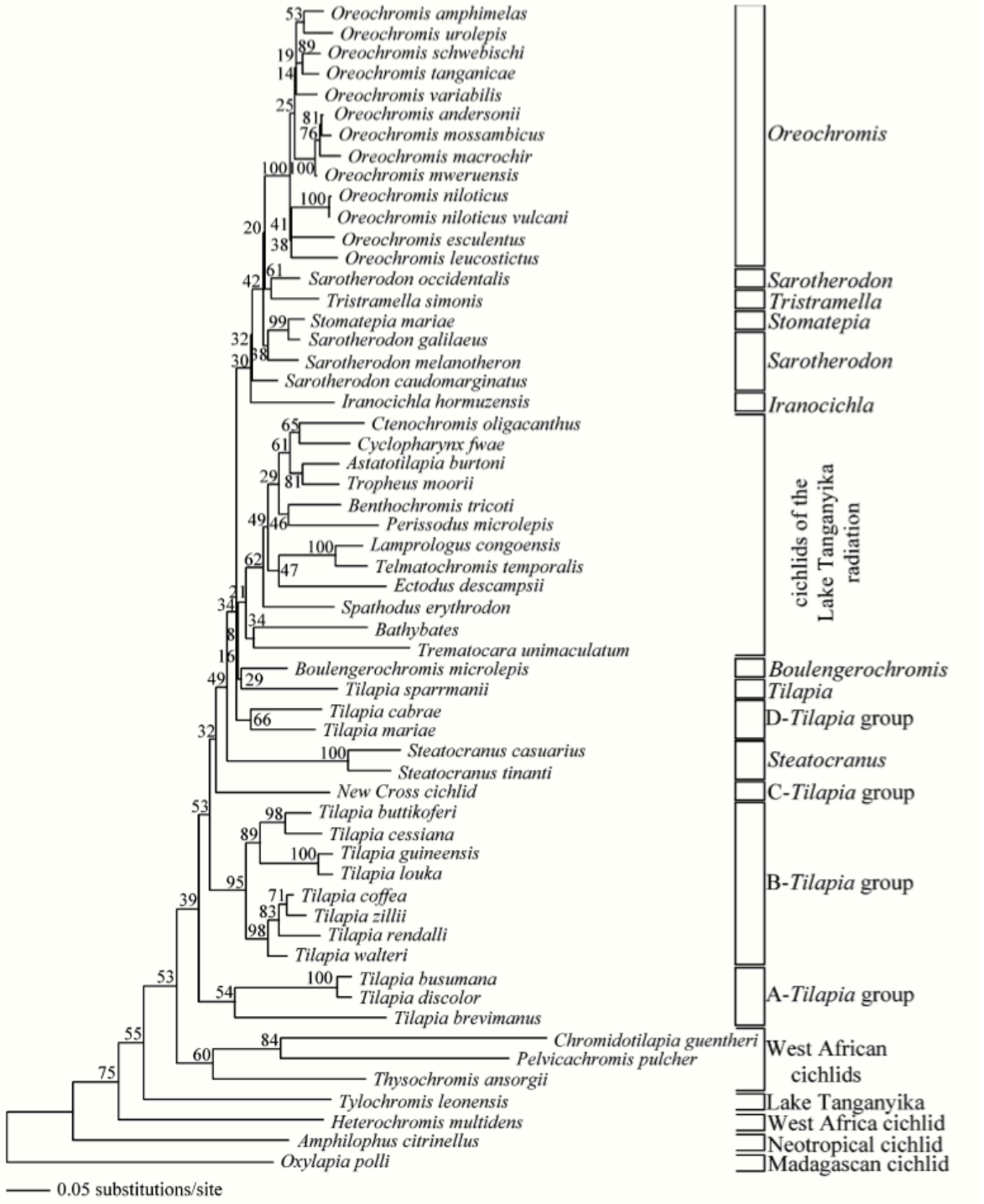
Molecular Signatures of Natural Selection

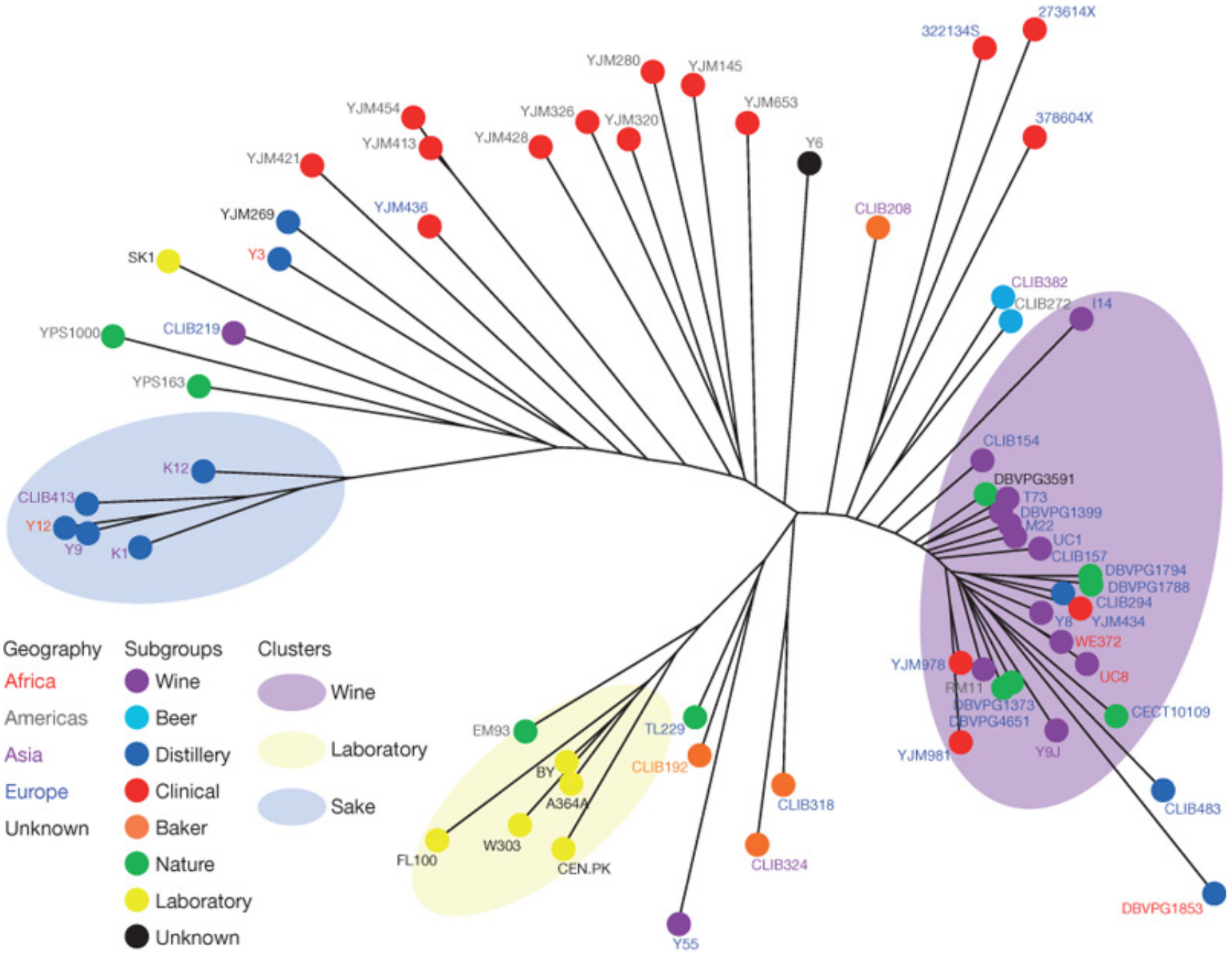
February 19, 2020

Genealogies

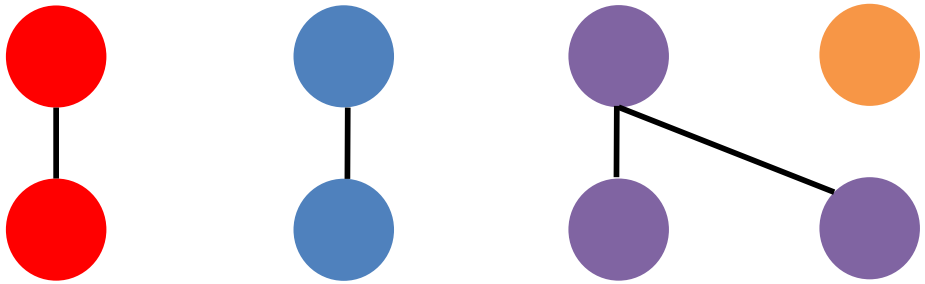
- Describe inheritance relationships among alleles
- Similar to phylogenetic tree

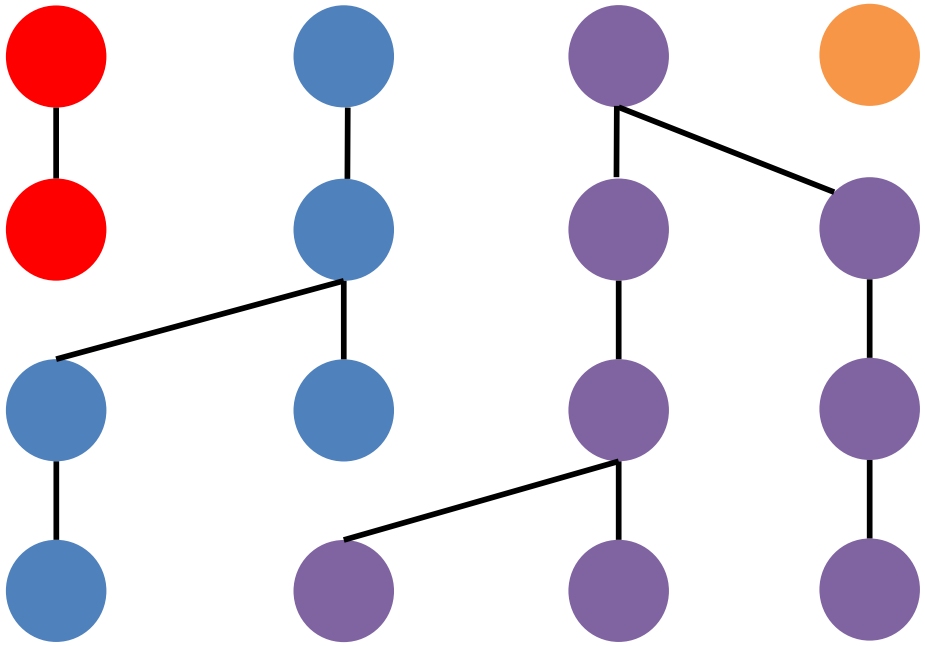

Oreochromis amphimelas
Oreochromis urolepis
Oreochromis schwebischi
Oreochromis tanganicae
Oreochromis variabilis
Oreochromis andersonii
Oreochromis mossambicus
Oreochromis macrochir
Oreochromis mweruensis
Oreochromis niloticus
Oreochromis niloticus vulcani
Oreochromis esculentus
Oreochromis leucostictus
Sarotherodon occidentalis
Tristramella simonis
Stomatepia mariae
Sarotherodon galilaeus
Sarotherodon melanotheron
Sarotherodon caudomarginatus
Iranocichla hormuzensis
Ctenochromis oligacanthus
Cyclopharynx fwae
Astatotilapia burtoni
Tropheus moorii
Benthochromis tricoti
Perissodus microlepis
Lamprologus congoensis
Telmatochromis temporalis
Ectodus descampsii
Spathodus erythrodon
Bathybates
Trematocara unimaculatum
Boulengerochromis microlepis
Tilapia sparrmanii
Tilapia cabrae
Tilapia mariae
Steatocranus casuarius
Steatocranus tinanti
New Cross cichlid
Tilapia buttikoferi
Tilapia cessiana
Tilapia guineensis
Tilapia louka
Tilapia coffea
Tilapia zillii
Tilapia rendalli
Tilapia walteri
Tilapia busumana
Tilapia discolor
Tilapia brevimanus
Chromidotilapia guentheri
Pelvicachromis pulcher
Thysochromis ansorgii
Tylochromis leonensis
Heterochromis multidens
Amphilophus citrinellus
Oxylapia polli
0.05 substitutions/site
Oreochromis
Sarotherodon
Tristramella
Stomatepia
Sarotherodon
Iranocichla
cichlids of the Lake Tanganyika radiation
Boulengerochromis
Tilapia
D-Tilapia group
Steatocranus
C-Tilapia group
B-Tilapia group
A-Tilapia group
West African cichlids
Lake Tanganyika
West Africa cichlid
Neotropical cichlid
Madagascan cichlid

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

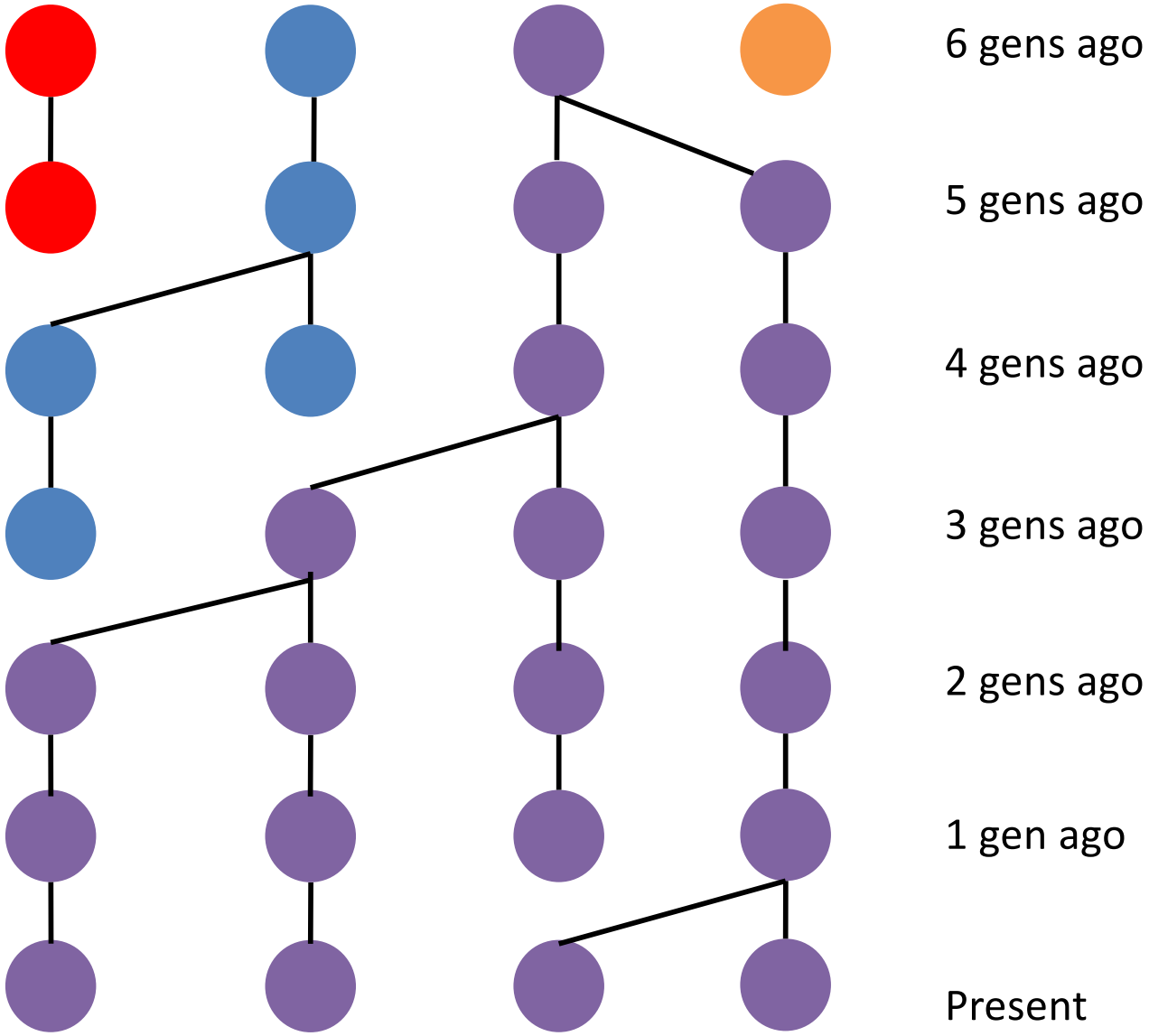


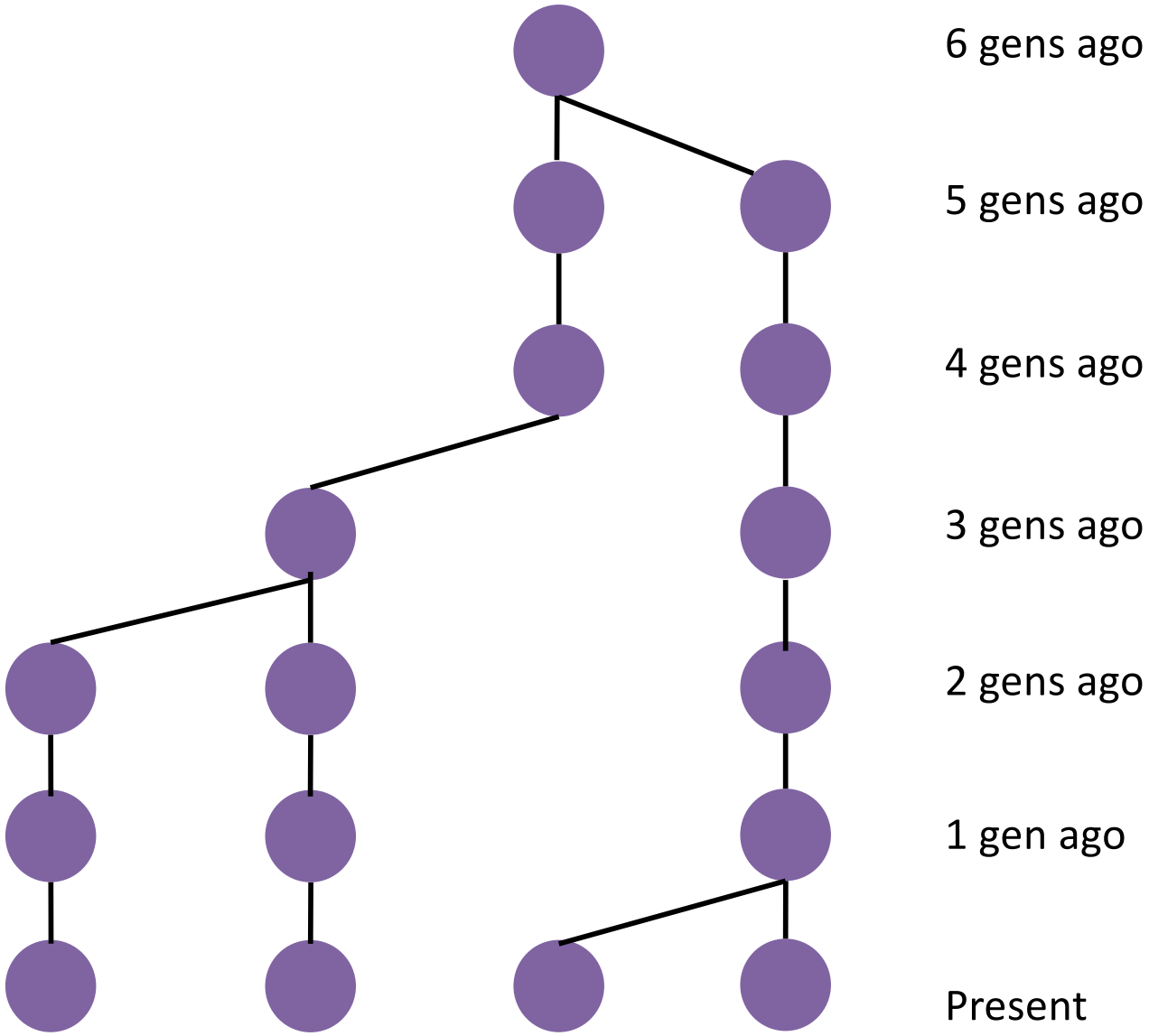
J Schacherer *et al. Nature* (2009) doi:10.1038/nature07670

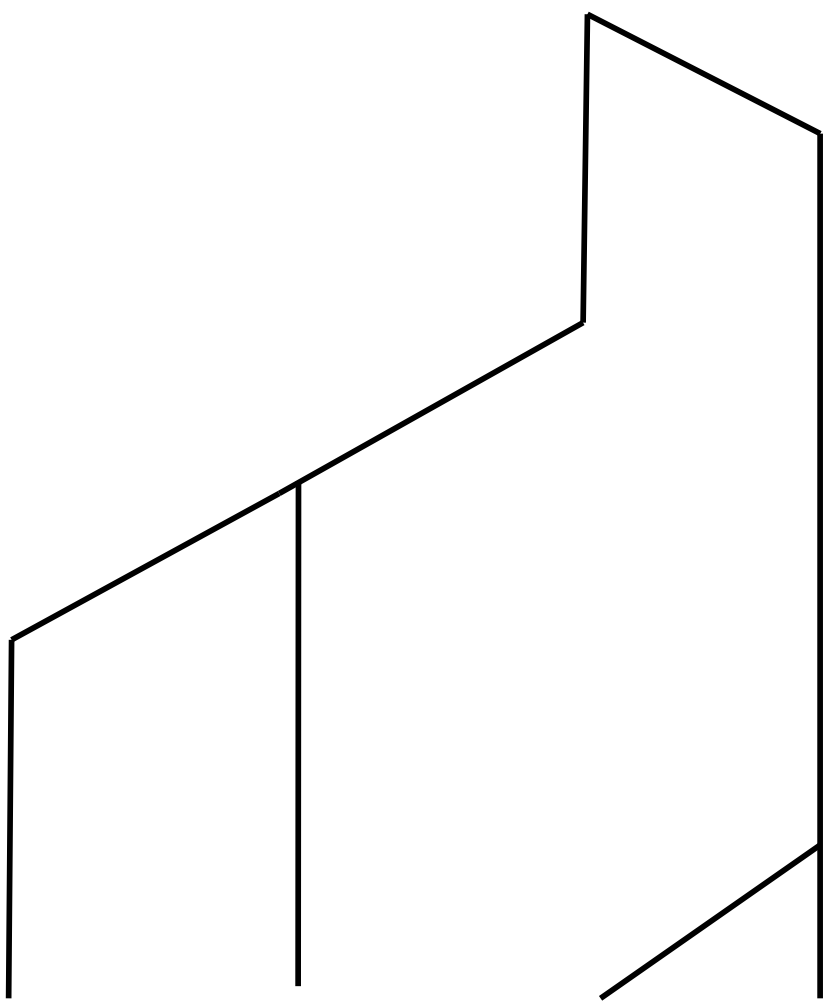




MRCA = 6 gen ago




6 gens ago
5 gens ago
4 gens ago
3 gens ago
2 gens ago
1 gen ago
Present



6 gens ago

5 gens ago

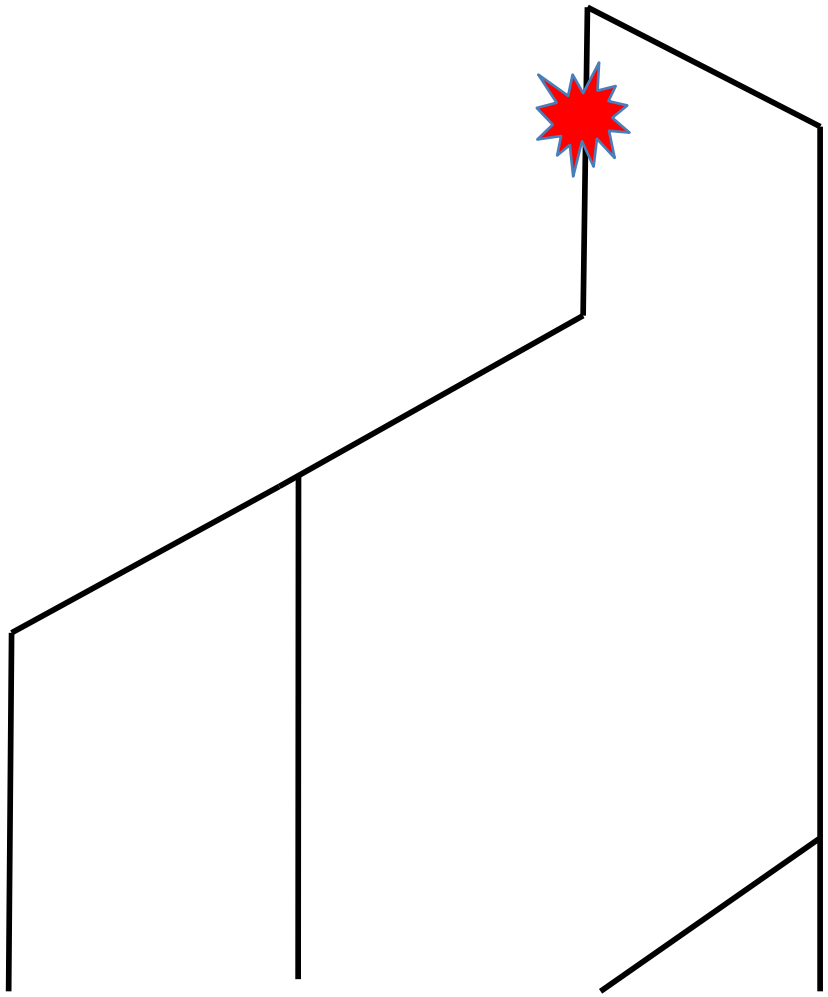
4 gens ago

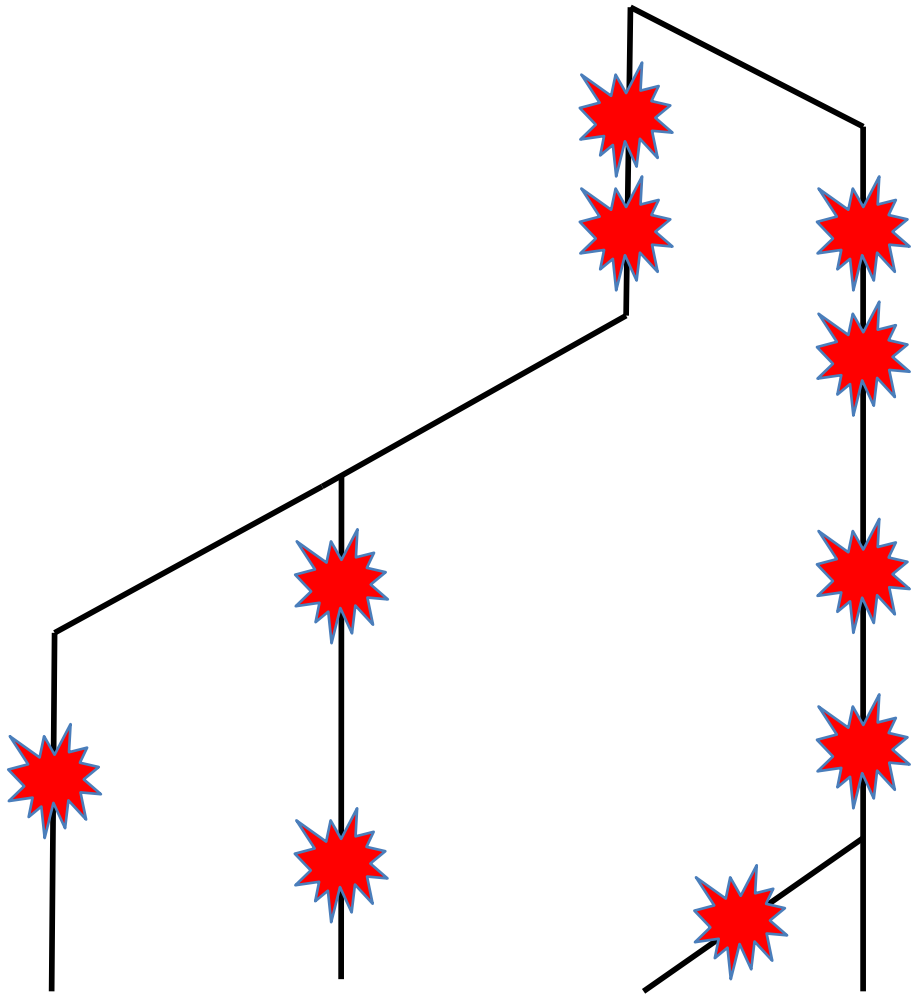
3 gens ago

2 gens ago

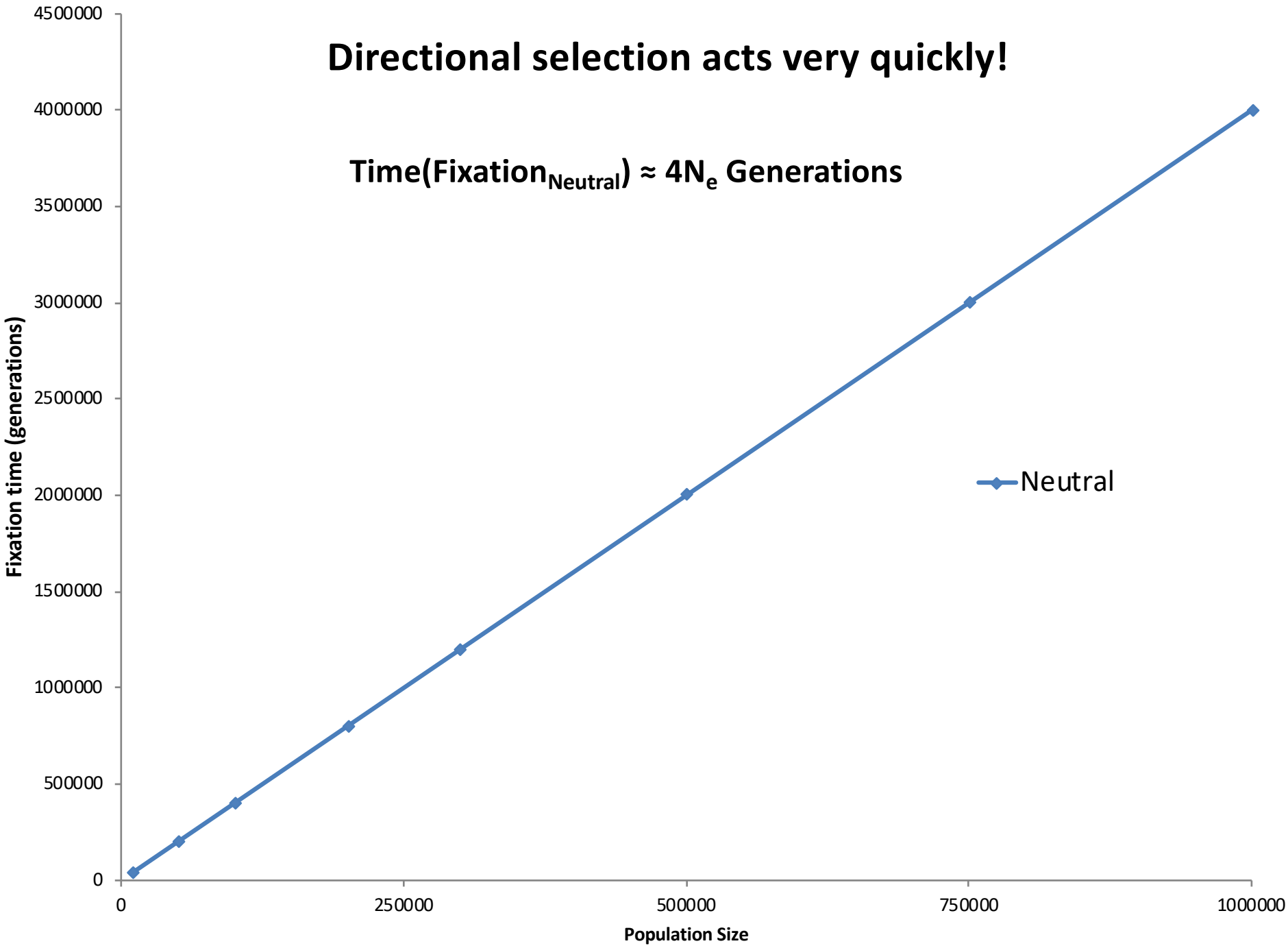
1 gen ago

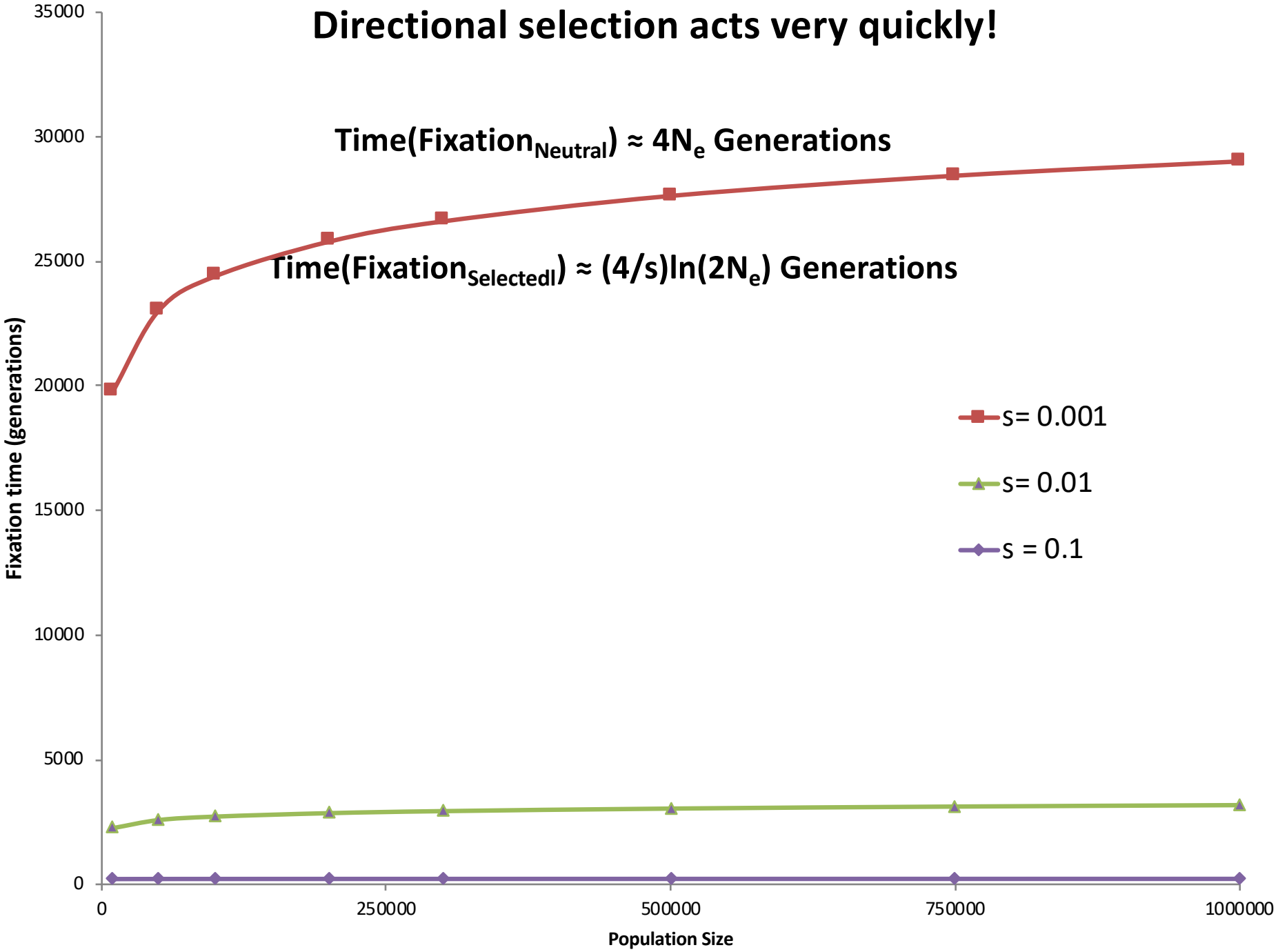
Present

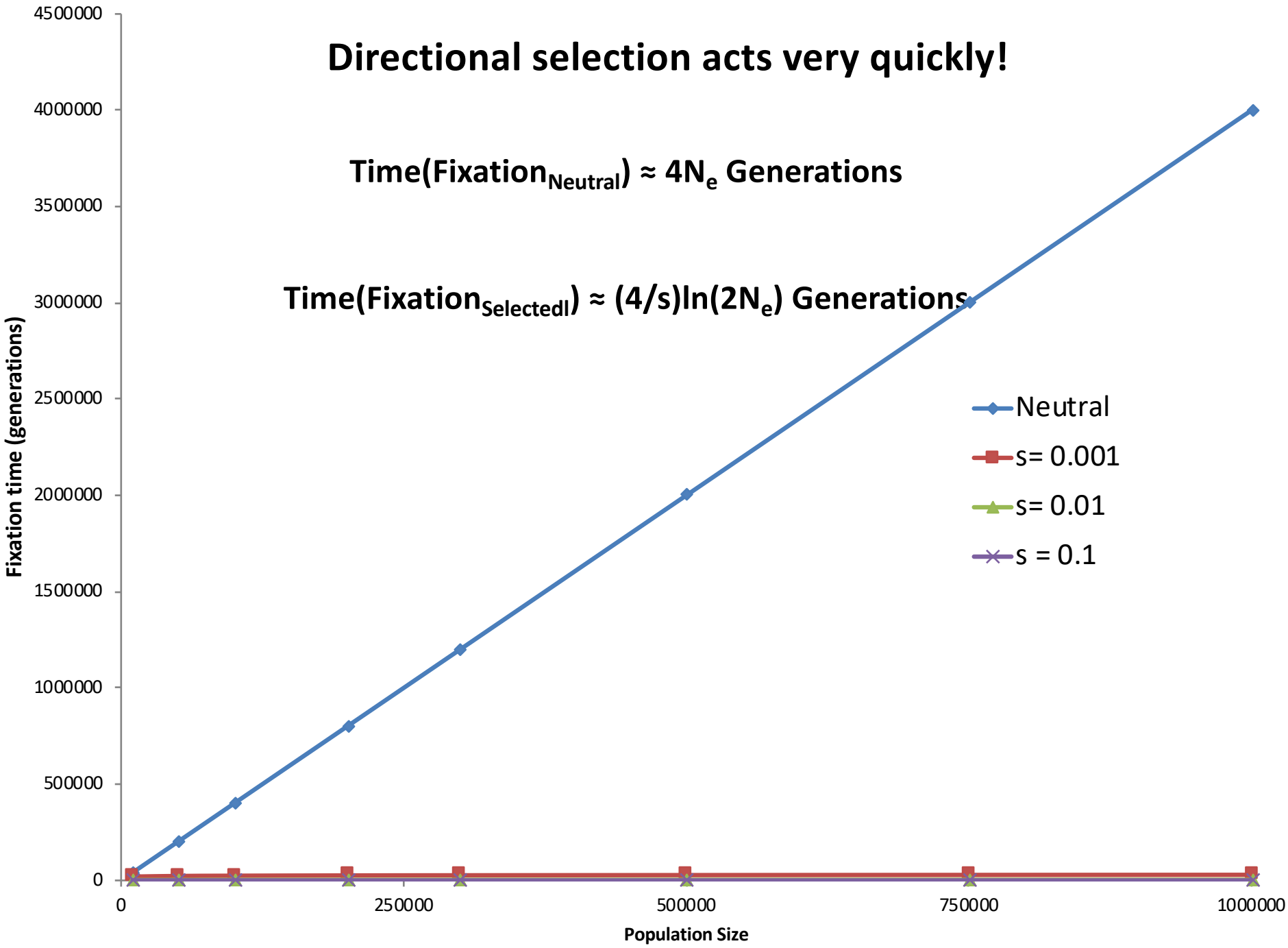




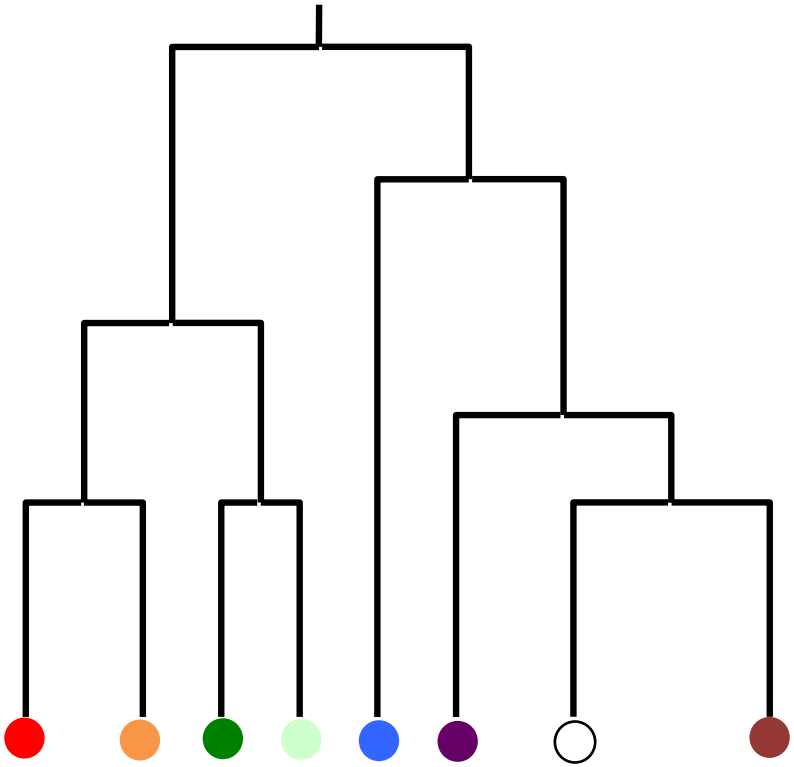
How does this change under directional selection?


Directional selection acts very quickly!
Time(Fixation$_{Neutral}$) ≈ 4N$_e$ Generations
Fixation time (generations)
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
250000
500000
750000
1000000
Population Size
Neutral


Directional selection acts very quickly!
Time(Fixation$_{Neutral}$) ≈ 4N$_e$ Generations
Time(Fixation$_{Selectedl}$) ≈ (4/s)ln(2N$_e$) Generations
Fixation time (generations)
35000
30000
25000
20000
15000
10000
5000
0
0
250000
500000
750000
1000000
Population Size
s= 0.001
s= 0.01
s = 0.1

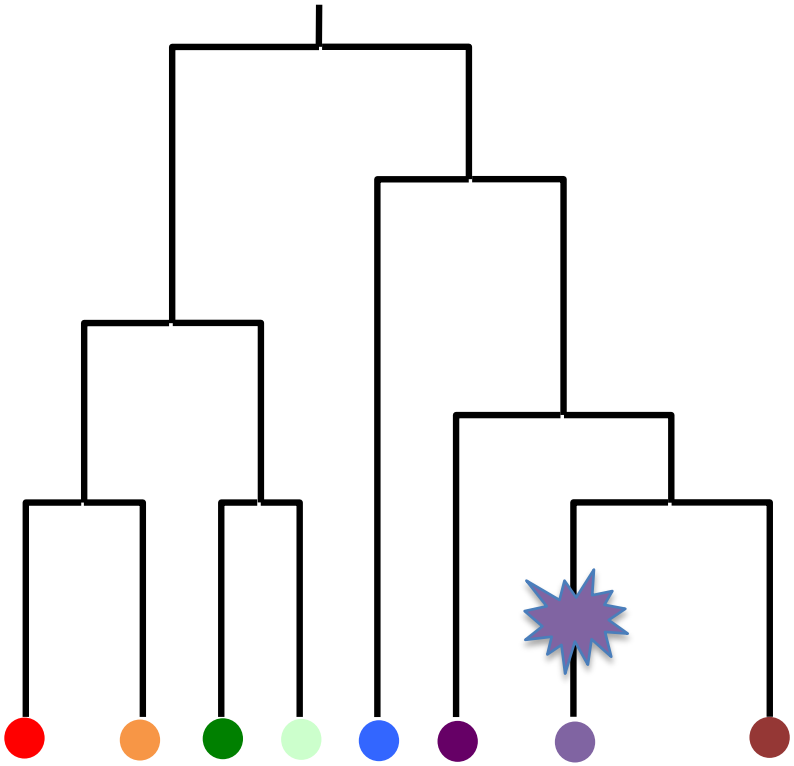

Directional selection acts very quickly!
Time(Fixation$_{Neutral}$) ≈ 4N$_e$ Generations
Time(Fixation$_{Selectedl}$) ≈ (4/s)ln(2N$_e$) Generations
Fixation time (generations)
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
250000
500000
750000
1000000
Population Size
Neutral
s= 0.001
s= 0.01
s = 0.1

The genealogical structure of a selective sweep



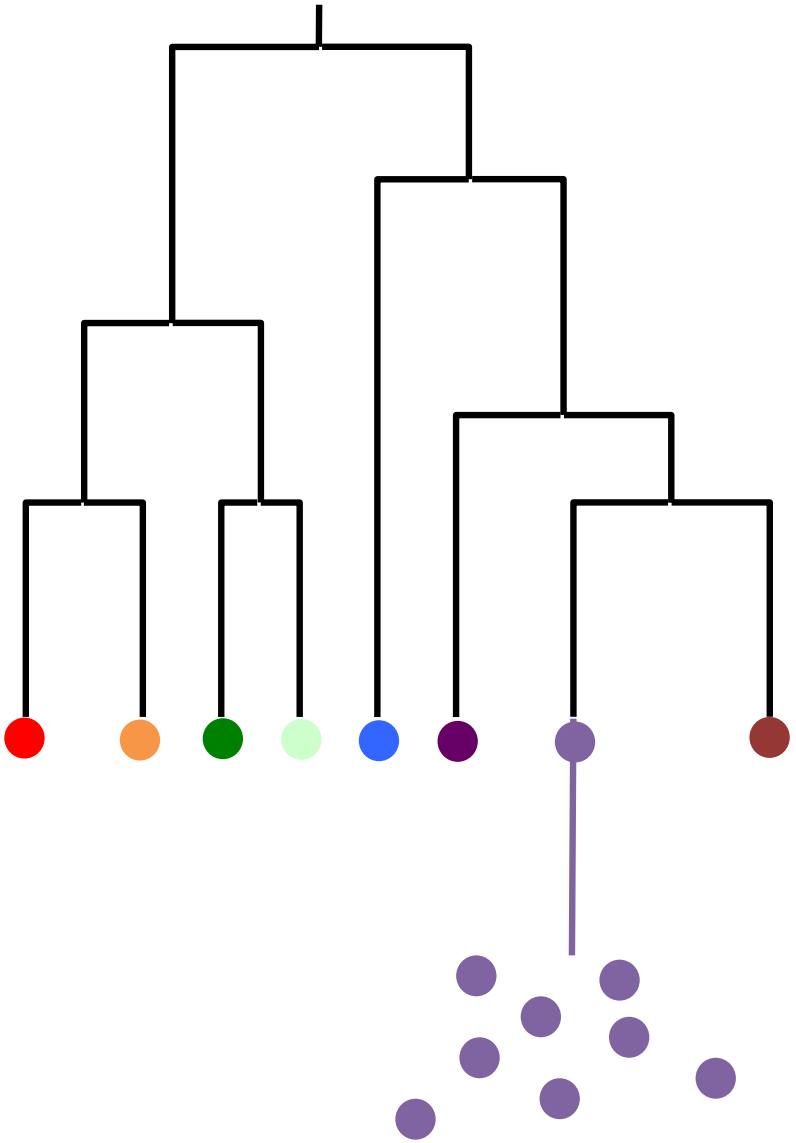
Population pre-selection

The genealogical structure of a selective sweep



Population pre-selection

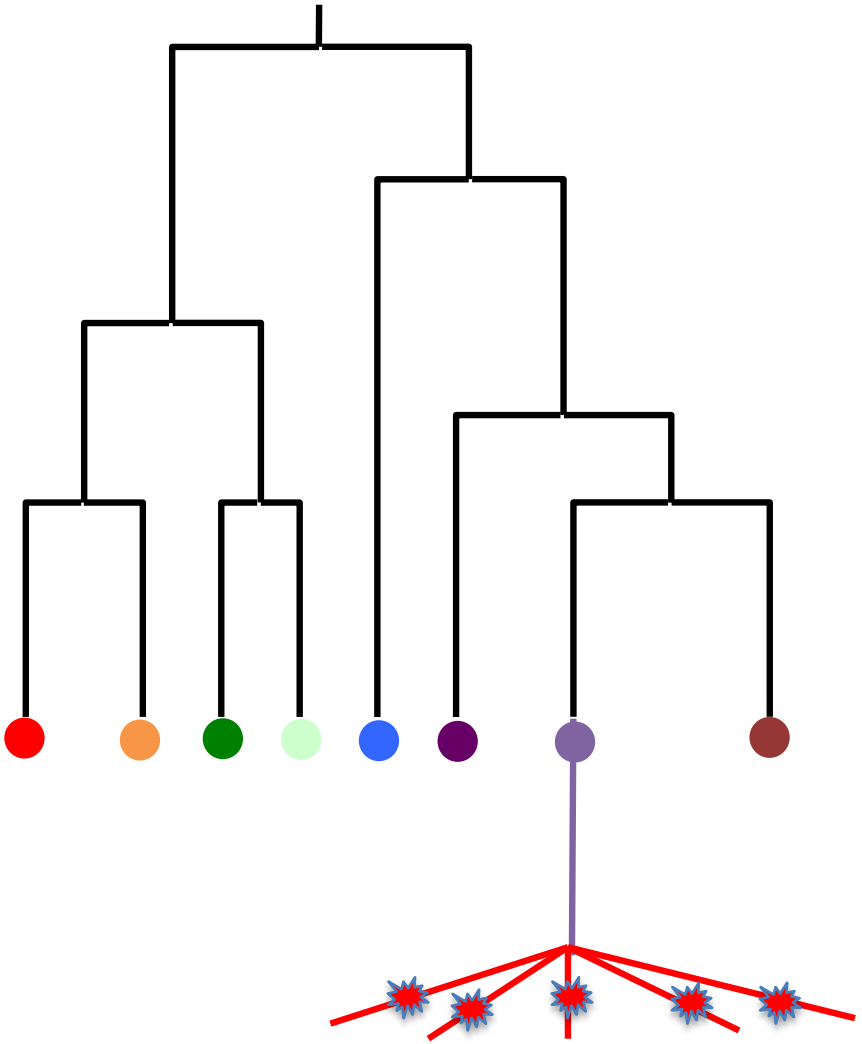
The genealogical structure of a selective sweep



Population pre-selection

During selection

The genealogical structure of a selective sweep



Population pre-selection

During selection

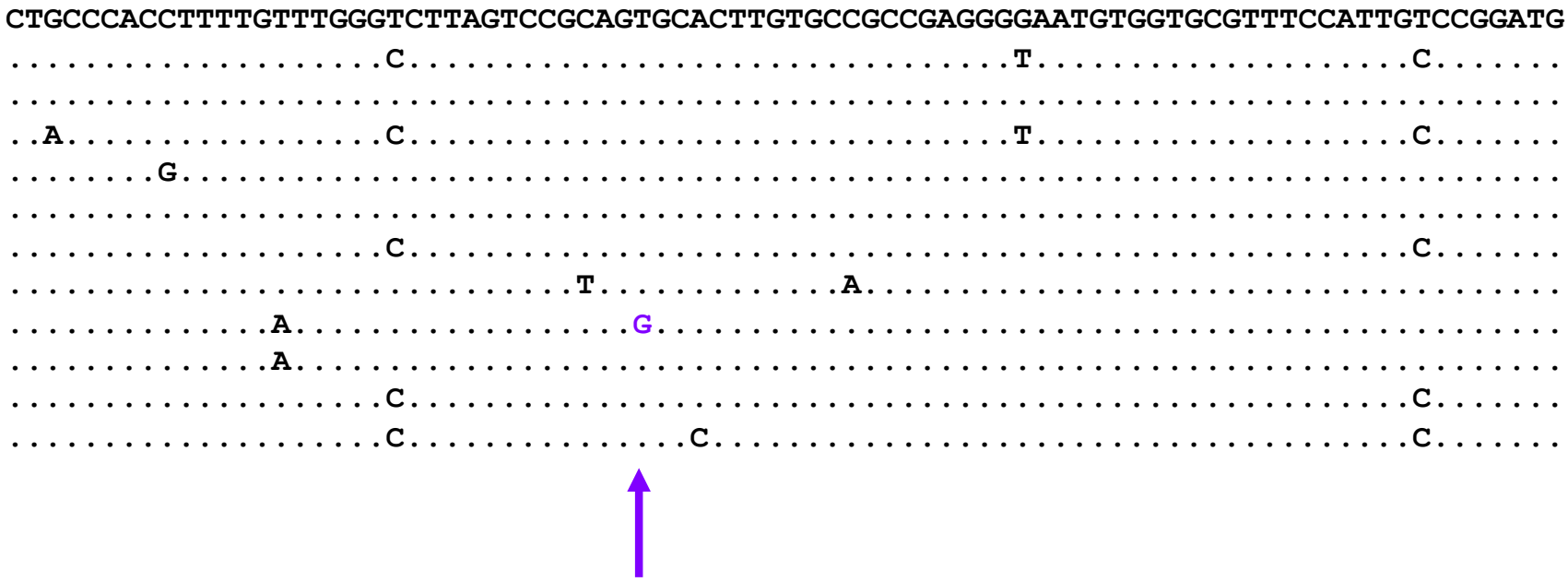
Mutational recovery post-selection

Effects of selection on sequences

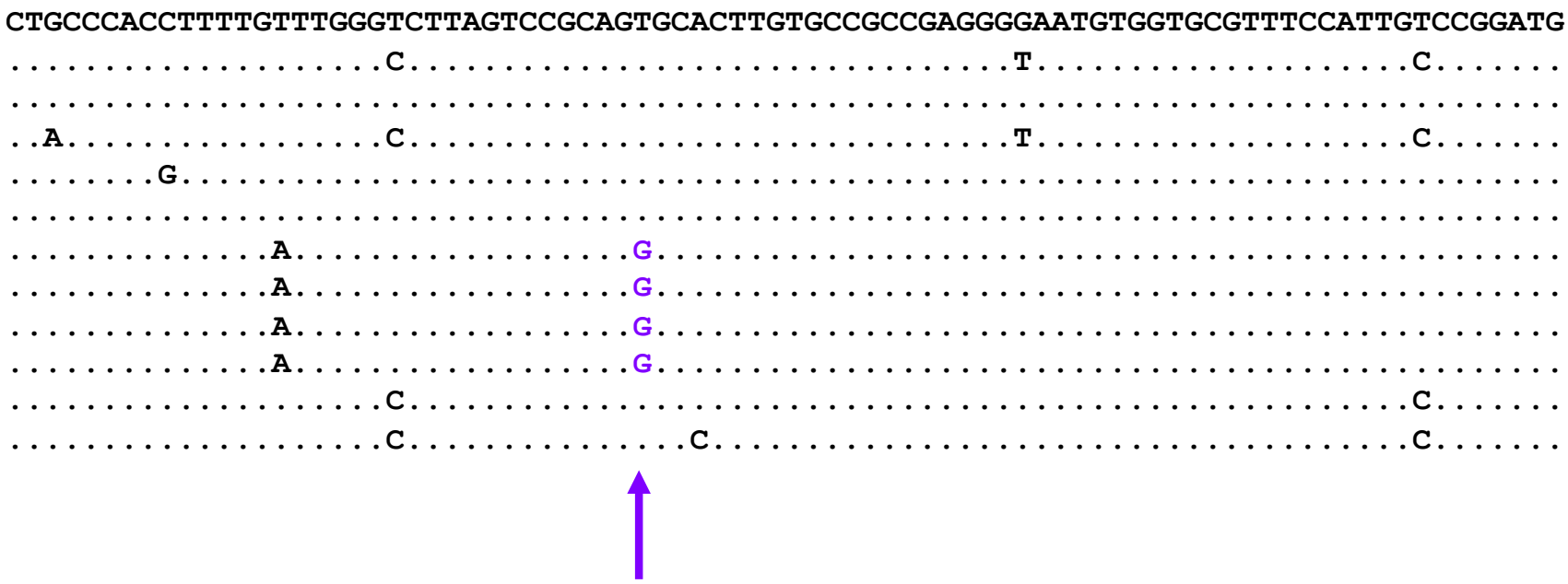
```
CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
....................C.................................T....................C.......
.....................................................................................
..A.................C.................................T....................C.......
........G.....................................................................
.....................................................................................
....................C.....................................................C.......
............................T............A.....................................
..............A......................................................................
..............A......................................................................
....................C.....................................................C.......
....................C..............C.....................................C.......
```

This is our population prior to selection

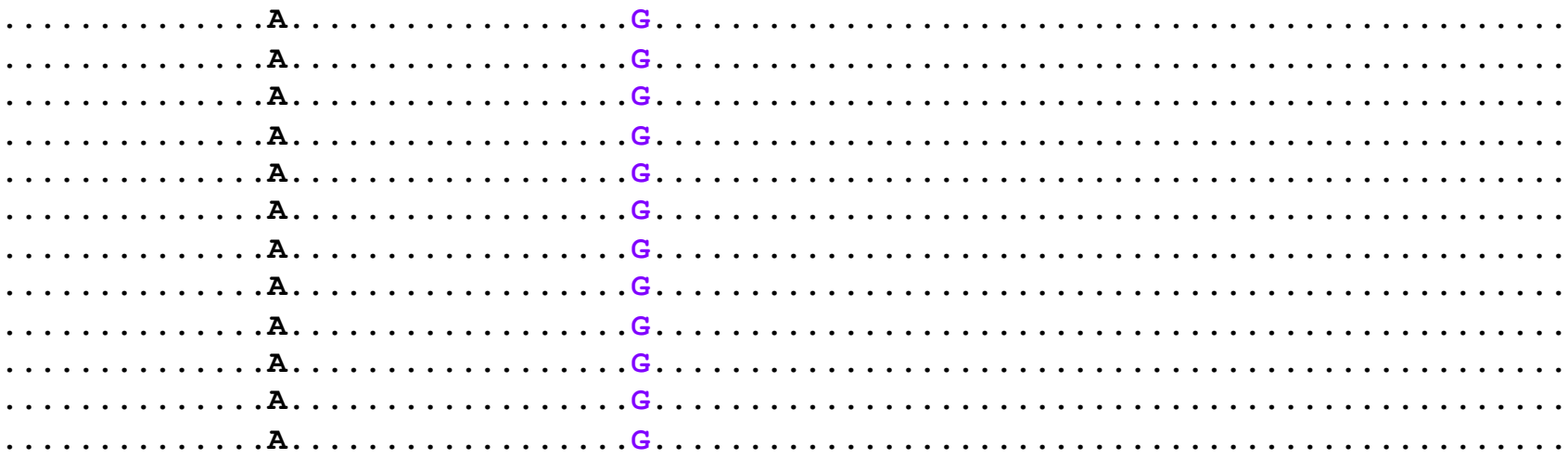
Effects of selection on sequences

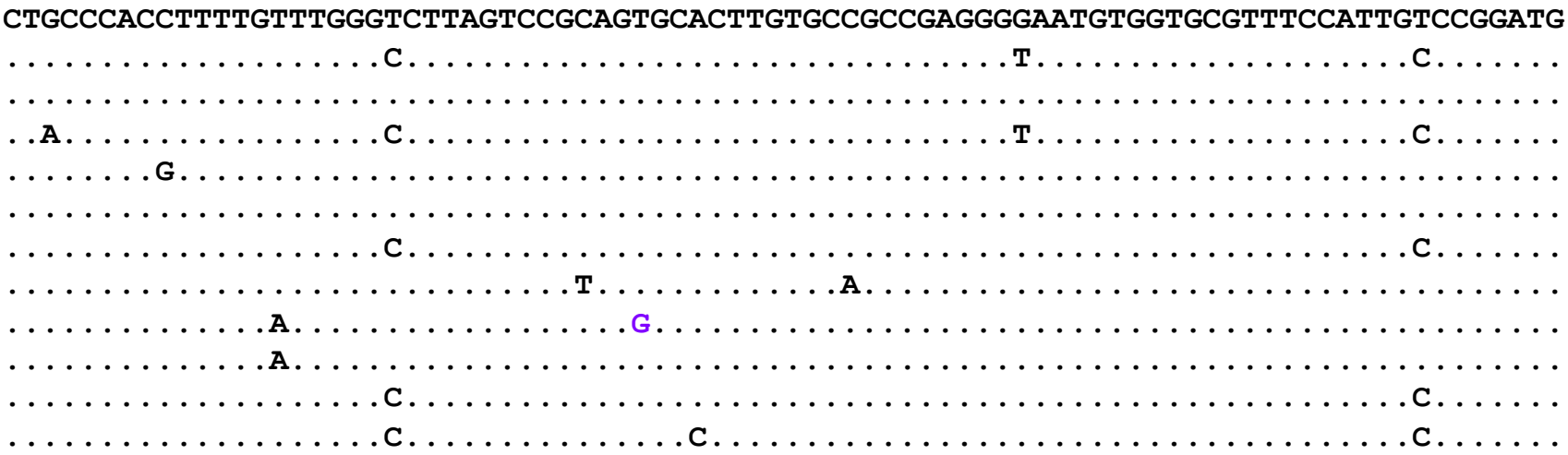


Effects of selection on sequences



Effects of selection on sequences




CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
C T C
A C T C
G
C C
T A
A G
A
C C
C C C

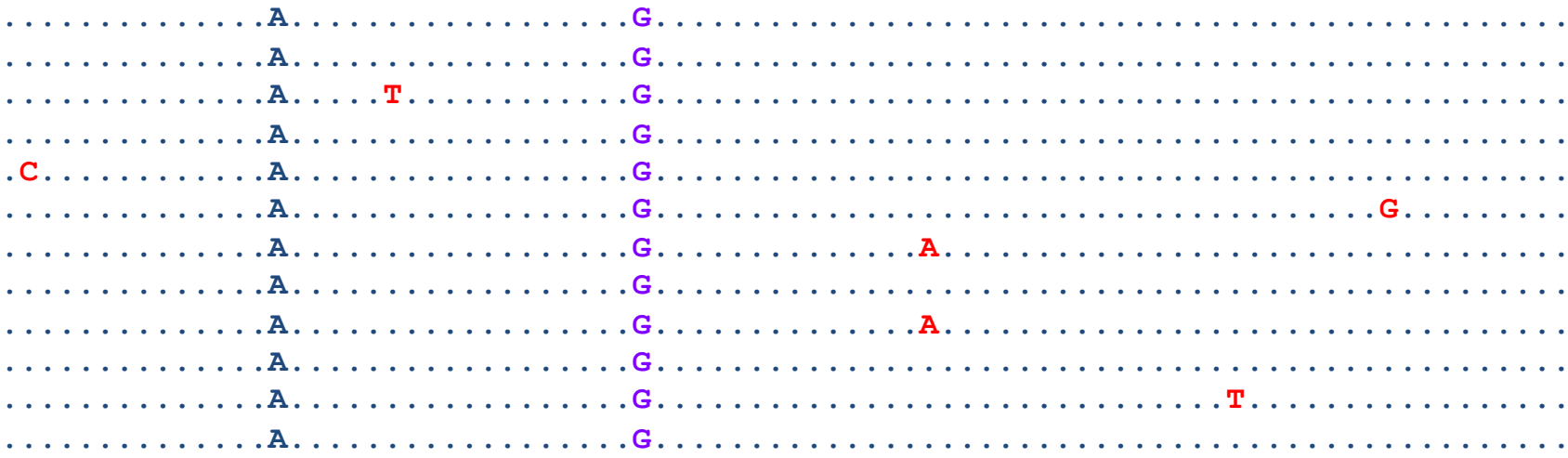
LOW recombination


CTGCCCACCTTTTGATTGGGTCTTAGTCCGCAGGGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
LOW recombination

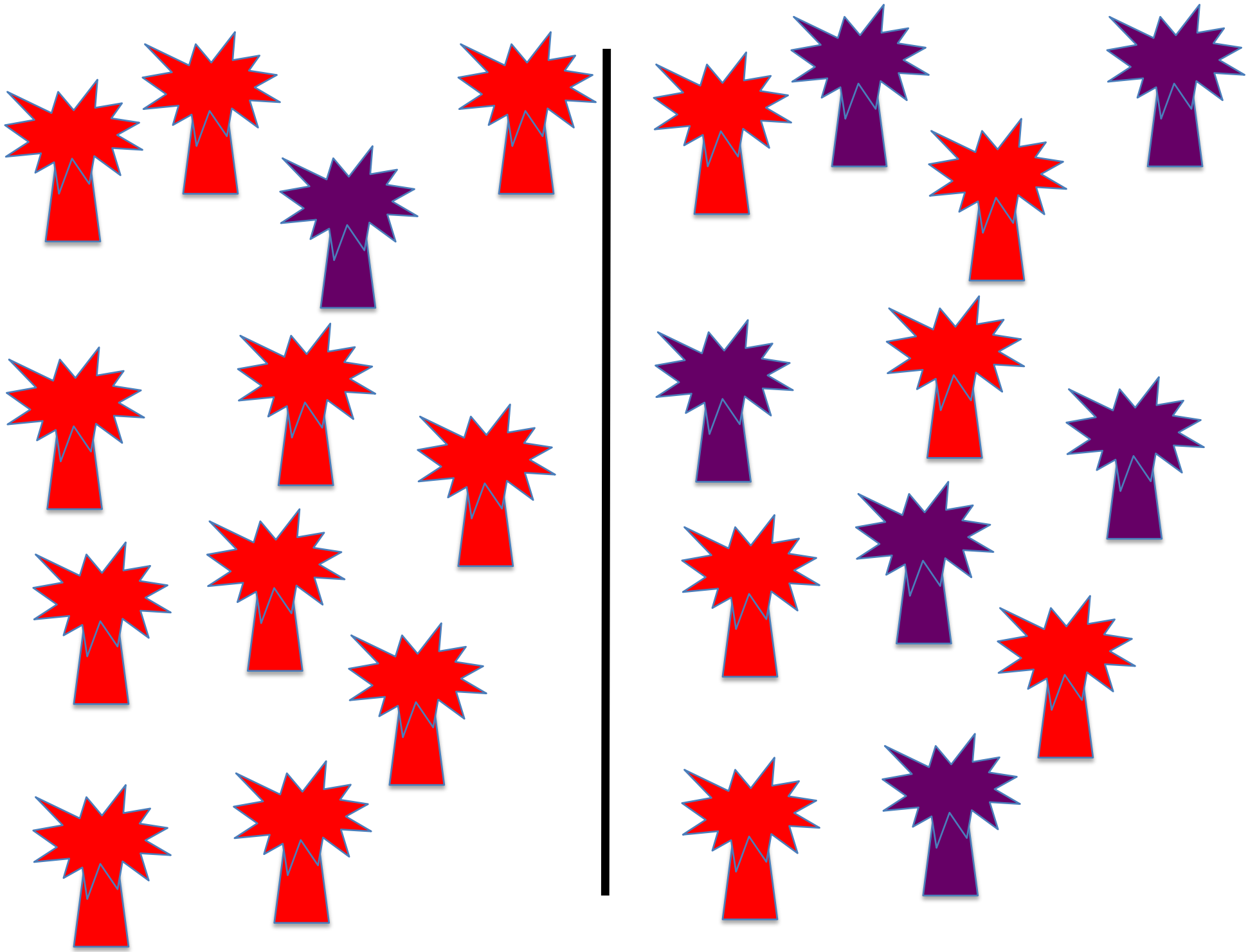

CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGGGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
HIGH recombination

Effects of selection on sequences

Effects of selection on sequences



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



Summarizing molecular data

```
#1    CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
#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

Θ

Haplotype number

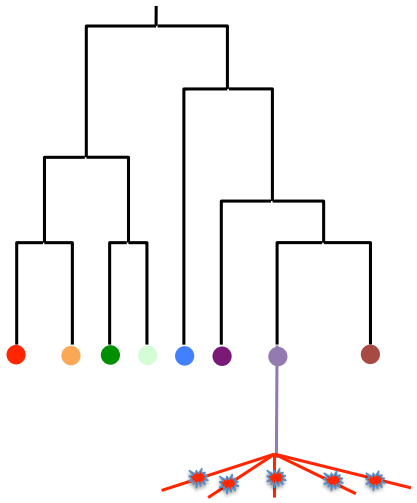
Haplotype diversity

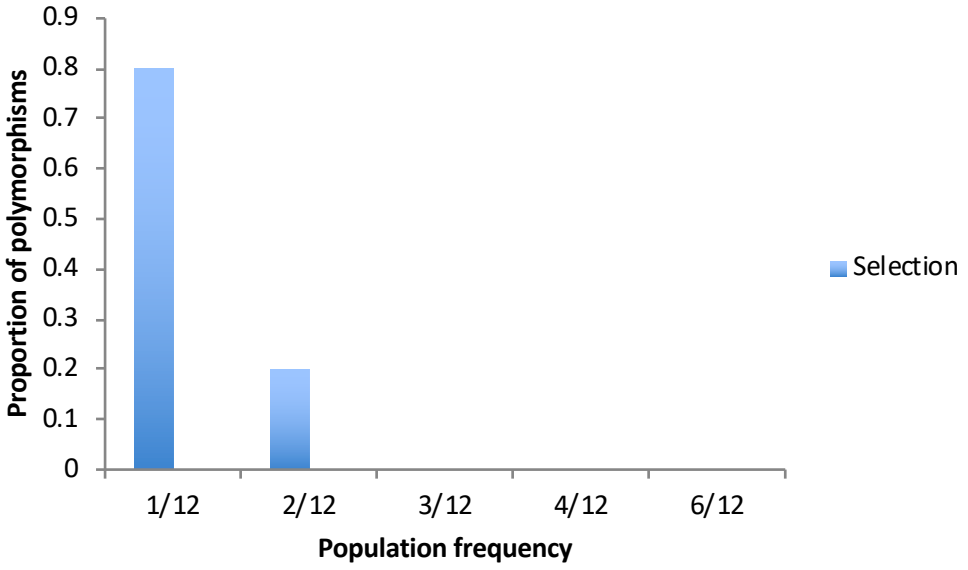
…

Using the Site Frequency Spectrum

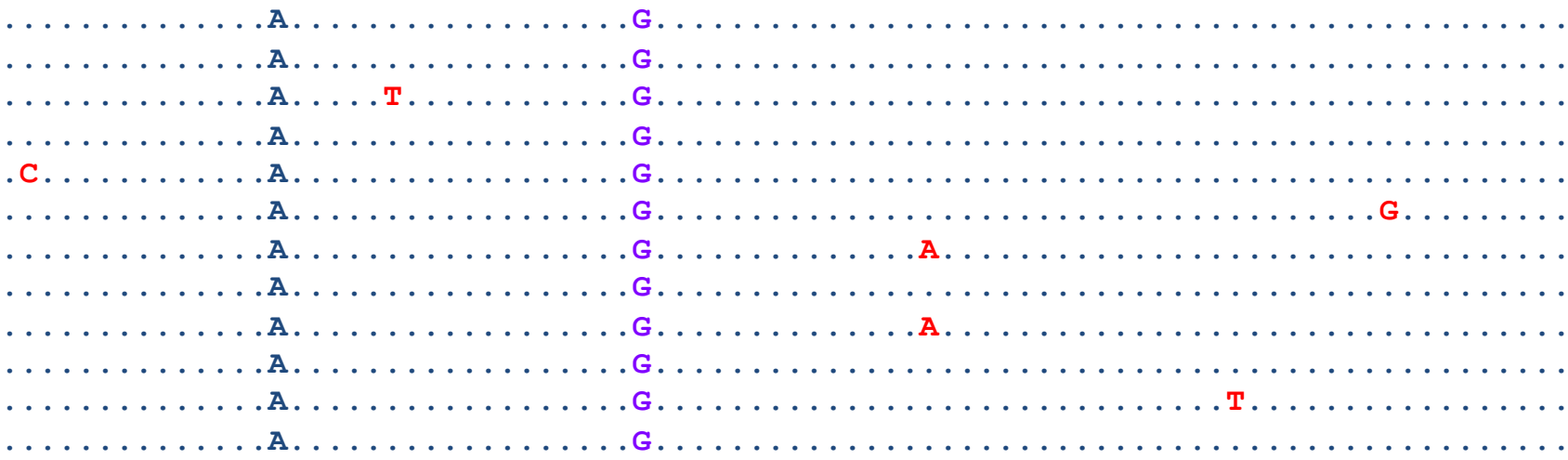
```
..............A..................G.....................................................
..............A..................G.....................................................
..............A.....T............G.....................................................
..............A..................G.....................................................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

1/12 1/12 2/12 1/12 1/12

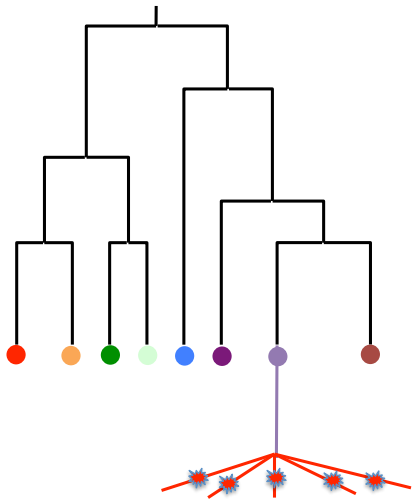


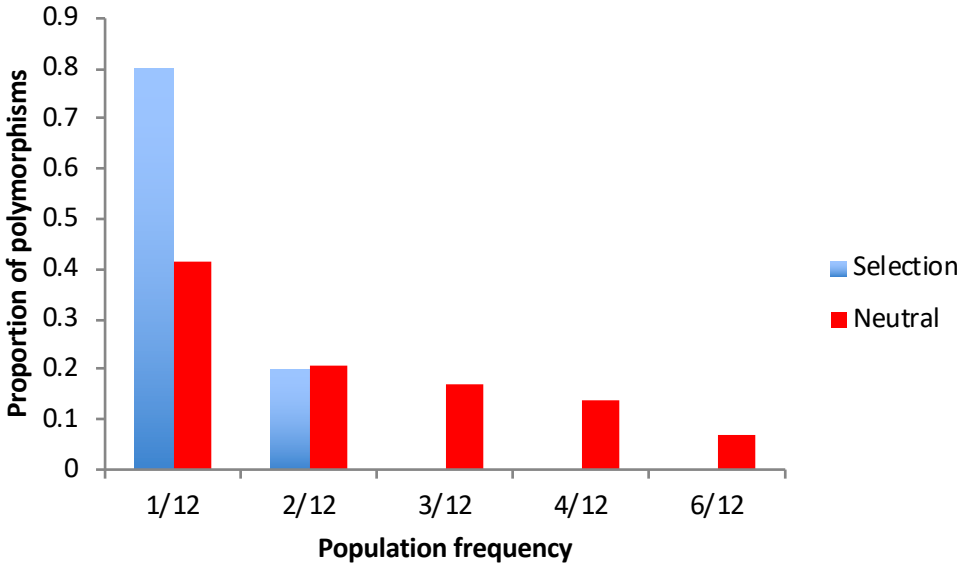


Using the Site Frequency Spectrum



1/12 1/12 2/12 1/12 1/12





Using estimates of theta

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

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

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

	a	b	c	d
a				
b				
c				
d				

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

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

	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	*j*
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			
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			
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

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

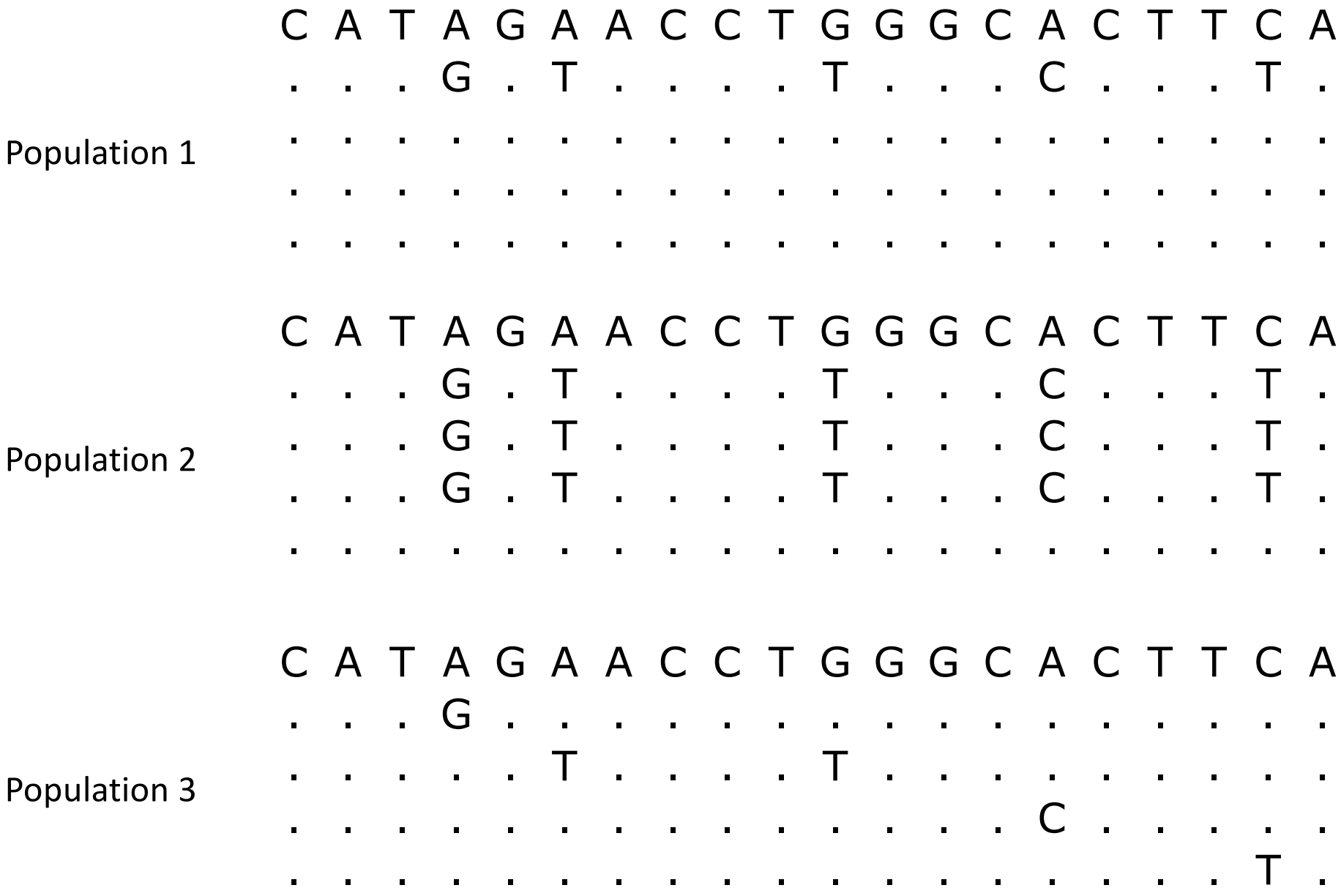
DNA Variation

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



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

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

Using estimates of theta

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

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    CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
#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

Θ

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…)

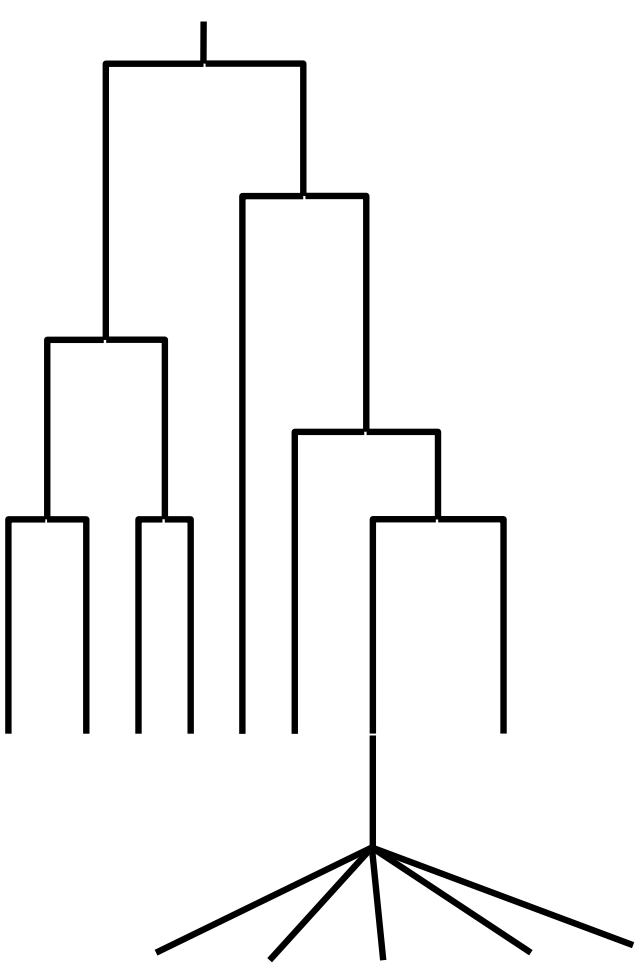
Modeling selection (Kim and Stephan 2002…)

…

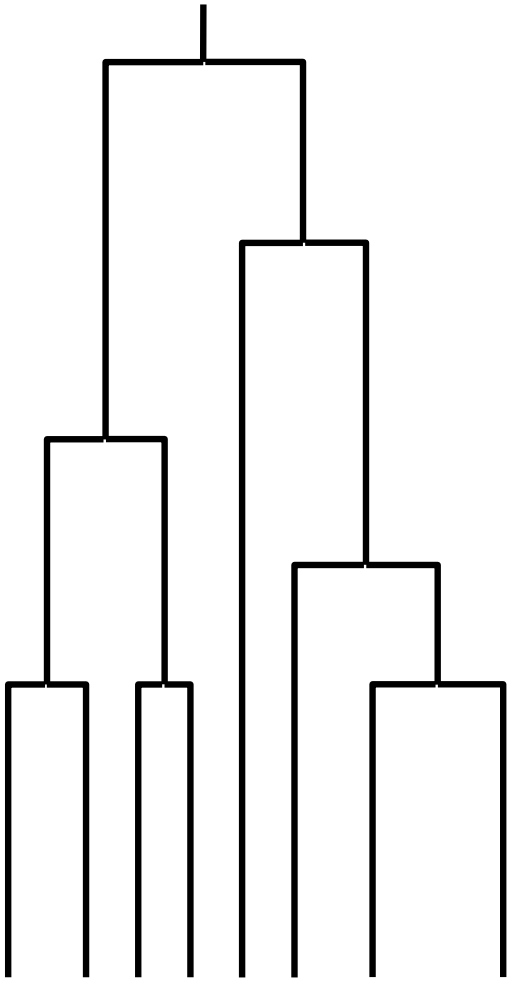
SFS-based tests sensitive to other factors

- Demography
  - Population bottlenecks
  - Population expansions
- Population substructure
- Migration

Comparing observed genealogy to neutral expectation

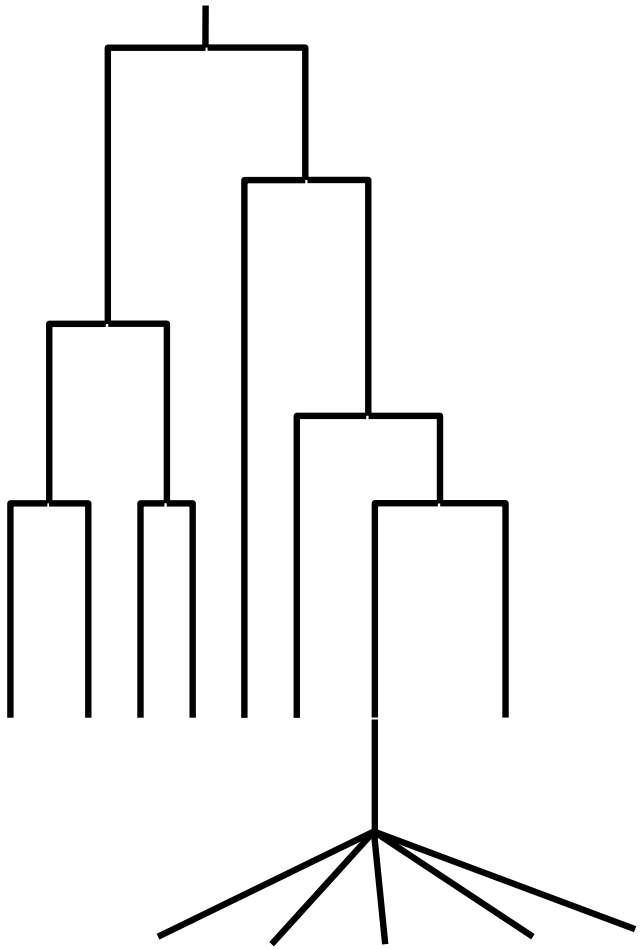


selective sweep

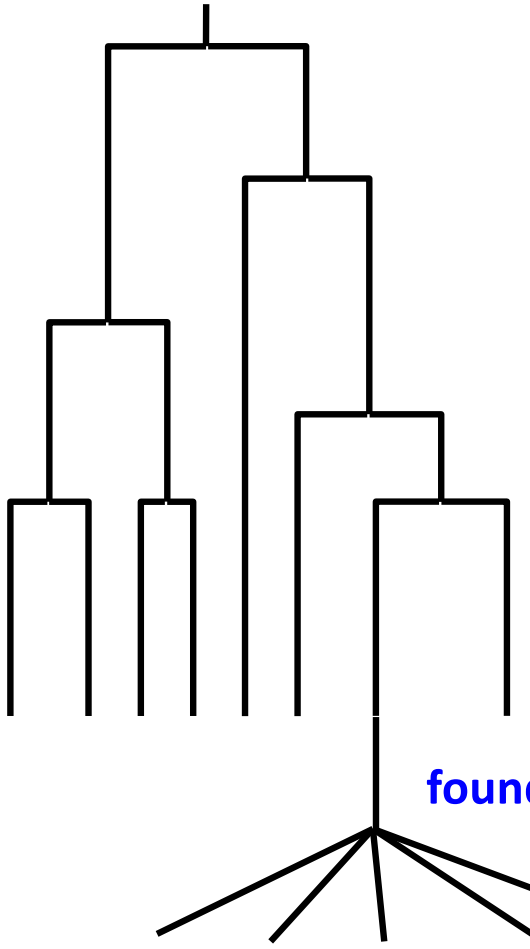


Neutrality

Genealogy may be impacted by forces other than selection



selective sweep

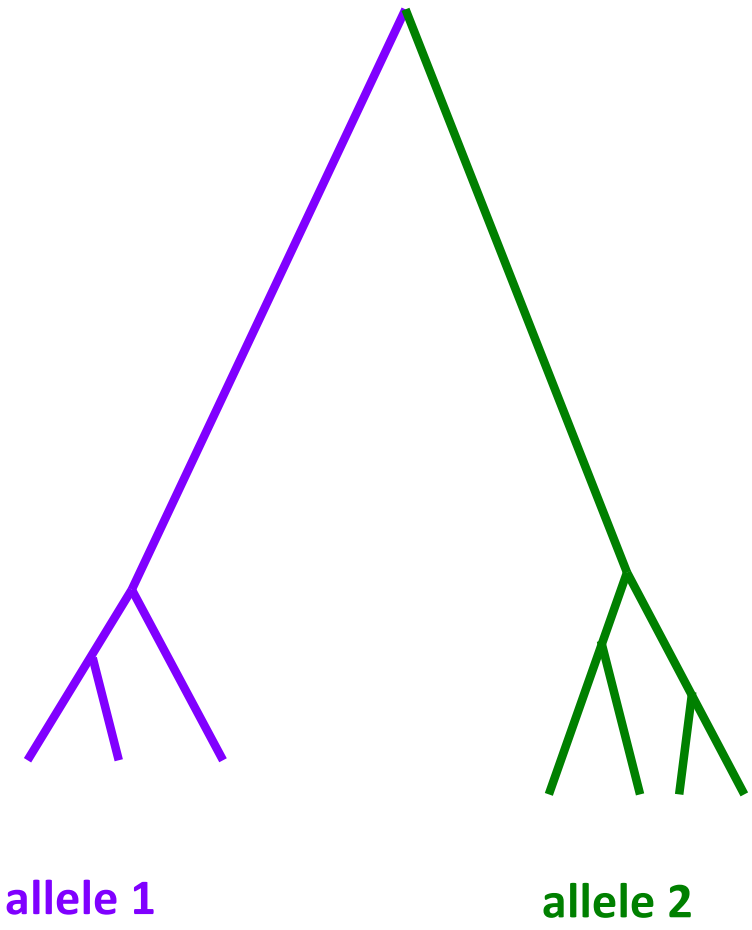


ancestral population

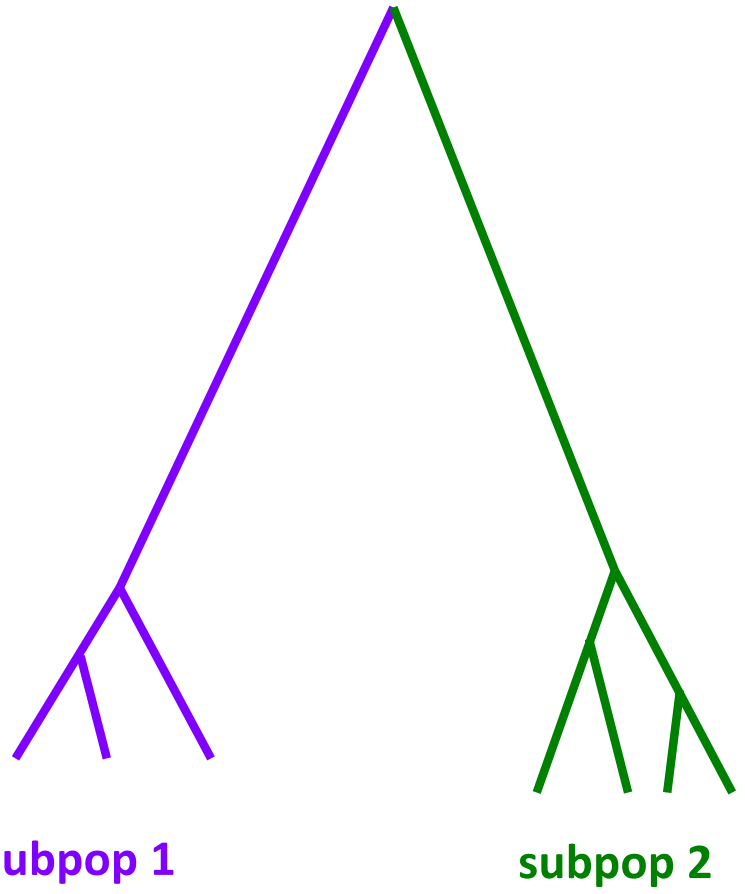
founding event

radiation / expansion

Genealogy may be impacted by forces other than selection



ancient balanced polymorphism



population structure

What if we have more data?

```
Indiv1    CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
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     CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
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     CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
Indiv2     ....................C.............................................C.......
Indiv3     ....................C.............................................C.......
Indiv4     ..A.................C.............................................C.......
Indiv5     ........G.....................................A.....................................
Indiv6     ....................C.............................................C.......
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    CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
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.....................
```

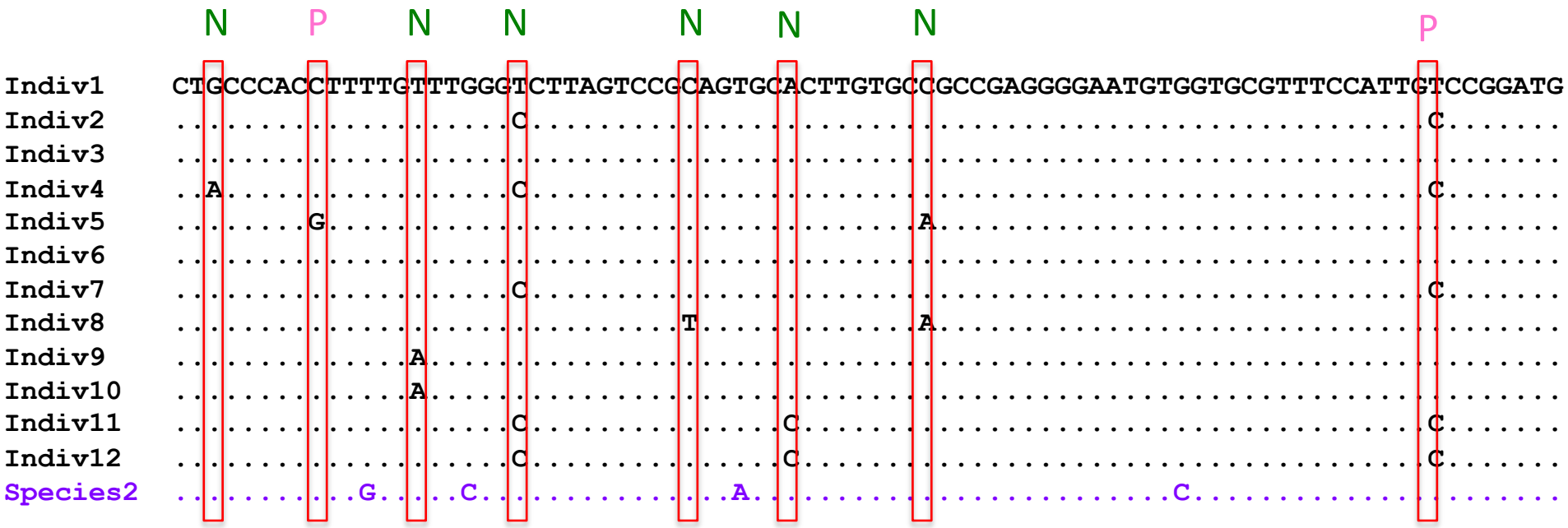
Neutral

Potentially Non-neutral

What if we have more data?

```
          N     P  N  N  P  N        N  N  N     N           N           P
Indiv1    CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
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?



Polymorphic

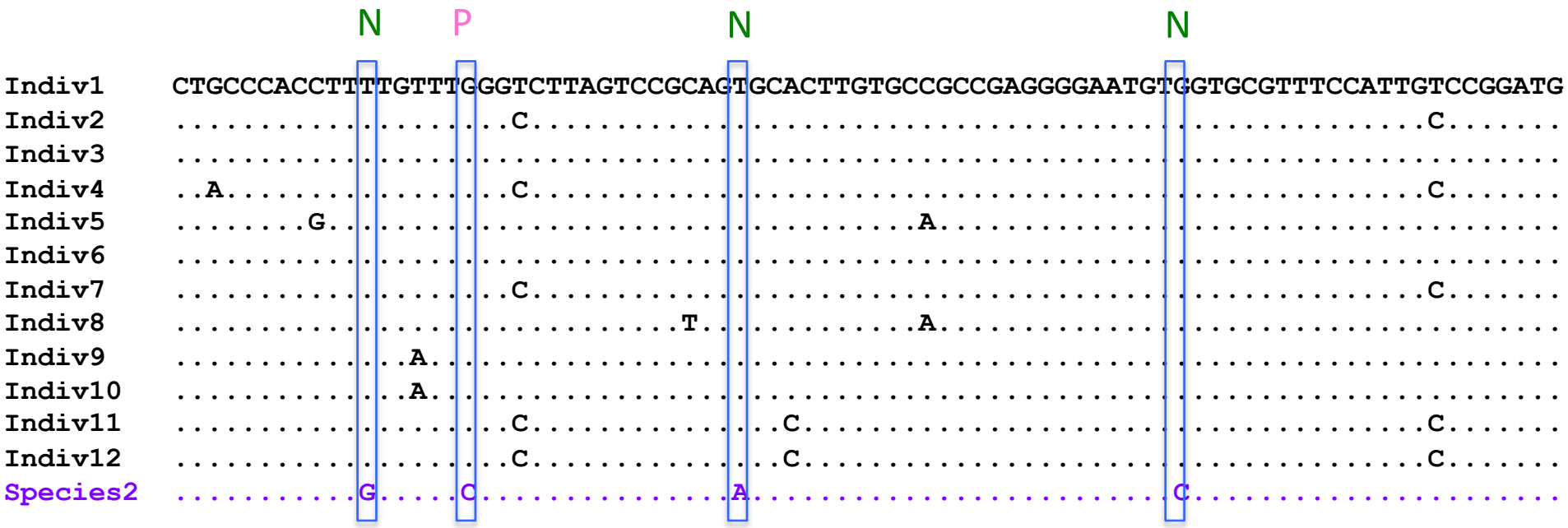
Divergent

Neutral

Potentially Non-neutral

	Poly	Div
Neutral	6	
Potentially Non-Neutral	2	

What if we have more data?



Polymorphic

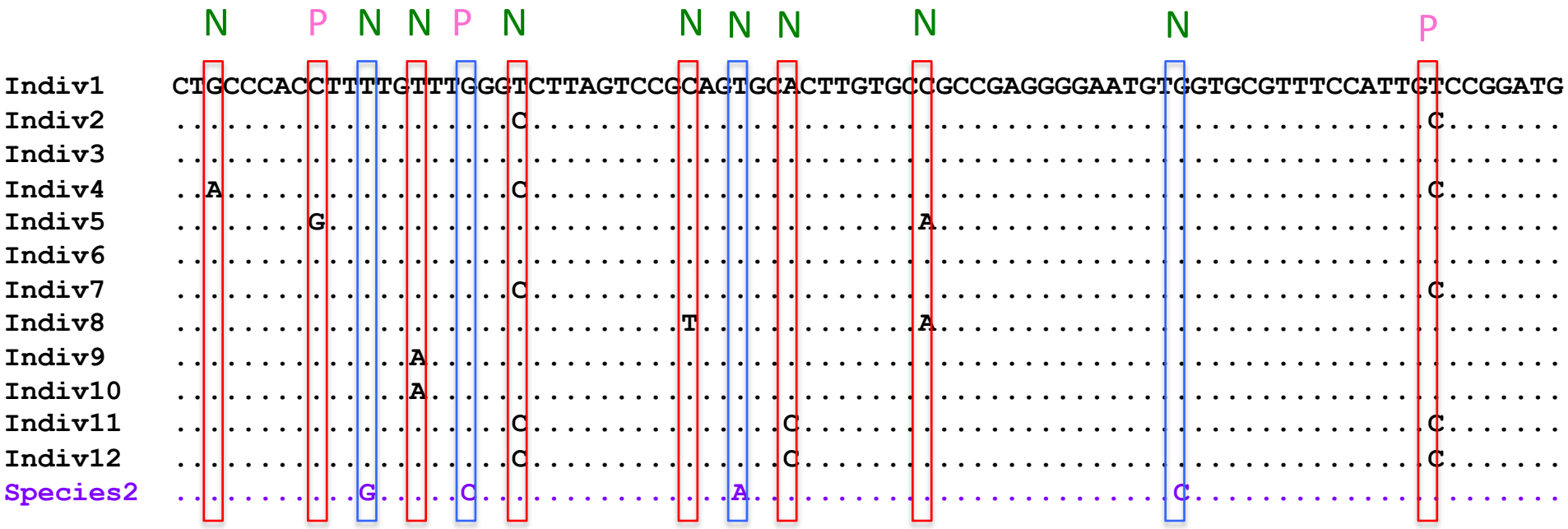
Divergent

Neutral

Potentially Non-neutral

	Poly	Div
Neutral	6	3
Potentially Non-Neutral	2	1

What if we have more data?



Polymorphic

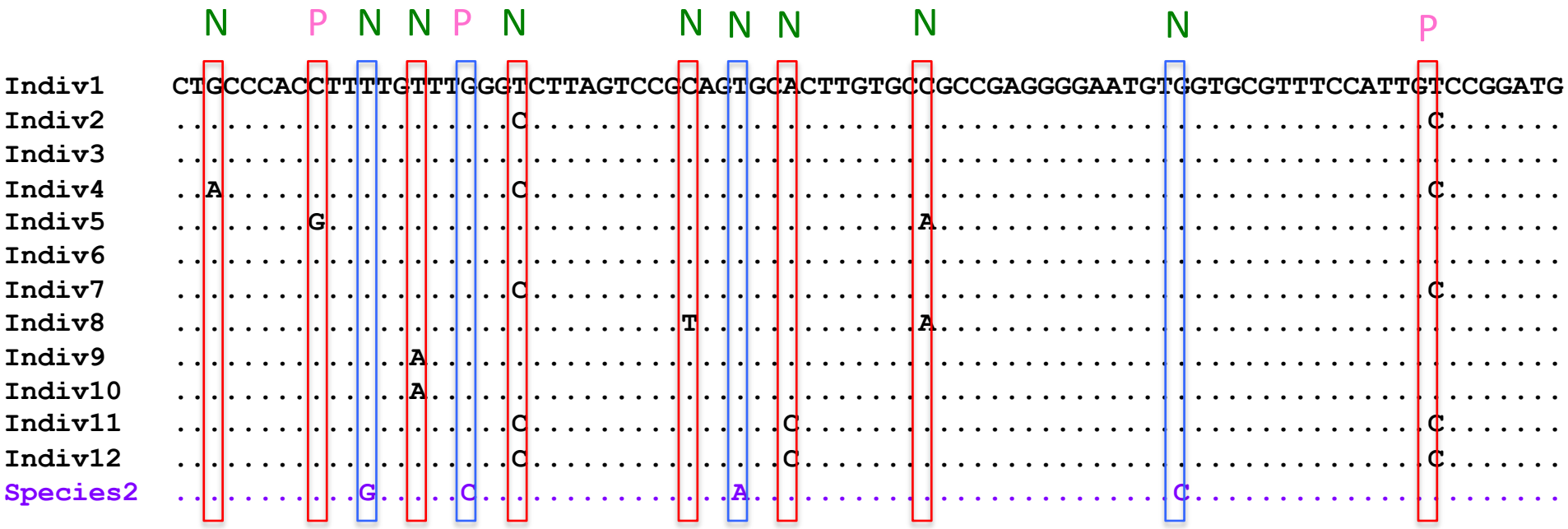
Divergent

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

	Poly	Div
Neutral	6	3
Potentially Non-Neutral	2	1

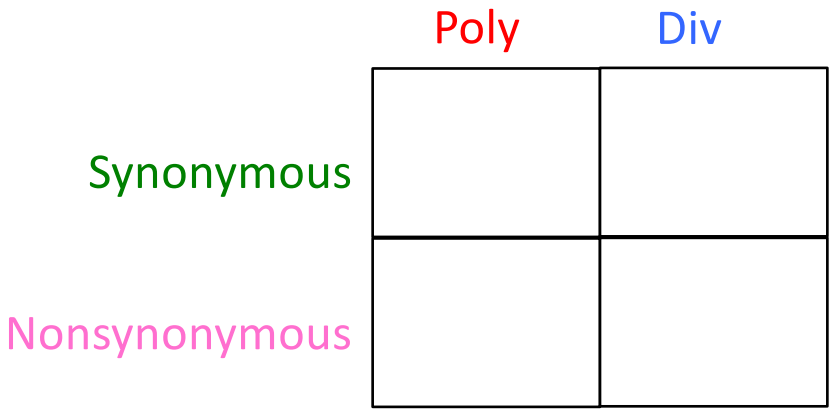
Test for statistical significance using FET

McDonald-Kreitman Test (1991)

Orginally

- -Neutral = synonymous sites
- -Potentially non-neutral = nonsynonymous sites

McDonald-Kreitman Test (1991)



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

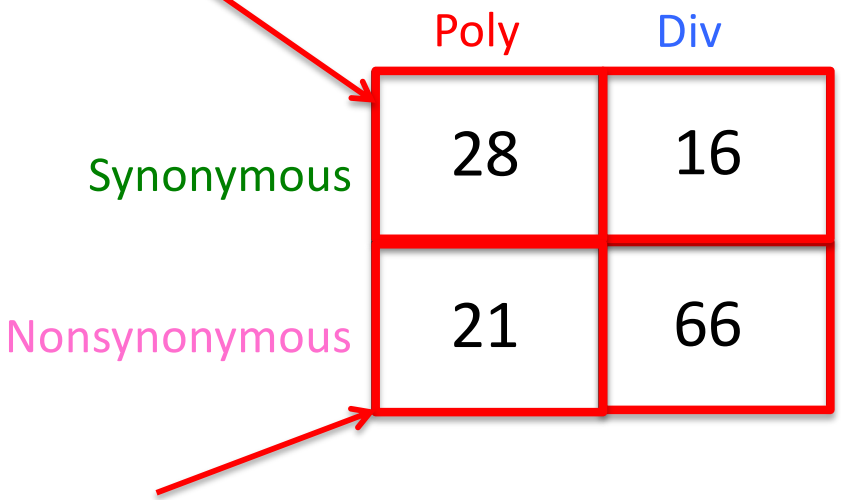
Excess fixed replacements

$P < 0.001$, FET

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

Can we reject...

Too much synonymous polymorphism?



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  CTGCCCACCTTTTGTTTGGGTCTTAGTCCGCAGTGCACTTGTGCCGCCGAGGGGAATGTGGTGCGTTTCCATTGTCCGGATG
Species2  ....................C.....................................................C.......
Species3  ..........A.....................................C..........G.................A
Species4  ..A................C.....................................................C.......
Species5  ........G.....................................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 = μ

$K_A < K_s$ 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	**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

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

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)

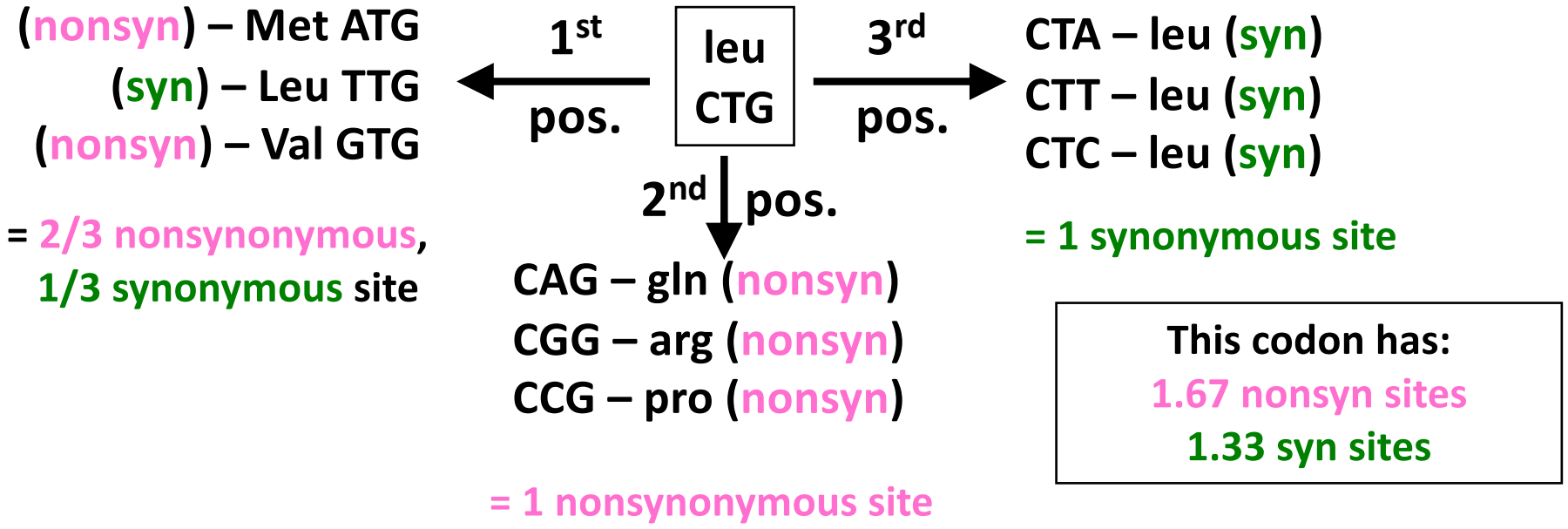
2. Count up the number of synonymous and nonsynonymous sites

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



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 = # nonsynonymous substitutions/ # nonsynonymous sites

K_S = # synonymous substitutions/ 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