

# Inférence bayésienne robuste et précise de généralogies pangénomiques pour de grands échantillons

Yun Deng<sup>1</sup>, Rasmus Nielsen<sup>\*2,3,4</sup>, et Yun S. Song<sup>\*2,5</sup>

<sup>1</sup>Centre de biologie computationnelle, Université de Californie, Berkeley, États-Unis

<sup>2</sup>Département de statistiques, Université de Californie, Berkeley, États-Unis

<sup>3</sup>Département de biologie intégrative, Université de Californie, Berkeley, États-Unis

<sup>4</sup>Centre de géogénétique, Université de Copenhague, Danemark

<sup>5</sup>Division d'informatique, Université de Californie, Berkeley, États-Unis

16 mars 2024

## Abstrait

L'Ancestral Recombination Graph (ARG), qui décrit l'histoire généralogique complète d'un échantillon de génomes, est un outil essentiel en génomique des populations et en recherche biomédicale. Les progrès récents ont augmenté l'évolutivité de la reconstruction ARG jusqu'à des dizaines ou des centaines de milliers de génomes, mais ces méthodes reposent sur des heuristiques, ce qui peut réduire la précision, en particulier en présence d'une mauvaise spécification du modèle. De plus, ils ne reconstruisent qu'une seule topologie ARG et ne peuvent pas quantifier l'incertitude considérable associée aux inférences ARG. Pour relever ces défis, nous présentons ici SINGER, une nouvelle méthode qui accélère l'échantillonnage ARG à partir de la distribution postérieure de deux ordres de grandeur, permettant une inférence précise et une quantification de l'incertitude pour les grands échantillons. Grâce à des simulations approfondies, nous démontrons la précision et la robustesse améliorées de SINGER pour modéliser les erreurs de spécification par rapport aux méthodes existantes. Nous illustrons l'utilité de SINGER en l'appliquant aux populations africaines dans le cadre du projet 1000 Genomes, identifiant les signaux d'adaptation locale et d'introgression archaïque, ainsi qu'un fort soutien au polymorphisme trans-espèce et à l'équilibrage de la sélection dans les régions HLA.

## 1. Introduction

De nombreux problèmes en génomique dépendent de méthodes informatiques permettant de déduire des informations généralogiques à partir d'une vaste collection de séquences d'ADN et d'interpréter les arbres reconstruits. En particulier, les approches généralogiques ont joué un rôle fondamental dans la compréhension de la variation génétique humaine.<sup>[28,46, 58]</sup> et ont jeté les bases de nombreuses méthodes informatiques utilisées dans la recherche biomédicale. Ces méthodes incluent le phasage des haplotypes lors du séquençage des génomes humains [3] et prédire la puissance statistique dans les études d'association à l'échelle du génome [29,53]. Chez les espèces à recombinaison, comme les humains, la relation généralogique ne peut pas être représentée par un seul arbre. Au lieu de cela, des millions d'arbres différents existent dans le génome, chaque position du génome ayant généralement son propre arbre qui ne diffère que très peu des arbres des sites voisins. L'ensemble de tous ces arbres, ainsi que l'ensemble des points de recombinaison qui créent de nouveaux arbres, est représenté par l'Ancestral

\*À qui la correspondance doit être adressée : rasmus\_nielsen@berkeley.edu , yss@berkeley.edu

Le graphe de recombinaison (ARG) et le modèle génératif correspondant sont appelés *le coalescent avec recombinaison*[13,21].

Bien que la simulation sous coalescence avec recombinaison soit simple [2,22,27], déduire des ARG à partir de données de variation génétique reste un problème difficile. Ce défi se pose parce que l'espace d'état est extrêmement vaste, tandis que les mutations qui informent la relation généalogique à une position donnée du génome sont limitées. Les ARG peuvent être construits de manière itérative en recherchant les branches et les moments auxquels les *n*La lignée rejoint l'ARG partiel pour la première fois *n*-1 génotypes, un processus appelé « threading ». En utilisant une approximation connue sous le nom de coalescence séquentielle de Markov (SMC) [36,37,57] et en formulant le problème du threading sous la forme d'un modèle de Markov caché (HMM), en combinaison avec une méthode intelligente de Monte Carlo par chaîne de Markov (MCMC), ARGweaver [44] peut échantillonner des ARG à l'échelle du génome à partir de la distribution postérieure approximative pour une taille d'échantillon modérée. Cependant, cela entraîne une surcharge de calcul importante, ce qui le rend peu pratique pour plus de quelques dizaines d'individus.

Récemment, des progrès significatifs ont été réalisés dans l'extension de la reconstruction ARG à des dizaines ou des centaines de milliers de génotypes. Des outils comme Relate [50] et tsinfer/tsdate [28,58] utilisent un HMM efficace de Li et Stephens [33] pour déduire une séquence de topologies d'arbres locaux le long du génome, suivie d'une estimation de la longueur des branches. ARG-Aiguille [60] construit des ARG en utilisant l'approche de thread susmentionnée, similaire à ARGweaver, ainsi que plusieurs heuristiques et approximations pour atteindre l'évolutivité. La capacité de déduire des ARG à l'échelle du génome pour de grands échantillons ouvre de nouvelles directions de recherche[16,17] et a permis une série d'applications basées sur l'ARG dans la recherche en génétique statistique et des populations [9,dix,14,19,24,25,41,47,51,56].

Malgré les progrès mentionnés ci-dessus, les méthodes d'inférence ARG actuelles présentent des limites importantes, qui entravent les analyses basées sur l'ARG des données de séquençage du génome entier (WGS). Premièrement, bien que les méthodes récentes soient hautement évolutives, cette vitesse accrue se fait au détriment de la précision dans des aspects clés de l'ARG reconstruit, tels que les temps de coalescence[59] et les événements de recombinaison [7]. Par exemple, Relate et tsinfer+tsdate présentent une précision considérablement réduite pour les temps de coalescence anciens, ce qui diminue leur efficacité dans les applications impliquant des temps anciens, telles que la détection de la sélection d'équilibrage. De plus, ni Relate ni tsinfer ne déduisent explicitement la position de la branche ou l'heure des événements de recombinaison, et il est difficile d'extraire les informations d'identité par descendance (IBD) de la représentation interne des données de Relate. Deuxièmement, les méthodes évolutives n'exploront généralement pas de topologies ARG alternatives, reconstruisant souvent une seule topologie ARG. Certains négligent également l'incertitude dans les estimations du temps de coalescence. Comme nous le démontrons dans cet article, un échantillonnage précis des topologies ARG peut considérablement améliorer l'inférence statistique, en particulier pour l'analyse d'arbres génétiques locaux, où une estimation ponctuelle peut être assez bruyante. Des applications telles que l'inférence d'ascendance locale, la détection d'introgression et l'analyse de sélection seraient difficiles sans intervalles de confiance appropriés. Par exemple, INDICES [51], une méthode basée sur la vraisemblance pour déduire les trajectoires de sélection et de fréquence allélique, bénéficie de l'incorporation d'un échantillon d'arbres locaux plutôt que de s'appuyer uniquement sur une estimation ponctuelle. Troisièmement, la plupart des méthodes de reconstruction ARG supposent un a priori simpliste, tel qu'une population panmictique de taille constante et l'absence de sélection, et ne sont pas robustes à la violation de ces hypothèses. Cette limitation a des implications conséquentes sur les applications, car de nombreuses populations humaines ont subi des changements démographiques complexes, notamment des goulots d'étranglement et des expansions récentes[32,48]. De même, l'influence de la sélection d'origine est omniprésente et façonne profondément le paysage de la diversité[38,40].

Pour résoudre les problèmes mentionnés ci-dessus, nous introduisons ici une nouvelle méthode d'inférence bayésienne ARG, SINGER (**S**ampling et **D**ANSferrage de **G**Enéalogies avec **R**écombinaison). Il réalise toutes les fonctionnalités d'ARGweaver – y compris l'échantillonnage postérieur basé sur MCMC, l'exploration de l'espace topologique, le suivi des événements de recombinaison, etc. – tout en étant au moins un ordre de grandeur plus rapide. Grâce à des simulations approfondies, nous démontrons que SINGER atteint une plus grande précision

dans plusieurs aspects cruciaux de l'inférence ARG par rapport aux méthodes concurrentes. Ils présentent également une plus grande robustesse face à diverses sources courantes de spécifications erronées du modèle. Nous soulignons l'utilité de notre méthode en l'appliquant aux populations africaines dans le cadre du projet 1000 Genomes, révélant des signaux d'adaptation locale et d'introgression archaïque, ainsi que des preuves solides de polymorphisme trans-espèces et de sélection équilibrée dans les régions HLA.

## 2 résultats

### 2.1 Un aperçu de l'algorithme SINGER

SINGER échantillonne les ARG de manière itérative en ajoutant un haplotype à la fois via une opération appelée threading [44]. Conditionné à un ARG partiel pour le premier  $n-1$  haplotypes, l'opération de threading échantillonne les points auxquels la lignée pour le  $n$  haplotype rejoint l'ARG partiel. SINGER résout ce problème de threading en construisant d'abord un HMM avec des branches comme états cachés et en échantillonnant une séquence de branches se joignant le long du génome depuis la partie postérieure, en utilisant un traçage stochastique (Figure 1 UN). Ensuite, SINGER construit un autre HMM avec des temps de jointure comme états cachés, conditionnés par ces branches de jointure échantillonées (Figure 1B). Nous appelons ces deux étapes « échantillonnage de branches » (Section 4.1) et « échantillonnage temporel » (Section 4.2), respectivement. Notons que cet algorithme de threading en deux étapes est approximatif. Cependant, comparé au HMM d'ARGweaver, qui traite chaque point de jonction de l'arborescence comme un état caché, l'algorithme de thread en deux étapes est beaucoup plus rapide car le nombre d'états cachés est considérablement plus petit.

De plus, pour explorer l'espace ARG (à la fois la topologie et la longueur des branches) en fonction de la distribution postérieure, nous réalisons MCMC en utilisant une nouvelle proposition que nous appelons Sub-Graph Pruning and Regrafting (SGPR). En bref, une opération SGPR élague d'abord un sous-graphe en introduisant une coupe sur un arbre choisi au hasard, puis étend la coupe vers la gauche et la droite (Figure 1D); dans la section supplémentaire A.4, nous montrons que la première étape de cette opération est équivalente à l'étape de retrait dans ce que l'on appelle le « mouvement de Kuhner » [30,35]. Cependant, notre opération SGPR diffère considérablement de celle de Kuhner en ce qui concerne l'étape de regreffe ; ce dernier échantillonne le précédent par simulation, tandis que notre méthode utilise l'algorithme de threading pour échantillonner le postérieur. La procédure de simulation du mouvement Kuhner ne prend pas en compte les données et il est donc peu probable qu'elle atteigne un état avec une probabilité améliorée ; en revanche, notre algorithme de threading encourage la compatibilité avec les données, ce qui conduit probablement à de meilleures mises à jour. Par rapport aux méthodes précédentes, le SGPR a ainsi un bien meilleur taux d'acceptation tout en introduisant d'importantes mises à jour de l'ARG (Supplémentaire Section A.4), ce qui entraîne un meilleur taux de convergence et un meilleur mélange du MCMC.

Enfin, pour ajuster certains biais résultant des approximations de l'algorithme, pour chaque intervalle d'amincissement de l'algorithme MCMC, nous redimensionnons les longueurs de branches (Section 4.3), en utilisant un algorithme que nous appelons « ARG re-scaling » (voir la section 4.3 pour plus de détails).

### 2.2 Benchmarks de performances sur les données de simulation

Nous avons d'abord comparé les performances de diverses méthodes d'inférence ARG à l'aide de données simulées.

**Installation.** Nous avons utilisé `msprime`[26] pour simuler des régions de 1 Mo pour des ensembles de 50 ou 300 haplotypes, en utilisant  $\mu=r=2\times 10^{-8}$ , où  $\mu$  et  $r$  désignent respectivement le taux de mutation et le taux de recombinaison par génération et par paire de bases (pb). Nous avons comparé les résultats de SINGER avec ceux des méthodes d'inférence ARG populaires, à savoir ARGweaver, Relate, tsinfer+tsdate et ARG-Needle. Pour Relate et tsinfer+tsdate, les résultats sont basés sur des moyennes a posteriori sur la topologie fixe estimée par ces méthodes. ARG-Needle utilise des moyennes postérieures de temps de jonction lors de l'enfilage. CHANTEUR

et ARGweaver peuvent faire la moyenne sur les topologies postérieures, et nos résultats pour ces méthodes sont basés sur 100 ARG échantillonnés. Pour garantir une comparaison équitable, nous avons utilisé les mêmes intervalles de déverminage et d'éclaircissement pour toutes les méthodes. La méthodologie détaillée peut être trouvée dans la section [4.5](#).

**Précision du temps de coalescence.** Pour évaluer l'exactitude des temps de coalescence déduits, nous avons comparé les temps de coalescence par paires de vérité terrain et ceux déduits pour 100 paires de noeuds feuilles choisis au hasard, comme dans YC Brandt et al. [\[59\]](#). Les temps de coalescence par paire jouent un rôle important dans l'application des ARG inférés, par exemple pour l'inférence démographique [\[50\]](#), GWAS [\[dix,60\]](#), et dans les études évolutionnistes [\[56\]](#). Pour les données simulées avec 50 haplotypes, SINGER avait une précision plus élevée que toutes les autres méthodes, ARGweaver et Relate ayant des performances similaires, et tsinfer+tsdate ayant les pires performances (Figure [2UN](#)). ARG-Needle ne fonctionne que pour plus de 300 séquences et n'a donc pas été inclus dans cette comparaison pour 50 haplotypes. Pour 300 haplotypes, nous avons comparé uniquement SINGER, Relate, tsinfer+tsdate et ARG-Needle, car la taille de l'échantillon est trop grande pour ARGweaver. SINGER a obtenu la plus grande précision, tandis que Relate et ARG-Needle ont obtenu des résultats similaires, tsinfer+tsdate ayant à nouveau obtenu les pires résultats (Figure supplémentaire [S17](#)). L'amélioration de SINGER par rapport à ARGweaver pourrait être due à une meilleure efficacité de mélange dans MCMC et à un schéma de discréétisation temporelle plus flexible (entraînant des bacs plus fins).

Nous avons également comparé l'approche coalescente par paires qui effectue une inférence pour chaque paire de séquences séparément, indépendamment des autres séquences. Plus précisément, nous avons considéré une méthode récemment proposée, Gamma-SMC [\[49\]](#), qui est ultra rapide. Comparaison de la figure supplémentaire [S10](#) avec figurine [2A](#) suggère que SINGER peut largement surpasser Gamma-SMC, alors que Relate et tsinfer+tsdate ne semblent pas s'améliorer par rapport à Gamma-SMC en termes de MSE ou de corrélation.

Une autre référence que nous avons considérée est le nombre moyen de lignées à l'échelle du génome en fonction du temps dans les arbres marginaux ; cette statistique est pertinente pour l'inférence de la démographie et de la sélection. ARGweaver a sous-estimé de nombreux temps de coalescence récents, ce qui a entraîné une baisse trop rapide du nombre de lignées par rapport aux attentes (Figure [2D](#)). Ceci est cohérent avec les résultats de la figure [2A](#), ainsi que les conclusions de Speidel et al. [\[50\]](#) qu'ARGweaver a tendance à sous-estimer les temps. Nous avons également constaté que tsdate a tendance à surestimer considérablement les temps de coalescence (Figure [2D](#)). En revanche, nous avons observé que les résultats de Relate et SINGER concordent bien avec les attentes (Figure [2D](#)).

**Précision de la topologie arborescente.** Pour évaluer la précision des topologies d'arbres déduites, nous avons utilisé la distance triplet, définie comme la fraction de sous-arbres à 3 feuilles présentant des topologies différentes dans une paire d'arbres donnée. Cette métrique est particulièrement intéressante car la précision des applications telles que l'imputation et l'inférence d'ascendance locale (avec deux populations sources) dépend de la précision des topologies de triplet. En moyenne, SINGER a atteint les distances triplet les plus faibles par rapport à la vérité terrain (Figure [2E](#)). Encore une fois, nous avons observé qu'ARGweaver n'est pas aussi précis que SINGER. Cette précision réduite peut être due au mélange MCMC moins efficace d'ARGweaver et à la présence de polytomies dans les arbres déduits par ARGweaver en raison de sa discréétisation temporelle.

**Robustesse face aux erreurs de spécification du modèle.** L'un des avantages de notre méthode est qu'elle est plus robuste aux erreurs de spécification du modèle ; en particulier, il est moins affecté par l'utilisation d'une mauvaise référence en termes de taille effective de la population  $N_e$  ou ne tenant pas compte des changements dans la taille de la population. Lorsque nous avons effectué une inférence à l'aide d'un  $N_e$  d'un facteur 5, les temps de coalescence déduits par SINGER étaient moins affectés que ceux déduits par Relate et tsinfer + tsdate, qui présentaient des biais systématiques à la baisse (Figure supplémentaire [S11](#)).

Nous avons également simulé les données sous un historique de taille de population déduit pour CEU [\[42,54\]](#), qui contient un goulot d'étranglement et une expansion récente. L'exécution de l'inférence ARG pour ces données a montré que

non seulement SINGER déduit plus précisément les temps de coalescence qu'ARGweaver, Relate et tsinfer+tsdate (Figure2B), mais qu'il capture également avec précision la bimodalité dans la distribution du temps de coalescence par paire provoquée par le goulot d'étranglement (Figure2C). Relate peut intégrer les changements de taille de la population, mais cela nécessite que l'utilisateur exécute un module distinct pour réestimer la longueur des branches et déduire lui-même les taux de coalescence, ce qui prend encore plus de temps que l'exécution de Relate lui-même. En revanche, SINGER ajuste automatiquement la longueur des branches avec le redimensionnement ARG et déduit avec précision les temps de coalescence sans pratiquement aucune surcharge de calcul supplémentaire, car l'algorithme de redimensionnement ARG est très efficace. Cependant, ARG-Needle nécessite de fournir au préalable un historique de taille afin d'ajuster ses temps de coalescence, et n'est pas capable de gérer un historique de taille inconnu. Nous soulignons qu'une méthode permettant de réestimer conjointement la longueur des branches et l'historique de la taille de la population est susceptible d'améliorer encore les résultats de SINGER.

**Comparaison d'exécution.** Comme décrit dans Méthodes, l'initialisation et les étapes MCMC ultérieures dans SINGER et ARGweaver utilisent toutes des algorithmes de threading. Par conséquent, nous avons comparé le temps d'exécution de l'opération de threading en fonction du nombre de feuilles dans l'ARG partiel. Comparé à ARGweaver, le threading de SINGER est d'environ  $\times 10$  plus rapide. De plus, comme décrit dans la section2.2, le schéma MCMC de SINGER est nettement plus efficace que celui d'ARGweaver, ce qui implique qu'il nécessite beaucoup moins d'itérations MCMC pour le mixage, réduisant ainsi encore plus le coût de calcul, ce qui entraîne une accélération totale de  $\sim 400\times$ . Nous avons également mis en œuvre des outils de parallélisation automatique dans SINGER pour échantillonner simultanément les ARG de plusieurs régions génomiques. Dans l'application de données réelles décrite ci-dessous, l'exécution de 10 000 itérations MCMC pour 200 séquences du génome entier africain a pris environ 2 jours sur des processeurs Intel Xeon E5-2643 v3 (3,4 Ghz) avec 120 cœurs au total.

**Précision des inférences de mutation et de recombinaison.** Nous avons également comparé l'estimation de l'âge des allèles à l'aide des ARG déduits. Nous avons omis ARG-Needle et ARGweaver dans ce benchmark, car ARG-Needle ne mappe pas les mutations aux généralogies déduites et il est difficile de récupérer la cartographie mutationnelle et les âges à partir de la sortie d'ARGweaver. Pour les données simulées avec 50 séquences, SINGER a nettement surpassé Relate et tsinfer+tsdate dans cette tâche (Figure3UN). Les résultats pour 300 séquences sont illustrés dans la Figure supplémentaireS18, ce qui suggère que SINGER reste plus précis que Relate et tsinfer+tsdate pour des échantillons de plus grande taille.

Dans un autre benchmark, nous avons comparé le nombre de points d'arrêt de recombinaison déduits à la vérité terrain dans chaque fenêtre de 5 Ko. Seuls ARGweaver et SINGER ont produit des estimations précises (Figure3B). Relate et tsinfer ont manqué de nombreux événements de recombinaison, une découverte cohérente avec des études antérieures.<sup>[7]</sup>.

Une autre façon d'évaluer l'exactitude des inférences de recombinaison consiste à examiner la distribution de la longueur de l'IBD par paire, car les segments de l'IBD sont perturbés par les recombinaisons. Dans cette analyse, ARGweaver et Relate ont été exclus ; le premier en raison des difficultés rencontrées pour extraire les informations IBD de sa sortie, et le second parce qu'il n'assure pas la persistance des noeuds parmi les arbres marginaux. Les résultats, illustrés dans la figure3C, montrent que SINGER a capturé avec précision la distribution de la longueur de l'IBD par paire, tandis que tsinfer a considérablement surestimé la longueur de l'IBD, ce qui s'aligne sur les résultats précédents (<sup>[7]</sup>).

**Comparaison de la convergence MCMC.** Chez l'homme et de nombreux autres organismes, la moyenne du taux de recombinaison à l'échelle du génome est similaire au taux de mutation moyen, ce qui entraîne une incertitude substantielle dans l'inférence ARG. Il est donc important d'obtenir des échantillons de la distribution a posteriori et de s'assurer que les incertitudes sont bien caractérisées, plutôt que de se fier uniquement à une estimation ponctuelle. Pour évaluer la convergence MCMC, nous avons obtenu 100 échantillons MCMC postérieurs de chacun des ARGweaver, Relate et SINGER, en utilisant les mêmes intervalles de déverminage et d'amincissement pour toutes les méthodes.

Pour évaluer l'efficacité de l'échantillonnage des ARG à partir de la distribution postérieure, nous avons utilisé le même point de référence que dans YC Brandt et al. [59]. Cela impliquait d'analyser des diagrammes de rangs de temps de coalescence par paire et de quantifier les écarts par rapport à la distribution uniforme, ce qui serait obtenu par un échantillonneur parfait à partir de la distribution postérieure [6,52]. Un diagramme de classement est un histogramme du rang d'un paramètre échantillonné à partir de l'échantillon antérieur par rapport à l'échantillon postérieur. Idéalement, un MCMC convergent et bien mélangé devrait produire des classements qui suivent une distribution uniforme. En revanche, un diagramme de rang en forme de U suggère un échantillonnage à partir d'une distribution sous-dispersée [6,52]. Nous avons observé que les tracés de classement pour SINGER sont beaucoup plus proches de la distribution uniforme par rapport à ceux d'ARGweaver et Relate (Figure 4B).

Le diagramme de classement est étroitement lié à la propriété de couverture des intervalles de crédibilité (IC) empiriques. Pour chaque position dans le génome et pour une paire d'haplotypes donnée, nous avons défini l'intervalle de crédibilité empirique à 90 % en sélectionnant le 5ème au 95ème centile des temps de coalescence par paire à partir des ARG échantillonnes (Figure 4UN). Cette approche a été appliquée de la même manière aux 70% et 50% intervalles crédibles. Le 90% intervalle crédible couvrait la vérité terrain en seulement 44% d'instances pour Relate et 54% pour ARGweaver. En revanche, la couverture était nettement meilleure pour SINGER, à 85% (Chiffre 4C). Les IC à d'autres niveaux ont également montré une couverture supérieure pour SINGER (Figure 4C).

De plus, notre référence a montré que même avec des intervalles d'amincissement 40 fois plus longs que SINGER, ARGweaver reste légèrement sous-performant dans l'inférence du temps de coalescence par paire et des propriétés de couverture des CI (Figure supplémentaire S12). Cela souligne l'avantage des mises à jour SGPR par rapport à l'algorithme MCMC d'ARGweaver. De plus, combiné à notre algorithme de threading plus rapide, cela implique qu'ARGweaver prend des centaines, voire des milliers de fois plus de temps pour atteindre des niveaux de performances comparables à ceux de SINGER. Pour Relate, même avec une longue éclaircie comme le suggère la documentation, la couverture CI reste nettement inférieure aux niveaux nominaux (Section supplémentaire B.3). Cela est probablement dû au fait que Relate échantillonne uniquement les temps de coalescence avec une topologie fixe, contrairement à SINGER, qui échantillonne à la fois les topologies et les temps de coalescence.

## 2.3 Applications aux données WGS africaines du projet 1000 Genomes

Nous avons appliqué SINGER à 200 séquences du génome entier de cinq populations indigènes africaines (GWD, YRI, ESN, LWK et MSL) dans le cadre du projet 1000 Genomes [4], avec 40 génomes tirés uniformément au hasard dans chaque population. Pour démontrer l'utilité de SINGER dans l'analyse génétique des populations, nous avons considéré l'inférence de quelques types différents de signaux évolutifs, notamment l'adaptation locale, la sélection d'équilibrage ancienne et l'introgression ancienne.

**Diagnostic des ARG échantillonnes par SINGER.** Nous avons examiné les ARG échantillonnes pour vérifier la convergence de la chaîne de Markov vers la stationnarité et pour nous assurer que les ARG utilisés dans nos analyses génétiques de population sont échantillonnes après un rodage approprié (Section supplémentaire C.3). Les chaînes ont généralement montré une bonne convergence vers la stationnarité (Figure supplémentaire S15). De plus, nous avons validé l'exactitude des ARG échantillonnes en comparant les temps de coalescence moyens par paires déduits (mis à l'échelle par  $4N_e\mu$ ) aux diversités empiriques moyennes par paires (basées sur les SNP) dans des fenêtres de 500 Ko, qui présentaient généralement une concordance élevée. En revanche, nous avons observé que l'ARG reconstruit par tsinfer+tsdate [58] a considérablement sous-estimé la variation de la diversité à l'échelle du génome (Figure supplémentaire S14). Cela est peut-être dû au fait que différentes régions génomiques ont des caractéristiques différentes.  $N_e$  valeurs résultant de différents niveaux de sélection d'arrière-plan [5,23]. Comme indiqué précédemment dans la section 2.2, SINGER est plus robuste à ce type de spécification erronée du modèle, même si un  $N_e$  est utilisé dans l'algorithme.

**Signatures de l'adaptation locale.** L'adaptation locale fait référence à des balayages sélectifs spécifiques à une population,

potentiellement due à des pressions sélectives exercées par des environnements locaux particuliers. Les régions génomiques impliquées dans l'adaptation locale présenteraient une diversité réduite pour les populations spécifiques soumises à la sélection, en particulier par rapport à d'autres populations sans balayage (Figure 5B). Cependant, l'inférence de la diversité locale basée sur SNP peut être plutôt bruyante à des échelles fines (Figure supplémentaire S13). D'un autre côté, avec des estimations de la longueur des branches à partir d'ARG déduits avec précision, la diversité à échelle fine peut être estimée avec plus de précision. Comme indiqué dans la section supplémentaire B.4 et figure supplémentaire S13, nous avons observé que SINGER produit des estimations plus précises de la diversité à petite échelle par rapport à Relate et tsinfer+tsdate. Cela offre une nouvelle opportunité d'utiliser les estimations basées sur l'ARG de la diversité à petite échelle spécifique à une population pour étudier l'adaptation locale. En particulier, nous notons que cette approche basée sur l'ARG peut toujours être appliquée pour détecter un balayage complet avec l'allèle bénéfique fixé dans la population, alors que des méthodes comme l'IHS [55] et le test proposé dans Relate [50] qui dépendent de variantes de ségrégation en souffriraient.

Pour trouver une réduction de la diversité locale spécifique à la population, nous avons divisé le génome en fenêtres de 1 Ko ne se chevauchant pas, puis, pour chaque fenêtre, avons calculé le rapport entre l'estimation de la diversité basée sur l'ARG pour l'échantillon combiné et l'estimation de la diversité spécifique à la population basée sur l'ARG, pour chacune des cinq populations ; les réductions de la diversité locale apparaîtraient sous forme de pics lorsque ces ratios seraient tracés le long du génome. La liste complète des régions affichant des niveaux élevés de ratio pour chaque population peut être téléchargée à partir des liens fournis dans la section Disponibilité des données. L'analyse du chevauchement avec les régions géniques a conduit à plusieurs découvertes intéressantes, dont quelques-unes sont mises en évidence dans la figure 5. Par exemple, nous avons découvert que le gène *MITF* connu une vague putative de dracunculose ; Il a été rapporté que ce gène est fonctionnellement lié à la pigmentation de la peau en codant pour un facteur de transcription induisant les mélanocytes [31]. Autour *MITF*, nous avons observé des différences substantielles dans l'estimation de la diversité locale basée sur l'ARG entre les cinq populations, ce qui concorde avec les découvertes précédentes concernant la variation de la pigmentation en Afrique [11]. Dans YRI, nous avons constaté que *SPCS3* peut avoir subi un balayage local ; ce gène code pour une protéine immunitaire censée avoir un impact sur la production de virions de flavivirus tels que le virus du Nil occidental et le virus de la fièvre jaune [61]. Ceci est concordant avec le rapport sur la propagation de ces maladies au Nigeria [1]. Enfin, nous avons constaté que *SCN9A*, qui code pour un canal sodium voltage-dépendant impliqué dans la perception de la douleur [45], a considérablement réduit la diversité du LWK par rapport aux autres populations.

**Sélection d'équilibrage dans le locus HLA.** Le locus de l'antigène leucocytaire humain (HLA) comprend un groupe de gènes sur le chromosome 6 humain qui codent pour des protéines transmembranaires qui présentent des peptides antigéniques aux cellules T. Cette région est connue pour être la région la plus diversifiée du génome humain, et on a émis l'hypothèse qu'elle est soumise à une sélection extrêmement équilibrée afin de maintenir sa grande diversité afin de faire face à divers défis immunitaires [12,34]. Il y a eu des preuves de polymorphisme trans-espèce pour certains allèles chez les primates, ce qui est par ailleurs très rare [12].

Les ARG déduits par SINGER montrent des temps de coalescence par paire extrêmement anciens dans le locus HLA, de nombreuses régions abritant des temps de coalescence plus anciens que le temps de divergence homme-chimpanzé (Figure 6C). Ce résultat est cohérent avec l'hypothèse de longue date d'une forte sélection équilibrée dans ce locus et avec les modèles observés de polymorphismes trans-espèces. Le temps de divergence entre les humains et les chimpanzés a été discutable, avec des estimations allant de 5 à 12 Mya [39]. Bien que de nombreux gènes dans la région HLA ne montrent pas de preuves solides de temps de coalescence plus anciens que la division homme-chimpanzé (par exemple, *TAP1*, *TAP2*, et *TAPBP*), beaucoup le font, notamment *HLA-A*, *HLA-DRB1*, et *HLA-DRB6*. Sans surprise, il n'y a pas de différences notables entre les cinq populations, le polymorphisme étant maintenu depuis l'Antiquité. En revanche, l'ARG déduit par [58] L'utilisation de tsinfer + tsdate n'affiche pas de temps de coalescence aussi extrêmes (Figure supplémentaire S19), ce qui pourrait être dû au manque de robustesse de la méthode pour modéliser une mauvaise spécification dans des régions s'écartant considérablement du principe de neutralité sélective.

**Introgression antique en Afrique.** Il a été émis l'hypothèse qu'il pourrait y avoir eu une introgression d'hominines archaïques « fantômes » inconnues dans d'anciens individus africains.[\[15\]](#). L'identification des étendues génomiques dans les génomes africains modernes résultant d'une telle introgression est une tâche difficile, en particulier lorsqu'il n'existe pas de génome connu des hominines sources. Cependant, les ARG peuvent faciliter cette tâche en utilisant l'observation suivante : pour une région génomique avec un tractus introgressé dans un haplotype donné, la coalescence entre cet haplotype et d'autres haplotypes sera épuisée dans l'intervalle entre le temps d'introgression et le temps de partage des humains modernes de la population « fantôme ». Ceci est similaire aux signaux de « branche longue » mentionnés dans [\[50\]](#), mais exprimé dans l'espace de coalescence par paires.

Cependant, les longues branches dans les ARG peuvent être plutôt sensibles à une inférence de topologie incorrecte ; plus précisément, la lignée introgressée peut se regrouper de manière incorrecte avec les lignées ancestrales de séquences non introgressées, détruisant ainsi la longue branche. Pour atténuer ce problème, nous proposons une nouvelle technique d'analyse d'introgression basée sur la carte thermique de distribution de coalescence. Pour toute séquence donnée, nous traçons la distribution du temps de coalescence par paires avec les séquences restantes dans chaque fenêtre de 10 Ko, comme illustré sur la figure[6A](#), où chaque colonne correspond à une fenêtre de 10 Ko. Nous avons constaté qu'il est utile d'utiliser une collection d'ARG échantillonnes dans la partie postérieure au lieu d'un seul ARG reconstruit, car la distribution de coalescence peut être plutôt bruyante dans cette dernière (Figure supplémentaire[S16](#)). L'utilisation d'échantillons ARG avec différentes topologies permet de lisser la carte thermique (Figure supplémentaire[S16](#)). Cette visualisation est similaire à celle proposée par Schweiger et Durbin [\[49\]](#).

Pour tester si un segment donné est issu d'un événement d'introgression ancien, nous pouvons rechercher un épuisement de la masse probable dans l'intervalle susmentionné et un enrichissement de la masse au-dessus de l'intervalle. Cette approche est plus robuste que de s'appuyer explicitement sur des branches longues, car un léger mauvais regroupement conduirait toujours à un épuisement probabiliste dans l'intervalle, alors que la branche longue serait complètement perturbée. À la suite de Durvasula et Sankararaman [\[8\]](#), nous avons utilisé respectivement 43 kya et 625 kya pour le temps d'introgression et le temps intermédiaire. Si nous traçons le rapport entre la probabilité de coalescence au-dessus du temps de fractionnement et celle dans l'intervalle entre l'introgression et les temps de fractionnement, les zones d'introgression devraient apparaître sous forme de pics, comme illustré dans (Figure[6B](#)), qui montre deux zones d'introgression archaïques potentielles d'une longueur de 210 kb et 90 kb (Figure[6UN](#)).

### 3 Discussion

Dans cet article, nous avons présenté SINGER, une nouvelle méthode d'inférence bayésienne conçue pour échantillonner efficacement les graphiques de recombinaison ancestrale à partir de la distribution postérieure. Ces méthodes mettent en œuvre un algorithme MCMC amélioré pour explorer l'espace ARG, permettant ainsi une caractérisation précise de l'incertitude dans les temps de coalescence et les topologies ARG. Notre approche représente le premier algorithme MCMC capable de s'adapter à au moins des centaines de séquences du génome entier tout en effectuant un échantillonnage postérieur complet des longueurs et des topologies des branches ARG. Par rapport à ARGweaver, notre approche bénéficie à la fois d'un algorithme de threading plus rapide et d'une exploration plus efficace de l'espace ARG, conduisant à un mixage MCMC amélioré. En estimant les quantités génétiques des populations clés – telles que les temps de coalescence, les topologies, les densités de recombinaison et l'âge des allèles – SINGER se compare favorablement aux méthodes d'inférence ARG existantes, notamment ARGweaver, Relate, tsinfer+tsdate et ARG-Needle. Comme nous l'avons démontré dans nos tests de référence, l'utilisation d'échantillons a posteriori peut sensiblement améliorer la précision de l'inférence et quantifier efficacement l'incertitude de l'estimation. Enfin et surtout, SINGER présente une plus grande robustesse aux sources courantes d'erreurs de spécification du modèle, telles que les changements de taille de population et la sélection de fond, sans qu'il soit nécessaire de les modéliser explicitement ou de les réestimer et de les réajuster ultérieurement.

Nous avons appliqué SINGER aux données WGS d'individus africains du projet 1000 Genomes, découvrant divers signaux de sélection et d'introgression archaïque. En utilisant des estimations de diversité à petite échelle spécifiques à une population dérivées de nos ARG déduits, nous avons identifié des régions génomiques montrant des preuves d'adaptation locale. De plus, nous avons trouvé des preuves solides d'un équilibrage de la sélection et des polymorphismes trans-espèces dans la région HLA, et avons cartographié les gènes associés aux pics de ces signaux. Enfin, nous avons utilisé une technique de visualisation, la carte thermique de distribution de coalescence, pour identifier les régions génomiques compatibles avec un modèle spécifique d'introgression archaïque.

Nous notons que notre approche proposée pour détecter les étendues d'introgression archaïques nécessite un modèle démographique d'introgression. Cependant, un débat est en cours concernant le moment, la force et même l'existence d'une introgression archaïque dans les populations africaines, comme le soulignent Ragsdale et al. [43]. À cet égard, les zones identifiées par notre approche ne seront fiables que si les périodes d'introgression et de division de la population supposées sont raisonnablement exactes. De plus, la détection de zones introgressées à l'aide d'ARG échantillonnés justifie un développement méthodologique plus approfondi. Notre carte thermique proposée de la distribution de coalescence résume les informations généalogiques au sein d'une collection d'ARG échantillonnés, fournissant une base pour le développement d'un algorithme systématique. Nous reportons aux recherches futures les défis liés à la sélection d'une valeur critique appropriée pour le taux de coalescence indiquant une introgression, ainsi qu'à la détermination des limites des zones introgressées.

Notre méthode présente quelques limites, ainsi que des stratégies potentielles pour y remédier. Premièrement, bien que SINGER soit nettement plus rapide et plus évolutif que les autres méthodes capables d'échantillonage postérieur complet des ARG, les applications de données réelles nécessitent souvent un grand nombre d'itérations MCMC pour déduire efficacement la distribution postérieure. Pour réduire cette charge de calcul, nous développons une méthode pour échantillonner les longueurs de branches ARG étant donné une topologie ARG fixe, qui pourrait ensuite être incorporée dans l'algorithme MCMC actuel de SINGER. De plus, la convergence MCMC a tendance à être plus lente pour les grands échantillons, il est donc nécessaire de concevoir des stratégies d'exploration ARG plus efficaces.

Deuxièmement, alors que SINGER montre une robustesse améliorée aux erreurs de spécification du modèle par rapport aux autres méthodes d'inférence ARG, en déduisant l'historique de la taille de la population et en réestimant la longueur des branches ARG, similaire à l'approche utilisée par Relate [50], peut améliorer encore la précision. Nous notons que l'algorithme de Relate permettant de déduire l'historique de la taille de la population et de réestimer la longueur des branches ne peut pas être facilement appliqué à d'autres méthodes d'inférence ARG, car il repose sur les structures de données spéciales de Relate. D'un autre côté, l'algorithme de tsdate pour échantillonner les longueurs de branches ARG fonctionne pour la structure de données de SINGER, mais il ne prend en charge que la taille de population constante. Plus généralement, conformément à ARGweaver-D Hubisz et al. [20], SINGER pourrait potentiellement être étendu pour inclure des modèles démographiques plus complexes.

Troisièmement, pour diviser le génome en compartiments, SINGER utilise actuellement des compartiments de taille égale pour simplifier la programmation, mais il serait préférable d'utiliser des tailles de compartiment dynamiques en fonction de la densité de recombinaison déduite. Le regroupement dynamique peut être particulièrement important pour les régions comportant des points chauds de recombinaison, car nous autorisons au plus une recombinaison entre des groupes adjacents et il serait donc moins probable de sous-estimer les recombinaisons en introduisant davantage de groupes. De plus, l'utilisation de bacs plus grands pour les régions à faibles taux de recombinaison réduira la durée d'exécution. Le paysage de recombinaison dans le génome humain comporte des points chauds ponctués séparés par des régions à taux de recombinaison plus faible, de sorte que le regroupement dynamique entraînera probablement des économies de calcul globales substantielles.

Enfin, notre méthode nécessite des génomes phasés en entrée et suppose un phasage précis. Malheureusement, il est souvent difficile d'obtenir des données progressives de haute qualité, en particulier pour les populations peu étudiées et les organismes non modèles. Par conséquent, l'amélioration des méthodes d'inférence ARG pour prendre en charge des données partiellement ou totalement non phasées augmenterait leur utilité. De plus, étant donné que tous les ensembles de données contiennent un certain degré d'erreurs de phase, la possibilité de corriger automatiquement les sites hétérozygotes mal phasés sur la base des ARG échantillonnés serait avantageuse. Cette fonctionnalité serait

particulièrement important dans les analyses conjointes des génomes anciens et modernes, car les échantillons anciens sont souvent mal ou non phasés.

## 4 méthodes

### 4.1 Échantillonnage des branches

Pour accélérer le calcul, nous partitionnons d'abord le génome en compartiments de taille égale, où la taille du compartiment est choisie pour être  $4 \times dix - 3/4 N_{er}$ . Nous construisons ensuite un HMM indexé par ces compartiments (loci) pour échantillonner les branches dans lesquelles la lignée pour le  $n$ lhaplotype rejoint l'ARG partiel pour le premier  $n-1$  haplotypes. L'espace d'état pour le lieu comprend toutes les branches de l'arbre marginal pour le locus dans l'ARG partiel et certaines branches de loci antérieurs. La définition précise de l'espace d'état peut être trouvée dans la section supplémentaire A.1. Pour le lieu, on utilise  $B$  pour désigner la branche sur laquelle se trouve la lignée du  $n$ lhaplotype se joint. Si l'ARG partiel ne contient pas déjà un événement de recombinaison entre les loci '-1et', alors la probabilité de transition du HMM est définie

comme

$$P(B=b_j | B_{-1}=b_{je}) = (1 - r_{je}) \delta_{je} + r \sum_{\substack{j \in S \\ k: b_k \in S}} \frac{q_j}{q_k},$$

où  $b_{je} \in S_{-1}$ ,  $b_j \in S$ , et  $r_{je}$  désigne la probabilité de recombinaison spécifique à la branche pour la branche  $b_{je}$ . La définition de  $r_{je}$ ,  $q_{je}$  peut être trouvé dans la section supplémentaire A.1.5. La structure de cette probabilité de transition est similaire à celle du modèle de Li-Stephens [33], mais avec des probabilités de recombinaison et de réadhésion spécifiques à une branche. Cela nous permet de réduire la complexité de calcul du HMM pour qu'elle soit linéaire par rapport au nombre d'états cachés, comme dans le modèle de Li-Stephens.

Nous limitons l'ARG final pour qu'il ait au plus un événement de recombinaison entre des locus adjacents. Par conséquent, si l'ARG partiel contient déjà un événement de recombinaison entre les loci '-1et', l'opération de threading n'est pas autorisée à introduire une recombinaison supplémentaire entre ces deux locus et la probabilité de transition dans ce cas est définie de manière similaire à celle de Rasmussen et al. [44]; les détails sont fournis dans la section supplémentaire A.1.4 et figure supplémentaire S3.

### 4.2 Échantillonnage temporel

Conditionnée sur une séquence de branches se joignant le long du génome résultant de l'algorithme d'échantillonnage de branches, l'étape d'échantillonnage temporel se déroule de la même manière que dans PSMC [32] mais avec la restriction que, pour chaque locus génomique  $\tau$ , le temps de coalescence doit se situer entre les deux points de terminaison de la branche de jointure pour le locus  $\tau$ . Pour accélérer le calcul, nous avons implémenté la technique de linéarisation de Harris et al. [18]. Les détails sont fournis dans la section supplémentaire A.2.

### 4.3 Redimensionnement ARG

Étant donné un ARG déduit, nous divisons l'axe temporel en fenêtres ne se chevauchant pas de telle sorte que la longueur totale des branches sur tous les arbres marginaux (pondérée par l'étendue de chaque arbre) dans chaque fenêtre temporelle soit la même ; par défaut, 100 fenêtres sont choisies. On compte ensuite le nombre de mutations tombant dans chacune de ces fenêtres. Si une mutation tombe sur une branche traversant plusieurs fenêtres, alors sa contribution au nombre de mutations pour chaque fenêtre est donnée par la proportion de la branche qui chevauche la fenêtre. Nous redimensionnons la taille de chaque fenêtre de telle sorte que le nombre attendu de mutations pour la fenêtre corresponde au nombre empirique. Il s'agit essentiellement d'une mise à l'échelle spécifique à la fenêtre pour mieux correspondre à l'horloge de mutation. De plus amples détails peuvent être trouvés dans la section supplémentaire A.3. Le redimensionnement ARG est effectué après l'initialisation et chaque étape d'éclaircissement.

## 4.4 Taille et regreffage de sous-graphe (SGPR)

Dans l'algorithme MCMC, nous proposons des mises à jour de l'ARG actuel en supprimant d'abord quelques branches suite à une coupure puis en regreffant à partir du point d'arrêt.

Pour retirer une branche d'un arbre marginal donné, on fait une coupe aléatoire sur l'arbre ; la probabilité qu'une branche donnée soit coupée est proportionnelle à sa longueur. Nous pouvons étendre la coupe vers la gauche et la droite le long du génome, en supprimant la branche partielle de la coupe jusqu'à son extrémité supérieure de la branche. Généralement, la coupure ne s'étendra pas sur l'ensemble du chromosome ; la largeur de l'extension sera plutôt la même que l'envergure du segment ancestral correspondant à la branche coupée. Les détails peuvent être trouvés dans la section supplémentaire A.4 et figure supplémentaire S8.

Pour regreffer la branche à partir du point d'arrêt, nous utilisons le même algorithme de threading que celui décrit précédemment, à la seule différence que nous ne considérons désormais que le sous-ARG au-dessus du point d'arrêt. Nous montrons dans la section supplémentaire A.4.3 que, en supposant que l'algorithme de threading échantillonne approximativement à partir du postérieur, la probabilité d'acceptation est généralement beaucoup plus élevée que celle des propositions précédentes [30,44], améliorant ainsi la convergence et le mixage dans MCMC.

## 4.5 Détails de la simulation et de l'analyse comparative

Toutes les simulations coalescentes de cet article ont été réalisées en utilisant *msprime* [26] avec  $r=m=2 \times 10^{-8}$  et  $N_e=1 \times 10^4$ . Nous avons simulé 50 ensembles de données contenant chacun 50 séquences sur une région de 1 Mo et 10 ensembles de données contenant chacun 300 séquences sur 1 Mo. Pour les simulations avec 50 séquences, nous avons également simulé avec l'historique de la taille de la population CEU. [https://github.com/PalamaraLab/ASMC\\_data/tree/main/demographies](https://github.com/PalamaraLab/ASMC_data/tree/main/demographies) estimé à partir de SMC++ [54]. Nous avons exécuté toutes les méthodes d'inférence avec ces vraies valeurs de paramètres.

En ce qui concerne l'échantillonnage MCMC, nous avons uniformisé le nombre d'itérations et le schéma d'éclaircissement pour toutes les méthodes. Nous avons prélevé 100 échantillons avec un intervalle d'amincissement fixé à 20 pour ARGweaver, Relate et SINGER, et utilisé 1 000 itérations pour le rodage.

Pour tous les tests de simulation impliquant ARGweaver et SINGER, des moyennes a posteriori ont été prises pour les statistiques d'intérêt, telles que le temps de coalescence par paire, l'âge des allèles, etc. Puisque Relate produit des moyennes d'itérations MCMC alors que les résultats tsdate sont des moyennes d'une table de probabilité, nous utilisons simplement leurs résultats à partir d'un seul produit car il s'agit en fait de moyennes postérieures.

## 4.6 Exécution de SINGER sur les données WGS africaines

Nous avons utilisé 200 génomes entiers de 5 populations autochtones africaines (GWD, YRI, ESN, LWK et MSL) dans le cadre du projet 1000 Genomes, avec 40 génomes tirés uniformément au hasard dans chaque population. Le taux de mutation par génération et par pb ( $\mu$ ) et le taux de recombinaison ( $r$ ) étaient tous deux réglés sur  $1.2 \times 10^{-8}$  dans CHANTEUR. Afin de déterminer la taille effective de la population  $N_e$ , nous avons comparé la différence moyenne empirique par paire ( $\pi \approx 0.001$ ) avec l'attente théorique ( $4N_e\mu$ ), ce qui a conduit à un choix de  $N_e=20,000$ . Nous avons exécuté SINGER pendant 10 000 itérations, les 4 000 premières itérations étant des rodages. Nous avons ensuite prélevé 100 échantillons du reste, en les éclaircissant toutes les 60 itérations.

## Disponibilité des données

Le code source de SINGER peut être téléchargé à partir de<https://github.com/popgenmethods/SINGER>. Nous avons téléchargé les échantillons ARG déduits (100 échantillons) et les cibles génétiques pour une adaptation locale à Zenedo sur<https://doi.org/10.5281/zenodo.10437053>,<https://doi.org/10.5281/zenodo.10467284>, et<https://doi.org/10.5281/zenodo.10467509>.

## Remerciements

Nous remercions Matthew Rasmussen et Melissa Hubisz pour leur correspondance utile concernant ARGweaver ; Vince Buffalo pour une discussion sur la sélection des antécédents ; Debora Brandt pour la discussion sur le pipeline de référence ; Joshua Schraiber pour ses conseils sur l'analyse des données ; Montgomery Slatkin et Alyssa Fortier pour une discussion sur HLA ; James Santangelo, Chuck Langley, Nathaniel Pope et Jules Perez pour avoir testé le logiciel. Cette recherche est financée en partie par une subvention R56-HG013117 du NIH.

## Les références

- [1] Adogo, L., Ogoh, M., 2020. Fièvre jaune au Nigéria : un examen de la situation actuelle. *Journal africain de microbiologie clinique et expérimentale* 21, 1–13.
- [2] Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, AP, Tsambos, G., Zhu, S., Eldon, B., Ellerman, EC, Galloway, JG, et al., 2022. Simulation efficace d'ascendance et de mutation avec msprime 1.0. *Génétique* 220, iyab229.
- [3] Browning, SR, Browning, BL, 2011. Phasement des haplotypes : méthodes existantes et nouveaux développements. *Nature Reviews Genetics* 12, 703-714.
- [4] Byrska-Bishop, M., Evani, US, Zhao, X., Basile, AO, Abel, HJ, Regier, AA, Corvelo, A., Clarke, WE, Musunuri, R., Nagulapalli, K., et al., 2022. Séquençage du génome entier à haute couverture de la cohorte élargie du projet de 1 000 génomes, comprenant 602 trios. *Cellule* 185, 3426-3440.
- [5] Charlesworth, B., Morgan, M., Charlesworth, D., 1993. L'effet des mutations délétères sur la variation moléculaire neutre. *Génétique* 134, 1289-1303.
- [6] Cook, SR, Gelman, A., Rubin, DB, 2006. Validation de logiciels pour modèles bayésiens utilisant des quantiles postérieurs. *Journal de statistiques informatiques et graphiques* 15, 675-692.
- [7] Deng, Y., Song, YS, Nielsen, R., 2021. La distribution des distances d'attente dans les graphiques de recombinaison ancestrale. *Biologie théorique des populations* 141, 34-43.
- [8] Durvasula, A., Sankararaman, S., 2020. Récupération des signaux d'introgression archaïque fantôme dans les populations africaines. *Avancées scientifiques* 6, eaax5097.
- [9] Fan, C., Cahoon, JL, Dinh, BL, Ortega-Del Vecchio, D., Huber, C., Edge, MD, Mancuso, N., Chiang, CW, 2023a. Un cadre basé sur la vraisemblance pour l'inférence démographique à partir d'arbres généalogiques. URL bioRxiv :<https://doi.org/10.1101/2023.10.10.561787>.
- [10] Fan, C., Mancuso, N., Chiang, CW, 2022. Une estimation généalogique des relations génétiques. *Le Journal américain de génétique humaine* 109, 812-824.
- [11] Fan, S., Spence, JP, Feng, Y., Hansen, ME, Terhorst, J., Beltrame, MH, Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., et coll., 2023b. Le séquençage du génome entier révèle une histoire démographique complexe de la population africaine et les signatures de l'adaptation locale. *Cellule* 186, 923-939.
- [12] Fortier, AL, Pritchard, JK, 2022. Polymorphisme trans-espèce ancien au niveau du complexe majeur d'histocompatibilité chez les primates. URL bioRxiv :<https://doi.org/10.1101/2022.06.28.497781>.
- [13] Griffiths, R., 1981. Modèles neutres à allèles multiples à deux locus avec recombinaison. *Biologie théorique des populations* 19, 169-186.
- [14] Guo, F., Carbone, I., Rasmussen, DA, 2022. Inférence phylogéographique sensible à la recombinaison utilisant le coalescent structuré avec recombinaison ancestrale. *Biologie computationnelle PLOS* 18, e1010422.
- [15] Hammer, MF, Woerner, AE, Mendez, FL, Watkins, JC, Wall, JD, 2011. Preuve génétique d'un mélange archaïque en Afrique. *Actes de l'Académie nationale des sciences* 108, 15123-15128.

- [16] Harris, K., 2019. D'une base de données de génomes à une forêt d'arbres évolutifs. *Génétique naturelle* 51, 1306-1307.
- [17] Harris, K., 2023. Utiliser d'énormes généalogies pour cartographier les variantes causales dans l'espace et le temps. *Nature Génétique* 55, 730-731.
- [18] Harris, K., Sheehan, S., Kamm, JA, Song, YS, 2014. Décoder les modèles de Markov cachés coalescents en temps linéaire, dans : *Research in Computational Molecular Biology : 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, États-Unis, 2-5 avril 2014, Actes 18*, Springer. pp. 100-114.
- [19] Hejase, HA, Mo, Z., Campagna, L., Siepel, A., 2022. Une approche d'apprentissage en profondeur pour l'inférence de balayages sélectifs à partir du graphe de recombinaison ancestral. *Biologie moléculaire et évolution* 39, msab332.
- [20] Hubisz, MJ, Williams, AL, Siepel, A., 2020. Cartographie du flux génétique entre les hominidés anciens grâce à l'inférence démographique du graphique de recombinaison ancestrale. *PLoS Génétique* 16, e1008895. URL :<http://dx.doi.org/10.1371/journal.pgen.1008895>, est ce que je:[10.1371/journal.pgen.1008895](https://doi.org/10.1371/journal.pgen.1008895).
- [21] Hudson, RR, 1983. Propriétés d'un modèle d'allèle neutre avec recombinaison intragénique. *Biologie théorique des populations* 23, 183-201.
- [22] Hudson, RR, 2002. Génération d'échantillons selon un modèle neutre de variation génétique Wright-Fisher. *Bioinformatique* 18, 337-338.
- [23] Hudson, RR, Kaplan, NL, 1995. Sélection de fond délétère avec recombinaison. *Génétique* 141, 1605-1617.
- [24] Ignatieva, A., Favero, M., Koskela, J., Sant, J., Myers, SR, 2023. La distribution de la durée des branches et la détection des inversions dans les graphes de recombinaison ancestrale. URL bioRxiv :<https://doi.org/10.1101/2023.07.11.548567>.
- [25] Ignatieva, A., Hein, J., Jenkins, PA, 2022. Recombinaison en cours dans le SRAS-CoV-2 révélée par la reconstruction généalogique. *Biologie moléculaire et évolution* 39, msac028.
- [26] Kelleher, J., Etheridge, AM, McVean, G., 2016. Simulation coalescente et analyse généalogique efficaces pour de grandes tailles d'échantillons. *Biologie computationnelle PLoS* 12, 1-22. est ce que je:[10.1371/journal.pcbi.1004842](https://doi.org/10.1371/journal.pcbi.1004842).
- [27] Kelleher, J., Thornton, KR, Ashander, J., Ralph, PL, 2018. Enregistrement efficace du pedigree pour une simulation rapide de la génétique des populations. *PLoS Biologie computationnelle* 14, e1006581.
- [28] Kelleher, J., Wong, Y., Wohns, AW, Fadil, C., Albers, PK, McVean, G., 2019. Déduire l'histoire du génome entier dans de grands ensembles de données de population. *Génétique naturelle* 51, 1330-1338.
- [29] Kruglyak, L., 2008. La voie vers des études d'association à l'échelle du génome. *Nature examine la génétique* 9, 314-318.
- [30] Kuhner, MK, Yamato, J., Felsenstein, J., 2000. Estimation du maximum de vraisemblance des taux de recombinaison à partir des données de population. *Génétique* 156, 1393-1401.
- [31] Levy, C., Khaled, M., Fisher, DE, 2006. MITF : maître régulateur du développement des mélanocytes et de l'oncogène du mélanome. *Tendances en médecine moléculaire* 12, 406-414.

- [32] Li, H., Durbin, R., 2011. Inférence de l'histoire de la population humaine à partir de séquences individuelles du génome entier. *Nature* 475, 493-496.
- [33] Li, N., Stephens, M., 2003. Modélisation du déséquilibre de liaison et identification des points chauds de recombinaison à l'aide de données de polymorphisme mononucléotidique. *Génétique* 165, 2213-2233.
- [34] Liu, B., Shao, Y., Fu, R., 2021. État actuel de la recherche sur HLA dans les maladies liées au système immunitaire. *Immunité, inflammation et maladie* 9, 340-350.
- [35] Mahmoudi, A., Koskela, J., Kelleher, J., Chan, Yb, Balding, D., 2022. Inférence bayésienne de graphes de recombinaison ancestrale. *Biologie computationnelle PLOS* 18, e1009960.
- [36] Marjoram, P., Wall, JD, 2006. Simulation « coalescente » rapide. *BMC Genetics* 7, Numéro d'article : 16.
- [37] McVean, GA, Cardin, NJ, 2005. Rapprochement du coalescent avec recombinaison. *Transactions philosophiques de la Royal Society B : Biological Sciences* 360, 1387-1393.
- [38] McVicker, G., Gordon, D., Davis, C., Green, P., 2009. Signatures génomiques généralisées de la sélection naturelle dans l'évolution des hominidés. *PLoS Génétique* 5, e1000471.
- [39] Moorjani, P., Amorim, CEG, Arndt, PF, Przeworski, M., 2016. Variation de l'horloge moléculaire des primates. *Actes de l'Académie nationale des sciences* 113, 10607-10612.
- [40] Murphy, DA, Elyashiv, E., Amster, G., Sella, G., 2022. Une variation à grande échelle des niveaux de diversité génétique humaine est prédite en purifiant la sélection sur les éléments codants et non codants. *Elife* 12, e76065.
- [41] Osmond, M., Coop, G., 2021. Estimation des taux de dispersion et localisation des ancêtres génétiques avec des généalogies à l'échelle du génome. URL bioRxiv :<https://doi.org/10.1101/2021.07.13.452277>.
- [42] Palamara, PF, Terhorst, J., Song, YS, Price, AL, 2018. L'inférence à haut débit des temps de coalescence par paires identifie les signaux de sélection et l'héritabilité enrichie de la maladie. *Nature Génétique* 50, 1311-1317.
- [43] Ragsdale, AP, Weaver, TD, Atkinson, EG, Hoal, EG, Möller, M., Henn, BM, Gravel, S., 2023. Une souche faiblement structurée pour les origines humaines en Afrique. *Nature* 617, 755-763.
- [44] Rasmussen, MD, Hubisz, MJ, Gronau, I., Siepel, A., 2014. Inférence à l'échelle du génome de graphiques de recombinaison ancestrale. *PLoS Génétique* 10, e1004342.
- [45] Reimann, F., Cox, JJ, Belfer, I., Diatchenko, L., Zaykin, DV, McHale, DP, Drenth, JP, Dai, F., Wheeler, J., Sanders, F., et al., 2010. La perception de la douleur est altérée par un polymorphisme nucléotidique dans SCN9A. *Actes de l'Académie nationale des sciences* 107, 5148-5153.
- [46] Rosenberg, NA, Nordborg, M., 2002. Arbres généalogiques, théorie de la coalescence et analyse des polymorphismes génétiques. *Nature Reviews Genetics* 3, 380-390.
- [47] Salehi Nowbandegani, P., Wohns, AW, Ballard, JL, Lander, ES, Bloemendal, A., Neale, BM, O'Connor, LJ, 2023. Modèles extrêmement clairsemés de déséquilibre de liaison dans des études d'associations ancestrales diverses. *Génétique de la nature*, 1-9.
- [48] Schiffels, S., Durbin, R., 2014. Déduire la taille de la population humaine et l'historique de séparation à partir de plusieurs séquences du génome. *Génétique naturelle* 46, 919-925.

- [49] Schweiger, R., Durbin, R., 2023. Inférence ultrarapide à l'échelle du génome des temps de coalescence par paires. *Recherche sur le génome* 33, 1–9.
- [50] Speidel, L., Forest, M., Shi, S., Myers, SR, 2019. Une méthode d'estimation généalogique à l'échelle du génome pour des milliers d'échantillons. *Génétique naturelle* 51, 1321-1329.
- [51] Stern, AJ, Wilton, PR, Nielsen, R., 2019. Une méthode approximative de pleine vraisemblance pour déduire les trajectoires de sélection et de fréquence allélique à partir des données de séquence d'ADN. *PLoS Génétique* 15, 1-32.
- [52] Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A., 2020. Validation des algorithmes d'inférence bayésienne avec un étalonnage basé sur la simulation. URL arXiv :<https://doi.org/10.48550/arXiv.1804.0678>.
- [53] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., Meyre, D., 2019. Avantages et limites des études d'association pangénomiques. *Nature Reviews Genetics* 20, 467-484.
- [54] Terhorst, J., Kamm, JA, Song, YS, 2017. Inférence robuste et évolutive de l'histoire de la population à partir de centaines de génomes entiers non phasés. *Nature Génétique* 49, 303-309.
- [55] Voight, BF, Kudaravalli, S., Wen, X., Pritchard, JK, 2006. Une carte de la sélection positive récente dans le génome humain. *PLoS Biologie* 4, e72.
- [56] Wang, S., Coop, G., 2022. Une histoire évolutive complexe des barrières génétiques au flux génétique chez les parulines en hybridation. URL bioRxiv :<https://doi.org/10.1101/2022.11.14.516535>.
- [57] Wiuf, C., Hein, J., 1999. La recombinaison en tant que processus ponctuel le long de séquences. *Biologie théorique des populations* 55, 248-259.
- [58] Wohns, AW, Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., McVean, G., 2022. Une généalogie unifiée des génomes modernes et anciens. *Sciences* 375, eabi8264.
- [59] YC Brandt, D., Wei, X., Deng, Y., Vaughn, AH, Nielsen, R., 2022. Évaluation des méthodes d'estimation des temps de coalescence à l'aide de graphiques de recombinaison ancestrale. *Génétique* 221, iyac044.
- [60] Zhang, BC, Biddanda, A., Gunnarsson, Á.F., Cooper, F., Palamara, PF, 2023. L'inférence à l'échelle d'une biobanque de graphiques de recombinaison ancestrale permet une analyse généalogique de traits complexes. *Génétique de la nature*, 1-9.
- [61] Zhang, R., Miner, JJ, Gorman, MJ, Rausch, K., Ramage, H., White, JP, Zuiani, A., Zhang, P., Fernandez, E., Zhang, Q., et al., 2016. Un criblage CRISPR définit une voie de traitement du peptide signal requise par les flavivirus. *Nature* 535, 164-168.

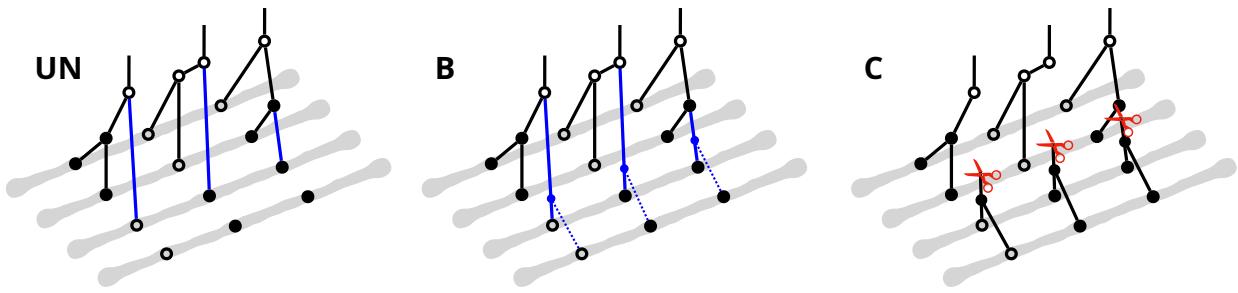


Figure 1 : Aperçu de la méthode. Les lignes grises représentent les haplotypes, tandis que les cercles indiquent les états alléliques des nœuds dans les arbres coalescents. Les cercles creux correspondent aux allèles ancestraux et les cercles pleins correspondent aux allèles dérivés. Dans les panneaux A, B et C, un ARG partiel pour les trois premiers haplotypes a déjà été construit et un quatrième haplotype est sur le point d'être enfilé sur cet ARG partiel. (A) L'étape initiale de l'enfilage du quatrième haplotype consiste à échantillonner la branche de jonction (surlignée en bleu) dans chaque arbre coalescent marginal de l'ARG partiel, un processus que nous appelons « échantillonnage de branche ». (B) Après la détermination des branches qui se joignent, l'étape suivante consiste à échantillonner le temps de jonction pour chacune de ces branches qui se joignent. Cette étape est appelée « échantillonnage temporel ». (C) Pour proposer une mise à jour d'un ARG dans MCMC, nous introduisons des coupes (illustrées par des ciseaux rouges) dans une séquence d'arbres coalescents marginaux pour élaguer les sous-arbres, puis les regreffons en résolvant le problème de filetage pour le sous-ARG ci-dessus les coupes. Cette proposition s'appelle « Élagage et regreffage de sous-graphes (SGPR) ».

**UN**

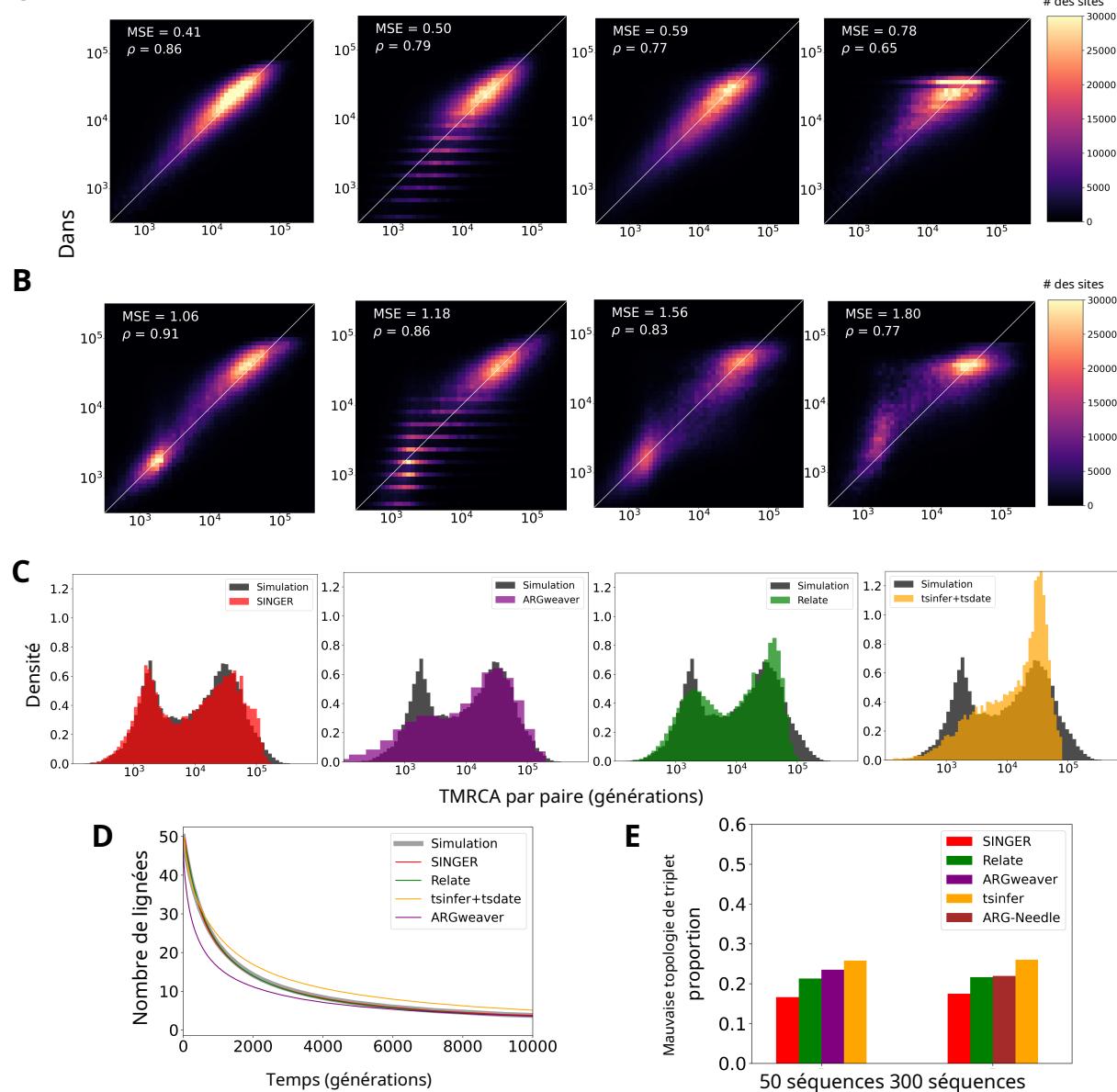


Figure 2 : Repères de performances sur le temps de coalescence et l'inférence de topologie. (A) Temps de coalescence par paire déduits par rapport à la vérité terrain dans des simulations impliquant 50 séquences dans un scénario de taille de population constante. (B) Semblable au panneau A mais pour les données simulées dans le cadre d'un historique de taille de population déduite pour le CEU. (C) Distribution déduite des temps de coalescence par paire (colorés) par rapport à la vérité terrain (noire) à partir de simulations sous la même démographie CEU que dans le panneau B. (D) Moyenne à l'échelle du génome du nombre de lignées en fonction du temps pour 50 séquences sous un historique de taille de population constante, comparées à la vérité terrain dans les simulations. (E) La proportion de topologies de triplet qui sont incorrectement déduites pour 50 et 300 séquences dans un historique de taille de population constante. En raison de contraintes d'exécution, ARGweaver n'est pas évalué pour 300 séquences.

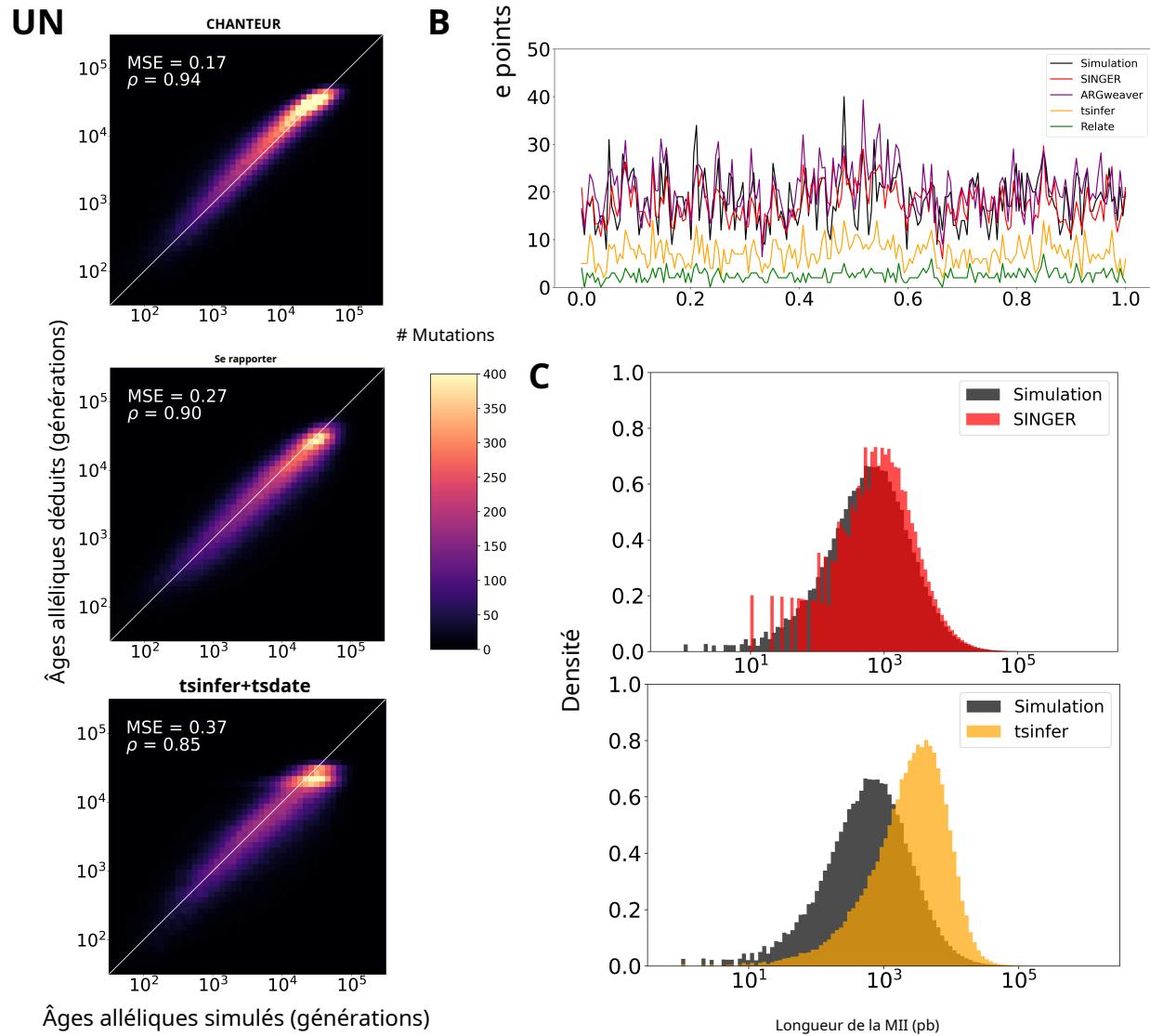


Figure 3 : Repères sur l'inférence de mutation et de recombinaison pour des données simulées avec 50 séquences et taille de population constante. (A) Âges des allèles déduits par rapport à la vérité terrain. (B) Nombre déduit de points d'arrêt de recombinaison dans des fenêtres génomiques de 5 Ko par rapport à la vérité terrain. (C) La distribution de longueur des IBD par paires dans les ARG déduits par rapport à la vérité terrain.

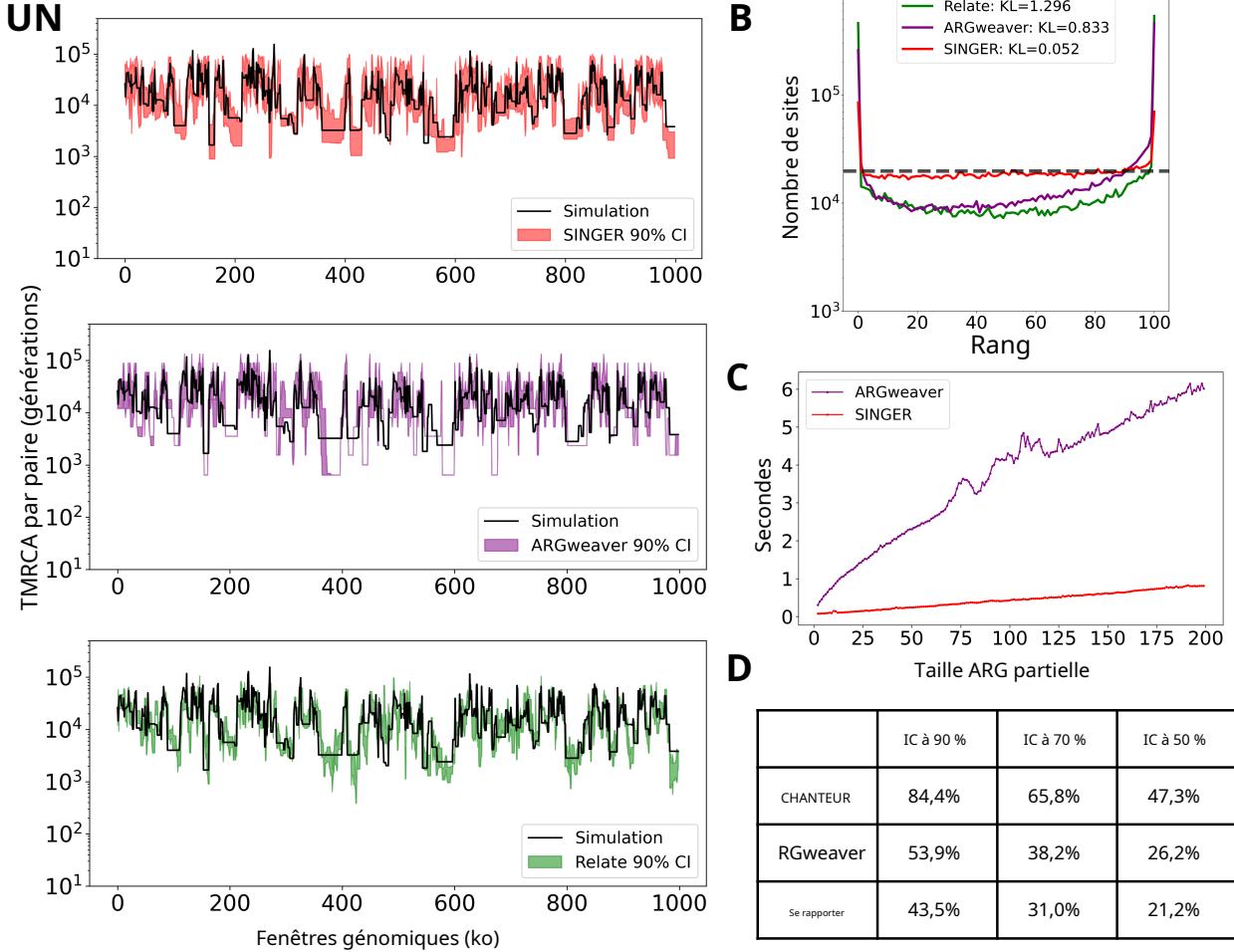


Figure 4 : Propriétés des échantillons et des environnements d'exécution ARG. (A) Intervalles empiriques crédibles à 90 % pour les temps de coalescence par paires, comme déduits par SINGER, ARGweaver et Relate. (B) Classer les tracés des temps de coalescence par paires dans les échantillons MCMC. Un échantillonneur parfait de la distribution postérieure obtiendrait la ligne pointillée plate, correspondant à la distribution uniforme. La divergence Kullback – Leibler (KL) est utilisée pour quantifier l'écart par rapport à la distribution uniforme. (C) Le temps d'exécution de l'algorithme de threading en fonction de la taille partielle de l'ARG (mesurée par le nombre de feuilles), pour ARGweaver et SINGER. (D) La couverture empirique du temps de coalescence par paire de la vérité terrain par l'intervalle crédible, pour différents niveaux nominaux.

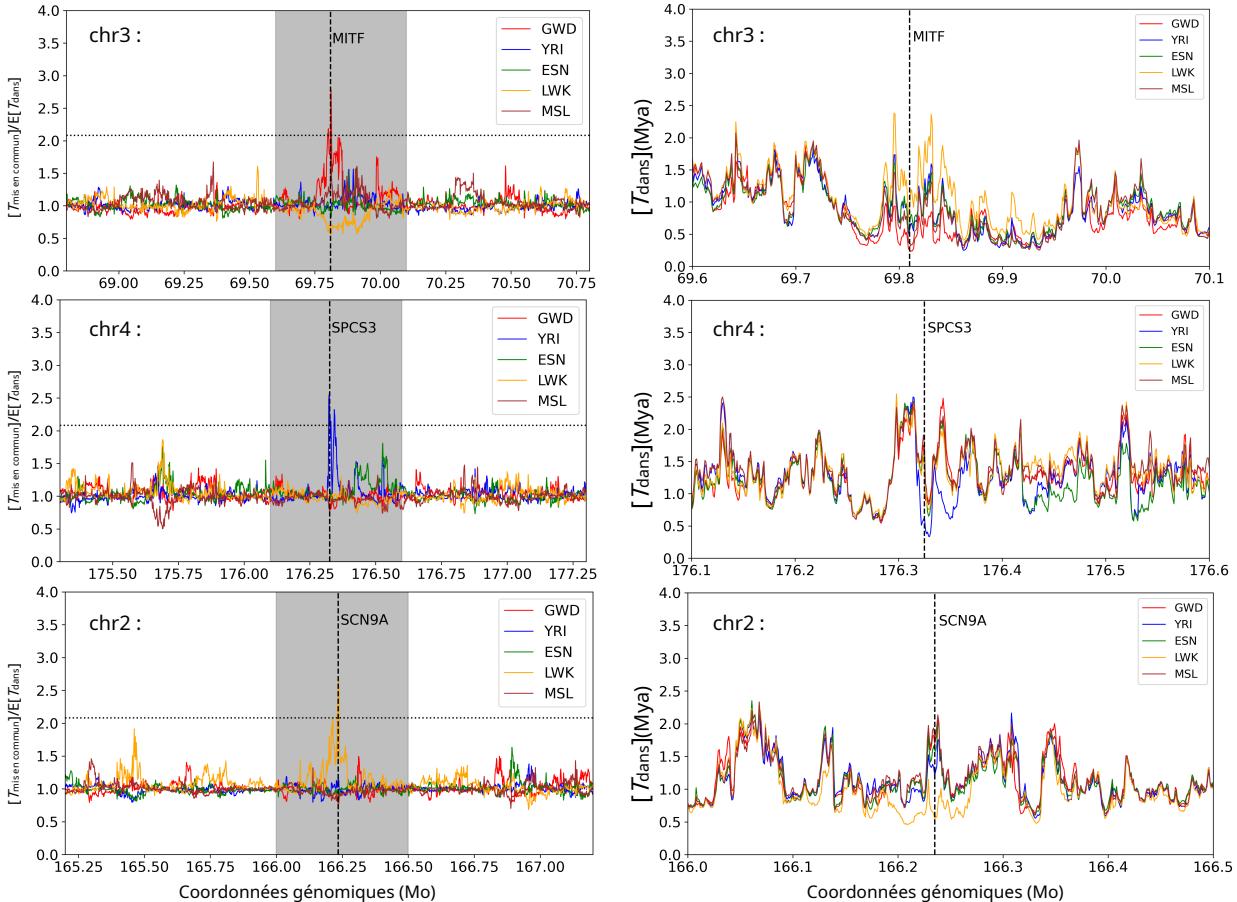


Figure 5 : Détection basée sur l'ARG de l'adaptation locale en Afrique. (A) Le rapport du temps moyen de coalescence par paire  $T_{\text{mis en commun}}$  dans l'échantillon regroupé (combinant les cinq populations) au temps de coalescence par paire moyen spécifique à la population  $T_{\text{dans}}$ , pour chaque fenêtre de 1 Ko. Dans chaque tracé, la ligne pointillée noire horizontale désigne le quantile à 99,99 % à l'échelle du génome, et la zone grisée correspond à une fenêtre de 50 Ko entourant le pic. Un pic significatif de ce rapport peut signaler une adaptation locale dans la population correspondante. Les positions de ces pics sont marquées par des lignes pointillées verticales et les gènes chevauchant ces signaux sont indiqués. (B) La moyenne  $T_{\text{dans}}$  pour chaque population, zoomé sur les régions grises mises en évidence dans le panneau A.

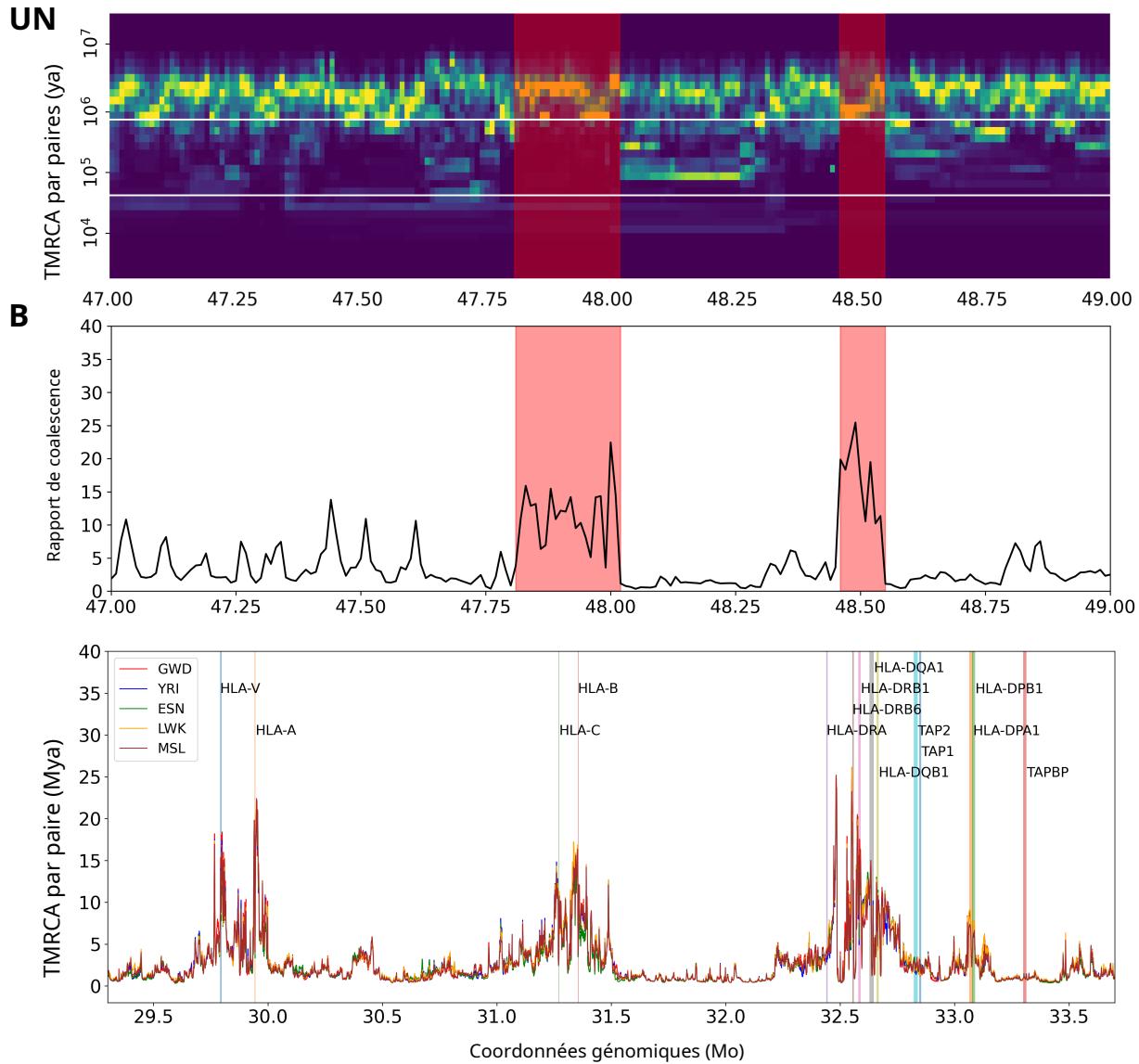


Figure 6 : Détection basée sur l'ARG des voies d'introgression archaïques et des signatures de sélection d'équilibre. (A) Identification de voies d'introgression archaïques potentielles. Pour un nœud feuille donné, ses temps de coalescence par paire avec tous les autres nœuds feuilles de l'arbre marginal sont résumés sous forme de distribution. Dans le graphique, chaque colonne représente une telle distribution à partir d'arbres marginaux dans une fenêtre de 10 Ko. Les deux lignes horizontales blanches délimitent l'intervalle entre le temps d'introgression et le temps intermédiaire. Un tract indiquant une introgression devrait présenter un épuisement des événements de coalescence au cours de cet intervalle et un enrichissement des événements de coalescence au-dessus du temps de partage. Les régions ombrées en rouge désignent des voies d'introgression putatives. (B) Le rapport de la densité de coalescence par paire au-dessus du temps de partage à celle comprise dans l'intervalle entre le temps d'introgression et le temps de partage. (C) Le temps moyen de coalescence par paire spécifique à une population  $T_{dans}$  dans le locus HLA. Il existe des signaux prononcés de polymorphisme trans-espèce et de sélection équilibrée pour plusieurs gènes HLA.