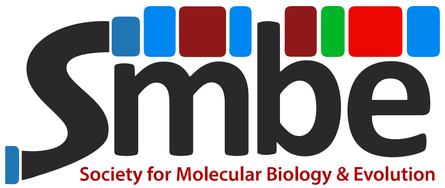
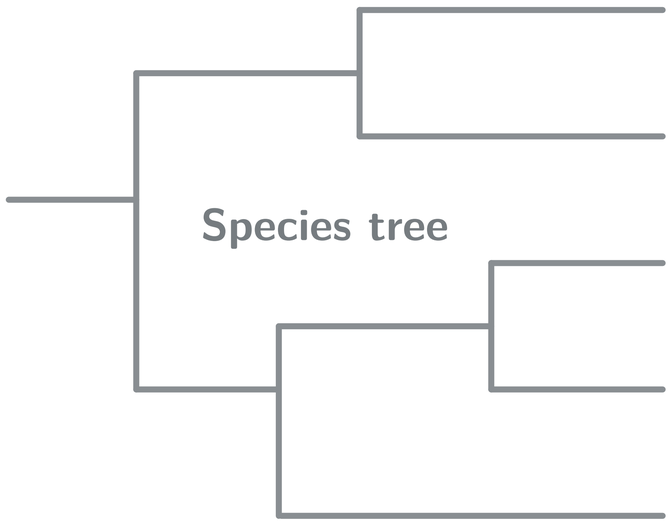
Inferring fluctuating population size and selection with phylogenetic codon models

Thibault Latrille - PhD student
Nicolas Lartillot - Research director



Theoretical and empirical motivations



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

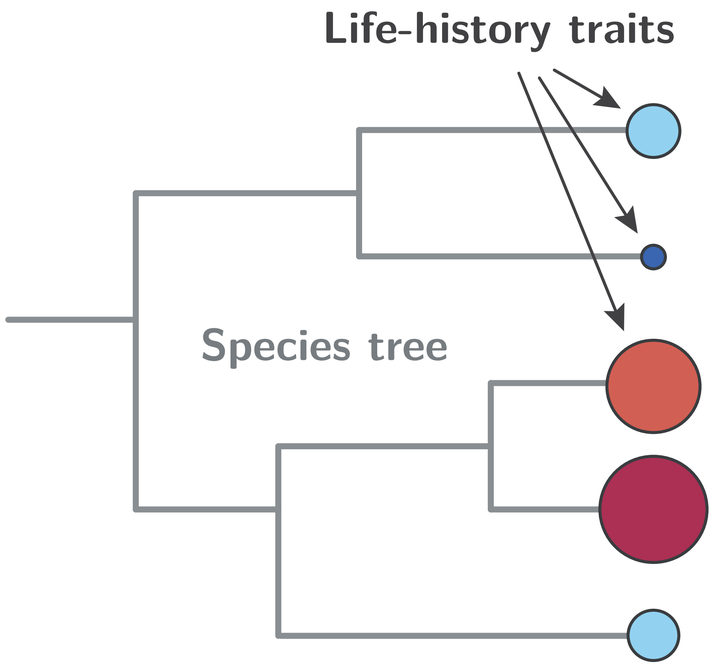
ATG|CGA|TAG|AAG|CTT|CGA|TCG

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

- **Can we quantify selection acting on protein coding and long term changes in effective population size, from a DNA alignment?**

- **Do we have enough signal for empirical estimation of selection and drift?**

Theoretical and empirical motivations



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

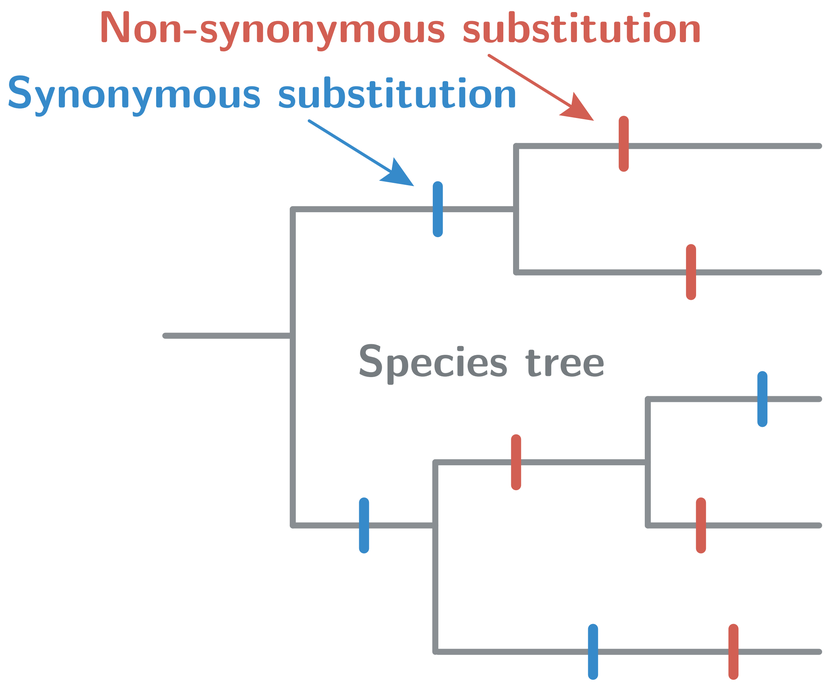
ATG|CGA|TCG|ATC|CAT|CGA|TCG

- **How are species life-history traits (longevity, maturity, weight, size, ...) related to population-genetics parameters (effective population size, mutation rate,...)?**

- **Can we reconstruct changes in effective population size instead of using dN/dS as a proxy?**

Stearns (1972); Lartillot & Poujol (2011); Weber et al (2014); Figuet (2016); James (2018).

Mapping substitutions along the species tree



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

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

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

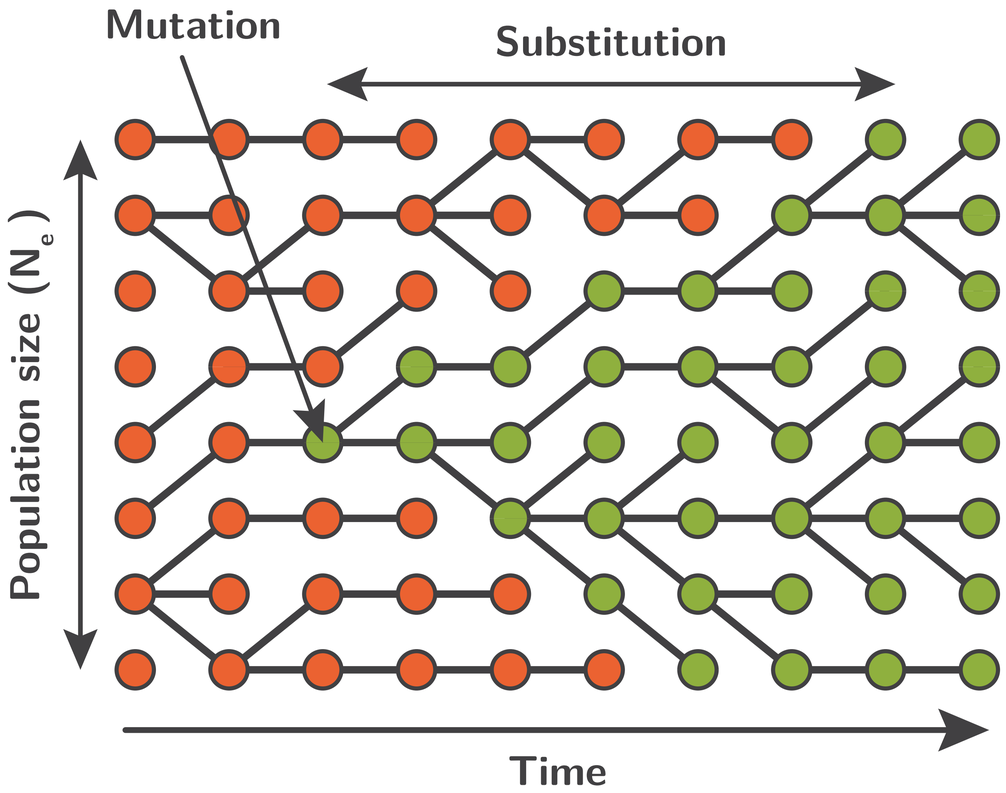
ATG|CGA|TAG|AAG|CTT|CGA|TCG

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

- **Synonymous mutations (not changing the protein) are assumed to be neutral.**
- **Non-synonymous mutations (changing the protein) are assumed to be under selective pressure at the amino-acid level.**

Stearns (1972); Lartillot & Poujol (2011); Weber et al (2014); Figuet (2016); James (2018).

Substitution is a mutation that reached fixation

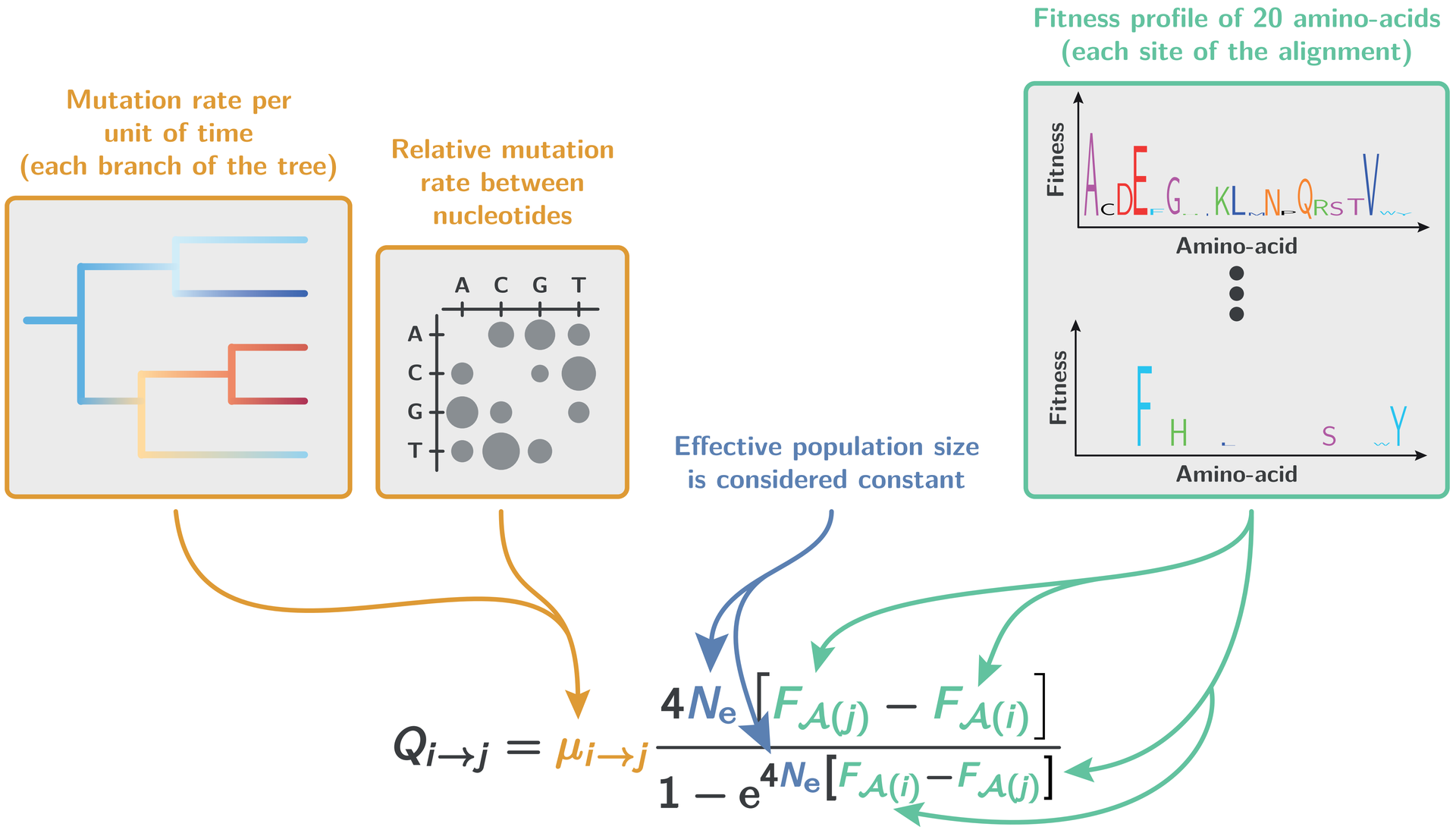


$$Q_{\mathsf{ATT}\to\mathsf{ATG}} = \mu_{\mathsf{T}\to\mathsf{G}} \frac{4N_e\left[F_{\mathsf{Met}} - F_{\mathsf{Ile}}\right]}{1 - e^{4N_e\left[F_{\mathsf{Ile}} - F_{\mathsf{Met}}\right]}}$$

- $Q_{\mathsf{ATT}\to\mathsf{ATG}}$ is the substitution rate from codon **ATT** to **ATG**.
- N_e is the effective population size.
- $\mu_{\mathsf{T}\to\mathsf{G}}$ is the mutation rate from nucleotide **T** to **G**.
- F_{Ile} (F_{Met}) is the fitness of Isoleucine (Methionine).

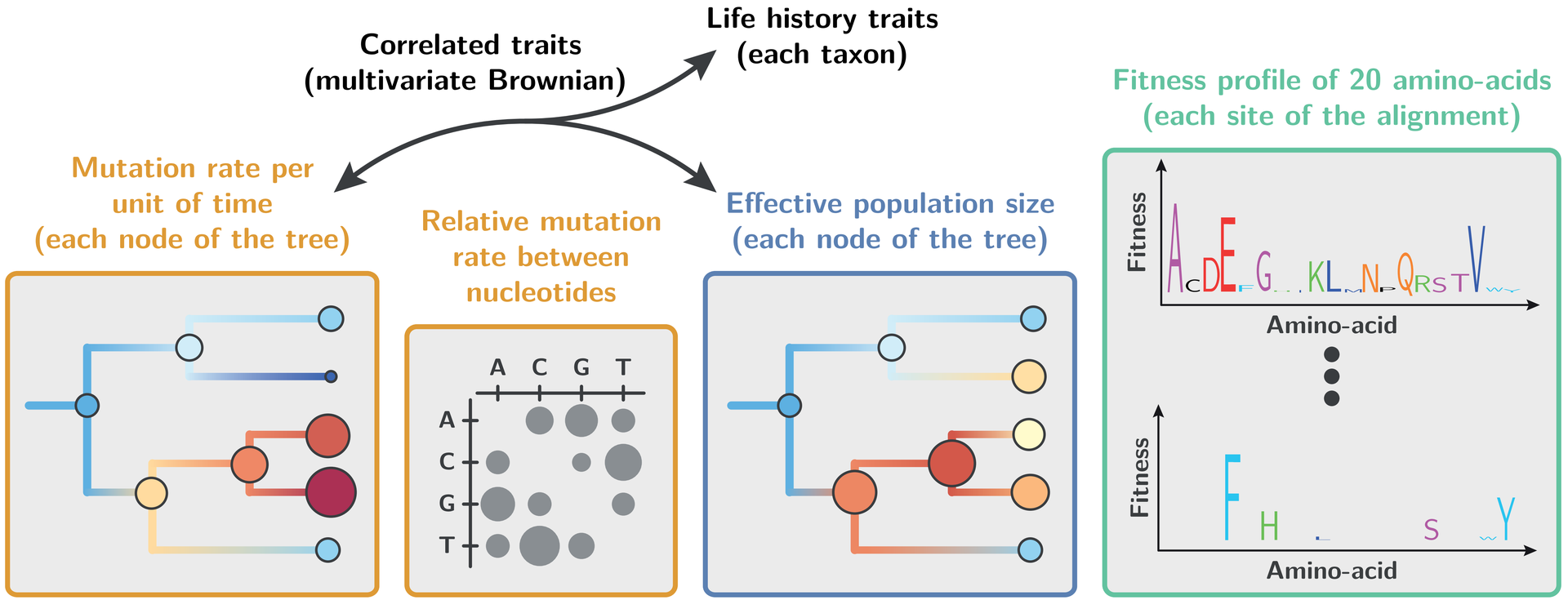
Kimura (1983); Ohta (1992); McClandish (2016).

Mechanistic mutation-selection codon models



Halpern & Bruno (1998); Tamuri *et al* (2014); Rodrique & Lartillot (2016).

Reconstructing changes in effective population size



$$Q_{i \to j} = \mu_{i \to j} \frac{4N_e \left[F_{A(j)} - F_{A(i)} \right]}{1 - e^{4N_e \left[F_{A(i)} - F_{A(j)} \right]}}$$

https://github.com/bayesiancook/bayescode/tree/chronogram

What need to be estimated?

Tree:

- **Age for each internal node of the dated tree.**

Mutation:

- **Mutation rate (per unit of time) for each node of the tree.**
- **Nucleotide relative rate matrix.**

Selection:

- **20 Amino-acid fitnesses for each profile category (K categories).**
- **Which profile category (1..K) for each site of the alignment.**

Drift:

- **Population size for each node of the tree.**

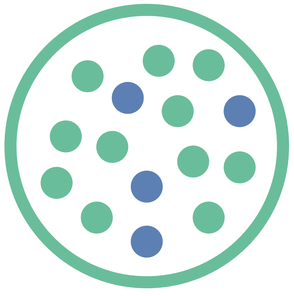
Traits:

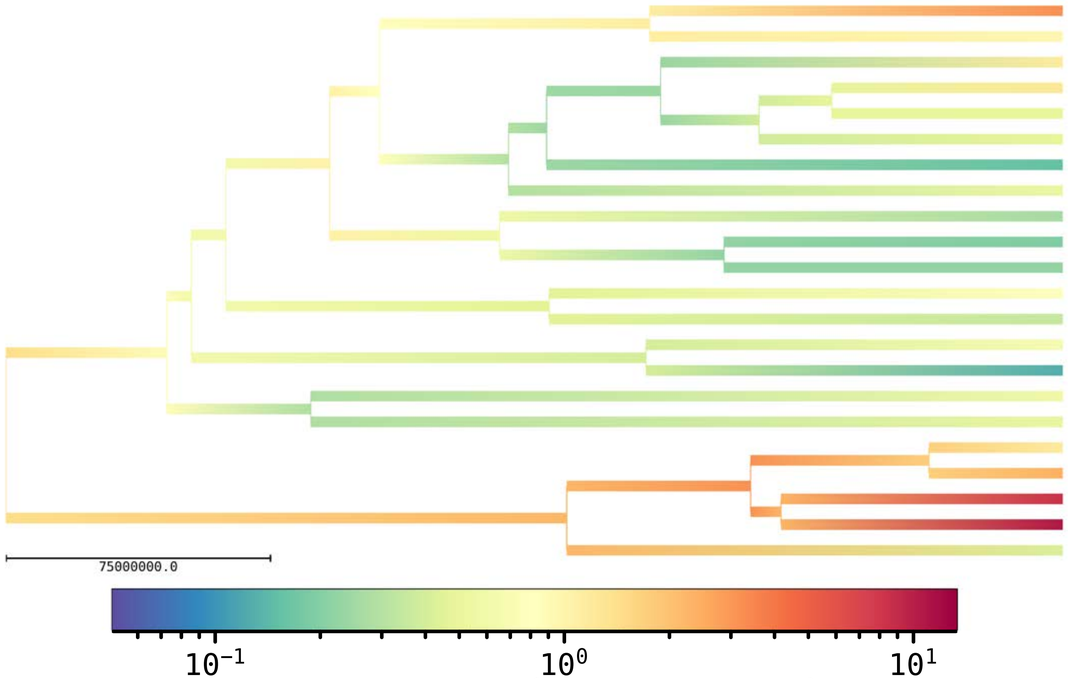
- **Life-history-traits for each node of the tree.**
- **Covariance matrix between traits (molecular and life-history).**

https://github.com/bayesiancook/bayescode/tree/chronogram

Wright-Fisher simulator along the phylogeny

Initial population size of 5000 individuals.
Mutation rate of 10^{-8} per generation per site.
Generation time of 20 years.
Sequence of 15 000 codon sites (50 exons).



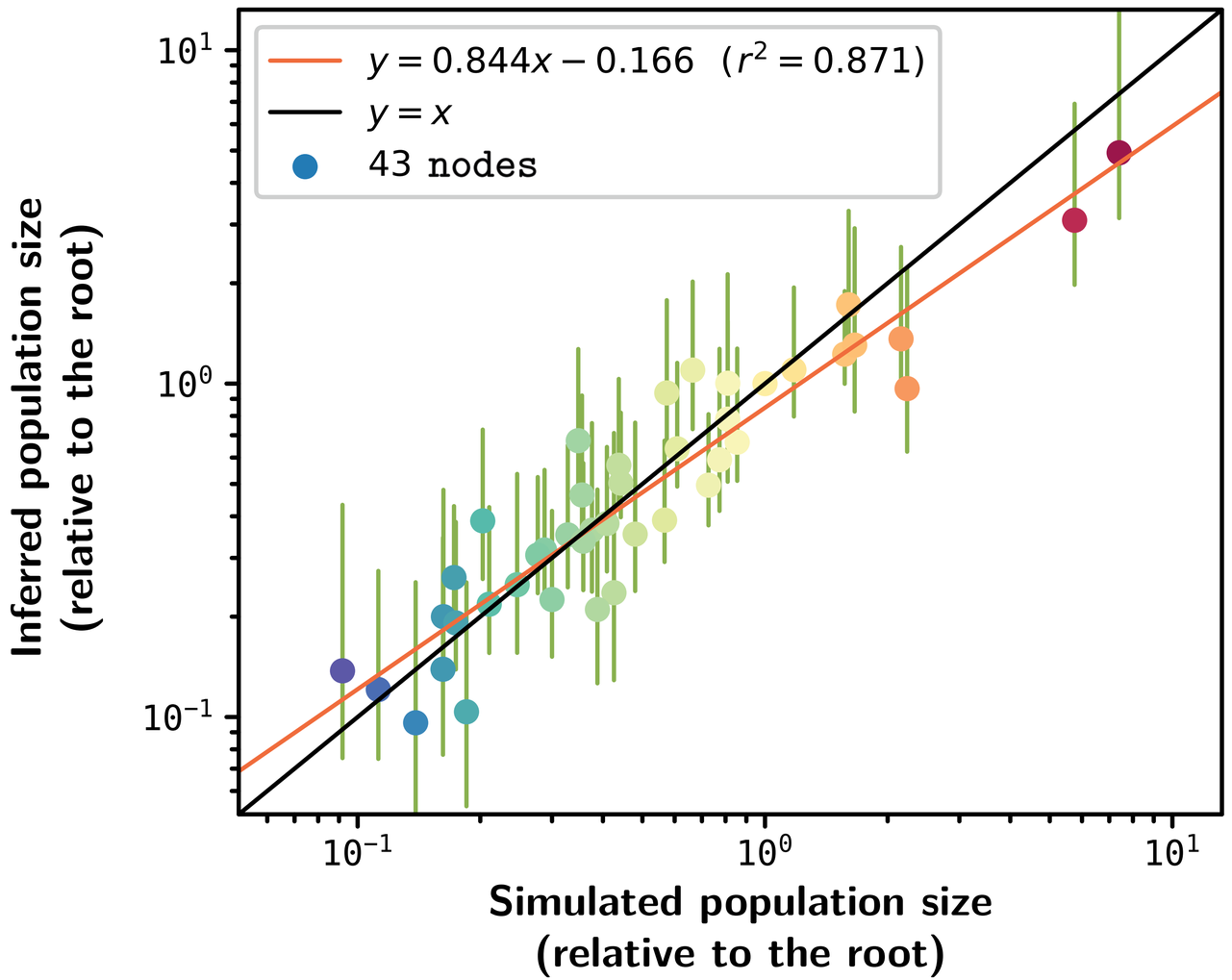


At each generation: mutation, selection & drift.

Alignment of protein coding DNA sequence

https://github.com/ThibaultLatrille/SimuEvol

Inference against simulated data (1/2)



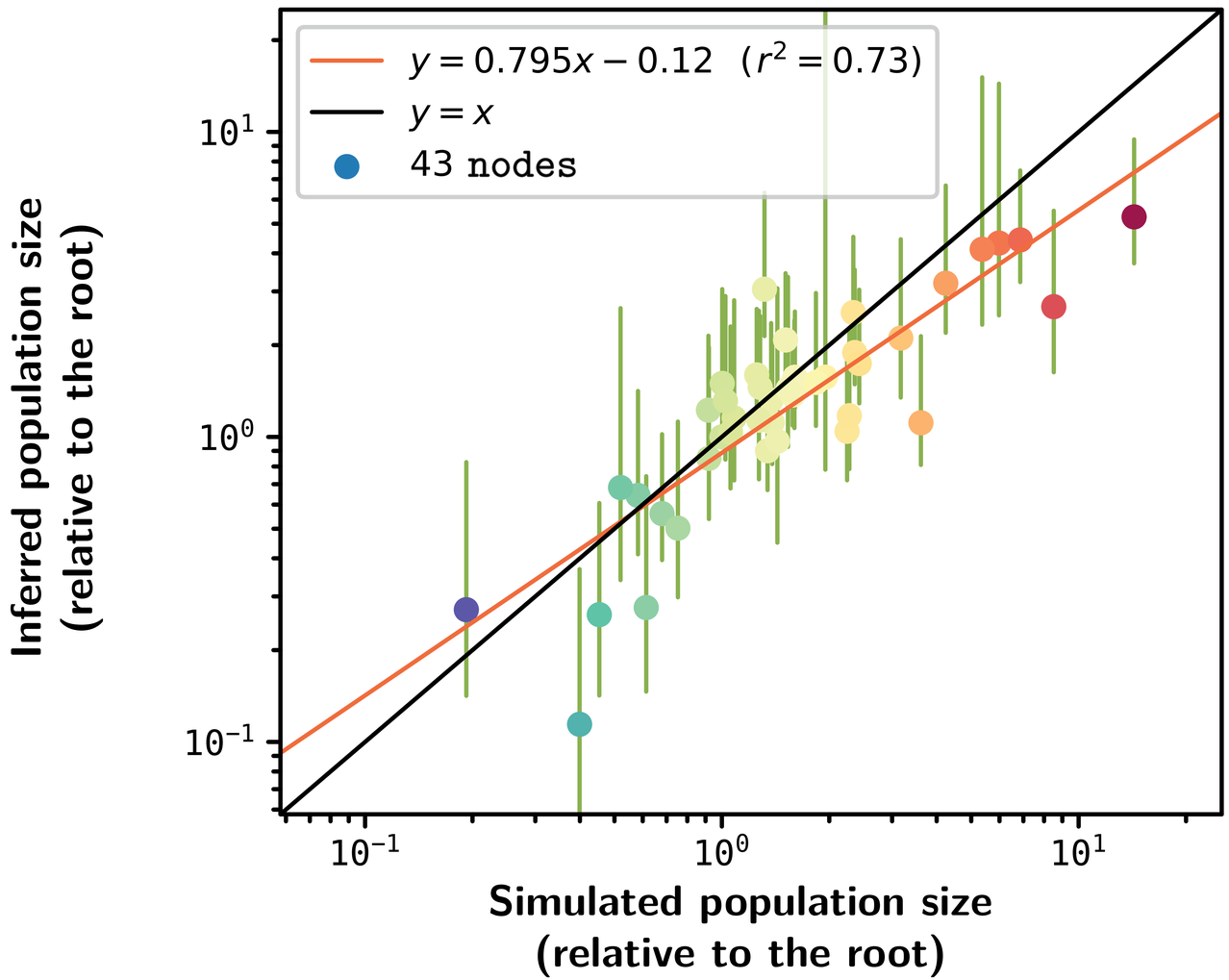
Taken into account by the simulator:

- **Long term fluctuation of population size, mutation rate per time, generation time;**
- **Fitness landscape for each site;**
- **Finite population size;**
- **Allele hitching;**

Not taken into account by the simulator:

- **Species tree ≠ gene tree;**
- **Epistasis;**
- **Fluctuating fitness landscape;**
- **Biased gene conversion;**
- **Selection on codon usage;**

Inference against simulated data (2/2)



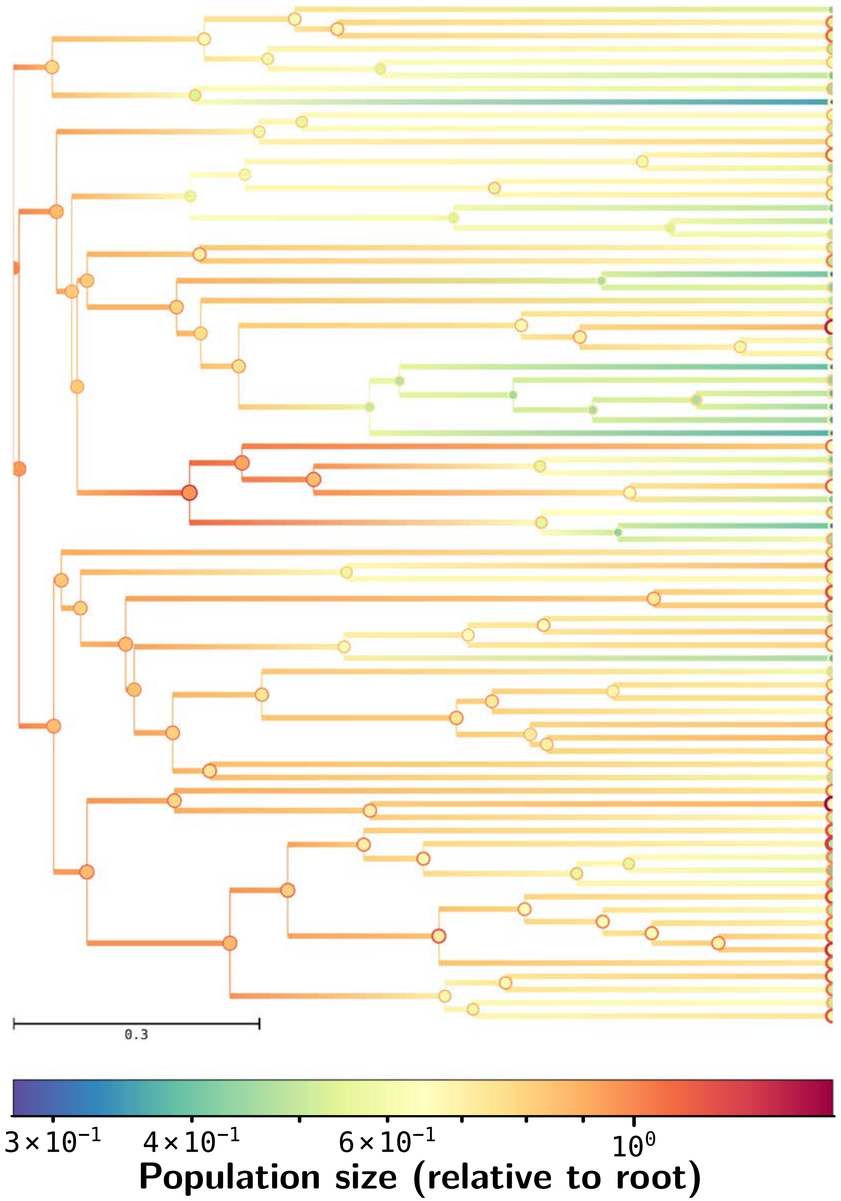
Taken into account by the simulator:

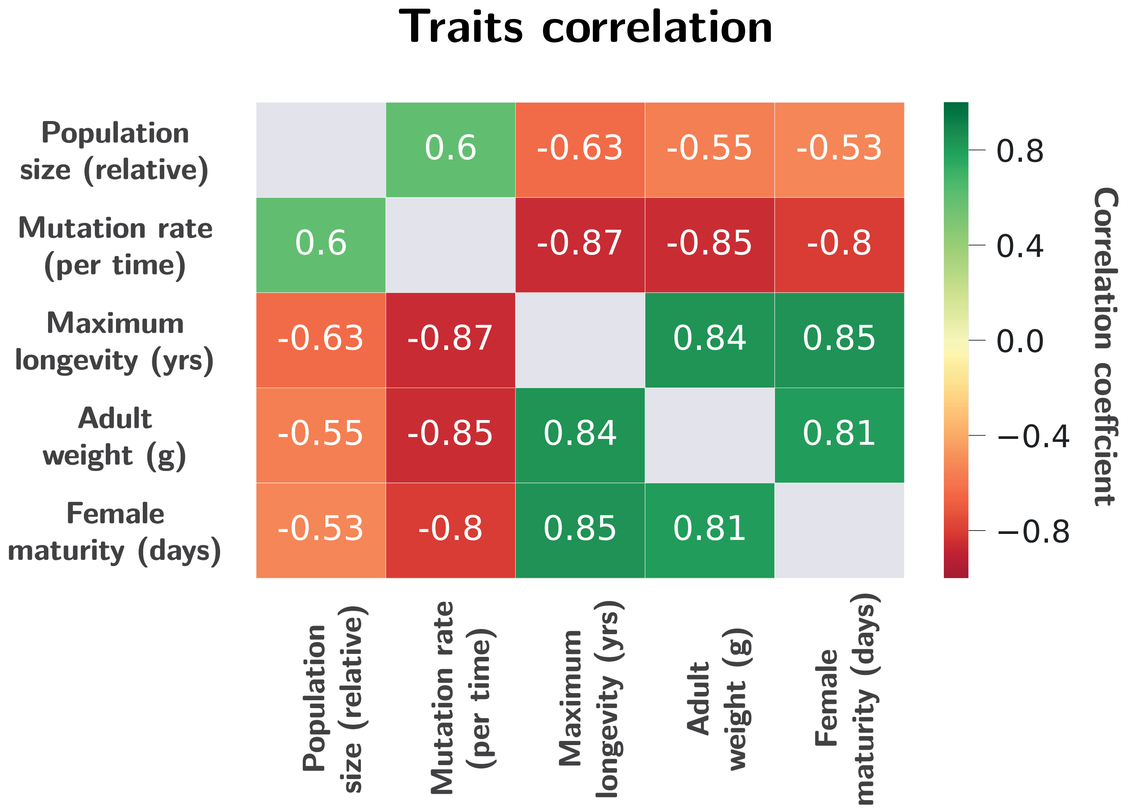
- **Long term fluctuation of population size, mutation rate per time, generation time;**
- **Fitness landscape for each site;**
- **Finite population size;**
- **Allele hitching;**
- **Short term fluctuation of population size;**

Not taken into account by the simulator:

- **Species tree ≠ gene tree;**
- **Epistasis;**
- **Fluctuating fitness landscape;**
- **Biased gene conversion;**
- **Selection on codon usage;**

Inference with mammalian empirical data





- **Concatenated random sample of 24 highly conserved coding sequences (>99% coverage) from OrthoMam database.**
- **Life-history traits extracted from AnAge database.**

Tacutu (2013), Scornavacca (2019)

Take home message

- **Fluctuating population size and selection can be inferred from protein coding DNA sequences using a phylogenetic approach.**

- **In mammals, population size correlates negatively with longevity, weight and maturity, and positively with mutation rate.**

- **The mechanistic mutation-selection model can be extended by taking into account polymorphism within species.**

- **Which mechanism could explain such a low variance of population size observed in empirical data? Epistasis, fluctuating selection, short-term fluctuation of population size...**

Thanks

Nicolas Lartillot, Vincent Lanore, Philippe Veber, Nicolas Rodrigue, Bastien Boussau, Florian Benitiere.

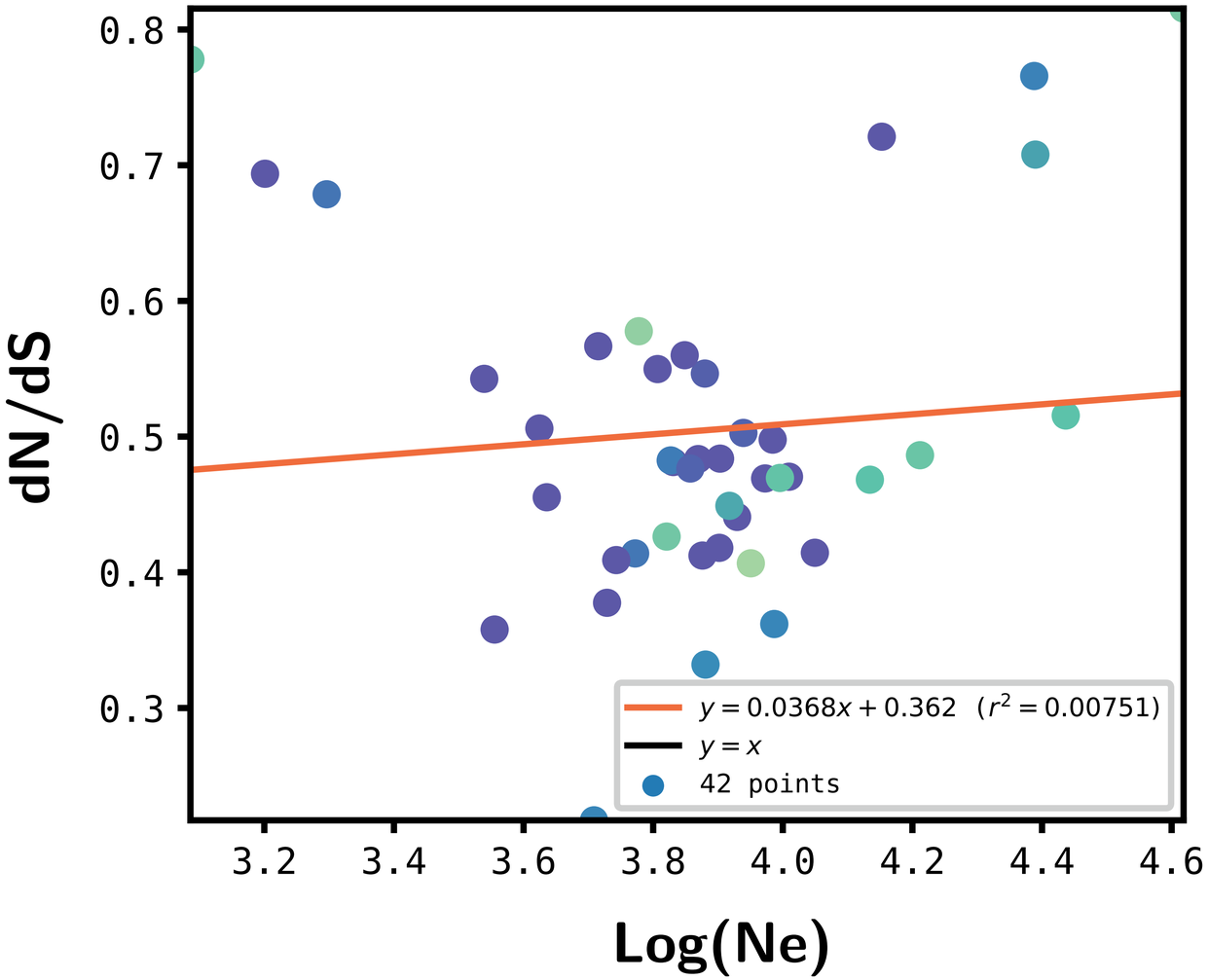
All lab members.

You for your attention!

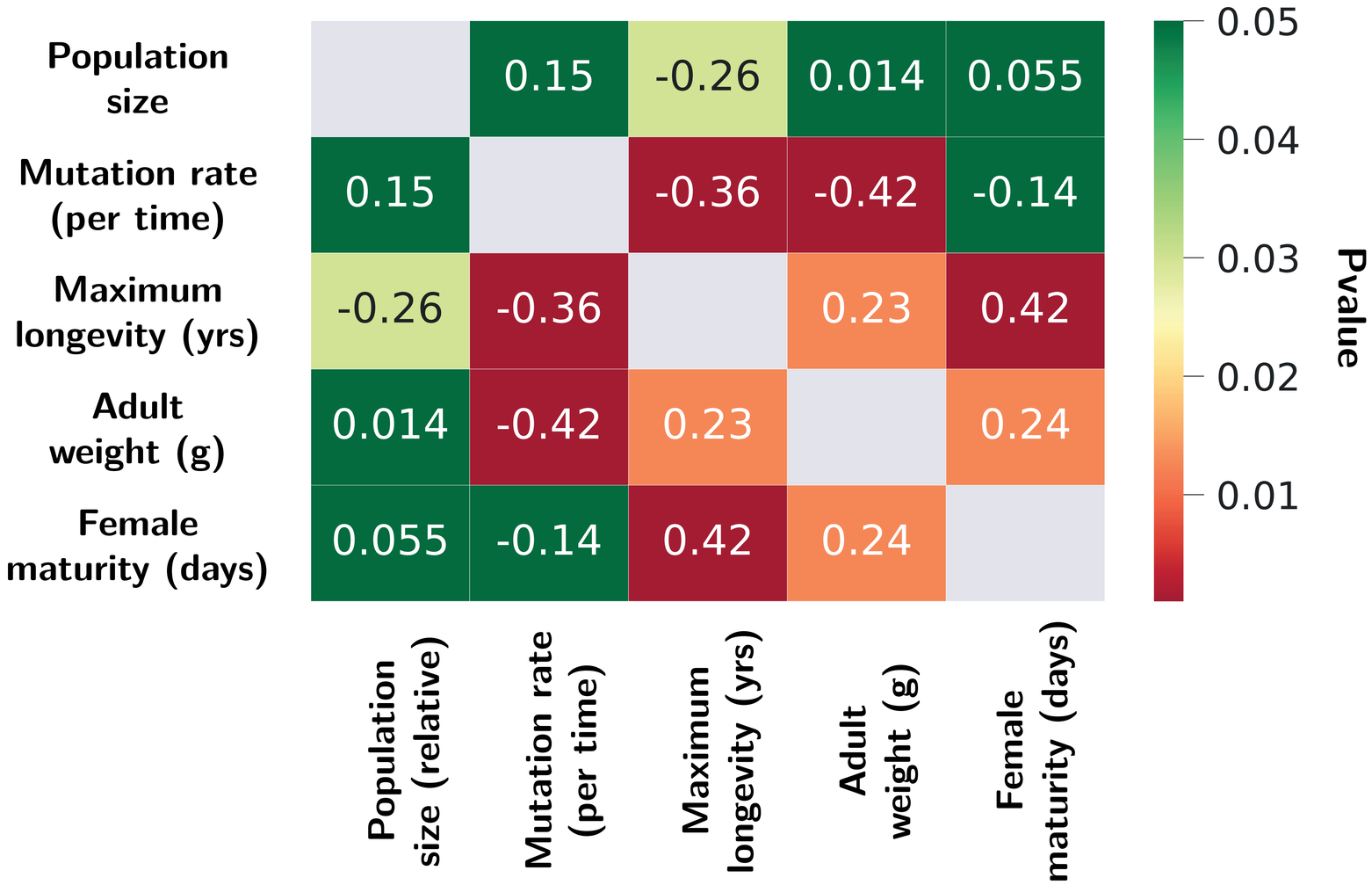
thibault.latrille@univ-lyon1.fr
@phylogenetrips

https://github.com/ThibaultLatrille/SimuEvol
https://github.com/bayesiancook/bayescode

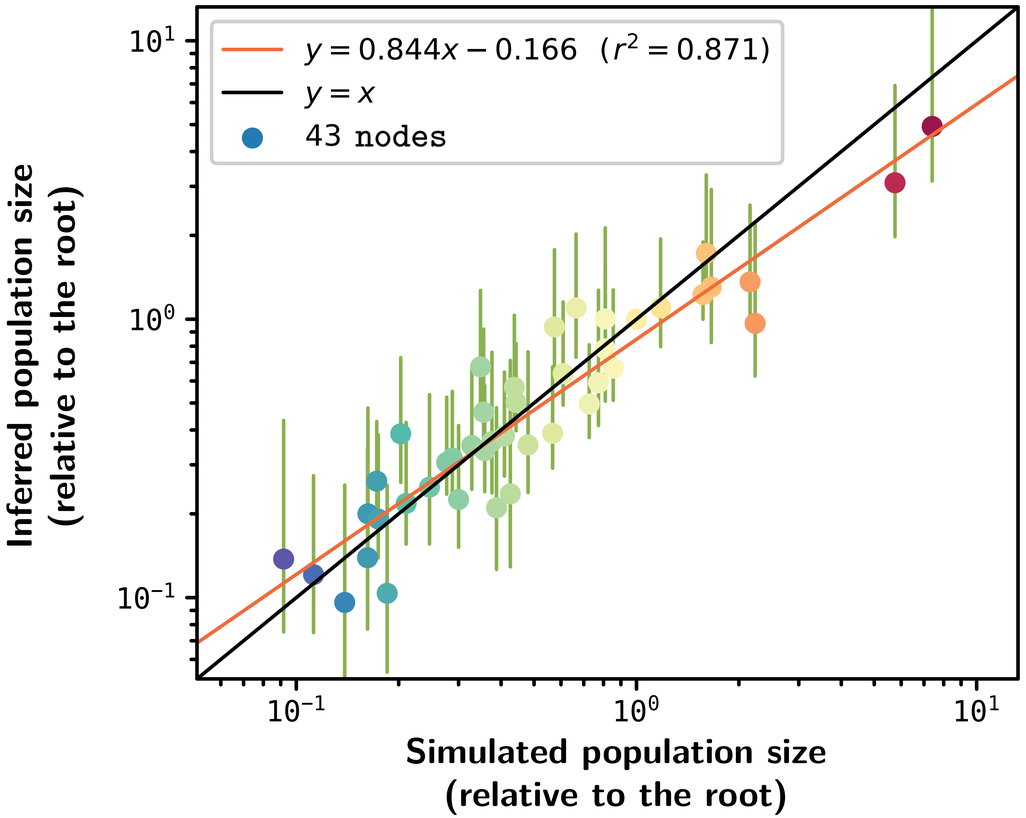
dN/dS is a poor estimate of population size

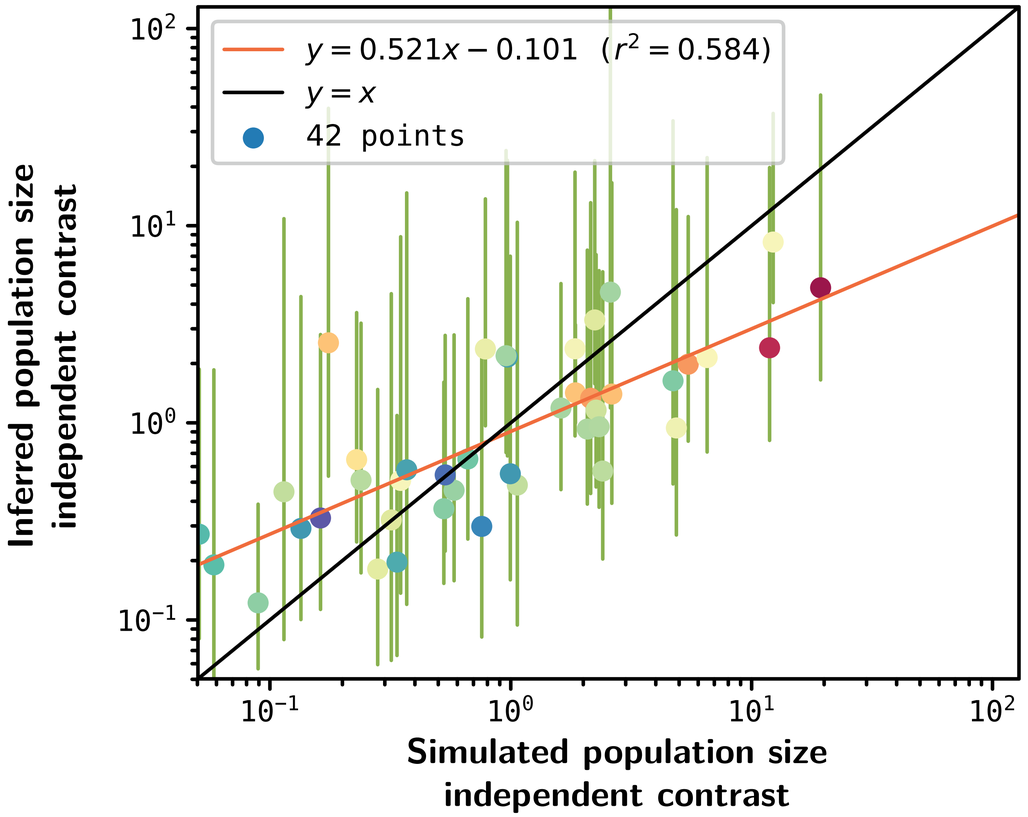


Partial correlation matrix

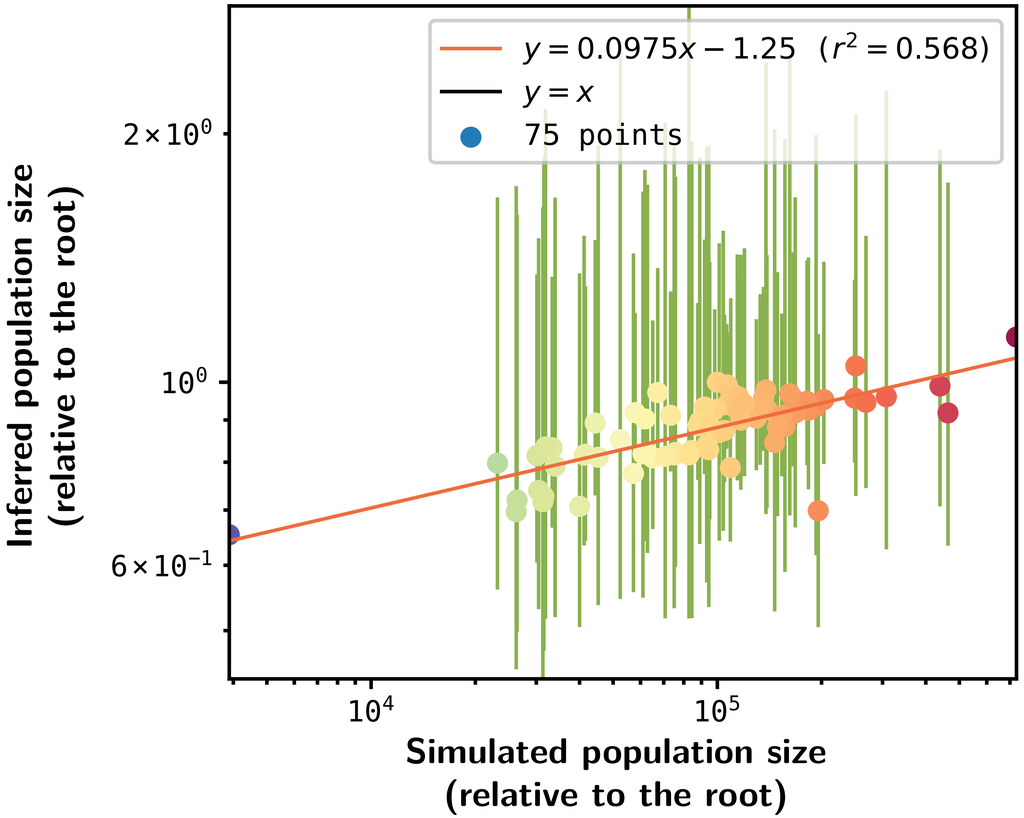


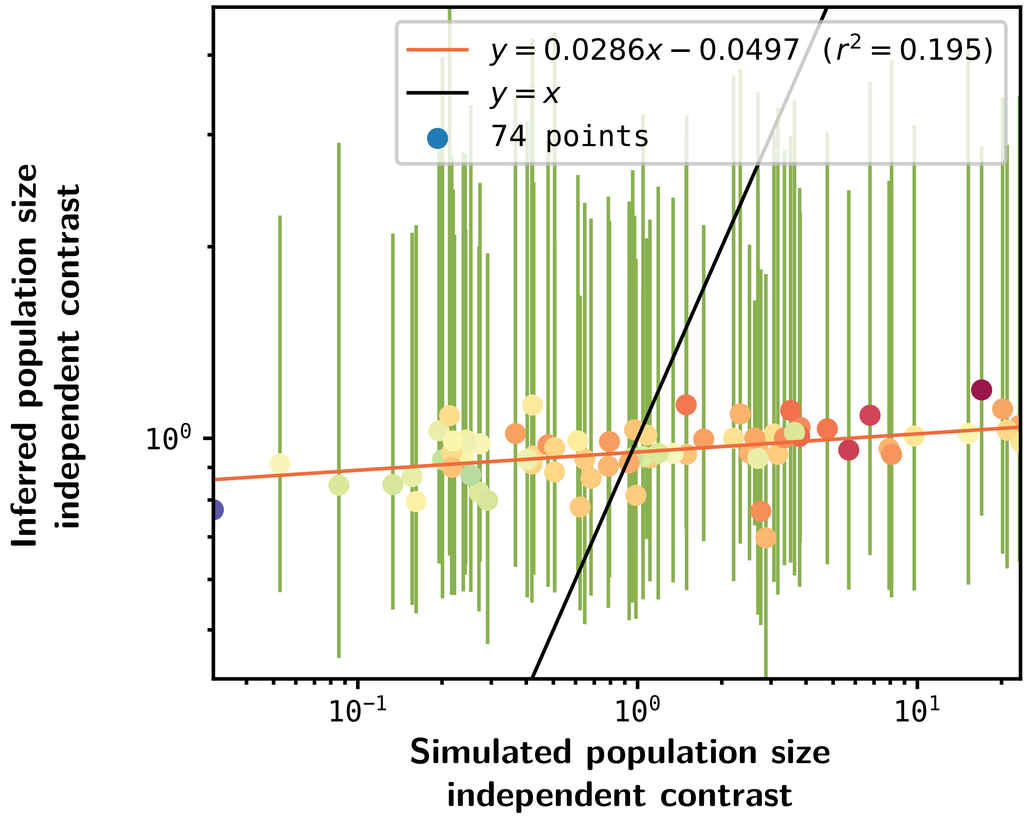
Comparing simulation and inference using independent contrast





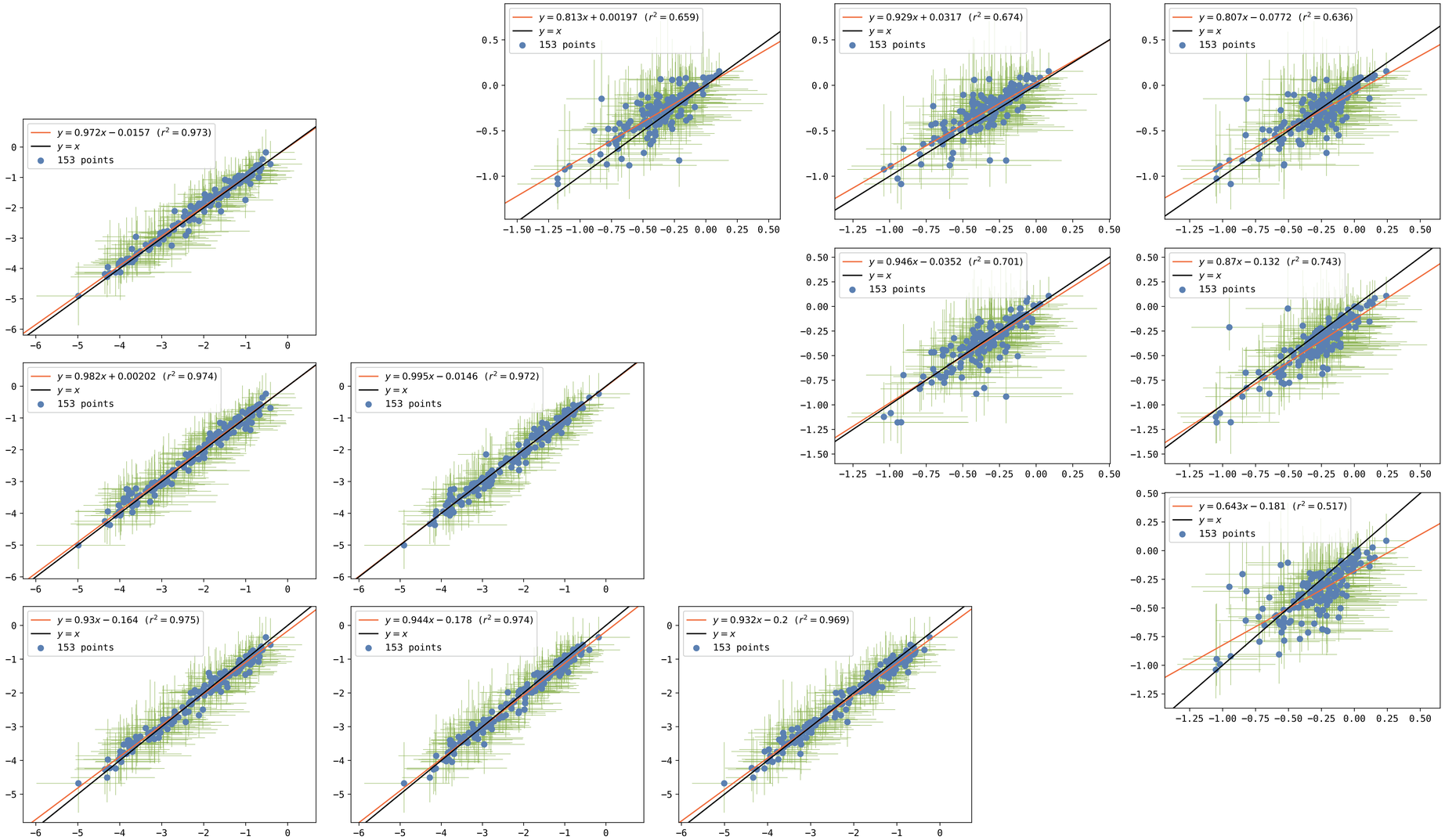
Estimating population size in the presence of epistasis





Mammalian experiment repeatability

Effective population size



Mutation rate per unit of time