Sequence analysis and genomics

7. Tests for selection

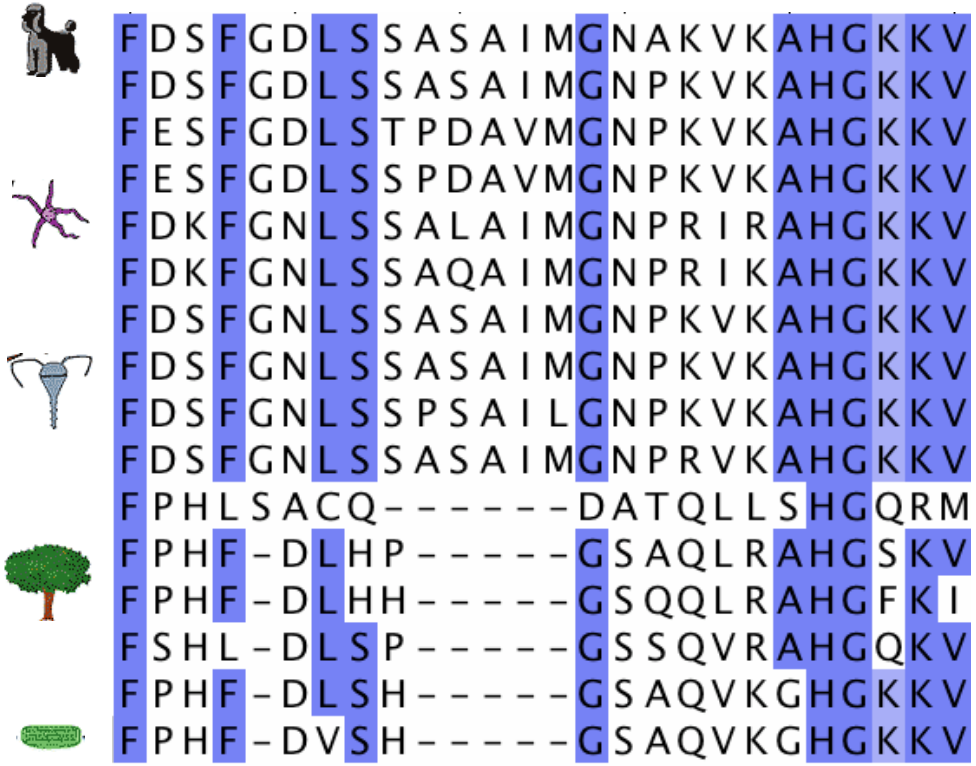
Dr. Katja Nowick

Group leader TFome and Transcriptome Evolution
Bioinformatics group
Paul-Flechsig-Institute for Brain Research
www. nowicklab.info
nowick@bioinf.uni-leipzig.de

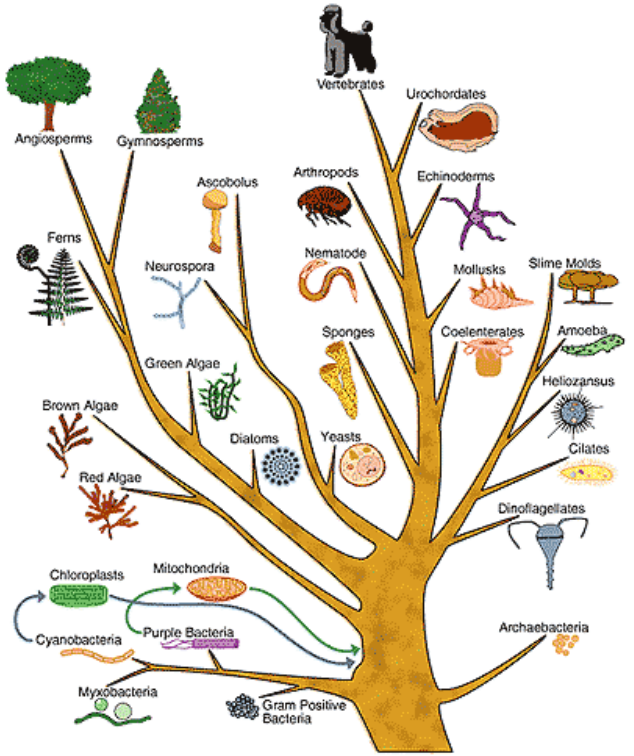
Phylogenetic analysis

Goal: Analysis of the relationship among a set of sequences that can be aligned

Multiple Sequence Alignment



Tree



Three major methods for phylogenetic analysis

Distance methods (e.g. Neighbor joining)

Progressive addition of the next most similar sequence to the tree; produces tree with shortest total branch length

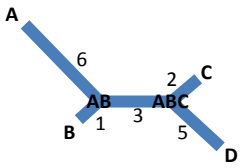
Distance matrix

	A	B	C	D
A	0	7	11	14
B	7	0	6	9
C	11	6	0	7
D	14	9	7	0

Q-Matrix

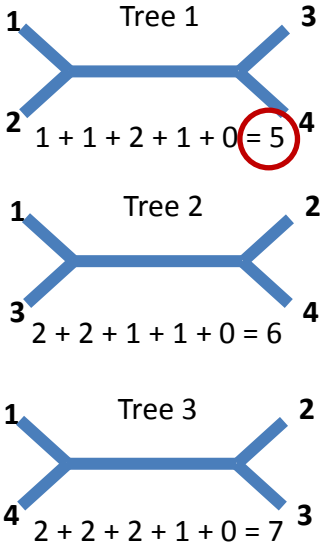
	A	B	C	D
A	0	−40	−34	−34
B	−40	0	−34	−34
C	−34	−34	0	−40
D	−34	−34	−40	0

Tree



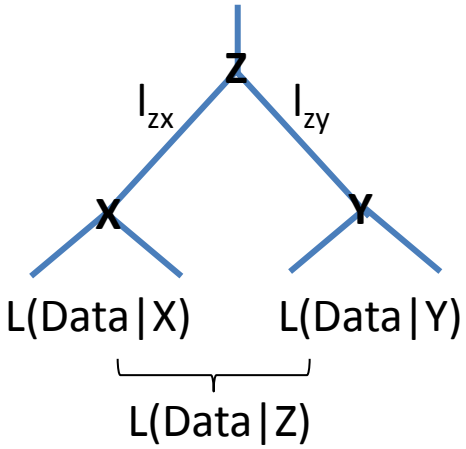
Maximum Parsimony

Predicts the tree that best fits the sequence variation in each column of the MSA; tree with minimum number of evolutionary steps to produce the observed sequences



Maximum Likelihood

Predicts the tree that best fits the sequence variation in each column of the MSA using an underlying model and probability calculations for nucleotide or aa changes; tree is the most likely arrangement of the branches for the observed sequences



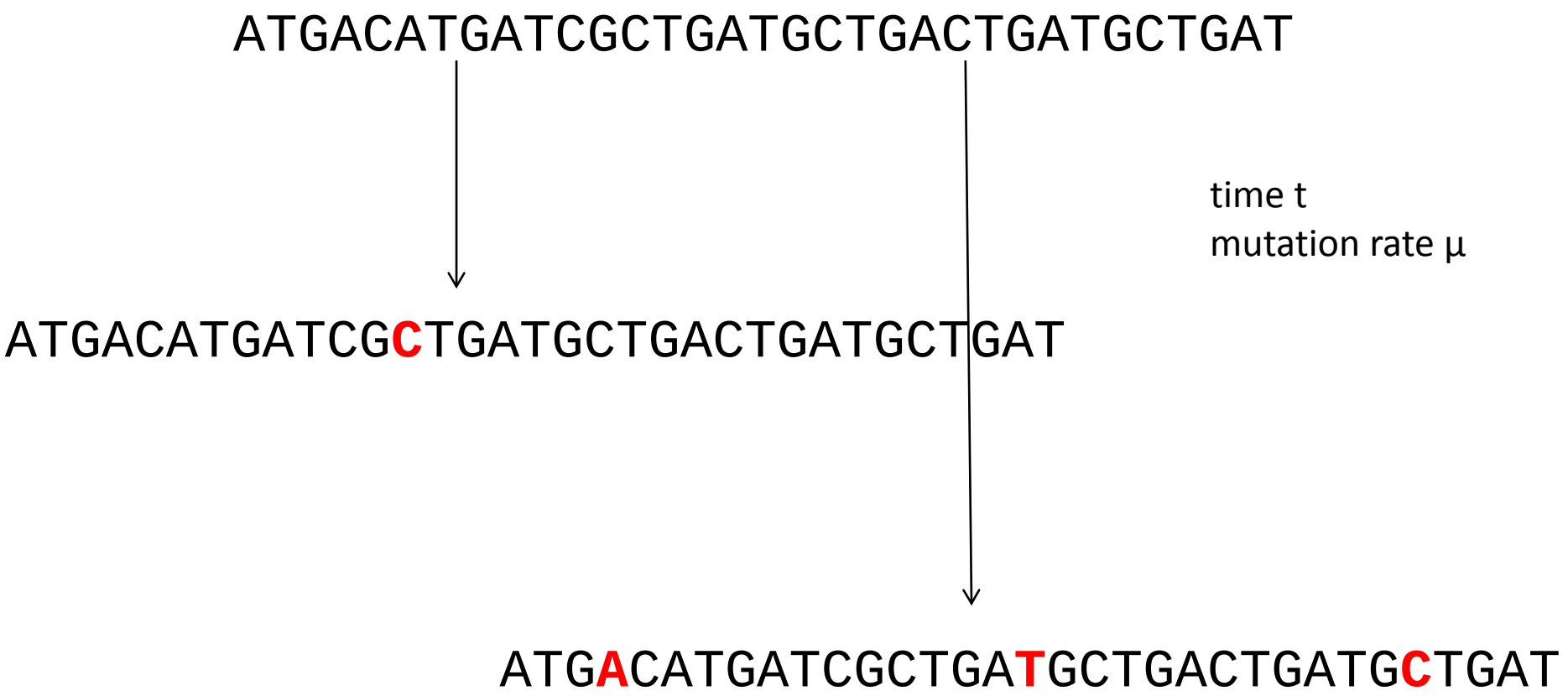
How reliable are my trees?

- Bootstrapping
- Use at least two of the major methods and compare the trees

Common phylogenetic programs

PHYLIP:	a collection of programs using distance, MP, or ML
PAUP:	originally only MP, later versions also distance and ML
PAML:	uses ML
TREE PUZZLE:	uses ML
MrBayes:	Bayesian estimates for the most likely tree

Tests for selection



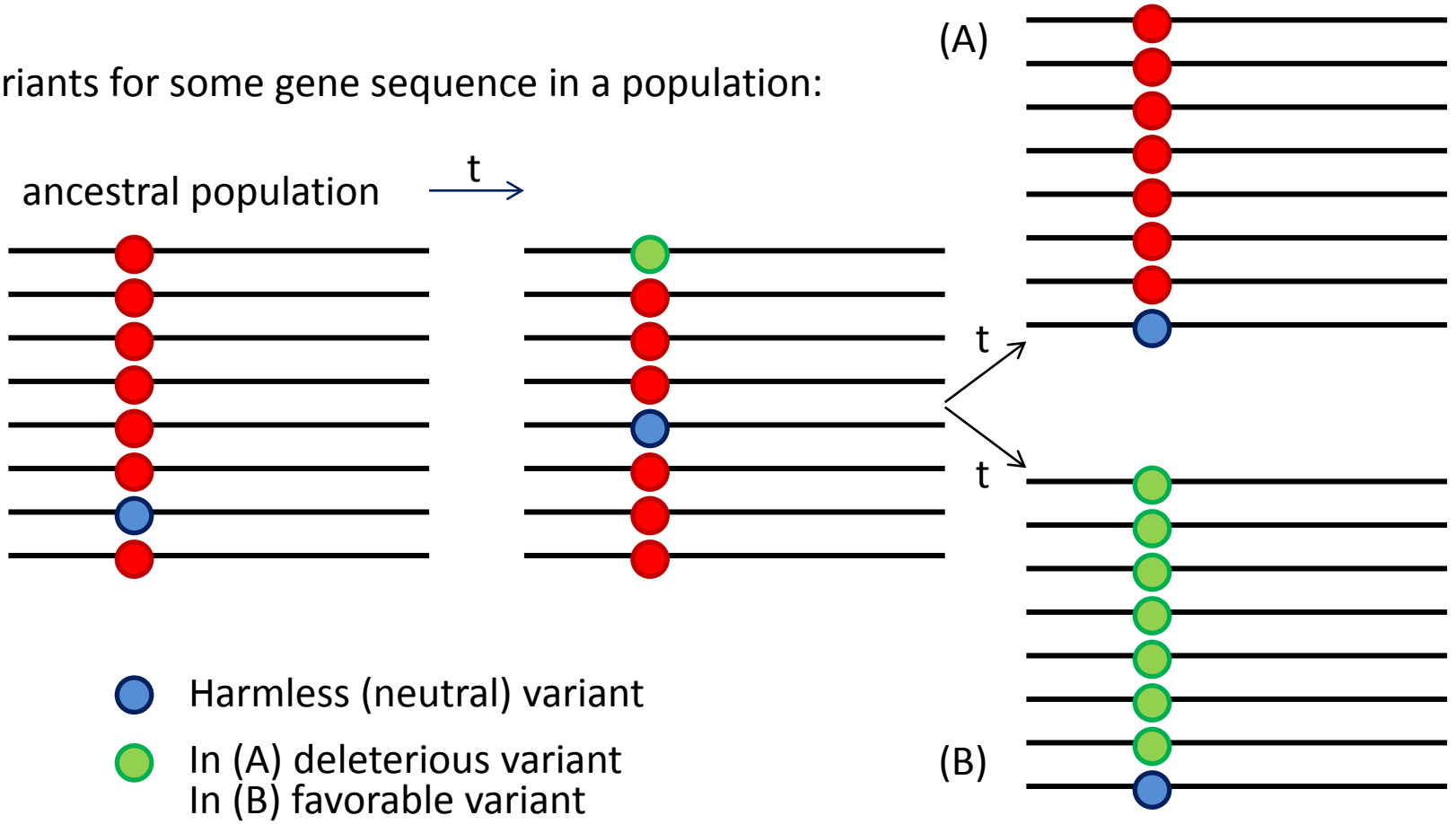
Number of substitutions depends on the time and mutation rate

Tests for selection

Mutations creates new variants

But which variants are passed on to the next generation depends on **selection**

Variants for some gene sequence in a population:



Tests for selection

Is my favorite gene X under selection or evolving neutrally?

- Positive selection (evolves faster, has changed more)

  Adaptation to a new environment
  New favorable trait

- Negative selection (evolves more slowly, seems highly conserved)

  Disease genes

- Balancing selection (two or more alleles are at high frequency)

  β-globin gene (sickle cell anemia / malaria resistance)

Tests for selection

Codon-based tests

For populations and between species
For proteins
Need to consider unequal mutation rate across genome

e.g. dN/dS
McDonald-Kreitman test

Allele frequency tests

For populations
For all types of sequences
Need to consider demography

e.g. Tajima's D
Hudson-Kreitman-Aguade (HKA) test

Need to compare observation to neutral expectation

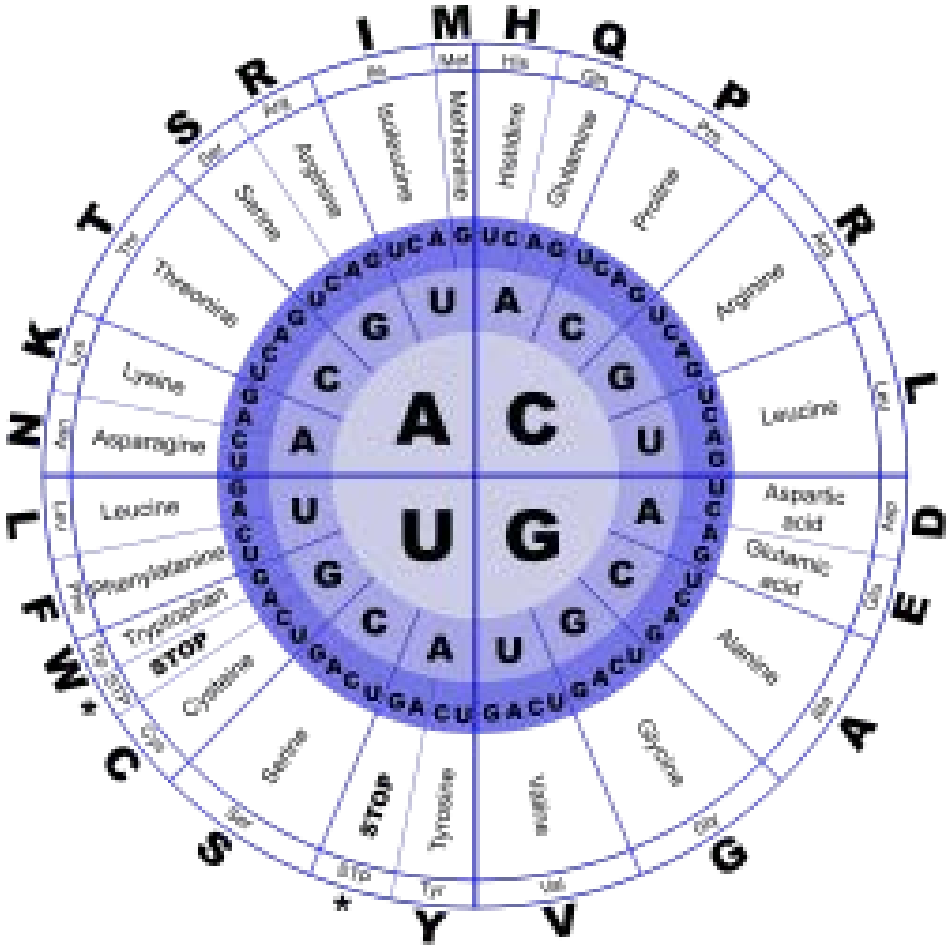
Codon-based tests

Selection acts on function (e.g. proteins)

Genetic code:

An amino acid is encoded by three nucleotides (= triplet, codon)

Most aa are encoded by more than one codon
The third codon position often doesn't matter
i.e. codons are synonymous



Mutation rate at synonymous positions is considered neutral
→ Is used to normalize , to compare mutation rates at non-synonymous sites to
Advantage of normalizing to dS: takes care of varying mutation rate across the genome

PAML

Software package Phylogenetic Analysis by Maximum Likelihood

Calculates ratios of non-synonymous vs. synonymous rates of mutations

Very powerful for estimation of parameters and hypothesis testing

Sophisticated substitution models (evolutionary models) for testing biological hypotheses

e.g. estimate synonymous and non-synonymous substitution rates

test for positive Darwinian evolution

Tree making algorithm rather primitive

Commonly people make the alignment by e.g. CLUSTALW, MUSCLE, T-COFFEE, DIALIGN

and then the tree with e.g. PHYLIP, PAUP, PhyML, RaxML

before applying PAML

PAML

Sophisticated substitution models (evolutionary models) for testing biological hypotheses:

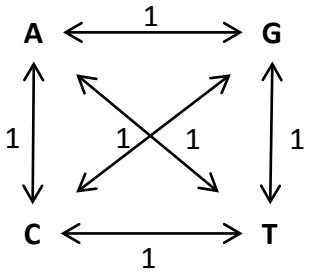
PAML implements JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), F81 (Felsenstein 1981), F84 (Felsenstein, DNAML program since 1984, PHYLIP Version 2.6), HKY85 (Hasegawa et al. 1984; Hasegawa et al. 1985), Tamura (1992), Tamura and Nei (1993), and REV, also known as GTR for general-time-reversible (Yang 1994b; Zharkikh 1994).

<u>Parameters for rate matrices, e.g.:</u>

JC69 :

$$Q = \begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{pmatrix} . & 1 & 1 & 1 \\ 1 & . & 1 & 1 \\ 1 & 1 & . & 1 \\ 1 & 1 & 1 & . \end{pmatrix} \end{array} \begin{matrix} T \\ C \\ A \\ G \end{matrix}$$

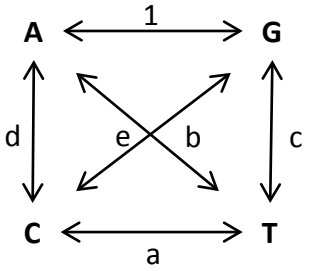
All substitutions have the same frequency



REV (GTR):

$$Q = \begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{pmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & \pi_G \\ c\pi_T & e\pi_C & \pi_A & . \end{pmatrix} \end{array} \begin{matrix} T \\ C \\ A \\ G \end{matrix}$$

Substitutions have different frequencies
π = frequency of that nucleotide in the sequence



Other models consider back mutations (useful for long evolutionary distances) or are basically intermediates between JC69 and REV

Codon-based tests

- Detecting positive/negative selection -

Codon substitution models:
The unit of evolution is the codon (not individual nucleotides or aa)
Most commonly used version is the model by Yang and Nielsen 1998 (YN00):
Chemical differences between amino acids are ignored

Because natural selection operates on the protein level, changes in nucleotides that lead to an amino acid change are under higher selective pressure

- Non-synonymous changes = changes that lead to aa change
- Synonymous changes do not lead to an aa change, i.e. are silent

dN (K_a) = rate of non-synonymous changes
# of ns-substitutions / # of ns-sites

dS (K_S) = rate of synonymous changes
# of s-substitutions / # of s-sites

Codon-based tests

- Detecting positive/negative selection -

dN (K_a) = rate of non-synonymous changes
# of ns-substitutions / # of ns-sites

dS (K_S) = rate of synonymous changes
# of s-substitutions / # of s-sites

ω = dN/dS (K_N/K_S)
used to infer direction and magnitude of selection on aa changes

$\omega < 1$ → negative purifying selection
$\omega = 1$ → neutral evolution
$\omega > 1$ → positive selection

Problem if species very closely related and almost no K_S → alternative K_i

Codon-based tests

- Detecting positive/negative selection -

Example: Genome wide scan for genes under positive selection in humans:

Nielsen et al. 2005

13731 human-chimpanzee orthologous genes

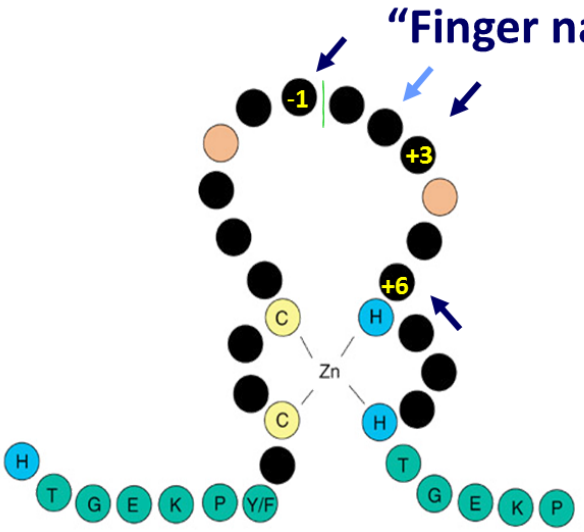
733 genes had $\omega > 1$

35 were significant in LRT against null hypothesis of $dN/dS = 1$

Many of them encode proteins of the immune system and olfactory receptors

Detecting adaptive evolution (positive selection)

However: ω of a gene is rarely larger than 1 !
Why?

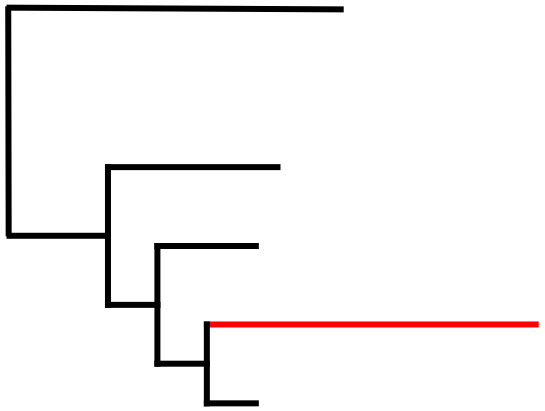


Usually only parts of a protein are under positive selection
while others are under strong purifying selection
to maintain the basic structure and function of a protein

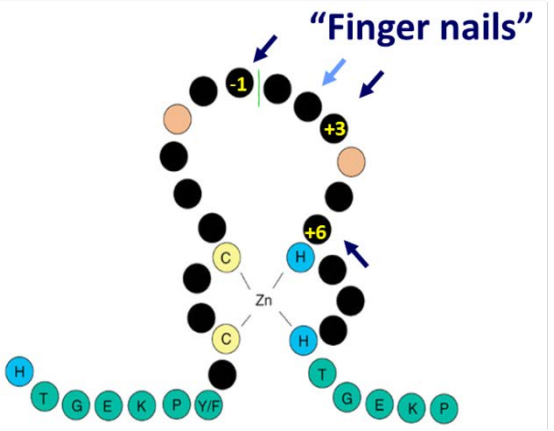
→ site models

Usually positive selection acts only during a certain evolutionary time window

→ branch models



Site models



Used to detect **positive selection acting on sites** (i.e. codons)
Let ω parameters vary across the different codons in the alignment
Assumes three different classes of sites: ω < 1, ω = 1, and ω > 1

Calculate likelihood for model "some codons have ω > 1"
and for model "no codon has ω > 1" (= null model)

Parameters in Site Models

Model	NSsites	p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

p = proportions of sites of a specific class

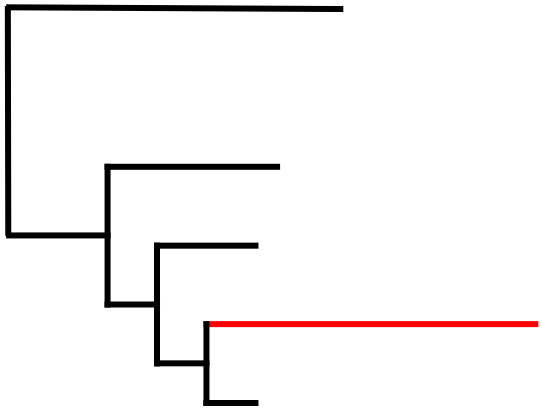
Typically people test M1a vs. M2a and M7 vs. M8

Make *Likelihood Ratio Test (LRT)* for the two models: LRT statistics = $2\Delta\ell$
then *Chi-squared test (X^2 –test)*: if $X^2 > 5.99$ → significant at 5% level
if $X^2 > 9.21$ → significant at 1% level

If LRT suggests positive selection, perform *Bayes empirical Bayes* method to calculate posterior probability for each site that it is a site from the class "positive selection" under M2a and M8

Branch models

Used to detect **positive selection acting on particular lineages**
Need to provide a tree
Let ω parameters vary across the different branches of the tree



Predefine the branch(es) that will be tested for positive selection
= **foreground branch(es)**
All other branches are **background branches**

LRT: for null model 0: all branches have the same ω
vs. alternative: foreground branch(es) have different ω

Branch-site models

Used to detect positive selection that affects only **certain sites** on **predefined lineages**
Need to provide a tree
The branch to test is called foreground branch, all others are background branches

Site Class	Proportion	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)\, p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

4 classes of sites:

Class 0: codons that are conserved throughout the tree with ω < 1
Class 1: codons that evolve neutrally throughout the tree with ω = 1
Class 2a: codons that are conserved in background but under positive selection in foreground branch
Class 2b: codons that are neutral in background but under positive selection in foreground branch

Calculate LRT for branch-site model A vs. two different null models:

Test 1: null hypothesis is site model M1a (only negative or neutral on any branch)
Test 2: null model is the branch-site model A with $\omega_2 = 1$ fixed (neutral evolution on foreground)

→ LRT with test 2 is the real tests for "positive selection on foreground branch"
test 1 would also be significant if relaxation of constraints

If LRT suggests positive selection for the foreground branch, calculate *Bayes empirical Bayes* posterior probability for each site that it is a site from the class "positive selection"

Branch-site models

- Some real data examples -

LRT Statistics ($2\Delta\ell$) and P Values for Different Branch-Site Tests of Positive Selection

Data Set	Number of Sequences	Number of Codons	Tests of this Paper: Test 1	Tests of this Paper: Test 2
Primate lysozyme *c* gene	19	130	3.36 ($P = 0.19$)	1.48 ($P = 0.22$)
Primate cancer gene BRCA1	8	1,160	8.10 ($P = 0.017$)	3.64 ($P = 0.056$)
Angiosperm phytochrome gene *phyA-C/F*	15	1,105	84.88 ($P < 0.001$)	19.88 ($P < 0.001$)
Angiosperm phytochrome gene *phyB-D/E*	11	1,105	57.90 ($P < 0.001$)	13.29 ($P < 0.001$)

Zhang et al., 2003

Lysozyme c: 19 primates

Positive selection is suspected along the branch to colobine monkeys (foreground branch)

leaf-eaters with foreguts where lysozyme may have acquired a new digestive function

was not significant

BRCA1: tumor suppressor gene

branch to humans and chimpanzees was suspected to be under positive selection

test 2 not significant, hence no positive selection

Phy: phytochrome genes involved in response to light in angiosperms

foreground branch separating A and C/F subfamilies and branch separating B and D/E subfamily significant

PAML

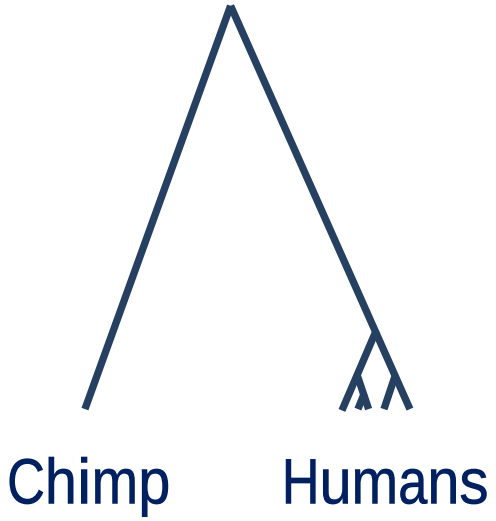
Other phylogenetic tests that are implemented in PAML programs:

- Comparison and tests of phylogenetic trees (baseml and codeml);
- Estimation of parameters in sophisticated substitution models, including models of variable rates among sites and models for combined analysis of multiple genes or site partitions (baseml and codeml);
- Likelihood ratio tests of hypotheses through comparison of implemented models (baseml, codeml, chi2);
- Estimation of divergence times under global and local clock models (baseml and codeml);
- Likelihood (Empirical Bayes) reconstruction of ancestral sequences using nucleotide, amino acid and codon models (baseml and codeml);
- Generation of datasets of nucleotide, codon, and amino acid sequence by Monte Carlo simulation (evolver);
- Estimation of synonymous and nonsynonymous substitution rates and detection of positive selection in protein-coding DNA sequences (yn00 and codeml).
- Bayesian estimation of species divergence times incorporating uncertainties in fossil calibrations (mcmctree).

McDonald-Kreitman test

Neutral evolution

Within species dN/dS = between species dN/dS



Positive selection

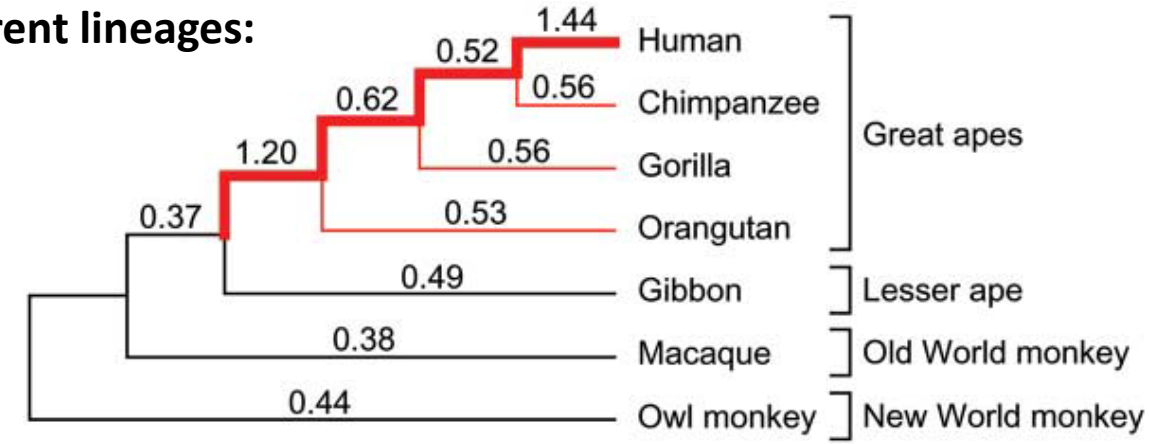
Within species dN/dS < between species dN/dS

Evolution of ASPM

Example: ASPM (Abnormal spindle-like microcephaly associated)
Implicated in determining brain size

Evans et al., 2003

ω for the different lineages:



→ Branch leading to great apes is faster: $p < 0.001$

McDonald Kreitman test
including 40 human individuals:

→ Excess of non-synonymous sites in human-chimp comparison vs. in polymorphisms among humans

Table 1. The numbers of non-synonymous (N) and synonymous (S) nucleotide changes in *ASPM*

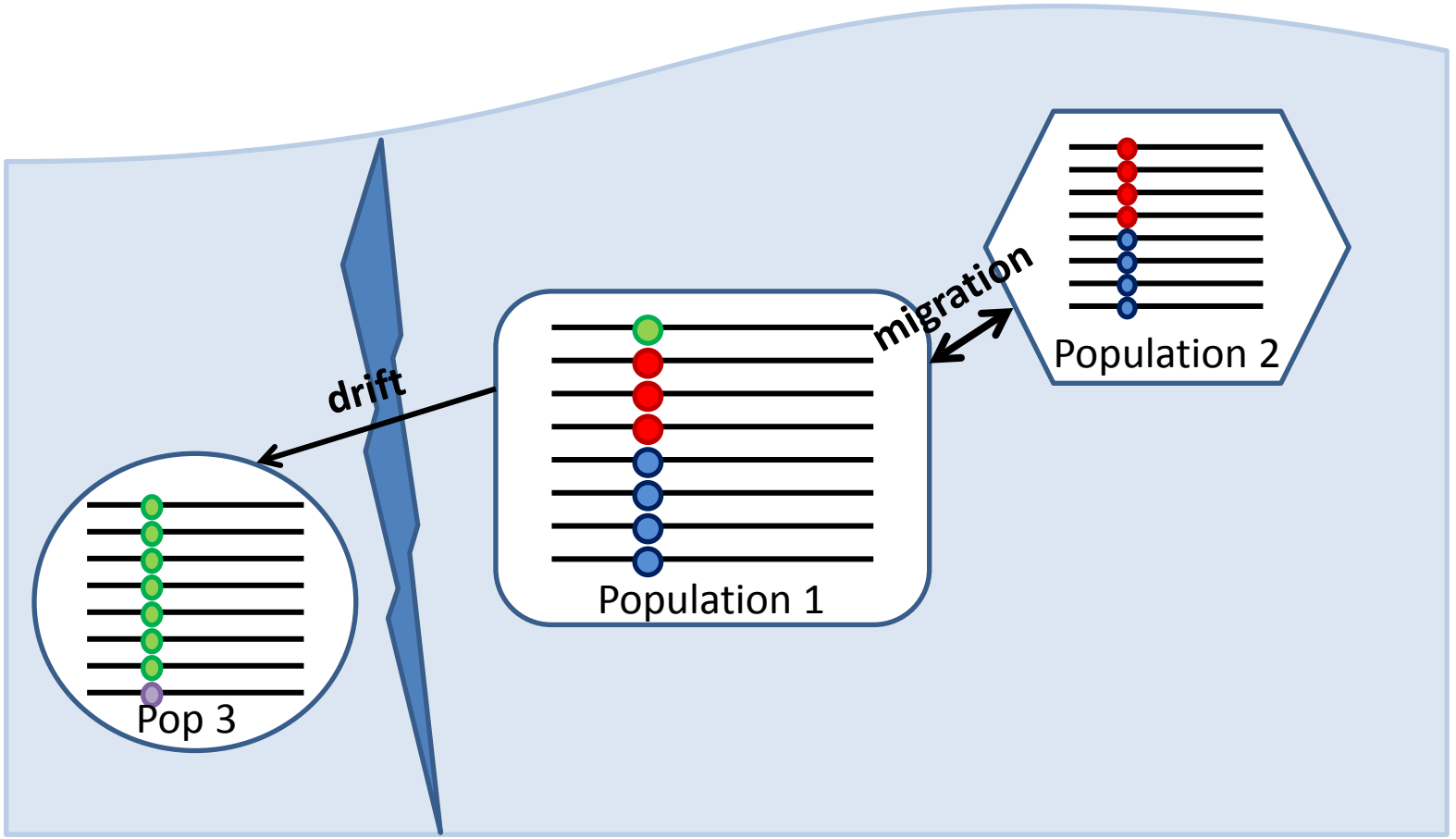
	N	S	P-value
Polymorphism within human populations ($n = 80$)	6	10	
Divergence between human and last human/chimpanzee ancestor	19	7	0.025

n denotes the number of haploid genomes sampled; P-value is from Fisher's exact test.

→ **ASPM evolves under positive selection on lineage to great apes and humans**

Allele frequency tests

Definition of a population:

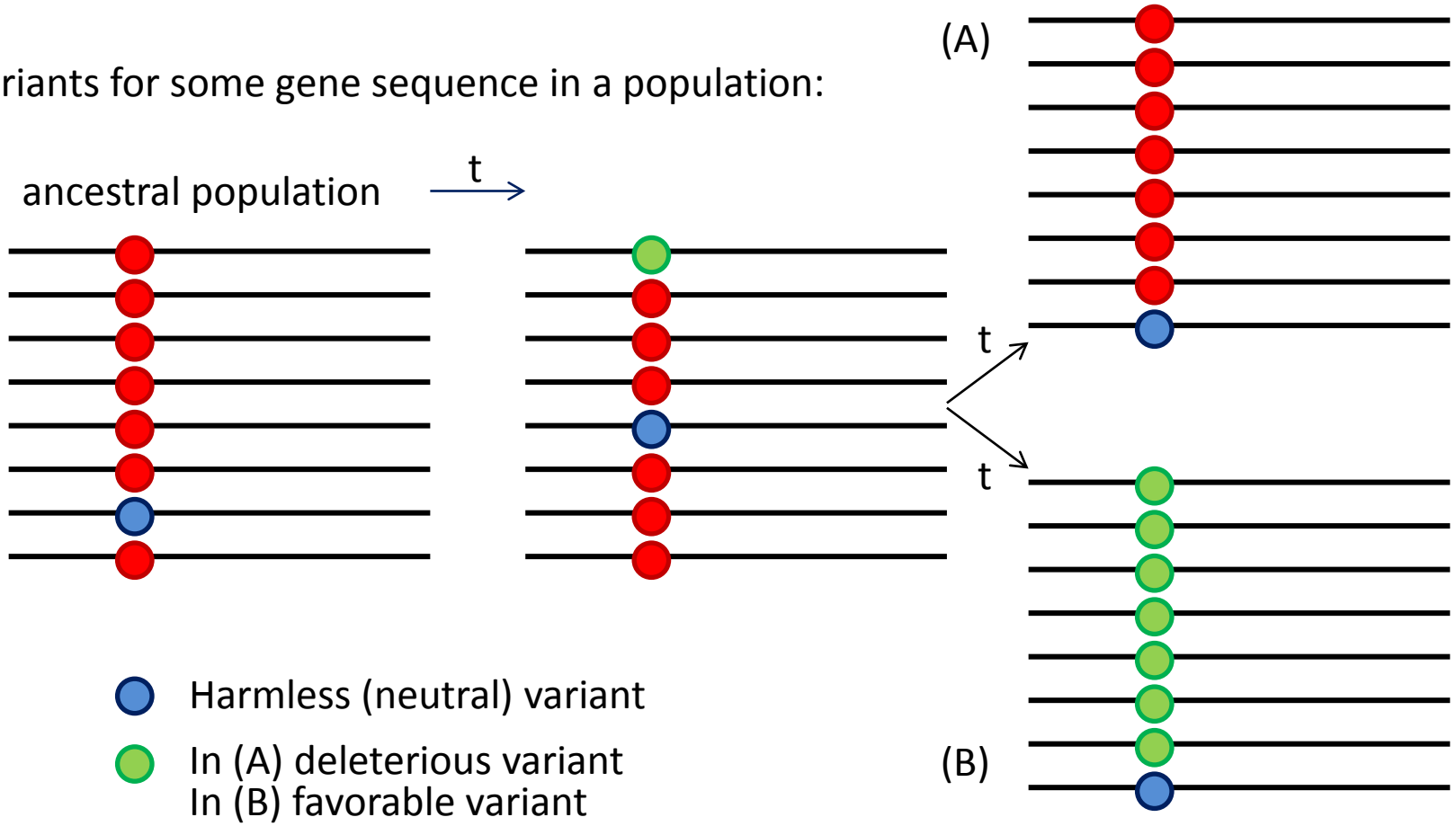


Individuals in a population have different alleles

Alleles come with a certain frequency = **site frequency spectrum**

Allele frequency tests

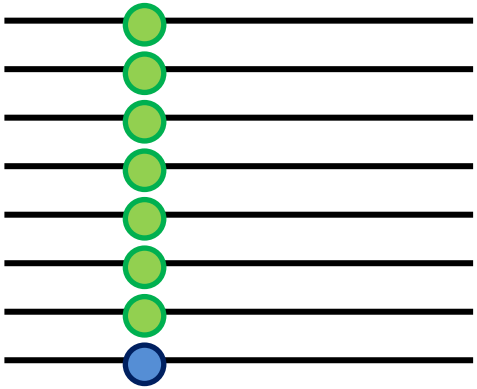
Variants for some gene sequence in a population:



Note: similar site frequency spectra can also be caused by changes in demography!

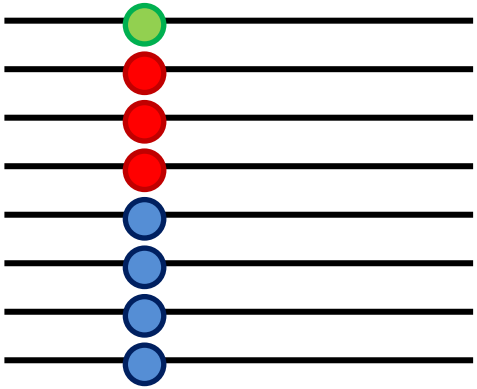
Allele frequency tests

Demographic vs. selection explanation



Low level of genetic diversity can be caused by

- Extensive drift (small population size)
- Low immigration
- Negative selection (removal of deleterious variants)
- Positive selection (sweep, fixation of favorable allele)

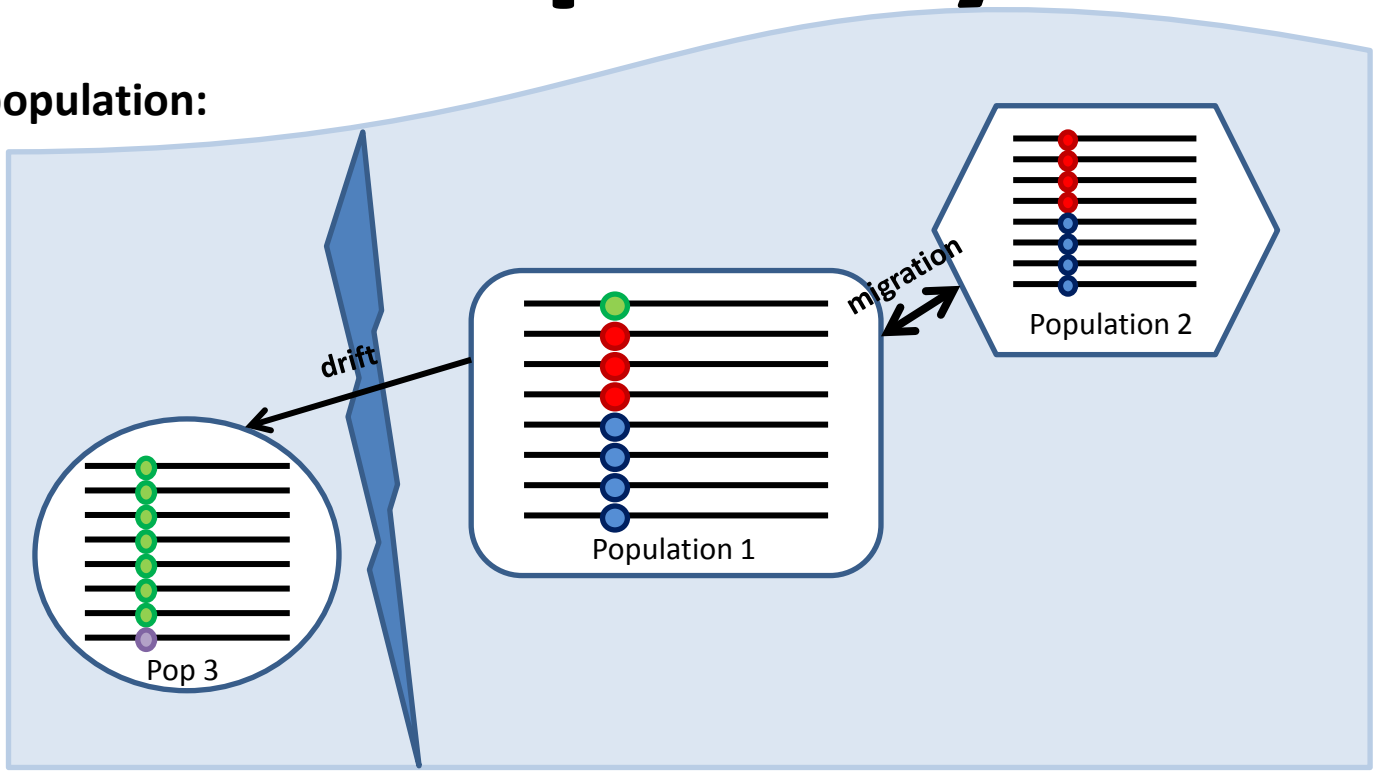


High level of genetic diversity can be caused by

- Large population size
- Extensive immigration
- Balancing selection (favoring increase in diversity)

Allele frequency tests

Definition of a population:



Most tests assume:

- No mutation
- No selection
- Infinite/constant population size (i.e. no migration)
- Random mating
- Diploid organisms

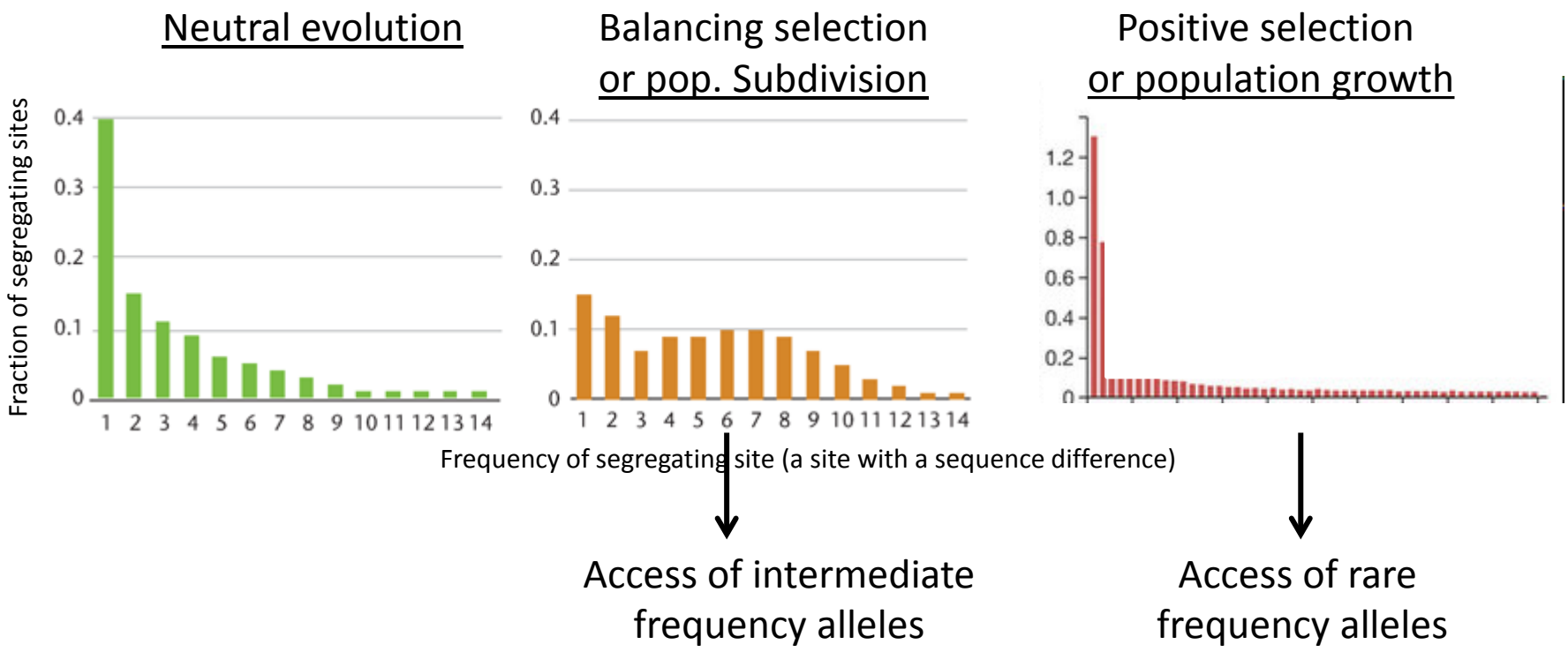
Note: all these assumptions are unrealistic!

Allele frequency tests

Principle of the tests:

Summary statistics of DNA sequence diversity

Compare observation vs. what is expected under neutral evolution



Tajima's D

S = # of segregating sites (sites that have a difference)
independent of allele frequency

π = nucleotide diversity
mean of pairwise differences (take all pairwise differences and divide by # of pairs)
depends on allele frequency

Differences between S and π tells you if the sequence is evolving neutrally
d = π-S
D = 0 → neutral evolution
D > 0 → balancing selection or population subdivision (a few alleles at high frequency)
D < 0 → positive selection or population growth (many rare alleles)

When is D different from zero?
Depends on how variable S and π are from sample to sample
→ d = π-S is divided by its standard deviation → D

Implemented in e.g. DNASP

Tajima's D calculation

$$
\begin{aligned}
a_1 &= \sum_{i=1}^{n-1} 1/i \\
a_2 &= \sum_{i=1}^{n-1} 1/i^2 \\
b_1 &= \frac{n+1}{3(n-1)} \\
b_2 &= \frac{2(n^2+n+3)}{9n(n-1)} \\
c_1 &= b_1 - 1/a_1 \\
c_2 &= b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2} \\
e_1 &= c_1/a_1 \\
e_2 &= c_2/(a_1^2 + a_2)
\end{aligned}
$$

Then Tajima's D is

$$
D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}}
$$

Tajima's D calculation

Example:

Position	12345	67890	12345	67890
Person Y	00000	00000	00000	00000
Person A	00100	00000	00100	00010
Person B	00000	00000	00100	00010
Person C	00000	01000	00000	00010
Person D	00000	01000	00100	00010

Four **polymorphic sites** (positions where someone differs from You):
positions 3, 7, 13 and 19

Tajima's D calculation

```
Position  12345 67890 12345 67890
Person Y  00000 00000 00000 00000
Person A  00100 00000 00100 00010
Person B  00000 00000 00100 00010
Person C  00000 01000 00000 00010
Person D  00000 01000 00100 00010
```

Number of polymorphisms between any two sequences:

Y vs. A: 3 polymorphisms

```
Person Y     00000 00000 00000 00000
Person A     00100 00000 00100 00010
```

Y vs. B: 2 polymorphisms

```
Person Y     00000 00000 00000 00000
Person B     00000 00000 00100 00010
```

⋮

C vs. D: 1 polymorphism

```
Person C     00000 01000 00000 00010
Person D     00000 01000 00100 00010
```

Tajima's D calculation

Number of polymorphisms between any two sequences:

Y vs. A: 3 polymorphisms

```
Person Y      00000 00000 00000 00000
Person A      00100 00000 00100 00010
```

⋮

C vs. D: 1 polymorphism

```
Person C      00000 01000 00000 00010
Person D      00000 01000 00100 00010
```

Average number of polymorphisms between two sequences:

$$\frac{3+2+2+3+1+3+2+2+1+1}{10} = 2$$

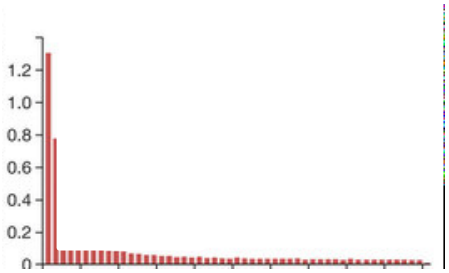
Number of segregating sites = number of polymorphic sites = 4

$$d = 2 - 4 = -2$$

D is then d divided by the square root of it's variance = the standard deviation

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

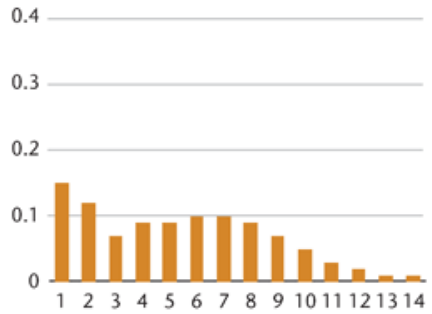
Tajima’s D interpretation



Interpretation of a strong negative Tajima's D:

Signifies an excess of low frequency polymorphisms relative to expectation

Indicating population size expansion (e.g., after a bottleneck or a selective sweep) and/or positive selection



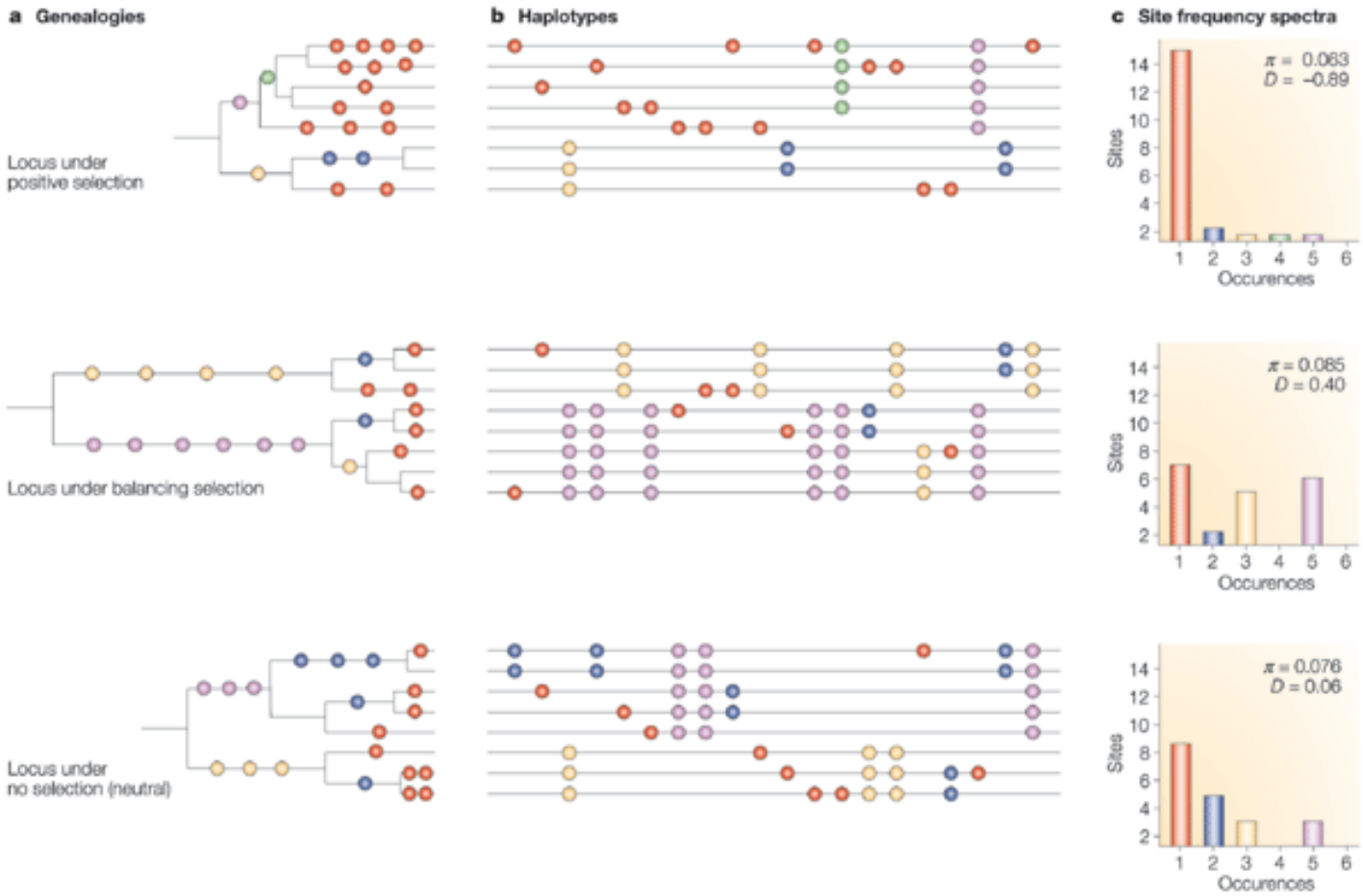
Interpretation of a strong positive Tajima's D:

Signifies low levels of both low and high frequency polymorphisms

Indicating a decrease in population size and/or balancing selection

Tajima's D interpretation

(Each locus has 20 segregating sites)



Positive selection results in lower level of sequence diversity and an excess of low-frequency variants (red) → negative value of Tajima's D

Balancing selection results in higher level of sequence diversity and an excess of intermediate-frequency variants (purple) → positive value of Tajima's D

The diversity estimate and site frequency spectrum of a neutral locus (bottom) can be used for comparison

Tajima's D statistics

Problems with the statistics, because the single value of D has no p-value attached

A very rough rule of thumb to significance is that values greater than +2 or less than -2 are likely to be significant
(because if the differences between π and S was normally distributed D would be between -2 and 2 in 95% of cases)

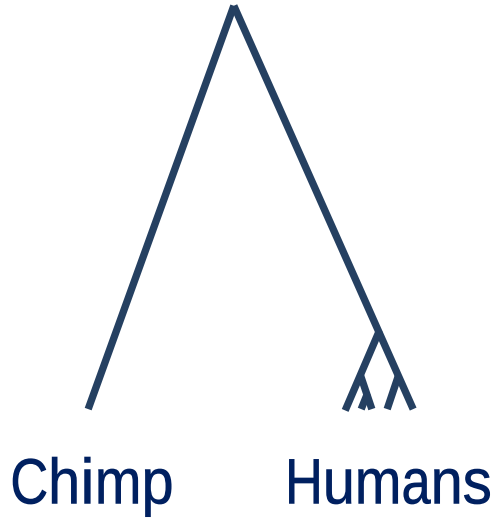
Typically people compare D between the gene of interest with the D of some other loci they think should be neutral or to some genome wide distribution of sliding windows

Neutral loci provide information about demography

Hudson-Kreitman-Aguade (HKA) test

Neutral evolution

Within species diversity = between species divergence



Positive selection

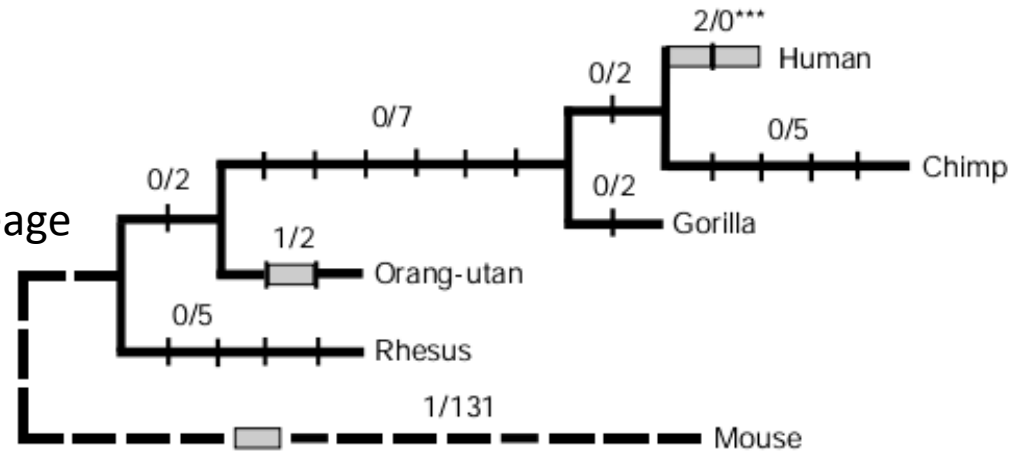
Within species diversity < between species divergence

Compare diversity/divergence ratio for your locus of interest
To diversity/divergence ratio for a neutral locus or genome wide average
Null model: Ratio for locus of interest is not different from neutral loci

Evolution of FOXP2

FOXP2: implicated in language acquisition
Enard et al., 2002

Codon-based test:
Shown are # of ns-sites/s-sites per lineage



Allele frequency tests:
Tajima's D = -2.20
To exclude that the strong negative Tajima's D of FOXP2 is just due to bottleneck
Tajima's D of FOXP2 was compared to Tajima's D of 313 human genes
Only one gene had a more negative Tajima's D than FOXP2

No. of chromosomes sequenced	40
Length covered (double stranded, all individuals)	14,063 bp
Divergence from the chimp sequence*	0.87%
No. of variable positions	47
Singletons (no. of variable sites occurring at frequency 1 and 39)	31
θ_W (nucleotide diversity based on the no. of polymorphic sites)	0.079%
θ_π (mean nucleotide diversity)	0.03%
θ_H (nucleotide diversity with more weight given to alleles at high frequency[17])	0.117%
D ($P < 0.01$)†	−2.20
H ($P < 0.05$)‡	−12.24