Identifying Selection in Experimental Evolution

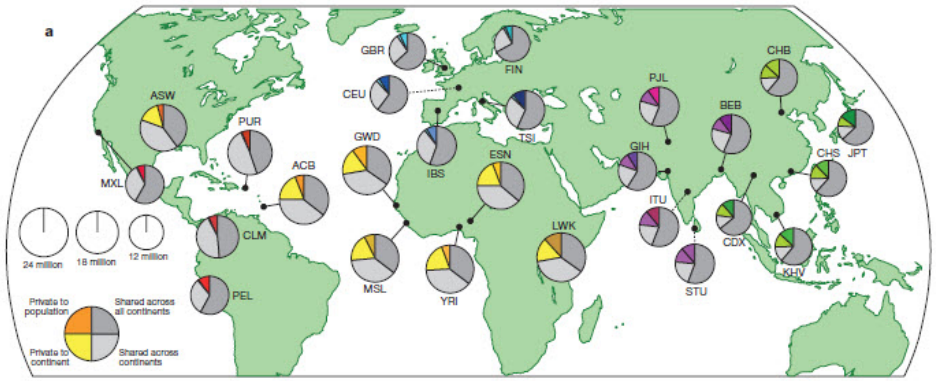
Arya Iranmehr
airanmehr@ucsd.edu

Bafna Lab
University of California, San Diego

January, 2017

Introduction

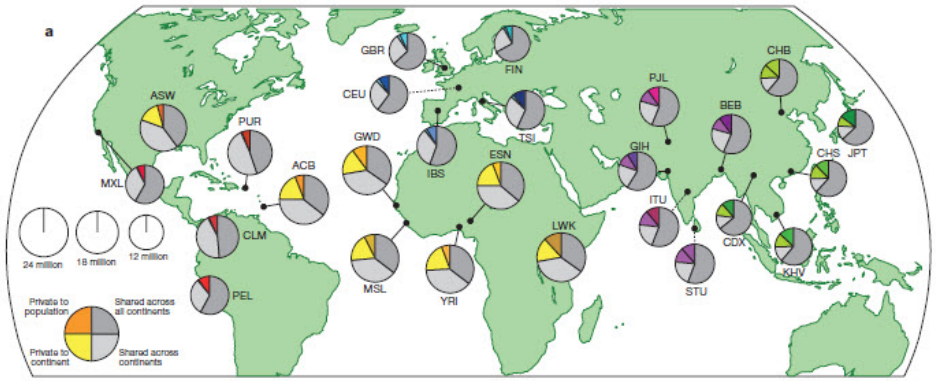
- Next generation sequencing has made whole-genome & whole-population sequencing possible.



www.1000genomes.org

Introduction

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

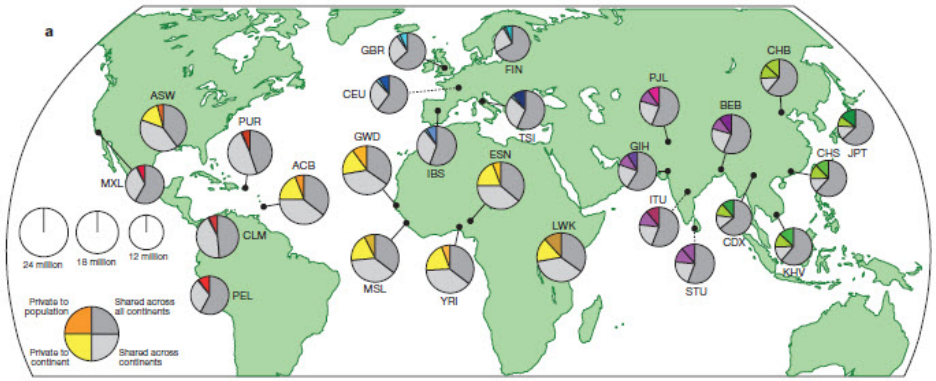


www.1000genomes.org

- For organisms with "short-generation-time", (e.g., yeast, *E. coli*, *D. melanogaster* etc.) it is also possible to collect time-series data of population.

Introduction

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



www.1000genomes.org

- For organisms with "short-generation-time", (e.g., yeast, *E. coli*, *D. melanogaster* etc.) it is also possible to collect time-series data of population.
- Given rise of these modern datasets (population longitudinal data), new techniques required to answer classical population genetics questions on real data.

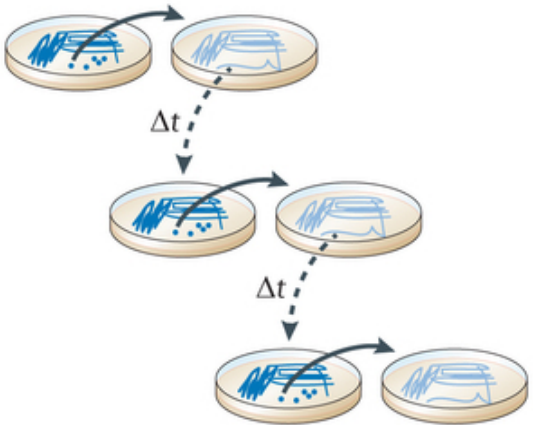
Goals

Design a method which

- Detect regions under selection.
- Localizing adaptive allele within the candidate region.
- Estimating selection parameters.

Experimental Evolution (EE)

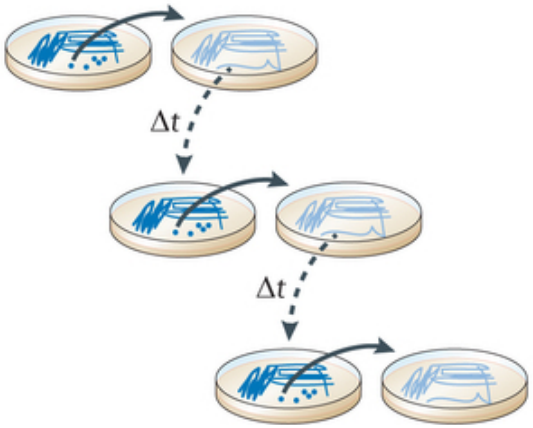
- EE is a long tradition in biology, which studies the phenotype in time by reducing environmental effects .



Nature Reviews Genetics 14, 827-839 (2013)

Experimental Evolution (EE)

- EE is a long tradition in biology, which studies the phenotype in time by reducing environmental effects .

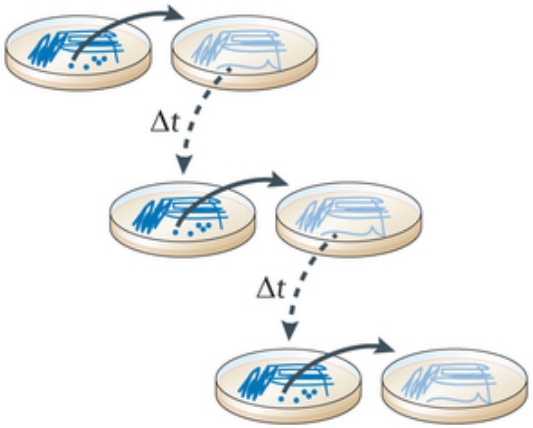


Nature Reviews Genetics 14, 827-839 (2013)

- In a controlled environment, EE evolves a homogeneous population.

Experimental Evolution (EE)

- EE is a long tradition in biology, which studies the phenotype in time by reducing environmental effects .



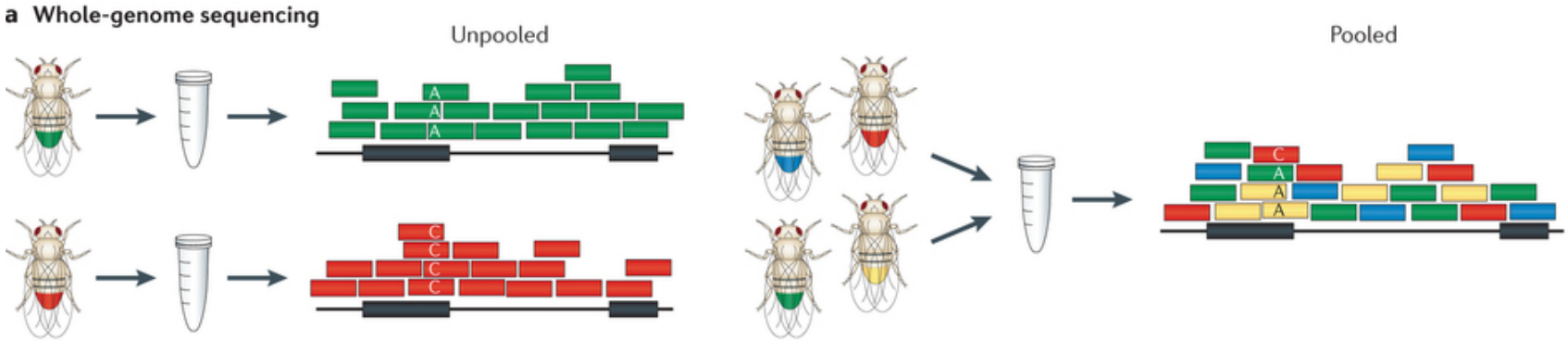
Nature Reviews Genetics 14, 827-839 (2013)

- In a controlled environment, EE evolves a homogeneous population.
- Let phenotype of interest be the response to a selection pressure, e.g., response to
  - antibiotic
  - low oxygen conditions
  - hot and cold temperatures
  - etc.

An experiment design for *D. melanogaster*

Whole-Genome Whole-Population Sequencing

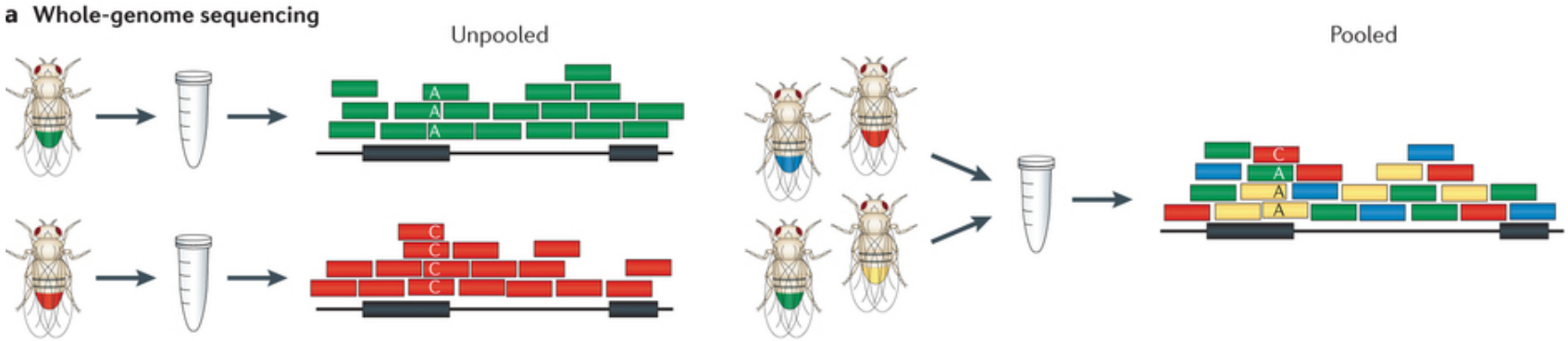
- Pooled-Sequencing



Nature Reviews Genetics 15, 749-763 (2014)

Whole-Genome Whole-Population Sequencing

- Pooled-Sequencing

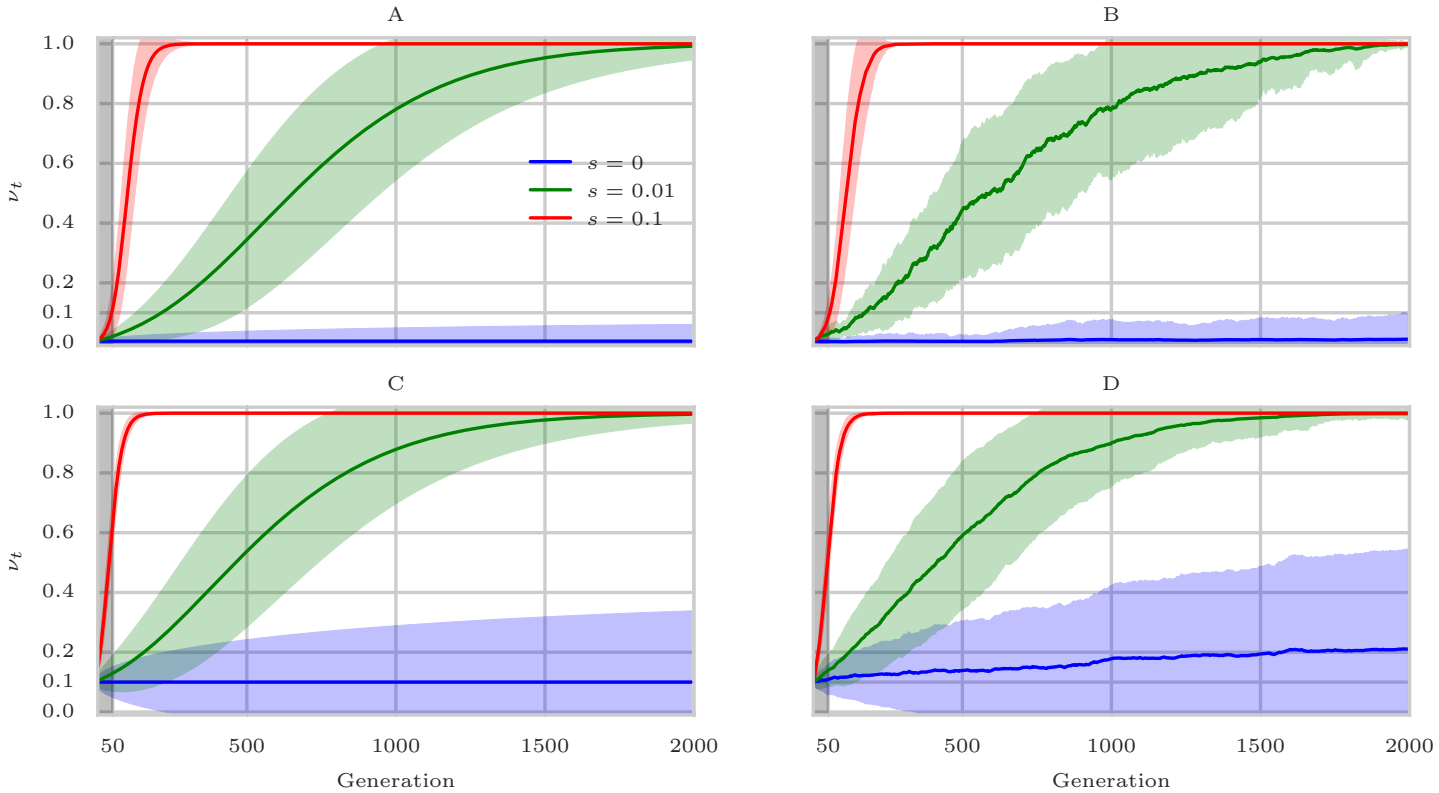


Nature Reviews Genetics 15, 749-763 (2014)

- Implication: only population allele frequency can be computed.

Dynamic of population allele frequency

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

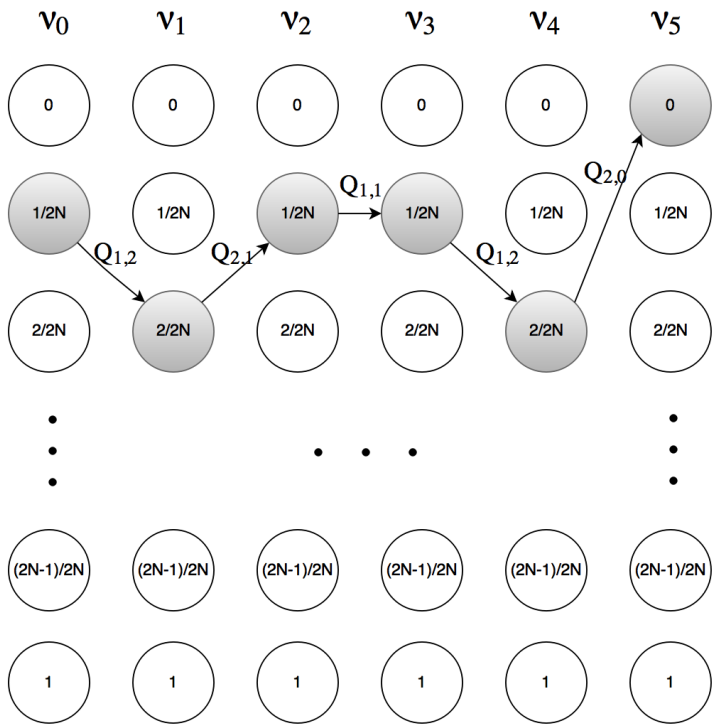


Simplified Model (I)

- Suppose we have sequenced a whole (diploid, size=N) population every generation (eg, for 6 generations) and exact allele frequency are given.

Simplified Model (I)

- Suppose we have sequenced a whole (diploid, size=N) population every generation (eg, for 6 generations) and exact allele frequency are given.
- 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.
- Neutral:

$$Q_{i,j} = \Pr(j; n = 2N, x = \nu_t = i/2N) = \binom{2N}{j} \nu_t^j (1-\nu_t)^{2N-j}$$

- Selection, for $w_{11} = 1+s,\ w_{01} = 1+hs, w_{00} = 1$

$$\hat{\nu}_{t+} = \mathbb{E}[\nu_{t+}|s,h,\nu_t] = \frac{w_{11}\nu_t^2 + w_{01}\nu_t(1-\nu_t)}{w_{11}\nu_t^2 + 2w_{01}\nu_t(1-\nu_t) + w_{00}(1-\nu_t)^2}$$

$$Q_{i,j}(s,h) = \Pr(j; n = 2N, x = \hat{\nu}_{t+})$$

Simplified Model (III)

- Likelihood of parameter can be easily computed

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

Simplified Model (III)

- 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 $\hat{s}$, $\hat{h}$.

Simplified Model (III)

- 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:

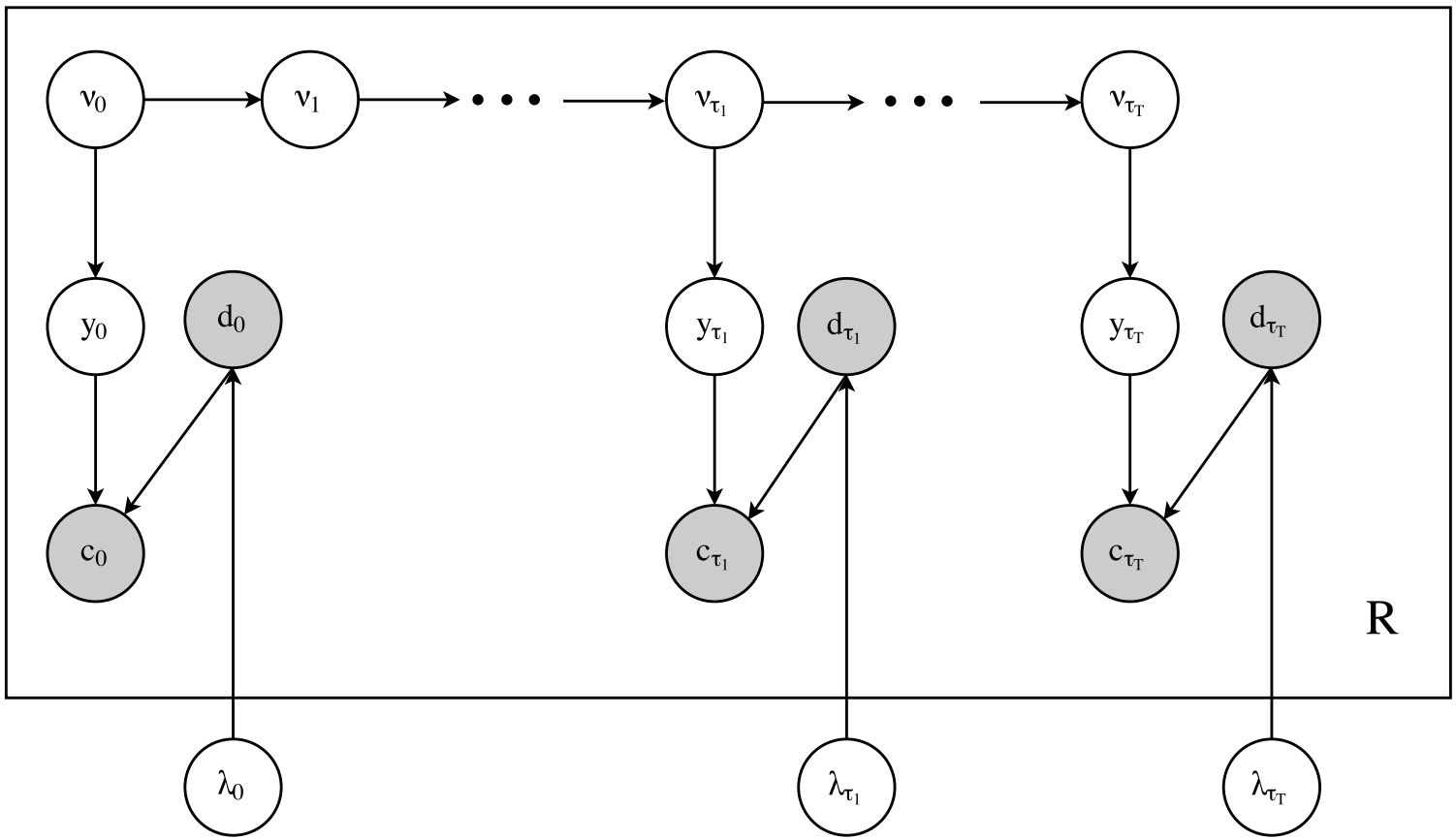
$$M = \frac{\text{likelihood of data as if being under selection with } \hat{s}, \hat{h}}{l\text{ikelihood of data as if being neutral}}$$
$$= \frac{\mathcal{L}(\hat{s}, \hat{h}|\{\nu_0, \ldots, \nu_5\})}{\mathcal{L}(0, 0|\{\nu_0, \ldots, \nu_5\})}$$

Model (complete)

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

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, and finite sample is taken for sequencing.



Generative Process

Generative Process 1: The Generative Process for Dynamic Pool-seq Data.

Input: $N, n, R, \{\lambda_{\tau_0}, \ldots, \lambda_{\tau_T}\}, \mathcal{T} = \{\tau_0, \ldots \tau_T\}$

Output: Time-series pool-seq data for R replicates of a single locus $\{\mathbf{c}^{(r)}\}$ and $\{\mathbf{d}^{(r)}\}$.

for $r \leftarrow 1$ **to** R **do**
  for $t \leftarrow \tau_0$ **to** τ_T **do**
    $2N\nu_t \sim \text{Binomial}(2N, \nu_{t-1})$;
    if $t \in \mathcal{T}$ **then**
      $d_t^{(r)} \sim \text{Poiss}(\lambda_{\tau_i})$;
      $2ny_t \sim \text{Binomial}(2n, \nu_t)$;
      $c_t^{(r)} \sim \text{Binomial}(d_t^{(r)}, y_t)$;
    end
  end
end

Composite Likelihood for a Region (I)

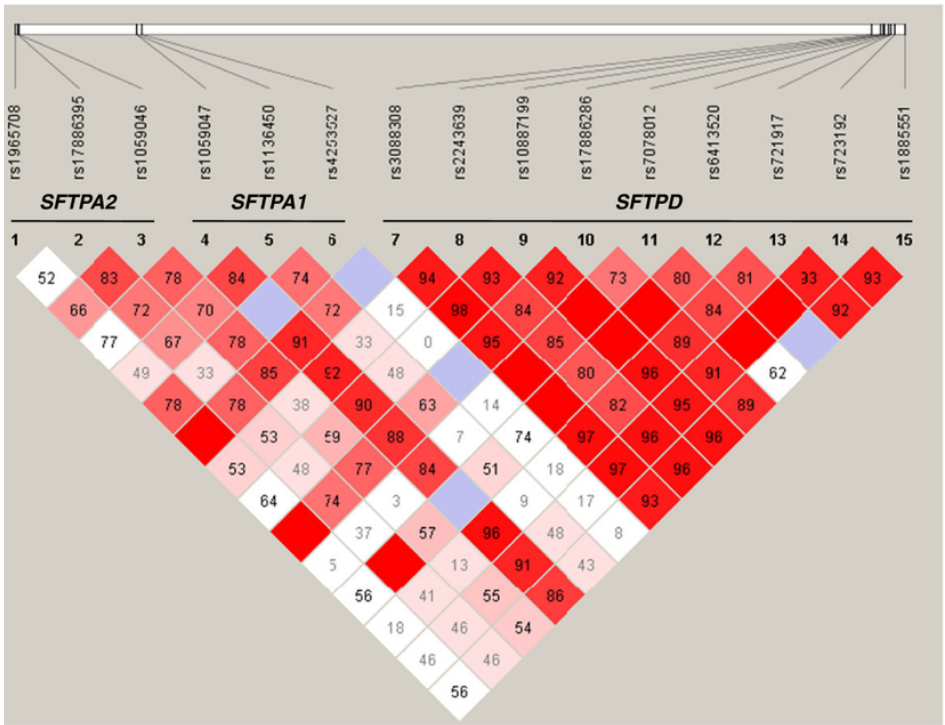
- So far we developed log-odds ratio statistics M (frequency data) and H (read count data) for each variant.

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).

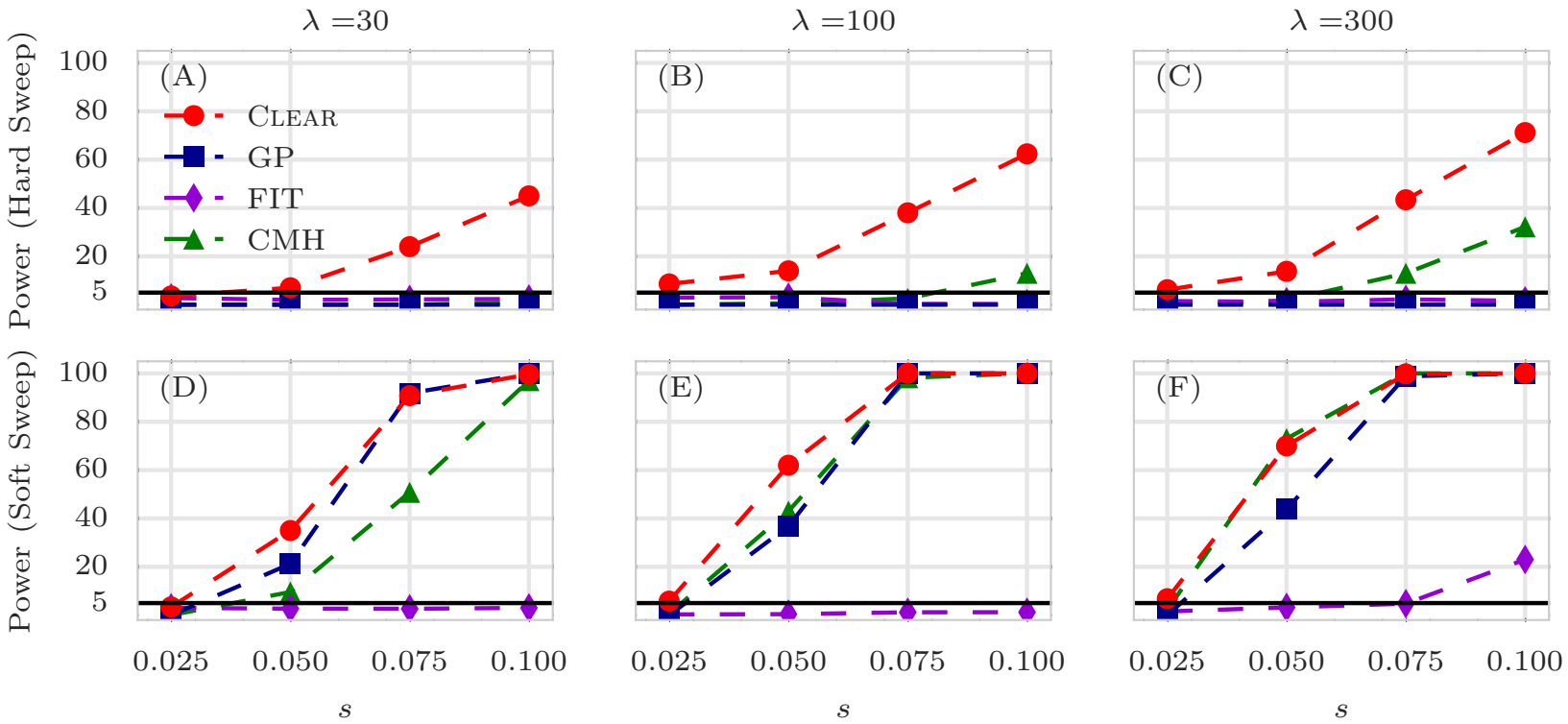
Composite Likelihood for a Region (II)

- 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.

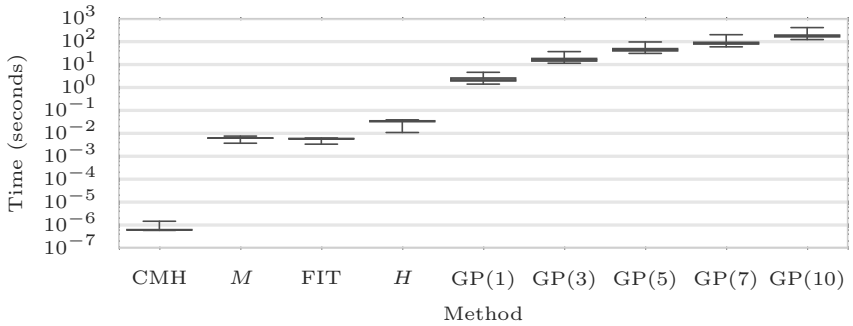
(ii) It can detect well even when coverage is low, i.e., favored allele frequency (1/200 in hard sweep) is below accuracy of sequencing (1/30).

Detecting regions under selection: Observations

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

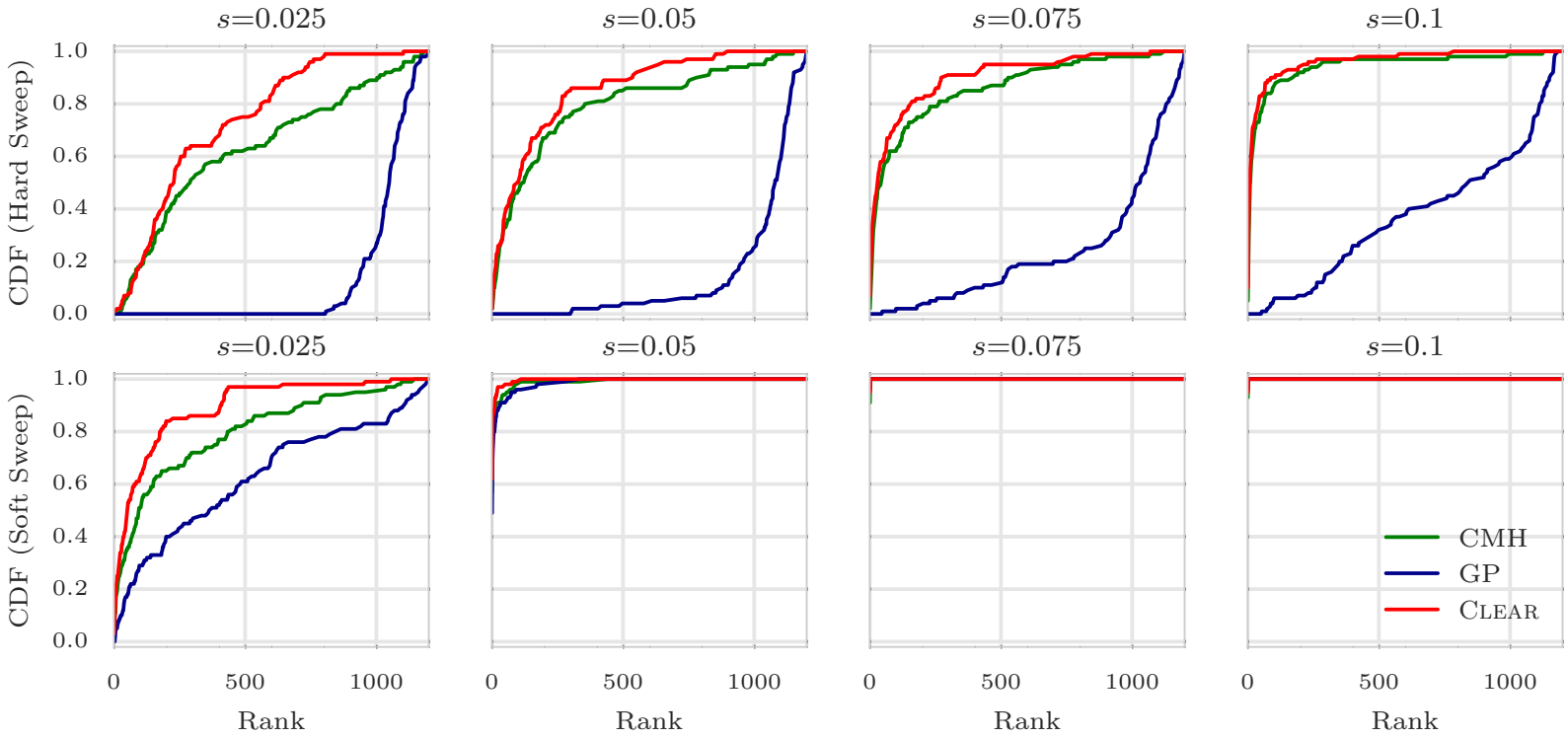
(ii) It can detect well even when coverage is low, i.e., favored allele frequency (1/200 in hard sweep) is below accuracy of sequencing (1/30).

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



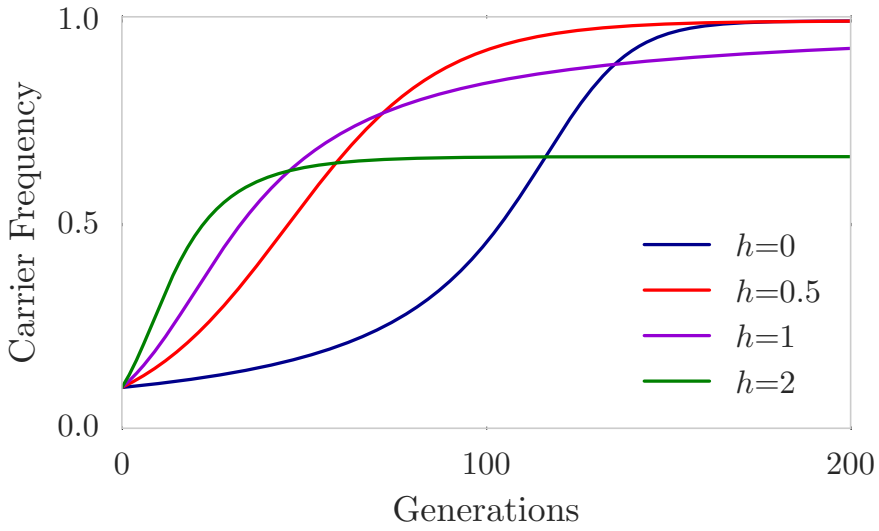
Localizing favored allele

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



Estimating parameters (I)

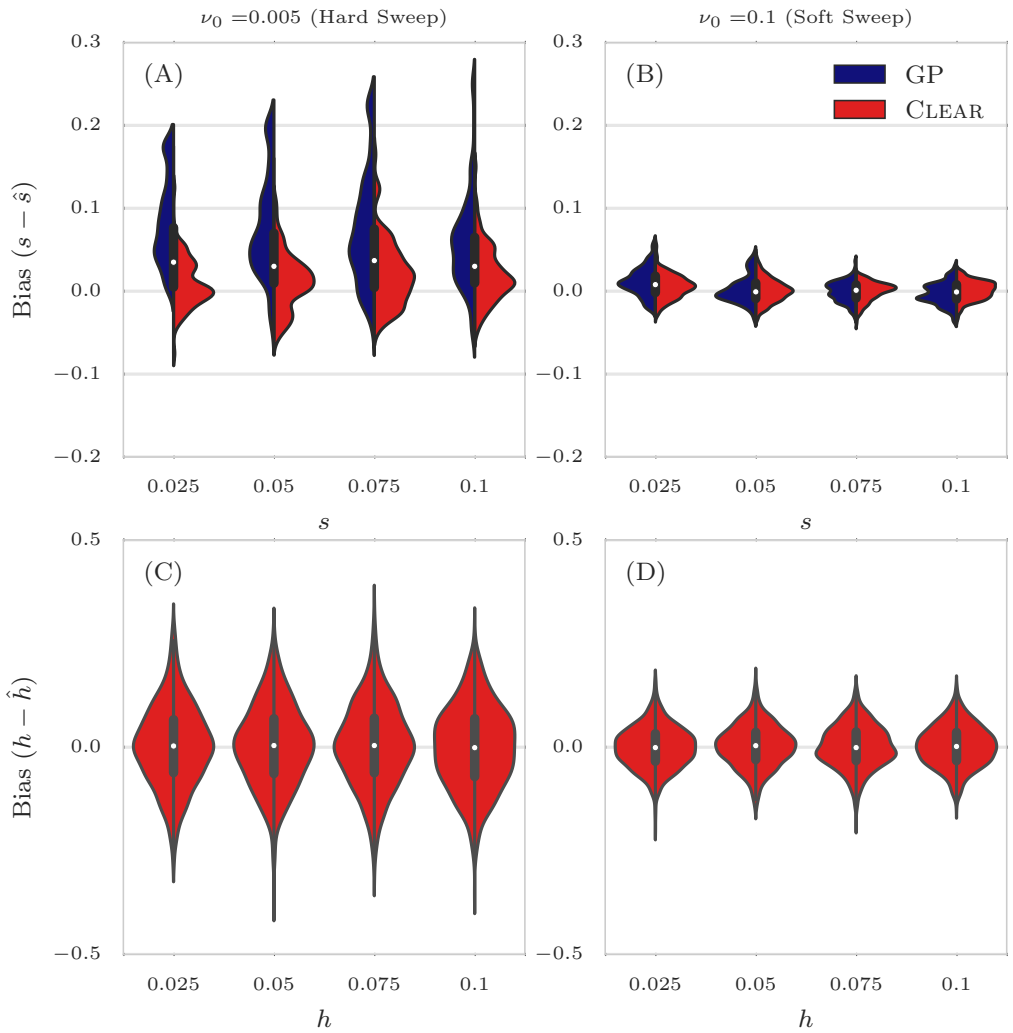
Our model estimates strength of selection s and overdominance h parameter for each variant.



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

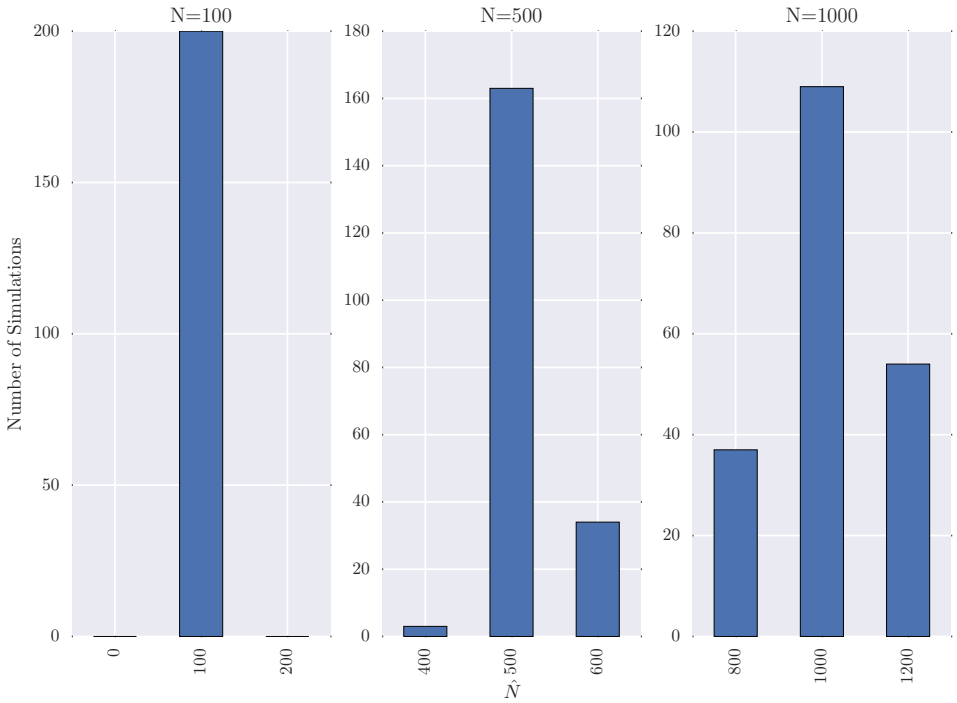
Estimating parameters (II)

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

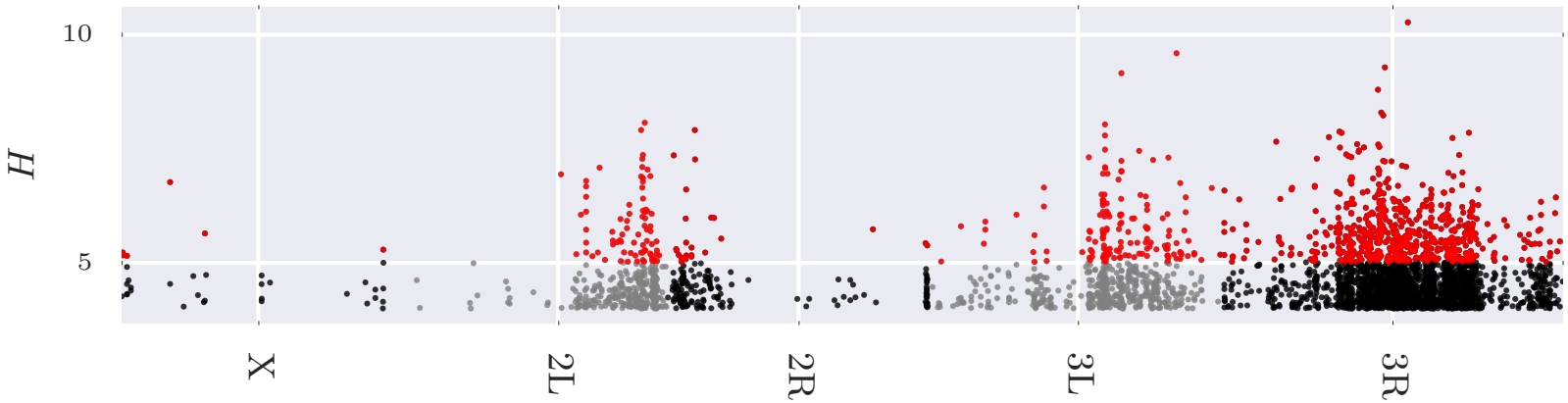
- A population of *D. melanogaster* is evolved for 59 generations, under alternative hot and cold temperatures.

Analysis of real data

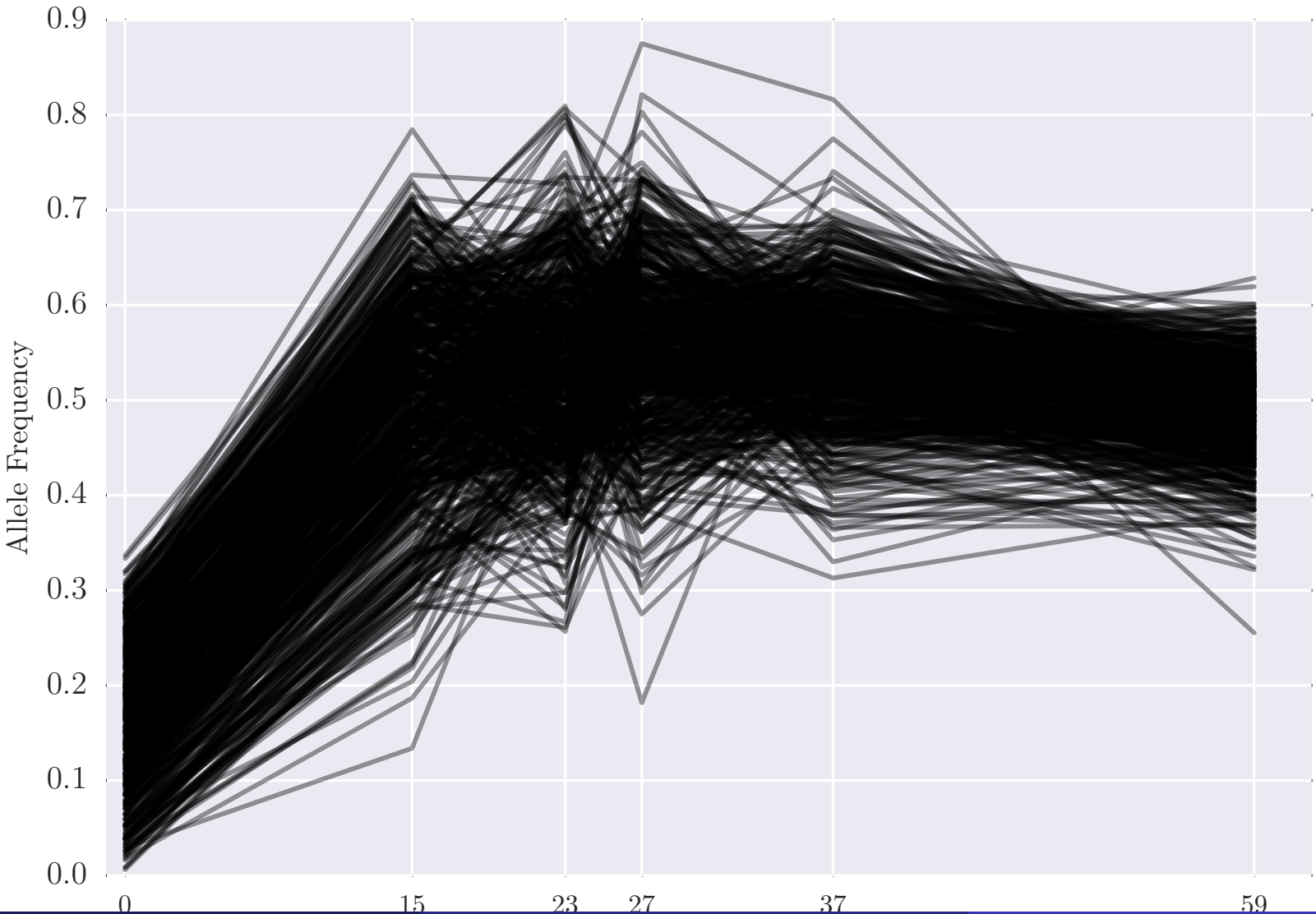
- A population of *D. melanogaster* is evolved for 59 generations, under alternative hot and cold temperatures.
- Coverage is different at generations and samples are not synchronized.

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



384 variants showing signature of overdominance



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.
- When initial frequency of the favored allele is low, stronger selection helps detecting selection but makes locating favored allele a harder task.
- Next step is to apply to new dataset with a well defined phenotype, e.g. response to hypoxia.

Thanks!