



HAL
open science

The role of admixture in the adaptation of human populations to changing environments

Sebastian Cuadros Espinoza

► **To cite this version:**

Sebastian Cuadros Espinoza. The role of admixture in the adaptation of human populations to changing environments. Populations and Evolution [q-bio.PE]. Sorbonne Université, 2022. English. NNT : 2022SORUS043 . tel-03715380

HAL Id: tel-03715380

<https://theses.hal.science/tel-03715380>

Submitted on 6 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sorbonne Université

Complexité du Vivant (ED 515)

Institut Pasteur, UMR 2000, Unité de Génétique Evolutive Humaine

**The role of admixture in the adaptation of human
populations to changing environments**

Par Sebastian Hans Cuadros Espinoza

Thèse de doctorat de Génétique des Populations

Dirigée par Etienne Patin

Présentée et soutenue publiquement le 29 mars 2022

Devant un jury composé de :

PR. CARINA SCHLEBUSCH

Rapporteur

PR. NICOLAS BIERNE

Rapporteur

DR. CECILE BERTHOULY-SALAZAR

Examinatrice

PR. RASMUS NIELSEN

Examineur

PR. GILLES FISCHER

Représentant de l'Université

PR. LLUIS QUINTANA-MURCI

Membre invité

DR. ETIENNE PATIN

Directeur de thèse



ନିମ୍ନ ଲିଖିତ କାର୍ଯ୍ୟକ୍ରମକୁ ଅନୁମୋଦିତ କରିବା ପାଇଁ ଆପଣଙ୍କୁ ଅନୁରୋଧ କରାଯାଉଛି।

Acknowledgements

First of all, I would like to thank all members of the jury: Nicolas Bierne, Carina Schlebusch, Cécile Berthouly-Salazar, Gilles Fischer and Rasmus Nielsen who have honoured me by evaluating this work and participating to my defence. It is a real pleasure to have you all as members of my jury.

I would then like to thank Lluís Quintana-Murci, for giving me the opportunity to do both my undergraduate internship and my doctoral project in his lab. It is truly a marvellous and exceptional work environment you got there Lluís, and I feel extremely lucky for having being part of it. You have my deepest gratitude.

I would also like to thank the members of my thesis committee, Paul Verdu, Olivier Tenaillon and Jean-Michel Gibert, for accepting of being part of it and for sharing their helpful advice during my PhD, thanks a lot!

I would also like to warmly thank those who put a smile on me when I go to work... (even if sometimes I could not really show it) my colleagues. And that goes for all of you, come and gone. Beginning with the 6th floor: Lara for being my local ancestry inference and population structure sensei; Javier for being my adaptive admixture Peruvian brother-in-arms (and for sharing that small yet golden Peruvian culture and slang nostalgia); Jacob for withstanding my complaints when I was feeling a bit down (that and my questionable-at-best taste in rap); Javi, Sara, Axel and Anthony for making these last months of PhD feel way more cheerful and less stressful than they should have been, and Guillaume, Mr. ABC himself, for being kind enough to share his infinite wisdom on Approximate Bayesian Computation and theoretical population genetics with me.

Then comes the less cool but nevertheless beloved 5th floor: Jérémy (alias Patte) and Alix for sharing their expertise on population models and coding as well as empathizing as fellow PhD strugglers; Maguelonne for kindly listening to me when I needed to “décompresser”; all the self-entitled “mean” office denizens: Marie, Lucas, Mary, Yann, Gaspard and Gaston for being the opposite of “mean” with me during my stay; and Christine, Aurélie, Li, Hélène and Maximus Decimus Meridius for being voices of reason amid shenaniganry. Seriously y’all, thanks for being so warmly welcoming, and for accepting me in all my wackiness,

weirdness and social awkwardness. I sincerely think it is the first time I have felt genuinely happy since leaving my homeland (took me six years to find it but I am glad it was with you).

And finally, of course, the boss himself, my PhD supervisor, Etienne Patin. I must say I have yet to figure your secret of being able to manage at least four completely different research projects and assisting to a bunch other scientific meeting all the while not eating or drinking anything during the whole day (I thought you were a photolithotroph but that idea was quickly discarded once I realized you worked very well shutters down... the chemolithotroph hypothesis needs to be explored further...). And to top it all off, you managed to do a bloody good job at everything. I feel so lucky for having worked with you all these past years. Sharing all the hardships and disappointments, but also all the excitements and surprises, it has truly been an adventure has it not? I guess having similar philosophies about work as well shared affinities towards scientific questions (and certain videogames) must have helped too. But let me tell you something I might have mentioned before, and I will do it again, you are inspirational. This has nothing to do with these shared affinities, nor with the titanic amount of work you can do on an everyday basis, because that is you, and is something I will probably never be able to pull off (after all, we can't have everything in this life, can't we?). No, I feel inspired by you for something you do and I, as well as any other person working in the field can do as well. And that is, no matter the project, or the funding, or the trends, or if the results are "novel" or not, to always strive towards two things: keep being curious and do sound science. And regarding that, I will always look up to you. So, thank you for showing me!

Summary

Acknowledgements	1
Summary	3
Résumé en langue française de la thèse	4
Introduction	4
Objectifs de la thèse	5
Résumé des résultats	6
Introduction	10
Chapter 1: Principles of population genetics.....	12
What are genes and how are they inherited?.....	12
Genes and populations: the concept of effective population size	13
The Hardy-Weinberg equilibrium model	14
Evolutionary forces	15
Modelling genes within populations	20
Chapter 2: Modelling and detecting positive selection	22
The classic sweep model	22
The soft sweep model.....	23
Detecting positive selection with allele frequency information.....	24
Detecting positive selection with haplotype information.....	26
Detecting positive selection: caveats and confounders	27
Chapter 3: Modelling and detecting admixture.....	29
The single pulse admixture model	29
Assessing admixture with allele frequency information (I).....	30
Assessing admixture with allele frequency information (II).....	32
Assessing admixture with haplotype information.....	34
Assessing admixture: towards more complex models	36
Chapter 4: Gene flow according to evolutionary biology	38
A barrier to speciation	38
A potential source of deleterious variation	38
A potential source of beneficial variation	39
Chapter 5: Admixture according to molecular anthropology.....	46
Initial studies in admixed populations.....	46
The ancient DNA revolution: admixture with archaic hominins	46
The ancient DNA revolution: admixture is everywhere	47
Adaptive admixture in modern humans	49
Chapter 6: Objectives of the thesis.....	52
Chapter 7: Results	53
Chapter 8: Discussion.....	112
Implications of the presented work: simulation analyses.....	112
Implications of the presented work: empirical data	112
Limitations of the presented work.....	114
Perspectives and future directions.....	115
References	117
Annexes	131
List of figures	155
List of tables.....	157

Résumé en langue française de la thèse

Introduction

L'histoire évolutive de l'espèce humaine est une histoire complexe, influencée par différents processus qui ont laissé leur trace sur les génomes des populations actuelles. Dès leur sortie du continent africain il y a plus de 60 000 ans, les populations humaines commencèrent à migrer et à se répandre sur toute la planète. Ces dispersions s'accrochèrent au cours des derniers 10 000 ans grâce à des avancées technologiques majeures, telles que l'agriculture et la domestication. Ces dispersions furent souvent accompagnées par des changements démographiques (changements de taille des populations, tels que des réductions de taille et des expansions), mais aussi par de mélanges génétiques (dits mélanges) avec des populations locales. Les avancées dans les champs de la génomique et de la paléogénomique ont révélé que des populations humaines ont rencontré et sont mélangées avec des populations d'homininés dits archaïques (par opposition à moderne) tels que Neandertal en Europe ou Denisova en Asie et Océanie, mais aussi que la plupart des populations actuelles sont le résultat d'événements de mélange complexes entre hommes modernes, à travers les dernières centaines de générations.

Les nombreuses migrations des populations humaines les ont mises en contact non seulement avec différentes populations, mais aussi avec différents environnements et contraintes environnementales, telles que des basses températures et faibles concentrations en oxygène, différentes ressources nutritionnelles limitées ou encore de nouveaux pathogènes. Des variations génétiques avantageuses ont augmenté en fréquence, par le biais de la sélection naturelle, et ont permis aux populations humaines de survivre et de s'adapter à ces nouvelles conditions. Une partie de ces variations génétiques avantageuses ne sont pas apparues dans les populations en expansion, mais ont été plutôt acquises par mélange avec d'autres populations. En effet, les mélanges avec Neandertal et Denisova ont apporté des adaptations génétiques affectant la réponse immunitaire, permettant ainsi aux populations humaines de mieux faire face aux nouveaux pathogènes rencontrés. Il a également été suggéré que les mélanges entre populations humaines modernes ont facilité la diffusion d'adaptations génétiques, telles que le catabolisme du lactose ou la résistance au paludisme dans des régions endémiques.

L'acquisition de mutations avantageuses à travers des événements de mélange entre populations, processus connu sous le terme de *mélange adaptatif*, n'est pas propre à l'espèce

humaine. Des exemples de métissages adaptatifs entre populations de différentes espèces (on parle alors d'*introgression adaptative*) ont été décrits chez des espèces animales, telles que le loup ou le moustique, mais aussi chez les plantes. Cependant, de tels exemples d'introgression sont relativement rares (du moins pour la plupart des espèces animales), et peuvent parfois se révéler infructueuses (à cause d'incompatibilités génétiques entre espèces différentes). Ceci n'est pas le cas pour le métissage entre populations humaines. Pourtant, le nombre de cas de métissages adaptatifs décrits dans la littérature est nettement inférieur au nombre de migrations et métissages qui ont été démontrés avec les données génétiques.

Le peu d'exemples de métissages adaptatifs au sein des population humaines vient en partie du fait que les méthodes de détection des signatures moléculaires du métissage adaptatif sont peu développées. En effet, la plupart de méthodes permettant d'inférer la sélection naturelle à partir de données génomiques font des hypothèses fortes sur les populations étudiées, et ne considèrent pas le cas des populations métissées. Cependant, les signatures moléculaires de sélection naturelle peuvent être différentes chez les populations métissées, voire même masquées par le métissage. Des méthodes spécifiques au cas des populations métissées existent, mais leur puissance statistique n'a jamais été évaluée en détails.

Objectifs de la thèse

C'est dans ce cadre que s'inscrit cette thèse. Les travaux présentés ici visent d'abord à caractériser les signatures moléculaires laissées par le métissage adaptatif, et à évaluer la puissance de différents types de méthodes pour les détecter. Pour cela, des génomes de populations humaines ont été simulés par ordinateur, tout en considérant des événements démographiques et de métissage complexes ainsi que d'autres considérations plus pratiques telles que la taille d'échantillonnage. A partir de ces génomes, les statistiques propres aux différentes méthodes ont été calculées et leur performance sous métissage adaptatif a été comparée. Ensuite, en utilisant les méthodes les plus performantes et les plus robustes aux facteurs générant des faux positifs, ces travaux ont permis de confirmer et inférer plusieurs événements de métissage adaptatifs au cours des derniers 5,000 ans, à partir de l'analyse des génomes d'au moins quinze populations métissées du monde entier.

Résumé des résultats

Pour déterminer les signatures moléculaires du métissage adaptatif, nous avons tout d'abord considéré trois types de scénarios de métissage avec sélection. Un premier scénario suppose qu'une variation génétique avantageuse apparaît par mutation dans une population et est ensuite transmise par métissage à une autre population (dite population source ; *scénario 1*). Le deuxième scénario suppose que cette variation génétique n'est plus avantageuse dans la population receveuse (*scénario 2*). Enfin le troisième scénario assume qu'une variation génétique apparaît par mutation dans une population et est transmise par métissage à une autre population, mais ne devient avantageuse que dans cette dernière (*scénario 3*).

Conceptuellement, le *scénario 1* correspond à la définition formelle de métissage adaptatif. Le *scénario 2* peut être vu comme une source de faux positifs, lorsqu'on cherche à détecter le métissage adaptatif, puisque les signaux de sélection naturelle propres à la population source peuvent se répercuter sur les génomes de la population métissée, sans pour autant que la mutation soit sélectionnée chez cette dernière. Le *scénario 3* correspond à un cas alternatif de métissage adaptatif, où la diversité génétique nouvellement acquise par métissage devient avantageuse uniquement à partir du métissage. Une fois ces trois cas de figure établis, nous avons évalué la performance de deux classes de méthodes permettant de détecter des mutations sous sélection positive. La première classe de méthodes « classiques » se fonde sur des comparaisons de fréquences alléliques entre populations (F_{ST}) ou la conservation d'homozygotie d'haplotypes (iHS), mais considèrent uniquement des populations homogènes, non métissées. La deuxième classe regroupe des méthodes adaptées aux populations métissées. Une première méthode, F_{adm} , se base sur la comparaison entre les fréquences alléliques observées et attendues dans la population métissée, l'attendu étant estimé à partir des fréquences alléliques dans les populations sources ayant contribué au métissage ; des fortes déviations de l'attendu théorique peuvent être interprétés comme un signal de métissage adaptatif. La seconde méthode, LAD, se base sur l'inférence de l'origine génétique locale (segments chromosomiques inférés comme provenant d'une des sources ayant contribué au métissage) le long des génomes des individus métissés, à partir de données haplotypiques ; des excès génomiques d'une origine génétique locale associée à une population source donnée peuvent constituer une preuve de métissage adaptatif.

Nous avons montré que les méthodes dites « classiques » se comportent de façon similaire dans les trois scénarios. Ces méthodes ne peuvent pas différencier des vrais signaux de

métissage adaptatif (*scénario 1* et *scénario 3*) de faux signaux provenant d'une population source (*scénario 2*). A l'inverse, les méthodes spécifiques aux populations métissées (F_{adm} et LAD) se comportent comme voulu : elles sont puissantes sous les *scénarios 1* et *3* mais n'ont aucune puissance pour détecter des signaux sous le *scénario 2*. Elles détectent donc spécifiquement les scénarios de sélection positive dans la population métissée, raison pour laquelle nous avons gardé uniquement ces deux méthodes par la suite.

Nous avons ensuite évalué la puissance de détection de F_{adm} et LAD dans différents scénarios de métissage adaptatif. En commençant par des considérations pratiques, nous avons étudié l'effet de la taille d'échantillon de la population métissée, mais aussi des populations considérées comme ayant contribué au métissage (c.-à-d., les populations sources). Nous avons constaté que, lorsque la population métissée a des petites tailles d'échantillon, la puissance de détection des deux statistiques diminue jusqu'à 40%, mais qu'uniquement F_{adm} était affecté par des petites tailles d'échantillon pour les populations sources. En effet, la puissance de LAD demeure inchangée même pour des échantillons de l'ordre de 20 individus.

Étant donné qu'il est souvent difficile d'obtenir des données génotypiques pour les véritables populations sources d'une population métissée, les généticiens des populations utilisent souvent comme substituts (appelées proxy) des populations actuelles apparentées, ce qui peut conduire à de faux signaux de métissage adaptatif. Nous avons étudié comment la puissance de détection est affectée par la divergence entre la véritable population source et une population proxy. Nous avons observé une différence de performance entre F_{adm} et LAD, ce dernier étant plus robuste à l'utilisation d'un proxy. LAD maintient un pouvoir de détection similaire même si le proxy et la source sont considérablement différents, alors que le pouvoir diminue de 25% pour F_{adm} .

Plusieurs études ont montré que les métissages entre populations humaines ont rarement été instantanés, comme on le suppose souvent par simplicité, et ont souvent impliqué plus de deux populations sources. Nous avons donc estimé la performance de F_{adm} et LAD sous des modèles de métissage plus complexes. Nous avons constaté que la puissance de détection n'est que modérément réduite dans le cadre d'un modèle de métissage continu sur plusieurs générations, ce qui suggère que nos estimations de puissance sont valables pour une variété de modèles de métissage.

En assumant un modèle de métissage instantané, nous avons examiné comment la puissance de détection est affectée par des paramètres clés du modèle de métissage adaptatif, tels que la force de la sélection, l'âge de l'événement de métissage et les taux de contribution génétique des populations sources. Ainsi, nous avons constaté que la puissance de détection augmente avec la force de la sélection, mais qu'elle est aussi déterminée par l'âge de l'événement de métissage. En effet plus celui est vieux, plus la sélection naturelle a du temps pour agir sur la variation avantageuse, et plus le signal sera fort. On observe ainsi des niveaux de puissance 4 fois plus élevés pour des métissages vieux de 2,000 ans, par rapport à des métissages vieux de 500 ans. De manière intéressante, nous avons observé que plus la contribution de la population source amenant l'adaptation génétique est importante, plus la puissance de détection est réduite.

Nous avons également estimé la puissance dans des scénarios où la démographie s'écarte d'un modèle de taille de population constante. En effet, il a été démontré que les événements démographiques, tels que les goulots d'étranglement, modifient les performances de plusieurs statistiques de neutralité. Nous avons constaté que la puissance de détection est peu affectée dans tous les modèles de population en expansion. En revanche, celle-ci est réduite de 50% dans des scénarios avec un fort goulot d'étranglement dans la population métissée, par rapport au scénario de taille constante. Enfin, la performance de F_{adm} et de LAD est peu affectée lorsqu'un goulot d'étranglement est introduit dans les populations sources.

Nous avons ensuite cherché à détecter des gènes candidats sous métissage adaptatif, en analysant, à l'aide de F_{adm} et LAD, les génomes de 15 populations métissées dans le monde, qui ont connu au moins un événement de mélange au cours des 5 000 dernières années. Pour maximiser la puissance de détection et faciliter la priorisation des gènes candidats, nous avons combiné les deux statistiques avec la méthode de Fisher. Fait important, nous avons constaté que la méthode de Fisher augmente le pouvoir de détection dans les scénarios défavorables, notamment lorsque la population métissée a subi un fort goulot d'étranglement, lorsque le métissage est relativement récent ou lorsqu'on utilise une population proxy fortement divergée de la vraie population source. Étant données les connaissances limitées sur les tailles historiques des populations étudiées, ce qui pourrait générer des faux positifs, nous avons appliqué la correction conservatrice de Bonferroni sur les valeurs obtenues par la méthode de Fisher.

Nos analyses ont permis de confirmer un certain nombre de signaux de métissage adaptatif précédemment rapportés. Parmi ceux-ci, nous avons trouvé le locus *HLA* de classe II dans les populations de langues bantoues du Gabon, le locus *HLA* de classe I chez les Mexicains, le locus *LCT/MCM6* associé à la persistance de la lactase chez les nomades Fulani du Burkina Faso et le gène *ACKRI* dans les populations d'origine africaine de Madagascar, du Sahel et du Pakistan. Ces résultats confirment ainsi que notre approche peut récupérer des signaux forts et bien documentés de métissage adaptatif.

Nous avons aussi identifié plusieurs nouvelles régions génomiques sous métissage adaptatif, parmi lesquelles le locus *MYH9/APOL1* chez les nomades Fulani. Des variations génétiques dans *APOL1* confèrent à la fois une protection contre la trypanosomiase humaine africaine et une susceptibilité aux maladies rénales communes chez les personnes d'origine africaine. Une autre région candidate contient le gène *PKN2*, qui présente un fort excès d'ascendance génétique papoue chez les Indonésiens de l'Est. *PKN2* joue un rôle dans les réponses de transduction des signaux cellulaires et serait impliqué dans la régulation du métabolisme du glucose dans les muscles squelettiques. Un dernier signal a été détecté dans *CXCL13* chez les pasteurs Nama d'Afrique du Sud. Le locus *CNOT6L/CXCL13* a déjà été trouvé comme étant associé de manière suggestive au risque de tuberculose dans les populations sud-africaines d'ascendance San. Cependant, nous avons constaté que ce signal pourrait être un faux positif, causé par l'utilisation d'un proxy de la population source San.

Enfin, nous avons détecté des signaux proches du seuil de détection, suggérant un métissage adaptatif sur des gènes précédemment décrits comme étant sous sélection positive, notamment le locus *MCM6/LCT* chez les Bakiga de langue bantoue en Ouganda et *TNFAIP3* chez les Indonésiens de l'Est. Ce dernier a non seulement été trouvé comme évoluant sous sélection positive chez les Papous mais aussi sous introgression adaptative issue de Denisova. *TNFAIP3* joue un rôle important dans la tolérance immunitaire de l'homme aux infections. Collectivement, ces résultats indiquent que des événements de métissage adaptatif ont eu lieu dans diverses populations métissées à travers le monde, et mettent en évidence le système immunitaire et le métabolisme des nutriments comme cibles importantes d'une adaptation génétique récente.

Introduction

Admixture (*noun*): Something that is added to something else.

The above definition comes from chemistry and is used commonly across multiple chemicals, structural, and engineering disciplines to refer to the product that results from the combination of various compounds, molecules, and substances. In the context of this work, it also refers to “something” that is added to “something else”, that “something” being individuals, or rather their genetic material. Although the definition of *admixture* in this context refers to a specific population genetics model, the concept behind this definition arose well before the field of population genetics itself, albeit erroneously and in a much darker context.

In Spanish, one early yet arguable translation of the word was “mestizaje” coming from the “mestizo” referring to the descendants of European settlers and Native Americans during the colonial period. An even more vicious term, miscegenation, was coined in a propaganda pamphlet published in 1863, during the American Civil War. Although the pamphlet itself was a hoax, meaning to sink Abraham Lincoln’s presidential campaign (presidential elections in 1864), the word continued to be used by advocates of racial superiority and purity. Much to their disappointment however, the first analysis of genetic data from human groups from all around the world in the 1970s would show that a genetic basis for racial purity, or for that matter, the concept of genetically defined races themselves, was non-existent. What they found instead was that most of the genetic variability in humans is found within populations rather than between populations. Furthermore, it was later shown, thanks to the advances in the extraction, purification and sequencing of DNA from ancient biological samples, that human groups met and interbred with other groups quite often since the very early stages of human history, even with other hominin species like Neanderthal. Genetic admixture has thus always been part of human history, and all humans are, by definition, admixed.

Our long history of admixture likely played an important role in genetic adaptation as groups of humans encountered environments with novel topographical (high altitude), chemical or microbial pressures. In these contexts, admixture provided the opportunity for inheritance of advantageous mutations. This phenomenon, termed *adaptive admixture*, is at the core of the work presented here, but to properly understand it, and study it, key concepts in the fields of population genetics, molecular anthropology and evolutionary biology are required. The next

chapters will thus be focused on introducing these concepts starting with population genetics, more precisely what we understand by genes, populations and genes within populations.

Chapter 1: Principles of population genetics

What are genes and how are they inherited?

The “proper” definition for the term *gene* could very well have its own chapter or even manuscript, however for the sake of simplicity (and convenience), we shall use the following one: *a basic unit of heredity that has a molecular support, DNA, that encodes the synthesis of a gene product*. Other definitions exist, much more focused on the molecular and functional structure, but this one will suffice. In fact, some would argue that the first phrase alone would suffice as it evokes the key concept of heredity. Indeed, it was through the study of trait inheritance that the concept of a gene was first elucidated by Gregor Mendel, who took a particular interest in garden peas, specifically in how variation in shapes and colours were transmitted from parents to offspring. Mendel’s Laws of inheritance were first published in 1866, and while they are now considered to be a seminal work in genetics, they were initially met with criticism and garnered little interest from the scientific community at the time, cited only three times in the thirty-five years following publication. Nevertheless, thanks to the work of Hugo de Vries and Carl Correns, Mendel’s Laws were rediscovered in the 1900s. These can be summed up as follows:

- Law of Segregation: variants of a gene, also called alleles, segregate into gametes. This means that, in diploid species, a heterozygous parent, carrying one copy of each allele, will produce gametes carrying either allele in equal frequency, and offspring are produced by the random union of parents’ gametes.
- Law of Independent Assortment: alleles from different genes assort independently into gametes. An important exception is genes that are linked, meaning close to each other in the genome (the concept of linkage is of particular importance and will be discussed further in Chapters 2 and 3).
- Law of Dominance and Uniformity: some alleles are dominant while others are recessive. A heterozygous parent will display the effects of the dominant allele. Important exceptions are alternative forms of dominance such as codominance or incomplete dominance.

Combined, these Laws define Mendelian inheritance. A few traits are known to follow Mendelian inheritance, but many traits have a much more complex genetic basis, namely

quantitative traits (in contrast to the discrete traits studied by Mendel). Furthermore, most traits are also influenced by the environment and such genotype-by-environment interactions are often non-trivial. Nevertheless, Mendel's Laws provide the necessary hypothetical framework for the later mathematical developments made by Fisher, Wright and Haldane that will become the basis for the field of population genetics. At this stage it becomes necessary, thus, to define what is meant by a population in the genetic sense.

Genes and populations: the concept of effective population size

Mark Stoneking's introductory book on molecular anthropology (Stoneking, 2016) defines a population as: "a spatial-temporal group of interbreeding individuals who share a common gene pool." Broken down, this describes a population of interbreeding individuals occupying a specific geographic area, during a specific time, and whose total collection of alleles can be referred to as a gene pool. This definition, while arguably imprecise, according to Stoneking himself (Stoneking, 2016), introduces the concept of a gene pool, which is a key variable in all evolutionary genetic models, or rather the size of said gene pool. Population geneticists often root models in reference to an "ideal" population, which is defined by the following criteria:

- Equal numbers of males and females
- All individuals are unrelated
- Equal probability of producing offspring for all individuals
- Constant size

Most natural populations do not match any of these criteria. For instance, some individuals can have a higher chance of having children than others, thus contributing more alleles to the next generation. Some individuals may also be related, thus having more alleles in common than if they were unrelated. In reality, there is less genetic variation than we would expect given the size of a population, and therefore the size of the gene pool transmitted to the next generation is always smaller than if the population was to be an ideal population. The notion of effective population size, noted as N_e , is thus introduced to link the ideal population with the real one. More precisely, it is the size of an ideal population whose gene pool would correspond to the one observed on the real population. It cannot be stressed enough how crucial

N_e is for population and evolutionary genetics, as the effects of all evolutionary forces can be written down as a function of N_e and it enables one to determine the impact that these forces have had or could have on the genetic variability of a population.

The Hardy-Weinberg equilibrium model

One way to understand how evolutionary forces can affect the gene pool of a population is to define a simple, yet highly unrealistic, model and see how the gene pool changes over time. Just as for the ideal population model, violations in the assumptions of this model would result in deviations from expected values, which are informative on what evolutionary forces may be operating. The values here are allele frequencies, or rather genotype frequencies, and the model assumptions of the studied population are the following:

- Discrete, non-overlapping generations
- Random mating
- Infinite population size
- No migration into or out of the population
- No new mutations occur
- The chances to survive to reproductive age and reproduce do not depend on the genotype

If we denote p and q the frequency of alleles A1 and A2 in a bi-allelic locus, then under these assumptions, the frequency of the genotypes A1|A1, A1|A2 and A2|A2 are p^2 , $2pq$ and q^2 respectively, and more importantly, these genotype frequencies, as well as the frequencies for the two alleles, do not change over time. These two facts are known as the Hardy-Weinberg principle, from G. H. Hardy and Wilhelm Weinberg, who independently published their results in 1908 in different journals(Hardy, 1908; Weinberg, 1908).

Evolutionary forces

Of the different evolutionary forces that violate the assumptions of the Hardy-Weinberg principle, we focus our discussion here on mutation, genetic drift, migration, and natural selection. The first evolutionary force, genetic mutation, is inherent to the molecular nature of DNA, as mutations naturally occur as errors during the replication and/or repair of DNA. Mutations can also occur if DNA is exposed to certain mutagenic agents, however the focus of the research here will consider these environmental mutagens as negligible factors. One particularity of genetic mutation is that rates of occurrence are extremely low (in humans, it is estimated that a mutation occurs with a probability of approximately 10^{-8} per base pair (Campbell et al., 2012; Lipson et al., 2015)), due to the high fidelity of the DNA replication and repair machinery. Because of this, the time it takes for allele frequencies to arrive to the new equilibrium imposed by mutation is extremely long, and the observable effect of mutation on allele frequencies is almost non-existent. Nevertheless, because it is the only evolutionary force that creates new variation, without it the other forces would not have anything to act on and evolution would not be possible.

Next in order of impact is genetic drift, which is a direct consequence of the infinite population size assumption being violated. In a finite-sized population, fluctuations in allele frequencies will occur due to random sampling of new individuals at each generation. In fact, the original definition of N_e , provided by Sewall Wright in 1931 (Wright, 1931), was introduced as a way of calculating genetic drift in a finite population. Low effective population sizes result in higher levels of genetic drift, which themselves result in accelerated loss of genetic variation and increased homozygosity due to inbreeding (Fig 1).

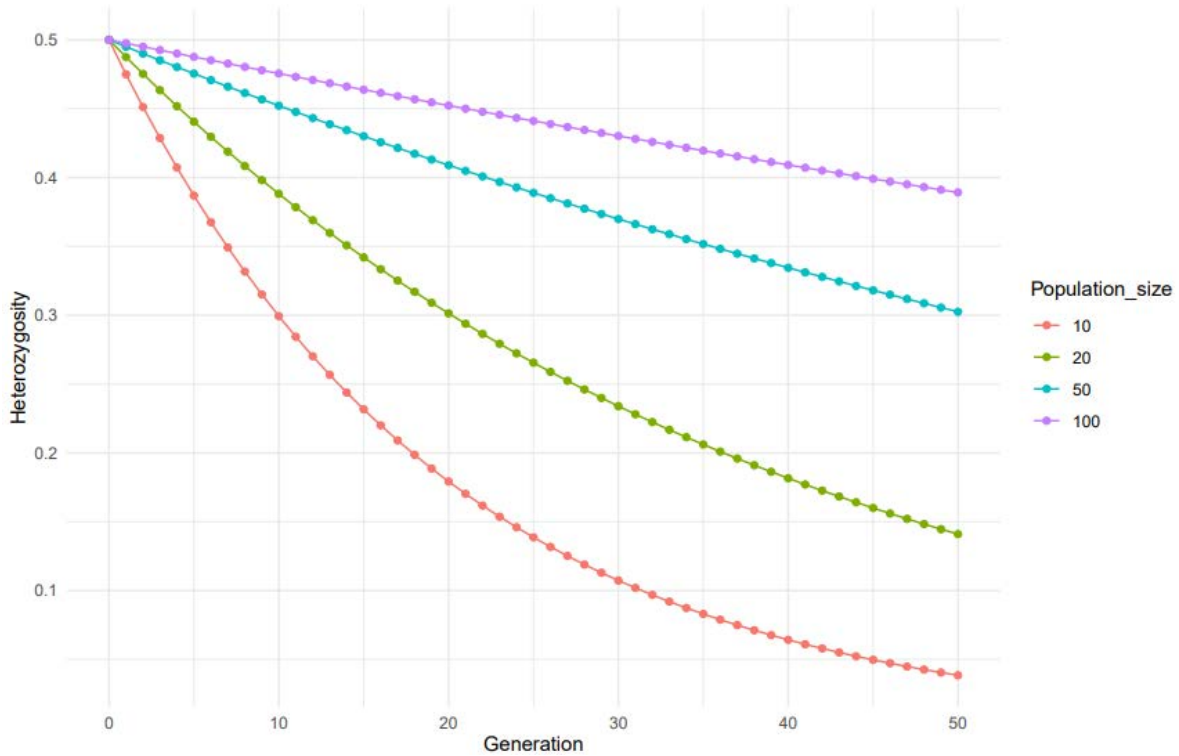


Fig. 1: Theoretical loss of heterozygosity over time (generations) as a function of population size.

An interesting property of genetic drift is how it balances with the other evolutionary forces in equilibria that can be quantitatively described. For instance, *eq. 1* describes the estimated heterozygosity at mutation-drift equilibrium, with μ being the mutation rate.

$$\hat{H} = \frac{4N_e\mu}{4N_e\mu + 1} \quad (\text{eq. 1})$$

The next two forces, migration and natural selection, are at the core of this thesis and will have their own chapter dedicated on how to model them and estimate them using genetic data from populations (Chapter 2 and Chapter 3 respectively). Here, however, a brief summary of the two will be made together with the description of equilibria with genetic drift.

We will start with migration, also known as gene flow, which is the movement of individuals (and by extension, alleles) between populations. In a simple model of two populations where one population acts as a source of individuals and the other as a sink, receiving the individuals, if we denote P, Q and p, q the frequencies of two alleles of a biallelic locus in the source and sink populations respectively, then eventually p, q will converge to P, Q . In a model with more populations, where all populations exchange individuals at the same rate

and reciprocally (also known as an island model), then the allele frequency in each population will eventually be identical to the average allele frequency over all populations. Gene flow will always reduce the allele frequency differences between populations, which over time will converge towards an average value. When genetic drift is taken into account, the migration-drift equilibrium can be described by eq. 2, where m is the migration rate and F the inbreeding coefficient, that is the probability of two alleles being identical by descent. It only takes a few migrants (scaled at the population level) to reduce F , or the effect of drift in the gene pool of a population.

$$\hat{F} = \frac{1}{4N_e m + 1} \quad (\text{eq. 2})$$

The final force to be considered, natural selection, operates at the phenotypic level but ultimately affects the gene pool. More precisely, selection operates in a way that individuals who carry certain alleles will be more fit to a certain environment and will have better chances of surviving and reproducing. By extension, their alleles will be more represented in the population gene pool. Overall, there are three modes of selection: directional selection, balancing selection, and disruptive selection. All three can be described through a simple framework, using a biallelic locus and assigning fitness (noted W) values for each of the genotypes: $W_{A1|A1} = 1$, $W_{A1|A2} = 1 + hs$, and $W_{A2|A2} = 1 + s$, where s is the selection coefficient and h is the degree of dominance (in the case of recessive inheritance, $h = 0$). These two parameters describe the mode of selection operating as well as how fast the alleles will attain their equilibrium frequency in a population (see Fig 2). In the case of directional selection, $|s| > 0$ and $|hs| < |s|$, meaning the fitness of the genotypes $A1|A2$ and $A2|A2$ are higher (or lower) than the remaining genotype. This ultimately results in the frequency of the $A2$ allele to increase (or decrease) in the population until reaching its equilibrium value, which in this case is either 1 (fixation) or 0 (loss). For balancing and disruptive selection however, $|hs| > |s|$, meaning the fitness of heterozygotes is higher (balancing) or lower (disruptive) than the fitness of the homozygotes. Because of this, the equilibrium frequency is neither 0 nor 1, and is instead defined by eq. 3, depending only on the dominance coefficient h .

$$\hat{p} = \frac{1-h}{1-2h} \quad (\text{eq. 3})$$

Nevertheless, in the case of disruptive selection, this is an “unstable” equilibrium, and any shift in allele frequency will eventually lead to either the fixation or the loss of the allele. Among the three modes of selection described, we will focus only on directional selection, more specifically on positive selection, which is one of the components of adaptive admixture, the other being admixture of course. On an additional, yet rather important, note, selection operates more or less efficiently depending on the levels of genetic drift. For example, in a population with low effective population size, changes in frequency for an adaptive allele will likely be governed by chance rather than by positive selection, whereas it would be the opposite in a population with a large effective population size. Because of this, the strength of natural selection is evaluated with a population-scaled selection coefficient: $N_e s$.

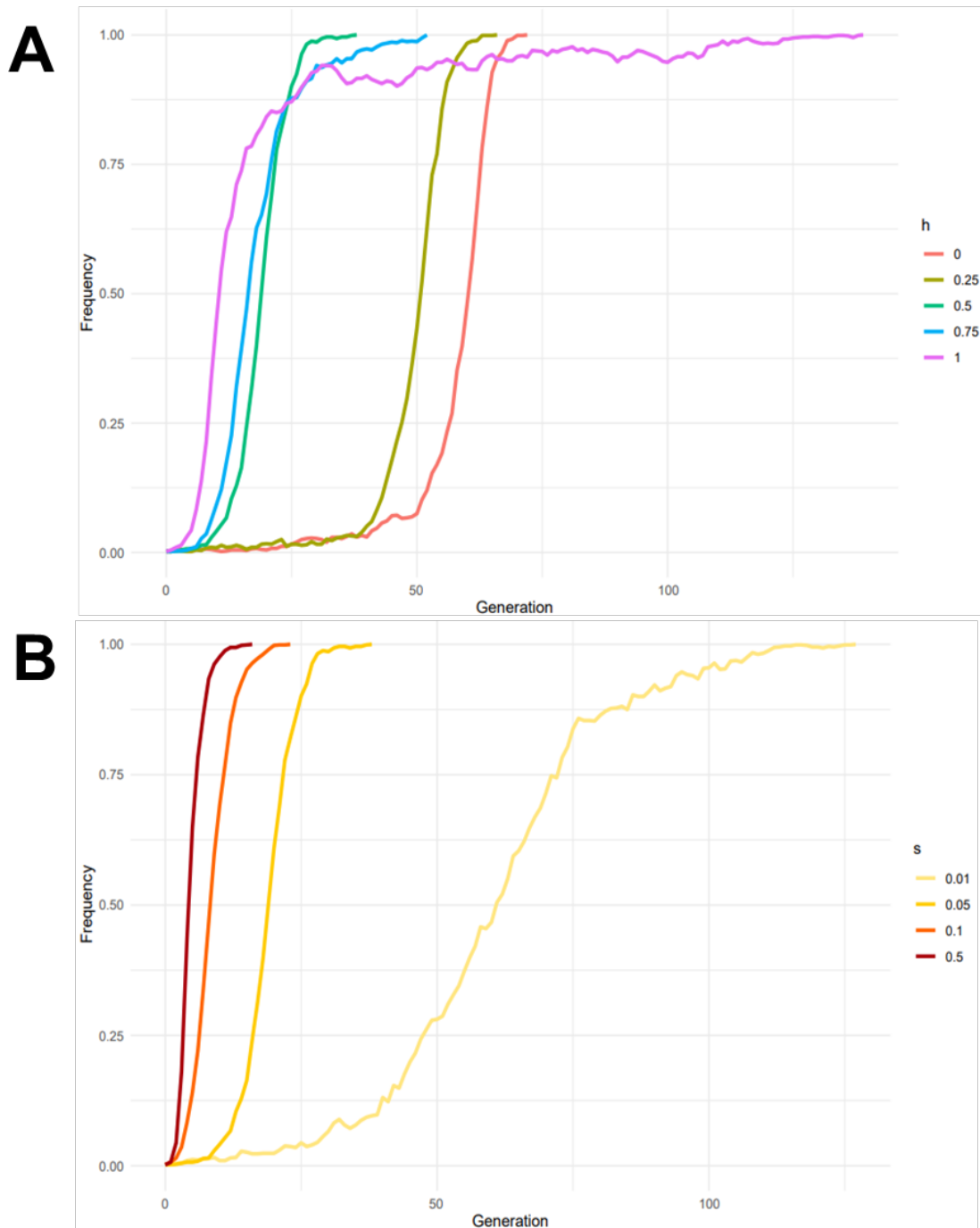


Fig. 2: Effect of (A) the dominance and (B) the selection coefficients on the frequency trajectory of an allele under positive selection, based on forward Wright-Fisher simulations. The selection coefficient determines the time it takes for the allele to go from very low frequency to very high frequency (i.e., the slope of the mid-section of the curve). The dominance coefficient, on the other hand, determines the time it takes for the allele to start sharply increasing in frequency, as well as the time it takes to arrive from high frequency to fixation. Simulations were computed using the SLiM engine (Haller and Messer, 2019) and following recipe 9.11 of the SLiM manual (Haller and Messer, 2016).

Modelling genes within populations

Before exploring how to model selection and migration, a general model describing how gene (or allele, to be precise) frequencies fluctuate within a population of finite size (by contrast to the Hardy-Weinberg model) is needed. The Wright-Fisher model (Fisher, 1952, 1922; Wright, 1931) provides exactly that, i.e., a forward in time mathematical description of random genetic drift. The model assumptions are the following:

- The population is panmictic (i.e all individuals can reproduce with every other individual with the same probability)
- Generations are non-overlapping (all individuals from one generation reproduce simultaneously and only contribute to the next generation).
- At each generation, offspring genotypes are formed by random sampling with replacement from individual genotypes of the previous generation.

Under such a model Given a biallelic locus in a haploid population of size N and noting the p and q the frequency of alleles A_1 and A_2 at the current generation, the probability that the A_1 allele will be present in k copies in the next generation is then defined by eq. 4

$$\binom{N}{k} p^k q^{N-k} \quad (\text{eq. 4})$$

Where $\binom{N}{k}$ denotes the binomial coefficient $\frac{N!}{k!(N-k)!}$ Mathematically speaking, eq.4

describes a binomial distribution of parameters N and p . Indeed, the sampling of each offspring from the previous generation of individuals is a Bernoulli trial, where a success is counted if the offspring carries the A_1 allele (which probability is p , the frequency of A_1 in the current generation). Because each trial is independent, the number of successes in N trials (the size of the population, since it is replaced at each generation) can be modelled by a binomial distribution. And because of this, two important properties of the Wright-Fisher model can be derived. First, the expectation of the number of A_1 alleles at each generation is equal to Np , and the frequency of the A_1 allele is equal to p . In other words, allele frequencies in a Wright-Fisher model do not change, on average, through generations. Second, the variance of the number of A_1 alleles is equal to Npq , and thus the variance of the frequency of the A_1 allele is equal to

$\frac{pq}{N}$. In other words, allele frequencies fluctuate between generations, with intensity inversely proportional to the population size. These results can be generalized to diploid systems (replacing N by $2N$), to non-constant population sizes (taking smaller or bigger samples at each generation), to multiallelic-sites (by using a multinomial distribution instead of a binomial distribution) and to multiple generations. For the latter, the variance of allele frequencies is proportional to $\frac{t}{N}$, with t being the number of generations. The Wright-Fisher can also integrate selection, mutation and recombination and there are even extensions of it to include non-overlapping generations (Moran, 1958). In the context of the work presented here, the default Wright-Fisher will be used for all analyses involving simulations.

Chapter 2: Modelling and detecting positive selection

The classic sweep model

Formally introduced in by Maynard Smith and Haigh in 1974 (Smith and Haigh, 1974), the classic sweep model provides a simple, yet useful description of the effect that positive selection has on nearby, neutral variation: a reduction in average heterozygosity. The model assumes that selection acts on a newly arising mutation (also known as *de novo* mutation), so that when it increases in frequency, closely linked alleles will also increase in frequency. This effect of selection on linked sites is known as “hitchhiking” and it is at the core of many methods used to infer positive selection, as noted later in this chapter. In the end, because there is a single starting copy of the beneficial mutation and because of hitchhiking, there will be only a single set of linked alleles (also known as haplotype) that will rise in frequency, therefore creating a sharp reduction in genetic diversity (thus the name “sweep”), although for these effects to be observed, the model implicitly assumes a relatively strong selection coefficient. The mutation in question must thus arise at the right moment, provide a strong increase in fitness to its carriers and, on top of that, not be eliminated by genetic drift (initial frequency of $1/2N$ in a diploid population). This is quite a lot of assumptions; consequently, classic sweep signals are expected to be rare. In humans for instance, the number of classic sweeps is estimated to be lower than 60 (Laval et al., 2021; Schrider and Kern, 2017). However, positive selection may also leave other types of signatures in the genome, for example when it acts on alleles already present at intermediate frequency in a population.

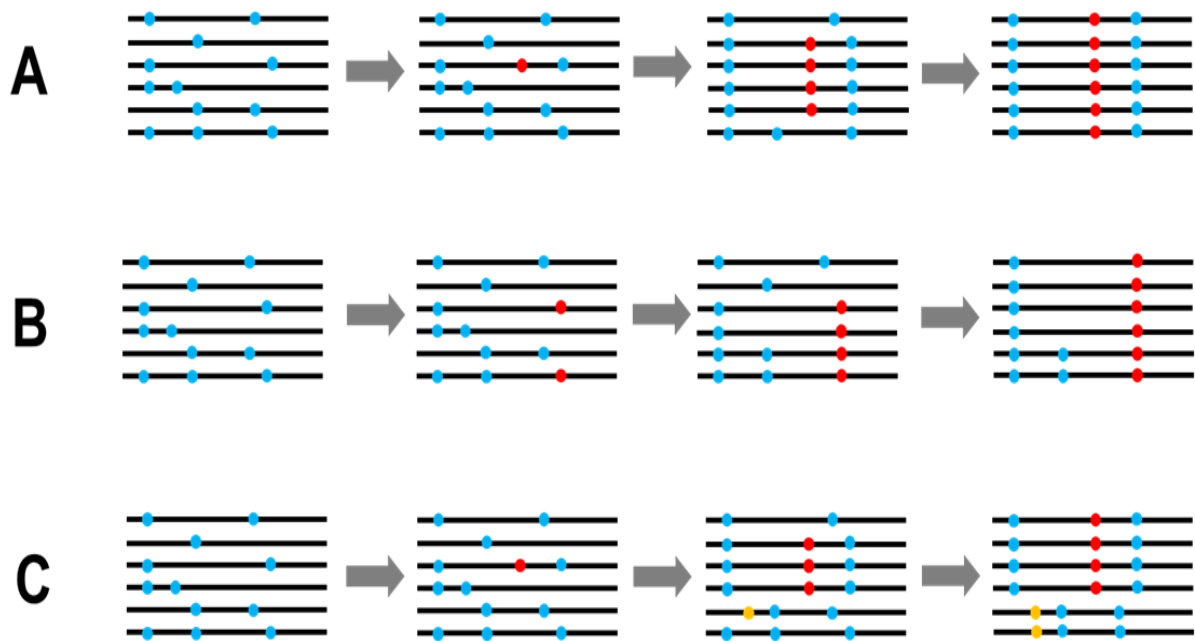


Fig. 3: Different types of selective sweeps. (A) Classic sweep: a beneficial *de novo* mutation (red dot) appears on one haplotype, and subsequently increases in frequency, bringing closed linked neutral alleles (light blue dots) to high frequencies too, until eventually reaching fixation. (B) Soft sweep from standing variation: an already existing mutation present on multiple haplotypes, becomes beneficial (red dots), bringing these haplotypes at high frequency. (C) Soft sweep from recurring mutations: similar to (A), but a second, beneficial mutation (yellow dot) appears on a different haplotype from that of the first mutation, which also rises in frequency. Gray arrows represent the passage of time.

The soft sweep model

Through a series of seminal papers in 2005 and 2006, Hermisson and Pennings challenged the classic sweep model by arguing that selection could also operate on alleles already present in a population, referred to as standing genetic variation (Fig. 3 B) (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006a, 2006b). The pattern produced by selection acting on standing variation would then be one of maintained levels of genetic variation. Indeed, because the beneficial mutation is present in multiple copies, there are multiple haplotypes that rise in frequency and surrounding genetic variation would not disappear as is the case in a classic, hard sweep (thus the name “soft sweep”). A similar pattern of “maintained” genetic variation can be achieved with recurrent beneficial mutations appearing in different haplotypes (Fig. 3C).

Much ink has been spilled over which of the two models is the most prevalent one in different species, notably humans (Harris et al., 2018; Jensen, 2014; Messer and Petrov, 2013; Schrider and Kern, 2017) and multiple methods have been developed to try to distinguish them (Garud et al., 2015; Schrider and Kern, 2017). This is beyond the scope of the present manuscript, however the concept of soft sweep, or rather selection on standing variation (since the nature of inferred sweeps is often ambiguous, see Harris et al. 2018), is of particular importance, because it is one possible signature adaptive admixture can produce if neutral genetic variation is acquired through gene flow and then becomes beneficial.

Detecting positive selection with allele frequency information

There are a multitude of methods to determine if a given genomic region is under positive selection. Here, I discuss the overall concept behind these methods and the type of information they use, as well as additional methods that will be later evaluated in the Results chapter.

The first class of methods use allele frequency information, either by directly analysing deviations in the frequency spectrum of a locus (Fay and Wu, 2000; Tajima, 1989a) or by comparing allele frequencies between populations. For the latter, a good example is F_{ST} . Introduced by Sewall Wright in 1950 (Wright, 1950) as part of the F -statistics, which measure the expected average heterozygosity for different degrees of population structure, F_{ST} measures the average heterozygosity at the subpopulation level relative to the total population. It can also be seen as the amount of genetic variance explained by population differences, relative to the total variance (eq. 5).

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1-\bar{p})} \quad (\text{eq. 5})$$

Because these quantities cannot easily be measured, multiple estimators of F_{ST} have been developed (Hudson et al., 1992; Nei, 1973; Weir and Cockerham, 1984). Although in the Results chapter, the evaluated F_{ST} is the one estimated by performing analysis of molecular variance (AMOVA) (Excoffier et al., 1992), here it is illustrated with the Hudson estimator

defined by eq. 6, where H_w is the number of differences within population and H_b the number of differences between populations.

$$F_{ST} = 1 - \frac{H_w}{H_b} \quad (\text{eq. 6})$$

Strong genetic differentiation between populations due to selection will result in high F_{ST} values at the beneficial allele, relative to neutral alleles. To identify regions under positive selection, the genome wide distribution of F_{ST} values is used. Because selection acts locally, affecting only specific loci (while drift affects all loci in a similar way), extreme values of the distribution could then be interpreted as the result of positive selection [Fig3] (Lewontin & Krakauer 1973, Akey 2002).

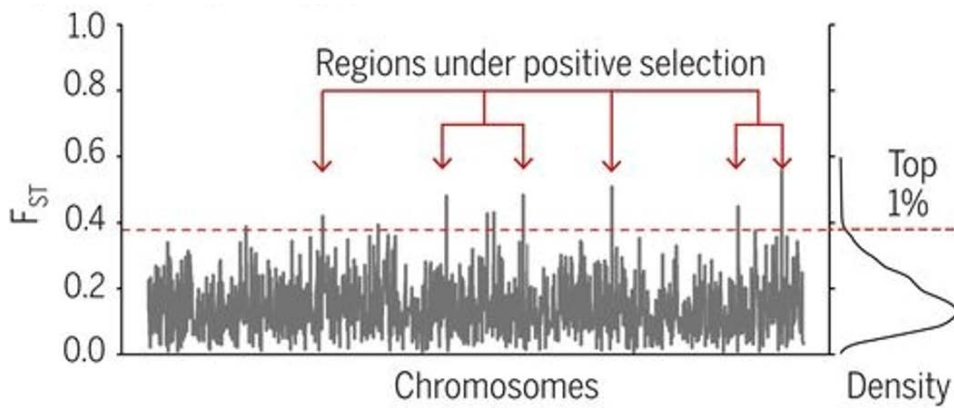


Fig. 4: Changes in F_{ST} along the genome. High F_{ST} values indicate loci that are potential targets of positive selection, based on a cut-off (red dashed line) arbitrarily set to be the top 1%. From Fan et al. 2016.

In a similar way, the difference in derived allele frequency (ΔDAF) statistic can be used to detect positive selection. The main difference being that alleles are “orientated” by assigning them either an “ancestral” or “derived”. This is accomplished by using a phylogenetic approach and multi-sequence alignment between sister species, to infer the common ancestral sequence. In humans, this was accomplished using the EPO (Enredo-Pecan-Ortheus)(Paten et al., 2008) multi-species whole genome sequence alignment from 12 primates (Herrero et al., 2016). As for F_{ST} , strong differences in derived allele frequencies between populations at given loci may indicate positive selection.

Detecting positive selection with haplotype information

A second class of methods use information of multiple linked sites, or haplotype information. Indeed, because of genetic hitchhiking, an extended region of homozygosity is created surrounding the beneficial mutation. This type of signature can be detected in the genome, as shown by Sabeti and colleagues (Sabeti et al., 2002), who defined a measure called extended haplotype homozygosity (EHH). Briefly, EHH is the probability that, at a given distance, x , from a core position, two chromosomes are homozygous at all SNPs situated in the interval defined by x and the core position. Under neutrality and with time, recombination would break down the haplotype in the interval, thus resulting in low EHH values for old, relatively high frequency haplotypes. Under positive selection however, recombination does not have enough time to break down the haplotype centred around the beneficial mutation, thus resulting in high EHH values for high frequency haplotypes. This statistic was later improved by Voight and colleagues (Voight et al., 2006), by evaluating the decay of EHH to the left and right of the core position. They introduced the integrated haplotype homozygosity (iHH) as the area under the curve defined by plotting EHH against distance from the core position. They then calculated the iHH value for the derived and the ancestral state of the core allele (see previous paragraph, here they used a chimpanzee sequence alignment to determine the state), and defined a new statistic, the integrated haplotype score (iHS), as seen in eq. 7

$$iHS = \ln \left(\frac{iHH_A}{iHH_D} \right) \quad (\text{eq. 7})$$

where iHH_A and iHH_D denote the iHH for the ancestral and derived alleles, respectively. When the rates of EHH decay are similar for both allele states, iHS is approximately equal to 0, while large negative or positive values indicate low rates of EHH decay (and therefore positive selection) for the derived or the ancestral allele, respectively. To adjust for differences in EHH decay due to the allele age (under neutrality, low frequency alleles tend to be younger and associated to long haplotypes), they standardized this score to obtain a final statistic with mean zero and variance equal to 1 (eq. 8).

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]} \quad (\text{eq. 8})$$

Both allele frequency and haplotype-based methods are powerful ways to detect recent and strong positive selection events, with signatures akin to a classic, hard sweep. However, as commented earlier on the chapter, these are not the only types of signatures that positive selection can produce. For instance, most of the methods previously presented have limited power to detect soft sweep-like signatures, produced by selection on standing variation or by polygenic selection (selection acting on multiple, often independent sites). Thus, the inference of selective events by only using these methods might result in several false negatives. In addition to this, other evolutionary forces might generate molecular signatures like those of positive selection, resulting in false positives (Jeffrey D Jensen et al., 2005; Tajima, 1989b).

Detecting positive selection: caveats and confounders

Changes in population size, specifically reductions in the case of bottlenecks or founder events, can affect levels of genetic diversity in a way that may be interpreted as signals of positive selection. This is the case for early developed tests such as Tajima's D , Fay and Wu's H statistics. Statistics based on genetic differentiation between populations such as F_{ST} may also produce high values in the case of strong, non-shared genetic drift, for instance if one of the analysed populations has had a small effective size. A simple way to avoid this problem is by using an outlier approach, comparing the value of a statistic in each locus to its genome-wide distribution and defining a threshold (usually the top 1% of the distribution) based on which a locus can be considered as a candidate for positive selection. However, this approach has several limitations. First, it assumes that selection acts in a locus specific way while other forces act at a genome-wide scale. Although this has a solid theoretical and empirical background (Kimura, 1968; Kimura and Ohta, 1971; Lewontin and Hubby, 1966; Ohta, 1973; Zuckerkandl and Pauling, 1965), recent works suggest that a much higher proportion of the genome might be functionally important, and thus potentially targeted by selection (Begun et al., 2007; Hahn, 2008; Kern and Hahn, 2018; Schrider and Kern, 2017; Sella et al., 2009). The genome-wide distribution of these statistics would therefore not only reflect demographic

events or strong drift but also selection. Second, an outlier approach cannot statistically differentiate the effects of positive selection from pure genetic drift or demography. All distributions will have a “top 1%”, and whether this reflects positive selection or a bottleneck cannot be known for certain, especially in cases of weak selective events and/or sharp reductions in population effective size. Third, because this approach is based on the position of a value relative to the whole distribution, evolutionary forces that increase the overall variance of the distribution can reduce detection power (by increasing false negatives).

Another evolutionary force that might make difficult the inference of positive selection is, quite ironically given the context of the work presented here, migration or more specifically, admixture. Although an admixture event does not result in an increase of false positives per se, it can obscure the typical signatures of a selective sweep. Indeed, newly arriving genetic variation from the source population can increase the levels of genetic diversity that were decreased by selection in the recipient population, and recombination with newly arriving haplotypes can break down haplotype homozygosity (Gravel, 2012; Lohmueller et al., 2011).

Finally, negative selection can further obscure the detection of positive selection. Deleterious mutations by themselves can interfere with the action of positive selection at nearby sites (Hill and Robertson, 1966) and reduce the fixation probability of a beneficial mutation, especially in the case where recombination levels are low (Birky and Walsh, 1988). More importantly, the effect of negative selection on linked sites (known as background selection) can reproduce, in regions of low recombination, signatures of reduced diversity at neutral genetic variation, like those produced by selective sweeps (Charlesworth et al., 1993).

Chapter 3: Modelling and detecting admixture

The single pulse admixture model

Although it has not been the most studied migration model in population genetics, the single pulse admixture model is often assumed when trying to estimate admixture proportions or when trying to describe the history of admixture through time (Verdu and Rosenberg, 2011). The model assumes that a given admixed population derives their genetic ancestry from at least two different source populations, resulting from a single instance of unidirectional gene flow (Fig. 5). Under this model, the allele frequencies in the admixed population, p_h , can be approximated as a linear combination of the allele frequencies in the source populations p_1 and p_2 , weighted by their contributions to the admixed population, α and $(1 - \alpha)$, referred to as admixture proportions (Bernstein, 1931) (eq. 9).

$$p_h = \alpha p_1 + (1 - \alpha) p_2 \quad (\text{eq. 9})$$

Importantly, even if allele frequencies at individual loci in the admixed population might drift away, on average, this approximation holds.

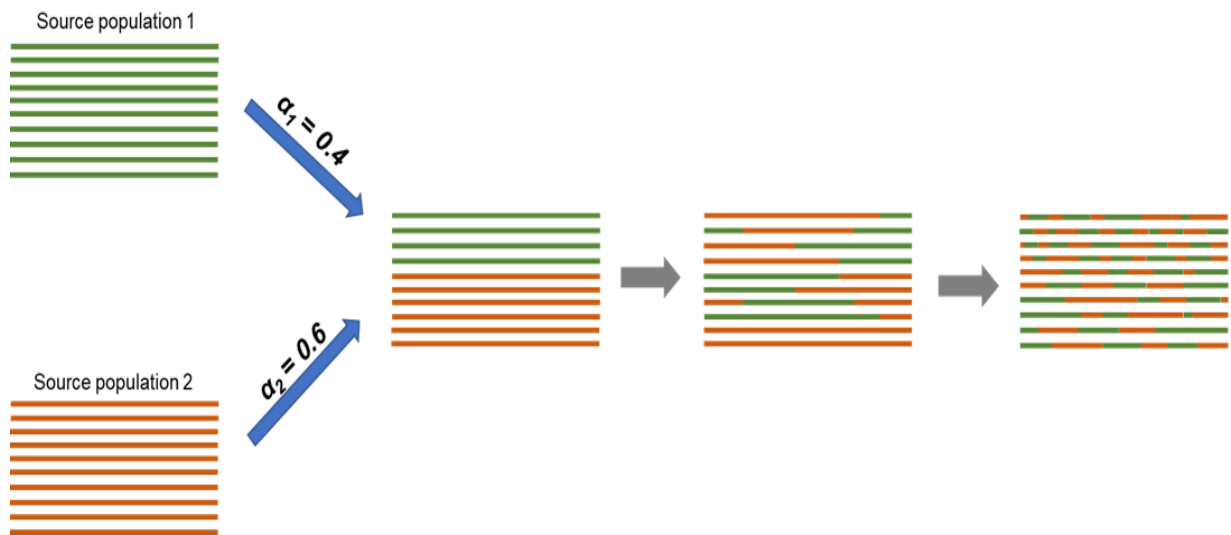


Fig. 5: Outcome of a single pulse admixture event. An admixed population is created from a single admixture event, by inheriting a fraction α_1 of haplotypes from one source population (source population 1 in green) and a fraction α_2 of haplotypes from another source population (source population 2 in orange). As generations pass, recombination events between the different haplotypes lead to variation in genetic ancestry along and between the haplotypes. Nevertheless, the overall proportion of green and orange segments in the admixed population still reflects the initial admixture proportions.

In human populations, this admixture model is quite unrealistic, not only because most contacts between previously isolated populations were in the form of colonization waves or forced displacements over time, but also because admixture in humans has been influenced by sociocultural rules and practices in contexts of discrimination, slavery or caste systems which could affect the patterns of genetic variation in admixed groups. Nevertheless, it has been shown that a single admixture event suffices to recapitulate the genetic diversity of admixed populations and estimate relatively accurately the average timing and admixture proportions in human populations (Hellenthal et al., 2014). For this reason, we will assume this simple model for most of the work discussed here.

Assessing admixture with allele frequency information (I)

Even though the analytical framework to estimate admixture proportions has been well established for over 90 years (Bernstein, 1931), a limiting factor for the first estimations (in human populations at least) was the availability of multi-locus genetic data. Not only that, but the framework assumed the exact contributing sources were known, which is rarely the case. It

would not be until the availability of high density, genomic data that an alternative framework of analyses, that did not make such assumptions, could be used. These are known as unsupervised analyses, because they do not require any prior information about population affiliation for the studied samples. There are two main types of unsupervised analyses. The first one is principal component analysis (PCA), which reduces the complexity of a large multidimensional dataset (a matrix of hundreds of thousands of genetic markers for hundreds of individuals for instance) by extracting principal components that explain the most the observed variability in the dataset, thus reducing dimensionality but retaining a maximum amount of information. Through PCA it has been shown that geography has had a very large influence in the genetic structure of closely related populations (Novembre et al., 2008). Moreover, it has also been proven that distances in a PCA plot can be correlated to genetic distances, and the proportion of variance explained by PCs equate F_{ST} (McVean, 2009). In that sense, individuals that are positioned in between two clusters (corresponding to two populations) along a principal component carry alleles at frequencies that are intermediate between those of the two clusters, which may be interpreted as resulting from an admixture event between these two populations.

The second type of unsupervised analyses are based on a clustering method developed by Pritchard and colleagues and implemented through a software, STRUCTURE (Pritchard et al., 2000). This analysis assumes that sampled individuals derive their genetic ancestry from a given number of unknown source populations, and then simultaneously estimates the allele frequencies of said source populations as well as the ancestry proportions for each sampled individual. How K , the number of source populations, also referred to as ancestry components, is determined depends on the method but usually multiple runs of the algorithm are made until finding the K values that best fits the data.

Both types of methods are great for visualizing and describing the genetic variability of a group of samples. However, in terms of result interpretation, there are several pitfalls (Lawson et al., 2018). Neither PCA nor STRUCTURE-like clustering give any information on the causes for the observed patterns. Even if these can be interpreted as genetic distances, in the case of PCA, whether these distances are due to pure genetic drift, a population bottleneck or an admixture event cannot be differentiated. Sampling strategy is of utmost importance, especially for STRUCTURE-like analyses since they can produce easily misinterpretable patterns when

an unsampled population has a strong contribution or a high level of shared ancestry with the rest of the samples.

Assessing admixture with allele frequency information (II)

The previous methods can provide information to emit hypotheses about admixture occurring in a given group. However, to formally test and validate these hypotheses, another class of methods were specifically developed, relying also on allele frequency information, called f -statistics (not to be confounded with Wright fixation indexes, denoted F -statistics). Developed by Nick Patterson and introduced by David Reich and colleagues (Patterson et al., 2012; Reich et al., 2009), these statistics are based on the concept of shared genetic drift between populations, which implies a shared evolutionary history. In particular, the f_3 statistic can be used as a formal test for admixture. It is defined by eq. 10

$$f_3(P_X; P_1, P_2) = E[(p_X - p_1)(p_X - p_2)] \quad (\text{eq. 10})$$

that is the product of the allele frequency differences between P_X and P_1 , and between P_X and P_2 , average across all sampled alleles. In terms of drift, this corresponds to the shared amount of drift between the (P_X, P_1) and (P_X, P_2) pairs. In the case of no admixture, the expected value of f_3 would reflect the amount of genetic drift specific to the lineage leading to population X since its divergence from the lineage(s) leading to populations P_1 and P_2 (Fig 4.A and B). In the case of admixture, for instance between P_1 and P_2 resulting in P_X , the shared drift is impacted by the inheritance by P_X of ancestry from P_1 and P_2 , such that the amount of shared drift between P_1 and P_2 (but not with P_X) negatively affects the f_3 statistic resulting in negative values (Fig 4 C). More intuitively, the f_3 statistic takes negative values when p_X is intermediate between p_1 and p_2 , which is expected under admixture (Bernstein 1931). However, p_X may become lower or larger than both p_1 and p_2 because of drift since admixture. Consequently, this statistic is particularly powerful when the divergence time between both source populations is particularly old (higher amounts of shared drift between P_1 and P_2 only) but decreases in performance if P_X has been subject to strong amounts of genetic drift (which could lead to positive f_3 values).

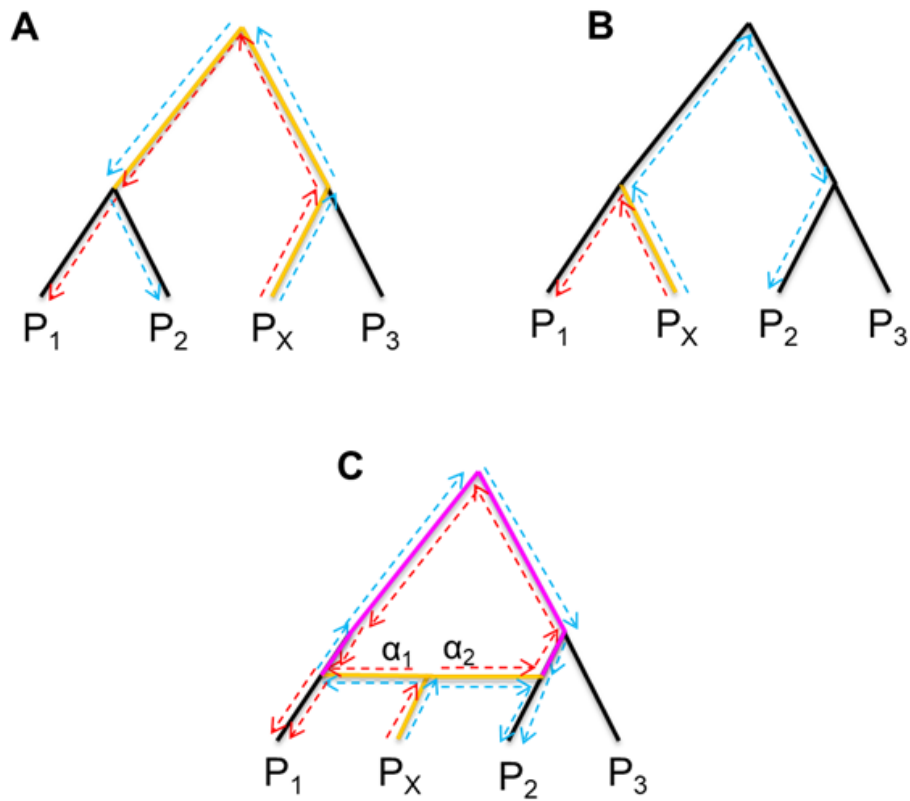


Fig. 6: The f_3 statistic used to infer admixture. Phylogenetic trees with topologies that may produce positive (A and B) or negative (C) values of the f_3 statistic. Red and blue arrows denote drift paths from the tested population (P_X) to population P_1 and P_2 respectively. The intersect of these drift paths, if they go on the same direction, (shared drift) contributes positively to the value of the f_3 statistic (orange-coloured branches). In the case of admixture, additional drift paths that can go in opposite directions are introduced, which contribute negatively to the f_3 statistic (magenta-coloured branches).

Assessing admixture with haplotype information

As with natural selection, analysing genetic data in the form of haplotypes provides additional layers of information about an admixture event. For instance, admixture can generate high amounts of LD between loci with different allele frequencies in the source populations (Nei and Li, 1973). Because of recombination, this LD (known as admixture LD or ALD) decays over time since the admixture event, which can be modelled to infer details about the timing of the event (Loh et al., 2013; Moorjani et al., 2011). More precisely, LD between markers, weighted by their allele frequency differences, can be modelled as a (negative exponential) function of genetic distance between said markers. When LD between markers at short genetic distances (which reflects more background LD than ALD) is excluded, the function reflects ALD decay, with the rate of decay proportional to the time since the admixture event occurred (Fig 5 A).

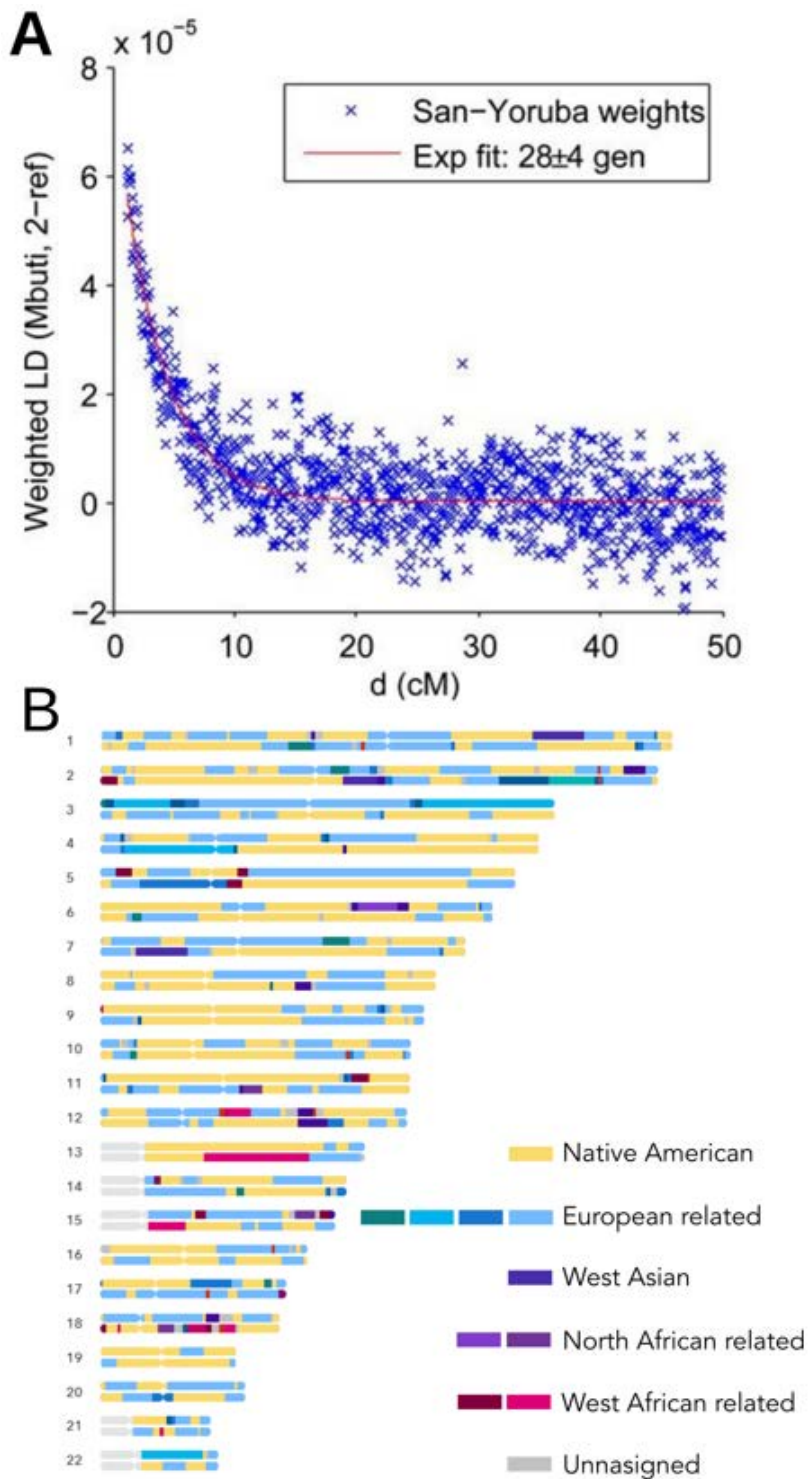


Fig. 7: Haplotype-based methods to assess admixture. (A) Weighted LD decay curve for Mbuti pygmies using San and Yoruba as reference populations. Red line corresponds to a fitted negative exponential, assuming an admixture time parameter value of 28 generations. From Loh et al. 2013. (B) Local ancestry inference of an admixed Peruvian individual (the author of this manuscript) using multiple reference populations. Obtained using a support vector machine-based method (23andme).

Another use of haplotypes is local ancestry inference (LAI), along individual genomes in an admixed population, also known as ancestry deconvolution. That is, given an individual admixed genome, determining which genomic regions were inherited from a given source that contributed to the admixture event (Fig 5B). There are different types of methods that can perform LAI: methods that explicitly incorporate LD and fit a Hidden Markov model (a type of statistical model where an observable process outcome, in this case alleles, depends on an unobserved process, ancestry) to the data (Baran et al., 2012; Guan, 2014; Price et al., 2009; Tang et al., 2006), discriminative methods that explicitly model the ancestry across a chromosome given haplotypes of known ancestry, through a window-based approach (Brisbin et al., 2012; Haasl et al., 2013; Hilmarsson et al., 2021; Maples et al., 2013), and methods based on chromosome painting, where the chromosomal segments of target individuals are described (or “painted”) based on a large panel of populations that may or may not be related to the admixture event (Lawson et al., 2012; Molinaro et al., 2021). The inferred ancestry blocks can then be used to learn about parameters of the admixture event such as the date (akin to ALD decay), the admixture proportions (Pool and Nielsen, 2009) or even the complexity of the event itself (Choin et al., 2021). Interestingly, LAI can also be used as an empirical way to detect positive selection occurring in an admixed population, as strong deviations in a particular ancestry at given regions of the genome could be interpreted as signals of positive selection for that ancestry (Tang et al., 2007). Nevertheless, this empirical analysis is an outlier approach and thus inherits all the drawbacks and limitations from it.

Assessing admixture: towards more complex models

As stated in the beginning of this chapter, the single pulse admixture model is both good (because it allows to infer parameters such as average proportions and average times of admixture in a rather simple framework) and bad (because it rarely reflects the actual admixture history of a population). More complex models exist, such as multiple discrete instances of admixture, or continuous non-symmetrical gene flow after an initial admixture event. There are even models that propose no individual admixed population but rather two populations that experience gene flow in an asymmetric, non-constant manner. In humans, even populations that were thought to have an admixture history akin to the single pulse model, such as admixed populations in South, Central and North America have shown a higher degree of admixture complexity (Chacón-Duque et al., 2018; Fortes-Lima et al., 2020). Inferring the exact, detailed admixture history is beyond the scope of the project presented here. However, accounting for

such complexity can be important when detecting positive selection in admixed populations, as we will see in the Results chapter.

Chapter 4: Gene flow according to evolutionary biology

A barrier to speciation

Gene flow has been viewed as a constraining force in evolution (Slatkin, 1987), that may prevent populations from evolving into two different species. Although the idea of isolation as a factor promoting evolution was described by Darwin (without explicitly affirming that isolation needs to happen for species to appear), it was formally advanced by Mayr's studies of "allopatric" mode of speciation: gene flow between populations of the same species can prevent local differentiation; if it is interrupted, populations can then evolve independently into separate species (Mayr, 1970, 1963, 1942). This view was further developed quantitatively by Ehrlich and Raven (Ehrlich and Raven, 1969) who asked just how much gene flow is needed to prevent speciation. As seen in Chapter 1, the exchange of only a few migrants between populations, independently of their size, suffices to reduce differentiation due to genetic drift. Nevertheless, speciation can occur in the context of gene flow (known as sympatric speciation), if different local adaptations arise and are closely linked to reproductive isolation or are correlated through a single gene, affecting both. More generally speaking, in the case where local adaptation is due to natural selection, gene flow can limit differentiation if the fraction of immigrants is higher than the fitness differences (Slatkin, 1987).

A potential source of deleterious variation

Genes under local adaptation can also act as a barrier to gene flow because, when introduced into the genome of other populations, they may become detrimental, and locally reduce the observed migration rates (Aeschbacher et al., 2017; Bengtsson, 1985). Why these genes become detrimental can be due to different reasons: hybrid sterility (e.g genetic adaptations through large genomic inversions), differences in environmental pressures between the populations experiencing gene flow, or negative interactions between previously isolated alleles that now share the same genetic background (also known as Dobzhansky-Müller incompatibilities). Nevertheless, to considerably reduce gene flow, the fitness of first-generation hybrids should be substantially lower and/or the loci affected by selection be numerous enough so that any other locus could be potentially linked to one of them (Barton and Bengtsson, 1986).

Deleterious variation can also arise, or rather accumulate, in populations with low effective population sizes. Because natural selection operates much less efficiently (see Chapter 1), newly appearing deleterious mutations might not be eliminated and rather increase in frequency purely by chance. A recent example of this has been reported in non-African human populations, which received gene flow from Neanderthal, Denisova or related archaic hominin groups. Indeed, certain regions of the genomes of these populations are depleted in archaic ancestry (Harris and Nielsen, 2016; Juric et al., 2016; Sankararaman et al., 2016, 2014). Given the functional importance of most of these regions and the low effective population sizes of archaic groups, it has been hypothesized that archaic hominin populations had accumulated a substantial load of deleterious variants, and that these variants were effectively purged once present in modern human populations due to their larger N_e (Harris and Nielsen, 2016; Juric et al., 2016). However, this hypothesis might not explain the totality of archaic ancestry depletion, such as in genes expressed in male tissues only, suggesting that hypotheses like male hybrid sterility may have come into play as well (Sankararaman et al., 2016).

A potential source of beneficial variation

While gene flow can serve to circulate deleterious variants, both through differences in local adaptation or effective population size, the circulation of beneficial variants can be seen as the other side of this same coin. Indeed, at the most basic level, gene flow increases the overall effective population size. This in turn not only leads to a higher efficiency of natural selection in purging deleterious variation and higher levels of heterozygosity (which may help, in a conservation context, rescue populations experiencing high levels of inbreeding), but may also generate novel genetic variation on which selection can act (Edelman and Mallet, 2021). Furthermore, under common environmental constraints, gene flow may directly enable the circulation of beneficial variation, something that was initially suggested by Anderson in 1949 (Anderson, 1949), who reasoned that hybrids may serve as a gateway for exchange of genetic adaptations between otherwise isolated species. However, theoretical difficulties for this type of adaptation, such as reduction in gene flow due to deleterious variation (see previous section) limited the scope of Anderson's theory. Nevertheless, work by Barton in 1979 (Barton, 1979) and Barton & Bengtsson in 1986 (Barton and Bengtsson, 1986) showed that alleles under strong positive selection (and providing a selective advantage in the heterozygotes) could spread

across hybrid zones. These zones would then serve as conduits for beneficial variation to circulate between locally adapted populations. The term *adaptive introgression* is employed when referring to this phenomenon (i.e the acquisition of beneficial variants through introgression). This does not imply however that the initial hybridization process is adaptive or that hybrids have higher fitness (a phenomenon known as heterosis or hybrid vigour). While heterosis can be observed in the first generation of hybrids, as heterozygous genotypes can have a higher fitness, the effects of adaptive introgression can only be observed in later generations, when the beneficial introgressed allele has not been lost and has rapidly increased in frequency, even when most of the initially introgressed variation have been lost due to linked negatively selected mutations (see previous paragraph).

Early examples of adaptive introgression came mostly from plant species such as sunflowers (Heiser, 1979; Heiser Jr., 1951), bitterbrush (Stutz and Thomas, 1964) and common groundsel (MONAGHAN and HULL, 1976). However, in most of these cases it was not possible to distinguish between adaptive introgression and alternative explanations such as convergent evolution or trans-species polymorphism (Hedrick, 2013; Kim and Rieseberg, 1999) (see Fig. 6). Other studies (Arnold et al., 1991) reasoned that, if they observe neutral molecular markers crossing hybrid zones, adaptive introgression is also plausible (since beneficial alleles would cross hybrid zones more easily than neutral ones). However, evidence was considered elusive due to the preponderance of neutral alleles relative to advantageous ones in hybrid zones (Rieseberg and Wendel, 1993).

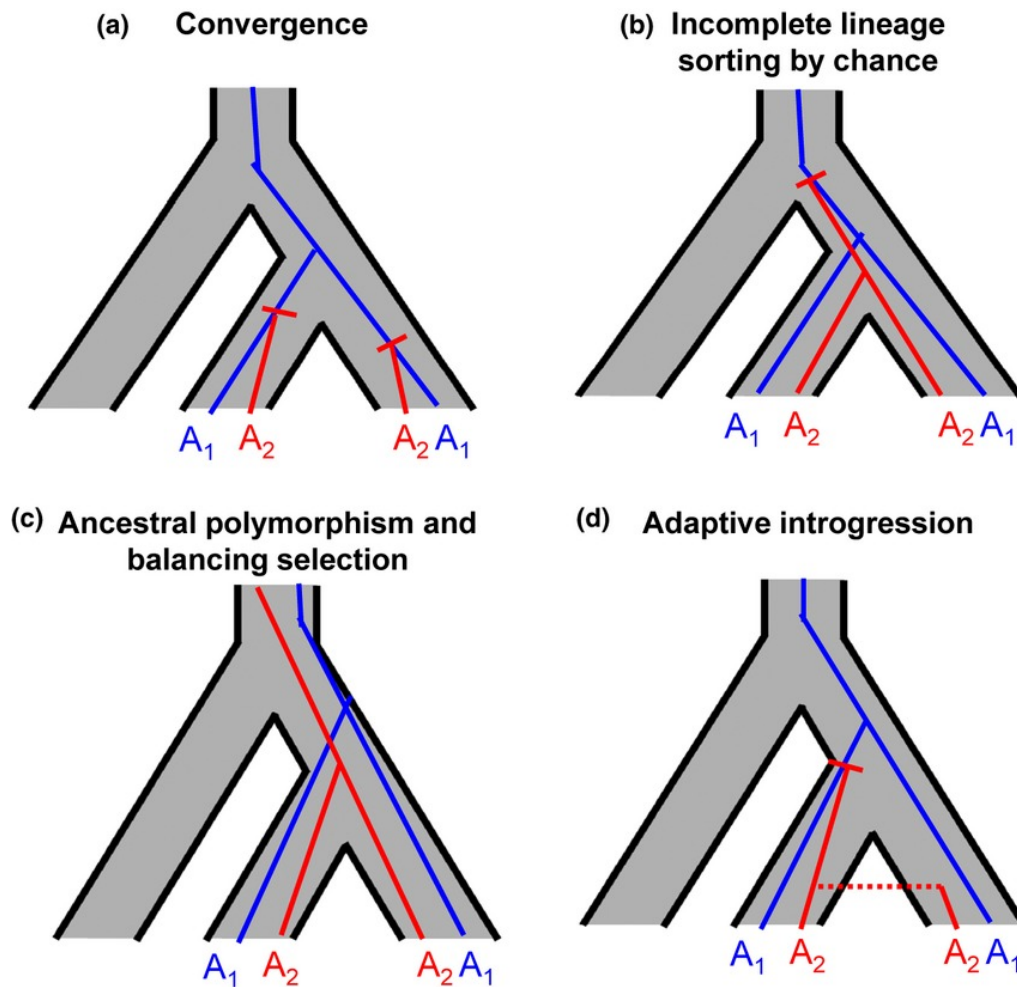












Fig. 8: Different evolutionary scenarios that can explain shared polymorphism between two groups. (a) Convergence: two identical, beneficial, derived alleles appear independently in both groups posterior to their divergence. Shared ancestral polymorphism is maintained by chance (b) or balancing selection (c). (d) Adaptive introgression: a beneficial allele appearing in one group is transmitted to another through an introgression event. From Hedrick’s review on adaptive introgression (Hedrick, 2013)














In animals, an early example was suggested by Lewontin & Birch in 1966 (Lewontin and Birch, 1966), in which expansion of the geographical range of the Australian fruit fly would have benefited of introgression from a closely related species. Nevertheless, the importance of adaptive introgression was often minimized, due to the often-low levels of introgression in animal species and due to the low fitness often observed in hybrid individuals (Hedrick, 2013).






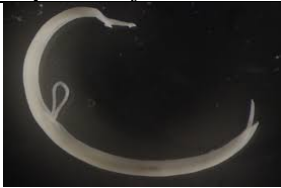






However, with the availability of high-density genetic data as well as new approaches that accurately assess adaptive introgressive variants, examples of this evolutionary phenomenon have multiplied in the recent years (Table 1), consolidating Anderson’s initial hypothesis of gene flow as an important source of variation available for evolution (Suarez-




Gonzalez et al., 2018b). This is of particular importance for species with unstable population structures (because of large-scale changes in geographic range or recurrent extinction/recolonisation events) that leads them to experience new or changing environments. Modern humans are a good example of this, as shown in the next chapter.

Table 1: Empirical examples of adaptive introgression

Organism		Trait(s)	Reference
 <i>Manacus candie</i>	 <i>Manacus vitellinus</i>	Plumage ornaments	(Parsons et al., 1993)
 <i>Mus musculus</i>	 <i>Mus spretus</i>	Xenobiotic resistance (rodenticides); olfaction	(Liu et al., 2015; Song et al., 2011)
 <i>Heliconius timareta</i>	 <i>Heliconius melpomene</i>	Wing coloration	(Pardo-Diaz et al., 2012)
 <i>Zea mays</i>	 <i>Zea mays ssp. mexicana</i>	Leaf sheath macrohairs pigmentation	(Hufford et al., 2013)
 <i>Anopheles coluzzi</i>	 <i>Anopheles gambiae</i>	Insecticide resistance	(Norris et al., 2015)

		Stress signalling and tolerance	(Arnold et al., 2016)
<i>Arabidopsis arenosa</i>	<i>Arabidopsis lyrata</i>		
		Phenology/Growth	(Lind-Riehl and Gailing, 2016)
<i>Qercus rubra</i>	<i>Qercus ellipsoidalis</i>		
		Phenology, biomass and ecophysiology	(Suarez-Gonzalez et al., 2018a)
<i>Populus balsamifera</i>	<i>Populus trichocarpa</i>		
		Long day tuberization	(Hardigan et al., 2017)
<i>Solanum tuberosum</i>	<i>Solanum microdontum</i>		
		Black coat colour (apparent)	(Anderson et al., 2009)
<i>Canis lupus</i>	<i>Canis familiaris</i>	Immune response (real)	(Schweizer et al., 2018)
		Hypoxia tolerance	(Miao et al., 2017)
		Domestication-related traits.	(Burgarella et al., 2018)
<i>Cenchrus americanus</i>	<i>Pennisetum glaucum monodii</i>		
		Milk production. Growth and height.	(Rochus et al., 2018)
<i>Ovis aries</i> breeds			

		<p>Hypoxia tolerance</p>	<p>(Hu et al., 2019)</p>
<p><i>Ovis aries</i></p>	<p><i>Ovis ammon</i></p>	<p>Winter coat colour change</p>	<p>(Jones et al., 2018)</p>
		<p>Tissue penetration; immune evasion</p>	<p>(Platt et al., 2019)</p>
		<p>Abiotic stresses (high-elevation environment)</p>	<p>(Ma et al., 2019)</p>
		<p>Body size and feed efficiency</p>	<p>(Barbato et al., 2020)</p>
		<p>Insecticide resistance</p>	<p>(Valencia-Montoya et al., 2020)</p>
			
<p><i>Helicoverpa zea</i></p>	<p><i>Helicoverpa armigera</i></p>		

		<p>Resistance to DDT exposure</p>	<p>(Svedberg et al., 2021)</p>
<p><i>Drosophila melanogaster</i> populations</p>		<p>Hypoxia tolerance</p>	<p>(Graham et al., 2021)</p>
			
<p><i>Anas flavirostris</i></p>	<p><i>Anas georgica</i></p>		

Chapter 5: Admixture according to molecular anthropology

Initial studies in admixed populations

Initial population genetics studies conceived *Homo sapiens* as a structured species composed of populations isolated by large geographical distances. Admixture was thought to be rare and confined to populations that descend from admixture events postdating colonisation processes. Nevertheless, admixed populations offered a good opportunity for population geneticists to estimate admixture proportions and admixture dynamics, using blood groups or immunoglobulins (Cavalli-Sforza and Bodmer, 1999; Chakraborty, 1986; Glass and Li, 1953; Long, 1991; Reed, 1969; Workman et al., 1963). Furthermore, admixed populations also offered the possibility to assess linkage between traits and genetic markers. In 1954, Rife showed that correlations between two genetic traits, in an admixed population, could be indicative of linkage (Rife, 1954). This would constitute the base for admixture mapping: a disease or a trait showing differences by ancestry can be correlated with local ancestry in admixed populations, which could be leveraged to map disease risk loci. However, advances in the field will be limited by the lack of availability of high-density genetic data (Shriner, 2013).

However, the way population geneticists evaluate and interpret admixture in human populations would change drastically in early 2010s, with a discovery that would take the field to the next level.

The ancient DNA revolution: admixture with archaic hominins

In 2010, Green and colleagues sequenced for the first time the Neandertal genome using DNA extracted from paleontological samples taken from Vindija Cave in Croatia (Green et al., 2010). This was revolutionary in many ways. Sequencing DNA from ancient samples (in this case over 38,000 years old) had long been a technological challenge due to the physical degradation, chemical modifications, and exogenous contaminants characteristic of ancient samples, yet the authors managed to use a targeted sequencing approach that yielded analysis-quality ancient DNA. Most importantly perhaps, the authors reported that the genome of

modern human populations contained traces of Neandertal DNA, suggesting an admixture event had occurred approximately 120,000 years ago. While it had long been known that the two species had co-existed in Europe and Western Asia during the late Pleistocene, the question of interbreeding remained controversial since it could previously only be addressed with morphological data and non-autosomal DNA. The analysis of autosomal ancient DNA brought the possibility to answer this question with almost certainty. Almost, because it was later shown that signals of admixture (referred to as introgression in this case because there are two different species) could have alternative explanations such as ancient population substructure or incomplete lineage sorting (see Fig. 6 in Chapter 4) (Eriksson and Manica, 2012). Nevertheless, with the development of new methods that distinguished between these alternative hypotheses, and with further studies providing additional evidence of introgression from other archaic hominins like Denisova (Meyer et al., 2012; Reich et al., 2010), it was clear that gene flow has been an important part in the history of non-African human populations. The analysis of DNA from ancient human samples would eventually show that this was a severe understatement: in reality, all present-day populations result from admixture events between at least two sources.

The ancient DNA revolution: admixture is everywhere

The advent of ancient DNA sequencing techniques saw an influx of population genetic methods in interdisciplinary archaeological and anthropological studies. Indeed, although some insights into human evolutionary history were gained by analysing DNA from present-day human populations, such as the strong support for the out-of-Africa hypothesis (Cann et al., 1987; Prugnolle et al., 2005a; Quintana-Murci et al., 1999; Underhill et al., 2000), or that the ancestral population of present day European and Asians experienced an initial bottleneck when arriving to Eurasia (Henn et al., 2012), many more questions at a finer resolution in space and time could not be properly answered. With ancient DNA, scientific interest for those questions was rekindled: it suffices to do a quick search of how many articles have been published with the words “the genetic history of” (a seemingly “click-bait,” yet surprisingly accurate, term associated with ancient DNA studies on human populations) included in their title since 2010 (Fig. 9 A) to see that, on average per year, that number was multiplied by almost 5 (almost 6 when using related search terms, Fig. 9 B) when compared to before 2010 (the year when the Neandertal draft sequence was published).

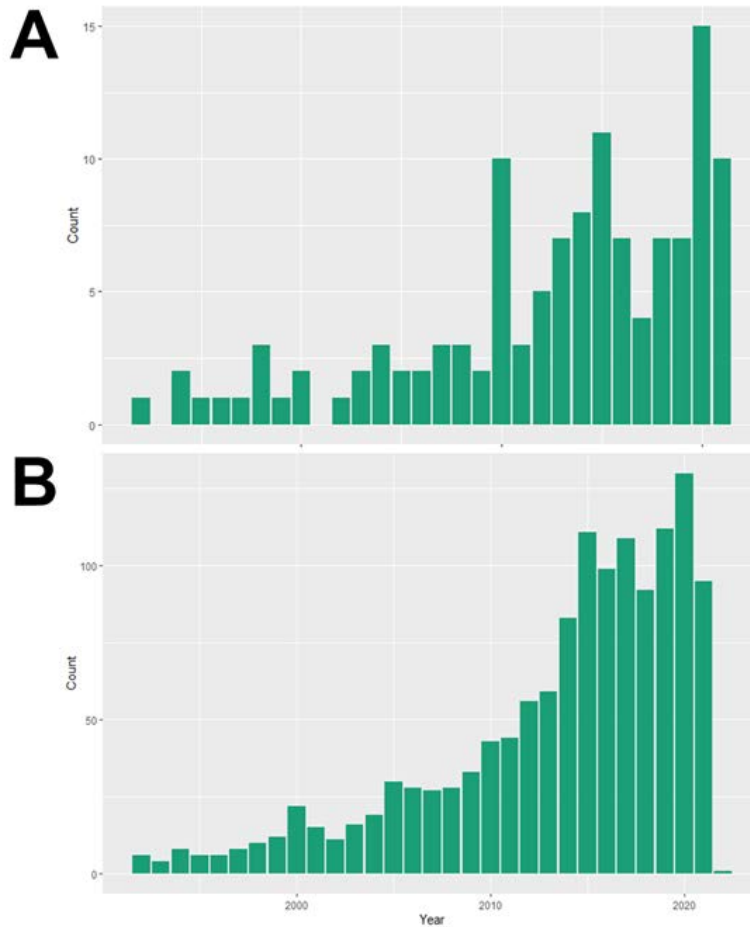


Fig. 9: Publication record of articles containing the terms (A) “the genetic history of” or (B) “ancient DNA” and “humans” in their title. Data from PubMed (publications from 1992 to 2021).

Overall, three major insights were gained by analysing this type of data. First, migrations were much more numerous and complex than previously thought (Liu et al., 2021) (See Fig 8). Second, there has been a great deal of population structure, even in early human evolutionary history (Skoglund and Mathieson, 2018). And third, most of present-day human populations derive from multiple and often complex admixture events (Korunes and Goldberg, 2021).

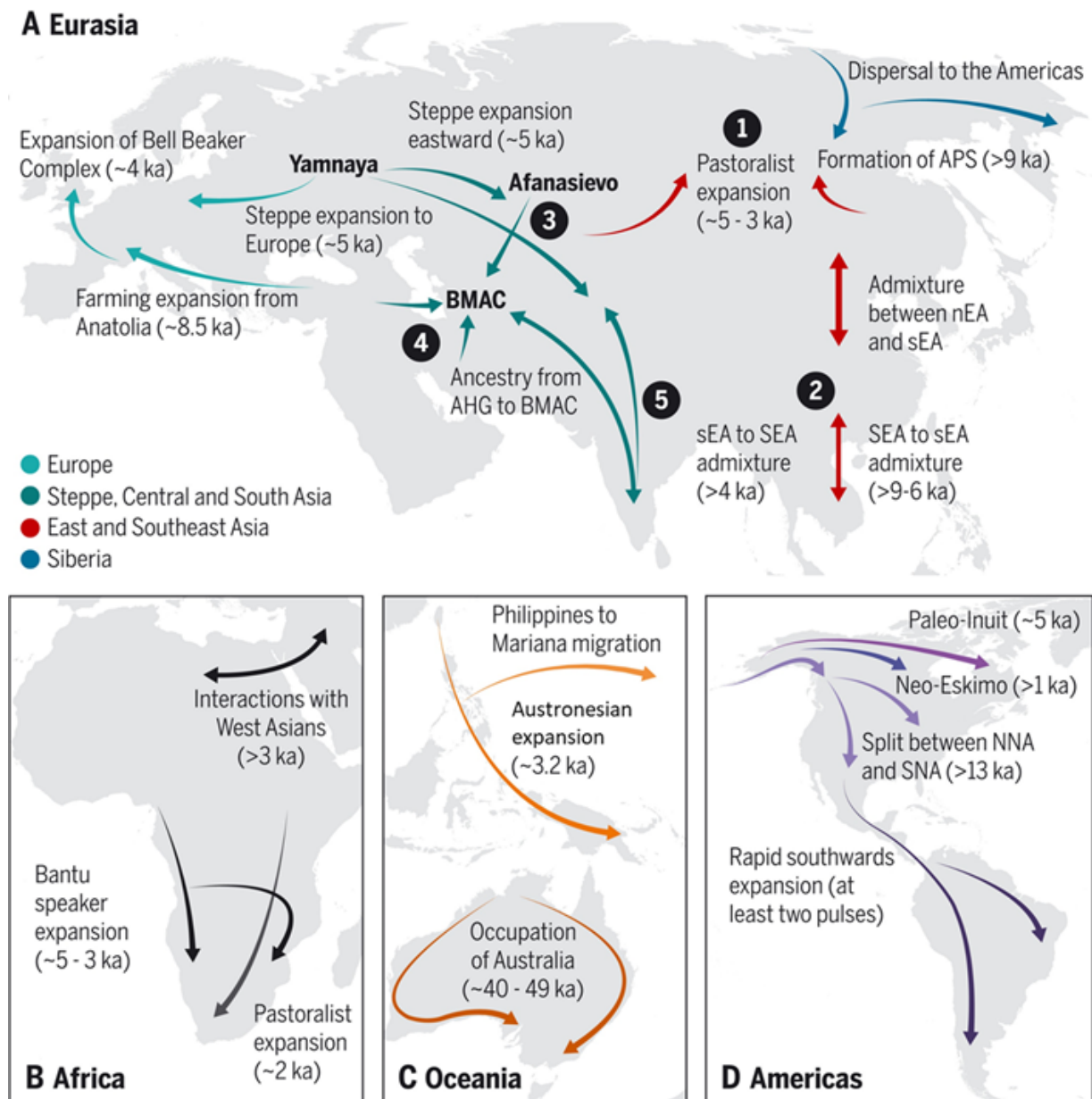


Fig. 10: Major population movements since the last glacial maximum inferred with ancient DNA. From Liu, Mao, Krause & Fu 2021 review on ancient human genomics (Liu et al., 2021).

Adaptive admixture in modern humans

Having established admixture as a rule rather than an exception in human evolutionary history, ancient DNA also opened the door to studying admixture as a driver for adaptive evolution in humans, as seen in the previous chapter. The year following the publication of the Neandertal genomic sequence, Abi-Rached and colleagues reported signatures of positive selection in class-I *HLA* haplotypes of Neandertal and Denisovan origin in Europeans, East Asians and Melanesians (Abi-Rached et al., 2011). Many more studies would follow,

suggesting adaptive introgression from archaic hominins had had an impact on diverse phenotypic traits in present-day modern human populations, such as pigmentation (Racimo et al., 2017), adaptation to life in high altitudes (Huerta-Sánchez et al., 2014) and immune response to pathogens (Choin et al., 2021; Deschamps et al., 2016; Enard and Petrov, 2018; Quach et al., 2016; Zeberg and Pääbo, 2020). Nevertheless, most of the admixture events characterized in human history occurred between human populations, especially during the last 10,000 years, due to numerous large-scale population dispersals enabled by technological and/or demographic changes (see Fig 8). In that sense, studies exploring adaptive admixture between human groups began to expand their scope beyond the classically studied admixed populations (admixed South and Central Americans and African Americans). An extensive recompilation of all adaptive admixture studies in modern humans is presented in Table 2. Not surprisingly, several of the reported genomic regions were previously described as targets for positive selection (Fan et al., 2016), however this consolidates the adaptive importance of said genes for the different environments that human populations have been exposed to since the last glacial maximum.

Table 2: Empirical examples of adaptive admixture in modern humans. The source putatively bringing the adaptive mutation is marked in bold.

Admixed population(s)	Source populations	Trait(s)	Reference
Puerto Ricans	Native Americans West Europeans West Africans	Immune response	(Ongaro et al., 2021; Tang et al., 2007)
Nama	Afro-Asiatic related group, San hunter-gatherers	Lactase persistence	(Breton et al., 2014)
Tibetans	Sherpa (inherited from Denisova-related group) Han Chinese-related group	High altitude adaptations	(Jeong et al., 2014)
East Sahelian populations (Sahelian Arabs and Nubians)	Near East populations North Africans East Africans	Malaria resistance	(Triska et al., 2015)

Colombians	Native Americans West Europeans West Africans	Immune response	(Deng et al., 2016; Norris et al., 2020; Ongaro et al., 2021; Rishishwar et al., 2015)
Cosmopolitan Mexicans	Mexican Native Americans West Europeans West-Africans	Immune response	(Deng et al., 2016; Norris et al., 2020; Ongaro et al., 2021; Zhou et al., 2016)
Makranis of Pakistan	Baluch East/Southeast Bantu	Malaria resistance	(Laso-Jadart et al., 2017)
Bakiga east Bantu	West Bantu East African pastoralists	Lactase persistence	(Patin et al., 2017)
West Bantu	West Africans West rainforest hunter-gatherers	Immune response	(Patin et al., 2017)
Malagasy	Southeast Asian Austronesian speakers, Southeast Bantu	Malaria resistance	(Pierron et al., 2018)
West Arabian Peninsula populations	East Africans Near East/Europeans related groups	Malaria resistance	(Fernandes et al., 2019)
Fulani	West Africans North Africans with European admixture	Lactase persistence	(Vicente et al., 2019)
Cabo Verde	West Africans West Europeans	Malaria resistance	(Hamid et al., 2021)
Polynesians	Papuans Austronesian speakers	Immune response	(Isshiki et al., 2020)

Chapter 6: Objectives of the thesis

The examples of adaptive admixture presented at the end of the previous chapter highlight not only the importance of studying admixture as a driver of natural selection, but also several weak points and limitations in trying to do so. A number of these studies derived their results from classic selection scans performed on admixed populations. However, as seen in Chapter 2, this can be problematic as these analyses may have reduced power in an admixed background. Furthermore, although local ancestry deviations are often interpreted as evidence for adaptive admixture, the thresholds used to characterize a deviation as “strong” are arbitrarily set most of the time.

Performing analyses that do not rely on statistical outliers, but rather explicitly model the studied admixture event could be one way of circumventing these issues, as demonstrated by examples stated in Table 2. However, most demographic histories are complex and accurately inferring model parameters or distinguishing between alternative population models can be difficult (Choin et al., 2021). Worse, a slight misspecification in the model used may result in a non-negligible number of false positives (Jeffrey D. Jensen et al., 2005; Schrider and Kern, 2016).

Using statistical outliers as a means to detect candidates for adaptive admixture is thus an important and necessary method for population geneticists. Nevertheless, if they are to be used, a correct assessment of statistical power as well as the effect potential confounders is necessary. The present work aims to evaluate this in two ways. First, by computing realistic simulations of admixed genomes under adaptive admixture or in the absence of positive selection and evaluating the power of different statistics to detect adaptive admixture. Next, by evaluating the effect of potential power-reducing factors such as limited sample sizes, or population demography; and finally, by analysing the genomes of 15 present-day, worldwide, admixed populations. The present work will then provide a more solid picture of how much admixture has contributed, as a source of beneficial variation, to the evolutionary history of humans in the last 10,000 years.

Chapter 7: Results

1 **The genomic signatures of natural selection**
2 **in admixed human populations**

3

4

5 Sebastian Cuadros Espinoza,^{1,2} Guillaume Laval,¹ Lluís Quintana-Murci,^{1,3} Etienne Patin^{1*}

6

7

8 ¹Institut Pasteur, Université de Paris, CNRS UMR2000, Human Evolutionary Genetics Unit,
9 Paris 75015, France

10 ²Sorbonne Université, Collège doctoral, Paris 75005, France

11 ³Chair Human Genomics and Evolution, Collège de France, Paris 75005, France

12

13 *Correspondence: epatin@pasteur.fr

14 Authors' emails: scuadros@pasteur.fr (S. C. E.), glaval@pasteur.fr (G. L.),

15 quintana@pasteur.fr (L. Q.-M.), epatin@pasteur.fr (E. P.)

16

17 **Abstract**

18 Admixture has been a pervasive phenomenon in human history, shaping extensively the
19 patterns of population genetic diversity. There is increasing evidence to suggest that
20 admixture can also facilitate genetic adaptation to local environments, i.e., admixed
21 populations acquire beneficial mutations from source populations, a process that we refer to
22 as *adaptive admixture*. However, the role of adaptive admixture in human evolution and the
23 power to detect it remain poorly characterized. Here, we use extensive computer simulations
24 to evaluate the power of several neutrality statistics to detect natural selection in the admixed
25 population, assuming multiple admixture scenarios. We show that statistics based on
26 admixture proportions, F_{adm} and LAD, show high power to detect mutations that are
27 beneficial in the admixed population, whereas other statistics, including iHS and F_{ST} , falsely
28 detect neutral mutations that have been selected in the source populations only. By combining
29 F_{adm} and LAD into a single, powerful statistic, we scanned the genomes of 15 worldwide,
30 admixed populations for signatures of adaptive admixture. We confirm that lactase
31 persistence and resistance to malaria have been under adaptive admixture in West Africans
32 and in Malagasy, North Africans and South Asians, respectively. Our approach also uncovers
33 new cases of adaptive admixture, including *APOLI* in Fulani nomads and *PKN2* in East
34 Indonesians, involved in resistance to infection and metabolism, respectively. Collectively,
35 our study provides evidence that adaptive admixture has occurred in human populations,
36 whose genetic history is characterized by periods of isolation and spatial expansions resulting
37 in increased gene flow.

38

39 **Introduction**

40 Over the last two decades, the search for molecular signatures of natural selection in the
41 human genome has played an integral part in understanding human evolution and population
42 differences in disease risk.¹⁻⁶ Genome scans for local adaptation have shed light on the
43 environmental pressures that populations have faced for the last 100,000 years, including
44 reduced exposure to sunlight, altitude-related hypoxia, new nutritional resources or exposure
45 to local pathogens. Candidate genes for local genetic adaptation have been identified based on
46 expected signatures of positive selection, such as extended haplotype homozygosity or strong
47 differences in allele frequencies between geographically diverse populations. In doing so,
48 selection studies have implicitly assumed that advantageous variation occurred in a single
49 population that has remained isolated from other populations since their separation. Yet,
50 ancient and modern genomics studies have clearly demonstrated that the last millennia of
51 human history have been characterized by large-scale spatial expansions, followed by
52 extensive gene flow.^{1,7,8} These findings indicate that most human populations descend from
53 admixture between formerly isolated groups, highlighting the need for detailed studies of the
54 expected genomic signatures of natural selection in admixed populations.

55 Several studies have searched for evidence of genetic adaptation in admixed populations
56 as a means to detect genes under positive selection in their ancestral sources, prior to
57 admixture.⁹⁻¹⁵ These studies showed that admixture can obscure signals of selective sweeps in
58 the source populations and proposed approaches to alleviate this problem, such as local
59 ancestry masking. Conversely, few studies have yet explored the patterns of diversity
60 expected under admixture with selection, as a means to detect genes under positive selection
61 in the admixed population since admixture.^{16,17} Studying the genomic signatures of *adaptive*
62 *admixture*, that is, positive selection in the admixed population of an allele that was beneficial
63 in one of its ancestral sources, could shed light on the role of gene flow in spreading
64 beneficial alleles among populations¹⁸ and the prevalence of recent, ongoing selection in
65 humans.

66 While an increasing number of studies have revealed how introgression from ancient
67 hominins, such as Neanderthals or Denisovans, facilitated genetic adaptation in modern
68 humans,¹⁹ the occurrence of adaptive admixture among modern humans remains largely
69 unexplored. Nonetheless, several empirical studies have reported candidate loci for positive
70 selection in admixed populations.^{16,20-38} A striking example is the Duffy-null $FY*B^{ES}$ allele,
71 which confers protection against *Plasmodium vivax* malaria.^{39,40} Selection signals have been

72 detected at the locus in diverse African-descent admixed populations from Madagascar, Cabo
73 Verde, Sudan and Pakistan,^{22,29,30,32,34} suggesting strong, ongoing selection owing to *vivax*
74 malaria in these regions. A variety of methods has been used to detect the signatures of
75 adaptive admixture, relying on classic neutrality statistics, such as iHS or F_{ST} , and deviations
76 from allele frequencies^{22,41} or admixture proportions^{21,24,27,30,31,34–38} expected under admixture
77 and neutrality.⁴² However, little is known about how these neutrality statistics behave under
78 scenarios of admixture with selection and, therefore, about the power of these statistics to
79 detect adaptive admixture. More worrying, it has been suggested that artifactual signals of
80 adaptive admixture can be observed because of errors in local ancestry inference (LAI) in
81 complex genomic regions^{43,44} and/or when the populations used as ancestral sources are poor
82 proxies of the true source populations.^{16,31} Lastly, reported signals of adaptive admixture are
83 still limited to few populations, relative to the large number of admixture events reported in
84 humans.^{1,7,8}

85 In this study, we compared the power of various neutrality statistics to detect adaptive
86 admixture, through computer simulations under different admixture with selection scenarios.
87 We then used a combination of the most powerful statistics to scan the genomes of 15
88 different admixed human populations from around the world and detect candidate loci for
89 adaptive admixture. In doing so, we confirm several, iconic signals of ongoing positive
90 selection since admixture and identify new cases that highlight pathogens as key drivers of
91 recent genetic adaptation in humans.

92

93 **Material and Methods**

94 **General simulation settings**

95 All the simulations were computed with the SLiM 3.2 engine⁴⁵ under the Wright-Fisher
96 model. Each simulation consisted of a 2-Mb long locus characterized by varying
97 recombination and mutation rates. For each simulation, we sampled the physical coordinates
98 of a random 2-Mb genomic window in the human genome, excluding telomeric and
99 centromeric regions, and assigned recombination rates based on the 1000 Genomes phase 3
100 genetic map⁴⁶ and mutation rates based on Francioli et al. mutation map.⁴⁷ To account for
101 background selection, which is thought to be prevalent in the human genome and could affect
102 the power of neutrality tests,⁴⁸ we simulated exon-like genetic elements positioned according
103 to the position of exons in the sampled 2-Mb genomic window. Each simulated exon is made
104 of positions under negative selection or under neutrality, mimicking non-synonymous and
105 synonymous positions, respectively. Deleterious mutations were set to occur three times more
106 frequently than neutral mutations, to account for codon degeneracy. The fitness effects of
107 deleterious mutations were sampled from the gamma distribution inferred in Europeans by
108 Boyko and colleagues.⁴⁹ For simulations that include positive selection, the beneficial
109 mutation was set to appear in the middle of the 2-Mb simulated locus and assumed to be
110 semi-dominant. Because we used computationally intensive forward-in-time simulations, we
111 rescaled population sizes and times according to N/λ and t/λ , with $\lambda = 10$, and used
112 rescaled mutation, recombination and selection parameters, $\lambda\mu$, λr and λs .⁴⁵ Of note, we
113 found that simulating background selection has little impact on the power to detect alleles
114 under strong positive selection in the admixed population ($s \geq 0.05$; data not shown).

115

116 **Admixture with selection models**

117 We performed simulations of a population that originates from admixture between two source
118 populations, referred to as P_1 and P_2 (Figure S1). We assumed that P_1 and P_2 contributed α_1
119 and α_2 admixture proportions to the admixed population, with $\alpha_1 + \alpha_2 = 1$. We also assumed
120 that P_1 and P_2 diverged T_{div} generations ago and the single-pulse admixture event occurred
121 T_{adm} generations ago. We simulated three scenarios of admixture with selection (Figures 1 and
122 S2). For *scenarios 1* and 2, a beneficial mutation was set to appear in the P_1 source population
123 and is transmitted to the admixed population with either the same selection coefficient
124 (*scenario 1*) or a selection coefficient set to 0 (*scenario 2*). For *scenario 3*, we adapted a
125 combination of recipes 9.6.2 and 14.7 from the SLiM manual,⁵⁰ introducing a set of “ancestry

126 marker” neutral mutations in the P_1 source population, and randomly choosing one of them to
127 become beneficial by setting its selection coefficient to $s > 0$ in the admixed population only.
128 We computed 500 simulations for each admixture with selection scenario, as well as 500
129 simulations for the null scenario (i.e., no positive selection). Because the goal of these
130 simulations was to compare the power of neutrality statistics to detect positive selection, only
131 the selection coefficient of the beneficial mutation s was given different values, ranging from
132 $s = 0.01$ to $s = 0.05$. All the other parameters were given fixed values: population sizes of
133 source and admixed populations $N = 10,000$; divergence time between source populations T_{div}
134 $= 2,000$ generations; admixture proportions $\alpha_1 = 0.35$ and $\alpha_2 = 0.65$; time of the single pulse
135 admixture event $T_{\text{adm}} = 70$ generations; time when the beneficial mutation appears $T_{\text{mut}} = 350$
136 generations ago.

137

138 **Power of explored neutrality statistics**

139 Neutrality statistics were computed for all genetic variants within the 2-Mb simulated loci
140 under no positive selection (H_0), and only for the selected mutation for simulated 2-Mb loci
141 under positive selection (H_1). We estimated detection power (i.e., the true positive rate, TPR)
142 for each statistic as the proportion of values under H_1 that are above a varying threshold value
143 under H_0 , corresponding to a given false positive rate (FPR). We computed F_{ST} , ΔDAF and
144 iHS using selink.⁵¹ We computed F_{ST} and ΔDAF between the admixed population and the
145 source population that does not experience positive selection. For iHS, we used a 200-kb
146 window and normalized the values by bins of similar derived allele frequency (DAF).

147 For the admixture-specific statistics, we introduced an allele frequency-based statistic,
148 F_{adm} , that measures the difference between x_i , the observed frequency of allele i in the
149 admixed population, and y_i , the expected allele frequency under admixture and neutrality. It
150 was shown that $y_i = \sum_p \alpha_p x_{i,p}$, which is the average of allele frequencies $x_{i,p}$ observed in the
151 source populations p weighted by estimated admixture proportions α_p , where $\sum_p \alpha_p = 1$
152 (ref.⁵²). Under neutrality, the squared difference between x_i and y_i , $(x_i - y_i)^2$, is the variance
153 of allele frequencies in the admixed population due to genetic drift.⁴² Thus, $(x_i - y_i)^2$ can be
154 interpreted as the genetic distance between the current admixed population and its ancestral
155 population at the time of admixture. Analogously to F_{ST} , this genetic distance can be used to
156 detect natural selection, as the change in frequency of a beneficial allele in time depends on
157 its selection coefficient.⁵³ F_{adm} is thus defined as follows:

158

159

$$F_{adm} = \frac{\sum_i (x_i - y_i)^2}{2(1 - \sum_i y_i^2)}$$

160

161 where $1 - \sum_i y_i^2$ is the expected heterozygosity in the admixed population, used here to allow
162 comparisons among SNPs.

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

When calculating F_{adm} in the simulated and observed data, the allele frequencies $x_{i,p}$ at the time of admixture were estimated by the allele frequencies in the current generation, which is accurate when genetic drift in source populations is weak or when admixture is recent. We used as admixture proportions the simulated proportions α_{sim} , for the simulated data, and the estimated proportions $\bar{\alpha}$, for the observed data, obtained by running ADMIXTURE v.1.23 (ref.⁵⁴; see section entitled ‘Empirical detection of adaptive admixture’). We verified with simulations that errors in the estimation of admixture proportions do not affect F_{adm} detection power (Figure S3A), by computing F_{adm} with α sampled from a normal distribution $\mathcal{N}(\mu = \alpha_{sim}, \sigma^2 = 0.026^2)$, 0.026 being the highest root-mean-square deviation of the ADMIXTURE estimation.⁵⁴ Additionally, we excluded sites where the observed allele frequency in the admixed population x_i is higher (or lower) than the maximum (or minimum) of the frequencies x_p in the source populations. Although this can reduce the detection power in *scenario 3*, this filter increases power for adaptive admixture scenario (Figure S3B), which is the focus of this study.

We also computed a LAI-based neutrality statistic, LAD, which measures the local ancestry deviation from the average genome-wide ancestry, defined as follows:

$$LAD_{w,p} = \alpha_{w,p} - \bar{\alpha}_p$$

where $\alpha_{w,p}$ is the admixture proportion from population p for a given window w , and $\bar{\alpha}_p$ is the estimated genome-wide admixture proportion. Natural selection has been proposed to bias the estimation of admixture proportions since the first estimates of this parameter were obtained.⁵⁵⁻⁵⁷ The rationale is that, when a beneficial allele is transmitted from a source population to the admixed population, estimated admixture proportions from this source population are expected to increase at the locus, relative to neutral loci. As single-marker estimates of admixture proportions are sensitive to errors in the estimation of allele frequencies, more powerful haplotype-based methods have preferentially been used to detect natural selection since admixture.³⁸

191 We used RFMix v1.5.4 to estimate local ancestry,⁵⁸ with default parameter values (except
192 for $-G$, which was replaced with the simulated T_{adm} value) and using the forward-backward
193 option with 3 expectation maximization steps. Because LAD is sensitive to phasing errors,⁵⁸
194 we incorporated potential phasing errors in our simulations by phasing, with SHAPEIT
195 v.4.2.1 (ref.⁵⁹), unphased diploid individuals obtained from the combination of two simulated
196 haploid individuals. Admixture proportions were estimated as the local ancestry inferred by
197 RFMix averaged across loci, for both the simulated and observed data.

198

199 **Sample size and source population choice scenarios**

200 We explored 5 different values of sample sizes for the two source populations and the
201 admixed population: $n = 20, 50, 100, 200$ and 500 individuals (Figures 2A and S4). When
202 exploring the values for a given population, sample sizes for the other two were fixed to $n =$
203 50 individuals. For the use of a proxy source population (Figure 2B), we simulated two
204 additional populations that diverge 400 generations ago from each of the two source
205 populations. We then used these proxy populations for F_{adm} and LAD calculations. To explore
206 the effect of the genetic distance (estimated by F_{ST}) between the proxy population and the true
207 source population on detection power, we set the population size of the proxies to 10,000,
208 4000, 1000 and 500, resulting in F_{ST} values of 0.005, 0.01, 0.02 and 0.03, respectively.

209 For the scenario of selection in the proxy source population only (Figure S5), we
210 simulated two additional populations that diverge 600 generations ago from each of the two
211 source populations. We randomly selected a mutation that occurred in the ancestral
212 population of the P_1 source population and its related proxy population and assigned it a
213 selection coefficient of $s = 0.02$ in the proxy population only, 599 generations ago. Under the
214 latter scenario, F_{adm} and LAD detect mutations that are not beneficial in the admixed
215 population and wrongly support positive selection in the P_2 source population (Figure S5B).
216 For comparison purposes, we thus compared F_{adm} and LAD distributions under this scenario
217 to those obtained under a simple scenario of adaptive admixture (*scenario 1*, Figure 1) where
218 the beneficial mutation is transmitted to the admixed population from the P_2 source
219 population (Figure S5D). In all these scenarios, the following parameters were given a fixed
220 value: $N_e = 10,000$; $T_{\text{div}} = 2,000$ generations; $\alpha_1 = 0.35$; $\alpha_2 = 0.65$; $T_{\text{adm}} = 70$ generations; $s =$
221 0.02 and $T_{\text{mut}} = 1,400$ generations ago.

222

223 **Complex admixture scenarios**

224 We estimated detection power under two additional admixture scenarios: a double pulse
225 model and a constant continuous model (Figure S6A,B). For these scenarios to be comparable
226 to the single pulse admixture scenario, we set the sum of the admixture proportions
227 contributed by each pulse to be equal to $\alpha_1 = 35\%$, and the average of the admixture dates to
228 be equal to 70 generations. Namely, under the double pulse model, the admixed population
229 originates from an admixture event that occurs 130 generations ago, between two source
230 populations with $\alpha_1 = 17.5\%$, and receives a second admixture pulse from P_1 10 generations
231 ago with $\alpha_1 = 17.5\%$. Under the constant continuous model, the admixed population also
232 originates from an admixture event occurring 130 generations ago, between two source
233 populations with $\alpha_1 = 35\% / 130 = 0.27\%$, when T_{adm} is not rescaled, and $\alpha_1 = 2.7\%$, when
234 T_{adm} is rescaled, but receives an additional pulse from P_1 of $\alpha_1 = 2.7\%$ at each generation until
235 present. In all scenarios, the following parameters were given a fixed value: population sizes
236 $N = 10,000$; $T_{\text{div}} = 2,000$ generations; $s = 0.02$ and $T_{\text{mut}} = 1,400$ generations ago.

237

238 **Admixture parameters**

239 Under the single pulse admixture model (Figure S1), we explored detection power as a
240 function of different model parameters (Figures 3 and S7-S11; Table S1). In total, 32,956
241 compatible parameter combinations were explored. The number of simulations was thus
242 reduced from 500 to 100, to limit computational burden. For the frequency of the beneficial
243 mutation on the source population at the time of admixture, instead of conditioning on the
244 frequency within simulations (which would have drastically increased computations), we
245 introduced the beneficial mutation T_{mut} generations ago, in the source population, based on
246 previous results.⁶⁰ For each statistic and parameter combination, we calculated the proportion
247 of simulated sites under selection that were recovered using a threshold of $\text{FPR} = 5\%$. We
248 then averaged the power across demographic parameter values to obtain a single value for
249 each combination of T_{adm} , α_1 and s . We performed a similar procedure to obtain a single value
250 for each combination of T_{adm} , α_1 , and one of the other parameters (e.g., T_{div} and N ; Figures S7-
251 S11).

252

253 **Non-stationary demography**

254 We estimated detection power under five alternative demographic scenarios (Figures 2C and
255 S6C), each with 500 simulations under adaptive admixture and 500 simulations with no
256 positive selection. Demographic scenarios include: (i) a recent expansion of the source
257 population, where the source population undergoes an expansion with a 5% growth rate since

258 T_{adm} , from an initial $N = 10,000$; (ii) a recent expansion of the admixed population, where the
259 admixed population undergoes an expansion with a 5% growth rate since T_{adm} , from an initial
260 $N = 10,000$; (iii) an old expansion of the source population, where the source population
261 undergoes an expansion with a 5% growth rate since $T_{\text{adm}} + 500$ generations, from an initial N
262 $= 10,000$; (iv) an old bottleneck in the source population, where the source population
263 undergoes a 10-fold size reduction from $T_{\text{div}} - 50$ to T_{div} , from an initial $N = 10,000$; and (v) a
264 recent bottleneck in the admixed population, where the admixed population undergoes a 10-
265 fold size reduction from $T_{\text{adm}} - 50$ to T_{adm} , from an initial $N = 10,000$. We compared these
266 scenarios to a constant population size scenario, with the same general parameters and the
267 size of all populations fixed to $N = 10,000$. In all scenarios, the following parameters were
268 given a fixed value: $T_{\text{div}} = 2,000$ generations; $\alpha_1 = 0.35$; $\alpha_2 = 0.65$; $T_{\text{adm}} = 70$ generations; $s =$
269 0.02 and $T_{\text{mut}} = 1,400$ generations ago.

270

271 **Empirical detection of adaptive admixture**

272 We analysed the genomes of 15 admixed populations to search for signals of adaptive
273 admixture. The datasets and references for all admixed and source populations can be found
274 in Table S2, as well as the final number of SNPs used after merging the datasets for admixed
275 and source populations. For each merged dataset, we: (i) excluded sites with a proportion of
276 missing genotypes $> 5\%$, using PLINK v.2.0 (ref.⁶¹) (ii) excluded A/T and C/G variant sites;
277 (iii) excluded first and second degree-related individuals (kinship coefficient > 0.08 computed
278 with KING v2.2.2; ref.⁶²) and (iv) performed phasing using SHAPEIT v.4.2.1, using default
279 parameter values. Additionally, we verified the validity of an admixture model for each set of
280 source/admixed populations (Table S2), by computing admixture f_3 statistics with admixr
281 package v.0.7.1 (ref.⁶³).

282 Admixture proportions were obtained by running ADMIXTURE v.1.23, considering the K
283 value producing the lowest cross-validation error and a set of “independent” SNPs obtained
284 by running the ‘--indep-pairwise’ command with PLINK v.2.0, with the following
285 parameters: 50-SNP window, 5-SNP step, and r^2 threshold of 0.5. We also verified for each
286 studied admixed population that the K value with the fewest cross-validation errors matches
287 the number of source populations. Local ancestry was inferred with RFMix v.1.5.4, after
288 excluding 2 Mb at telomeres and centromeres of each chromosome, as well as monomorphic
289 sites and singletons, and using default parameter values except for the generation time ‘-G’,
290 which was given a value based on literature (Table S2).

291 We combined the SNP ranks for these two statistics using Fisher's method,⁶⁴ defined as
292 follows:

$$293 \quad X_{2k}^2 = -2 \sum_{i=1}^k \ln(r_i)$$

294 where r_i is defined as the rank of a given SNP for the statistic i , divided by the total number of
295 analysed SNPs (i.e., the empirical P -value), and $k = 2$ is the number of statistics.

296 Using simulations, we verified that this statistic followed a chi-squared distribution with
297 $2k = 4$ degrees of freedom under no positive selection (Figure 4A), including when the
298 admixed population experienced a 10-fold bottleneck. In these simulations, we used the same
299 parameter values as those in Figure 2C for the "constant size" and "bottleneck in the admixed
300 population" scenarios. Statistical significance was defined based on Bonferroni correction: we
301 considered a P -value threshold of 0.05 divided by the number of 0.2-cM RFMix windows
302 analysed (all SNPs within the same window had the same local ancestry value), which
303 yielded, on average, a P -value threshold of 3.5×10^{-6} (Table S3). To reduce the number of
304 false positives due to positive selection in a proxy source population only (Figure S5), we
305 computed iHS in the two source populations and excluded from the list of candidate genes
306 any locus that includes SNPs with both a $|iHS| > 2$ in one source population and an excess of
307 local ancestry from the other source population (Figure S5). To annotate the different signals
308 that passed this threshold, we chose the protein coding gene within 250-kb of the variant with
309 the highest V2G score.⁶⁵

310

311 **Results**

312 **Power estimation under different models of admixture with selection**

313 To estimate the power to detect positive selection in admixed populations, we performed
314 extensive forward-in-time simulations of a population that originates from admixture between
315 two source populations (Figure S1). We introduced a beneficial mutation in one of the source
316 populations, with a varying selection coefficient (Material and Methods). We considered three
317 different scenarios of admixture with selection (Figures 1A and S2). *Scenario 1* corresponds
318 to adaptive admixture, where the admixed population inherits an allele that is beneficial in
319 one of its source populations: the mutation is under positive selection in the source
320 population, is transmitted to the admixed population and remains beneficial – with the same
321 selection coefficient – in the admixed population. In *scenario 2*, the beneficial allele is under
322 positive selection in the source population, is transmitted to the admixed population and
323 becomes neutral in the admixed population only. We simulated this scenario to verify if some
324 neutrality statistics wrongly support positive selection in the admixed population because of a
325 residual signal inherited from the source population. At the same time, this scenario is also
326 useful to evaluate the power to detect residual signals of positive selection in the admixed
327 population, as a means to detect genes under positive selection in source populations that no
328 longer exist in an unadmixed form.^{9–15} Finally, in *scenario 3*, a neutral mutation in the source
329 population becomes beneficial in the admixed population only, at the time of admixture. This
330 case is used to determine how neutrality statistics behave when natural selection operates
331 since admixture on standing neutral variation.

332 We evaluated the performance, under each scenario, of three classic neutrality statistics,
333 F_{ST} , ΔDAF and iHS , as well as two statistics that are specifically designed to detect selection
334 in an admixed population: F_{adm} , which is proportional to the squared difference between the
335 observed and the expected allele frequency in the admixed population,^{22,42,57} and LAD, the
336 difference between the admixture proportion at the locus and its genome-wide average,³⁸
337 estimated based on local ancestry inference (LAI) by RFMix (Material and Methods).⁵⁸
338 Receiver operating characteristic (ROC) curves indicate that both the classic neutrality
339 statistics and F_{adm} and LAD are powerful to detect adaptive admixture (*scenario 1*) when the
340 selection coefficient $s = 0.05$ (>70% detection power for a false positive rate (FPR) of 5%;
341 Figures 1B and S2), in agreement with a previous study.¹⁶ Nevertheless, the power of F_{ST} ,
342 ΔDAF and iHS is also high when the mutation is beneficial in the source population and is no
343 longer selected in the admixed population (*scenario 2*), indicating that these statistics wrongly

344 detect selection in the source population as selection in the admixed population. In contrast,
345 F_{adm} and LAD detect adaptive admixture specifically, as their power under *scenario 2* is low
346 or nil (Figure 1B). Of note, our simulations also imply that the power of classic statistics is
347 substantial when using the admixed population as a means to detect selection in the source
348 populations (>65% detection power when $s = 0.05$ and $\text{FPR} = 5\%$). Finally, LAD and iHS
349 showed a reduced power to detect selection in the admixed population when the mutation is
350 neutral in the source populations (*scenario 3*), relative to the adaptive admixture case
351 (*scenario 1*). This may stem from the fact that, under *scenario 3*, the beneficial mutation has
352 been selected for fewer generations than in *scenario 1*, resulting in a weaker signal for classic
353 statistics. Furthermore, this scenario is similar to selection on standing variation, where the
354 adaptive mutation may be present on several haplotypes, making it harder to detect.⁶⁶

355 Collectively, our simulations indicate that F_{adm} and LAD are the only studied statistics
356 that have substantial power to specifically detect strong, ongoing selection in the admixed
357 population and have more power to detect adaptive admixture than post-admixture selection
358 on standing variation. Because our objective is to detect the signatures of positive selection in
359 the admixed population, and not in the source populations, we based all subsequent analyses
360 on the F_{adm} and LAD statistics.

361

362 **Effects of the study design**

363 We investigated how sample size and the choice of source populations affect the power of
364 F_{adm} and LAD to detect adaptive admixture signals (Material and Methods). We explored
365 sample sizes ranging from $n = 20$ to $n = 500$, for both the admixed and the source populations.
366 We found that $n = 100$ already provides optimal power, because the variance of neutrality
367 statistics is virtually unchanged when $n \geq 100$ (Figures 2A and S4A). Conversely, we found
368 that when $n < 50$, sampling error increases the variance of F_{adm} and LAD null distributions, by
369 as much as 5 times, and ultimately decreases detection power by up to 40% ($\text{FPR} = 5\%$).
370 Interestingly, LAD detection power is not affected when the sample size of the source
371 populations is decreased, even when $n = 20$ (Figure S4B). Consistently, RFMix accuracy was
372 shown to be only minimally reduced when the sample size of reference panels is as small as n
373 $= 3$, as it uses both source and admixed individuals for LAI.⁵⁸

374 Because obtaining genotype data for the true source populations of an admixed population
375 is difficult, if not impossible, population geneticists often use related, present-day populations
376 as proxies, which may lead to false adaptive admixture signals.^{16,31} We explored how
377 detection power is affected by the genetic distance between the true source population and a

378 related population used as a proxy for F_{adm} and LAD computations (Material and Methods).
379 We observed a difference in performance between F_{adm} and LAD, the latter being more robust
380 to the use of a proxy (Figure 2B). LAD maintains similar detection power even if the
381 divergence between the true and proxy populations is $F_{ST} = 0.01$, whereas power decreases by
382 25% for F_{adm} . Such a difference in power may result from the nature of the two statistics. In
383 the case of F_{adm} , the expected allele frequency is directly estimated from the allele frequencies
384 observed in the proxy, and these frequencies are decreasingly correlated with those in the true
385 source population, as their divergence increases. On the other hand, LAD is derived from LAI
386 by RFMix, which has been shown to be robust to the use of proxy reference populations.⁵⁸

387 Nonetheless, we identified a potentially problematic scenario for both F_{adm} and LAD
388 involving population proxies: when the selection event occurs specifically in the proxy source
389 population (i.e., the mutation is not selected in both the true source and the admixed
390 populations; Figure S5), spurious deviations in local ancestry and in allele frequencies were
391 observed in the admixed population. Specifically, this generates an excess of local ancestry
392 from the other source population and expected allele frequencies higher than those observed
393 in the admixed population (Figures S5A and S5B). We found that this scenario produces
394 weaker LAD values (i.e., lower detection power) but stronger F_{adm} values (i.e., higher
395 detection power), relative to an adaptive admixture event (Figure S5C-F). To remediate this,
396 we performed a selection scan in the proxy population using a single-population statistic, iHS,
397 and excluded the top 1% values. In doing so, we managed to exclude approximately 90% of
398 the outlier values of F_{adm} and LAD generated by this scenario. More importantly, because
399 there is no correlation between iHS in the source population and F_{adm} or LAD in the case of
400 adaptive admixture, none of the outlier values generated by a true adaptive admixture event
401 were excluded by this analysis step (Figure S5G,H).

402

403 **Effects of the admixture model and non-stationary demography**

404 Several studies have shown that admixture in humans has often involved multiple admixture
405 pulses from two or more source populations.^{8,51,67-71} We thus estimated the detection
406 performance of F_{adm} and LAD under admixture models that are more complex than the single
407 admixture pulse. We found that the power to detect adaptive admixture is only moderately
408 reduced under a two-pulse admixture model or a constant, continuous admixture model: the
409 true positive rate (TPR) decreases by <11% at a FPR = 5%, relative to the single pulse model
410 (Figure S6A,B). This suggests that our power estimations are valid for a variety of admixture
411 models.

412 Assuming a single-pulse admixture model, we then explored how detection power is
413 impacted by key parameters of the adaptive admixture model, including the strength of
414 selection s , the admixture time T_{adm} , the admixture proportion α and the divergence time
415 between source populations T_{div} (Figures 3 and S7-S11; Table S1). As expected, we found
416 that detection power is high only when the selection coefficient s is strong; the TPR is up to
417 94% and 27% when $s = 0.05$ and 0.01 , respectively (FPR = 5%; Figure 3). Power is also
418 determined by the admixture time T_{adm} , as it affects the duration of selection; the TPR is up to
419 94% and 21% when $T_{\text{adm}} \geq 70$ and ≤ 20 generations, respectively. Interestingly, we observed
420 that the higher the admixture proportion α (from the source population where the selected
421 mutation appeared), the lower the detection power. Power decreases particularly when $\alpha >$
422 0.65 , probably because of a threshold effect: if the beneficial allele is at high frequency and
423 e.g., $\alpha = 0.9$, there is little room for the observed allele frequency or local ancestry to deviate
424 from its expectation, making it hard to detect. Finally, as the divergence time between source
425 populations decreases, the detection power of LAD is reduced by $\sim 15\%$ (51% vs. 67%, when
426 $T_{\text{div}} = 500$ or 2,000 generations, respectively; Figure S7), whereas that of F_{adm} is not affected
427 (59% vs. 55%, when $T_{\text{div}} = 500$ or 2,000 generations, respectively). The reduced detection
428 power of LAD is probably due to the decreased accuracy of RFMix when T_{div} decreases.⁷²

429 We also estimated power under scenarios where demography deviates from a constant
430 population size model. Indeed, demographic events, such as bottlenecks, have been shown to
431 alter the performance of several neutrality statistics.⁷³⁻⁷⁹ We simulated 5 demographic
432 scenarios, including 10-fold bottlenecks and 5% growth rate expansions in either the admixed
433 or the source populations (Material and Methods). We found that detection power is
434 minimally affected under all expansion models (TPR decrease of 5% at a FPR = 5%; Figure
435 2C). In contrast, detection power is reduced by as much as 50% under the scenario where a
436 10-fold bottleneck is introduced in the admixed population, relative to the stationary model.
437 This is probably explained by the increased variance of F_{adm} and LAD null distributions under
438 this scenario (Figure S6C). Finally, detection power of both F_{adm} and LAD is minimally
439 affected when the 10-fold bottleneck is introduced in the source populations, either few
440 generations after their divergence or before the admixture pulse (TPR decrease of 5% at a
441 FPR = 5%; Figure 2C), suggesting that both statistics are relatively robust to increased genetic
442 drift occurring in the source populations.

443

444 **Empirical detection of adaptive admixture in humans**

445 We next sought to detect candidate genes for adaptive admixture in humans, by scanning,
446 with both F_{adm} and LAD statistics, the genomes of 15 worldwide populations (Table S2) that
447 have experienced at least one admixture event in the last 5,000 years (i.e., the upper detection
448 limit set for accurate local ancestry inference⁸⁰). To improve detection power and facilitate
449 candidate prioritization, we combined the empirical P -values of both statistics with Fisher's
450 method,⁶⁴ used here as a combined test for positive selection since admixture. We confirmed
451 with simulations that the Fisher's score follows a χ^2 distribution with 4 degrees of freedom
452 under the null hypothesis of absence of positive selection and when assuming different
453 demographic scenarios (Figure 4A). Consistently, we found that F_{adm} and LAD statistics are
454 not correlated under the null hypothesis (Spearman's coefficient = 0.03), whereas they are
455 correlated under adaptive admixture (Spearman's coefficient = 0.96). Importantly, we found
456 that Fisher's method increases detection power under unfavourable scenarios, relative to each
457 individual statistic (Figure 4B). In particular, Fisher's method improves power when the
458 admixed population experienced a 10-fold bottleneck, when admixture is recent ($T_{\text{adm}} = 10$
459 generations) or when using a proxy population that experienced strong drift (F_{ST} with the true
460 source population = 0.02). Given the limited knowledge on the past population sizes of the
461 studied populations, which could increase FPR (Figure S6C), we applied a conservative
462 Bonferroni correction on Fisher's P -values, considering the number of RFMix genomic
463 windows as the number of tests (all SNPs within a given window have the same value for
464 LAD). This yielded a P -value threshold of approximately $P = 3.5 \times 10^{-6}$ (Table S3). Finally, we
465 verified that the empirical distribution of Fisher's P -values is uniform in all studied
466 populations and found an excess of low P -values for several populations (Figure S12),
467 suggesting that adaptive admixture has occurred in these groups.

468 Our genome scans identified a number of previously reported signals of adaptive
469 admixture. Among these, we found the *HLA* class II locus in Bantu-speaking populations
470 from Gabon³¹ (Figures 5A and 5C; top ranking SNP identified in *HLA-DPA1* [MIM: 142880];
471 $P = 7.9 \times 10^{-8}$; expected frequency of 0.33 vs. observed frequency of 0.70), the *HLA* class I
472 locus in Mexicans^{27,35,37,81} (Figure S13; top ranked SNP identified in *ABCF1*; $P = 2.2 \times 10^{-6}$;
473 expected frequency of 0.013 vs. observed frequency of 0.039), the lactase persistence-
474 associated *LCT/MCM6* locus [MIM: 223100] in the Fulani nomads of Burkina Faso⁸² (Figure
475 6A; top ranked SNP identified in *CCNT2* [MIM: 603862]; $P = 1.1 \times 10^{-6}$; expected frequency
476 of 0.12 vs. observed frequency of 0.47), and the *ACKR1* gene (previously referred as *DARC*
477 [MIM: 613665]) in African-descent populations from Madagascar, the Sahel and
478 Pakistan.^{29,30,34} (Figures 5B, 5D and S13). For the latter locus, the top-ranking variant is

479 *rs12075* in the Malagasy ($P = 3.4 \times 10^{-9}$; expected frequency 0.45 vs. observed frequency of
480 0.93), as previously found.³⁰ This variant, also known as the Duffy-null *FY*B^{ES}* allele [MIM:
481 110700], confers resistance against *Plasmodium vivax* infection in sub-Saharan Africans.^{39,40}
482 Together, these results confirm that our conservative approach can recover strong, well-
483 documented signals of adaptive admixture.

484

485 **New candidate genes for adaptive admixture**

486 We found several novel candidate loci for adaptive admixture (Figures 6 and S14), among
487 which the *MYH9/APOL1* [MIM: 603743] locus in the Fulani (Figures 6A and 6C; $P = 1.3 \times 10^{-7}$;
488 top ranked SNP in *IFT27* [MIM: 615870]; expected frequency of 0.15 vs. observed
489 frequency of 0.45). Common *APOL1* variants confer both protection against human African
490 trypanosomiasis (HAT, or sleep sickness) and susceptibility to common kidney diseases
491 [MIM: 612551] in African-descent individuals.⁸³ Another candidate is the *PKN2* [MIM:
492 602549] locus in East Indonesians ($P = 1.1 \times 10^{-6}$; top ranked SNP in *ZNF326* [MIM: 614601];
493 expected frequency of 0.27 vs. observed frequency of 0.46), which shows a large excess of
494 Papuan ancestry (Figure 6B and 6D). *PKN2* plays a role in cellular signal transduction
495 responses and has been reported as involved in the regulation of glucose metabolism in
496 skeletal muscle.⁸⁴ A nearby locus, *LRRC8B* [MIM: 612888], has been reported as a candidate
497 for positive selection in Solomon Islanders,⁵¹ although it did not show signals for adaptive
498 admixture in this population. A unique, strong signal was detected at the *ARRDC4/IGF1R*
499 [MIM: 147370] locus in Solomon islanders ($P = 7.4 \times 10^{-9}$; top ranked SNP close to *ARRDC4*;
500 expected frequency of 0.09 vs. observed frequency of 0.58), where an excess of East Asian-
501 related ancestry was observed (Figures S14B and S14F). This locus was previously identified
502 as a candidate for positive selection in Near and western Remote Oceanians.⁵¹ *ARRDC4* is an
503 arrestin that plays important roles in glucose metabolism and immune response to enterovirus
504 infection,⁸⁵ whereas *IGF1R*, the receptor for the insulin-like growth factor, is a key
505 determinant of body size and growth.^{86,87} A last example is *CXCL13* [MIM: 605149] in the
506 Nama pastoralists from South Africa (Figures S14A and S14E; $P = 2.3 \times 10^{-6}$; top ranked SNP
507 identified in *CXCL13*; expected frequency of 0.51 vs. observed frequency of 0.80). The
508 *CNOT6L/CXCL13* locus has previously been reported as suggestively associated with
509 tuberculosis (TB) risk in South African populations with San ancestry.⁸⁸ However, we found
510 that the top-ranking variants show outlier extended haplotype homozygosity in the Ju'hoansi
511 San, used as source population (iHS = -3.12), while European ancestry is in excess at the

512 locus in the Nama, suggesting a spurious signal due to positive selection in the proxy source
513 population (Figure S5).

514 Lastly, we detected suggestive signals of adaptive admixture at genes shown to be strong
515 candidates for positive selection, including the *MCM6/LCT* locus in the Bantu-speaking
516 Bakiga of Uganda (Figure S15; $P = 4.3 \times 10^{-6}$; top ranked SNP in *CCNT2*; expected frequency
517 of 0.15 vs. observed frequency of 0.31) and *TNFAIP3* [MIM: 191163] in East Indonesians,
518 who show an excess of Papuan-related ancestry at the locus (Figure 6B; $P = 5.0 \times 10^{-6}$; top
519 ranked SNP in *TNFAIP3*; expected frequency of 0.27 vs. observed frequency of 0.43). The
520 *TNFAIP3* locus has not only been reported as evolving under positive selection in Papuans⁵¹
521 but also as adaptively introgressed from Denisovans.^{51,89–91} *TNFAIP3* plays an important role
522 in human immune tolerance to pathogen infections.⁹² Collectively, these results indicate that
523 adaptive admixture has occurred in various admixed populations around the world, and
524 highlight the immune system and nutrient metabolism as important targets of recent genetic
525 adaptation.

526

527 **Discussion**

528 In this study, we evaluated the power of several neutrality statistics to detect loci under
529 positive selection in admixed populations and used these statistics to explore cases of adaptive
530 admixture in the genomes of 15 worldwide human populations. Although F_{adm} and LAD, or
531 closely related statistics based on the difference between observed and expected allele
532 frequencies and admixture proportions, have been used in several empirical studies, their
533 power has not been thoroughly evaluated. Here, we showed that these statistics are powerful
534 to detect adaptive admixture and have no power to detect residual signals of positive selection
535 in the source populations. Thus, F_{adm} and LAD are suited to search for loci under positive
536 selection in admixed populations since admixture, particularly when selection is strong (i.e., s
537 ≥ 0.05), admixture is relatively old (i.e., $T_{\text{adm}} > 2,000$ years) and the admixture proportion is
538 moderate-to-low (i.e., $\alpha < 0.6$). Notably, we found that power is marginally affected when
539 admixture has been recurrent, a feature that is convenient given the difficulty to distinguish
540 between single-pulse, double-pulse or more complex admixture models from the genetic
541 data.^{8,51,67-71} Furthermore, F_{adm} is more powerful than LAD when selection occurs in the
542 admixed population only and when the divergence time between source populations is low
543 ($T_{\text{div}} = 500$ generations), whereas LAD is more powerful than F_{adm} when source sample sizes
544 are low (i.e., $n = 20$) and when the true and proxy source populations are distantly related
545 (i.e., $F_{\text{ST}} \geq 0.01$; Table S4). The latter result is consistent with the known robustness of LAI to
546 cases where the populations used as reference sources are poor proxies of the true source
547 populations.⁵⁸ Nonetheless, caution must be taken when handling population proxies, as
548 selection occurring only in the proxy population can produce artifactual genomic signals, for
549 both LAD and F_{adm} , that might be misinterpreted as adaptive admixture.^{16,31,51} We suggest
550 that performing selection scans on the proxy source populations can help distinguish false
551 from true adaptive admixture signals. We also caution that F_{adm} calculation relies on the
552 accurate estimation of admixture proportions, which can be biased under certain scenarios.⁹³
553 Finally, we found that combining F_{adm} and LAD statistics into a unique statistic, based on the
554 Fisher's method, provides well-calibrated P -values under different models and substantially
555 increases power under several realistic admixture with selection scenarios, relative to
556 individual statistics.

557 When applying this combined method on the empirical data, we identified several
558 previously reported candidate variants for adaptive admixture. These include the *ACKRI*
559 Duffy-null allele detected in admixed populations from Madagascar,²⁰ the Sahel³⁴ and

560 Pakistan,²⁹ the lactase persistence -13910 C>T *LCT* allele in the Fulani from West Africa⁸²
561 and *HLA* alleles in Bantu-speaking populations from western Central Africa³¹ and
562 Mexicans.^{27,35,37,81} These candidate loci were detected previously based on LAD only, or in
563 combination with classic neutrality statistics. However, the detection of natural selection with
564 the LAD statistic has previously been questioned, because deviations in local ancestry can be
565 explained as artifacts of long-range linkage disequilibrium (LD), which was not properly
566 modelled by the first-generation LAI methods.⁴³ Our analyses reveal that these genomic
567 regions not only show outlier LAD values, but also outlier F_{adm} values. Because F_{adm} only
568 depends on allele frequencies at the SNP of interest, these results support the view that the
569 observed signals of adaptive admixture are true and unlikely to be explained by incorrectly
570 modelled LD.

571 Our results also highlight novel signals of adaptive admixture, such as the *APOLI/MYH9*
572 locus in the Fulani nomads of West Africa. Interestingly, an *APOLI* haplotype of non-African
573 origin, named G3, was shown to be under positive selection in the Fulani of Cameroon,⁹⁴ in
574 line with the excess of non-African ancestry that we detected at the locus in the Fulani from
575 Burkina Faso. Nevertheless, the physiological effect of the G3 variants is still debated:
576 experimental work suggests that the G3 haplotype has no lytic activity against *Trypanosoma*
577 parasites and is not associated with increased susceptibility to common kidney diseases in
578 African Americans.⁹⁵ Alternatively, the significant excess of non-African ancestry observed
579 at the locus may be due to strong negative selection against HAT-resistance *APOLI* alleles
580 (i.e., G1 and G2 haplotypes), in regions where the incidence of sleeping sickness is low, such
581 as Burkina Faso.⁹⁶ As they do not confer a selective advantage in *Trypanosoma brucei*-free
582 regions, the G1 and G2 haplotypes only strongly increase the risk for chronic kidney
583 diseases⁸³ and thus become disadvantageous. Further epidemiological and experimental work
584 will be needed to confirm this hypothesis.

585 In accordance with our simulation study, several of the putatively selected alleles
586 detected here are known to be under strong positive selection in humans, including alleles in
587 *ACRKI*,⁹⁷⁻⁹⁹ *LCT*^{100,101} or *HLA*.⁸¹ Given that we focused on admixture events occurring
588 during the five last millennia, only alleles that confer a very strong selective advantage can
589 leave detectable signatures in the genomes of the studied admixed individuals. In addition to
590 their confirmatory nature, these results improve our understanding of the selective advantage
591 conferred by these well-known beneficial alleles. First, because F_{adm} and LAD detect natural
592 selection since admixture only, selection studies in recently admixed populations represent a
593 valuable tool to detect recent ongoing selection. Second, admixed and source populations

594 have often lived in different environments, so evolutionary studies of adaptive admixture can
595 help refine correlations between signatures of natural selection and environmental pressures.
596 An illustrative example is the Duffy-null $FY*B^{ES}$ allele, which is fixed or nearly fixed in most
597 sub-Saharan African populations.⁹⁹ It has long been proposed that natural selection has
598 favoured this allele because it protects against malaria due to *Plasmodium vivax*.¹⁰² Indeed,
599 cellular experiments have shown that the parasite depends on the ACKR1 protein for
600 erythrocytic infection.^{39,40} However, recent studies have casted doubt on this result, because
601 *P. vivax* has been detected in $FY*B^{ES}$ homozygous carriers,^{103,104} suggesting that parasite
602 invasion is possible when its human receptor ACKR1 is absent. We and others have found
603 signatures of adaptive admixture for the $FY*B^{ES}$ allele in African-descent admixed
604 populations from Madagascar,^{12,20} Cabo Verde,²³ the Sahel³⁴ and Pakistan,¹⁹ but not in North
605 Americans or South Africans.^{16,31} Evidence of ongoing positive selection for Duffy negativity
606 is thus confined to regions where the current incidence of *P. vivax* malaria is estimated to be
607 high.¹⁰⁵ These findings thus support the view that resistance to *vivax* malaria is the main
608 evolutionary force driving the frequency of the $FY*B^{ES}$ allele in humans.

609 Overall, our study reports evidence that recent admixture has facilitated human genetic
610 adaptation to varying environmental conditions. It has been proposed that gene flow can
611 promote rapid evolution when the demographic structure of a species is unstable.¹⁸ Our
612 findings support this view, as *Homo sapiens* is a structured species that has settled a large
613 variety of ecological niches and has undergone large-scale, massive dispersals followed by
614 extensive gene flow.^{1,7,8} We thus anticipate that more cases of adaptive admixture in humans
615 will soon be uncovered, thanks to methodological and technological advances. Importantly,
616 given the highly conservative nature of our approach, it is very likely that we do not recover
617 variants that have probably been weakly to mildly selected since admixture, such as *TNFAIP3*
618 in Indonesian populations of Papuan-related ancestry^{51,89-91} or the *MCM6/LCT* locus in the
619 Bantu-speaking Bakiga from Uganda.³¹ The use of new, accurate LAI methods^{80,106} and the
620 development of novel powerful neutrality statistics, such as the integrated decay in ancestry
621 tracts (iDAT),²³ and model-based probabilistic frameworks¹⁰⁷ are promising paths to improve
622 the power to detect adaptive admixture, while better accounting for the demography of
623 admixed populations. Furthermore, many human traits are known to be highly polygenic,
624 suggesting that polygenic adaptation is a key driver of phenotypic evolution,¹⁰⁸ highlighting
625 the need for new methods to detect polygenic selection since admixture.¹⁰⁹ Finally, genomic
626 studies of adaptive admixture are expected to be more powerful when admixture is ancient,
627 but statistical tests for admixture in modern genomes have low power when admixture time is

628 older than 5,000 years.⁸ Ancient genomics studies offer a great opportunity to circumvent this
629 limitation, by revealing how human populations interacted in the past and how beneficial
630 alleles have spread in time and space.^{41,110}

631 **Supplemental Data**

632 Supplemental data include fifteen figures and four tables.

633

634 **Acknowledgements**

635 We thank all volunteers participating in this research; Sophie Créno and the HPC Core
636 Facility of Institut Pasteur (Paris) for the management of computational resources; the two
637 anonymous reviewers for their useful comments; Omar Alva Sanchez, Denis Pierron, Thierry
638 Letellier, Mario Vicente, Carina Schlebusch, Andres Moreno-Estrada, Andres Ruiz-Linares
639 and the Health Aging and Body Composition (Health ABC) Study for kindly providing access
640 to their data. We also thank Javier Bougeard, Lara Rubio Arauna, Jérémy Choin, Maxime
641 Rotival, Paul Verdu and Olivier Tenaillon for helpful discussions. S.C.-E. is supported by
642 Sorbonne Université Doctoral College, the Inception program (Investissement d’Avenir grant
643 ANR-16-CONV-0005) and the Institut Pasteur. The laboratory of Human Evolutionary
644 Genetics is supported by the Institut Pasteur, the Collège de France, the CNRS, the Fondation
645 Allianz-Institut de France, the French Government’s Investissement d’Avenir programme,
646 Laboratoires d’Excellence ‘Integrative Biology of Emerging Infectious Diseases’ (ANR-10-
647 LABX-62-IBEID) and ‘Milieu Intérieur’ (ANR-10-LABX-69-01), the Fondation de France
648 (n°00106080), the Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339214)
649 and the French National Research Agency (ANR-19-CE35-0005).

650

651 **Declaration of interests**

652 The authors declare no competing interests.

653

654 **Web Resources**

655 SLiM software, <https://messerlab.org/slim/>

656 RFMix software, https://www.dropbox.com/s/cmq4sadh9gozi9/RFMix_v1.5.4.zip

657 PLINK software, <https://www.cog-genomics.org/plink/2.0/>

658 SHAPEIT software, <https://odelaneau.github.io/shapeit4/>

659 ADMIXTURE software, <https://dalexander.github.io/admixture/index.html>

660 admixr R package, <https://cran.r-project.org/web/packages/admixr/index.html>

661 selink software, <https://github.com/h-e-g/selink>

662 OMIM, <http://www.omim.org/>

663 1000 Genomes Phase 3 and HGDP genomic data, <https://www.internationalgenome.org/data>

664 Estonian Biocentre public genomic data, <https://evolbio.ut.ee>
665 Jakobsson Lab genomic data, <http://jakobssonlab.iob.uu.se/data/>
666 European Genome-Phenome archive, <https://ega-archive.org/>
667 The dbGAP database, <https://dbgap.ncbi.nlm.nih.gov/>

668

669 **Data and Code Availability**

670 Accession numbers for the genotype data used in this study are listed in Table S2. All SLiM
671 parameter files can be found here: <https://github.com/h-e-g/ADAD>.

672

673 **References**

- 674 1. Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., and Willerslev, E.
675 (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310.
- 676 2. Novembre, J., and Di Rienzo, A. (2009). Spatial patterns of variation due to natural
677 selection in humans. *Nat Rev Genet* 10, 745–755.
- 678 3. Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary
679 Genetics. *Cell* 177, 184–199.
- 680 4. Rees, J.S., Castellano, S., and Andrés, A.M. (2020). The Genomics of Human Local
681 Adaptation. *Trends in Genetics* 36, 415–428.
- 682 5. Fan, S., Hansen, M.E.B., Lo, Y., and Tishkoff, S.A. (2016). Going global by adapting
683 local: A review of recent human adaptation. *Science* 354, 54–59.
- 684 6. Mathieson, I. (2020). Human adaptation over the past 40,000 years. *Current Opinion in*
685 *Genetics & Development* 62, 97–104.
- 686 7. Pickrell, J.K., and Reich, D. (2014). Toward a new history and geography of human genes
687 informed by ancient DNA. *Trends in Genetics* 30, 377–389.
- 688 8. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S.
689 (2014). A Genetic Atlas of Human Admixture History. *Science* 343, 747–751.
- 690 9. Johnson, N.A., Coram, M.A., Shriver, M.D., Romieu, I., Barsh, G.S., London, S.J., and
691 Tang, H. (2011). Ancestral Components of Admixed Genomes in a Mexican Cohort.
692 *PLoS Genetics* 7, e1002410.
- 693 10. Vicuña, L., Fernandez, M.I., Vial, C., Valdebenito, P., Chaparro, E., Espinoza, K.,
694 Ziegler, A., Bustamante, A., and Eyheramendy, S. (2019). Adaptation to Extreme
695 Environments in an Admixed Human Population from the Atacama Desert. *Genome*
696 *Biology and Evolution* 11, 2468–2479.
- 697 11. Yelmen, B., Mondal, M., Marnetto, D., Pathak, A.K., Montinaro, F., Gallego Romero, I.,
698 Kivisild, T., Metspalu, M., and Pagani, L. (2019). Ancestry-Specific Analyses Reveal
699 Differential Demographic Histories and Opposite Selective Pressures in Modern South
700 Asian Populations. *Molecular Biology and Evolution* 36, 1628–1642.
- 701 12. Lohmueller, K.E., Bustamante, C.D., and Clark, A.G. (2011). Detecting Directional
702 Selection in the Presence of Recent Admixture in African-Americans. *Genetics* 187, 823–
703 835.
- 704 13. Reynolds, A.W., Mata-Míguez, J., Miró-Herrans, A., Briggs-Cloud, M., Sylestine, A.,
705 Barajas-Olmos, F., Garcia-Ortiz, H., Rzhetskaya, M., Orozco, L., Raff, J.A., et al. (2019).

- 706 Comparing signals of natural selection between three Indigenous North American
707 populations. *PNAS* *116*, 9312–9317.
- 708 14. Ávila-Arcos, M.C., McManus, K.F., Sandoval, K., Rodríguez-Rodríguez, J.E., Villa-Islas,
709 V., Martin, A.R., Luisi, P., Peñaloza-Espinosa, R.I., Eng, C., Huntsman, S., et al. (2020).
710 Population History and Gene Divergence in Native Mexicans Inferred from 76 Human
711 Exomes. *Molecular Biology and Evolution* *37*, 994–1006.
- 712 15. Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T.,
713 Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A., et al. (2013). Genetic
714 Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular*
715 *Biology and Evolution* *30*, 1877–1888.
- 716 16. Bhatia, G., Tandon, A., Patterson, N., Aldrich, M.C., Ambrosone, C.B., Amos, C.,
717 Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., et al. (2014). Genome-wide Scan of
718 29,141 African Americans Finds No Evidence of Directional Selection since Admixture.
719 *The American Journal of Human Genetics* *95*, 437–444.
- 720 17. Refoyo-Martínez, A., Fonseca, R.R. da, Halldórsdóttir, K., Árnason, E., Mailund, T., and
721 Racimo, F. (2019). Identifying loci under positive selection in complex population
722 histories. *Genome Res.* *29*, 1506–1520.
- 723 18. Slatkin, M. (1987). Gene flow and the geographic structure of natural populations.
724 *Science* *236*, 787–792.
- 725 19. Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for
726 archaic adaptive introgression in humans. *Nat Rev Genet* *16*, 359–371.
- 727 20. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub,
728 Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian Genetic Diversity
729 Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool.
730 *The American Journal of Human Genetics* *91*, 83–96.
- 731 21. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer,
732 M., Bustamante, C.D., and Ostrer, H. (2010). Genome-wide patterns of population
733 structure and admixture among Hispanic/Latino populations. *PNAS* *107*, 8954–8961.
- 734 22. Hodgson, J.A., Pickrell, J.K., Pearson, L.N., Quillen, E.E., Prista, A., Rocha, J., Soodyall,
735 H., Shriver, M.D., and Perry, G.H. (2014). Natural selection for the Duffy-null allele in
736 the recently admixed people of Madagascar. *Proceedings of the Royal Society B:*
737 *Biological Sciences* *281*, 20140930.

- 738 23. Breton, G., Schlebusch, C.M., Lombard, M., Sjödin, P., Soodyall, H., and Jakobsson, M.
739 (2014). Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern
740 African Khoe Pastoralists. *Current Biology* 24, 852–858.
- 741 24. Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D.B., Pritchard,
742 J.K., Beall, C.M., and Di Rienzo, A. (2014). Admixture facilitates genetic adaptations to
743 high altitude in Tibet. *Nature Communications* 5, 3281.
- 744 25. Macholdt, E., Lede, V., Barbieri, C., Mpoloka, S.W., Chen, H., Slatkin, M., Pakendorf,
745 B., and Stoneking, M. (2014). Tracing Pastoralist Migrations to Southern Africa with
746 Lactase Persistence Alleles. *Current Biology* 24, 875–879.
- 747 26. Rishishwar, L., Conley, A.B., Wigington, C.H., Wang, L., Valderrama-Aguirre, A., and
748 King Jordan, I. (2015). Ancestry, admixture and fitness in Colombian genomes. *Scientific*
749 *Reports* 5, 12376.
- 750 27. Zhou, Q., Zhao, L., and Guan, Y. (2016). Strong Selection at MHC in Mexicans since
751 Admixture. *PLoS Genetics* 12, e1005847.
- 752 28. Busby, G., Christ, R., Band, G., Leffler, E., Le, Q.S., Rockett, K., Kwiatkowski, D., and
753 Spencer, C. (2017). Inferring adaptive gene-flow in recent African history. *BioRxiv*
754 205252.
- 755 29. Laso-Jadart, R., Harmant, C., Quach, H., Zidane, N., Tyler-Smith, C., Mehdi, Q., Ayub,
756 Q., Quintana-Murci, L., and Patin, E. (2017). The Genetic Legacy of the Indian Ocean
757 Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan.
758 *The American Journal of Human Genetics* 101, 977–984.
- 759 30. Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-loth, V., Sanchez, J., Alva, O.,
760 Arachiche, A., Boland, A., Olaso, R., Deleuze, J.-F., et al. (2018). Strong selection during
761 the last millennium for African ancestry in the admixed population of Madagascar. *Nature*
762 *Communications* 9, 932.
- 763 31. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G.,
764 Perry, G.H., Barreiro, L.B., Froment, A., et al. (2017). Dispersals and genetic adaptation
765 of Bantu-speaking populations in Africa and North America. *Science* 356, 543–546.
- 766 32. Hamid, I., Korunes, K.L., Beleza, S., and Goldberg, A. (2021). Rapid adaptation to
767 malaria facilitated by admixture in the human population of Cabo Verde. *eLife* 10,
768 e63177.
- 769 33. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li,
770 S., Jongh, M.D., Singleton, A., Blum, M.G.B., et al. (2012). Genomic Variation in Seven

- 771 Khoe-San Groups Reveals Adaptation and Complex African History. *Science* 338, 374–
772 379.
- 773 34. Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., and Pereira, L. (2015).
774 Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and*
775 *Evolution* 7, 3484–3495.
- 776 35. Deng, L., Ruiz-Linares, A., Xu, S., and Wang, S. (2016). Ancestry variation and
777 footprints of natural selection along the genome in Latin American populations. *Scientific*
778 *Reports* 6, 21766.
- 779 36. Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., and Jin, L. (2012). Genome-wide
780 detection of natural selection in African Americans pre- and post-admixture. *Genome*
781 *Research* 22, 519–527.
- 782 37. Norris, E.T., Rishishwar, L., Chande, A.T., Conley, A.B., Ye, K., Valderrama-Aguirre,
783 A., and Jordan, I.K. (2020). Admixture-enabled selection for rapid adaptive evolution in
784 the Americas. *Genome Biology* 21, 29.
- 785 38. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G.,
786 and Risch, N.J. (2007). Recent Genetic Selection in the Ancestral Admixture of Puerto
787 Ricans. *The American Journal of Human Genetics* 81, 626–633.
- 788 39. Miller, L.H., Mason, S.J., Clyde, D.F., and McGinniss, M.H. (1976). The Resistance
789 Factor to *Plasmodium vivax* in Blacks. *New England Journal of Medicine* 295, 302–304.
- 790 40. Tournamille, C., Colin, Y., Cartron, J.P., and Le Van Kim, C. (1995). Disruption of a
791 GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–
792 negative individuals. *Nature Genetics* 10, 224–228.
- 793 41. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A.,
794 Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide
795 patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503.
- 796 42. Long, J.C. (1991). The Genetic Structure of Admixed Populations. *Genetics* 127, 417–
797 428.
- 798 43. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D.,
799 Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-Range LD Can Confound
800 Genome Scans in Admixed Populations. *The American Journal of Human Genetics* 83,
801 132–135.
- 802 44. Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Zaitlen, N., Eng, C.,
803 Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2013). Analysis of

- 804 Latino populations from GALA and MEC studies reveals genomic loci with biased local
805 ancestry estimation. *Bioinformatics* 29, 1407–1415.
- 806 45. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the
807 Wright–Fisher Model. *Molecular Biology and Evolution* 36, 632–637.
- 808 46. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R.,
809 Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference
810 for human genetic variation. *Nature* 526, 68–74.
- 811 47. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn,
812 C.M., Swertz, M., Wijmenga, C., van Ommen, G., et al. (2015). Genome-wide patterns
813 and properties of de novo mutations in humans. *Nature Genetics* 47, 822–826.
- 814 48. Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch,
815 N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., et al. (2011). Natural Selection
816 Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the
817 Human Genome. *PLoS Genetics* 7, e1002326.
- 818 49. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D.,
819 Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008).
820 Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome.
821 *PLoS Genetics* 4, e1000083.
- 822 50. Haller, B.C., and Messer, P.W. SLiM: An Evolutionary Simulation Framework. 660.
- 823 51. Choin, J., Mendoza-Revilla, J., Arauna, L.R., Cuadros-Espinoza, S., Cassar, O., Larena,
824 M., Ko, A.M.-S., Harmant, C., Laurent, R., Verdu, P., et al. (2021). Genomic insights into
825 population history and biological adaptation in Oceania. *Nature* 592, 583–589.
- 826 52. Bernstein, F. (1931). Die geographische Verteilung der Blutgruppen und ihre
827 anthropologische Bedeutung. In *Comitato Weak Representations of the Data. Italiano per*
828 *Lo Studio Dei Problemi Della Popolazione*, (Roma: Istituto Poligrafico dello Stato), pp.
829 227–243.
- 830 53. Charlesworth, B., and Charlesworth, D. (2010). *Elements of Evolutionary Genetics* (W.
831 H. Freeman).
- 832 54. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of
833 ancestry in unrelated individuals. *Genome Research* 19, 1655–1664.
- 834 55. Workman, P.L., Blumberg, B.S., and Cooper, A.J. (1963). Selection, Gene Migration and
835 Polymorphic Stability in a U. S. White and Negro Population. *The American Journal of*
836 *Human Genetics* 15, 429–437.
- 837 56. Reed, T.E. (1969). Caucasian genes in American Negroes. *Science* 165, 762–768.

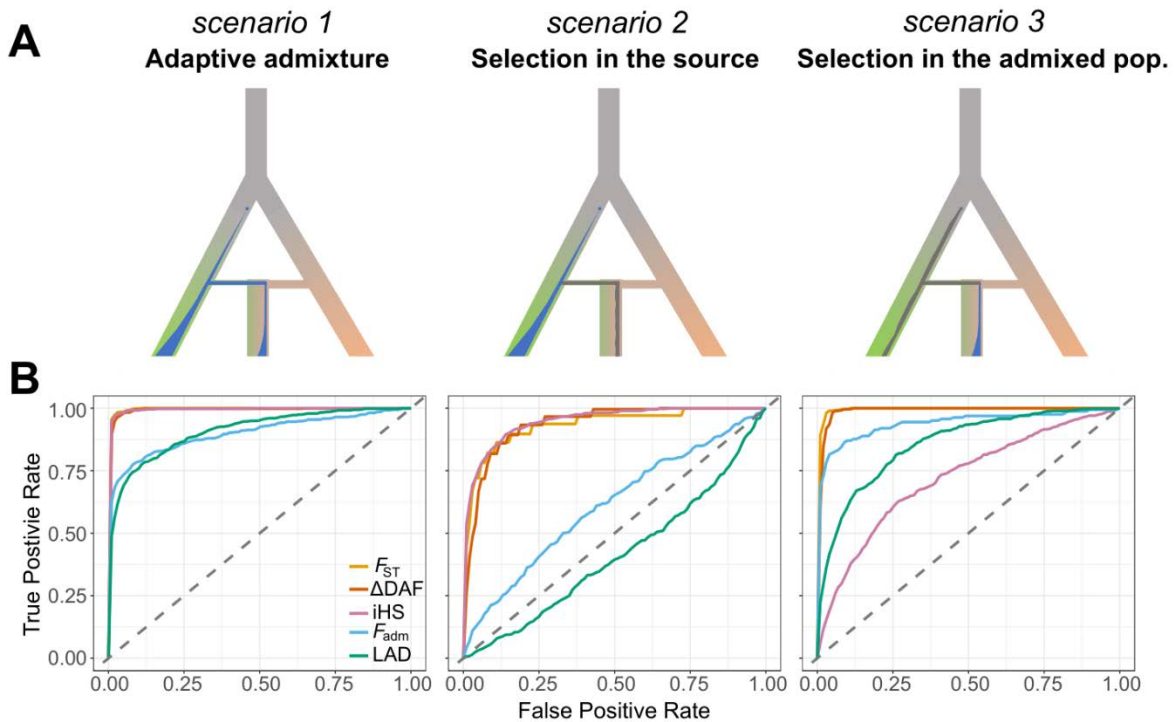
- 838 57. Cavalli-Sforza, L.L., and Bodmer, W.F. (1971). *The Genetics of Human Populations*
839 (Freeman & Co).
- 840 58. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A
841 Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The*
842 *American Journal of Human Genetics* 93, 278–288.
- 843 59. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T.
844 (2019). Accurate, scalable and integrative haplotype estimation. *Nat Commun* 10, 5436.
- 845 60. Stephan, W., Wiehe, T.H.E., and Lenz, M.W. (1992). The effect of strongly selected
846 substitutions on neutral polymorphism: Analytical results based on diffusion theory.
847 *Theoretical Population Biology* 41, 237–254.
- 848 61. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015).
849 Second-generation PLINK: rising to the challenge of larger and richer datasets.
850 *GigaScience* 4.
- 851 62. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M.
852 (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*
853 26, 2867–2873.
- 854 63. Petr, M., Vernot, B., and Kelso, J. (2019). admixr — R package for reproducible analyses
855 using ADMIXTOOLS. *Bioinformatics* 35, 3194–3195.
- 856 64. Fisher, R.A. (1925). *Statistical Methods for Research Workers* (Oliver and Boyd,
857 Edinburgh).
- 858 65. Ghossaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A.,
859 Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. (2021). Open Targets
860 Genetics: systematic identification of trait-associated genes using large-scale genetics and
861 functional genomics. *Nucleic Acids Research* 49, D1311–D1320.
- 862 66. Peter, B.M., Huerta-Sanchez, E., and Nielsen, R. (2012). Distinguishing between
863 Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS Genetics*
864 8, e1003011.
- 865 67. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J.,
866 Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., et al. (2016). The Great
867 Migration and African-American Genomic Diversity. *PLoS Genetics* 12, e1006059.
- 868 68. Fortes-Lima, C.A., Laurent, R., Thouzeau, V., Toupance, B., and Verdu, P. (2021).
869 Complex genetic admixture histories reconstructed with Approximate Bayesian
870 Computation. *Molecular Ecology Resources* 21, 1098–1117.

- 871 69. Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I.,
872 Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., et al. (2016). A genomic
873 history of Aboriginal Australia. *Nature* 538, 207–214.
- 874 70. Medina, P., Thornlow, B., Nielsen, R., and Corbett-Detig, R. (2018). Estimating the
875 Timing of Multiple Admixture Pulses During Local Ancestry Inference. *Genetics* 210,
876 1089–1107.
- 877 71. Pickrell, J.K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf,
878 B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa.
879 *PNAS* 111, 2632–2637.
- 880 72. Molinaro, L., Marnetto, D., Mondal, M., Ongaro, L., Yelmen, B., Lawson, D.J.,
881 Montinaro, F., and Pagani, L. (2021). A Chromosome-Painting-Based Pipeline to Infer
882 Local Ancestry under Limited Source Availability. *Genome Biology and Evolution* 13.
- 883 73. Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by
884 DNA Polymorphism. *Genetics* 123, 585–595.
- 885 74. Tajima, F. (1989). The Effect of Change in Population Size on DNA Polymorphism.
886 *Genetics* 123, 597–601.
- 887 75. Wakeley, J., and Aliacar, N. (2001). Gene Genealogies in a Metapopulation. *Genetics*
888 159, 893–905.
- 889 76. Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133,
890 693–709.
- 891 77. Przeworski, M. (2002). The Signature of Positive Selection at Randomly Chosen Loci.
892 *Genetics* 160, 1179–1189.
- 893 78. Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M.,
894 Cavalli-Sforza, L.L., Feldman, M.W., and Pritchard, J.K. (2009). The Role of Geography
895 in Human Adaptation. *PLoS Genetics* 5, e1000500.
- 896 79. Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On Detecting
897 Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology*
898 *and Evolution* 31, 1275–1291.
- 899 80. Dias-Alves, T., Mairal, J., and Blum, M.G.B. (2018). Loter: A Software Package to Infer
900 Local Ancestry for a Wide Range of Species. *Molecular Biology and Evolution* 35, 2318–
901 2326.
- 902 81. Meyer, D., C. Aguiar, V.R., Bitarello, B.D., C. Brandt, D.Y., and Nunes, K. (2018). A
903 genomic perspective on HLA evolution. *Immunogenetics* 70, 5–27.

- 904 82. Vicente, M., Priehodová, E., Diallo, I., Podgorná, E., Poloni, E.S., Černý, V., and
905 Schlebusch, C.M. (2019). Population history and genetic adaptation of the Fulani nomads:
906 inferences from genome-wide data and the lactase persistence trait. *BMC Genomics* 20,
907 915.
- 908 83. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I.,
909 Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Knob, A.L.U., et al. (2010). Association
910 of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science* 329,
911 841–845.
- 912 84. Ruby, M.A., Riedl, I., Massart, J., Åhlin, M., and Zierath, J.R. (2017). Protein kinase N2
913 regulates AMP kinase signaling and insulin responsiveness of glucose metabolism in
914 skeletal muscle. *American Journal of Physiology-Endocrinology and Metabolism* 313,
915 E483–E491.
- 916 85. Meng, J., Yao, Z., He, Y., Zhang, R., Zhang, Y., Yao, X., Yang, H., Chen, L., Zhang, Z.,
917 Zhang, H., et al. (2017). ARRDC4 regulates enterovirus 71-induced innate immune
918 response by promoting K63 polyubiquitination of MDA5 through TRIM65. *Cell Death*
919 *Disease* 8, e2866.
- 920 86. Wit, J.M., and Walenkamp, M.J. (2013). Role of insulin-like growth factors in growth,
921 development and feeding. *World Review of Nutrition and Dietetics* 106, 60–65.
- 922 87. Warrington, N.M., Beaumont, R.N., Horikoshi, M., Day, F.R., Helgeland, Ø., Laurin, C.,
923 Bacelis, J., Peng, S., Hao, K., Feenstra, B., et al. (2019). Maternal and fetal genetic effects
924 on birth weight and their relevance to cardio-metabolic risk factors. *Nature Genetics* 51,
925 804–814.
- 926 88. Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., van Helden, P.D., Mulder, N.J., Price,
927 A.L., and Hoal, E.G. (2014). Genome-wide association study of ancestry-specific TB risk
928 in the South African Coloured population. *Human Molecular Genetics* 23, 796–809.
- 929 89. Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M.,
930 Dannemann, M., Grote, S., McCoy, R.C., Norton, H., et al. (2016). Excavating Neandertal
931 and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352, 235–239.
- 932 90. Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akey,
933 J.M. (2016). Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa
934 Environments. *Current Biology* 26, 3375–3382.
- 935 91. Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J.,
936 Mondal, M., Pagani, L., Ricaut, F.-X., Stoneking, M., et al. (2019). Multiple Deeply
937 Divergent Denisovan Ancestries in Papuans. *Cell* 177, 1010-1021.e32.

- 938 92. Zammit, N.W., Siggs, O.M., Gray, P.E., Horikawa, K., Langley, D.B., Walters, S.N.,
939 Daley, S.R., Loetsch, C., Warren, J., Yap, J.Y., et al. (2019). Denisovan, modern human
940 and mouse TNFAIP3 alleles tune A20 phosphorylation and immunity. *Nature*
941 *Immunology* 20, 1299–1310.
- 942 93. Toyama, K.S., Crochet, P.-A., and Leblois, R. (2020). Sampling schemes and drift can
943 bias admixture proportions inferred by structure. *Molecular Ecology Resources* 20, 1769–
944 1785.
- 945 94. Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C.A., Froment, A.,
946 Nyambo, T.B., Omar, S.A., Wambebe, C., et al. (2013). Identifying Darwinian Selection
947 Acting on Different Human APOL1 Variants among Diverse African Populations. *The*
948 *American Journal of Human Genetics* 93, 54–66.
- 949 95. Limou, S., Nelson, G.W., Lecordier, L., An, P., O’hUigin, C.S., David, V.A., Binns-
950 Roemer, E.A., Guiblet, W.M., Oleksyk, T.K., Pays, E., et al. (2015). Sequencing rare and
951 common APOL1 coding variants to determine kidney disease risk. *Kidney International*
952 88, 754–763.
- 953 96. Franco, J.R., Simarro, P.P., Diarra, A., and Jannin, J.G. (2014). Epidemiology of human
954 African trypanosomiasis. *Clinical Epidemiology* 6, 257–275.
- 955 97. Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the Signature of Natural Selection
956 in Humans: Evidence from the Duffy Blood Group Locus. *The American Journal of*
957 *Human Genetics* 66, 1669–1679.
- 958 98. Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. (2002). Complex Signatures of
959 Natural Selection at the Duffy Blood Group Locus. *The American Journal of Human*
960 *Genetics* 70, 369–383.
- 961 99. McManus, K.F., Taravella, A.M., Henn, B.M., Bustamante, C.D., Sikora, M., and
962 Cornejo, O.E. (2017). Population genetic analysis of the *DARC* locus (Duffy) reveals
963 adaptation from standing variation associated with malaria resistance in humans. *PLoS*
964 *Genetics* 13, e1006560.
- 965 100. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S.,
966 Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent
967 adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* 39, 31–40.
- 968 101. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A.,
969 Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic Signatures of Strong
970 Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*
971 74, 1111–1120.

- 972 102. Livingstone, F.B. (1984). The Duffy blood groups, *vivax* malaria, and malaria selection
973 in human populations: a review. *Human Biology* 56, 413–425.
- 974 103. Ménard, D., Barnadas, C., Bouchier, C., Henry-Halldin, C., Gray, L.R., Ratsimbaoa, A.,
975 Thonier, V., Carod, J.-F., Domarle, O., Colin, Y., et al. (2010). *Plasmodium vivax* clinical
976 malaria is commonly observed in Duffy-negative Malagasy people. *PNAS* 107, 5967-
977 5971.
- 978 104. Popovici, J., Roesch, C., and Rougeron, V. (2020). The enigmatic mechanisms by which
979 *Plasmodium vivax* infects Duffy-negative individuals. *PLoS Pathogens* 16, e1008258.
- 980 105. Battle, K.E., Lucas, T.C.D., Nguyen, M., Howes, R.E., Nandi, A.K., Twohig, K.A.,
981 Pfeiffer, D.A., Cameron, E., Rao, P.C., Casey, D., et al. (2019). Mapping the global
982 endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal
983 modelling study. *The Lancet* 394, 332–343.
- 984 106. Guan, Y. (2014). Detecting Structure of Haplotypes and Local Ancestry. *Genetics* 196,
985 625–642.
- 986 107. Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M., and Ramachandran,
987 S. (2018). Localization of adaptive variants in human genomes using averaged one-
988 dependence estimation. *Nature Communications* 9, 703.
- 989 108. Sella, G., and Barton, N.H. (2019). Thinking About the Evolution of Complex Traits in
990 the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human*
991 *Genetics* 20, 461–493.
- 992 109. Racimo, F., Berg, J.J., and Pickrell, J.K. (2018). Detecting Polygenic Adaptation in
993 Admixture Graphs. *Genetics* 208, 1565–1584.
- 994 110. Dehasque, M., Ávila-Arcos, M.C., Díez-del-Molino, D., Fumagalli, M., Guschanski, K.,
995 Lorenzen, E.D., Malaspinas, A.-S., Marques-Bonet, T., Martin, M.D., Murray, G.G.R., et
996 al. (2020). Inference of natural selection from ancient DNA. *Evolution Letters* 4, 94–108.
997



1000

1001

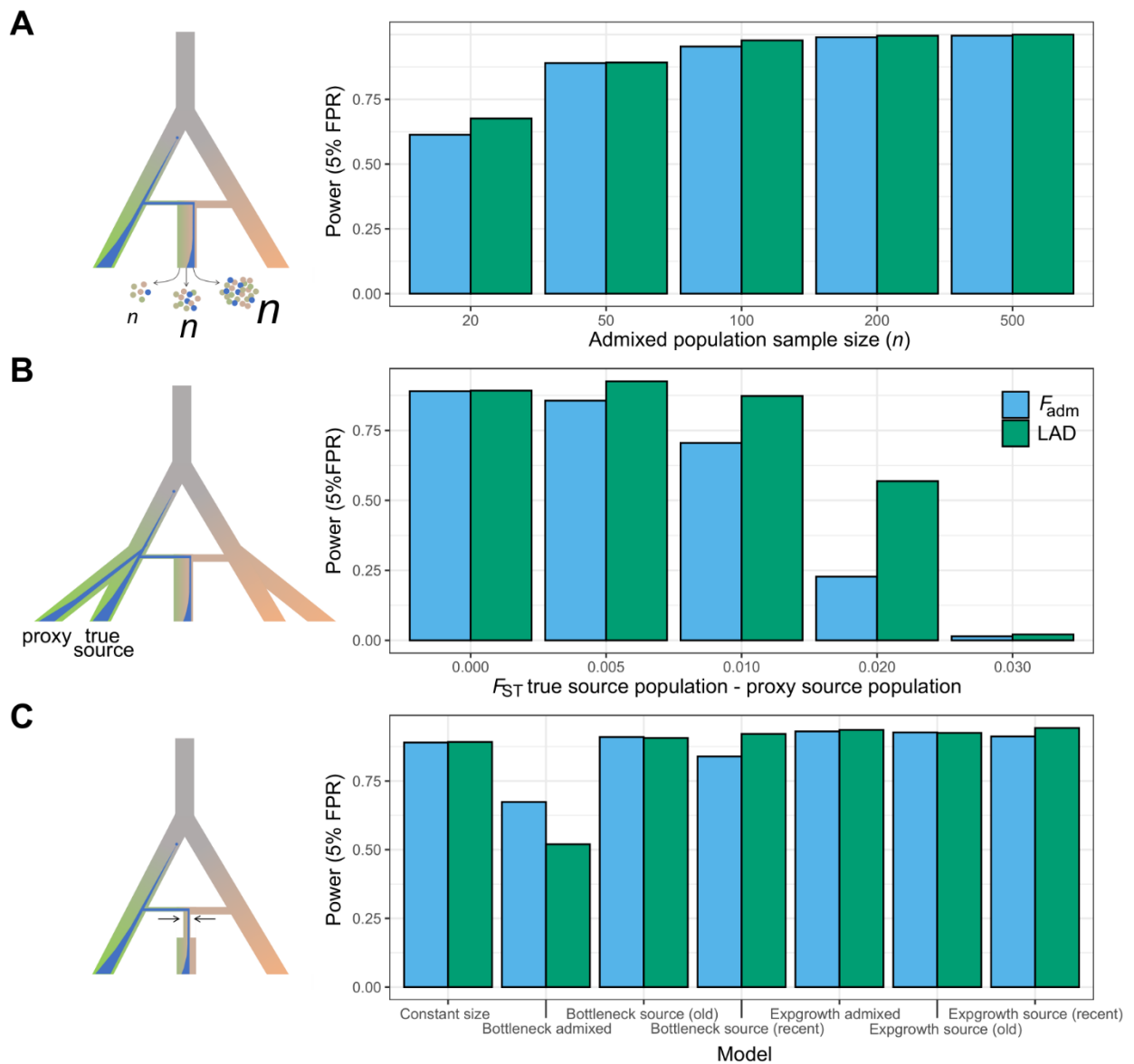
1002 **Figure 1. Performance of neutrality statistics under different scenarios of admixture**
 1003 **with selection.**

1004 (A) Explored scenarios of admixture with selection (from left to right): adaptive admixture,
 1005 positive selection in the source population only and positive selection in the admixed
 1006 population only. The blue and gray points indicate the appearance of a new beneficial and
 1007 neutral mutations, respectively. The blue and gray areas indicate changes in frequency of the
 1008 beneficial and neutral mutation, respectively.

1009 (B) Receiver operating characteristic (ROC) curves comparing the performance of classic
 1010 neutrality statistics F_{ST} , *iHS*, ΔDAF and the admixture-specific statistics F_{adm} and LAD,
 1011 across the 3 explored scenarios. The selection coefficient was fixed to $s = 0.05$, to highlight
 1012 the differences between statistics and between models (see Figure S2 for lower s values).

1013 False positive rate (FPR) is the fraction of simulated neutral sites that are incorrectly detected
 1014 as adaptive, and true positive rate (TPR) is the fraction of simulated adaptive mutations that
 1015 are correctly detected as under selection.

1016



1017

1018

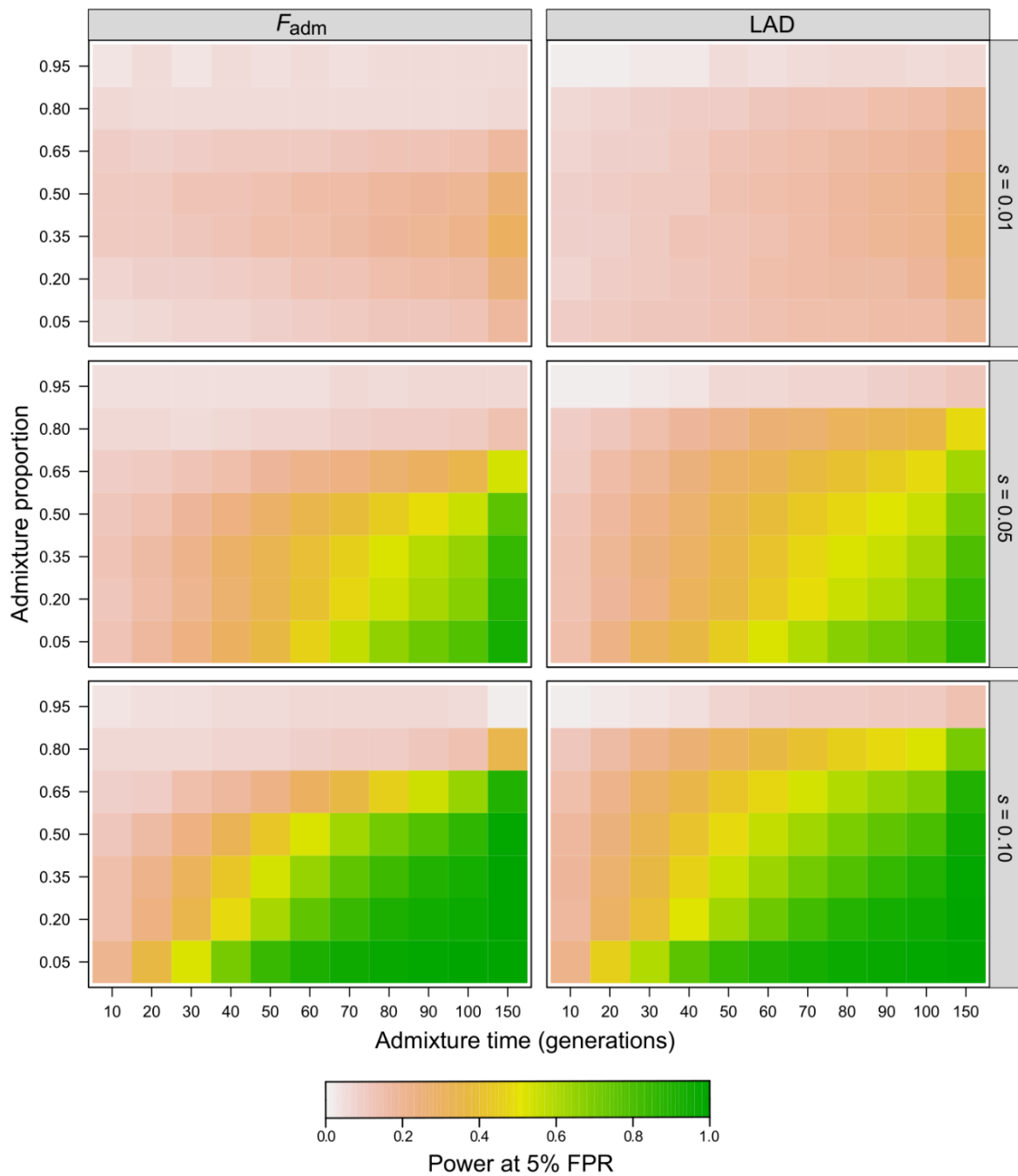
1019 **Figure 2. Effects of study design on the power to detect adaptive admixture.**

1020 (A) Effects of the sample size of the admixed population (n) on the detection power of F_{adm}
 1021 and LAD, at a fixed FPR = 5%. The simulated model is shown on the left, including different
 1022 values of n .

1023 (B) Effects of the use of proxy source populations on the detection power of F_{adm} and LAD, at
 1024 a fixed FPR = 5%. The genetic distance between the true source and its proxy was measured
 1025 by F_{ST} . The simulated model is shown on the left, including the true source and its proxy.

1026 (C) Effects of non-stationary demography on the detection power of F_{adm} and LAD, at a fixed
 1027 FPR = 5%. The simulated model is shown on the left, including a bottleneck in the admixed
 1028 population.

1029



1030

1031

1032 **Figure 3. Effects of model parameters on the power to detect adaptive admixture.**

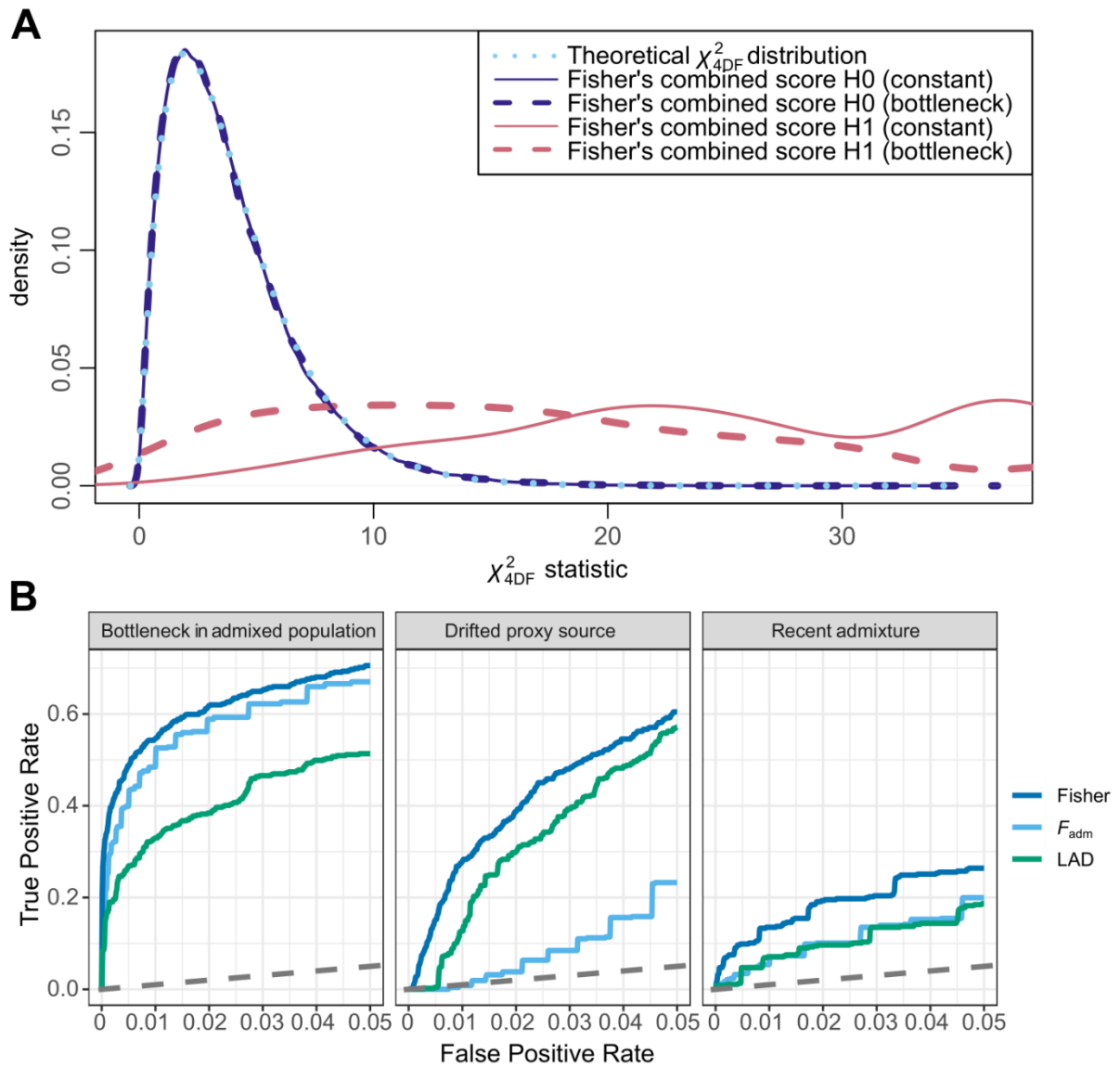
1033 Color represents average detection power for a fixed FPR = 5% across parameter

1034 combinations. The effects of other parameters, such as population sizes and divergence time

1035 (Figure S1 and Table S1), are shown in Figures S7-S11.

1036

1037



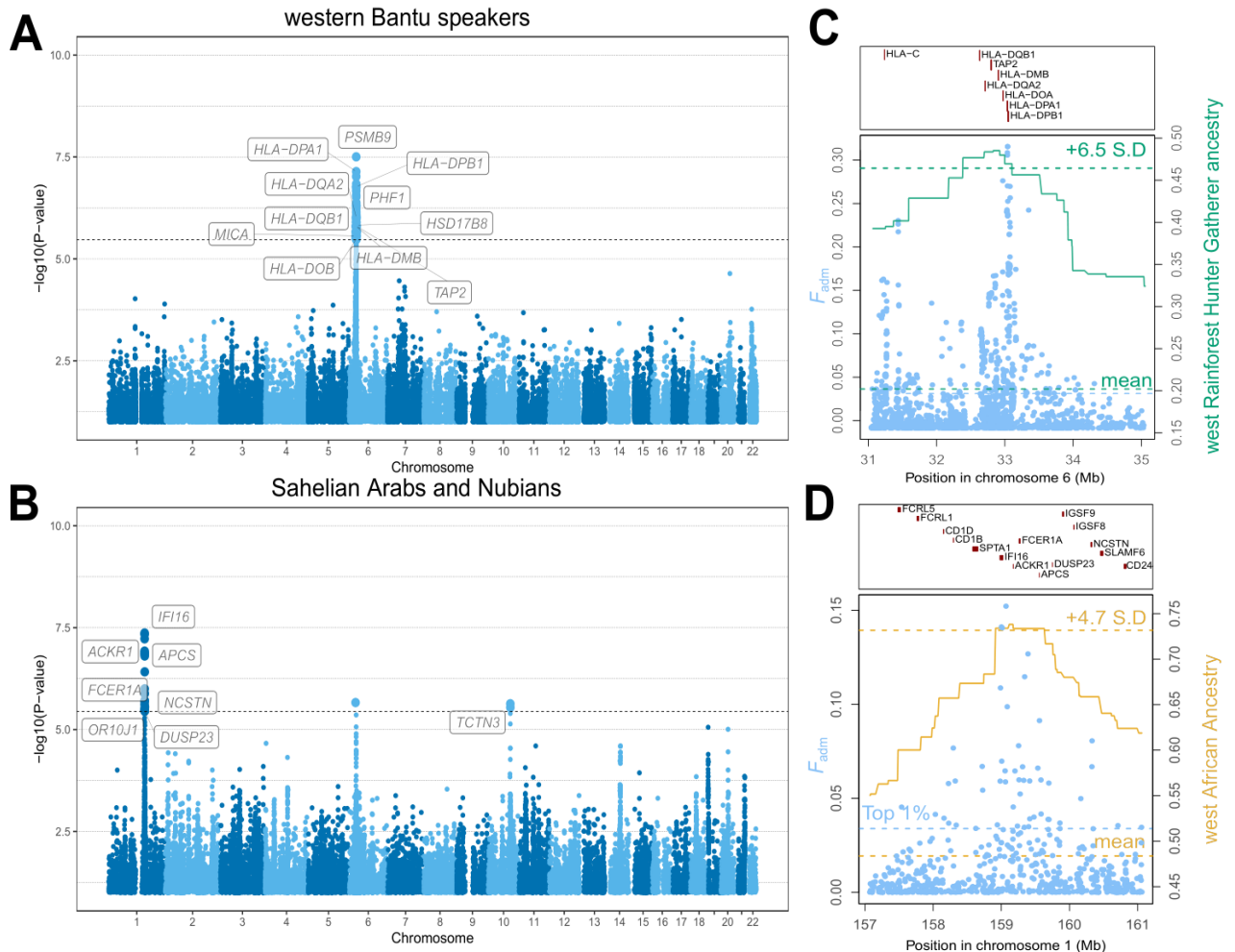
1038

1039

1040 **Figure 4. Performance of the Fisher's method to detect adaptive admixture.**

1041 (A) Distributions of the combined Fisher's score under the null hypothesis of no positive
 1042 selection (H0, blue lines) and under adaptive admixture (H1, pink lines), compared to the
 1043 theoretical χ^2 distribution with 4 degrees of freedom (dotted light blue line). Solid and dashed
 1044 lines indicate distributions under a constant population size and a 10-fold bottleneck in the
 1045 admixed population.

1046 (B) ROC curves for F_{adm} , LAD and the combined Fisher's score under unfavourable scenarios
 1047 for detecting adaptive admixture: a 10-fold bottleneck introduced in the admixed population,
 1048 the use of a proxy source population having experienced strong drift (F_{ST} between the true
 1049 source and proxy populations of 0.02) and recent admixture ($T_{adm} = 10$ generations). Only FPR
 1050 $< 5\%$ are shown.



1051
 1052
 1053
 1054
 1055
 1056
 1057
 1058
 1059
 1060
 1061
 1062
 1063
 1064
 1065
 1066

Figure 5. Iconic genomic signals of adaptive admixture.

(A) Genome-wide signals of adaptive admixture in Bantu-speaking populations from Gabon.

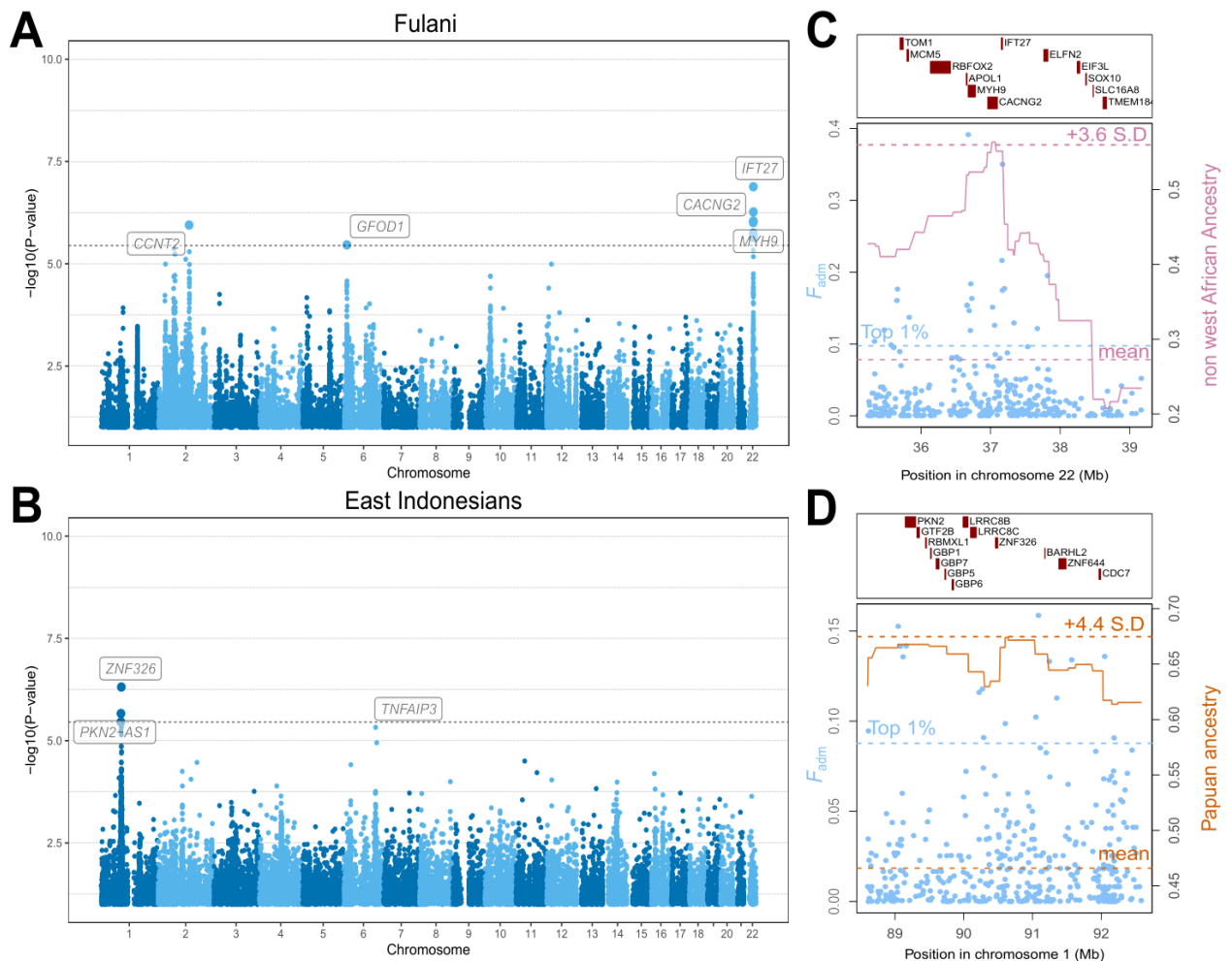
(B) Genome-wide signals of adaptive admixture in Sahelian Arabs and Nubians.

(A-B) Highlighted blue points indicate variants that passed the Bonferroni significance threshold (shown by a horizontal dotted line). Gene labels were attributed based on the gene with the highest V2G score within 250-kb of the candidate variant.

(C) Local signatures of adaptive admixture for the *HLA* region in Bantu-speaking populations from Gabon.

(D) Local signatures of adaptive admixture for the *ACKR1* region in Sahelian Arabs and Nubians.

(C-D) Light blue points indicate F_{adm} values for individual variants. The green and gold solid lines indicate average local ancestry from African rainforest hunter-gatherers and West Africans respectively.



1067

1068

1069 **Figure 6. Newly discovered genomic signals of adaptive admixture.**

1070 (A) Genome-wide signals of adaptive admixture in the Fulani nomads of West Africa.

1071 (B) Genome-wide signals of adaptive admixture in East Indonesians.

1072 (A-B) Highlighted blue points indicate variants that pass the Bonferroni significance

1073 threshold (shown by a horizontal dotted line). Gene labels were attributed based on the gene

1074 with the highest V2G score within 250-kb of the candidate variant.

1075 (C) Local signatures of adaptive admixture for the *IFT27/MYH9/APOL1* region in the Fulani

1076 nomads.

1077 (D) Local signatures of adaptive admixture for the *PKN2/LRR8CB* region in East Indonesians.

1078 (C-D) Light blue points indicate F_{adm} values for individual variants. The pink and orange

1079 solid lines indicate the local ancestry from Europeans and North Africans, and Papuans,

1080 respectively.

Supplementary Figures

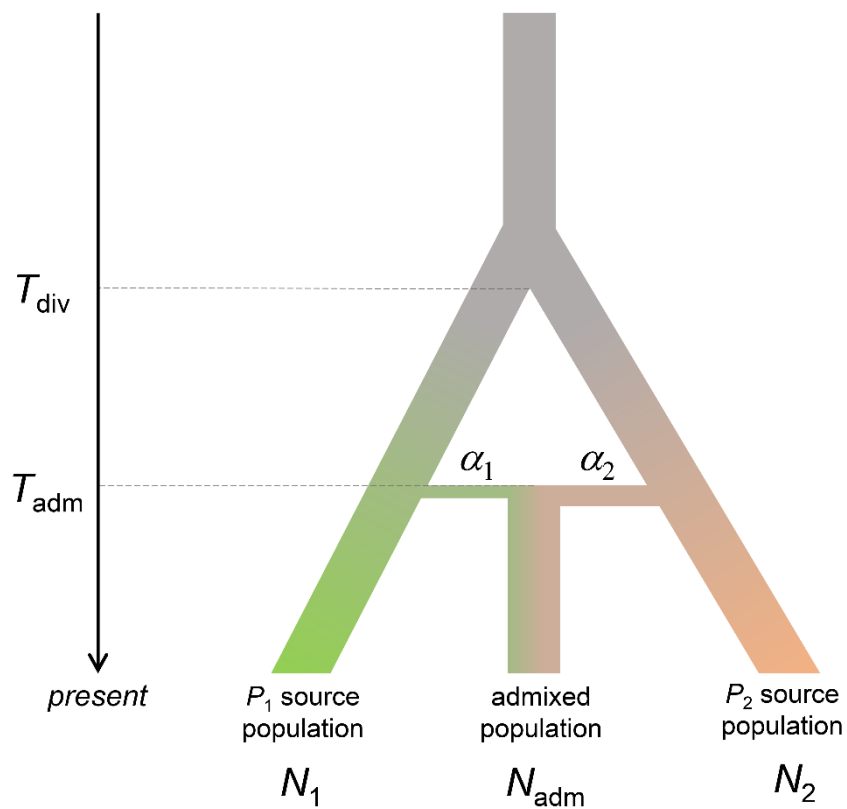


Figure S1. The simulated single-pulse admixture model.

The admixed population originates from admixture between two source populations, referred to as P_1 and P_2 . P_1 and P_2 contribute α_1 and α_2 admixture proportions to the admixed population, with $\alpha_1 + \alpha_2 = 1$. P_1 and P_2 diverge T_{div} generations ago and the admixture event occurs T_{adm} generations ago. The population sizes of the admixed population and of P_1 and P_2 source populations are N_{adm} , N_1 and N_2 , respectively.

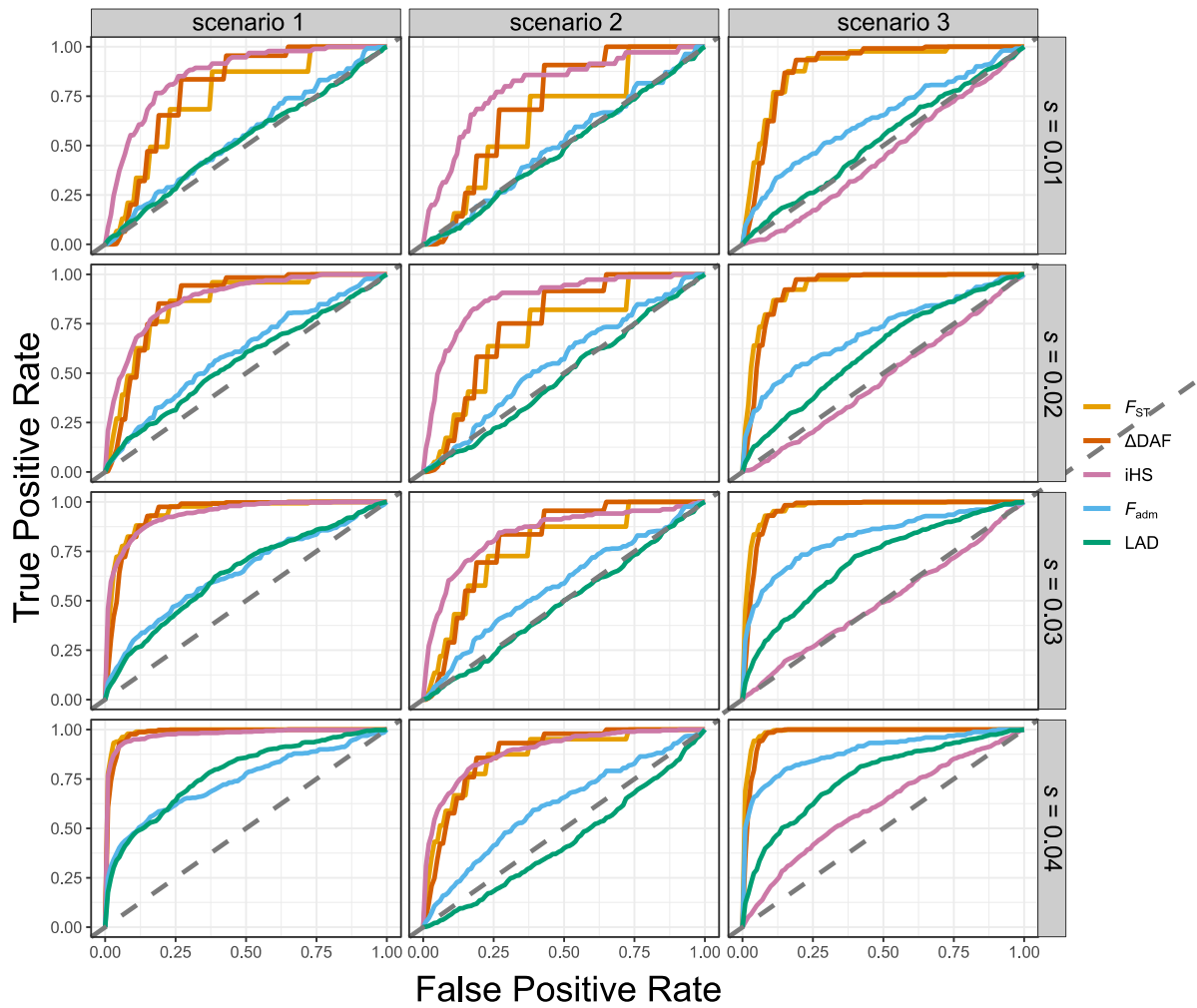


Figure S2. Performance of neutrality statistics under different scenarios of admixture with selection, assuming different selection coefficients.

Receiver operating characteristic (ROC) curves comparing the performance of the classic neutrality statistics F_{ST} , ΔDAF and iHS and the admixture-specific statistics F_{adm} and LAD, across the 3 explored admixture with selection scenarios, with varying selection coefficients $s \in \{0.01, 0.02, 0.03, 0.04\}$.

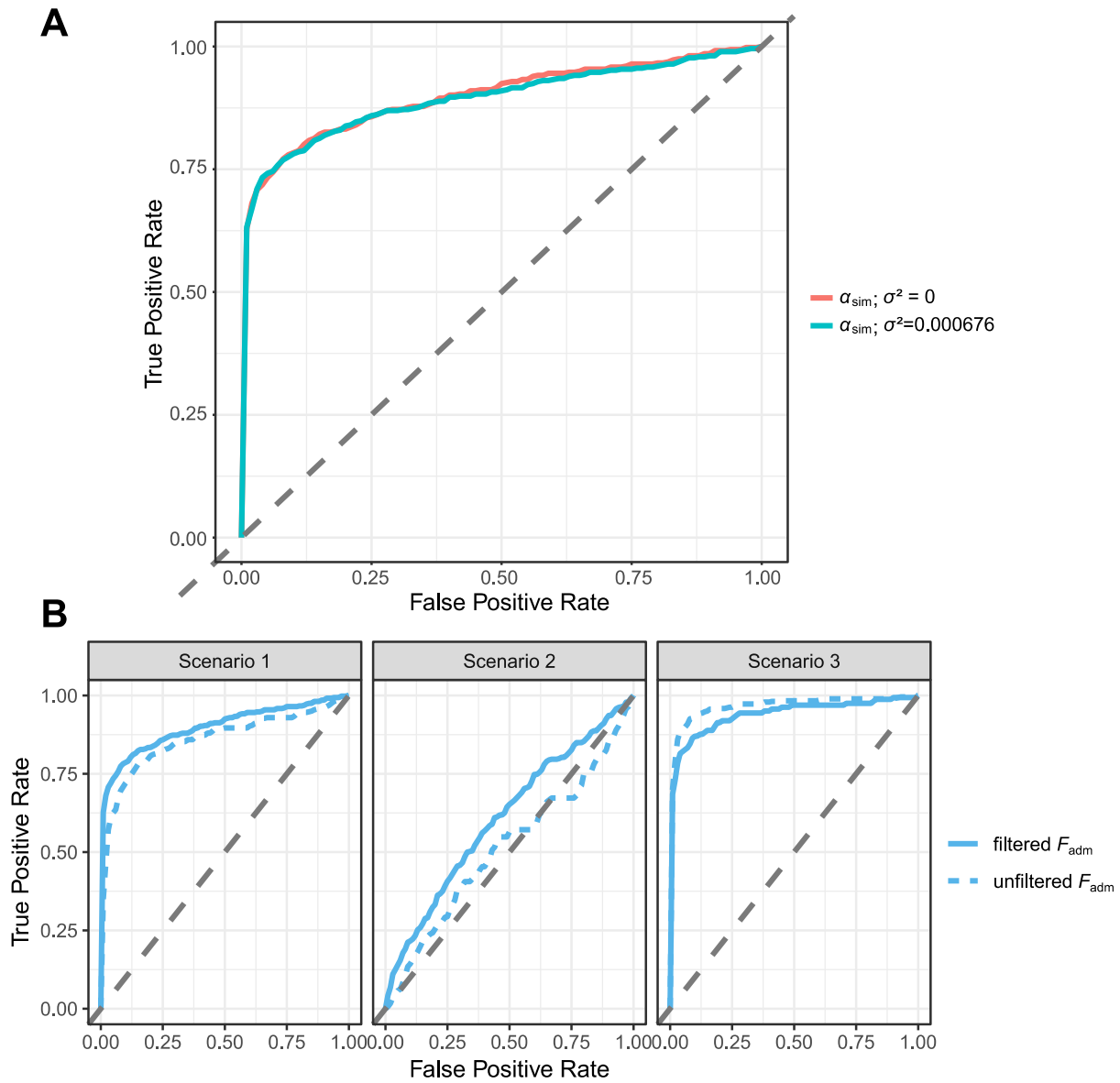


Figure S3. Performance of F_{adm} when using simulated admixture proportions with error and when applying or not an allele frequency filter.

(A) Receiver operating characteristic (ROC) curves comparing the performance of F_{adm} when using the simulated admixture proportions α_{sim} or α sampled from a normal distribution $\mathcal{N}(\mu = \alpha_{\text{sim}}, \sigma^2 = 0.026^2 = 0.000676)$, 0.026 being the highest root-mean-square deviation of the ADMIXTURE estimation.⁵⁴

(B) Receiver operating characteristic (ROC) curves comparing the performance of F_{adm} , with and without applying an allele frequency filter based on the source populations (see Material & Methods), under the 3 explored admixture with selection scenarios.

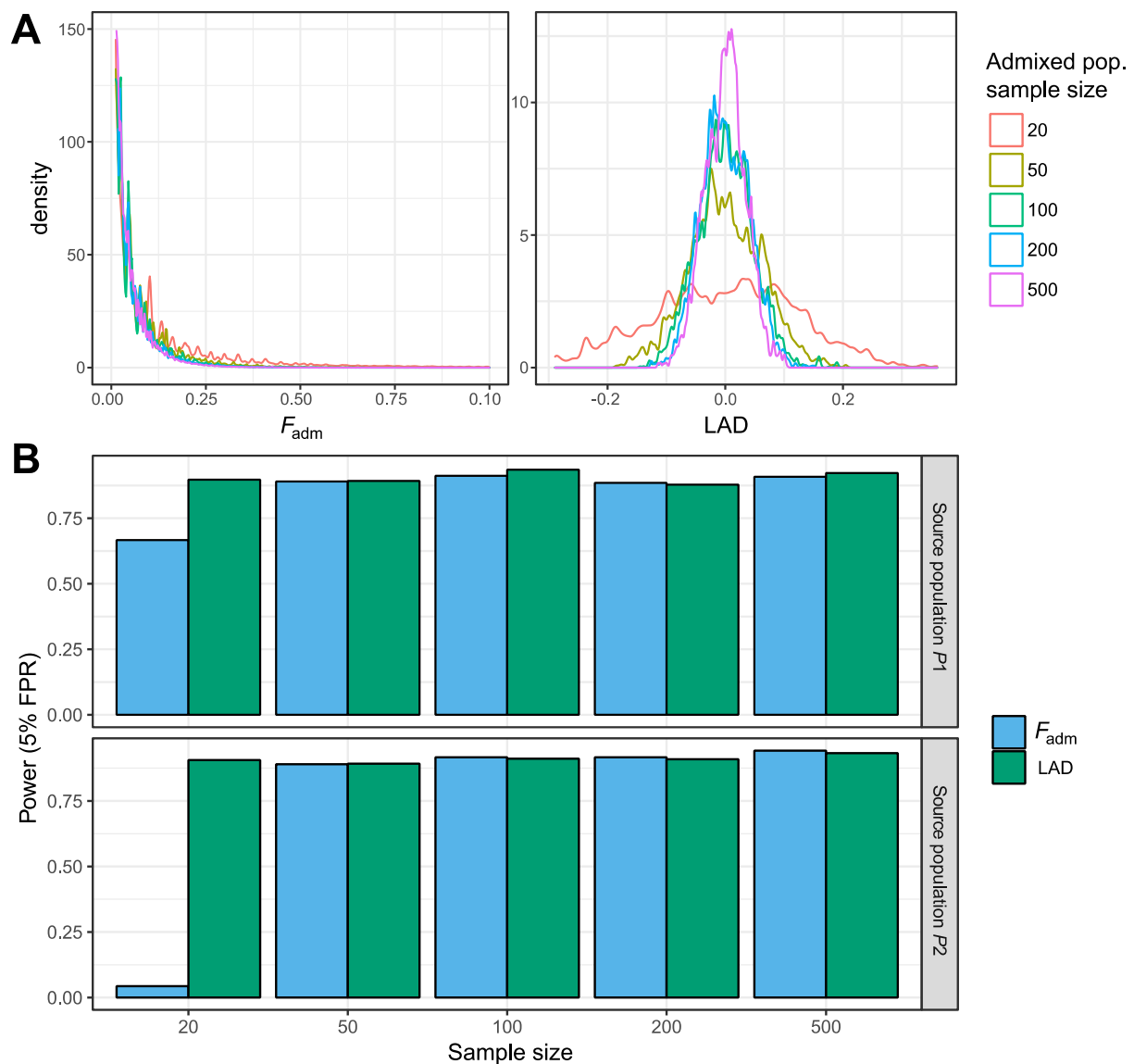


Figure S4. Effects of sample size on the power of F_{adm} and LAD statistics.

(A) Distributions under the null hypothesis (no positive selection) of F_{adm} and LAD, with varying sample sizes for the admixed population.

(B) Effect of the sample size of the source populations on the detection power of F_{adm} and LAD.

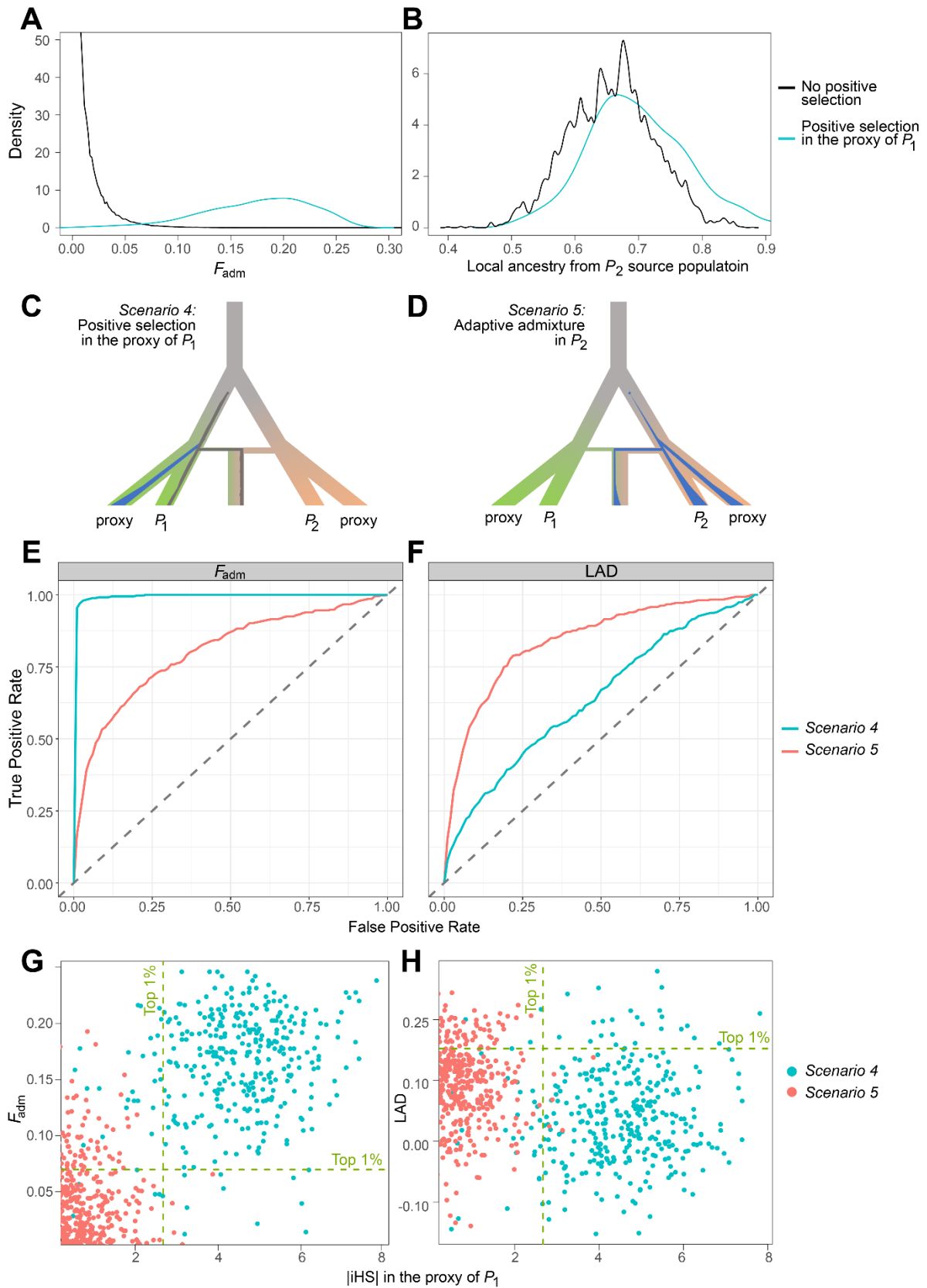


Figure S5. False positive signals due to selection in the proxy source population.

(A) Distributions of F_{adm} when there is or not positive selection in the proxy of the P_1 source population.

(B) Distributions of local ancestry in the admixed population from the P_2 source population, when there is or not positive selection in the proxy source population.

(C) The simulated model, assuming positive selection only in the proxy of the P_1 source population.

(D) The simulated model, assuming adaptive admixture in the P_2 source population. The scenario was simulated for comparison purposes.

(C-D) The blue and gray points indicate the appearance of a new beneficial and neutral mutations, respectively. The blue and gray areas indicate changes in frequency of the beneficial and neutral mutation, respectively.

(E-F) ROC curves for (E) F_{adm} and (F) LAD comparing the scenario where there is positive selection in the proxy of P_1 only (*scenario 4*; Figure S5C) and the scenario where there is a adaptive admixture in P_2 (*scenario 5*; Figure S5D).

(G–H) Absolute iHS values for the selected mutation in the proxy of the P_1 source population vs. (G) F_{adm} and (H) LAD values in the admixed population, when there is selection in this proxy of P_1 only (*scenario 4*; Figure S5C), or when there is adaptive admixture in P_2 (*scenario 5*; Figure S5D). Dashed green lines represent the 99th percentiles (based on the null model simulations) for absolute iHS (vertical) and F_{adm} or LAD (horizontal). Excluding values that are above the absolute iHS 99th percentile excludes approximately 90% of the extreme F_{adm} and LAD values under selection in this proxy of P_1 only (*scenario 4*) but, importantly, does not exclude any extreme value generated under the true adaptive admixture scenario (*scenario 5*).

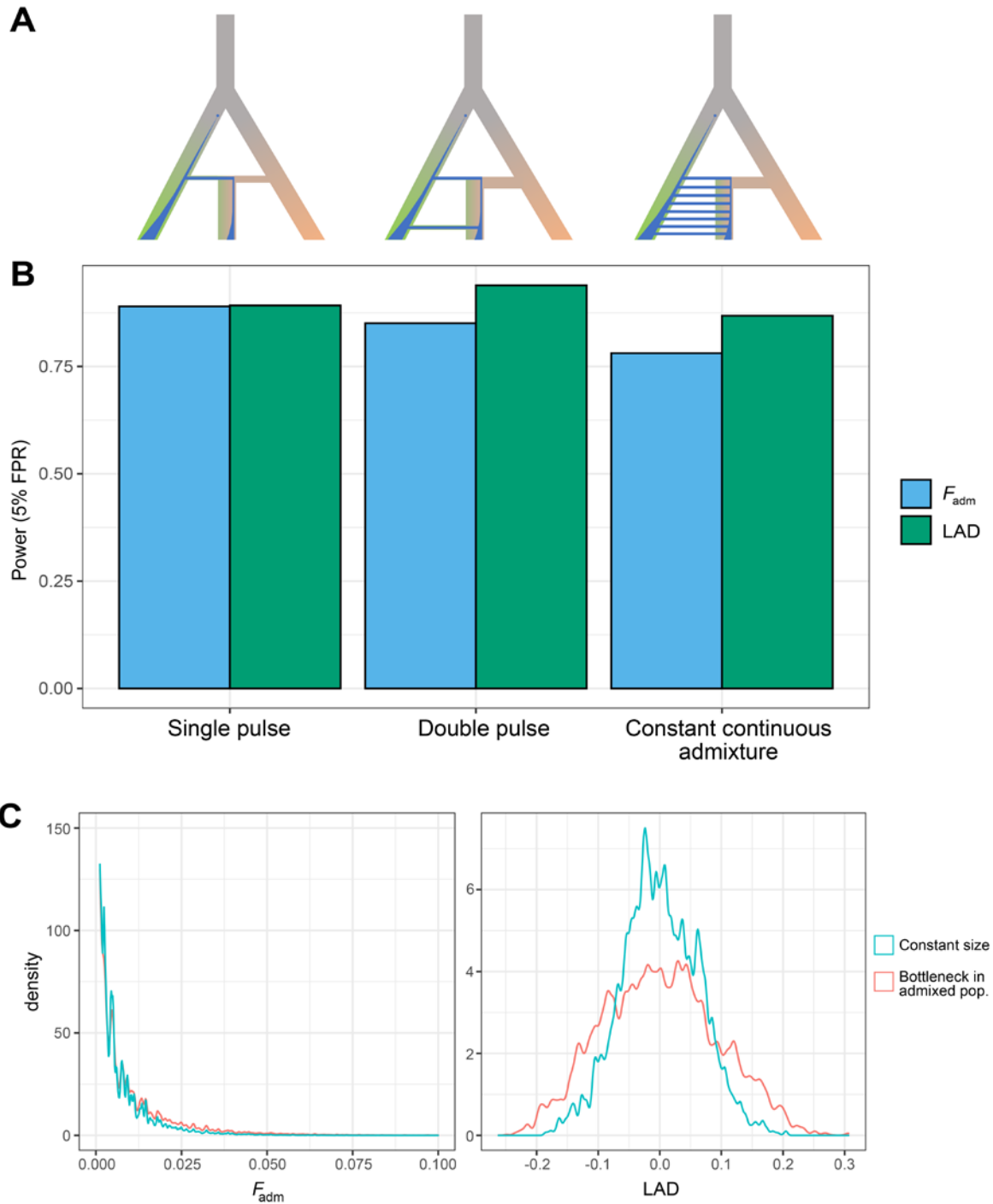


Figure S6. Effects of complex admixture and non-stationary demography on the power to detect adaptive admixture.

(A) The different simulated admixture models: a single pulse admixture model, a double pulse admixture model and a constant continuous admixture model. For these scenarios to be comparable, we set the sum of the admixture proportions contributed by each pulse to be equal to $\alpha_1 = 35\%$, and the average of the admixture dates to be equal to 70 generations (Material and Methods).

(B) Detection power of F_{adm} and LAD under the three different admixture scenarios (FPR = 5%; Material & Methods).

(C) Distributions of F_{adm} and LAD under the null hypothesis (no positive selection), with or without a 10-fold bottleneck in the admixed population.

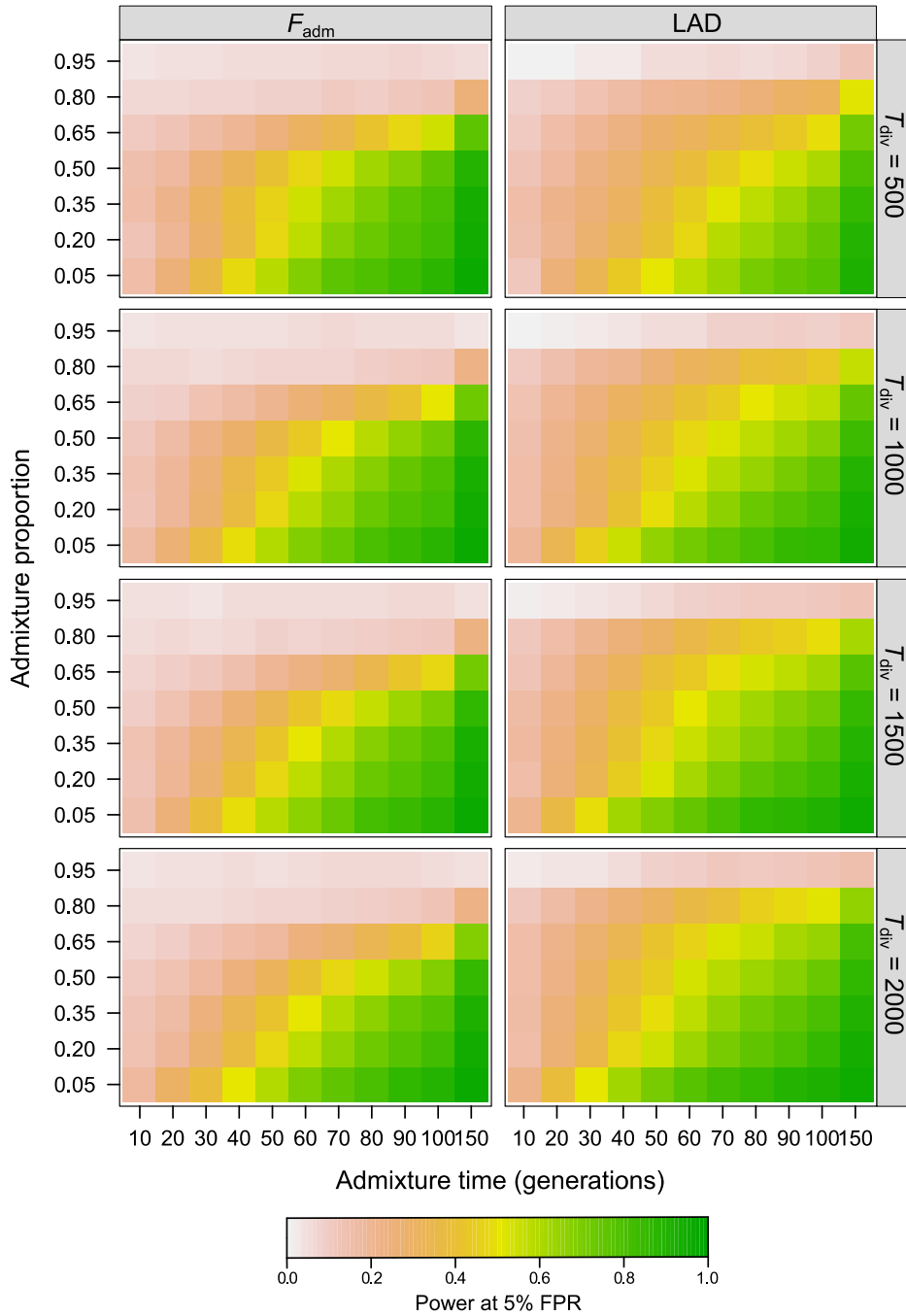


Figure S7. Effects of the divergence time between source populations on the power to detect adaptive admixture.

Effects on the detection power of F_{adm} and LAD of admixture time T_{adm} , admixture proportion α and the divergence time between source populations T_{div} . Colour indicates average detection power for a FPR = 5% threshold, across combinations of the remaining parameters. Because T_{div} is the upper limit of the time at which the beneficial mutation appears T_{mut} , we assumed for these simulations $T_{mut} < 500$ generations and $s \in \{0.05; 0.10\}$.

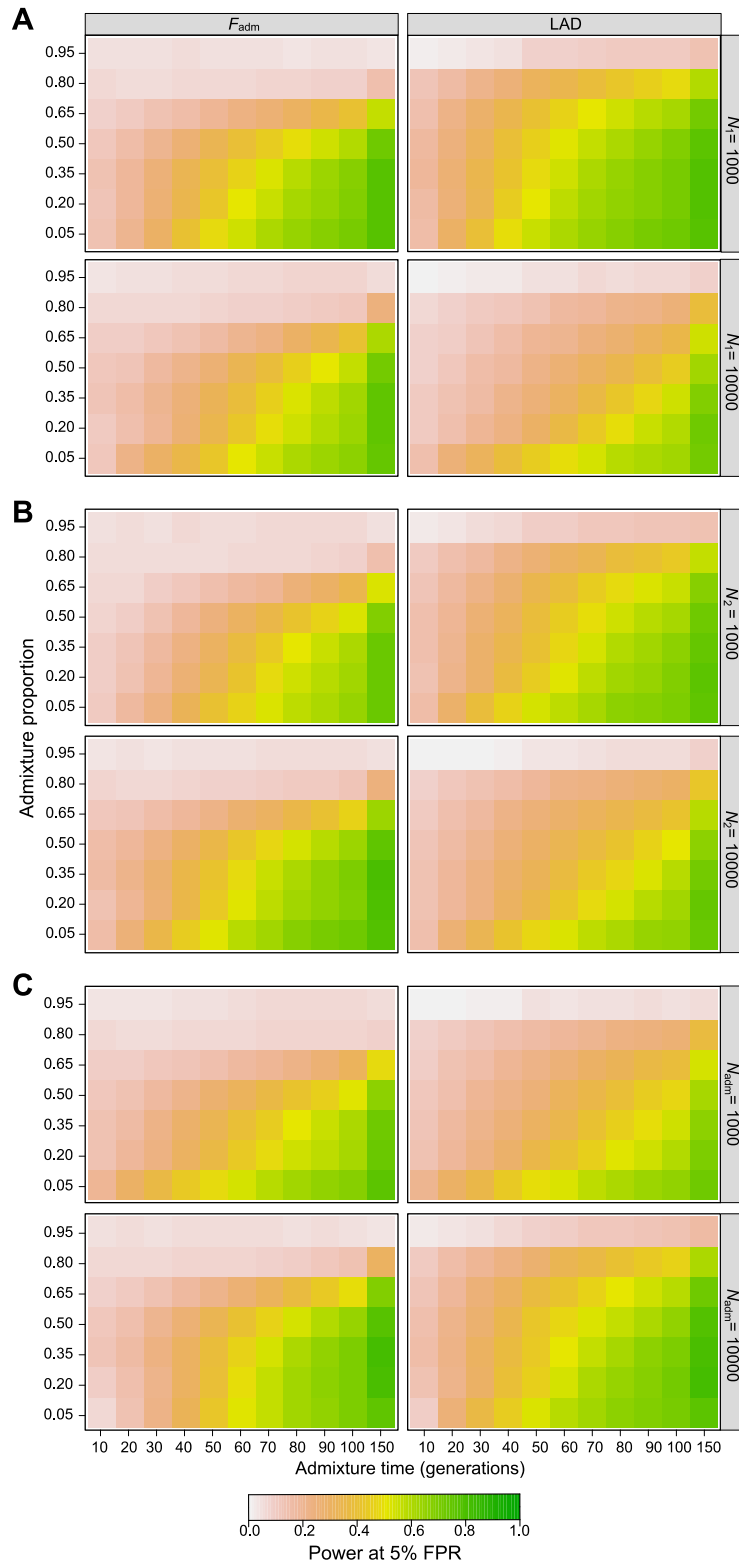


Figure S8. Effects of population sizes on the power to detect adaptive admixture.

Effects on the detection power of F_{adm} and LAD of admixture time T_{adm} , admixture proportion α and (A) N_1 , (B) N_2 and (C) N_{adm} , the population sizes of source population P_1 , source population P_2 and the admixed population, respectively (Figure S1). Colour indicates average detection power for a FPR = 5% threshold, across combinations of the remaining parameters.

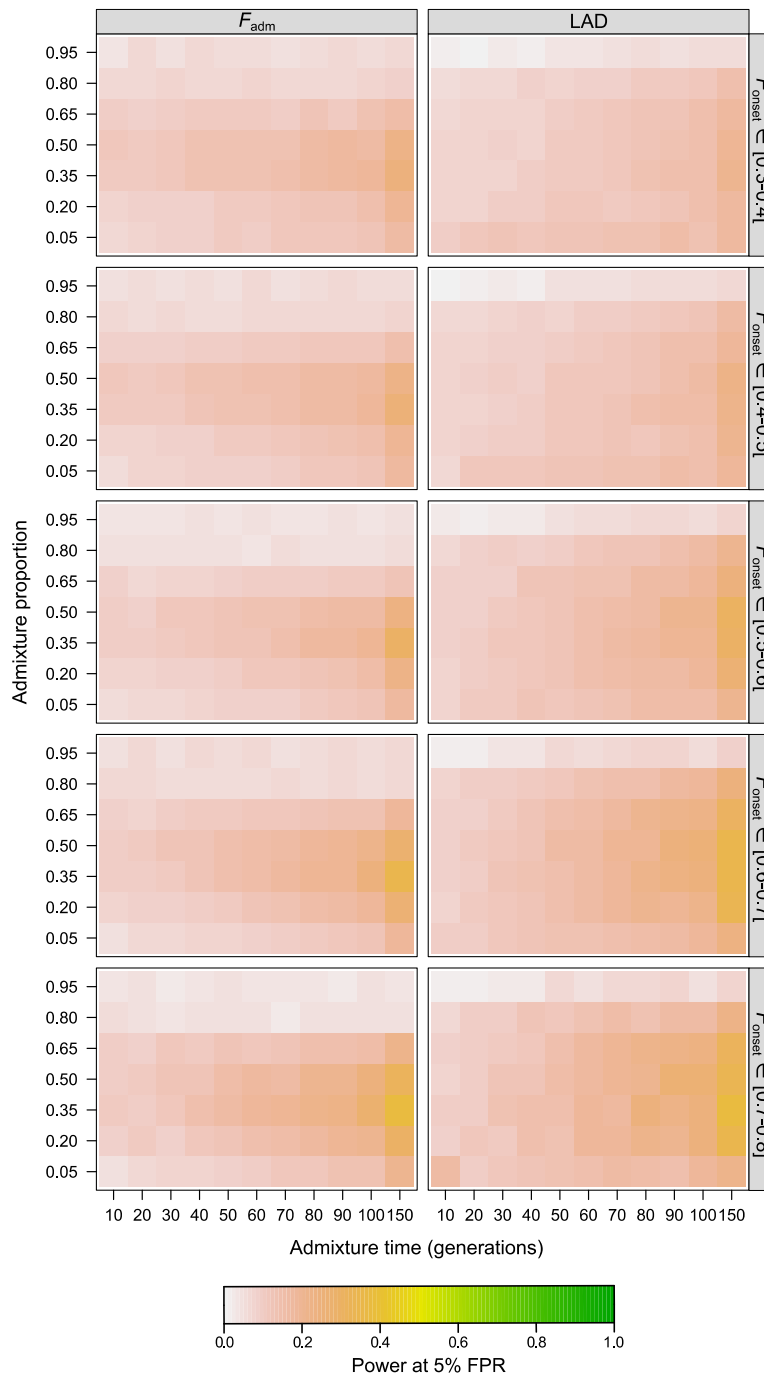


Figure S9. Effects of the frequency of the beneficial mutation ($s = 0.01$) on the power to detect adaptive admixture.

Effects on the detection power of F_{adm} and LAD of admixture time T_{adm} , admixture proportion α and F_{onset} , the frequency of the beneficial mutation in the source population at the time of admixture T_{adm} . Colour indicates average detection power for a FPR = 5% threshold, across combinations of the remaining parameters.

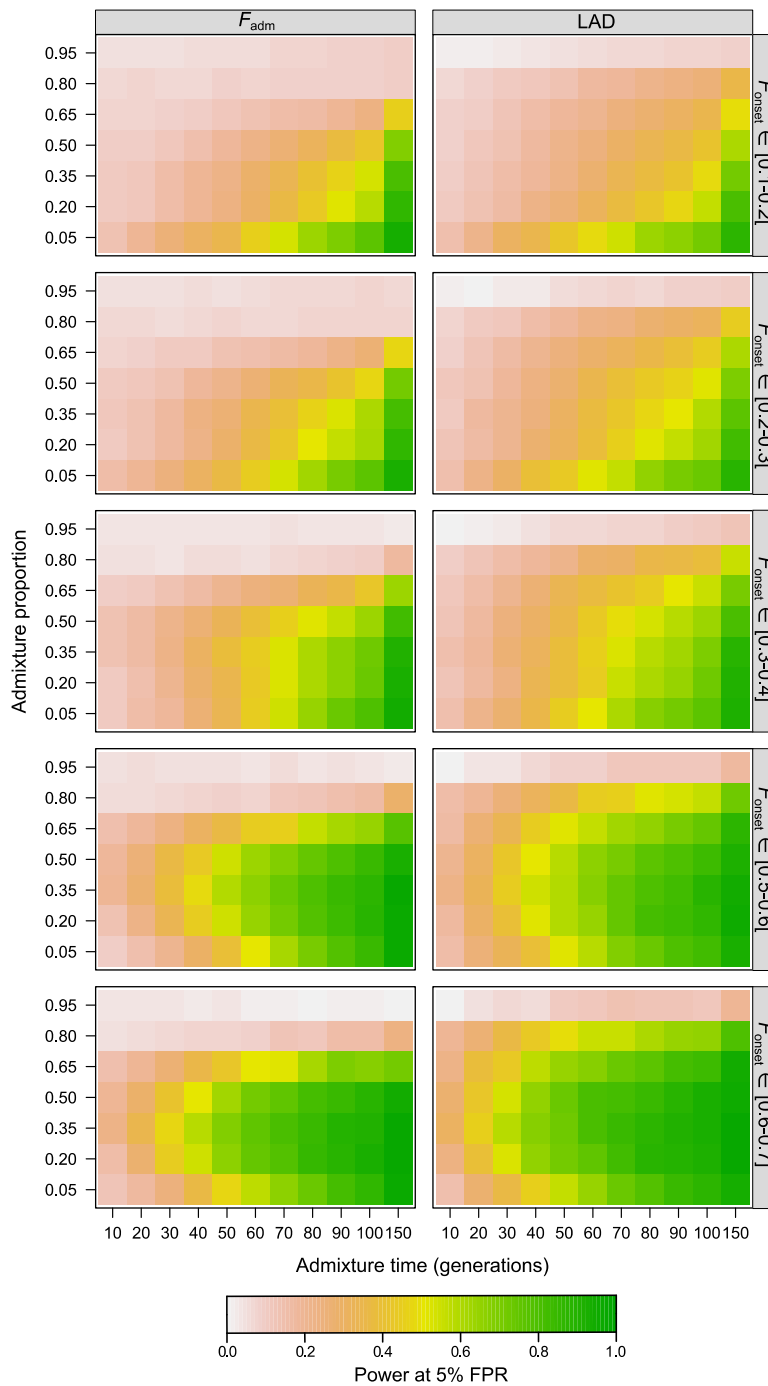


Figure S10. Effects of the frequency of the beneficial mutation ($s = 0.05$) on the power to detect adaptive admixture.

Effects on the detection power of F_{adm} and LAD of admixture time T_{adm} , admixture proportion α and F_{onset} , the frequency of the beneficial mutation in the source population at the time of admixture T_{adm} . Colour indicates average detection power for a FPR = 5% threshold, across combinations of the remaining parameters.

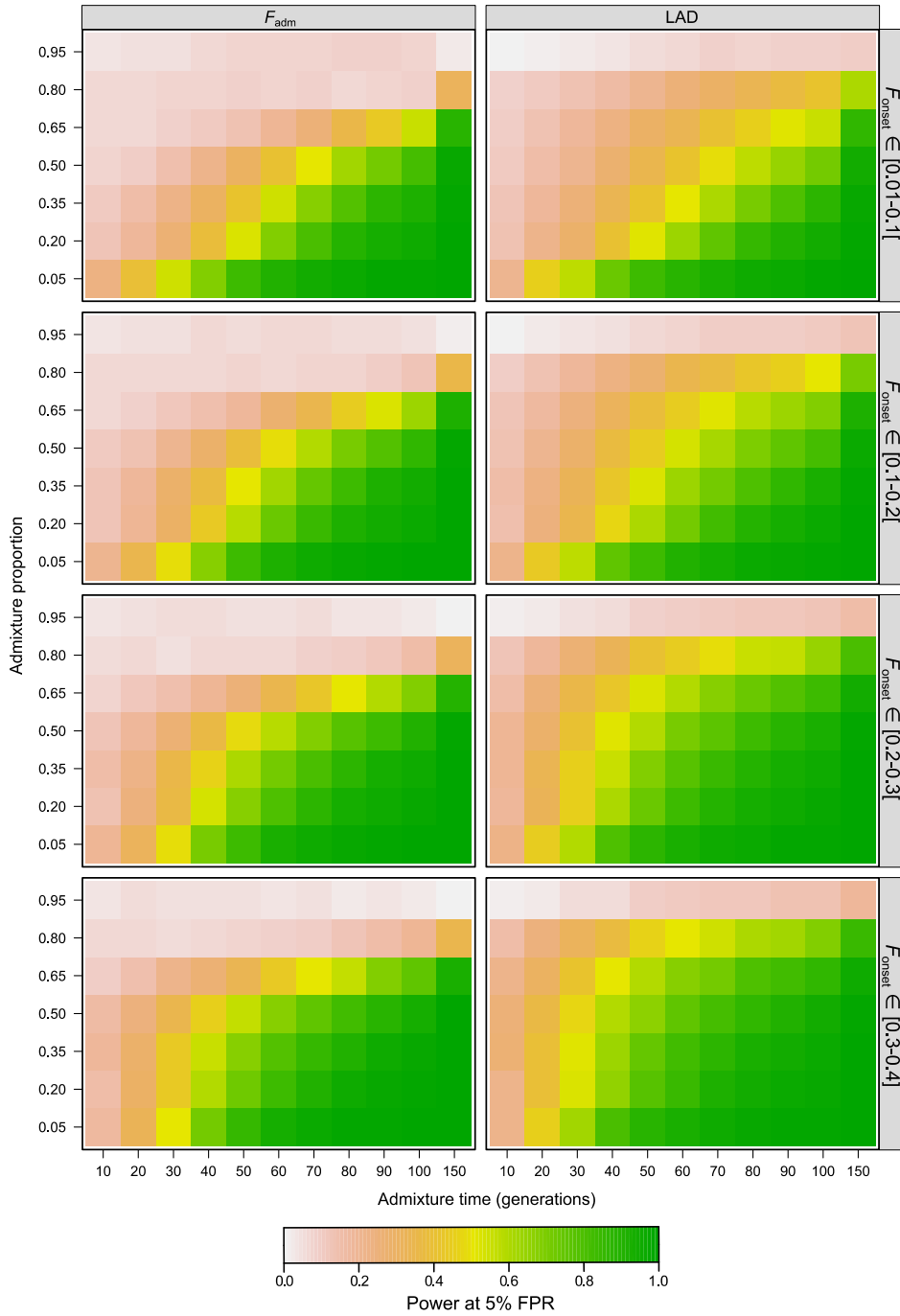


Figure S11. Effects of the frequency of the beneficial mutation ($s = 0.10$) on the power to detect adaptive admixture.

Effects on the detection power of F_{adm} and LAD of admixture time T_{adm} , admixture proportion α and F_{onset} , the frequency of the beneficial mutation in the source population at the time of admixture T_{adm} . Colour indicates average detection power for a FPR = 5% threshold, across combinations of the remaining parameters.

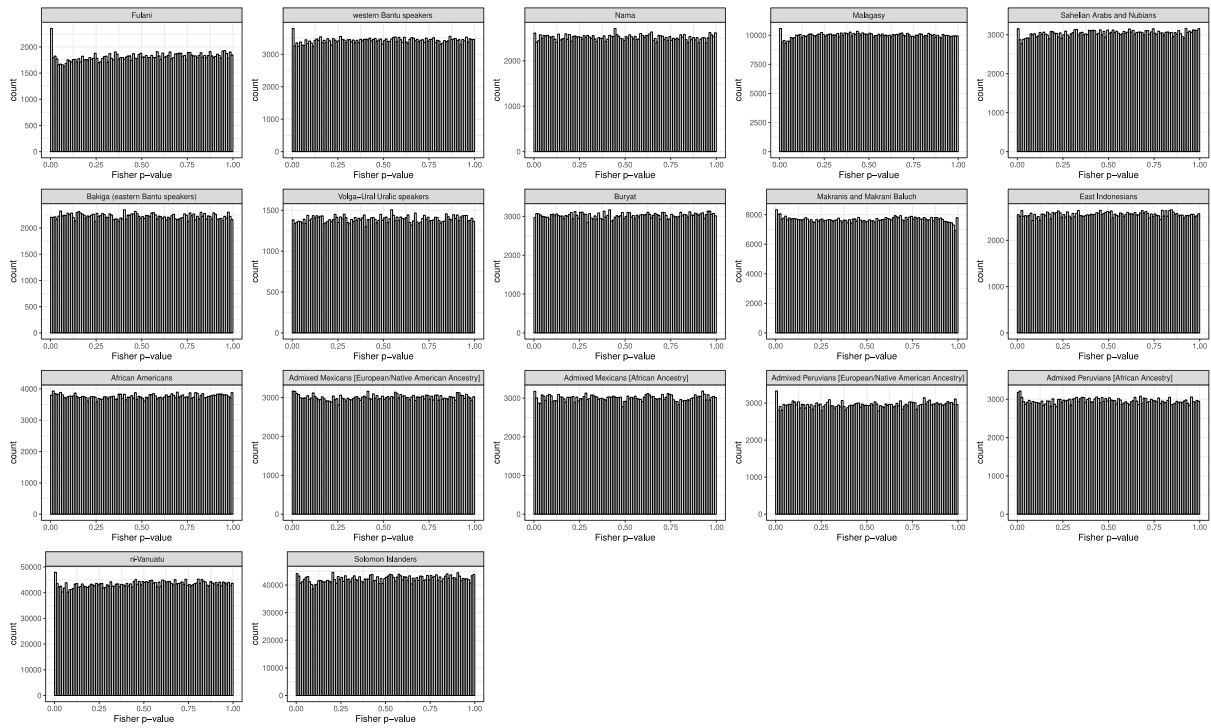


Figure S12. Distributions of Fisher's combined P -values in the empirical data.

Histograms of combined P -values using Fisher's method, for the 15 analysed admixed populations. The P -values are uniformly distributed, except for certain populations where there is an excess of small P -values, corresponding to the populations where signals for adaptive admixture were found.

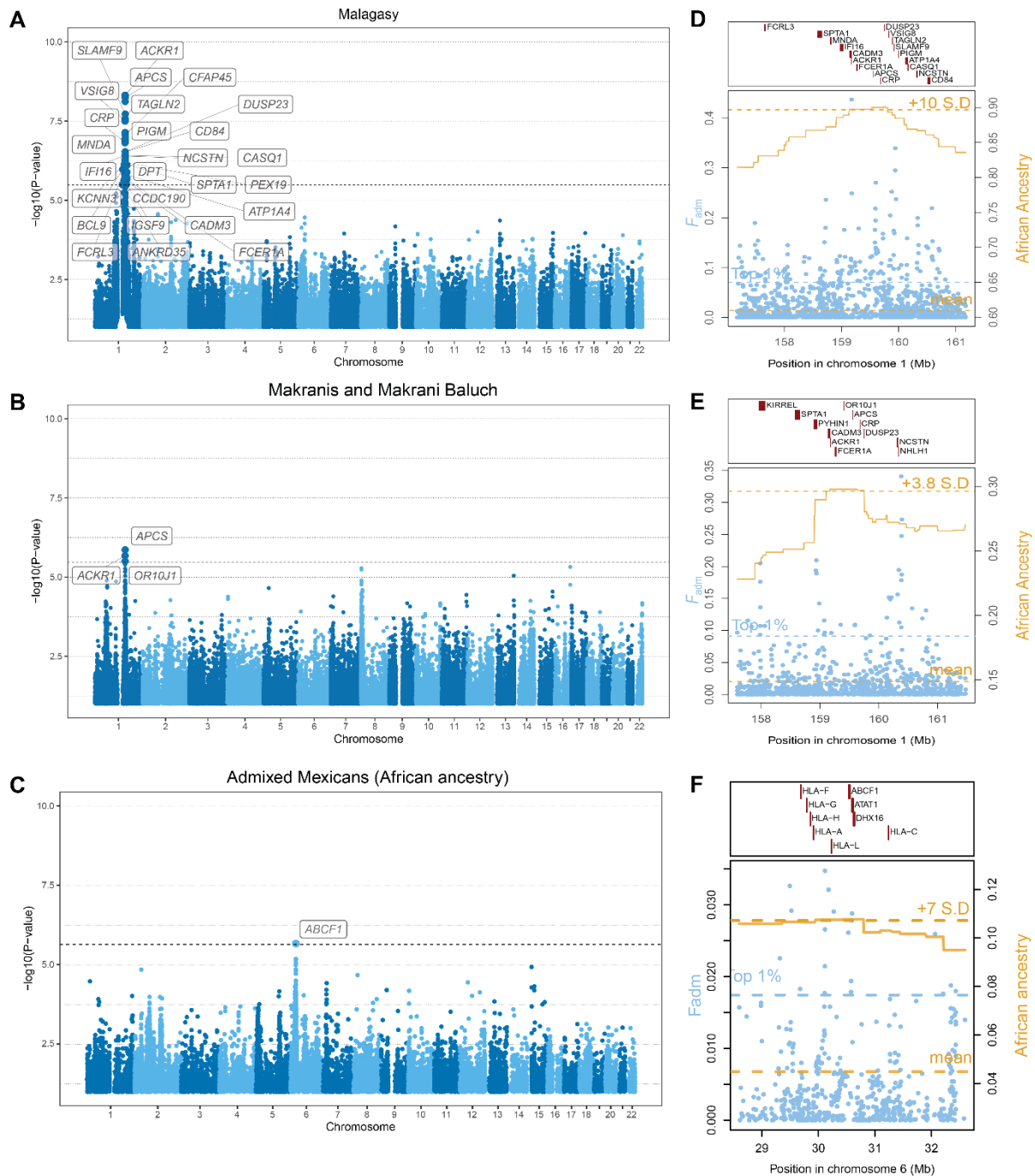


Figure S13. Other previously reported genomic signals of adaptive admixture.

(A) Genome-wide signals of adaptive admixture in Malagasy populations from Madagascar.

(B) Genome-wide signals of adaptive admixture in African-descent Makranis and Makrani Baluch from Pakistan.

(C) Genome-wide signals of adaptive admixture in admixed Mexicans (African ancestry).

(A-C) Highlighted blue points indicate variants that passed the Bonferroni significance threshold (shown by a horizontal dotted line). Gene labels were attributed based on the gene with the highest V2G score within 250-kb of the candidate variant.

(D) Local signatures of adaptive admixture for the *ACKRI* region in Malagasy from Madagascar.

(E) Local signatures of adaptive admixture for the *ACKRI* region in Makranis and Makrani Baluch from Pakistan.

(F) Local signatures of adaptive admixture for the *HLA* class I region in admixed Mexicans.

(D-F) Light blue points indicate F_{adm} values for individual variants. The gold solid line indicates the average African local ancestry.

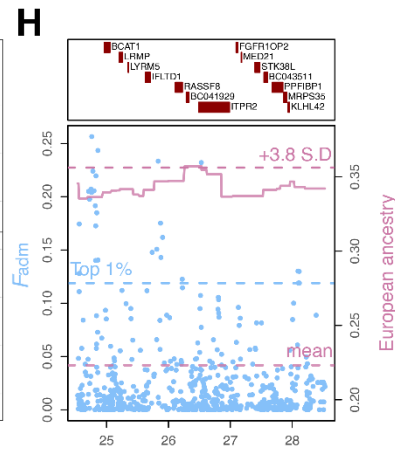
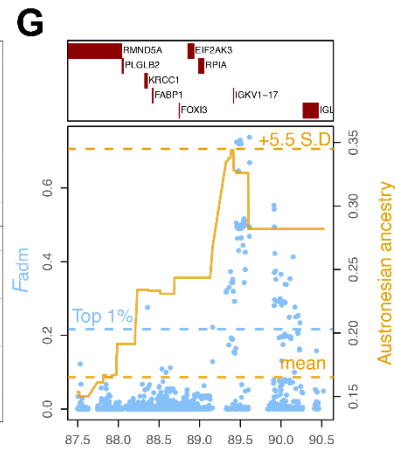
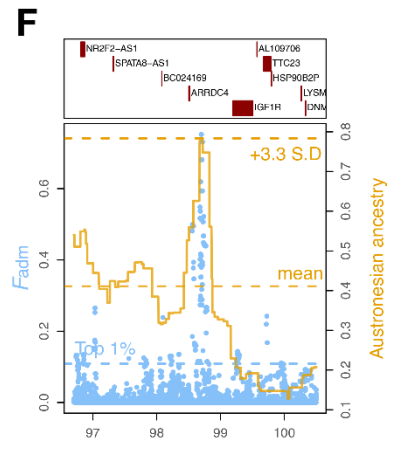
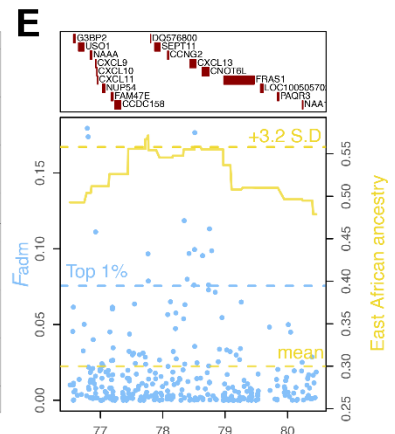
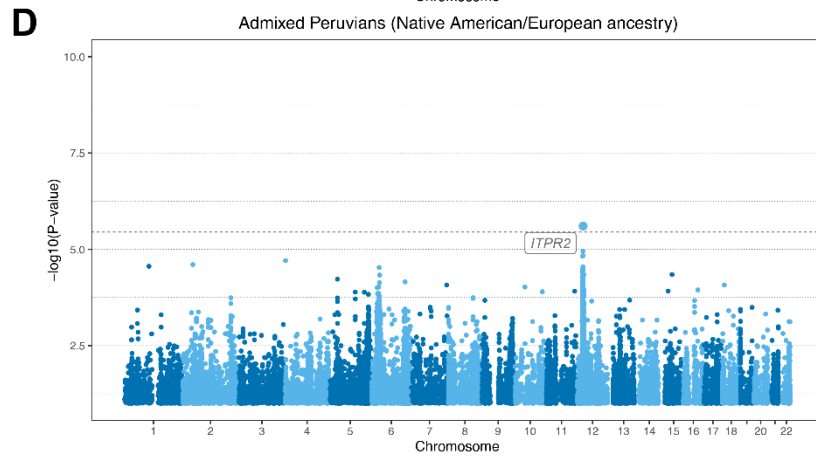
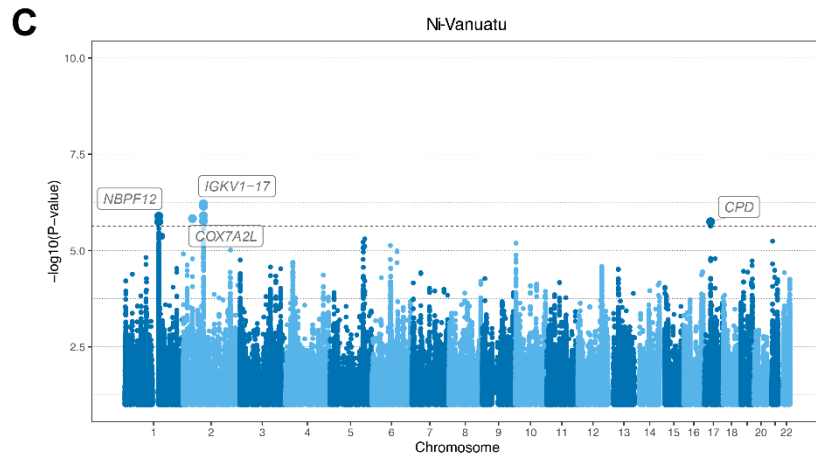
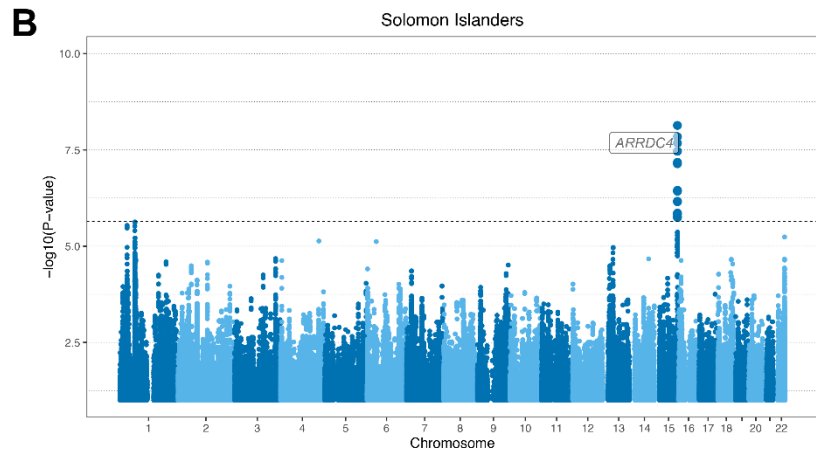
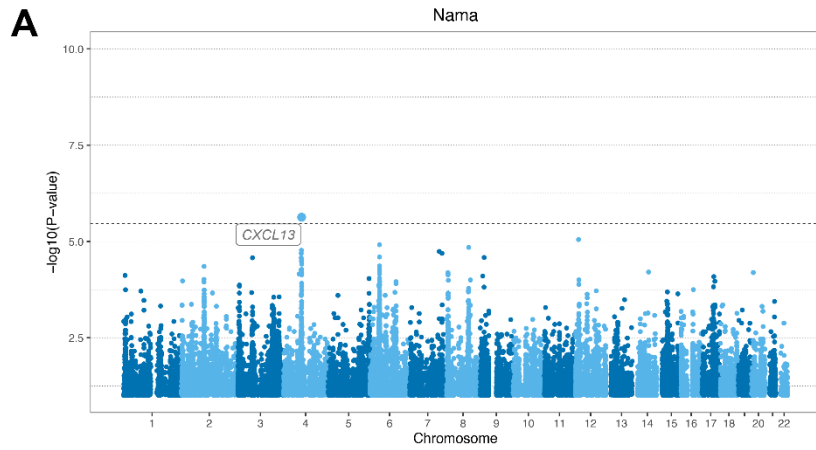


Figure S14. Other novel genomic signals of adaptive admixture.

(A) Genome-wide signals of adaptive admixture in the Nama from South Africa.

(B) Genome-wide signals of adaptive admixture in Solomon Islanders.

(C) Genome-wide signals of adaptive admixture in Vanuatu Islanders.

(D) Genome-wide signals of adaptive admixture in admixed Peruvians.

(A-D) Highlighted blue points indicate variants that passed the Bonferroni significance threshold (shown by a horizontal dotted line). Gene labels were attributed based on the gene with the highest V2G score within 250-kb of the candidate variant.

(E) Local signatures of adaptive admixture for the *CNOT6L/CXCL13* region in the Nama from South Africa.

(F) Local signatures of adaptive admixture for the *ARRDC4* region in Solomon Islanders.

(G) Local signatures of adaptive admixture for the *IGKVI-17* region in Vanuatu Islanders.

(H) Local signatures of adaptive admixture for the *ITPR2* region in admixed Peruvians.

(E-H) Light blue points indicate F_{adm} values for individual variants. The yellow, gold and pink solid lines indicate average local ancestry from East Africans, Austronesians and Europeans respectively.

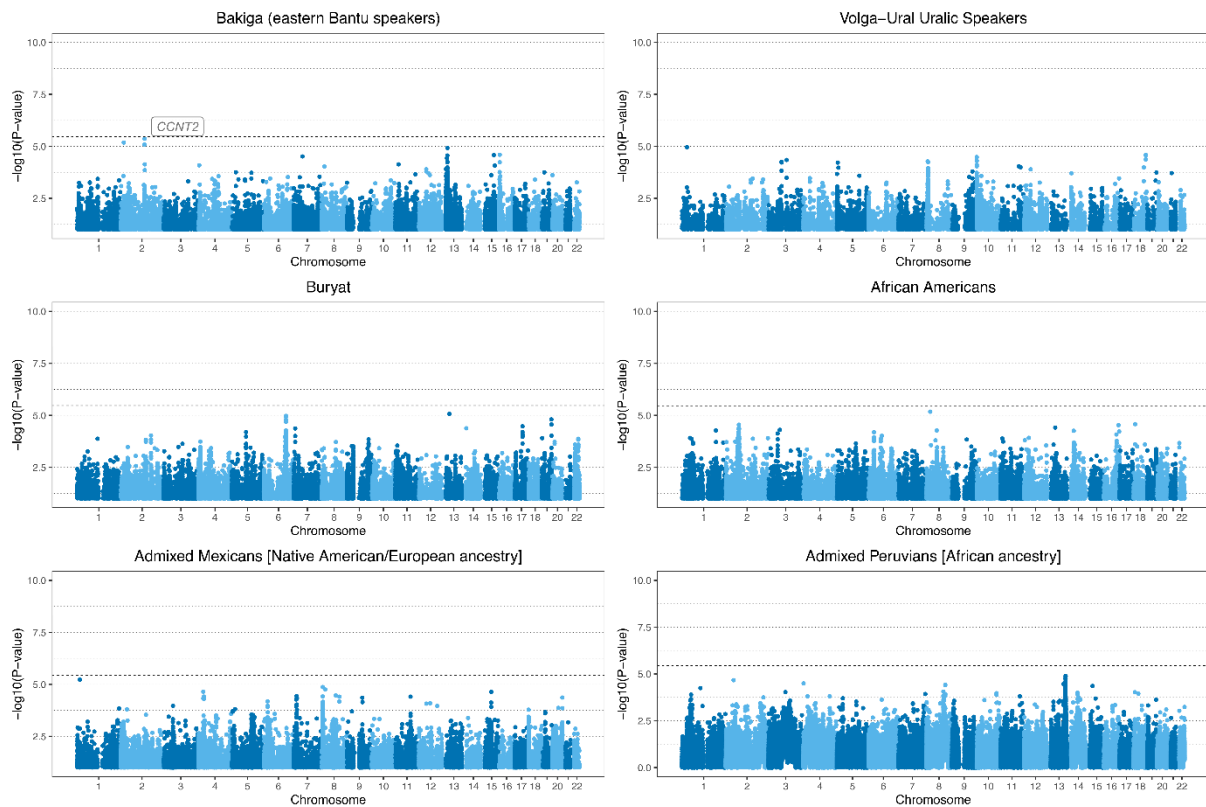


Figure S15. Genome scans for populations where there is no evidence for adaptive admixture.

Manhattan plots of $-\log_{10}(P\text{-values})$ for the combined Fisher's method, in the remaining 6 admixed populations where no variant passes the Bonferroni significance threshold (shown by a horizontal dotted line).

Chapter 8: Discussion

Implications of the presented work: simulation analyses

Based on the simulation analyses and subsequent power evaluation, three key points can be made to population geneticists studying selection in admixed populations. First, population bottlenecks occurring in the admixed population will reduce detection power. Given that this is not an unreasonable assumption in most modern human populations, changes in population size should be inferred and modelled, when possible and when not, more stringent thresholds defining genomic outliers should be employed. A possible solution is to use of methods that combine multiple statistics such as Fisher's method, which maintain a decent detection power at stringent thresholds.

Second, source population proxies should be chosen with utmost care. Failure to do so may not only result in a decrease in detection power and multiple false negatives, but also and more worryingly, an increase in false positives, should positive selection be acting exclusively on one of the source proxies.

And third, power is severely limited for very recent admixture events (less than 500 years before present). Although this concerns only a few admixed populations, this is further complicated due to the lack of accurate source proxies for these. In the case of African Americans, failure to account for accurate African-ancestry proxies produced false signals of post-admixture selection (Patin et al., 2017). For populations with Native American ancestry, this is even more problematic given that accurate proxies no longer exist in an unadmixed form (and even when considering ancient DNA, pre-colonization native-American populations were highly structured (Posth et al., 2018)).

Implications of the presented work: empirical data

All the candidates identified and described in this manuscript were described before as either under positive selection or under adaptive admixture. This should not come as a surprise, given that the methods we used – as well as those used by others to identify these loci – are powerful to detect mutations that have conferred a (very) strong selective advantage. This is the

case for the *ACKRI* locus (or more specifically the Duffy-null *FY*B^{ES}* allele) found in multiple admixed populations carrying Central African ancestry in northeast Africa, Madagascar and Pakistan. Selection coefficients associated with this beneficial mutation have been estimated to be higher than 0.2 (Pierron et al., 2018). However, the *LCT* locus, for which beneficial mutations selection coefficients were estimated to be higher than 0.04 (Breton et al., 2014; Vicente et al., 2019) was not an outlier in the Nama from South Africa or the Bakiga, a Bantu-speaking population from East Africa, even though it was reported as such in previous studies (Breton et al., 2014; Patin et al., 2017). There are multiple reasons that might explain this observation. First, the overall high conservative nature of the multiple testing correction used (Bonferroni correction), as the locus in the Bakiga is very close to the significance threshold. Second, the low sample size for the Nama population (limited to only 20 individuals) could have reduced detection power. And finally, it is possible that the signal at the *LCT* locus on the Nama may not be as strong as in other populations with East-African ancestry (Breton et al., 2014).

The identified candidates provide insights into the environmental pressures the analysed admixed populations have been subject to since admixture, especially pathogen pressures, since several outliers are involved in immune response. For instance, the strong signals at the *ACKRI* locus in multiple populations suggest ongoing selection for *Plasmodium vivax* malaria resistance, as their location coincides with the distribution of *Plasmodium vivax* and malaria cases (World Health Organization, 2017). Conversely, the signal found for non-west African ancestry at the *APOLI* locus in the Fulani of Burkina Faso may suggest that this population may have not been extensively exposed to African Trypanosomiasis, given that the G1 and G2 haplotypes of West African origin confer not only resistance to this disease but increased risk for kidney disease (Franco et al., 2014; Genovese et al., 2010; Ko et al., 2013; Limou et al., 2014). Also, and quite puzzlingly, the identified false signal of adaptive admixture at the *CXCL13* gene in the Nama was initially interpreted as selection for non-San ancestry in this region, given that San-related ancestry is positively correlated with an increased risk for tuberculosis (Chimusa et al., 2014). However, this signal was driven by positive selection in the proxy used for San ancestry (the Ju'hoansi) in the Nama. The context behind this observation remains to be explored. Finally, the *HLA* locus identified in two populations where it was reported previously (Lindo et al., 2016; Patin et al., 2017; Zhou et al., 2016) highlights this region as a hotspot for genetic adaptation in immune response against pathogens (Patin et al., 2017;

Prugnolle et al., 2005b) with genetic diversity maintained through balancing selection in the form of positive frequency-dependent selection (Alter et al., 2017).

Limitations of the presented work

Beyond the outlier *vs.* model-based approach debate that has been highlighted previously (see Chapters 2, 5 and 6), our analyses on adaptive admixture do present limitations. Some aspects of the admixture model that hold true when studying modern human populations were not explored. For instance, gene flow between source populations was not considered, even though it was probably the case in most analysed populations. The expectation from this is a potential reduction in detection power, especially for the LAD statistic. Indeed, gene flow between parental sources reduces genetic divergence between the two populations, which could be seen as both having a more recent split time, which in turn has been shown to reduce detection power (due to a higher variance in the local ancestry inference). Also, admixture events involving more than two sources were not explored, although the two statistics employed to detect adaptive admixture, F_{adm} and LAD, can be easily extended to more than two sources and we expect most of the observations regarding detection power to hold in these cases. Additionally, how the accuracy of local ancestry inference can affect LAD detection power were not thoroughly explored. For example, LAI accuracy can be reduced when studying old admixture events or admixture events between closely related sources (Yelmen et al., 2021b) and recent LAI methods show higher accuracy than that used in our study. Furthermore, we expect that scenarios resulting in directional biases in admixture proportion estimates (e.g., unbalanced sample sizes (Toyama et al., 2020)) could also decrease F_{adm} detection power.

Importantly, even though the present work qualitatively highlights the importance of admixture in human adaptation, it does not quantitatively answer how much of the genetic adaptations in human evolutionary history were enabled and enhanced by admixture. First, because old admixture with selection events cannot be analysed using the presented methods (especially local ancestry) and present-day data (ancient DNA is often needed in these cases). Second, some of the admixture events explored here are quite complex (Siberian admixed populations for instance), and reliable source proxies cannot always be obtained (again this would benefit from incorporating ancient DNA data). And third, conceptually speaking, the effect of admixture on the fixation probability of a beneficial mutation was not explored.

Although previous theoretical works have derived fixation probabilities for newly arising beneficial mutations in deme-structured populations experiencing reciprocal gene flow (Maruyama and Kimura, 1980; Slatkin, 1981), the fixation probability of a beneficial mutation introduced through single or recurrent admixture events in a given population has not been fully explored. Because the fixation probability depends on the initial frequency, which in the case of adaptive admixture will be higher than $\frac{1}{2N}$, the fixation probability should be higher for a mutation under adaptive admixture. Nevertheless, this is probably more complex, due to the effects of linked variation, which might be deleterious (especially in the case of low population sizes due to a higher mutational load, see Chapter 5).

Perspectives and future directions

Most of the adaptive admixture analyses conducted here, as well as elsewhere, have focused essentially on the extremes of the distribution of selection coefficients. This is natural given that these are the loci where statistical power is the highest. However, by doing that, a fair number of regions in the genome that may have been weak-to-moderately advantageous have been excluded. One example of this are traits with a polygenic architecture, which is governed by a multitude of sites with individually weak effects. Selection occurring in these traits, known as polygenic selection, would thus pass completely undetected. Some methods have been specifically developed to study polygenic selection, and in theory, could be applied to admixed populations if the admixture proportions and individual effect sizes of each site could be correctly estimated in both the source populations and the admixed population. However, these methods have been shown to be sensitive to fine scale population structure, something that is not easy to properly account for. Nevertheless, detection power for weak-to-moderate positive selection could be improved through a combination of hyper-realistic simulations through generative adversarial networks (Wang et al., 2021; Yelmen et al., 2021a) that cannot be told apart from real data, an accurate demographic inference and admixture specific statistics, such as F_{adm} , LAD or the newly proposed iDAT (a combination between local ancestry and derived allele information (Hamid et al., 2021)) integrated in a probabilistic framework such as SWIFr (Sugden et al., 2018).

Ancient DNA, as mentioned many times in this manuscript, can improve the resolution of population genetic analyses, including the inference of positive selection and adaptive admixture. The use of contemporary populations as proxy source populations, which can be one of the biggest hurdles when studying past admixture events, can be bypassed with ancient samples, although ancient DNA data only partly circumvent the problem, due to the limited number of high-quality samples (something that has been improving at a rapid pace) and due to geographical discontinuities. Perhaps one of the biggest opportunities ancient DNA brings, in terms of positive selection, is the study of time-series type data. This allows population geneticists to directly observe changes in allele frequencies over time, thus allowing them to estimate the onset and intensity of past selection events (Dehasque et al., 2020). Time-series data is not a new concept, as it was inherent to “evolve and resequence” experiments, where under controlled laboratory conditions, individuals from an evolving population could be sampled and sequenced at different points in time but were limited to species with short generation times (Turner and Miller, 2012) or with an asexual mode of reproduction (Bennett et al., 1990; Good et al., 2017). With ancient DNA, sampling multiple time points in sexually reproducing species became possible and combined with ecological data, has been used to infer past selection pressures and even temporal changes in said pressures in domestic species such as horses (Librado et al., 2017), maize (Ramos-Madrigal et al., 2016) and dogs (Ollivier et al., n.d.). In humans, time series datasets from ancient DNA have been used to infer selection due to dietary changes associated with domestication (Mathieson et al., 2015; Mathieson and Mathieson, 2018; Sverrisdóttir et al., 2014) or due to exposure to pathogens (Kerner et al., 2021).

References

- Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S.G.E., Maiers, M., Guethlein, L.A., Tavoularis, S., Little, A.-M., Green, R.E., Norman, P.J., Parham, P., 2011. The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science*. <https://doi.org/10.1126/science.1209202>
- Aeschbacher, S., Selby, J.P., Willis, J.H., Coop, G., 2017. Population-genomic inference of the strength and timing of selection against gene flow. *PNAS* 114, 7061–7066. <https://doi.org/10.1073/pnas.1616755114>
- Alter, I., Gragert, L., Fingerson, S., Maiers, M., Louzoun, Y., 2017. HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. *PLOS Computational Biology* 13, e1005693. <https://doi.org/10.1371/journal.pcbi.1005693>
- Anderson, E., 1949. *Introgressive hybridization*. J. Wiley, New York.
- Anderson, T.M., vonHoldt, B.M., Candille, S.I., Musiani, M., Greco, C., Stahler, D.R., Smith, D.W., Padhukasahasram, B., Randi, E., Leonard, J.A., Bustamante, C.D., Ostrander, E.A., Tang, H., Wayne, R.K., Barsh, G.S., 2009. Molecular and Evolutionary History of Melanism in North American Gray Wolves. *Science*. <https://doi.org/10.1126/science.1165448>
- Arnold, B.J., Lahner, B., DaCosta, J.M., Weisman, C.M., Hollister, J.D., Salt, D.E., Bomblies, K., Yant, L., 2016. Borrowed alleles and convergence in serpentine adaptation. *PNAS* 113, 8320–8325. <https://doi.org/10.1073/pnas.1600405113>
- Arnold, M.L., Buckner, C.M., Robinson, J.J., 1991. Pollen-mediated introgression and hybrid speciation in Louisiana irises. *PNAS* 88, 1398–1402. <https://doi.org/10.1073/pnas.88.4.1398>
- Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., Rodriguez-Santana, J., Burchard, E.G., Halperin, E., 2012. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367. <https://doi.org/10.1093/bioinformatics/bts144>
- Barbato, M., Hailer, F., Upadhyay, M., Del Corvo, M., Colli, L., Negrini, R., Kim, E.-S., Crooijmans, R.P.M.A., Sonstegard, T., Ajmone-Marsan, P., 2020. Adaptive introgression from indicine cattle into white cattle breeds from Central Italy. *Sci Rep* 10, 1279. <https://doi.org/10.1038/s41598-020-57880-4>
- Barton, N., Bengtsson, B.O., 1986. The barrier to genetic exchange between hybridising populations. *Heredity* 57, 357–376. <https://doi.org/10.1038/hdy.1986.135>
- Barton, N.H., 1979. The dynamics of hybrid zones. *Heredity* 43, 341–359. <https://doi.org/10.1038/hdy.1979.87>
- Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E., Langley, C.H., 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in *Drosophila simulans*. *PLOS Biology* 5, e310. <https://doi.org/10.1371/journal.pbio.0050310>
- Bengtsson, B.O., 1985. The flow of genes through a genetic barrier, in: Greenwood, J.J., Harvey, P.H., Slatkin, M. (Eds.), *Evolution Essays in Honour of John Maynard Smith*. Cambridge University Press, pp. 31–42.
- Bennett, A.F., Dao, K.M., Lenski, R.E., 1990. Rapid evolution in response to high-temperature selection. *Nature* 346, 79–81. <https://doi.org/10.1038/346079a0>

- Bernstein, F., 1931. Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Presented at the Comitato Italiano per lo Studio dei Problemi della Popolazione, Istituto Poligrafico dello Stato, Rome.
- Birky, C.W., Walsh, J.B., 1988. Effects of linkage on rates of molecular evolution. *PNAS* 85, 6414–6418. <https://doi.org/10.1073/pnas.85.17.6414>
- Breton, G., Schlebusch, C.M., Lombard, M., Sjödin, P., Soodyall, H., Jakobsson, M., 2014. Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. *Current Biology* 24, 852–858. <https://doi.org/10.1016/j.cub.2014.02.041>
- Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., Bustamante, C.D., 2012. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *hbio* 84, 343–364. <https://doi.org/10.3378/027.084.0401>
- Burgarella, C., Cubry, P., Kane, N.A., Varshney, R.K., Mariac, C., Liu, X., Shi, C., Thudi, M., Couderc, M., Xu, X., Chitikeni, A., Scarcelli, N., Barnaud, A., Rhoné, B., Dupuy, C., François, O., Berthouly-Salazar, C., Vigouroux, Y., 2018. A western Sahara centre of domestication inferred from pearl millet genomes. *Nat Ecol Evol* 2, 1377–1380. <https://doi.org/10.1038/s41559-018-0643-y>
- Campbell, C.D., Chong, J.X., Malig, M., Ko, A., Dumont, B.L., Han, L., Vives, L., O’Roak, B.J., Sudmant, P.H., Shendure, J., Abney, M., Ober, C., Eichler, E.E., 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* 44, 1277–1281. <https://doi.org/10.1038/ng.2418>
- Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. *Nature* 325, 31–36. <https://doi.org/10.1038/325031a0>
- Cavalli-Sforza, L.L., Bodmer, W.F., 1999. *The Genetics of Human Populations*. Courier Corporation.
- Chacón-Duque, J.-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., Hünemeier, T., Ramallo, V., Silva de Cerqueira, C.C., Hurtado, M., Villegas, V., Granja, V., Villena, M., Vásquez, R., Llop, E., Sandoval, J.R., Salazar-Granara, A.A., Parolin, M.-L., Sandoval, K., Peñaloza-Espinosa, R.I., Rangel-Villalobos, H., Winkler, C.A., Klitz, W., Bravi, C., Molina, J., Corach, D., Barrantes, R., Gomes, V., Resende, C., Gusmão, L., Amorim, A., Xue, Y., Dugoujon, J.-M., Moral, P., González-José, R., Schuler-Faccini, L., Salzano, F.M., Bortolini, M.-C., Canizales-Quinteros, S., Poletti, G., Gallo, C., Bedoya, G., Rothhammer, F., Balding, D., Hellenthal, G., Ruiz-Linares, A., 2018. Latin Americans show widespread *Converso* ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun* 9, 5388. <https://doi.org/10.1038/s41467-018-07748-z>
- Chakraborty, R., 1986. Gene admixture in human populations: Models and predictions. *American Journal of Physical Anthropology* 29, 1–43. <https://doi.org/10.1002/ajpa.1330290502>
- Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics* 134, 1289–1303.
- Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., van Helden, P.D., Mulder, N.J., Price, A.L., Hoal, E.G., 2014. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human Molecular Genetics* 23, 796–809. <https://doi.org/10.1093/hmg/ddt462>
- Choin, J., Mendoza-Revilla, J., Arauna, L.R., Cuadros-Espinoza, S., Cassar, O., Larena, M., Ko, A.M.-S., Harmant, C., Laurent, R., Verdu, P., Laval, G., Boland, A., Olaso, R.,

- Deleuze, J.-F., Valentin, F., Ko, Y.-C., Jakobsson, M., Gessain, A., Excoffier, L., Stoneking, M., Patin, E., Quintana-Murci, L., 2021. Genomic insights into population history and biological adaptation in Oceania. *Nature* 592, 583–589. <https://doi.org/10.1038/s41586-021-03236-5>
- Dehasque, M., Ávila-Arcos, M.C., Díez-del-Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E.D., Malaspinas, A.-S., Marques-Bonet, T., Martin, M.D., Murray, G.G.R., Papadopulos, A.S.T., Therkildsen, N.O., Wegmann, D., Dalén, L., Foote, A.D., 2020. Inference of natural selection from ancient DNA. *Evolution Letters* 4, 94–108. <https://doi.org/10.1002/evl3.165>
- Deng, L., Ruiz-Linares, A., Xu, S., Wang, S., 2016. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci Rep* 6, 21766. <https://doi.org/10.1038/srep21766>
- Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.-L., Patin, E., Quintana-Murci, L., 2016. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *The American Journal of Human Genetics* 98, 5–21. <https://doi.org/10.1016/j.ajhg.2015.11.014>
- Edelman, N.B., Mallet, J., 2021. Prevalence and Adaptive Impact of Introgression. *Annual Review of Genetics* 55, 265–283. <https://doi.org/10.1146/annurev-genet-021821-020805>
- Ehrlich, P.R., Raven, P.H., 1969. Differentiation of Populations. *Science*. <https://doi.org/10.1126/science.165.3899.1228>
- Enard, D., Petrov, D.A., 2018. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* 175, 360-371.e13. <https://doi.org/10.1016/j.cell.2018.08.034>
- Eriksson, A., Manica, A., 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *PNAS* 109, 13956–13960. <https://doi.org/10.1073/pnas.1200567109>
- Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* 131, 479–491.
- Fan, S., Hansen, M.E.B., Lo, Y., Tishkoff, S.A., 2016. Going global by adapting local: A review of recent human adaptation. *Science* 354, 54–59. <https://doi.org/10.1126/science.aaf5098>
- Fay, J.C., Wu, C.-I., 2000. Hitchhiking Under Positive Darwinian Selection. *Genetics* 155, 1405–1413. <https://doi.org/10.1093/genetics/155.3.1405>
- Fernandes, V., Brucato, N., Ferreira, J.C., Pedro, N., Cavadas, B., Ricaut, F.-X., Alshamali, F., Pereira, L., 2019. Genome-Wide Characterization of Arabian Peninsula Populations: Shedding Light on the History of a Fundamental Bridge between Continents. *Molecular Biology and Evolution* 36, 575–586. <https://doi.org/10.1093/molbev/msz005>
- Fisher, R.A., 1952. Statistical methods in genetics. *Heredity* 6, 1–12. <https://doi.org/10.1038/hdy.1952.1>
- Fisher, R.A., 1922. On the Dominance Ratio. *Proceedings of the Royal Society of Edinburgh* 42, 321–341.
- Fortes-Lima, C.A., Laurent, R., Thouzeau, V., Toupance, B., Verdu, P., 2020. Complex genetic admixture histories reconstructed with Approximate Bayesian Computations. <https://doi.org/10.1101/761452>
- Franco, J.R., Simarro, P.P., Diarra, A., Jannin, J.G., 2014. Epidemiology of human African trypanosomiasis. *CLEP* 6, 257–275. <https://doi.org/10.2147/CLEP.S39728>

- Garud, N.R., Messer, P.W., Buzbas, E.O., Petrov, D.A., 2015. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLOS Genetics* 11, e1005004. <https://doi.org/10.1371/journal.pgen.1005004>
- Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Knob, A.L.U., Bernhardt, A.J., Hicks, P.J., Nelson, G.W., Vanhollenbeke, B., Winkler, C.A., Kopp, J.B., Pays, E., Pollak, M.R., 2010. Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science*. <https://doi.org/10.1126/science.1193032>
- Glass, B., Li, C.C., 1953. The dynamics of racial intermixture—an analysis based on the American Negro. *Am J Hum Genet* 5, 1–20.
- Good, B.H., McDonald, M.J., Barrick, J.E., Lenski, R.E., Desai, M.M., 2017. The dynamics of molecular evolution over 60,000 generations. *Nature* 551, 45–50. <https://doi.org/10.1038/nature24287>
- Graham, A.M., Peters, J.L., Wilson, R.E., Muñoz-Fuentes, V., Green, A.J., Dorfsman, D.A., Valqui, T.H., Winker, K., McCracken, K.G., 2021. Adaptive introgression of the beta-globin cluster in two Andean waterfowl. *Heredity* 127, 107–123. <https://doi.org/10.1038/s41437-021-00437-6>
- Gravel, S., 2012. Population Genetics Models of Local Ancestry. *Genetics* 191, 607–619. <https://doi.org/10.1534/genetics.112.139808>
- Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspina, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušić, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., Rasilla, M. de la, Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A Draft Sequence of the Neandertal Genome. *Science*. <https://doi.org/10.1126/science.1188021>
- Guan, Y., 2014. Detecting Structure of Haplotypes and Local Ancestry. *Genetics* 196, 625–642. <https://doi.org/10.1534/genetics.113.160697>
- Haasl, R.J., McCarty, C.A., Payseur, B.A., 2013. Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. *Eur J Hum Genet* 21, 554–562. <https://doi.org/10.1038/ejhg.2012.258>
- Hahn, M.W., 2008. Toward a Selection Theory of Molecular Evolution. *Evolution* 62, 255–265. <https://doi.org/10.1111/j.1558-5646.2007.00308.x>
- Haller, B.C., Messer, P.W., 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* 36, 632–637. <https://doi.org/10.1093/molbev/msy228>
- Haller, B.C., Messer, P.W., 2016. SLiM: An Evolutionary Simulation Framework 710.
- Hamid, I., Korunes, K.L., Beza, S., Goldberg, A., 2021. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *eLife* 10, e63177. <https://doi.org/10.7554/eLife.63177>
- Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J.C., Douches, D.S., Farré, E.M., Veilleux, R.E., Buell, C.R., 2017. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *PNAS* 114, E9999–E10008. <https://doi.org/10.1073/pnas.1714380114>
- Hardy, G.H., 1908. Mendelian proportions in a mixed population. *Science* 28, 49–50.

- Harris, K., Nielsen, R., 2016. The Genetic Cost of Neanderthal Introgression. *Genetics* 203, 881–891. <https://doi.org/10.1534/genetics.116.186890>
- Harris, R.B., Sackman, A., Jensen, J.D., 2018. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLOS Genetics* 14, e1007859. <https://doi.org/10.1371/journal.pgen.1007859>
- Hedrick, P.W., 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology* 22, 4606–4618. <https://doi.org/10.1111/mec.12415>
- Heiser, C.B., 1979. Hybrid Populations of *Helianthus Divaricatus* and *H. Microcephalus* After 22 Years. *TAXON* 28, 71–75. <https://doi.org/10.2307/1219560>
- Heiser Jr., C.B., 1951. HYBRIDIZATION IN THE ANNUAL SUNFLOWERS: *HELIANTHUS ANNUUS* x *H. DEBILIS* VAR. *CUCUMERIFOLIUS*. *Evolution* 5, 42–51. <https://doi.org/10.1111/j.1558-5646.1951.tb02758.x>
- Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., Myers, S., 2014. A Genetic Atlas of Human Admixture History. *Science* 343, 747–751. <https://doi.org/10.1126/science.1243518>
- Henn, B.M., Cavalli-Sforza, L.L., Feldman, M.W., 2012. The great human expansion. *PNAS* 109, 17758–17764. <https://doi.org/10.1073/pnas.1212380109>
- Hermisson, J., Pennings, P.S., 2005. Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics* 169, 2335–2352. <https://doi.org/10.1534/genetics.104.036947>
- Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., Flicek, P., 2016. Ensembl comparative genomics resources. *Database* 2016, bav096. <https://doi.org/10.1093/database/bav096>
- Hill, W.G., Robertson, A., 1966. The effect of linkage on limits to artificial selection. *Genet Res* 8, 269–294.
- Hilmarsson, H., Kumar, A.S., Rastogi, R., Bustamante, C.D., Montserrat, D.M., Ioannidis, A.G., 2021. High Resolution Ancestry Deconvolution for Next Generation Genomic Data. <https://doi.org/10.1101/2021.09.19.460980>
- Hu, X.-J., Yang, J., Xie, X.-L., Lv, F.-H., Cao, Y.-H., Li, W.-R., Liu, M.-J., Wang, Y.-T., Li, J.-Q., Liu, Y.-G., Ren, Y.-L., Shen, Z.-Q., Wang, F., Hehua, Ee., Han, J.-L., Li, M.-H., 2019. The Genome Landscape of Tibetan Sheep Reveals Adaptive Introgression from Argali and the History of Early Human Settlements on the Qinghai–Tibetan Plateau. *Molecular Biology and Evolution* 36, 283–303. <https://doi.org/10.1093/molbev/msy208>
- Hudson, R.R., Slatkin, M., Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132, 583–589. <https://doi.org/10.1093/genetics/132.2.583>
- Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z.X.P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, Jian, Wang, Jun, Nielsen, R., 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. <https://doi.org/10.1038/nature13408>
- Hufford, M.B., Lubinsky, P., Pyhäjärvi, T., Devengenzo, M.T., Ellstrand, N.C., Ross-Ibarra, J., 2013. The Genomic Signature of Crop-Wild Introgression in Maize. *PLOS Genetics* 9, e1003477. <https://doi.org/10.1371/journal.pgen.1003477>
- Isshiki, M., Naka, I., Watanabe, Y., Nishida, N., Kimura, R., Furusawa, T., Natsuhara, K., Yamauchi, T., Nakazawa, M., Ishida, T., Eddie, R., Ohtsuka, R., Ohashi, J., 2020. Admixture and natural selection shaped genomes of an Austronesian-speaking

- population in the Solomon Islands. *Sci Rep* 10, 6872. <https://doi.org/10.1038/s41598-020-62866-3>
- Jensen, J.D., 2014. On the unfounded enthusiasm for soft selective sweeps. *Nat Commun* 5, 5281. <https://doi.org/10.1038/ncomms6281>
- Jensen, Jeffrey D., Kim, Y., DuMont, V.B., Aquadro, C.F., Bustamante, C.D., 2005. Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics* 170, 1401–1410. <https://doi.org/10.1534/genetics.104.038224>
- Jensen, Jeffrey D., Kim, Y., DuMont, V.B., Aquadro, C.F., Bustamante, C.D., 2005. Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics* 170, 1401–1410. <https://doi.org/10.1534/genetics.104.038224>
- Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D.B., Pritchard, J.K., Beall, C.M., Di Rienzo, A., 2014. Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* 5, 3281. <https://doi.org/10.1038/ncomms4281>
- Jones, M.R., Mills, L.S., Alves, P.C., Callahan, C.M., Alves, J.M., Lafferty, D.J.R., Jiggins, F.M., Jensen, J.D., Melo-Ferreira, J., Good, J.M., 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*. <https://doi.org/10.1126/science.aar5273>
- Juric, I., Aeschbacher, S., Coop, G., 2016. The Strength of Selection against Neanderthal Introgression. *PLOS Genetics* 12, e1006340. <https://doi.org/10.1371/journal.pgen.1006340>
- Kern, A.D., Hahn, M.W., 2018. The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution* 35, 1366–1371. <https://doi.org/10.1093/molbev/msy092>
- Kerner, G., Laval, G., Patin, E., Boisson-Dupuis, S., Abel, L., Casanova, J.-L., Quintana-Murci, L., 2021. Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years. *The American Journal of Human Genetics* 108, 517–524. <https://doi.org/10.1016/j.ajhg.2021.02.009>
- Kim, S.-C., Rieseberg, L.H., 1999. Genetic Architecture of Species Differences in Annual Sunflowers: Implications for Adaptive Trait Introgression. *Genetics* 153, 965–977. <https://doi.org/10.1093/genetics/153.2.965>
- Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature* 217, 624–626. <https://doi.org/10.1038/217624a0>
- Kimura, M., Ohta, T., 1971. Protein Polymorphism as a Phase of Molecular Evolution. *Nature* 229, 467–469. <https://doi.org/10.1038/229467a0>
- Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C.A., Froment, A., Nyambo, T.B., Omar, S.A., Wambebe, C., Ranciaro, A., Hirbo, J.B., Tishkoff, S.A., 2013. Identifying Darwinian Selection Acting on Different Human APOL1 Variants among Diverse African Populations. *The American Journal of Human Genetics* 93, 54–66. <https://doi.org/10.1016/j.ajhg.2013.05.014>
- Korunes, K.L., Goldberg, A., 2021. Human genetic admixture. *PLOS Genetics* 17, e1009374. <https://doi.org/10.1371/journal.pgen.1009374>
- Laso-Jadart, R., Harmant, C., Quach, H., Zidane, N., Tyler-Smith, C., Mehdi, Q., Ayub, Q., Quintana-Murci, L., Patin, E., 2017. The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan. *The American Journal of Human Genetics* 101, 977–984. <https://doi.org/10.1016/j.ajhg.2017.09.025>
- Laval, G., Patin, E., Boutillier, P., Quintana-Murci, L., 2021. Sporadic occurrence of recent selective sweeps from standing variation in humans as revealed by an approximate

- Bayesian computation approach. *Genetics* 219, iyab161.
<https://doi.org/10.1093/genetics/iyab161>
- Lawson, D.J., Hellenthal, G., Myers, S., Falush, D., 2012. Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics* 8, e1002453.
<https://doi.org/10.1371/journal.pgen.1002453>
- Lawson, D.J., van Dorp, L., Falush, D., 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* 9, 3258.
<https://doi.org/10.1038/s41467-018-05257-7>
- Lewontin, R.C., Birch, L.C., 1966. Hybridization as a Source of Variation for Adaptation to New Environments. *Evolution* 20, 315–336. <https://doi.org/10.2307/2406633>
- Lewontin, R.C., Hubby, J.L., 1966. A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF DROSOPHILA PSEUDOOBSCURA. *Genetics* 54, 595–609.
<https://doi.org/10.1093/genetics/54.2.595>
- Librado, P., Gamba, C., Gaunitz, C., Sarkissian, C.D., Pruvost, M., Albrechtsen, A., Fages, A., Khan, N., Schubert, M., Jagannathan, V., Serres-Armero, A., Kuderna, L.F.K., Povolotskaya, I.S., Seguin-Orlando, A., Lepetz, S., Neuditschko, M., Thèves, C., Alquraishi, S., Alfarhan, A.H., Al-Rasheid, K., Rieder, S., Samashev, Z., Francfort, H.-P., Benecke, N., Hofreiter, M., Ludwig, A., Keyser, C., Marques-Bonet, T., Ludes, B., Crubézy, E., Leeb, T., Willerslev, E., Orlando, L., 2017. Ancient genomic changes associated with domestication of the horse. *Science*.
<https://doi.org/10.1126/science.aam5298>
- Limou, S., Nelson, G.W., Kopp, J.B., Winkler, C.A., 2014. APOL1 Kidney Risk Alleles: Population Genetics and Disease Associations. *Advances in Chronic Kidney Disease, Focal Segmental Glomerulosclerosis* 21, 426–433.
<https://doi.org/10.1053/j.ackd.2014.06.005>
- Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J.S., Willerslev, E., DeGiorgio, M., Malhi, R.S., 2016. A time transect of exomes from a Native American population before and after European contact. *Nat Commun* 7, 13175. <https://doi.org/10.1038/ncomms13175>
- Lind-Riehl, J., Gailing, O., 2016. Adaptive variation and introgression of a CONSTANS-like gene in North American red oaks. *Forests* 8, 3. <https://doi.org/10.3390/f8010003>
- Lipson, M., Loh, P.-R., Sankararaman, S., Patterson, N., Berger, B., Reich, D., 2015. Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. *PLOS Genetics* 11, e1005550.
<https://doi.org/10.1371/journal.pgen.1005550>
- Liu, K.J., Steinberg, E., Yozzo, A., Song, Y., Kohn, M.H., Nakhleh, L., 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *PNAS* 112, 196–201.
<https://doi.org/10.1073/pnas.1406298111>
- Liu, Y., Mao, X., Krause, J., Fu, Q., 2021. Insights into human history from the first decade of ancient human genomics. *Science*. <https://doi.org/10.1126/science.abi8202>
- Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., Berger, B., 2013. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics* 193, 1233–1254. <https://doi.org/10.1534/genetics.112.147330>
- Lohmueller, K.E., Bustamante, C.D., Clark, A.G., 2011. Detecting Directional Selection in the Presence of Recent Admixture in African-Americans. *Genetics* 187, 823–835.
<https://doi.org/10.1534/genetics.110.122739>
- Long, J.C., 1991. The genetic structure of admixed populations. *Genetics* 127, 417–428.
<https://doi.org/10.1093/genetics/127.2.417>

- Ma, Y., Wang, J., Hu, Q., Li, J., Sun, Y., Zhang, L., Abbott, R.J., Liu, J., Mao, K., 2019. Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. *Commun Biol* 2, 1–12. <https://doi.org/10.1038/s42003-019-0445-z>
- Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D., 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* 93, 278–288. <https://doi.org/10.1016/j.ajhg.2013.06.020>
- Maruyama, T., Kimura, M., 1980. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *PNAS* 77, 6710–6714. <https://doi.org/10.1073/pnas.77.11.6710>
- Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E.R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J.L., de Castro, J.M.B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M.A.R., Roodenberg, J., Vergès, J.M., Krause, J., Cooper, A., Alt, K.W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., Reich, D., 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. <https://doi.org/10.1038/nature16152>
- Mathieson, S., Mathieson, I., 2018. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution* 35, 2957–2970. <https://doi.org/10.1093/molbev/msy180>
- Mayr, E., 1970. *Populations, Species, and Evolution*. Harvard University Press.
- Mayr, E., 1963. *Animal Species and Evolution*. Harvard University Press.
- Mayr, E., 1942. *Systematics and the origin of species*. Columbia University Press.
- McVean, G., 2009. A Genealogical Interpretation of Principal Components Analysis. *PLOS Genetics* 5, e1000686. <https://doi.org/10.1371/journal.pgen.1000686>
- Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution* 28, 659–669. <https://doi.org/10.1016/j.tree.2013.08.003>
- Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C. de, Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*. <https://doi.org/10.1126/science.1224344>
- Miao, B., Wang, Z., Li, Y., 2017. Genomic Analysis Reveals Hypoxia Adaptation in the Tibetan Mastiff by Introgression of the Gray Wolf from the Tibetan Plateau. *Molecular Biology and Evolution* 34, 734–743. <https://doi.org/10.1093/molbev/msw274>
- Molinaro, L., Marnetto, D., Mondal, M., Ongaro, L., Yelmen, B., Lawson, D.J., Montinaro, F., Pagani, L., 2021. A Chromosome-Painting-Based Pipeline to Infer Local Ancestry under Limited Source Availability. *Genome Biology and Evolution* 13, evab025. <https://doi.org/10.1093/gbe/evab025>
- MONAGHAN, J.L., HULL, P., 1976. Differences in Vegetative Characteristics Among Four Populations of *Senecio vulgaris* L. Possibly Due to Interspecific Hybridization. *Annals of Botany* 40, 125–128. <https://doi.org/10.1093/oxfordjournals.aob.a085103>
- Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., Reich, D., 2011. The History of African Gene Flow into

- Southern Europeans, Levantines, and Jews. *PLOS Genetics* 7, e1001373.
<https://doi.org/10.1371/journal.pgen.1001373>
- Moran, P. a. P., 1958. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* 54, 60–71.
<https://doi.org/10.1017/S0305004100033193>
- Nei, M., 1973. Analysis of Gene Diversity in Subdivided Populations. *PNAS* 70, 3321–3323.
- Nei, M., Li, W.-H., 1973. LINKAGE DISEQUILIBRIUM IN SUBDIVIDED POPULATIONS. *Genetics* 75, 213–219. <https://doi.org/10.1093/genetics/75.1.213>
- Norris, E.T., Rishishwar, L., Chande, A.T., Conley, A.B., Ye, K., Valderrama-Aguirre, A., Jordan, I.K., 2020. Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biology* 21, 29. <https://doi.org/10.1186/s13059-020-1946-2>
- Norris, L.C., Main, B.J., Lee, Y., Collier, T.C., Fofana, A., Cornel, A.J., Lanzaro, G.C., 2015. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *PNAS* 112, 815–820.
<https://doi.org/10.1073/pnas.1418892112>
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., Bustamante, C.D., 2008. Genes mirror geography within Europe. *Nature* 456, 98–101.
<https://doi.org/10.1038/nature07331>
- Ohta, T., 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246, 96–98.
<https://doi.org/10.1038/246096a0>
- Ollivier, M., Tresset, A., Bastian, F., Lagoutte, L., Axelsson, E., Arendt, M.-L., Bălăşescu, A., Marshour, M., Sablin, M.V., Salanova, L., Vigne, J.-D., Hitte, C., Hänni, C., n.d. Amy2B copy number variation reveals starch diet adaptations in ancient European dogs. *Royal Society Open Science* 3, 160449. <https://doi.org/10.1098/rsos.160449>
- Ongaro, L., Mondal, M., Flores, R., Marnetto, D., Molinaro, L., Alarcón-Riquelme, M.E., Moreno-Estrada, A., Mabunda, N., Ventura, M., Tambets, K., Hellenthal, G., Capelli, C., Kivisild, T., Metspalu, M., Pagani, L., Montinaro, F., 2021. Continental-scale genomic analysis suggests shared post-admixture adaptation in the Americas. *Human Molecular Genetics* 30, 2123–2134. <https://doi.org/10.1093/hmg/ddab177>
- Pardo-Diaz, C., Salazar, C., Baxter, S.W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W.O., Jiggins, C.D., 2012. Adaptive Introgression across Species Boundaries in *Heliconius* Butterflies. *PLOS Genetics* 8, e1002752.
<https://doi.org/10.1371/journal.pgen.1002752>
- Parsons, T.J., Olson, S.L., Braun, M.J., 1993. Unidirectional Spread of Secondary Sexual Plumage Traits Across an Avian Hybrid Zone. *Science*.
<https://doi.org/10.1126/science.260.5114.1643>
- Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E., 2008. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* 18, 1814–1828. <https://doi.org/10.1101/gr.076554.108>
- Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., Heyer, E., Massougbedji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J.B., Fernandes, V., Pereira, L., Veen, L.V. der, Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J.-M., Quintana-Murci, L., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356, 543–546.
<https://doi.org/10.1126/science.aal1988>
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient Admixture in Human History. *Genetics* 192, 1065–1093. <https://doi.org/10.1534/genetics.112.145037>

- Pennings, P.S., Hermisson, J., 2006a. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Molecular Biology and Evolution* 23, 1076–1084. <https://doi.org/10.1093/molbev/msj117>
- Pennings, P.S., Hermisson, J., 2006b. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLOS Genetics* 2, e186. <https://doi.org/10.1371/journal.pgen.0020186>
- Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olasso, R., Deleuze, J.-F., Ricaut, F.-X., Rakotoarisoa, J.-A., Radimilahy, C., Stoneking, M., Letellier, T., 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nature Communications* 9, 932. <https://doi.org/10.1038/s41467-018-03342-5>
- Platt, R.N., II, McDew-White, M., Le Clec'h, W., Chevalier, F.D., Allan, F., Emery, A.M., Garba, A., Hamidou, A.A., Ame, S.M., Webster, J.P., Rollinson, D., Webster, B.L., Anderson, T.J.C., 2019. Ancient Hybridization and Adaptive Introgression of an Invadysin Gene in Schistosome Parasites. *Molecular Biology and Evolution* 36, 2127–2142. <https://doi.org/10.1093/molbev/msz154>
- Pool, J.E., Nielsen, R., 2009. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics* 181, 711–719. <https://doi.org/10.1534/genetics.108.098095>
- Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T.C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., Broomandkhoshbacht, N., Cooper, A., Culleton, B.J., Ferraz, T., Ferry, M., Furtwängler, A., Haak, W., Harkins, K., Harper, T.K., Hünemeier, T., Lawson, A.M., Llamas, B., Michel, M., Nelson, E., Oppenheimer, J., Patterson, N., Schiffels, S., Sedig, J., Stewardson, K., Talamo, S., Wang, C.-C., Hublin, J.-J., Hubbe, M., Harvati, K., Nuevo Delaunay, A., Beier, J., Francken, M., Kaulicke, P., Reyes-Centeno, H., Rademaker, K., Trask, W.R., Robinson, M., Gutierrez, S.M., Prufer, K.M., Salazar-García, D.C., Chim, E.N., Müller Plumm Gomes, L., Alves, M.L., Liryo, A., Inglez, M., Oliveira, R.E., Bernardo, D.V., Barioni, A., Wesolowski, V., Scheifler, N.A., Rivera, M.A., Plens, C.R., Messineo, P.G., Figuti, L., Corach, D., Scabuzzo, C., Eggers, S., DeBlasis, P., Reindel, M., Méndez, C., Politis, G., Tomasto-Cagigao, E., Kennett, D.J., Strauss, A., Fehren-Schmitz, L., Krause, J., Reich, D., 2018. Reconstructing the Deep Population History of Central and South America. *Cell* 175, 1185-1197.e22. <https://doi.org/10.1016/j.cell.2018.10.027>
- Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., Myers, S., 2009. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLOS Genetics* 5, e1000519. <https://doi.org/10.1371/journal.pgen.1000519>
- Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959. <https://doi.org/10.1093/genetics/155.2.945>
- Prugnolle, F., Manica, A., Balloux, F., 2005a. Geography predicts neutral genetic diversity of human populations. *Current Biology* 15, R159–R160. <https://doi.org/10.1016/j.cub.2005.02.038>
- Prugnolle, F., Manica, A., Charpentier, M., Guégan, J.F., Guernier, V., Balloux, F., 2005b. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Current Biology* 15, 1022–1027. <https://doi.org/10.1016/j.cub.2005.04.050>
- Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., Deschamps, M., Naffakh, N., Duffy, D., Coen, A., Leroux-Roels, G., Clément, F., Boland, A., Deleuze, J.-F., Kelso, J., Albert, M.L.,

- Quintana-Murci, L., 2016. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* 167, 643–656.e17. <https://doi.org/10.1016/j.cell.2016.09.024>
- Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., Santachiara-Benerecetti, A.S., 1999. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet* 23, 437–441. <https://doi.org/10.1038/70550>
- Racimo, F., Marnetto, D., Huerta-Sánchez, E., 2017. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular Biology and Evolution* 34, 296–317. <https://doi.org/10.1093/molbev/msw216>
- Ramos-Madrugal, J., Smith, B.D., Moreno-Mayar, J.V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M.T.P., Wales, N., 2016. Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Current Biology* 26, 3195–3201. <https://doi.org/10.1016/j.cub.2016.09.036>
- Reed, T.E., 1969. Caucasian Genes in American Negroes. *Science*. <https://doi.org/10.1126/science.165.3895.762>
- Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M., Pääbo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053–1060. <https://doi.org/10.1038/nature09710>
- Reich, D., Thangaraj, K., Patterson, N., Price, A.L., Singh, L., 2009. Reconstructing Indian population history. *Nature* 461, 489–494. <https://doi.org/10.1038/nature08365>
- Rieseberg, L.H., Wendel, J.F., 1993. Introgression and Its Consequences in Plants. *Botany Publication and Papers*.
- Rife, D.C., 1954. Populations of hybrid origin as source material for the detection of linkage. *Am J Hum Genet* 6, 26–33.
- Rishishwar, L., Conley, A.B., Wigington, C.H., Wang, L., Valderrama-Aguirre, A., King Jordan, I., 2015. Ancestry, admixture and fitness in Colombian genomes. *Scientific Reports* 5, 12376. <https://doi.org/10.1038/srep12376>
- Rochus, C.M., Tortereau, F., Plisson-Petit, F., Restoux, G., Moreno-Romieux, C., Tossier-Klopp, G., Servin, B., 2018. Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. *BMC Genomics* 19, 71. <https://doi.org/10.1186/s12864-018-4447-x>
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. <https://doi.org/10.1038/nature01140>
- Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., Reich, D., 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354–357. <https://doi.org/10.1038/nature12961>
- Sankararaman, S., Mallick, S., Patterson, N., Reich, D., 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology* 26, 1241–1247. <https://doi.org/10.1016/j.cub.2016.03.037>
- Schrider, D.R., Kern, A.D., 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* 34, 1863–1877. <https://doi.org/10.1093/molbev/msx154>

- Schrider, D.R., Kern, A.D., 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics* 12, e1005928. <https://doi.org/10.1371/journal.pgen.1005928>
- Schweizer, R.M., Durvasula, A., Smith, J., Vohr, S.H., Stahler, D.R., Galaverni, M., Thalmann, O., Smith, D.W., Randi, E., Ostrander, E.A., Green, R.E., Lohmueller, K.E., Novembre, J., Wayne, R.K., 2018. Natural Selection and Origin of a Melanistic Allele in North American Gray Wolves. *Molecular Biology and Evolution* 35, 1190–1209. <https://doi.org/10.1093/molbev/msy031>
- Sella, G., Petrov, D.A., Przeworski, M., Andolfatto, P., 2009. Pervasive Natural Selection in the *Drosophila* Genome? *PLOS Genetics* 5, e1000495. <https://doi.org/10.1371/journal.pgen.1000495>
- Shriner, D., 2013. Overview of Admixture Mapping. *Current Protocols in Human Genetics* 76, 1.23.1-1.23.8. <https://doi.org/10.1002/0471142905.hg0123s76>
- Skoglund, P., Mathieson, I., 2018. Ancient Genomics of Modern Humans: The First Decade. *Annu Rev Genomics Hum Genet* 19, 381–404. <https://doi.org/10.1146/annurev-genom-083117-021749>
- Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. *Science* 236, 787–792. <https://doi.org/10.1126/science.3576198>
- Slatkin, M., 1981. FIXATION PROBABILITIES AND FIXATION TIMES IN A SUBDIVIDED POPULATION. *Evolution* 35, 477–488. <https://doi.org/10.1111/j.1558-5646.1981.tb04911.x>
- Smith, J.M., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 13.
- Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M.W., Kohn, M.H., 2011. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology* 21, 1296–1301. <https://doi.org/10.1016/j.cub.2011.06.043>
- Stoneking, M., 2016. *An Introduction to Molecular Anthropology*. Wiley.
- Stutz, H., Thomas, L., 1964. HYBRIDIZATION AND INTROGRESSION IN COWANIA AND PURSHIA. <https://doi.org/10.1111/j.1558-5646.1964.tb01590.x>
- Suarez-Gonzalez, A., Hefer, C.A., Lexer, C., Douglas, C.J., Cronk, Q.C.B., 2018a. Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P. trichocarpa*. *New Phytologist* 217, 416–427. <https://doi.org/10.1111/nph.14779>
- Suarez-Gonzalez, A., Lexer, C., Cronk, Q.C.B., 2018b. Adaptive introgression: a plant perspective. *Biology Letters* 14, 20170688. <https://doi.org/10.1098/rsbl.2017.0688>
- Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M., Ramachandran, S., 2018. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat Commun* 9, 703. <https://doi.org/10.1038/s41467-018-03100-7>
- Svedberg, J., Shchur, V., Reinman, S., Nielsen, R., Corbett-Detig, R., 2021. Inferring Adaptive Introgression Using Hidden Markov Models. *Molecular Biology and Evolution* 38, 2152–2165. <https://doi.org/10.1093/molbev/msab014>
- Sverrisdóttir, O.Ó., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J.M., Arsuaga Ferreras, J.L., Götherström, A., Thomas, M.G., 2014. Direct Estimates of Natural Selection in Iberia Indicate Calcium Absorption Was Not the Only Driver of Lactase Persistence in Europe. *Molecular Biology and Evolution* 31, 975–983. <https://doi.org/10.1093/molbev/msu049>
- Tajima, F., 1989a. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123, 585–595.
- Tajima, F., 1989b. The Effect of Change in Population Size on DNA Polymorphism. *Genetics* 123, 597–601.

- Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., Risch, N.J., 2007. Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics* 81, 626–633. <https://doi.org/10.1086/520769>
- Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N., 2006. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *Am J Hum Genet* 79, 1–12.
- Toyama, K.S., Crochet, P.-A., Leblois, R., 2020. Sampling schemes and drift can bias admixture proportions inferred by structure. *Molecular Ecology Resources* 20, 1769–1785. <https://doi.org/10.1111/1755-0998.13234>
- Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., Pereira, L., 2015. Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and Evolution* 7, 3484–3495. <https://doi.org/10.1093/gbe/evv236>
- Turner, T.L., Miller, P.M., 2012. Investigating Natural Variation in *Drosophila* Courtship Song by the Evolve and Resequence Approach. *Genetics* 191, 633–642. <https://doi.org/10.1534/genetics.112.139337>
- Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonn  Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L., Oefner, P.J., 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet* 26, 358–361. <https://doi.org/10.1038/81685>
- Valencia-Montoya, W.A., Elfekih, S., North, H.L., Meier, J.I., Warren, I.A., Tay, W.T., Gordon, K.H.J., Specht, A., Paula-Moraes, S.V., Rane, R., Walsh, T.K., Jiggins, C.D., 2020. Adaptive Introgression across Semipermeable Species Boundaries between Local *Helicoverpa zea* and Invasive *Helicoverpa armigera* Moths. *Molecular Biology and Evolution* 37, 2568–2583. <https://doi.org/10.1093/molbev/msaa108>
- Verdu, P., Rosenberg, N.A., 2011. A General Mechanistic Model for Admixture Histories of Hybrid Populations. *Genetics* 189, 1413–1426. <https://doi.org/10.1534/genetics.111.132787>
- Vicente, M., Priehodova, E., Diallo, I., Podgorna, E., Poloni, E.S., ˇCerny, V., Schlebusch, C.M., 2019. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genomics* 20, 915. <https://doi.org/10.1186/s12864-019-6296-7>
- Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A Map of Recent Positive Selection in the Human Genome. *PLOS Biology* 4, e72. <https://doi.org/10.1371/journal.pbio.0040072>
- Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H.H., Mathieson, I., Mathieson, S., 2021. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources* 21, 2689–2705. <https://doi.org/10.1111/1755-0998.13386>
- Weinberg, W., 1908.  ber den Nachweis der Vererbung beim Menschen. *W rttemb.* 64, 369–382.
- Weir, B.S., Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38, 1358–1370. <https://doi.org/10.2307/2408641>
- Workman, P.L., Blumberg, B.S., Cooper, A.J., 1963. Selection, Gene Migration and Polymorphic Stability in a U. S. White and Negro Population. *Am J Hum Genet* 15, 429–437.
- World Health Organization, 2017. World malaria report 2017. Geneva.
- Wright, S., 1950. Genetical Structure of Populations. *Nature* 166, 247–249. <https://doi.org/10.1038/166247a0>
- Wright, S., 1931. Evolution in Mendelian Populations. *Genetics* 16, 97–159.

- Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., Jay, F., 2021a. Creating artificial human genomes using generative neural networks. *PLOS Genetics* 17, e1009303. <https://doi.org/10.1371/journal.pgen.1009303>
- Yelmen, B., Marnetto, D., Molinaro, L., Flores, R., Mondal, M., Pagani, L., 2021b. Improving Selection Detection with Population Branch Statistic on Admixed Populations. *Genome Biology and Evolution* 13, evab039. <https://doi.org/10.1093/gbe/evab039>
- Zeberg, H., Pääbo, S., 2020. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612. <https://doi.org/10.1038/s41586-020-2818-3>
- Zhou, Q., Zhao, L., Guan, Y., 2016. Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics* 12, e1005847. <https://doi.org/10.1371/journal.pgen.1005847>
- Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology* 8, 357–366. [https://doi.org/10.1016/0022-5193\(65\)90083-4](https://doi.org/10.1016/0022-5193(65)90083-4)

Annexes

Genomic insights into population history and biological adaptation in Oceania

<https://doi.org/10.1038/s41586-021-03236-5>

Received: 20 May 2020

Accepted: 13 January 2021

Published online: 14 April 2021

 Check for updates

Jeremy Choin^{1,2,16}, Javier Mendoza-Revilla^{1,16}, Lara R. Arauna^{1,16}, Sebastian Cuadros-Espinoza^{1,3}, Olivier Cassar⁴, Maximilian Larena⁵, Albert Min-Shan Ko⁶, Christine Harmant¹, Romain Laurent⁷, Paul Verdu⁷, Guillaume Laval¹, Anne Boland⁸, Robert Olaso⁸, Jean-François Deleuze⁸, Frédérique Valentin⁹, Ying-Chin Ko¹⁰, Mattias Jakobsson^{5,11}, Antoine Gessain⁴, Laurent Excoffier^{12,13}, Mark Stoneking¹⁴, Etienne Patin^{1,17}✉ & Lluís Quintana-Murci^{1,15,17}✉

The Pacific region is of major importance for addressing questions regarding human dispersals, interactions with archaic hominins and natural selection processes¹. However, the demographic and adaptive history of Oceanian populations remains largely uncharacterized. Here we report high-coverage genomes of 317 individuals from 20 populations from the Pacific region. We find that the ancestors of Papuan-related ('Near Oceanian') groups underwent a strong bottleneck before the settlement of the region, and separated around 20,000–40,000 years ago. We infer that the East Asian ancestors of Pacific populations may have diverged from Taiwanese Indigenous peoples before the Neolithic expansion, which is thought to have started from Taiwan around 5,000 years ago^{2–4}. Additionally, this dispersal was not followed by an immediate, single admixture event with Near Oceanian populations, but involved recurrent episodes of genetic interactions. Our analyses reveal marked differences in the proportion and nature of Denisovan heritage among Pacific groups, suggesting that independent interbreeding with highly structured archaic populations occurred. Furthermore, whereas introgression of Neanderthal genetic information facilitated the adaptation of modern humans related to multiple phenotypes (for example, metabolism, pigmentation and neuronal development), Denisovan introgression was primarily beneficial for immune-related functions. Finally, we report evidence of selective sweeps and polygenic adaptation associated with pathogen exposure and lipid metabolism in the Pacific region, increasing our understanding of the mechanisms of biological adaptation to island environments.

Archaeological data indicate that Near Oceania, which includes New Guinea, the Bismarck archipelago and the Solomon Islands, was peopled around 45 thousand years ago (ka)⁵. The rest of the Pacific—known as Remote Oceania, and including Micronesia, Santa Cruz, Vanuatu, New Caledonia, Fiji and Polynesia—was not settled until around 35 thousand years later. This dispersal, associated with the spread of Austronesian languages and the Lapita cultural complex, is thought to have started in Taiwan around 5 ka, reaching Remote Oceania by about 0.8–3.2 ka⁶. Although genetic studies of Oceanian populations have revealed admixture with populations of East Asian origin^{7–13}, attributed to the Austronesian expansion, questions regarding the peopling history of Oceania remain. It is also unknown how the settlement of the Pacific was accompanied by genetic adaptation to

island environments, and whether archaic introgression facilitated this process in Oceanian individuals, who present the highest levels of combined Neanderthal and Denisovan ancestry worldwide^{14–17}. We report here a whole-genome-based survey that addresses a wide range of questions relating to the demographic and adaptive history of Pacific populations.

Genomic dataset and population structure

We sequenced the genomes of 317 individuals from 20 populations spanning a geographical transect that is thought to underlie the peopling history of Near and Remote Oceania (Fig. 1a and Supplementary Note 1). These high-coverage genomes (around 36×) were

¹Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris, France. ²Université Paris Diderot, Sorbonne Paris Cité, Paris, France. ³Sorbonne Université, Collège doctoral, Paris, France. ⁴Oncogenic Virus Epidemiology and Pathophysiology, Institut Pasteur, UMR 3569, CNRS, Paris, France. ⁵Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden. ⁶Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. ⁷Muséum National d'Histoire Naturelle, UMR7206, CNRS, Université de Paris, Paris, France. ⁸Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France. ⁹Maison de l'Archéologie et de l'Ethnologie, UMR 7041, CNRS, Nanterre, France. ¹⁰Environment-Omics-Disease Research Center, China Medical University and Hospital, Taichung, Taiwan. ¹¹Science for Life Laboratory, Uppsala University, Uppsala, Sweden. ¹²Institute of Ecology and Evolution, University of Bern, Bern, Switzerland. ¹³Swiss Institute of Bioinformatics, Lausanne, Switzerland. ¹⁴Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. ¹⁵Collège de France, Paris, France. ¹⁶These authors contributed equally: Jeremy Choin, Javier Mendoza-Revilla, Lara R. Arauna. ¹⁷These authors jointly supervised this work: Etienne Patin, Lluís Quintana-Murci. ✉e-mail: epatin@pasteur.fr; quintana@pasteur.fr

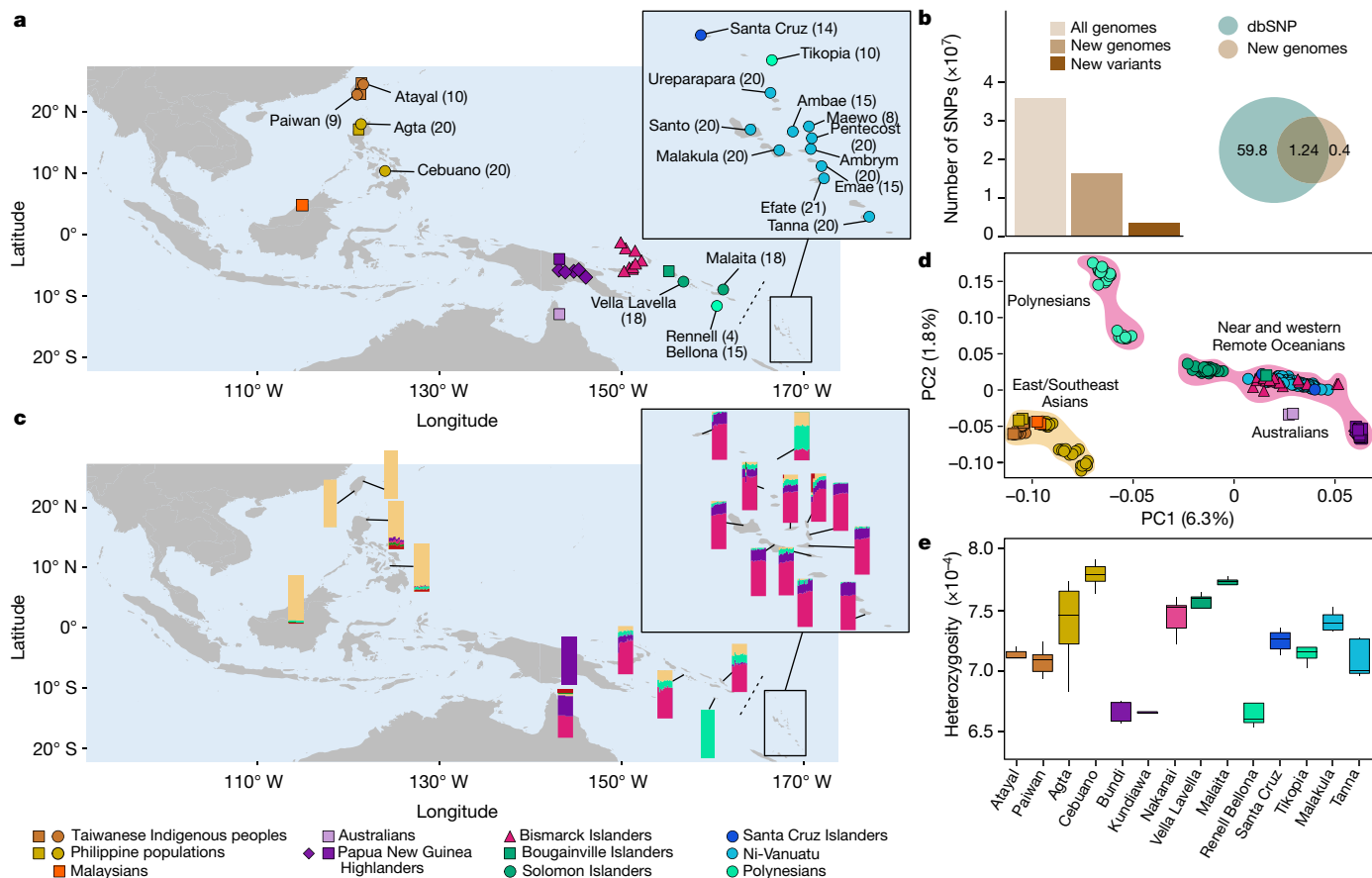


Fig. 1 | Whole-genome variation in Pacific Islanders. **a**, Location of studied populations. The indented map is a magnification of western Remote Oceania. Circles indicate newly generated genomes. Sample sizes are indicated in parentheses. Squares, triangles and diamonds indicate genomes from Mallick et al.¹⁹, Vernot et al.¹⁶ and Malaspina et al.¹⁸, respectively. **b**, The number of SNPs (left), expressed in tens of millions, and comparison with dbSNP (right). New variants are SNPs that are absent from available datasets^{16,18,19} and dbSNP. **c**, ADMIXTURE ancestry proportions at $K = 6$ (lowest cross-validation error; for

all K values, see Extended Data Fig. 1). ADMIXTURE results for Australian populations are discussed in Supplementary Note 3. **d**, PCA of Pacific Islanders and East Asian individuals. The proportion of variance explained is indicated in parentheses. **e**, Population levels of heterozygosity (for all populations, see Supplementary Fig. 9). Population samples were randomly down-sampled to obtain equal sizes ($n = 5$). The line, box, whiskers and points indicate the median, interquartile range, $1.5 \times$ the interquartile range and outliers, respectively. **a**, **c**, Maps were generated using the maps R package⁵¹.

analysed with the genomes of selected populations—including Papua New Guinean Highlanders and Bismarck Islanders^{16,18,19}—and archaic hominins^{20–22} (Supplementary Note 2 and Supplementary Table 1). The final dataset involves 462 unrelated individuals, including 355 individuals from the Pacific region, and 35,870,981 single-nucleotide polymorphisms (SNPs) (Fig. 1b). Using ADMIXTURE, principal component analysis (PCA) and a measure of genetic distance (F_{ST}), we found that population variation is explained by four components, associated with (1) East and Southeast Asian individuals; (2) Papua New Guinean Highlanders; (3) Bismarck Islanders, Solomon Islanders and ni-Vanuatu; and (4) Polynesian outliers (here ‘Polynesian individuals’) (Fig. 1c, d, Extended Data Fig. 1 and Supplementary Note 3). The largest differences are between East and Southeast Asian individuals and Papua New Guinean Highlanders, the remaining populations show various proportions of the two components, supporting the Austronesian expansion model^{8,10,11}. Strong similarities are observed between Bismarck Islanders and ni-Vanuatu, consistent with an expansion from the Bismarck archipelago into Remote Oceania at the end of the Lapita period^{8,10}. Levels of heterozygosity differ markedly among Oceanian populations (Kruskal–Wallis test, $P = 1.4 \times 10^{-12}$) (Fig. 1e), and correlate with individual admixture proportions ($\rho = 0.89$, $P < 2.2 \times 10^{-16}$). The lowest heterozygosity and highest linkage disequilibrium were observed in Papua New Guinean Highlanders and Polynesian individuals, which

probably reflect low effective population sizes. Notably, F -statistics show a higher genetic affinity of ni-Vanuatu from Emäe to Polynesian individuals, relative to other ni-Vanuatu, which suggests gene flow from Polynesia^{6,23}.

The settlement of Near and Remote Oceania

To explore the peopling history of Oceania, we investigated a set of demographic models—driven by several evolutionary hypotheses—with a composite likelihood method²⁴ (Supplementary Note 4). We first determined the relationship between Papua New Guinean Highlanders and other modern and archaic hominins, and replicated previous findings¹⁸ (Extended Data Fig. 2a and Supplementary Table 2). We next investigated the relationship between Near Oceanian groups, assuming a three-epoch demography with gene flow. Observed site frequency spectra were best explained by a strong bottleneck before the settlement of Near Oceania (effective population size (N_e) = 214; 95% confidence interval, 186–276). The separation of Papua New Guinean Highlanders from Bismarck and Solomon Islanders dated back to 39 ka (95% confidence interval, 34–45 ka), and that of Bismarck Islanders from Solomon Islanders to 20 ka (95% confidence interval, 16–30 ka) (Fig. 2a, Supplementary Tables 3, 4), shortly after the human settlement of the region around 30–45 ka^{5,6}.

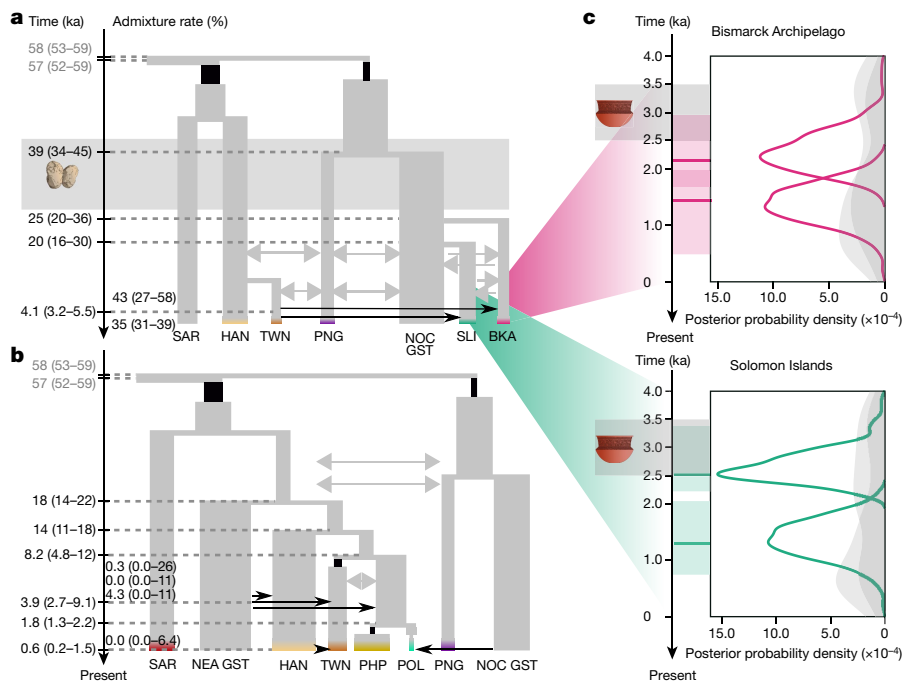


Fig. 2 | Demographic models of the human settlement of the Pacific.

a, Maximum-likelihood model for Near Oceanian populations. Point estimates of parameters and 95% confidence intervals are reported in Supplementary Table 4. The grey area indicates the archaeological period for the settlement of Near Oceania. **b**, Maximum-likelihood model for Formosan-speaking (TWN) and Malayo-Polynesian-speaking (PHP and POL) populations. Point estimates of parameters and 95% confidence intervals are reported in Supplementary Table 7 ('3-pulse model'). **a**, **b**, **BKA**, Bismarck Islanders; **HAN**, Han Chinese individuals; **NEA GST**, a northeast Asian unsampled population; **NOC GST**, a Near Oceanian meta-population; **PHP**, Philippine individuals; **PNG**, Papua New Guinean Highlanders; **POL**, Polynesian individuals from the Solomon Islands; **SAR**, Sardinian individuals; **SLI**, Solomon Islanders; **TWN**, Taiwanese Indigenous peoples. Rectangle width indicates the estimated effective population size. Black rectangles indicate bottlenecks. One- and

two-directional arrows indicate asymmetric and symmetric gene flow, respectively; grey and black arrows indicate continuous and single-pulse gene flow, respectively. The 95% confidence intervals are indicated in parentheses. We assumed a mutation rate of 1.25×10^{-8} mutations per generation per site and a generation time of 29 years. We limited the number of parameter estimations by making simplifying assumptions concerning the recent demography of East-Asian-related and Near Oceanian populations in **a** and **b**, respectively (Supplementary Note 4). Sample sizes are reported in Supplementary Note 4. **c**, Posterior (coloured lines) and prior (grey areas) distributions for the times of admixture between Near Oceanian and East-Asian-related populations, under the double-pulse most-probable model, obtained by ABC (Supplementary Notes 5, 6). Point estimates and 95% credible intervals are indicated by horizontal lines and rectangles, respectively. The grey rectangle indicates the archaeological period of the Lapita cultural complex in Near Oceania²⁷.

We then incorporated western Remote Oceanian populations into the model, represented by ni-Vanuatu individuals from Malakula. We estimated that the ancestors of ni-Vanuatu individuals received migrants from the Bismarck that contributed more than 31% of their gene pool (95% confidence interval, 31–48%) less than 3 ka (Extended Data Fig. 2b and Supplementary Table 5), which is consistent with ancient DNA results^{8–10}. However, the best-fitted model revealed that the Papuan-related population who entered Vanuatu less than 3 ka was a mixture of other Near Oceanian sources^{8,23}: the Papuan-related ancestors of ni-Vanuatu diverged from Papua New Guinean Highlanders and later received approximately 24% (95% confidence interval, 14–41%) of Solomon Islander-related lineages. Interestingly, we found a minimal (<3%) direct contribution of Taiwanese Indigenous peoples to ni-Vanuatu individuals, dating back to around 2.7 ka (95% confidence interval, 1.1–7.5 ka). This suggests that the East-Asian-related ancestry of modern western Remote Oceanian populations has mainly been inherited from admixed Near Oceanian individuals.

Insights into the Austronesian expansion

We characterized the origin of the East Asian ancestry in Oceanian populations by incorporating Philippine and Polynesian Austronesian speakers into our models (Supplementary Note 4). Assuming isolation with migration, we estimated that Taiwanese Indigenous peoples and Malayo-Polynesian speakers (Philippine Kankanaey and Polynesian

individuals from the Solomon Islands) diverged around 7.3 ka (95% confidence interval, 6.4–11 ka) (Extended Data Fig. 2c), in agreement with a recent genetic study of Philippine populations²⁵. Similar estimates were obtained when modelling other Austronesian-speaking groups (>8 ka) (Supplementary Table 6). These dates are at odds with the out-of-Taiwan model—that is, a dispersal event starting from Taiwan around 4.8 ka that brought agriculture and Austronesian languages to Oceania^{2–4}. However, unmodelled gene flow from northeast Asian populations into Austronesian-speaking groups²⁶ could bias parameter estimation. When accounting for such gene flow, we obtained consistently older divergence times than expected under the out-of-Taiwan model⁴, but with overlapping confidence intervals (approximately 8.2 ka; 95% confidence interval, 4.8–12 ka) (Fig. 2b and Supplementary Tables 7–9). Although this suggests that the ancestors of Austronesian speakers separated before the Taiwanese Neolithic², given the uncertainty in parameter estimation, further investigation is needed using ancient genomes.

We next estimated the time of admixture between Near Oceanian individuals and populations of East Asian origin under various admixture models, using an approximate Bayesian computation (ABC) approach (Supplementary Notes 5, 6 and Supplementary Table 10). We found that a two-pulse model best matched the summary statistics for Bismarck and Solomon Islanders. The oldest pulse occurred after the Lapita emergence in the region around 3.5 ka²⁷ (2.2 ka (95% credible interval, 1.7–3.0) and 2.5 ka (95% credible interval, 2.2–3.4) for Bismarck

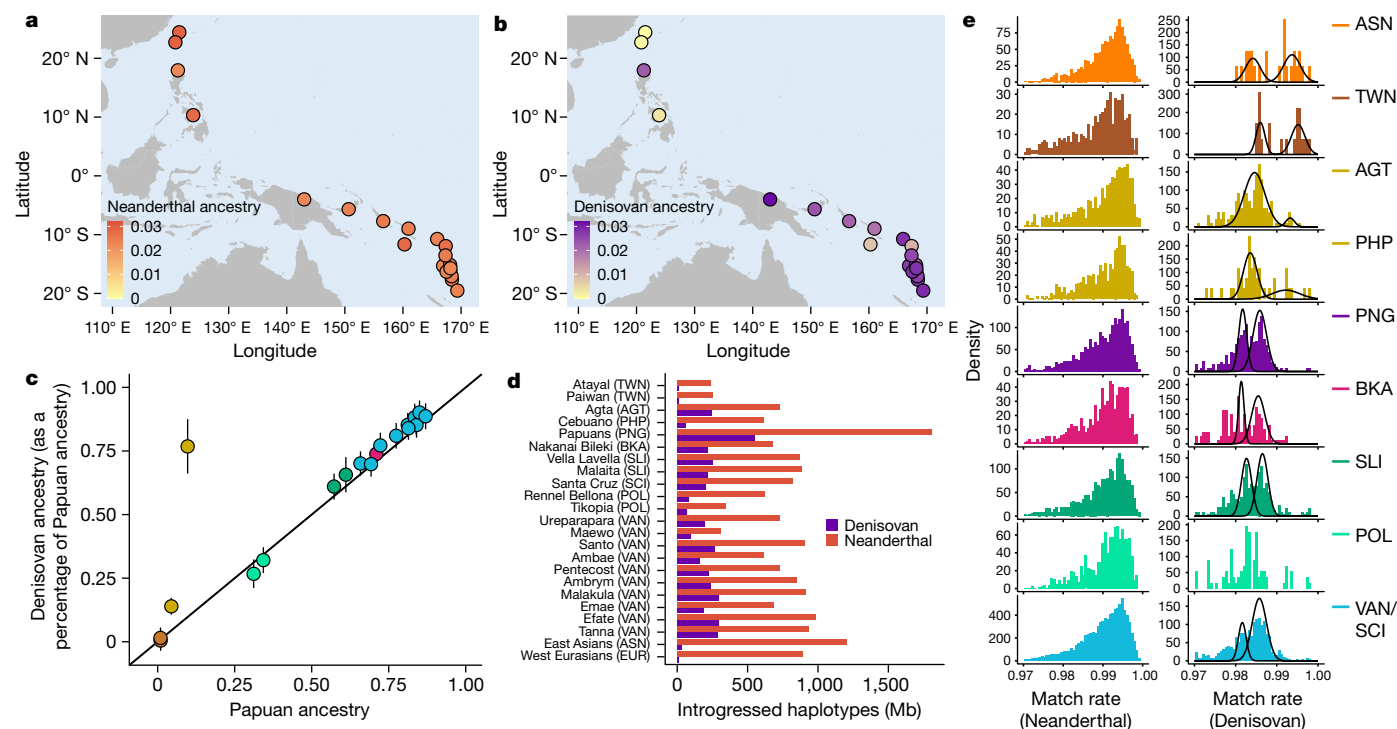


Fig. 3 | Neanderthal and Denisovan introgression across the Pacific. **a, b**, Estimates of Neanderthal (**a**) and Denisovan (**b**) ancestry on the basis of f_4 -ratio statistics. Maps were generated using the maps R package³¹. **c**, Correlation between Papuan ancestry and Denisovan ancestry (as a percentage of Papuan ancestry; $n = 20$ populations). The black line is the identity line. Bars denote 2 s.e. of the estimate. **d**, Cumulative length of the high-confidence archaic haplotypes retrieved in Pacific, East Asian and west Eurasian populations. **e**, Match rate to the Vindija Neanderthal (left) and Altai

Denisovan (right) genomes, based on long (>2,000 sites), high-confidence archaic haplotypes, to remove false-positive values attributable to incomplete lineage sorting. Fitted density curves for populations with significant bimodal match rate distributions are shown. AGT, Philippine Agta; ASN, East Asian individuals (Simons Genome Diversity Project samples only¹⁹); EUR, western Eurasian individuals; SCI, Santa Cruz Islanders; VAN, ni-Vanuatu. The remaining acronyms are as in Fig. 2. Population sample sizes are reported in Supplementary Table 1.

and Solomon Islanders, respectively) (Fig. 2c). This reveals that the separation of Malayo-Polynesian peoples from Taiwanese Indigenous peoples was not followed by an immediate, single admixture episode with Near Oceanian populations, suggesting that Austronesian speakers went through a maturation phase during their dispersal.

Neanderthal and Denisovan heritage

Pacific Islanders have substantial Neanderthal and Denisovan ancestry, as indicated by PCA, D -statistics and f_4 -ratio statistics (Supplementary Note 7). Whereas Neanderthal ancestry is homogeneously distributed (around 2.2–2.9%), Denisovan ancestry differs markedly between groups (approximately 0–3.2%) and is highly correlated with Papuan-related ancestry^{14,15} ($R^2 = 0.77, P < 2.1 \times 10^{-7}$) (Fig. 3a–c). A notable exception is the Philippine Agta (who self-identify as ‘Negritos’) and, to a lesser extent, the Cebuano, who have high Denisovan but little Papuan-related ancestry ($R^2 = 0.99, P < 2.2 \times 10^{-16}$, after excluding Agta and Cebuano).

To explore the sources of archaic ancestry, we inferred high-confidence introgressed haplotypes (Fig. 3d and Supplementary Note 8) and estimated haplotype match rates to the Vindija Neanderthal and Altai Denisovan genomes. Neanderthal match rates were unimodal in all groups (Fig. 3e) and Neanderthal segments significantly overlapped between population pairs (permutation-based $P = 1 \times 10^{-4}$) (Supplementary Notes 9–11), which is consistent with a unique introgression event in the ancestors of non-African populations from a single Neanderthal population. Conversely, different peaks were apparent for Denisovan-introgressed segments (Fig. 3e and Extended Data Fig. 3). A two-peak signal was not only detected in East Asian individuals (around 98.6% and about 99.4% match rate to the Denisovan

genome) as previously reported²⁸, but was also found in Taiwanese Indigenous peoples, Philippine Cebuano and Polynesian individuals. Haplotypes with a match of approximately 99.4% were significantly longer than those with a match of approximately 98.6% (one-tailed Mann–Whitney U -test; $P = 5.14 \times 10^{-4}$), suggesting that—in East Asian populations—introgression from a population closely related to the Altai Denisovan occurred more recently than introgression from the more-distant archaic group.

We also observed two Denisovan peaks in Papuan-related populations²⁹ (Gaussian mixture model $P < 1.68 \times 10^{-4}$) (Supplementary Table 11), with match rates of around 98.2% and 98.6% (Fig. 3e). Consistently, we confirmed using ABC that Papua New Guinean Highlanders received two distinct pulses (posterior probability = 99%) (Supplementary Note 12). Haplotypes with an approximately 98.6% match were of similar length in all populations (Kruskal–Wallis test, $P > 0.05$), whereas haplotypes with a match of around 98.2% were significantly longer in Papuan-related populations than those with a match of about 98.6% in other populations (Supplementary Note 10). ABC parameter inference supported a first pulse around 46 ka (95% credible interval, 39–56 ka), from a lineage that diverged 222 ka from the Altai Denisovan (95% credible interval, 174–263 ka) (Supplementary Note 12 and Supplementary Table 12) and a second pulse into Papuan-related populations around 25 ka (95% confidence interval, 15–35 ka) from a lineage that separated 409 ka from the Altai Denisovan (95% credible interval, 335–497 ka). This model was more-supported than a previously reported model in which the pulse from distantly related Denisovans occurred around 46 ka²⁹ (ABC posterior probability = 99%) (Supplementary Note 12). Our results document multiple interactions of Denisovans with the ancestors of Papuan-related groups and a deep structure of introgressing archaic humans.

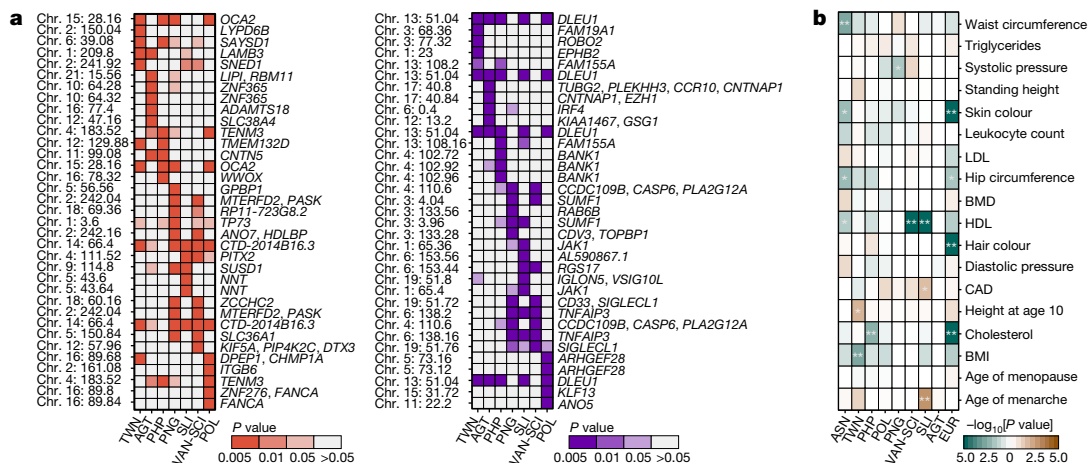


Fig. 4 | Mechanisms of genetic adaptation to Pacific environments.

a, Genomic regions showing the strongest evidence of adaptive introgression from Neanderthals (red) and Denisovans (purple). Each row is a 40-kb window, each column is a Pacific population group, and each cell is coloured according to whether the window is in the top 0.5%, 1%, 5%, >5% of the empirical distributions of the adaptive introgression Q_{95} and U -statistics (Supplementary Note 14). The starting position and genes of each genomic window are indicated. Only the five most extreme windows are shown for each population group. All results are reported in Supplementary Note 14 and Supplementary Tables 14, 15.

For the Philippine Agta, we also observed two Denisovan-related peaks, with match rates of around 98.6% and 99.4% (Fig. 3e). We found that the 99.4% peak is probably due to gene flow from East Asian populations (Supplementary Note 10). Introgressed haplotypes in the Agta overlap significantly with those in Papuan-related populations (Supplementary Note 11), but their high Papuan-independent Denisovan ancestry (Fig. 3c) suggests additional interbreeding. This, together with the discovery of *Homo luzonensis* in the Philippines³⁰, prompted us to search for introgression from other archaic hominins. Using the S' method²⁸, and filtering Neanderthal and Denisovan haplotypes, we retained 59 archaic haplotypes spanning a total of 4.99 megabases (Mb), around 50% of which were common to most groups (Extended Data Fig. 4 and Supplementary Note 13). Focusing on the Agta and Cebuano, we retained only around 1 Mb of introgressed haplotypes that were private to these groups. This suggests that *Homo luzonensis* made little or no contribution to the genetic make-up of modern humans or that this hominin was closely related to Neanderthals or Denisovans.

The adaptive nature of archaic introgression

Although evidence of archaic adaptive introgression exists^{31,32}, few studies have evaluated its role in Oceanian populations. We first tested 5,603 biological pathways for enrichment in adaptive introgression signals (Supplementary Notes 14, 15). For Neanderthal and Denisovan segments, a significant enrichment was observed for 24 and 15 pathways, respectively, of which 9 were related to metabolic and immune functions (Supplementary Tables 13–18). Focusing on Neanderthal adaptive introgression, we replicated genes such as *OCA2*, *CHMP1A* or *LYPD6B*^{31,32} (Fig. 4a). We also identified previously unreported signals in genes relating to immunity (*CNTN5*, *IL10RA*, *TIAMI* and *PRSS57*), neuronal development (*TENM3*, *UNC13C*, *SEMA3F* and *MCPH1*), metabolism (*LIPI*, *ZNF444*, *TBC1D1*, *GPBP1*, *PASK*, *SVEP1*, *OSBPL10* and *HDLBP*) and dermatological or pigmentation phenotypes (*LAMB3*, *TMEM132D*, *PTCHI*, *SLC36A1*, *KRT80*, *FANCA* and *DBNDD1*) (Extended Data Fig. 5), further supporting the notion that Neanderthal variants, beneficial or not, have influenced numerous human phenotypes^{31–33}.

For Denisovans, we replicated signals for immune-related (*TNFAIP3*, *SAMSNI*, *ROBO2* and *PELI2*)^{29,31} and metabolism-related (*DLEU1*, *WARS2*

CCDC109B is also known as *MCUB*, *KIAA1467* is also known as *FAM234B*, *FAM19A1* is also known as *TAF1*, *MTERFD2* is also known as *MTERF4*, *RPI1-723G8.2* is also known as *LINCOI899*. **b**, Signals of polygenic adaptation. Blue and brown colours indicate the $-\log_{10}(P \text{ value})$ for a significant decrease (trait $iHS > 0$) or increase (trait $iHS < 0$) in the candidate trait. * $P < 0.025$; ** $P < 0.005$. BMD, heel-bone mineral density; BMI, body mass index; CAD, coronary atherosclerosis; HDL high-density lipoprotein levels; LDL, low-density lipoprotein levels. **a, b**, Population acronyms are as in Figs. 2, 3.

and *SUMF1*)^{29,32} genes. Our most-extreme candidates comprise 14 previously unreported signals in genes relating to the regulation of innate and adaptive immunity, including *ARHGFE28*, *BANK1*, *CCR10*, *CD33*, *DCC*, *DDX60*, *EPHB2*, *EVI5*, *IGLON5*, *IRF4*, *JAK1*, *LRR8C8* and *LRR8D*, and *VSIG10L* (Fig. 4a and Supplementary Table 15). For example, *CD33*—which mediates cell–cell interactions and keeps immune cells in a resting state³⁴—contains an approximately 30-kb-long haplotype with seven high-frequency, introgressed variants, including an Oceanian-specific nonsynonymous variant (rs367689451-A; derived allele frequency (DAF) > 66%) (Extended Data Fig. 5) predicted to be deleterious (SIFT score = 0). Similarly, *IRF4*—which regulates Toll-like receptor signalling and interferon responses to viral infections³⁵—has an around 29-kb-long haplotype containing 13 high-frequency (DAF > 64%) variants in the Agta. These results suggest that Denisovan introgression has facilitated human adaptation by serving as a reservoir of resistance alleles against pathogens.

Genetic adaptation to island environments

Finally, we searched for signals of classic sweeps and polygenic adaptation in Pacific populations (Supplementary Notes 16–18 and Supplementary Tables 19–25). We found 44 sweep signals common to all Papuan-related groups (empirical $P < 0.01$) (Extended Data Fig. 6), including the *TNFAIP3* gene, which was identified as adaptively introgressed from Denisovans³¹ (Extended Data Fig. 7). The strongest hit (empirical $P < 0.001$) included *GABRP*, which mediates the anticonvulsive effects of endogenous pregnanolone during pregnancy³⁶, and *RANBP17*, which is associated with body mass index and high-density lipoprotein cholesterol³⁷ (Extended Data Fig. 8a, b). The highest score identified a nonsynonymous, probably damaging variant (rs79997355) in *GABRP* at more than 70% frequency in Papua New Guinean Highlanders and ni-Vanuatu, and low frequency (less than 5%) in East and Southeast Asian populations. Among population-specific signals, *ATG7*, which regulates cellular responses to nutrient deprivation³⁸ and is associated with blood pressure³⁹, presented high selection scores in Solomon Islanders.

Among populations with high East Asian ancestry, we identified 29 shared sweep signals ($P < 0.01$) (Extended Data Fig. 9). The highest

scores ($P < 0.001$) overlapped with an approximately 1-Mb haplotype containing multiple genes, including *ALDH2*. *ALDH2* deficiency results in adverse reactions to alcohol and is associated with increased survival in Japanese individuals⁴⁰. The *ALDH2* rs3809276 variant occurs in more than 60% and less than 15% in East-Asian-related and Papuan-related groups, respectively. We also detected a strong signal around *OSBPL10*, associated with dyslipidaemia and triglyceride levels⁴¹ and protection against dengue⁴², which we found to have been adaptively introgressed from Neanderthals (Extended Data Fig. 7). Population-specific signals included *LHFPL2* in Polynesian individuals (Extended Data Fig. 8c, d), variation in which is associated with eye macula thickness—a highly variable trait involved in sharp vision⁴³. *LHFPL2* variants reach around 80% frequency in Polynesian individuals, but are absent from databases, highlighting the need to characterize genomic variation in understudied populations.

Because most adaptive traits are expected to be polygenic⁴⁴, we tested for directional selection of 25 complex traits with a well-studied genetic architecture⁴⁵, by comparing the integrated haplotype scores (iHS) of trait-associated alleles to those of matched, random SNPs⁴⁶. Focusing on European individuals as a control, we found signals of polygenic adaptation for lighter skin and hair pigmentation but not for increased height (Fig. 4b), as previously reported^{46,47}. In Pacific populations, we detected a strong signal for lower levels of high-density lipoprotein cholesterol in Solomon Islanders and ni-Vanuatu ($P = 1 \times 10^{-5}$).

Implications for human history and health

The peopling of Oceania raises questions about the ability of our species to inhabit and adapt to insular environments. Using current estimates of the human mutation rate and generation time¹⁸ (Supplementary Note 4 and Supplementary Tables 2–7), we find that the settlement of Near Oceania 30–45 ka^{5,6} was rapidly followed by genetic isolation between archipelagos, suggesting that navigation during the Pleistocene epoch was possible but limited. Furthermore, our study reveals that genetic interactions between East Asian and Oceanian populations may have been more complex than predicted by the strict out-of-Taiwan model⁴, and suggests that at least two different episodes of admixture occurred in Near Oceania after the emergence of the Lapita culture^{11,27}. Our analyses also provide insights into the settlement of Remote Oceania. Ancient DNA studies have proposed that Papuan-related peoples expanded to Vanuatu shortly after the initial settlement, replacing local Lapita groups^{8,10,23}. We suggest that most East-Asian-related ancestry in modern ni-Vanuatu individuals results from gene flow from admixed Near Oceanian populations, rather than from the early Lapita settlers. These results, combined with evidence of back migrations from Polynesia^{6,10,23}, support a scenario of repeated population movements in the Vanuatu region. Given that we explored a relatively limited number of models, archaeological, morphometric and palaeogenomic studies are required to elucidate the complex peopling history of the region.

The recovery of diverse Denisovan-introgressed material in our dataset, together with previous studies^{28,29}, shows that modern humans received multiple pulses from different Denisovan-related groups (Extended Data Fig. 10). First, we estimate that the East-Asian-specific pulse²⁸, derived from a clade closely related to the Altai Denisovan, occurred around 21 ka. The geographical distribution of haplotypes from this clade indicates that it probably occurred in mainland East Asia. Second, another clade distantly related to Altai Denisovans^{28,29} contributed haplotypes of similar length to Near Oceanian populations, East Asian populations and Philippine Agta. Because our models do not support a recent common origin of Near Oceanian and East Asian populations, we suggest that East Asian populations inherited these archaic segments indirectly, via gene flow from a population ancestral to the Agta and/or Near Oceanian populations. Assuming a pulse into the ancestors of Near Oceanian individuals, we date this introgression to around 46 ka, possibly in Southeast Asia, before migrations to

Sahul. Third, another pulse^{28,29}—which was specific to Papuan-related groups—is derived from a clade more distantly related to Altai Denisovans. We date this introgression to approximately 25 ka, suggesting it occurred in Sundaland or further east. Archaic hominins found east of the Wallace line include *Homo floresiensis* and *Homo luzonensis*^{30,48}, suggesting that either these lineages were related to Altai Denisovans, or Denisovan-related hominins were also present in the region. The recent dates of Denisovan introgression that we detect in East Asian and Papuan populations indicate that these archaic humans may have persisted as late as around 21–25 ka. Finally, the high Denisovan-related ancestry in the Agta^{14,15} suggests that they experienced a different, independent pulse. Collectively, our analyses show that interbreeding between modern humans and highly structured groups of archaic hominins was a common phenomenon in the Asia–Pacific region.

This study reports more than 100,000 undescribed genetic variants in Pacific Islanders at a frequency of more than 1%, some of which are expected to affect phenotype variation. Candidate variants for positive selection are observed in genes relating to immunity and metabolism, which suggests genetic adaptation to pathogens and food sources that are characteristic of Pacific islands. The finding that some of these variants were inherited from Denisovans highlights the importance of archaic introgression as a source of adaptive variation in modern humans^{29,31,32,49}. Finally, the signal of polygenic adaptation related to levels of high-density lipoprotein cholesterol suggests that there are population differences in lipid metabolism, potentially accounting for the contrasting responses to recent dietary changes in the region⁵⁰. Large genomic studies in the Pacific region are required to understand the causal links between past genetic adaptation and present-day disease risk, and to promote the translation of medical genomic research in understudied populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-021-03236-5>.

- Gosling, A. L. & Matisoo-Smith, E. A. The evolutionary history and human settlement of Australia and the Pacific. *Curr. Opin. Genet. Dev.* **53**, 53–59 (2018).
- Hung, H.-C. & Carson, M. T. Foragers, fishers and farmers: origins of the Taiwanese Neolithic. *Antiquity* **88**, 1115–1131 (2014).
- Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
- Bellwood, P. *First Farmers: the Origins of Agricultural Societies* (Blackwell, 2005).
- O’Connell, J. F. et al. When did *Homo sapiens* first reach Southeast Asia and Sahul? *Proc. Natl Acad. Sci. USA* **115**, 8482–8490 (2018).
- Kirch, P. V. *On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact* (Univ. California Press, 2017).
- Wollstein, A. et al. Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992 (2010).
- Lipson, M. et al. Population turnover in Remote Oceania shortly after initial settlement. *Curr. Biol.* **28**, 1157–1165 (2018).
- Skoglund, P. et al. Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).
- Posth, C. et al. Language continuity despite population replacement in Remote Oceania. *Nat. Ecol. Evol.* **2**, 731–740 (2018).
- Pugach, I. et al. The gateway from Near into Remote Oceania: new insights from genome-wide data. *Mol. Biol. Evol.* **35**, 871–886 (2018).
- Bergström, A. et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160–1163 (2017).
- Ioannidis, A. G. et al. Native American gene flow into Polynesia predating Easter Island settlement. *Nature* **583**, 572–577 (2020).
- Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015).
- Reich, D. et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).
- Vernot, B. et al. Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
- Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
- Malaspinas, A. S. et al. A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016).

19. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
20. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
21. Prüfer, K. et al. A high-coverage Neanderthal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
22. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
23. Lipson, M. et al. Three phases of ancient migration shaped the ancestry of human populations in Vanuatu. *Curr. Biol.* **30**, 4846–4856 (2020).
24. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
25. Larena, M. et al. Multiple migrations to the Philippines during the last 50,000 years. *Proc. Natl Acad. Sci. USA*, <https://doi.org/10.1073/pnas.2026132118> (2021).
26. Yang, M. A. et al. Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288 (2020).
27. Rieth, T. M. & Athens, J. S. Late Holocene human expansion into Near and Remote Oceania: a Bayesian model of the chronologies of the Mariana Islands and Bismarck Archipelago. *J. Island Coast. Archaeol.* **14**, 5–16 (2019).
28. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61 (2018).
29. Jacobs, G. S. et al. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell* **177**, 1010–1021 (2019).
30. Détroit, F. et al. A new species of *Homo* from the Late Pleistocene of the Philippines. *Nature* **568**, 181–186 (2019).
31. Gittelman, R. M. et al. Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Curr. Biol.* **26**, 3375–3382 (2016).
32. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
33. Simonti, C. N. et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).
34. Vitale, C. et al. Surface expression and function of p75/AIRM-1 or CD33 in acute myeloid leukemias: engagement of CD33 induces apoptosis of leukemic cells. *Proc. Natl Acad. Sci. USA* **98**, 5764–5769 (2001).
35. Negishi, H. et al. Negative regulation of Toll-like-receptor signaling by IRF-4. *Proc. Natl Acad. Sci. USA* **102**, 15989–15994 (2005).
36. Hedblom, E. & Kirkness, E. F. A novel class of GABA_A receptor subunit in tissues of the reproductive system. *J. Biol. Chem.* **272**, 15346–15350 (1997).
37. Hoffmann, T. J. et al. A large multiethnic genome-wide association study of adult body mass index identifies novel loci. *Genetics* **210**, 499–515 (2018).
38. Lee, I. H. et al. Atg7 modulates p53 activity to regulate cell cycle and survival during metabolic stress. *Science* **336**, 225–228 (2012).
39. Giri, A. et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
40. Sakaue, S. et al. Functional variants in *ADH1B* and *ALDH2* are non-additively associated with all-cause mortality in Japanese population. *Eur. J. Hum. Genet.* **28**, 378–382 (2020).
41. Perttilä, J. et al. *OSBPL10*, a novel candidate gene for high triglyceride trait in dyslipidemic Finnish subjects, regulates cellular lipid metabolism. *J. Mol. Med.* **87**, 825–835 (2009).
42. Sierra, B. et al. *OSBPL10*, *RXRA* and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed Cubans. *PLoS Pathog.* **13**, e1006220 (2017).
43. Gao, X. R., Huang, H. & Kim, H. Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort. *Hum. Mol. Genet.* **28**, 1162–1172 (2019).
44. Sella, G. & Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
45. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
46. Field, Y. et al. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
47. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
48. Brown, P. et al. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055–1061 (2004).
49. Gouy, A. & Excoffier, L. Polygenic patterns of adaptive introgression in modern humans are mainly shaped by response to pathogens. *Mol. Biol. Evol.* **37**, 1420–1433 (2020).
50. Gosling, A. L., Buckley, H. R., Matisoo-Smith, E. & Merriman, T. R. Pacific populations, metabolic disease and 'just-so stories': a critique of the 'thrifty genotype' hypothesis in Oceania. *Ann. Hum. Genet.* **79**, 470–480 (2015).
51. R Core Team. R: A language and environment for statistical computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, 2013).

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Article

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample collection and approvals

Samples were obtained from 317 adult volunteers in Taiwan, the Philippines, the Solomon Islands and Vanuatu from 1998 to 2018. DNA was extracted from blood, saliva or cheek swabs (Supplementary Note 1). Informed consent was obtained from each participant, including consent for genetics research, after the nature and scope of the research was explained in detail. The study received approval from the Institutional Review Board of Institut Pasteur (2016-02/IRB/5), the Ethics Commission of the University of Leipzig Medical Faculty (286-10-04102010), the Ethics Committee of Uppsala University 'Regionala Etikprövningsnämnden Uppsala' (Dnr 2016/103) and from the local authorities, including the China Medical University Hospital Ethics Review Board, the National Commission for Culture and the Arts (NCCA) of the Philippines, the Solomon Islands Ministry of Education and Training and the Vanuatu Ministry of Health (Supplementary Note 1). The consent process, sampling and/or subsequent validation in the Philippines were performed in coordination with the NCCA and, in Cagayan valley region, with local partners or agencies, including Cagayan State University, Quirino State University, Indigenous Cultural Community Councils, Local Government Units and/or regional office of National Commission on Indigenous Peoples. More details about the sampling in the Philippines can be found in ref. ²⁵. Research was conducted in accordance with: (i) ethical principles set forth in the Declaration of Helsinki (version: Fortaleza October 2013), (ii) European directives 2001/20/CE and 2005/28/CE, (iii) principles promulgated in the UNESCO International Declaration on Human Genetic Data and (iv) principles promulgated in the Universal Declaration on the Human Genome and Human Rights.

Whole-genome sequencing data

Whole-genome sequencing was performed on the 317 individual samples (Supplementary Table 1), with the TruSeq DNA PCR-Free or Nano Library Preparation kits (Illumina). After quality control, qualified libraries were sequenced on a HiSeq X5 Illumina platform to obtain paired-end 150-bp reads with an average sequencing depth of 30× per sample. FASTQ files were converted to unmapped BAM files (uBAM), read groups were added and Illumina adapters were tagged with Picard Tools version 2.8.1 (<http://broadinstitute.github.io/picard/>). Read pairs were mapped onto the human reference genome (hs37d5), with the 'mem' algorithm from Burrows–Wheeler Aligner v.0.7.13⁵² and duplicates were marked with Picard Tools. Base quality scores were recalibrated with the Genomic Analysis ToolKit (GATK) software v.3.8⁵³.

Whole-genome data for Bismarck Islanders¹⁶ were processed in the same manner as the newly generated genomes, while for Papua New Guinean Highlanders¹⁸ and other populations of interest¹⁹, raw BAM files were converted into uBAM files, and processed as described above. Variant calling was performed following the GATK best-practice recommendations⁵⁴. All samples were genotyped individually with 'HaplotypeCaller' in gvcf mode. The raw multisample VCF was then generated with the 'GenotypeGVCFs' tool. Using BCFtools v.1.8 (<http://www.htslib.org/>), we applied different hard quality filters on invariant and variant sites, based on coverage depth, genotype quality, Hardy–Weinberg equilibrium and genotype missingness (Supplementary Note 2). The sequencing quality was assessed by several statistics (that is, breadth of coverage 10×, transition/transversion ratio and per-sample missingness) computed with GATK⁵⁴ and BCFtools. Heterozygosity was assessed with PLINK v.1.90^{55,56} and cryptically related samples were

detected with KING v.2.1⁵⁷. Previously unknown SNPs were identified by comparison with available datasets^{16,18,19} and dbSNP⁵⁸.

Genetic structure analyses

PCAs were performed with the 'SmartPCA' algorithm implemented in EIGENSOFT v.6.1.4⁵⁹. The genetic structure was determined with the unsupervised model-based clustering algorithm implemented in ADMIXTURE⁶⁰, which was run—assuming $K=1$ to $K=12$ —100 times with different random seeds. Linkage disequilibrium (r^2) between SNP pairs was estimated with Haploview⁶¹, which was averaged per bin of genetic distance using the 1000 Genomes Project phase 3 genetic map⁶². F_{ST} values were estimated by analysis of molecular variance (AMOVA) as previously described⁶³ (Supplementary Note 3).

Demographic inference

Demographic parameters were estimated with the simulation-based framework implemented in fastsimcoal v.2.6²⁴. We filtered out sites (1) within CpG islands⁶⁴; (2) within genes; and (3) outside of Vindija Neanderthal and Altai Denisovan accessibility masks. These masks exclude sites (1) at which at least 18 out of 35 overlapping 35-mers are mapped elsewhere in the genome with zero or one mismatch; (2) with coverage of less than 10; (3) with mapping quality less than 25; (4) within tandem repeats; (5) within small insertions or deletions; and (6) within coverage filters stratified by GC content. For each demographic model, we performed 600,000 simulations, 65 conditional maximization cycles and 100 replicate runs starting from different random initial values. We limited overfitting by considering only site frequency spectrum (SFS) entries with more than five counts for parameter estimation. We optimized the fit between expected and observed SFS values following a previously described approach^{18,65,66}. Specifically, we first calculated and optimized the likelihood with all of the SFS entries for the first 25 cycles. We then used only polymorphic sites for the remaining 40 cycles. We obtained maximum-likelihood estimates of demographic parameters, by first selecting the 10 runs with the highest likelihoods from the 100 replicate runs. To account for the stochasticity that is inherent to the approximation of the likelihood using coalescent simulations, we re-estimated the likelihood of each of the 10 best runs, using 100 expected SFS obtained using 600,000 simulations. Finally, we re-estimated again the likelihood of the three runs with the highest average, this time using 10^7 simulations, and considered the run with the highest likelihood as the maximum-likelihood run. We corrected for the different numbers of SNPs in the expected and observed SFS, by rescaling parameters by a rescaling factor defined as $S_{\text{obs}}/S_{\text{exp}}$: the N_e and generation times were multiplied by the rescaling factor, whereas migration rates were divided by the rescaling factor. For all inferences, we considered a mutation rate of 1.25×10^{-8} mutations per generation per site^{19,67} and a generation time of 29 years⁶⁸. We also provide estimates of divergence and admixture times assuming a mutation rate of 1.4×10^{-8} mutations per generation per site⁶⁹ (Supplementary Tables 3–7). Model assumptions and parameter search ranges can be found in Supplementary Note 4.

We checked the fit of each best-fit model, by comparing all entries of the observed SFS against simulated entries, averaged over 100 expected SFS obtained with fastsimcoal2²⁴ (Supplementary Note 4). We also compared observed and simulated F_{ST} values, computed with vcfTools v.0.1.13⁷⁰, for all population pairs. We checked that parameter estimates were not affected by background selection and biased gene conversion (Supplementary Note 4). We calculated confidence intervals with a nonparametric block bootstrap approach; we generated 100 bootstrapped datasets by randomly sampling with replacement the same number of 1-Mb blocks of concatenated genomic regions as were present in the observed data. For each bootstrapped dataset, we obtained multi-SFS with Arlequin v.3.5.2.2⁷¹ and re-estimated parameters with the same settings as for the observed dataset, with 20 replicate runs.

Finally, to obtain the 95% confidence intervals, we calculated the 2.5% and 97.5% percentile of the estimate distribution obtained by nonparametric bootstrapping.

For model selection, classical model choice procedures, such as the likelihood ratio tests, could not be used because the likelihood function used in fastsimcoal2²⁴ is a composite likelihood (owing to the presence of linked SNPs in the data). Instead, we compared the likelihoods of the most likely runs between the alternative models, estimated from 600,000 simulations. We also compared the distribution of the \log_{10} (likelihood) of the observed SFS based on 100 expected SFS computed with 10^7 coalescent simulations, using parameters maximizing the likelihood under each scenario. A model was considered the most likely if its mean \log_{10} (likelihood) was 50 units larger than that of the second most likely model⁶⁶. We estimated by simulations that this criterion results in an 81% probability to select the true model (Supplementary Note 4).

We evaluated the accuracy of demographic parameter estimation, using a parametric bootstrap approach. We simulated, with fastsimcoal2²⁴, x 1-Mb DNA loci, with x chosen to obtain the same numbers of segregating SNPs and monomorphic sites as in the observed data, assuming parameters maximizing the likelihood under each model. We then generated 20 simulated SFS by random sampling and used bootstrapped SFS to re-estimate parameters under the same settings as for the original dataset (65 expectation conditional maximization cycles, 600,000 simulations and 100 runs per simulated SFS). We calculated the mean, median and the 2.5% and 97.5% percentiles of the distribution of parameter estimates obtained by parametric bootstrapping, and checked that they included the true (simulated) parameter value.

Admixture models

We applied two ABC approaches⁷² to test for different admixture models for Near Oceanian populations and estimated parameters under the most probable model. Model choice and posterior parameter estimation by ABC are based on summary statistics⁷³. The first approach, developed in the MetHis method⁷⁴, is based on the moments of the distribution of admixture proportions and explicit forward-in-time simulations that follow a general mechanistic admixture model⁷⁵. The second approach uses—as summary statistics—the moments of the distribution of the length of admixture tracts^{76,77}. We assumed three competing models of admixture: a single-pulse, a two-pulse or a constant-recurring model (Supplementary Notes 5, 6). We checked a priori the goodness-of-fit of simulated and observed statistics with the gfit function implemented in the abc R package⁷⁸. Method performance was assessed by estimating the error rates by cross-validation, and by checking a posteriori that the statistics simulated under the most probable model closely fitted the observed statistics.

For the MetHis approach, we simulated 100,000 independent SNPs segregating in the two source populations with fastsimcoal2²⁴, under the refined demographic model for Near Oceanian populations (Fig. 2a). From the foundation of the admixed population to the present generation, the forward-in-time evolution of the 100,000 SNPs in the admixed population was simulated with MetHis⁷⁴, under the classical Wright–Fisher model. For model choice, we conducted 10,000 independent simulations under each of the three competing models. On the basis of 30,000 simulations, we used the random-forest ABC approach⁷⁹ implemented in the abcrf R package. For the best scenario identified, we conducted an additional 20,000 simulations with MetHis. We then used all 30,000 simulations computed under the winning scenario for joint posterior parameter estimation, with the neural-network ABC approach implemented in the abc R package⁷⁸. The performance of the method is described in Supplementary Note 5.

For the approach based on admixture tract length, we performed—under each alternative admixture model—5,000 simulations of 100 5-Mb linked DNA loci with fastsimcoal2²⁴, assuming a variable recombination rate sampled from the 1000 Genomes Project phase 3 genetic map⁶².

We performed 10,000 additional simulations for parameter estimation under the winning model. As summary statistics, we used the mean and variance, across the 100 5-Mb regions, of the mean, minimum and maximum of the distribution of the length of admixture tracts across Near Oceanian populations. The six resulting summary statistics were computed based on local ancestry inference, with RFMix v.1.5.4⁸⁰, which was run with three expectation-maximization steps, a window of 0.03 cM, and Taiwanese Indigenous peoples and Papua New Guinean Highlanders as source populations. The performance of the method is described in Supplementary Note 6. We used the logistic multinomial regression and the neural-network ABC methods implemented in the abc R package⁷⁸ for model choice and parameter estimation, respectively.

Archaic introgression

Before performing archaic introgression analyses, we masked our whole-genome sequencing dataset for regions non-accessible in archaic genomes. We merged the masked dataset with the high-coverage genomes of Vindija and Altai Neanderthals and the Altai Denisovan^{20–22}. We assessed introgression between archaic hominins and modern humans with D -statistics⁸¹. We computed a D -statistic of the form $D(X, \text{West Eurasians/East Asians/Africans}; \text{Neanderthal Vindija, chimpanzee})$ and $D(X, \text{West Eurasians/East Asians/Africans}; \text{Neanderthal Vindija, Denisova Altai})$ to test for introgression from Neanderthal; and D -statistics of the form $D(X, \text{West Eurasians/East Asians}; \text{Denisova Altai, chimpanzee})$ and $D(X, \text{West Eurasians/East Asians}; \text{Denisova Altai, Neanderthal Vindija})$ to test introgression from Denisovans. The last two D -statistics were used to account for the more-recent common ancestor between Neanderthals and Denisovans. We computed f_4 -ratios to estimate the proportion of genome-wide Neanderthal and Denisovan introgression in a modern human population (Supplementary Note 7). All D - and f_4 -ratio statistics were computed with 'qpDstat' and 'qpF4ratio' implemented in ADMIXTOOLS v.5.1.1⁸¹. A weighted-block jackknife procedure dropping 5-cM blocks of the genome in each run was used to compute standard errors.

We used two statistical methods to identify archaic sequences in modern human genomes. The first, S-prime (S'), identifies introgressed sequences without the use of an archaic reference genome²⁸. For the identification of S' introgressed segments in Pacific genomes, we only considered variants with a frequency less than 1% in African individuals from the Simons Genome Diversity Project (SGDP) dataset¹⁹, and segments were detected in each population separately. Genetic distances between sites were estimated from the 1000 Genomes Project phase 3 genetic map⁶². After retrieving empirical S' scores, we estimated a null distribution of S' scores by simulating—with fastsimcoal2²⁴—2,500 10-Mb genomic regions under the best-fitted demographic model for western Remote Oceanian populations (Supplementary Note 4). We fixed all parameters to maximum-likelihood estimates, but removed the simulated introgression pulses from Neanderthals and Denisovans. On the basis of these null distributions of S' scores, we estimated the threshold giving a false-positive rate of less than 0.01, to retain significantly introgressed S' haplotypes (Supplementary Note 8).

The second method, based on conditional random fields (CRF), identifies introgressed archaic haplotypes in phased genomic data, using a reference archaic genome^{17,82}. We phased the data with SHAPEIT2^{83,84}, using 200 conditioning states, 10 burn-in steps and 50 Markov chain Monte Carlo main steps, for a window length of 0.5 cM and an effective population size of 15,000. For the detection of Neanderthal-introgressed haplotypes, we used as reference panels the Vindija Neanderthal genome and SGDP African individuals¹⁹ merged with the Altai Denisovan genome. To detect Denisovan-introgressed haplotypes, we used as reference panel the Altai Denisovan genome and SGDP African individuals¹⁹ merged with the Vindija Neanderthal genome. Results from the two independent runs were analysed jointly to keep those containing alleles with a marginal posterior probability

Article

$P_{\text{Neanderthal}} \geq 0.9$ and $P_{\text{Denisova}} < 0.5$ as Neanderthal-introgressed haplotypes and those containing alleles with $P_{\text{Denisova}} \geq 0.9$ and $P_{\text{Neanderthal}} < 0.5$ as Denisovan-introgressed haplotypes.

We computed a match rate between each detected S' or CRF segment and the Vindija Neanderthal and Altai Denisovan genomes as previously described²⁸ (Supplementary Note 9). We considered that a site matches if the putative introgressed allele is observed in the archaic genome. The match rate was calculated as the number of matches divided by the total number of compared sites. Because longer S' haplotypes carry more information on the archaic origin of introgressed segments, we computed only match rates for S' haplotypes with more than 40 unmasked sites. For the statistical assessment and assignment of introgressed haplotypes to different Denisovan components, we fitted single Gaussian versus two-component Gaussian mixtures to the Denisovan match rate distributions (Supplementary Note 10).

We estimated the sharing of introgressed haplotypes between populations by first retaining S' introgressed haplotypes with a score $>190,000$ and a length of at least 40 kb (Supplementary Note 11). We then classified each haplotype as of either Neanderthal or Denisovan origin, as previously described²⁸. For each haplotype present in a given population, we then estimated the fraction of base-pair overlap with the haplotypes present in a second population, with respect to the length of the segments in the first. As a test statistic, we computed the proportion of segments with a fraction of base-pair overlap greater than 0.5. We assessed significance by performing 10,000 bootstrap iterations, in which we randomly placed introgressed segments with the same number and of the same length as observed along the callable genome (around 2.1 Gb). For each population pairwise comparison, we reported the highest P value of the two. All P values were adjusted for multiple testing with the Benjamini–Hochberg method.

We formally tested for the presence of two distinct Denisovan lineages in Papuan-related populations with an ABC approach⁷², by performing 50,000 independent simulations of 64 DNA sequences of 10 Mb each with fastsimcoal2²⁴. We simulated the demographic model for Near Oceanian populations (Fig. 2a), introducing one or two Denisovan pulses into the Papua New Guinean branch, and a population resize in Papua New Guinea to capture the demographic effect of the agricultural transition¹² (Supplementary Note 12). As summary statistics, we used the moments of the distribution of the S' scores, S' haplotype length and S' match rate to the Altai Denisovan genome. We determined which of the single- and double-pulse introgression models was the most probable, using a logistic multinomial regression algorithm with a tolerance rate set to 5%. We estimated the performance of our ABC model choice by cross-validation. Parameter estimation under the double-pulse winning model was performed on the basis of an additional 150,000 independent simulations, using the neural network algorithm with a tolerance rate set to 5%. We used the same procedure to test whether our two-pulse model, in which the pulse from a more-distant Denisovan lineage occurs later than the other pulse, fits the data better than a previous model in which the pulse from a more-distant Denisovan lineage occurs earlier than the other pulse²⁹. Introgression parameter values were sampled from uniform priors limited by the previously obtained 95% confidence intervals (Supplementary Note 12).

We investigated whether Pacific populations had received gene flow from an unknown archaic hominin, by retaining S' haplotypes unlikely to be of Neanderthal or Denisovan origin, through the removal of Neanderthal and Denisovan haplotypes inferred by the CRF approach (Supplementary Note 13). We characterized these S' haplotypes further by estimating their match rates to the Vindija Neanderthal and Altai Denisovan genomes and retaining only those with a match rate of less than 1% to either of these archaic hominins. The remaining S' haplotypes represent putatively introgressed material from outside the Neanderthal and Denisovan branch.

Adaptive introgression

Candidate regions for adaptive introgression were detected on the basis of the number and derived allele frequency of sites common to modern and archaic humans (Supplementary Note 14), with Q95 and U -statistics³². We computed these statistics in 40-kb non-overlapping windows along the genome of all target populations, using SGP African individuals¹⁹ as the outgroup. We used the chimpanzee reference genome to determine the ancestral or derived states of alleles, removed sites with any missing genotypes, and discarded genomic windows with fewer than five sites. Candidate genomic windows were defined as those with both U and Q95 statistics in the top 0.5% of their respective genome-wide distributions.

We assessed the enrichment of introgressed genes in various biological pathways, including the Kyoto Encyclopedia of Genes and Genomes (KEGG)⁸⁵, Wikipathways⁸⁶, the genome-wide association studies (GWAS) catalogue⁸⁷, Gene Ontology⁸⁸, and manually curated lists of innate immunity genes⁸⁹ and virus-interacting proteins⁹⁰. We merged Pacific populations into three population groups (Supplementary Note 15). We assessed statistical significance using a resampling-based enrichment test that compares the number of introgressed genes in a given gene set to that observed in randomly sampled sets of genes that are matched for different genomic features (that is, recombination rate, PhastCons⁹¹, combined annotation-dependent depletion (CADD) scores⁹², density of DNase I segments⁹³ and number of SNPs). We also determined whether a given gene set was enriched in adaptively introgressed genes, by comparing the number of genes overlapping an adaptively introgressed segment in the gene set with that observed in randomly sampled sets of matched genes. Adaptively introgressed segments were defined as those intersecting with genomic windows with Q95 and U -statistics in the top 5% of their respective genome-wide distributions.

Classic sweeps

For the detection of classic sweep signals, we combined the inter-population locus-specific branch lengths (LSBL)⁹⁴ and cross-population extended haplotype homozygosity (XP-EHH)⁹⁵ statistics into a Fisher's score (F_{CS}). We estimated the F_{CS} as the sum of the $-\log_{10}$ (percentile rank of the statistic for a given SNP) of all statistics, and defined 'outlier SNPs' as those with a F_{CS} among the 1% highest genome-wide. Putatively selected regions were defined as genomic windows with a proportion of outlier SNPs within the 1% highest genome-wide, after partitioning all windows into five bins based on the number of SNPs. The test, reference and outgroup populations used are described in Supplementary Note 16. LSBL and XP-EHH statistics were computed with the optimized, window-based algorithms implemented in selink (<https://github.com/h-e-g/selink>).

Polygenic adaptation

We searched for evidence of polygenic adaptation, using an approach testing whether the mean integrated haplotype score (iHS) of trait-increasing alleles differed significantly from that of random SNPs with a similar allele frequency^{46,96}. We obtained GWAS summary statistics for 25 candidate complex traits from the UK Biobank database⁴⁵, including traits relating to morphology, metabolism and immunity, as these phenotypic traits are strong candidates for responses, through natural selection, to changes in climatic, nutritional and pathogenic environments. We classified SNPs as 'trait-increasing' or 'trait-decreasing' based on UK Biobank effect size (β) estimates. We computed iHS with selink, for each SNP and population, and standardized scores in 100 bins of DAF. We then polarized the iHS, such that positive iHS values indicated directional selection of the trait-decreasing allele, whereas negative iHS values indicated directional selection of the trait-increasing allele. We called the resulting statistic the polarized trait iHS (tiHS).

For each trait, we assessed significance keeping only unlinked trait-associated variants (Supplementary Note 18). We then compared the mean t_iHS of the x independent, trait-associated alleles with the mean t_iHS of 100,000 random samples of x SNPs with similar DAF, genomic evolutionary rate profiling (GERP) score and surrounding recombination rate, to account for the effects of background selection. We considered that directional selection has increased (or decreased) a given trait if less than 2.5% (or 0.5%) of the resampled sets had a mean t_iHS that is lower (or higher) than that observed. We adjusted P values for multiple testing with the Benjamini–Hochberg method. The false-positive rate of the approach at a P value of 2.5% (or 0.5%) was estimated by resampling (Supplementary Note 18).

Because this approach assumes that alleles affecting traits are the same in Oceanian and European populations and that they affect traits in the same direction, we used another approach, which tests for the co-localization of selection signals and trait-associated genomic regions. We partitioned the genome into 100-kb non-overlapping contiguous windows and considered a window to be associated with a trait if at least one SNP within the window was genome-wide significant ($P < 5 \times 10^{-8}$). For each window, we estimated the mean t_iHS for each population. We then tested whether the mean t_iHS of trait-associated windows was greater than that for a null distribution, obtained from 100,000 sets of randomly sampled windows, each set being matched to trait-associated windows in terms of mean GERP score, recombination rate, DAF and number of SNPs.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The whole-genome sequencing dataset generated and analysed in this study is available from the European Genome-Phenome Archive (EGA; <https://www.ebi.ac.uk/ega/>), under accession code EGAS00001004540. Data access and use is restricted to academic research in population genetics, including research on population origins, ancestry and history. The SGDP genome data were retrieved from the EBI European Nucleotide Archive (accession codes PRJEB9586 and ERP010710). The genome data from Malaspinas et al.¹⁸ were retrieved from the EGA (accession code EGAS00001001247). The genome data from Vernot et al.¹⁶ were retrieved from dbGAP (accession code phs001085.v1.p1).

Code availability

Neutrality statistics were computed with the optimized, window-based algorithms implemented in selink (<https://github.com/h-e-g/selink>). All other custom-generated computer codes or algorithms used in this study are available on GitHub (<https://github.com/h-e-g/evocoeania>).

52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
54. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
55. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
57. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
58. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
59. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
60. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
61. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
62. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
63. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
64. Meyer, L. R. et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41**, D64–D69 (2013).
65. de Manuel, M. et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).
66. Sikora, M. et al. The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).
67. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
68. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
69. Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
70. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
71. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
72. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
73. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
74. Fortes-Lima, C. A., Laurent, L., Thouzeau, V., Toupance, B. & Verdu, P. Complex genetic admixture histories reconstructed with approximate Bayesian computations. *Mol. Ecol. Resour.* <https://doi.org/10.1111/1755-0998.13325> (2021).
75. Verdu, P. & Rosenberg, N. A. A general mechanistic model for admixture histories of hybrid populations. *Genetics* **189**, 1413–1426 (2011).
76. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
77. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
78. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
79. Pudlo, P. et al. Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859–866 (2016).
80. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RfMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
81. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
82. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
83. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
84. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
85. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
86. Kutmon, M. et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **44**, D488–D494 (2016).
87. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
88. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
89. Deschamps, M. et al. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* **98**, 5–21 (2016).
90. Enard, D. & Petrov, D. A. Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell* **175**, 360–371 (2018).
91. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
92. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
93. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
94. Shriver, M. D. et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* **1**, 274–286 (2004).
95. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
96. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
97. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).

Acknowledgements We thank all volunteers and Indigenous communities participating in this research; S. Créno and the HPC Core Facility of Institut Pasteur (Paris) for the management of computational resources; F. Mendoza de Leon Jr, NCCA chairperson 2010–2016, for his support; C. Ebeo, O. Casel, K. Pullupul Hagada, D. Guilay, A. Manera and R. Quilang of Cagayan State University, Lahaina Sue Azarcon and Samuel Benigno of Quirino State University and the regional and provincial offices of the National Commission for Indigenous Peoples (NCIP)–Cagayan Valley for their support and assistance. J.C. is supported by the INCEPTION programme ANR-16-CONV-0005 and the Ecole Doctorale FIRE-CRI-Programme Bettencourt and L.R.A. by a Pasteur-Roux-Cantarini fellowship. The CNRGH sequencing platform was

Article

supported by the France Génomique National infrastructure, funded as part of the « Investissements d'Avenir » programme managed by the Agence Nationale pour la Recherche (ANR-10-INBS-09). M.J. is supported by the Knut and Alice Wallenberg foundation. M.S. is supported by the Max Planck Society. The laboratory of L.Q.-M. is supported by the Institut Pasteur, the Collège de France, the CNRS, the Fondation Allianz-Institut de France and the French Government's Investissement d'Avenir programme, Laboratoires d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (ANR-10-LABX-62-IBEID) and 'Milieu Intérieur' (ANR-10-LABX-69-01).

Author contributions E.P. and L.Q.-M. conceived and supervised the project; J.C. led and performed the processing of the genetic data as well as the analyses of population structure and demographic inference; J.M.-R. led and performed the analyses of archaic and adaptive introgression; L.R.A. led and performed the analyses of genetic adaptation; S.C.-E., R.L. and P.V. performed the analyses of admixture models; E.P. coordinated all genetic analyses; O.C., M.L., A.M.-S.K., Y.-C.K., M.J., A.G. and M.S. collaborated with local groups to collect population

samples; C.H., A.B., R.O. and J.-F.D. coordinated and performed sample preparation and sequencing; F.V. provided the archaeological and anthropological context; G.L. and L.E. provided the theoretical and methodological context; J.C., J.M.-R., L.R.A., E.P. and L.Q.-M. wrote the manuscript, with critical input from all authors.

Competing interests The authors declare no competing interests.

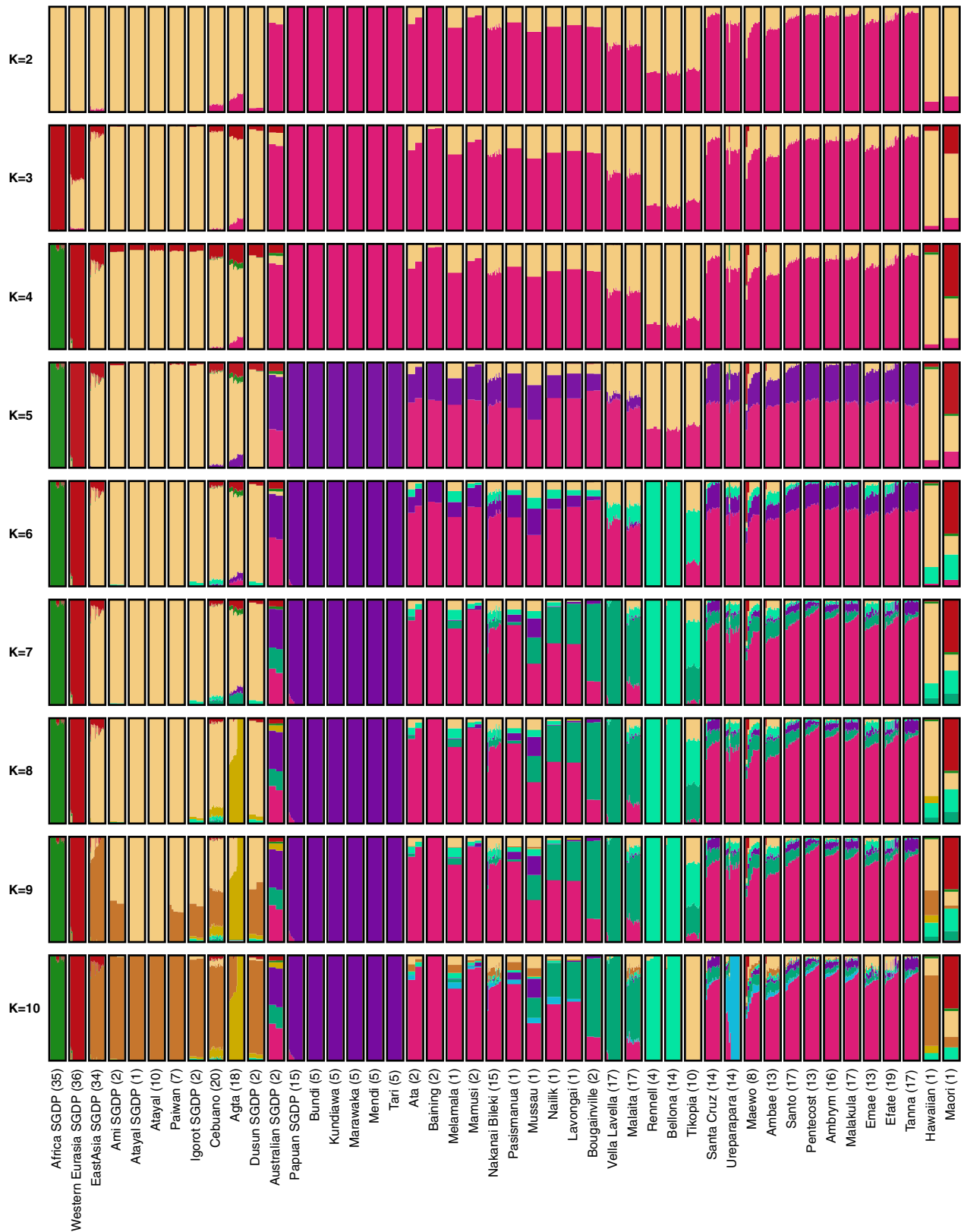
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-021-03236-5>.

Correspondence and requests for materials should be addressed to E.P. or L.Q.-M.

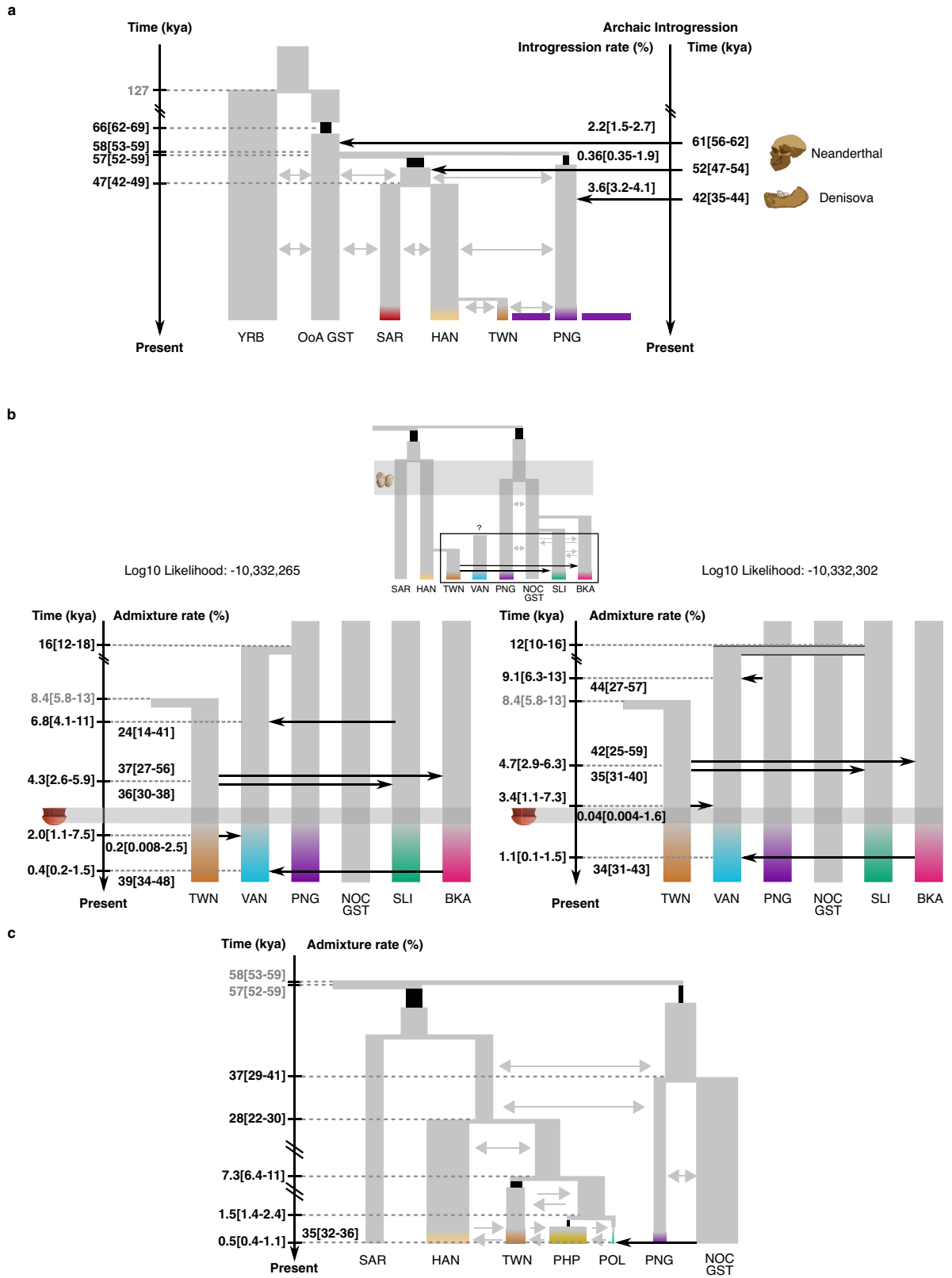
Peer review information *Nature* thanks Patrick Kirch, Cosimo Posth and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Genetic structure of Pacific populations.
 ADMIXTURE ancestry components are shown from $K=2$ (top) to $K=10$ (bottom) for the 462 unrelated individuals. The lowest cross-validation error

was obtained at $K=6$ (Supplementary Fig. 5). Populations are delimited by black borders. Population width is not proportional to population sample size, which is indicated in parentheses.

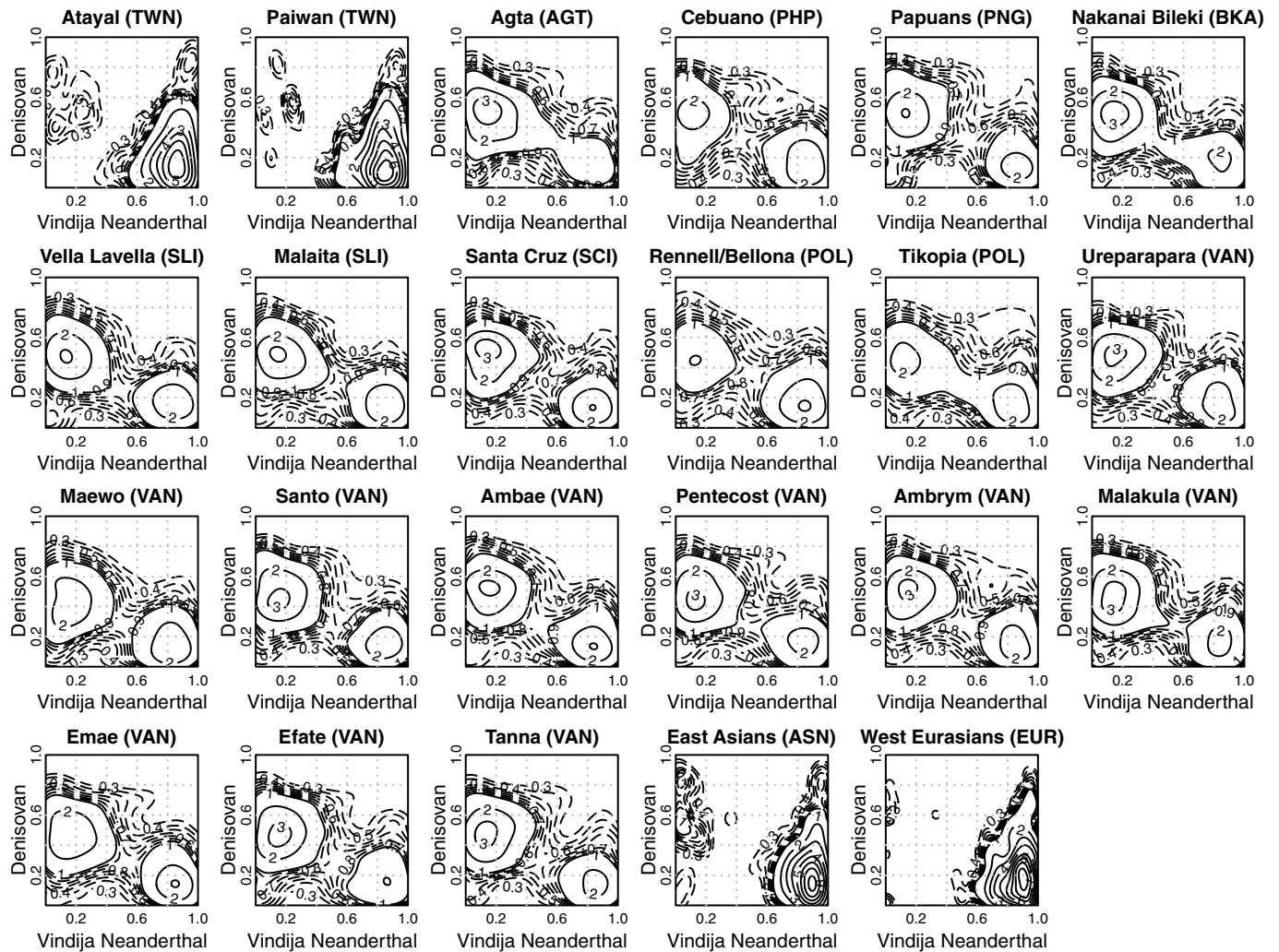


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Demographic models for Pacific populations.

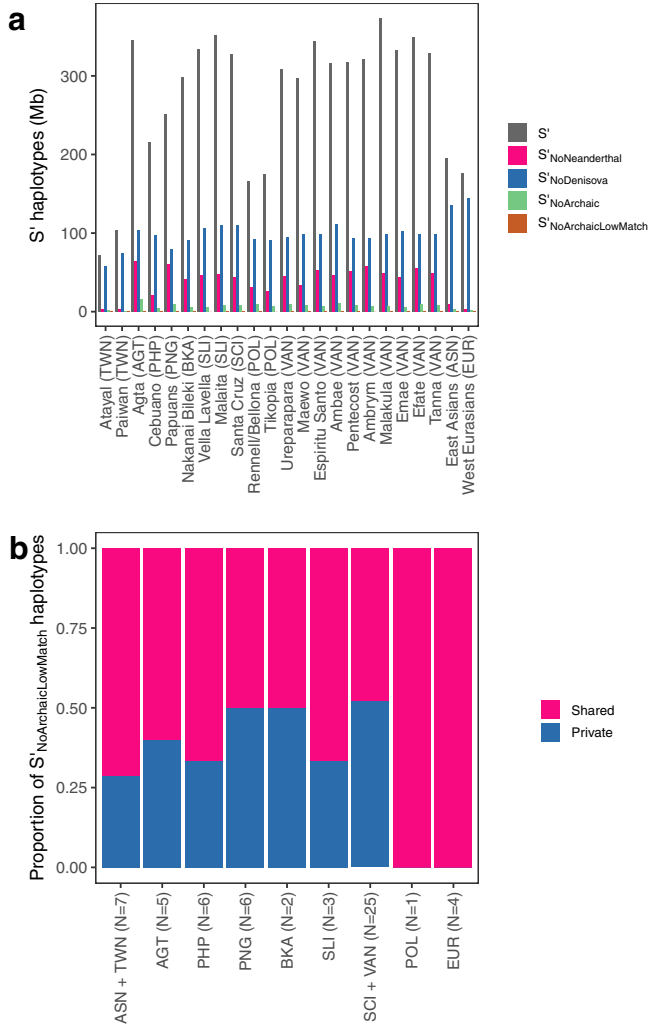
a, Maximum-likelihood demographic model for baseline populations. Point estimates of parameters and 95% confidence intervals are shown in Supplementary Table 2. **b**, Maximum-likelihood demographic models for western Remote Oceanian individuals (VAN). The likelihoods of the two models are not considered to be different. Point estimates of parameters and 95% confidence intervals are shown in Supplementary Table 5. The (VAN, PNG) model (left) assumes that the ni-Vanuatu diverged from Papua New Guinean Highlanders and then received gene flow from Solomon Islanders, Bismarck Islanders and Austronesian-speaking Taiwanese Indigenous peoples. The (VAN, SLI) (right) model assumes that the ni-Vanuatu diverged from the Solomon Islanders and then received gene flow from the other three groups. For the sake of clarity, only Taiwanese Indigenous, Near Oceanian and western Remote Oceanian populations are shown. **c**, Maximum-likelihood model for Austronesian-speaking populations, represented by Taiwanese Indigenous, Philippine Kankanaey and Tikopia Polynesian individuals. BKA, Bismarck Islanders; HAN, Han Chinese individuals (China); NOC GST, a meta-population

of Near Oceanian individuals; OoA GST, an unsampled population to represent the Out-of-Africa exodus; PHP, Philippine individuals; PNG, Papua New Guinean Highlanders; POL, Polynesian individuals from the Solomon Islands; SAR, Sardinian individuals (Italy); SLI, Solomon Islanders; TWN, Taiwanese Indigenous peoples; VAN, ni-Vanuatu; YRB, Yoruba individuals (Nigeria). We assumed a mutation rate of 1.25×10^{-8} mutations per generation per site and a generation time of 29 years. Single-pulse introgression rates are reported as a percentage. The 95% confidence intervals are shown in square brackets. The larger the rectangle width, the larger the estimated effective population size (N_e), except for **b**. Bottlenecks are indicated by black rectangles. Grey and black arrows represent continuous and single pulse gene flow, respectively. One- and two-directional arrows indicate asymmetric and symmetric gene flow, respectively. We limited the number of parameter estimations by making simplifying assumptions regarding the recent demography of East-Asian-related and Near Oceanian populations in **a** and **c**, respectively (Supplementary Note 4). Sample sizes are described in Supplementary Note 4.

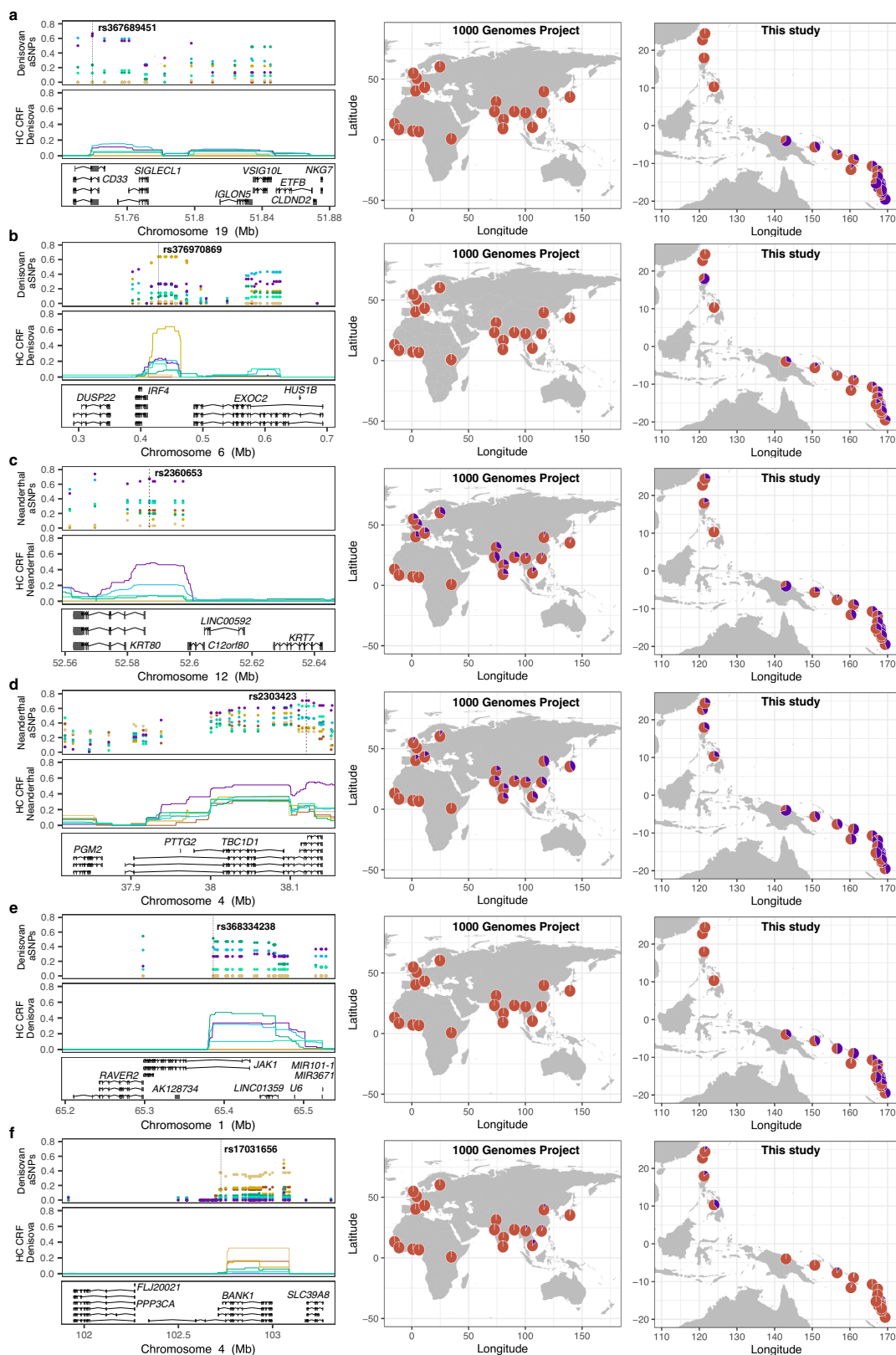


Extended Data Fig. 3 | Match rate of introgressed S' haplotypes in Pacific populations to the Vindija Neanderthal and Altai Denisovan genomes. The match rate is the proportion of putative archaic alleles matching a given archaic genome, excluding sites at masked positions. Only S' haplotypes with

more than 40 sites outside archaic genome masks were included in the analysis. The numbers indicate the height of the density corresponding to each contour line. Contour lines are shown for multiples of 1 (solid lines) and multiples of 0.1 between 0.3 and 0.9 (dashed lines).

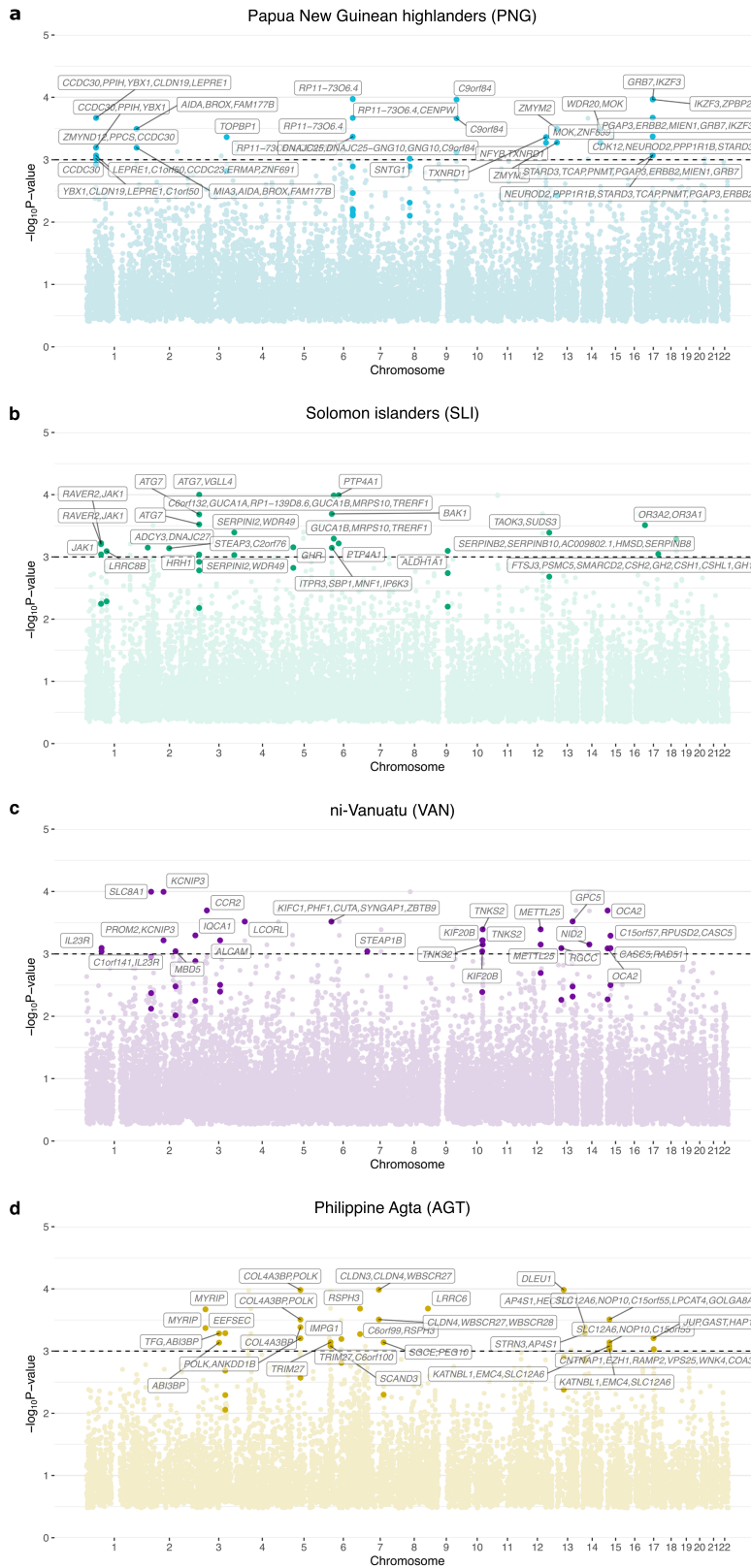


Extended Data Fig. 4 | Detection of introgressed haplotypes from an unknown archaic hominin. a. Cumulative length of S' haplotypes retrieved among modern human populations (S'), after removing Neanderthal CRF haplotypes ($S'_{\text{NoNeanderthal}}$) or Denisovan CRF haplotypes ($S'_{\text{NoDenisova}}$) or both ($S'_{\text{NoArchaic}}$), and removing from the $S'_{\text{NoArchaic}}$ haplotypes those with a match rate higher than 1% to either the Vindija Neanderthal or Altai Denisovan genomes ($S'_{\text{NoArchaicLowMatch}}$). These S' haplotypes are, therefore, putatively introgressed haplotypes from hominins outside of the Neanderthal and Denisovan branch (Supplementary Note 13). **b.** Proportion of $S'_{\text{NoArchaicLowMatch}}$ haplotypes common or private (that is, unique) to populations. Total numbers of $S'_{\text{NoArchaicLowMatch}}$ haplotypes are shown above the population labels.



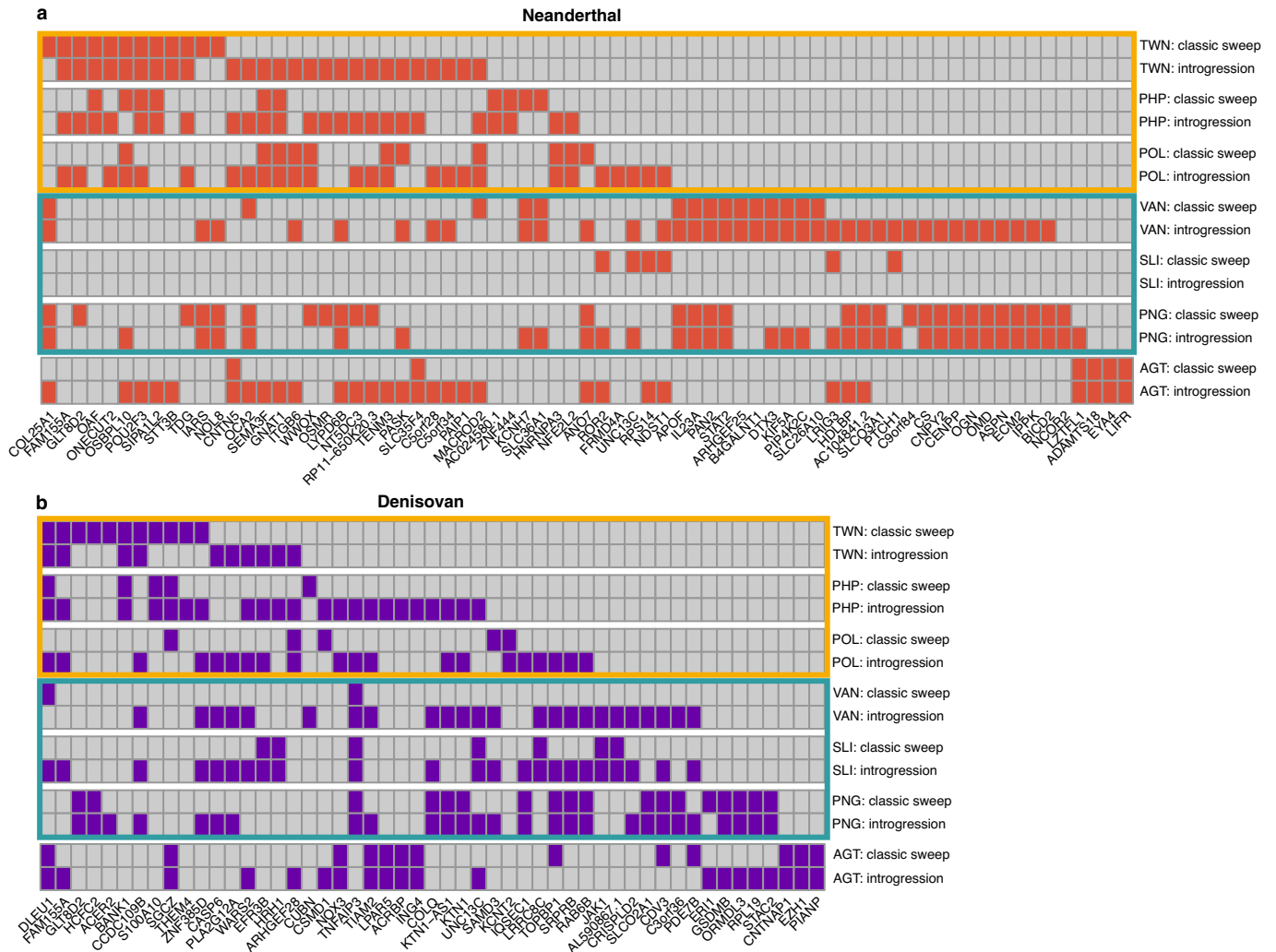
Extended Data Fig. 5 | Examples of candidate loci for adaptive introgression in Pacific populations. **a**, Adaptive introgression of Denisovan origin at the *CD33* locus. **b**, Adaptive introgression of Denisovan origin at the *IRF4* locus. **c**, Adaptive introgression of Neanderthal origin at the *KRT80* locus. **d**, Adaptive introgression of Neanderthal origin at the *TBC1D1* locus. **e**, Adaptive introgression of Denisovan origin at the *AK1* locus. **f**, Adaptive introgression of Denisovan origin at the *BANK1* locus. **a–f**, Left, local Manhattan plot showing the derived allele frequency of archaic SNPs (aSNPs), the

proportion of high-confidence introgressed haplotypes (HC CRF) and the gene isoforms at the locus (in Mb, based on hg19 coordinates). Middle, derived allele frequencies of the top archaic SNP in 1000 Genomes Project phase 3 populations (excluding recently admixed populations). Right, derived allele frequencies of the top archaic SNP in populations from this study. Colours in the left panels indicate populations as in Fig. 1. Pie charts indicate the derived allele frequency in purple, and are centred on the approximate geographical location of each population. Maps were generated using the maps R package³¹.



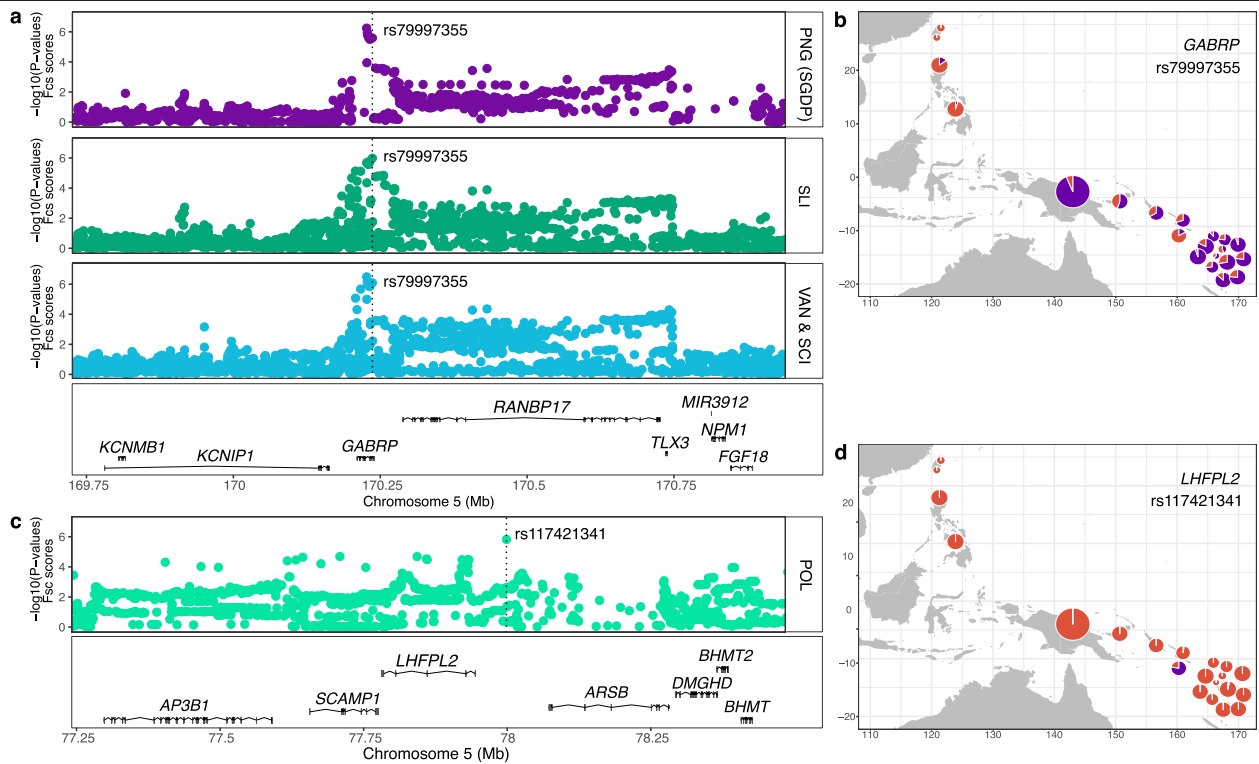
Extended Data Fig. 6 | Classic sweep signals detected in Papuan-related populations. a–d, Manhattan plots of classic sweep signals in Papua New Guinean Highlanders (a), Solomon Islanders (b), ni-Vanuatu (c) and Philippine

Agta (d). a–d, The y axis shows the $-\log_{10}(P)$ value for the number of outlier SNPs per window. Each point is a 100-kb window. The names of genes associated with windows with significant sweep signals are shown.



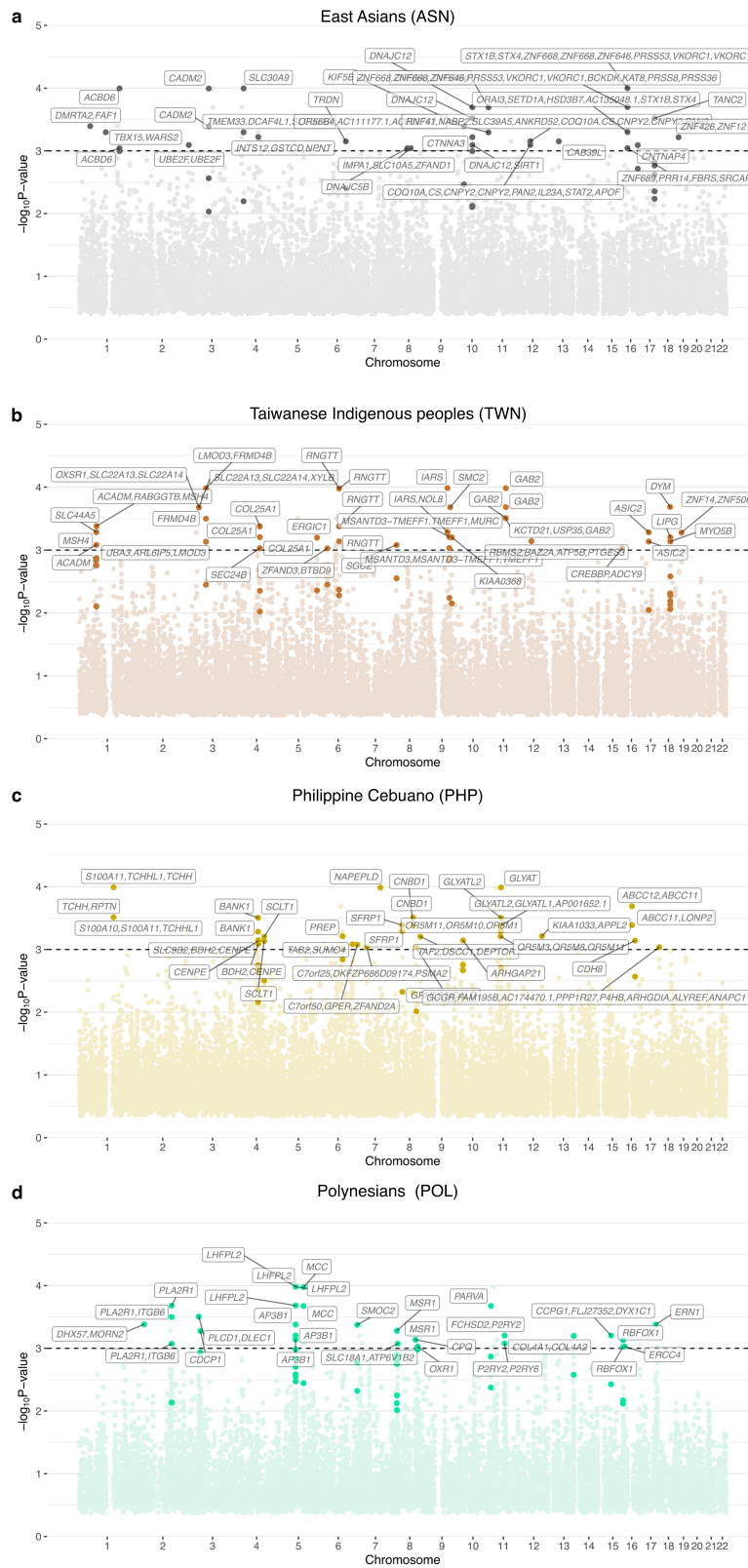
Extended Data Fig. 7 | Classic sweeps and adaptive archaic introgression.
a, b, Coloured squares indicate genomic regions displaying signals of both a selective sweep and adaptive introgression from Neanderthals (**a**) or

Denisovans (**b**). Yellow and blue frames indicate genomic regions identified in East-Asian- and Papuan-related populations, respectively. AGT, Philippine Agta; PHP, Philippine individuals.



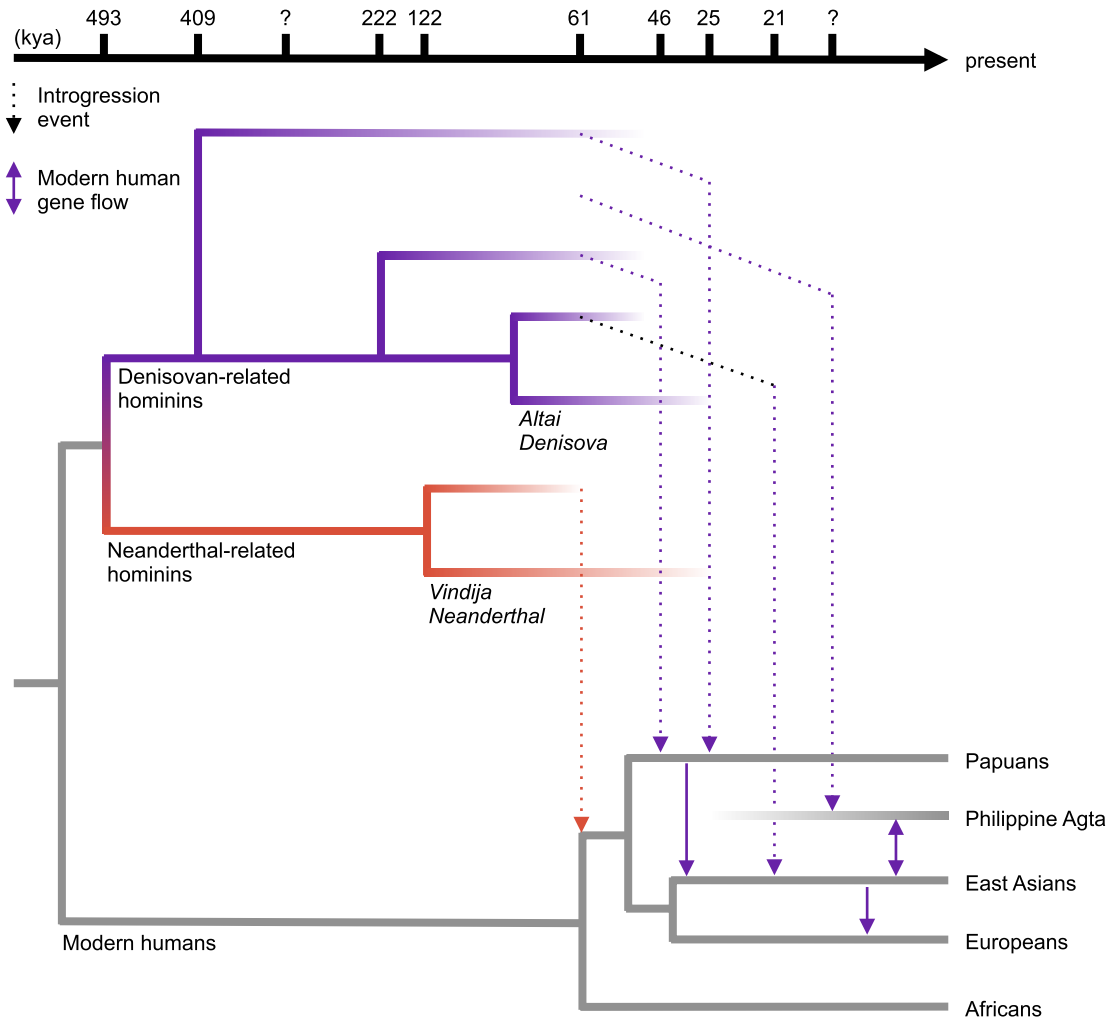
Extended Data Fig. 8 | Examples of candidate loci for classic sweeps in Pacific populations. **a, c**, Sweep signals detected in Papuan-related populations at the *GABRP* locus (**a**) and in Polynesian populations at the *LHFPL2* locus (**c**). Manhattan plots show the $-\log_{10}(P\text{value})$ of the Fisher's scores for each SNP (Supplementary Note 16). **b, d**, Maps showing the population allele frequencies for candidate SNPs rs79997355 (*GABRP*) (**b**) and rs117421341

(*LHFPL2*) (**d**). Pie charts indicate the derived allele frequency in purple, in which the radius is proportional to the sample size (Supplementary Table 1). The pie charts for the populations of Santa Cruz and Vanuatu were moved from their sampling locations for convenience. Maps were generated using the maps R package⁵¹.



Extended Data Fig. 9 | Classic sweep signals detected in East-Asian-related populations. Manhattan plots of classic sweep signals in East Asian individuals (a), Taiwanese Indigenous peoples (b), Philippine Cebuano (c) and Polynesian

individuals (d). **a-d**, The y axis shows the $-\log_{10}(P\text{value})$ for the number of outlier SNPs per window. Each point is a 100-kb window. The names of genes associated with windows with significant sweep signals are shown.



Extended Data Fig. 10 | Schematic model of the history of archaic introgression in modern humans. The phylogenetic tree depicts relationships among archaic and modern humans. Estimates for the splits between archaic, introgressing populations and for introgression episodes are shown. Five introgression events are consistent with our data: a Neanderthal introgression event into the common ancestors of non-African individuals around 61 ka; a Denisovan introgression event into the ancestors of Papuan individuals approximately 46 ka, which is shared with the ancestral Indigenous

Australian individuals and Philippine Agta populations^{14,15,17,97}; a Denisovan introgression event that occurred only in the ancestors of Papuan individuals around 25 ka; a Denisovan introgression event in the ancestors of East Asian individuals around 21 ka, the legacy of which is also observed in Philippine Agta and western Eurasian individuals due to subsequent gene flow (solid purple arrows); and a Denisovan introgression event into the ancestors of the Philippine Agta at an unknown date.

List of figures

Fig. 1: Theoretical loss of heterozygosity over time (generations) as a function of population size.	16
Fig. 2: Effect of (A) the dominance and (B) the selection coefficients on the frequency trajectory of an allele under positive selection, based on forward Wright-Fisher simulations. The selection coefficient determines the time it takes for the allele to go from very low frequency to very high frequency (i.e., the slope of the mid-section of the curve). The dominance coefficient, on the other hand, determines the time it takes for the allele to start sharply increasing in frequency, as well as the time it takes to arrive from high frequency to fixation. Simulations were computed using the SLiM engine and following recipe 9.11 of the SLiM manual (Haller & Messer 2021).....	19
Fig. 3: Different types of selective sweeps. (A) Classic sweep: a beneficial <i>de novo</i> mutation (red dot) appears on one haplotype, and subsequently increases in frequency, bringing closed linked neutral alleles (light blue dots) to high frequencies too, until eventually reaching fixation. (B) Soft sweep from standing variation: an already existing mutation present on multiple haplotypes, becomes beneficial (red dots), bringing these haplotypes at high frequency. (C) Soft sweep from recurring mutations: similar to (A), but a second, beneficial mutation (yellow dot) appears on a different haplotype from that of the first mutation, which also rises in frequency. Gray arrows represent the passage of time.....	23
Fig. 4: Changes in F_{ST} along the genome. High F_{ST} values indicate loci that are potential targets of positive selection, based on a cut-off (red dashed line) arbitrarily set to be the top 1%. From Fan et al. 2016.	25
Fig. 5: Outcome of a single pulse admixture event. An admixed population is created from a single admixture event, by inheriting a fraction α_1 of haplotypes from one source population (source population 1 in green) and a fraction α_2 of haplotypes from another source population (source population 2 in orange). As generations pass, recombination events between the different haplotypes lead to variation in genetic ancestry along and between the haplotypes. Nevertheless, the overall proportion of green and orange segments in the admixed population still reflects the initial admixture proportions.	30
Fig. 6: The f_3 statistic used to infer admixture. Phylogenetic trees with topologies that may produce positive (A and B) or negative (C) values of the f_3 statistic. Red and blue arrows denote drift paths from the tested population (P_X) to population P1 and P2 respectively. The intersect of these drift paths, if they go on the same direction, (shared drift) contributes positively to the value of the f_3 statistic (orange-coloured branches). In the case of admixture, additional drift paths that can go in opposite directions are introduced, which contribute negatively to the f_3 statistic (magenta-coloured branches).....	33
Fig. 7: Haplotype-based methods to assess admixture. (A) Weighted LD decay curve for Mbuti pygmies using San and Yoruba as reference populations. Red line corresponds to a fitted negative exponential, assuming an admixture time parameter value of 28 generations. From Loh et al. 2013. (B) Local ancestry inference of an admixed Peruvian individual (the author of this manuscript) using multiple reference populations. Obtained using a support vector machine-based method (23andme).....	35
Fig. 8: Different evolutionary scenarios that can explain shared polymorphism between two groups. (a) Convergence: two identical, beneficial, derived alleles appear independently in both groups posterior to their divergence. Shared ancestral polymorphism is maintained by chance (b) or balancing selection (c). (d) Adaptive introgression: a beneficial allele appearing in one group is transmitted to another through an introgression event. Shamelessly ripped from Hedrick's review on adaptive introgression (Hedrick 2013)	41

Fig. 9: Publication record of articles containing the terms (A) “the genetic history of” or (B) “ancient DNA” and “humans” in their title. Data from PubMed (publications from 1992 to 2021). 48

Fig. 10: Major population movements since the last glacial maximum inferred with ancient DNA. Shamelessly ripped from Liu, Mao, Krause & Fu 2021 review on ancient human genomics..... 49

List of tables

Table 1: Empirical examples of adaptive introgression	42
Table 2: Empirical examples of adaptive admixture in modern humans. The putative source bringing the adaptive mutation is marked in bold.	50

Le rôle du métissage dans l'adaptation des populations humaines à des environnements changeants.

Résumé :

Au cours de l'histoire humaine, le métissage a été un phénomène récurrent, qui a laissé une empreinte profonde sur la diversité génétique des populations. Plusieurs études ont suggéré que le métissage aurait également pu faciliter l'adaptation génétique de l'espèce humaine aux environnements locaux qu'elle a colonisés : des populations auraient acquis des mutations avantageuses par métissage avec des populations déjà adaptées à leur milieu de vie. Cependant, l'importance de ce phénomène, dénommé métissage adaptatif, ainsi que la puissance statistique d'en détecter les signatures génomiques, restent très peu connus. Dans cette thèse, j'ai utilisé des simulations informatiques intensives afin de caractériser, pour différents tests statistiques, la puissance de détecter des gènes sous métissage adaptatif, tout en considérant des situations réalistes, telles que la présence de la sélection de fond, des changements démographiques et des scénarios de métissage complexes.

Nous avons montré que deux statistiques en particulier, F_{adm} et LAD, utilisant les fréquences alléliques et les proportions de métissage attendues sous neutralité, ont une puissance élevée de détecter des mutations sous métissage adaptatif, alors que d'autres statistiques classiques, iHS et F_{ST} , détectent à tort des mutations qui n'ont été sélectionnées que dans les populations parentales. En combinant F_{adm} et LAD en une seule statistique, nous avons analysé les génomes de quinze populations métissées du monde entier, afin d'identifier des signatures génomiques de métissage adaptatif. Nous avons confirmé que la persistance de la lactase et la résistance au paludisme sont des traits qui ont été sous métissage adaptatif chez des populations métissées d'Afrique de l'Ouest, et de Madagascar, Afrique du Nord et Asie du Sud, respectivement. Notre approche a également permis d'identifier de nouveaux cas de métissage adaptatif, dont le locus *APOLI/MYH9* chez les nomades Fulani et le locus *PKN2* chez des populations d'Indonésie de l'Est, deux locus impliqués dans l'immunité et le métabolisme. Pour conclure, notre étude montre que le phénomène de métissage adaptatif a effectivement eu lieu chez les populations humaines, dont l'histoire génétique est marquée par des périodes d'isolement et d'expansions suivis de mélanges intensifs.

Mots clés : métissage, génétique des populations, sélection naturelle

The role of admixture in the adaptation of human populations to changing environments

Abstract:

Admixture has been a recurrent phenomenon through human history, having deeply shaped the genetic diversity of human populations. Numerous studies have suggested that admixture could have also contributed to the evolution of genetic adaptations to new, local environments that humans encountered and colonized: populations would have acquired beneficial mutations through admixture with local populations already adapted to the newly encountered environment. However, the importance of this phenomenon in human evolution, known as adaptive admixture, as well as the statistical power to detect its genomic signatures, remain poorly understood. In this thesis, by using intensive computer simulations, we have characterized the statistical power, for different methods, to detect genes under adaptive admixture while considering realistic parameter settings, such as background selection, changes in demography and complex admixture scenarios. We have shown that two statistics in particular, F_{adm} and LAD, which use allele frequency information and admixture estimates under neutrality, have high power to detect mutations under adaptive admixture, whereas other classic statistics, iHS and F_{ST} , falsely detect mutations under positive selection but in the source populations only. By combining F_{adm} and LAD in a single statistic, we have analysed the genomes of fifteen worldwide admixed populations to identify genomic signatures of adaptive admixture. We have confirmed that lactase persistence and malaria resistance are traits that have been under adaptive admixture in admixed populations of West Africa, Madagascar, North Africa and South Asia respectively. In addition, our approach has identified new cases of adaptive admixture, among which the *APOLI/MYH9* locus in the Fulani nomads and the *PKN2* locus in populations of East Indonesia, which are loci involved in immunity and metabolism respectively. To conclude, our study has shown that adaptive admixture has effectively occurred in human populations, whose history is marked by periods of isolation and expansion followed by intensive admixture.

Keywords: admixture, population genetics, natural selection