Comment identifier la part de l'introgression qui est due à la sélection ?

Présenté par : Jules Romieu

Sous la direction de : François Rousset
Raphaël Leblois
Miguel de Navascués
Pierre-André Crochet

Membres du jury : Sylvain Glémin
Bertrand Servin
Cécile Berthouly-Salazar
Sophie Donnet
Maud Tenaillon
François Rousset

8 avril 2025

Plan :

1. **Introduction**
   - Définitions des processus étudiés
   - Objectifs et organisation de la thèse

2. **Présentation du modèle démo-génétique général**

3. **Comparaison des méthodes existantes d'inférence de l'introgression adaptative**

4. **Test d'une nouvelle méthode d'inférence de l'introgression adaptative**

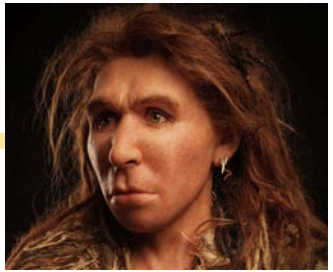
Contexte : Flux de gène entre espèces

Homme moderne

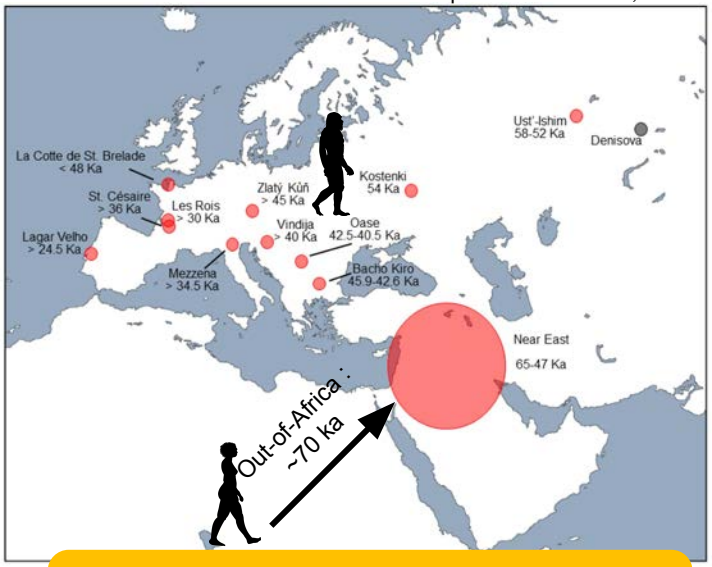
40-65 kya

Hybridation

Néandertal



Modifié d'après Churchill *et al.*, 2022



Dates et localités possibles d'hybridation entre l'homme et Néandertal

Contexte : Flux de gène entre espèces

Homme moderne

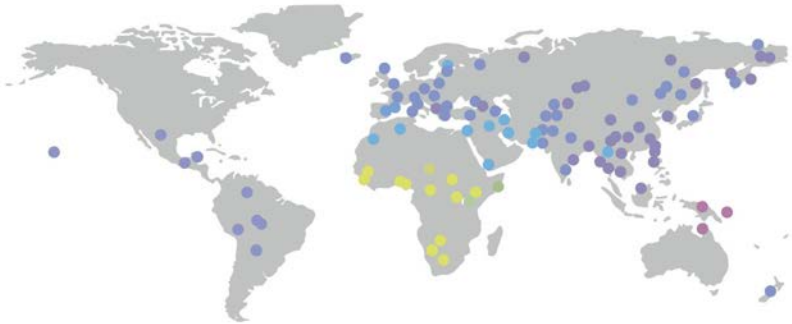
Néandertal

40-50 kya

Temps

Présent

Hybridation



Pourcentage d'ADN de Néandertal chez l'homme à travers le globe

Contexte : Flux de gène entre espèces

Homme moderne

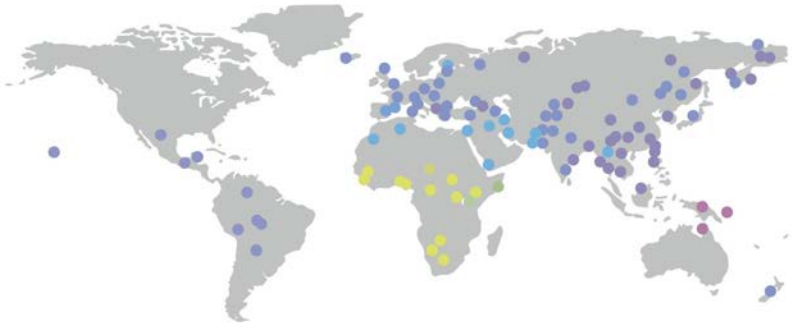
Néandertal

40-50 kya

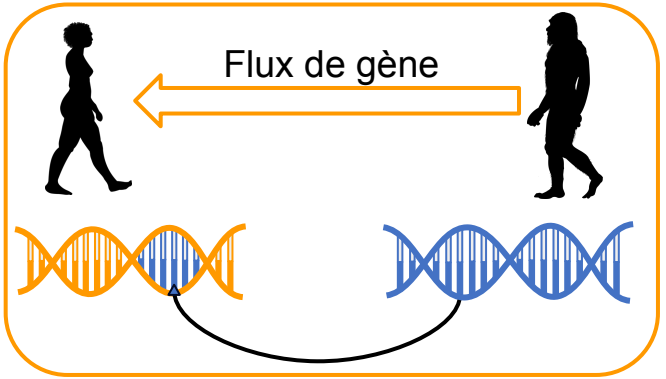
Temps

Présent

Hybridation



Pourcentage d'ADN de Néandertal chez l'homme à travers le globe



Contexte : Flux de gène entre espèces

Homme moderne

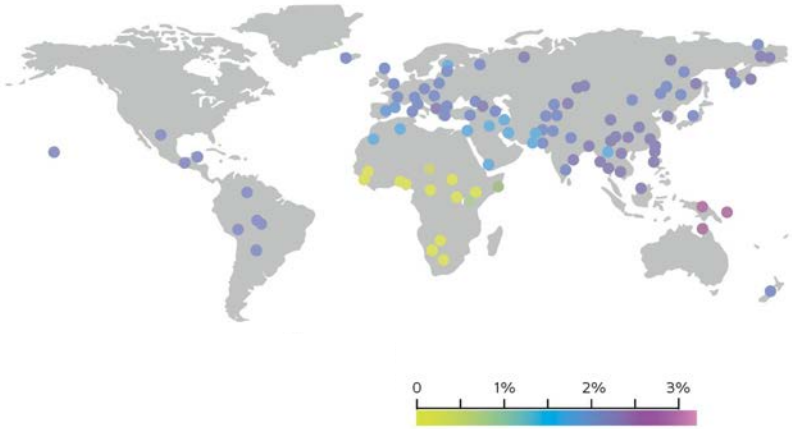
Néandertal

40-50 kya

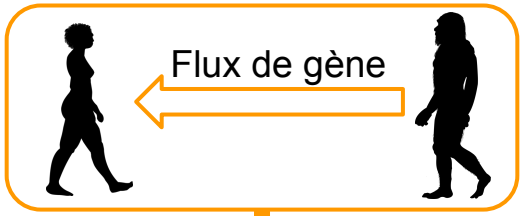
Temps

Présent

Hybridation



Pourcentage d'ADN de Néandertal chez l'homme à travers le globe



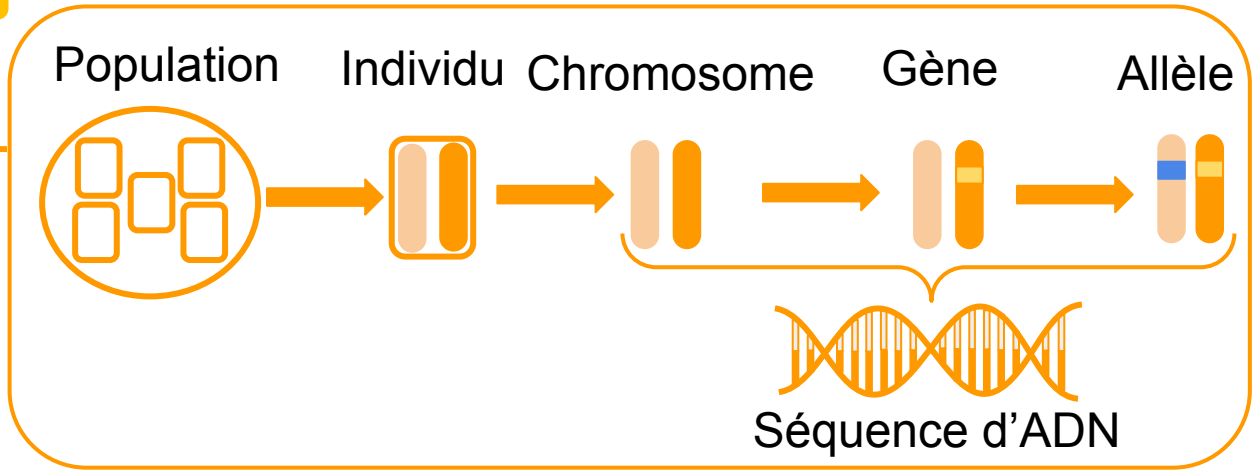
Hybridation et flux de gènes

- Source de diversité génétique
- Potentiellement adaptative

Impact sur l'adaptation et l'apparition des espèces ?

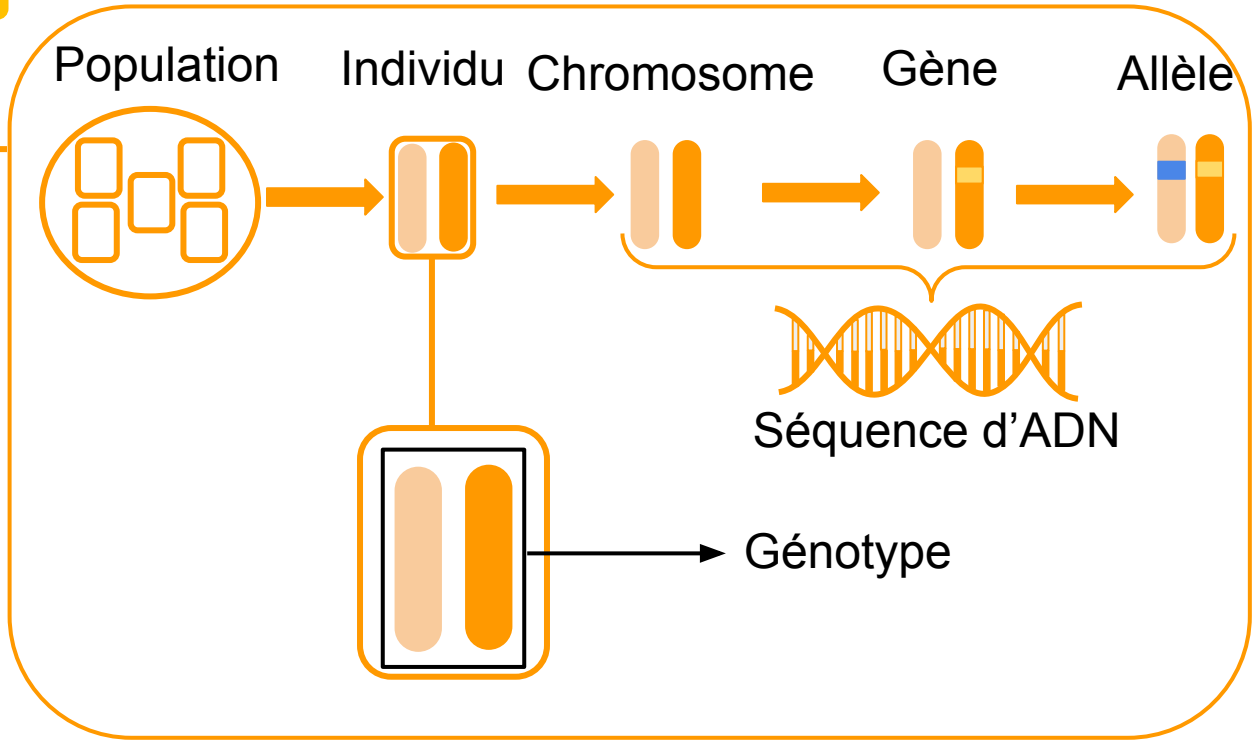
Contexte : Étudier la diversité génétique des populations

Espèce : *Homo sapiens*

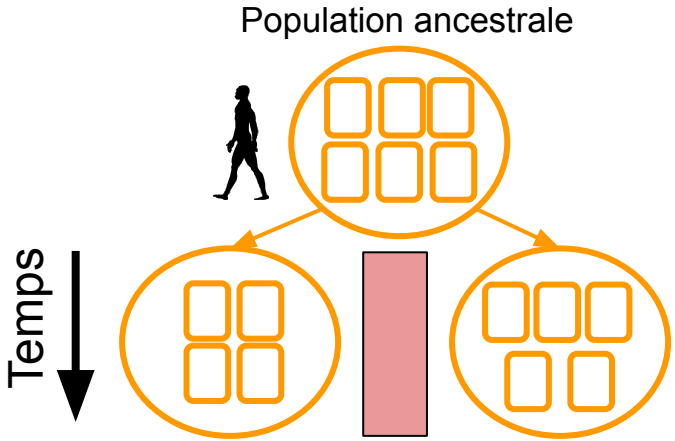


Contexte : Étudier la diversité génétique des populations

Espèce : *Homo sapiens*

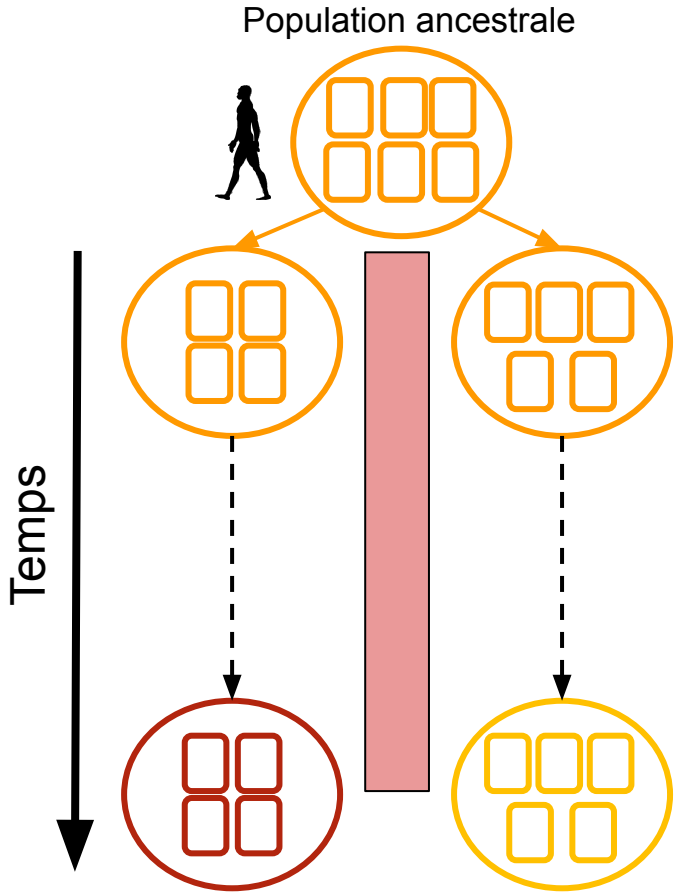


Processus d’isolement reproducteur :



1 **Apparition d’une barrière** :

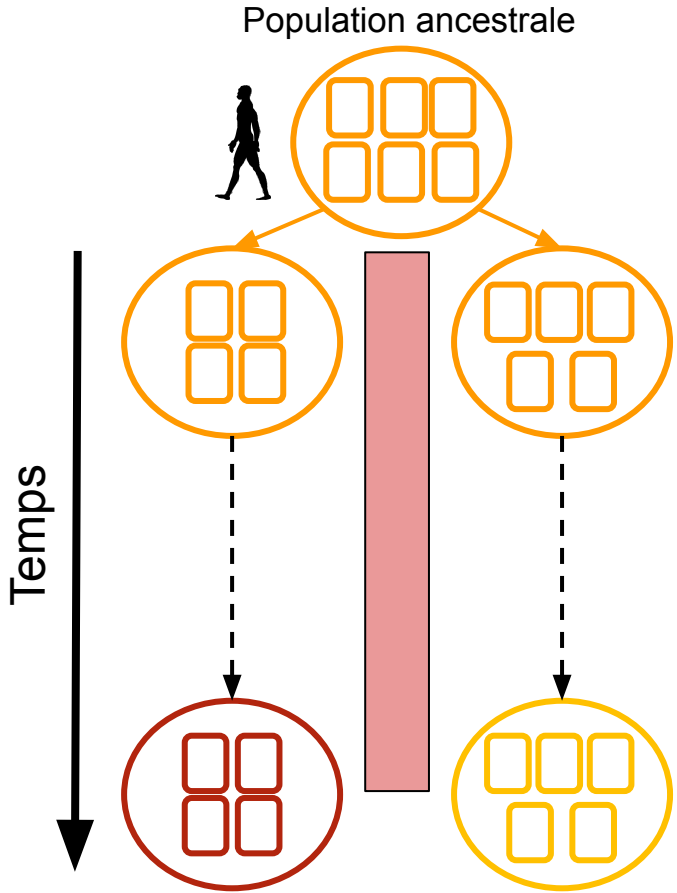
Processus d'isolement reproducteur :



1 Apparition d'une barrière :

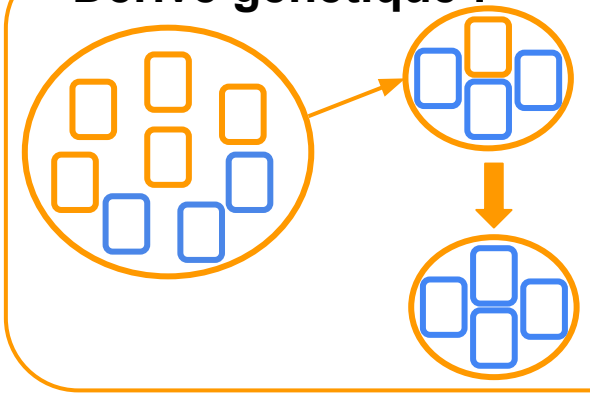
2 Forces évolutives

Processus d'isolement reproducteur :



2 **Forces évolutives**

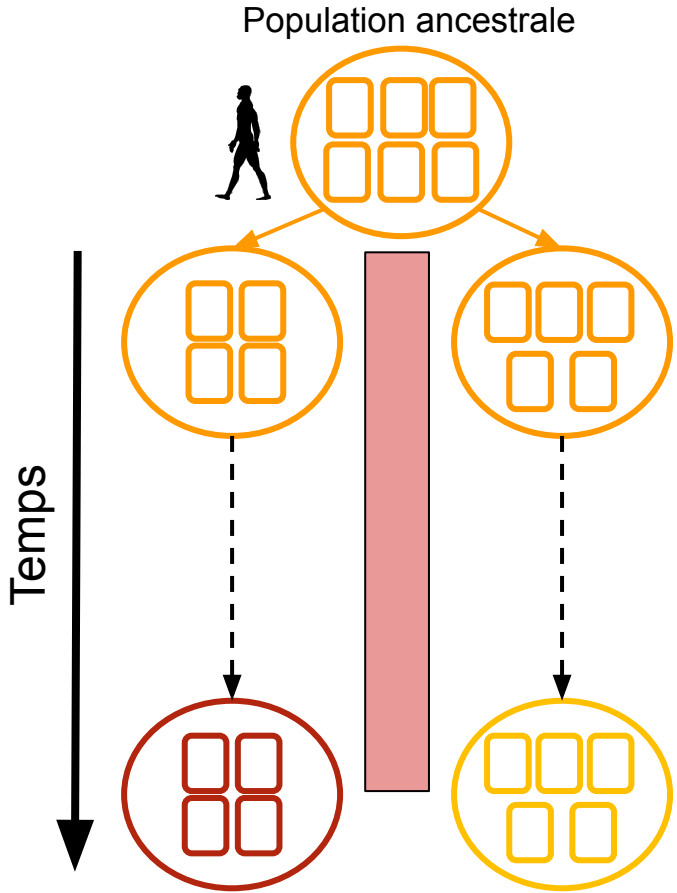
Dérive génétique :



Définition

Changement dans les fréquences alléliques due à la reproduction au hasard dans la population (certains individus se reproduisent plus que d'autres)

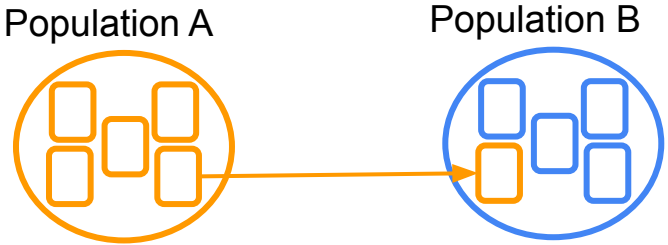
Processus d'isolement reproducteur :



1

2 **Forces évolutives**

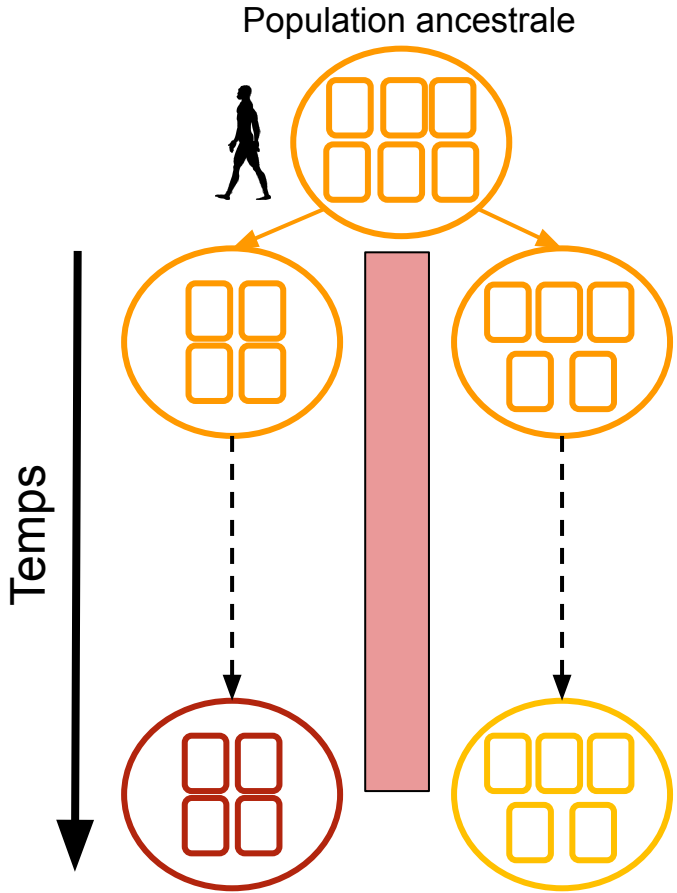
Migration :



Définition

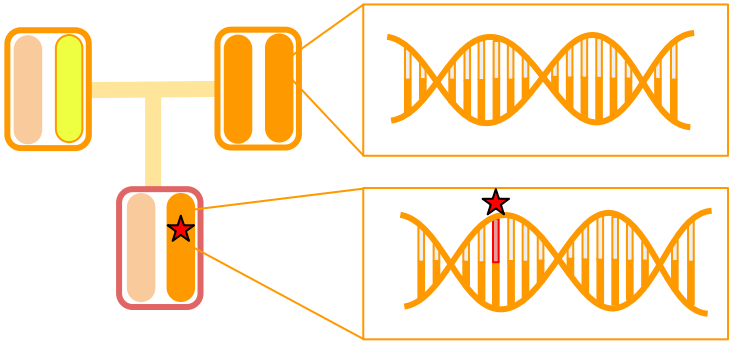
Déplacement des individus entre populations aboutissant à une homogénéisation de la diversité entre populations

Processus d'isolement reproducteur :



2 **Forces évolutives**

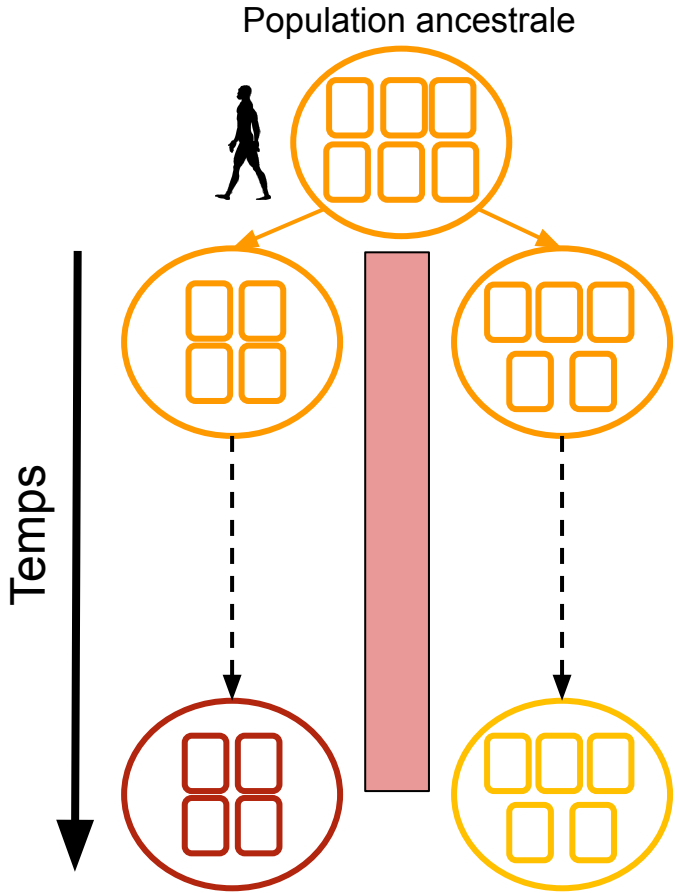
Mutation :



Définition

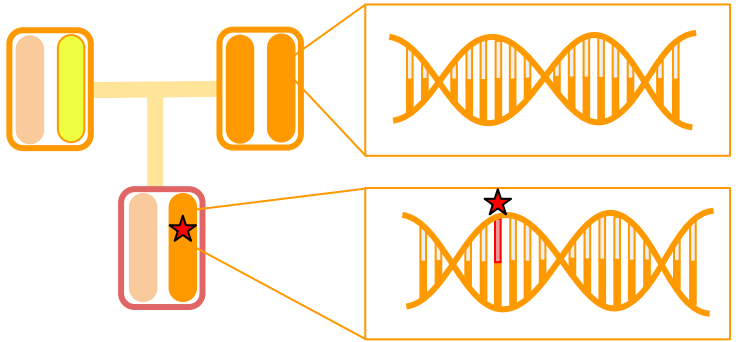
Modification des séquences d'ADN lors de la reproduction. Seule source de variation génétique

Processus d'isolement reproducteur :

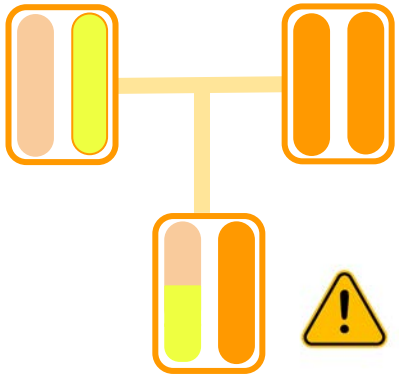


2 **Forces évolutives**

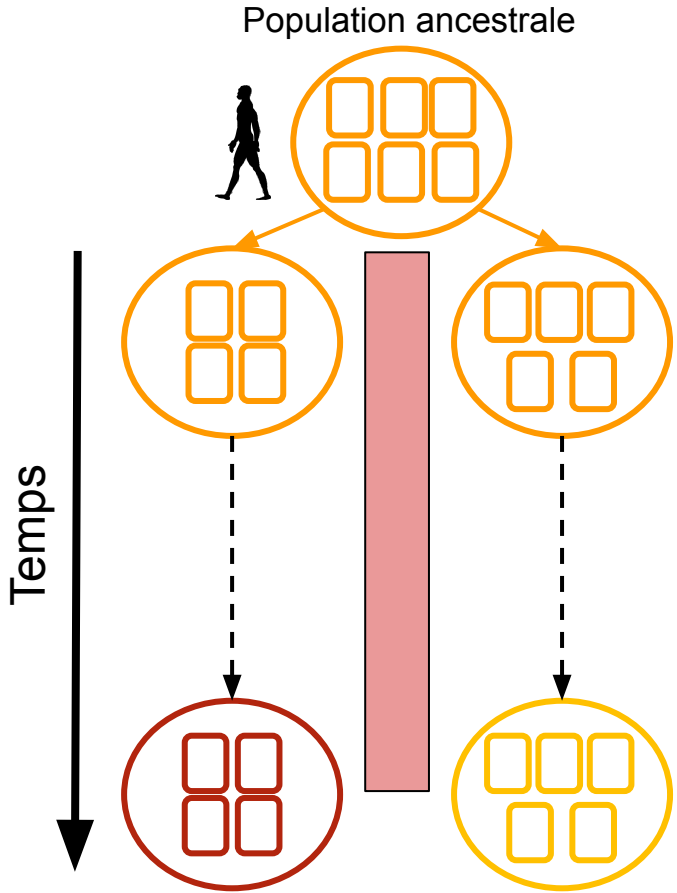
Mutation :



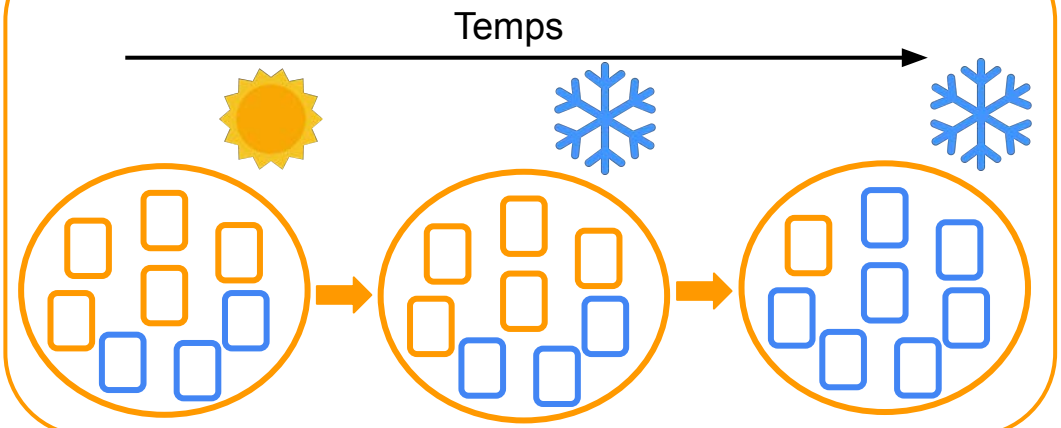
Recombinaison :



Processus d'isolement reproducteur :



Sélection naturelle :

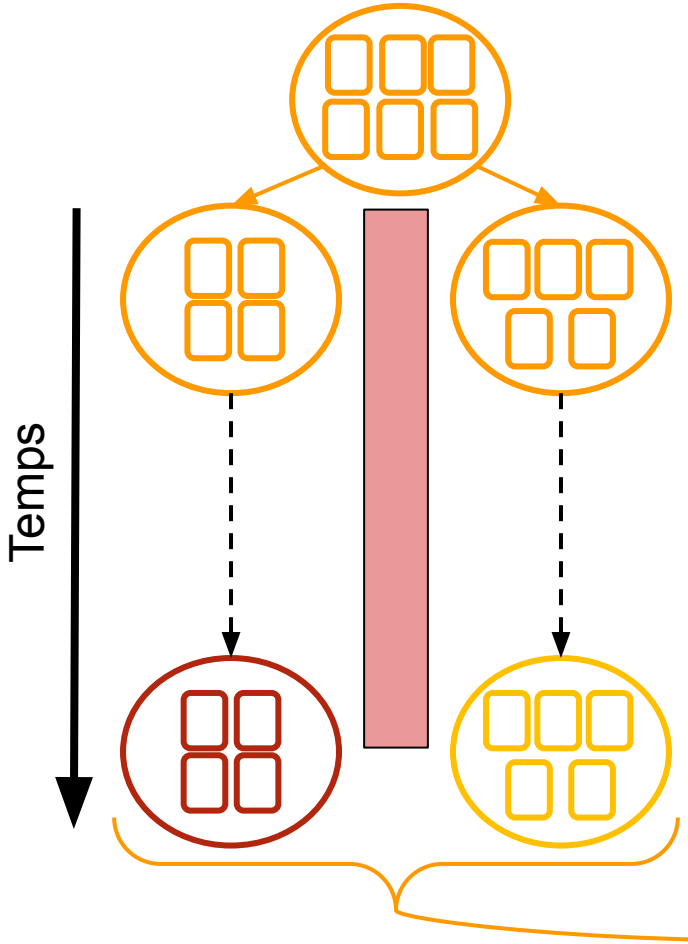


2 **Forces évolutives**

Définition

Changement dans les fréquences alléliques d'une population du fait des effets avantageux de certaines mutations sur la reproduction des individus qui les portent

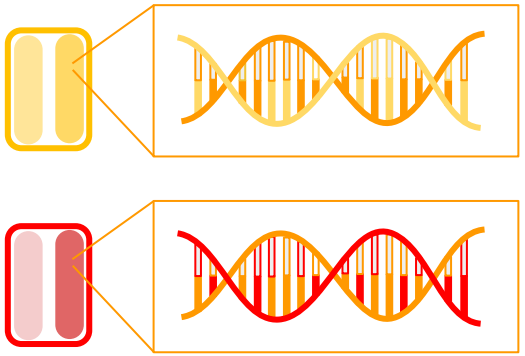
Processus d'isolement reproducteur :



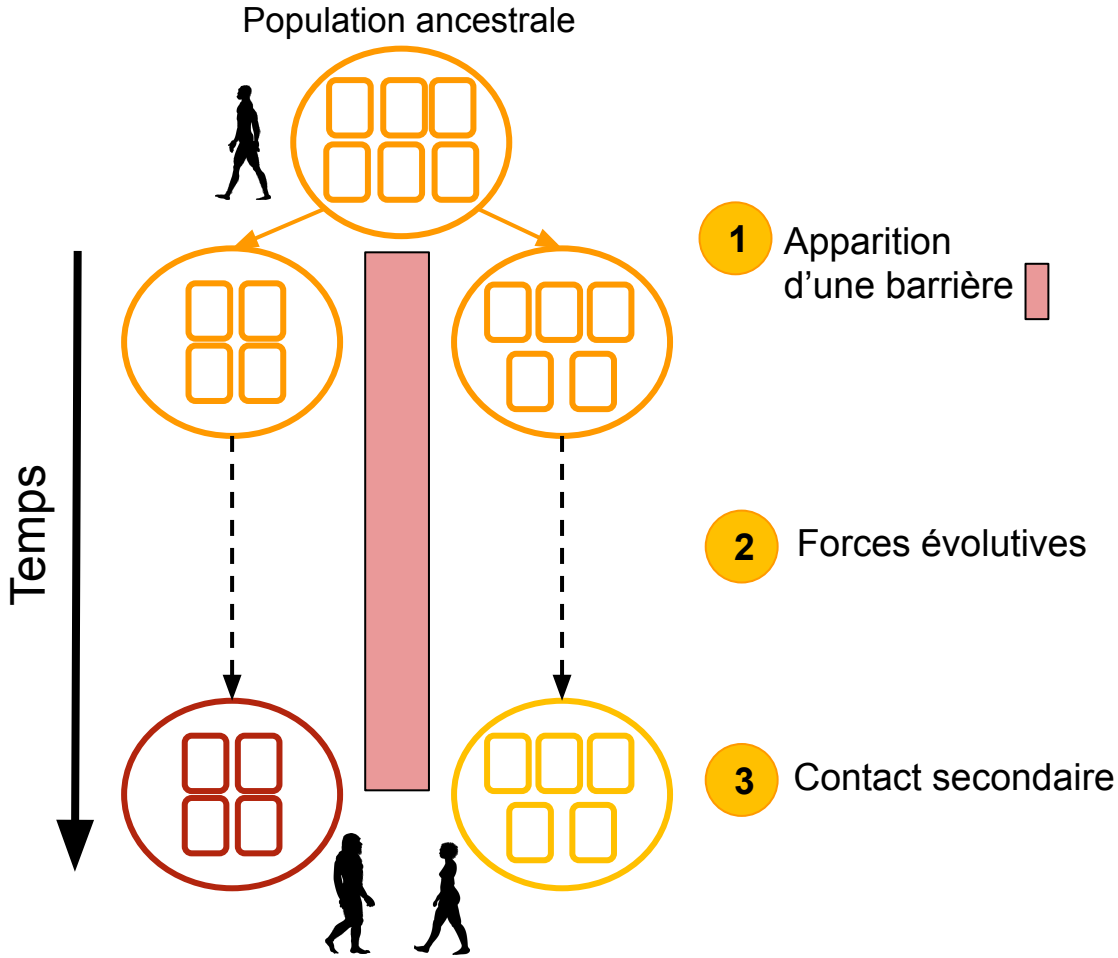
1 Apparition d'une barrière

2 Forces évolutives

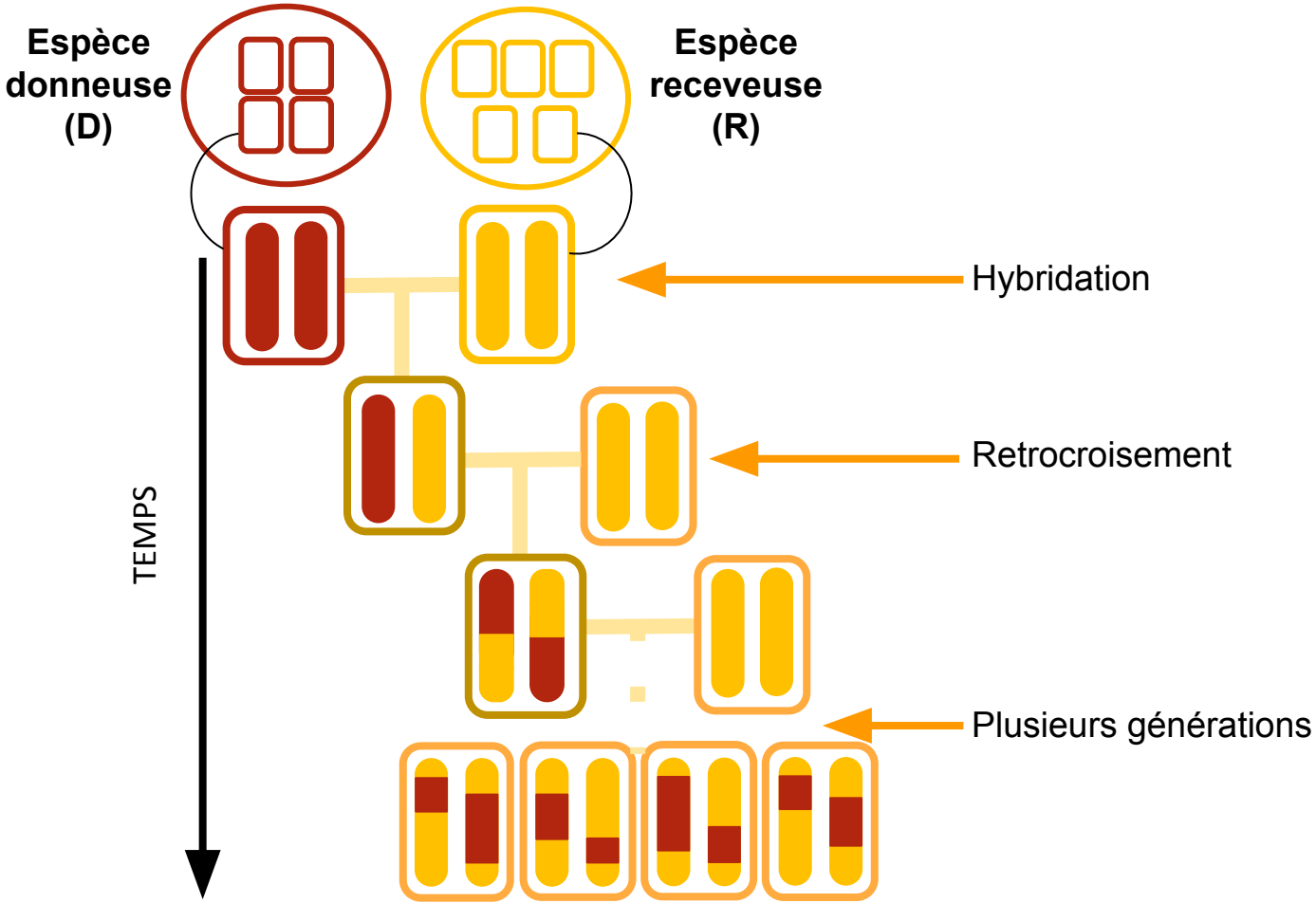
Divergence génétique :



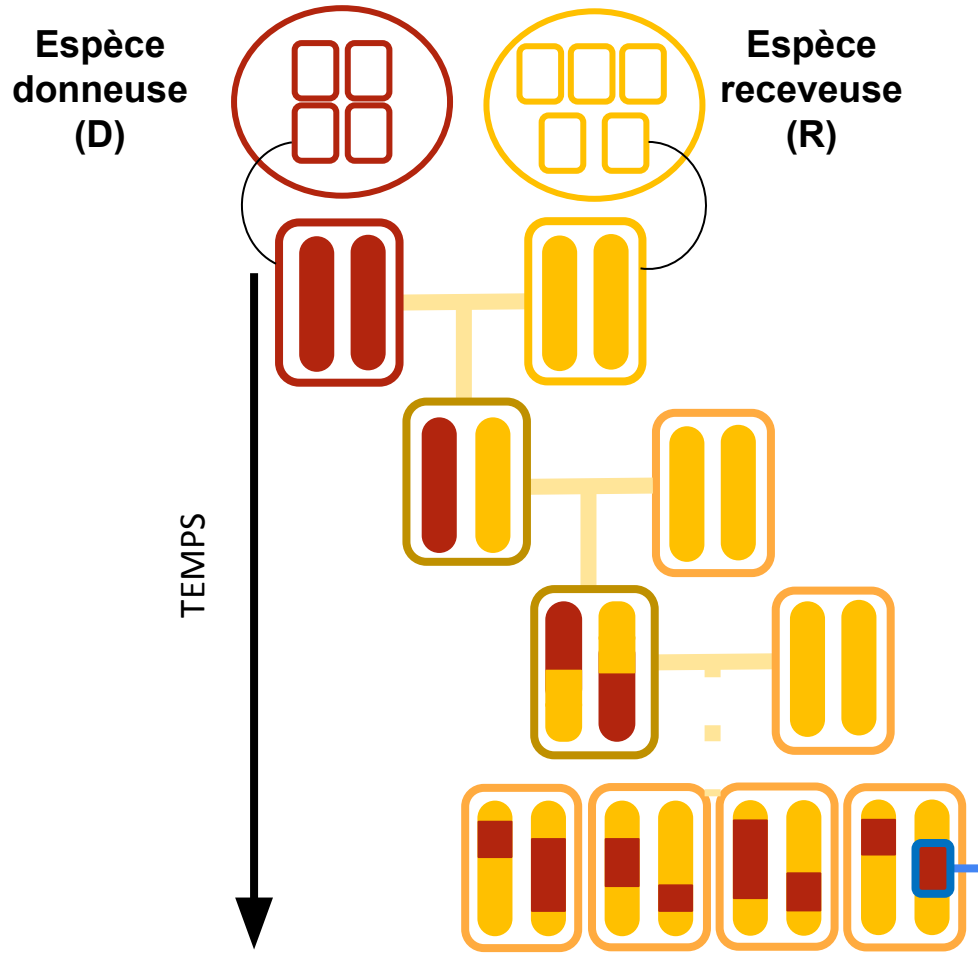
Processus d'isolement reproducteur :



Processus d'introgression :



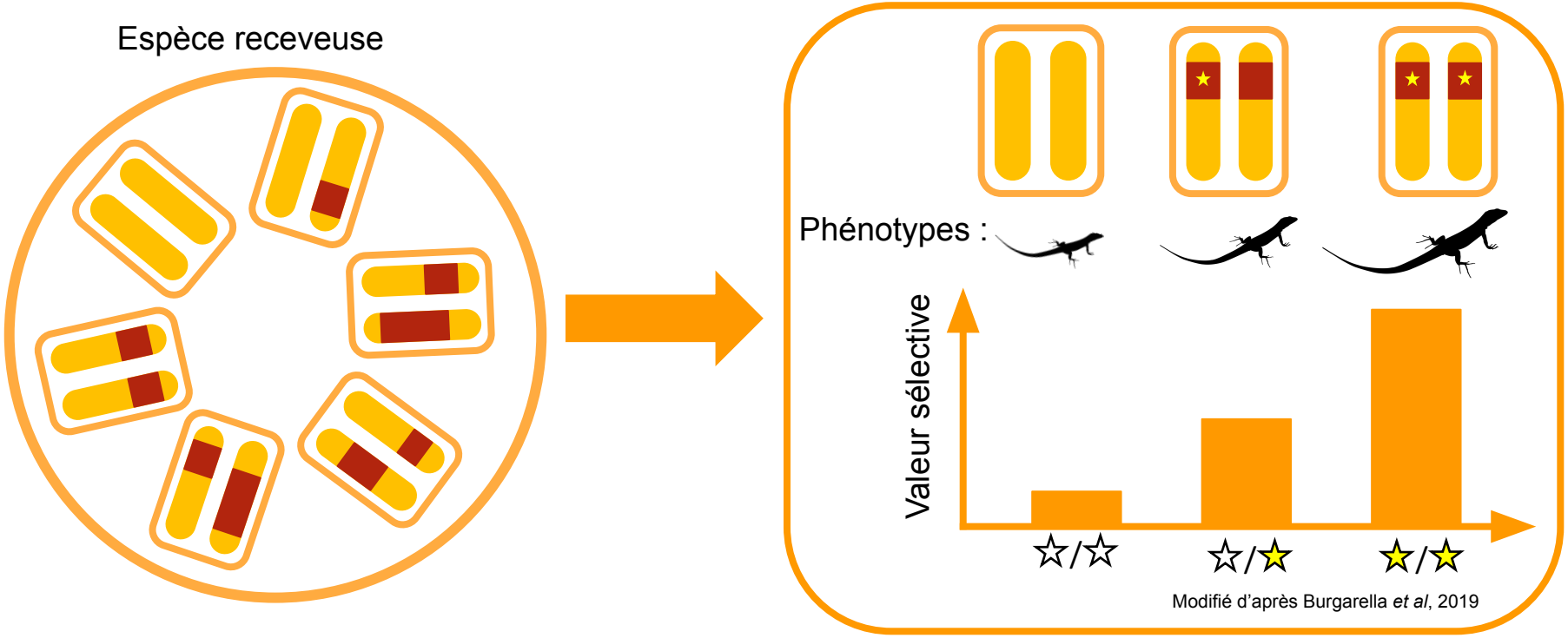
Processus d'introgression :



Introgression (I)

Incorporation de matériel génétique d'une espèce à une autre par hybridation et rétrocroisements répétés.

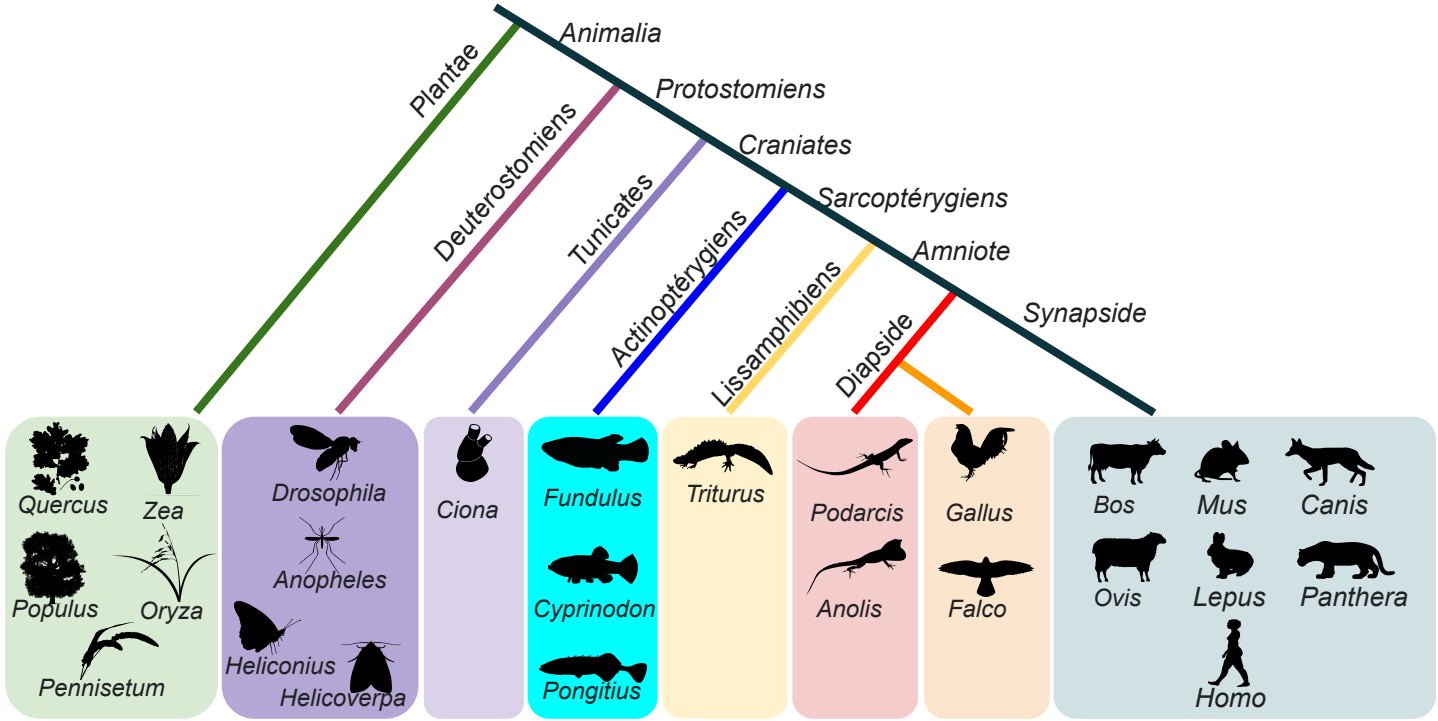
Introgression adaptative :



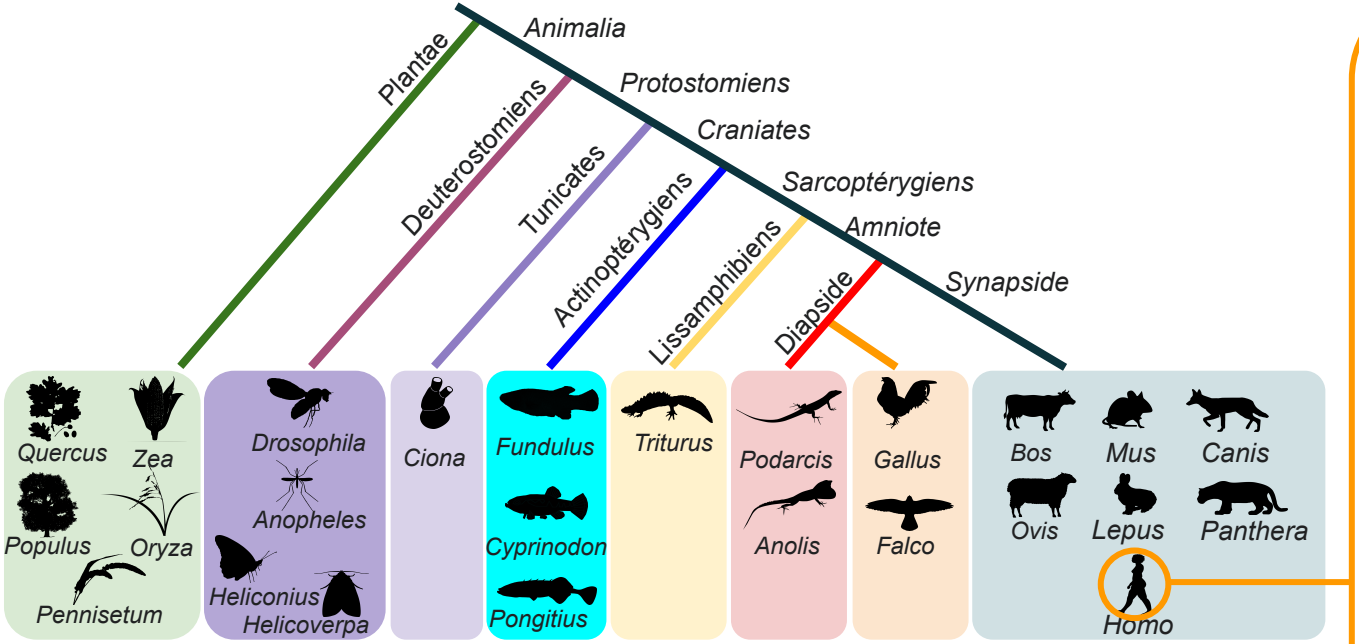
Introgression adaptative (IA)

Introgression d'un allèle bénéfique entraînant une augmentation de la valeur sélective des individus porteurs de cet allèle

Distribution de l'IA dans le règne du vivant :

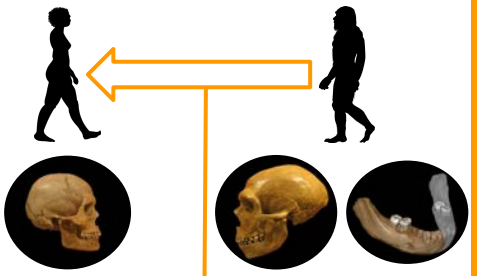


Exemple d'IA dans le règne du vivant



Introgression ancienne

40 - 56 kya



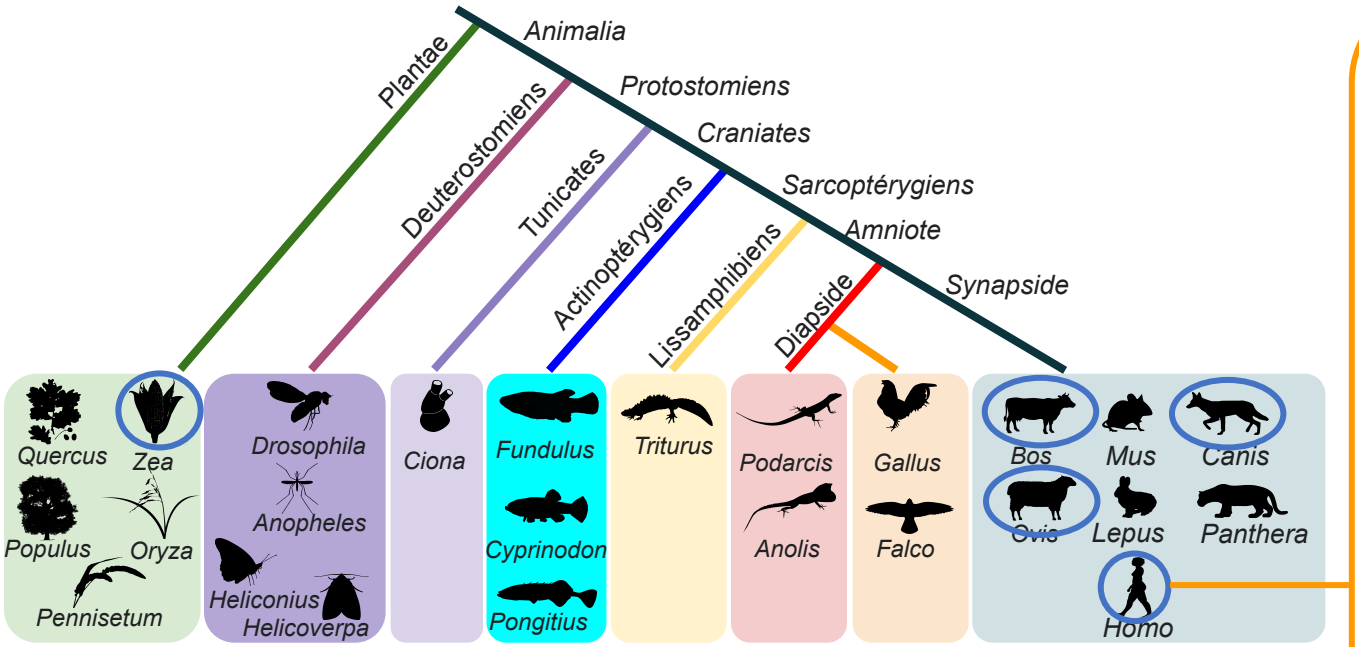
Homo sapiens

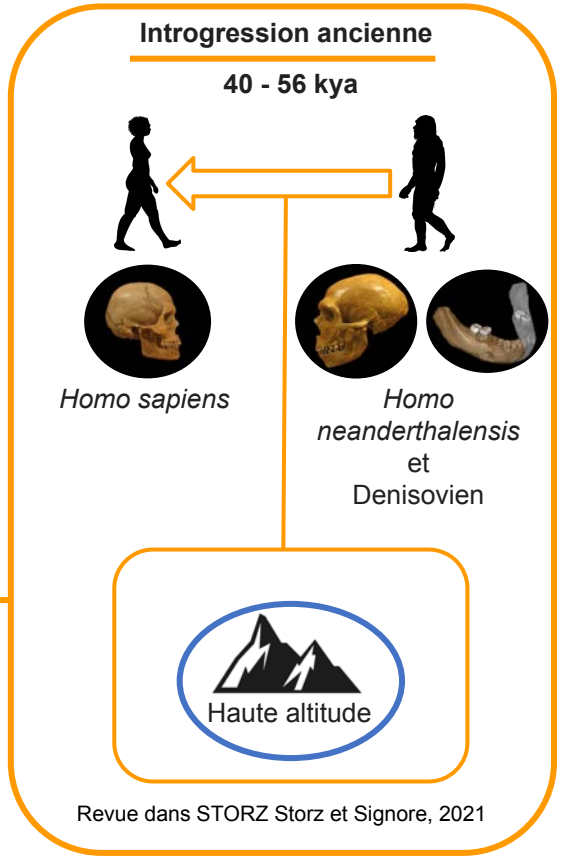
Homo neanderthalensis et Denisovien

Haute altitude

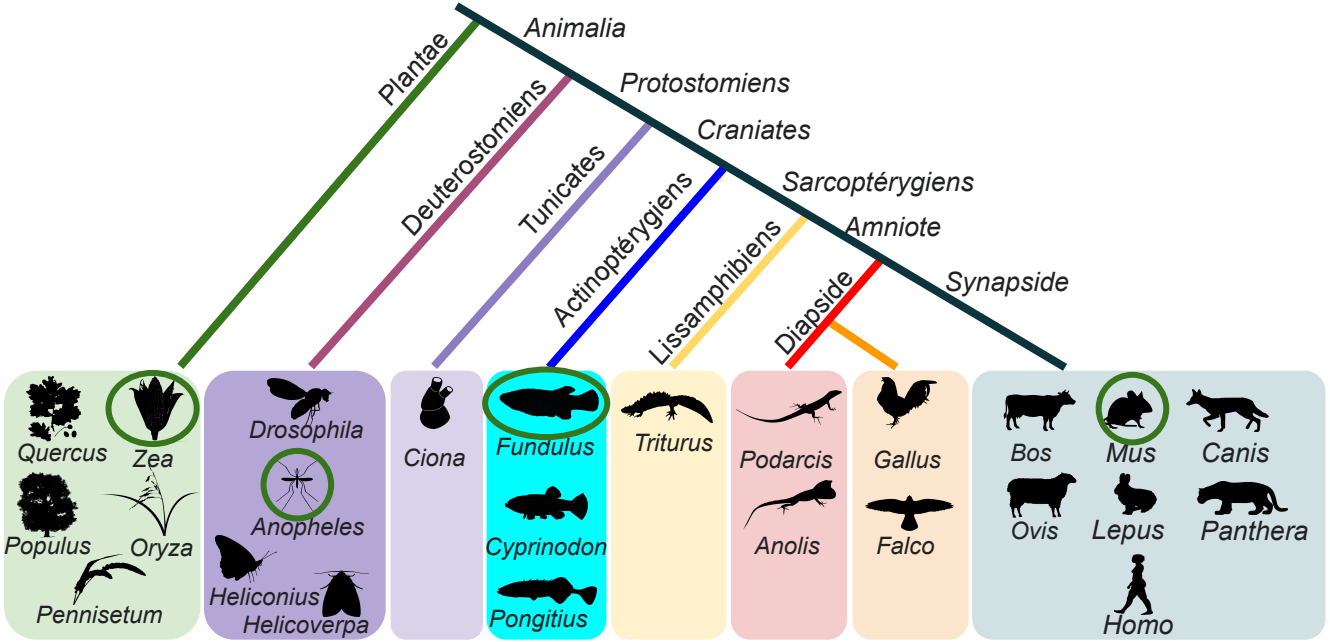
Huerta-Sánchez *et al.*, 2014; Hackinger *et al.*, 2016; Hu *et al.*, 2017

Exemple d'IA dans le règne du vivant





Exemple d'IA dans le règne du vivant



Dégradation environnementale

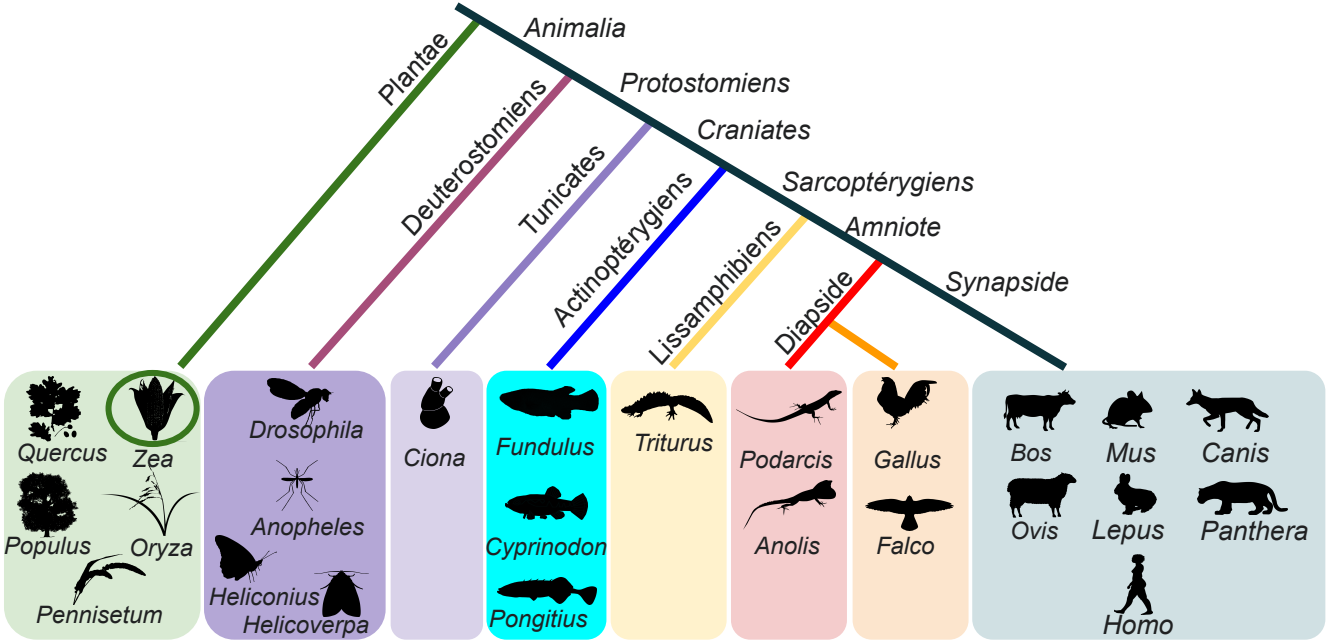
Rodenticides — Song *et al.*, 2011

Toxines — Oziolor *et al.*, 2019

Insecticides — Hanemaaijer *et al.*, 2019

Herbicides — Le Corre *et al.*, 2020

Enjeux de l'étude de l'IA :



Incorporation de gènes cultivés dans les plantes sauvages :

- Expansion des mauvaises herbes résistantes
- Invasion de gènes modifiés génétiquement

Implication en agronomie

Plantes sauvages ← **Plantes cultivées**

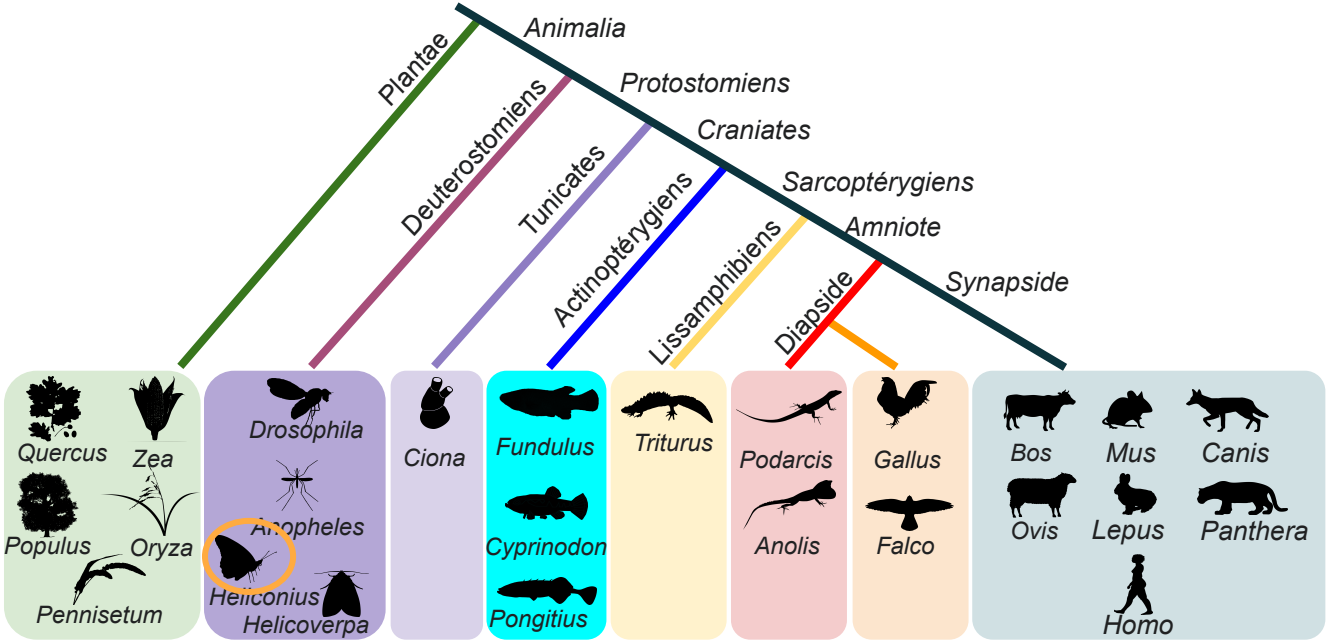
Téosintes européennes

Maïs

Résistance aux herbicides passant des plantes cultivées aux plantes sauvages

Le Corre *et al.*, 2020

Enjeux de l'étude de l'IA :



L'introgression adaptative peut-elle faciliter la spéciation ?

- Par l'acquisition de gènes entraînant l'isolement reproducteur

Implication en biologie évolutive

0.7%

Espèce receveuse ← **Espèce donneuse**

Heliconius pardalinus

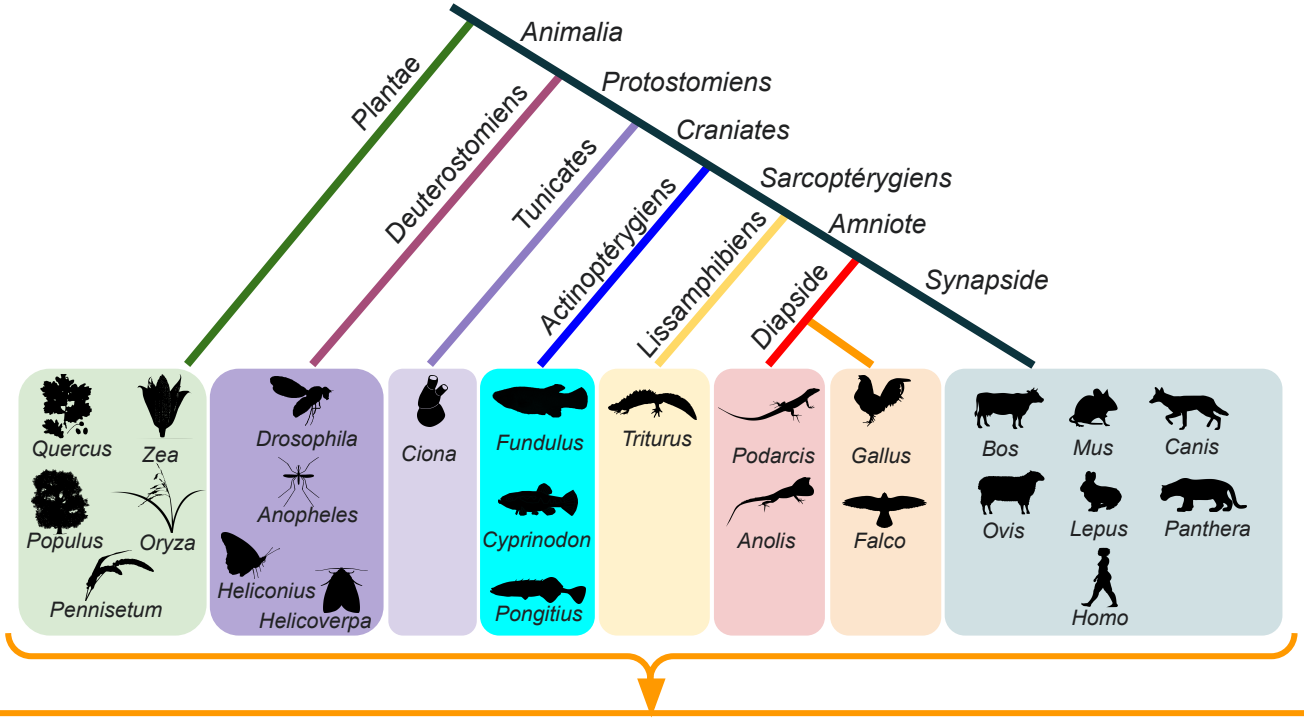
H. melpomene

H. elevatus

Acquisition de pattern d'aile utilisés pour la reproduction

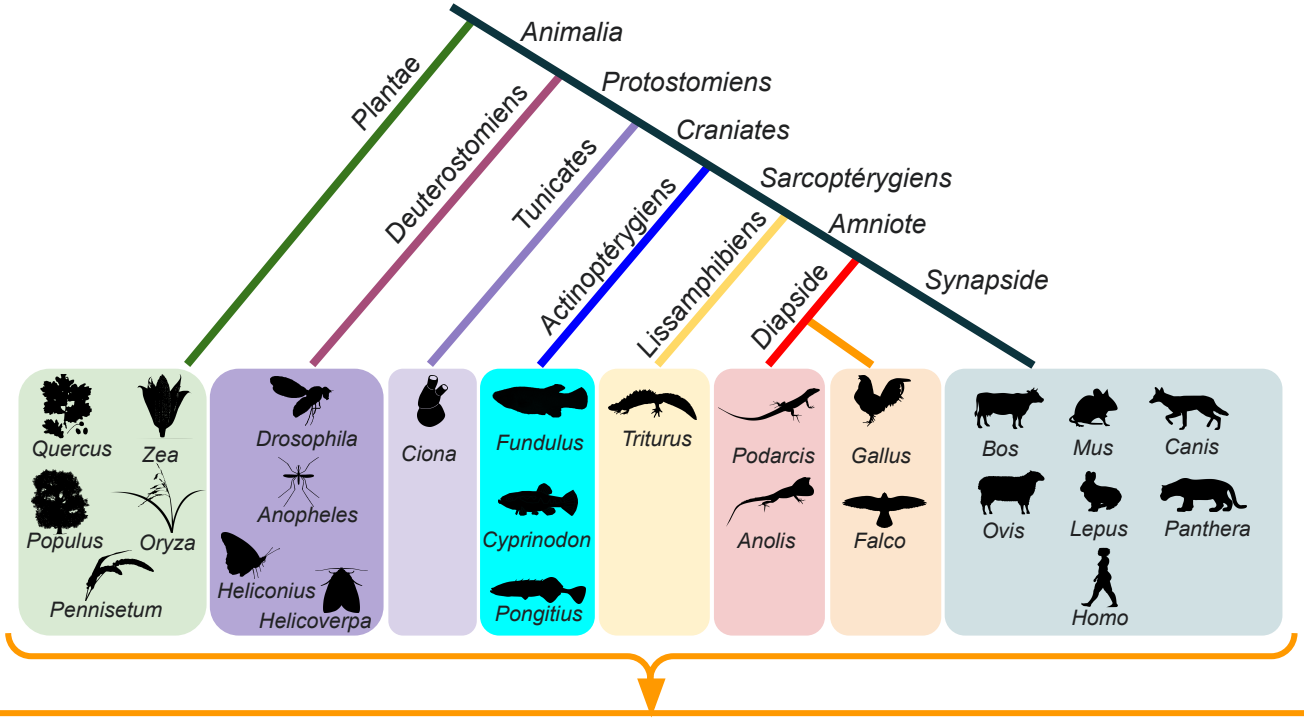
Padro-Diaz *et al.*, 2012 ; Rosser *et al.*, 2024

Prévalence de l'IA dans le règne du vivant :



- Inférence de l'IA en deux temps : (1) Introgression (2) Sélection
- Méthodes récentes : Uniquement de la classification (IA vs non-IA)

Prévalence de l'IA dans le règne du vivant :



- Inférence de l'IA en deux temps : (1) Introgression (2) Sélection
- Méthodes récentes : Uniquement de la classification (IA vs non-IA)

Besoin de nouvelles méthodes pour quantifier l'IA dans les génomes

Objectif de la thèse :

Identifier la part d'introgression qui est due à la sélection ?

Données génétiques

Méthodes d'inférence de l'IA

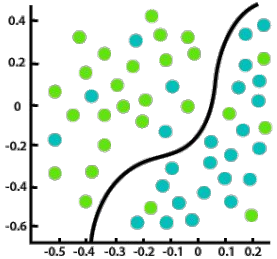
Deux approches possibles :

1. **Utiliser les méthodes existantes d'inférence de l'IA**
2. **Développer de nouvelles méthodes d'inférence de l'IA**

Organisation de la thèse :

1 **Comparer les méthodes de classification de l'IA existantes**

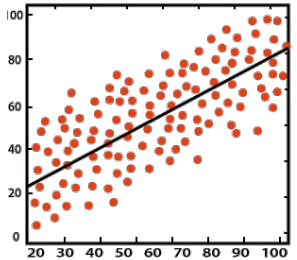
→ 3 méthodes et une statistique résumante



Classification

2 **Développer et tester une nouvelle méthode d'inférence de l'IA**

→ 1 méthode d'**estimation** par simulation des valeurs de paramètres



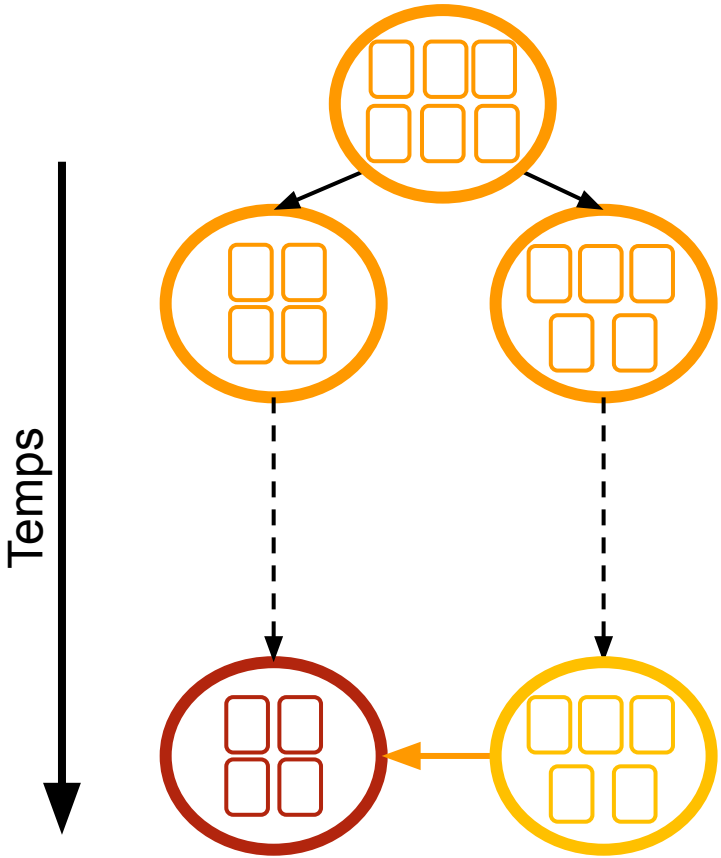
Regression

3 **Utiliser une approche de test par simulation**

→ Utilisation de données génétiques tests et d'entrainement des méthodes simulées

Simuler les données génétiques :

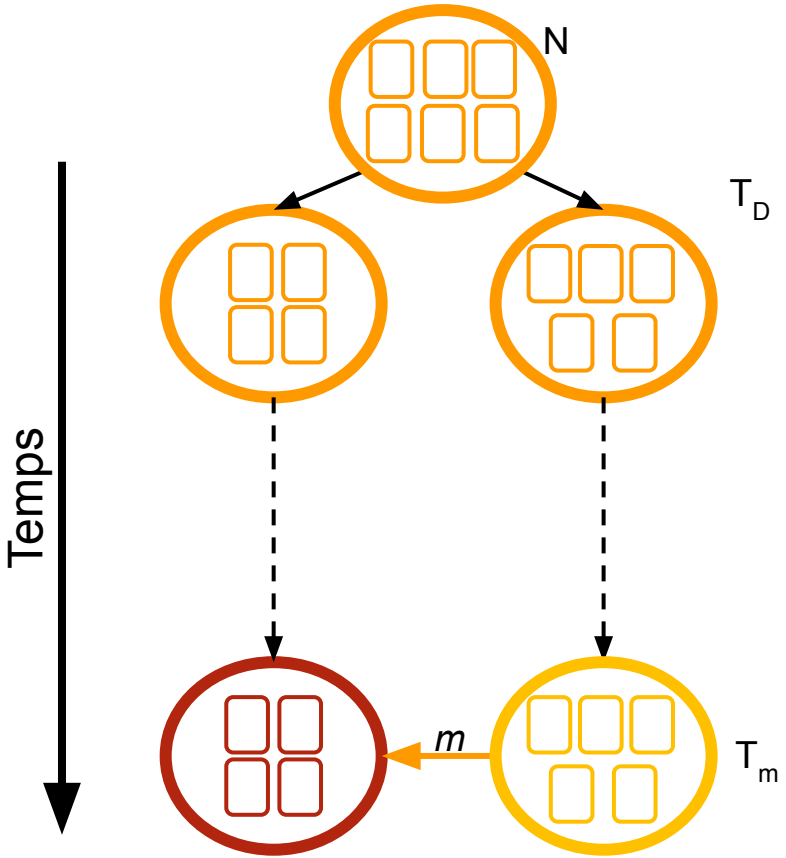
Le modèle démographique :



Paramètres définissent le modèle qui génère les données génétiques

Simuler les données génétiques :

Le modèle démographique :

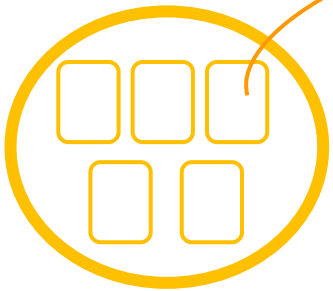


Paramètres du modèle démographique :

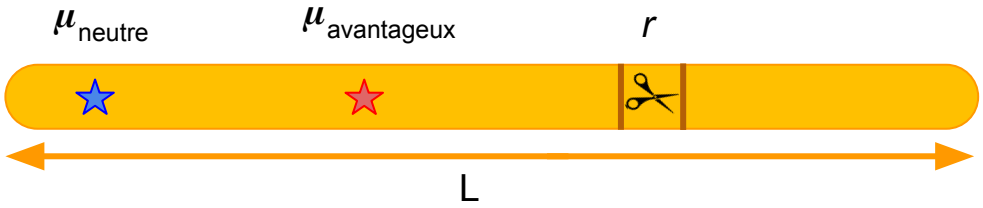
N = Taille de la population (nombre d'individus)
T_D = Temps de divergence (en génération)
T_m = Temps de migration (en génération)
m = Taux de migration

Simuler les données génétiques

La structure du génome :

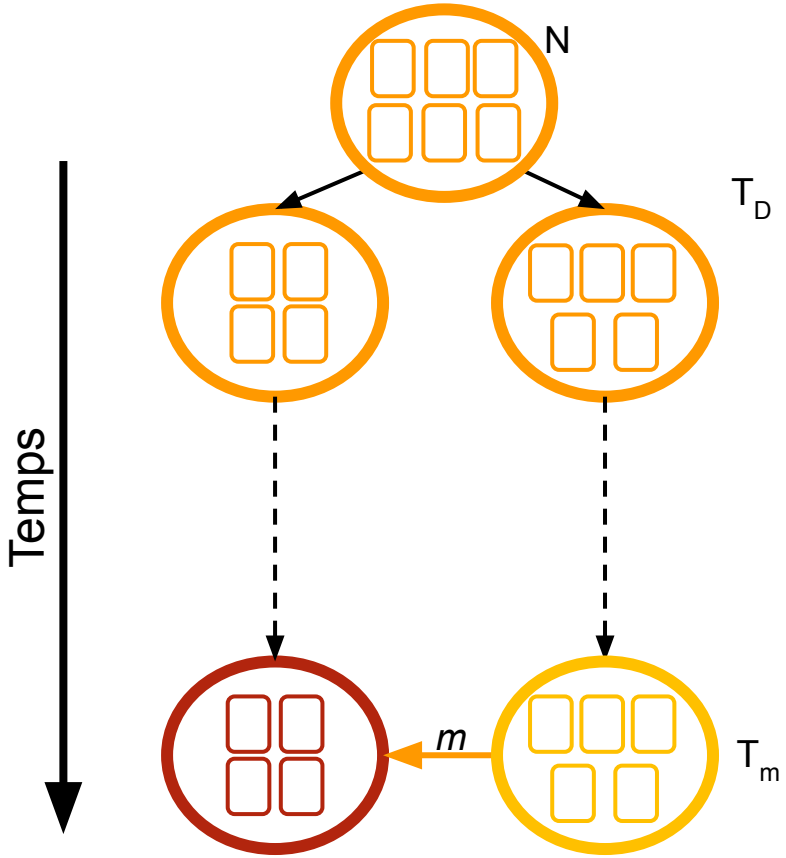


Paramètres définissant la structure du génome :



μ_{neutre} = Taux de mutation neutre
$\mu_{avantageux}$ = Taux de mutation avantageux
s = coefficient de sélection
r = Taux de recombinaison
L = Longueur du génome (en nucléotide)

Simuler les données génétiques :



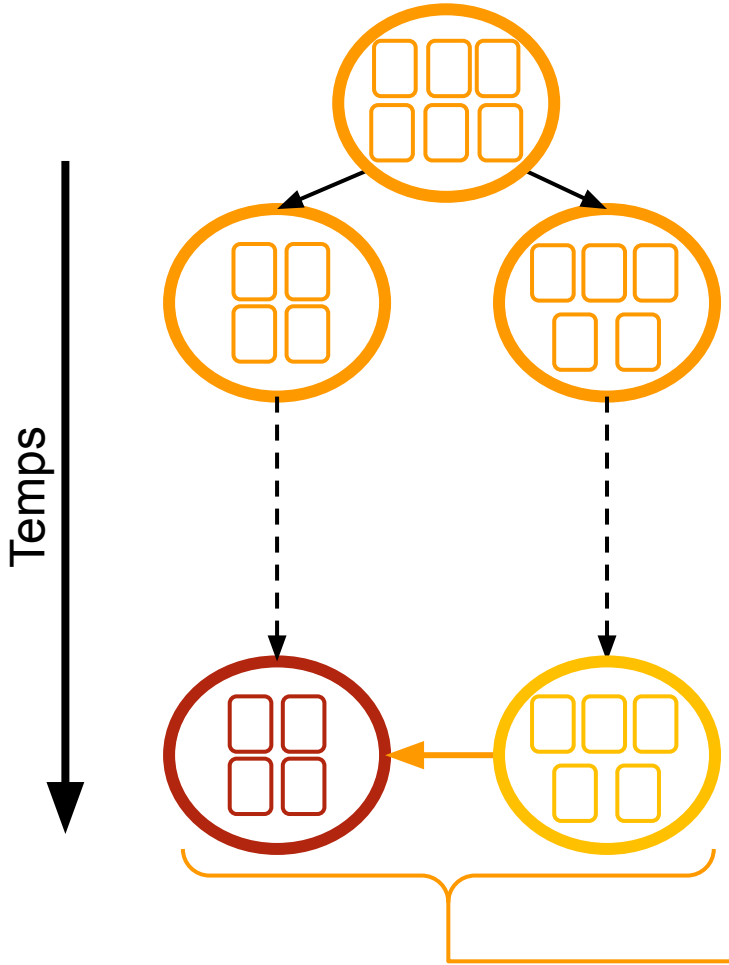
Paramètres du modèle démographique :

N = Taille de la population (nombre d'individus)
T_D = Temps de divergence (en génération)
T_m = Temps de migration (en génération)
m = Taux de migration

Paramètres de la structure du génome

Fixer les valeurs permet de définir un scenario démo-génétique donné

Cadre d'inférence :



Données génétiques simulées

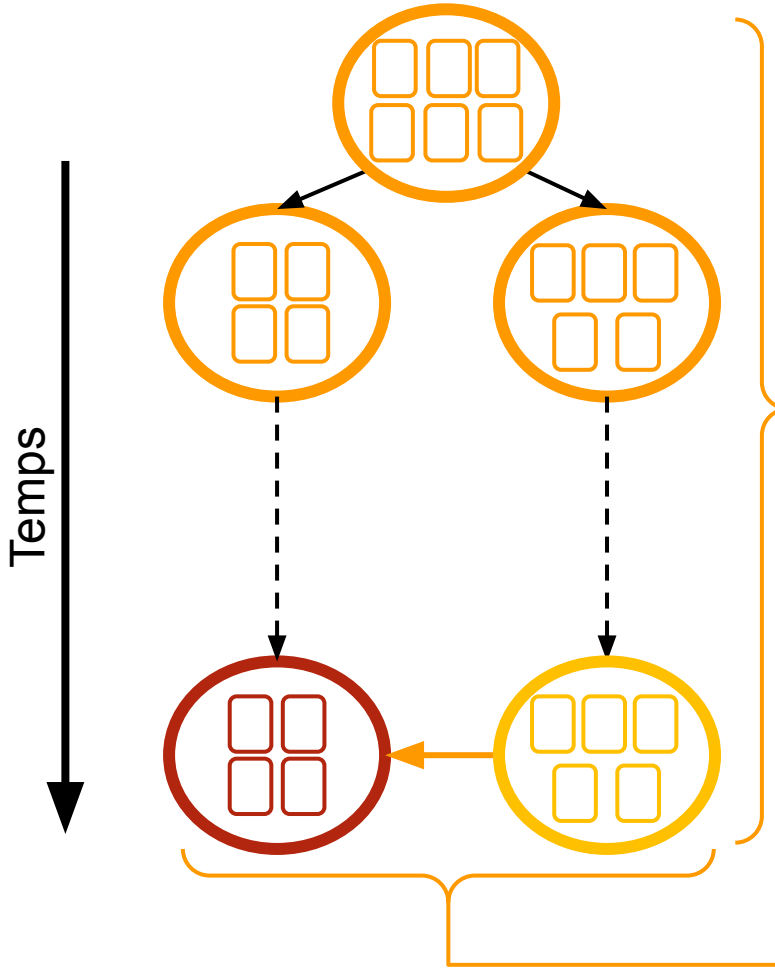
Matrice génotypique :

ID des sites

Echantillons

	0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	A	A	C

Cadre d’inférence :

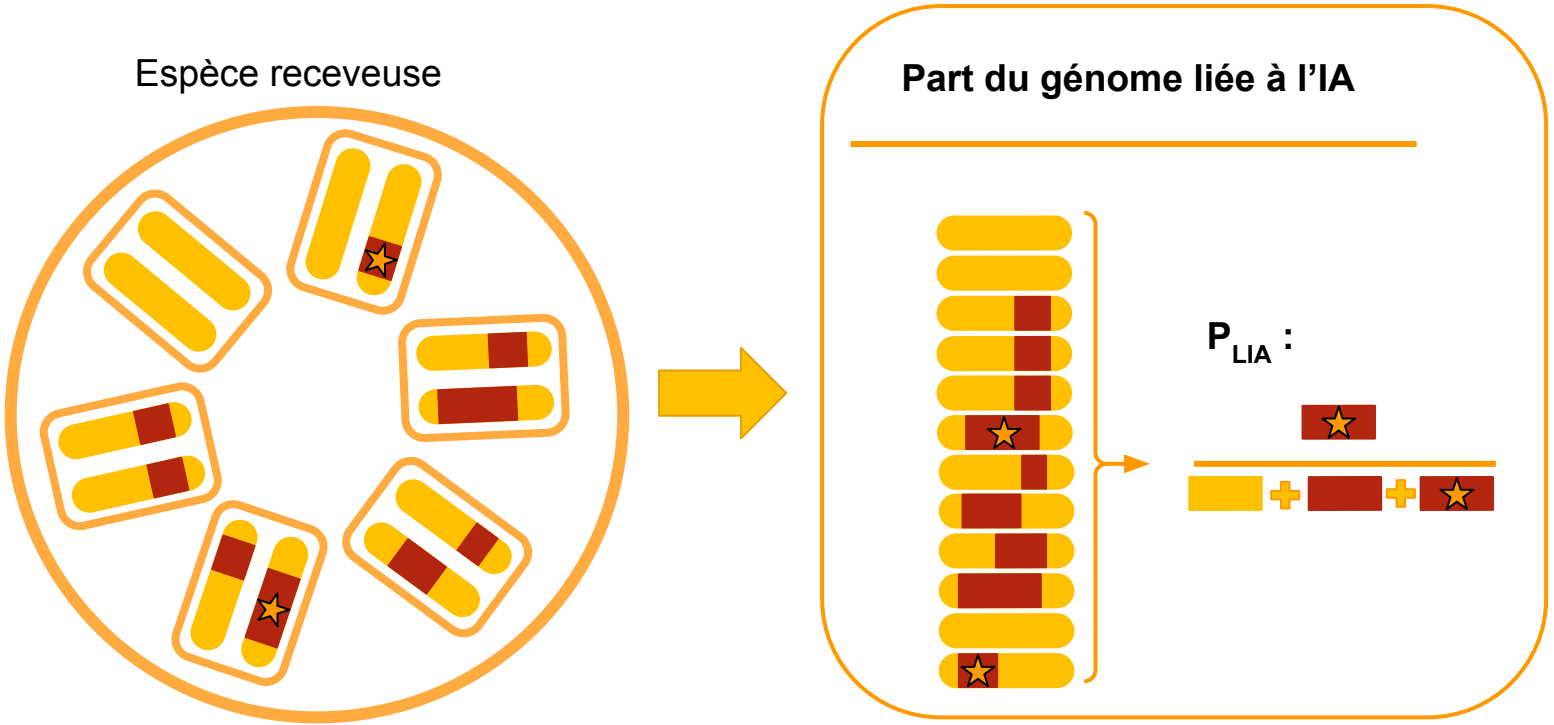


Variables latentes

- Aléatoires pour des paramètres donnés
- Non observables dans les données

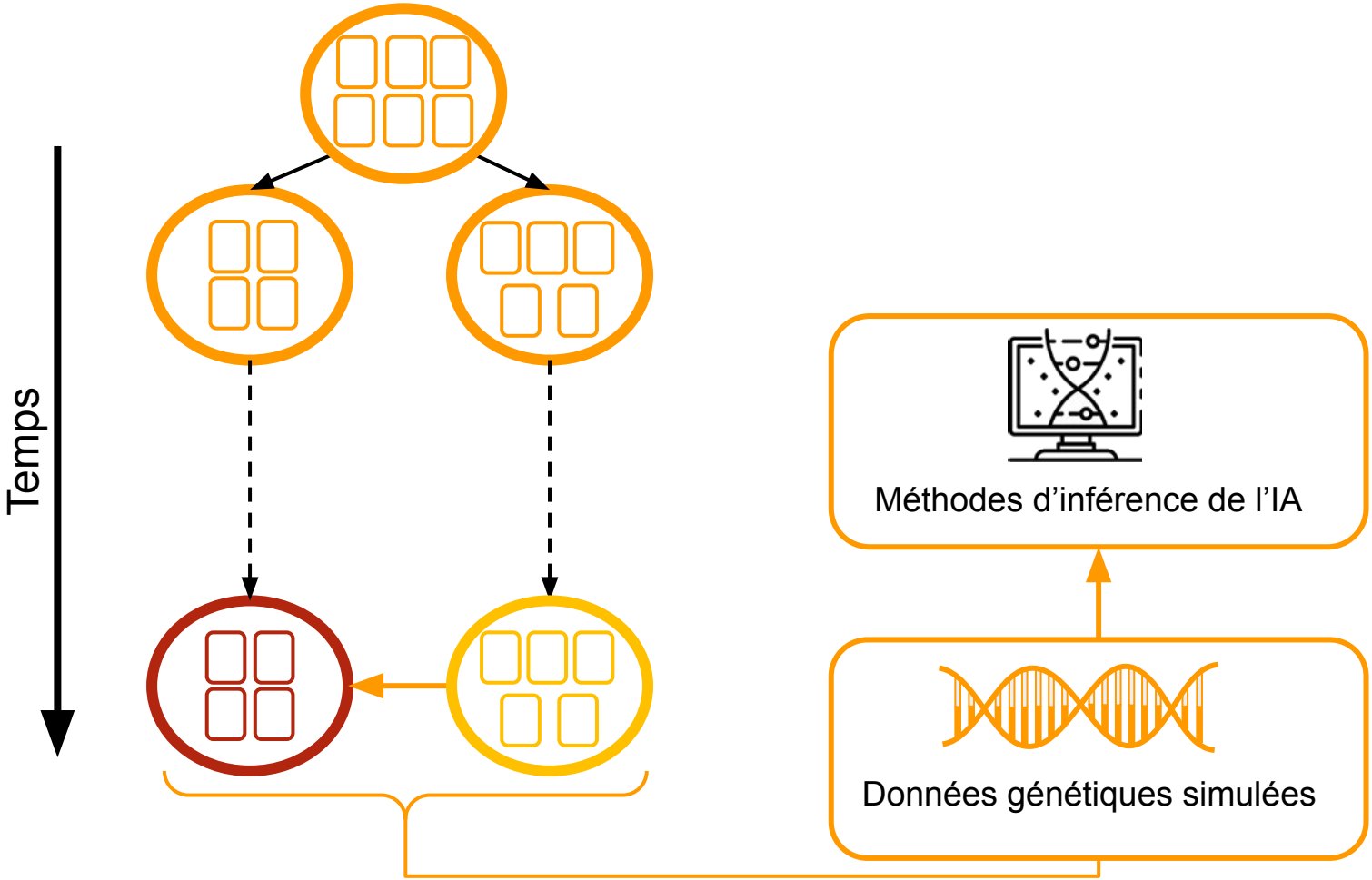
Données génétiques simulées

Exemple de variable latente :

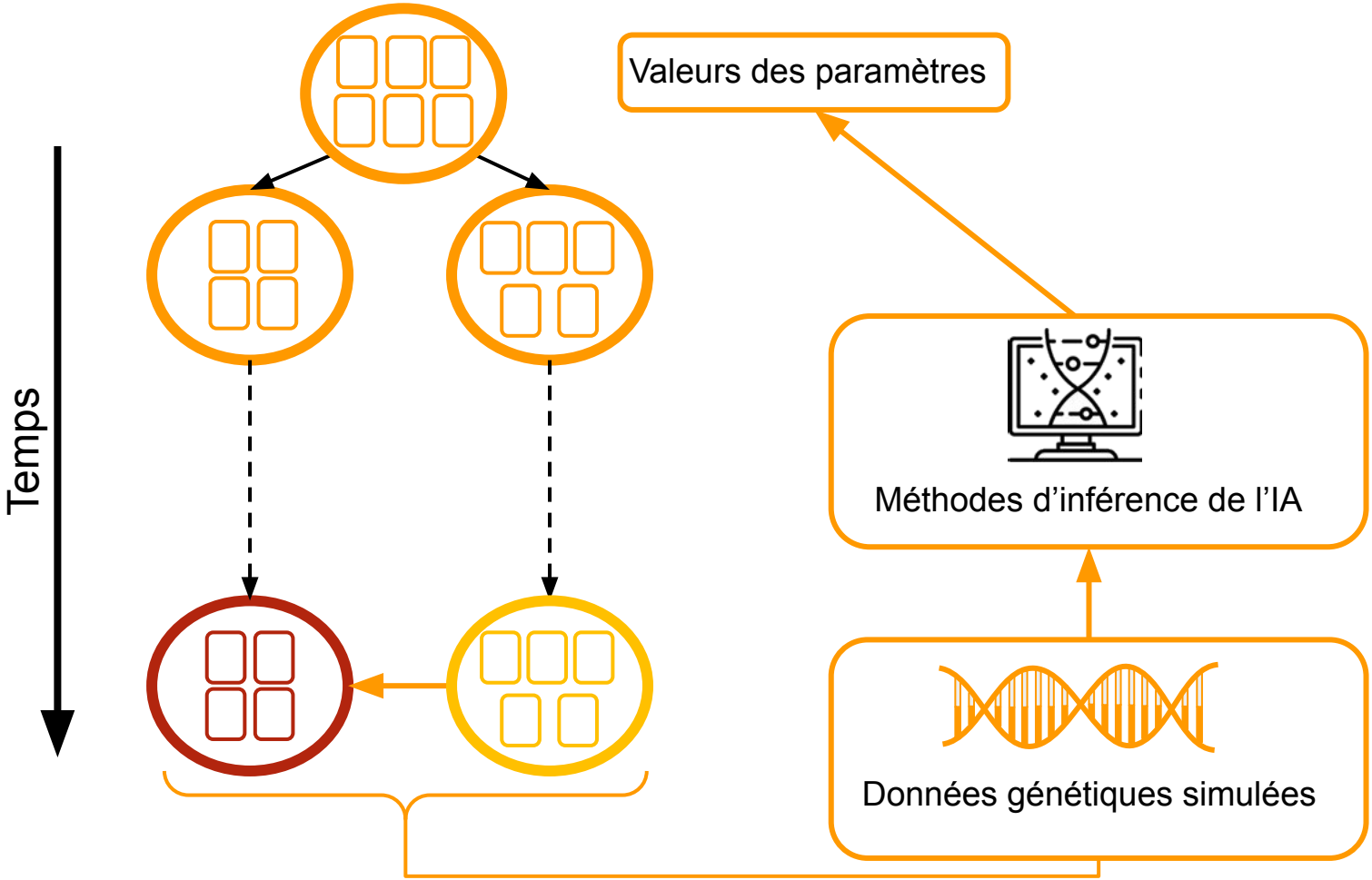


Dépend du coefficient de sélection (*s*) mais varie pour une même valeur de *s*

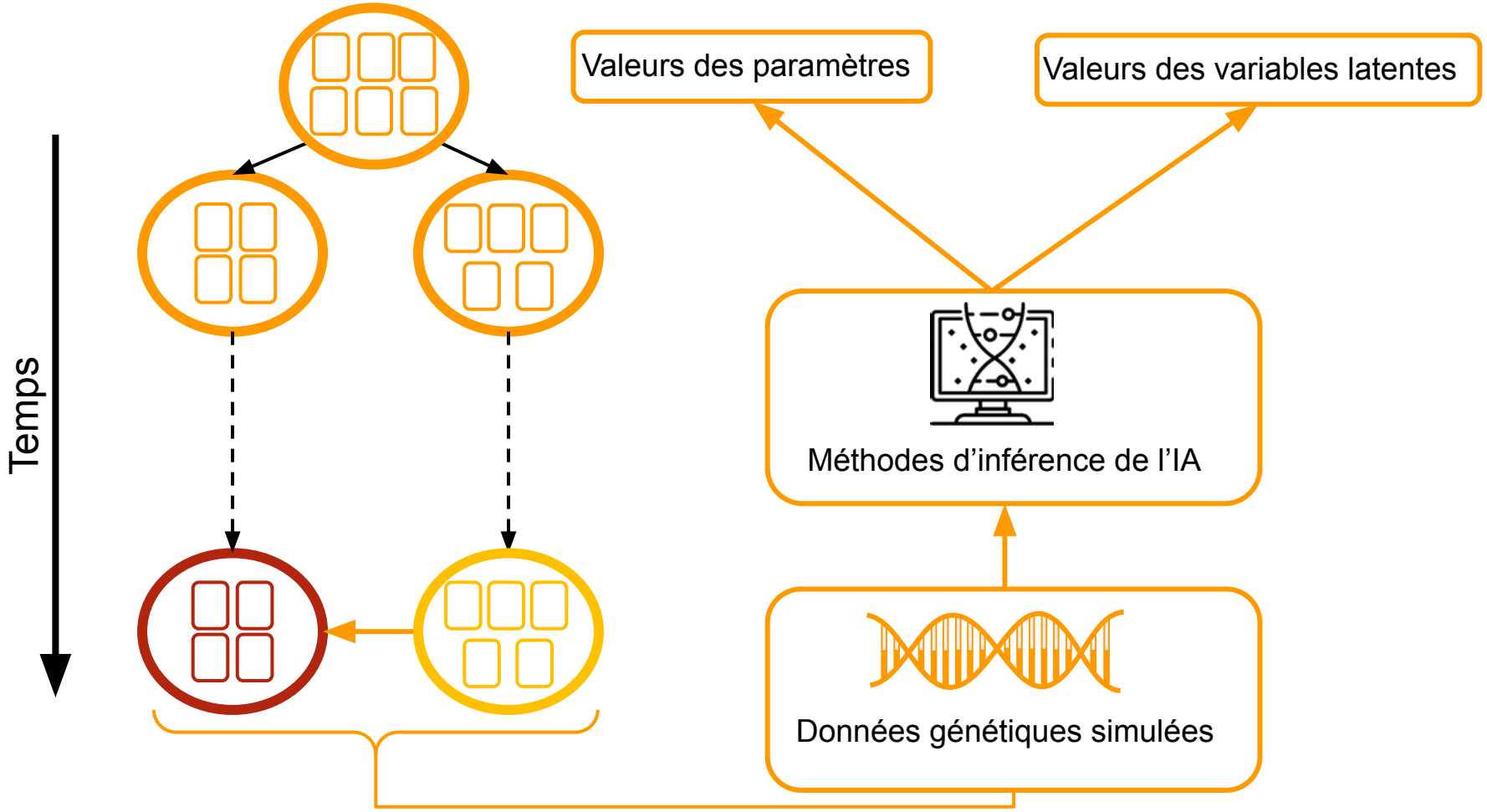
Cadre d’inférence :



Cadre d’inférence :

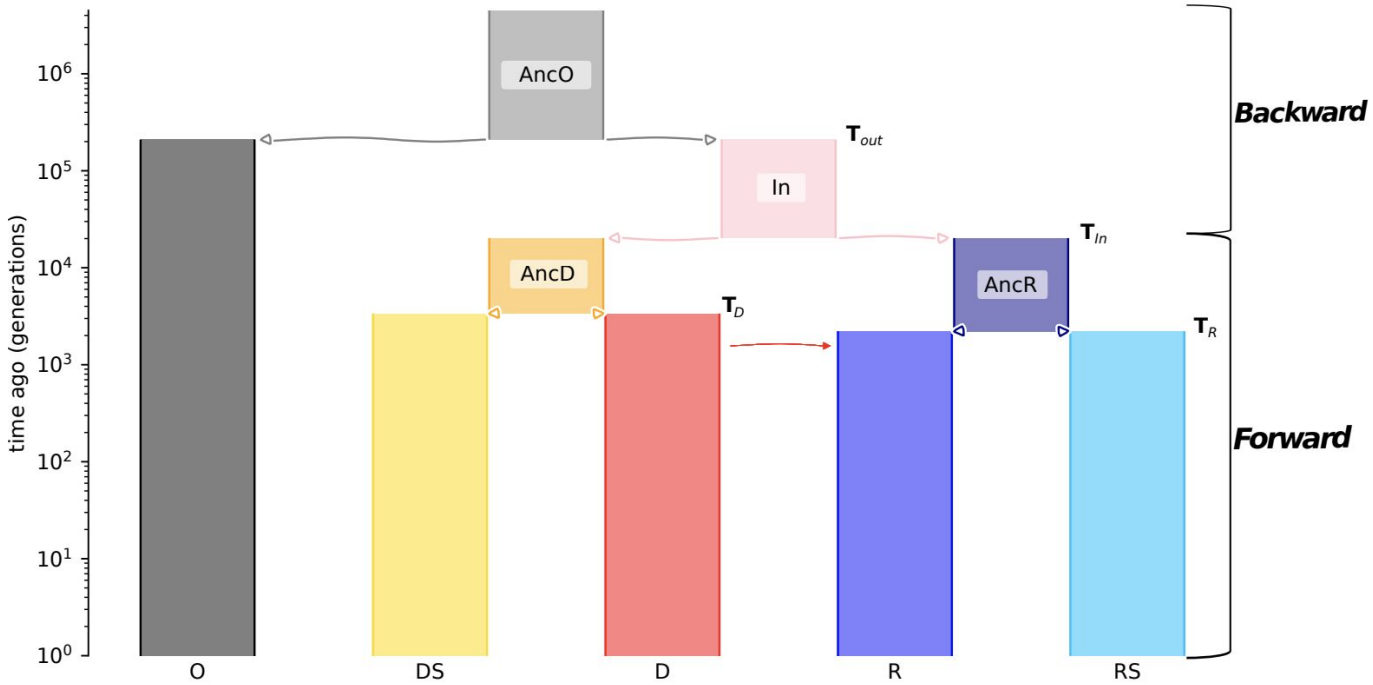


Cadre d’inférence :



Modèle démo-génétique général

Modèle démographique général



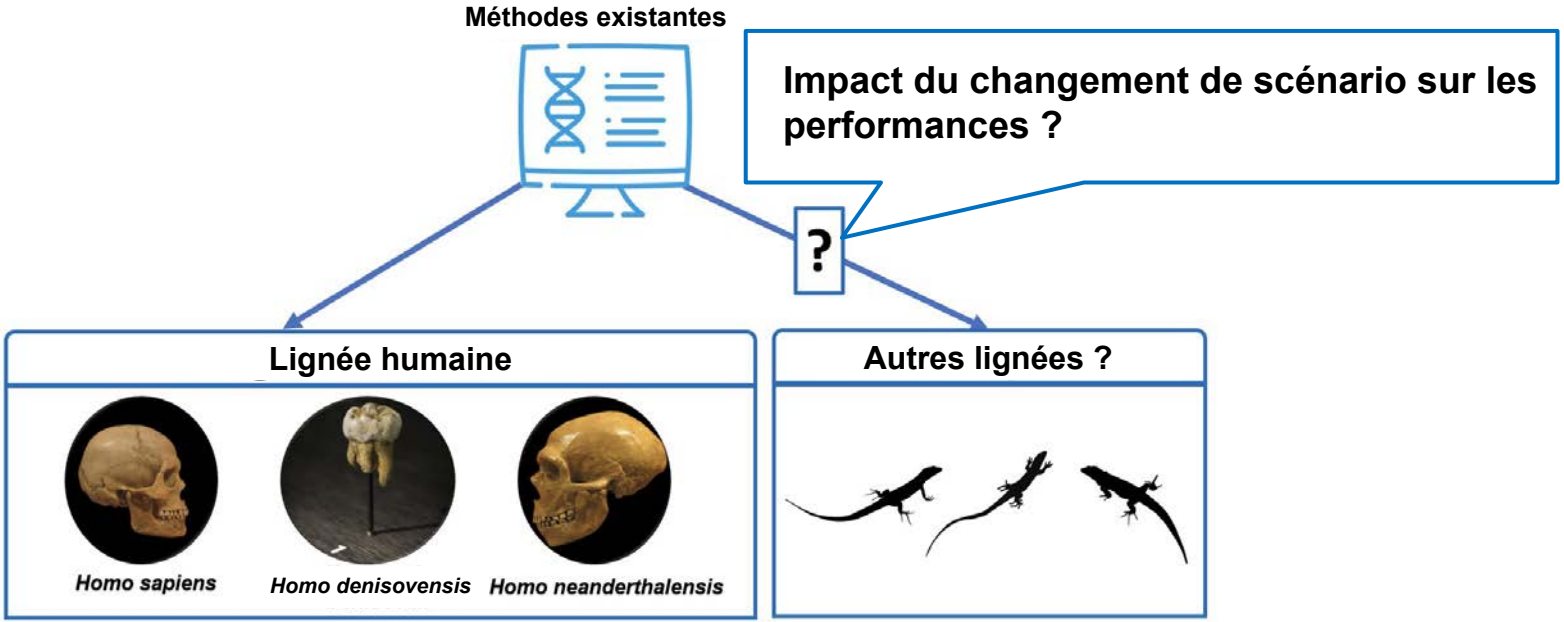
En résumé :

- Combinaison *backward*/*forward*
- 5 populations actuelles
- 1 Événement de migration (D ⇒ R)

Comparer les performances des méthodes d’inférence de l’IA existantes

Pourquoi évaluer les méthodes d'inférence existantes ?

1 ***Testées uniquement sur des données génétiques issues de scénario Humain***

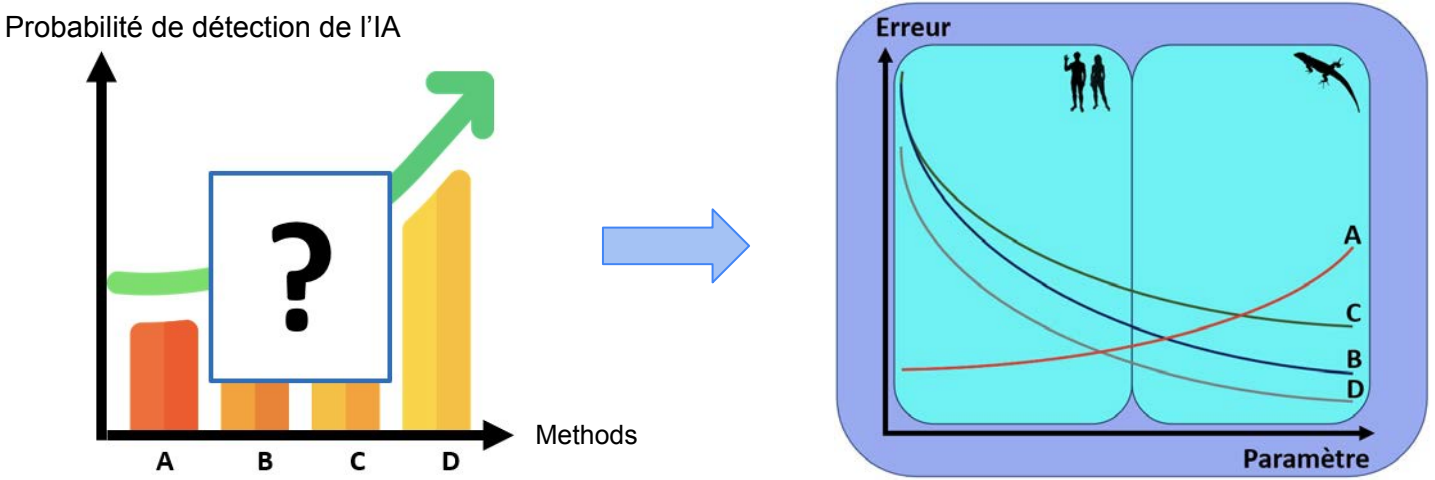


Pourquoi évaluer des méthodes existantes ?

1 ***Testées uniquement sur des données génétiques issues de scénario Humain***

2 ***Peu d'études comparent les performances des méthodes entre elles***

↳ *Identifier les forces et les faiblesses de chaque méthode*



Méthodes d'inférence de l'IA testées

Q95
(Racimo *et al.*, 2017)

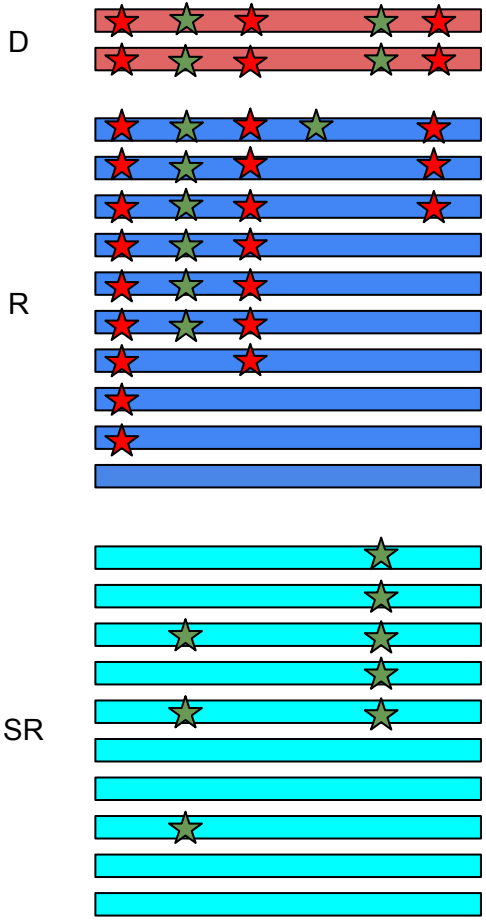
Statistique résumante

Fenêtres

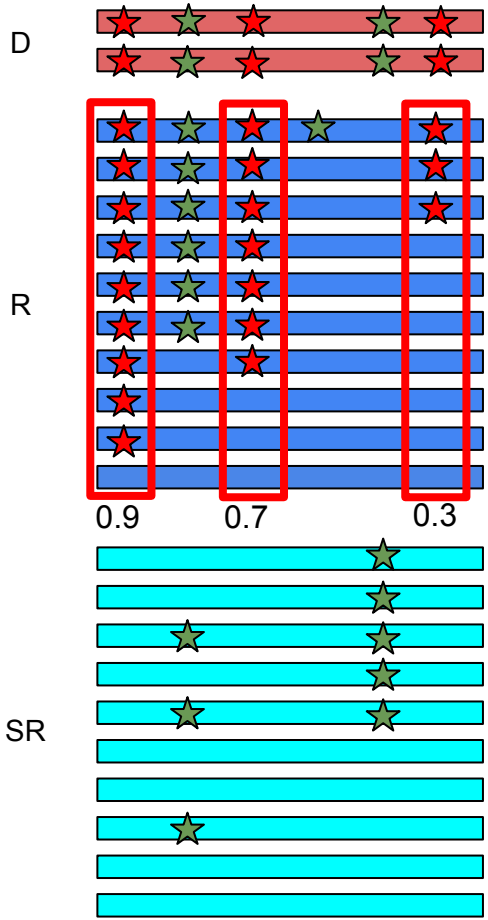
Hautes fréquences des allèles partagés uniquement entre D et R

Utilise les échantillons de 3 populations

Exemple de statistique résumante : le Q95(1%,100%)

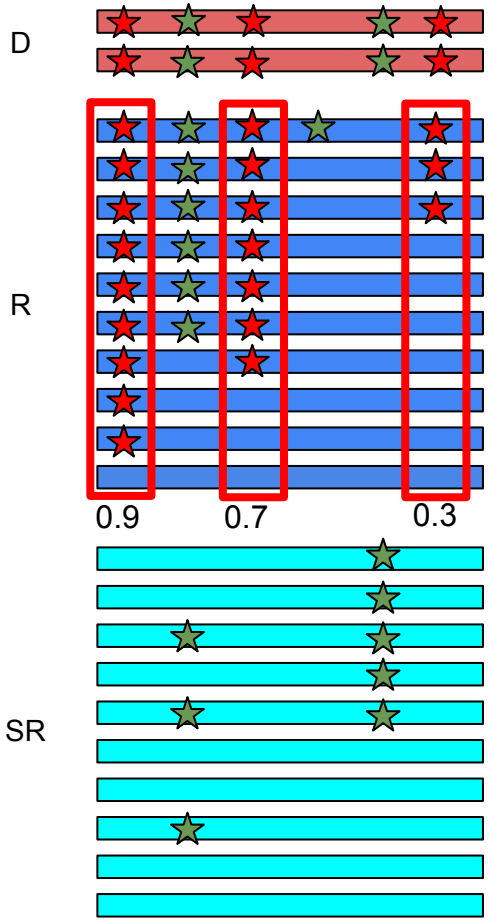


Exemple de statistique résumante : le Q95(1%,100%)



Allèles dérivés dans R : (1) 100% dans D (2) < ou = 1% dans SR

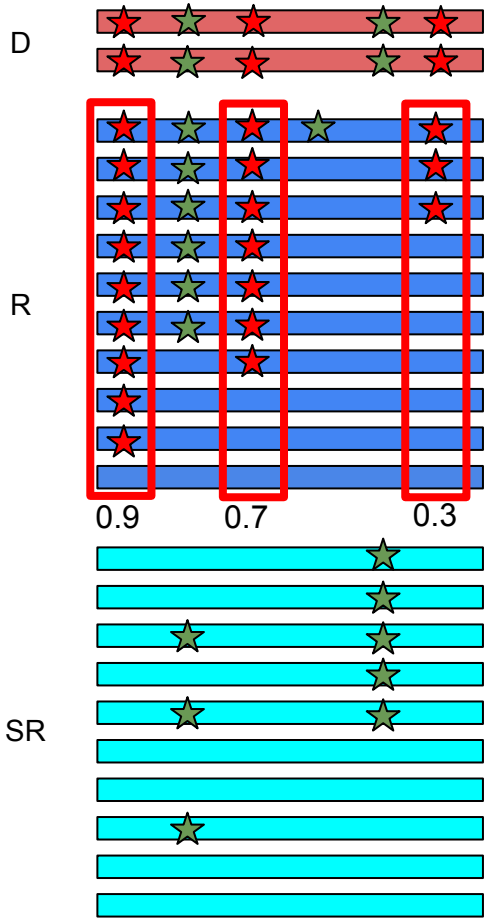
Exemple de statistique résumante : le Q95(1%,100%)



Allèles dérivés dans R : (1) 100% dans D (2) < ou = 1% dans SR

Calculer le quantile à 95% de la distribution de ces fréquences d'allèle partagés entre D et R : **quantile à 95%(0.3, 0.7, 0.9) = 0.88**

Exemple de statistique résumante : le Q95(1%,100%)



Allèles dérivés dans R : (1) 100% dans D (2) < ou = 1% dans SR

Calculer le quantile à 95% de la distribution de ces fréquences d'allèle partagés entre D et R : **quantile à 95%(0.3, 0.7, 0.9) = 0.88**

Q95(1%,100%) = 0.88

Philosophie :

Régions sous IA : présence de sites avec des allèles partagés entre D et R à fortes fréquences, mais à de très faibles fréquences, voir non présents dans SR

Méthodes d'inférence de l'IA testées

Q95
(Racimo *et al.*, 2017)

VolcanoFinder
(Setter *et al.*, 2020)

Statistique résumante

Fenêtres

Hautes fréquences des allèles partagés uniquement entre D et R

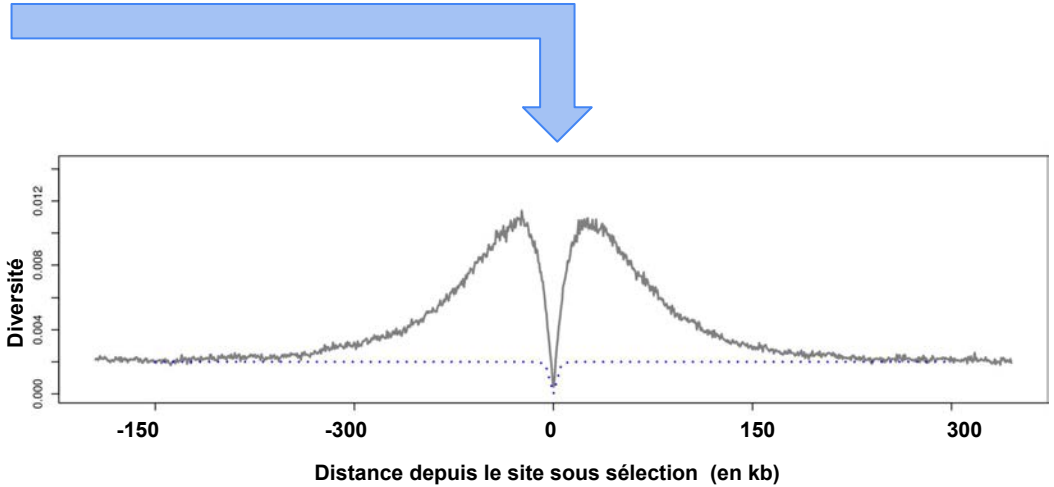
Utilise les échantillons de 3 populations

Basée sur un CLR

Sites

Excès d'allèles à des fréquences intermédiaires

Utilise uniquement les échantillons de la population R

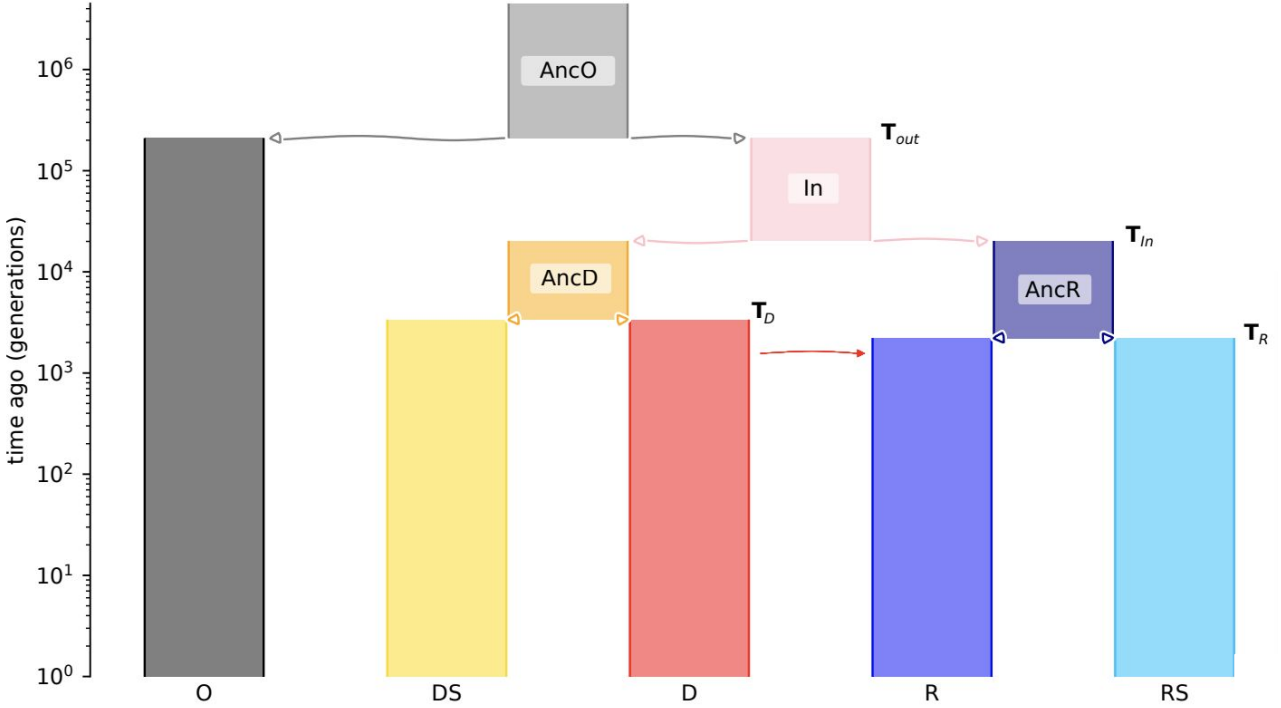


Méthodes d'inférence de l'IA testées

Méthodes de classification basées sur des simulations

Q95 (Racimo *et al.*, 2017)	**VolcanoFinder** (Setter *et al.*, 2020)	**genomatnn** (Gower *et al.*, 2021)	**MaLAdapt** (Zhang *et al.*, 2023)
Statistique résumante	**Basée sur un CLR**	***Machine learning*** (CNN)	***Machine learning*** (Extra Trees Classifier)
Fenêtres	Sites	Fenêtres	Fenêtres
Hautes fréquences des allèles partagés uniquement entre D et R	Excès d'allèles à des fréquences intermédiaires	Boite noire (entraînée sur les matrices génotypiques)	Entraînée sur un ensemble de statistiques résumantes
Utilise les échantillons de 3 populations	Utilise uniquement les échantillons de la population R	Utilise les échantillons de 3 populations	Utilise les échantillons de 3 populations

Scénarios démographiques tests :



Temps en génération

T_{out}	209 891
T_{In}	20 225
T_D	2 218
T_R	3 375
T_m	1 566

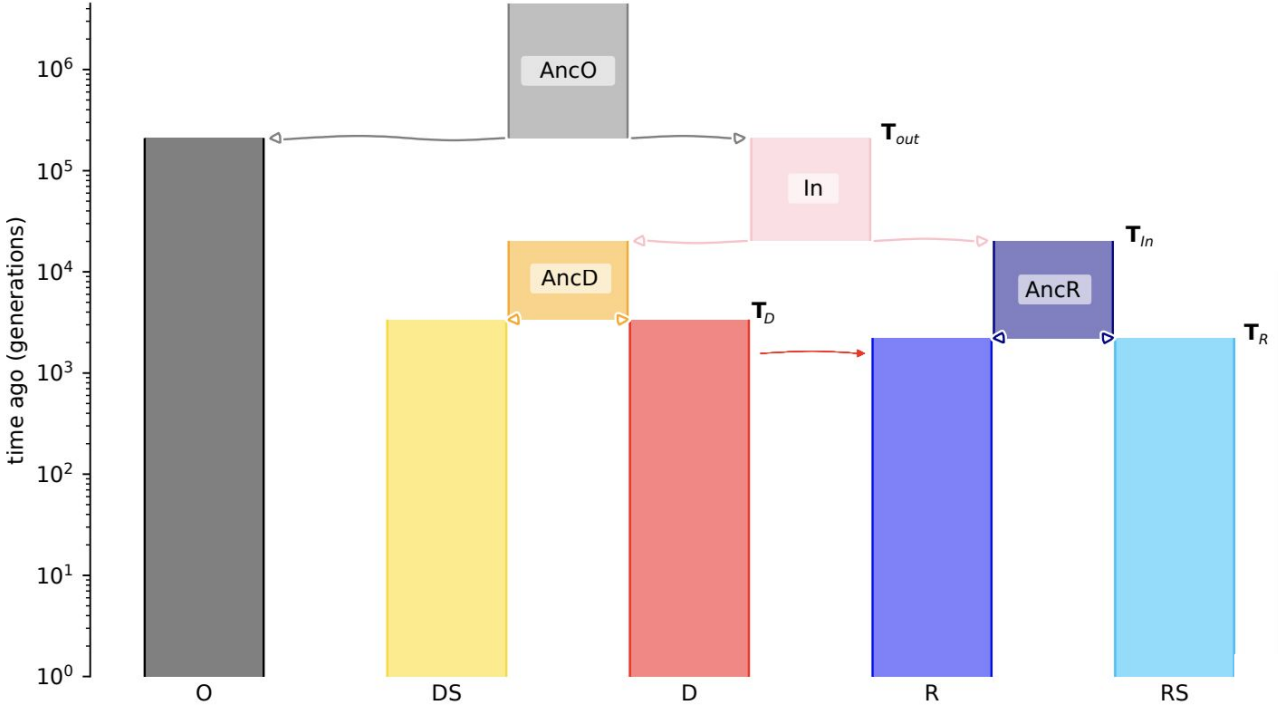
Pan troglodytes

H. neanderthalensis

H. sapiens

Référence

Scénarios démographiques tests :



Temps en génération

		
T_{out}	209 891	340 000
T_{In}	20 225	54 000
T_D	2 218	34 000
T_R	3 375	20 000
T_m	1 566	15 000

Pan troglodytes

U. americanus

H. neanderthalensis

U.maritimus

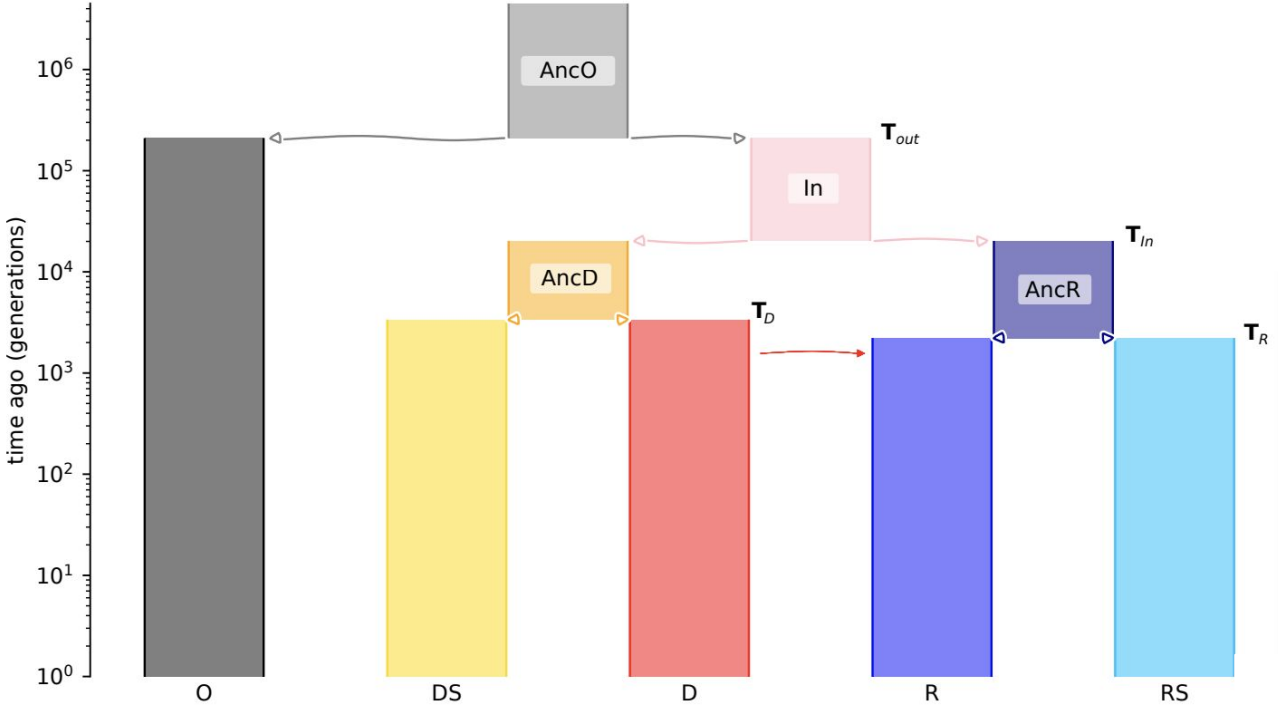
H. sapiens

U.arctos

Référence

Temps de migration ancien

Scénarios démographiques tests :



Temps en génération

T_{out}	209 891	340 000	4 000 000
T_{In}	20 225	54 000	2 400 000
T_D	2 218	34 000	4 000
T_R	3 375	20 000	40 000
T_m	1 566	15 000	2 000

Pan troglodytes

U. americanus

P.murallis

H. neanderthalensis

U.maritimus

P.bocagei

H. sapiens

U.arctos

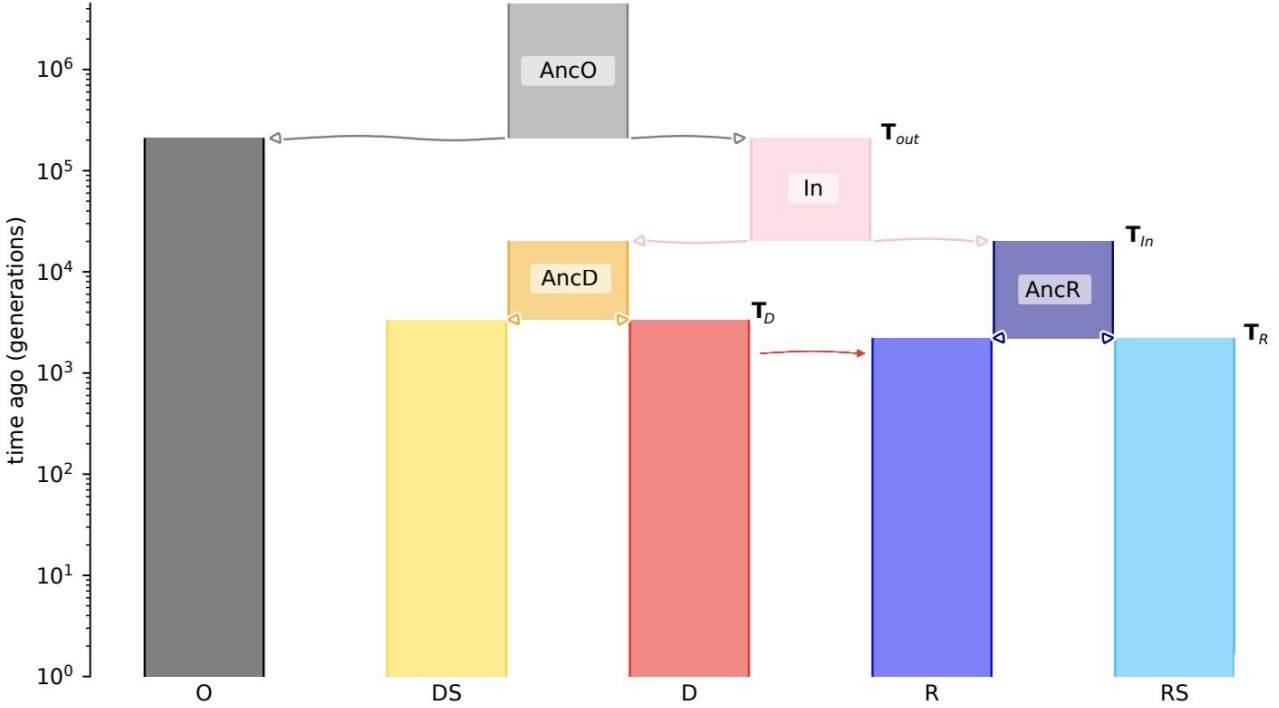
P.carbonelli

Référence

Temps de migration ancien

Temps de divergence ancien

Scénarios démographiques tests :



Objectif 1 :

Test de l'impact des variations de différents paramètres :

- Force de la sélection (*s*)
- Force de la migration (*m*)
- Tailles des populations (*N*)
- Présence de hotspots de recombinaison (*r*)
- Taille des échantillons

Pan troglodytes

U. americanus

P.murallis

H. neanderthalensis

U.maritimus

P.bocagei

H. sapiens

U.arctos

P.carbonelli

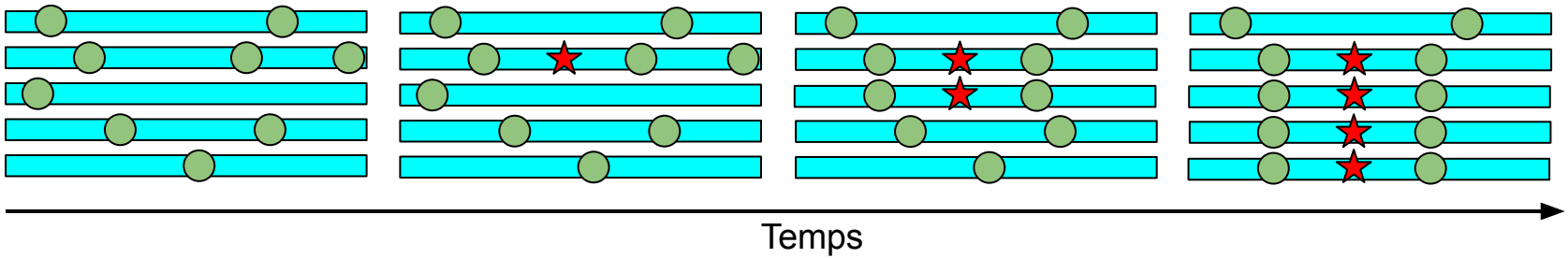
Référence

Temps de migration ancien

Temps de divergence ancien

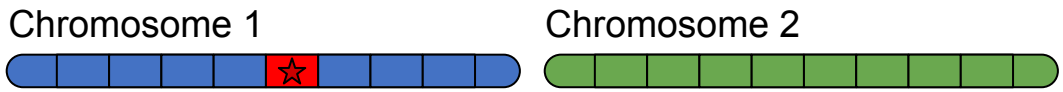
Impact de l’auto-stop sur les inférences :

Exemple de l’effet de auto-stop :



Impact de l'auto-stop sur les inférences :

Structure du génome :



Fenêtre de 50kb

Fenêtre d'introgression adaptative

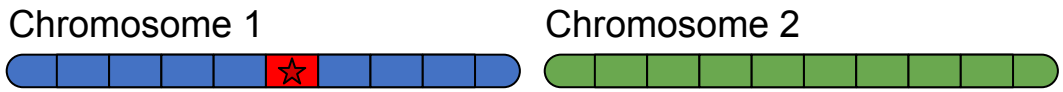
Fenêtre adjacente

Fenêtre du chromosome 2

Mutation avantageuse introgressée

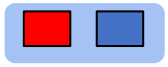
Impact de l’auto-stop sur les inférences :

Structure du génome :

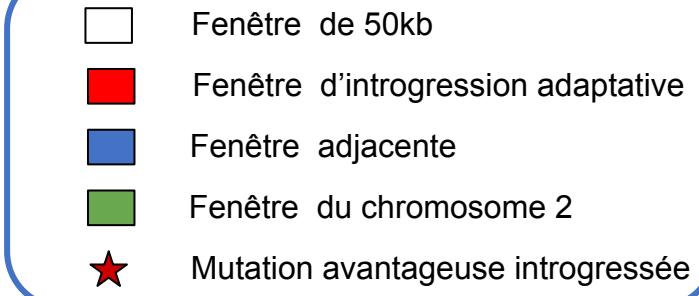


Jeux de données d'entraînement des méthodes :

Adjacente
(style MaLAdapt)



Second chromosome
(style genomatnn)



Impact de l'auto-stop sur les inférences :

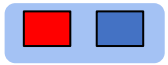
Structure du génome :

Chromosome 1

Chromosome 2

Jeux de données d'entraînement des méthodes :

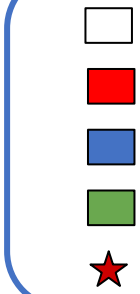
Adjacente
(style MaLAdapt)



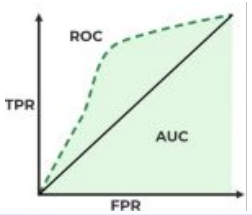
Second chromosome
(style genomatnn)

Objectif 2 :

Effet sur les performances de jeux de données tests avec différents types de fenêtres ?



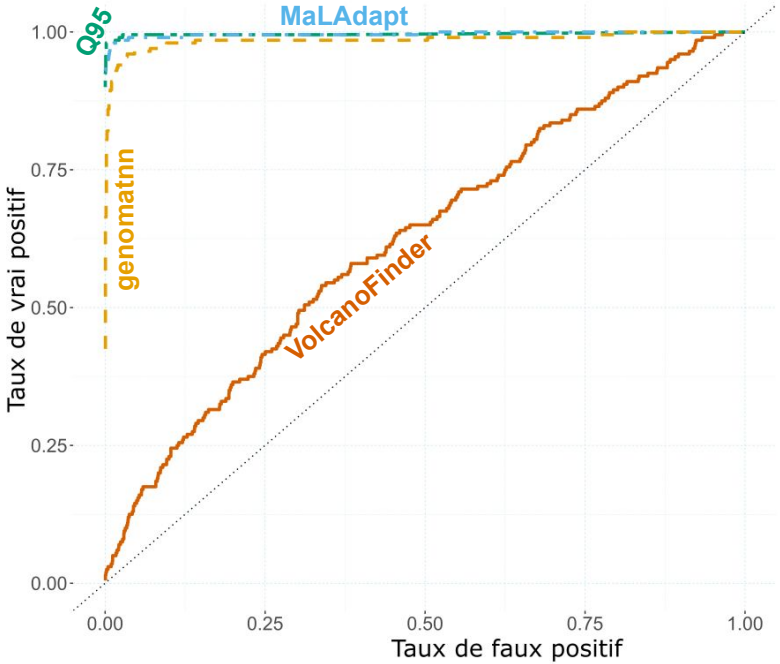
Résumé de l'approche de test par simulation :

Simulation	Ensemble de données test	Méthode de classification	Métriques de classification
Backward + *Forward* Modèle mutationnel avantageux : 	Données génétiques simulées Variables latentes Classes connues : Fenêtres non-IA Fenêtres IA	• **Q95** (statistique résumante) • **VolcanoFinder** (composite likelihood) • **genomatnn** (Deep learning) • **MaLAdapt** (Machine learning)	Classes connues contre classes prédites 

Résultats : Impact du type de fenêtres

Humain de référence

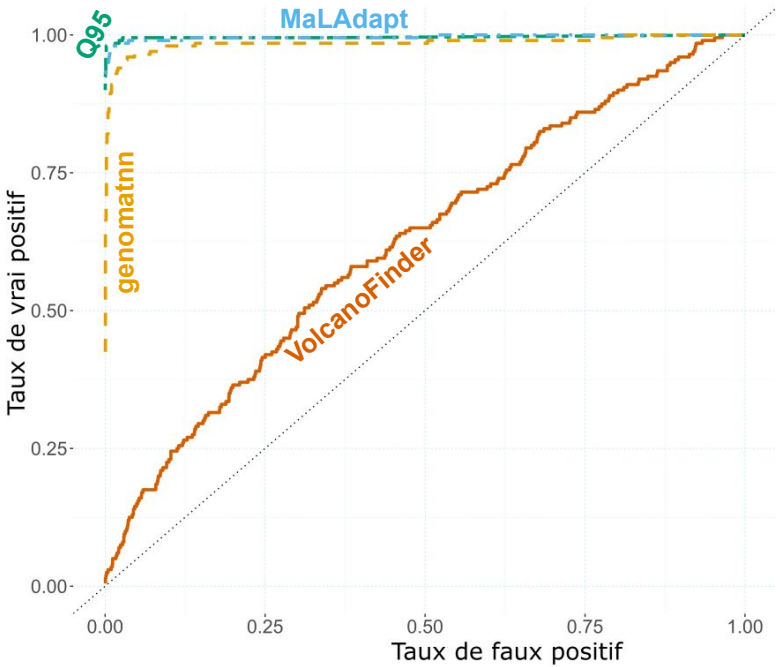
Fenêtres IA vs second chromosome



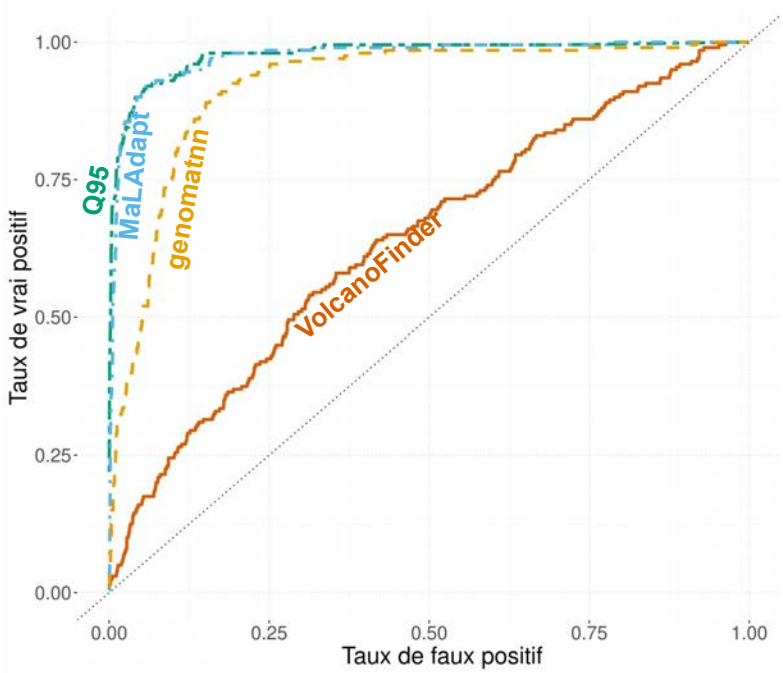
Résultats : Impact du type de fenêtres

Humain de référence

Fenêtres IA vs second chromosome



Fenêtres IA vs Adjacentes

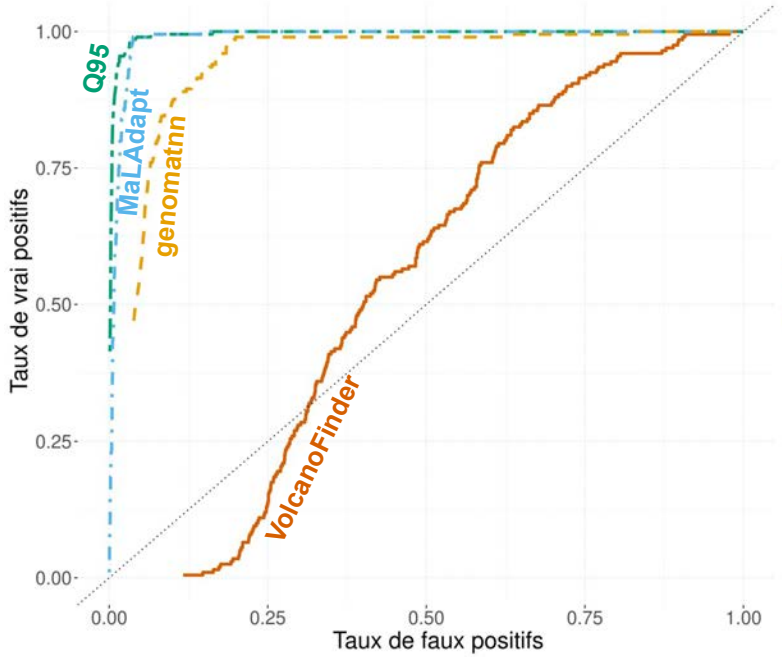


- Effet de l'auto-stop : Baisse des performances (toutes les méthodes sauf VolcanoFinder)
- Effet de la misspecification : Baisse des performances (genomatnn)

Résultats : Impact de la force de la sélection

Podarcis : fenêtres IA vs Adjacentes

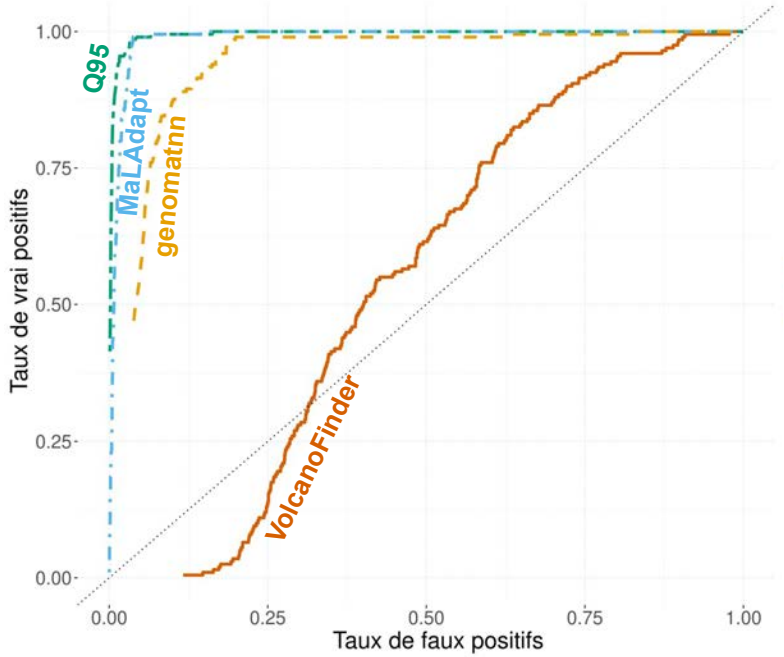
s intermédiaire (0.01)



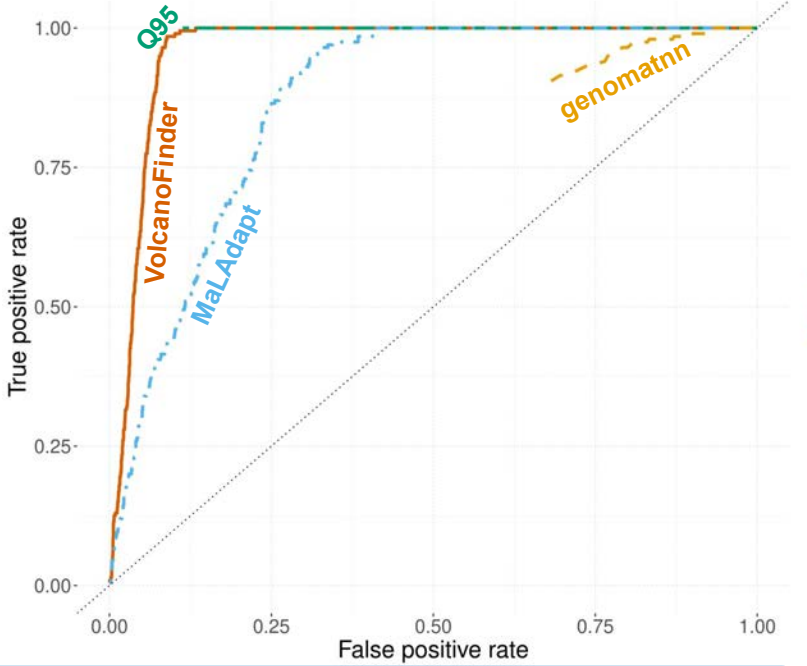
Résultats : Impact de la force de la sélection

Podarcis : fenêtres IA vs Adjacentes

s intermédiaire (0.01)



s fort (0.1)



Effet de l'autostop :

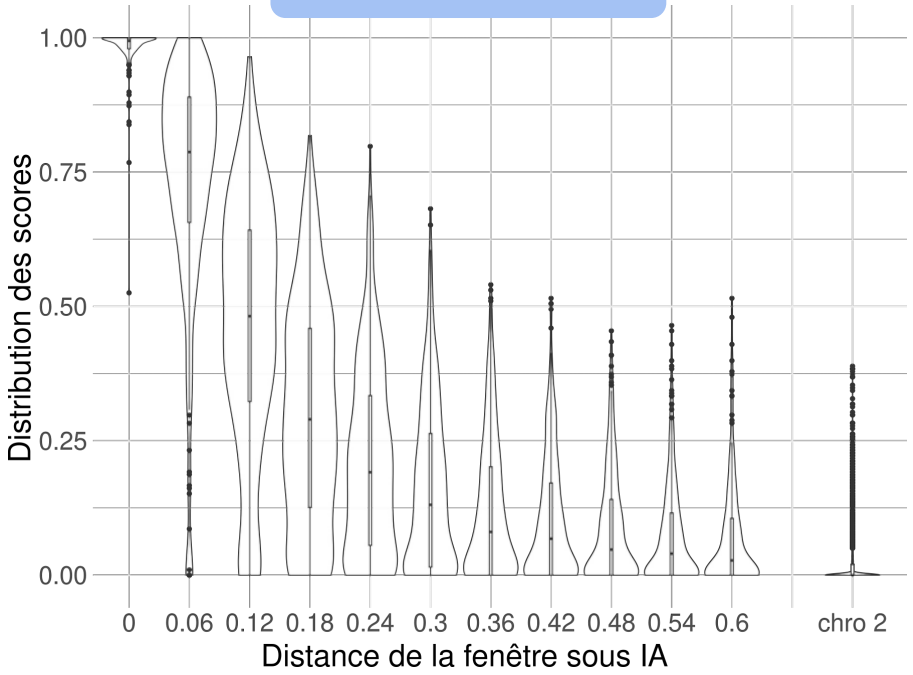
- Baisse des performances de MaLAdapt et genomatnn dû à *l'autostop*
- Augmentation des performances pour VolcanoFinder

Résultats : Impact de la force de la sélection

Podarcis : Distributions des scores avec la distance à la fenêtre sous IA

Q95

s intermédiaire (0.01)

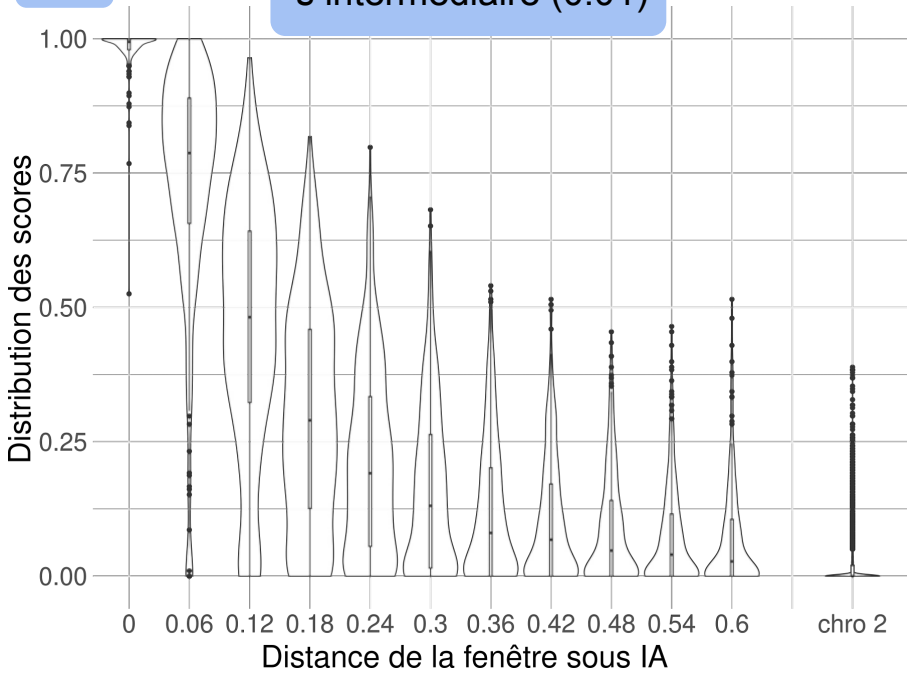


Résultats : Impact de la force de la sélection

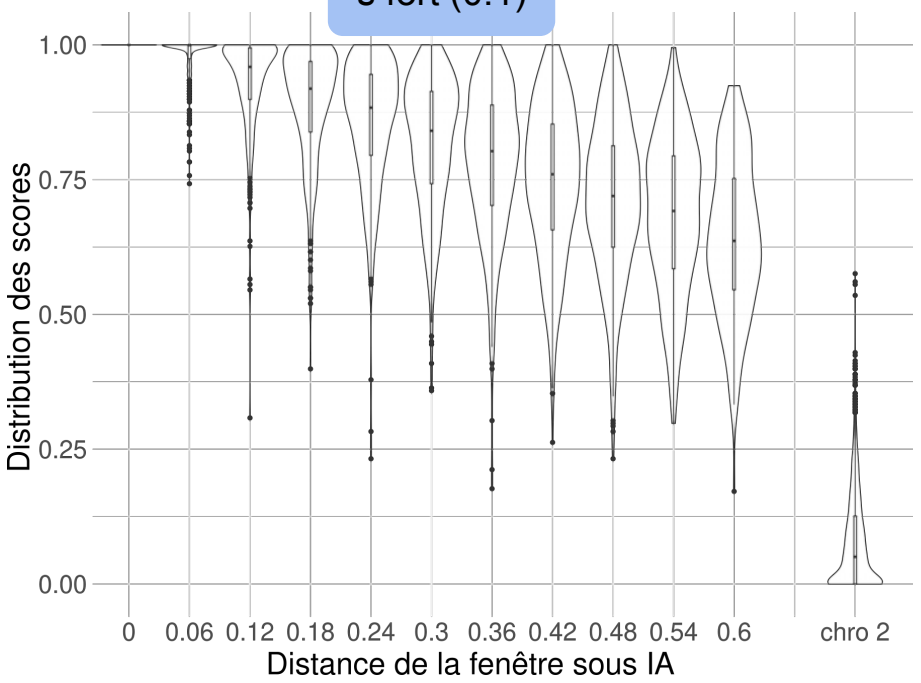
Podarcis : Distributions des scores avec la distance à la fenêtre sous IA

Q95

s intermédiaire (0.01)



s fort (0.1)



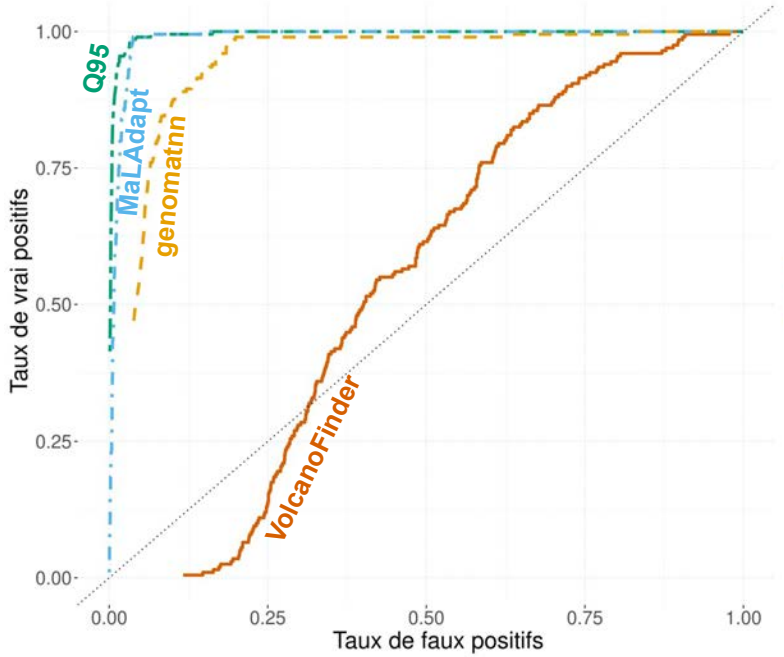
Effet de l'autostop :

- Difficulté de différencier les distributions des scores des fenêtres IA/non-IA quand *s* fort
- *s* fort : valeurs hautes même pour les fenêtres éloignées

Résultats : Impact de la force de la sélection

Podarcis : fenêtres IA vs Adjacentes

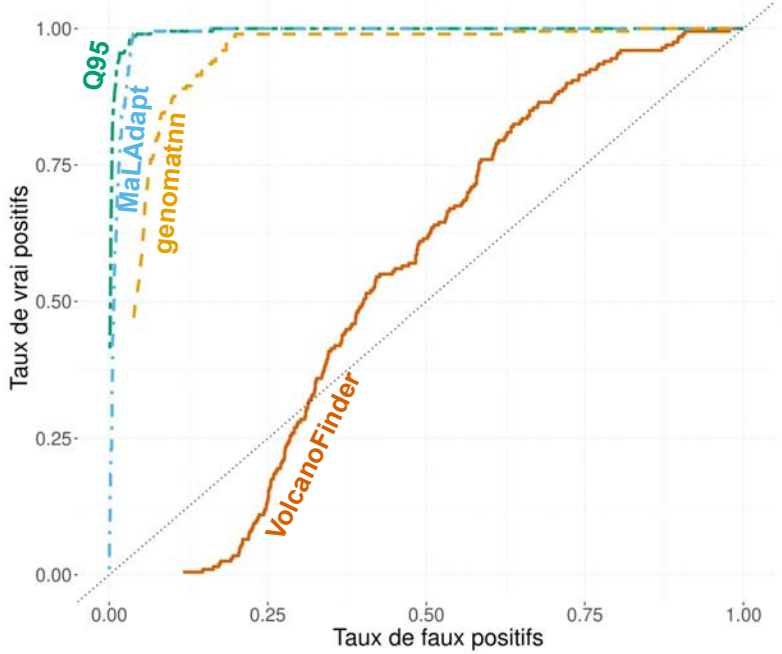
s intermédiaire (0.01)



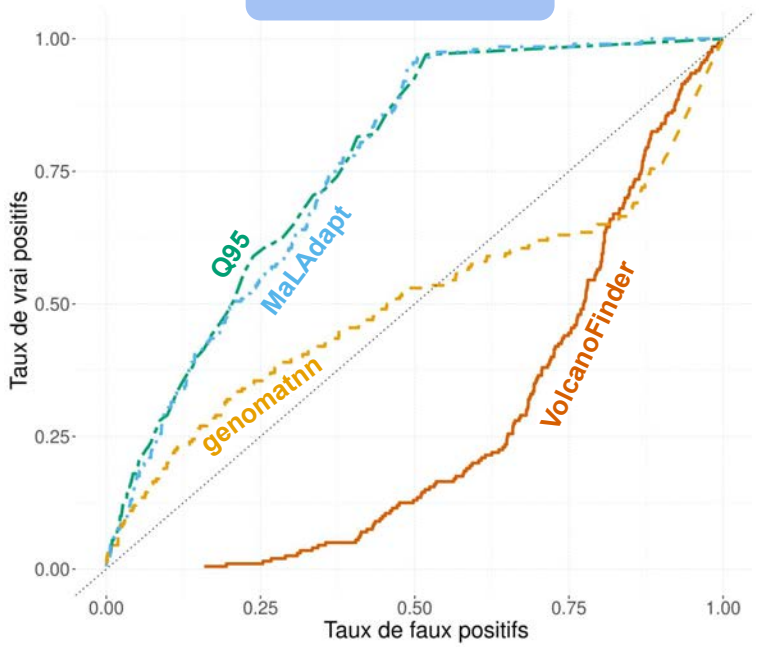
Résultats : Impact de la force de la sélection

Podarcis : fenêtres IA vs Adjacentes

s intermédiaire (0.01)



s faible (0.001)



Signal de sélection faible : Baisse des performances de toutes les méthodes

Résultats et discussions :

- Méthode avec les meilleures performances pour nos tests :

Q95
(Racimo et al., 2017)

Résultats et discussions :

- Méthode avec les meilleures performances pour nos tests :

Q95
(Racimo et al., 2017)

- Sélection forte : Baisse des performances dû à l'auto-stop (sauf pour VolcanoFinder) illustre l'importance de **prendre en compte les fenêtres adjacentes**

- **Pire performance** : Sélection faible et petites tailles de populations
- Légère baisse des performances : présence de *hostpot* et migration ancienne (sauf genomatnn)
- Légère augmentation des performances : Divergence ancienne et haut taux de migration (sauf genomatnn)

Résultats et discussions :

- Méthode avec les meilleures performances pour nos tests :

Q95
(Racimo et al., 2017)

- Sélection forte : Baisse des performances dû à l'auto-stop (sauf pour VolcanoFinder) illustre l'importance de **prendre en compte les fenêtres adjacentes**

- **Pire performance** : Sélection faible et petites tailles de populations
- Légère baisse des performances : présence de *hotspot* et migration ancienne (sauf genomatnn)
- Légère augmentation des performances : Divergence ancienne et haut taux de migration (sauf genomatnn)

- Important de prendre en compte des fenêtres rendant compte du déséquilibre de liaison (style adjacente) dans l'entraînement des méthodes

Résultats et discussions :

- Méthode avec les meilleures performances pour nos tests :

Q95
(Racimo et al., 2017)

- Sélection forte : Baisse des performances dû à l'auto-stop (sauf pour VolcanoFinder) illustre l'importance de **prendre en compte les fenêtres adjacentes**

- **Pire performance** : Sélection faible et petites tailles de populations
- Légère baisse des performances : présence de *hostpot* et migration ancienne (sauf genomatnn)
- Légère augmentation des performances : Divergence ancienne et haut taux de migration (sauf genomatnn)

- Important de prendre en compte des fenêtres rendant compte du déséquilibre de liaison (style adjacente) dans l'entraînement des méthodes

- Méthodes de classification par simulation (genomatnn et MaLAdapt) : les mauvaises spécifications des modèles démo-génétiques ont un fort impact sur les performances

Nouvelle méthode d’inférence de la part de l’introgression qui est adaptative

Pourquoi développer une nouvelle méthode d'inférence ?

1. Toutes les méthodes existantes sont des méthodes de classification

   ↳ Problèmes inhérents aux méthodes de classification (déséquilibre entre les classes, nécessite d'établir un seuil au score etc)

   ↳ Difficulté de déduire directement une proportion d'IA en se basant sur les fenêtres inférées

   ↳ Identifie uniquement les fenêtres *outliers*

Pourquoi développer une nouvelle méthode d'inférence ?

1. Toutes les méthodes existantes sont des méthodes de classification

   ↳ Problèmes inhérents aux méthodes de classification (déséquilibre entre les classes, nécessite d'établir un seuil au score etc)

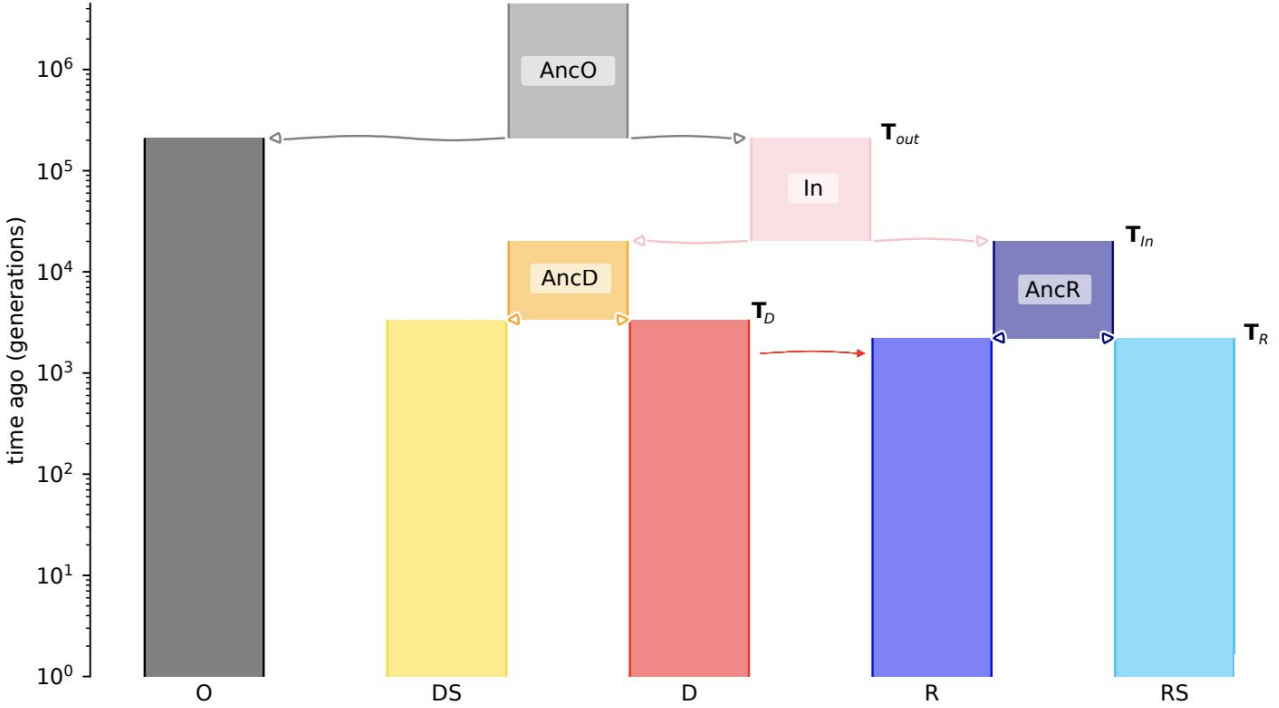
   ↳ Difficulté de déduire directement une proportion d'IA en se basant sur les fenêtres inférées

   ↳ Identifie uniquement les fenêtres *outliers*

2. Utiliser des méthodes d'estimation permet d'avoir des estimations de valeurs continues et de leurs intervalles de confiance (IC) associés

   ↳ **Objectifs : Estimation de la part du génome qui est liée à l'IA**

Scénarios démo-génétique tests



Pan troglodytes | *H. neanderthalensis* | *H. sapiens*

Temps en génération

T_{out}	209 891
T_{In}	20 225
T_D	2 218
T_R	3 375
T_m	1 566

Structure du génome

1 Chromosome

5 Mb

Approche d'inférence par simulation

ID des sites

Echantillons

	0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	A	A	C

Transformation données génétiques en statistiques résumantes

Approche d'inférence par simulation

ID des sites

Echantillons

	0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	A	A	C

Transformation données génétiques en statistiques résumantes

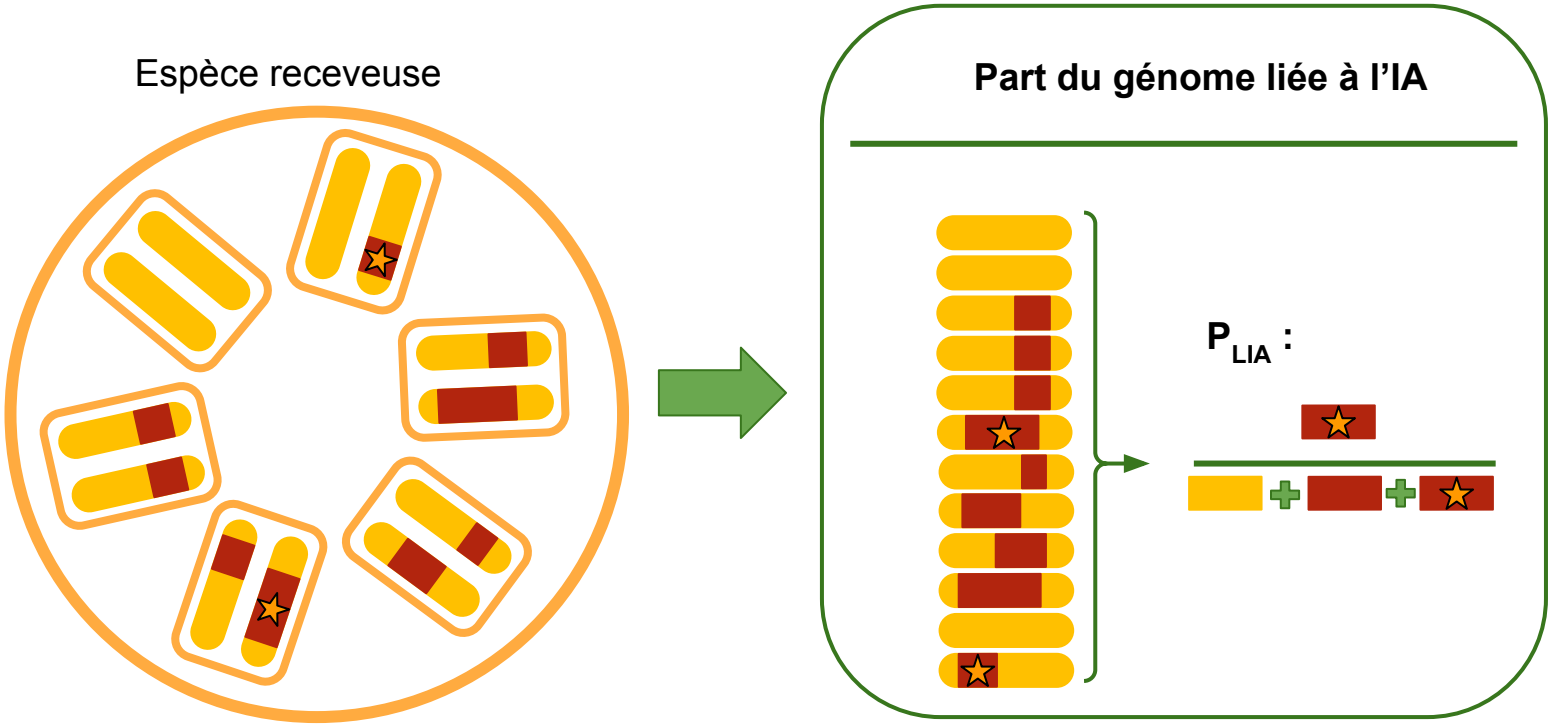
Capturent les signaux d'introgression, de sélection et d'introgression adaptative

Par fenêtre de 50kb

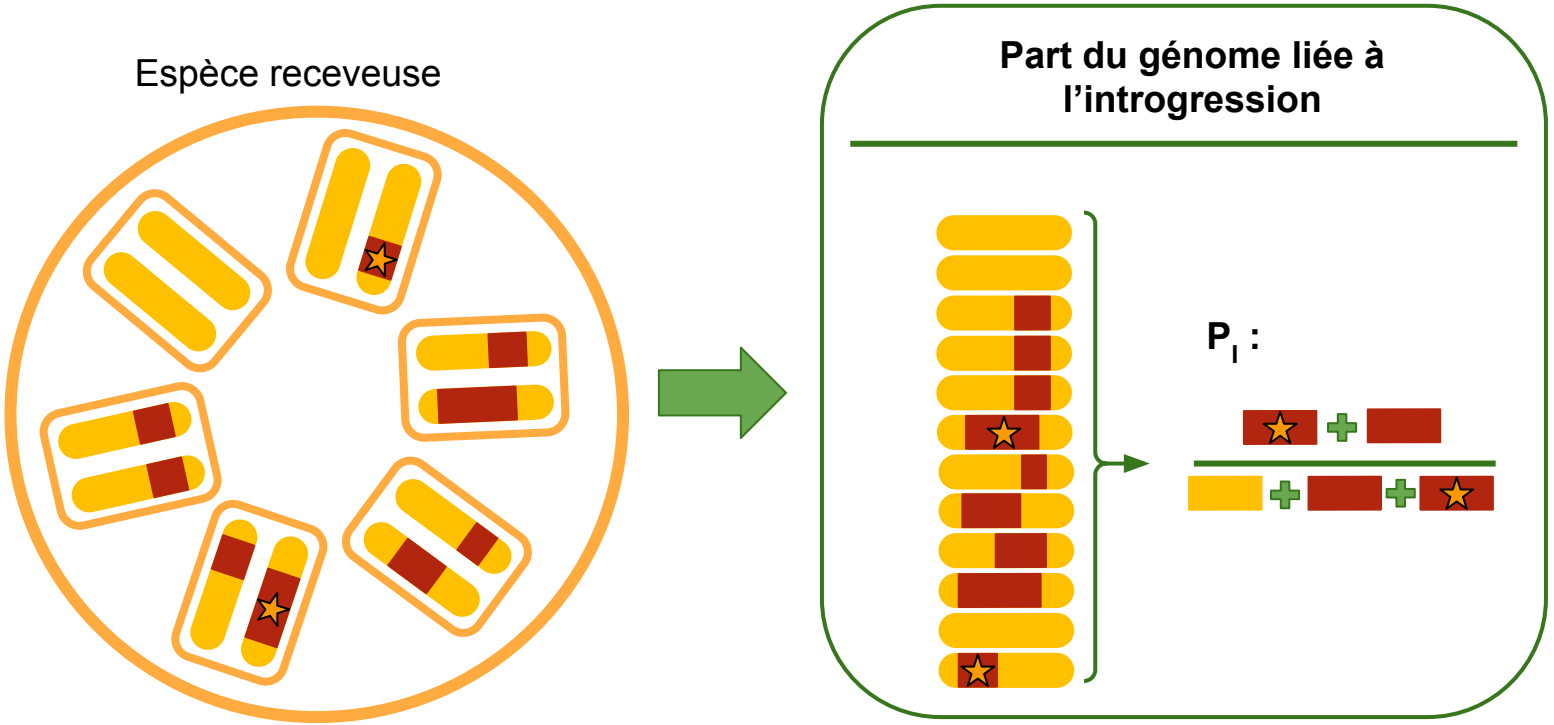
Résumé sur l'ensemble du génome

Moyenne et variance ainsi que d'autres métriques caractérisant leurs distributions

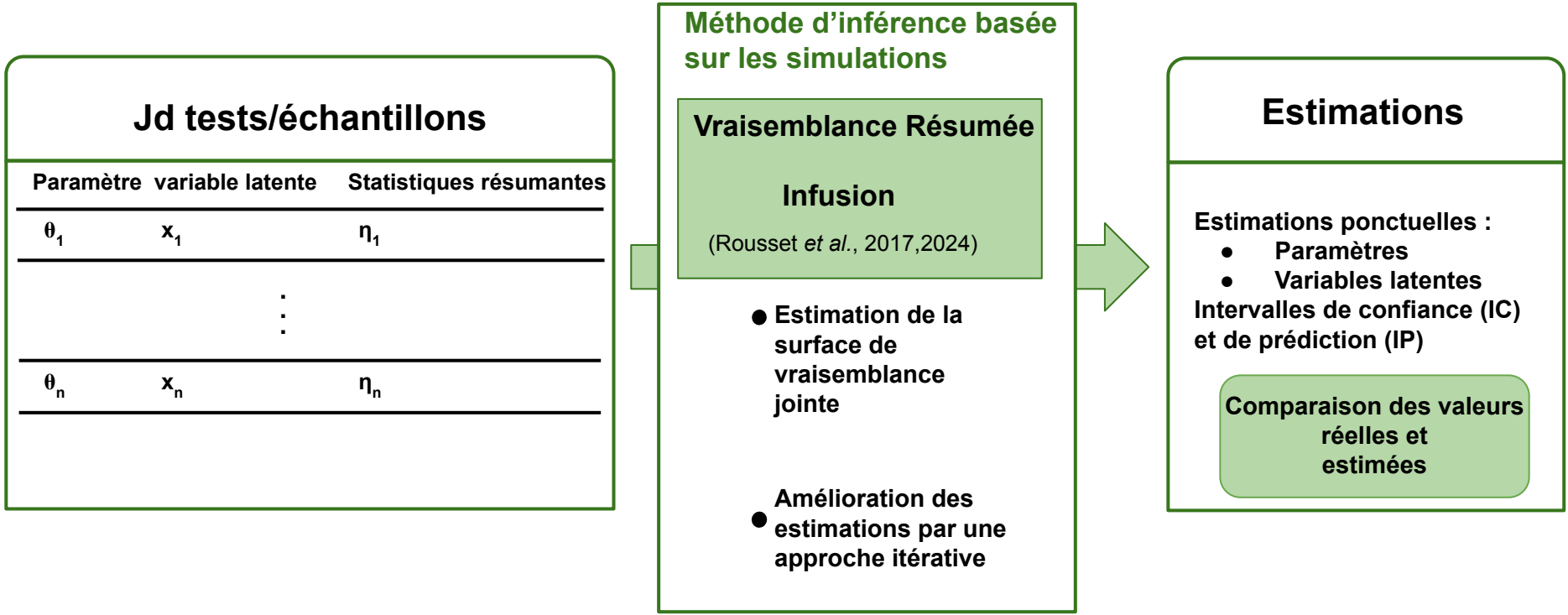
Les variables latentes : Part du génome liée à l'IA



Les variables latentes : Part du génome liée à l'introgression

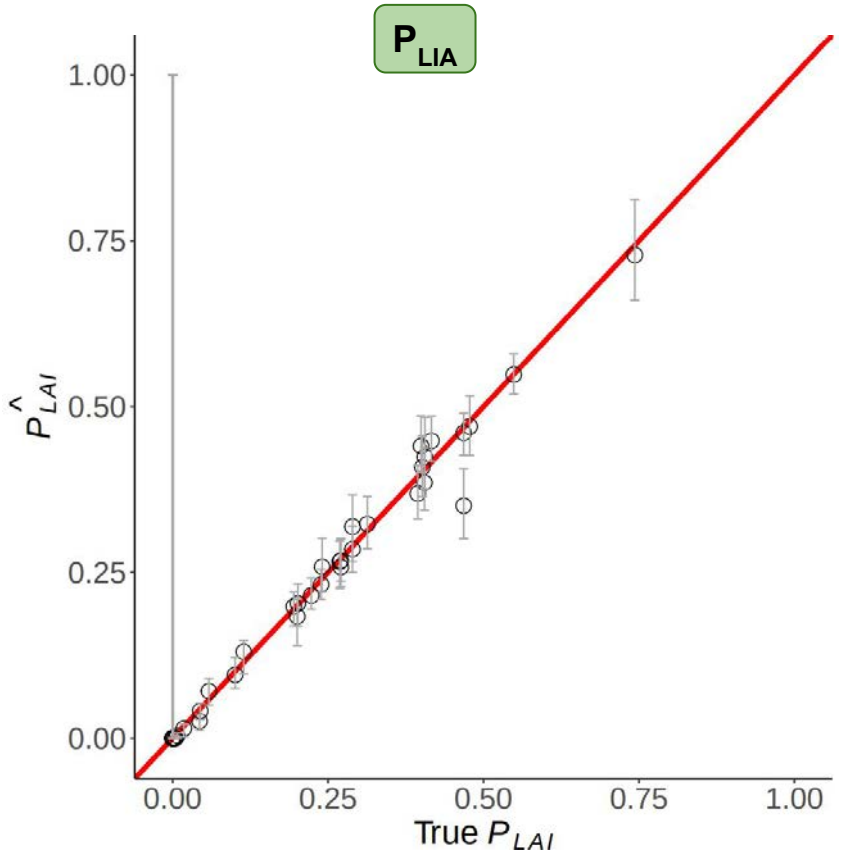


Approche d'inférence par simulation



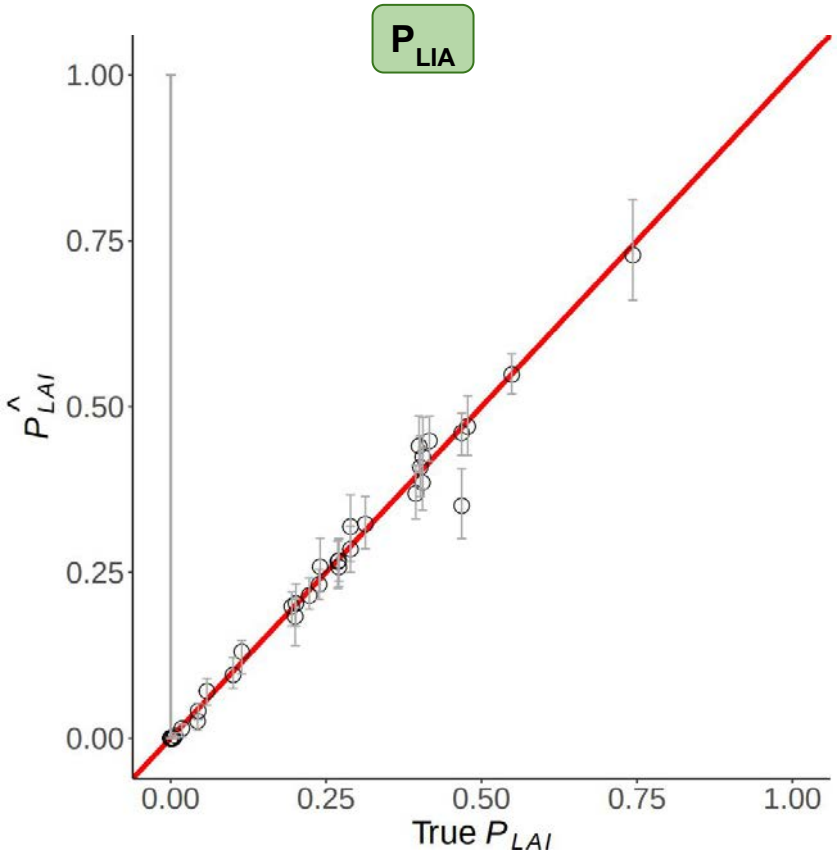
Résultats : Jeux de données avec *s* et *m* variables

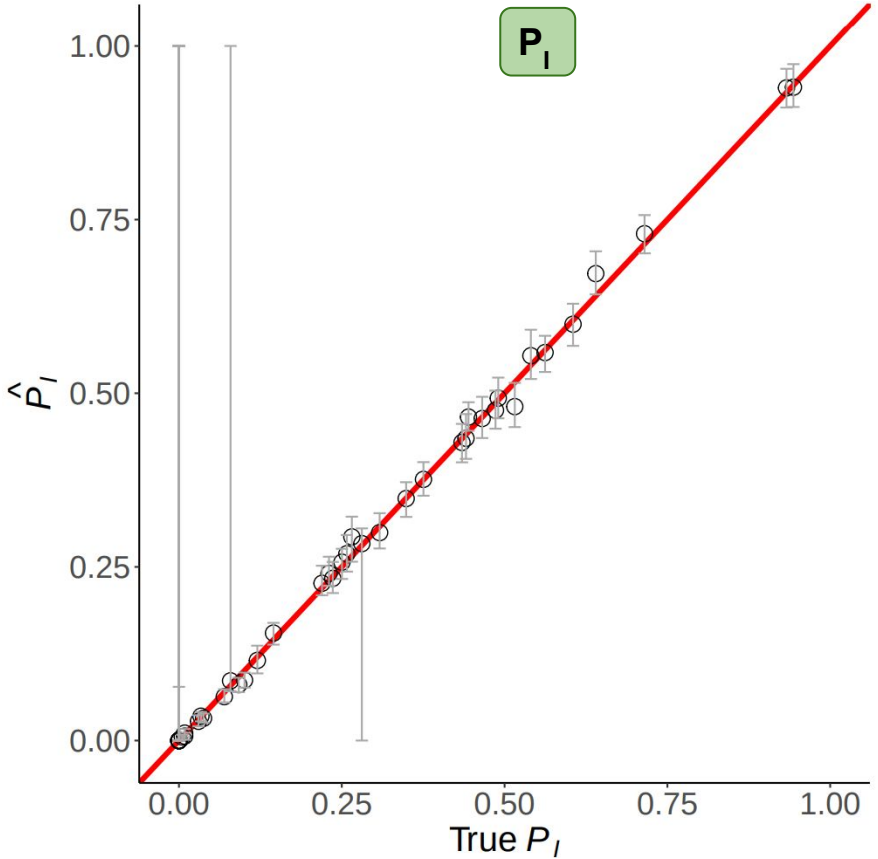
Estimation des variables latentes



Résultats : Jeux de données avec *s* et *m* variables

Estimation des variables latentes

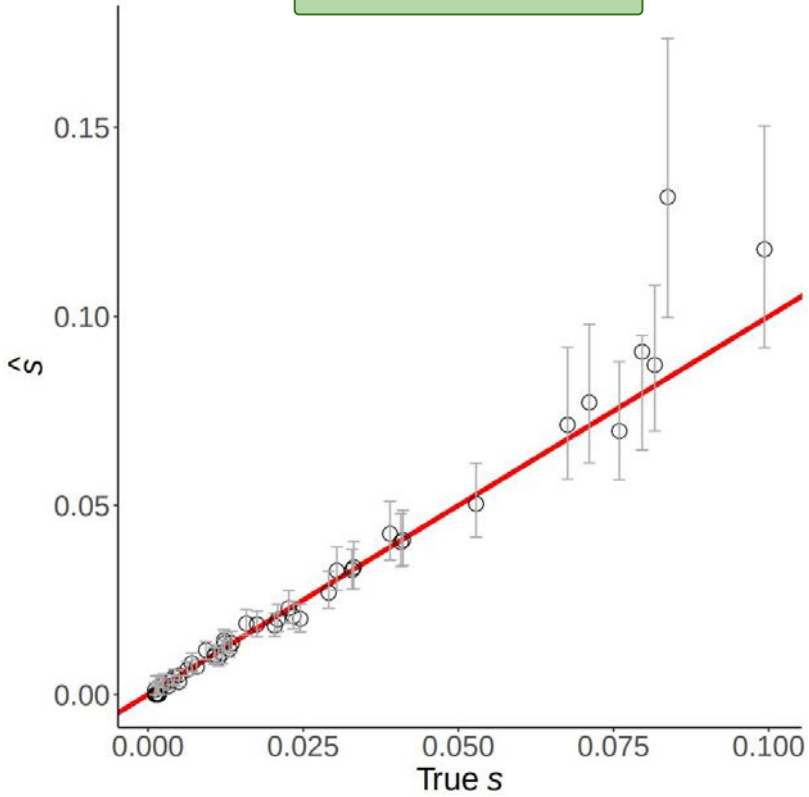




Résultats : Jeux de données avec *s* et *m* variables

Estimation des paramètres

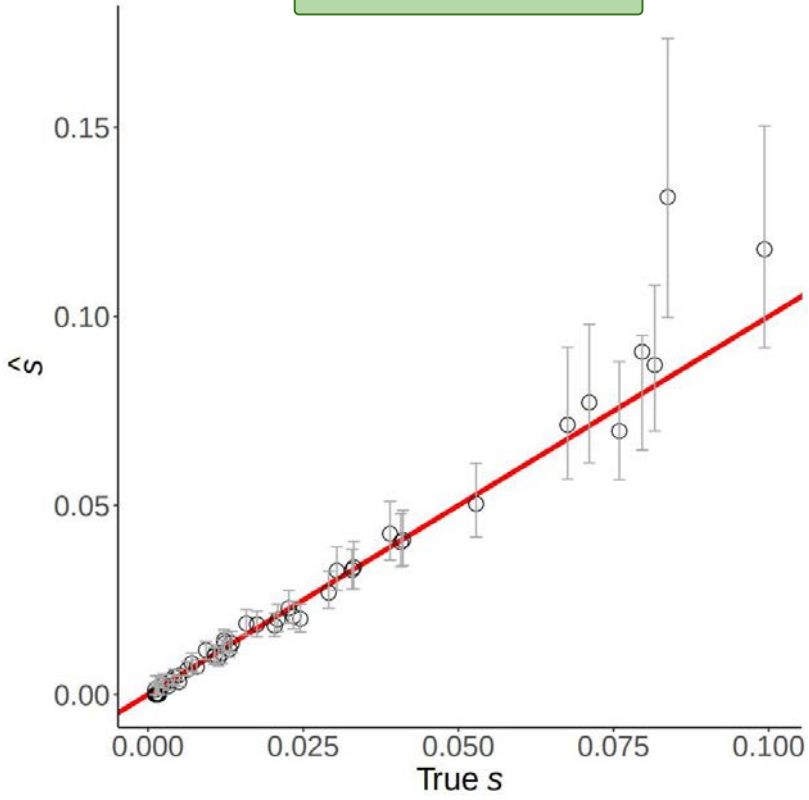
Coefficient de sélection



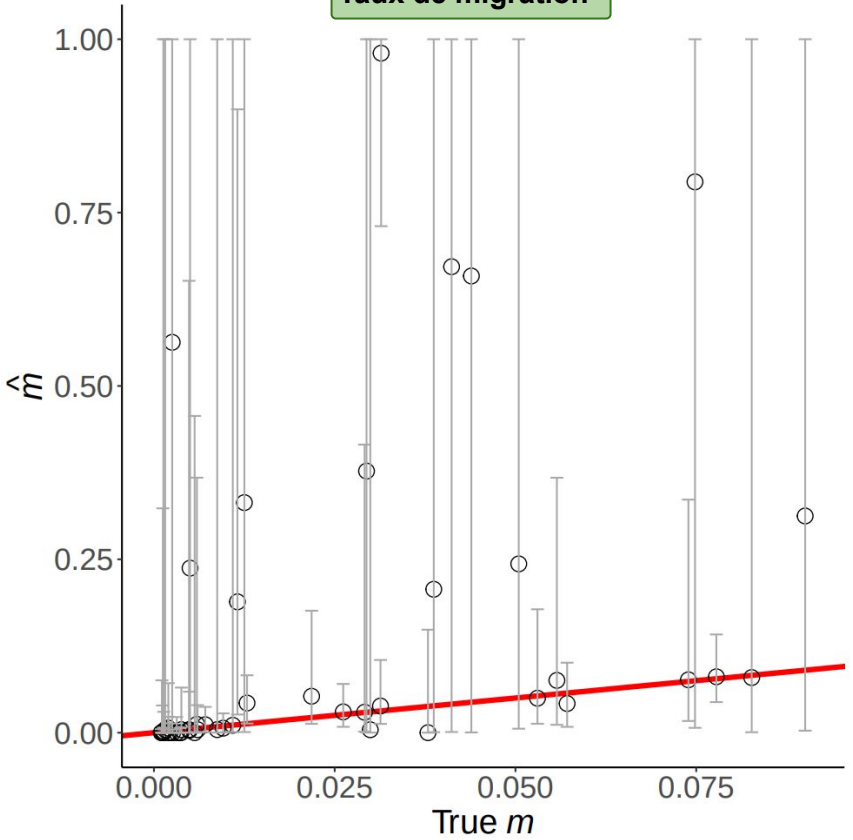
Résultats : Jeux de données avec *s* et *m* variables

Estimation des paramètres

Coefficient de sélection



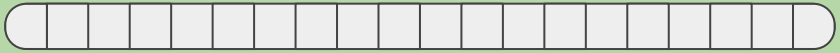
Taux de migration



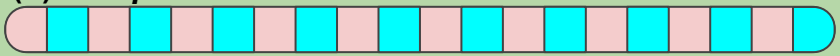
Test de robustesse : hétérogénéité du taux de recombinaison

Test à l'hétérogénéité du taux de recombinaison :

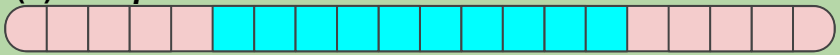
(a) Sans *coldspot*



(b) *Coldpsots* alternés



(c) *Coldpsot* étendu



Légende :

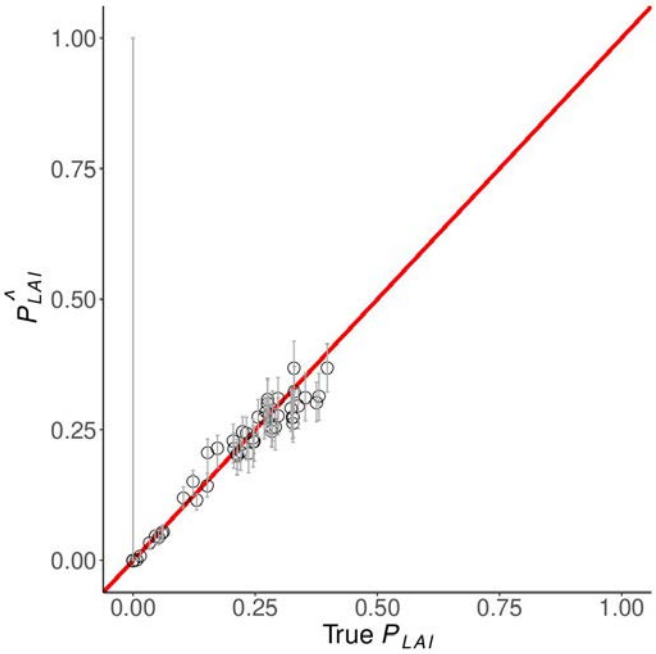
- Région de 50Kb
- Région de faible recombinaison (*coldspot*)
- Région de haute recombinaison
- Référence (recombinaison constante)

*Exemple de génome de 1Mb

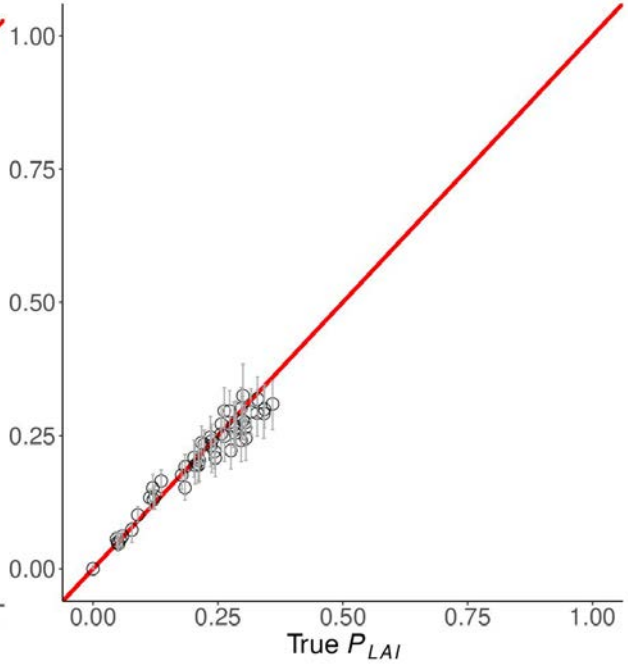
Résultats : Test à l'hétérogénéité du taux de recombinaison

Estimation des variables latentes

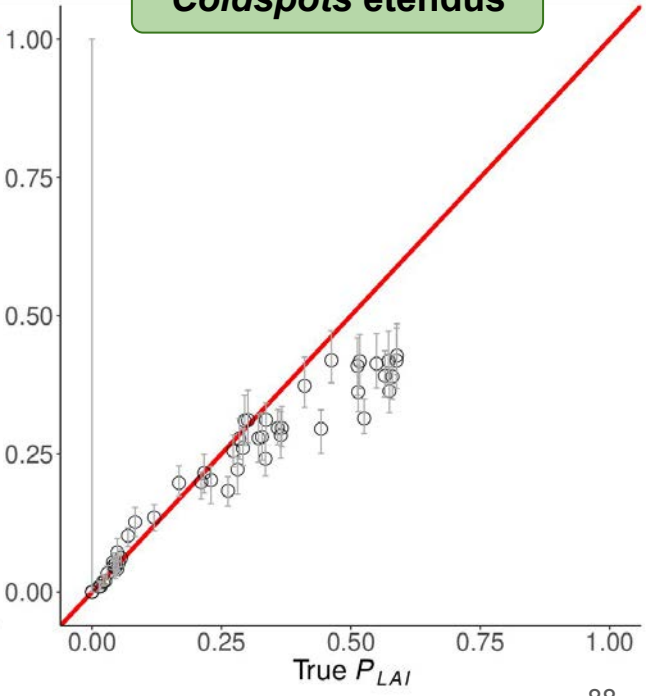
Sans *coldspot*



***Coldspots* alternés**



***Coldspots* étendus**



Discussions et conclusion :

- **Estimations des variables latentes :**
  - Bonnes précision des estimations des variables latentes liées à la proportion d'I et d'IA ($P_I$et P_{LIA})

- **Estimation des paramètres :**
  - Bonne précision des estimations du coefficient de sélection (*s*)

    ⇒ Effet de l'estimation conjointe de plus de paramètres (ex : temps de divergence)

- **Impact de l'hétérogénéité du taux de recombinaison :**
  - Baisse des performances d'estimation avec l'augmentation de l'hétérogénéité du taux de recombinaison

    ⇒ Prendre en compte l'effet de la recombinaison plus explicitement dans la méthode

Conclusion générale

- **Comparaison des méthodes d'inférences existantes de l'IA :**
  - *Benchmark* complet des méthodes de classification existantes
  - Meilleures méthodes pour nos tests :

Q95
(Racimo *et al.*, 2017)

- **Nouvelle méthode d'inférence de l'IA :**

  - Première méthode d'inférence des valeurs de variables latentes liées à l'IA
  - Estimation précise des variables latentes liées à la proportion d'I et d'IA

- **Ouvertures :**
  - Effet sur les inférences de la présence de populations D et R structurées ?

Article | Published: 13 December 2024

Ignoring population structure in hominin evolutionary models can lead to the inference of spurious admixture events

Rémi Tournebize & Lounès Chikhi

Nature Ecology & Evolution **9**, 225–236 (2025) | Cite this article

Conclusion générale

- **Comparaison des méthodes d'inférences existantes de l'IA :**
  - *Benchmark* complet des méthodes de classification existantes
  - Meilleures méthodes pour nos tests :

Q95
(Racimo *et al.*, 2017)

- **Nouvelle méthode d'inférence de l'IA :**
  - Première méthode d'inférence des valeurs de variables latentes liées à l'IA
  - Estimation précise des variables latentes liées à la proportion d'I et d'IA

- **Ouvertures :**
  - Effet sur les inférences de la présence de populations D et R structurées ?
  - Comment adapter la méthode à l'analyse de données réelles ?

- 1.5Gb
- 19 chromosomes

Podarcis sp

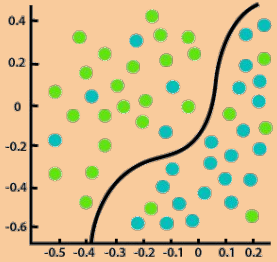
Remerciements

Merci pour votre attention !
(J'espère que la présentation s'est déroulée sans lézard pour vous)

Remerciements

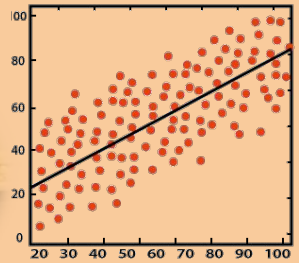
Remerciements spéciaux :

- Ghislain Camarata
  Collaboration pour la **partie 1**



Classification

- Mathieu Uhl
  Collaboration pour la **partie 2**



Regression

Remerciements

Liste non exhaustive des remerciement :

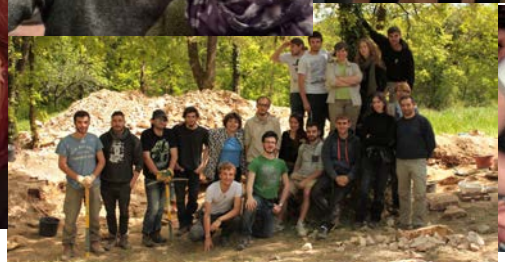
- Merci aux membres du jury
- Merci à mes encadrants
- Merci à mes parents
- Merci à tous les copains
- Merci aux co-bureaux
- Merci aux doctorants du CBGP
- Merci à tous ceux que j'ai oublié mais qui le méritent










Remerciements

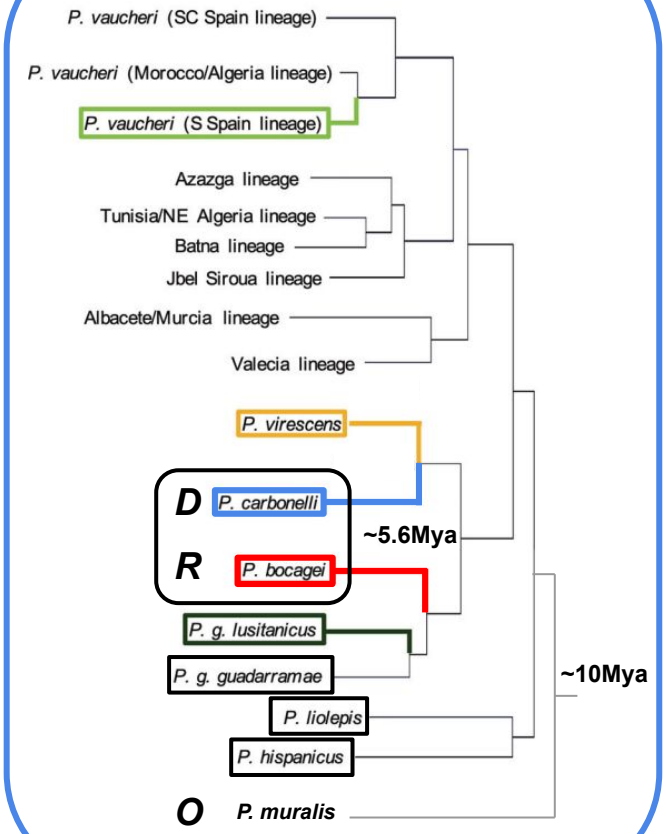
Liste non exhaustive des remerciement :

- Merci aux membres du jury
- Merci à mes encadrants
- Merci à mes parents
- Merci à tous les copains
- Merci aux co-bureaux
- Merci aux doctorants du CBGP
- Merci à tous ceux que j’ai oublié mais qui le méritent
- Merci à mon chat

Informations supplémentaires

Définir un modèle démographique pour *Podarcis*

Phylogénie

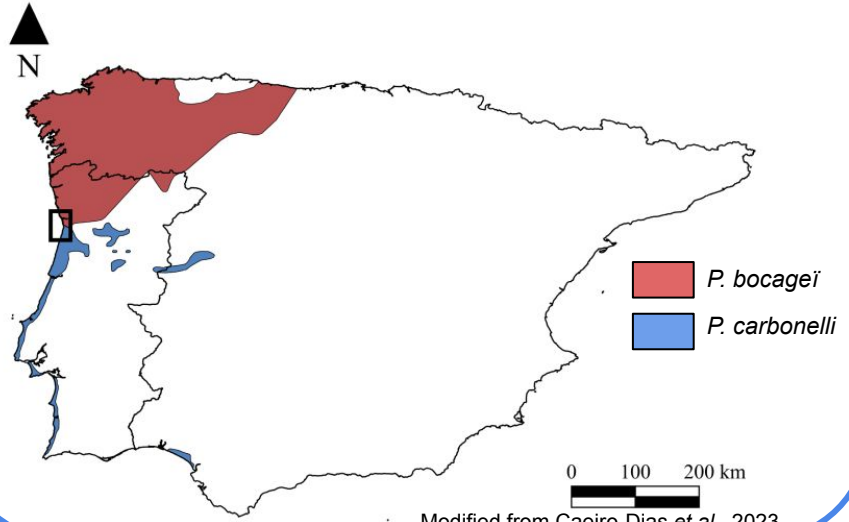


Modified from Kaliontzopoulou *et al.*, 2011

Divergence des population définies par le LGM

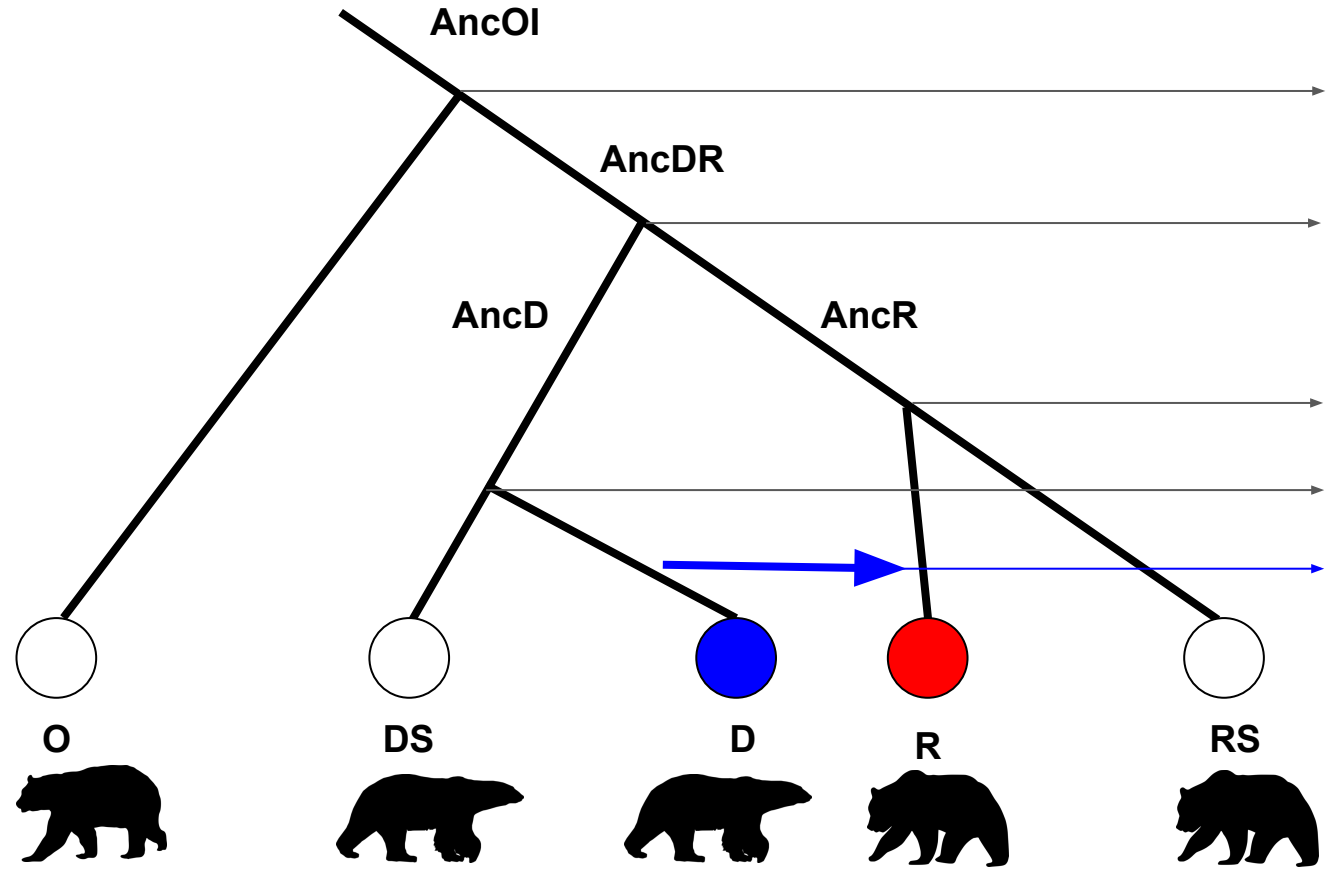
- D : Avant pour *P. carbonelli*
- R : Après pour *P. bocageï*

Current distribution



Modified from Caeiro-Dias *et al.*, 2023

Scenario Ursus : *U.maritimus to U.arctos*



3.4 Ma (2.0-4.7) Liu *et al.*, 2014 / 5-4 Ma Lan *et al.*, 2022
(between 1.34 and 5 Ma in all the publications)

~0.54 Ma Jong et *al.*, 2023
(between 0.264-1.6 Ma in all the publications)

~0.35 Ma (0.39-0.29) Jong et *al.*, 2023

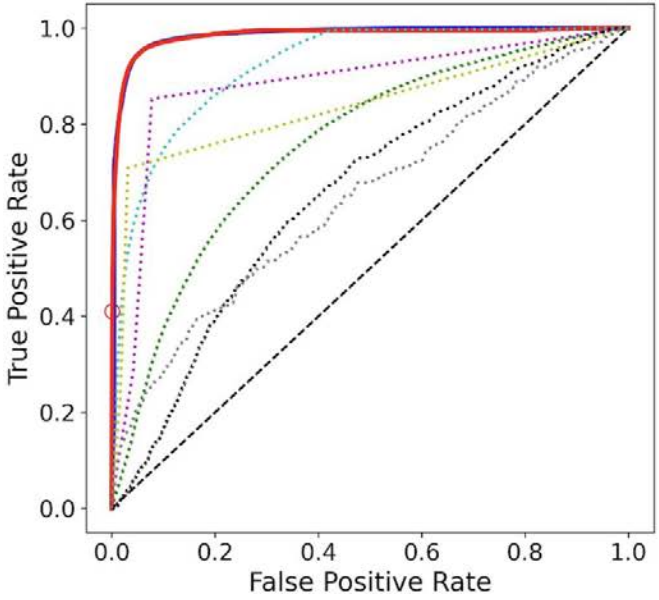
> **0.2 Ma**, Lan *et al.*, 2022
(between 0.1 and 0.33 Ma in all the publication)

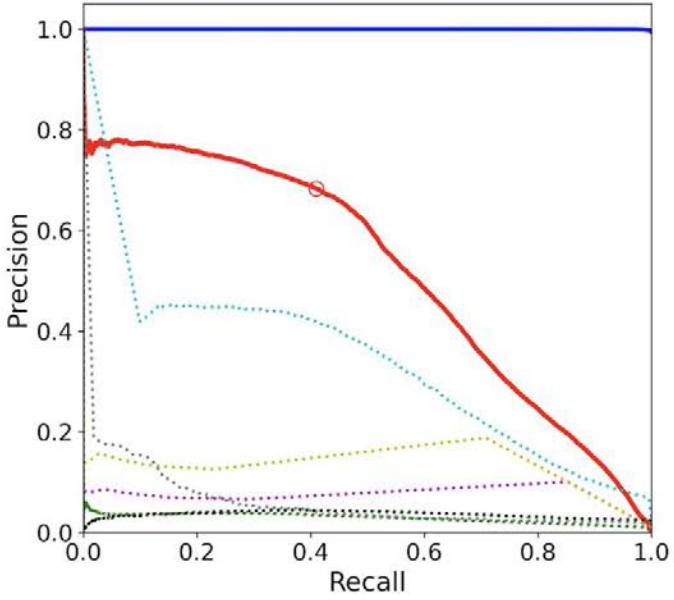
~0.150 Ma (0,148-0,319 Ma) Liu *et al.*, 2014
(between 0.1 and 0.32 Ma in all the publication)

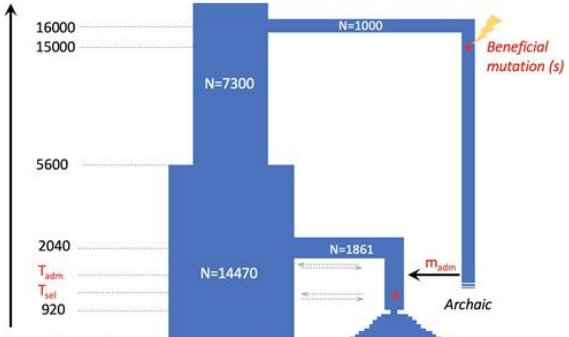
Admixture may occur before the split between polar bear populations

generation length :
- Polar bear = 11.35 years
- Brown bear = **10 years**

Comparaison des performance des méthode : MaLAdapt

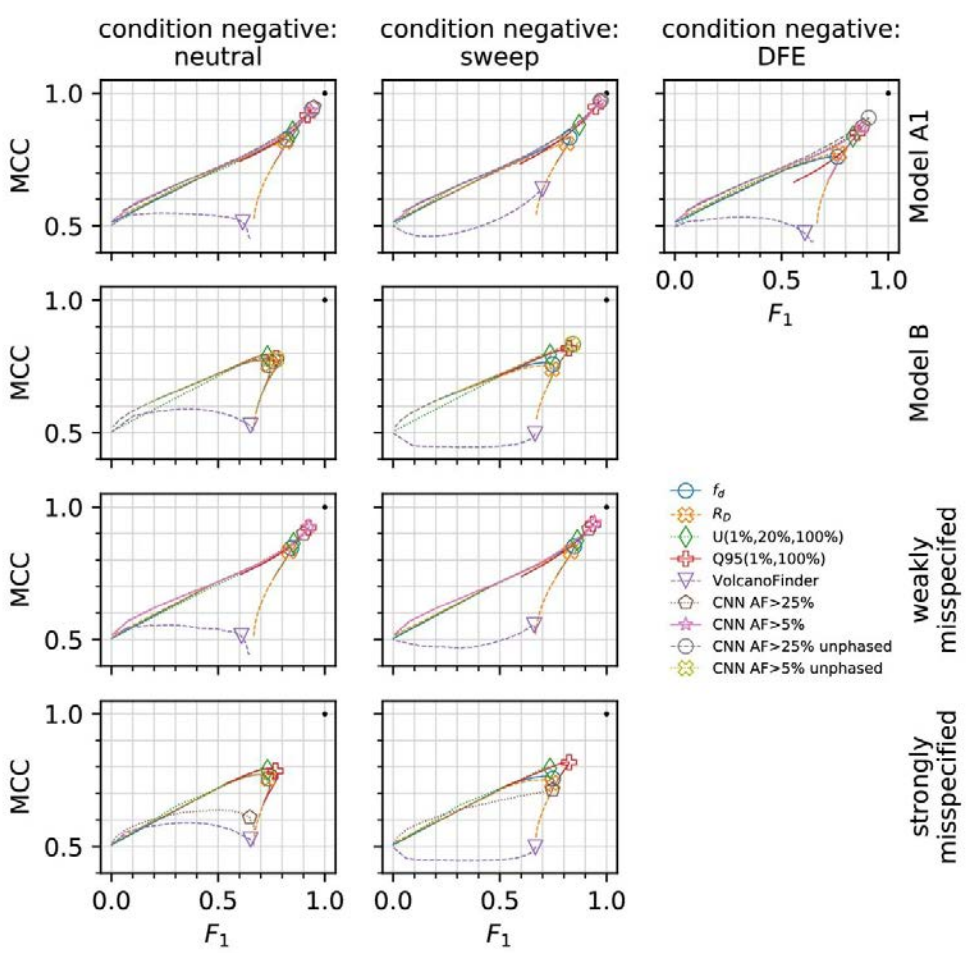


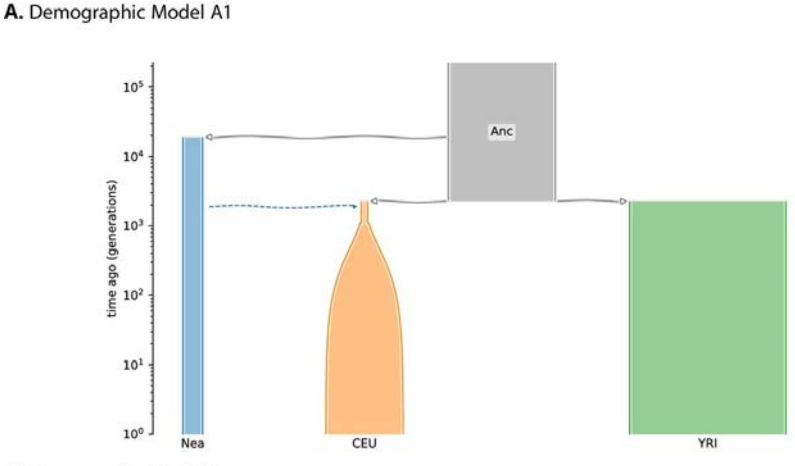




N = Différent ent P
m = {0.01, 0.02, 0.05, 0.1}
μ = 1.08e-8
r = map de recombination
s = [1e-4, 1e-2]
L = 5Mb
1 mutation dans D
1 pulse
Mutation délétère

Comparaison des performance des méthode : genomatnn



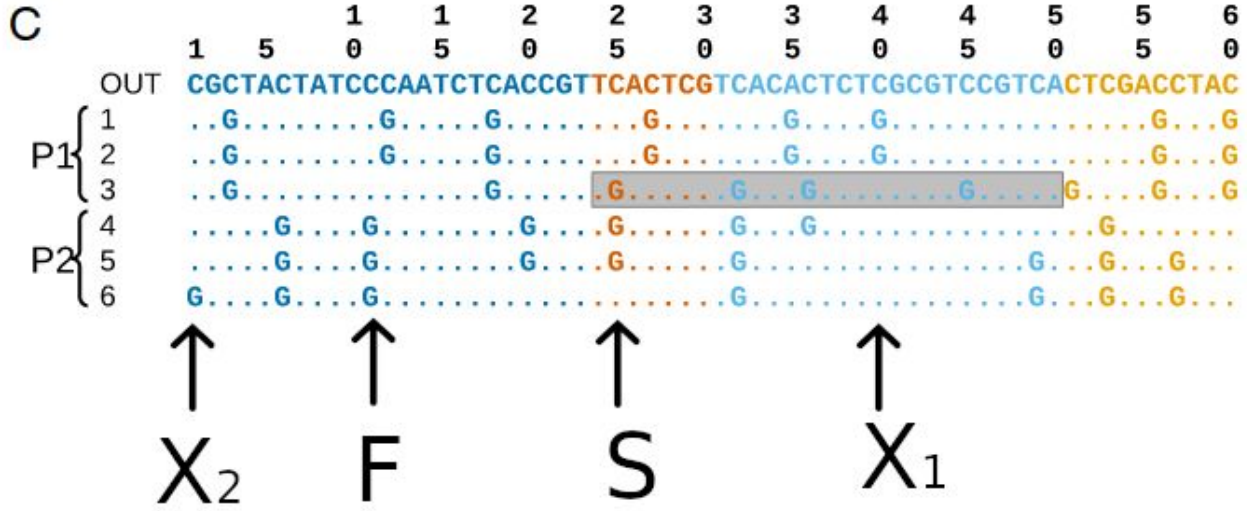


N = Différent entre P
m = 0.0225
μ = 1.29e-8
r = carte de recombination
s = [1e-4, 1e-1]
L = 100kb
1 mutation dans D
1 pulse

Statistiques résumées

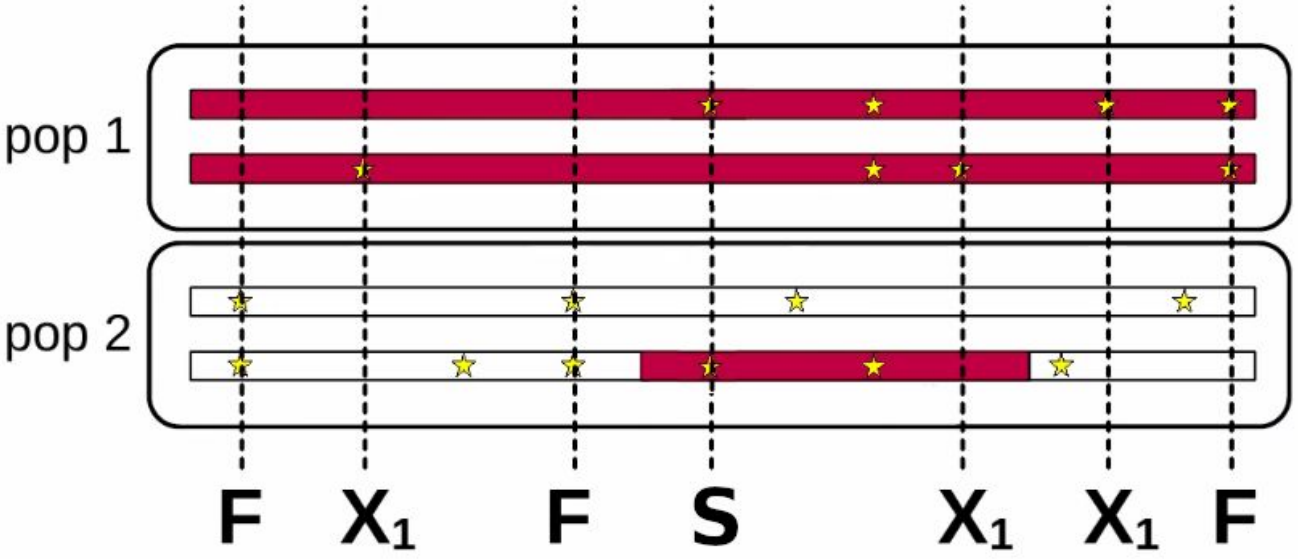
Statistique	Description	Signal capté	Référence
θ_π	Estimateur de la diversité	Diversité génétique	Tajima, 1983
θ_W	Estimateur de la diversité	Diversité génétique	Watterson, 1975
θ_H	Estimateur de la diversité	Diversité génétique	Fay and Wu, 2000
R_D	Rapport de divergence	Introgression	Racimo et al., 2017
$Q95(w, y)$	Quantiles des fréquences alléliques dérivées	Introgression adaptative	Racimo et al., 2017
$U(w, x, y)$	Nombre d'allèles partagés	Introgression adaptative	Racimo et al., 2017
Dxy	Divergence entre 2 populations	Introgression	Nei and Li, 1979
RND	Rapport de divergence	Introgression	Feder et al., 2005
d_f	Corrélations entre fréquences allèliques de 3 pop	Introgression	Pfeifer and Kapan, 2019
$f3$	Corrélations entre fréquences allèliques de 3 pop	Introgression	Patterson et al., 2012
$f4$	Corrélations entre fréquences allèliques de 4 pop	Introgression	Patterson et al., 2012
D	Corrélations entre fréquences allèliques de 4 pop	Introgression	Green et al., 2010
f_D	Corrélations entre fréquences allèliques de 4 pop	Introgression	Martin et al., 2015
FST	Différenciation des populations	//	Weir and Cockerham, 1984
Wx, Rf, Rs	Disposition spatiale des différences fixes, des polymorphismes partagés et exclusifs	Introgression	Navascués et al., 2014
Zns	Moyenne des DL par pair de variant	Desequilibre de liaison	Kelly, 1997
$H1$	Homozygotie de l'haplotype	Selection	Garud et al., 2015
$H2$	Homozygotie de l'haplotype	Selection	Garud et al., 2015
$H12$	Homozygotie de l'haplotype	Selection	Garud et al., 2015
$H123$	Homozygotie de l'haplotype	Selection	Garud et al., 2015
$H2/H1$	Homozygotie de l'haplotype	Selection	Garud et al., 2015

Nouvelles statistiques résumantes



Type de site	Description
X1	Polymorphisme exclusif de la population 1
X2	Polymorphisme exclusif de la population 2
S	Sites polymorphiques partagés
F	Différence fixées

Nouvelles statistiques résumantes



Introgression de P1 dans P2

Avec de l'introgression :

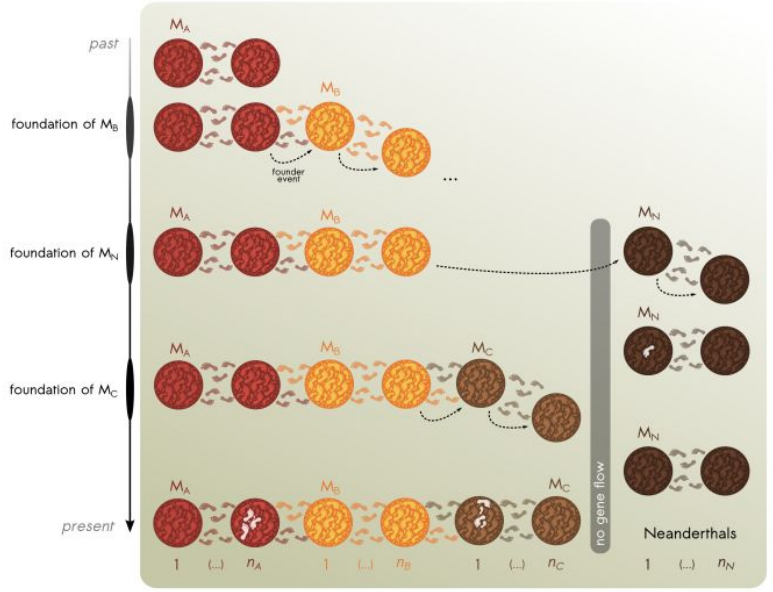
• Sites S mais rarement des sites F

Sans Introgression :

• Sites F et S ségrégé dans un petit nombre de groupes

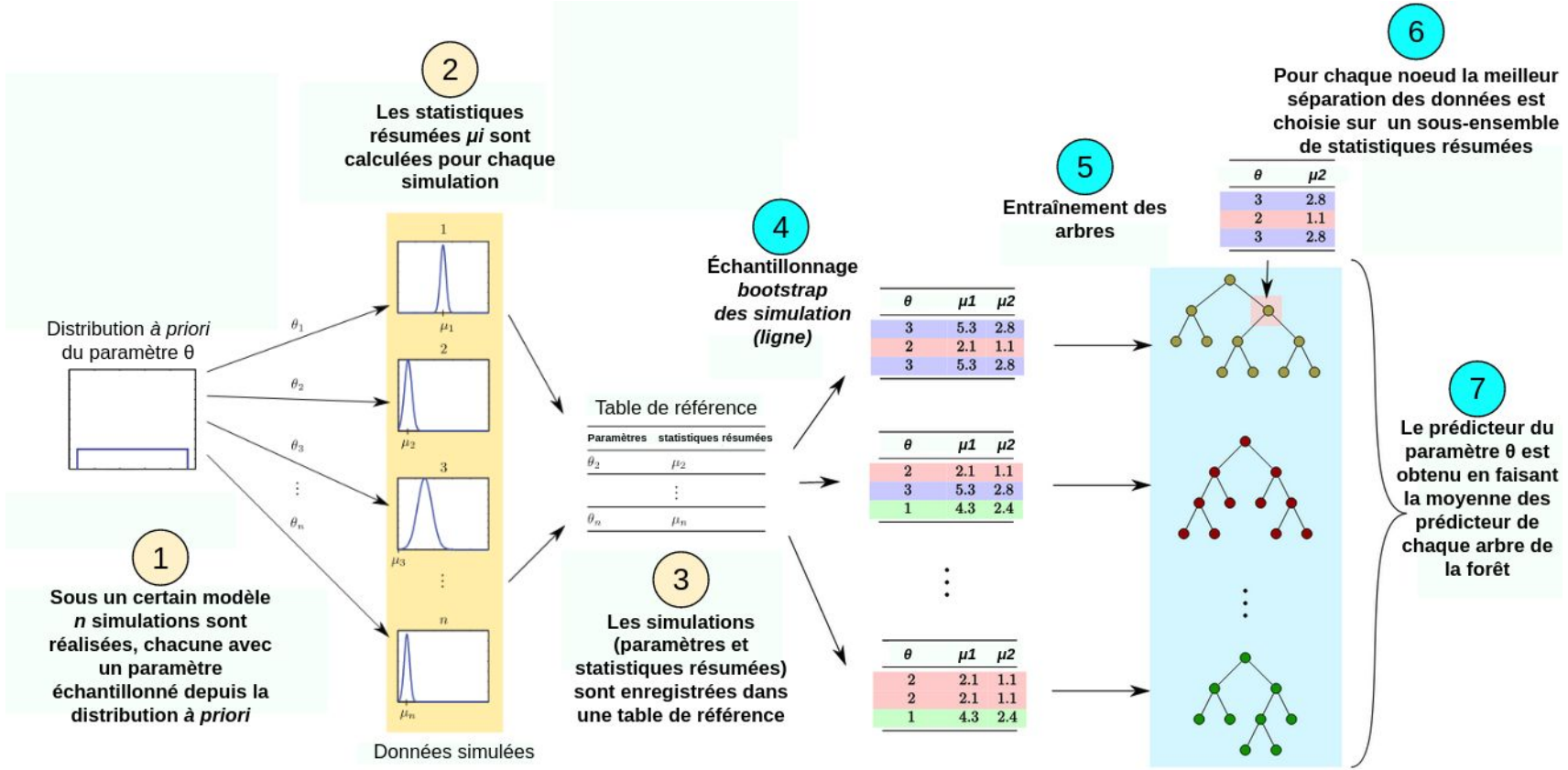
• Polymorphisme exclusif (X1 si migration de P1 à P2)

Impact de population spatialement structurées sur l'inférence de l'I



Simplified representation of the 1D structured model considered in this study. Time flows from top (past) to bottom (present), with an initial metapopulation M_A consisting of n_A (=10) demes exchanging migrants with their neighbours. At some point in the past, the rightmost deme of M_A founds a new metapopulation M_B of n_B (=10) demes, with which it will continue exchanging migrants till the present. Later, the rightmost deme of M_B founds the metapopulation M_N of n_N (=10) demes which will become Neanderthals. The M_N metapopulation will never exchange migrants with any other deme from the other metapopulations. Closer towards the present, the rightmost deme of M_B founds M_C which corresponds to the expansion of *H. sapiens* towards Eurasia. White feet represent the sampled metapopulations (not the specific demes) for respective sampling times. The location of the sampled demes (within the corresponding metapopulations) is a random variable (see Notes S1.1). Fifty individuals are sampled in M_A and in M_C to represent modern-day YRI and CEU samples respectively. For the Neanderthals (M_N), one individual is sampled at 50 kya. More details about the models can be found in the Materials and Methods section and in Notes S1.1.

Description de l'ABC-RF



Algorithme Vraisemblance Résumée dans Infusion

Algorithme 3 Méthode de la vraisemblance résumée intégrée dans `Infusion`

N_{ini}, Nombre d'échantillons à simuler pour la tableau d'entrainement initiale. N_{it} nombre d'itérations à réaliser. $i_{\Theta}^{(0)}(\boldsymbol{\theta})$ la distribution instrumentale initiale. D_{train}, le tableau d'entrainement.

Étape 1 : Création du tableau d'entrainement initial

Pour $i \leftarrow 1$ **to** N_{ini} **Faire**

Simule $\boldsymbol{\theta}_i \sim i_{\Theta}^{(0)}(\boldsymbol{\theta})$;

Simule $S_i \sim P(S; \boldsymbol{\theta}_i)$;

Calcule $\mathbf{T}_{S_i}$;

Ajoute $\boldsymbol{\theta}_i$, $\mathbf{T}_{S_i}$ à D_{train};

Fin Pour

Étape 2 : Statistiques projetées

Entraine un modèle de forêts aléatoire pour chaque paramètre θ en utilisant $\mathbf{T}_S$ comme variables prédictives.

Utilise les prédictions de la forêt aléatoire comme statistique projeté $p[\mathbf{T}_S]$.

Étape 3 : Première estimation de la vraisemblance résumée

Estime la densité conjointe $\hat{P}_{\mathbf{T},\Theta}\{p[\mathbf{T}_D], \boldsymbol{\theta}\}$ en utilisant un modèle MGM.

Calcule la vraisemblance résumée :

$$\hat{L}\{\boldsymbol{\theta}; p[\mathbf{T}_D]\} = \frac{\hat{P}_{\mathbf{T},\Theta}\{p[\mathbf{T}_D], \boldsymbol{\theta}\}}{i_{\Theta}(\boldsymbol{\theta})}. \tag{2.11}$$

Trouve $\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \hat{L}\{\boldsymbol{\theta}; p[\mathbf{T}_D]\}$.

Étape 4 : Phase itérative

Pour $i \leftarrow 1$ **to** N_{it} **Faire**

Met à jour $i_{\Theta}(\boldsymbol{\theta})$ pour concentrer l'échantillonnage des paramètres dans la région de haute vraisemblance.

Tire de nouvelles valeurs de $\boldsymbol{\theta}$ dans cette région en utilisant la distribution instrumentale mise à jour.

Simule de nouveaux échantillons S, calcule $\mathbf{T}_S$, et projette en $p[\mathbf{T}_S]$.

Ajoute ces données à la table d'entrainement.

Ré-estime la densité conjointe $\hat{P}_{\mathbf{T},\Theta}\{p[\mathbf{T}_D], \boldsymbol{\theta}\}$ et met à jour $\hat{L}\{\boldsymbol{\theta}; p[\mathbf{T}_D]\}$.

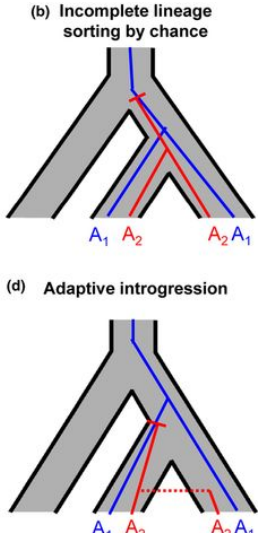
Fin Pour

Retourne $\hat{\boldsymbol{\theta}}$ et ses intervalles de confiances.

ILS

ILS : Le phénomène qui entraîne que deux ou plusieurs lignées échouent a coalescer dans leur population ancestrale la plus récente

- Les arbres des gènes individuels sont discordants avec l'histoire de l'espèce
  - Lorsque l'ILS se produit, il devient possible que l'ordre des événements de coalescence diffère d l'ordre des séparation des populations/espèces dans la phylogénie



Impact de population spatialement structurées sur l'inférence de l'I

"*Une partie du signal d'admixture pourrait s'expliquer par la structure spatiale de la population qui conduit à l'isolement par la distance et au triage incomplet des lignées (ILS) dans certaines populations Hs.*"

"*Les modèles actuels d'admixture ne parvenaient pas à expliquer certain patrons de diversité génétique qui peuvent être expliqués sans admixture mais avec uniquement la structure de la population.*"

"Dans ces modèles structurés mais sans admixture, les segments d'ADN identifiés comme introgressés sont le résultat d'une ascendance partagée et sont hérités d'un ancêtre commun structuré à la fois pour Hn et Hs."

EX : Eriksson and Manica (2012)

La statistique D n'est pas robuste à la structure spatiale de la population

"Ils ont estimé des valeurs de D similaires à celles observées dans les données réelles, mais uniquement avec un modèle prenant en compte une structure spatiale de la population (stepping stone 1D)."

"Les arbres génétiques échantillonnés dans des populations structurées ont des propriétés qu'aucun modèle panmictique ne peut reproduire."

"Des longs fragments interprétés comme introgressés de Neandertal peuvent être détectés sans mélange Hn réel, une fois que la structure spatiale est prise en compte."

Des variables latentes

“Une variable latente est une variable non observée et aléatoire sous une distribution donnée”

“Un paramètre n'est en comparaison pas variable sous une distribution donnée puisqu'on la justement fixé pour identifier une distribution dans la famille”