BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH

Department of STATISTICS

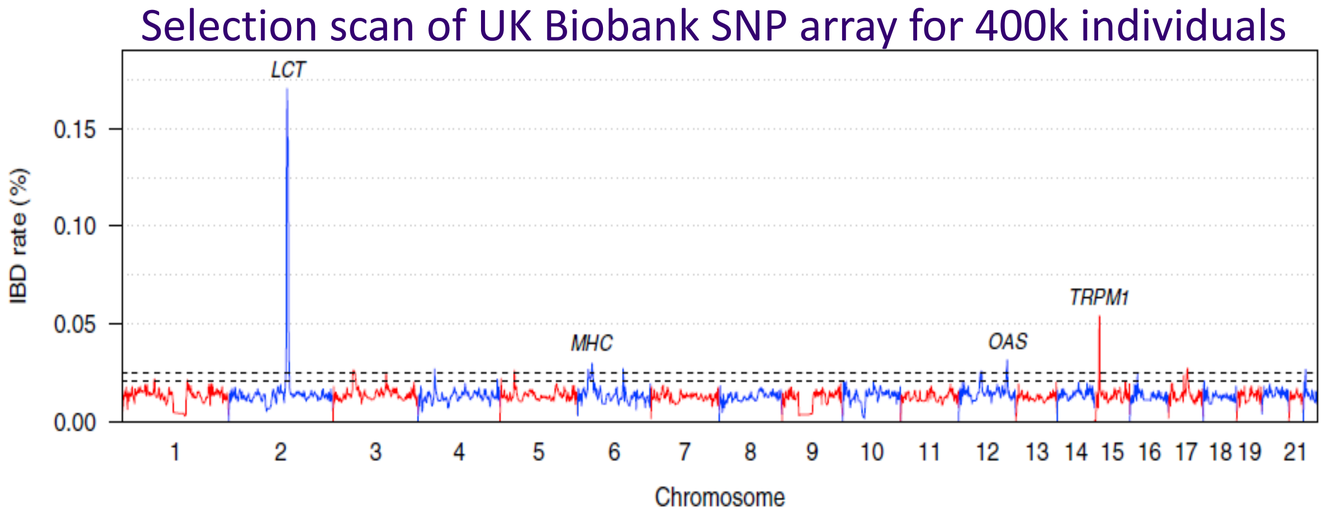
Inferring incomplete sweep parameters for recent adaptive selection using identity-by-descent (IBD) segments

Presenter: Seth D. Temple (PhD student), Advisor: Sharon R. Browning

Highlights

- Method to infer incomplete sweep parameters, selection coefficient and variant frequency, using identity-by-descent (IBD) segments
  - Does not require knowledge of the causal variant; can be untyped in dataset
  - Implies curve of variant frequency back in time, tracking human evolution in past 10,000 years

Selection scan of UK Biobank SNP array for 400k individuals



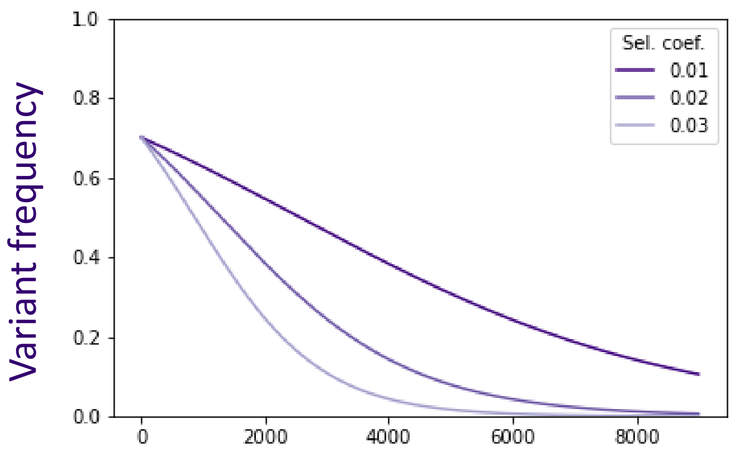
Rate of IBD segments along autosomes (Figure 4, Browning & Browning, 2020)

Introduction

What is an incomplete sweep?

- New adaptive allele ↗ in frequency.
- Consider *(1+s) : 1* fitness advantage; **s is (constant) selection coefficient.**
- Current variant frequency is p_0.
- E.g., mutant for lactase persistence at high freq. in world populations.

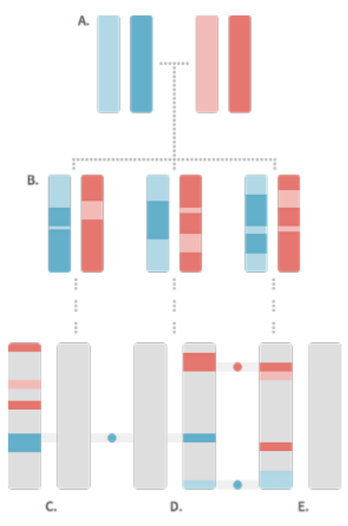
Incomplete sweeps of varying s



Years ago (1 generation = 30 years)

What is an IBD segment?

- Genomic region inherited from a common ancestor without recombination



Schematic for IBD segments (Photo from Ancestry DNA (Ball et al., 2016))

IBD segments give evidence for selection

- Extended haplotypes at high freq. indicate recent positive selection.
- Selection creates substructure s.t. effective size of haploids w/ helpful mutant shrinks quicker, making excess IBD.
- Model relating IBD segments to demographic history promotes principled data analysis.

Inference Method

- *Variant frequency (sample mean).* Label phased haploids into groups *A* and *B* based on short identity-by-state (IBS) segments, defining a local network.
- *Selection coef.* (*moment moment*). Choose s such that empirical IBD rate matches expected IBD rate, where IBD rate is *P(seg >= X* cM *)*.
- *Parametric bootstrap.* Simulate segments from fast *n*-coalescent.

Simulation Study

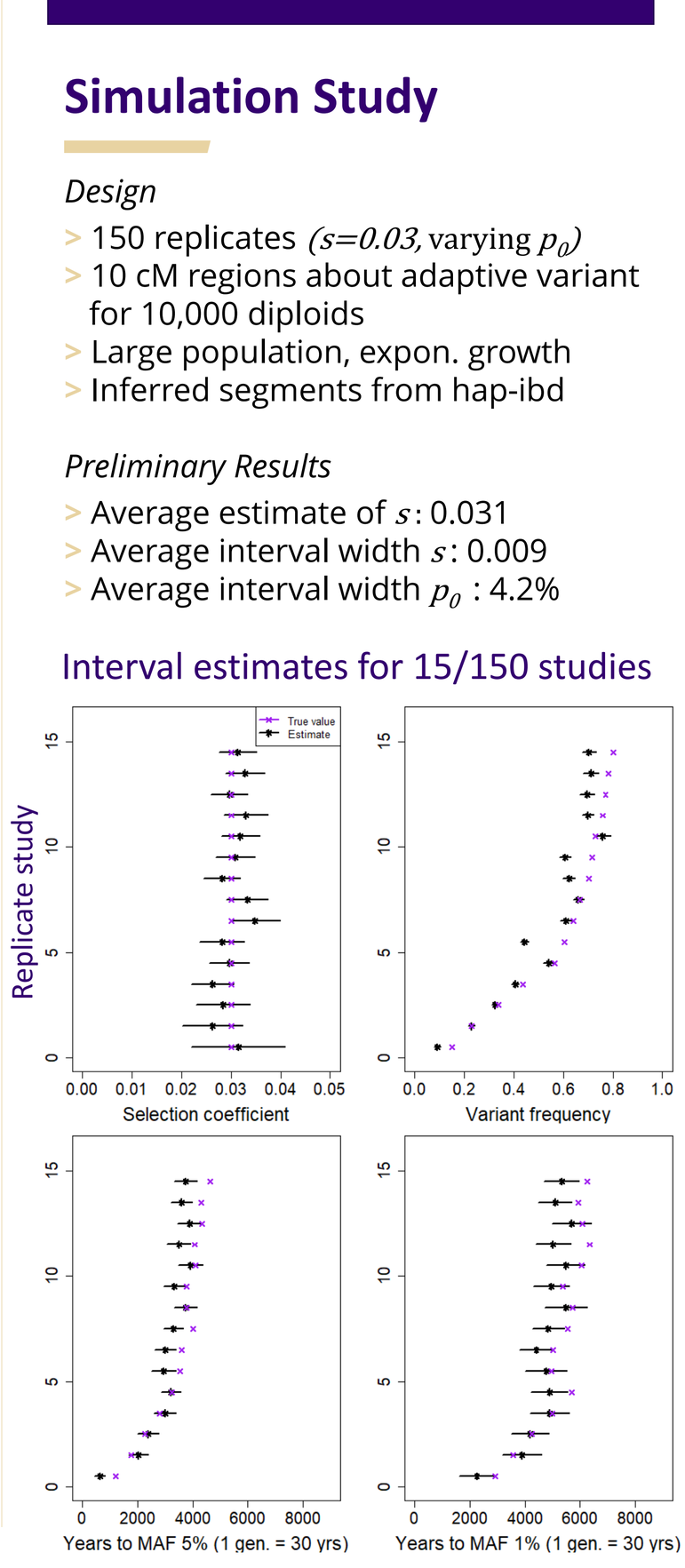
Design

- 150 replicates *(s=0.03,* varying p_0*)*
- 10 cM regions about adaptive variant for 10,000 diploids
- Large population, expon. growth
- Inferred segments from hap-ibd

Preliminary Results

- Average estimate of s: 0.031
- Average interval width s: 0.009
- Average interval width p_0 : 4.2%

Interval estimates for 15/150 studies



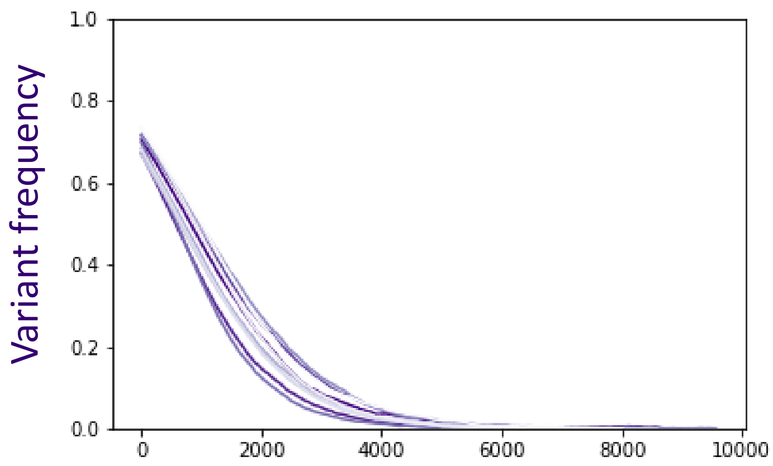
Real Data Analysis

Above: UK Biobank selection scan highlighted genomic regions with excess IBD sharing. Estimate s, p_0 for these suggestive regions.

Preliminary result for LCT region

- Common variant -13.910:T not typed
- Peak IBD rate is 0.1702%
- Selection coefficient: 0.032 +- 0.006
- Variant frequency: 71.8% +- 4.8%
- Time to 1% MAF: 5,300 +- 530 YA
- Prior studies report s between 0.01 - 0.09 w/ interval +- 0.005 - 0.05, timing between 2,000 - 20,000 YA.

Plausible trajectories for LCT selection



Years ago (1 generation = 30 years)

Conclusions

- Selection impacts IBD segments through population structure.
- Our IBD method to study positive selection provides inference with low bias, tight intervals.
- Implicated genes involved in immunity, nutrition, pigmentation.
- Downstream analysis may fine-map causal variants.

References, Acknowledgements

Financial support of the NDSEG fellowship from the US DoD and the NIH T32 predoctoral traineeship in statistical genetics

- Ball, C. A., et al. *Ancestry.com*. 2016. 1-46.
- Browning, R. Sharon, & Brian L. Browning. 2020. *AJHG* 107 (5): 895-910.
- Zhou, Ying, et al. 2020. *AJHG* 106 (4): 426-37.

UNIVERSITY of WASHINGTON