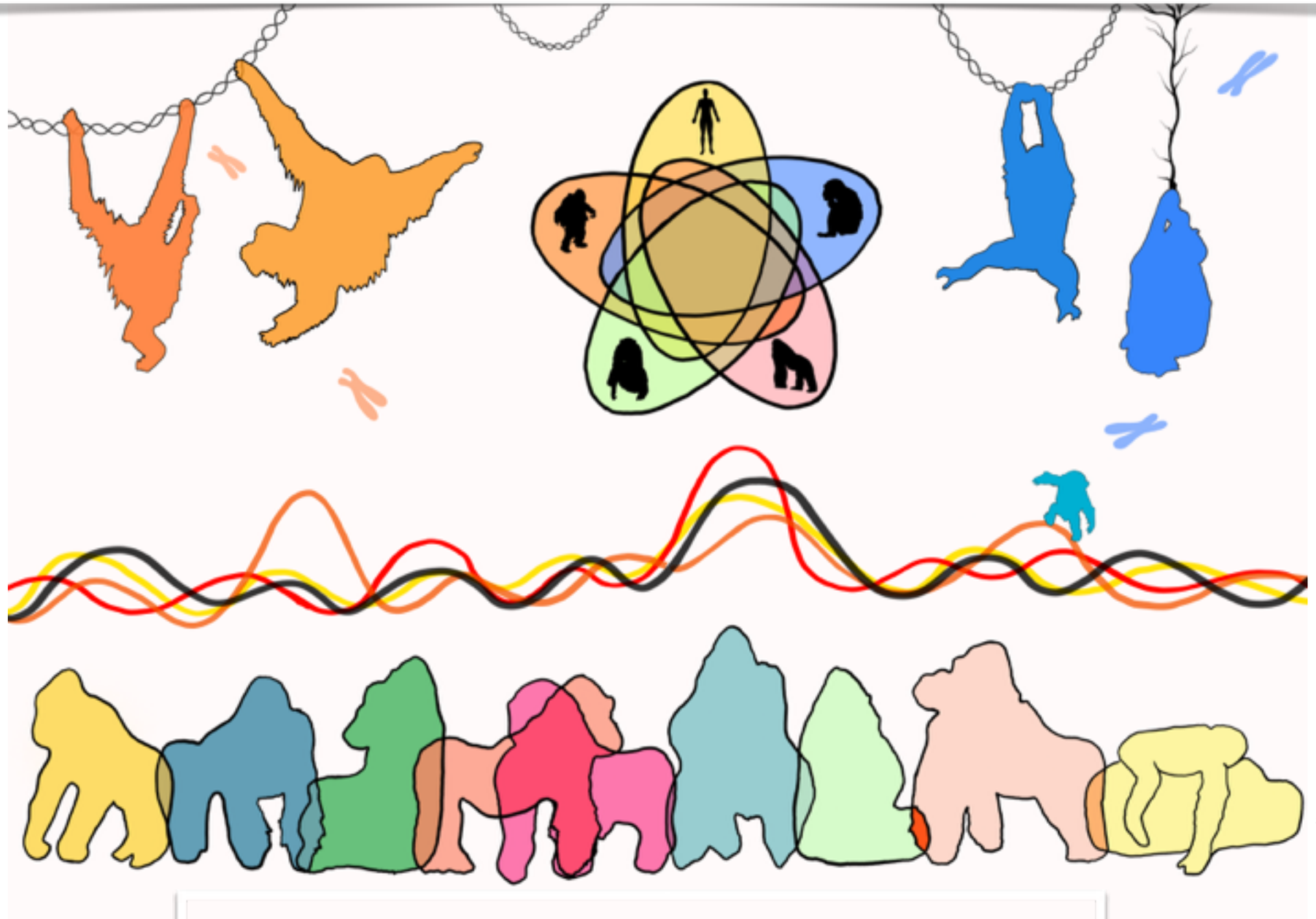
Species divergence (our closest living relatives)



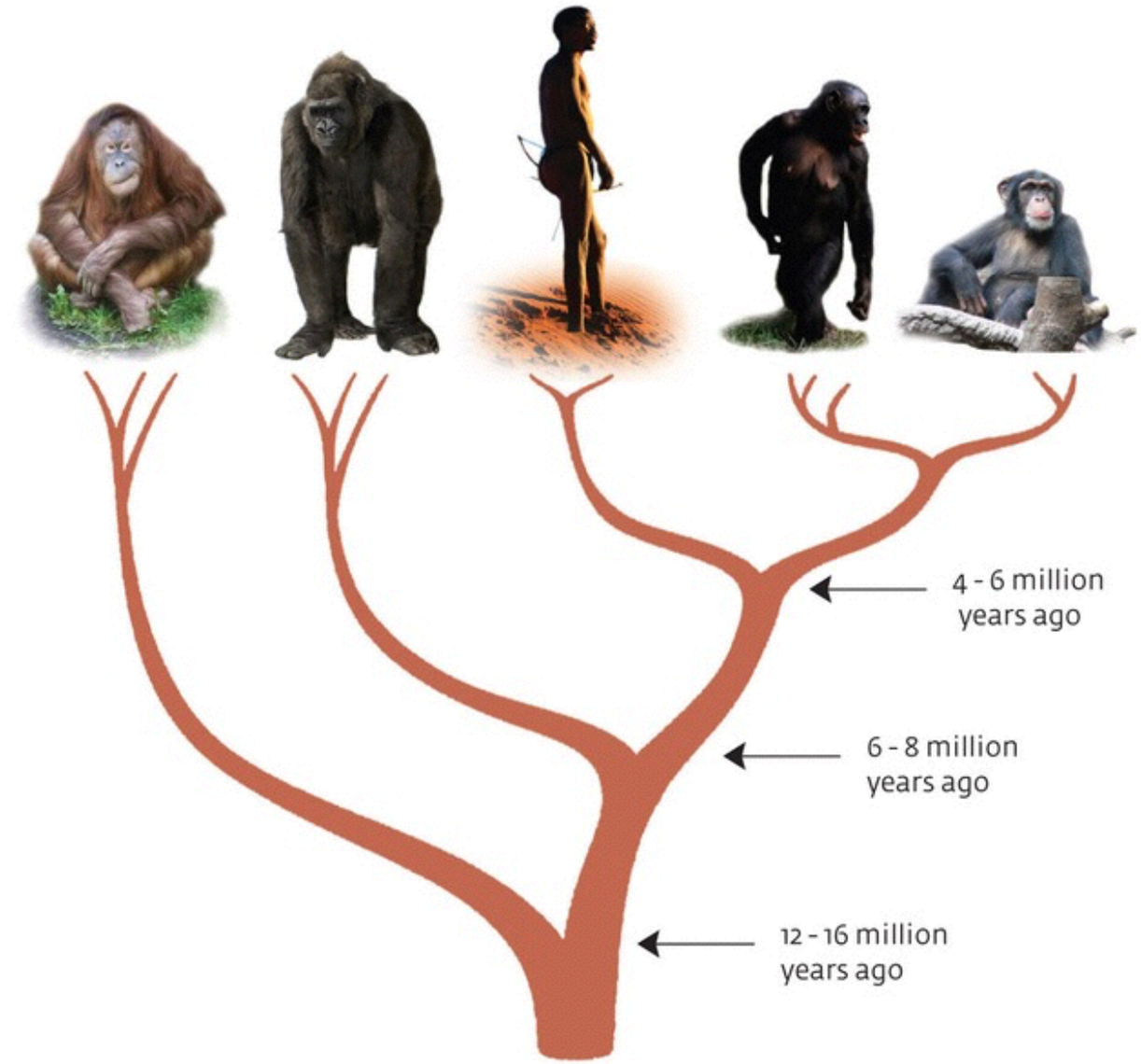
Alex Cagan

Aida Andrés
University College London

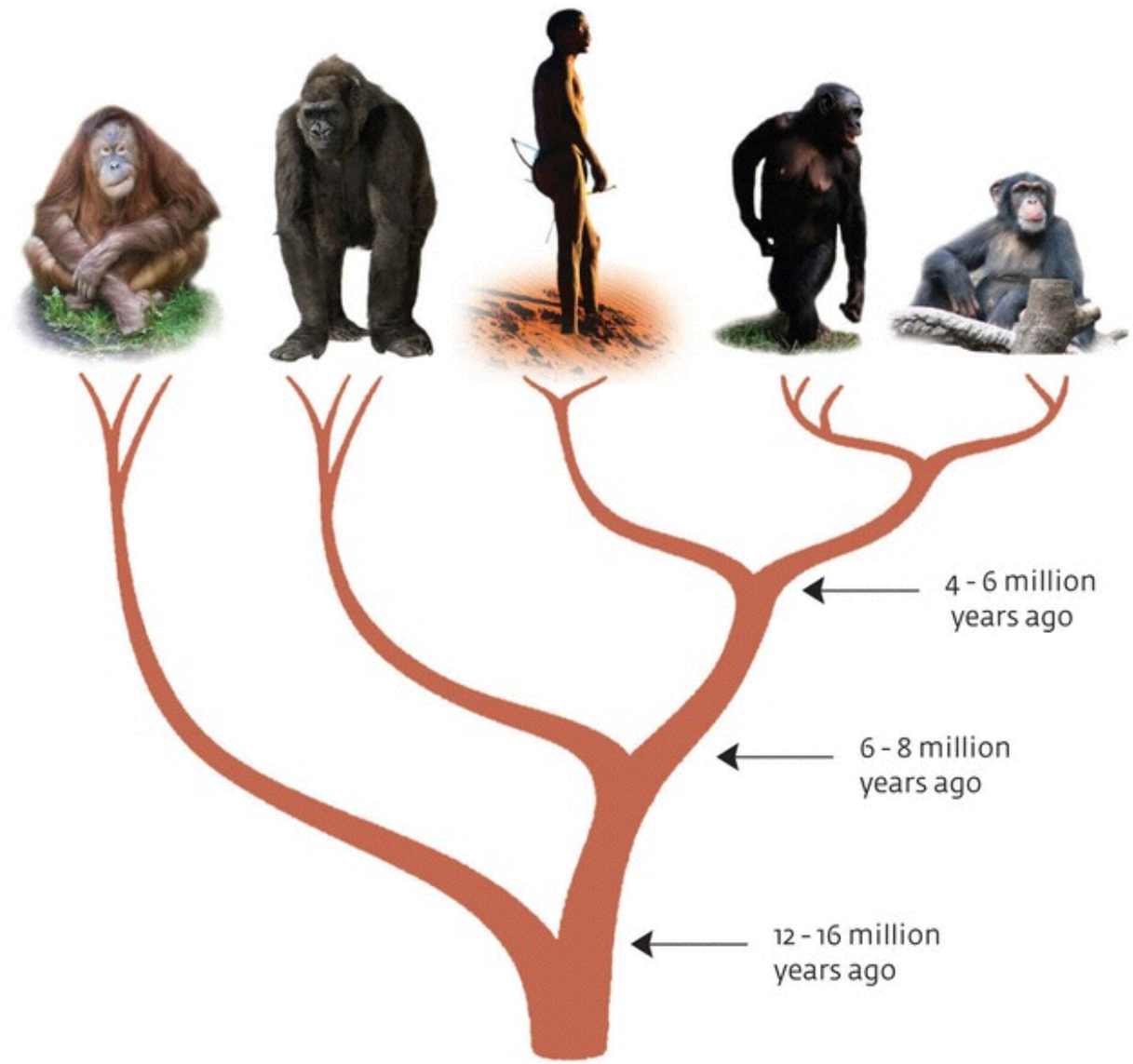
Prado-Martinez *et al.*, Science 2013; De Manuel *et al.*, Science 2016; *e.g.* 1000 Genomes dataset

Differ in:

- Demographic history
- Social patterns
- Mating behaviour
- Environment
- Diet
- Size
- Locomotion



Extremely closely related



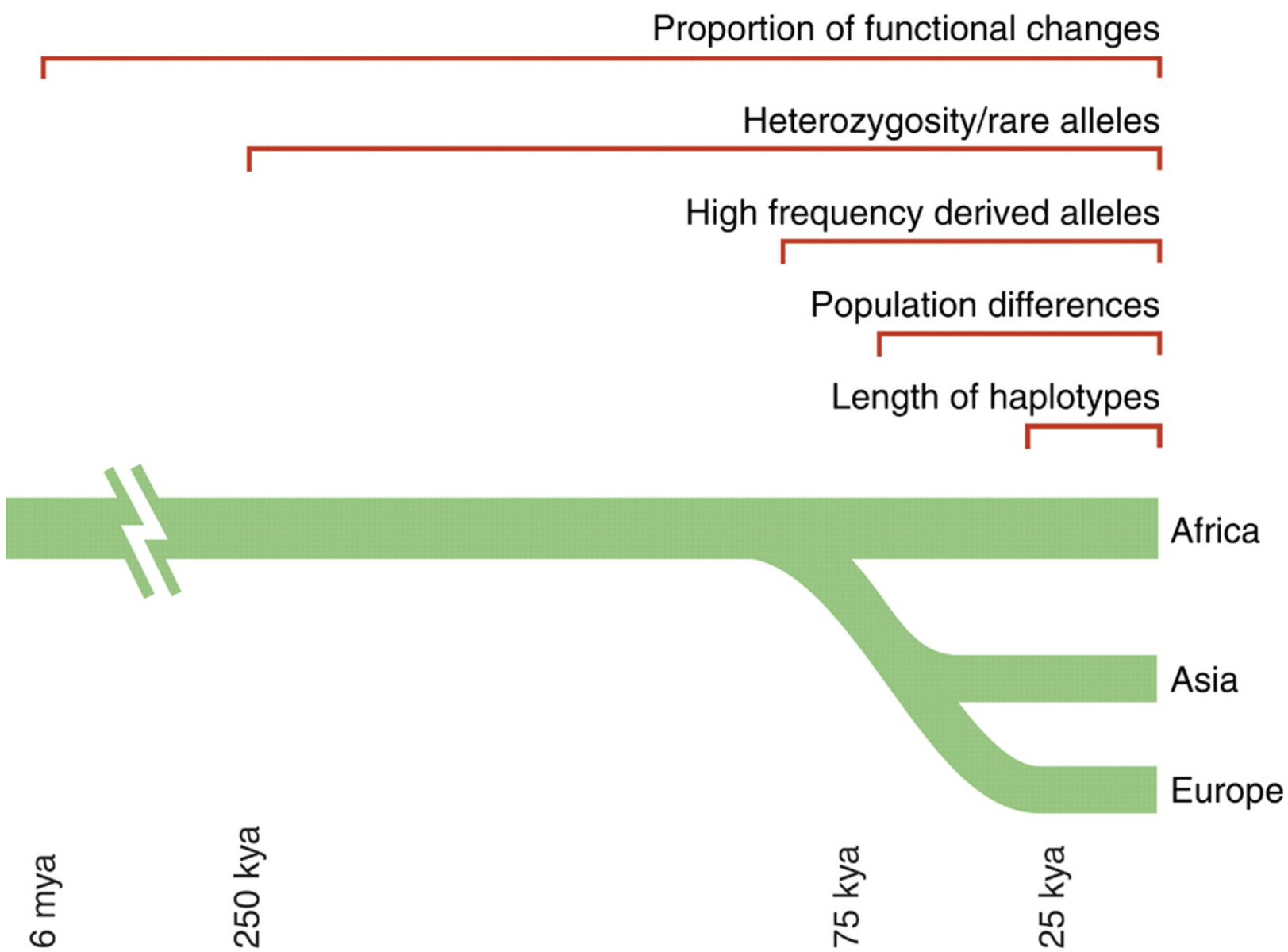
Selection at different older frames?

Selection over long evolutionary times?

Evolutionary context

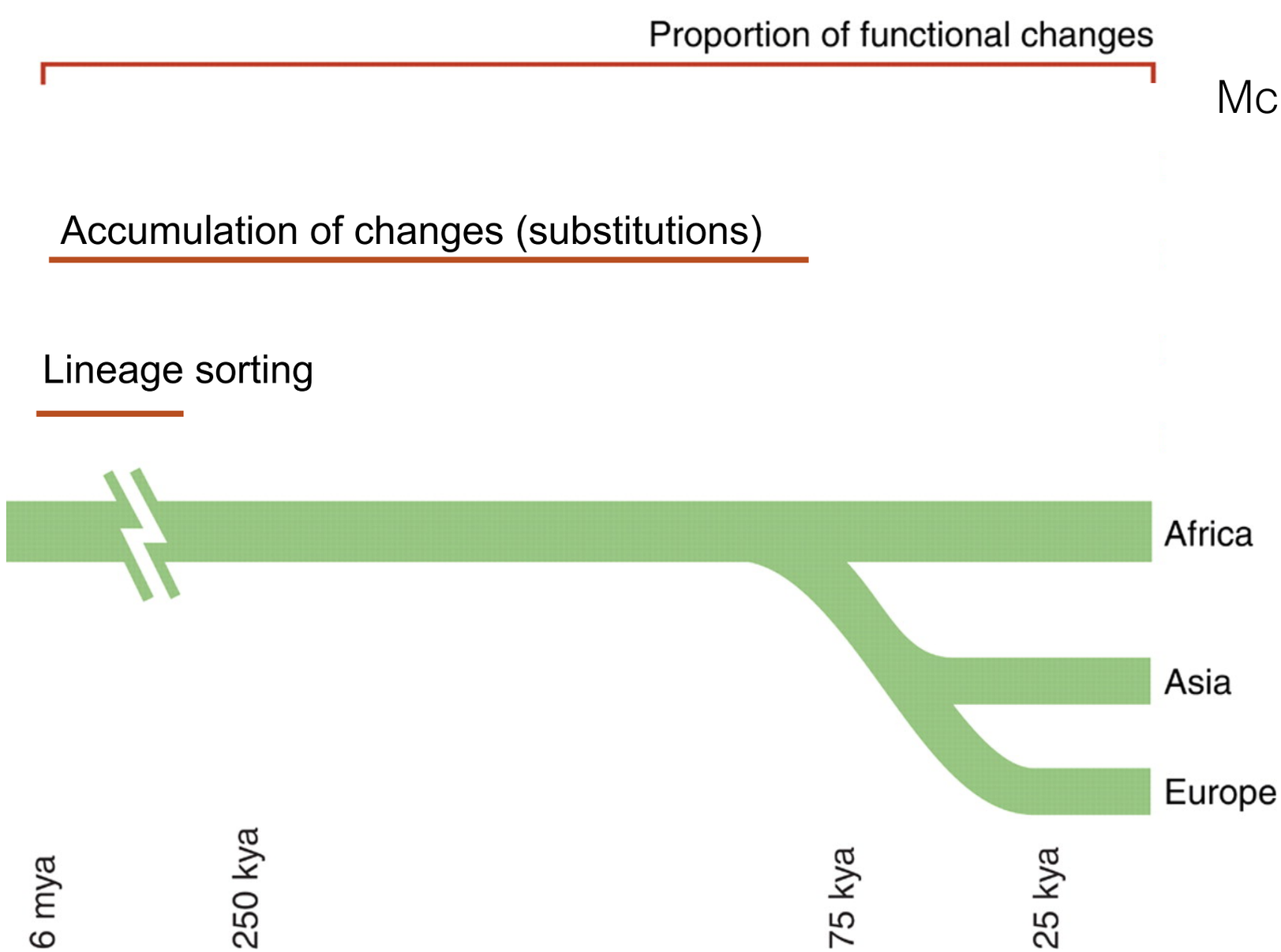
Direct advantages of including information on divergence

Time Scales for the Signatures of Selection



Sabeti et al., Science, 2006

Time Scales for the Signatures of Selection



Sabeti et al., Science, 2006

Protein-coding evolution: Ka/Ks

non-synonymous

synonymous

	AGA									UUA					AGC					
	AGG									UUG					AGU					
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

$$\frac{Ka}{Ks} = \frac{\text{proportion of NS changes}}{\text{proportion of S changes}}$$

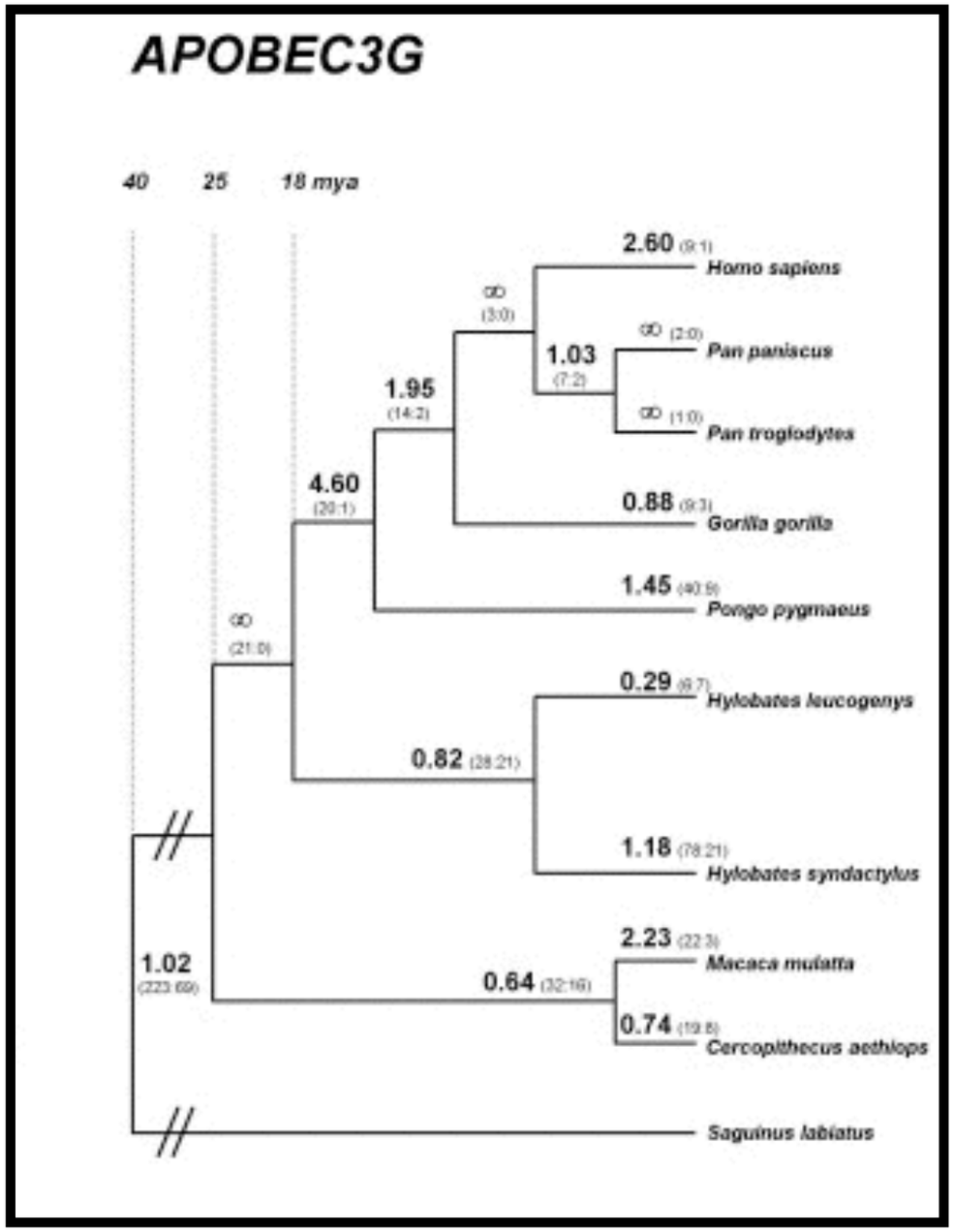
dN/dS (PAML)

Maximum likelihood approach

Along the full sequence

For specific lineages

Across all lineages



Ortiz *et al.,* Retrovirology 2006

dN/dS (PAML)

Maximum likelihood approach

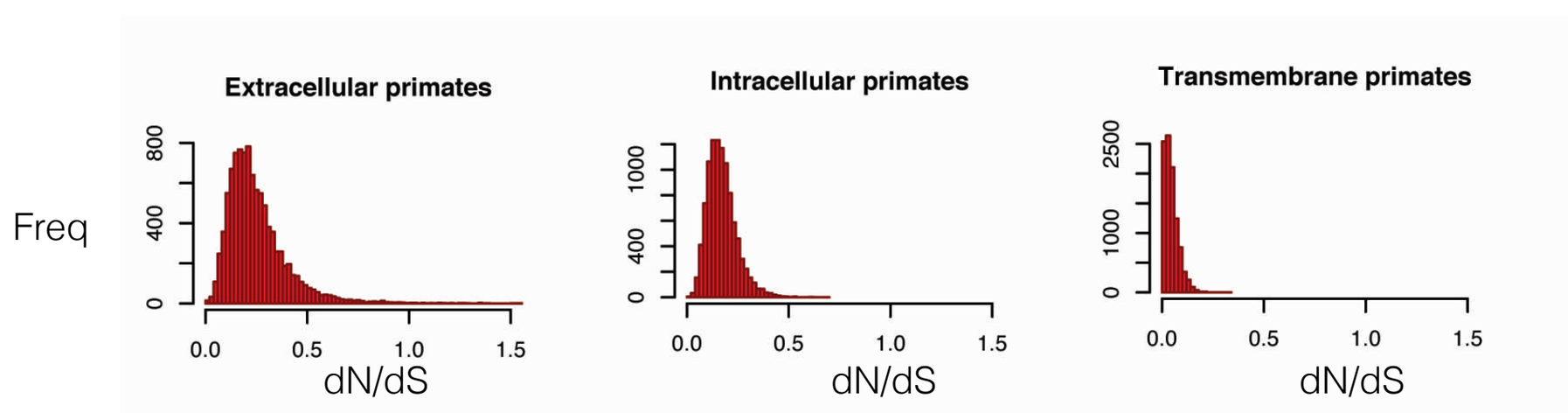
Along the full sequence

For specific lineages

Across all lineages

Per codon

In particular protein sections (domains)



Andrés et al., MBE 2007

dN/dS (PAML)

Maximum likelihood approach

Along the full sequence

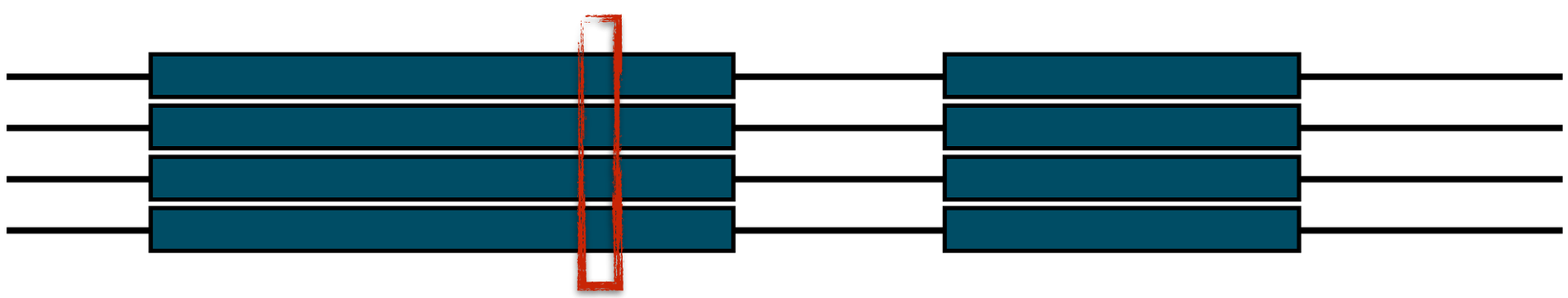
For specific lineages

Across all lineages

Per codon

In particular protein sections (domains)(exons)

For specific codons



dN/dS (PAML)

Maximum likelihood approach

- Along the full sequence
  - For specific lineages
  - Across all lineages
- Per codon
  - In particular protein sections (domains)
  - For specific codons
  - For specific codons *and* lineages

Calculate the likelihood of models with positive selection in particular lineages/codons and identify putatively selected codons

McDonald & Kreitman (MK)

	Fixed	Polymorphic
Synonymous	D_s	P_s
Nonsynonymous	D_n	P_n

Test

Assumptions:

- Rapid fixation of advantageous alleles
- Rapid removal of deleterious alleles
- Similar drift in both

McDonald and Kreitman, 1991

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).

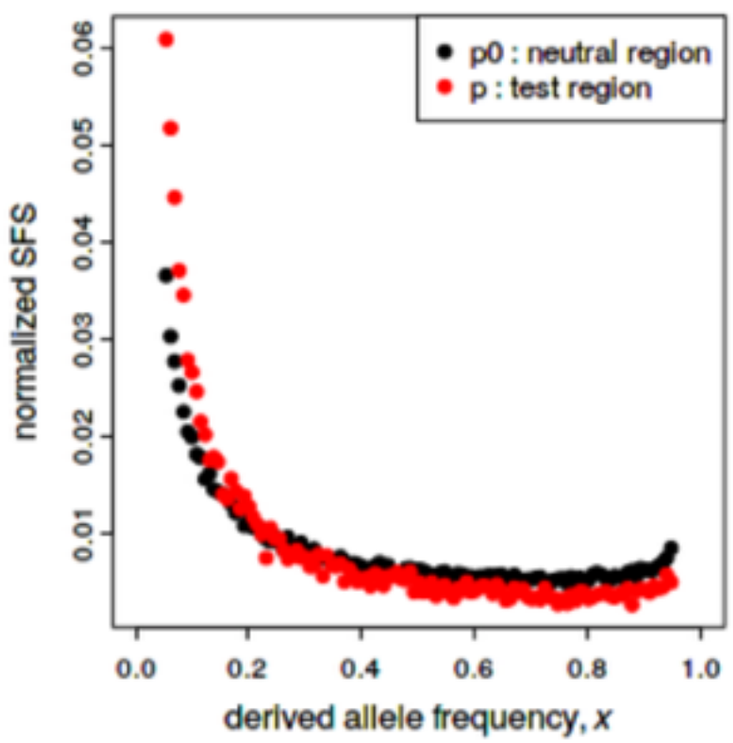
	Fixed	Polymorphic
Synonymous	D_s	P_s
Nonsynonymous	D_n	P_n

$$\alpha = 1 - \frac{d_0}{d}\frac{p}{p_0},$$

Smith and Eyre-Walker 2002; Eyre-Walker 2006

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).



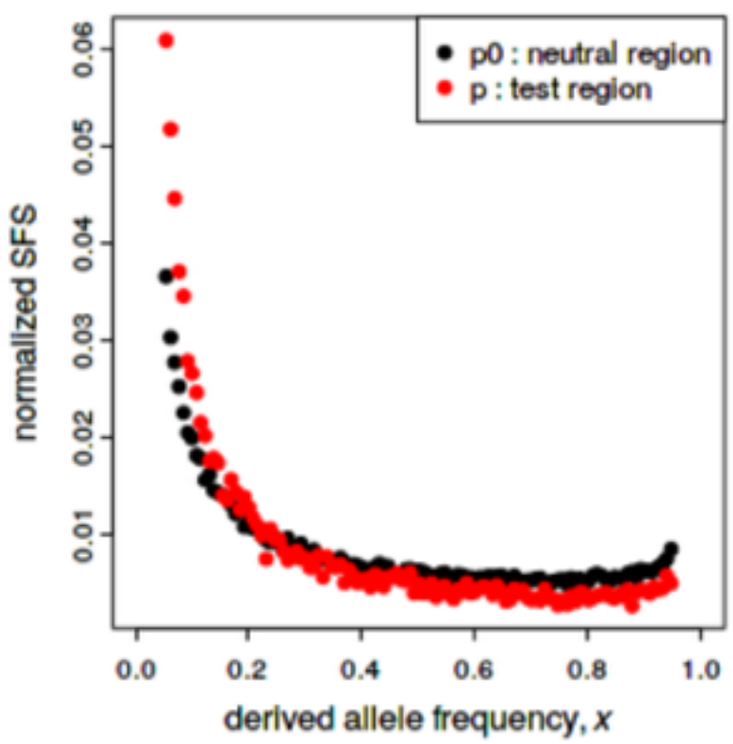
Issue: accumulation of polymorphic slightly deleterious alleles

Solution: removing low-frequency alleles

Issue: how to choose right freq?

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).



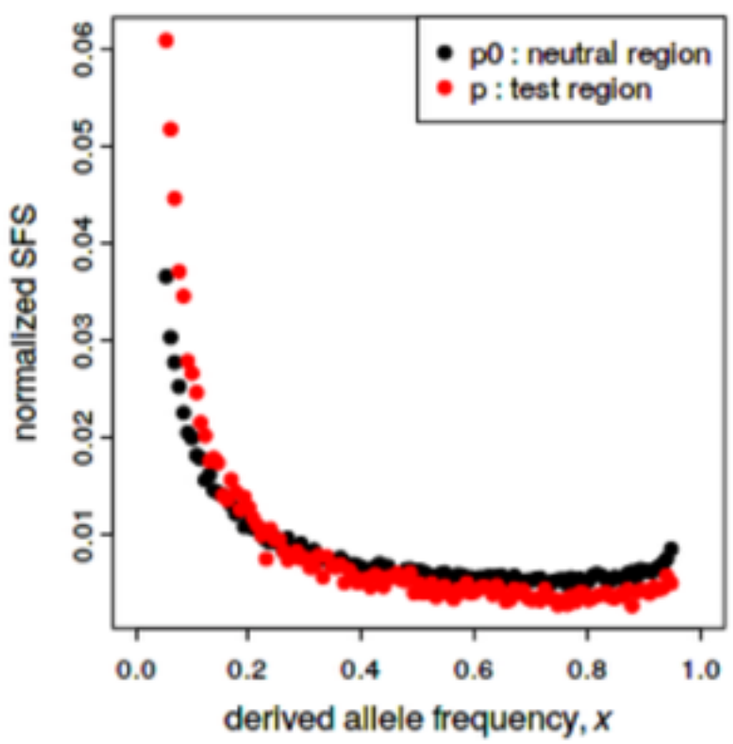
Issue: accumulation of polymorphic slightly deleterious alleles

Solution: simultaneous estimate of the DFE of new mutations using the SFS, and alpha
e.g. DFE-alpha

Eyre-Walker and Keighley MBE 2009

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).



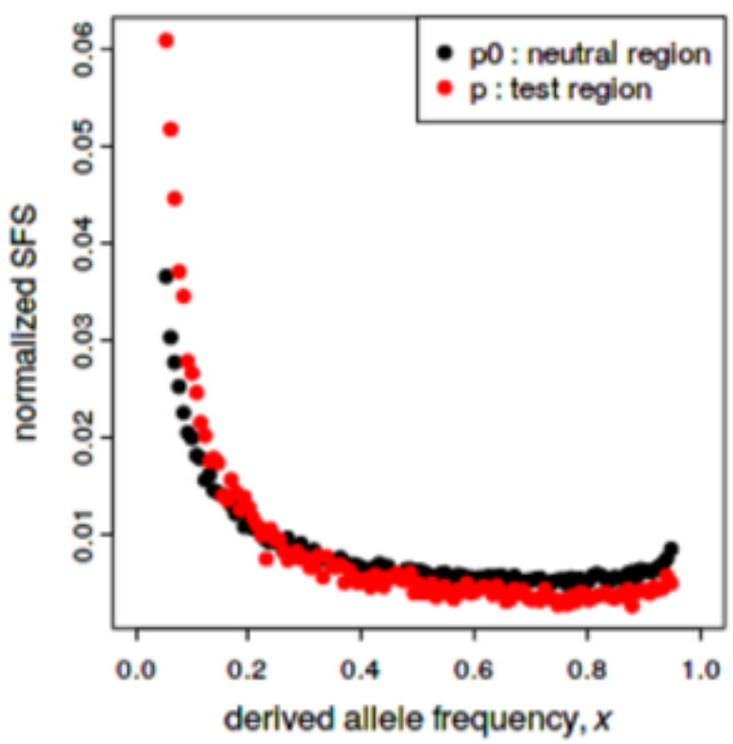
Issue: accumulation of polymorphic slightly deleterious alleles

Solution: simultaneous estimate of the DFE of new mutations using the SFS, and alpha
e.g. DFE-alpha

Issue: demography and linked selection (background selection and genetic draft)

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).



Issue: accumulation of polymorphic slightly deleterious alleles

Issue: demography and linked selection (background selection and genetic draft)

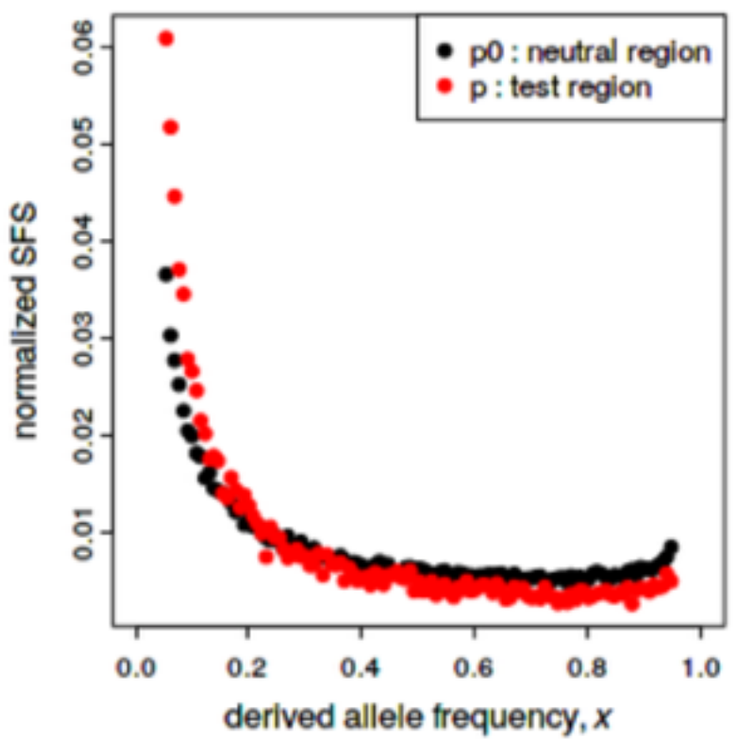
Solution: asymptoticMK

Messer and Petrov, PNAS 2013

McDonald & Kreitman (MK)

Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).

asymptoticMK



$$\alpha(x) = 1 - \frac{d_0}{d}\frac{p(x)}{p_0(x)}.$$

Messer and Petrov, PNAS 2013

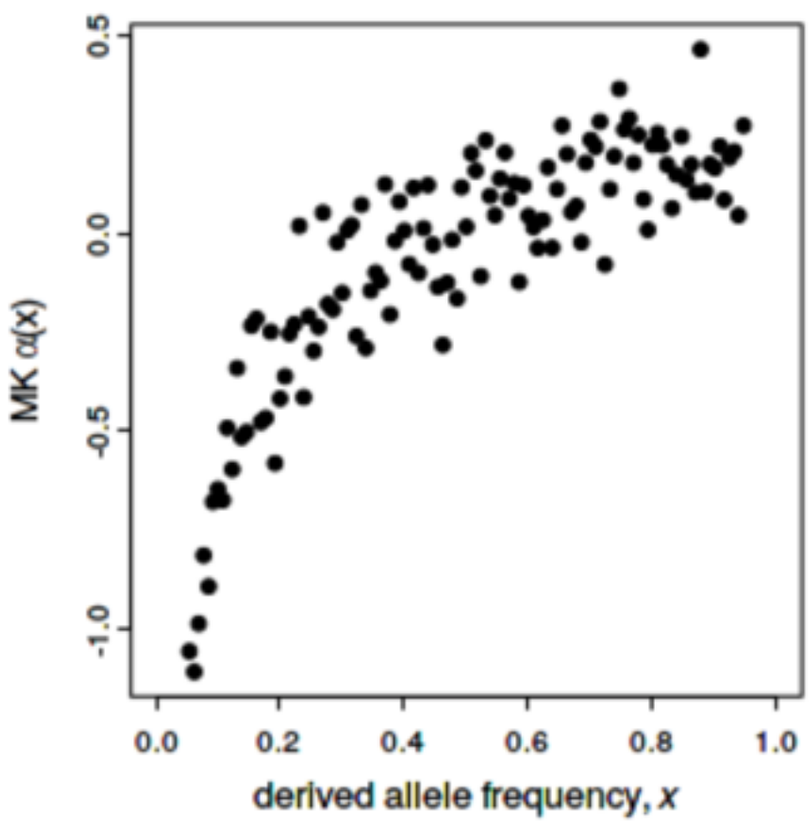
McDonald & Kreitman (MK)

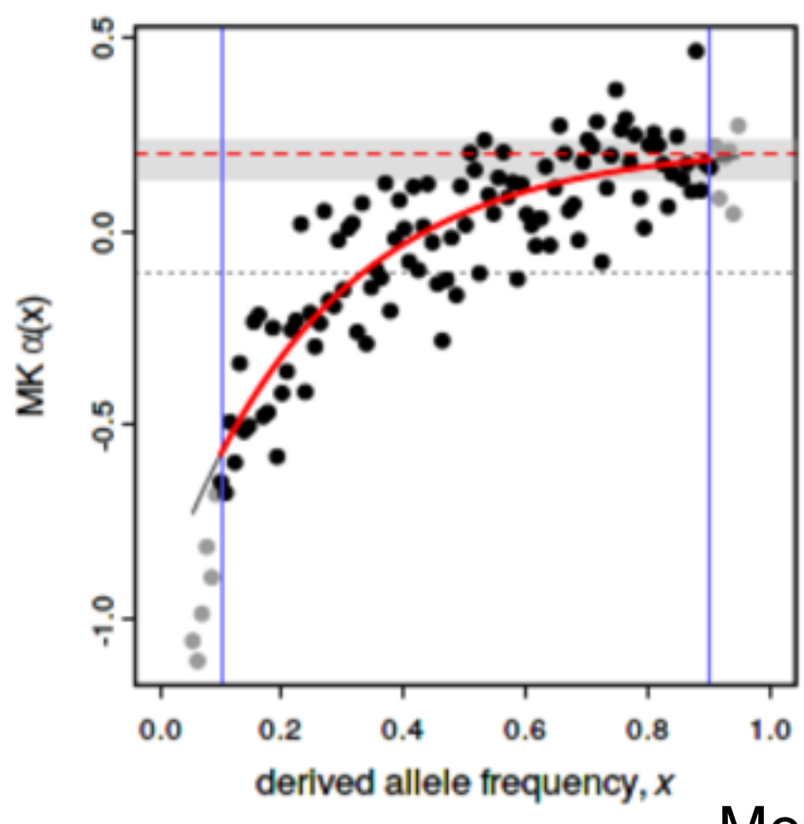
Estimate the proportion of non-synonymous substitutions driven by positive selection (alpha).

asymptoticMK

$$\alpha(x) = 1 - \frac{d_0}{d}\frac{p(x)}{p_0(x)}.$$

Fit an exponential function and extrapolate x=1





Messer and Petrov, PNAS 2013

asymptoticMK

asymptoticMK: Asymptotic McDonald–Kreitman Test

By Benjamin C. Haller & Philipp W. Messer. Copyright © 2017 Philipp Messer.

See below for background and usage information. If you use this service, please cite our paper:

[not yet published, please check back for a citation...]

Submit your data:

d :

d_0 :

Input file : Choose File SFSasympMK_...an_no1.txt

(Tab-delimited with named columns for x, p, and p_0) [sample]

x interval to fit : [0.10 , 0.90]

Submit

	Fixed	Polymorphic
Synonymous	d0	pS
Nonsynonymous	d	pN

Haller and Messer, G3 2017

HKA (Hudson, Kreitman and Aguadé)

	Fixed	Polymorphic
Locus 1	*D1*	*P1*
Locus 2	*D2*	*P2*

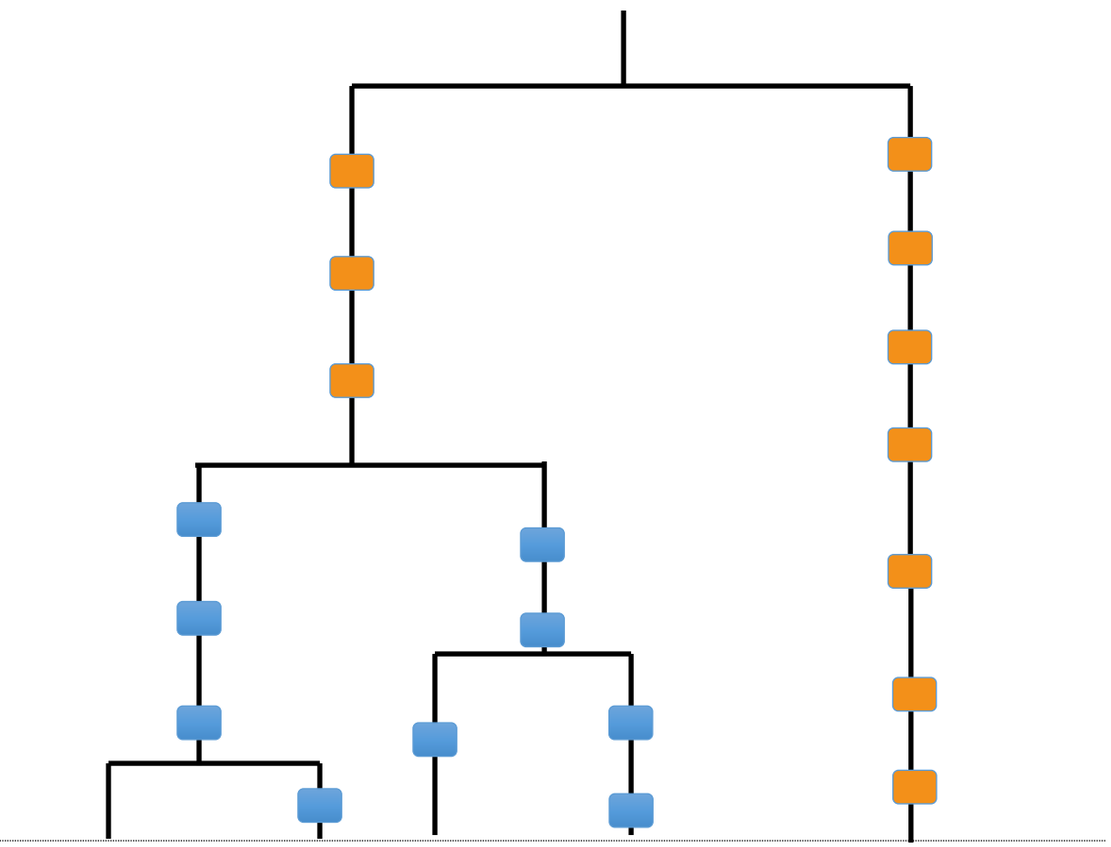
P1 / D1

P2 / D2

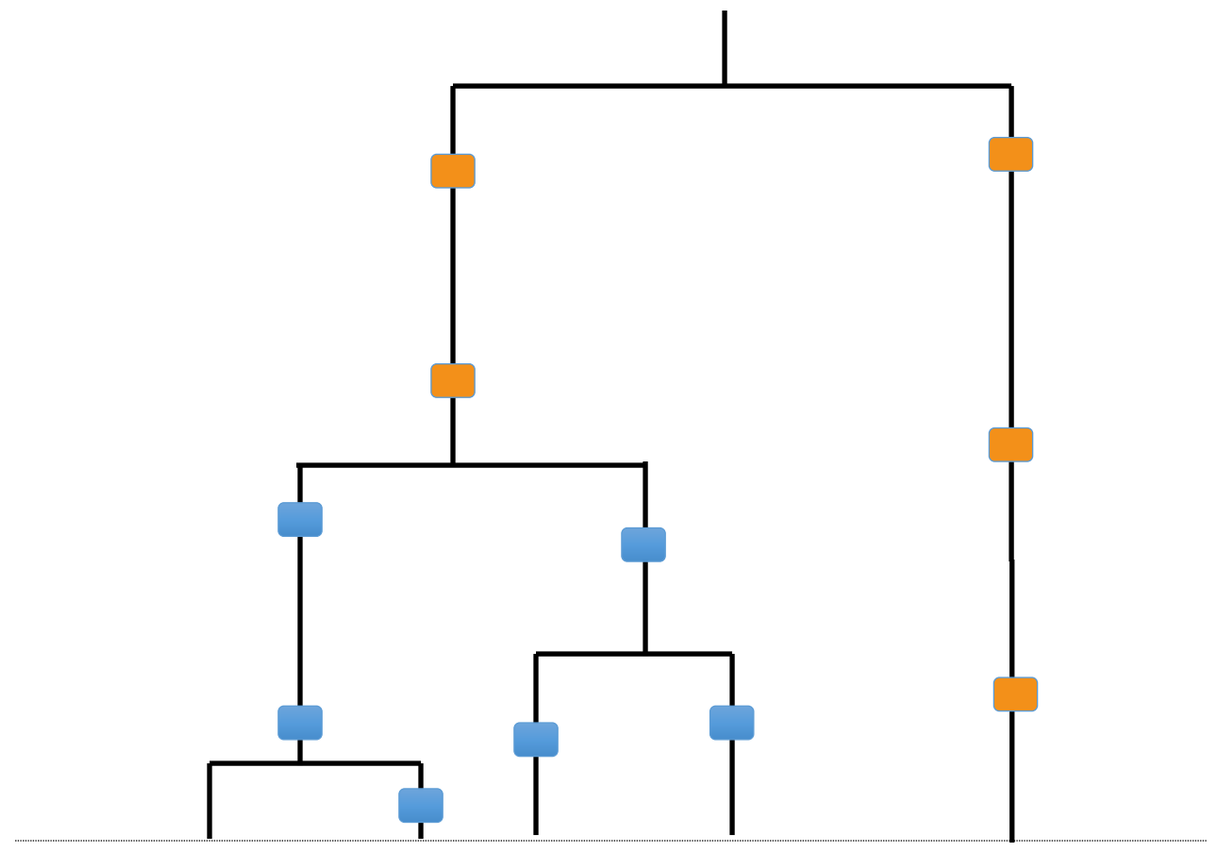
Hudson, Kreitman and Aguadé, Genetics 1987

HKA (Hudson, Kreitman and Aguadé)

Neutral



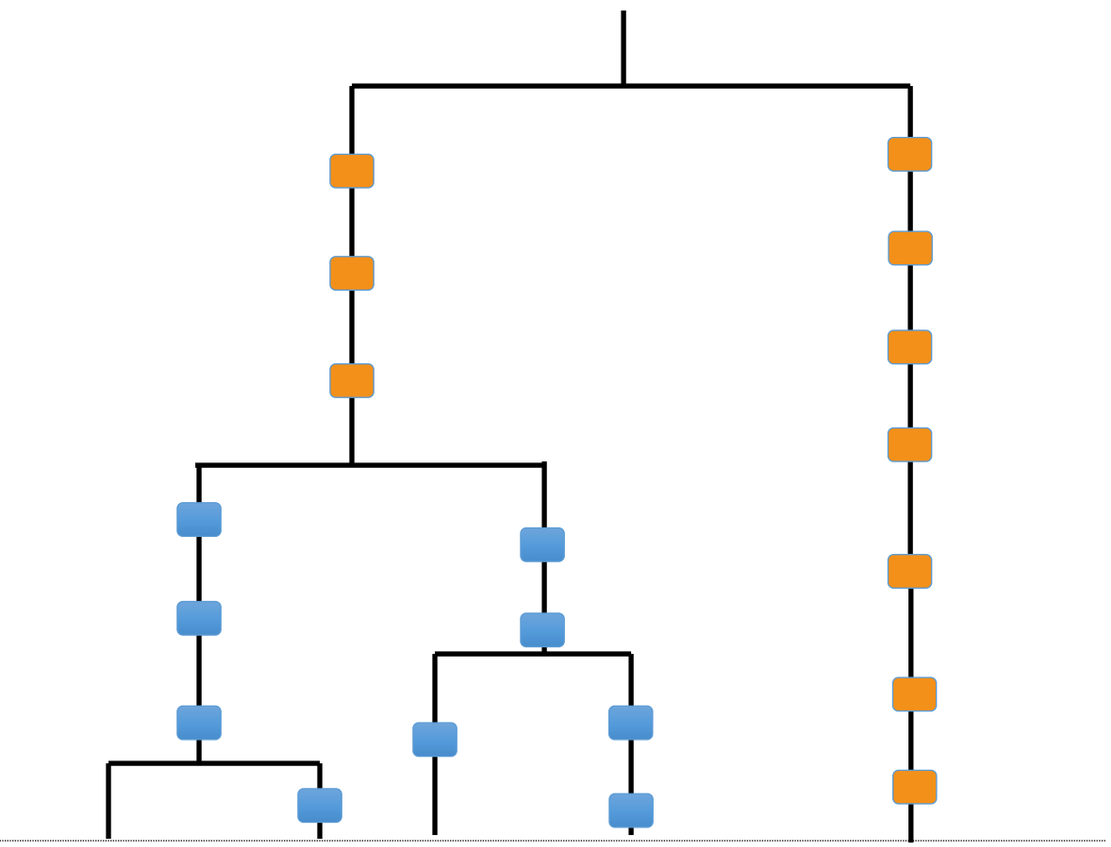
SNPs/FDs



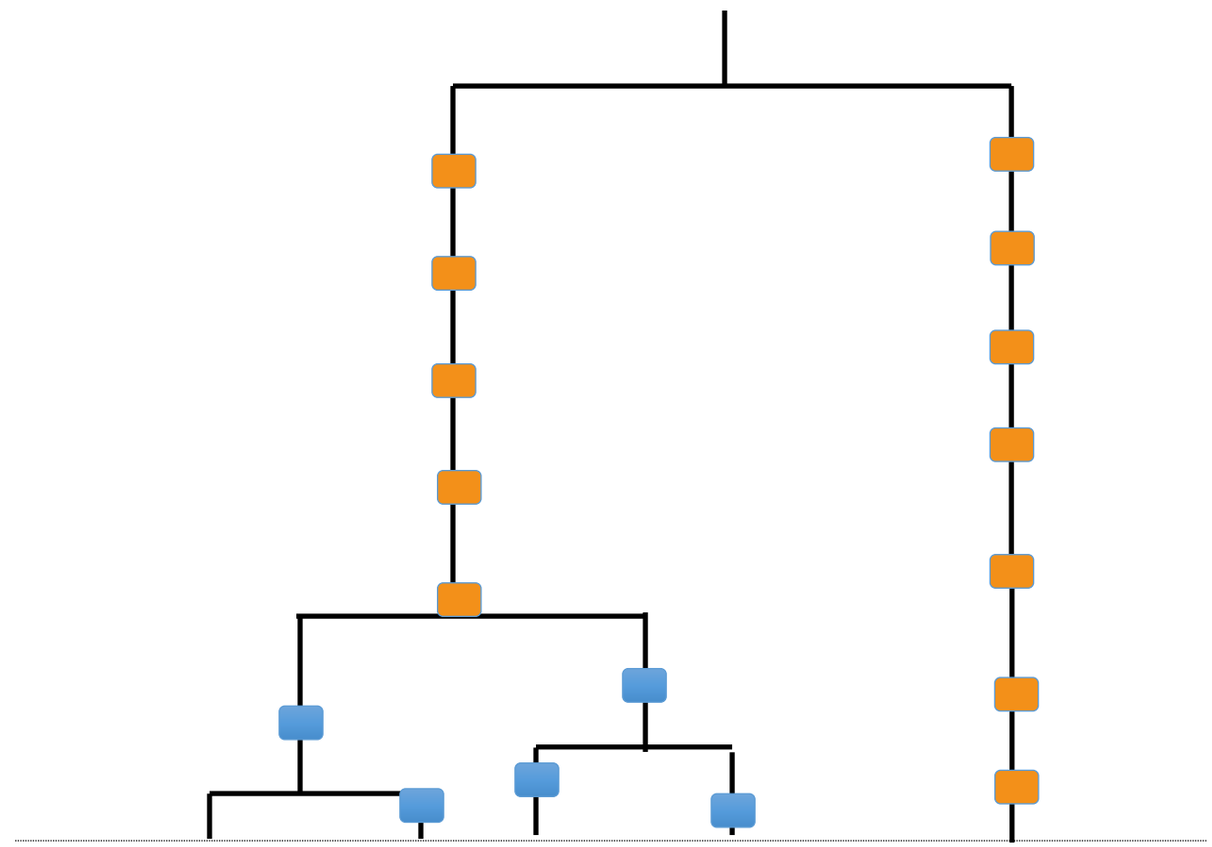
SNPs/FDs

HKA (Hudson, Kreitman and Aguadé)

Neutral



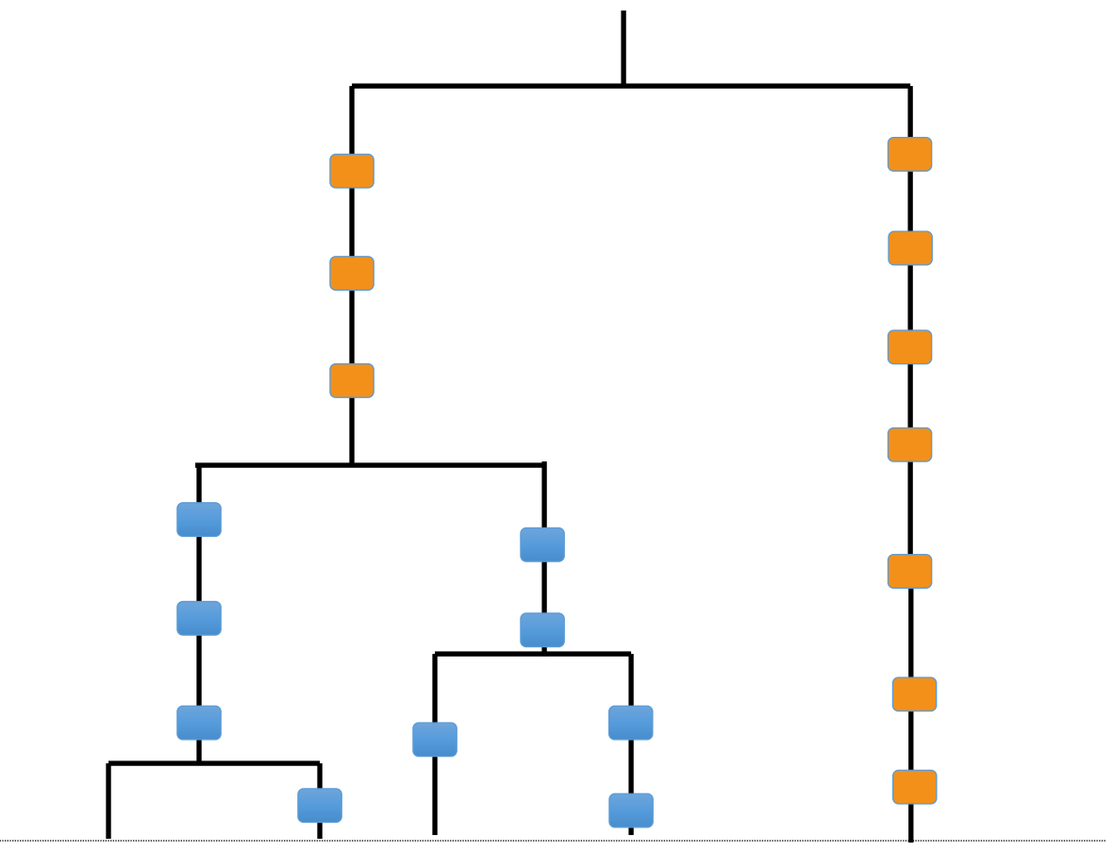
Positive selection



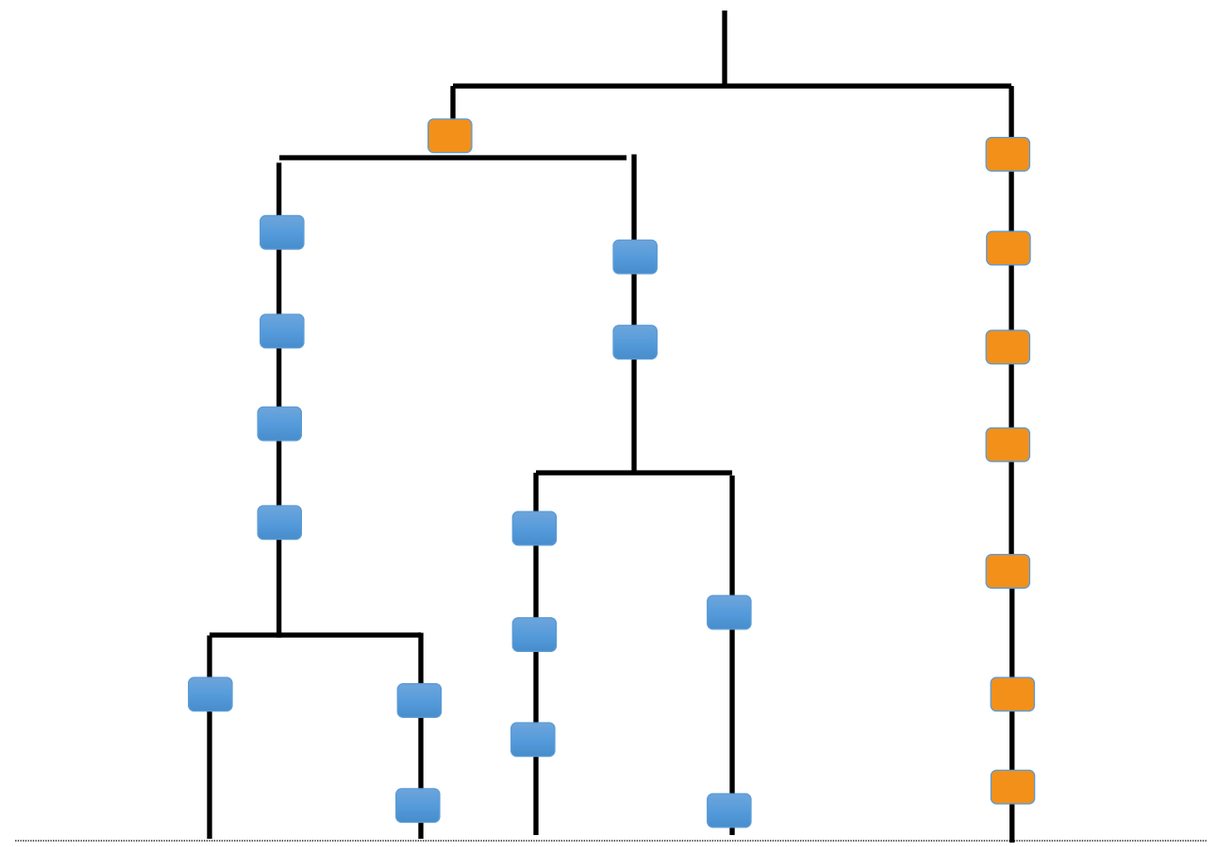
SNPs/FDs > SNPs/FDs

HKA (Hudson, Kreitman and Aguadé)

Neutral



Balancing selection



SNPs/FDs < SNPs/FDs

HKA (Hudson, Kreitman and Aguadé)

	Fixed	Polymorphic
Locus 1	*D1*	*P1*
Locus 2	*D2*	*P2*

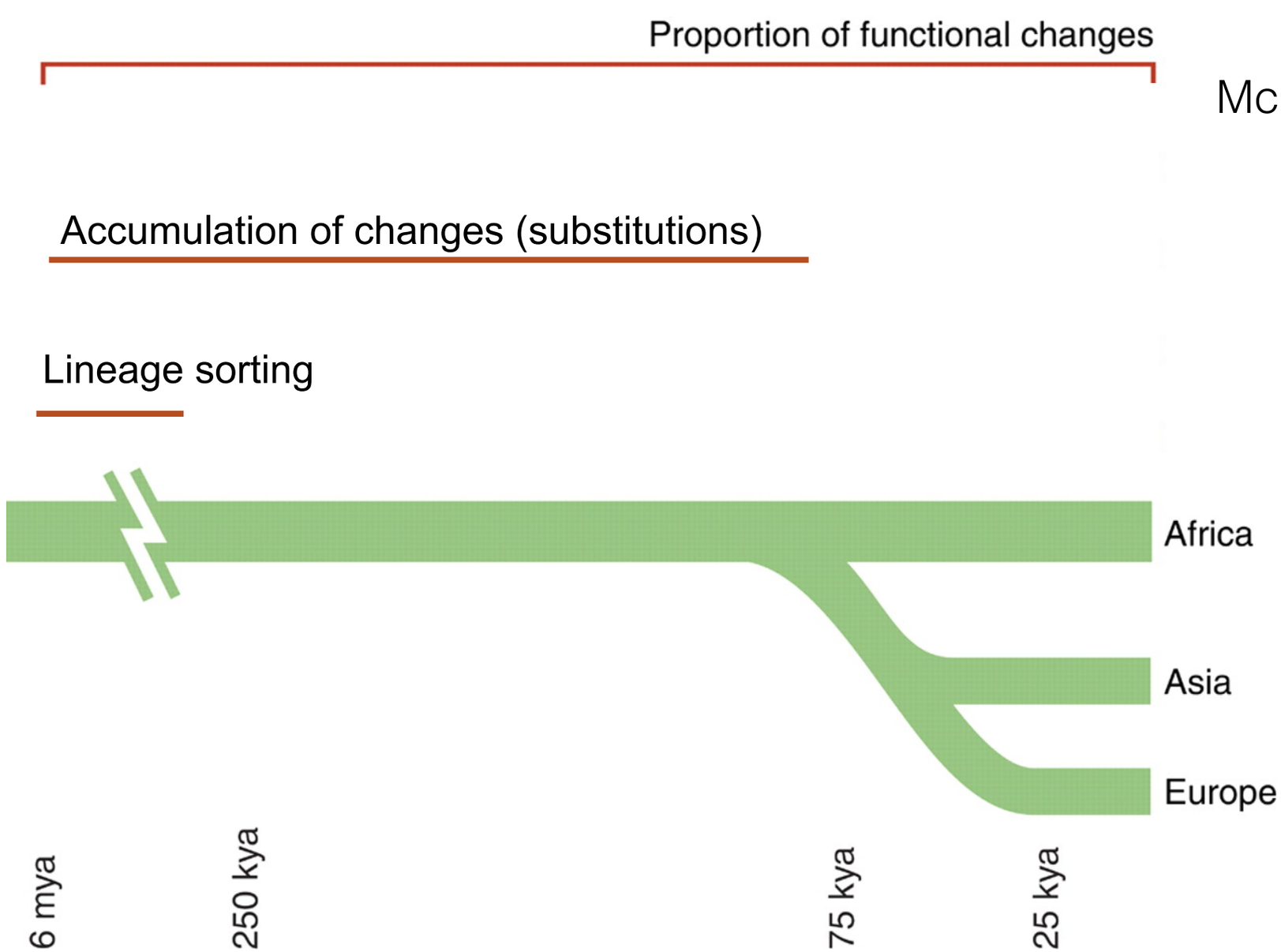
P1 / D1

P2 / D2

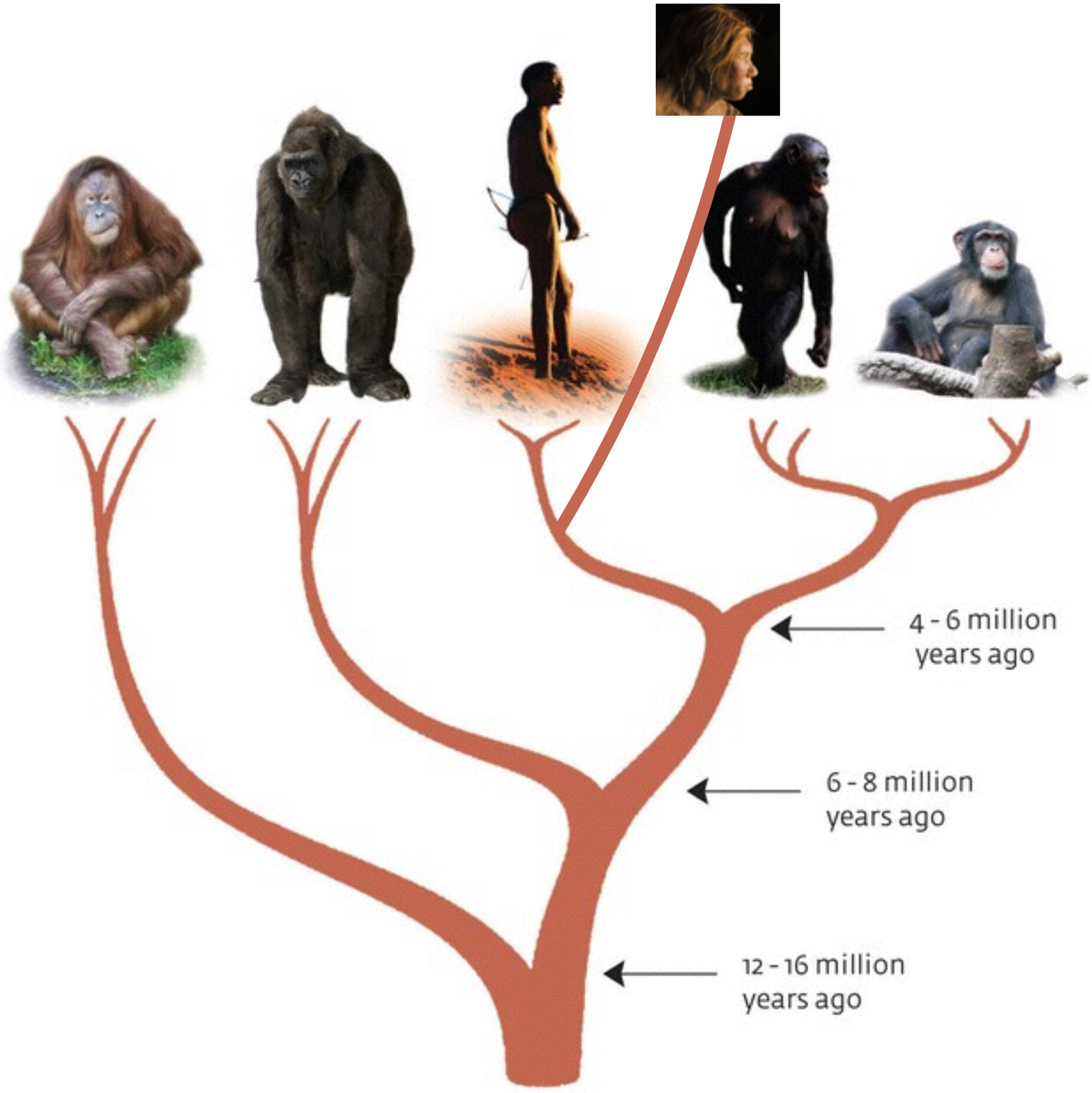
Test

Hudson, Kreitman and Aguadé, Genetics 1987

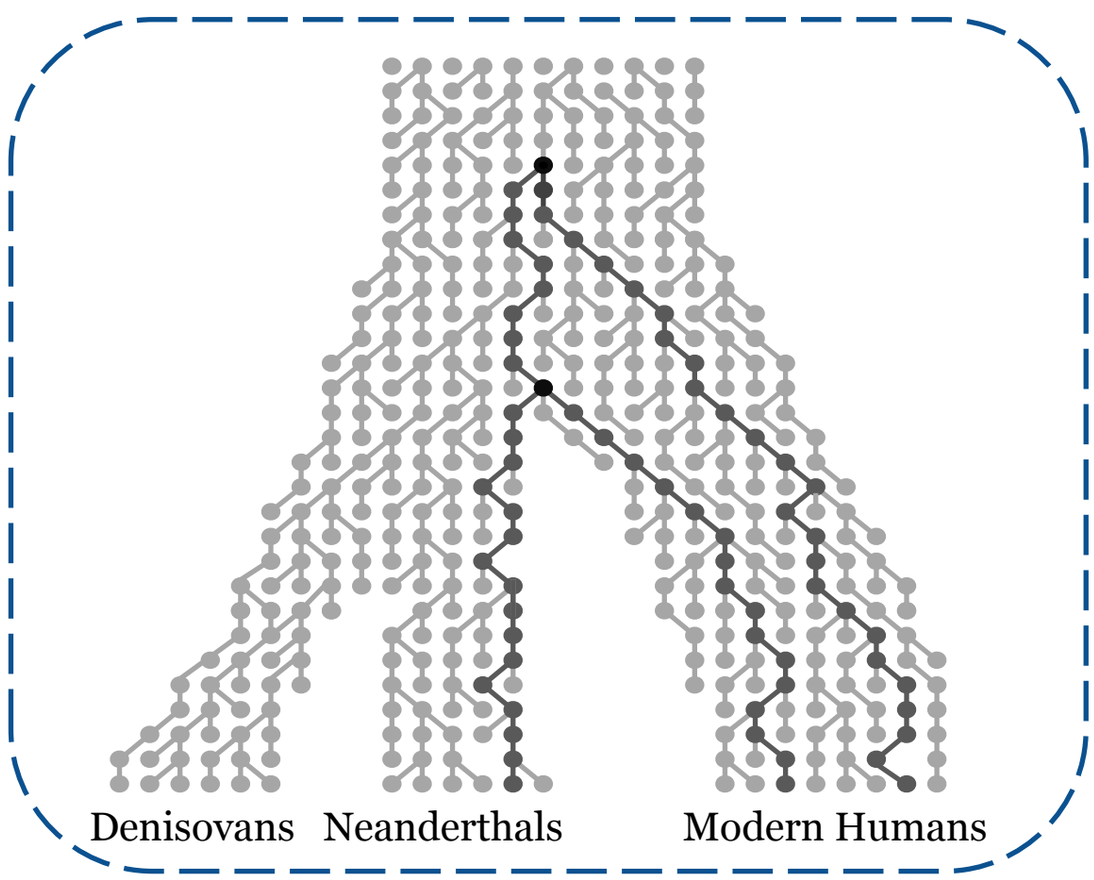
Time Scales for the Signatures of Selection



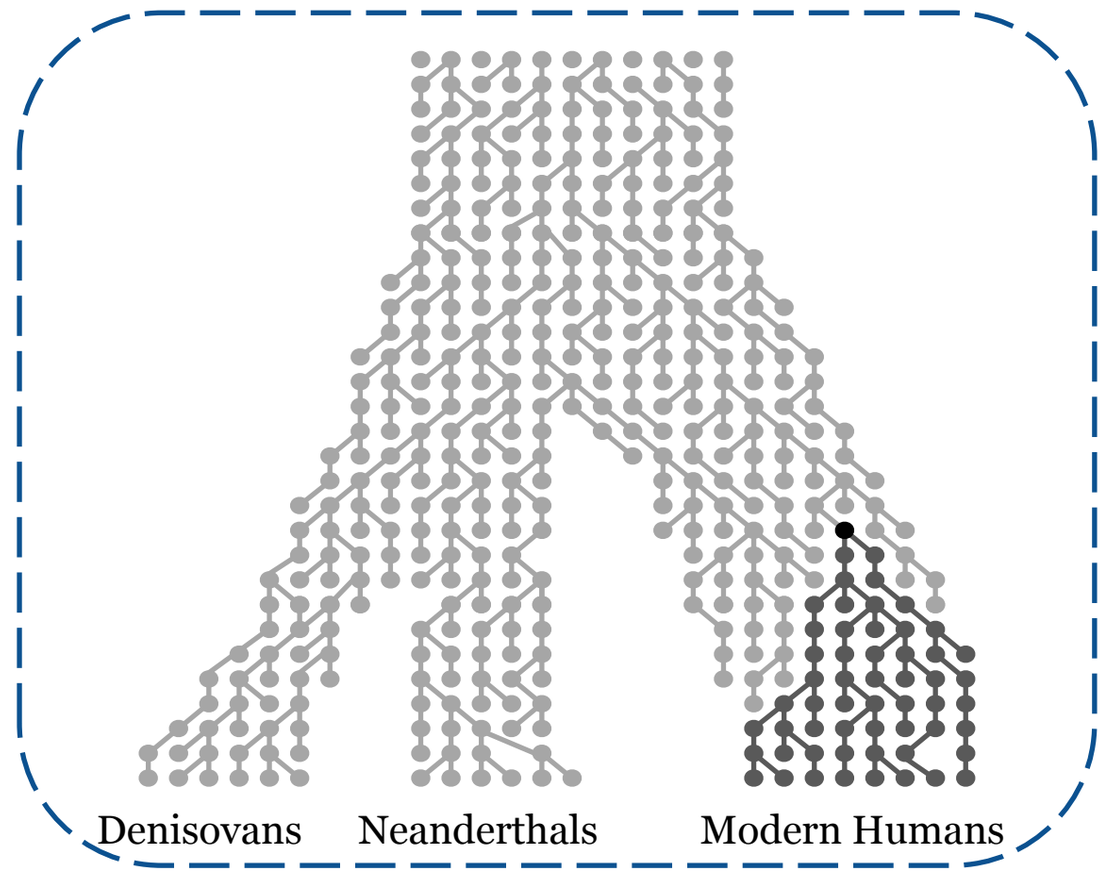
Sabeti et al., Science, 2006


4 - 6 million
years ago
6 - 8 million
years ago
12 - 16 million
years ago

Changes in Local Genealogies along the Genome

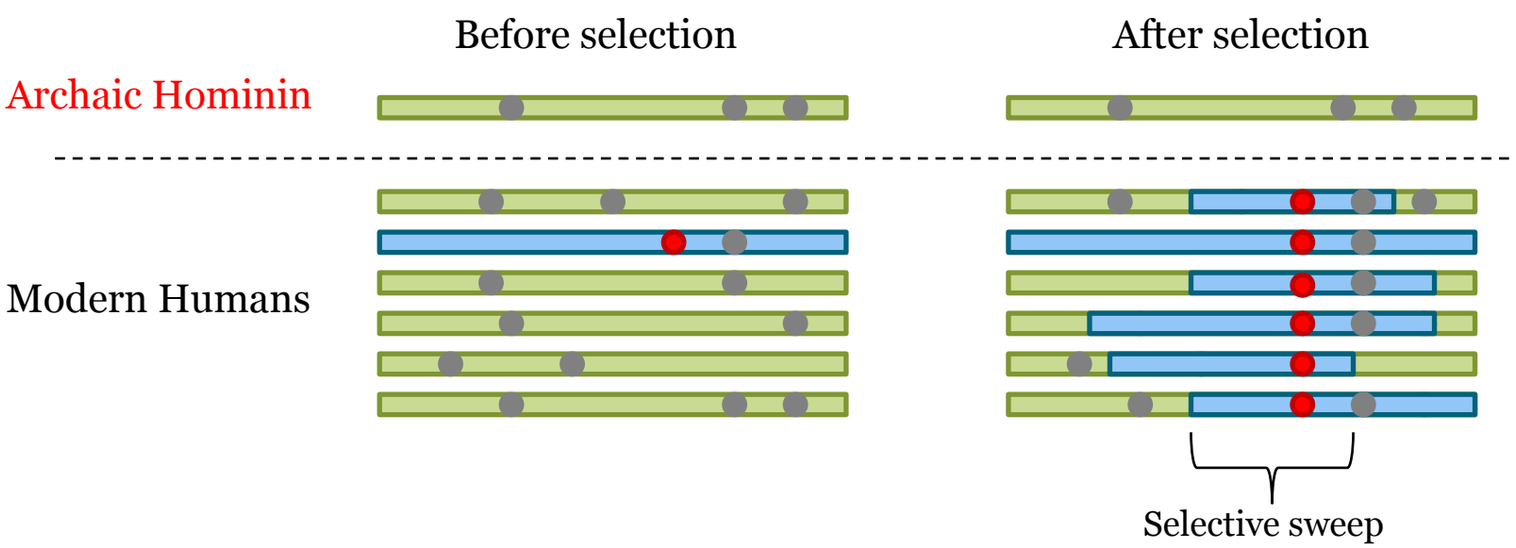


Internal regions (~90%)

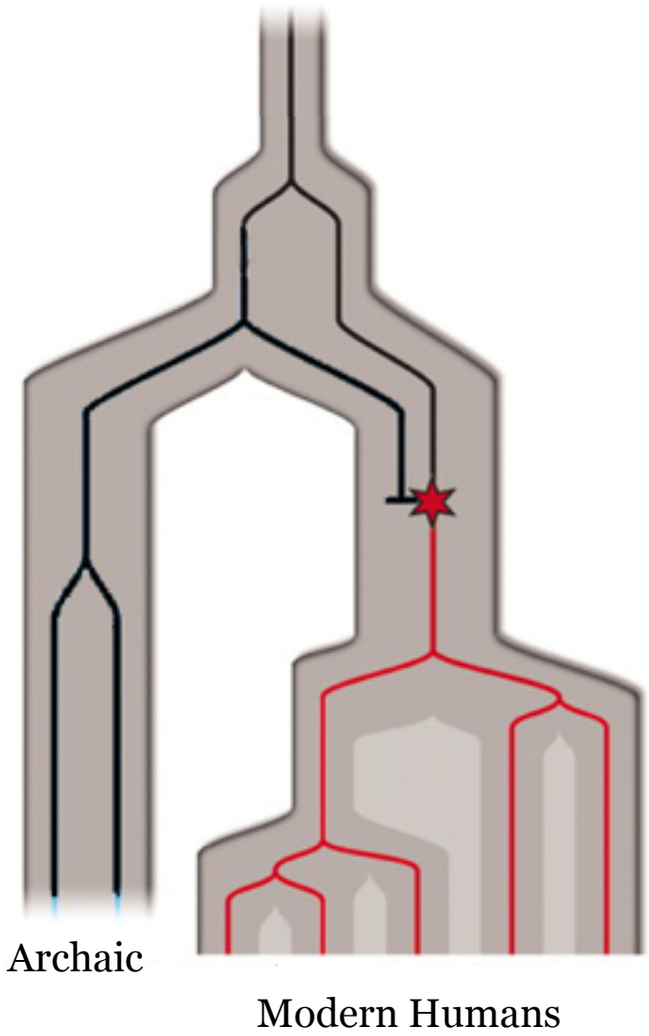


External regions (~10%)

Signal of Positive Selection

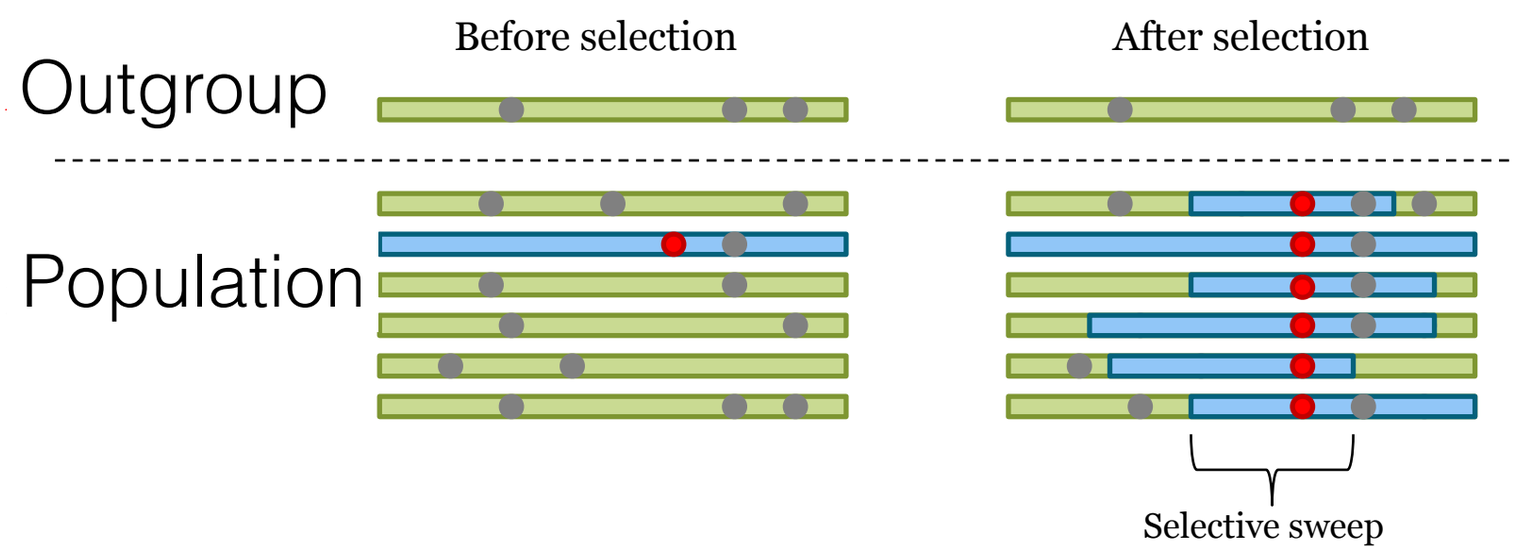


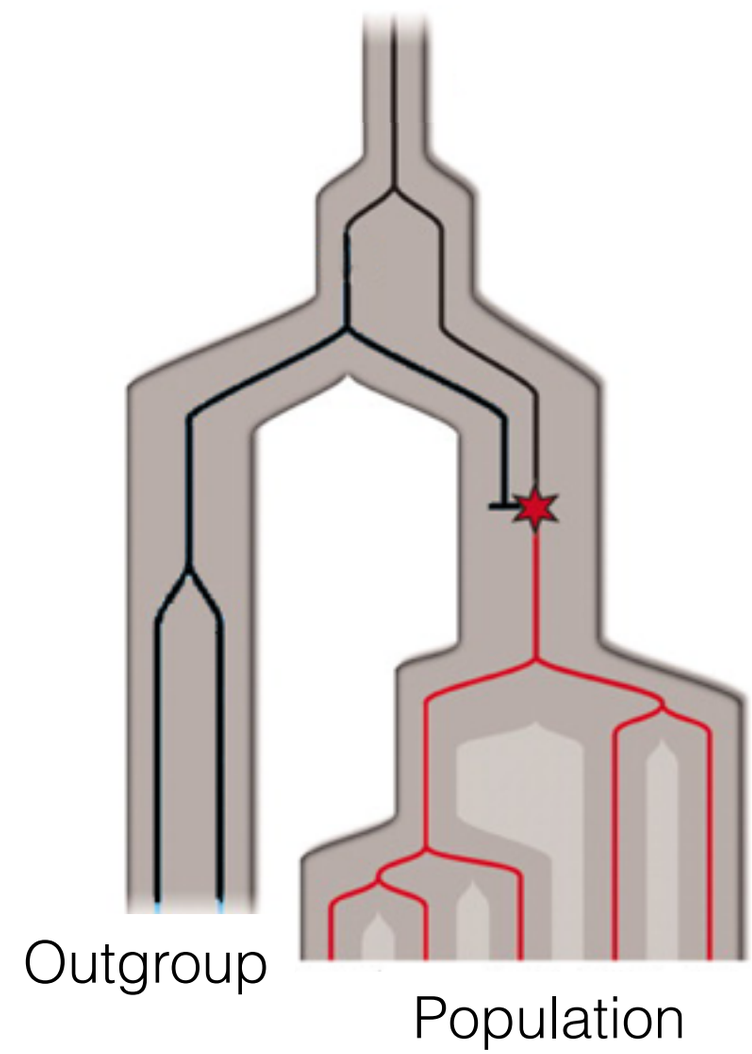
Length



Adapted from Green et al., Science, 2010

Signal of Positive Selection





We can use several outgroup individuals

Right only for some species/population pairs

Cagan et al., MBE 2016

Extended lineage sorting (ELS)

Data

Genotypes in the outgroup individual (A,D)

Allele frequency in your population.

Extended lineage sorting (ELS)

Relevant parameters

- Probability of the outgroup to share a derived allele in the population.
  - If external region, the probability is 0 (but there may be errors, e.g. genotyping)
  - If internal region, if the site is fixed in the population the probability is 1 (but there may be errors)
  - If internal region, if the site is polymorphic the probability depends on the age of the allele (proxy frequency)

Extended lineage sorting (ELS)

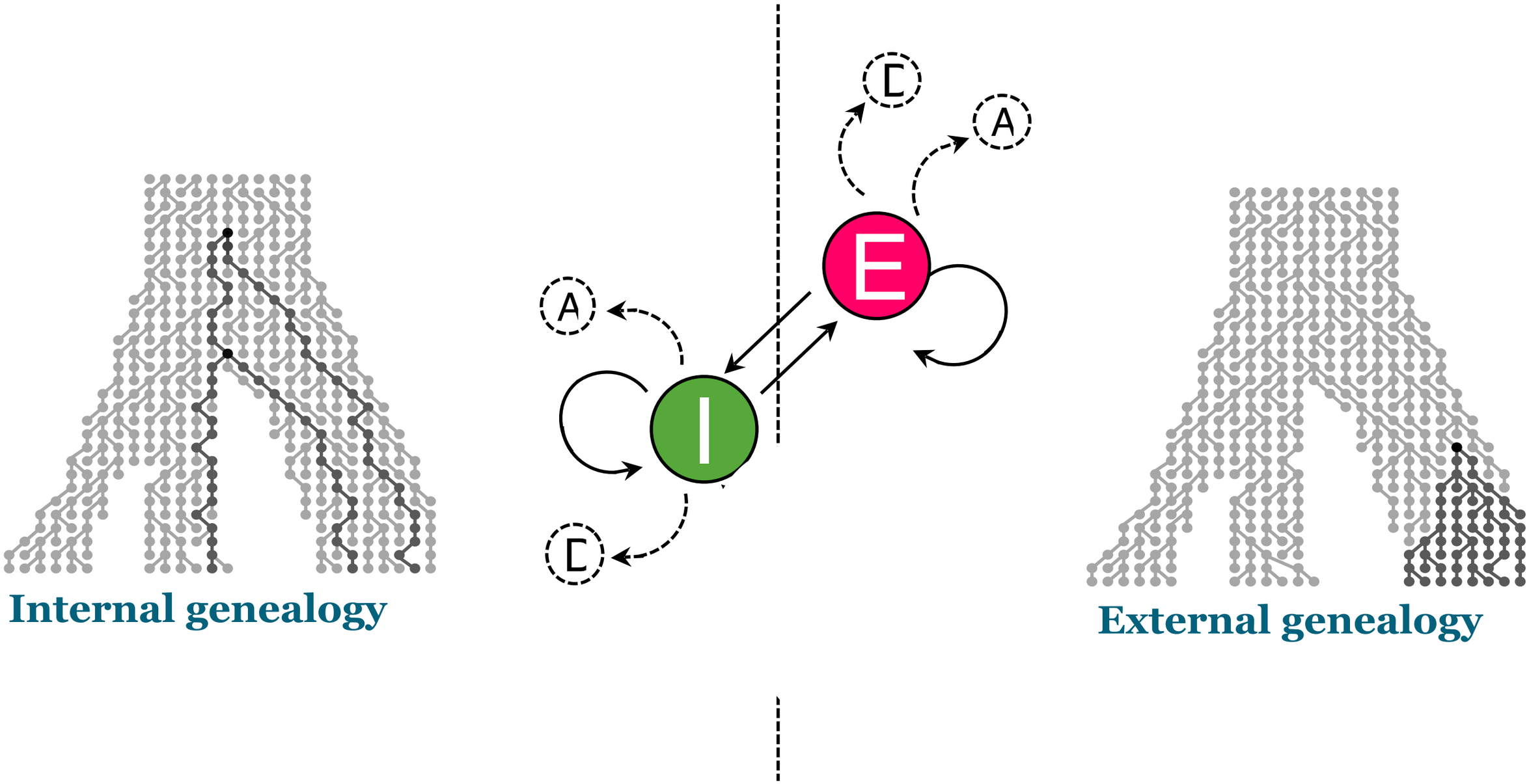
Relevant parameters

Probability of the outgroup to share a derived allele in the population.

Length of internal and external regions

Length of ELS regions

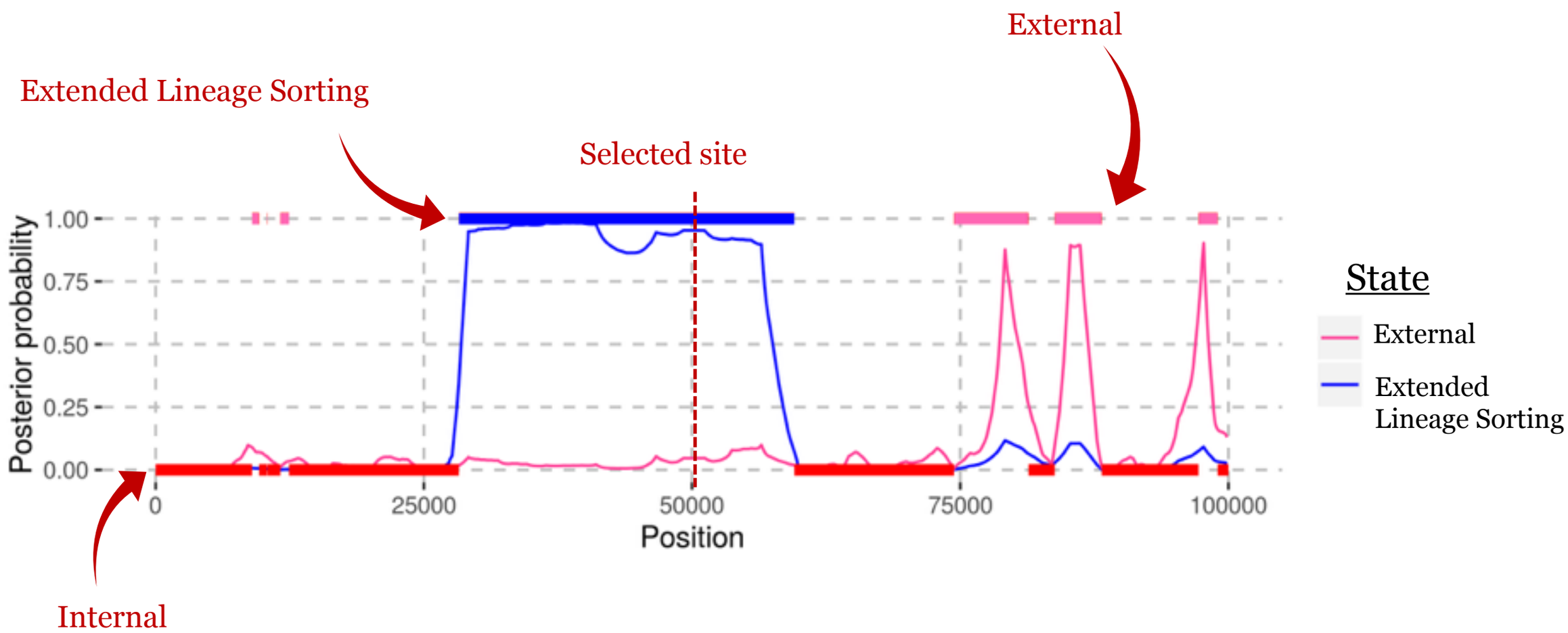
A Hidden Markov Model to Detect Extended Lineage Sorting



Estimate for each position the probability of each state

Peyregne et al., Genome Research 2017

Detection of Extended Lineage Sorting



Peyregne et al., Genome Research 2017

We obtain

Statistical signatures of natural selection

Thanks to

Stephane Peyregne

Gabriel Santpere

Joshua Schmidt

Philip Messer

asymptoticMK

asymptoticMK: Asymptotic McDonald–Kreitman Test

By Benjamin C. Haller & Philipp W. Messer. Copyright © 2017 Philipp Messer.

See below for background and usage information. If you use this service, please cite our paper:

[not yet published, please check back for a citation...]

Submit your data:

d :

d_0 :

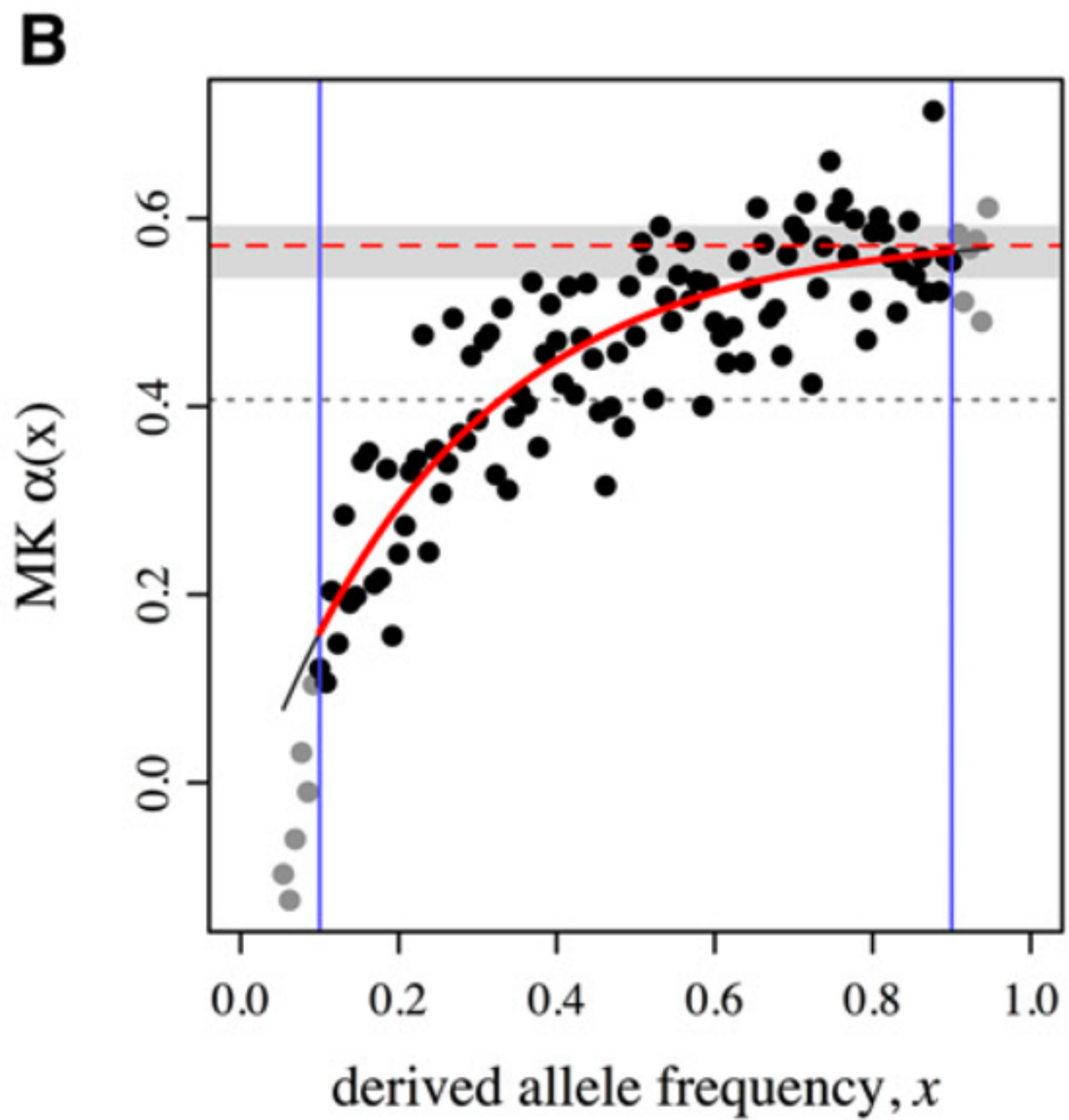
Input file : Choose File SFSasympMK_...an_no1.txt

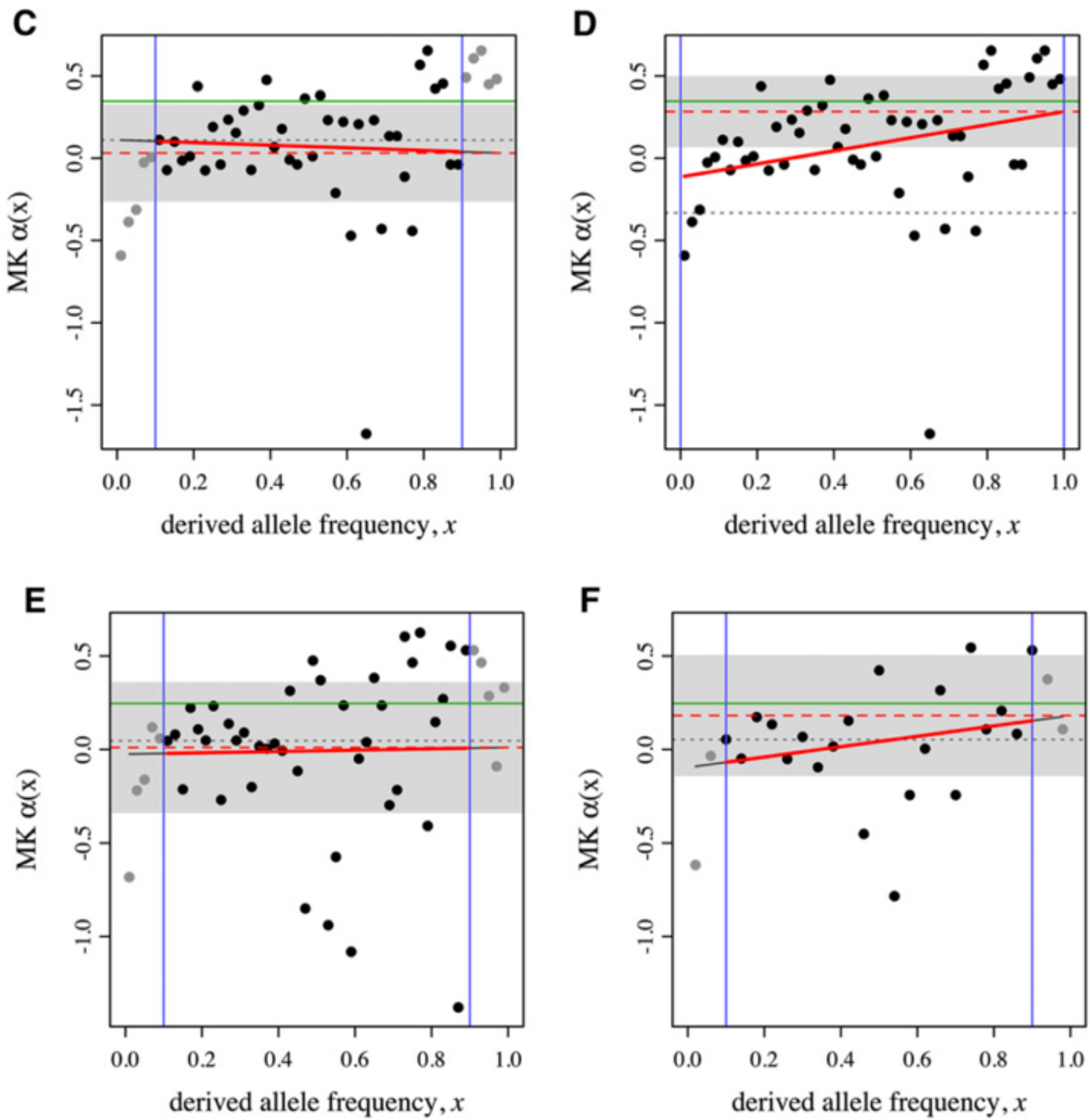
(Tab-delimited with named columns for x, p, and p_0) [sample]

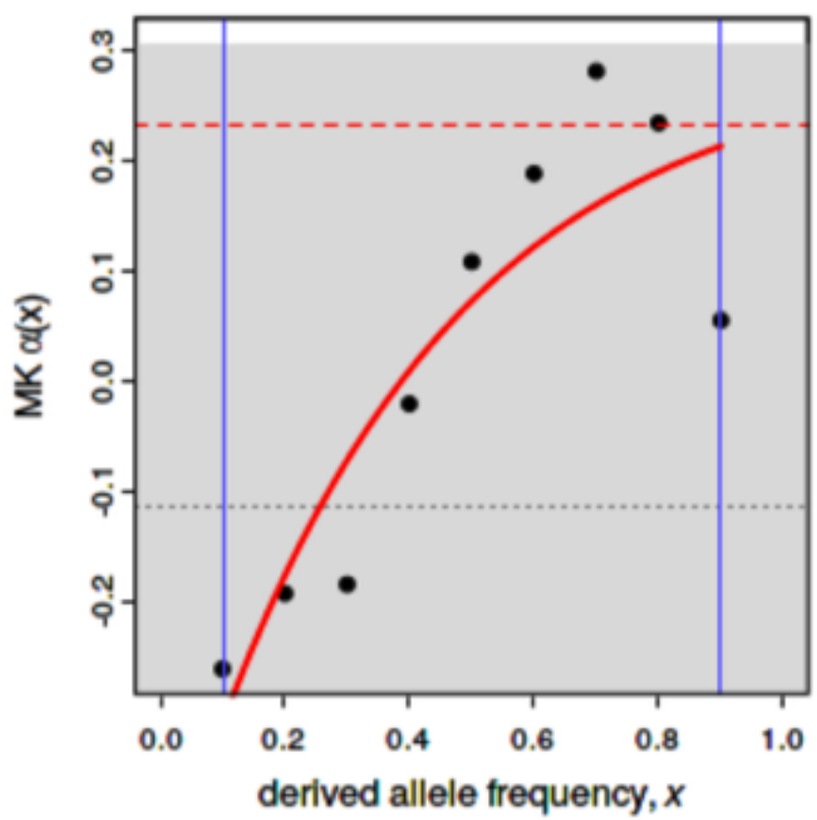
x interval to fit : [0.10 , 0.90]

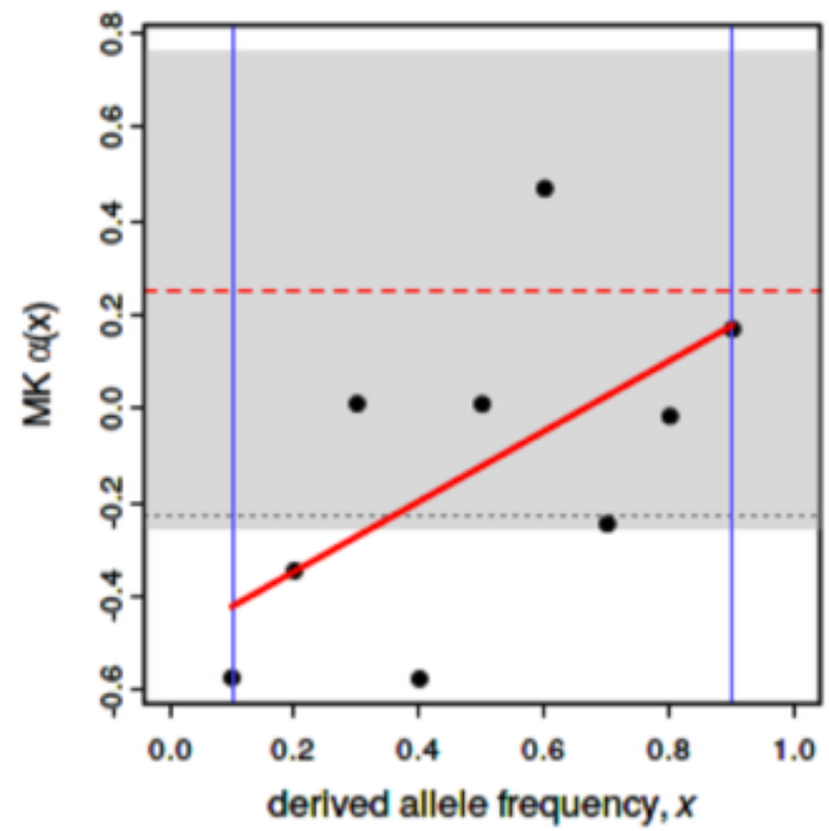
Submit

asymptoticMK




C
D
E
F
MK α(x)
derived allele frequency, x
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.2
0.4
0.6
0.8
1.0


MK α(x)
0.3
0.2
0.1
0.0
-0.1
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
derived allele frequency, x


MK α(x)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.2
0.4
0.6
0.8
1.0
derived allele frequency, x

For individual genes?

For sets of genes?

Extended lineage sorting (ELS)

Process

- For efficiency purposes we have the genetic data & configuration file, and the parameters previously inferred.

- Run ELS_HMM to calculate the probability at each site of it being internal, external or ELS.

- Combine sites into internal, external and ELS regions

- Use length to prioritise ELS regions & visualise

- Compare across outgroup individuals

Peyregne et al., Genome Research 2017