UNIVERSITÉ DE MONTPELLIER

ÉCOLE DOCTORALE GAIA

Habilitation à diriger des recherches (HDR)

# Inférence en génétique spatiale des populations vue par le prisme de la coalescence

Raphaël LEBLOIS, CBGP - INRAE

INRAE
la science pour la vie, l'humain, la terre

CBGP
Centre de Biologie pour la Gestion des Populations

UNIVERSITÉ DE MONTPELLIER, ÉCOLE DOCTORALE GAIA

Habilitation à diriger des recherches (HDR)

# Inférence en génétique spatiale des populations vue par le prisme de la coalescence

Raphaël LEBLOIS, CBGP - INRAE

Soutenue le 8 avril 2024 devant le jury composé de

| | | |
|---|---|---|
| **Lounès CHIKHI** | Directeur de Recherche, CNRS Toulouse | Examinateur |
| **Flora JAY** | Chargée de Recherche, Univ. Paris Saclay | Examinatrice |
| **Mathieu JORON** | Directeur de Recherche, CNRS Montpellier | Examinateur |
| **Amaury LAMBERT** | Professeur, Collège de France | Rapporteur |
| **Anna-Sapfo MALASPINAS** | Professeure, Univ. de Lausanne | Rapportrice |
| **Denis ROZE** | Directeur de Recherche, CNRS Roscoff | Rapporteur |
| **Céline SCORNAVACCA** | Directrice de Recherche, CNRS Montpellier | Examinatrice |
| **Laure SEGUREL** | Chargée de Recherche, CNRS Lyon | Examinatrice |
| Membre invité: **John NOVEMBRE** | Professeur, Université de Chicago | |

# Plan de l'exposé

1. Parcours: recherche, enseignements, encadrements,...
2. Introduction (coalescence, dispersion, inférence)
3. La régression : simple, efficace mais limitée
4. Vraisemblance: puissante mais complexe, et finalement limitée
5. Inférence par simulation: puissante, flexible mais encore lourde...
6. Conclusions et perspectives

# Parcours

# Parcours

M1 M2 Thèse | Post-doc | MCF | chargé de recherche

1999 2004 2006 2010

25 ans sur l'inférence démographique à partir de données génétiques…


invasion du crapaud de la canne à sucre en Australie

Spécialisation sur la dispersion

à petite échelle géographique et évolutive

# Parcours: recherche, enseignements, encadrements,...

- Inférences démographiques à partir de données génétiques

  interface biologie évolutive, écologie moléculaire, statistique, bio-informatique

  → comprendre le fonctionnement démographique des populations
  → logiciels libres (open source) “faciles à utiliser”

# Parcours: recherche, enseignements, encadrements,...

- Inférences démographiques à partir de données génétiques
- Enseignement (15 - 30 H/an, M1, M2, ED)
- Encadrement (6 M1, 18 M2, 5 doctorants, 3 post-doctorants)

# Parcours: recherche, enseignements, encadrements,...

- Inférences démographiques à partir de données génétiques
- Enseignement (15 - 30 H/an, M1, M2, ED)
- Encadrement (6 M1, 18 M2, 5 doctorants, 3 post-doctorants)
- Tâches collectives (CU, outils de calcul, colloques ex: MCEB)

+ Référent DD du CBGP depuis 3 ans

# Parcours: recherche, enseignements, encadrements,...

- Inférences démographiques basées sur la coalescence
- Enseignement (15 - 30 H/an, M1, M2, ED)
- Encadrement (6 M1, 18 M2, 5 doctorants, 3 post-doctorants)
- Référent DD

+ tout ce travail est entièrement collaboratif (étudiants + FR, PAC, AE, RV, MN, MG,...)

# Introduction

# Inférence

# en génétique spatiale des populations

# vue par le prisme de la coalescence

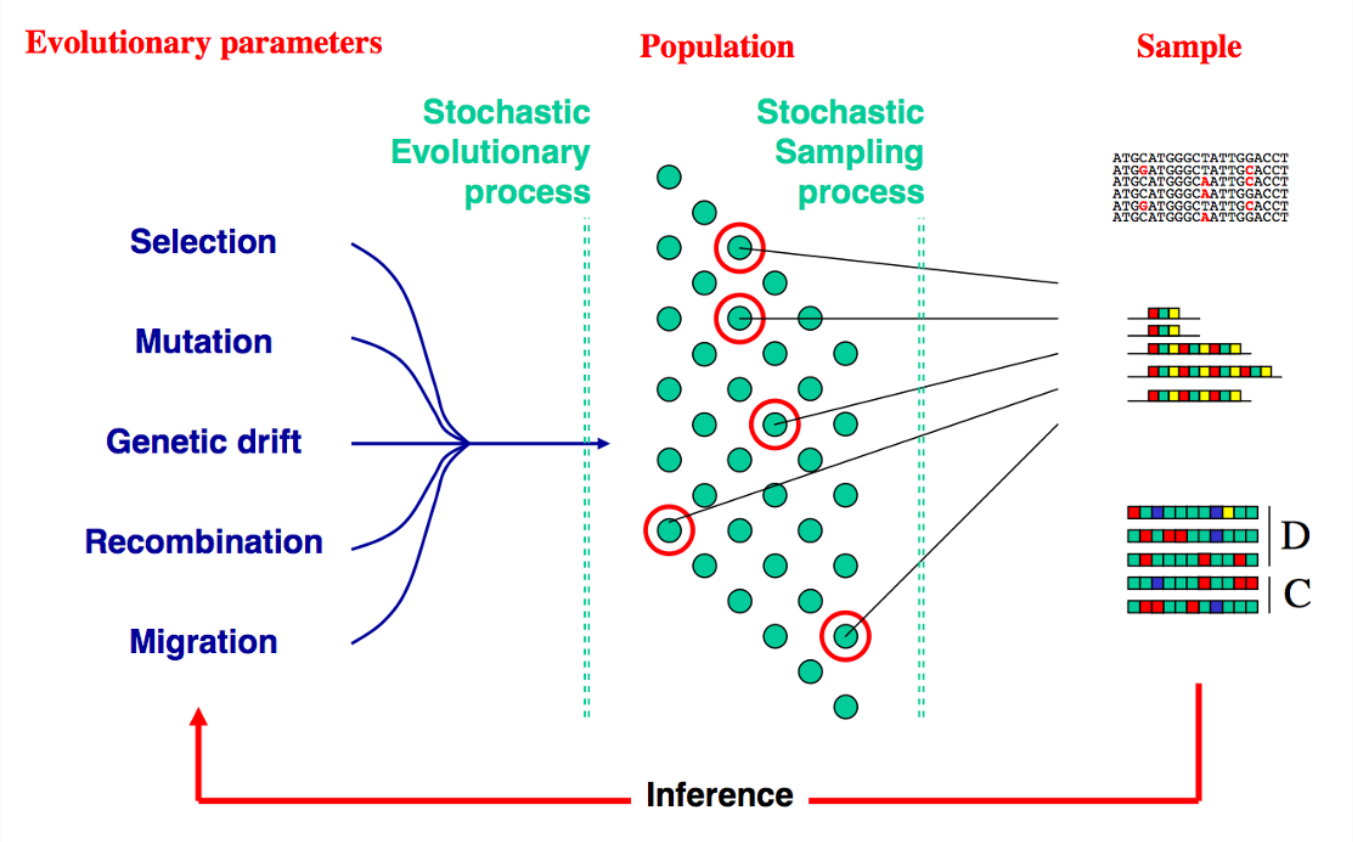
Développer et tester des logiciels implémentant des méthodes d'inférence des processus démographiques **locaux et récents**, notamment les caractéristiques de dispersion et de densité, à partir de données génétiques en populations naturelles

pour des estimation de plus en plus complètes, précises mais toujours **ROBUSTES**

# Inférence
# en génétique spatiale des populations
# vue par le prisme de la coalescence

La **robustesse** est un facteur clé des analyses de biologie évolutive en populations naturelles

Principe de l'inférence en génétique des populations




→ Il faut des données, des modèles démo-génétiques, des méthodes d'inférence statistique

# Introduction

Paramètres démographiques : tailles de population, **densités**, **taux de dispersion** (migration), **distributions de dispersion**, … et leurs changements au cours du temps

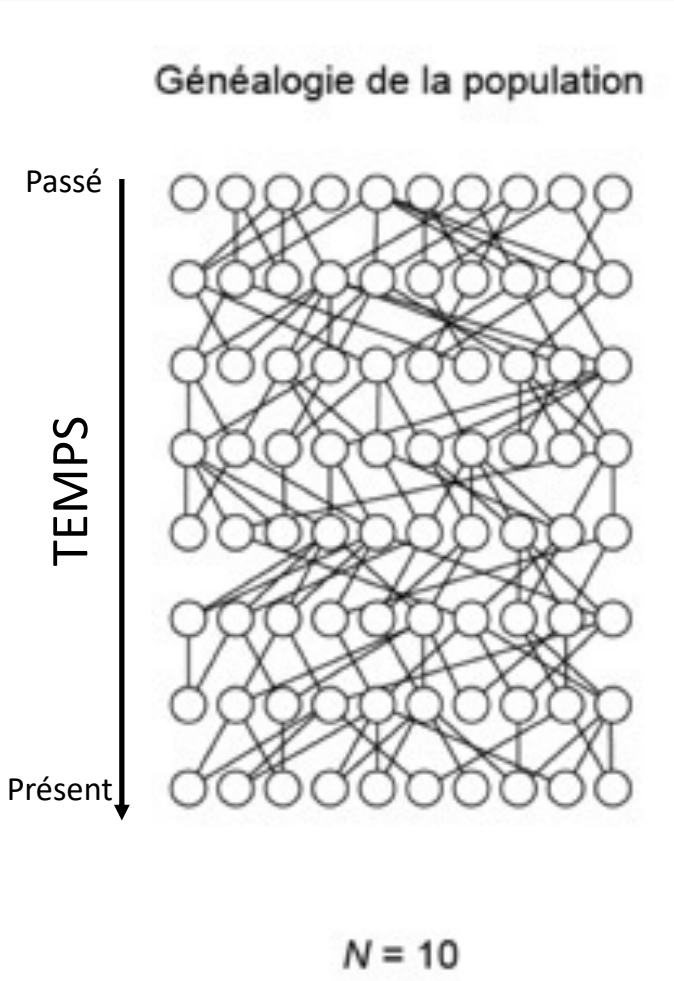
- Facteurs clés de l'adaptation locale des populations à leur environnement (en interaction avec les paramètres sélectifs)
- Fort intérêt pour la gestion des populations:

  biologie de la conservation, bio-invasions, agro-écologie…
  - "connectivité" entre populations
  - vitesse des invasions

# Inférence
# en génétique spatiale des populations
# vue par le prisme de la coalescence

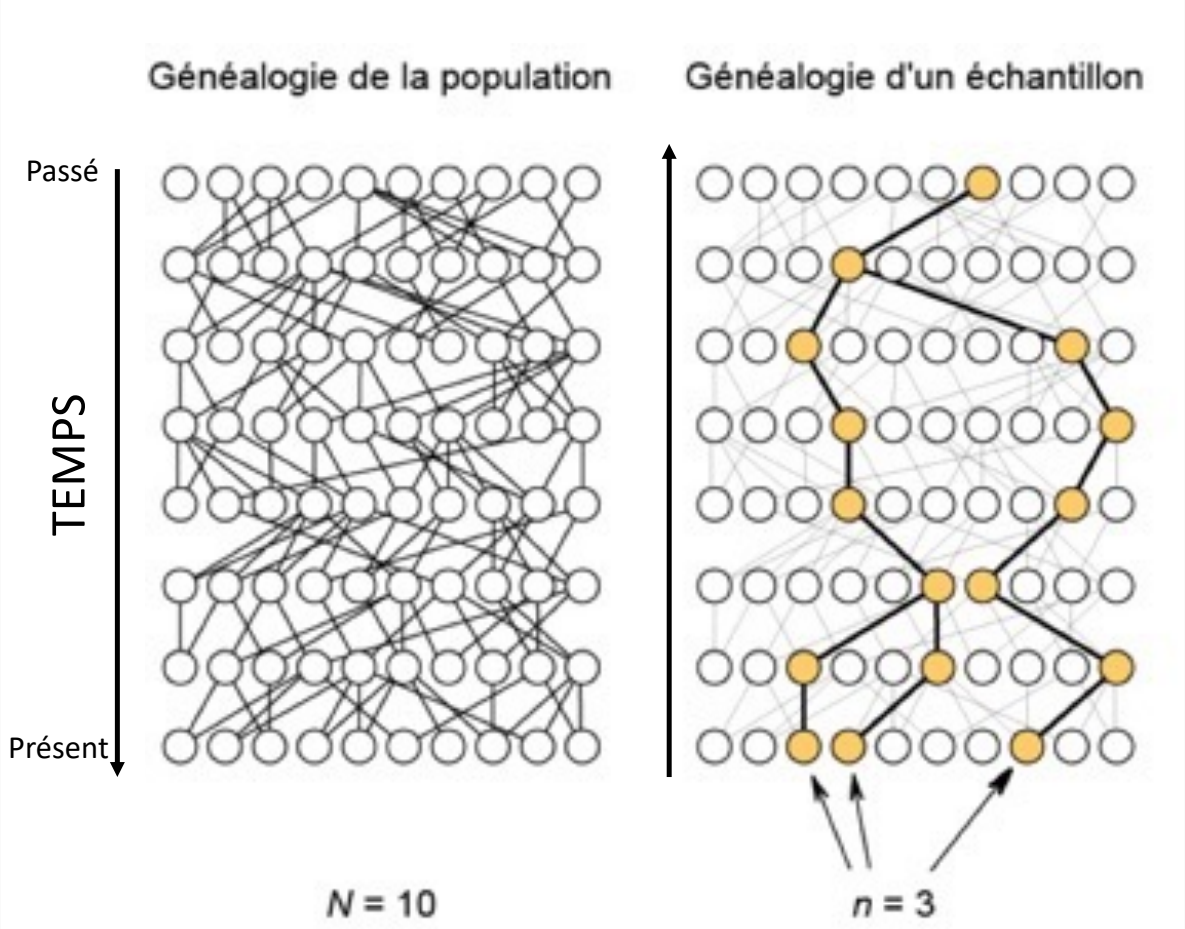
- Théorie de la coalescence
- Génétique spatiale des populations et dispersion
- Méthodes d'inférence statistique

# La coalescence



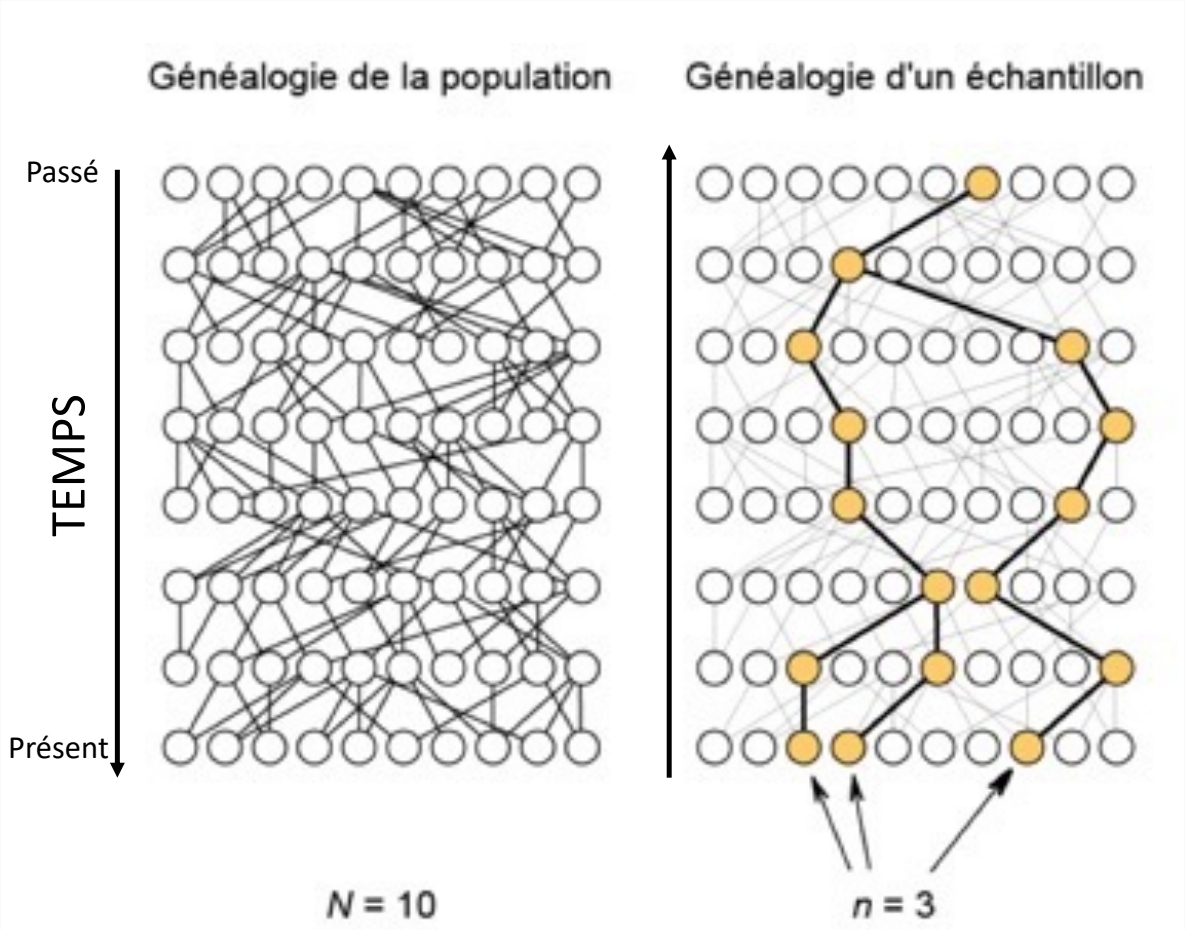

Approche classique ”en avant”: Reconstruire la transmission des lignées génétiques par descendance à chaque génération, dans le sens de l’évolution et dans toute la population

# La coalescence génération par génération




La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

# La coalescence génération par génération

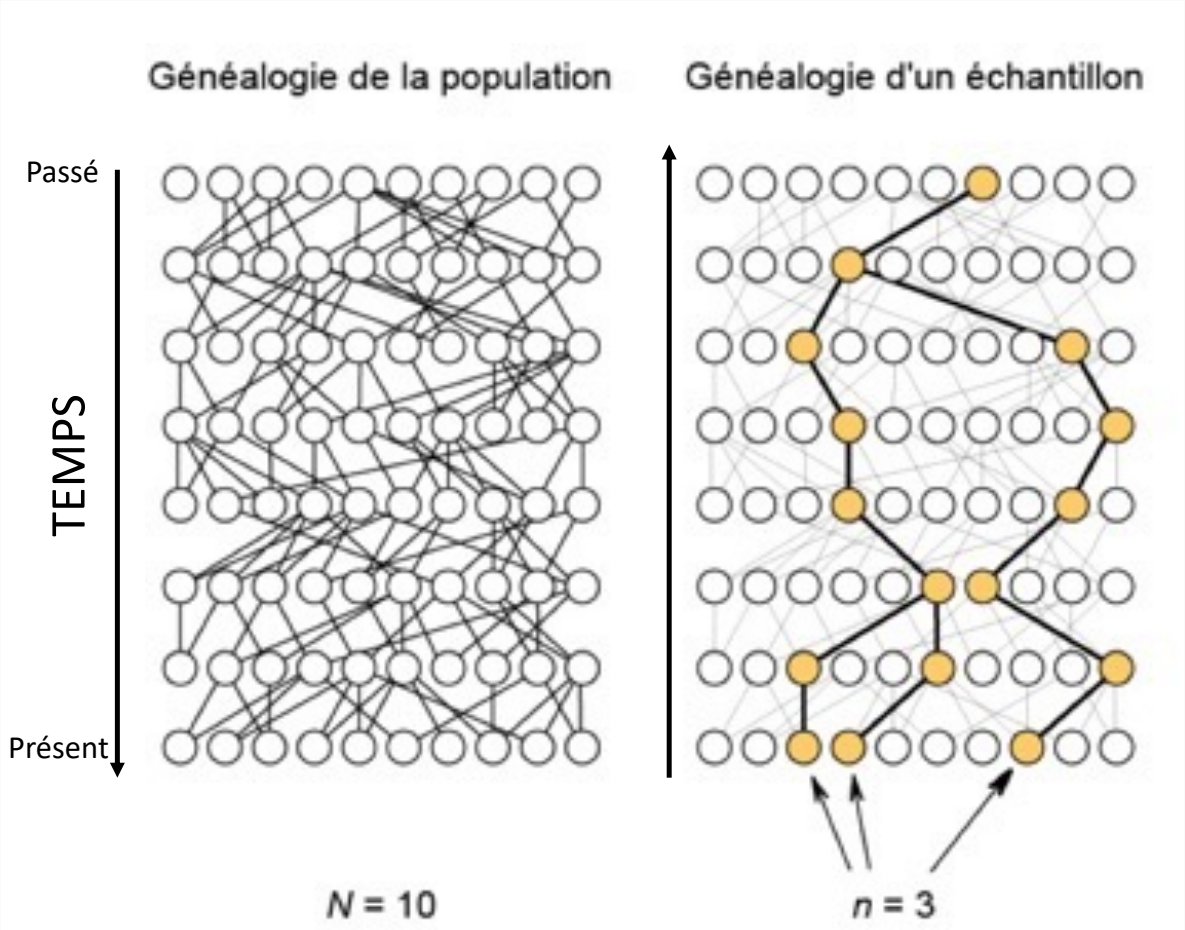



coa $N$? (dispersion $m$ ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

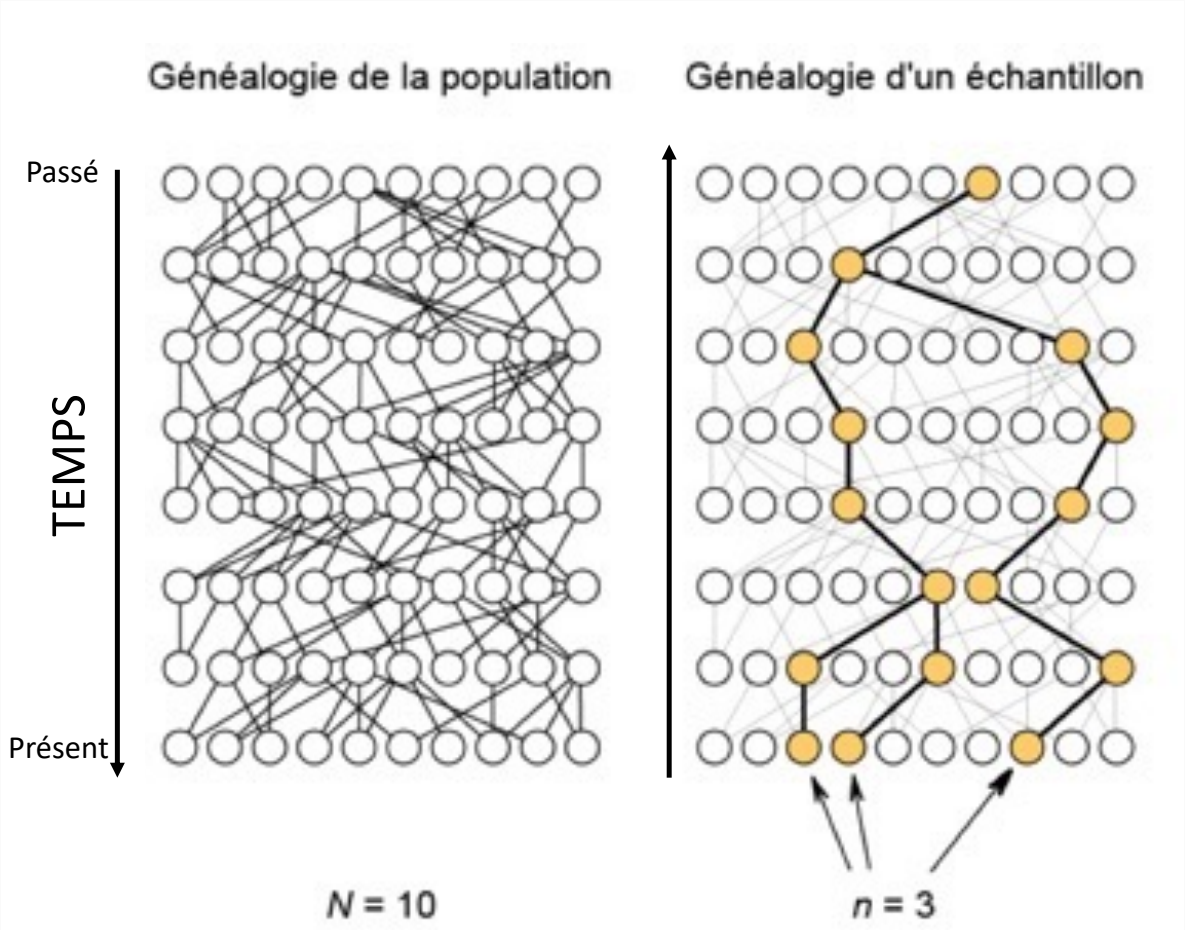



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

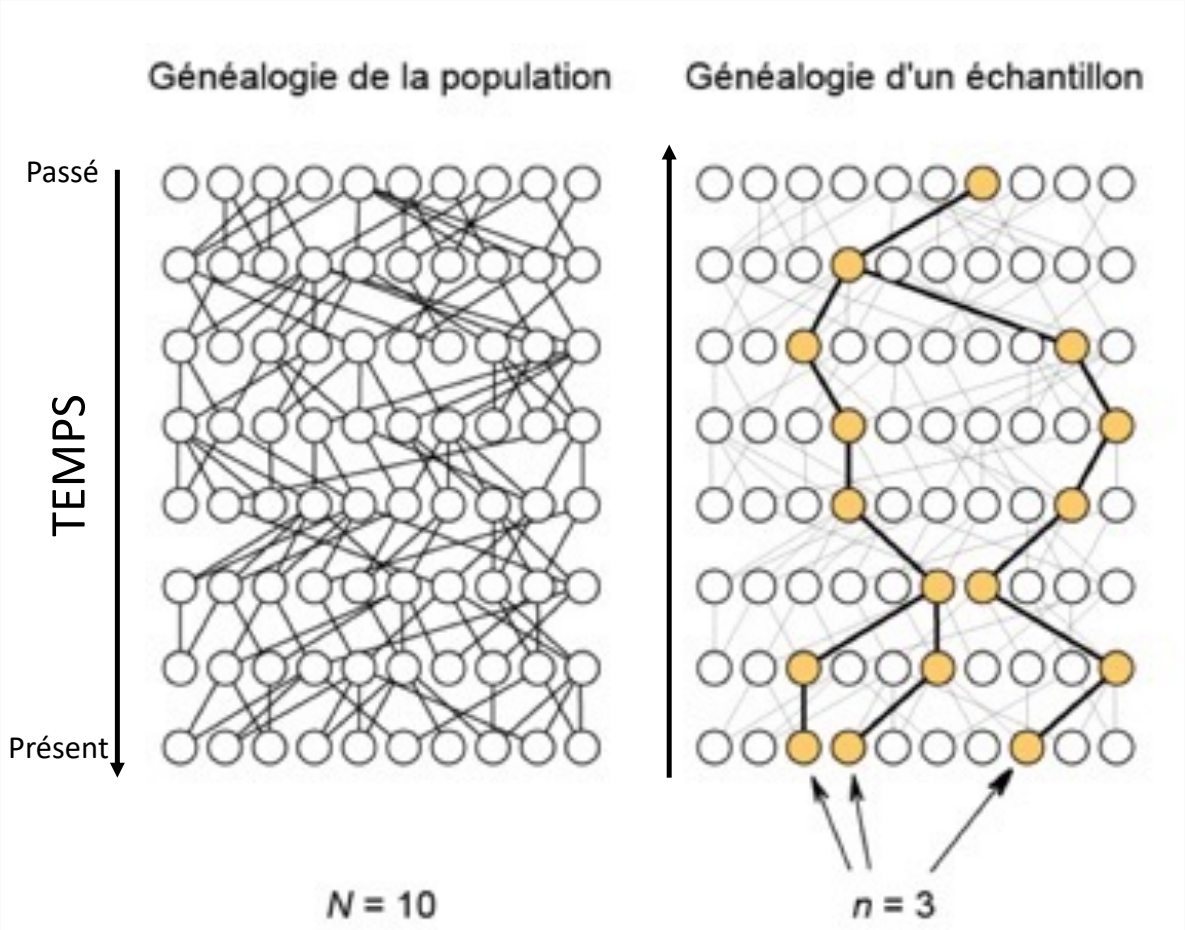



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

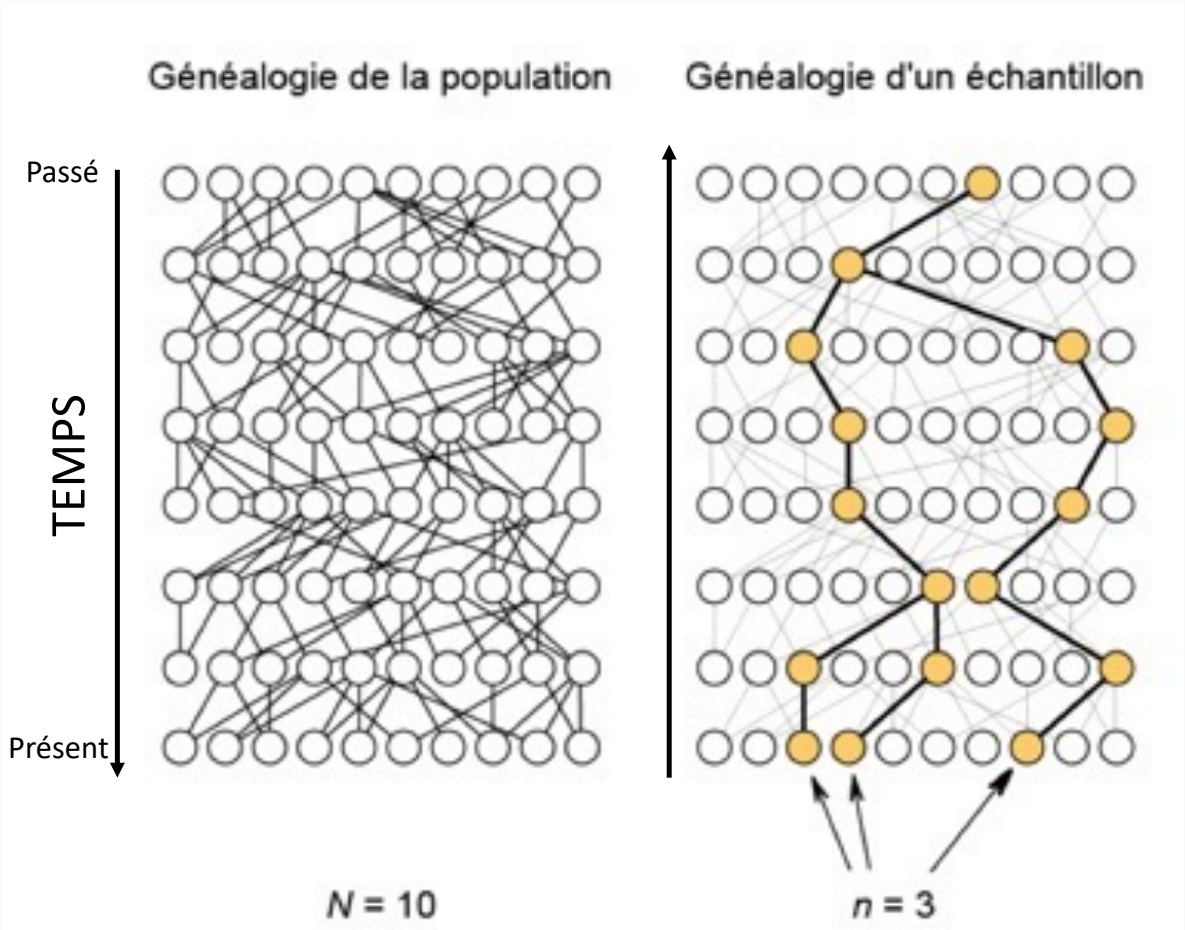



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

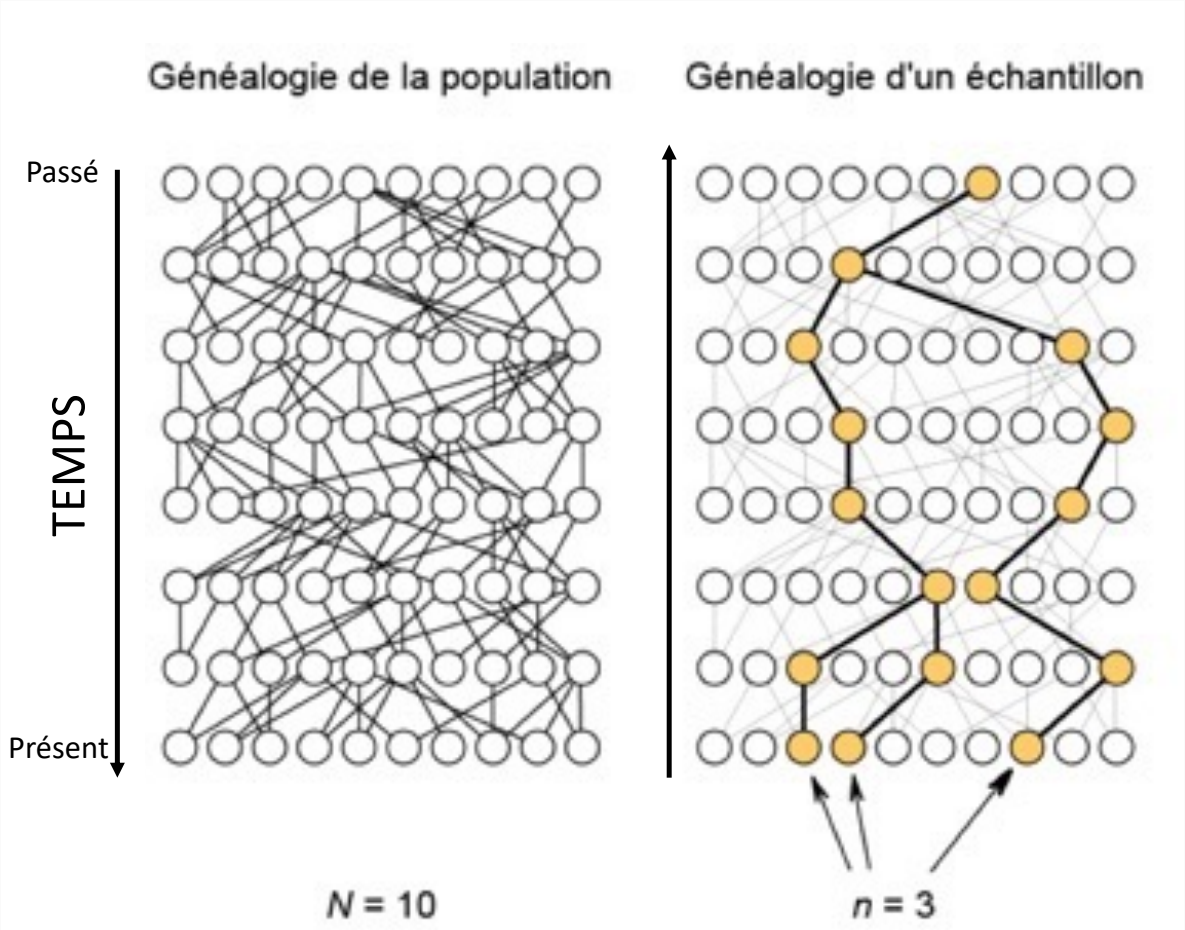



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

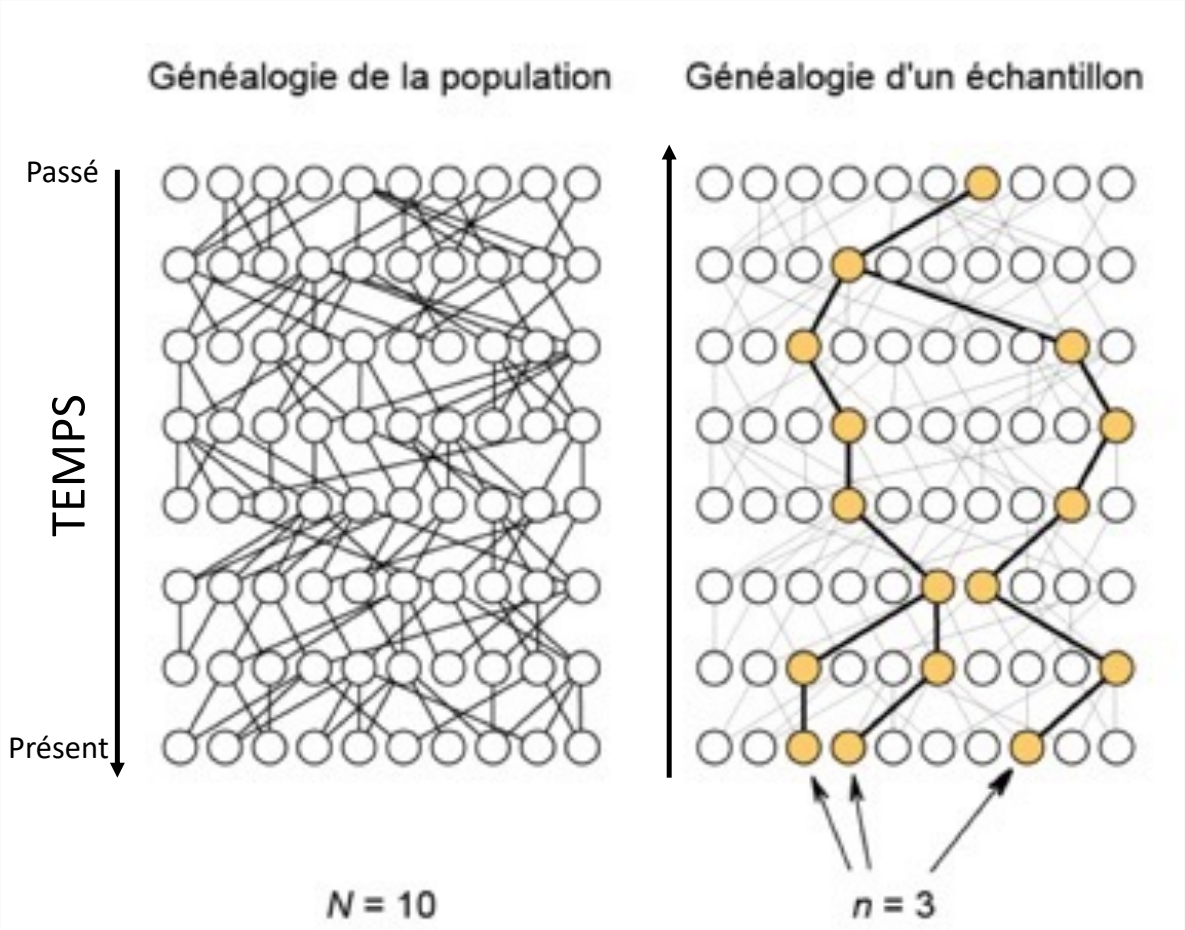



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

# La coalescence génération par génération

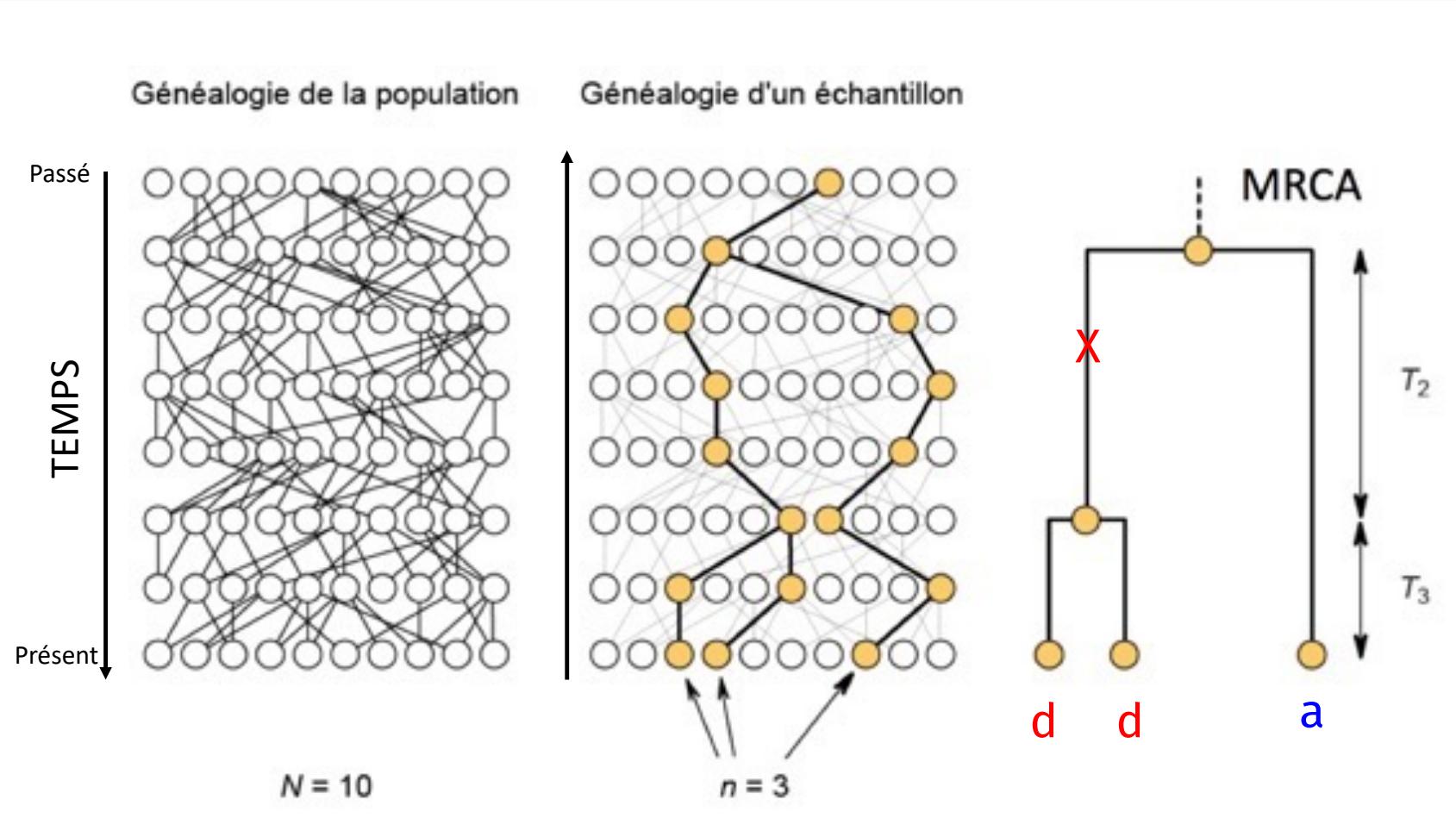



coa? (dispersion ?)

La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

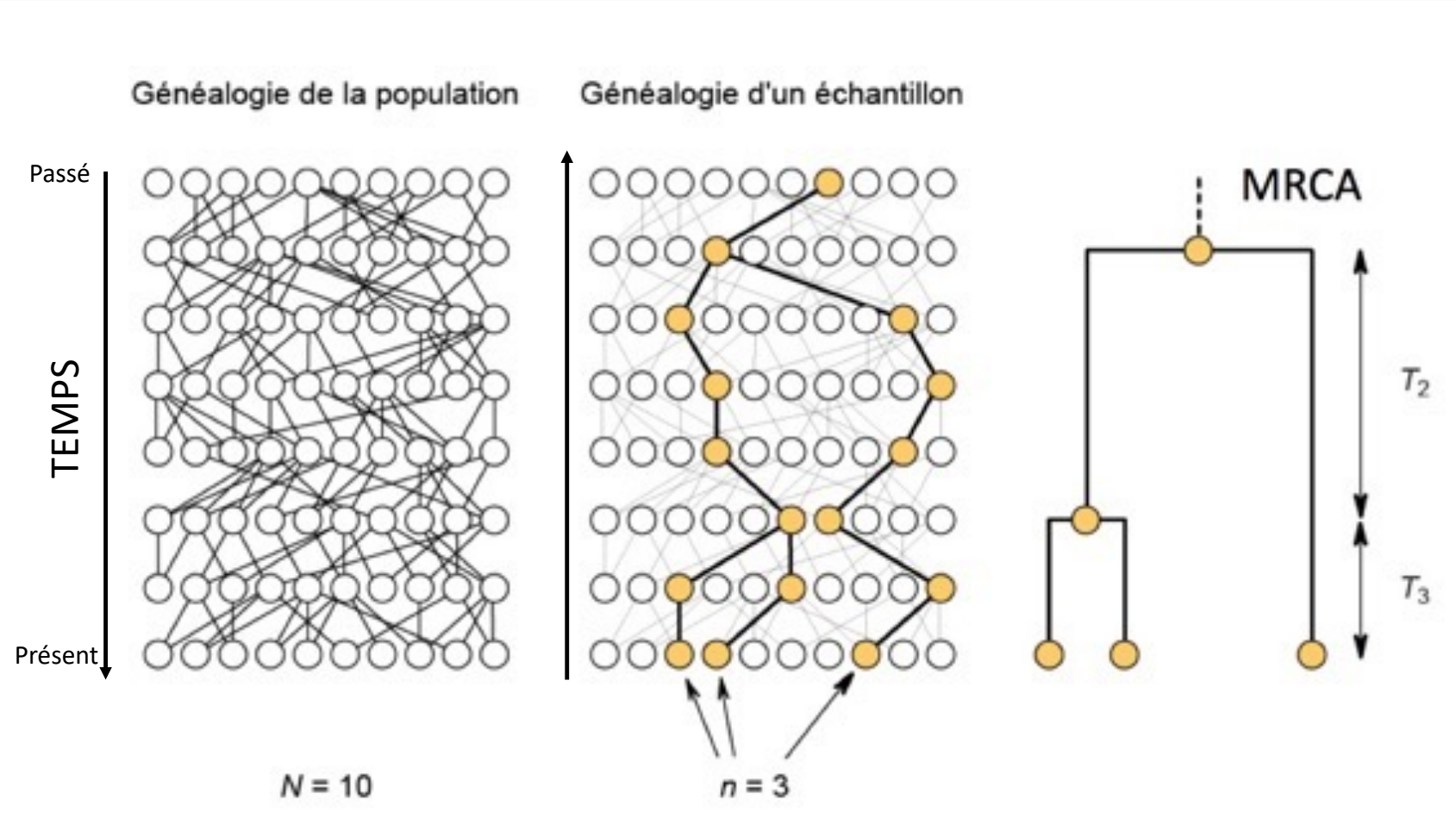
# La coalescence génération par génération




La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération

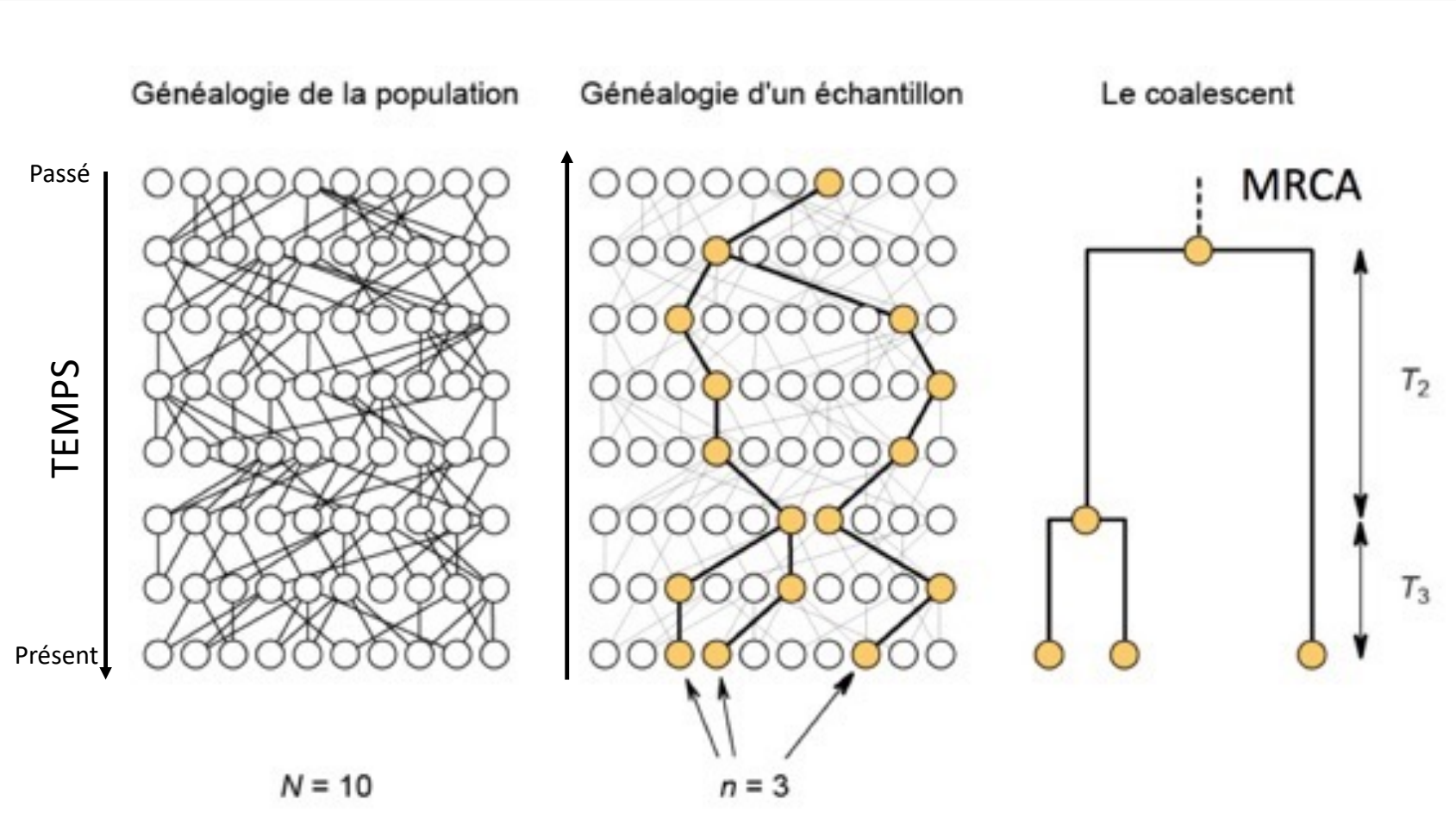
# La coalescence génération par génération




La coalescence (”en arrière”) : Reconstruire la généalogie des lignées ancestrales d’un échantillon

Approche exacte génération par génération (rapide et flexible)

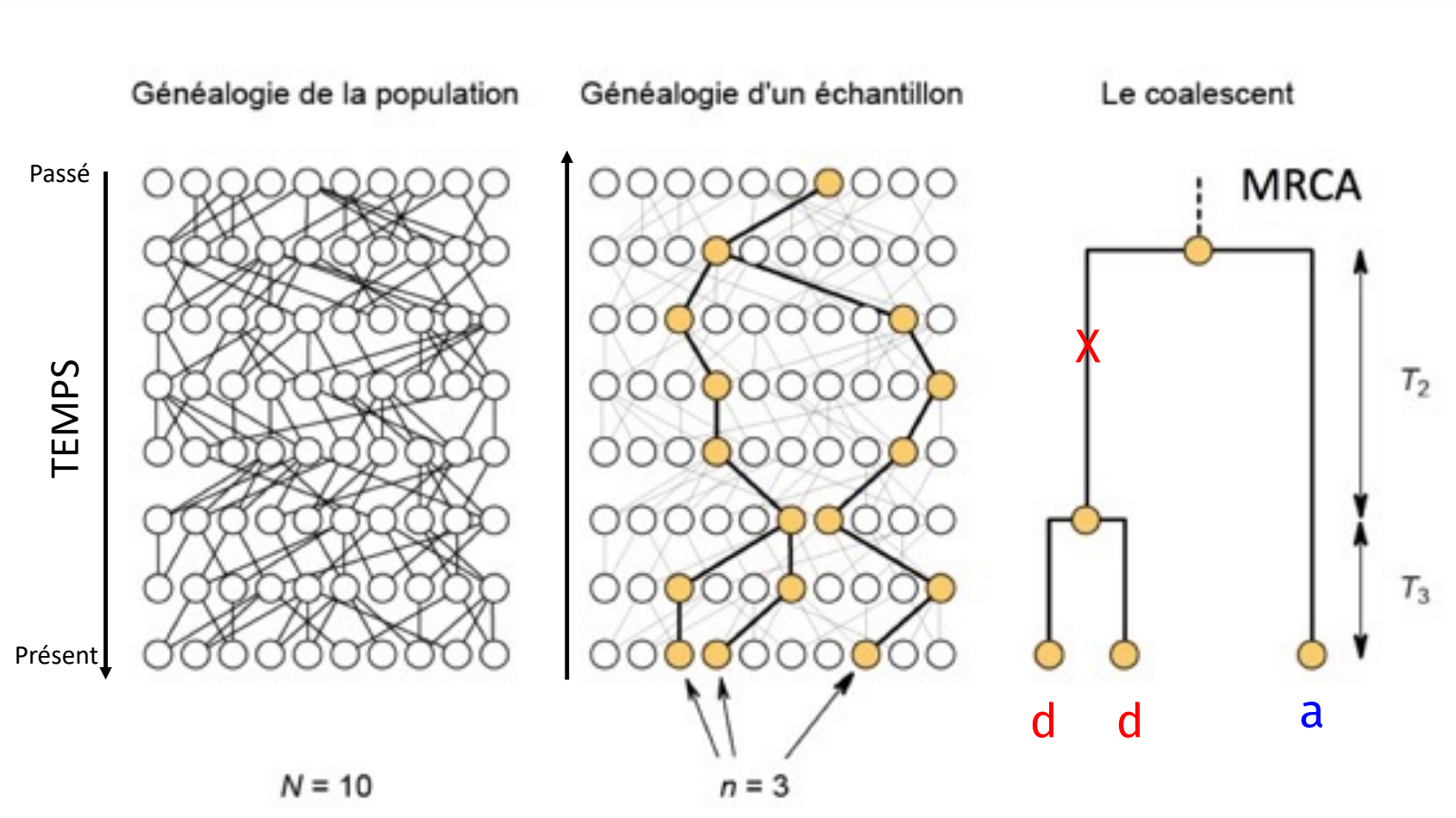
# La coalescence et le n-coalescent




Le $n-$coalescent (”en arrière”) : Reconstruire les temps de coalescence des lignées ancestrales d’un échantillon

Sous l’hypothèse $N \longrightarrow \infty$ : $\Pr(T_k = t) \approx \frac{k(k-1)}{2N} e^{-t\frac{k(k-1)}{2N}}$

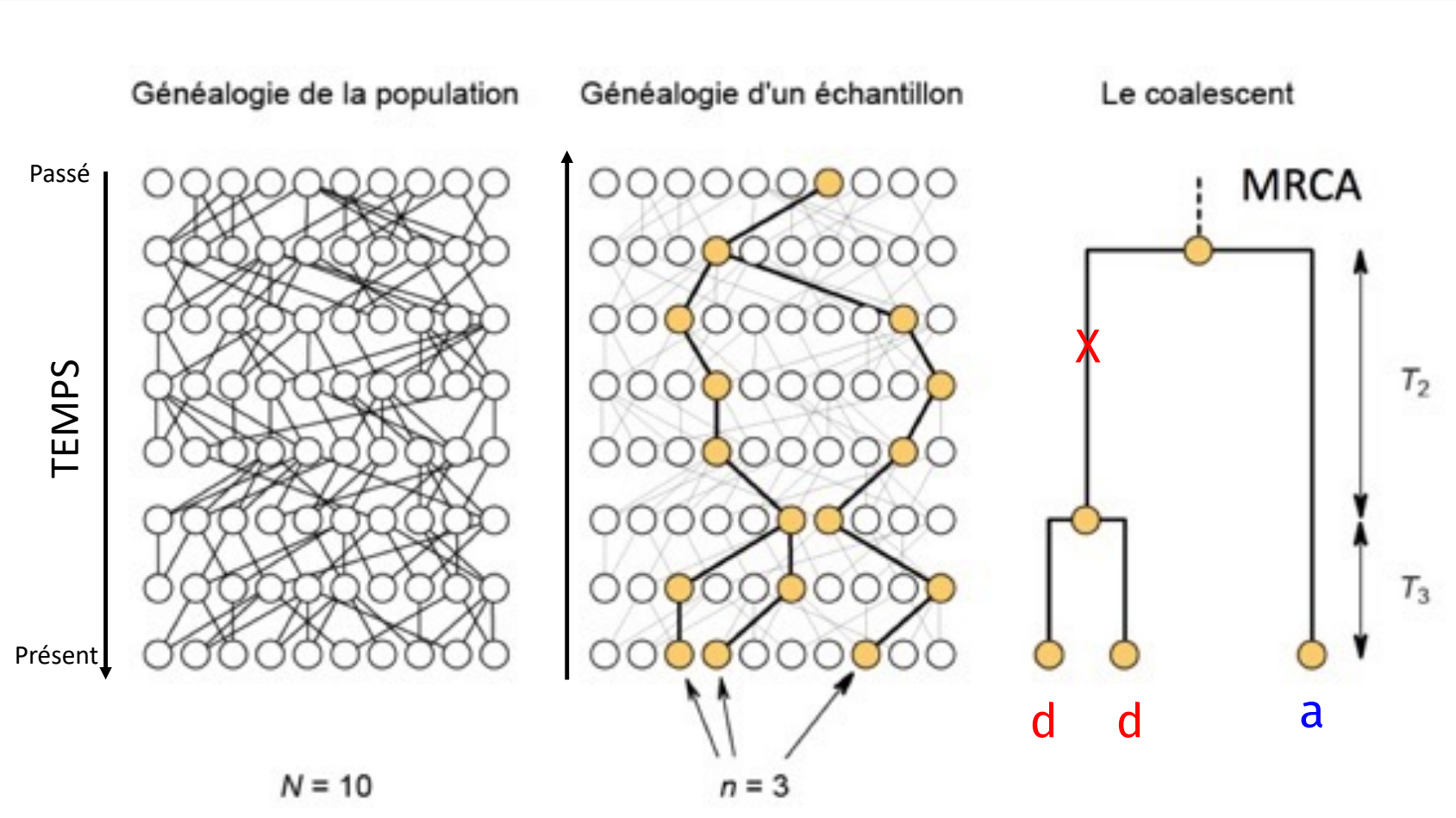
# La coalescence et le n-coalescent




Le $n-$coalescent (”en arrière”) : Reconstruire les temps de coalescence des lignées ancestrales d’un échantillon

$n-$ coalescent plus rapide mais moins flexible

# La coalescence et le n-coalescent




Le $n-$coalescent (”en arrière”) : Reconstruire les temps de coalescence des lignées ancestrales d’un échantillon

Information des données = évènements de coalescence et mutations

# La coalescence et le n-coalescent

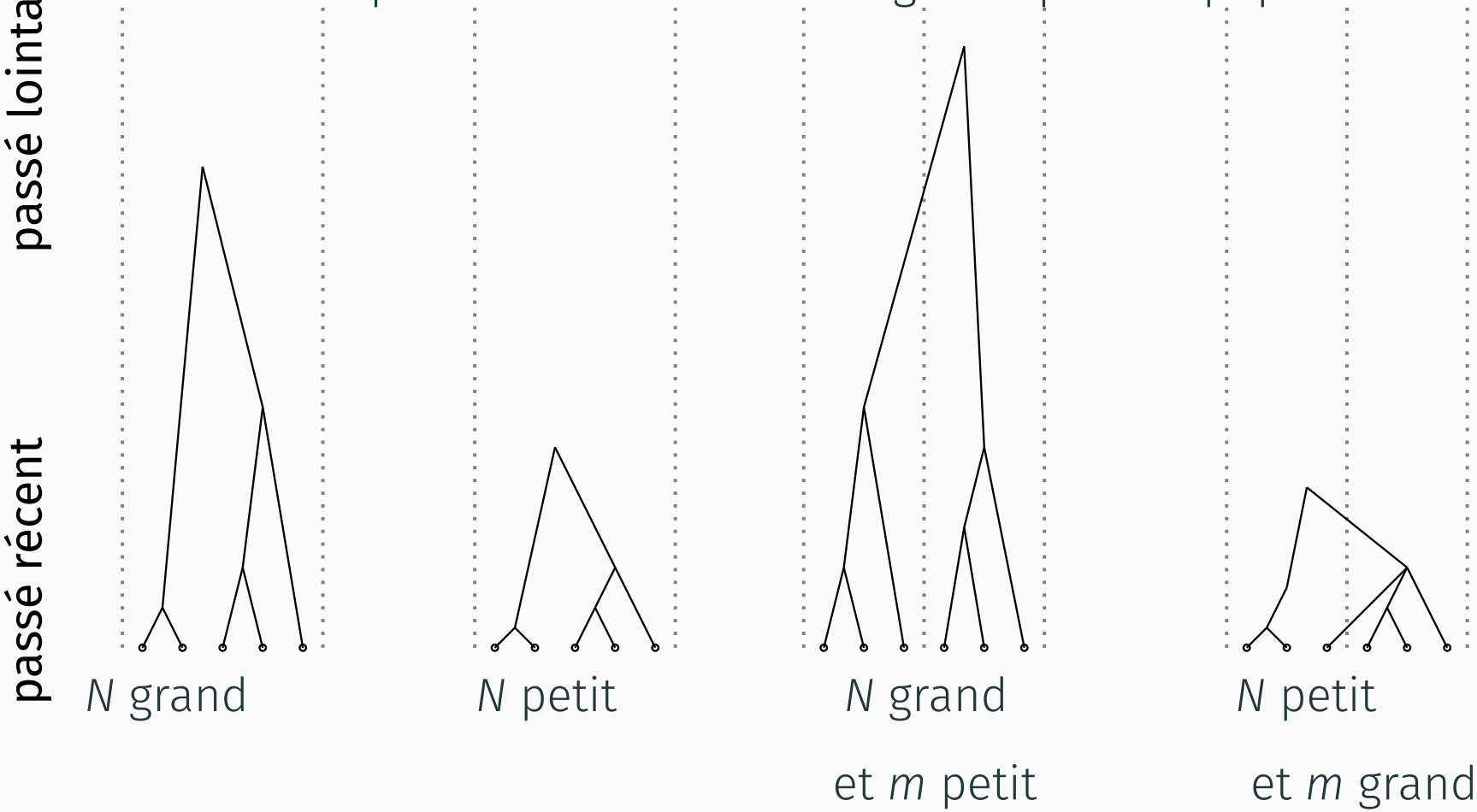
La théorie de la coalescence permet notamment:

- Simulations très efficaces
- intuition patrons et inférences en génétique des populations

# La coalescence et le n-coalescent

La théorie de la coalescence permet notamment:

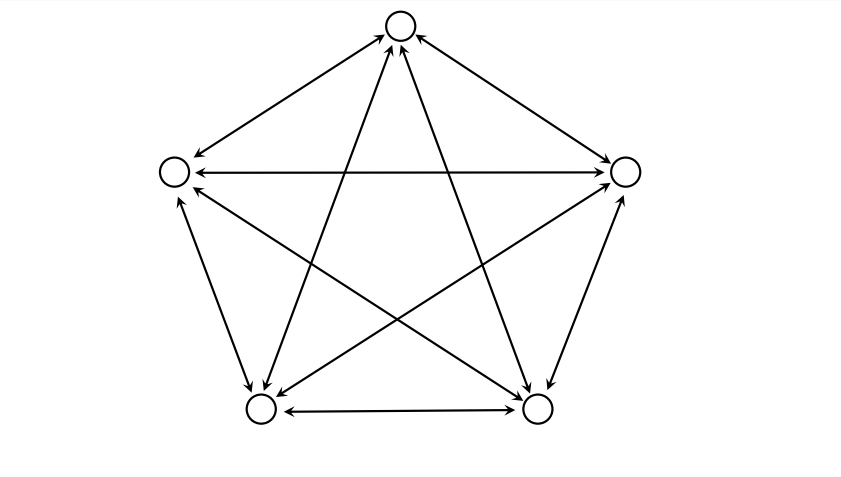
- Simulations très efficaces
- intuition patrons et inférences en génétique des populations



# Dispersion et isolement par la distance

- Théorie de la coalescence
- **Génétique spatiale des populations et dispersion**
- Méthodes d'inférence statistique

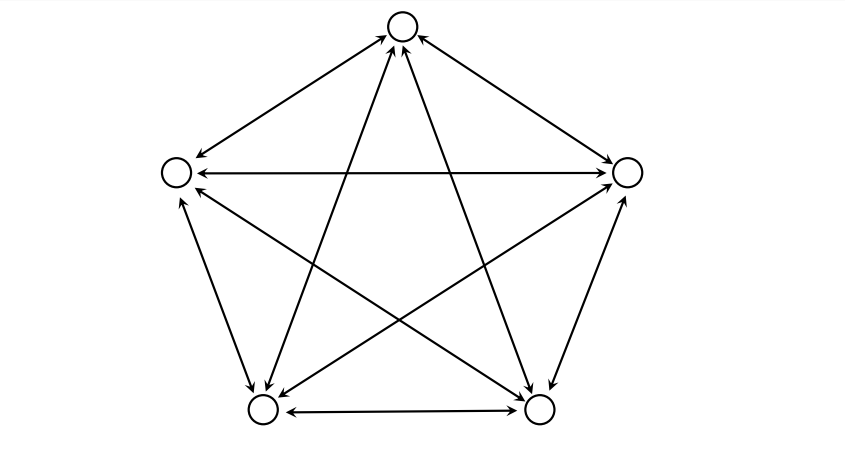
# Dispersion et isolement par la distance



Les modèles de populations structurées classiques (ex: modèle en îles) ne sont pas assez réalistes pour étudier la dispersion à petite échelle géographique  car

1. la dispersion est limitée dans l'espace chez une majorité d'espèces
2. les individus ne sont pas toujours clairement distribués en sous-populations panmictiques (habitat continu)

# Dispersion et isolement par la distance



Les modèles de populations structurées classiques (ex: modèle en îles) ne sont pas assez réalistes pour étudier la dispersion à petite échelle géographique

L'**isolement par la distance (IBD)** est un modèle simple considérant une dispersion localisé dans l'espace, et un habitat continu

# Dispersion et isolement par la distance

Isolement par la distance (IBD; Wright, 1943) :

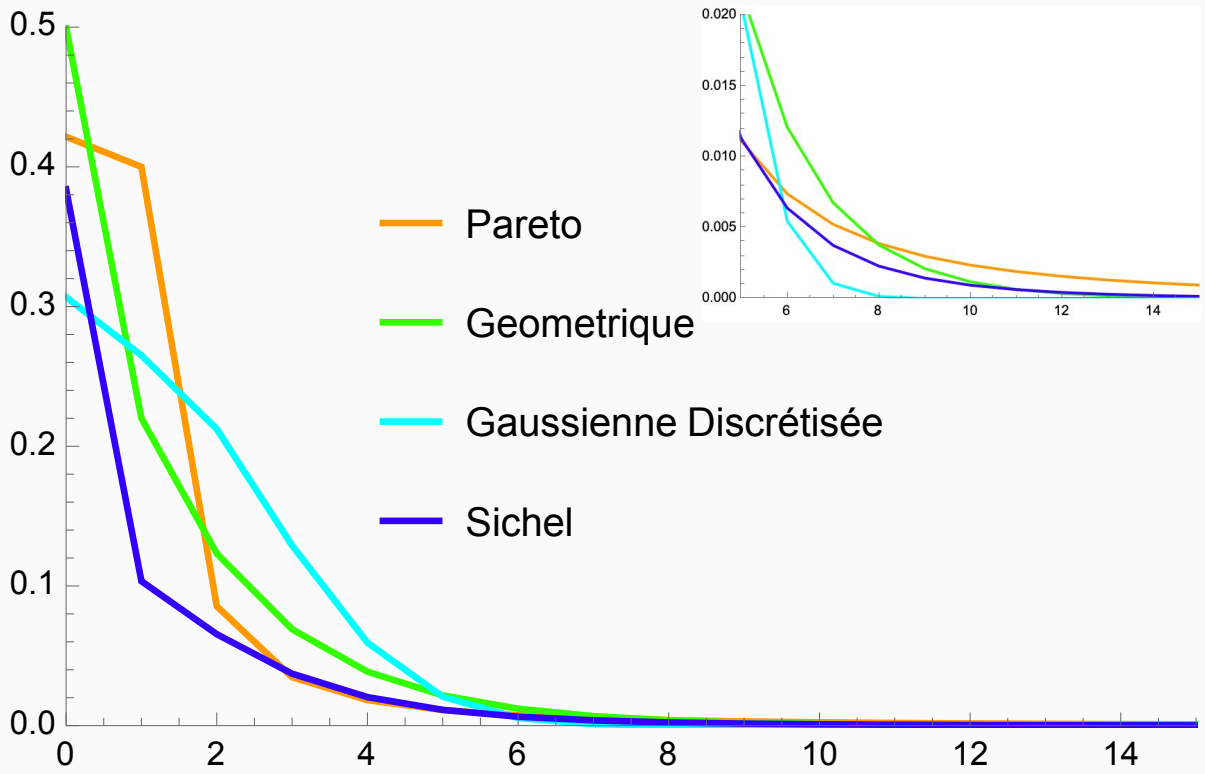
- Principe : dispersion spatialement limitée

# Dispersion et isolement par la distance

Isolement par la distance (IBD; Wright, 1943) :

- Principe : dispersion spatialement limitée

$\rightarrow$ **distribution de dispersion parent-descendant**

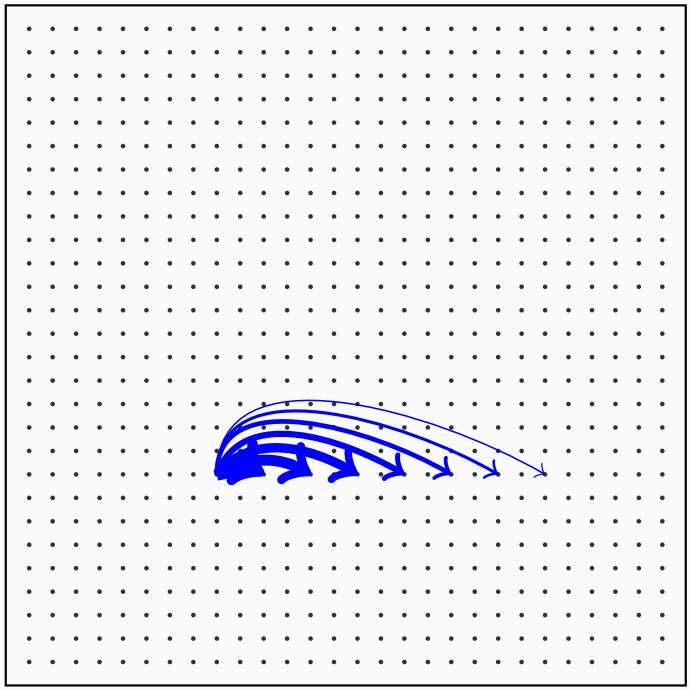


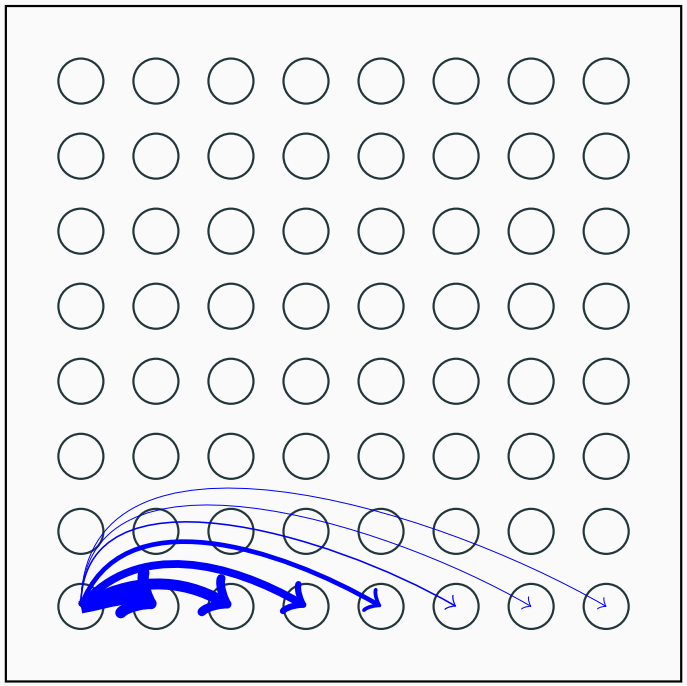

Exemples de distributions de dispersion

# Dispersion et isolement par la distance

Isolement par la distance (IBD; Wright, 1943) **sur un réseau**:

- Dèmes panmictiques, des individus ou des couples répartis sur une grille homogène, avec une distribution de dispersion parent-descendant

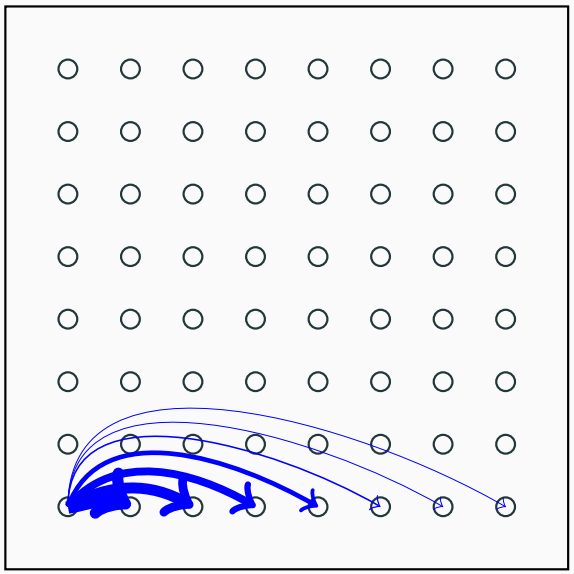




2 modèles selon la distribution spatiale des individus et des habitats

# Dispersion et isolement par la distance

IBD en dèmes ou en “habitat continu” sur un réseau



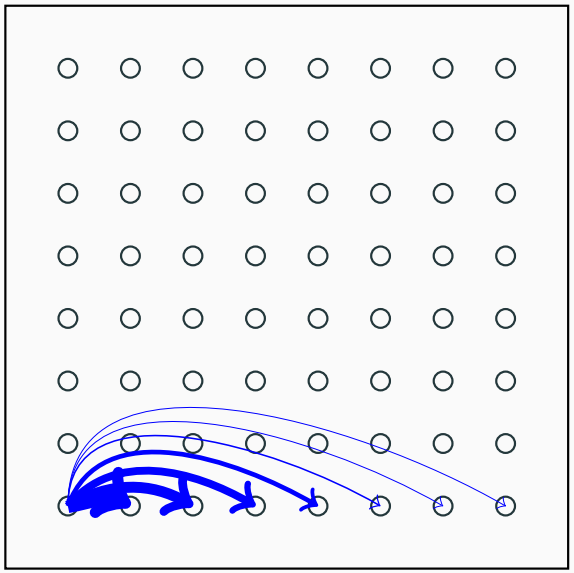
- Taille du lattice/habitat: $n_x$ ($n_y$)
- **Taille des dème**: $N$ (=1,2 en “habitat continu”)
- **Taux de dispersion**: $m$
- **Distrib. de dispersion**: toutes (e.g. géométrique)
- Forme de la distribution : 1 à 3 paramètres (ex. $g_{geom}$)
- **Densité** (ou unité du réseau) : $D$

- Taux de mutation = $\mu$

# Dispersion et isolement par la distance

IBD en dèmes ou en “habitat continu” sur un réseau

Paramètres “composites”



- $\sigma^2$ = carré moyen de la dispersion parent-descendant
- $D\sigma^2$ (parfois $4\pi D\sigma^2$)
-
-
-

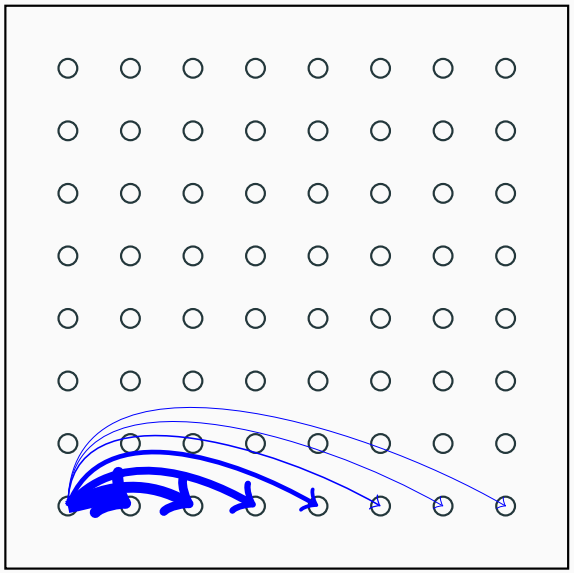
$(4\pi)D\sigma^2$ caractérise (l’inverse de) la force de l’isolement par la distance

fort $D\sigma^2 \rightarrow$ faible IBD, faible $D\sigma^2 \rightarrow$ fort IBD

# Dispersion et isolement par la distance

IBD en dèmes ou en “habitat continu” sur un réseau

Paramètres “composites”



- $\sigma^2$ = carré moyen de la dispersion parent-descendant
- $\mathbf{D}\sigma^2$ (parfois $4\pi D\sigma^2$)
- $\theta_{d(eme)} = 4N\mu$
- $\theta_{g(lobal)} = 4n_x n_y N\mu$
- $2Nm$ nombre d’émigrants par génération

$(4\pi)D\sigma^2$ caractérise (l’inverse de) la force de l’isolement par la distance

fort $D\sigma^2 \rightarrow$ faible IBD, faible $D\sigma^2 \rightarrow$ fort IBD

# Méthodes d'inférence

- Théorie de la coalescence
- Génétique spatiale des populations et dispersion
- **Méthodes d'inférence statistique**

# Méthodes d'inférence

- Méthode des moment

  Exprimer une quantité $Q$ en fonction du paramètre $P$ d'intérêt du modèle
  $\rightarrow Q = f(P)$

  Calculer $Q$ sur les données $D$
  $\rightarrow \hat{Q} = g(D)$

  Utiliser ces deux relations pour estimer le paramètre d'intérêt
  $\rightarrow \hat{P} = f^{-1}(\hat{Q})$

- données fortement réduites (à $\hat{Q}$)
- peu de paramètres estimables

# Méthodes d'inférence

- Méthode des moment


$n = 10$ tirage de 4 Faces (t=1) et 6 Piles (t=0), $p$ ?

$$M = \frac{\sum_1^n t}{n}$$

$$E[M] = p$$

$$\hat{p} = \frac{\sum_1^n t}{n} \rightarrow \hat{p} = 0.4 \text{ [ boostrap CI ]}$$

# Méthodes d'inférence

- Méthode des moment

  - forte réduction des données, peu de paramètres estimables

  Exemple classique en génétique des populations :

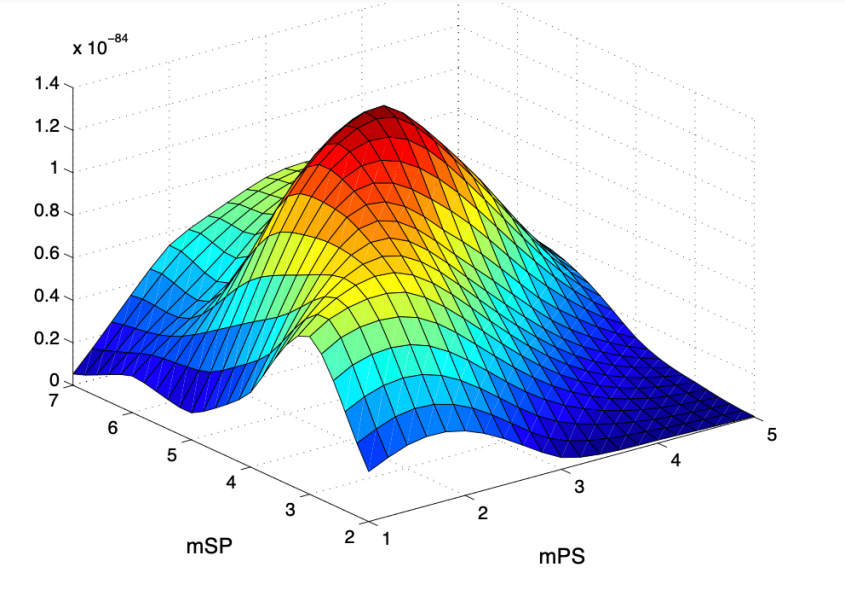
$$F_{ST} \approx \frac{1}{1+2Nm} \rightarrow \widehat{Nm} = \frac{1/\widehat{F_{ST}}-1}{2}$$

# Méthodes d'inférence

- Méthode des moment
- Maximum de Vraisemblance

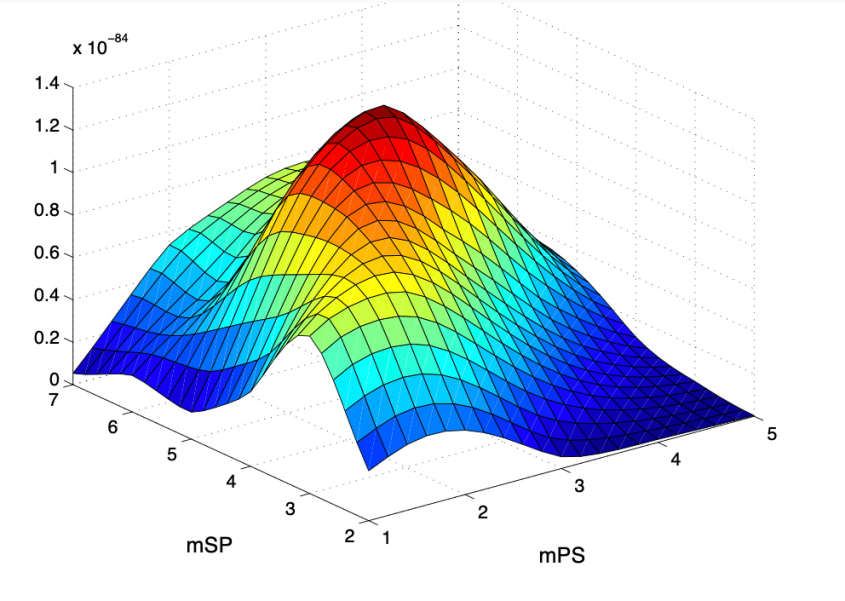
# Méthodes d'inférence

- Méthode des moment
- Maximum de Vraisemblance
  Calculer (ou estimer) la vraisemblance $\mathcal{L}(\mathcal{P}; D)$ des données $D$ en tout point de l'espace des paramètres $\mathcal{P}$



# Méthodes d'inférence

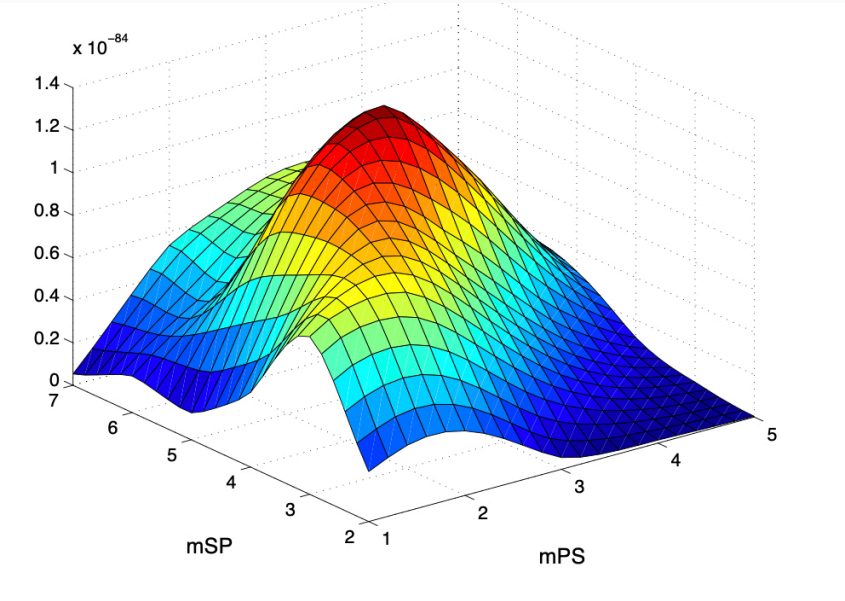
- Méthode des moment
- Maximum de Vraisemblance
  Calculer (ou estimer) la vraisemblance $\mathcal{L}(\mathcal{P}; D)$ des données $D$ en tout point de l'espace des paramètres $\mathcal{P}$




L'estimateur par maximum de vraisemblance (MLE) est le point de paramètre pour lequel l'échantillon observé est le plus probable.

# Méthodes d'inférence

- Méthode des moment
- Maximum de Vraisemblance
  Calculer (ou estimer) la vraisemblance $\mathcal{L}(\mathcal{P}; D)$ des données $D$ en tout point de l'espace des paramètres $\mathcal{P}$




Meilleure approche statistique pour l'inférence :

- Utilise toute l'information des données
- Permet d'estimer "tous" les paramètres d'un modèle

# Méthodes d'inférence

- Méthode des moment
- Maximum de Vraisemblance
- Inférence par simulation

# Méthodes d'inférence

- Méthode des moment
- Maximum de Vraisemblance
- Inférence par simulation

Lorsque la vraisemblance n'est pas calculable

Comparaison entre données réelles et données simulées

Réduction des données en un ensemble de statistiques résumantes

A partir d'un grand nombre de simulation (table d'apprentissage), on retient les paramètres qui génèrent les statistiques résumantes les plus proches des observées (distance, Random-Forest, réseau de neurones,...)

# Méthodes d'inférence

- **Méthode des moment**

  simple mais limitée à certains paramètres

  - réduction forte des données,

- **Maximum de Vraisemblance**

  complexe et pas toujours calculable

  + meilleure approche d'inférence

- **Inférence par simulation**

  alternative très flexible à la vraisemblance

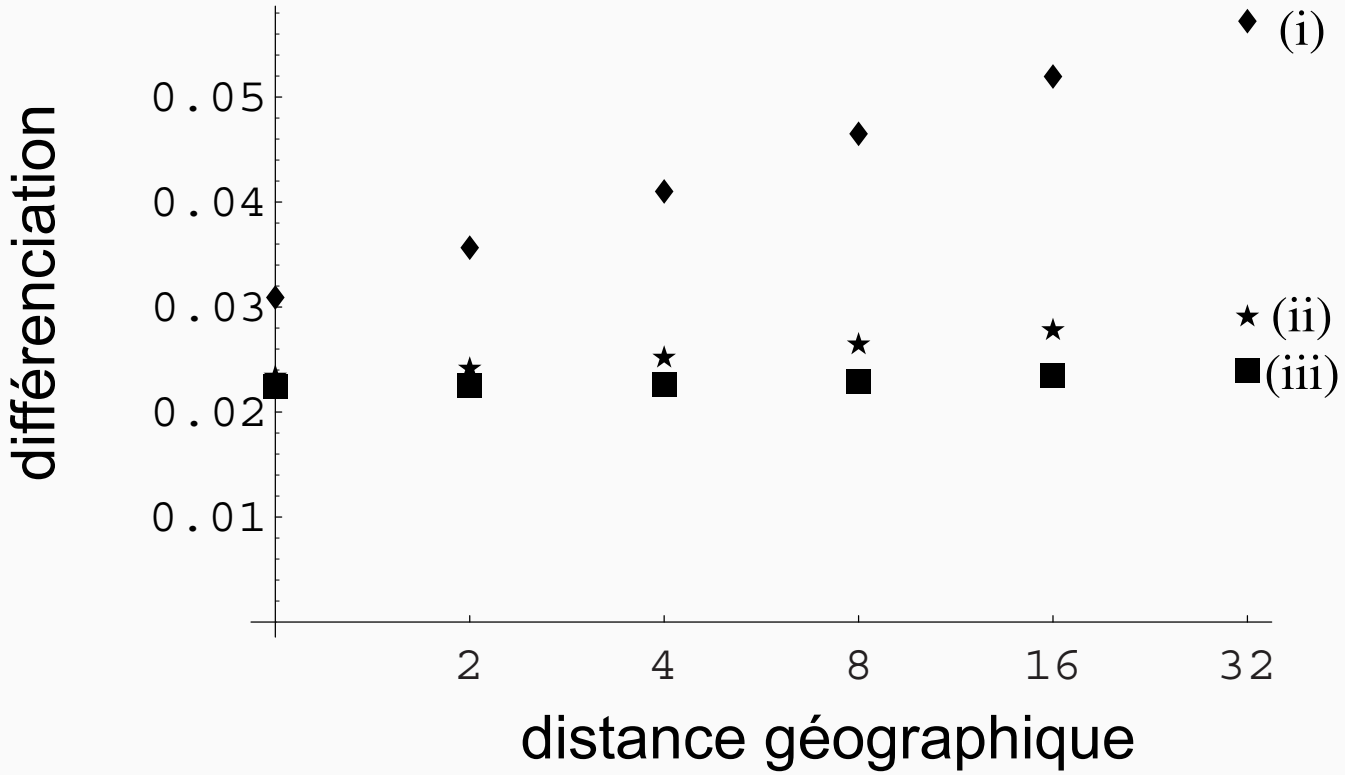
  - réduction des données
  - cadre statistique pas toujours optimal

# La régression : simple, efficace mais limitée

# La régression : simple

IBD = dispersion limitée dans l'espace

$\rightarrow$ différenciation entre population (ou entre individus) augmente avec la distance géographique



# La régression : simple

IBD = dispersion limitée dans l'espace

$\rightarrow$ différenciation entre population (ou entre individus) augmente avec la distance géographique

Fondement de la méthode de la régression (Rousset, 1997; Rousset, 2000)

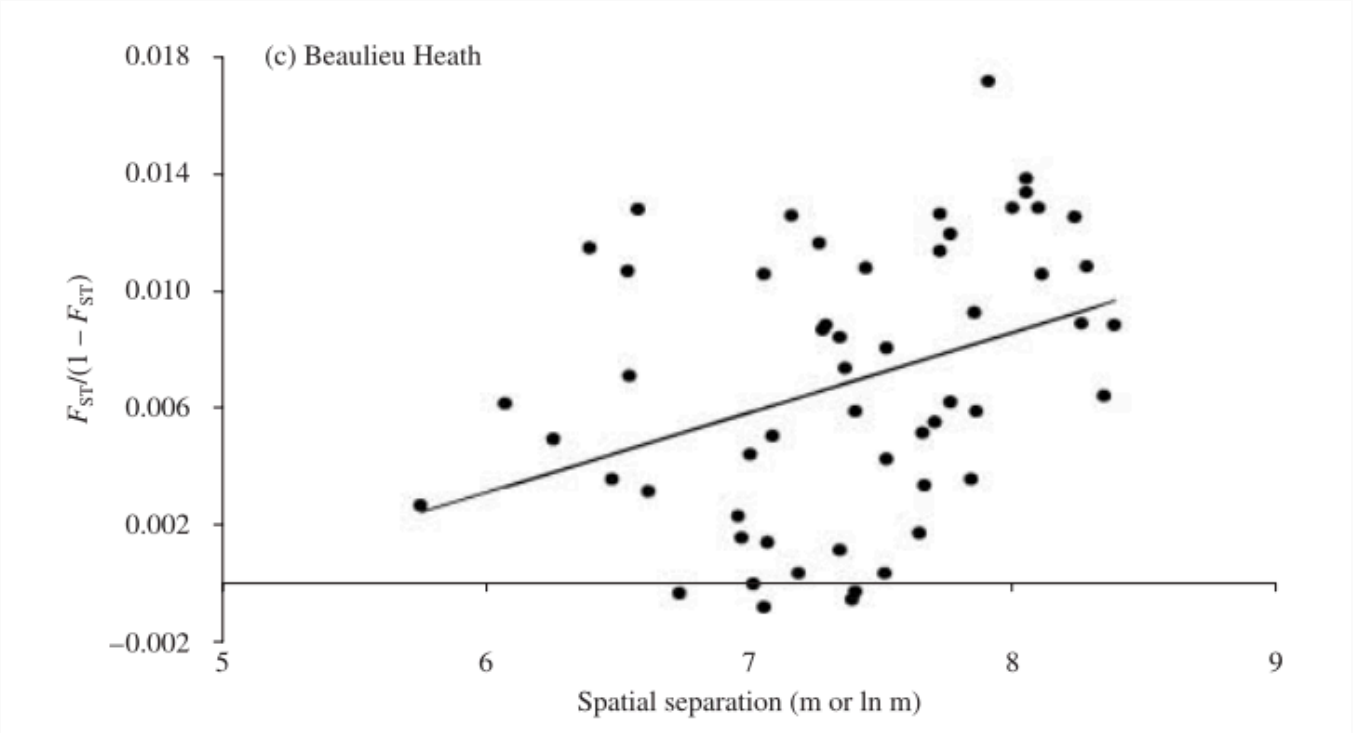
$\rightarrow$ Relation linéaire entre un paramètre de différenciation et la distance géographique entre échantillons

$$\frac{F_{ST}}{1 - F_{ST}} \text{ or } a_r \approx \frac{ln(\text{distance géo})}{4\pi D\sigma^2} + constante \tag{1}$$

# La régression : simple

méthode de la régression (Rousset, 1997; Rousset, 2000)

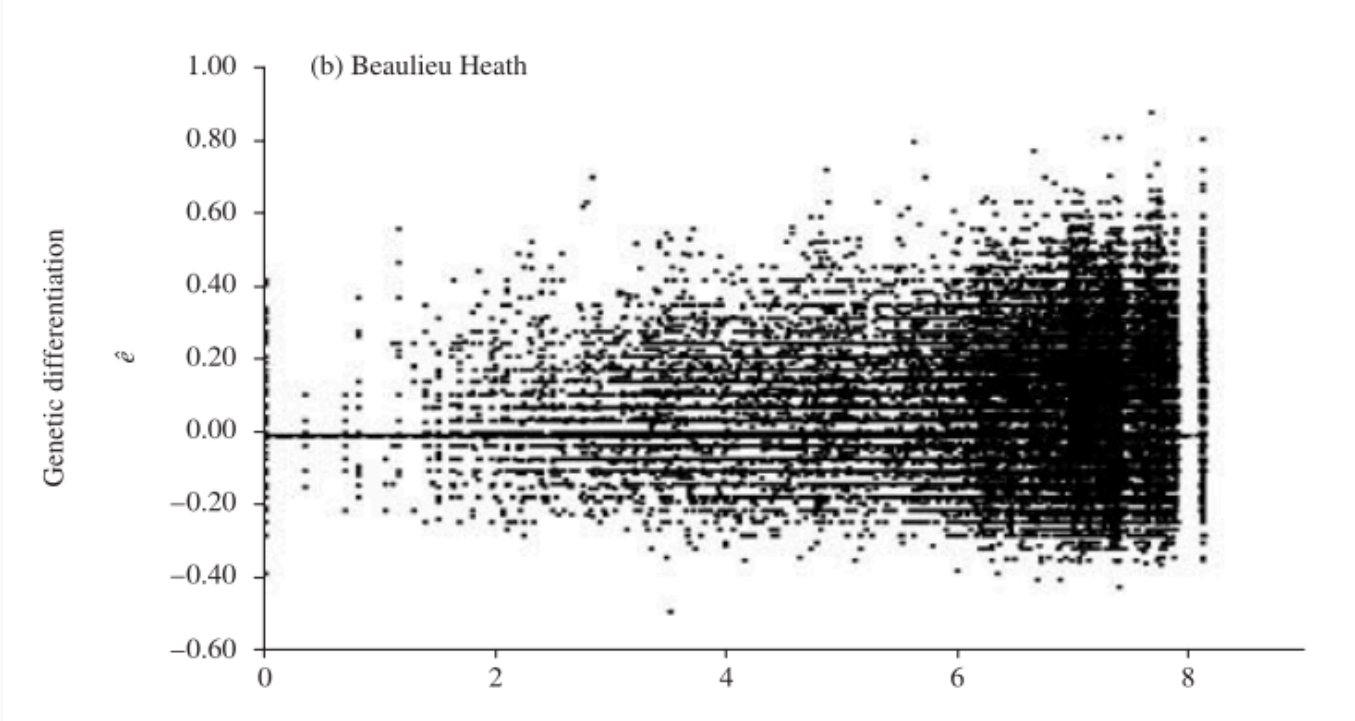
En pratique, 1/pente est un estimateur du produit $D\sigma^2$



# La régression : simple

méthode de la régression (Rousset, 1997; Rousset, 2000)

En pratique, 1/pente est un estimateur du produit $D\sigma^2$



# La régression : simple, efficace

Méthode de la régression de Rousset $\rightarrow \widehat{D\sigma^2}$

- Largement testée par simulation (Leblois, Estoup, and Rousset, 2003; Leblois, Rousset, and Estoup, 2004)

$\rightarrow$ suggère qu'elle permet une **estimation robuste du produit $D\sigma^2$ local et actuel**

# La régression : simple, efficace

Méthode de la régression de Rousset $\rightarrow \widehat{D\sigma^2}$

- Largement testée par simulation (Leblois, Estoup, and Rousset, 2003; Leblois, Rousset, and Estoup, 2004)

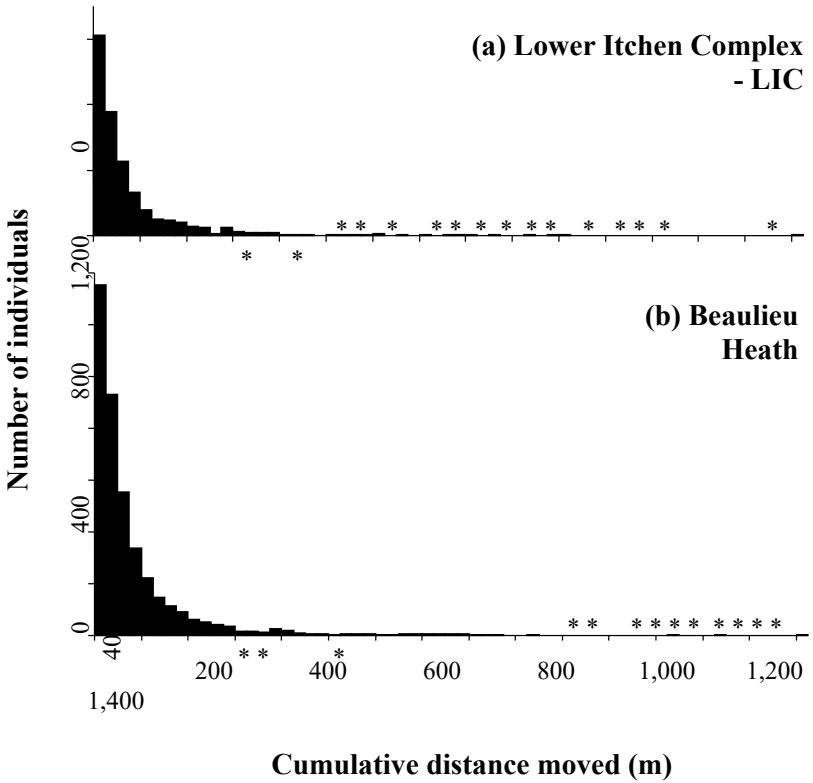
$\rightarrow$ suggère qu'elle permet une **estimation robuste du produit $D\sigma^2$ local et actuel**

- (le plus important) elle a permis de faire des **comparaisons entre estimations génétiques et démographiques** de $D\sigma^2$

# Estimations génétiques vs. démographiques

## Comparisons between genetic and demographic estimates

- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)




**Demographic data (CMR)**

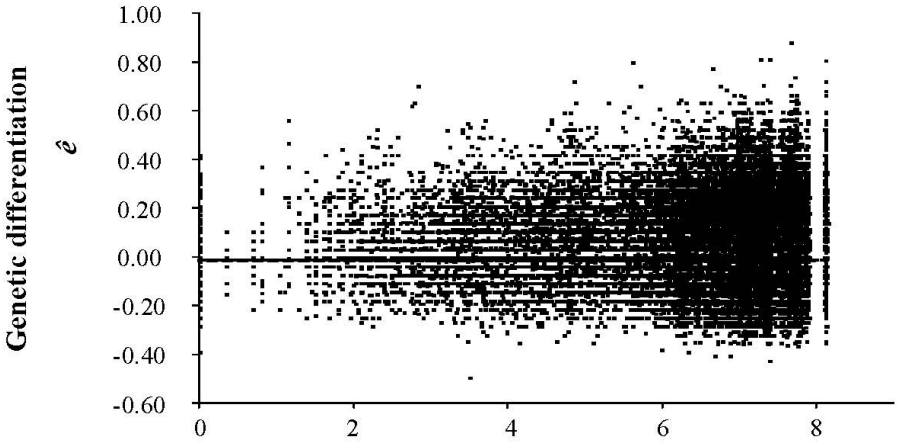
➠ Census density and distribution of dispersal

## Comparisons between genetic and demographic estimates

- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)

**Genetic data : 700 individuals genotyped at 13 microsatellite loci**

➠ indirect estimates of $D\sigma^2$



**Comparisons between genetic and demographic estimates**

- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)

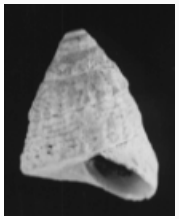
| | $D\sigma^2$ estimates | |
|---|---|---|
| | Direct (demographic) | Indirect (genetic) |
| Site 1 | 277 | 222 |
| Site 2 | 249 | 259 |
| Site 3 | 555 | 753 |

**very good agreement between demographic and genetic estimates**

1

# Estimations génétiques vs. démographiques

**Comparisons between genetic and demographic estimates**



| | Direct (Demography) | Indirect (genetic) |
|---|---|---|
| American Marten | 7.5 | 3.8 |
| Kangaroo rats | 1.43 | 2.58 |
| intertidal snails | 2.4 | 3.6 |
| Forest lizards | 11.5 | 5.5 |
| Humans in the rainforest | 29.3 | 21.1 |
| Legumin | 9.6 | 13.9 |

very good agreement between demographic and genetic estimates for all available data sets with demographic and genetic data at a local geographical scale

➠ **validate the regression method and isolation by distance models**

**IBD seems to be relevant models for the inference of demographic parameters at small geographic and temporal scale**

# Conclusion régression : simple, efficace

Méthode de la régression de Rousset $\rightarrow$ **(robuste)** $\widehat{D\sigma^2}$

Interprétation de la robustesse des inférence grâce à la théorie de la coalescence:

- petites tailles de populations ($N \rightarrow$ un individu ou un couple)
- forts taux de dispersion

$\rightarrow$ majorité de l'arbre de coalescence est dans un passé récent

$\rightarrow$ faible influence des processus démographiques passés, de l'hétérogénéité de l'habitat, et des processus mutationnels et adaptatifs

# Conclusion régression : simple, efficace mais limitée

Méthode de la régression de Rousset $\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$

- $D\sigma^2$ peu “parlant” pour les biologistes
- utilisation sub-optimale des données

  résumées dans la pente uniquement
- IBD homogène dans l’espace et constant dans le temps

# Conclusion régression : simple, efficace mais limitée

Méthode de la régression de Rousset $\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$

- $D\sigma^2$ peu "parlant" pour les biologistes
- utilisation sub-optimale des données

  résumées dans la pente uniquement
- IBD homogène dans l'espace et constant dans le temps

D'un point vue pratique ("empirique") :

- énormément appliquée

$\rightarrow$ valide qu'une grande majorité des espèces sont en IBD

$\rightarrow$ l'IBD est souvent (beaucoup) plus fort qu'attendu (a priori biologique)

# Conclusion régression : simple, efficace mais limitée

Méthode de la régression de Rousset $\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$

Pour aller plus loin:

- estimer $D$ et $\sigma^2$ séparément
- estimer d'autres paramètres du modèle IBD

# Conclusion régression : simple, efficace mais limitée

Méthode de la régression de Rousset $\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$

Pour aller plus loin:

- estimer $D$ et $\sigma^2$ séparément
- estimer d'autres paramètres du modèle IBD

Théoriquement, l'inférence basée sur la vraisemblance pourrait le permettre...

# Vraisemblance basée sur la coalescence

# Maximum de vraisemblance : complexe...

1995-2000: deux approches pour estimer la vraisemblance d'un échantillon génétique par coalescence (Felsenstein et al., 1999; Griffiths and Tavaré, 1994)

$$\mathcal{L}(\mathcal{P}; D) = \int_G \Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P}),$$

$$\hat{\mathcal{L}}(\mathcal{P}; D) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\Pr(D|G_i; \mathcal{P}) \Pr(G_i; \mathcal{P})}{f(G_i)}.$$

# Maximum de vraisemblance : complexe...

1995-2000: deux approches pour estimer la vraisemblance d'un échantillon génétique par coalescence (Felsenstein et al., 1999; Griffiths and Tavaré, 1994)

- toutes basées sur les approximations du $n-$coalescent

# Maximum de vraisemblance : complexe...

1995-2000: deux approches pour estimer la vraisemblance d'un échantillon génétique par coalescence (Felsenstein et al., 1999; Griffiths and Tavaré, 1994)

→ adaptation des algorithmes d'échantillonnage préférentiel (IS, Stephens and Donnelly, 2000; de Iorio and Griffiths, 2004)

implémentation de différents **modèles**, dont **IBD 1D and 2D**, dans le logiciel *MIGRAINE* (Rousset and Leblois, 2007; Rousset and Leblois, 2012; Rousset, Beeravolu, and Leblois, 2018)

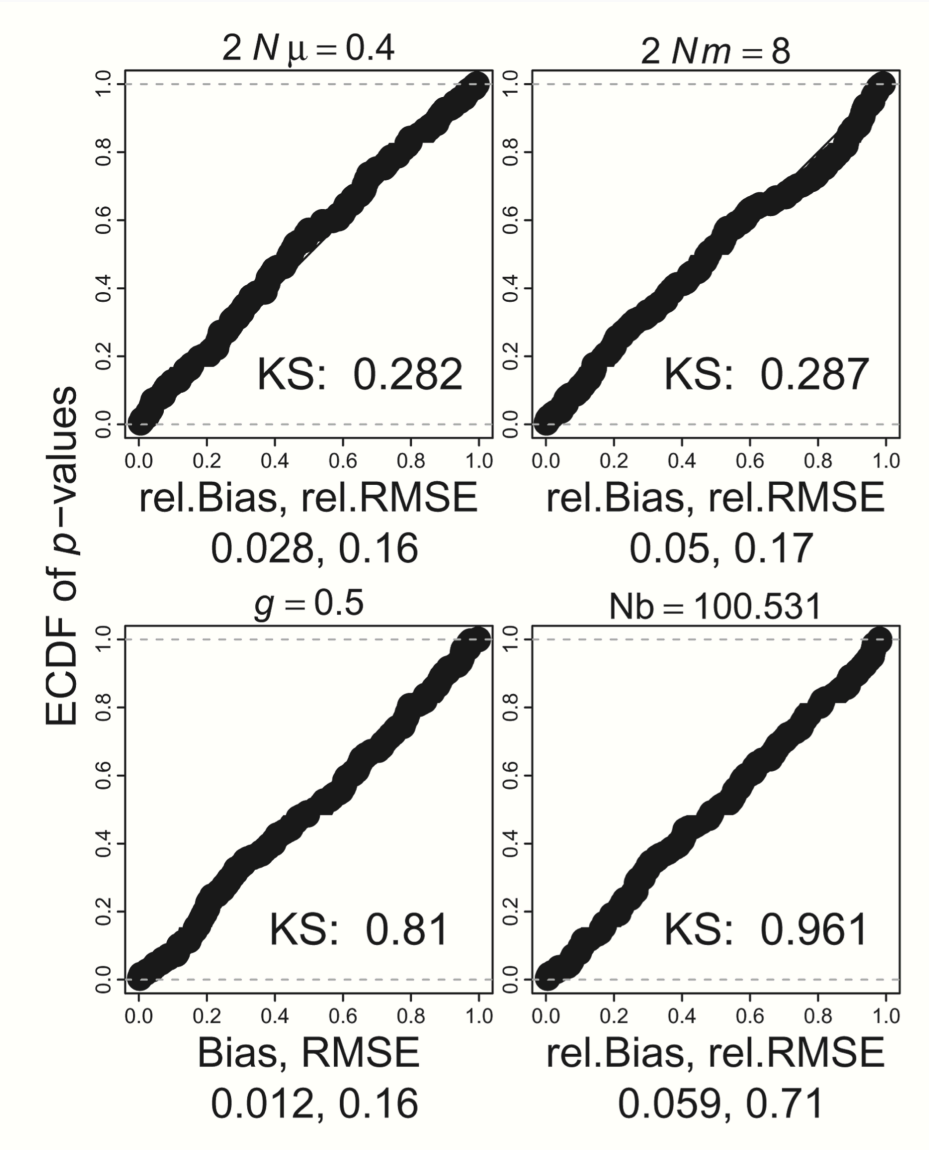
# Maximum de vraisemblance : complexe, puissante...

Des résultats encourageants:

- permet d'estimer d'autres paramètres : $Nm$, $g$, $N\mu$, $D\sigma^2$

# Maximum de vraisemblance : complexe, puissante...

Des résultats encourageants:



# Maximum de vraisemblance : complexe, puissante...

Des résultats encourageants:

- permet d'estimer d'autres paramètres : $Nm$, $g$, $N\mu$, $D\sigma^2$

mais malheureusement peu robustes

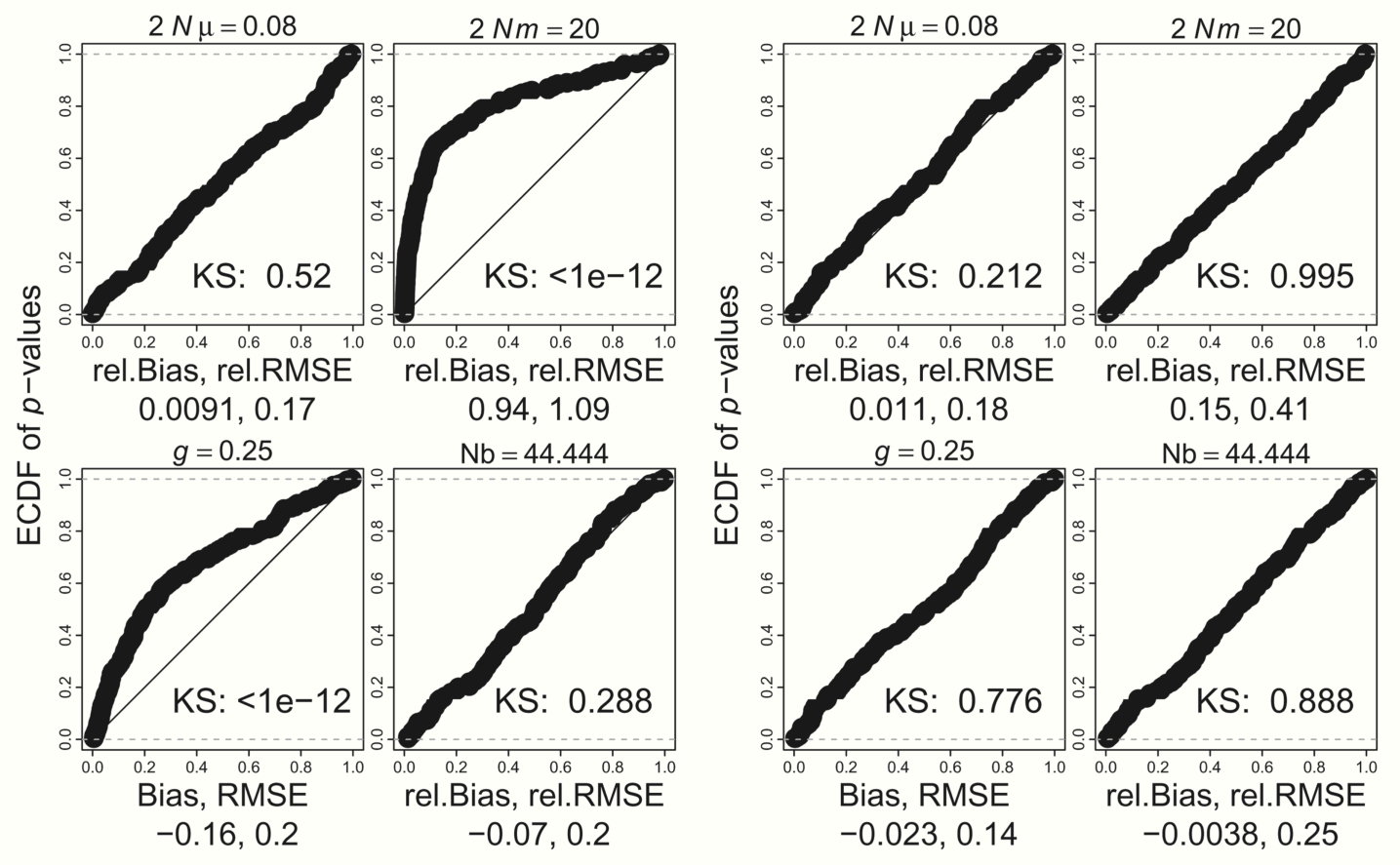
# Maximum de vraisemblance : complexe, puissante mais limitée

Forte influence des hypothèses du $n-$coalescent:

- fort taux de migration $\rightarrow$ bias sur $Nm$ et $g$

# Maximum de vraisemblance : complexe, puissante mais limitée

$N = 40\ \mu = 10^{-3}\ m = 10^{-3}$ $N = 40,000\ \mu = 10^{-6}\ m = 10^{-6}$

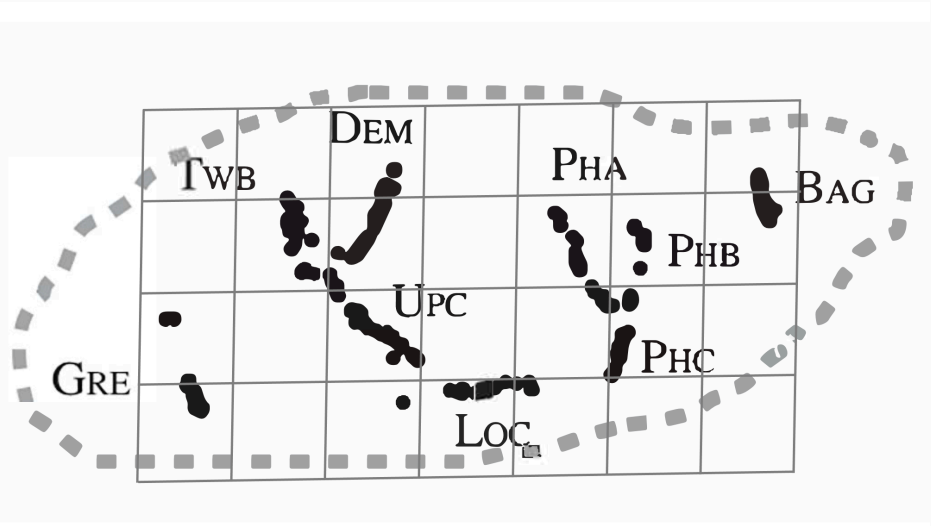



$\widehat{Nm}$ et $\widehat{g}$ biaisés mais $\widehat{D\sigma^2}$ est robuste

# Maximum de vraisemblance : complexe, puissante mais limitée

Long temps de calcul & $n-$coalescent:

- pas d'habitat continu ($N \not\to 1$ ou 2)
- difficulté grille de dèmes vs. densité irrégulière



# Maximum de vraisemblance : complexe, puissante mais limitée

Forte influence des hypothèses du $n-$coalescent:

- fort taux de migration $\rightarrow$ bias sur $Nm$ et $g$

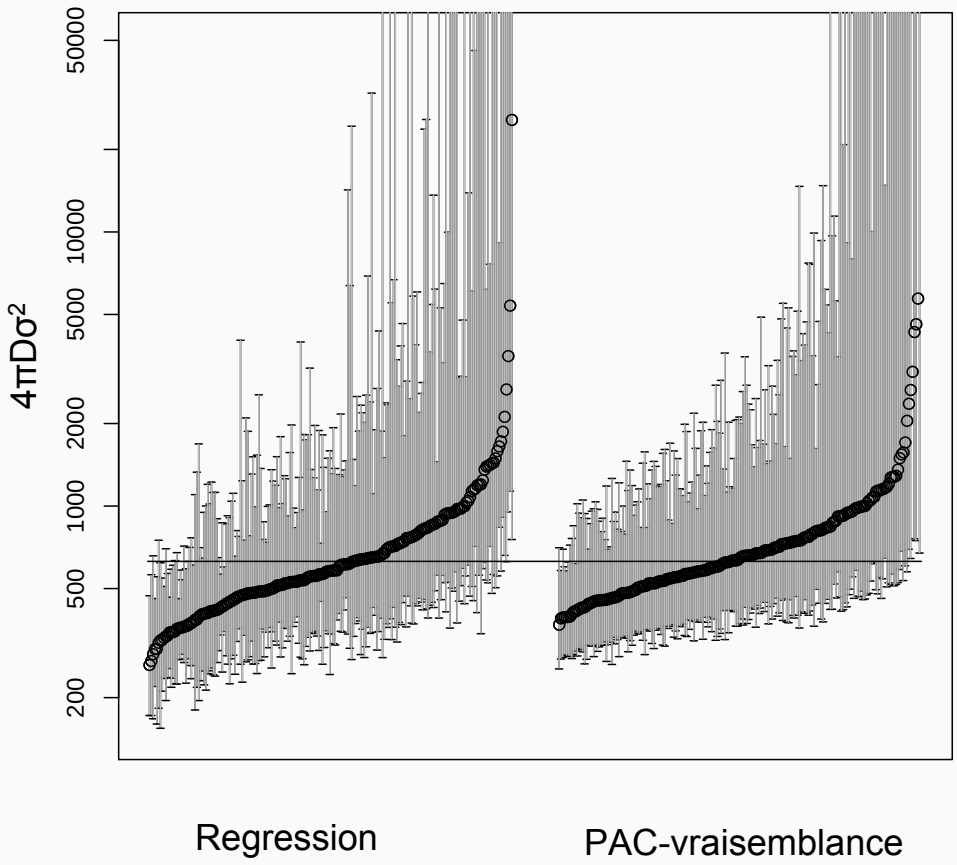
Long temps de calcul & $n-$coalescent:

- pas d'habitat continu ($N \not\rightarrow 1$ ou 2)
- difficulté grille de dèmes vs. densité irrégulière

$\rightarrow$ seule l'estimation de $D\sigma^2$ est robuste, "interprétable",

et plus précise qu'avec la régression (ouf !)

# Maximum de vraisemblance : complexe, puissante mais limitée




$$\frac{\text{RMSE}_{\text{vraisemblance}}}{\text{RMSE}_{\text{régression}}}$$ entre 0.25 et 0.66

# Conclusion vraisemblance : complexe, puissante mais limitée

Maximum de vraisemblance basé sur la coalescence

$\rightarrow \widehat{Nm}, \widehat{g}, \widehat{N\mu}, \widehat{D\sigma^2}$ **(en conditions favorables)**

$\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$ **(en général)**

+ utilisation optimale des données complètes

- $n-$coalescent inadéquate, long temps de calculs
- $D\sigma^2$ toujours peu “parlant” pour les biologistes
- difficile d’ajouter des hétérogénéités spatiales et/ou temporelles

# Conclusion vraisemblance : complexe, puissante mais limitée

Maximum de vraisemblance basé sur la coalescence

$\rightarrow \widehat{Nm}, \widehat{g}, \widehat{N\mu}, \widehat{D\sigma^2}$ **(en conditions favorables)**

$\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$ **(en général)**

- $n-$coalescent inadéquate, long temps de calculs
- $D\sigma^2$ toujours peu "parlant" pour les biologistes
- difficile d'ajouter des hétérogénéités spatiales et/ou temporelles

Pour aller plus loin:

- estimer $D$ et $\sigma^2$ séparément
- estimation **(robuste?)** d'autres paramètres du modèle IBD
- considérer des habitats hétérogènes dans l'espace

# Conclusion vraisemblance : complexe, puissante mais limitée

Maximum de vraisemblance basé sur la coalescence

$\rightarrow \widehat{Nm}, \widehat{g}, \widehat{N\mu}, \widehat{D\sigma^2}$ **(en conditions favorables)**

$\rightarrow$ **(uniquement)** $\widehat{D\sigma^2}$ **(en général)**

Pour aller plus loin:

- estimer $D$ et $\sigma^2$ séparément
- estimation **(robuste?)** d'autres paramètres du modèle IBD
- considérer des habitats hétérogènes dans l'espace

Théoriquement, l'inférence par simulation pourrait le permettre...

# Inférence par simulation: puissante, flexible mais encore lourde...

# Inférence par simulation: flexible

Initiée en génétique des populations (2000's)

Notamment au CBGP (*DIY-ABC*, *ABC-RF*, Arnaud Estoup)

++ n'importe quel modèle (avec simulations assez efficaces...)

– trouver les bonnes statistiques (réduire à minima les données)

# Inférence par simulation: flexible

Inférence par simulation sous IBD

- Simulations en IBD: *GSpace* (et *IBDSim*)
- calcul des statistiques résumantes: *GsumStat*
- Inférence statistique : *Summary Likelihood* (paquet R *Infusion*)

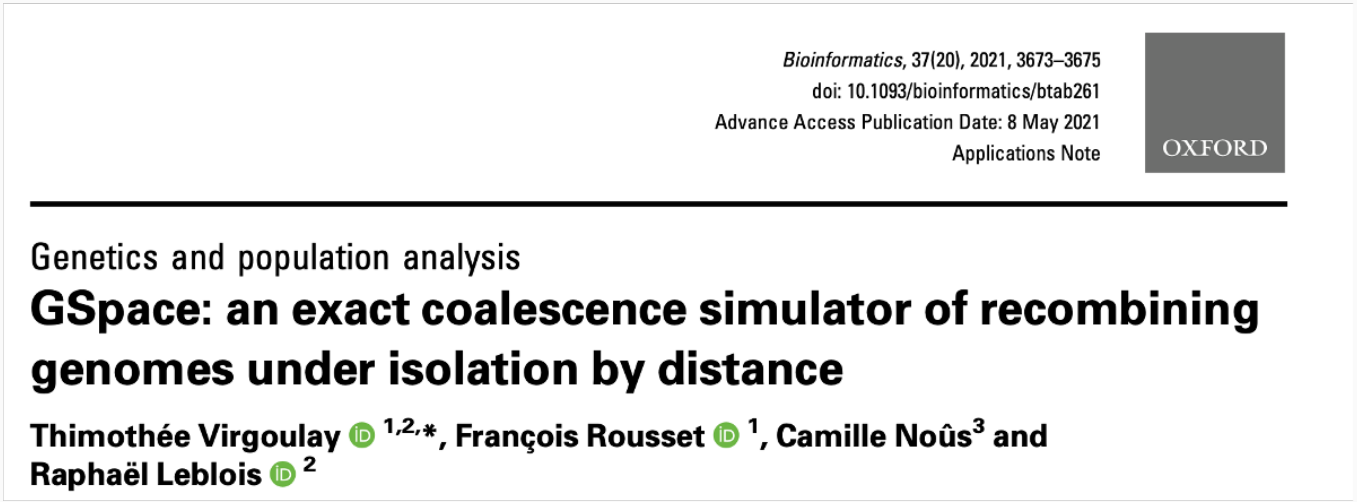
Encapsulé dans un paquet R *gspace2infr* pour faciliter

→ les inférences pour les non-spécialistes

→ les tests par simulation de précision, robustesse, et couverture des IC

# Inférence par simulation: flexible

Simulations sous IBD avec *GSpace* (Virgoulay et al. 2021)



*Bioinformatics*, 37(20), 2021, 3673–3675
doi: 10.1093/bioinformatics/btab261
Advance Access Publication Date: 8 May 2021
Applications Note

OXFORD

Genetics and population analysis

**GSpace: an exact coalescence simulator of recombining genomes under isolation by distance**

**Thimothée Virgoulay [1,2,*], François Rousset [1], Camille Noûs[3] and Raphaël Leblois [2]**

- algorithme exact de coalescence génération par génération
- avec recombinaison ($\rightarrow$ simulation de génomes)
- populations structurées spatialement
- en dèmes ou en habitat continu sur un réseau
- avec des distributions de dispersion

# Inférence par simulation: flexible

Calcul des statistiques résumantes sur données simulées et réelles avec *GsumStat*

- non spatiales: $N_a$, $H_e = 1 - Q_0$, $F_{IS}$, $F_{ST}$, AFS (SFS),
- spatiales: $Q_r$, **pente et intercept des régressions** $\frac{F_{ST}}{1-F_{ST}}$ et $a_r$

+ recombinaison

# Inférence par simulation: flexible

Inférence avec *Infusion* (*SL*, Rousset et al., 2017)

- Alternative à l'ABC/ABC-RF (comparaison en cours)

- Principe : Estimation de la surface de "vraisemblance-résumée" à partir de la distribution jointe des paramètres et des statistiques résumantes en utilisant des mélanges de gaussiennes multivariées

+ procédure itérative pour densifier les simulations sur les zones d'intérêt

# Inférence par simulation: flexible

**premiers tests par simulation** sous IBD entre individus en habitat continu ($N = 2$)

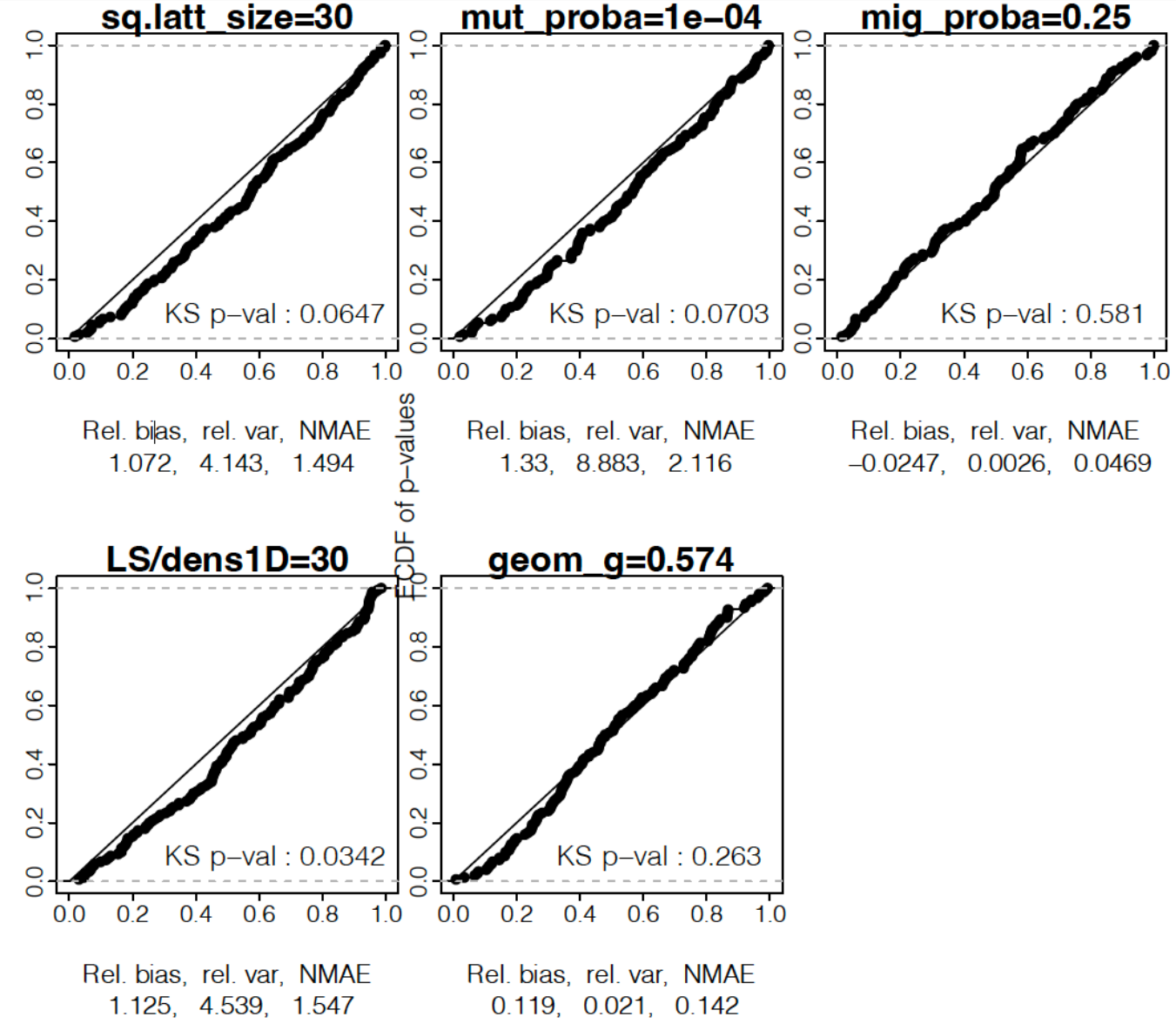
**petit jeu de données génétiques**: 100 individus échantillonnés au centre d'un habitat carré de 30x30,

génotypés a 20 marqueurs indépendants multialléliques (microsatellites)

ou 100 bi-alléliques répartis sur 10 chromosomes (SNPs)

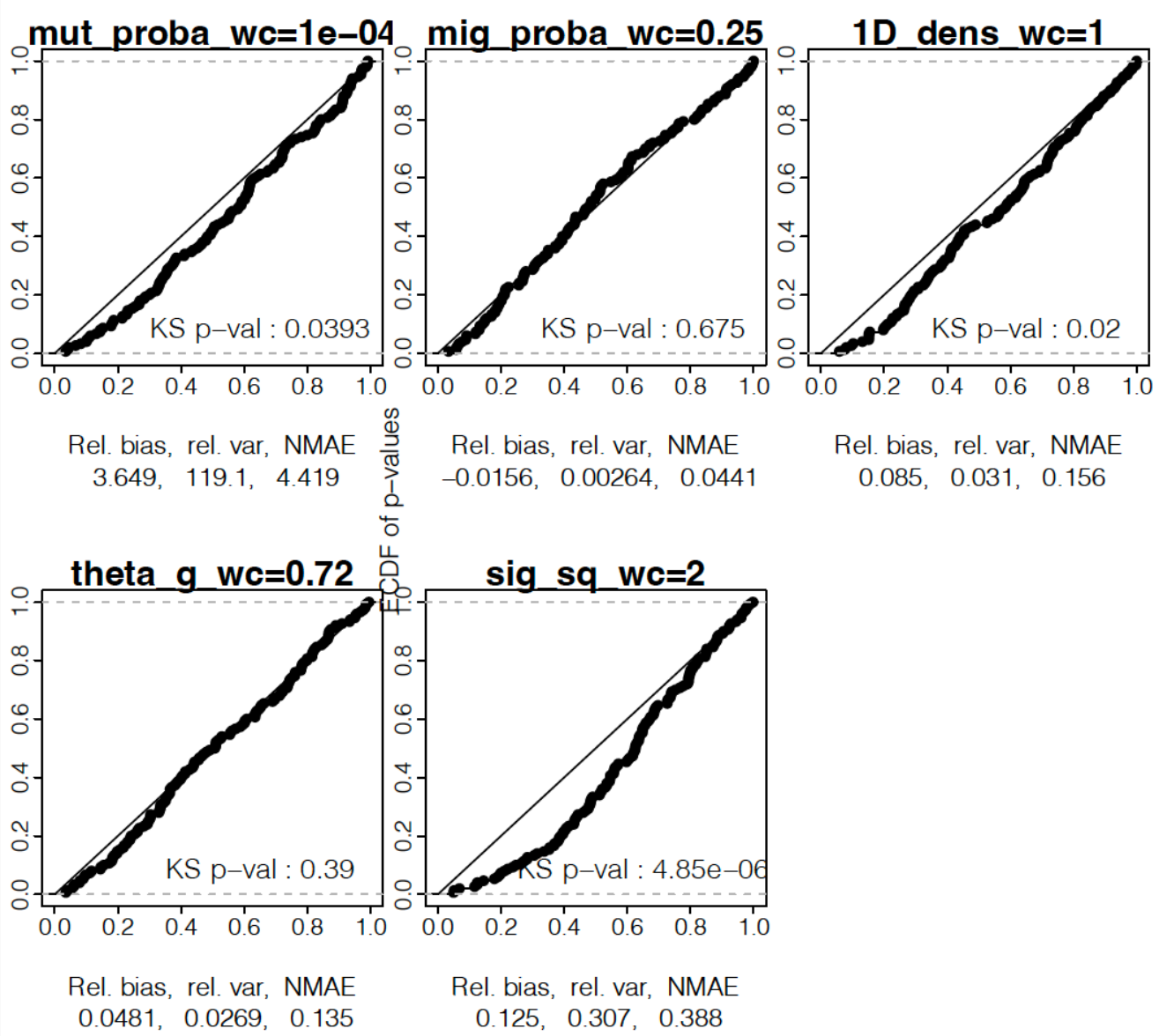
# Inférence par simulation: flexible, puissante

## Paramètres canoniques



# Inférence par simulation: flexible, puissante

## Paramètres composites



# Inférence par simulation: flexible, puissante

**premiers tests par simulation** sous IBD en habitat continu

**petit jeu de données génétique**: 10x10 individus, habitat carré de 30x30, 20 marqueurs multi-alleliques indépendants

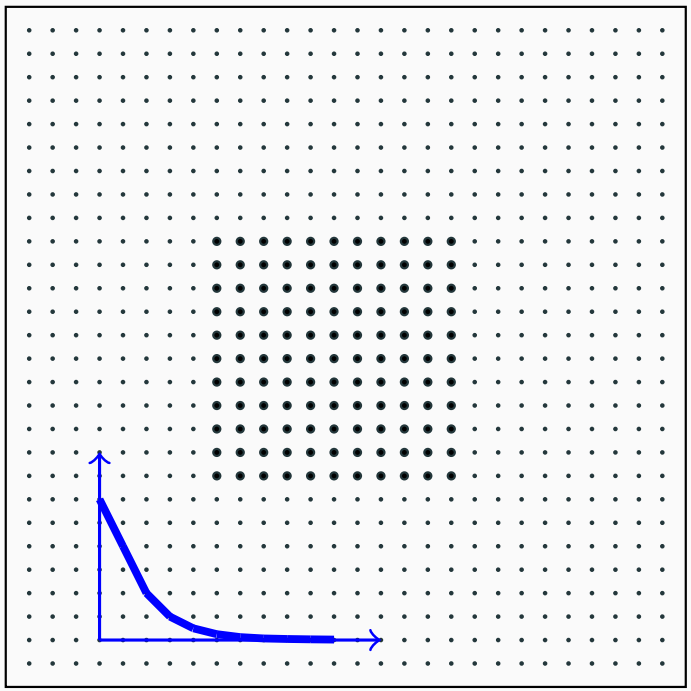
Inférences plus ou moins précises selon les paramètres:

- 5% $\leq$ bias & RMSE $\leq$ 20% : $m$, $g$, $D$, $\theta_{pop}$ et $\sigma^2$

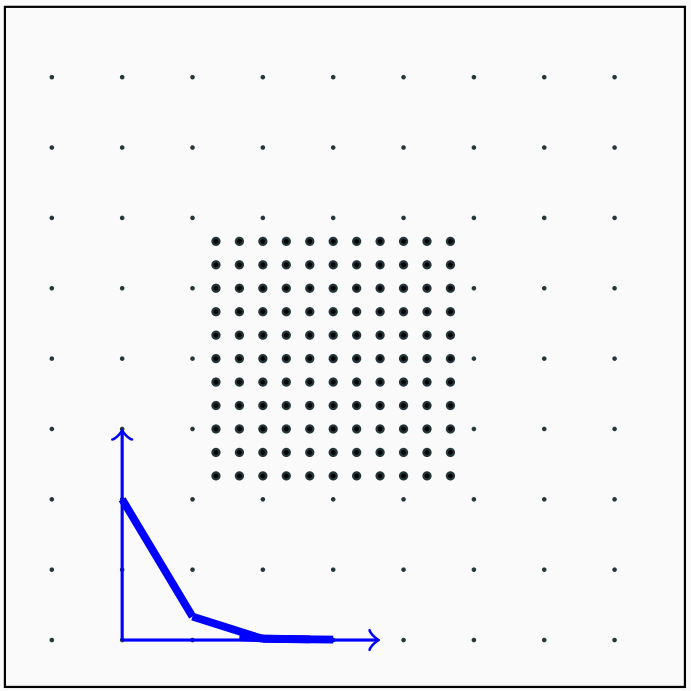
- beaucoup moins d'information sur la taille totale de la population ($N_{pop} = 2 * n_x^2$) et le taux mutation $\mu$ indépendamment (dépend de l'échelle d'échantillonnage)

# Inférence par simulation: puissante, flexible mais encore lourde...

Limite (importante ?): difficile de gérer les variations de densité à cause de la nature discrète du réseau et des distributions de dispersion



Densité = 1

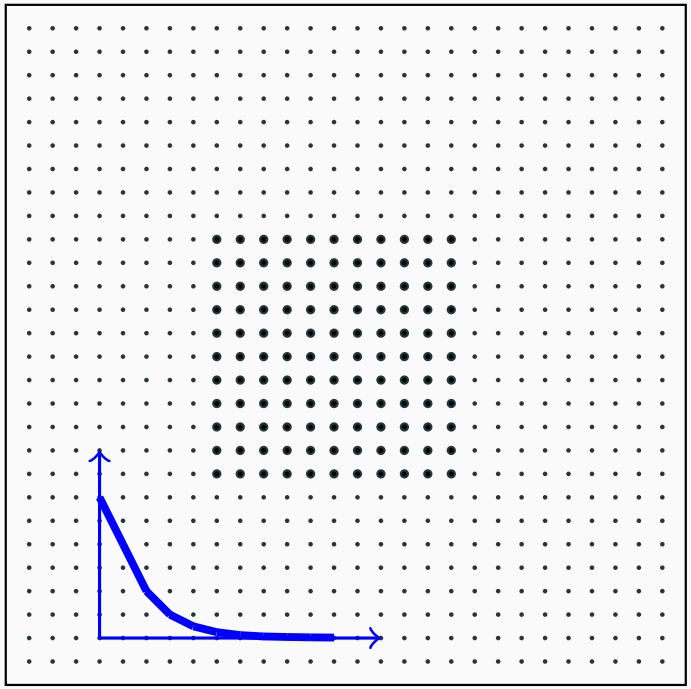


Densité = 1/3

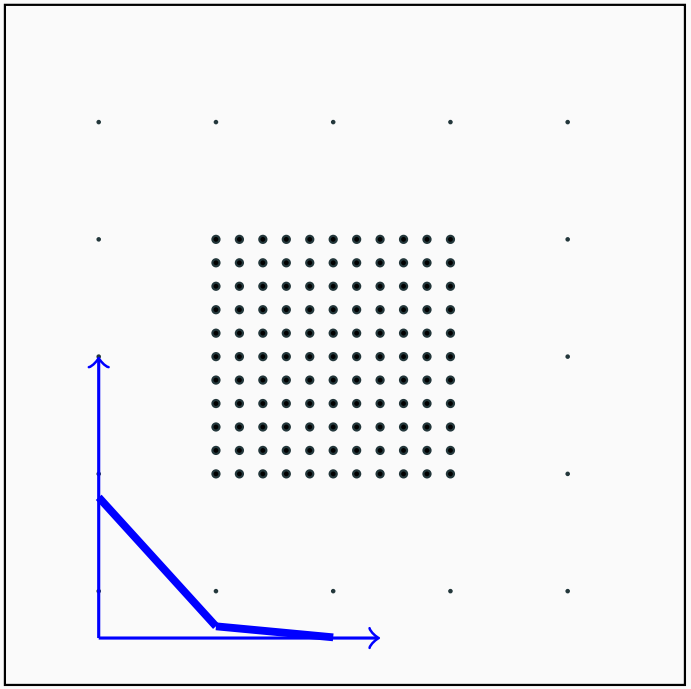
Impossible de garder $\sigma^2$ constant avec une baisse de densité

# Inférence par simulation: puissante, flexible mais encore lourde...

Limite (importante ?): difficile de gérer les variations de densité à cause de la nature discrète du réseau et des distributions de dispersion



Densité = 1



Densité = 1/5

Impossible de garder $\sigma^2$ constant avec une baisse de densité

# Inférence par simulation: puissante, flexible mais encore lourde...

Tests par simulation

+ Résultats très encourageants
- Temps de simulation très longs
- Nature discrète du réseau

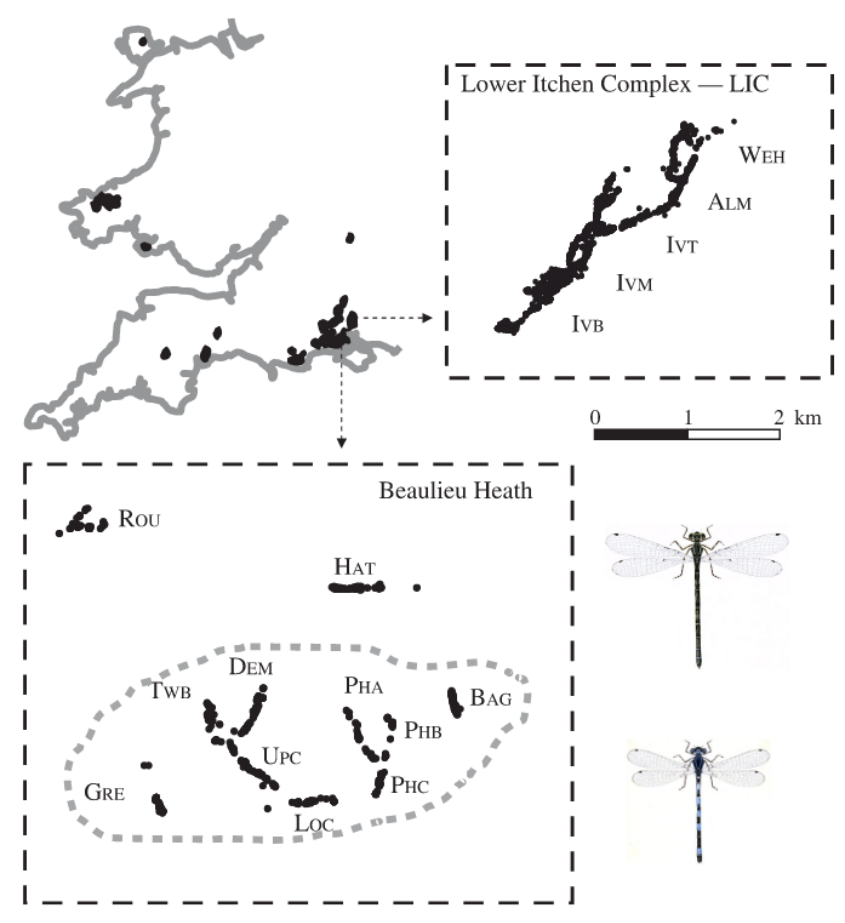
quelques difficultés à analyser des données réelles

- Modèle constant dans le passé
  $\rightarrow$ pas de polymorphisme ancestral

# Inférence par simulation: puissante, flexible mais encore lourde...

quelques difficultés à analyser des données réelles

- Densité homogène dans l'espace
  $\rightarrow$ avec quelle densité (démo) faut il comparer



# Inférence par simulation: puissante, flexible mais encore lourde...

Tests par simulation

+ Résultats très encourageants
- Temps de simulation très longs
- Nature discrète du réseau

quelques difficultés à analyser des données réelles

- Modèle constant dans le passé
  → pas de polymorphisme ancestral
- Densité homogène dans l'espace
  → avec quelle densité (démo) faut il comparer

Comparer précision et robustesse avec la régression et le maximum de vraisemblance sur données simulées et réelles

# Conclusions et perspectives

# Conclusions et perspectives

Inférence de la dispersion et de la densité en populations naturelles

- Pertinence du modèle IBD (habitat continu, $N = 1, 2$)
  - \+ assez “réaliste” pour des inférences précises et robustes
  - \- limites du modèle en réseau
- Intérêts et limites des différentes approches d’inférence
- Complémentarité des tests par simulation et sur données réelles
- Inférence de $D$, $\sigma^2$, $m$, ($N$, $\mu$) vs. $D\sigma^2$, $Nm$, ($N\mu$)
  - coalescence exacte vs. approximation du $n-$coalescent
  - contraste dérive locale et globale en IBD

# Conclusions et perspectives

Inférence de la dispersion et de la densité en populations naturelles

- Pertinence du modèle IBD (habitat continu, $N = 1, 2$)
- Intérêts et limites des différentes approches d'inférence
- Complémentarité des tests par simulation et sur données réelles
- Inférence de $D$, $\sigma^2$, $m$, ($N$, $\mu$) vs. $D\sigma^2$, $Nm$, ($N\mu$)

La **robustesse** est un facteur crucial à prendre en compte lors de l'évaluation de méthodes d'inférence en génétique des populations

# Conclusions et perspectives

Inférence de la dispersion et de la densité en populations naturelles

- Pertinence du modèle IBD (habitat continu, $N = 1, 2$)
- Intérêts et limites des différentes approches d'inférence
- Complémentarité des tests par simulation et sur données réelles
- Inférence de $D$, $\sigma^2$, $m$, ($N$, $\mu$) vs. $D\sigma^2$, $Nm$, ($N\mu$)

Objectif bientôt atteint:

La combinaison de méthodes d'**inférence par simulation** avec le modèle d'**IBD individuel en habitat continu** permet des l'**estimation de beaucoup plus de paramètres** que les approches développées précédemment et peut facilement être étendue à des **modèles plus complexes**

# Conclusions et perspectives

Inférence de la dispersion et de la densité en populations naturelles

- Pertinence du modèle IBD (habitat continu, $N = 1, 2$)
- Intérêts et limites des différentes approches d'inférence
- Complémentarité des tests par simulation et sur données réelles
- Inférence de $D$, $\sigma^2$, $m$, ($N$, $\mu$) vs. $D\sigma^2$, $Nm$, ($N\mu$)

Intérêt d'un projet de recherche sur le temps long

# (Une infinité de ) Perspectives...

Inférence de la dispersion et de la densité en populations naturelles

- Nouveaux modèles, nouvelles méthodes, nouvelles statistiques

# (Une infinité de ) Perspectives...

Inférence de la dispersion et de la densité en populations naturelles

- Population ancestrale panmictique (temps de calcul & polymorphisme ancestral)
- Tester robustesse aux déséquilibres démographiques

  ou inférer les variations démographiques passées (tailles d'habitat, densité) conjointement avec la dispersion
- Hétérogénéités spatiales de la densité

$\rightarrow$ Applications en conservation, bio-invasion, agro-écologie

# (Une infinité de ) Perspectives...

Inférence de la dispersion et de la densité en populations naturelles

- Population ancestrale panmictique (temps de calcul & polymorphisme ancestral)
- Tester robustesse aux déséquilibres démographiques

  ou inférer les variations démographiques passées (tailles d'habitat, densité) conjointement avec la dispersion
- Hétérogénéités spatiales de la densité

$\rightarrow$ Applications en conservation, bio-invasion, agro-écologie

... encore de nombreuses années de recherche collaborative avec plein d'étudiant.e.s ...

# AU DELÀ DE LA SCIENCE...

**Limiter le fort biais de genre** dans la discipline

**protéger notre planète**, **nos libertés académiques et sociales**, **lutter contre les inégalités** et promouvoir la **coopération** au lieu de la compétition

si l'on veut que nos étudiant.e.s puissent s'appuyer dans le futur sur ce qu'on leur apprend aujourd'hui

**Limiter le fort biais de genre** dans la discipline

**protéger notre planète**, **nos libertés académiques et sociales**, **lutter contre les inégalités** et promouvoir la **coopération** au lieu de la compétition

si l'on veut que nos étudiant.e.s puissent s'appuyer dans le futur sur ce qu'on leur apprend aujourd'hui

# Merci à toutes et tous

Jury, public, François, Arnaud, étudiant.e.s, CBGP, famille, amis, de Grabels...

# References



De Iorio, Maria and Robert C. Griffiths (2004). **"Importance sampling on coalescent histories".** In: *Advances in Applied Probability* 36, pp. 417–433.


Felsenstein, Joseph et al. (1999). **"Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data".** In: *Statistics in Molecular Biology and Genetics*. Ed. by Françoise Seillier-Moiseiwitsch. Vol. 33. Hayward, California: Institute of Mathematical Statistics, pp. 163–185.


Griffiths, R.C. and Simon Tavaré (1994). **"Ancestral inference in population genetics".** In: *Statistical Science* 9, pp. 307–319.

# References II


- Leblois, Raphael, Arnaud Estoup, and François Rousset (2003). **"Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance".** In: *Molecular Biology and Evolution* 20, pp. 491–502.
- Leblois, Raphaël, François Rousset, and Arnaud Estoup (2004). **"Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data".** In: *Genetics* 166, pp. 1081–1092.
- Rousset, François (1997). **"Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance".** In: *Genetics* 145, pp. 1219–1228.

# References III


- Rousset, François (2000). **“Genetic differentiation between individuals”.** In: *Journal of Evolutionary Biology* 13, pp. 58–62.
- Rousset, François, Champak Reddy Beeravolu, and Raphaël Leblois (2018). **“Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations”.** In: *Journal de la société Française de Statistique* 159.3, pp. 142–166. ISSN: 2102-6238.
- Rousset, François and Raphaël Leblois (2007). **“Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification”.** In: *Molecular Biology and Evolution* 24, pp. 2730–2745.

# References iv


- Rousset, François and Raphaël Leblois (2012). **"Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals".** In: *Molecular Biology and Evolution* 29, pp. 957–973.
- Rousset, François et al. (2017). **"The summary-likelihood method and its implementation in the Infusion package".** In: *Molecular Ecology Resources* 17.1, pp. 110–119.
- Stephens, Matthew and Peter Donnelly (2000). **"Inference in molecular population genetics (with discussion)".** In: *Journal of the Royal Society of Statistics* 62, pp. 605–655.
- Wright, Sewall (1943). **"Isolation by distance".** In: *Genetics* 28, pp. 114–138.