Modèles hiérarchiques Bayésiens pour la génomique des populations

Mathieu Gautier

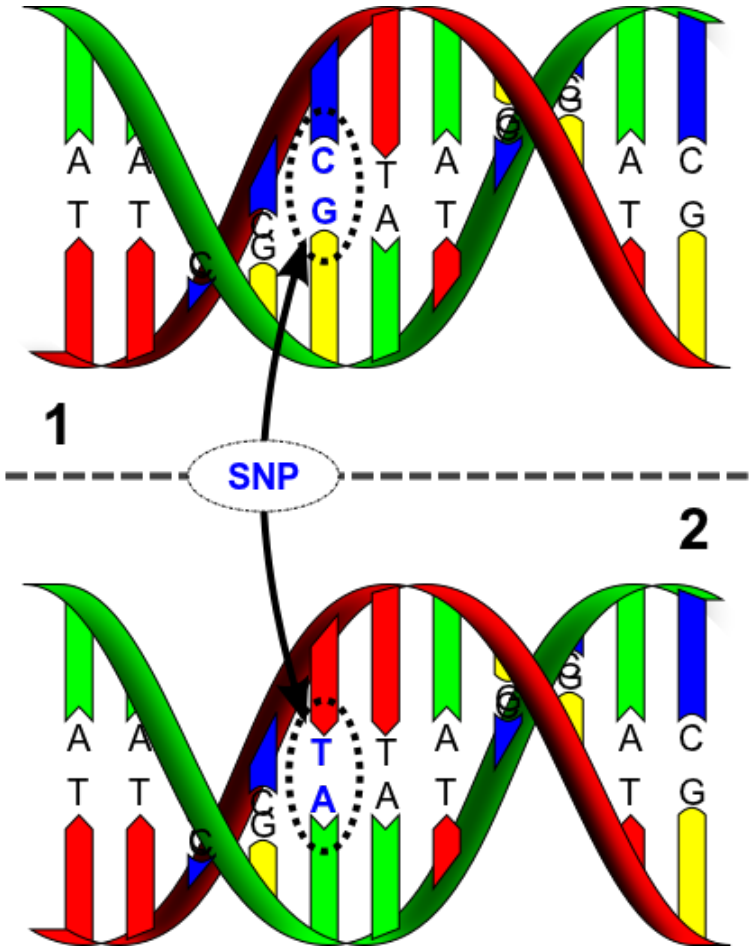
UMR INRA/CIRAD/IRD/SupAgro CBGP

28 Novembre 2014

La révolution génomique du haut-débit

- 2001 : Publication des deux premiers assemblages du génome humain
- Énorme contribution dans l'amélioration des technologies et des stratégies de séquençage et de génotypage
- Démocratisation des technologies
  - De nombreux génomes séquencés pour des coûts en baisse
    - 2001 : génome humain : 3.10^9\$ (7 ans)
    - 2007 : génome équin : 30.10^6\$ (18 mois)
    - 2013 : 1 génome humain reséquencé pour 1,000 \$
  - Caractérisation facilitée de la variabilité génétique
    - 1,000 genome project chez l'homme (McVean *et al.*, 2011), le riz (McNally *et al.*, 2014), la vache (Hayes *et al.*, 2014)...
    - Possibilités de disposer d'un grand nombre de marqueurs dans les espèces non modèles
- La disponibilité en marqueurs n'est plus limitante

Exemple de marqueurs populaires : Les SNPs



- Polymorphisme nucléotidique ponctuel

  (Single Nucleotide Polymorphism)
- Le plus souvent bi-allélique

  ($\mu \simeq 10^{-8}$)
- Répartition homogène dans le génome et très fréquents

  (*e.g.* environ 1 à 2 par kb chez le bovin ou l'homme soit plusieurs millions en tout)
- Génotypage entièrement automatisable

  (marqueur phare de l'ère du haut débit mais potentiellement soumis à des biais de recrutement)

La variabilité génétique au sein des populations

Les forces évolutives agissant sur l'évolution des fréquences alléliques

- Mutation (et recombinaison à l'échelle haplotypique) ⇒ source de la variabilité
- Dérive génétique (taille finie des populations) ⇒ érosion de la variabilité
- Migration ⇒ favorise le maintien de la variabilité (flux/échange d'allèles)
- Sélection ⇒ favorise le maintien/fixation de variants favorables

Influences sur les patrons génomique de variabilité *(Cavalli-Sforza, 1966)*

- Facteurs démographiques (dérive, flux de gène) ⇒ effet global ("Diversité génétique neutre")
- Sélection (mutation et recombinaison) ⇒ effet local ("Diversité génétique adaptative")

⇒ *Les patrons de variabilité génomiques informent sur les processus historiques et biologiques (approches indirectes)*

Inférer l'histoire démographique des populations

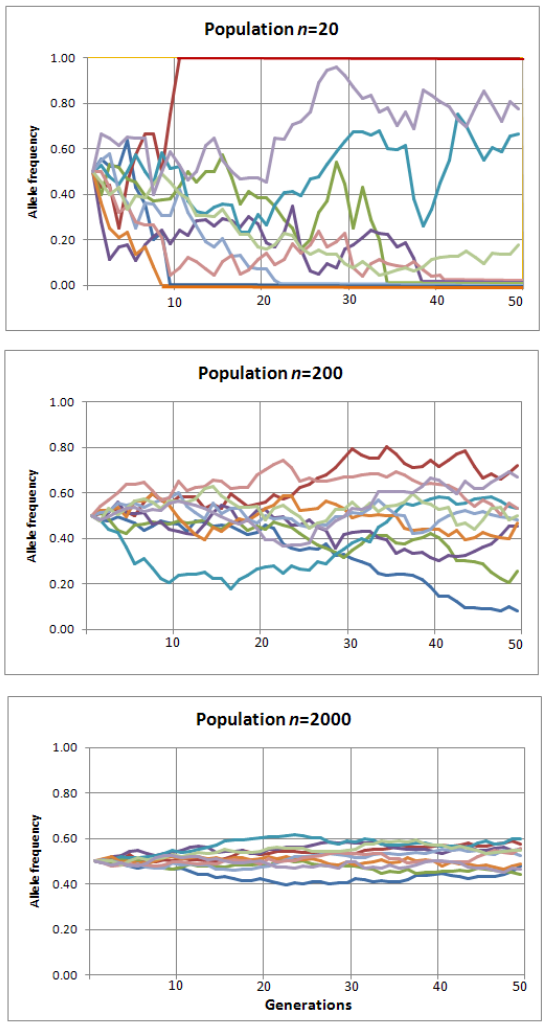
Objectifs

- Inférer l'histoire démographique des populations à partir des données et d'un modèle démo-génétique décrivant le processus biologique supposé
- Comparer différents scénarios possibles et estimer leurs paramètres

Deux types d'approches

- Likelihood-free (le modèle demo-genetique est simulé)
  - ABC : *e.g.* DIY–ABC (Cornuet, Estoup & co)
- Likelihood approaches (le modèle demo-genetique est caractérisé analytiquement)
  - Modélisation "Backward" (coalescent) : *e.g.* MIGRAINE (Leblois, Rousset), LAMARCK (Kuhner & co), IM (Hey, Nielsen & co), MSVAR (Beaumont)
  - Modélisation "Forward" : *e.g.* PHYLIP (Felsenstein & co), ∂a∂i (Gutenkunst *et al.*, 2010), TREEMIX (Pritchard & Pickrell, 2012), KIM_TREE (Gautier & Vitalis, 2013)

Le modèle (démographique) en pure dérive (Wright/Fisher)



Hypothèses

- Les populations (j) évoluent pendant t générations (non chevauchantes) en complet isolement depuis leur population ancestrale

Évolution des fréquences alléliques

- $\mathbb{E}[\alpha_j] = \pi$ et $\mathbb{V}[\alpha_j] = \pi(1-\pi)[1 - \left(1 - \frac{1}{2N_j}\right)^t]$
- $\lim_{\frac{t}{2N} \to 0} \mathbb{V}[\alpha_j] = \pi(1-\pi)\tau$ avec $\tau := \frac{t}{2N}$

Modèle en déséquilibre

- $\mathbb{V}[\alpha_j]$ est fonction croissante de t ($V_{max} = \pi(1-\pi)$)
- $\mathbb{P}_{fix} = \pi$ et $\mathbb{P}_{lost} = 1 - \pi$

Approximations de la Distribution

A) le modèle $\mathcal{F}$

$$\alpha|\pi, c \sim \beta\left(a_\beta = \pi_i \frac{1-c}{c}, b_\beta = (1-\pi)\frac{1-c}{c}\right)$$

- **Balding et al., 1996** ; Falush et al., 2003, Gautier et al., 2010 ; Sìren et al., 2011
- $\mathbb{E}[\alpha|\pi, c] = \frac{a_\beta}{a_\beta + b_\beta} = \pi$ and $\mathbb{V}[\alpha|\pi, c] = \frac{a_\beta + b_\beta}{(a_\beta + b_\beta)(a_\beta + b_\beta + 1)} = c\pi(1-\pi)$

B) le modèle $\mathcal{N}_\mathcal{T}$

$$\alpha|\pi, c \sim \mathbb{N}_{[0,1]}\left(\pi, c\pi(1-\pi)\right)$$

- **Nicholson et al. (2002)** ;Gautier et al. (2010), Coop et al. (2010), Pickrell and Pritchard (2012)
- $\mathbb{E}[\alpha|\pi, c] = \pi$ and $\mathbb{V}[\alpha|\pi, c] \simeq c\pi(1-\pi)$ (troncature)

C) le modèle $\mathcal{K}$: Approximation de diffusion (Kimura, 1955,1964)

Distribution des fréquences allélique sous l'approximation de diffusion

Hypothèses et notations

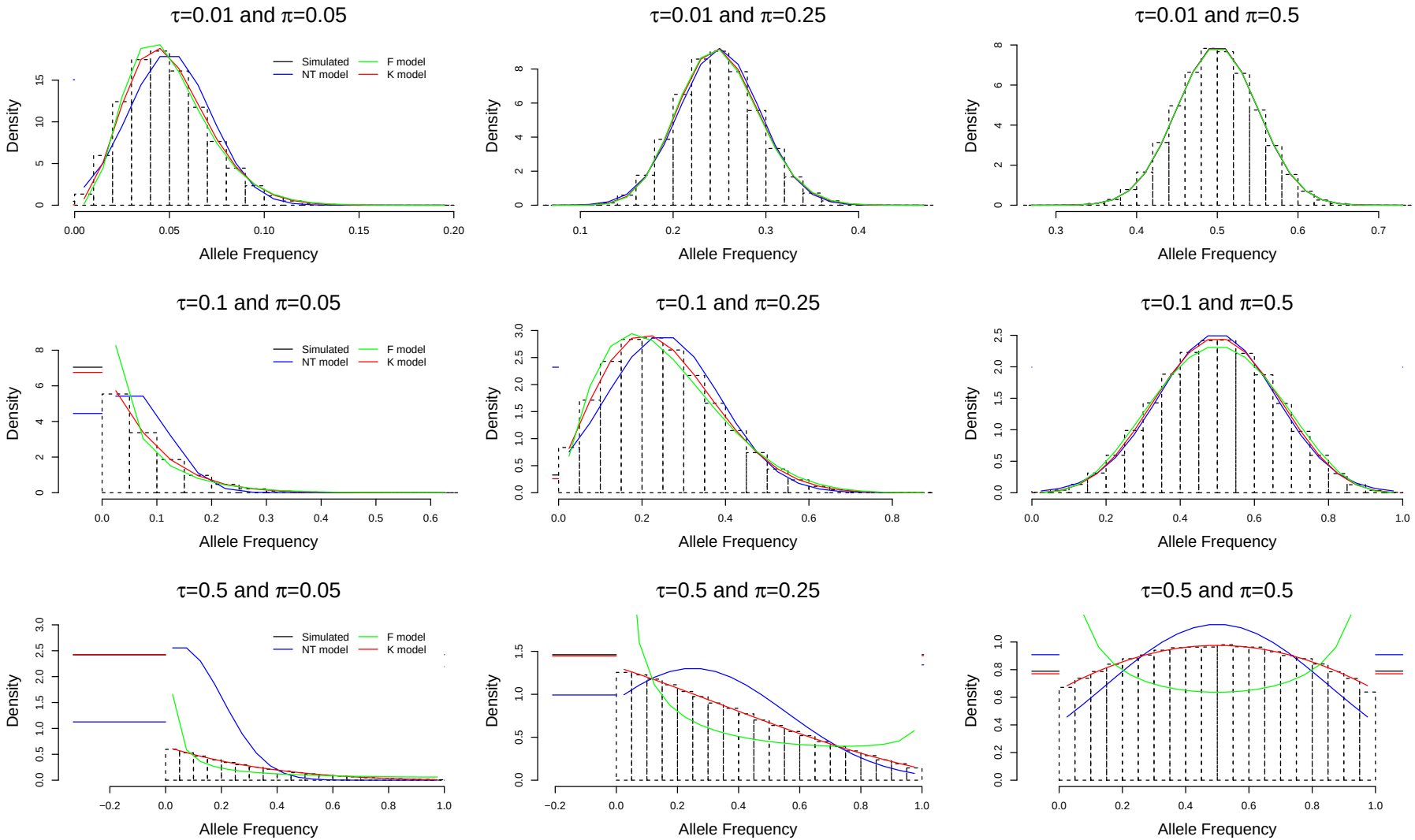
- Si $\frac{1}{2N_e} \to 0$, α varie selon un processus de Markov à temps continu
- Equation Forward de Kolmogorov : $\phi(\alpha \mid \pi, t)$, $\frac{\partial \phi}{\partial t} = \frac{1}{4N_e} \frac{\partial^2}{\partial \alpha^2} (\alpha(1-\alpha)\phi)$

Solutions (Kimura, 1955, 1964)

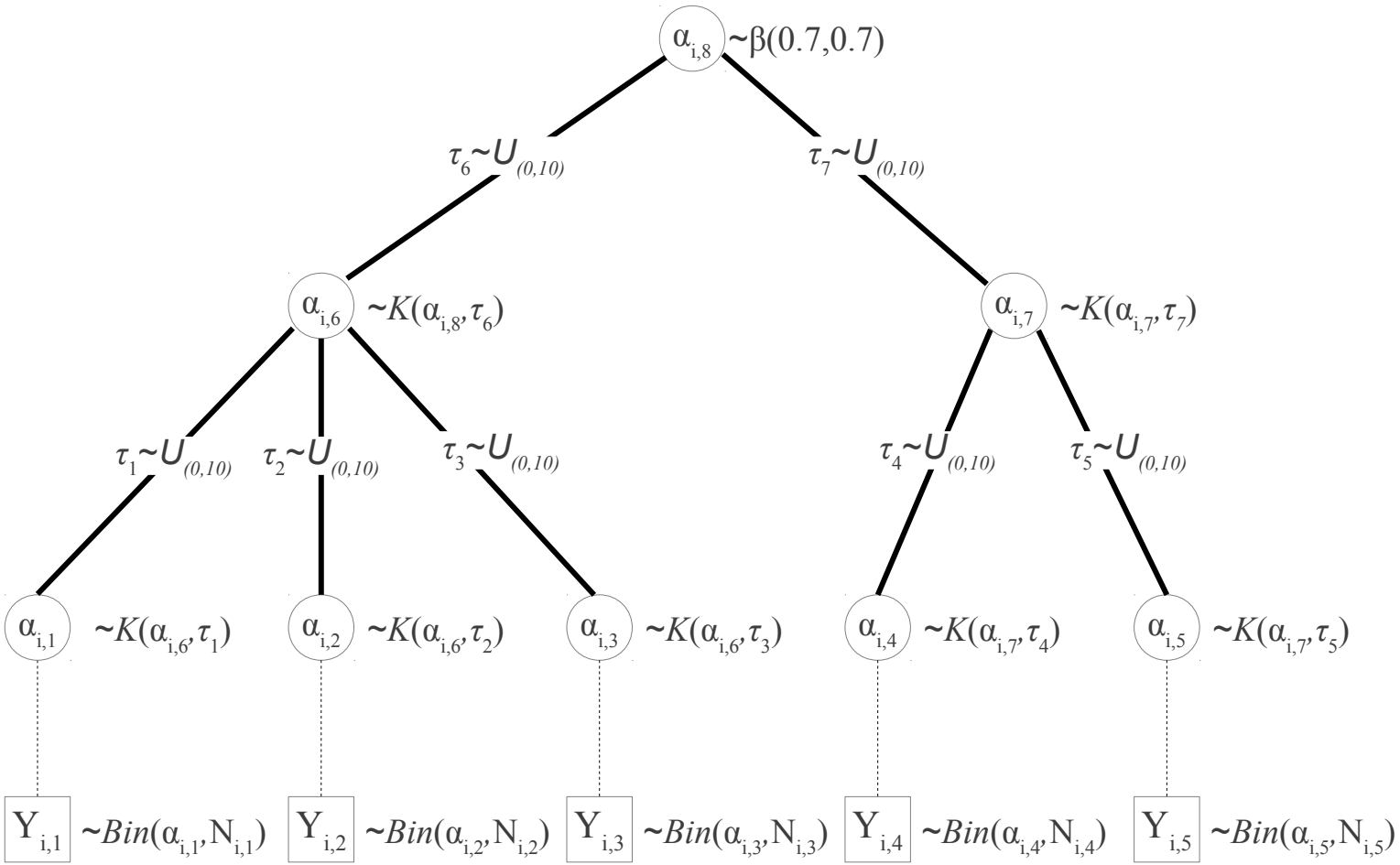
$$\begin{cases} \pi(\alpha_{ij} \mid \pi_i, \tau_j) & = (1-w_{ij}^2) \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} T_{l-1}^1(w_{ij}) T_{l-1}^1(z_i) e^{-\frac{1}{2} l(l+1)\tau_j} \quad \text{si } 0 < \alpha_{ij} < 1 \\ \mathrm{P}(\alpha_{ij} = 0 \mid \pi_i, c_j) & = (1-\pi_i) + \frac{(1-z_i)^2}{2} \sum_{l=1}^{\infty} (-1)^i \frac{2l+1}{l(l+1)} T_{l-1}^1(-z_i) e^{-\frac{1}{2} l(l+1)\tau_j} \\ \mathrm{P}(\alpha_{ij} = 1 \mid \pi_i, c_j) & = \pi_i + \frac{(1-z_i)^2}{2} \sum_{l=1}^{\infty} (-1)^i \frac{2l+1}{l(l+1)} T_{l-1}^1(z_i) e^{-\frac{1}{2} l(l+1)\tau_j} \end{cases}$$

- $\tau_j = \frac{t}{2N_j}$ (avec tailles variables $\tau_j = N_{j,0}^{-1} + \sigma^2 \sum_{k=0}^{t} N_{j,k}^{-1}$)
- $w_{ij} = 1 - 2\alpha_{ij}$, $z_i = 1 - 2\pi_i$ et $T_{l-1}^1(x)$ sont des polynômes de Gegenbauer satisfaisant la récursion : $T_0^1 = (x)$, $T_1^1(x) = 3x$ et $T_n^1(x) = \frac{1}{n}\left[(2x(n+\frac{1}{2})T_{n-1}^1(x) - (n+1)T_{n-2}^1(x)\right]$ quand $n \geq 2$

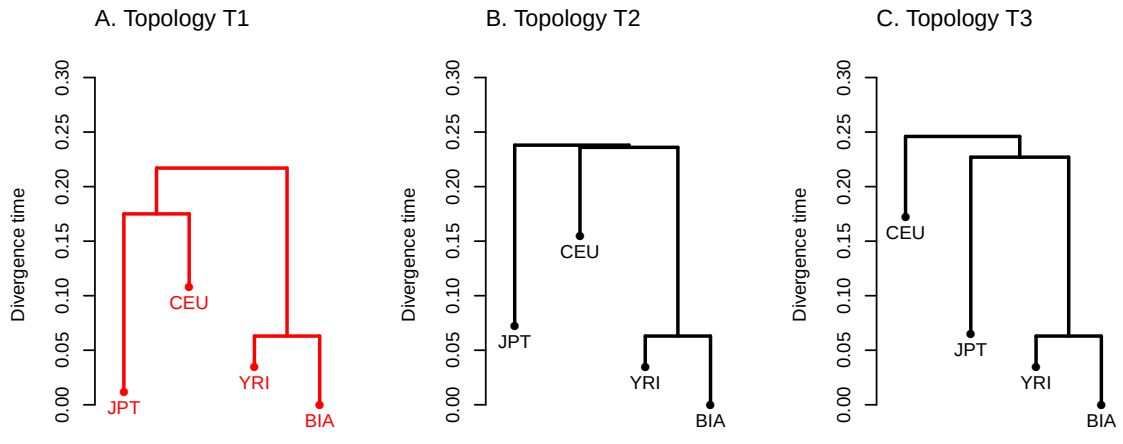
Comparaisons des trois distributions ($\mathcal{N}_T$,$\mathcal{F}$ et $\mathcal{K}$)



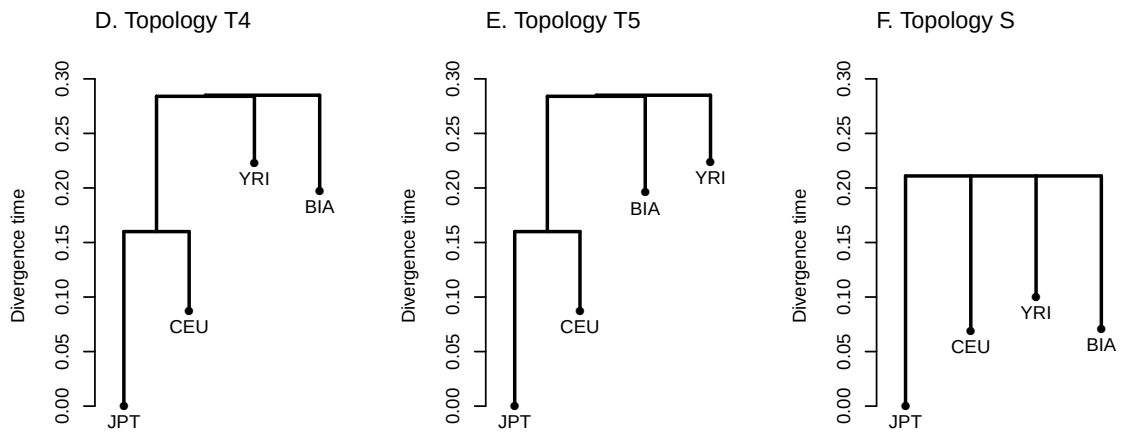
Le modèle KIM_TREE (Gautier & Vitalis, MBE, 2013)



Application (comparaison de modèles) : 4 populations humaines, 450,000 SNPs



DIC_{T1} = 9 463 457 | $DIC_{T2} - DIC_{T1}$ = 3 867 | $DIC_{T3} - DIC_{T1}$ = 5 698



$DIC_{T4} - DIC_{T1}$ = 32 017 | $DIC_{T5} - DIC_{T1}$ = 32 161 | $DIC_{S} - DIC_{T1}$ = 5 297

Conclusions/perspectives sur le modèle KimTree

KimTree permet :

- Une estimation precise des temps de divergence sur un arbre de population
- De comparer différents scénarios (arbre bi- ou multi-furqué)

KimTree est robuste à :

- Flux de gène faible à modéré (écart au modèle WF)
- Biais de recrutement des SNPs (sous réserve que le panel de découverte soit représentatif de l'ensemble de l'arbre)

Développements en cours

- Introduction d'évènements d'admixture
- Estimation des sex-ratios efficaces
- Identification de SNP outliers (PPP-values et calibration)

Identification de locus outliers : PPP-value (Gautier *et al.*, 2010)

Ecart au modèle (H_0 : échangeabilité des loci)

- Mesure de discrépance : $T(y_{ij}, \pi_i, c_j) = \sum_{j=1}^{J} \frac{\left[y_{ij} - \mathbb{E}(y_{ij} \mid \pi_i, c_j)\right]^2}{\mathbb{V}(y_{ij} \mid \pi_i, c_j)}$

  avec $\mathbb{E}(y_{ij} \mid \pi_i, c_j) = \pi_i$ et $\mathbb{V}(y_{ij} \mid \pi_i, c_j) = \frac{\pi_i(1-\pi_i)(1+(n_{ij}-1)c_j)}{n_{ij}}$

- $\mathrm{P}_i = \mathbb{P}\left[T(y_{ij}^r, \pi_i, c_j) > T(y_{ij}, \pi_i, c_j) \mid y_{ij}\right]$

Implémentation (MCMC)

- A chaque itération t, on échantillonne $y_{ij}^r \sim Bin(n_{ij}, \alpha_{ij}^t)$
- On calcule : $\mathrm{P}_i^t = \begin{cases} 1 & \text{si } \sum_{j=1}^{J} \left[T_t(y_{ij}^r, \pi_i^t, c_j^t) - T_t(y_{ij}, \pi_i^t, c_j^t)\right] > 0 \\ 0 & \text{sinon} \end{cases}$
- $\widehat{\mathrm{P}_i} = \frac{1}{N} \sum_{t=1}^{N} \mathrm{P}_i^t$

Calibration (simulations sous le modèle d'inférence)

Signatures de Sélection

Approches "empiriques" (gros jeux de données)

- Outliers=extreme de la distribution empirique observée
- PB : Contrôle des faux positifs/faux négatifs

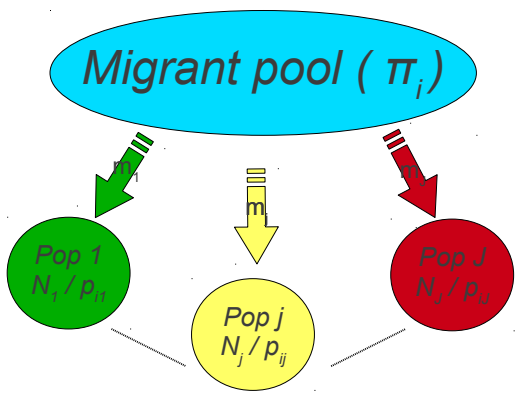
Approches (Bayésienne) par modélisation

- Modélisation de la distribution des fréq. alleliques sour un modèle démographique donnée (e.g. Equilibre migration/dérive, pure dérive)
  $\Longrightarrow$ Mesure d'un écart au modèle neutre pour chaque marqueur (effet locus (*e.g.* Beaumont & Balding, 2004 ; Riebler *et al.*, 2008 ; Foll & Gaggiotti, 2009) ; PPPvalues (Gautier *et al.*, 2010) ; ...)
- Modélisation explicite de la sélection locus/population spécifique
  $\Longrightarrow$ Modèle en île à l'équilibre migration/dérive/sélection (Wright, 1931)

Le modèle SELESTIM (Vitalis *et al.*, Genetics, 2014)

Modèle démographique de Wright (1931)

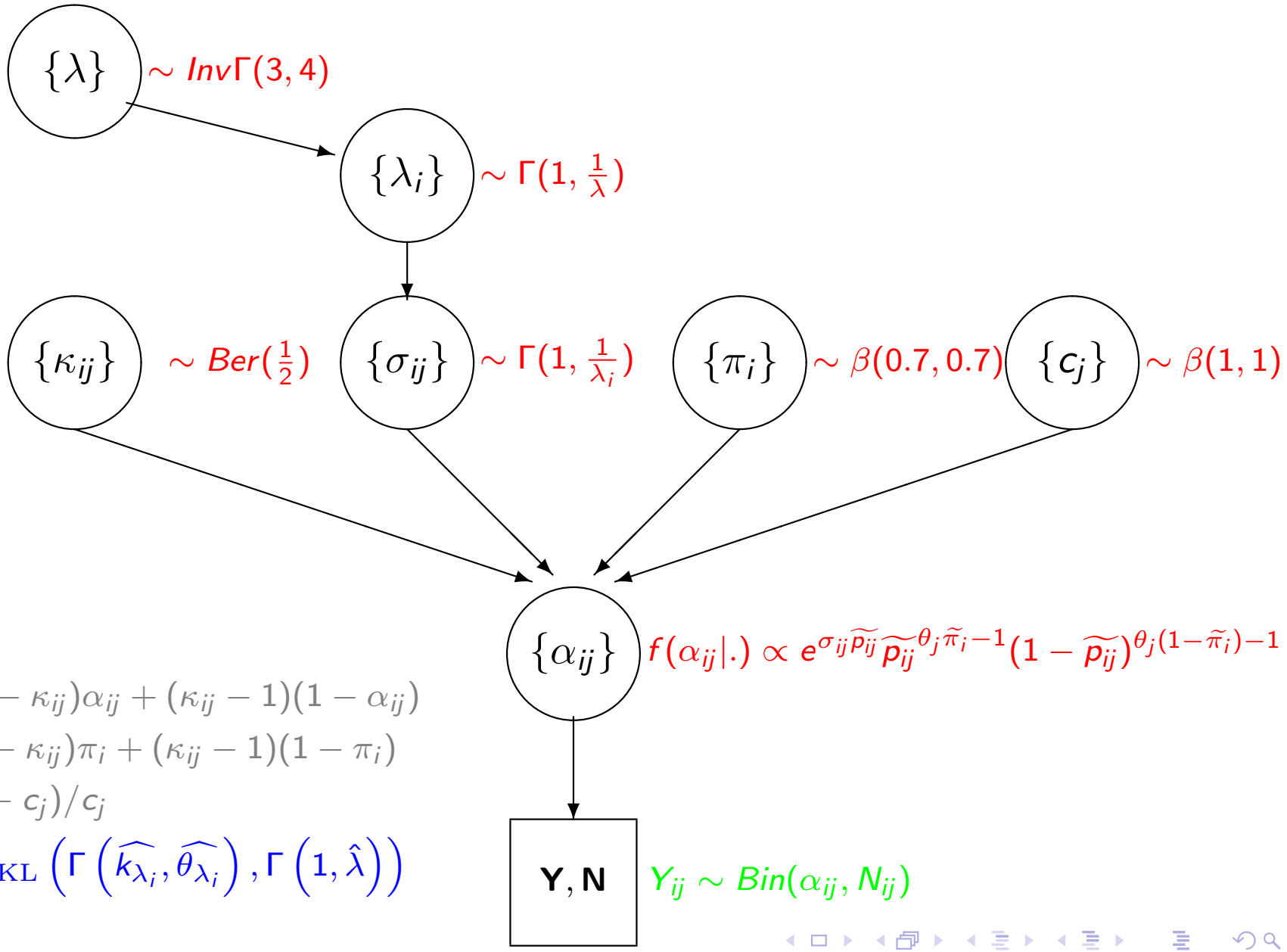
- Modèle à l'équilibre (migration/dérive/sélection), sans mutation
- Estimation des coefficients de sélection σ_{ij}
- Signal de sélection $\Longrightarrow \sigma_{ij} \gg 0$



A l'équilibre migration-dérive-sélection *i.e.* perte d'allèles due à la dérive et/ou à la sélection intra-dème = gain par migration)

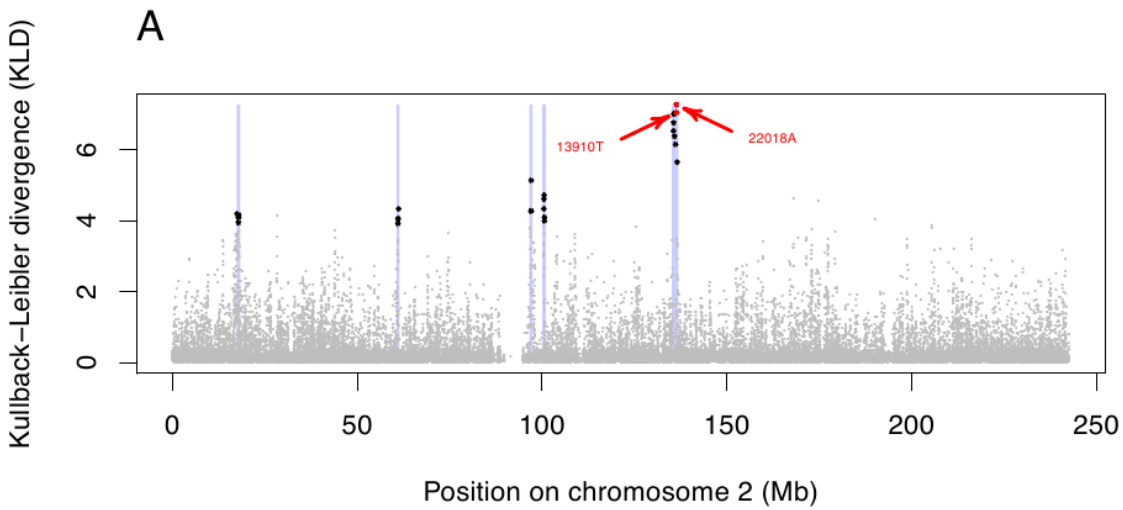
- $f(\alpha_{ij}) \propto e^{\sigma_{ij}\widetilde{p_{ij}}} \widetilde{p_{ij}}^{\theta_j \widetilde{\pi_i} - 1} (1 - \widetilde{p_{ij}})^{\theta_j (1 - \widetilde{\pi_i}) - 1}$
- Si $\sigma_{ij} = 0$, $\alpha_{ij} \sim \beta \left(\theta_j \pi_i, \theta_j (1 - \pi_i) \right)$

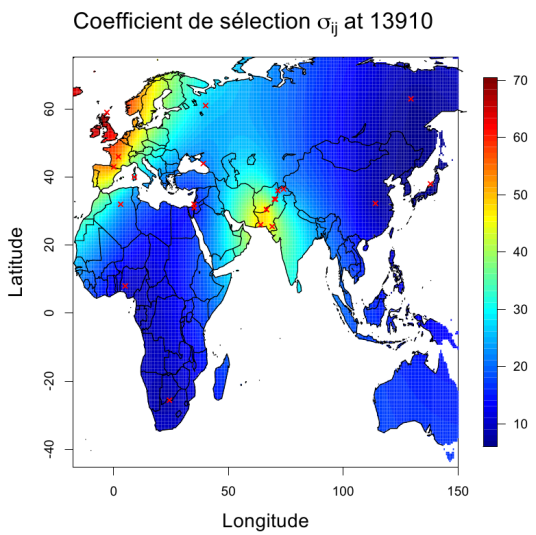
Le modèle SELESTIM (Vitalis *et al.*, Genetics, 2014)

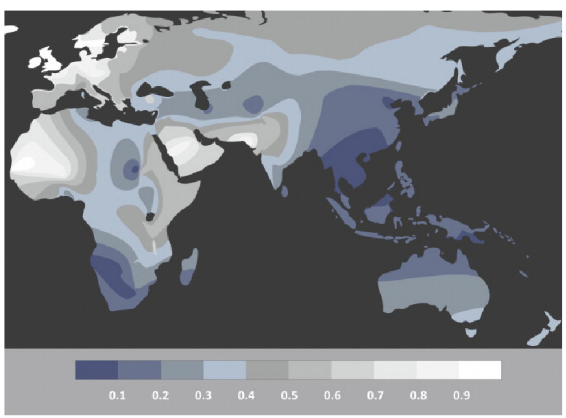


- $\widetilde{p_{ij}} = (2-\kappa_{ij})\alpha_{ij} + (\kappa_{ij}-1)(1-\alpha_{ij})$
- $\widetilde{\pi}_i = (2-\kappa_{ij})\pi_i + (\kappa_{ij}-1)(1-\pi_i)$
- $\theta_j = (1-c_j)/c_j$
- KLD=$\mathrm{D_{KL}}\left(\Gamma\left(\widehat{k_{\lambda_i}}, \widehat{\theta_{\lambda_i}}\right), \Gamma\left(1, \hat{\lambda}\right)\right)$

Application : 23 human populations, 52,631 SNPs sur HSA2

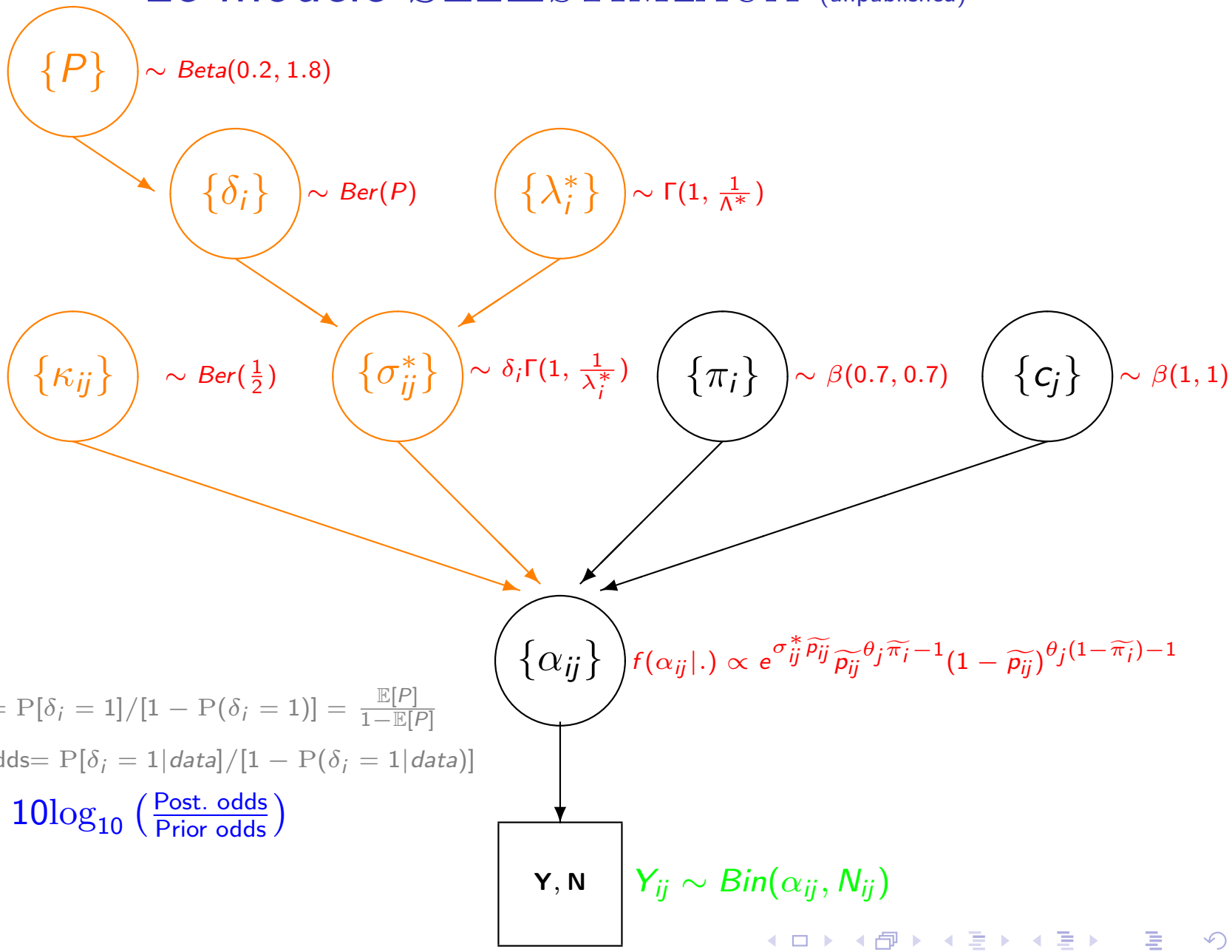






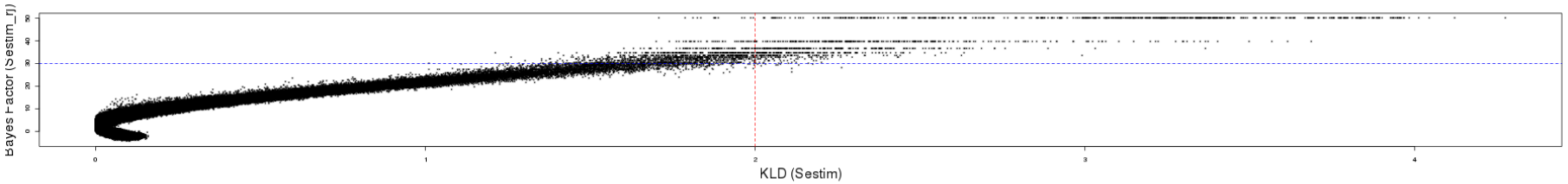
Distribution of lactase persistence phenotype (Itan *et al*. 2010)

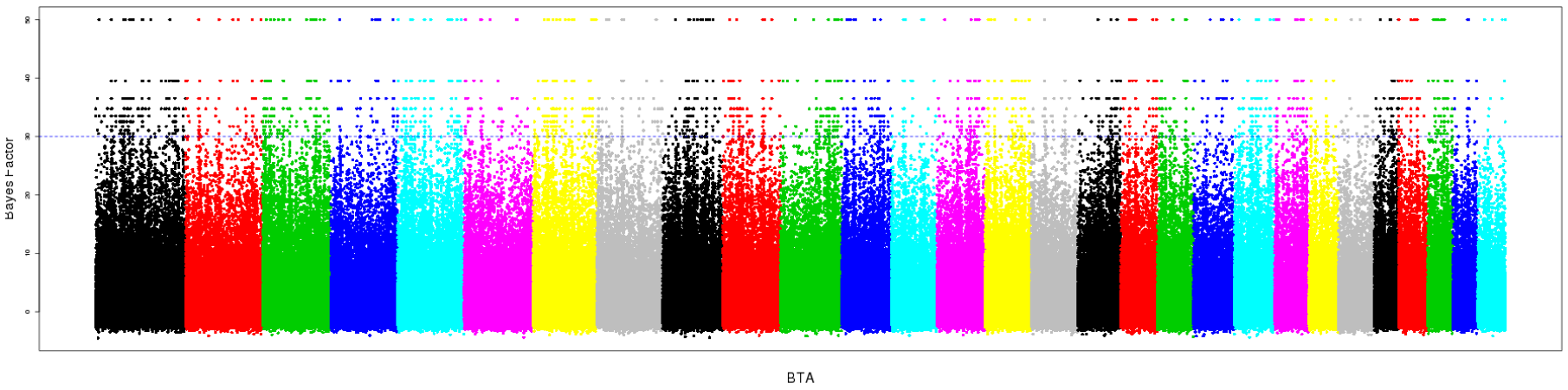
Le modèle SELESTIM_AUX (unpublished)

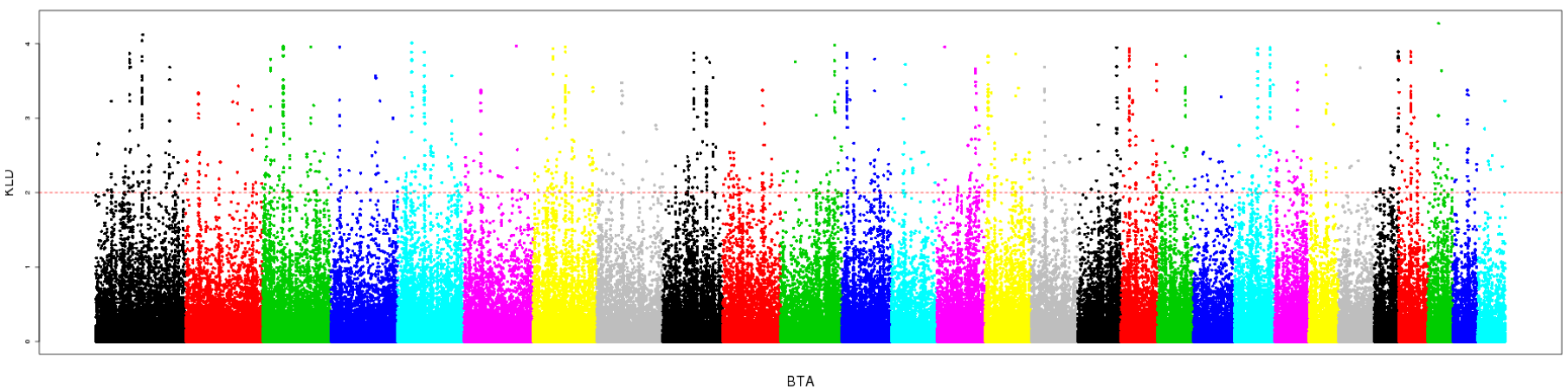


- Prior odds$= \mathrm{P}[\delta_i = 1]/[1 - \mathrm{P}(\delta_i = 1)] = \frac{\mathbb{E}[P]}{1 - \mathbb{E}[P]}$
- Posterior odds$= \mathrm{P}[\delta_i = 1|data]/[1 - \mathrm{P}(\delta_i = 1|data)]$
- $\mathrm{BF}_{dB} = 10 \log_{10} \left(\frac{\text{Post. odds}}{\text{Prior odds}} \right)$

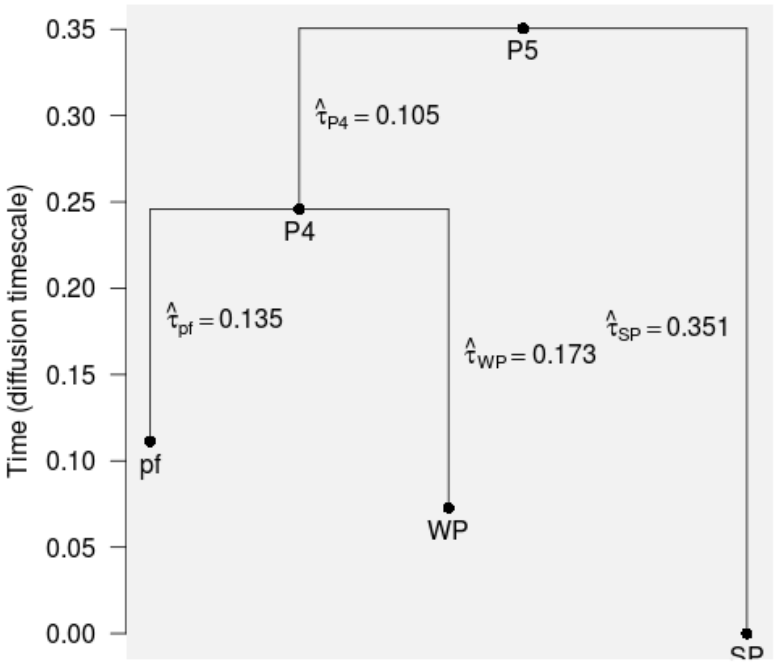
Application : 6 pop. bovines africaines, 658,520 SNPs

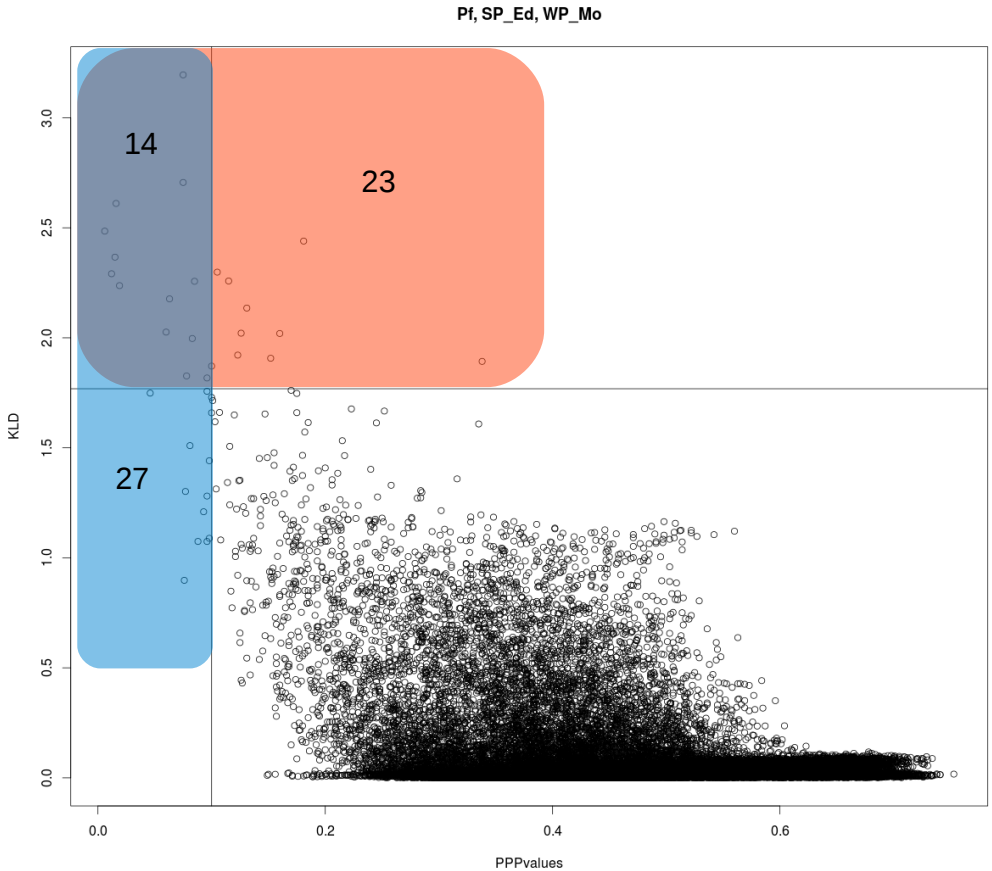






PPPvalues (Kimtree) VS KLD (SelEstim) : 3 pop. PPM, 53,520 SNPs





Développements en cours autour de ce type de modèle

Utilisation de modèle démographique plus complexe

- Populations structurées
- Généralisation multi-allélique

Exploitation de l'information apportée par l'organisation spatiale des marqueurs le long du génome

- Post-traitement (e.g. lissage)
- Intégration de la dépendance spatiale des marqueurs dans les modèles (AR, HMM)
- Utilisation de données de séquence (extension multi-allélique)

Remerciements

- Renaud Vitalis
- Mark Beaumont
- Kevin Dawson
- Jean-Louis Foulley