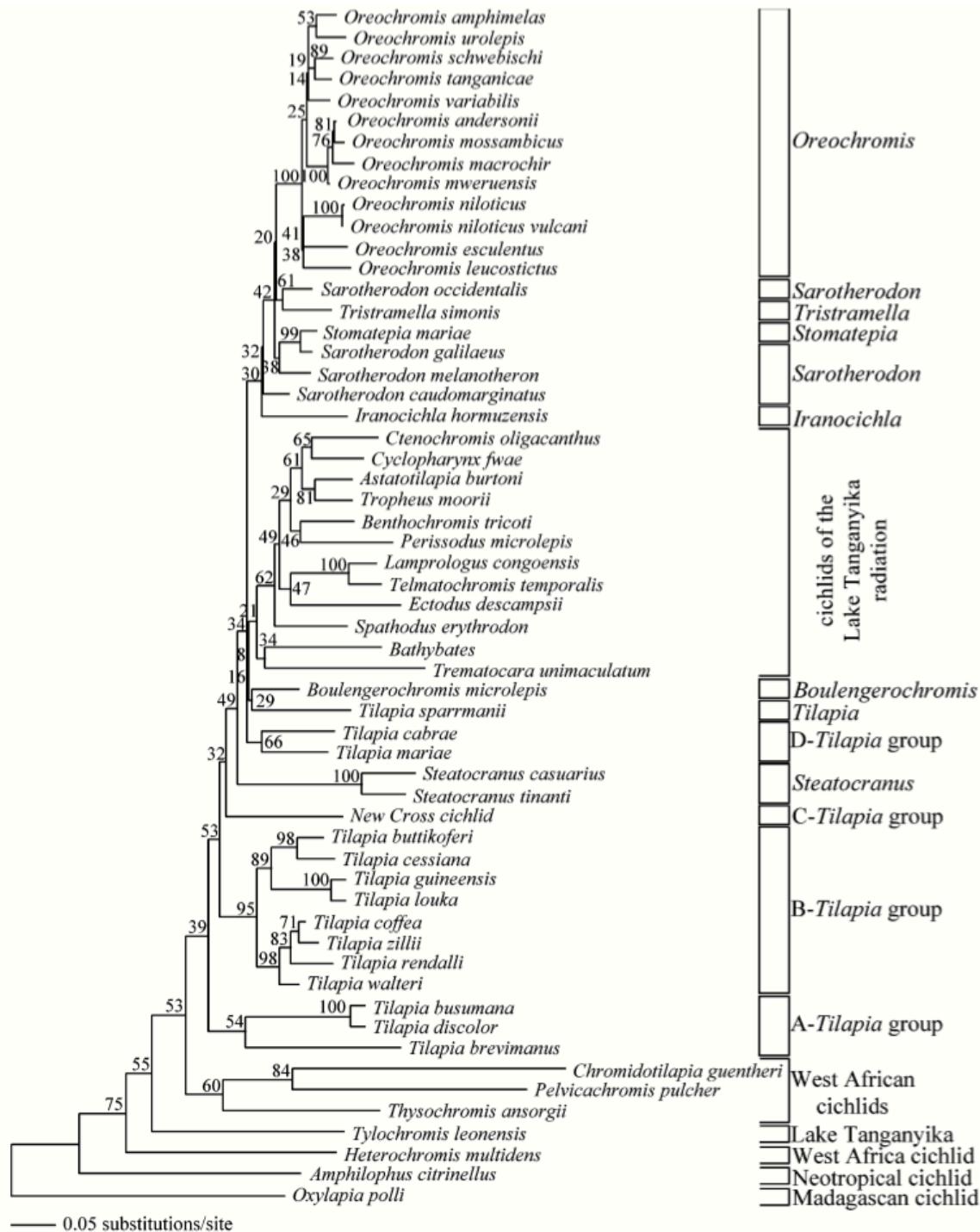


Molecular Signatures of Natural Selection

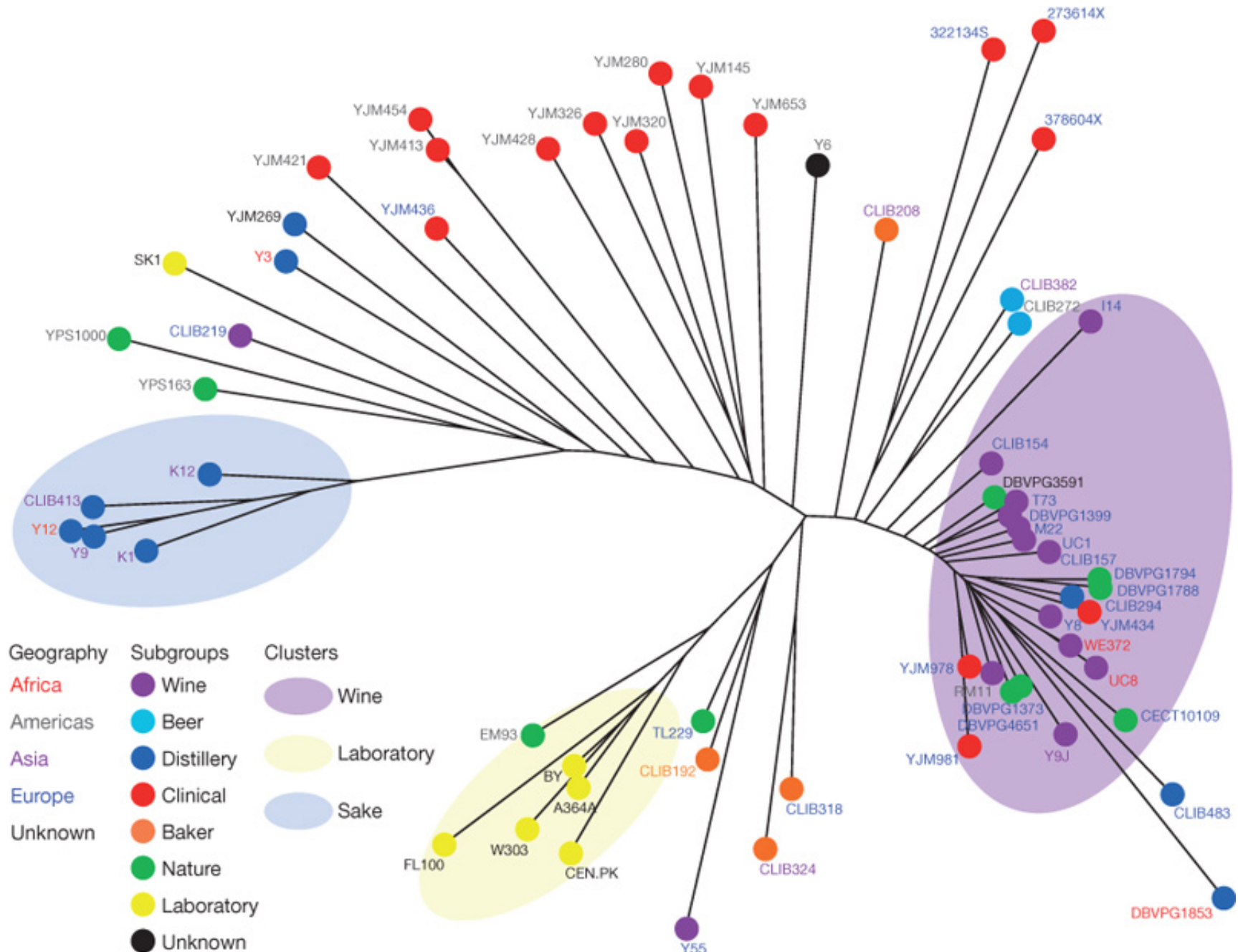
Genealogies

- Describe inheritance relationships among alleles
- Similar to phylogenetic tree

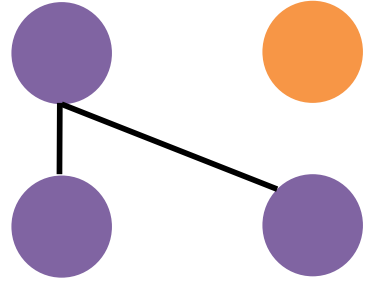
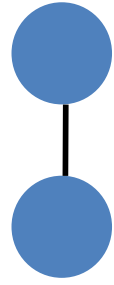
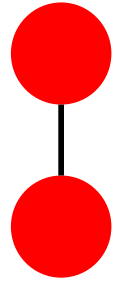


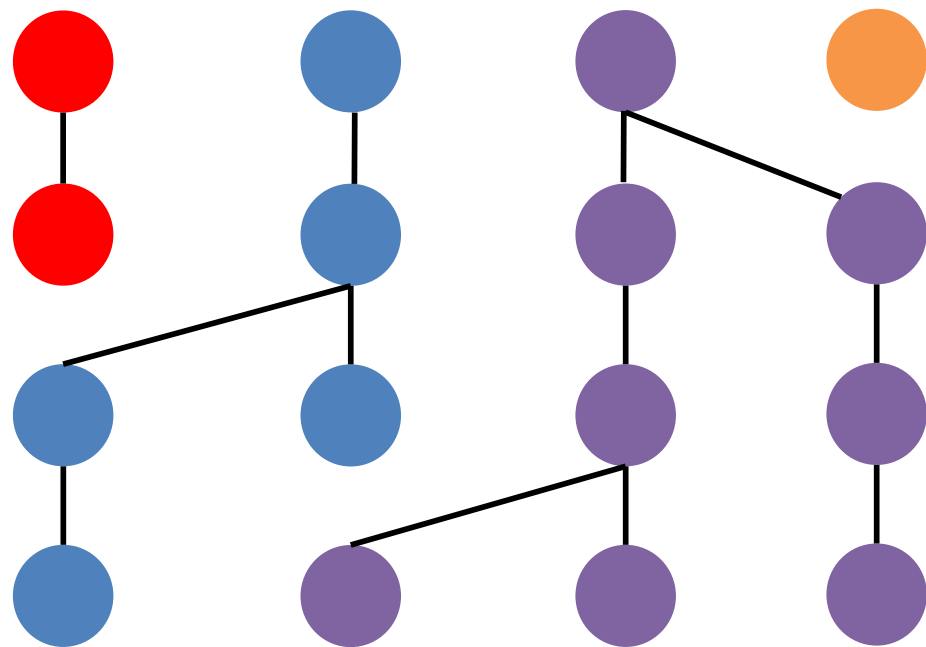
— 0.05 substitutions/site

Neighbour-joining tree of 63 *S. cerevisiae* strains.

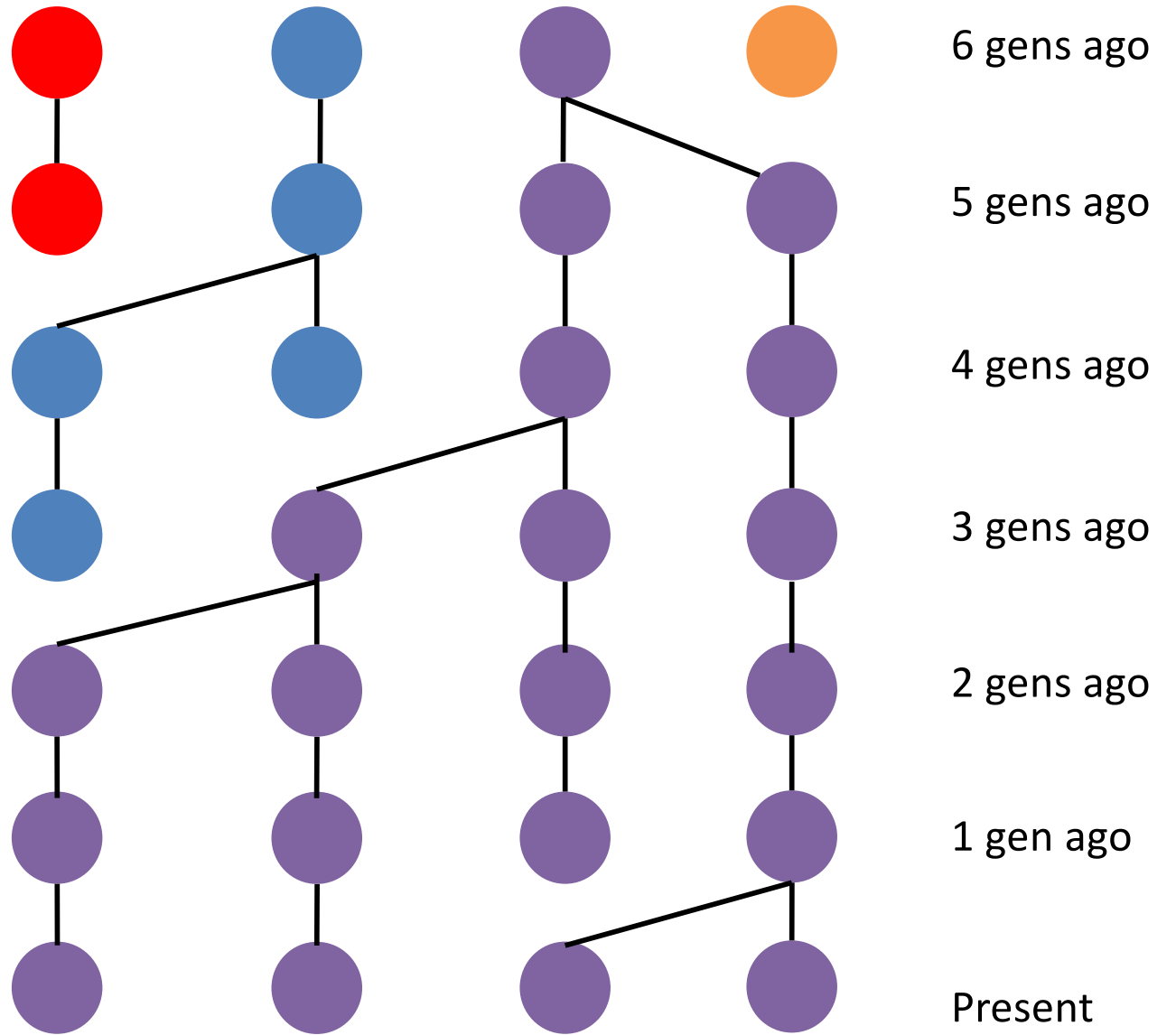


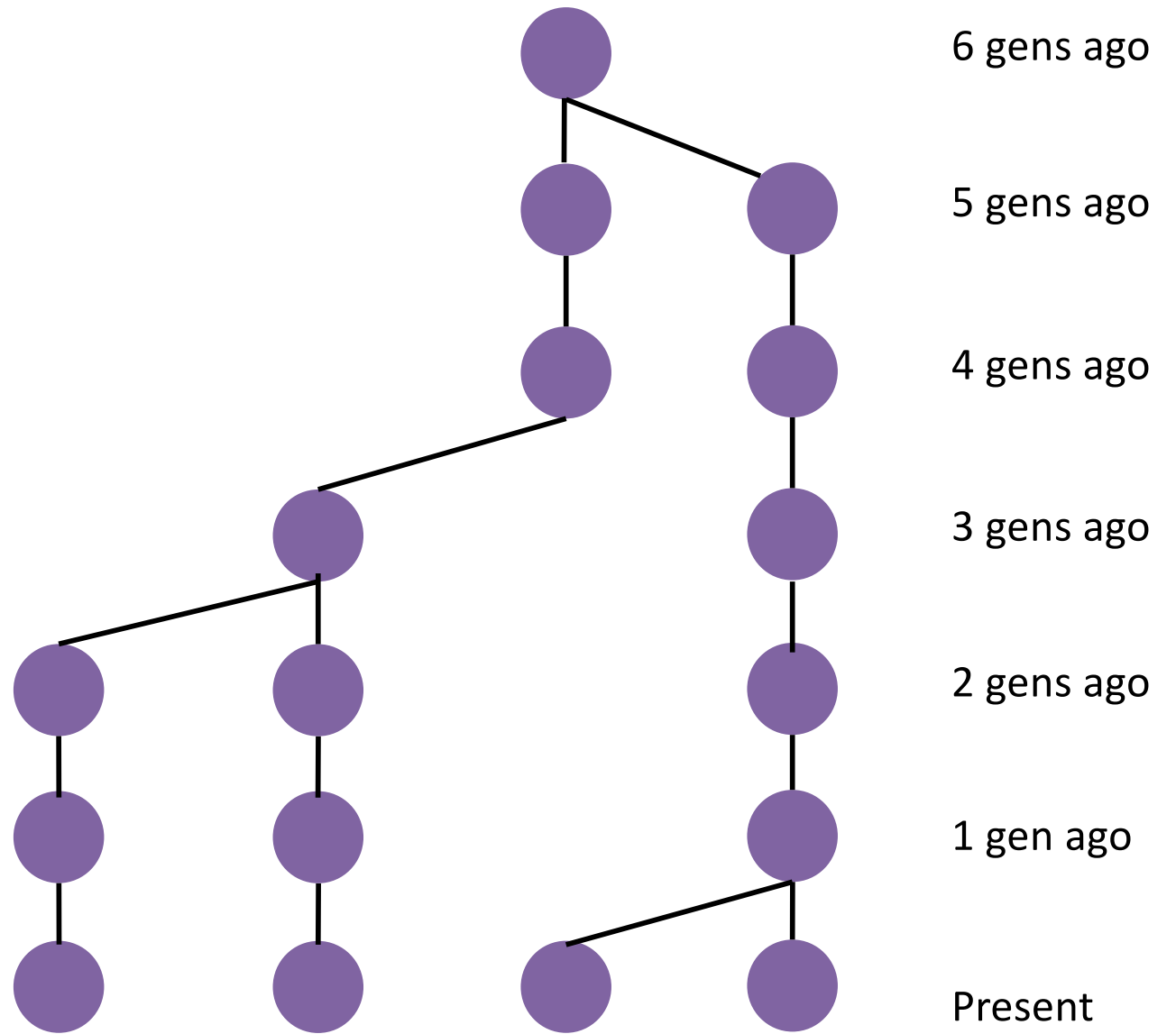


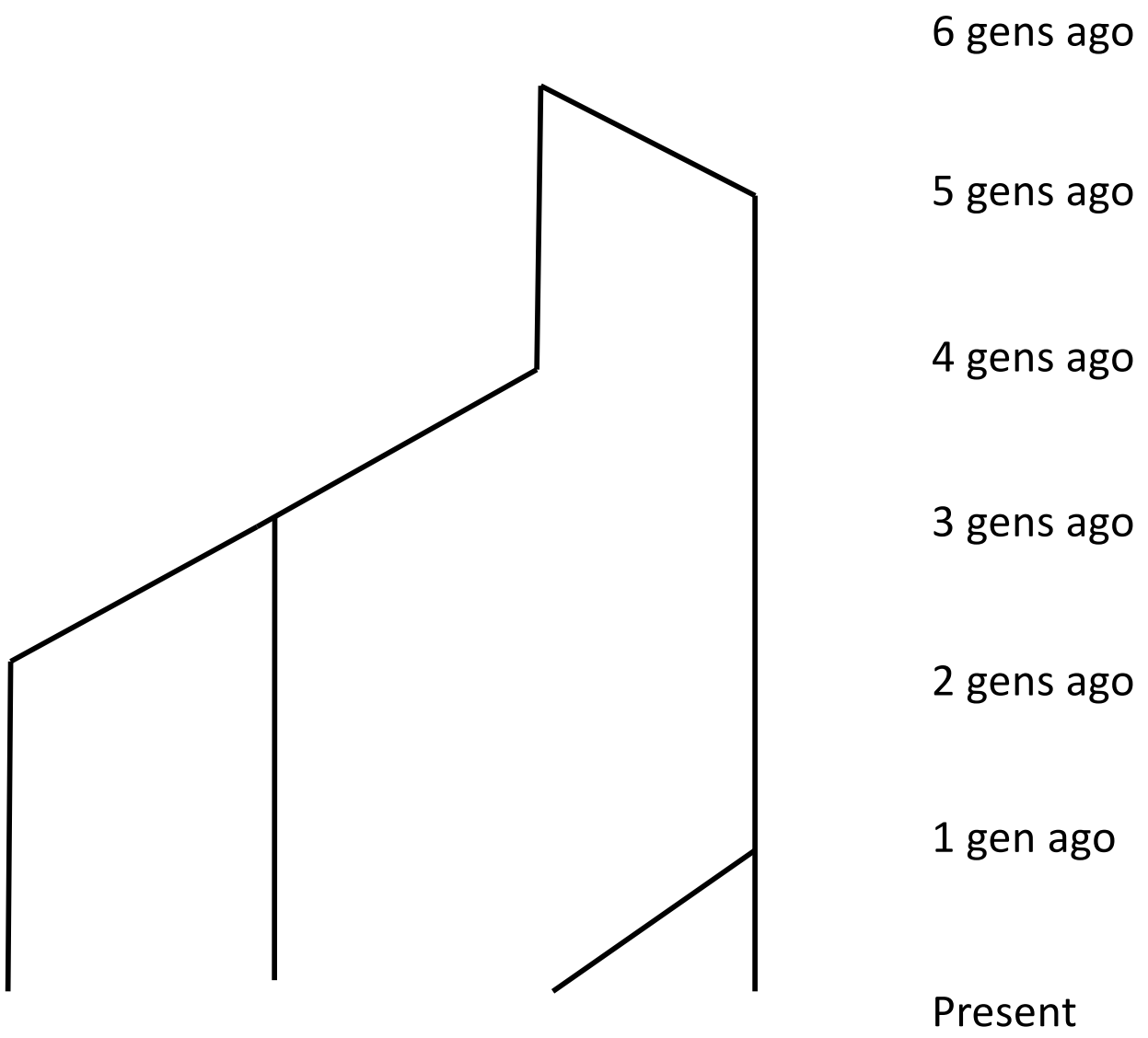


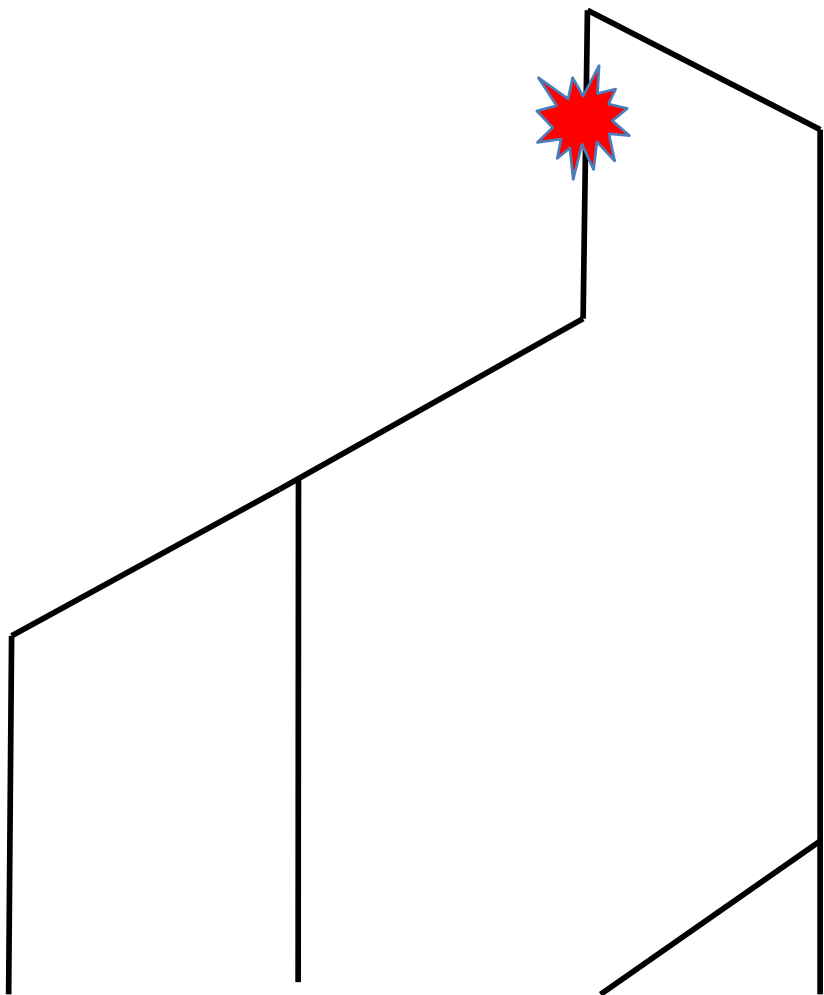


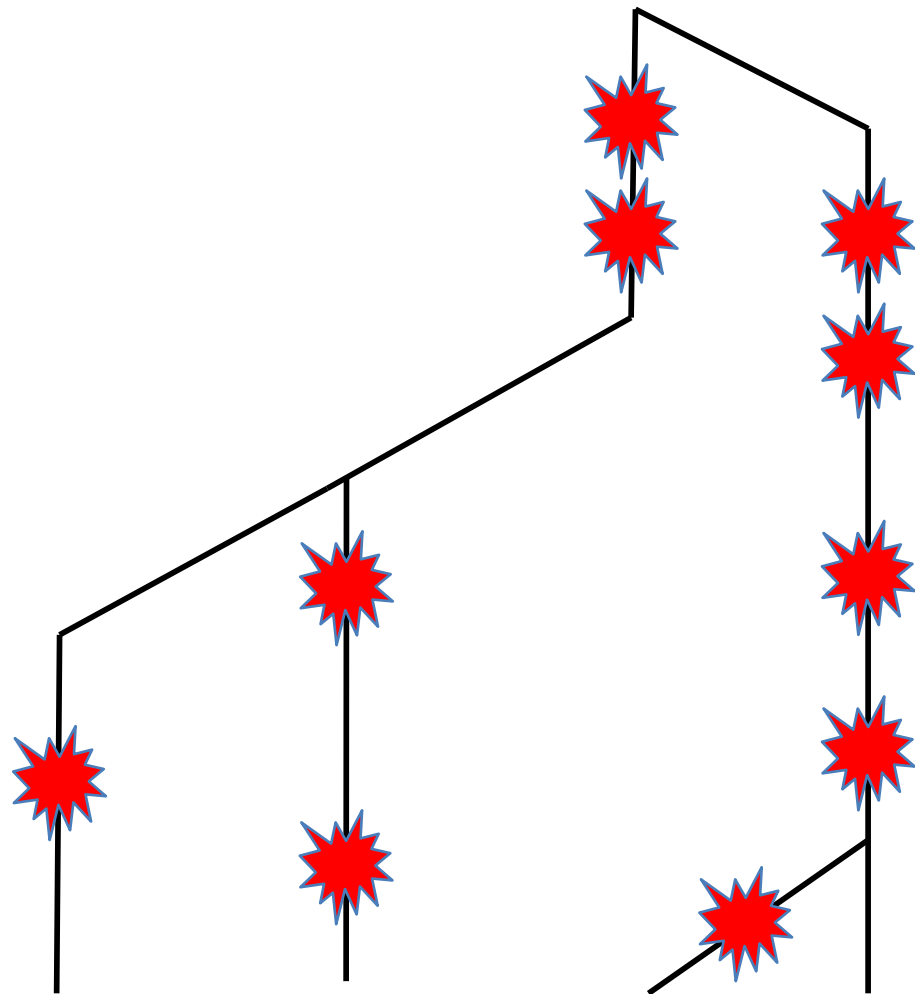
MRCA = 6 gen ago







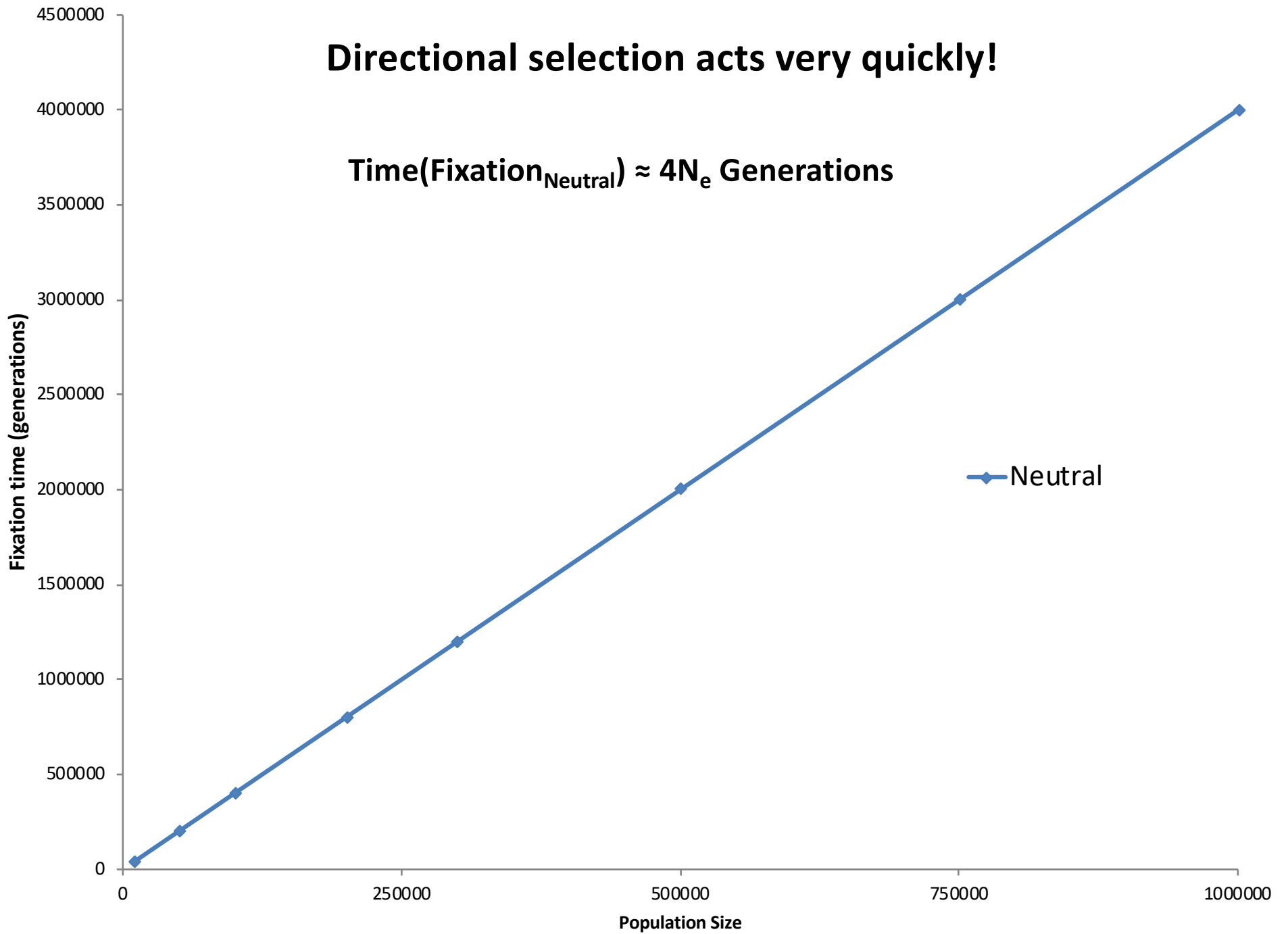




**How does this change under
directional selection?**

Directional selection acts very quickly!

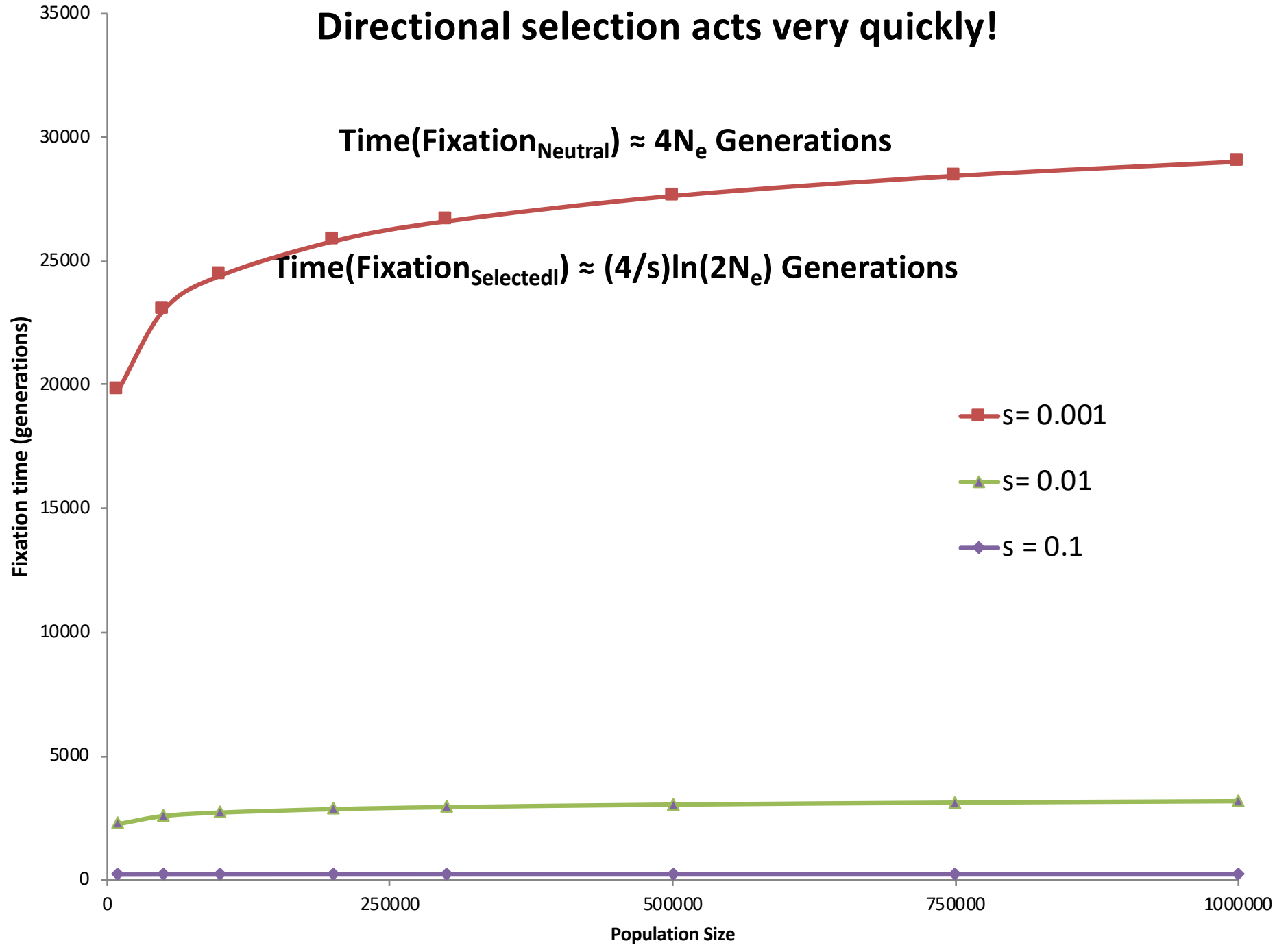
Time(Fixation_{Neutral}) $\approx 4N_e$ Generations



Directional selection acts very quickly!

$\text{Time}(\text{Fixation}_{\text{Neutral}}) \approx 4N_e \text{ Generations}$

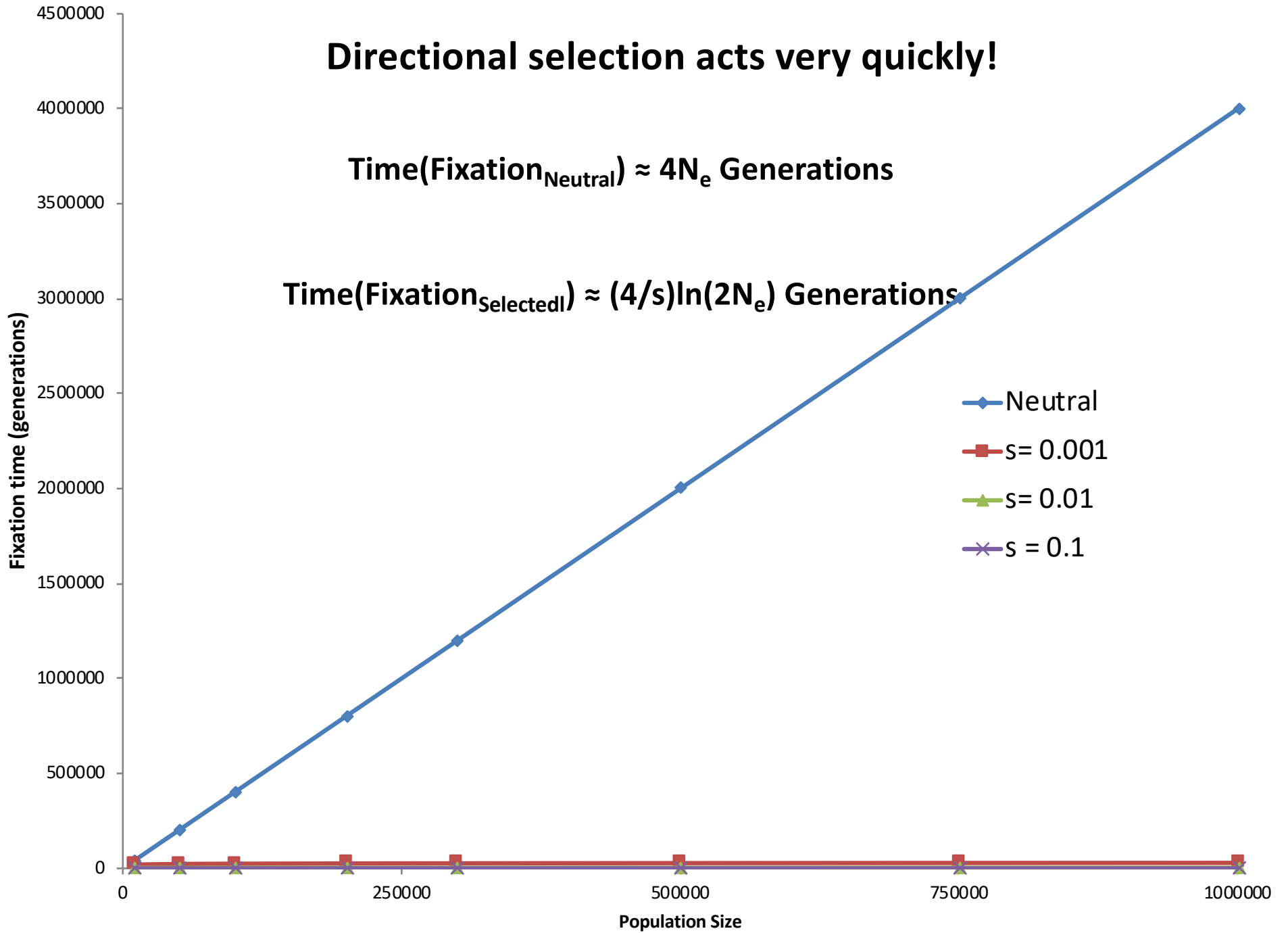
$\text{Time}(\text{Fixation}_{\text{Selected}}) \approx (4/s)\ln(2N_e) \text{ Generations}$



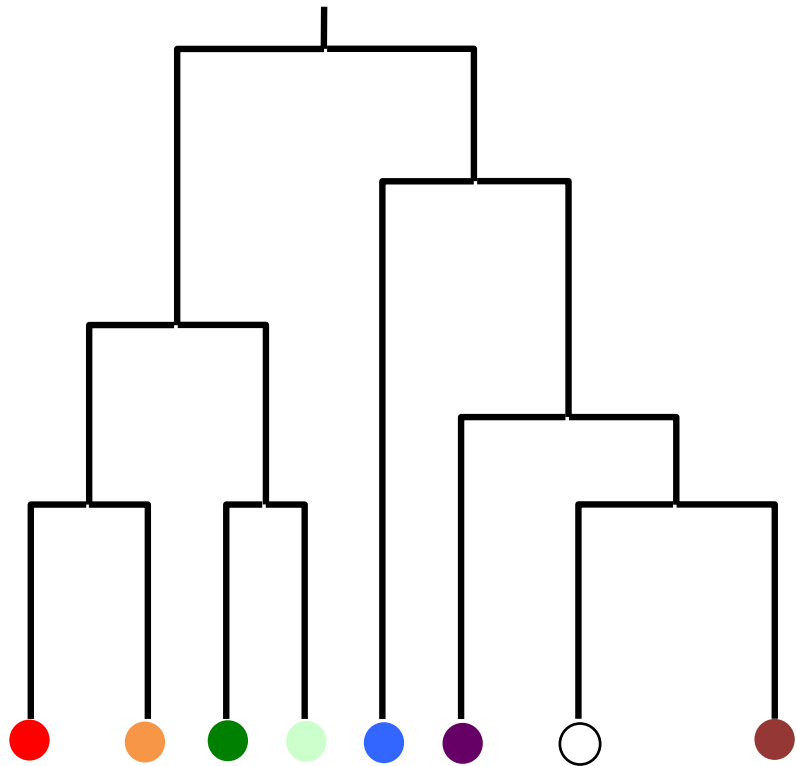
Directional selection acts very quickly!

$$\text{Time(Fixation}_{\text{Neutral}}) \approx 4N_e \text{ Generations}$$

$$\text{Time(Fixation}_{\text{Selected}}) \approx (4/s)\ln(2N_e) \text{ Generations}$$

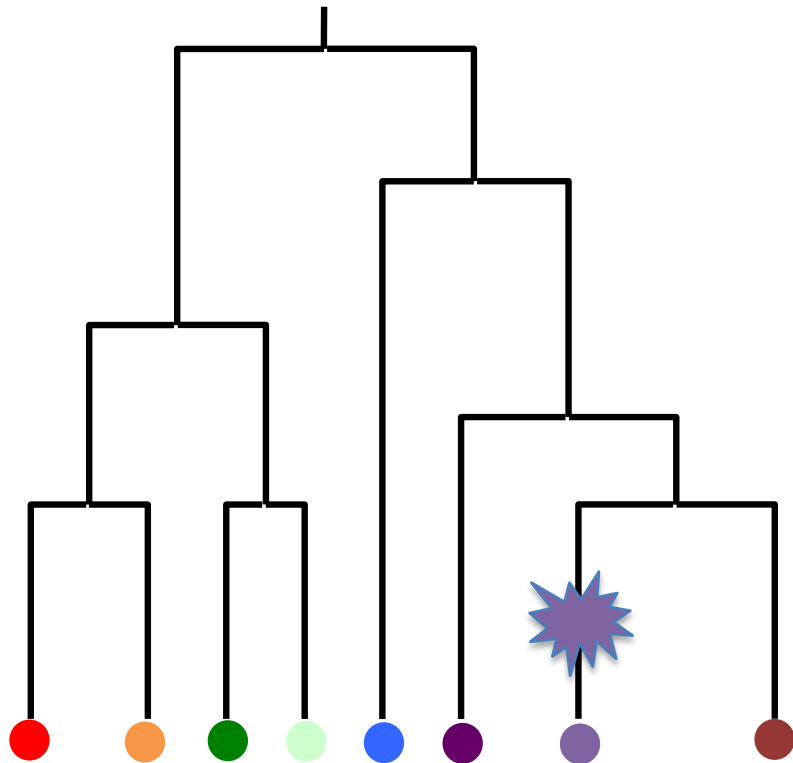


The genealogical structure of a selective sweep



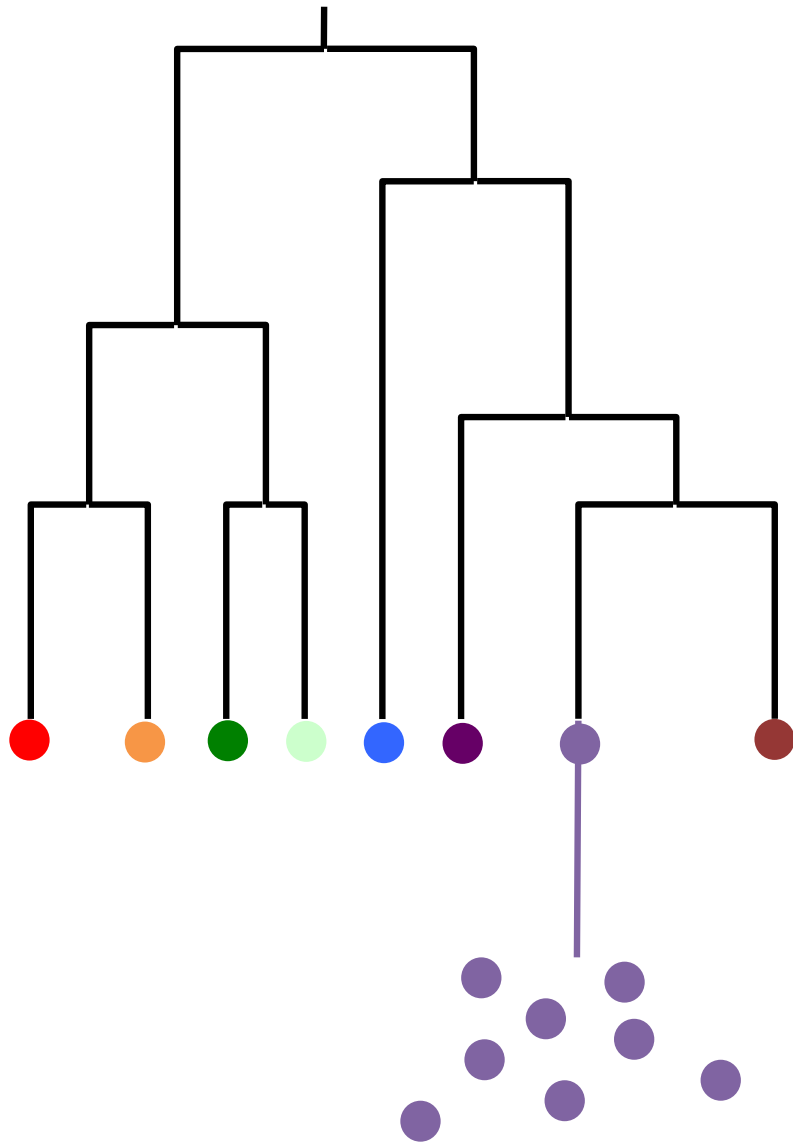
Population pre-selection

The genealogical structure of a selective sweep



Population pre-selection

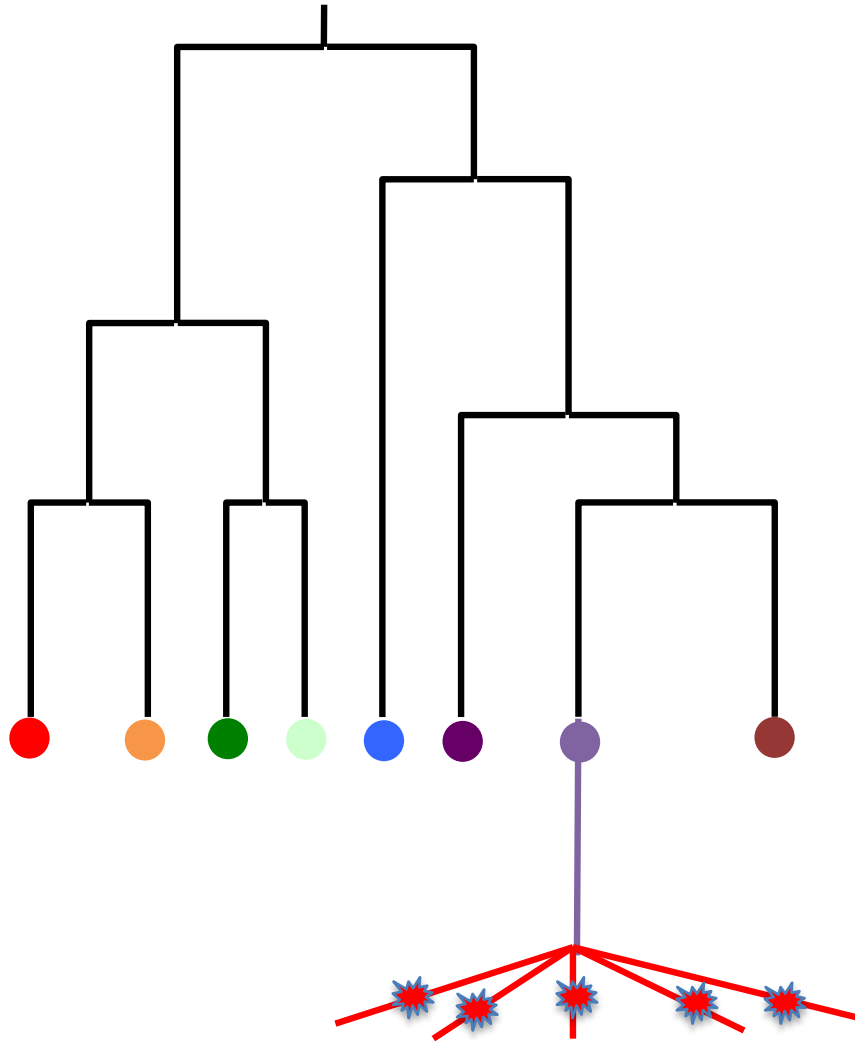
The genealogical structure of a selective sweep



Population pre-selection

During selection

The genealogical structure of a selective sweep



Population pre-selection

During selection

Mutational recovery post-selection

Effects of selection on sequences

CTGCCACCTTTTGTGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGC GTTCCATTGTCCGGATG
.....C.....T.....C.....
.....
.A.....C.....T.....C.....
.....G.....
.....
.....C.....C.....
.....T.....A.....
.....A.....
.....A.....
.....C.....C.....
.....C.....C.....C.....

This is our population prior to selection

Effects of selection on sequences

CTGCCACCTTTTGTGGTCTTAGTCCGCAGTGCACCTTGCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
.....C.....T.....C.....
.....
.A.....C.....T.....C.....
.....G.....
.....
.....C.....C.....
.....T.....A.....
.....A.....G.....
.....A.....
.....C.....C.....
.....C.....C.....



Effects of selection on sequences

CTGCCACCTTTTGTGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
.....C.....T.....C.....
.....
.A.....C.....T.....C.....
.....G.....
.....
.....A.....G.....
.....A.....G.....
.....A.....G.....
.....A.....G.....
.....C.....C.....
.....C.....C.....



CTGCCACCTTTTGTGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG

.....C.....T.....C.....
.....
.A.....C.....T.....C.....
.....G.....
.....
.....C.....C.....
.....T.....A.....
.....A.....G.....
.....A.....
.....C.....C.....
.....C.....C.....C.....

CTGCCACCTTTTGATTGGGTCTTAGTCCGCAGGGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCATTGTCCGGATG
.....A.....G.....C.....
.....A.....G.....
..A.....A.....G.....C.....
.....A.....G.....
.....A.....G.....
.....A.....G.....C.....
.....A.....G.....
.....A.....G.....
.....A.....G.....
.....A.....G.....C.....
.....A.....G.....C.....

LOW recombination

CTGCCACCTTTTGATTGGGTCTTAGTCCGCAGGGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
A.....G.....C.....
A.....G.....
 .A.....A.....G.....C.....
A.....G.....
A.....G.....
A.....G.....C.....
A.....G.....
A.....G.....
A.....G.....C.....
A.....G.....C.....

LOW recombination

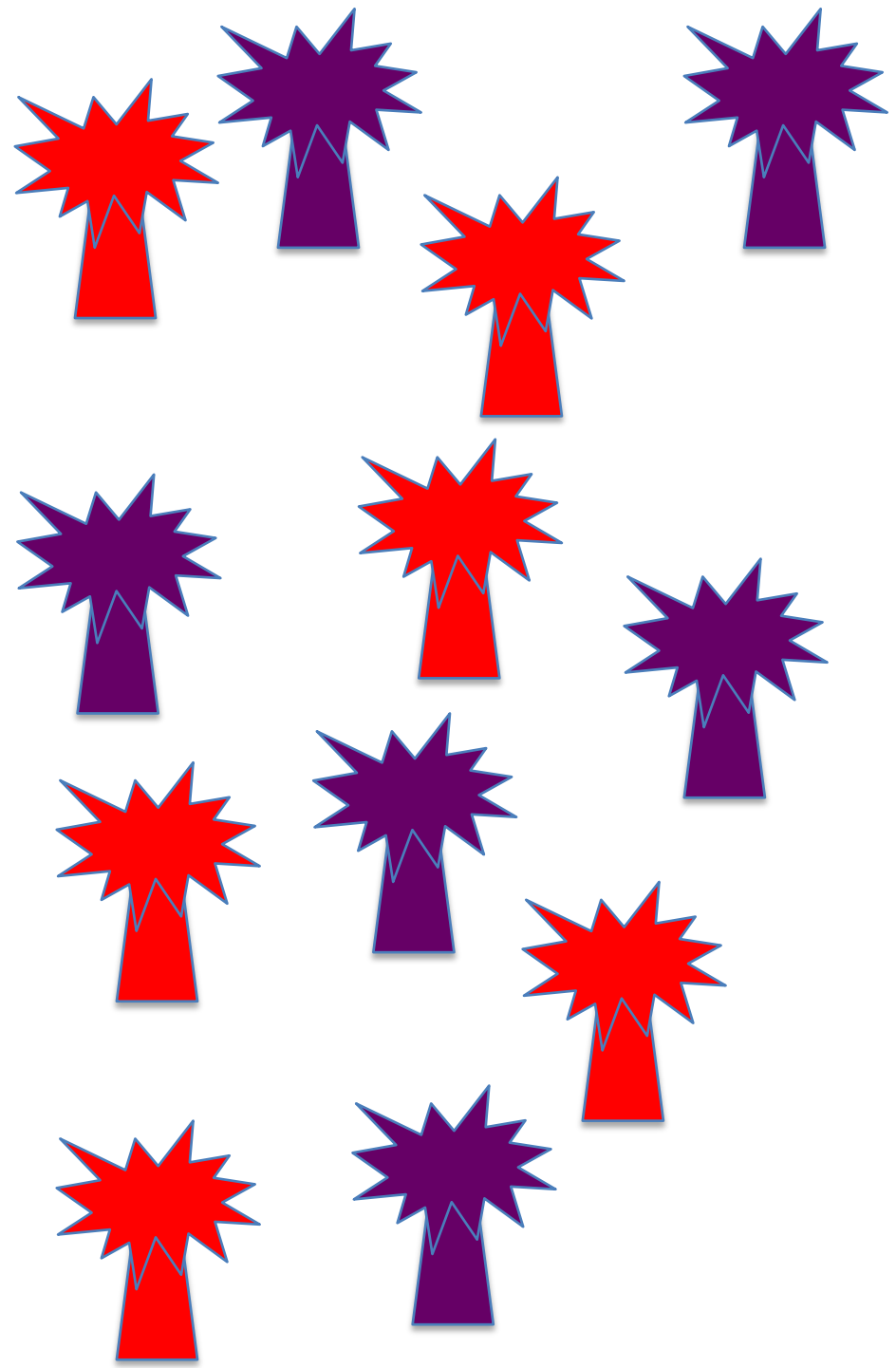
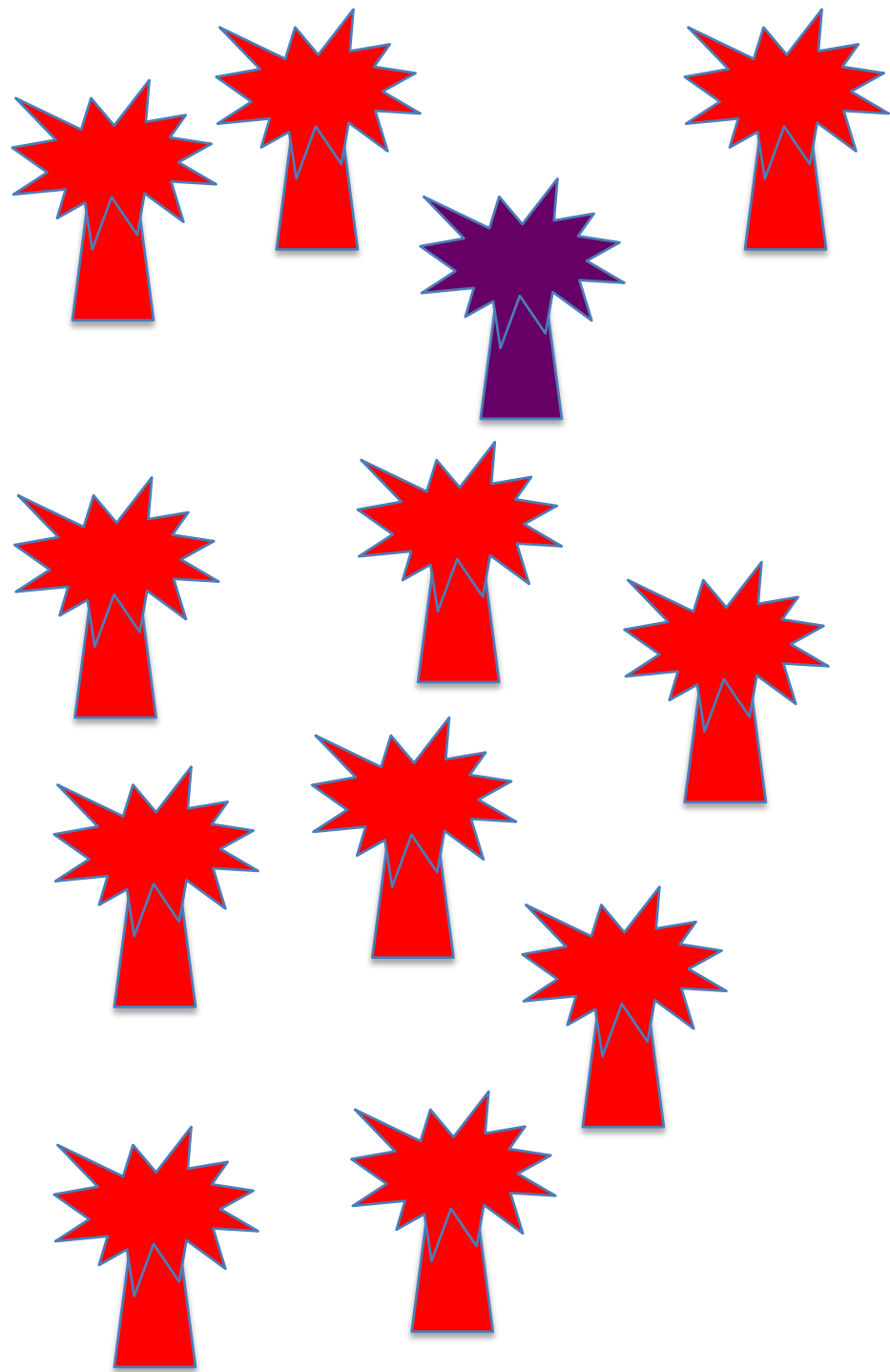
CTGCCACCTTTTGTTTGGGTCTTAGTCCGCAGGGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
C.....G.....T.....C.....
G.....
 .A.....C.....G.....T.....C.....
G.....
C.....G.....C.....
A.....G.....A.....
A.....G.....
C.....G.....C.....
C.....G.....C.....

HIGH recombination

Effects of selection on sequences

.....A.....G.....
.....A.....G.....
.....A.....T.....G.....
.....A.....G.....
C.....A.....G.....
.....A.....G.....G.....
.....A.....G.....A.....
.....A.....G.....A.....
.....A.....G.....
.....A.....G.....T.....
.....A.....G.....

How do we use these features to detect departures from neutrality using molecular data?



Summarizing molecular data

```
#1 CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
#2 .....C.....T.....C.....
#3 .....
#4 ..A.....C.....T.....C.....
#5 .....G.....A.....
#6 .....
#7 .....C.....C.....
#8 .....T.....A.....
#9 .....A.....
#10 .....A.....
#11 .....C.....C.....C.....
#12 .....C.....C.....C.....
```

Site frequency spectrum

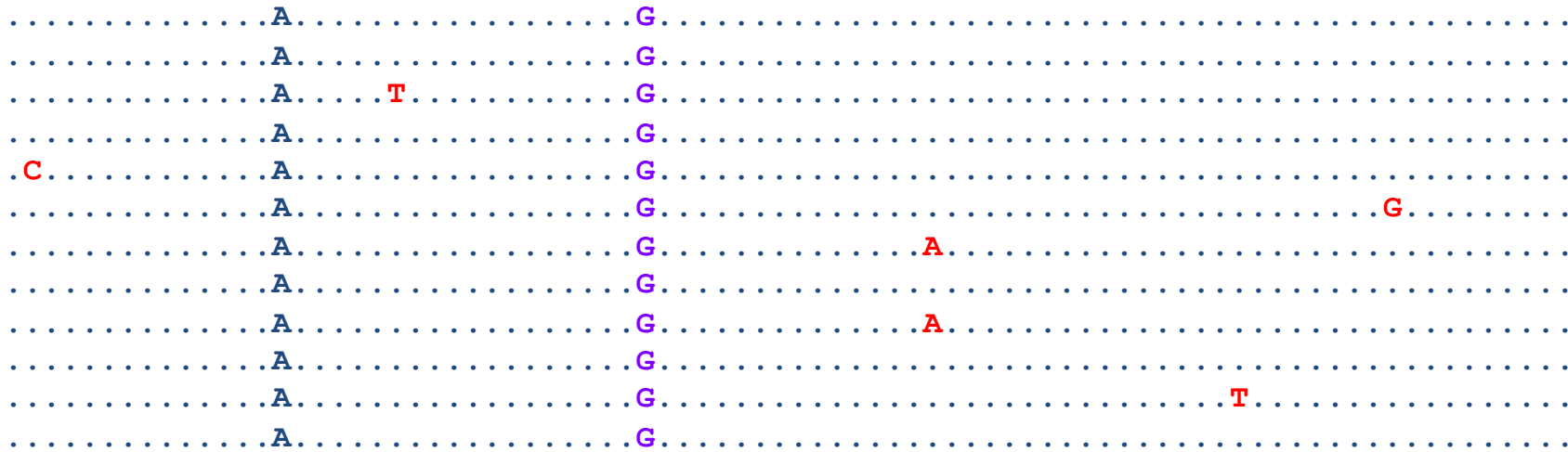
Ⓜ

Haplotype number

Haplotype diversity

...

Using the Site Frequency Spectrum



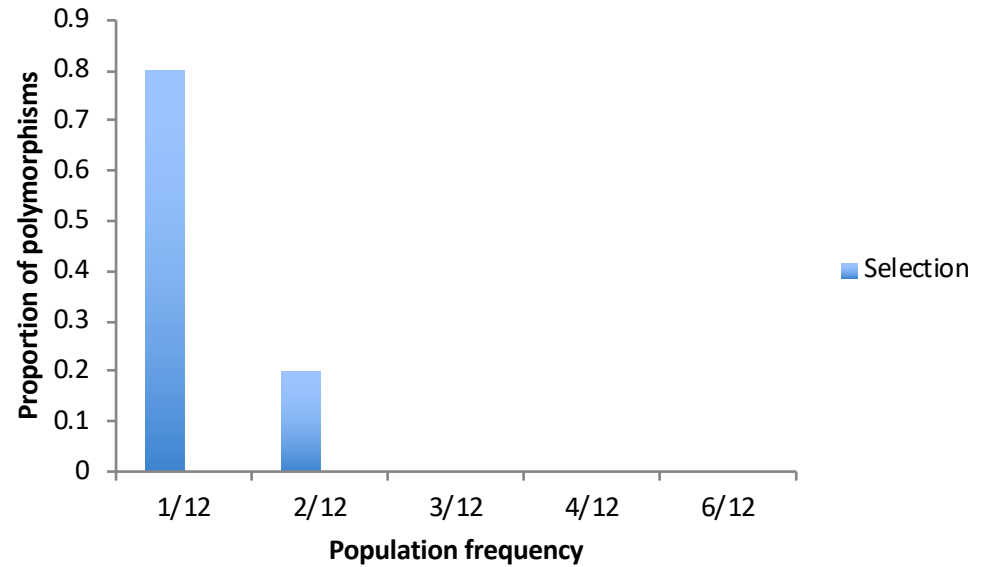
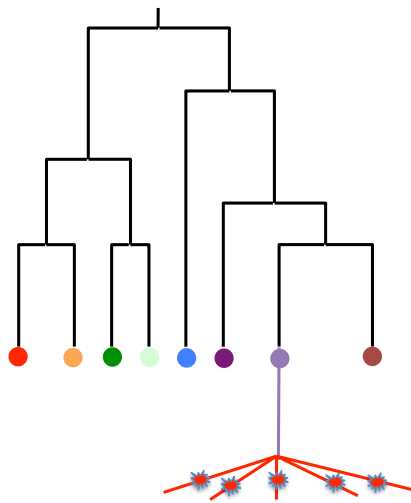
1/12

1/12

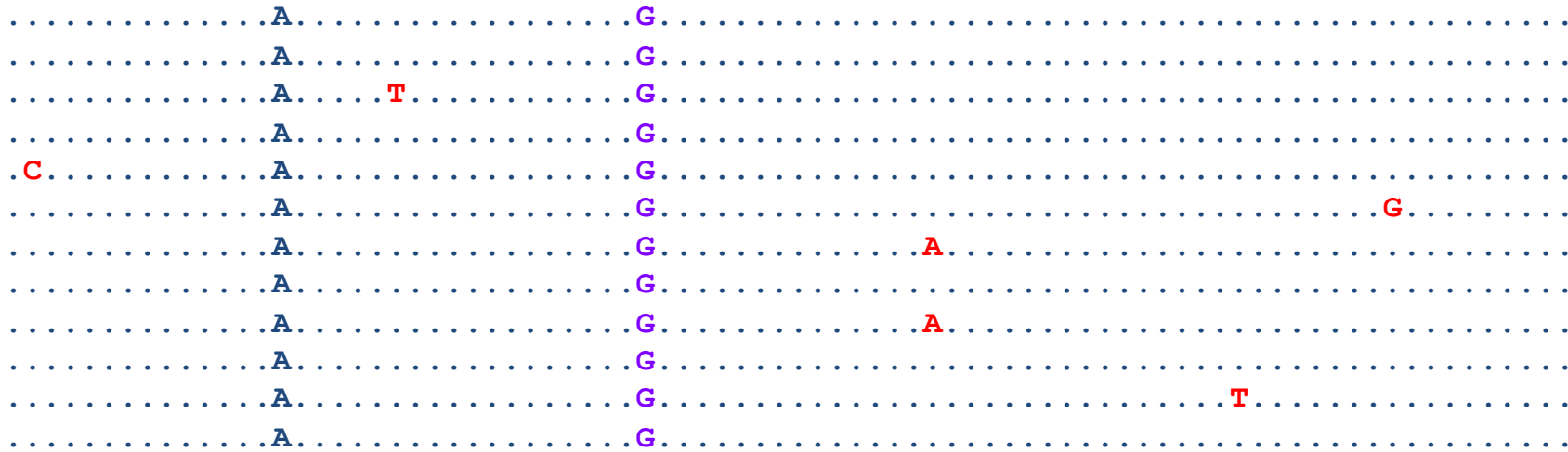
2/12

1/12

1/12



Using the Site Frequency Spectrum



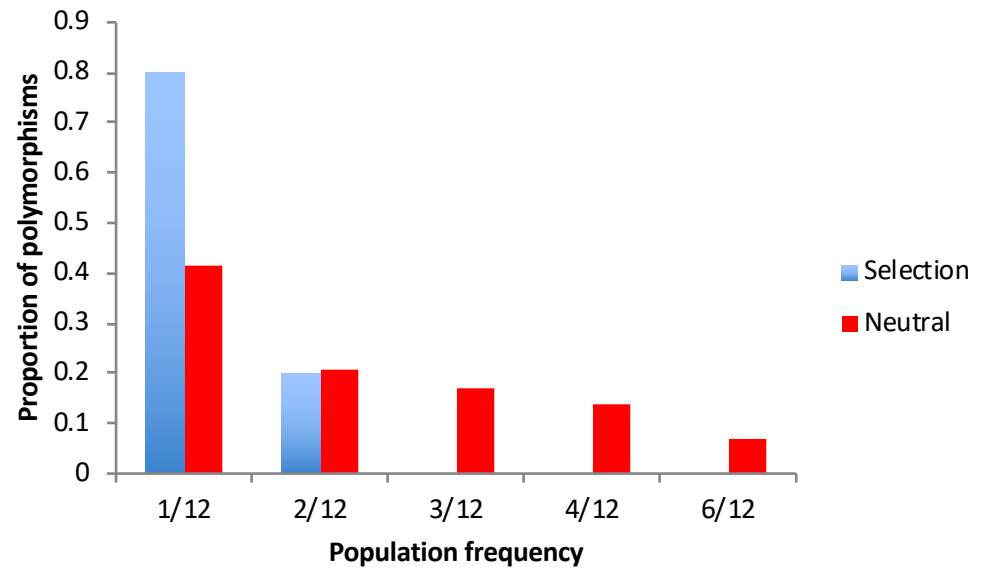
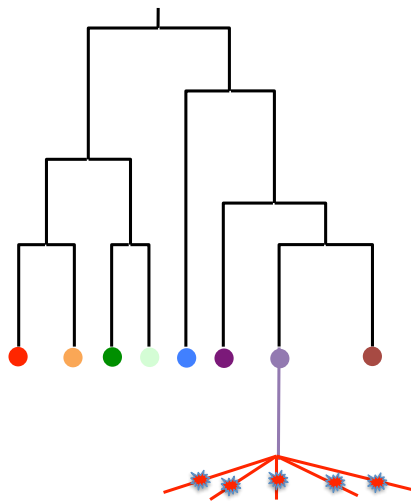
1/12

1/12

2/12

1/12

1/12



Using estimates of theta

$$E(\pi) = \theta \qquad E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

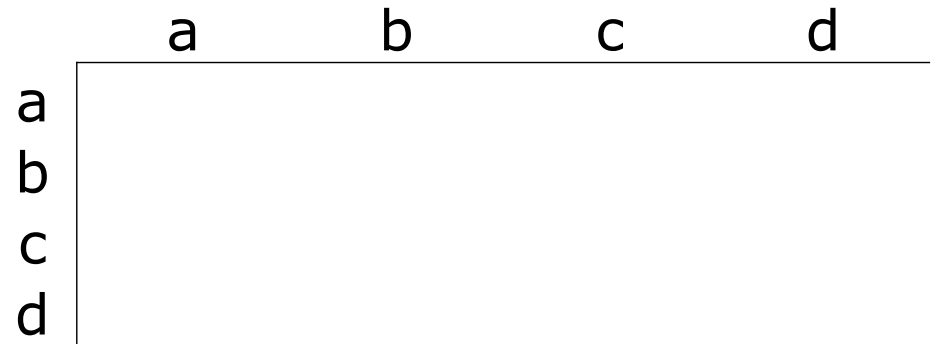
DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$



DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b		0		
c			0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b		0		
c			0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c			0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c			0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05		0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05		0	
d				0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05		0	
d	0.1			0

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

<i>i</i>	<i>j</i>
a	a
a	b
a	c
a	d
b	a
b	b
b	c
b	d
c	a
c	b
c	c
c	d
d	a
d	b
d	c
d	d

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

i	j	\hat{p}_i	\hat{p}_j	π_{ij}
a	a			
a	b			
a	c			
a	d			
b	a			
b	b			
b	c			
b	d			
c	a			
c	b			
c	c			
c	d			
d	a			
d	b			
d	c			
d	d			

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

i	j	\hat{p}_i	\hat{p}_j	π_{ij}
a	a	0.2	0.2	0
a	b			
a	c			
a	d			
b	a			
b	b			
b	c			
b	d			
c	a			
c	b			
c	c			
c	d			
d	a			
d	b			
d	c			
d	d			

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

i	j	\hat{p}_i	\hat{p}_j	π_{ij}
a	a	0.2	0.2	0
a	b	0.2	0.4	0.15
a	c	0.2	0.2	0.05
a	d	0.2	0.2	0.1
b	a	0.4	0.2	0.15
b	b	0.4	0.4	0
b	c	0.4	0.2	0.2
b	d	0.4	0.2	0.25
c	a	0.2	0.2	0.05
c	b	0.2	0.4	0.2
c	c	0.2	0.2	0
c	d	0.2	0.2	0.05
d	a	0.2	0.2	0.1
d	b	0.2	0.4	0.25
d	c	0.2	0.2	0.05
d	d	0.2	0.2	0

Calculating π

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

	a	b	c	d
a	0			
b	0.15	0		
c	0.05	0.2	0	
d	0.1	0.25	0.05	0

- $p_a = 1/5 = 0.2$
- $p_b = 2/5 = 0.4$
- $p_c = 1/5 = 0.2$
- $p_d = 1/5 = 0.2$

i	j	\hat{p}_i	\hat{p}_j	π_{ij}	$\hat{p}_i \hat{p}_j \pi_{ij}$
a	a	0.2	0.2	0	0
a	b	0.2	0.4	0.15	0.012
a	c	0.2	0.2	0.05	0.002
a	d	0.2	0.2	0.1	0.004
b	a	0.4	0.2	0.15	0.012
b	b	0.4	0.4	0	0
b	c	0.4	0.2	0.2	0.016
b	d	0.4	0.2	0.25	0.02
c	a	0.2	0.2	0.05	0.002
c	b	0.2	0.4	0.2	0.016
c	c	0.2	0.2	0	0
c	d	0.2	0.2	0.05	0.004
d	a	0.2	0.2	0.1	0.004
d	b	0.2	0.4	0.25	0.02
d	c	0.2	0.2	0.05	0.002
d	d	0.2	0.2	0	0

Sum = 0.112

$$\pi = \frac{5}{4} (0.112) = 0.14$$

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

DNA Variation

Individual Allele

1	a	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
2	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.
3	c	.	.	.	G	.	A	G	.	.	.	A	.	.	.	C	.
4	d	.	.	.	G	.	A	T	.	.	.	A	.	.	.	C	.
5	b	.	.	.	A	.	T	G	.	.	.	C	.	.	.	T	.

- π = Average pairwise difference among alleles

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

- Θ_w = Watterson's estimator $\Theta_w = S / \sum_{i=1}^{N-1} \frac{1}{i}$

	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
	.	.	.	G	.	T	T	.	.	.	C	.	.	.	T	.
Population 1

	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
	.	.	.	G	.	T	T	.	.	.	C	.	.	.	T	.
Population 2	.	.	.	G	.	T	T	.	.	.	C	.	.	.	T	.
	.	.	.	G	.	T	T	.	.	.	C	.	.	.	T	.
	.	.	.	G	.	T	T	.	.	.	C	.	.	.	T	.

	C	A	T	A	G	A	A	C	C	T	G	G	G	C	A	C	T	T	C	A
	.	.	.	G
Population 3	T	T
	C
	T

$$\pi = \frac{N}{N-1} \sum_{ij} \hat{p}_i \hat{p}_j \pi_{ij}$$

$$\Theta_w = S / \sum_{i=1}^{N-1} \frac{1}{i}$$

Using estimates of theta

$$E(\pi) = \theta \qquad E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima's D

Tajima (1989)

$$D = (\pi - \theta_W) / \text{stdev}(\pi - \theta_W)$$

Under neutrality, $D = 0$

Tajima's D

Tajima (1989)

$$D = (\pi - \theta_W) / \text{stdev}(\pi - \theta_W)$$

Under neutrality, $D = 0$

Under directional selection, $D < 0$

Tajima's D

Tajima (1989)

$$D = (\pi - \theta_W) / \text{stdev}(\pi - \theta_W)$$

Under neutrality, $D = 0$

Under directional selection, $D < 0$

Under balancing selection, $D > 0$

Summarizing molecular data

```
#1 CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
#2 .....C.....T.....C.....
#3 .....
#4 ..A.....C.....T.....C.....
#5 .....G.....A.....
#6 .....
#7 .....C.....C.....
#8 .....T.....A.....
#9 .....A.....
#10 .....A.....
#11 .....C.....C.....C.....
#12 .....C.....C.....C.....
```

Site frequency spectrum

Ⓜ

Haplotype number

Haplotype diversity

...

Polymorphism-Based Tests of Neutrality

Site Frequency Spectrum

- Tajima's D
- Fu and Li's F (1993)
- Fay and Wu's H (2000)

...

Haplotype tests (Depaulis and Veuille 1998;
Andolfatto et al. 1999; Hudson et al. 1994...)

Modeling selection (Kim and Stephan 2002...)

...

SFS-based tests sensitive to other factors

Demography

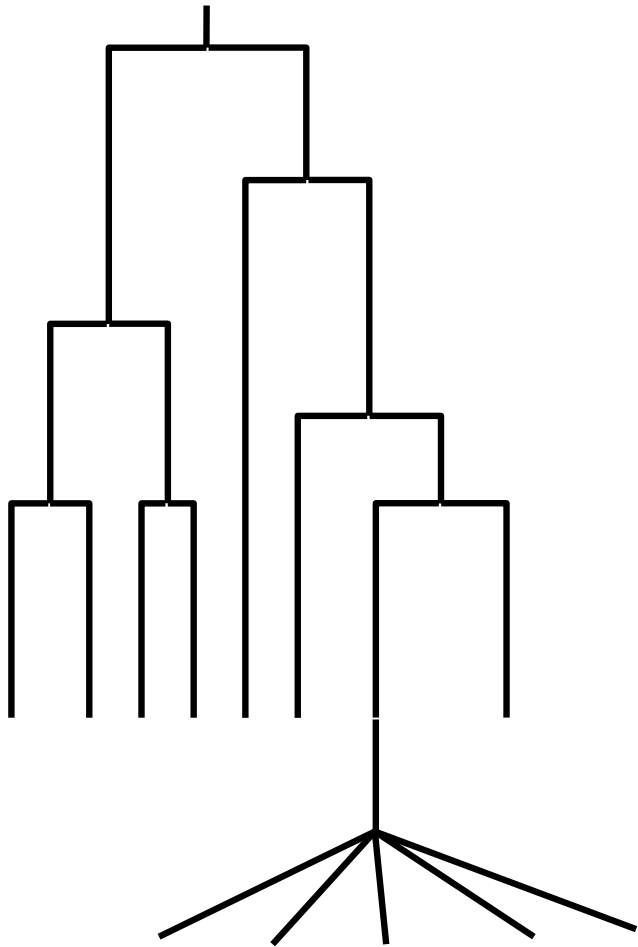
Population bottlenecks

Population expansions

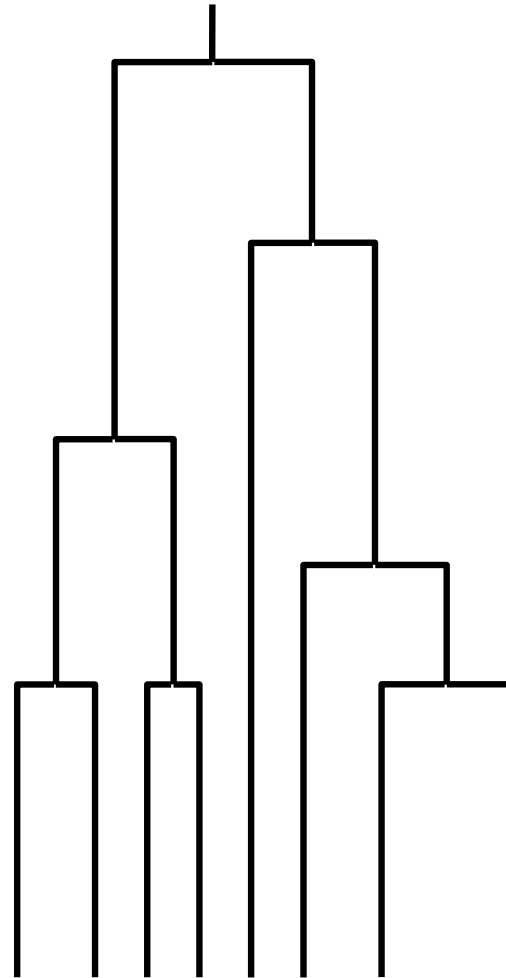
Population substructure

Migration

Comparing observed genealogy to neutral expectation

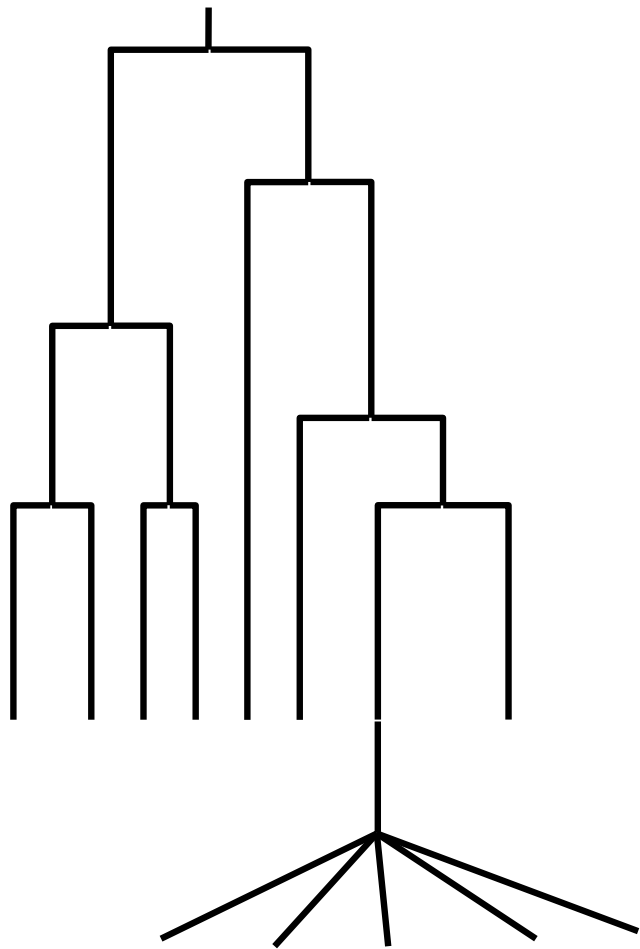


selective sweep

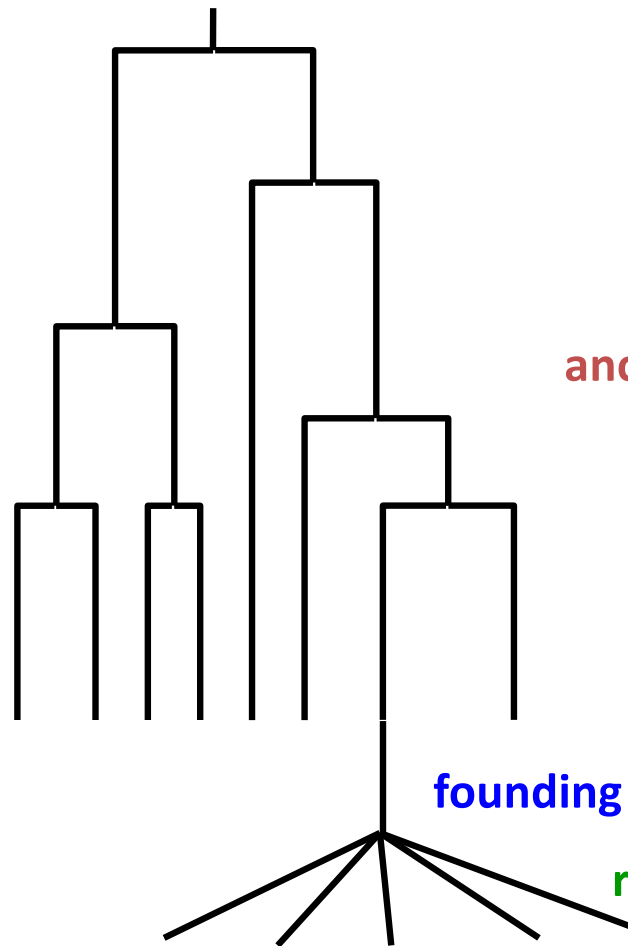


Neutrality

Genealogy may be impacted by forces other than selection



selective sweep

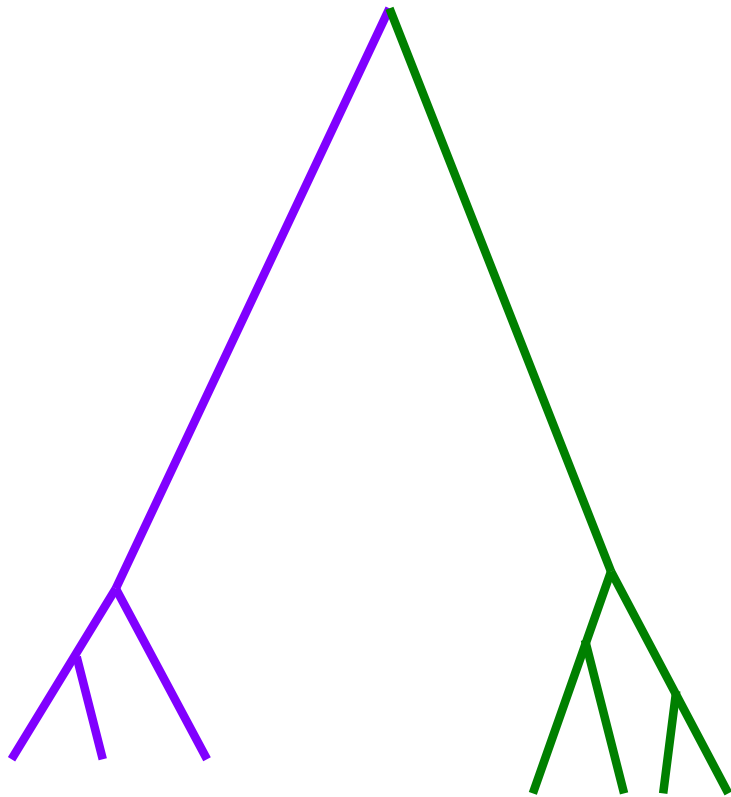


ancestral population

founding event

radiation / expansion

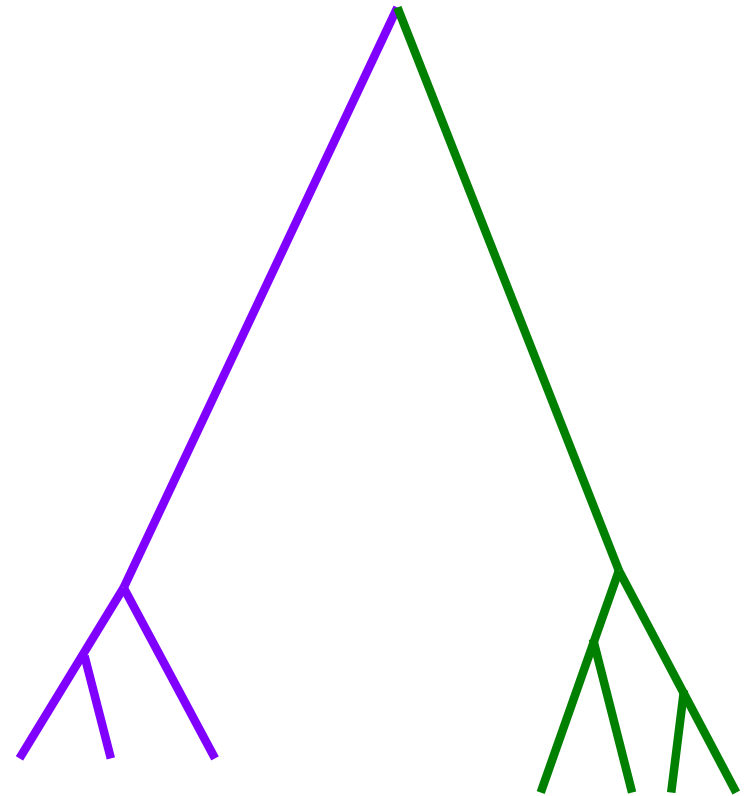
Genealogy may be impacted by forces other than selection



allele 1

allele 2

ancient balanced polymorphism



subpop 1

subpop 2

population structure

What if we have more data?

```
Indiv1 CTGCCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
Indiv2 .....C.....C.....
Indiv3 .....
Indiv4 ..A.....C.....C.....
Indiv5 .....G.....A.....
Indiv6 .....
Indiv7 .....C.....C.....
Indiv8 .....T.....A.....
Indiv9 .....A.....
Indiv10 .....A.....
Indiv11 .....C.....C.....C.....
Indiv12 .....C.....C.....C.....
```


What if we have more data?

```
Indiv1 CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
Indiv2 .....C.....C.....
Indiv3 .....
Indiv4 ..A.....C.....C.....
Indiv5 .....G.....A.....
Indiv6 .....
Indiv7 .....C.....C.....
Indiv8 .....T.....A.....
Indiv9 .....A.....
Indiv10 .....A.....
Indiv11 .....C.....C.....C.....
Indiv12 .....C.....C.....C.....
Species2 .....G.....C.....A.....C.....
```

McDonald-Kreitman Test (1991)

- Polymorphism vs divergence
- Divide mutations
 - Neutral Changes
 - Potentially non-neutral changes
- Non-neutral/neutral = Non-neutral/neutral

What if we have more data?

```
Indiv1 CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
Indiv2 .....C.....C.....
Indiv3 .....
Indiv4 ..A.....C.....C.....
Indiv5 .....G.....A.....
Indiv6 .....
Indiv7 .....C.....C.....
Indiv8 .....T.....A.....
Indiv9 .....A.....
Indiv10 .....A.....
Indiv11 .....C.....C.....C.....
Indiv12 .....C.....C.....C.....
Species2 .....G.....C.....A.....C.....
```

What if we have more data?

N P N N P N N N N N N P

Indiv1	CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
Indiv2C.....C.....
Indiv3
Indiv4	.A.....C.....C.....
Indiv5G.....A.....
Indiv6
Indiv7C.....C.....
Indiv8T.....A.....
Indiv9A.....
Indiv10A.....
Indiv11C.....C.....C.....
Indiv12C.....C.....C.....
Species2G.....C.....A.....C.....



Neutral



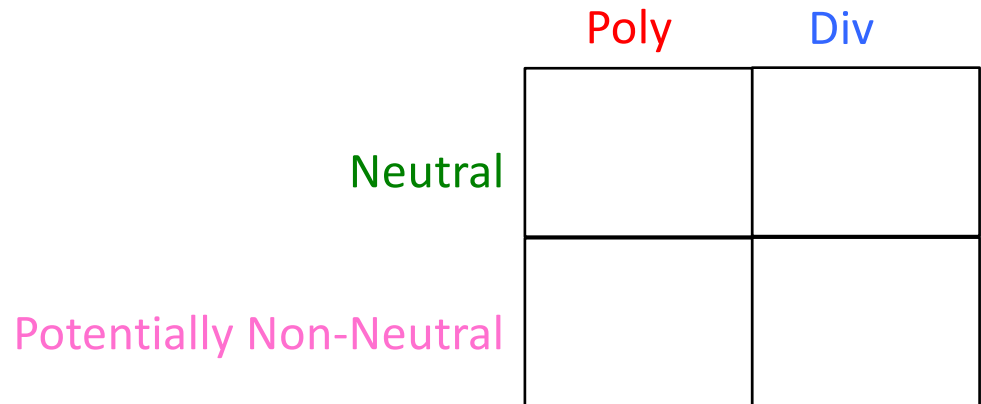
Potentially Non-neutral

What if we have more data?

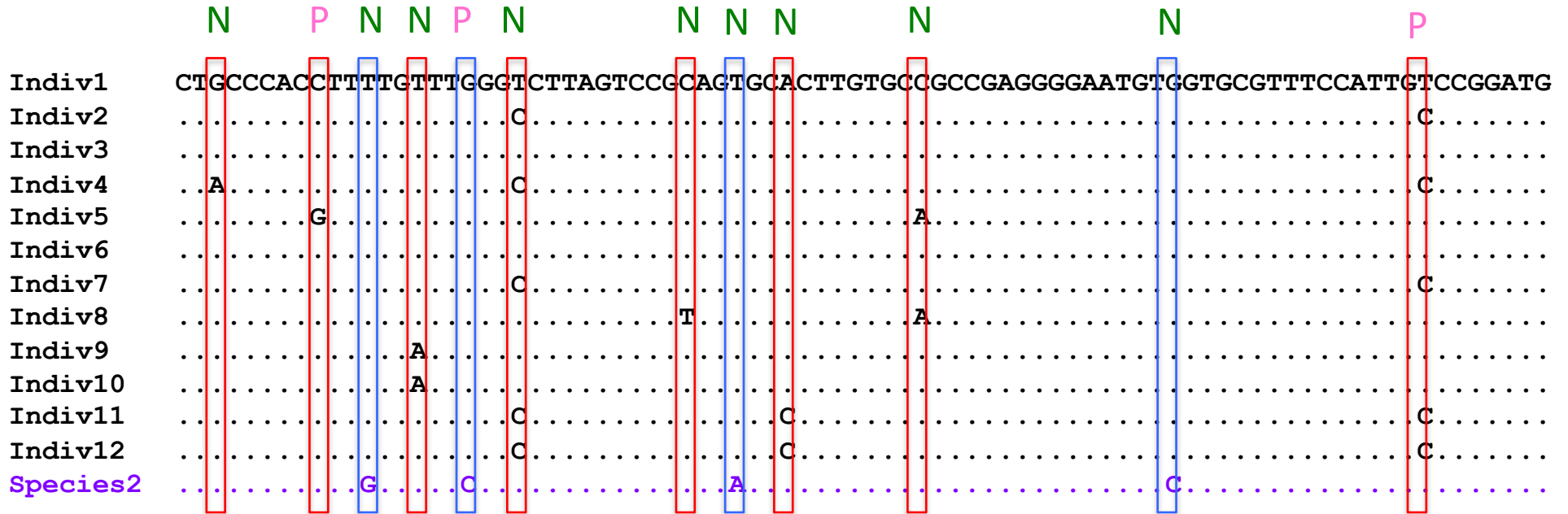
N P N N P N N N N N N P

Individ1 CTGCCACCTTTTGTGGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
 Individ2C.....C.....
 Individ3
 Individ4 ..A.....C.....C.....
 Individ5G.....A.....
 Individ6
 Individ7C.....C.....
 Individ8T.....A.....
 Individ9A.....
 Individ10A.....
 Individ11C.....C.....C.....
 Individ12C.....C.....C.....
 Species2G.....C.....A.....C.....

- Polymorphic
- Divergent
- Neutral
- Potentially Non-neutral



What if we have more data?



- █ Polymorphic
- █ Divergent
- █ Neutral
- █ Potentially Non-neutral

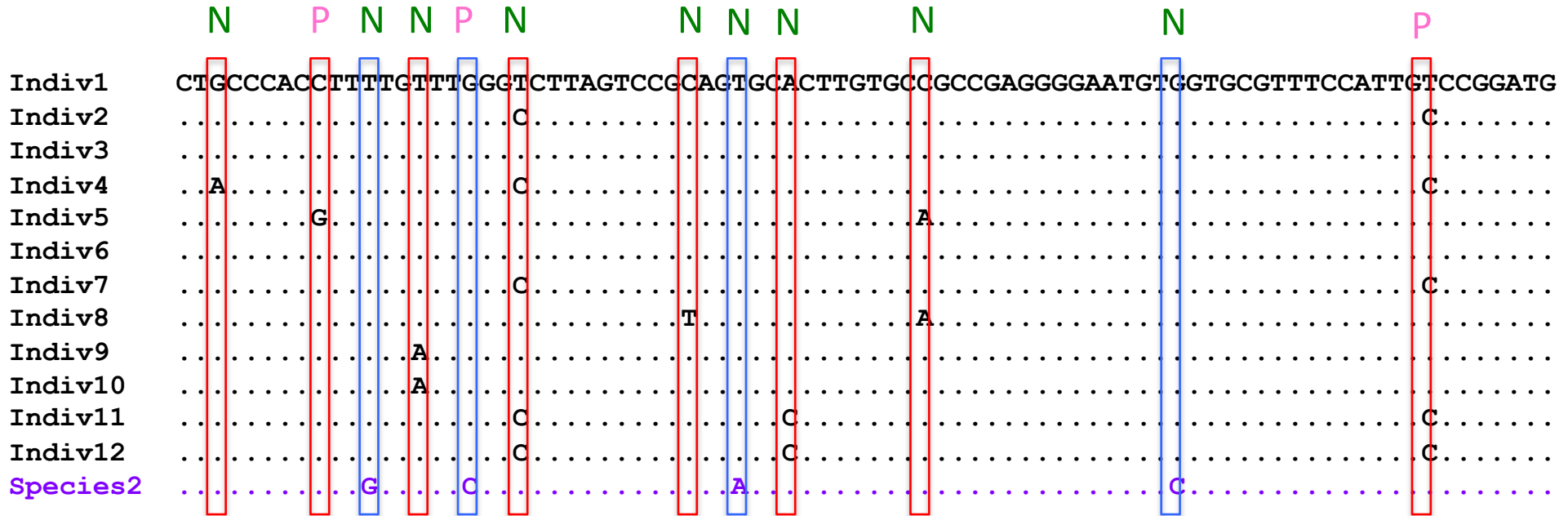
	Poly	Div
Neutral	6	3
Potentially Non-Neutral	2	1

Ratio Neutral/Potentially Non-Neutral

3

3

What if we have more data?



- █ Polymorphic
- █ Divergent
- █ Neutral
- █ Potentially Non-neutral

Neutral

Potentially Non-Neutral

	Poly	Div
Neutral	6	3
Potentially Non-Neutral	2	1

Test for statistical significance using FET

McDonald-Kreitman Test (1991)

Originally

-Neutral = synonymous sites

-Potentially non-neutral = nonsynonymous sites

McDonald-Kreitman Test (1991)

	Poly	Div
Synonymous		
Nonsynonymous		

McDonald-Kreitman Test (1991)

	Poly	Div
Synonymous	28	16
Nonsynonymous	21	12
Ratio	1.3	1.3

Acp23D4b from *Drosophila yakuba* (n = 10) and *D. teissieri* (n = 1)

McDonald-Kreitman Test (1991)

	Poly	Div
Synonymous	28	16
Nonsynonymous	21	66

P < 0.001, FET

Acp23D4b from *Drosophila yakuba* (n = 10) and *D. teissieri* (n = 1)

McDonald-Kreitman Test (1991)

	Poly	Div
Synonymous	28	16
Nonsynonymous	21	66

Excess fixed replacements
 $P < 0.001$, FET

Acp23D4b from *Drosophila yakuba* (n = 10) and *D. teissieri* (n = 1)

Can we reject...

Too much synonymous polymorphism?

	Poly	Div
Synonymous	28	16
Nonsynonymous	21	66

Too little synonymous divergence?

Excess fixed replacements

$P < 0.001$, FET

Too little nonsynonymous polymorphism?

Acp23D4b from *Drosophila yakuba* (n = 10) and *D. teissieri* (n = 1)

McDonald-Kreitman Test (1991)

- Originally
 - Neutral = synonymous sites
 - Potentially non-neutral = nonsynonymous sites
- Now
 - Neutral = synonymous sites
 - Potentially non-neutral = intronic sites...
- Limitations
 - Neutral sites may not be neutral
 - Will only detect recurrent selection

What if we only had divergence data?

Species1 CTGCCACCTTTTGTGGTCTTAGTCCGCAGTGCACCTTGTGCCGCCGAGGGGAATGTGGTGCCTTCCATTGTCCGGATG
Species2C.....C.....
Species3A.....C.....G.....A
Species4 ..A.....C.....C.....
Species5G.....A.....

Define: K_A = rate of substitution at nonsynonymous sites

Define: K_S = rate of substitution at synonymous sites

$K_A = K_S$ Neutral expectation

Neutral rate of evolution = $k = \mu$

$K_A < K_S$ Purifying selection (most common)

Nonsynonymous mutations are deleterious

$K_A > K_S$ Frequent directional selection

Nonsynonymous mutations are fixed adaptively

How to calculate K_A , K_S ?

species 1 **CGG ACA CTG**

species 2 **AGG ATA CTC**

Nei-Gojobori Method (1983)

1. Count up the number of **synonymous** and **nonsynonymous** substitutions

species 1	C GG	A CA	C T G
	arg	thr	leu
species 2	A GG	A T A	C T C
	arg	ile	leu

Nei-Gojobori Method (1983)

1. Count up the number of **synonymous** and **nonsynonymous** substitutions

species 1	C GG	A CA	C T G
	arg	thr	leu
species 2	A GG	A T A	C T C
	arg	ile	leu

Nei-Gojobori Method (1983)

1. Count up the number of **synonymous** and **nonsynonymous** substitutions

species 1	C GG	A CA	C T G
	arg	thr	leu
species 2	A GG	A TA	C T C
	arg	ile	leu

Nei-Gojobori Method (1983)

1. Count up the number of **synonymous** and **nonsynonymous** substitutions

species 1	CGG	ACA	CTG
	arg	thr	leu
species 2	AGG	ATA	CTC
	arg	ile	leu

Nei-Gojobori Method (1983)

1. Count up the number of **synonymous** and **nonsynonymous** substitutions

species 1	CGG	ACA	CTG
	arg	thr	leu
species 2	AGG	ATA	CTC
	arg	ile	leu

2 synonymous differences

1 nonsynonymous difference

Nei-Gojobori Method (1983)

2. Count up the number of **synonymous** and **nonsynonymous** sites

species 1

CGG	ACA	CTG
arg	thr	leu

species 2

AGG	ATA	CTC
arg	ile	leu

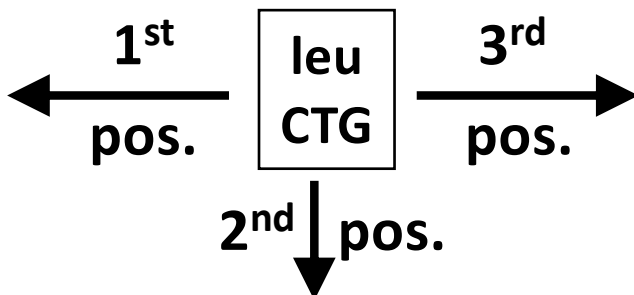
Nei-Gojobori Method (1983)

2. Count up the number of **synonymous** and **nonsynonymous** sites

species 1	CGG	ACA	CTG
	arg	thr	leu
species 2	AGG	ATA	CTC
	arg	ile	leu

(nonsyn) – Met ATG
 (syn) – Leu TTG
 (nonsyn) – Val GTG

= 2/3 nonsynonymous,
 1/3 synonymous site



CAG – gln (nonsyn)
 CGG – arg (nonsyn)
 CCG – pro (nonsyn)

= 1 nonsynonymous site

CTA – leu (syn)
 CTT – leu (syn)
 CTC – leu (syn)


= 1 synonymous site

This codon has:
 1.67 nonsyn sites
 1.33 syn sites

Nei-Gojobori Method (1983)

2. Count up the number of **synonymous** and **nonsynonymous** sites

species 1	CGG	ACA	CTG
	arg	thr	leu
species 2	AGG	ATA	CTC
	arg	ile	leu



This codon might have different proportion **synonymous** versus **nonsynonymous** in species 1 versus species 2

Most common approach is to take the average across both species

Nei-Gojobori Method (1983)

3. Compare rates of substitution at **synonymous**, **nonsynonymous** sites

species 1	CGG	ACA	CTG
	arg	thr	leu
species 2	AGG	ATA	CTC
	arg	ile	leu

$K_A = \# \text{ nonsynonymous substitutions} / \# \text{ nonsynonymous sites}$

$K_S = \# \text{ synonymous substitutions} / \# \text{ synonymous sites}$

$K_A = K_S$ Neutral expectation

$K_A < K_S$ Purifying selection

$K_A > K_S$ Frequent directional selection

Tests of Neutrality

- Polymorphism-based tests
 - Site frequency spectrum
 - Tajima's D
- Polymorphism and Divergence
 - McDonald-Kreitman test
- Divergence-based
 - K_A versus K_S
- Each of these methods has limitations
 - Power
 - Sensitivity to many population processes