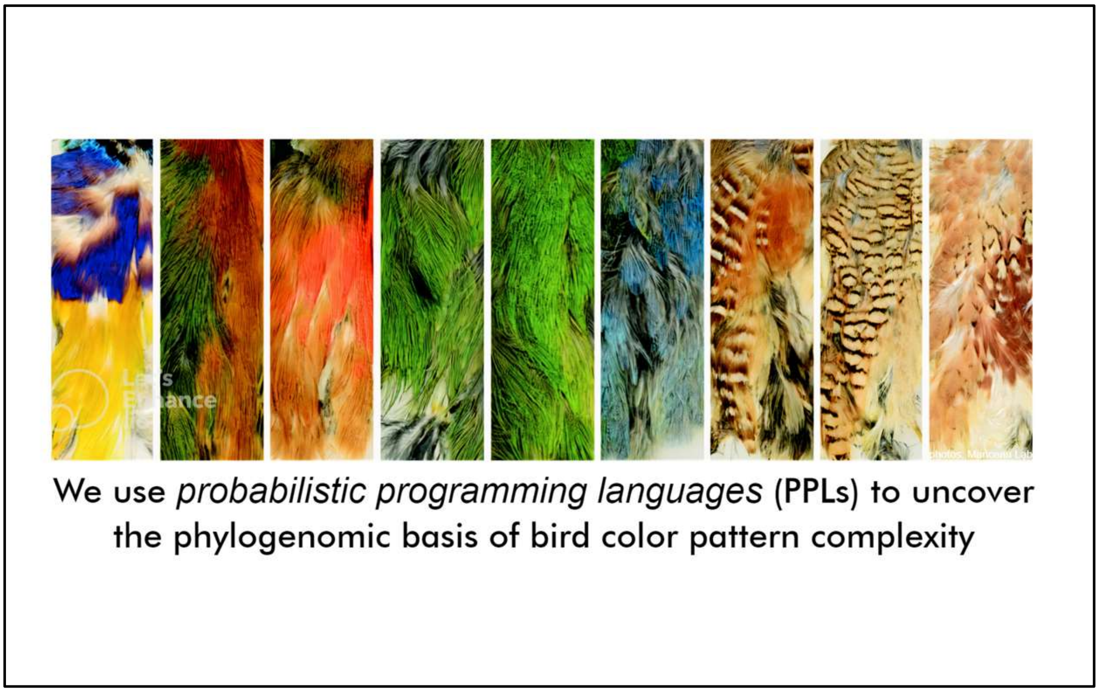
# A Probabilistic Programming Approach to Investigate the Coevolution of Genes and Phenotypes in Birds

Viktor Senderov, Amaury Lambert, Marie Manceau, Carole Desmarquet, Caitlyn Jean-Baptiste, Ingrid Lafontaine, Hélène Morlon

ISMB/ECCB 2023 - Lyon
2023-07-27

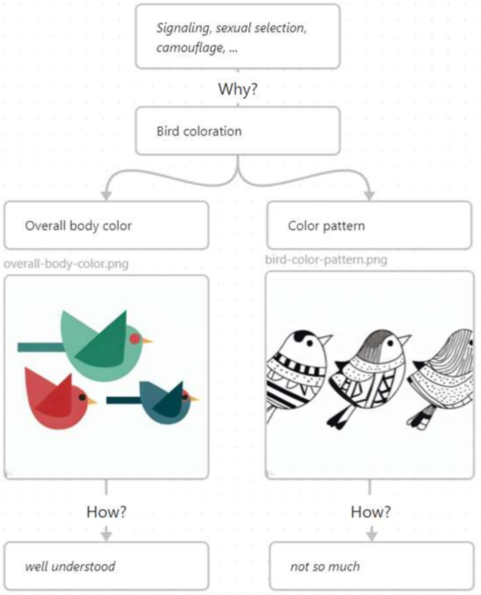


We use *probabilistic programming languages* (PPLs) to uncover the phylogenomic basis of bird color pattern complexity

These are some of the extraordinary colors of the Estrildid family of finches.

## Talk outline

1. Overview of bird coloration
2. Mathematical modeling
3. Probabilistic programming languages (PPLs)



*Signaling, sexual selection, camouflage, ...*

Why?

Bird coloration

Overall body color

Color pattern

How?

How?

*well understood*

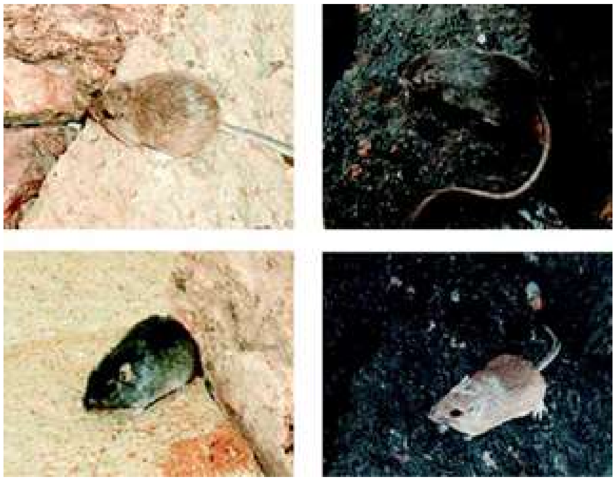
*not so much*

What is the cause of the extraordinary diversity in bird coloration?

The answer to the question “*What is the cause of the extraordinary diversity of bird coloration?”* may depend on whether we are interested in the evolutionary drivers of bird color diversity or whether we are interested in the mechanisms by which bird coloration is formed.

The evolutionary drivers have been studied extensively and range from communication and sexual selection to camouflage and predator avoidance.

We also understand quite a bit about the developmental mechanisms by which overall bird color is produced. In fact color can be generated through deposition of pigments or by structural features that interfere with light.



(Photo: Hopi E. Hoekstra) Light and dark rock pocket mice on light granite and dark basalt rocks

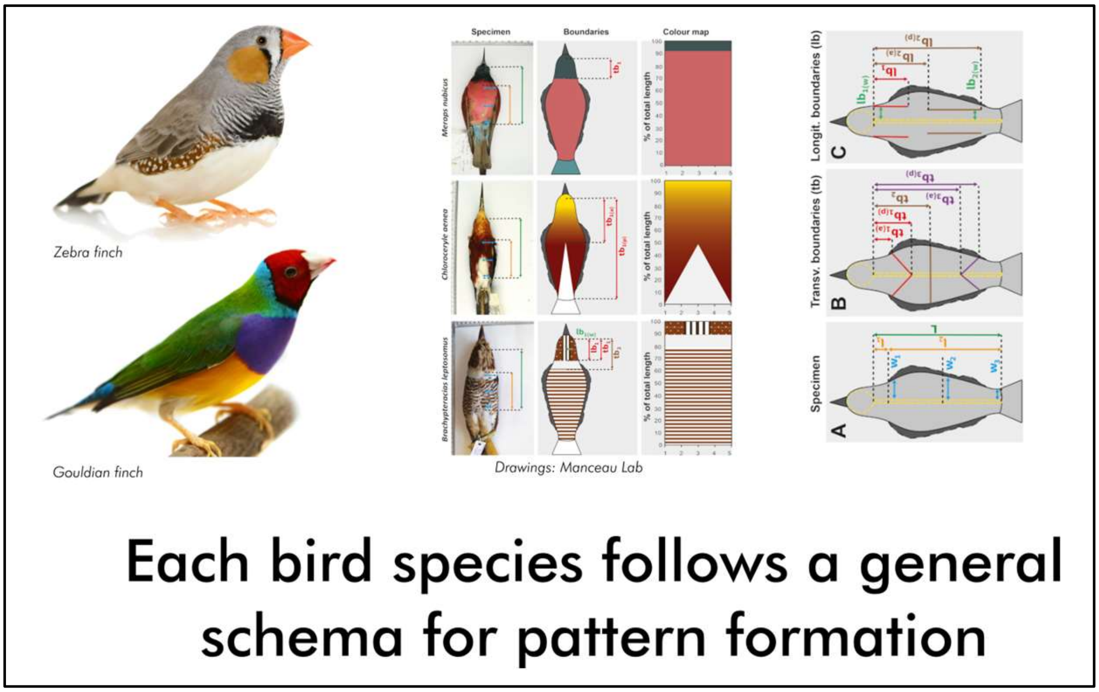
The darker color of some populations of Rock pocket mice that inhabit dark lava substrates is caused by coding changes in *Melanocortin 1* (*Mc1r*).

overall-body-color.png
bird-color-pattern.png

What is the cause of the extraordinary diversity in bird coloration?

Our study focuses on the mechanistic understanding of how bird color *patterns* are produced. A useful analogy is to think of a children's coloring book.

We believe that color patterns are evolutionary more conserved than color itself. For example, a group of related birds may have different colors but still share a similarly in the size and shape of the color domains.



Zebra finch

Gouldian finch

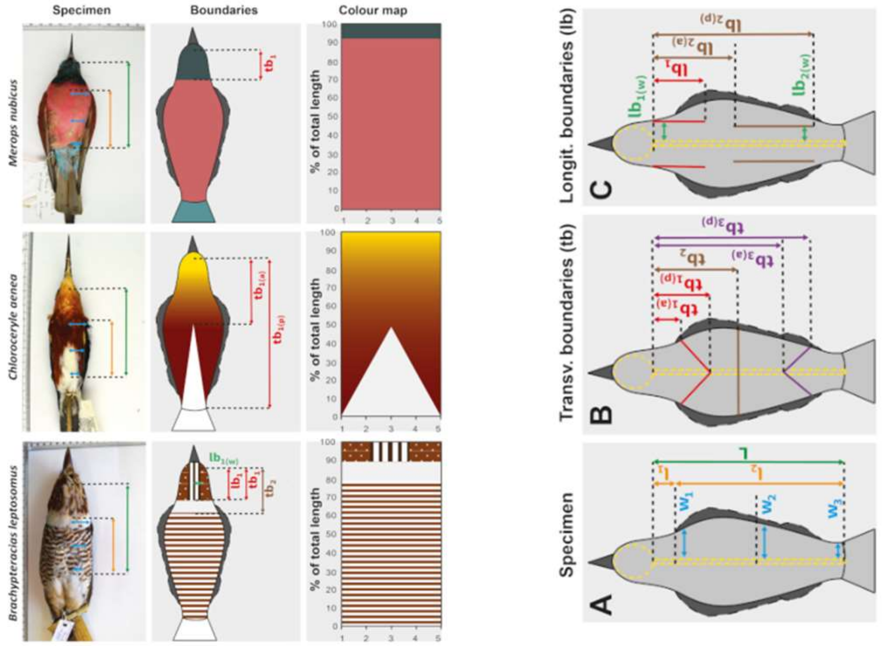
Drawings: Manceau Lab

Each bird species follows a general schema for pattern formation

But given the apparently overly complex coloration how do we even make sense of it?

It turns out that even though the appearance of birds is very diverse their color patterns can be systematized in a "color map."

For example the two finches that we see here, the Zebra finch and the Gouldian finch, may look very different, but their coloration is roughly divided in the same domains, separated by borders of roughly the same shape.

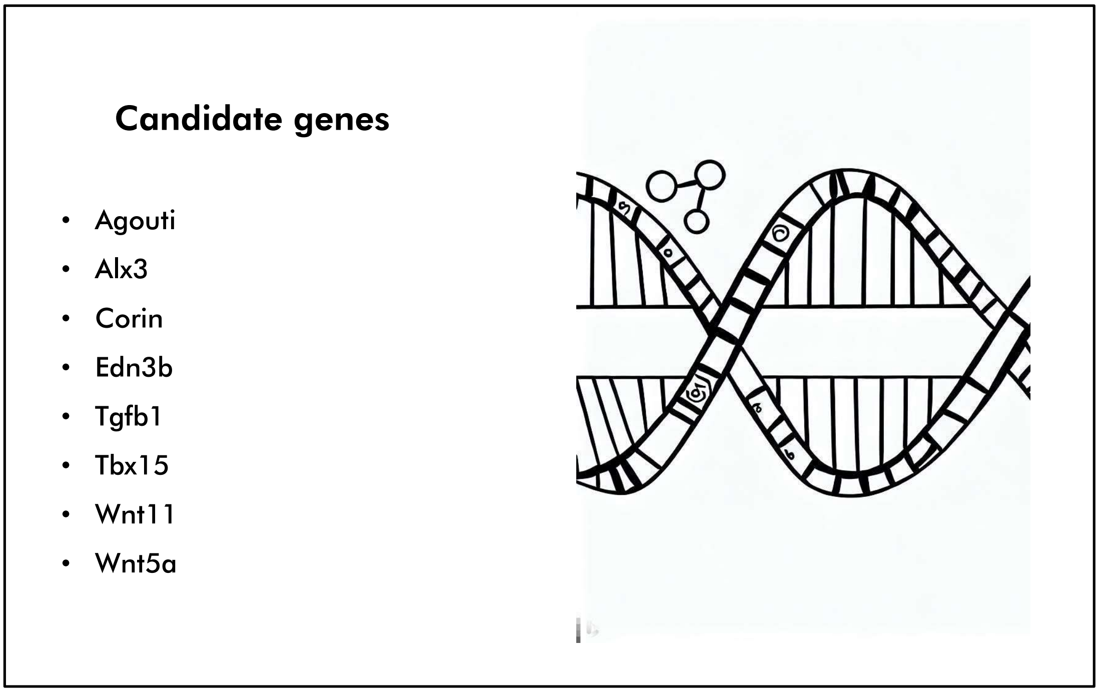


Drawings: Manceau Lab

We photographed and measured around 350 bird species representing all major avian families and recorded their color patterns by measuring the number of domains, the shape of the borders between them, the color of each domain, as well as specific characters such as the presence of stripes, patches, etc.

Our objective is to apply phylogenomic comparative methods and link this data to genomic sequences using information about the phylogeny of these species.

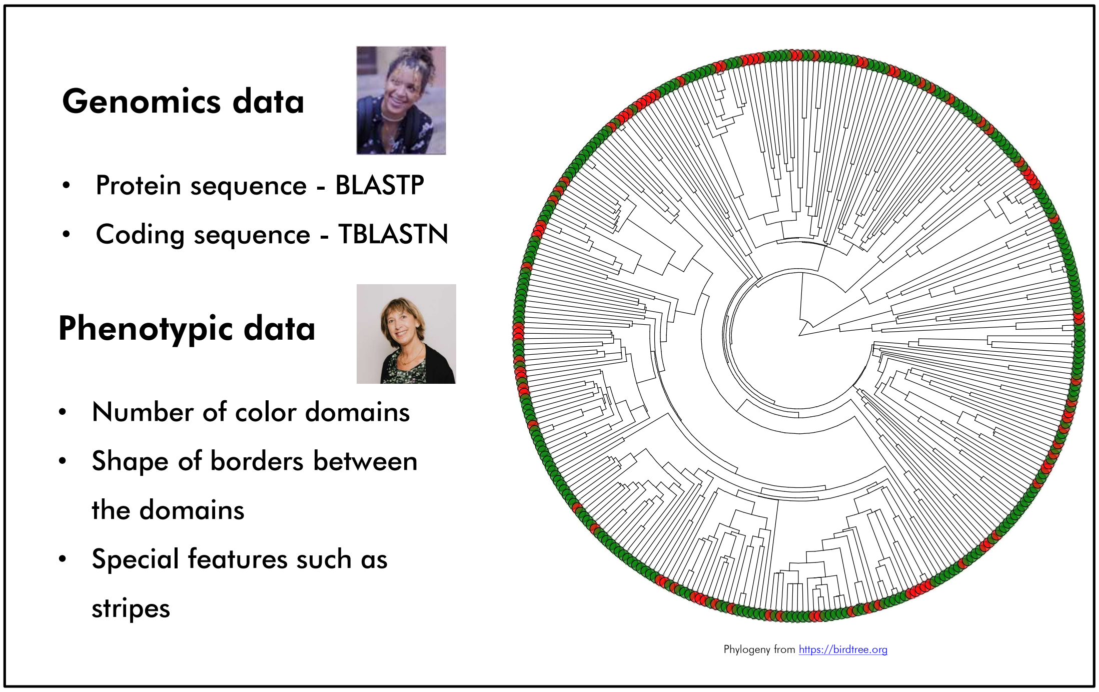
Thus, we want to see which genes are implicated in the evolution of distinct patterns on an evolutionary scale, which will then spur further experimental studies with bird embryos to identify the molecular mechanisms these genes use to form the patterns.



## Candidate genes

- Agouti
- Alx3
- Corin
- Edn3b
- Tgfb1
- Tbx15
- Wnt11
- Wnt5a

- *Agouti and Agouti-related Protein*, which controls the amount and distribution of eumelanin (brown/black) and pheomelanin (yellow/red) pigmentation in the mammalian coat.
- *Alx3*, which encodes a nuclear protein that functions as a transcriptional regulator involved in cell-type differentiation and development.
- *Edn3b*, which acts upstream of or within melanocyte differentiation and pigment cell development.



**Genomics data**

- Protein sequence - BLASTP
- Coding sequence - TBLASTN

**Phenotypic data**

- Number of color domains
- Shape of borders between the domains
- Special features such as stripes

Phylogeny from https://birdtree.org

The next step was to download genetic data for these genes for the 350 or so chosen bird species.

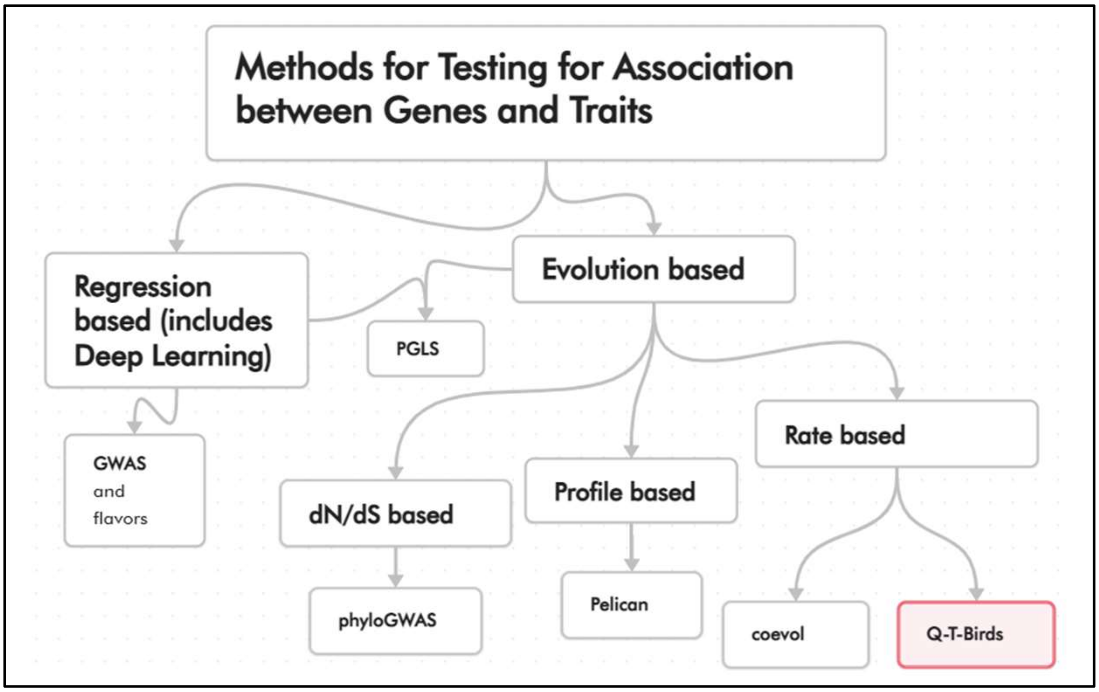
We made a choice to phenotype only such bird species for which there is publicly available genomic data.

We downloaded the protein genomes from NCBI and ran BLASTP with a reference sequence from the domestic chicken *Gallus gallus domesticus (L. 1758)*

Then we repeated the procedure with the coding DNA sequence, but used TBLASTN instead.

This data is then combined with the phenotypic data and placed on the tips of a phylogenetic tree downloaded from https://birdtree.org/.

To summarize: our data is a phylogenetic tree, the tips of which have been enriched with categorical variables representing a phenotypic character (e.g. number of color domains, presence of stripes, etc.) and with an aligned nucleic or protein sequence for the genes of interest.



Methods for Testing for Association between Genes and Traits

Regression based (includes Deep Learning)

Evolution based

PGLS

GWAS and flavors

dN/dS based

Profile based

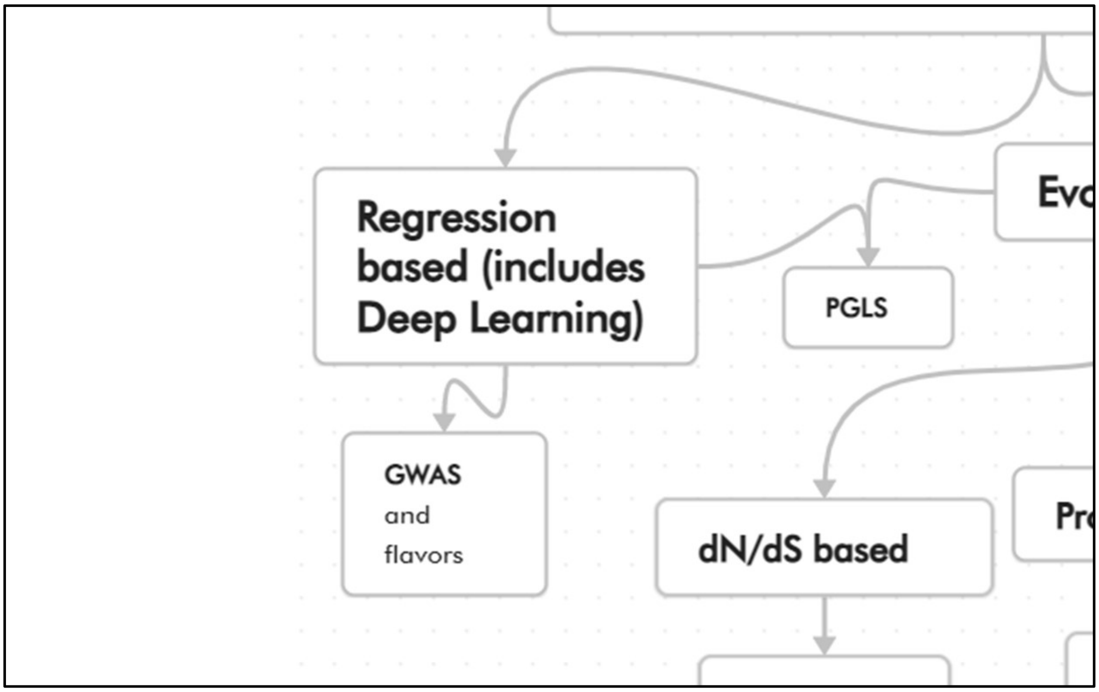
Rate based

phyloGWAS

Pelican

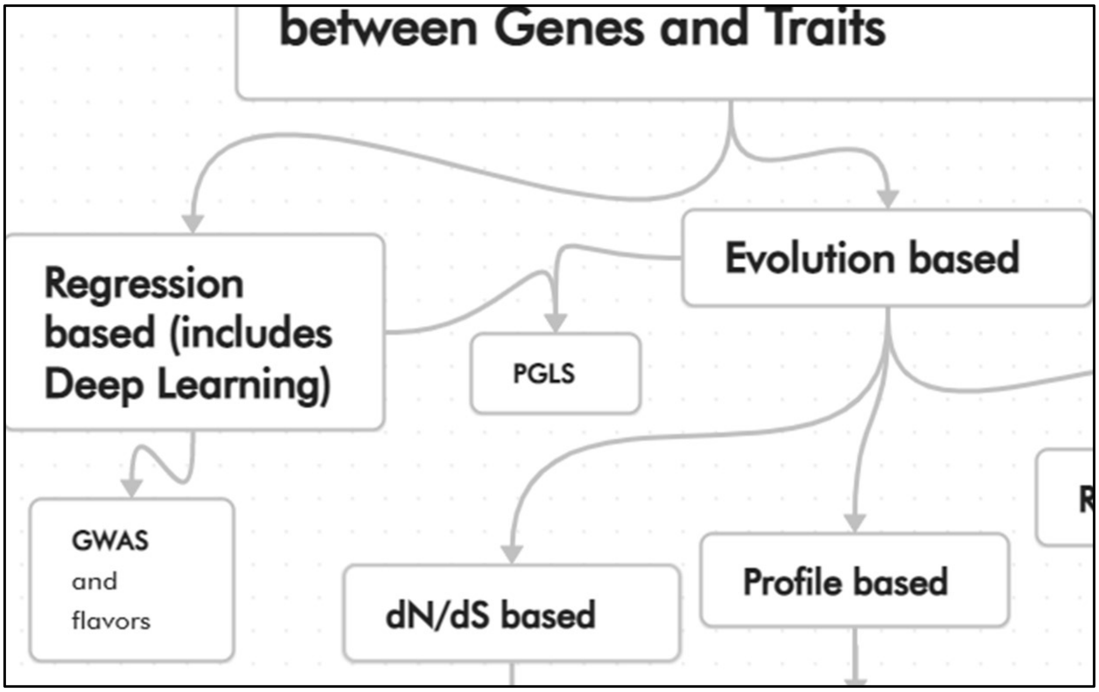
coevol

Q-T-Birds



*GWAS – genome-wide association study* is a regression method used often in biomedicine and relies on the assumption that the samples are independently identically distributed (i.i.d.).

In phylogenetic data the data points are clearly not i.i.d. as the evolutionary tree represents their degree of relatedness. For this and for other reasons such as phenotypic plasticity, genetic linkage, differential selection pressures, etc. models have been proposed that take the evolution process into account.



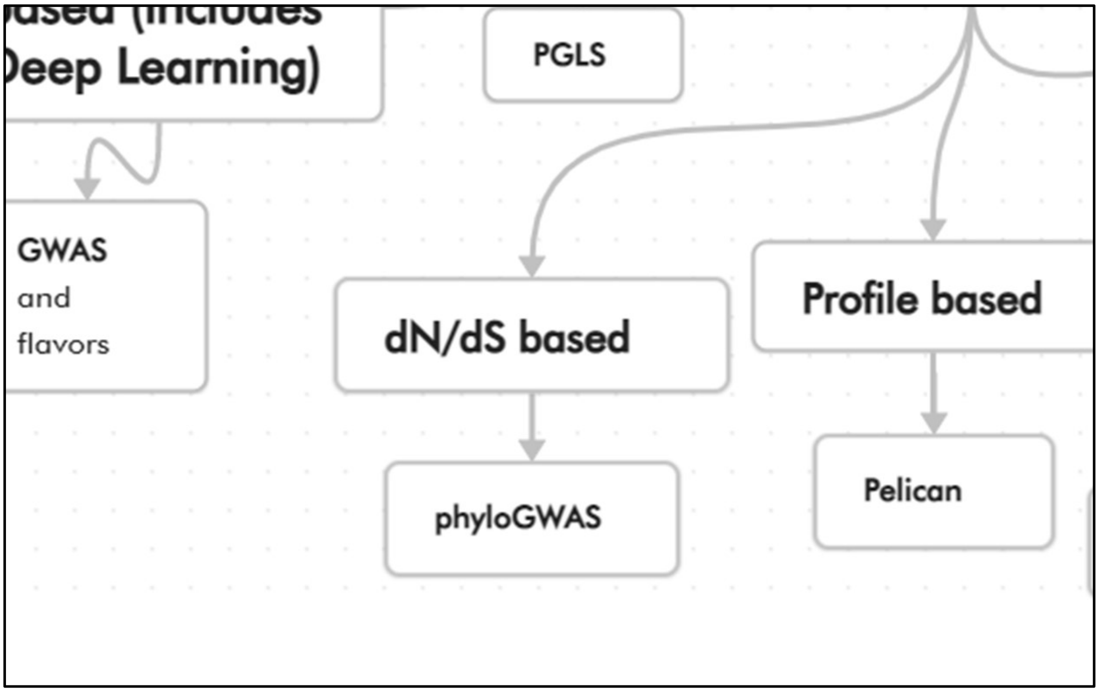
Regression based.

*PGLS (phylogenetic generalized least squares)* is a regression approach that accounts for the non-independence of the samples. However, it is not a mechanistic approach as it does not attempt to model the mechanism of the evolutionary process itself.

---

*Details:* Even though it can be, PGLS usually it is not used for testing gene-phenotype association but rather for figuring out the relationship between two traits (say body size and temperature) while controlling for the phylogenetic relationships in the data.

While regression can be used for prediction—i.e. given a new predictor value, we can compute a new response value—it is a hard problem as we need to estimate the magnitude of the effect of the predictor. This can be further confounded by the fact that the true relationship between the predictor and the response may be non-linear and unknown.

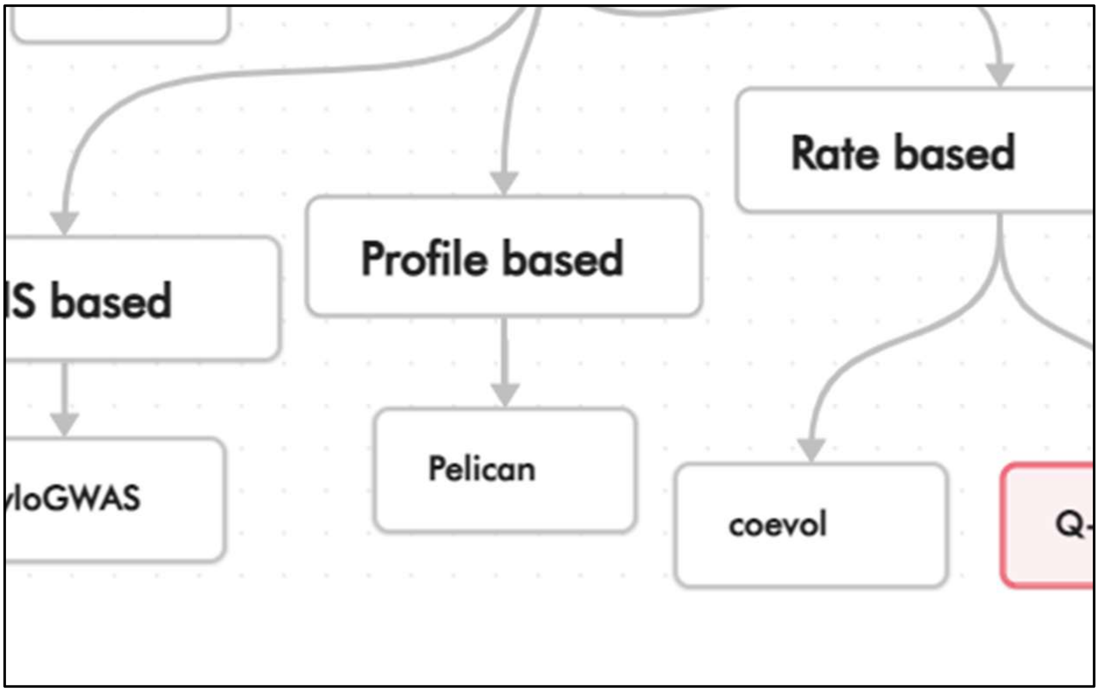


Methods based on the ratio of non-synonymous over synonymous mutations (dN/dS).

The ratio is used to detect genes that are under selection for the given phenotype of interest.  However, it is not always possible to tell which direction the selection has—i.e. whether the phenotype is beneficial or deleterious.  In order to account for that another family of methods have been proposed called…

---

*Details:* The dN/dS ratio is the ratio of protein changing over protein non-changing mutations on an evolutionary branch. A dN/dS of more than 1 suggests that the site is rapidly evolving. To link genotypes to phenotypes using dN/dS, we would look for genes with high dN/dS values in the phenotype of interest. However, high dN/dS ratios can indicate positive selection, neutral evolution, or even negative selection under certain scenarios, making the interpretation sometimes challenging.

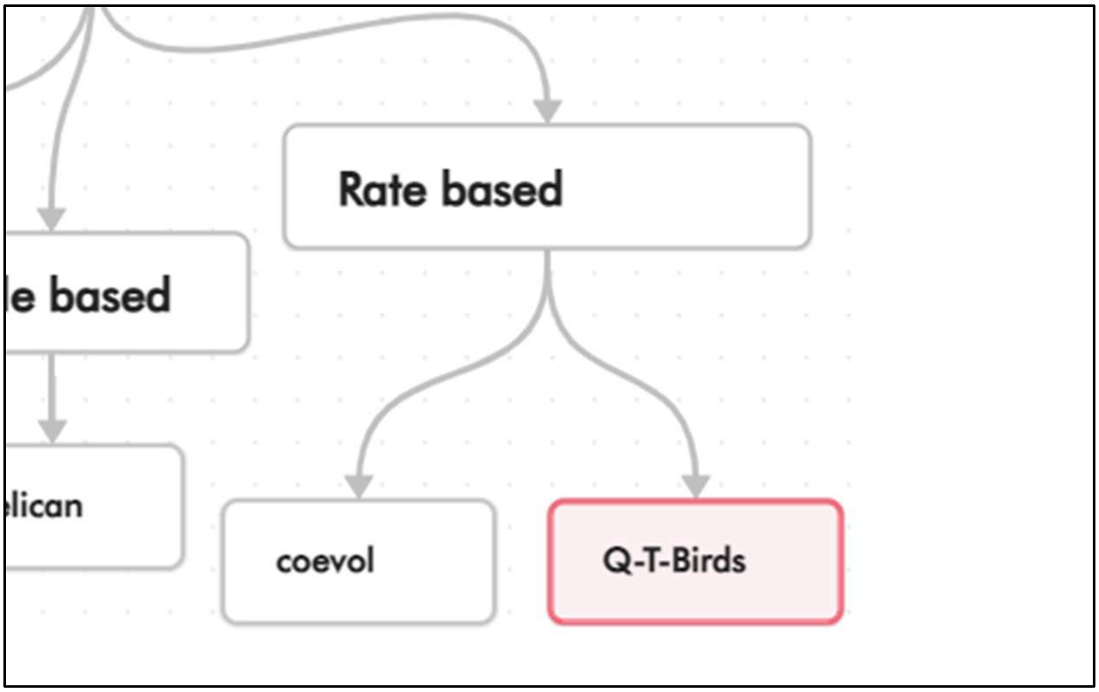


Profile-based methods.

An example of such methods is Pelican, which is a re-implementation of an older program, done by Douchemin et al., here in Lyon.  A disadvantage of Pelican is that it tests for association between sites and genotypes, not whole genes.

---

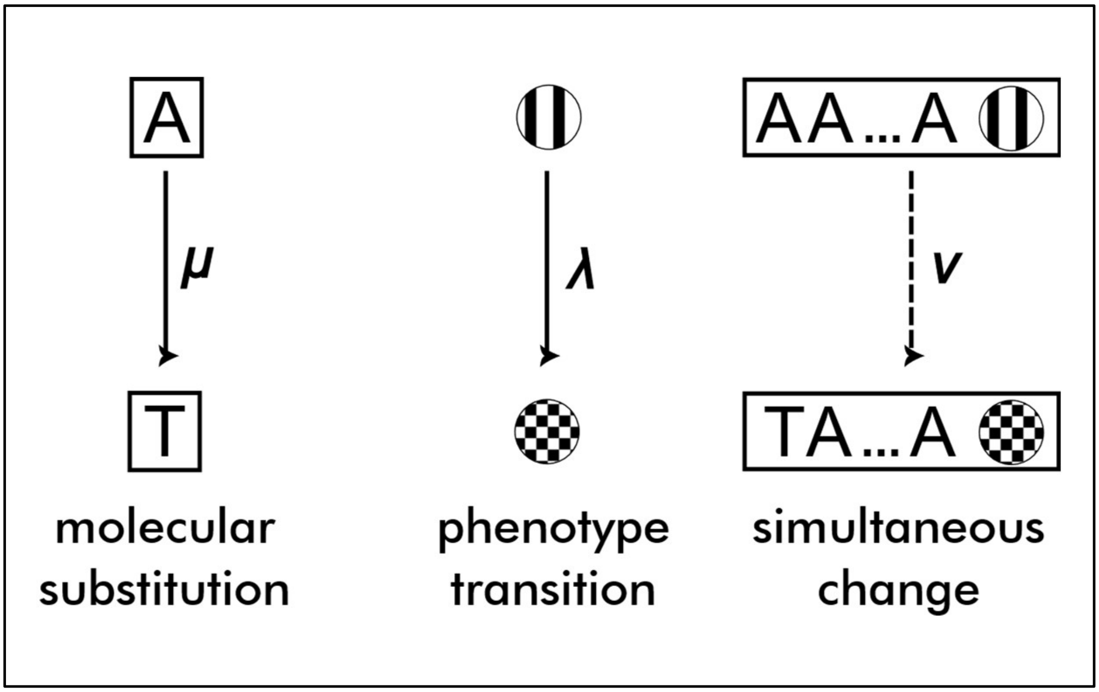
*Details:* In those methods a fitness profile is used to represent the relative fitness of each amino acid at a given site.



We believe that testing for the association between a whole gene and a phenotype may be more informative and we may detect even very weak signals, and that's why we have developed the QT-Birds method.

---

*Details:* our method belongs to a group of methods that link the phenotype to the evolutionary rate at which the genotypes evolve.



In the QT-Birds algorithm we model the association of genes and phenotypes directly by introducing three rates. For a given pair (gene, phenotyepe), we introduce:

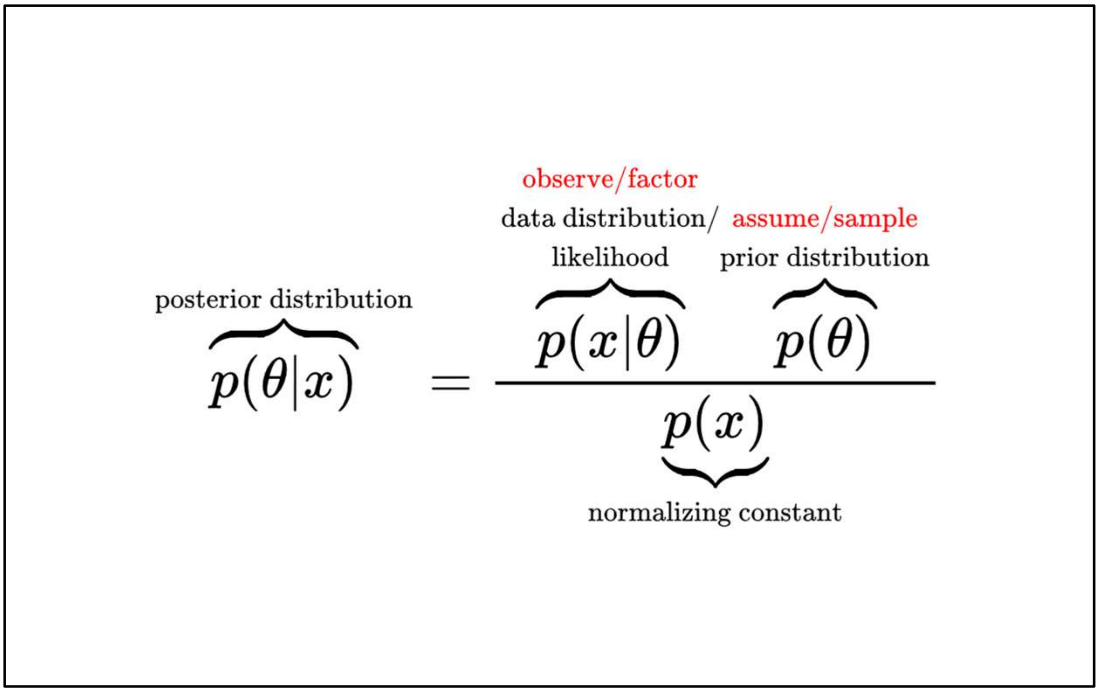
λ - rate of phenotypic change
μ - rate of molecular change
ν - rate of joint phenotypic-molecular change

We model the evolution of genotypes and phenotypes in time on the tree generatively using a Poisson process. Those rates are the parameters of the Poisson distributions. If a λ-change the phenotype shifts, but not the gene; if a μ-change occurs, on of the sites of the gene shifts; and, finally, if a ν-change occurs, then simultaneously the phenotype and the genotype change.

For genes and phenotypes that are not linked we expect to detect the rate ν to be close to 0.
For genes and phenotypes that are linked, we expect to detect higher ν-values.

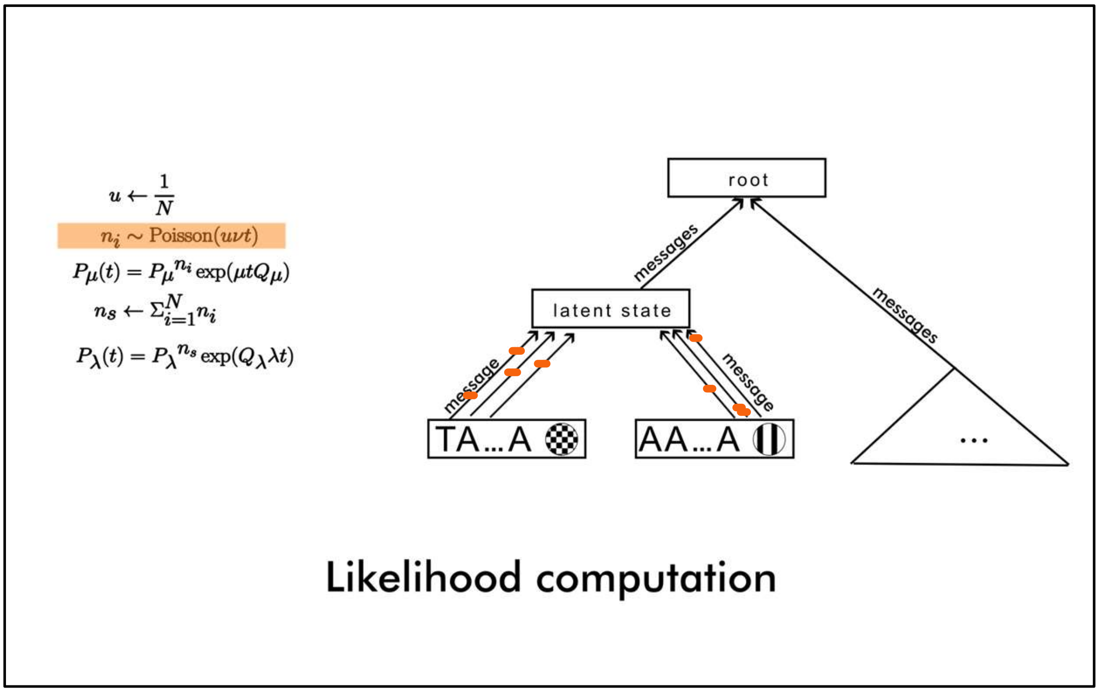
---

*Details:* In the present version of the algorithm we assume these rates to be constant on the tree, however they are unique for each site-character pair.



$$\overbrace{p(\theta|x)}^{\text{posterior distribution}} = \frac{\overbrace{p(x|\theta)}^{\substack{\text{observe/factor} \\ \text{data distribution/} \\ \text{likelihood}}} \quad \overbrace{p(\theta)}^{\substack{\text{assume/sample} \\ \text{prior distribution}}}}{\underbrace{p(x)}_{\text{normalizing constant}}}$$

To infer the rates we use Bayesian inference, which requires a way of specifying the conditional likelihood of the tree, given the observed data.



$$u \leftarrow \frac{1}{N}$$

$$n_i \sim \mathrm{Poisson}(u\nu t)$$

$$P_\mu(t) = {P_\mu}^{n_i} \exp(\mu t Q_\mu)$$

$$n_s \leftarrow \Sigma_{i=1}^{N} n_i$$

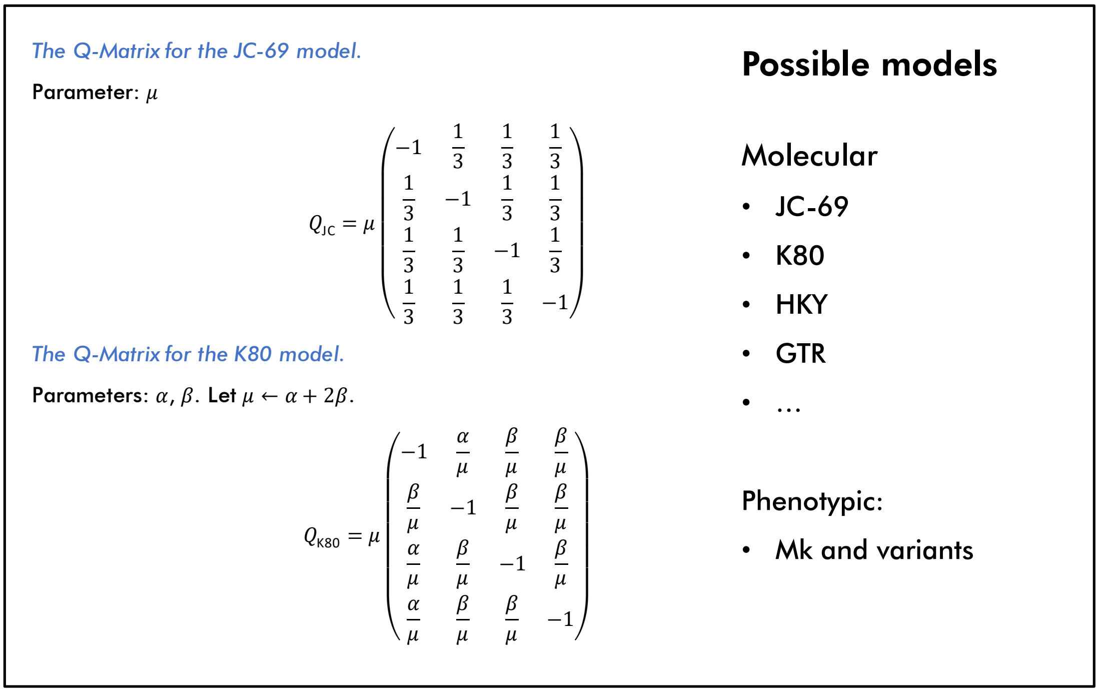
$$P_\lambda(t) = {P_\lambda}^{n_s} \exp(Q_\lambda \lambda t)$$

## Likelihood computation

But to specify the likelihood in a tractable way, we need to make use of a technique called data-augmentation.

We simulate the simultaneous jumps on the tree (orange dots on the graph) and then we use a message passing algorithm to update our belief about the probability distribution of the nucleotides and phenotypes in the latent nodes of the tree.

The messages are transition probability matrices after time t obtained from the continuous time Markov Chain process, and the data augmented discrete jumps.



*The Q-Matrix for the JC-69 model.*

Parameter: $\mu$

$$Q_{\mathrm{JC}} = \mu \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}$$

*The Q-Matrix for the K80 model.*

Parameters: $\alpha, \beta$. Let $\mu \leftarrow \alpha + 2\beta$.

$$Q_{\mathrm{K80}} = \mu \begin{pmatrix} -1 & \frac{\alpha}{\mu} & \frac{\beta}{\mu} & \frac{\beta}{\mu} \\ \frac{\beta}{\mu} & -1 & \frac{\beta}{\mu} & \frac{\beta}{\mu} \\ \frac{\alpha}{\mu} & \frac{\beta}{\mu} & -1 & \frac{\beta}{\mu} \\ \frac{\alpha}{\mu} & \frac{\beta}{\mu} & \frac{\beta}{\mu} & -1 \end{pmatrix}$$

## Possible models

Molecular

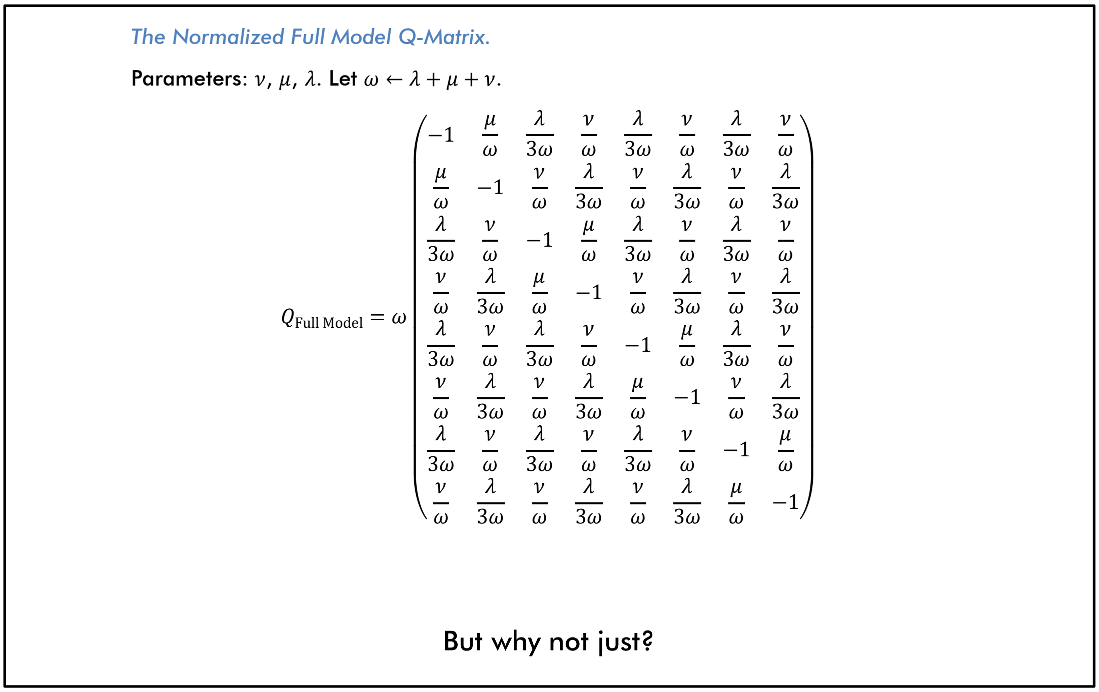
- JC-69
- K80
- HKY
- GTR
- …

Phenotypic:

- Mk and variants

Note that the individual processes are not constrained to symmetric processes such as Jukes-Cantor or Mk. In fact λ and μ are the total rates, while the individual transitions may have their own rates. In the example I've shown that, instead of the Jukes-Cantor model, you can use the Kimura model, which provides for different substitution rates for transitions (within pyrimidine and within purine changes) and transversions (changes between a pyrimidine and a purine).

The total joint transition rate is ν is not broken up into further rates, however. If a simultaneous jump is sampled from it, then the actual transitions that occur are governed by the instantaneous probability matrices obtained by the molecular and the phenotypic processes. Another assumption is that the molecular rate does not differ between sites of the same gene, although it can differ between genes.



*The Normalized Full Model Q-Matrix.*

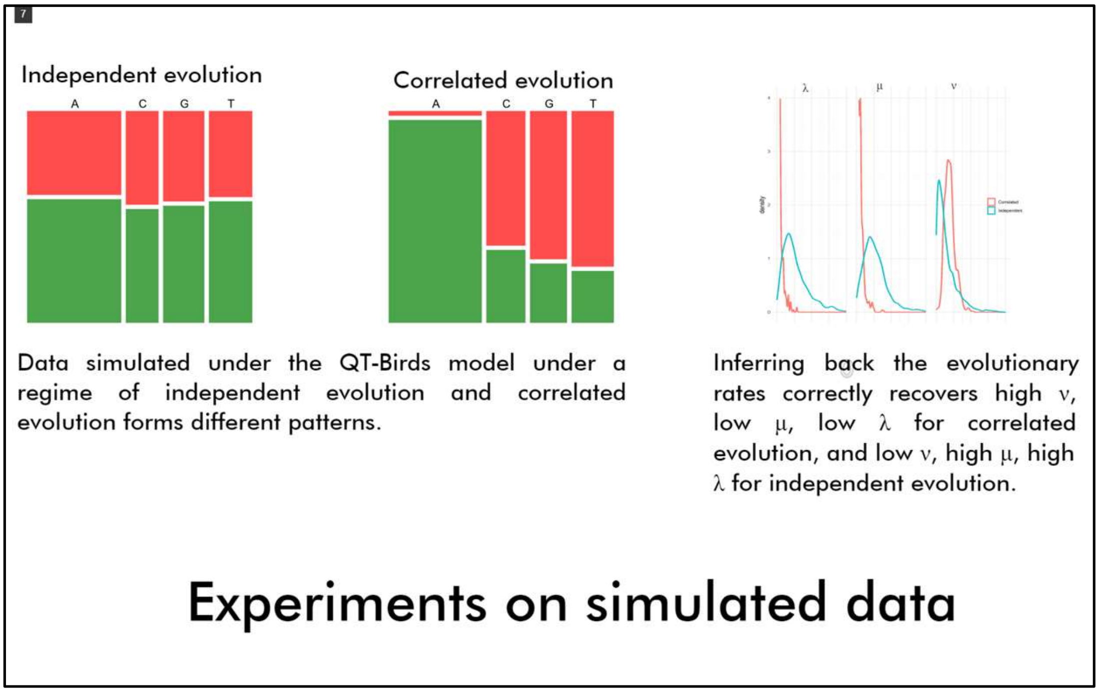
Parameters: $\nu$, $\mu$, $\lambda$. Let $\omega \leftarrow \lambda + \mu + \nu$.

$$Q_{\text{Full Model}} = \omega \begin{pmatrix} -1 & \frac{\mu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} \\ \frac{\mu}{\omega} & -1 & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} \\ \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & -1 & \frac{\mu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} \\ \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\mu}{\omega} & -1 & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} \\ \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & -1 & \frac{\mu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} \\ \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\mu}{\omega} & -1 & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} \\ \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & -1 & \frac{\mu}{\omega} \\ \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\nu}{\omega} & \frac{\lambda}{3\omega} & \frac{\mu}{\omega} & -1 \end{pmatrix}$$

But why not just?

If we didn’t use the data augmentation I described before, we would need to compute a large and intractable Q-matrix for the joint process.

Here I am giving the Q-Matrix for a 2state Mk model combined with a 4 state JC-69 model for a single nucleotide. This is an 8x8 matrix. However, as the length of the locus increases, the dimensionality of the full Q-matrix becomes intractable as it is the product of the dimensions of the individual sites and the character. For example for a gene of length 100 and a binary character, the dimension will be 2^201.



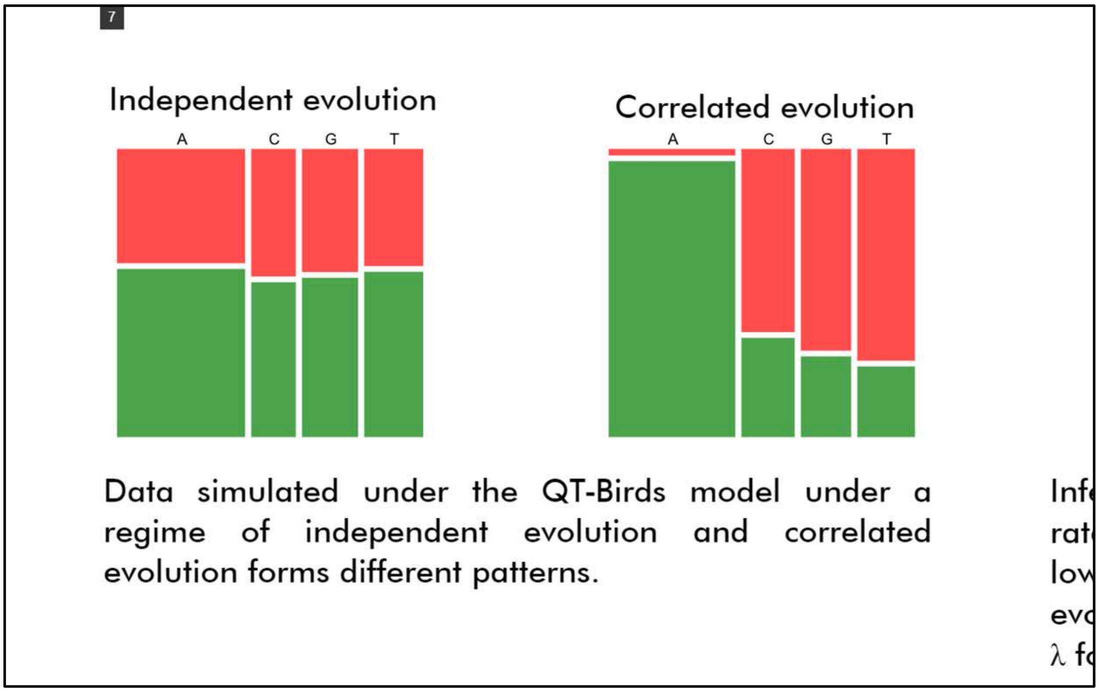
Independent evolution

Correlated evolution

Data simulated under the QT-Birds model under a regime of independent evolution and correlated evolution forms different patterns.

Inferring back the evolutionary rates correctly recovers high ν, low μ, low λ for correlated evolution, and low ν, high μ, high λ for independent evolution.

# Experiments on simulated data

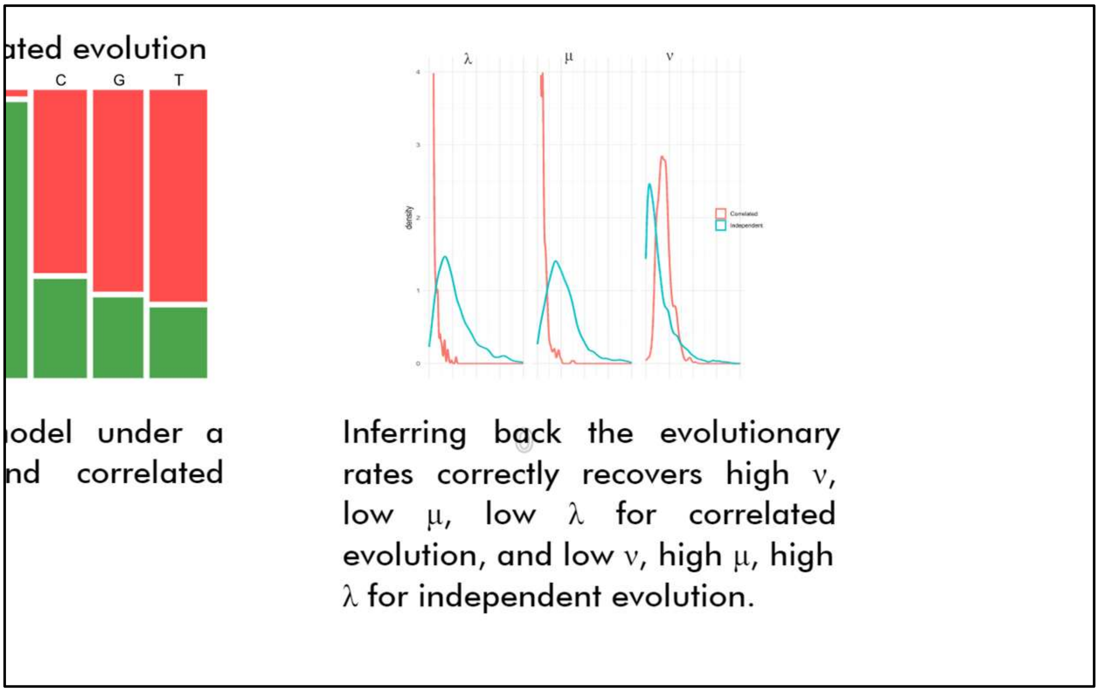


Independent evolution

Correlated evolution

Data simulated under the QT-Birds model under a regime of independent evolution and correlated evolution forms different patterns.

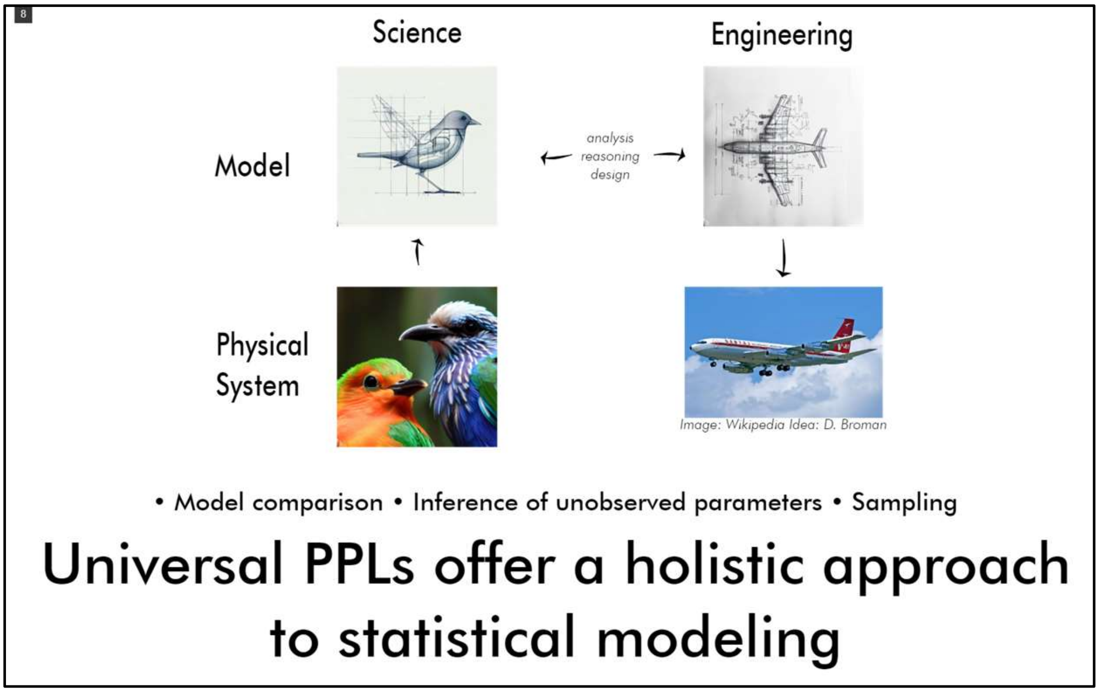
Inf
rat
low
evo
λ f



ated evolution

odel under a
nd correlated

Inferring back the evolutionary rates correctly recovers high ν, low μ, low λ for correlated evolution, and low ν, high μ, high λ for independent evolution.



8

Science

Engineering

Model

← analysis reasoning design →

Physical System

Image: Wikipedia Idea: D. Broman

• Model comparison • Inference of unobserved parameters • Sampling

**Universal PPLs offer a holistic approach to statistical modeling**

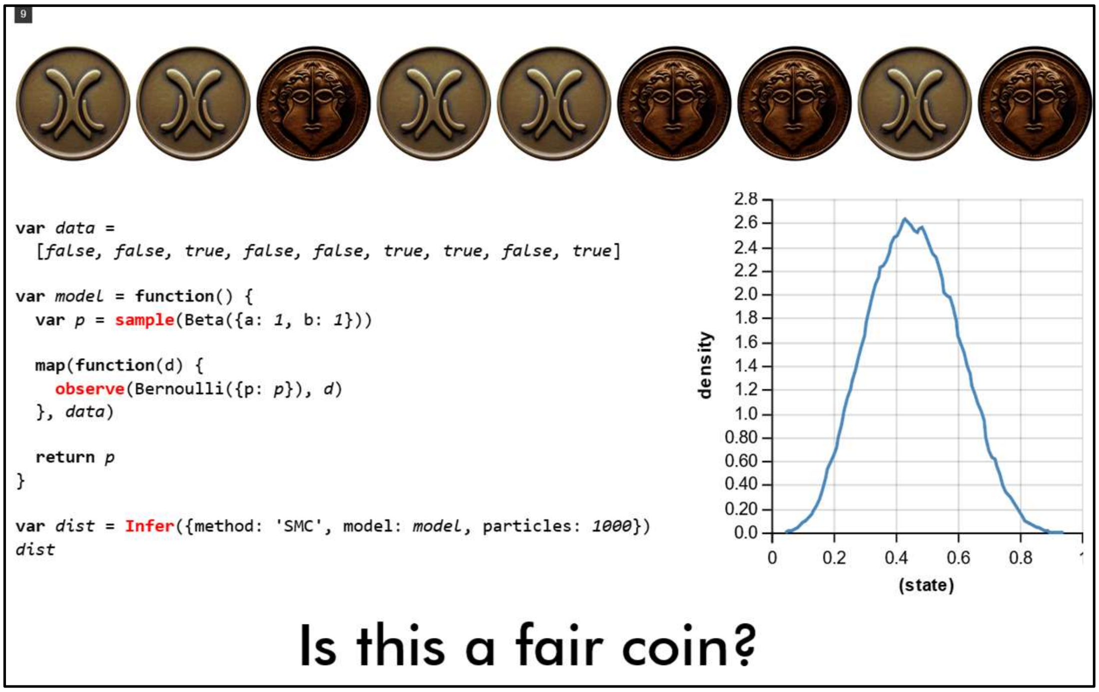
But let's step back for a minute and discuss how we use models to learn about unknown things.

Often it is the case that we have a physical system that we would like to study. And even though sometimes we can build the system and study it directly, often it is not possible or too costly. That's why we create a model of the system and we do our analysis, reasoning, and design on the model.

In our case the evolutionary past is long gone and we only see shadows of it in the fossil record, in the DNA, and in the present day phenotypes. So we need to create a model.

A probabilistic programming language is a special kind of programming language. A program written in a PPL exactly encodes a statistical probability distribution, conditioned on observed data and a given model.

Furthermore, using advanced compiler techniques, without any extra effort but writing the generative model we can sample from it, do model comparison, and infer unobserved parameters.

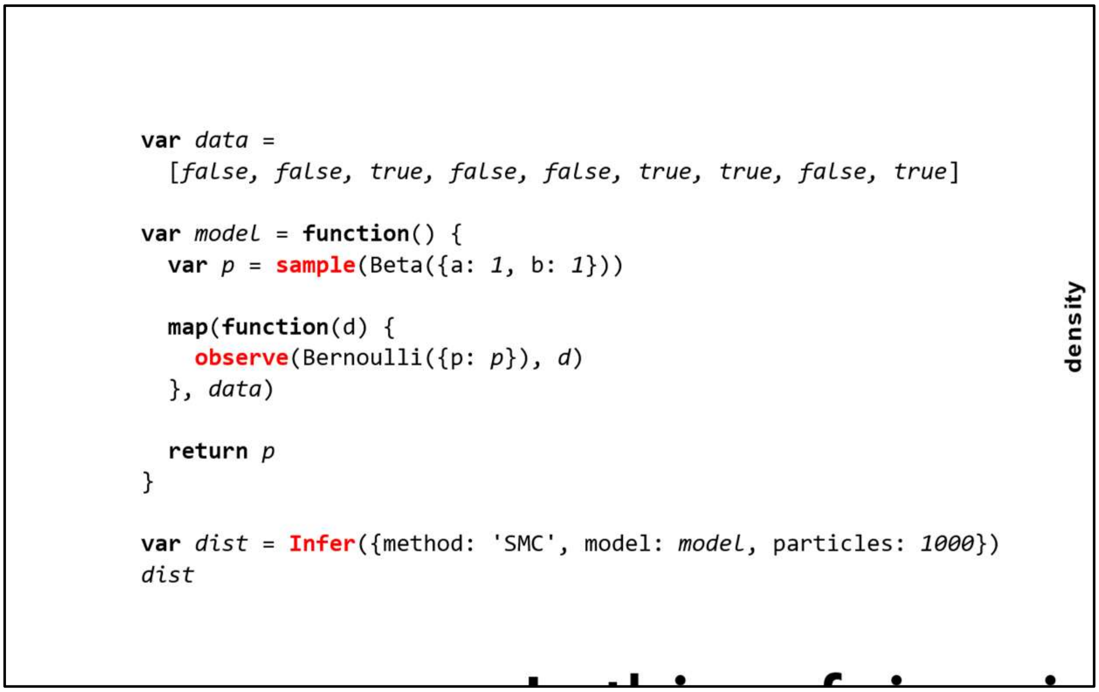


```
var data =
  [false, false, true, false, false, true, true, false, true]

var model = function() {
  var p = sample(Beta({a: 1, b: 1}))

  map(function(d) {
    observe(Bernoulli({p: p}), d)
  }, data)

  return p
}

var dist = Infer({method: 'SMC', model: model, particles: 1000})
dist
```

Is this a fair coin?

To give you a taste of what a PPL looks in practice, let's take a look at the "Hello, world" of probabilistic programming, namely is a given coin fair.

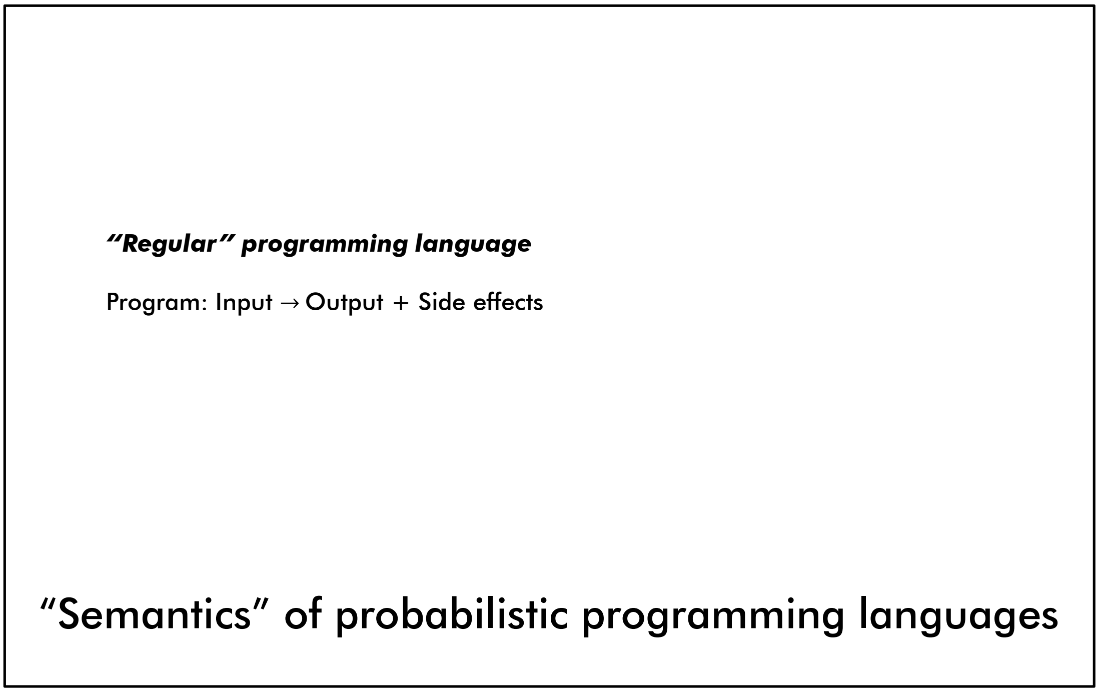
A PPL looks just like any regular language -- this one is called WebPPL and happens to look like JavaScript -- with the addition of probabilistic constructs.



```
var data =
  [false, false, true, false, false, true, true, false, true]

var model = function() {
  var p = sample(Beta({a: 1, b: 1}))

  map(function(d) {
    observe(Bernoulli({p: p}), d)
  }, data)

  return p
}

var dist = Infer({method: 'SMC', model: model, particles: 1000})
dist
```

sample - to generate samples from a prior distribution
observe - to reweigh or condition the likelihood of the computation on observed data
Infer - a higher order function to utilize a number of statistical sampling procedures -- here Sequential Monte Carlo -- to compute the posterior distribution of the hidden parameters under the model, conditioned on our data

An important thing to note here is that the user does not encode the SMC -- it is provided automatically by the compiler.

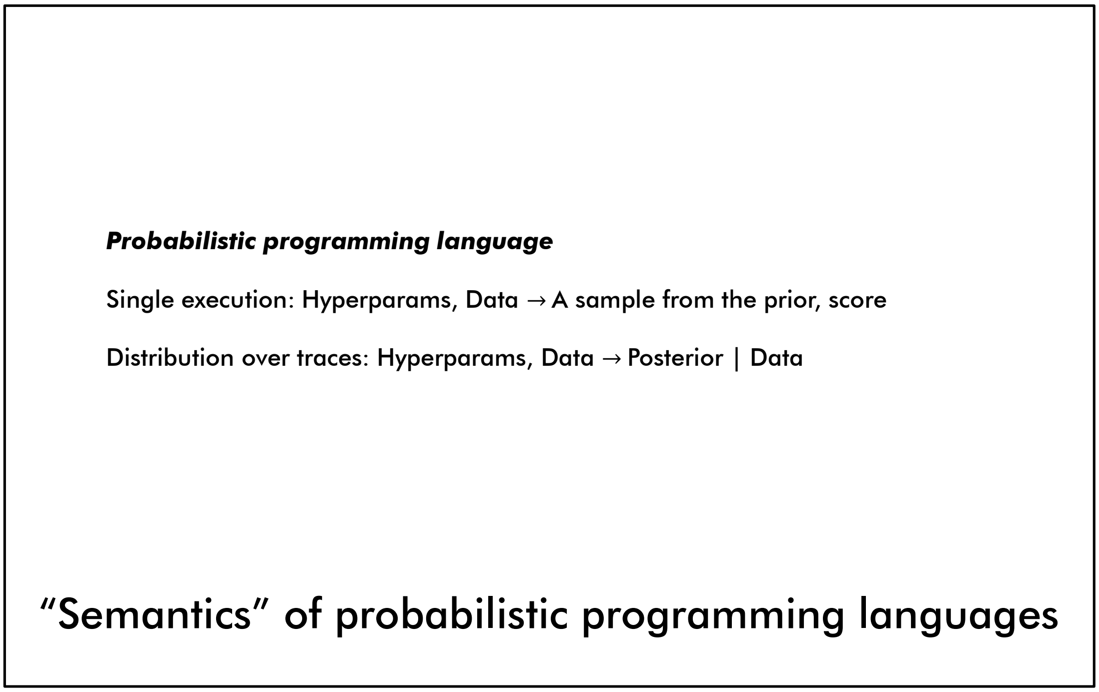


***"Regular" programming language***

Program: Input → Output + Side effects

## "Semantics" of probabilistic programming languages

Note that it is very easy to tweak a probabilistic program to simply simulate data under the model, i.e. do a prior simulation.

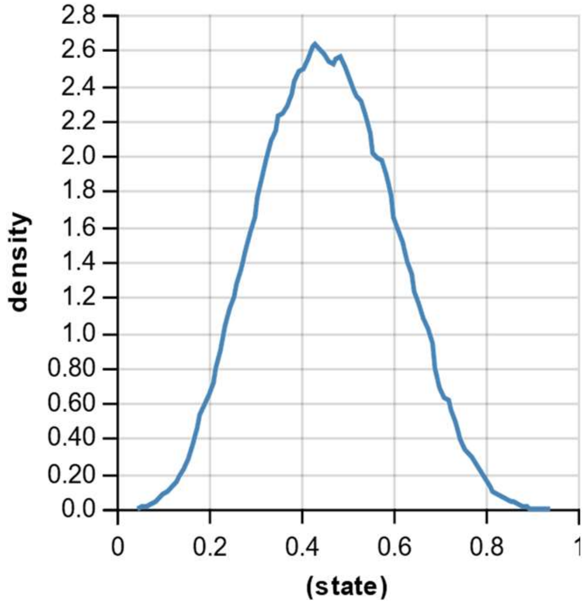


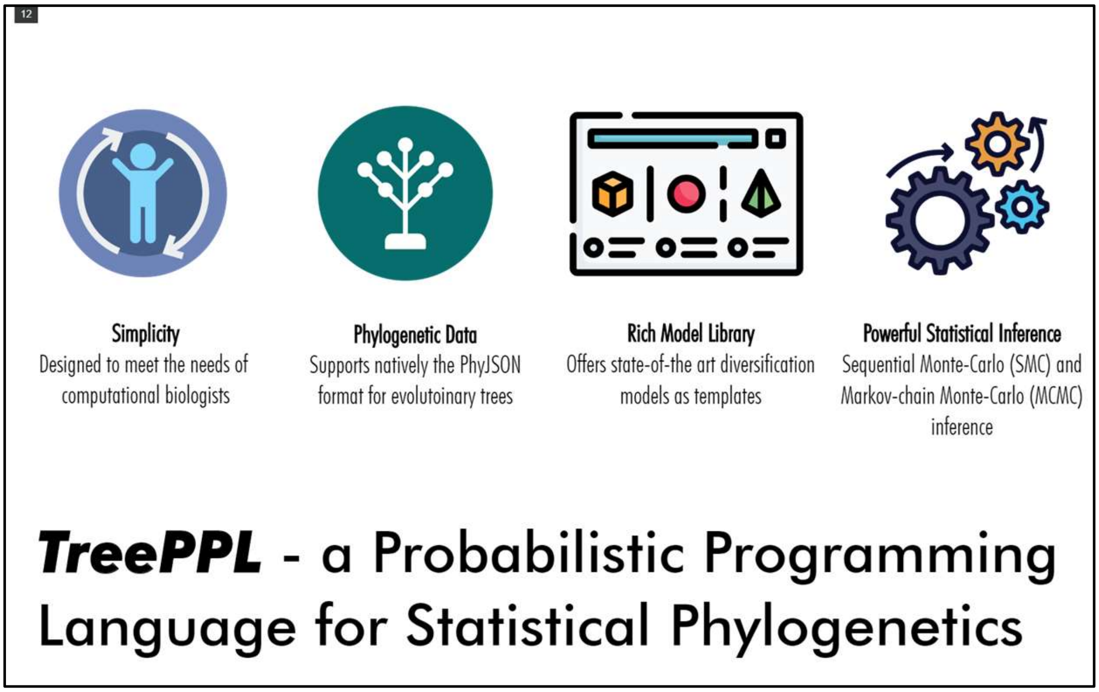
***Probabilistic programming language***

Single execution: Hyperparams, Data → A sample from the prior, score

Distribution over traces: Hyperparams, Data → Posterior | Data

# “Semantics” of probabilistic programming languages


density
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.80
0.60
0.40
0.20
0.0
0
0.2
0.4
0.6
0.8
1
(state)



12

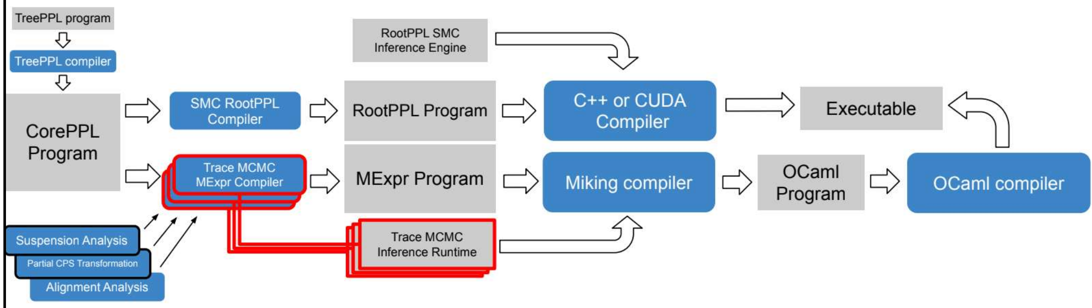
**Simplicity**
Designed to meet the needs of computational biologists

**Phylogenetic Data**
Supports natively the PhyJSON format for evolutoinary trees

**Rich Model Library**
Offers state-of-the art diversification models as templates

**Powerful Statistical Inference**
Sequential Monte-Carlo (SMC) and Markov-chain Monte-Carlo (MCMC) inference

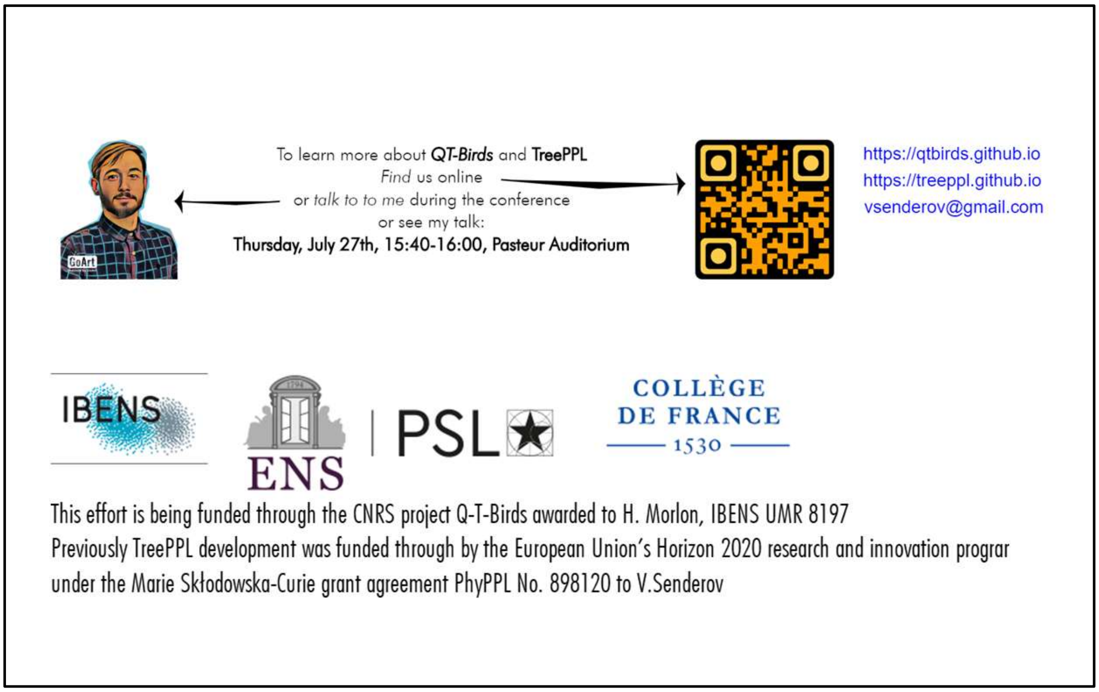
# ***TreePPL*** - a Probabilistic Programming Language for Statistical Phylogenetics



Conjugacy analysis
(aumatic static delayed sampling)

Lightweight importance sampling
Sequential Monte Carlo: Bootstrap Particle Filter,
Sequential Monte Carlo: Alive particle filter
Lightweight MCMC
Trace MCMC
Particle MCMC

# Bonus material



To learn more about ***QT-Birds*** and **TreePPL**
*Find* us online
or *talk to to me* during the conference
or see my talk:
**Thursday, July 27th, 15:40-16:00, Pasteur Auditorium**

https://qtbirds.github.io
https://treeppl.github.io
vsenderov@gmail.com

GoArt
IBENS
ENS
PSL
COLLÈGE DE FRANCE 1530

This effort is being funded through the CNRS project Q-T-Birds awarded to H. Morlon, IBENS UMR 8197
Previously TreePPL development was funded through by the European Union's Horizon 2020 research and innovation prograr
under the Marie Skłodowska-Curie grant agreement PhyPPL No. 898120 to V.Senderov