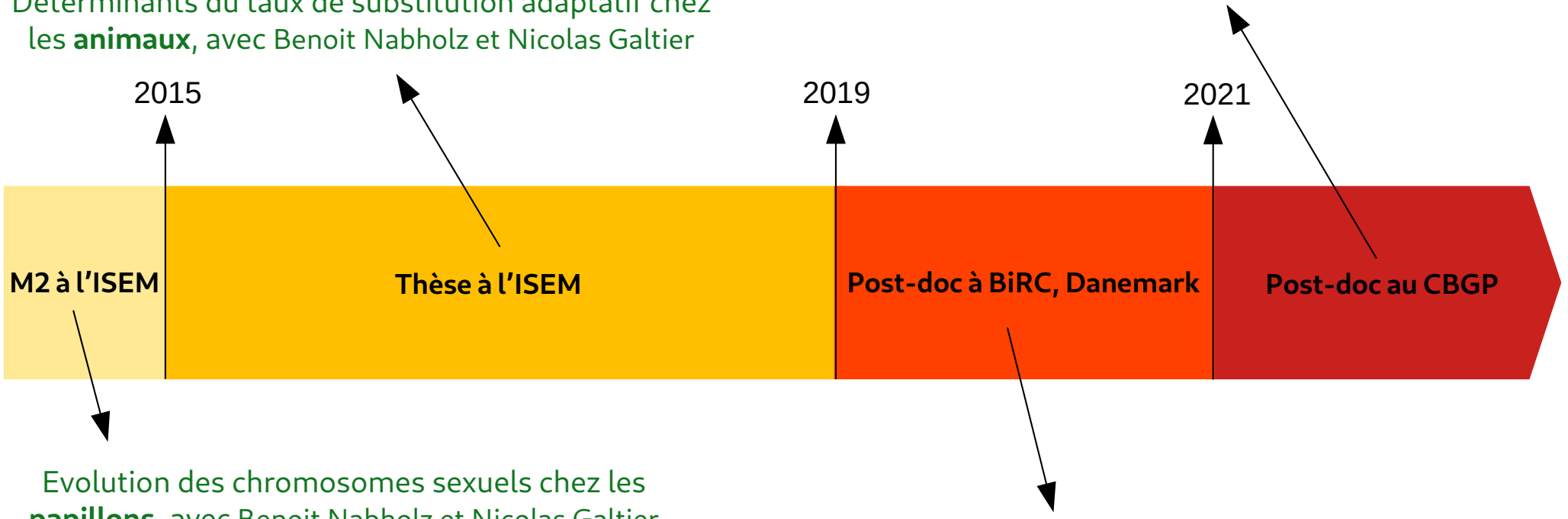
POPULATION SIZE, INCOMPLETE LINEAGE SORTING AND SELECTION IN ANIMAL GENOMES

Marjolaine Rousselle

marjolaine.rousselle@inrae.fr
@MarjoRousselle

Séminaire du CBGP-19/10/2021

Mon expérience :



Déterminants du taux de substitution adaptatif chez les **animaux**, avec Benoit Nabholz et Nicolas Galtier

Histoire démographique et dynamique de la différenciation vers la spéciation chez le **puceron du pois**, avec Emmanuelle Jousselin, Carole Smadja, Mathieu Gautier et Renaud Vitalis

Evolution des chromosomes sexuels chez les **papillons**, avec Benoit Nabholz et Nicolas Galtier

Tri de ligné incomplet et évolution des chromosomes sexuels chez les **primates**, avec Mikkel H. Schierup

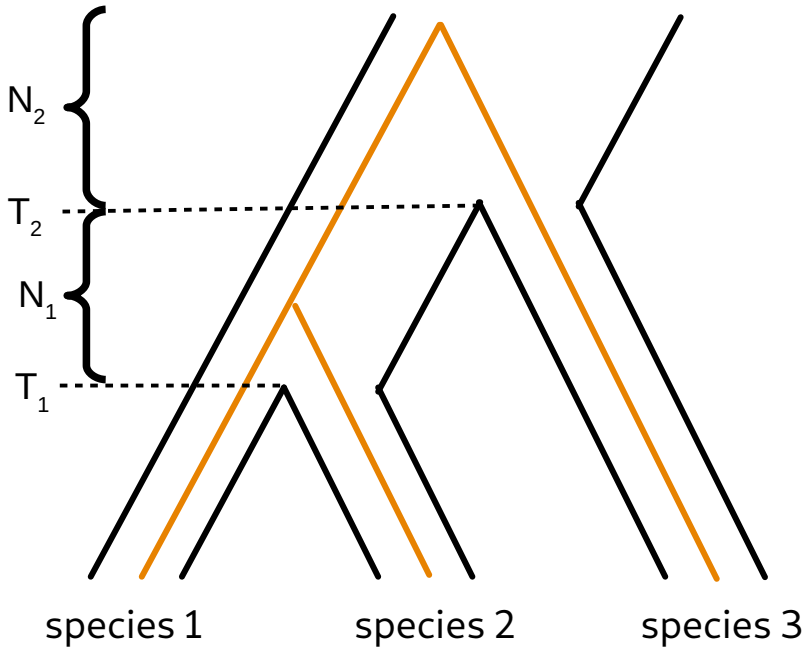
POPULATION SIZE, INCOMPLETE LINEAGE SORTING AND SELECTION IN ANIMAL GENOMES

Part 1 : Reconstruction of **ancestral population sizes** and **speciation times** in the primate phylogeny by studying the **genealogy fo sequences along the genome**

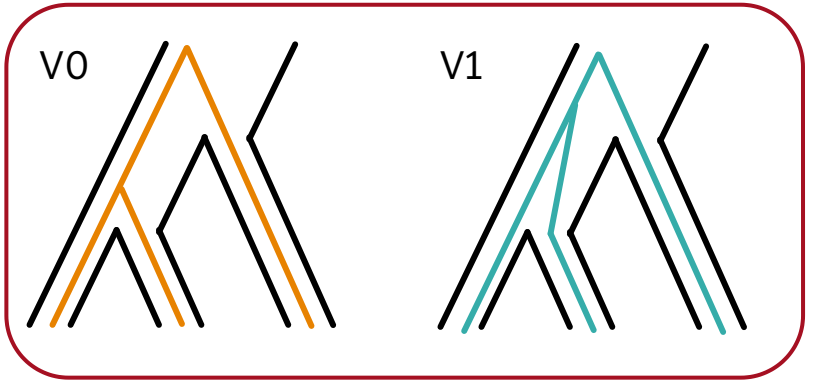
→ Incomplete lineage sorting (ILS)

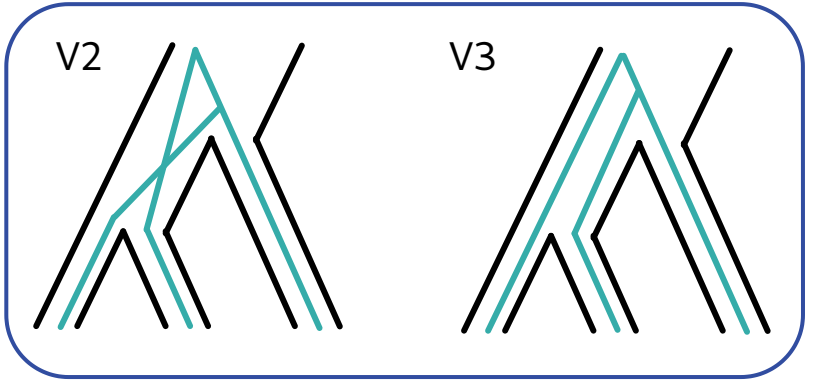
Deep coalescence

→ incomplete lineage sorting



Canonical topology

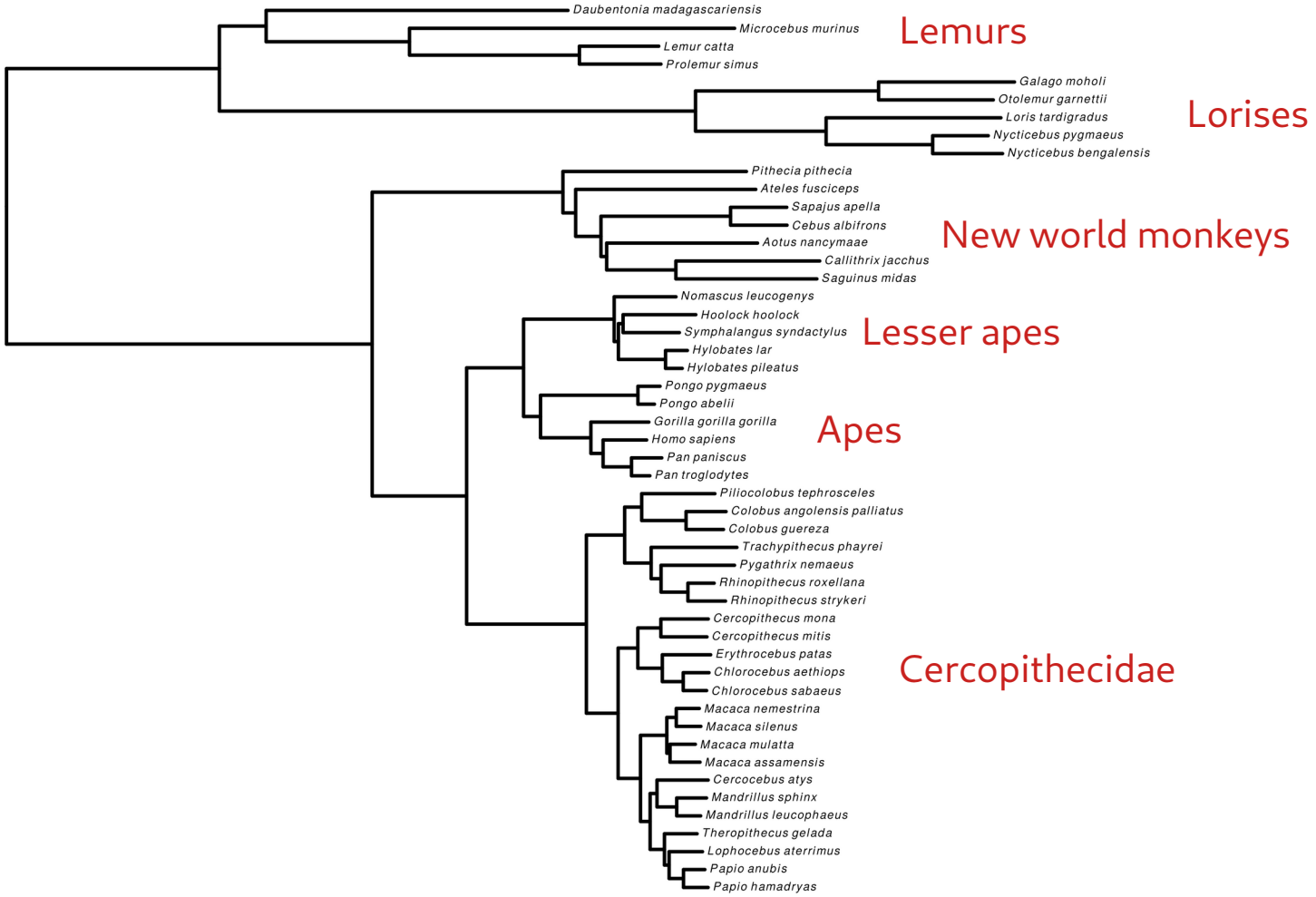




Alternative topologies

$$P_{incongruence} = \frac{2}{3} e^{\frac{T_2 - T_1}{2N_1 g}}$$

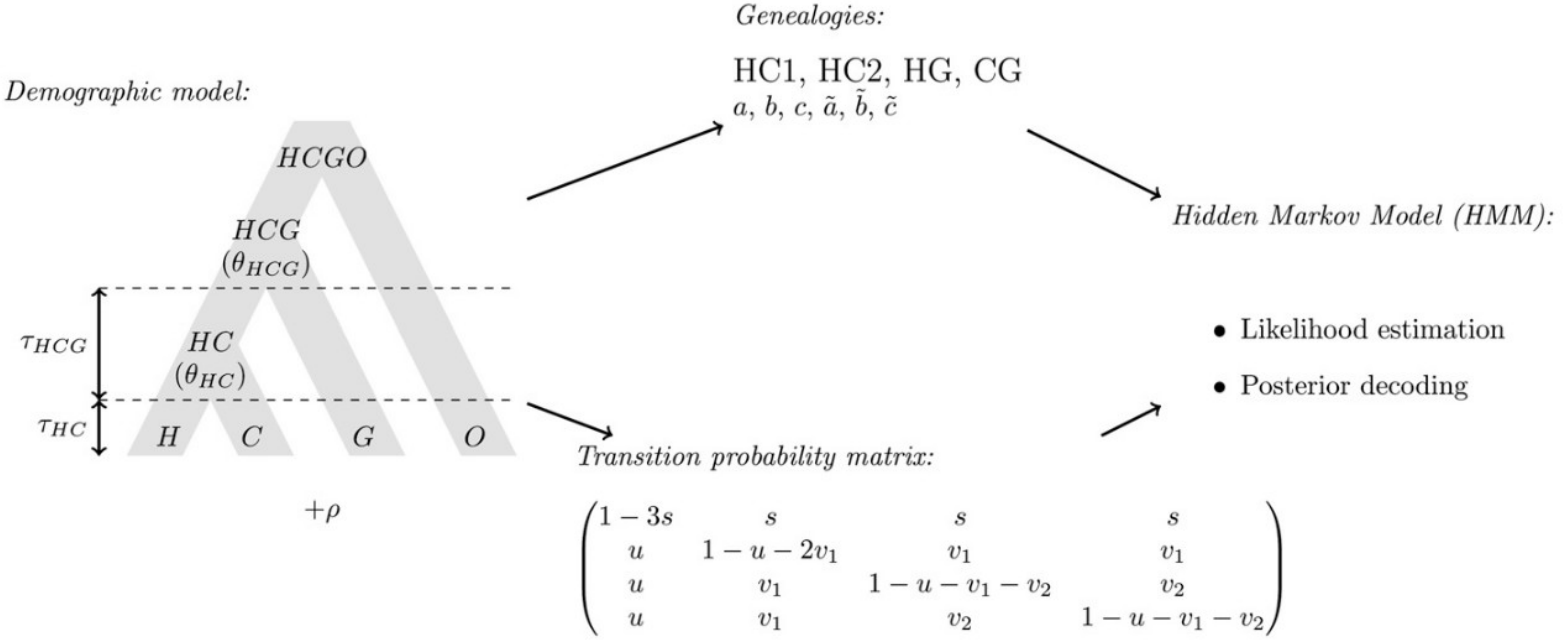
Data set : **genome-wide alignment of 46 primates**



Guojie Zhang,
Copenhagen University

& cie

CoalHMM framework



From Dutheil et al. 2009

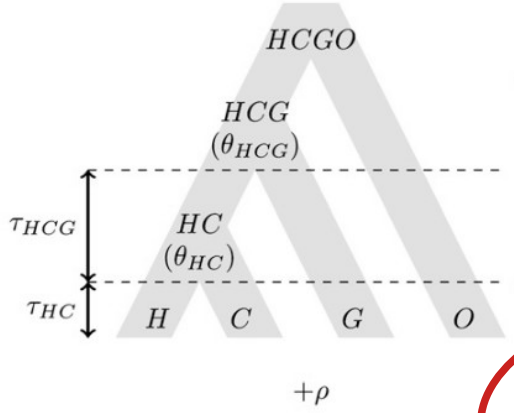
CoalHMM framework

Genealogies:

HC1, HC2, HG, CG

$a, b, c, \tilde{a}, \tilde{b}, \tilde{c}$

Demographic model:



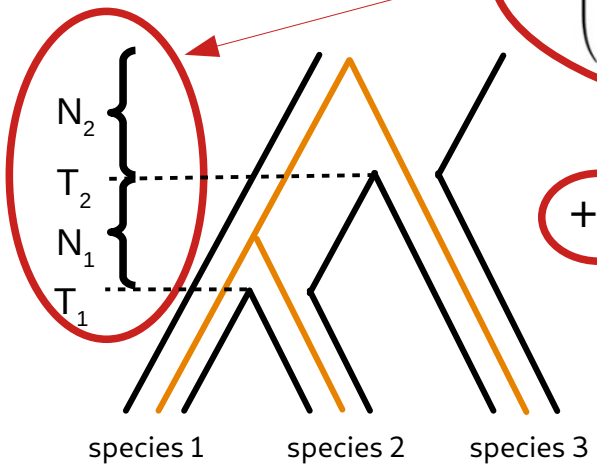
Hidden Markov Model (HMM):

- Likelihood estimation
- Posterior decoding

Transition probability matrix:

$$\begin{pmatrix} 1-3s & s & s & s \\ u & 1-u-2v_1 & v_1 & v_1 \\ u & v_1 & 1-u-v_1-v_2 & v_2 \\ u & v_1 & v_2 & 1-u-v_1-v_2 \end{pmatrix}$$

From Dutheil et al. 2009



+ρ

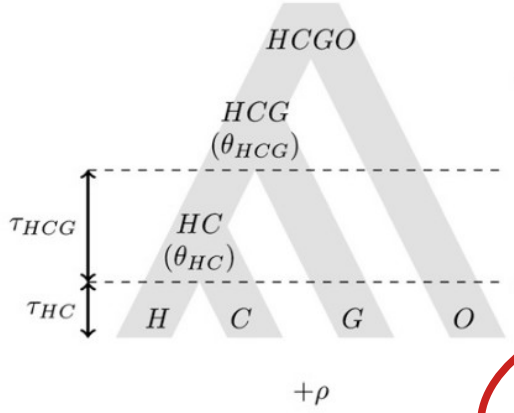
CoalHMM framework

Genealogies:

HC1, HC2, HG, CG

$a, b, c, \tilde{a}, \tilde{b}, \tilde{c}$

Demographic model:



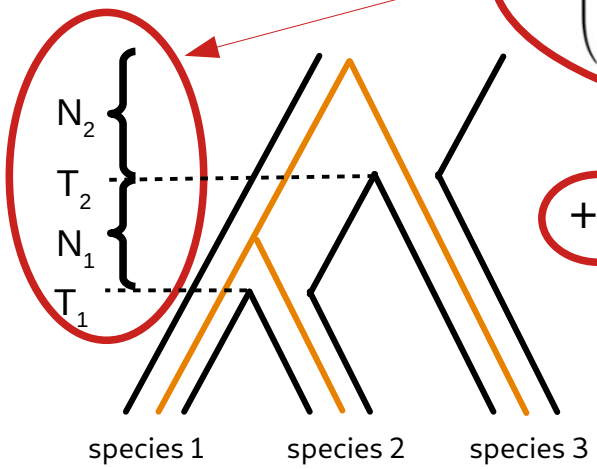
Hidden Markov Model (HMM):

- Likelihood estimation
- Posterior decoding

Transition probability matrix:

$$\begin{pmatrix} 1-3s & s & s & s \\ u & 1-u-2v_1 & v_1 & v_1 \\ u & v_1 & 1-u-v_1-v_2 & v_2 \\ u & v_1 & v_2 & 1-u-v_1-v_2 \end{pmatrix}$$

From Dutheil et al. 2009

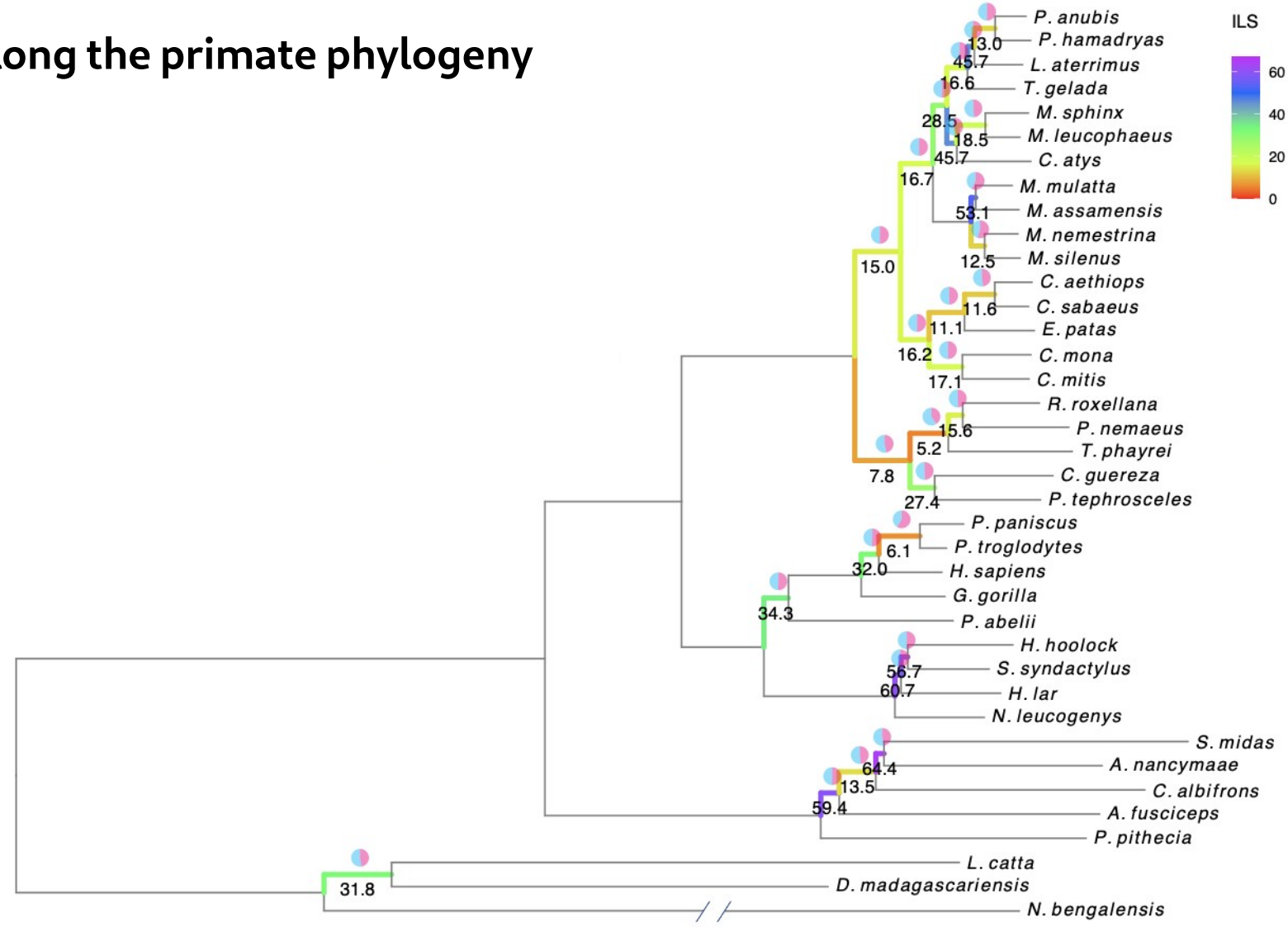


+ρ

Iker Rivas-González

rivasiker / autocoalhmm

Pervasive ILS along the primate phylogeny



Ancestral population sizes and speciation times

→ debiasing
→ μ and g to get absolute estimates

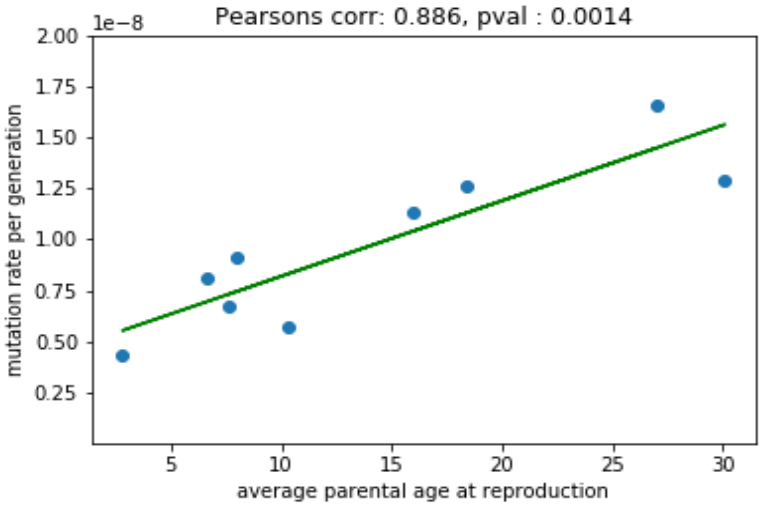
-Using only 4 possible genealogies=bias coalescent parameters (from Dutheil et al. 2009)
→ **Simulations + random forest to predict the biases on data**

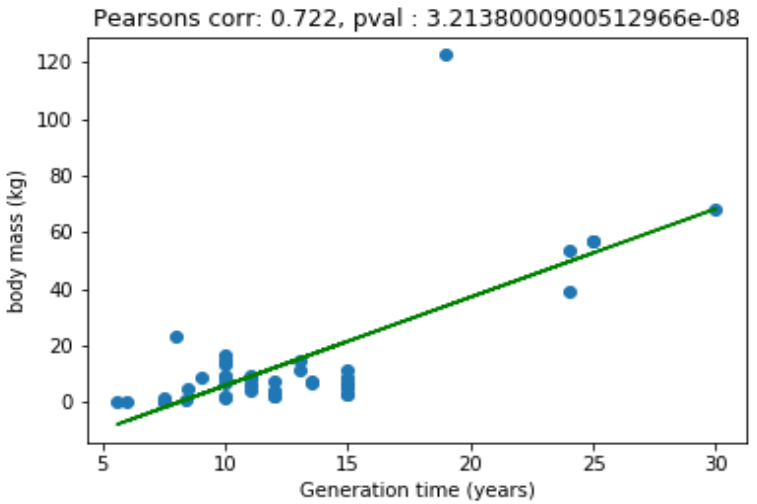
-What yearly mutation rate to use?
→ Reconstruction of **yearly mutation rate** for **extent** and **ancient** species via the reconstruction of **body size** and **generation time.**

Ancestral population sizes and speciation times

-What yearly mutation rate to use?

→ Reconstruction of **yearly mutation rate** for **extent** and **ancient** species via the reconstruction of **body size** and **generation time.**



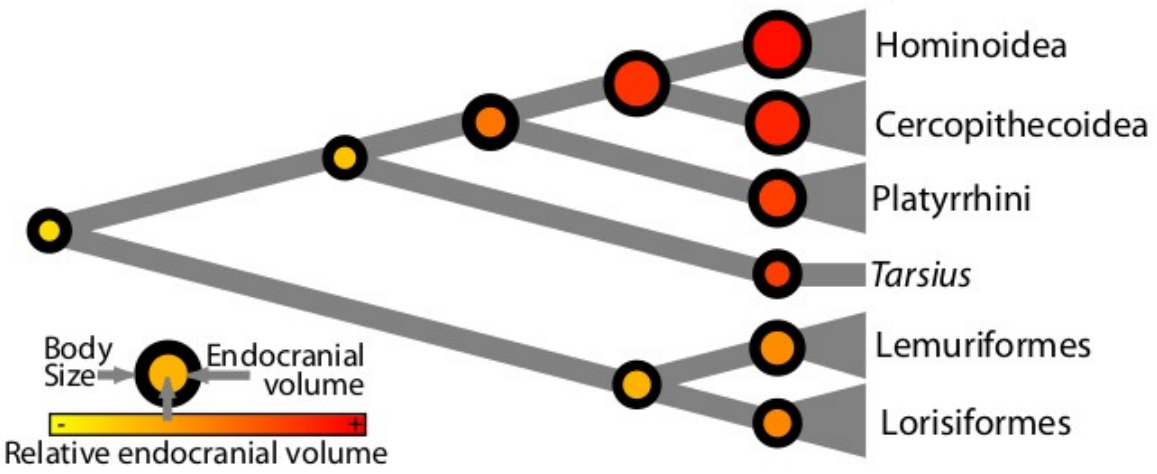


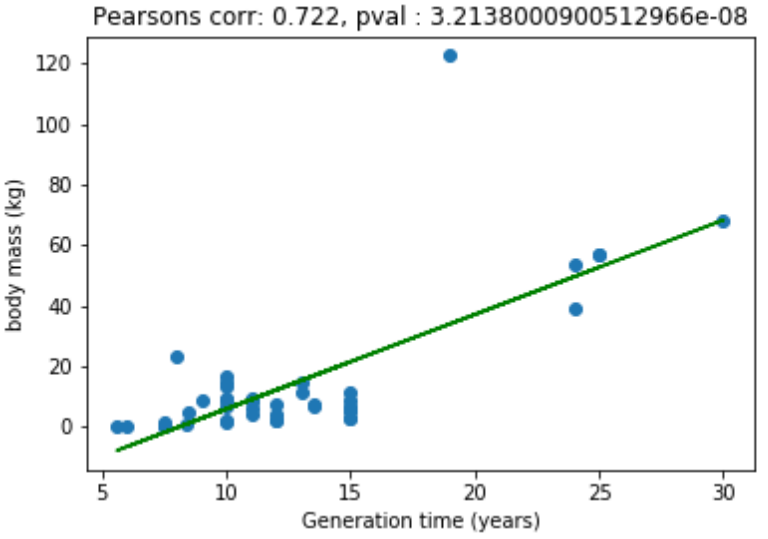
Mutation rates values from pedigree studies in ***Homo sapiens***, ***Pan troglodytes***, ***Gorilla gorilla***, ***Pongo abelii***, ***Papio anubis***, ***Macaca mulatta***, ***Aotus nancymaae***, ***Callithrix jacchus*** *and* ***Chlorocebus aethiops*** (*Microcebus murinus* discarded)

Generation times from the IUCN red list database and body sizes from the databse in Galan-Acedo et al. 2019

What yearly mutation rate to use?

Reconstruction of yearly mutation rate for **extent and ancient** species via the reconstruction of **body size** and **generation time**.

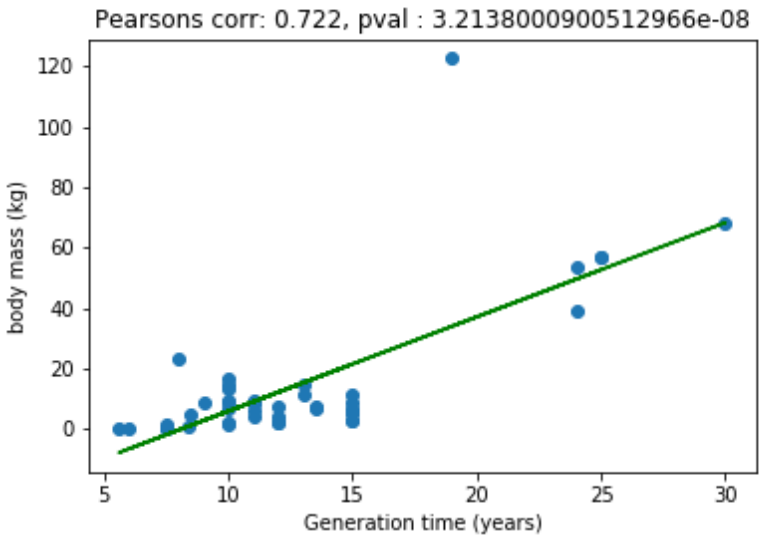


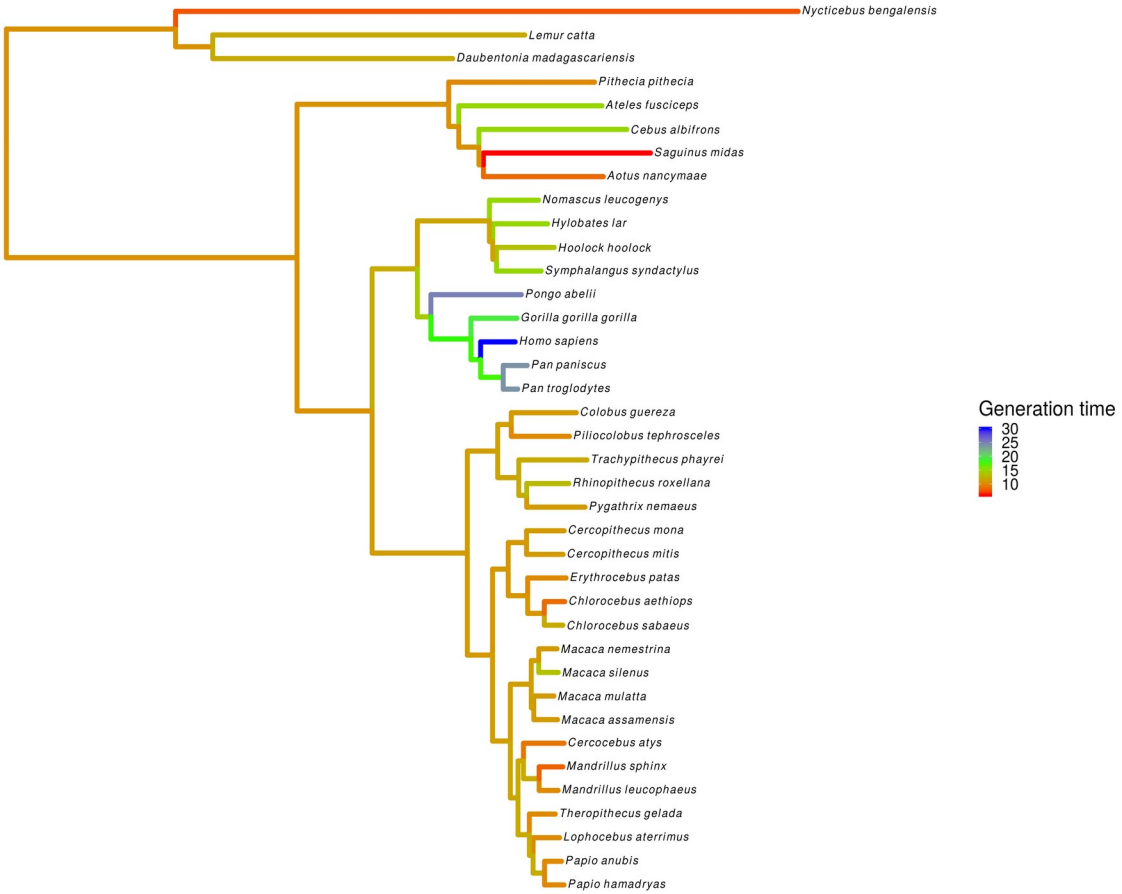


From Steiper and Seiffert 2012

What yearly mutation rate to use?

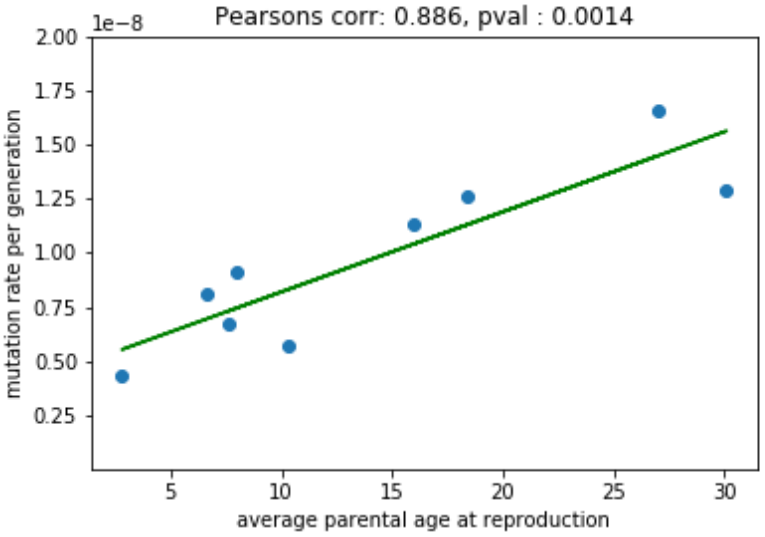
Reconstruction of yearly mutation rate for **extent and ancient** species via the reconstruction of **body size** and **generation time**.

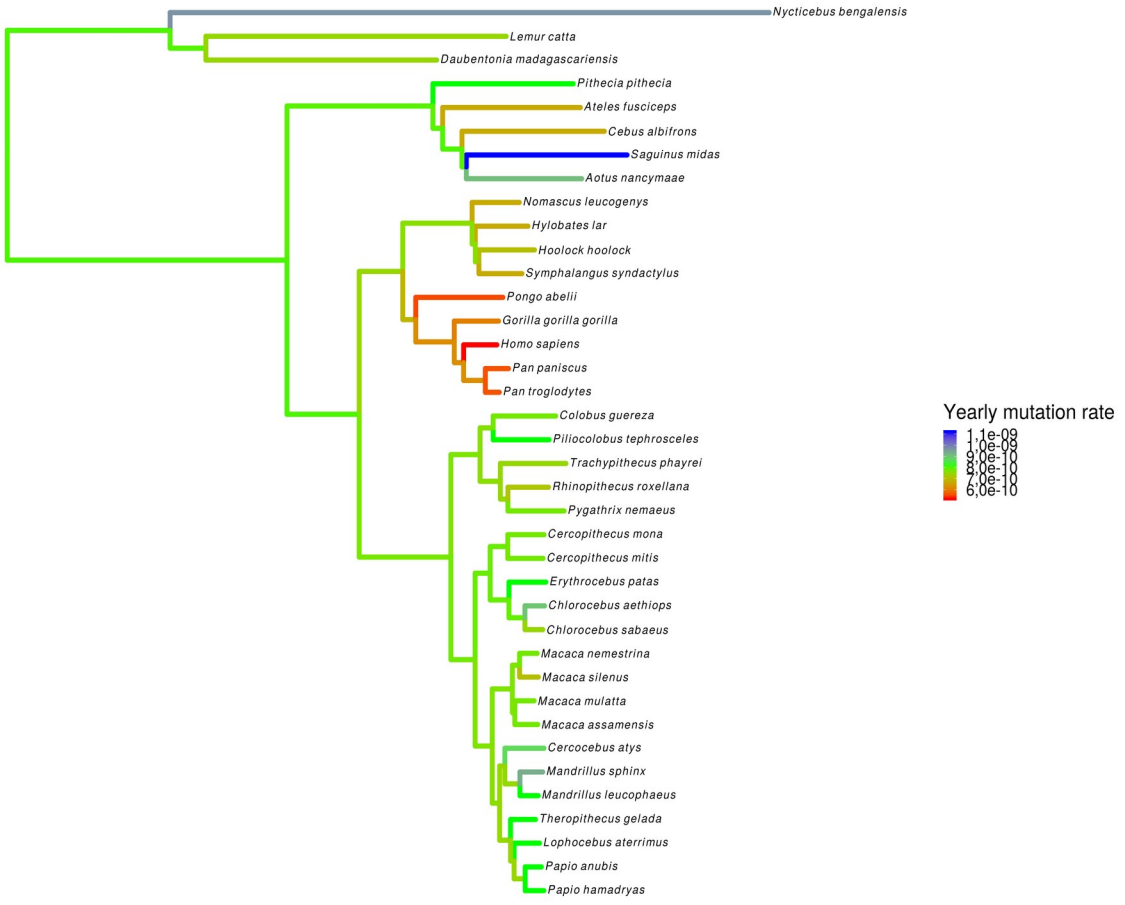




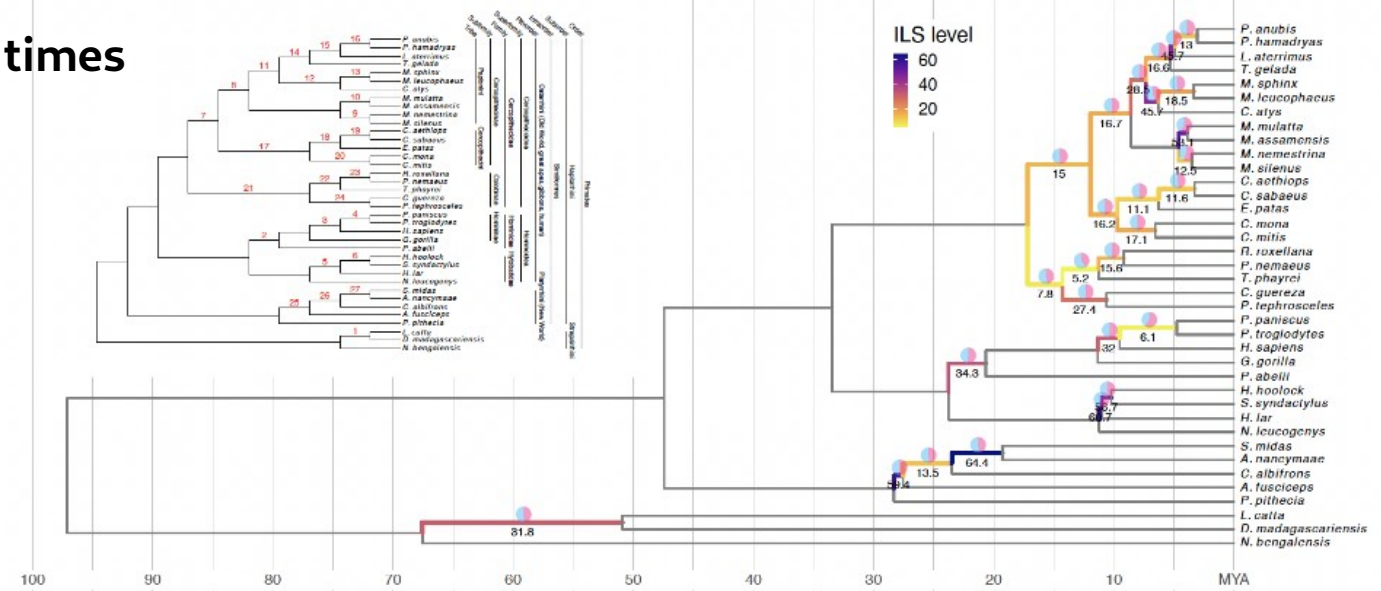
What yearly mutation rate to use?

Reconstruction of yearly mutation rate for **extent and ancient** species via the reconstruction of **body size** and **generation time**.

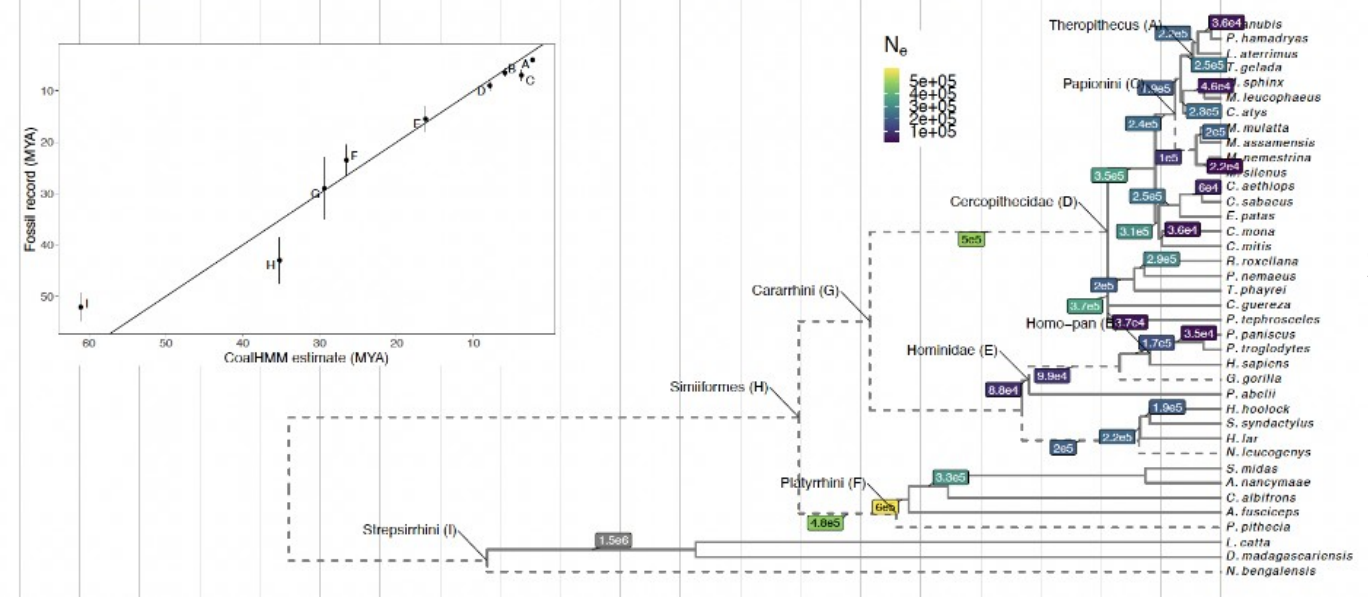




Speciation times

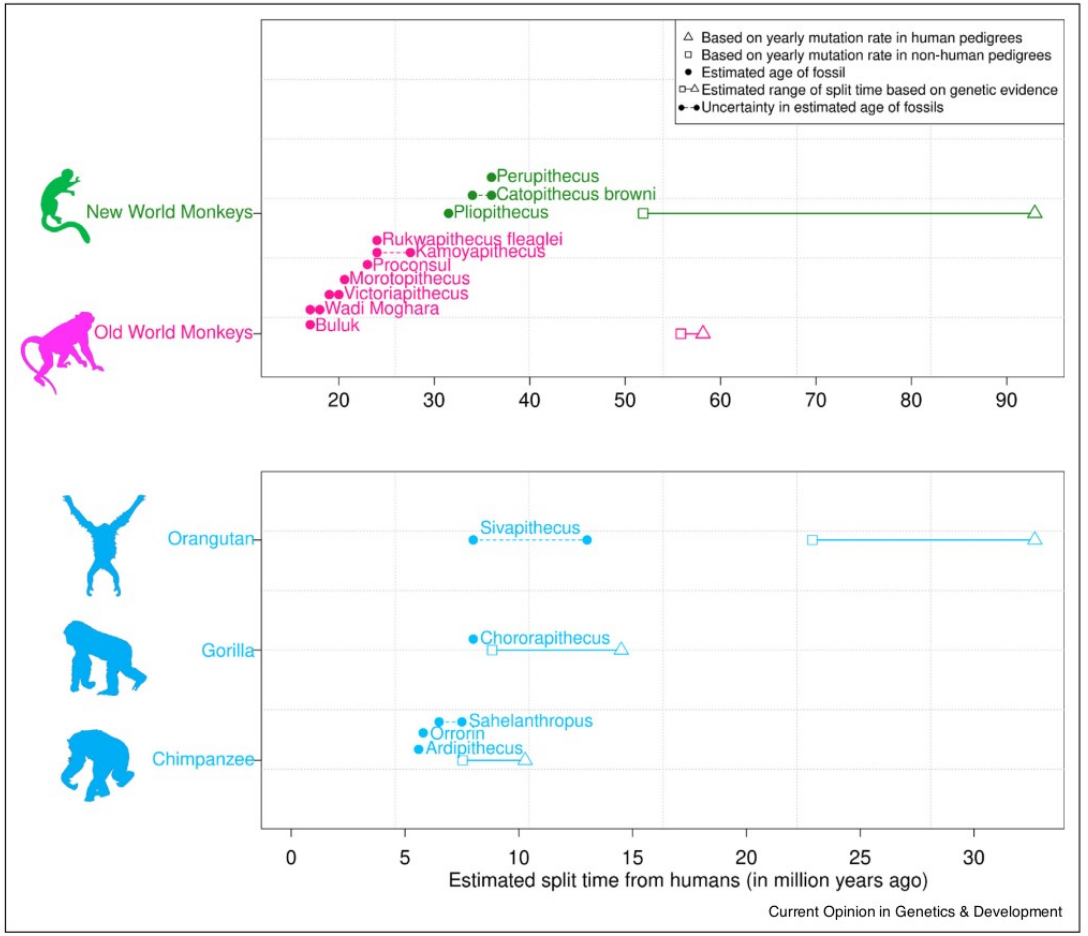


Divergence times



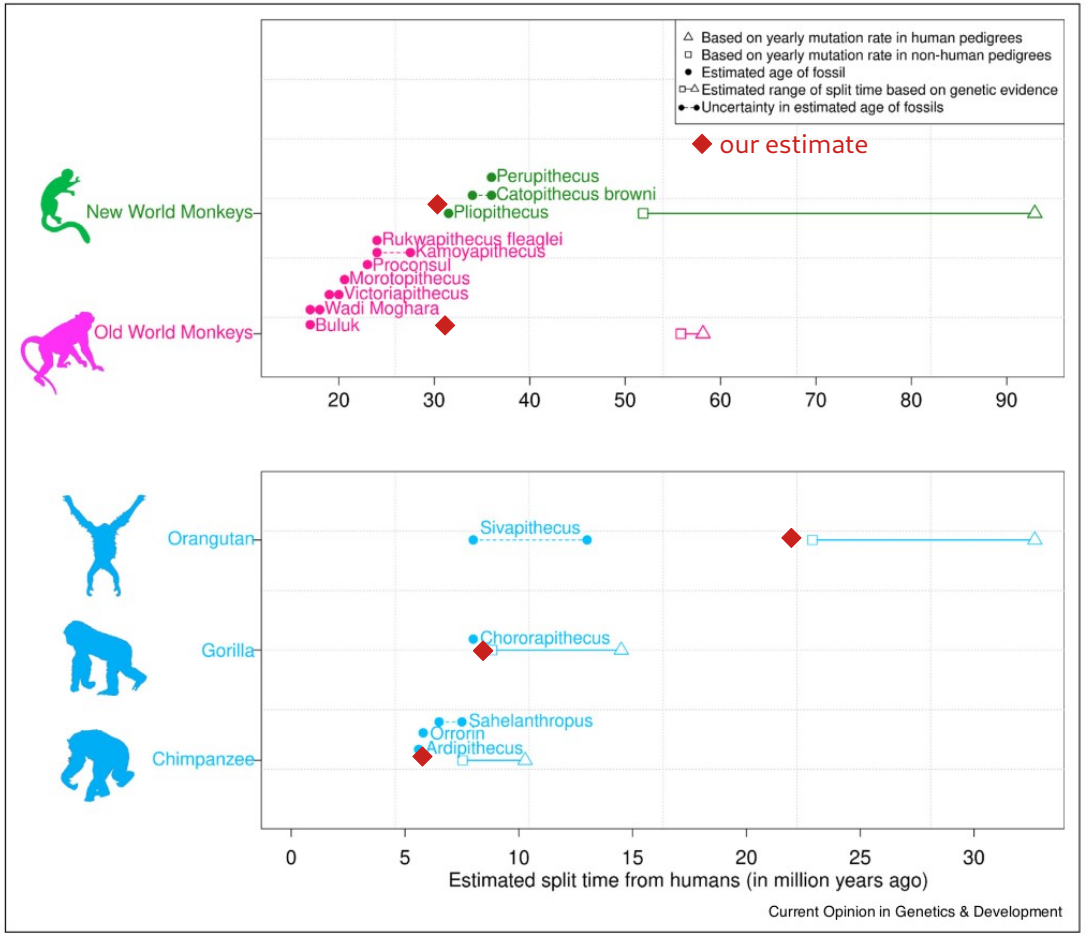
Speciation times

Speciation times & fossil record



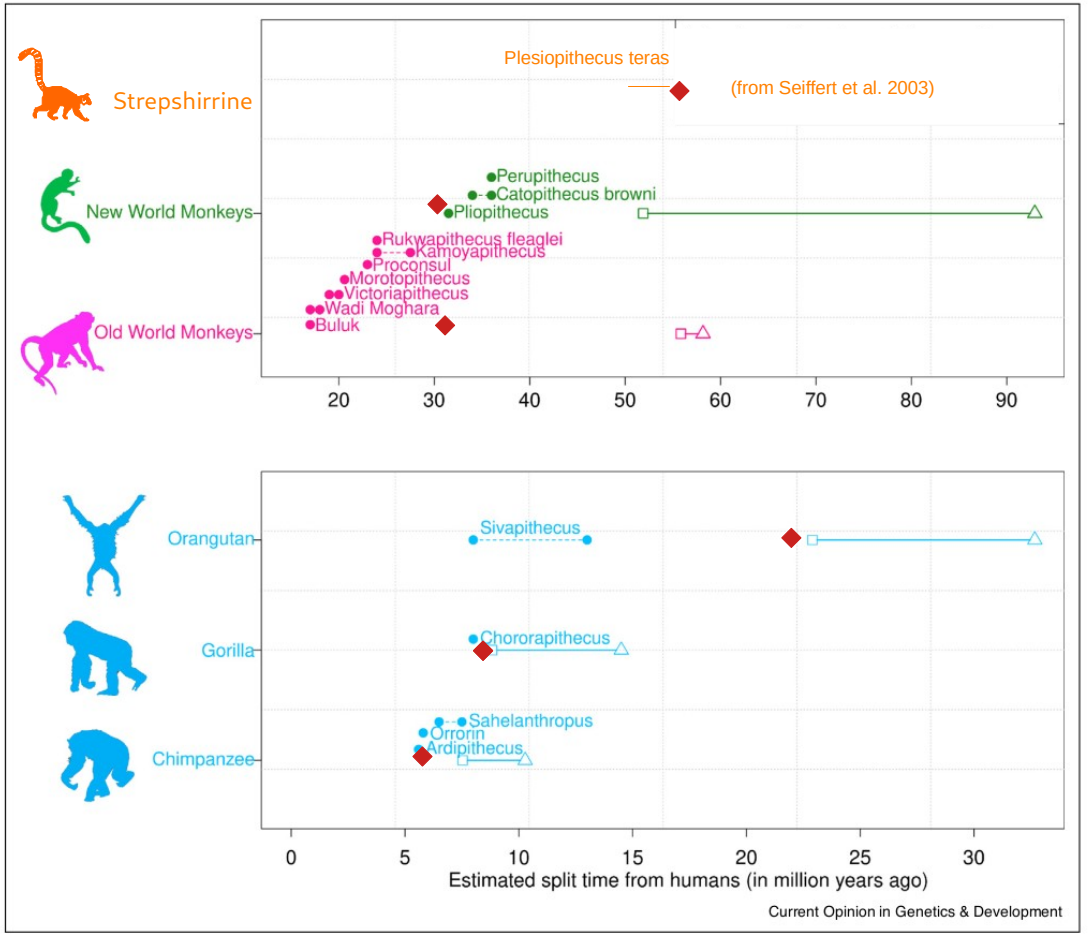
From Chintalapati & Moorjani, 2020

Speciation times & fossil record



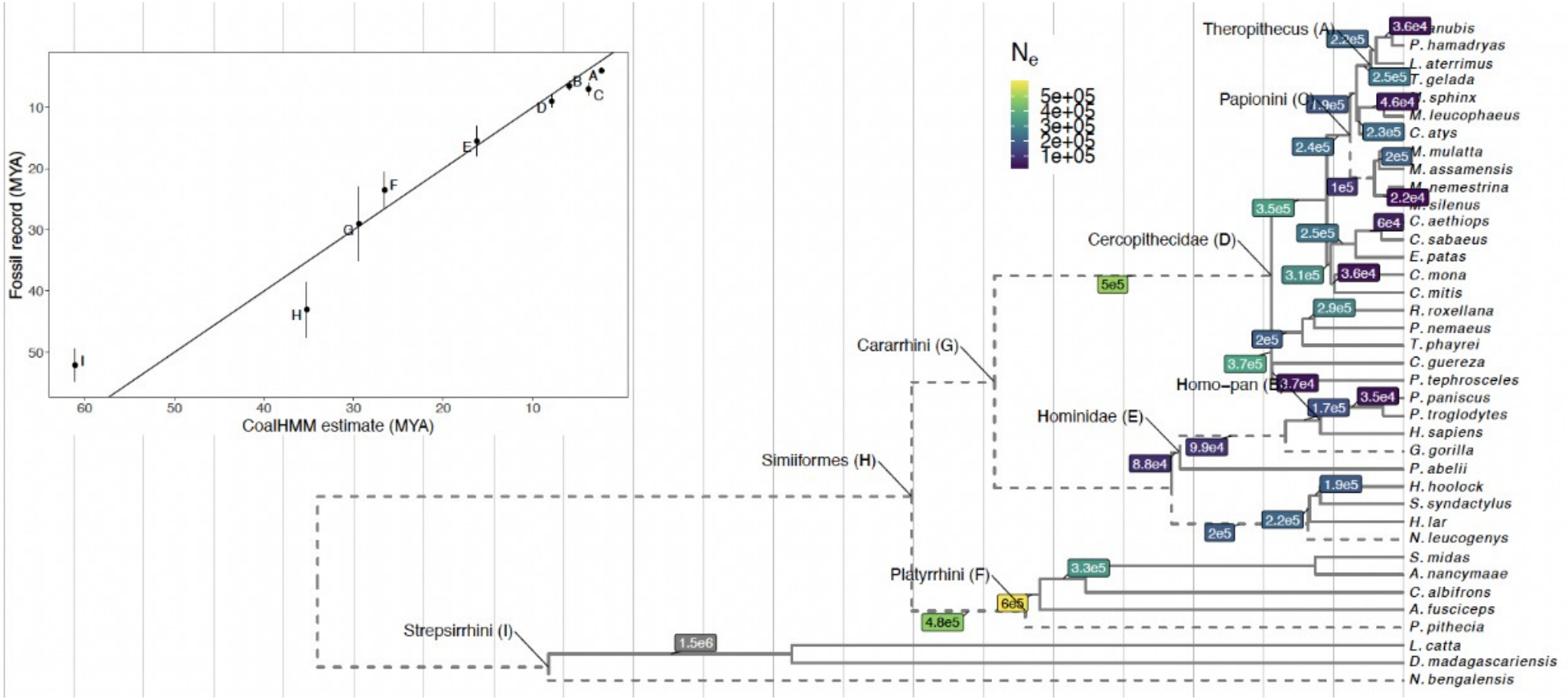
Modified from Chintalapati & Moorjani, 2020

Speciation times & fossil record



Modified from Chintalapati & Moorjani, 2020

Ancestral population sizes



Take home messages :

- High percentage of ILS in short branches of the primate phylogeny

-ILS signal and coalescent theory allows a robust reconstruction of ancestral population parameters.

- **High variation in ancestral population sizes**, up to N of 1.5 millions

- Speciation times globally consistent with the fossil record

Aknowledgements

Iker Rivas-González

Mikkel H. Schierup

Guojie Zhang

And collaborators :
Fang Li
Long Zhou
Josefin Stiller
Dongdong Wu
Kasper Munch
Julien Dutheil

Part 2 : Influence of population size of on the **adaptive substitution rate** in eukaryotes.

→ Mc-Donald & Kreitman like test

Mc-Donald and Kreitman test :

→ What is the proportion of mutations that have been fixed by:
positive selection *vs.* drift ?

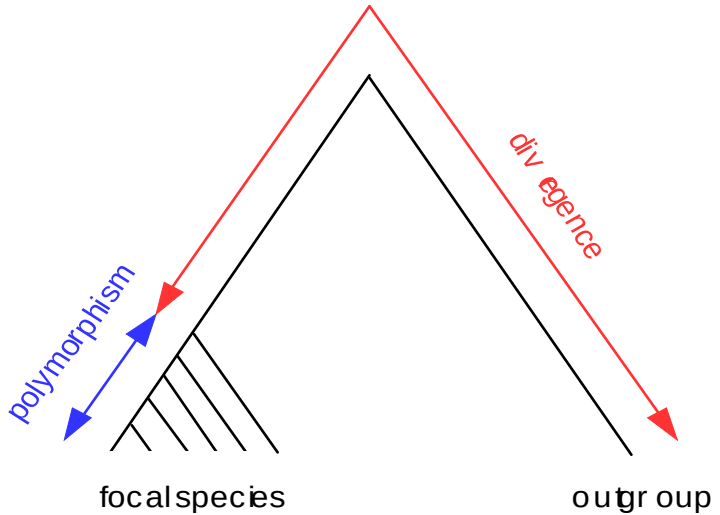
↳ **α = proportion of adaptive substitution** in protein sequences in the lineage of a species

McDonald & Kretiman 1991

The Mc-Donald & Kreitman approach

Comparison of :

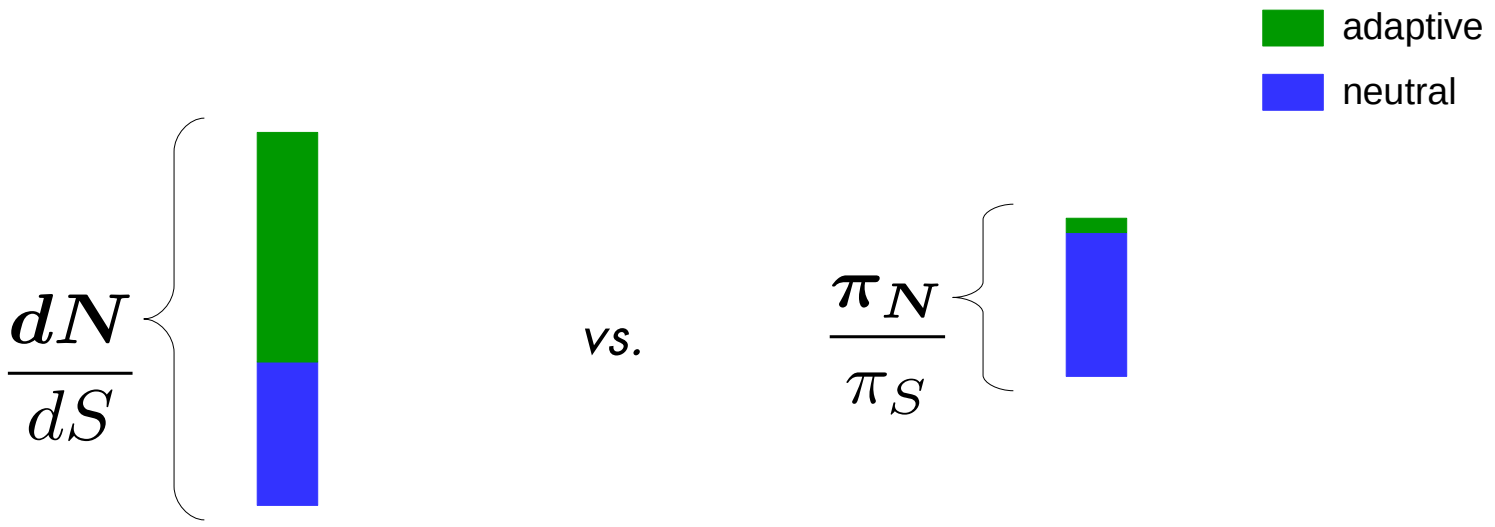
- inter-specific and intra-specific data
  - ↑ divergence
  - ↑ polymorphism



- two types of mutations : synonymous *vs.* non-synonymous
  - ↑ neutral
  - ↑ potentially selected

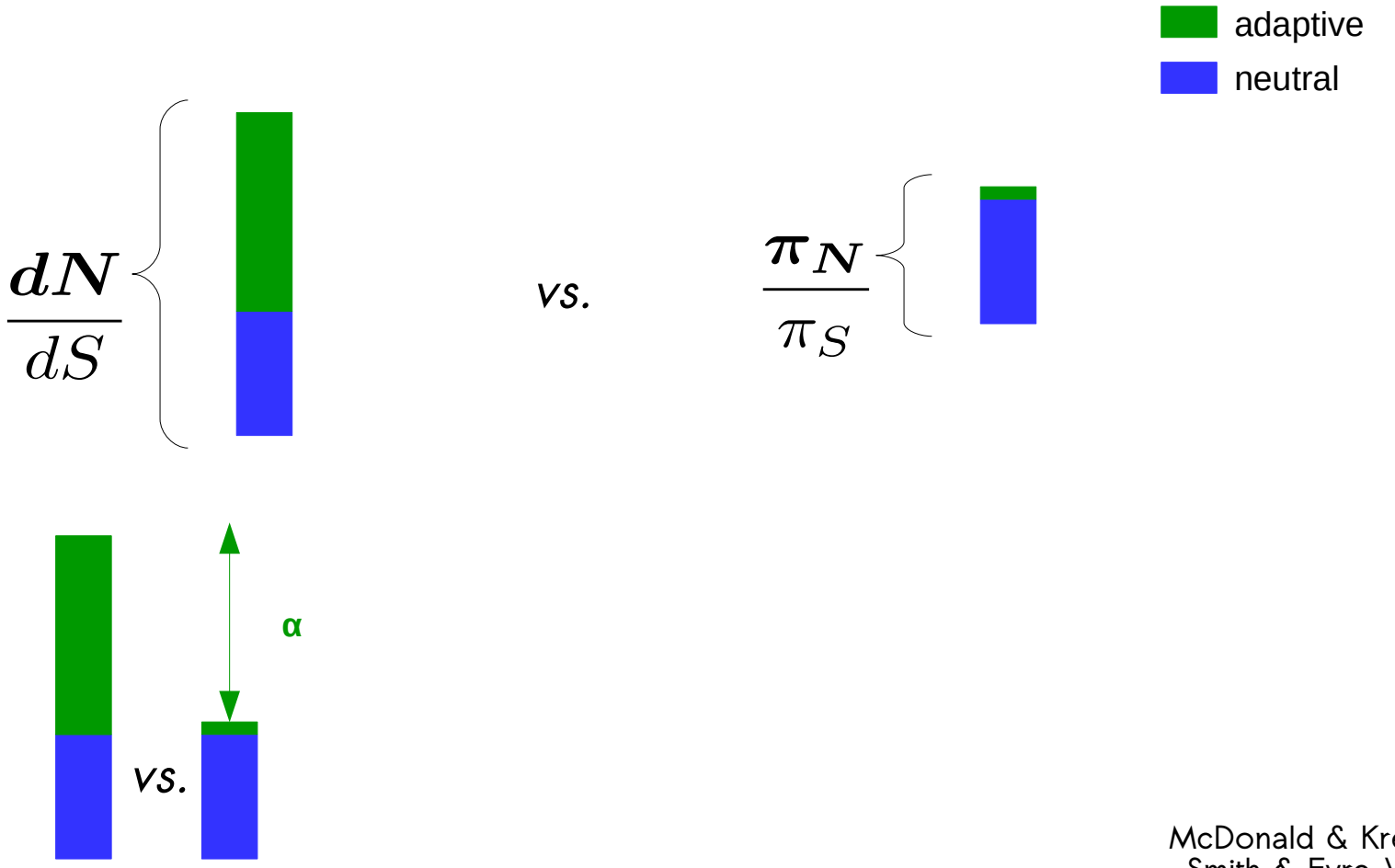
McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

The McDonald & Kreitman approach



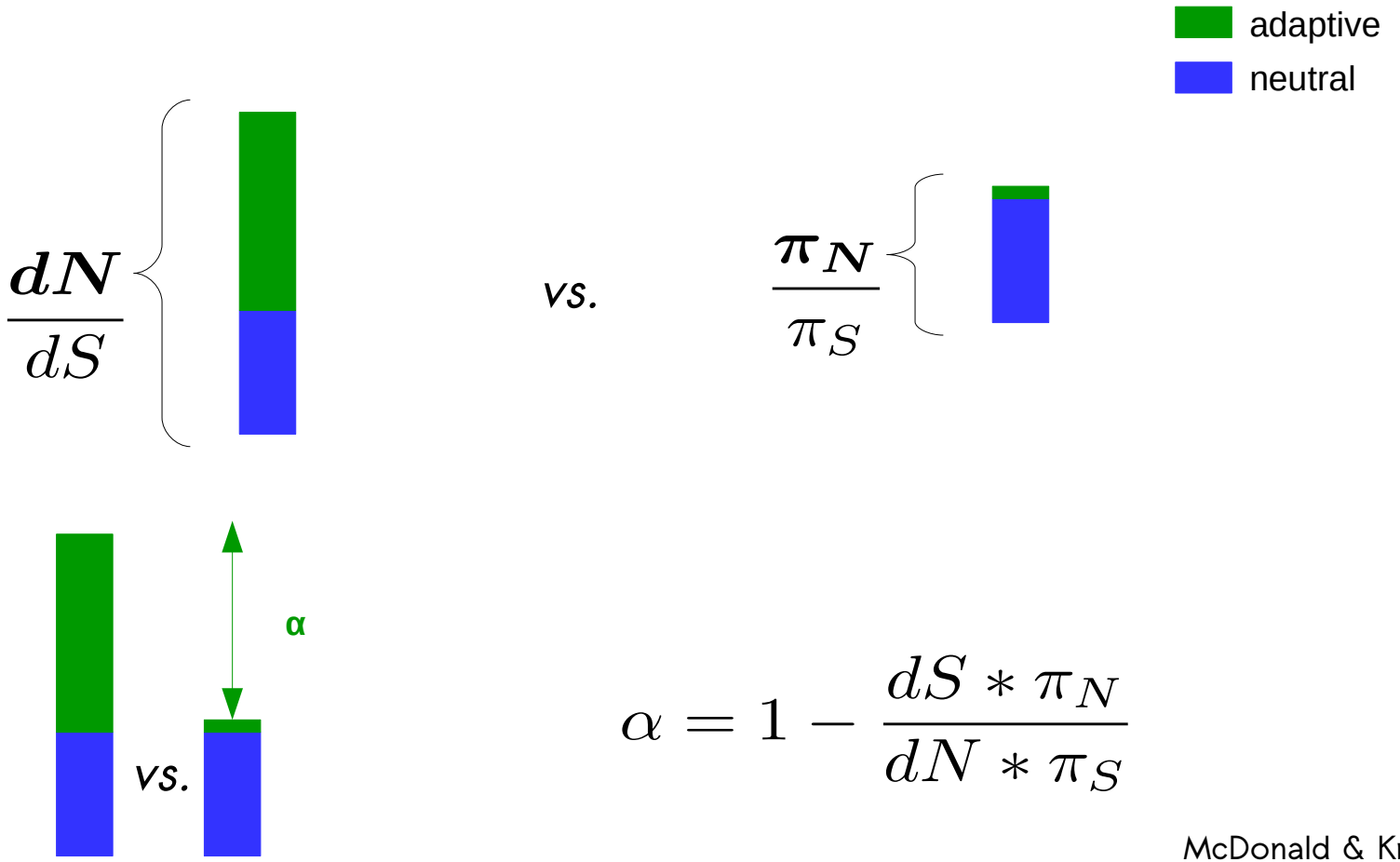
McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

The McDonald & Kreitman approach



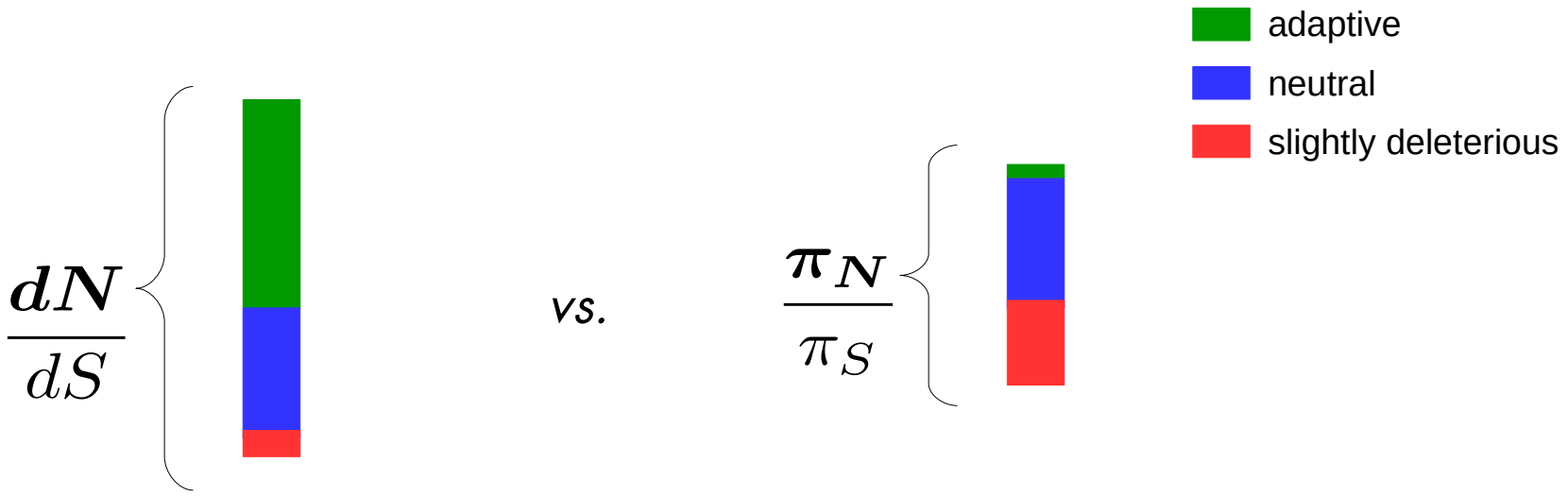
McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

The McDonald & Kreitman approach



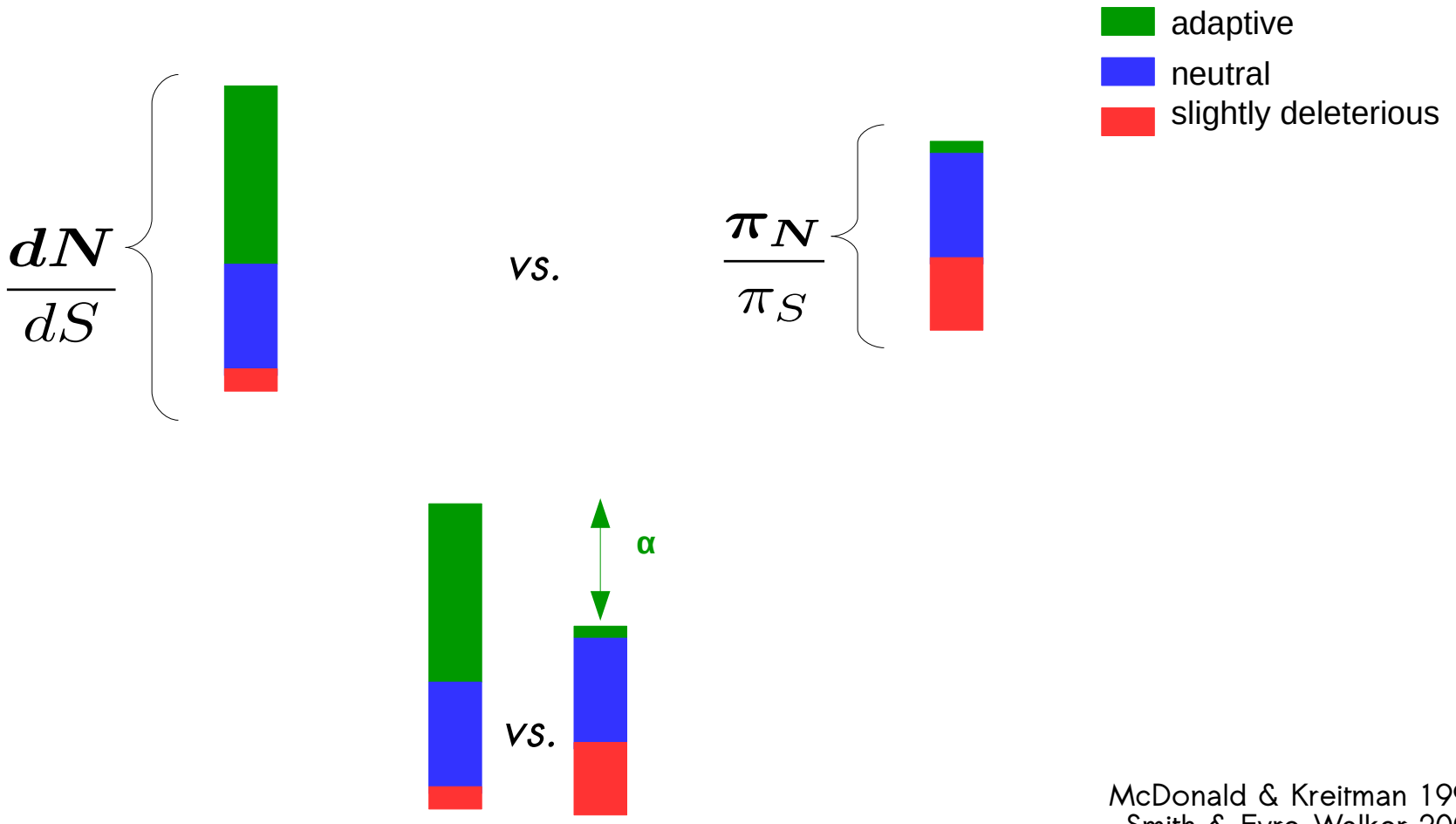
McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

The issues with the McDonald & Kreitman approach



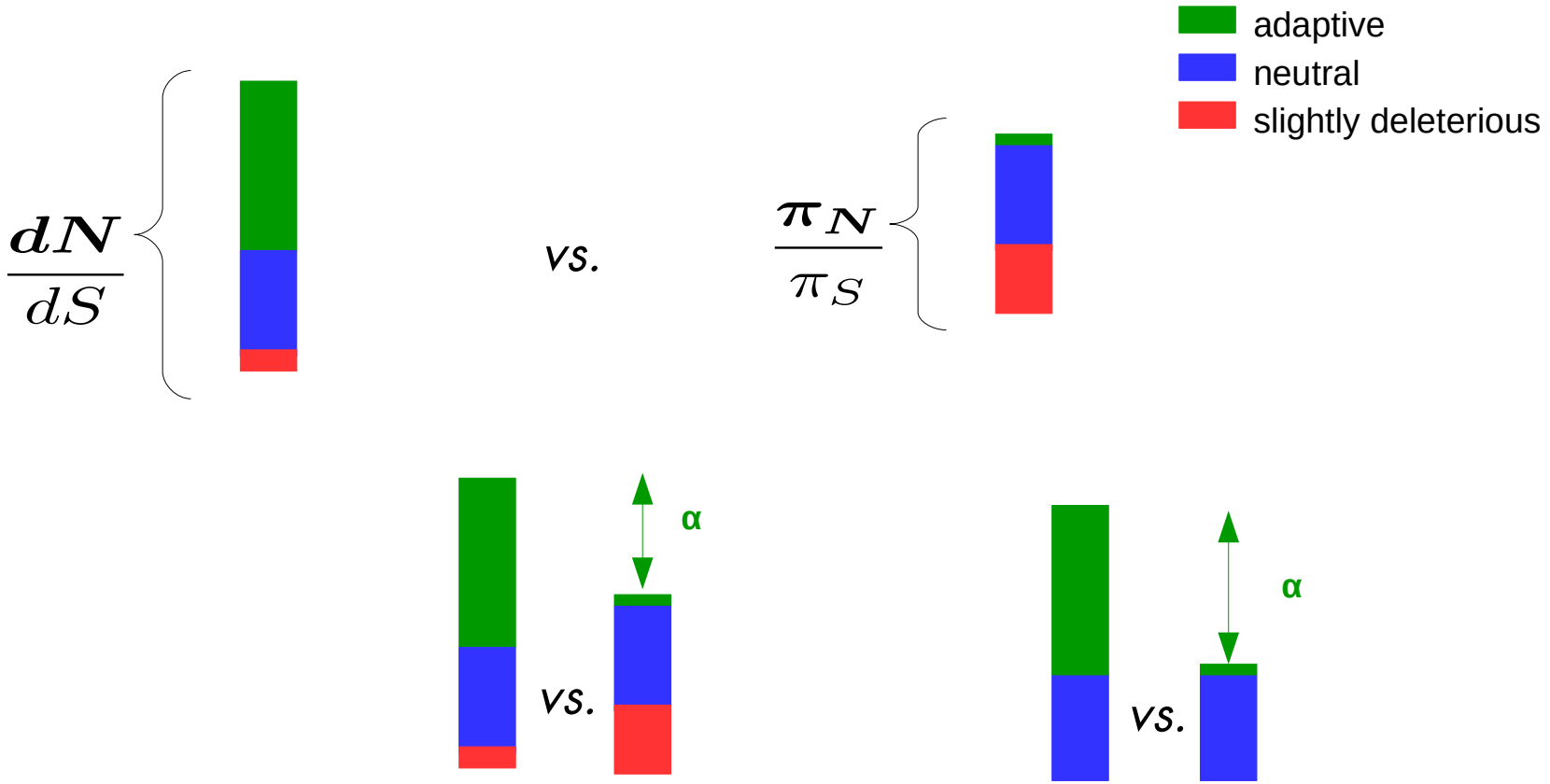
McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

The issues with the McDonald & Kreitman approach



McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

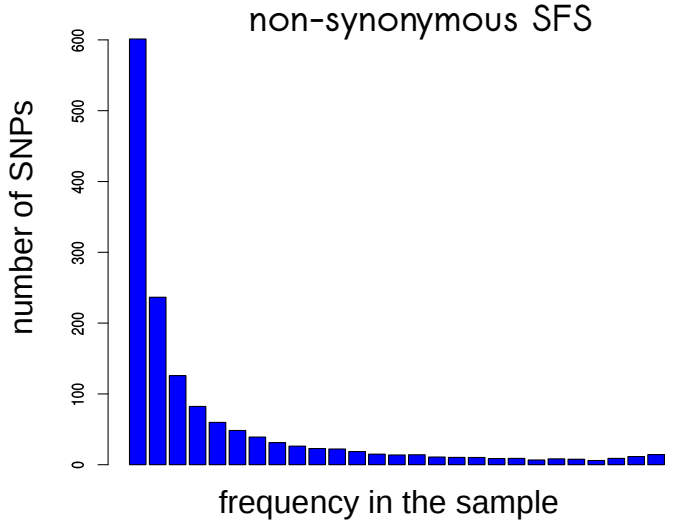
The issues with the McDonald & Kreitman approach

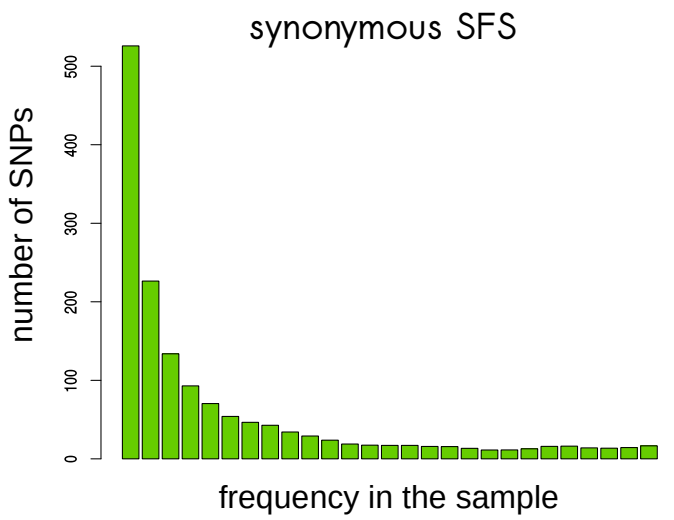


McDonald & Kreitman 1991
Smith & Eyre-Walker 2002

DFE-α approach

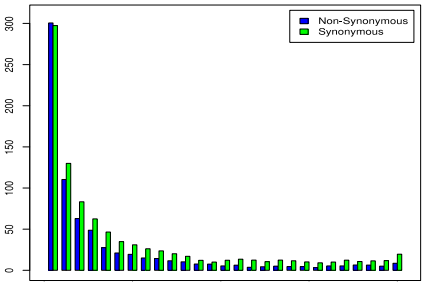
$$\frac{\pi_N}{\pi_S}$$

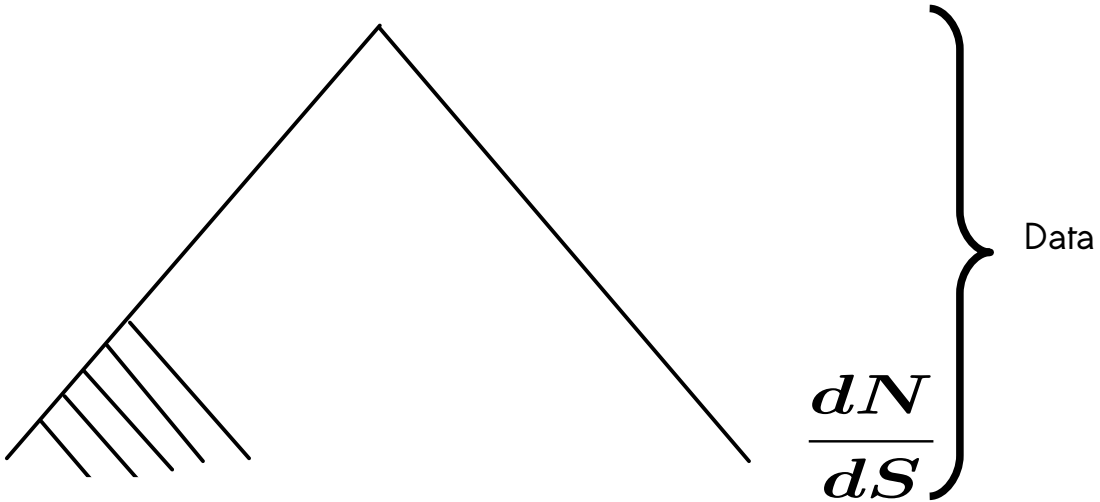




Eyre-Walker et al. 2006

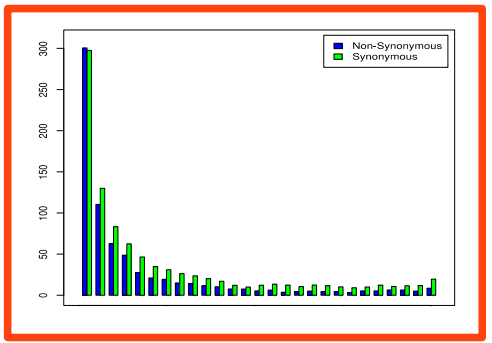
DFE-α approach

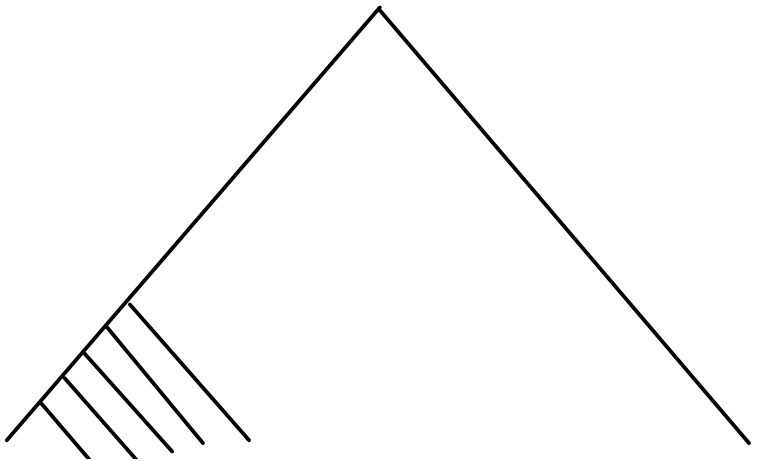




Eyre-Walker et al. 2006
Eyre-Walker & Keightley 2009
Galtier 2016
Tataru et al. 2017

DFE-α approach

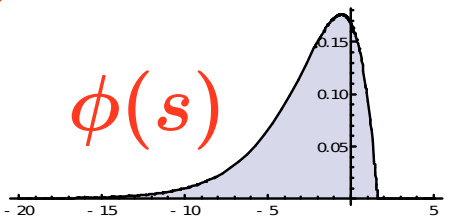




$\frac{dN}{dS}$

Data

$\theta = 4N_e\mu$



$\phi(s)$

$r_1, r_2, r_3, \ldots, r_{2n-1}$

Adjusted parameters (ML)

$\widehat{\omega_{na}}$

$$\frac{dN}{dS} - \widehat{\omega_{na}} = \omega_a$$

$$\alpha = \frac{\omega_a}{\omega_a + \widehat{\omega_{na}}}$$

Eyre-Walker et al. 2006
Eyre-Walker & Keightley 2009
Galtier 2016
Tataru et al. 2017

Results of the McDonald & Kreitman-like approaches

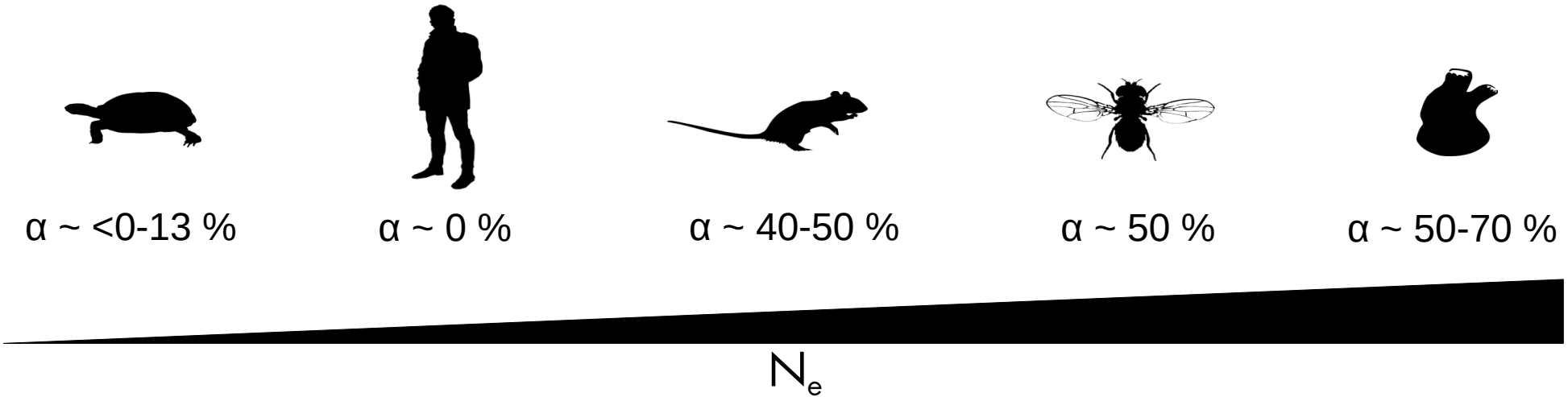
α ~ 0 %

α ~ 50 %

N_e

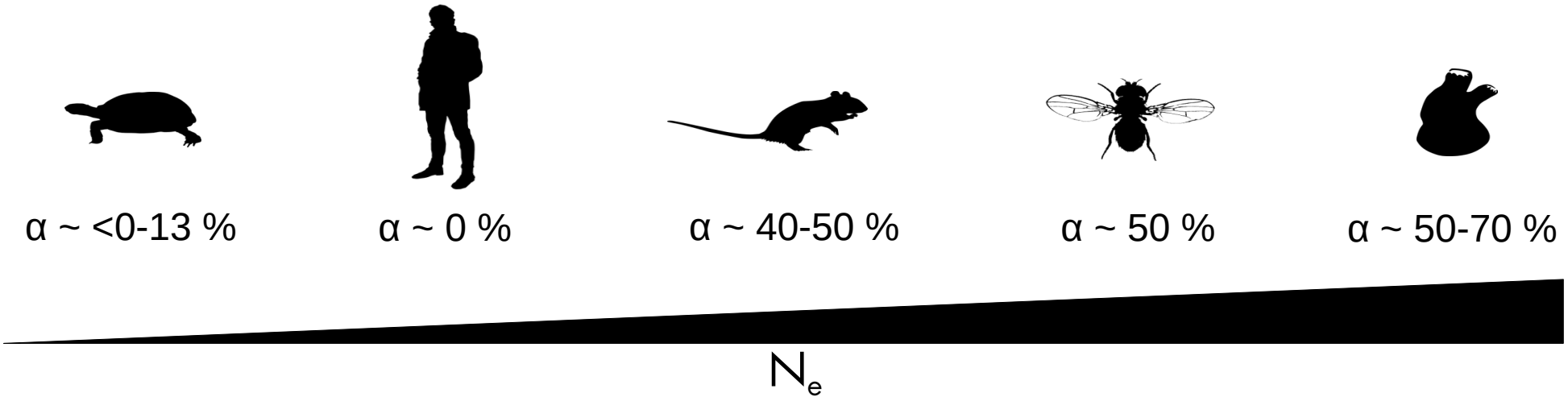
Smith & Eyre-Walker 2002
Zhang & Li 2005
Bustamante et al. 2005

Results of the McDonald & Kreitman-like approaches



Smith & Eyre-Walker 2002
Zhang & Li 2005
Bustamante et al. 2005
Halligan et al. 2010
Tsagkogeorga et al. 2012
Loire et al. 2013

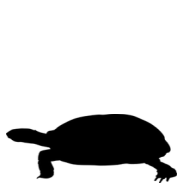
Results of the McDonald & Kreitman-like approaches



1. Large populations → more beneficial mutations
2. Adaptive substitution rate ~ $N_e s \mu_a$

Smith & Eyre-Walker 2002
Zhang & Li 2005
Bustamante et al. 2005
Halligan et al. 2010
Tsagkogeorga et al. 2012
Loire et al. 2013

Results of the McDonald & Kreitman-like approaches


α ~ <0-13 %


α ~ 0 %


α ~ 40-50 %


α < 0 %?


α ~ 50 %

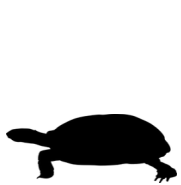

α ~ 50-70 %

N_e ?

$$\alpha = \frac{\omega_a}{\omega_a + \widehat{\omega_{na}}}$$

Charlesworth & Eyre-Walker 2006
Liti et al. 2009
Deinunm et al. 2015

Inconsistent observations






α ~ <0-13 % α ~ 0 % α ~ 40-50 % α < 0 %? α ~ 50 % α ~ 50-70 %

N_e ?

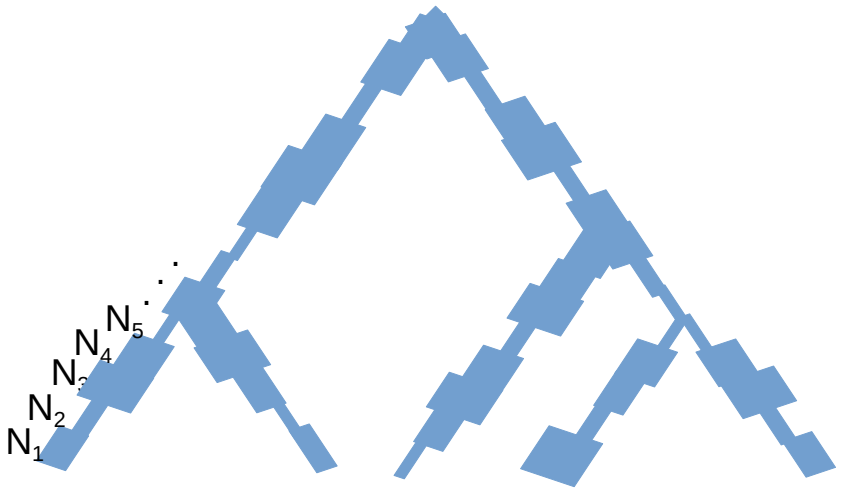
- Positive correlation between ω_a and N_e
  → 13 eukaryotes and six sunflowers
- No correlation between ω_a and N_e
  → 44 eukaryotes and two *Drosophila* species

Strasburg et al. 2010
Jensen & Bachtrog 2011
Gossman et al. 2012
Galtier 2016

Controlling the biases in the DFE-α method

-gBGC : using only GC-conservative mutations
→ drawback: reduces the dataset by 90 %

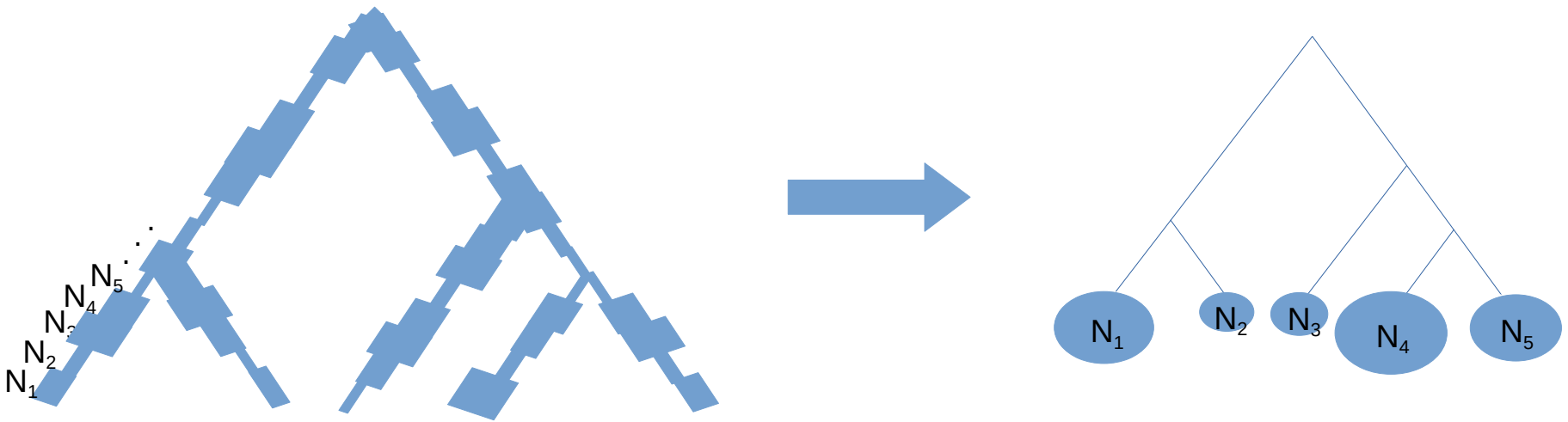
-Long-term demographic fluctuations :
→ long-term fluctuations of the selective/drift regime in the divergence of those species ~ observed variation of the recent selective/drift regime between closely related species



Controlling the biases in the DFE-α method

-gBGC : using only GC-conservative mutations
→ drawback: reduces the dataset by 90 %

-Long-term demographic fluctuations :
→ long-term fluctuations of the selective/drift regime in the divergence of those species ~ observed variation of the recent selective/drift regime between closely related species



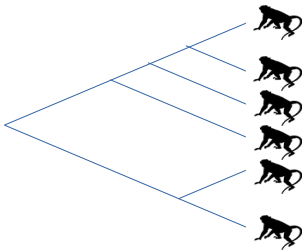
Controlling the biases in the DFE-α method : dataset

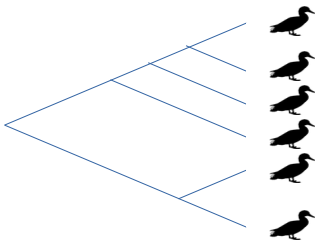

Marie-Ka Tilak

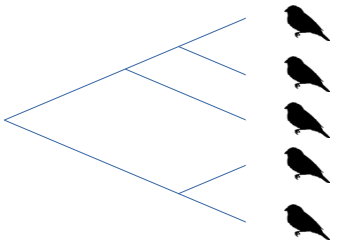

Émeric Figuet

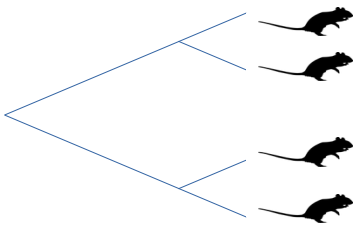

Paul Simion

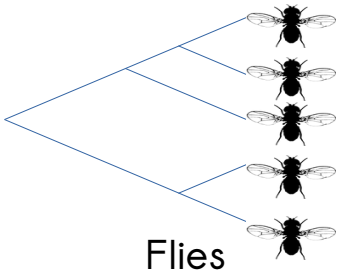
Existing datasets :


Primates
(Catarrhinae)

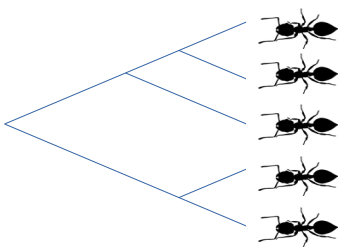

Fowls
(Galloanserae)

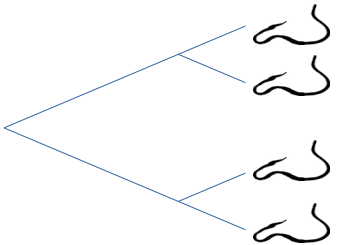

Passerines
(Passeriformes)

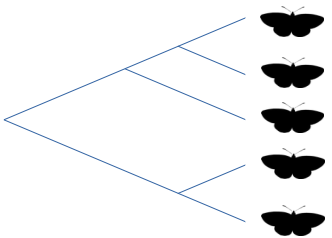

Muroids
(Muroidea)

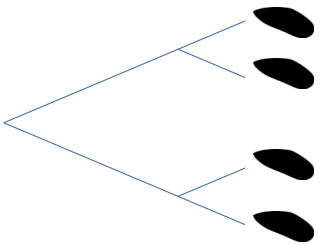

Flies
(Drosophila)

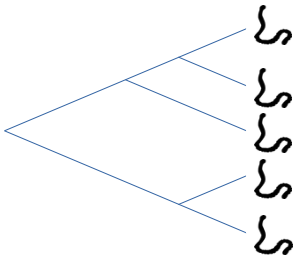
Newly generated datasets via exon capture :


Ants
(*Formica*)

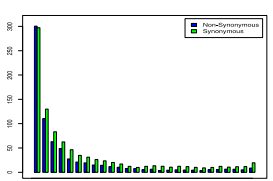

Earth worms
(Lumbricinae)

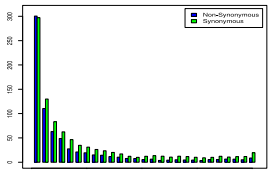

Butterflies
(Satyrini)

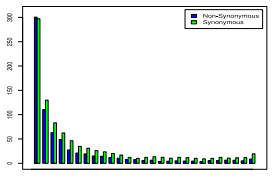

Mussels
(*Mytilus*)

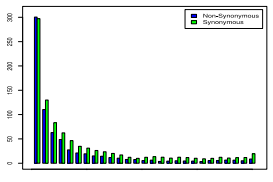

Ribbon worms
(*Lineus*)

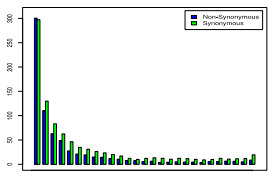
Controlling the biases in the DFE-α method : 1st strategy



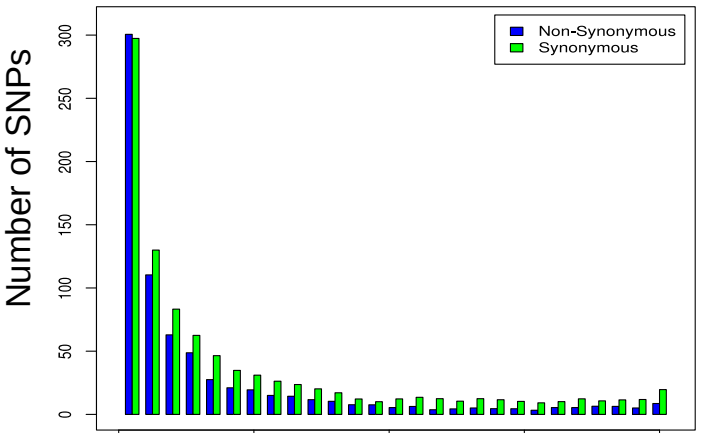






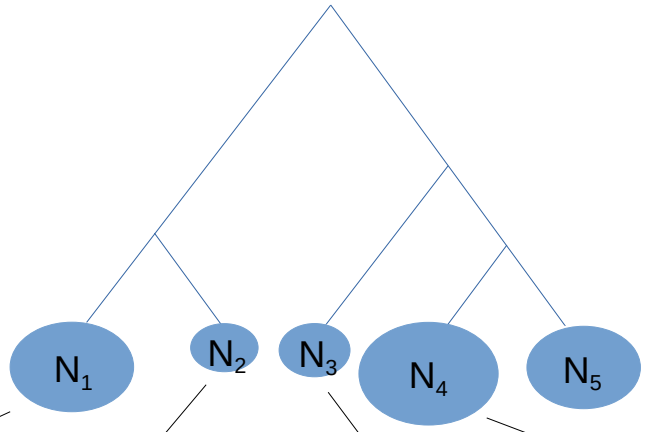


Pooled SFS



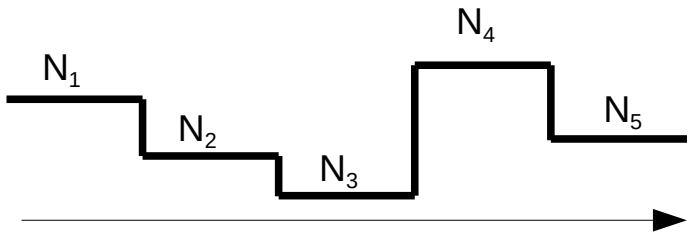
+ total $\frac{dN}{dS}$ → classic DFE-α approach

Controlling the biases in the DFE-α method : 2nd strategy

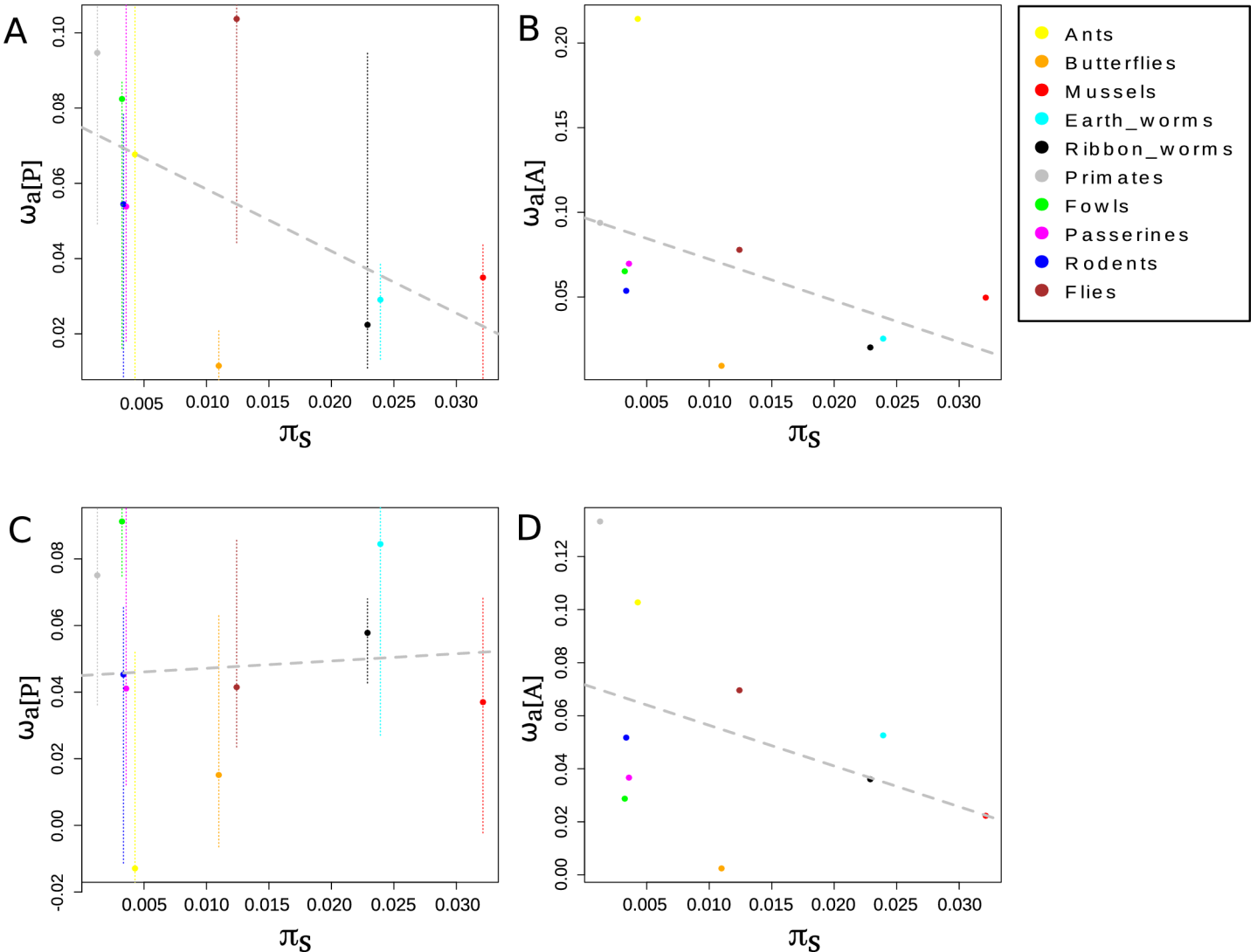


$$\overline{\omega_{na}} = \frac{1}{5}\omega_{na}[N_1] + \frac{1}{5}\omega_{na}[N_2] + \frac{1}{5}\omega_{na}[N_3] + \frac{1}{5}\omega_{na}[N_4] + \frac{1}{5}\omega_{na}[N_5]$$

$$\overline{\omega_a} = \frac{dN}{dS} - \overline{\omega_{na}}$$



Results :


A
B
C
D
ωa[P]
ωa[A]
πS
Ants
Butterflies
Mussels
Earth_worms
Ribbon_worms
Primates
Fowls
Passerines
Rodents
Flies

A negative relationship between ω_a and N_e ?

Theory says that there should be a positive relationship between ω_a and N_e because :

1. Large populations → more beneficial mutations
2. Adaptive substitution rate ~$N_e s \mu_a$

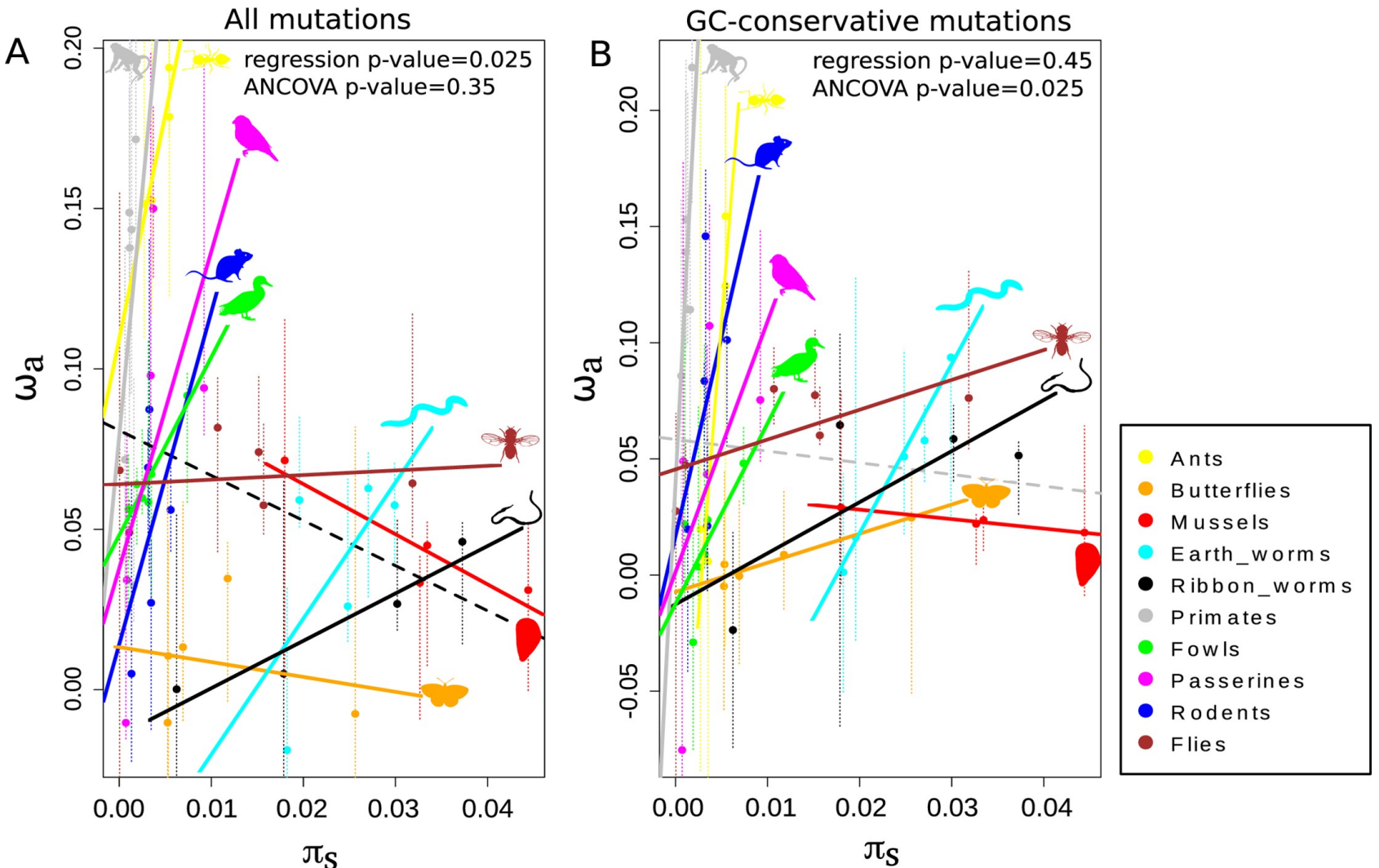
Only valid if

→ the input of adaptive mutation is limited

→ the DFE is independant on Ne

Kimura 1983
Gossman et al. 2012

Is adaptation limited by mutation?



A negative relationship between ω_a and N_e ?

Theory says that there should be a positive relationship between ω_a and N_e because :

1. Large populations → more beneficial mutations
2. Adaptive substitution rate ~$N_e s \mu_a$

Only valid if

→ the input of adaptive mutation is limited → **we show that this is not always true and it depends on Ne**

→ the DFE is independant on Ne

Kimura 1983
Gossman et al. 2012

A negative relationship between ω_a and N_e ?

Theory says that there should be a positive relationship between ω_a and N_e because :

1. Large populations → more beneficial mutations
2. Adaptive substitution rate ~$N_e s \mu_a$

Only valid if

→ the input of adaptive mutation is limited → **we show that this is not always true and it depends on Ne**

→ the DFE is independant on Ne → **Fisher's geometrical model arguments**

Kimura 1983
Gossman et al. 2012

Fisher's geometrical model predictions

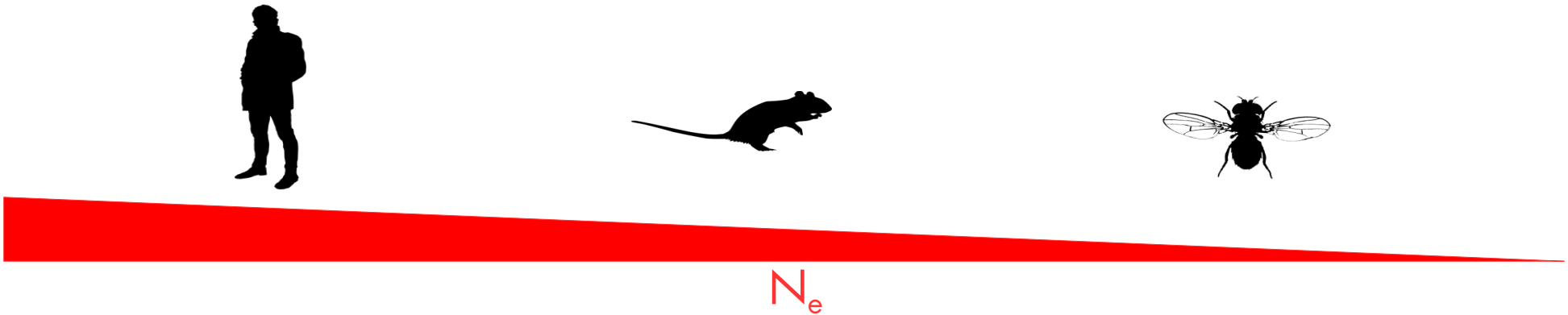
Low population size

1. Maladaptation
↓
larger distance to the fitness optimum
↓
larger proportion of adaptive mutations

2. High longevity
↓
longer generation time
↓
higher rate of environmental change
↓
higher ω_a

Orr 2000
Fernandez & Lynch 2011
Lourenço et al. 2013

Take home messages :



-When comparing distantly related species with different life history traits and different distance to their fitness optimum, we do not expect ω_a to correlate positively with Ne

-Adaptation is limited by the supply of new mutations (Ne.μ) only in low Ne taxa

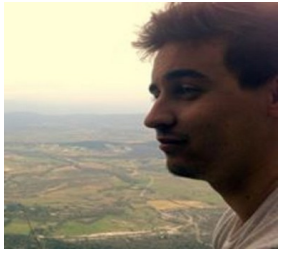
My thesis directors


My collaborators





My internship students




POPULATION SIZE, INCOMPLETE LINEAGE SORTING AND SELECTION IN ANIMAL GENOMES

Marjolaine Rousselle

marjolaine.rousselle@inrae.fr
@MarjoRousselle

Merci de votre attention !

CBGP
Centre de Biologie pour la Gestion des Populations

Séminaire du CBGP-19/10/2021