Phylogénies non binaires en épidémiologie

Patrick Hoscheit, Tim Vaughan, Oliver Pybus

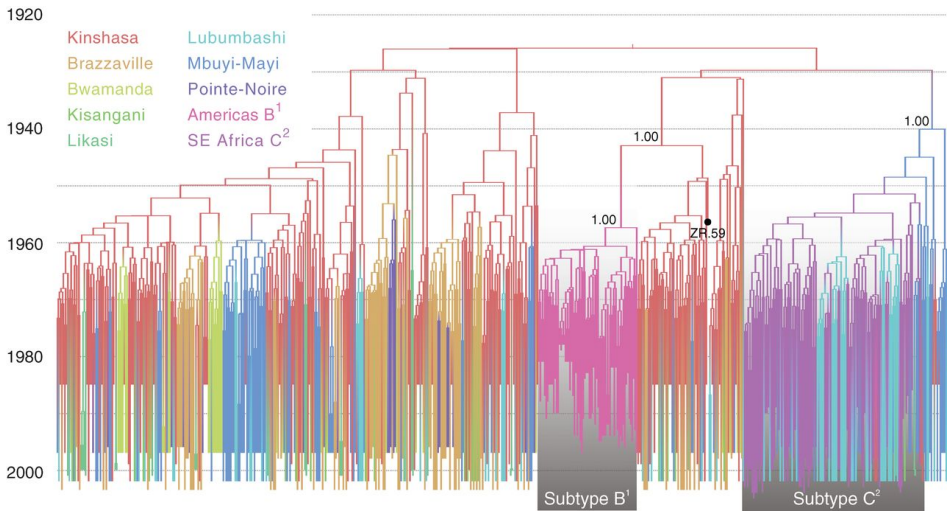
AG du département MIA
22 mai 2019

Méthodes phylodynamiques

Paradigme

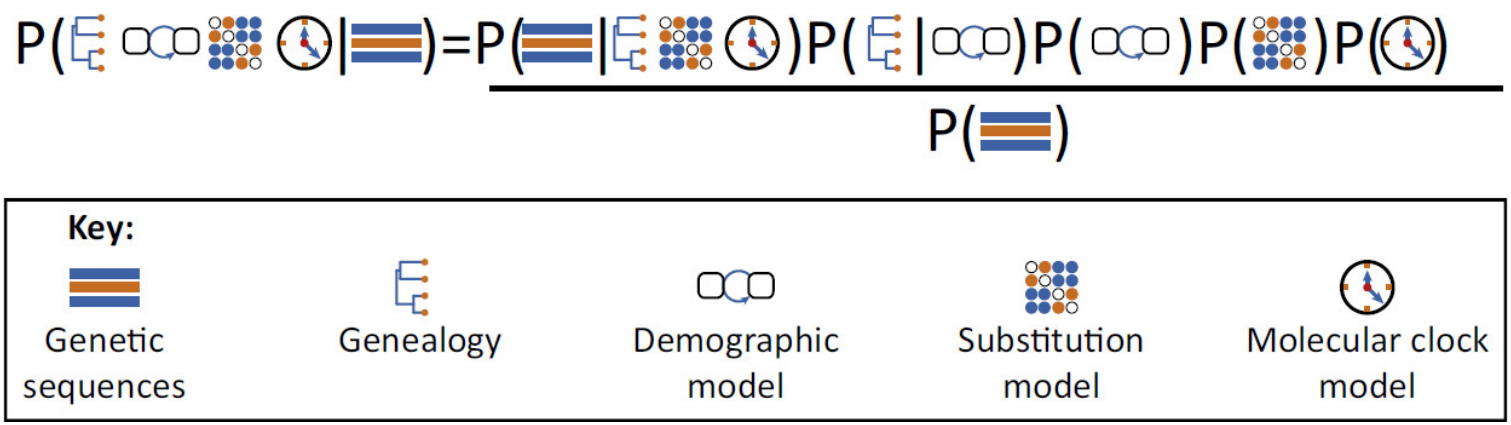
L'épidémiologie d'un virus est intimement reliée à son évolution. On peut se servir de l'une pour comprendre l'autre.

- Explosion des données moléculaires liée au faible coût du séquençage
- Disponibilité d'importantes ressources de calcul



[Faria et al. 2014]

Analyse bayésienne



[Du Plessis, Stadler 2015]

- Vraisemblance phylogénétique calculée par l'algorithme de Felsenstein
- Échantillonnage de la loi *a posteriori* par MCMC (BEAST, BEAST2)
- Modèles d'arbres paramétriques (type SIR) ou non paramétriques (modèles *sky**)
- Possibilité d'inclure d'autres données (spatiales, immunologiques,…)

Multifurcations

Traditionnellement, on distingue :

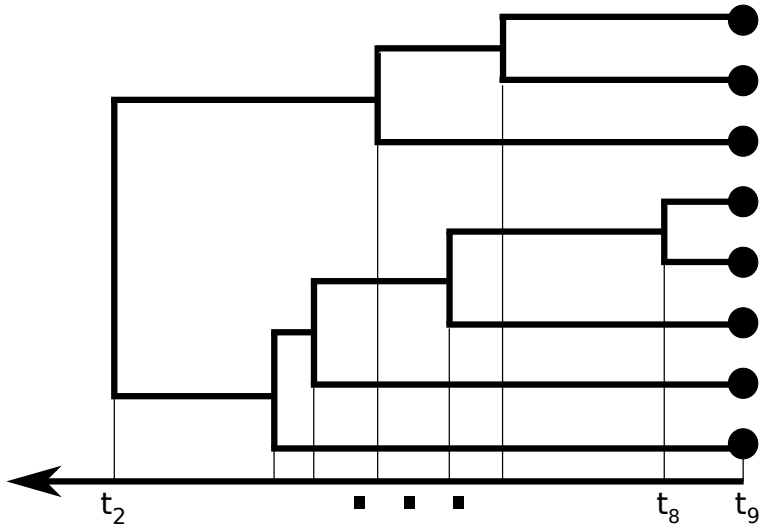
- Multifurcations *douces* : reflètent un défaut de résolution phylogénétique
- Multifurcations *fortes* : réprésentent l'occurence (quasi-)simultanée de plusieurs événements de diversification (spéciation ou infection)

Multifurcations dures peuvent apparaître dans plusieurs contextes :

- Lois de reproduction fortement hétérogènes (huîtres, morues, superpropagateurs)
- Sélection forte et répétée
- Radiation adaptative (controversé)
- Suréchantillonnage

Quels modèles phylodynamiques pour tester et prendre en compte l'existence potentielle de multifurcations dures ?

Coalescent de Kingman



- Processus en temps continu $(\mathcal{T}^n_{\text{King}}(t),\ t \geq 0)$ débute avec n lignées
- Chaque paire de lignées coagule à taux 1, indépendamment des autres paires.
- Population variable déforme ce processus de coalescence
- Vraisemblance d'un arbre sachant la taille de population efficace :

$$\mathcal{L}(\mathcal{T}|N_e) = \prod_i \frac{\binom{i}{2}}{N_e(t_i)} \exp\left(-\int_{t_i}^{t_{i+1}} \frac{\binom{i}{2}}{N_e(t)} dt\right)$$

Λ-coalescents

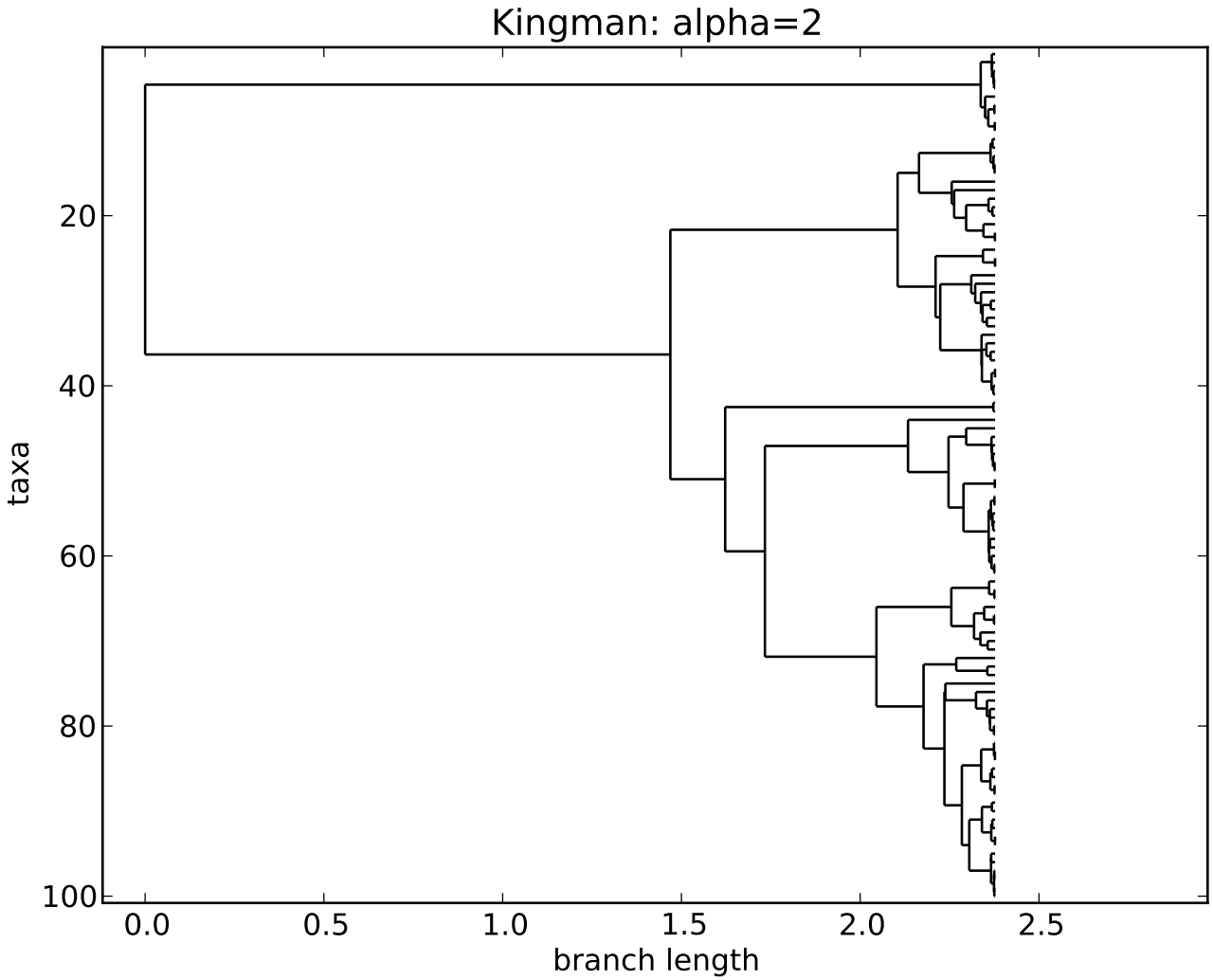
Définition (Pitman 1999)

Soit Λ une mesure de probabilités sur $[0,1]$, et soit $n \geq 0$. Le n-Λ-coalescent est l'arbre aléatoire débutant avec n lignées, et tel que, s'il y a $p \leq n$ lignées à un instant donné, alors tout k-uplet (avec $2 \leq k \leq p$) coagule indépendamment des autres au taux

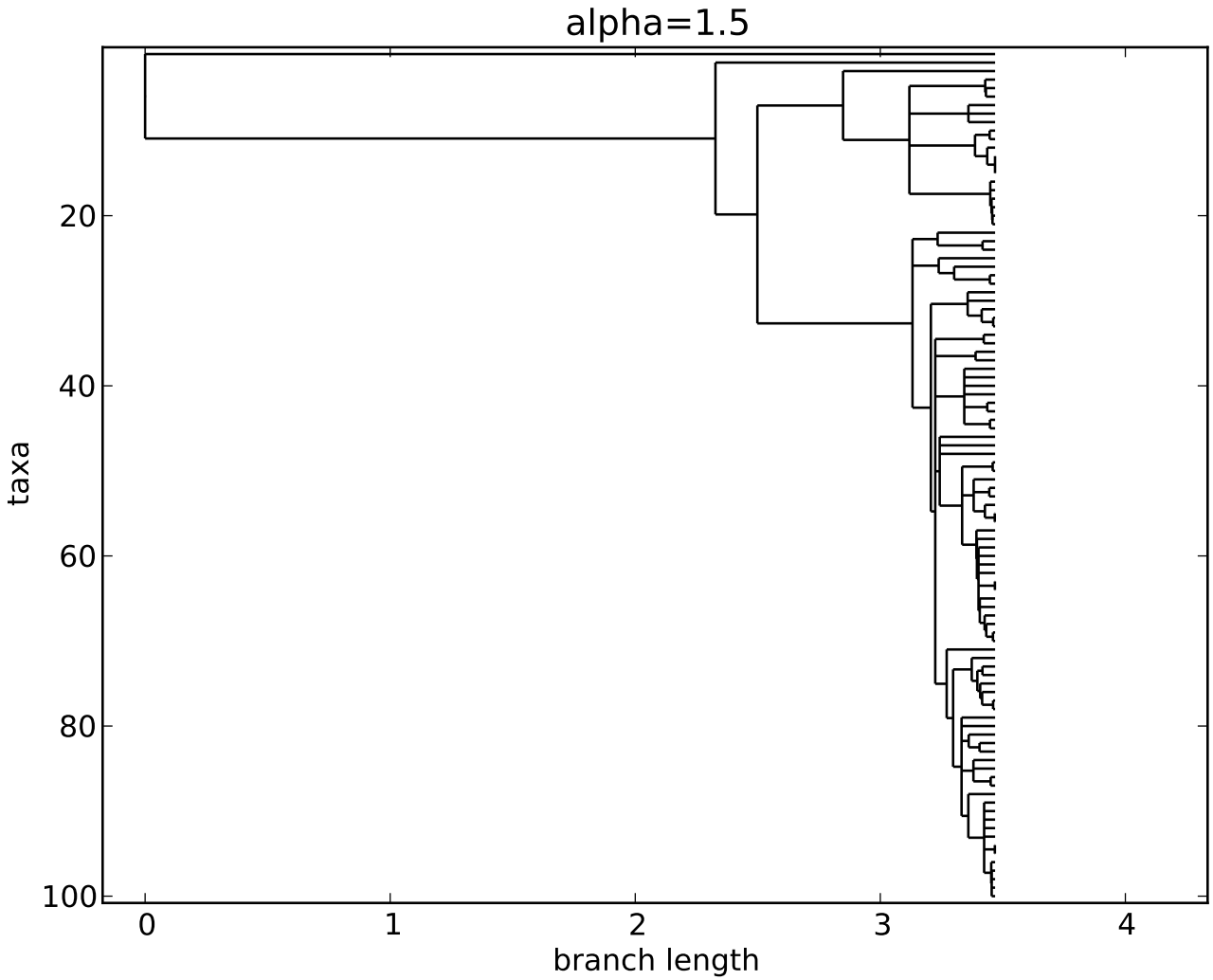
$$\lambda_{p,k} = \int_0^1 x^{k-2}(1-x)^{p-k}\Lambda(dx).$$

- Permet des coalescences non-binaires
- Cas particuliers : $\Lambda = \delta_0$ (coalescent de Kingman), $\Lambda = \delta_1$ (coalescent en étoile) et $\Lambda = \mathbf{1}_{[0,1]}(x)dx$ (coalescent de *Bolthausen-Sznitman*)
- Ici, famille à un paramètre : $\Lambda(dx) = x^{1-\alpha}(1-x)^{\alpha-1}dx$, $\alpha \in [1,2]$

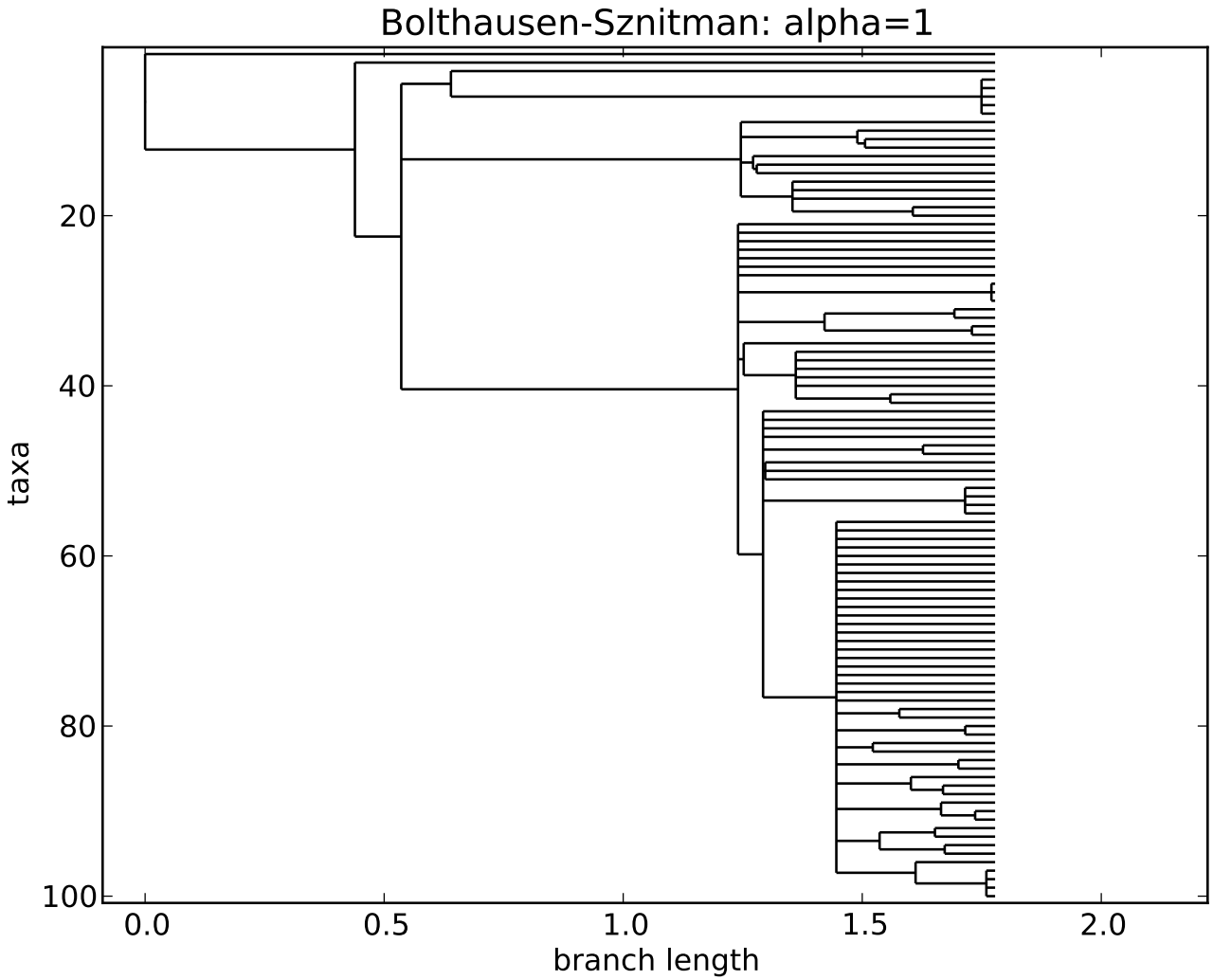
Coalescent de Kingman



Beta(0.5,1.5)-coalescent



Beta(1,1)-coalescent



Processus Λ-skyline

Vraisemblance d'un arbre sachant la taille de population efficace :

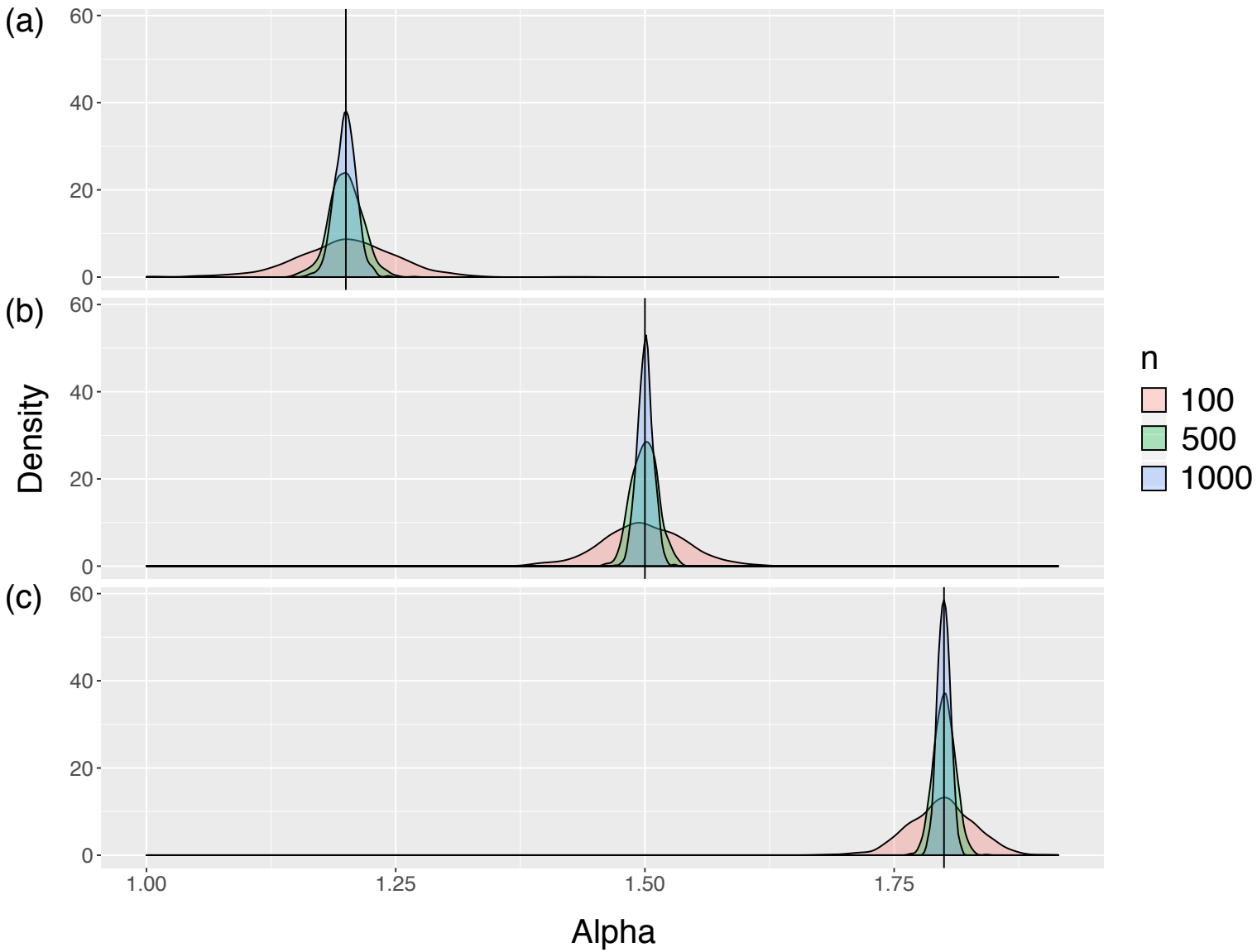
$$\mathcal{L}(\mathcal{T}|N_e,\alpha) = \prod_i \frac{\binom{n_i}{k_i}\lambda_{n_i,k_i}(\alpha)}{N_e(t_i)} \exp\left(-\int_{t_i}^{t_{i+1}} \frac{\sum_{k=2}^{n_i}\binom{n_i}{k}\lambda_{n_i,k}(\alpha)}{N_e(t)}dt\right)$$

- Estimateur MV de la taille de population, supposée constante sur les intervalles inter-coalescence

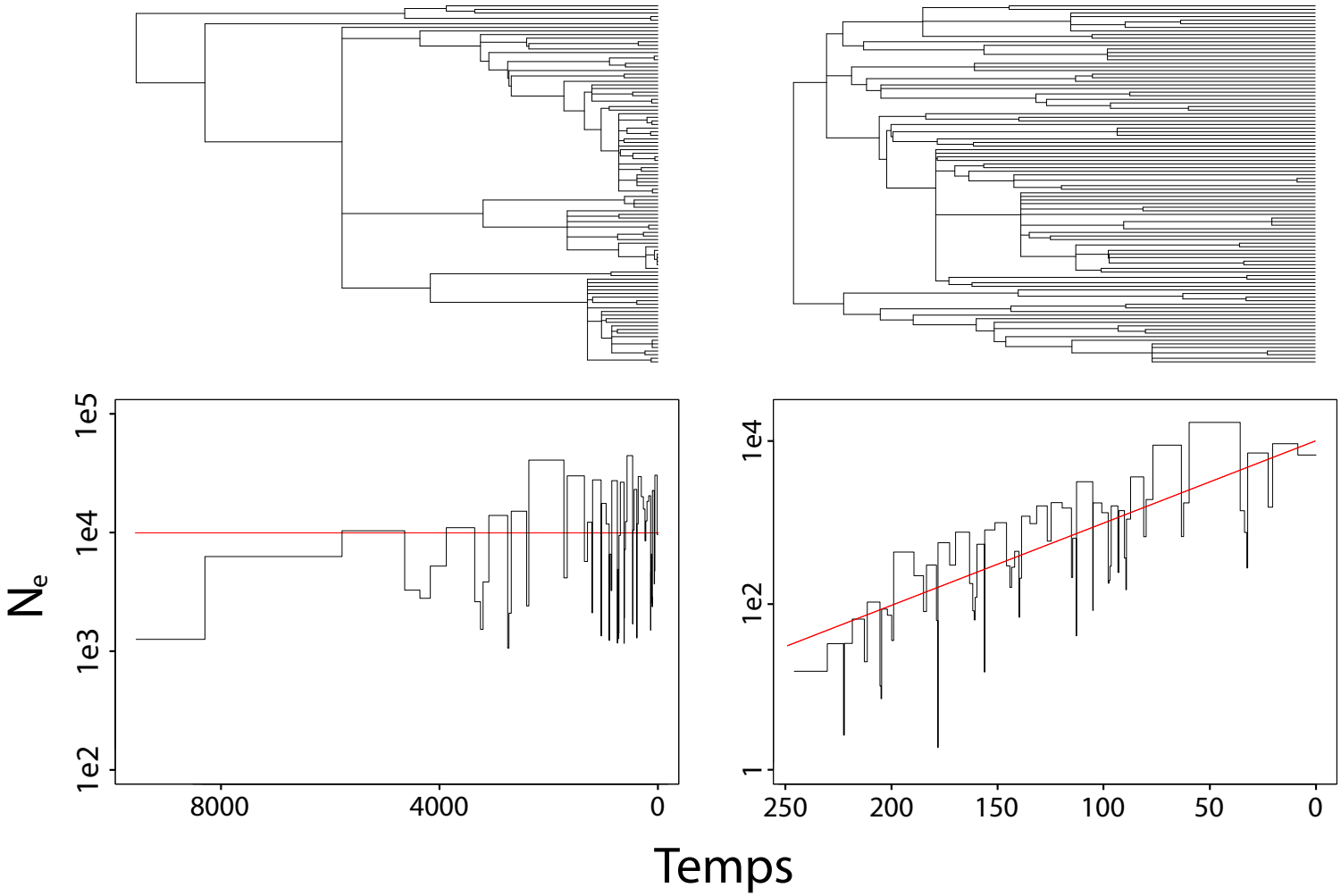
$$\hat{N}_e(t) = \sum_{k=2}^{n_i}\binom{n_i}{k}\lambda_{n_i,k}(\alpha)(t_i - t_{i+1}),\ t \in [t_{i+1}, t_i)$$

- Permet d'estimer $\alpha \in [1,2]$ pour un arbre donné
- Validation en simulant 1000 arbres avec $\alpha = 1.2, 1.5$ et 1.8 (N_e constant)

Données simulées



Données simulées



Superpropagation

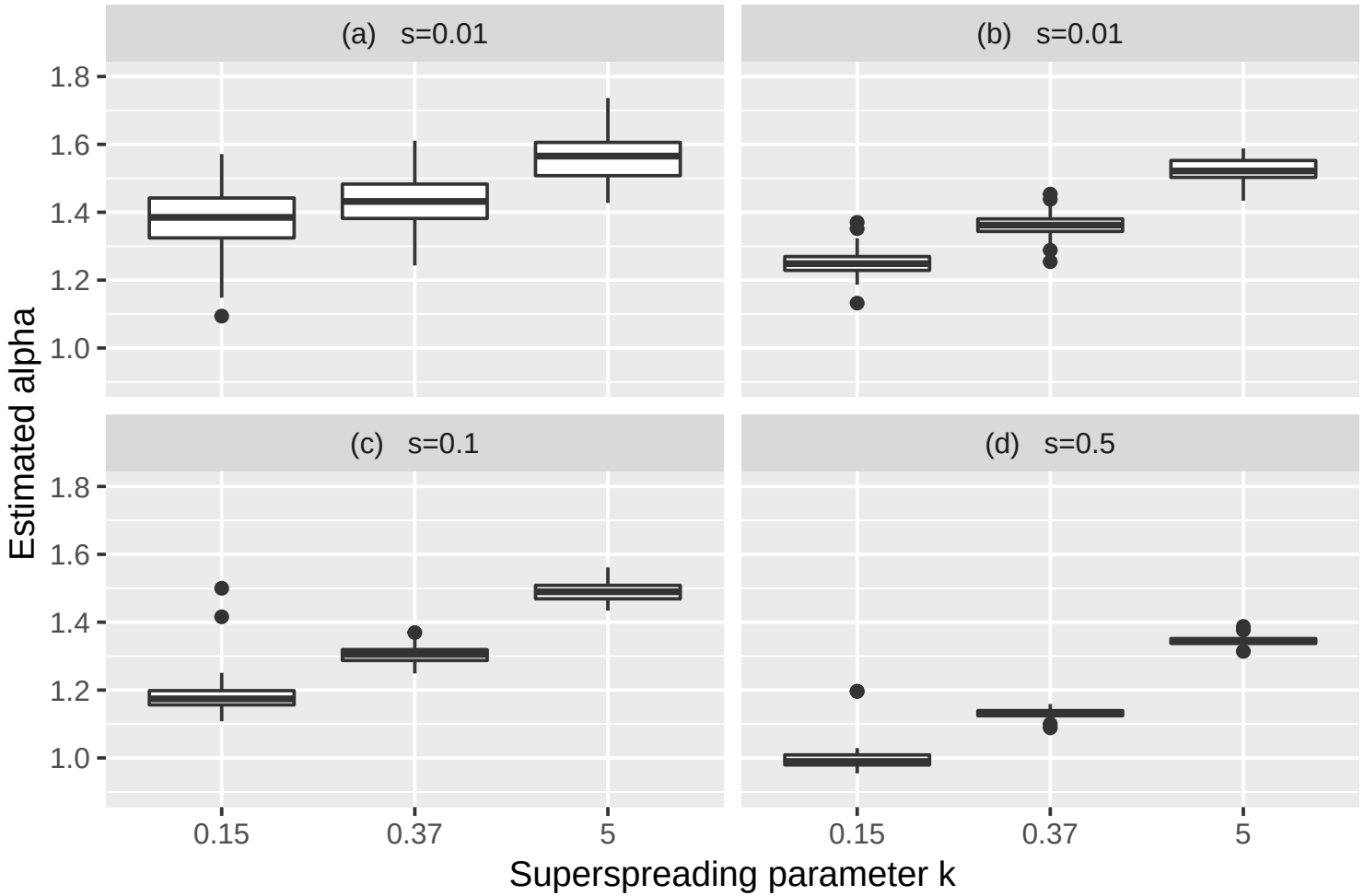
Modèle SIR de Lloyd-Smith :

- Chaque individu infectieux est contagieux pendant un temps exponentiel
- ν_i potentiel de contagion de l'individu i, de loi Gamma$(R_0/k, k)$
- Sachant ν_i, l'individu infectieux i transmet l'infection à Z_i susceptibles, avec Z_i de loi Po(ν_i)

Pour R_0 donné, paramètre k mesure le potentiel d'apparition d'individus *superpropagateurs*

- Fort potentiel (SARS, rougeole,...) : $k < 1$
- Faible potentiel (infections contrôlées) : $k > 1$

Arbres de Lloyd-Smith



Implémentation BEAST2

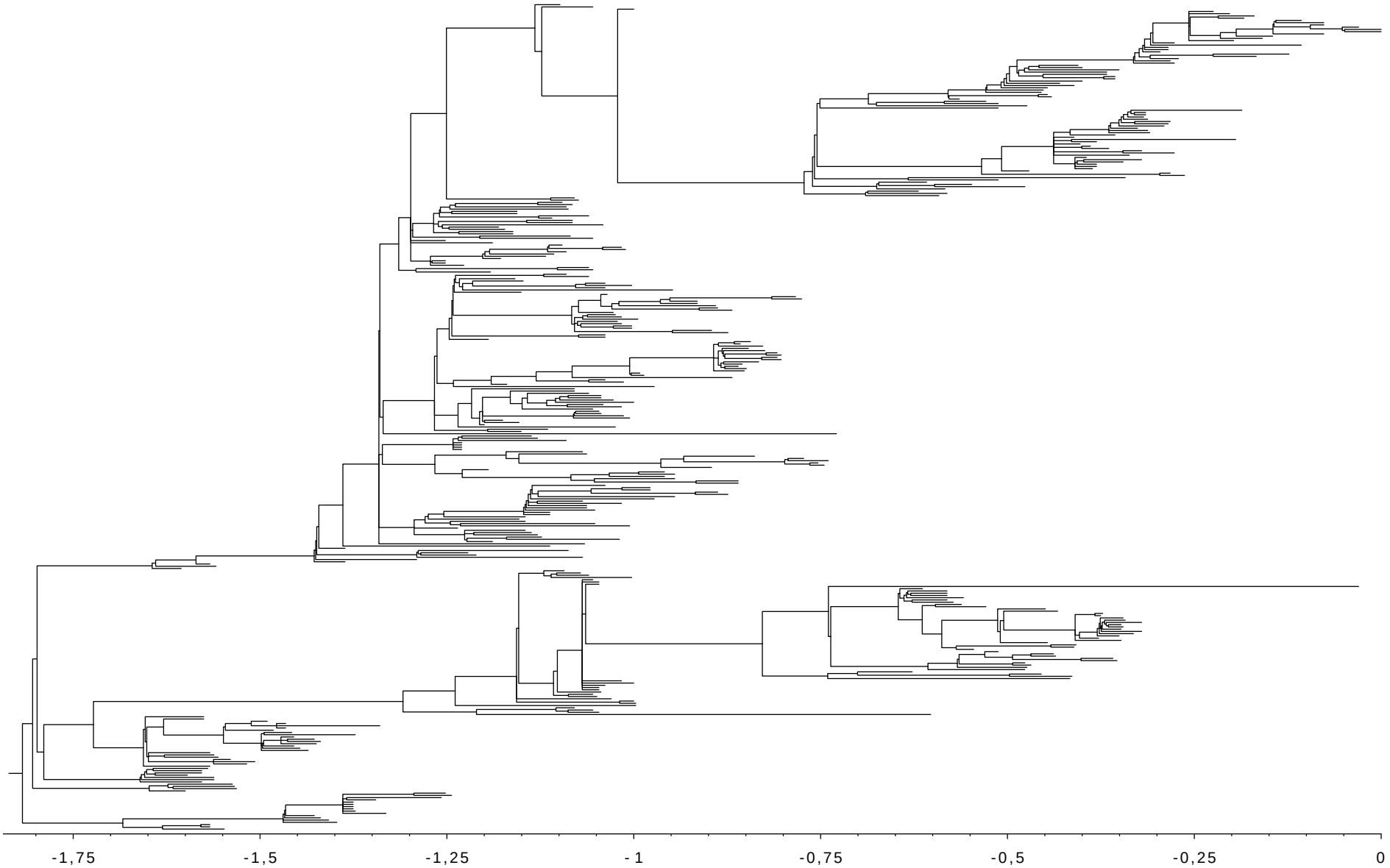
Réalisée par Tim Vaughan (ETH Zürich)

- Nécessité de définir des opérateurs MCMC adaptés aux arbres avec multifurcations
- Validation par données simulées
- Prototype disponible sur `github.com/tgvaughan/pitchfork`

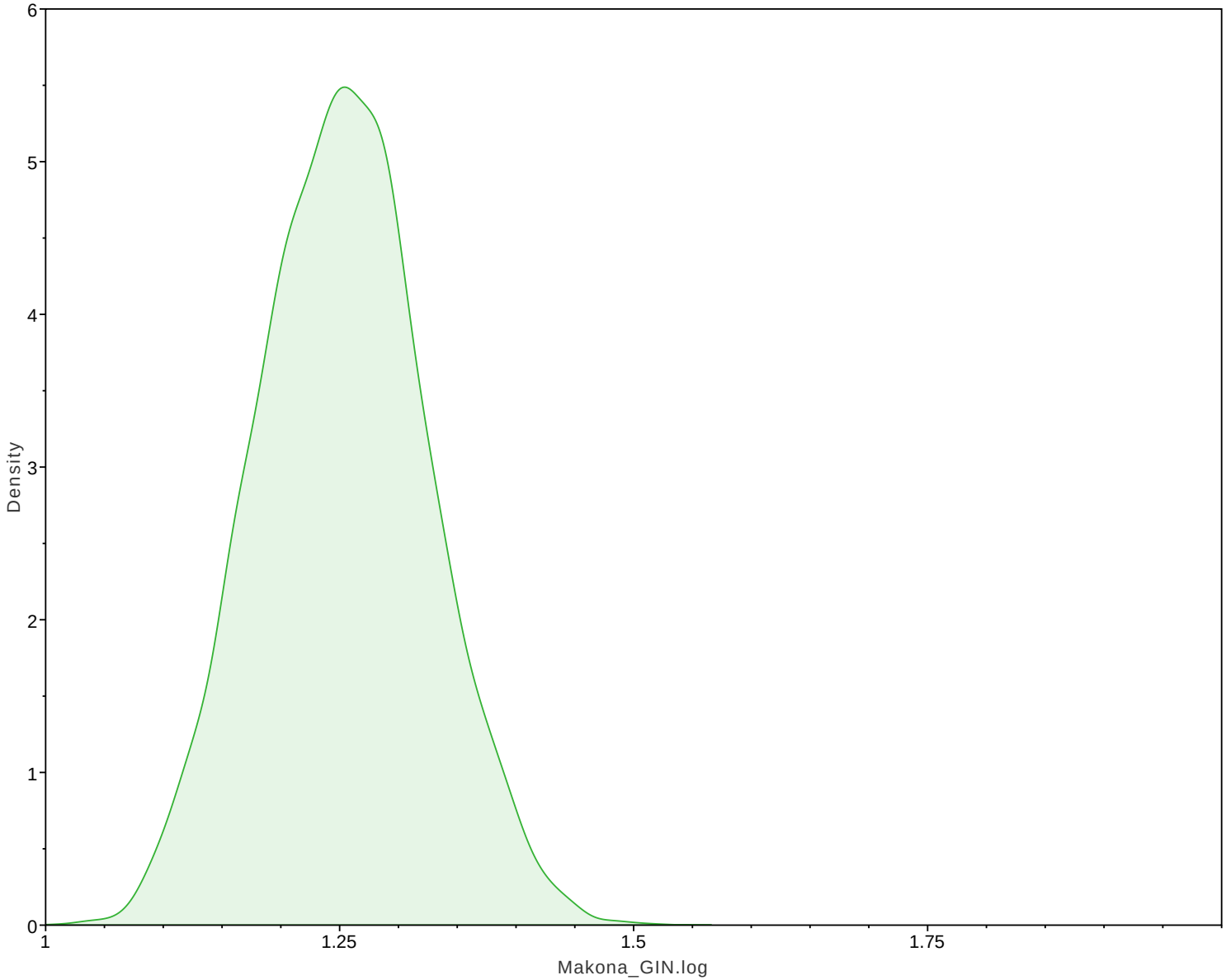
Application à un échantillon de séquences Ebola

- Épidémie en Guinée en 2014-2015
- 368 séquences complètes (~ 19.000 pb)

Phylogénie Ebola



Distribution *a posteriori* de α



Questions ouvertes

- Interprétation de N_e, en particulier en contexte épidémiologique
- Mesures mixtes $\Lambda(dx) = a\delta_0 + (1-a)\text{Beta}(2-\alpha, \alpha)(dx)$
- Lien avec les processus de branchement
- Nouvelle mesure Λ qui incorpore les paramètres du modèle de Lloyd-Smith
- Autres contextes d'apparition de multifurcations : expansion spatiale, sélection forte, …

Merci de votre attention !