Modélisation de l'articulation des mécanismes sélectifs et neutres dans l'évolution des séquences d'ADN codant pour des protéines

30 Novembre 2020

Thibault Latrille

- **Nicolas LARTILLOT, Directeur de recherche, CNRS/LBBE (Lyon)**
- **Céline BROCHIER-ARMANET, Professeure, Université Claude Bernard Lyon 1**
- **Julien Yann DUTHEIL, Research Group Leader, Max Planck Institute (Allemagne)**
- **Richard GOLDSTEIN, Professeur, University College London (Royaume-Uni)**
- **Carina Farah MUGAL, Chercheure, Uppsala University (Suède)**

LBBE LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE

UNIVERSITÉ DE LYON

cnrs

CC-BY-SA 2020

Modelling the interplay between selective and neutral mechanisms in the evolution of protein-coding DNA sequences

Introduction: dissecting the thesis title

I. Inferring mutation in presence of selection

II. Inferring genetic drift in presence of mutation and selection

III. Rate of evolution as a function of genetic drift

Conclusion

Introduction

Modelling the interplay

between selective and neutral mechanisms

in the evolution

of protein-coding DNA sequences.

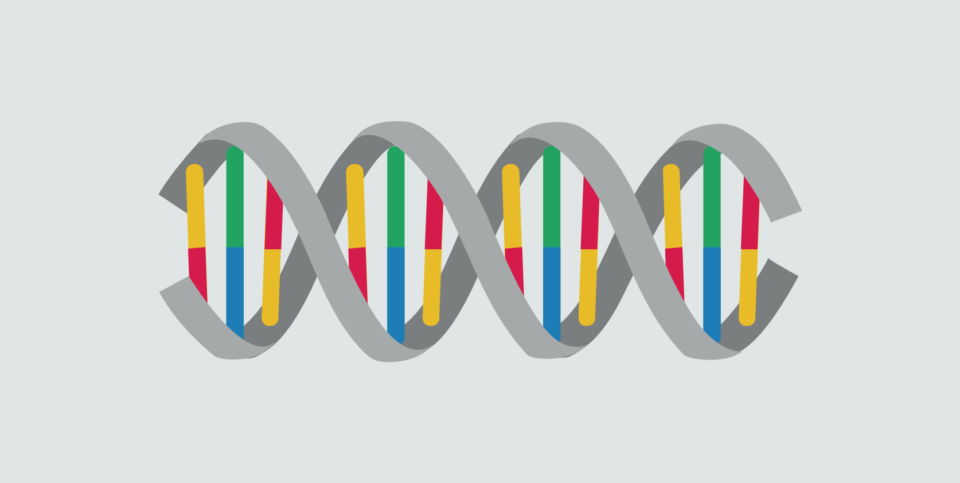
- **The introduction will consist in dissecting the title of this thesis, bottom up.**

Protein-coding DNA sequences

Replication

DNA

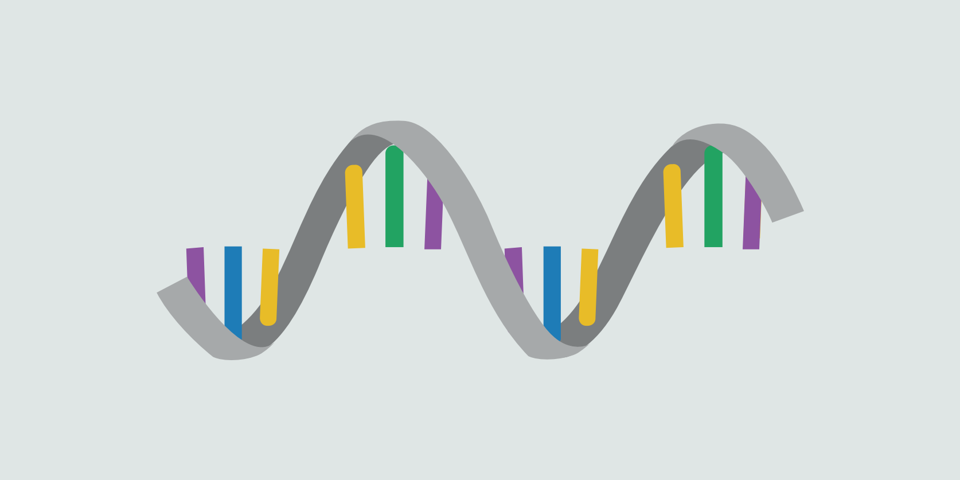
4-nucleotide alphabet



Transcription

RNA

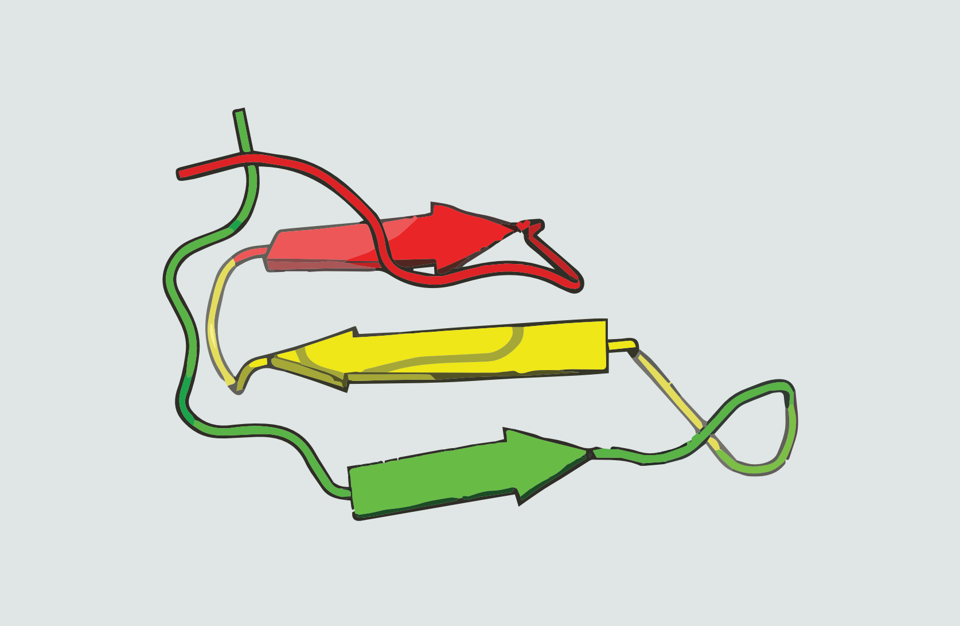
4-nucleotide alphabet



Translation

Protein

20-amino-acid alphabet



ATG|CTC| ... |CTA|CGC

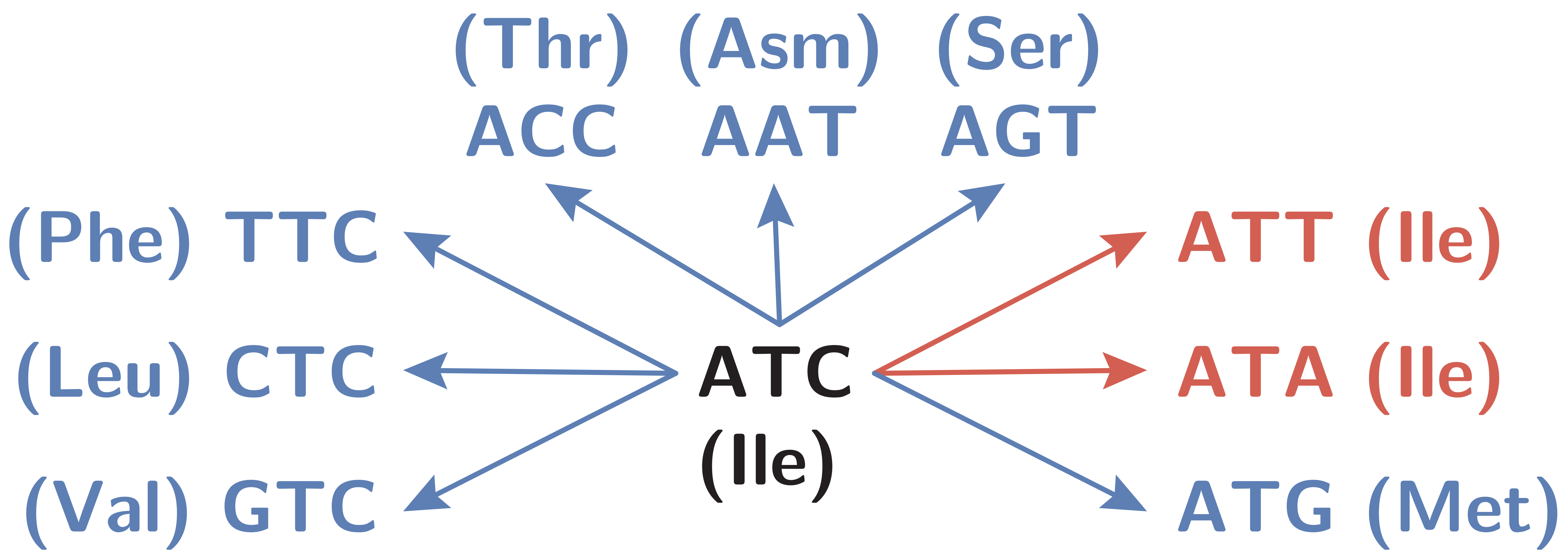
		T		C		A		G	
T	TTT	Phenylalanine (Phe/P)	TCT	Serine (Ser/S)	TAT	Tyrosine (Tyr/Y)	TGT	Cysteine (Cys/C)	T
	TTC		TCC		TAC		TGC		C
	TTA	Leucine (Leu/L)	TCA		TAA	Stop (Ochre)	TGA	Stop (Opal)	A
	TTG		TCG		TAG	Stop (Amber)	TGG	Tryptophan (Trp/W)	G
C	CTT		CCT	Proline (Pro/P)	CAT	Histidine (His/H)	CGT	Arginine (Arg/R)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Glutamine (Gln/Q)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Isoleucine (Ile/I)	ACT	Threonine (Thr/T)	AAT	Asparagine (Asn/N)	AGT	Serine (Ser/S)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lysine (Lys/K)	AGA	Arginine (Arg/R)	A
	ATG	Methionine (Met/M)	ACG		AAG		AGG		G
G	GTT	Valine (Val/V)	GCT	Alanine (Ala/A)	GAT	Aspartic acid (Asp/D)	GGT	Glycine (Gly/G)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glutamic acid (Glu/E)	GGA		A
	GTG		GCG		GAG		GGG		G

Genetic code table (4^3=64 codons)

Methionine|Leucine| ... |Leucine|Alanin

Franklin & Gosling (1953); Watson & Crick (1953); Wilkins *et al* (1953); Crick (1958); Crick (1970).

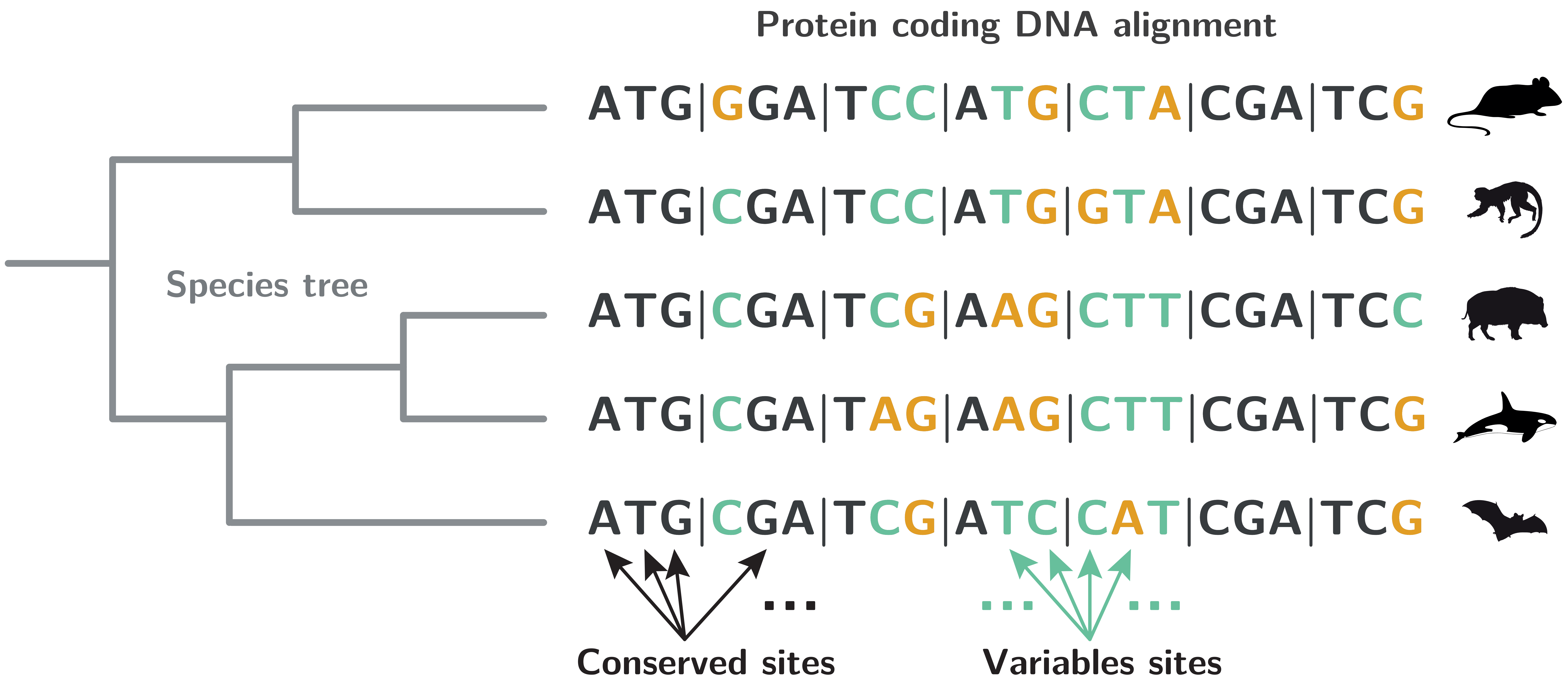
DNA mutations change the protein, or not.



	T		C		A		G		
T	TTT	Phenylalanine (Phe/P)	TCT	Serine (Ser/S)	TAT	Tyrosine (Tyr/Y)	TGT	Cysteine (Cys/C)	T
	TTC		TCC		TAC		TGC		C
	TTA	Leucine (Leu/L)	TCA		TAA	Stop (Ochre)	TGA	Stop (Opal)	A
	TTG		TCG		TAG	Stop (Amber)	TGG	Tryptophan (Trp/W)	G
C	CTT		CCT	Proline (Pro/P)	CAT	Histidine (His/H)	CGT	Arginine (Arg/R)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Glutamine (Gln/Q)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Isoleucine (Ile/I)	ACT	Threonine (Thr/T)	AAT	Asparagine (Asn/N)	AGT	Serine (Ser/S)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lysine (Lys/K)	AGA	Arginine (Arg/R)	A
	ATG	Methionine (Met/M)	ACG		AAG		AGG		G
G	GTT	Valine (Val/V)	GCT	Alanine (Ala/A)	GAT	Aspartic acid (Asp/D)	GGT	Glycine (Gly/G)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glutamic acid (Glu/E)	GGA		A
	GTG		GCG		GAG		GGG		G

- Non-synonymous mutations change the protein.
- Synonymous mutations do not change the protein.

Evolution of protein-coding DNA sequences



Protein coding DNA alignment

Species tree

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

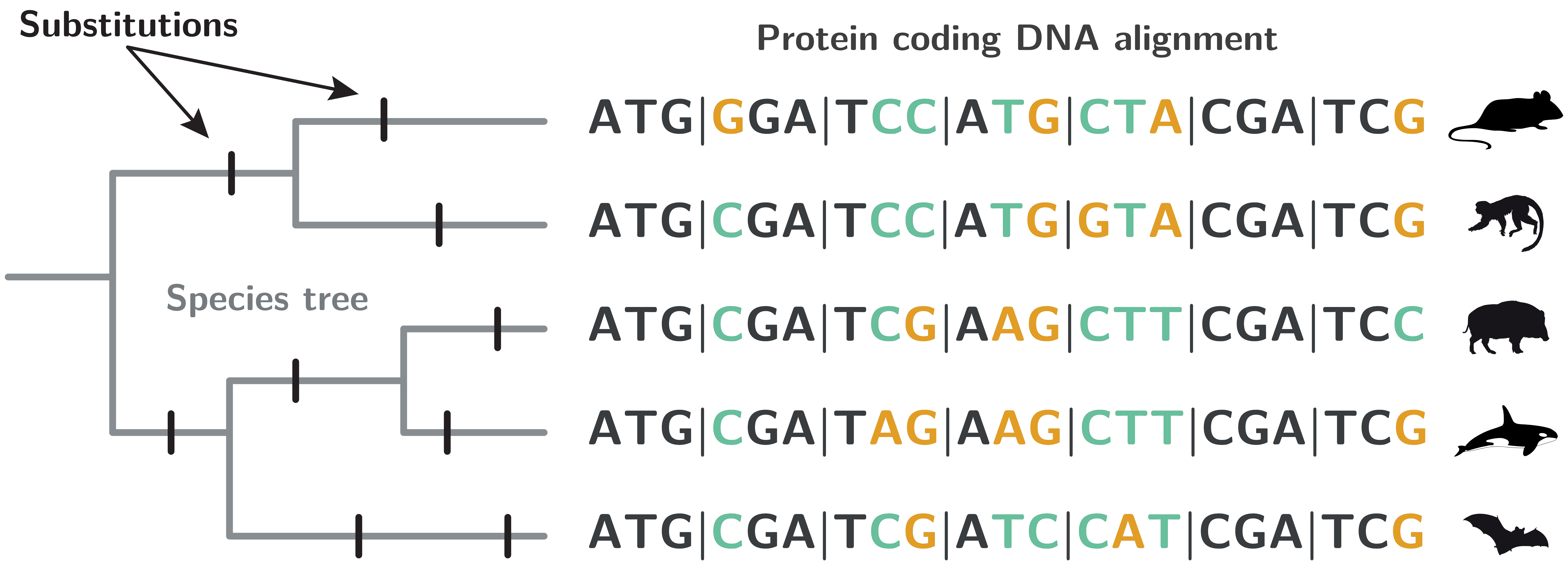
Conserved sites

Variables sites

- **Sequences from the same gene in different species are aligned.**

Zuckerland & Pauling (1995).

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

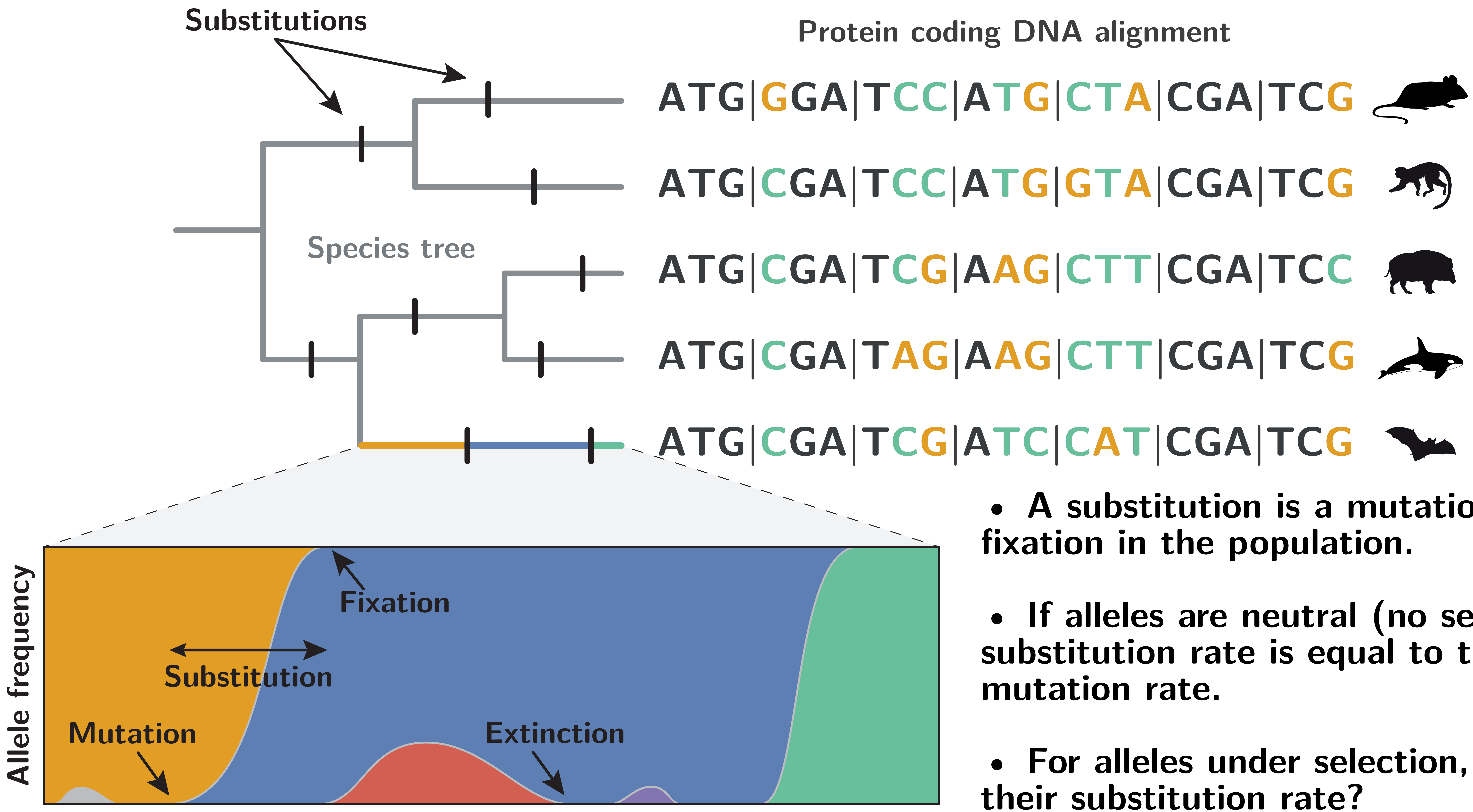
ATG|CGA|TAG|AAG|CTT|CGA|TCG

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

- **Differences correspond to point substitution events happening in the ancestral branches**

Felsenstein (1981).

History of substitutions along the species tree


Substitutions
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
Species tree
Allele frequency
Fixation
Substitution
Mutation
Extinction

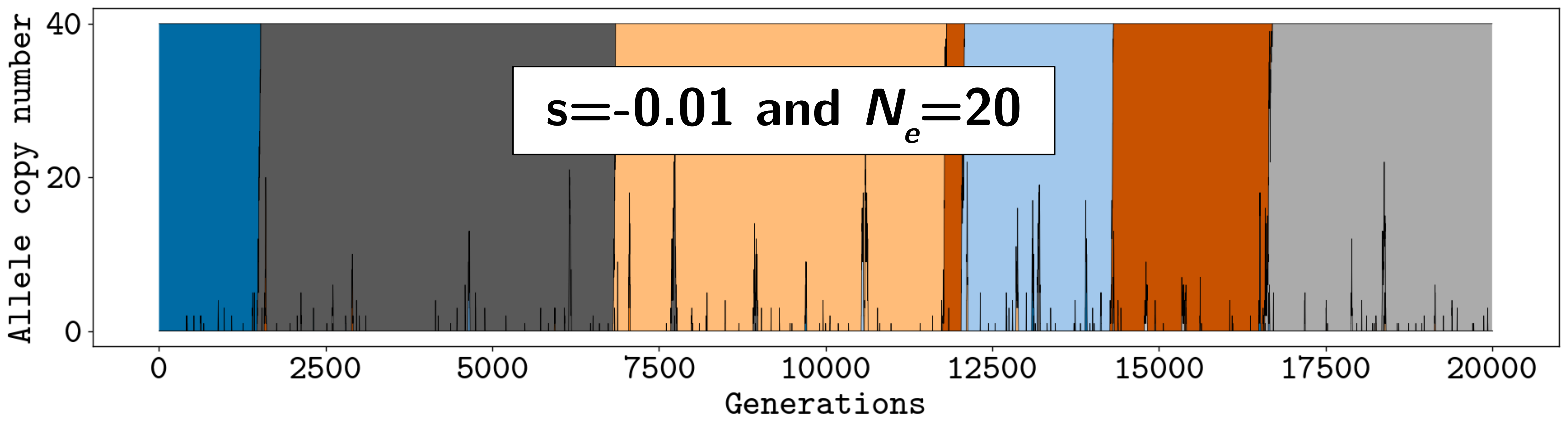
- A substitution is a mutation that reached fixation in the population.

- 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?

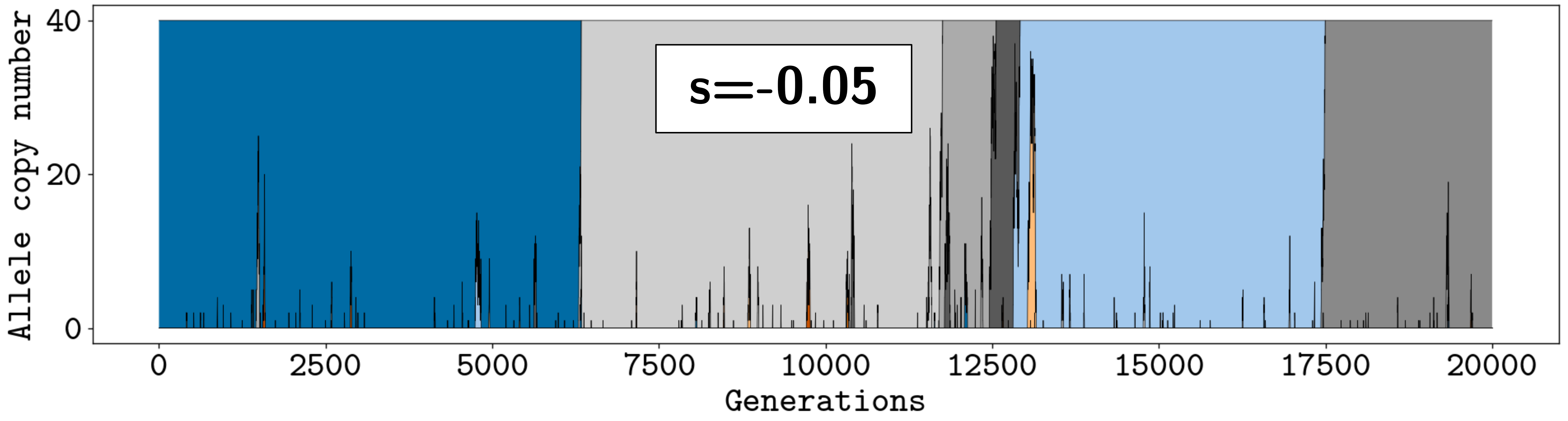
Felsenstein (1981); Kimura (1983); Ohta (1992).

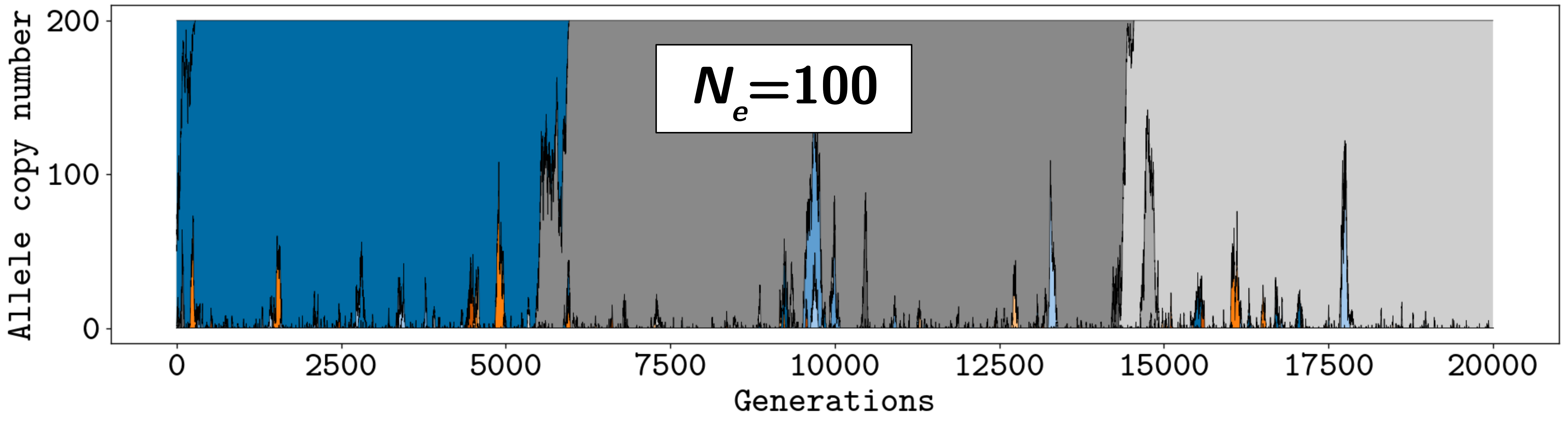
The effect of selection and genetic drift

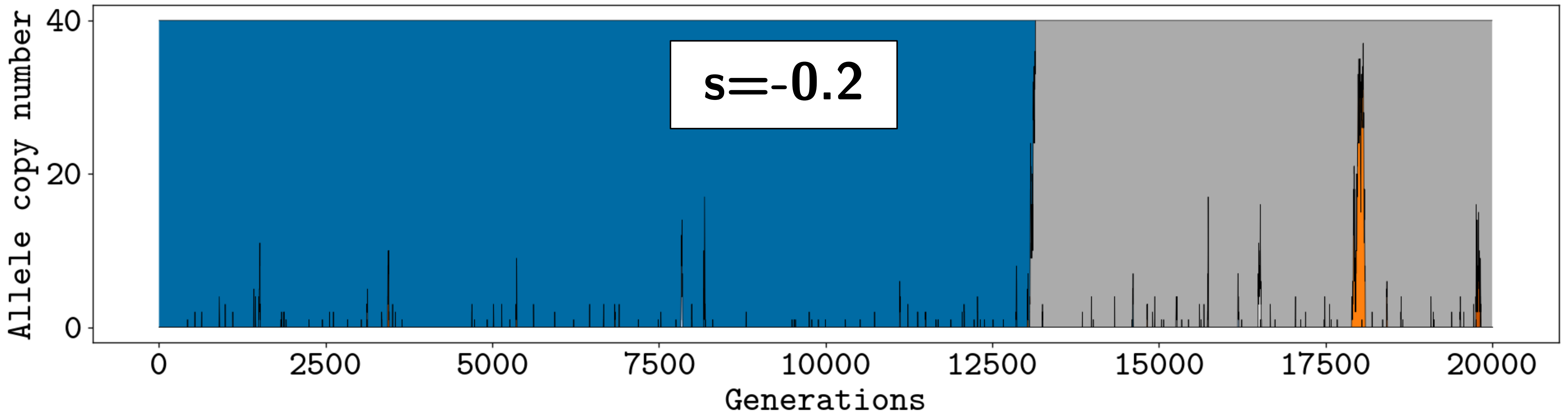


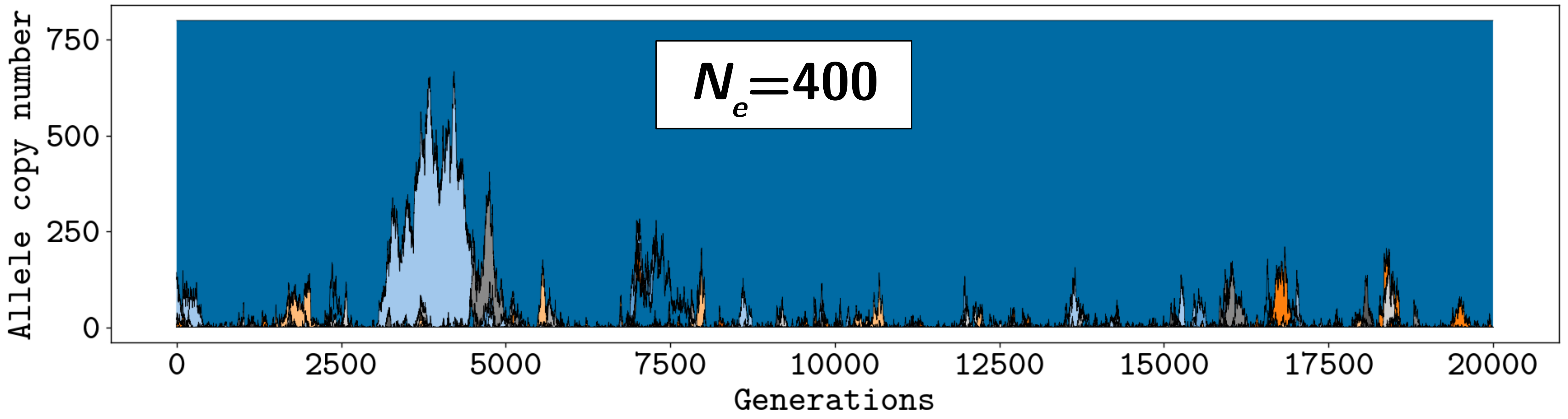
Stronger negative selection coefficient

Increased population size (N_e)





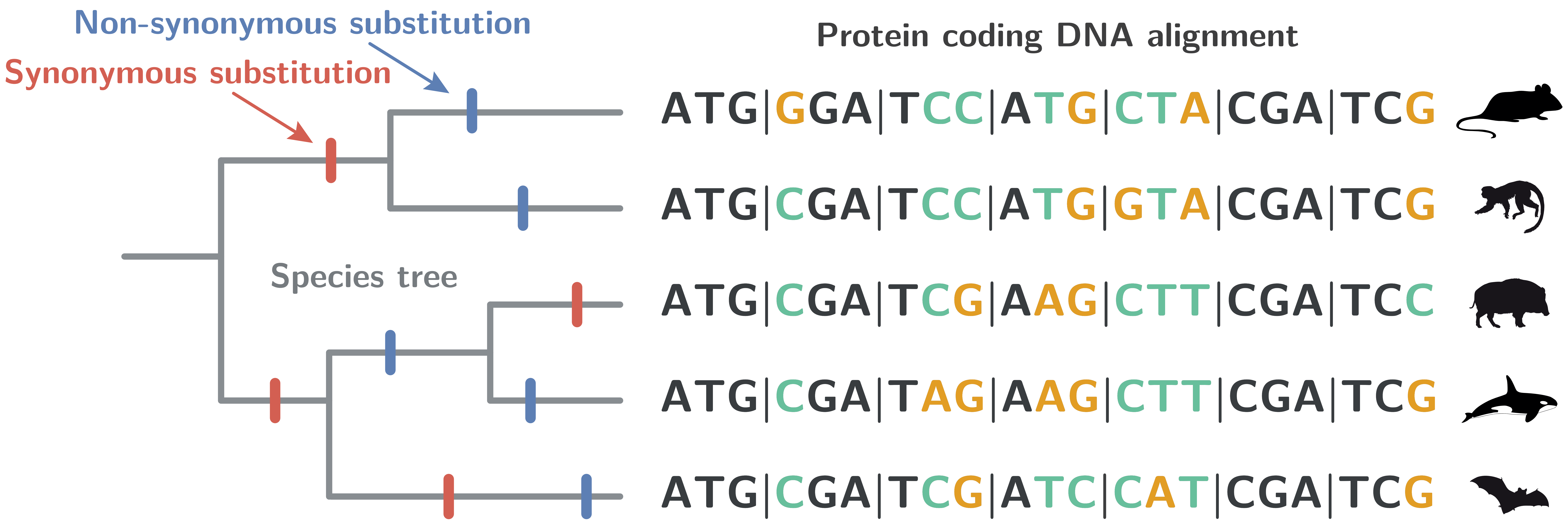




- **Stronger negative selection coefficient results in a decrease of the fixation probability.**
- **Effective population size (N_e) acts as a magnifier of selection.**

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 estimating the strength of selection exercised on proteins.

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

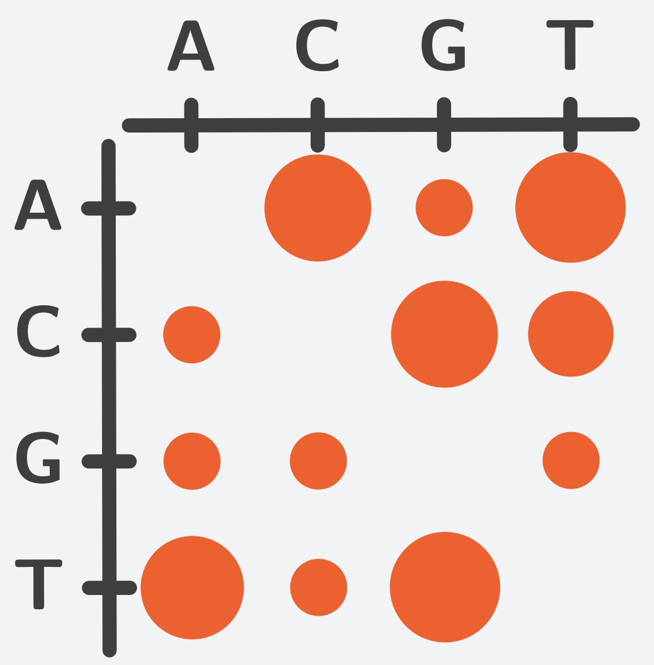
ω-based phylogenetic codon models

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

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

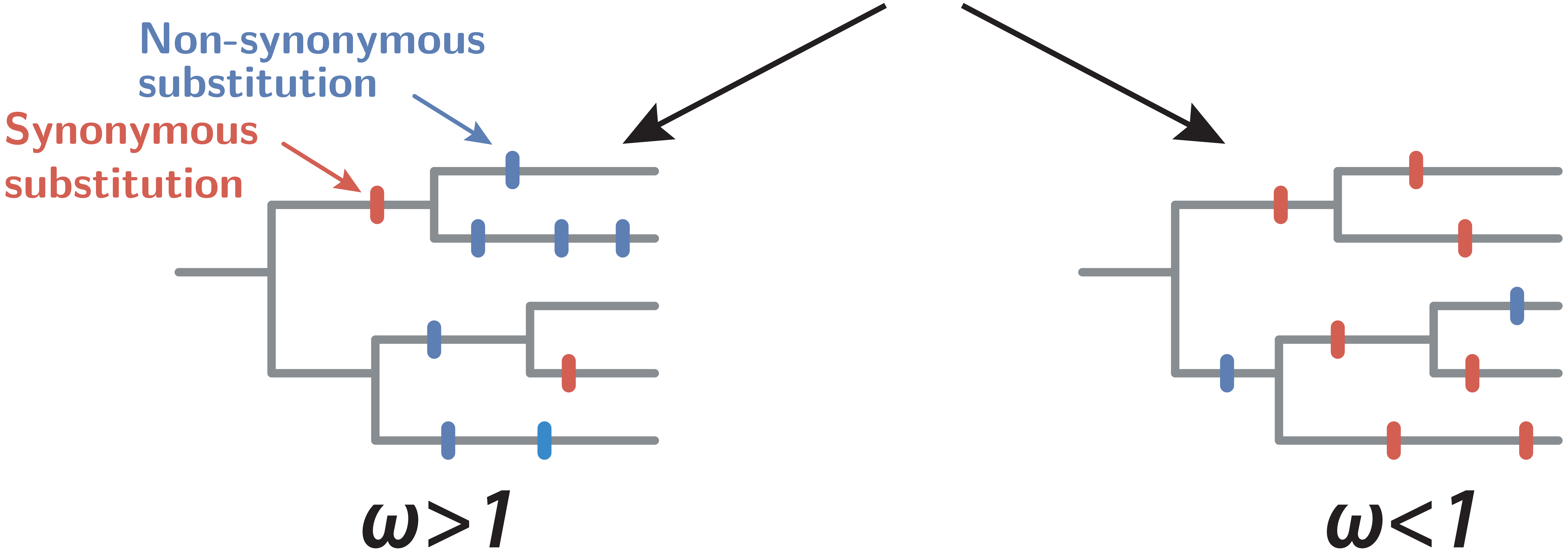
ω

Scaling factor exerced on non-synonymous mutations



$\mu_{i,j}$

Mutation rates between nucleotides

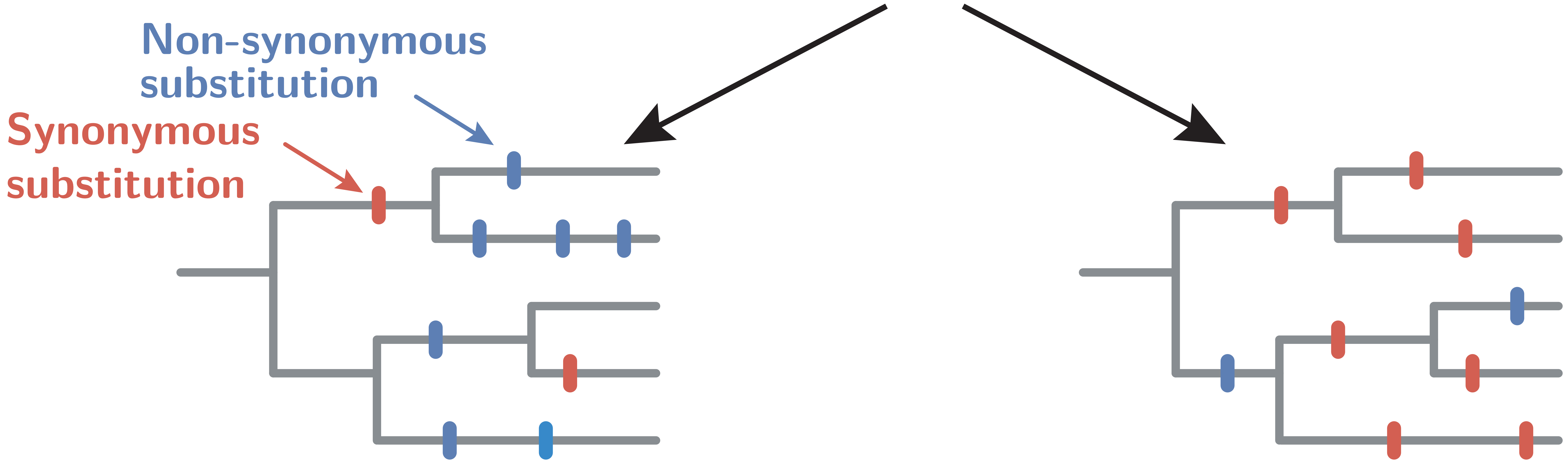


- ω can be interpreted as the average fixation probability of non-synonymous mutations, relative to neutral mutations.

ω-based phylogenetic codon models

ω

Scaling factor exerced on non-synonymous mutations



ω>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).

ω<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).

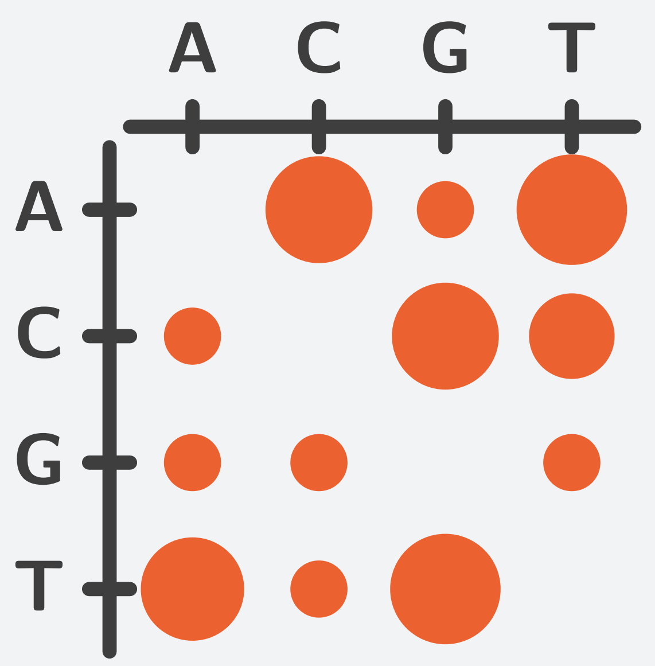
- **Weaker selection for long-lived and bigger species.**

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

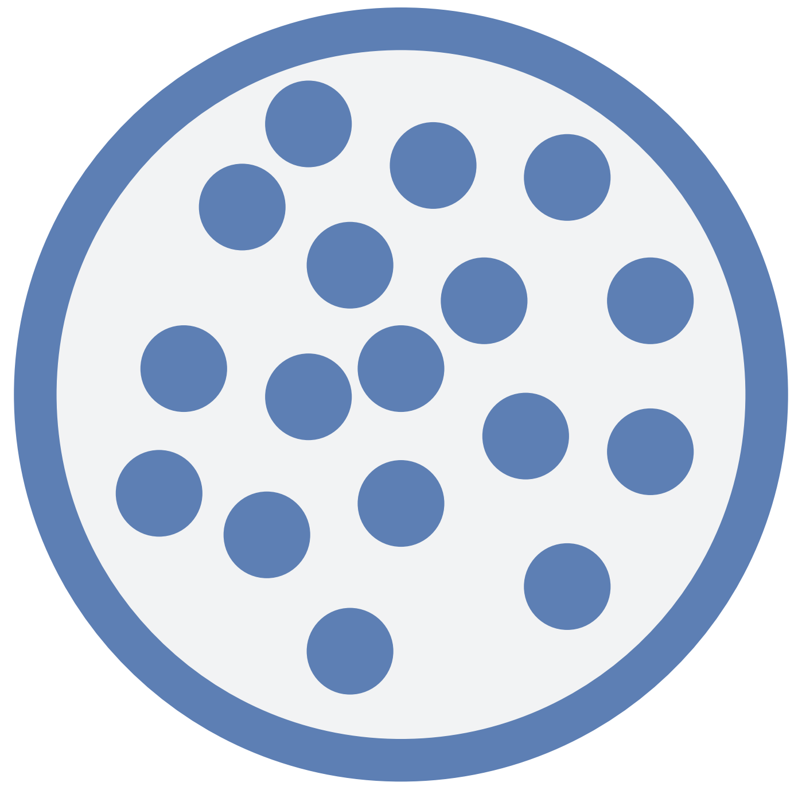
Mutation-selection phylogenetic codon models

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

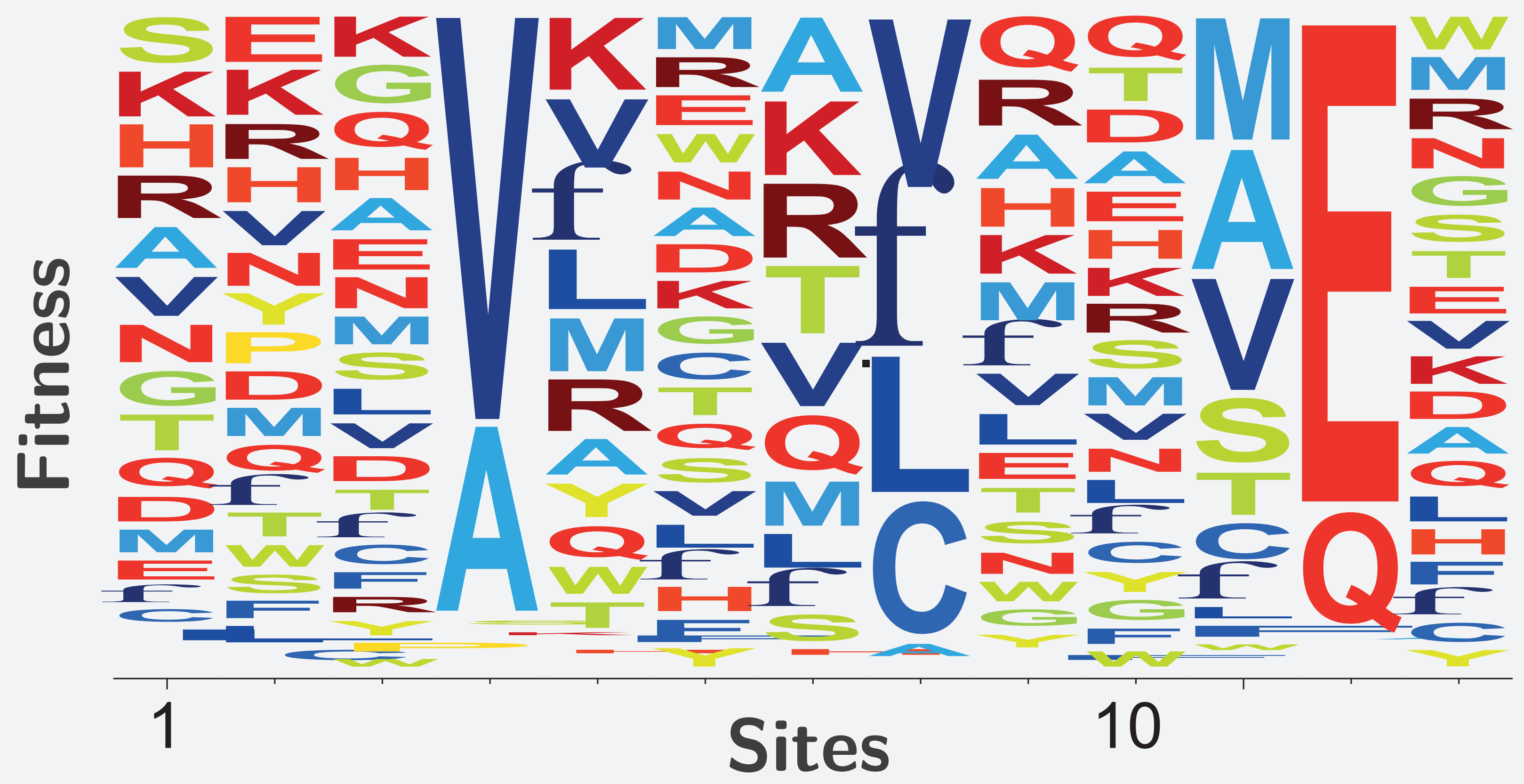
$$\begin{cases} Q_{i,j} & = \mu_{i,j} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} & = \mu_{i,j} \dfrac{4N_\mathrm{e}\left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)}\right)}{1 - \mathrm{e}^{4N_\mathrm{e}\left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)}\right)}} \text{ if codons } i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



Mutation rates between nucleotides



Effective population size (N_e)

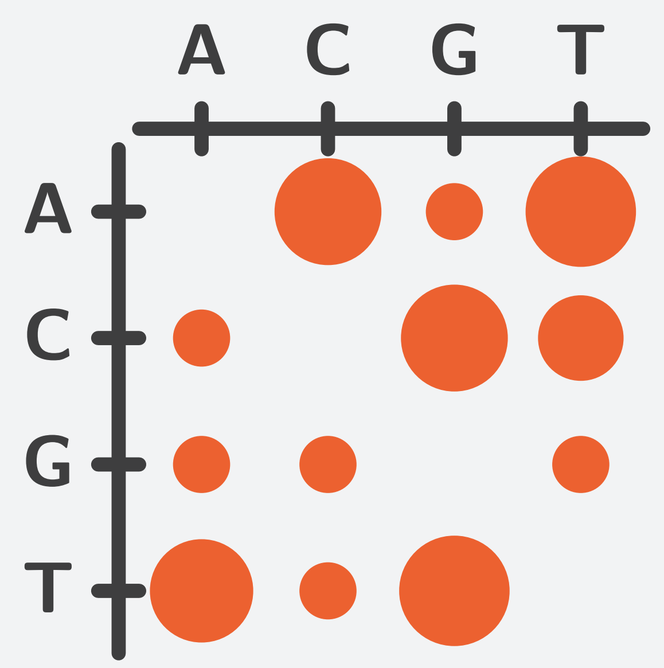


Site-specific amino-acid fitness profiles

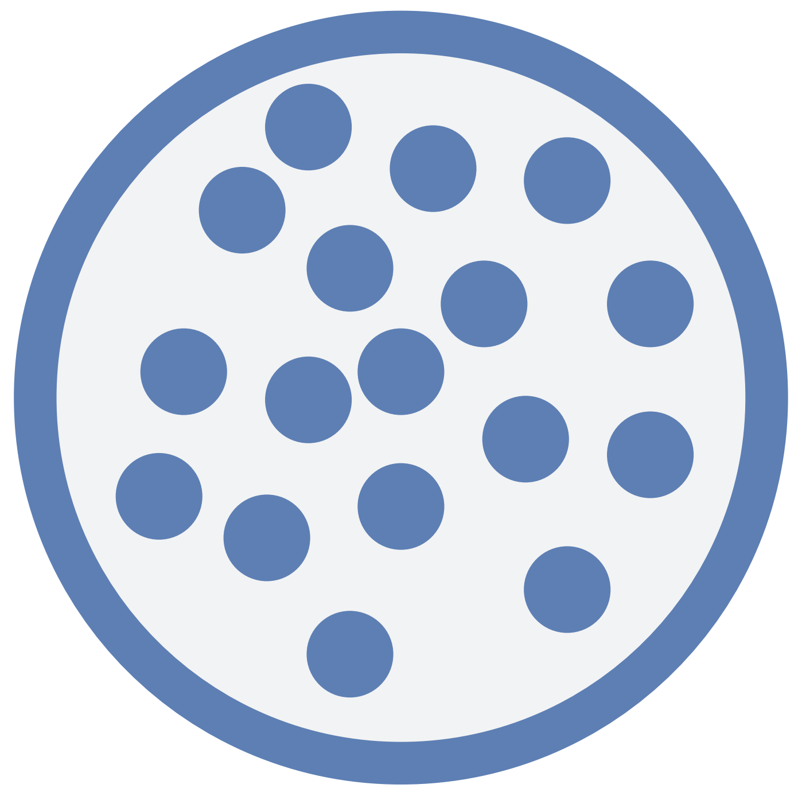
- **Selection on non-synonymous mutations depends on the local physico-chemical properties of amino acids involved in the mutation.**

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

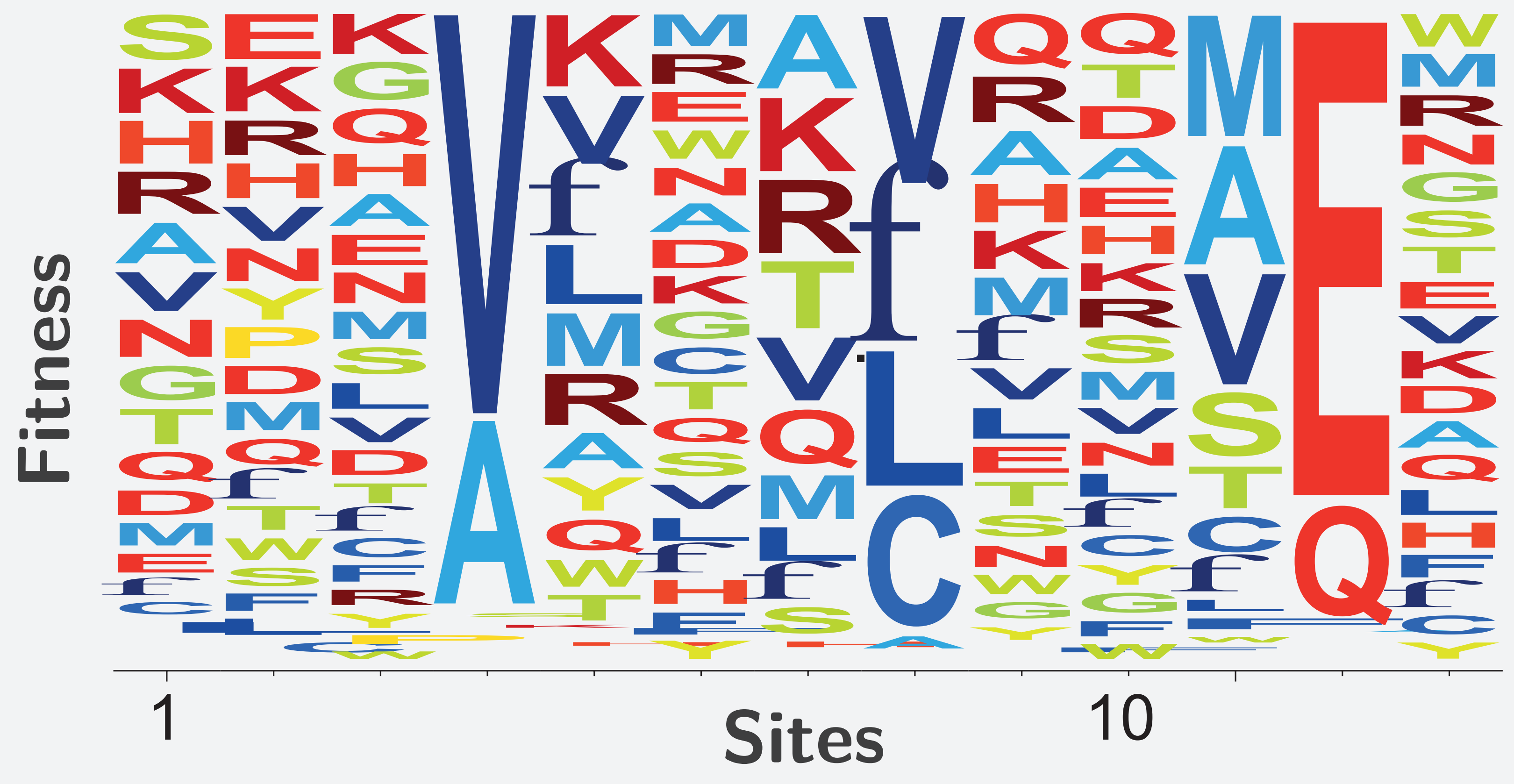
Mutation-selection phylogenetic codon models



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 (2016); Bloom (2016); Rodrigue *et al* (2020).

- **Detecting convergent evolution.**

→ Parto & Lartillot (2017).

CC-BY-SA 2020

Substitutions are the result of the interplay between:

- **Mutations (creation of new variants)**
- **Selection (filtering variants)**
- **Genetic drift (amount of randomness)**

Can mutation, selection and genetic drift be disentangled with phylogenetic codon models?

Part I.
Can ω-based codon models disentangle mutation and selection?

Simulations | ω-based codon models | Empirical analyses

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

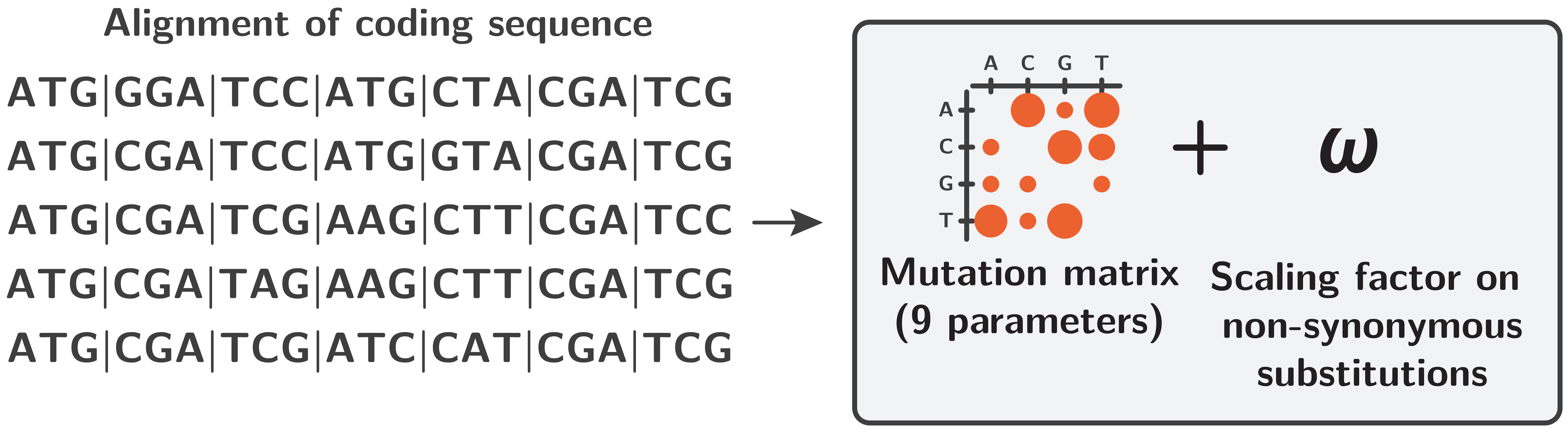
Simulations | Mutation-selection codon models | Empirical analyses

Part III.
Can the relationship between ω and N_e be derived generally at mutation-selection balance?

Simulations | Theory | Protein thermodynamic stability

Part I.
Can ω-based codon models disentangle mutation and selection?

Mutation and selection are modelled separately in *ω*-based codon models



Alignment of coding sequence

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

Mutation matrix (9 parameters)

Scaling factor on non-synonymous substitutions

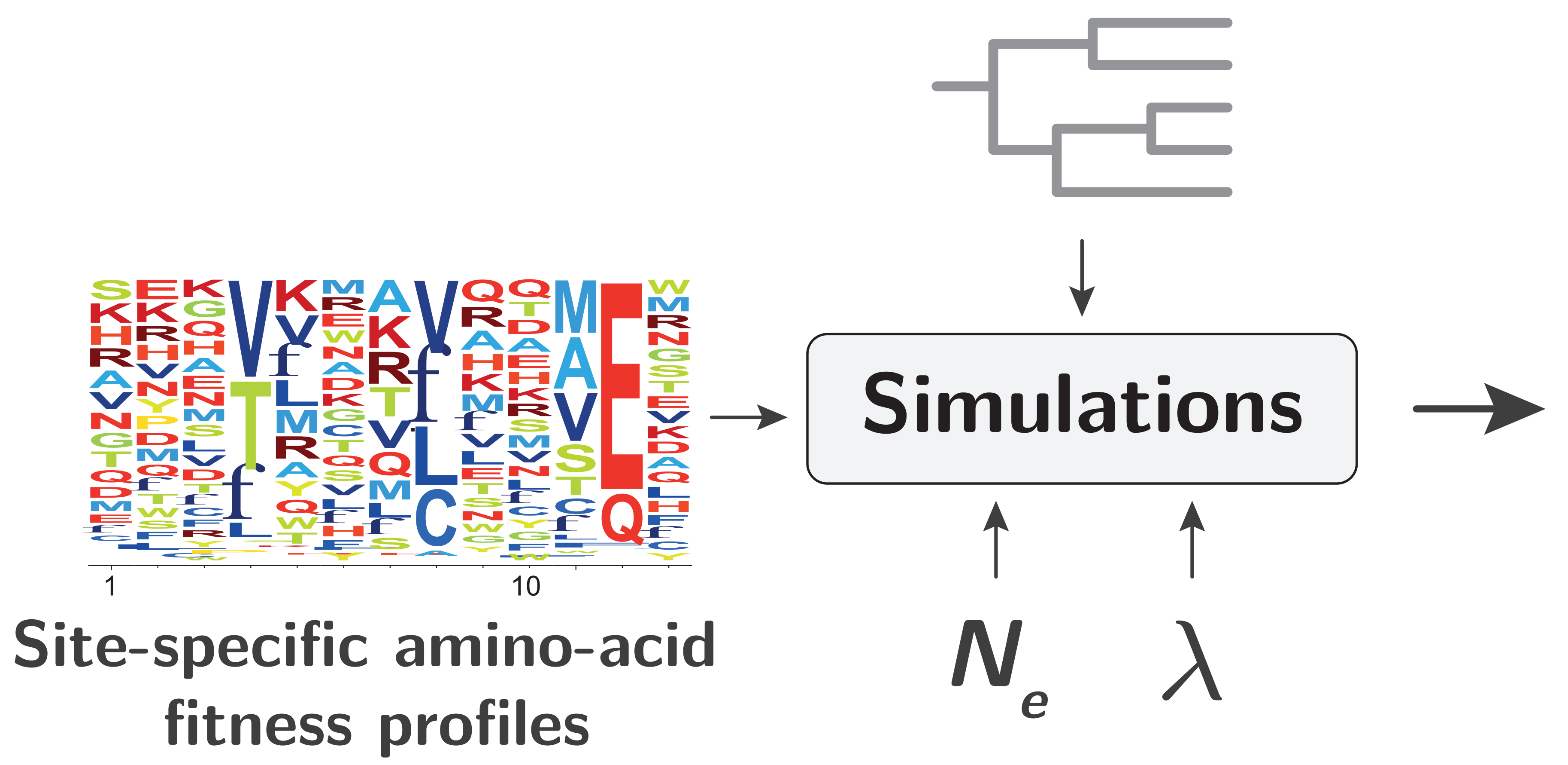
- *ω*-based codon models estimates the strength of selection for a given gene, or a given site.

- These models seek to capture mutation at the level of nucleotide and selection at the level of amino-acids.

- Can *ω*-based codon models disentangle mutation and selection?

Goldman & Yang (1994); Muse & Gaut (1994); Singler & Hickey (2008); Rodrigue *et al* (2008).

Observed bias in the nucleotide composition is weaker than the underlying mutational bias



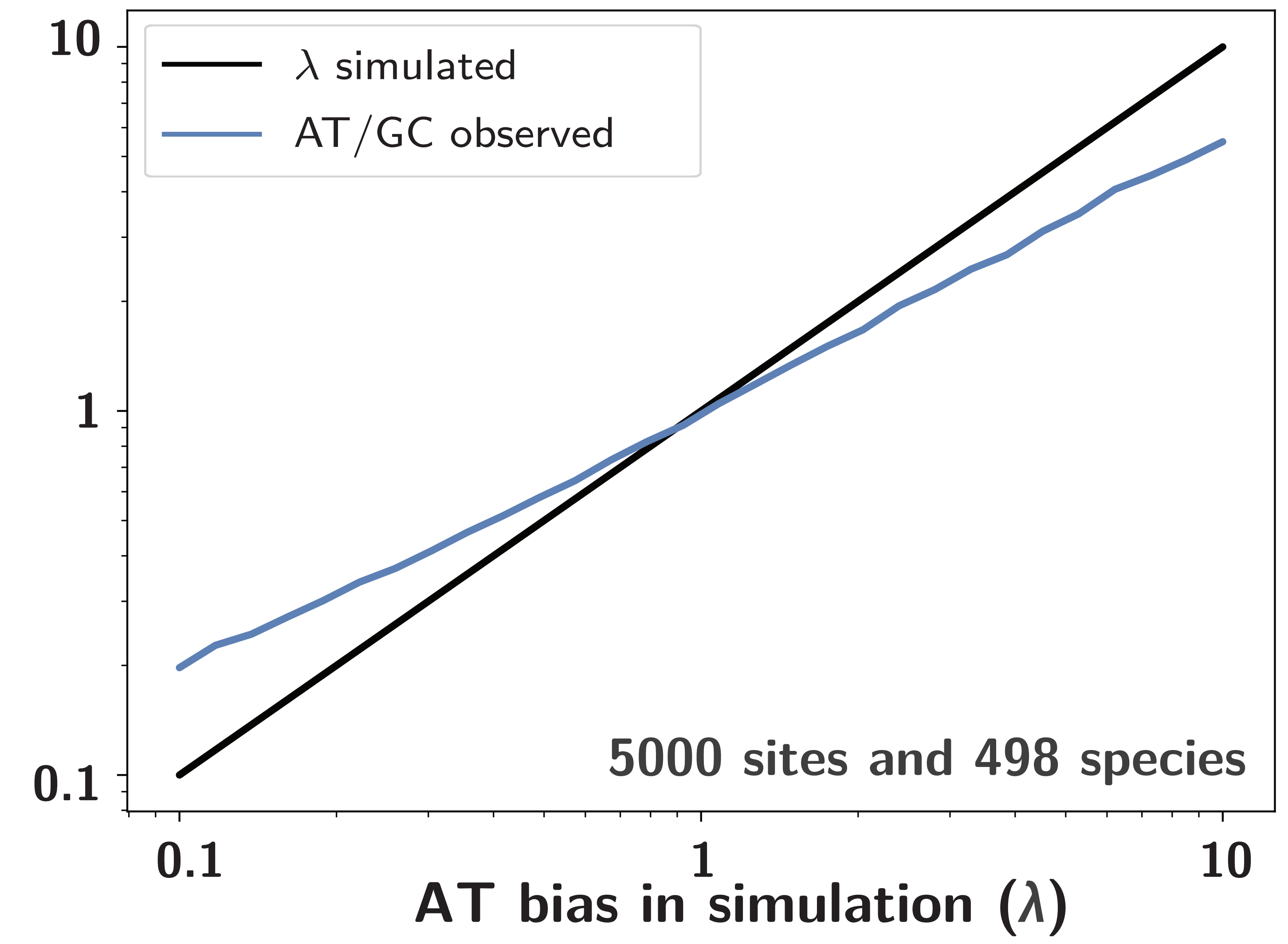
Site-specific amino-acid fitness profiles

Simulations

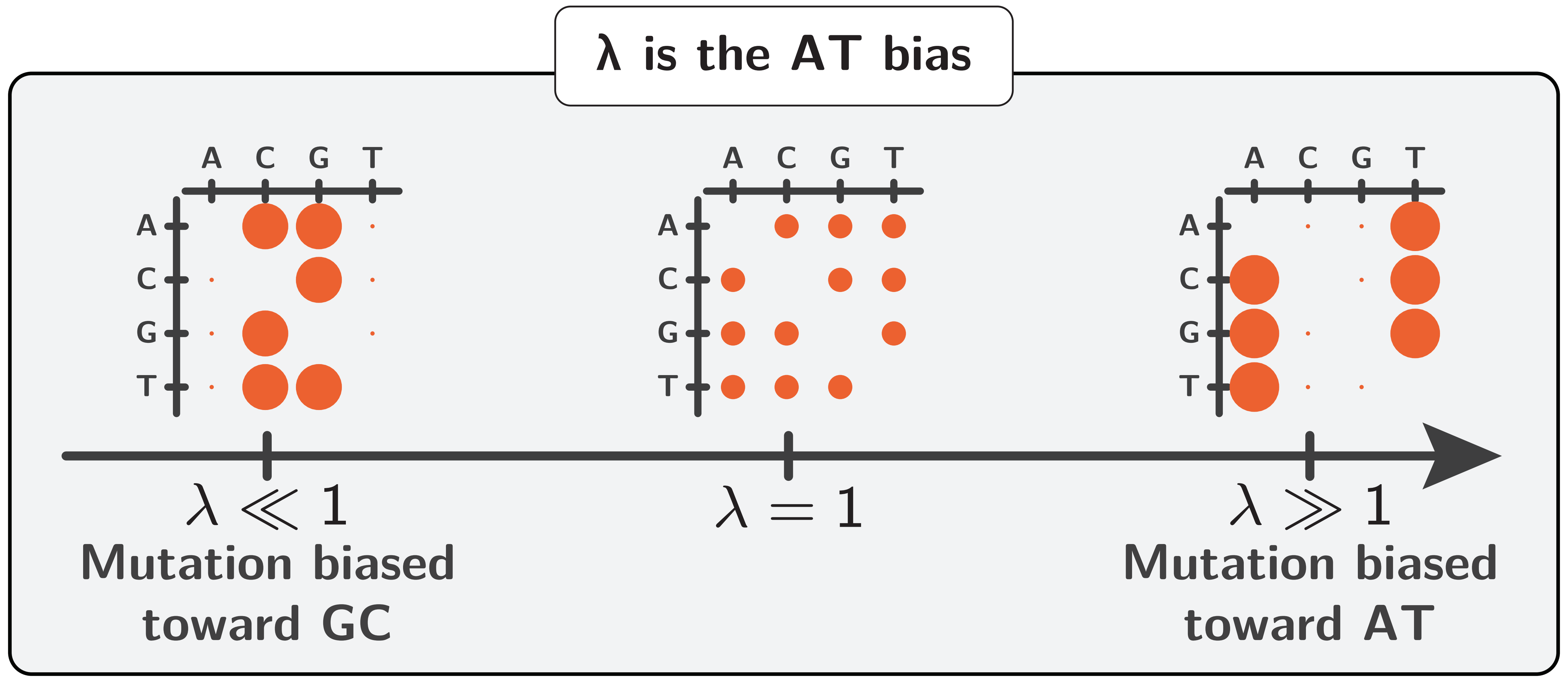
N_e λ

Alignment of coding sequence

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



5000 sites and 498 species



λ is the AT bias

$\lambda \ll 1$
Mutation biased toward GC

$\lambda = 1$

$\lambda \gg 1$
Mutation biased toward AT

https://github.com/ThibaultLatrille/NucleotideBias

ω-based codon models do not reliably estimate the mutational bias

Site-specific amino-acid fitness profiles

Simulations

N_e λ

Alignment of coding sequence

```
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
```

λ is the AT bias

$\lambda \ll 1$ Mutation biased toward GC

$\lambda = 1$

$\lambda \gg 1$ Mutation biased toward AT

Muse & Gaut model

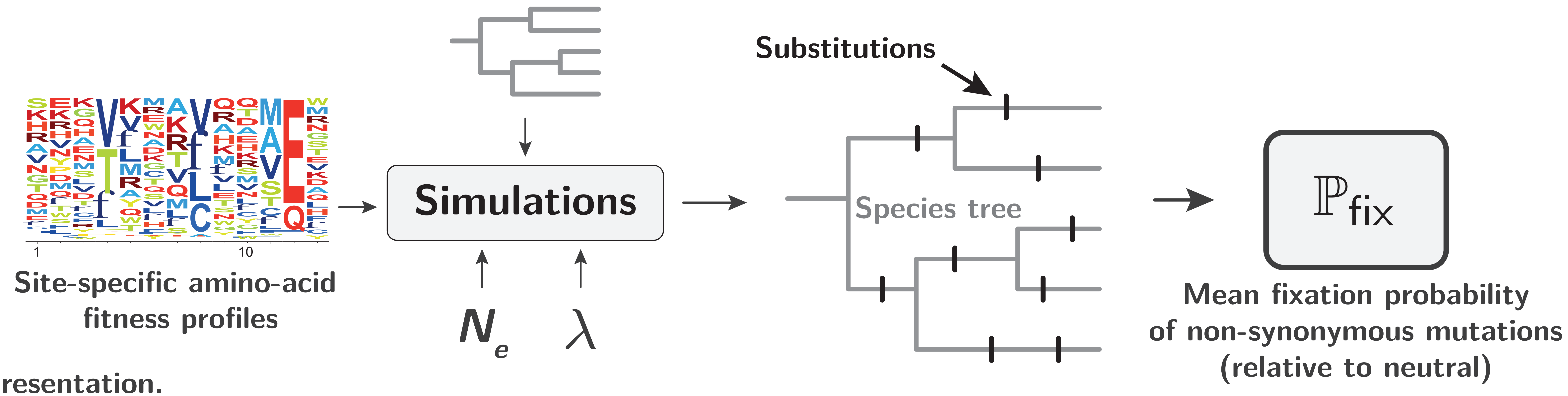
Mutation matrix (9 parameters) + ω Scaling factor on non-synonymous substitutions

$\widehat{\lambda}$

5000 sites and 498 species

https://github.com/ThibaultLatrille/NucleotideBias

Selection is opposed to the mutational bias



Site-specific amino-acid fitness profiles

Simulations

N_e λ

Substitutions

Species tree

$\mathbb{P}_{\text{fix}}$

Mean fixation probability of non-synonymous mutations (relative to neutral)

Cartoon representation.

Simulations, 5000 sites and 498 species.

$\mathbb{P}_{\text{fix}}$

$\mathbb{P}_{\text{fix}}(\text{AT} \rightarrow \text{GC})$

$\mathbb{P}_{\text{fix}}(\text{AT} \rightarrow \text{GC}) / \mathbb{P}_{\text{fix}}(\text{GC} \rightarrow \text{AT})$

1.0
0.8
0.6
0.4
0.2
0.0
Low N_e
High N_e
0.1
1
10
Mutation biased toward GC
λ
Mutation biased toward AT
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4

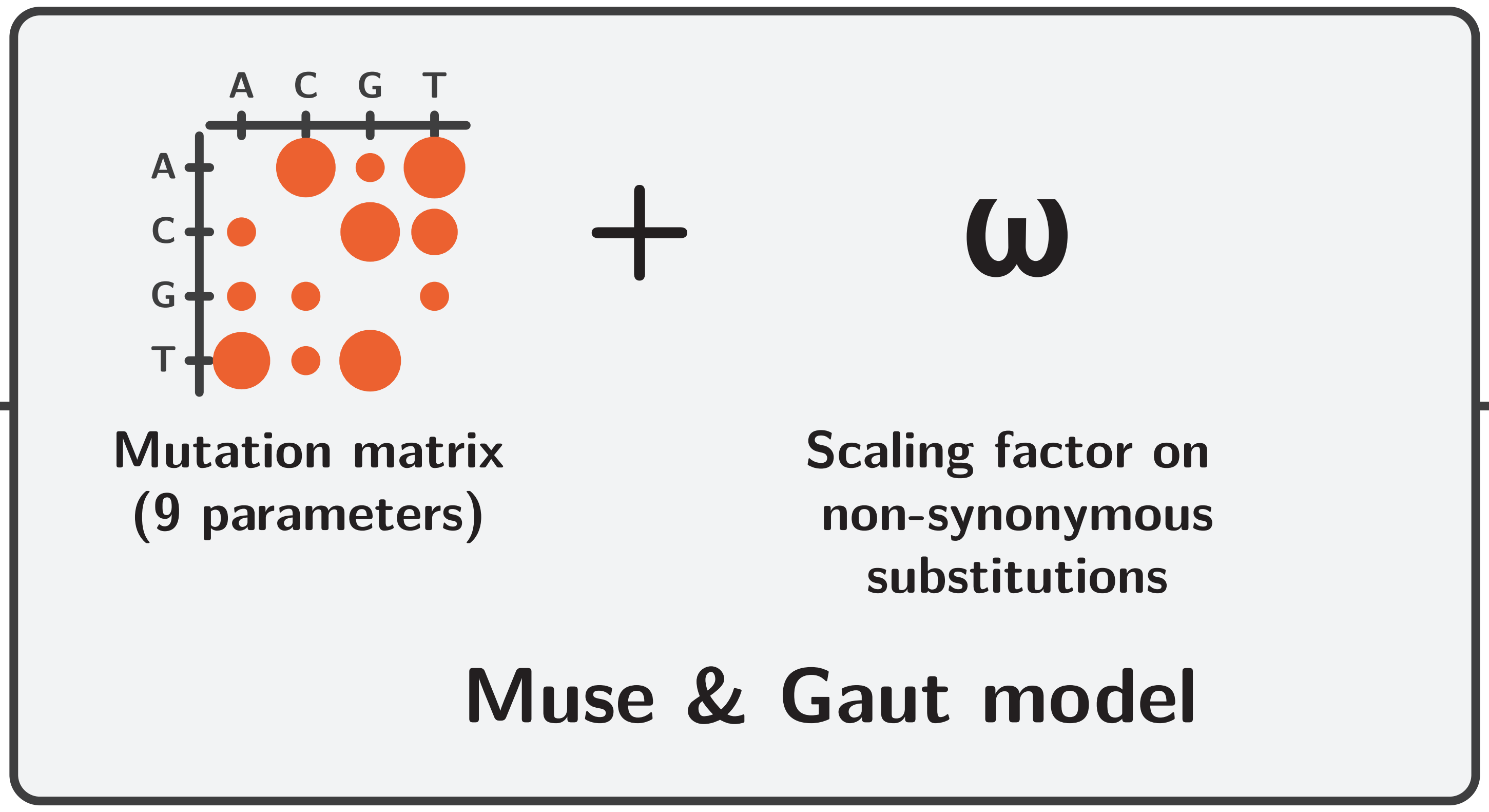
https://github.com/ThibaultLatrille/NucleotideBias

Modelling selection in different directions allows to infer reliably the mutation biases.

Empirical experiments

Influenza Nucleoprotein 498 sites, 180 strains	*E-coli* Lactamase 263 sites, 85 strains
$\widehat{\lambda}$=1.39 $\widehat{\omega}$=0.085	$\widehat{\lambda}$=0.85 $\widehat{\omega}$=0.29

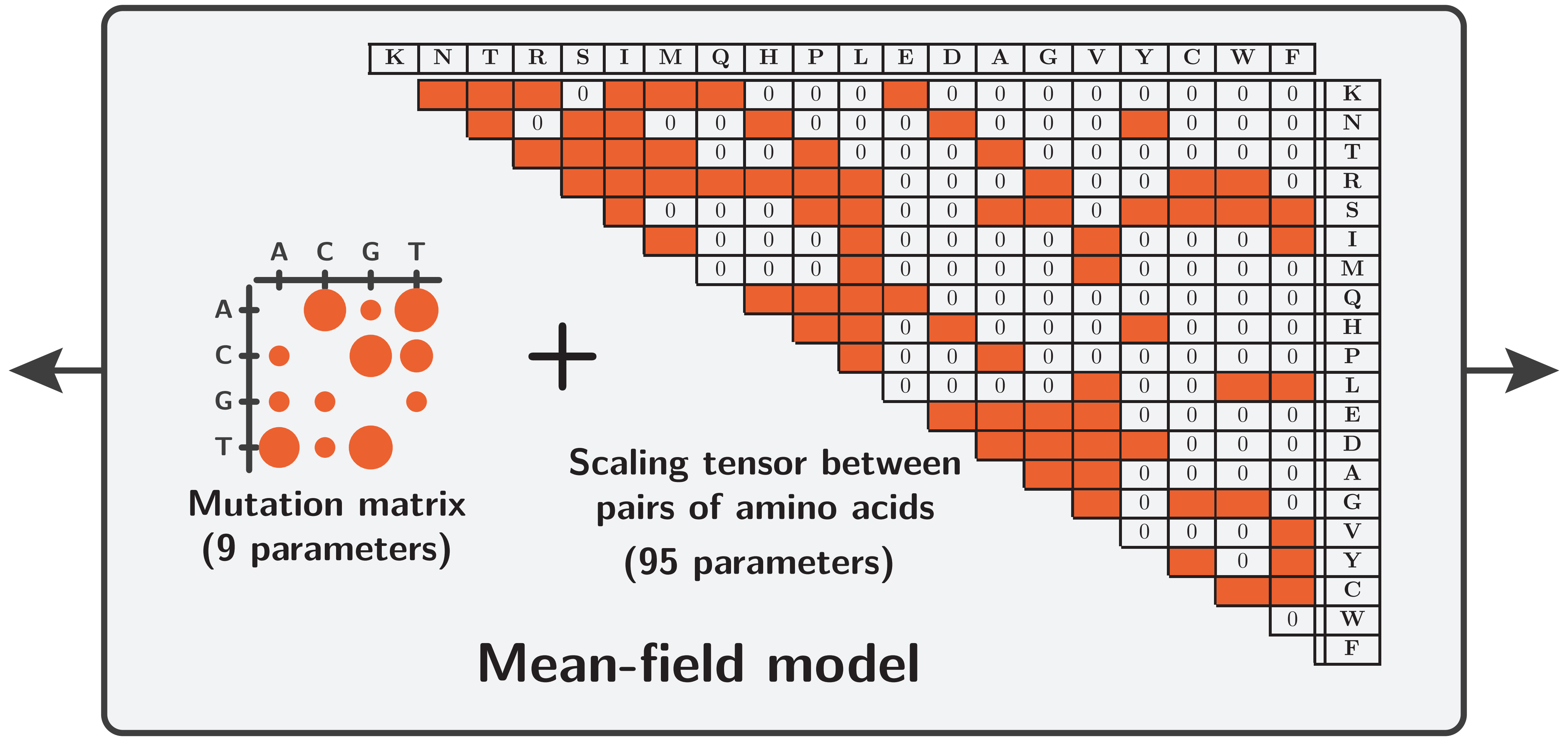
$\widehat{\lambda}$=1.64	$\widehat{\lambda}$=0.68
$\widehat{\omega}$=0.086	$\widehat{\omega}$=0.30
$\widehat{\omega}_{AT\to GC}$=0.14	$\widehat{\omega}_{AT\to GC}$=0.31
$\widehat{\omega}_{GC\to AT}$=0.10	$\widehat{\omega}_{GC\to AT}$=0.44
$\widehat{\omega}_{AT\to GC}/\widehat{\omega}_{GC\to AT}$=1.36	$\widehat{\omega}_{AT\to GC}/\widehat{\omega}_{GC\to AT}$=0.71



Mutation matrix (9 parameters)

Scaling factor on non-synonymous substitutions

Muse & Gaut model

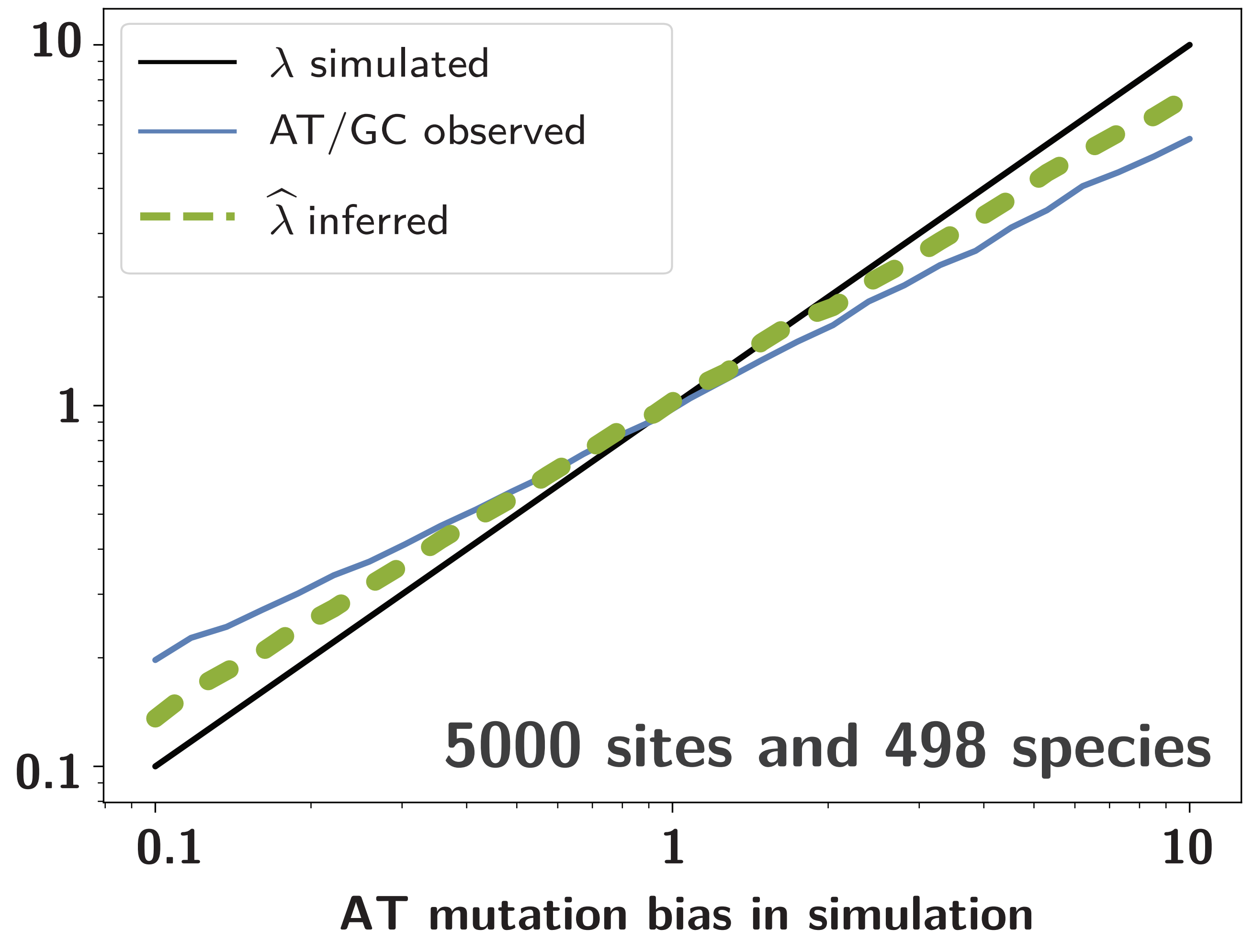


Mutation matrix (9 parameters)

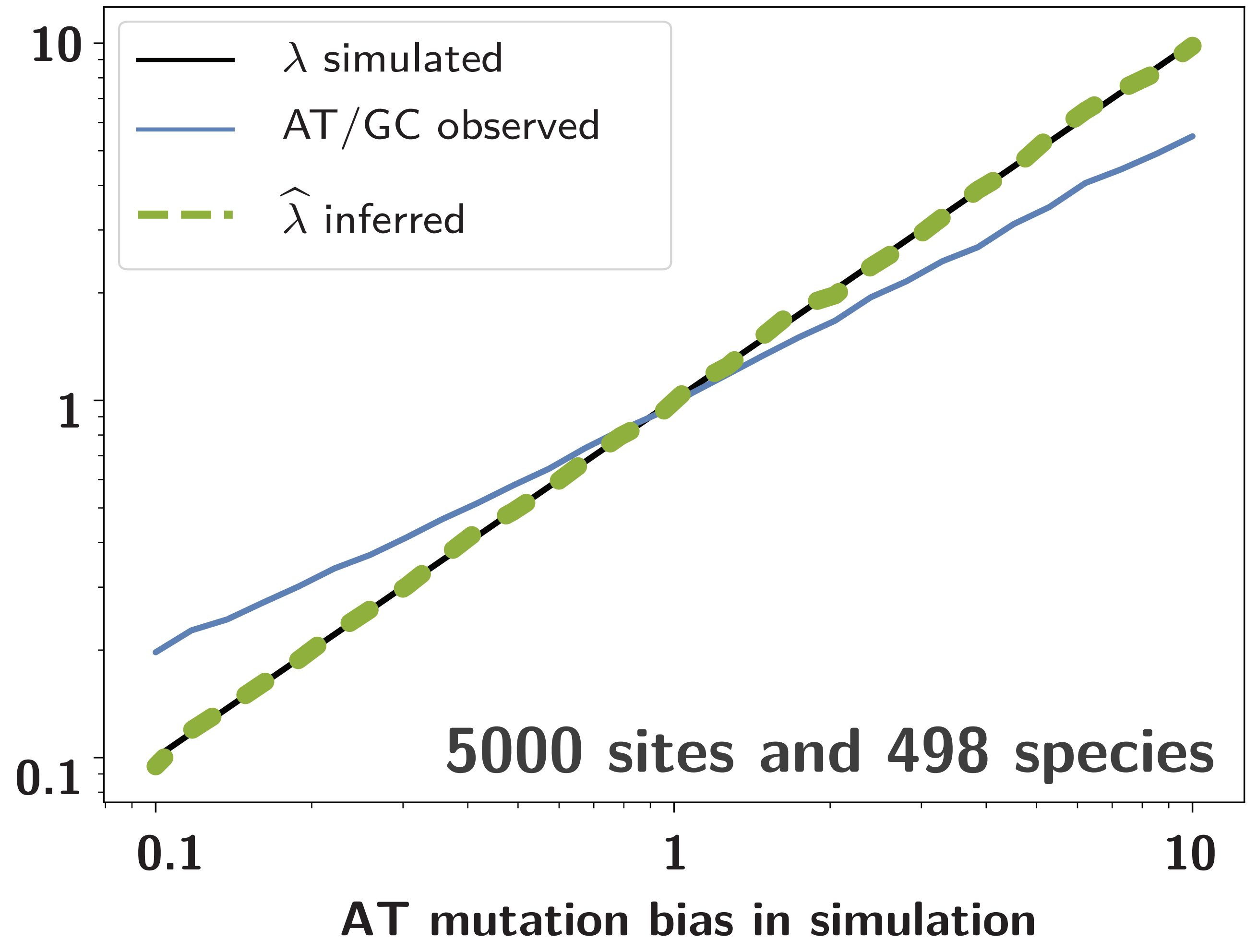
Scaling tensor between pairs of amino acids (95 parameters)

Mean-field model

Simulated experiments



5000 sites and 498 species



5000 sites and 498 species

https://github.com/ThibaultLatrille/NucleotideBias

Can ω-based codon models disentangle mutation and selection?

- ω-based codon models with a single parameter of selection do not reliably estimate mutational biases.

- Mutational bias is balanced by a fixation bias (selection) in the opposite direction.

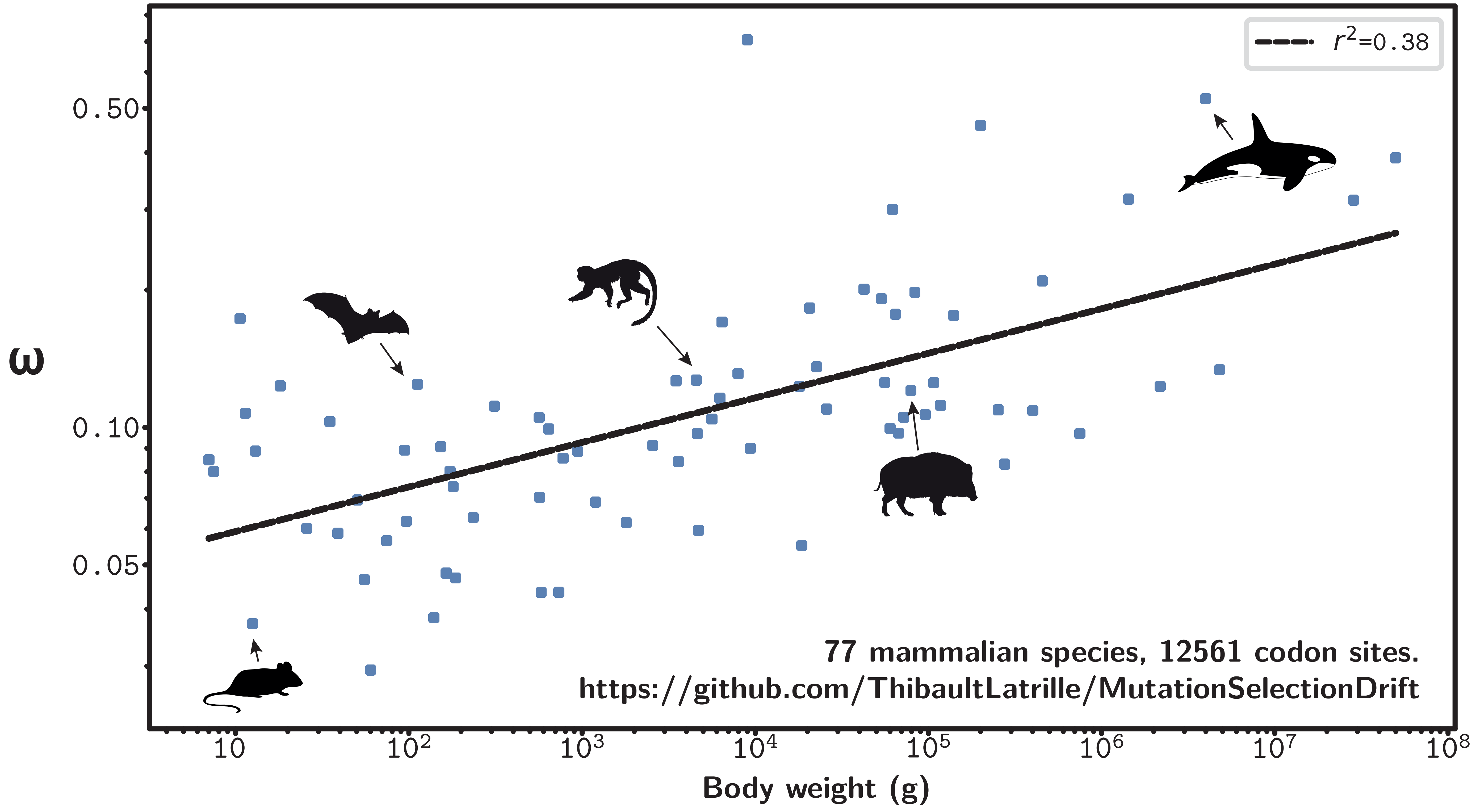
- Inference of mutational bias requires to model fixation bias in different direction.

- Estimation of GC-biased gene conversion requires to disentangle mutation and selection reliably.

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

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.



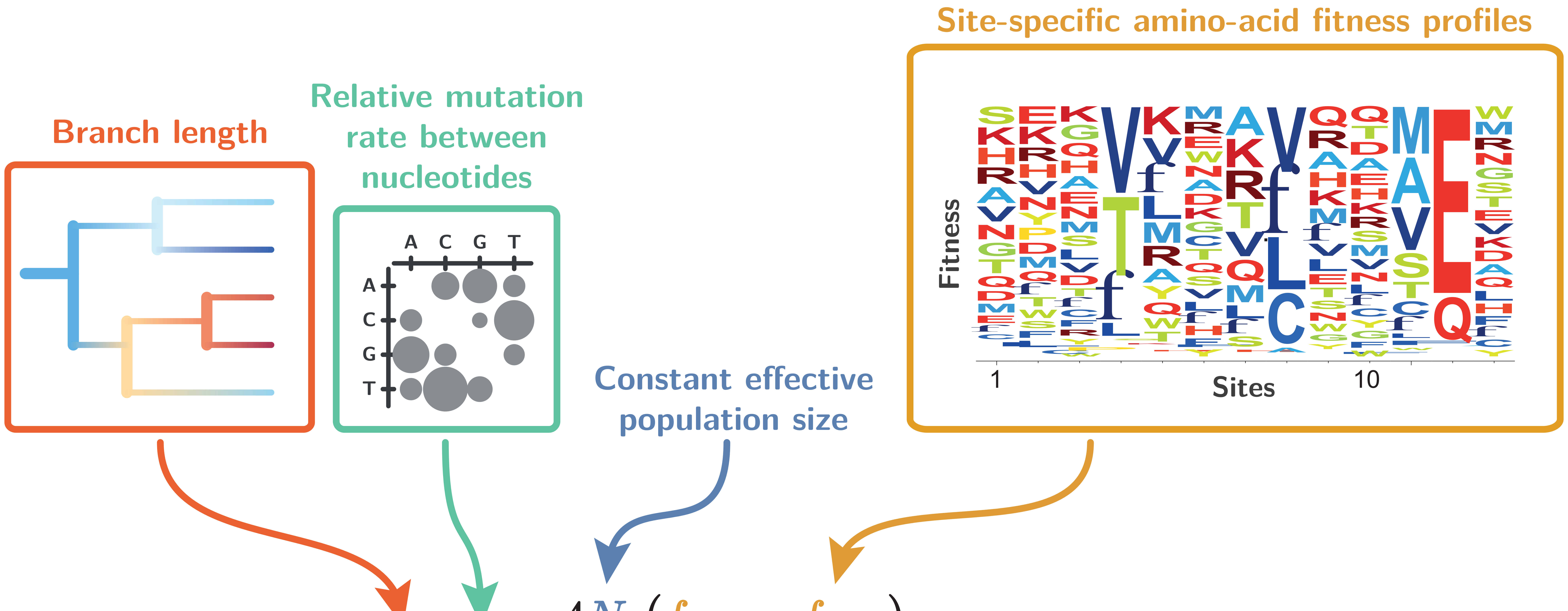
77 mammalian species, 12561 codon sites.
https://github.com/ThibaultLatrille/MutationSelectionDrift

- 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

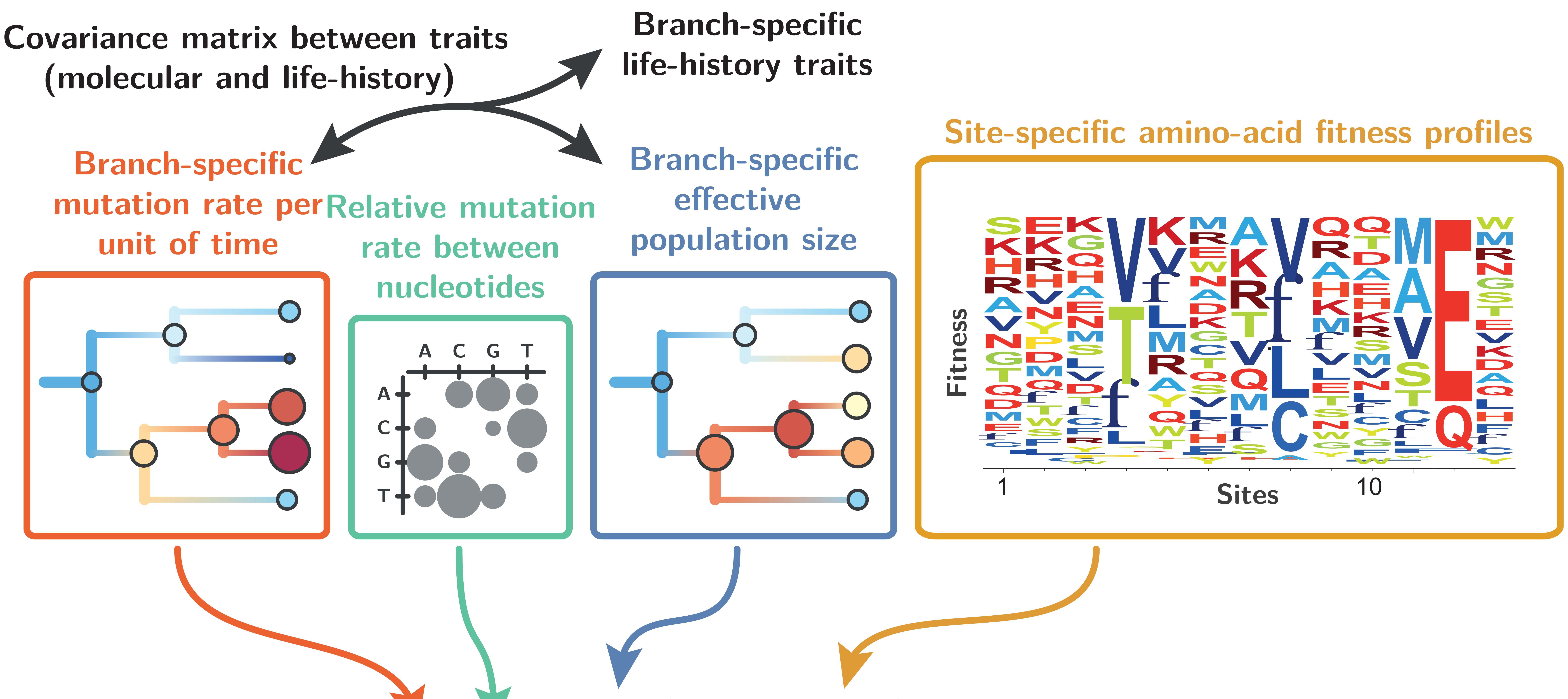
Site-specific amino-acid fitness profiles

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

- **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

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

- **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

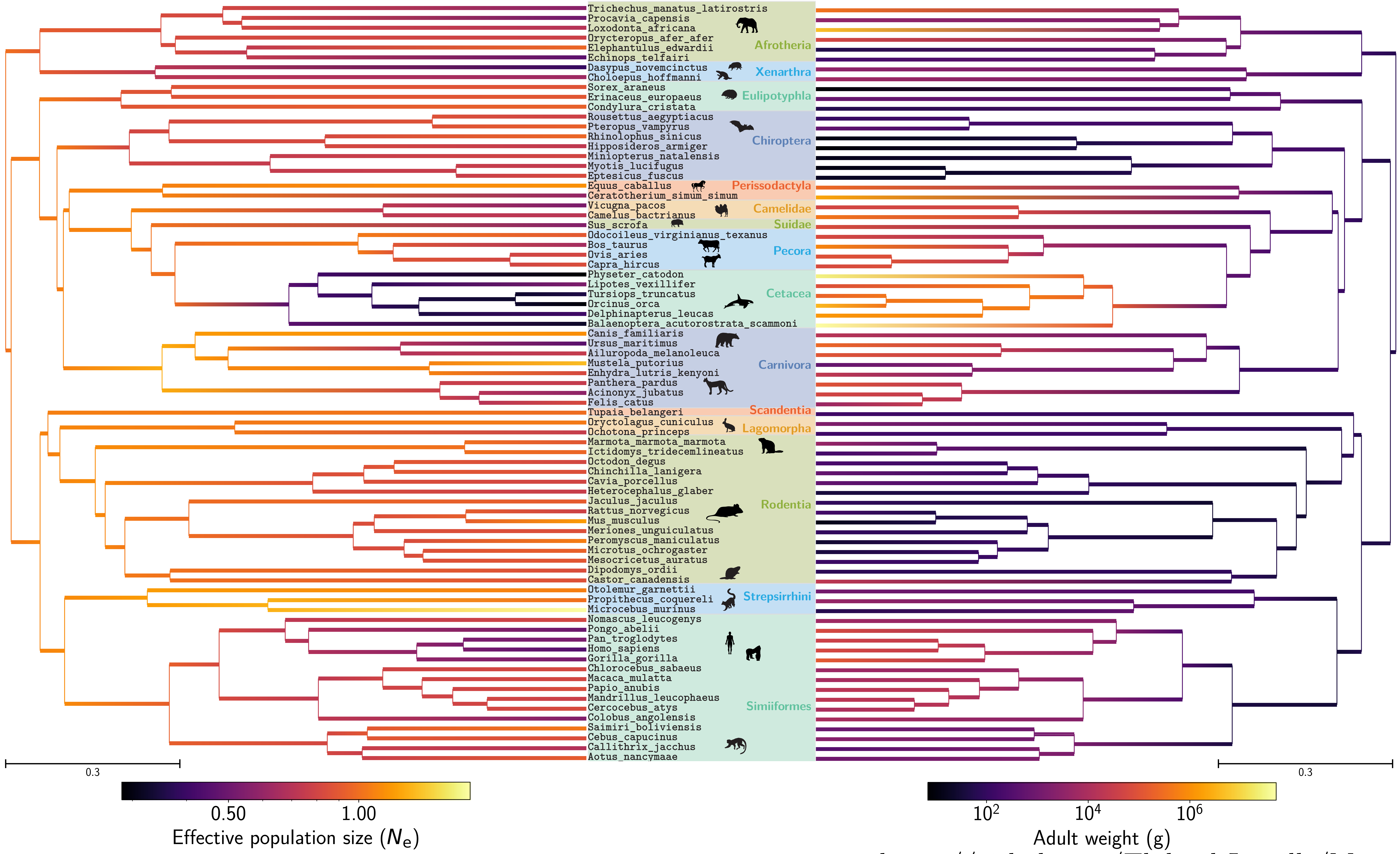
Reconstructed life-history traits, effective population size (N_e), and mutation rate for each branch of the tree.

Site-specific amino-acid fitness profiles

Mutation matrix

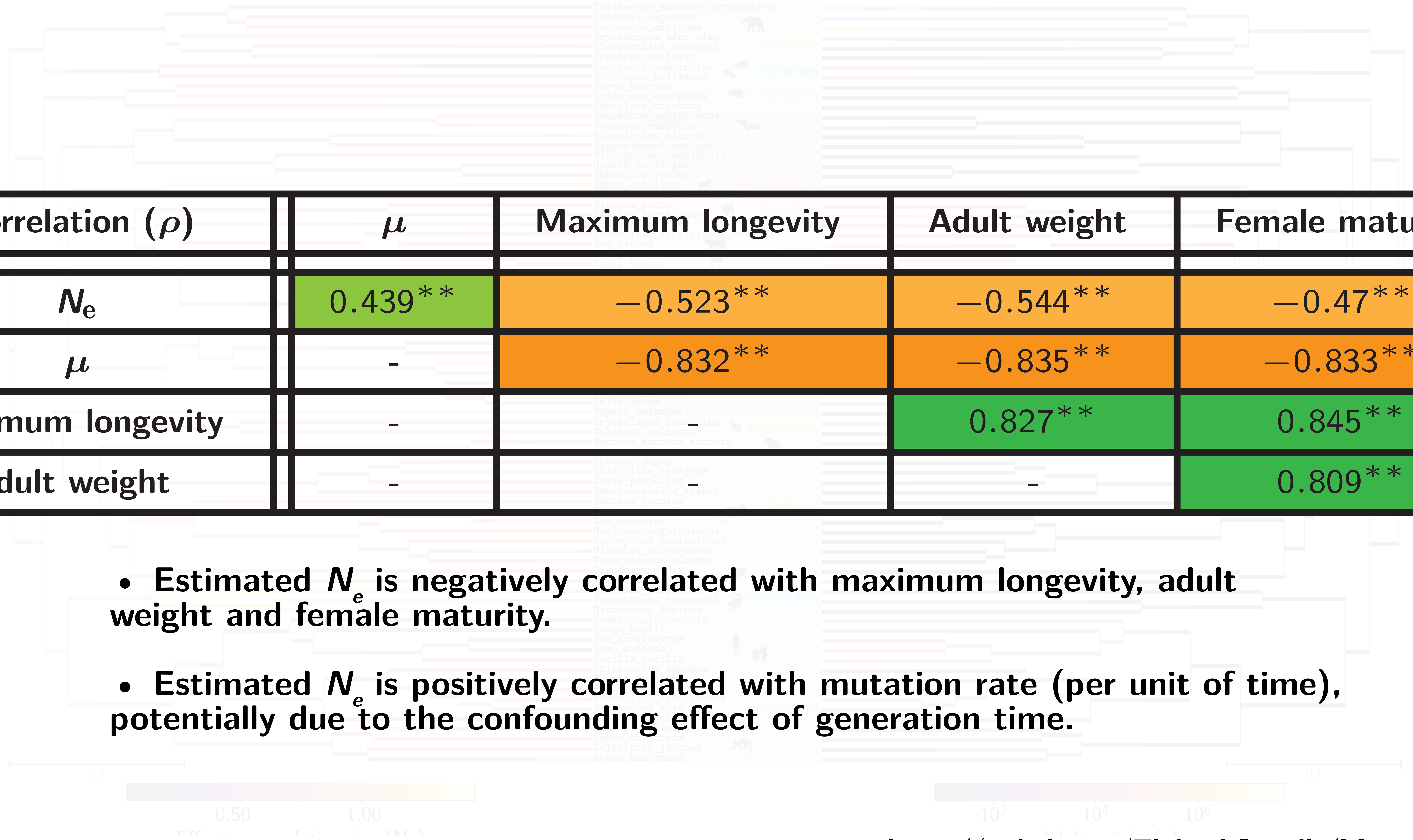
https://github.com/bayesiancook/bayescode

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



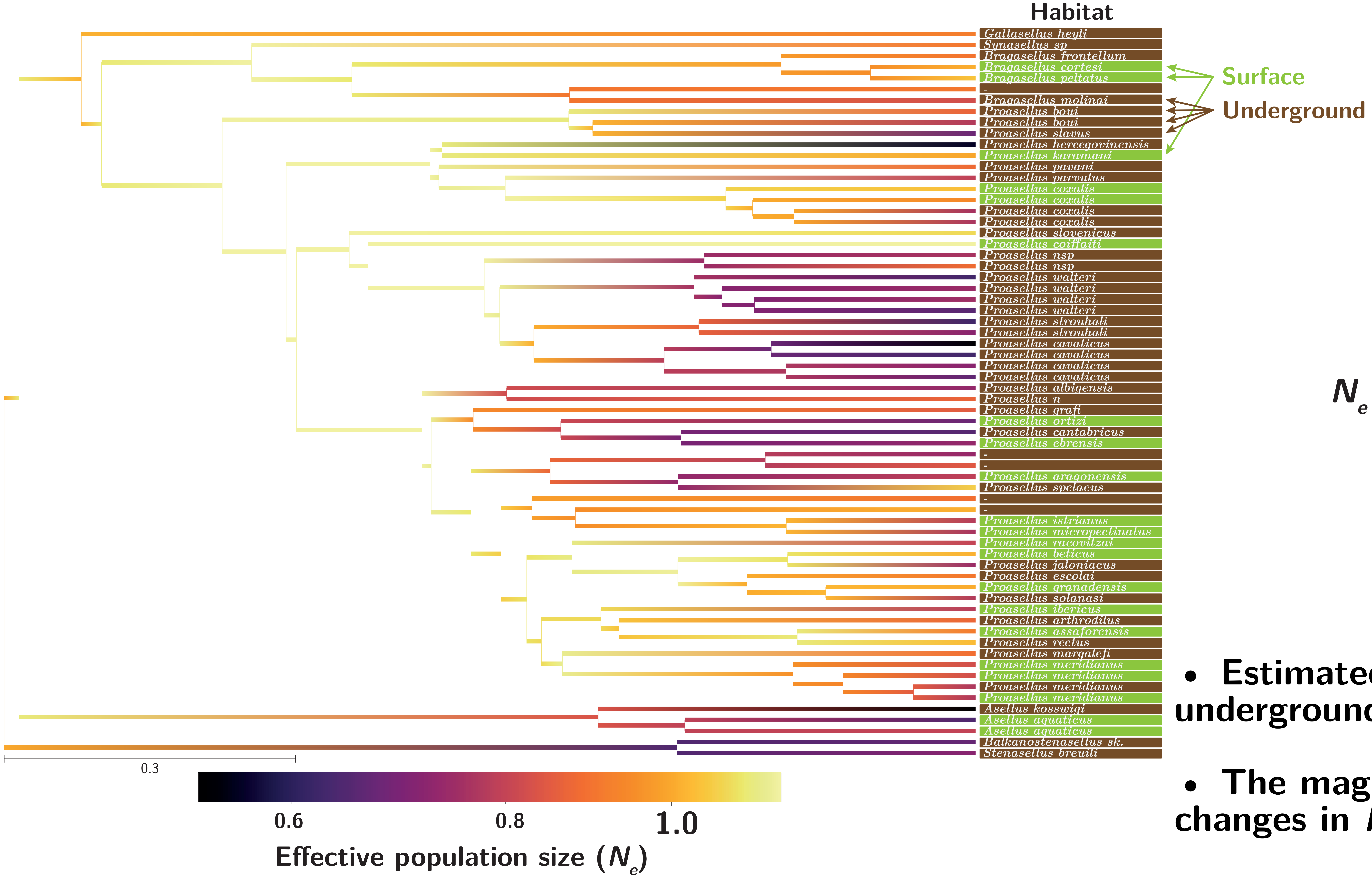
Correlation (ρ)	μ	Maximum longevity	Adult weight	Female maturity
N_e	0.439^{**}	-0.523^{**}	-0.544^{**}	-0.47^{**}
μ	-	-0.832^{**}	-0.835^{**}	-0.833^{**}
Maximum longevity	-	-	0.827^{**}	0.845^{**}
Adult weight	-	-	-	0.809^{**}

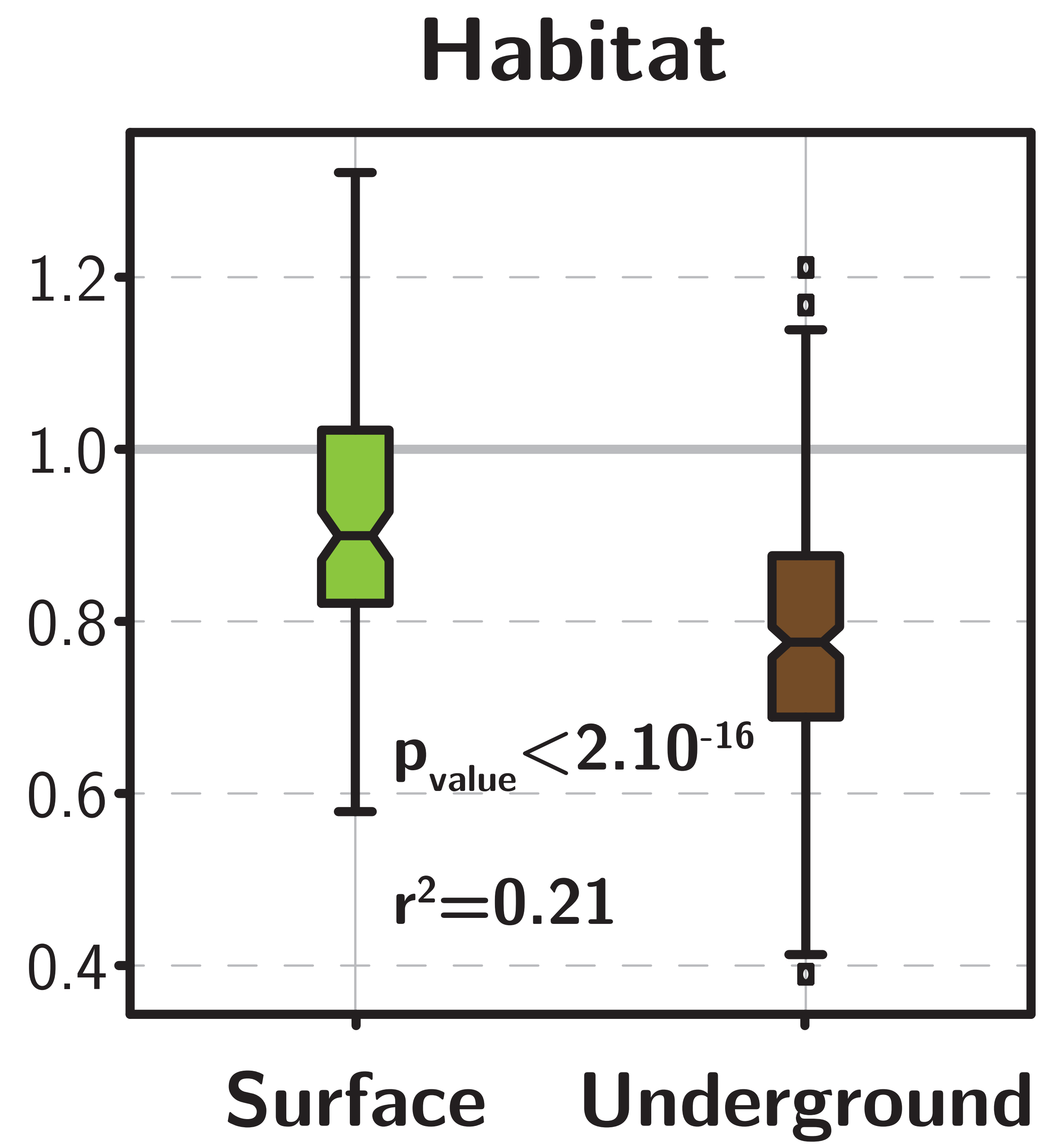
- **Estimated N_e is negatively correlated with maximum longevity, adult weight and female maturity.**

- **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, …

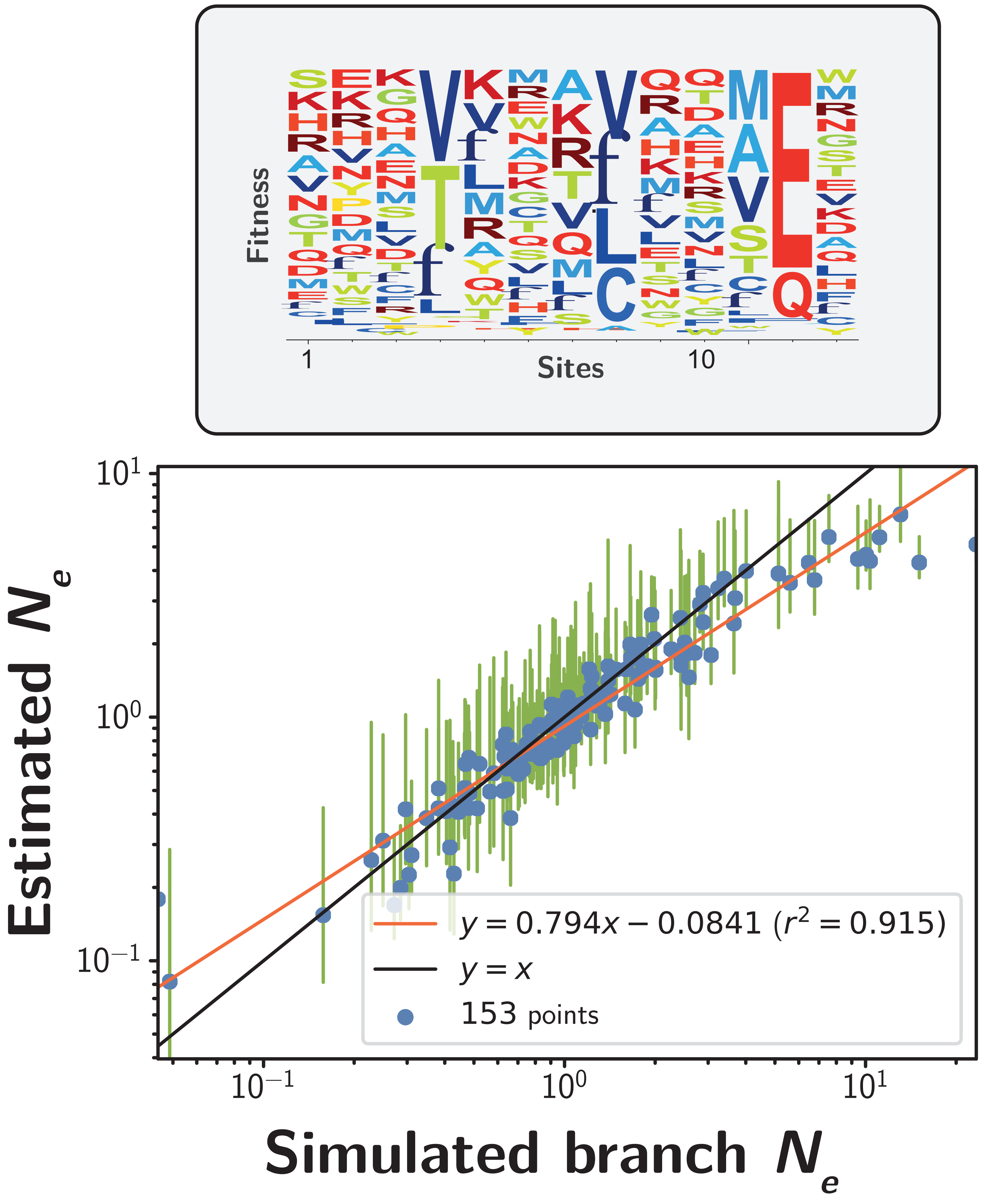
Nucleotide mutation rate matrix

Fluctuation of population size, mutation rate and generation time

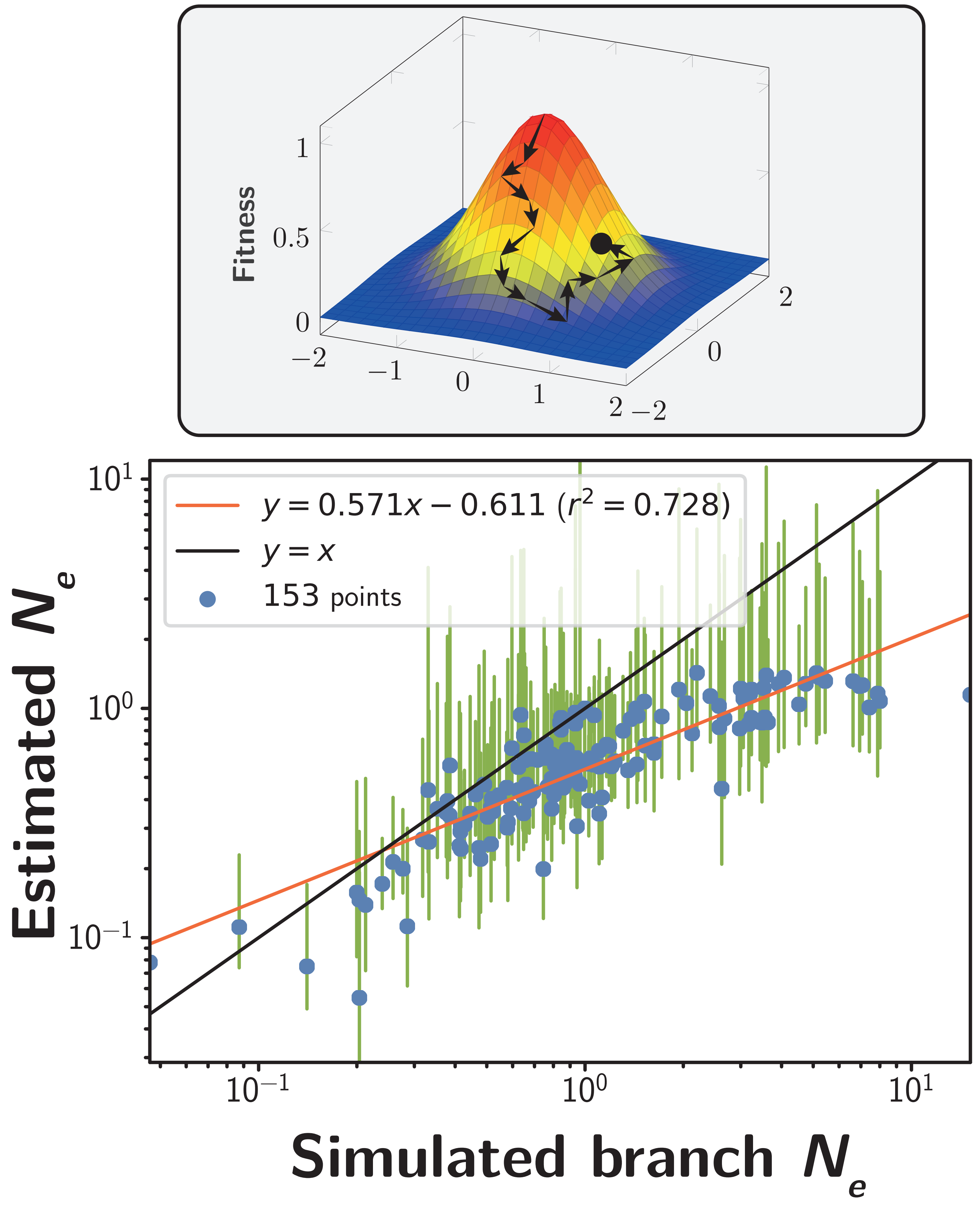
Comparing inference and simulation

N_e cannot be realiably estimated in the presence of epistasis

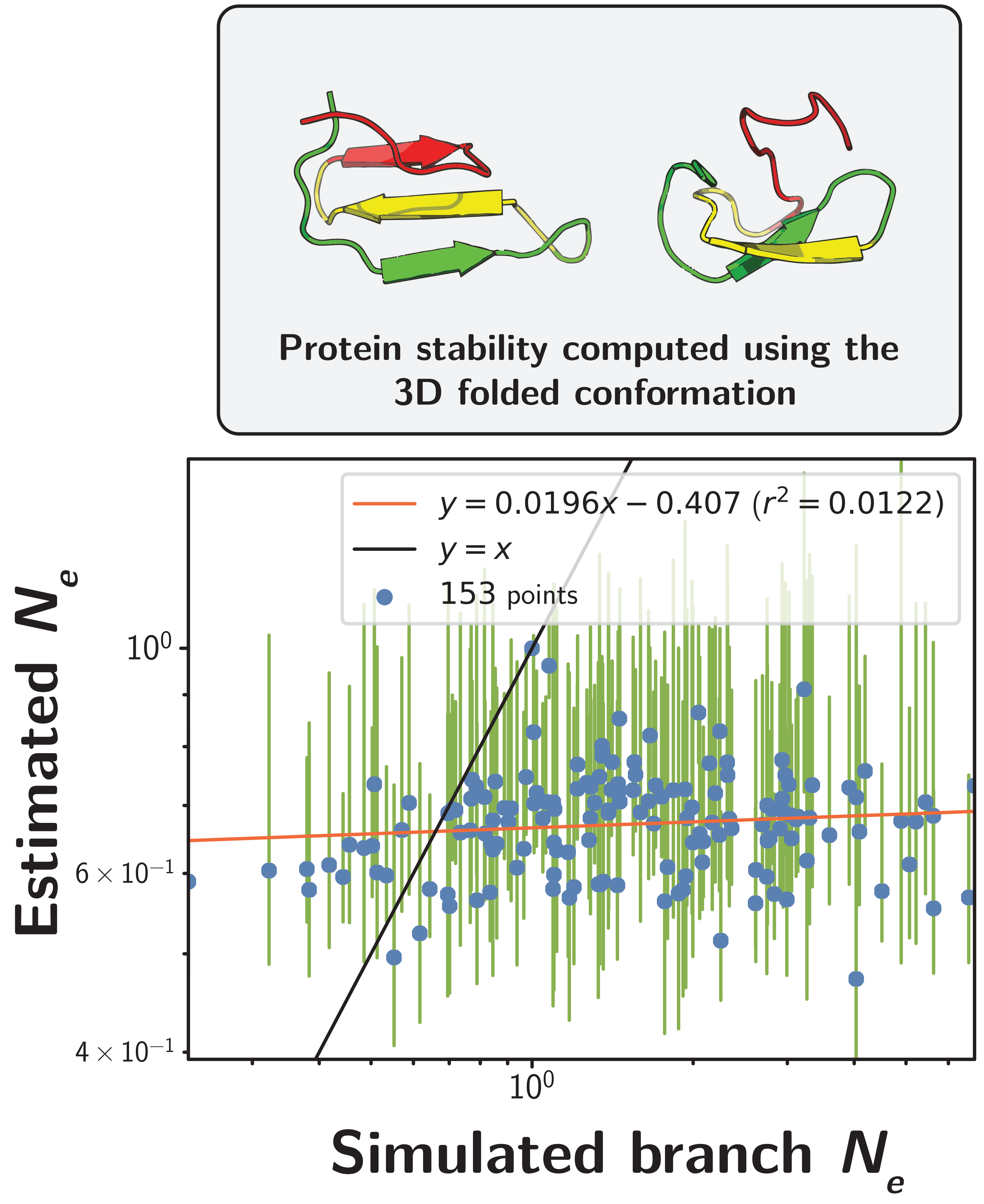
Site-specific amino-acid fitness profiles



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.**

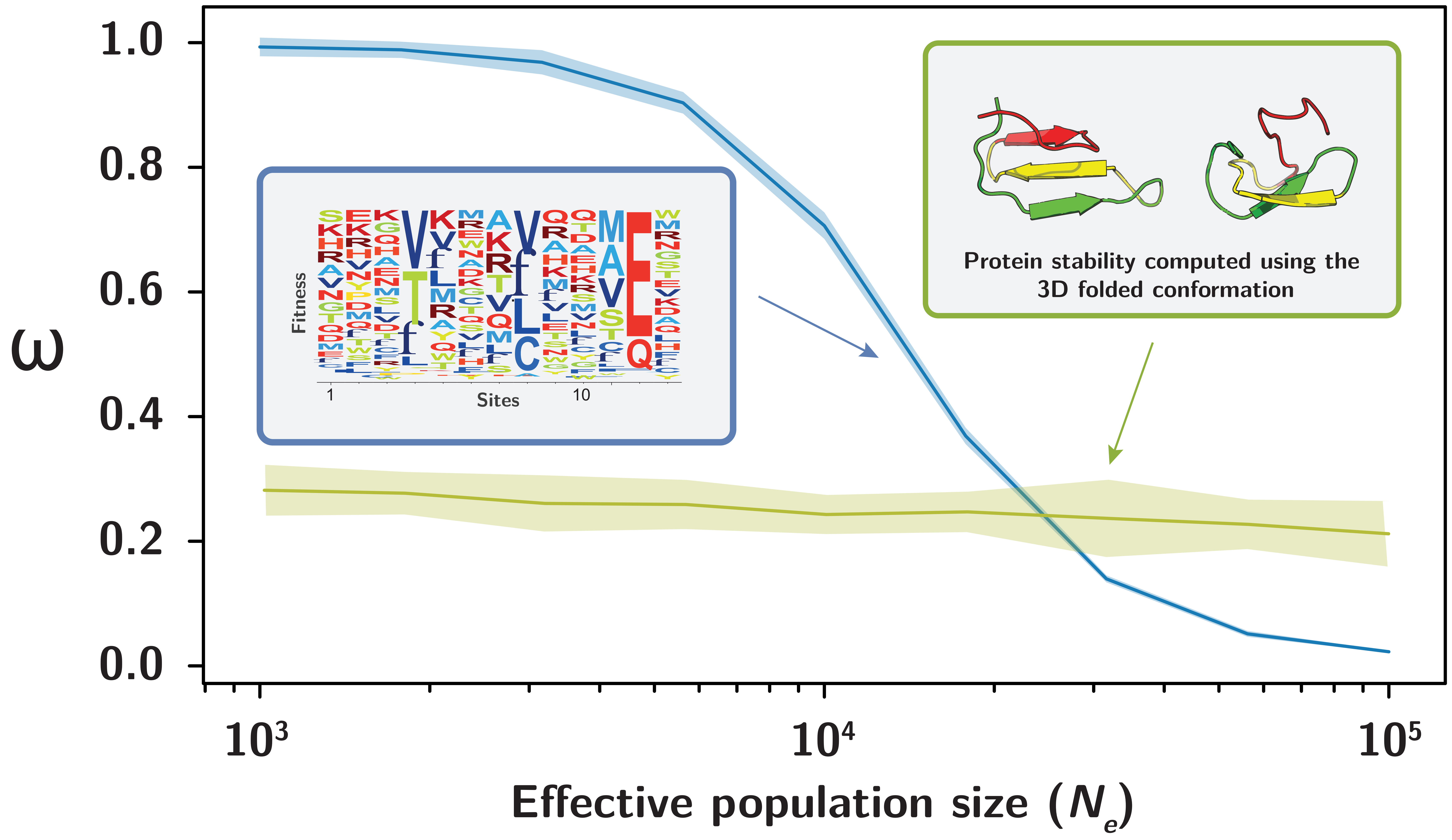
- **The changes in N_e along lineages are in the expected direction, but the range of estimated N_e is lower than expected.**

- **Which mechanism could explain such a low variance of N_e estimated in empirical data?**

- **Epistasis appears to be a reasonable explanation.**

III.
Can the relationship between ω and N_e be derived generally at mutation-selection balance?

Relationship between ω and N_e

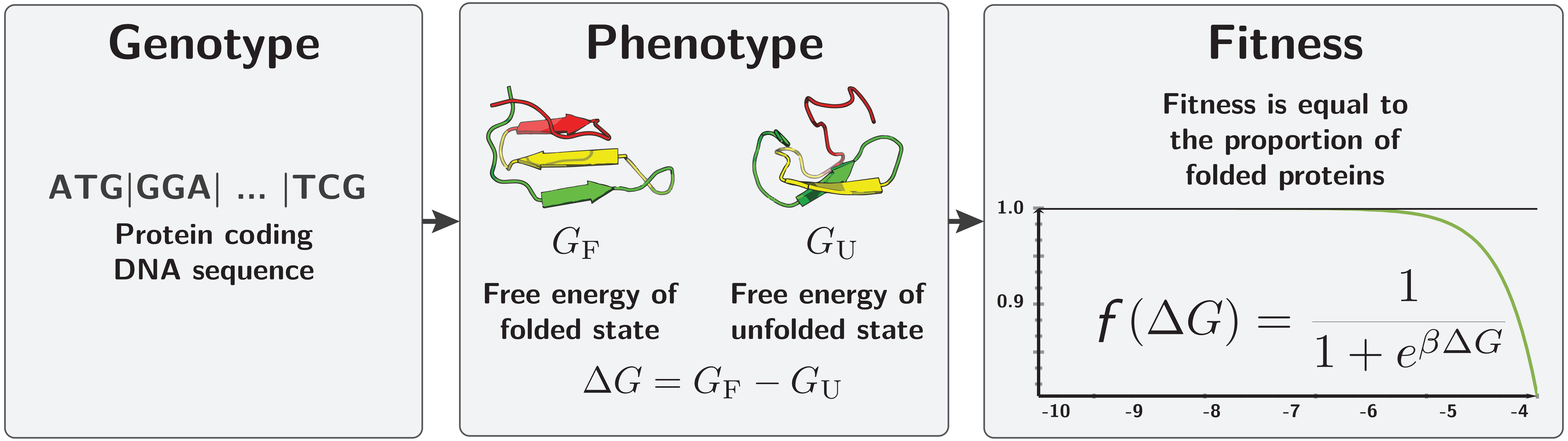


Protein stability computed using the 3D folded conformation

- Can we determine the relationship between ω and N_e in the case of fitness determined by protein stability?

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

Fitness as the proportion of folded proteins



Genotype

ATG|GGA| ... |TCG

Protein coding DNA sequence

Phenotype

G_{F}

Free energy of folded state

G_{U}

Free energy of unfolded state

$$\Delta G = G_{\mathrm{F}} - G_{\mathrm{U}}$$

Fitness

Fitness is equal to the proportion of folded proteins

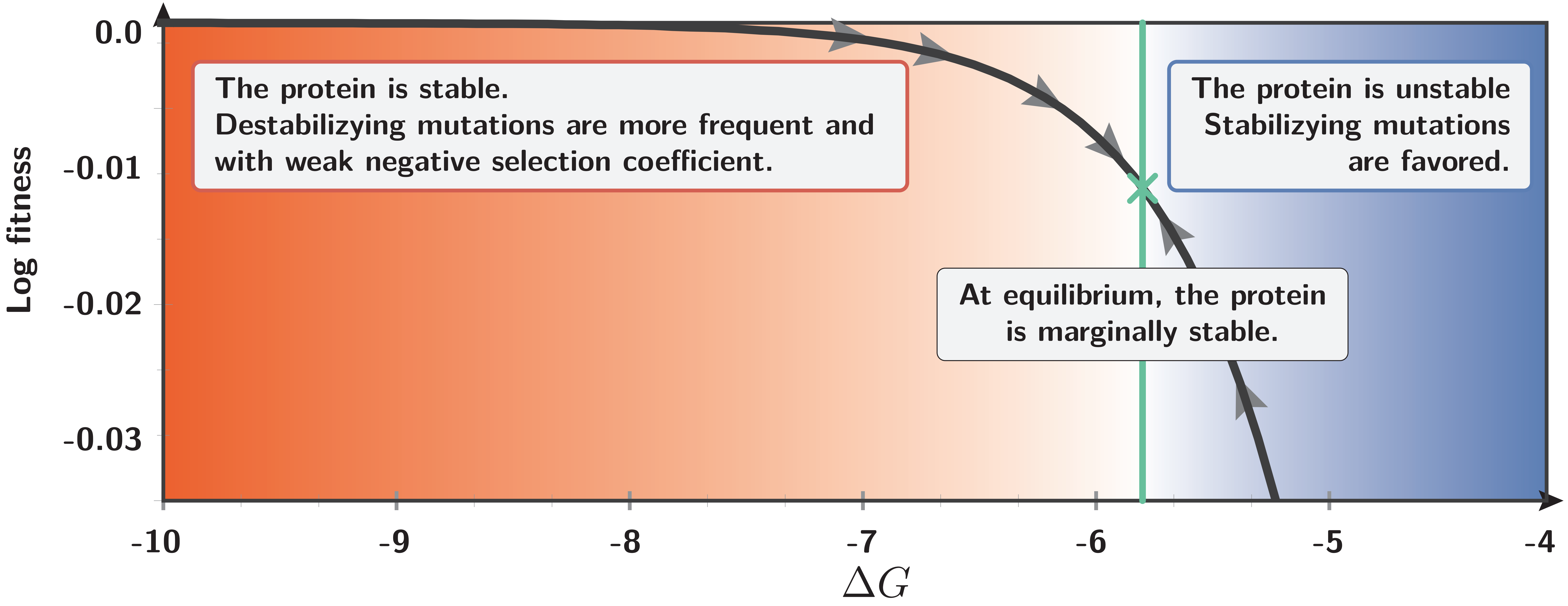
$$f(\Delta G) = \frac{1}{1 + e^{\beta \Delta G}}$$

β is the inverse of the temperature ($\beta = 1/T$)

- **Free energy of is computed using the 3D conformations and pairwise contact potential energies between neighboring amino-acid residues.**

Miyazawa and Jernigan (1985), Williams et al (2006), Goldstein (2011), Pollock et al (2012)

Proteins are marginally stable at mutation-selection balance



The protein is stable.
Destabilizying mutations are more frequent and with weak negative selection coefficient.

The protein is unstable
Stabilizying mutations are favored.

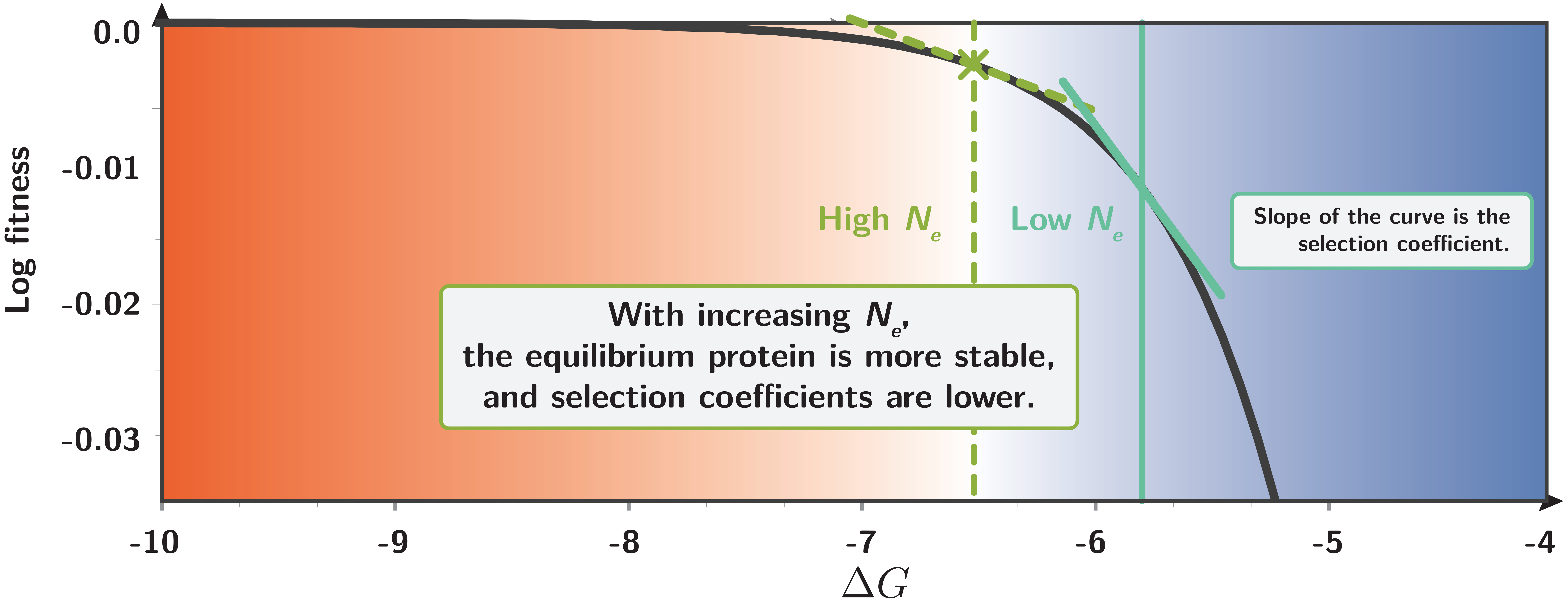
At equilibrium, the protein is marginally stable.

- The optimal stability of proteins is never achieved.

- Marginal stability is the default expectation of the mutation-selection balance even under directional selection for stability.

Taverna & Goldstein (2002

Equilibrium response to a change in N_e



Slope of the curve is the selection coefficient.

With increasing N_e, the equilibrium protein is more stable, and selection coefficients are lower.

- Selection coefficient is dependent on the position in the fitness landscape.

- If the distribution of phenotypic changes is independent of the underlying phenotype, then ω is independent of N_e.

- Can we derive the relationship between N_e and ω as a function of the microscopic molecular parameters of the model?

Cherry (1998); Goldstein (2013).

1D linear model of protein stability

- n is the number of sites in the protein.
- β is the temperature (equals to 1.686 mol/kcal at 25°C).
- $\Delta\Delta G > 0$ (in kcal/mol) is the expected change in free energy (between folded and unfolded states) for a destabilizing mutation.

Genotype

n sites

Stable site

Unstable site

One stable amino acid and 19 unstable amino acids at each site

Phenotype

$$x = \frac{\text{🟧}}{n}$$

Each unstable site add $\Delta\Delta G$ to ΔG

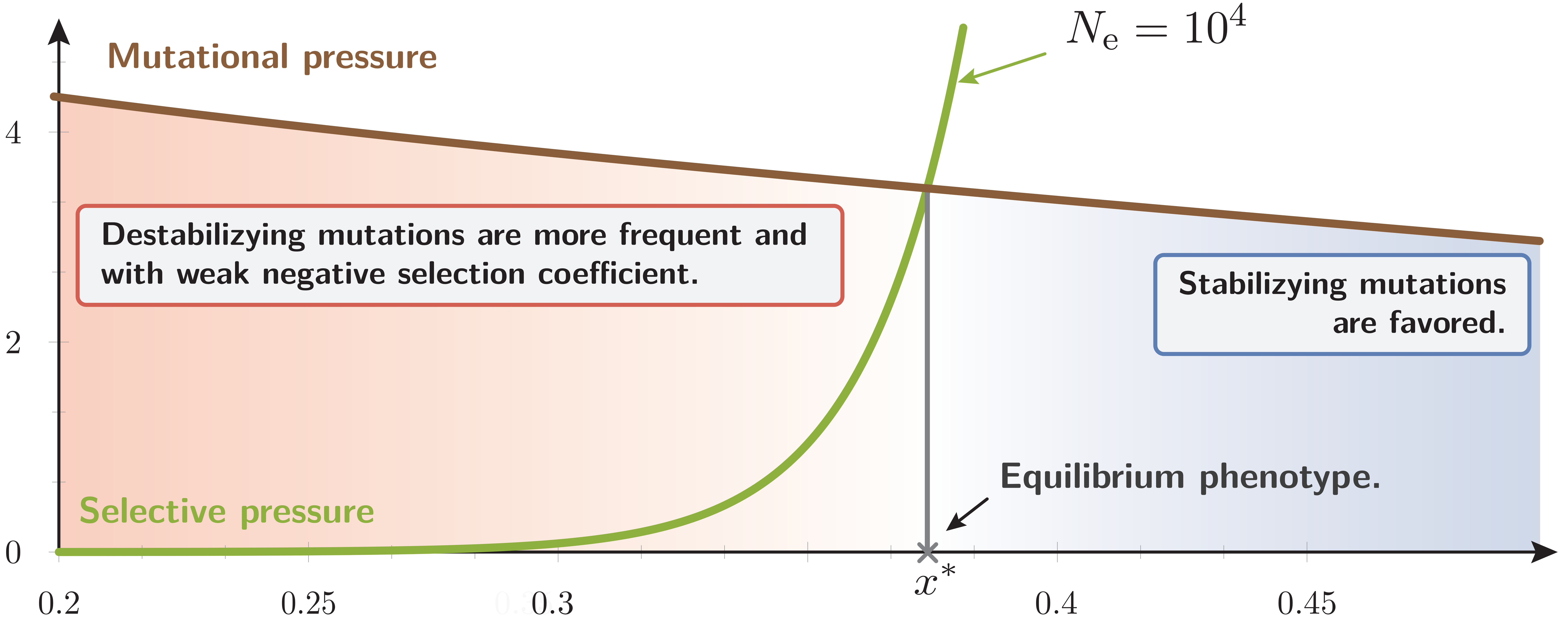
Fitness

$$\frac{1}{1 + e^{\beta(\Delta G_{\text{min}} + xn\Delta\Delta G)}}$$

f(x)
x
0
0.2
0.4
0.6
0.8
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

- **What is the equilibrium phenotype at mutation-selection balance?**
- **What is the resulting ω?**

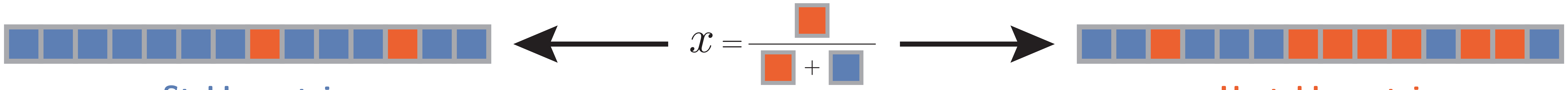
What is the phenotype at equilibrium?



Destabilizing mutations are more frequent and with weak negative selection coefficient.

Stabilizing mutations are favored.

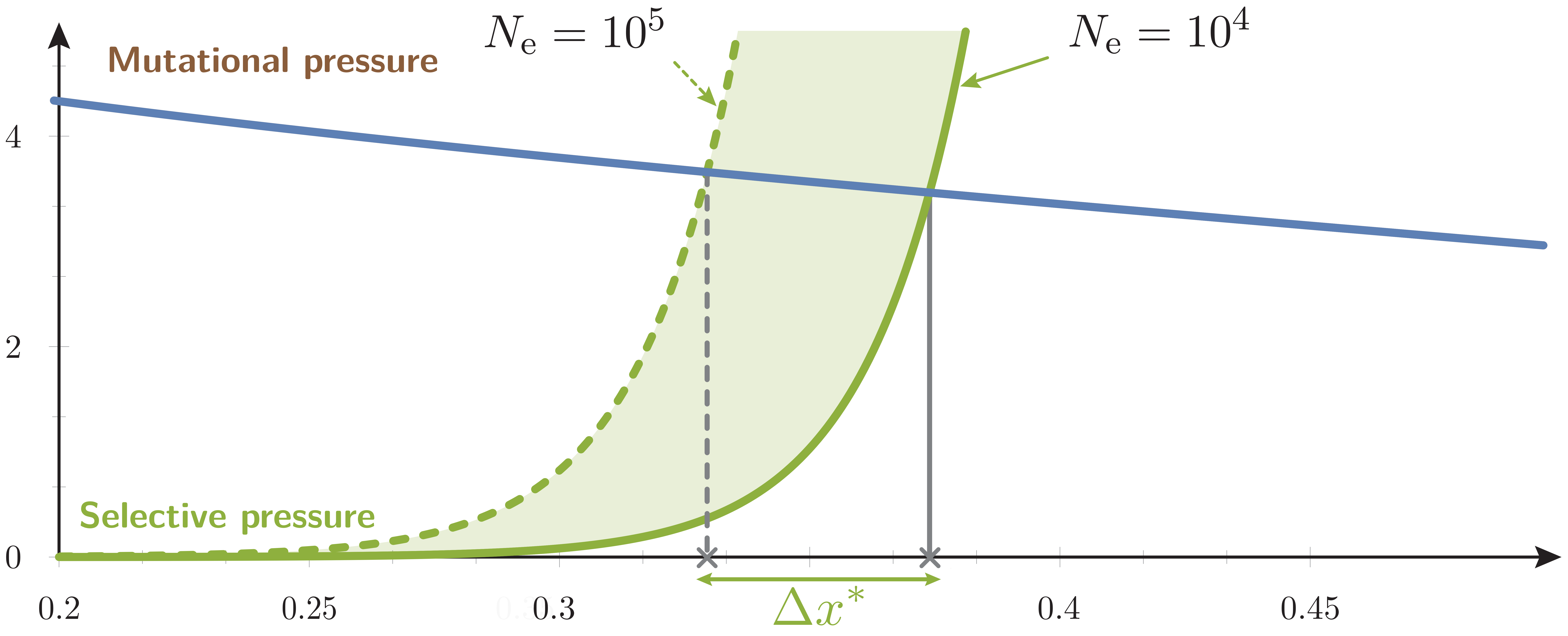
Equilibrium phenotype.

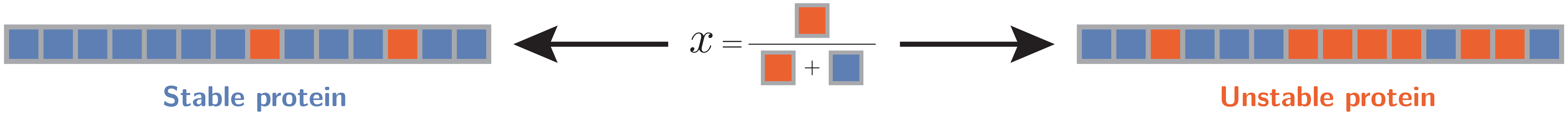


Stable protein

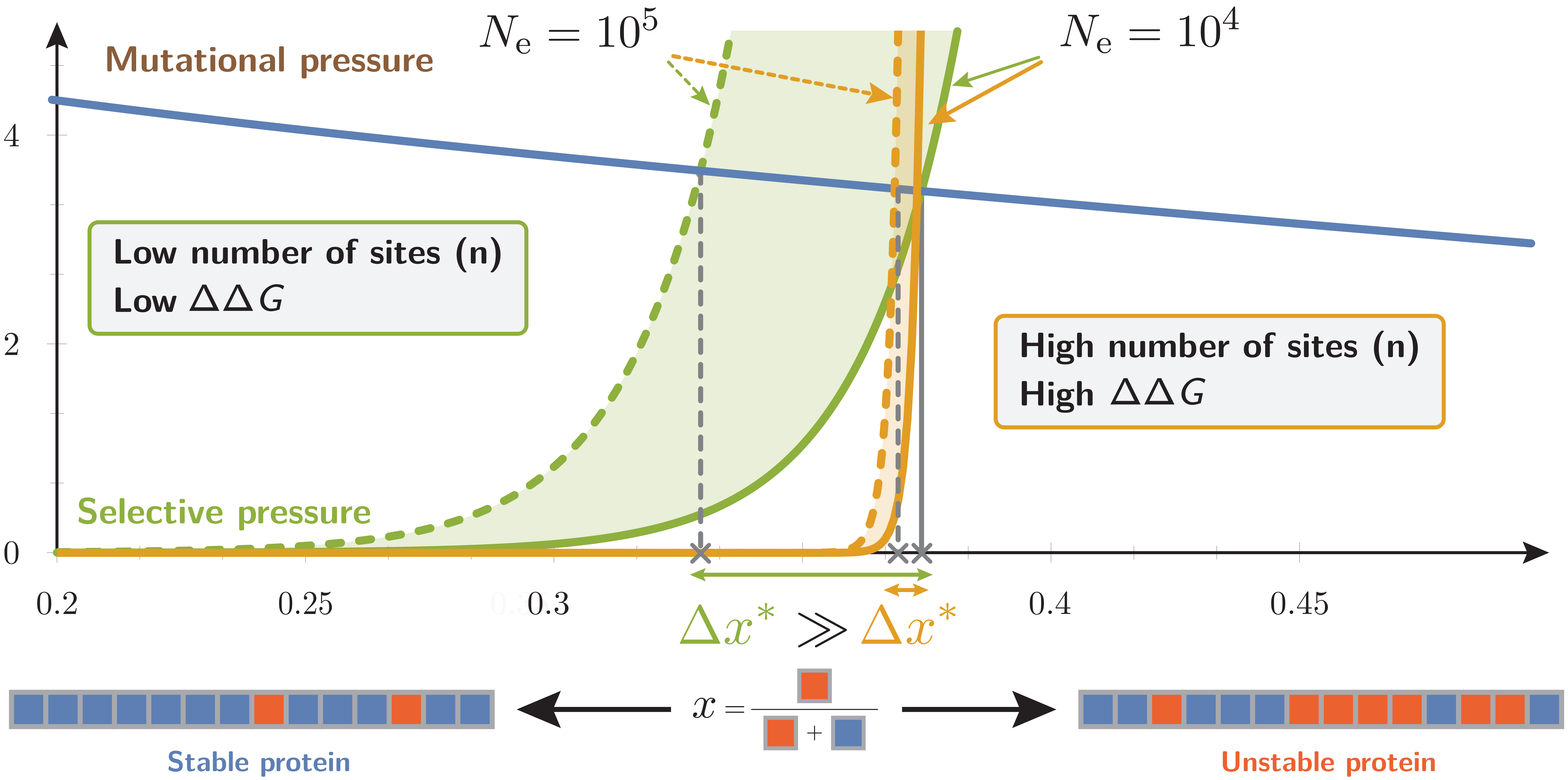
Unstable protein

What is the new phenotype at equilibrium after a change in N_e?





What is the new phenotype at equilibrium after a change in N_e for a sharp fitness function?



Low number of sites (n)
Low $\Delta\Delta G$

High number of sites (n)
High $\Delta\Delta G$

ω as a function of N_e

At equilibrium (x^*), the response in ω to changes in N_e is:

$$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_\text{e})} \simeq -\frac{\dfrac{\partial \ln f(x^*)}{\partial x^*}}{\dfrac{\partial^2 \ln f(x^*)}{\partial {x^*}^2}} \simeq -\frac{1}{\beta n \Delta\Delta G}.$$

- n is the number of sites in the protein.
- β is the temperature (equals to 1.686 mol/kcal at 25°C).
- $\Delta\Delta G > 0$ (in kcal/mol) is the expected change in free energy (between folded and unfolded states) for a destabilizing mutation.

https://github.com/ThibaultLatrille/GenotypePhenotypeFitness

ω as a function of protein expression level (y)

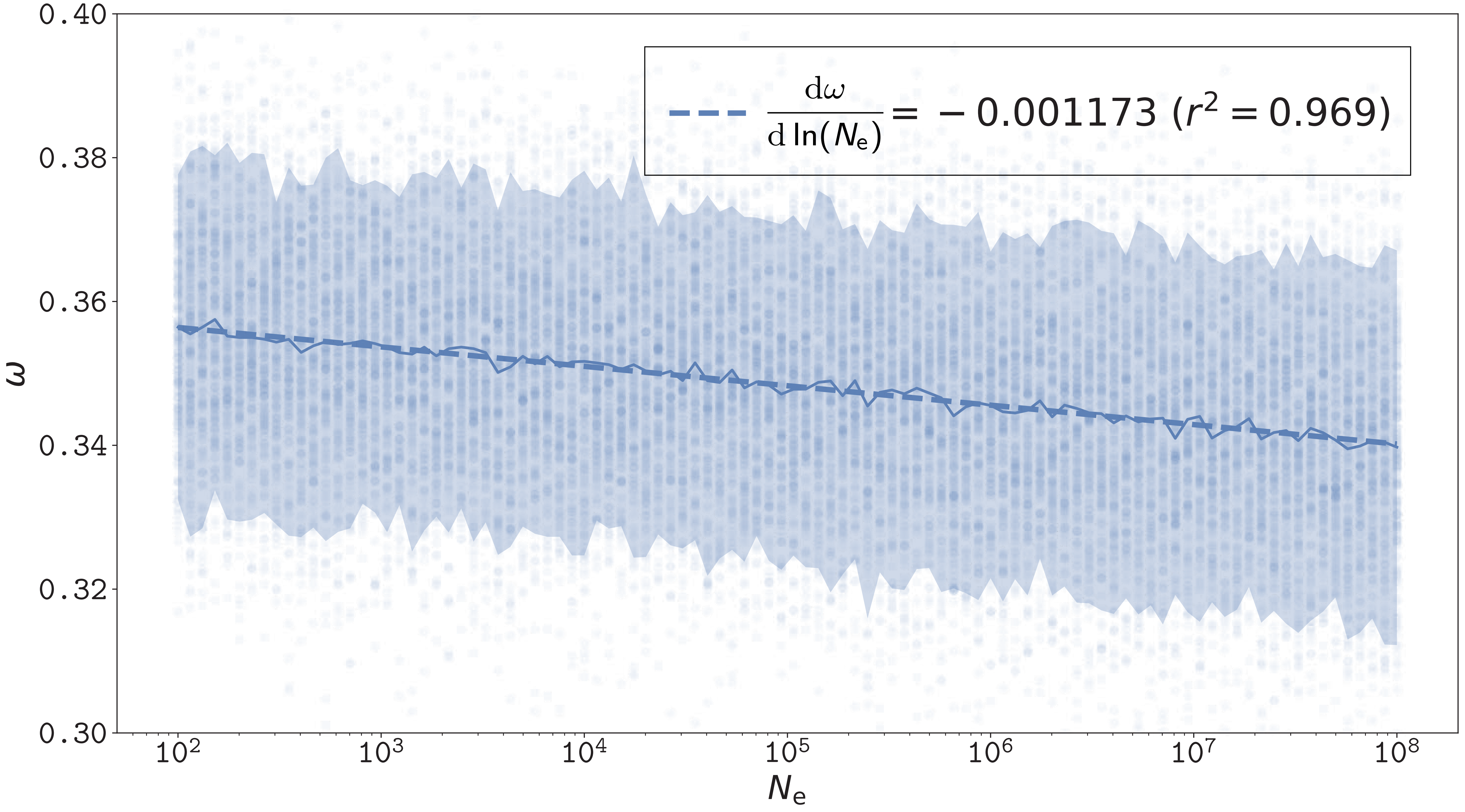
- If misfolded proteins are toxic, the decrease in fitness is proportional to the number of misfolded proteins.
- Hence, the decrease in fitness is proportional to protein expression level (y).
- As a result, selective pressure is proportional to both N_{e} and y.

The response in ω to changes in protein expression level (y) is also:

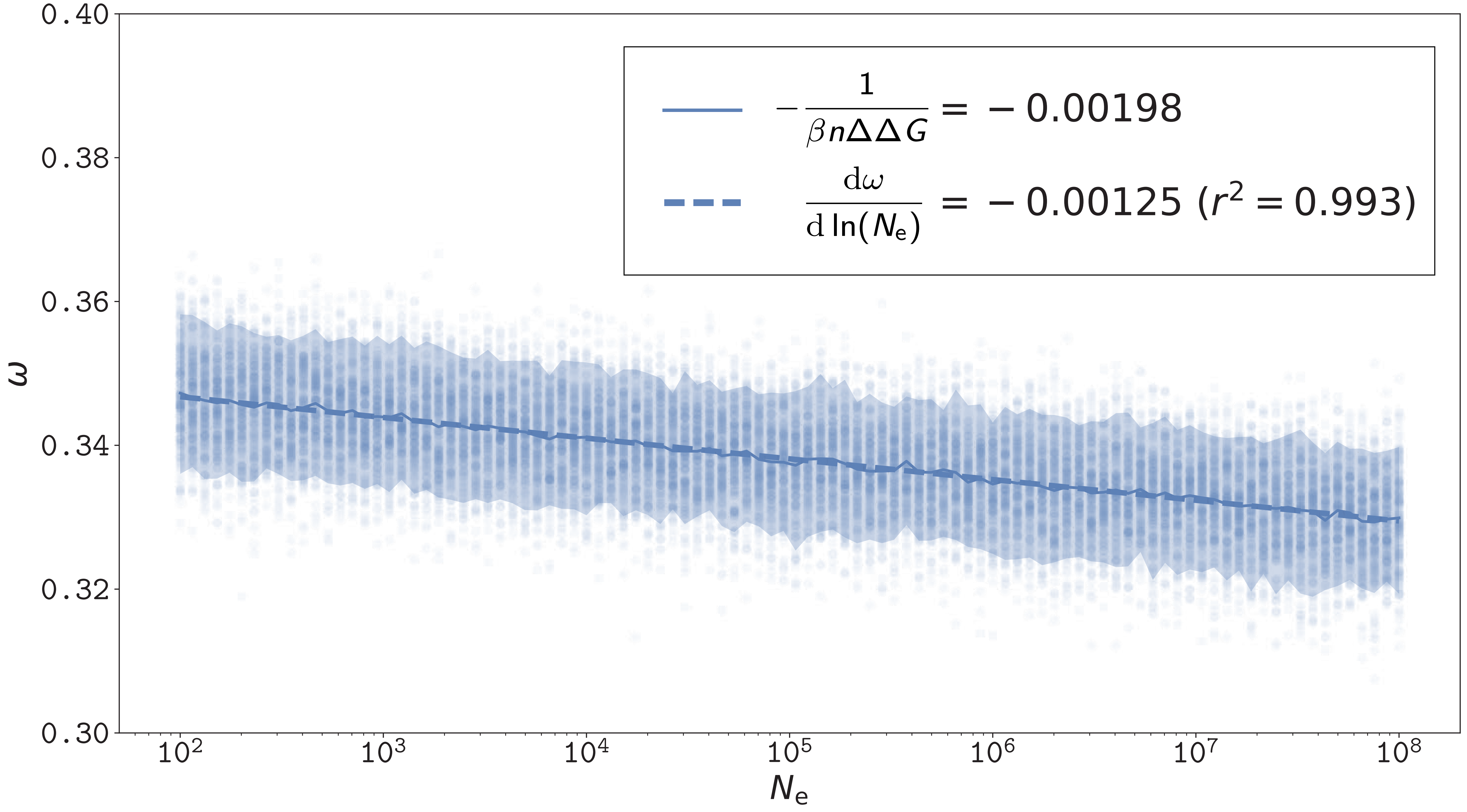
$$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(y)} \simeq \frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_{\mathrm{e}})} \simeq -\frac{1}{\beta n \Delta\Delta G}.$$

Confirmation of the theoretical results with simulations

Simulations with 3D protein model



Simulations with 1D protein model



- Parameters are $\Delta G_{\mathrm{min}} = -118$, $\Delta\Delta G = 1$, $n = 300$, $\beta = 1.686$.
- Theoretical slope is -0.00198 and observed is -0.00126

https://github.com/ThibaultLatrille/GenotypePhenotypeFitness

Interpreting theoretical results in the light of empirical data

Molecular parameters $\Delta\Delta G \simeq 1$ $n = 300$ $\beta = 1.686$	**ω function of N_e (diversity estimate) in primates**	**ω function of expression level in different Archaea & Bacteria**	**ω function of expression level in different Eukaryotes**
$-\frac{1}{\beta n \Delta\Delta G}$	$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_\mathrm{e})}$	$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(y)}$	$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(y)}$
-0.002	-0.04	$[-0.046; -0.021]$	$[-0.026; -0.004]$

- Weak predicted linear response of ω to changes in either N_e or expression level.
- Models based on the probability of folding are at odds with empirically results.
- Other aspects of protein biophysics could be explored such as protein-protein interactions.

Zeldovich *et al* (2007), Goldstein (2013), Zhang & Yang (2015), Brevet & Lartillot (2020).

V. Conclusion

Modelling the interplay between selective and neutral mechanisms

- Can ω-based codon models disentangle mutation and selection?

→ No, if a single ω.
→ Yes, if ω in different directions.

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

→ N_e estimation in the right direction.
→ The magnitude of estimated N_e is lower than expected, probably due to mis-specification of the mutation-selection model.

- Can the response ω to changes in N_e be derived generally at mutation-selection balance?

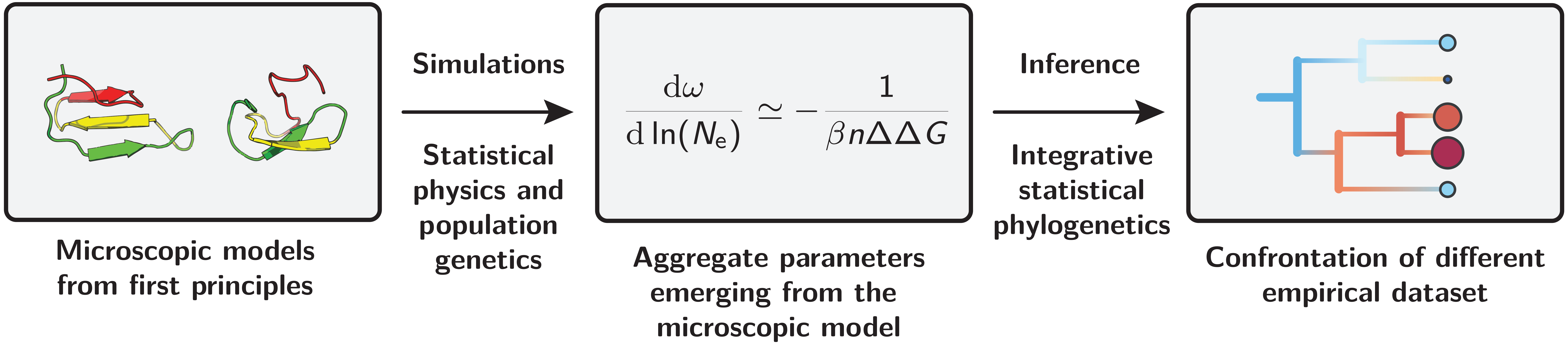
→ Yes, under a linear 1D model of fitness based on protein stability.
→ Weaker dependency of ω to changes in N_e as the number of sites increases.
→ Response of ω to changes in N_e and protein expression level is equal.

Inference framework

- Mechanistic mutation-selection codon models are complex and heavily parameterized, but are still relying on strong assumptions broken in practice.

- Phenomenological models (ω-based) are more easily fitted to the data, but require careful definition and parameterization.

- Aggregate parameters (ω) can be derived out of population-genetic (N_e) and molecular parameters ($\Delta\Delta$G, β...).



Microscopic models from first principles

Simulations

Statistical physics and population genetics

$$\frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_\mathrm{e})} \simeq -\frac{1}{\beta n \Delta\Delta G}$$

Aggregate parameters emerging from the microscopic model

Inference

Integrative statistical phylogenetics

Confrontation of different empirical dataset

Thank you

Advisor

Nicolas Lartillot

PhD committee

Céline Brochier-Armanet
Julien Yann Dutheil
Richard Goldstein
Carina Farah Mugal

PhD guides

Benoit Nabholz
Christophe Douady
Tristan Lefébure
Laurent Gueguen
Laurent Duret
Nicolas Rodrigue

LBBEers

Anamaria Necsulea
Damien de Vienne
Dominique Mouchiroud
Hélène Badouin
Annabelle Haudry
Bastien Boussau
Éric Tannier
Vincent Daubin
...

Teachers

Marie Sémon
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

The nest

Judith Estrada Meza
Eric & Anne
Iris, Myriam & Samuel
Lamonerie & Latrille
...

Migrating finches

Wandrille Duchemin
Adrián Arellano Davín
Aline Muyle
Héloïse Phillippon
Pierre Garcia
Monique Aouad
Anne Oudard
Frédéric Jauffrit
Maud Gautier
Samuel Barreto
Vincent Lanore
François Gindraud
Diego Hartasánchez Frenk
...

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
...