Population Genetics:
Estimating diversity & selection tests
Introductory

– Part 1 –

Neutral mechanisms of evolution

– Part 2 –

Theoretical foundations of population genetics

– Part 3 –

Estimating diversity

– Part 4 –

Detecting selection

– Part 1 –

Neutral mechanisms of evolution

Neutral mechanisms - mutations

Source of variation – mutations

- Single nucleotide substitutions (SNP)
- Insertions & deletions (INDEL)
- Large-scale rearrangements – duplications, translocations

ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT

Neutral mechanisms - mutations

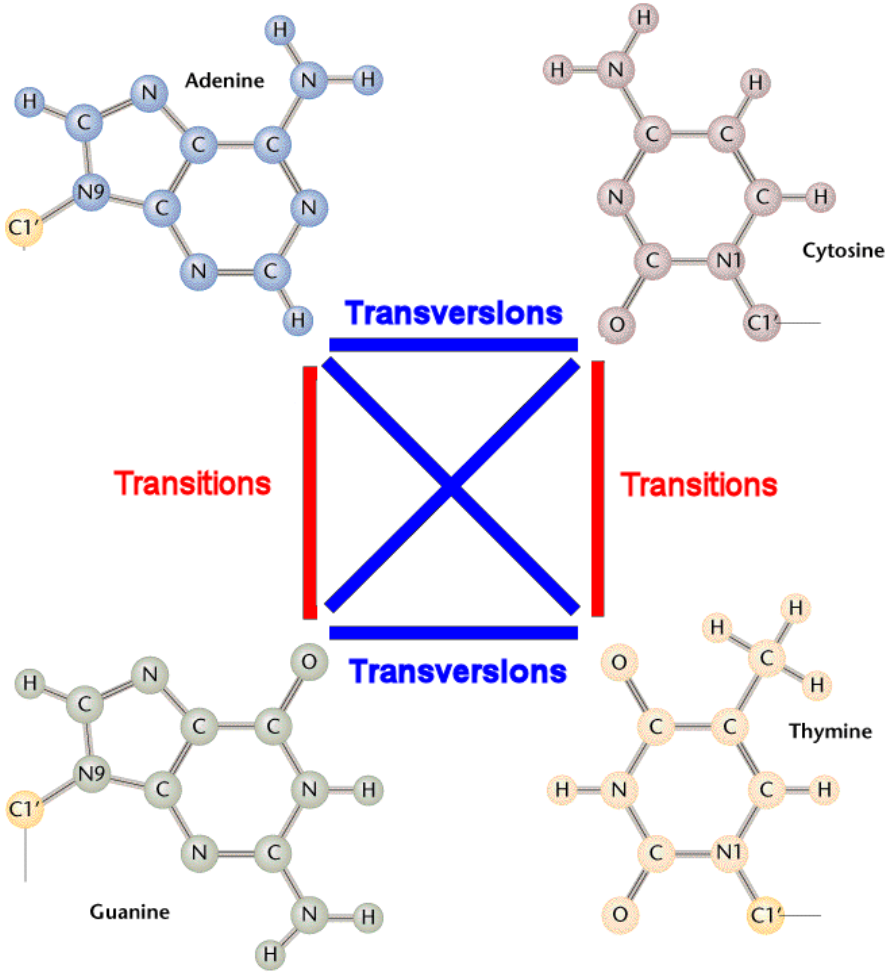
Source of variation – mutations

- Single nucleotide substitutions (SNP)
- Insertions & deletions (INDEL)
- Large-scale rearrangements – duplications, translocations

ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCG**G**ACGTCGATCGCT

Mutations – transitions & transversions

ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCGTACGTCGATCGCT
ATTCGCTGTCCG**G**ACGTCGATCGCT



Mutations – effect on protein sequence

Non-synonymous

I R C P Y V D R

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG **G**AC GTC GAT CGC

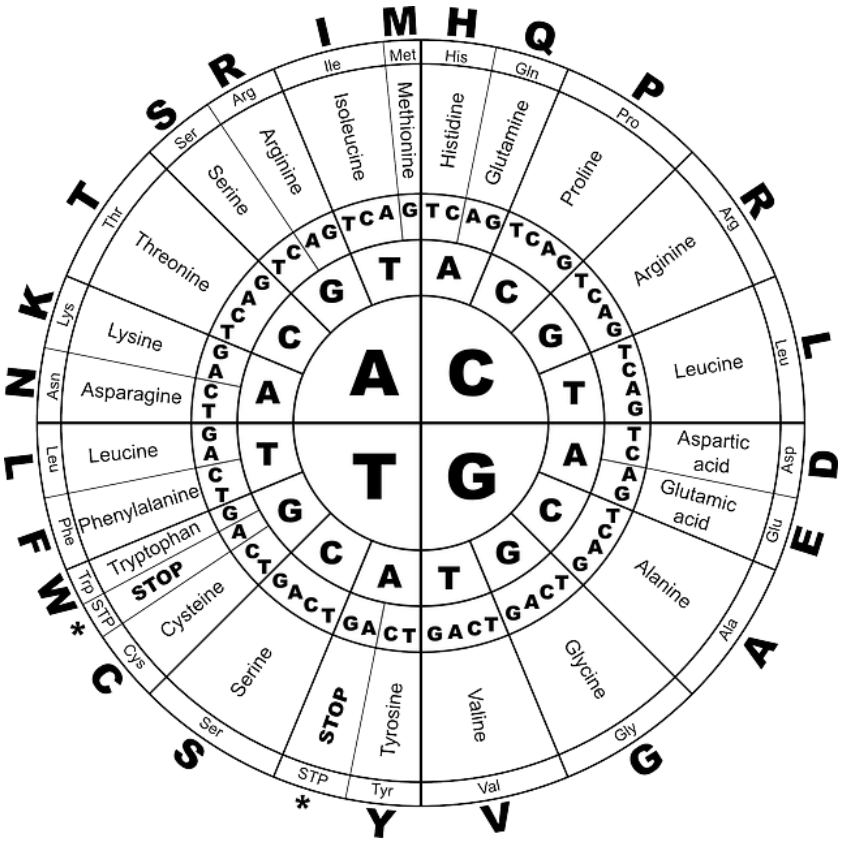
I R C P **D** V D R

Synonymous

I R C P Y V D R

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TA**T** GTC GAT CGC

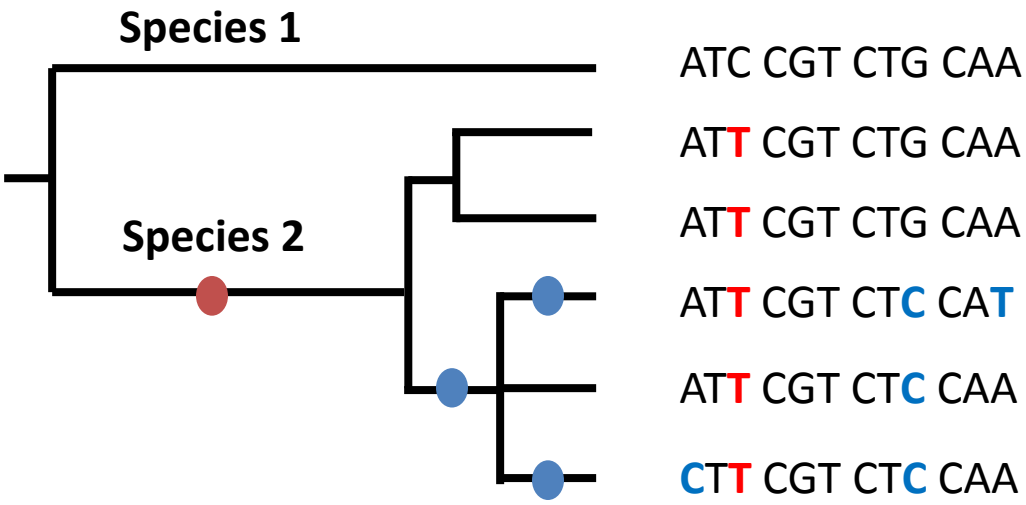
I R C P Y V D R



Mutations – effect on fitness / reproductive success

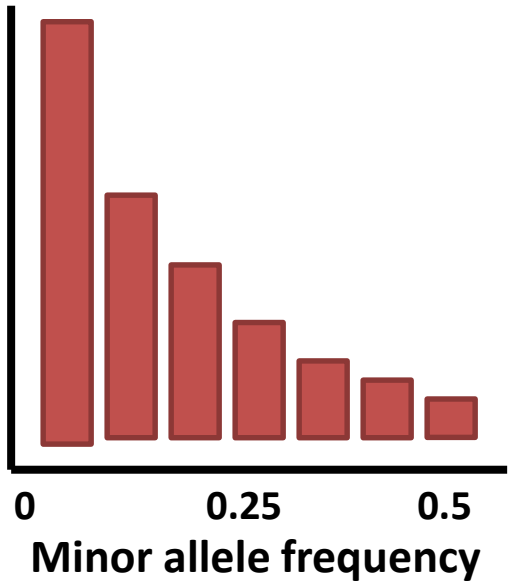
- **Neutral**
- **Deleterious**
- **Advantageous**

Mutations – polymorphisms & substitutions



Mutations – minor allele frequency

Folded site frequency spectrum




```
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG GAC GTC GAT CGC
ATA CGC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
```

Mutations – allele frequency

Species1 ATT C**A**C TGT CCG TAC GTC GAT CGC
Species2 ATT C**A**C TGT CCG TAC GTC GAT CGC

ATT C**G**C TGT CCG TAC GTC GAT CGC
ATT C**G**C TGT CCG TAC GTC GAT CGC
ATT C**G**C TGT CCG TAC GTC GAT CGC
ATT C**G**C TGT CCG **G**AC GTC GAT CGC
AT**A** C**G**C TGT CCG **G**AC GTC GAT CG**A**
AT**A** CAC TGT CCG **G**AC GTC G**C**T CG**A**

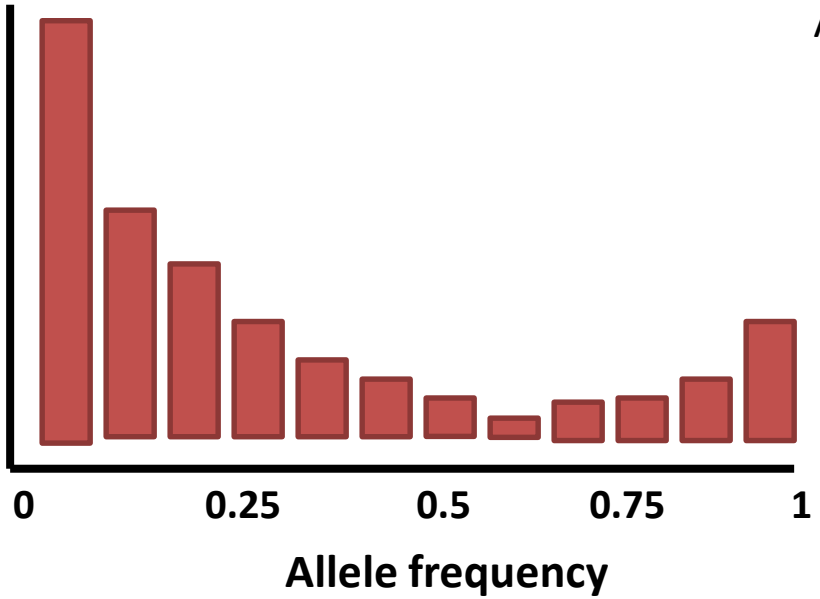
ancestral

derived

0.83 0.5 0.17

high medium low

Unfolded site frequency spectrum (SFS)



Mutations - take home

- **Mutations come in various flavors**

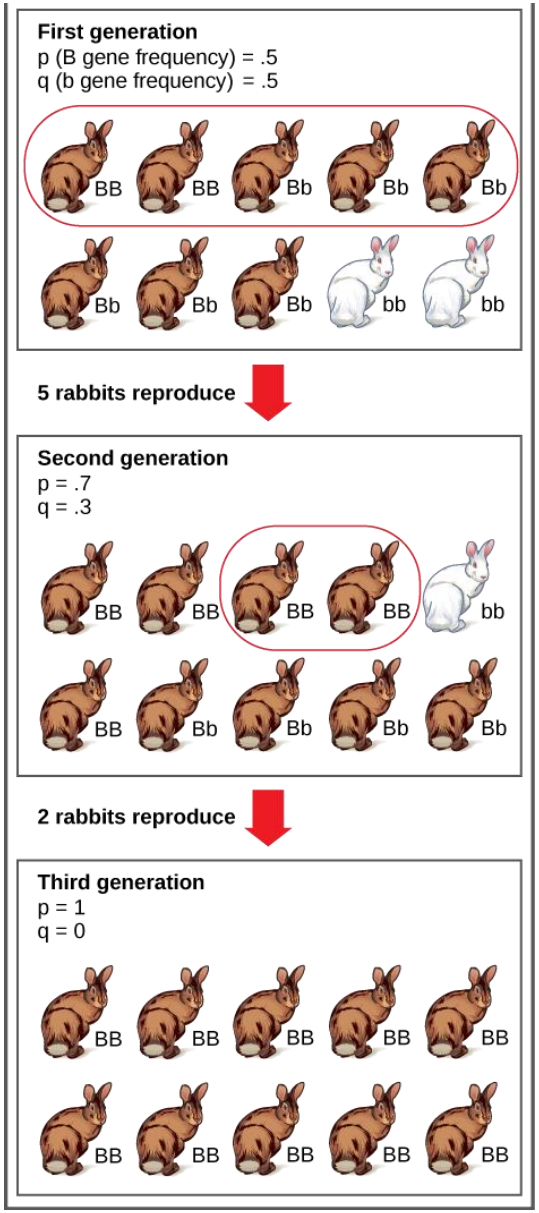
- **Population geneticists try to understand processes that affect allele frequencies in populations.**

Neutral mechanisms – Genetic drift

- Genetic drift is change in allele frequencies in a population from generation to generation that occurs due to chance events (sampling error).

- It's a stochastic process.

Even if we know everything about a population and its biology, we cannot predict the state of the population in the future.



Drift - Exercise 1.

R script - drift_neutral.R

Let's start with 50 alleles / mutations at frequency of 0.5.

What happens to those alleles in a population of 100 individuals after 500 generations due to genetic drift?

What happens in a much larger population?

Repeat several times.

Drift - Exercise 1.

R script - drift_neutral.R

Let's start with 50 alleles / mutations at frequency of 0.5.

What happens to those alleles in a population of 100 individuals after 500 generations due to genetic drift?

What happens in much larger populations?

Take home: Effect of genetic drift depends on population size – strong in small and weak in large populations.

Drift - Exercise 2.

What happens to a single mutation in a population of 1000 individuals, which occurred only in 1 individual, after 500 generations due to genetic drift?

Repeat several times.

Let's try 100 mutations at frequency 1/individual. What happens to most them after?

Drift - Exercise 2.

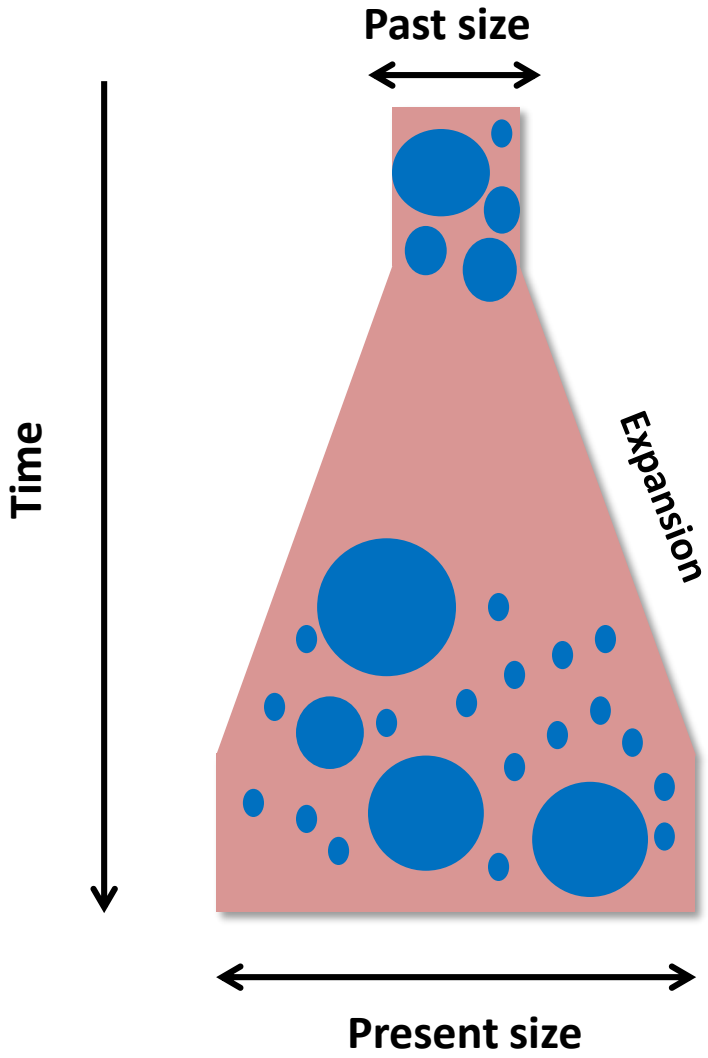
What happens to a single mutation in a population of 1000 individuals, which occurred only in 1 individual, after 500 generations due to genetic drift?

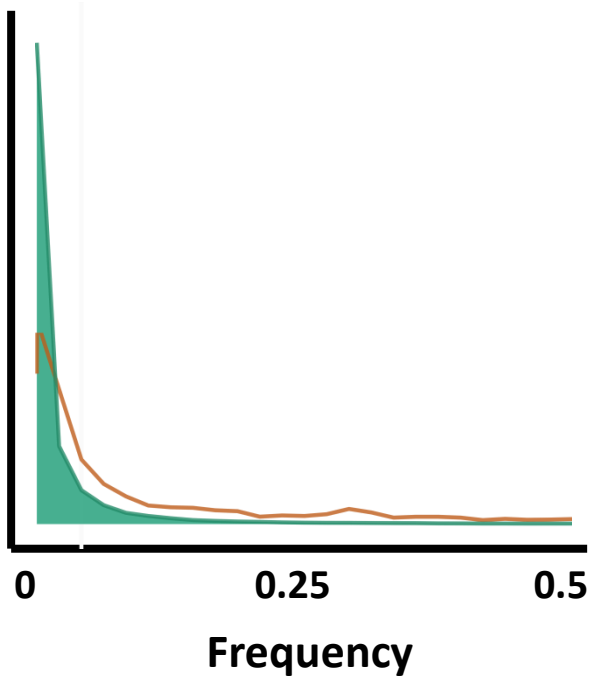
Let's try 100 mutations at frequency 1/individual. What happens to most them after?

Take home:

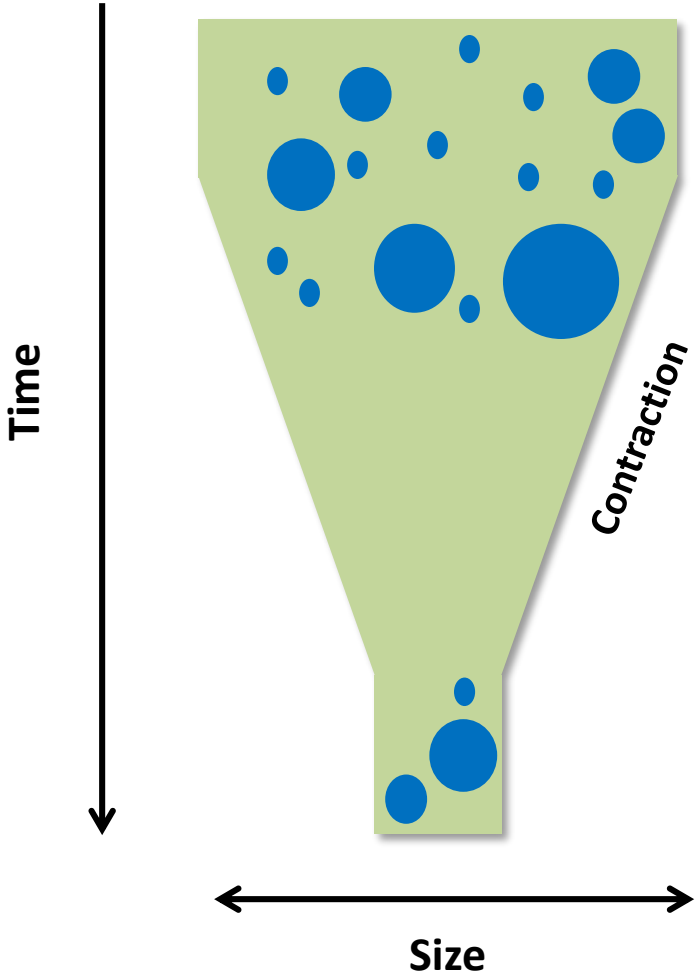
- **Most novel mutations are lost just due to the drift.**
- **It takes a lot of mutations and time for neutral mutations to reach substantial frequencies.**
- **Intermediate frequency neutral mutations are usually much older than rare mutations.**

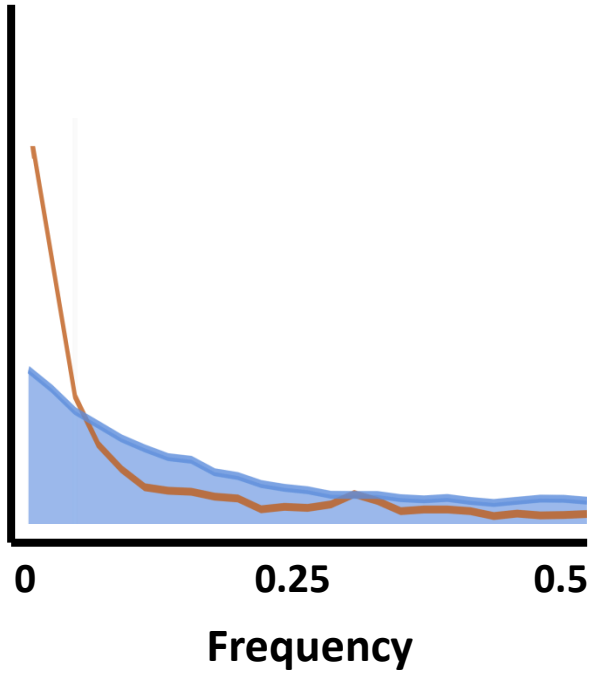
Demography – population growth



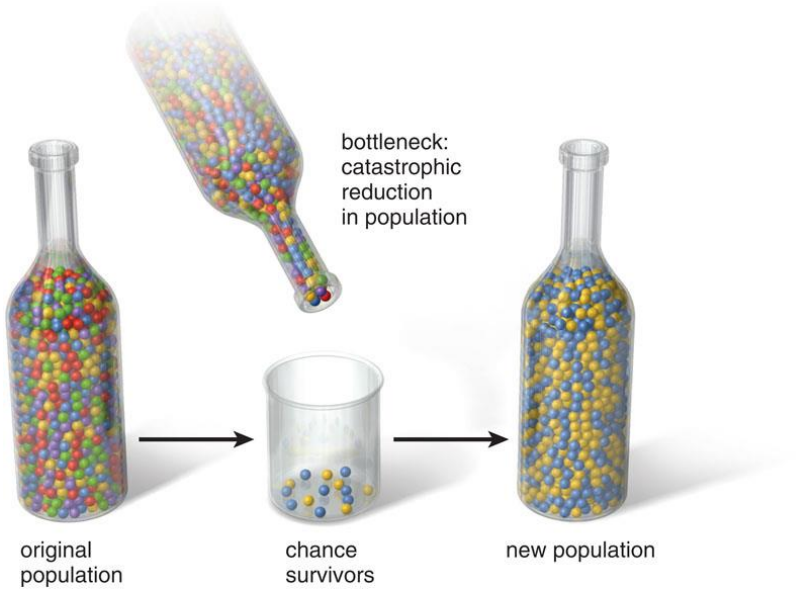


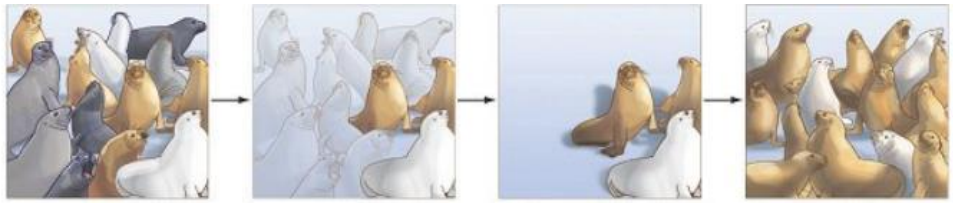
Demography – population contraction





Demography – bottleneck



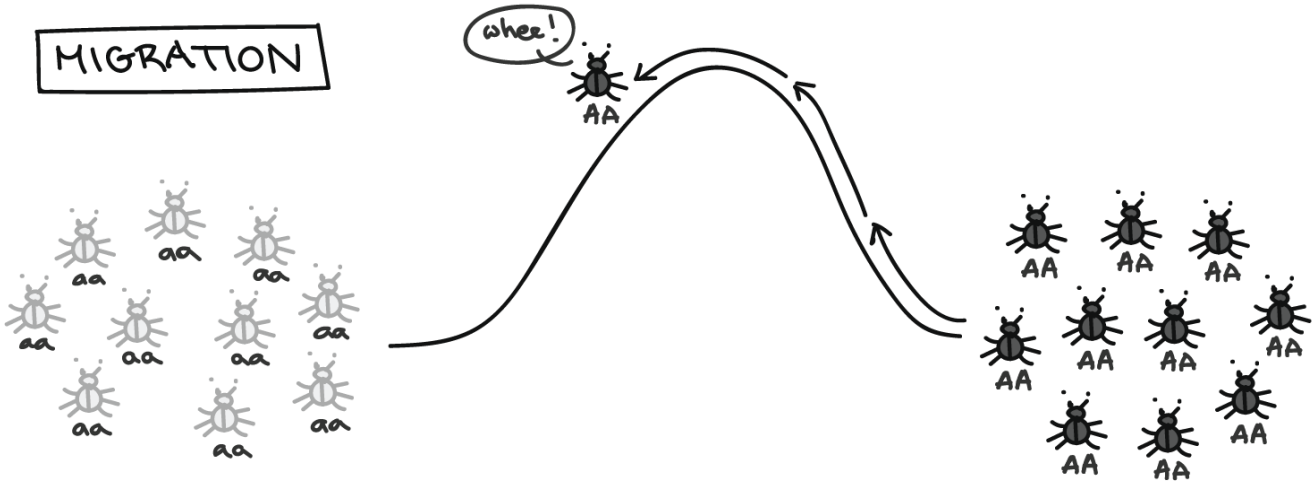


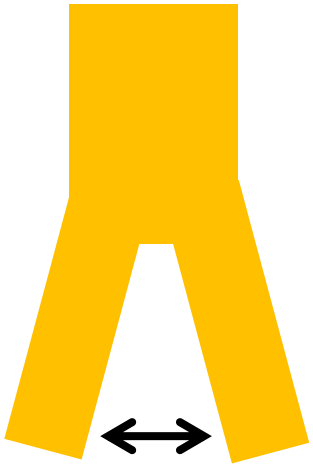
©Arie Zackay

- **Domestication**
- **Post-glacial expansion**

Neutral mechanisms – migration

a.k.a. gene flow





Makes populations genetically more similar

Demography & migration - take home

- **Historical demographic events and migration strongly affect allele frequency distribution.**
- **They often mimic signatures of selection.**
- **If unknown or unaccounted for may lead to wrong conclusions.**

– Part 2 –

Theoretical foundations of population genetics

Hardy-Weinberg Equilibrium Model (1908)

Godfrey Hardy

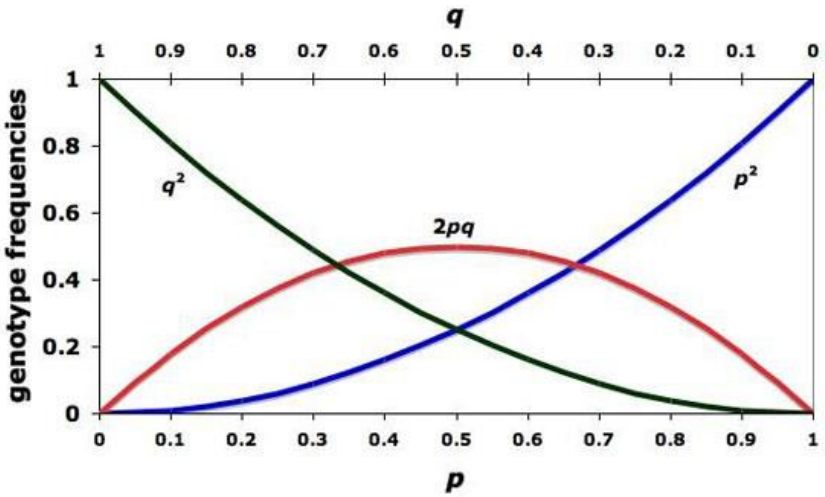
Idealized population

- Infinite size (**no drift**)
- Random mating
- Non-overlapping generations
- No sex-ratio bias
- No selection
- No mutation
- No migration
- No recombination

Wilhelm Weinberg

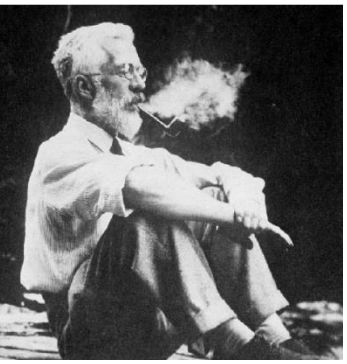
$$p + q = 1$$

$$p^2 + 2pq + q^2 = 1$$



Within a large population the **allele frequencies remain constant** from one generation to the next unless the equilibrium is disturbed by migration, genetic mutations, or selection – **population in equilibrium**

Wright-Fisher Neutral Model (1930 - 1931)

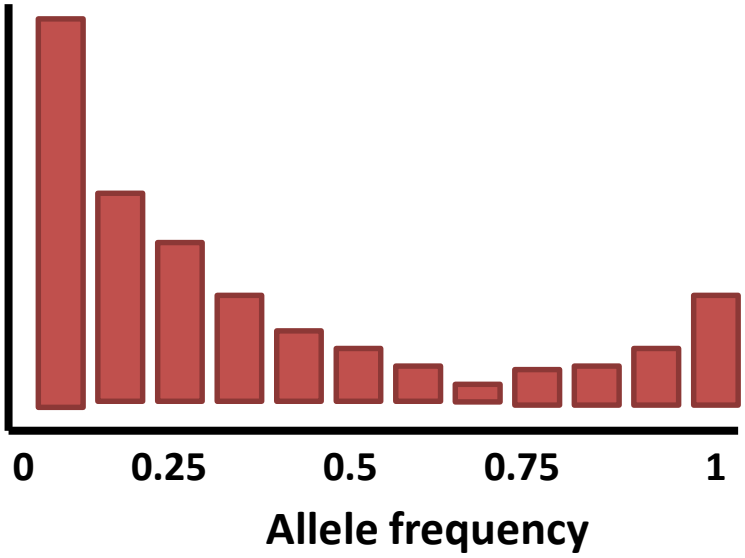


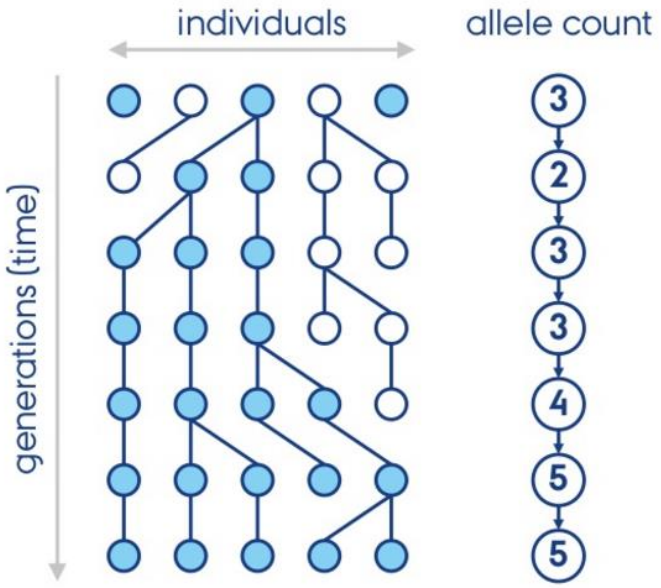
Ronald Fisher

Evolution of allele frequency forward in time at a bi-allelic locus

- Finite & constant size
- Random mating
- …

Sewall Wright





Census & effective population size

N – census population size, number of individuals

Ne – effective population size

Size of a Wright-Fisher population in which genetic drift occurs at the levels observed in an actual population

Theta – population mutation rate

$$\Theta = 4N_e\mu$$

Θ - population mutation rate, diversity
N_e - effective population size
μ - mutation rate

Theoretical models- take home

- **The theoretical models (HW, WF) describe the behaviour of allele frequencies in neutral idealized populations**

- **Though such populations don't exist in nature, the models provide reference points and mathematical framework of population genetics**

– Part 3 –

Estimating diversity

Which of the populations is more diverse?

Case 1

GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC

n = 12

Case 2

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC

n = 6

Number of segregating sites, S

Case 1

```
GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC
```

n = 12

Case 2

```
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
```

n = 6

Haplotypes

Case 1

GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC

n = 12
S = 8

Case 2

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC

n = 6
S = 4

Watterson's estimator Θw (Θs)

Case 1

GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC

n = 12
S = 8

Case 2

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC

n = 6
S = 4

$$\Theta w = \frac{S}{\sum_{i=2}^{n} \frac{1}{i-1}}$$

Could be normalized for the length of the analysed genomic region including invariant sites

Nei's nucleotide diversity (Θπ or π) (1979)

Masatoshi Nei

The average number of pairwise differences per sequence in the sample

$$\theta_\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} S_i \, i(n-i)$$

Could be normalized for the length of the analysed genomic region including invariant sites

Nei's nucleotide diversity (Θπ or π)

Case 1

GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC

n = 12
S = 8

Case 2

ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC

n = 6
S = 4

$$\frac{Sum\ of\ pairwise\ differences}{Number\ of\ comparisons}$$

Estimating diversity - take home

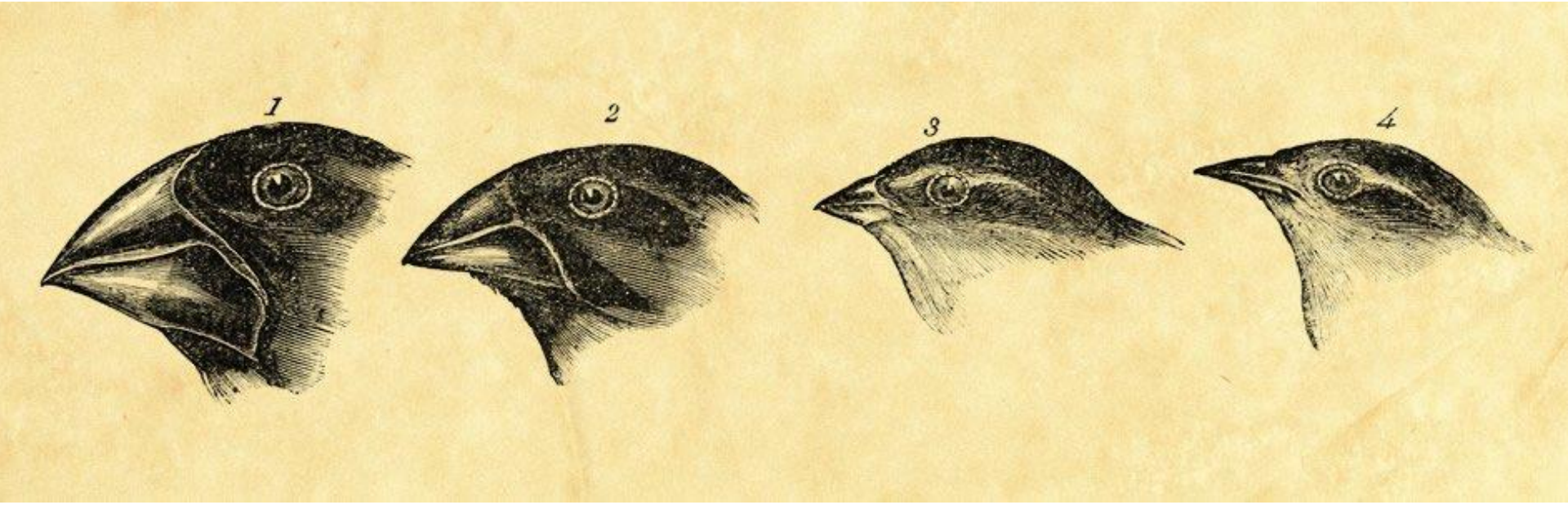
- **Various parameters are used to describe diversity**

- **Using Thetas, populations differing in size and distributions of allele frequencies can be compared**

– Part 4 –

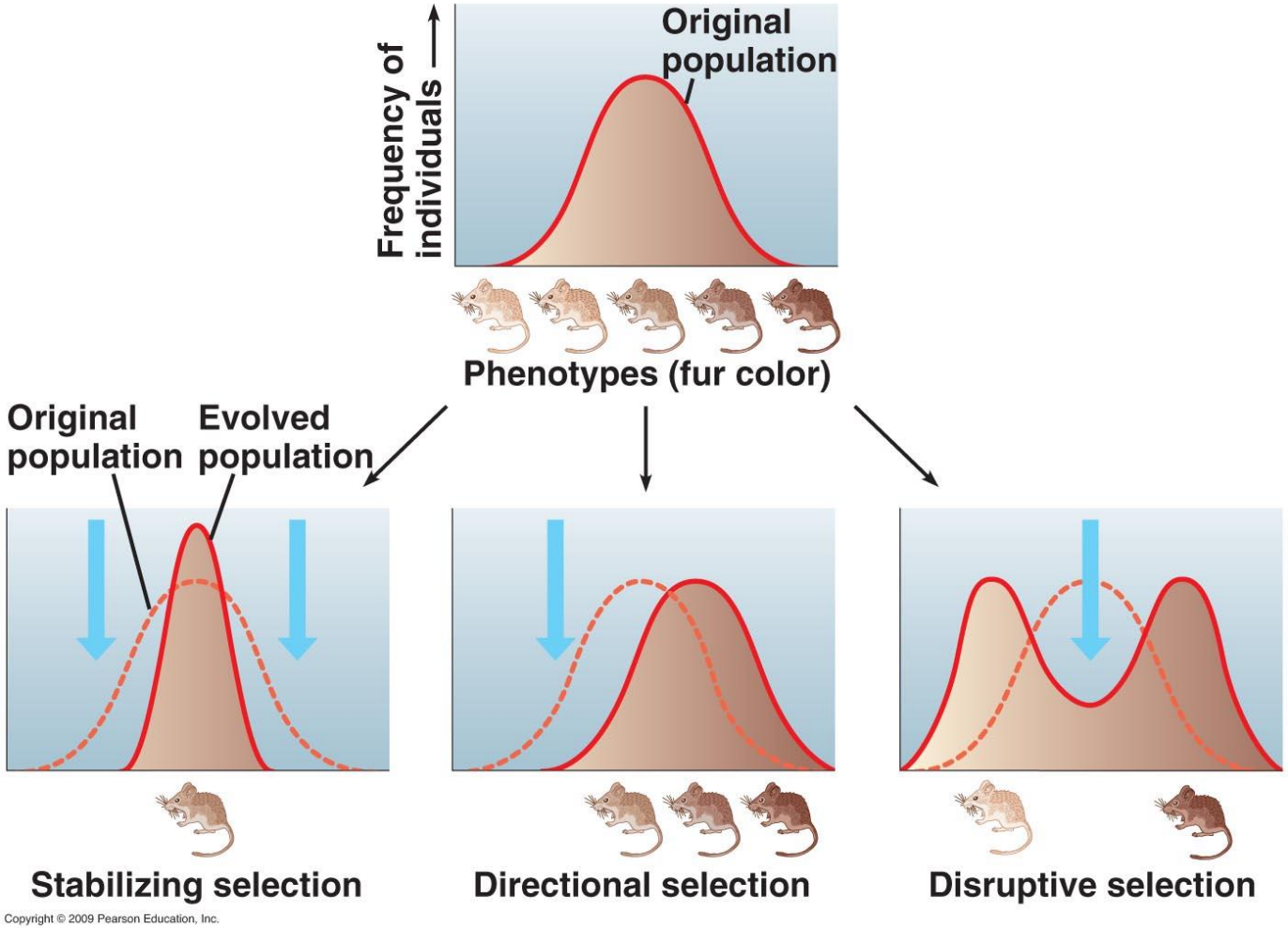
Detecting selection

What is natural selection?



Natural selection is the process by which heritable traits become either more or less common in a population as a result of their impact on reproductive success.

Selection effects on phenotype



Selection effects on genotype

Positive (Darwinian) selection

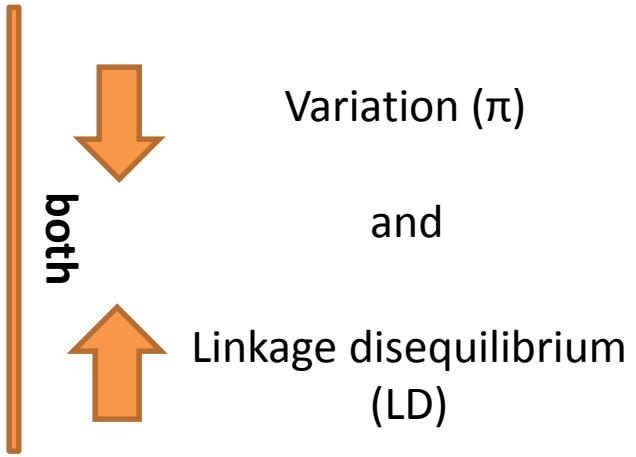
- increases frequencies of advantageous alleles

Model: **"Selective sweep"**

Negative (purifying) selection

- decrease frequencies of detrimental alleles

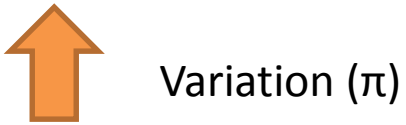
Model: **"Background selection"** (BGS)



Balancing selection

- any kind of selection that maintains 2 or more alleles in a population

	anemia	**malaria**
HbA HbA	none	severe
HbA HbS	mild	less severe
HbS HbS	severe	?



Selection effects on genotype

Positive (Darwinian) selection
- increases frequencies of advantageous alleles

Model: **"Selective sweep"**

Negative (purifying) selection
- decrease frequencies of detrimental alleles

Model: **"Background selection"** (BGS)

both

↓ Variation (π)

and

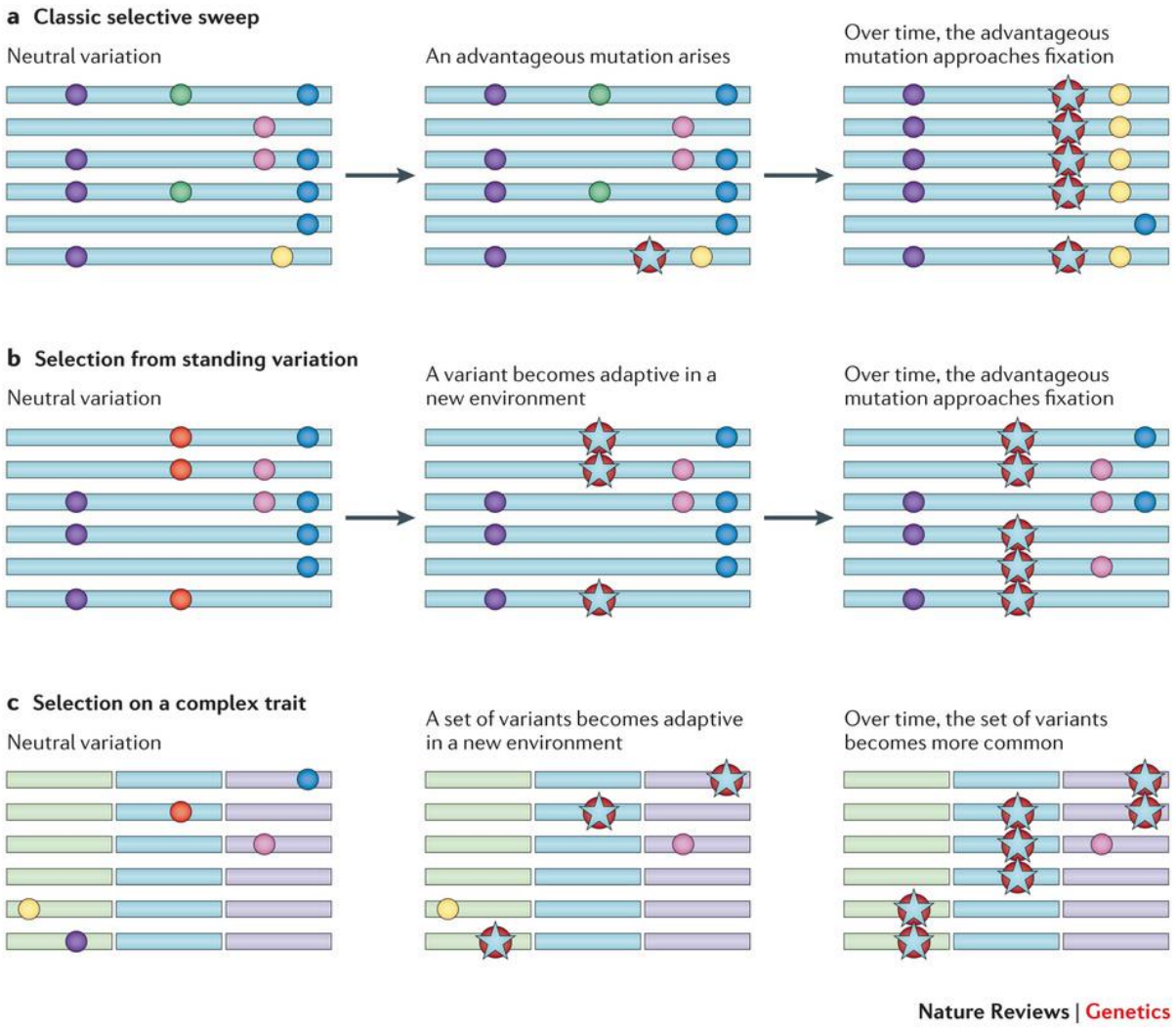
↑ Linkage disequilibrium (LD)

Balancing selection
- any kind of selection that maintains 2 or more alleles in a population

	anemia	**malaria**
HbA HbA	none	severe
HbA HbS	mild	less severe
HbS HbS	severe	?

↑ Variation (π)

Selective sweep – 4 scenarios

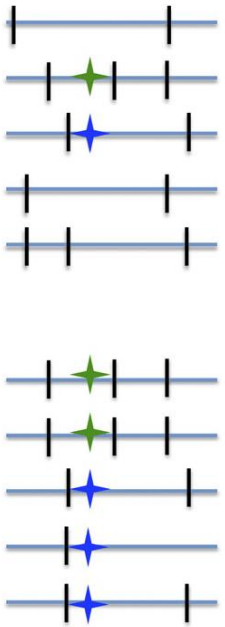


Hard

Soft

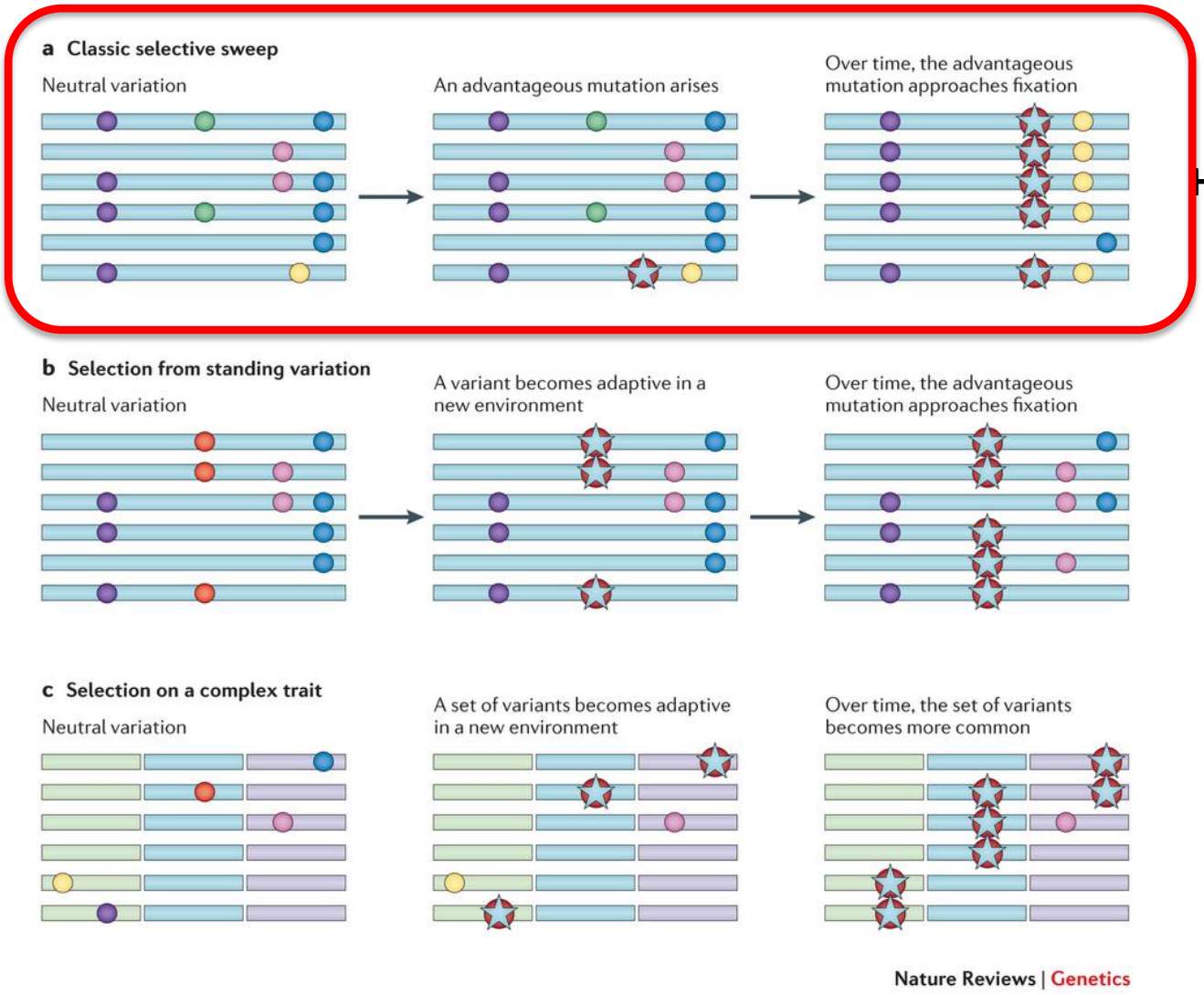
+

Selection on multiple new mutations (soft selective sweep)



Scheinfeldt & Tishkoff 2013

Selective sweep – 4 scenarios

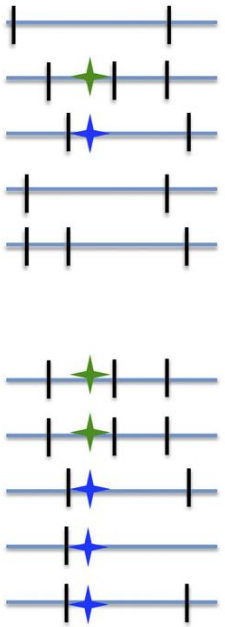


Hard

Soft

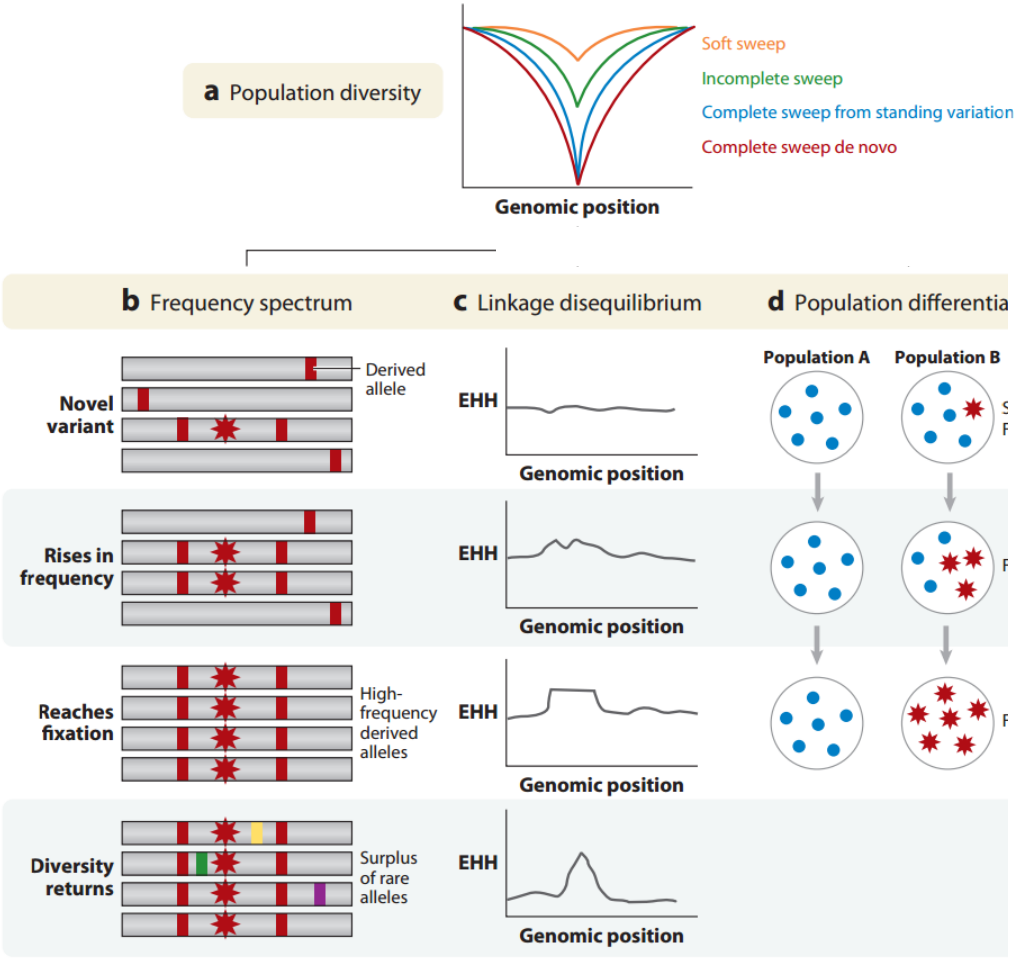
+

Selection on multiple new mutations (soft selective sweep)



Scheinfeldt & Tishkoff 2013

Signatures of a selective sweep



Vitti et al. Annual review of genetics. 2013

↓ Variation (π)

and

↑ Linkage disequilibrium (LD)

Selection tests - variety

Selection in a single population

- Deviations in allele frequencies (Tajima's D, Fay&Wu's H, Composite likelihood ratio)
- Deviation in linkage disequilibrium (OmegaPlus, Extended haplotype homozygosity EHH, iHS)

Selection between populations

- Population differentiation (Lewontin-Krakauer test, OutFLANK, PCAadapt, hapFLK, SelEstim, BayPass, LFMM, Bayenv)
- Linkage disequilibrium (XP-EHH)
- Combined LD + SFS (XP-CLR)

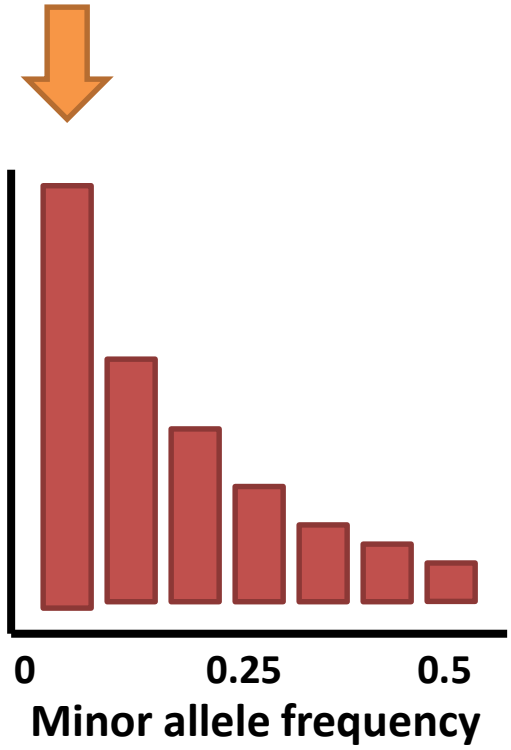
Selection between species

- Contrasting synonymous and non-synonymous sites (Ka/Ks, McDonald-Kreitman test)
- Contrasting polymorphism and divergence (Hudson-Kreitman-Aguade test)

Composite of multiple signals (CMS) test

Selection using allele frequency - Tajima's D (1989)

Reveals enrichment or depletion of rare alleles



Fumio Tajima

$$Tajima's\ D = \Theta\pi - \Theta w$$

observed

expected

Tajima's D - Exercise

Case 1

1 ATT CGC TGT CCG TAC GTC GAT CGC
2 ATT CGC TGT CCG TAC GTC GAT CGC
3 ATT CGC TGT CCG TAC GTC GAT CGC
4 ATT CGC TGT CCG **G**AC GTC GAT CGC
5 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
6 AT**A** C**A**C TGT CCG **G**AC GTC G**C**T CGC

Tajima's D - Exercise

Case 2

1 ATT CGC TGT CCG TAC GTC GAT CGC
2 ATT CGC TGT CCG TAC G**C**C GAT CGC
3 ATT CGC TGT CCG TAC GTC GAT CGC
4 ATT CGC TGT CCG TAC GTC GAT CGC
5 ATT C**C**C TGT CCG TAC GTC GAT CGC
6 AT**A** CGC TGT CCG TAC GTC G**C**T CGC

Tajima's D - Exercise

Case 3

1 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
2 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
3 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
4 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
5 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
6 AT**A** C**A**C TGT CCG **G**AC GTC G**C**T CGC

Tajima's D - interpretation

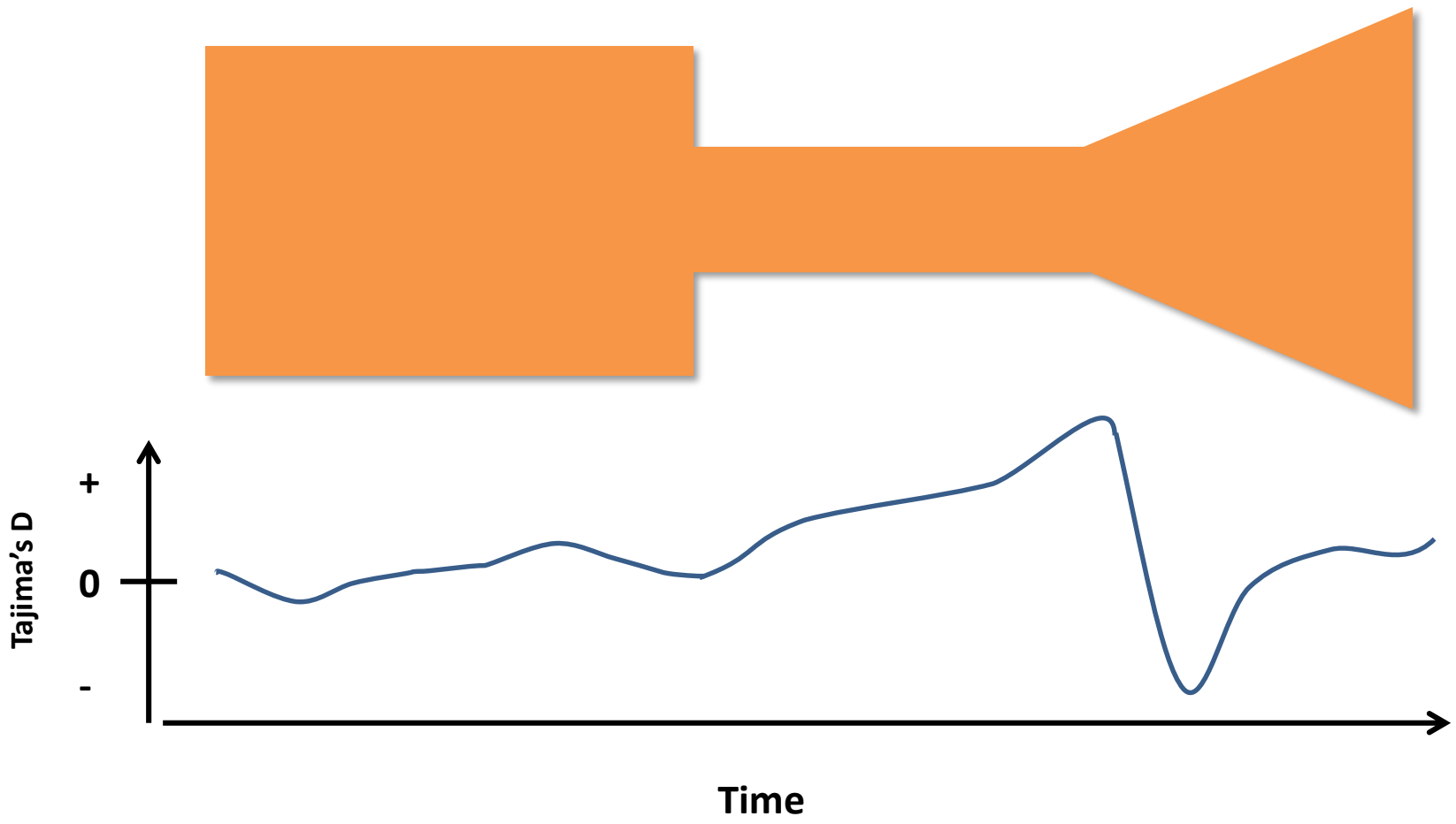
"0" – neutral evolution

"negative" – **selection removing variation** or **recent population expansion**

"positive" – **selection maintaining variation** or **recent population contraction**

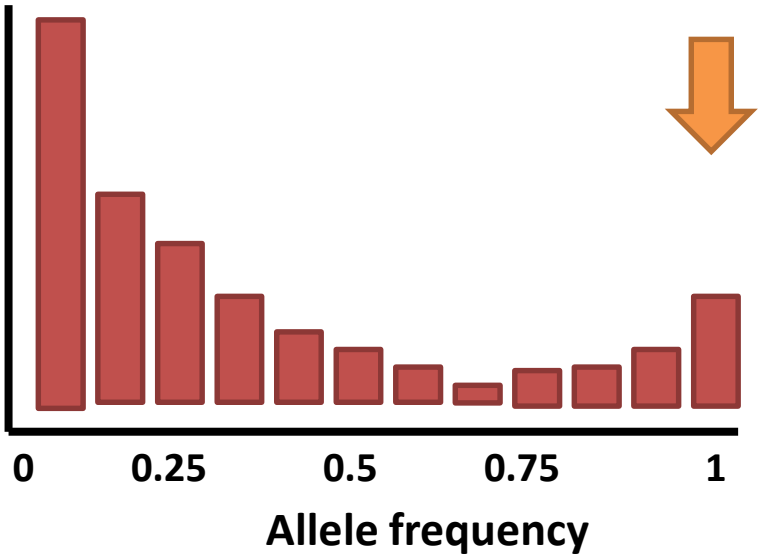
Not robust to demographic changes -> knowing demographic history is critical to detect selection.

Tajima's D in a bottleneck



Fay & Wu's H – enrichment of high frequency derived alleles

Derived (youngest) alleles reach high frequencies slowly & randomly.



"negative" – excess of high-frequency derived alleles, selective sweep

"positive" (very rare) – selection maintaining Variation

Only very recent sweeps.

May be affected by population structure.

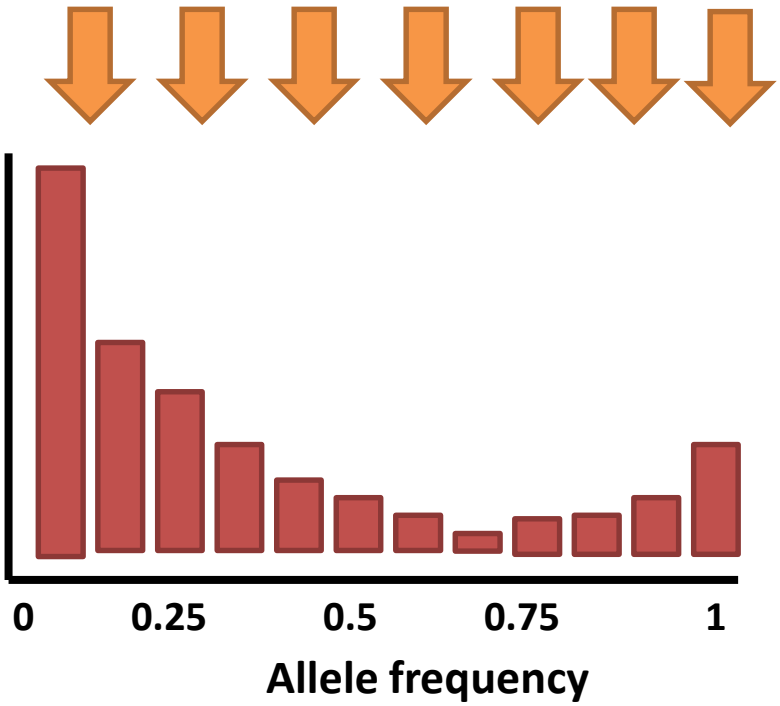
Testing full SFS – composite likelihood ratio test

Composite likelihood ratio test (CLR)

Nielsen (2005) modification uses genome-wide SFS as a null hypothesis

Software: SweepFinder & SweeD

Still may be affected by demography and SNP ascertainment bias

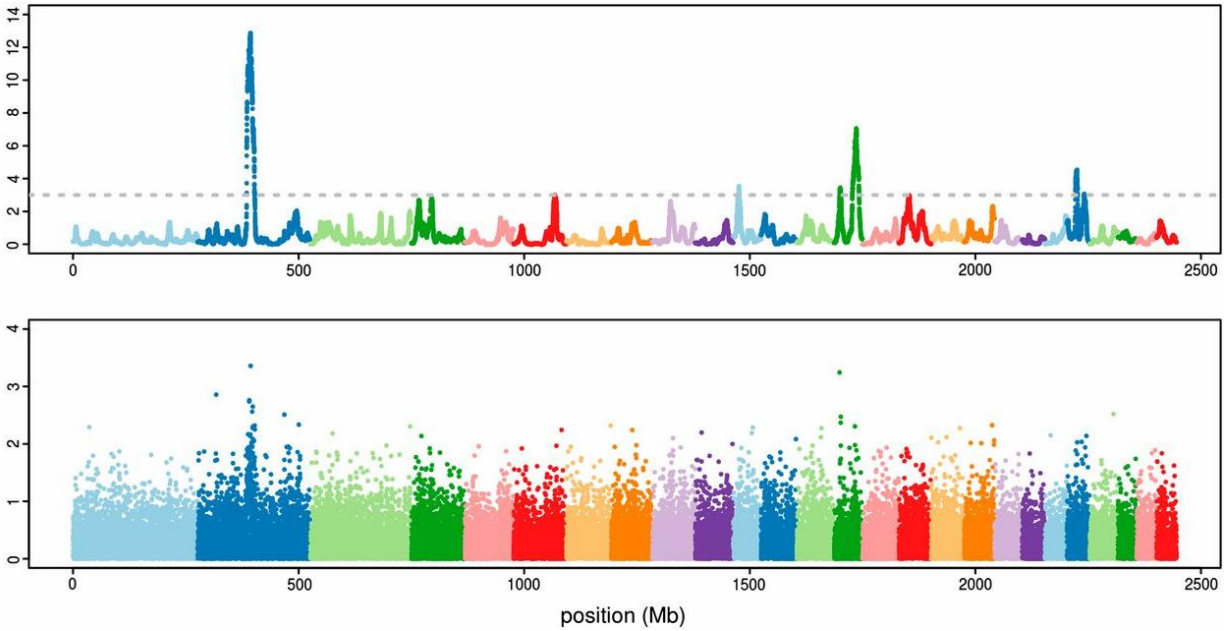


Genome scans – detecting outliers

Genome scans – calculating diversity statistics from the genome-wide genotyping or re-sequencing data in sliding windows.

Assumption: **genome-wide variation** represents **neutral** history whereas **outliers** are instances of **selective sweep**.

Problem: how to define outliers



Genome scans – detecting outliers

Take top 0.1%, 1%, 5% of the distribution – naïve, wrong but still gets published

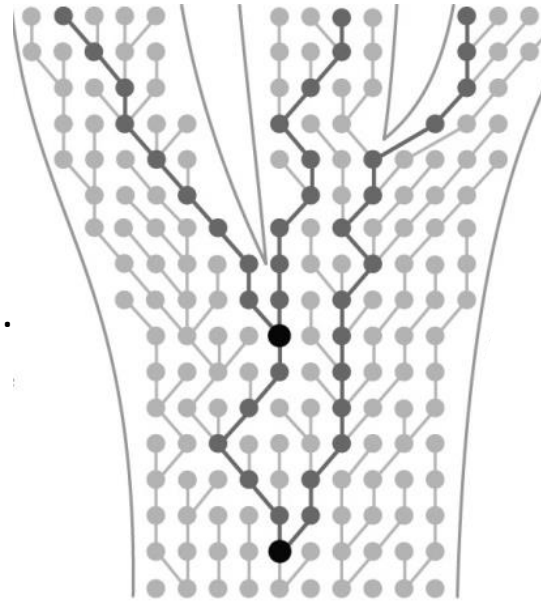
Genome scans – detecting outliers

Simulate neutral datasets with diversity parameters of your population (Θ and size)

- using WF model if statistic believed to be robust to demography (CLR)

- using estimated demographic model for your population (Tajima's D, FayWu's H)

Calculate statistic on the neutral simulated dataset (null distribution) and define thresholds for your experimental data.

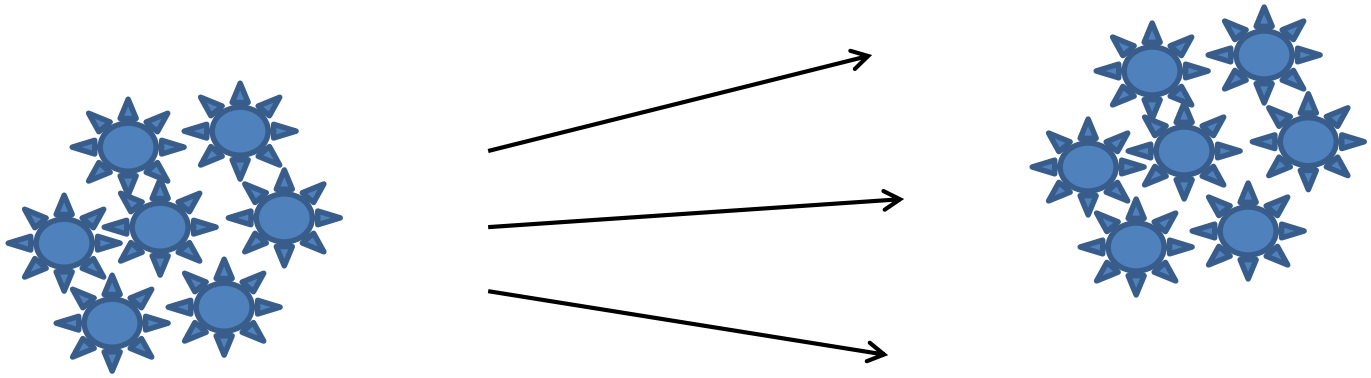


Reverse (limited but fast) and forward (versatile but slow) simulators.

Overview of simulation tools:

Hoban et al. Nature Reviews Genetics 13.2 (2012): 110-122.

Selection using population divergence

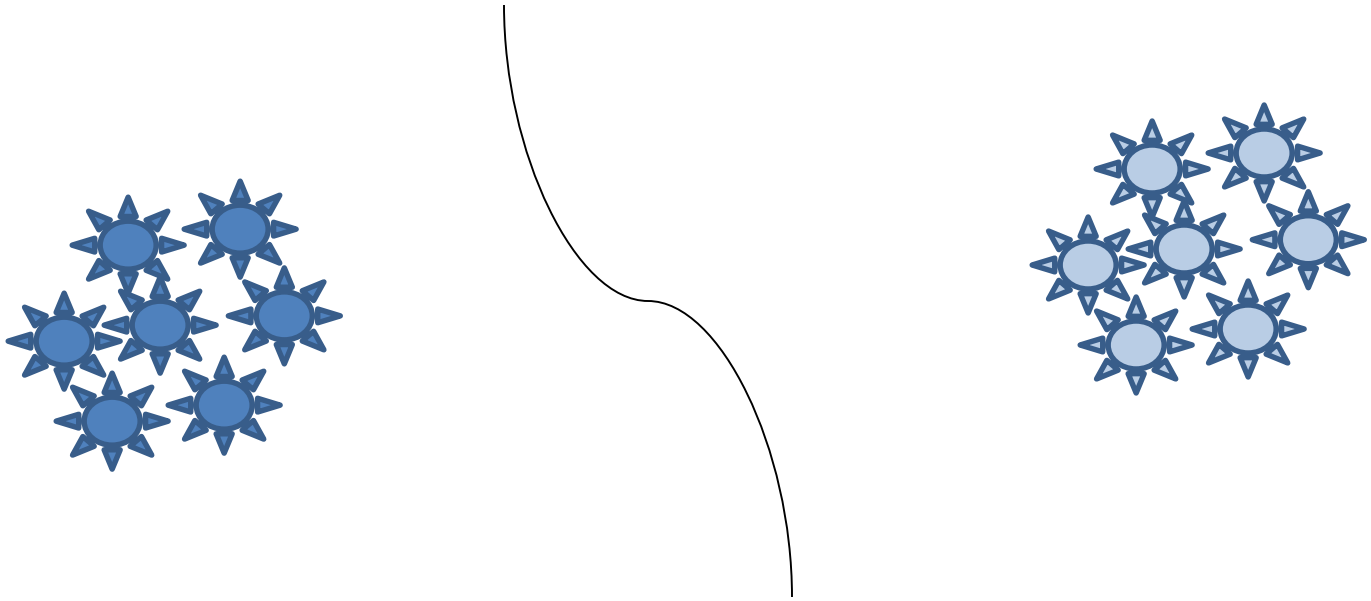


Cold

Warm

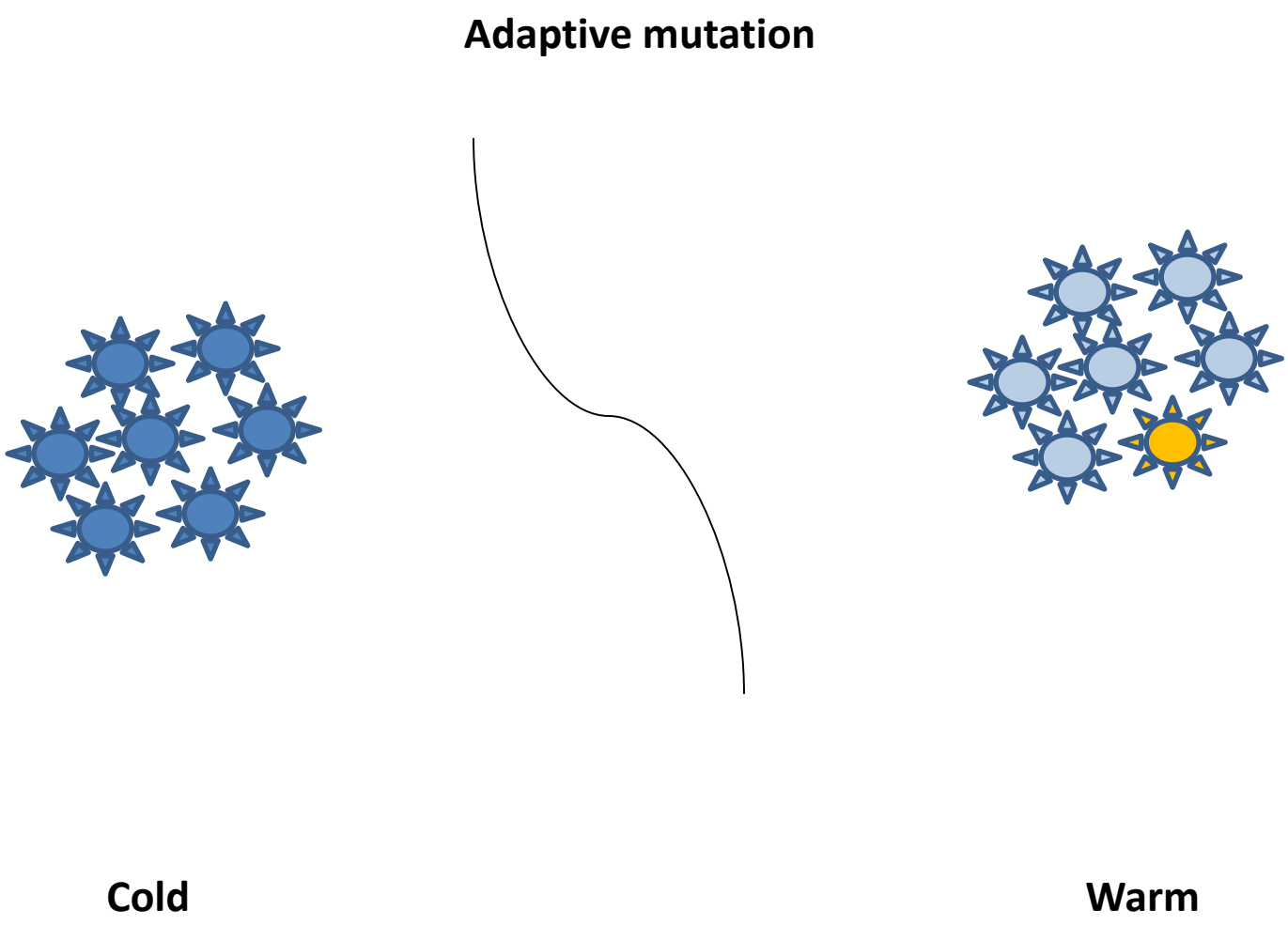
Selection using population divergence

Non adapted populations



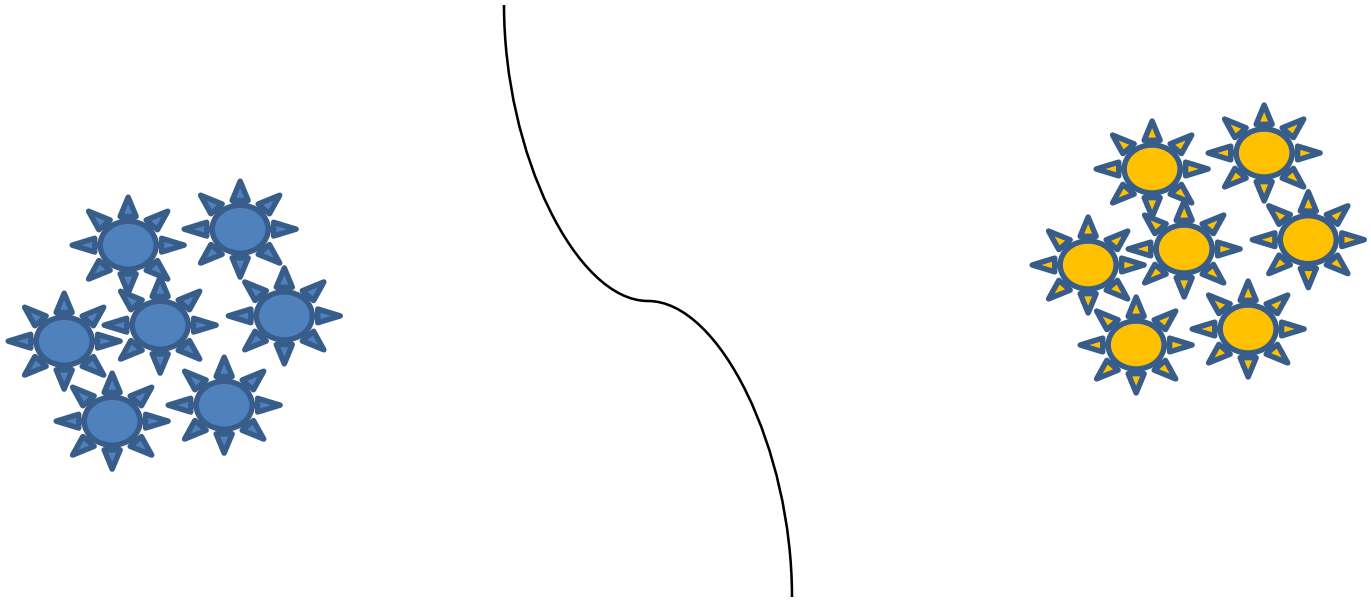
Cold **Warm**

Selection using population divergence



Selection using population divergence

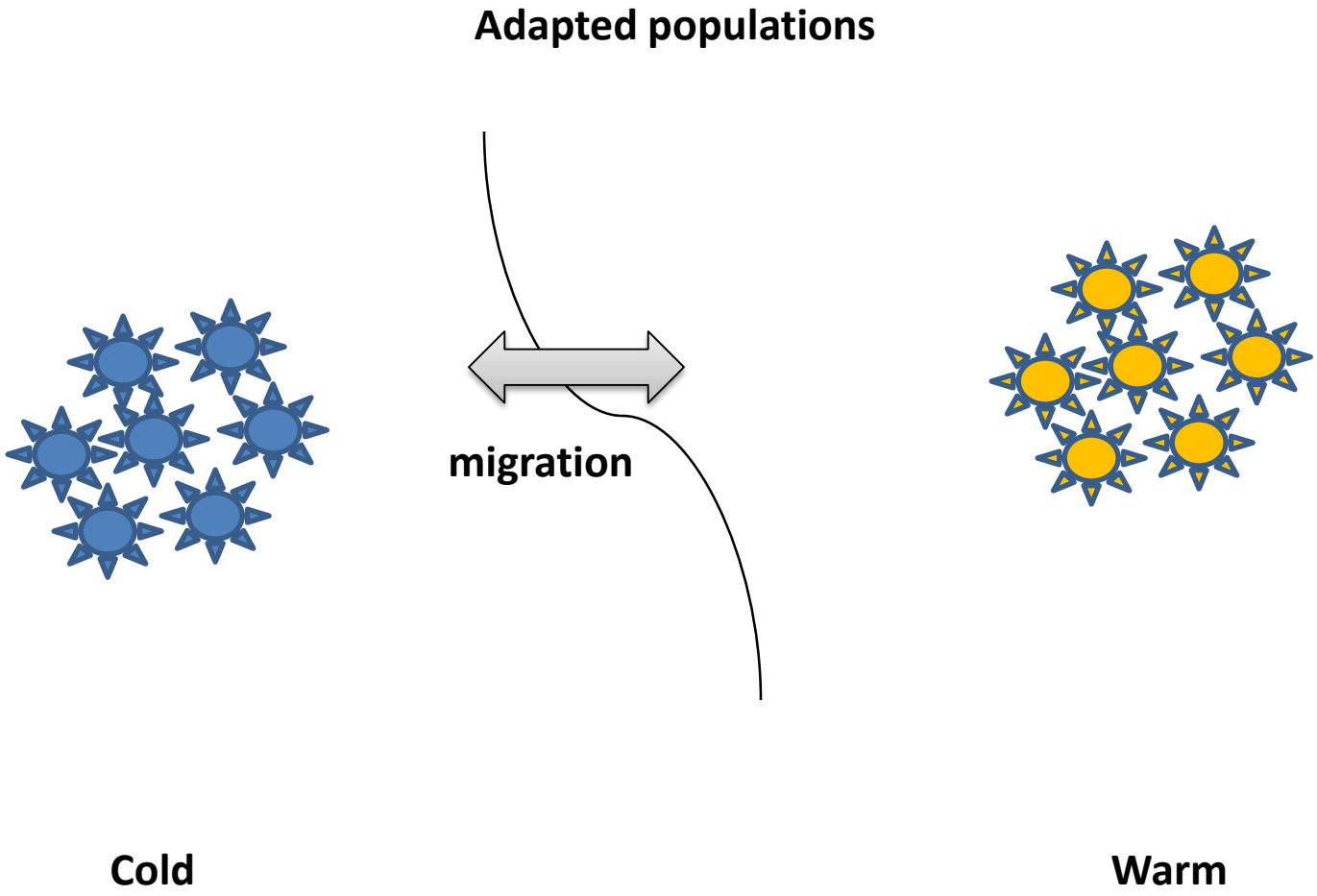
Adapted populations



Cold

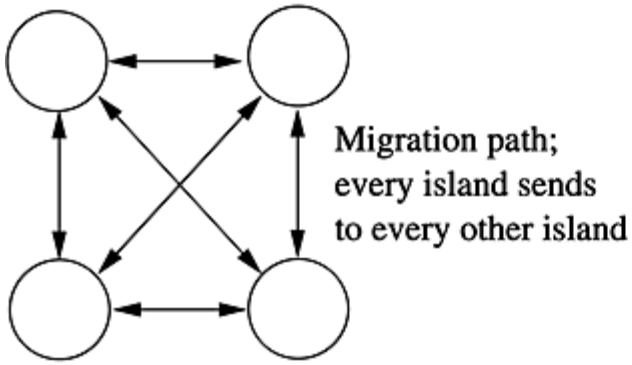
Warm

Selection using population divergence

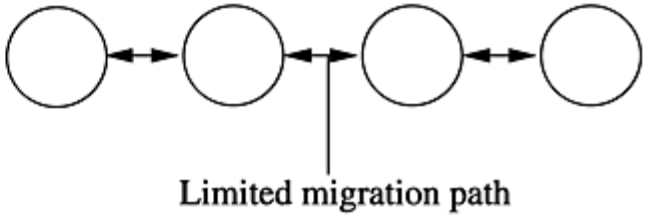


Theoretical models of population structure

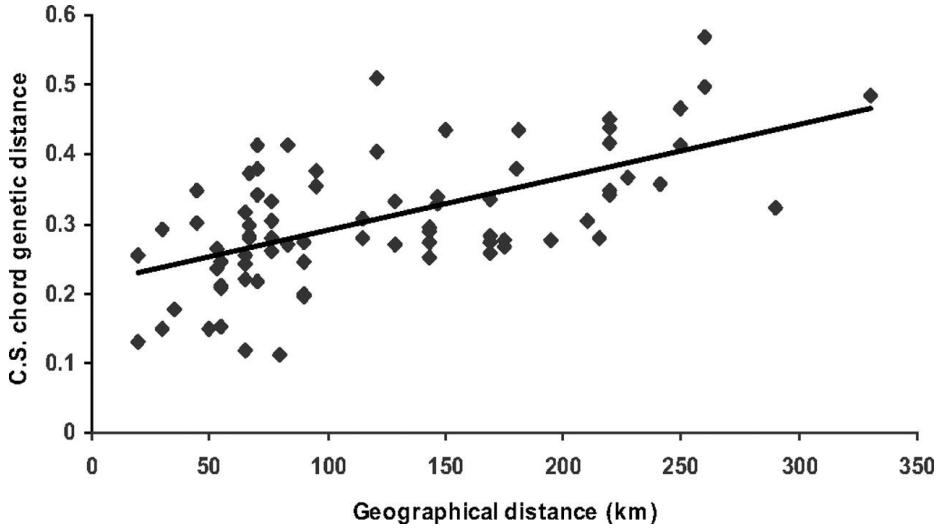
Island model



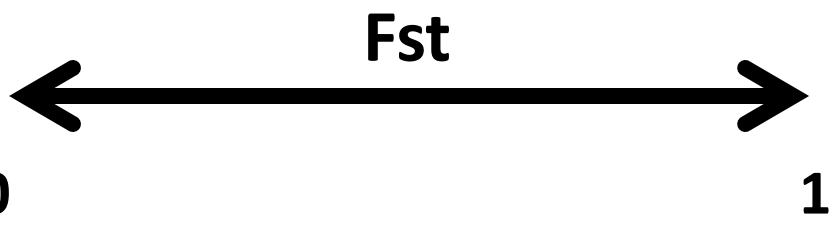
Stepping stone model



Isolation by distance



Measuring population divergence – fixation index



No difference

A A
A A
A A
T T
T T
T T

A 0.5 0.5
T 0.5 0.5

$$Fst = \frac{Ht - Hs}{Ht}$$

Ht – total heterozygosity
Hs – subpopulation heterozygosity

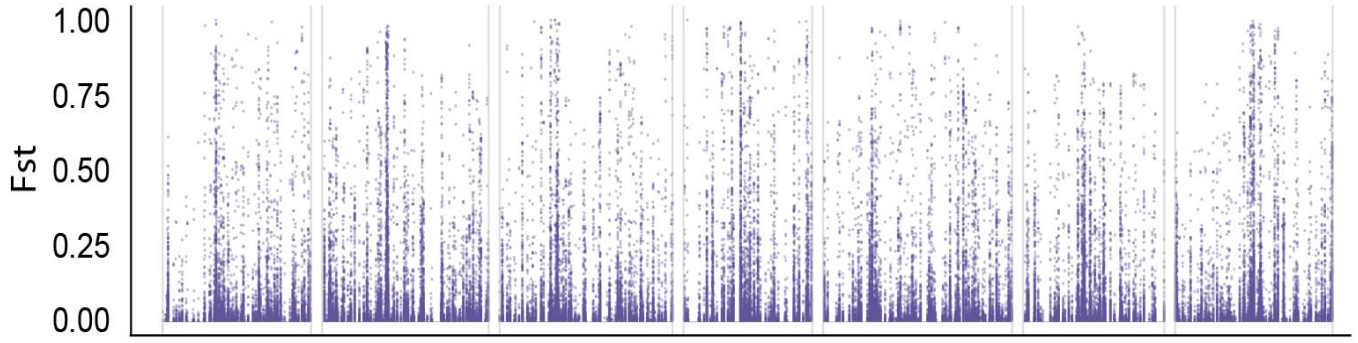
Complete difference

A T
A T
A T
A T
A T
A T

A 1 0
T 0 1

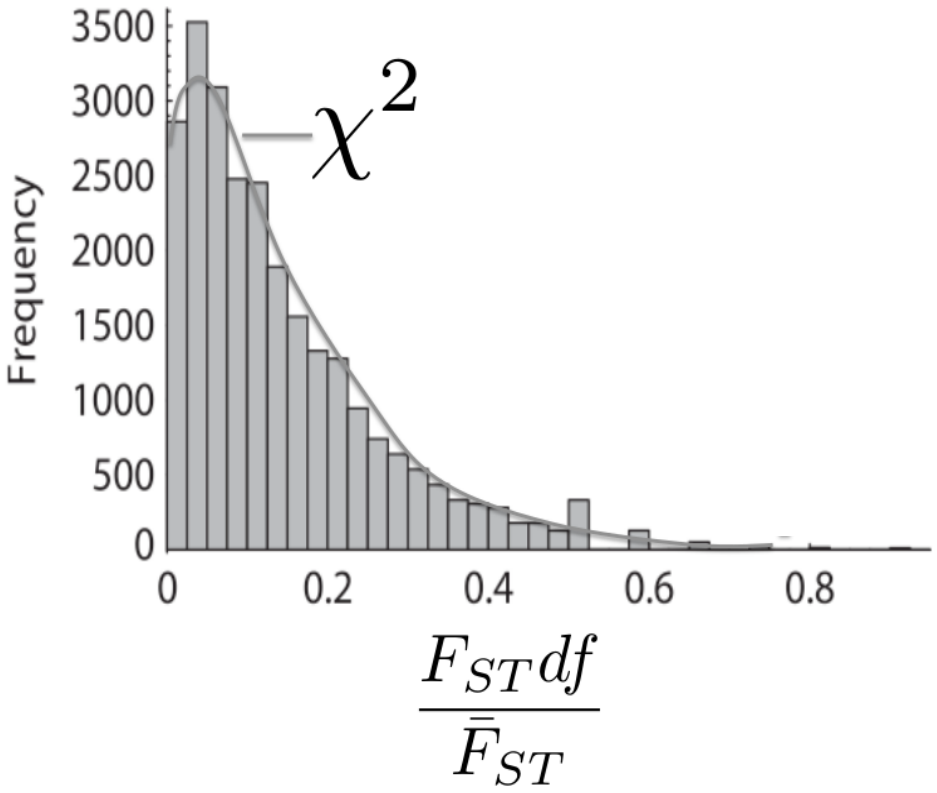
Genome-wide Fst scans – how to define outliers

Regions of exceptional differentiation indicate selection on adaptive mutation

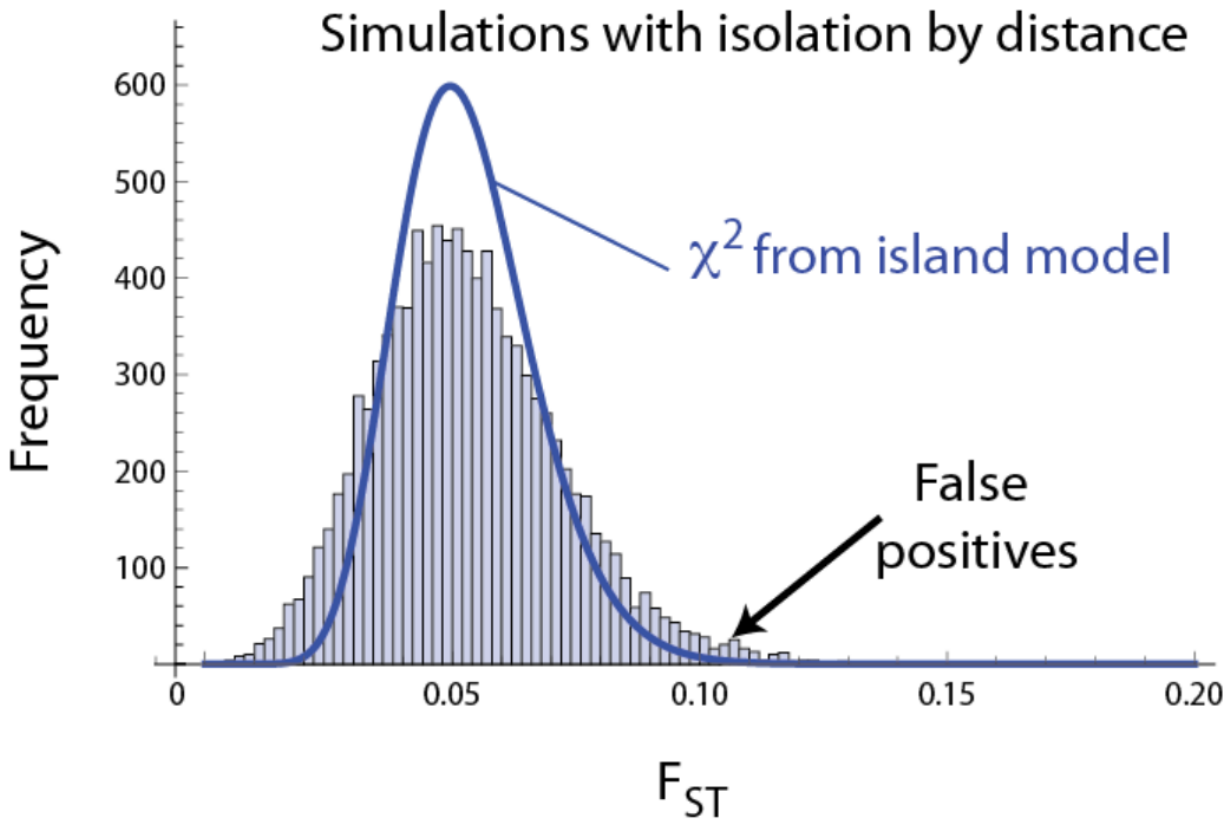


Lewontin-Krakauer (LK) test

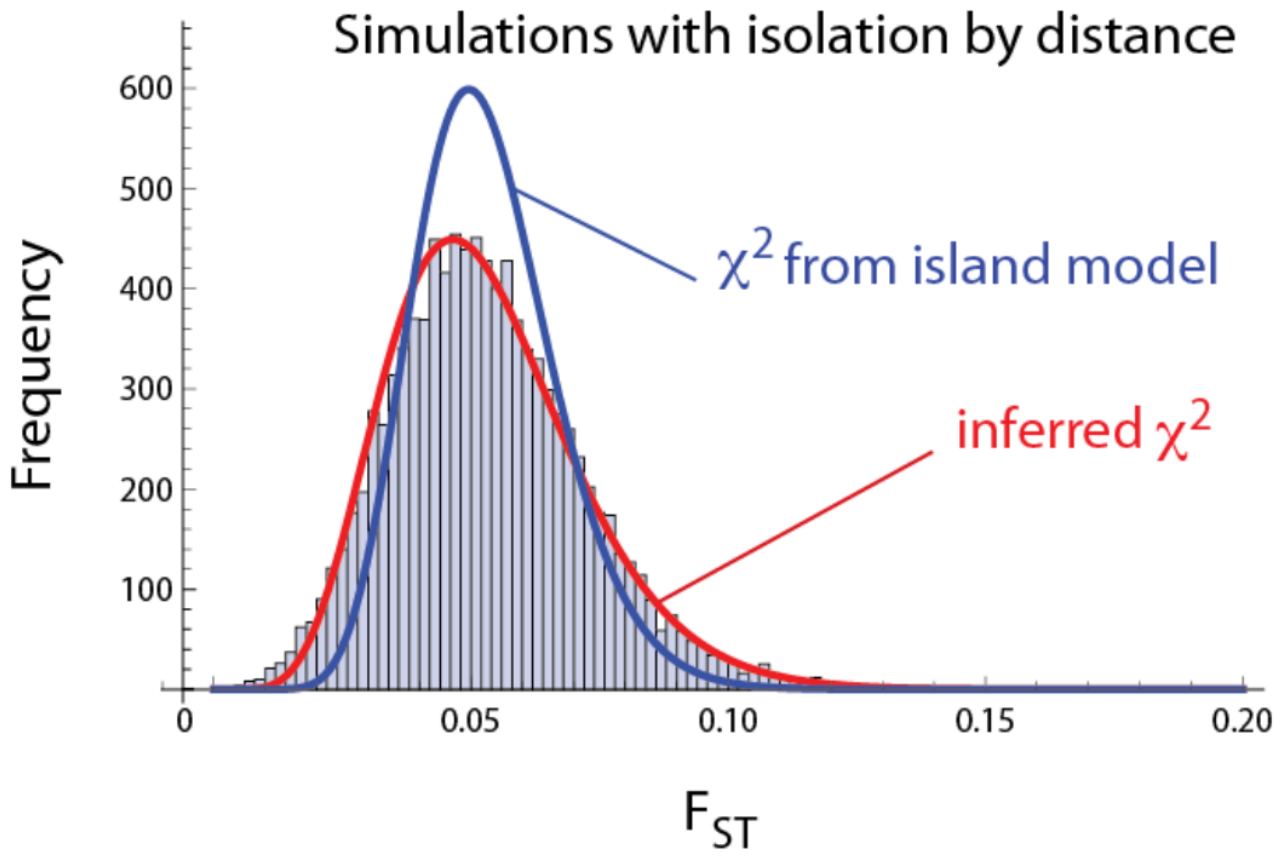
Assumes island model



LK test, isolation by distance – different distribution



Using genome-wide Fst to estimate distribution

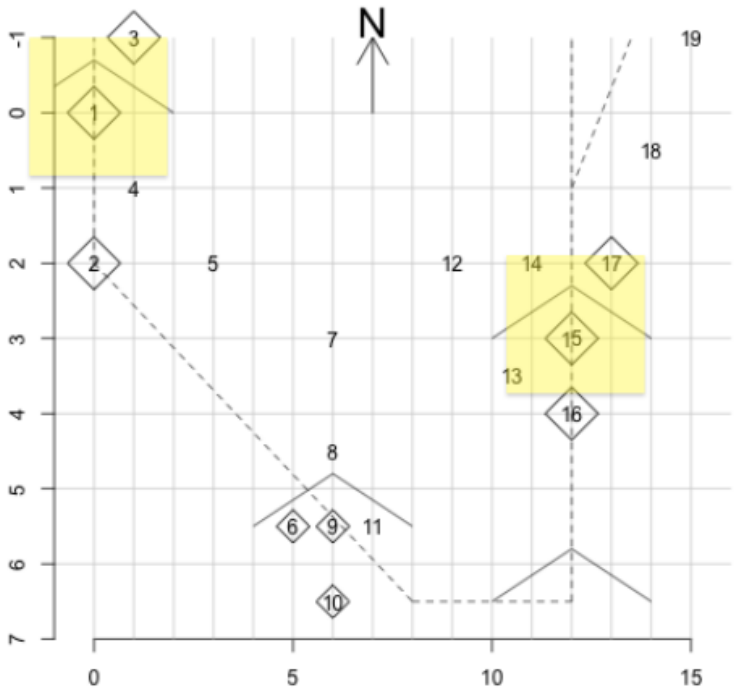


Detecting Fst outliers with OutFLANK – example

Simulations by Katie Lotterhos

Generation 0, refuges filled (Pops 1-3 and 15-17)
Generation 3000, expand to melted summit (Pop 6,9,10)
Generation 4000, all populations available

Isolation by distance model



Detecting Fst outliers with OutFLANK – exercise

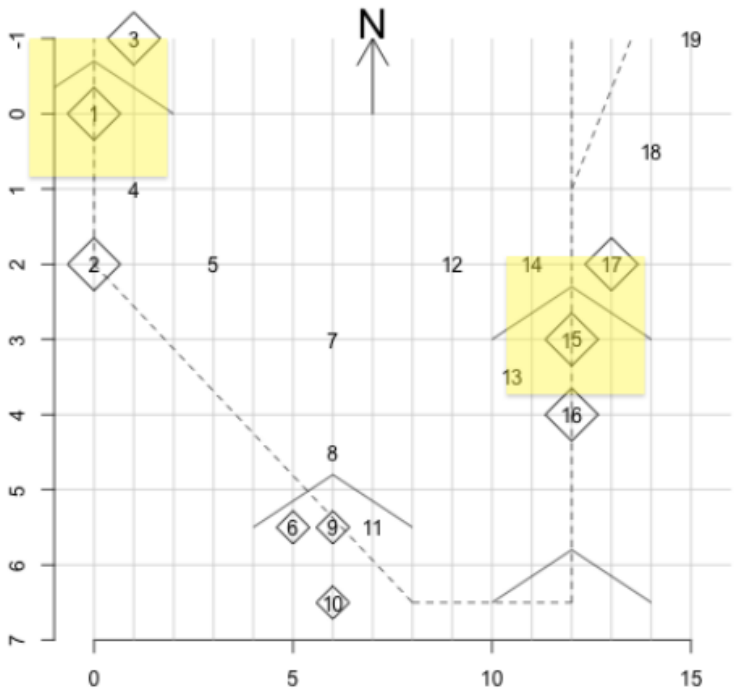
Simulations by Katie Lotterhos

Generation 0, refuges filled (Pops 1-3 and 15-17)
Generation 3000, expand to melted summit (Pop 6,9,10)
Generation 4000, all populations available

Island model migration

12 QTL

663 706 923 1163 1347 1378
1639 1666 1825 2133 2556 2871



Detecting Fst outliers with OutFLANK – take home

OutFLANK

- Low false positive rate (false positive neutral / total neutral) – true positives
- Low power in complex demographic and selection scenarios

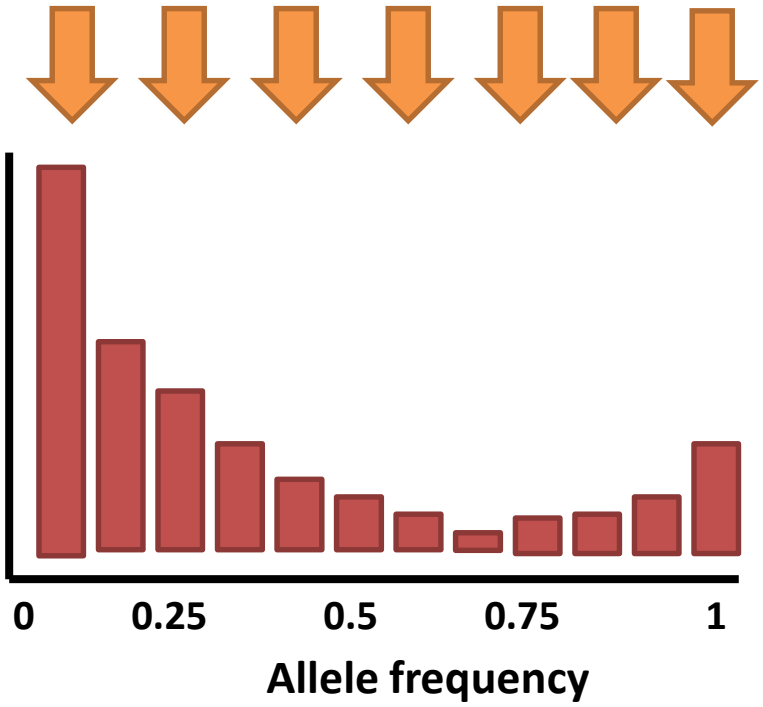
Which means it rarely detects signals but if it does those are likely true outliers

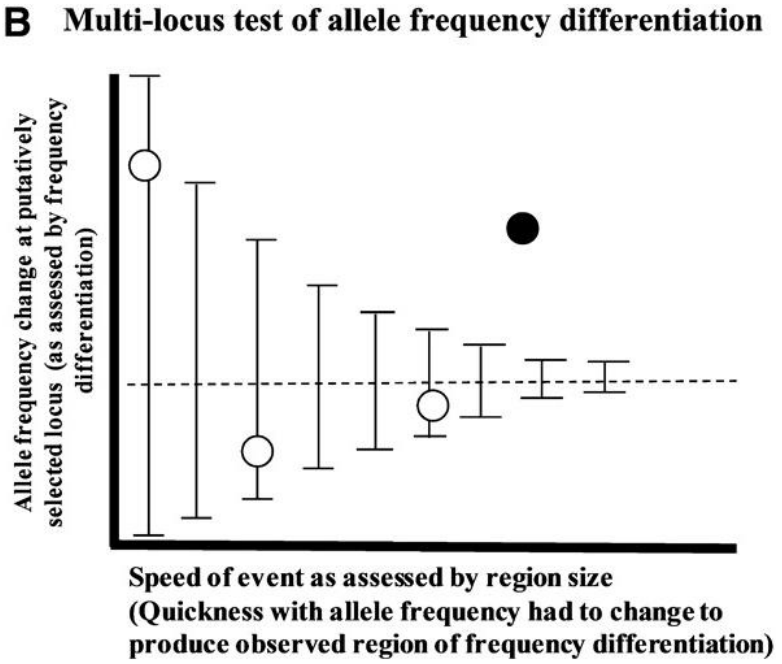
- Other approaches:

PCAadapt
hapFLK
SelEstim
BayPass
LFMM
Bayenv

Cross-population CLR (XP-CLR)

- considered robust to demographic history and SNP ascertainment bias
- requires well resolved physical & genetic maps
- requires dense genotyping or re-sequencing assay
- compares divergence only between two populations





Chen et al. Genome Res. 2010

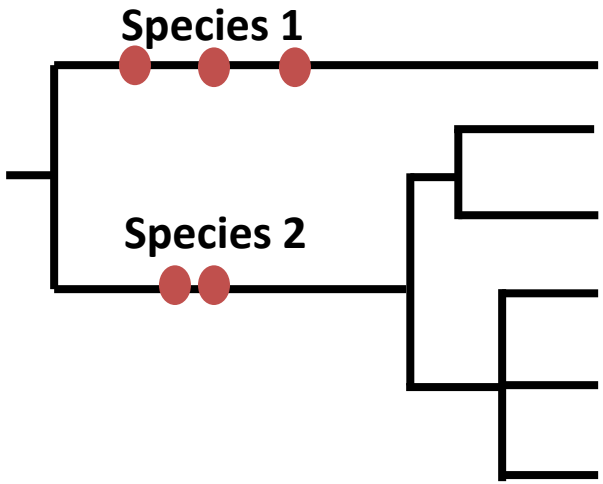
Detecting selection at macroevolutionary scale (between species)

The Neutral theory of molecular evolution (Kimura)

most of the variation seen at the molecular level is selectively neutral

Most of the observed mutations are neutral and their frequency governed by drift.

In genes, most non-synonymous mutations are neutral, which behave just like synonymous mutations.



$Ka/Ks = Dn/Ds = \omega$

$Ka/Ks \sim 0$ – neutral

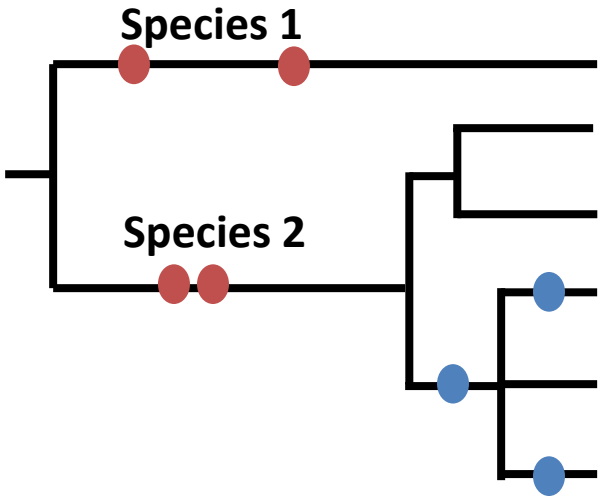
$Ka/Ks > 0$ – diversifying selection

$Ka/Ks < 0$ – purifying selection (against nonsyn)

McDonald – Kreitman test (MK)

McDonald

Neutral theory predicts that ratio of non-synonymous to synonymous changes should be constant through time.



Kreitman

Contrasts “present” (within species) with “historical” (between species) non-syn:syn ratios.

McDonald – Kreitman test (MK) example

```
          *       *     *
Species 1 CTT ACT TAT ACC CGT
          CTG ACT TAT ACC CGT
          CTG ACT TCT ACC CGT
          CTG ACT TCT ACA CGT
          *       *
Species 2 ATG ACC TCT ACC CGT
```

	Fixed between	Polymorphism within
non-syn	**1 (A)**	**1 (B)**
syn	**1 (C)**	**2 (D)**
	historical	present day

Ratio is statistically tested using G

A/C = B/D, if all non-syn are neutral

A/C > B/D – non-syn substitutions were advantageous & selected

A/C < B/D – maladaptive recessive non-syn mutations persist within species

MK test – Exercise 1

Adaptive protein evolution at the *Adh* locus in *Drosophila*

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

image source: beerymethod.com

TABLE 1 Variable nucleotides from the coding region of the *Adh* locus in *D. melanogaster*, *D. simulans* and *D. yakuba*

	Con.	*D. melanogaster* a b c d e f g h i j k l	*D. simulans* a b c d e f	*D. yakuba* a b c d e f g h i j k l		
781	G	TTTTTTTTTTTT	------	------------	Repl.	Fixed
789	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
808	A	------------	------	GGGGGGGGGGGG	Repl.	Fixed
816	G	TTTT-------T	TTTTTT	------------	Syn.	Poly.
834	T	------------	CC---C	------------	Syn.	Poly.
859	C	------------	------	GGGGGGGGGGGG	Repl.	Fixed
867	C	------------	------	GGGGGAGGGGGG	Syn.	2 Poly.
870	C	TTTTTTTTTTTT	------	------------	Syn.	Fixed
950	G	------------	-A----	------------	Syn.	Poly.
974	G	------------	T-TTTT	------------	Syn.	Poly.
983	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1019	C	------------	------	----A-------	Syn.	Poly.
1031	C	------------	------	--------A---	Syn.	Poly.
1034	T	------------	------	-CCCCC--C-CC	Syn.	Poly.
1043	C	------------	------	----A-------	Syn.	Poly.
1068	C	TT----------	------	------------	Syn.	Poly.
1089	C	------------	AAAAAA	------------	Repl.	Fixed
1101	G	------------	------	AAAAAAAAAAAA	Repl.	Fixed
1127	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1131	C	------------	------	----T-------	Syn.	Poly.
1160	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1175	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1178	C	------------	------	----------A-	Syn.	Poly.
1184	C	------------	------	GGGGGGGGGGGG	Syn.	Fixed
1190	C	------------	------	--A---------	Syn.	Poly.
1196	G	------------	------	TTTT-TTT-T--	Syn.	Poly.
1199	C	-T----------	------	------------	Syn.	Poly.
1202	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1203	C	------------	--T---	------------	Syn.	Poly.
1229	T	--CCCCCCCCCC	------	------------	Syn.	Poly.
1232	T	------------	------	AAAAAAAAAAAA	Syn.	Fixed
1235	C	----------A-	------	------------	Syn.	Poly.
1244	C	------------	------	--A---------	Syn.	Poly.
1265	C	------------	------	GGGGGGGGGGGG	Syn.	Fixed
1271	A	------------	-T-T--	------------	Syn.	Poly.
1277	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1283	C	AA----------	------	------------	Syn.	Poly.
1298	C	------------	------	TTTTTTTTTTTT	Syn.	Fixed
1304	C	------------	----T-	------------	Syn.	Poly.
1316	C	------------	--TT--	TTTTTTTTTTTT	Syn.	Poly.
1425	C	AA----------	------	------------	Syn.	Poly.
1431	T	CC----------	------	CCCCCCCCCCCC	Syn.	Poly.
1443	C	------GGGGGG	------	------------	Syn.	Poly.
1452	C	-----TTTTTTT	------	------------	Syn.	Poly.
1490	A	------CCCCCC	------	------------	Repl.	Poly.
1504	C	TTTTTTTTTTTT	------	------------	Syn.	Fixed
1518	C	-----TTTTTTT	------	------------	Syn.	Poly.
1524	T	------------	------	GGGGGGGGGGGG	Syn.	Fixed
1527	C	TTTTTT------	------	----T-------	Syn.	Poly.
1530	G	------------	------	---A--------	Syn.	Poly.
1545	T	------------	------	CCCCCCCCCCCC	Syn.	Fixed
1548	C	------------	------	----A-------	Syn.	Poly.
1551	C	------------	---T--	------------	Syn.	Poly.
1555	C	-----------T	------	------------	Repl.	Poly.
1557	C	AAAAA-------	------	------------	Syn.	Poly.
1560	G	------------	---A--	------------	Syn.	Poly.
1573	G	------------	------	CCCCCCCCCCCC	Repl.	Fixed
1581	C	------------	------	TTTTTTTTTTTT	Syn.	Fixed
1584	C	------------	------	GGGGGGGGGGG-	Syn.	Poly.
1590	C	TTTTTTTTTTTT	TTT---	------------	Syn.	Poly.
1596	G	--AA-A-----A	------	------------	Syn.	Poly.
1611	A	------------	------	TTTTTTTTTTTT	Syn.	Fixed
1614	C	------------	-G----	-------T----	Syn.	2 Poly.
1635	C	------------	------	--T-T-------	Syn.	Poly.
1657	A	------------	------	TTTTTTTTTTTT	Repl.	Fixed
1665	C	------------	------	---------T--	Syn.	Poly.

McDonald – Kreitman test (MK) - exercise 2

Calculate MK test in pairwise comparisons between species (3 pairs)

McDonald – Kreitman test (MK) - exercise 2

Calculate MK test in pairwise comparisons between species (3 pairs)

Siddiq, Mohammad A., et al. "Experimental test and refutation of a classic case of molecular adaptation in Drosophila melanogaster." Nature Ecology & Evolution 1 (2017): 0025.

Selection tests – take home

•Mutation, drift, migration, recombination & selection are ongoing evolutionary processes affecting allele frequencies

•Discerning selection signatures from neutral is challenging

•Multiple tests exist but none yet provide definitive answers….

…. but rather help formulate evolutionary hypotheses, which can be experimentally tested

Diversity estimates – 2 scenarios

-- Exercise handout --

```
GTT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TCG CCG TTC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CAC TGT CCG TAC GCC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATT CGC TGT CCG TAC GTC
ATA CGC TGT CCG GAC GTC
ATA CAC TGT CCG GAC GTC
```

n =
S =
Hap =
Θw =
Θπ =

```
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATT CGC TGT CCG TAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GAT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
ATA CAC TGT CCG GAC GTC GCT CGC
```

n =
S =
Hap =
Θw =
Θπ =

Tajima's D – 3 scenarios

-- Exercise handout --

Example 1

1 ATT CGC TGT CCG TAC GTC GAT CGC
2 ATT CGC TGT CCG TAC GTC GAT CGC
3 ATT CGC TGT CCG TAC GTC GAT CGC
4 ATT CGC TGT CCG **G**AC GTC GAT CGC
5 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
6 AT**A** C**A**C TGT CCG **G**AC GTC G**C**T CGC

Θw = *Θπ =* *Tajima's D =*

Example 2

1 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
2 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
3 ATT CGC TGT C**T**G TAC G**C**C GAT CGC
4 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
5 AT**A** CGC TGT CCG **G**AC GTC GAT CGC
6 AT**A** C**A**C TGT CCG **G**AC GTC G**C**T CGC

Θw = *Θπ =* *Tajima's D =*

Example 3

1 ATT CGC TGT CCG TAC GTC GAT CGC
2 ATT CGC TGT CCG TAC G**C**C GAT CGC
3 ATT CGC TGT CCG TAC GTC GAT CGC
4 ATT CGC TGT CCG TAC GTC GAT CGC
5 ATT C**C**C TGT CCG TAC GTC GAT CGC
6 AT**A** CGC TGT CCG TAC GTC G**C**T CGC

Θw = *Θπ =* *Tajima's D =*