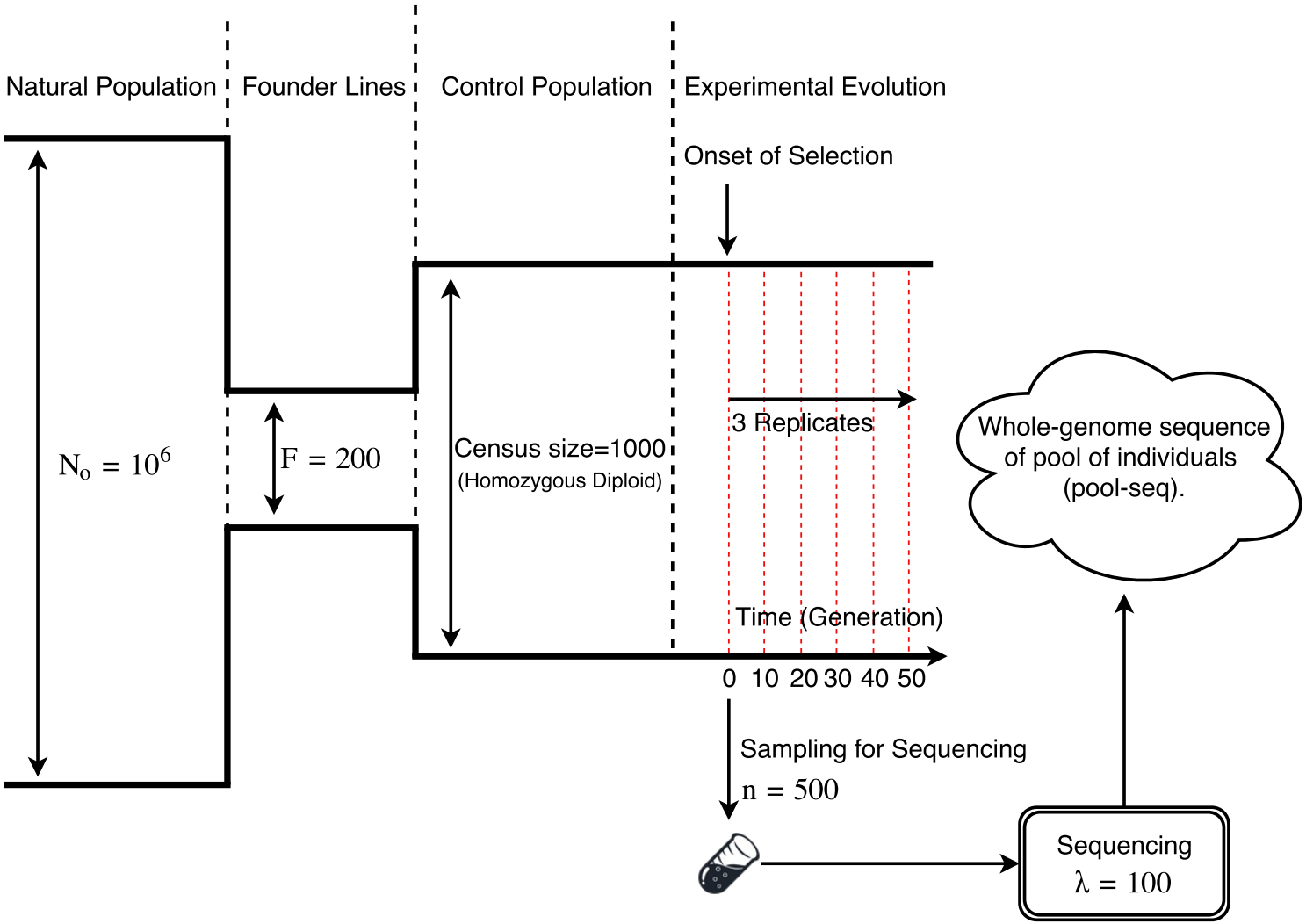
# Identifying Selection in Experimental Evolution

Arya Iranmehr
airanmehr@ucsd.edu

Bafna Lab
University of California, San Diego

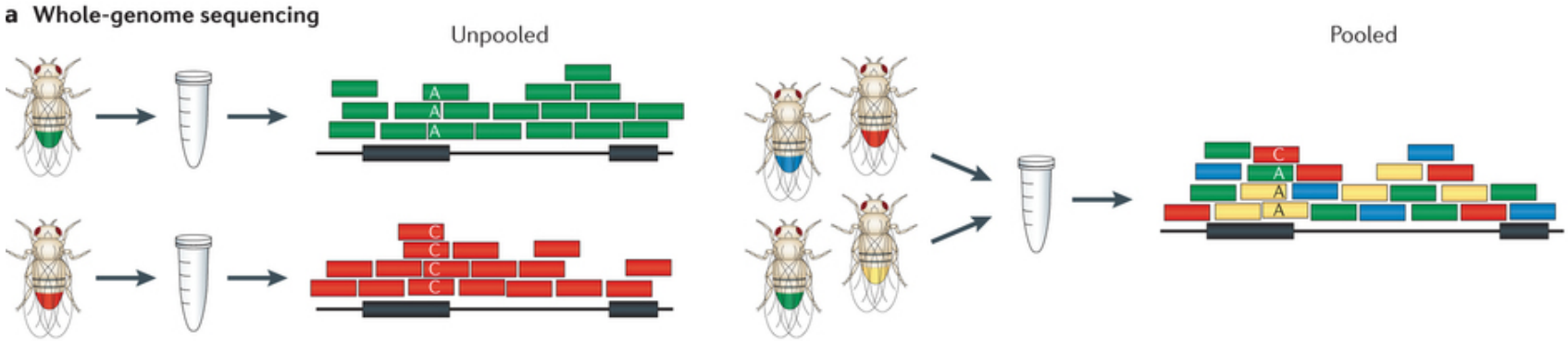
March, 2017

# An experiment design for *D. melanogaster*



# Whole-Genome Whole-Population Sequencing

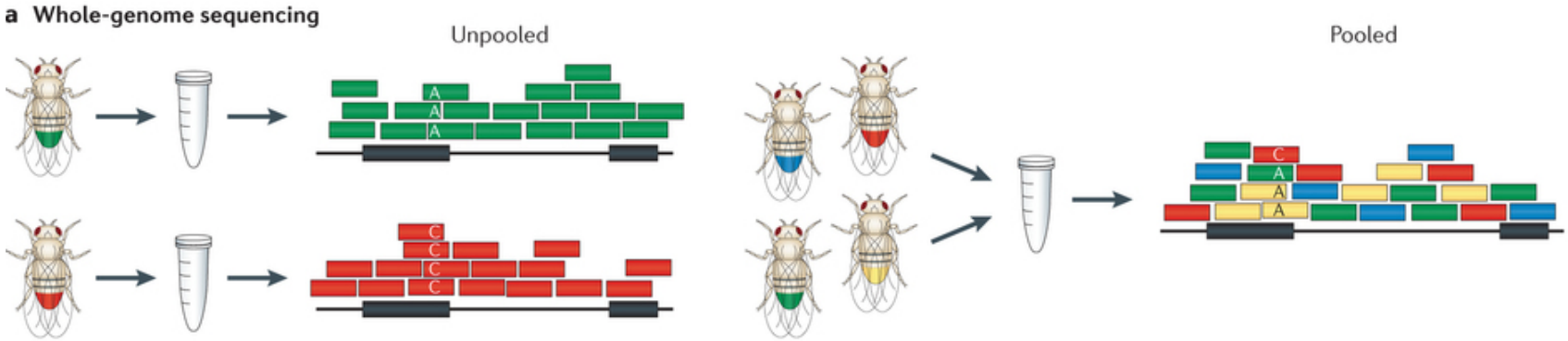
- Pooled-Sequencing




Nature Reviews Genetics 15, 749-763 (2014)

# Whole-Genome Whole-Population Sequencing

- Pooled-Sequencing




Nature Reviews Genetics 15, 749-763 (2014)

- Implication: only population allele frequency can be computed.

# *D. melanogaster* vs Bacteria

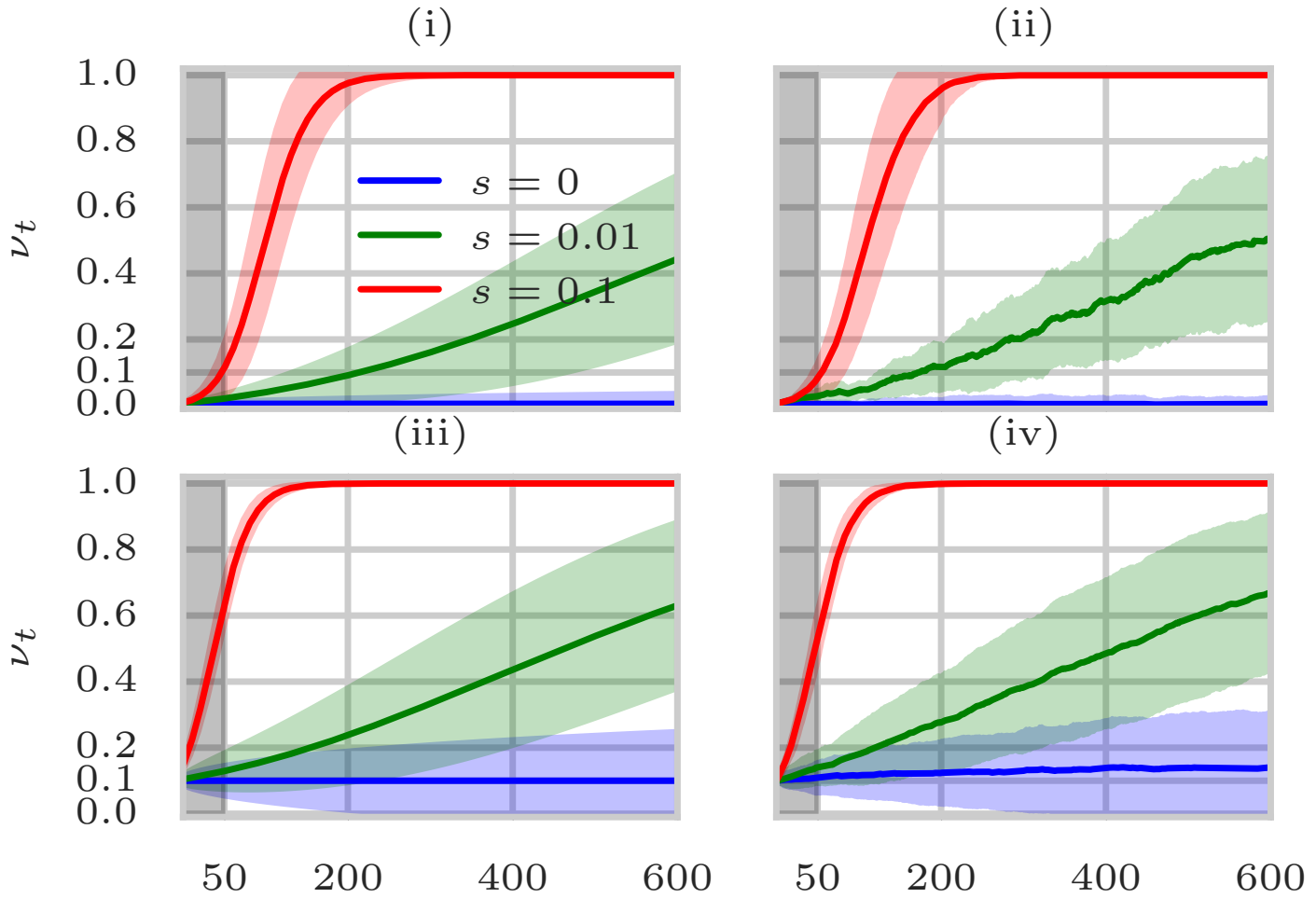
(i) Population size: $N_{Drosophila} \ll N_{Bacteria}$
Among other consequences: Mechanism of adaptation is standing variation in *D. melanogaster* while it is *de novo* mutation in Bacteria.

(ii) Reproduction: *D. melanogaster* has sexual reproduction (with crossovers) that helps localizing selection.

(iii) In both cases we are interested in detecting partial sweeps.

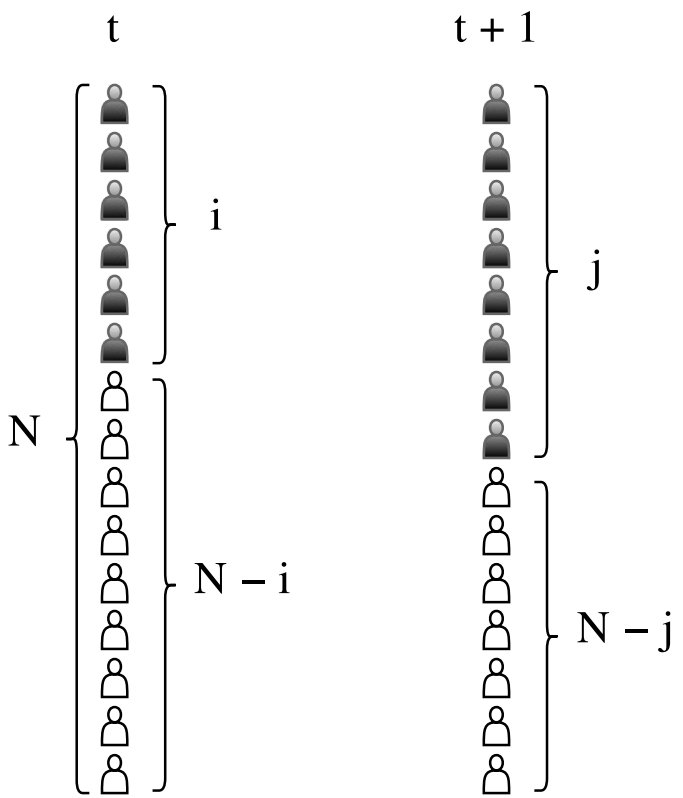
# Dynamic of population allele frequency

under different initial conditions and *selection strengths* frequency change differently



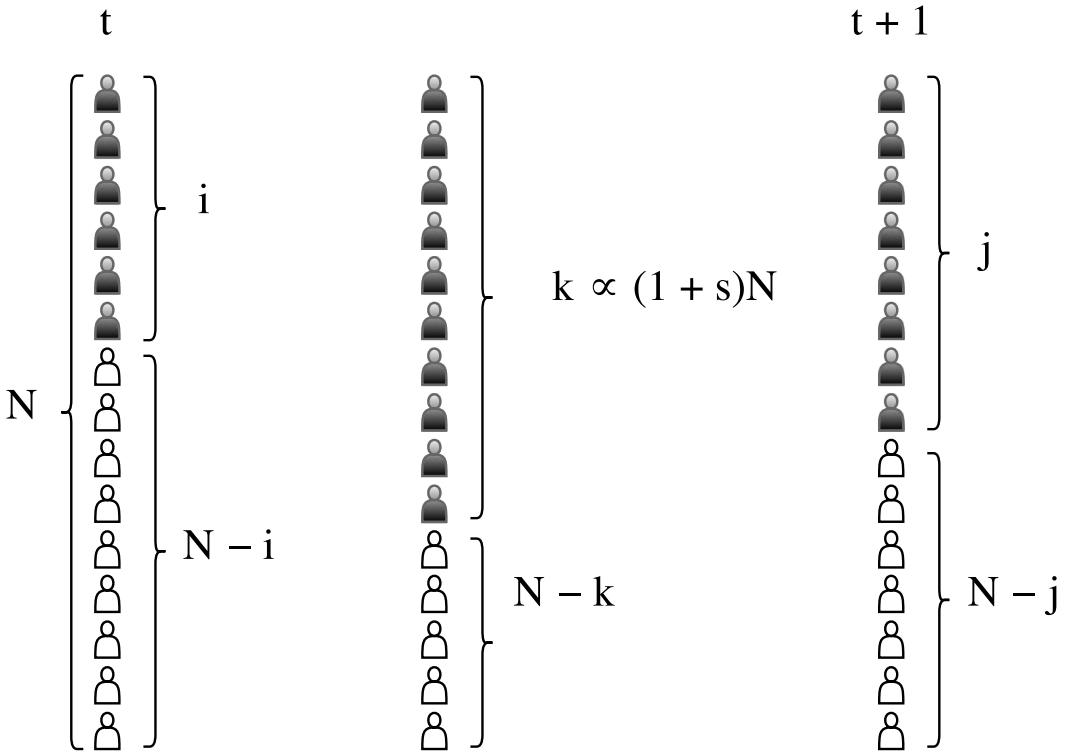
# Binomial Sampling

- In a finite population, we can model change in frequency of an allele 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. $\Pr(i \to j) = B(j; N, k/N)$

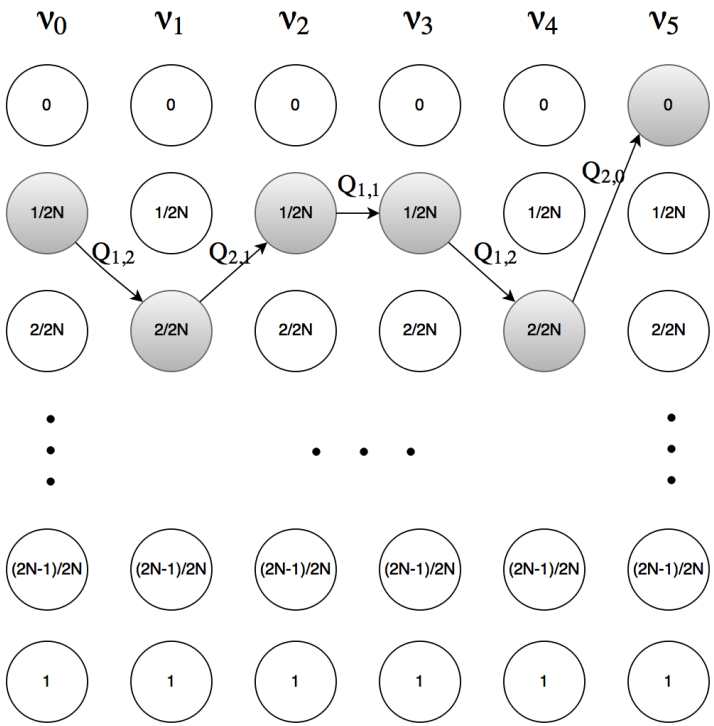


# Simplified Model (I)

- 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.
- A Markov chain, can compute likelihood of a trajectory for a given $N$ and $s$ (a $N \times N$ 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, $M$ statistic for each SNP:

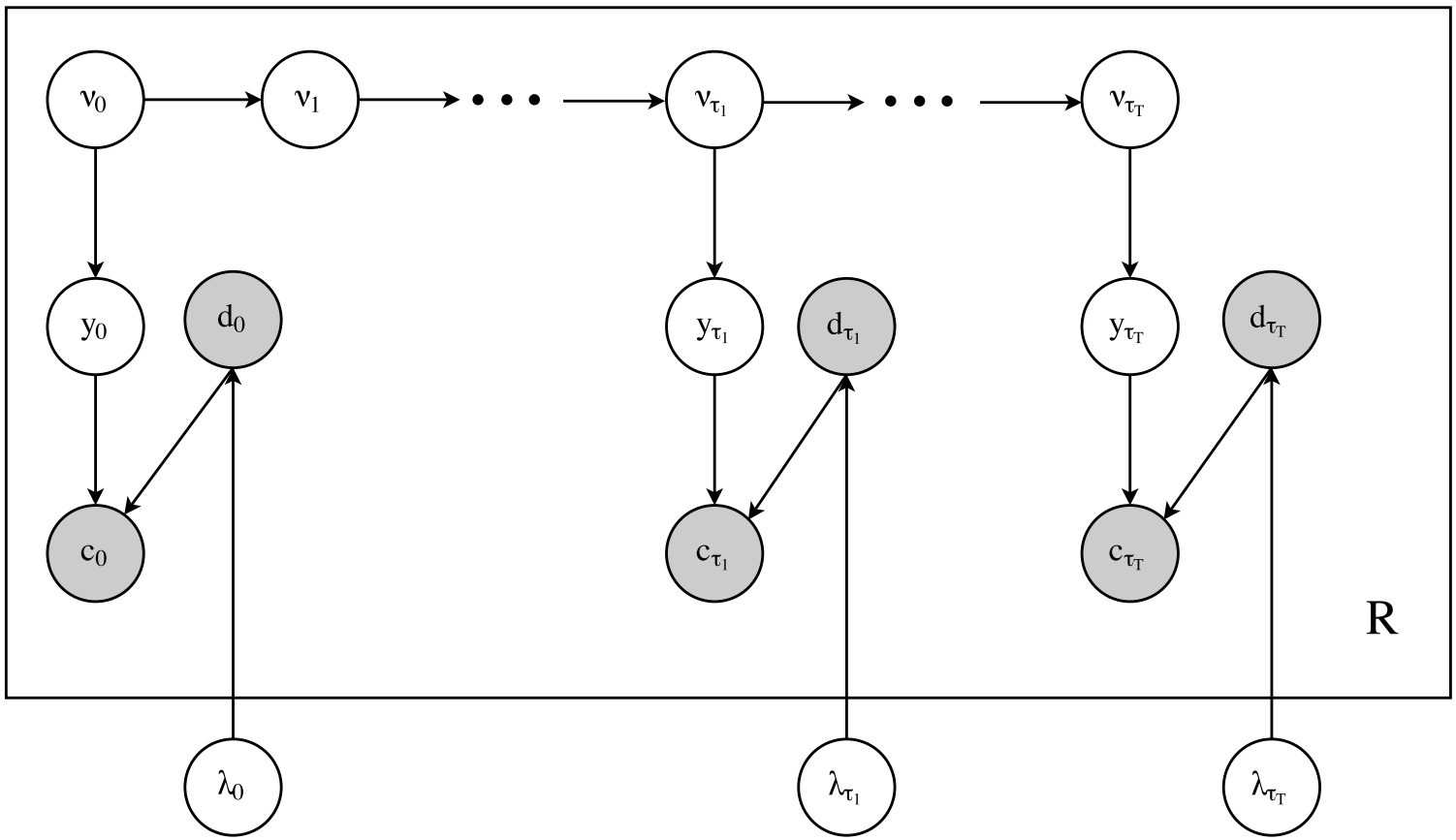
$$M = \frac{\text{likelihood of data as if being under selection with } \hat{s}, \hat{N}}{\text{likelihood of data as if being neutral with } \hat{N}}$$

# 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.



# Composite Likelihood for a Region (I)

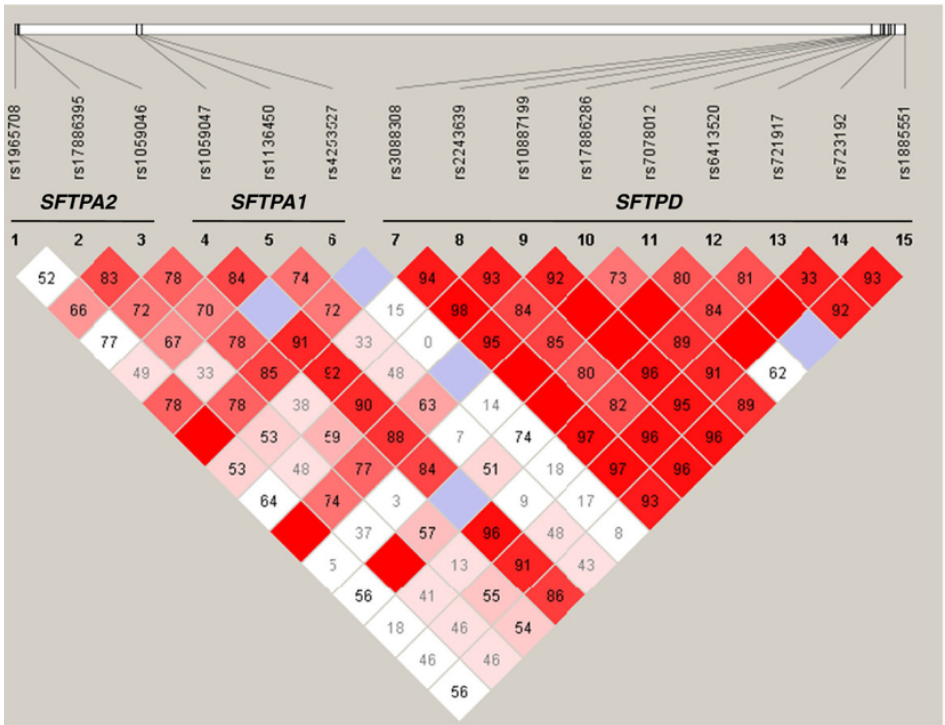
- So far we developed log-odds ratio statistics $M$ (frequency data) and $H$ (read count data) for each variant.

# Composite Likelihood for a Region (I)

- So far we developed log-odds ratio statistics $M$ (frequency data) and $H$ (read count data) for each variant.
- For a small region with $L$ variants we can simply take the max score in the region, which is prone to false positives.

# Composite Likelihood for a Region (I)

- So far we developed log-odds ratio statistics $M$ (frequency data) and $H$ (read count data) for each variant.
- For a small region with $L$ variants we can simply take the max score in the region, which is prone to false positives.
- We know that nearby variants can be correlated, esp. when selection is going on




Crit Care. 2014 Jun 20;18(3):R127

# Composite Likelihood for a Region (II)

- Computing joint likelihoods of SNPs is infeasible (haplotypes are required) and intractable (requires estimating covariance).

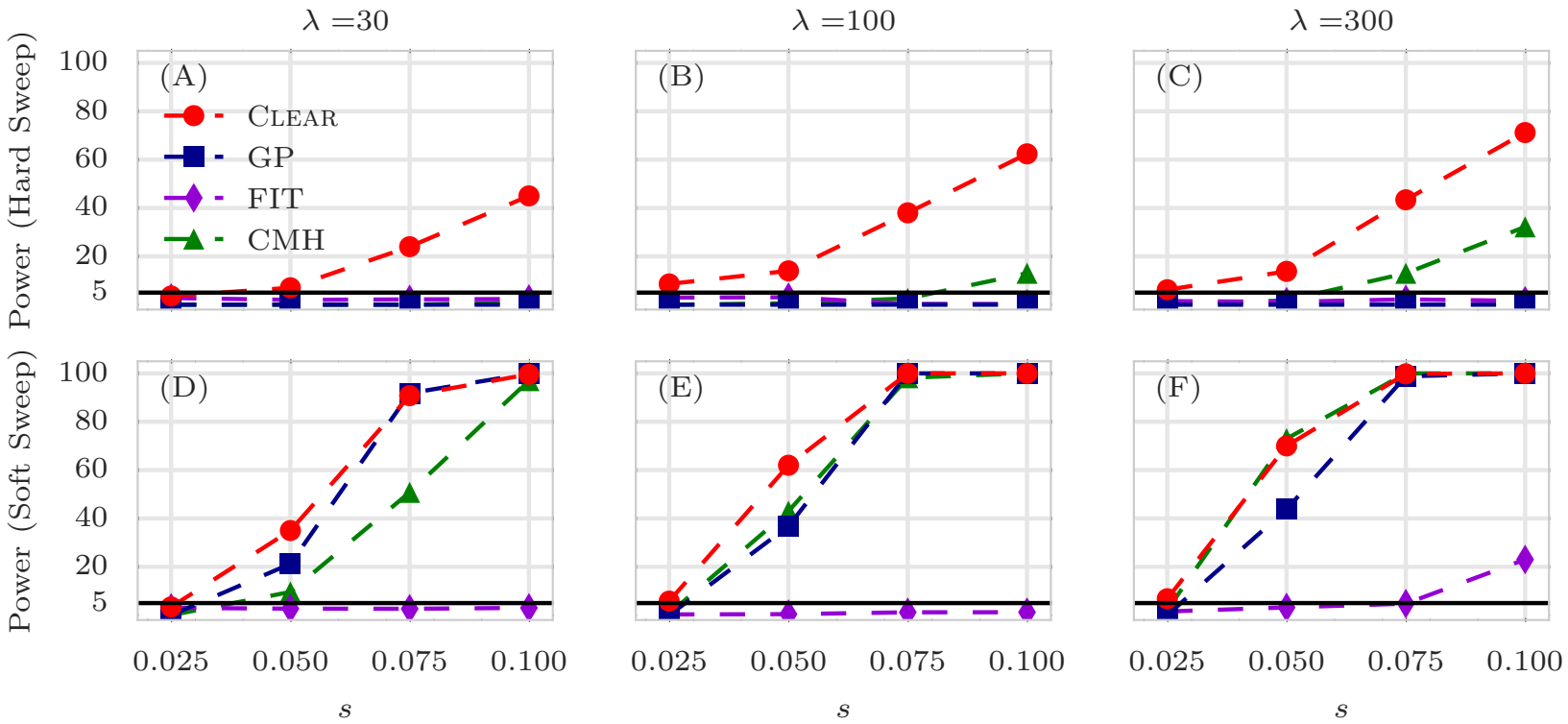
# Composite Likelihood for a Region (II)

- 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.



# Detecting regions under selection: Observations

(i) Provides better and much robust performances to change of coverage.

# Detecting regions under selection: Observations

(i) Provides better and much robust performances to change of coverage.

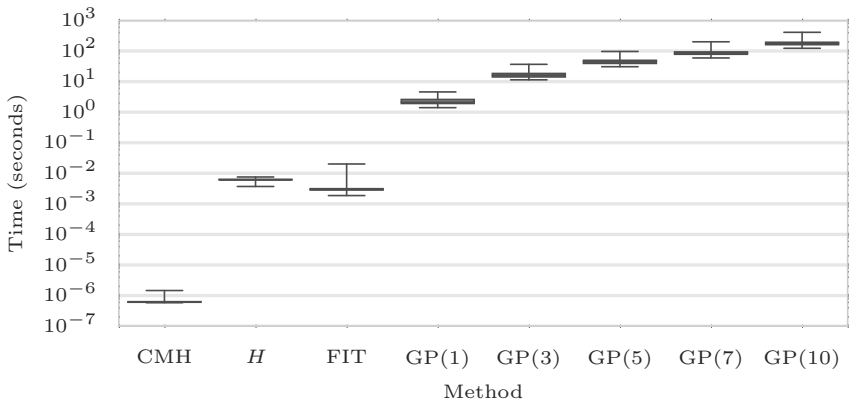
(ii) It can detect well even when coverage is low, i.e., favored allele frequency (1/200 in hard sweep) is below accuracy of sequencing (1/30).

# Detecting regions under selection: Observations

(i) Provides better and much robust performances to change of coverage.

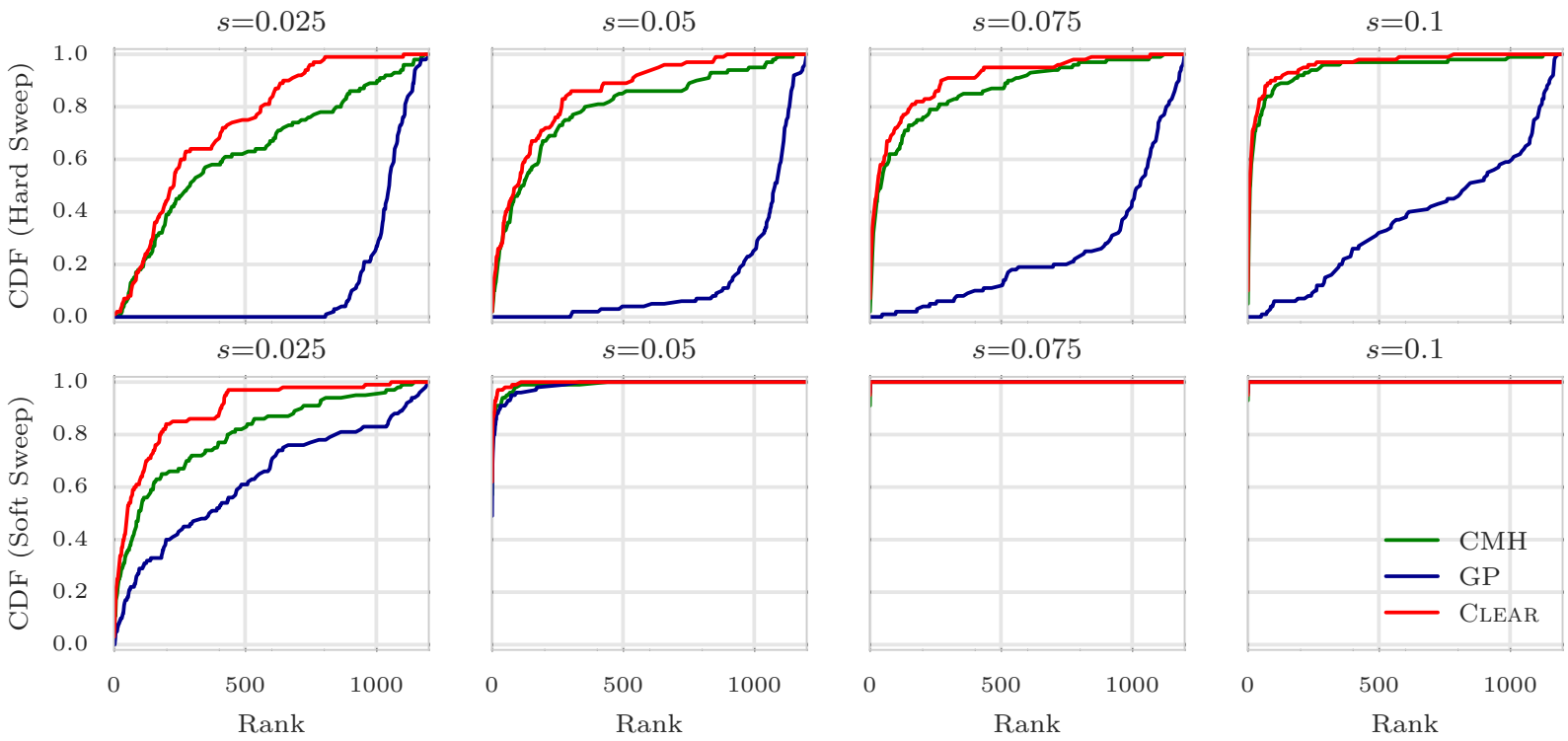
(ii) It can detect well even when coverage is low, i.e., favored allele frequency (1/200 in hard sweep) is below accuracy of sequencing (1/30).

(iii) Run time is better or comparable with others.



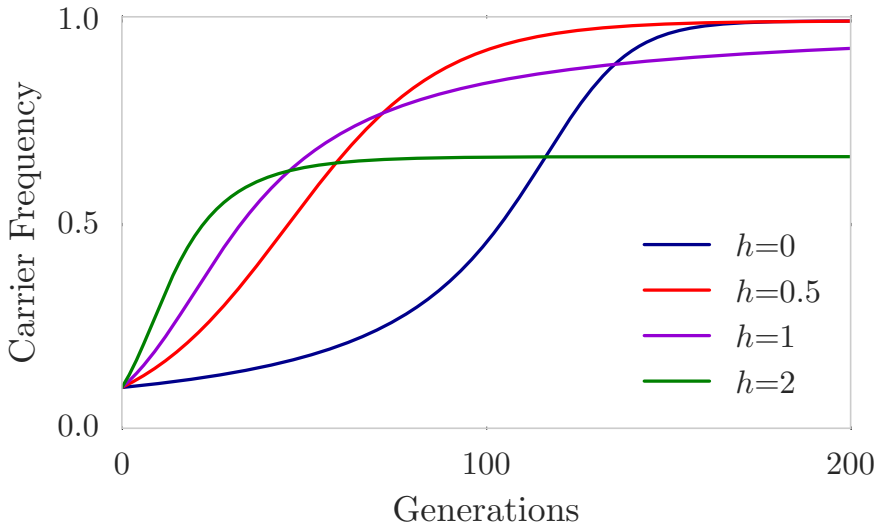
# Localizing favored allele

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



# Estimating parameters (I)

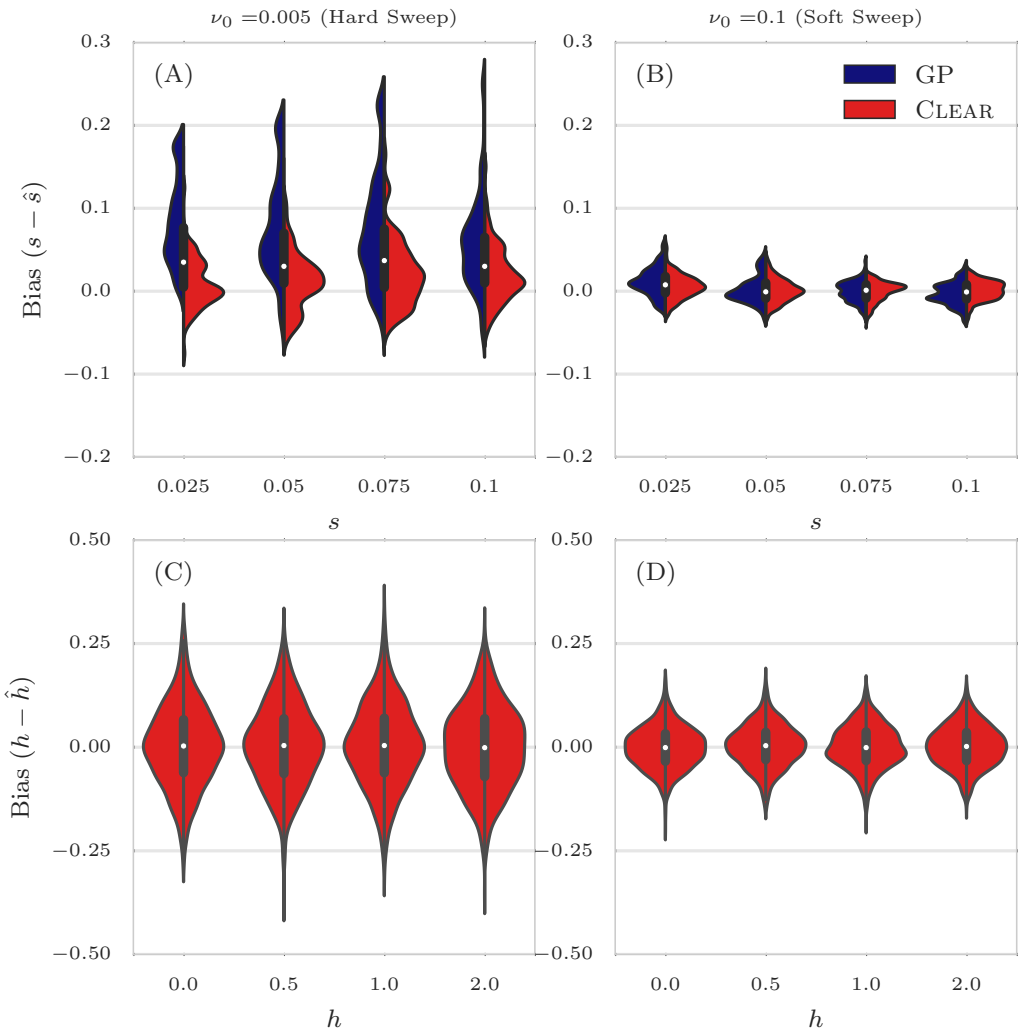
Our model estimates strength of selection $s$ and overdominance $h$ parameter for each variant.




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

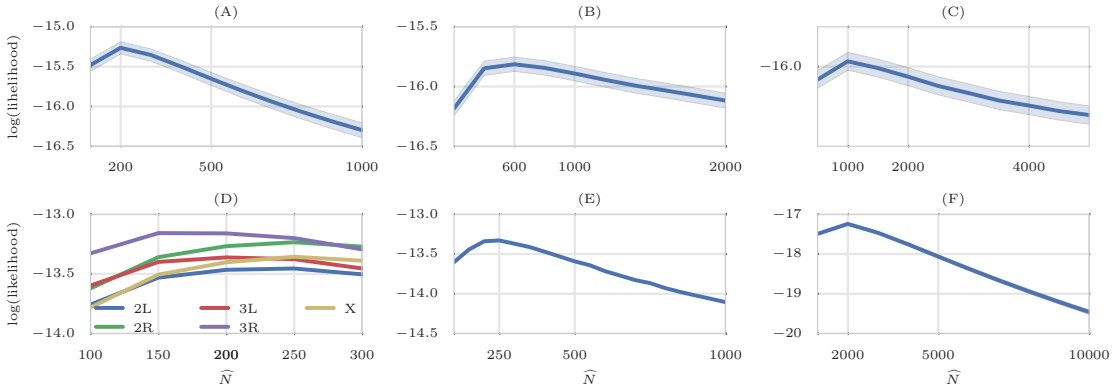
# Estimating parameters (II)

Distribution of bias of parameters in 500 simulations.



# Estimating parameters (III)

Assuming majority of the variants evolving neutrally, we can fit population size $N$ on neutral model, i.e. $Q(0, 0, 2N)$



# Hypothesis Testing

The statistical procedure involves:

(i) Estimating population size, $\widehat{N}$, over the whole genome.

(ii) Estimating selection parameters for given $\widehat{N}$

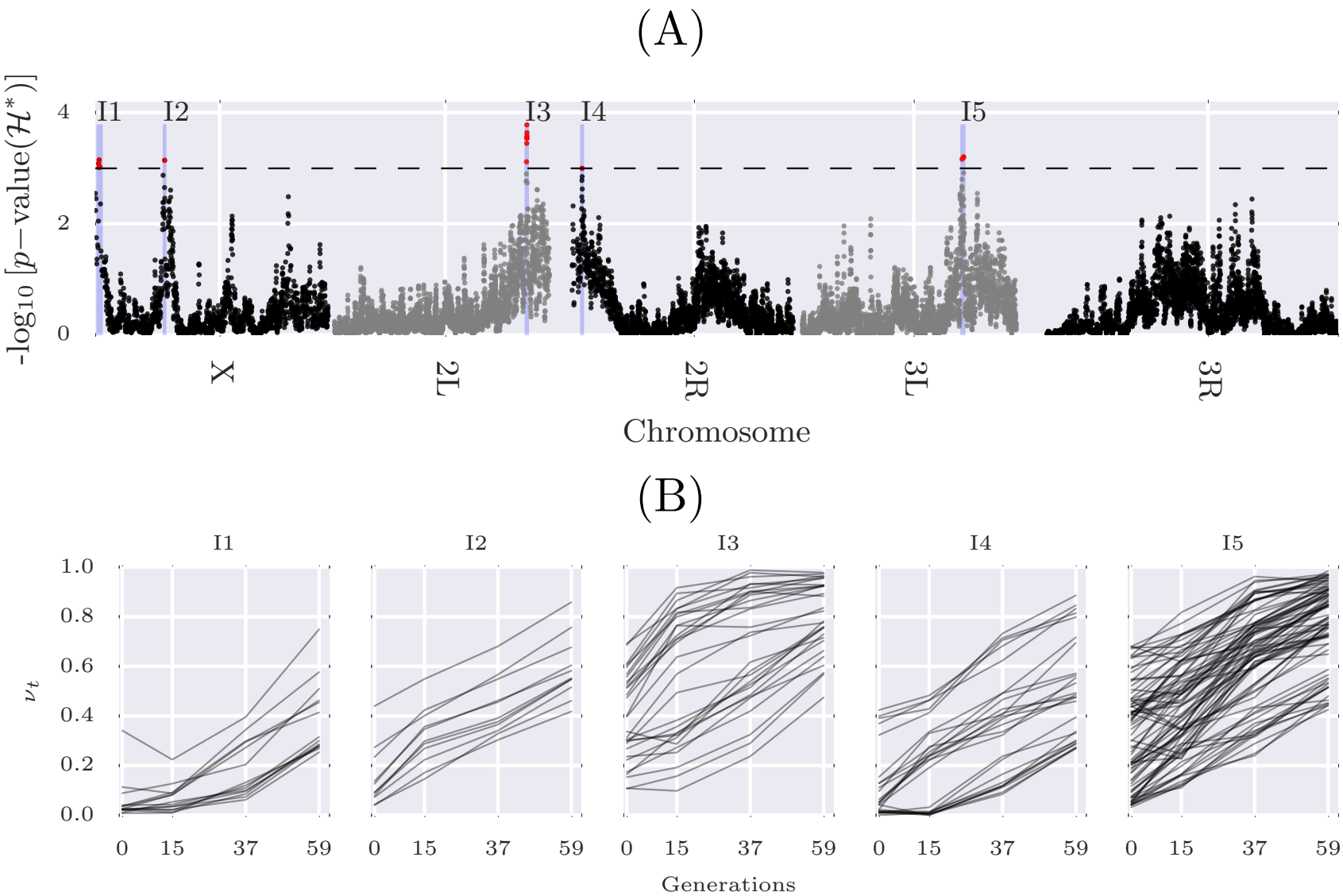
(iii) Computing likelihood statistics.

(iv) Hypothesis testing:The null distribution of likelihood ratio statistics are computed on a set of single locus drift simulations with population size of $\widehat{N}$. $p$-values and FDR is computed accordingly.

# 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
- $\hat{N} = 200$

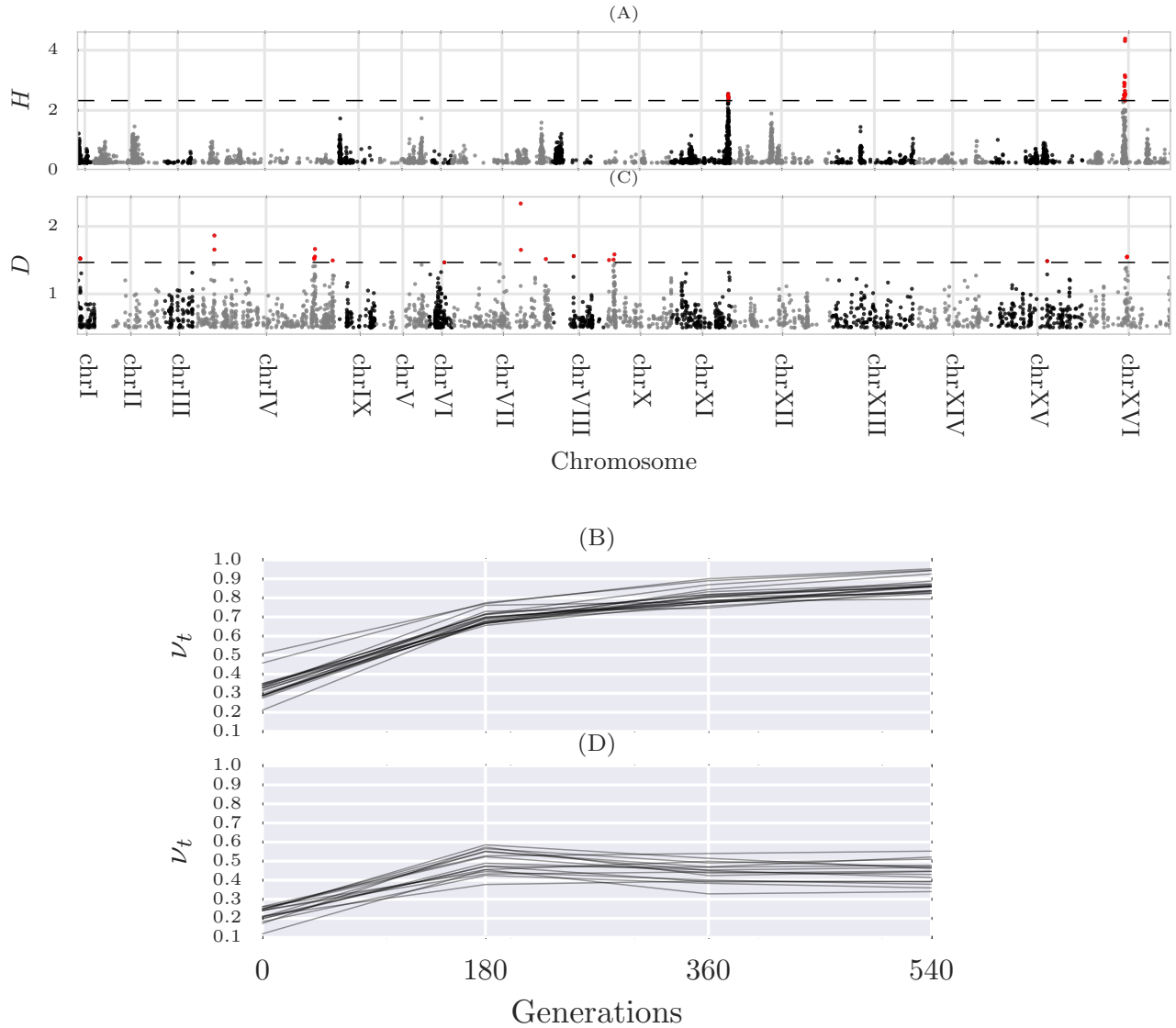
# *D. melanogaster*



# 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!