Population-genetics based modelling of phylogenetic codon models

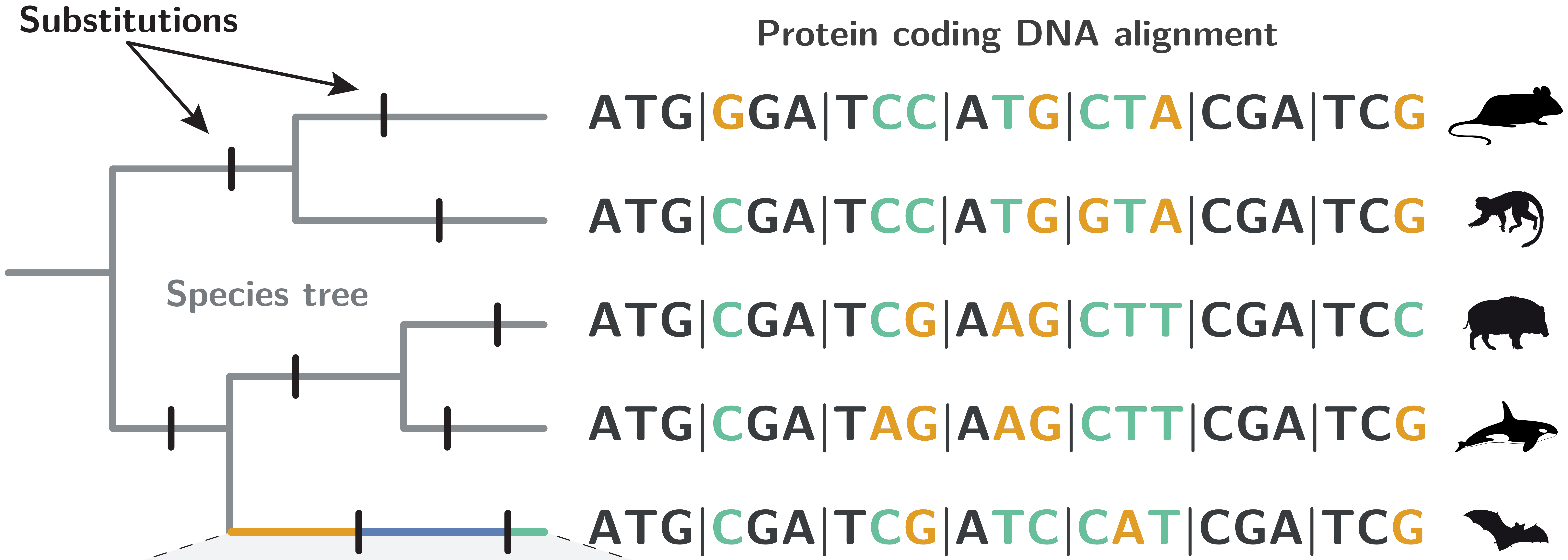
June 24, 2021

Thibault Latrille - teaching assistant at LBBE
Nicolas Lartillot - CNRS research director at LBBE

Lab meeting

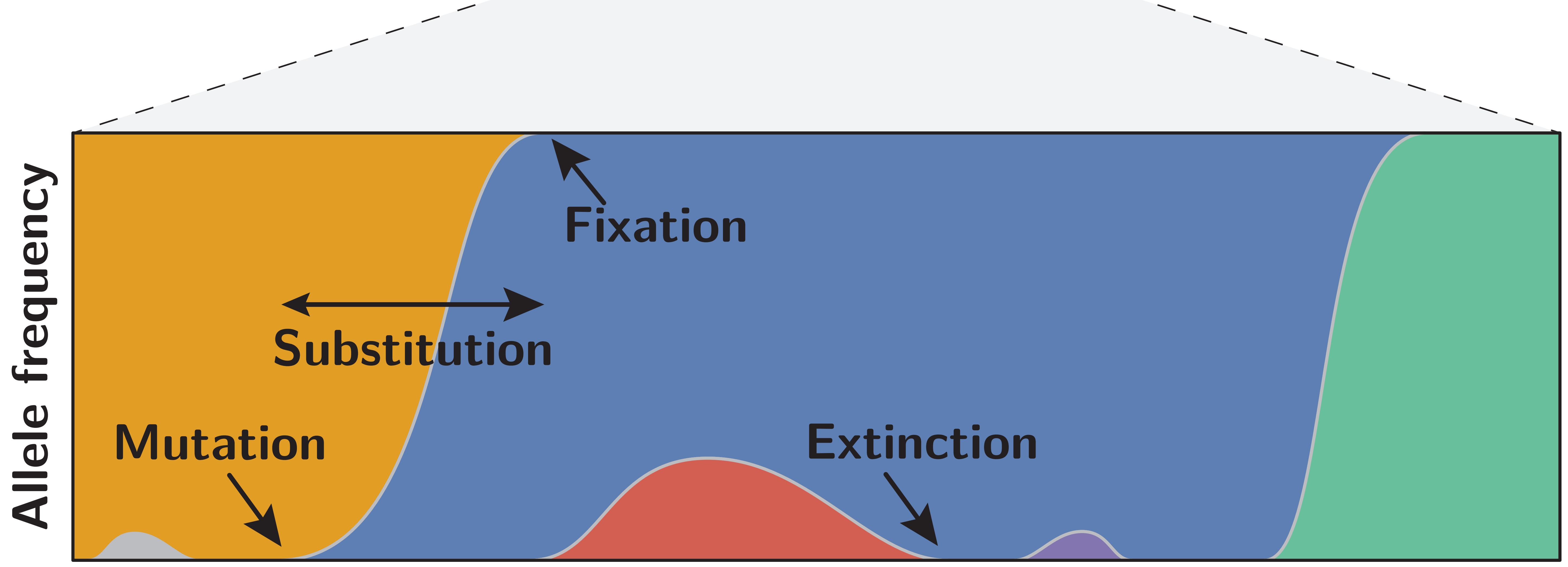
LBBE LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE
UNIVERSITÉ DE LYON
cnrs
ENS DE LYON
thibault.latrille@ens-lyon.org

History of substitutions along the species tree



Protein coding DNA alignment

ATG|GGA|TCC|ATG|CTA|CGA|TCG
ATG|CGA|TCC|ATG|GTA|CGA|TCG
ATG|CGA|TCG|AAG|CTT|CGA|TCC
ATG|CGA|TAG|AAG|CTT|CGA|TCG
ATG|CGA|TCG|ATC|CAT|CGA|TCG



- A substitution is a mutation that reached fixation in the population.
- If alleles are neutral (no selection), the substitution rate is equal to the underlying mutation rate.
- For alleles under selection, what determines their substitution rate?

Felsenstein (1981); Kimura (1983); Ohta (1992).

The effect of selection and genetic drift

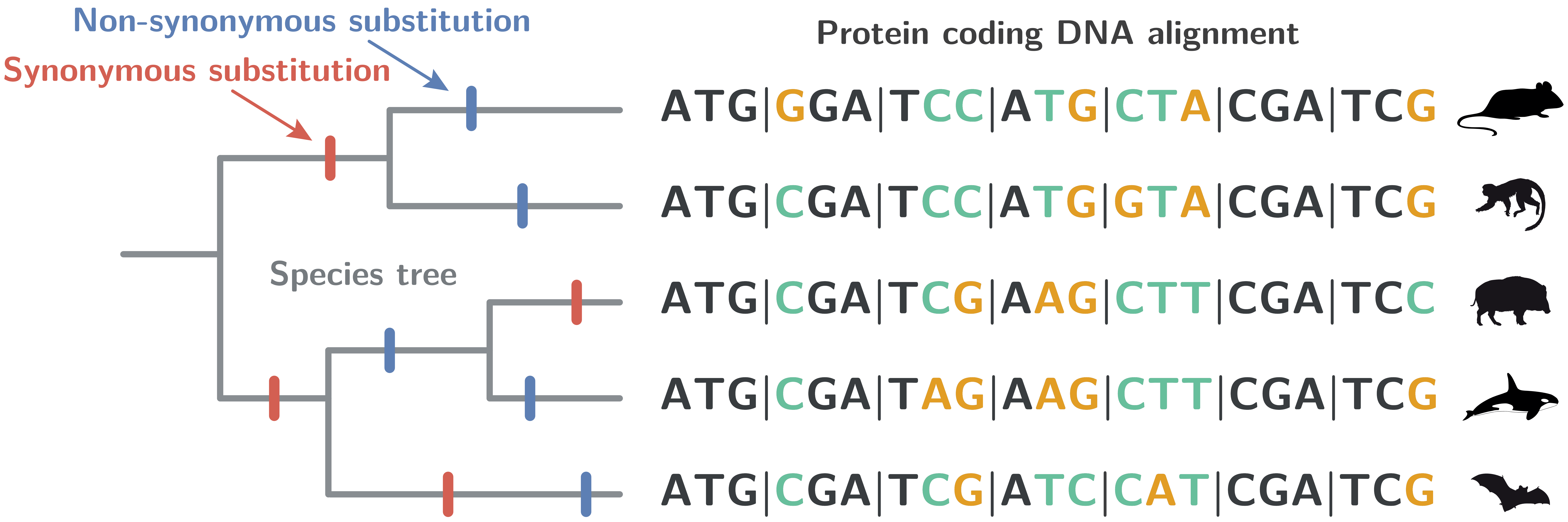
Stronger negative selection coefficient

Increased population size (N_e)

- **Stronger negative selection coefficient results in a decrease of the fixation probability.**
- **Effective population size (N_e) acts as a magnifier of selection.**

https://github.com/ThibaultLatrille/WrightFisher

Codon models take advantage of the genetic code



Non-synonymous substitution

Synonymous substitution

Species tree

Protein coding DNA alignment

ATG|GGA|TCC|ATG|CTA|CGA|TCG

ATG|CGA|TCC|ATG|GTA|CGA|TCG

ATG|CGA|TCG|AAG|CTT|CGA|TCC

ATG|CGA|TAG|AAG|CTT|CGA|TCG

ATG|CGA|TCG|ATC|CAT|CGA|TCG

- Non-synonymous substitutions are reflecting the effect of mutation, selection and drift.

- Synonymous substitutions are considered selectively neutral, reflecting the mutational processes.

- Contrasting non-synonymous and synonymous substitution rates allows to estimate the strength of selection exercised on proteins.

King & Jukes (1969); Kimura (1983); Goldman & Yang (1994); Muse & Gaut (1994).

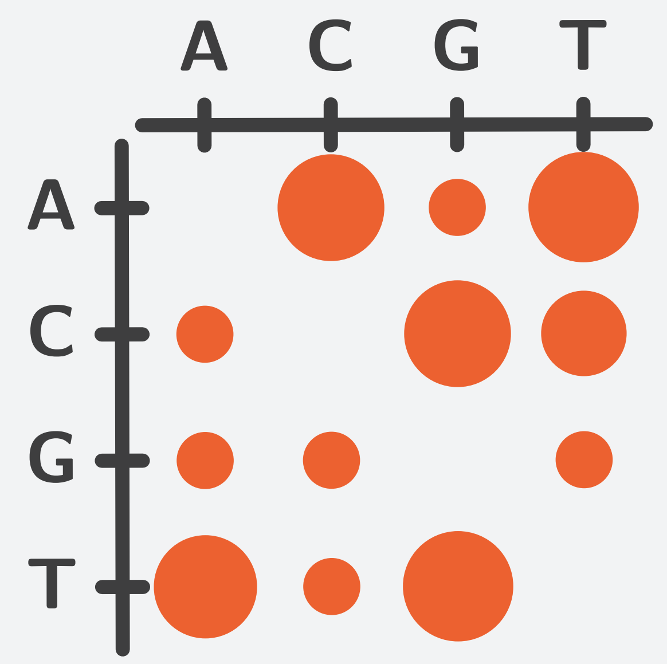
ω-based phylogenetic codon models

- $Q_{i,j}$ is the substitution rate from codon i to j.

$$\begin{cases} Q_{i,j} = \mu_{i,j} & \text{if codons } i \text{ and } j \text{ are synonymous} \\ Q_{i,j} = \omega\, \mu_{i,j} & \text{if codon } i \text{ and } j \text{ are non-synonymous.} \end{cases}$$

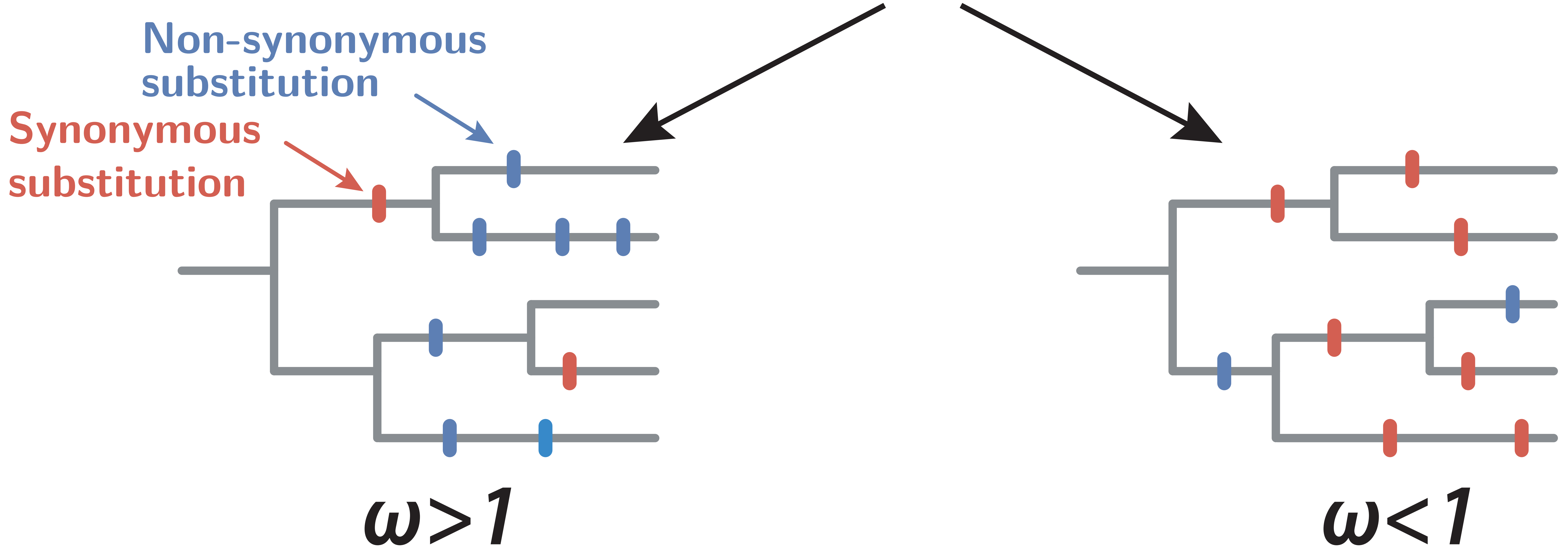
ω

Scaling factor exerced on non-synonymous mutations



$\mu_{i,j}$

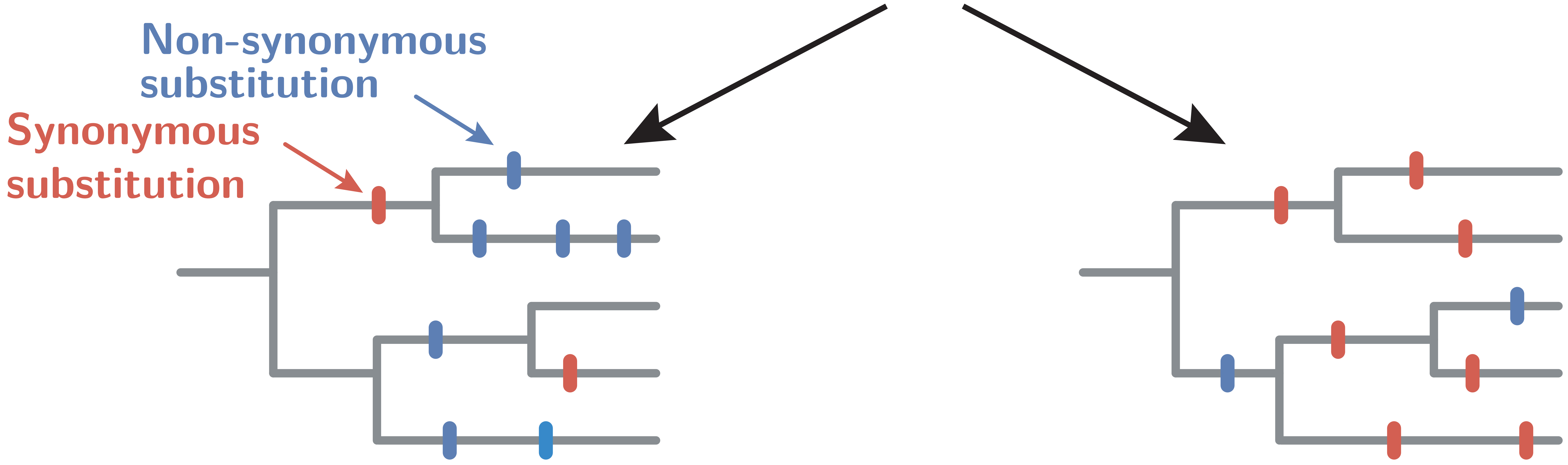
Mutation rates between nucleotides



ω-based phylogenetic codon models

ω

Scaling factor exerced on
non-synonymous mutations



$\omega > 1$

- **Detecting fast evolving genes.**

→ Kosiol *et al* (2008).

- **Detecting rapidly changing sites.**

→ Nieslen & Yang (1998); Enard *et al* (2016).

- **Decting burst of evolution.**

→ Yang & Nielsen (1998); Zhang & Nielsen (2005).

$\omega < 1$

- **Stronger selection for highly expressed proteins.**

→ Drummond (2005); Zhang & Yang (2015).

- **More constrains for buried sites inside a protein.**

→ Ramsey *et al* (2011); Echave *et al* (2016).

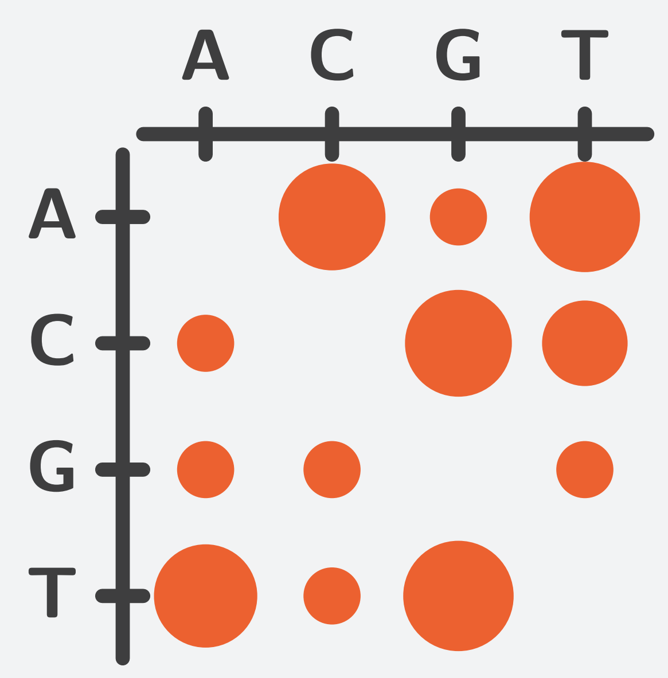
- **Weaker selection for long-lived and bigger species.**

→ Popadin *et al* (2007); Lanfear et al (2010).

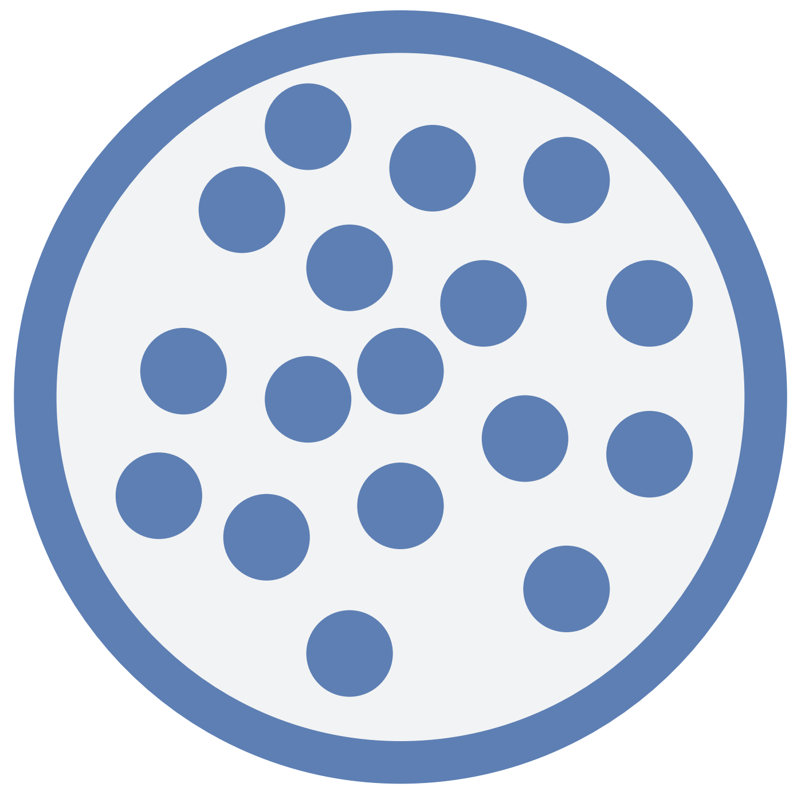
Mutation-selection phylogenetic codon models

- $Q_{i,j}$ is the substitution rate from codon i to j.

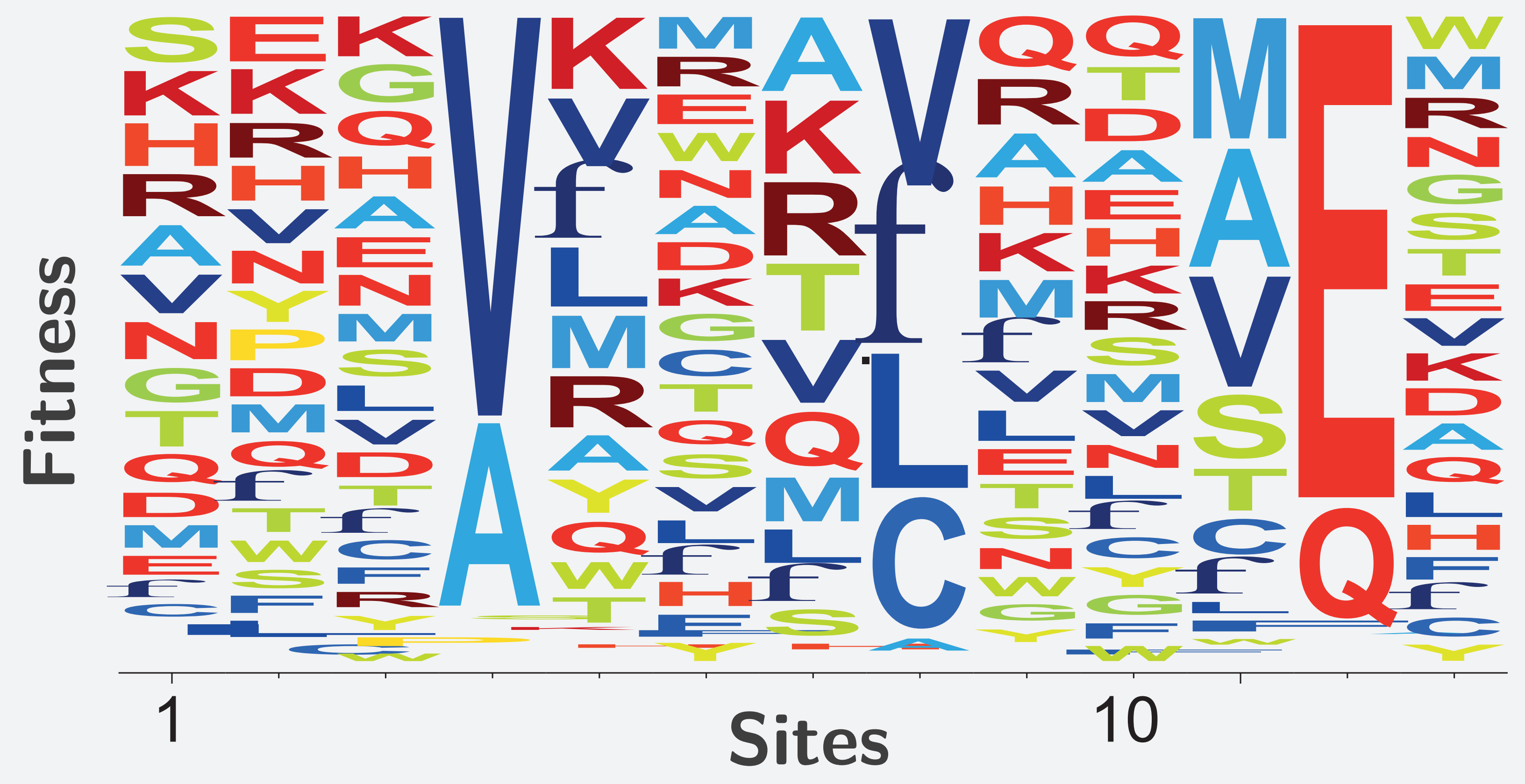
$$\begin{cases} Q_{i,j} &= \mu_{i,j} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} &= \mu_{i,j} \dfrac{4N_{\mathrm{e}}\left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)}\right)}{1 - \mathrm{e}^{4N_{\mathrm{e}}\left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)}\right)}} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases}$$



Mutation rates between nucleotides



Effective population size (N_e)

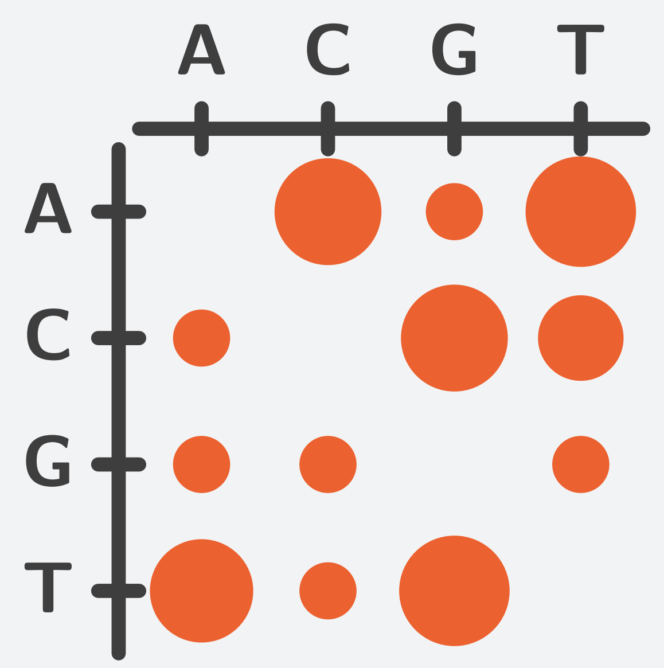


Site-specific amino-acid fitness profiles

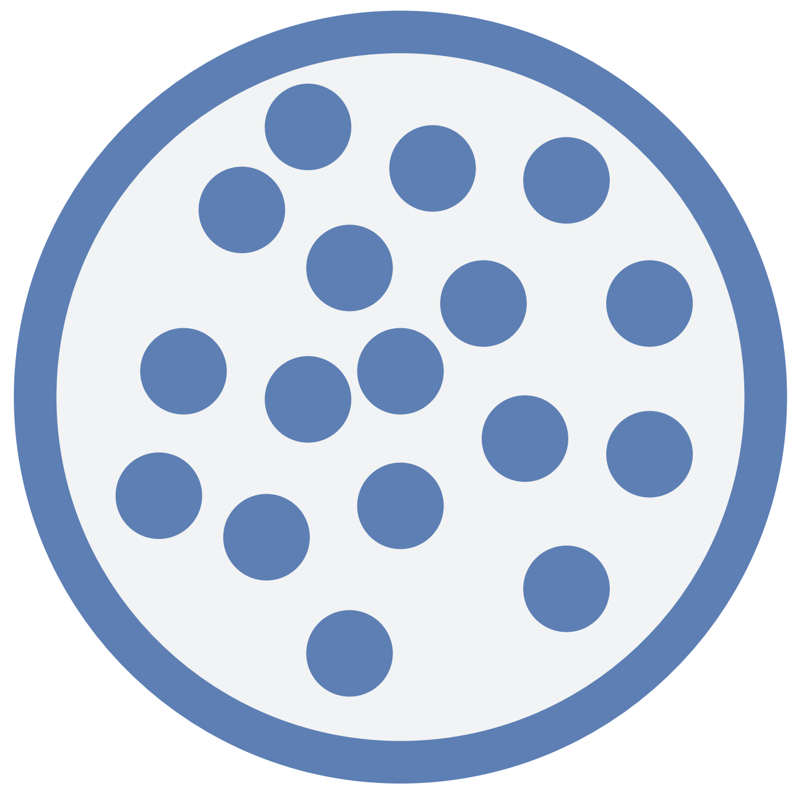
- **Selection on non-synonymous mutations depends on the local physico-chemical properties of amino acids involved in the mutation.**

- **Positive selection in one direction is balanced by purifying selection in the opposite direction.**

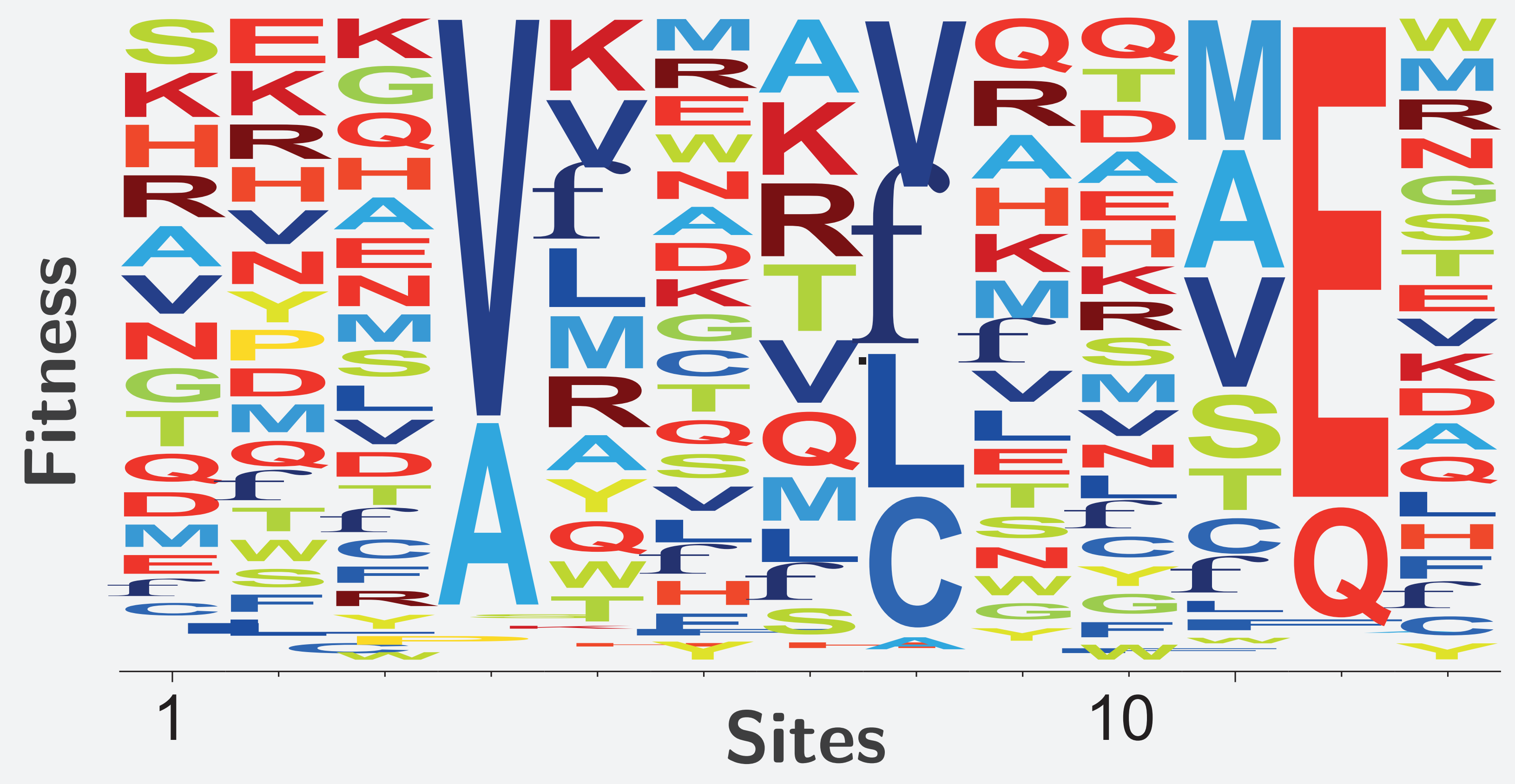
Mutation-selection phylogenetic codon models



Mutation rates between nucleotides



Effective population size (N_e)



Site-specific amino-acid fitness profiles

- **Estimating fitness profiles inside a protein.**

→ Halpern & Bruno (1998); Rodrigue *et al* (2010); Tamuri & Goldstein (2012).

- **Probability of fixation of non-synonymous mutation induced by the model at mutation-selection balance.**

→ Spielman & Wilke (2015); Dos Reis (2015), Jones *et al* (2016).

- **Nearly-neutral model for more sensitive tests of positive selection.**

→ Rodrigue & Lartillot (2017); Bloom (2016); Rodrigue *et al* (2020).

- **Detecting convergent evolution.**

→ Parto & Lartillot (2017).

Part I.

Can mutation-selection codon models estimate variations in N_e along the phylogeny?

Divergence

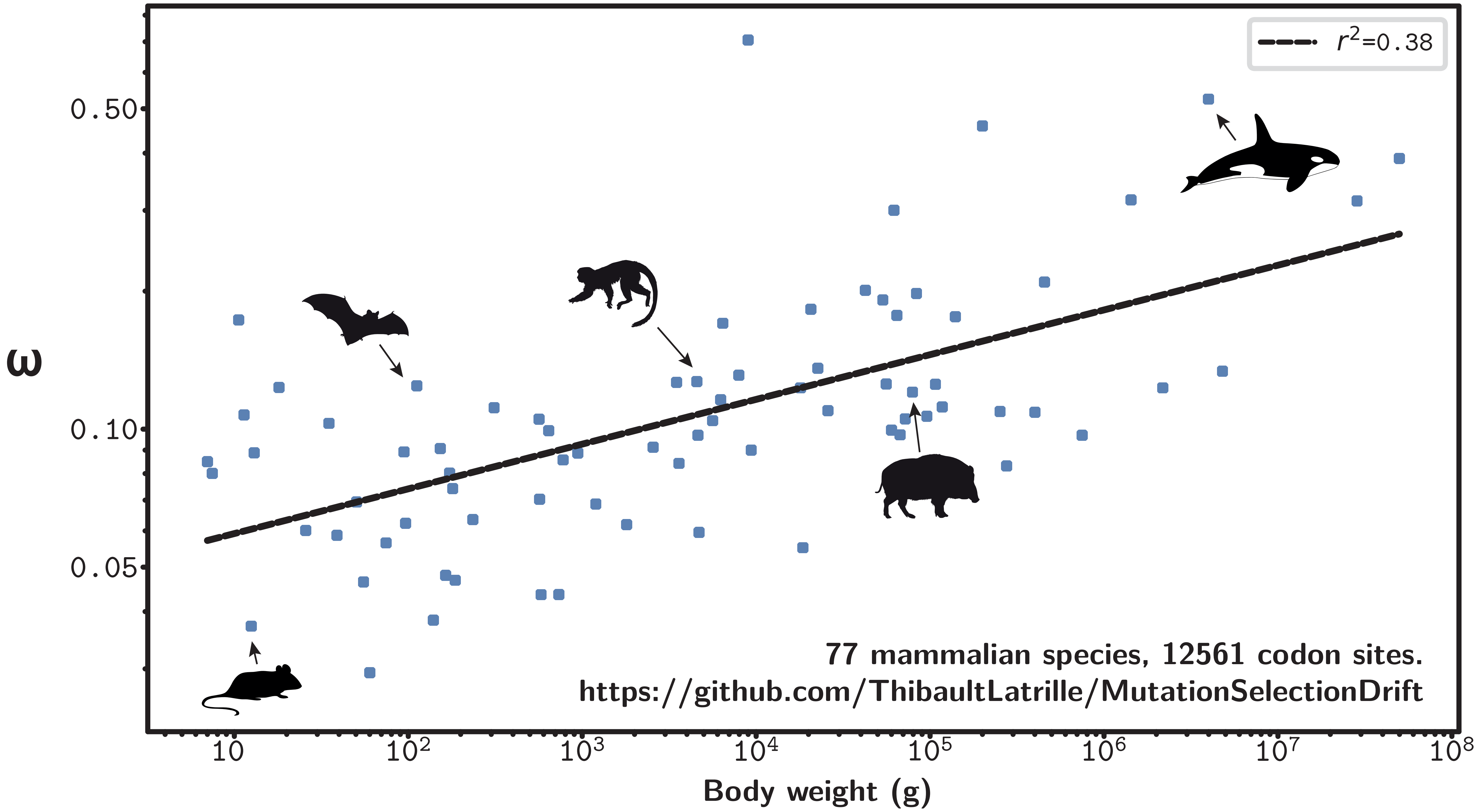
Mutation-selection codon models

Empirical analyses

Simulations

Can ω-based codon models estimate variations in N_e along the phylogeny?

- ω is used as a proxy for N_e in phylogenetic analyses.



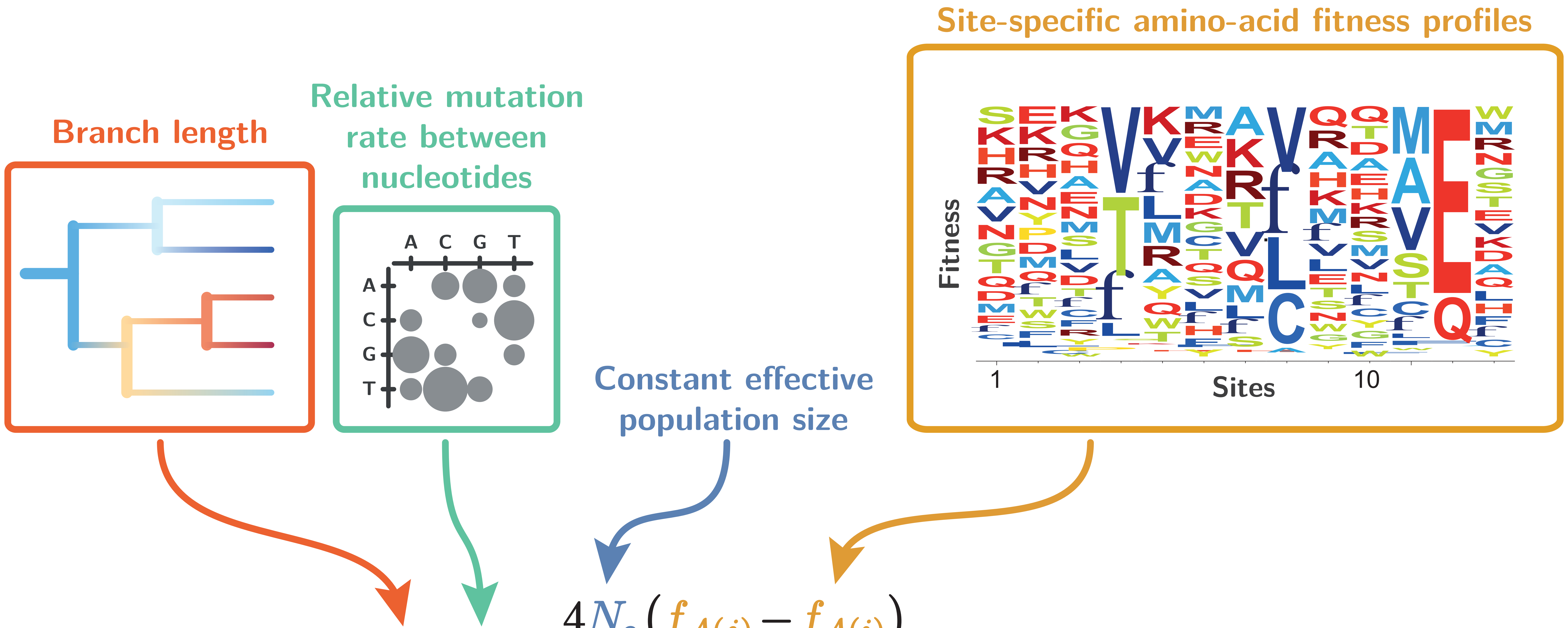
77 mammalian species, 12561 codon sites.
https://github.com/ThibaultLatrille/MutationSelectionDrift

- Used to relate N_e to species life-history traits (longevity, maturity, weight, body size, ...) and ecological traits (habitat, ...).

- Mutation-selection codon models can be parameterized directly with N_e, allowing to revisit these studies.

Popadin *et al* (2007); Lanfear *et al* (2010); Lartillot & Poujol (2011); Lartillot & Delsuc (2012); Romigiuer *et al* (2014); Galtier (2016).

Current mutation-selection codon models assume a constant N_e along the phylogeny



Branch length

Relative mutation rate between nucleotides

Constant effective population size

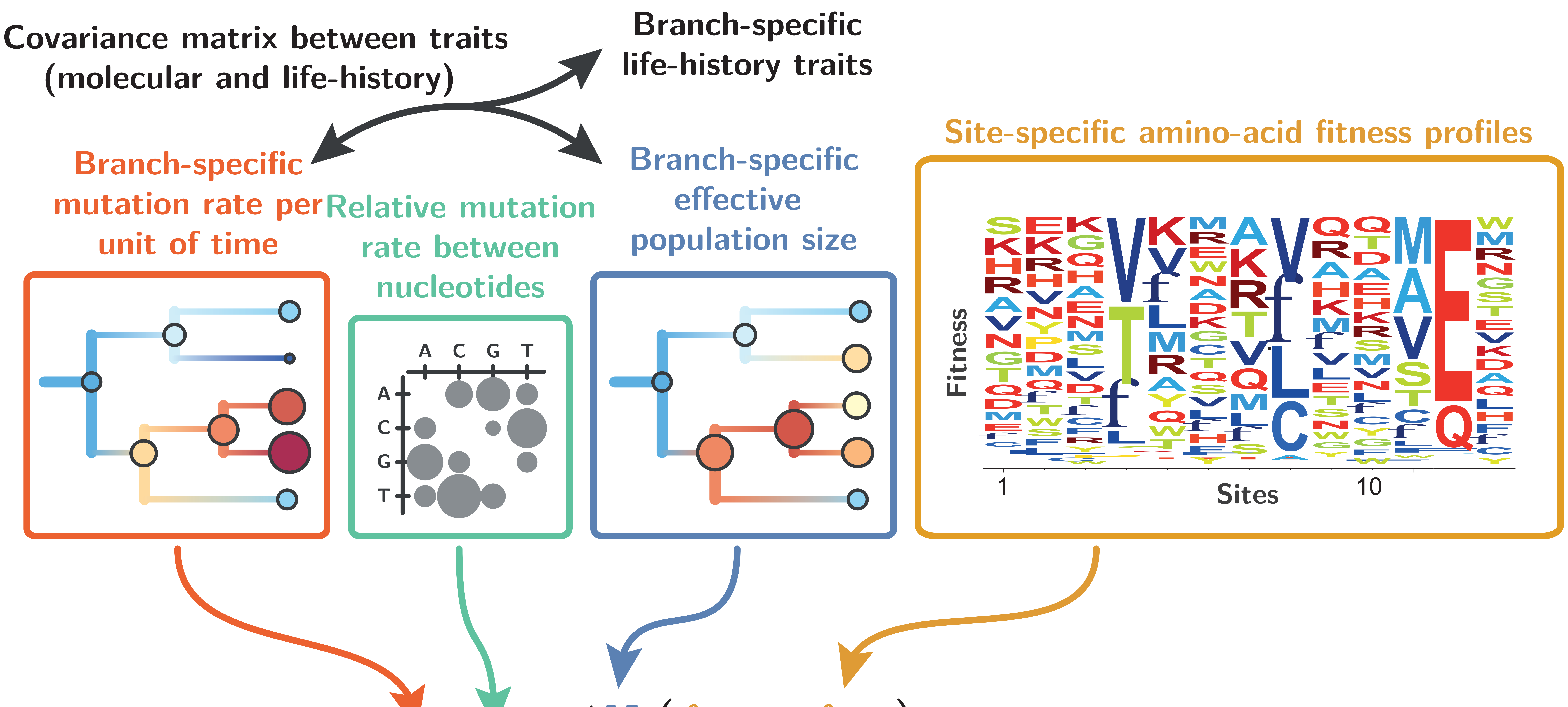
Site-specific amino-acid fitness profiles

$$Q_{i,j} = \mu R_{i \to j} \frac{4N_e\left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)}\right)}{1 - e^{4N_e\left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)}\right)}}$$

- **Selection is heterogeneous between amino acids and along the sequence.**
- **N_e is considered fixed along the different lineages.**

Halpern & Bruno (1998); Rodrique *et al* (2010); Rodrigue & Lartillot (2014); Tamuri *et al* (2014).

Mutation-selection codon models with N_e variations along the phylogeny



Covariance matrix between traits (molecular and life-history)

Branch-specific life-history traits

Branch-specific mutation rate per unit of time

Relative mutation rate between nucleotides

Branch-specific effective population size

Site-specific amino-acid fitness profiles

$$Q_{i,j} = \mu R_{i \to j} \frac{4N_e \left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)} \right)}{1 - e^{4N_e \left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)} \right)}}$$

- **Mutation-selection codon model that estimates selection along the DNA sequence, and N_e along the branches of the tree.**

https://github.com/bayesiancook/bayescode

Input and ouput of the Bayesian framework

Input

Life-history traits

Protein coding DNA alignment

Tree topology

ATG|GGA|TCC|ATG|CTA|CGA|TCG
ATG|CGA|TCC|ATG|GTA|CGA|TCG
ATG|CGA|TCG|AAG|CTT|CGA|TCC
ATG|CGA|TAG|AAG|CTT|CGA|TCG
ATG|CGA|TCG|ATC|CAT|CGA|TCG

Bayesian inference model

Output

Dated tree

Covariance matrix between traits

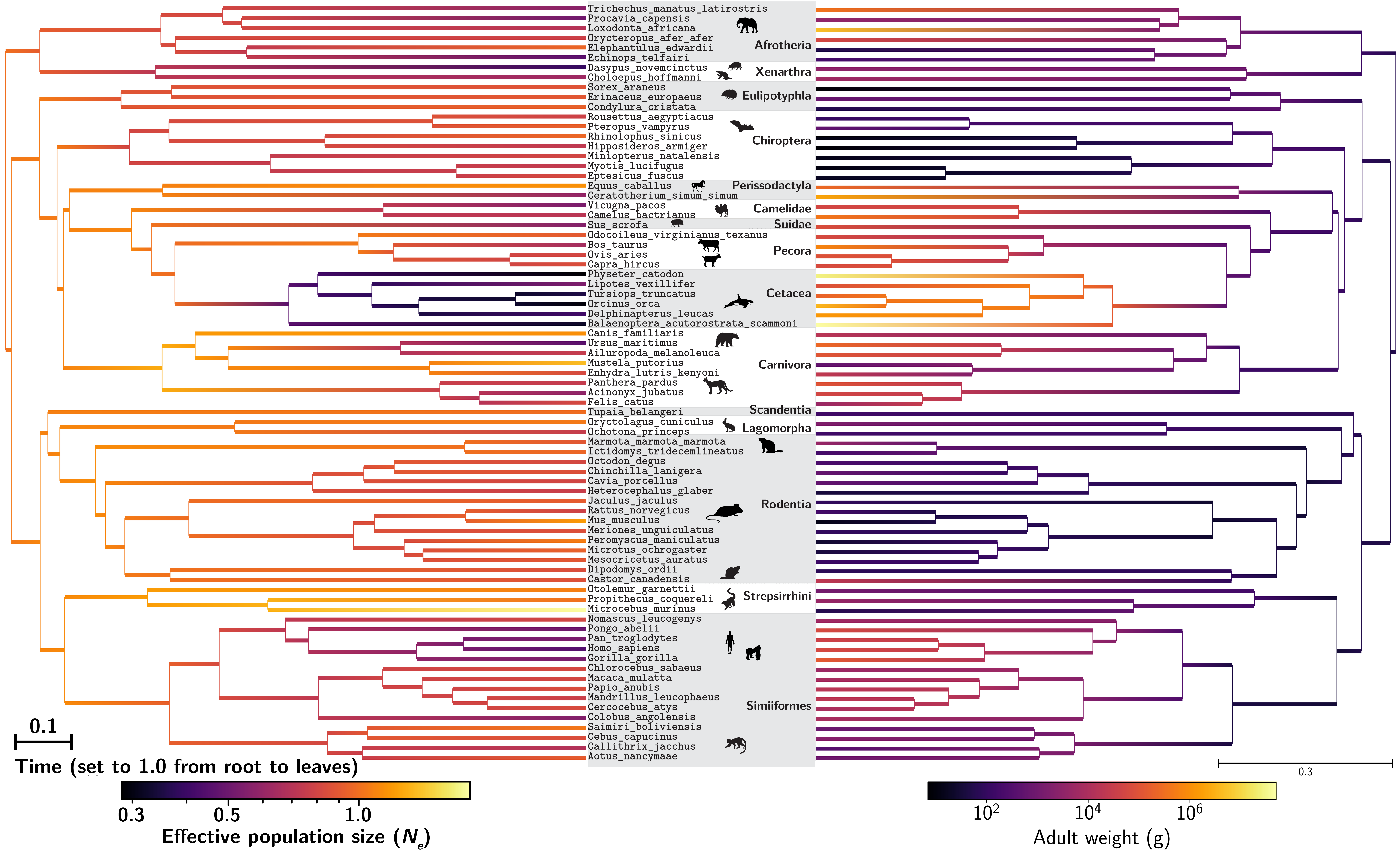
Reconstructed life-history traits, effective population size (N_e), and mutation rate for each branch of the tree.

Site-specific amino-acid fitness profiles

Mutation matrix

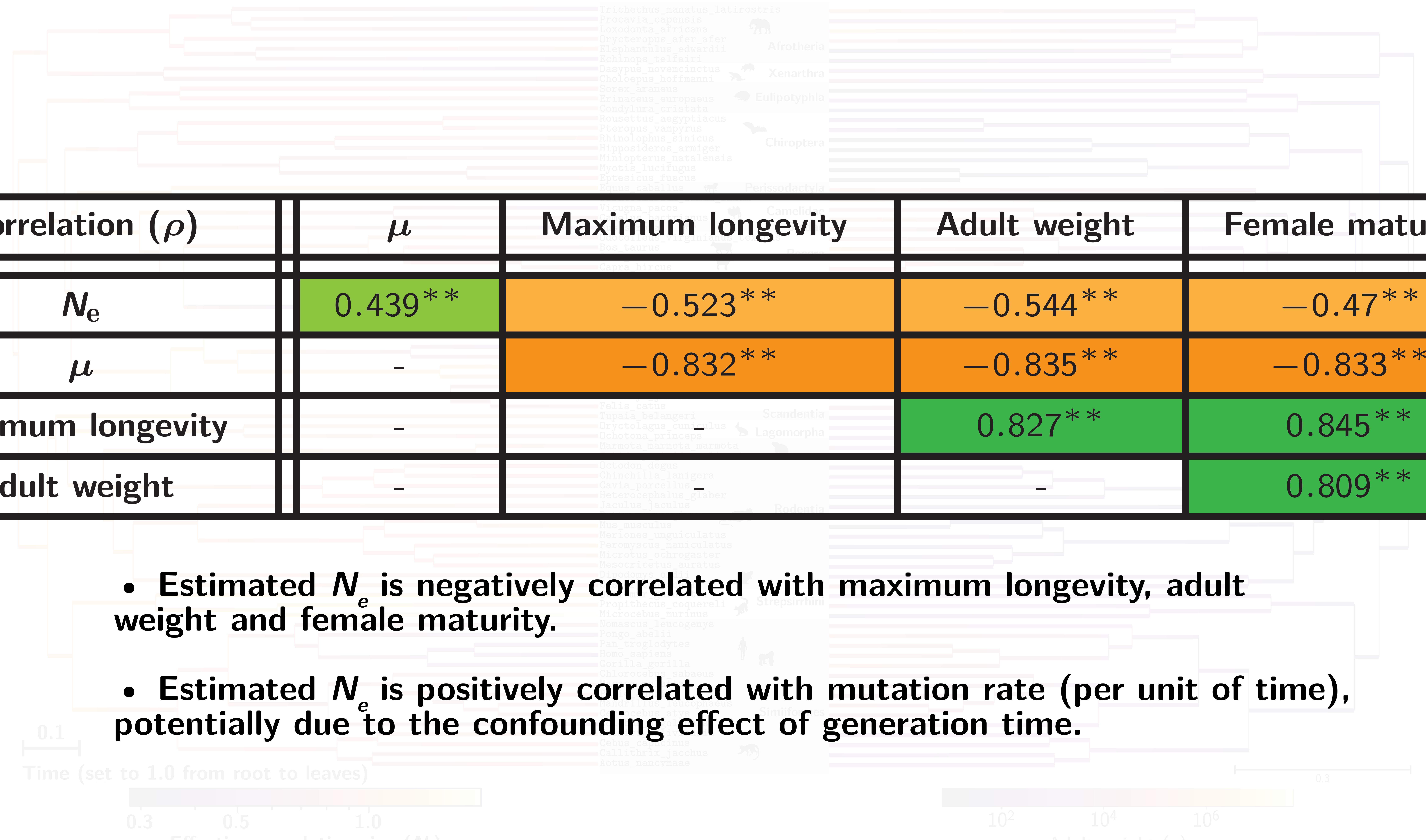
https://github.com/bayesiancook/bayescode

Reconstructing long term changes of N_e in mammals



https://github.com/ThibaultLatrille/MutationSelectionDrift

Estimated N_e is related to life-history traits in mammals

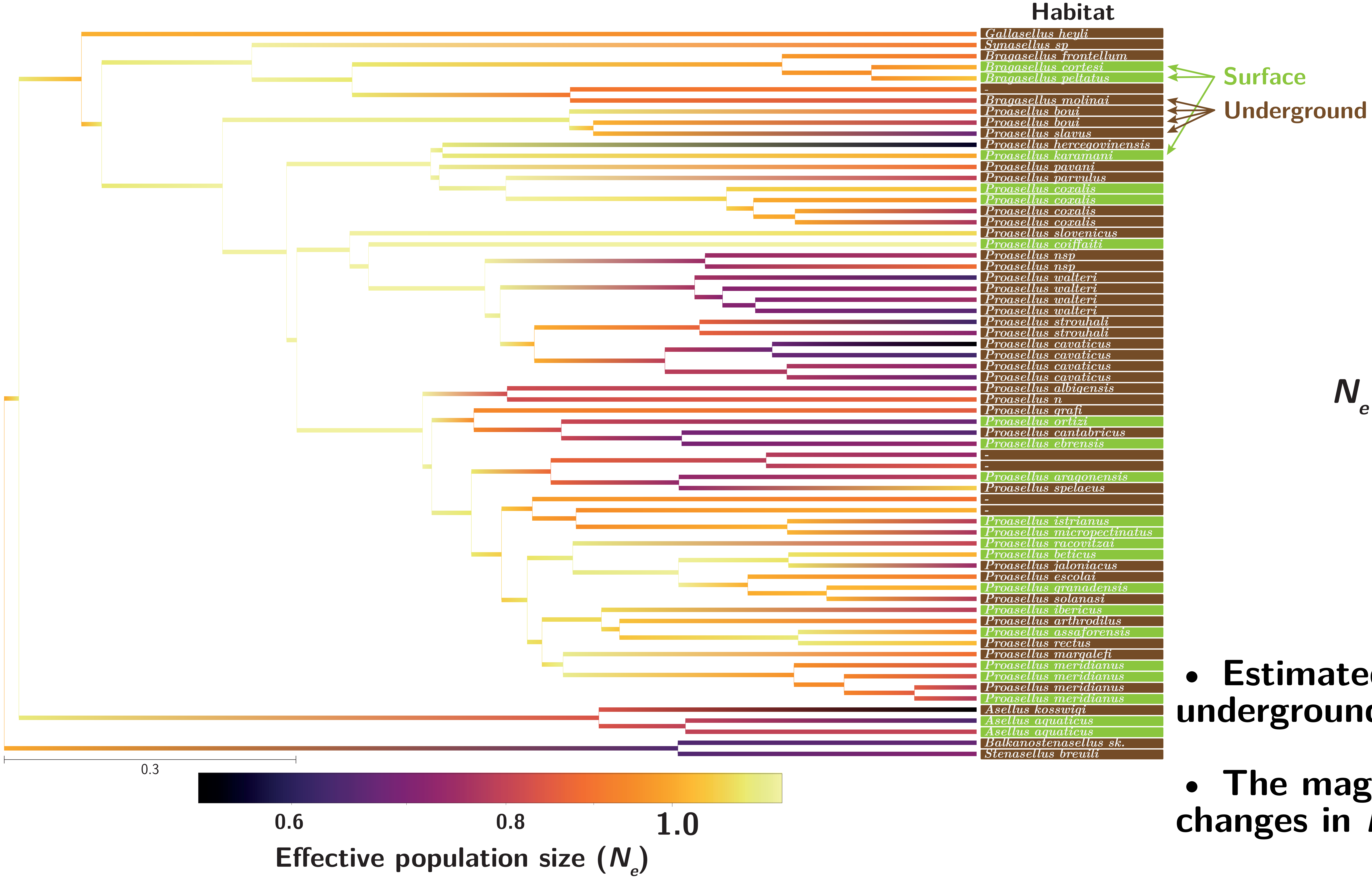


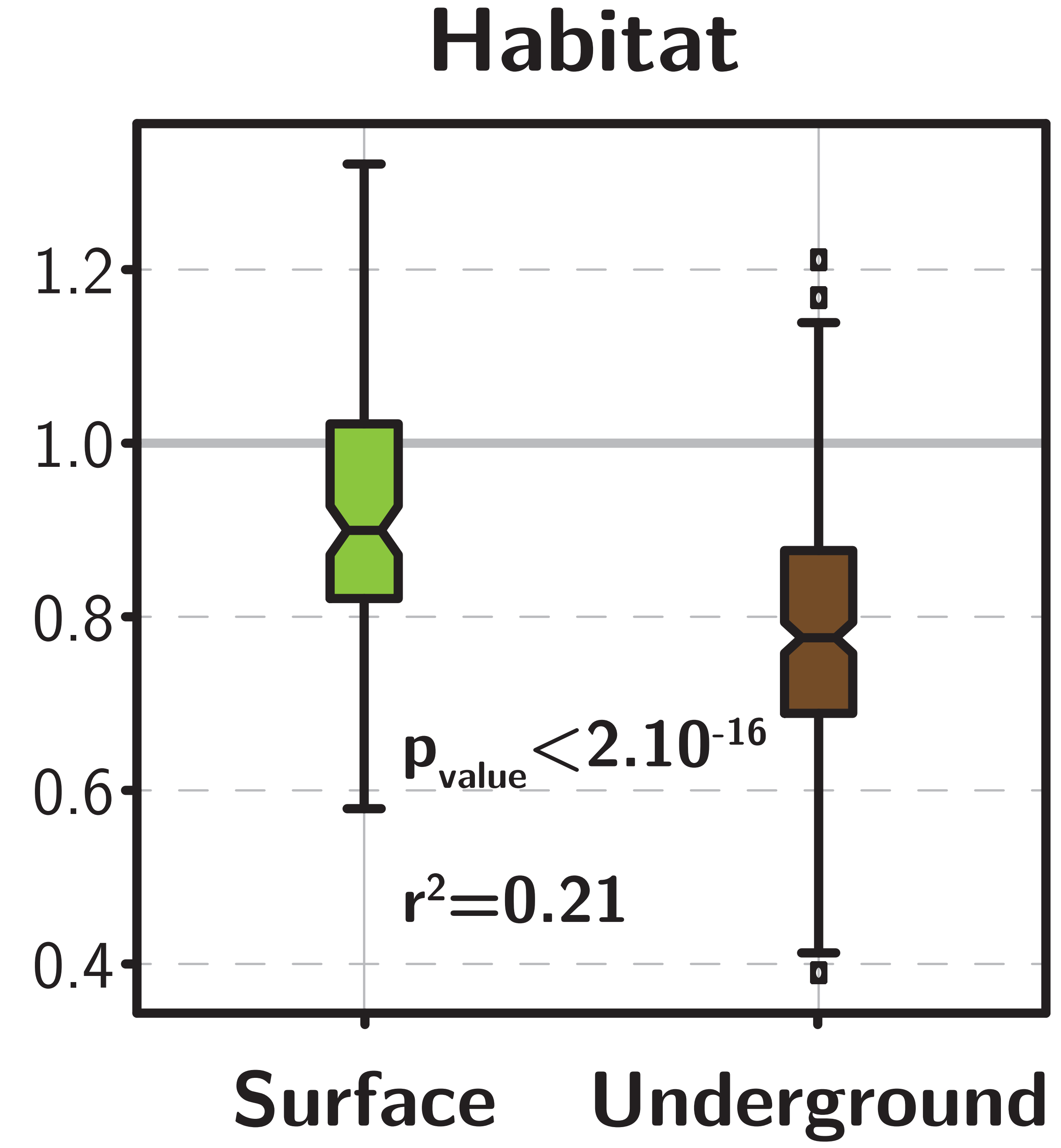
Correlation (ρ)	μ	Maximum longevity	Adult weight	Female maturity
N_e	0.439^{**}	-0.523^{**}	-0.544^{**}	-0.47^{**}
μ	-	-0.832^{**}	-0.835^{**}	-0.833^{**}
Maximum longevity	-	-	0.827^{**}	0.845^{**}
Adult weight	-	-	-	0.809^{**}

- **Estimated N_e is negatively correlated with maximum longevity, adult weight and female maturity.**
- **Estimated N_e is positively correlated with mutation rate (per unit of time), potentially due to the confounding effect of generation time.**

https://github.com/ThibaultLatrille/MutationSelectionDrift

Estimated N_e is related to ecological traits in isopods





- **Estimated N_e is lower for underground species.**
- **The magnitude of estimated changes in N_e is low.**

https://github.com/ThibaultLatrille/MutationSelectionDrift

Capderrey et al (2013); Eme et al (2013); Saclier et al (2018)

Validating the inference model against simulated alignments

Fitness landscape

Species tree

Mutation-Selection models

Simulation along the phylogeny

ATG|GGA|TCC|ATG|CTA|CGA|TCG
ATG|CGA|TCC|ATG|GTA|CGA|TCG
ATG|CGA|TCG|AAG|CTT|CGA|TCC
ATG|CGA|TAG|AAG|CTT|CGA|TCG
ATG|CGA|TCG|ATC|CAT|CGA|TCG

Inference of Mutation rate, N_e, ...

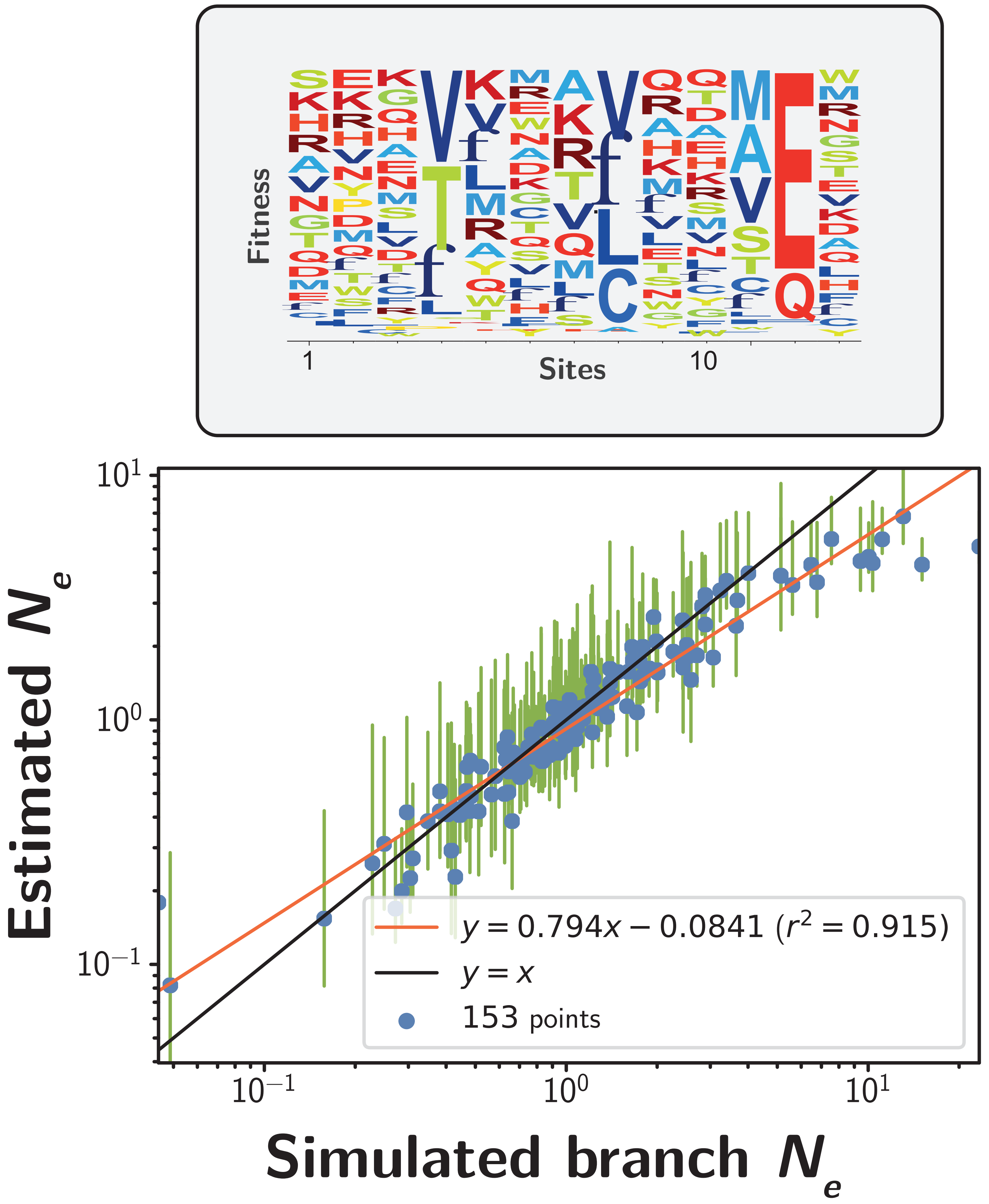
Nucleotide mutation rate matrix

Fluctuation of population size, mutation rate and generation time

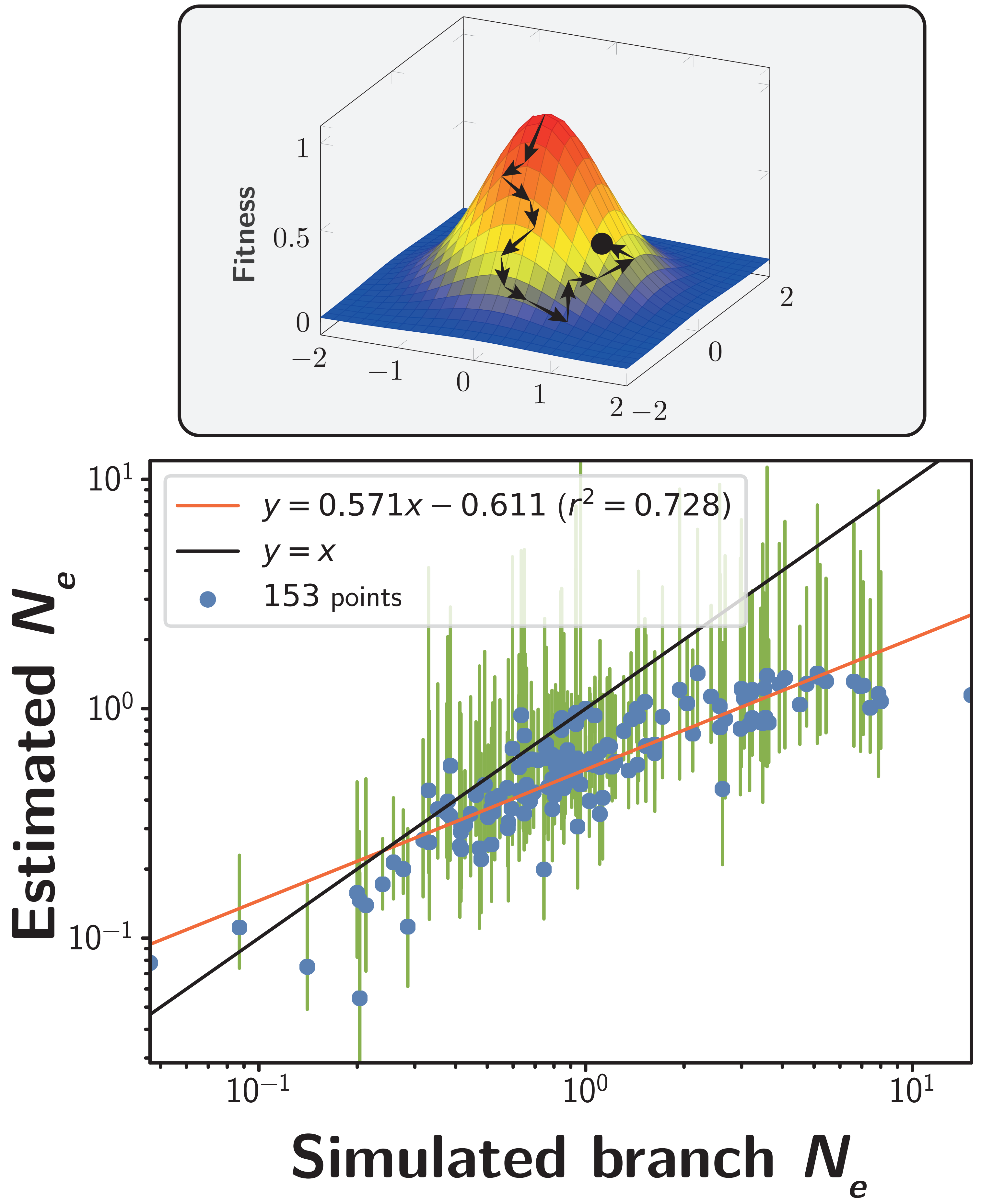
Comparing inference and simulation

N_e cannot be realiably estimated in the presence of epistasis

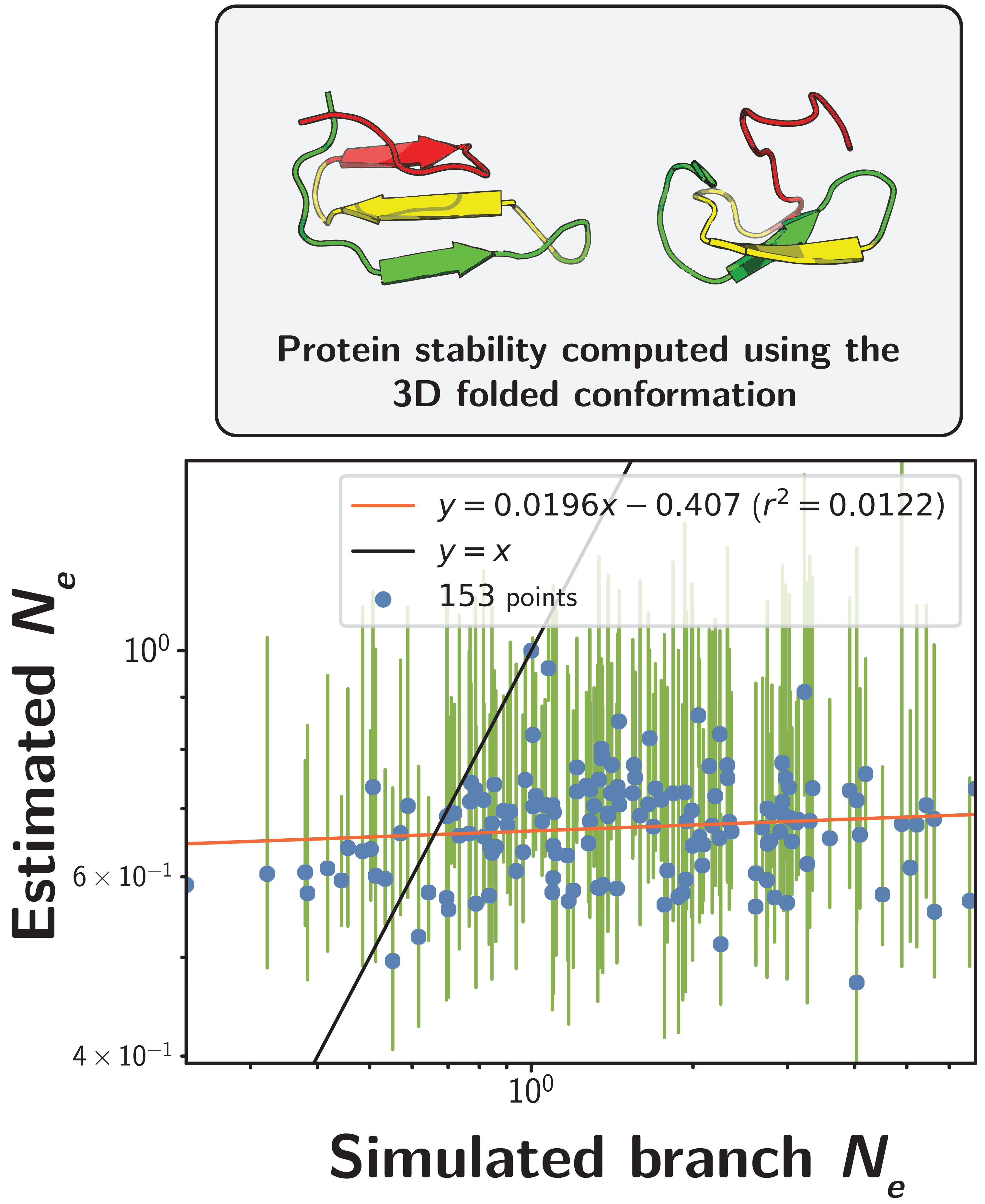
Site-specific amino-acid fitness profiles



Fisher geometric fitness landscape



Protein stability fitness landscape



Protein stability computed using the 3D folded conformation

Increased epistatic interactions between sites

Harder to estimate the underlying population size (N_e)

https://github.com/ThibaultLatrille/MutationSelectionDrift

Can mutation-selection codon models estimate changes in N_e along the phylogeny?

- In mammals, estimated N_e correlates negatively with longevity, weight and maturity, and positively with mutation rate.
- In isopods, underground lineages have a lower estimated N_e.
- The changes in N_e along lineages are in the expected direction, but the range of estimated N_e is lower than expected.
- Which mechanism could explain such a low variance of N_e estimated in empirical data?
- Epistasis appears to be a reasonable explanation.

Acknowledgements

Advisor

Nicolas Lartillot

Collaborators & reviewers

Benoit Nabholz (ISEM)
Christophe Douady (LEHNA)
Tristan Lefébure (LEHNA)
Nicolas Rodrigue (Ottawa Univ.)
Julien Yann Dutheil (Max Planck Inst.)
Richard Goldstein (UCL)
Carina Farah Mugal (Uppsala Univ.)

LBBEers

Laurent Gueguen
Laurent Duret
Anamaria Necsulea
Céline Brochier-Armanet
Damien de Vienne
Dominique Mouchiroud
Hélène Badouin
Bastien Boussau
Éric Tannier
Vincent Daubin
...

Teachers

Marie Sémon
Annabelle Haudry
Carine Rey
Corentin Dechaut
Vincent Lacroix
Arnaud Mary
Catherine Moulia
...

IT team

Bruno Spataro
Stéphane Delmotte
Simon Penel
Adil El Filali
Vincent Miele
Aurélie Siberchicot
Philippe Veber

The A team

Nathalie Arbasetti
Odile Mulet-Marquis
Laetitia Catouaria
Aurélie Zerfass

Finches

Florian Bénitière
Alexandre Laverré
Alexia Nguyen Trung
Djivan Prentout
Théo Tricou
Marina Brasó Vives
Claire Gayral
Hugo Menet
Louis Duchemin
Antoine Villié
Alice Genestier
Julien Joseph
Elise Say-Sallaz
...