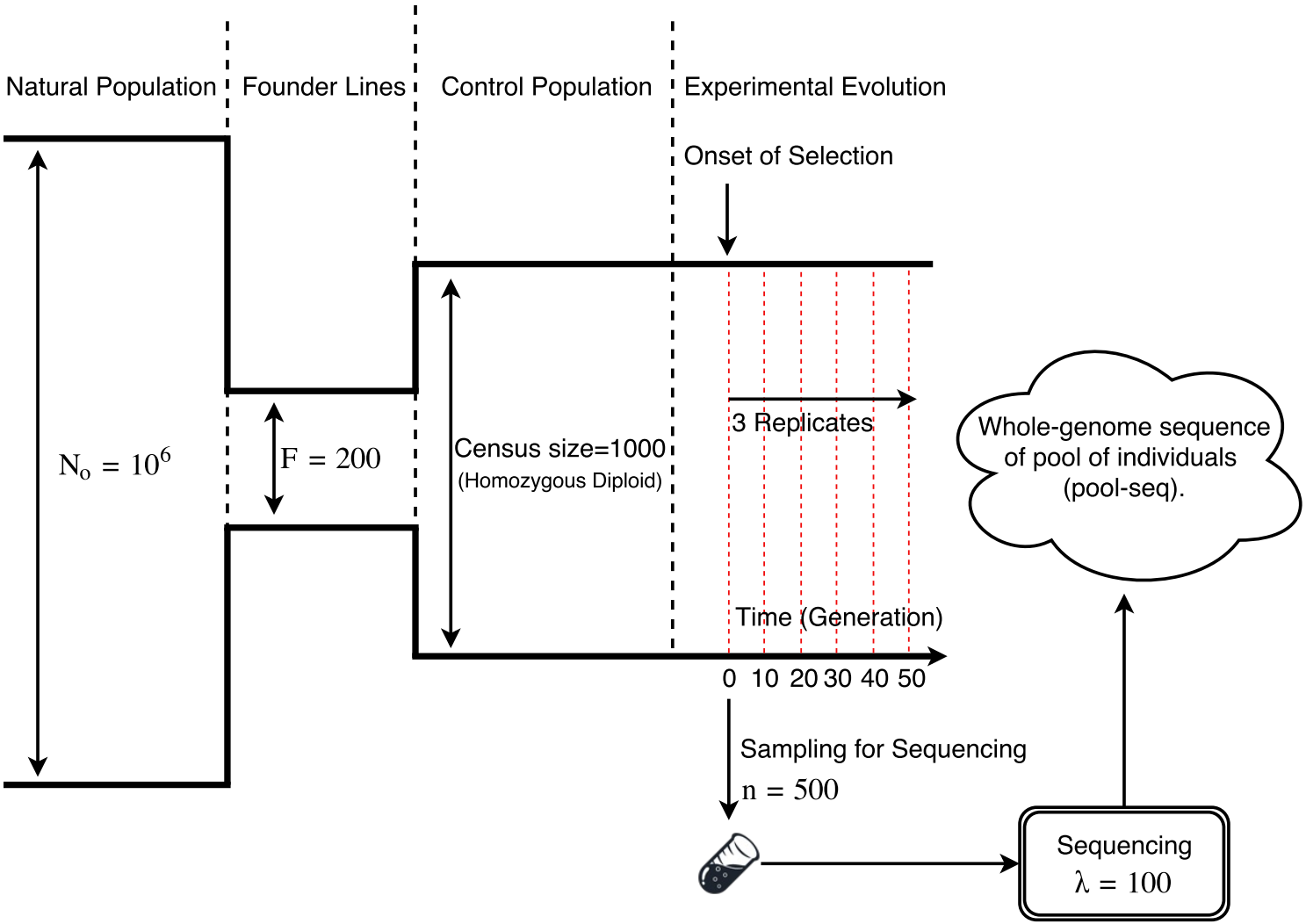
# Identifying Selection in Experimental Evolution

Arya Iranmehr
airanmehr@ucsd.edu

Bafna Lab
University of California, San Diego

March, 2017

# An experiment design for *D. melanogaster*

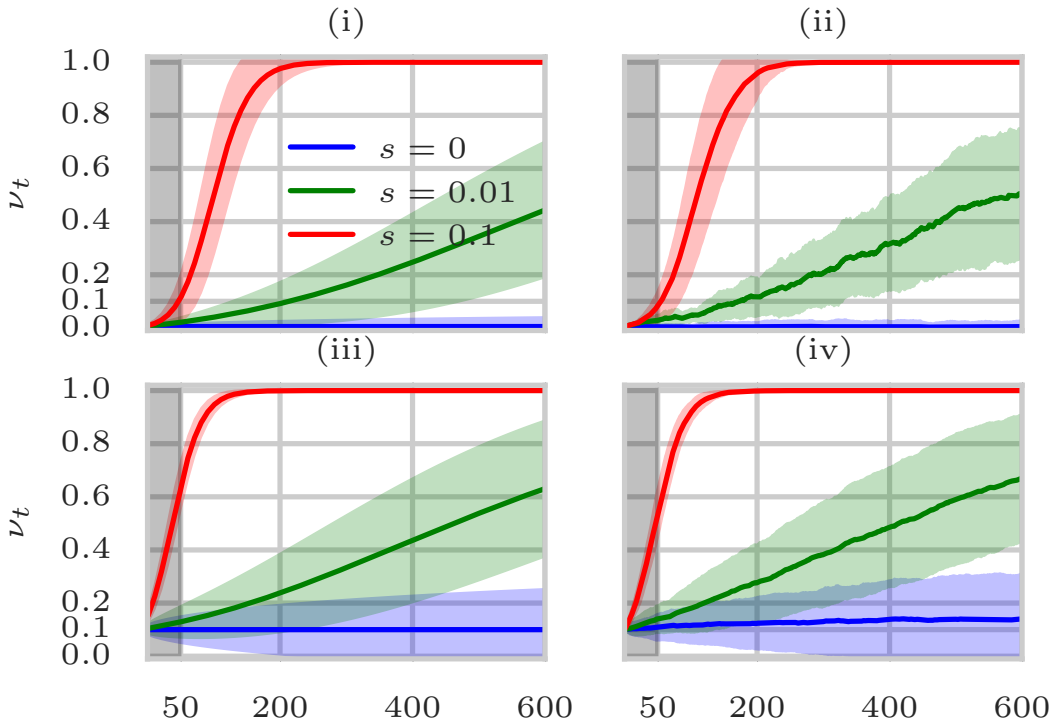


# Challenges (I)

- Small population size $\Rightarrow$ strong drift

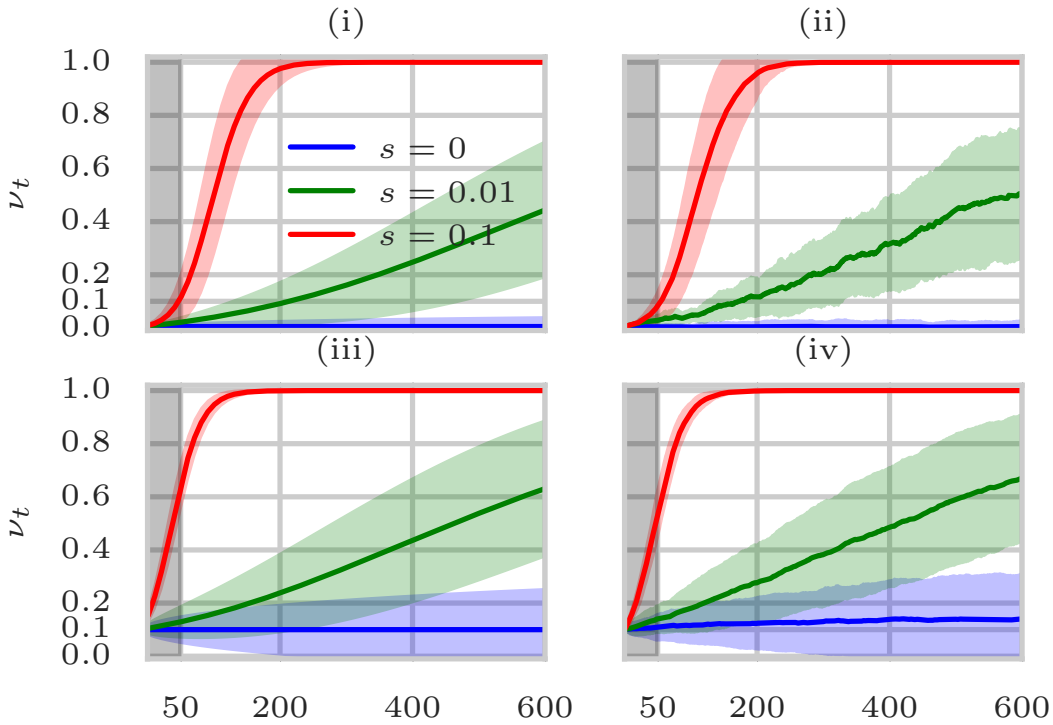
# Challenges (I)

- Small population size ⇒ strong drift
- Partial sweep



# Challenges (I)

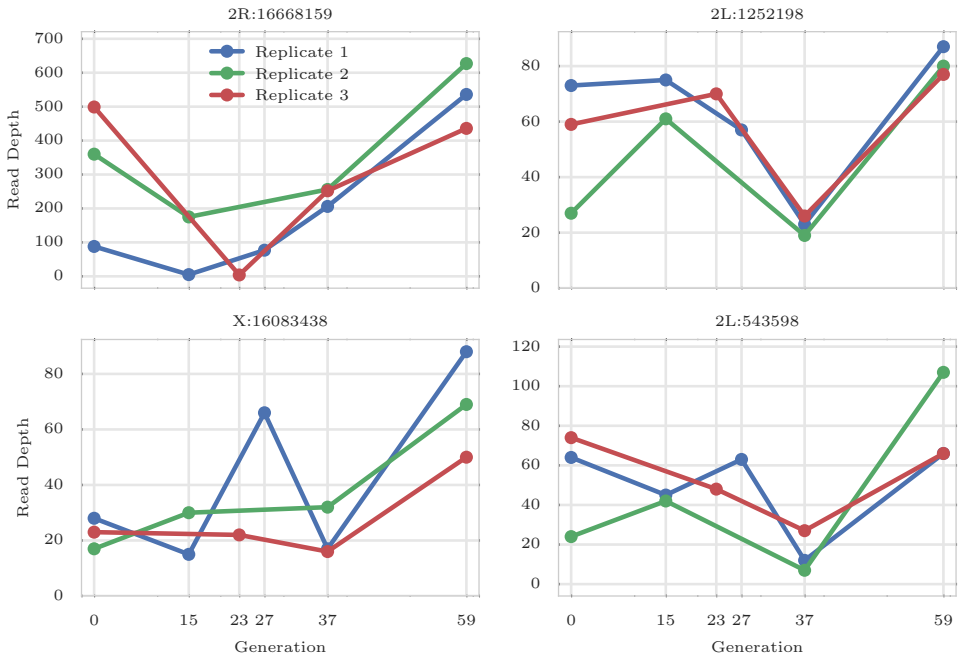
- Small population size ⇒ strong drift
- Partial sweep




- Strong selection ⇒ Short fixation time ⇒ High LD ⇒ Difficult to locate favored allele

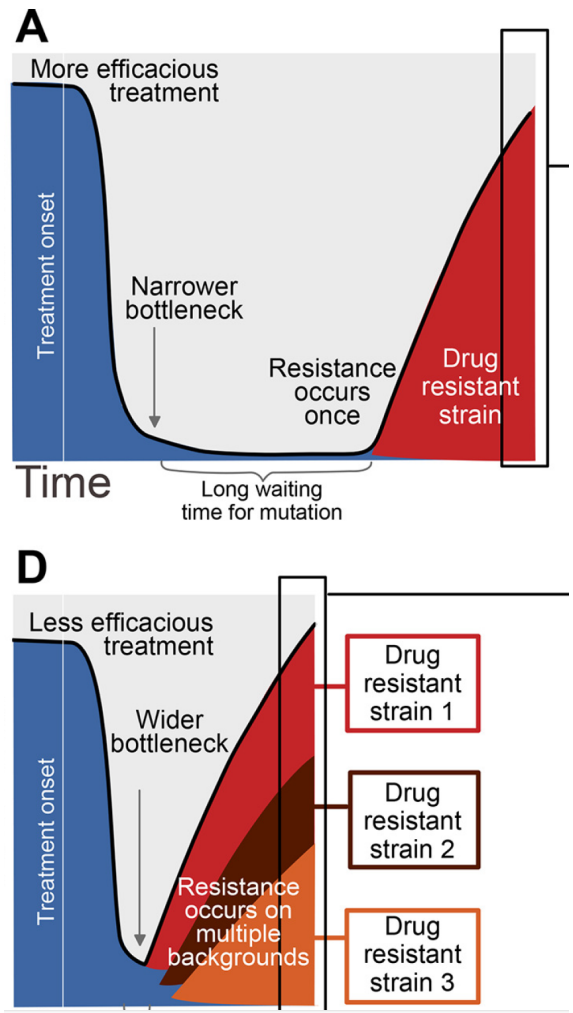
# Challenges (II)

- Pool-seq data: Heterogeneous coverage for a variant



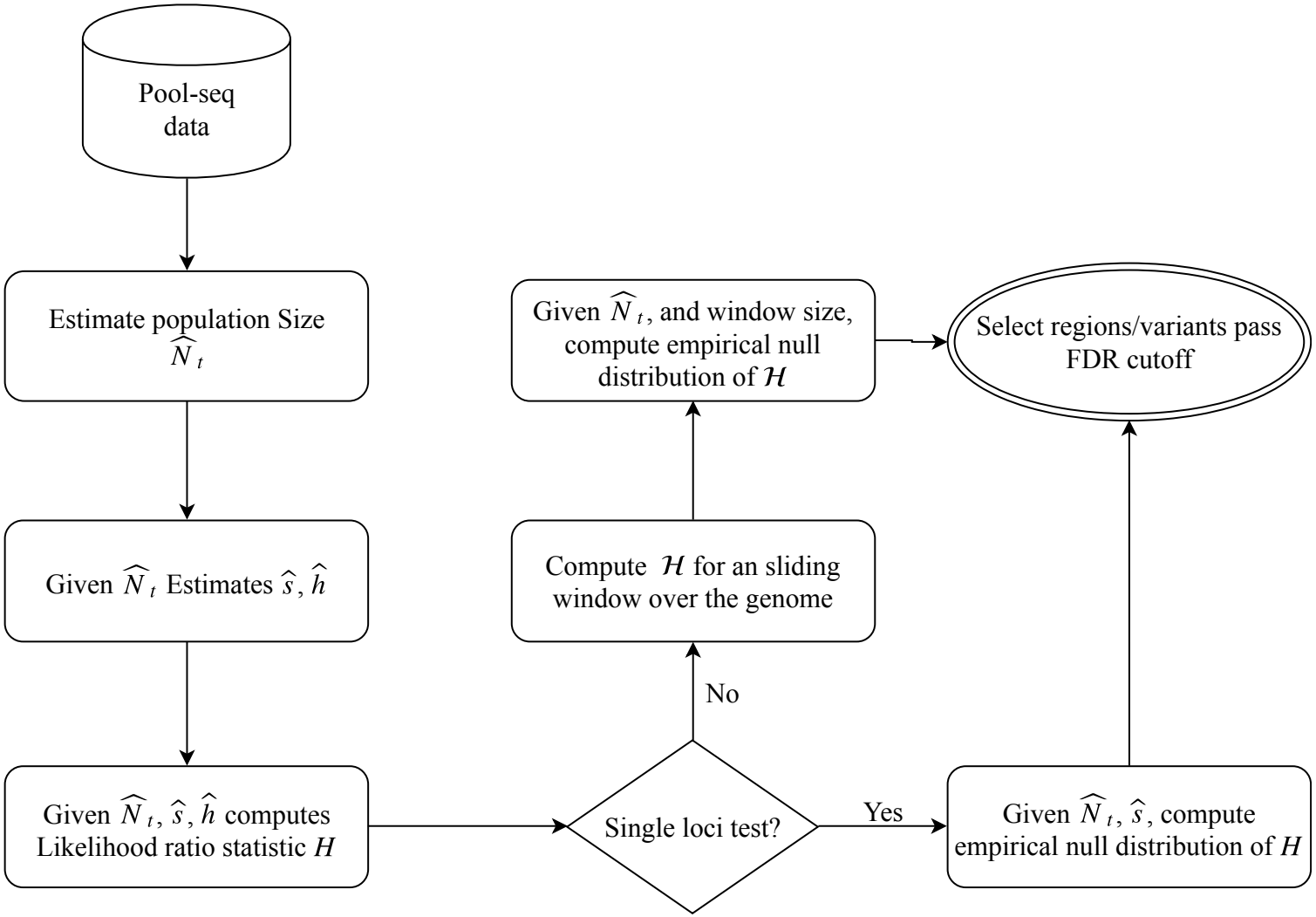
# Challenges (III)

- Selection + Demography




Feder *et al.* 2016

# CLEAR procedure

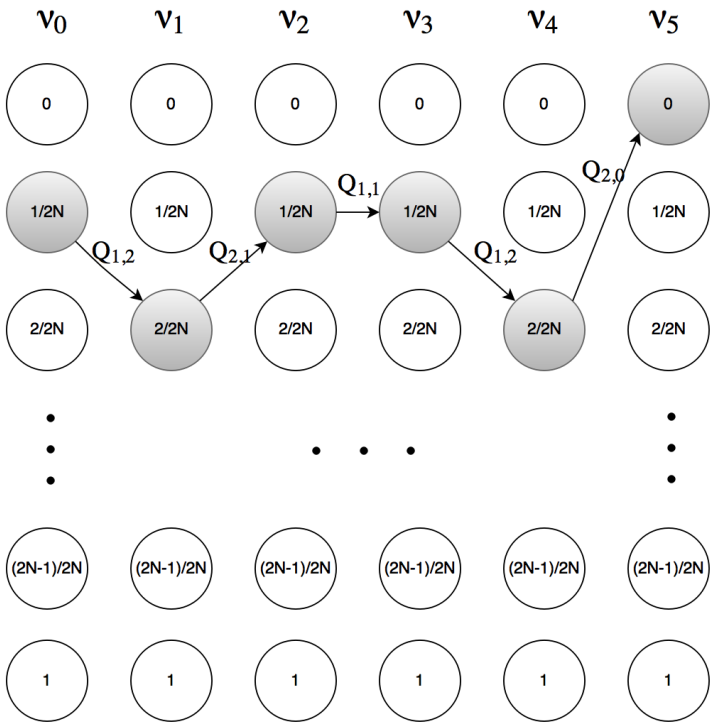


# Simplified Model (I)

- Suppose we have sequenced a whole (diploid, size=$N$) population every generation and exact allele frequency are given.

# Simplified Model (I)

- Suppose we have sequenced a whole (diploid, size=$N$) population every generation and exact allele frequency are given.
- Wright-Fisher Markov chain, computes likelihood of a trajectory for a given $N$ (a $2N \times 2N$ transition matrix $Q$)




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

# Likelihood ratio test

- find $\hat{N}$ and $\hat{s}$ that maximizes likelihood of data.
- compute likelihood ratio, $H$ statistic for each SNP:

$$H = \frac{\text{likelihood of data as if being under selection with } \hat{s}, \hat{N}}{\text{likelihood of data as if being neutral with } \hat{N}}$$

# Model (complete)

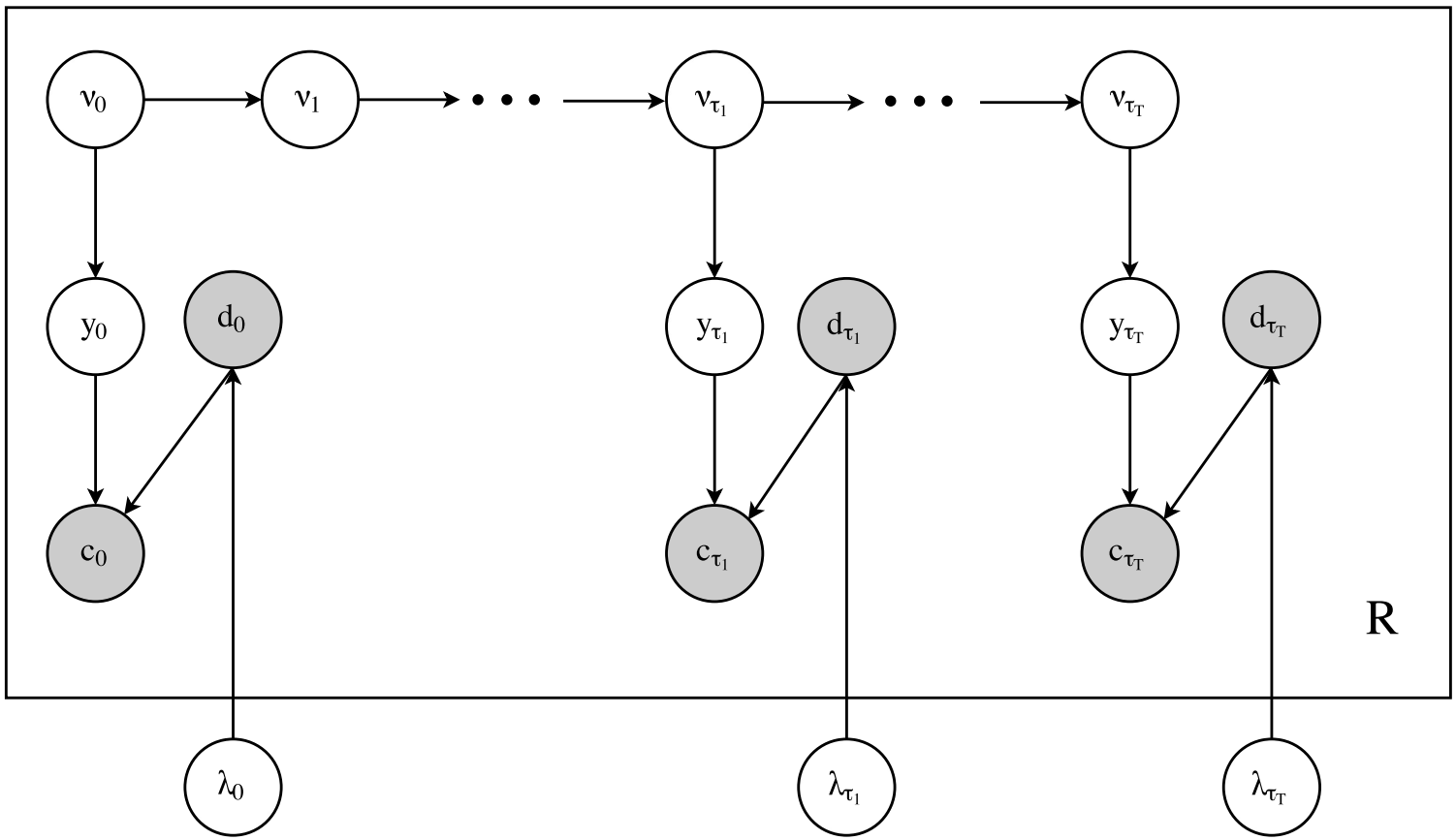
- In reality, population is sequenced after some ($\tau$) generations.
  solution: use $Q^\tau$ in computing likelihoods.

# Model (complete)

- In reality, population is sequenced after some ($\tau$) generations. solution: use $Q^\tau$ in computing likelihoods.
- Allele frequencies are unknown, and depth of each variant can be different, and finite sample is taken for sequencing.

# Model (complete)

- In reality, population is sequenced after some ($\tau$) generations.
  solution: use $Q^\tau$ in computing likelihoods.
- Allele frequencies are unknown, and depth of each variant can be different, and finite sample is taken for sequencing.



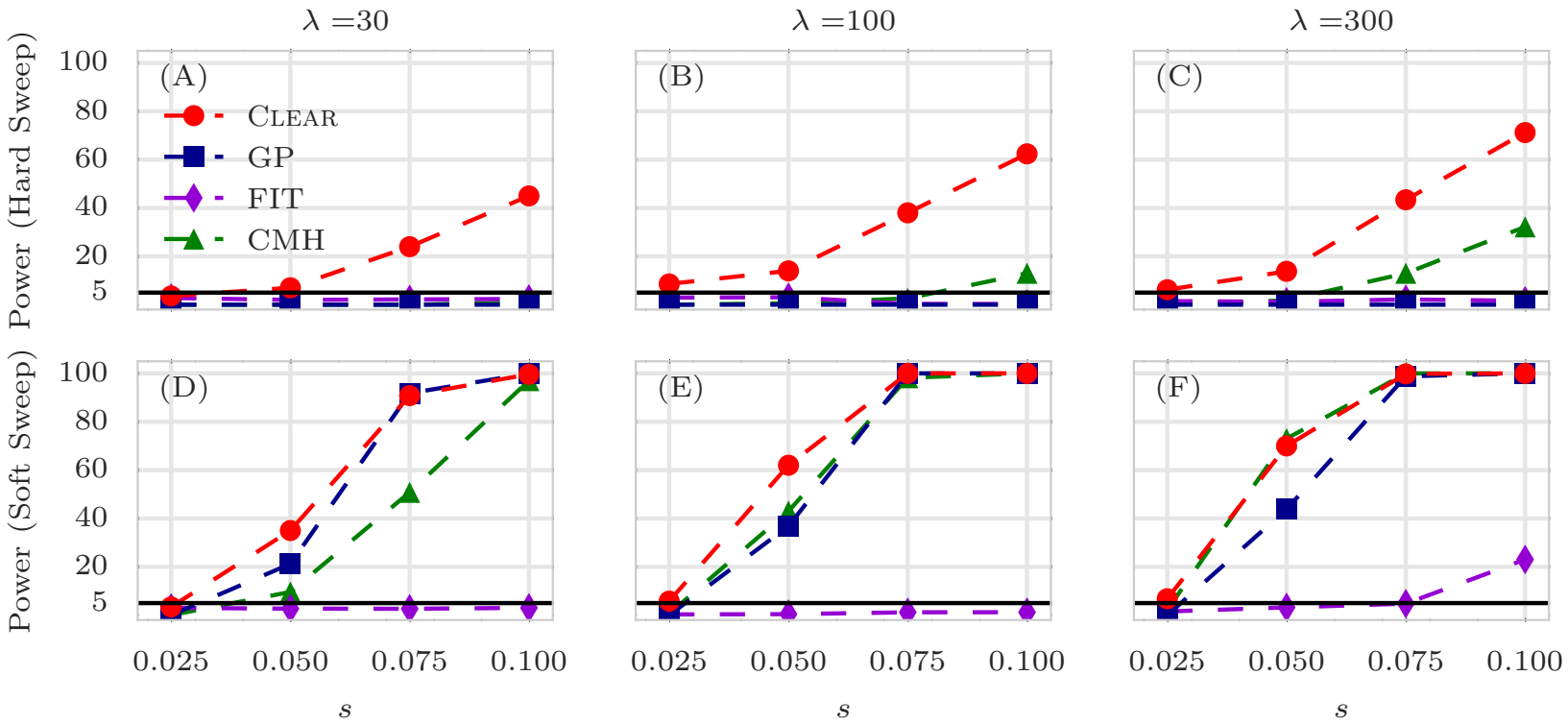
# Composite Likelihood for a Region

- Computing joint likelihoods of SNPs is infeasible (haplotypes are required) and intractable (requires estimating covariance).
- A heuristic is to compute composite (aka, pseudo) likelihood of the region $L$ to reduce false-positives

$$\mathcal{H} = \frac{1}{|L|} \sum_{\ell \in L} H_\ell$$

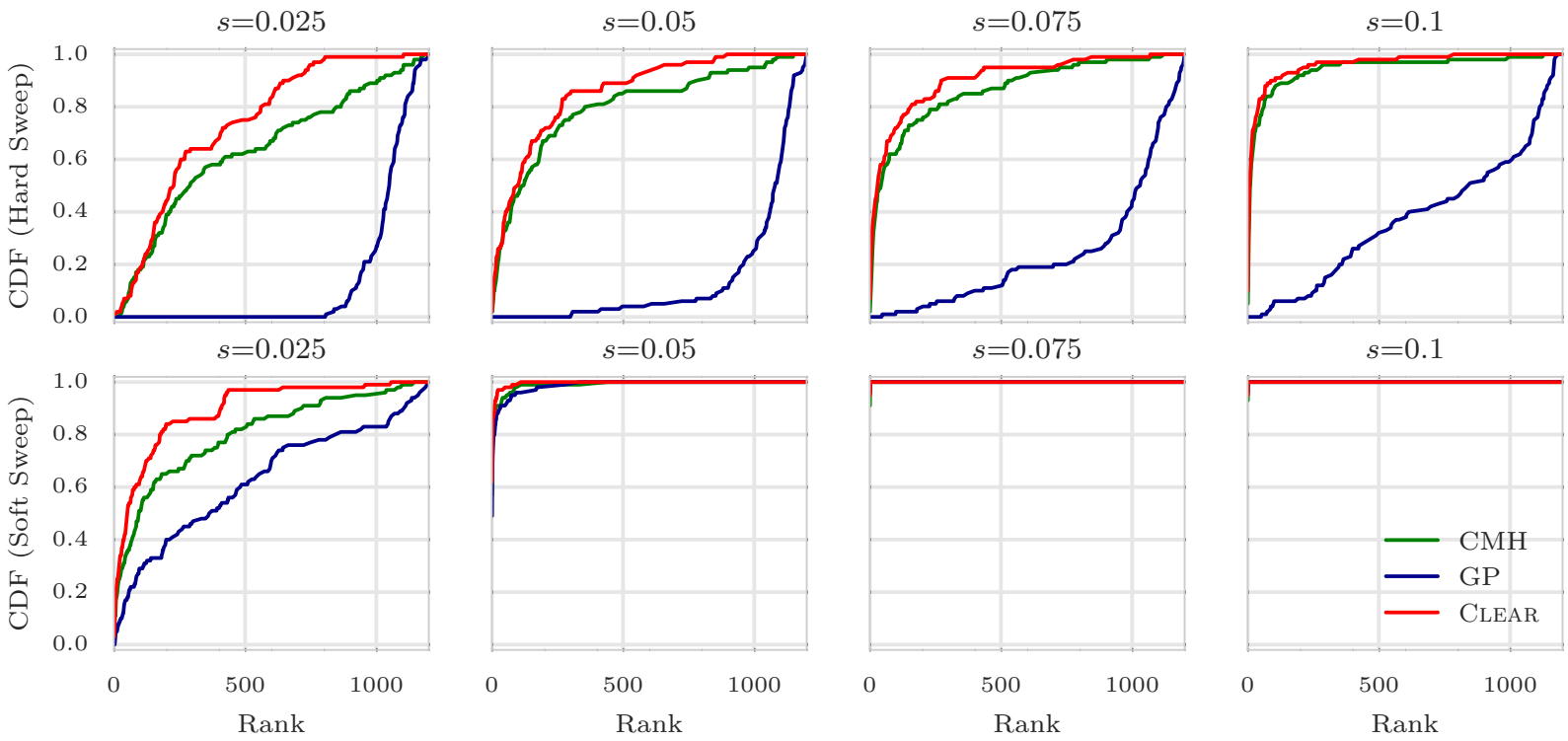
# Performance in Detecting Regions under Selection

Each point represent power (TPR when FPR$\leq$0.05) of detection in 1000 simulations (500 neutral, 500 selection) of a 50Kbp window, for different coverages.



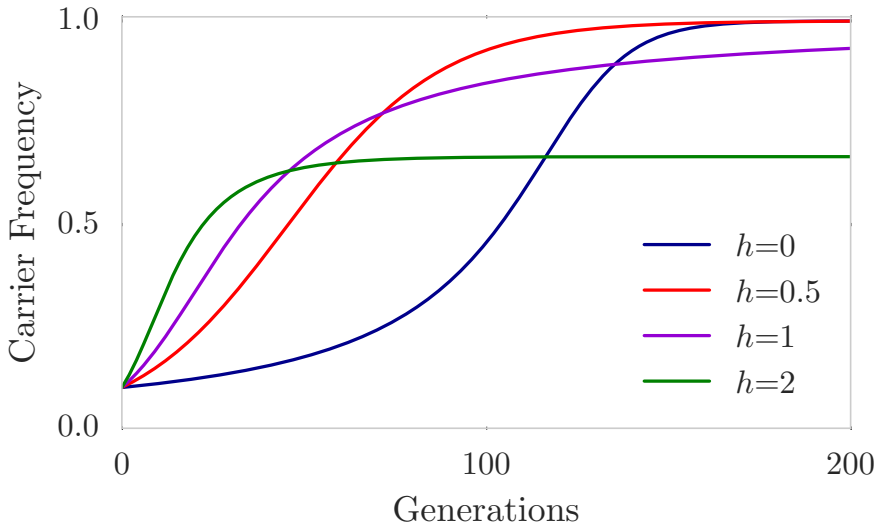
# Localizing favored allele

Each curve depicts cumulative distribution of the rank of favored allele among ($\approx 1150$) variants, in 500 simulations.



# Estimating parameters (I)

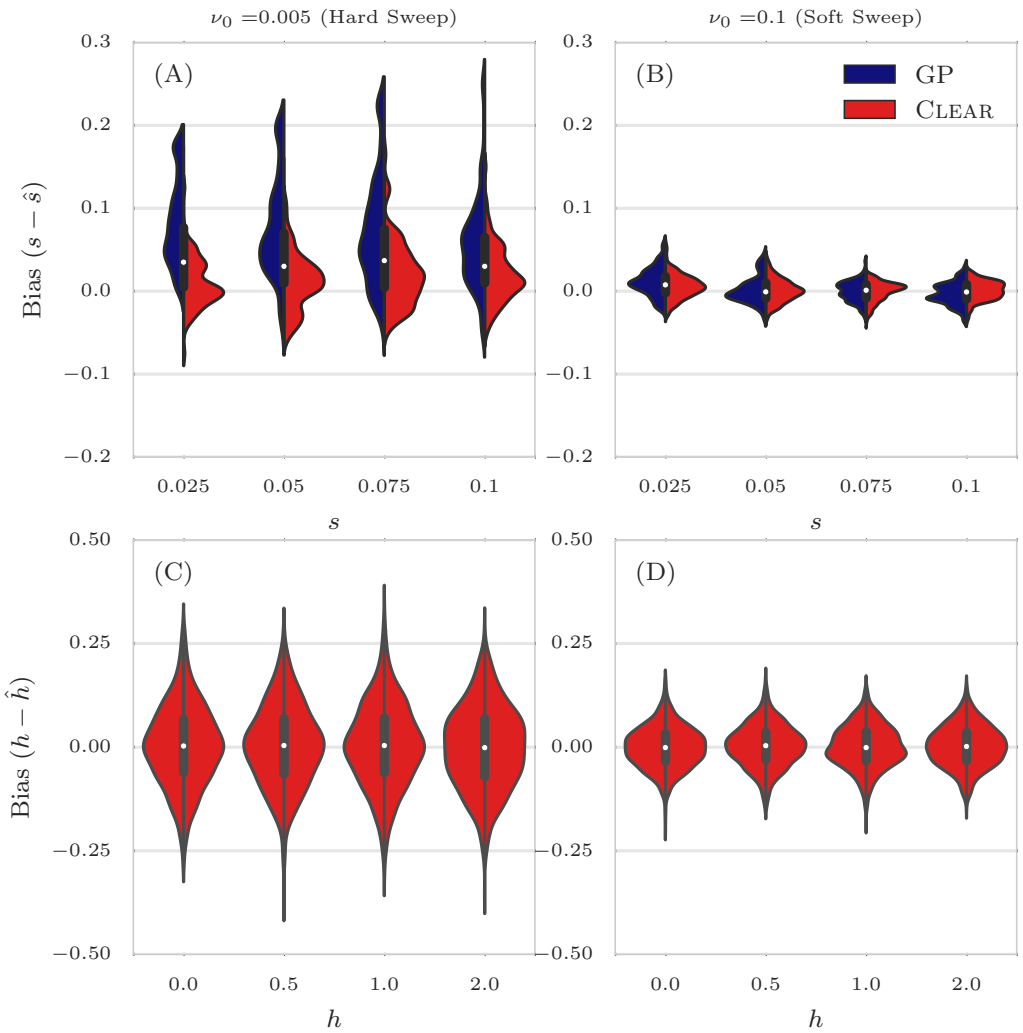
Our model estimates strength of selection $s$ and overdominance $h$ parameter for each variant.



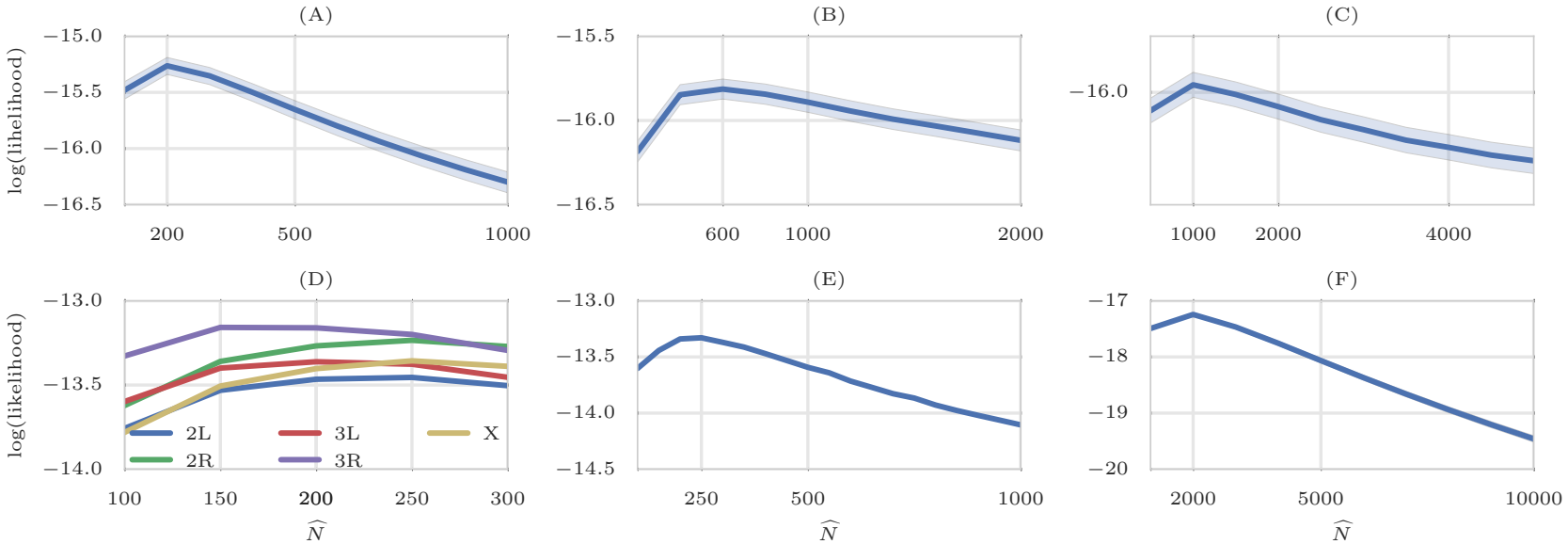

- $h = 0$ : recessive adaptive allele
- $h = 0.5$ : directional selection
- $h = 1$: dominant adaptive allele
- $h > 1$ :overdominance

# Estimating parameters (II)

Distribution of bias of parameters in 500 simulations.



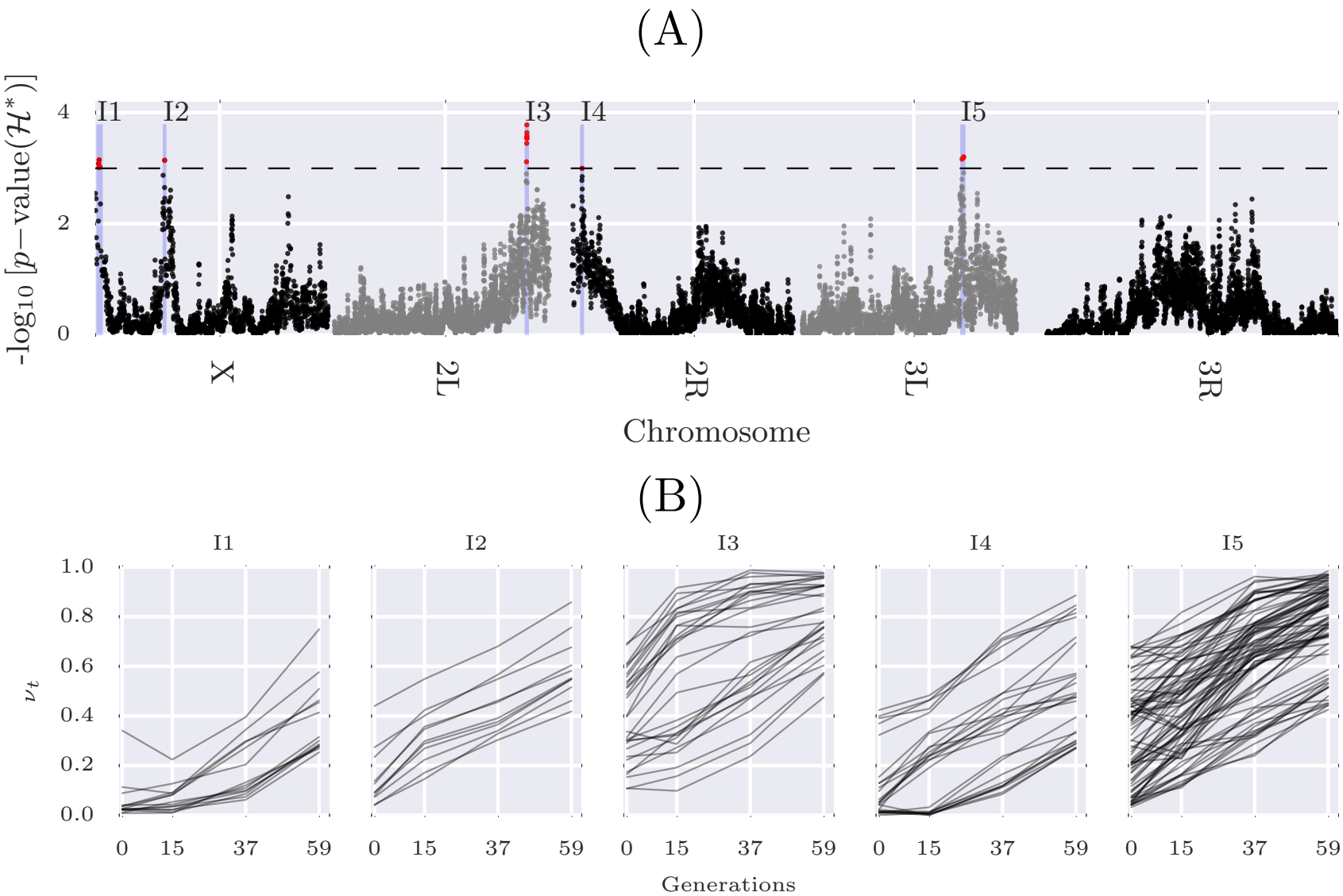
# Estimating parameters (III)



# Analysis of real data

- A population of *D. melanogaster* is evolved for 59 generations, under alternative hot and cold temperatures.
- Coverage is different at generations and samples are not synchronized.
- Genome scan for sliding window size=50Kbp, steps=10Kbp
- $\hat{N} = 200$

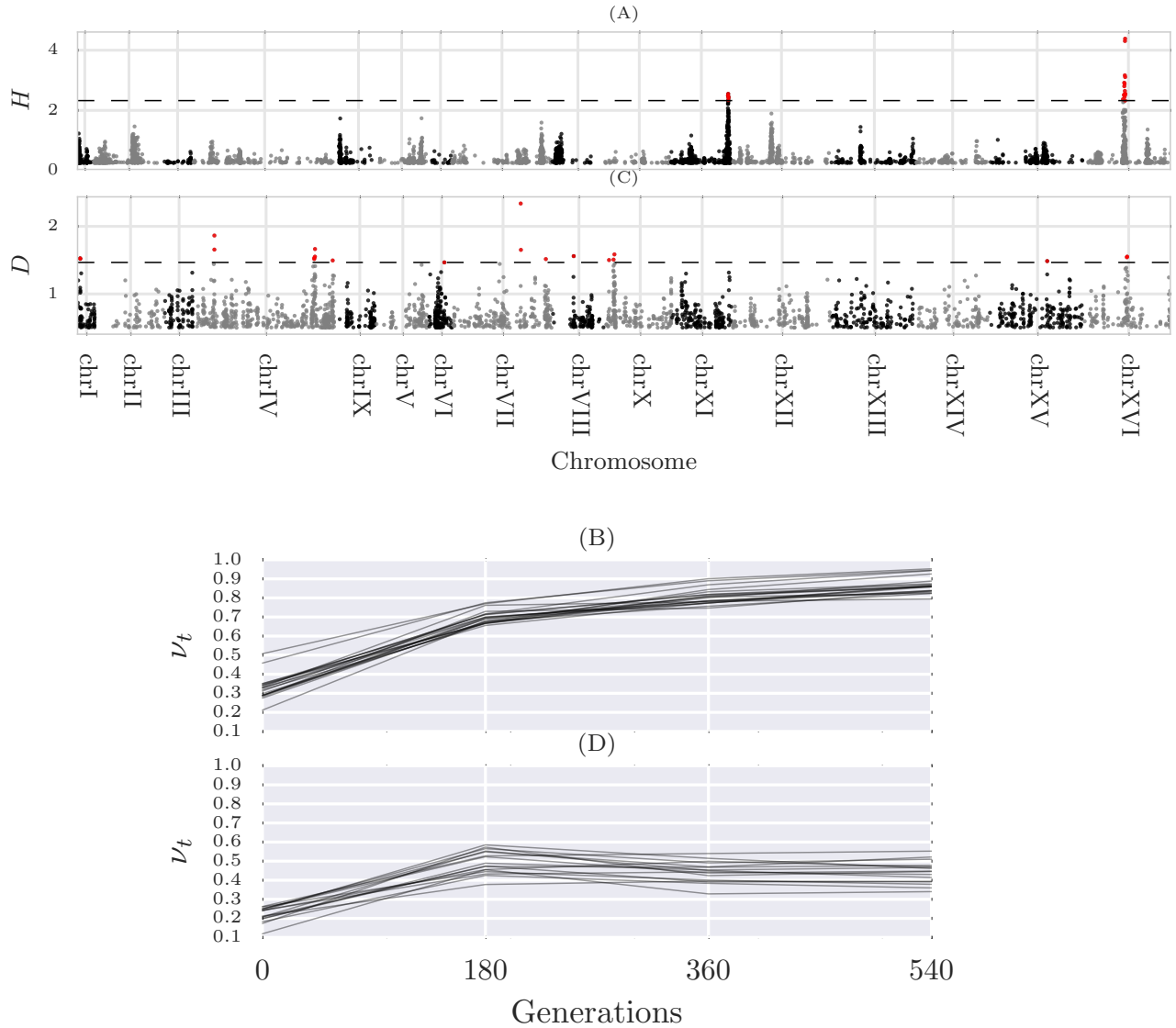
# *D. melanogaster*



# Outcrossing Yeast populations

- 12 replicates of Yeast populations (census size $10^7 - 10^9$) are E&Red for 540 generations.
- $\hat{N} = 2000$
- two regions violating FDR cutoff are found.

# Outcrossing Yeast populations



# Discussion

- An efficient method for analyzing full time-series read-count data is proposed.

# Discussion

- An efficient method for analyzing full time-series read-count data is proposed.
- By computing composite likelihood $\mathcal{H}$ statistic is more robust to false positives.

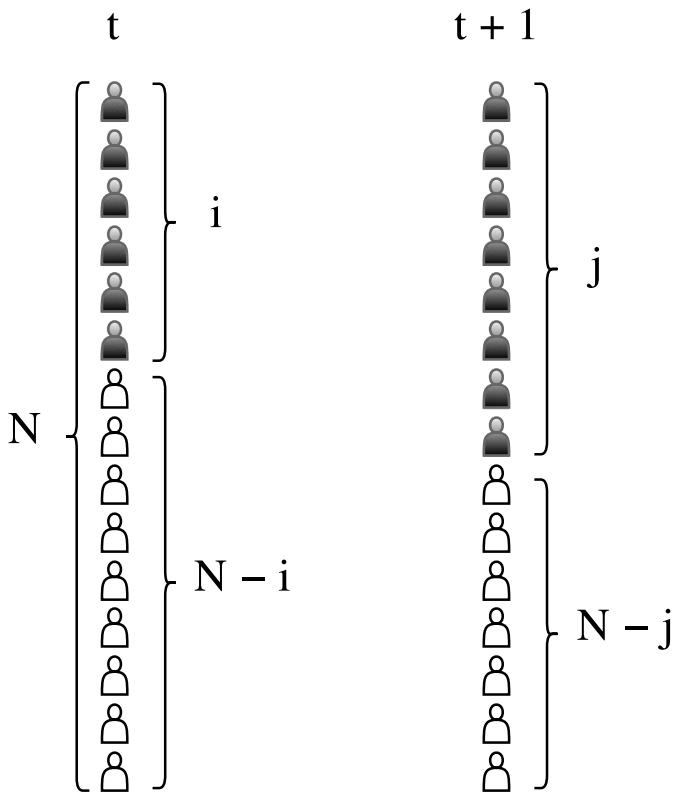
# Discussion

- An efficient method for analyzing full time-series read-count data is proposed.
- By computing composite likelihood $\mathcal{H}$ statistic is more robust to false positives.
- We can infer demographic changes as well as selection for and experiment.

Thanks!

# Modeling genetic drift via Binomial sampling

- Drift: rate of sampling remain constant $\Pr(i \to j) = B(j; N, i/N)$



# Binomial Sampling with Selection

- In selection, we sample favored allele proportional to $1 + s$, and the alternate allele with weight 1. $\Pr(i \to j) = B(j; N, k/N)$