Une approche Bayésienne de l'étude des populations bactériennes pour données méta-génomiques

David Abrial,Xavier Bailly

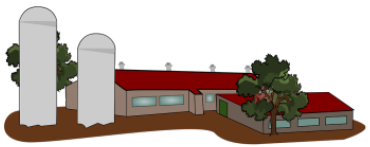
INRA de Clermont-Ferrand
Unité Epia - CATI IMOTEP

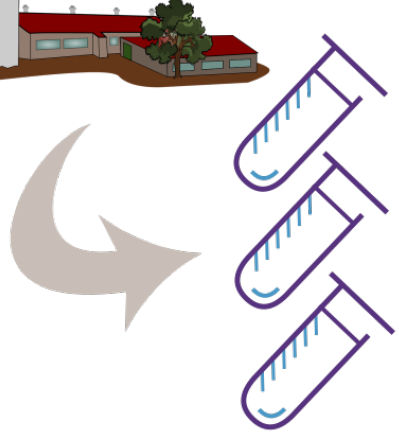
Mars 2019

Plan

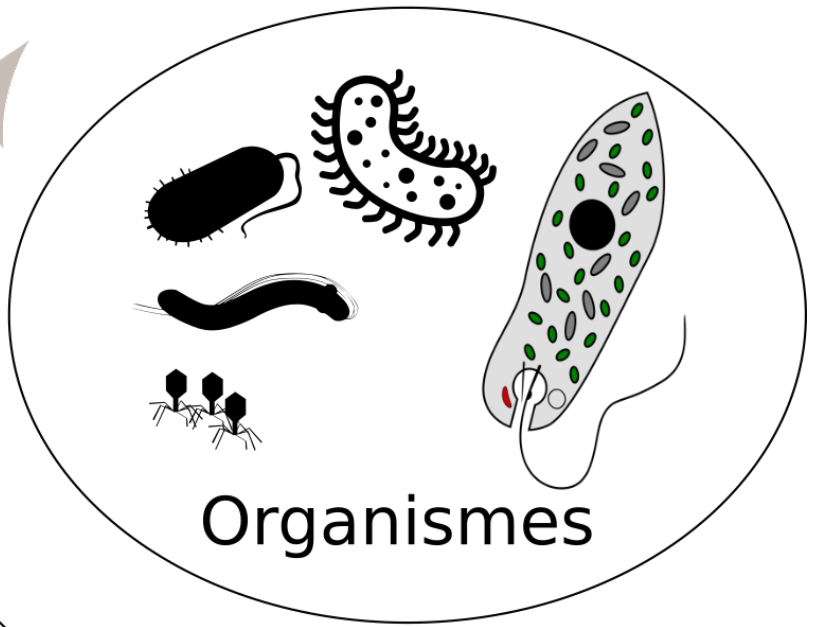
- Motivations
- Notre modèle
- État d'avancement et perspectives

Approche méta-génomique

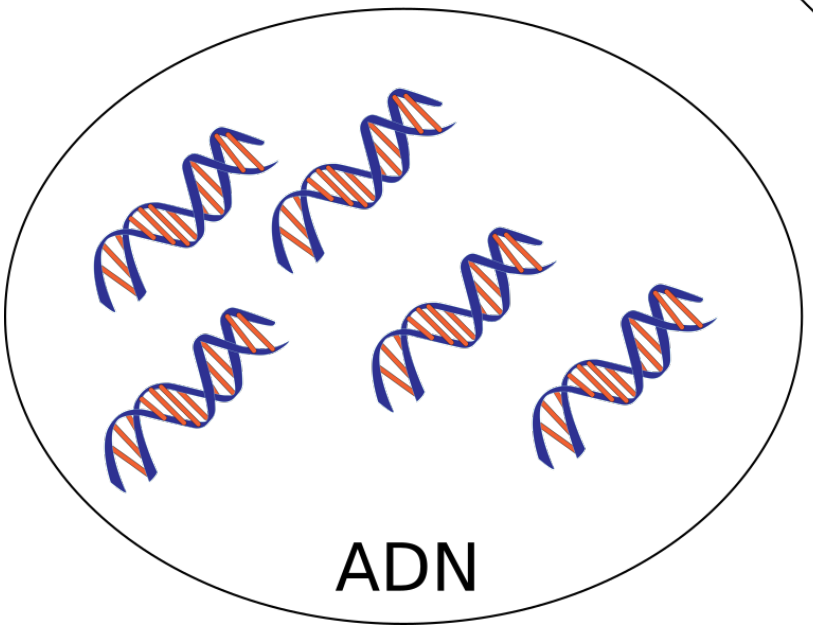




Echantillons



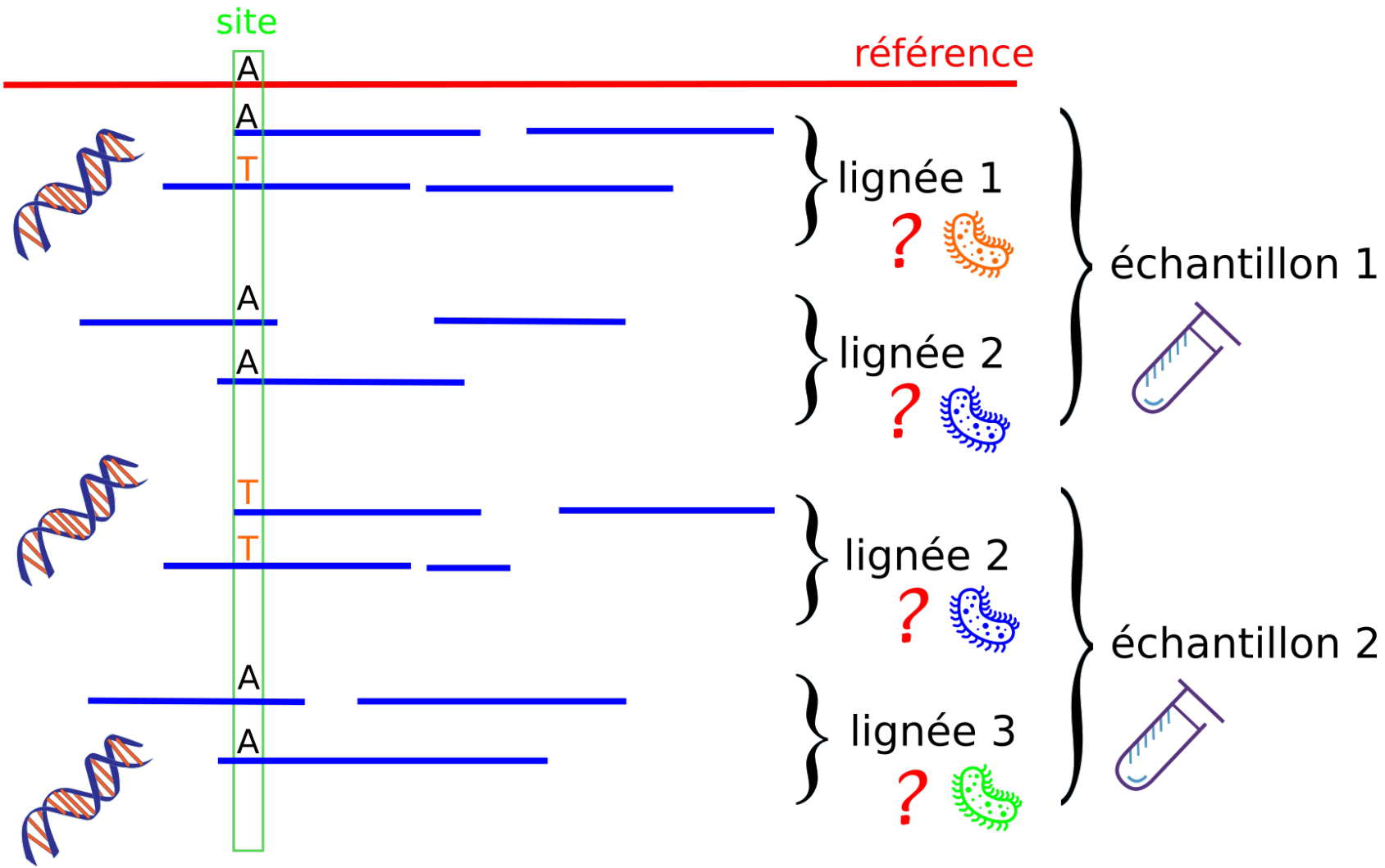
Organismes



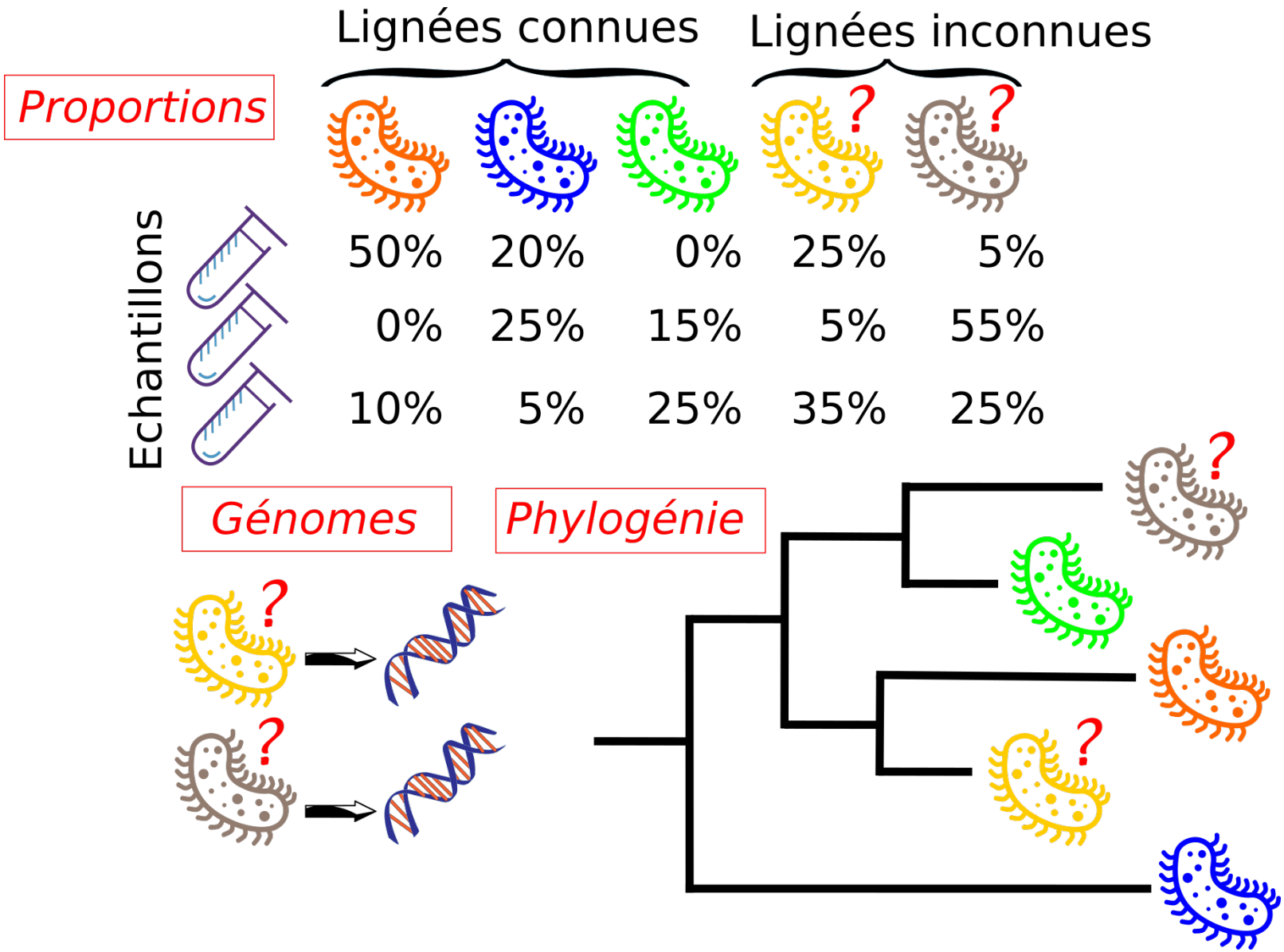
ADN

- extraction
- purification
- séquençage

Les fréquences alléliques



Les objectifs



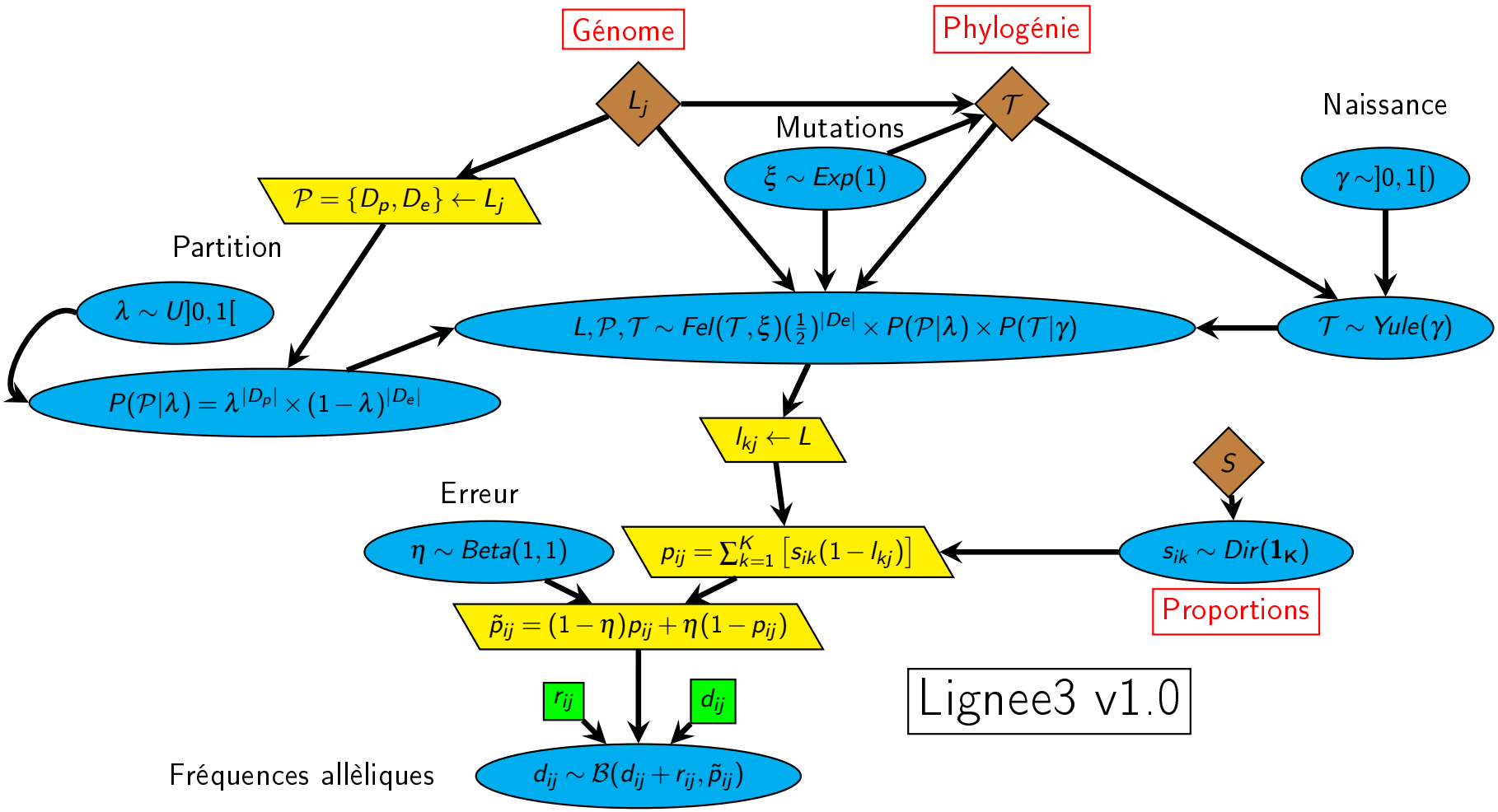
Un modèle hiérarchique bayésien 1/2

O'Brien et al. 2014 *'A bayesian approach to inferring the phylogenetic structure of communities from metagenomic data'*, **Genetics, 197,925-937.**

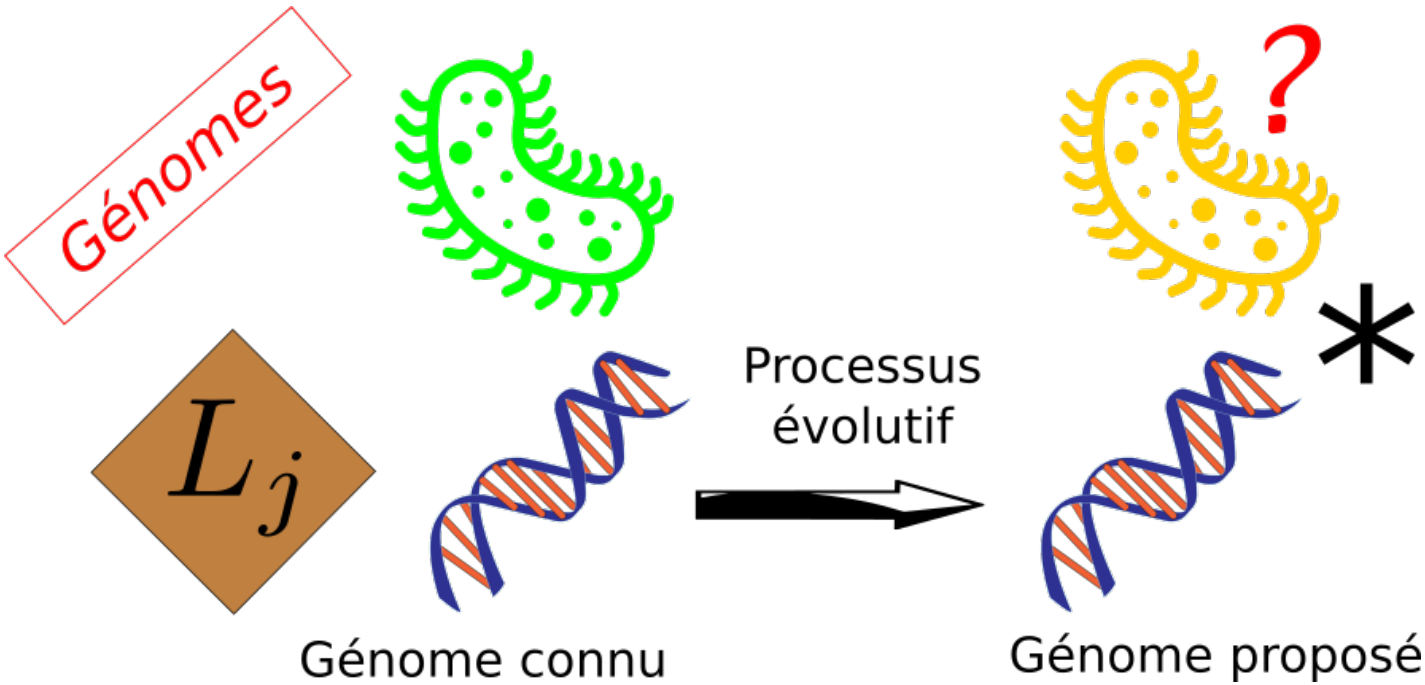
Modifications :

- le génome binaire est remplacé par un codage en nucléotides
- le triplet génome/arbre/partition est proposé puis accepté ou rejeté dans son ensemble
- un modèle évolutif est utilisé pour proposer un génome
- pour l'arbre, un *a priori* basé sur un processus de pure naissance (Yule) remplace le coalescent

Notre modèle hiérarchique bayésien 2/2

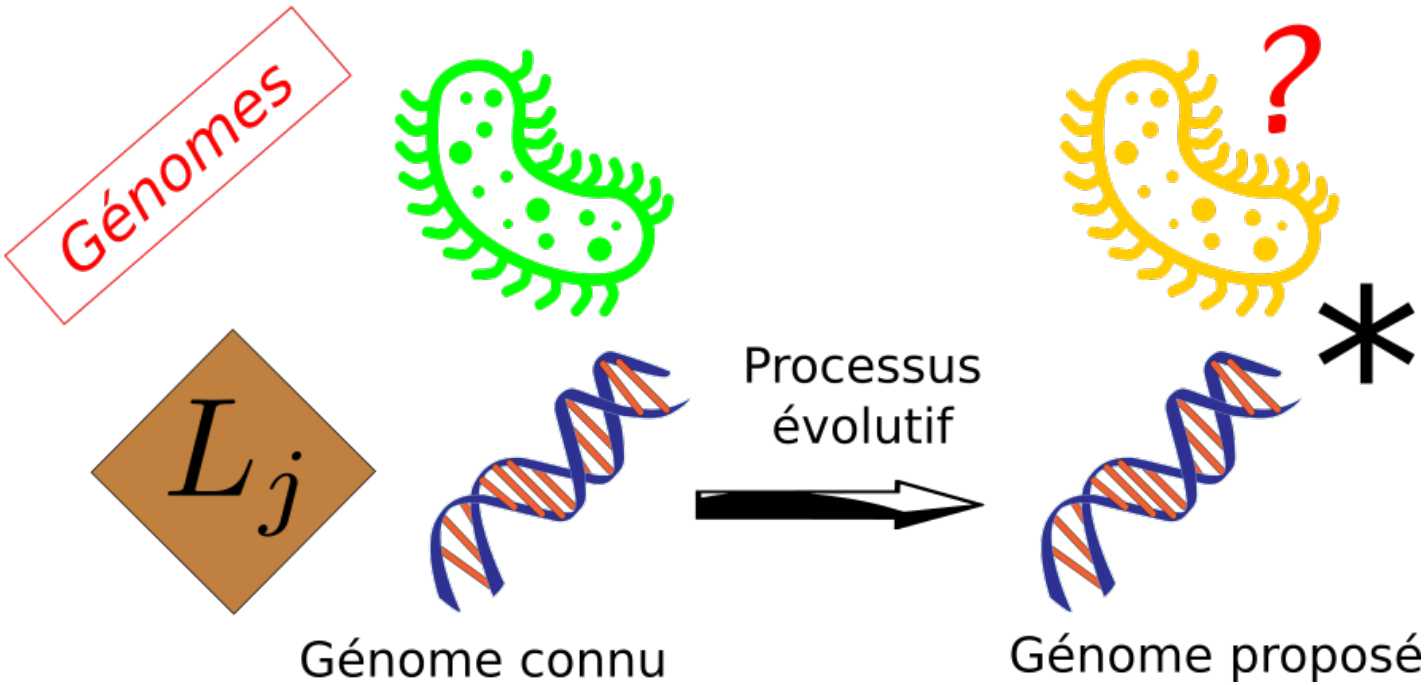


Proposition génotype 1/2



- chaque site évolue selon un poids ϕ combinaison de l'**entropie de Shannon** et d'une pondération de site
- l'entropie prend en compte la variabilité des fréquences alléliques
- la pondération de site favorise les sites variants dans l'ADN des échantillons et pas celui des lignées observées

Proposition génotype 2/2

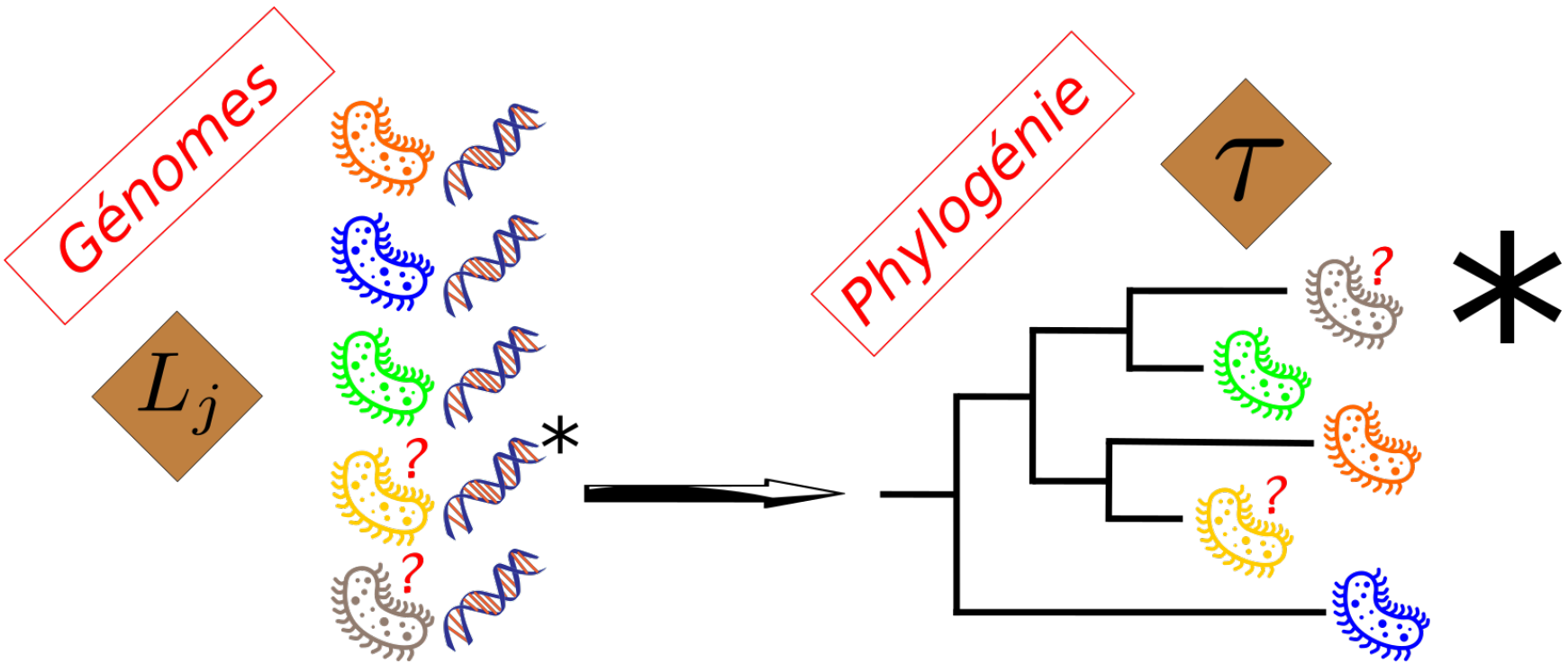


- le **modèle évolutif** est celui de Juke et Cantor 1969

$$P_{x_j} = exp(\phi Qt)_{x_j}$$

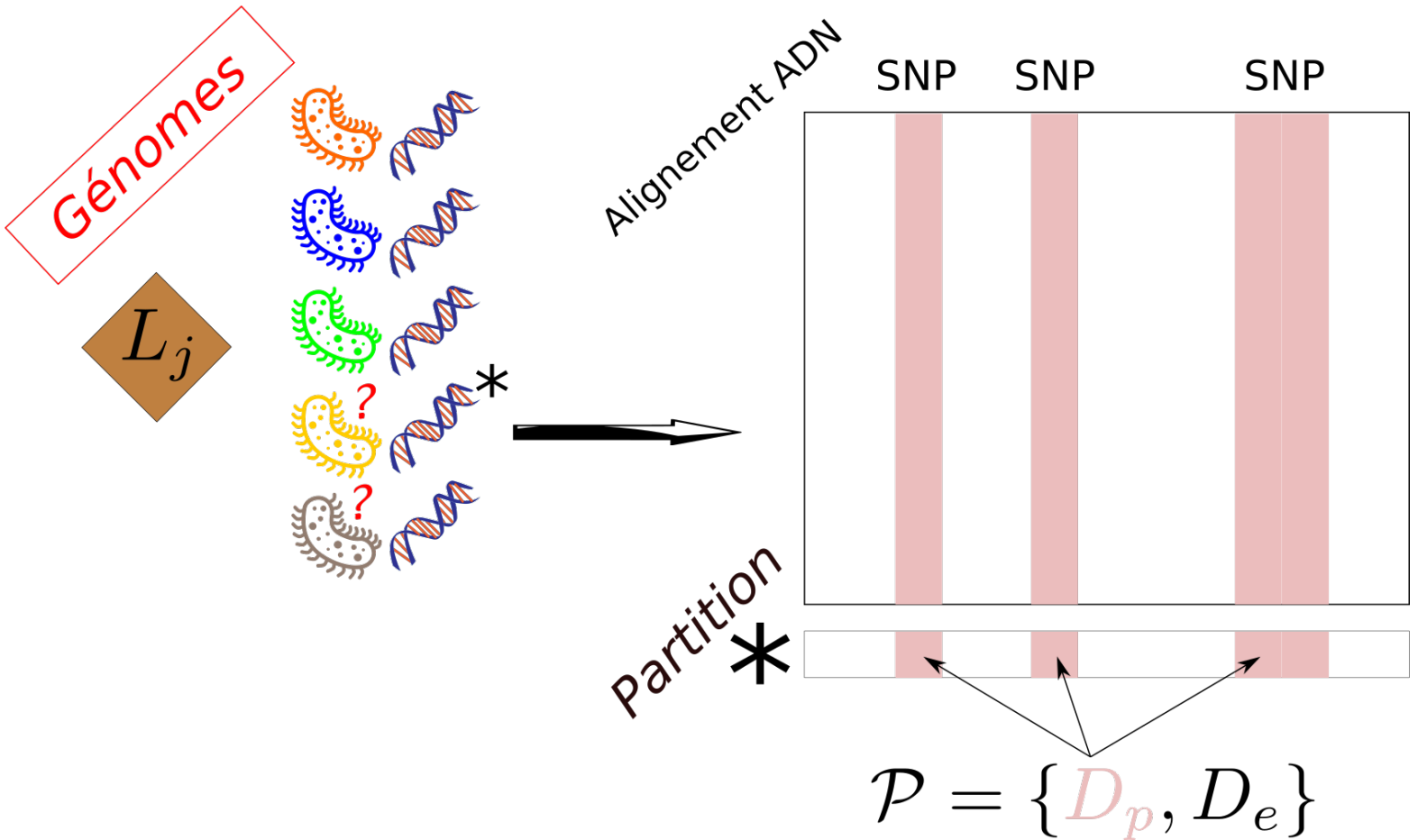
- le temps d'évolution t est inféré à partir de la phylogénie des lignées connues

Proposition phylogénie



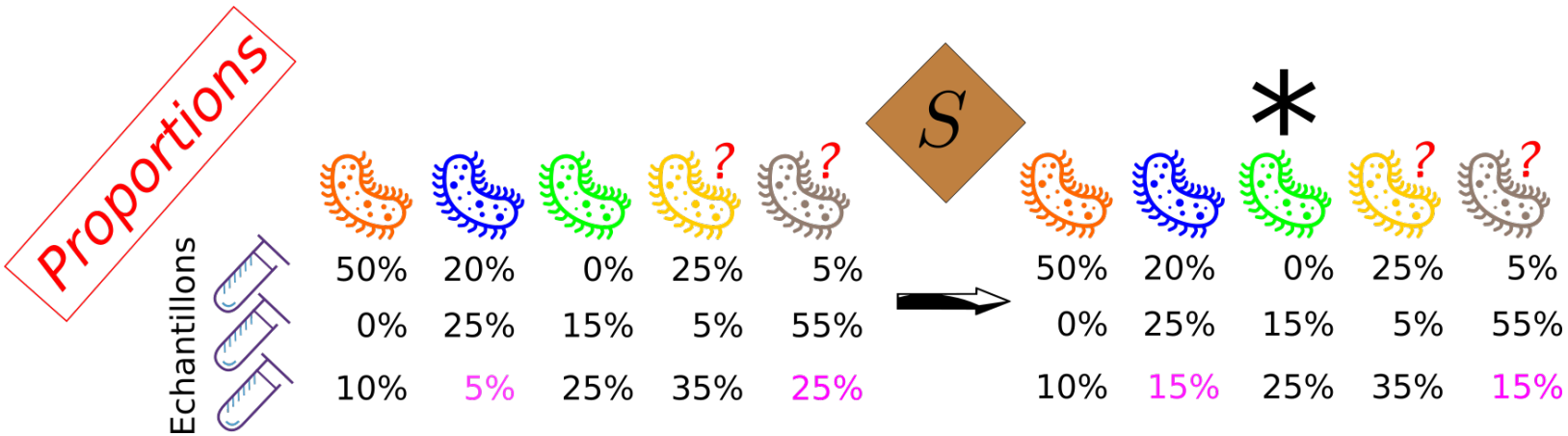
- l'arbre proposé est estimé par **maximum de vraisemblance** selon le modèle de Juke et Cantor 1969

Proposition partition



- le génotype proposé implique une proposition pour la partition des sites en sites variants et non variants

Proposition proportions



- on propose une nouvelle matrice des proportions en modifiant 2 probabilités pour un échantillon tiré aléatoirement
- une loi Dirichlet peut également être utilisée pour modifier toute la distribution

Estimation : Metropolis-Hastings

- les probabilités d'acceptation-rejet doivent être calculées à chaque itération, exemple pour génotype/arbre/partition :

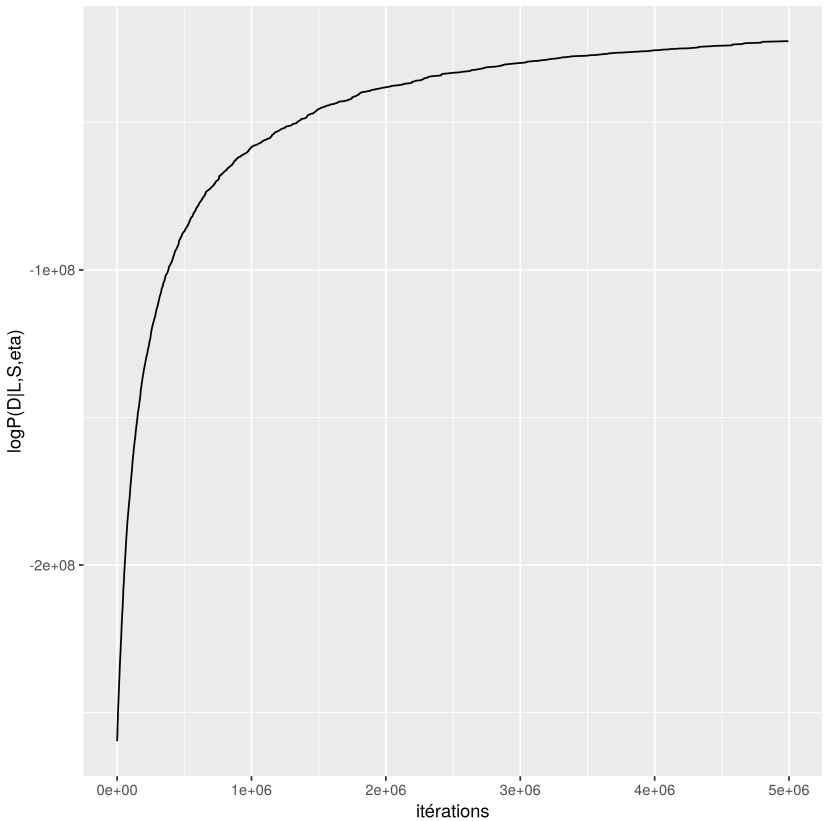
$$\alpha = \alpha_1 \times \alpha_2$$

$$\alpha_1 = \frac{P(D|L^*,S,\eta)}{P(D|L,S,\eta)} \times \frac{P(L^*|\mathcal{P}^*,\mathcal{T}^*,\xi)}{P(L|\mathcal{P},\mathcal{T},\xi)} \times \frac{P(\mathcal{P}^*|\lambda)}{P(\mathcal{P}|\lambda)} \times \frac{P(\mathcal{T}^*)}{P(\mathcal{T})}$$

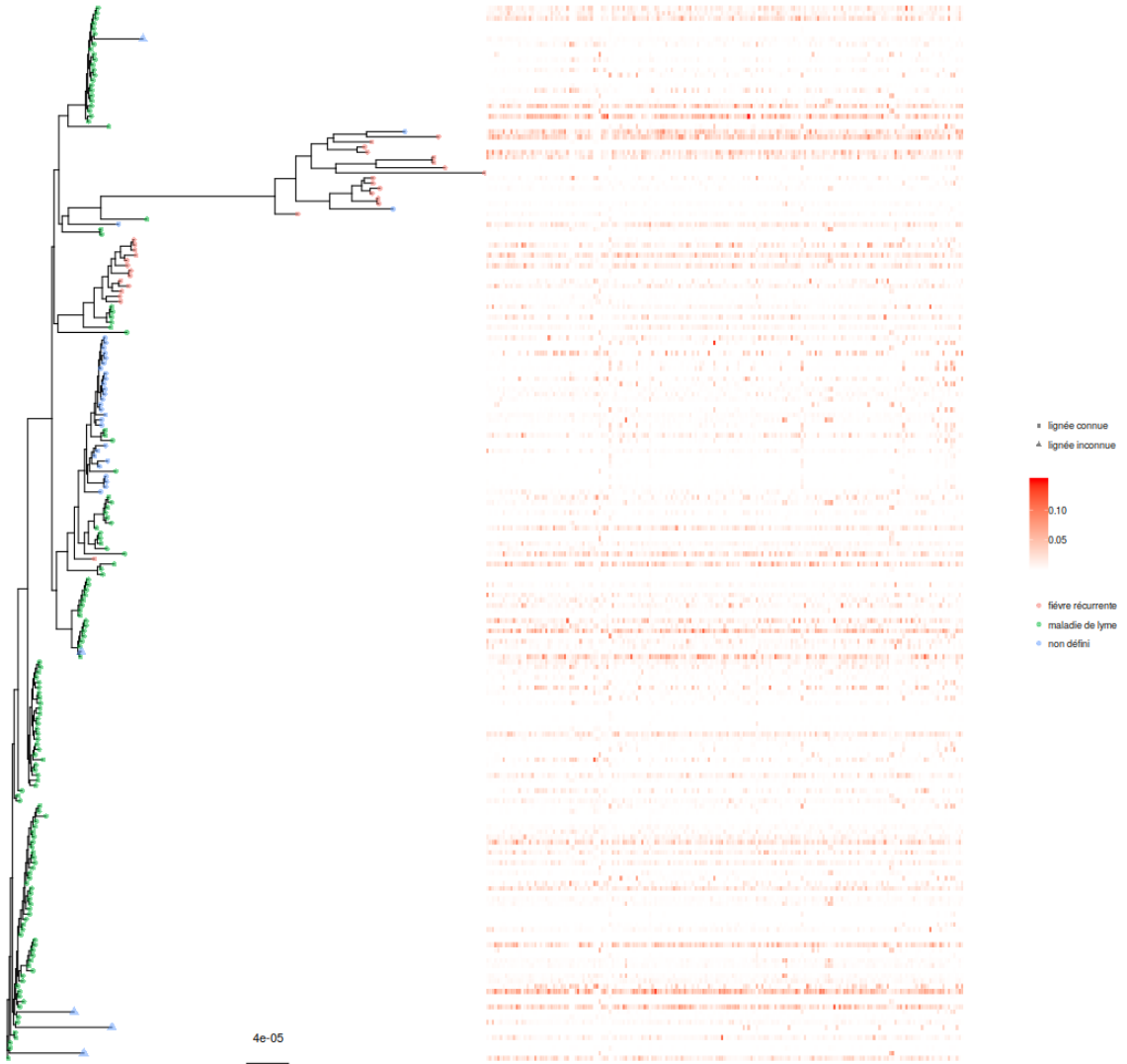
$$\alpha_2 = \frac{q(L|L^*)}{q(L^*|L)} \times \frac{q(\mathcal{T}|\mathcal{T}^*)}{q(\mathcal{T}^*|\mathcal{T})} \times \frac{q(\mathcal{P}|\mathcal{P}^*)}{q(\mathcal{P}^*|\mathcal{P})}$$

Premiers résultats

- données de séquençage (Margos, G) : référence le gène ribosomale rplB de *Borrelia garinii* NP8, 179 échantillons, 200 lignées connues (pubmlst ; Millins, C.) et 5 inconnues.
- 5E6 itérations, 15 jours de calcul ! ! !



Premiers résultats



Perspectives et logiciel

- **Algorithme**
  - $5E6$ itérations on déjà été conduites, objectif $50E6$! !
  - la convergence n'est pas encore satisfaisante (chaîne d'arbre et grande matrice).
  - le pilotage de l'algorithme doit être amélioré.
  - les lois *a priori* doivent être testées
  - parallélisation par propositions multiples.
- **Validation du modèle**
  - tester sur données simulées.
  - comparaison avec le modèle de O'Brien 2014 ou d'autres.
  - étude du choix du nombre de lignées inconnues (inférence ?)
- **Le logiciel *lignee3***
  - C++, GSL 2.4 ; Bio++ 2.4
  - CodeLite, GIT, valgrind 3.13, Doxygen 1.8