Statistical Inference Using Identity-by-Descent Segments

Seth D. Temple

Ph.D. Dissertation Defense

August 14, 2024

Agenda

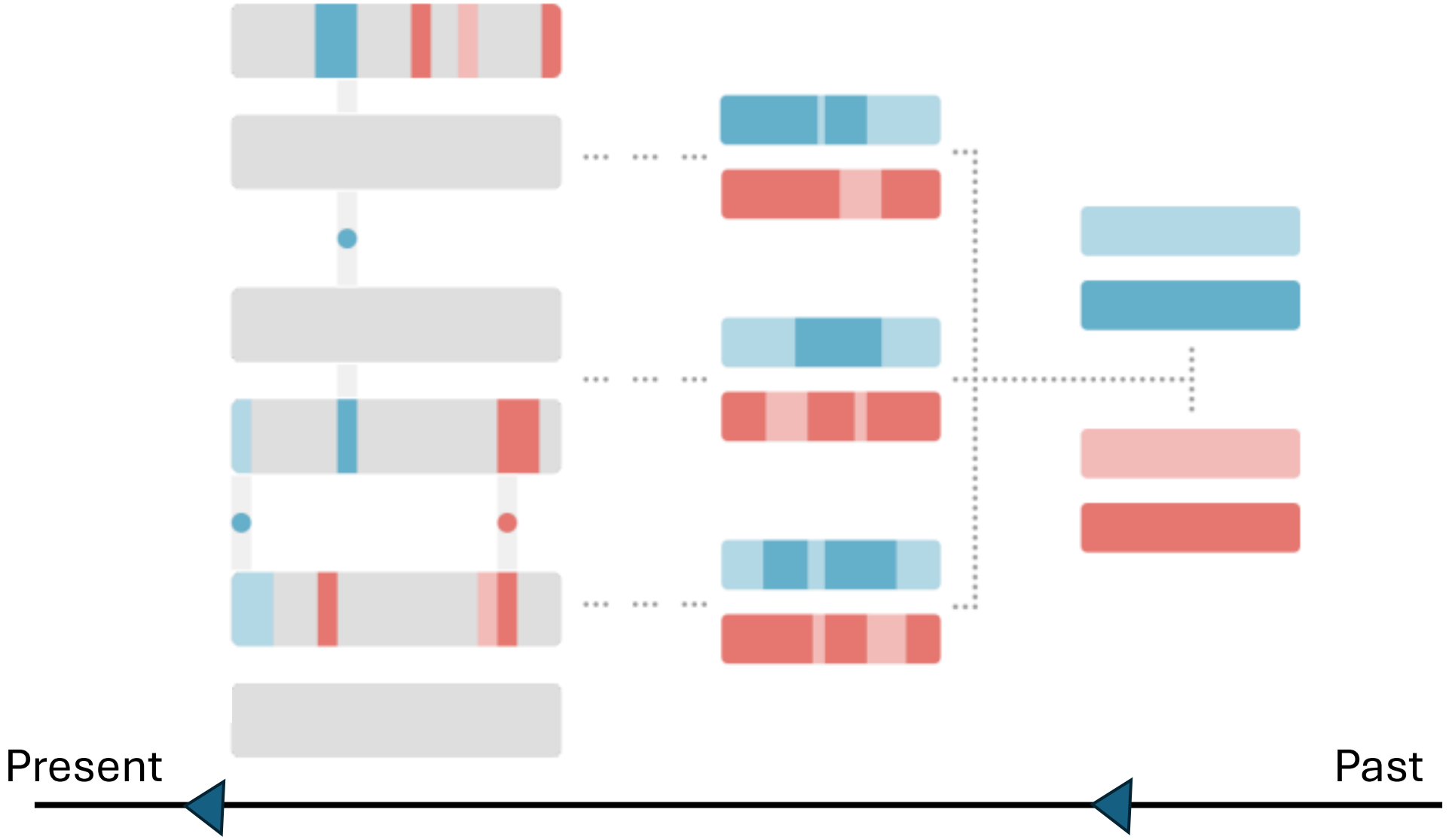
1. Adaptive evolution and identity-by-descent (IBD)
2. Asymptotic normality of IBD rate
3. Genome-wide threshold for IBD selection scan
4. Modeling selective sweeps w/ IBD data (general exam)

Motivation, details, & simulation studies in Topics 2, 3, & 4.

Case study

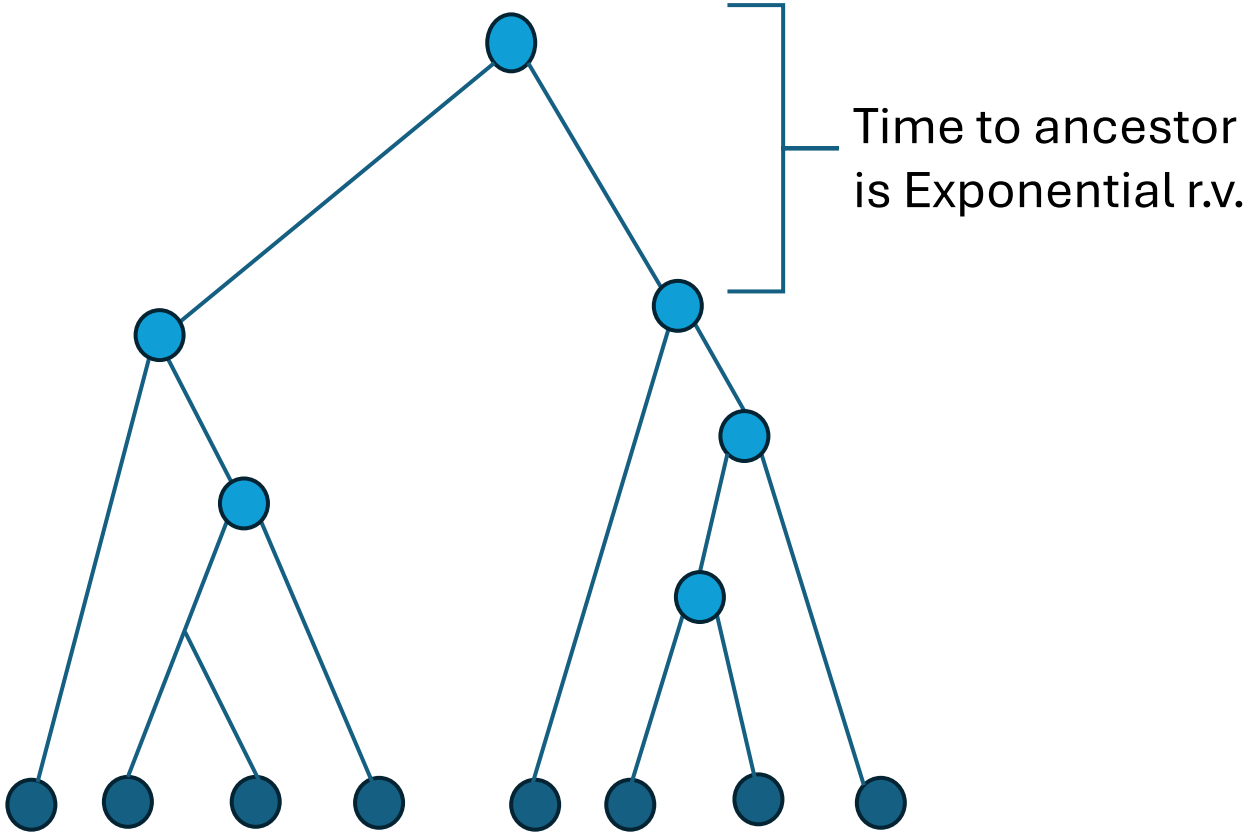
1. Change in environment: pollution
2. Black color phenotype is advantageous
3. Alleles leading to black phenotype increase in frequency

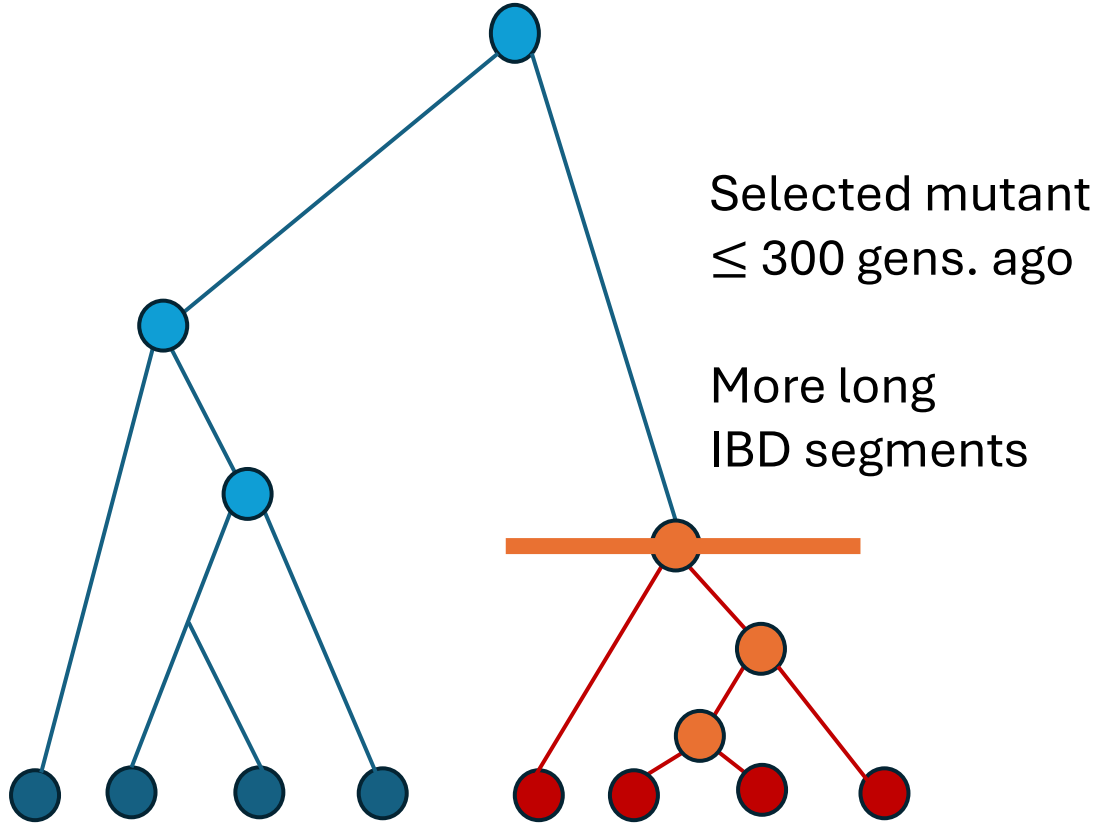
Identity-by-descent segments



Credit: AncestryDNA

Coalescent at a locus





Current sample

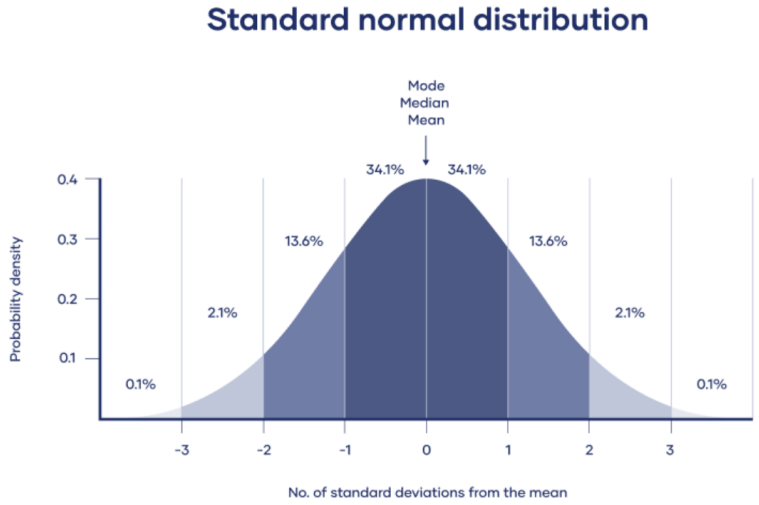
Common ancestor

Selected allele

Primary aims

In studying positive selection with IBD segments, we want to:

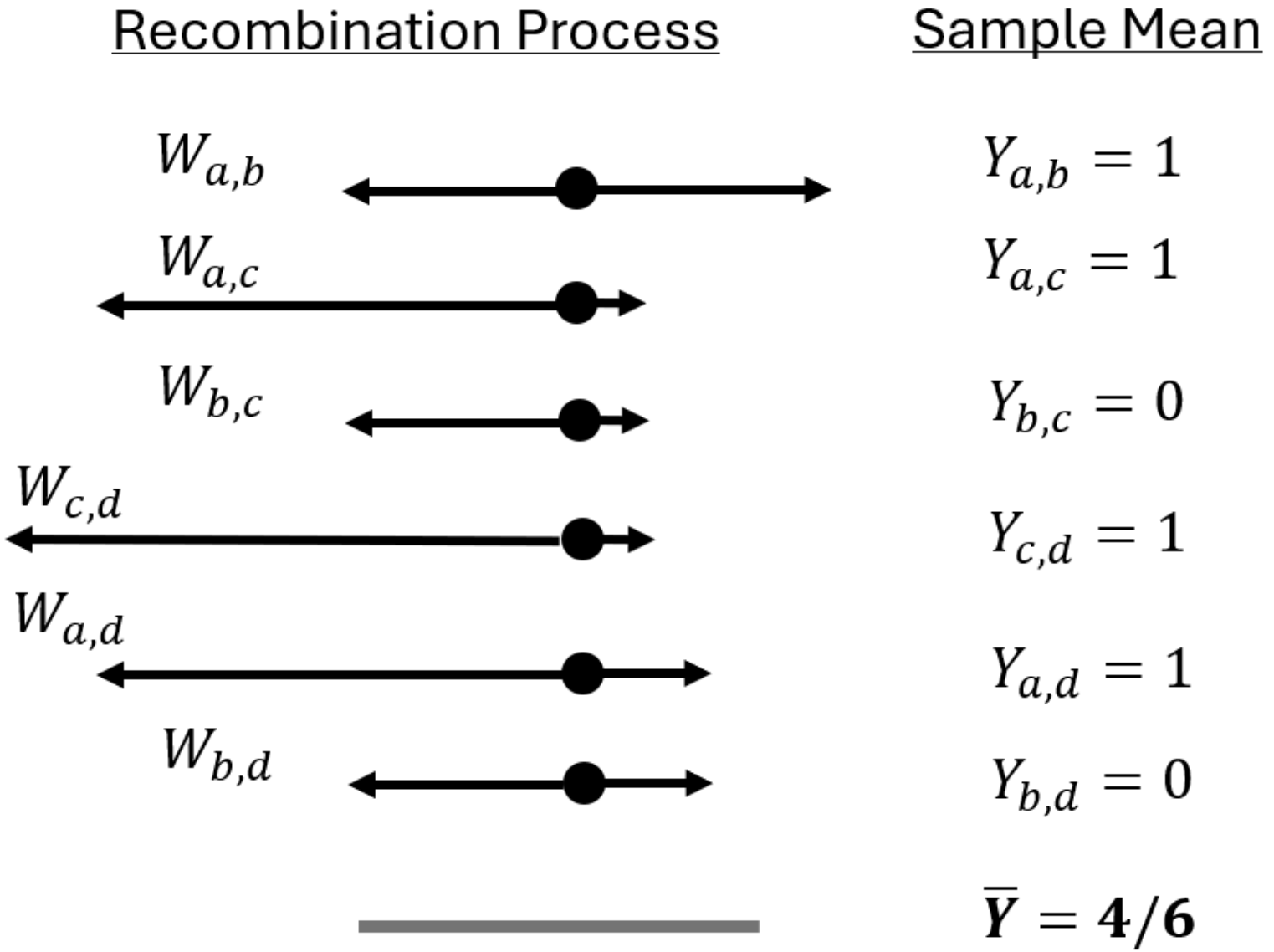
1. Propose genome-wide scan threshold that controls family-wise error rate (FWER)
2. Estimator $\hat{s}$ for selection coefficient that has desirable properties (general exam)
    - Confidence intervals (CIs) have proper coverage



Asymptotic normality of the detectable IBD rate

Temple and Thompson (2024) pre-print

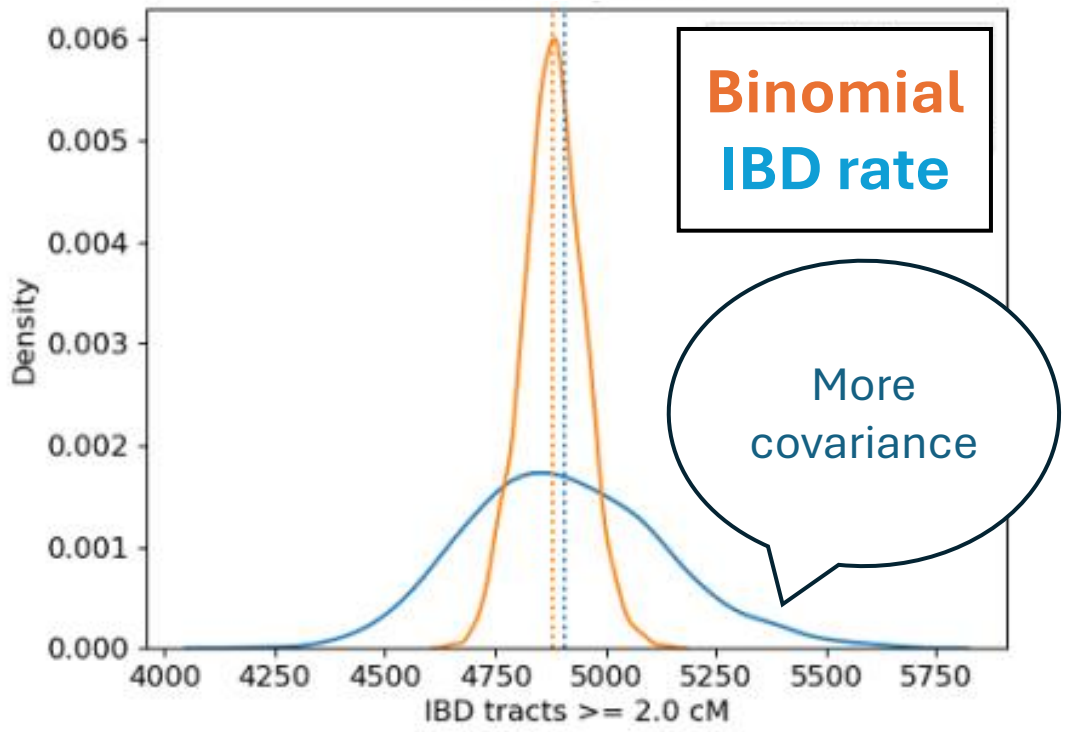
Example:
4 haplotypes



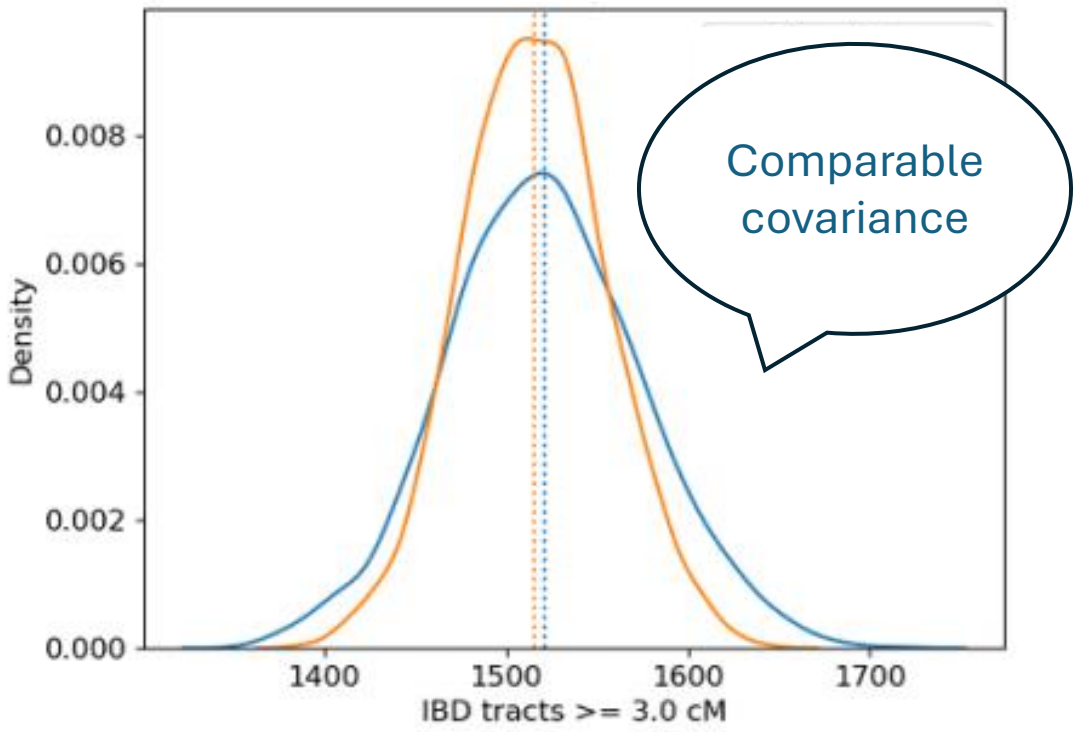
$W's$ are Erlang(shape 2) with rate depending on common ancestor time

Segment threshold w is important scale factor

2.0 cM threshold



3.0 cM threshold



The binomial model concerns ***independent*** binary observations.
"Success" prob. is $\approx$ (w cM $\times$ population size N)^{-1} .

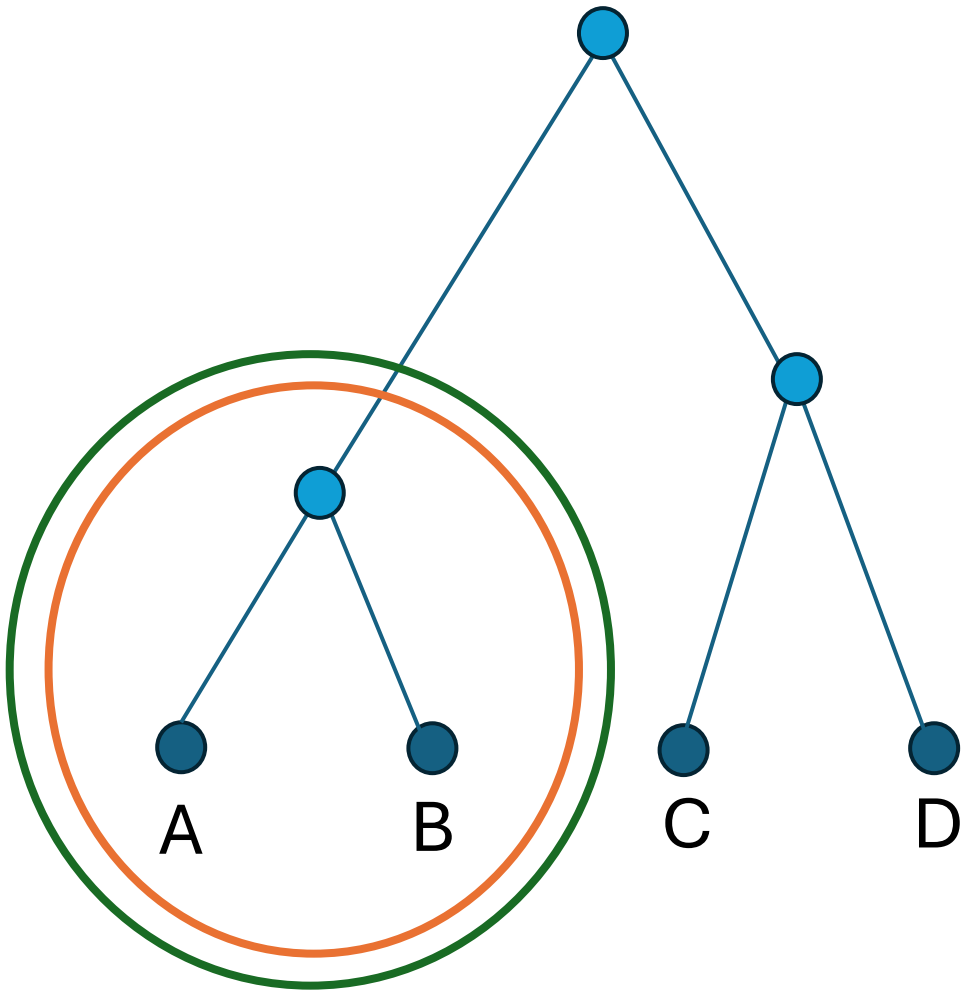
Seminal Theorem.

Under the following interpretable conditions (+1 other), the IBD rate is normally distributed.

1. Large sample size $\boldsymbol{n}$ & population size $\boldsymbol{N} \times w$
2. Population size much smaller than sample size ^2
   - You have enough data
3. Sample size much smaller than scaled population size
   - Vanishing covariance terms

Our proof uses a result from Chandrasekhar et al. (2023)
These are little $o(\cdot)$ statements

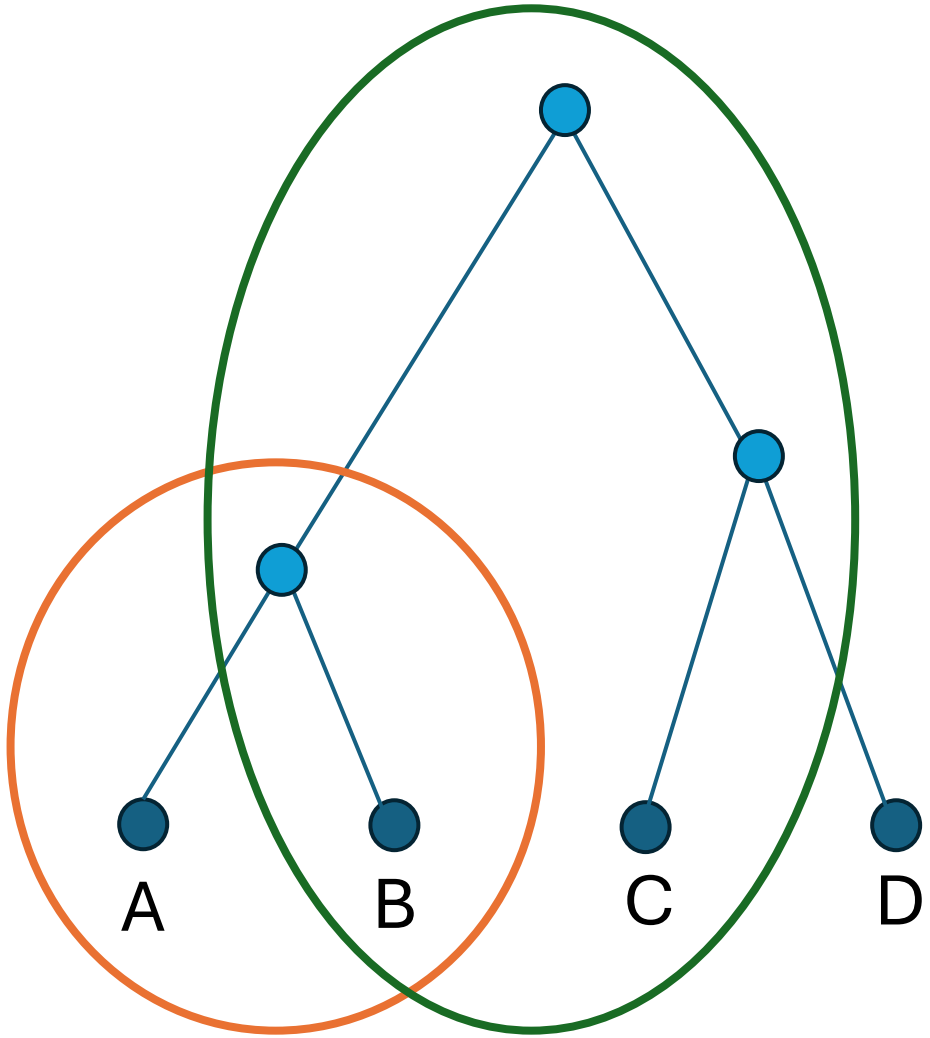
Covariance (IBD seg. **AB**, IBD seg. **AB**)



There are $\sim n^2$ of these terms

Each covariance of order $O((Nw)^{-2})$

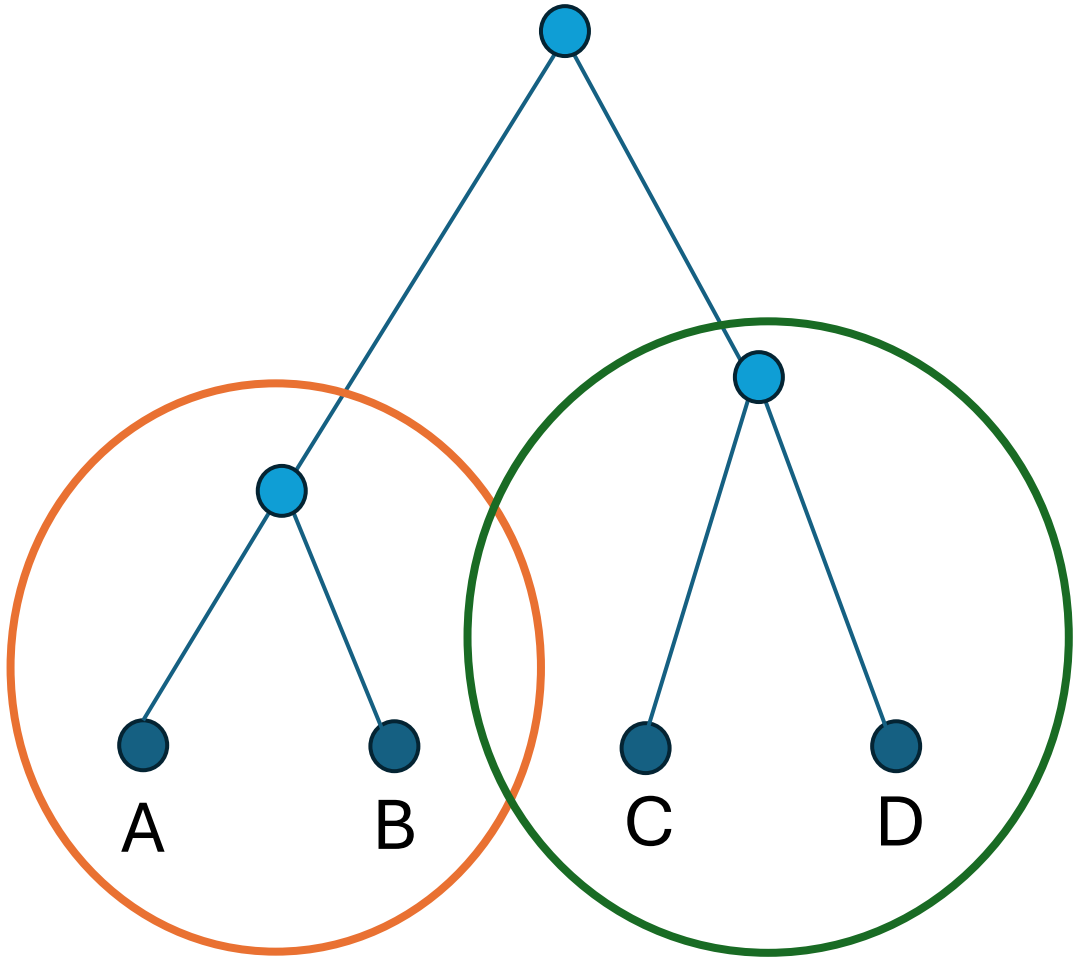
Covariance (IBD seg. **AB**, IBD seg. **BC**)



There are $\sim n^3$ of these terms

Each covariance of order $O((Nw)^{-3})$

Covariance (IBD seg. **AB**, IBD seg. **CD**)

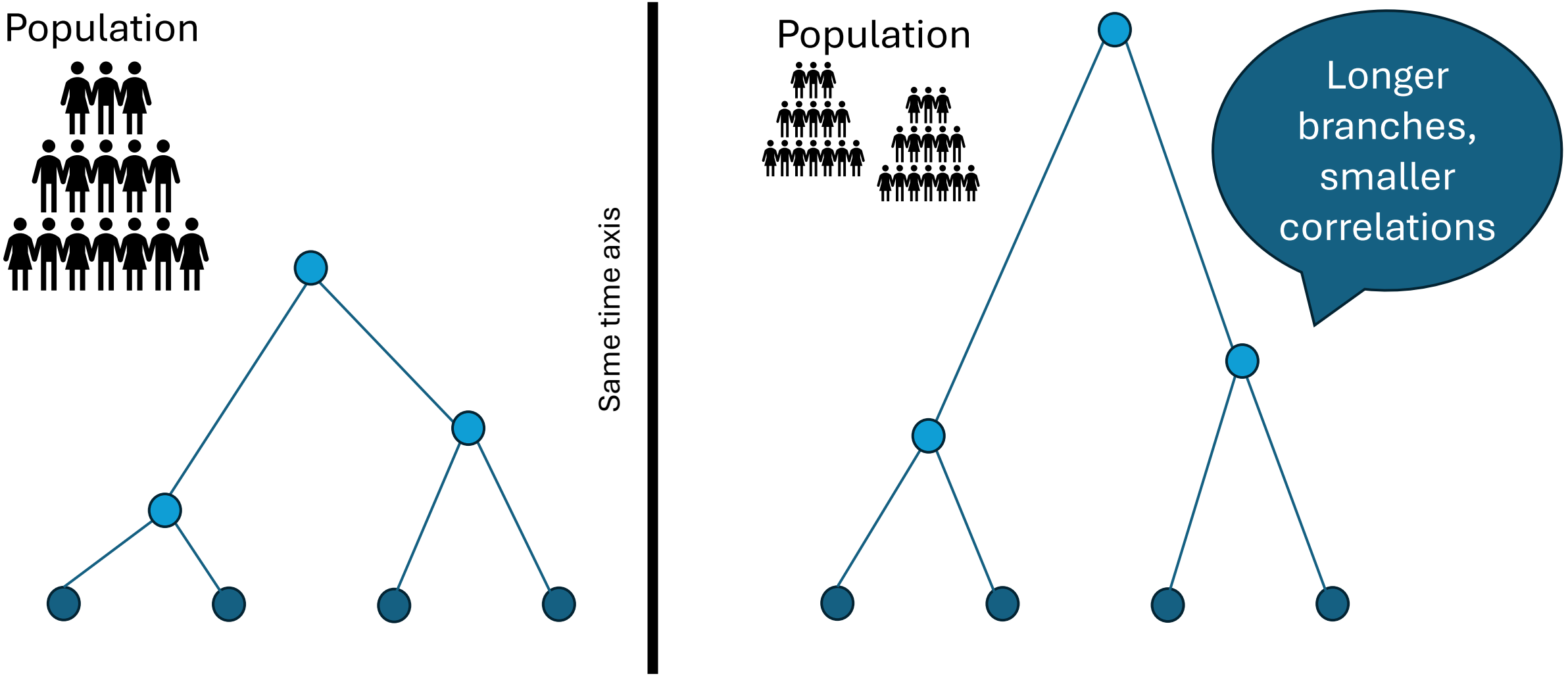


There are $\sim n^4$ of these terms

There are two possible trees

Each covariance of order $O((Nw)^{-4})$

Scaled population much larger than sample



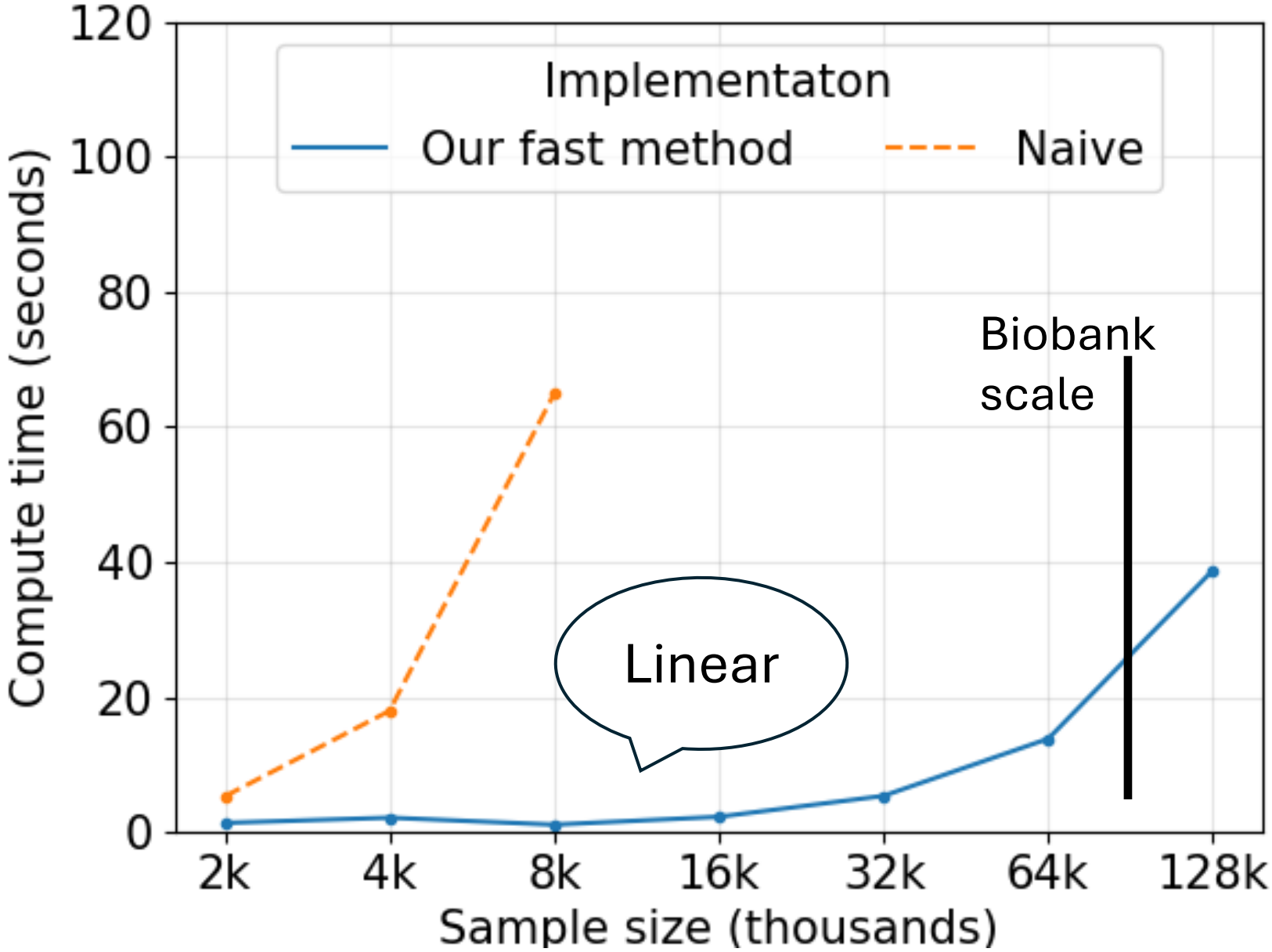
Covariances depend on Nw

Algorithm to simulate IBD segments

1. Simulate a coalescent tree
2. At each coalescent event :
    i. Draw recombination endpoints
    ii. Compare endpoints of haplotypes
    iii. Track the long enough segments

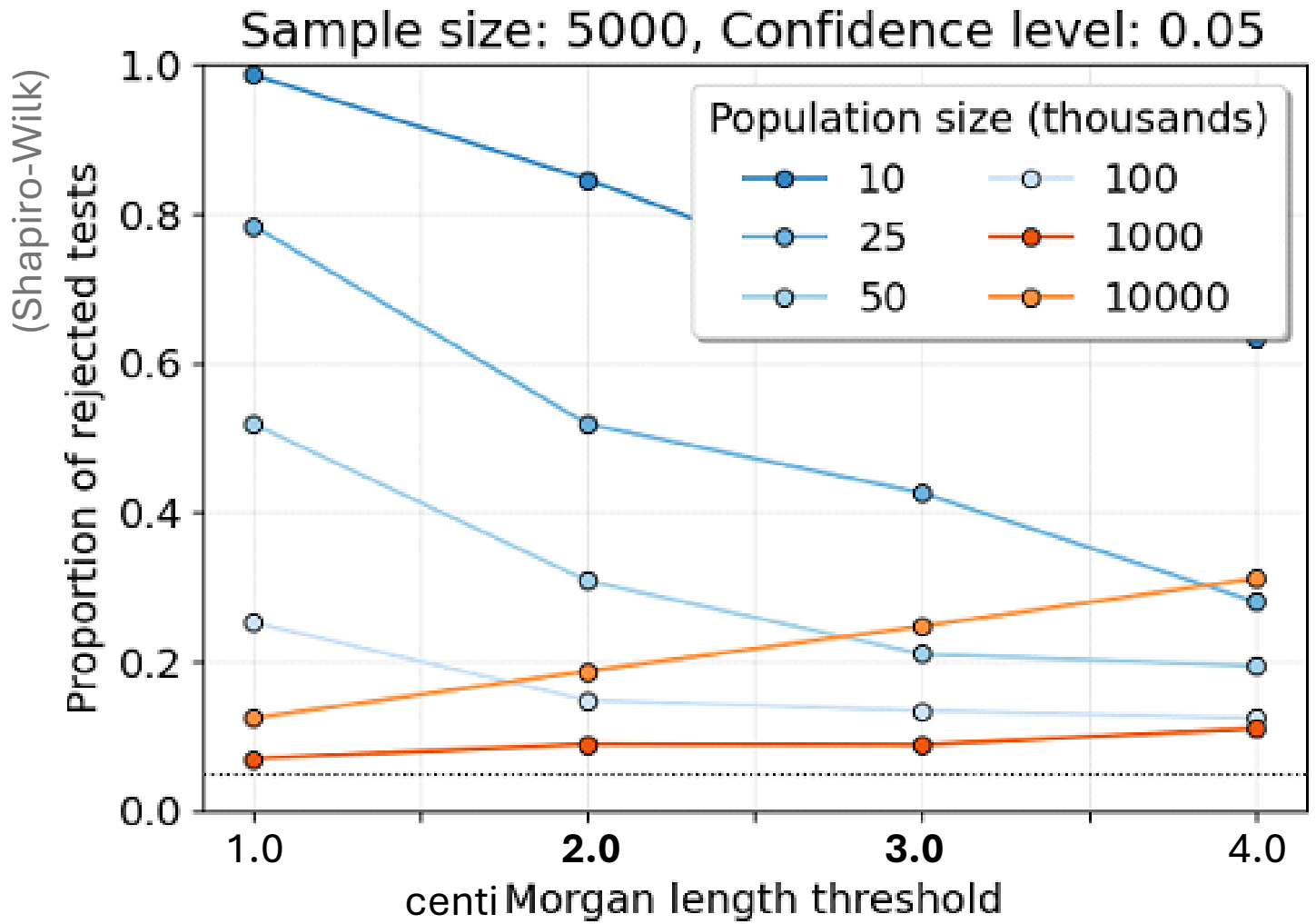
Probabilistic speed ups

- Why?
  - Billions of simulations to test distributional properties
  - Increasing sample size ^ 2
- Pruning
  - If haplotype segment shorter than w threshold, stop making comparisons
- Merging
  - If haplotype segments have same endpoints, treat them henceforth as same


Implementaton
Our fast method
Naive
Biobank
scale
Linear
Compute time (seconds)
0
20
40
60
80
100
120
2k
4k
8k
16k
32k
64k
128k
Sample size (thousands)

Central limit works, kind of

* Theorem says we should reject less when population size or cM threshold increase



Comments:

- cM threshold fixed by user
- Tradeoff between sample and population size
- When we reject null: heavier tailed distr.

Contribution to the field

- Normality motivates,
  i. Multiple testing
  ii. Confidence intervals

- Existing literature
  - Central limits hard to get in genetics
  - Shai et al. (2014) : data “looks” normally distributed
  - Palamara et al., Field et al. : data “looks” gamma distributed

$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$

μ is genome-wide IBD rate

We use a Z-test

Genome-wide threshold for IBD-based selection scan

Temple and Browning, *in progress*

"Scan" is multiple hypothesis tests along genome

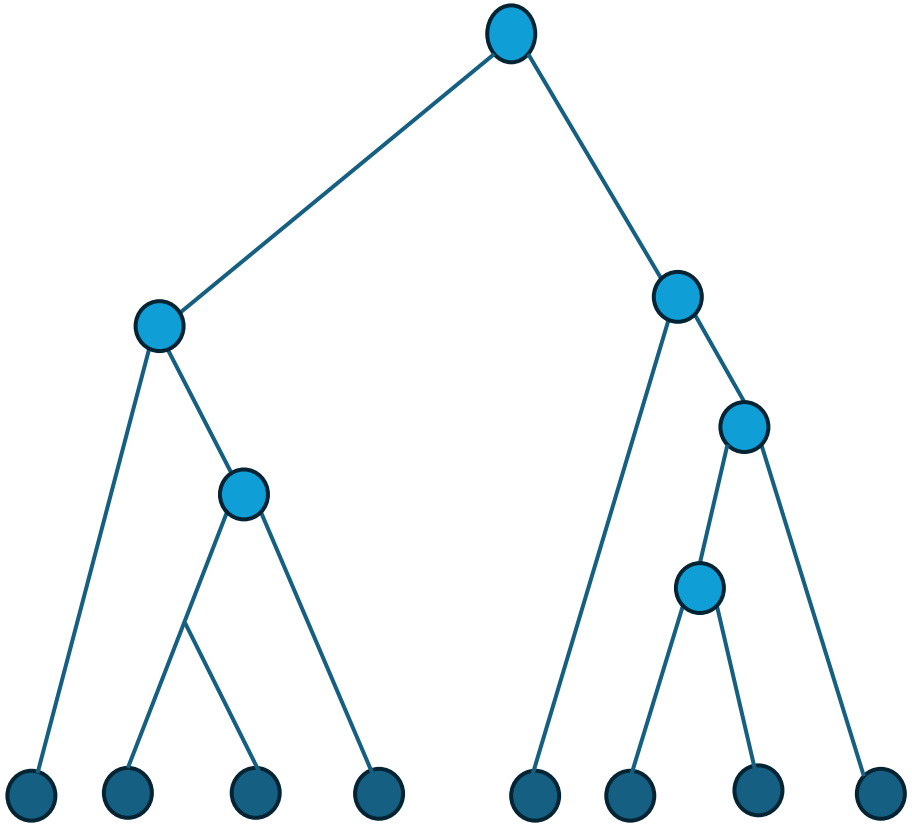
FWER existing literature

- Standard deviation thresholds
  (Sabeti et al., Voight et al., Browning x2, Temple et al.)
- Bonferroni correction
  (Palamara et al.)
- GWAS significance level 5e-8
  (Field et al., Speidel et al., Taliun et al.)

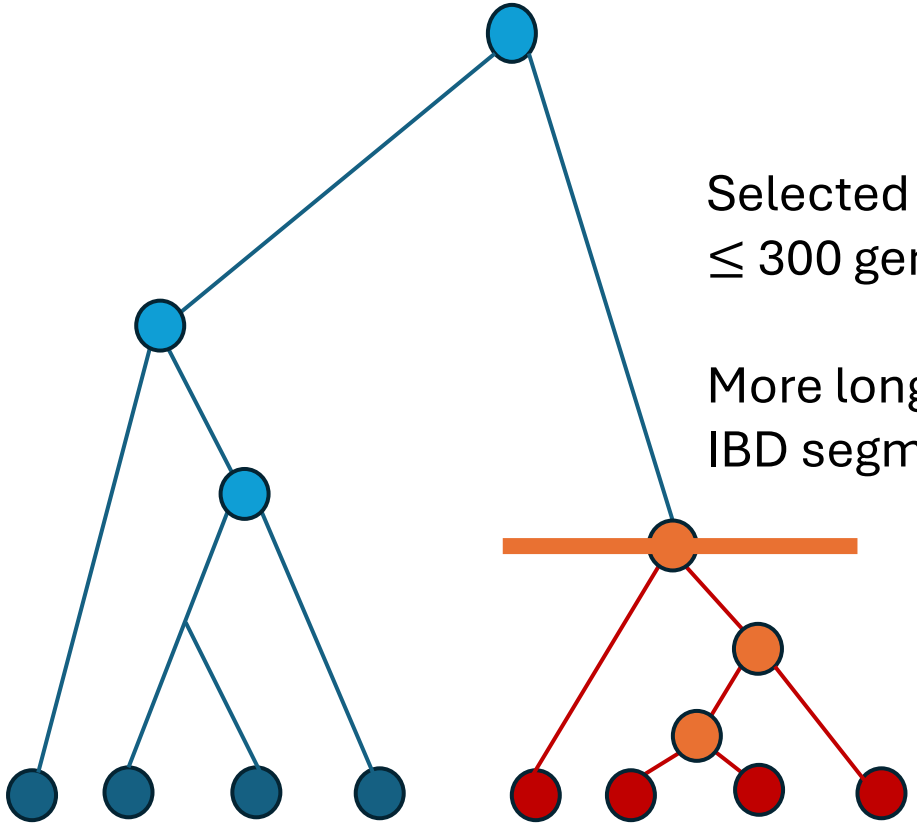
Limitation: we don’t know if any of these control FWER

Recall

Neutral locus



Selected locus

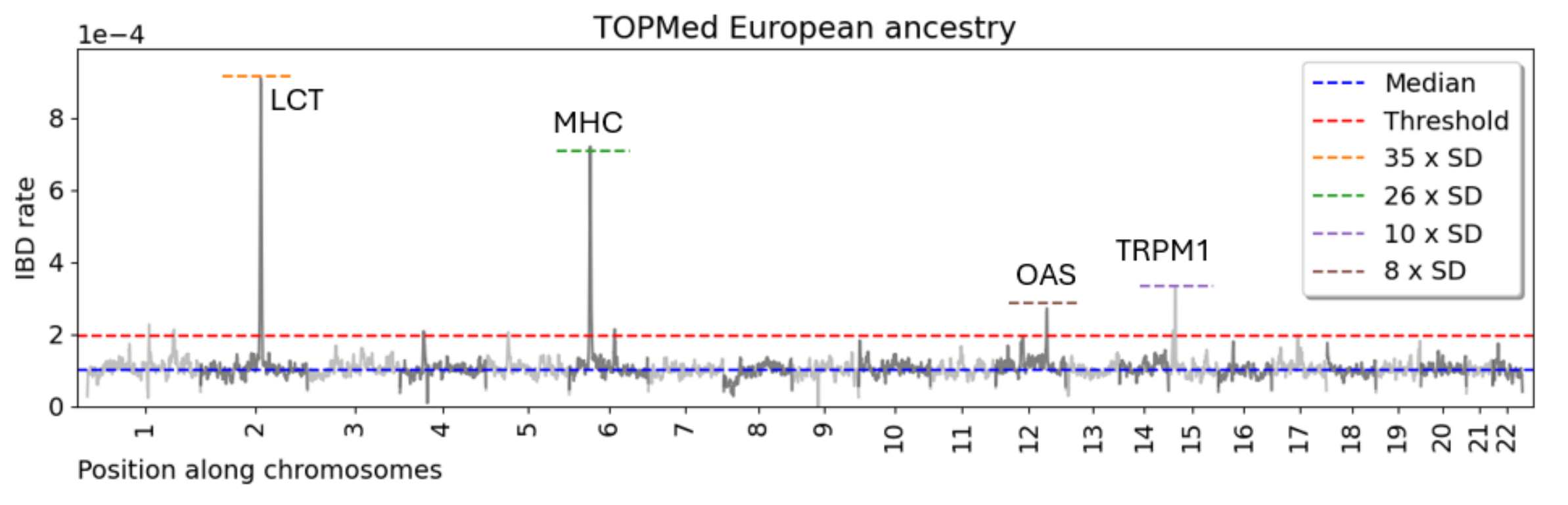


Current sample

Common ancestor

Selected allele

4 standard deviations



Modeling the Ornstein-Uhlenbeck process

Properties

- Normality at each locus
- Spatial homogeneity
- Mean-reverting
- Markov property
- Covariance form ↓↓↓

$$\textbf{Cov (IBD at locus 1, IBD at locus 2)} = \exp(-\theta \times \Delta)$$

Decay parameter θ ; spacing between loci Δ

Multiple testing methods

1. Standardize over genome
2. Fit regression : log Covariance $= -\theta \times \Delta$'s
3. Multiple testing
    i. Approximation method
    (Siegmund and Yakir (2007))

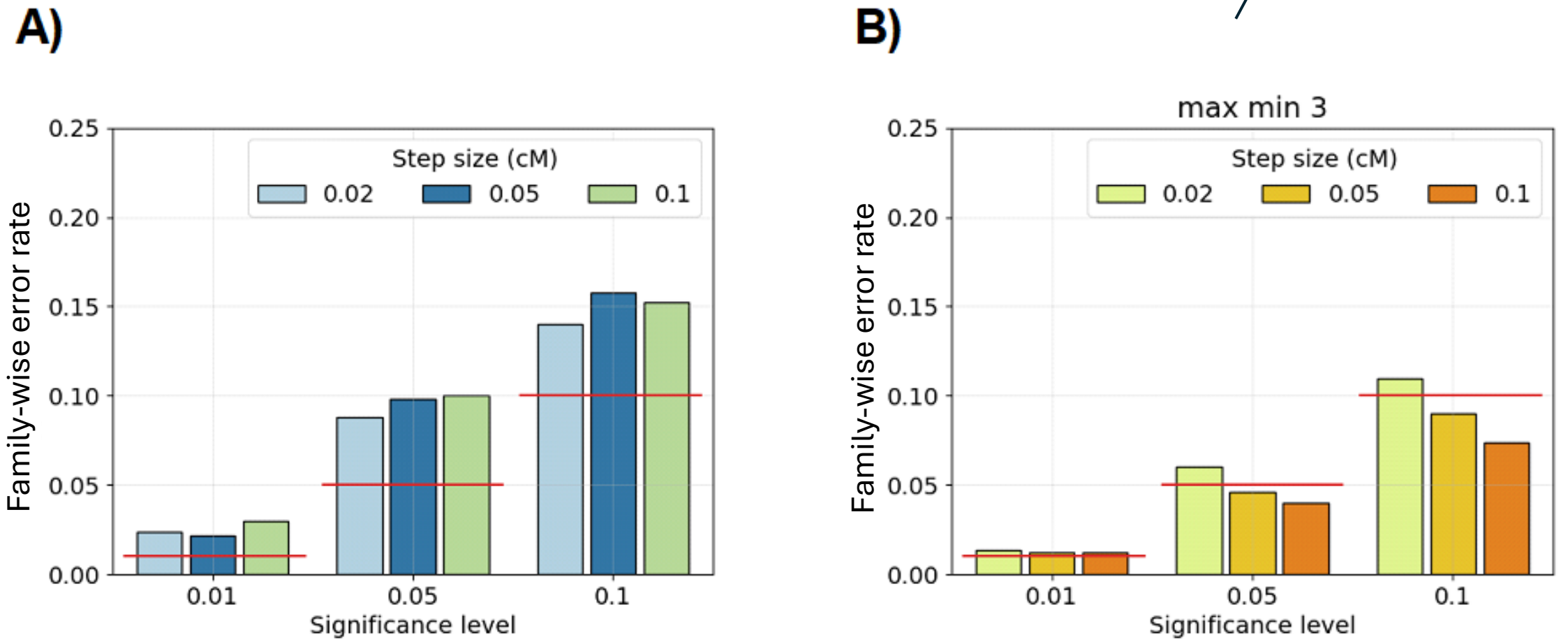
Significance levels when multiple testing

Nominal	Genome-wide (this work)	Bonferroni
0.01	1.08e-6	2.08e-7
0.05	**6.24e-6**	**1.04e-6**
0.10	1.36e-5	2.08e-6

- Our method is designed to control FWER
- Bonferroni method may be very conservative

Anti-conservative scan

Compute min IBD rate over windows size 3



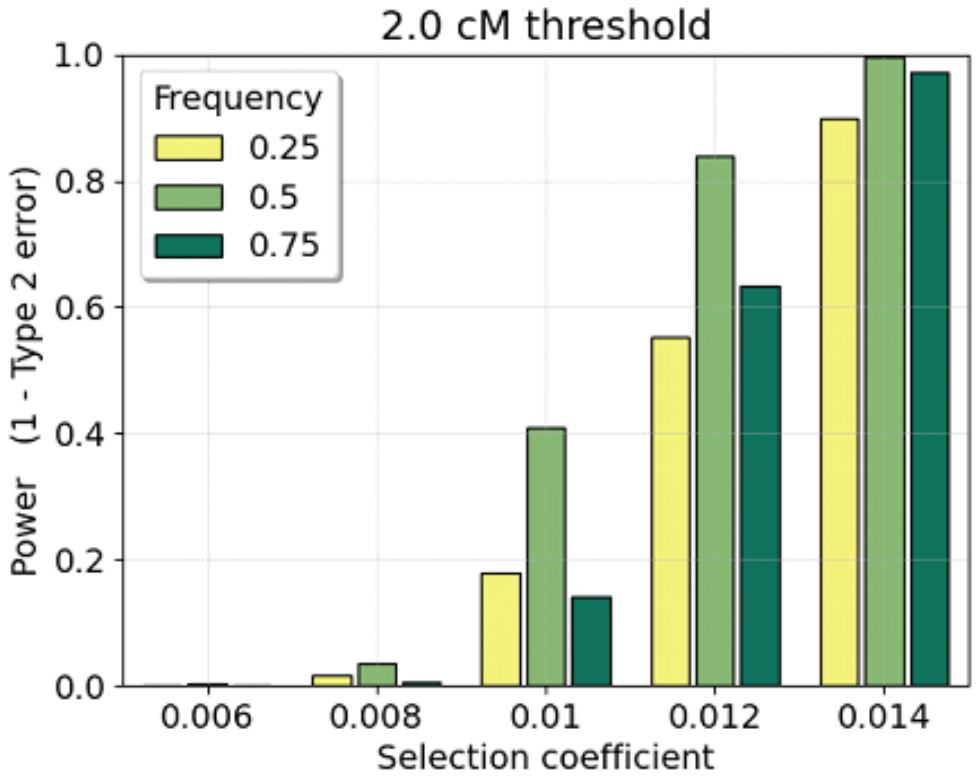
"Scan" is multiple hypothesis tests along genome

We appropriately handle the step size Δ

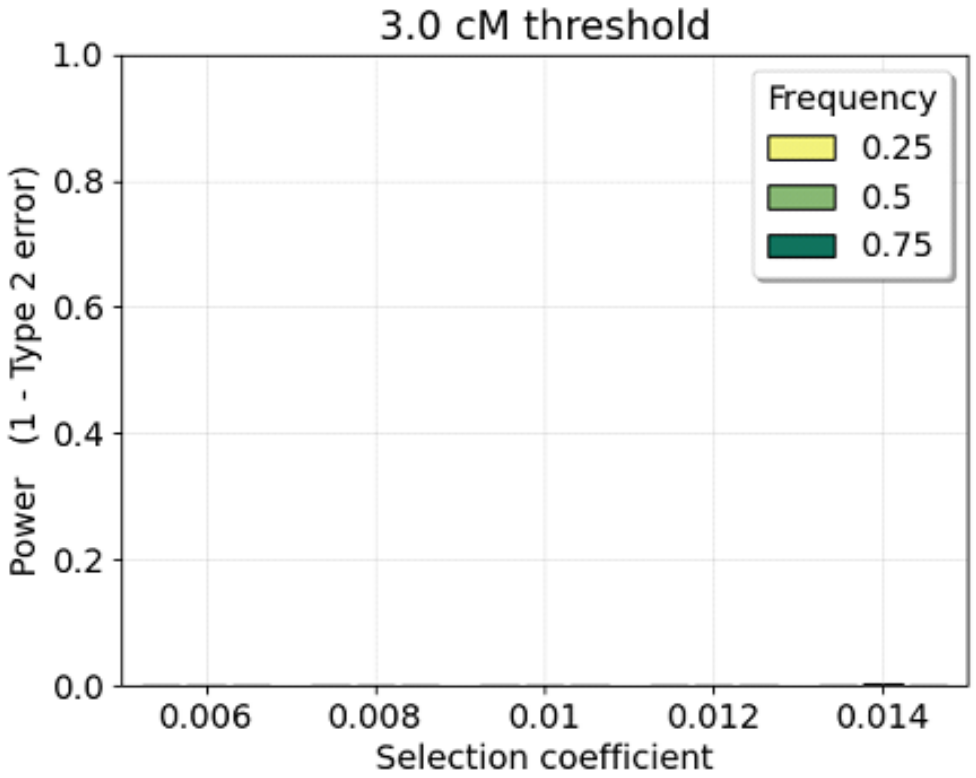
- Precision in locating selected locus
  - Palamara et al. (2018) use ad-hoc 0.05 cM spacing
  - Chen et al. (2023) use ad-hoc 1.0 cM spacing
- Adjust for genome size
  - 5e-8 significance level based on humans
- Choice impacts number of tests / threshold

Power simulations

A)



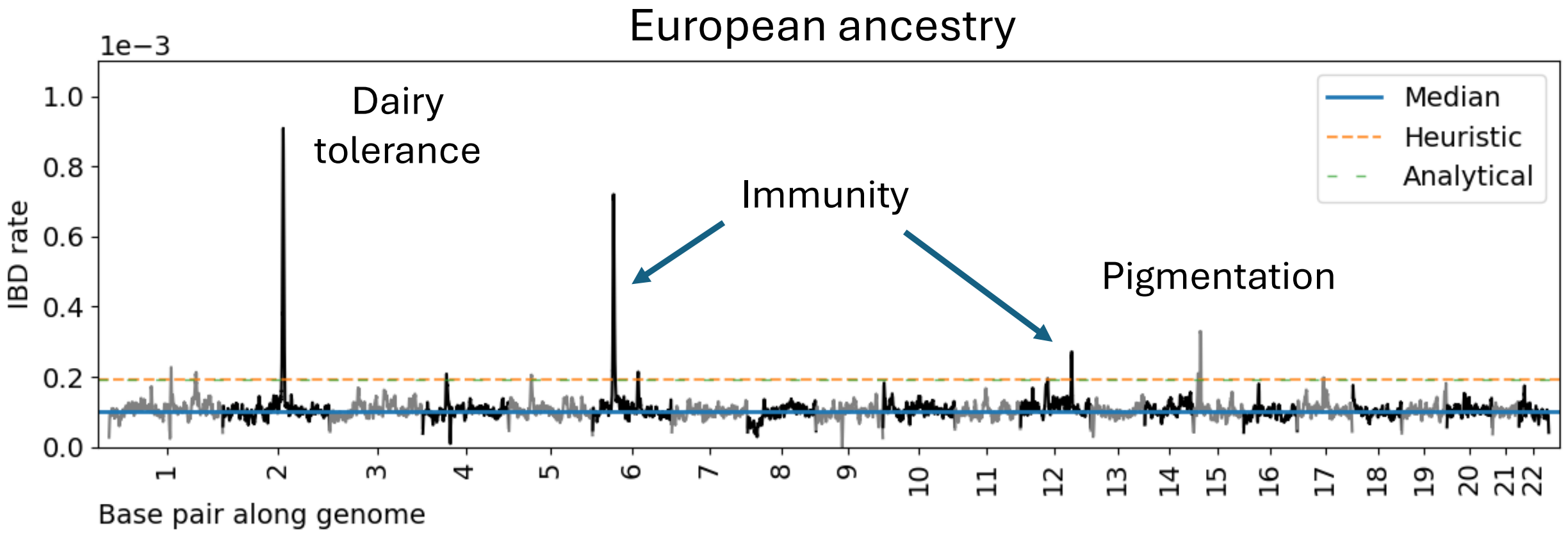
B)



Selection scan in three different ancestry groups

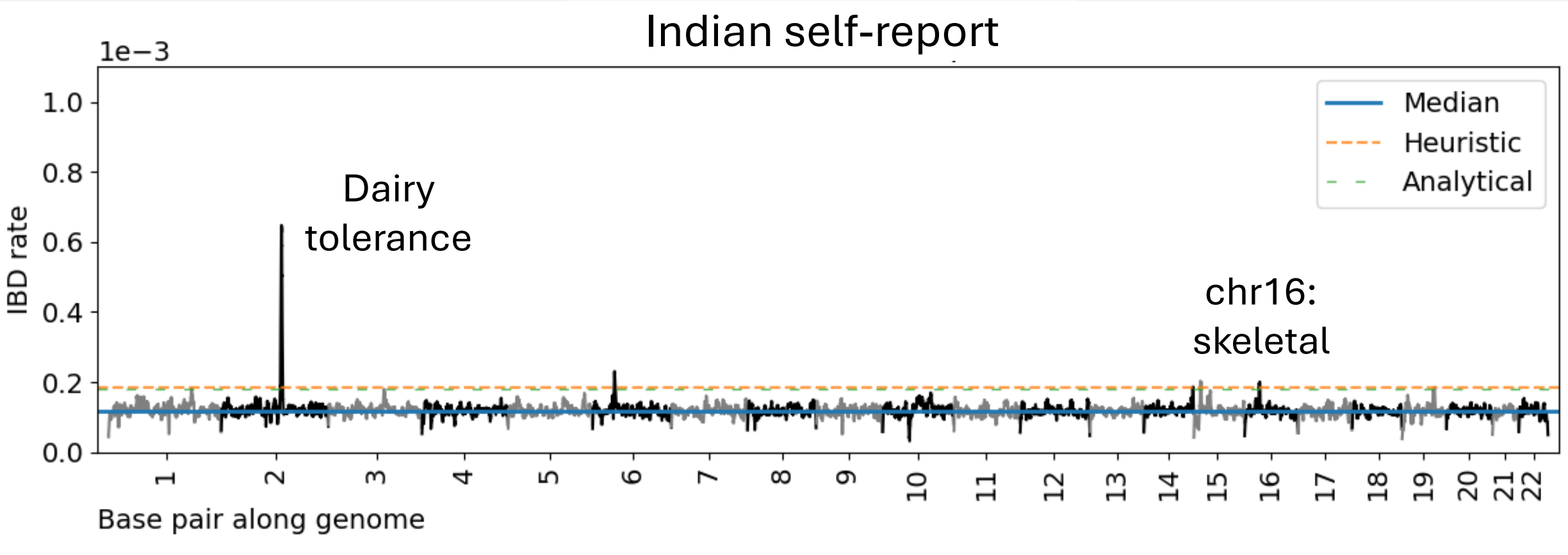
European, South Asian, African

Analytical: our multiple testing method
Heuristic: 4σ above the median
Thresholds on top of each other



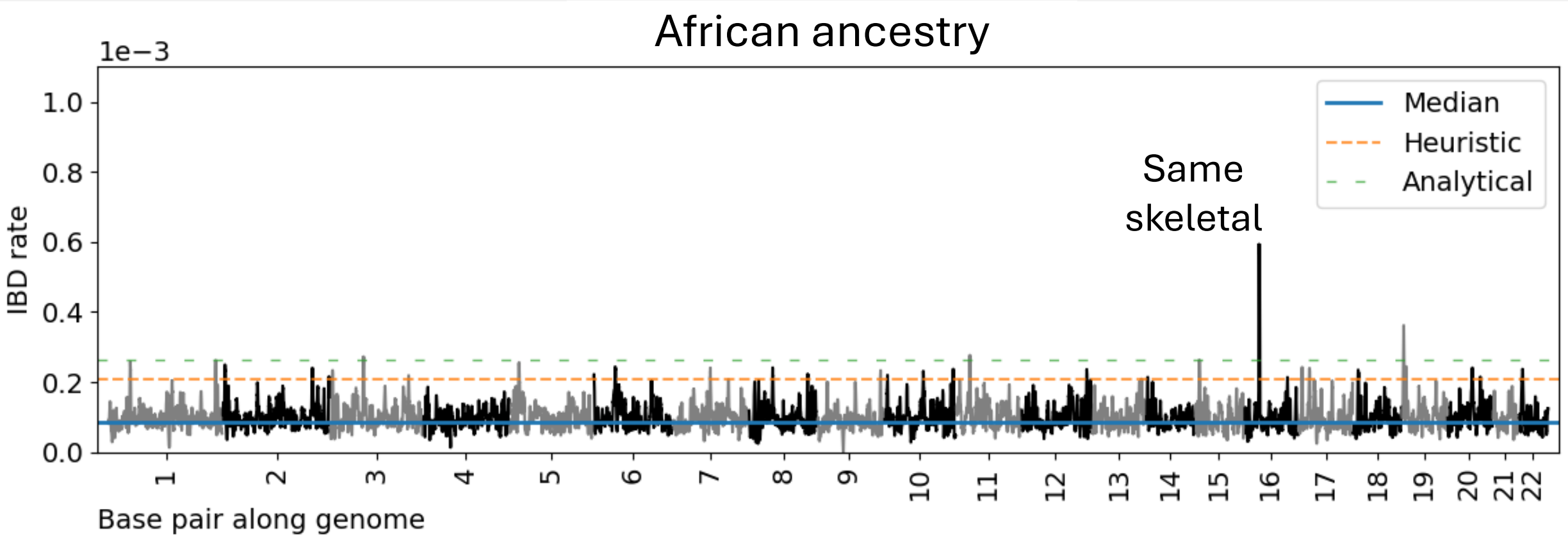
Replicated in UKBB White British samples
Replicated with different sample subsets

Analytical: our multiple testing method
Heuristic: 4σ above the median
Thresholds on top of each other



LCT result is consistent w/ existing literature

Analytical: our multiple testing method
Heuristic: 4σ above the median
Thresholds **are not** on top of each other



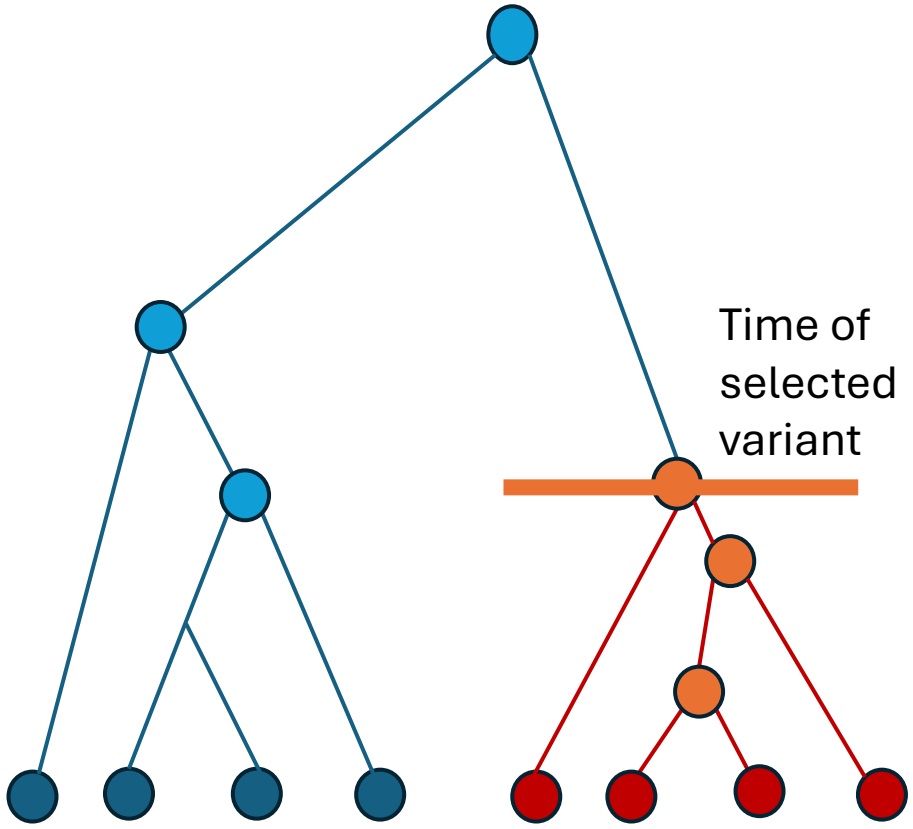
Replicated in UKBB Black British samples

Modeling selective sweeps using IBD segments

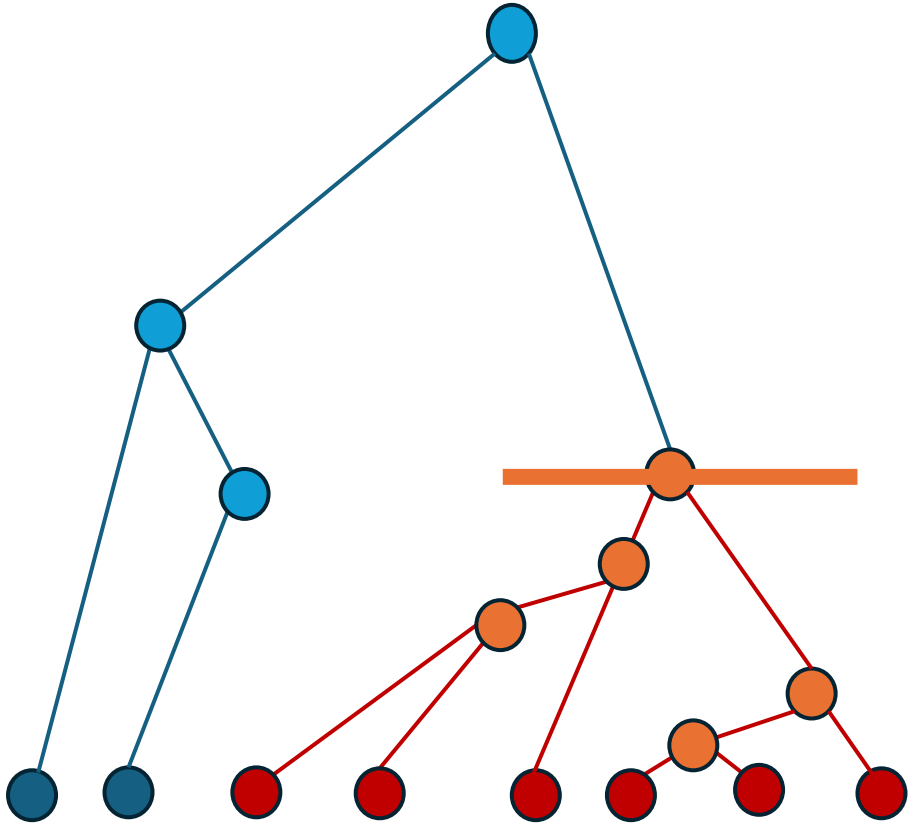
Temple, Waples, and Browning (2023)
To appear in *American Journal of Human Genetics*

Recall

Weaker selection coefficient s



Stronger selection coefficient s



Current sample

Common ancestor

Selected allele

Updates

Selection coefficient estimator

- Compared favorably to other methods (tree inference, NNs)
- More empirical coverage simulations
- Simulation study indicating that our estimator is sufficient
- Robustness, interpretation under time-varying selection

Conclusion: an entire workflow

1. Detect selection : excess IBD rate
2. Estimate freq., location of selection
3. Estimate selection coefficient
4. Confidence intervals for $\hat{s}$
    - Our central limit + Delta method

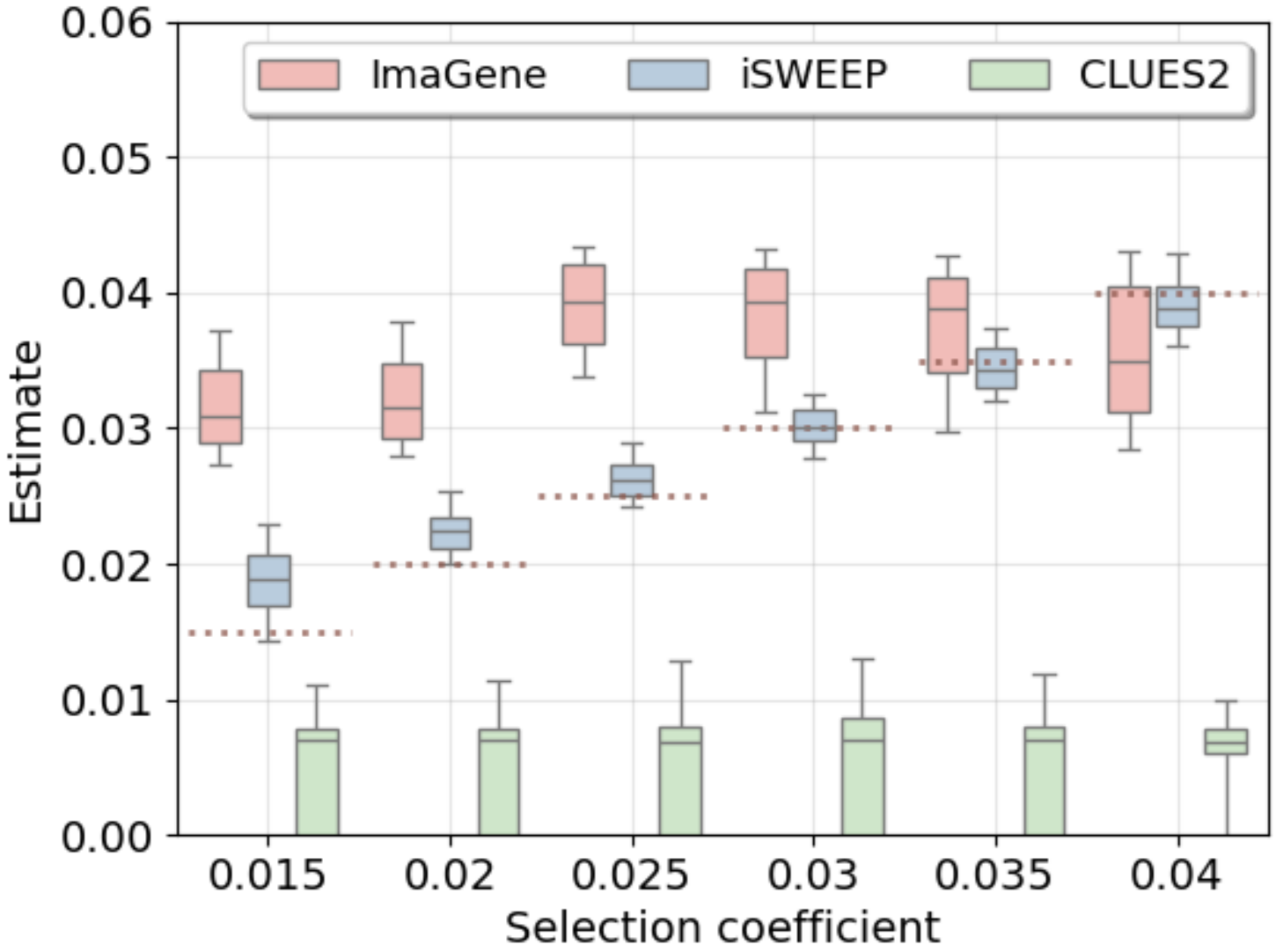
Acknowledgments

- Committee: Sharon, Elizabeth, Kelley, Amy
- Co-authors: Sharon, **Ryan**, Elizabeth
- Lab members: **Ruoyi**, Nobu, Robert
- Other StatGen: Elizabeth Blue, Ellen Wijsman, Tim Thornton
- Other STAT: Ellen, Daniela, Ema, office staff

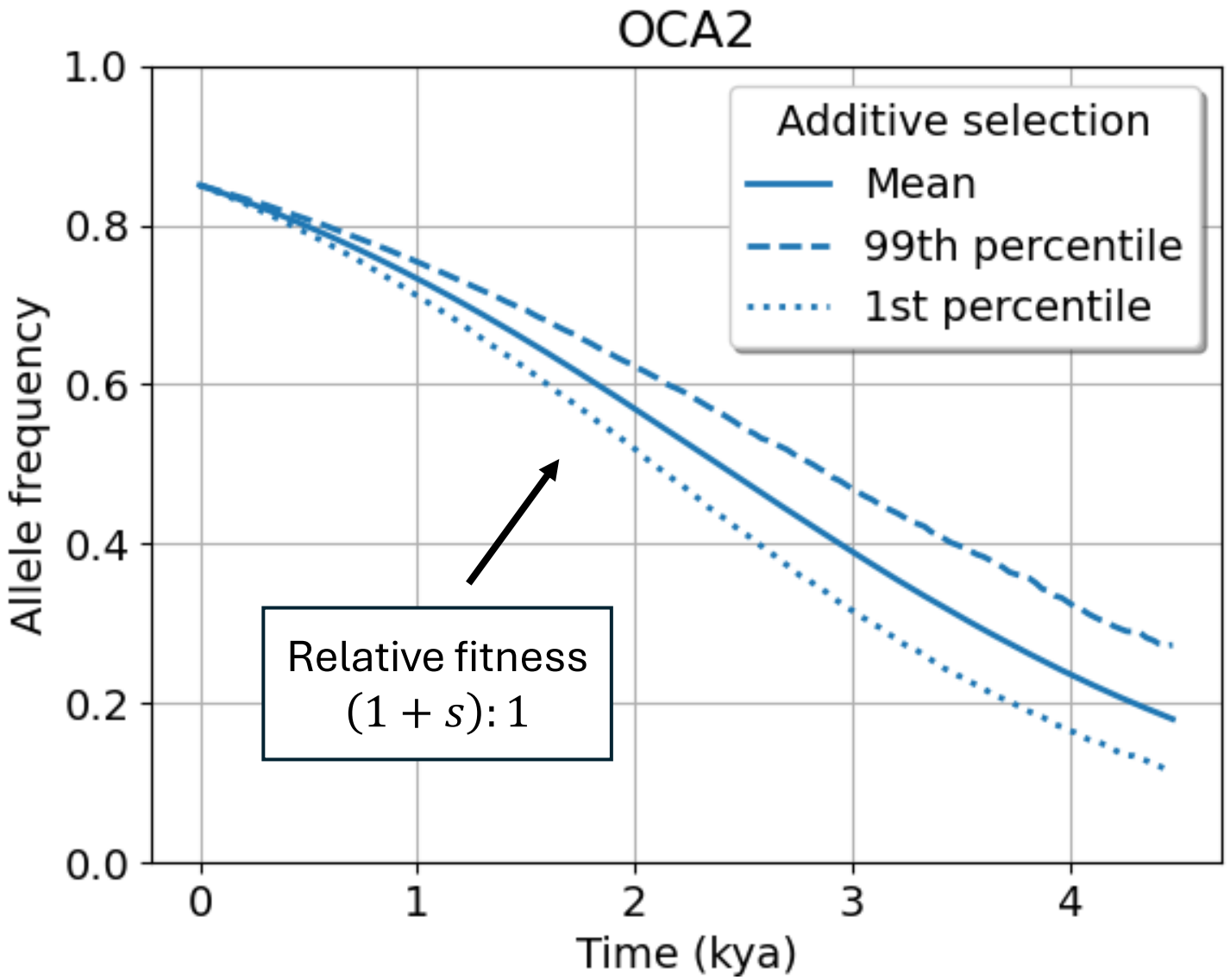
Supplementary Material

Main figures in *AJHG* paper

Comparing other methods



Real data



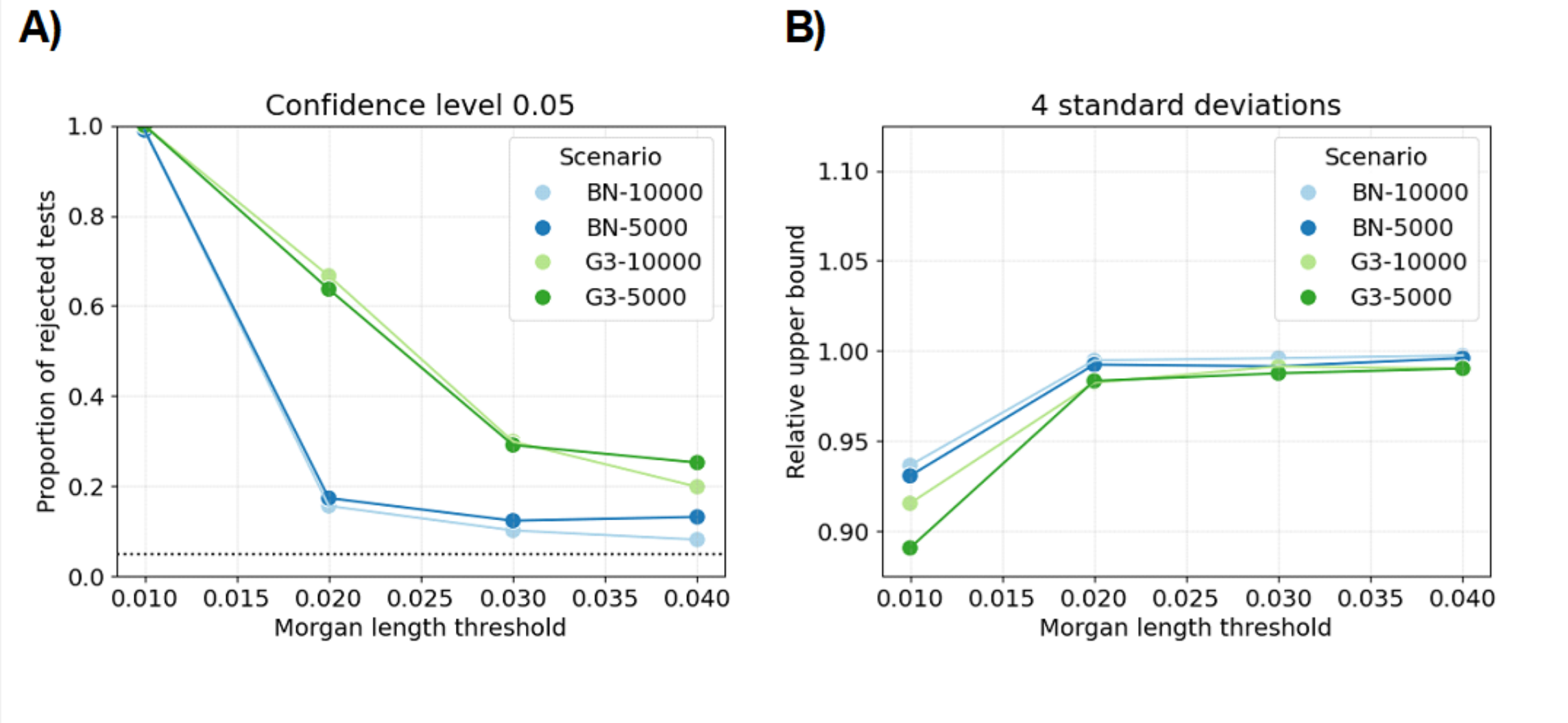
Present ← Past

Appendix:
Central limit theorem

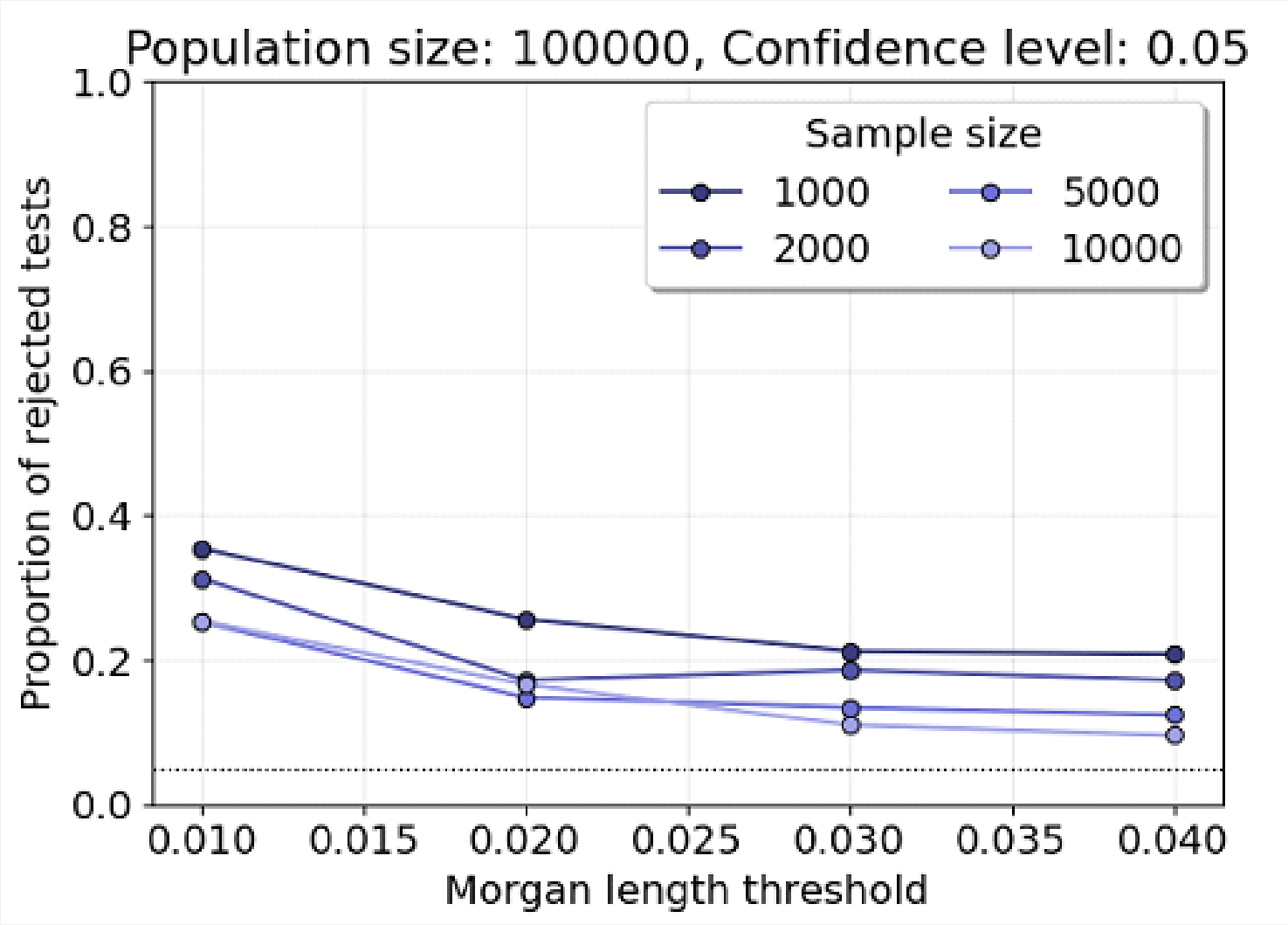
Theorem 3.1. *The mean-centered and suitably scaled IBD rate statistic* $\bar{\mathbf{Z}}_{\binom{n}{2},N}$ *converges in distribution to the standard normal distribution for* n *and* Nw *tending to infinity when the following are true:*

1. $Nw = o(n^2)$, *scaled population size is small relative to the number of pairs;*
2. $n = o(Nw)$, *sample size is small relative to scaled population size;*
3. $\mathbb{E}[Z_{a,b} \times \mathbf{Z}_{-a,b} | \mathbf{Z}_{-a,b}] \geq 0$ *for all* $\mathbf{Z}_{-a,b}$.

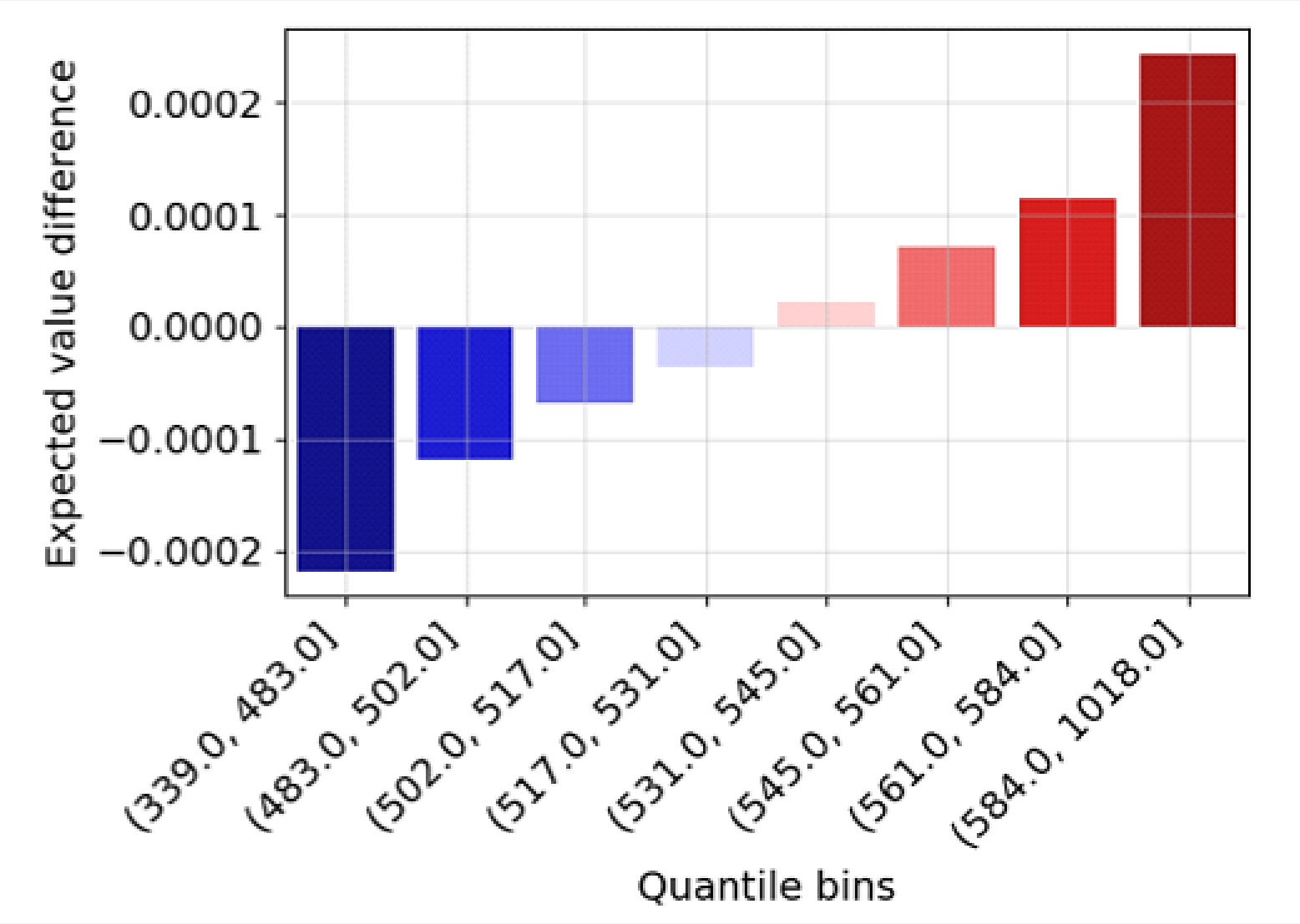
Demography CLT



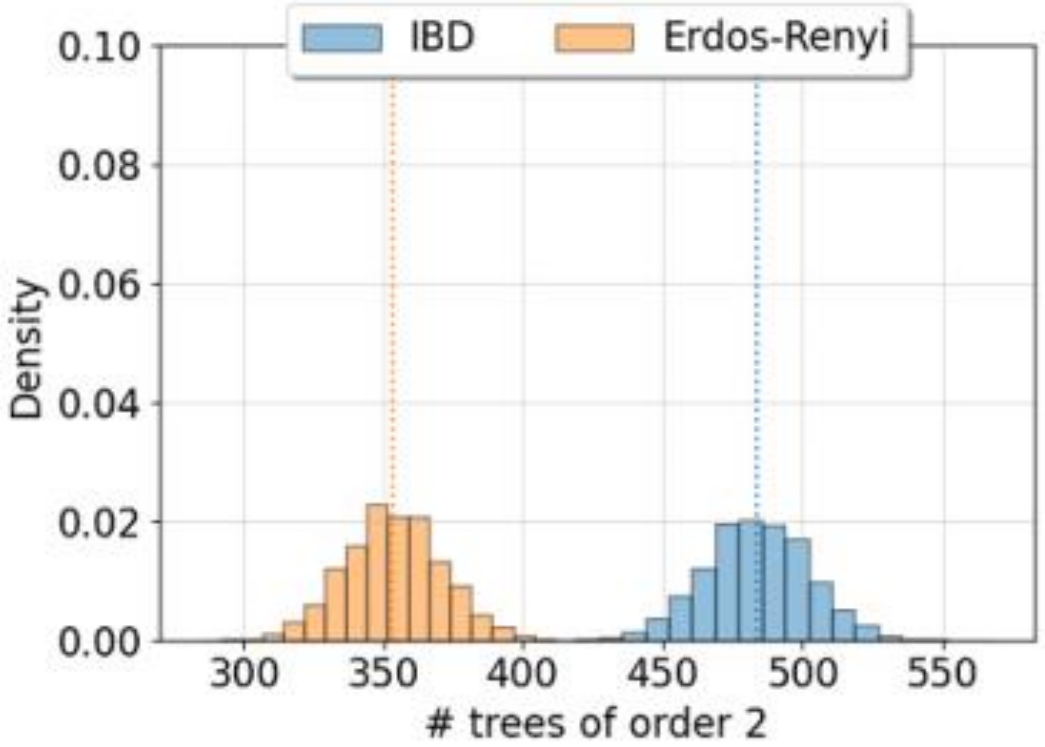
CLT: sample size is less important

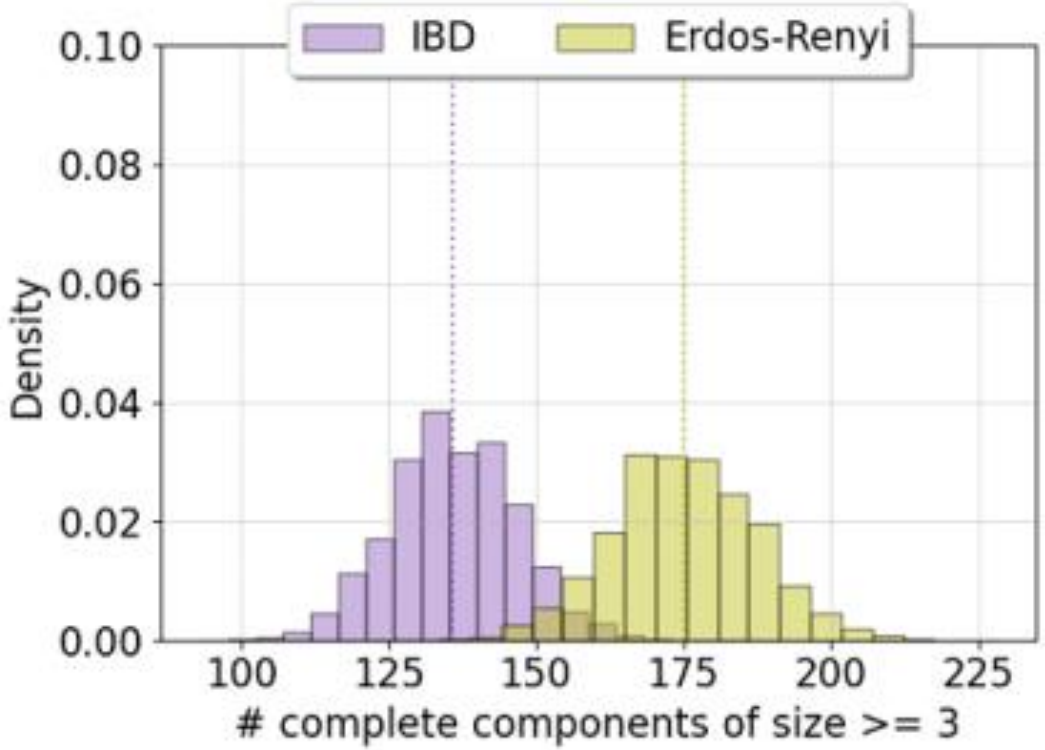


Third condition (Monte Carlo)



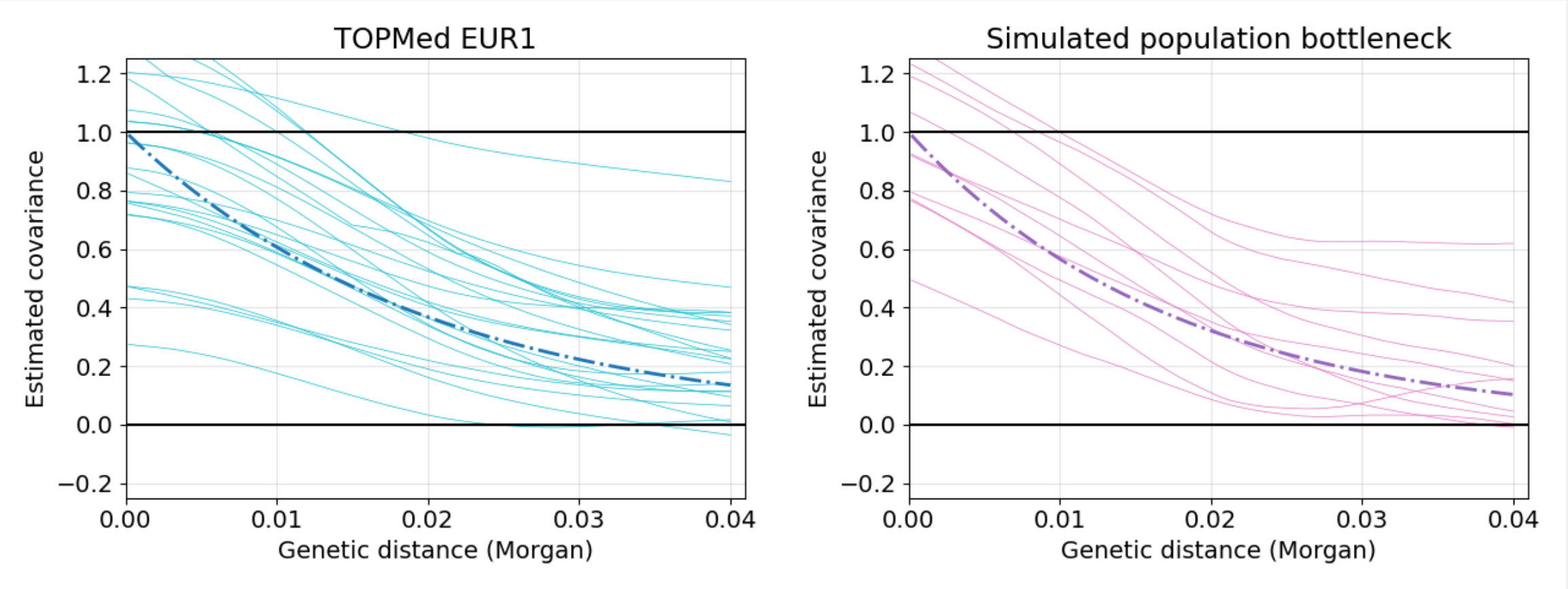
Features of IBD networks



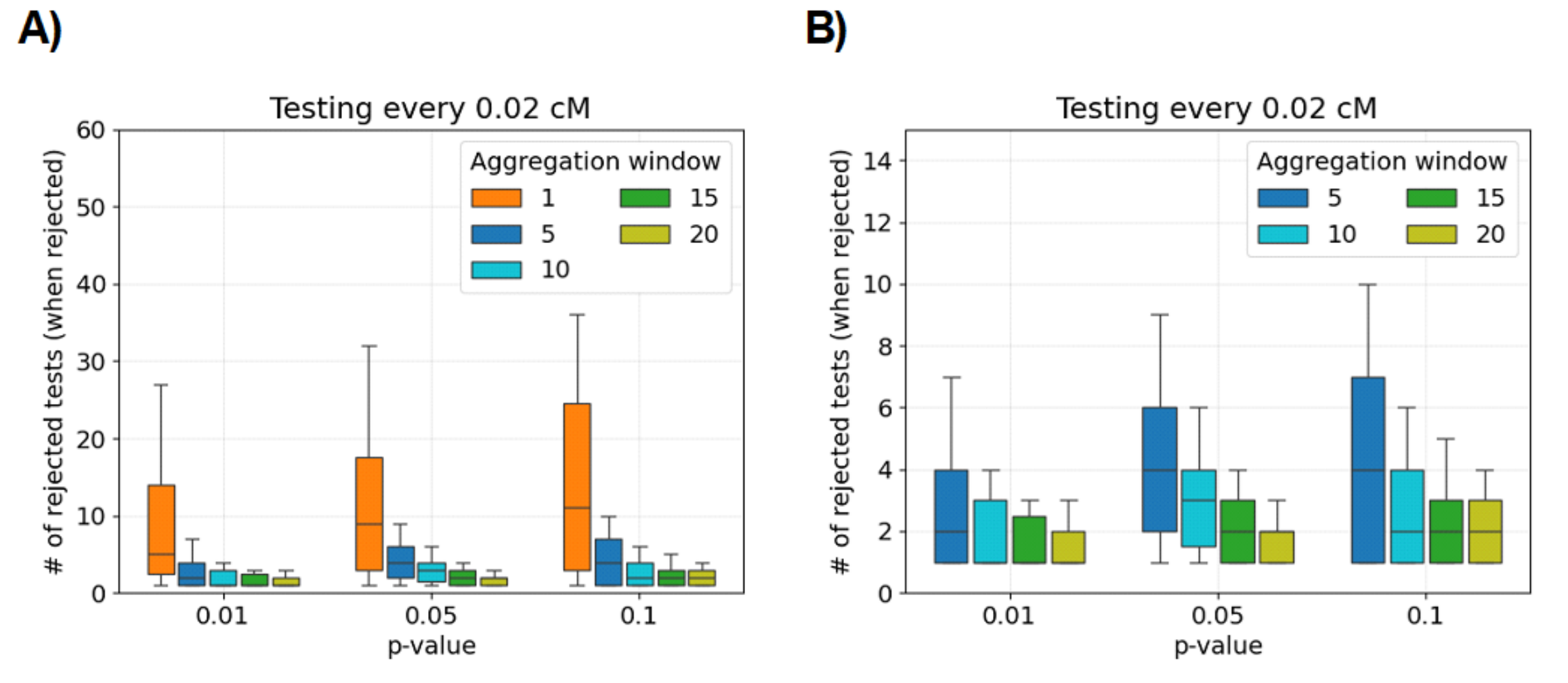


Appendix: Multiple testing

Estimating θ



Rejections are next to each other



Siegmund and Yakir (2007) approximation

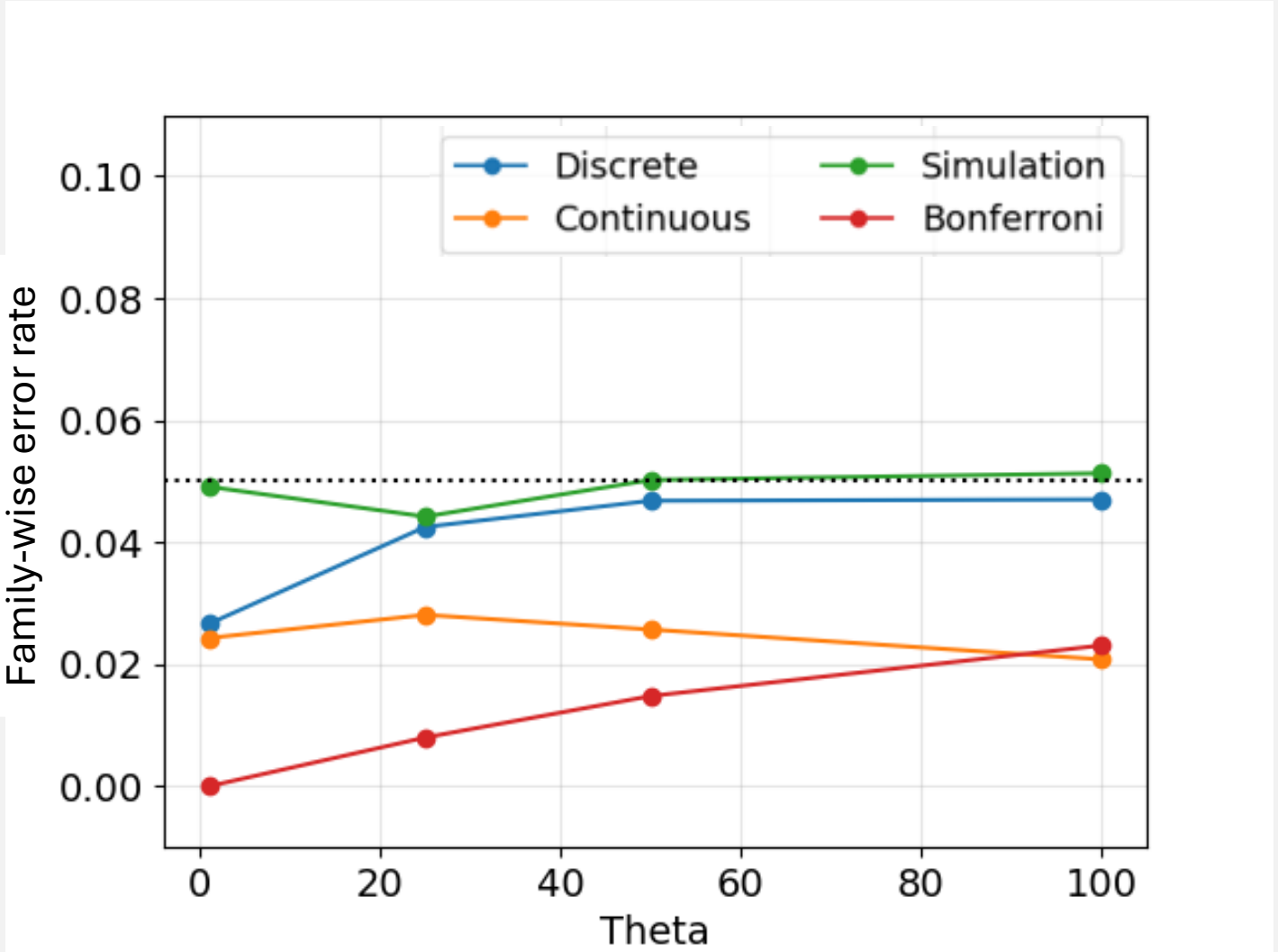
$$P(\max_{1\le m\le M} \tilde{\mathbf{Z}}_m \ge z) \approx 1 - \exp(-C[1-\Phi(z)] - \theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{2\theta\Delta\}^{1/2})),$$

Independent tests

First Markov hitting time

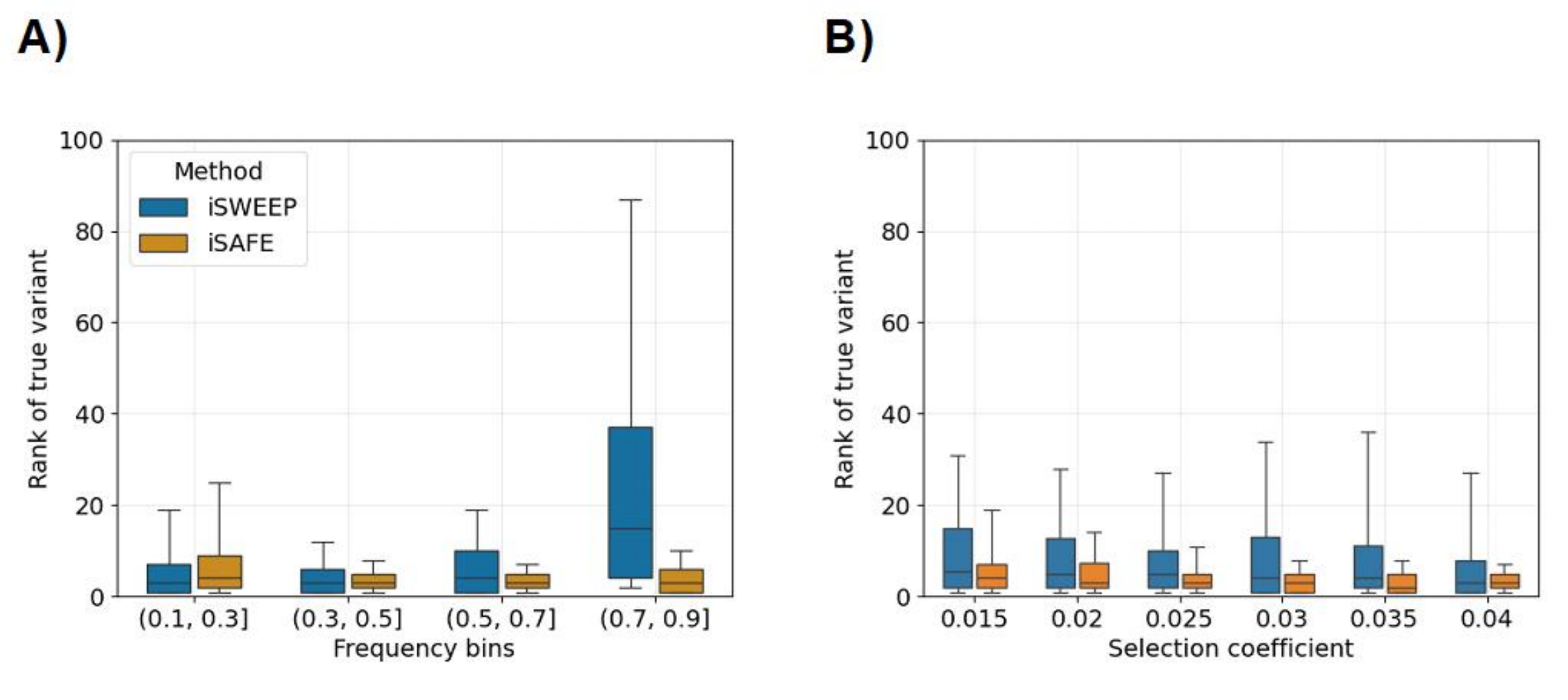
- C : number of chromosomes
- L : total length of genome
- Δ : spacings
- θ : exponential decay

Simulating true OU process

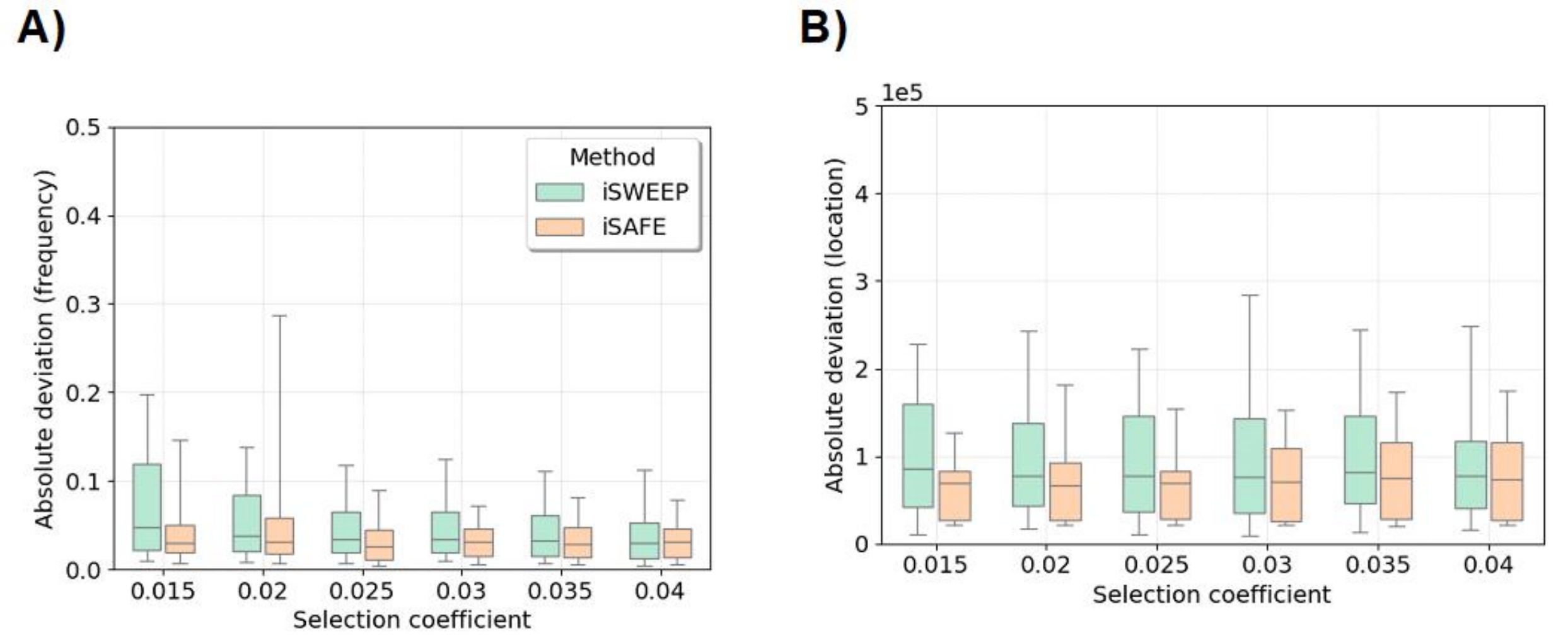


Appendix:
Methods for genetic data

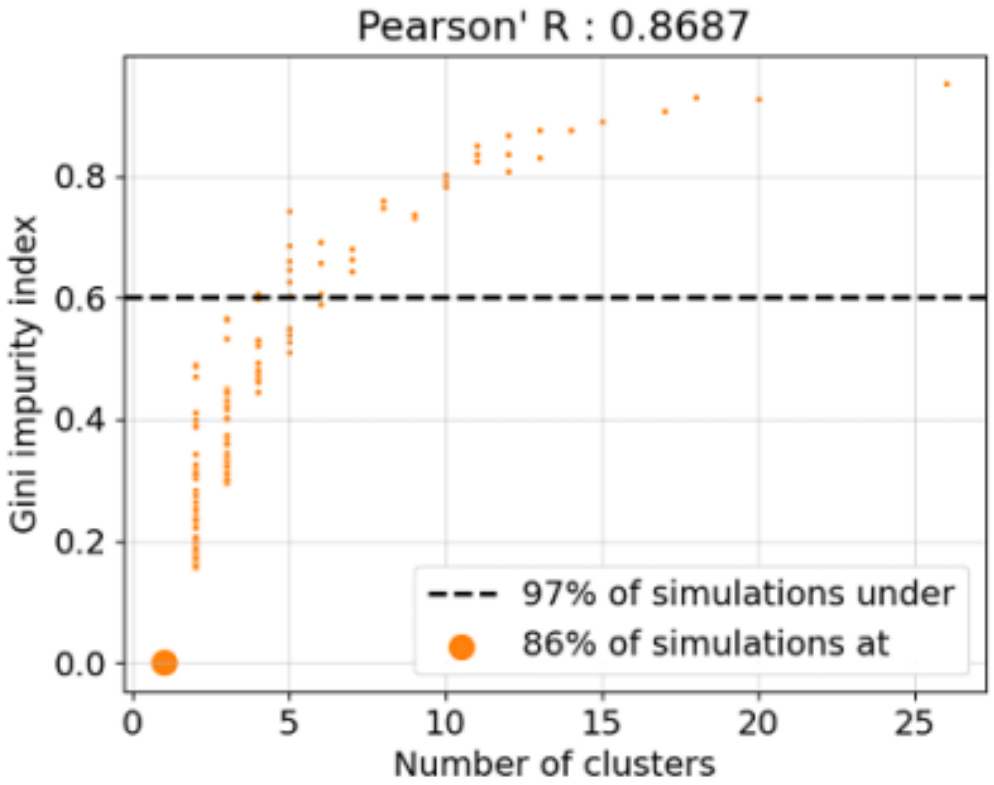
Ranking

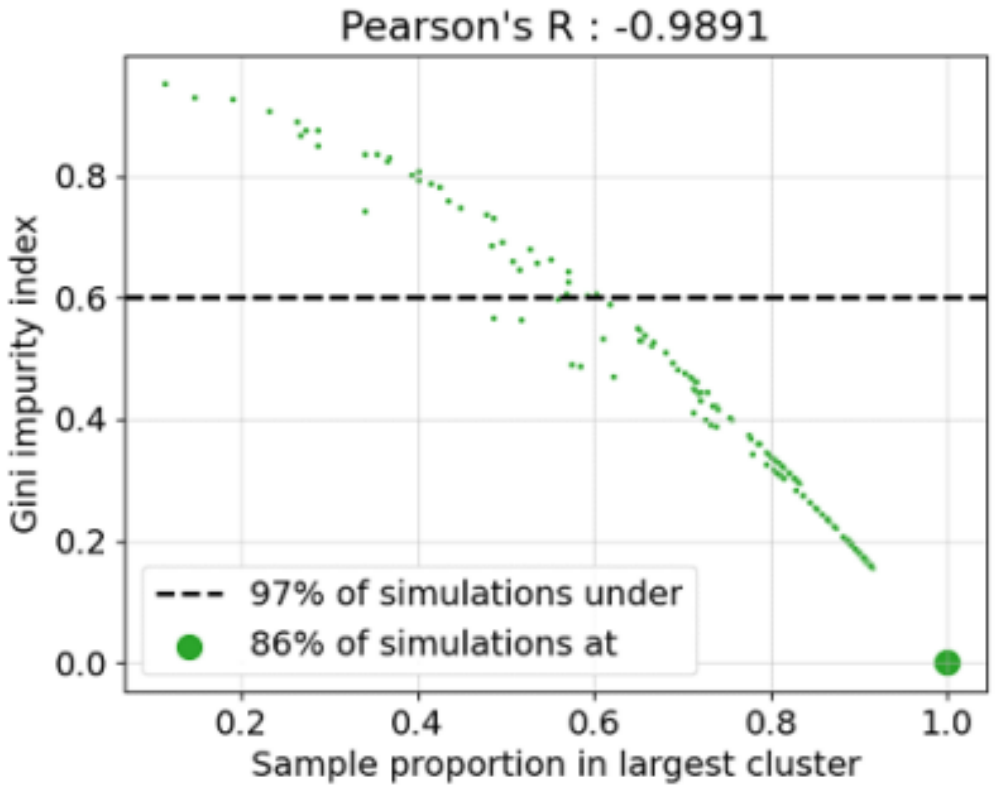

A)
Method
iSWEEP
iSAFE
Rank of true variant
100
80
60
40
20
0
(0.1, 0.3]
(0.3, 0.5]
(0.5, 0.7]
(0.7, 0.9]
Frequency bins
B)
Rank of true variant
100
80
60
40
20
0
0.015
0.02
0.025
0.03
0.035
0.04
Selection coefficient

Frequency & Location

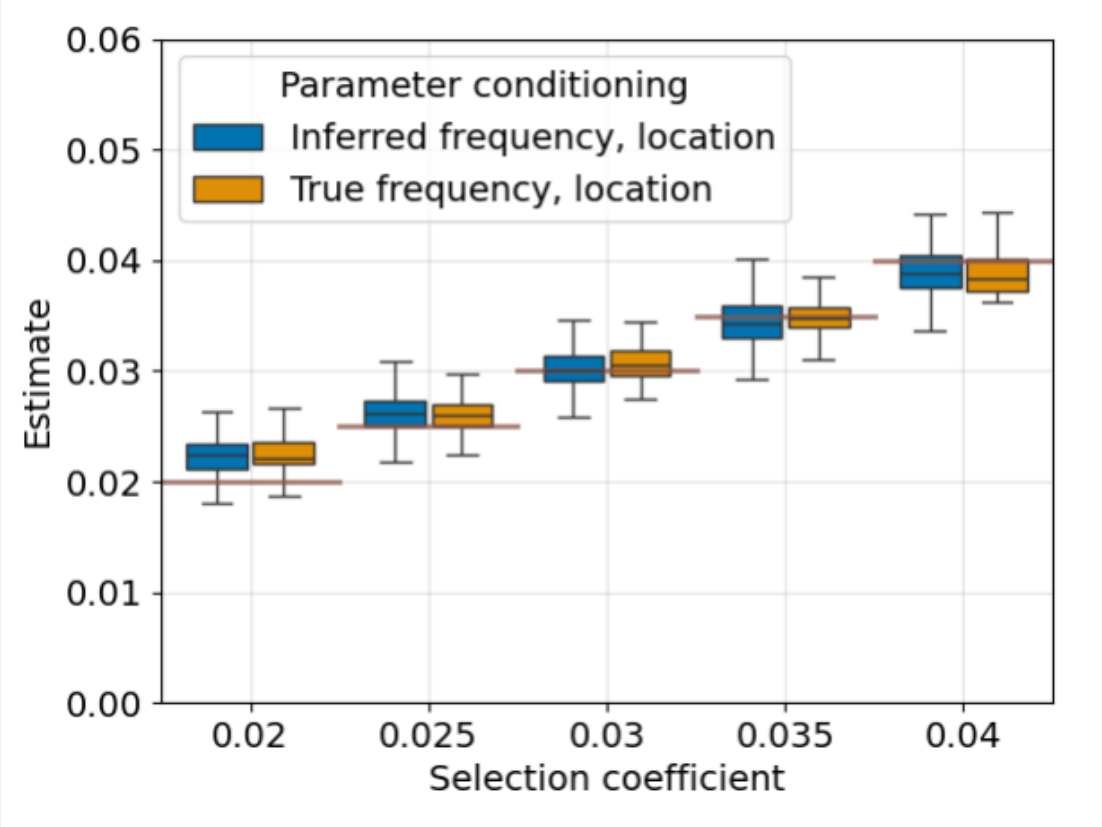


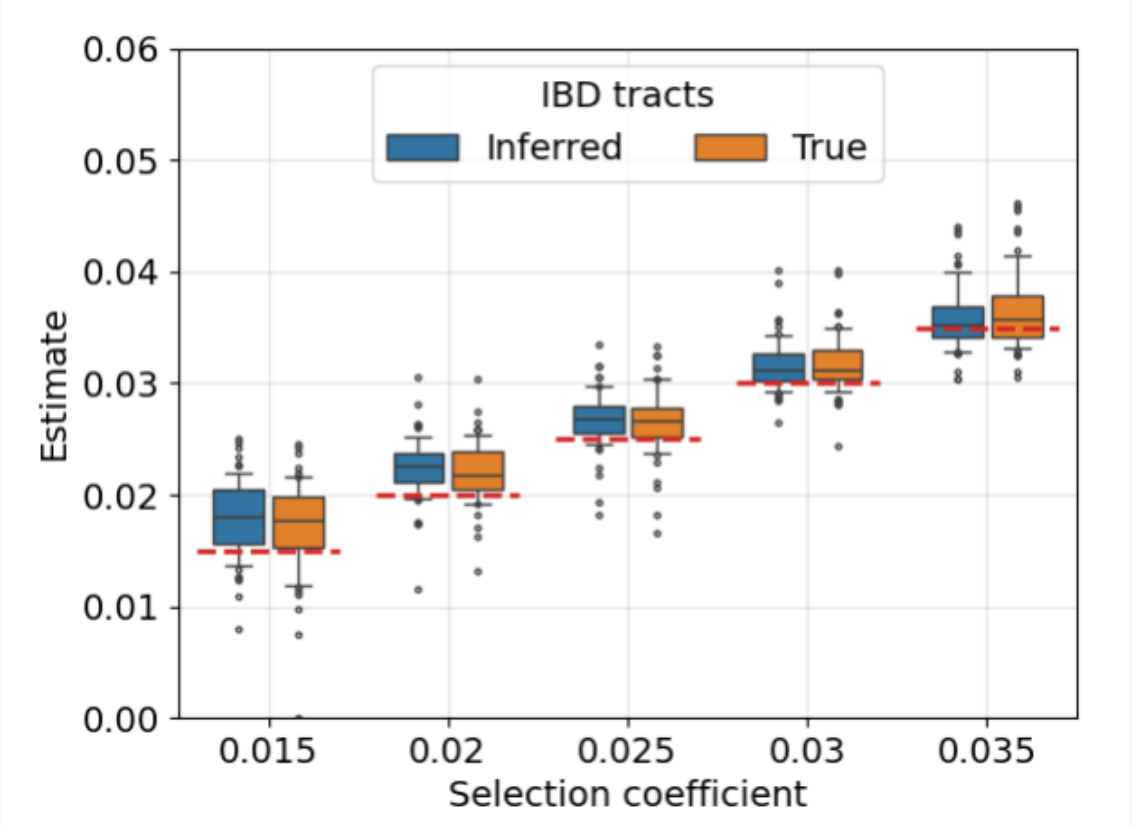
Diagnostic check for sweeps





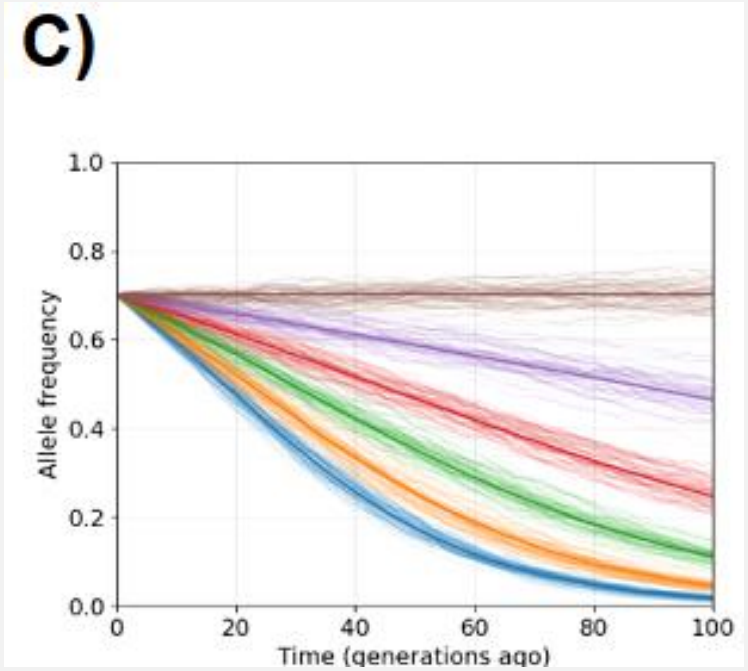
Robust to inferring IBD in data

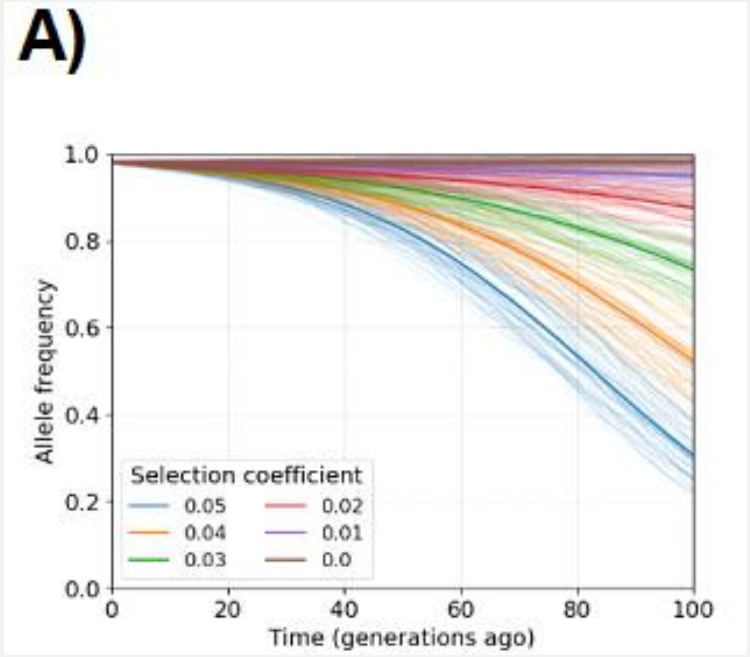


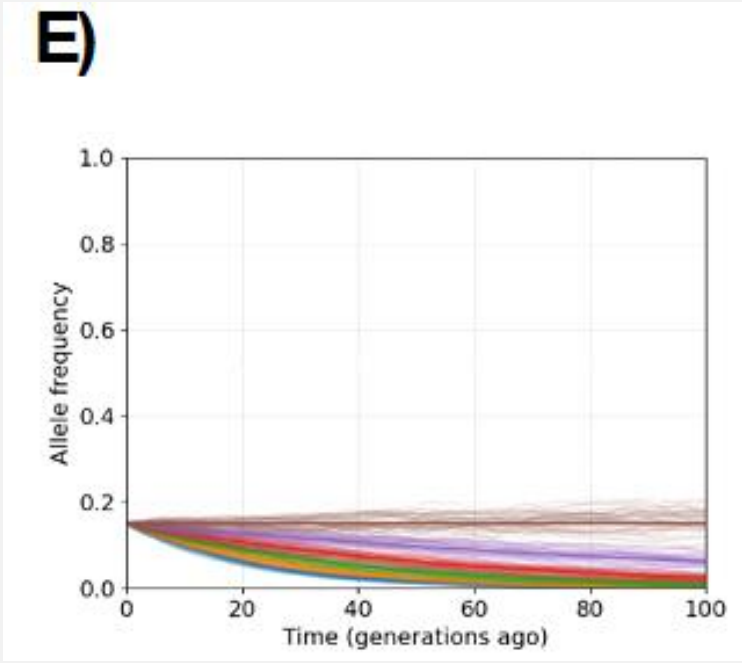


Appendix:
Selection coefficient

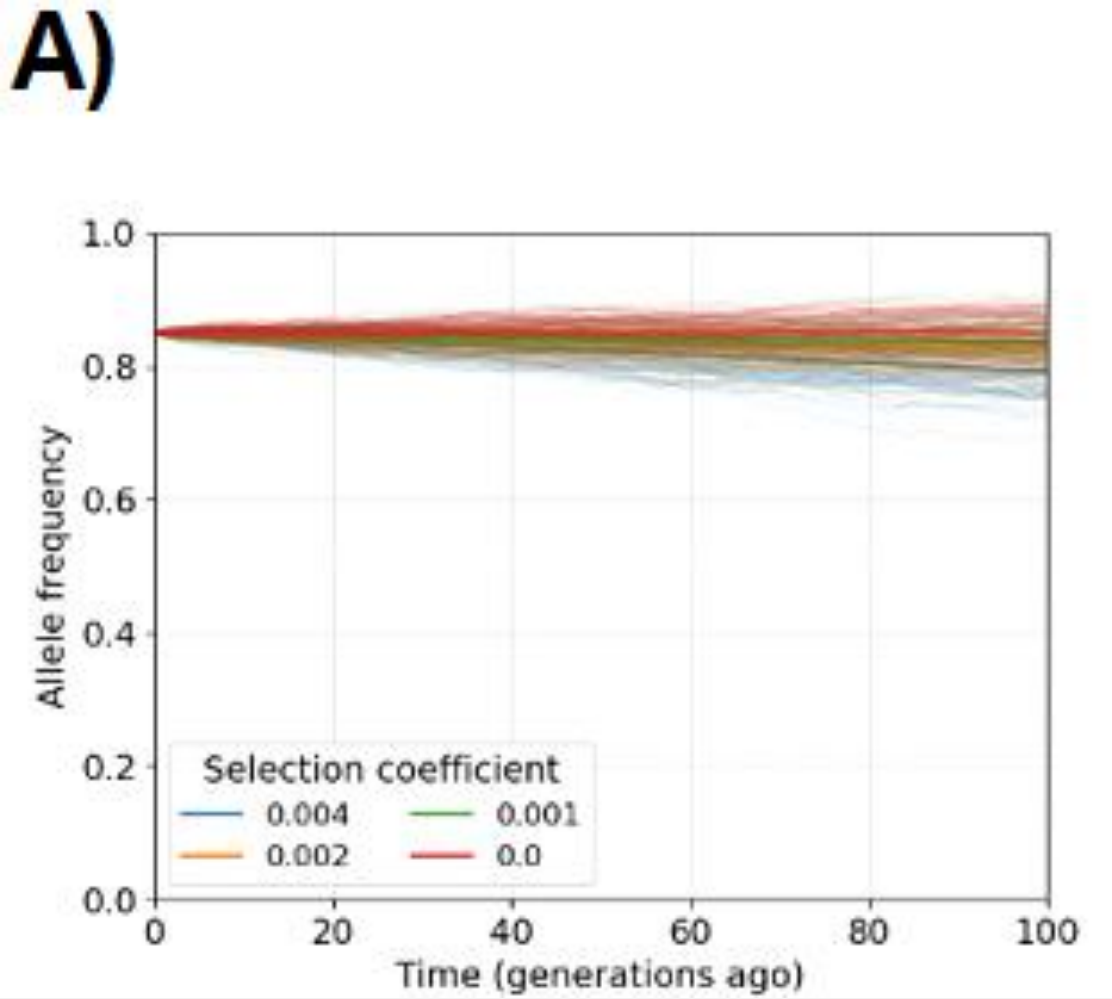
Strong selection

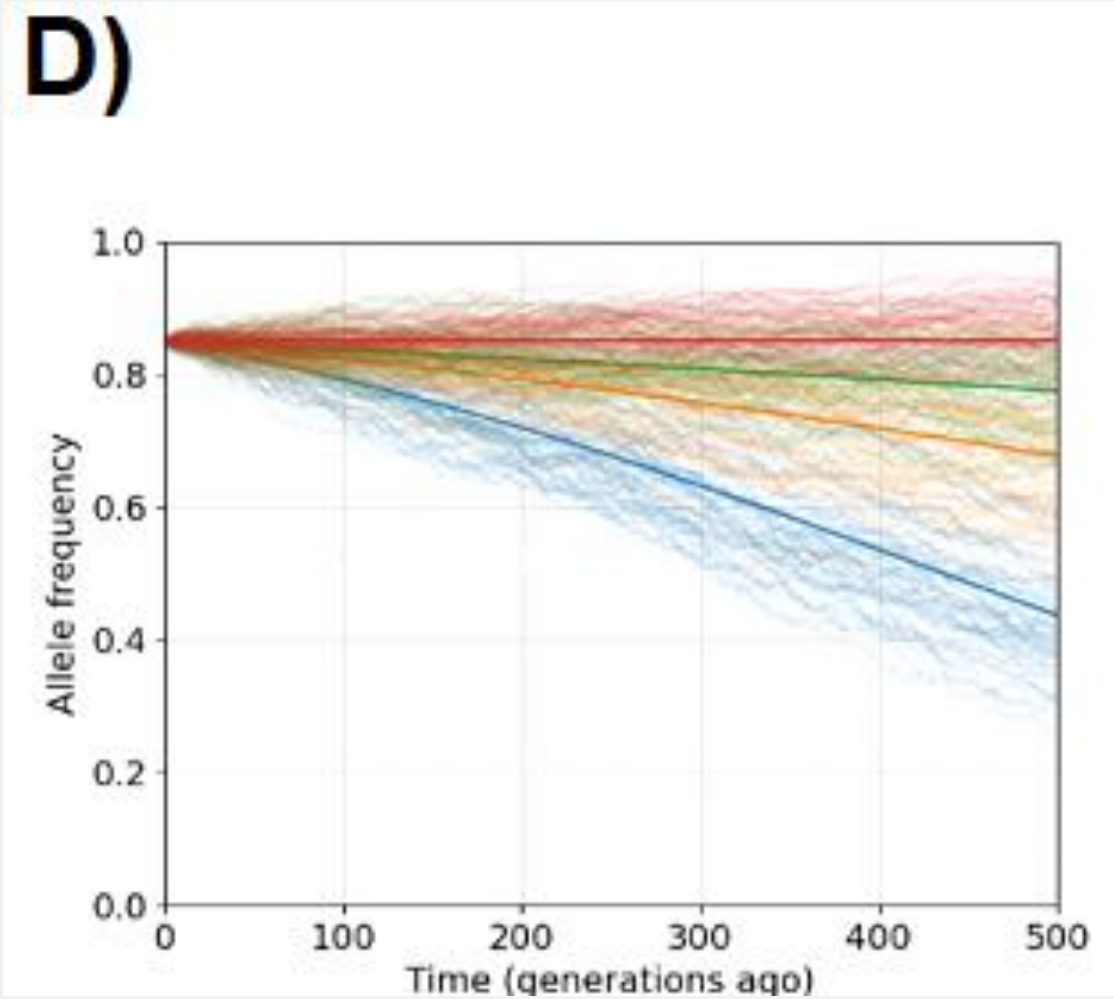


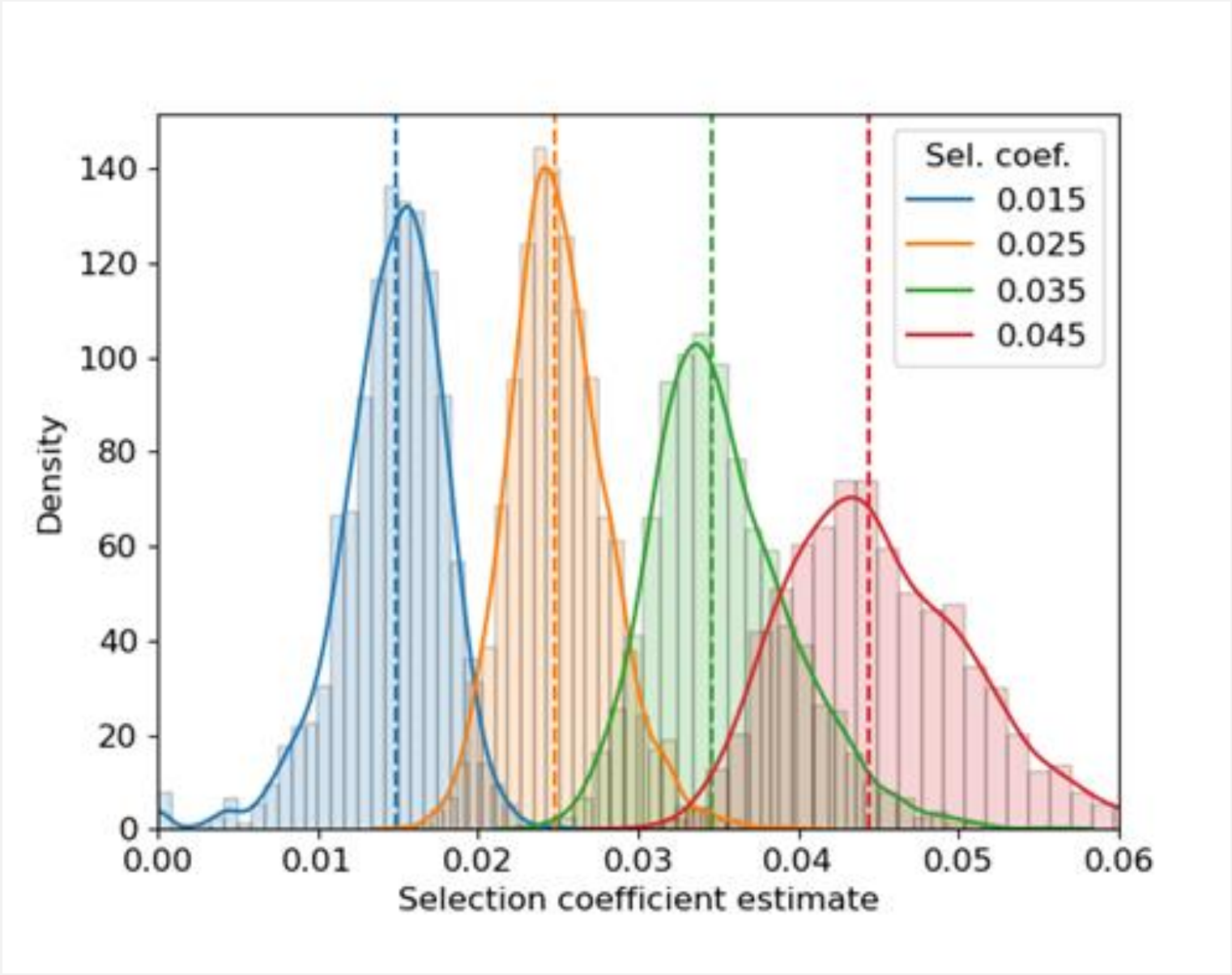


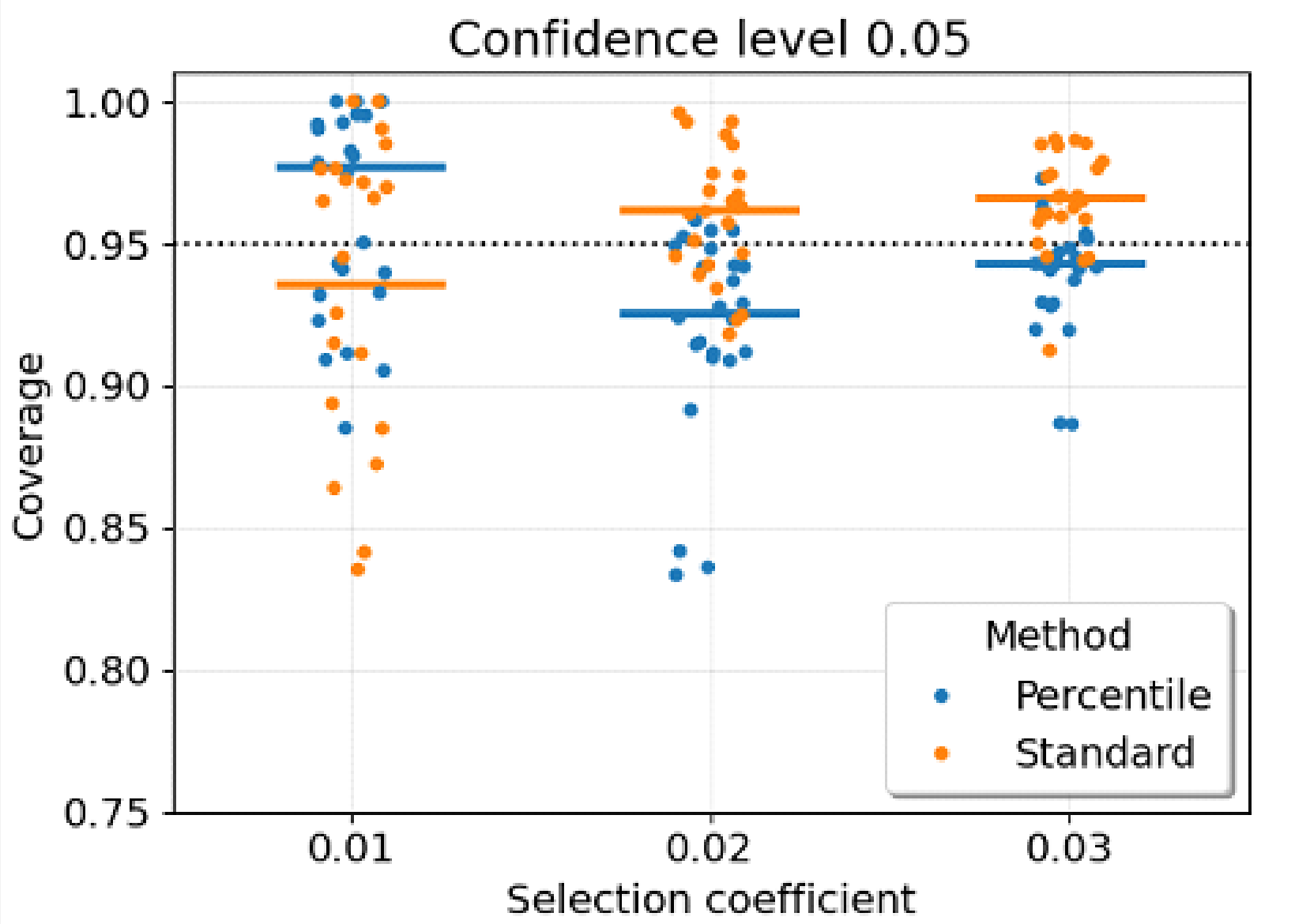


Weak selection

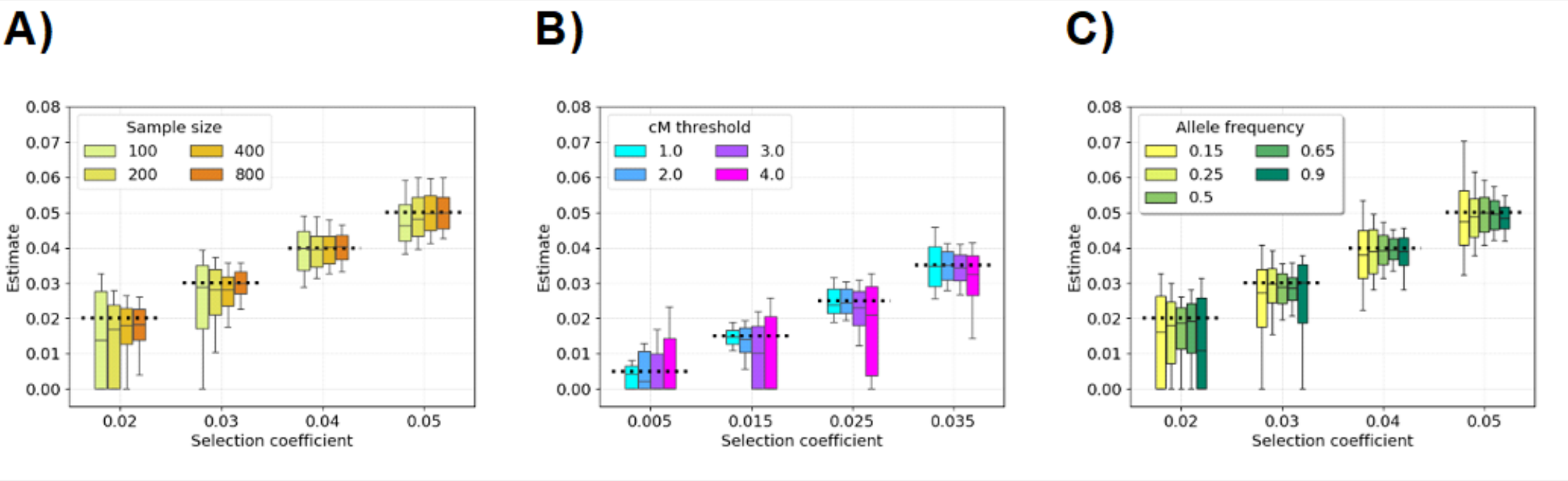


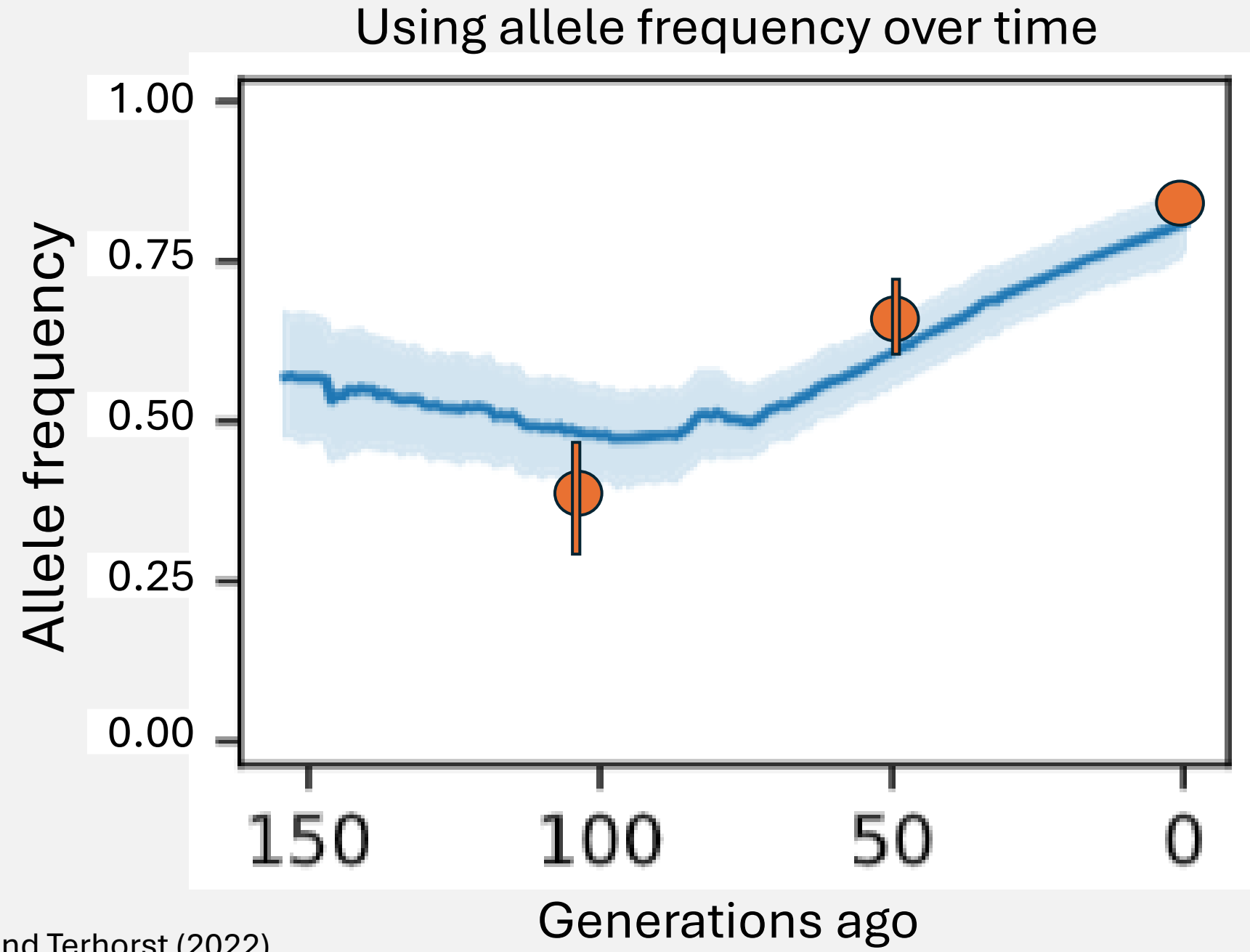



Sel. coef.
0.015
0.025
0.035
0.045
Density
0
20
40
60
80
100
120
140
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Selection coefficient estimate


Confidence level 0.05
1.00
0.95
0.90
0.85
0.80
0.75
Coverage
0.01
0.02
0.03
Selection coefficient
Method
Percentile
Standard

Experimental design

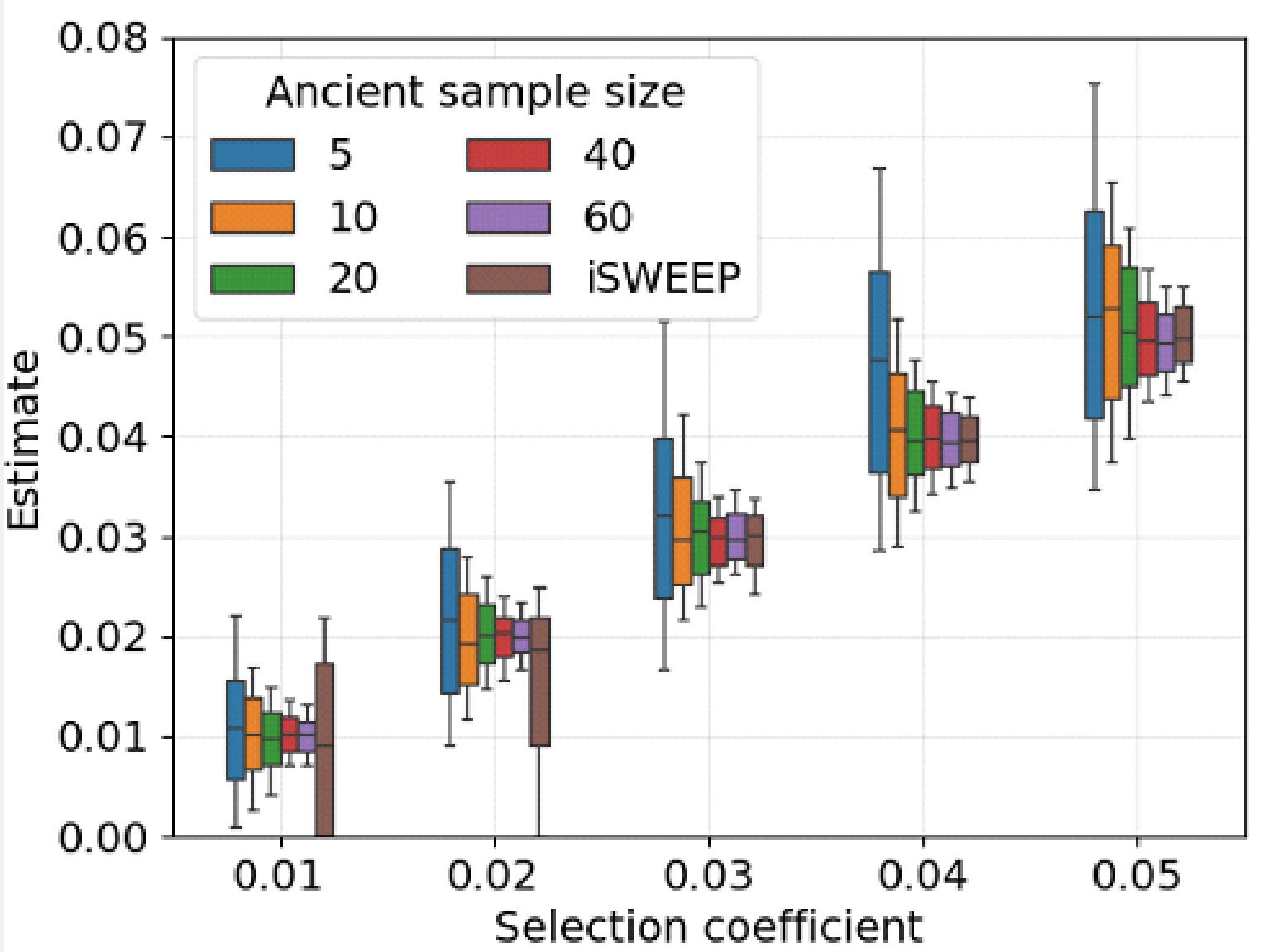




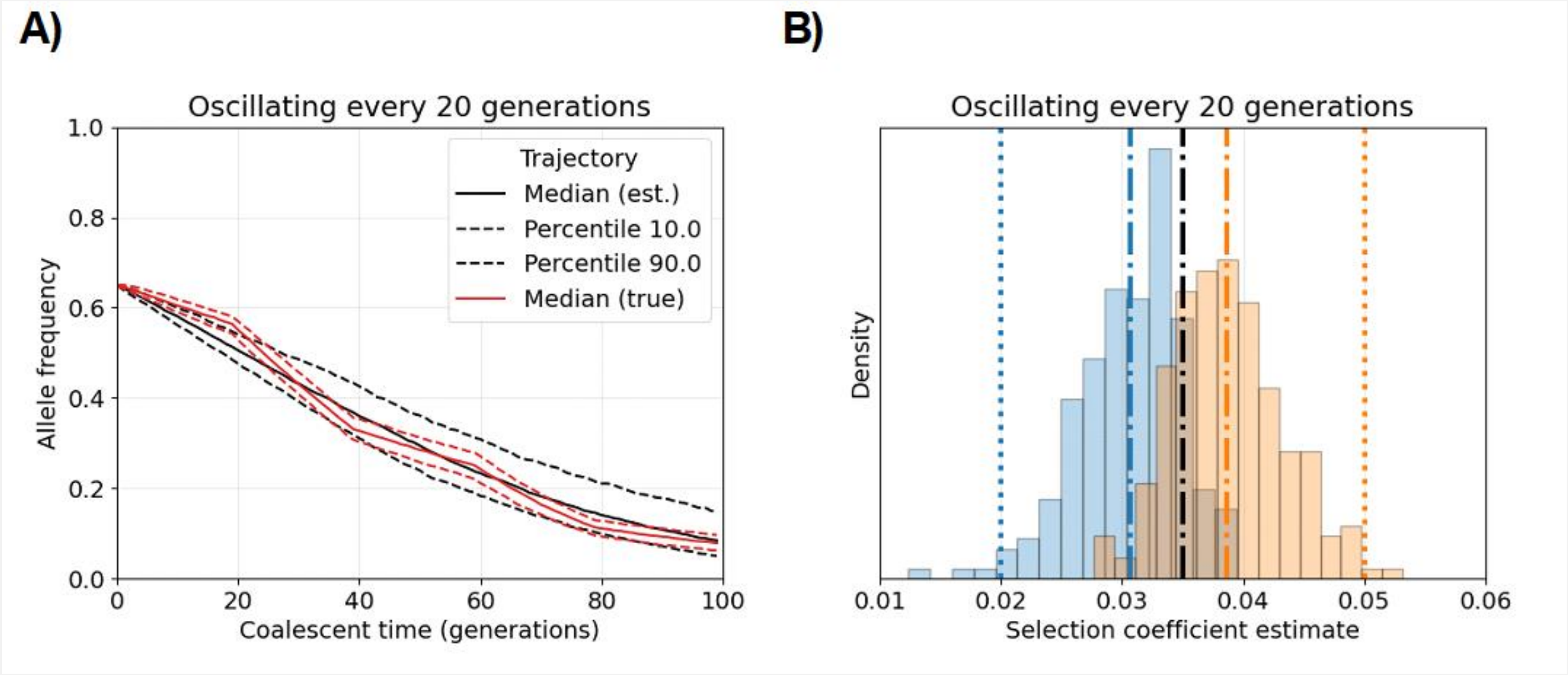
Our estimates (approx.)

Mathieson and Terhorst (2022)

Comparing against allele frequency method



Mathieson and McVean (2013)


A)
Oscillating every 20 generations
Trajectory
Median (est.)
Percentile 10.0
Percentile 90.0
Median (true)
Allele frequency
Coalescent time (generations)
B)
Oscillating every 20 generations
Density
Selection coefficient estimate