Detection of genomic footprints of natural selection

Genomic approaches to variation and adaptation: a road map
– 9 November 2020 –

Thibault Leroy
thibault.leroy@univie.ac.at

Genetic basis of adaptive evolution, an important topic in evolutionary biology!

Different methods depending on the levels of divergence:

Long-time scales	Short-time scales
Different species (divergence)	Different populations
Substitutions	Polymorphisms
Individual-level data	Population-level data
Protein-coding sequences	Whole genome sequences (if possible)

— substitutions

— polymorphisms

Species 1

…ACGTATGTGCGTGGTAGCCTAG…

…ACGTACGTGCGTGGTAGCCTGG…

…ACGTATGTGCGTGGTAGCCTAG…

…ACGTACGTGCGTGGTAGCCTAG…

Species 2

…AAGTACGTGCGCGGTAGGCTAG…

…AAGTACGTGCGCGGTAGCCTAG…

…AAGTACGTGCGCGGTAGCCTAG…

…AAGTACGTGCGCGGTAGCCTAG…

Genetic basis of adaptive evolution, an important topic in evolutionary biology!

Different methods depending on the levels of divergence:

Long-time scales	Short-time scales
Different species (divergence)	Different populations
Substitutions	Polymorphisms
Individual-level data	Population-level data
Protein-coding sequences	Whole genome sequences (if possible)

— substitutions

— polymorphisms

Species 1

…ACGTATGTGCGTGGTAGCCTAG…
…ACGTACGTGCGTGGTAGCCTGG…
…ACGTATGTGCGTGGTAGCCTAG…
…ACGTACGTGCGTGGTAGCCTAG…

Species 2

…AAGTACGTGCGCGGTAGGCTAG…
…AAGTACGTGCGCGGTAGCCTAG…
…AAGTACGTGCGCGGTAGCCTAG…
…AAGTACGTGCGCGGTAGCCTAG…

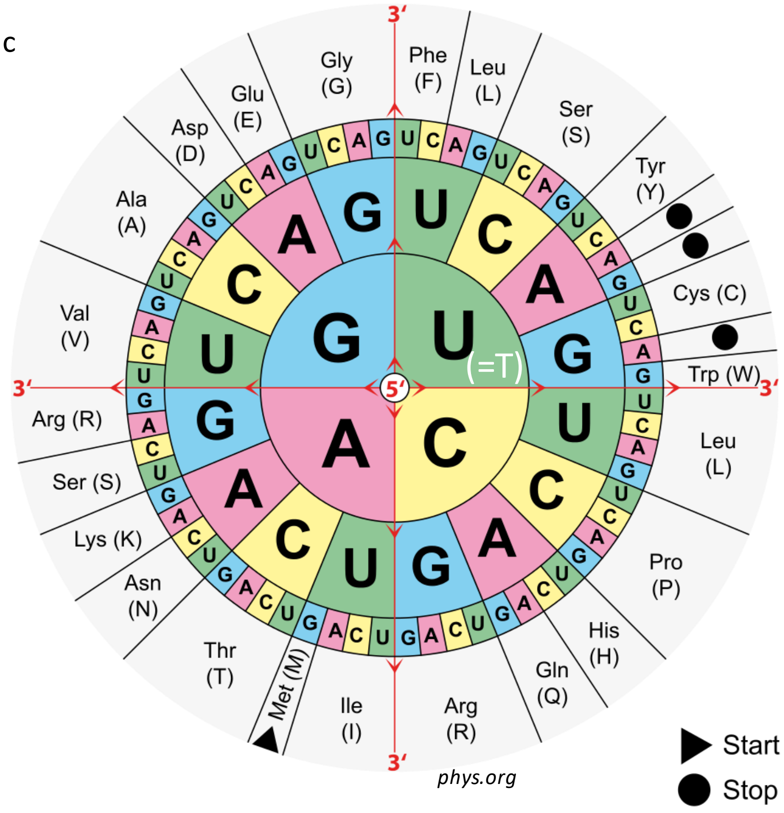
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

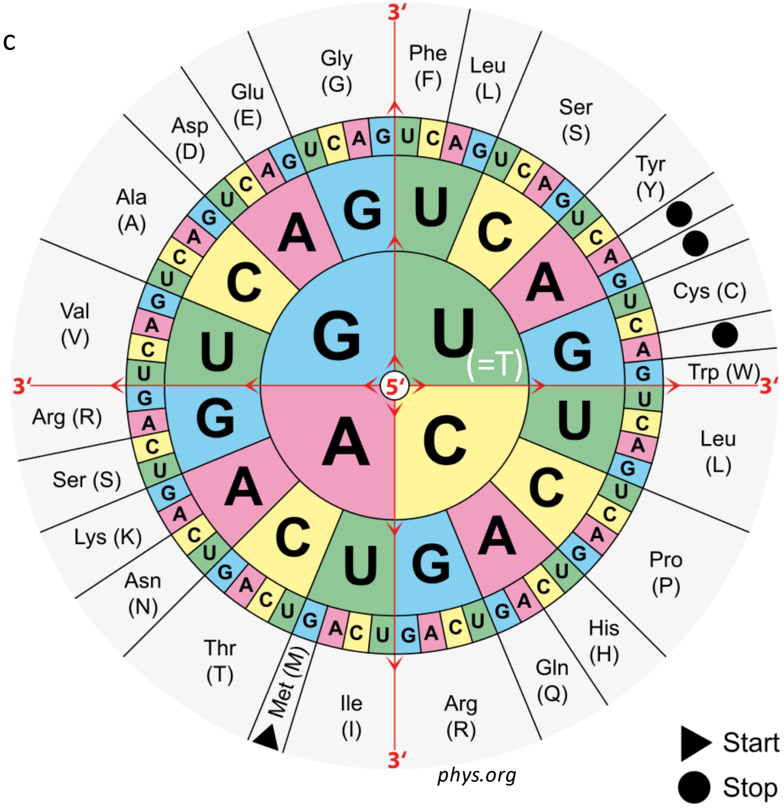
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

↓

ACG = Thr

CCG = Pro

GCG = Ala

TCG = Ser

All mutations at this position will change the amino acid!

Syn sites = 0

Non-Syn sites =1

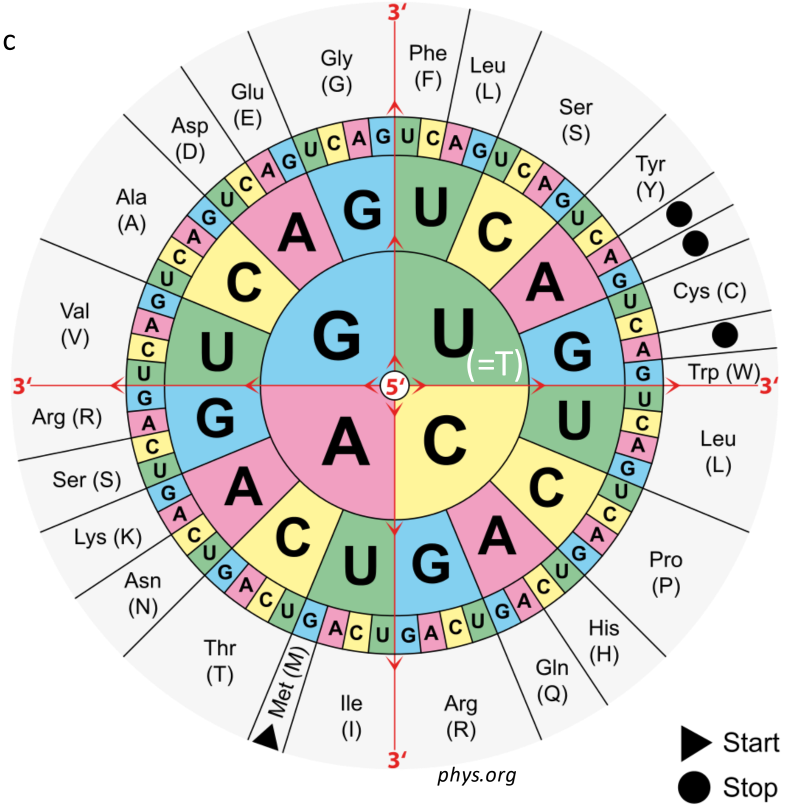
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

ACG = Thr
AGG = Arg
ATG = Met
AAG = Lys

All mutations at this position will change the amino acid!

Syn sites = 0
Non-Syn sites =2

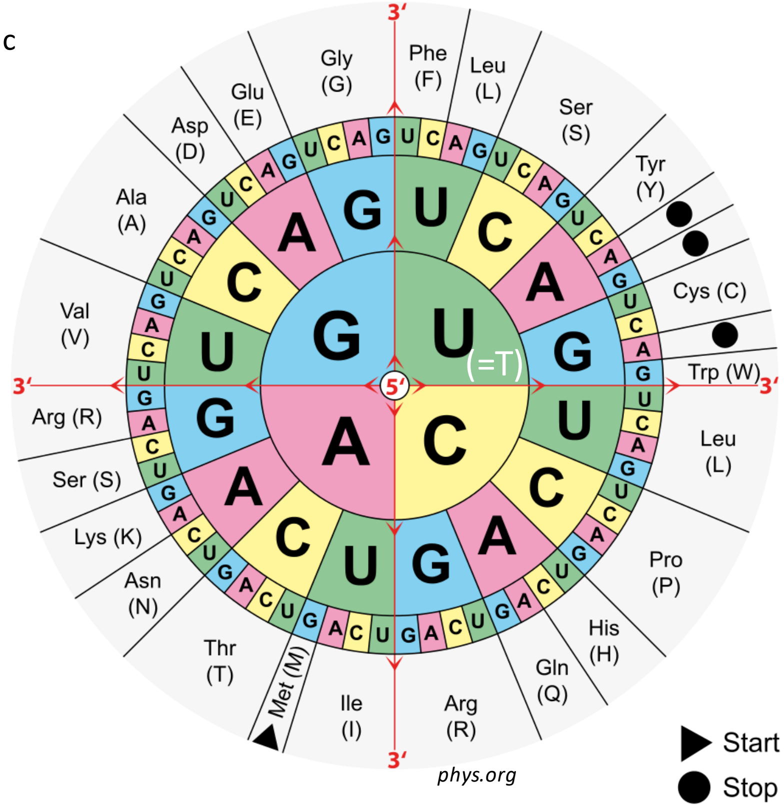
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

↓

ACG = Thr
ACC = Thr
ACT = Thr
ACA = Thr

All mutations at this position will NOT change the amino acid!

Syn sites = 1
Non-Syn sites =2

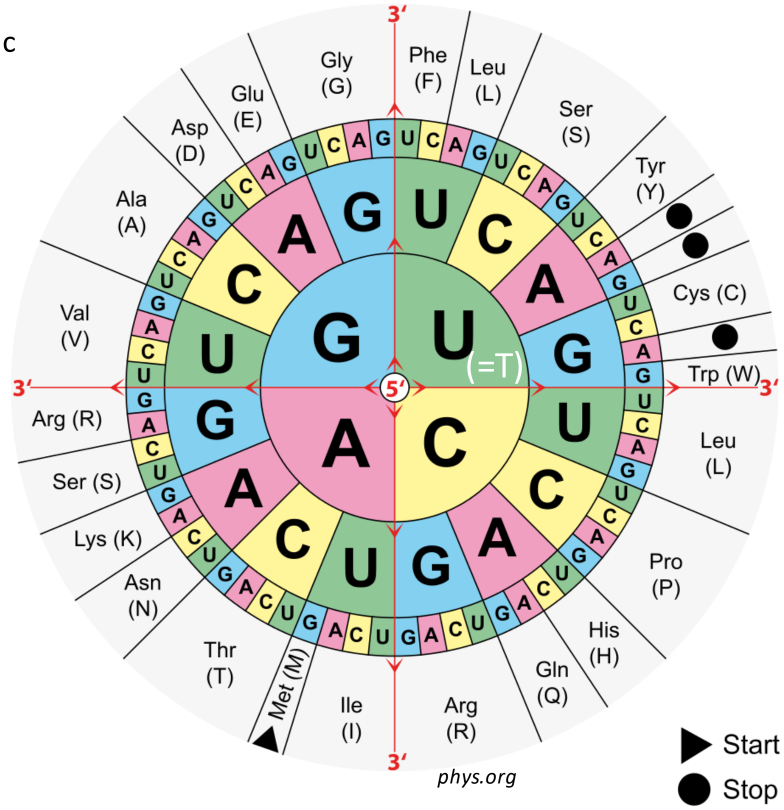
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT ...

↓

TTT = Phe
ATT = Ile
CTT = Leu
GTT = Val

All mutations at this position will change the amino acid!

Syn sites = 1
Non-Syn sites =3

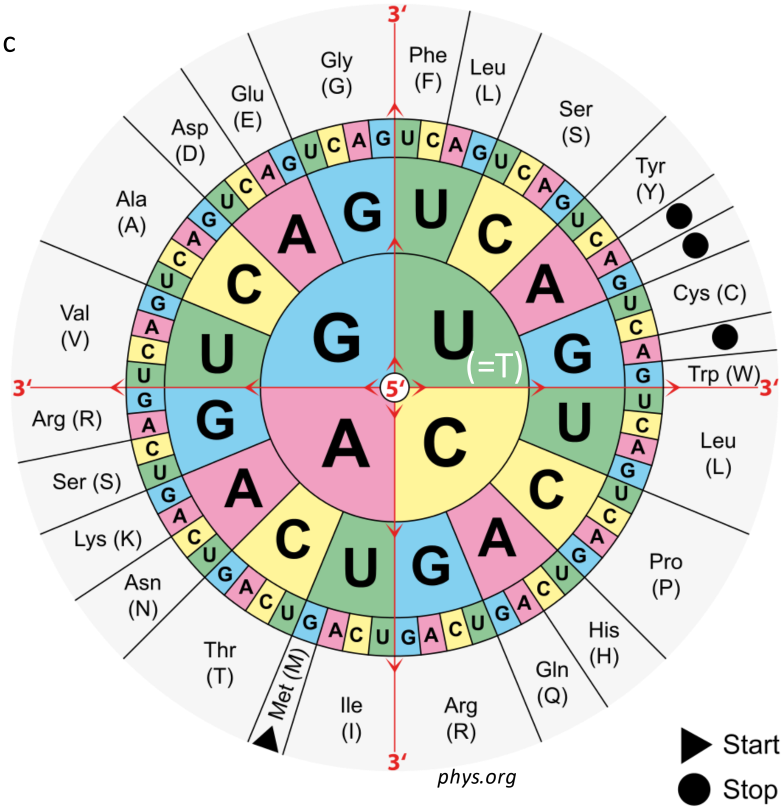
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

↓

TTT = Phe
TAT = Tyr
TCT = Ser
TGT = Cys

All mutations at this position will change the amino acid!

Syn sites = 1
Non-Syn sites =4

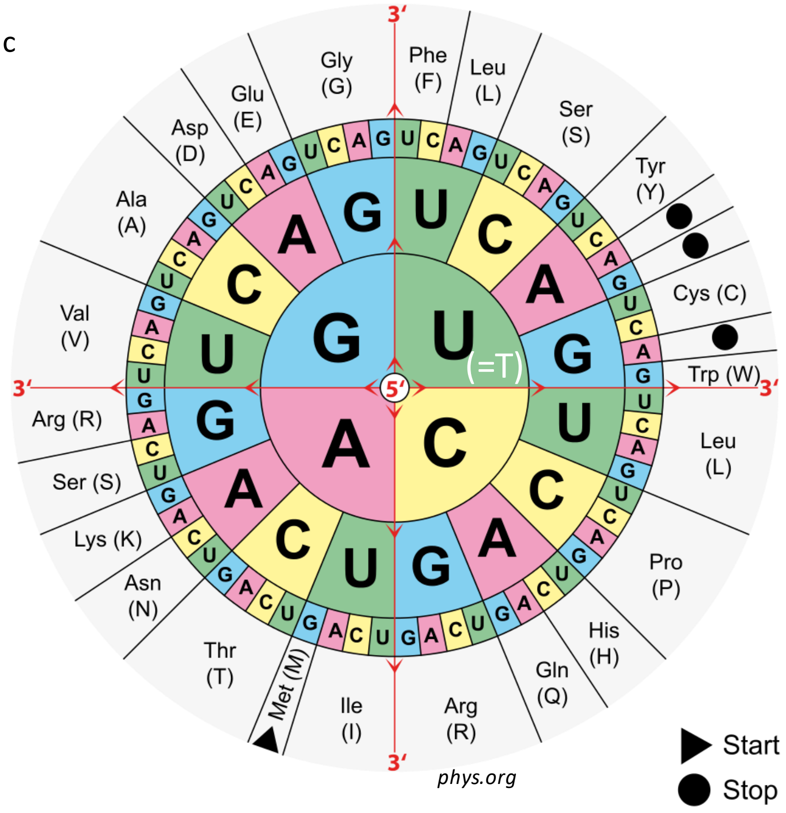
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of nonsynonymous substitutions **per non-synonymous site** (d_N) to the number of synonymous substitutions **per synonymous site** (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous sites:

= which mutations could potentially lead to a synonymous or potentially a non-synonymous change (=expectation)

ACG TTT …

↓

TTT = Phe

TTC = Phe

TTG = Leu

TTA = Leu

2/3 mutations at this position will change the amino acid!

Syn sites = 1.33

Non-Syn sites =4.66

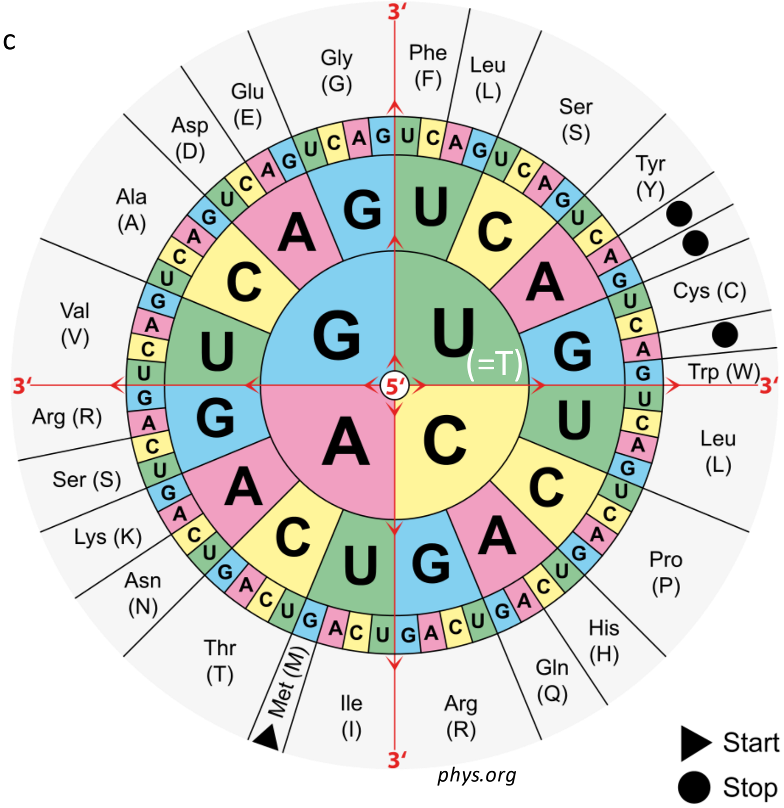
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of **nonsynonymous substitutions** per non-synonymous site (d_N) to the number of **synonymous substitutions** per synonymous site (d_S)

Genetic code (RNA)



Non-synonymous vs. synonymous substitutions:

=observed

Leucine codons:

CTT, CTC, CTA, CTG, TTA, TTG

Genetic variation:

CTT <-> CTA, CTT <-> CTC, CTT -<-> CTG, CTC <-> CTG, TTA <->TTG, CTA <-> TTA, CTG <-> TTG

→ All these mutations will not change the amino acid (synonymous mutations)

These synonymous substitutions **are not affecting the amino acid sequences and are** (assumed to be) **NOT subject to natural selection**

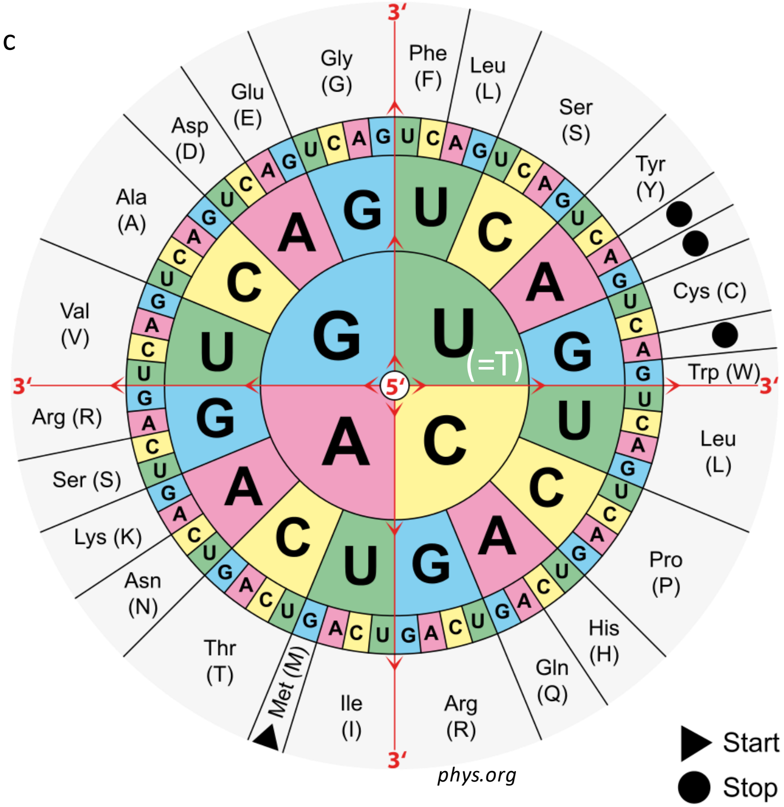
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of **nonsynonymous substitutions** per non-synonymous site (d_N) to the number of **synonymous substitutions** per synonymous site (d_S)

Genetic code (RNA)



phys.org

Non-synonymous vs. synonymous substitutions:

=observed

Any substitutions that causes an amino acid change is a non-synonymous substitution

Genetic variation (e.g.):

TTA ->TTC i.e. Leucine -> Phenylalanine

These synonymous substitutions **change the sequence of the protein sequence and can therefore be subjected to natural selection**

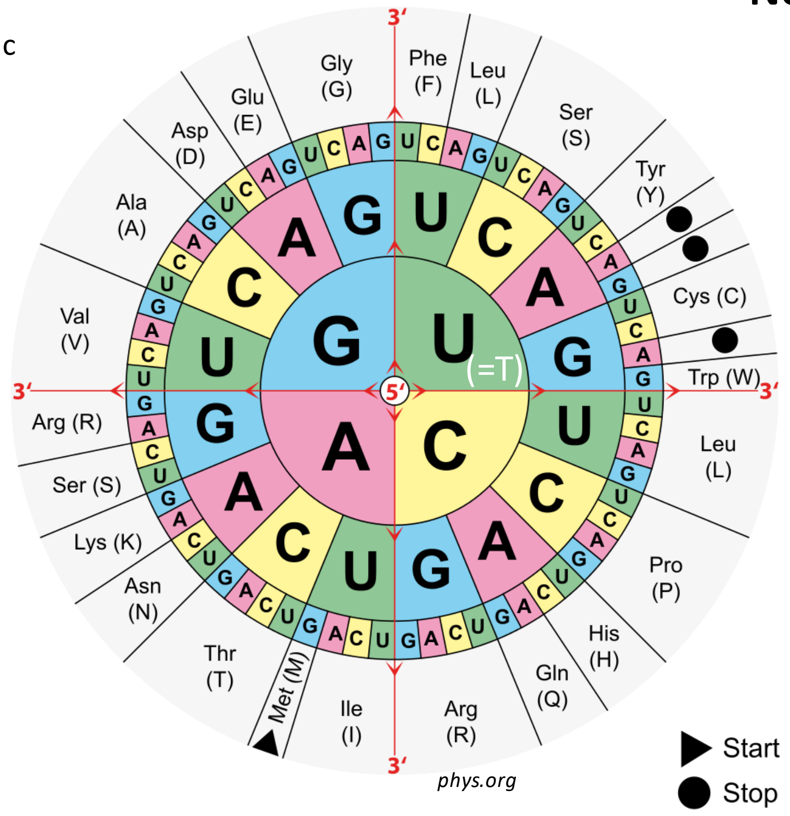
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

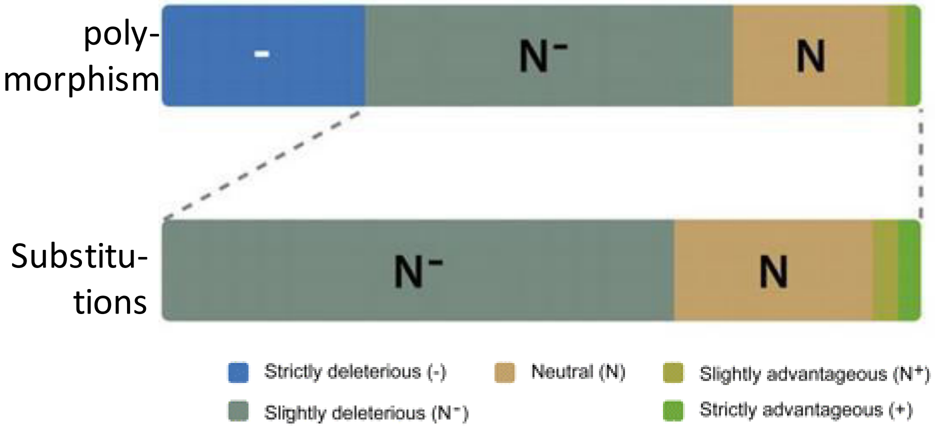
More precisely, this ratio is the number of **nonsynonymous substitutions** per non-synonymous site (d_N) to the number of **synonymous substitutions** per synonymous site (d_S)

Genetic code (RNA)



Non-synonymous vs. synonymous substitutions:

In general, few non-synonymous mutations are adaptive, most mutations on protein-coding genes are either neutral or deleterious



Razeto-Barry et al. 2012 Genetics

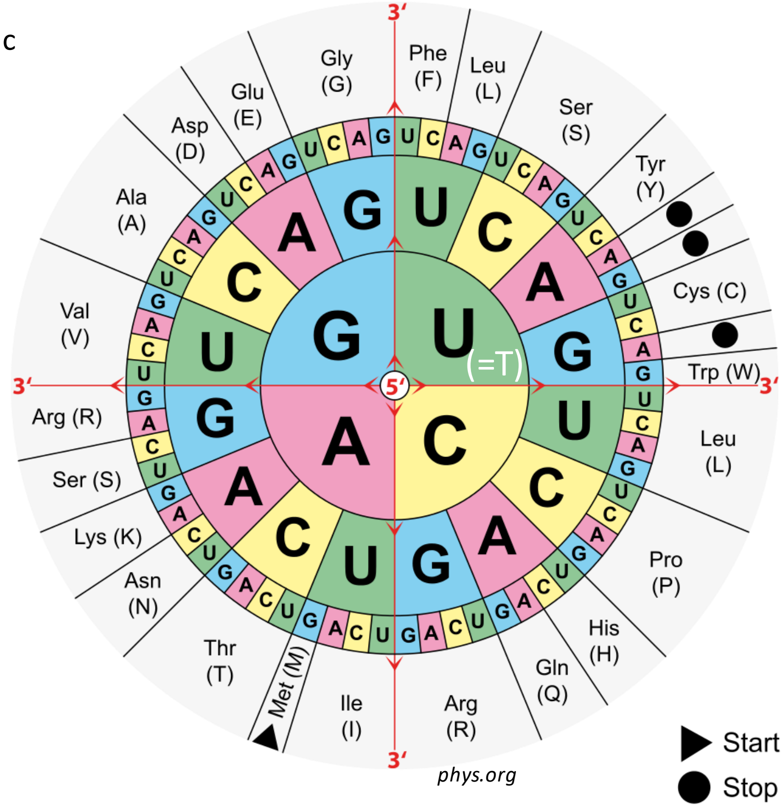
Long-time scales

d_N/d_S ratio

Evolutionary pressures on **proteins** are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (d_N/d_S, also known as K_a/K_s or ω).

More precisely, this ratio is the number of **nonsynonymous substitutions** per non-synonymous site (d_N) to the number of **synonymous substitutions** per synonymous site (d_S)

Genetic code (RNA)



The expectation for the d_N/d_S ratio is then:

$d_N/d_S \sim 1$ **Neutral evolution**

$d_N/d_S < 1$ **Purifying selection (negative selection)**

Non-synonymous mutations are selected **against**

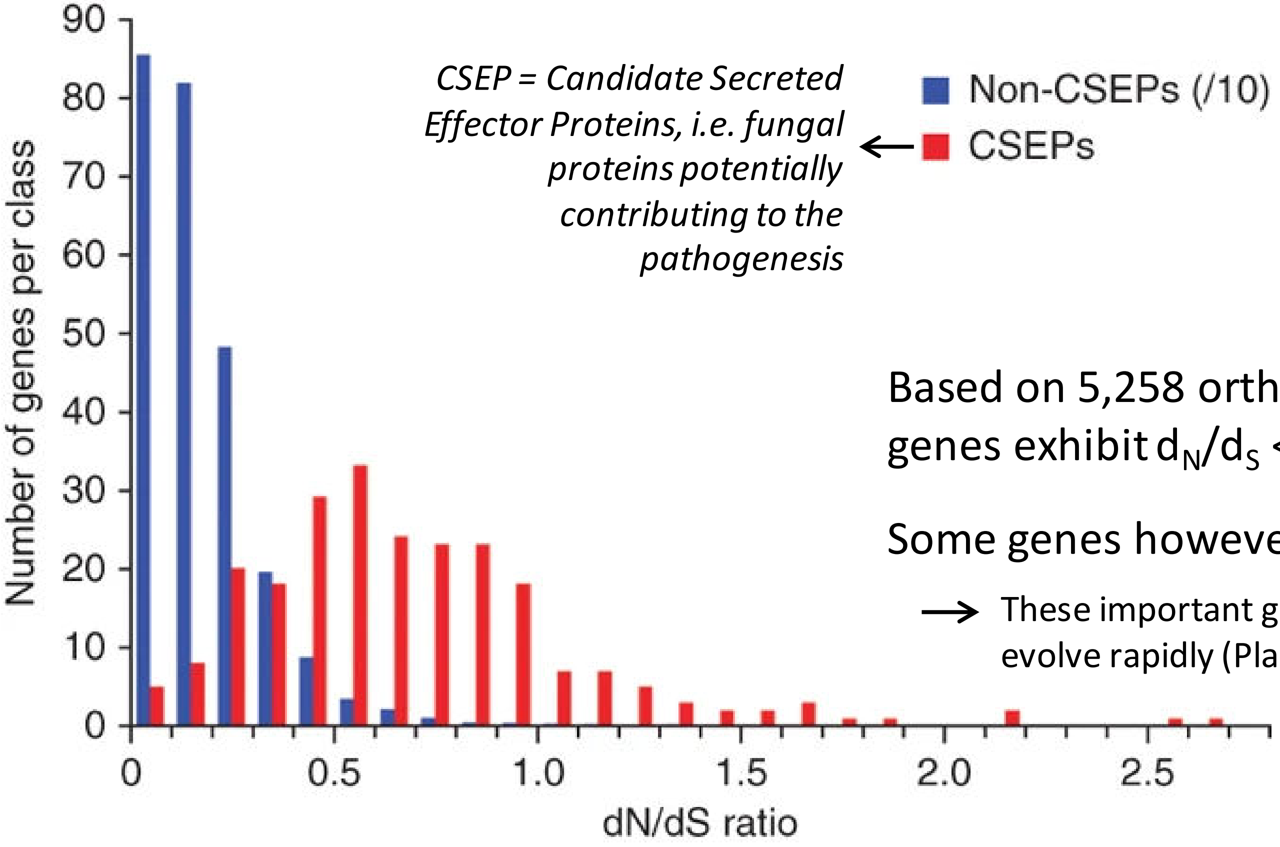
$d_N/d_S > 1$ **Positive selection (advantageous mutations)**

Non-synonymous mutations are selected **for** (at least some)

Long-time scales

d_N/d_S ratio: example

Divergence between two cereal powdery mildews (fungal disease) *Blumeria graminis forma specialis tritici* vs. *Blumeria graminis forma specialis hordei*



CSEP = Candidate Secreted Effector Proteins, i.e. fungal proteins potentially contributing to the pathogenesis ←

Based on 5,258 orthologous genes, most genes exhibit $d_N/d_S << 1$ (average 0.24)

Some genes however exhibit $d_N/d_S>1$

→ These important genes are under selection pressure to evolve rapidly (Plant-Pathogen arms races)

Wicker et al. 2013 Nature Genetics

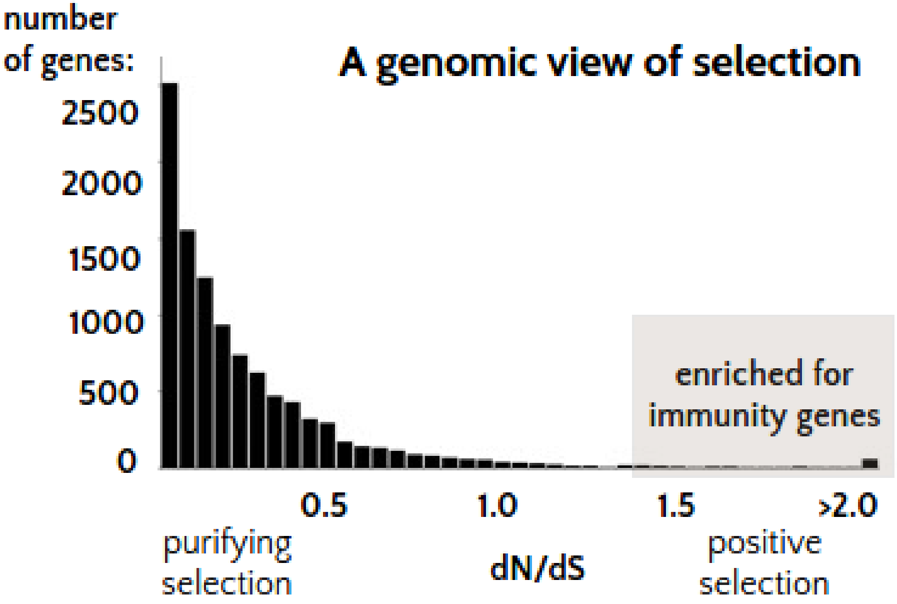
Long-time scales

Human-Chimpanzee d_N/d_S

→ Average d_N/d_S ~ 0.23

→ Genes with $d_N/d_S > 1$ involved in some functions e.g. resistance to pathogens/parasites

Divergence: ~6.5 mya



The histogram above groups genes by dN/dS, the ratio of rates of non-synonymous (dN) and synonymous (dS) codon changes in comparisons between human, chimp, and rhesus. Immunity genes locked in molecular arms races can evolve rapidly under extreme positive selection; dN/dS ›2.

cellvolution.org, Univ. Utah

Long-time scales

McDonald-Kreitman test: background

d_N/d_S is a very conservative test potentially leading to many false negatives

e.g. some mutations were positively selected but the rest of the sequence is strongly constrained. Overall the gene will exhibit $dN/dS \leq 1$

The idea introduced by John H. McDonald & Martin Kreitman is to compare divergence data (i.e. substitutions) with within-species genetic variation (i.e. polymorphisms)

Long-time scales

McDonald-Kreitman test: background

d_N/d_S is a very conservative test potentially leading to many false negatives

e.g. some mutations were positively selected but the rest of the sequence is strongly constrained. Overall the gene will exhibit dN/dS ≤ 1

The idea introduced by John H. McDonald & Martin Kreitman is to compare divergence data (i.e. substitutions) with within-species genetic variation (i.e. polymorphisms)

Following the Neutral Theory, the ratio of non-syn to syn changes is predicted to be roughly constant through time
(*i.e.* ratio within species ~ ratio between species)

Why?

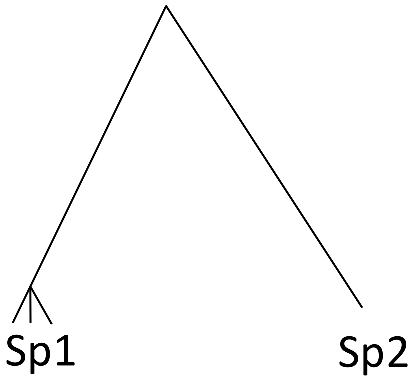
Nonsyn/Syn changes (polymorphism) $= \dfrac{4N\mu_N \sum_{i=1}^{n-1} \frac{1}{i}}{4N\mu_S \sum_{i=1}^{n-1} \frac{1}{i}} = \boxed{\dfrac{\mu_N}{\mu_S}}$

Nonsyn/Syn changes (substitutions) $= \dfrac{2\mu_N t}{2\mu_S t} = \boxed{\dfrac{\mu_N}{\mu_S}}$

Long-time scales

McDonald-Kreitman test: background

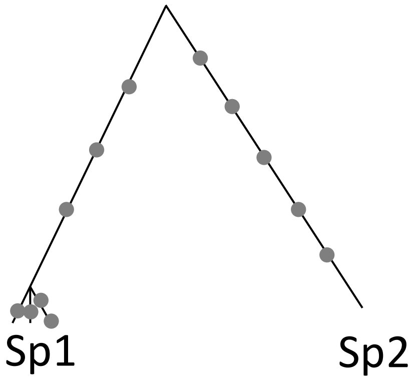
As a consequence we can estimate the ratio from both within (polymorphism) and between species (substitutions). Within-species data provide information about ‘present’ while between species provide information about ‘past divergence’



Long-time scales

McDonald-Kreitman test: background

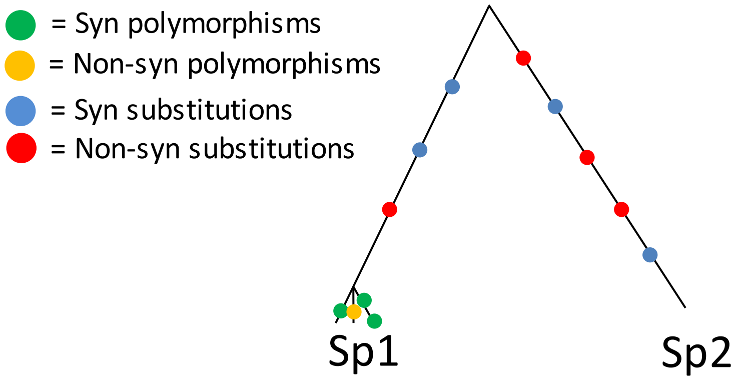
As a consequence we can estimate the ratio from both within (polymorphism) and between species (substitutions). Within-species data provide information about ‘present’ while between species provide information about ‘past divergence’



Long-time scales

McDonald-Kreitman test: background

As a consequence we can estimate the ratio from both within (polymorphism) and between species (substitutions). Within-species data provide information about ‘present’ while between species provide information about ‘past divergence’

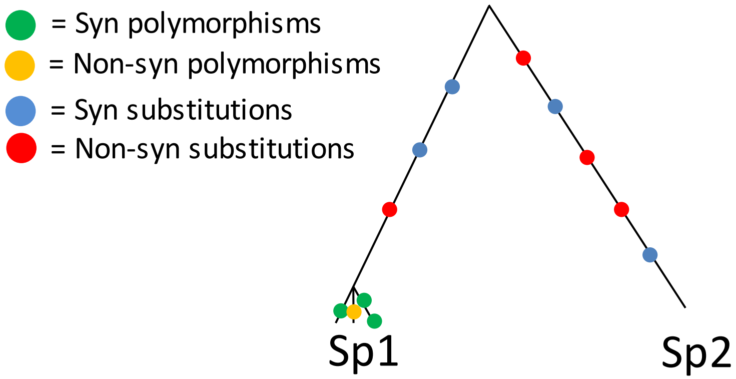


Long-time scales

McDonald-Kreitman test: background

As a consequence we can estimate the ratio from both within (polymorphism) and between species (substitutions). Within-species data provide information about 'present' while between species provide information about 'past divergence'

	substitutions	polymorphisms
Non-syn	D_N	P_N
Syn	D_S	P_S


🟢 = Syn polymorphisms
🟡 = Non-syn polymorphisms
🔵 = Syn substitutions
🔴 = Non-syn substitutions

For a given gene:

D_S: the number of synonymous substitutions 🔵

D_N: the number of non-synonymous substitutions 🔴

P_S: the number of synonymous polymorphisms 🟢

P_N: the number of non-synonymous polymorphisms 🟡

Interpretation:

$D_N/D_S = P_N/P_S$ -> consistent with neutrality

$D_N/D_S > P_N/P_S$ -> more nonsyn changes between species (positive selection)

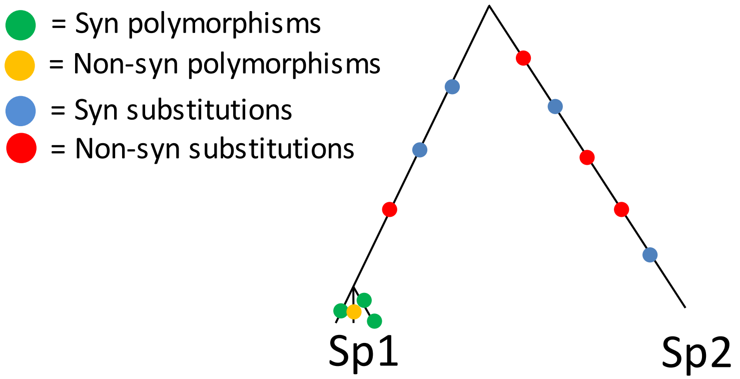
$D_N/D_S < P_N/P_S$ -> less nonsyn changes between species (negative selection)

Long-time scales

McDonald-Kreitman test: background

As a consequence we can estimate the ratio from both within (polymorphism) and between species (substitutions). Within-species data provide information about ‘present’ while between species provide information about ‘past divergence’

	substitutions	polymorphisms
Non-syn	D_N: 4* 🔴	P_N: 1* 🟡
Syn	D_S: 4* 🔵	P_S: 3* 🟢
	D_N/D_S =1	P_N/P_S =1/3



For a given gene:

D_S: the number of synonymous substitutions 🔵

D_N: the number of non-synonymous substitutions 🔴

P_S: the number of synonymous polymorphisms 🟢

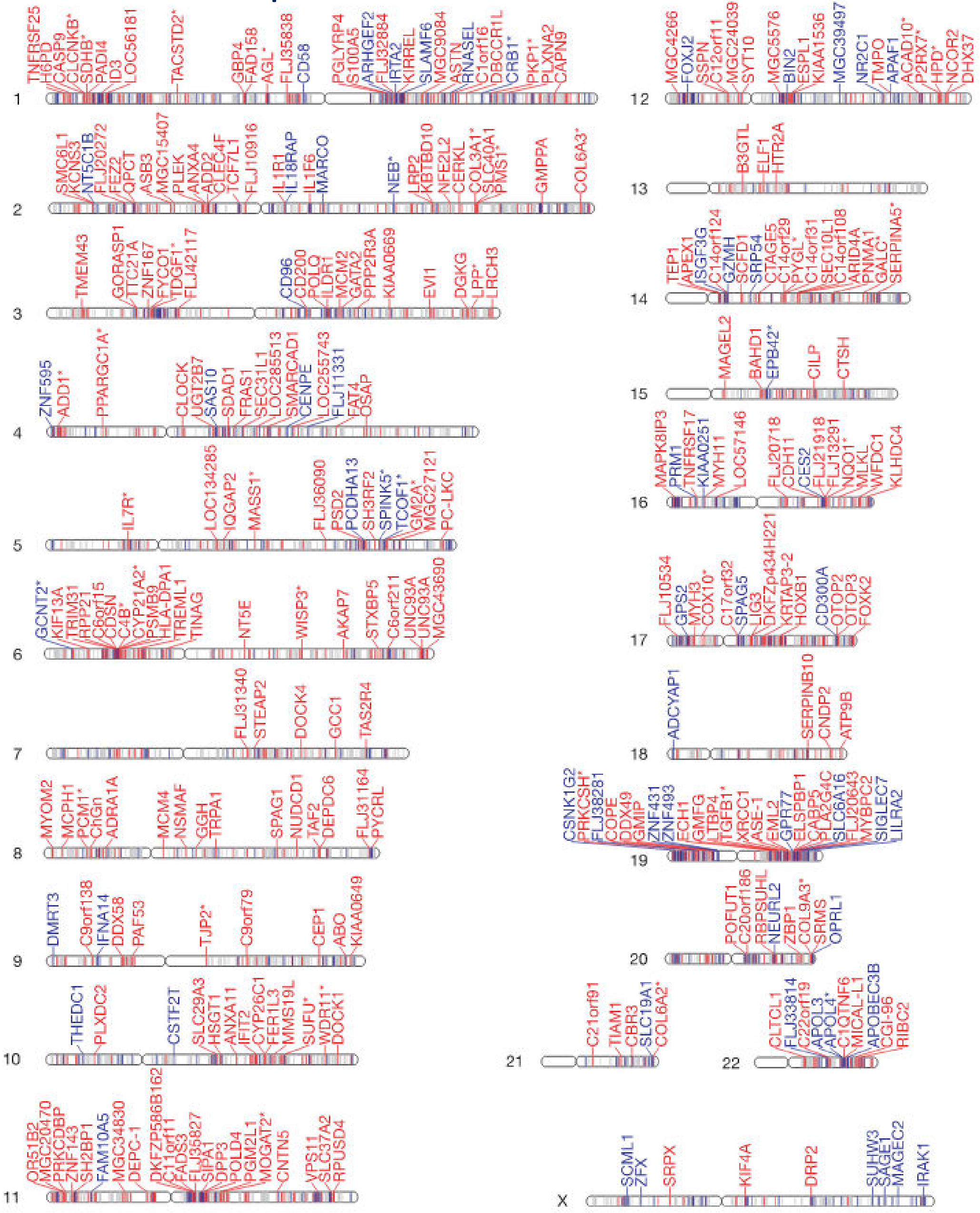
P_N: the number of non-synonymous polymorphisms 🟡

$$\mathbf{D_N/D_S > P_N/P_S}$$

Then contingency tests based on these 2x2 tables can be performed to test the significance (such as chi-squared tests)

Long-time scales

McDonald-Kreitman test: example



- Human-Chimp comparison (39 humans, 1 chimp, 11,000 genes)
- 304 genes with evidence of positive selection (blue)

  ‘a small minority of non-neutral genes are facing positive selection’
- 813 genes with evidence of negative selection (red)

Bustamante et al. 2005 Nature

Summary (long-time scales only)

d_N/d_S and MK tests use sequence data from divergent taxa allowing to identify genes with a lot of non-synonymous substitutions that were selected for (*i.e.* positive selection)

Tests can be performed on some candidate proteins (e.g. one or few genes with a specific function) or to scan all genes of a given species to identify genes that were under selection

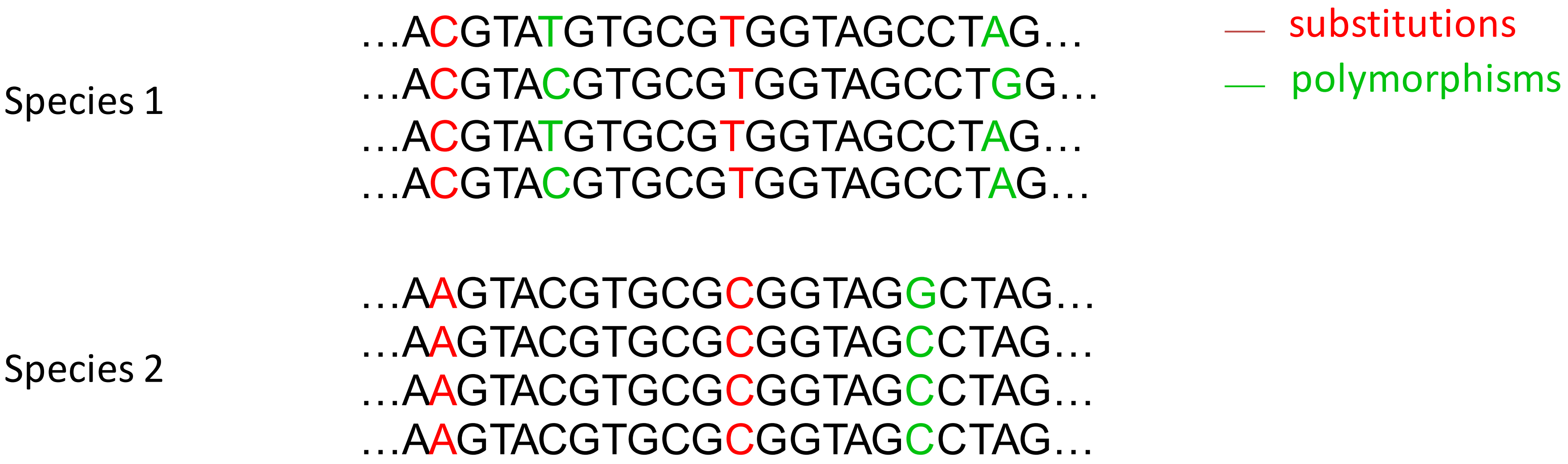
In the vast majority of species, the proportion of genes exhibiting signatures of positive selection is low, at least as compared to those evolving under negative selection, consistent with the general hypothesis of a strong evolutionary constraint on proteins

Extensions of the MK test over the last two decades to take into account short-term demographic variation and the presence of slightly deleterious mutations
(e.g. Moutinho et al. 2019 *Evolutionary Ecology* for a review)

Genetic basis of adaptive evolution, an important topic in evolutionary biology!

Different methods depending on the levels of divergence:

Long-time scales	Short-time scales
Different species (divergence)	Different populations
Substitutions	Polymorphisms
Individual-level data	Population-level data
Protein-coding sequences	Whole genome sequences (if possible)



Species 1

…ACGTATGTGCGTGGTAGCCTAG…
…ACGTACGTGCGTGGTAGCCTGG…
…ACGTATGTGCGTGGTAGCCTAG…
…ACGTACGTGCGTGGTAGCCTAG…

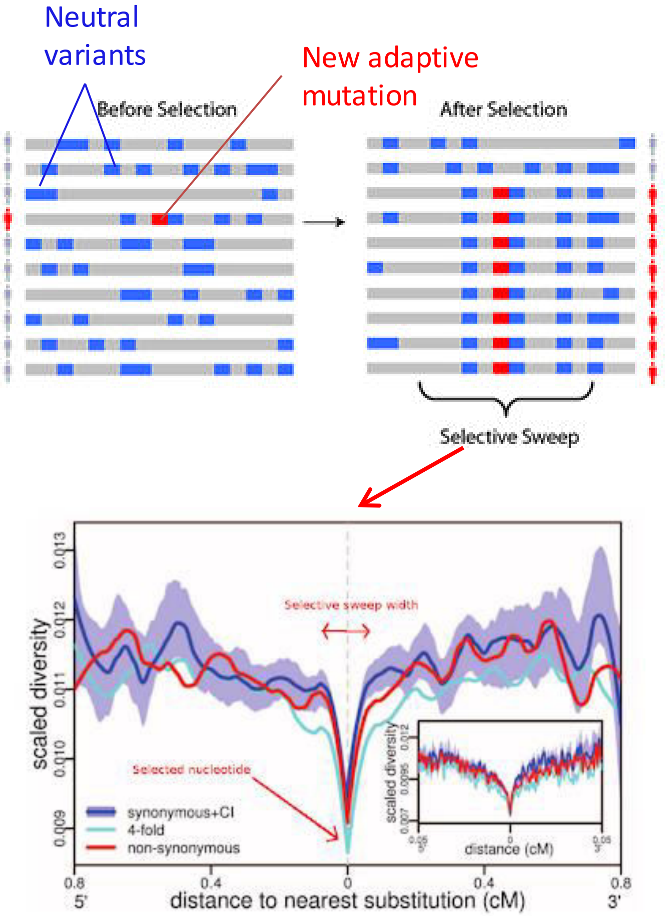
Species 2

…AAGTACGTGCGCGGTAGGCTAG…
…AAGTACGTGCGCGGTAGCCTAG…
…AAGTACGTGCGCGGTAGCCTAG…
…AAGTACGTGCGCGGTAGCCTAG…

— substitutions
— polymorphisms

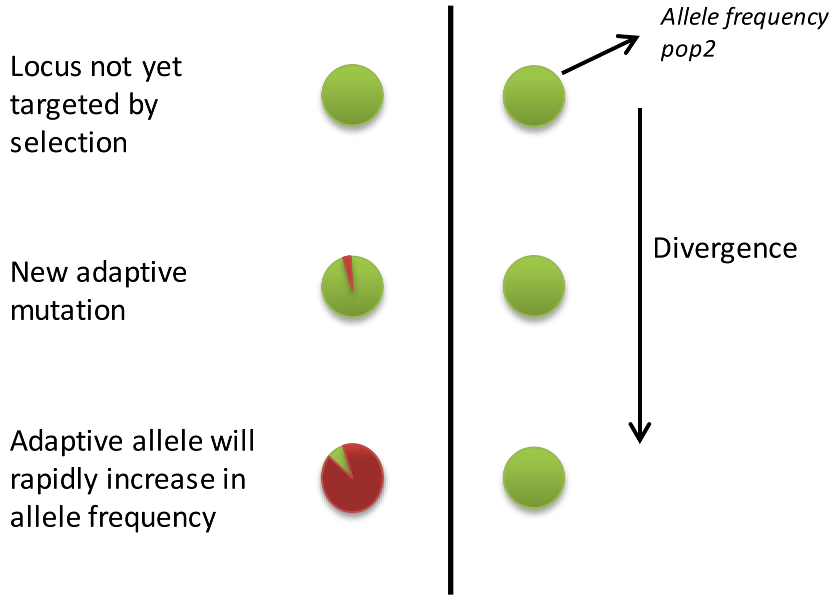
Short-time scales, methods are divided into two main groups:

Selective sweeps
(within-population variation)



Reduction of the diversity at the selected locus (+ its linked neutral variants)

Genetic differentiation
(between populations)



Locus not yet targeted by selection

New adaptive mutation

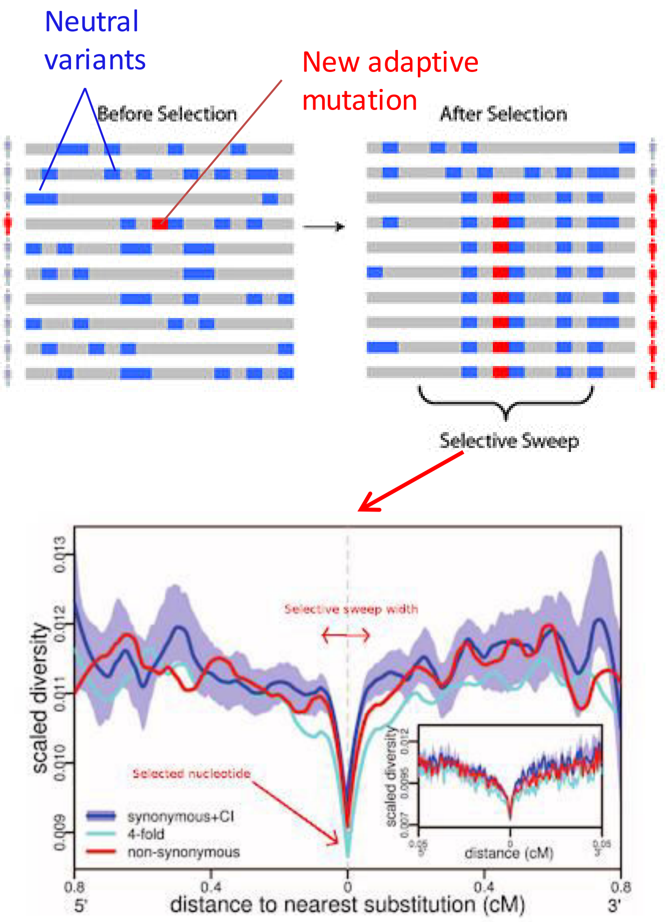
Adaptive allele will rapidly increase in allele frequency

Extreme allele frequency differences between the two populations at the selected locus

SNP in close vicinity to the targeted SNPs also exhibit strong differences in allele frequency

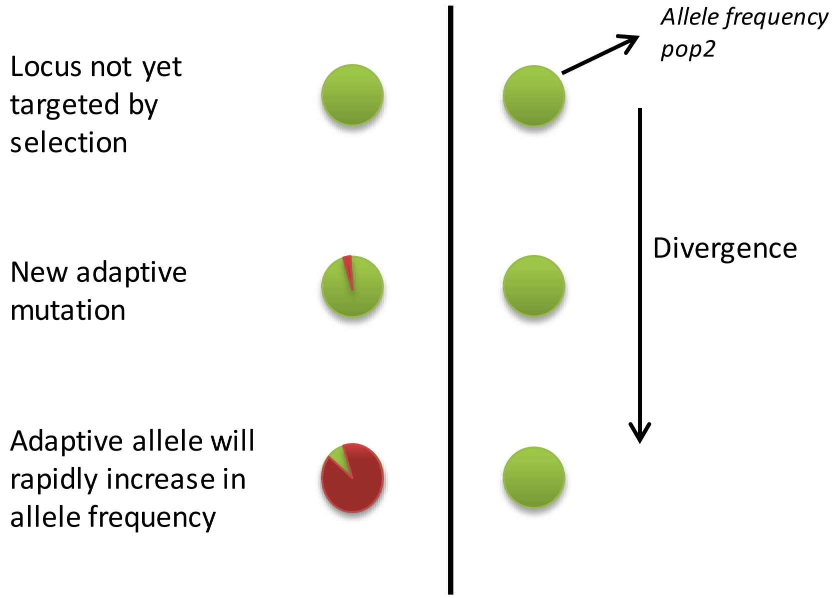
Short-time scales, methods are divided into two main groups:

Selective sweeps
(within-population variation)



Reduction of the diversity at the selected locus (+ its linked neutral variants)

Genetic differentiation
(between populations)



Locus not yet targeted by selection

New adaptive mutation

Adaptive allele will rapidly increase in allele frequency

Allele frequency pop2

Divergence

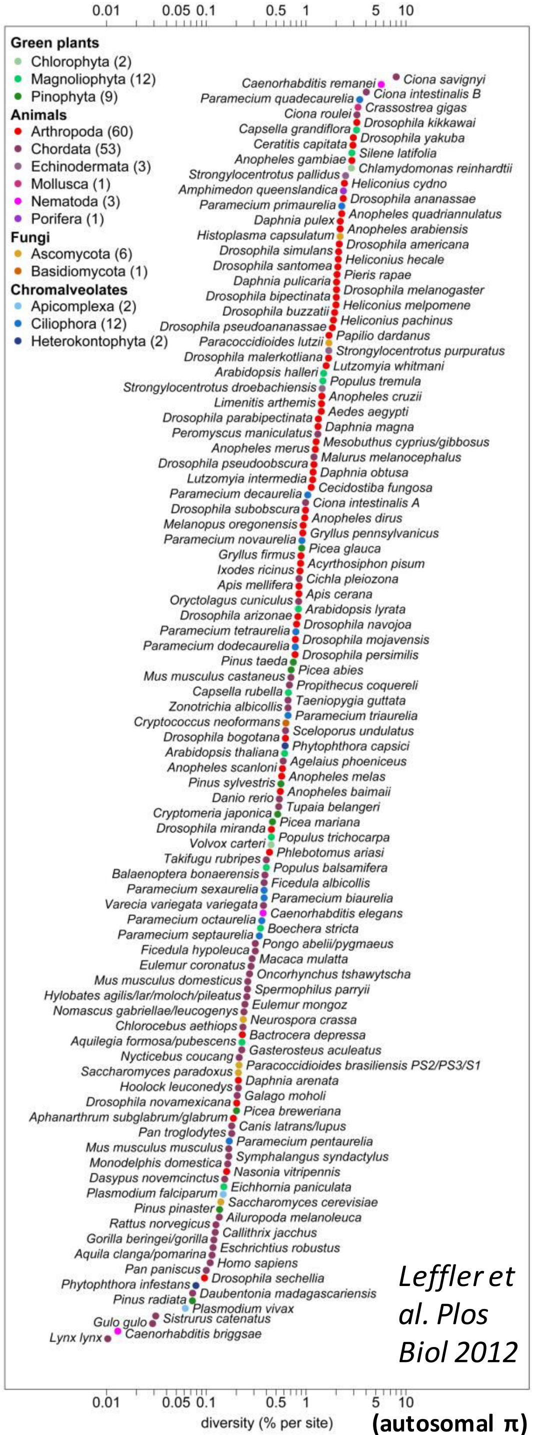
Extreme allele frequency differences between the two populations at the selected locus

SNP in close vicinity to the targeted SNPs also exhibit strong differences in allele frequency

Nucleotide diversity indices (a reminder!)

Genetic diversity is highly variable among the tree of life!

Species with large population sizes or elevated mutation rates exhibit higher genetic diversity (=4Neμ)



Leffler et al. Plos Biol 2012

(autosomal π)

Nucleotide diversity indices and Tajima’s D

Genetic diversity is highly variable among the tree of life!

Species with large population sizes or elevated mutation rates exhibit higher genetic diversity (=4Neµ)

Two different measures:

- Average number of differences between pairs of sequences => π
- Total number of segregating sites (S) => S/harmonic number => θ

1:AGATCGCTGCAAT
2:AGATCGCTTCAAT
3:AGATCGCTTCAAT
4:AGATCGCTTCGAT
5:AGATCGCTTCGAG

At equilibrium (constant population size), we expect θ = π
=> Tajima’s D = π – θ = 0

Nucleotide diversity indices and Tajima's D

Genetic diversity is highly variable among the tree of life!

Species with large population sizes or elevated mutation rates
exhibit higher genetic diversity (=4Neμ)

Two different measures:

- Average number of differences between pairs of sequences => π
- Total number of segregating sites (S) => S/harmonic number => θ

1:TCATCGCTGCAAT
2:TCATCGCTTCAAT
3:TCATCGCTTCAAT
4:TCATCGCTTCGAT
5:TCATCGCTTCGAG

S=3; Harmonic number= $\sum_{i=1}^{n-1} \frac{1}{i}$ $\Rightarrow 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$

θ=S/Harmonic number=3/2.083=1.44

Pairwise number of differences:
1vs.2 = 1; 1vs.3=1; 1vs.4=2; 1vs.5=3; 2vs.3 =0; 2vs.4=1;
2vs.5=2; 3vs.4=1; 3vs.5=2; 4vs.5=1
Average: 1.4 per sequence (1.4/13 => 0.11 per base pair)

At equilibrium (constant population size), we expect θ = π
=> Tajima's D = π – θ = 0

Nucleotide diversity indices and Tajima's D

Genetic diversity is highly variable among the tree of life!

Species with large population sizes or elevated mutation rates
exhibit higher genetic diversity (=4Neμ)

Two different measures:

- Average number of differences between pairs of sequences => π
- Total number of segregating sites (S) => S/harmonic number => θ

S=3; Harmonic number= $\sum_{i=1}^{n-1} \frac{1}{i}$ $\Rightarrow 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$

1:AAATACCAACAAC
2:AAATACCATCAAC
3:AAATACCATCAAG
4:AAATACCATCAAC
5:AAATACCATCGAC

θ=S/Harmonic number=3/2.083=1.44

Pairwise number of differences:
1vs.2 = 1; 1vs.3=2; 1vs.4=1; 1vs.5=2; 2vs.3 =1; 2vs.4=0; 2vs.5=1; 3vs.4=1; 3vs.5=2; 4vs.5=1
Average: 1.2 per sequence (*i.e.* 1.2/13 => 0.09 per base pair)

At equilibrium (constant population size), we expect θ = π
=> Here θ > π; Tajima's D < 0 **Excess of rare alleles** as compared to the expectation!

Nucleotide diversity indices and Tajima's D

Genetic diversity is highly variable among the tree of life!

Species with large population sizes or elevated mutation rates exhibit higher genetic diversity (=4Neµ)

Two different measures:

- Average number of differences between pairs of sequences => π
- Total number of segregating sites (S) => S/harmonic number => θ

S=3; Harmonic number= $\sum_{i=1}^{n-1} \frac{1}{i}$ $\Rightarrow 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$

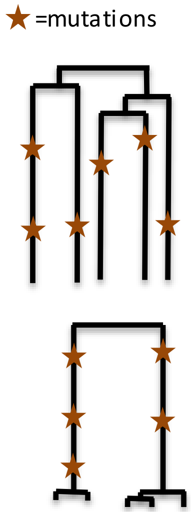
1:AGATCGCTCCAAG
2:AGATCGCTCCTAA
3:AGATCGCTACTAA
4:AGATCGCTACAAA
5:AGATCGCTACAAG

θ=S/Harmonic number=3/2.083=1.44

Pairwise number of differences:
1vs.2 = 2; 1vs.3=3; 1vs.4=2; 1vs.5=1; 2vs.3 =1; 2vs.4=2; 2vs.5=3; 3vs.4=1; 3vs.5=2; 4vs.5=1
Average: 1.8 per sequence (*i.e.* 1.8/13 => 0.14 per base pair)

At equilibrium (constant population size), we expect θ = π
=> Here θ < π; Tajima's D > 0 **Deficit of rare alleles** as compared to the expectation!

How to interprete Tajima's D deviations?



	Demographic effects	Selection
D<0 **(=excess of rare alleles)**	Population expansion	Recent selective sweep (i.e. effect of an advantageous allele)
D>0 **(=deficit of rare alleles)**	Bottleneck (i.e. sudden population contraction)	Balancing selection (i.e. multiple alleles are maintained)

Demographic effects are expected to similarly affect the whole genome (i.e. most genes show consistent deviations from D=0), while selection affect some specific genes

How to interprete Tajima’s D deviations?

Ex. African rice

Oryza barthii (Wild ancestor)

Domestication →

Oryza glaberrima (domesticated species)

X 23 individuals from the centre of domestication

X 25 individuals

For each species, I computed θ, π and Tajima’s D for all 100 kb sliding windows spanning the 12 *Oryza* chromosomes

Leroy & Rougemont, in press

How to interprete Tajima's D deviations?

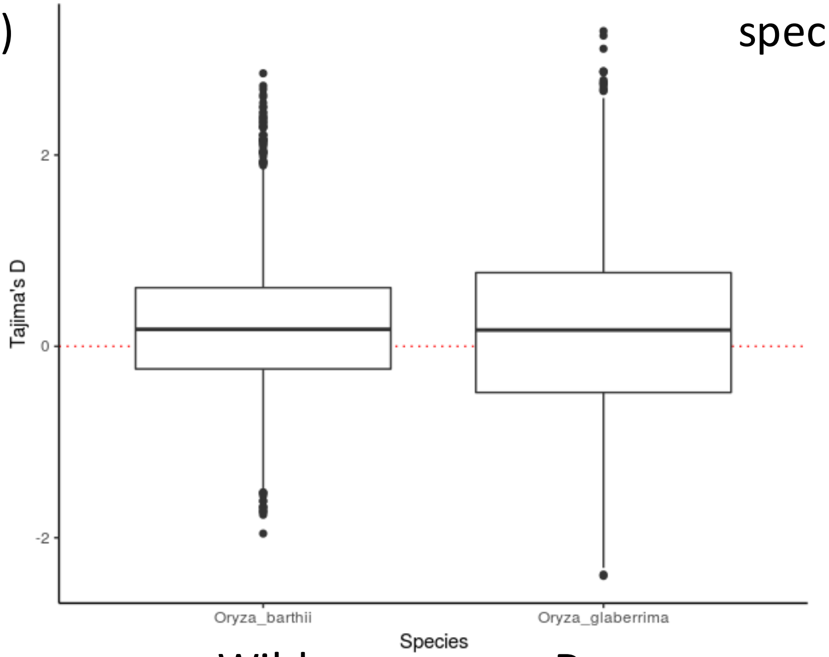
Ex. African rice

Oryza barthii (Wild ancestor)

Domestication →

Oryza glaberrima (domesticated species)

Tajima's D



Wild

Domesticated

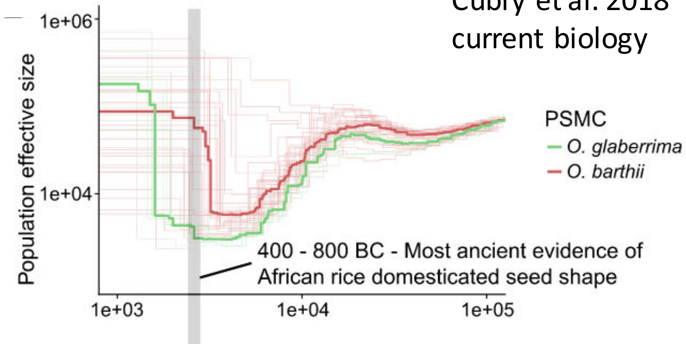
↑ **Most genomic windows exhibit slightly positive Tajima's D values (both species)**

=> Demographic effect

Leroy & Rougemont, in press

How to interprete Tajima's D deviations?

Ex. African rice


1e+06
1e+04
Population effective size
1e+03
1e+04
1e+05
PSMC
O. glaberrima
O. barthii
400 - 800 BC - Most ancient evidence of African rice domesticated seed shape

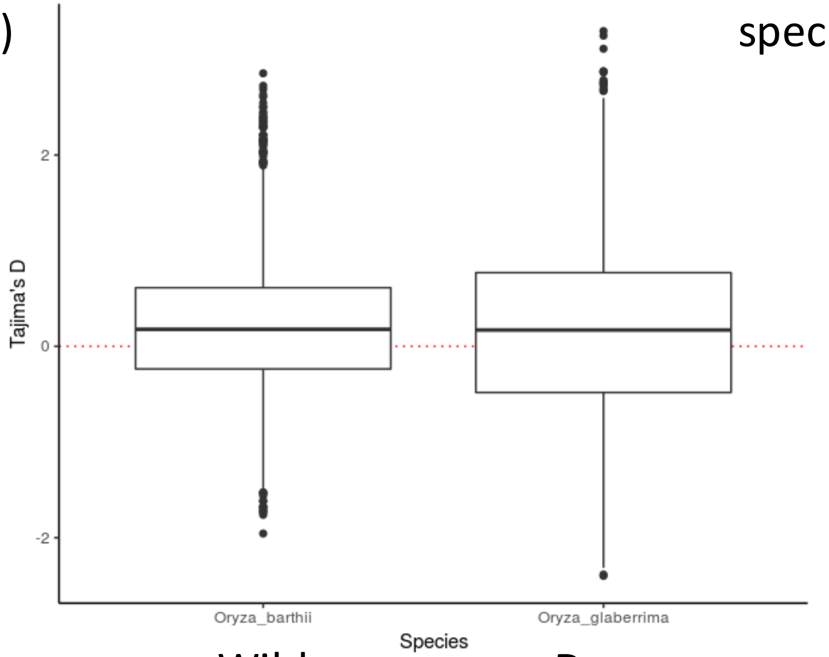
Cubry et al. 2018 current biology

Oryza barthii (Wild ancestor)

Domestication →

Oryza glaberrima (domesticated species)

Tajima's D



Wild

Domesticated

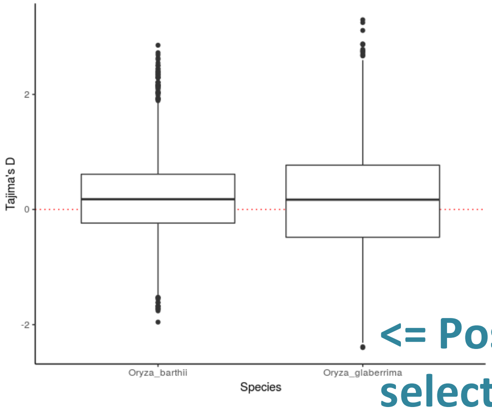
↑ **Most genomic windows exhibit slightly positive Tajima's D values (both species)**

=> Demographic effect

Leroy & Rougemont, in press

How to interprete Tajima's D deviations?

Demographic effect: 'the core of the distribution'

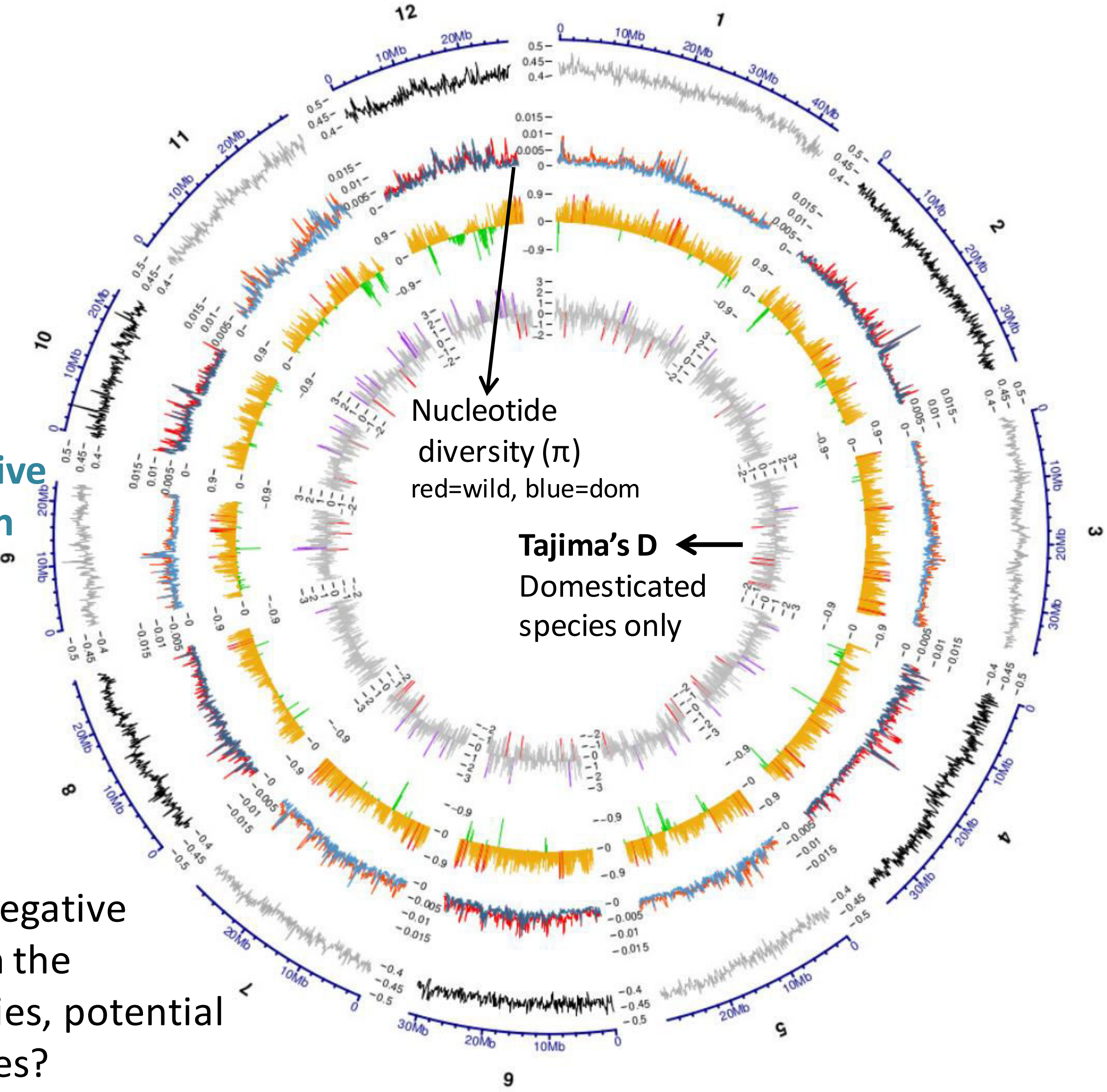


<= Positive selection

Selection: 'the outliers'!

In practice, we often use a simple rule, +2/-2 to identify 'potential selected genes'

→ Some genes with negative Tajima's D values in the domesticated species, potential domestication genes?

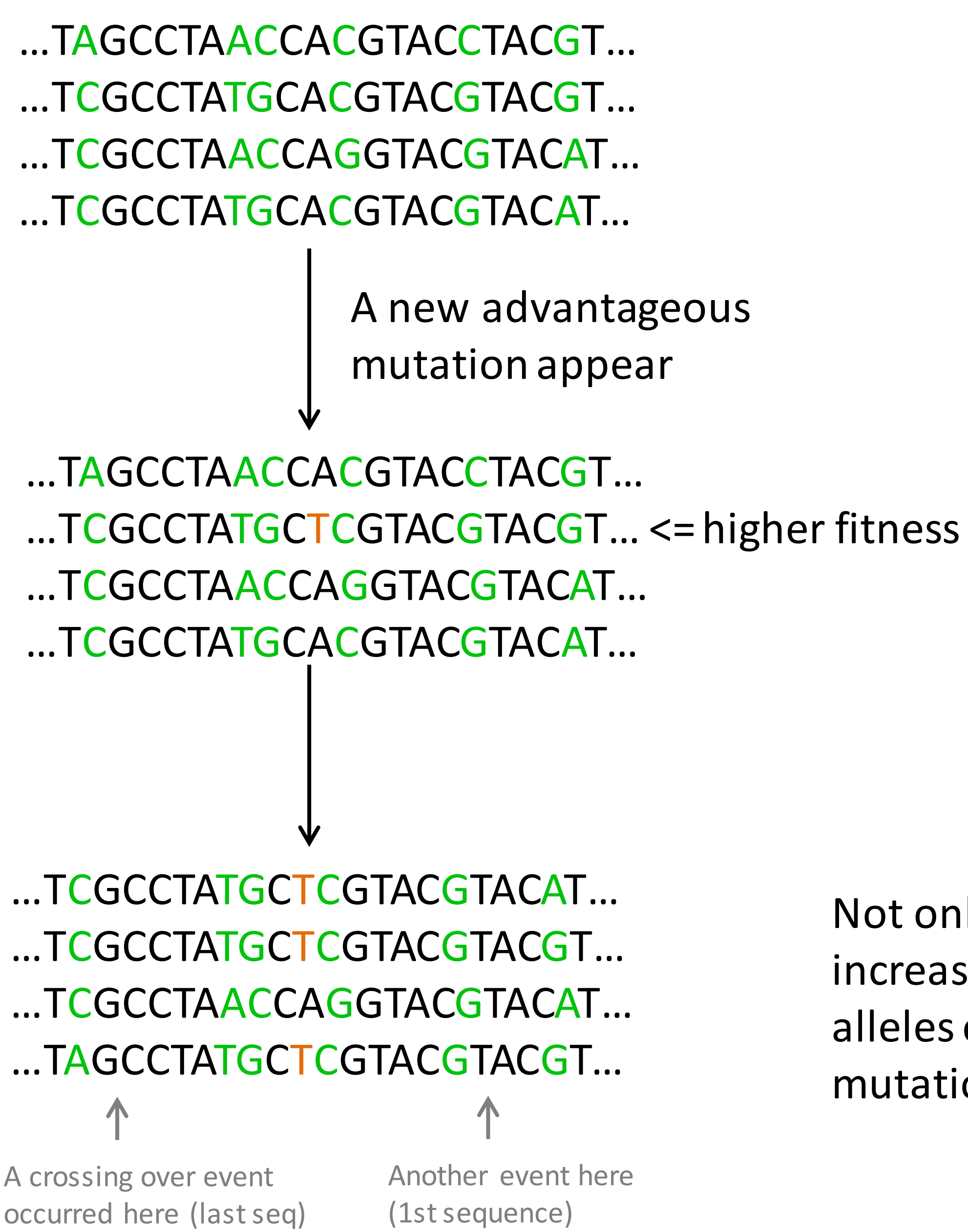


Nucleotide diversity (π)
red=wild, blue=dom

Tajima's D
Domesticated species only

Leroy & Rougemont, in press

Why advantageous alleles generate regions of low diversity?



...TAGCCTAACCACGTACCTACGT...
...TCGCCTATGCACGTACGTACGT...
...TCGCCTAACCAGGTACGTACAT...
...TCGCCTATGCACGTACGTACAT...

A new advantageous
mutation appear

...TAGCCTAACCACGTACCTACGT...
...TCGCCTATGCTCGTACGTACGT... <= higher fitness
...TCGCCTAACCAGGTACGTACAT...
...TCGCCTATGCACGTACGTACAT...

...TCGCCTATGCTCGTACGTACAT...
...TCGCCTATGCTCGTACGTACGT...
...TCGCCTAACCAGGTACGTACAT...
...TAGCCTATGCTCGTACGTACGT...

A crossing over event
occurred here (last seq)

Another event here
(1st sequence)

Not only the beneficial mutation increase in frequency, but also alleles of this individual near the mutation!

Why advantageous alleles generate regions of low diversity?

...TAGCCTAACCACGTACCTACGT...
...TCGCCTATGCACGTACGTACGT...
...TCGCCTAACCAGGTACGTACAT...
...TCGCCTATGCACGTACGTACAT...

↓ A new advantageous mutation appear

...TAGCCTAACCACGTACCTACGT...
...TCGCCTATGCTCGTACGTACGT... <= higher fitness
...TCGCCTAACCAGGTACGTACAT...
...TCGCCTATGCACGTACGTACAT...

↓

...TCGCCTATGCTCGTACGTACGT...
...TCGCCTATGCTCGTACGTACAT...
...TCGCCTATGCTCGTACCTACAT...
...TAGCCTATGCTCGTACGTACGT...

Until fixation!

Why advantageous alleles generate regions of low diversity?

…TAGCCTAACCACGTACCTACGT…
…TCGCCTATGCACGTACGTACGT…
…TCGCCTAACCAGGTACGTACAT…
…TCGCCTATGCACGTACGTACAT…

A new advantageous
mutation appear

…TAGCCTAACCACGTACCTACGT…
…TCGCCTATGCTCGTACGTACGT… <= higher fitness
…TCGCCTAACCAGGTACGTACAT…
…TCGCCTATGCACGTACGTACAT…

…TCGCCTATGCTCGTACGTACGT…
…TCGCCTATGCTCGTACGTACAT…
…TCGCCTATGCTCGTACCTACAT…
…TAGCCTATGCTCGTACGTACGT…

Until fixation!

Why advantageous alleles generate regions of low diversity?

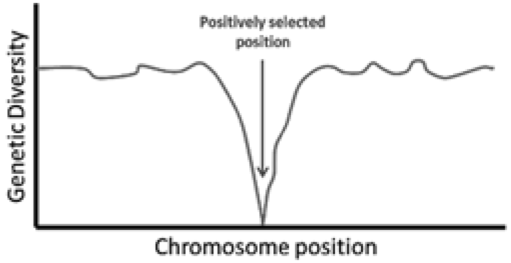
…TAGCCTAACCACGTACCTACGT…
…TCGCCTATGCACGTACGTACGT…
…TCGCCTAACCAGGTACGTACAT…
…TCGCCTATGCACGTACGTACAT…

Before

…TCGCCTATGCTCGTACGTACAT…
…TCGCCTATGCTCGTACGTACGT…
…TCGCCTATGCTCGTACCTACAT…
…TAGCCTATGCTCGTACGTACGT…

After

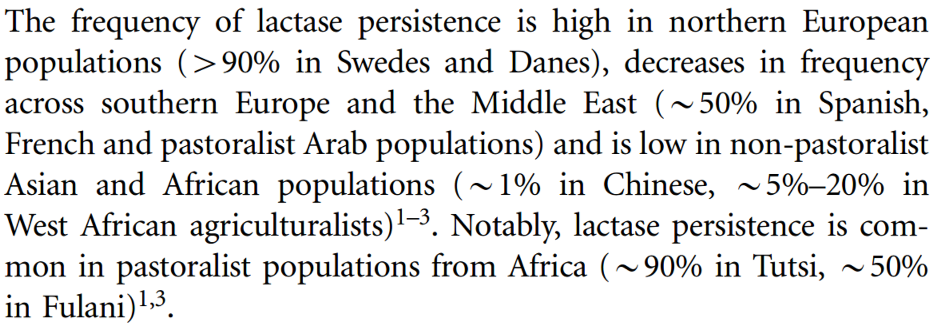
→ Reduced levels of nucleotide diversity around the advantageous allele + excess of rare alleles (*i.e.* D<0)
(a selective sweep)



The extent of the selective sweep depends on the balance between the intensity of natural selection ('how advantageous is the allele') and the local recombination rate

Example of selective sweeps in humans

Lactase persistence = ability to digest milk as adults in humans



The frequency of lactase persistence is high in northern European populations (>90% in Swedes and Danes), decreases in frequency across southern Europe and the Middle East (~50% in Spanish, French and pastoralist Arab populations) and is low in non-pastoralist Asian and African populations (~1% in Chinese, ~5%–20% in West African agriculturalists)[1–3]. Notably, lactase persistence is common in pastoralist populations from Africa (~90% in Tutsi, ~50% in Fulani)[1,3].

Long tracks without genetic variations in lactase-persistent individuals (selective sweep to continue to digest milk)

This is an example (among few) of a selective sweep detected in humans ('a hard sweep')

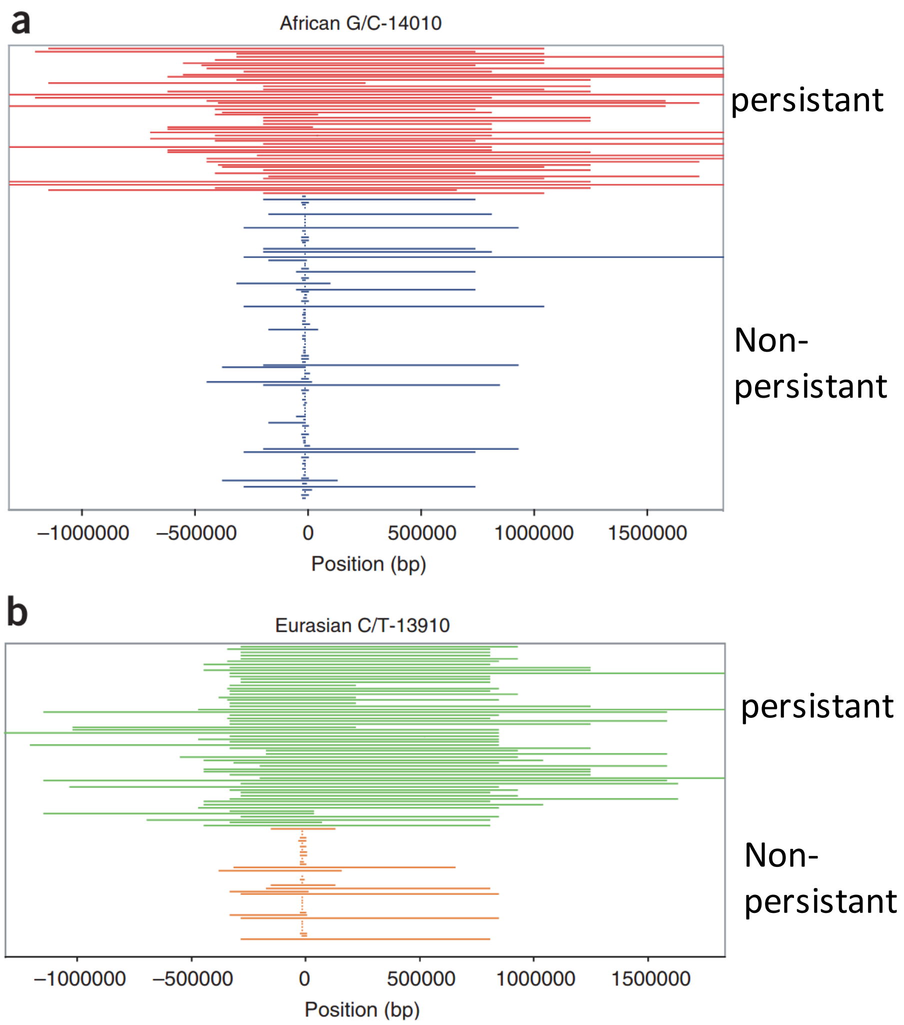
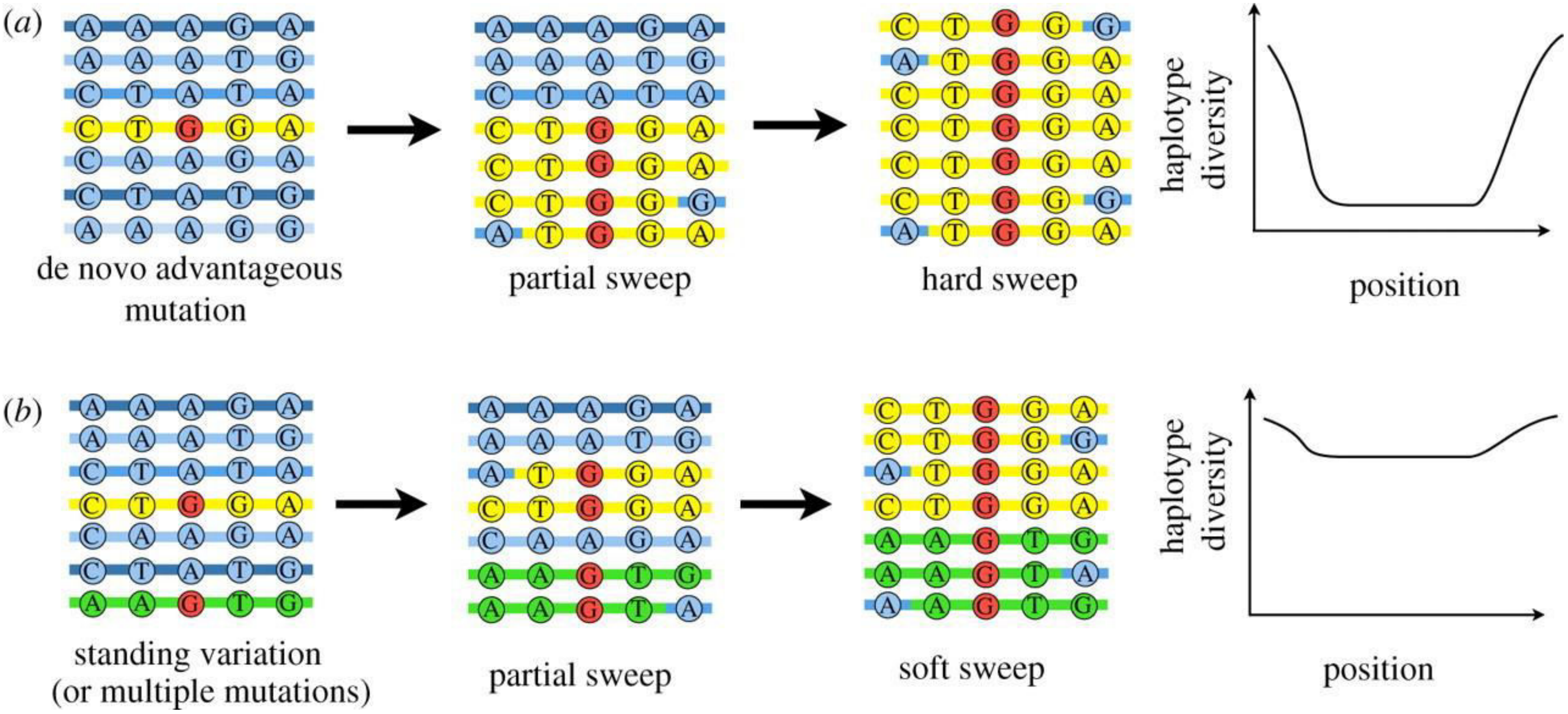


Figure 6 Comparison of tracts of homozygous genotypes flanking the lactase persistence–associated SNPs. (**a**) Kenyan and Tanzanian C-14010 lactase-persistent (red) and non-persistent G-14010 (blue) homozygosity tracts. (**b**) European and Asian T-13910 lactase-persistent (green) and C-13910 non-persistent (orange) homozygosity tracts, based on the data from ref. 14. Positions are relative to the start codon of *LCT*. Note that some tracks are too short to be visible as plotted.

Tishkoff et al. 2007 Nature Genetics

Soft sweeps vs. hard sweeps



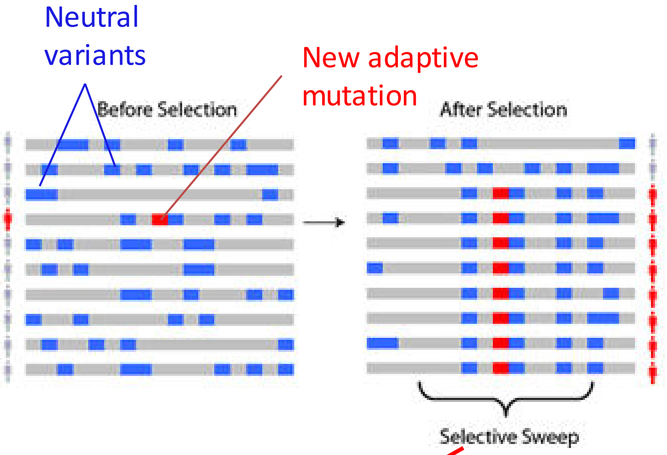
Novembre & Han 2012, Phil. Trans. R. Soc. B

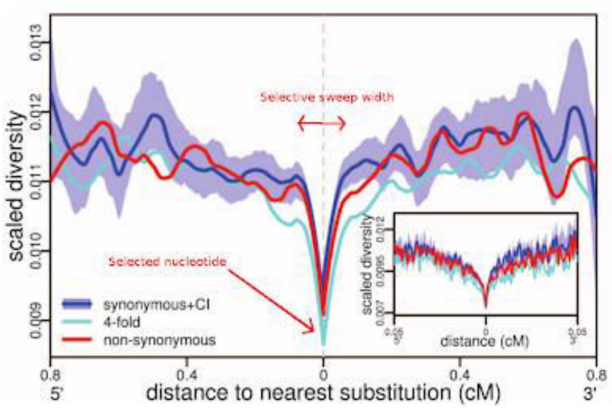
Some recent studies suggested that soft sweeps are probably more frequent, but this statement is still debated because soft sweep detection can generate a lot of false positives…

Short-time scales, methods are divided into two main groups:

Selective sweeps

(within-population variation)

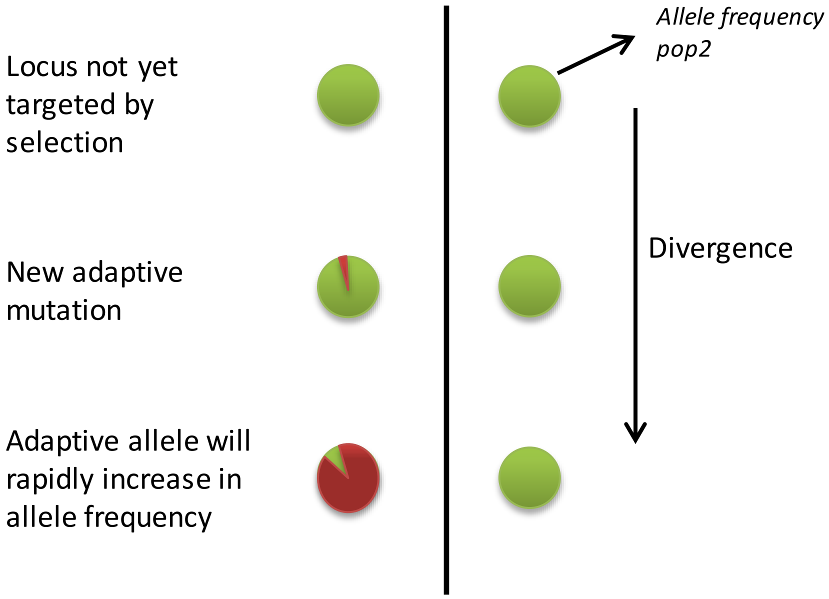




Reduction of the diversity at the selected locus (+ its linked neutral variants)

Genetic differentiation

(between populations)



Locus not yet targeted by selection

New adaptive mutation

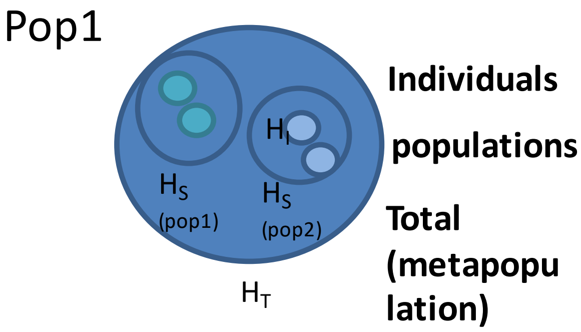
Adaptive allele will rapidly increase in allele frequency

Extreme allele frequency differences between the two populations at the selected locus

SNP in close vicinity to the targeted SNPs also exhibit strong differences in allele frequency

Fixation indices (F-statistics, F_{ST} in particular) <-> inbreeding

In nature, individuals rarely mate completely at random because of some geographically or ecologically-restricted mating among individuals. Such a non-random population mating drive differentiation among populations over the whole genome (i.e. population structure).

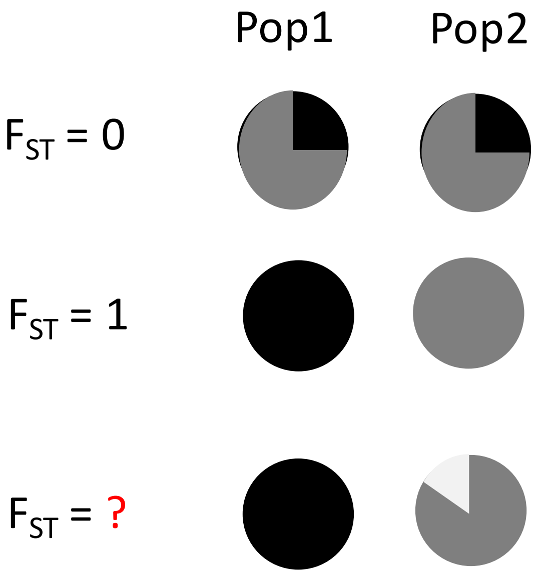


F_{ST} = deviation in allele frequencies among populations relative to the expectation assuming panmixtia (random mating)

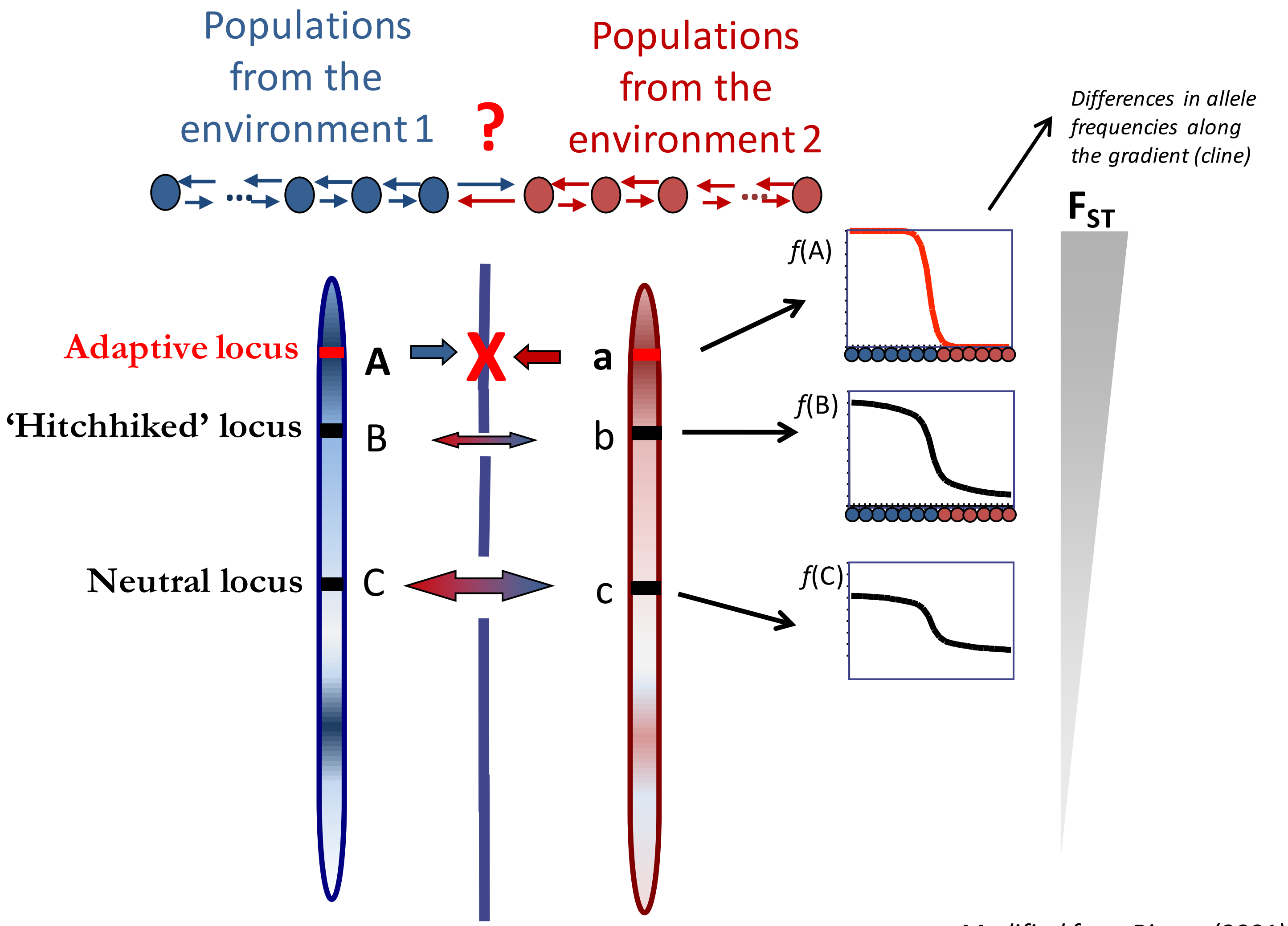
$$F_{ST} = (H_T - H_S)/H_T$$
$$= 1 - H_S/H_T$$

(with $H_s=2p_{S(pop)}q_{S(pop)}$ & $H_T=2p_{Total}q_{Total}$)

across multiple populations: average H_S
(here 2 pops: average between $H_{S(pop1)}$ & $H_{S(pop2)}$)



Genetic differentiation



Populations from the environment 1

?

Populations from the environment 2

Differences in allele frequencies along the gradient (cline)

F_{ST}

Adaptive locus

‘Hitchhiked’ locus

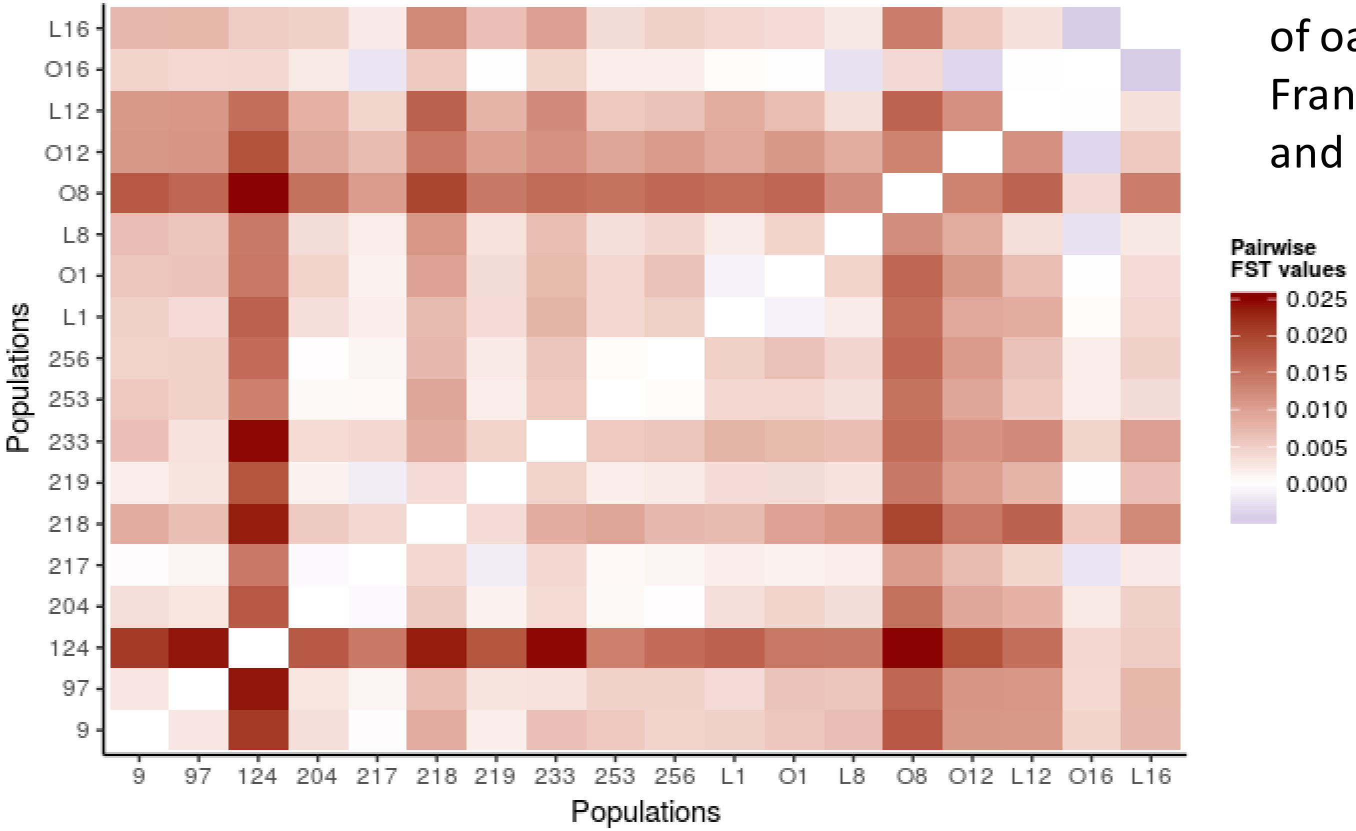
Neutral locus

Modified from Bierne (2001)

Among population variation in F_{ST}

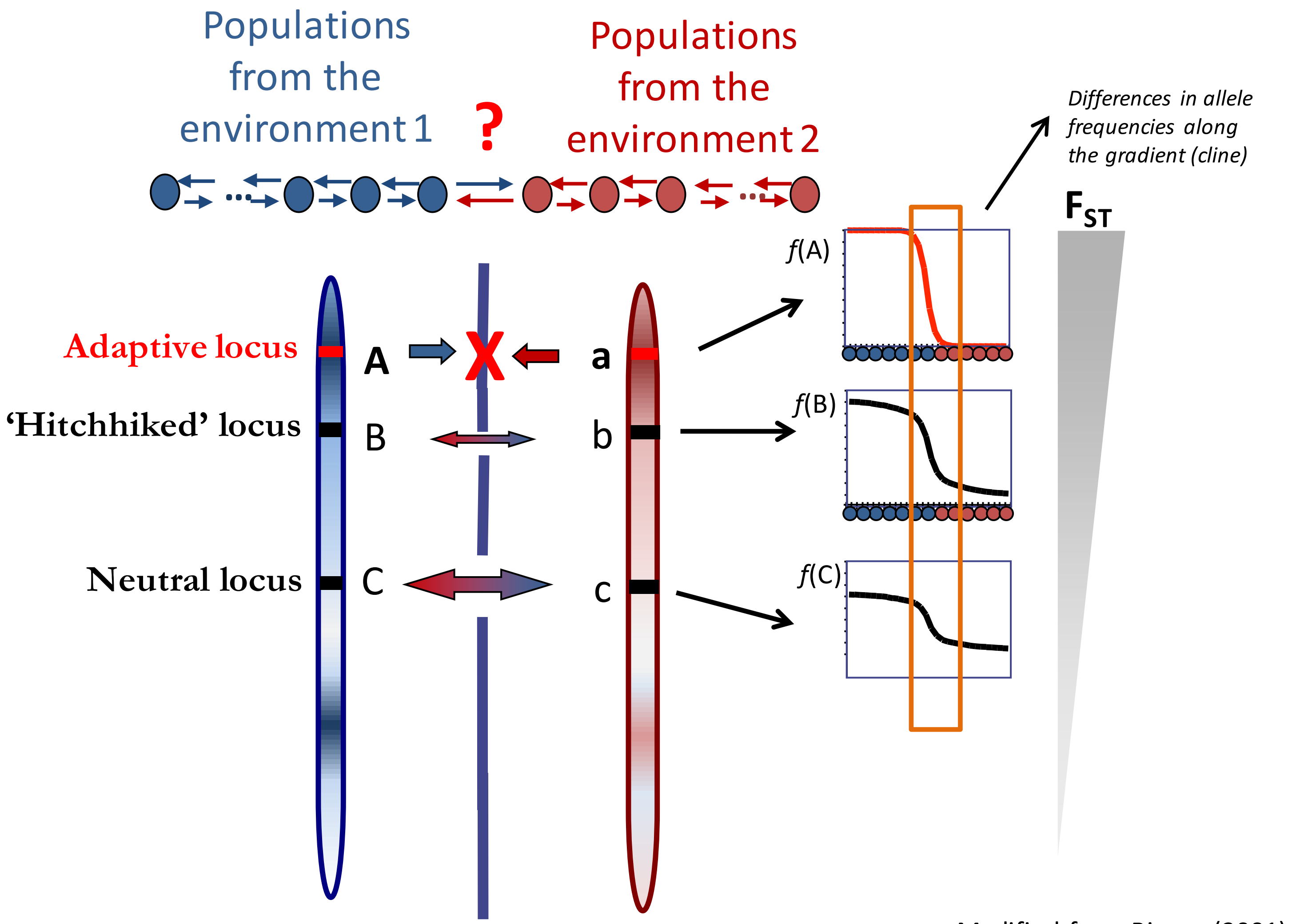
Given that the large majority of SNPs in the genome are neutral, the pairwise population differentiations computed over the whole dataset are representative of the population structure (*i.e.* past or present departure from panmixia of a given population <-> demographic history)

100k SNP, 18 pops of oaks over France, Germany and Ireland



Leroy et al. 2020 New Phytologist 226: 1171-1182

Genetic differentiation

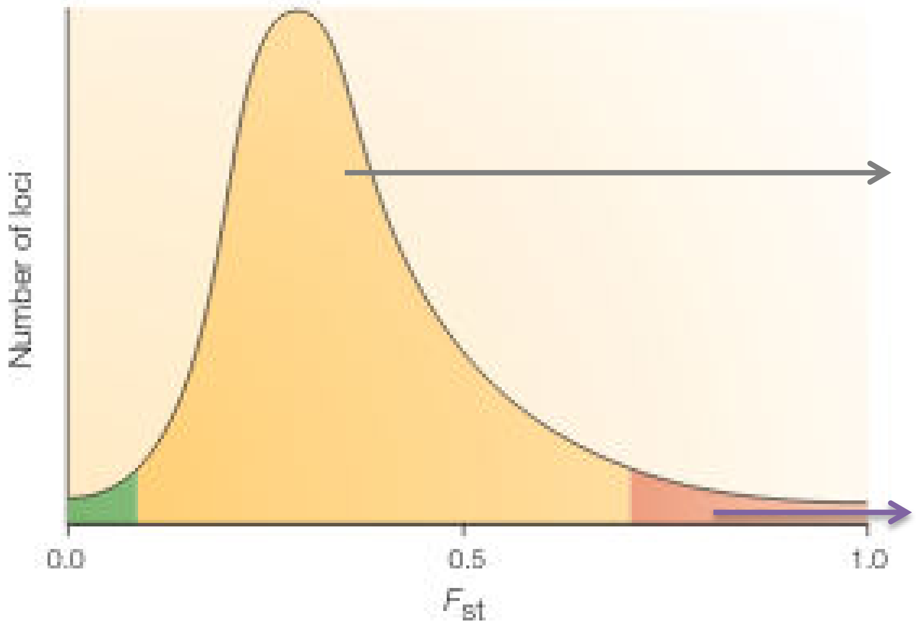

Populations from the environment 1
?
Populations from the environment 2
Differences in allele frequencies along the gradient (cline)
F_ST
Adaptive locus
A
a
'Hitchhiked' locus
B
b
Neutral locus
C
c
f(A)
f(B)
f(C)

Modified from Bierne (2001)

Reciprocally, if we want to identify some potential adaptive locus, we can focus on SNPs exhibiting the highest F_{ST} values!

Among locus variation in F_{ST}

Empirical distribution of F_{ST} among all genotyped loci



Neutral (informative of the population structure/demographic history)

Genes under positive (diversifying) selection?

Figure 2 | **Identifying outlier behaviour.** A hypothetical distribution of F_{st} (genetic divergence) and F_{is} (deviation from Hardy–Weinberg proportions) among neutral loci that are sampled from across the genome. Locus-specific effects lead to a few outlier loci with a highly divergent F_{st} or F_{is} value relative to most other loci across the genome. Modified with permission from REF. 1 © (2001) Annual Reviews.

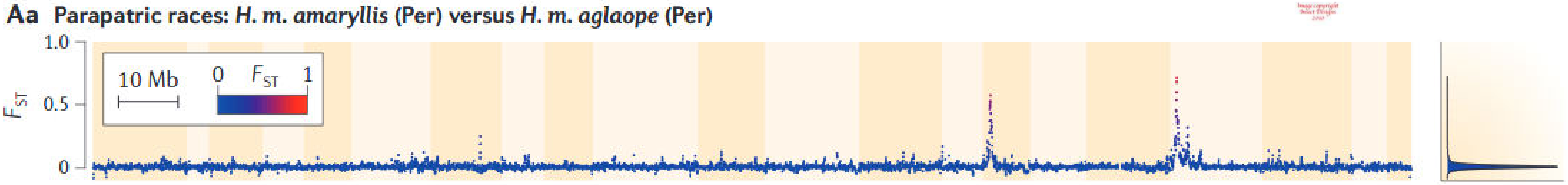
Lewontin and Krakauer's (LK) test for the heterogeneity of the F_{ST} index across loci (Lewontin & Krakauer, 1973 Genetics)

Loci targeted by natural selection can be on both tailed of the distribution ('outlier loci'):

Very low F_{ST} levels = putative loci under balancing selection (less differentiation than expected for a neutral marker)

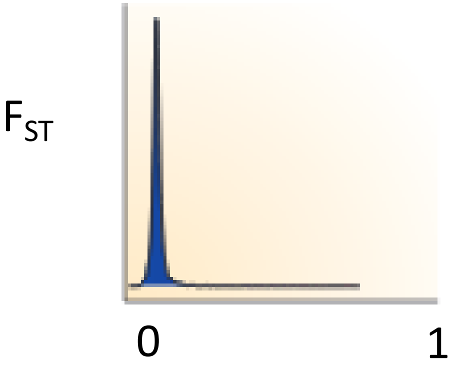
Very high F_{ST} levels = putative loci under positive selection (more differentiation than expected for a neutral marker)

Among locus variation in Fst



Aa Parapatric races: *H. m. amaryllis* (Per) versus *H. m. aglaope* (Per)

This plot showing the variation of the differentiation along chromosomes is called a 'Manhattan plot'



→ Almost all SNPs exhibit Fst values close to 0 (*i.e.* almost no population structure)

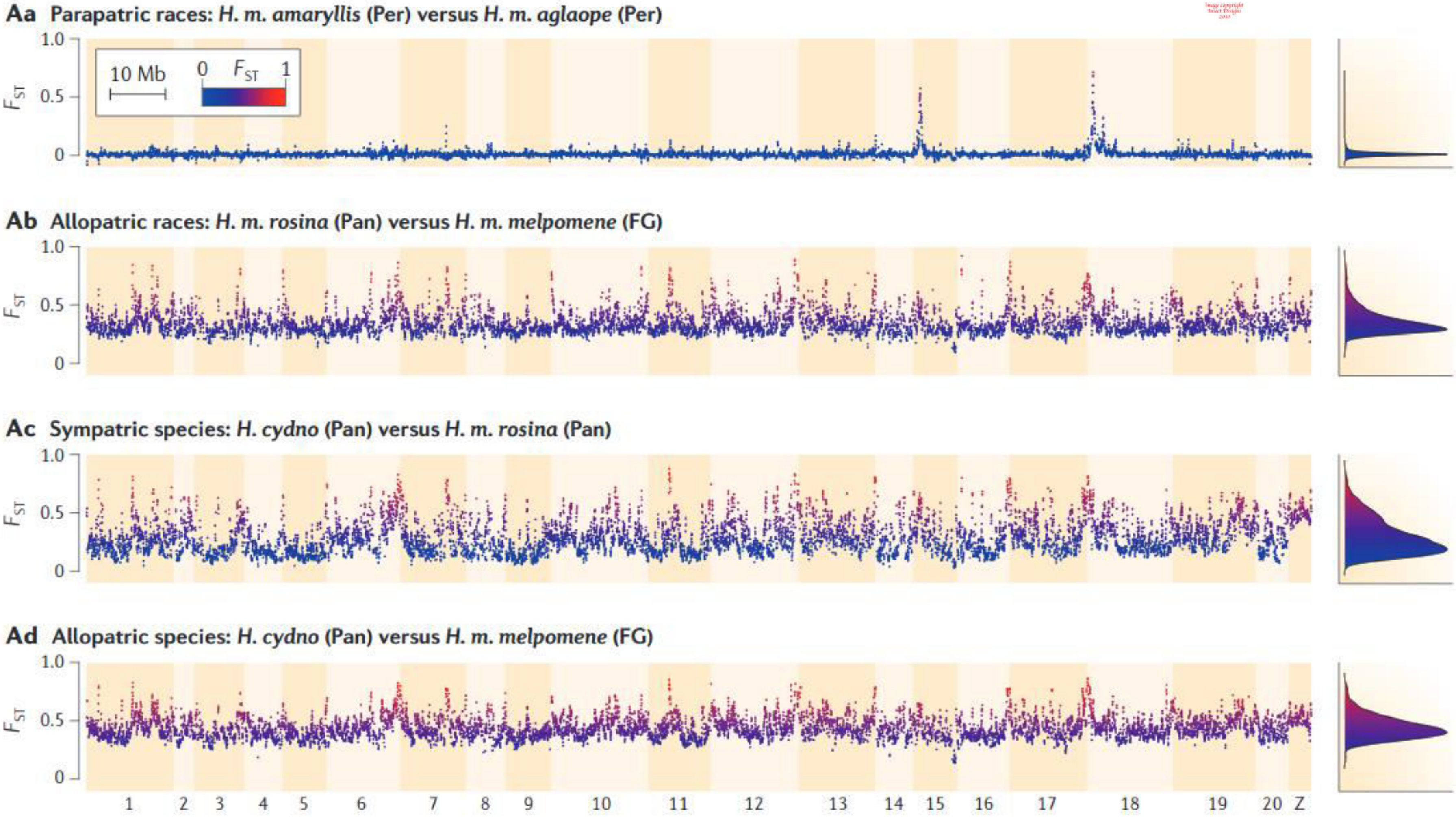
→ Very long tail of the distribution ('clear outliers')

→ These outliers colocate in a few narrow regions of high differentiation, which represent interesting regions to identify the genetic basis for reproductive isolation between these two parapatric populations

→ Ideal situation, but rarely observed in practice!

Seehausen et al. 2014 Nature Review Genetics

Among locus variation in Fst



Aa Parapatric races: *H. m. amaryllis* (Per) versus *H. m. aglaope* (Per)

Ab Allopatric races: *H. m. rosina* (Pan) versus *H. m. melpomene* (FG)

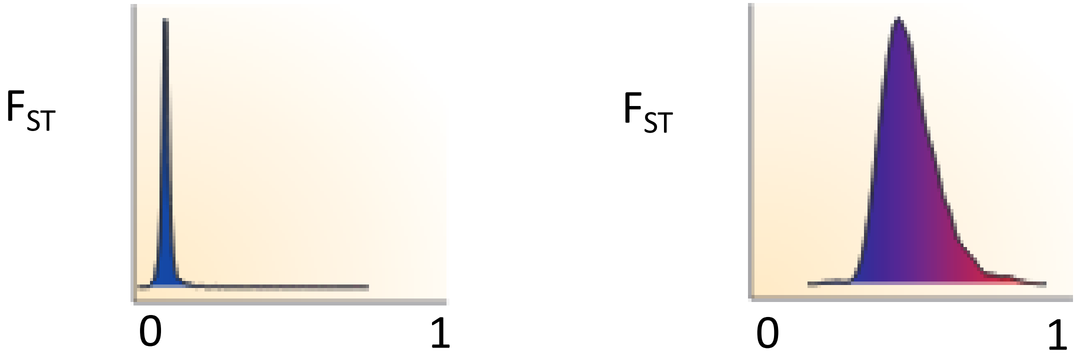
Ac Sympatric species: *H. cydno* (Pan) versus *H. m. rosina* (Pan)

Ad Allopatric species: *H. cydno* (Pan) versus *H. m. melpomene* (FG)

The plot showing the variation of the differentiation along chromosomes are called 'Manhattan plots'

Seehausen et al. 2014 Nature Review Genetics

Defining the threshold to identify the genes potentially under selection is tricky!



Which proportion of the genome is really under positive selection? 0.1%, 1%, 5%, more ?

If we a priori choose a threshold of 1%, i.e. we assume that 1% of the genome is under selection. In this case, I will consider SNPs that are in the top 1% of the F_{ST} distribution!

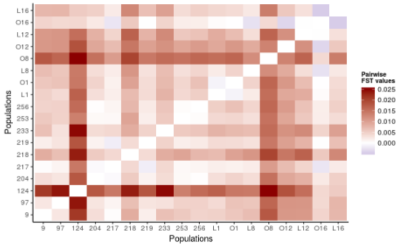
Problem 1: if 5% of the genome is under positive selection, a lot of selected SNPs will be falsely considered as neutral (false negatives).

Problem 2: in an even worst case, assume now that the populations evolve under strict neutrality (no genes are under selection), all the SNPs considered as outliers are in reality false positives

Such a strategy based on an assumed proportion is inadequate!

The general strategy is to generate a neutral expectation

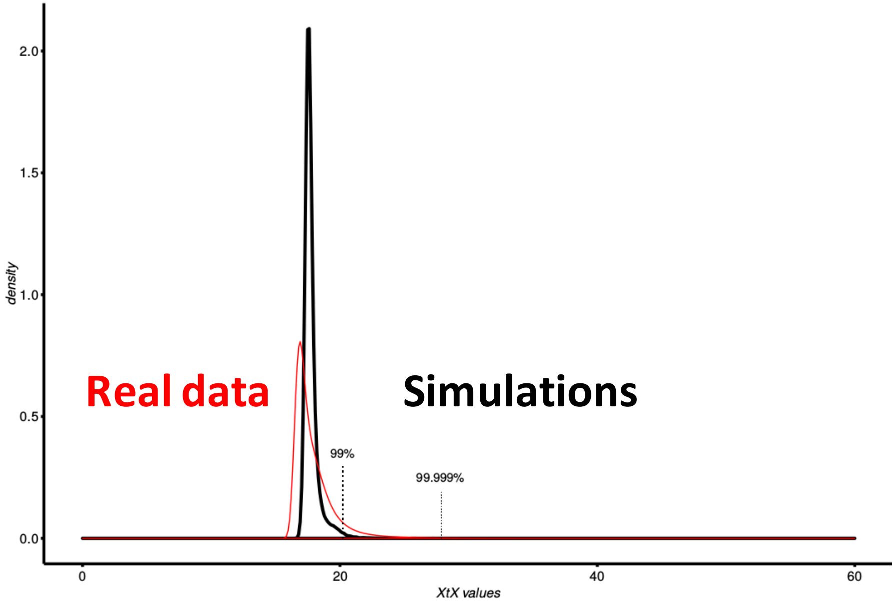
Strategy 1: perform neutral simulations assuming the observed levels of population structure



Perform simulations (so-called “Pseudo-Observed Datasets”, PODs) assuming the observed levels of population structure

All performed **simulations assume strict neutrality**

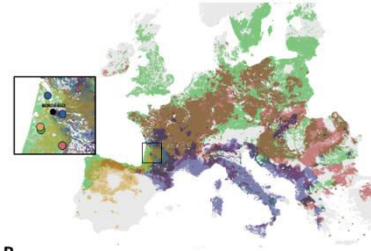
Thanks to these simulations we can therefore generate **the expected distribution of the metrics (e.g. F_{ST}) without selection** and then by comparing to the observed distribution, identify potential outliers



e.g. 18 oak pops, 3,090 SNPs among the 1,349,416 investigated SNPs exhibit values that are higher than the highest F_{ST} value observed for the simulations

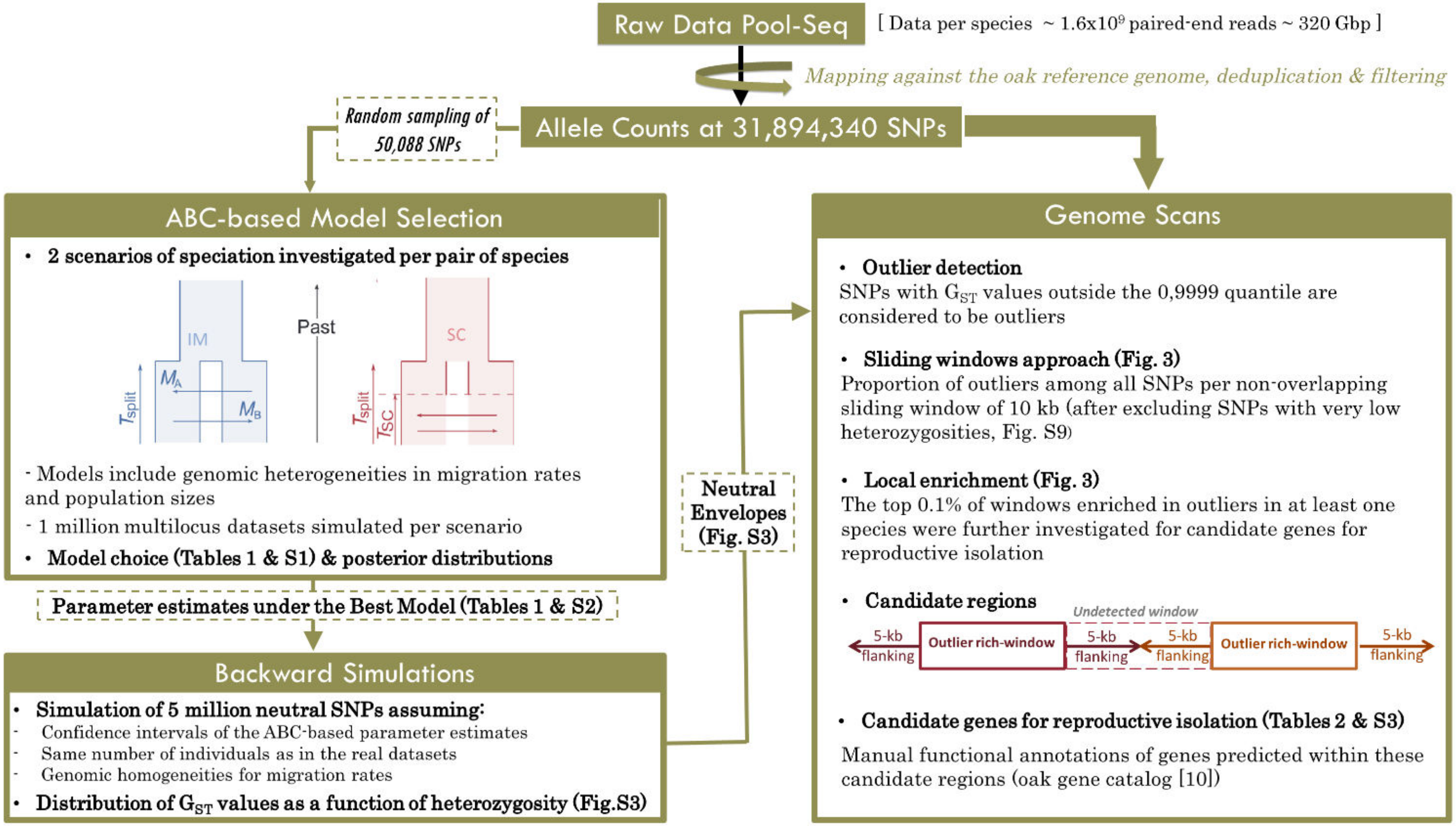
Assuming this criteria
3,090 / 1,349,416
=> 0.23% of the genome is under selection

Leroy et al. 2020 New Phytologist 226: 1171-1182



The general strategy is to generate a neutral expectation

Strategy 2: First, reconstruct the demographic history of a given species and then perform neutral simulations under this best demographic scenario



Raw Data Pool-Seq

[Data per species ~ 1.6×10^9 paired-end reads ~ 320 Gbp]

Mapping against the oak reference genome, deduplication & filtering

Allele Counts at 31,894,340 SNPs

Random sampling of 50,088 SNPs

ABC-based Model Selection

- **2 scenarios of speciation investigated per pair of species**

- Models include genomic heterogeneities in migration rates and population sizes
- 1 million multilocus datasets simulated per scenario

- **Model choice (Tables 1 & S1) & posterior distributions**

Parameter estimates under the Best Model (Tables 1 & S2)

Backward Simulations

- **Simulation of 5 million neutral SNPs assuming:**
  - Confidence intervals of the ABC-based parameter estimates
  - Same number of individuals as in the real datasets
  - Genomic homogeneities for migration rates
- **Distribution of G_{ST} values as a function of heterozygosity (Fig.S3)**

Neutral Envelopes (Fig. S3)

Genome Scans

- **Outlier detection**

  SNPs with G_{ST} values outside the 0,9999 quantile are considered to be outliers

- **Sliding windows approach (Fig. 3)**

  Proportion of outliers among all SNPs per non-overlapping sliding window of 10 kb (after excluding SNPs with very low heterozygosities, Fig. S9)

- **Local enrichment (Fig. 3)**

  The top 0.1% of windows enriched in outliers in at least one species were further investigated for candidate genes for reproductive isolation

- **Candidate regions**

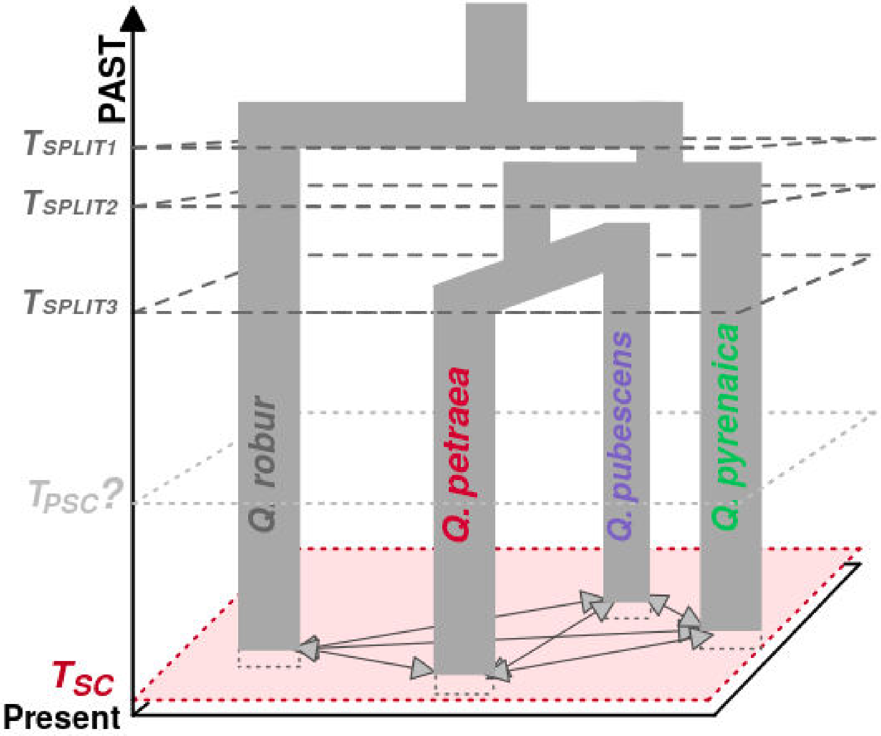
- **Candidate genes for reproductive isolation (Tables 2 & S3)**

  Manual functional annotations of genes predicted within these candidate regions (oak gene catalog [10])

Leroy et al. 2020 New Phytol, 226: 1183-1197

The general strategy is to generate a neutral expectation

Strategy 2: First, reconstruct the demographic history of a given species and then perform neutral simulations under this best demographic scenario

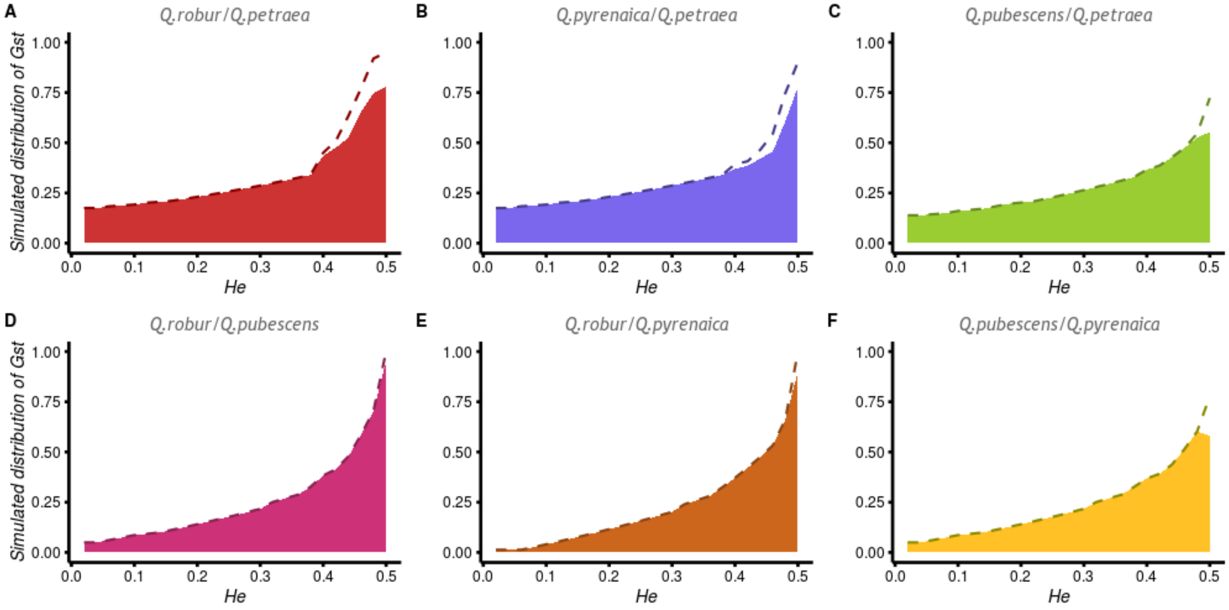


→ Best scenario identified using ABC (recent secondary contact between all species)

Leroy et al. 2017 New Phytologist

The general strategy is to generate a neutral expectation

Strategy 2: First, reconstruct the demographic history of a given species and then perform neutral simulations under this best demographic scenario



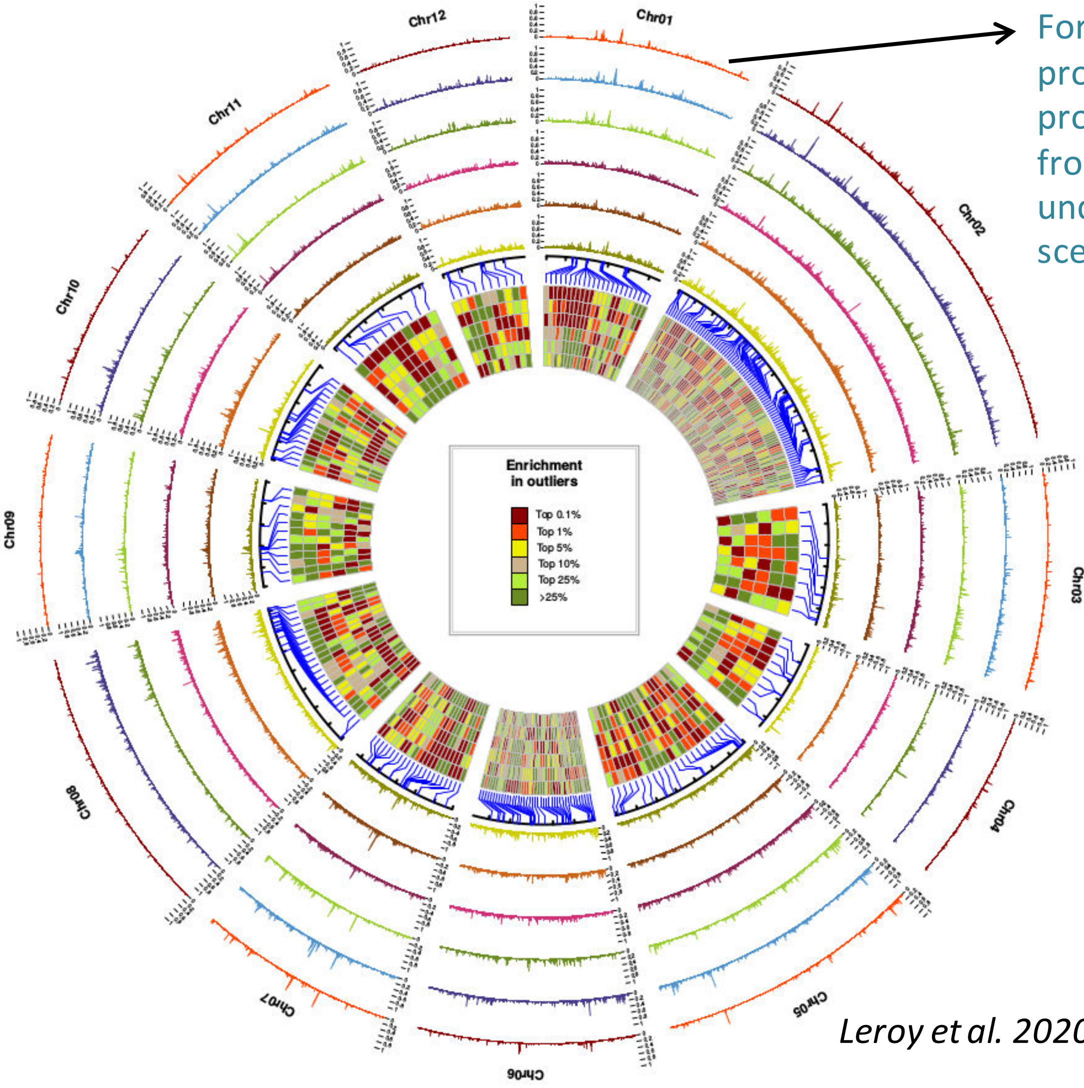
→ Generate neutral distribution based on the simulations under the best demographic scenario

→ Identify SNPs that exhibit values higher than this ‘neutral envelope’

Leroy et al. 2020 New Phytol, 226: 1183-1197

The general strategy is to generate a neutral expectation

Strategy 2: First, reconstruct the demographic history of a given species and then perform neutral simulations under this best demographic scenario



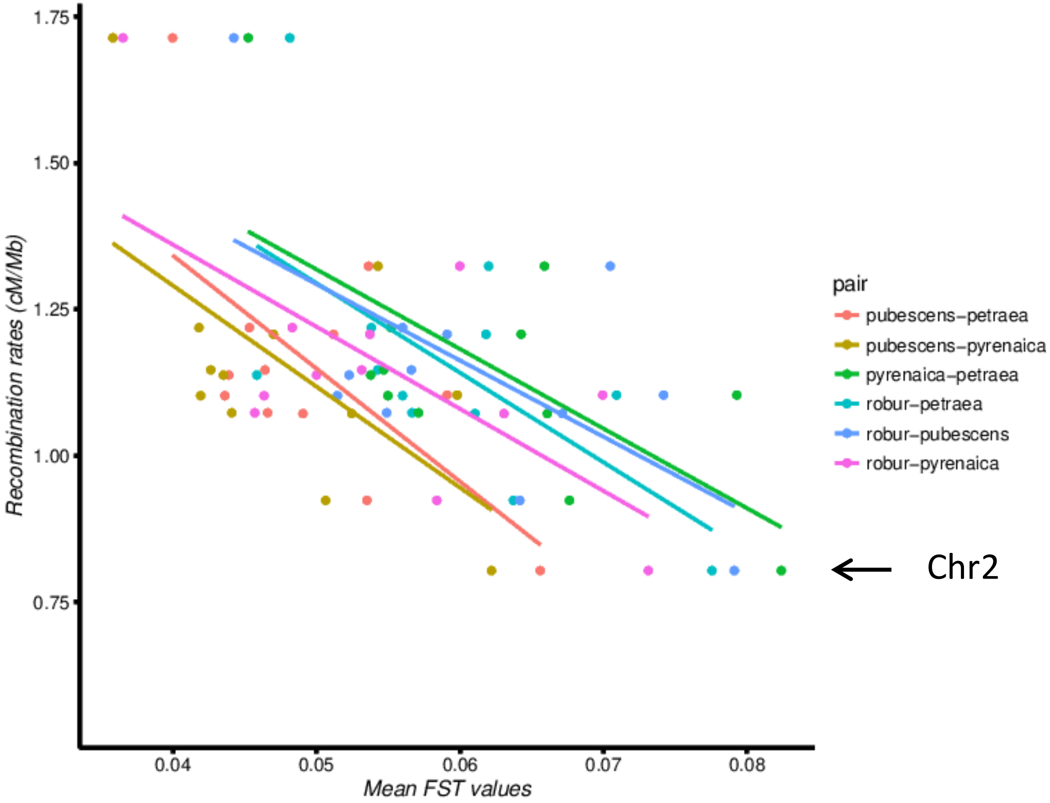
→ For each species pair, proportion of 'outliers', i.e. proportion of SNPs deviating from neutral expectations under the best demographic scenario

→ Identify narrow regions with elevated differentiation levels

→ Identify candidate genes in these narrow regions

Leroy et al. 2020 New Phytol, 226: 1183-1197

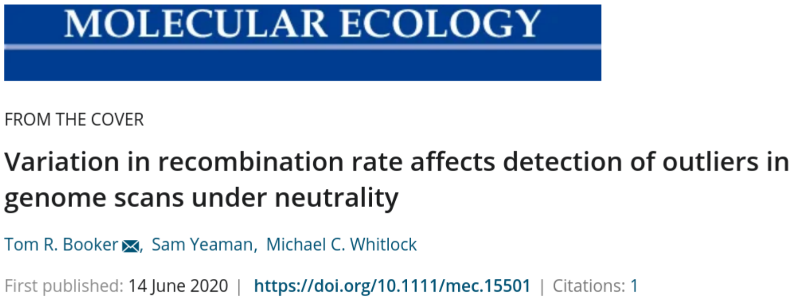
Variation of local recombination rate: another issue!



Leroy et al. 2020 New Phytol, 226: 1183-1197

Some other sources of variation (local or interchromosomal differences in recombination rates, effective population size variations…) are generally not taken into account!

That is now changing, because we more and more know that the neutral F_{ST} distribution also highly depends on the recombination rate!



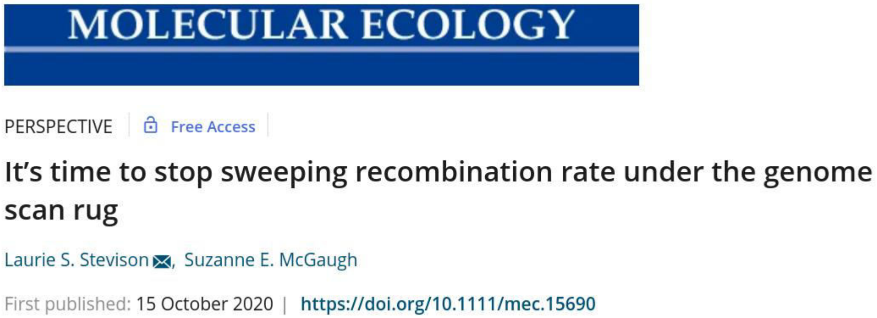
MOLECULAR ECOLOGY

FROM THE COVER

Variation in recombination rate affects detection of outliers in genome scans under neutrality

Tom R. Booker, Sam Yeaman, Michael C. Whitlock

First published: 14 June 2020 | https://doi.org/10.1111/mec.15501 | Citations: 1



MOLECULAR ECOLOGY

PERSPECTIVE | Free Access

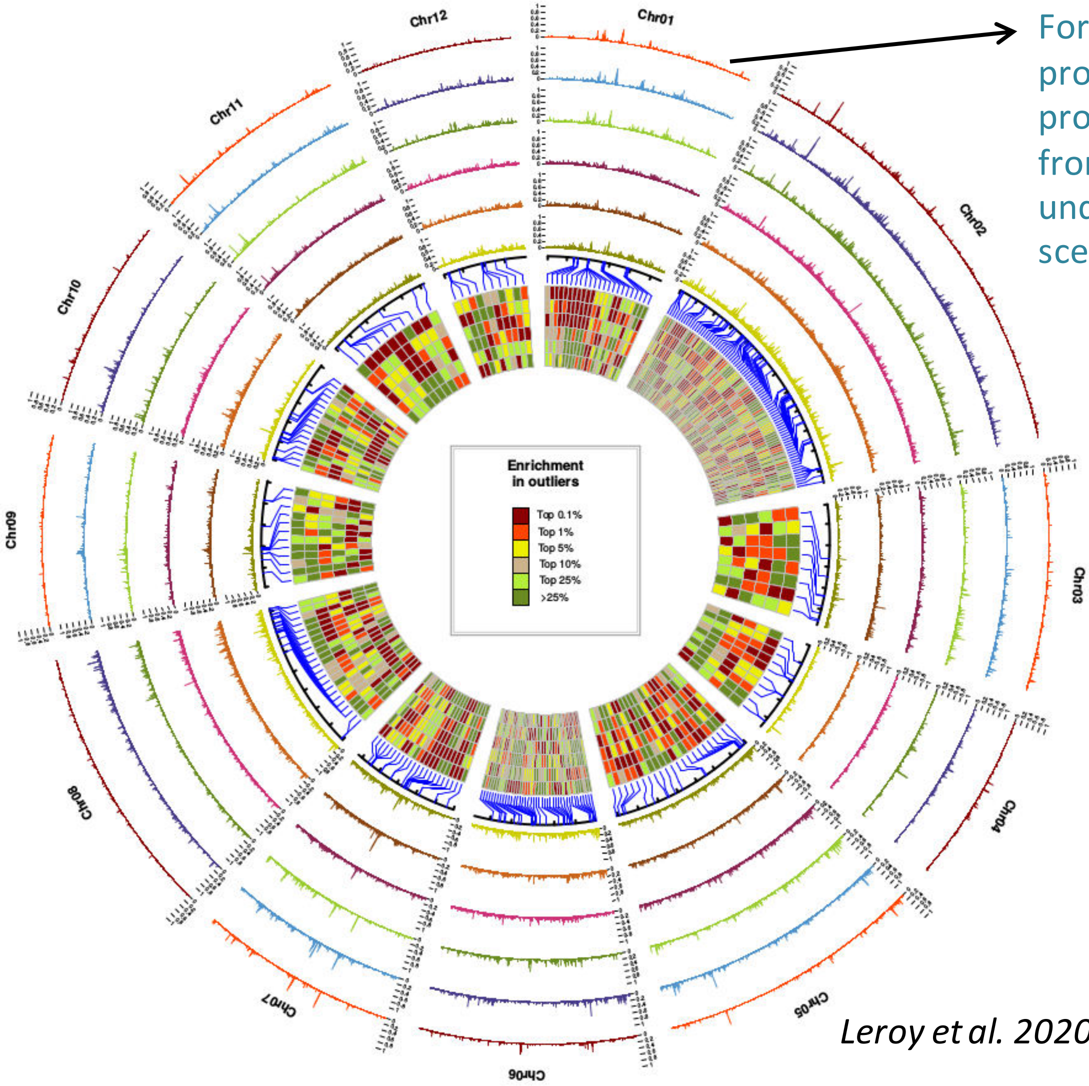
It's time to stop sweeping recombination rate under the genome scan rug

Laurie S. Stevison, Suzanne E. McGaugh

First published: 15 October 2020 | https://doi.org/10.1111/mec.15690

The general strategy is to generate a neutral expectation

Strategy 2: First, reconstruct the demographic history of a given species and then perform neutral simulations under this best demographic scenario



→ For each species pair, proportion of 'outliers', i.e. proportion of SNPs deviating from neutral expectations under the best demographic scenario

→ A lot of regions identified on the chromosome 2

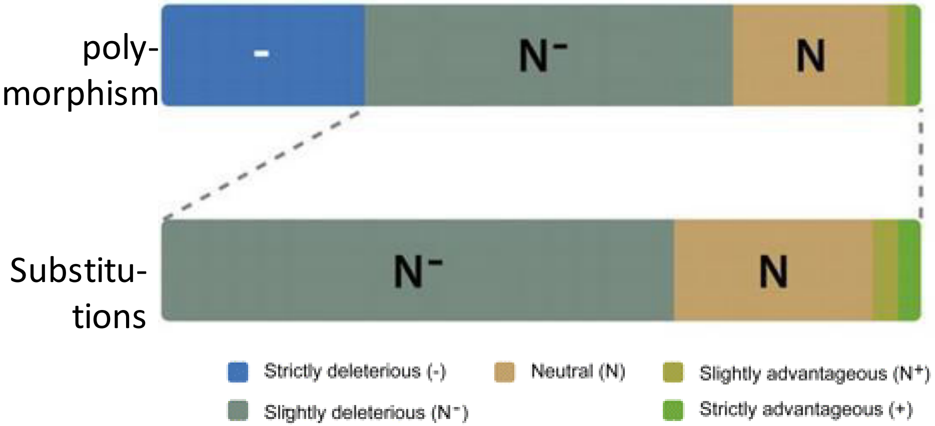
→ False positives because of the lower recombination rate?

Leroy et al. 2020 New Phytol, 226: 1183-1197

Summary

- Most non-synonymous mutations are neutral or deleterious, some can be advantageous

- Advantageous mutations are more frequently observed among substitutions than among polymorphisms because advantageous mutations rapidly fix in the population and are therefore ephemeral in the polymorphism (Reciprocally deleterious mutations are more frequent in the polymorphism)



- Substitution data are informative about historical selection, while polymorphism data are more informative about recent/ongoing selection

- Can be investigated with very different kinds of data, from a handful of genes from two or few species (substitutions) to whole-genome sequence of one or many populations (polymorphisms)!

- Selective sweep methods (incl. Tajima's D) only require data from a single population, 'FST scans' require at least 2 populations

- Identifying footprints of selection remains a complex task (e.g. detecting soft sweeps, neutral envelopes)