Detecting Selection in Dynamic Data

Arya Iranmehr
airanmehr@ucsd.edu

Bafna Lab
University of California, San Diego

March, 2016

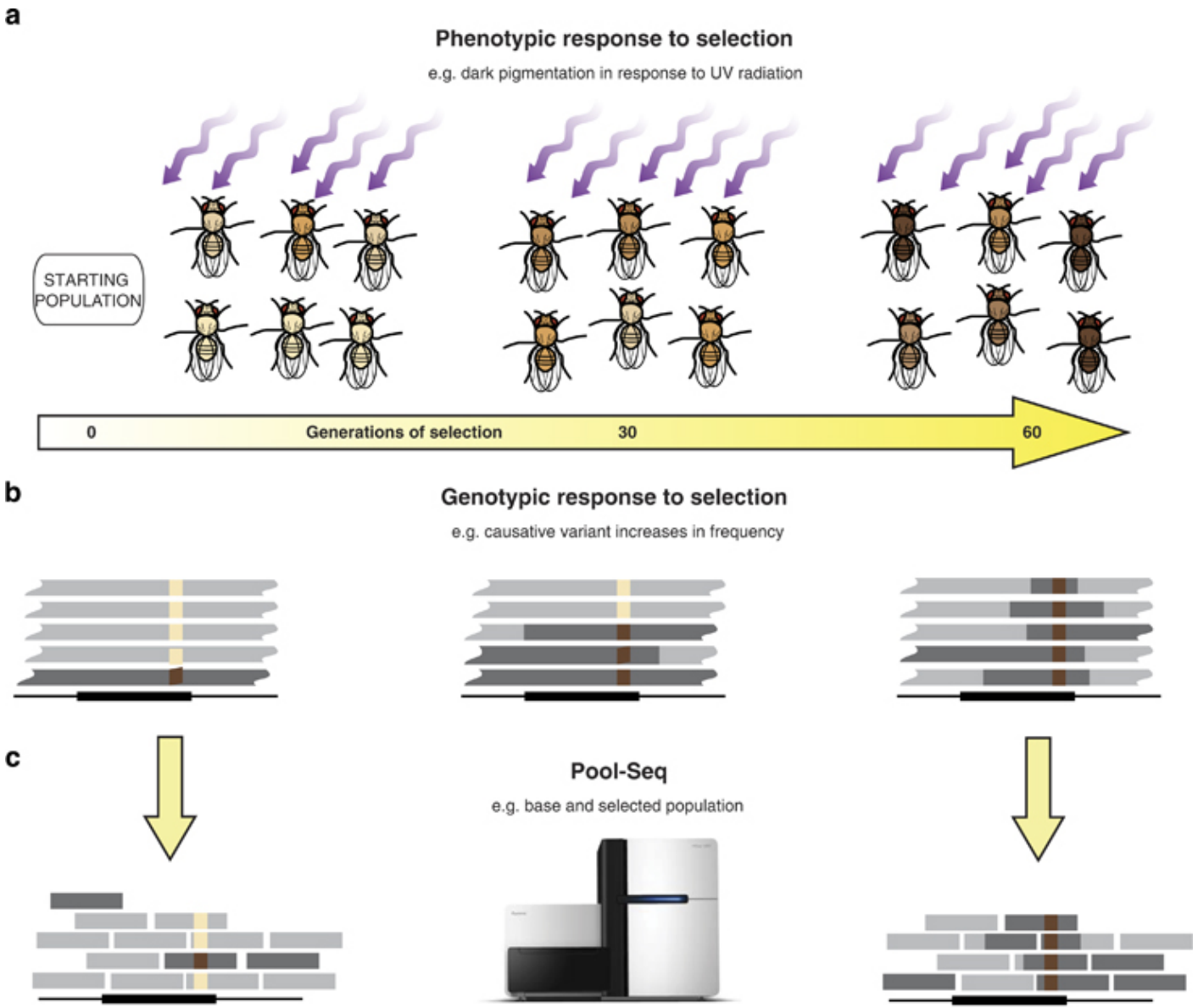
Introduction

- We are interested in (quickly) identifying genetic adaptations, in different organisms in find genes (or alleles) that are beneficial w.r.t. a selective pressure.
- Examples of interesting selective pressures:
  - harsh environmental conditions
  - antibiotics
  - chemotherapy
  - anti-viruses
  - pesticides
  - etc.

Datasets

- **Classical Static Datasets** Given a single observation of Allele Frequency Spectrum(AFS vector) or SNP matrix of a region (say 50Kbp) a variety of tests applies.
- **Modern Dynamic Datasets**, multiple observations are available.

Experimental Evolution



Goals

- Given the genome time series data (allele frequencies) we are interested in
  - I Detecting selection
  - II Locating the favored mutation
  - III Estimating strength of selection

Simulations

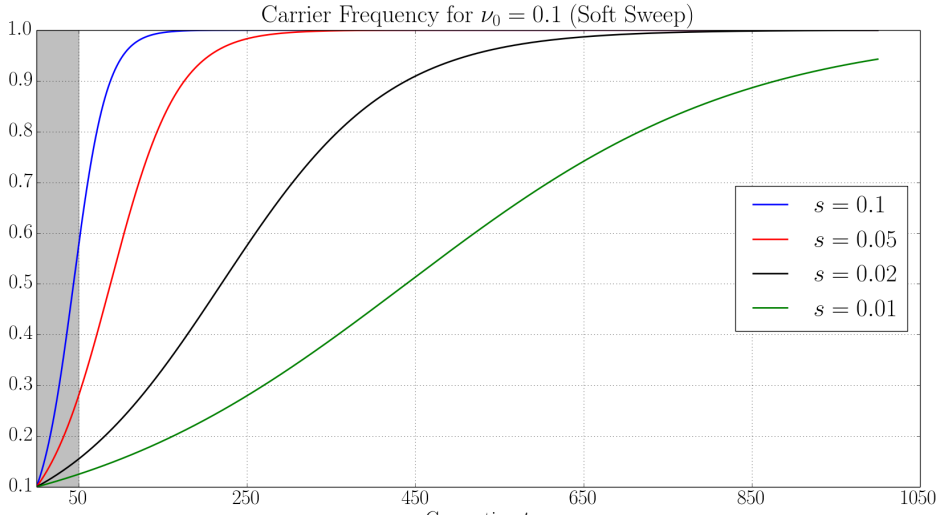
(i) For each simulation, population of $F = 200$ founder lines is created in `msms` program with parameters
- window size $L = 50$Kbp.
- scaled mutation rate $\theta = 200$. (≈ 1100 SNPs)

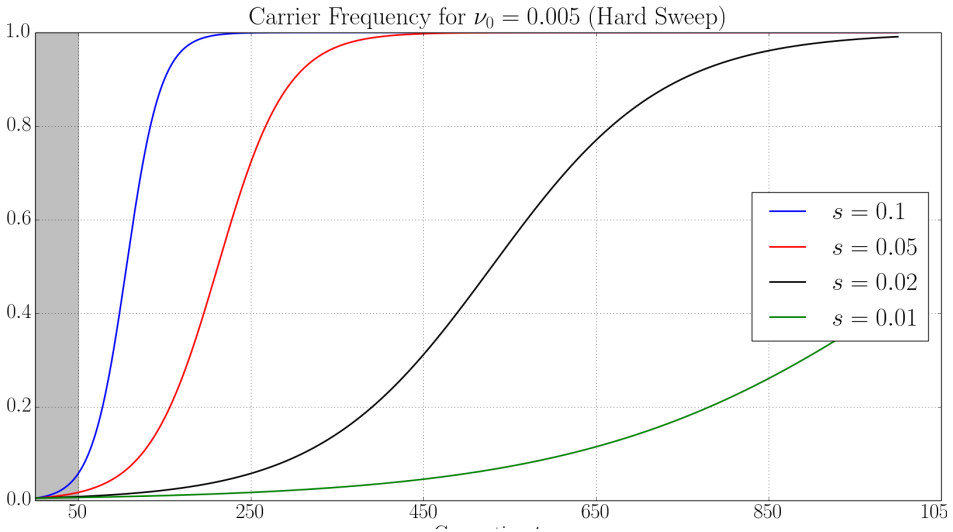
(ii) Using $F = 200$ founder lines a population diploid is created.

(iii) Using forward simulator population is evolved and AF is sampled every 10 generation for 50 generations.

(iv) This procedure is repeated 100 times for each $s = \{0.1, 0.05, 0.02, 0.01, 0\}$.

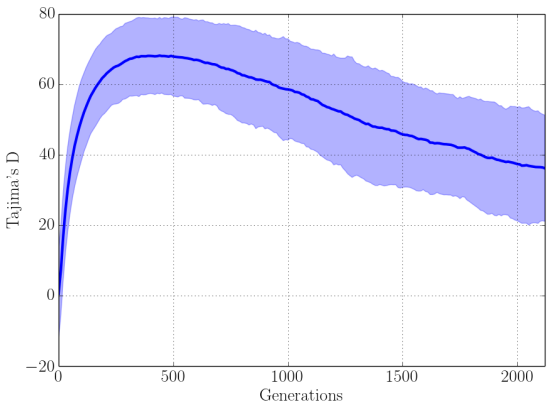
Dynamic of Allele Frequency under Selection

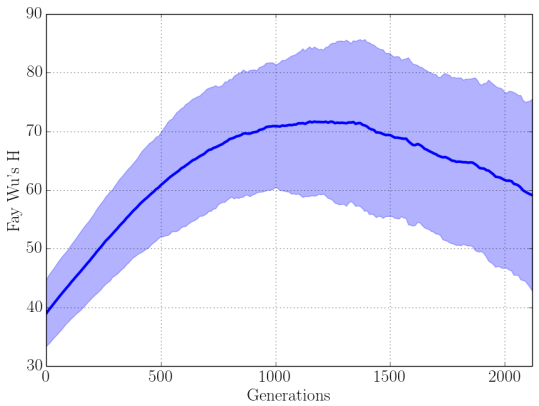


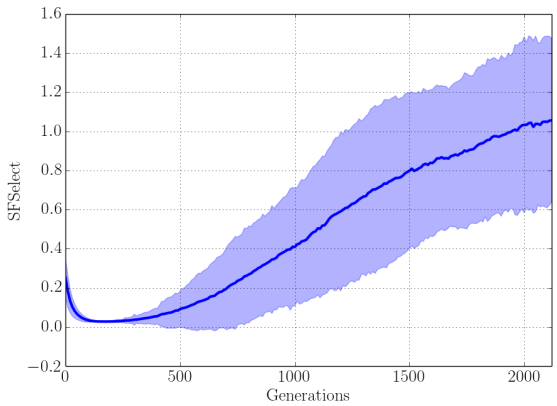


Dynamic of AFS under Neutral Evolution + Bottleneck

- side-effect of not tracking new mutations and restricting population to founder-lines.







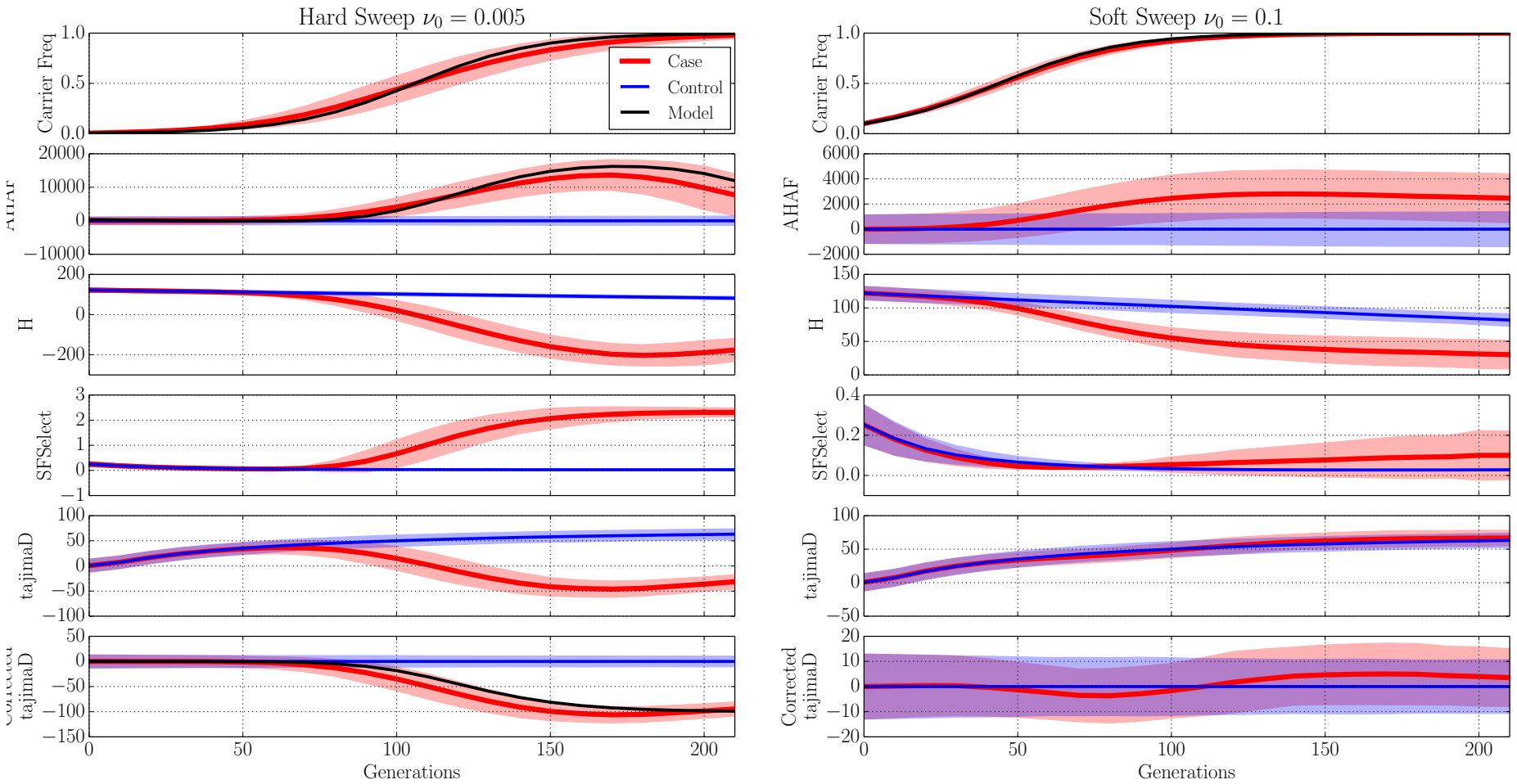


Figure : Mean and 95% CI of 1000 simulations for strong selection.

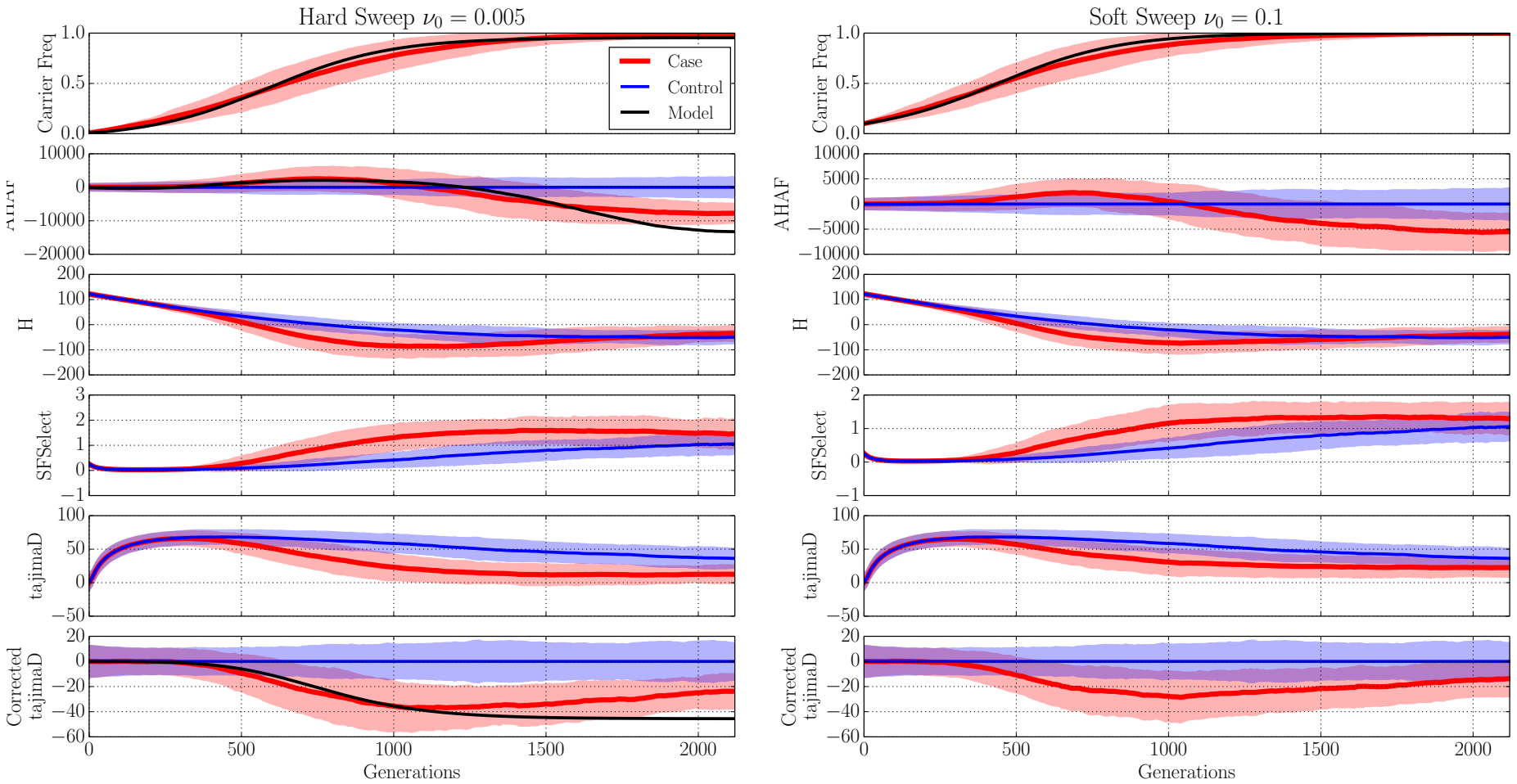
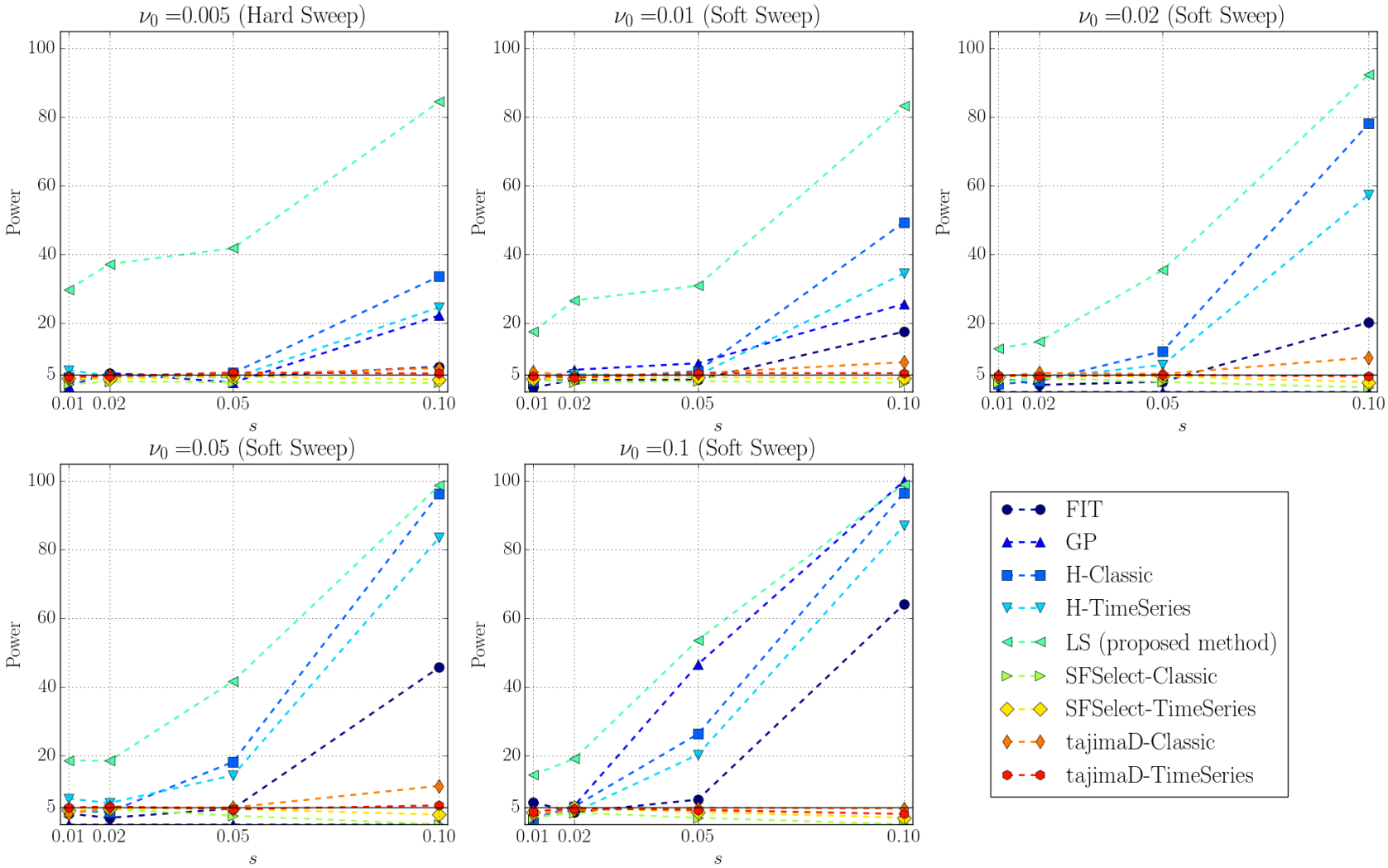


Figure : Mean and 95% CI of 1000 simulations for weak selection.

Experiments

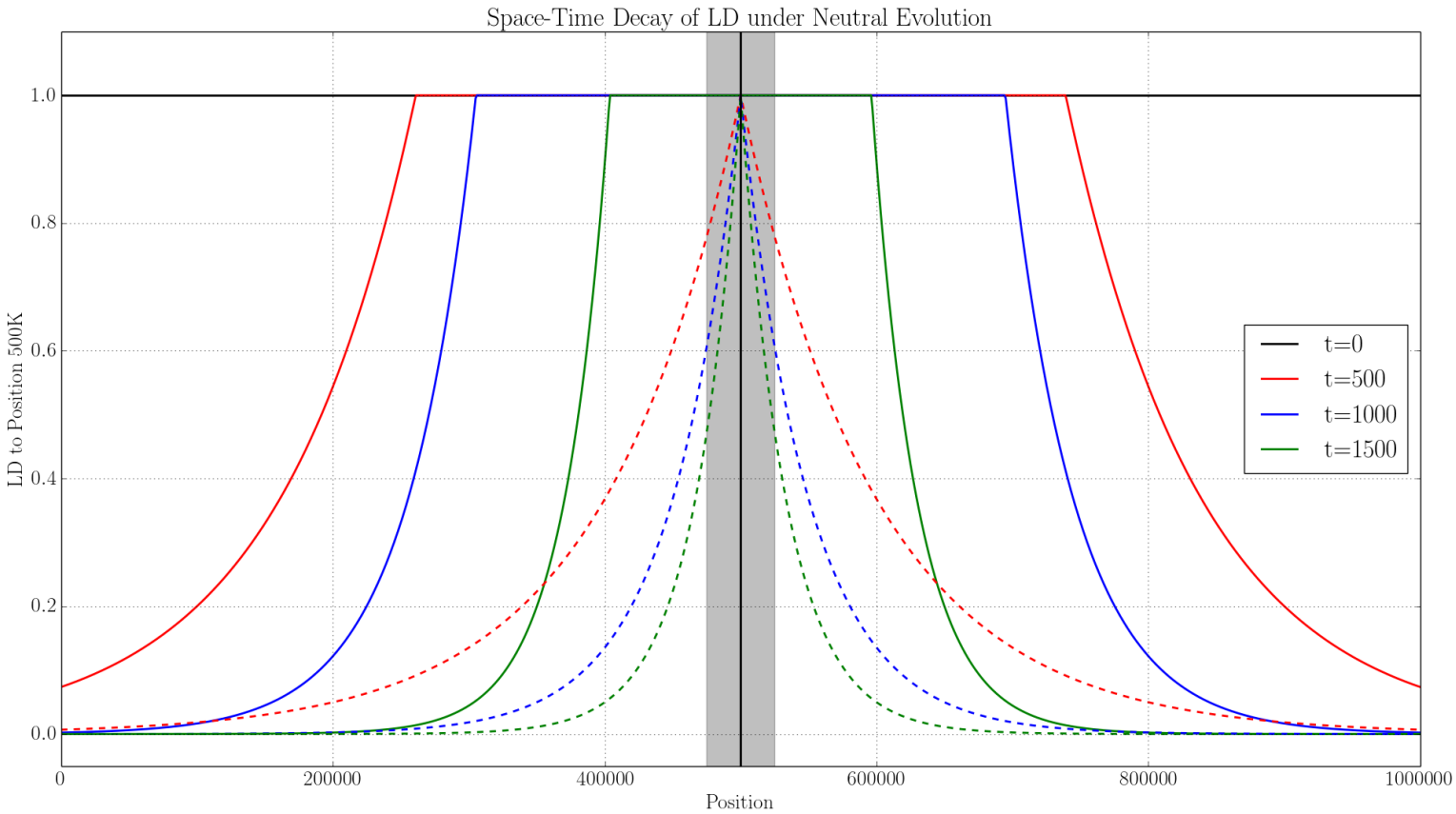
- for each setting we performs 200 simulations, (100 neutral and 100 selection), and predictive performance of all the methods in detecting selection.
- We computed ROC curve and defined power of a method as area under ROC curve when False-Positive rate is less than 0.1.

Power of detection

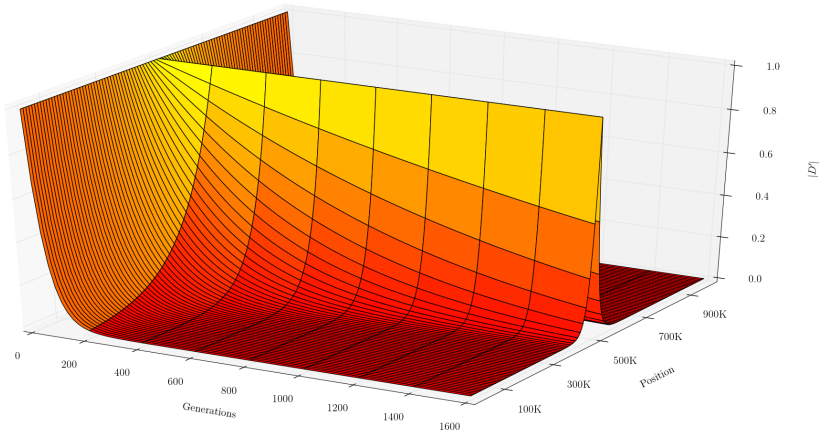


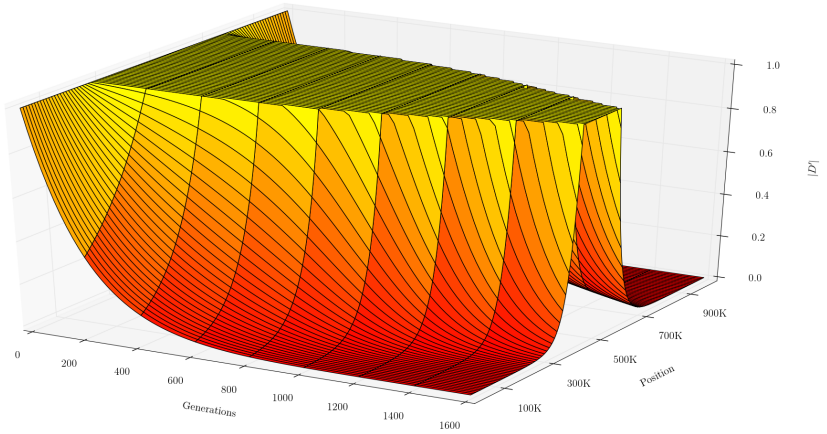
Locating selection I

Strong LD in a window

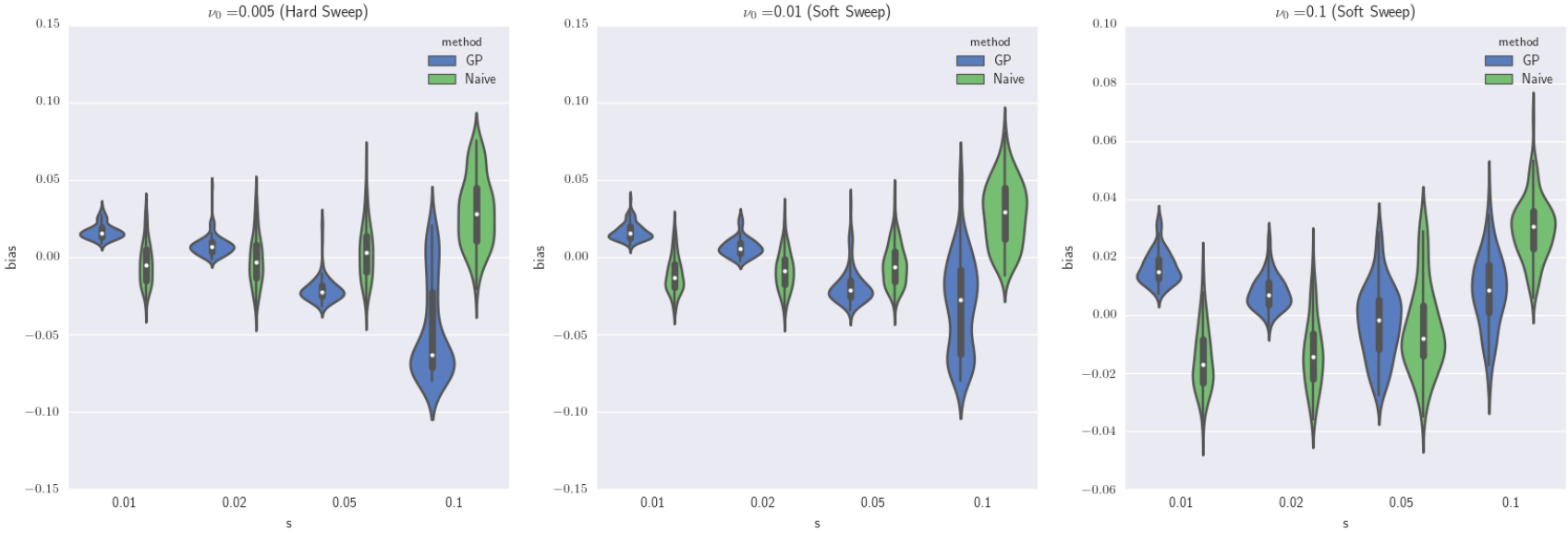


Locating selection II





Power of detection



Summary

- Behaviour of different population statistics in time is studied.
- A method based on single locus AF is proposed and shown to have superior performance than multi-locus and traditional methods.
- Analysis of real data and locating genes in the genome is our next step.

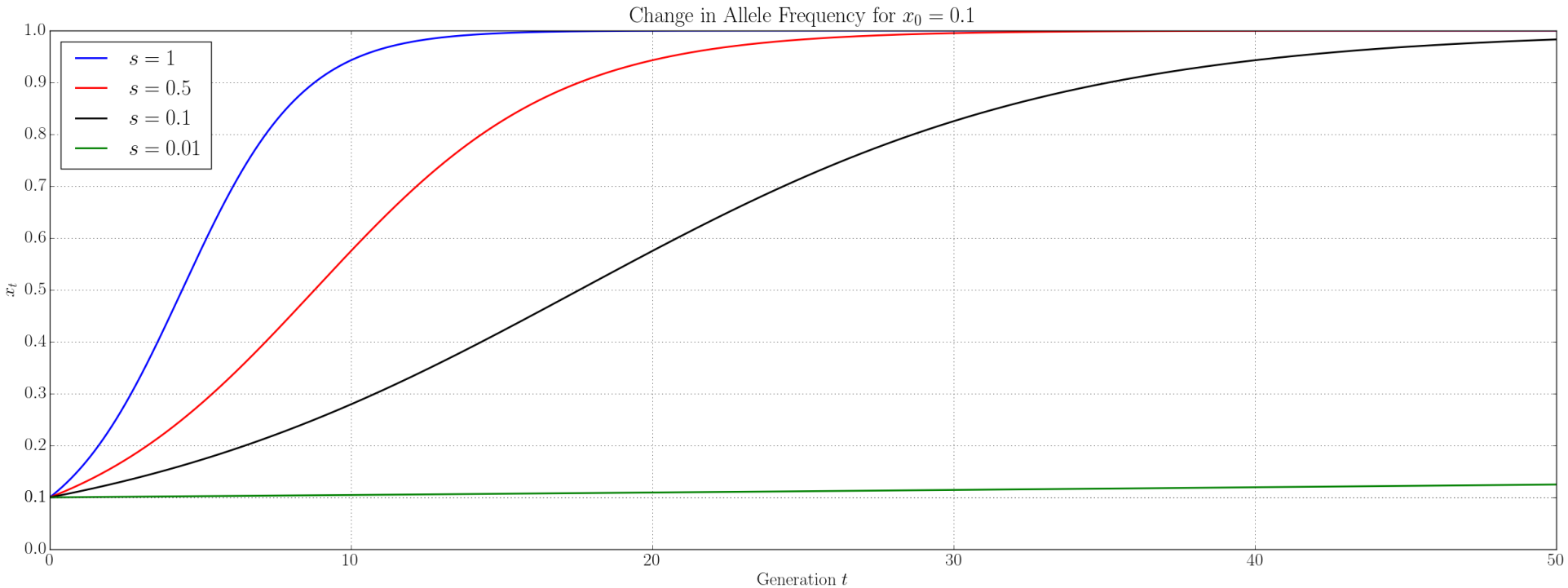
Thanks!

Single Locus Model: Carrier Frequency

- By differentiating update equations ($x_{t+1} = x_t + \frac{sx_t(1-x_t)}{2+2sx_t}$ w.r.t. t and solving differential equation, we have

$$\nu_t = \sigma\left(st/2 + \eta(\nu_0)\right)$$

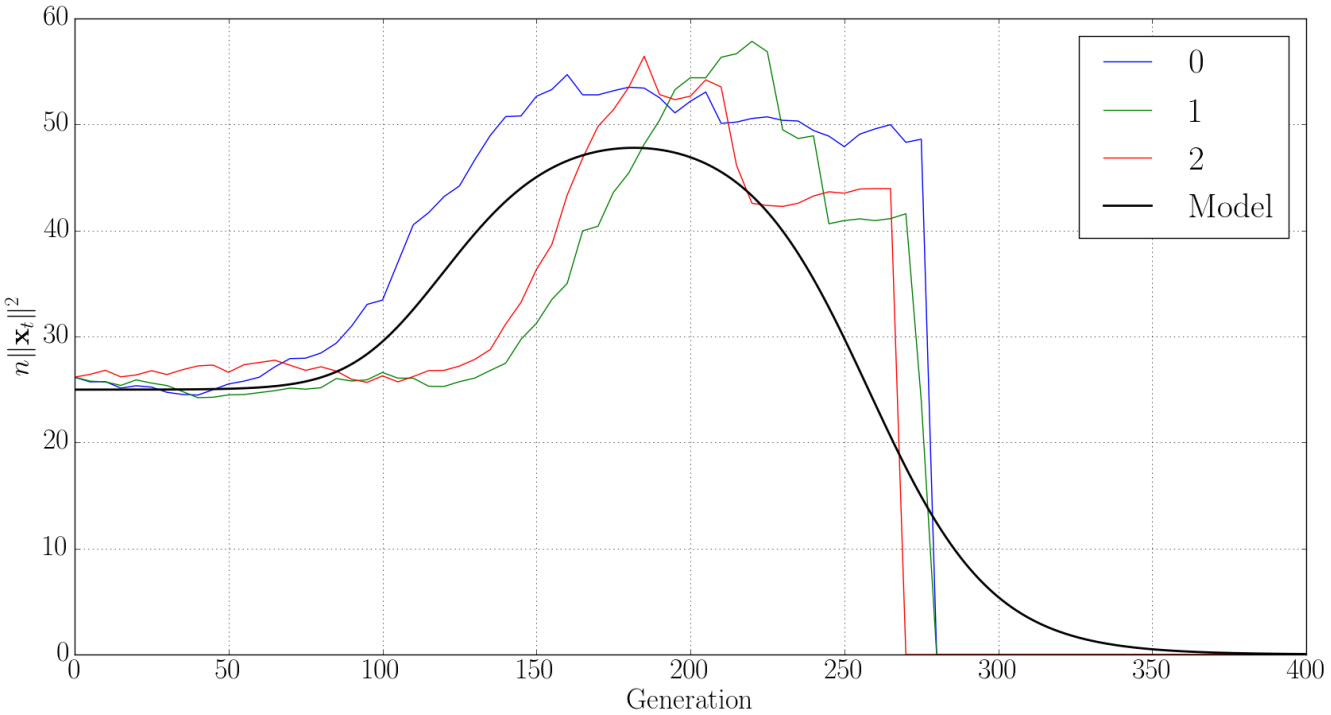
where $\sigma(.)$ is logistic function and $\eta = \sigma^{-1}$ is the logit function and ν_t is the frequency of the carrier at time t.



Multi Locus Model: Average Haplotype Allele Frequency

$$\mathbb{E}[\text{1-HAF}(t)] = \|\mathbf{x}_t\|^2 \approx \theta\nu_t \left(\frac{\nu_t+1}{2} - \frac{1}{(1-\nu_t)n+1}\right) + \theta(1-\nu_t)\left(\frac{n+1}{2n} - \frac{1}{(1-\nu_t)n+1}\right)$$

where $\mathbf{x}_t$ is vector of AF at time t and $\nu_t = \sigma\left(st/2 + \eta(\nu_0)\right)$.



Critics to GP

- Its too complex, likely to overfit with small number of iid replicates.
- Although the likelihood model is based on different parameters, in practice, it can learn only one parameter at a time.
- not tractable, its time complexity is quartic!
- worse, each iteration requires `maxGeneration` recursion which makes it very hard to analyse late epochs of sweeps.
- In addition to PoolSeq data, it requires initial population haplotypes.
- Despite its elegant theory, it has not compared with classical methods.
- In practice, single locus scan is performed and multi-locus (with 3-7 seg. sites.) model is fitted at regions of interests.