Natural Selection

Rasmus Nielsen

UC-Berkeley

Haploid Model

	A	a
Frequency	p	q
Relative Fitness	$w_A = 1$	$w_a = 1 + s$

Average fitness: $\overline{w} = w_A p + w_a q$

Allele frequency in next generation: $p' = \frac{w_A p}{\overline{w}}$

Change over many generations: $p_t = \frac{p_0}{p_0 + q_0(1+s)^t}$

Diploid Model

	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Relative Fitness	$w_{AA} = 1$	$w_{Aa} = 1 + hs$	$w_{aa} = 1 + s$

Average fitness: $\overline{w} = w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2$

Allele frequency in next generation: $p' = \frac{w_{AA}p^2 + w_{Aa}pq}{\overline{w}} = \frac{\overline{w}_A p}{\overline{w}}$

Directional selection: $w_{AA} \geq w_{Aa} \geq w_{aa}$ (neutrality: $w_{AA} = w_{Aa} = w_{aa}$)

Diploid Model

	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Relative Fitness	$w_{AA} = 1$	$w_{Aa} = 1 + hs$	$w_{aa} = 1 + s$

Average fitness: $\overline{w} = w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2$

Allele frequency in next generation: $p' = \frac{w_{AA}p^2 + w_{Aa}pq}{\overline{w}} = \frac{\overline{w}_A p}{\overline{w}}$

Directional selection: $w_{AA} \geq w_{Aa} \geq w_{aa}$ (neutrality: $w_{AA} = w_{Aa} = w_{aa}$)

Overdominance: $w_{AA} < w_{Aa} > w_{aa}$ (stable equilibrium)

Diploid Model

	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Relative Fitness	$w_{AA} = 1$	$w_{Aa} = 1 + hs$	$w_{aa} = 1 + s$

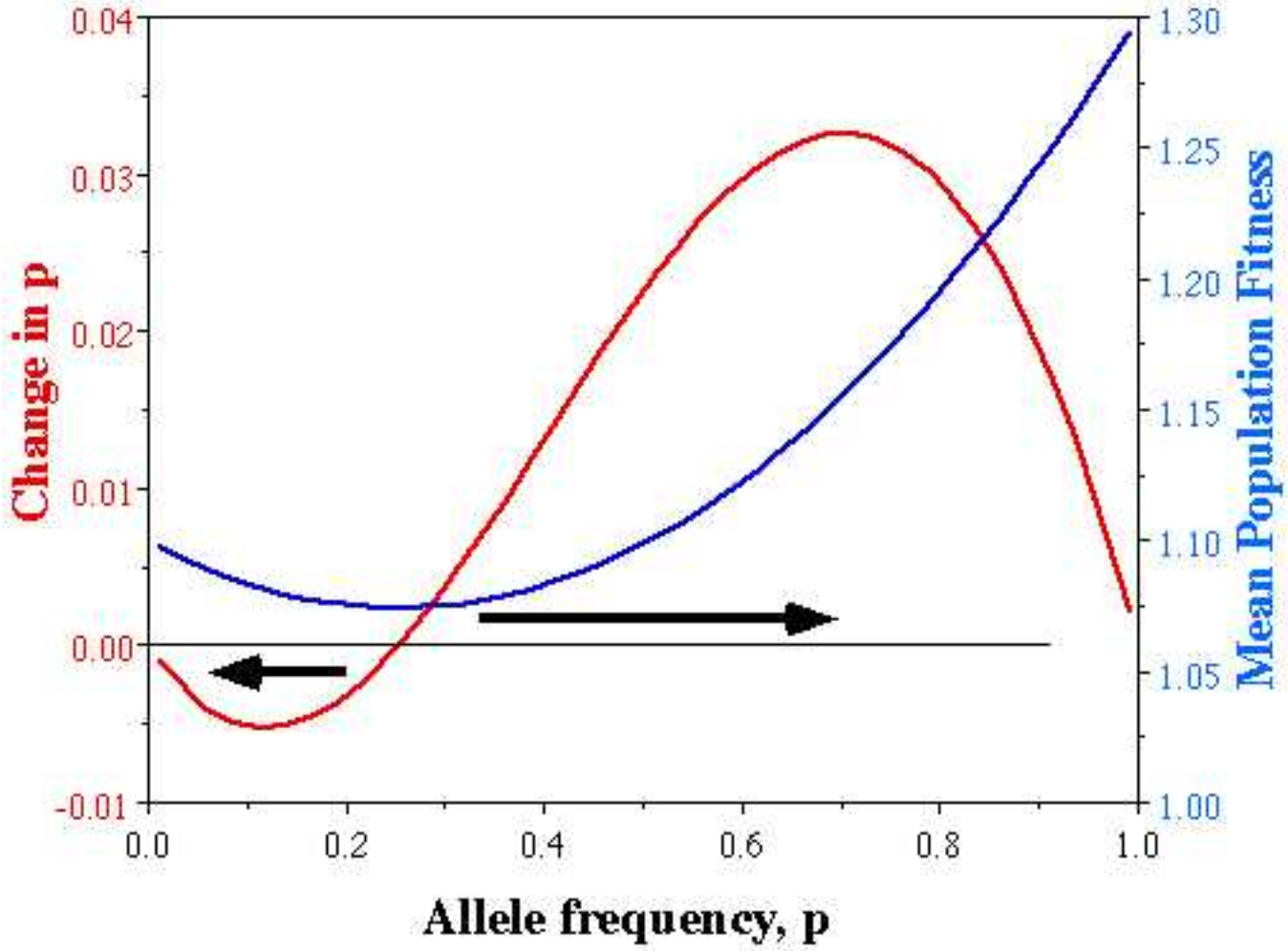
Average fitness: $\overline{w} = w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2$

Allele frequency in next generation: $p' = \frac{w_{AA}p^2 + w_{Aa}pq}{\overline{w}} = \frac{\overline{w}_{AA}p}{\overline{w}}$

Directional selection: $w_{AA} \geq w_{Aa} \geq w_{aa}$ (neutrality: $w_{AA} = w_{Aa} = w_{aa}$)

Overdominance: $w_{AA} < w_{Aa} > w_{aa}$ (stable equilibrium)

Underdominance: $w_{AA} > w_{Aa} < w_{aa}$ (either A or a will be fixed)


0.04
0.03
0.02
0.01
0.00
-0.01
Change in p
1.30
1.25
1.20
1.15
1.10
1.05
1.00
Mean Population Fitness
0.0
0.2
0.4
0.6
0.8
1.0
Allele frequency, p

Types of Selection

- Fertility versus viability selection
- Frequency dependent selection and fluctuating selection.
- Sexual selection
- Kin selection
- Group selection

Diploid Model

	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Relative Fitness	$w_{AA} = 1$	$w_{Aa} = 1 + hs$	$w_{aa} = 1 + s$

Average fitness: $\overline{w} = w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2$

Allele frequency in next generation: $p' = \frac{w_{AA}p^2 + w_{Aa}pq}{\overline{w}} = \frac{\overline{w}_{AA}p}{\overline{w}}$

Directional selection: $w_{AA} \geq w_{Aa} \geq w_{aa}$ (neutrality: $w_{AA} = w_{Aa} = w_{aa}$)

Overdominance: $w_{AA} < w_{Aa} > w_{aa}$ (stable equilibrium)

Underdominance: $w_{AA} > w_{Aa} < w_{aa}$ (either A or a will be fixed)

Typically we assume fitnesses of *AA*, *Aa* and *aa* to be 1, $1 - s$, and $(1 - s)^2 \approx 1 - 2s$, respectively.

Wright-Fisher (viability selection):

$$x' = \frac{x}{x + (1 - x)(1 - s)} = \frac{x}{1 - (1 - x)s} = x + x(1 - x)s + o(s)$$

Diffusion processes

$$E(Y_{1/N} - Y_0) = x(1-x)\gamma \cdot \frac{1}{2N} + o\left(N^{-1}\right)$$

$$\mathrm{var}\left(Y_{1/N} - Y_0\right) = x(1-x) \cdot \frac{1}{2N} + o\left(N^{-1}\right)$$

$$Lf = \frac{1}{4N} x(1-x) \frac{d^2}{dx^2} f + sx(1-x) \frac{d}{dx} f$$

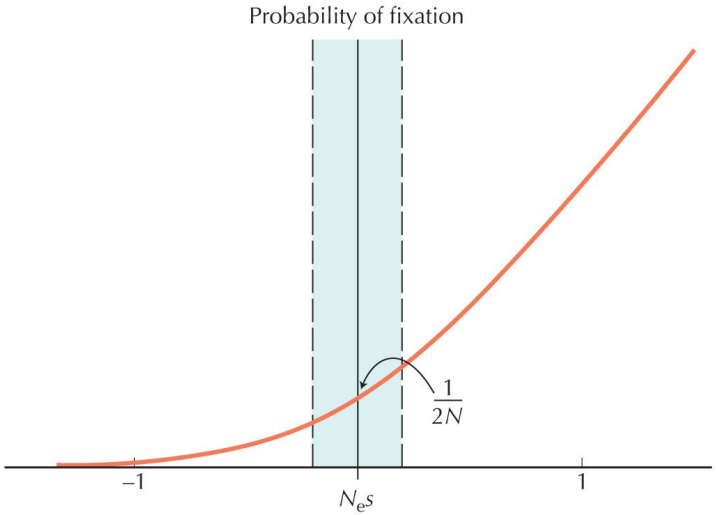


FIGURE 18.6. The probability that a single copy of an allele with selective advantage s will be fixed in a population of effective size N_e is $2s(N_e/N)/(1 - \exp(-4N_e s))$, where N is the actual number of individuals. The graph shows this probability plotted against $N_e s$, for $N_e = N$. If the allele is strongly favored ($N_e s >> 1$), then $P \sim 2s(N_e/N)$. If $N_e s$ is small, then drift is much stronger than selection ($1/2N_e >> s$), and so the allele is effectively neutral (*shaded strip*). Because each of the $2N$ genes in the population has the same chance of ultimately fixing, $P \sim 1/2N$ (see p. 425). Finally, if the allele is deleterious ($N_e s << -1$), then the probability of fixation becomes very small: $P \sim 2|s|(N_e/N)\exp(-4N_e|s|)$, where $|s|$ is the positive magnitude of selection (i.e., $-s$ if $s < 0$).

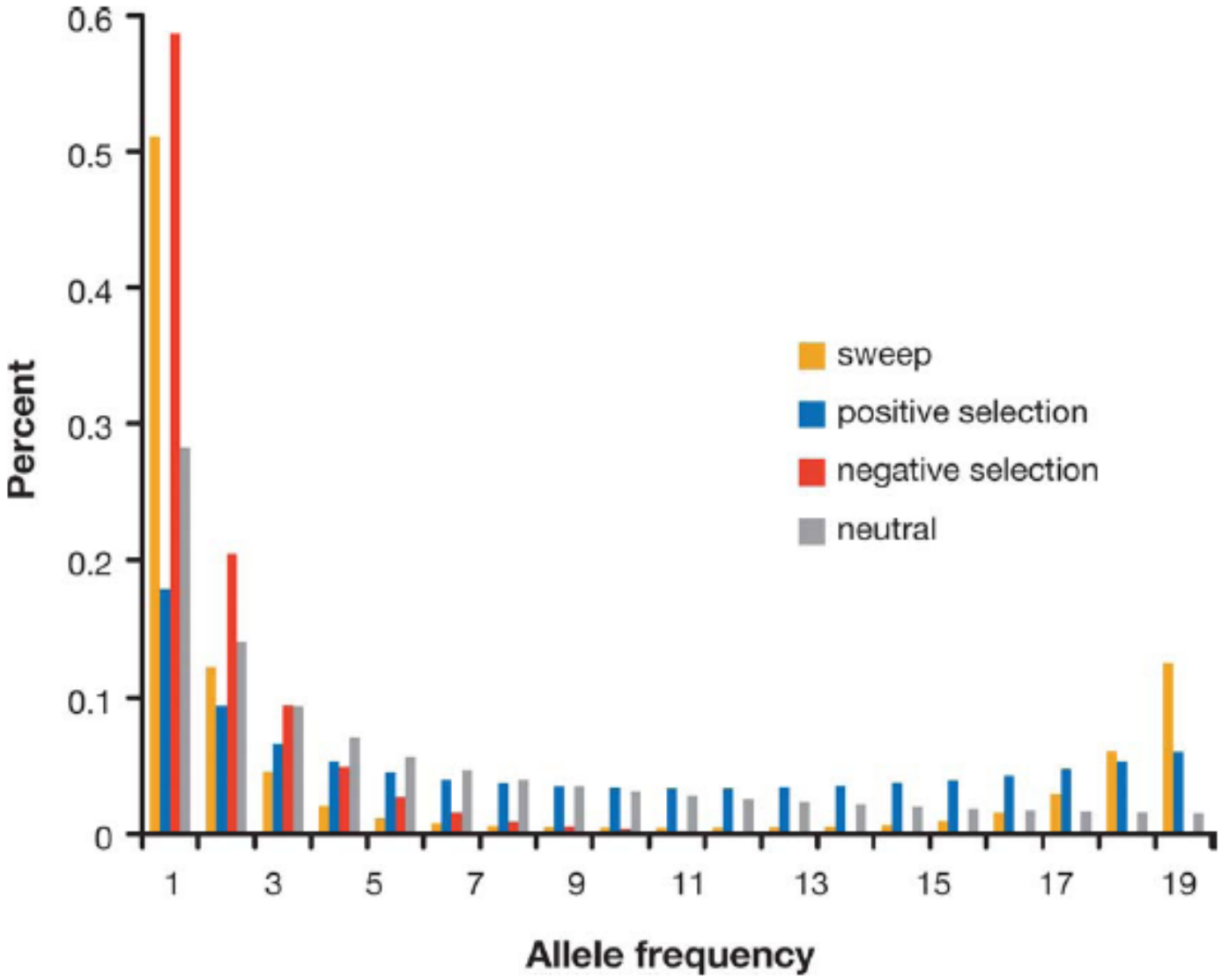
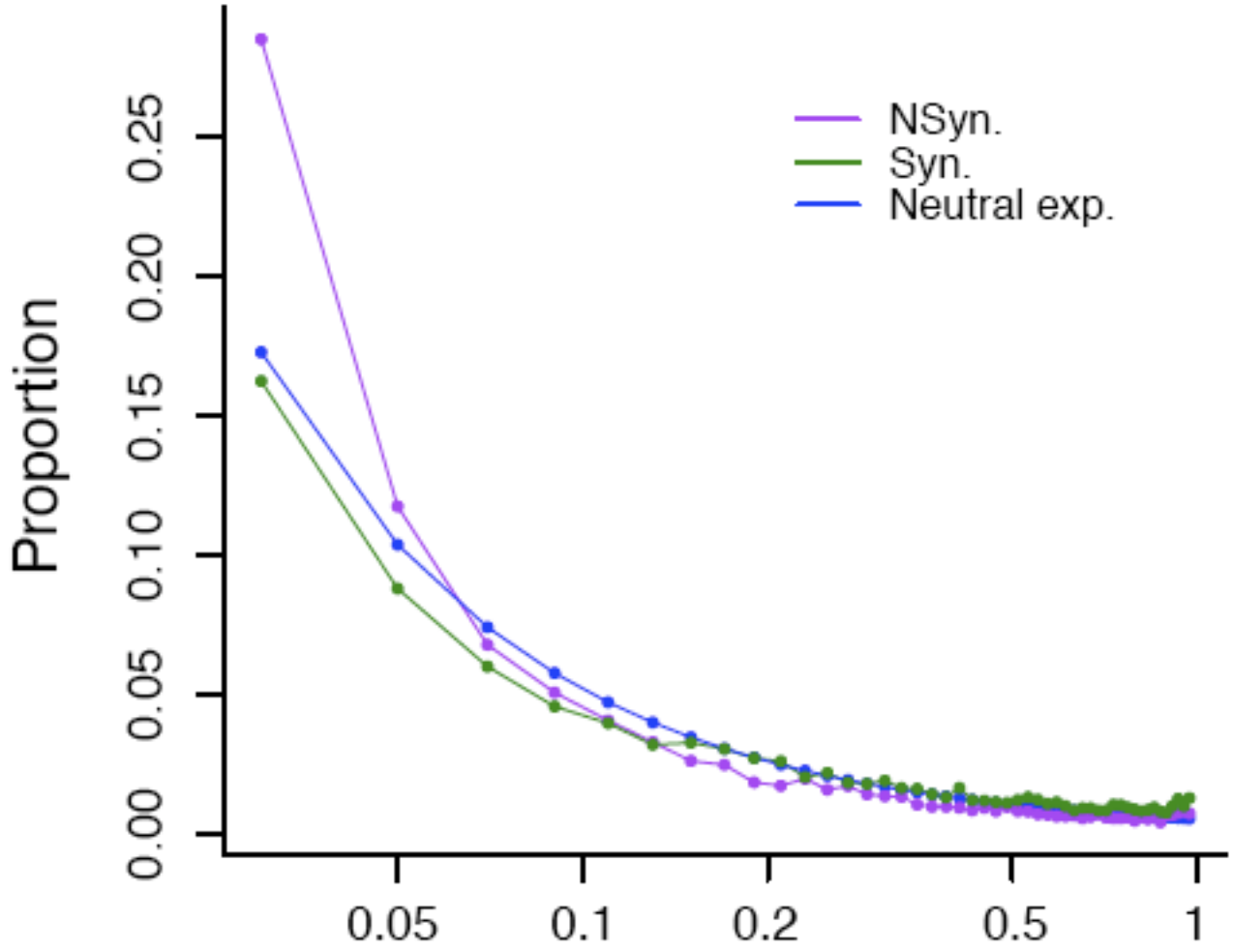

0.6
0.5
0.4
0.3
0.2
0.1
0
Percent
sweep
positive selection
negative selection
neutral
1
3
5
7
9
11
13
15
17
19
Allele frequency

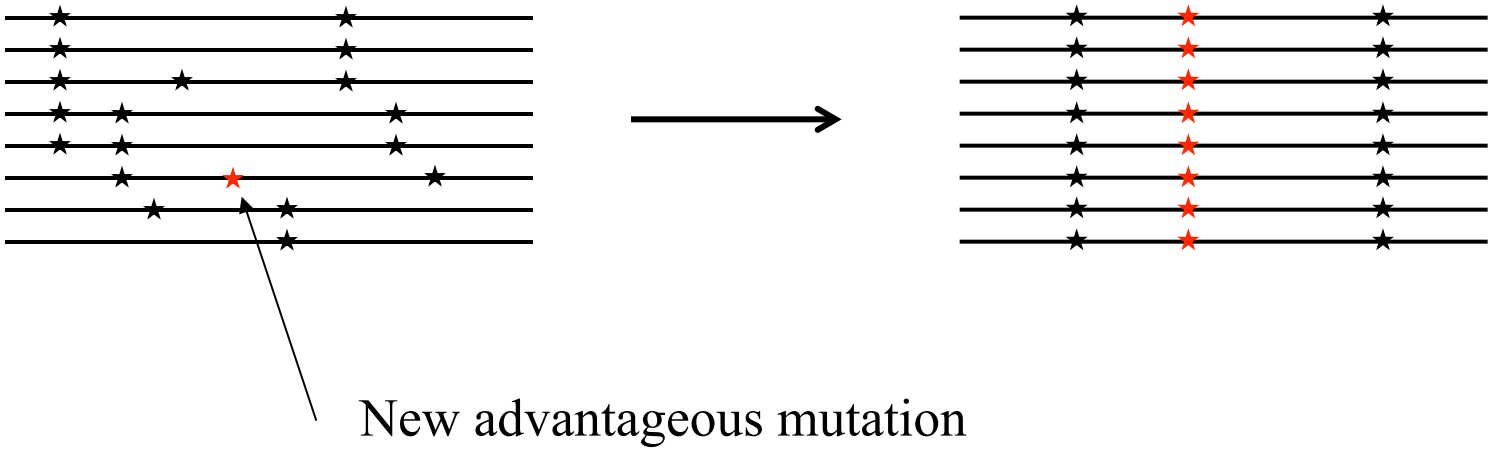
Figure 2

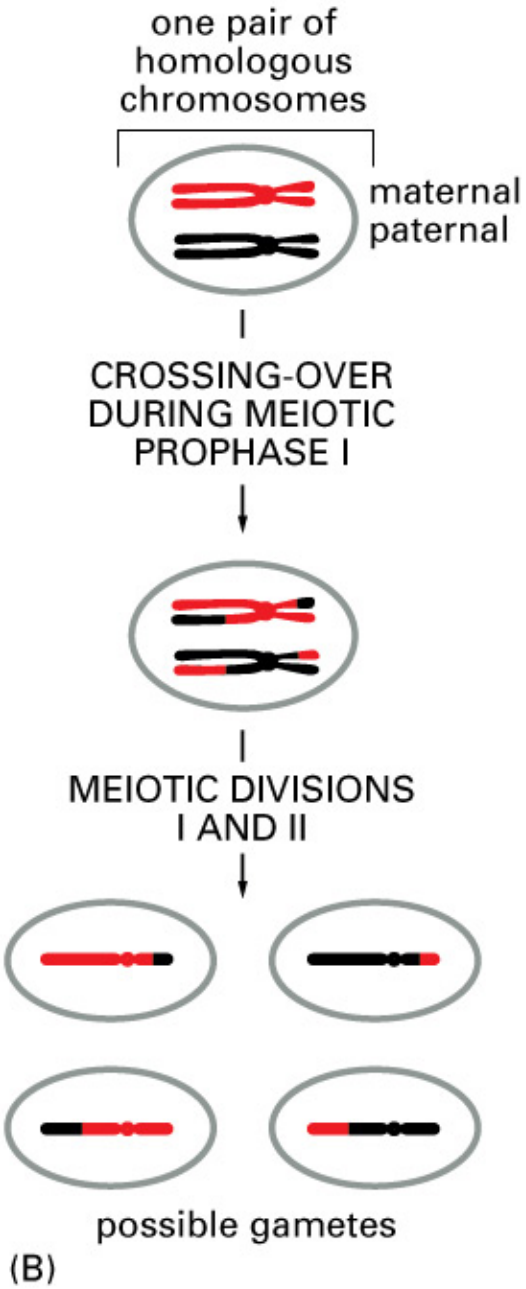

Proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.05
0.1
0.2
0.5
1
NSyn.
Syn.
Neutral exp.

Deleterious mutations

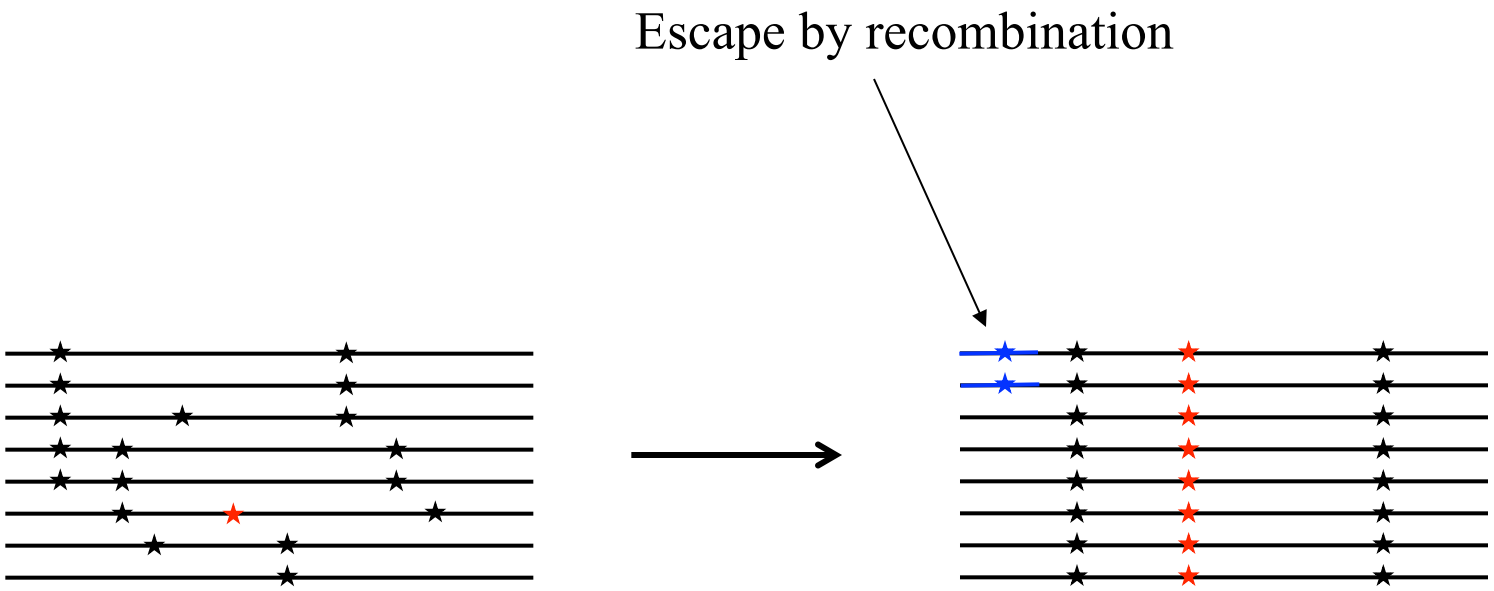
80% of all new non-synonymous SNPs must have selection coefficients in the range where they are negatively selected but not so deleterious that they will never be found in the population.

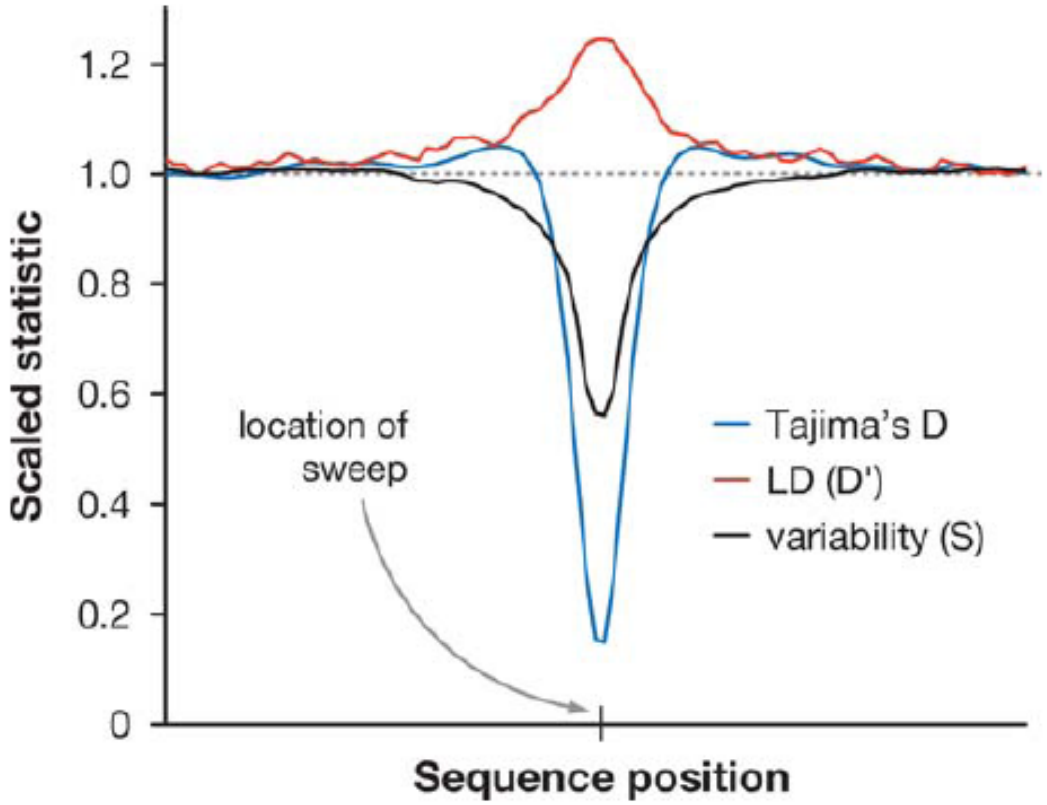
Selective Sweeps

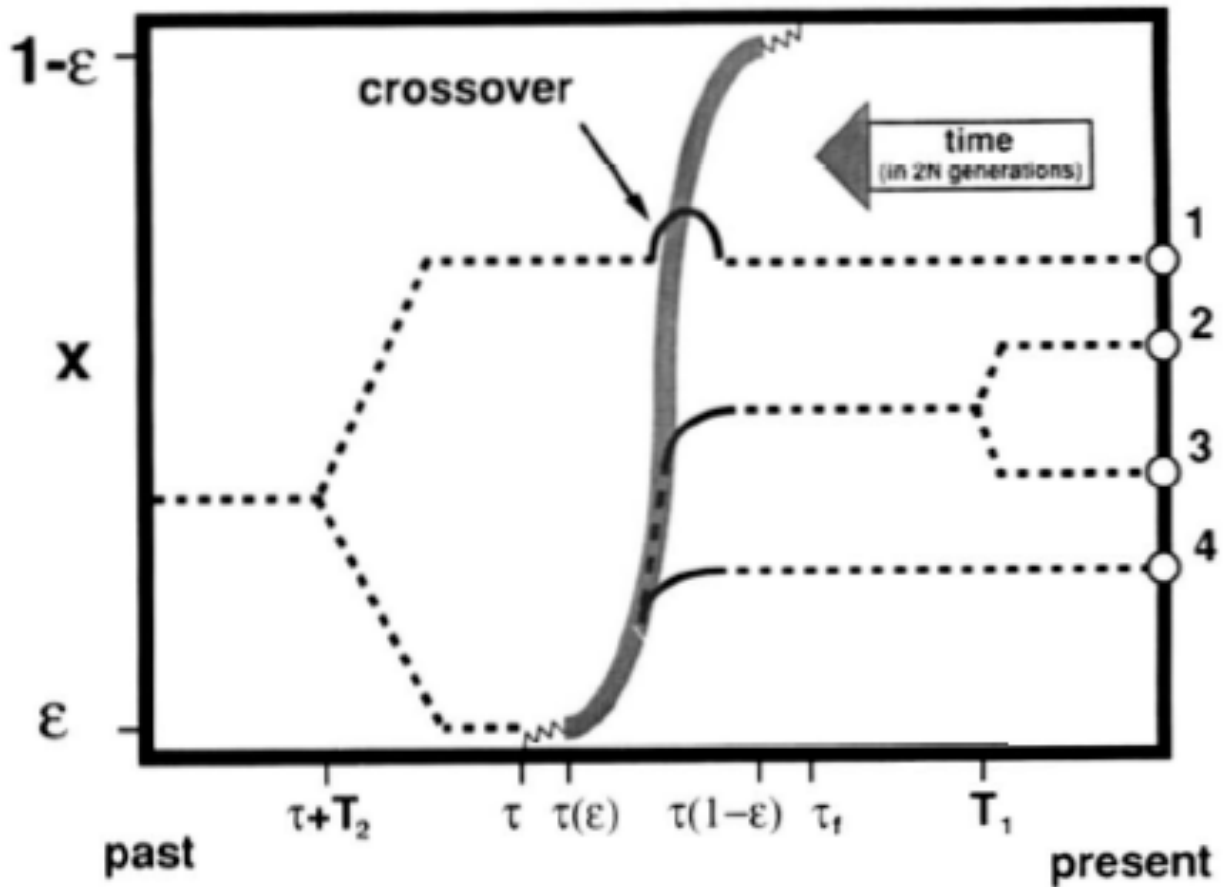



one pair of homologous chromosomes
maternal
paternal
CROSSING-OVER DURING MEIOTIC PROPHASE I
MEIOTIC DIVISIONS I AND II
possible gametes
(B)

Selective Sweeps

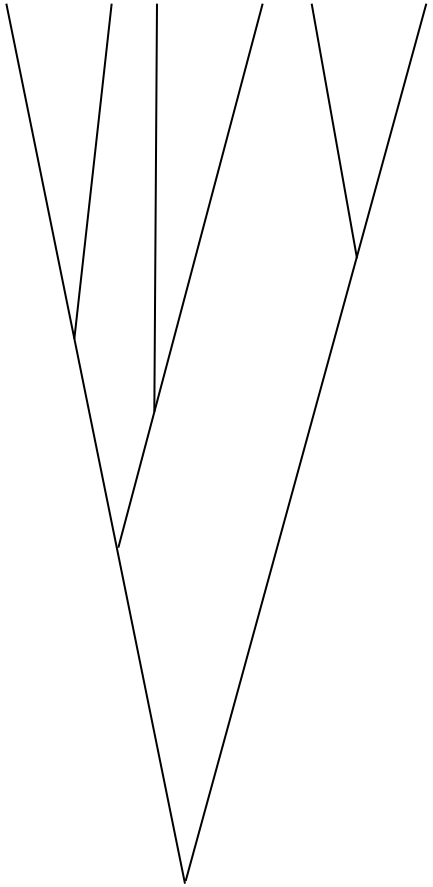



1.2
1.0
0.8
0.6
0.4
0.2
0
Scaled statistic
location of sweep
Tajima's D
LD (D')
variability (S)
Sequence position



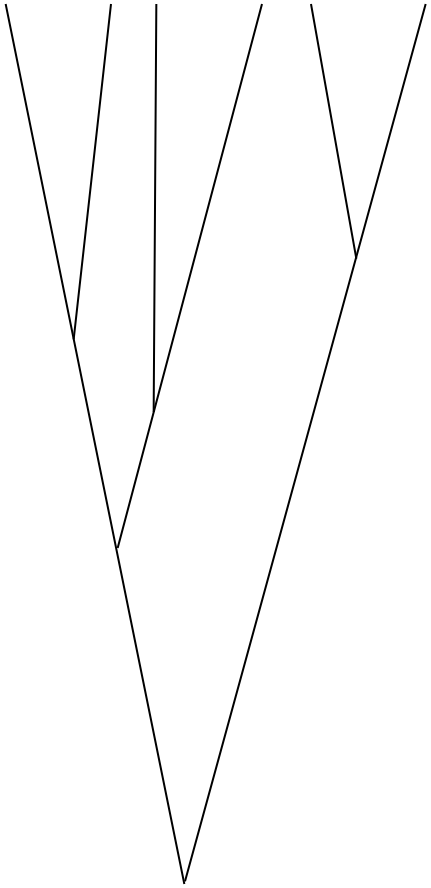
From Kaplan, Hudson and Langley (1989)

Coalescence Tree

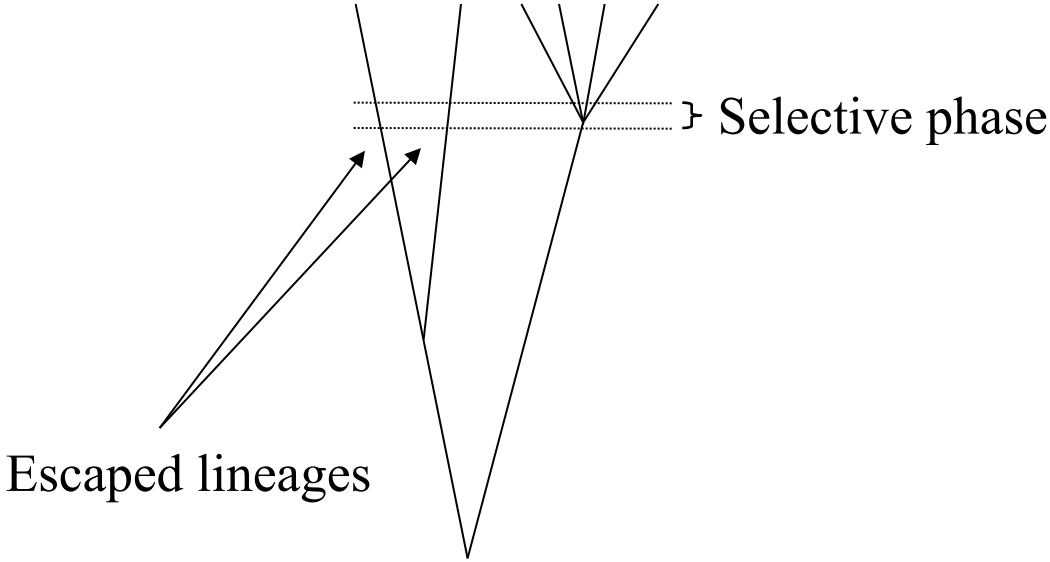


Coalescence Models

Neutral coalescence tree



Coalescence tree with sweep



Kaplan, Hudson and Langley (1989)

Three phases:

(1) When the selected allele is at low frequency ($x(t) < \varepsilon$), $2Nx(t)$ is modelled using a supercritical branching process.

(2) At intermediate frequencies, the the change in frequency is modelled deterministically:

$$\frac{dx(t)}{dt} = 2Nsx(t)\big(1 - x(t)\big)$$

(3) When the selected allele is at high frequencies ($x(t) > 1 - \varepsilon$), 1 then $2N(1 - x(t))$ follows a subcritical branching process.

$X(t)$: frequency of mutant at time t

N: population size

s: selection coefficient

Coalescent Process ($0 < x(t) < 1$)

For a sample of n gene copies in a neutral locus, the coalescent process has state space on $\{(i, j) : 1 \leq i + j \leq n\}$ and follows a time-inhomogeneous Markov jump process, which jumps from state (i, j) at rates

$$q_{i-1,j}(x(t)) = \binom{i}{2} x(t)^{-1}$$

$$q_{i,j-1}(x(t)) = \binom{j}{2} (1 - x(t))^{-1}$$

$$q_{i+1,j-1}(x(t)) = jRx(t)$$

$$q_{i-1,j+1}(x(t)) = iR(1 - x(t))$$

$R = 2Nr$, r = recombination rate per generation

Various Extensions and Simplifications

- Stephan, Wiehe and Lentz (1992), Wiehe and Stephan (1993), Kim and Stephan (2002) and others ignore the stochastic phases.
- Durrett and Schweinsberg (2004) showed that under this assumption and $N \to \infty$, $r\ln(2N)/s \to a$, $s(\ln N)^2 \to \infty$ then for $j > 1$

$$p_{k,k-j+1} \to \binom{k}{j} p^j (1-p)^{k-j}, \text{ where } p = e^{-a}$$

Where $p_{n,k}$ is the probability that n ancestral lineages are reduced to k ancestral lineages after a selective sweep (looking backwards in time).

- More theory on multiple mergers: Barton (1998) Durrett and Schweinsberg (2004), Etheridge et al. (2006), Pfaffelhuber et al. (2006) and others.

Krone and Neuhauser (1997)
Ancestral Selection Graph

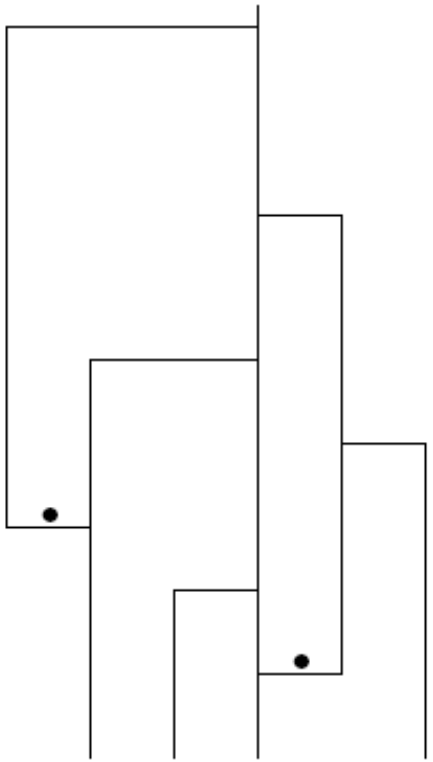


Fig. 6.1. A realization of the ancestral selection graph for a sample of size 4.

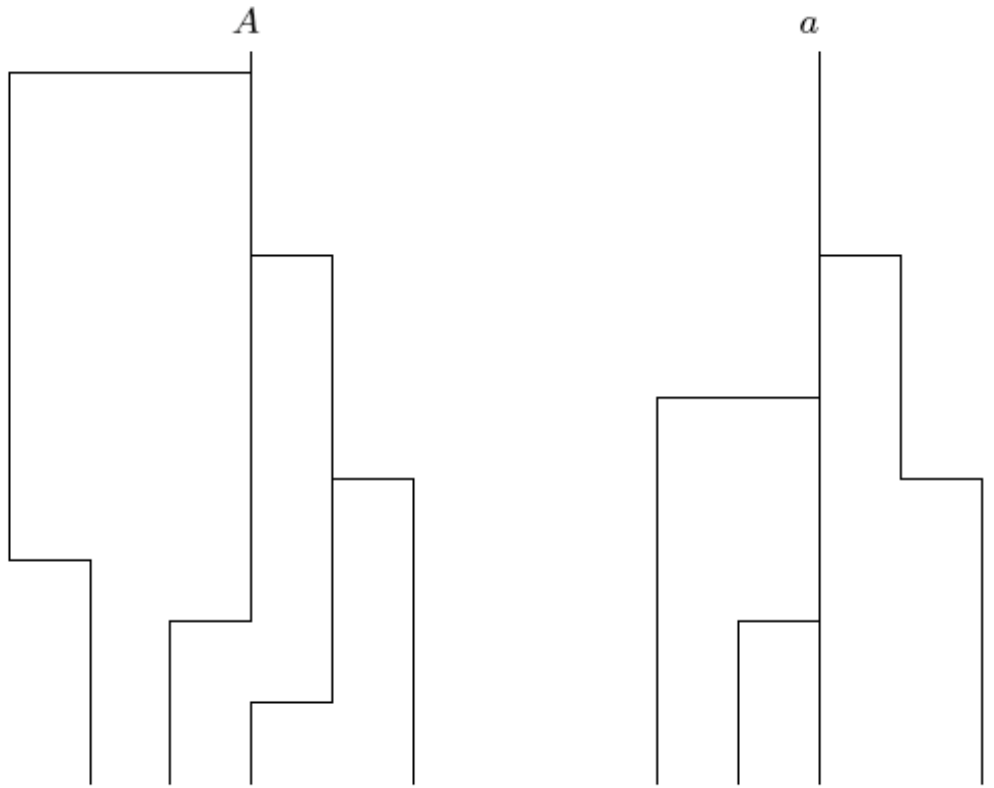


Fig. 6.2. The actual genealogy depends on the state of the ultimate ancestor.

Background Selection

- Charlesworth, Morgan and Charlesworth (1993): the shape of the genalogy will be just as in the neutral case but with

$$N_e = Ne^{-u/2sh}$$

  sc $\pi/\pi_0 \approx e^{-u/2sh}$, where *u* is the deleterious mutation rate.

- Adding recombination, Nordborg, Charlesworth and Charlesworth (1996):

$$\frac{\pi}{\pi_0} \approx \exp\left(-\int_0^L \frac{u(x)sh}{2(sh+|M(x)-M(y)|)^2}\,dx\right)$$

Motoo Kimura (1969): The Neutral Theory

Molecular evolution is dominated by mutation and genetic drift. Selection plays only a minor role.

To tests this hypothesis – and to detect selection in general – a number of Tests of Neutrality have been developed

Evolutionary factor	Intraspecific variability[a]	Interspecific variability	Ratio of interspecific to intraspecific variability	Frequency spectrum
Increased mutation rate	Increases	Increases	No effect	No effect
Negative directional selection	Reduced	Reduced	Reduced if selection is not too strong.	Increases the proportion of low frequency variants
Positive directional selection	May increase or decrease	Increased	Increased	Increases the proportion of high frequency variants
Balancing selection	Increases	May increase or decrease	Reduced	Increases the proportion of intermediate frequency variants
Selective sweep (linked neutral sites)	Decreased	No effect on mean rate of substitution, but the variance increases.	Increased	Mostly increases the proportion of low frequency variants.

First base	Second base: U			Second base: C			Second base: A			Second base: G			Third base
U	UUU	Phenylalanine	F	UCU	Serine	S	UAU	Tyrosine	Y	UGU	Cysteine	C	U
	UUC	Phenylalanine	F	UCC	Serine	S	UAC	Tyrosine	Y	UGC	Cysteine	C	C
	UUA	Leucine	L	UCA	Serine	S	UAA	Stop		UGA	Stop		A
	UUG	Leucine	L	UCG	Serine	S	UAG	Stop		UGG	Tryptophan	W	G
C	CUU	Leucine	L	CCU	Proline	P	CAU	Histidine	H	CGU	Arginine	R	U
	CUC	Leucine	L	CCC	Proline	P	CAC	Histidine	H	CGC	Arginine	R	C
	CUA	Leucine	L	CCA	Proline	P	CAA	Glutamine	Q	CGA	Arginine	R	A
	CUG	Leucine	L	CCG	Proline	P	CAG	Glutamine	Q	CGG	Arginine	R	G
A	AUU	Isoleucine	I	ACU	Threonine	T	AAU	Asparagine	N	AGU	Serine	S	U
	AUC	Isoleucine	I	ACC	Threonine	T	AAC	Asparagine	N	AGC	Serine	S	C
	AUA	Isoleucine	I	ACA	Threonine	T	AAA	Lysine	K	AGA	Arginine	R	A
	AUG	Start (Methionine	M)	ACG	Threonine	T	AAG	Lysine	K	AGG	Arginine	R	G
G	GUU	Valine	V	GCU	Alanine	A	GAU	Aspartic Acid	D	GGU	Glycine	G	U
	GUC	Valine	V	GCC	Alanine	A	GAC	Aspartic Acid	D	GGC	Glycine	G	C
	GUA	Valine	V	GCA	Alanine	A	GAA	Glutamic Acid	E	GGA	Glycine	G	A
	GUG	Valine	V	GCG	Alanine	A	GAG	Glutamic Acid	E	GGG	Glycine	G	G

RNA Codon Amino acid Abbreviation

Copyright © 2004 Pearson Prentice Hall, Inc.

Nonsynonymous/synonymous rate ratio: d_N/d_S

d_N = number of nonsynonymous mutations per nonsynonymous site.

d_S = number of synonymous mutations per synonymous site.

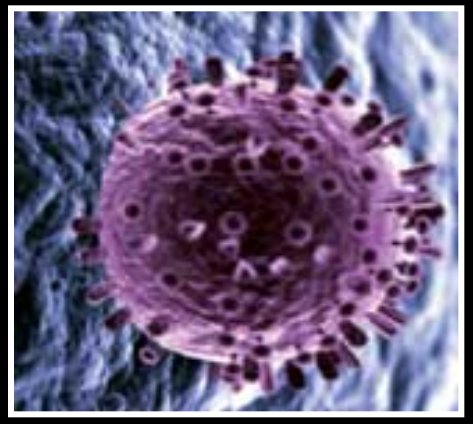
$d_N/d_S < 1$: Negative selection

$d_N/d_S = 1$: Neutrality (no selection)

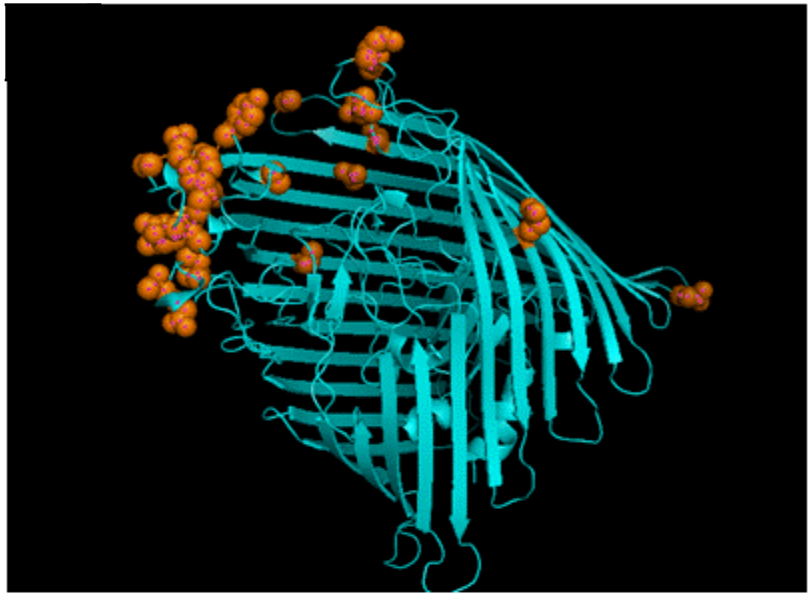
$d_N/d_S > 1$: Positive selection

Selection for avoidance of immune recognition in viruses

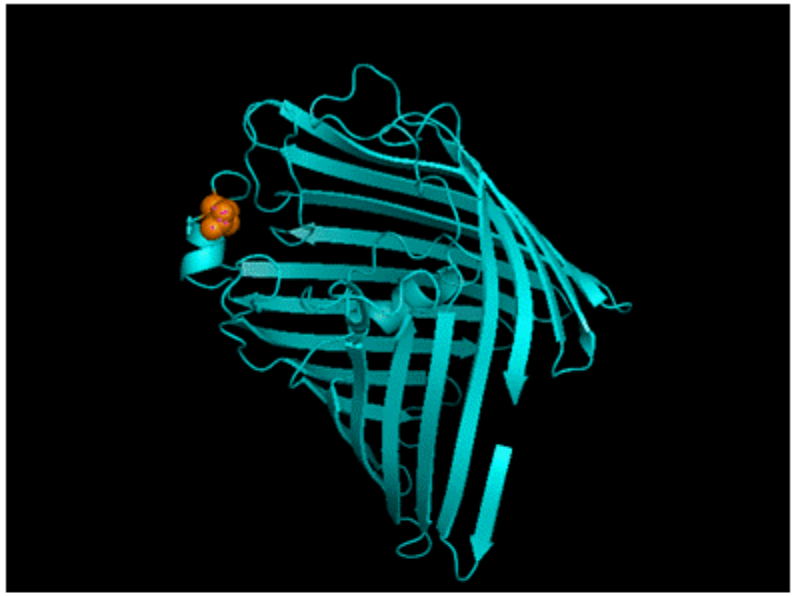
Influenza
hemagglutinin
molekyle



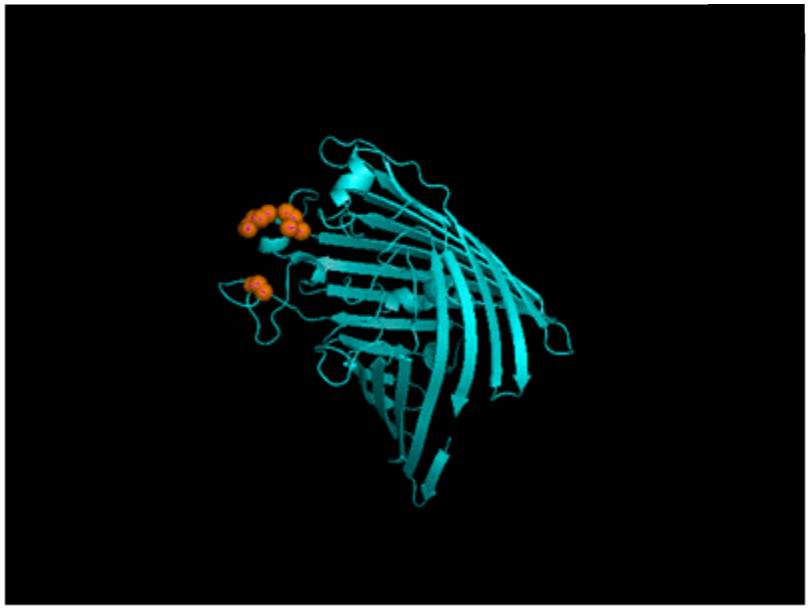
Gene categories that show evidence of positive selection in *E. coli*			
EcoCyc category	**Description**	**Category size**	**Fisher p-value**
BC-8	Extrachromosomal	142	0.00E-00
BC-8.3	Transposon related	48	0.00E-00
BC-4.1.B	Beta barrel porins	20	2.73E-14
BC-8.1	Prophage genes and phage related functions	112	4.91E-14
BC-7.4	Outer membrane	63	2.14E-10
BC-1.6.3.2	Core region	11	2.54E-07
BC-1.5.3.11	Menaquinone, ubiquinone	28	1.03E-05
BC-1.6.11	Glycoprotein (incl. some fimbrie and curlin protein)	12	4.21E-05
BC-8.4	Colicin related	10	0.00366



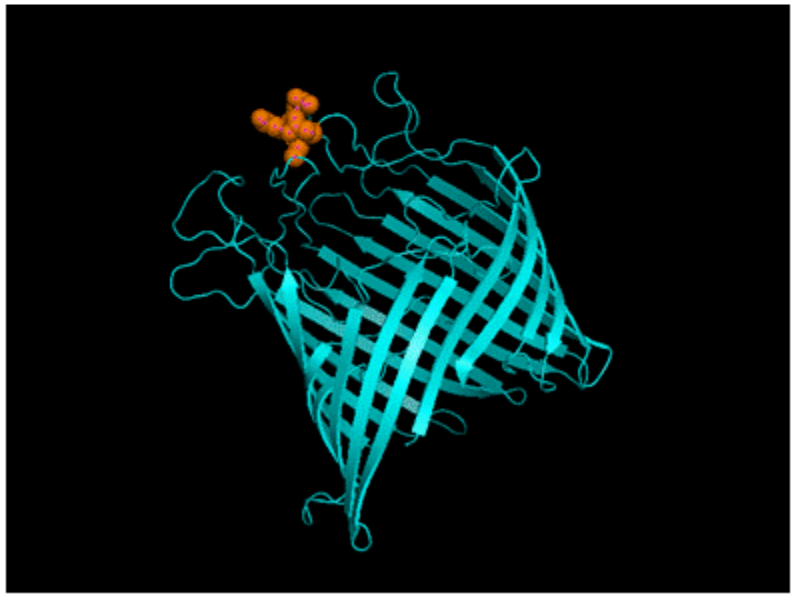
FhuA



OmpF



OmpC



LamB

Biological process	**Number of genes**	***p*-value**
Immunity and defense	417	0.0000
T-cell mediated immunity	82	0.0000
Chemosensory perception	45	0.0000
Biological process unclassified	3069	0.0000
Olfaction	28	0.0004
Gametogenesis	51	0.0005
Natural killer cell mediated immunity	30	0.0018
Spermatogenesis and motility	20	0.0037
Inhibition of apoptosis	40	0.0047
Interferon-mediated immunity	23	0.0080
Sensory perception	133	0.0160
B-cell- and antibody-mediated immunity	57	0.0298

McDonald-Kreitman test

	Within species	**Between species**
Nonsynonymous	**A**	**B**
Synonymous	**C**	**D**

Neutrality: $\frac{A}{B} = \frac{C}{D}$

Test using test of homogeneity.

1
OR51B2
MGC20470
PRKCDBP
ZNF143
SH2BP1
FAM10A5
MGC34830
DEPC(1
DKFZP586B162
C11orf11
FADS3
FLJ35827
SIPA1
DPP3
POLD4
PGM2L1
MOGAT2*
CNTN5
VPS11
SLC37A2
RPUSD4
10
THEDC1
PLXDC2
CSTF2T
SLC29A3
HSGT1
ANXA11
IFIT2
CYP26C1
FER1L3
MMS19L
SUFU*
WDR11*
DOCK1
9
DMRT3
C9orf138
IFNA14
DDX58
PAF53
TJP2*
C9orf79
CEP1
ABO
KIAA0649
8
MYOM2
MCPH1
PCM1*
ChGn
ADRA1A
MCM4
NSMAF
GGH
TRPA1
SPAG1
NUDCD1
TAF2
DEPDC6
FLJ31164
PYCRL
7
FLJ31340
STEAP2
DOCK4
GCC1
TAS2R4
6
GCNT2*
KIF13A
TRIM31
RPP21
C6orf15
CDSN
C4B*
CYP21A2*
PSMB9
HLA(DPA1
TREML1
TINAG
NT5E
WISP3*
AKAP7
STXBP5
C6orf211
UNC93A
UNC93A
MGC43690
5
IL7R*
LOC134285
IQGAP2
MASS1*
FLJ36090
PSD2
PCDHA13
SH3RF2
SPINK5*
TCOF1*
GM2A*
MGC27121
PC(LKC
4
ZNF595
ADD1*
PPARGC1A*
CLOCK
UGT2B7
SAS10
SDAD1
FRAS1
SEC31L1
LOC285513
SMARCAD1
CENPE
LOC255743
FLJ11331
FAT4
OSAP
3
TMEM43
GORASP1
TTC21A
ZNF167
FYCO1
TDGF1*
FLJ42117
CD96
CD200
POLQ
ILDR1
MCM2
GATA2
PPP2R3A
KIAA0669
EVI1
DGKG
LPP*
LRCH3
2
SMC6L1
KCNS3
NT5C1B
FLJ20272
FEZ2
QPCT
ASB3
MGC15407
PLEK
ANXA4
ADD2
CLEC4F
TCF7L1
FLJ10916
IL1R1
IL18RAP
IL1F6
MARCO
NEB*
LRP2
KBTBD10
NFE2L2
CERKL
COL3A1*
SLC40A1
PMS1*
GMPPA
COL6A3*
1
TNFRSF25
H6PD
CASP9
CLCNKB*
SDHB*
PADI4
ID3
LOC56181
TACSTD2*
GBP4
FAD158
AGL*
FLJ35838
CD58
PGLYRP4
S100A5
ARHGEF2
FLJ32884
IRTA2
KIRREL
SLAMF6
MGC9084
ASTN
RNASEL
C1orf16
DBCCR1L
CRB1*
PKP1*
PLXNA2
CAPN9
X
SCML1
ZFX
SRPX
KIF4A
DRP2
SUHW3
SAGE1
MAGEC2
IRAK1
21
C21orf91
TIAM1
CBR3
SLC19A1
COL6A2*
22
CLTCL1
FLJ33814
C22orf19
APOL3
APOL4*
C1QTNF6
MICAL(L1
APOBEC3B
CGI(96
RIBC2
20
POFUT1
C20orf186
RBPSUHL
NEURL2
ZBP1
COL9A3*
SRMS
OPRL1
19
CSNK1G2
PRKCSH*
FLJ38281
COPE
DDX49
GMIP
ZNF431
ZNF493
ECH1
GMFG
LTBP4
TGFB1*
XRCC1
ASE(1
EML2
GPR77
ELSPBP1
CABP5
PLA2G4C
SLC6A16
FLJ20643
MYBPC2
SIGLEC7
LILRA2
18
ADCYAP1
SERPINB10
CNDP2
ATP9B
17
FLJ10534
GPS2
MYH3
COX10*
C17orf32
SPAG5
LIG3
DKFZp434H221
KRTAP3(2
HOXB1
CD300A
OTOP2
OTOP3
FOXK2
16
MAPK8IP3
PRM1
TNFRSF17
KIAA0251
MYH11
LOC57146
FLJ20718
CDH11
CES2
FLJ21918
FLJ13291
NQO1*
MLKL
WFDC1
KLHDC4
15
MAGEL2
BAHD1
EPB42*
CILP
CTSH
14
TEP1
APEX1
ISGF3G
C14orf124
GZMH
SCFD1
SRP54
CTAGE5
C14orf29
PYGL*
C14orf31
SEC10L1
C14orf108
ARID4A
PNMA1
GALC*
SERPINA5*
13
B3GTL
ELF1
HTR2A
12
MGC4266
FOXJ2
SSPN
C12orf11
MGC24039
SYT10
MGC5576
BIN2
ESPL1
KIAA1536
MGC39497
NR2C1
TMPO
APAF1
ACAD10*
P2RX7*
HPD*
NCOR2
DHX37

The HKA test (Hudson-Kreitman-Aquade)

In the HKA test, the levels of polymorphism and divergence in two or more loci are considered:

	Locus 1	Locus 2
Segregating sites	S_1	S_2
Fixed differences	F_1	F_2

$$X^2 = \sum_{i=1}^{2} \frac{\left(F_i - \hat{E}(F_i)\right)^2}{\hat{V}(F_i)} + \sum_{i=1}^{2} \frac{\left(S_i - \hat{E}(S_i)\right)^2}{\hat{V}(S_i)}$$

Dmd intron 7 and 44 in humans

Nachman and Crowell 2000

HKA tests comparing *Dmd* intron 7 *vs.* intron 44, *Homo vs. Pan*

Geographic region	Locus	*S*	*D*	HKA χ^2	*P* value
Africa	Intron 7	6	39		
	Intron 44	15	27	2.51	NS
Europe	Intron 7	1	39		
	Intron 44	10	27	5.10	<0.05
Asia	Intron 7	0	39		
	Intron 44	10	27	7.01	<0.01
Americas	Intron 7	3	39		
	Intron 44	9	27	2.52	NS
World	Intron 7	9	39		
	Intron 44	19	27	3.08	0.08
non-Africa	Intron 7	4	39		
	Intron 44	15	27	5.01	<0.05

NS, not significant.

Tajima's *D* Test (1989)

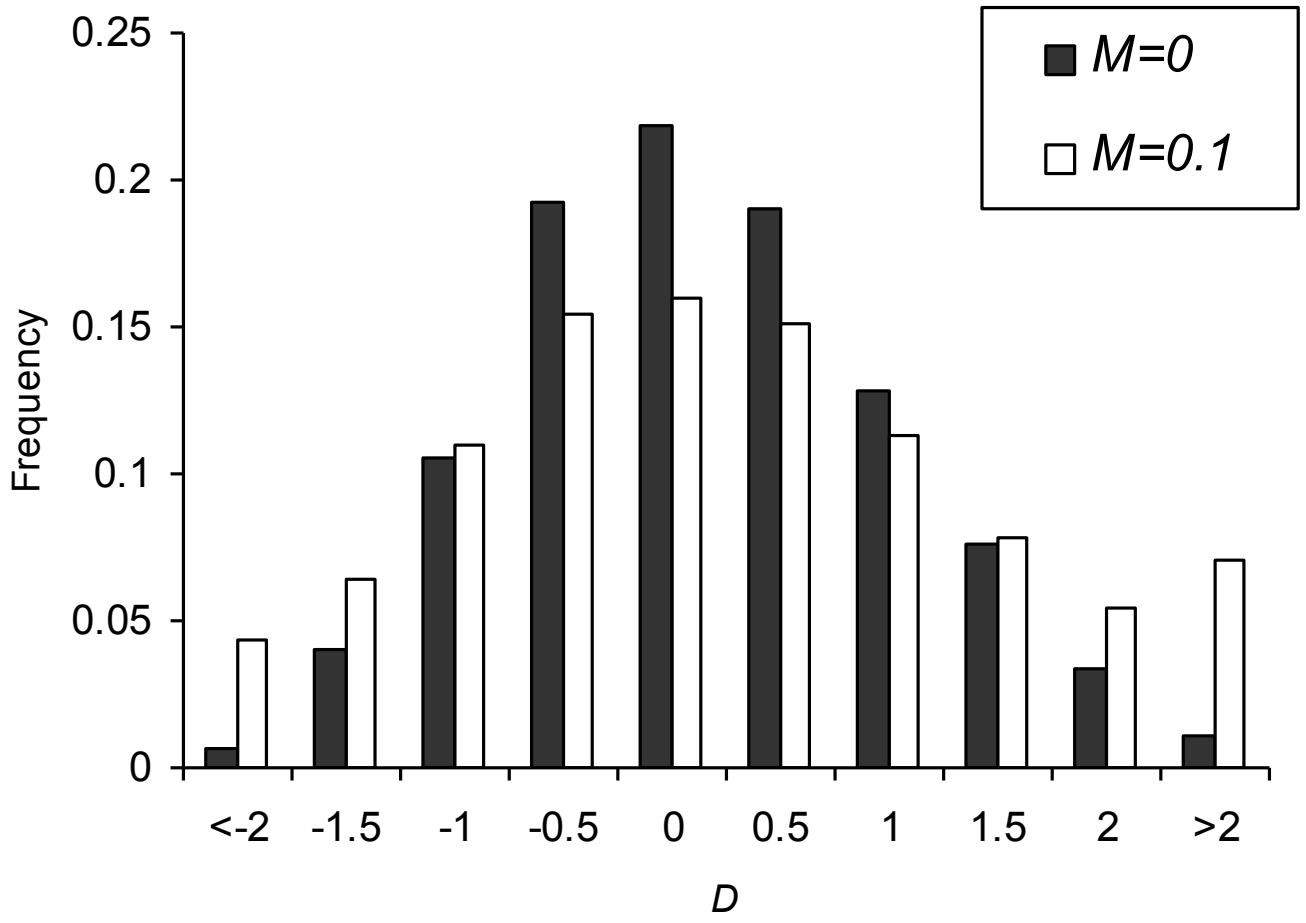
There are two common unbiased method of moments estimators of θ under the infinite sites model: Watterson's estimator, based on the number of segregating sites and Tajima's estimator, based on the average number of pairwise differences:

$$\hat{\theta}_W = S / \left(\sum_{i=1}^{n-1} 1/i \right) \quad \text{and,} \qquad \hat{\theta}_T = \sum_{i,j:i \neq j} k_{ij} \Big/ \binom{n}{2}$$

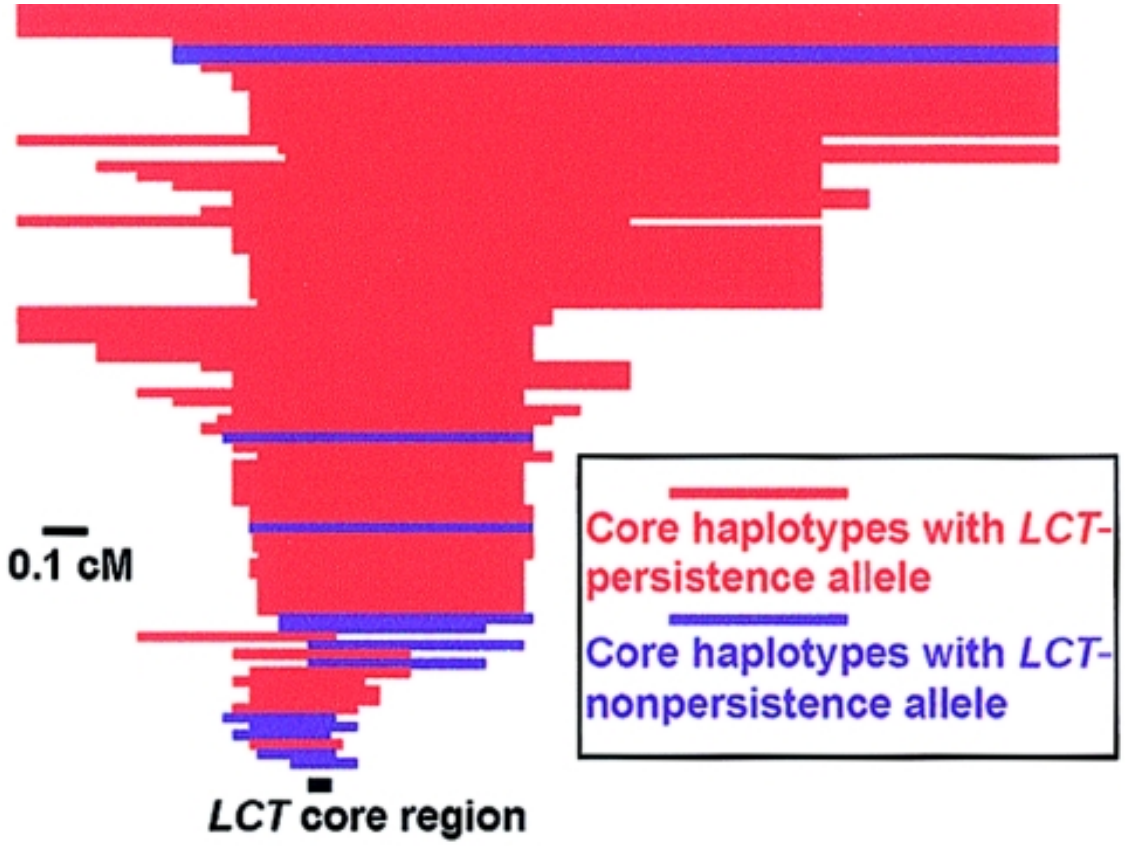
where S is the number of segregating (variable) sites and k_{ij} is the number of nucleotide differences between sequence i and j in the sample. Notice that is given by the average number of pairwise differences which in much of the literature is denoted by π. Tajima suggested using

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}(\hat{\theta}_T - \hat{\theta}_W)}}$$

as a test statistic when testing the neutral model.


0.25
0.2
0.15
0.1
0.05
0
Frequency
M=0
M=0.1
<-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
>2
D

Haplotype homozygosity



Selective Sweeps