

The hitchhiking effect of beneficial mutations  
and inference of adaptive evolution

Yuseob Kim

School of Life Sciences  
Arizona State University

## Nature of beneficial mutations

- Rate
- Fitness effect (selection coefficient,  $s > 0$ )
- Genomic location
- Molecular nature
- Phenotypic effect

Beneficial mutations are difficult to observe:

RNA Virus (VSV):  $\sim 6 \times 10^{-8}$  /genome/generation

(Miralles et al. 1999)

*E. coli*:  $\sim 4 \times 10^{-9}$  /genome/generation

(Imhof and Schloetterer 2001)

Drosophila:  $10^{-6} \sim 10^{-4}$  /genome/generation

( $\sim 2 \times 10^{-3}$  *substitutions* /population/generation; Smith and Eyre-Walker 2002)

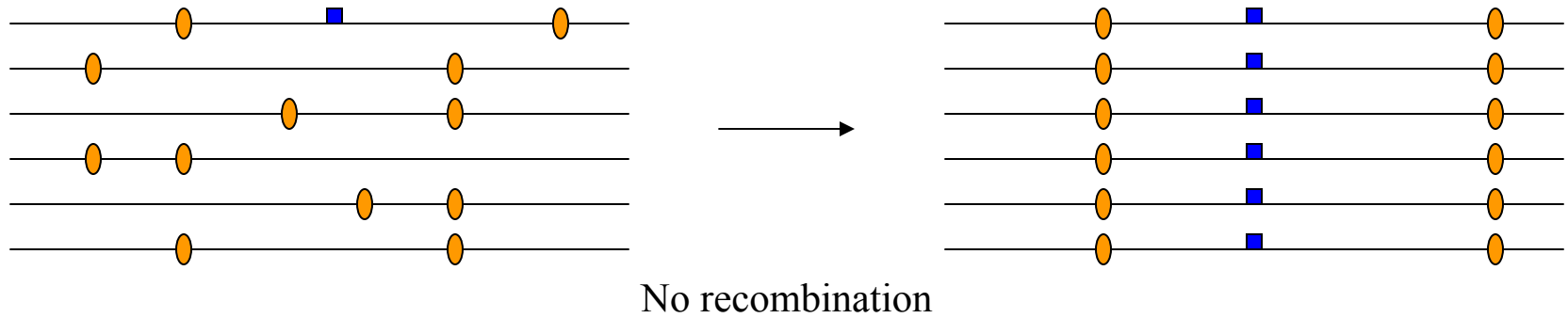
Rare mutations to beneficial alleles:

How can we study beneficial mutations in plants and animals?

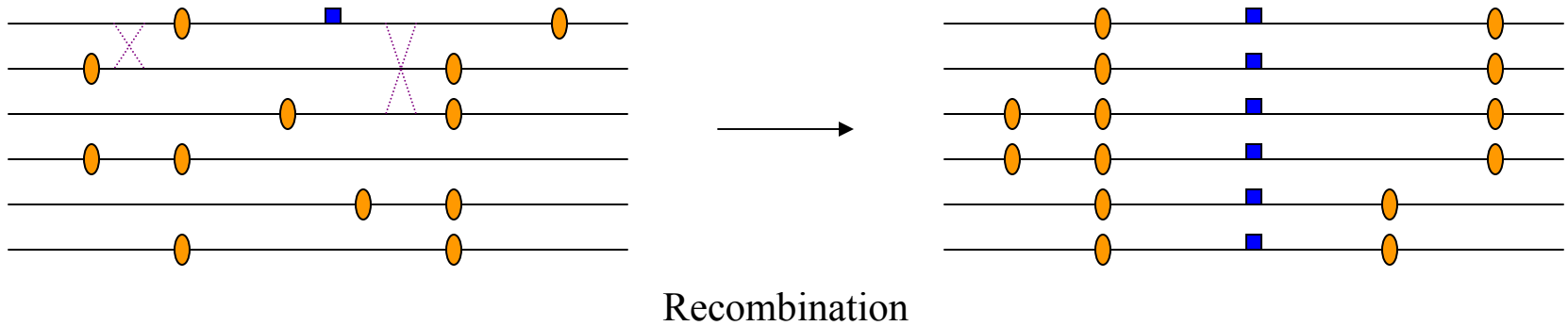
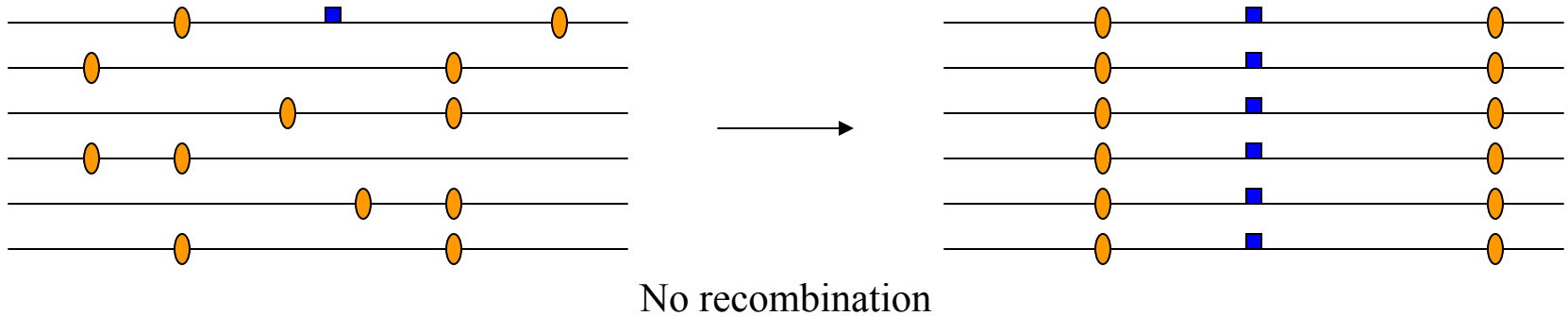
Solution:

Inferring the evolutionary history  
from DNA sequence variation and diversity

# “Hitchhiking” effect of beneficial mutations or “Selective sweeps”

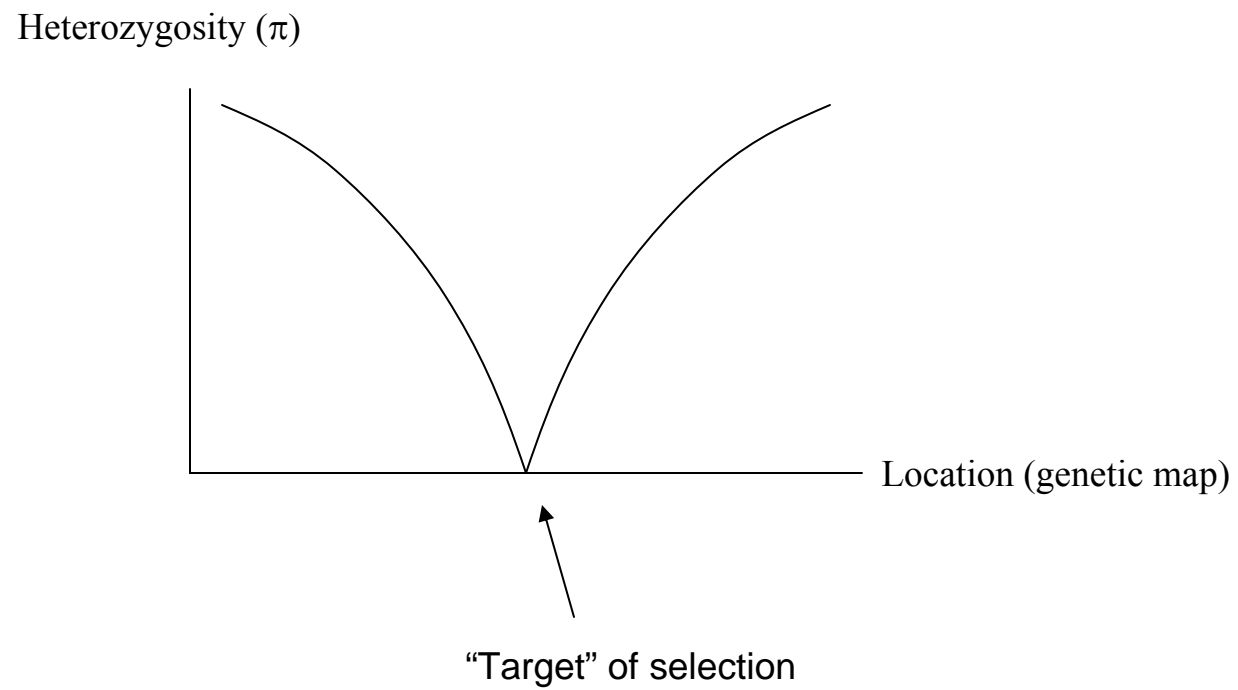


# “Hitchhiking” effect of beneficial mutations or “Selective sweeps”



# Hitchhiking effect of a beneficial mutation

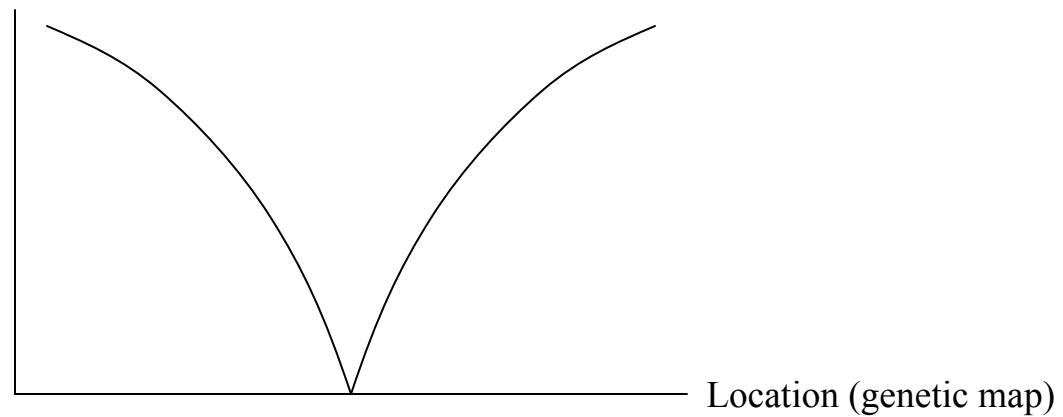
Maynard Smith and Haigh (1974)



# Hitchhiking effect of a beneficial mutation

Maynard Smith and Haigh (1974)

Heterozygosity ( $\pi$ )



⇒ Mapping genes of recent adaptive change

## Hitchhiking effect – example

Non-glutinous rice



10~30% of the total starch  
is amylose

Glutinous rice



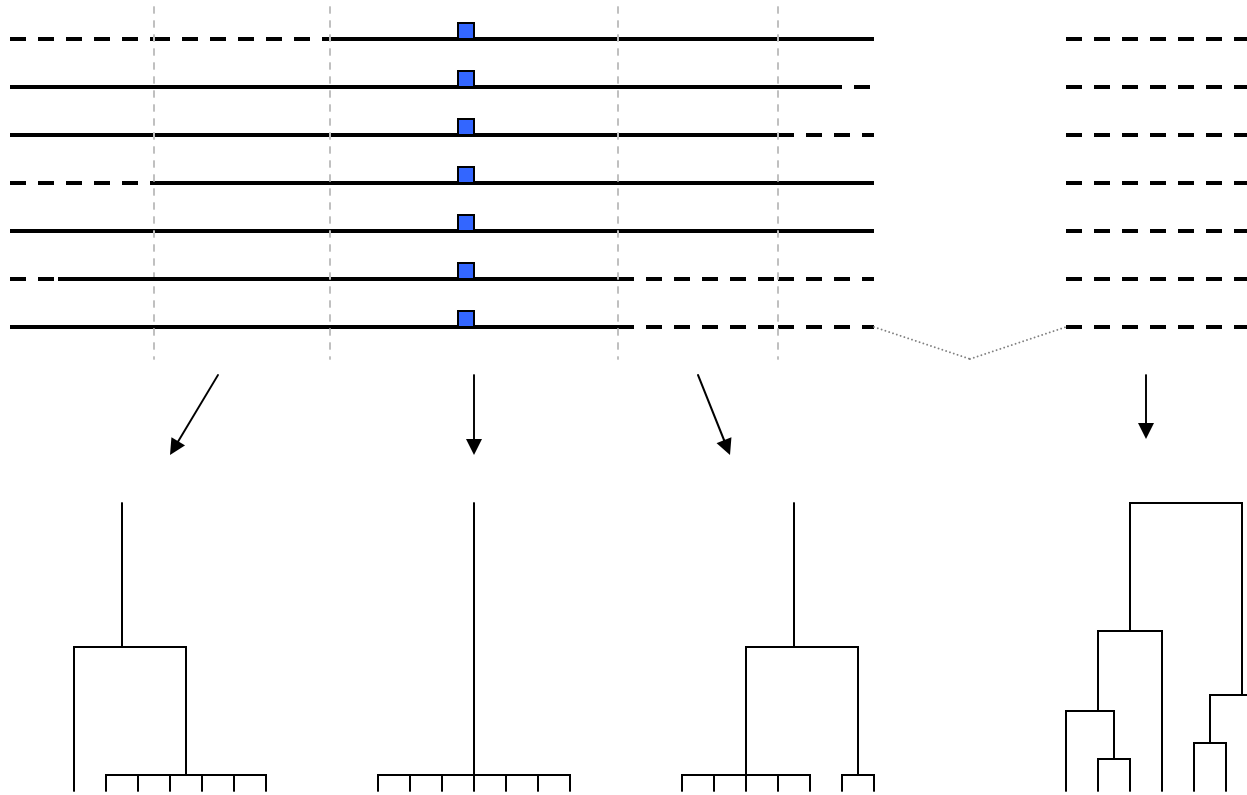
0% of the total starch  
is amylose



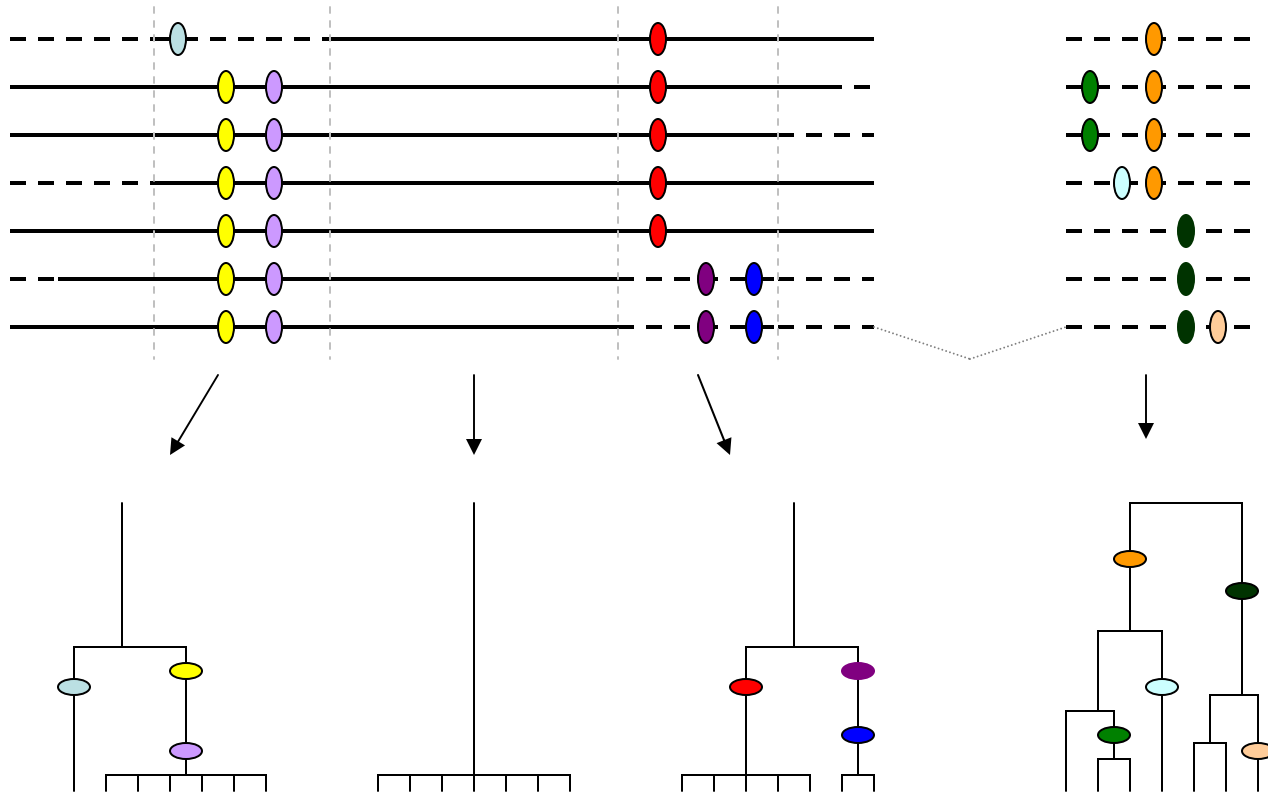
# Outline

1. Signatures of selective sweeps
2. Sampling probabilities
3. Complication with demography

# A selective sweep - genealogical interpretation



# A selective sweep - genealogical interpretation

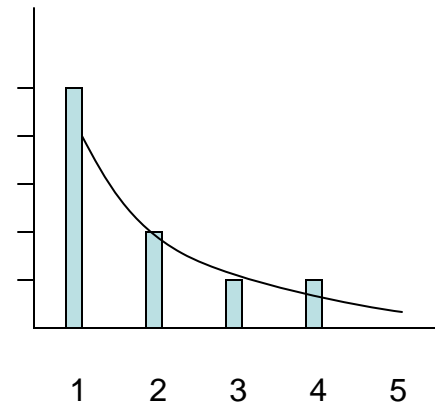


## Expected signatures of selective sweeps (Transient patterns)

- 1) Spatial distribution: a “pocket” of reduced variation
- 2) Frequency spectrum: high & low frequency mutant alleles
- 3) Linkage disequilibrium (LD)

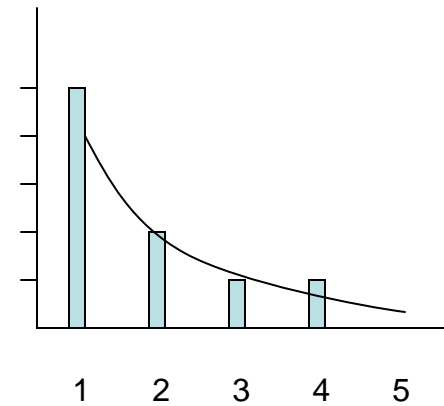
# Frequency spectrum - Neutral

-----G-----T-----  
---A-----T-----C---  
-----C---T---T-----  
-G-----C-----G-----C-  
-----T-----C-----  
-----T-----C-----



## Frequency spectrum - Neutral

-----G-----T-----  
---A-----T-----C---  
-----C---T---T-----  
-G-----C-----G-----C-  
-----T-----C-----  
-----T-----C-----



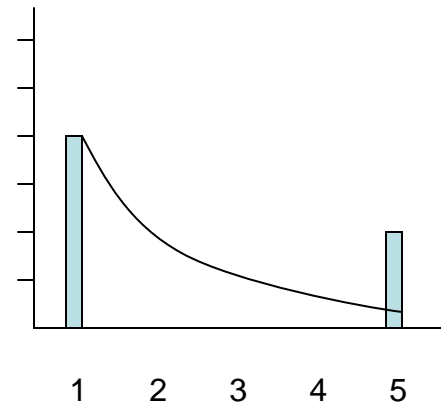
$$E[\pi] = E[\theta_W] = E[\theta_H] = 4N\mu$$

$$\text{Tajima's } D \sim \pi - \theta_W = 0$$

$$\text{Fay \& Wu's } H \sim \pi - \theta_H = 0$$

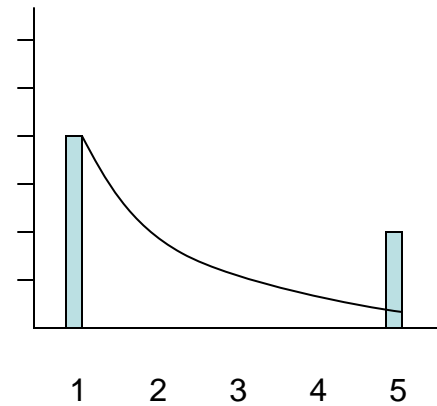
# Frequency spectrum - Selective sweep

-----G-----T-----  
-----G-----G-----  
-----G-----T-----  
--C--A-----T-----  
-----G-----T-----  
-----G-----T-----T-----



## Frequency spectrum - Selective sweep

-----G-----T-----  
-----G-----G-----  
-----G-----T-----  
--C--A-----T-----  
-----G-----T-----  
-----G-----T-----T-----

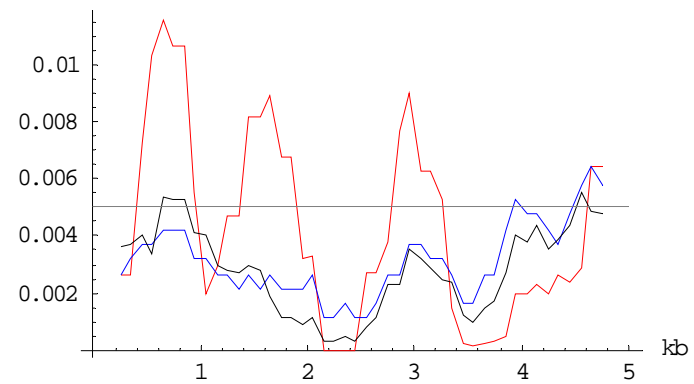
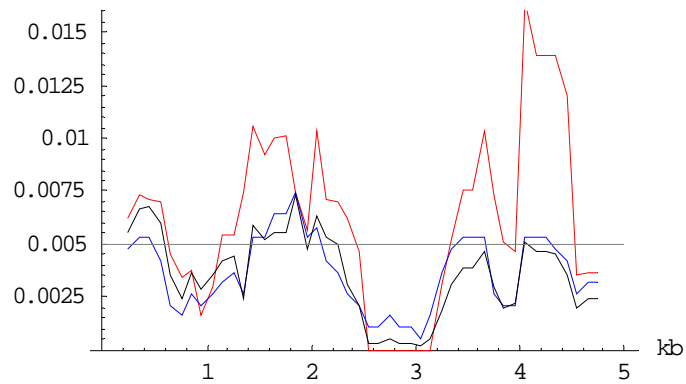
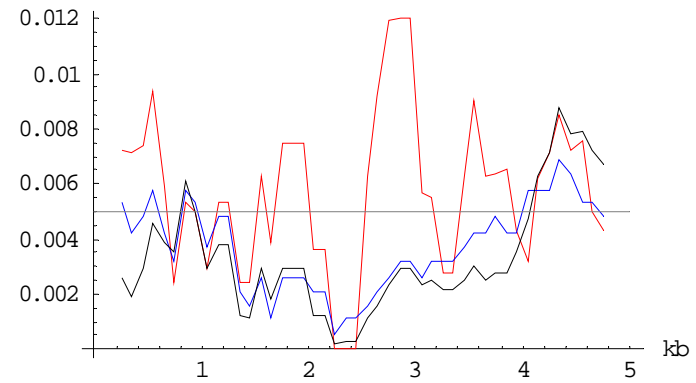
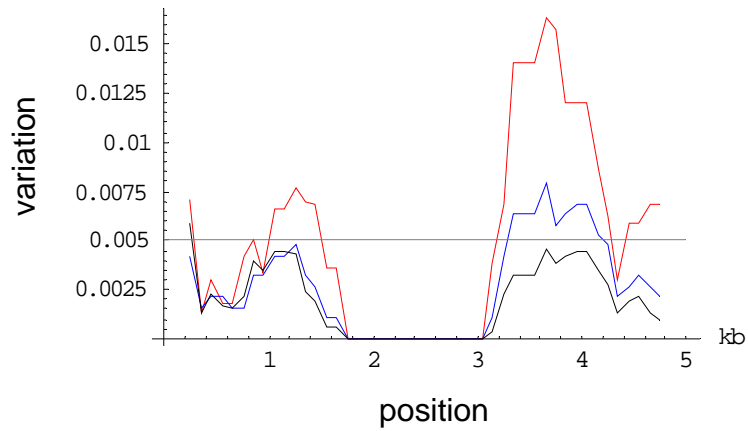


$$E[\pi] < E[\theta_W], \quad \text{Tajima's } D < 0$$

$$E[\pi] < E[\theta_H], \quad \text{Fay \& Wu's } H < 0$$

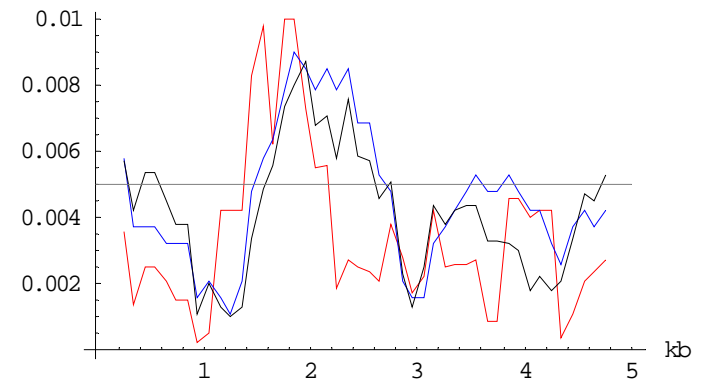
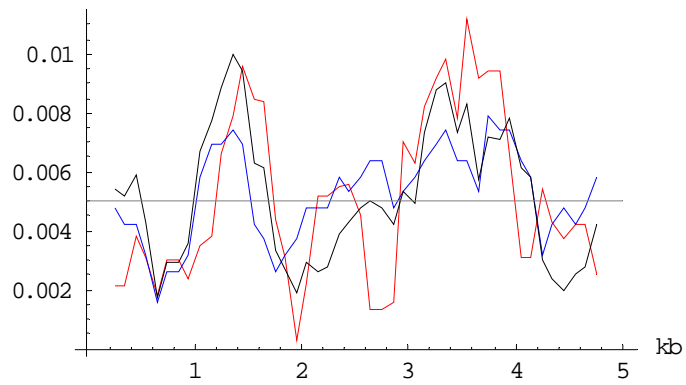
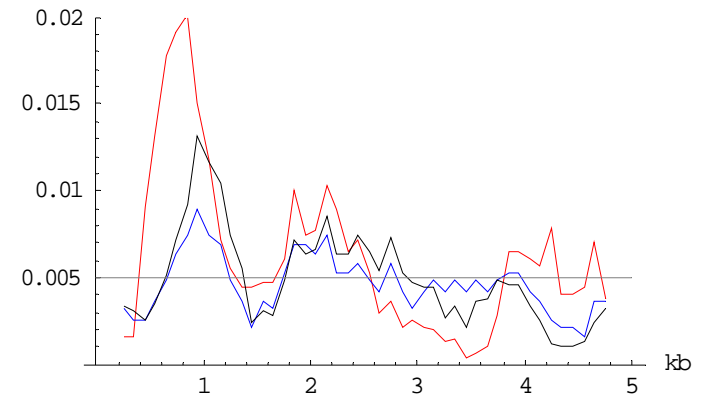
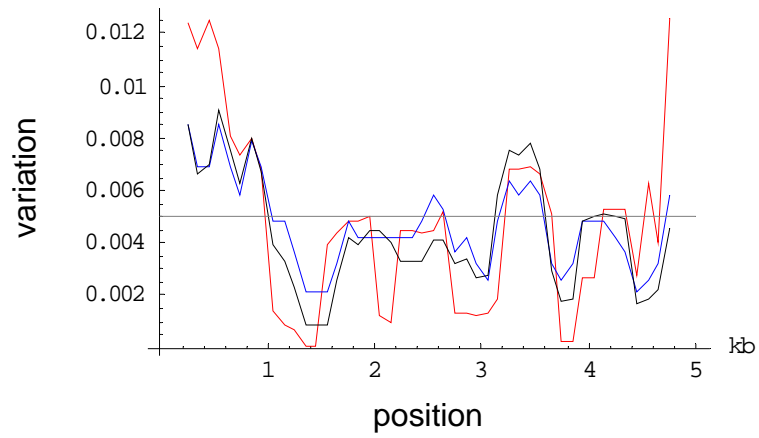
Selective sweeps (  $4Nr = 500$ ,  $\alpha = 2Ns = 1000$ ,  $\tau = 0.001$ ,  $n = 25$  )

—  $\pi$  —  $\theta_W$  —  $\theta_H$

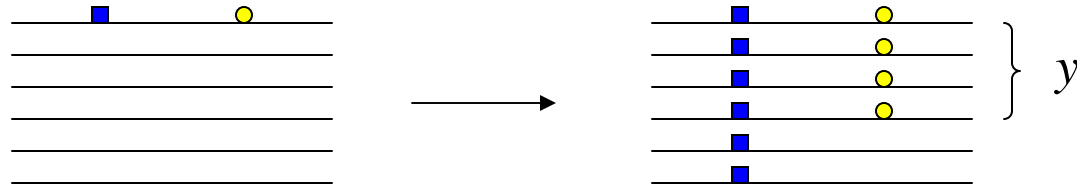


# Neutral variation ( $4Nr = 500$ , $n = 25$ )

—  $\pi$  —  $\theta_W$  —  $\theta_H$

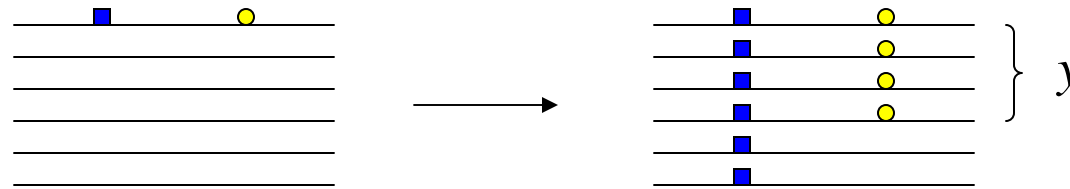


# Frequency spectrum after a sweep

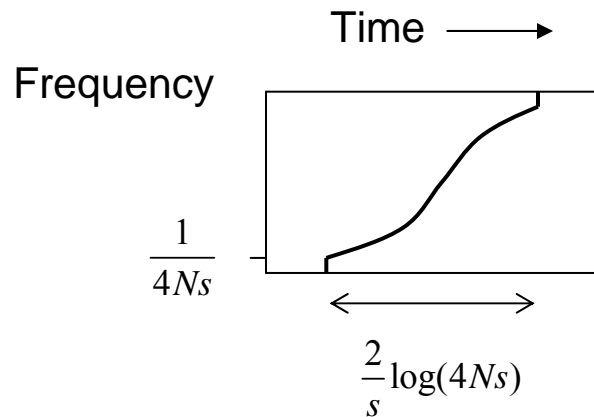


$$y \approx (4Ns)^{-r/s}$$

## Frequency spectrum after a sweep



$$y \approx (4Ns)^{-r/s}$$

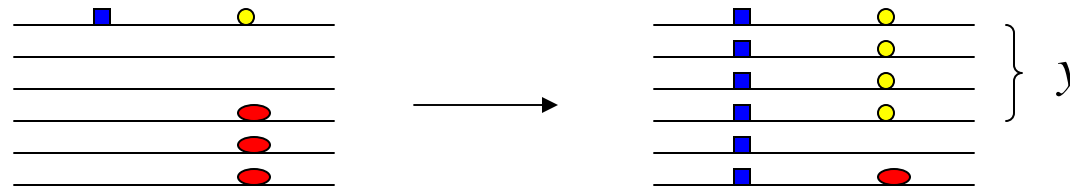


Residual association between alleles

$$\approx \exp\left(-r \times \frac{2}{s} \log(4Ns) \times \frac{1}{2}\right) = (4Ns)^{-r/s}$$

(Barton 2000)

## Frequency spectrum after a sweep

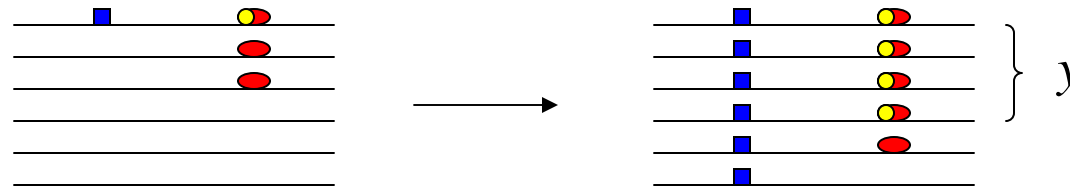


$$y \approx (4Ns)^{-r/s}$$

Transformation of allele frequency by hitchhiking:

$$p \rightarrow \begin{cases} (1-y)p \text{ with probability } 1-p \\ y+(1-y)p \text{ with probability } p \end{cases} \quad (\text{Gillespie 2000})$$

## Frequency spectrum after a sweep

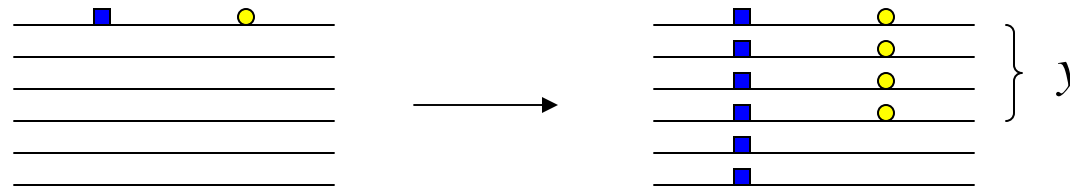


$$y \approx (4Ns)^{-r/s}$$

Transformation of allele frequency by hitchhiking:

$$p \rightarrow \begin{cases} (1-y)p & \text{with probability } 1-p \\ y+(1-y)p & \text{with probability } p \end{cases} \quad (\text{Gillespie 2000})$$

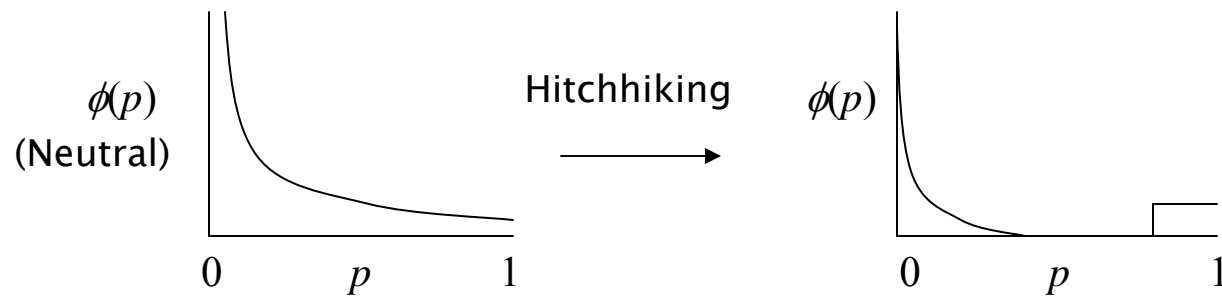
## Frequency spectrum after a sweep



$$y \approx (4Ns)^{-r/s}$$

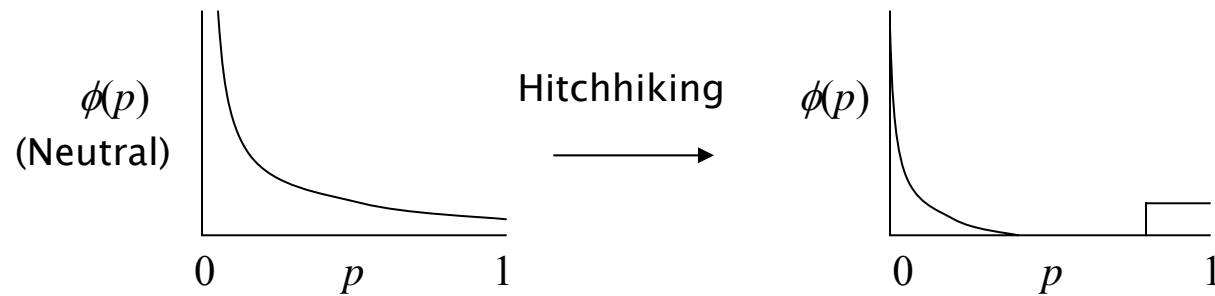
Transformation of allele frequency by hitchhiking:

$$p \rightarrow \begin{cases} (1-y)p & \text{with probability } 1-p \\ y+(1-y)p & \text{with probability } p \end{cases} \quad (\text{Gillespie 2000})$$



$$\phi(p) = \frac{\theta}{p}$$

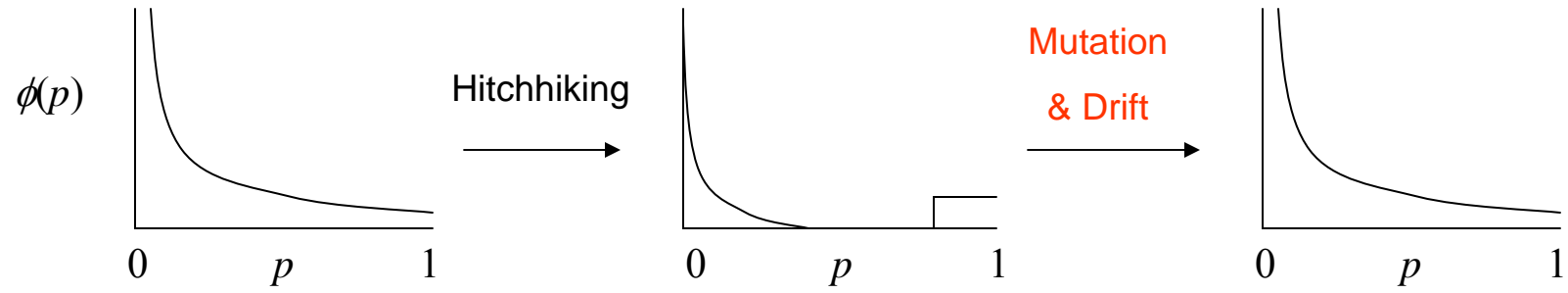
## Frequency spectrum after a sweep



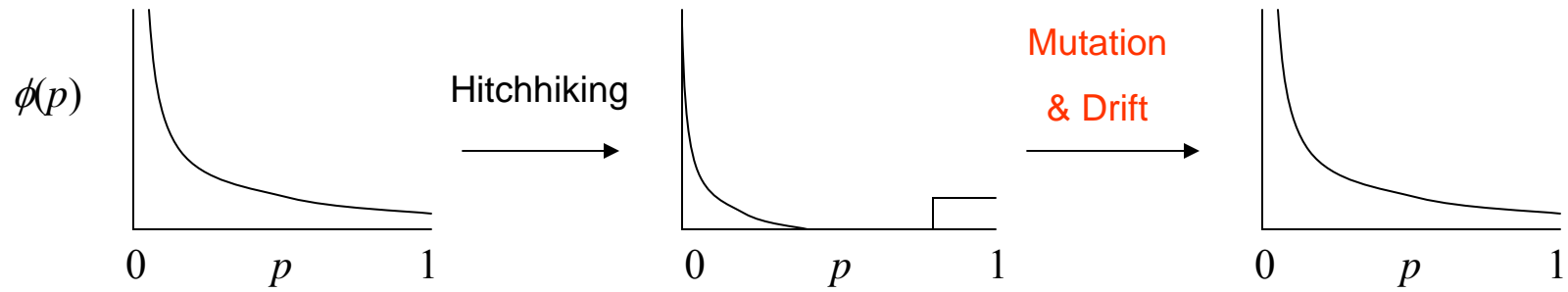
Probability of polymorphism with  $k$  neutral variants ( $0 < k < n$ ):

$$P_k = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \phi(x) dx$$

# Frequency spectrum after a sweep



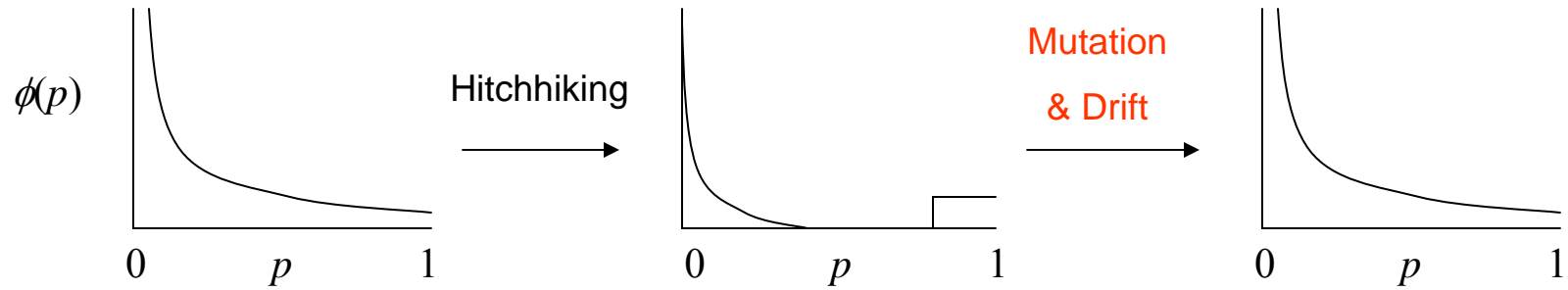
# Frequency spectrum after a sweep



Allele frequency:  $p \xrightarrow{t \text{ generations}} X$

$$E[X^k | p, t]?$$

## Frequency spectrum after a sweep



Allele frequency:  $p \xrightarrow{t \text{ generations}} X$

$\nearrow E[X^k | p, t]?$

$$\frac{dE[X^k]}{dt} = \frac{k(k-1)}{4N} \{E[X^{k-1}] - E[X^k]\} \quad (\text{Kimura 1955})$$

Solution:

$$E[X^k | p, t] = \sum_{i=1}^k C_i^{(k)}(p) e^{-\frac{i(i-1)}{4N}t}$$

where

$$C_i^{(k)}(p) = C_i^{(k-1)}(p) \frac{k(k-1)}{k(k-1) - i(i-1)}, \quad (1 \leq i \leq k)$$

$$C_k^{(k)}(p) = p^k - \sum_{i=1}^{k-1} C_i^{(k)}(p), \quad (k \geq 2)$$

$$C_1^{(1)}(p) = p$$

Sampling probability at  $t$  generations after a selective sweep

$$P_k = P_k(t, r) = \int_0^1 Q_k(p) \phi(p) dp + \theta R_k,$$

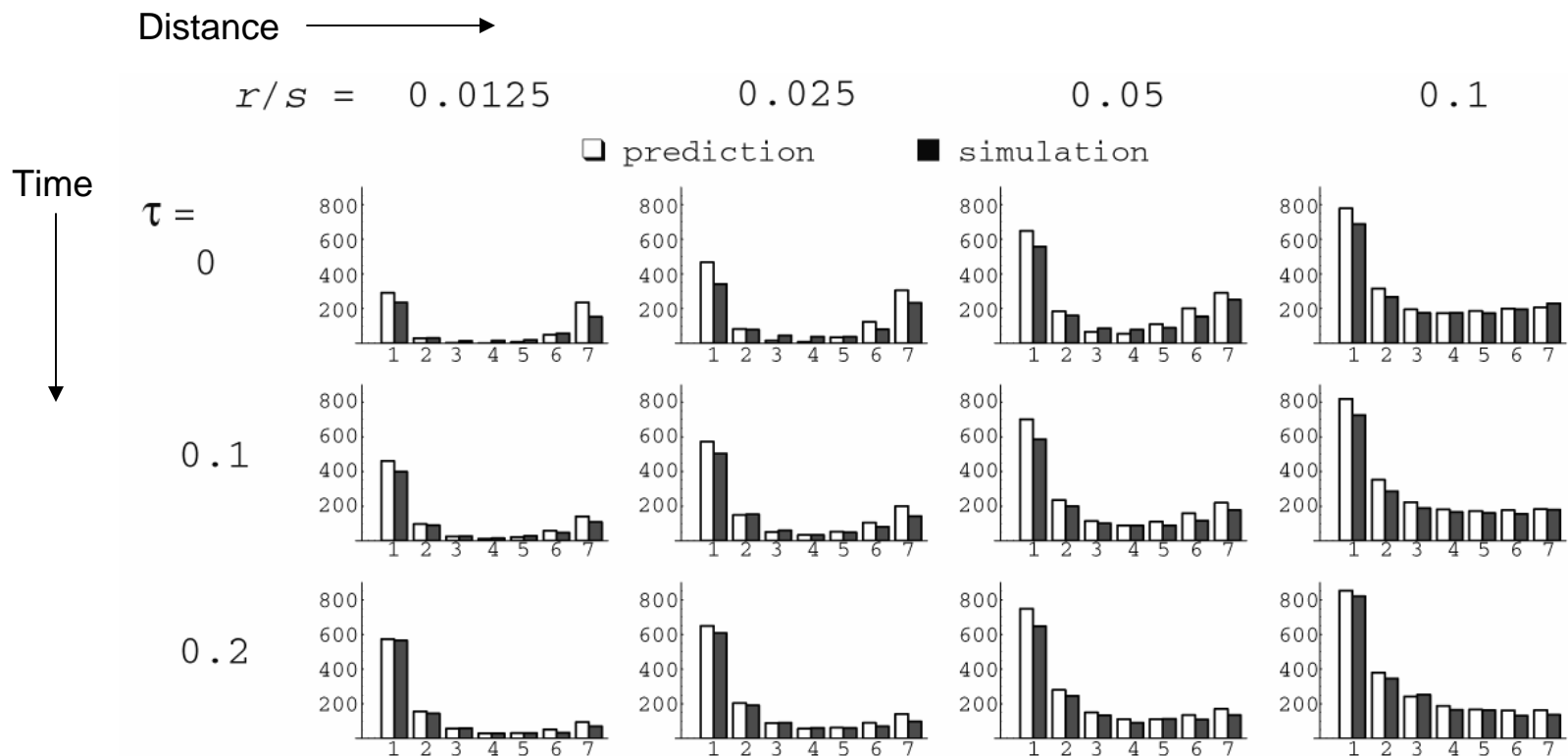
where

$$Q_k(p) = \binom{n}{k} \sum_{i=0}^n \sum_{j=0}^i c_{ij}^{(k, n-k)} \{p(y + (1-y)p)^j + (1-p)(1-y)^j p^j\} e^{-\frac{i(i-1)}{4N}t},$$

$$R_k = \binom{n}{k} \left\{ \frac{t}{4N} c_{11}^{(k, n-k)} + \sum_{i=2}^n \sum_{j=0}^i \frac{c_{ij}^{(k, n-k)}}{i(i-1)} (1 - e^{-\frac{i(i-1)}{4N}t}) \right\}.$$

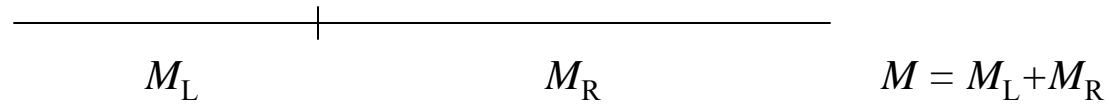
(Kim 2006)

# Sampling probability at $t$ generations after a selective sweep



$(n = 8, N = 10^5, \alpha = 1000)$

## Frequency spectrum under recurrent selective sweeps



$\lambda$ : substitution rate per base (selection)

$\rho$ : recombination rate per base

$\theta$ : scaled mutation rate ( $4N\mu$ )

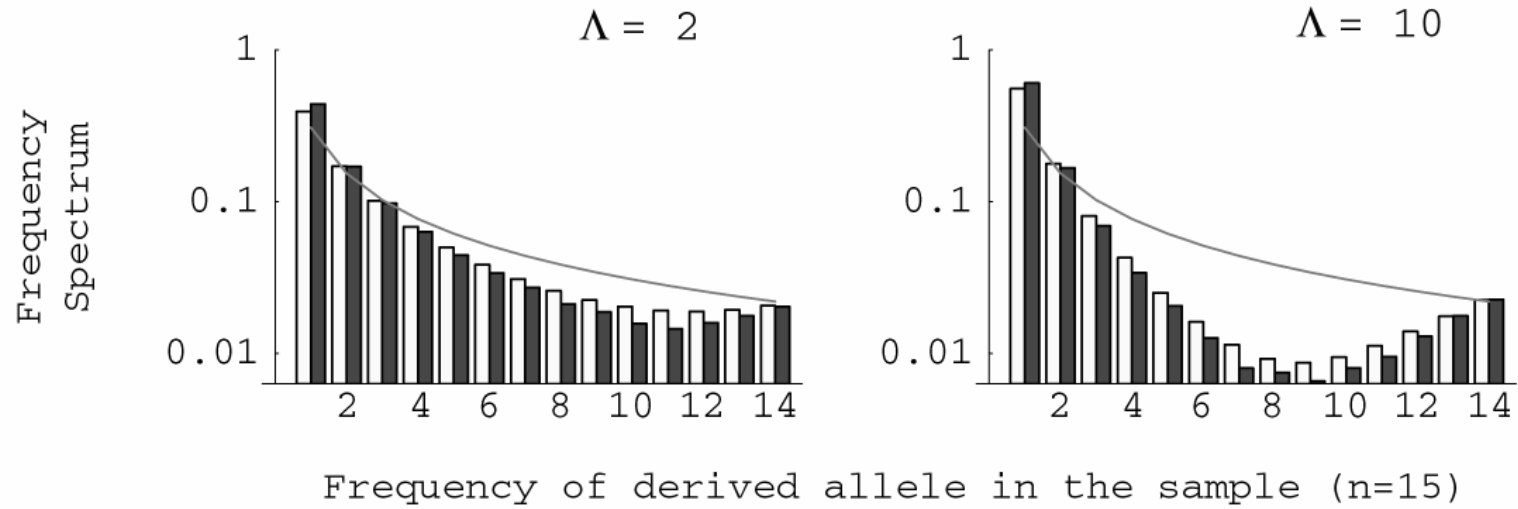
$$\bar{P}_k \approx \frac{1}{M} \int_{-M_L}^{M_R} \int_0^{\infty} P_k(t, \rho | m) \lambda M e^{-\lambda M t} dt dm$$

$$= \int_0^1 \bar{Q}_k(p) \bar{\phi}(p) dz + \theta \bar{R}_k$$

$$\approx \sum_{j=1}^{n-1} \bar{Q}_k \binom{j}{n} \bar{P}_j + \theta \bar{R}_k$$

## Frequency spectrum under recurrent selective sweeps

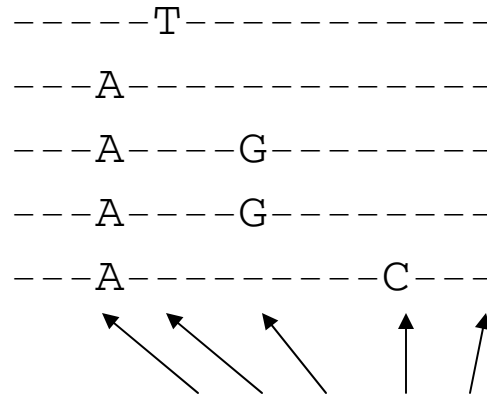
□ Prediction    ■ Simulation



$$\Lambda = 4NM\lambda, \alpha = 2000, 4NM\rho = 400, M_L = M_R = 10^5$$

(Kim 2006)

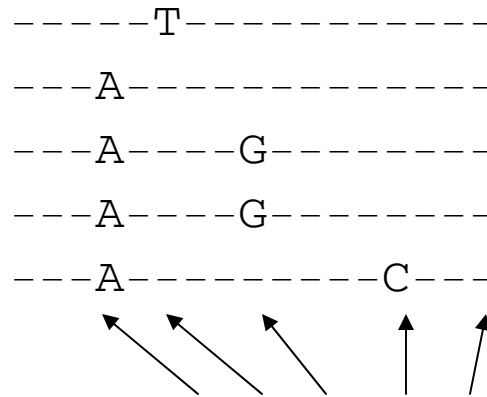
# (Composite) Likelihood of the data



$$\text{Composite Likelihood } (CL_1) = P_4 P_1 P_2 P_1 P_0 \dots$$

(Kim and Stephan 2002)

## (Composite) Likelihood of the data



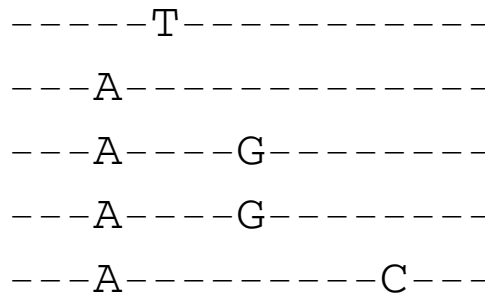
$$\text{Composite Likelihood } (CL_1) = P_4 P_1 P_2 P_1 P_0 \dots$$

(Kim and Stephan 2002)

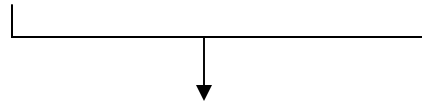
⇒ Information from  
Spatial distribution + Frequency spectrum

(Composite) Likelihood of the data:  
Including Linkage Disequilibrium (LD)

(Kim and Nielsen 2004)



$$\text{Composite Likelihood } (CL_2) = P_{4,1,0} P_{4,2,2} P_{1,2,0} P_{2,1,0} P_{4,1,1} \dots$$



Determined by simulations

## Likelihood ratio test

Model 0: Neutrality  $\longrightarrow$   $CL_N$

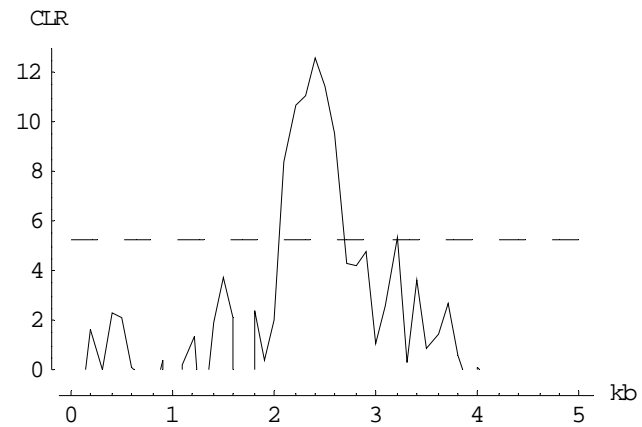
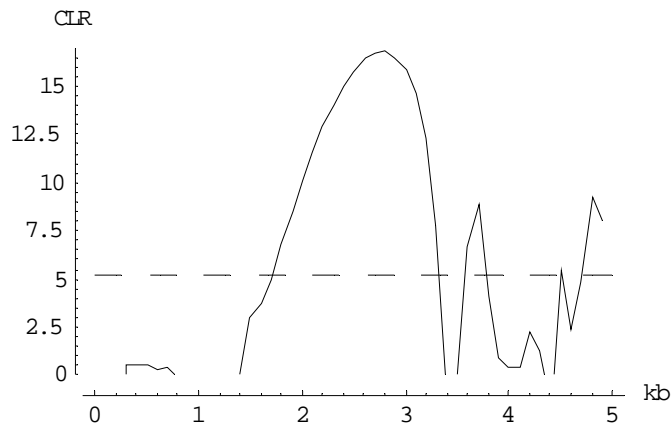
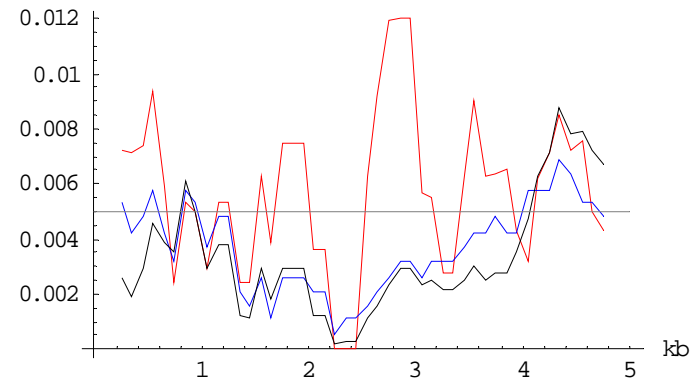
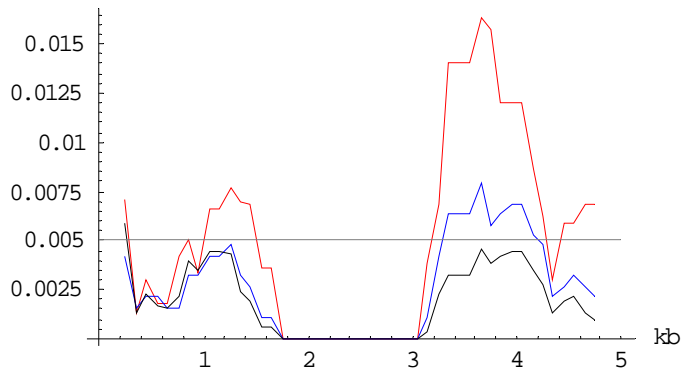
Model 1: Selective sweep ( $\alpha=2N_s, X, \dots$ )  $\longrightarrow$   $CL_S$

$$\Lambda = \ln(CL_S/CL_N)$$

Null distribution of  $\Lambda$   $\longleftarrow$  Neutral Simulations

# Selective sweeps ( $R = 500$ , $\alpha = 1000$ , $\tau = 0.001$ , $n = 25$ )

—  $\pi$     —  $\theta_W$     —  $\theta_H$



# Application to real data:

## *Acp26a* region, *D. melanogater* North Carolina population

Aguadé et al. 1992

TABLE 2  
Polymorphism at the *Mat26A* region of *D. melanogaster*

Allele	Location																																						
	-22	1	52	109	157	151	176	217	270	307	316	383	406	417	693	701	743	813	1120	1155	1260	1308	1332	1378	1383	1418	1421	1491	1499	1503	1533	1557	1572	1578	1575	1582			
con	T	C	A	C	G	C	T	C	C	G	G	A	A	C	T	GC	G	G	T	T	T	C	A	A	G	G	T	T	C	G	A	A	T	C	-				
1	*	*	*	*	*	*	*	*	G	*	*	G	*	T	*	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	A	*	*	*	*	G	*	*	C	TA	A	*	C	*	C	T	*	G	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3	A	T	*	T	*	*	*	A	*	A	*	G	T	*	*	**	*	A	C	C	*	*	C	G	A	C	G	*	A	*	*	G	C	T	*				
4	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	**	*	*	*	*	*	*	C	T	C	G	*	*	*	A	*	A	C	*	*	-	+		
5	A	T	C	*	A	A	*	*	G	*	*	*	C	*	*	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
6	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	**	*	*	*	*	*	C	T	C	G	*	*	*	A	*	A	C	*	*	-	+			
7	A	T	C	*	A	A	*	*	G	*	*	*	C	*	*	**	*	*	*	C	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
8	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	**	*	*	*	*	*	C	T	C	G	*	*	*	A	*	A	C	*	*	-	+			
9	A	T	C	T	*	*	A	A	*	*	*	G	*	*	*	**	*	*	C	C	*	*	C	*	*	*	*	A	*	A	C	*	*	-	+				
10	A	T	C	*	A	A	*	*	G	*	*	*	C	*	*	**	A	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
con	s	s	s	s	ns	ns	ns	ns	s	ns	ns	ns	ns	s	s	ns	ns	s	s	s	s	s	ns	ns	s	s	s	s	s	s	s	s	s	s	s	s	s		
					Asn	Gln	Leu	Pro		Asp	Asp	Asn	Ile			Ser	Arg						Glu	Ile															
					Ser	Lys	Glu	Thr		Asn	Asn	Ser	Leu			Ile	Lys						His	Val															

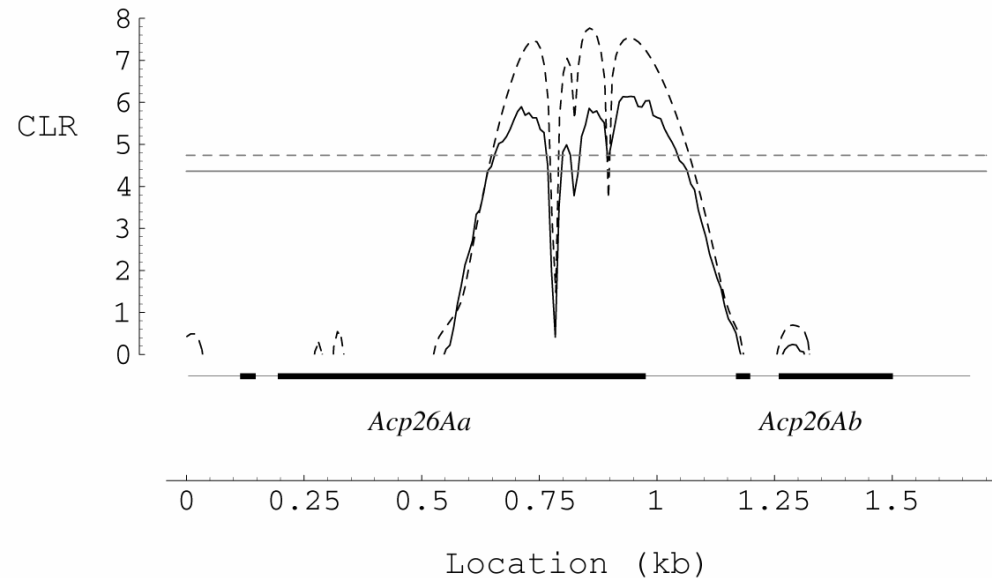
con, consensus; e1a, exon 1 gene *Mat26Aa*; ia, intron gene *Mat26Aa*; e2a, exon 2 gene *Mat26Aa*; ib, intron gene *Mat26Ab*; e2b, exon 2 gene *Mat26Ab*; \*, same than consensus; +, presence; -, absence; s, silent; ns, nonsilent; del, deletion.

## Composite likelihood ratio test:

*Acp26a* region of *D. melanogaster* (Aguade *et al.* 1992)

----- CLR<sub>1</sub> (spatial + spectrum) = 7.8 (p<0.001;  $\alpha = 100.9$ ,  $X = 857$ )

———— CLR<sub>2</sub> (spatial + spectrum + LD) = 6.2 (p<0.007;  $\alpha = 110.2$ ,  $X = 943$ )



(Kim and Nielsen 2004)

## Estimating “genomic” parameters of directional selection

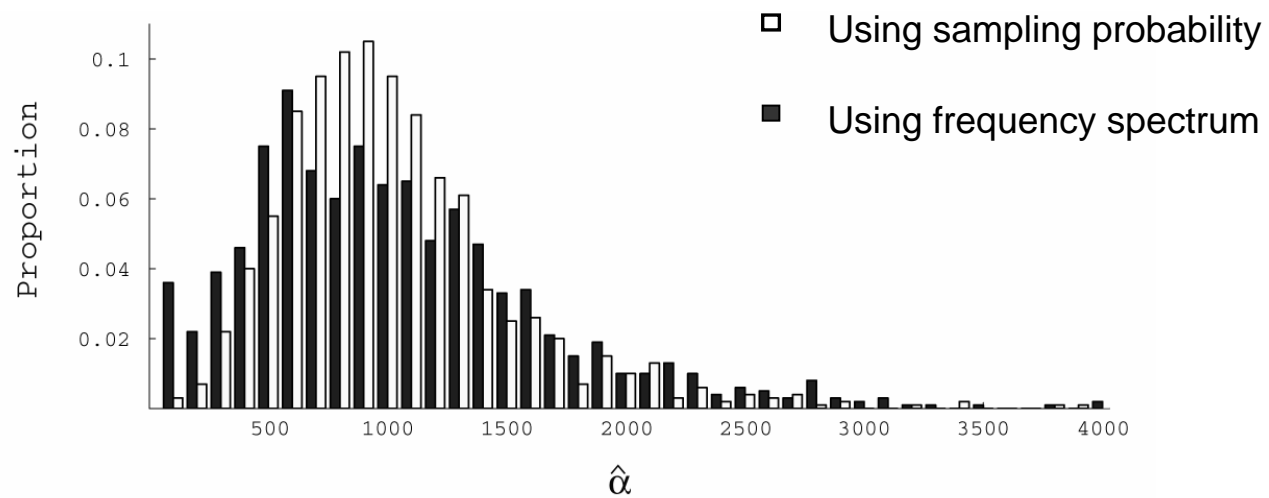
- Data from  $L$  independent loci
- $j$ th locus with frequency spectrum  $S_{j1}, S_{j2}, \dots, S_{jn-1}$ .

New composite likelihood, 
$$CL(\alpha, \lambda) = \prod_{j=1}^L \prod_{i=1}^{n_j-1} P_i^{S_{ji}}$$

## Estimating “genomic” parameters of directional selection

Simulated data set:  $L = 30$ ,  $n = 15$ ,  $\alpha = 1000$ ,  $4N\lambda = 4 \times 10^{-5}$

### Estimation of $\alpha$ when the rate of sweep is known

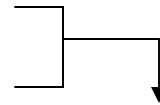


# Effect of demography

Real populations: bottleneck, expansion, subdivision, migration

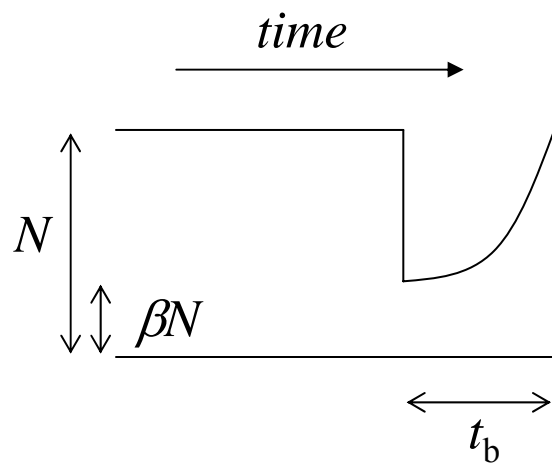
Effect: - Skew of frequency spectrum

- Increased noise in spatial pattern

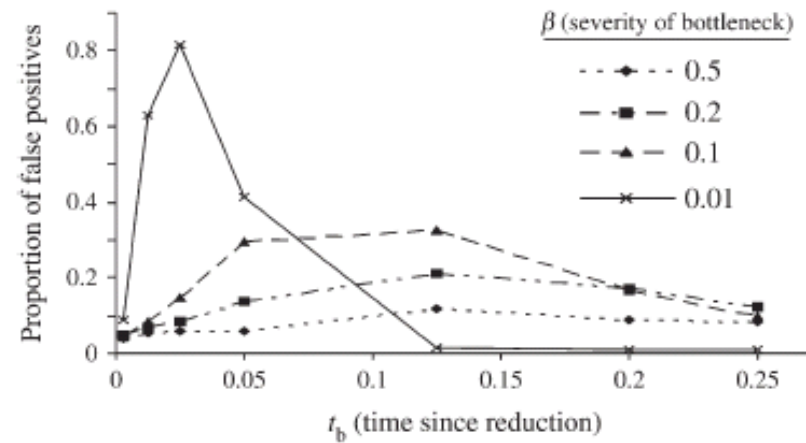


Masking true sweep patterns  
Creating “sweep-like” patterns

## False positive signals of sweeps due to bottlenecks

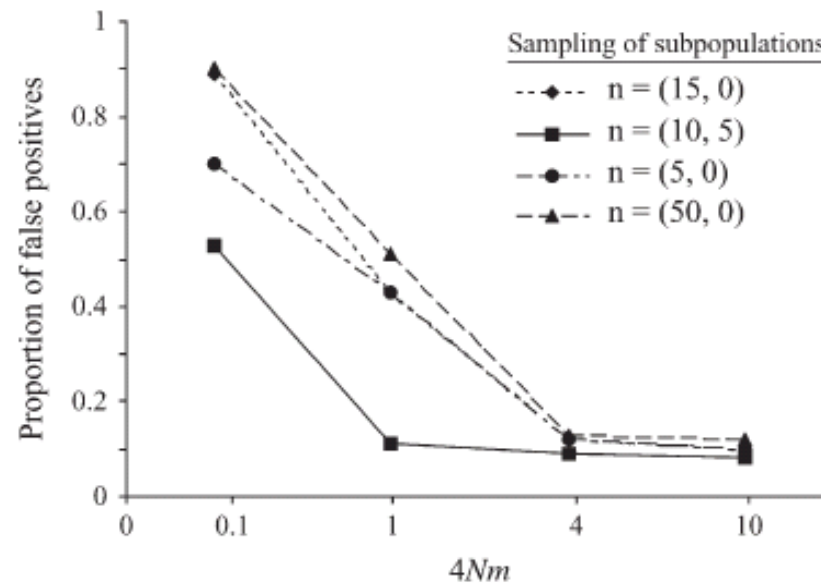


Kim&Stephan test  
(Jensen et al. 2005)



## False positive signals of sweeps due to population subdivision

Two subpopulations with migration → Kim&Stephan test  
(Jensen et al. 2005)

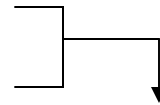


## Effect of demography

Real populations: bottleneck, expansion, subdivision, migration

Effect: - Skew of frequency spectrum

- Increased noise in spatial pattern



Masking true sweep patterns

Creating “sweep-like” patterns

Solutions: 1. Goodness-of-fit test

2. Composite likelihood with “background” spectrum

3. Exact modeling

## Goodness-of-fit test

(Jensen et al., 2005)

Kim & Stephan 2002: Likelihood ratio  $\Lambda_{KS} = \ln \frac{CL_S(\hat{\alpha}, \hat{X})}{CL_N}$



If significant, calculate  $\Lambda_{GOF} = \ln \frac{CL_A(\hat{P})}{CL_S(\hat{\alpha}, \hat{X})}$

where  $CL_A(\hat{P}) = \prod_{k=0}^{n-1} \hat{P}_k^{x_k}$ ,  $\hat{P}_k = \frac{x_k}{L}$

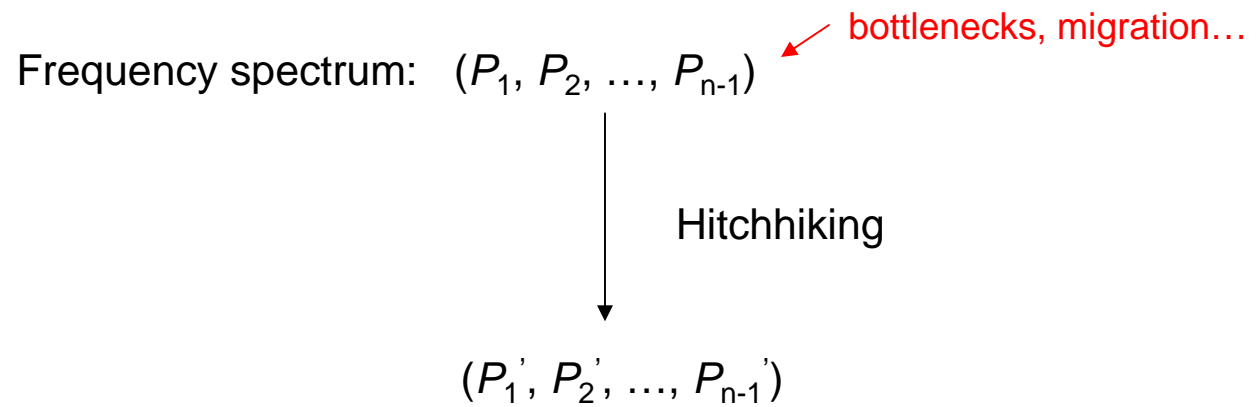
( $x_k$ : number of sites with frequency  $k$  out of  $n$ )



Compare  $\Lambda_{GOF}$  from test data and  
from simulated data (selective sweep with  $\hat{\alpha}, \hat{X}$ )

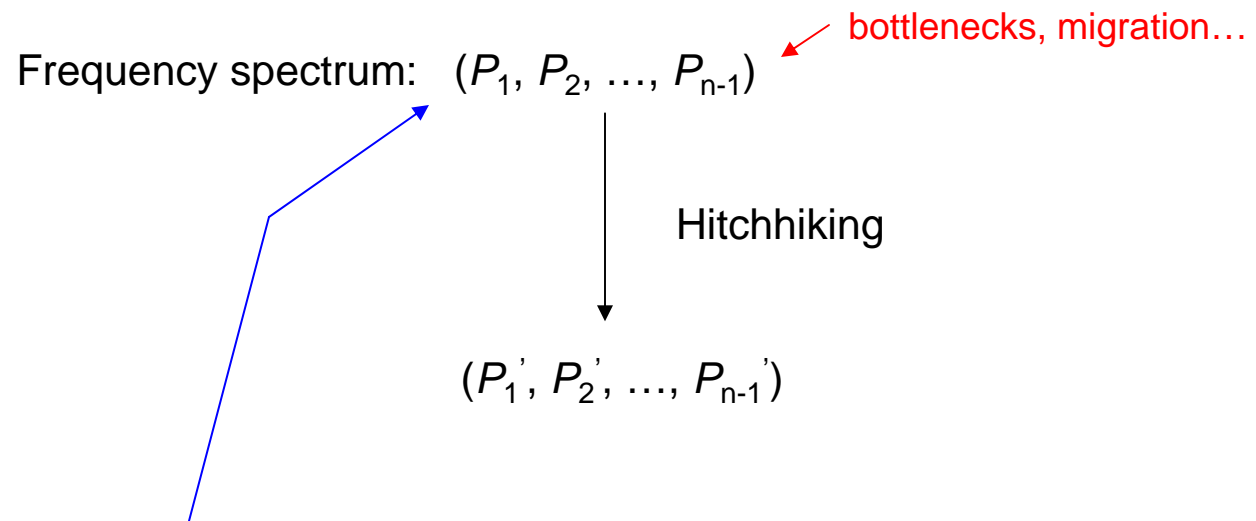
# Composite likelihood with “background” spectrum

(Nielsen et al. 2005)



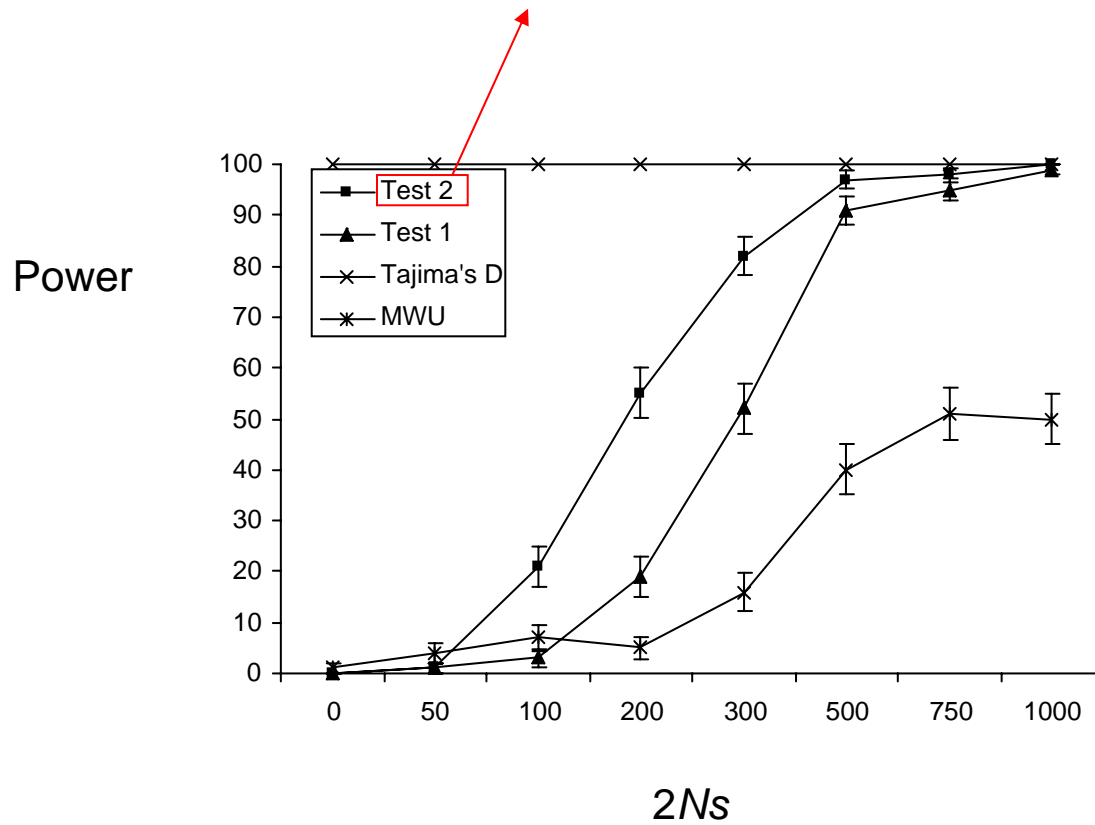
# Composite likelihood with “background” spectrum

(Nielsen et al. 2005)



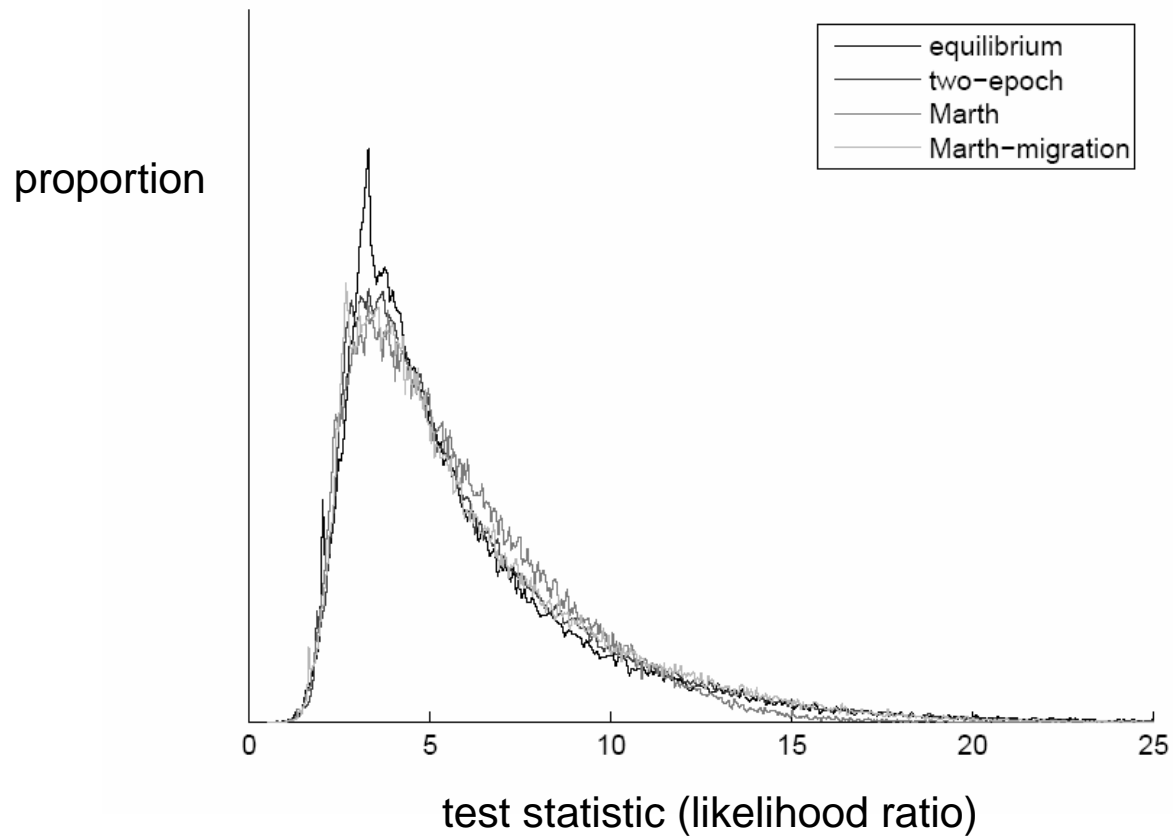
Estimated from genome-wide SNP data

# Composite likelihood with “background” spectrum (Test 2)



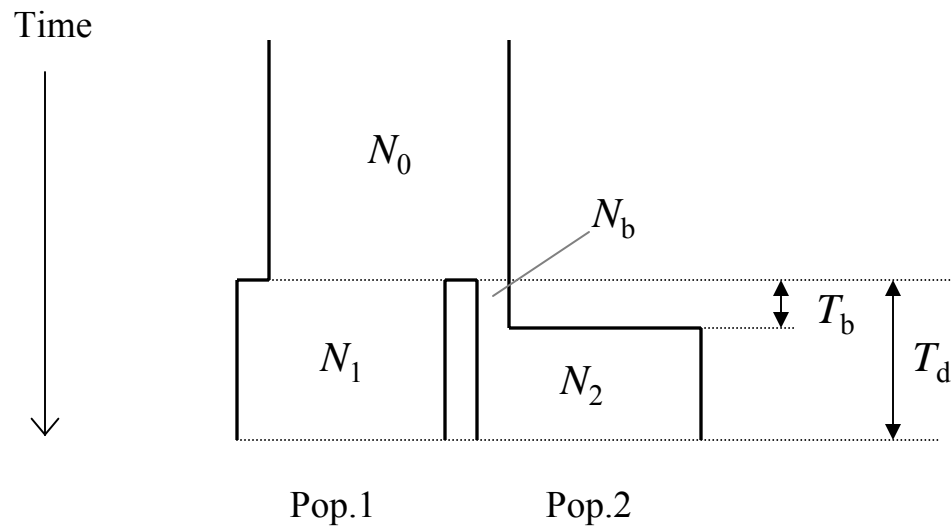
(Nielsen et al. 2005)

Composite likelihood with “background” spectrum:  
Null distribution of test statistic – robustness to demography



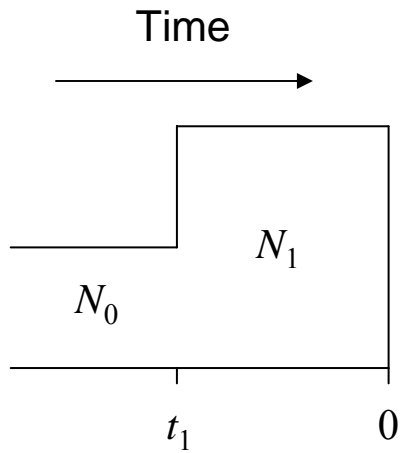
(Nielsen et al. 2005)

## Modeling selective sweeps under complex demography



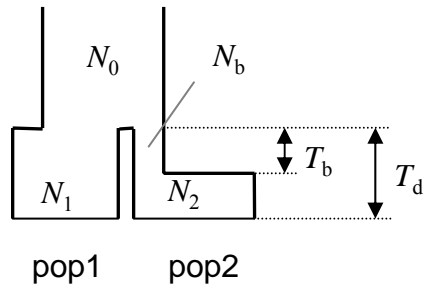
$$P_{ij} = \text{Prob}[i \text{ mutants in Pop.1 and } j \text{ mutants in Pop.2}]$$

## Scaling of time

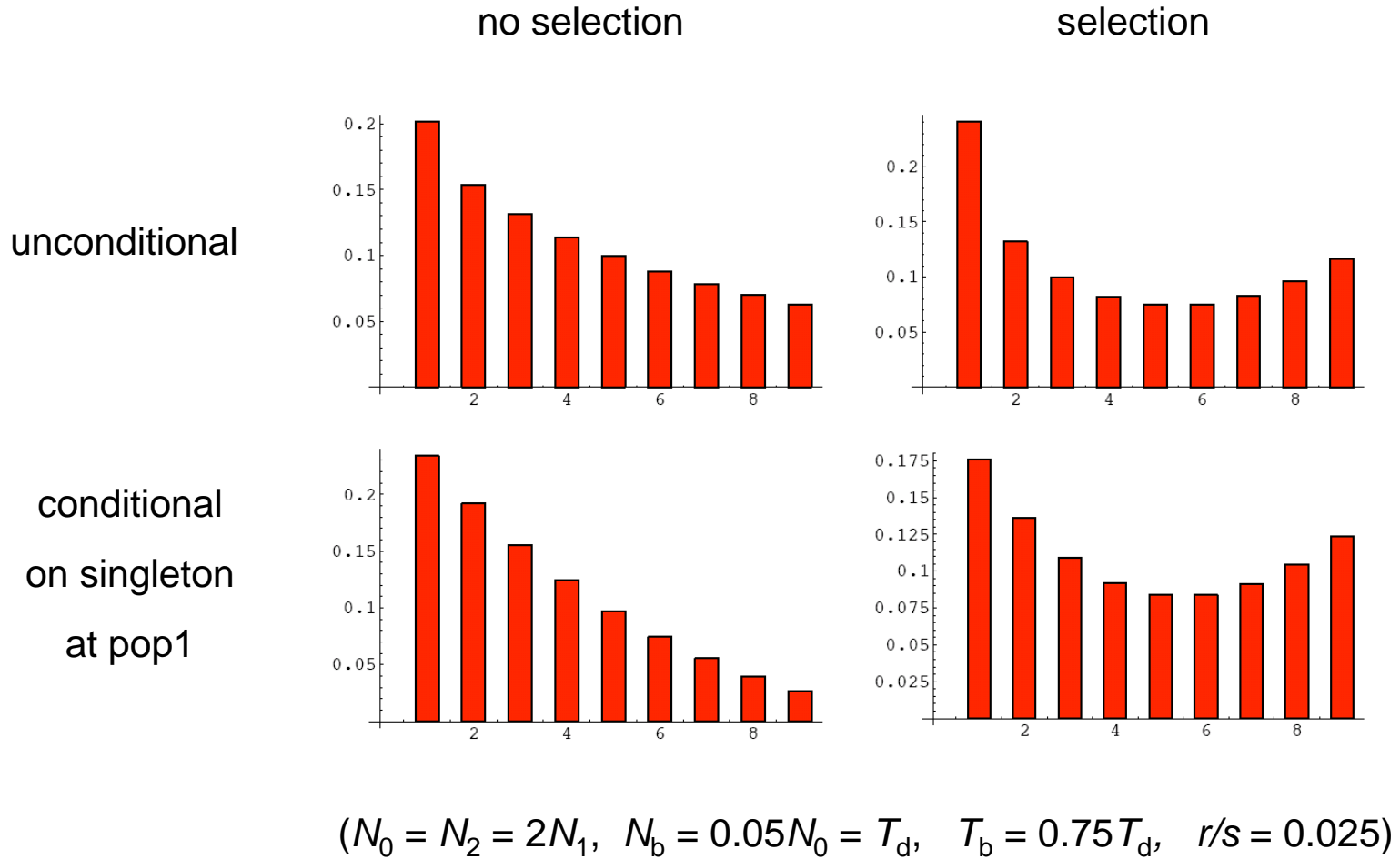


$$E[X^k | p, t] = \sum_{i=1}^k C_i^{(k)}(p) e^{-i(i-1)\tau}$$

$$\tau = \begin{cases} \frac{t-t_1}{4N_0} + \frac{t_1}{4N_1}, & (t \geq t_1) \\ \frac{t}{4N_1}, & (t < t_1) \end{cases}$$



## Frequency spectrum at pop2



# Further Problems

Single sweeps – approximations for population subdivision

Recurrent sweeps – any demographic model

Soft sweeps

Other models of adaptive substitutions

## Acknowledgement

- Hideki Innan
  - Jeff Jensen
  - Rasmus Nielsen
  - Carlos Bustamante
- 
- National Science Foundation
  - Arizona State University
    - ICMS

## Acknowledgement

- Hideki Innan
  - Jeff Jensen
  - Rasmus Nielsen
  - Carlos Bustamante
- 
- National Science Foundation
  - Arizona State University
    - ICMS
- Postdoctoral position available!