Detecting Selection in Experimental Evolution Experiment

Arya Iranmehr
airanmehr@ucsd.edu

Bafna Lab
University of California, San Diego

September, 2016

Introduction

- Next generation sequencing has made whole-genome & whole-population sequencing possible.

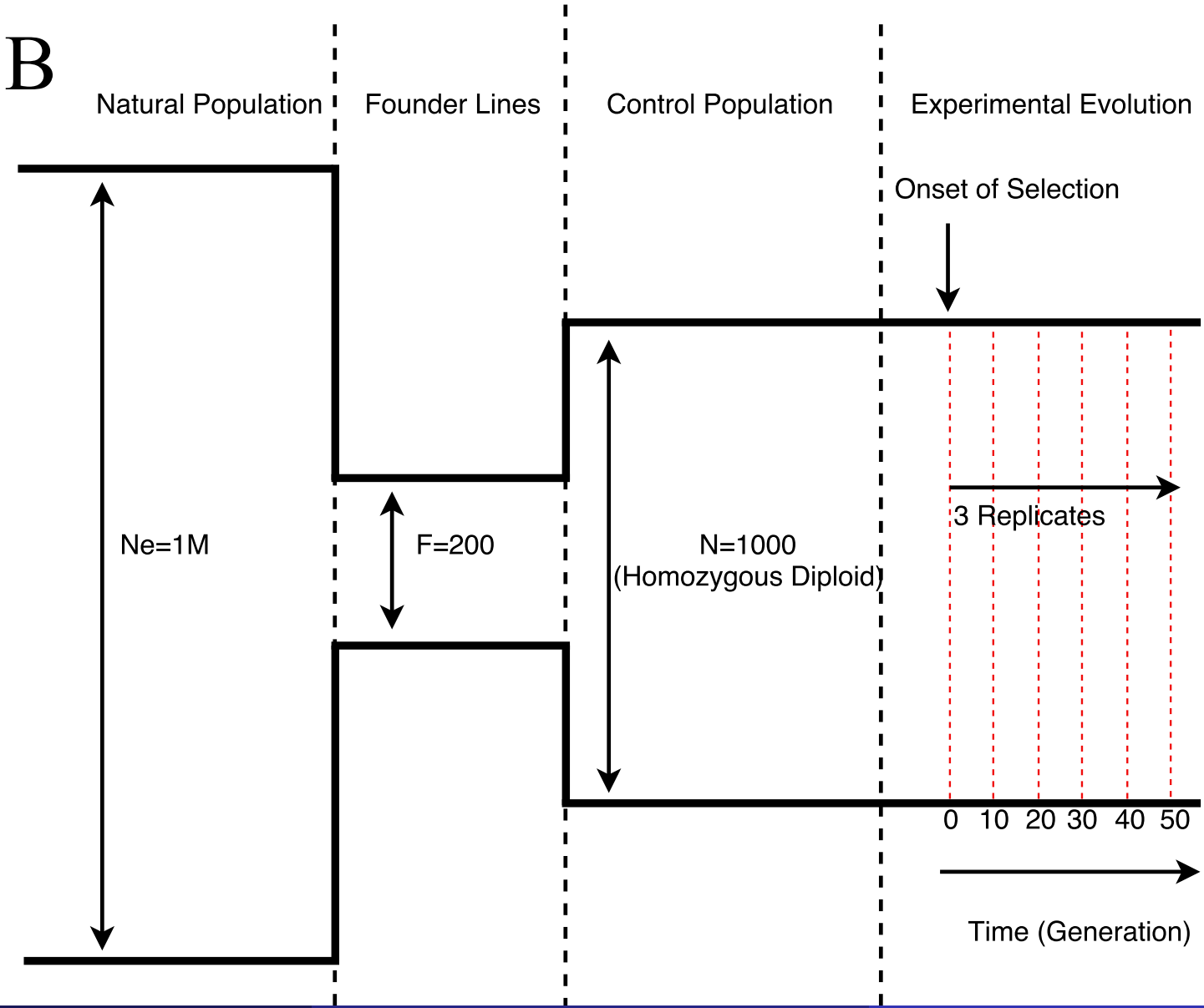
Introduction

- Next generation sequencing has made whole-genome & whole-population sequencing possible.
- For organisms with "short-generation-time", (e.g., yeast, *E. coli*, *D. melanogaster* etc.) it is possible to collect time-series data of population.

Introduction

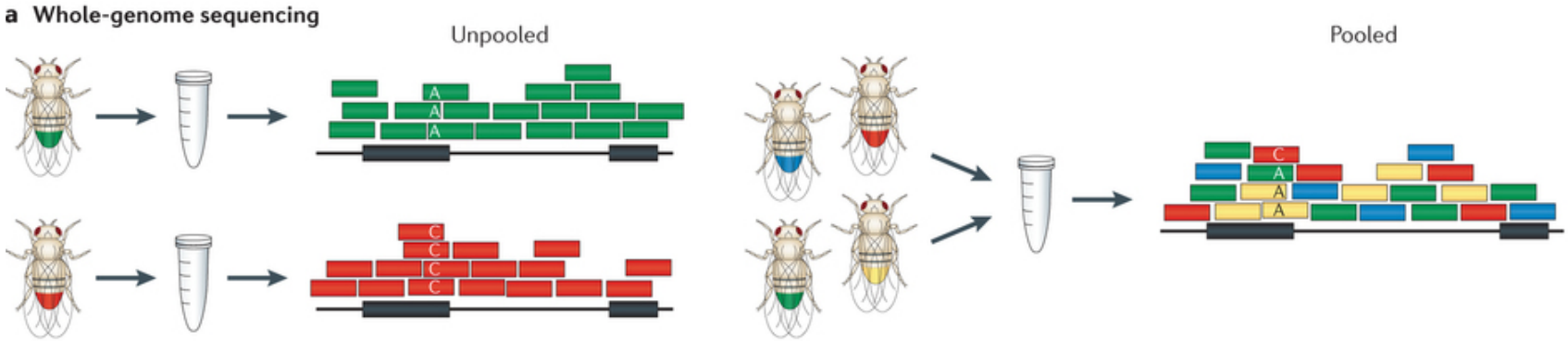
- Next generation sequencing has made whole-genome & whole-population sequencing possible.
- For organisms with "short-generation-time", (e.g., yeast, *E. coli*, *D. melanogaster* etc.) it is possible to collect time-series data of population.
- Given rise of these modern datasets (population longitudinal data),

Experiment design



Whole-Genome Whole-Population Sequencing

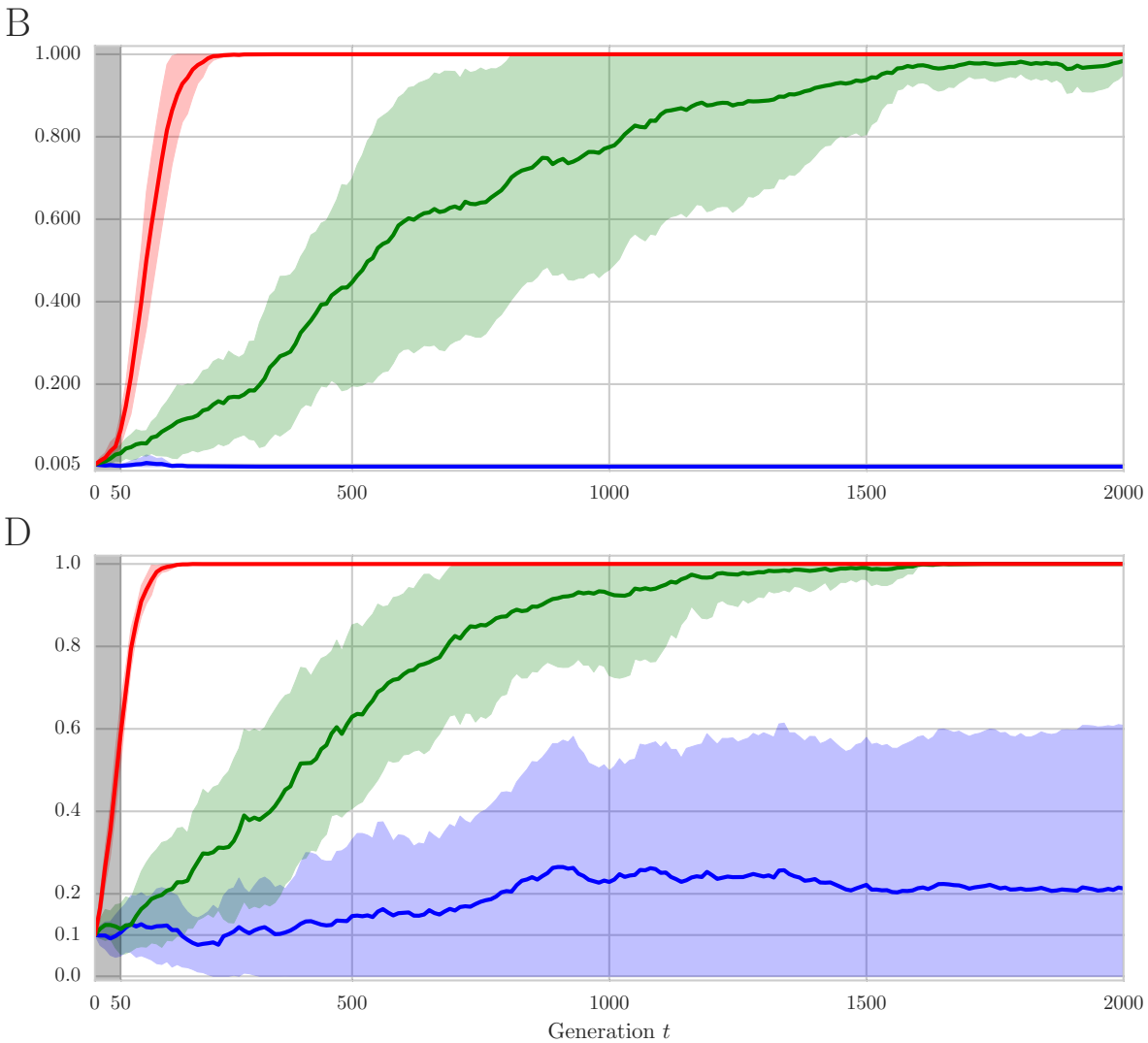
- Pooled-Sequencing



Nature Reviews Genetics 15, 749-763 (2014)

- Implication: only population allele frequency can be computed.

Dynamic of population allele frequency



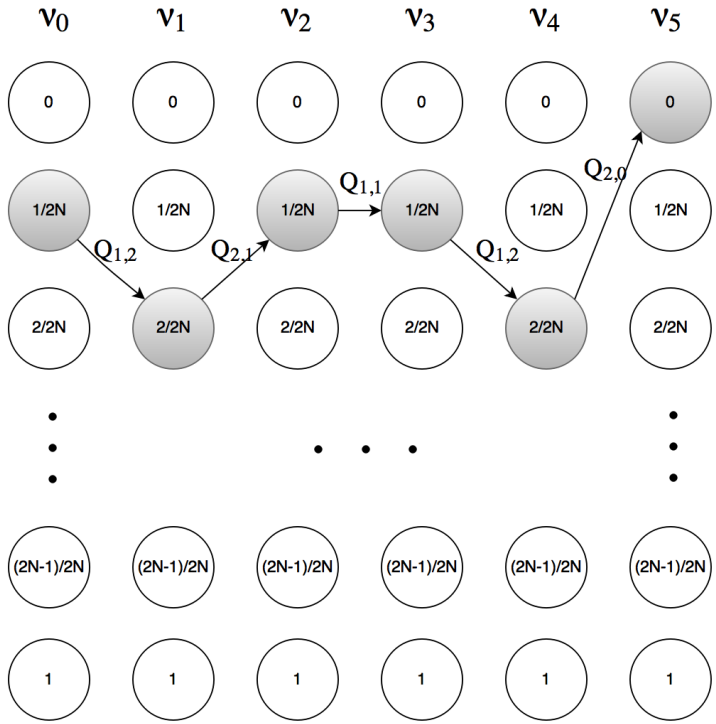
Goals

Design a method which

- Detect regions under selection.
- Localizing adaptive allele within the region.
- Estimating model parameter.

Simplified Model (I)

- Suppose we have sequenced a whole (diploid, size=N) population every generation (eg, for 6 generations) and we exact allele frequency.
- The a discrete-time discrete-state model, Markov chain, can generate such a data.



$$P(\nu_0, \dots, \nu_5) = Q_{1,2}\ Q_{2,1}\ Q_{1,1} Q_{1,2}\ Q_{2,0}$$

Simplified Model (II)

- Where $Q_{i,j}(s,h)$ is the probability of going from frequency $i/(2N)$ to $j/(2N)$ when selection strength is s and over dominance is h.
- Likelihood of parameter can be easily computed

$$\mathcal{L}(s,h|\{\nu_0,\ldots,\nu_5\}) = \Pr(\{\nu_0,\ldots,\nu_5\}|Q(s,h))$$

- perform maximum likelihood to find $\hat{s}$, $\hat{h}$
- compute likelihood ratio, M statistic for each SNP:

$$M = \frac{\text{likelihood of data as if being under selection with } \hat{s},\hat{h}}{l\text{ikelihood of data as if being neutral}}$$
$$= \frac{\mathcal{L}(\hat{s},\hat{h}|\{\nu_0,\ldots,\nu_5\})}{\mathcal{L}(0,0|\{\nu_0,\ldots,\nu_5\})}$$

Model (complete)

- In reality, population is sequenced after some (τ) generations.
  solution: use Q^τ in computing likelihoods.
- Allele frequencies are unknown, and depth of each variant can be different.
  solution: extend Markov chain to an HMM by specifying emission probabilities

$$d \sim \text{Poisson}(\text{Coverage})$$

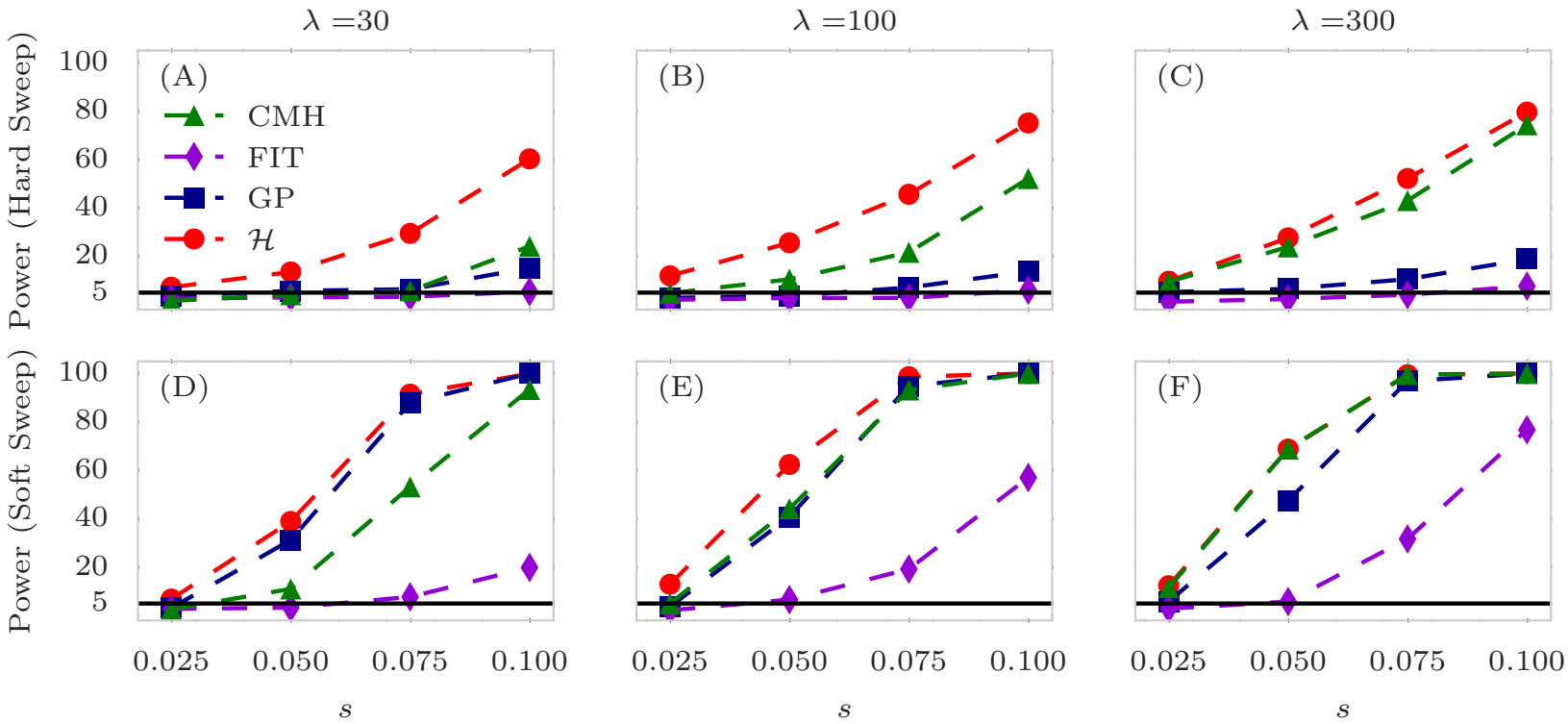
$$c \sim \text{Bionomial}(N = d, \theta = \nu)$$

Composite Likelihood

- In general there are non random

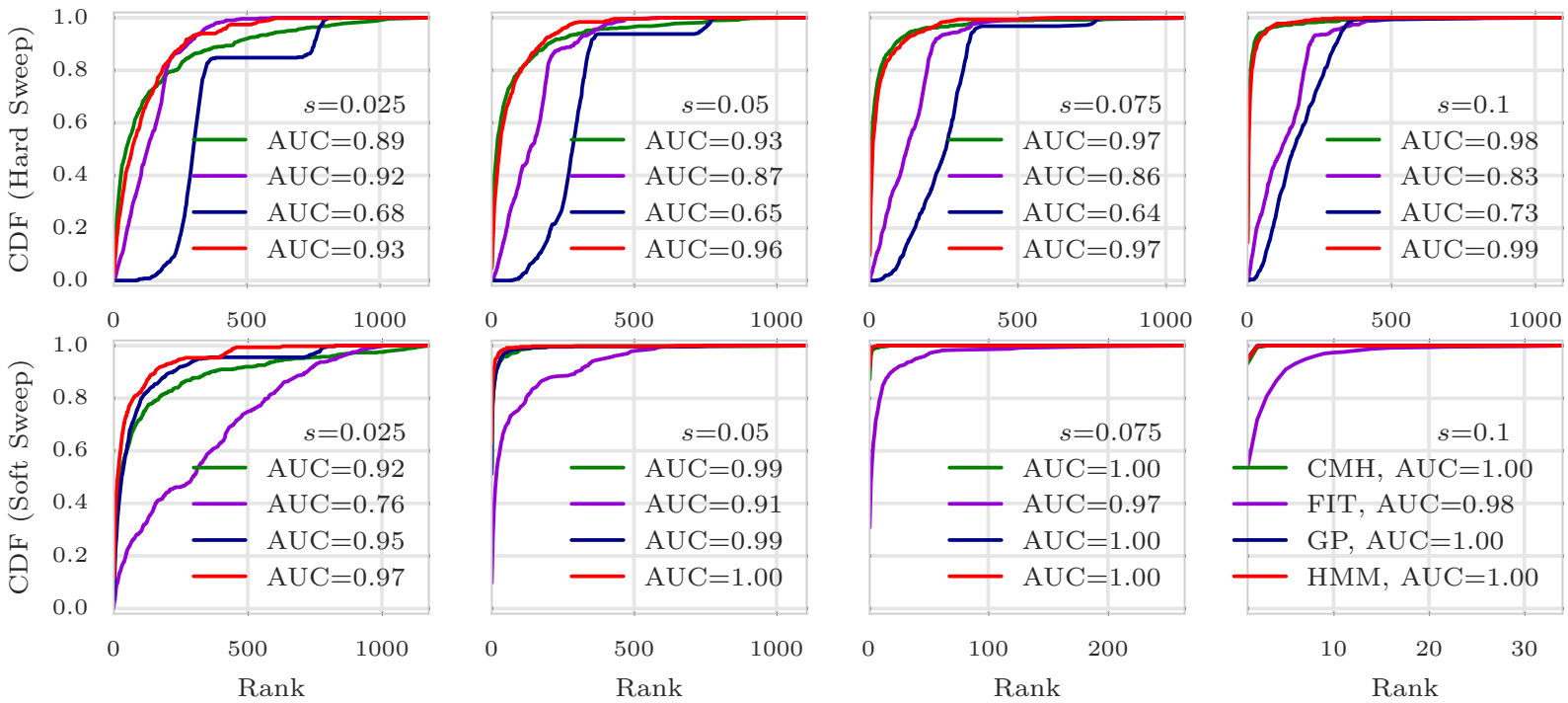
Detecting regions under selection

Each point represent power of detection in 1000 simulation (500 neutral, 500 selection) of a 50Kbp window.

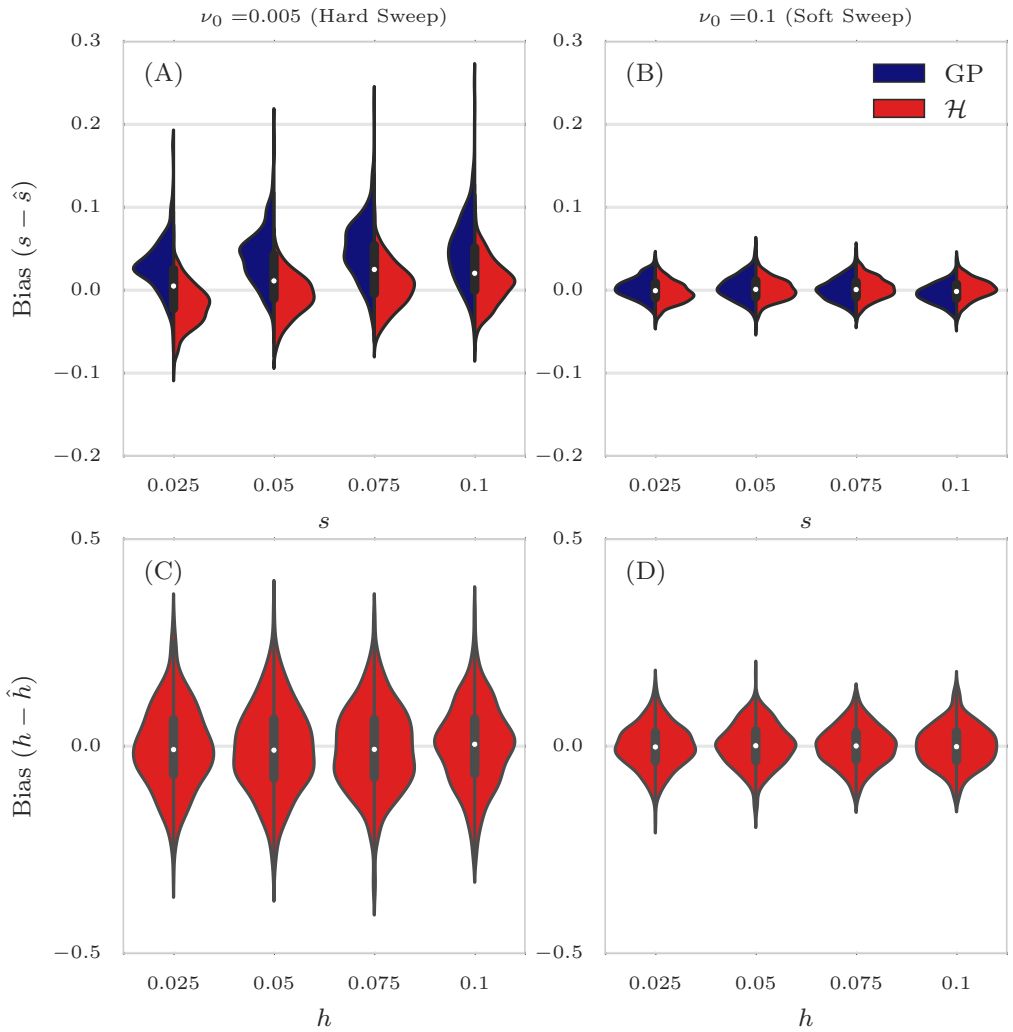


Localizing favored allele

Genome scan for sliding window size=50Kbp, steps=10Kbp



Estimating parameters



Detecting regions under selection in real data

Genome scan for sliding window size=50Kbp, steps=10Kbp