p \mathbf{1}_{ep} + \boldsymbol{\varepsilon}$$

Exchange the summations

Recall $\mathbf{G}_p = \sum_{e:p\in e} \mathbf{Z}_e \mathbf{1}_{ep}$ and $\mathbf{y} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$

Exchange the inner and the outer summation

$$\sum_{e=1}^{E} \sum_{p:p\in e} \mathbf{Z}_e \boldsymbol{\beta}_p \mathbf{1}_{ep} + \boldsymbol{\varepsilon}$$

Exchange the summations

Recall $\mathbf{G}_p = \sum_{e:p\in e} \mathbf{Z}_e \mathbf{1}_{ep}$ and $\mathbf{y} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$

Pull out $\mathbf{Z}_e$ and group the positions nested in $p : p \in e$

$$\begin{aligned} &\sum_{e=1}^{E} \mathbf{Z}_e \left(\sum_{p:p\in e} \boldsymbol{\beta}_p \mathbf{1}_{ep} \right) + \boldsymbol{\varepsilon} \\ &= \sum_{e=1}^{E} \mathbf{Z}_e \boldsymbol{\upsilon}_e + \boldsymbol{\varepsilon} \\ &= \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \end{aligned}$$

Exchange the summations

Recall $\mathbf{G}_p = \sum_{e:p\in e} \mathbf{Z}_e \mathbf{1}_{ep}$ and $\mathbf{y} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$

Pull out $\mathbf{Z}_e$ and group the positions nested in $p : p \in e$

$$\begin{aligned} &\sum_{e=1}^{E} \mathbf{Z}_e \left(\sum_{p:p\in e} \boldsymbol{\beta}_p \mathbf{1}_{ep} \right) + \boldsymbol{\varepsilon} \\ &= \sum_{e=1}^{E} \mathbf{Z}_e \boldsymbol{\upsilon}_e + \boldsymbol{\varepsilon} \\ &= \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \end{aligned}$$

$\boldsymbol{\upsilon}$ is a random variable made up of mutation-driven random variables $\mathbf{1}_{ep}$!

Random effects are independent

Random effects are independent

- Independent entries of random effects is a key assumption of linear mixed models

Random effects are independent

- Independent entries of random effects is a key assumption of linear mixed models
- This can be *proved* in ARG-LMM, instead of assuming it

$$\mathrm{Cov}(\mathbf{u}_e, \mathbf{u}_{e'}) = \sum_{p \in e, e'} \boldsymbol{\beta}_p^2 \mathrm{Cov}(\mathbf{1}_{ep}, \mathbf{1}_{e'p})$$

Random effects are independent

- Independent entries of random effects is a key assumption of linear mixed models
- This can be *proved* in ARG-LMM, instead of assuming it

$$\mathrm{Cov}(\mathbf{u}_e, \mathbf{u}_{e'}) = \sum_{p \in e, e'} \boldsymbol{\beta}_p^2 \mathrm{Cov}(\mathbf{1}_{ep}, \mathbf{1}_{e'p})$$

- The covariance between the indicators are higher-order terms of mutation rates, so we ignore it (Wakeley 2008)

$$\begin{aligned}\mathrm{Cov}(\mathbf{1}_{ep}, \mathbf{1}_{e'p}) &= \mathrm{E}[\mathbf{1}_{ep}\mathbf{1}_{e'p}] - \mathrm{E}[\mathbf{1}_{e'p}]\mathrm{E}[\mathbf{1}_{ep}] \\ &= 0 - l_e u_{ep} l_{e'} u_{e'p} \approx 0\end{aligned}$$

where l_e is the (time-)length of edge e.

The marginal distribution of $\mathbf{u}_e$?

The marginal distribution of $\mathbf{u}_e$?

- The Gaussian prior on random effects is a popular choice

The marginal distribution of $\mathbf{u}_e$?

- The Gaussian prior on random effects is a popular choice
- One might be tempted to invoke the central limit theorem to $\mathbf{u}_e$ (sum of indicators)

$$\mathbf{u}_e \Bigg/ \sqrt{l_e \boldsymbol{s}_e} \cdot \sqrt{\frac{1}{\boldsymbol{s}_e} \sum_{p:p \in e} \boldsymbol{\beta}_p^2 u_{ep}} \to N(0, 1^2) \text{ as } \boldsymbol{s}_e \to \infty$$

where $\boldsymbol{s}_e$ is the span (in base pairs) of edge e

The marginal distribution of $\mathbf{u}_e$?

- The Gaussian prior on random effects is a popular choice
- One might be tempted to invoke the central limit theorem to $\mathbf{u}_e$ (sum of indicators)

$$\mathbf{u}_e \Bigg/ \sqrt{l_e \boldsymbol{s}_e} \cdot \sqrt{\frac{1}{\boldsymbol{s}_e} \sum_{p:p \in e} \boldsymbol{\beta}_p^2 u_{ep}} \to N(0, 1^2) \text{ as } \boldsymbol{s}_e \to \infty$$

  where $\boldsymbol{s}_e$ is the span (in base pairs) of edge e

- The convergence is unlikely to be fast enough given the small value of $\mathbf{E}[\mathbf{1}_{ep}]$ (Berry-Esseen).

The marginal distribution of $\mathbf{u}_e$?

- The Gaussian prior on random effects is a popular choice
- One might be tempted to invoke the central limit theorem to $\mathbf{u}_e$ (sum of indicators)

$$\mathbf{u}_e \Big/ \sqrt{l_e \boldsymbol{s}_e} \cdot \sqrt{\frac{1}{\boldsymbol{s}_e} \sum_{p:p \in e} \boldsymbol{\beta}_p^2 u_{ep}} \to N(0, 1^2) \text{ as } \boldsymbol{s}_e \to \infty$$

  where $\boldsymbol{s}_e$ is the span (in base pairs) of edge e

- The convergence is unlikely to be fast enough given the small value of $\mathbf{E}[\mathbf{1}_{ep}]$ (Berry-Esseen).
- Fortunately, the variance is computable and is

$$\text{Var}(\mathbf{u}_e) = l_e \boldsymbol{s}_e \cdot \frac{1}{\boldsymbol{s}_e} \sum_{p:p \in e} \boldsymbol{\beta}_p^2 u_{ep}$$

More on $\mathrm{Var}(\mathbf{u}_e)$

More on $\mathrm{Var}(\mathbf{u}_e)$

- The weight $\mathrm{Var}(\mathbf{u}_e)$ has two components

More on $\mathrm{Var}(\mathbf{u}_e)$

- The weight $\mathrm{Var}(\mathbf{u}_e)$ has two components
- The area $l_e s_e$

More on $\mathrm{Var}(\mathbf{u}_e)$

- The weight $\mathrm{Var}(\mathbf{u}_e)$ has two components
- The area $l_e \boldsymbol{s}_e$
- Mutation rate-weighted squared average of effect sizes

$$\tau_e^2 = \frac{1}{\boldsymbol{s}_e} \sum_{p: p \in e} \boldsymbol{\beta}_p^2 u_{ep}$$

More on $\mathrm{Var}(\mathbf{u}_e)$

- The weight $\mathrm{Var}(\mathbf{u}_e)$ has two components
- The area $l_e \boldsymbol{s}_e$
- Mutation rate-weighted squared average of effect sizes

$$\tau_e^2 = \frac{1}{\boldsymbol{s}_e} \sum_{p:p \in e} \boldsymbol{\beta}_p^2 u_{ep}$$

- As a measure of functional significance, variance components are confounded by the area

Fixed effects are constant under neutrality

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

$$[\mathbf{Zf}]_n = \sum_{e=1}^{E} \mathbf{Z}_{ne} \mathbf{E}\left[\sum_{p:p \in e} \boldsymbol{\beta}_p \mathbf{1}_{ep}\right]$$

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

$$\sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \left(\sum_{e:p \in e} \mathbf{Z}_{ne} l_e \right) = \sum_{p=1}^{P} \boldsymbol{\beta} u_p \cdot 2t_{\text{root},p} = \text{const. resp. to } n$$

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

$$\sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \left(\sum_{e:p\in e} \mathbf{Z}_{ne} l_e \right) = \sum_{p=1}^{P} \boldsymbol{\beta} u_p \cdot 2t_{\text{root},p} = \text{const. resp. to } n$$

- An intercept is enough to account for the fixed effects $\mathbf{Zf}$ under this condition

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

$$\sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \left(\sum_{e:p \in e} \mathbf{Z}_{ne} l_e \right) = \sum_{p=1}^{P} \boldsymbol{\beta} u_p \cdot 2t_{\text{root},p} = \text{const. resp. to } n$$

- An intercept is enough to account for the fixed effects $\mathbf{Zf}$ under this condition
- The assumption is standard in neutral settings

Fixed effects are constant under neutrality

- Suppose that $u_{ep} = u_p$, i.e., the mutation rate is constant across edges for a given position

$$\sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \left(\sum_{e:p\in e} \mathbf{Z}_{ne} l_e \right) = \sum_{p=1}^{P} \boldsymbol{\beta} u_p \cdot 2t_{\text{root},p} = \text{const. resp. to } n$$

- An intercept is enough to account for the fixed effects $\mathbf{Zf}$ under this condition
- The assumption is standard in neutral settings

Conjecture: selection $\Rightarrow$ fixed effects?

Becareful of pseudoreplication

Becareful of pseudoreplication

- Non-overlapping samples are not independent

Becareful of pseudoreplication

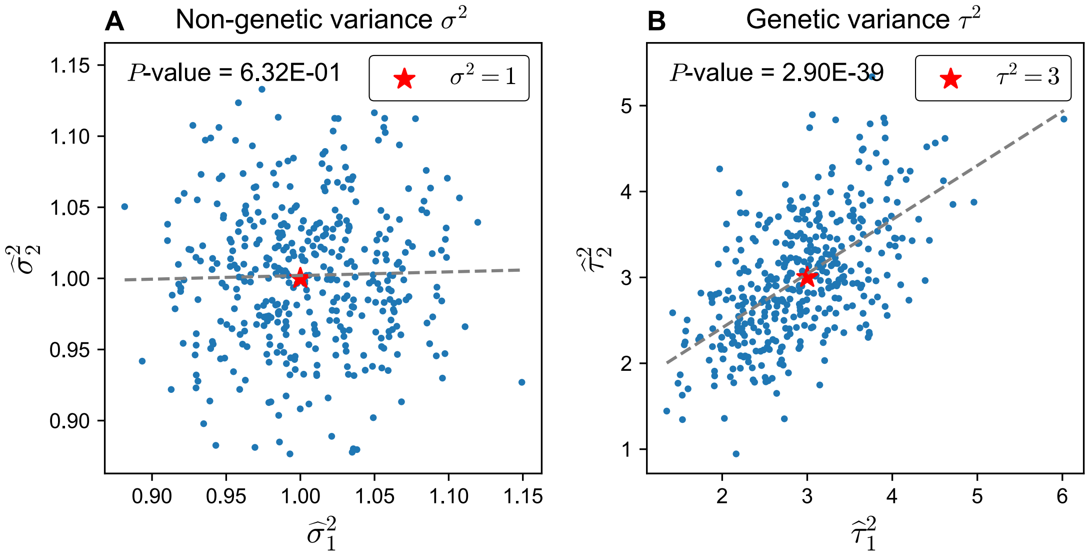
- Non-overlapping samples are not independent
- Everyone shares some amount of mutational history

Becareful of pseudoreplication

- Non-overlapping samples are not independent
- Everyone shares some amount of mutational history
- Parameter estimates (e.g. variance components) are correlated

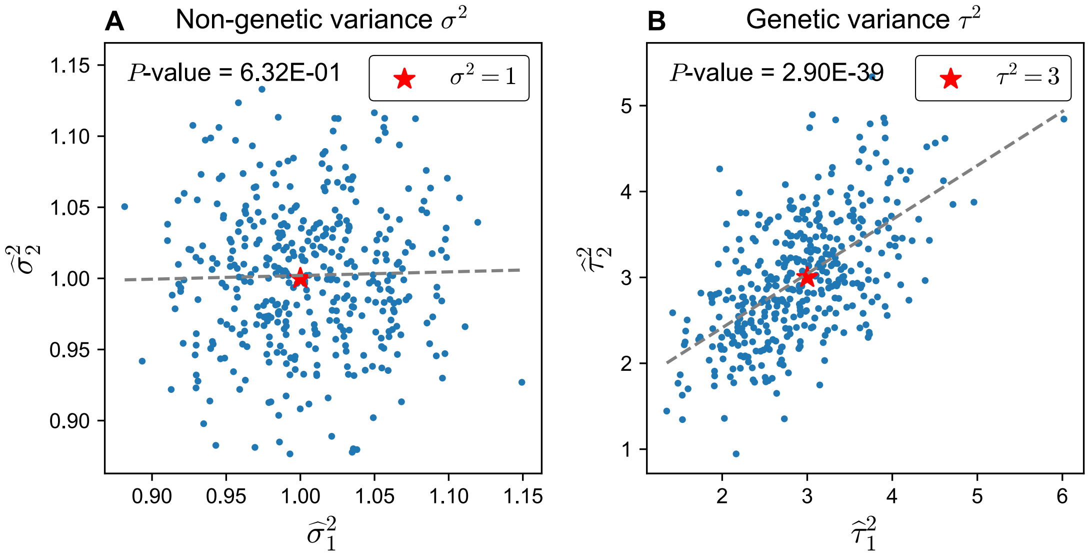
Becareful of pseudoreplication

- Non-overlapping samples are not independent
- Everyone shares some amount of mutational history
- Parameter estimates (e.g. variance components) are correlated



Becareful of pseudoreplication

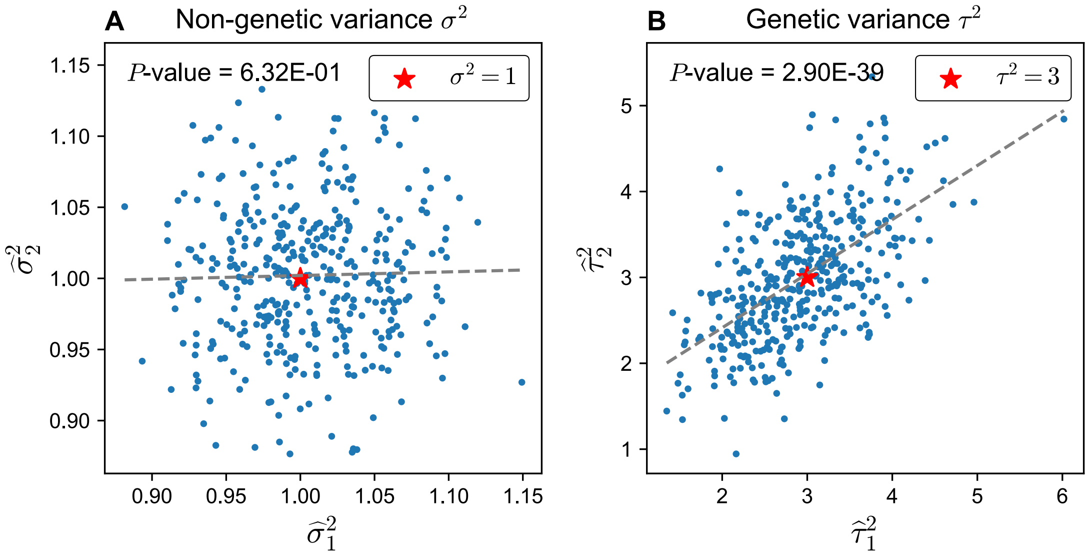
- Non-overlapping samples are not independent
- Everyone shares some amount of mutational history
- Parameter estimates (e.g. variance components) are correlated



This is also the very reason why BLUP works

Becareful of pseudoreplication

- Non-overlapping samples are not independent
- Everyone shares some amount of mutational history
- Parameter estimates (e.g. variance components) are correlated



We are all correlated!

Missing heritability?

Figure from Whalebone Magazine

Missing heritability?

ARG-LMM variance component only reflects mutational variability

Figure from Whalebone Magazine

Missing heritability?

ARG-LMM variance component only reflects mutational variability

- Pedigree-based heritability captures Mendelian segregation and mutation is ignored

Figure from Whalebone Magazine

Missing heritability?

ARG-LMM variance component only reflects mutational variability

- Pedigree-based heritability captures Mendelian segregation and mutation is ignored
- ARG-LMM’s generative model only has mutation and no segregation

Figure from Whalebone Magazine

Missing heritability?

ARG-LMM variance component only reflects mutational variability

- Pedigree-based heritability captures Mendelian segregation and mutation is ignored
- ARG-LMM's generative model only has mutation and no segregation
- Why compare quantities stemming from different random forces? (Zhang et al. 2023)

Figure from Whalebone Magazine

Missing heritability?

ARG-LMM variance component only reflects mutational variability

- Pedigree-based heritability captures Mendelian segregation and mutation is ignored
- ARG-LMM's generative model only has mutation and no segregation
- Why compare quantities stemming from different random forces? (Zhang et al. 2023)

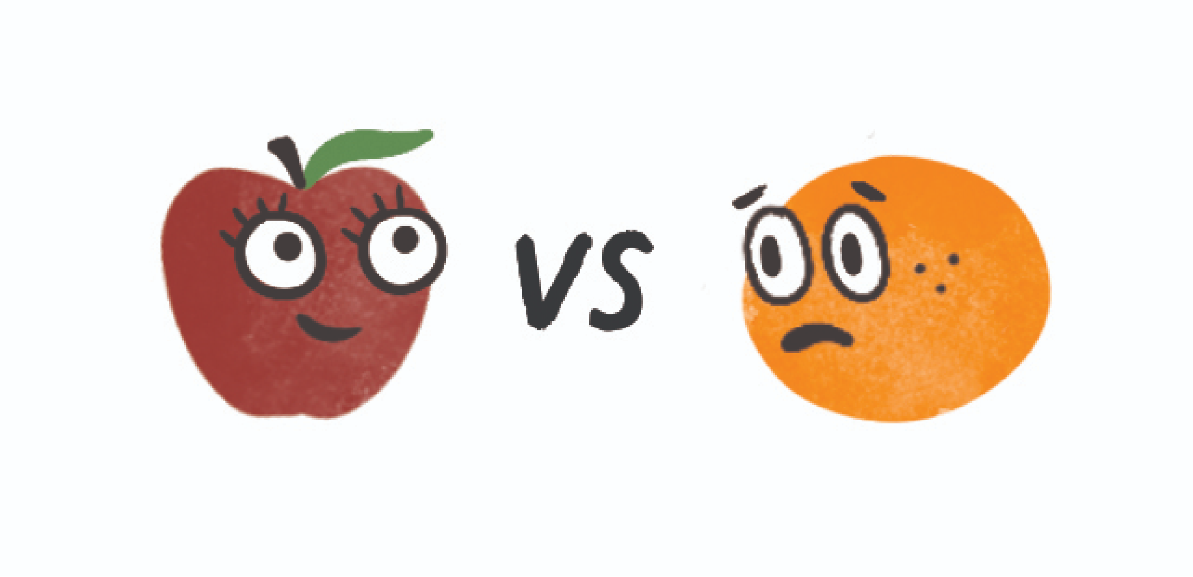
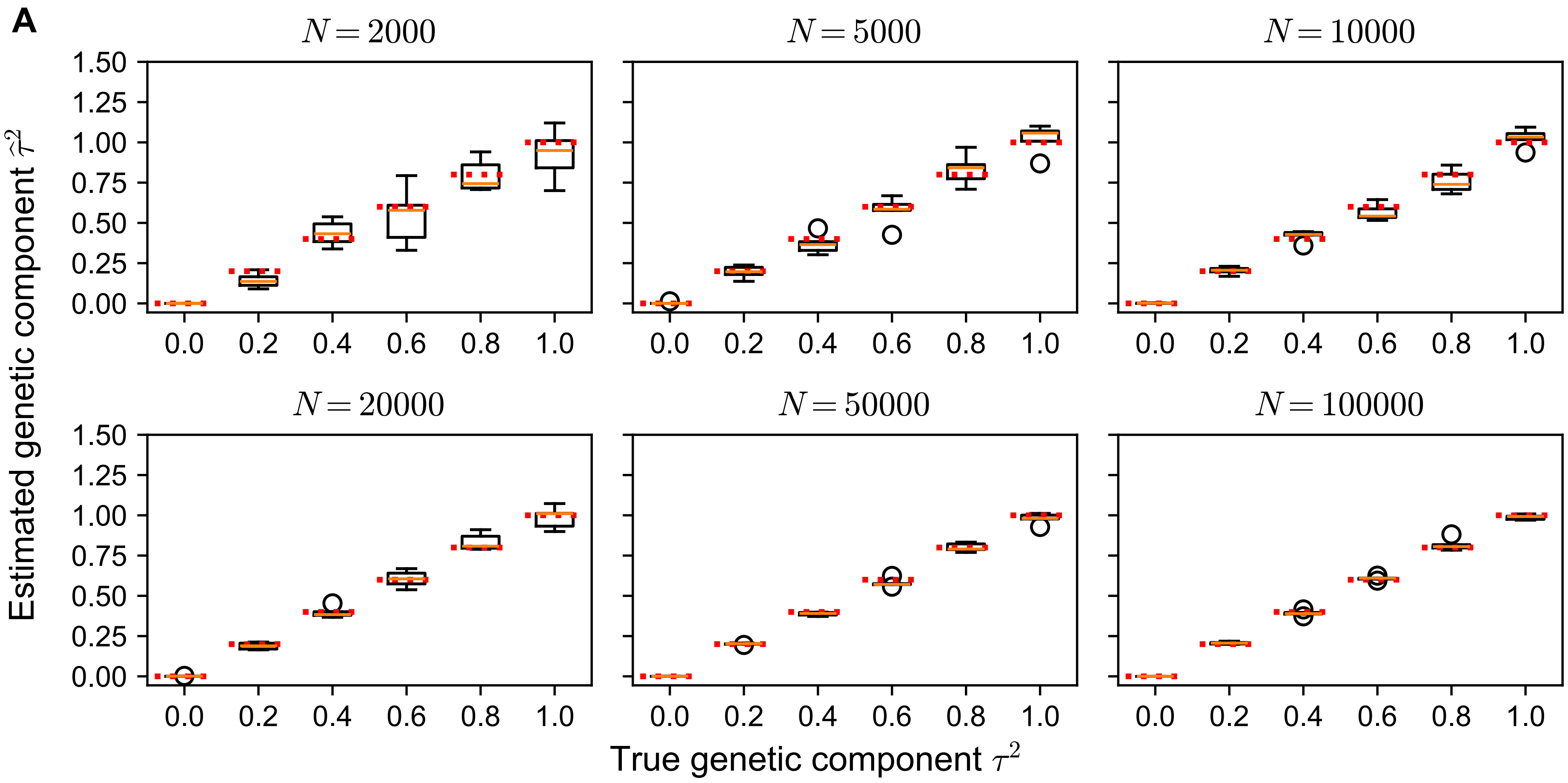


Figure from Whalebone Magazine

Estimation quality of variance components

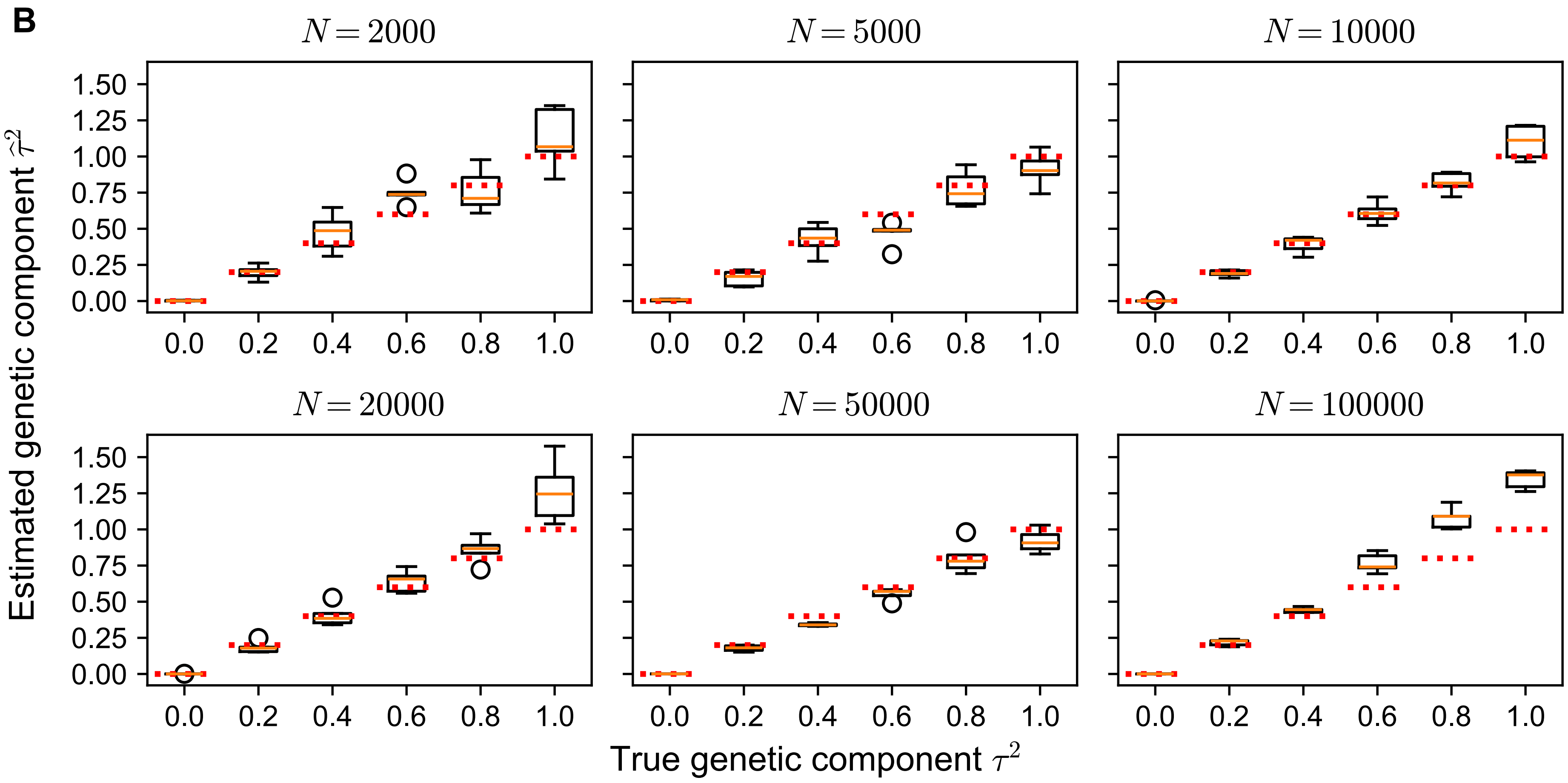
Estimation quality of variance components

Simulated trees



Estimation quality of variance components

Inferred (tsinfer+tsdate) trees



There are many genetic variances

There are many genetic variances

- ARG-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{ARG})$$

There are many genetic variances

- ARG-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{ARG})$$

- Pedigree-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{Pedigree})$$

There are many genetic variances

- ARG-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{ARG})$$

- Pedigree-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{Pedigree})$$

- Demography-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{Demography})$$

There are many genetic variances

- ARG-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{ARG})$$

- Pedigree-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{Pedigree})$$

- Demography-conditioned variance

$$\mathrm{Var}(\mathbf{y} \mid \mathrm{Demography})$$

- Time conditioning (reference population)