Genomic insight into the history of Oceanian populations : implication for human evolution and health

Jérémy Choin

Université de Paris
Frontière de l'Innovation en Recherche et Education (ED 474)

Institut Pasteur, UMR 2000, Génétique Evolutive Humaine

Genomic insight into the history of Oceanian populations: implication for human evolution and health

par Jérémy Choin

Thèse de doctorat de GÉNÉTIQUE, OMIQUES, BIOINFORMATIQUE ET BIOLOGIE DES SYSTÈMES

Dirigée par Lluis Quintana-Murci et par Antoine Gessain

Présentée et soutenue publiquement le 25 Octobre 2021

Devant un jury composé de :

Pr. Elizabeth (Lisa) Matisoo-Smith (PR Université d'Otago)	Rapporteur
Pr. Johanness Krause (PR Université de Jena)	Rapporteur
Dr. Heidi Colleran (CR Max Planck Institute for Evolutionary Anthropology)	Examinatrice
Pr. Guillaume Achaz (PR Université de Paris)	Examinateur
Dr. Etienne Patin (CR-HDR CNRS)	Membre invité
Pr. Lluis Quintana-Murci (PR Institut Pasteur)	Directeur de thèse
Pr. Antoine Gessain (PR Institut Pasteur)	Co-directeur de thèse

A James, Monique et Brigitte qui ont fait de moi ce que je suis,
A Koko pour m'avoir montré le chemin,

Acknowledgments/Remerciements

First of all, I would like to warmly thank all members of the jury: Lisa Matisoo-Smith, Johanness Krause, Heidi Colleran and Guillaume Achaz who have honoured me by evaluating this work and participating to my PhD viva. It is a real pleasure to have you all as members of my jury. I am particularly grateful to Lisa Matisoo-Smith who despite the time zone shift and distance between New Zealand and Paris has accepted to review my work.

Je tiens ensuite à remercier Lluis Quintana-Murci, mon directeur de thèse. Oui Lluis, "on est content". Merci de m'avoir ouvert les portes de ton laboratoire et pour la confiance que tu m'a accordée tout au long de ces quatre années de thèse. Je te suis reconnaissant pour ton incontestable disponibilité, ton encadrement et tes précieux conseils qui m'ont fait grandir. Merci de m'avoir poussé à toujours me dépasser et donner le meilleur de moi-même. Merci de m'avoir rappelé au quotidien que le travail paie (presque) toujours. Un chapitre se ferme mais l'histoire continue et j'espère sincèrement que nos chemins se croiseront à nouveau.

Je souhaite témoigner ma reconnaissance envers Antoine Gessain, mon co-directeur de thèse. Je le remercie chaleuresement pour nos échanges et pour l'intérêt qu'il porte à la génétique évolutive humaine bien que ce ne soit pas son domaine d'expertise. Merci Antoine et Olivier pour ce travail colossale d'échantillonage mené au Vanuatu.

Je remercie également mes tuteurs Laurent Excoffier et Frédérique Valentin pour leur souci du bon déroulement de ma thèse. Merci Laurent de m'avoir accueilli quelques semaines dans ton laboratoire et de m'avoir éclairé sur l'utilisation de *Fastsimcoal*. Frédérique, merci de tes conseils, de ton aide et de m'avoir fait découvrir un peu de ton univers, celui de l'anthropologie physique.

Mes pensées vont maintenant vers mes chèrs collègues. Tout d'abord Christine, que je remercie pour les manip', notamment pour toutes les extractions d'ADN. Merci Christine pour ces petites discussions matinales qui m'ont permis de commencer mes journées en douceur. Guillaume et Maxime que je remercie pour leur aide en statistiques. Etienne pour son encadrement. Merci beaucoup Etienne pour ta bienveillance, ton temps et ta patience. J'ai appris tellement de choses avec toi! Merci d'avoir contribué (et de toujours contribuer) au succès de chaque projet du laboratoire. Mary, Yann, Gaspard et Sebas pour les afterworks qui m'ont aidé à décompresser. A ma petite Magou, qui m'a supporté à côté d'elle et pour les vendredis soirs "Chinois et vin rouge". Je remercie chaleuresement

Lara R. Arauna et Javier Mendoza-Revilla de la "Pacific team" qui ont été mes coéquipiers et qui ont grandement contribué à la réussite de la première partie de ma thèse. Merci Lara et Javier pour votre patience et compréhension. C'est une vraie chance que d'avoir travaillé avec vous. Enfin, je remercie tous les autres membres de l'équipe Génétique Evolutive Humaine (encore présents ou anciens membres) qui ont contribué, de près ou de loin, à ma vie de "thésard".

Je voudrais remercier Ying-Chin Ko, Mark Stoneking, Maximilian Larena et Mattias Jakobsson d'avoir partagé de précieux échantillons ainsi que toutes les personnes volontaires qui ont généreusement donné de leur sang ou de leur salive. Sans eux, ce travail de thèse n'aurait pas pu se faire.

Un grand merci à la famille DANESE-LEMOINE (Anna, Marie-Laure, Claude et Etienne) de m'avoir ouvert leur porte et offert un toit au début de ma thèse. Merci Anna, Marie-Laure et Claude pour toutes ces discussions très enrichissantes (accompagnées d'un bon verre de vin) et les concerts de Jazz. Merci du fond du coeur pour votre grande générosité.

Je remercie mes amis et notamment Marie. La science nous a d'abord rapproché puis nos points communs, nos valeurs communes. Merci Marie pour ton grand coeur et ta bonne humeur. Merci de toujours me donner le sourir, merci d'être là tout simplement. Je tiens également à remercier Rodrigo pour son soutien quotidien et de m'accepter tel que je suis. Merci du fond du coeur.

Enfin, je remercie ma famille pour leur présence. Je pense à mes grands-parents et à ma mère qui m'ont toujours encouragé, soutenu et aidé. Ce travail vous est dédié.

Merci! Thank you!

Summary

Titre: Approches en génétique des populations pour comprendre l'histoire des populations d'Océanie.

Résumé:

Introduction

Au cours des 125 000 dernières années, l'homme moderne (*Homo sapiens*) s'est répandu sur tous les continents et s'est installé dans divers écosystèmes, aussi extrêmes que le désert du Sahara, le cercle polaire arctique ou l'Himalaya. Les données archéologiques et linguistiques ont fourni des informations précieuses sur le rythme des dispersions humaines à travers le monde, cependant de nombreuses questions restent ouvertes : les populations ont-elles migré avec leurs langues et leurs modes de vie ? Les cultures humaines définies par l'archéologie reflètent-elles des entités génétiques distinctes ? Les dispersions humaines se sont-elles accompagnées d'un mélange génétique avec des groupes locaux d'humains archaïques ou modernes ? Comment les humains se sont-ils génétiquement adaptés aux environnements nouvellement colonisés ? L'avènement récent des technologies de séquençage à haut débit permet désormais d'aborder ces questions dans les moindres détails, à travers la caractérisation complète de la diversité génétique des populations humaines vivant à differentes époques (époque actuelle ou passée). Les approches de génétiques des populations sont ainsi très complémentaires aux études archéologiques, anthropologiques et linguistiques. En effet, elles permettent d'expliquer d'autres facettes de l'histoire complexe des populations humaines.

Quatre forces évolutives façonnent la diversité génétique d'une population : la mutation qui crée la variation génétique, la dérive génétique qui tend à augmenter la différenciation génétique des petites populations, les migrations ou le flux de gènes qui homogénéise les populations et la sélection naturelle qui permet aux populations de s'adapter à leurs environnements (c'est-à-dire la sélection naturelle positive) et de purger les mutations délétères du génomes (c'est-à-dire la sélection négative). Les différents évènements caractérisant l'histoire des peuples humains – les fluctuations au cours du temps de la taille des populations, les évènements de métissage et d'introgression avec des populations humaines aujourd'hui éteintes ou les événements d'adaptation génétique à de nouveaux environnements - façonnent leur diversité génétique. Ainsi, la génétique des populations peut être utilisé afin de reconstruire le passé démographique des diffèrent peuples et de mettre en lumière les fonctions biologiques qui ont contribué à leur adaptation pour, in fine, mieux comprendre leur susceptibilité face aux maladies.

L'Océanie

La région de l'Océanie couvre plus de 8 500 000 km² de surface terrestre répartie entre l'Australie/Papouasie-Nouvelle-Guinée et l'île de Pâques (l'île la plus à l'est du triangle polynésien). Cette région du monde est peuplée par un peu plus de 40 000 000 habitants, représentant uniquement environ 0,5 % de la population mondiale. Cependant, cette région possède une incroyable diversité culturelle et linguistique avec environ 1 750

langues différentes (25 % des langues mondiales, en excluant les langues parlées en Australie) réparties en deux groupes : les langues austronésiennes et les langues papoues. En plus de ces langues, les peuples insulaires d'Océanie parlent également le français, l'anglais et différentes langues créoles ou « pidgin » comme le bichlamar qui est l'une des langues officielles du Vanuatu.

Qui sont les peuples du Pacific ? D'où viennent-ils ? Ces questions ont suscité l'intérêt des premiers explorateurs scientifiques européens. En 1852, Dumont d'Urville, un navigateur botaniste français, divise l'Océanie en trois régions afin de considérer la diversité phénotypique et culturelle rapportée par les explorateurs européens : Mélanésie, Polynésie et Micronésie. Cependant, cette classification géographique de l'Océanie a été faite dans le cadre de la théorie raciale et deux de ces régions, la Mélanésie et la Micronésie ne reflètent aucune réalité culturelle. Pour rendre compte de la richesse culturelle et linguistique observée en Océanie ainsi que des différences d'origines et d'histoire des peuples, les chercheurs préfèrent aujourd'hui utiliser un autre découpage : l'Océanie proche, qui regroupe les îles peuplées pendant la période du Pléistocène, entre 50 000 et 25 000 ans et l'Océanie lointaine regroupant les îles peuplées durant la période de l'Holocène, il y a environ 3 000 ans. Les données archéologiques, anthropologiques et linguistiques recueillies depuis le XXe siècle ont fourni un éclairage crucial sur l'histoire du peuplement, la répartition géographique des premiers peuples d'Océanie, leurs sociétés, leurs différents modes de vie ainsi que sur les anciens réseaux d'échanges. L'ensemble de ces données multidisciplinaires ont surtout permis de proposer des hypothèses et des modèles – encore actuellement débattus – sur l'histoire du peuplement de l'Océanie proche et lointaine.

Objectifs de la thèse

L'Océanie est composée de milliers d'îles regroupées en deux grands ensembles, caractérisés par deux vagues de peuplement distinctes : l'Océanie proche et l'Océanie lointaine. Le premier ensemble, comprenant la Nouvelle-Guinée, l'archipel Bismarck et les Îles Salomon, a été peuplé par l'homme moderne (*Homo sapiens*) il y a environ 40 000 ans. Le deuxième, incluant toutes les autres îles d'Océanie, n'a été peuplé qu'il y a un peu plus de 3 000 ans et ce par l'expansion de peuples parlant des langues austronésiennes, probablement originaires de Taiwan (modèle «Out-Of-Taiwan »). Ce projet de thèse vise à reconstituer l'histoire génétique des populations insulaires d'Océanie, dans le but de reconstruire leur passé démographique pour à terme, mieux comprendre leur rapport face aux maladies. Ainsi, nous avons séquencé l'ADN de 317 individus autochtones répartis en 20 populations et couvrant l'ensemble des régions géographiques à la base de l'histoire du peuplement de l'Océanie proche et lointaine. Plus précisément, mon projet de thèse vise à (i) caractériser la diversité génétique des populations d'Océanie, (ii) retracer tous les différents événements constituant leur histoire démographique, et enfin (iii) évaluer la purge des mutations délétères (c'est-à-dire des mutations pouvant provoquer des maladies) dans ces populations. Dans sa globalité, cette étude nous a permis d'en apprendre davantage sur l'histoire génétique de l'Océanie, une région du monde qui a été largement oubliée des études génétiques.

Résumé des résultats

Histoire démographique des populations du pacifique

Pour reconstruire l'histoire du peuplement et du passé démographique des populations insulaires du Pacifique, nous avons conjointement inféré les paramètres caractérisant leur histoire démographique en utilisant des spectres de fréquences alléliques multidimensionnels et une approche basée sur le maximum de vraisemblance. Tout d'abord, nous avons exploré différentes topologies d'arbres et estimé les paramètres démographiques des groupes d'Océanie proche (c'est-à-dire les groupes de Nouvelle-Guinée, de l'archipel Bismarck et des Îles Salomon). Cette étude a révélé que le peuplement de cette région avait été accompagné d'un effet fondateur très fort, environ 5 fois supérieur à celui observé pour le peuplement de l'Eurasie. Cette étude a aussi permis d'inférer une divergence ancienne des différentes populations de cette région remontant à la période du Pléistocène supérieur, il y 20 000 à 45 000 ans. Ces résultats indiquent un isolement génétique rapide des différents groupes de l'Océanie proche, après le peuplement initial daté d'environ 45 000 ans (données archéologiques, (O'Connell et al. 2018a; O'Connell et Allen 2015)).

Nous avons également testé différentes topologies et estimé des paramètres démographiques pour les peuples de l'Océanie lointaine de l'ouest (Vanuatu). Nous avons ainsi confirmé l'expansion récente (inférieure à 3 000 ans) de groupes originaires de l'Archipel Bismarck vers l'Océanie lointaine de l'ouest, notamment en direction des îles du Vanuatu, en accord avec des études récentes d'ADN ancien (Posth et al. 2018; Lipson et al. 2018). Ces résultats ont également suggéré des contacts complexes et multiples entre des groupes d'Asie de l'est et des groupes d'Océanie proche, en désaccord avec l'hypothèse « Out-of-Taiwan ». En raison d'un manque de continuité entre les premiers ni-Vanuatu et les ni-Vanuatu actuels, comme le montrent les études d'ADN ancien et les études craniométriques (Posth et al. 2018 ; Lipson et al. 2018 ; Valentin et al. 2016), l'interprétation des modèles démographiques utilisant l'ADN moderne est très limitée.

En supposant un modèle d'isolement avec migration, nous avons estimé que les peuples autochtones taïwanais et les locuteurs malayo-polynésiens ont divergé il y a environ 7 300 ans, en contradiction avec le modèle « Out-of-Taiwan » - hypothèse qui prédit un évènement de dispersion de Taïwan il y a environ 4 800 ans et qui aurait apporté à la fois l'agriculture et les langues austronésiennes en Océanie (Bellwood 1997). Nous avons confirmé ces temps de divergence anciens, même en considérant des flux de gènes dans les groupes parlant des langues austronésiennes, mais avec des intervalles de confiance plus larges. Ces résultats suggèrent une structure de population des locuteurs austronésiens qui prédate l'apparition de l'agriculture à Taïwan. Cependant, en raison de la grande incertitude dans les estimations, d'autres analyses utilisant des génomes anciens sont nécessaires.

En somme, ces analyses ont permis d'affiner notre compréhension de l'histoire démographique et adaptative des peuples des îles d'Océanie, une région du monde longtemps absente des études de génétique des populations.

Efficacité de la sélection naturelle dans les populations du Pacifiques

D'un point de vue théorique, la génétique des populations prédit que, pour des populations de petite taille, l'efficacité de la sélection est réduite, conduisant ainsi à une plus forte accumulation de mutations pouvant causer des maladies rares ou fréquentes (Simons et al. 2014; Balick et al. 2015). Si certaines études épidémiologiques ont révélé des cas de pathologies à des fréquences anormalement élevées dans des populations insulaires (O'Brien et al. 1988; Carr, Morton, and Siegel 1971; Eickhoff and Beighton 1985), rares sont celles qui ont formellement démontré une augmentation du fardeau de mutations délétères dans les populations humaines ayant connu de forts effets fondateurs suivis d'un isolement. Enfin, des études (Organisation Mondiale de la Santé) ont également mis en évidence la forte prévalence de maladies métaboliques, notamment la goutte, le diabète et l'obésité dans les populations Océaniennes. Toutefois, les forces évolutives en œuvres sont actuellement débattues (Gosling et al. 2015) : hypothèse du phénotype économe (i.e. sélection naturelle) versus hypothèse du phénotype dérivant (i.e. dérive génétique). Ainsi, il apparaît essentiel d'analyser la façon dont la démographie et la sélection naturelle ont façonné la diversité génétique de ces populations afin d'améliorer notre compréhension des différences de susceptibilité aux maladies entre populations de régions du monde jusqu'à présent très peu étudiées.

Au cours des dix dernières années, plusieurs études ont étudié l'impact de la démographie sur le fardeau des mutations délétères chez l'homme (Lopez et al. 2018a; Simons and Sella 2016; Simons et al. 2014; Do et al. 2015; Henn et al. 2016b; Henn et al. 2015b; Lohmueller et al. 2008; Lohmueller 2014; Fu et al. 2013; Pedersen et al. 2017a; Font-Porterias et al. 2021). Bien qu'il y ait de plus en plus de preuves suggérant un impact négligeable du goulot d'étranglement associé à la sortie d'Afrique (« Out-of-Africa ») sur le fardeau des mutations délétères additives (Lopez et al. 2018a; Simons and Sella 2016; Simons et al. 2014; Do et al. 2015), de fortes réductions de la taille des populations, comme celles subies par les Inuits du Groenland, peuvent avoir impacté le nombre et la fréquence des mutations délétères récessives (Pedersen et al. 2017a). Dans ce contexte, compte tenu de leur histoire de peuplement caractérisée par de forts effets fondateurs en série, les populations des îles du Pacifique Sud offrent un excellent modèle pour évaluer dans quelle mesure ces processus démographiques spécifiques ont eu un impact sur l'apparition et la distribution de mutations délétères dans le génome humain.

Nous avons étudié le fardeau des mutations délétères et l'efficacité de la sélection chez les populations insulaires du Pacifique en utilisant des séquences « génome entier ». Par rapport à d'autres populations non africaines, les polynésiens et les papous portent moins de mutations délétères - y compris les mutations « perte de fonction » (LoF) - mais qui ont tendance à ségréger à des fréquences plus élevées, probablement en raison d'une forte dérive génétique. Nous avons ensuite cherché à savoir si l'histoire démographique des populations insulaires du Pacifique avait eu un impact sur leur fardeau de mutations délétères. Pour ce faire, nous avons estimé la distribution des effets de fitness des nouvelles mutations délétères, ainsi que le fardeau génétique des papous, des peuples des îles Salomon, des ni-Vanuatu et des polynésiens. Nos résultats montrent que, malgré leurs différences marquées de régimes démographiques, seules des différences subtiles dans la capacité de la sélection naturelle à purger les allèles délétères sont observées entre les

océaniens et les autres populations humaines.

En somme, ces résultats suggèrent que la forte dérive génétique agissant sur certains groupes océaniens a eu des conséquences limitées sur l'efficacité de la sélection naturelle. Cependant, des analyses complémentaires, telles que des simulations, sont nécessaires pour évaluer la trajectoire du fardeau génétique au cours du temps ainsi que pour examiner en détail l'impact du récent évènement de métissage asiatique sur le fardeau de mutations délétères des populations insulaires du Pacifique.

Mots clefs: Océanie ; génomique ; histoire démographique ; sélection naturelle ; fardeau de mutations délétères

Title: Genomic insight into the history of Oceanian populations: implication for human evolution and health.

Abstract: Oceania is key to understand human evolution history, as contemporary Pacific islanders descend from two highly divergent, ancestral groups, who represent the early out-of-Africa dispersal > 45,000 years ago and the most recent expansion into empty territories < 1,000 years ago. Ultimately, the region of Oceania is of major importance for addressing questions related to human dispersal and natural selection processes. The improvement of DNA sequencing methods, combined with the development of mathematical and statistical frameworks, can provide insight into both the way natural selection removes disease-causing mutations from human populations and their potential to adapt to a broad range of climatic, nutritional, and pathogenic conditions. Oceania, owing to its insular environment, provides with an excellent model to test important hypotheses for the study of human genetic diversity and medical research. In this context, the aims of this thesis are to bring knowledge on the demographic past of Oceanian islanders and to the question of how population size changes and admixture affect the burden of deleterious mutations in these populations. To do so, we have sequenced the whole genomes of 317 individuals from 20 populations that cover the geographic transect at the basis of the peopling history of Near and Remote Oceania. Specifically, this thesis aims to (i) characterize the genetic diversity of these populations at high-resolution, (ii) reconstruct their past demographic history in terms of divergence, migration and population-size changes, and finally (iv) evaluate their burden of deleterious mutations. All combined, this thesis project increased our understanding of the genomic history of Oceania, a region of the world that has been largely neglected in genomic studies.

Keywords: Oceania; genomics; demographic history; natural selection; burden of deleterious mutations

List of Figures

List of abbreviations

A	Adenine
ABC	Approximate Bayesian Computation
AFS	Allele Frequency Spectrum
APT	Austronesian Painting Tradition
BMI	Body Mass Index
C	Cytosine
DNA	Deoxyribonucleic Acid
G	Guanine
GWAS	Genome wide association study
LCC	Lapita Cultural Complex
LGM	Last Glacial Maximum
LoF	Loss-of-Function
MP	Malayo-Polynesian
mtDNA	Mitochondrial genome
MRCA	Most Recent Common Ancestor
N_e	Effective population size
NGS	Next Generation Sequencing
NRY	Non-combining region of Y
SNP	Single Nucleotide Polymorphism
SFS	Site Frequency Spectrum
T	Thymine
WGS	Whole Genome Sequencing
WHO	World Health Organization

Contents

Introduction

The region of Oceania covers more than 8,500,000 km² of land surface between Australia/Papua New Guinea and Easter Island, the easternmost island of the Polynesian Triangle. More than 40,000,000 inhabitants populate this region, which represents only around 0.5% of the total world population. However, this region has an incredible cultural and linguistic diversity with around 1,750 different native languages (25% of the worldwide languages, excluding Australian languages) divided into two groups: Austronesian and Papuan languages. In addition, to these native languages, Oceanian islanders speak also French, English and different creole or pidgin languages such as the Bislama which is one of the official languages of the Vanuatu.

What are the origins of the Pacific peoples? Who are they? What is their peopling history? These burning questions generated the interest of the first European scientific explorers. In 1852 Dumont d'Urville, a French botanist navigator, divided Oceania into three regions in order to consider the phenotypic and cultural diversity reported by Europeans explorers: Melanesia, Polynesia and Micronesia. However, this geographical classification of Oceania was made in the context of the racial theory and two of these regions, Melanesia and Micronesia do not reflect any cultural reality. To account for the cultural and linguistic richness observed in Oceania as well as for the differences in origins and in peopling history, scholars today, prefer to use another division of the region: Near Oceania, which groups islands settled during the Pleistocene period around 50,000-30,000 years ago and Remote Oceania for islands peopled during the Holocene period, around 3,000 years ago. Archaeological, anthropological and linguistic data collected since the 20th century undoubtedly provided crucial insight into the time of settlement, the geographic distribution of the first Near and Remote Oceanians, their societies, their different lifestyles and also their dynamic trading networks. Importantly, hypotheses and models – still currently debated - about the peopling history of Near and Remote Oceania were drawn from these multidisciplinary data.

Genetic approaches are very complementary to archaeology, anthropology and linguistics and can explain other facets of the complex history of human populations. However, Oceanian islanders are underrepresented in genetic databases and very little is known about their current and past genetic diversity. The different events characterizing the past history of human groups - population size changes over time, admixture, introgression events with now extinct hominins or events of genetic adaptation to new environment – shape their genetic diversity. The advance of technologies to access DNA sequences, together with the development and improvement of mathematical algorithms, now allow

population geneticists to trace back all these different events from both modern and ancient DNA data. Therefore, evolutionary genetic approaches can be used to reconstruct the demographic past of Near and Remote Oceanian islanders and to hypothesized on the biological functions that contributed to their adaptation and, ultimately, to better understand their present-day relation to diseases.

Chapter 1

Settlement and peoples of Oceania

The current island of New Guinea is politically divided into two regions, the eastern part belongs to the independent state of Papua New Guinea (Eastern New Guinea, the Bismarck Archipelago and Bougainville) and the western half (the provinces of Papua and West Papua) is part of Indonesia. Eastern New Guinea, the Bismarck Archipelago and the Solomon Islands form the geographical, archaeological, linguistic and anthropological entity of Near Oceania, the first and only region of Oceania settled during the Pleistocene period (Figure 1.1).

1.1 Oceania in the Pleistocene

During the Pleistocene period (2,580,000 to 11,700 years ago), the current territories of New Guinea, Australia and Tasmania were connected into a single landmass named Sahul. This ancient continent was separated by around 100 km of water (i.e. the Wallace's Line) from island Indonesia and mainland Southeast Asia, which were gathered in a single continent known as Sunda (Figure 1.1). The sea level was 150 meters lower than it is today, facilitating the settlement of Sahul through Sunda by *Homo sapiens* between around 50,000 and 65,000 years ago (O'Connell et al. 2018a; O'Connell and Allen 2015; Clarkson et al. 2017). Northern and Eastern islands lying off New Guinea, namely the Bismarck Archipelago and the Solomon Islands respectively, were never connected to Sahul. When and how were Sahul, the Bismarck Archipelago and the Solomon Islands settled? Who were the first settlers of these regions?

1.1.1 Sahul: the initial settlement

Archaeological materials indicate that the peopling of Near Oceania began with the settlement of the ancient continent of Sahul. Multiple routes taken by the first settlers were hypothesized including the northern and southern routes that are today favoured by the scientific community but highly debated (Kealy, Louys, and O'Connor 2017, 2018; Bird et al. 2018). The northern route hypothesis assumes a peopling of Sahul via Sulawesi and New Guinea and the southern route, via Flores Island/Timor and Australia. Paleogeographic studies propose two advantages of the northern route with respect to the southern route: (i) distances between islands are shorter and (ii) backward voyages were possible using winds and water currents. However, the archaeological record is older in northwestern Australia (southern Sahul) (Clarkson et al. 2017) than New Guinea (northern Sahul) favoring the southern route (O'Connell et al. 2018a; O'Connell and Allen 2015).

Besides the route(s) taken by the first settlers, recent works based on mathematical models and simulations also shed light on the nature of these voyages: it is unlikely that Sahul was settled by accident but instead, the voyages were planned and deliberated, involving a founder population of a least 1,300 individuals (Bird et al. 2019; Bradshaw et al. 2019; Bradshaw et al. 2021). These first settlers of the region lived in small structured groups and were highly mobile. They were likely initially "strand looper" foragers who hunted and gathered maritime resources along the shore but also rapidly exploited plant resources such as yams in the Highlands of New Guinea (Summerhayes et al. 2010).

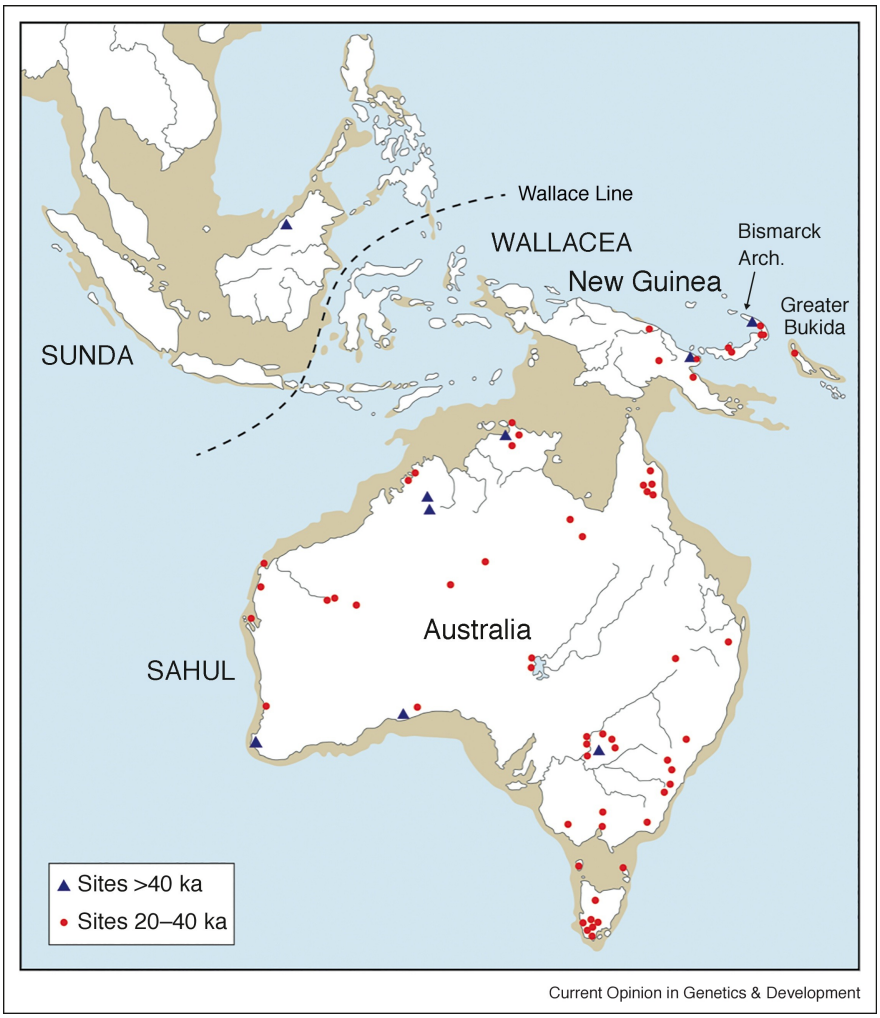


Figure 1.1: **Sunda and Sahul in the Pleistocene**. Map showing Sunda and Sahul landmasses before (light brown) and after (white) Holocene sea level changes. The distribution of Pleistocene archaeological sites is represented by red dots and blue triangles. The map is from (Gosling and Matisoo-Smith 2018b)

1.1.2 The Bismarck Archipelago: behaviour changes and networks

The Bismarck Archipelago - composed of volcanic islands and covered by a dense tropical rain forest - was settled around 40,000 years ago shortly after northern Sahul (New Guinea) (Leavesley and Chappell 2004). Three islands of the archipelago were never connected to each other by land: New Britain, New Ireland and Manus (Figure 1.2). The Pleistocene archaeology of the Bismarck Archipelago can be divided into two broad periods. The first period covers the initial settlement (40,000 years ago) until the Last Glacial Maximum (LGM) around 22,000 years ago and the second period started after the LGM until the end of the Pleistocene period around 11,700 years ago.

Before 20,000 years ago, the first settlers of the Bismarck Archipelago, were small, highly mobile groups of hunter gatherers, similar to those in Sahul. They moved in search of food resources such as shellfish, rats and reptiles. After 20,000 years, the different groups developed networks where they exchanged food and goods; in other words, resources "were moved to people" (Gosden 1995; Leavesley 2006). Indeed, there is archaeological evidence of connectivity from that period between Bismarck islands but also between islands of the Bismarck Archipelago and Northern Sahul (New Guinea). These connections led, for example, to the introduction of the cuscus (*Phalanger orientalis*) to New Ireland from New

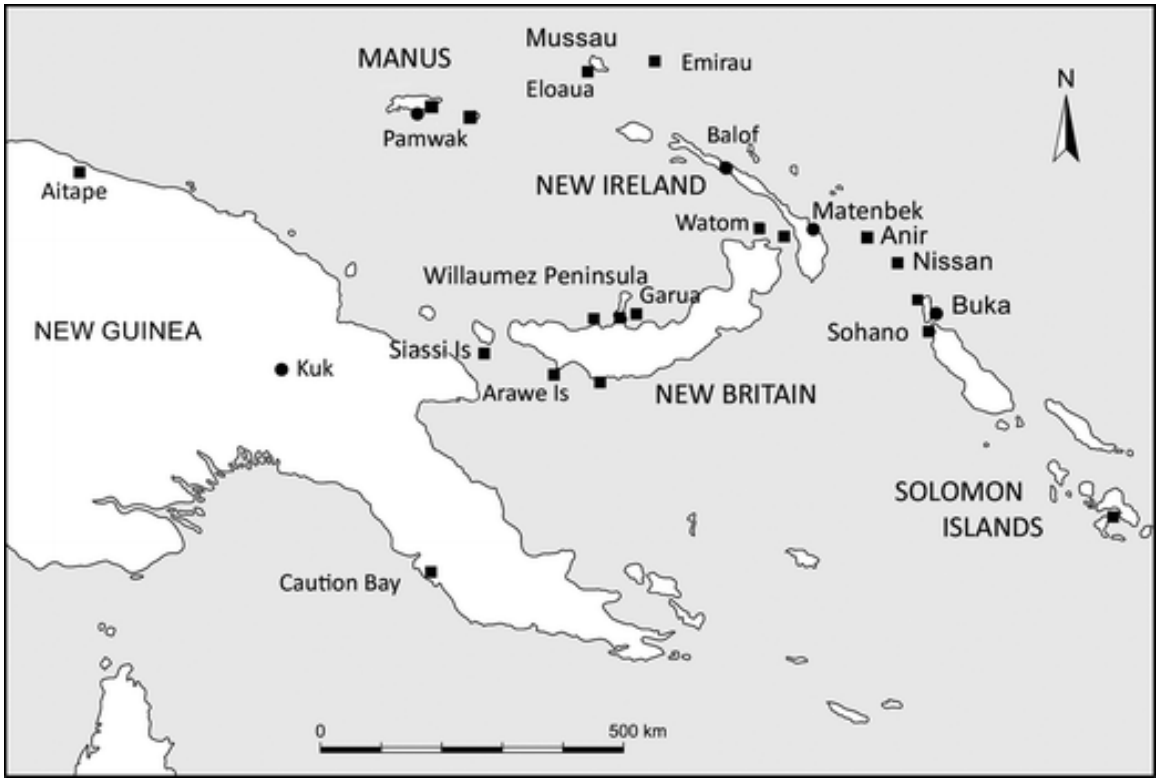


Figure 1.2: **Map of the Bismarch Archipelago (Specht et al. 2014)**.

Guinea via New Britain (Leavesley 2005). The cuscus became the main resource in the archipelago. Between 20,000 and 18,000 years ago, the first extractions of obsidian stones occurred in New Britain, with evidence of exchanges to New Ireland (Summerhayes 2009). Later, around 12,000 years ago, obsidian, together with other animals such as the bandicoot and another species of cuscus (*Spilocuscus kraemeri*), were introduced to Manus island from New Guinea (Spriggs 1997).

1.1.3 The Solomon Islands: An isolated archipelago?

The Solomon Islands encompasses hundreds of inhabited islands covered of tropical rainforest and mountains. The main islands of this archipelago are Bougainville (politically part of Papua New Guinea), Vella Lavella, New Georgia, Kolombanga, Choiseul, Santa Isabel, Guadalcanal, Malaita and Makira (Figure 1.3). Most of these islands were combined in a single landmass named Greater Bukida or Greater Bougainville until the end of the Pleistocene period (Walter and Sheppard 2009).

Archaeology for the Pleistocene of the Solomon Islands is only represented by one site in the island of Buka (Kilu sites, western Solomon Islands) discovered in 1988. This site revealed an early settlement of the western Solomon Islands by *Homo sapiens* from 29,000 year ago, after the peopling of Sahul and the Bismarck Archipelago (Wickler and Spriggs

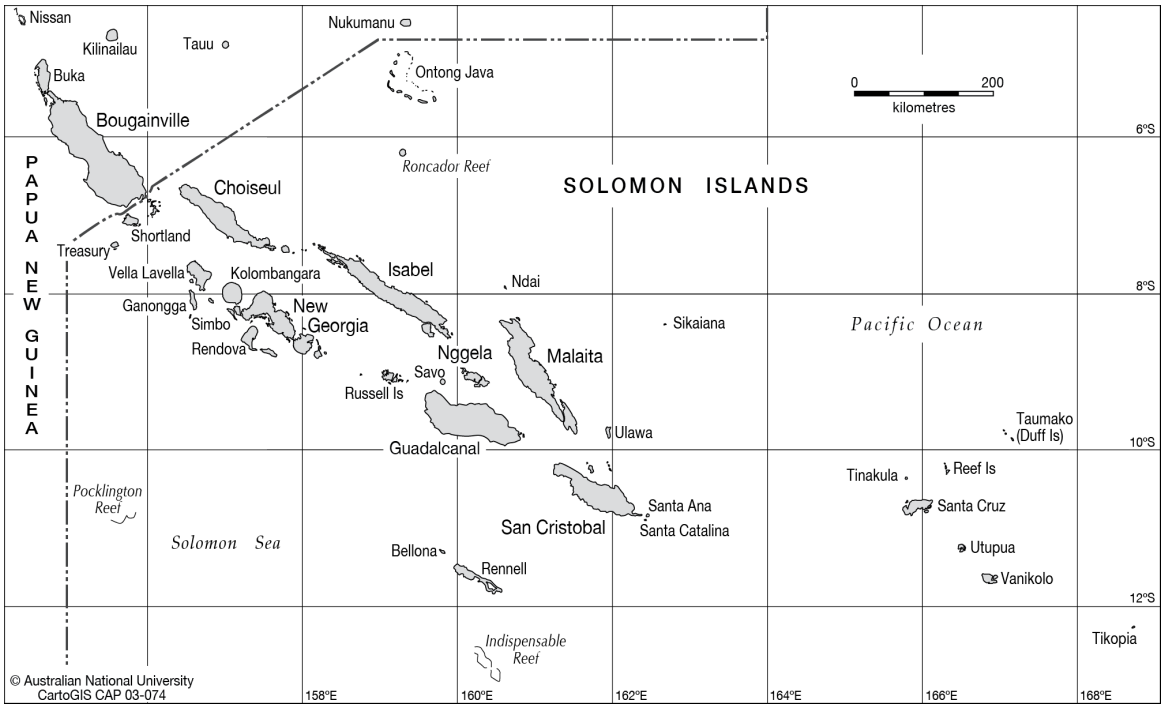


Figure 1.3: **Map of the Solomon Islands (http://asiapacific.anu.edu.au)**.

1988). This settlement would have involved a sea crossing of less than 200 km from New Ireland (Bismarck Archipelago). Little is known about the first settlers of the Solomon Islands and there is no archaeological evidence of the changes in behaviour and networks observed in the Bismarck Archipelago and New Guinea. The first and only evidence of connexion with the Bismarck Archipelago is the presence of *Canarium* charcoals dated from the end of the Pleistocene (Walter and Sheppard 2009).

1.2 Oceania in the Holocene

While the different islands that compose the region of Near Oceania were peopled during the Pleistocene period, Remote Oceania remained uninhabited until the late Holocene period. Who were the first settlers of the remote islands of Oceania and where did they come from? Which route(s) did they take? How did they settle Remote Oceania?

1.2.1 Lapita: the first Remote Oceanians

Remote Oceania comprises the islands of Micronesia, the Reef/Santa Cruz, the Vanuatu archipelago, New Caledonia, Fiji, and the different Polynesian islands (Figure 1.4). This region was settled only from around 3,200 years ago by seafarers, associated with the spread of the **Lapita Cultural Complex** (LCC) and **Austronesian languages**. I will focus here mainly on the peopling history of the western part of this region, which includes islands from the Reef/Santa Cruz to Fiji.

The Lapita Cultural Complex

Up to now, 293 Lapita sites have been found across the Pacific region covering a geographic transect including New Guinea, the Bismarck Archipelago, the Solomon Islands, Vanuatu, New Caledonia, Fidji, Tonga, Samoa as well as Wallis and Futuna (Bedford and Spriggs 2019). The earliest site is dated to 3500-3200 years ago in the Bismarck Archipelago (Mussau island), likely the homeland of the Lapita Cultural Complex (Rieth and S. 2017). This oldest Lapita site coincides with the most massive volcanic eruption that occurred in New Britain island (Bismarck Archipelago) and named the W-K2 event, around 3,600 years ago. Novel archaeological artefacts were found above the W-K2 tephra as well as evidence for a change in settlement pattern, which, altogether reflect a sharp cultural change soon after the volcanic eruption (Kirch 2017).

This new cultural assemblage is mainly characterised by a specific type of decorated potteries known as dentate-stamped pottery with a large spatiotemporal variation in

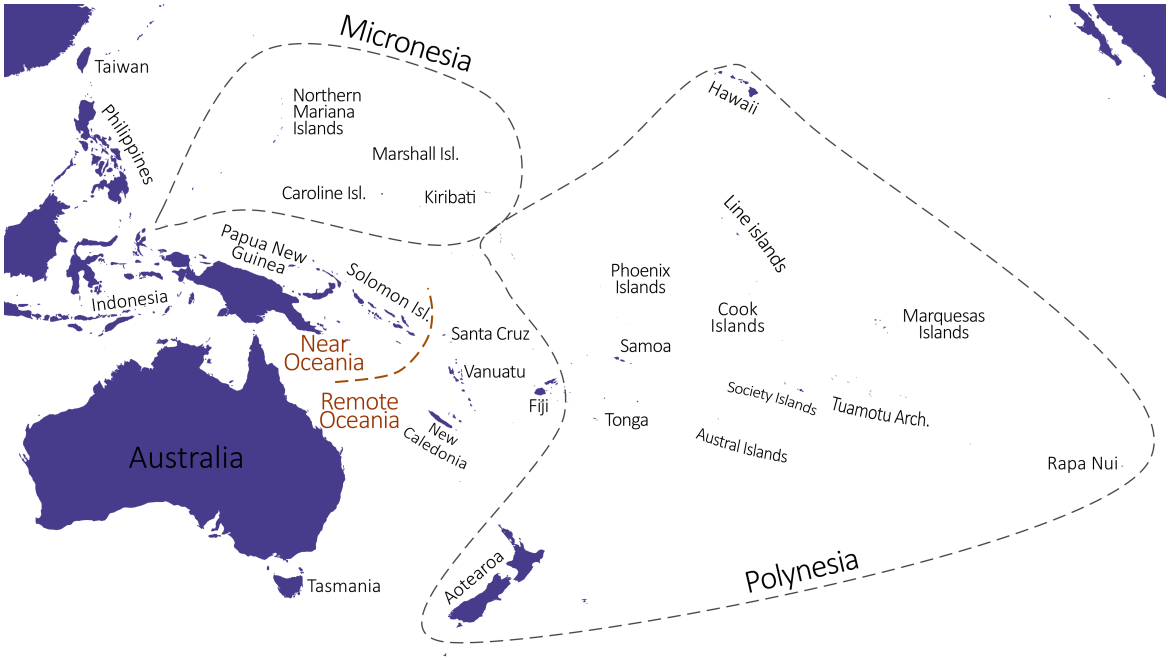


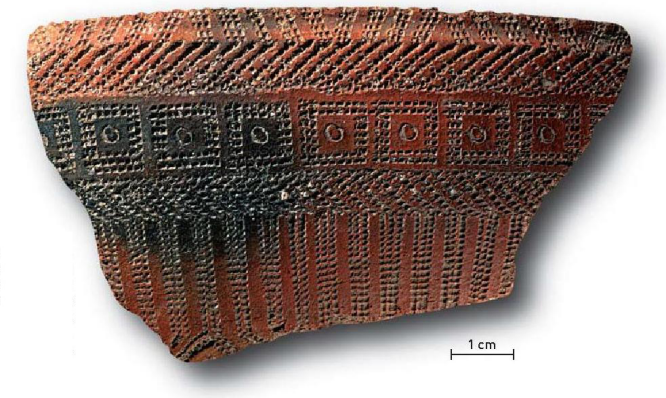
Figure 1.4: **Map of Near and Remote Oceania**. Brown dashed line indicates the limit between Near and Remote Oceania

motifs, decorations, style and form (Figure 1.5a). The specificity of this type of pottery relies in its production with the use of toothed tools to stamp complex patterns into the wet clay. Besides potteries, the Lapita culture is also characterized by long-distance transfers of obsidian (Figure 1.5b), the use of shell-based ornaments and tools such as arm rings, necklaces, food scrapers, fishhooks and adzes (Figure 1.5c) (Noury and Galipaud 2011).

Archaeological sites were mainly located along the coasts or in small offshore islands reflecting a preference of the Lapita people for small and uninhabited areas. There is evidence that the first Lapita settlers transported with them domesticated animals and plants such as taro, yams, coconuts, pigs, dogs and rats, which indicate horticulture and gardening practices (Kirch 2017). However, isotopic measures of human and pig bone collagen from archaeological sites in Vanuatu, revealed that initially, Lapita people were likely "strand loopers", who mainly lived along the shore and consumed maritime resources and to a lesser extent wild animals as well as domesticated plants and animals (Kinaston, Buckley, et al. 2014; Kinaston, Bedford, et al. 2014).

The Austronesian languages

The Austronesian language family comprises around 1,200 languages spoken by more than 380 million people in the world. Austronesian speakers form the largest expanded diaspora, which covers territories ranging from Madagascar, Island Southeast Asia, Near Oceania and Remote Oceania up to the Polynesian Triangle. The world's largest language



(a) Lapita pottery from the Bismarck Archipelago

(b) Obsidian from Vanuatu

(c) Adzes from the Solomon Islands

Figure 1.5: **Archaeological elements of the Lapita Cultural Complex**

density per capita is located in Vanuatu where 138 Austronesian languages are spoken, corresponding to about one language for 1,700 speakers (Klamer 2019; Blust 2019, 2009).

The study of the vocabulary, mainly cognates (i.e. words of the same origin) together with phonology (i.e. sounds), indicates that (i) all Austronesian languages derived from a same ancestral language named Proto-Austronesian and that (ii) all Austronesian languages spoken outside Taiwan belong to the same group known as Malayo-Polynesian (Blust 2019). Austronesian languages are categorized into 10 primary subgroups: Atayalic, East Formosan, Puyama, Paiwan, Rukai, Tsouic, Bunun, western Plains, northwest Formosan and Malayo Polynesian. The first 9 groups are found in Taiwan and are gathered into a group named "Formosan".

The Malayo-Polynesian (MP) languages, spoken outside Taiwan, are themselves divided into western MP and central-eastern MP. Western MP speakers are located in the Philippines, western Indonesia, mainland Southeast Asia, Madagascar as well as in some Micronesia islands. Central-eastern MP languages are found in eastern Indonesia as well as in Near and Remote Oceania. Centre-eastern MP languages are subdivided into deeper levels as shown in Figure 1.6. It is worth mentioning that although scholars broadly use this tree of the Austronesian language family, some branches are still currently debated (Blust 2009). For example, Blust and other linguists suggest that languages grouped into western MP do not belong to a unique subgroup but instead, correspond to multiple branches or subgroups that do not fall within the central-eastern MP cluster (Ross 1995; Blust 1999). Similarly, the catalogue of worldwide languages and dialects, Glottolog (https://glottolog.org/), does not consider the subgroup western MP but instead distinguishes a total of 25 subgroups of MP including central-eastern MP, Central MP, eastern MP, S Halmahera W New Guinea and Oceanic.

1.2.2 Origin of the Lapita and settlement of Remote Oceania

From the “express-train” to the “Out-of-Taiwan” model

In 1988, Jared M. Diamond proposed the hypothesis of the “express-train” to explain the origin of the Polynesian populations (Diamond 1988). This model stipulates that a group of people associated with the Lapita culture, spread rapidly over around 4,500 km, from the Bismarck Archipelago to Samoa, the hypothesized cradle of the ancestral Polynesian population. These sea travellers brought with them animals, plants and also agriculture. However, Diamond did not address the question of the origin of the first Lapita people (“Where west of the Bismarcks did the train start and what were its intermediate stations?”, (Diamond 1988)).

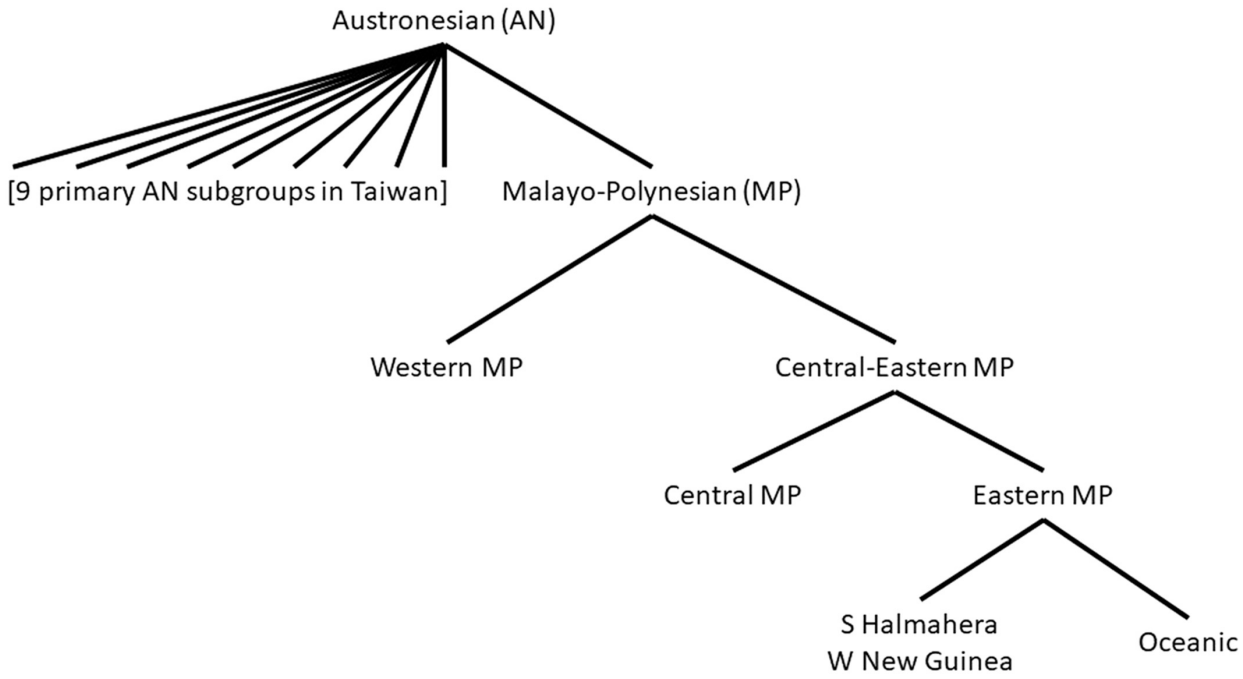


Figure 1.6: **Tree of Austronesian languages (Blust 2009)**.

The strong correlation between sites where Lapita potteries were found and the geographic repartition and structure of the Austronesian languages prompted the archaeologist Peter Bellwood to hypothesize that the Lapita Cultural Complex derived from the vast Austronesian expansion in Oceania (Bellwood 1997). In this context, both Peter Bellwood and the linguist Robert Blust, refined Diamond's "express-train" model by providing a Taiwanese origin for the Proto-Austronesians and thus by correlation, for the Proto-Lapita culture. This refers to the "Out-of-Taiwan" model. More precisely, this model proposes that the ancestors of Polynesians and their culture expanded through a wave-of-advance mode. This expansion involved rice farmers who spoke Austronesian languages from southern China, around 6,000 years ago, and arrived to Taiwan around 4,000 years ago. The expansion continued into the Philippines, Islands Southeast Asia to reach eventually Remote Oceania around 3,000 ago and eastern Polynesian islands less than 1,000 years ago. Through their migrations, Austronesian speakers would have also replaced local populations who first settled the islands of Southeast Asia during the Pleistocene period.

Under the "Out-of-Taiwan" hypothesis, the first settlers of Remote Oceania, originating from Taiwan, would have brought with them a whole package composed of new technologies and navigation skills, horticulture and agriculture practices, the Austronesian language, and also their genes.

The "Slow-boat" and the "Triple-I" models

I will present here two alternative models that has been propose to the explain origin(s) of the Lapita Cultural Complex in the Bismarck Archipelago: the “slow-boat” model (Oppenheimer and Richards 2001; Kayser et al. 2000) and the “Voyaging Corridor Triple I” model (Green 2003).

The “slow-boat” model was initially proposed through the study of the genetic diversity of the Y chromosome, which is male-specific and uniparentally-inherited (Oppenheimer and Richards 2001; Kayser et al. 2000). This model stipulates that the Lapita Cultural Complex and the genetic makeup of the first Remote Oceanians would have emerged through intensive exchanges and gene flow between eastern Indonesians and non-Austronesian-speaking groups from the Bismarck Archipelago and the Solomon Islands starting between 6,000 and 3,500 years ago. This long process, which occurred in a voyaging corridor between Tropical Island Southeast Asia and Near Oceania would have been followed by a rapid peopling of Remote Oceania islands, around 3,100 years ago.

Kayser and colleagues in 2000 (Kayser et al. 2000) were the first to propose this model, or at least to give it a name (“[…] we propose a new model of Polynesian origin that we call the slow-boat model”), and placed the origin of the ancestors of Polynesians in "Asia/Taiwan" in agreement with the “Out-of-Taiwan” model, but a concomitant study (Su et al. 2000) also based on Y chromosomes, proposed an origin in Islands Southeast Asia rather than Taiwan, as also supported by other geneticists (Oppenheimer and Richards 2001) and some anthropologists and archaeologists (Terrell 2004; Torrence and Swadling 2008).

From 1991, the archaeologist Roger Green proposed another model named “Voyaging corridor Triple I” (Green 2003), which can be seen as an extension of the “slow-boat” model. Triple-I stands for intrusion, integration and innovation. With this model, Green hypothesized a diverse origin for the different components that characterize the Lapita Cultural Complex: some of them were introduced from Island Southeast Asia to the Bismarck Archipelago, some were innovations made locally by Lapita people in Near Oceania, and finally some elements of non-Austronesian speaking communities were incorporated into early Lapita groups. Consequently, under this model, the Lapita Cultural Complex, which includes the Austronesian language, horticulture and agriculture practices as well as genes, was not brought together in a single package, through a single migration wave.

The post-Lapita period in the Vanuatu

500 years after the initial settlement of the Vanuatu, the dented-stamped pottery disappeared and was replaced by other forms of pottery (Spriggs 1997). For example, in central Vanuatu, the Lapita pottery style was replaced by the Mangaasi style, which

is characterized by its reddish colour and incised and applied relief decorations (Spriggs 1997). In addition to the end of the Lapita pottery, the long-distance trade of obsidian also disappeared and the study of Vanuatu burials revealed a change in diet and funerary practices (Summerhayes 2010; Valentin et al. 2014). Secondary migrations occurring shortly after the initial settlement have been hypothesized to explain this cultural change observed in the Vanuatu. In line with this, a craniometric study carried out by Valentin et al. in 2016 (Valentin et al. 2016) indicates that Ni-Vanuatu from the Lapita period are morphologically closer to present day East Asians and Polynesians while the latest post-Lapita Ni-Vanuatu present a stronger affinity to present-day Australo-Papuans. The authors suggested that migrations from Papuan-related groups into Ni-Vanuatu can explain these observations, starting around 200 years after the initial settlement of western Remote Oceania (i.e. 3,200 years ago for the Vanuatu).

European contacts

When looking at the name of some islands of Vanuatu, it is apparent that Europeans played a role in the history of the Archipelago. In 1606, with the support of the Castilian Crown (King Philip III), Pedro Fernández de Quirós and his crew anchored at Big Bay on an island that he named Austrialia del Espiritu Santo (nowadays known as Espiritu Santo in central Vanuatu) in honour of the Spanish Habsburg monarch's Royal house of Austria. Despite limited but violent contacts with Ni-Vanuatu (e.g. kidnappings and beheadings), Quirós claimed this land in the name of the Spanish Crown, as well as the Catholic Church in order to "take Christianity to the heathens of the unknown Terra australis" (Luque and Mondragon 2005). Pedro Fernández de Quirós's voyage lasted around a year, including a month spent in Espiritu Santo. He visited different islands before entering the land of what is nowadays Espiritu Santo, such as Taumako in the Solomon Islands. Nevertheless, limited elements of the Oceanian cultures, languages and lifestyles can be drawn from this voyage. Perhaps because it was not the priority or because the relationship with others was different at that time: "the convoluted procedures and overall behaviour of the Spanish men in Big Bay were neither the result of one man's extravagant religiosity nor simply of Spanish arrogance, but encompass overlapping medieval, renaissance and (to a lesser degree) baroque legal and cultural canons which have hitherto been glossed in scholarly analyses of the earliest European explorations of Oceania." (Luque and Mondragon 2005).

It was only during the mid-18th century that scientists joined French and British expeditions to the Pacific. At that time, the Age of Enlightenment, both France and Britain were powerful expending empires that placed science at the centre of the society. It is crucial to say that the perception of human societies during the 18th century was stadial. Indeed, in the mid-18th century French and Scottish philosophers developed the stadial

theory also named the four stages theory, notably influenced by contacts with indigenous peoples from the Americas. According to this theory, societies go through four different stages or ages: (i) the age of hunters or savagery, (ii) the age of pastoralism or barbarism, (iii) the age of agriculture or civilisation and (iv) the age of commercial societies or Europe (Schorr 2018). In this context, Louis-Antoine de Bougainville from France (1768) and James Cooks from Britain (1774) undertook their voyage to the Pacific.

For the first time, observing, recording and collecting data were at the centre of the expedition. Contrary to Pedro Fernández de Quirós's voyage, Bougainville and Cook brought precious, though subjective, descriptions and information about the fauna and flora and Oceanian cultures, societies and peoples. They both indisputably contributed to major scientific breakthroughs (at least from the European point of view), notably in cartography, with the mapping of Pacific Islands, navigation and naturalism. In 1774, James Cook explored the islands of what is nowadays Vanuatu and named this archipelago New Hebrides.

European contacts and influences in the Vanuatu increased from 1839 with the beginning and intensification of Protestant and Catholic missionaries, first in the South of the Archipelago, mainly in Tanna, Aneityum and Erromango islands. The first contacts turned most of the time violent, with the murders of Europeans (e.g. John Williams and James Harris in 1839) at Dillon's Bay in Erromango island because Ni-Vanuatu rejected missionaries (Flexner and Spriggs 2015). Instead of discouraging Europeans, missionaries reached their height with the idea of bringing "light to the dark isles" (Flexner and Spriggs 2015) in a region of the world peopled by "savages" who used black magic and cannibalistic rituals (Copeland 1866).

Overall, the process of conversion was long, especially outside the New Hebrides (Vanuatu) because of the non-acceptance of Europeans missionaries and strengthened by competition between Catholic and Protestant missionaries. In the New Hebrides, both coexisted but with different ways of converting Ni-Vanuatu. Anglican missionaries adopted a strategy that I would personally name "from the inside": young people were taken from a location (e.g., a Vanuatu island), placed in schools located in another place (e.g. in New Zealand), and were then placed back in their original communities, to convert their relatives. Catholics adopted a strategy "from the outside", where they preferred to maintain a permanent presence at different strategic places to convert most of the communities, involving some recently converted Polynesians in the process (Flexner 2013).

In Melanesian practices, referred also as kastom (pidgin word for custom), spirituality, the supernatural and thus religion is part of the Melanesian identity and the daily life,

including politics and economics. This link between religion, identity and daily life choices is so tight that missionaries failed to deeply change Ni-Vanuatu practices, but local communities incorporated elements of Christianity into kastom, mainly material things: traditional dresses were replaced by imported European clothes, pottery was in part locally replaced by iron cooking vessels and the most prized item in 1860 in Northern Vanuatu was empty bottles. Missionaries also impacted the daily life and societies of local communities by, for example, setting labour tasks and changing the gender role (toward a male authority versus female domestic tasks) (Flexner 2016; Bedford and Spriggs 2008). This incorporation of the Christianity into kastom is still visible nowadays, as attested by James L. Flexner and Matthew Spriggs: "many Ni-Vanuatu still see supernatural causes at work in instances of illness or death" (Flexner and Spriggs 2015).

Social structure was also impacted by the arrival of western traders of sandalwood who, after having exhausted sandalwood resources, traded young Ni-Vanuatu men to work in sugar plantation in Australia, Fiji and New Caledonia, a practice referred as blackbirding (Docker 1970). One major consequence of this trade was a massive depopulation of Vanuatu islands, coupled with an increased mortality due to European diseases transmitted to local populations (e.g. measles, influenza and cholera) (Flexner 2016). In the mid-19th century, ca. 5,000-7,000 individuals peopled the island of Erromango in South Vanuatu (Gordon 1863) while Colley and Ash estimated a population of ca. 600 inhabitants in 1967 (Colley and Ash 1971).

From the 20th century onwards, the New Hebrides became an Anglo-French condominium (1906) and played a strategic role during the World War II, notably with the presence of American soldiers to prevent Japanese army from gaining a foothold after the attack of Perl harbour in 1941. The New Hebrides obtained their independence in 1980 and the archipelago was renamed Vanuatu (Vanua "land" and tu "be independent") by local communities (Flexner 2016). Nowadays 86% of Ni-Vanuatu are Christians (Vanuatu National Statistics Office) and a part of them, in southern islands (TAFEA province) consider missionary sites as being part of their culture, history and heritage (Flexner and Spriggs 2015): "In our fieldwork experiences, we have found that people will unironically express their sincere Christian faith, and then invite visitors for a traditionally prepared shell of kava".

Chapter 2

Using evolutionary genetic approaches to learn about human demographic history

Archaeology, anthropology and linguistics have provided valuable insight into the peopling history of Near and Remote Oceania and the lifestyles of the different populations that settled these regions. However, cultural, linguistic and genetic studies do not always tell the same story and all have their own limitations. For example, because populations tend to move, genetic continuity between past and present-day groups of a region is not necessarily observed. Consequently, the different demographic events estimated with genetic approaches would not reflect the population history of the initial ancestral

population. Cultural practices and the language can be transmitted not only vertically (from one generation to another), as genes, but can also be transmitted horizontally, through the transmission of ideas (Diamond and Bellwood 2003). Another important point is the fact that population geneticists study the dynamics of genetic interactions between populations, while current archaeological research tends to focus on internal changes rather than the impact of movements of people (Veeramah 2018). Hence, the information provided by genomic data is complementary to other disciplines and can add another dimensionality to the history of a population.

For decades, evolutionary genetic approaches, combined with the rapid and dramatic progress of sequencing technologies and methods, have allowed the detailed reconstruction of human population history, such as the estimation of populations size changes over time, divergence time, admixture, introgression events with now extinct hominins and events of genetic adaptation to new environments (Dannemann and Racimo 2018; Gosling and Matisoo-Smith 2018b; Marchi, Schlichta, and Excoffier 2021; Patin and Quintana-Murci 2018; Rotival, Cossart, and Quintana-Murci 2021). We will see in this chapter how genomes are used to trace back the demographic history of human populations and what genomes of Oceanian groups revealed about their past history.

2.1 Theory and basic principles of population genetics

2.1.1 Variations in the human genome: mutations and recombination

The DNA (deoxyribonucleic acid) is a macromolecule found within cells and composed of linked nucleotides that are commonly represented by four letters. A nucleotide is composed of a sugar, a phosphate group and a nitrogenous base. Four canonical nucleotides are found in the DNA: Adenine (A) and Guanine (G) are the two purines and Cytosine (C) and Thymine (T), the two pyrimidines. Billions of linked nucleotides form the DNA sequence (Watson and Crick 1953). In humans, less than 3% of the DNA contains genes that encode proteins (Dunham et al. 2012).

Genetic mutations, i.e., changes in the DNA nucleotide sequence, result in different versions of a same genetic position (i.e. a locus), the alleles, that segregate in the population. A mutation that occurs in a gene and that changes its final product, the protein amino-acid sequence, is named a non-synonymous mutation. On the contrary, a genic mutation that does not change the protein sequence is called a synonymous mutation. From an evolutionary perspective, molecular evolution corresponds to the changes in frequency through time of the different alleles that constitute the genetic diversity of a specific population or group. Only mutations located in the DNA of reproductive cells (germinal mutations) are transmitted to the next generation and participate to the genetic diversity of a population.

Mutations are divided into three classes based on the number or the size of the modification: substitutions (point mutations), insertions and deletions and chromosomal rearrangements. I will focus here on point mutations because they are the most frequent and are broadly used in the population genetics field. A substitution, also named single nucleotide polymorphism (SNP), corresponds to the modification of a single position (one nucleotide) of the DNA owing to either an error during the DNA replication or errors introduced by the DNA maintenance machinery while fixing physical or chemical alterations (e.g. UV exposure). In the human genome, transitions (i.e., the change of a purine (pyrimidine) by another purine (pyrimidine)) are observed at least twice as more as transversions (i.e., the change of a purine (pyrimidine) by another pyrimidine (purine)). The rate of substitutions per site and per generation is expected to be 10^{-8}, but the mutation rate can go up to 10^{-5} substitutions per site and per generation depending on the genomic region (e.g. CpG sites) (Campbell et al. 2012; Lipson et al. 2015; Walser, Ponger, and Furano

2008; Seplyarskiy and Sunyaev 2021).

The development and improvement of sequencing technologies allow to obtain the whole sequence of the individual's DNA. Among the 3.5 billion positions in the human genome, around 4 millions are polymorphic between two individuals, i.e., 1 substitution variation is expected every 1,000 positions between two randomly chosen individuals (Genomes Project et al. 2015; International HapMap 2003; Karczewski et al. 2020; Bergstrom et al. 2020). The dbSNP, database of single nucleotide polymorphisms and short insertions/deletions counts more than 683 million variants detected in world-wide human populations (2019, build 153, (Sherry et al. 2001)).

Humans have two sets of chromosomes inherited from their two parents (one set per parent). During the creation of reproductive cells, also named gametes (i.e sperm and oocyte), homologous chromosomes align and pair with each other through the formation of DNA junctions that result in the exchange of the genetic information, a process referred to as meiotic **recombination**. Recombination generates at each meiosis unique combinations of alleles, called haplotypes, which are different from the combinations of alleles inherited from the parents (Figure 2.1). The number of recombination events per generation between two given positions of a chromosome, i.e., the recombination rate, increases with increasing chromosomal distance between the two positions. Two SNPs of the same chromosome are said to be in linkage disequilibrium when the recombination rate between these SNPs is low. As a result, alleles are not transmitted independently to the next generation but rather in blocks where genetic recombination is low. Hence, the frequency of a mutation depends on the frequency of other mutations located on the same haplotype. Recombination tends to dissociate mutations found in a same genetic region (i.e. decrease of the linkage disequilibrium). In addition to create haplotype diversity, recombination, through time, also tends to break long haplotypes into smaller ones (Figure 2.1).

The recombination rate, like the mutation rate, varies greatly along the genome, characterized by "hotspots" and "coldspots" of recombination. This rate depends on the genomic context, such as the percentage of G and C nucleotides, the number of transposable elements in the region and the presence of binding sites for PRDM9, a DNA-binding protein that promotes recombination (Genomes Project et al. 2015; Myers et al. 2005).

To summarize, the genome is organised in blocks or haplotypes made of alleles in high linkage disequilibrium. Each haplotype block is separated by hotspots of recombination. The mutation and recombination are critical events that create genetic variations in a given population. This variability constitutes a substrate on which other evolutionary forces can

act.

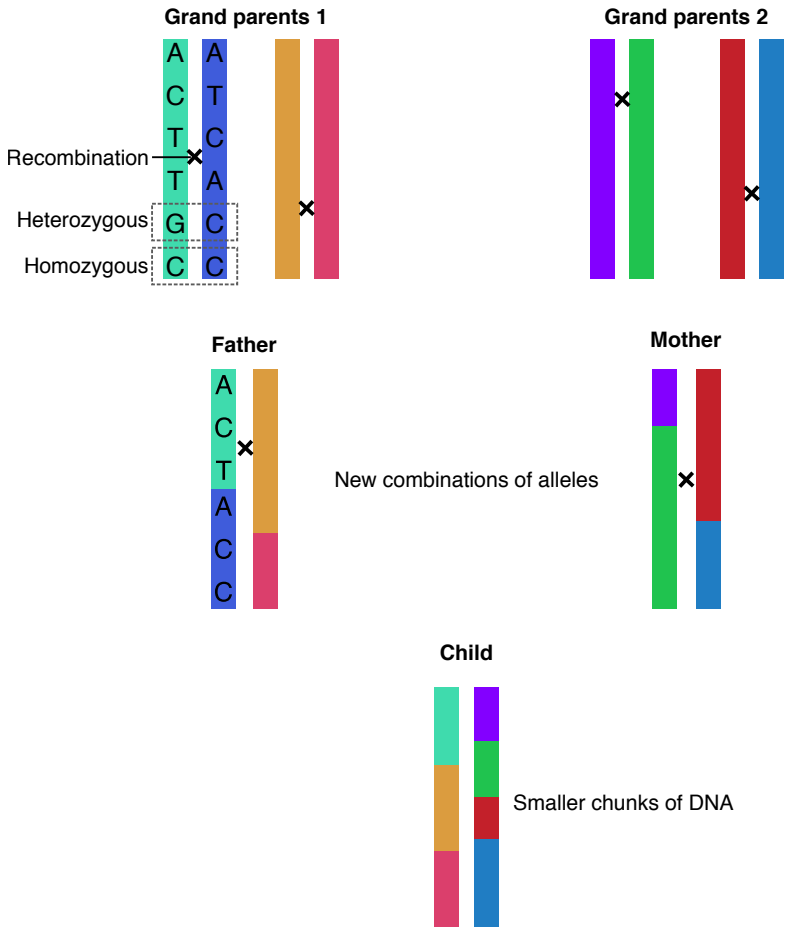


Figure 2.1: **Genetic recombination**. Schematic representation of recombination events and its impact on the size of the haplotypes (coloured bars) through times (two generations).

2.1.2 Demographic history: genetic drift, gene flow and isolation

Among all mutations found in a genome, most are neither beneficial nor deleterious for the carrier, meaning that they do not have any phenotypic consequences. Such mutations are said to evolve under **neutrality**, i.e., in absence of natural selection (chapter 3). If the population size is small, a stochastic process known as **genetic drift** drives their frequency over time (Kimura 1991; Wright 1931). Allele frequencies at generation g+1 are different from those at generation g, because siblings are a non-representative sample of the adult population. Under genetic drift, the probability of a neutral mutation to be fixed in a population (or to be eliminated) depends on its initial frequency. Considering an isolated population of effective size N_e (i.e. number of individuals who contribute genetically to the next generation), a new neutral mutation that appears in this population has a fixation

probability equal to $1/(2N_e)$ and takes an average of $4N_e$ generations to reach fixation (Hartl and Clark 2007). Consequently, in a population with large N_e, the strength of the genetic drift is weak, leading to a stability of allelic frequencies during a long period of time. On the contrary, in small populations (small N_e), the genetic drift is stronger causing sharp allele frequency variations from one generation to another.

Under neutrality, levels of genetic diversity of a population, θ, is proportional to the effective population size: $\theta = 4N_e\mu$, μ being the mutation rate. Changes in population size that occur over time, i.e., the demographic history, have ultimately an impact on the genetic diversity of a given population, by affecting the strength of genetic drift. Hence, human populations that experienced contrasting demographic histories, such as bottlenecks, founder effects, population expansions and contractions, differ in their current levels of genetic diversity.

Human populations can also exchange migrants, resulting in gene flow. Contrary to genetic drift, gene flow reduces levels of genetic differentiation between populations. Populations that are geographically closer tend to be less genetically differentiated, because of more recent divergence and/or substantial gene flow between them. The isolation-by-distance model explains this correlation between genetic differentiation and geographic distance (Wright 1943; Cavalli-Sforza and Feldman 2003).

To sum up, the levels of genetic differentiation between human groups depend on their demographic history affecting the strength of genetic drift, as well as their level of genetic isolation. The resulting distribution of genetic variation within and between populations is called the genetic structure of human populations.

2.1.3 Different types of genetic data

Each human cell includes 22 homologous pairs of autosomes, which are inherited from each parent, one pair of sex chromosomes (i.e., a maternal X and a paternal X or Y chromosomes), and a mitochondrial genome (mtDNA), a small circular genome of 16,569 base pairs found in mitochondria. Genetic markers on these different chromosomes can be used to reconstruct different aspects of the demographic past of human populations. Indeed, because mitochondria are only transmitted from the mother to the child, the study of mtDNA mutations is thus used to trace back the maternal lineages. Similarly, the study of mutations on the Y chromosome, which is only carried by men trace back the history of the male lineages. Historically, uniparentally-inherited chromosomes were broadly used to reconstruct the demographic history of human populations because they are short, thus easy to sequence and non-recombining.

The genetic information, can be retrieved using a variety of technologies such as genotyping or SNP arrays, which capture the information at pre-ascertained SNPs, and next generation sequencing (NGS), for instance whole genome sequencing (WGS), which provides the information for the full DNA sequence. Markers found in SNP arrays are either (i) old variants that segregate at intermediate or high frequency in different human populations or (ii) SNPs that segregate mainly in well-studied populations (e.g. Europeans). This ascertainment bias can lead to an underestimation of the genetic diversity, mainly because of the exclusion of rare and private mutations. The WGS strategy gives access to mutations that segregate at very low frequencies and thus reduces the ascertainment bias. However, whole genome sequences are obtained at a higher cost, around $1,000 for one individual, limiting the sample size of populations to study.

Consequently, the choice of the strategy to access the genetic information depends on the scientific questions, the methods and the study populations (and the budget): do the sampled individuals belong to under-studied populations? Are the methods that I want to apply sensitive to the number of rare variants? Can I correct for the ascertainment bias? Do I need a large sample size or/and a high variant density?

2.2 Demographic inference

We saw in the previous sections how variations in effective population size changes and gene flow or admixture affect patterns of neutral genetic diversity in the genome of human populations. We will see now how to infer the demographic parameters that characterize their demographic past.

2.2.1 The coalescent theory

A pair of alleles sampled in present-day individuals is made of an ancestral form and a derived form that appeared in the population some generations ago. All current copies of the derived allele are thought to descend from the same mutational event in the past, and to have thus been inherited from the same common ancestor. The coalescent theory (Kingman 2000) describes how the alleles observed in a sample may have originated from a common ancestor, called the most recent common ancestor (MRCA). Looking backward in time, the coalescent model is a random process that merges the two copies of an allele at a given generation into one ancestral copy inherited from the previous generation (i.e., a coalescence event). Considering not only a pair of alleles or genes but a subset of the observed current genetic diversity of a population, the coalescent theory

can estimate the topology of the gene genealogy. This topology is then used to estimate demographic parameters such as the divergence time between human populations, or changes in effective population size trough time (Rosenberg and Nordborg 2002).

Assuming a panmictic (i.e. random mating) and isolated population and in absence of recombination and natural selection, the probability that two lineages coalesce in the previous generation is $1/(2N_e)$. The probability that these two lineages did not coalesce in the previous generation is $1-1/(2N_e)$. Hence, the probability that two lineages coalesce at generation *t* is given by the following geometric distribution: $Pc(t) = (1-\frac{1}{2N_e})^{(t-1)}(\frac{1}{2N_e})$. Considering k allele copies, the probability that at least two allele copies coalesce in the previous generation is $\frac{k(k-1)}{4N_e}$. The time to first coalescence thus follows a geometric distribution with $E[T_k] = \frac{4N_e}{k(k-1)}$ generations, which indicates that the first coalescence event in a sample of *k* allele copies occurs on average $\frac{4N_e}{k(k-1)}$ generations ago. The expected time to the MRCA, meaning the time for which all lineages coalesce into a unique ancestral allele is equal to $E[T_{total}] = 4N_e \sum_{i=2}^{k} \frac{1}{i-1}$.

In case of a population collapse, the genetic diversity decreases, thus lineages coalesce (backward in time) more rapidly in the population, which leads to an acceleration of the coalescence rate. On the contrary, with a recent population expansion, a large fraction of the genetic diversity is composed of new variants that are carried by only few samples (e.g. singletons) and there are more genetic lineages because of a higher effective population size. Therefore, under a recent expansion, the probability that two alleles coalesce decreases leading to a slowdown of the coalescence rate.

2.2.2 Joint estimation of demographic parameters

Methods implemented in software such as PSMC (Li and Durbin 2011), which estimates effective population size changes in time and divergence times, or GLOBTROTTER (Hellenthal et al. 2014), which infers the time of admixture events, are used to estimate simple demographic models. They usually estimate a demographic parameter of interest, assuming that all other parameters are constant or null. For example, PSMC (Li and Durbin 2011) estimates the effective population size through time, without considering gene flow between populations. However, an increase in effective population size could be due either to (i) an increase of the census size or (ii) an increase in the rate of new incoming migrants (admixture). Furthermore, population stratification, meaning a population that is composed of sub-groups that exchange varying levels of gene flow, can produce a signal of bottleneck (Walhund effect, (Nielsen and Beaumont 2009; Chikhi et al. 2010)). Similarly, tools that estimate the time of admixture assume a constant effective population

size of the parental and the admixed populations. These methods work better when admixture occurred less than 5,000 years ago. They are able to estimate older events but with less accuracy because, with time, ancestry segments are shorter and thus more difficult to detect. These methods also estimate admixture times less accurately when admixture is continuous rather than discrete, especially if the events are recent (Hellenthal et al. 2014; Pugach et al. 2018b). Indeed, it is usually assumed that genetic interactions between groups occurred in pulses, meaning very rapid and short contacts that lasted one generation. However, when working with human populations, we expect to observe more complex modes of gene flow, such as continuous and repeated genetic interactions between two or several human groups. Although these algorithms are robust to the violation of some assumptions, depending on the populations studied and the questions that are addressed, the interpretation of the results can be convoluted and/or limited.

One way to overcome these issues is to jointly estimate the parameters characterizing the demographic history of the studied populations using simulations. The parameters can be inferred using different statistical frameworks such as the maximum likelihood framework, which searches for the set of parameters that best explain the observed data via maximization of a likelihood function (Gutenkunst et al. 2009; Excoffier et al. 2021) or an approximate Bayesian Computation Approach (ABC) (Beaumont, Zhang, and Balding 2002; Cooke and Nakagome 2018), which relies on the comparison of observed and simulated genetic data, in the form of summary statistics, to estimate demographic parameters.

Depending on the scientific questions that are addressed, only a subset of the parameters that characterized the demographic history of the studied populations can be estimated. The other parameters are referred to as "nuisance parameters" because they are not estimated but instead, they are just considered (they are allowed to vary) in order to not bias the inference of the parameters of interest (e.g., considering gene flow between two populations to not bias the estimation of their divergence time and their effective population size).

Site frequency spectrum and the maximum likelihood framework

The site frequency spectrum (SFS) or allele frequency spectrum (AFS) corresponds to the distribution of allele frequencies in a given population. The SFS can be obtained using frequencies of either the derived allele ("unfolded" SFS) or using the minor allele ("folded" SFS). It is also possible to compute the SFS for more than one population at the same time through joint or multidimensional SFS (Excoffier et al. 2013).

The different demographic events that populations experienced have an impact on the

shape of the SFS. For example, under a recent expansion, the effective population size increases, which inflates the number of rare mutations segregating in the population. Under such a scenario the SFS is characterized by an excess of rare mutations and a deficit of fixed mutations compared to what is expected under a stationary demography. Conversely, a drop in effective population size - due to a population contraction or a founder event - increases the strength of genetic drift, which leads to a greater fixation or elimination of alleles at low and intermediate frequencies, as well as a global loss of genetic diversity. Under such a scenario, the SFS is characterized by an excess of fixed mutations and a deficit of rare mutations compared to what is expected under stationary demography. When investigating the demographic history of human populations, it is paramount to compute the SFS using only neutral mutations, or the least selected mutations (as neutral as possible), usually by removing mutations inside genes, since natural selection can mimic the impact of demography on the shape of the SFS.

The inference of demographic parameters using the SFS can be done, for example, using the maximum likelihood framework and coalescent simulations implemented in the *Fastsimcoal2* tool (Excoffier et al. 2021; Excoffier et al. 2013). This algorithm estimates the likelihood of the observed SFS (one-dimensional, joint or multidimensional SFSs) given the expected SFS generated under a set of demographic parameters. These expected SFSs are approximated from a number of coalescent simulations provided by the user (usually > 100,000 simulations). The algorithm starts from initial values of the parameters taken randomly from a distribution. Then, through a series of cycles, the algorithm calculates the likelihood for different parameter values to finally find the set of parameters that maximizes the likelihood. This algorithm needs to be repeated several times (i.e. multiple runs) starting from different initial values to ensure that the likelihood converges toward the global maximum of the likelihood function and not just to local maxima.

However, depending on the complexity of the model tested (i.e. the number of demographic parameters to infer), the algorithm may not converge. When inferring demographic parameters with this method, the user should thus take a step-by-step approach, starting with very simple models and make them more and more complex. It is also important to run additional tests to check for the robustness of the inference, e.g., by increasing both the number of simulations used to approximate SFS or increasing the number of runs and cycles.

Approximate Bayesian Computation

Approximate Bayesian Computation approaches (Beaumont, Zhang, and Balding 2002; Cooke and Nakagome 2018), based on the Bayesian statistical framework, are often used

to estimate parameters of models for which the likelihood function is too complex to be evaluated. This framework relies on the simulation of genetic data under different demographic models, such as different topologies of the population tree, and values of the corresponding parameters (prior distributions). Summary statistics are then computed from the simulations and compared with the observed summary statistics. The comparison is classically done through the calculation of a distance between observed and simulated summary statistics. Nowadays, new approaches can be used to compare simulated and observed data such as machine learning approaches. The closest simulations are then used to compute posterior distributions of demographic parameters to estimate.

The summary statistics should be tested a priori because not all are informative for estimating demographic parameters. For example, the SFS and derived summary statistics (e.g. Tajima's D, θ_w, θ_π) have been shown to be informative to infer effective population size and divergence times (Cooke and Nakagome 2018; Fagundes et al. 2007; Veeramah et al. 2012) while the length of haplotypes is informative to date events of admixture (Gravel 2012; Liang and Nielsen 2014). The evaluation of the summary statistics as well as the accuracy, the sensitivity and the specificity of the ABC are essential but computationally demanding analyses.

2.3 What did the genomes of Pacific islanders reveal about their history?

2.3.1 Deep population structure of Near Oceania

The first genetic studies of Near Oceanians were mainly based on a subset of genetic markers contained in the hypervariable regions of the mitochondrial DNA (mtDNA) (Redd and Stoneking 1999; Huoponen et al. 2001; Betty et al. 1996; Stoneking et al. 1990). These studies unravelled the deep coalescent age of Australian and New Guinean lineages, which was interpreted as evidence for multiple settlements of Sahul, followed by a rapid genetic isolation between groups. However, the complete sequencing of the mitochondrial DNA showed that northern and southern Sahul, corresponding to current New Guinea and Australia respectively, were settled by a common founder population dated back to 50,000 years ago (Hudjashov et al. 2007). More recently, Pedro and colleagues (Pedro et al. 2020), based on 379 whole mtDNA sequences of Australians and Near Oceanians, argued for at least two concomitant waves of settlement around 50,000 years ago, through two different routes (northern and southern routes), followed by a period of 20,000 years of

genetic isolation. The study of Y chromosome variations also indicated deep population structure and old divergence times within Near Oceanians (Kayser 2010; Bergstrom et al. 2016).

Wollstein and colleagues in 2010 (Wollstein et al. 2010), followed by Malaspinas et al. in 2016 (Malaspinas et al. 2016a), provided the first genetic demographic models of Oceanians using an autosomal genotyping and a whole-genome sequencing strategy, respectively. In the latter study, the authors co-estimated effective population size changes and divergence time between Oceanians and non-Oceanian groups assuming a model of isolation followed by migrations. Using coalescent simulations and the maximum-likelihood framework, they found that present-day Australians and New Guineans derived from the same Out-of-Africa migration as their Eurasian neighbours, dating back to around 60,000-104,000 years ago. They also estimated that all present-day Australians derived from the same ancestral population, suggesting a unique wave of settlement for southern Sahul. Finally, they dated an old divergence time between Australians and New Guineans, around 20,000 to 45,000 years ago which points toward a deep structure of Sahul populations.

2.3.2 Which model for the peopling of Remote Oceania?

Archaeologists, anthropologists and linguists proposed different scenarios for the origin of the proto-Lapita and the first Bismarck Lapita societies. Although the archaeological data point to the Green's Triple-I model (Green 2003), a Taiwanese versus Island Southeast Asian origin for the proto-Lapita is debated (Gray, Drummond, and Greenhill 2009; Terrell 2004; Torrence and Swadling 2008). Does genetics also favour Green's Triple-I model (Green 2003)? Do genomic studies point to a specific geographic area for the origin of the proto-Lapita? Are the Lapita people entering Remote Oceania already admixed?

Using animals and plants to trace back population movements in the Pacific

In the 1990s, strengthened by issues in obtaining DNA samples from Oceanian individuals, Lisa Matisoo-Smith proposed a new approach to trace back the migration routes taken by Oceanian seafarers: the use of the DNA of animals and plants they transported with them (Matisoo-Smith 1994; Matisoo-Smith et al. 1999; Matisoo-Smith 2015). This approach is referred to as the "commensal model".

The study of mtDNA variation of the Pacific rat (Rattus exulans) (Matisoo-Smith and Robins 2004), supports the Triple-I model for the origin of Lapita cultural complex and discards the "Express-Train" model and the "Bismarck Archipelago Indigenous Inhabitants" model, which stipulates that the Lapita cultural complex emerged locally

from the Bismarck Archipelago without any migration wave from East/Southeast Asia. Similarly, the study of ancient pig bones, using both ancient DNA and morphometry, placed the origin of Oceanian pigs in mainland Southeast Asia (coast of Vietnam). This study also revealed that Oceanian pigs are not closely related to present-day pigs from China, Taiwan and the Philippines, which suggests that the spread of the Austronesian languages from Taiwan ("Out-of-Taiwan" model) does not correlate with the movement of pigs in Oceania (Larson et al. 2007). On the other hand, a recent genetic study of the paper mulberry (Broussonetia papyrifera) used for textile production, indicates an exclusive Taiwanese origin of this plant (Olivares et al. 2019). Other commensal plants and animals have been studied such as the dog, the chicken or the taro (Zhang et al. 2020; Thomson et al. 2014). Altogether, these different studies attest of a diverse origin of the different domesticated animals and plants transported by the first settlers of Remote Oceania and suggest multiple migrations and complex interactions between East/Southeast Asians and Oceanians.

Y chromosome and mtDNA tell a different story

Early works based on uniparentally inherited genetic markers shed light on a specific set of four mutations in the control region of the mtDNA that characterized the haplogroup "B4a1a1a", also known as the "Polynesian motif" (Sykes et al. 1995; Melton et al. 1995; Redd et al. 1995). This mtDNA haplogroup is found at very high frequency in Polynesian groups and is also present in Micronesia and in Near Oceania, mainly in the Bismarck Archipelago. Although the Polynesian motif is absent in Taiwan, the Philippines and China, related B4 lineages are found in these three regions.

The geographic distribution of the Polynesian motif was initially interpreted as in favour of the "Express-Train" and "Out-of-Taiwan" models, to explain the origin of Polynesians (Redd et al. 1995). However, Richards et al. in 1998 (Richards, Oppenheimer, and Sykes 1998), combining the geographic distribution with the estimated age of the Polynesian motif and founder events, proposed an alternative interpretation: the Polynesian motif originates from Island Southeast Asia between 5,500 years ago and 34,500 years ago, before the arrival of Taiwanese farmers in Indonesia around 4,000 years ago. Soares et al. (Soares et al. 2011), through the study of the full mtDNA sequence of 157 Pacific islanders, argued that the so-called Polynesian motif arose around 6,500 years ago, before the Lapita period and likely within the Bismarck Archipelago. The motif then spread westward to Islands Southeast Asia around 5,000 years ago and eastward to Remote Oceania around 3,500 years ago. Although the authors rejected both a Taiwanese and an Island Southeast Asian origin of the ancestors of Polynesians, they hypothesized a model of non-demic diffusion of Austronesian languages (here diffusion of the language with very limited population

movements) from Taiwan to other Pacific islands from around 4,000 years ago.

Y chromosome lineages (NRY) of Remote Oceanian islanders are mainly of Papuan-related origin (K, M, S and C NRY branches, (Kayser 2010; Mona et al. 2007; Scheinfeldt et al. 2006)), but also of East Asian-related origin, such as the O lineages (Kayser 2010). This suggests an appreciable contribution of Papuan-related groups to the ancestors of Remote Oceanians ("slow-boat" model). The discrepancy between East Asian maternal markers and Papuan-related paternal markers (mtDNA versus NRY) has been interpreted as sex-specific migrations: women of East Asian-related ancestry migrated and admixed with local Papuan-related men. Hage and Marck in 2003 (Hage and Marck 2003), attributed this difference between maternal and paternal markers to the effect of matrilocal residence and matrilineal descent structure of Lapita societies (Jordan et al. 2009).

New insight into the settlement of Remote Oceania using ancient DNA and autosomal markers

The genetic studies based on autosomal markers (microsatellites and SNPs) of modern Oceanian individuals confirmed the admixed nature of some Austronesian-speaking groups from Near and Remote Oceania. They also show that Polynesian groups present the highest level of East Asian-related ancestry, around 80%, with only 20% of Papuan-related ancestry supporting an East/Southeast Asian origin of the proto Lapita people (Wollstein et al. 2010; Friedlaender et al. 2008). The date of this admixture was estimated to occur around 3,000 years ago, using different methods (Pugach et al. 2018b; Wollstein et al. 2010; Pugach et al. 2011). Taken together, these studies strengthened the view that the Lapita people admixed first in Near Oceania before entering and peopling the pristine islands of Remote Oceania.

However, in 2016, scientists from the Harvard Medical school published for the first time the ancient DNA sequence of three individuals from the Vanuatu and one from Tonga dating to the Lapita period (Skoglund et al. 2016). Surprisingly, this study revealed that the initial settlers of Remote Oceania were of almost complete East Asian ancestry, as also suggested by craniometric data (Valentin et al. 2016). Based on these results, the authors suggested that the first people to migrate to Remote Oceania did not mix with Near Oceanian Papuan-related groups, as previously thought. The authors suggested that the Papuan-related ancestry observed in modern individuals reflect more recent or post-Lapita migrations of Papuan-related groups to Remote Oceania.

This hypothesis was confirmed by two ancient DNA studies (Posth et al. 2018; Lipson et al. 2018) that generated a time-transect dataset composed of Lapita and post-Lapita individuals from different islands of Vanuatu. These studies point towards a secondary

wave of settlement after the initial settlement of Remote Oceania, albeit Lipson et al. (Lipson et al. 2018) estimated a non-zero proportion of Papuan-related ancestry in some Lapita individuals. These two ancient DNA studies also suggest that the second settlement occurred before the end of the Lapita period, in the late Lapita period around 2,700 years ago and that the Papuan-related groups involved in the event are closer to group that live nowadays in the Bismarck archipelago.

Pugach et al (Pugach et al. 2018b), using a SNP array dataset composed of 823 Pacific individuals, found that the peopling of Remote Oceania did not follow a simple linear wave-of-advance scenario as suggested by the "Out-of-Taiwan" model. By comparing the level of Bismarck-related ancestry between groups, they found that populations from Santa Cruz Islands were closer to populations from the Bismarck Archipelago than to any other Solomon islanders, in agreement with previous genetic analyses based on mtDNA (Duggan et al. 2014). Pugach and colleagues thus suggested that the peopling of Remote Oceania occurred in a "leapfrog" manner, bypassing most of the Solomon Islands. This "leapfrog" hypothesis was first proposed by Peter Sheppard in 2011 to explain the absence of Early Lapita pottery in the archaeological record of western and central Solomon Islands as well as the presence of Bismarck obsidian only in Santa Cruz islands (Sheppard 2011).

Chapter 3

Implication of the population history on health and diseases in Oceania

In the previous chapter, I briefly described the molecular signatures left by genetic drift and gene flow – and thus the demographic history – on the patterns of genetic variation. I also evoked how we can use the genetic data to infer the different demographic parameters characterizing the population history of human groups. I will now describe another evolutionary force, natural selection. Four evolutionary forces shape the genetic diversity of a population: the mutation that creates the genetic variation, the genetic drift that tends to increase genetic differentiation of small populations, migrations or gene flow that homogenises populations (Chapter 2) and natural selection that (i) allows populations to adapt to their environments (i.e. positive natural selection) or (ii) purges deleterious

mutations (i.e. negative selection). In this chapter I will focus on both negative and positive natural selection.

The effective population size determines the strength of genetic drift acting on the genomes of populations. However, in theory, due to the codon degeneracy (i.e. redundancy of the genetic code), around 1/3 of new mutations in genes are expected to be synonymous and around 2/3 non-synonymous. A large fraction of new mutations that arise in human genes have thus the potential to reduce the fitness (i.e. the survival probability and the reproductive success) of individuals that carry these mutations and contribute to disease susceptibility. In small populations, one expects the frequency of such mutations to be under the control of genetic drift, and can thus theoretically be found at intermediate frequency even if they are deleterious. Therefore, understanding the joint effects of demographic history and natural selection on deleterious mutations appears crucial to better understand the between-population differences in the susceptibility to common and rare diseases.

3.1 The burden of deleterious mutations in humans

Strongly deleterious mutations such as mutations that appear in genes involved in fundamental developmental processes, but also in functions such as innate immunity, are under strong purifying selection (Quintana-Murci 2019; Quintana-Murci and Clark 2013) and are therefore rapidly eliminated from the population. However, a large fraction of deleterious mutations corresponding to weakly deleterious mutations are eliminated (also by purifying selection) at a slower rate and can persist for some generations in the population.

3.1.1 The link between demography and efficacy of natural selection

Each individual carries thousands of deleterious mutations, most of them in the heterozygous state, that have not yet been eliminated by natural selection. The rate at which deleterious mutations are fixed in the population is named the efficacy of natural selection and increases with the product $N_e s$, where N_e is the effective population size and *s* the selection coefficient, which measures relative change in fitness conferred by mutations (Charlesworth 2009). Mutations with a selection coefficient lower than $1/N_e$, are considered to be "nearly neutral" mutations, meaning that the frequency of such mutations fluctuates in the population following random genetic drift expectations. Hence, a same mutation that would be quickly eliminated by natural selection in large populations (large N_e) can reach intermediate frequency - although deleterious - in small populations (small N_e).

To sum up, in theory, deleterious mutations have more chances to increase in frequency and reach fixation in populations that experienced strong founder events or bottlenecks, where the efficacy of natural selection to remove deleterious mutations is expectedly lower. On the contrary, populations that experienced a recent expansion would have a higher efficacy of natural selection, but more rare, deleterious mutations that recently appeared in the population.

3.1.2 The genetic load

The genetic load (L) measures the reduction in fitness of an average genotype found in a population compared to the maximal or optimal fitness, which by convention is set to 1: $L = \frac{W_{max} - \widehat{W}}{W_{max}}$ where W_{max} is the optimal fitness and $\widehat{W}$ the mean fitness

of the individual. The main factor contributing to the genetic load is attributed to the reduction in fitness that is due to the accumulation of deleterious mutations in genomes, also known as mutational load. The genetic load includes other elements such as the inbreeding load, which corresponds to the increase in the number of recessive mutations in the homozygous state carried by the children of consanguineous marriages increasing inbreeding depression (i.e.,increase of genetic load because of parental relatedness)

When fixing W_{max} to 1 in the equation above $L = 1 - \widehat{W}$. Assuming one deleterious mutation that segregates at frequency *p* reducing the fitness of carriers by *s* in the homozygous state and by *hs* in the heterozygous state we obtain $L = 2p(1-p)hs + p^2$, with *h* the dominance coefficient, i.e., relationships between alleles and their effect on the phenotype (dominant/recessive). Nevertheless, in humans, it is not possible to measure the fitness of individuals and very little is known about the distribution of the dominance coefficient *h*. To circumvent these issues, one can use empirical proxies (i.e. measures or statistics derived from empirical data), to evaluate the mutational load in different human populations and consider that all deleterious mutations follow the same model of dominance. In this thesis, I will refer mainly to two models of dominance: either all mutations are under a semi-dominant model, also named additive model, where the heterozygous carriers have an intermediate phenotype ($h = 0.5$), or under a recessive model, where the reduction in fitness is only seen when the deleterious allele is in the homozygous state ($h = 0$).

3.1.3 Approximation of the mutational load in human populations

Different statistics have been proposed to approximate the mutational load in human populations (Lohmueller 2014) such as the ratio of non-synonymous/synonymous mutations (P_n/P_s) (Lohmueller et al. 2008; Henn et al. 2015a) or the number of heterozygous and derived homozygous genotypes per individual (Lohmueller et al. 2008). In 2016, Simons and Sella (Simons and Sella 2016) found that under an additive model of dominance, the number of derived alleles ($N_{alleles} = 2N_{homozygous} + N_{heterozygous}$) carried by individuals is the only statistic that directly correlates with the mutational load and is not biased by demographic events such as bottlenecks (Figure 3.1). Based on this statistic, the same authors found that recent demographic events did not significantly impact the load in humans, meaning that no differences are observed between human groups. For instance, no differences are observed between sub-Saharan Africans and Eurasians, despite the additional bottleneck experienced by the latter (the Out-Of-Africa bottleneck). One likely explanation for these results is that under an additive model, the proportion of segregating mutations, both neutral and weakly deleterious, and the frequency of these

mutations vary in the opposite directions, maintaining the individual burden of deleterious mutations almost constant (Simons and Sella 2016; Simons et al. 2014). Similar conclusions were drawn by Do et al. (Do et al. 2015) using a similar statistic named $R_{x/y}$ (ratio of the count of derived alleles between two individuals from population x and y). They concluded that Africans and Europeans did not present any significant differences in the ability of natural selection to remove deleterious mutations. These results are at odds with previous studies (Lohmueller et al. 2008; Henn et al. 2016a; Fu et al. 2013) that found, based on other statistics, that recent demographic events impacted the burden of deleterious mutations and the efficacy of natural selection in humans.

During a bottleneck, heterozygosity decreases due to the loss of mutations, so the number of homozygous alleles increases. This means that, contrary to additive alleles, the count of recessive deleterious mutations, thus the recessive load, is more likely to be affected by recent demographic changes. Studies revealed that mutations with a strong impact on

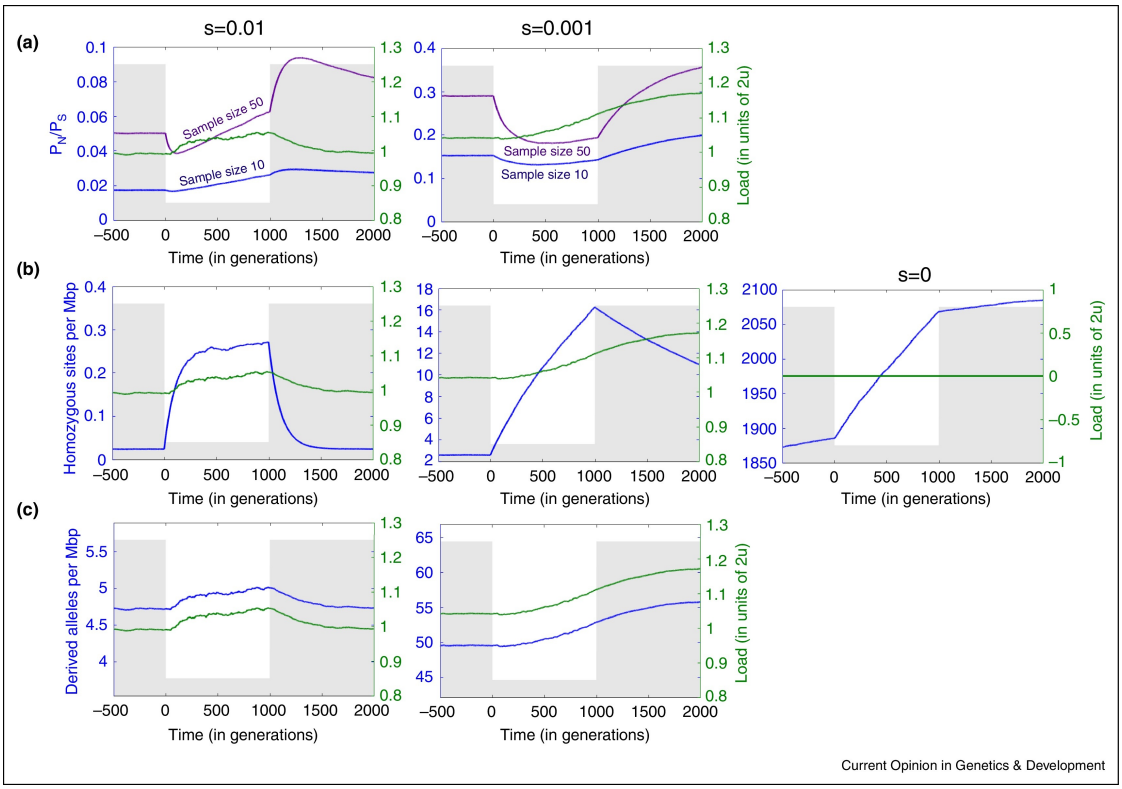


Figure 3.1: **Proxies for load (Simons and Sella 2016).** Additive load computed from simulations (green lines) with bottleneck (population size in gray varies from 10,000 to 1000 at time 0 and recovers a 1000 generations later) compared with different proxies used to calculate the mutational load using different samples sizes (blue and purple lines): (a) ratio non-synonymous/synonymous, (b) number of homozygous sites and (c) number of derived alleles. Only the number of derived alleles directly correlates with the mutational load and is not biased by the bottleneck (demographic event).

the phenotype, such as Loss-of-Function (LoF) mutations, are more likely to be recessive (Wright 1929; Simmons and Crow 1977; Phadnis and Fry 2005). Populations marked by strong founder events such as the Finnish or Ashkenazi Jews harbour significantly less LoF mutations than other populations (Narasimhan et al. 2016). The study of endogamous groups, e.g. Pakistani populations, also indicates the presence of fewer segregating LoF mutations highlighting the likely role of recent inbreeding in purging recessive alleles (Tadmouri et al. 2009; Bittles and Hamamy 2010; Garcia-Dorado 2008). Finally, Lopez et al. (Lopez et al. 2018b), found evidence for significant differences in the burden of deleterious mutations under a recessive model of dominance between African Rainforest hunter-gatherers, African farmers and Europeans. Similarly, Pedersen et al. (Pedersen et al. 2017b) found that Greenlandic Inuit harbour a higher recessive load than some continental populations because of their prolonged and extreme bottleneck.

3.2 Genetic adaptation to environments

A small fraction of the human genome, the genes and their regulatory regions, can harbour mutations that are under positive natural selection because they increase fitness by affecting traits. If a phenotype confers an advantage in a given environment, the carriers of the mutation(s) involved in such a phenotype would have a higher survival probability and reproductive success (i.e., higher fitness). Consequently, the mutation(s) would increase in the population more rapidly than expected under genetic drift alone.

A large number of population and evolutionary genetic studies have provided new insight into genomic regions that have been targeted by natural positive selection, ultimately contributing to the adaptive history of modern human populations (Barreiro et al. 2008; Barreiro and Quintana-Murci 2010; Bersaglieri et al. 2004; Fan et al. 2016; Hamblin and Di Rienzo 2000; Karlsson, Kwiatkowski, and Sabeti 2014; Lee et al. 2012; Malaspinas et al. 2016a; Quintana-Murci 2016, 2019; Quintana-Murci and Clark 2013; Sabeti et al. 2007; Tishkoff et al. 2007; Voight et al. 2006). For instance, genetic variants responsible for lactase persistence in adulthood present strong signals of positive selection in the genome of Europeans and East Africans (Bersaglieri et al. 2004; Tishkoff et al. 2007). Pathogen exposure also played a key role in the genetic adaptation of human populations (Barreiro and Quintana-Murci 2010; Quintana-Murci 2016, 2019; Quintana-Murci and Clark 2013) such as a genetic mutation at the *ACKR1* locus conferring resistance to malaria in Africa (Barreiro et al. 2008; Hamblin and Di Rienzo 2000; Quintana-Murci 2019) or in genes involved in the NF-κB signaling pathway conferring a resistance to cholera in population from Bangladesh (Lee et al. 2012; Karlsson, Kwiatkowski, and Sabeti 2014). There is also

evidence for genetic adaptation to climates, such as desert arid climate as reported for Aboriginal Australians in this case, mutations in the *NET01* and *KCNJ2* genes (Malaspinas et al. 2016a). Other examples of human local genetic adaptations are shown in Figure 3.2.

3.2.1 The classic sweep model

The "classic selective sweep" model, also named the "hard selective sweep" model, refers to a process in which a new and strongly beneficial mutation appears and increases rapidly in frequency to ultimately reach fixation in a given population (Pritchard, Pickrell, and Coop 2010). I have previously mentioned that mutations in the genome are genetically linked to each other (i.e. linkage disequilibrium) and form haplotypes. Under the "classic selective sweep" model, the strongly beneficial mutation will appear on a specific genetic background or haplotype that contains neutral mutations. Due to linkage disequilibrium, not only the beneficial mutation but the whole haplotype will disproportionately be transmitted to next generations following a mechanism known as "genetic hitch-hiking". The haplotype will increase so fast, that the recombination will not have time to break it into smaller haplotypes. As a result, one would expect to find around the selected locus (i)

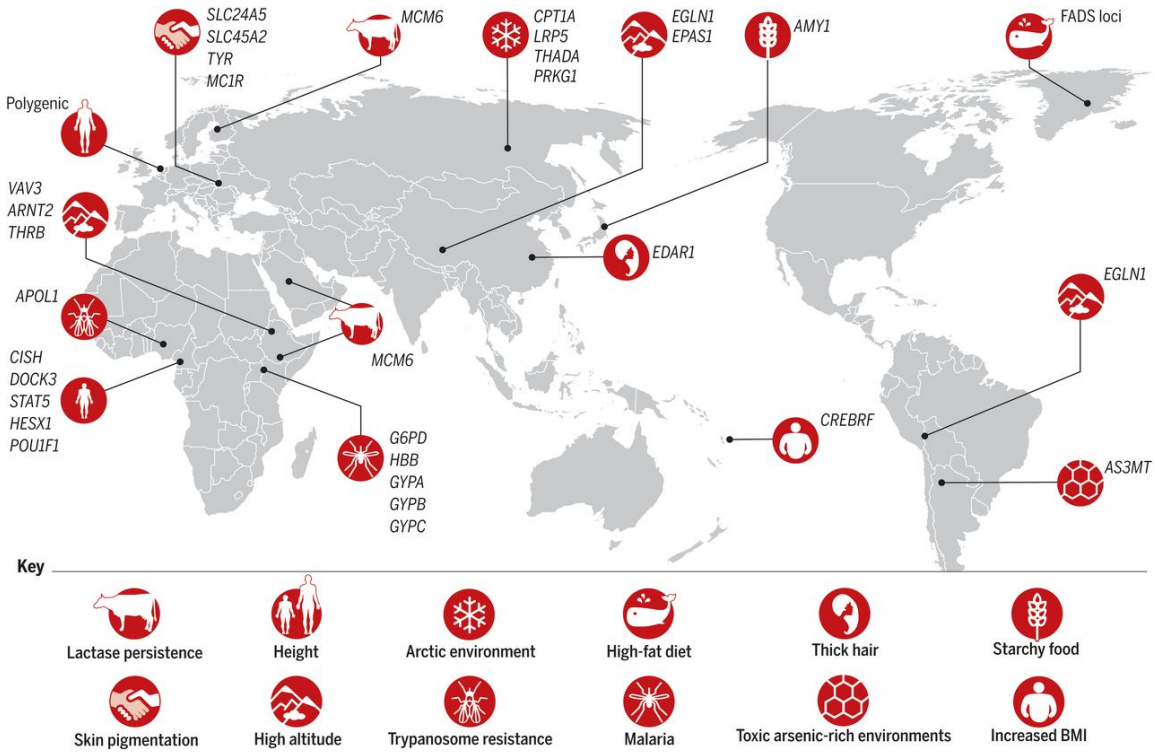


Figure 3.2: **Human local adaptation to their environments (Fan et al. 2016).** Examples of genes and phenotypes targeted by positive natural selection. Phenotypes with associated targeted genes are labelled according to the nature of selected traits.

a drop in genetic diversity, (ii) derived alleles at high frequency and (ii) haplotypes that are conserved over large genomic distances.

To be adaptive, a new mutation must appear in the right genomic region(s) and at the right moment. Consequently, "classic sweep" signals are expected to be rare. There is indeed compelling evidence to suggest that human genetic adaptation over the last 250,000 years involved only a low number of classic sweeps, suggesting that other modes of natural selection did occur (Pritchard, Pickrell, and Coop 2010; Schrider and Kern 2017).

3.2.2 Selection on standing variation and polygenic adaptation

Under a "selection on standing variation" model of genetic adaptation, the environmental pressure postdates the occurrence of the mutation(s) (Przeworski, Coop, and Wall 2005). In a specific environment, only a subset of mutations is beneficial, most of the genetic variation is neutral. Most of these mutations appeared in human ancestral populations and segregate under genetic drift in populations at different frequencies. Following a change in environmental pressure (e.g. settlement of a new geographic region) part of the standing variation can become advantageous, because they confer an advantage in face of this new environment; the frequency of these specific mutations would no longer be driven by genetic drift only but also by natural selection. As these mutations already exist in the population, the adaptive process will be faster than under the classic sweep model of natural selection.

There is increasing evidence to suggest that most traits in humans are polygenic, and each of the associated genes appear to have a small effect on the ultimate phenotype. However, the classic model of positive selection (the "hard sweep" model), assumes that selection targets de novo mutations that strongly impact adaptive traits (e.g., lactase persistence). A more realistic model, the polygenic model of selection, predicts that weak positive selection targets several genomic regions associated with complex traits or diseases (Pritchard, Pickrell, and Coop 2010). This alternative model of selection, also known as polygenic adaptation, is a process in which alleles associated with a specific complex trait and used to segregate only by genetic drift in a population become advantageous due to a change in environmental pressures (Pritchard, Pickrell, and Coop 2010; Peter, Huerta-Sanchez, and Nielsen 2012).

Molecular signatures of positive selection can be detected with a number of statistics based on several, different aspects of the data; the site frequency spectrum (Nielsen et al. 2005), genetic differentiation among populations (e.g F_{ST} or PBS statistics, (Shriver et al. 2004; Yi et al. 2010)) and haplotype homozygosity (e.g XP-EHH and iHS statstics, (Sabeti et al.

2007; Voight et al. 2006)). Under the hard sweep model, we expect a selected allele to be highly frequent in a specific population and be carried by haplotypes conversed over long genomic distances, due to hitchhiking effects (Fig. 3.2). However, for complex traits or diseases, many genetic variants are involved in the phenotype and each of them has a small contribution to the variance of the trait. Consequently, under a model of polygenic adaptation, we expect a subtle shift in allelic frequency in a specific population and in this case, the selective event follows a "soft-sweep model" (Pritchard, Pickrell, and Coop 2010)(Fig. 3.2).

3.2.3 Adaptive admixture and adaptive introgression

Adaptive admixture corresponds to another regime of natural selection in which beneficial mutations are transmitted from a population to another via gene flow. In human populations, admixture events are pervasive and thus have the potential to play a key role in the rapid genetic adaptation of human populations (Racimo et al. 2015; Gower et al. 2021; Patin et al. 2017; Hamid et al. 2021; Jeong et al. 2014). I will refer in this manuscript (Chapter 5, Article) to adaptive admixture when the two populations belong to the same species (i.e two modern human populations) and to adaptive introgression when they are from two different species or human lineages, like between archaic hominins (i.e., Neanderthal and Denisova) and *Homo sapiens*.

3.3 Metabolic disorders in the Pacific

According to the World Health Organization (WHO), most of the top 10 countries with the highest rate of obesity are found in Pacific Islands. In some islands of Polynesia and Micronesia, more than 70% of the population is obese (e.g Nauru, Samoa, Tonga) and obesity represents up to 75% of the causes of death (Fig. 3.3). More specifically, metabolic disorders such as Type 2 diabetes and Gout are highly prevalent in the Oceanian regions (Gosling et al. 2015).

3.3.1 Population history

The first Europeans who arrived in the Pacific islands described autochthonous people as "healthy", "muscular" and "strong" indicating that traditional food and diet were appropriate for the lifestyle of Pacific islanders (Fisk 1966). Since 1963, the Pacific region has experienced a sharp nutrition transition owing to the global trade and globalization.

Imported food has either replaced part of the local food, especially carbohydrate sources (root crops, fruits and vegetables have been replaced by imported flour, rice, meat, alcohol and milk) or has been added to local fat sources (e.g. imported vegetable oil or butter added to coconuts). More generally, the consumption of fat increased, for example, in French Polynesia, by 80% between 1963 and 2000 (Gosling et al. 2015; Hughes and Lawrence 2005; Fisk 1966).

Over the last 50 years, Pacific islanders migrated from rural to urban regions and nowadays more than half of the population lives in urban areas where they have a more sedentary lifestyle and practice less physical activity. As a consequence, the highest rate of obesity is found in urban centers such as New Zealand and a survey from 1998 in the Vanuatu islands indicates that although people living in rural areas absorbed more calories than people from urban areas, they are less obese mainly because they consume five times less imported fat products (Hughes and Lawrence 2005).

Despite the lower levels of metabolic disorders, they are still present in rural regions where people maintain a more traditional lifestyle and remain more isolated from the globalization (Gosling et al. 2015; Gosling, Matisoo-Smith, and Merriman 2014). Furthermore, bone lesions resembling that of gout arthritis have been also identified in the first Lapita settlers of the Vanuatu dated to around 3,000 years ago (Buckley 2007). Together, these observations suggest that, in addition to environmental factors, Pacific islanders could also be more biologically susceptible to metabolic disorders because of their genetic background and specific population history (both demographic and adaptive history).

3.3.2 A case of "maladaptation"?

The geneticist James Neel in 1962 proposed the hypothesis of the thrifty gene or thrifty genotype to explain the high prevalence of Type 2 diabetes observed in contemporary societies (Neel 1962). This hypothesis stipulates that mutations found in genes involved in fat storage were under positive natural selection because they conferred an advantage in period of food privation. Because of changes in diet and lifestyle (caloric and food excess), the genetic variants that were formerly advantageous are nowadays detrimental and are associated with metabolic disorders.

Focusing on Oceania, studies (Diamond 2003; Bindon and Baker 1997) argued that the voyages in canoes associated with the settlement of remote islands as well as between-island connexions (trade) were accompanied by food privation and a high mortality rate. In this context, people aboard canoes who carried thrifty alleles would have

had a higher survival probability. Candidate thrifty alleles have been proposed, such as a mutation located in the *CREBRF* gene, which shows a signature of positive selection and is associated with increased Body Mass Index (BMI) and fat storage in Samoans (Minster et al. 2016; Loos 2016). However, the same mutation protects against Type 2 diabetes (Minster et al. 2016; Krishnan et al. 2018) and is also associated with taller stature in Samoans and Maori (Carlson et al. 2020). Because of these pleiotropic effects, it is not possible to know which of these traits was/were likely targeted by natural selection and, thus, whether the CREBRF mutation is indeed a thrifty allele.

After more than 60 years of research, the thrifty gene hypothesis is still currently highly debated. For example, Ayub et al. (Ayub et al. 2014) found no evidence of positive selection at loci associated with Type 2 diabetes in African, European and East Asian groups. Although the thrifty gene hypothesis is commonly used to explain the high prevalence of metabolic disorders found in Oceania, very few studies succeeded in formally assessing and identifying genetic variants associated with metabolic disorders specifically in Oceanian groups (most of associated variants come from European-based association studies) and presenting signatures of natural selection.

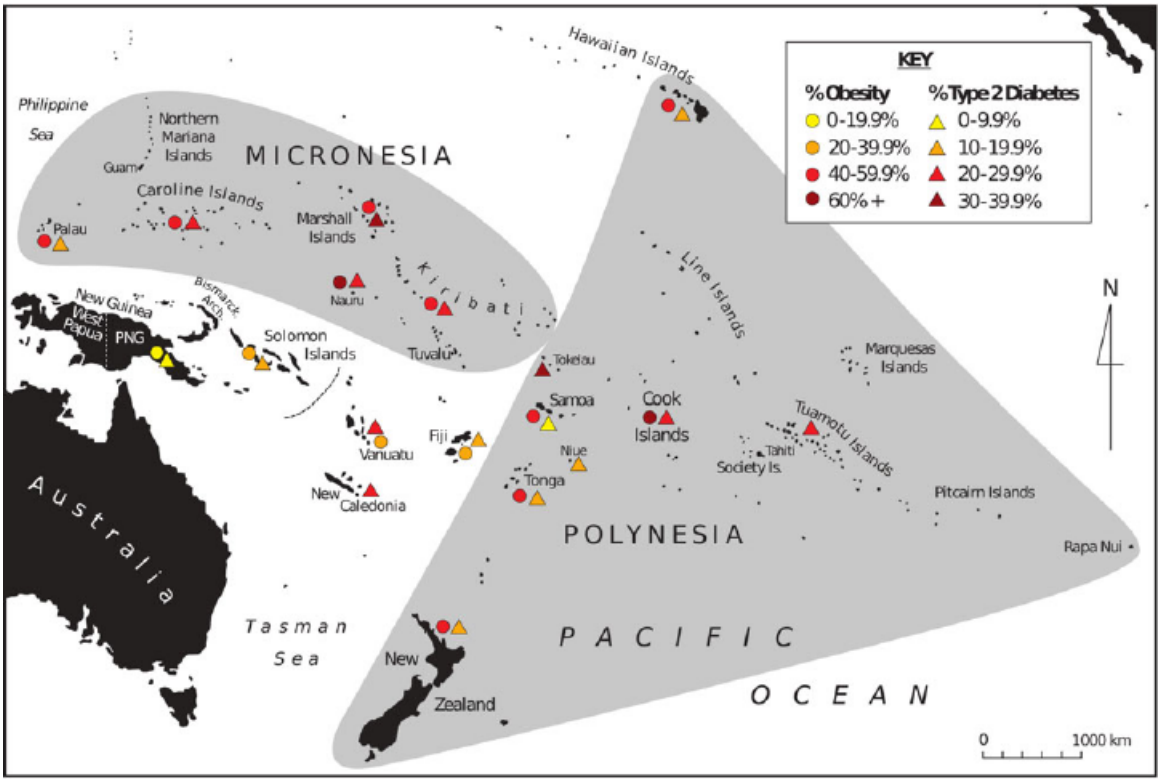


Figure 3.3: **Prevalence of obesity and Type 2 diabetes in Oceania (Gosling et al. 2015)**.

Chapter 4

Objectives of the thesis

During the last 125,000 years, modern humans (*Homo sapiens*) spread across all continents and settled in diverse ecosystems, as extreme as the Sahara Desert, the Arctic Circle, or the Himalayas. Archaeological and linguistic data have provided valuable insights into the tempo of human dispersals across the globe but many questions remain open: did populations expand together with their languages and lifestyles? Do human cultures defined by archaeology reflect distinct genetic entities? Were human dispersals accompanied by genetic admixture with local groups of archaic or modern humans? How have humans genetically adapted to the newly colonized environments? The recent advent of high-throughput sequencing technologies now allows tackling these questions in great detail, through the full characterization of the genetic diversity of human populations living in current and ancient times.

These massive sequence-based datasets can be interpreted in light of theoretical frameworks in population genetics, developed from well-known mathematical frameworks, such as the coalescent theory or the diffusion approximation to the discrete generational model. Combining whole-genome sequencing data with robust statistical and mathematical frameworks in population genetics thus allows infer demographic models that best explain current patterns of genetic variation.

The region of Oceania, composed of thousands of scarcely populated islands, provides with an excellent model system to test important hypotheses in human evolution, population genetics and evolutionary biology. This project aims to reconstruct the genetic history of Oceanian Islanders, with the goal of dissecting their demographic past and to ultimately better understand their present-day relation to disease. Specifically, my PhD project aims to (i) characterize the genetic diversity of Oceanian populations, which are under-represented in genomic studies, (ii) trace back all the different events constituting

their demographic history (see Chapter 5), and finally (iii) evaluate the purge of deleterious mutations (i.e., mutations that could cause diseases) in these populations (see Chapter 6).

To do so, I have first set up the necessary pipeline to process the high-coverage sequencing of 317 new whole genomes. I combined multiple bioinformatics tools to align sequencing reads, call genetic variants and genotypes, and check sample and variant quality. Once high-quality data was obtained, my next steps have been the detailed characterisation of the genetic diversity and structure of Oceanians, to ultimately jointly infer the demographic parameters characterizing their population history. Specifically, I inferred the demographic models of (i) Near Oceanians, (ii) western Remote Oceanians and (iii) East/Southeast Asian ancestors of Near and Remote Oceanians. I explored and evaluated a large range of possible demographic scenarios using the maximum likelihood framework and SFS-based parameter estimations implemented in *Fastsimcoal2* (Excoffier et al. 2013; Excoffier et al. 2021). Secondly, I started to evaluate the burden of deleterious mutations of Pacific islanders. To do so, I estimated and compared the efficacy of natural selection and the mutational load between Pacific and reference populations.

Chapter 5

Result 1
Demographic history and genetic adaptation of Pacific islanders

5.1 Context

As seen in chapters 1 and 2, the islands of the Pacific are classified into Near Oceania and Remote Oceania. These two-sub regions of Oceania differ in their geographic location and peopling history. The oldest archaeological sites are found in Near Oceania (New Guinea, the Bismarck Archipelago and the Solomon Islands) and studies indicate a settlement of this region between around 45,000 years ago for Northern Sahul and 25,000-30,000 years ago for western Solomon Islands. The descendants of this Pleistocene occupation are Papuan-speaking communities that live today in New Guinea and islands lying off its northeast coast. The peopling of the rest of the Pacific, known as Remote Oceania and including the Reef/Santa Cruz islands, Vanuatu, New Caledonia, Fiji, Micronesia and Polynesia, only occurred recently in the Holocene. This dispersal which has been associated with the expansion of Austronesian languages and the Lapita Cultural Complex, was proposed to originate around 5,000 years ago from Taiwan and reach western Remote Oceania by around 3,200 years ago, and the Polynesian Triangle by 1,000–700 years aga (i.e. the "Out-of-Taiwan" model).

Ancient DNA studies in Remote Oceania, primarily in Vanuatu and Tonga, have reported virtually no Papuan ancestry in individuals from the Lapita period, and supported a second movement of Papuan-like people likely from the Bismarck Archipelago, shortly after the initial Lapita settlement (Lipson et al. 2018; Posth et al. 2018; Skoglund et al. 2016). Genetic studies of modern Oceanians have reported varying levels of Papuan- and East Asian-related ancestry across islands (Friedlaender et al. 2008; Pugach et al. 2018b; Wollstein et al. 2010). However, the detail characterization of the demographic history (i.e. effective population size, divergence times, mode and tempo of gene flow) of Near and Remote Oceanians as well as the different biological functions that contributed to their adaptation remain poorly defined. Additionally, some Oceanian groups have retained the highest worldwide levels of combined Denisovan and Neanderthal ancestry (Qin and Stoneking 2015; Reich et al. 2011; Vernot et al. 2016; Sankararaman et al. 2016), but it is still unclear when and how this introgression occurred and whether it facilitated local adaptation. To date, genetic studies of this region have focused on geographically-restricted datasets and/or ascertained SNP arrays (Friedlaender et al. 2008; Pugach et al. 2018b; Wollstein et al. 2010), limiting our ability to unbiasedly study the genomic history of Near and Remote Oceania and the legacy of archaic admixture across Oceanians.

In this article I mainly led and performed the processing of the high-coverage whole genome sequences, the analyses related to the description of the dataset, the population

structure (see **Genomic dataset and population structure** and related Supplementary information) and demographic inference of Pacific islanders (see **The settlement of Near and Remote Oceania**, **Insights into the Austronesian expansion** and related Supplementary information).

5.2 Article

Article

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 Cassar[4], Maximilian Larena[5], Albert Min-Shan Ko[6], Christine Harmant[1], Romain Laurent[7], Paul Verdu[7], Guillaume Laval[1], Anne Boland[8], Robert Olaso[8], Jean-François Deleuze[8], Frédérique Valentin[9], Ying-Chin Ko[10], Mattias Jakobsson[5,11], Antoine Gessain[4], Laurent Excoffier[12,13], Mark Stoneking[14], Etienne Patin[1,17] ✉ & Lluis 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 processes[1]. 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 ka[6]. 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

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

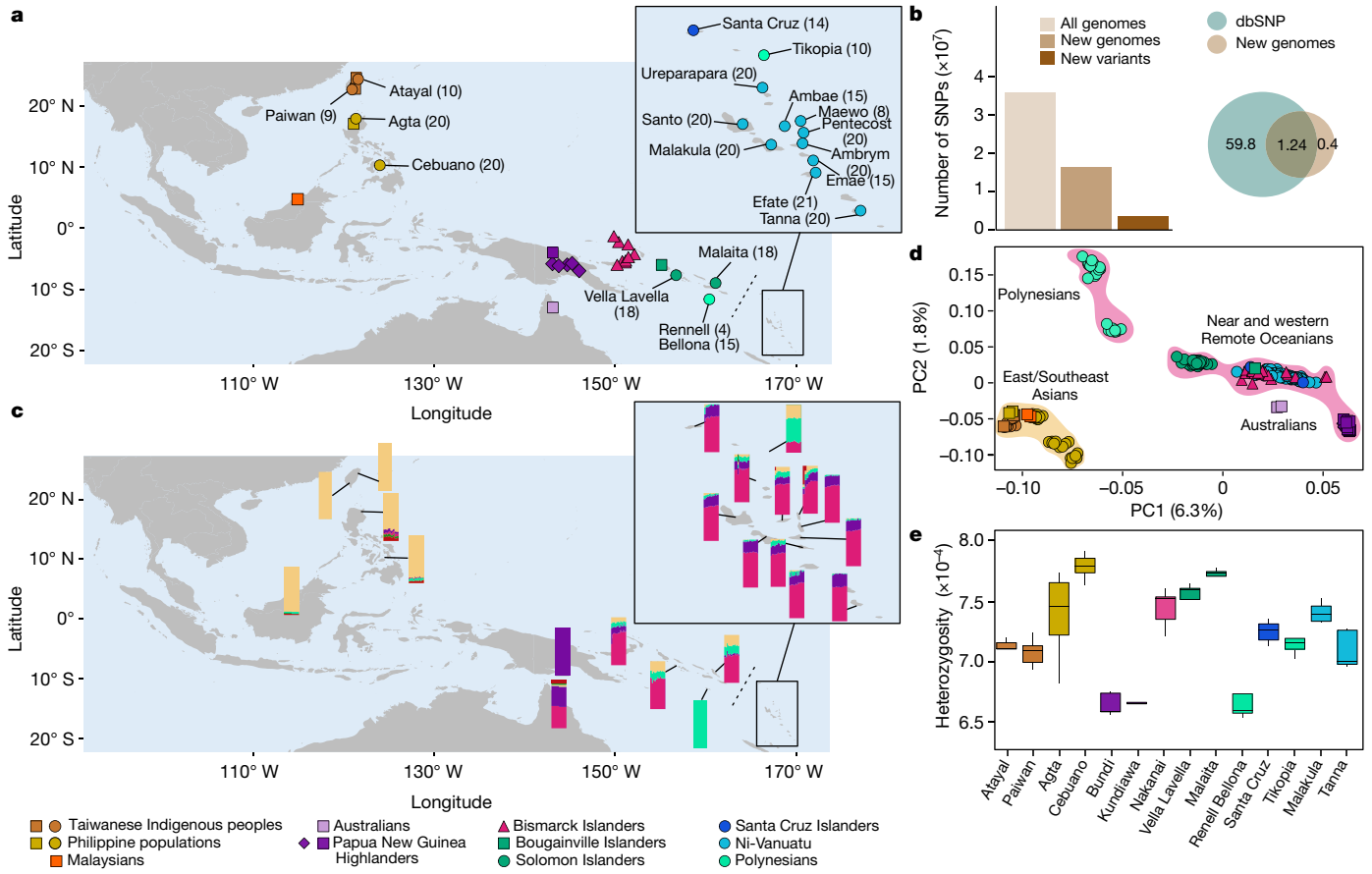


Fig. 1 | Whole-genome variation in Pacific Islanders. a, Location of studied populations. The indented map is a magnification of western Remote Oceania. Circles indicate newly generated genomes. Sample sizes are indicated in parentheses. Squares, triangles and diamonds indicate genomes from Mallick et al.[19], Vernot et al.[16] and Malaspinas 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× the interquartile range and outliers, respectively. **a**, **c**, Maps were generated using the maps R package[51].

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 Emae 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 method[24] (Supplementary Note 4). We first determined the relationship between Papua New Guinean Highlanders and other modern and archaic hominins, and replicated previous findings[18] (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].

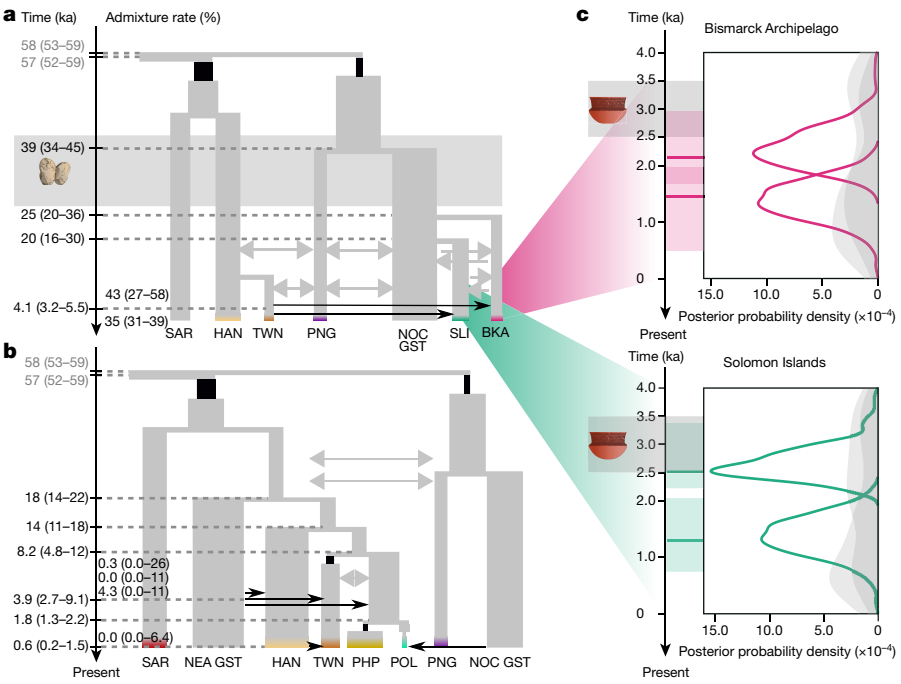


Fig. 2 | Demographic models of the human settlement of the Pacific. **a**, Maximum-likelihood model for Near Oceanian populations. Point estimates of parameters and 95% confidence intervals are reported in Supplementary Table 4. The grey area indicates the archaeological period for the settlement of 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 Oceania[27].

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 populations[25]. 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 groups[26] could bias parameter estimation. When accounting for such gene flow, we obtained consistently older divergence times than expected under the out-of-Taiwan model[4], 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 Neolithic[2], 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 ka[27] (2.2 ka (95% credible interval, 1.7–3.0) and 2.5 ka (95% credible interval, 2.2–3.4) for Bismarck

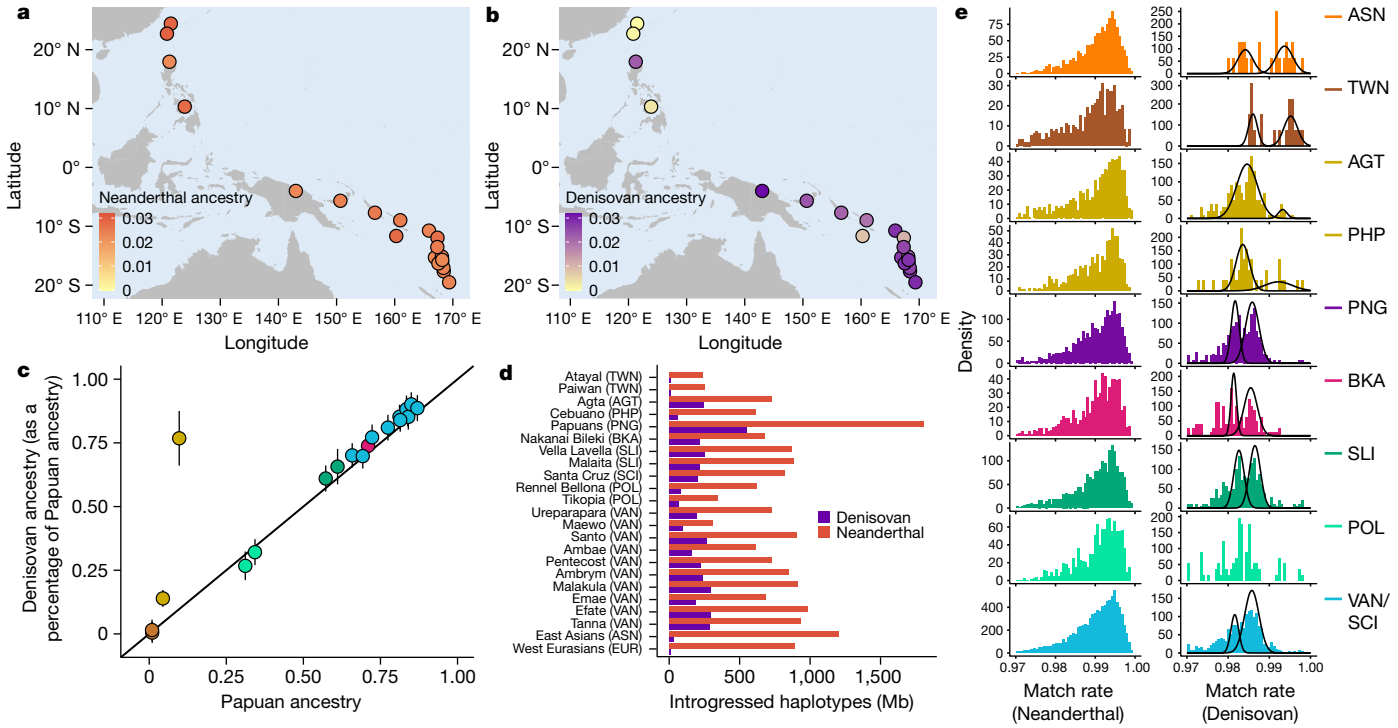


Fig. 3 | Neanderthal and Denisovan introgression across the Pacific. **a**, **b**, Estimates of Neanderthal (**a**) and Denisovan (**b**) ancestry on the basis of f_4-ratio statistics. Maps were generated using the maps R package[51]. **c**, Correlation between Papuan ancestry and Denisovan ancestry (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 only[19]); 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 reported[28], 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 populations[29] (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 ka[29] (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.

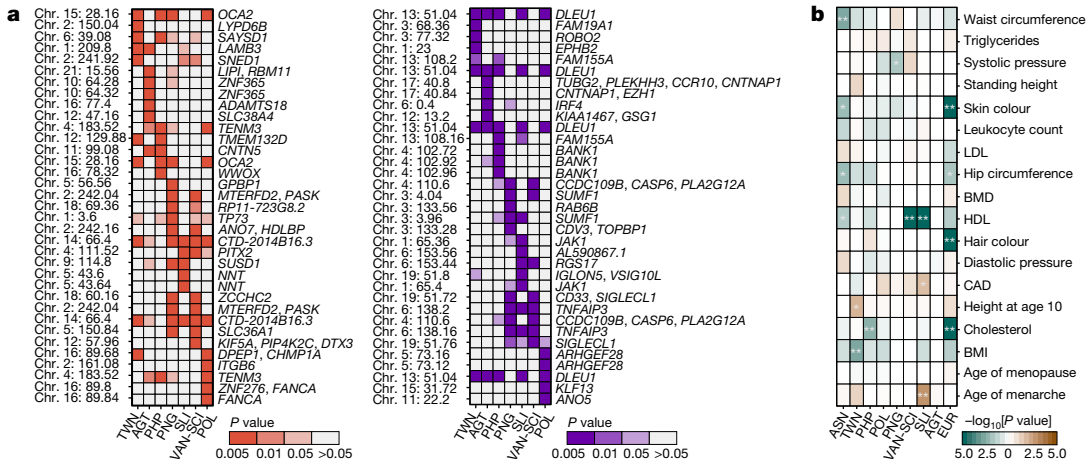


Fig. 4 | Mechanisms of genetic adaptation to Pacific environments. **a**, Genomic regions showing the strongest evidence of adaptive introgression from Neanderthals (red) and Denisovans (purple). Each row is a 40-kb window, each column is a Pacific population group, and each cell is coloured according to whether the window is in the top 0.5%, 1%, 5%, >5% of the empirical distributions of the adaptive introgression Q95 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. *CCDC109B* is also known as *MCUB*, *KIAA1467* is also known as *FAM234B*, *FAM19A1* is also known as *TAFA1*, *MTERFD2* is also known as *MTERF4*, *RP11-723G8.2* is also known as *LINC01899*. **b**, Signals of polygenic adaptation. Blue and brown colours indicate the $-\log_{10}$(*P* 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.

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 Philippines[30], prompted us to search for introgression from other archaic hominins. Using the S′ method[28], 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*, *TIAM1* 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*, *PTCH1*, *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*, *SAMSN1*, *ROBO2* and *PELI2*)[29,31] and metabolism-related (*DLEU1*, *WARS2* 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 *ARHGEF28*, *BANK1*, *CCR10*, *CD33*, *DCC*, *DDX60*, *EPHB2*, *EVI5*, *IGLON5*, *IRF4*, *JAK1*, *LRRC8C* and *LRRC8D*, and *VSIG10L* (Fig. 4a and Supplementary Table 15). For example, *CD33*—which mediates cell–cell interactions and keeps immune cells in a resting state[34]—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 infections[35]—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 Denisovans[31] (Extended Data Fig. 7). The strongest hit (empirical *P* < 0.001) included *GABRP*, which mediates the anticonvulsive effects of endogenous pregnanolone during pregnancy[36], and *RANBP17*, which is associated with body mass index and high-density lipoprotein cholesterol[37] (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 deprivation[38] and is associated with blood pressure[39], 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 individuals[40]. 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 levels[41] and protection against dengue[42], 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 vision[43]. *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 polygenic[44], we tested for directional selection of 25 complex traits with a well-studied genetic architecture[45], by comparing the integrated haplotype scores (iHS) of trait-associated alleles to those of matched, random SNPs[46]. 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 time[18] (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 model[4], 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 pulse[28], 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 region[50]. 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.

1. 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).
2. Hung, H.-C. & Carson, M. T. Foragers, fishers and farmers: origins of the Taiwanese Neolithic. *Antiquity* **88**, 1115–1131 (2014).
3. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
4. Bellwood, P. *First Farmers: the Origins of Agricultural Societies* (Blackwell, 2005).
5. 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).
6. Kirch, P. V. *On the Road of the Winds: An Archeological History of the Pacific Islands before European Contact* (Univ. California Press, 2017).
7. Wollstein, A. et al. Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992 (2010).
8. Lipson, M. et al. Population turnover in Remote Oceania shortly after initial settlement. *Curr. Biol.* **28**, 1157–1165 (2018).
9. Skoglund, P. et al. Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).
10. Posth, C. et al. Language continuity despite population replacement in Remote Oceania. *Nat. Ecol. Evol.* **2**, 731–740 (2018).
11. Pugach, I. et al. The gateway from Near into Remote Oceania: new insights from genome-wide data. *Mol. Biol. Evol.* **35**, 871–886 (2018).
12. Bergström, A. et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160–1163 (2017).
13. Ioannidis, A. G. et al. Native American gene flow into Polynesia predating Easter Island settlement. *Nature* **583**, 572–577 (2020).
14. Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015).
15. 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).
16. Vernot, B. et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
17. 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).
18. 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 Neandertal 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).

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.0.7.13[52] and duplicates were marked with Picard Tools. Base quality scores were recalibrated with the Genomic Analysis ToolKit (GATK) software v.3.8[53].

Whole-genome data for Bismarck Islanders[16] were processed in the same manner as the newly generated genomes, while for Papua New Guinean Highlanders[18] and other populations of interest[19], raw BAM files were converted into uBAM files, and processed as described above. Variant calling was performed following the GATK best-practice recommendations[54]. 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 GATK[54] and BCFtools. Heterozygosity was assessed with PLINK v.1.90[55,56] and cryptically related samples were detected with KING v.2.1[57]. Previously unknown SNPs were identified by comparison with available datasets[16,18,19] and dbSNP[58].

Genetic structure analyses

PCAs were performed with the 'SmartPCA' algorithm implemented in EIGENSOFT v.6.1.4[59]. The genetic structure was determined with the unsupervised model-based clustering algorithm implemented in ADMIXTURE[60], 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 Haploview[61], which was averaged per bin of genetic distance using the 1000 Genomes Project phase 3 genetic map[62]. F_{ST} values were estimated by analysis of molecular variance (AMOVA) as previously described[63] (Supplementary Note 3).

Demographic inference

Demographic parameters were estimated with the simulation-based framework implemented in fastsimcoal v.2.6[24]. We filtered out sites (1) within CpG islands[64]; (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_{obs}/S_{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 years[68]. We also provide estimates of divergence and admixture times assuming a mutation rate of 1.4×10^{-8} mutations per generation per site[69] (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 fastsimcoal2[24] (Supplementary Note 4). We also compared observed and simulated F_{ST} values, computed with vcftools v.0.1.13[70], 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.3.5.2.2[71] 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 fastsimcoal2[24] 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 model[66]. 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 fastsimcoal2[24], 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 approaches[72] 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 statistics[73]. The first approach, developed in the MetHis method[74], is based on the moments of the distribution of admixture proportions and explicit forward-in-time simulations that follow a general mechanistic admixture model[75]. 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 package[78]. 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 fastsimcoal2[24], 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 MetHis[74], 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 approach[79] 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 package[78]. 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 fastsimcoal2[24], assuming a variable recombination rate sampled from the 1000 Genomes Project phase 3 genetic map[62]. 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.1.5.4[80], 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 package[78] 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-statistics[81]. We computed a D-statistic of the form D(X, West Eurasians/East Asians/Africans; Neanderthal Vindija, chimpanzee) and D(X, West Eurasians/East Asians/Africans; Neanderthal Vindija, Denisova Altai) to test for introgression from Neanderthal; and D-statistics of the form D(X, West Eurasians/East Asians; Denisova Altai, chimpanzee) and D(X, West Eurasians/East Asians; 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.5.1.1[81]. 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 genome[28]. 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) dataset[19], and segments were detected in each population separately. Genetic distances between sites were estimated from the 1000 Genomes Project phase 3 genetic map[62]. After retrieving empirical S' scores, we estimated a null distribution of S' scores by simulating—with fastsimcoal2[24]—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 individuals[19] merged with the Altai Denisovan genome. To detect Denisovan-introgressed haplotypes, we used as reference panel the Altai Denisovan genome and SGDP African individuals[19] 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

$P_{Neanderthal} \geq 0.9$ and $P_{Denisova} < 0.5$ as Neanderthal-introgressed haplotypes and those containing alleles with $P_{Denisova} \geq 0.9$ and $P_{Neanderthal} < 0.5$ 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 described[28] (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 described[28]. 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 approach[72], by performing 50,000 independent simulations of 64 DNA sequences of 10 Mb each with fastsimcoal2[24]. 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 transition[12] (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 pulse[29]. 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*-statistics[32]. We computed these statistics in 40-kb non-overlapping windows along the genome of all target populations, using SGDP African individuals[19] 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], Wikipathways[86], the genome-wide association studies (GWAS) catalogue[87], Gene Ontology[88], and manually curated lists of innate immunity genes[89] and virus-interacting proteins[90]. 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, PhastCons[91], combined annotation-dependent depletion (CADD) scores[92], density of DNase I segments[93] 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 database[45], 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 tiHS of the x independent, trait-associated alleles with the mean tiHS 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 tiHS 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 tiHS for each population. We then tested whether the mean tiHS 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/evoceania).

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

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.

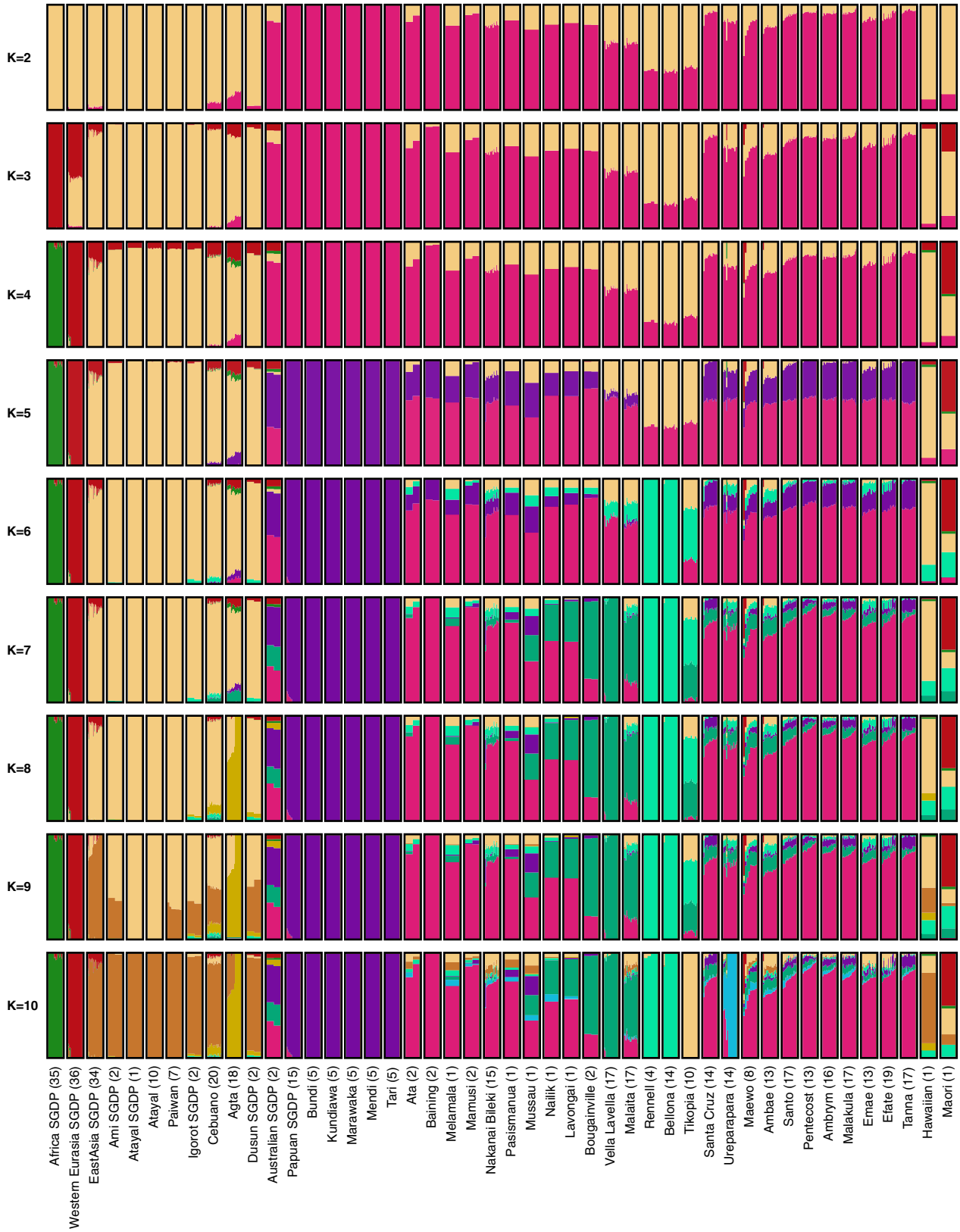
Competing interests The authors declare no competing interests.

Additional information

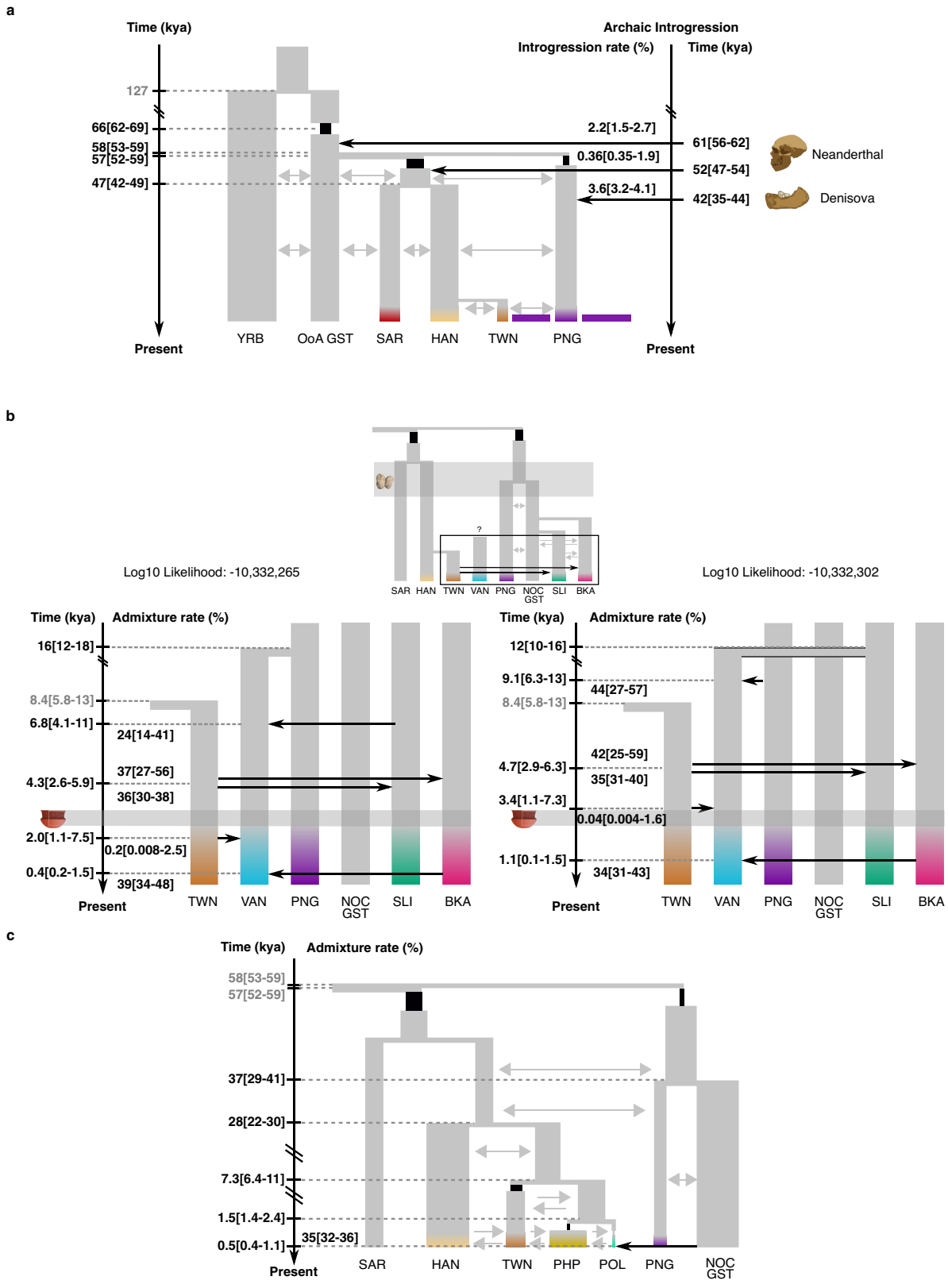
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03236-5.

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.

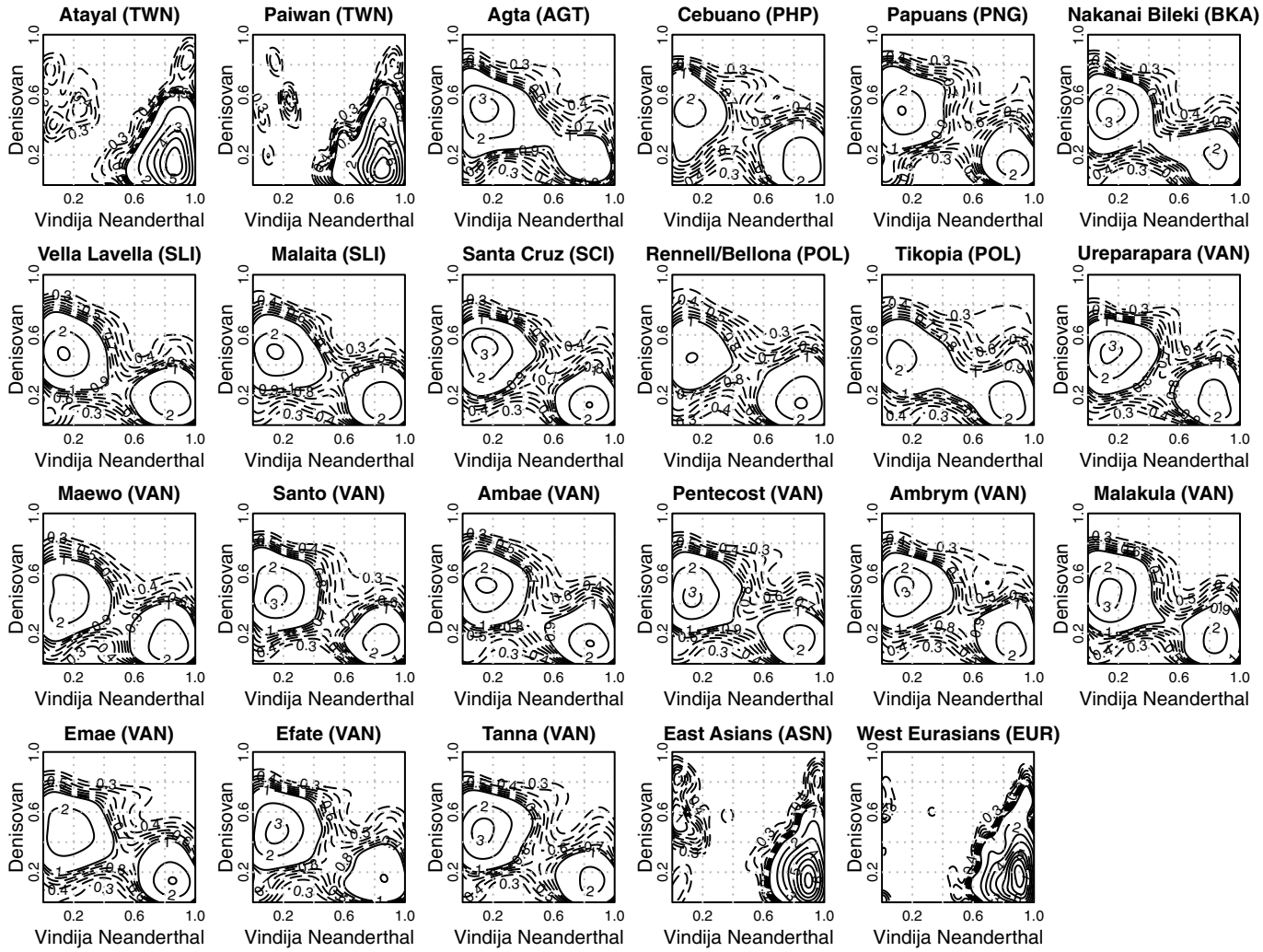


Extended Data Fig. 1 | Genetic structure of Pacific populations. ADMIXTURE ancestry components are shown from $K = 2$ (top) to $K = 10$ (bottom) for the 462 unrelated individuals. The lowest cross-validation error was obtained at $K = 6$ (Supplementary Fig. 5). Populations are delimited by black borders. Population width is not proportional to population sample size, which is indicated in parentheses.

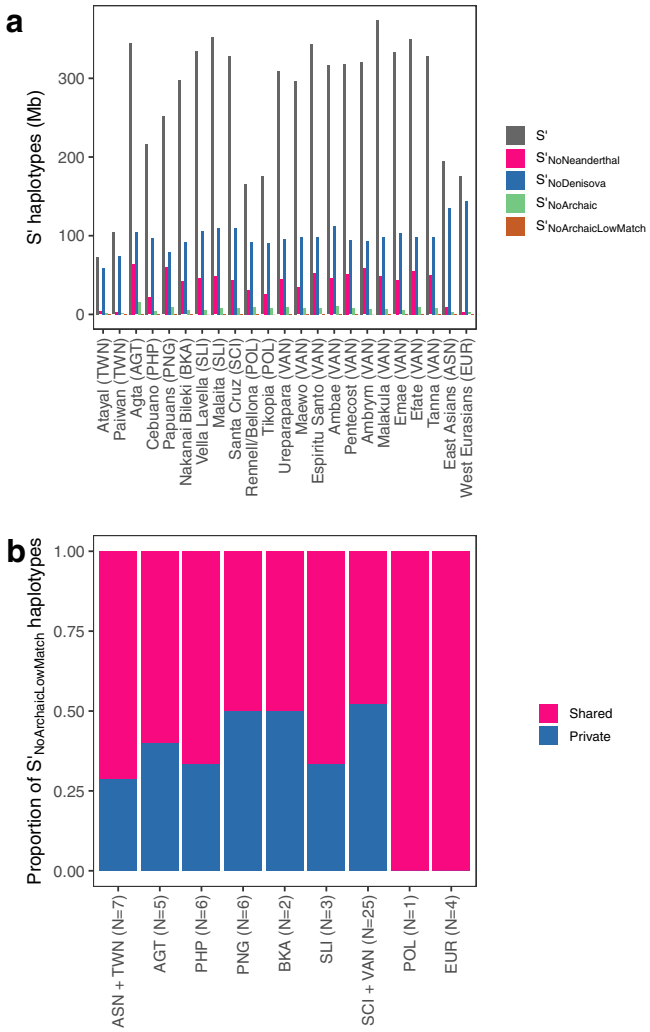


Extended Data Fig. 2 | See next page for caption.

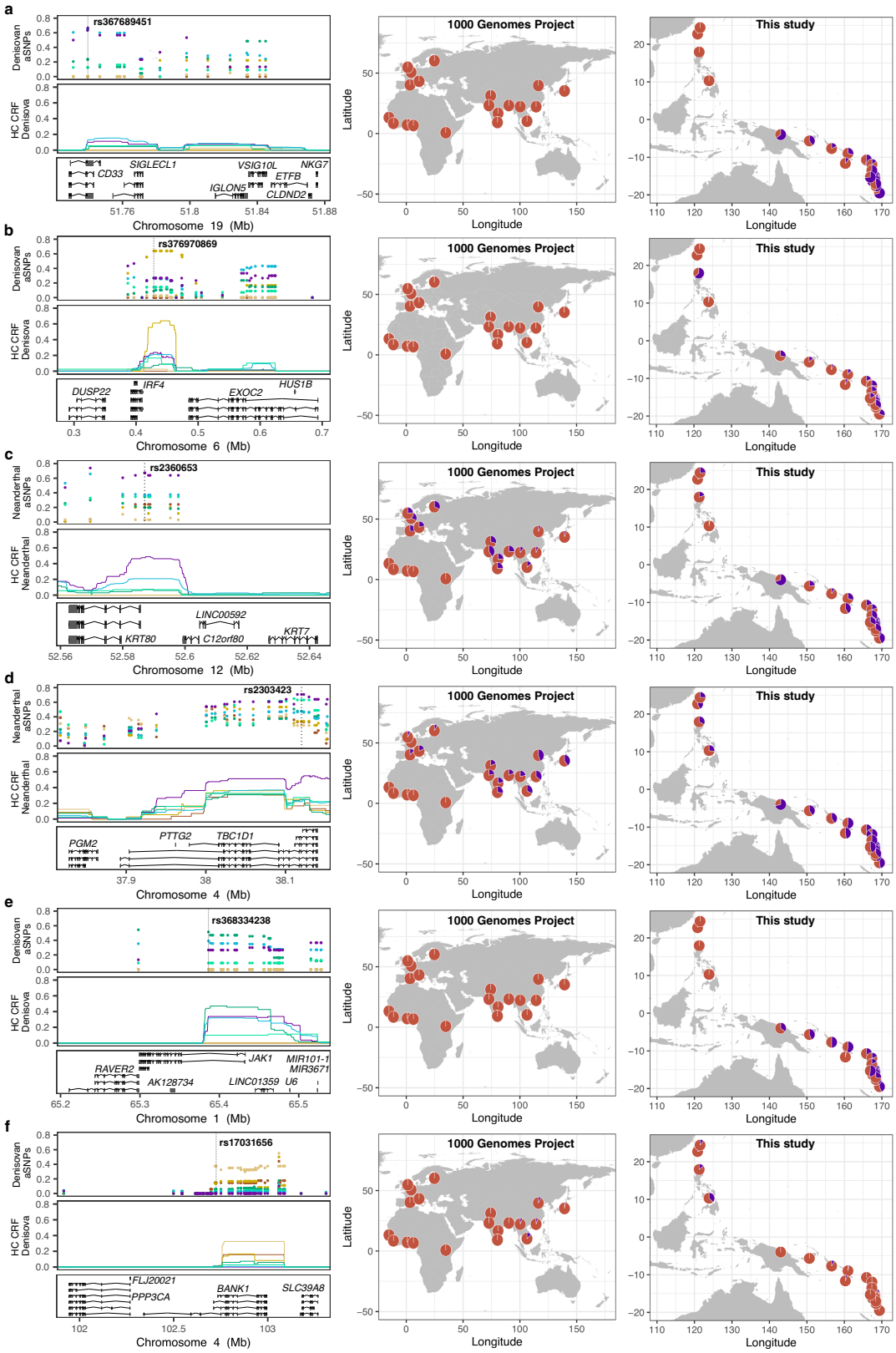
Extended Data Fig. 2 | Demographic models for Pacific populations. **a**, Maximum-likelihood demographic model for baseline populations. Point estimates of parameters and 95% confidence intervals are shown in Supplementary Table 2. **b**, Maximum-likelihood demographic models for western Remote Oceanian 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.



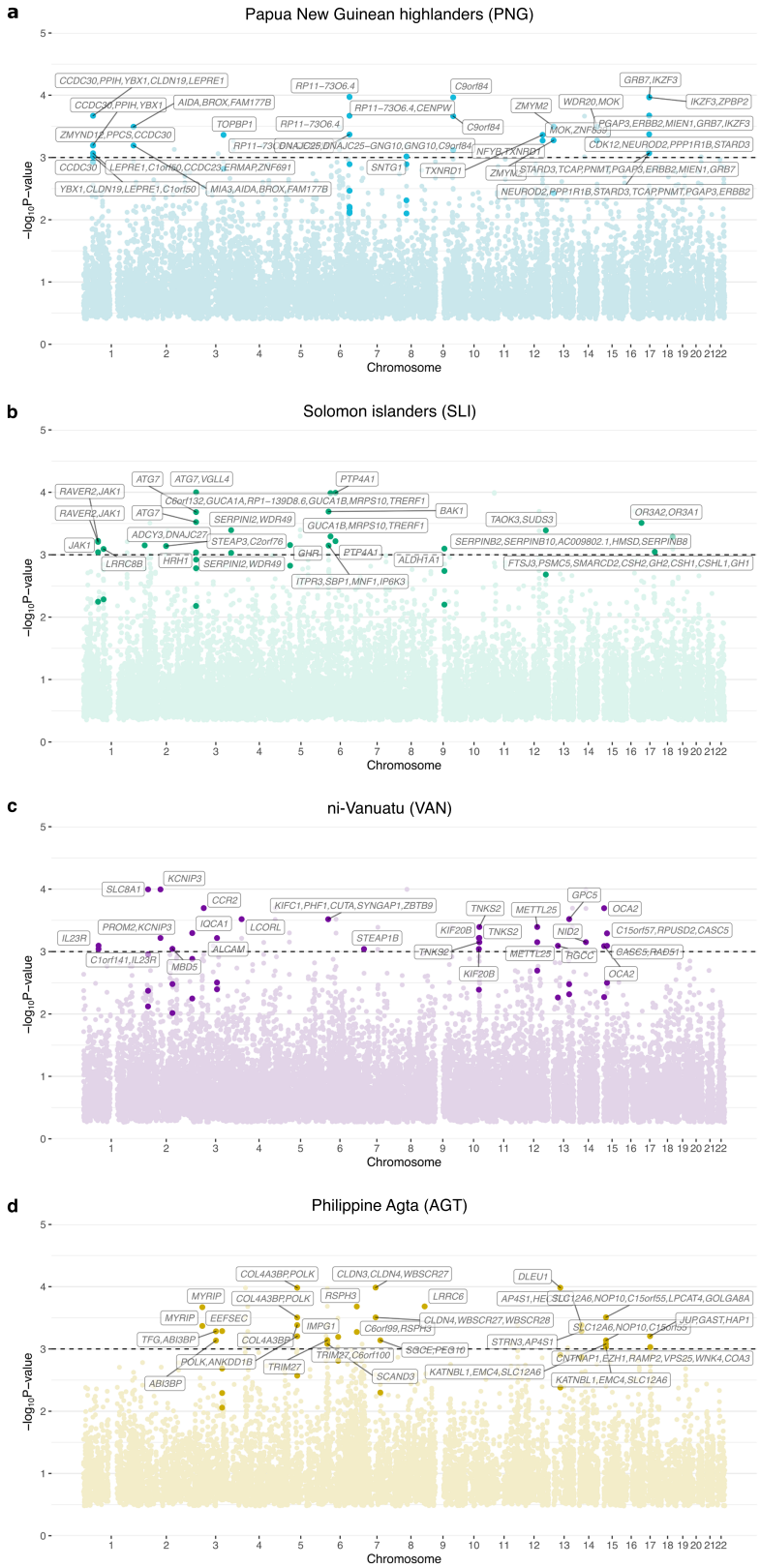
Extended Data Fig. 3 | Match rate of introgressed S′ haplotypes in Pacific populations to the Vindija Neanderthal and Altai Denisovan genomes. The match rate is the proportion of putative archaic alleles matching a given archaic genome, excluding sites at masked positions. Only S′ haplotypes with more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



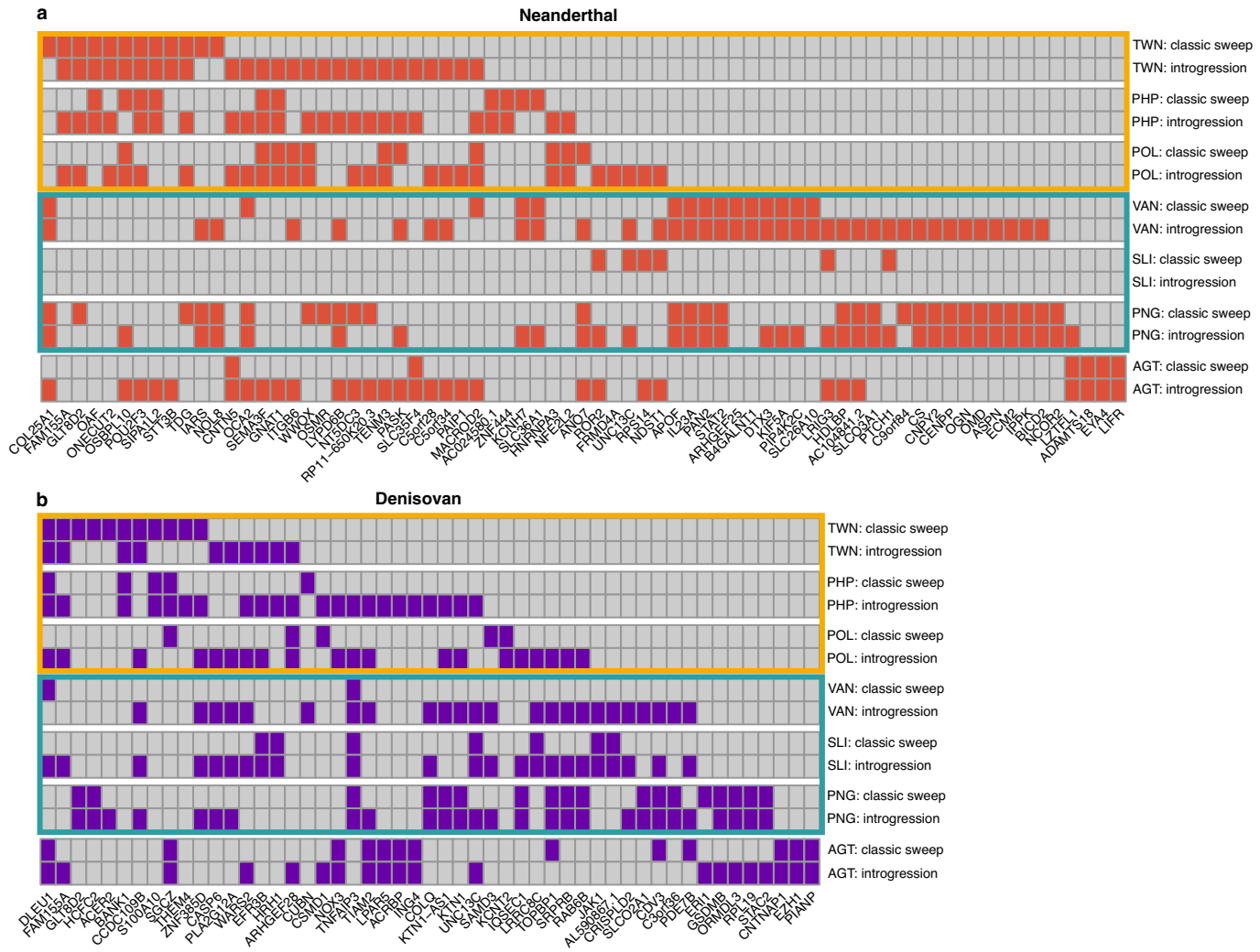
Extended Data Fig. 4 | Detection of introgressed haplotypes from an unknown archaic hominin. a, Cumulative length of S′ haplotypes retrieved among modern human populations (S′), after removing Neanderthal CRF haplotypes ($S'_{NoNeanderthal}$) or Denisovan CRF haplotypes ($S'_{NoDenisova}$) or both ($S'_{NoArchaic}$),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



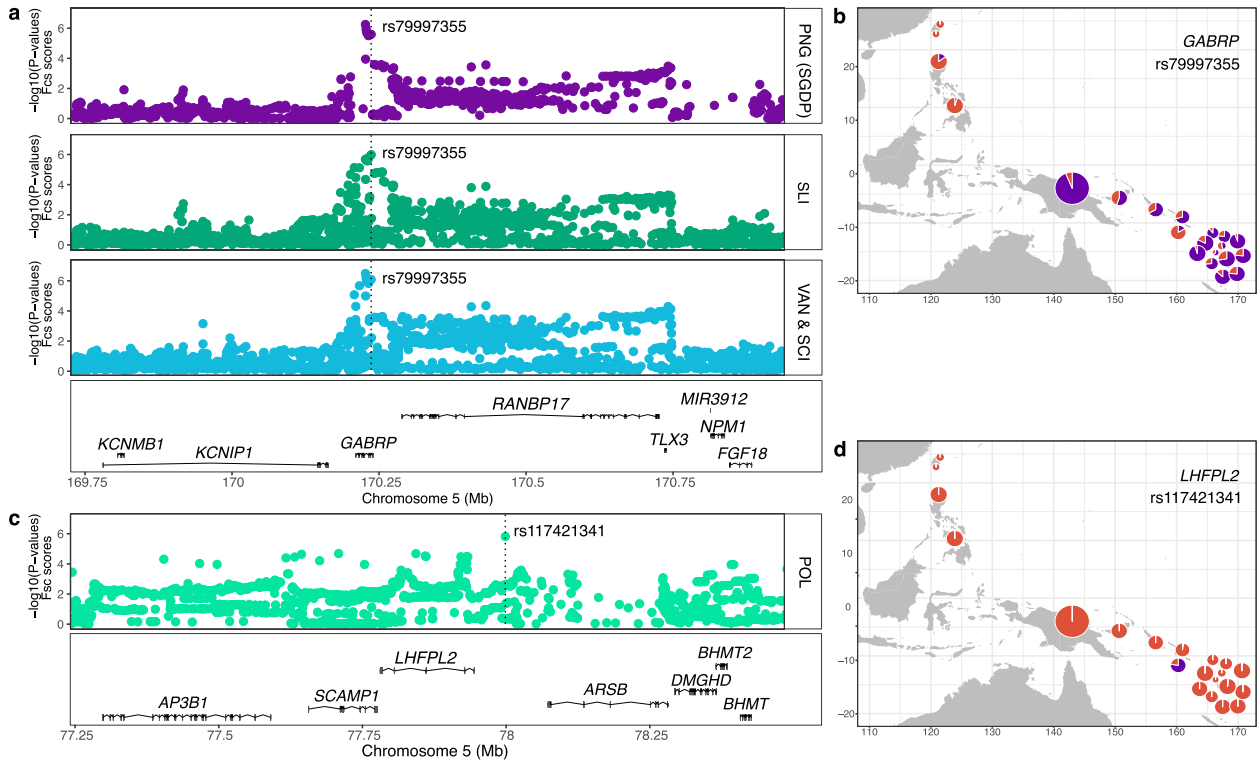
Extended Data Fig. 5 | Examples of candidate loci for adaptive introgression in Pacific populations. a, Adaptive introgression of Denisovan origin at the *CD33* locus. **b**, Adaptive introgression of Denisovan origin at the *IRF4* locus. **c**, Adaptive introgression of Neanderthal origin at the *KRT80* locus. **d**, Adaptive introgression of Neanderthal origin at the *TBC1D1* locus. **e**, Adaptive introgression of Denisovan origin at the *JAK1* 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 package[51].



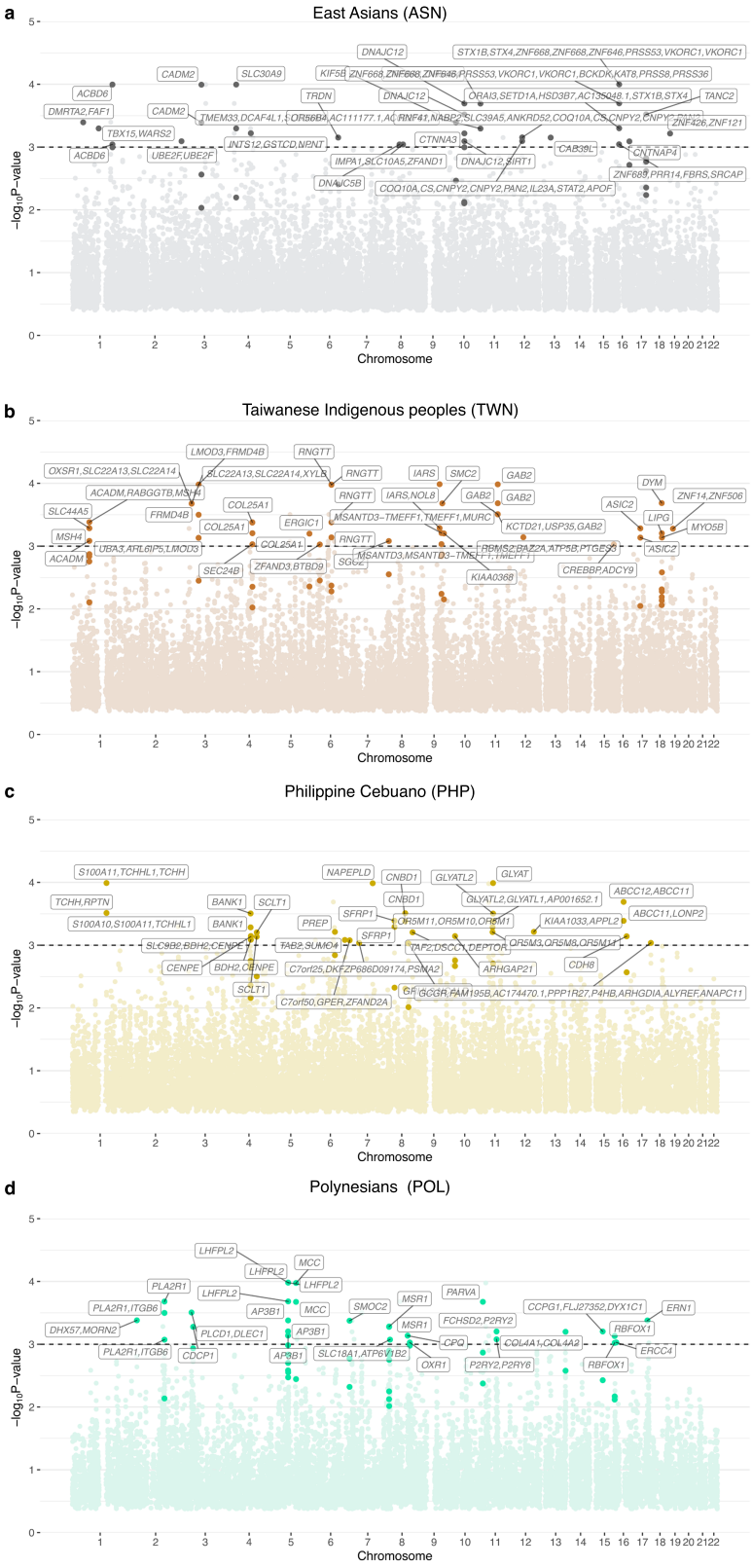
Extended Data Fig. 6 | Classic sweep signals detected in Papuan-related populations. a–d, Manhattan plots of classic sweep signals in Papua New Guinean Highlanders (**a**), Solomon Islanders (**b**), ni-Vanuatu (**c**) and Philippine Agta (**d**). **a–d**, The *y* axis shows the $-\log_{10}$(*P* value) 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.



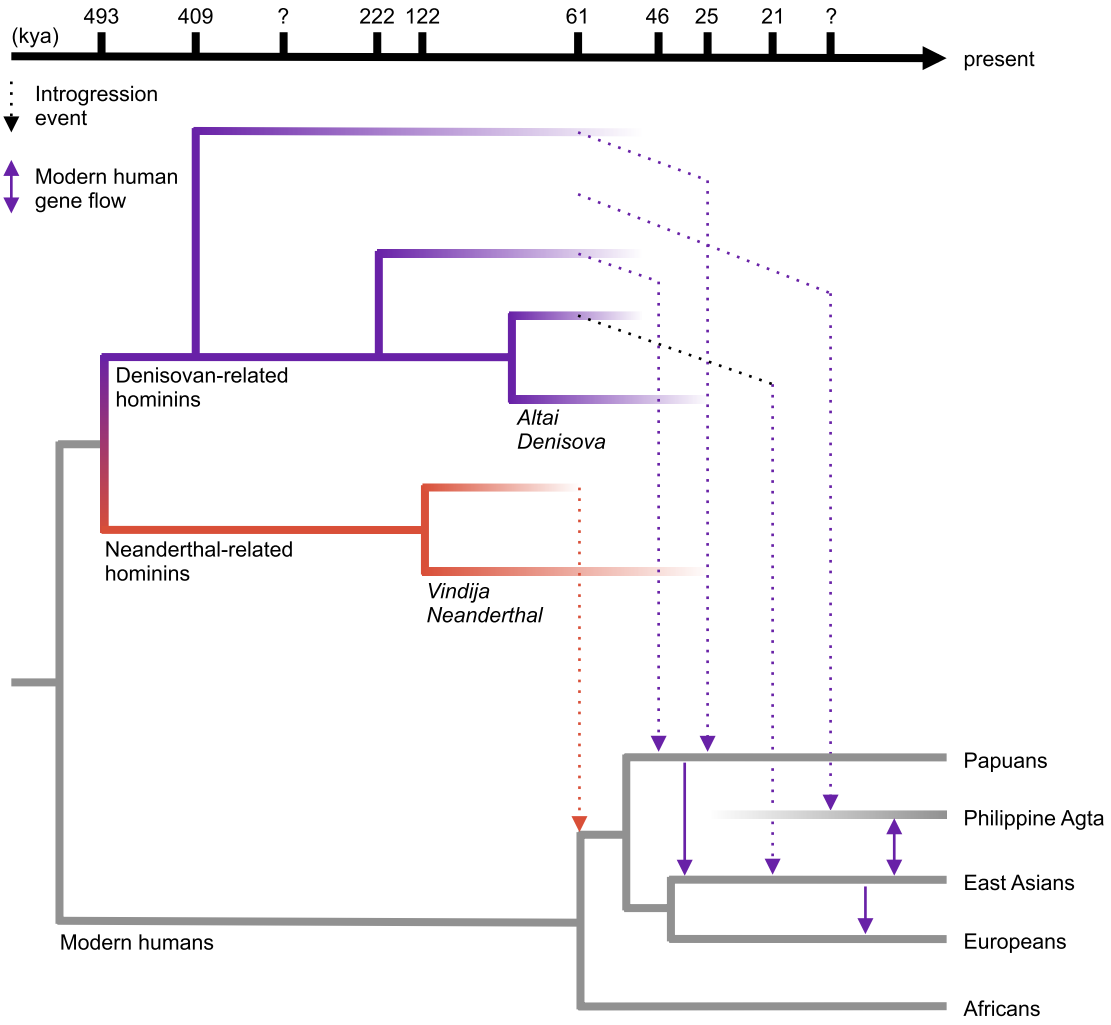
Extended Data Fig. 7 | Classic sweeps and adaptive archaic introgression. **a**, **b**, Coloured squares indicate genomic regions displaying signals of both a selective sweep and adaptive introgression from Neanderthals (**a**) or Denisovans (**b**). Yellow and blue frames indicate genomic regions identified 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 shows the $-\log_{10}$(*P* 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 package[51].



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.

nature research

Corresponding author(s): Lluis Quintana-Murci

Last updated by author(s): Dec 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Picard Tools v.2.8.1, BWA v.0.7.13, GATK v.3.8, vcftools 0.1.13, BCFtools v.1.8, PLINK v.1.9, KING v.2.1
Data analysis	EIGENSOFT v.7.2.1, ADMIXTURE v.1.22, PONG v.1.4, Haploview v.4.2, SHAPEIT2, fastsimcoal v.2.6, R v.3.4 or later, abc R package v.2.1, MetHis v.1.0, ADMIXTOOLS v.5.1.1, S-prime v.07Dec18.5e2, CRF (Sankararaman et al., Nature 2014), selink v.2 (www.github.com/h-e-g/selink), Arlequin v.3.5.2.2, other custom-generated scripts are deposited on GitHub (www.github.com/h-e-g/evoceania)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The whole genome dataset generated and analysed in this study is available from the European Genome-Phenome Archive (EGA), under accession code EGAS00001004540. The SGDP genome data were retrieved from the EBI European Nucleotide Archive (accession numbers: PRJEB9586 and ERP010710). The genome data from Malaspinas et al., Nature 2016 were retrieved from EGA (accession number: EGAD00001001634). The genome data from Vernot et al., Science 2016 were retrieved from dbGAP (accession number: phs001085.v1.p1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We sequenced the genome of >300 Pacific Islanders at high coverage (>30x), to describe the genetic diversity of human populations from this understudied region. Population genetics analyses were used to infer (i) the genetic structure, (ii) demographic history, (iii) the levels of archaic introgression and (iv) candidate loci and traits under positive selection in Near and Remote Oceanians.
Research sample	We sequenced the genome of 317 individuals from 20 human populations that were chosen to cover a geographic transect thought to underlie the peopling history of Near and Remote Oceania. This includes Taiwan, the Philippines, the Solomon Islands, Santa Cruz and the Vanuatu Archipelago. These newly sampled populations were analysed in combination with other populations from the Asia-Pacific region for which genomes are available, including Papua New Guinea, the Bismarck Archipelago and East Asia. Sampled individuals are meant to represent Near Oceanians (Papua New Guineans, Bismarck and Solomon islanders), western Remote Oceanians (ni-Vanuatu), Austronesian-speaking groups (Taiwanese aborigines, Philippine Cebuano), Polynesian-speaking populations (Polynesian outliers from the Solomon Islands), and Philippine 'Negritos' (Philippine Agta). The study sample was also chosen to characterize in great detail the genomic diversity of human populations that are understudied in human genomics.
Sampling strategy	Populations were sampled to cover a geographic transect thought to underlie the peopling history of Near and Remote Oceania. Sampling of related individuals was avoided, because relatedness can confound population genetics analyses. The ethno-linguistic group of sampled individuals was defined based on the self-declared group of their parents and grand-parents. An average of n = 16 unrelated individuals were sampled per population. Sample size for demographic inference with fastsimcoal2 is usually n = 5 (Malaspinas et al., Nature 2016). For archaic introgression and positive selection analyses, power mainly depends on other factors than sample size, but it is commonly accepted that n = 20 provides high power (Pickrell et al., Genome Res 2009). We thus merged closely-related populations into population groups for these analyses.
Data collection	All demographic information was collected through a structured questionnaire and/or ethnographic interviews. DNA was obtained from peripheral whole blood by venepuncture, or saliva by Oragene kits and cheek swabs. The sampling survey of Taiwanese aborigines was conducted by Albert Ko (Institute of Vertebrate Paleontology and Paleoanthropology, China). The sampling survey of Solomon Islanders was conducted by Mark Stoneking (Max Planck Institute for Evolutionary Anthropology, Germany). The sampling survey of Ni-Vanuatu was conducted by Olivier Cassar and Antoine Gessain (Institut Pasteur, Paris). The sampling survey of Philippine Negritos was conducted by Maximilian Larena (Human Evolution, Department of Organismal Biology, Uppsala University, Sweden).
Timing and spatial scale	The sampling survey of Taiwanese aborigines was conducted between 1998 and 2001. The sampling survey of Solomon islanders was conducted in August and September 2004. The sampling survey of ni-Vanuatu was conducted between April 2003 and August 2005. The sampling survey of Philippine Negritos was conducted between 2015 and 2018. The timing of sampling surveys was determined based on logistic requirements that depended on the accessibility of sampling sites and financial resources.
Data exclusions	Samples were excluded if they showed evidence of (i) DNA contamination, (ii) parental relatedness, (iii) relatedness to other samples, or (iv) genetic ancestry from populations outside of Oceania and East/Southeast Asia. All exclusion criteria were pre-established.
Reproducibility	We compared genotype calls obtained by next-generation sequencing to SNP genotyping arrays for the same individuals (unpublished data) and found very high concordance rates (>99.99%). No other experimental data were collected.
Randomization	To avoid batch effects, individuals were randomized according to their population of origin, across library preparation batches.
Blinding	Blinding was not relevant in this study because no condition or status was compared across sampled individuals.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Age, gender, ethno-linguistic group and genotypic information were collected for all human research participants. Participants include 173 males and 44 females, and were from 18 to 76 years of age. Ethno-linguistic groups are described in Supplementary Table 1. Genotyping rate was >95% for all participants, except one.
Recruitment	In each population, only unrelated volunteers with a self-reported ethno-llinguistic group were recruited from local villages. Sampling of related individuals was avoided because relatedness can confound population genetics analyses. The ethno-linguistic group of sampled individuals was defined based on the self-declared group of their parents and grand-parents. We do not anticipate any bias in our results that could be due to this recruitment strategy.
Ethics oversight	The study received approval from the Institutional Review Board of Institut Pasteur (n°2016-02/IRB/5), the Ethics Commission of the University of Leipzig Medical Faculty (n°286-10-04102010), the Ethics Committee of Uppsala University "Regionala Etikprövningsnämnden Uppsala" (Dnr 2016/103), as well as from local authorities including the Vanuatu Ministry of Health, the China Medical University Hospital Ethics Review Board, the National Commission for Culture and the Arts of the Philippines (in accordance with the provisions of Philippine Republic Act 7356, or the Law Creating the NCCA), and the Solomon Islands Ministry of Education and Training.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

nature portfolio

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

Supplementary information

Genomic insights into population history and biological adaptation in Oceania

In the format provided by the authors and unedited

Supplementary Information

Genomic insights into population history and biological adaptation in Oceania

Jeremy Choin, Javier Mendoza-Revilla, Lara R. Arauna, Sebastian Cuadros-Espinoza, Olivier Cassar, Maximilian Larena, Albert Min-Shan Ko, Christine Harmant, Romain Laurent, Paul Verdu, Guillaume Laval, Anne Boland, Robert Olaso, Jean-François Deleuze, Frédérique Valentin, Ying-Chin Ko, Mattias Jakobsson, Antoine Gessain, Laurent Excoffier, Mark Stoneking, Etienne Patin & Lluis Quintana-Murci

The **Supplementary Information** file contains: Supplementary Notes 1-18, which contain additional information on the methodology, results and discussion of the main text; Supplementary Figs. 1-82, which present additional results and quality checks; and Supplementary References.

Table of Contents

Supplementary Note 1: Population Sampling

DNA sampling in the Vanuatu

The Vanuatu archipelago is located in the Southwest Pacific and is part of Remote Oceania. Vanuatu contains 83 islands and forms a Y-shaped chain that spans nearly 1,100 km. Its current estimated population is 307,815. Indigenous Melanesians, called ni-Vanuatu, constitute 98.5% of the population. The sampling survey of ni-Vanuatu was conducted between April 2003 and August 2005 by Olivier Cassar and Antoine Gessain (Institut Pasteur, Paris) in remote villages located on 18 islands[1]. To avoid relatedness among individuals, couples were identified through ethnographic interviews, in English or Melanesian Pijin, and were preferentially sampled. Sex, age and living place, as well as date and place of blood collection, were collected through a structured questionnaire. 5-ml blood samples were obtained by venepuncture and transferred to the Institut Pasteur of New Caledonia, where plasma and buffy coats were isolated, frozen, and stored at –80°C. DNA was purified from frozen buffy coats at the Institut Pasteur of Paris, using QIAamp DNA Blood Mini Kit protocol, and eluted in AE buffer. DNA concentration was quantified with the Invitrogen Qubit 3 Fluorometer using the Qubit dsDNA broad-range assay. Prior to library preparation, DNA integrity was checked on agarose gels.

DNA sampling in Taiwan

Taiwanese indigenous peoples, also called Taiwanese aborigines, are ethnic groups that represent 2.4% of the total population of Taiwan, and are thought to have inhabited the island for at least 5,000 years ago (5 ka)[2,3]. Furthermore, archaeological remains suggest that Taiwan could have been settled as early as 20–30 ka[4]. Details about sampling of Taiwanese indigenous peoples (i.e., Paiwan and Atayal) and DNA extraction can be found elsewhere[5]. Briefly, samples were collected from 1998 to 2001 in indigenous villages, and their ethno-linguistic group was defined based on the group of their parents, using a structured questionnaire. Genomic DNA was extracted from peripheral whole blood by wizard genomic DNA purification kit (QIAGEN-Gentra Puregene Blood Kit) following standard laboratory protocols, and stored at –20°C. DNAs were made available by Ying-Chin Ko (Environment-Omics-Disease Research Center, China Medical University and Hospital, Taiwan). Prior to library preparation, DNA integrity was checked on agarose gels.

DNA sampling in the Philippines

The Philippines are an archipelago of 7,641 islands situated in Island Southeast Asia (ISEA), at the crossroads of historic human migrations in the Asia-Pacific region. Modern humans have inhabited the Philippine islands for ~47 ka[6], and it is thought that ancestors of Aeta, Ayta and Agta foragers (the so-called Philippine 'Negritos') are the archipelago's earliest inhabitants[7,8]. A large–scale sampling campaign was conducted by Maximilian Larena (Human Evolution, Department of Organismal Biology, Uppsala University, Sweden) from 2015 to 2018. Briefly, saliva samples were collected with the Oragene Saliva Collection Kit (DNA Genotek Inc, Canada). Only unrelated individuals, or only one individual from sets of individuals who self-reported to be up to 2nd-degree relatives, were included in the study. In addition, only individuals who self-reported to have all of their 4 grandparents to come from the same ethnic group were included in the study. The Philippine Negritos included in the study were asked with regards to the acceptability of the term 'Negrito'; all participants self-identify as Negritos and accept this term. The saliva samples were processed for DNA extraction at the Mattias Jakobsson Laboratory (Department of Organismal Biology, Uppsala University, Sweden), using the prepIT DNA isolation kit (DNA Genotek Inc., Canada). Prior to library preparation, DNA integrity was checked on agarose gels.

DNA sampling in the Solomon Islands

The Solomon Islands Archipelago consists of six major islands and >900 smaller islands lying to the east of Papua New Guinea and northwest of Vanuatu. It is believed that the

archipelago was first settled by modern humans ~30 ka[9]. The present-day population is constituted of 95.3% and 3.1% of peoples of Melanesian and Polynesian origins, respectively, the latter most likely originating from back migrations from Polynesia[10]. Cheek swab samples were collected across the Solomon Islands in August and September 2004. Details about sampling can be found elsewhere[11]. Self-described information on the birthplace, language, and ethnicity of each donor was obtained. DNA was extracted from the cheek swabs as described previously[12]. Six island populations were included in the current study, to represent Austronesian-speaking groups (Malaita and Santa Cruz Islands), Papuan-speaking groups (Vella Lavella Island) and Polynesian-speaking groups (so-called 'Polynesian outliers'; Bellona, Rennell and Tikopia Islands). Individuals associated to the Tikopia Island were recent migrants who traced their ancestry exclusively to Tikopia, but were sampled in Tikopian communities from other Solomon Islands. DNAs were made available by Mark Stoneking (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany). Prior to library preparation, DNA integrity was checked on agarose gels.

Ethical statements

All participants were volunteers of at least 18 years of age. Informed consent was obtained from each volunteer participant, including consent for genetic research, after the nature and scope of the study was explained in detail. The study received approval from the Institutional Review Board of Institut Pasteur (n°2016-02/IRB/5), the Ethics Commission of the University of Leipzig Medical Faculty (n°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 of the Philippines (in accordance with the provisions of Philippine Republic Act 7356, or the Law creating the NCCA), the Solomon Islands Ministry of Education and Training, and the Vanuatu Ministry of Health. 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. The present study was conducted in full respect of the legal and ethical requirements and guidelines for good clinical practice, in accordance with national and international rules. Namely, 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, (iv) principles promulgated in the Universal Declaration on the Human Genome and Human Rights, (v) the principle of respect for human dignity and the principles of non-exploitation, non-discrimination and non-instrumentalisation, (vi) the principle of individual autonomy, (vii) the principle of justice, namely with regard to the improvement and protection of health and (viii) the principle of proportionality. The rights and welfare of the subjects have been respected, and the hazards have not outweighed the benefits of the study.

Supplementary Note 2: Whole-genome Sequencing

Library preparation

Whole-genome sequencing (WGS) was performed on a total of 317 individuals from Taiwan (i.e., the Paiwan and the Atayal), the Philippines (i.e., the Agta and the Cebuano), the Solomon Islands (i.e., populations from Malaita, Vella Lavella, Rennell, Bellona, Tikopia and Santa Cruz islands), and 10 islands from the Vanuatu archipelago (Supplementary Table 1). WGS was performed at the CNRGH (*Centre National de Recherche en Génomique Humaine, Institut de Biologie François Jacob*, Evry, France). For 298 samples, a PCR-free library preparation was obtained with the Illumina TruSeq DNA PCR-free Library Preparation Kit from 1µg of genomic DNA. For the remaining 19 samples, a PCR-based library preparation was obtained with the Illumina TruSeq DNA Nano Library Preparation Kit from 100 ng of genomic DNA (Supplementary Table 1). After normalisation and quality control, qualified libraries were sequenced on a HiSeqX5 Illumina platform (Illumina Inc., CA, USA) to obtain paired-end 150-bp reads. One lane of HiSeqX5 flow cell was produced for each blood-derived DNA sample. Additional sequencing was produced for saliva-derived DNA samples, to reach an average sequencing depth of 30×.

Read mapping and variant calling

Sequence quality parameters were assessed throughout the sequencing run. *FASTQ* files for each sample were generated using the standard Illumina pipeline. *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 then mapped onto the human reference genome hs37d5, using the 'mem' algorithm from Burrows–Wheeler Aligner version 0.7.13[13], and duplicates were marked with Picard Tools. Base quality scores were recalibrated using GATK version 3.8[14]. WGS data from Vernot *et al*.[15] were processed as the newly-generated genomes. WGS data from Malaspinas *et al*.[16] and the SGDP[17] were first converted to raw *BAM* files into *uBAM* files, and then processed as previously described.

Variant calling was performed following the GATK Best Practices recommendations (https://software.broadinstitute.org/gatk/best-practices/), and using GATK version 3.8[18]. All samples were genotyped individually using 'HaplotypeCaller' in *gvcf* mode. We turned off the PCR indel correction of 'HaplotypeCaller' ('-pcr_indel_model NONE') for the 298 samples prepared following a PCR-free protocol, as well as for 122 out of the 133 samples from SGDP[17] that were prepared following a PCR-free protocol. A final step of joint genotyping was performed to create a raw *multisample VCF*, using the 'GenotypeGVCFs' tool with the option '-allSites', to include homozygous reference sites.

Ancestral state annotation

The ancestral state for any given site was defined as the allele present in the chimpanzee reference genome (panTro4) aligned against hg19 (ref.[19]), which was downloaded from the UCSC platform[20]. Sites not present in the chimpanzee genome, or containing alleles that did not match either the reference or alternative allele, were discarded.

Quality-control filters

Methods. We split the dataset into two *VCFs*: one with only autosomal homozygous reference sites (i.e., invariant sites) and a second with only autosomal variant sites. Variant sites were first filtered using GATK 'VQSR'[21] with a truth sensitivity cut-off of 99.5 for SNPs ('--ts_filter_level 99.5') and 99 for INDELs ('--ts_filter_level 99'). A series of hard filters were applied on invariant and variant sites, using BCFtools version 1.8 (http://www.htslib.org/); we set as missing all genotypes with (i) a depth (DP) < 10 or (ii) DP > twice the sample coverage, and (iii) a genotype quality (GQ/RGQ) < 30. Additional, *ad hoc* filters were applied (Supplementary Fig. 1). At *Level 1*, we removed sites that were missing in more than 5% of the samples and/or were in Hardy–Weinberg disequilibrium (*P*-value < 10^{-4}) in at least one of

the populations. At *Level 2*, we removed sites that were missing in at least one sample (i.e., 0% of missingness) and/or were in Hardy–Weinberg disequilibrium (*P*-value < 10^{-4}) in at least one of the populations. At *Level 3a*, we removed, in addition to *Level 2* filters, sites (i) within CpG islands, obtained from the UCSC table browser[20]; (ii) within genes, obtained from Ensembl BioMart version 97; and (iii) outside of Vindija Neanderthal and Altai Denisovan accessibility masks, downloaded from: http://ftp.eva.mpg.de/neandertal/Vindija/FilterBed/. These masks exclude sites (i) where at least 18 of 35 overlapping 35-mers are mapped elsewhere in the human genome with zero or one mismatch; (ii) with minimum coverage of 10; (iii) with mapping quality < 25; (iv) with tandem repeats; (v) with indels; and (vi) with coverage filters stratified by GC content. At *Level 3b*, we only removed, in addition to *Level 2* filters, sites (i) within CpG islands, (ii) within segmental duplications (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ %20genomicSuperDups.txt.gz), and (iii) where at least 18 of 35 overlapping 35-mers are mapped elsewhere in the human genome with zero or one mismatch. At *Level 3b'*, in addition to *Level 3b*, we excluded sites outside of Vindija Neanderthal, Altai Neanderthal and Altai Denisova accessibility masks.

Per sample heterozygosity was computed with PLINK version 1.90[22,23] with the '--het' argument. We defined as a heterozygosity outlier, a sample presenting a level of heterozygosity at least 3 standard deviations (SD) lower or higher than the population mean, reflecting high parental relatedness or contamination, respectively. To identify cryptically-related samples, we used kinship values inferred by KING version 2.1[24]. We considered a pair of samples as related if they presented a kinship coefficient > 0.08, a threshold that is slightly more stringent than a second-degree of relatedness as defined by KING[24]. To maximise sample size, we excluded related samples using an iterative approach, as described elsewhere[25].

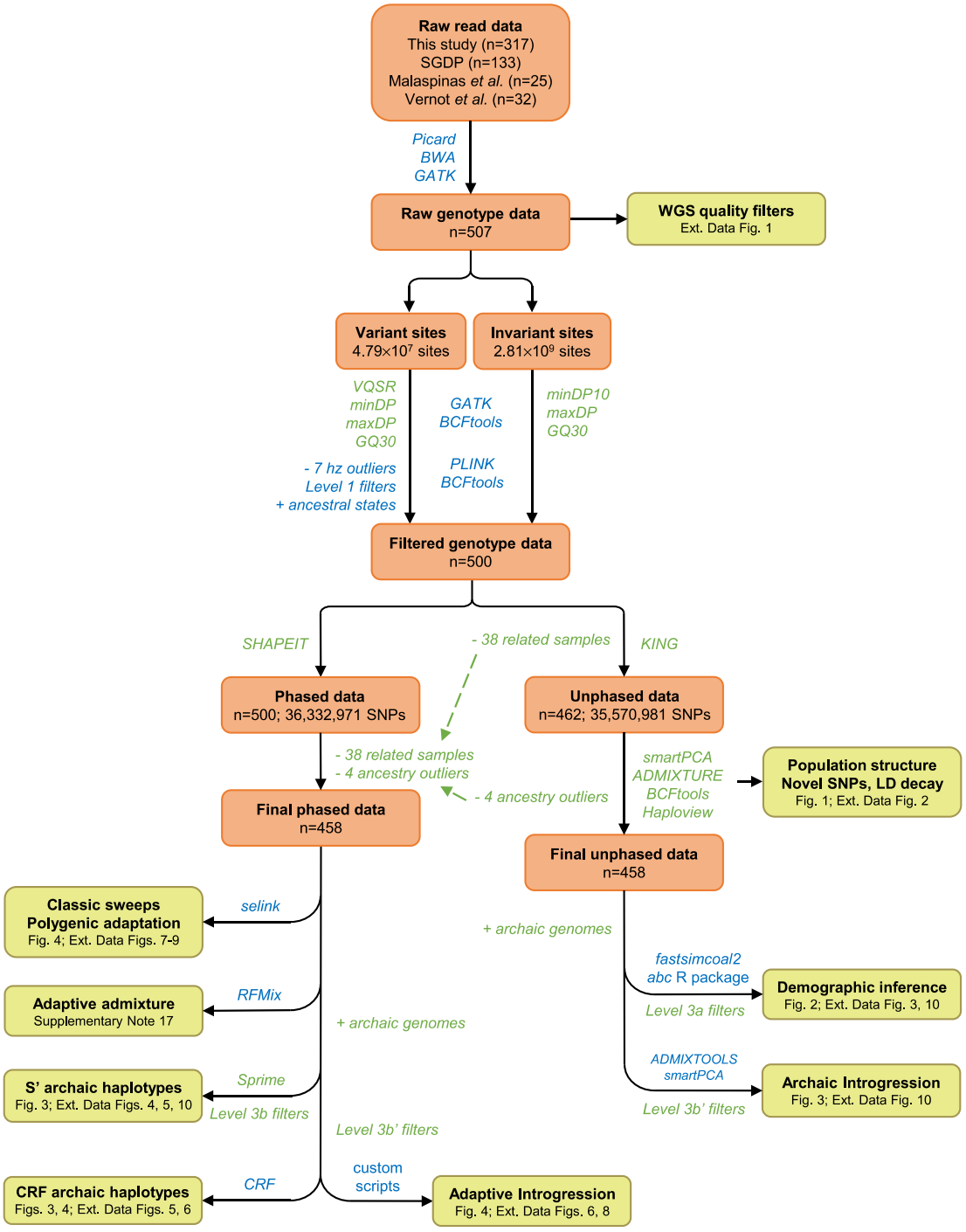
Results. Out of the 317 whole-genome sequenced samples, we identified 3 samples as potentially contaminated (heterozygosity > mean + 3SD), and 4 samples that presented evidence for parental relatedness (heterozygosity < mean – 3SD) (Supplementary Fig. 2a,b). Among the remaining individuals, and based on kinship coefficients, we inferred 21 pairs of samples that were 1st-degree related (0.177 < kinship < 0.354), 19 2nd-degree related (0.0884 < kinship < 0.177), and 3 ambiguous between 2nd-degree and 3rd-degree related (0.08 < kinship < 0.0884) (Supplementary Fig. 2b). In total, we removed 39 samples from our collection of 317 newly-generated genomes, including 7 samples with outlier heterozygosity, and 32 cryptically related samples. In addition, we removed 6 samples from Vernot *et al*.[15] that our analysis identified as related, leading to a final dataset of 462 unrelated samples. Among these, a total of 36,339,995 bi-allelic SNPs were identified, 35,870,981 of which segregate in the sample (i.e., the two alternative alleles are observed in the sample).

Data quality checks

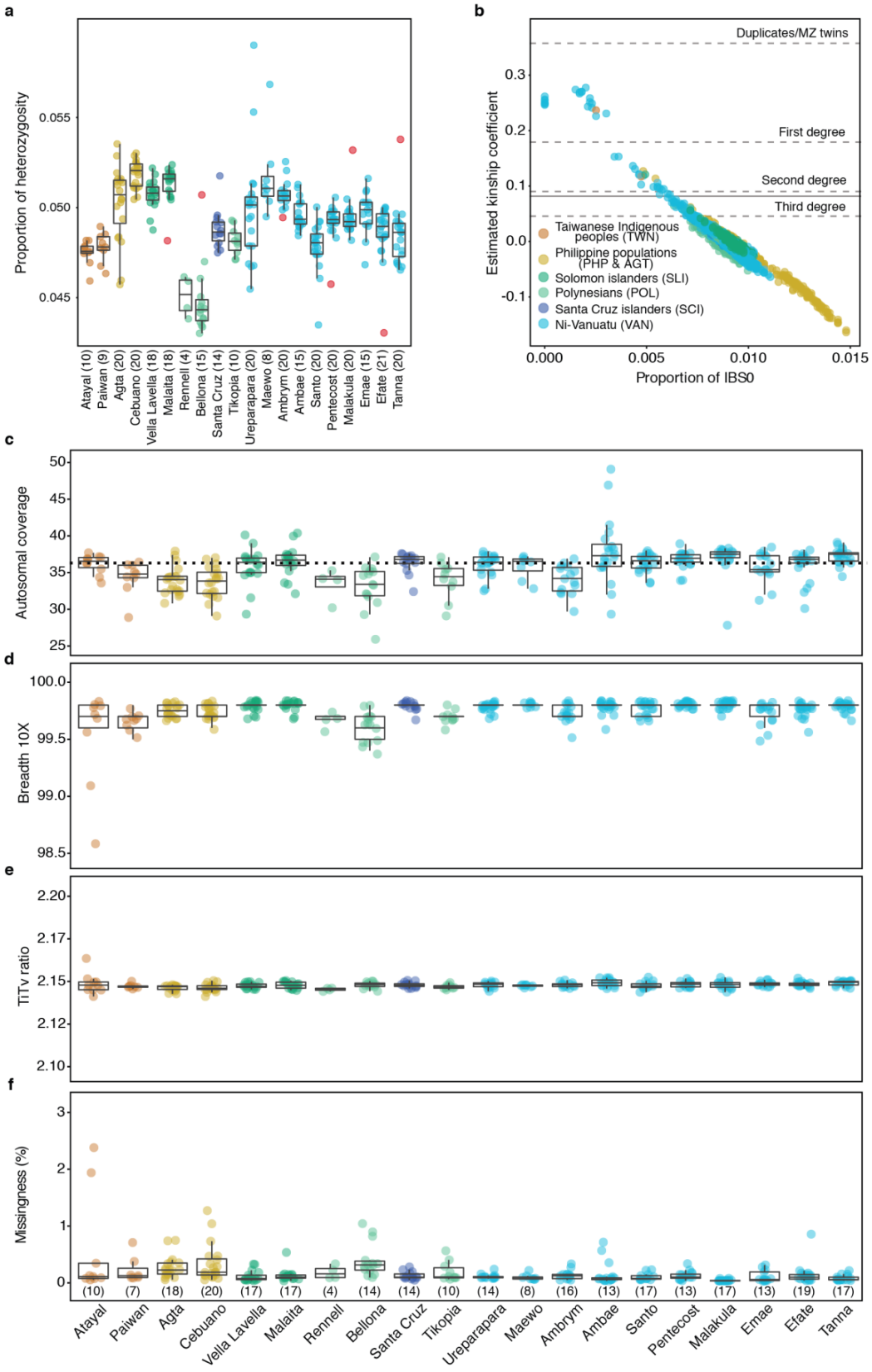
Methods. Sequencing quality was first assessed by a set of statistics, using 'GATK DepthOfCoverage' version 3.8[18] and BCFtools version 1.8 (http://www.htslib.org/) on individual *BAM* files and on *VCF* files (507 samples, 39,035,215 bi-allelic SNPs, no fixed homozygous reference site). Specifically, we evaluated the percentage of the genome covered at least at 10× (breadth of coverage 10× on the *BAM* files) as well as *VCF*-based mean coverage per individual. A second round of quality checks was performed on a dataset filtered at *Level 1* (Supplementary Fig. 1), after removing heterozygosity outliers and individuals presenting signs of cryptic relatedness (i.e., 462 samples). We inspected the Transition/Transversion ratio (Ti/Tv), which is insensitive to ancestry and should be ~2.0-2.1 for whole genome sequencing[26], and the per-sample missingness (i.e., the number of sites missing over the total number of sites in the *VCF*).

Results. The mean coverage per sample ranged between 26× and 49×, with a median of 36×, and the breadth of coverage at 10× varied between 94.7% and 99.8% (Supplementary Fig. 2c,d). The value of the Ti/Tv ratio was homogenous across samples and was ≈2.1 (2.14-2.16) (Supplementary Fig. 2e). No individual genome presented a missingness >5%

(Supplementary Fig. 2f), except the Atayal B00FLGA (missingness = 6.9%). Quality statistics indicated that the newly-generated whole genomes were of high quality.



Supplementary Figure 1. Analysis flowchart of the whole-genome dataset. Salmon and tan boxes indicate datasets and analyses, respectively. Blue and green text indicates computer programs and filters applied on the whole-genome data, respectively.



Supplementary Figure 2. Processing and quality of the whole-genome dataset. a, Individual heterozygosity. Red dots indicate outliers relative to the mean heterozygosity of the population. **b**,

Cryptic relatedness between individuals. Dashed lines indicate kinship thresholds for the 1st, 2nd and 3rd degrees of relatedness, according to KING (ref.[56]). The solid line indicates the threshold used to consider a pair of individuals as cryptically related at either the 1st or 2nd degree. **c**, Autosomal depth of coverage. The dashed line indicates the median coverage (36×). **d**, Breadth of coverage 10× (i.e., percentage of the genome covered at >10×). **e**, Transition-Transversion (Ti/Tv) ratio. **f**, Missingness (i.e., percentage of missing genotypes). **d, f,** To facilitate visualisation, the Atayal sample B00FLGA was not plotted (breadth 10×=94.7, missingness=6.9%). **a-d**, Per-population sample size is shown in brackets in **a**. **e-f**, Per-population sample size is shown in brackets in **f**. **a,c-f**, The line, box, whiskers and points respectively indicate the median, interquartile range (IQR), 1.5*IQR and outliers.

Novel SNPs relative to other datasets

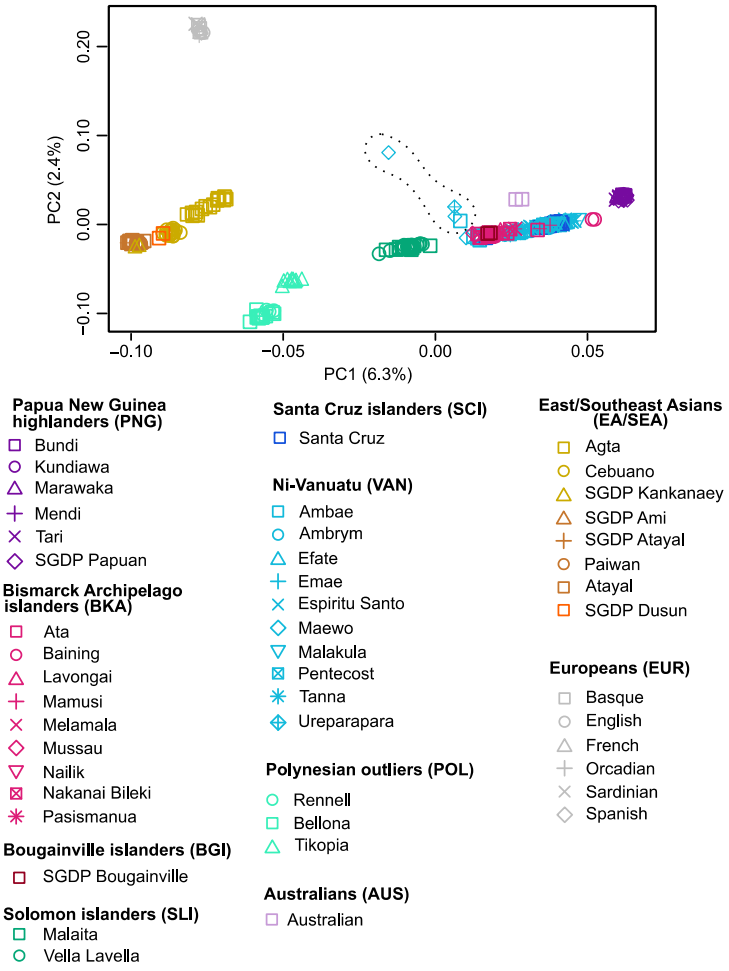
We generated 4 different datasets with PLINK version 1.90 (refs.[22,23]): (i) 25 Papuans from Malaspinas *et al*.[16], (ii) 26 individuals from the Bismarck Archipelago from Vernot *et al*.[15], (iii) 133 samples from the SGDP[17], and (iv) 278 unrelated samples sequenced for this study. For each dataset, we removed invariant sites and counted the number of variant sites, i.e., all remaining bi-allelic SNPs. Using BCFtools 'isec' (http://www.htslib.org/), we intersected variant sites found in the new genomes with (i) the three other datasets[15-17], (ii) SNPs found in dbSNP database build 152[27] (downloaded from https://www.ncbi.nlm.nih.gov/snp/), and (iii) the union of the four. We considered a SNP as novel if its chromosomal position and both its reference and alternative alleles were not found in the intersected dataset ('bcftools isec --collapse none --complement'). Novel SNPs were then divided into three categories based of their minor allele frequency (MAF): (i) rare variants (MAF < 1%), (ii) intermediate variants (1% < MAF < 5%), and (iii) common variants (MAF > 5%).

Supplementary Note 3: Genetic Structure and Diversity

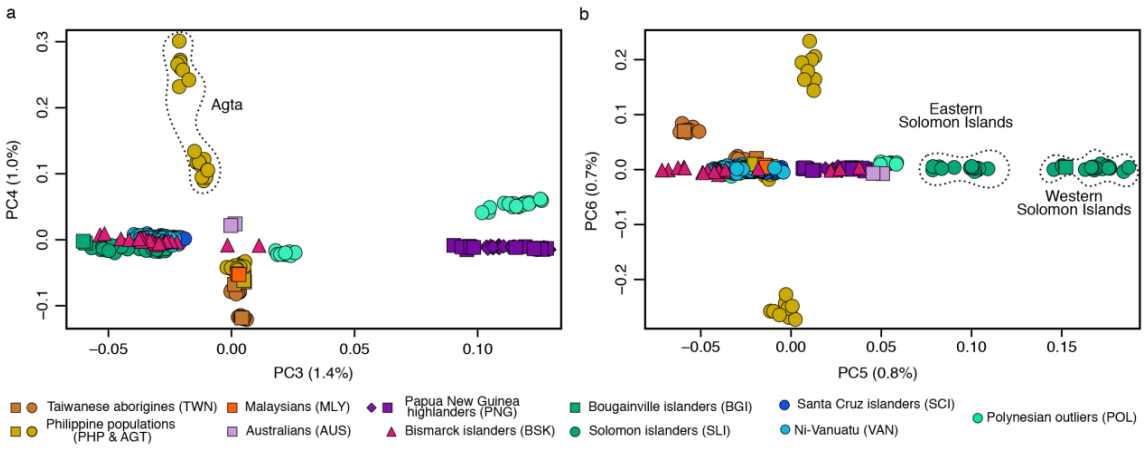
Principal component analyses

Methods. PCA were performed on the *Level 1* dataset (Supplementary Note 2). A first PCA comprised 355 unrelated individuals from the Pacific region, as well as a group of Europeans from SGDP[17], to detect recent admixture of Oceanians with Europeans. The second PCA was performed on the same dataset, excluding Europeans and four Vanuatu samples presenting non-negligible proportions of African or European ancestry (Supplementary Table 1). Additional variant pruning was performed with PLINK version 1.90 (refs.[22,23]), excluding variants with MAF < 5% and pairs of common SNPs in strong linkage disequilibrium (LD) using the '--indep-pairwise 50 5 0.5' argument. PCA was performed using the 'SmartPCA' algorithm implemented in EIGENSOFT version 6.1.4 (ref.[28]).

Results. In a PCA of populations from the Pacific, East/Southeast Asia and Europe, PC1, which explains 6.3% of the variance, separates Papua New Guinean highlanders (PNG) from East/Southeast Asians and Europeans, whereas PC2 (2.4% of the variance) separates East/Southeast Asians from Europeans (Supplementary Fig. 3). Notably, four individuals from Vanuatu show suggestive evidence for European or African ancestry. In a PCA of populations from the Pacific and East Asia only, PC1 separates PNG from East/Southeast Asians (6.3% of the variance), whereas PC2 (1.8% of the variance) separates East/Southeast Asians from Polynesian-speaking populations of the Solomon Islands, i.e., Polynesian outliers (Fig. 1d). Populations from the Bismarck Archipelago, the Solomon Islands and Vanuatu and Polynesian outliers form a cline between PNG and East/Southeast Asians on PC1, suggesting varying levels of East Asian-related ancestry (Fig. 1). PC3 (1.4% of variance) separates PNG and Polynesian outliers from all other populations, whereas PC4 (1.0% of variance) separates the Philippine Agta from East/Southeast Asians (Supplementary Fig. 4). PC5 (0.8% of the variance) separates western and eastern Solomon islanders from all other groups, suggesting contrasting demographic pasts in western and eastern Solomon Islands[10,29,30]. Finally, PC6 (0.7% of the variance) separates the Philippine Agta into two populations (Supplementary Fig. 4). Together, these results indicate that population genomic variation in the Pacific is best explained by four genetic clusters, associated with (i) East/Southeast Asians including Austronesian speakers, (ii) PNG, (iii) Bismarck, Solomon and Vanuatu islanders, and (iv) Polynesian outliers. The largest differences are between East/Southeast Asians and PNG, while the remaining populations show varying proportions of the two components, in agreement with an expansion of East Asian-related Austronesian speakers across the region, followed by admixture[30-32].



Supplementary Figure 3. PCA of whole genomes of populations from the Pacific, East/Southeast Asia and Europe. These include Europeans[17], East/Southeast Asians[17], indigenous Australians[17], Papua New Guinea highlanders[16,17], Bismarck islanders[15], as well as the populations studied here (i.e., Taiwanese indigenous peoples, Philippine populations, Solomon islanders and ni-Vanuatu). The variance explained by each PC is indicated in brackets. The dashed area indicates samples with non-negligible proportions of African or European ancestry (Supplementary Table 1).

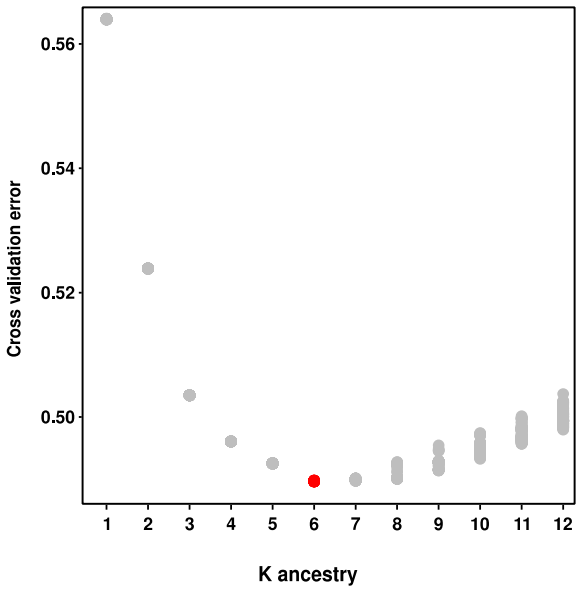


Supplementary Figure 4. PCA of whole genomes of Pacific populations. **a**, PC3 versus PC4, **b**, PC5 versus PC6. The variance explained by each PC is given in brackets.

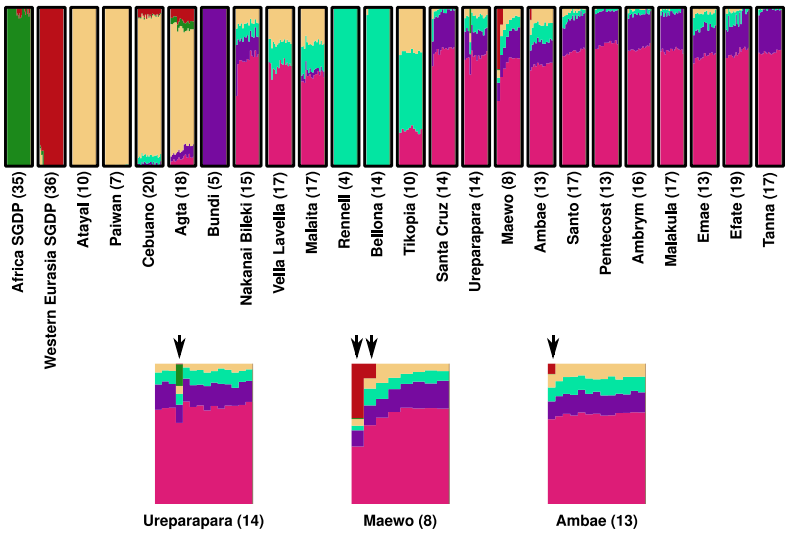
Admixture analyses

Methods. Genetic clustering analysis was performed on the *Level 1* dataset (Supplementary Note 2). To estimate the proportions of *K* genetic components for each individual, we used the unsupervised model-based clustering algorithm ADMIXTURE[33]. We ran the algorithm with *K* = 1 to *K* = 12, 100 times with different random seeds, including the argument 'cv' to output cross-validation errors. Results were visualised using 'PONG' version 1.4[34] in major-mode. All f_4-statistics were computed with ADMIXTOOLS version 5.1.1. To estimate allele sharing of Vanuatu populations with Polynesian outliers, we computed f_4-statistics of the form f_4 (Polynesian outliers, Taiwanese indigenous peoples; ni-Vanuatu, Mbuti) for each Vanuatu population. We grouped the Polynesian outliers from the dataset (Tikopia, Rennell and Bellona islanders) into a single group, and grouped Atayal and Paiwan as a single Taiwanese indigenous group. Given that Polynesian outliers appear to descend from admixture between an East Asian-related and a Papuan-related population, as shown by previous studies[35,36] and here, differences among Vanuatu populations in their affinity to Polynesians could be driven by the Papuan-related ancestry proportions of each Vanuatu population. To correct for this potential confounder, we computed another f_4-statistic of the form f_4(Papuans, Taiwanese indigenous peoples; ni-Vanuatu, Mbuti), which tests the affinities of each Vanuatu population with Papuans and Taiwanese indigenous peoples.

Results. In agreement with PCA results, ADMIXTURE at *K* = 2 identified two genetic components that are maximised in (i) East Asia, the Philippines and Polynesia, and (ii) PNG, Bismarck, Solomon and Vanuatu islanders. Varying proportions of the East Asian component were estimated across Near and Remote Oceanians (Extended Data Fig. 1), which has been attributed to the Holocene expansions of Austronesian speakers likely originating from Taiwan[30-32]. At increasing *K* values, ADMIXTURE identified new components that are maximized in Africans, Europeans and PNG, mirroring human structure at the worldwide scale[17]. At *K* = 6, for which the cross-validation error was minimal (Supplementary Fig. 5), a component specific to Polynesian outliers was apparent, suggesting that genetic drift, probably because of serial founder events, has increased genetic differentiation of these groups with respect to other Oceanian populations. Of note, four individuals from Vanuatu (i.e., from Ureparapara, Maewo and Ambae islands) showed non-negligible proportions of African or European ancestry (Supplementary Figs. 3 and 6 and Supplementary Table 1), and were discarded from subsequent analyses. At *K* = 7, ADMIXTURE analyses supported a component that is maximized in Solomon islanders, in agreement with PCA (Supplementary Fig. 4b). At *K* = 8, a component specific to the Philippine Agta was found, suggesting a history of genetic isolation from other Philippine populations, as previously suggested[7]. Nevertheless, we caution that the low sample size of the Agta in our dataset might result in the underestimation, by ADMIXTURE, of their genetic differentiation from other populations. Intriguingly, for *K* > 5, indigenous Australians show a pattern consistent with admixture between different components maximized in Near Oceanians. We suggest that this pattern is probably due to increased genetic drift in the latter groups[37], as previously suggested[30], and to their low sample size. For all *K* values explored, we observed strong genetic affinities between individuals of the Bismarck archipelago and of Vanuatu and Santa Cruz islands (Fig. 1c,d, Extended Data Fig. 1), consistent with a post-Lapita expansion of Bismarck islanders into Remote Oceania[31,32,38].



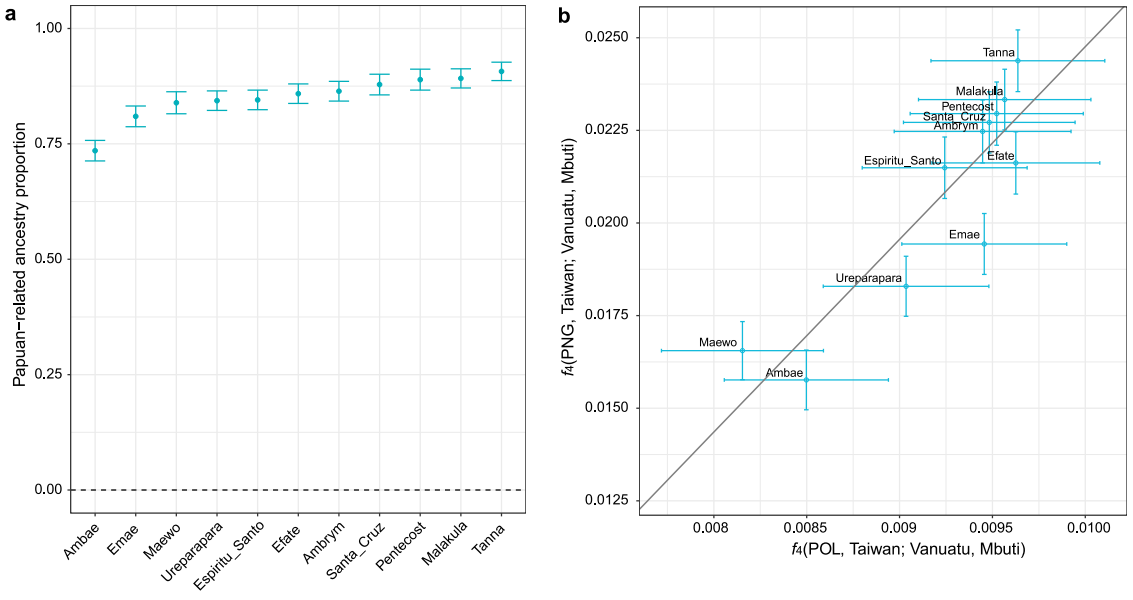
Supplementary Figure 5. ADMIXTURE cross-validation error. A hundred independent runs were performed with different random seeds, for each *K* prior value. The red dot indicates the *K* value with the lowest cross validation error (*K* = 6).



Supplementary Figure 6. European and African components in Oceanians. ADMIXTURE results at *K* = 6 for the 20 study populations, together with selected populations from Malaspinas *et al*.[16] (Bundi PNG), Vernot *et al*.[15] (Nakanai Bileki from the Bismarck Archipelago) and SGDP[17] (Africans and western Eurasians). Detailed ADMIXTURE results are shown in the bottom, where individuals showing non-negligible proportions of African or European components are indicated by an arrow.

At *K* = 6, we observed various levels of Polynesian ancestry among Vanuatu populations (Extended Data Fig. 1). Interestingly, Polynesian outliers – Polynesian-speaking people living outside Polynesia – are known to reside in different Vanuatu islands, including Emae, Mele, Ifira, Futuna and Aniwa[39]. To test if some Vanuatu populations show increased Polynesian-related ancestry, we computed f_4-statistics[40] that estimate allele sharing of Vanuatu populations with either Polynesians or Papuans, to account for the various proportions of Papuan-related ancestry in the ni-Vanuatu (Supplementary Fig. 7a). In contrast with other

Vanuatu populations, Emae islanders showed higher allele sharing with Polynesians than expected, given their Papuan-related ancestry (Supplementary Fig. 7b). This suggests that migrations from Polynesia were more frequent in Emae, relative to the other western Remote Oceanian islands. This is in agreement with linguistic and genetic evidence suggesting back migrations from Polynesia into Vanuatu islands where Polynesian outliers reside[39,41].

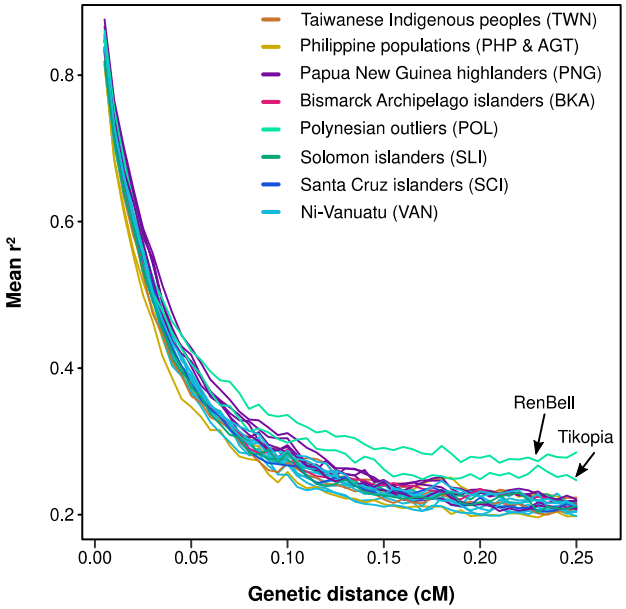


Supplementary Figure 7. Proportions of Papuan-related and Polynesian-related ancestry in western Remote Oceanian populations. **a**, Various proportions of Papuan-related ancestry among western Remote Oceanian populations. Estimates were obtained with a f_4-ratio of the form f_4(Han, Mbuti; Vanuatu, Papuans) / f_4(Han, Mbuti; Taiwanese indigenous peoples, Papuans). Bars indicate two standard errors. Standard errors were calculated using a weighted-block jackknife procedure dropping 5-cM blocks of the genome in each run. The sample size (n) of each population is detailed in Supplementary Table 1. **b**, Allele sharing of western Remote Oceanian populations with Polynesian outliers, accounting for Papuan-related ancestry. For each western Remote Oceanian population, allele sharing with Polynesian outliers, relative to Taiwanese indigenous peoples, is shown against allele sharing with Papuans, relative to Taiwanese indigenous peoples. Bars indicate two standard errors for all f_4-statistic estimates. The black line indicates the regression line of a linear model of all populations (n=11 populations).

Linkage disequilibrium decay

Methods. Linkage disequilibrium (LD) between pairs of SNPs was estimated based on r^2 values with Haploview[42]. As r^2 is sensitive to sample size, we randomly sampled 5 individuals per population in the *Level 1* dataset (Supplementary Note 2). We then removed bi-allelic SNPs with a population MAF < 5%. r^2 values were computed for every pair of SNPs using a 1-Mb sliding window approach, and were averaged per bin of genetic distance using the 1000 Genomes Project Phase 3 genetic map (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140404_data_for_phase3_paper/shapeit2_scaffolds/)[43].

Results. Polynesian outliers, originating from Rennell, Bellona and Tikopia islands, showed slower LD decay with genetic distance, relative to other populations (Supplementary Fig. 8). As LD decay depends on effective population size[44], these results suggest a lower N_e for Polynesian outliers, which may be attributed to the serial founder events experienced by these populations following the settlement of Polynesia, and/or back migrations from Polynesia to Near Oceania[10].

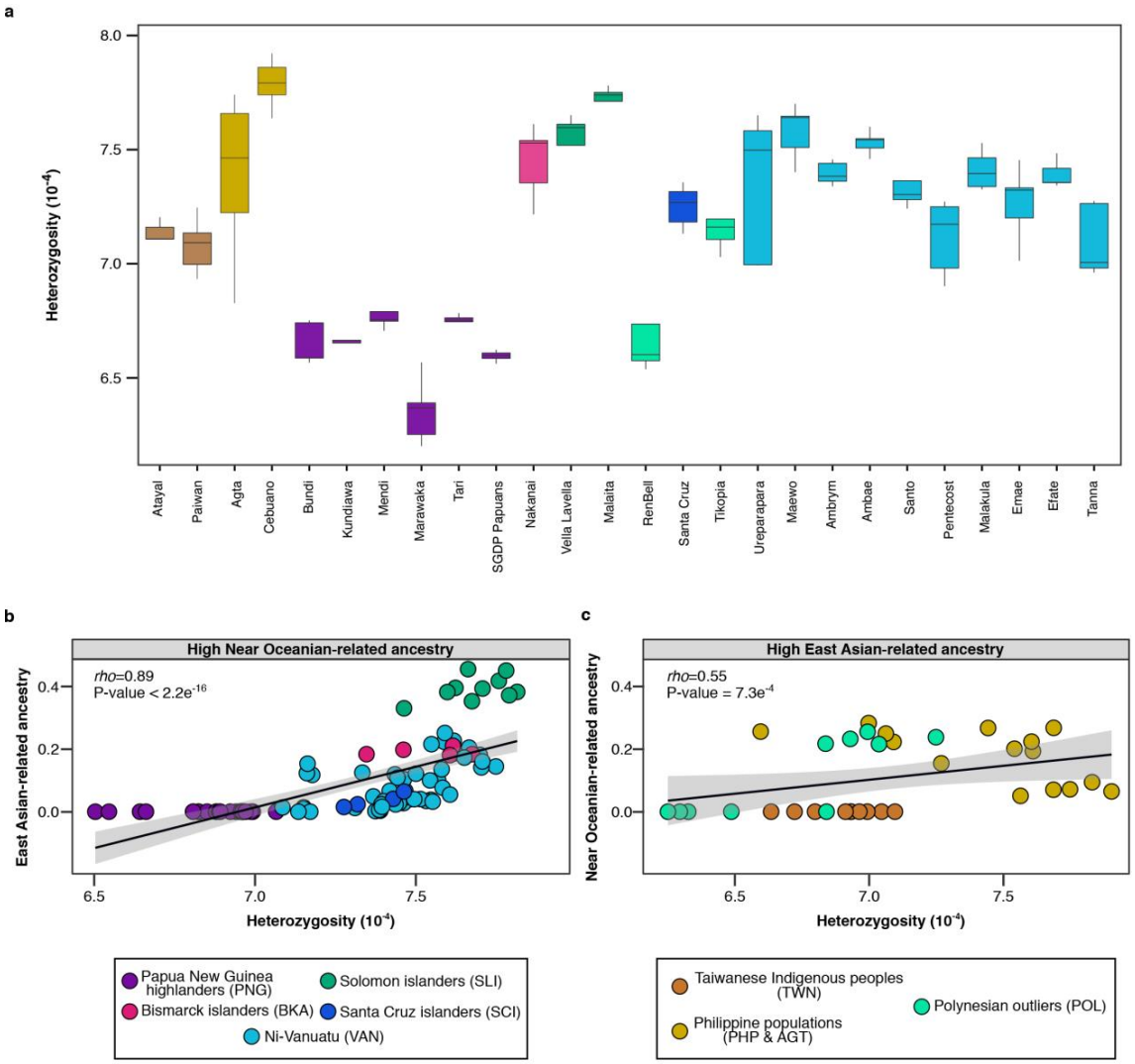


Supplementary Figure 8. Linkage disequilibrium decay with genetic distance in Pacific populations. Each population is composed of 5 samples, which were randomly chosen. RenBell indicates Polynesians from Rennell and Bellona.

Heterozygosity and admixture proportions

Methods. We computed heterozygosity, for each population with a sample size > 5 and for each sample, as the number of heterozygous sites divided by the total number of callable positions (i.e., all variant and invariant sites) using BCFtools version 1.8. Individual admixture proportions were obtained from ADMIXTURE at K = 6 (Fig. 1c).

Results. Levels of heterozygosity differed markedly across populations (Kruskal Wallis P-value = 1.4×10^{-12}; Fig. 1e and Supplementary Fig. 9a). Pacific populations with the lowest heterozygosity were those showing no evidence of admixture (Fig. 1c and Extended Data Fig. 1), including PNG and Taiwanese indigenous peoples. The Kundiawa and the Bundi showed heterozygosity levels comparable to those of other groups of PNG, suggesting they well represent other Papua New Guinean populations. Populations with low heterozygosity also include Polynesian outliers (Supplementary Fig. 9a), who likely experienced founder events following the settlement of Polynesia, and/or back migrations from Polynesia to Near Oceania[10]. Heterozygosity of Polynesian speakers from Rennell and Bellona Islands was substantially lower than that of Polynesian speakers from Tikopia, suggesting increased genetic drift in the former, in agreement with their higher levels of LD (Supplementary Fig. 8), ADMIXTURE results[37] at K = 6 (Fig. 1c and Extended Data Fig. 1), and previous observations based on levels of LD, runs of homozygosity and genetic differentiation[30]. This suggests that Polynesian outliers experienced founder effects of various intensities, following their back migrations from Polynesia to the Solomon Islands. Finally, we observed a significant correlation between heterozygosity and East Asian-related admixture proportions in Oceanians (Supplementary Fig. 9b; r^2 = 0. 89, P-value < 2.2×10^{-16}), indicating that admixture is a key determinant of heterozygosity levels in the region.

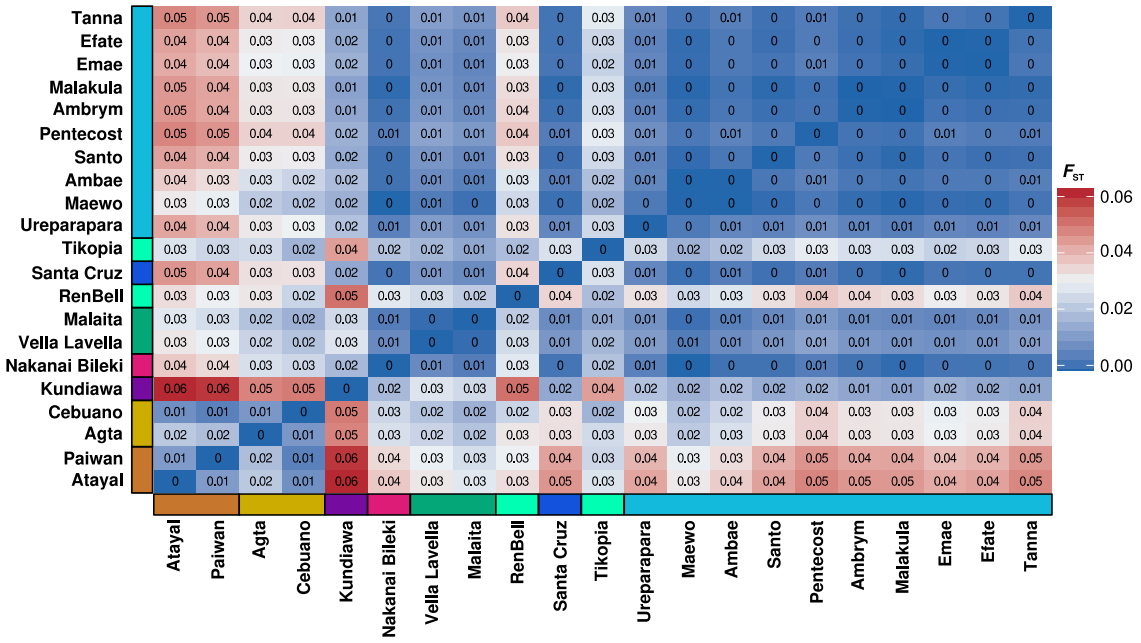


Supplementary Figure 9. Heterozygosity and estimated admixture proportions in Pacific populations. **a**, Population levels of heterozygosity. For each population with a sample size ≥ 5, five samples were randomly sampled to obtain equal sizes. The line, box, whiskers and points respectively indicate the median, interquartile range (IQR), 1.5*IQR and outliers of per-population heterozygosity levels. **b**, Heterozygosity of Oceanians against their estimated proportion of East Asian-related ancestry. **c**, Heterozygosity of East Asian-related Pacific populations, against their estimated proportion of Papuan-related ancestry. **b-c**, Each point represents an individual. Colours indicate the population group of origin. Individual admixture proportions were obtained from ADMIXTURE at K = 6 (Fig. 1c). Spearman's coefficient ρ and corresponding P-value are shown. The black line indicates the regression line of a linear model and the grey zone the 95% CI (n=95 and n=35 individuals for **b** and **c**, respectively).

AMOVA-based F_{ST}

Methods. F_{ST} values were estimated using the Analysis of Molecular Variance (AMOVA) method (Φ_{ST} in ref.[45]). Values were computed with a home-made perl script (available on www.github.com/h-e-g/evoceania).

Results. F_{ST} values indicated low genetic differentiation among Vanuatu islands and between Vanuatu and Bismarck archipelago populations (Supplementary Fig. 10), as shown by the PCA (Fig. 1d). The highest genetic differentiation was detected between PNG and Taiwanese indigenous peoples.



Supplementary Figure 10. AMOVA-based F_{ST} among Pacific populations. F_{ST} matrix for all possible pairs of populations. Colour bars by the population names indicate population affiliations according to Fig. 1. Colour scale indicates lower (F_{ST} < 0.01; in blue) and higher genetic differentiation (F_{ST} > 0.04; in red). RenBell indicates Polynesian outliers sampled from Rennell and Bellona Islands.

Supplementary Note 4: Demographic Inference

Parameter estimation

To infer the demographic history of populations from the Pacific, we used datasets filtered at *Level 3a*, and annotated for the ancestral state (Supplementary Note 2). Demographic parameter estimation was performed using the simulation-based framework[46] implemented in *fastsimcoal* version 2.6 (http://cmpg.unibe.ch/software/fastsimcoal2/). This method estimates the multinomial likelihood of the observed multidimensional Site Frequency Spectrum (SFS) O, given the expected SFS E approximated from coalescent simulations of a given model under specific parameter values θ, following refs.[46,47]. The multinomial likelihood is computed as:

$$L_{full} = P(O|\theta) \propto P_0^{L-S}(1-P_o)^S \prod_{i=1}^{n-1} \hat{e_i}^{o_i}$$

where O = {o_1, ..., o_{n-1}} are entries of the observed SFS, E = {e_1, ..., e_{n-1}} are entries of the expected SFS, P_0 is the probability that no mutation occurred on the expected mean coalescent tree, S is the total number of polymorphic sites and L is the total length of the surveyed sequence. *fastsimcoal2* starts with initial random parameter values sampled from a specified distribution and performs a series of expectation conditional maximization (ECM) optimization cycles. To avoid local maxima, the same demographic scenario is simulated several times with varying starting points of the algorithm (i.e., random seed and initial values). We performed 600,000 simulations, 65 ECM cycles, and 100 replicate runs (unless specified) starting from different random initial values. To limit overfitting, only SFS entries with more than 5 counts were considered for parameter estimations ('-C 5').

To maximize the fit between the expected and observed SFS, we used the approach described in refs.[16,48,49]. Specifically, the likelihood L_{full} was first computed and optimized using all entries of the SFS (i.e. both invariant and variant sites with entry counts > 5) for the first 25 cycles and then $L_{SFS} \propto \prod_{i=1}^{n-1} \hat{e_i}^{o_i}$ was optimized, using only variant sites (with entry counts > 5) for the remaining 40 cycles.

To obtain the maximum-likelihood (ML) estimates of demographic parameters for a given model, we first selected the 10 runs, among the 100 replicate runs, with the highest likelihood. To account for the stochasticity 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 refined the likelihood of the three runs with the highest average, re-estimated $\log_{10}$(likelihood) using 10^7 simulations, and considered the run with the highest likelihood as the ML run. To correct for the different number of SNPs in the expected and observed SFS, we rescaled the parameters by a rescaling factor (RF) defined as S_{obs}/S_{exp}: N_e and generation times were multiplied by RF, while migration rates were divided by RF.

For all time parameter estimates, we assumed a generation time of 29 years[50] and a constant mutation rate of 1.25×10^{-8} mutation/generation/site, i.e., the rate of *de novo* mutations estimated from deep sequencing of family pedigrees, and used in several recent population genomics studies[16,17,51]. We decided to use this mutation rate because we built a demographic model of Oceanian populations by adding newly-studied populations to the previous 'Out-of-Africa' model obtained by Malaspinas *et al.*[16], where a constant mutation rate of 1.25×10^{-8} mutation/generation/site was also assumed. We note that another study has estimated a higher mutation rate of $1.3–1.8\times10^{-8}$ mutation/generation/site, based on the comparison of high-coverage ancient and modern human genomes[52]. To account for uncertainty in mutation rate estimations, we also provide, for all divergence and admixture times, estimates assuming a mutation rate of 1.40×10^{-8} mutation/generation/site (Supplementary Tables 3-7).

Confidence intervals

We calculated confidence intervals with a non-parametric block bootstrap approach. We first generated 100 bootstrapped datasets by randomly sampling with replacement the same number of 1-Mb blocks of concatenated genomic regions as in the observed data. Then, for each bootstrapped dataset, we obtained multi-SFS with Arlequin version 3.5.2.2 (http://cmpg.unibe.ch/software/arlequin35/, ref.[53]), and re-estimated parameters using the same settings as for the original dataset, but with 20 replicate runs instead of 100. To obtain the 95% confidence intervals, we calculated the 2.5% and 97.5% percentiles of the estimate distribution obtained by non-parametric bootstrap.

Model selection

For model selection, because the likelihood function is a composite likelihood (due to the presence of linked SNPs in our datasets), we did not use classical model choice procedures such as the likelihood ratio tests or Akaike Information Criterion (AIC). Instead, we estimated the difference between models in the expected $\log_{10}$(likelihood) of the observed SFS, referred as *initial* likelihood, which is approximated from 600,000 simulations. Furthermore, we also re-estimated a hundred times the $\log_{10}$(likelihood L_{SFS}) of the observed SFS, from 100 expected SFS computed with 10^7 coalescent simulations, instead of 600,000 simulations, and using parameters that maximized the likelihood under each scenario (i.e., run with the highest likelihood). These likelihoods are referred as *re-estimated* likelihoods. Their distribution reflects the stochasticity inherent to the approximation of the likelihood using coalescent simulations. We considered that a model is the most likely if (i) the *initial* expected $\log_{10}$(likelihood) of the observed SFS under this model is higher than that of the alternative models, and (ii) the difference between the mean of the 100 *re-estimated* $\log_{10}$(likelihoods) of this model and that of other models (Δ maximum $\log_{10}$(likelihood); ΔML) is greater than 50 (see ref.[49]). Finally, for some of our model comparisons, we also estimated the probability that the true model is selected, using simulated SFS as observed SFS (see section "Refining the demographic history of Near Oceania"). The true positive rate of the model selection was computed as $TPR = \frac{\mathrm{n}_{\Delta\mathrm{ML} \geq +50}}{(\mathrm{n}_{\Delta\mathrm{ML} \geq +50} + \mathrm{n}_{\Delta\mathrm{ML} \leq -50})}$, where ΔML = Likelihood$_{\text{True model}}$ – Likelihood$_{\text{Alternative model}}$, $\mathrm{n}_{\Delta\mathrm{ML} \geq +50}$ is the number of pseudo-observed SFS for which the true model is favoured, and $\mathrm{n}_{\Delta\mathrm{ML} \leq -50}$ is the number of pseudo-observed SFS for which the alternative model is favoured.

Model fitting

To identify entries of the expected SFS that show a poor fit with the observed SFS, we compared all entries of the observed multidimensional SFS against simulated entries, averaged over 100 SFS expected under the most likely model, obtained with *fastsimcoal2*[46]. Entries with the worst fit were defined as those that exhibit a difference between the expected and observed SFS larger than 500 units (i.e., $|(m_i \log_{10}(p_i)) - (m_i \log_{10}(m_i/L))| > 500$, where m_i is the observed count at the i-th entry, p_i is the expected SFS at the i-th entry and L is the total number of polymorphic sites). In addition, we also compared observed *vs.* simulated F_{ST} for all pairs of populations, computed with vcftools version 0.1.13[54]. Specifically, we computed Weir and Cockerham's F_{ST} for each 1-Mb block of concatenated genomic regions in the observed data, and averaged values across blocks. In parallel, we simulated 1,000 times x 1-Mb DNA loci, x being the number of 1-Mb blocks in the observed data, using *fastsimcoal2* under the best-fitted model. We assumed a mutation rate of 1.25×10^{-8} mutation/generation/site[17,51] and a recombination rate obtained from the 1000 Genomes Phase 3 genetic map[43]. We then verified that the observed genomic average of F_{ST} was included in the distribution of 500 averages of x randomly sampled, simulated 1-Mb DNA loci. For the baseline model that includes archaic introgression (see section below), we also compared observed *vs.* simulated f_4-ratio statistics. Namely, we simulated with *fastsimcoal2* 500 independent sets of 10-Mb genomic regions and sampled simulated individuals from African, East Asian, PNG, Neanderthal and Denisovan populations so that

sample sizes were equal to those of the observed data. We then estimated Denisovan ancestry in the simulated PNG population using the following f_4-ratio statistic: f_4(*Africans, Neanderthal; East Asians, PNG*) / f_4(*Africans, Neanderthal; East Asians, Denisova*) (Supplementary Note 7), and verified that the observed f_4-ratio statistic was included in the distribution of simulated values.

Estimation accuracy

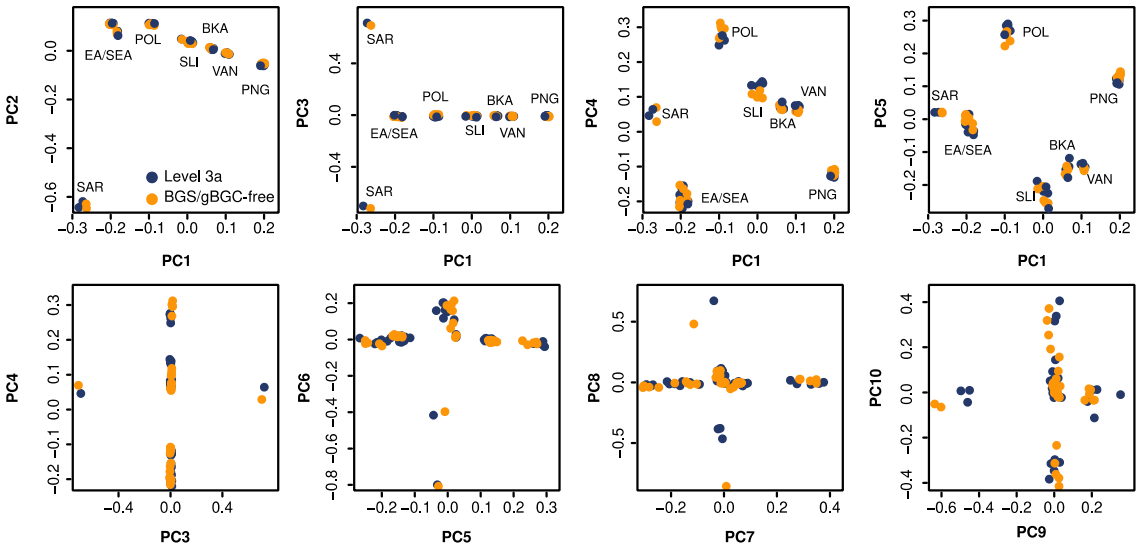
For the most complex models (see sections "Refining the demographic history of Near Oceania", "The demographic history of western Remote Oceania" and "Refining the sources of East Asian ancestry among Oceanians"), we evaluated the accuracy of the estimations of demographic parameters with *fastsimcoal2*, with a parametric bootstrap approach. Specifically, we simulated, with *fastsimcoal2*, *x* 1-Mb DNA loci, *x* being chosen to obtain the same number of segregating SNPs and number of invariant sites as in the observed data, assuming parameters that maximized the likelihood under each model. We assumed a mutation rate of 1.25×10^{-8} mutation/generation/site[17,51] and used a recombination rate obtained from the 1000 Genomes Phase 3 genetic map[43]. Twenty simulated SFS were then generated with Arlequin version 3.5.2.2 (http://cmpg.unibe.ch/software/arlequin35/)[53]. We re-estimated the parameters using each of the 20 simulated SFS and the same settings as for the original dataset (65 ECM cycles, 600,000 simulations and 100 runs per simulated SFS). Then, we calculated the mean, median and the 2.5% and 97.5% percentiles of the distribution of parameter estimates obtained by parametric bootstrap, and verified that they include the true (simulated) parameter value.

Background selection and GC-gene biased conversion

Rationale. Demographic inference assumes that the genome is mainly evolving under neutrality, but this assumption may be violated because of background selection (BGS; i.e., loss of neutral mutations linked with deleterious alleles due to negative selection) and GC-biased gene conversion (gBGC; i.e., increase in frequency of GC alleles due to recombination)[55]. To account for this, we excluded sites within CpG islands and genes (*Level 3a* filters) for demographic inference. However, linked selection might affect sites in intergenic regions, particularly in low-recombining regions. We thus compared the genetic structure of Pacific populations for different sets of variants filtered, or not, for low-recombining regions and high gBGC sites.

Methods. We compared the PCA of two datasets that include the same individuals (i.e., those used for the demographic inference, to the exclusion of archaic hominins) but include two different sets of SNPs. The first set is composed of SNPs that passed the *Level 3a* filters, whereas the second, referred to as *BGS/gBGC-free*, includes SNPs that passed the filters described in ref.[55]. Specifically, we kept (i) sites with no missingness, (ii) sites with a local recombination rate > 1.5 cM/Mb using 1000 Genomes Project Phase 3 genetic map[43], and (iii) sites with mutation types C↔G and A↔T (i.e., unbiased Weak ↔ Weak, Strong ↔ Strong alleles)[55]. As the first set presented ~17× more SNPs than the second set (3,800,502 vs 218,074 SNPs), we randomly selected 218,074 SNPs in the first set, for comparison purposes. The two PCA were then computed using 'SmartPCA' algorithm implemented in 'EIGENSOFT' version 6.1.4 (ref.[28]).

Results. Highly similar results were obtained by PCA using *Level 3a* and *BGS/gBGC-free* datasets (Supplementary Fig. 11). This suggests that *Level 3a* filters were sufficiently stringent to remove most sites influenced by BGS or gBGC, and that our demographic models of Pacific populations should not be strongly affected by these evolutionary forces.



Supplementary Figure 11. Impact of background selection (BGS) and GC-gene biased conversion (gBGC) on the genetic structure of Pacific populations. First ten PCs of a PCA of two different datasets, one including all the *Level 3a* SNPs used for demographic inference (in blue), the other including SNPs that were further filtered for high BGS and gBGC genomic regions (*BGS/gBGC-free* in orange). Both datasets include the 41 samples from 10 populations that were used for demographic inference. Papua New Guinea highlanders (PNG) are represented by Kundiawa and Bundi populations, the ni-Vanuatu (VAN) by Malakula islanders, the Bismarck islanders (BKA) by the Nakanai Bileki, Solomon islanders (SLI) by Vella Lavella islanders, Polynesian outliers (POL) by Bellona islanders, East/Southeast Asians (EA/SEA) by Han Chinese, Paiwan Taiwanese and Philippine Kankanaey and Europeans by Sardinians (SAR). Proportions of variance explained by PC1 (PC2) were 8.3% (6.5%) and 8.6% (6.5%) for *Level 3a* and *BGS/gBGC-free* datasets, respectively.

Baseline demographic model of human populations

Demographic modelling and hypotheses. To build a demographic model of Oceanian populations, we started by confirming the 'Out-of-Africa' model and re-estimating the parameters obtained by Malaspinas *et al.*[16]. This model includes modern populations from Africa, Europe, East Asia and Oceania in isolation with migration, as well as archaic hominins, to model archaic introgression. This baseline model served as a scaffold on which newly-studied populations were subsequently added. We chose to follow this rationale to limit the number of parameters to estimate, as we fixed several parameters of the baseline model (e.g., those related to demographic events that predate the settlement of Oceania) in the subsequent models.

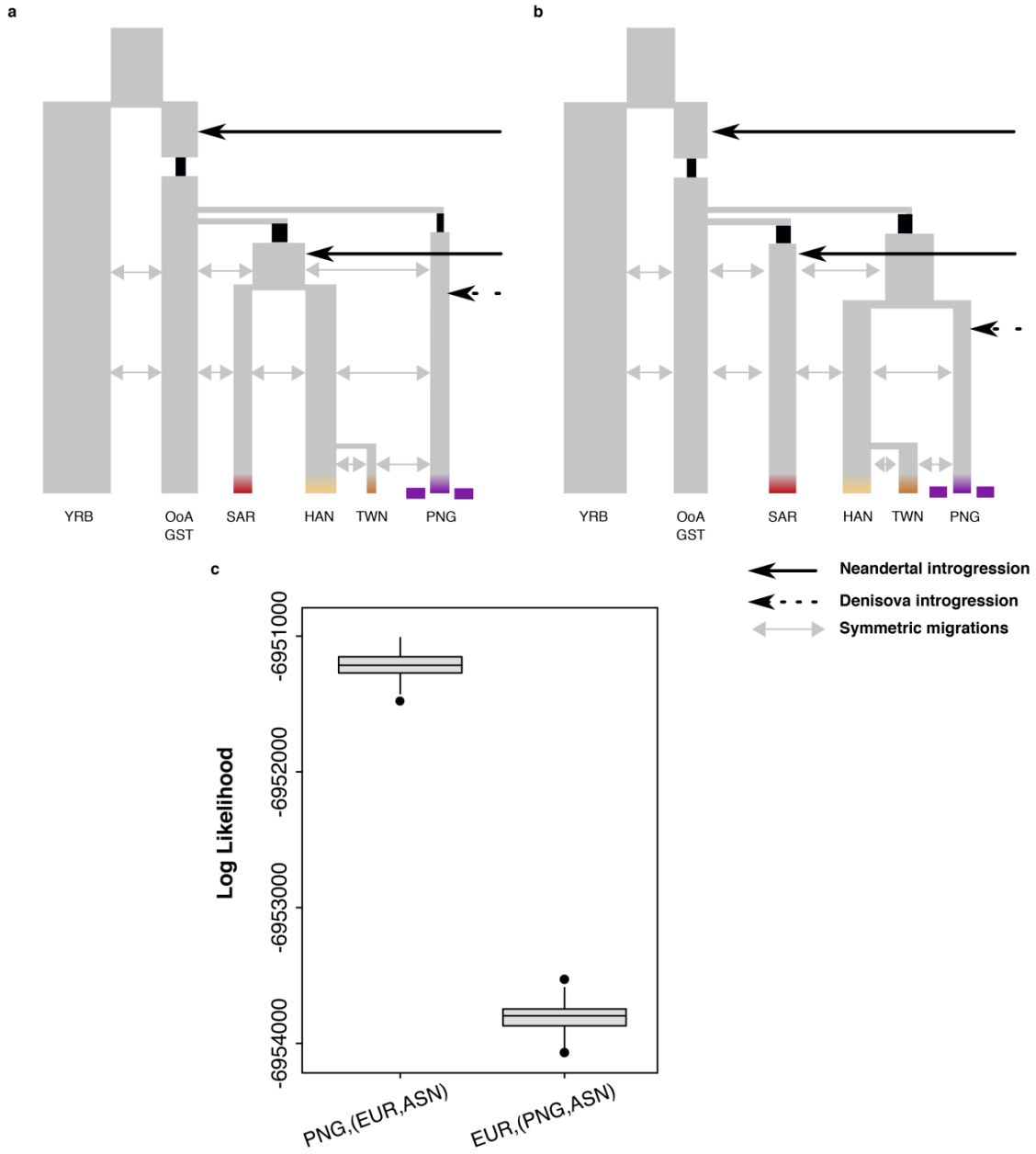
Our model differs in several aspects from those used by Malaspinas *et al.*[16]. We used the Vindija Neanderthal instead of the Altai Neanderthal, because the Vindija Neanderthal was shown to be more closely related to the Neanderthals who interbred with modern humans[56]. Near and Remote Oceanians are thought to descend from at least two parental populations that relate to present-day Papua New Guineans and Austronesian speakers from Taiwan[30-32]. Because our study focuses mainly on the history of Oceania, we replaced the indigenous Australians by PNG, and added to the model the Taiwanese indigenous peoples, to represent Austronesian speakers[31,32,38,57]. We assumed that the ancestors of Taiwanese indigenous peoples separated from the ancestors of mainland Han Chinese, in agreement with archaeological and genetic evidence[57,58]. Furthermore, to leverage the WGS data obtained for several related PNG populations, PNG were modelled following a continent-island model, where the continent represents a meta-population that sends migrants to islands, constituted by the two sampled populations (i.e., the Bundi and Kundiawa). We assumed that lineages first coalesced within islands during 100 generations until all remaining lineages are transferred to the continent[46].

To reduce the parameter space, all parameters relating to sub-Saharan Africans were fixed to previous ML estimates[16] (i.e., N_e, divergence times and migration rates), whereas all other parameters were re-estimated. The search ranges of divergence times of Denisovans and Neanderthals were set to the confidence intervals estimated in refs.[17,56,59]. Sampling time of the Altai Denisovan and Vindija Neanderthal were fixed to 2,800 and 2,000 generations, respectively[56]. We accounted for archaic introgression in non-African populations, by estimating the time and proportion of (i) Neanderthal introgression in the common ancestors of all non-African populations, (ii) Neanderthal introgression in the common ancestors of Eurasians, and (iii) Denisovan introgression in the ancestors of PNG. Following a hypothesis-free approach, we tested two tree topologies, to evaluate whether East Asians (Han and Taiwanese indigenous peoples) share more recent common ancestors with Europeans (PNG, (EUR, ASN)) or with Papuans (EUR, (PNG, ASN)) (Supplementary Fig. 12a,b). We note that we did not interpret the divergence time between Han Chinese and Taiwanese indigenous peoples in the baseline model, as this was the purpose of a more detailed model of populations contributing East Asian-related ancestry to Oceanians (see sections 'The sources of East Asian ancestry among Oceanians' and 'Refining the sources of East Asian ancestry among Oceanians'). All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

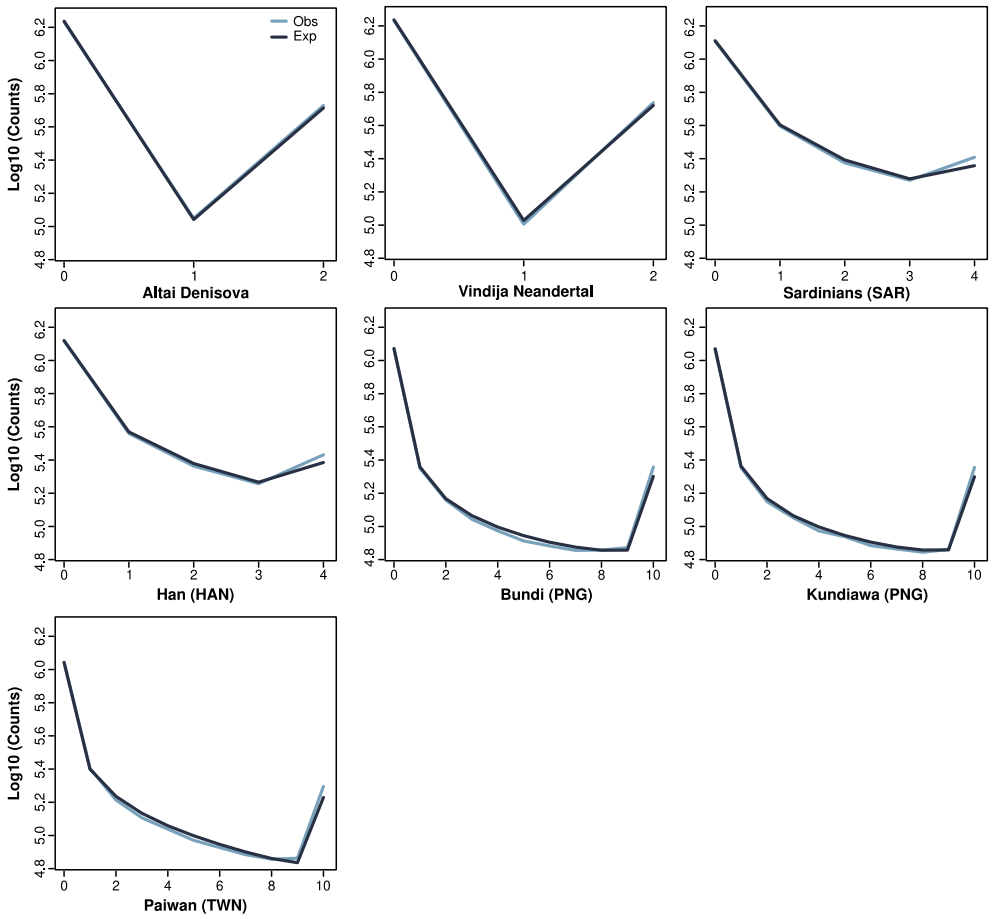
Dataset. We used 2 datasets with different populations for replication. Dataset 1 includes 2 SGDP Sardinians (SAR)[17], 2 SGDP Han Chinese (HAN)[17], 5 Paiwan (Taiwanese indigenous peoples, TWN), and 5 Bundi and 5 Kundiawa from Papua New Guinea (PNG)[16]. Dataset 2 includes the same populations, except that the 5 Paiwan were replaced by 2 Philippine Kankanaey (PHP)[17], representing an Austronesian-speaking community from the Philippines. The two datasets were merged with the two high-coverage genomes of Vindija Neanderthal[56] and Altai Denisovan[60], filtered at *Level 3a* and annotated for the ancestral state (Supplementary Note 2). Datasets were then decomposed into blocks of 1-Mb concatenated genomic regions, and multi-SFS were generated with Arlequin version 3.5.2.2 (ref.[53]).

Results. We found that the (PNG, (EUR, ASN)) model, where East Asians (ASN) share more recent common ancestors with Europeans (EUR) than with Papuans (PNG), was significantly more likely than the alternative (EUR, (PNG, ASN)) model (ΔML > 1,000 $\log_{10}$ units for both *initial* and *re-estimated* likelihoods, see section 'Model selection'), confirming previous results[16] (Supplementary Fig. 12c). Under the most likely model, we estimated a strong Out-of-Africa bottleneck in the ancestral population of all non-Africans (N_e = 411, 95% CI: 364–7,950; intensity = 24%, 95% CI: 1%–27%). We found a substantial population reduction associated with the peopling of Eurasia (N_e = 1,822, 95% CI: 395–2,174; intensity = 5.5%, 95% CI: 4.6%–25%) and in PNG (N_e = 247, 95% CI: 140–285; intensity = 40%, 95% CI: 35%–71%). Neanderthal introgression in the ancestral population of non-Africans was estimated to occur 61 ka (95% CI: 56–62 ka) with a rate of ~2% (95% CI: 1.5%–2.7%) (Extended Data Fig. 2a; Supplementary Table 2). Neanderthal introgression in the ancestral Eurasian population was estimated to occur ~52 ka (95% CI: 47–54 ka) with a rate of ~0.36% (95% CI: 0.36%–1.86%). Finally, Denisovan introgression into the ancestral population of PNG occurred ~42 ka (95% CI: 35–44 ka) with a rate of ~3.6% (95% CI: 3.2%–4.1%) (Extended Data Fig. 2a; Supplementary Table 2). We estimated a divergence time between ancestors of Eurasians and PNG ~57 ka (95% CI: 53–60 ka). Remarkably, and despite the differences of our model to that of Malaspinas *et al.*[16], most of our point estimates fell within the CIs previously reported. Furthermore, point estimates of demographic parameters were similar when using, instead of Taiwanese indigenous peoples (TWN), the Philippine Kankanaey (PHP) to represent Austronesian speakers (Supplementary Table 2). Altogether, our baseline model confirms previous findings, and recapitulates important aspects of the demographic history of populations involved in the settlement of Near and Remote Oceania, i.e., East Asians and PNG.

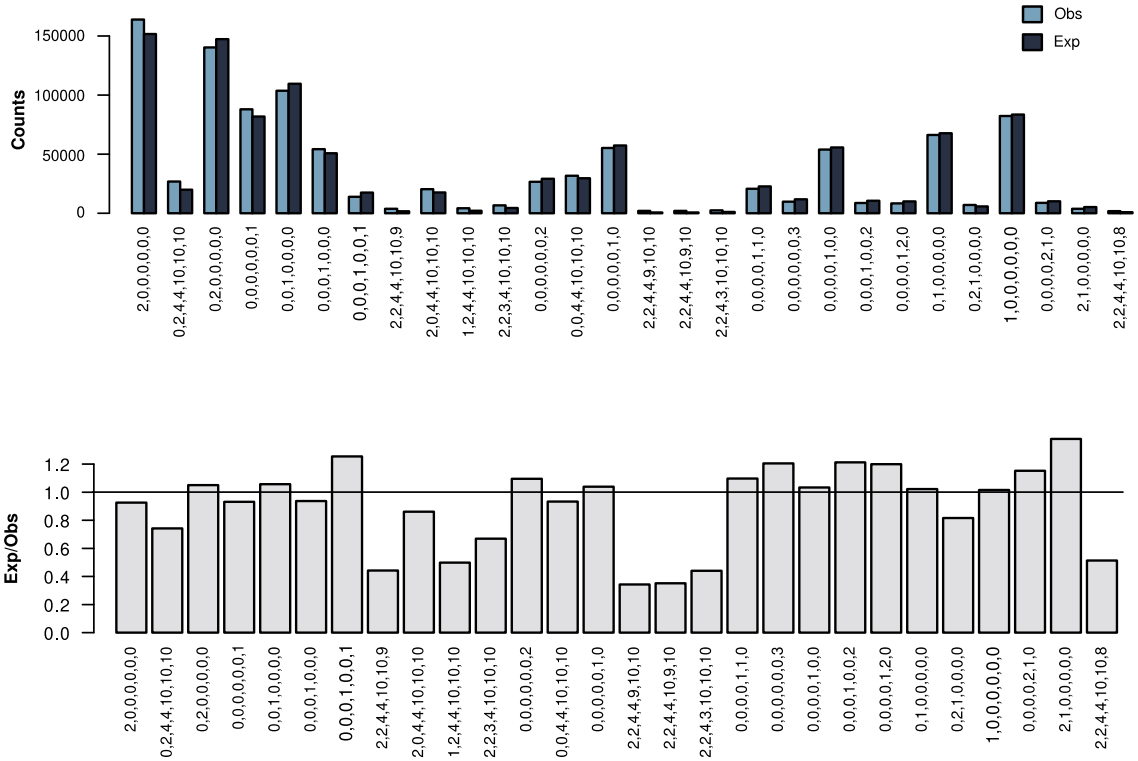
Model fitting. Overall, we obtained a very good fit of expected and observed marginal SFS, indicating that the model and parameter estimates well reproduce the data (Supplementary Fig. 13). The entries of the SFS with the poorest fit were those where the derived allele is fixed in archaic and most modern human populations (Supplementary Fig. 14). This is most probably due to ancestral state misspecification, which is not expected to affect the parameters that we aim to estimate, such as divergence times among modern human populations. Other entries with a relatively poor fit were those where the derived allele segregates, or is fixed, in archaic hominins but is absent from modern humans, probably because some parameters were constrained to previously estimated values (i.e., divergence times related to Denisovans and Neanderthals[17,56,59]). We also tested if simulated data under the best-fitted model well reproduce the observed data for summary statistics related to archaic introgression. Namely, we compared the observed f_4-ratio statistic for Denisovan introgression (Supplementary Note 7) to that estimated from simulations with *fastsimcoal2*[46] under the best-fitted model. Simulated statistics were very close (mean f_4-ratio = 0.029; median f_4-ratio = 0.024; IQR = 0.042) to the observed value (f_4-ratio = 0.032) in PNG, confirming the accuracy of our baseline model concerning Denisovan introgression. Finally, we checked that genetic differentiation among modern human populations, measured by Weir and Cockerham's F_{ST}, was well reproduced by the best-fitted model. We observed a very good fit between observed and expected F_{ST} (Supplementary Fig. 15), validating further our baseline model.



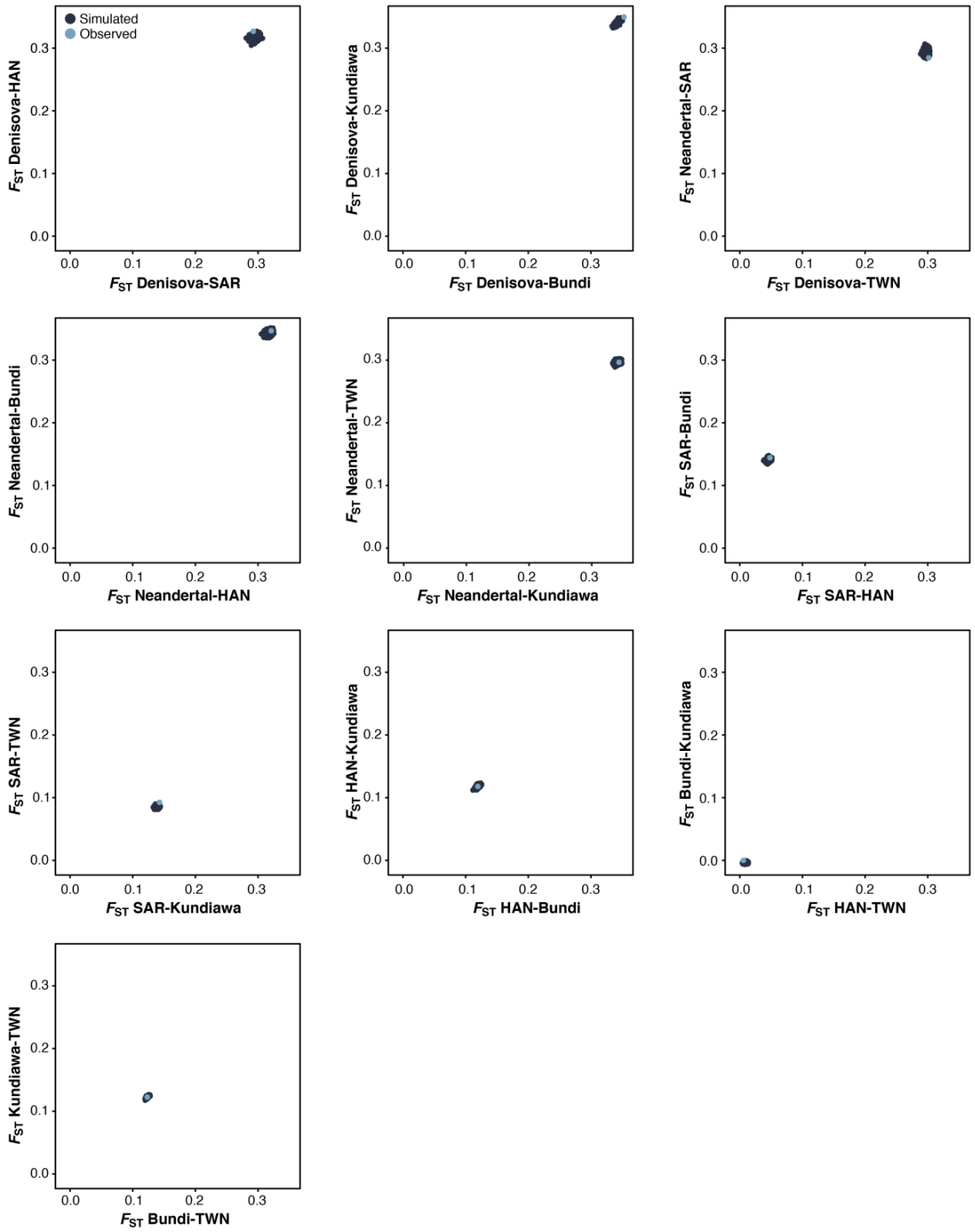
Supplementary Figure 12. Alternative topologies for the baseline model. **a**, The (PNG, (Europe, East Asia)) model assumes that East Asians share a more recent common ancestor with Europeans. **b**, The (Europe, (PNG, East Asia)) model assumes that East Asians share a more recent common ancestor with Papuans. **a**,**b**, OOA GST indicates the unsampled African population who left Africa, SAR indicates Sardinians[17], HAN indicates Han Chinese[17], TWN indicates Taiwanese indigenous peoples, and PNG indicates Papua New Guinean highlanders (Bundi and Kundiawa[16]). In both models, the ancestors of the two groups separate independently from Africans, as in ref.[16]. For convenience, only modern human populations are represented. Grey arrows indicate symmetric migrations between modern humans. Solid black arrows represent Neanderthal introgression into the common ancestors of all non-African populations and **a**, Eurasians or **b,** Europeans. The dashed black arrow indicates Denisovan introgression into the ancestral population of PNG. Bottlenecks are indicated by black rectangles. **c**, Likelihood distribution of the two alternative topologies in **a** and **b**. The line, box, whiskers and points respectively indicate the median, interquartile range (IQR), 1.5*IQR and outliers of the *re-estimated* likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario.



Supplementary Figure 13. Fitting of the SFS for the baseline model. The marginal one-dimensional SFS of the observed data (in blue) is compared to the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 simulations, using parameters that best fit the data under the (PNG, (Europe, East Asia)) model.



Supplementary Figure 14. SFS entries with the worst fit for the baseline model. Differences in the number of counts between the observed (blue) and expected (black) SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of the number of counts for the i^{th} entry in the expected and observed SFS. Entries are indicated by columns and correspond to the counts of the derived allele in Denisova ($2n$ = 2), Vindija Neanderthal ($2n$ = 2), Sardinians (SAR, $2n$ = 4), Han (HAN, $2n$ = 4), Bundi (PNG, $2n$ = 10), Kundiawa (PNG, $2n$ = 10) and Paiwan (TWN, $2n$ = 10) (from bottom to top).



Supplementary Figure 15. Observed versus simulated F_{ST} for each pair of populations used in the baseline model. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under the best parameters inferred for the baseline model, and were compared with observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.

The demographic history of Near Oceania

Demographic modelling and hypotheses. A Late Pleistocene occupation by modern humans has been documented in Near Oceania (New Guinea, the Bismarck Archipelago and the Solomon Islands)[61,62], but the early demographic history of Near Oceanians remains largely unknown. Archaeological evidence supports the existence of a Holocene expansion from East/Southeast Asia associated with the peopling of Remote Oceania. This recent expansion is thought to be at the origin of the Lapita cultural complex and the spread of Austronesian languages in Near and Remote Oceania[2,62]. This hypothesis is supported by previous genetic studies[30-32,36,38] and our analyses (Supplementary Note 3), indicating that Oceanians descend from two ancestral populations related to present-day PNG and East/Southeast Asians.

To gain insight into the peopling history of Near Oceania, we sought to model, in addition to baseline populations (see section 'Baseline demographic model of human populations'), two representative populations from Near Oceania, i.e., populations from the Bismarck Archipelago (BKA) and Solomon Islands (SLI). Following a hypothesis-free approach, we tested different topologies: (i) PNG diverged first, followed by the separation of islanders from the two other archipelagos (PNG,(BKA,SLI)), (ii) Solomon islanders diverged first, followed by the Bismarck Archipelago islanders and PNG (SLI,(PNG,BKA)), and (iii) Bismarck Archipelago islanders diverged first, followed by Solomon islanders and PNG (BKA,(PNG,SLI)) (Supplementary Fig. 16a). To account for admixture with populations of East Asian origin during the Holocene (i.e., attributed to the Austronesian expansion[2,62]), we modelled pulses of gene flow from Taiwanese indigenous peoples (TWN) to Bismarck and Solomon islanders (Supplementary Note 3), as Taiwanese indigenous peoples are considered a good proxy of Austronesian-speaking peoples entering Oceania[38]. Finally, archaeological studies suggest an extensive exchange network, notably between Papua New Guinea and the Bismarck Archipelago from 20 ka[63,64]. We therefore considered gene flow within Near Oceania by simulating asymmetrical migration following a stepping-stone model (i.e., between PNG and BKA, as well as between the BKA and SLI).

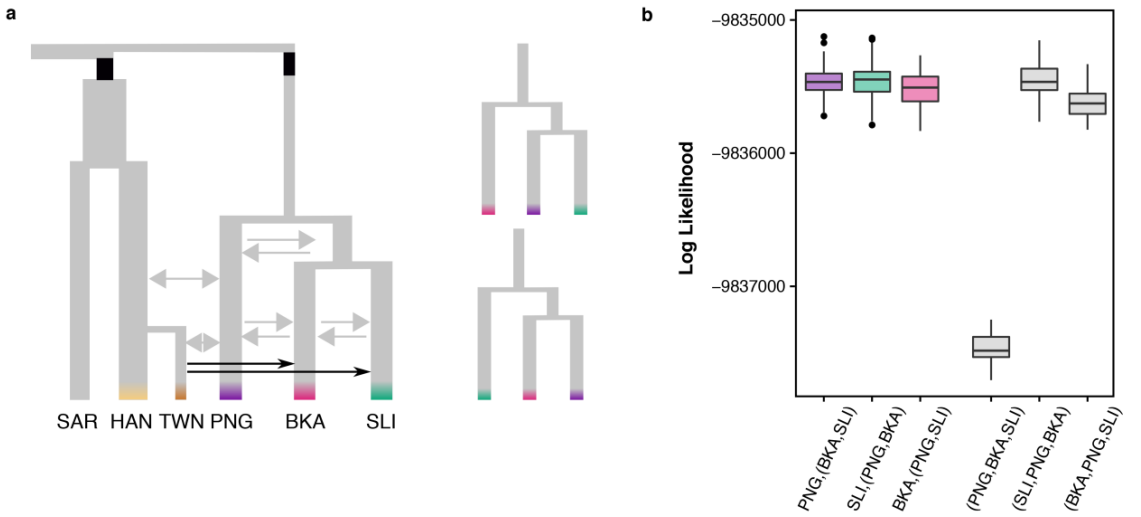
All parameters related to Eurasians and archaic hominins, together with parameters for events that predate the divergence between Eurasians and Near Oceanians, were fixed to the values obtained in the best-fitted baseline model (Supplementary Table 2). The rate of Denisovan introgression into the ancestral population of Near Oceanians was also fixed. To obtain parameter estimates for each demographic scenario, we selected the run, among 150 runs, that yielded the highest likelihood. All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

Dataset. We modified our baseline model by adding 5 Nakanai Bileki (Bismarck Archipelago; BKA) and 5 individuals from Vella Lavella, or Malaita for replication (Solomon Islands; SLI). To decrease the dimensionality of the multi-SFS, we excluded, from the SFS data, the 2 Sardinians (SAR), the 5 Bundi (keeping 5 Kundiawa samples to represent PNG) and the two archaic genomes, although the corresponding populations were simulated in the model by fixing their demographic parameters to the values obtained in the best-fitted baseline model. The multi-SFS was generated as for the baseline model.

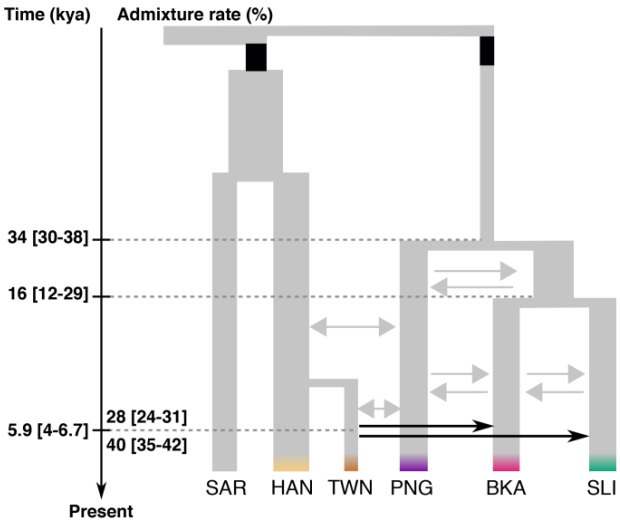
Results. The distributions of *re-estimated* likelihoods of the three models were largely overlapping, indicating no marked differences between the three tested topologies (Supplementary Fig. 16b). On average, a difference of 2 and 52 $\log_{10}$ likelihood units was observed between the (PNG,(BKA,SLI)) and (SLI,(PNG,BKA)) models, and between (PNG,(BKA,SLI)) and (BKA,(PNG,SLI)), respectively. Of note, we found that the set of parameters maximizing the likelihood under the (SLI,(PNG,BKA)) topology was compatible with a divergence of the three groups at the same time; the likelihood distribution of the (SLI,(PNG,BKA)) and (SLI,PNG,BKA) models were similar ($\Delta ML = 7$; Supplementary Fig. 16b). These results suggest that either PNG diverged first or that the three groups diverged simultaneously.

Based on the (PNG,(BKA,SLI)) model, which shows the highest average likelihood, we estimated a divergence between PNG and the ancestral population of the Bismarck

Archipelago (BKA) and Solomon Islands (SLI) ~34 ka (95% CI: 29.9–37.8 ka), followed by the divergence of the populations from the two archipelagos ~16 ka (95% CI: 12.3–29.0 ka, Supplementary Fig. 17 and Supplementary Table 3). We estimated that admixture from Taiwanese indigenous peoples (TWN) into the two archipelagos occurred ~6 ka (95% CI: 4.0–6.7 ka) and contributed ~28% (95% CI: 23.6%–31%) to Bismarck Archipelago islanders (BKA), and ~40% (95% CI: 34.7%–42.1%) to Solomon islanders (SLI), in agreement with the higher East Asian admixture proportion estimated for the latter (Fig 1c and Extended Data Fig. 1). Finally, strong migration ($2Nm > 1$) was observed between PNG and the Bismarck archipelago islanders ($2Nm_{PNG > BKA} = 2.20$, 95% CI: 1.88–4.21; $2Nm_{BKA > PNG} = 1.00$, 95% CI: 0.0004–1.11) as well as between Bismarck Archipelago and Solomon islanders ($2Nm_{SLI > BKA} = 2.58$, 95% CI: 0.14–5.05) (Supplementary Table 3). This suggests substantial gene flow between Near Oceanians, in agreement with archaeological data suggesting extensive exchange networks in the region starting 20 ka[63,64]. Importantly, similar estimates of demographic parameters were obtained when using population samples from Malaita, instead of Vella Lavella, to represent the Solomon Islands (SLI; Supplementary Table 3).

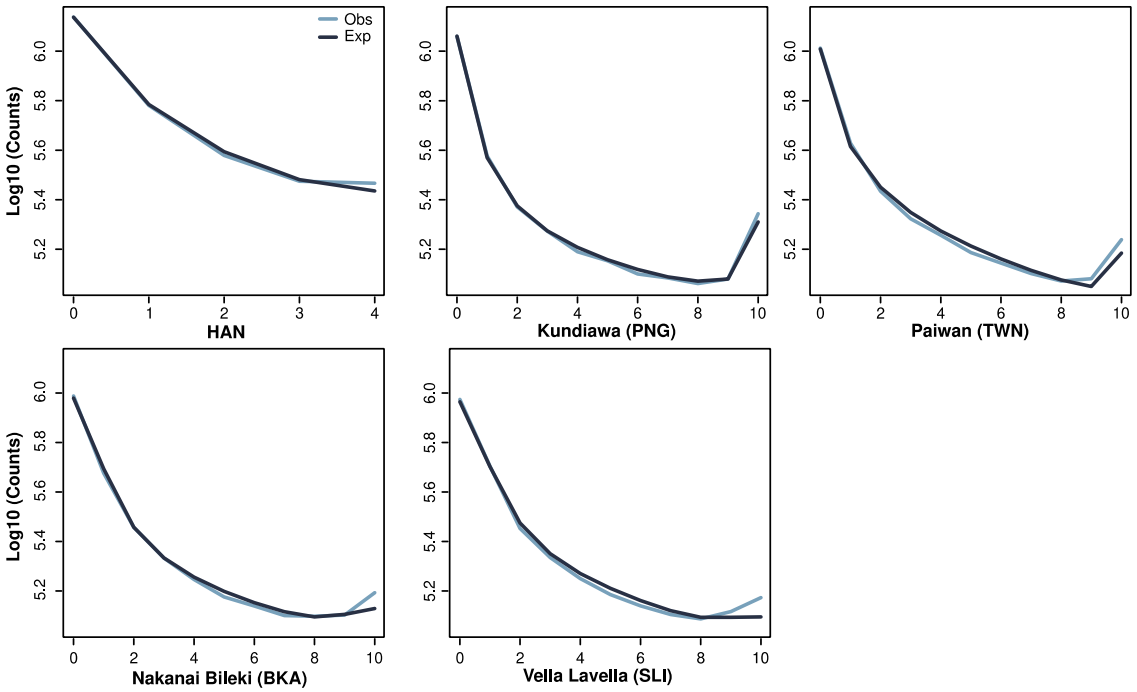


Supplementary Figure 16. Alternative topologies for Near Oceanians. **a**, Schematic representation of the three topologies tested. Model to the left corresponds to (PNG,(BKA,SLI)) and models to the right give a simplification of the two other topologies, (BKA,(PNG,SLI)) (top right) and (SLI,(PNG,BKA)) (bottom right). For the sake of clarity, only the populations from Eurasia and Near Oceania are shown. Grey arrows indicate migrations estimated in these models (one arrow for symmetric and two arrows for asymmetric gene flow). Black arrows indicate a single-pulse gene flow from Taiwanese indigenous peoples into the Bismarck Archipelago and the Solomon Islands (modelling the Austronesian expansion to Near Oceania). Bottlenecks are indicated by black rectangles. SAR indicates Sardinians, HAN indicates Han Chinese, TWN indicates Taiwanese indigenous peoples, BKA indicates Bismarck islanders, SLI indicates Solomon islanders, and PNG indicates Papua New Guinean highlanders. **b**, Likelihood distribution of the three topologies tested (left) and corresponding nested models where the three groups diverged simultaneously (right). The line, box, whiskers and points respectively indicate the median, IQR range, 1.5*IQR and outliers of the likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario. For the nested models, we used the same set of parameters as for the corresponding topology, except that we set the latest split among Near Oceanians at one generation apart from the oldest split.

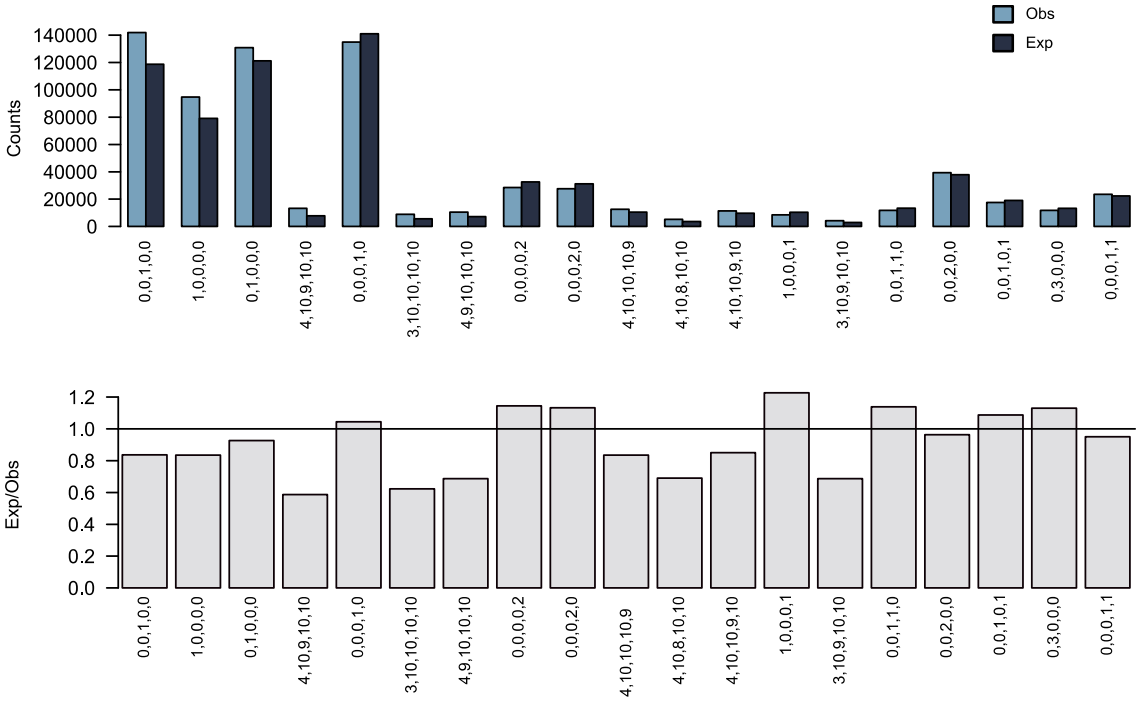


Supplementary Figure 17. Best-fitted model for Near Oceanians. SAR indicates Sardinians, TWN indicates Taiwanese indigenous peoples, BKA indicates Bismarck islanders, SLI indicates Solomon islanders, and PNG indicates Papua New Guinean highlanders. For the sake of clarity, only the populations from Eurasia and Near Oceania are shown. Grey arrows indicate migrations estimated in these models (one arrow for symmetrical and two arrows for asymmetrical gene flow). Black arrows indicate gene flow pulses from Taiwanese indigenous peoples into the Bismarck and the Solomon islanders (modelling the Austronesian expansion to Near Oceania). Bottlenecks are indicated by black rectangles. Estimated times are given in ka using a generation time of 29 years. Admixture proportions are given in %. 95% CIs are given in square brackets. The larger the rectangle width, the larger the effective population size (N_e). Bottlenecks are indicated by black rectangles. Point estimates of parameters and corresponding 95% CIs are given in Supplementary Table 3.

Model fitting. We obtained a good fit of expected and observed marginal SFS (Supplementary Fig. 18). The worst entries were those for high-frequency derived alleles, particularly in Near Oceanians. The entries of the joint SFS with the poorest fit were also those where the derived allele is fixed in most modern human samples (Supplementary Fig. 19). As for the baseline model, this is probably due to ancestral state misspecification. We observed a very good fit between observed and expected F_{ST} values (Supplementary Fig. 20), indicating that the model and parameter estimates well reproduce this aspect of the data.

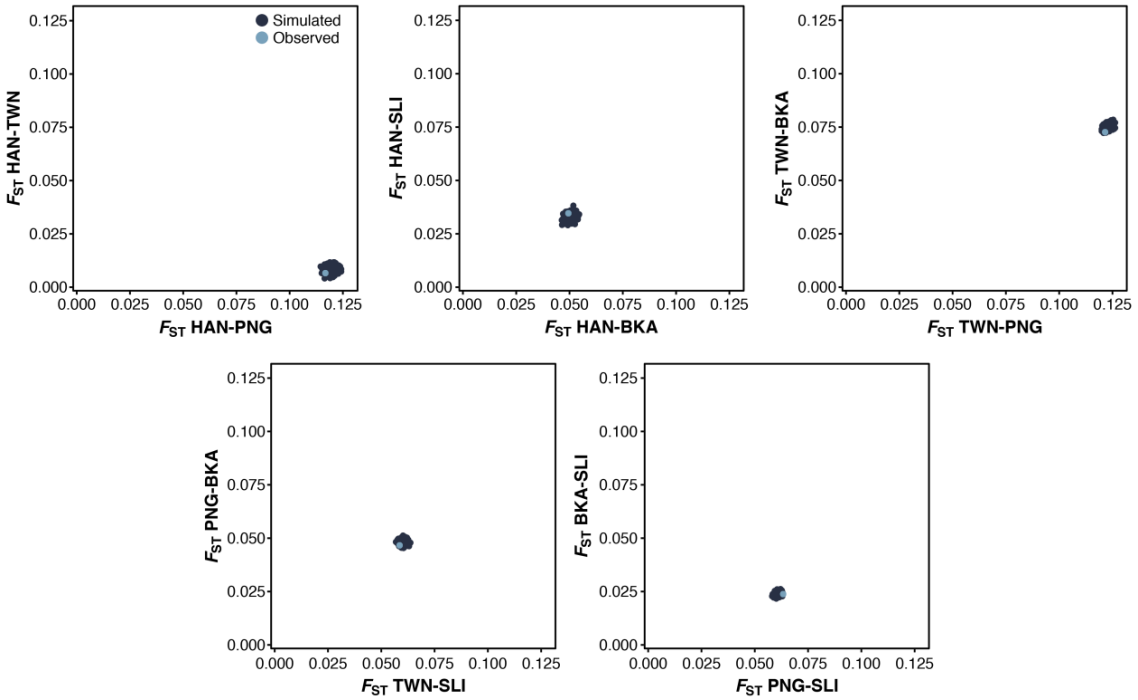


Supplementary Figure 18. Fitting of the SFS for the best-fitted model for Near Oceanians. The marginal one-dimensional SFS of the observed data (in blue) is compared to the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 simulations using parameters that best fit the data under the (PNG,(BKA,SLI)) model.



Supplementary Figure 19. SFS entries with worst fit for the best-fitted model for Near Oceanians. Differences in the number of counts between the observed (in blue) and expected (in black) SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele

in Han Chinese (HAN, 2*n* = 4), Kundiawa (PNG, 2*n* = 10), Paiwan (TWN, 2*n* = 10), Nakanai Bileki (BKA, 2*n* = 10) and Vella Lavella (SLI, 2*n* = 10) (from bottom to top).



Supplementary Figure 20. Observed versus simulated F_{ST} for each pair of populations used for the model for Near Oceanians. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under the best parameters inferred for the Near Oceanian model and were compared with F_{ST} obtained from the empirical data used for parameter inference (light blue).

Refining the demographic history of Near Oceania

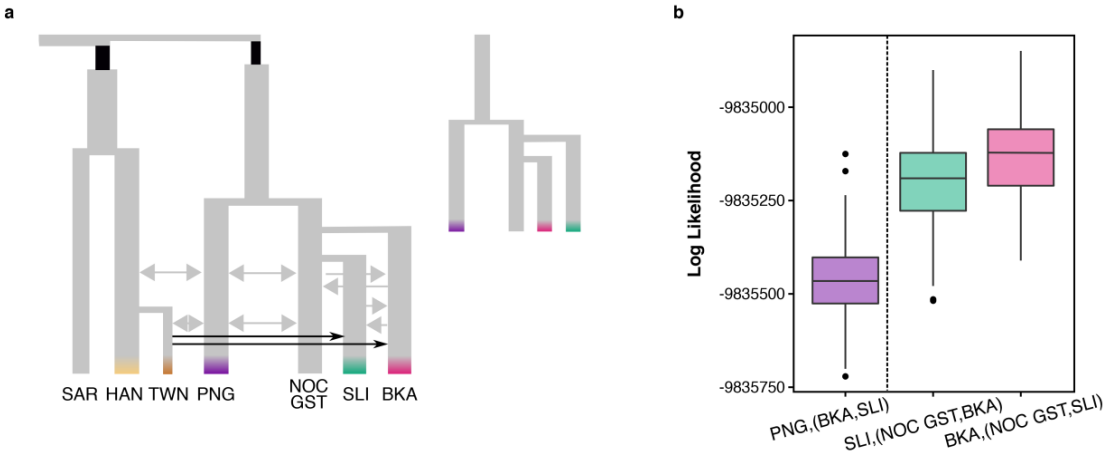
Demographic modelling and hypotheses. Several studies have documented genetic structure among populations from Near Oceania[30,65]. To account for such population structure, we modified our best-fitted model inferred in the section 'The demographic history of Near Oceania' (Supplementary Fig. 17) to include a ghost population representing a Near Oceanian meta-population. We thus simulated a ghost, unsampled population representing a Near Oceanian meta-population (NOC Ghost) and assumed that PNG and populations from the Bismarck Archipelago (BKA) and the Solomon Islands (SLI) diverged from it. Following a hypothesis-free approach, we tested two alternative models: (i) Bismarck islanders diverged from the NOC Ghost before Solomon islanders (BKA,(NOC GST, SLI)) or (ii) Solomon islanders diverged before Bismarck islanders (SLI,(NOC GST, BKA)) (Supplementary Fig. 21a). To allow comparison between models without (Supplementary Fig. 17) or with (Supplementary Fig. 21a) a NOC Ghost, we modified the parameters in the latter model so that the total number of parameters was the same for both models. The dataset used was the same multi-SFS as in the model for Near Oceanians (see section 'The demographic history of Near Oceania'). All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

Results. We found that the model that best fitted the data was the (BKA,(NOC GST, SLI)) model, where Bismarck islanders (BKA) diverged from the NOC Ghost before Solomon islanders (SLI) (ΔML ≥ 65 $\log_{10}$ units for both *initial* and *re-estimated* likelihoods, Supplementary Fig. 21b). We estimated the true positive rate of our model choice procedure (ΔML ≥ 50 $\log_{10}$ units, see section 'Model selection'), by using simulated SFS as pseudo-

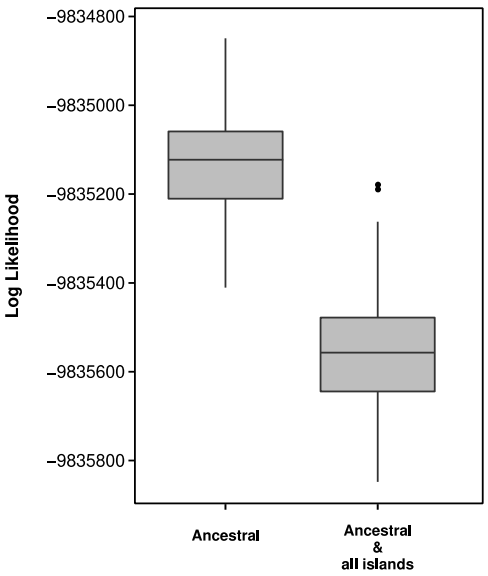
observed SFS. Namely, we obtained SFS under the (BKA,(NOC GST, SLI)) model by simulations, and estimated the likelihood of the pseudo-observed SFS under the (BKA,(NOC GST, SLI)) model, or the (SLI,(NOC GST, BKA)) alternative model. Out of 40 pseudo-observed SFS, the true model has $\Delta ML \geq 50$ $\log_{10}$ units in 81% of the cases, suggesting a reasonable true positive rate.

Under the most supported (BKA,(NOC GST, SLI)) model, we found that ancestors of Near Oceanians experienced a strong population reduction, which is >5x stronger than in Eurasians (N_e = 214, 95% CI: 186–276; intensity = 47%, 95% CI: 36–54, Fig. 2a, Supplementary Table 4). We checked whether this bottleneck signal is better explained by individual bottlenecks in each Near Oceanian population, by estimating the likelihood of an alternative model where each Near Oceanian population independently experiences a bottleneck, whose duration is fixed to 100 generations after each split. We found that the model without population-specific bottlenecks was the most likely (ΔML = 428 $\log_{10}$ units, Supplementary Fig. 22), further supporting the occurrence of a strong bottleneck in the ancestors of all Near Oceanians before the settlement of Oceania. We estimated a divergence between PNG and the NOC Ghost at ~40 ka (95% CI: 34–45 ka; Fig. 2a, Supplementary Table 4), between the NOC Ghost and Bismarck islanders (BKA) at ~25 ka (95% CI: 20–36 ka), and between the NOC Ghost and Solomon islanders (SLI) at ~20 ka (95% CI: 15.8–29.8 ka). We dated admixture between Taiwanese indigenous peoples (TWN) and the populations from the two archipelagos at ~4 ka (95% CI: 3.2–5.5 ka), with a contribution of ~43% (95% CI: 27%–58%) to Bismarck Archipelago islanders (BKA), and ~35% (95% CI: 31.7%–38.7%) to Solomon islanders (SLI). Comparable estimates of demographic parameters were obtained when using samples from Malaita instead of Vella Lavella, to represent the Solomon Islands (Supplementary Table 4), except for the divergence of Solomon islanders from the NOC Ghost and the gene flow pulse rates. This is in agreement with the suggested differences in the peopling history of the eastern, relative to western, Solomon Islands[29,30]. Furthermore, we evaluated the accuracy of our parameter estimation by parametric bootstrap, and found that the mean and median of the parameter estimates are very close to the true values and are all included in the 95%CIs (Supplementary Table 8 'Near Oceania'), except for one migration rate parameter. Together, these results indicate that the settlement of Near Oceania was rapidly followed by genetic isolation among archipelagos, and suggest that populations from the Solomon Islands diverged more recently or, at least, at the same time than those from the Bismarck Archipelago.

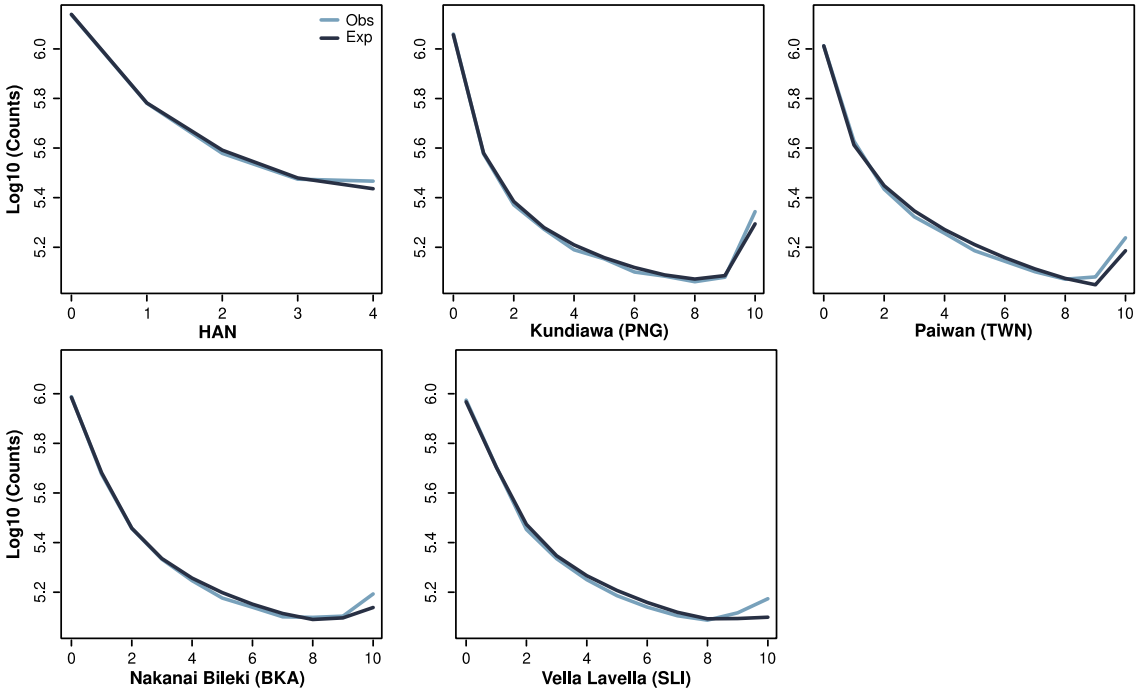
Model fitting. We obtained a better fit of expected and observed marginal SFS in this refined model, compared to the model without the NOC Ghost (Supplementary Figs. 21b and 23). The worst entries were again those for high-frequency derived alleles. The entries of the joint SFS with the poorest fit were also those where the derived allele is fixed in most modern human samples (Supplementary Fig. 24). As for the baseline model, this is probably due to ancestral state misspecification. We observed a very good fit between observed and expected F_{ST} (Supplementary Fig. 25), indicating that the model and parameter estimates well reproduce this aspect of the data.



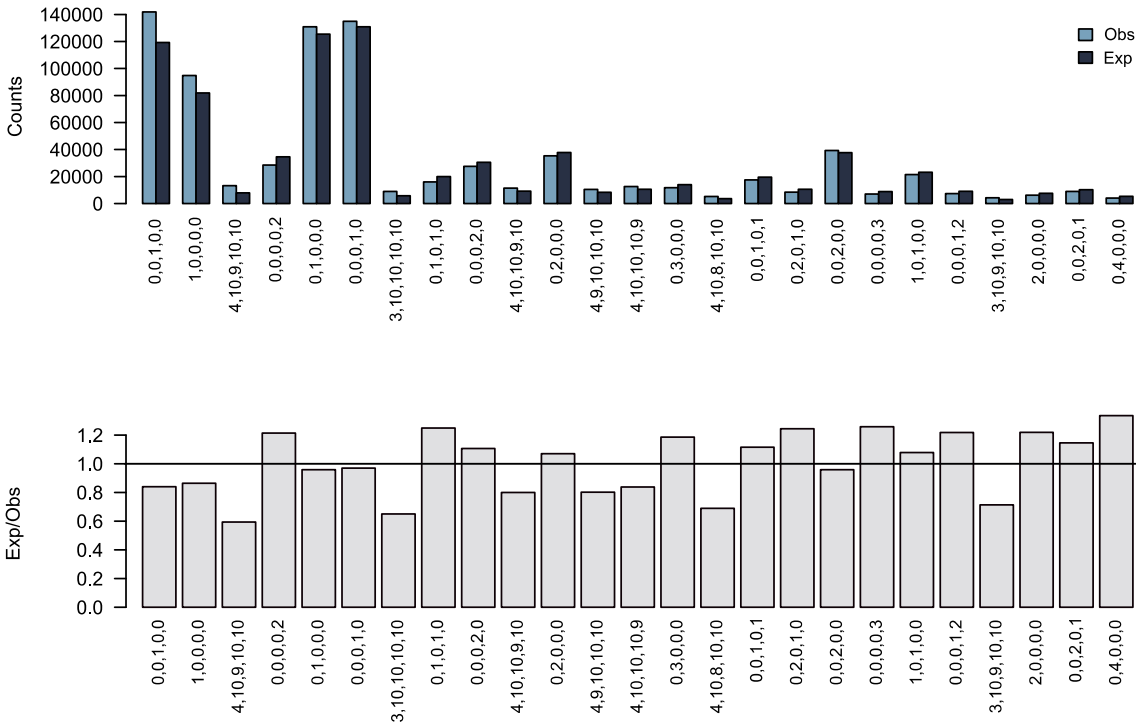
Supplementary Figure 21. Alternative refined models for Near Oceanians**. a**, Schematic representation of the two alternative refined models tested. SAR indicates Sardinians, HAN indicates Han Chinese, TWN indicates Taiwanese indigenous peoples, BKA indicates Bismarck islanders, SLI indicates Solomon islanders, PNG indicates Papua New Guinean highlanders and NOC GST indicates an unsampled population from Near Oceania. The model to the left corresponds to the topology where the Bismarck Archipelago diverged from the NOC Ghost before the Solomon Islands (BKA,(NOC GST, SLI)), while the smaller model to the right presents a simplification of the model where the Solomon islands diverged before the Bismarck Archipelago (SLI,(NOC GST,BKA)). For the sake of clarity, only the topologies for Eurasia and Near Oceania regions are shown. Grey arrows indicate migrations estimated in these models (one arrow for symmetric and two arrows for asymmetric gene flow). Black arrows indicate single pulse gene flow from Taiwanese indigenous peoples into the Bismarck Archipelago and the Solomon Islands (modelling Austronesian expansions to Near Oceania). Bottlenecks are indicated by black rectangles. **b**, Likelihood distribution of the alternative models. The line, box, whiskers and points respectively indicate the median, IQR range, 1.5*IQR and outliers of the likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario. The (PNG,(BKA,SLI)) model does not include a "NOC Ghost" (Supplementary Figs. 16-20, Supplementary Table 3), the (SLI,(NOC GST, BKA)) model is that where Solomon islanders diverged from the NOC Ghost before Bismarck islanders, and the (BKA,(NOC GST,SLI)) model is that where the Bismarck islanders diverged from the NOC Ghost before Solomon islanders.



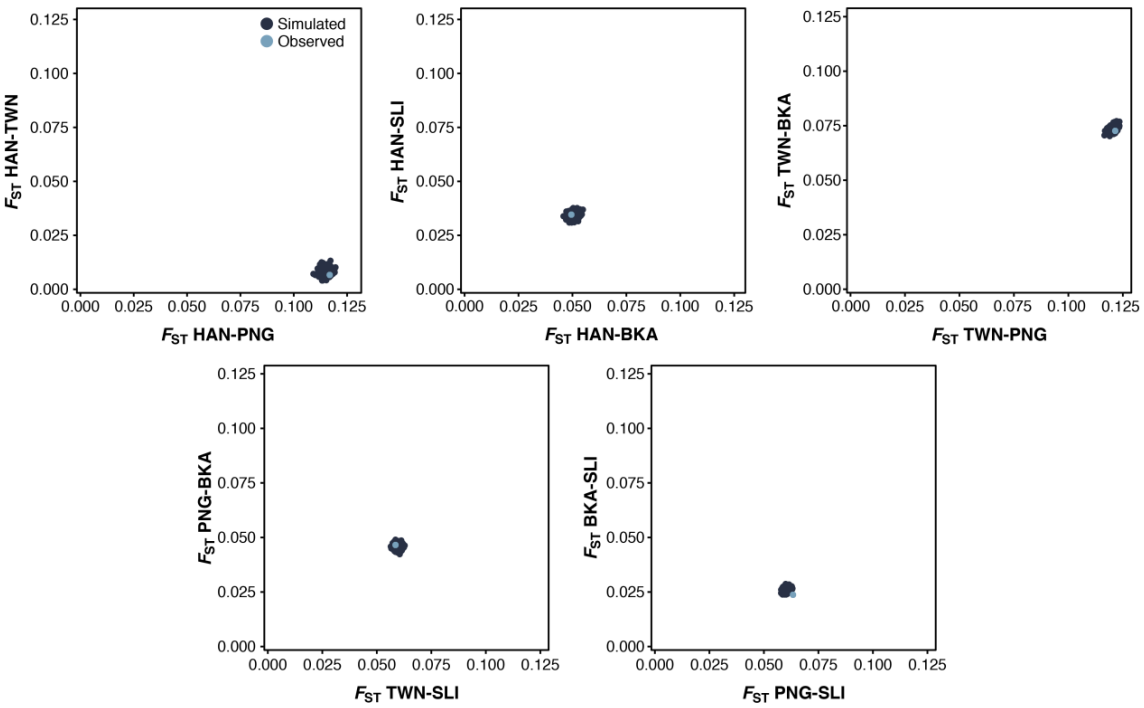
Supplementary Figure 22. Likelihood of refined models for Near Oceanians with or without population-specific bottlenecks. The line, box, whiskers and points respectively indicate the median, IQR range, 1.5*IQR and outliers of the likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario. On the x-axis, the “Ancestral” model corresponds to the best-fitted model inferred with a bottleneck only in the ancestral population of Near Oceanians ((BKA,(NOC GST,SLI); Fig. 2a and Supplementary Fig. 21), and “Ancestral & all Islands” to a model with a bottleneck in the ancestral population of all Near Oceanians, as well as independent bottlenecks in each of the Near Oceanian populations.



Supplementary Figure 23. Fitting of the SFS of the refined model for Near Oceanians**.** We compared marginal 1-dimensional SFS of the observed data (in blue) and the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 simulations using parameters that best fit the data under the (BKA,(NOC GST,SLI)) model.



Supplementary Figure 24. SFS entries with worst fit of the refined model for Near Oceanians. Differences in the number of counts between the observed (in light blue) and expected (in dark blue) SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ unit of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele in Han Chinese (HAN, 2*n* = 4), Kundiawa (PNG, 2*n* = 10), Paiwan (TWN, 2*n* = 10), Nakanai Bileki (BKA, 2*n* = 10) and Vella Lavella (SLI, 2*n* = 10) (from bottom to top).



Supplementary Figure 25. Observed versus simulated F_{ST} for each pair of populations used for the refined demographic history of Near Oceania. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under the best parameters inferred for the refined model for Near Oceanians and were compared with observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.

The demographic history of western Remote Oceania

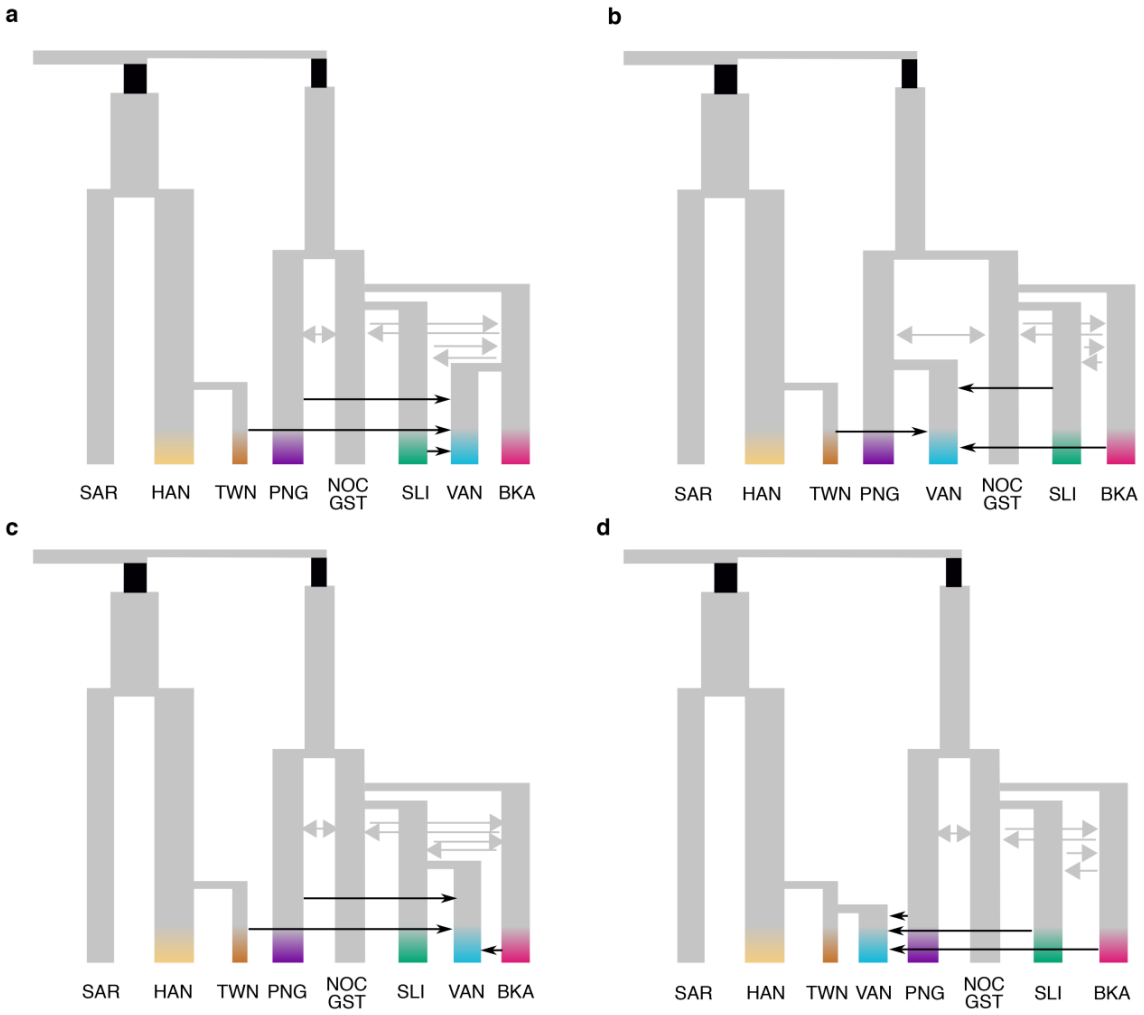
Demographic modelling and hypotheses. Evidence for the first human occupation of western Remote Oceania is dated to only ~3.5 ka, and is associated with the Lapita cultural complex[66]. A seminal ancient DNA study of three individuals from Vanuatu from the Lapita period reported their high genetic affinity to Austronesian-speaking populations, supporting an initial settlement of Vanuatu by Austronesian-related peoples[38]. Two more recent studies, primarily in Vanuatu, showed that such an initial settlement was rapidly followed by a partial population replacement by Papuan-related peoples, who share genetic affinities with populations from the Bismarck Archipelago[31,32].

To gain insight into the demographic history of western Remote Oceanians, we sought to model, in addition to baseline and Near Oceanian populations (see 'Refining the demographic history of Near Oceania'), a representative population of western Remote Oceanians, i.e., the ni-Vanuatu from Malakula or Emae Islands (VAN). Following a hypothesis-free approach, we modelled western Remote Oceanians as a population that diverges from any Near Oceanian population or from Taiwanese indigenous peoples (i.e., a proxy of Austronesian-speaking people expanding to Oceania), and subsequently receives separate gene flow pulses from any of these populations. Specifically, we tested four alternative models for the origins of ni-Vanuatu: (i) the (VAN,BKA) model assumes that the ni-Vanuatu (VAN) diverged from Bismarck islanders (BKA) and then received separate gene flow pulses from PNG, the Solomon islanders (SLI) and Taiwanese indigenous peoples (TWN), (ii) the (VAN,PNG) model assumes that the ni-Vanuatu diverged from PNG and then received separate gene flow from the three other populations, (iii) the (VAN,SLI) model assumes that the ni-Vanuatu diverged from the Solomon islanders and then received separate gene flow from the three other populations and finally, (iv) the (VAN,TWN) model assumes that the ni-Vanuatu diverged from Austronesian-speaking Taiwanese indigenous peoples (TWN) and then received separate gene flow from the three other populations (Supplementary Fig. 26). The intensity of the gene flow pulses was sampled from a log-uniform, so that low values are more probable than high values, to avoid difficulties of interpretation (i.e., a pulse of high intensity is equivalent to a population split in the model). Furthermore, we acknowledge that the genetic contribution of these different populations may have been inherited by western Remote Oceanians through a single migration event from an admixed population. However, we did not test such scenarios, because it would require exploring a large number of possible models. In light of this limitation, we interpret these results with caution.

The divergence time between the ni-Vanuatu (VAN) and the other populations was constrained to occur after ~700 generations ago (~20 ka, i.e., the divergence between Solomon islanders and the NOC Ghost). The time of gene flow from Taiwanese indigenous peoples (TWN) to both Near Oceanian populations was constrained to occur after ~300 generations (~9 ka, i.e., the HAN-TWN divergence). There was no *a priori* on the chronology of the three gene flow pulses into the ni-Vanuatu. All parameters estimated in the best-fitted refined model for Near Oceanians were fixed to ML estimates (see section 'Refining the demographic history of Near Oceania'), except migrations between Near Oceanians, as well as the time and proportion of admixture with Austronesian-speaking Taiwanese indigenous peoples in Bismarck and the Solomon islanders. All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

Dataset. We used the same dataset as in 'The demographic history of Near Oceania', except that we added 5 ni-Vanuatu individuals from Malakula or 5 ni-Vanuatu individuals from

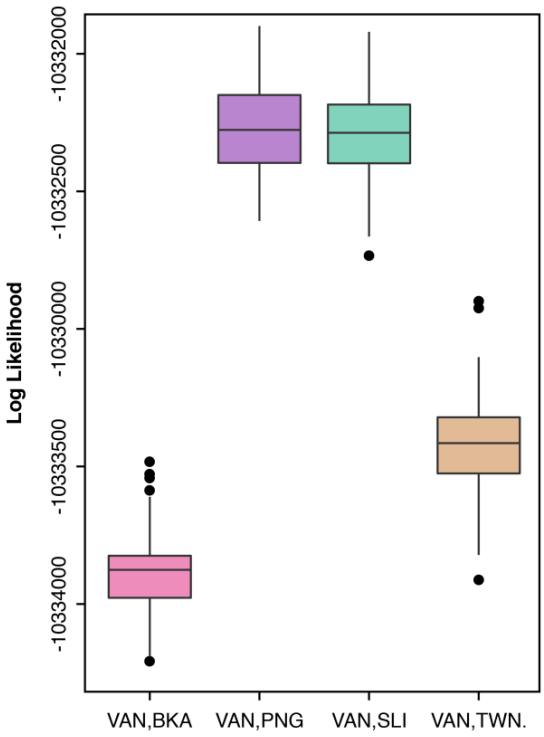
Emae, for replication, as populations representative of western Remote Oceanians (VAN), and we no longer included Han Chinese (HAN) in the multi-SFS. The multi-SFS were generated as for the baseline model.



Supplementary Figure 26. Alternative demographic models for western Remote Oceanians. **a**, The ni-Vanuatu diverged from the Bismarck Archipelago and then received gene flow from PNG, the Solomon Islands and Taiwanese indigenous peoples (VAN,BKA). **b**, The ni-Vanuatu diverged from PNG and then received gene flow from the three others groups (VAN,PNG). **c**, The ni-Vanuatu diverged from the Solomon Islands and then received gene flow from the three others groups (VAN,SLI). **d**, The ni-Vanuatu diverged from Taiwanese indigenous peoples and then received gene flow from the three other groups (VAN,TWN). **a-d**, SAR indicates Sardinians, HAN indicates Han Chinese, TWN indicates Taiwanese indigenous peoples, BKA indicates Bismarck islanders, SLI indicates Solomon islanders, PNG indicates Papua New Guinean highlanders, VAN indicates ni-Vanuatu, and NOC GST indicates an unsampled population from Near Oceania. For the sake of clarity, only populations from Eurasia, and Near and western Remote Oceania are shown. Grey arrows indicate migrations that are estimated in these models (single and double arrows for asymmetric and symmetric gene flow, respectively). Black arrows indicate gene flow pulses into the ni-Vanuatu.

Results. Among the four tested models, the *re-estimated* likelihood distributions of the (VAN,PNG) and (VAN,SLI) models were the highest and were largely overlapping, indicating no marked differences between these two models (ΔML = 37 $\log_{10}$ units, Supplementary Fig.

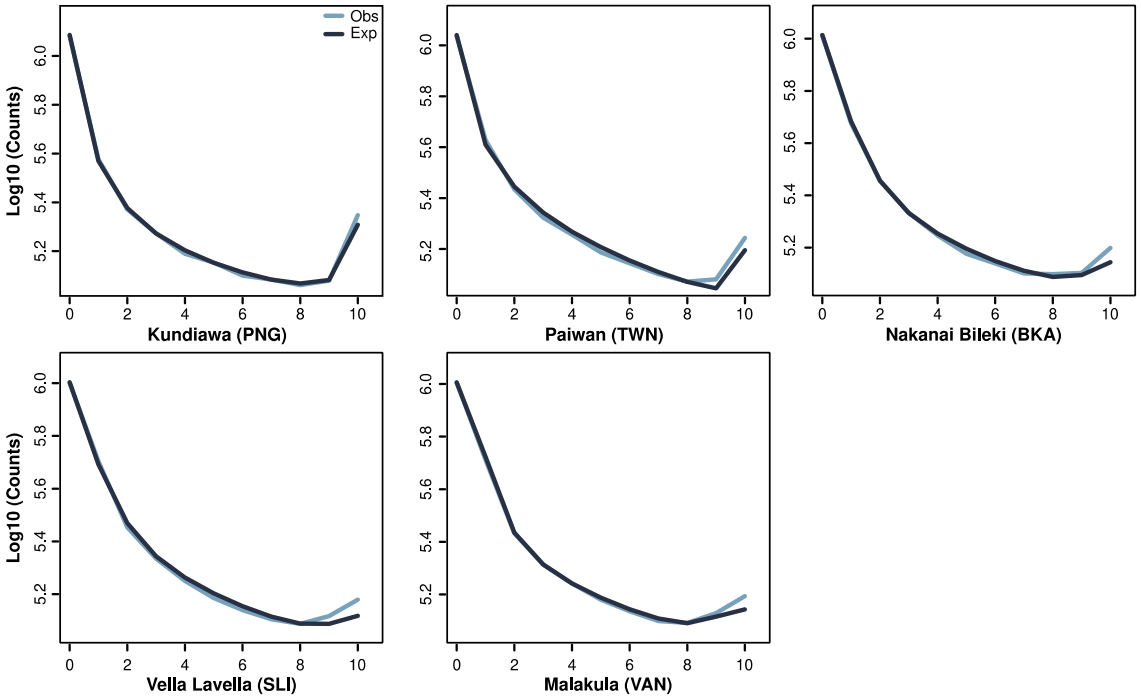
27). The model with the highest mean *re-estimated* likelihood was the (VAN,PNG) model, which assumes that the ni-Vanuatu (VAN) diverged from PNG and then received gene flow from the Solomon islanders (SLI), Taiwanese indigenous peoples (TWN) and Bismarck islanders (BKA) (Supplementary Fig. 26b). We estimated that the ancestral population of the ni-Vanuatu (VAN) diverged from PNG ~16 ka (95% CI: 12–18 ka, Extended Data Fig. 2b, Supplementary Table 5). It later received ~24% (95% CI: 14%–41%) of lineages from Solomon islanders (SLI) ~7 ka (95% CI: 4.1–11 ka). Under the second most likely (VAN, SLI) model (Supplementary Fig. 26c), the ancestral population of the ni-Vanuatu (VAN) diverged from Solomon islanders (SLI) ~12 ka (95% CI: 10–16 ka, Extended Data Fig. 2b, Supplementary Table 5), and later received ~44% (95% CI: 27%–57%) of PNG lineages ~9 ka (95% CI: 6.3–13 ka). This suggests that the Papuan-related population entering Vanuatu at the end of the Lapita period was different and more diverse than the Bismarck islanders modelled in our study[32,41]. Importantly, under both the (VAN,PNG) and (VAN,SLI) models, the ni-Vanuatu (VAN) received < 3% of lineages from Austronesian-speaking Taiwanese indigenous peoples (TWN) ~2–3 ka, and ~34–39% of Bismarck Archipelago (BKA) lineages < 2 ka (Extended Data Fig. 2b and Supplementary Table 5). This result was confirmed when modelling ni-Vanuatu from Emae, instead of Malakula (Supplementary Table 5). Furthermore, we found that the accuracy of parameter estimations in this model was high, using parametric bootstrap (Supplementary Table 8 'Remote Oceania'). Collectively, our findings support a very low, direct genetic contribution of Taiwanese indigenous peoples to the ni-Vanuatu, suggesting that the bulk of the East Asian ancestry detected in present-day western Remote Oceanians was inherited from already admixed Near Oceanians.



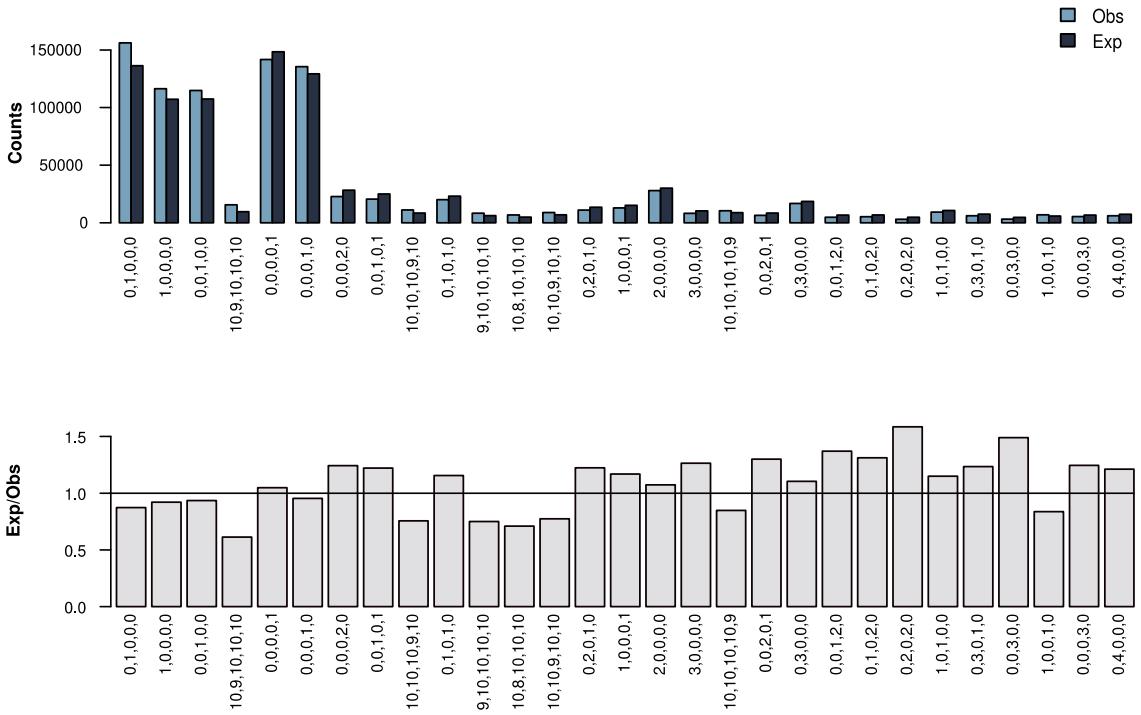
Supplementary Figure 27. Likelihood distribution of alternative models for western Remote Oceanians. The line, box, whiskers and points respectively indicate the median, IQR range, 1.5*IQR and outliers of the likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario.

Model fitting. We obtained a very good fit of expected and observed marginal SFS, except for high-frequency derived alleles (Supplementary Fig. 28). The entries of the joint SFS with the poorest fit were also those where the derived allele is fixed in most modern human samples

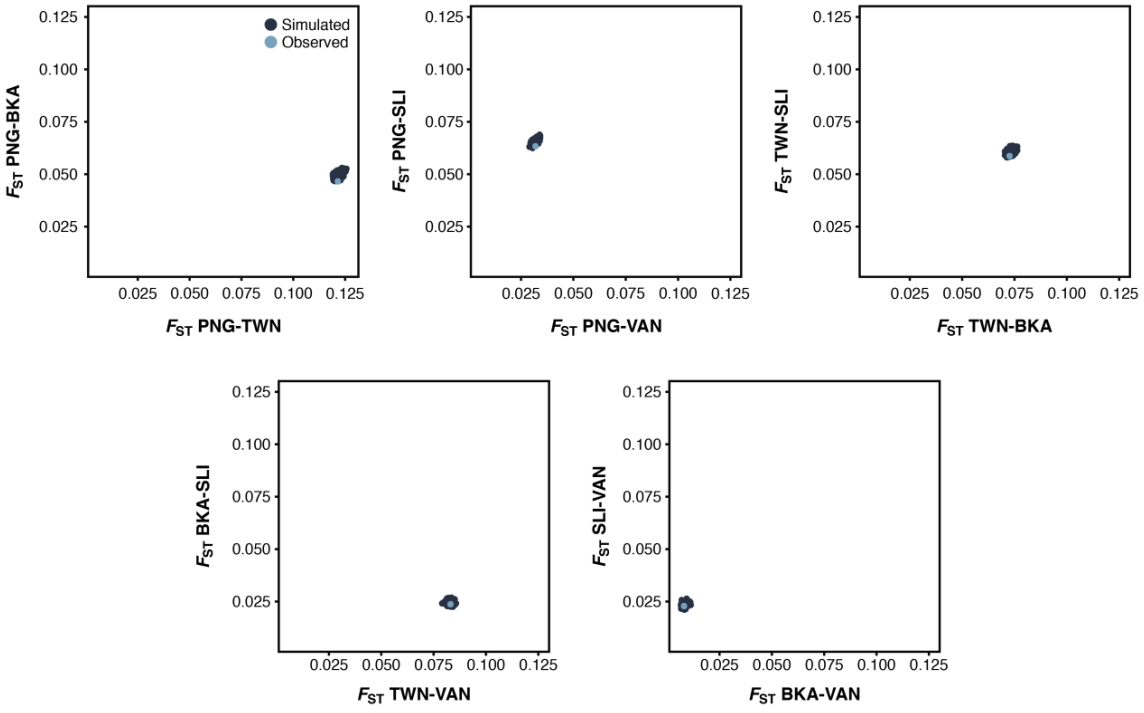
(Supplementary Fig. 29). As for the baseline model, this is probably due to ancestral state misspecification. We observed a very good fit between observed and expected F_{ST} (Supplementary Fig. 30), indicating that the model and parameter estimates well reproduce this aspect of the data.



Supplementary Figure 28. Fitting of the SFS of the model for western Remote Oceanians. We compared marginal 1-dimensional SFS of the observed data (in blue) and the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 simulations using parameters that best fit the data under the best model (ni-Vanuatu diverged from PNG and received gene flow from the three other groups (VAN,PNG)).



Supplementary Figure 29. SFS entries with worst fit for the best-fitted model for western Remote Oceanians. Differences in the number of counts between the observed and expected SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele in Kundiawa (PNG, 2*n* = 10), Paiwan (TWN, 2*n* = 10), Nakanai Bileki (BKA, 2*n* = 10), Vella Lavella (SLI, 2*n* = 10) and Malakula islanders (VAN, 2*n* = 10) (from bottom to top).



Supplementary Figure 30. Observed versus simulated F_{ST} for each pair of populations used for the model for western Remote Oceanians. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under parameters inferred for the model of western Remote Oceania (VAN, PNG), and were compared to observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.

The sources of East Asian ancestry among Oceanians

Demographic modelling and hypotheses. To gain insights into the genetic history of populations contributing East Asian-related ancestry to Oceanians, we sought to model, in addition to baseline populations from East Asia (Han Chinese, HAN, and Taiwanese indigenous peoples, TWN), a Malayo-Polynesian-speaking population from the Philippines (PHP). To represent the latter population, we used the Kankanaey because they show little Philippine Agta 'Negrito' ancestry[17], unlike the Cebuano (Extended Data Fig. 1). Although we could have modelled gene flow from the Agta to the Cebuano, we decided instead to use the Kankanaey, to keep the model as simple as possible and to limit the number of parameters to estimate. We first sought to estimate the divergence time between Austronesian-speaking populations from Taiwan and the Philippines, i.e., between Taiwanese indigenous peoples (i.e., Formosan speakers) and the Philippine Kankanaey (i.e., Malayo-Polynesian speakers). To do so, we assumed that Taiwanese indigenous peoples (TWN) and Philippine Kankanaey (PHP) are sister groups that have evolved under isolation with asymmetric migration (IM). We checked whether our results extend to Austronesian-speaking populations outside of the Philippines, by adding to the previous model, two Oceanian populations with high levels of East Asian-related ancestry, i.e., Polynesian outliers (POL; Polynesians share high genetic affinities with ancient DNA samples from the Lapita period[31,32,38]), and Solomon islanders (SLI) (Fig. 1, Extended data Fig. 1).

We fixed parameters relating to events that predate the divergence between Eurasians and Near Oceanians, as well as parameters specific to Europeans, Kundiawa PNG and archaic introgression from Neanderthal and Denisova, to the point estimates obtained in previous models (Supplementary Tables 2 and 4). We assumed that Han Chinese (HAN) diverged from the ancestors of Austronesian-speaking populations, followed by the divergence of Formosan-speaking Taiwanese indigenous peoples (TWN) and Malayo-Polynesian speakers (PHP and POL). This tree topology is supported by significant *D*-statistic results ($Z > 2$, Supplementary Table 9) and phylo-linguistic analyses of Austronesian languages[2]. For all models, we considered migrations between populations following a stepping-stone model. In the models including Polynesian outliers (POL) or Solomon islanders (SLI), we accounted for population structure in Near Oceanians by simulating an unsampled population representing Near Oceanians (NOC Ghost, see 'Refining the demographic history of Near Oceania'), which diverged from PNG. We simulated a gene flow pulse from the NOC Ghost to Polynesian outliers (POL), to account for their Papuan-related ancestry (Fig. 1, Extended Data Fig. 1). In the model including Near Oceanians from the Solomon Islands, Solomon islanders (SLI) diverged from the NOC Ghost and later received a gene flow pulse from another unsampled ghost population, which diverged from Austronesian-speaking Taiwanese indigenous peoples. This 'SEA Ghost' population represents an East Asian-related population that migrated to Near Oceania and admixed with autochthonous groups. All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

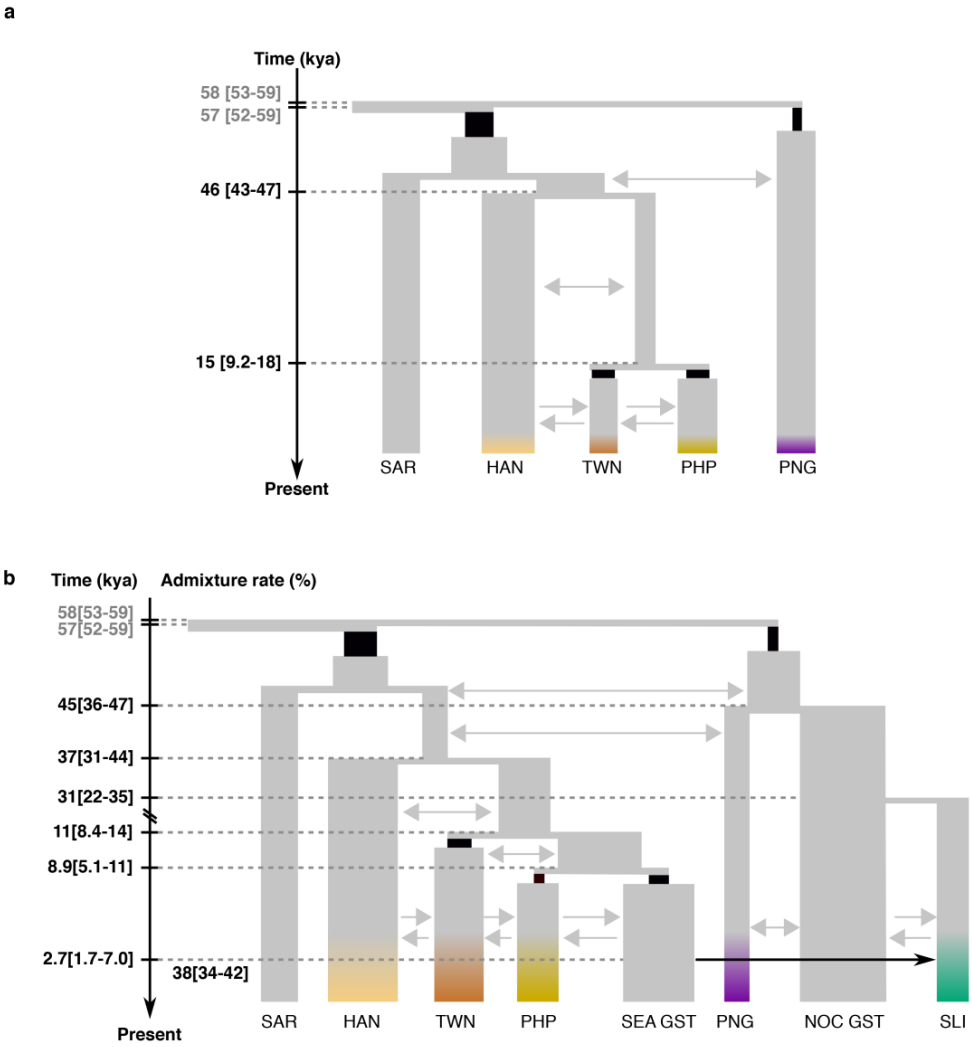
Dataset. SFS data for Malayo-Polynesian speakers included (i) 2 Kankanaey from the Philippines (PHP; Supplementary Fig. 31a), (ii) 5 Polynesian outliers from Tikopia Island (POL; Extended Data Fig. 2c), (iii) 5 Polynesian outliers from Bellona Island (POL) as a replicate, or (iv) 5 Solomon islanders from Vella Lavella (SLI; Supplementary Fig. 31b). For the baseline populations, SFS data for 2 Han Chinese (HAN), 5 Paiwan (Taiwanese

indigenous peoples, TWN) and 5 Kundiawa (PNG) were used. The multi-SFS was generated as for the 'Baseline demographic model of human populations'.

Results. In all models, we estimated that the ancestors of Han Chinese (HAN) and Taiwanese indigenous peoples (TWN) separated >20 ka (Supplementary Table 6), in contrast with our baseline model (Supplementary Table 2), where constant population sizes and symmetric gene flow were assumed for model simplicity. Our models for East Asian-related populations therefore suggest a relatively ancient structure among continental East Asia and Taiwan, which was first settled 20–30 ka[4]. Alternatively, these results suggest that gene flow from a non-modelled population into Han Chinese and/or Taiwanese indigenous peoples artificially inflates divergence time estimates (see 'Refining the sources of East Asian ancestry among Oceanians').

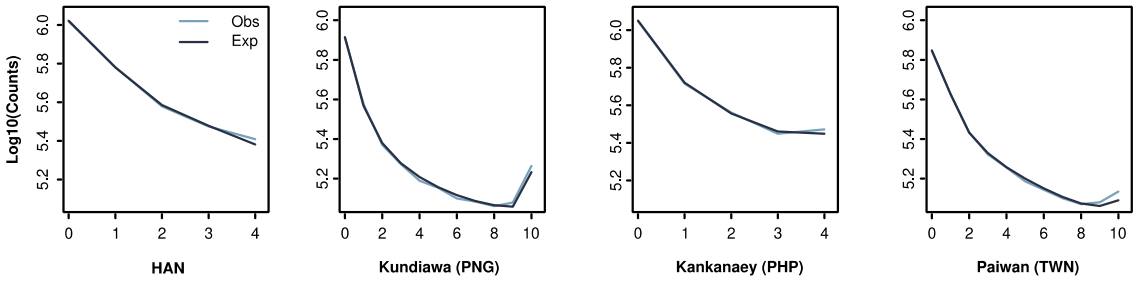
We estimated the divergence between Taiwanese indigenous peoples (TWN) and Philippine Kankanaey (PHP) at ~15 ka (95% CI: 9.2–18 ka), under an isolation-with-migration model (Supplementary Fig. 31a). When modelling Austronesian-speaking populations outside the Philippines, we estimated the divergence between Formosan-speaking Taiwanese indigenous peoples (TWN) and Malayo-Polynesian-speaking groups (i.e., Philippine Kankanaey, PHP, and Polynesian outliers, POL) to ~7.3 ka (95% CI: 6.4–11 ka), when using the Tikopia population to represent Polynesians (Extended Data Fig. 2c, Supplementary Table 6). When replicating this model using Polynesian outliers from Bellona Island, the divergence was dated to ~11 ka (Supplementary Table 6). Collectively, these estimations suggest that population differentiation among Austronesian-speaking populations predates the emergence of agriculture in Taiwan, which is thought to have started ~4,8 ka[2,62]. To confirm these estimations, we also used another model where a population from the Solomon Islands (SLI) receives gene flow from an unsampled East Asian-related source (SEA Ghost; Supplementary Fig. 31b). The divergence between Taiwanese indigenous peoples (TWN) and the source of the East Asian-related ancestry in Solomon islanders (SEA Ghost) was dated to ~11 ka (95% CI: 8.4–14 ka, Supplementary Table 6), reinforcing the notion that ancestors of Formosan- and Malayo-Polynesian-speaking populations were isolated before the emergence of agriculture in Taiwan[2,62].

We estimated that Polynesian outliers (POL) received a pulse of gene flow from Near Oceania ~0.5 ka (95% CI: 0.4–1.1 ka, Extended Data Fig. 2c) that contributed ~35% (95% CI: 32%–36%) to their gene pool, in agreement with ADMIXTURE results (Extended Data Fig. 1). Conversely, we estimated that Solomon islanders (SLI) received a pulse of gene flow from an East Asian-related source ~2.7 ka (95% CI: 1.7–7.0 ka, Supplementary Fig. 31b) that contributed ~38% (95% CI: 34%–42%) to their gene pool, which we interpret as the signature of the demic diffusion of the Lapita cultural complex to the region[66,67]. Finally, we found that the effective population size of Polynesian outliers (POL) was highly reduced (N_e = 134, 95% CI: 119–230), suggesting the occurrence of strong bottlenecks during the settlement of Polynesia[68] and/or the subsequent back migrations to the Solomon Islands[10]. Furthermore, we estimated a stronger founder effect in Polynesian outliers from Bellona, relative to Tikopia (Supplementary Table 6), in agreement with our empirical observations (Supplementary Note 3). This indicates that Polynesian groups experienced founder effects of various intensities following their back migrations to the Solomon Islands.

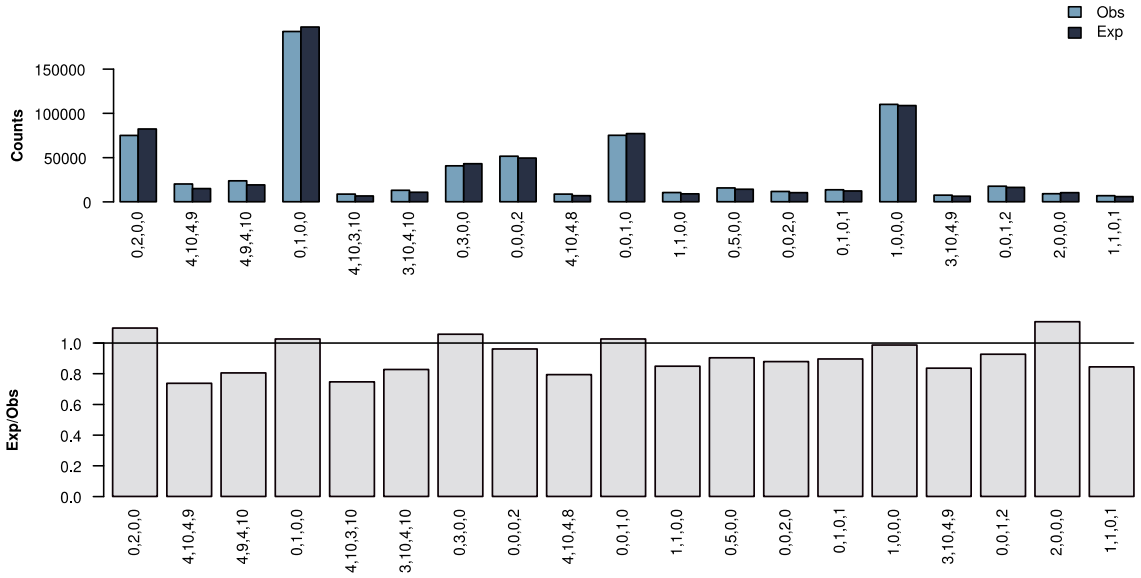


Supplementary Figure 31. Demographic models for East Asian-related populations of the Pacific. **a**, Best-fitted model for Taiwanese indigenous peoples and Philippine Kankanaey. **b**, Best-fitted model for East Asian-related populations contributing to Near Oceanians. **a-c**, SAR indicates Sardinians, HAN indicates Han Chinese, TWN indicates Taiwanese indigenous peoples, PHP indicates the Kankanaey from the Philippines, SEA GST indicates an unsampled population that represents a Southeast Asian-related population contributing to Near Oceanians, PNG indicates Papua New Guinean highlanders, NOC GST indicates a meta-population of Near Oceanians and SLI indicates Solomon islanders. Point estimates of all parameters and corresponding 95% CIs are given in Supplementary Table 6. Timing of events is given in ka, assuming a generation time of 29 years. Single pulse admixture rates are reported in %. 95% CIs are given in square brackets. The larger the rectangle width, the larger the effective population size (N_e). Bottlenecks are indicated by black rectangles. Grey and black arrows represent continuous and single pulse gene flow, respectively. Uni- and bi-directional arrows indicate estimated symmetric and asymmetric migrations.

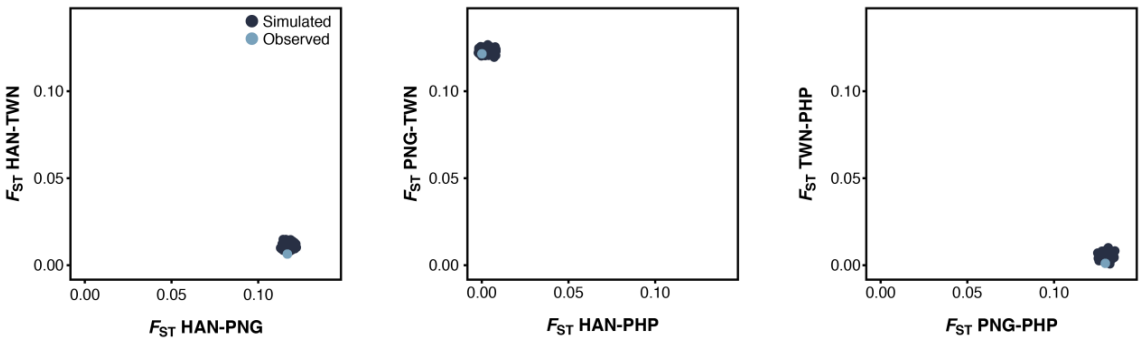
Model fitting. We obtained a very good fit of expected and observed marginal SFS, except for high-frequency derived alleles (Supplementary Figs. 32 and 35). The entries of the joint SFS with the poorest fit were also those where the derived allele is fixed most modern human samples (Supplementary Figs. 33 and 36). As for the baseline model, this is probably due to ancestral state misspecification. We observed a very good fit between observed and expected F_{ST} (Supplementary Figs. 34 and 37), indicating that the model and parameter estimates well reproduce this aspect of the data.



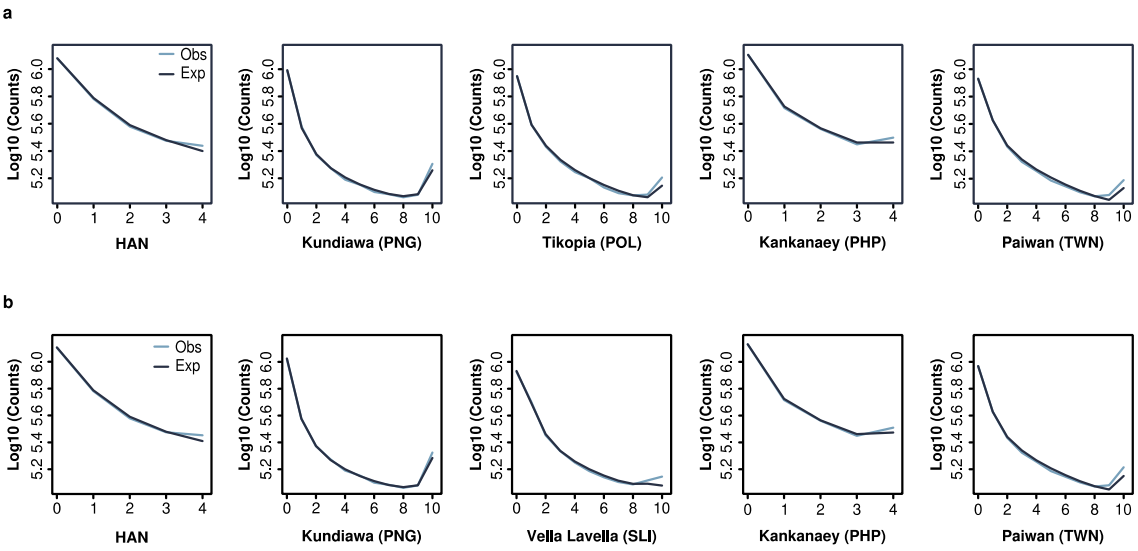
Supplementary Figure 32. Fitting of the SFS for the model for Taiwanese indigenous peoples and Philippine Kankanaey. We compared marginal 1-dimensional SFS of the observed data (in blue) and the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 simulations using parameters that best fitted the data.



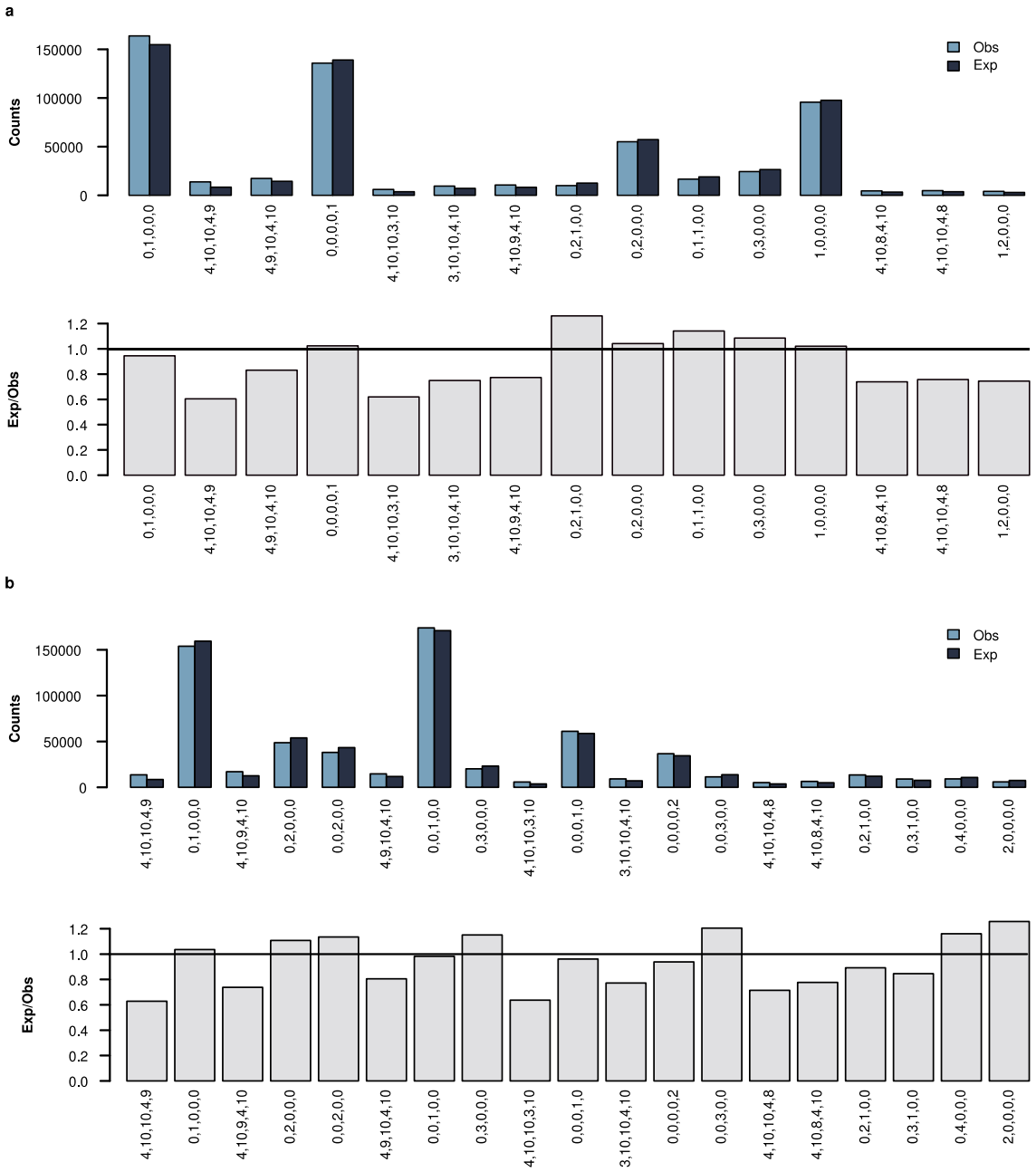
Supplementary Figure 33. SFS entries with worst fit for the model of Taiwanese indigenous peoples and Philippine Kankanaey. Differences in the number of counts between the observed and expected SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele in Han Chinese (HAN, $2n = 4$), Kundiawa (PNG, $2n = 10$), Kankanaey (PHP, $2n = 4$) and Paiwan (TWN, $2n = 10$) (from bottom to top).



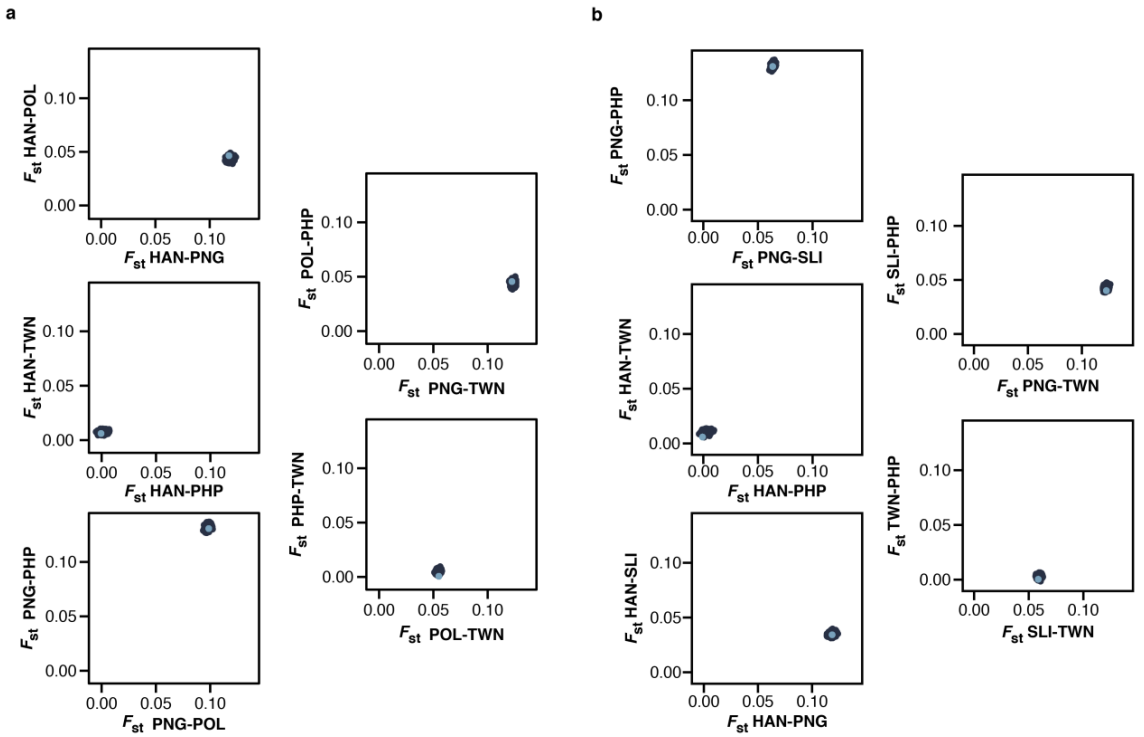
Supplementary Figure 34. Observed versus simulated F_{ST} for each pair of populations used for the model for Taiwanese indigenous peoples and Philippine Kankanaey. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under parameters inferred for the best-fitted model, and were compared to observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.



Supplementary Figure 35. Fitting of the SFS of the model for East Asian-related populations contributing to Near Oceanians. **a,** Model with Polynesian outliers (Tikopia (POL)) and **b**, with Near Oceanians (Vella Lavella (SLI)). We compared marginal 1-dimensional SFS of the observed data (in blue) and the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 coalescent simulations using parameters that best fit the data under the best models.



Supplementary Figure 36. SFS entries with the worst fit of the model for East Asian-related populations contributing to Oceanians. **a,** Model with Polynesian outliers (Tikopia (POL)) and **b**, with Near Oceanians (Vella Lavella (SLI)). Differences in the number of counts between the observed and expected SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele in Han (HAN, 2*n* = 4), Kundiawa (PNG, 2*n* = 10), Bellona (POL, 2*n* = 10) or Vella Lavella (SLI, 2*n* = 10), Kankanaey (PHP, 2*n* = 4) and Paiwan (TWN, 2*n* = 10) (from bottom to top).



Supplementary Figure 37. Observed versus simulated F_{ST} for each pair of populations used in the model for East Asian-related populations contributing to Oceanians. **a,** Model with Polynesian outliers (Tikopia (POL)) and **b**, with Near Oceanians (Vella Lavella (SLI)). Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under parameters inferred for the best models and were compared to observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.

Refining the sources of East Asian ancestry among Oceanians

Demographic modelling and hypotheses. Our models for East Asian-related populations from the Pacific suggest a divergence of Taiwanese indigenous peoples and Malayo-Polynesian speakers that occurred earlier than 5 ka, at odds with a demic diffusion of agriculture and Austronesian languages from Taiwan to Oceania ~4.8 ka[2,69,70]. A possible caveat of these models is that the modelled populations, including Han Chinese, Taiwanese indigenous peoples, Philippine Kankanaey and Polynesians may have received gene flow from a non-modelled, distantly-related population, which could bias upward divergence time estimates. In this context, a recent ancient DNA study has found evidence for gene flow from Northeast into Costal Southeast Asia after the Neolithic[57]. Based on this, we modified our model (Extended Data Fig. 2c) by adding an unsampled population that represents northeastern Asian groups (NEA Ghost; Supplementary Fig. 38). We considered two alternative models of gene flow from Northeast to East/Southeast Asians. The first model, referred to as the '3-pulse' model, includes gene flow from the NEA Ghost to Han Chinese (HAN), to Taiwanese indigenous peoples (TWN) and to the ancestral population of Malayo-Polynesian speakers (i.e., Philippine Kankanaey, PHP, and Polynesians, POL; Fig. 2b). The second model, referred to as the '2-pulse' model, includes gene flow from the NEA Ghost to Han Chinese (HAN) and to the ancestral population of Austronesian speakers (here, TWN, PHP and POL; Supplementary Fig. 38c). To enable model comparison, we estimated the same number of parameters for the '2-pulse' and '3-pulse' models; for the '3-pulse' model (Fig. 2b), we assumed that admixture of Northeast Asians with each of the three East/Southeast Asian groups occurred at the same time, whereas for the '2-pulse' model (Supplementary Fig. 38c), we allowed for different times of admixture.

We also reasoned that the well-established introgression event from Denisovans into East Asians (Extended Data Fig. 10; ref.[71]) could affect parameter estimation. We thus allowed for gene flow from the Altai Denisovan to the ancestral population of East/Southeast Asian groups. We also allowed for gene flow between (i) Europeans (Sardinians, SAR) and the NEA Ghost population and (ii) between Taiwanese indigenous peoples (TWN) and the ancestors of Malayo-Polynesian-speaking populations, as well as a single pulse admixture from Han Chinese (HAN) to Taiwanese indigenous peoples (TWN), to account for the recent expansion of Han Chinese to Taiwan[72]. For both the '2-pulse' and '3-pulse' models, our demographic parameters of interest were (i) the divergence times between all East/Southeast Asian populations, (ii) the contribution of the NEA Ghost to East/Southeast Asian populations, (iii) the migration rate between Taiwanese indigenous peoples (TWN) and the ancestors of Malayo-Polynesian-speaking populations, (iv) the time and rate of admixture from Han Chinese (HAN) to Taiwanese indigenous peoples (TWN), and (v) the effective population sizes of East/Southeast Asians.

Before estimating these parameters of interest, we used a simplified version of the '2-pulse' and '3-pulse' models, where we assumed no gene flow from the NEA Ghost to other groups, to estimate the time and rate of the Denisovan introgression into the ancestral population of East/Southeast Asian groups, as well as the migration between Europeans (Sardinians, SAR) and the NEA Ghost population (Supplementary Fig. 38a). To limit the number of parameters to estimate, parameters relative to events predating the divergence between Eurasians and Near Oceanians, as well as parameters specific to Europeans, Papuans (Kundiawa, KUN and NOC Ghost) and archaic introgression from the Altai Denisovan (to the ancestor of Papuan groups) and Neanderthal, were fixed to the point estimates obtained in previous models (Supplementary Tables 2, 4 and 6). Similarly, to account for the Papuan-related ancestry found in Polynesians (POL) (Fig. 1, Extended Data Fig. 1), we also fixed the rate and time of admixture from the NOC Ghost into Polynesians (POL), based on the point estimate previously obtained (see section 'The sources of East Asian ancestry among Oceanians'; Extended Data Fig. 2c and Supplementary Table 6 'PHP-POL'). All *fastsimcoal2* input files can be found on GitHub (www.github.com/h-e-g/evoceania).

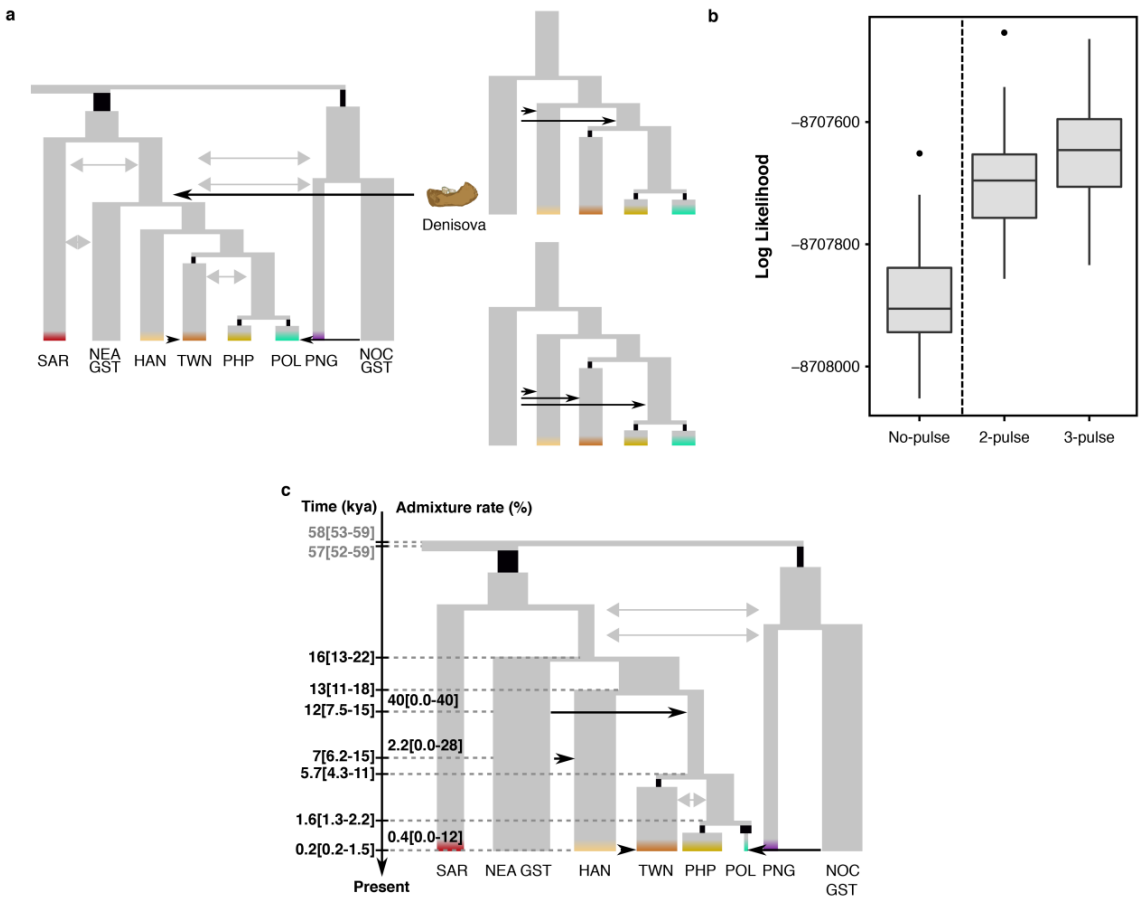
Dataset. SFS data for Malayo-Polynesian speakers included (i) 2 Kankanaey from the Philippines (PHP) and (ii) 5 Polynesian outliers from Tikopia (POL). For baseline populations, SFS data for 1 Sardinian (SAR), 2 Han Chinese (HAN), 5 Paiwan (Taiwanese indigenous peoples, TWN) and 1 Kundiawa (PNG) were used. The multi-SFS was generated as for the 'Baseline demographic model of human populations'.

Results. We found stronger support for the '3-pulse' model, relative to the '2-pulse' model ($\Delta ML = 53$ $\log_{10}$ units, based on the mean *re-estimated* likelihoods; Supplementary Fig. 38b). Under the '3-pulse' model, we estimated that the ancestors of Northeast Asians and East/Southeast Asians diverged ~18 ka (95% CI: 14–22 ka), and ancestors of Han Chinese (HAN) diverged from the ancestors of Formosan (TWN) and Malayo-Polynesian speakers (PHP and POL) ~14 ka (95% CI: 11–18 ka; Fig. 2b and Supplementary Table 7 '3-pulse'). Similar divergence times were obtained under the '2-pulse' model (Supplementary Fig. 38c and Supplementary Table 7 '2-pulse'). Taiwanese indigenous peoples (TWN) diverged from Malayo-Polynesian speakers (PHP and POL) ~8.2 ka (95% CI: 4.8–12.0 ka) under the '3-pulse' model, and ~5.7 ka (95% CI: 4.3–11 ka) under the '2-pulse' model. We found that the accuracy of parameter estimations in the '3-pulse' model was good, using parametric bootstrap (Supplementary Table 8 'East Southeast Asia'), except for the admixture rates from Northeast Asians to East/southeast Asian groups. These results suggest that modelling gene flow from an unsampled population representing Northeast Asians does not largely affect the divergence time between Taiwanese indigenous peoples and Malayo-Polynesian speakers. Collectively, despite a large confidence interval, our most likely model suggests that the ancestors of present-day Austronesian speakers separated before the Taiwanese Neolithic[69], questioning the strict Out-of-Taiwan model[70]. However, further investigation will

be needed to evaluate whether other models can better explain the patterns of genetic diversity observed in the region. These limitations indicate that archaeological and paleogenomic studies will be required to better understand the complex peopling history of the Pacific.

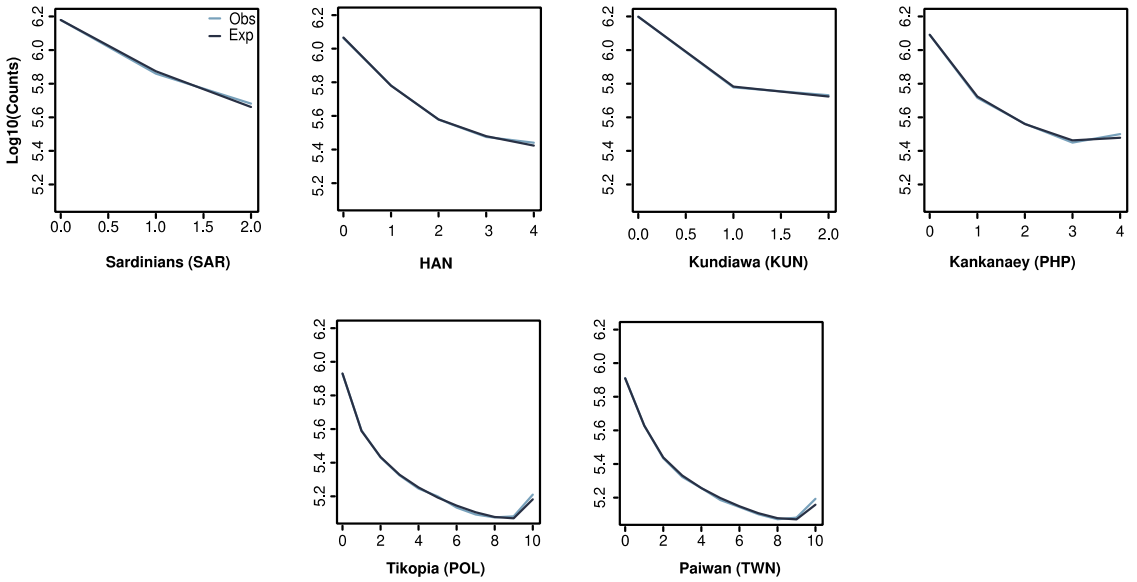
Finally, we confirmed that the effective population size of Polynesian outliers (POL) was highly reduced, based on both the '3-pulse' and '2-pulse' models (N_e ~130, 95% CIs: [107–156] and [110–162]; Supplementary Table 7 '2-pulse' and '3-pulse'), suggesting strong population bottlenecks during the settlement of Polynesia[68] and/or the subsequent back migrations to the Solomon Islands[10].

Model fitting. We obtained a very good fit of expected and observed marginal SFS, except for high-frequency derived alleles (Supplementary Fig. 39). The entries of the joint SFS with the poorest fit were singletons and doubletons in the Sardinians (SAR), probably because we fixed all parameters related to this population, and entries where the derived allele is fixed in most modern human samples (Supplementary Fig. 40). As for the baseline model, this is probably due to ancestral state misspecification. We observed a very good fit between observed and expected F_{ST} (Supplementary Fig. 41), indicating that the model and parameter estimates well reproduce this aspect of the data, except for the pairwise comparison with Sardinians (SAR) and Formosan and Malayo-Polynesian speakers, suggesting, again, that fixed parameter values for Sardinians (SAR) reduce fitting.

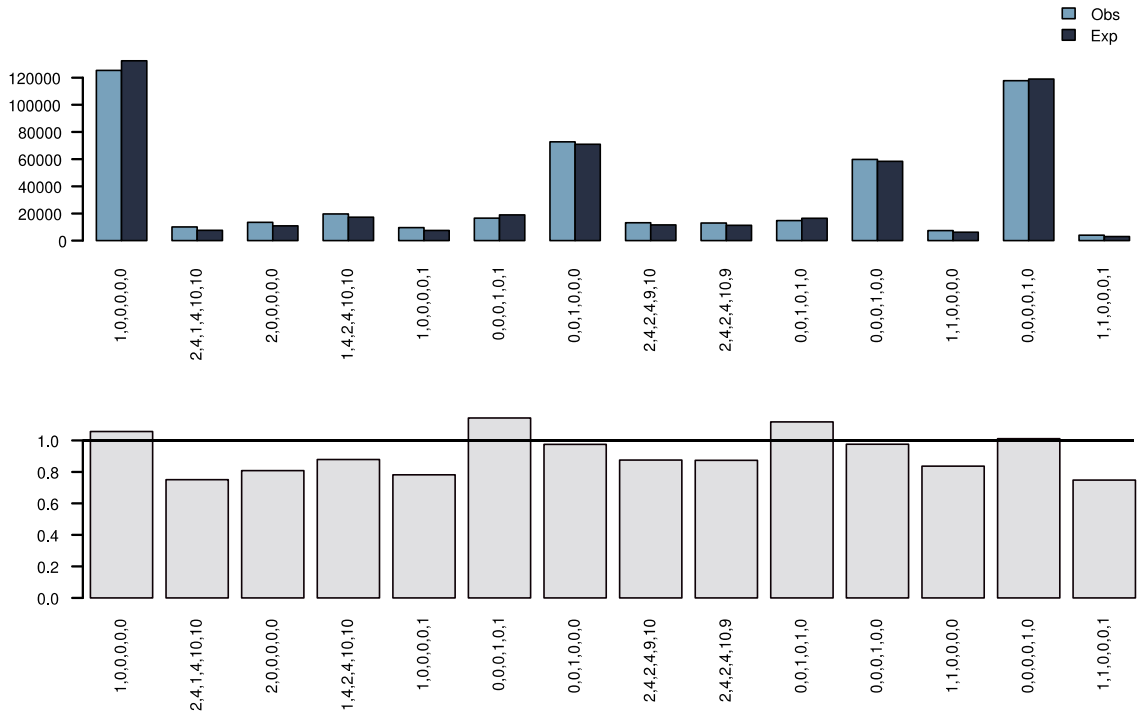


Supplementary Figure 38. Alternative refined models of sources of East Asian-related ancestry among Oceanians, with or without gene flow from Northeast Asians to East/Southeast Asians. **a**, Schematic representation of alternative models for Formosan- and Malayo-Polynesian-speaking populations, with (models to the right) or without (model to the left, "No-pulse") gene flow from the NEA Ghost to the different groups of East/Southeast Asians (HAN, TWN, PHP and POL). Models in the top

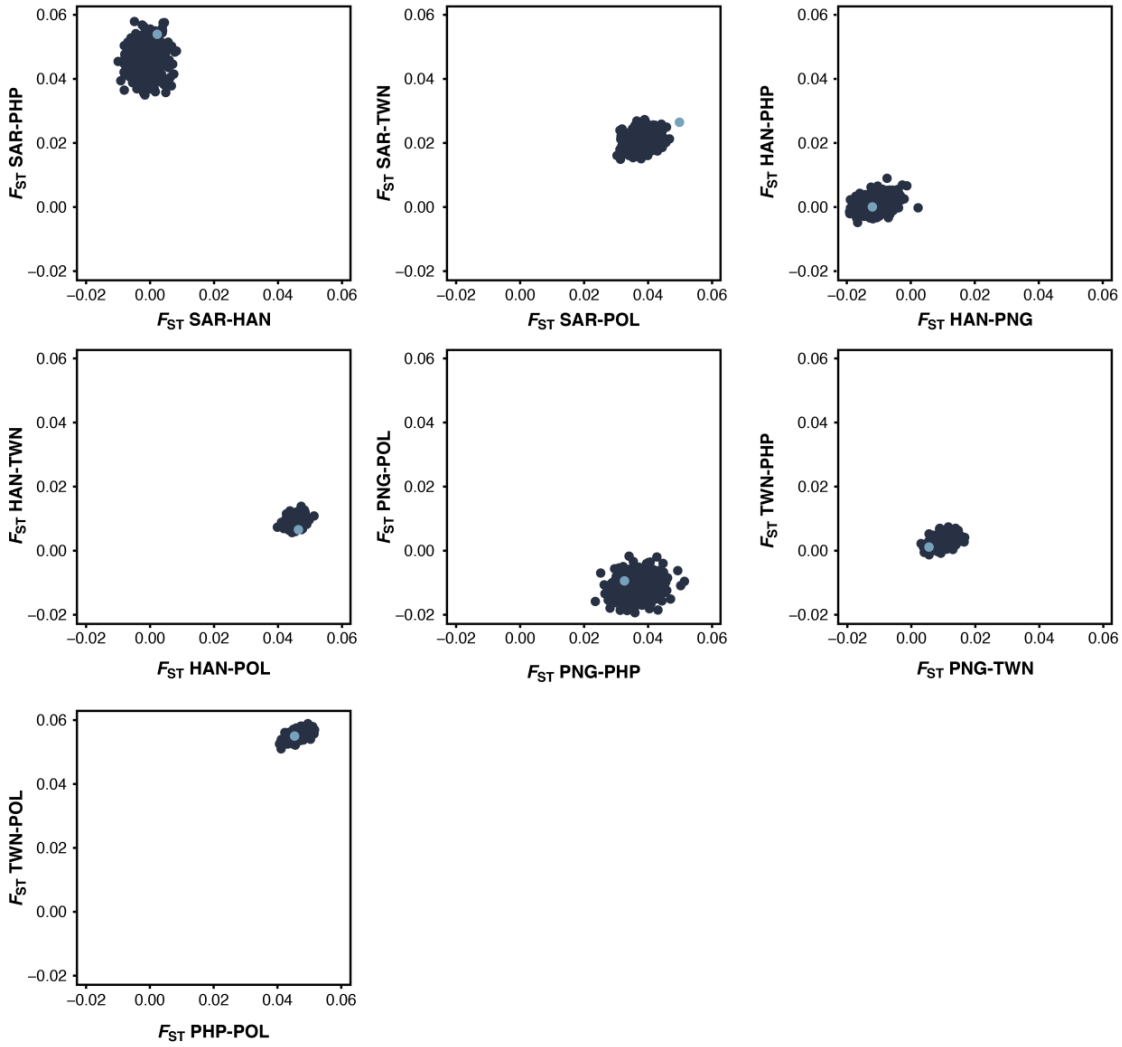
right and the bottom right corners represent the “2-pulse” and “3-pulse” models, respectively. **b**, Likelihood distribution of the three alternative models. The line, box, whiskers and points respectively indicate the median, IQR range, 1.5*IQR and outliers of the likelihood distributions obtained from 100 expected SFS computed with 10^7 coalescent simulations and using parameters that maximized the likelihood under each scenario. **c**, Maximum-likelihood 2-pulse model for Formosan- and Malayo-Polynesian-speaking populations. Point estimates of all parameters and 95% CIs are given in Supplementary Table 7. Formosan speakers are represented by the Paiwan from Taiwan (TWN) and Malayo-Polynesian-speaking populations by Philippine Kankanaey (PHP) and Polynesians from Tikopia (POL). Point estimates of all parameters and 95% CIs are given in Supplementary Table 7. The 95% CIs are given in square brackets. SAR indicates Sardinians, NEA GST indicates a Northeast Asian unsampled population, HAN indicates Han Chinese, TWN indicates Taiwanese indigenous peoples, PHP indicates the Kankanaey from the Philippines, POL indicates Polynesians from Tikopia, NOC GST indicates a Near Oceanian meta-population and PNG indicates Papua New Guinean highlanders. The larger the rectangle width, the larger the estimated effective population size (N_e). Bottlenecks are indicated by black rectangles. Bi-directional arrows indicate symmetric gene flow, and grey and black arrows represent continuous and single pulse gene flow, respectively. We assumed a mutation rate of 1.25×10^{-8} mutation/generation/site and a generation time of 29 years. We limited the number of parameters to be estimated, by making simplifying assumptions concerning the recent demography of Near Oceanian populations (Supplementary Note 4). Sample sizes are described in Supplementary Note 4. The admixture pulses from NEA Ghost were constrained to occur after the divergence between Han Chinese (HAN) and the ancestral population of Austronesian speakers (Supplementary Table 7). Time axes are not at scale.



Supplementary Figure 39. Fitting of the SFS of the refined model of sources of East Asian-related ancestry among Oceanians. We compared marginal 1-dimensional SFS of the observed data (in blue) and the averaged expected SFS (in black) obtained from 100 SFS approximated with 10^7 coalescent simulations using parameters that best fit the data under the best models.



Supplementary Figure 40. SFS entries with the worst fit of the refined model of sources of East Asian-related ancestry among Oceanians. Differences in the number of counts between the observed and expected SFS for entries harbouring a discrepancy of more than 500 $\log_{10}$ units of likelihood. The plot at the bottom gives the relative fit computed as the ratio of number of counts for the i^{th} entry in the expected and observed SFS. Entries are given in column and corresponds to number of counts of the derived allele in Sardinians (SAR, 2*n* = 2), Han (HAN, 2*n* = 4), Kundiawa (PNG, 2*n* = 2), Kankanaey (PHP, 2*n* = 4), Tikopia (POL, 2*n* = 10) and Paiwan (TWN, 2*n* = 10) (from bottom to top).



Supplementary Figure 41. Observed versus simulated F_{ST} for each pair of populations used in the refined model of sources of East Asian-related ancestry among Oceanians. Simulated pairwise F_{ST} (dark blue) were obtained with 500 simulations under parameters inferred for the best models and were compared to observed F_{ST} (light blue) obtained from the empirical data used for parameter inference.

Supplementary Note 5: East Asian Admixture in Near Oceania

Rationale

Gene flow from East Asia into Remote Oceania was previously dated to 1.5–2.5 ka, and was attributed to the expansions of Austronesian speakers into the Pacific starting from Taiwan ~5 ka[31,32,38]. Yet, the methods used in these studies assumed that gene flow was instantaneous. Likewise, in our ML model for Near Oceanians (see 'Refining the demographic history of Near Oceania'), we assumed that gene flow occurred as a single, instantaneous pulse, to simplify parametrization, and estimated that admixture occurred ~4 ka (95%CI: 3.2–5.5 ka) (Fig. 2a). We reasoned that this assumption may be unrealistic, and could bias the estimation of the time of the gene flow pulse(s). Indeed, a recent study has suggested that the discrepancies between admixture time estimates obtained by different methods could be explained by the occurrence of several pulses of gene flow in Near and Remote Oceanians[30]. To determine the mode and tempo of admixture in Near Oceanians, we applied an approximate Bayesian computation (ABC) approach[73], developed in the MetHis method[74], to estimate the posterior probability of three competing admixture models: a single-pulse, a two-pulse or a constant-recurring model of admixture. MetHis relies on explicit forward-in-time simulations of complex admixture histories following a general mechanistic admixture model[75].

Simulation setting

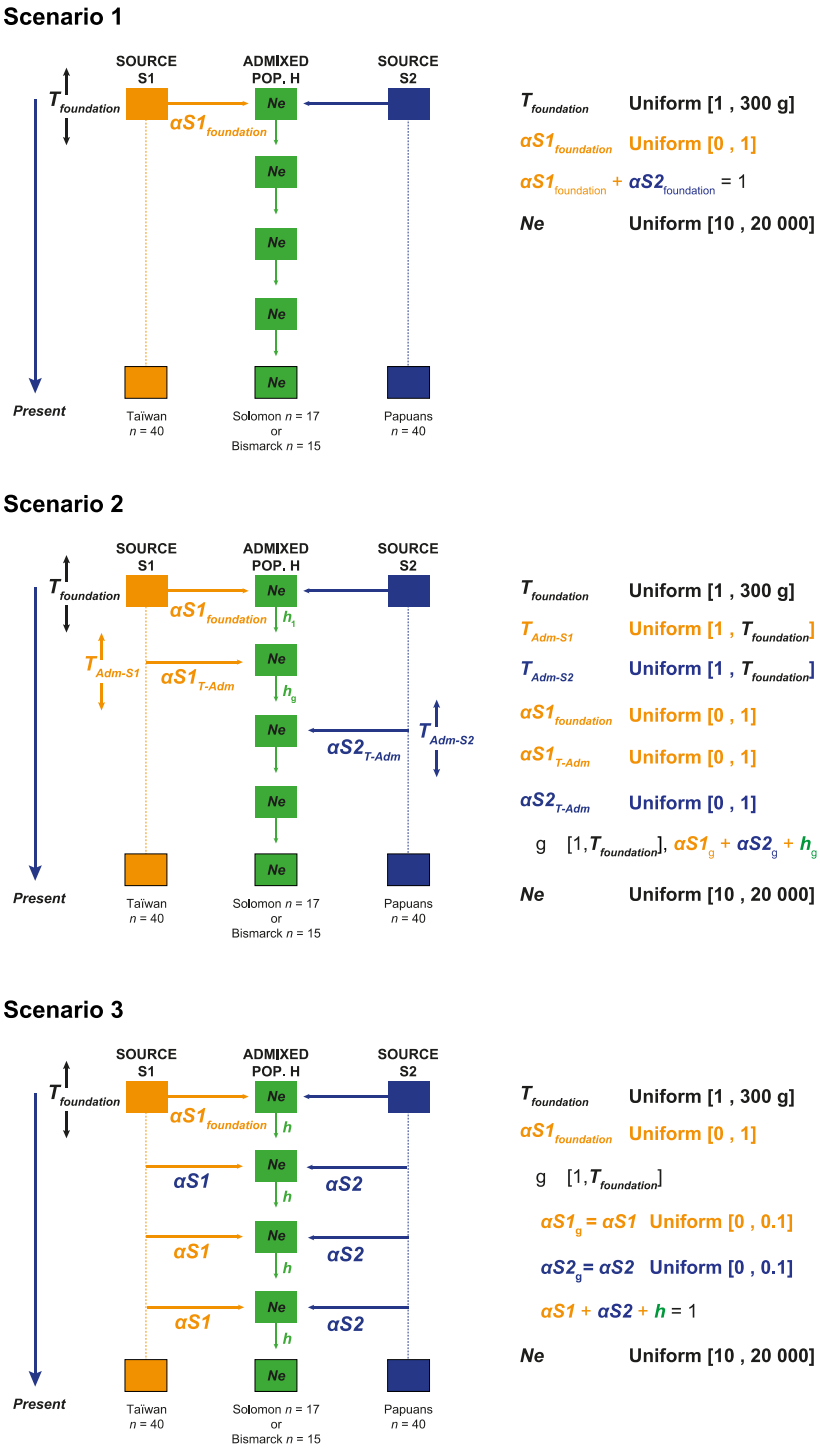
We considered three competing scenarios for the admixture history of the Bismarck Archipelago and the Solomon Islands, respectively (Supplementary Fig. 42). For all three models, we considered that (i) the admixed population H (Bismarck or Solomon islanders) is founded from an admixture event between source populations S1 (Taiwanese indigenous peoples) and S2 (PNG) occurring at time $T_{foundation}$ before present, with a proportion $\alpha S1_{foundation}$ from S1 and $1 - \alpha S1_{foundation}$ from S2; (ii) the effective population size N_e of the admixed population H is constant from $T_{foundation}$ to the present; (iii) for simplicity, both source populations are large populations at the drift-mutation equilibrium throughout the admixture process; and (iv) mutation is neglected throughout the admixture process.

Under the single admixture pulse model (Scenario 1; Supplementary Fig. 42), we considered that the admixed population H is founded from a single pulse of admixture occurring at $T_{foundation}$. No subsequent event of admixture from either S1 or S2 occurs between $T_{foundation}$ and the present. Under the two admixture pulse models (Scenario 2; Supplementary Fig. 42), we considered that the source population S1 can contribute an additional pulse of admixture to the gene pool of population H, occurring at time $T_{Adm\text{-}S1}$ with a proportion $\alpha S1_{T\text{-}Adm}$. Separately, we considered that source population S2 can also contribute an additional pulse of admixture to the gene pool of the admixed population H, occurring at time $T_{Adm\text{-}S2}$ with a proportion $\alpha S2_{T\text{-}Adm}$. Finally, under the constant-recurring admixture model (Scenario 3, Supplementary Fig. 42), we considered that, from $T_{foundation}$ to the present, source populations S1 and S2 contribute to the gene pool of population H with proportions $\alpha S1$ and $\alpha S2$, respectively, at each generation.

Prior distributions for each parameter are provided in Supplementary Fig. 42, for the three competing scenarios considered. Note that, for all three scenarios, following model definitions[75], at each generation g after $T_{foundation}$, admixture proportions $\alpha S1_g$ and $\alpha S2_g$ from source populations S1 and S2 satisfy $\alpha S1_g + \alpha S2_g = 1 - h_g$, where h_g is the contribution of the admixed population H to itself at the following generation, such that h_g is in [0,1].

We simulated 100,000 independent SNPs segregating in the two source populations until $T_{foundation}$ with *fastsimcoal2*[46], under the refined demographic model for Near Oceanians (Fig. 2a). From $T_{foundation}$ to the present, forward-in-time evolution of the 100,000 SNPs in the admixed population H was simulated with MetHis[74], under the classical Wright-Fisher model. Namely, at each generation, the two parents of each individual in the admixed population H were randomly drawn from source populations S1 and S2, and the admixed population H, with probabilities $\alpha S1_g$, $\alpha S2_g$, and $h_g = 1 - \alpha S1_g - \alpha S2_g$. At the end of each MetHis

simulation, *n* = 40 individuals were randomly drawn from source populations S1 and S2, respectively, *n* = 15 individuals from the admixed population H for the Bismarck Archipelago and *n* = 17 individuals for the Solomon Islands, as for the observed data (Supplementary Table 1). All individuals were sampled to be unrelated, by explicitly flagging individual genealogies during the last two generations of the simulations.



Supplementary Figure 42. Schematic representation of the three scenarios for the admixture history of Near Oceanians. Scenarios include the single-pulse model (Scenario 1), the two-pulse model (Scenario 2) and the constant-recurring model (Scenario 3). Prior distributions are indicated on the right.

Summary Statistics and ABC implementation

Using MetHis, we computed the following summary statistics for all subsequent ABC analyses: pairwise F_{ST}[76] between S1 and H and between S2 and H; f_3(H;S1,S2)[40], the mean and variance of the inbreeding coefficient *F* among individuals in population H, as implemented in vcftools[54]; the mean and variance of SNP-by-SNP heterozygosities in population H; the mean Allele-Sharing Dissimilarity (ASD) between S1 and H, S2 and H, and within population H, and the mean, variance, kurtosis, skewness, and mode of the distribution of the estimated admixture proportions from source population S1 across admixed individuals, as well as the minimum, maximum and all 10% percentiles of the distribution. Admixture proportions were estimated based on the individual-pairwise ASD matrix[77] calculated for the 100,000 SNPs. We projected the ASD matrix in two dimensions using Multi-Dimensional Scaling[74], and considered, as an estimate of the admixture proportions for each admixed individual, the relative distance between the individual and the centroids of the two source populations.

We considered the machine-learning ABC pipeline for scenario choice and posterior parameter estimation, as described in MetHis[74]. For ABC scenario choice, we conducted 10,000 independent simulations under each of the three competing scenarios described above (Supplementary Fig. 42). We identified the most probable scenario with the Random-Forest ABC approach[78] implemented in the *abcrf* R package, based on 30,000 simulations. For the best scenario identified, we conducted 20,000 additional simulations with MetHis. The total 30,000 simulations were then used for joint posterior parameter estimation, using the Neural-Network ABC approach implemented in the *abc* R package[79].

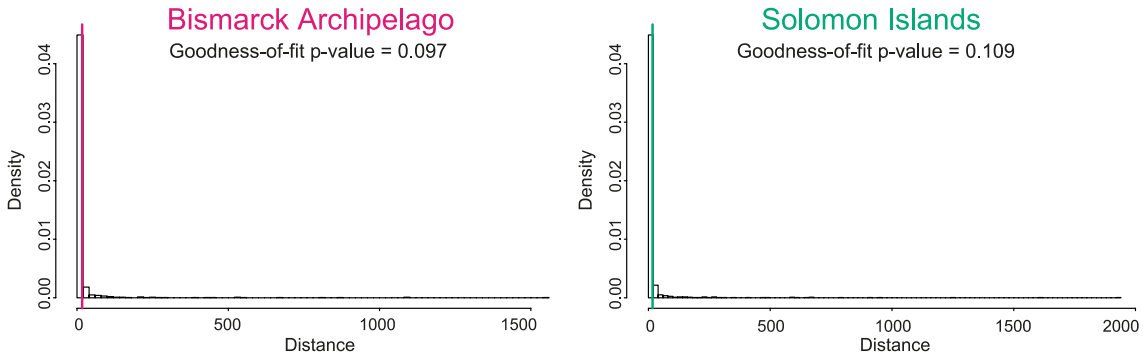
Method performance

We first plotted each prior distribution of summary statistics, and visually checked that the observed summary statistics for Bismarck and the Solomon islanders fell within the simulated distributions. We then performed a goodness-of-fit approach using the *gfit* function from the *abc* R package[79], with 100 replicates and tolerance rate set to 0.01.

To estimate the error rate of our scenario-choice approach, we used the *abcrf* function of the *abcrf* R package. Specifically, we obtained the cross-validation table and associated prior error rate, by using an out-of-bag approach, considering the same prior probability for the three competing scenarios. We performed scenario-choice prediction and estimation of posterior probabilities of the winning scenario with the *predict.abcrf* function in the same R package, using the complete simulated reference table for training the Random-Forest algorithm. We did so for the Bismarck and Solomon admixed populations separately. Both analyses were performed considering 1,000 decision trees in the forest, after visually checking that error rates converged appropriately, with the *err.abcrf* function.

To estimate the error rate of our parameter estimation approach, we first had to determine the parameters of the Neural-Network ABC approach. Indeed, there are no absolute rules for choosing the tolerance rate and number of neurons in the hidden layer most conservative to conduct posterior-parameter estimations in Neural-Network ABC[74,79,80]. To do so, we used the cross-validation procedure implemented with the *cv4abc* function from the *abc* package for tolerance rates of 10% (3000 closest simulations to the target data) or 1% (300 closest simulations to the target data), and a number of neurons in the hidden layer of the neural network ranging from 4 to 6 (one minus the number of parameters in the winning scenario) considering, in-turn and "out-of-bag", 100 random simulations as pseudo-observed target data and the remaining 29,900 simulations in the reference table under the winning scenario. For each analysis, we considered a "logit" transformation of parameters bounded by their respective prior ranges. All other neural-network parameters were left to default values. The cross-validation parameter prediction error was then calculated across the 100 separate posterior estimations for pseudo-observed datasets for each pair of tolerance rate and number of neurons, and for each parameter θ_i, as $\sum_1^{100}\left(\hat{\theta}_i - \theta_i\right)^2 / \left(100 \times Variance(\theta_i)\right)$, using the median point estimate for each parameter and the *summary.cv4abc* function in the

abc package. This allowed to compare errors for scenario parameters across Neural-Network tolerance rates and numbers of hidden neurons.



Supplementary Figure 43. Goodness-of-fit of the simulated vs. observed summary statistics used in the ABC approach. *P*-values were computed from a null distribution obtained by using simulated summary statistics as pseudo-observed summary statistics, and 100 replicates.

Results

Independently of the admixture scenario considered (Supplementary Fig. 42), the simulation scheme used for our ABC approach was able to produce vectors of summary statistics that are consistent with the observed data, for both the Bismarck and the Solomon cases (goodness-of-fit *P*-value > 0.05; Supplementary Fig. 43).

Although the different admixture models are nested for certain parts of the parameter space, the MetHis – RF-ABC framework could distinguish *a priori* among the three competing scenarios substantially more frequently than by chance (Supplementary Fig. 44a). We found a cross-validation out-of-bag prior error rate of 46.82%, compared to an expected 66.66%, and a substantial majority of votes for the correct true scenario, for every predicted scenario.

a

Predicted Scenario \ True Scenario	Scenario 1	Scenario 2	Scenario 3
Scenario 1	51.6%	26.7%	21.7%
Scenario 2	20.6%	46.3%	33.1%
Scenario 3	14.5%	23.3%	61.7%

b

	Scenario 1	Scenario 2	Scenario 3
BKA (n = 15)	319	533	148
SLI (n = 17)	296	554	150

Supplementary Figure 44. Choice of the admixture scenario for Near Oceanians by MetHis RF-ABC. **a,** Cross-validation prediction votes. **b,** Prediction votes for admixture scenarios of Near Oceanians from the Bismarck Archipelago (BKA) and the Solomon Islands (SLI) by RF-ABC.

Based on this prior analysis, we conducted separate RF-ABC scenario-choice predictions for populations of the Bismarck Archipelago and of the Solomon Islands. For both admixed populations, Scenario 2 was favoured with a large majority of the Random-Forest votes (Supplementary Fig. 44b). Furthermore, the associated posterior probabilities of Scenario 2 were 50.51% and 53.19%, for Bismarck and Solomon populations, respectively, supporting Scenario 2 as the best choice. Under the scenario with the highest posterior probability, we next estimated admixture parameters. We first tested different parameters for the Neural-network ABC approach, and showed that, *a priori*, 4 neurons in the hidden layer and a 10% tolerance rate minimized the average parameter prediction error (Supplementary Table 10). Considering these parameters, and logit transformations of all parameters bounded by their respective priors, we obtained posterior densities for each parameter, median and mean point-estimates, as well as 90% Credibility Intervals (CI), for the Bismarck Archipelago and Solomon Island populations (Supplementary Table 10). We found that the cross-validation error was relatively large for all admixture parameters, and the 90% CIs covered most of the prior distributions, suggesting that our estimations were not accurate. This may stem from the limited information contained in the summary statistics used by MetHis[74] when sample size is low, and calls for other approaches to accurately estimate admixture parameters based on other aspects of the data.

Supplementary Note 6: Dating East Asian Gene Flow

Rationale

We found that admixture patterns among Near Oceanians are more compatible with a double-pulse than a single-pulse model of admixture, using ABC[74] (Supplementary Note 5). We estimated admixture times under the double-pulse scenario, but obtained large 90% credible intervals that cover most of the parameter prior distributions, when using moments of the distribution of admixture proportions as summary statistics (Supplementary Note 5). We thus reasoned that other aspects of the genetic data should be used as summary statistics. In particular, the mathematical relationship between the length of admixture tracts and the time of admixture is well documented, although it is limited to simple admixture models[81,82]. We thus evaluated and tested another ABC approach that estimates admixture times under different admixture scenarios, using moments of the distribution of the length of admixture tracts as summary statistics.

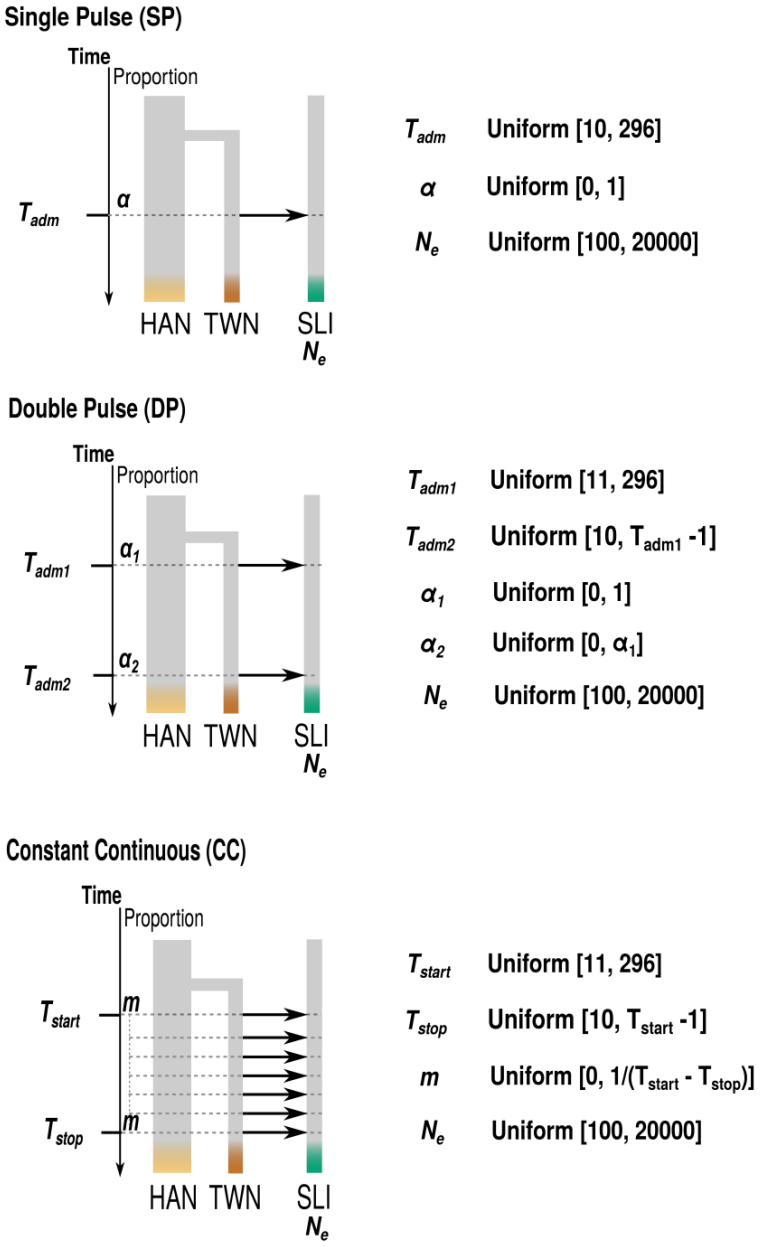
Simulation setting

We modified our refined ML demographic model for Near Oceanians (Fig. 2a) to include a single-pulse (SP), a double-pulse (DP) or a constant-continuous (CC) gene flow from East Asians into Bismarck or Solomon islanders, separately (Supplementary Fig. 45). Specifically, we assumed in the SP model that the Taiwanese indigenous peoples contributed a proportion α to the Near Oceanian population, instantaneously at time T_{adm}. In the DP model, we assumed that Taiwanese indigenous peoples contributed α_1 and α_2 admixture proportions to Near Oceanians at two different times T_{adm1} and T_{adm2}, respectively. In the CC model, Taiwanese indigenous peoples contributed to Near Oceanians with a constant rate m starting at T_{start} and stopping at T_{stop}. In the three models, we set migration rates among Near Oceanians and between Near Oceanians and other populations to zero, one generation before T_{adm} (SP model), T_{adm1} (DP model), or T_{start}. (CC model).

The prior distributions for each model parameter are described in Supplementary Fig. 45. For the DP model, the time of the second pulse T_{adm2} was sampled from a uniform distribution dependent on the sampled value of the time of the first pulse T_{adm1}, so that $T_{adm1} > T_{adm2}$. Likewise, for the CC model, T_{stop} was sampled so that $T_{start} > T_{stop}$. Because our main goal was to estimate the time of gene flow, and because we aimed to assess the effect of admixture proportions without inferring them, we decided to use the parameters α, α_1, α_2 and m as nuisance parameters for each model. The effective population size of the recipient population (N_e) was also considered as a nuisance parameter. For each simulation, we simulated 100 5-Mb independent DNA loci with *fastsimcoal2*[46], assuming a variable recombination rate sampled from the 1000 Genomes Phase 3 genetic map[43].

Summary statistics and implementation

Based on previous work[81,82], we used, as ABC summary statistics, moments of the distribution of the length of admixture tracts across Near Oceanian individuals. Namely, we computed, for each observed or simulated 5-Mb genomic region, the mean, minimum and maximum of the length of admixture tracts across individuals. We also computed the mean and the variance, across genomic regions, of these three summary statistics. The six resulting statistics were obtained from local ancestry inference with RFMix v1.5.4[83]. RFMix was run with 3 Expectation-Maximization (EM) steps, a window of 0.03 cM, and Taiwanese indigenous peoples and PNG as source populations, as for the observed data (Supplementary Note 17). Summary statistics were computed with custom R scripts. All the ABC analyses were performed using functions of the *abc* R package[79]. For model choice, we performed 5,000 simulations under each alternative model, and used the logistic multinomial regression method implemented in the *postpr* function and a 5% tolerance rate. For parameter estimation, we performed 10,000 additional simulations under the most probable model, and used the Neural network method implemented in the *abc* function, using default numbers of hidden layers and neurons and a 1% tolerance rate.



Supplementary Figure 45. Three models of East Asian-related gene flow into Near Oceania, considering Solomon islanders (SLI) as the recipient population. The same models were used for a Bismarck Archipelago population as the recipient population. Prior parameter distributions for the haplotype-based ABC approach are shown on the right.

Method performance

To check *a priori* if simulations generally reproduced the observed data, we first checked whether the summary statistics for the observed data were in the boundaries of those for the simulated data. We sampled 100 5-Mb genomic windows in the genomes of individuals from the Bismarck Archipelago or the Solomon Islands, and computed the mean and variance of the 100 observed summary statistics. We then compared observed means and variances to means and variances of summary statistics computed for 100 simulated 5-Mb DNA loci.

To estimate the performance of model selection by ABC, we used a “leave-one-out” cross validation procedure: for each gene flow model, a simulation was selected as a validation simulation, while the rest were used as training simulations, 100 times.

To estimate the accuracy of parameter estimation by ABC, we performed a “leave-one-out” cross-validation analysis and an accuracy test, to confirm that simulated parameter values were correctly estimated. Accuracy indices were computed as follows:

Prediction error $PE = \frac{\frac{1}{S}\sum_{i=1}^{S}\left(\hat{\theta}_i-\theta_i\right)^2}{var(\theta_i)}$

Relative estimation bias $rEB = \frac{1}{S}\sum_{i=1}^{S}\frac{\left(\hat{\theta}_i-\theta_i\right)^2}{\theta_i}$

95% credible interval $95\%COV = \frac{1}{S}\sum_{i=1}^{S}1(q_1 < \theta_i < q_2)$

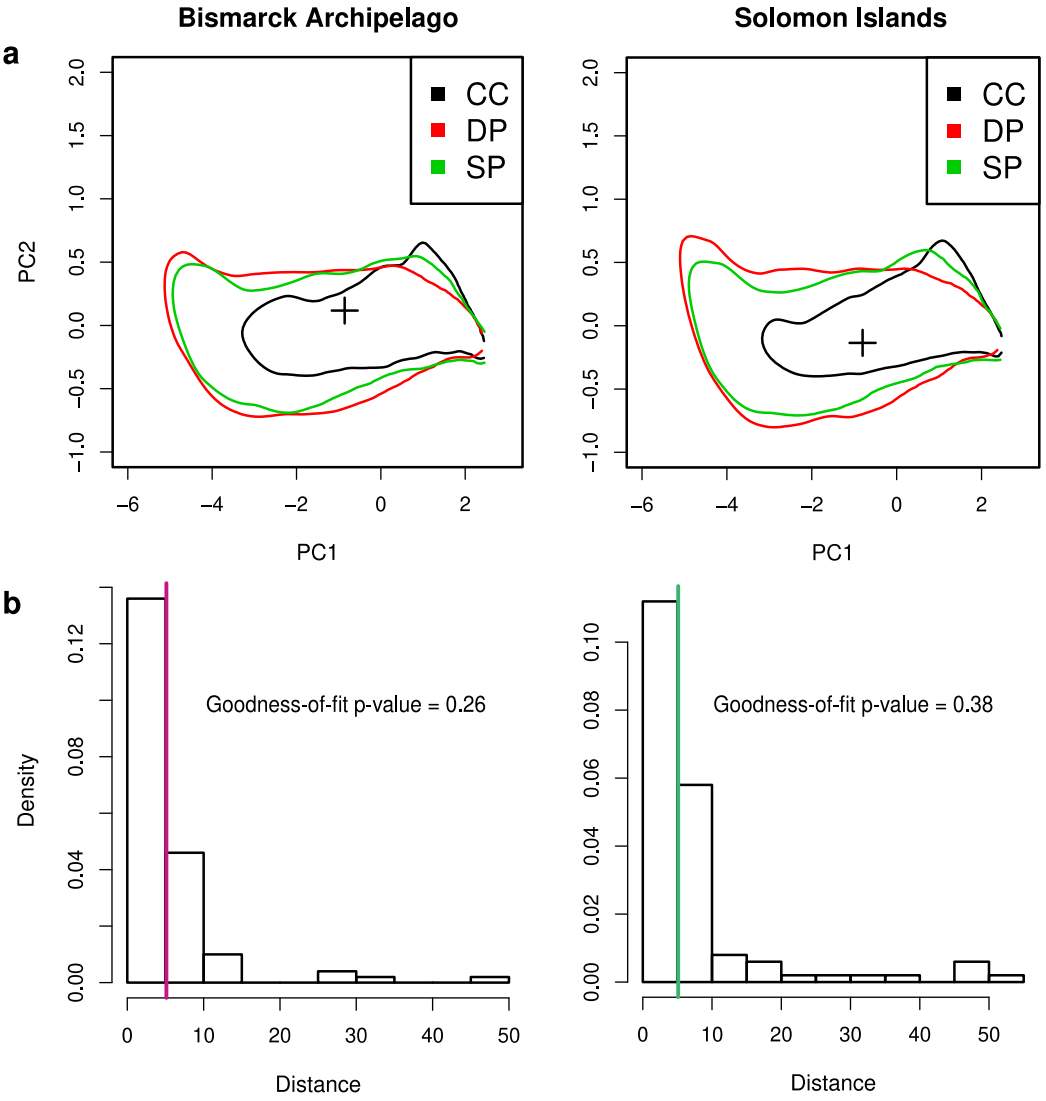
where $\theta_i$and $\hat{\theta}_i$ are the true and ABC estimated values of parameter θ for the i^{th}simulated dataset, S is the number of simulated data, 1(*C*) the indicative function (equal to 1 when *C* is true, 0 otherwise) and q_1 and q_2 the respective 0.025 and 0.975 quantiles. These accuracy indices were computed using S = 300 simulated data. Finally, we performed posterior predictive checks by re-simulating 1,000 datasets of 100 5-Mb regions, using parameter estimates sampled from the 95% percentile of their approximate posterior distribution. Nuisance parameters (i.e., N_e, α) were sampled from uniform prior distributions. We then compared simulated to observed summary statistics.

Results

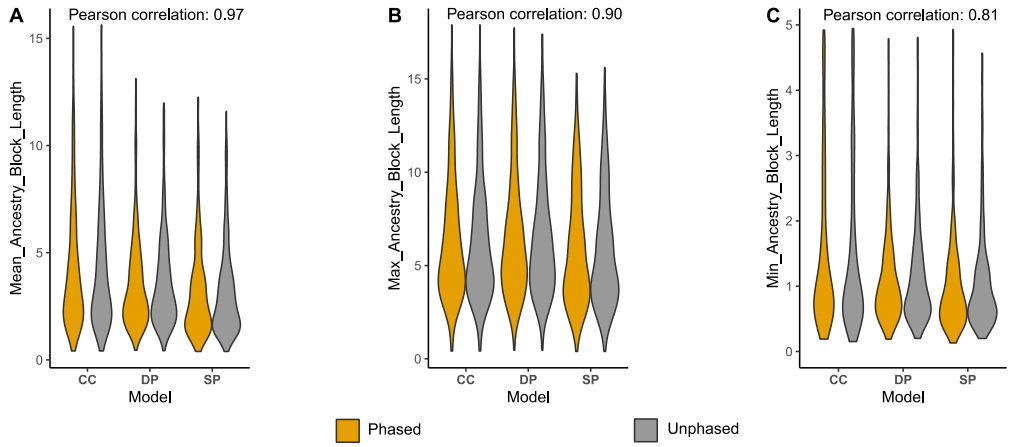
We first checked whether our simulations generally reproduce the observed data, and found that they were in good agreement for the six summary statistics used, based on the length of admixture tracts detected in Near Oceanians (Supplementary Fig. 46). We also checked that our summary statistics were not sensitive to phasing errors. To do so, we compared statistics computed from simulations where the phase was known, to the same statistics computed from the same simulations, but where the data was unphased and then phased with SHAPEIT2[84,85], under the same conditions as in the observed data. Summary statistics were generally unchanged (Supplementary Fig. 47).

Based on cross validation, we estimated the probability to correctly choose the SP, DP and CC models, and found that the error in model choice was minimal for the DP model and maximal for the SP model (Supplementary Fig. 48a,b). Then, to identify the most probable gene flow model for Near Oceanians, we compared the observed tract length distributions against simulations under the three competing scenarios. In agreement with the MetHis ABC approach[74], which is based on other aspects of the data (Supplementary Note 5), we found that summary statistics for the Bismarck and Solomon islanders were closest to those under the DP model (Supplementary Fig. 48c,d). Taken together, these results support two separate epochs of gene flow from East Asian-related populations into Near Oceanians, in both the Bismarck Archipelago and the Solomon Islands.

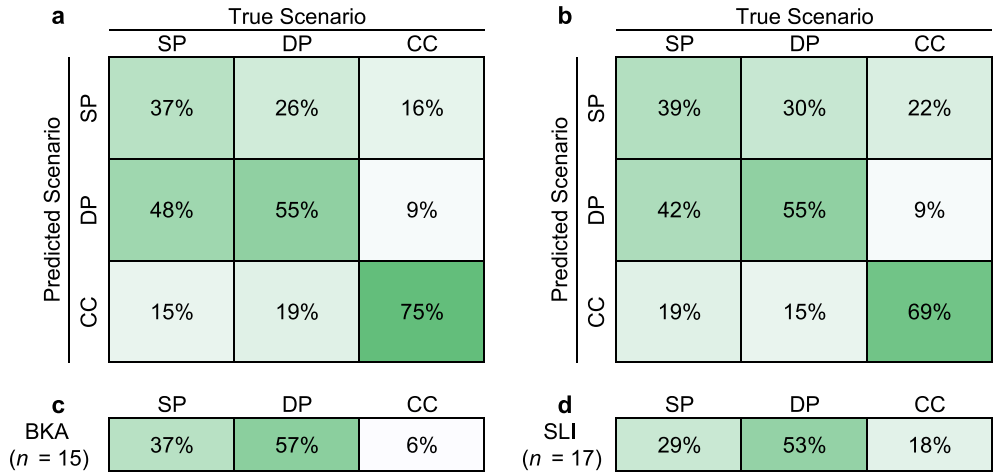
We next estimated the performance of our ABC method to estimate the times of the two gene flow pulses in Near Oceanians. We found that the time of gene flow pulses was more accurately estimated for recent times (up to ~100 generations) (Supplementary Fig. 49). The estimation of the time of the oldest pulse T_{adm1} was generally more accurate than T_{adm2}. Nevertheless, we observed a low prediction error and low positive relative biases of T_{adm1} and T_{adm2} for both the Bismarck and Solomon Archipelagos.



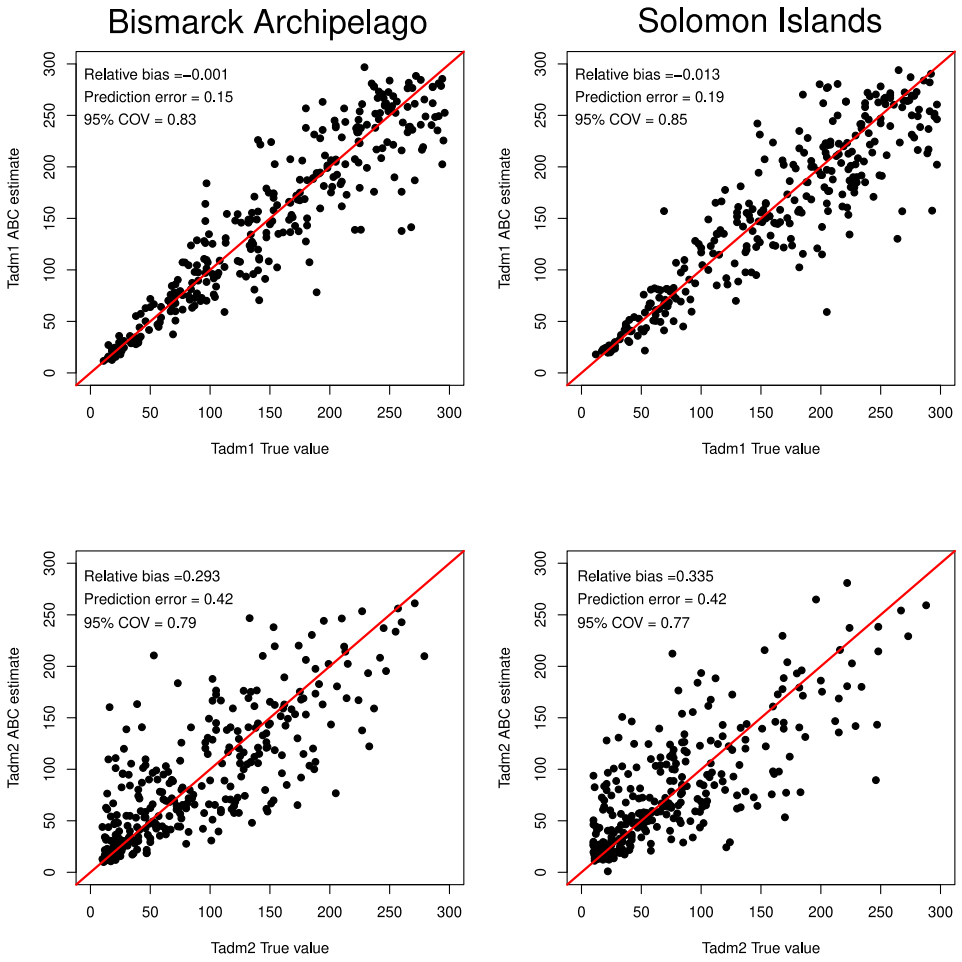
Supplementary Figure 46. A priori check of the summary statistics used to estimate the times of admixture in Near Oceanians by ABC. **a,** PCA of the six ABC summary statistics obtained for the three simulated models of gene flow (90% coloured contours) and the observed data (black plus sign). SP, DP and CC indicate single-pulse, double-pulse, and constant-continuous models of gene flow. **b,** Goodness-of-fit of the simulated models of gene flow with the observed summary statistics. *P*-values were computed from a null distribution obtained by using simulated summary statistics as pseudo-observed summary statistics, and 100 replicates.



Supplementary Figure 47. Limited effects of haplotype phasing on admixture tract length statistics. The mean, maximum and minimum length of admixture tracts were computed on simulations where haplotype phase is known (in orange) and on the same simulations, but where haplotypes were unphased and reconstructed with SHAPEIT2 (in grey).

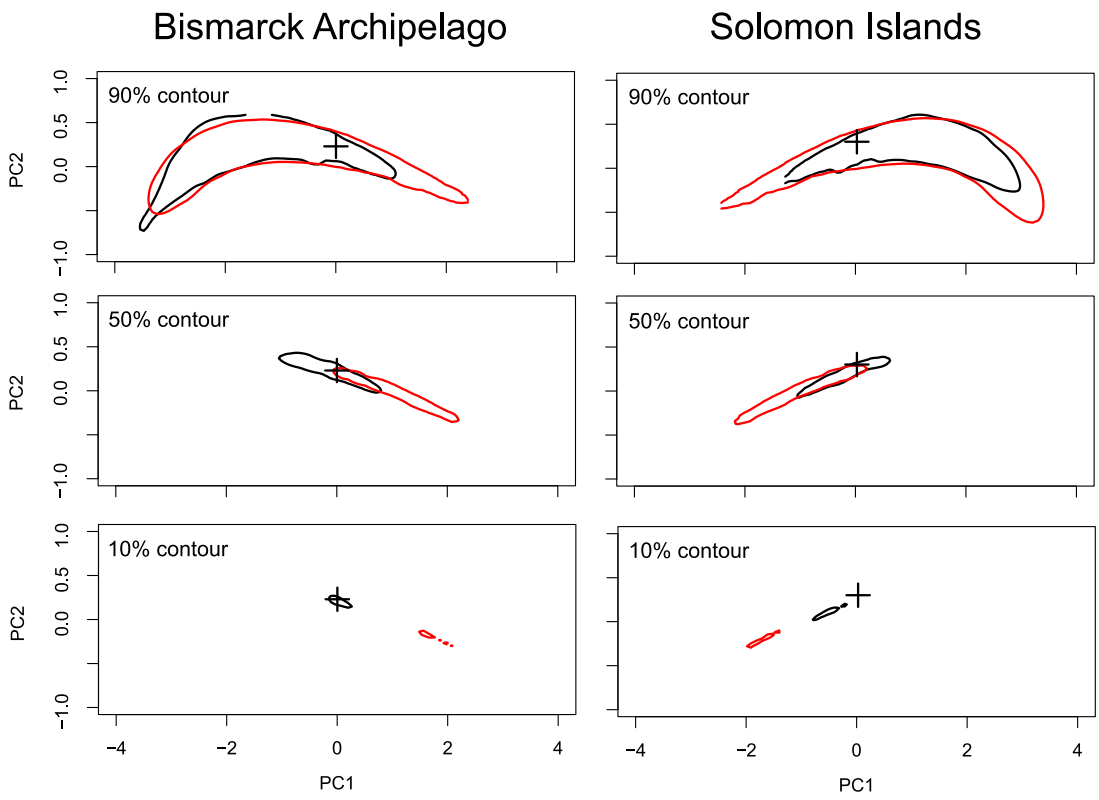


Supplementary Figure 48. Choice of the gene flow model for Near Oceanians by ABC based on admixture tract length. **a, b,** Confusion matrix for the choice of gene flow models by ABC for **a,** Bismarck Archipelago and **b,** Solomon Islands. **c, d,** Approximate posterior probabilities of the three competing gene flow models for **c,** Bismarck Archipelago (BKA) and **d,** Solomon Islands (SLI).



Supplementary Figure 49. Performance of the ABC estimation of the times of gene flow T_{adm1} and T_{adm2} in the double-pulse gene flow model, for the Bismarck Archipelago and the Solomon Islands.

Finally, we estimated the time of the two gene flow pulses in Near Oceanians, under the most probable DP model (Supplementary Figs. 44 and 48c,d). Assuming a 29-year generation time, the time of the oldest pulse was dated to ~2.3 ka in Near Oceanians (2.2 [95% CI: 1.7–3.0] and 2.5 [95% CI: 2.2–3.4] ka for Bismarck and Solomon islanders, respectively; Fig. 2c), following the emergence of the Lapita cultural complex in the region ~3.5 ka[67]. The time of the most recent pulse was estimated to ~1.4 ka (1.4 [95% CI: 0.4-2.0] and 1.3 [95% CI: 0.7-2.0] ka for Bismarck and Solomon islanders, respectively). Posterior predictive checks further confirmed that the estimates of the times of admixture were in good agreement with the observed data (Supplementary Fig. 50). These results collectively suggest recurrent genetic interactions between East Asian-related populations and the ancestors of present-day Near Oceanians, and support that the admixture events followed the Lapita period, in agreement with the Austronesian origin of this cultural complex[10].



Supplementary Figure 50. Posterior predictive checks of the double-pulse model for populations of the Bismarck Archipelago and the Solomon Islands. PCA of summary statistics for the observed data (black '+' sign) and simulated data (90%, 50%, and 10% contours), using prior (red) or posterior (black) distributions of estimated parameters.

Supplementary Note 7: Estimating Levels of Archaic Introgression

For all the analyses presented in this section, we used the dataset merged with the high-coverage genomes of Vindija and Altai Neanderthals[56,59], and that of the Altai Denisovan[60], filtered at *Level 3b'* (Supplementary Note 2).

Projected Principal Component Analysis

Methods. To assess the relationship between modern humans and archaic hominins, we computed a PCA on the chimpanzee, Vindija Neanderthal, Altai Neanderthal and Altai Denisovan genomes, and projected modern human samples onto the plane defined by the first two principal components. The PCA was carried out using the 'SmartPCA' algorithm implemented in EINGENSOFT program version 7.2.1 (ref.[28]).

Results. All modern human samples were located at the centre of the PCA plot (Supplementary Fig. 51a). When zooming into the central portion of the projected PCA plot, modern human populations separate into different clusters, relative to the chimpanzee and archaic hominins. The first PC (explaining 62% of the variance) separated Africans from non-Africans, and showed that non-Africans have a greater affinity towards Neanderthal and Denisovan. The second PC (explaining 32% of the variance) separated the Altai and Vindija Neanderthals from the Altai Denisovan. The second PC revealed a clear genetic affinity of Eurasians (East Asians and West Eurasians) towards Neanderthals, and Pacific populations towards Denisovan. Notably, there is a clear cline of Denisovan-related ancestry in Near and Remote Oceanians, as well as the Agta, and to a lesser extent the Cebuano population, from the Philippines.

***D*- and f_4-ratio statistics**

Methods. To formally assess introgression between archaic hominins and modern humans, we computed *D*-statistics[40]. The ancestral state for any given site was defined as the allele present in the chimpanzee reference genome[19]. Sites that were not present in the chimpanzee genome, or that contained alleles that did not match either the reference or alternative allele in the chimpanzee genome, were discarded, leaving a total of 13,027,305 bi-allelic SNPs for further analysis.

To test for introgression between Neanderthal and modern humans, we computed a *D*-statistic of the form *D(X, West Eurasians/East Asians/Africans; Vindija Neanderthal, Chimpanzee).* This statistic measures if a target population *X* shares more derived alleles with the Vindija Neanderthal compared to West Eurasians, East Asians or Africans. We computed a second *D*-statistic of the form *D(X, West Eurasians/East Asians/Africans; Vindija Neanderthal, Denisova)* that measures derived allele sharing with the Vindija Neanderthal or Altai Denisovan compared to West Eurasians, East Asians, or Africans.

Likewise, to formally assess introgression between Denisovan and modern humans, we computed a *D*-statistic of the form *D(X, West Eurasia/East Asia; Denisova, Chimpanzee)*. Similarly, we computed a second *D*-statistic of the form *D(X, West Eurasia/East Asia; Altai Denisovan, Vindija Neanderthal).* We considered populations showing significant allele sharing (|Z-scores| > 2) as evidence of Neanderthal or Denisovan introgression.

To estimate the genome-wide proportion of Neanderthal ancestry for a target population X, we used the following f_4-ratio statistic:

$$P_N(X) = \frac{f_4(Chimpanzee, Neanderthal\ Altai, Africans, X)}{f_4(Chimpanzee, Neanderthal\ Altai; Africans, Neanderthal\ Vindija)}$$

However, this statistic can be inflated by unaccounted Denisovan ancestry. To circumvent this, we repeated the analysis by focusing only on sites where the Denisovan genome is homozygous ancestral as in ref.[56]. This additional filter removed around 10% of the sites.

Similarly, to estimate the genome-wide proportion of Denisovan ancestry for a target population X, we used the f_4-ratio statistic of the following form:

$$P_D(X) = \frac{f_4(Africans, Neanderthal\ Vindija;\ East\ Asians, X)}{f_4(Africans, Neanderthal\ Vindija; East\ Asians, Denisovan)}$$

This f_4-ratio statistic can correctly infer genome-wide proportions of Denisovan ancestry in Oceanians, using East Asians to correct for the levels of Neanderthal ancestry in Oceanians (see ref.[86]).

It has been previously proposed that the Denisovan ancestry in Oceanians was acquired from the ancestors of PNG[5,86], given that the amount of Papuan-related ancestry in these populations is highly correlated with their Denisovan ancestry. We tested this hypothesis, by estimating the amount of Denisovan ancestry as a fraction of Papuan-related ancestry using the following f_4-ratio statistic:

$$P_{DasP}(X) = \frac{f_4(Africans, Denisovan;\ East\ Asians, X)}{f_4(Africans, Denisovan; East\ Asians, Papuans)}$$

To estimate the amount of PNG ancestry, we used the following f_4-ratio statistic:

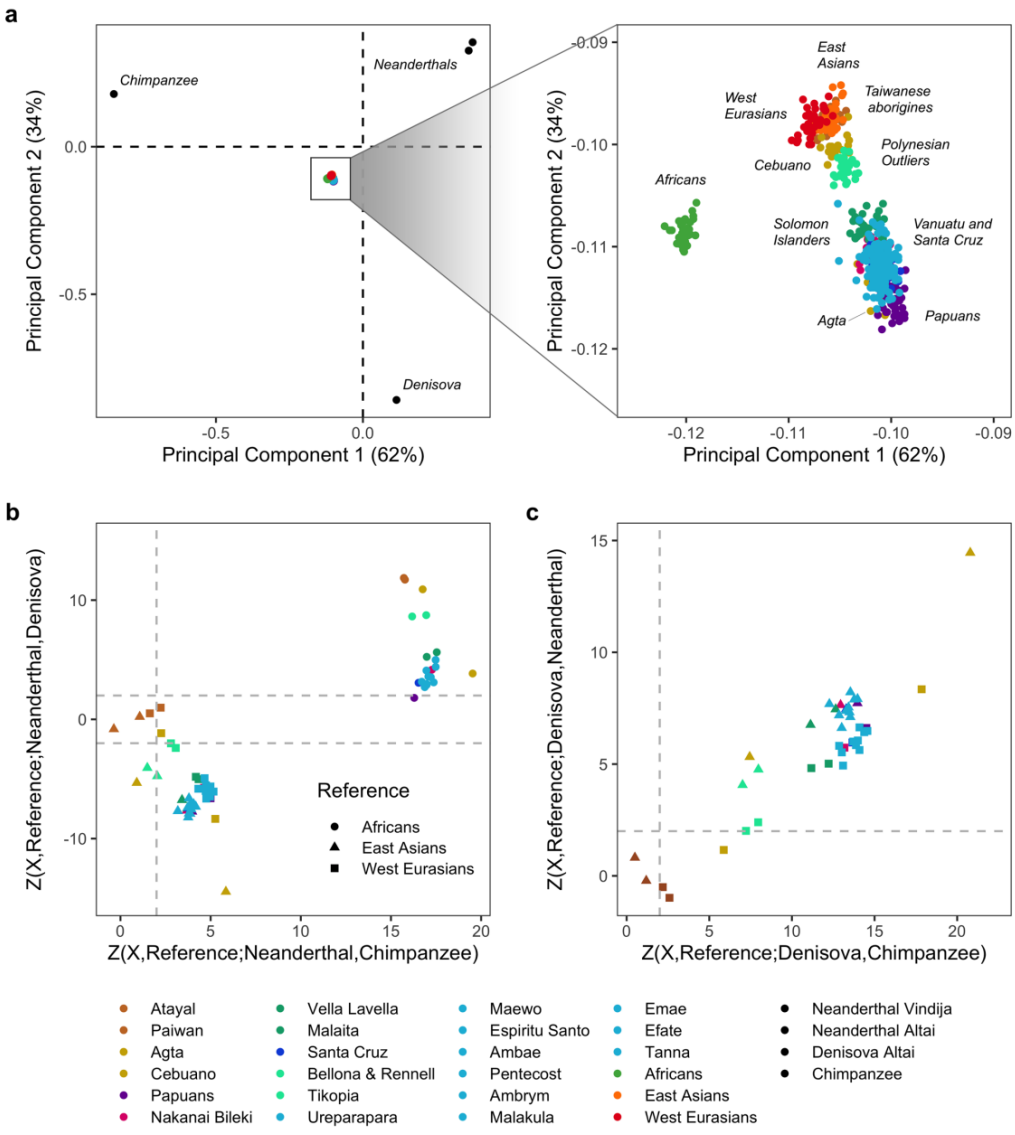
$$P_P(X) = 1 - \frac{f_4(Africans, Australians;\ X, Papuans)}{f_4(Africans, Australians; East\ Asians, Papuans)}$$

For these analyses, we considered all African, West Eurasian, and East Asian individuals from the SGDP (see Table S1 in ref.[17]). All *D*- and f_4-ratio statistics were computed using 'qpDstat' and 'qpF4ratio' algorithms implemented in ADMIXTOOLS version 5.1.1[40], respectively. A weighted-block jackknife procedure that drops 5-cM blocks of the genome in each run was used to compute standard errors. To assess the correlation between Papuan-related ancestry and Denisovan ancestry as a fraction of the Papuan-related ancestry, we fitted a linear regression model using ordinary least squares using R version 3.4.4.

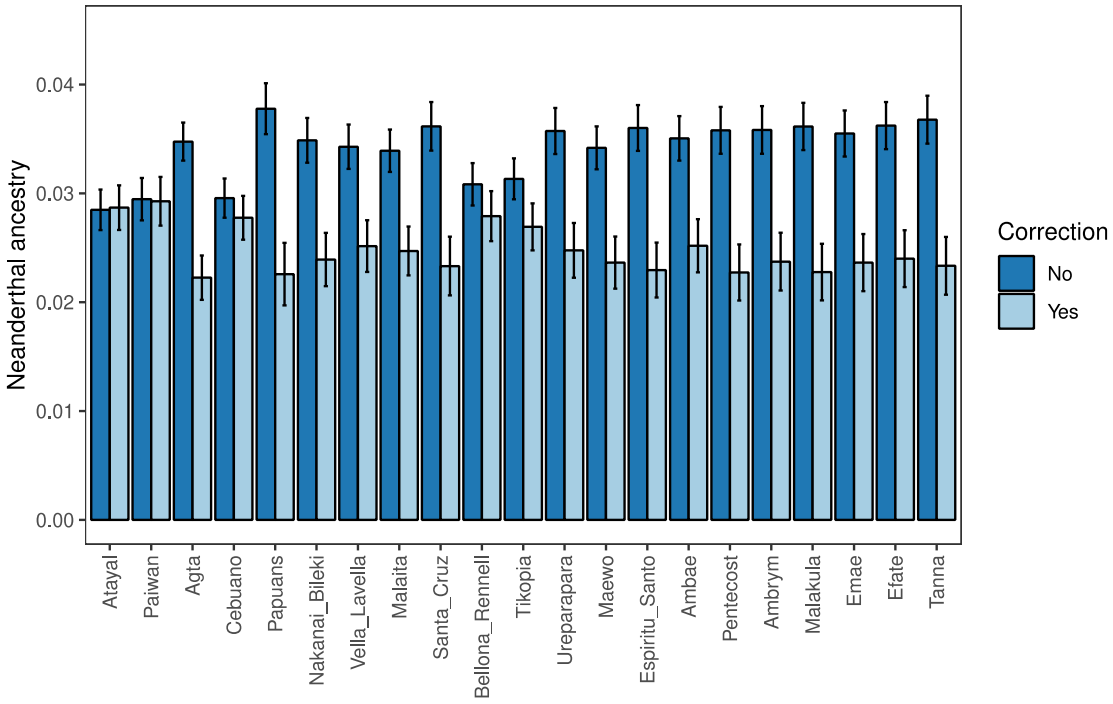
Results. As expected, *D*-statistics showed that all Pacific populations share more derived alleles with the Vindija Neanderthal, compared to Africans (x-axis *Z*-score > 15 for all comparisons; Supplementary Fig. 51b). Likewise, all populations, with the exception of the Atayal population from Taiwan, share more derived alleles with Neanderthals compared to West Eurasians (x-axis *Z*-score > 2), as previously reported[87-90]. We detected higher derived allele sharing with the Vindija Neanderthal in the Agta as well as in Near and Remote Oceanians compared to East Asians (x-axis Z-score >2). However, this was driven by higher allele sharing with the Denisovan with respect to the Vindija Neanderthal (y-axis Z-score < -2). For Denisovan ancestry, *D*-statistics showed that, with the exception of the Taiwanese Atayal and Paiwan, all populations share more derived alleles with the Altai Denisovan (x-axis Z-score > 5), and that this was not driven by higher derived allele sharing with the Vindija Neanderthal (y-axis Z-score > 2; Supplementary Fig. 51c). This was most apparent when using West Eurasians as reference populations, which have virtually no Denisovan ancestry[17].

The estimated genome-wide Neanderthal ancestry levels varied between 2.8% and 3.8% across Pacific populations, by using f_4-ratio statistics (Supplementary Fig. 52). However, after restricting the analysis to Denisovan ancestral homozygous sites, Neanderthal ancestry estimates were significantly lower in Near and Remote Oceanian populations (ranging from 2.2% to 2.8%), but differed minimally (<0.01%) in East Asian-related populations (e.g. Atayal and Paiwan Taiwanese indigenous peoples), who are expected to have low levels of Denisovan ancestry (Supplementary Fig. 52). Overall, we found that Neanderthal ancestry is homogenously distributed across Pacific populations, with values ranging from 2.2% to 2.9%

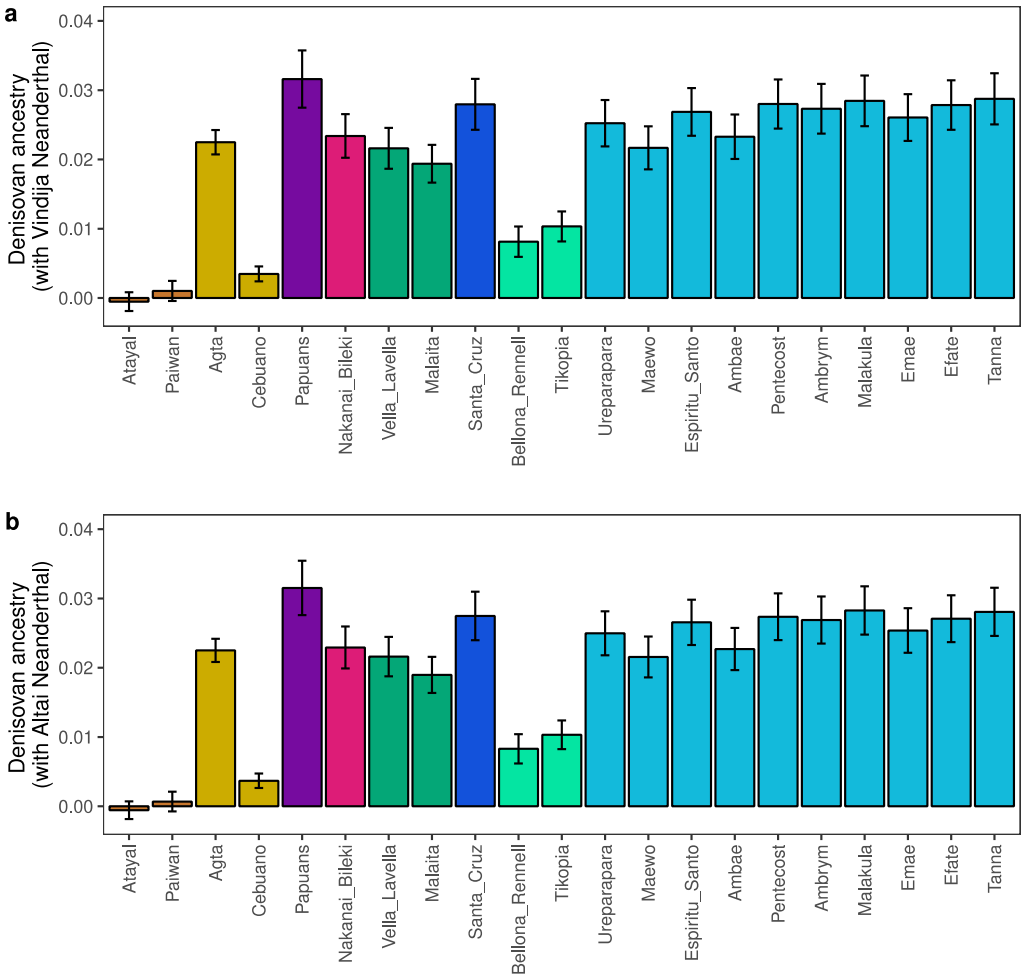
(Fig. 3a). The highest estimates were found in East Asian populations, such as the Atayal and Paiwan from Taiwan, in accordance with previous reports[87-90]. Conversely, Denisovan ancestry was heterogeneously distributed, with values ranging from 0% to 3.2% (Fig. 3b, Supplementary Fig. 53), and was maximal in Near and Remote Oceanians, and the Philippine Agta. The use of the Altai Neanderthal as a sister group to the Altai Denisovan, instead of the Vindija Neanderthal, yielded similar f_4-ratio estimates (Fig. 3b, Supplementary Fig. 53).



Supplementary Figure 51. Genetic affinities of modern humans to archaic hominins. **a**, Principal Component Analysis (PCA) of modern human populations projected onto the first two PCs defined by the chimpanzee, the Altai Neanderthal, the Vindija Neanderthal, and the Altai Denisovan genomes. The right panel represents a zoomed-in version of the PCA plot on the left. **b,** Derived allele sharing of Pacific populations to the Vindija Neanderthal. *Z*-score of a *D*-statistic of the form *D(X, East Asia/West Eurasia/Africa; Neanderthal, Chimpanzee)* is shown against *Z*-score of *D(X, East Asia/West Eurasia/Africa; Neanderthal, Denisova)*. **c,** Derived allele sharing of Pacific populations to Altai Denisovan. *Z*-score of a *D*-statistic of the form *D(X, East Asia/West Eurasia; Denisova, Chimpanzee)* is shown against *Z*-score of *D(X, East Asia/West Eurasia; Denisova, Neanderthal)*. Dotted lines indicate significant derived allele sharing (|*Z*-score| >2). Population sample sizes are reported in Supplementary Table 1.



Supplementary Figure 52. Genome-wide levels of Neanderthal ancestry when accounting, or not, for Denisovan ancestry. Levels of Neanderthal ancestry were estimated via the f_4-ratio statistic. Dark blue bars indicate Neanderthal ancestry estimates using all sites, whereas light blue bars, estimates after restricting sites to those where the Altai Denisovan is homozygous ancestral. The estimates in populations known to carry high levels of Denisovan ancestry (Near and Remote Oceanians) are significantly lower after the correction. Error bars represent 2 standard deviations from the point estimate computed via a weighted-block jackknife procedure. Population sample sizes can be found in Supplementary Table 1.



Supplementary Figure 53. Genome-wide levels of Denisovan ancestry. Levels of Denisovan ancestry estimated via f_4-ratio statistic using **a**, the Vindija Neanderthal or **b**, Altai Neanderthal as sister group to the Altai Denisovan. Error bars represent 2 standard deviations from the point estimate computed via a weighted-block jackknife procedure. Population sample sizes are reported in Supplementary Table 1.

Supplementary Note 8: Detecting Introgressed Archaic Haplotypes

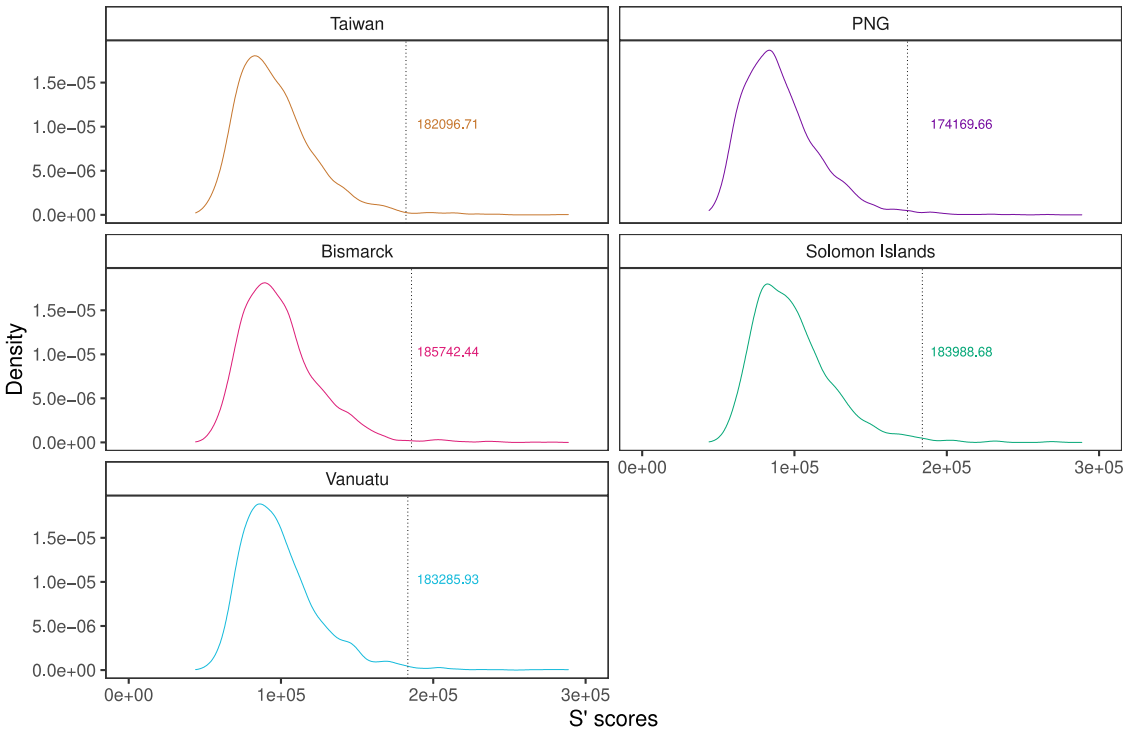
To identify archaic sequences introgressed into modern human genomes, we used two statistical methods that have been shown to be powerful in this regard[71,87,91].

S' reference-free method

Methods. We first used a recently developed method, S-prime (S'), which seeks to identify introgressed sequences from archaic hominins without using an archaic reference genome[71]. The S′ method has been shown to have increased power with respect to other archaic reference-free methods and is suitable for large-scale genome-wide data. S′ is designed to detect divergent haplotypes whose variants are in strong LD and that are absent (or at very low frequency) in a population that has not received introgression (i.e., the outgroup). To identify S′ introgressed segments in genomes of Pacific populations, we considered only variants with allele frequency < 1% in the 35 Africans from the SGDP dataset[17]. As S′ is an archaic reference-free method, we did not apply the merged Vindija Neanderthal, Altai Neanderthal, and Altai Denisovan mask filters, but otherwise kept sites that passed all filters at *Level 3b* (Supplementary Note 2). Note, however, that when comparing S′ introgressed haplotypes with an archaic genome, we did apply these masks (see section below). To estimate genetic distances between sites, we used the 1000 Genomes Phase 3 genetic map[43]. To avoid potential effects of population structure, we performed our analysis separately by population. However, due to the small sample size of SGDP populations[17], we combined all East Asian samples (excluding Taiwanese indigenous peoples) as well as all West Eurasian samples, and considered them as two different population groups. As the S′ approach does not allow for missing data, we further filtered sites with at least one missing genotype, leaving a total of 26,734,553 bi-allelic SNPs.

After retrieving empirical S′ scores from our modern human genomes, we used simulations to estimate our false positive rate (FPR) to detect S′ introgressed haplotypes. We estimated a null distribution of S′ scores by simulating genomic sequence data using the coalescent-based simulation software *fastsimcoal2*[46]. We used the demographic model for western Remote Oceanians (Extended Data Fig. 2b, Supplementary Table 5) with parameters fixed to ML point estimates, except that we removed all archaic introgression pulses (i.e. Neanderthal and Denisovan). Using this demographic model, we extracted a sample of 20 individuals from each of the populations representing East Asians, Taiwanese indigenous peoples, PNG, Bismarck and Vanuatu islanders, and 35 individuals from the population representing Africans. We used a sample size of 35 African individuals, as all S′ analyses were conducted using 35 Africans from the SGDP as outgroup population. Note that this demographic model is a null demographic model (i.e., without archaic introgression) for all the analysed populations in this study. The null S′ distribution was obtained from simulations of 2500 independent sets of 10-Mb genomic regions.

Results. We observed that the S′ statistic is highly robust to different demographic scenarios, as attested by the S′ score distributions that are very similar across populations (Supplementary Fig. 54). The highest estimated 99th percentile of the simulated S′ scores across populations was 185,742, which was found in the simulated population representing Bismarck islanders. We therefore decided to use a conservative S′ score of 190,000 to identify significantly introgressed haplotypes, which would be equivalent to a FPR < 0.01.



Supplementary Figure 54. S′ score distribution under a demographic model without archaic introgression. We computed a null distribution of S′ scores by simulating 2,500 independent sets of 10-Mb genomic regions for five different populations. Details of the ML demographic model are described in Supplementary Note 4 (Extended Data Fig. 2b, Supplementary Table 5). The dotted line indicates the 99th percentile of the S′ distribution, with the corresponding value shown in each panel.

Conditional Random Fields method

Methods. We applied a method based on Conditional Random Fields (CRF) to identify introgressed archaic haplotypes in our phased genomic data[87,91]. In contrast to S′, which relies on simulations to determine the significance of introgressed segments, the CRF method is able to incorporate the parametric assumption directly into a probabilistic framework. The CRF method uses information of an outgroup population (i.e. a population that did not experience archaic introgression), archaic genomes (Neanderthal or Denisovan), and genomes from a population that harbours introgressed sequences. Under this framework, each site along the genome is included as a random variable with two states: introgressed or non-introgressed (thus of modern human origin). Emission probabilities that incorporate different genomic features of a tested haplotype are used to evaluate whether a particular site has a higher probability of being of archaic or modern human origin. CRF inferences require estimating model parameters, which were fixed for the values previously estimated[87].

To estimate the genetic distance between sites, we used the 1000 Genomes Phase 3 genetic map[43], as for S′ analyses. The ancestral state for any given site was defined as the allele present in the chimpanzee reference genome[19]. Sites that were not present in the chimpanzee genome, or that contained alleles that did not match either the reference or alternative alleles in the chimpanzee genome, were discarded. We phased the data using SHAPEIT2[84,85] with 200 conditioning states, 10 burn-in steps and 50 MCMC main steps, for a window length of 0.5 cM and an effective population size of 15,000. Missing sites below the 5% threshold were imputed during the phasing. We did not allow for missing sites in the Neanderthal and Denisovan genomes before phasing. After phasing, a total of 18,949,412 bi-allelic SNPs were used for further analysis.

Following ref.[91], we inferred archaic ancestry in two steps:

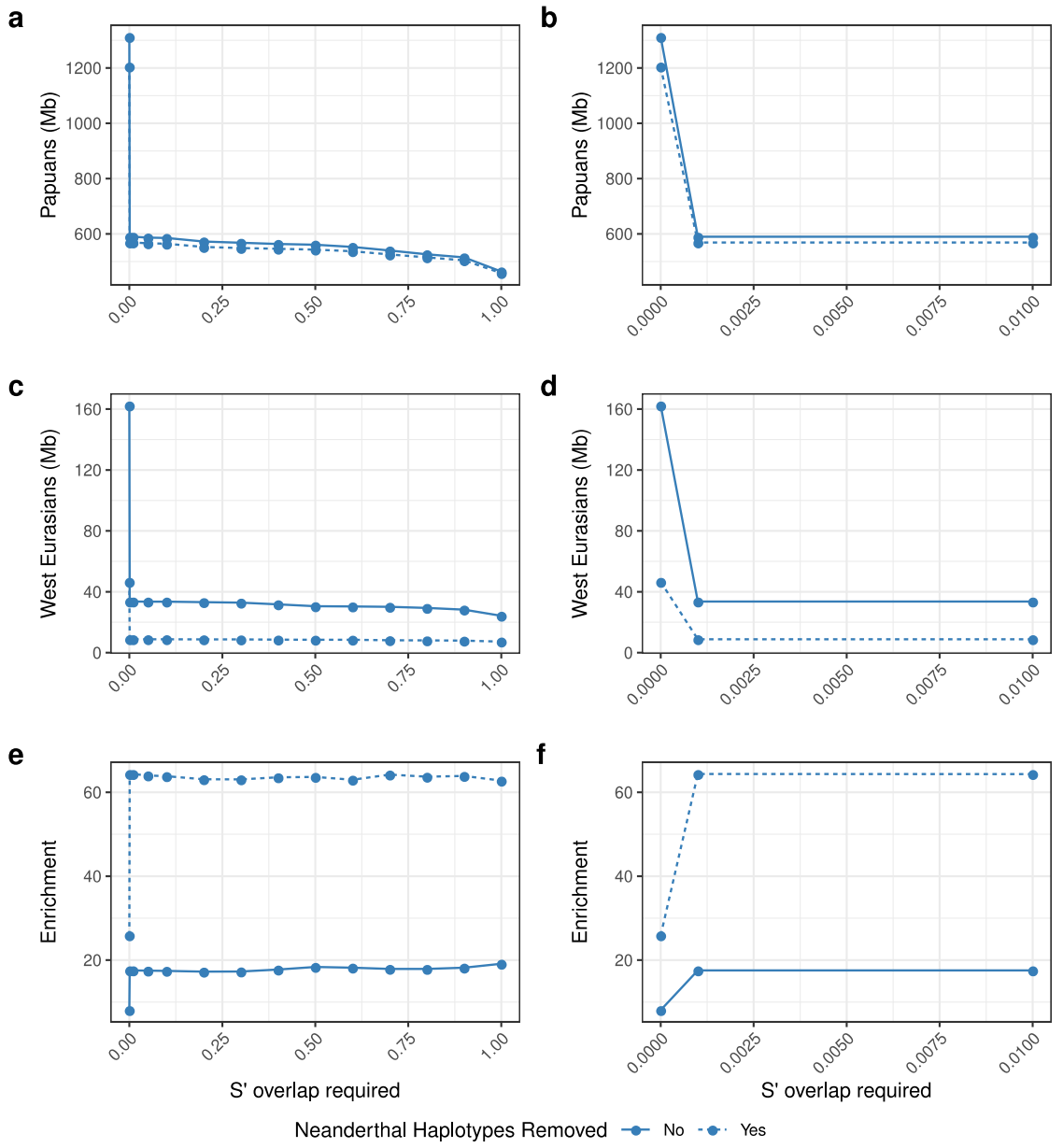
1. To infer Neanderthal ancestry, one reference panel consisted of the Vindija Neanderthal genome, while the other consisted of 35 Africans from SGDP[17] merged with the Altai Denisovan genome.
2. To infer Denisovan ancestry, one reference panel consisted of the Denisovan genome, while the other consisted of 35 Africans from SGDP[17] merged with the Vindija Neanderthal genome.

Note that the CRF method was run independently to infer Denisovan and Neanderthal haplotypes. However, given that Denisovans and Neanderthals share a more recent common ancestor than with any modern human population, there is a probability of the same introgressed segment in a particular modern human haplotype to be assigned to both Neanderthal and Denisovan ancestry. To avoid such cases, we decided to use the posterior probabilities from both CRF runs to call archaic introgressed sites. Specifically, we considered Neanderthal introgressed haplotypes as those containing alleles with (i) Neanderthal marginal posterior probability ≥ 0.9 and (ii) Denisovan marginal posterior probability < 0.5. Likewise, we considered Denisovan haplotypes as those containing alleles with (i) Denisovan marginal posterior probability ≥ 0.9 and (ii) Neanderthal marginal posterior probability < 0.5.

Combining S′ and CRF methods

Methods. It has recently been shown that combining different methods that detect archaic introgressed sequences can increase the detection rate of truly introgressed haplotypes (i.e., decrease the FPR)[92]. We therefore sought to assess the specificity of combining the CRF and S′ methods by comparing the amount of total retrieved Denisovan haplotypes in PNG and West Eurasians, as these populations carry the highest and lowest amount of Denisovan ancestry in our dataset, respectively. Namely, we estimated the ratio of the total amount of Denisovan haplotypes retrieved in PNG to that found in West Eurasians, as a means to explore the amount of truly introgressed archaic segments in a given population[92]. Specifically, for Denisovan-introgressed haplotypes detected by the CRF method, we estimated the ratio of remaining haplotypes in PNG and West Eurasians after keeping only those haplotypes with a fraction of base-pair overlap higher than 0 (i.e., without considering S′ haplotypes), 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 (i.e. complete overlap) with S′ haplotypes. We also explored our strategy of considering the estimated Neanderthal CRF posterior probabilities when calling Denisovan introgressed haplotypes (described above).

Results. Our analysis showed that accounting for the estimated posterior probabilities of a Neanderthal haplotype reduces the total amount of Denisovan haplotypes retrieved in PNG by only ~20% (Supplementary Fig. 55a,b). Conversely, the total amount of Denisovan haplotypes in West Eurasians was reduced by up to ~70% (Supplementary Fig. 55c,d). This shows that our strategy of considering the posterior probability of Neanderthal haplotypes does remove incorrectly inferred archaic and/or ambiguous haplotypes (i.e., haplotypes with similar posterior probabilities for Neanderthal and Denisovan introgression). Our analysis also showed that using even the most lenient thresholds of overlap between the CRF and S′ methods (i.e., a base-pair overlap of only 0.1%) can result in an approximate 60-fold increase of Denisovan segments in PNG relative to West Eurasians, while still retaining a high amount of introgressed segments (Supplementary Fig. 55e,f). In light of these results, we decided to keep for each Denisovan or Neanderthal introgressed haplotype detected by the CRF method (using the procedure outlined above), only those that have a fraction of base-pair overlap higher than 0.1% with a significant S′ haplotype.



Supplementary Figure 55. Effects of analysis settings on the detection rate of Denisovan introgressed haplotypes. For all analyses, the total amount of CRF Denisovan haplotypes was obtained when filtering (solid line) or not (dashed line) for high-probability Neanderthal haplotypes. **a,** Cumulative length of Denisovan CRF haplotypes in PNG using different overlapping thresholds with S′ segments. **b,** Zoomed-in version of panel **a**. **c,** Cumulative length of Denisovan CRF identified haplotypes in West Eurasians, using different overlapping thresholds with S′ segments. **d,** Zoomed-in version of panel **c**. **e,** Ratio of the cumulative length of Denisovan haplotypes in PNG versus West Eurasians using the same parameters as in **a** and **b**. **f,** Zoomed-in version of panel **e**.

Supplementary Note 9: Match Rates of Archaic Haplotypes

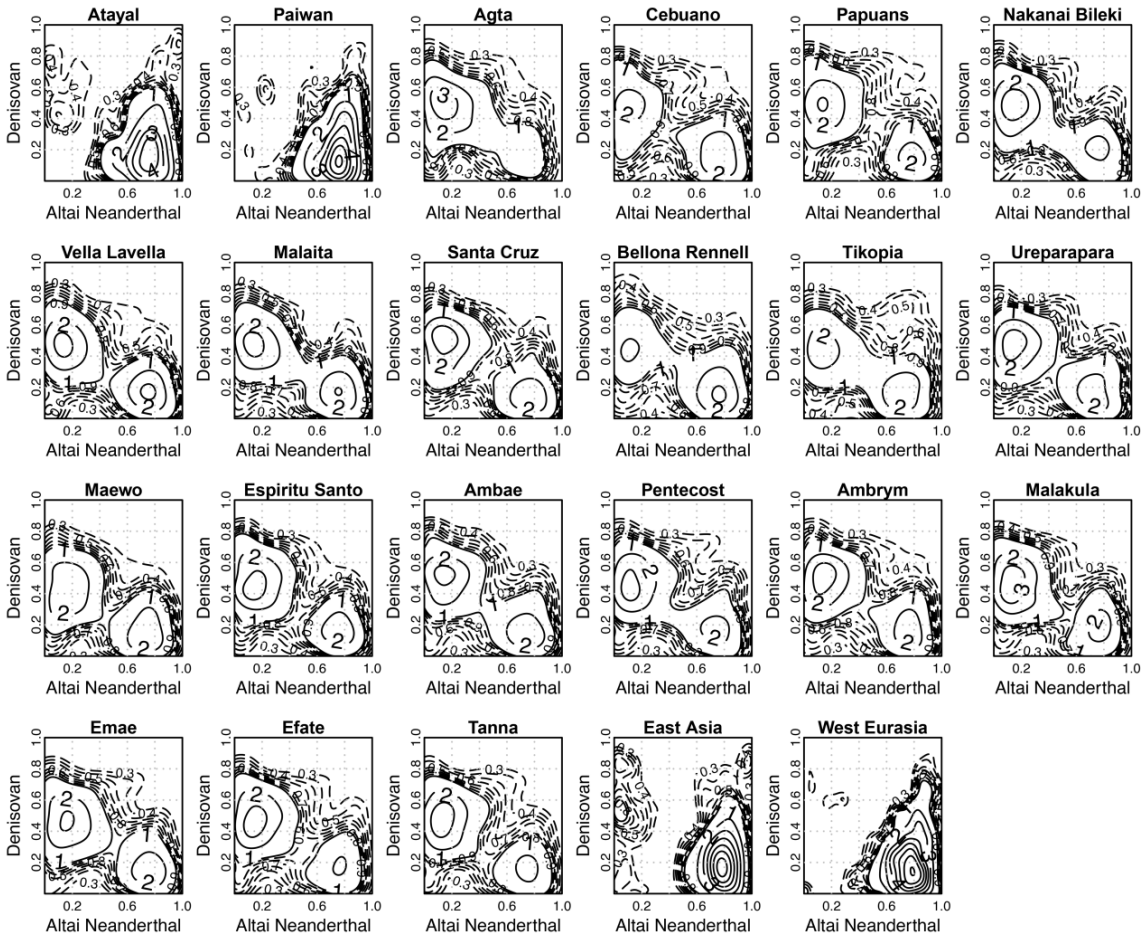
Methods

After retaining introgressed S′ haplotypes (Supplementary Note 8), we sought to compare these haplotypes to archaic genomes. Following a previous study[71], we computed a match rate between each S′ haplotype and the Vindija Neanderthal and Altai Denisovan genomes, using putatively introgressed alleles (i.e., absent in Africans). We considered that a site matches if the putative introgressed allele is present in the archaic genotype, and mismatches otherwise. The match rate was calculated as the number of matches divided by the total number of compared sites (i.e., matches and mismatches). To eliminate potentially unreliable genomic regions, owing to poor mappability or low coverage, we computed match rates using sites that pass all filters at *Level 3b'* (Supplementary Note 2). As longer S′ haplotypes carry more information on the archaic origin of introgressed segments, we only computed match rates for S′ haplotypes with more than 40 (unmasked) sites. To visualize match rates to Neanderthal and Denisovan genomes, we computed two-dimensional probability densities for the contour density plots, using the *kde2d* function from the MASS package in R version 3.4.4 with default parameters, but restricting the contour lines to the range of interest.

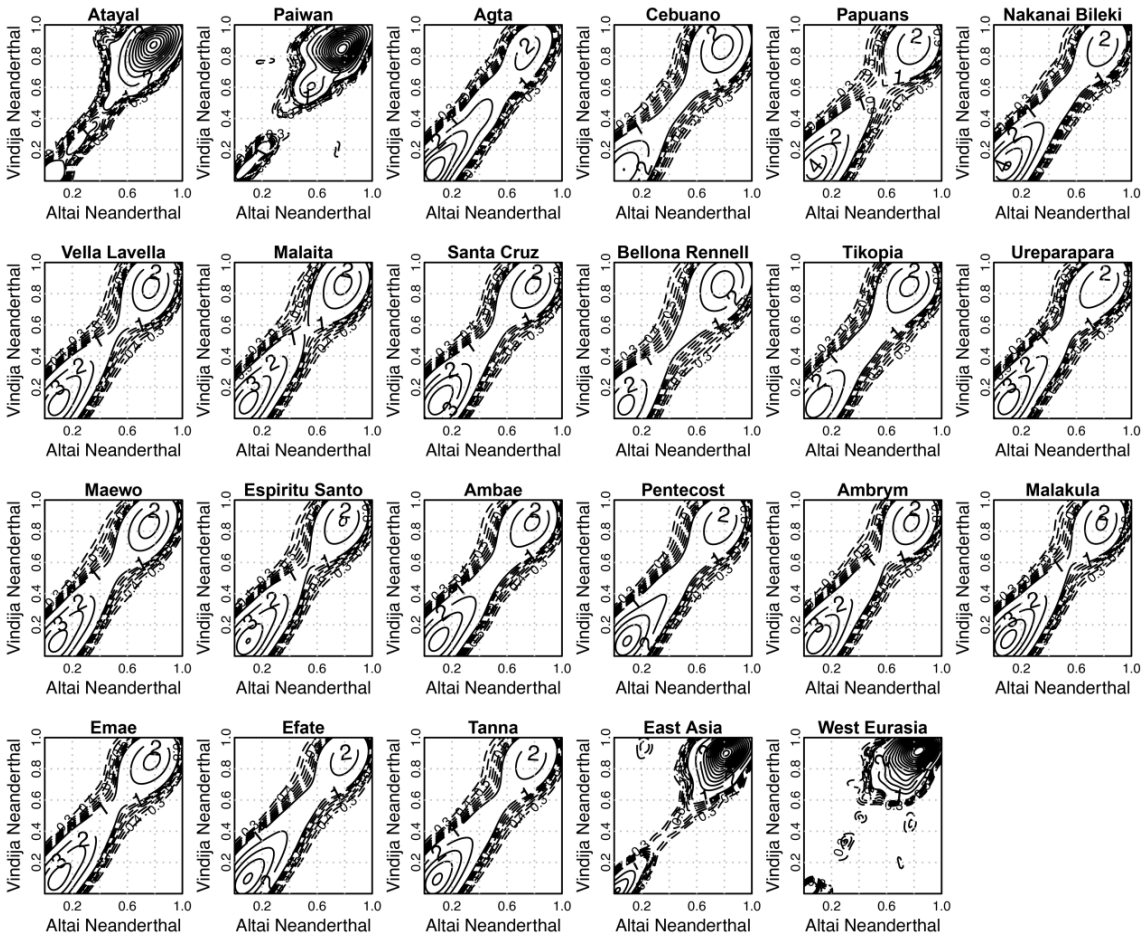
Results

Across all populations, we observed a dense cluster of S′ haplotypes with high match rate to the Vindija Neanderthal and low match rate to the Altai Denisovan (Extended Data Fig. 3). These haplotypes likely represent Neanderthal introgressed sequences. The mode of the match rate to the Vindija Neanderthal is ~0.9, which is higher than the one reported (~0.8) in ref.[71]. This is likely due to our use of a more closely related Neanderthal genome (i.e. the Vindija Neanderthal) to the actual Neanderthal population that introgressed with modern humans. Accordingly, when using the Altai Neanderthal in our match rate estimations, we obtained a mode at ~0.8 (Supplementary Fig. 56).

We also observed another cluster of S′ haplotypes with very low match rate to the Vindija Neanderthal (and Altai Neanderthal), but with a higher match rate to the Altai Denisovan (mode of ~0.5), which likely represents Denisovan introgressed haplotypes (Extended Data Fig. 3 and Supplementary Fig. 56). This cluster is most apparent in all populations but West Eurasians, where we observed only a very small cluster of Denisovan haplotypes. This observation is likely due to recent East Eurasian ancestry in some of these individuals, as previously observed[17,91]. The populations that carry this shared signal of Denisovan introgression include the Atayal and Paiwan Taiwanese indigenous peoples, the Cebuano and Agta from the Philippines, Polynesian outliers, and Near and Remote Oceanians. Notably, we also replicated a second signal of Denisovan introgression (mode at ~0.78) in East Asians[71], which is also present in the Atayal from Taiwan. Lastly, the match rates using the Altai Neanderthal and Vindija Neanderthal were, as expected, highly correlated (Supplementary Fig. 57). Nevertheless, we observed that introgressed haplotypes in Taiwanese indigenous peoples, East Asians, and West Eurasians are slightly more similar to the Vindija Neanderthal genome, in agreement with previous observations[56].



Supplementary Figure 56. Match rate of introgressed S′ haplotypes to the Altai Neanderthal and Altai Denisovan genomes. The match rate is the proportion of putative archaic alleles that match a given archaic genome, excluding sites at masked positions. Only S′ haplotypes with at least 40 sites not masked in the Vindija Neanderthal and Altai Denisovan genomes are included in the match rate calculations. Numbers inside the contour plots 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).



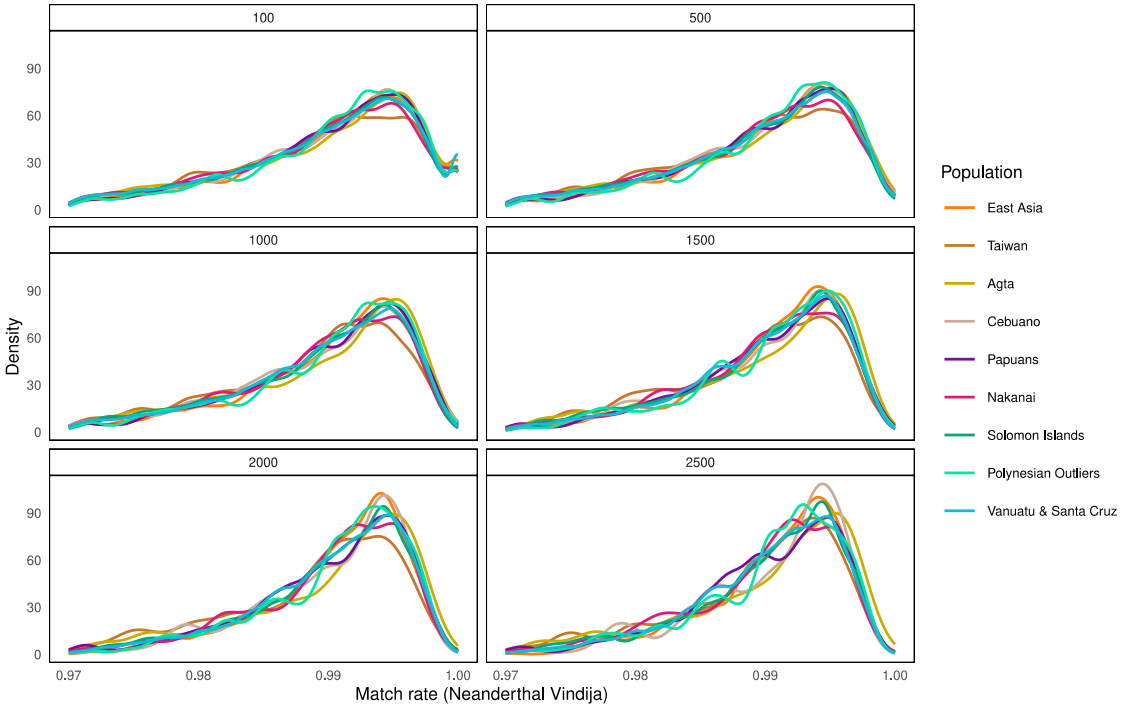
Supplementary Figure 57. Match rate of introgressed S′ haplotypes to the Altai Neanderthal and Vindija Neanderthal genomes. The match rate is the proportion of putative archaic alleles that match a given archaic genome, excluding sites at masked positions. Only S′ haplotypes with at least 40 sites not masked in the Vindija Neanderthal and Altai Denisovan genomes are included in the match rate calculations. Numbers inside the contour plots 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).

Match rates using high-confidence introgressed haplotypes

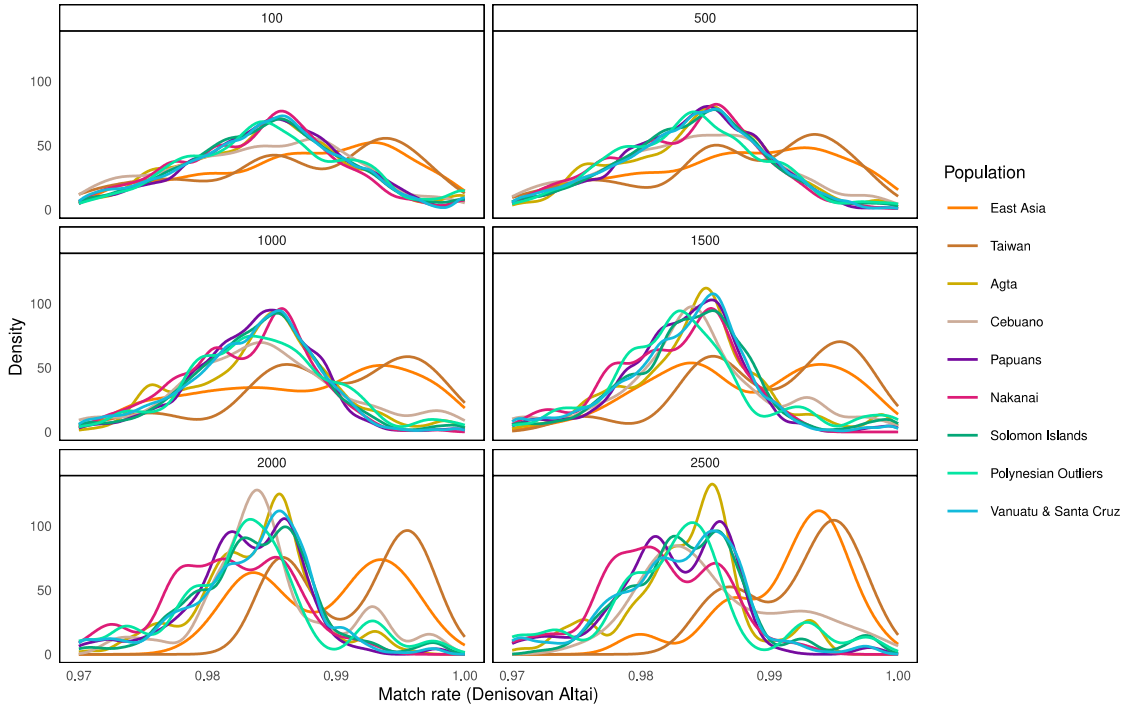
Methods. Similarly to the analysis above, we compared high-confidence introgressed haplotypes (i.e., CRF haplotypes intersecting with those detected by S′; Supplementary Note 8) to archaic genomes. This analysis was based on match rate estimates using only sites that pass all filters at *Level 3b'* (Supplementary Note 2). An important difference between the CRF and S′ haplotypes is that the latter are composed of putative archaic sites only (i.e., absent in Africans) whereas the former do not. The CRF haplotypes are therefore not only composed of the introgressed alleles, but also of alleles that are likely to be old and shared across modern humans and archaic hominins. The match rates of introgressed CRF haplotypes to the Neanderthal and Denisovan genomes, are thus much higher than those of S′ haplotypes.

Results. Using haplotypes composed of at least 100, 500, 1000, 1500, 2000 and 2500 sites, we observed a single dominant peak of match rate to the Vindija Neanderthal, across all populations (Supplementary Fig. 58). This pattern, which is in line with our observations based on S′ haplotypes (Extended Data Fig. 3), supports a unique introgression event from a

single Neanderthal population that likely occurred in the ancestors of non-Africans, as recently documented[71,92]. Likewise, we plotted the match rate to the Altai Denisovan genome (Supplementary Fig. 59). In contrast to the match rate distributions for Vindija Neanderthal, we observed several Denisovan peaks across populations, which vary depending on the length of the introgressed haplotypes considered. The Denisovan peak that is most similar to the Altai Denisovan genome was apparent in East Asians as well as in Pacific populations with high East Asian-related ancestry, even when considering haplotypes with 100 sites. The two distinct Denisovan peaks recently reported in PNG[92] were only apparent when using haplotypes with at least 2,000 sites (Fig. 3e and Supplementary Fig. 59).



Supplementary Figure 58. Match rate of high-confidence introgressed haplotypes to the Vindija Neanderthal genome. The match proportion is the proportion of alleles that match the Vindija Neanderthal genome, excluding sites at masked positions. Each panel is labelled with the minimum number of sites in the introgressed haplotypes used to compute the density distributions.



Supplementary Figure 59. Match rate of high-confidence introgressed haplotypes to the Altai Denisovan genome. The match proportion is the proportion of alleles that match the Vindija Neanderthal genome, excluding sites at masked positions. Each panel is labelled with the minimum number of sites in the introgressed haplotypes used to compute the density distributions.

Supplementary Note 10: Defining Different Denisovan Components

Rationale

To assign introgressed haplotypes to different Denisovan components, which likely reflect population structure among the Denisovan-related groups that contributed ancestry to modern humans, we fitted single Gaussian versus two-component Gaussian mixtures to the Denisovan match rate distributions.

Denisovan components in East Asians and Taiwanese peoples

We first focused on the Denisovan match rate distribution observed in East Asians and Taiwanese indigenous peoples (Atayal and Paiwan). These populations displayed a strong signal of bimodality (Fig. 3e); one mode was observed at <99% match rate to the Altai Denisovan genome, which overlaps the distribution observed in most Southwest, Near and Remote Oceanian populations, while a second mode was found at ~99.5% match rate, which is private to East Asians and Pacific populations with high East Asian-related ancestry. Note that, for this analysis, we used introgressed haplotypes with a Denisovan match rate >98% in a combined dataset including East Asian, Atayal, and Paiwan introgressed haplotypes that contained at least 100 SNPs because (i) the bimodal distribution was apparent and stable with this threshold, and (ii) using more SNPs would result in a lower number of Denisovan introgressed haplotypes for downstream analyses.

Fitting single versus two-component Gaussian mixture model strongly supported the bimodal distribution (Likelihood ratio [LHR] = 35.4; *P*-value = 9.89×10^{-8}, Supplementary Table 11). The two Gaussians are distributed according to N($\mu = 0.985$, $\sigma^2 = 7.73\times10^{-6}$) and N($\mu = 0.994$, $\sigma^2 = 8.03\times10^{-6}$). As expected, we confirmed that a bimodal match rate distribution was also strongly supported when using introgressed haplotypes that contain even higher number of SNPs, given that longer introgressed haplotypes enable better differentiation of the various introgression components (Supplementary Table 11). We used the two-mixture Gaussian model to assign introgressed Denisovan haplotypes to the two distinct Denisovan components using a probability higher than 0.80. This resulted in a classification of 219 out of 246 Denisovan introgressed haplotypes.

We next tested whether the length of the Denisovan introgressed haplotypes is significantly different, which could reflect different pulses of Denisovan introgression occurring at different times. Although the length of the Denisovan introgressed haplotypes that are most similar to the Altai Denisovan were longer (median = 131.1kb compared with median = 93.4kb), the difference was not statistically significant (Two-sided Mann Whitney *U*-test, *P*-value > 0.05). A possible explanation is that the two pulses occurred very closely in time; however, it is likely that the low number of high-confidence Denisovan introgressed haplotypes detected in East Asian and Taiwanese indigenous populations also limits our power to find significant differences. In light of this, we repeated the analysis using Denisovan introgressed haplotypes detected only by the CRF method, i.e. without intersecting these with the S′ haplotypes. In agreement with our previous analysis, a bimodal distribution was strongly supported (LHR = 132.83; *P*-value = 2.22×10^{-16}). The two Gaussians were distributed according to N($\mu = 0.985$,$\sigma^2 = 9.65\times10^{-6}$) and N($\mu = 0.994$,$\sigma^2 = 7.35\times10^{-6}$), similarly to our previous estimates. Assigning Denisovan haplotypes to these distributions using a probability higher than 0.80 resulted in a classification of 618 out of 679 CRF Denisovan haplotypes. Using these segments, we found that the Denisovan haplotypes with a match rate of ~99.4% to the Altai Denisovan were significantly longer than those with a match rate ~98.5% (median = 99.3kb compared with median = 72.7kb, One-tailed Mann-Whitney *U*-test, *P*-value = 5.14×10^{-4}). This supports a scenario in which introgression from an archaic population closely related to the Altai Denisovan occurred later in time than that from a more distant Denisova-related population.

Denisovan components in the Philippine Agta

We next focused on the match rate distribution in the Agta from the Philippines. We used introgressed haplotypes with a Denisovan match rate >98% that contained at least 2000 SNPs, because structure within Denisovan components was only apparent using this minimum number of SNPs (Supplementary Fig. 59). Fitting single versus two-component Gaussian mixture model strongly supported the bimodal distribution (LHR = 22.2; *P*-value = 5.79×10^{-5}, Supplementary Table 11). The two Gaussians are distributed according to N(μ = 0.985, $\sigma^2 = 6.36\times10^{-6}$) and N(μ = 0.993, $\sigma^2 = 1.09\times10^{-6}$). We note that the two distributions are highly similar to those observed in the East Asian and Taiwanese indigenous populations. This signal may therefore be attributed to gene flow from Austronesian-speaking groups, carrying the high-match Denisovan component, to the Philippine Agta. Interestingly, when removing the Denisovan segments that overlapped with those detected in East Asian and Taiwanese indigenous populations, the Gaussian mixture model did not support bimodality (LHR = 6.93; *P*-value = 0.07). The single Gaussian is distributed according to N(μ = 0.985, $\sigma^2 = 7.47\times10^{-6}$), which overlaps with the components found broadly across East and Southeast Asians, and Near and Remote Oceanians. If additional interbreeding has occurred between the ancestors of the Agta and Denisovan-related archaic groups, it is possible that the Austronesian gene flow into the Agta diluted most of this introgression signal. Further analysis of additional, multiple Philippine populations, with lower levels of Austronesian-related ancestry will be needed to support this hypothesis.

Denisovan components in Papuan-related groups

We then focused on the bimodal Denisovan match rate distribution observed in PNG and populations with high Papuan-related ancestry (Fig. 3e). We used introgressed haplotypes with a Denisovan match rate >98% and <99% that contained at least 2,000 SNPs, because the bimodal distribution of interest was only apparent at this range and using this minimum number of SNPs (Supplementary Fig. 59). As for our previous analysis in the Agta, we also considered Denisovan introgressed segments originating from recent Austronesian gene flow in populations from the Solomon Islands, the Vanuatu archipelago, Santa Cruz and Polynesian outliers, by removing Denisovan segments that overlapped with those detected in East Asians and Taiwanese indigenous peoples. We then fitted single vs. two-component Gaussian mixture models to Denisovan match rate distributions, in each population separately. Two-component Gaussian distributions were supported in all populations, except in the Polynesian outliers (Supplementary Table 11). The low number of Denisovan introgressed segments in Polynesian outliers (N = 50, the lowest among all populations) could have reduced the power to detect distinct Denisovan components.

Notably, the match rate distributions among populations were extremely similar, with the first component showing a mean of ~98.2%, and the second component showing a mean of ~98.6% match rate to the Altai Denisovan. We then classified Denisovan introgressed haplotypes using a probability higher than 0.80 and compared their length, in each population separately. The length of the haplotypes were significantly different in PNG (median = 435kb vs. 363kb, Two-sided Mann-Whitney *U*-test, *P*-value = 1.64×10^{-3}), Solomon islanders (median = 435kb vs. 373kb, Two-sided Mann-Whitney *U*-test, *P*-value = 1.92×10^{-4}), ni-Vanuatu and Santa Cruz islanders (median = 435kb vs. 372kb, Two-sided Mann-Whitney *U*-test, *P*-value = 8.21×10^{-15}), but not in the Bismarck archipelago islanders (Two-sided Mann-Whitney *U*-test, *P*-value > 0.05). We note that our observation in PNG is different from that recently reported[92], where the two Denisovan components were found to have the same median length. However, we confirmed our observation when varying the number of SNPs required to define introgressed haplotypes (Supplementary Table 11). In the populations where we could detect a significant difference, the Denisovan haplotypes with less similarity to the Altai Denisovan genome are longer, supporting a scenario where the pulse from a more distantly related Denisovan group occurred into PNG later in time.

Finally, we also tested whether the length of the Denisovan introgressed haplotypes from the two distinct Denisovan components in Papuans were significantly different to those found in East Asians, Taiwanese indigenous peoples, and the Philippine Agta. Given that the bimodal distribution in Papuans was only apparent using a minimum number of 2,000 SNPs

per haplotype, we used this number when classifying Denisovan haplotypes. While the length of the haplotypes in Papuans, East Asians, Taiwanese indigenous peoples, and Agta with a Denisovan match rate of ~98.5–98.6% were not significantly different (Kruskal-Wallis rank sum test, *P*-value = 0.176), putatively introgressed haplotypes in Papuans with Denisovan match rate ~98.2% were significantly different (Kruskal-Wallis rank sum test, *P*-value = 8.93×10^{-5}). Specifically, those with a ~98.2% match rate in Papuans are longer compared to those with a match rate of ~98.5-98.6% in East Asians and Taiwanese indigenous peoples (median = 435kb compared with median = 370kb, One-tailed Mann-Whitney *U*-test, *P*-value = 7.69×10^{-3}), and the Agta (median = 435kb compared with median = 357kb, One-tailed Mann-Whitney *U*-test, *P*-value = 2.08×10^{-5}). Collectively, these results suggest that the ancestors of modern humans from the Pacific experienced at least three independent introgression events from Denisovan-related archaic hominins, one being specific to East Asian-related populations, one being specific to Papuan-related populations, and one detected among East Asian- and Papuan-related populations, as well as the Agta from the Philippines.

Supplementary Note 11: Detecting shared archaic introgression

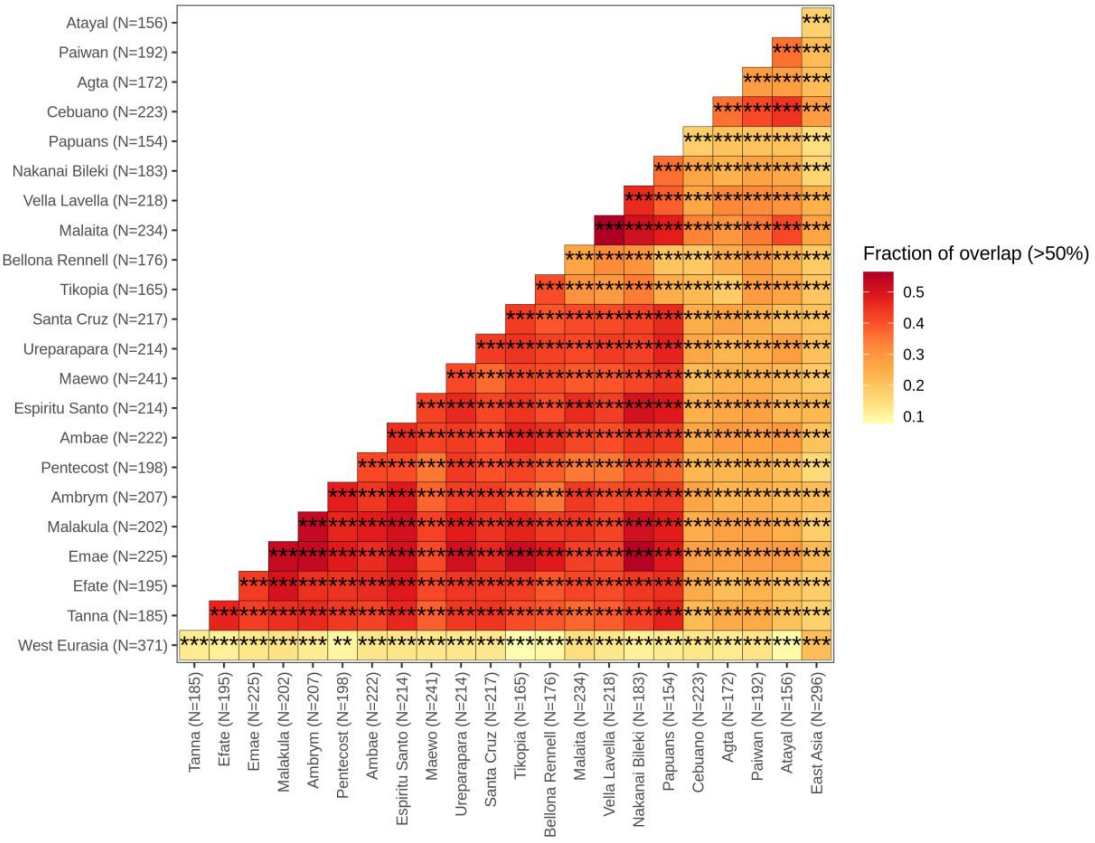
Method

We evaluated the extent to which Pacific populations share a common history of archaic introgression, by computing a statistic measuring the overlap of introgressed haplotypes detected in two human populations. Our rationale was that, if two populations inherited their archaic ancestry through an introgression event in their common ancestors, the introgressed haplotypes would tend to be observed in the same genomic positions. We based this analysis on S′ introgressed haplotypes because this method relies on tiling across individuals (i.e. detecting introgressed haplotypes at the population level) and is therefore suitable for comparing introgressed haplotype between populations. For each population, we first retained only S′ introgressed haplotypes with a score >190,000 and a length of at least 40kb in order to retain truly introgressed segments (Supplementary Note 8). We then classified each haplotype as either of Neanderthal or Denisovan origin, as in ref.[71]. The Neanderthal haplotypes are those with a match rate equal or higher than 0.6 to the Vindija Neanderthal and less than 0.4 to the Altai Denisovan. The Denisovan haplotypes are those with a match rate equal or higher than 0.4 to the Altai Denisovan and less than 0.4 to the Vindija Neanderthal. 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. The fraction of base-pair overlap is estimated with respect to the length of the segments in the first population. For example, for a 100Kb segment identified in the first population that has a 25Kb overlap with a segment in the second population, the base-pair overlap fraction is equal to 0.25. Note that this statistic is not symmetrical as two populations can have different numbers of introgressed segments. As a test statistic, we computed the proportion of segments that have a fraction of base-pair overlap higher than 0.5. To assess significance, we performed 10,000 bootstrap iterations where we randomly placed introgressed segments of the same number and length as those observed along the callable genome (~2.1 Gbp). In order to report a single *P*-value for each pairwise comparison, we took the highest *P*-value for each comparison. All *P*-values were then adjusted for multiple testing by the Benjamini-Hochberg method.

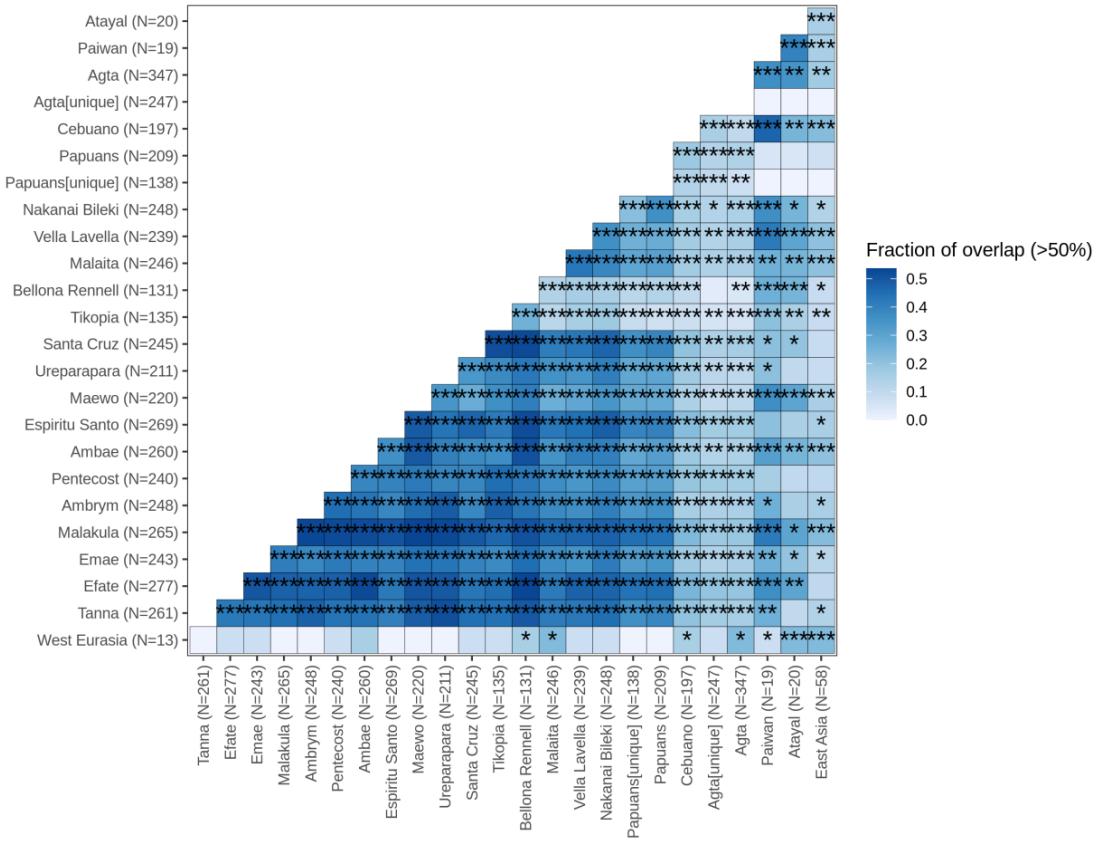
Results

Neanderthal-introgressed haplotypes overlapped significantly between all pairs of populations (FDR < 0.005 for all comparisons; Supplementary Fig. 60). This indicates that there was likely a single Neanderthal introgression event in the common ancestors of all non-Africans, in line with our observations based on Neanderthal match rate distributions (Fig. 3e and Supplementary Fig. 58), as well as previous studies[71,92]. We note that there is a tendency for higher overlap between closely related populations, suggesting that our statistic is to some extent affected by population structure. Conversely, Denisovan-introgressed haplotypes did not show a significant overlap between all pair of populations, suggesting independent introgression events (Supplementary Fig. 61). We found that Denisovan-introgressed haplotypes in Papuans were not significantly shared with East Asians, Taiwanese indigenous peoples, or West Eurasians. This result is not likely explained by a lower statistical power due to the lower number of introgressed segments, as we were able to detect sharing between some Vanuatu populations (which carry similar number of Denisovan introgressed segments as Papuans) and East Asian-related populations. This result suggests that Papuans and East Asians inherited at least part of their Denisovan ancestry through independent introgression events. This analysis is also in line with our previous observation based on match rate distributions (Fig. 3e and Supplementary Fig. 59), where East Asian populations show a Denisovan component absent from Papuans, and Papuans show a component absent from East Asians. We also found that Denisovan-introgressed haplotypes in West Eurasians overlap significantly with those present in East Asian and Taiwanese indigenous populations. One plausible hypothesis is that the Denisovan ancestry in West Eurasians was acquired through recent gene flow from East

Eurasians, after Denisovan introgression in East Asians[91]. Lastly, we also found that Denisovan haplotypes in Papuans and the Agta are significantly shared (FDR = 1.42×10^{-4}), suggesting an introgression event in their common ancestors, or in the ancestors of the Agta, followed by gene flow from the Agta to Papuan groups. As the Agta carry high levels of East Asian-related ancestry (Extended Data Fig. 1), we also repeated this analysis by removing Denisovan haplotypes in Agta that overlap with those found in East Asians. Interestingly, we found that the Denisovan introgressed haplotypes in the Agta were still significantly shared with Papuans (FDR = 9.40×10^{-4}). Removing Denisovan introgressed haplotypes in Papuans that are shared with East Asians and Taiwanese indigenous peoples also resulted in a significant sharing (FDR = 1.42×10^{-4}). Overall, these results suggest that at least some of the Denisovan ancestry present in the Agta was acquired through an introgression event shared with Papuans, occurring in their common ancestors or in the ancestors of the Agta. Further analysis in additional, multiple Philippine populations, with higher levels of Denisovan ancestry and a lower degree of East Asian-related ancestry, will be required to test these alternative scenarios.



Supplementary Figure 60. Sharing of Neanderthal-introgressed haplotypes between Pacific populations. Each cell shows the fraction of Neanderthal-introgressed haplotypes that overlap more than 50% between populations. Numbers above each population label indicate the total number of Neanderthal-introgressed haplotypes. Significance is indicated by stars, with *FDR < 0.001, **FDR < 0.01, and *FDR < 0.05.



Supplementary Figure 61. Sharing of Denisovan-introgressed haplotypes between Pacific populations. Each cell shows the fraction of Denisovan-introgressed haplotypes that overlap more than 50% between populations. Numbers above each population label indicate the total number of Denisovan-introgressed haplotypes. Agta[unique] and Papuans[unique] indicate Denisovan-introgressed haplotypes in the Agta and Papuans that do not overlap with those found in East Asians and Taiwanese indigenous populations. Significance is indicated by stars, with *FDR < 0.001, **FDR < 0.01, and *FDR < 0.05.

Supplementary Note 12: Multiple Denisovan Sources in Papuans

Rationale

It has been recently proposed that modern Papuans inherited their Denisovan ancestry through two Denisovan introgression events[92]. Our analyses also suggest the presence of two Denisovan components, based on the distribution of match rates to the Altai Denisovan genome (Fig 3e and Supplementary Fig. 59). However, a recent study did not find evidence of two distinct Denisovan lineages in Papuans and argued for a single Denisovan pulse[93]. These conflicting observations prompted us to formally test the two competing models, using an ABC approach[73] based on summary statistics computed from the S′ statistic.

Simulation setting

We used the demographic model for western Remote Oceanians (Extended Data Fig. 2b, Supplementary Table 5) with parameters fixed to ML point estimates, but adding a single (SP) or double (DP) pulse of Denisovan introgression in the Papuan branch. We also changed the sampling time of the Altai Denisovan in the model, because the age of the Altai Denisovan fossil was recently revised using a Bayesian age modelling approach that combines chronometric, stratigraphic and genetic data[94]. As the study did not provide point estimates, we used an age of 63.9 ka, which represents the centre of the reported date interval of 51.6–76.2 ka (at 95% probability). We also included a population resize in Papuans to capture the effect of the agricultural transition in Papua New Guinea[65]. Including this extra parameter was needed to obtain simulation-based summary statistics that matched our observed (empirical) summary statistics (see below). We note that our aim was not to infer this parameter and as such, the population resize in Papuans is considered a nuisance parameter. Specifically, in the SP model, we assumed a Denisovan introgression that occurred T_{DR2} generations ago (ga), that contributed α_{DR2} of Denisovan ancestry, and that involved a Denisovan lineage that diverged $T_{DR2\text{-}DenisovanAltai}$ ga from the Altai Denisovan. We refer to this Denisovan lineage as Denisovan-related lineage 2 (DR2). In the DP model, in addition to the parameters presented above, we included a second Denisovan introgression that occurred T_{DR3} ga, that contributed α_{DR3} of Denisovan ancestry, and that involved a Denisovan lineage that diverged $T_{DR3\text{-}DenisovanAltai}$ ga from the Altai Denisovan. We refer to this Denisovan lineage as Denisovan-related lineage 3 (DR3). The prior distributions for each parameter are shown in Supplementary Table 12. To differentiate between the SP and DP models, we simulated a total of 50,000 independent sets of 64 10-Mb genomic sequences per model, with *fastsimcoal2*[46]. For parameter estimation, we further computed 150,000 extra independent simulations under the best supported model.

Summary statistics

As ABC summary statistics, we used moments of the distribution of S′ scores, S′ haplotype length, and S′ match rate to the Altai Denisovan genome. As we were interested in the Denisovan introgression pulses, we restricted our analysis to Denisovan-introgressed haplotypes by retaining only those haplotypes with a match rate to the simulated Altai Denisovan genome ≥ 0.2 and < 0.3 to the simulated Vindija Neanderthal genome, as in ref.[71]. Specifically, for each statistic, we computed the minimum, median, mean, first interquantile, third interquantile, maximum, and the variance. To capture information occurring from two distinct Denisovan populations, we also fitted two Gaussian distributions on the Denisovan match rate distribution, and then computed the same summary statistics presented above for each of the classified components, using a probability classification threshold of at least 0.8. Given that in our empirical data, we only retained S′ haplotypes with an S′ score higher than 190,000 (to lower the number of false-positives) (Supplementary Note 8), we also filtered the simulated introgressed haplotypes based on this criterion before computing summary statistics. ABC was then performed using the *abc* R package[79]. To differentiate between the SP and DP models, we used a logistic multinomial regression. For parameter estimation, we used a neural network using default parameters of hidden layers and neurons.

Goodness-of-fit and method performance

Prior to model choice and parameter estimation, we checked whether the SP and DP models provided a good fit to the observed data. We performed a goodness-of-fit test using 100 replicates for each model, and a tolerance set to 5%. As an additional (and graphical) procedure, we also performed Principal Component Analysis (PCA) of all summary statistics, using the first two PCs, and displayed the 90% envelope of the first two PCs for each model. To evaluate the performance of model selection and parameter inference, we used a "leave-one-out" cross validation approach using 100 replicates.

Results

Our goodness-of-fit test showed that only the DP model was able to produce S′-based summary statistics that are consistent with the observed data (goodness-of-fit SP model, *P*-value<0.01; goodness-of-fit DP model, *P*-value=0.24). Similarly, the 90% envelope of the first two PCs, computed from all summary statistics for each model separately, showed that a high proportion of summary statistics simulated under the DP model were more similar to the observed values (Supplementary Fig. 62a). Cross-validation via 100 independent simulations for difference tolerance rates showed that our ABC approach was able to distinguish between these two different models with high accuracy (>83%; Supplementary Fig. 62b), the highest accuracy being obtained at a tolerance level of 5%. ABC model selection at a tolerance level of 5% showed that the DP model was strongly favoured (posterior probability = 99%). We also conducted a second ABC analysis, using only the S′-based summary statistics computed from Denisovan match rates. We conducted this second analysis in order to test whether SP and DP models could produce distinct Denisovan match rate distributions. Interestingly, the goodness-of-fit test showed that both models were able to produce summary statistics that are consistent with the observed data (goodness-of-fit for the SP model, *P*-value=0.19; goodness-of-fit for the DP model, *P*-value=0.47). The 90% envelope of the first two PCs showed that summary statistics simulated under the DP model were more similar to the observed values (Supplementary Fig. 63a), in line with our previous observation. Encouragingly, the cross-validation analysis based on this subset of summary statistics showed that the ABC approach was still able to distinguish between these two models, albeit with slightly lower accuracy (>76%; Supplementary Fig. 63b). By performing ABC model selection at a tolerance level of 0.005, 0.01, and 0.05 we found that the DP model was strongly favoured, resulting in a posterior probability of 99% for the three tolerance rates. Together, our analyses support that the Denisovan ancestry in Papuans arose from (at least) two distinct introgression events from two Denisovan-related populations.

We next focused on the DP model to infer the different parameters of the two distinct Denisovan introgression pulses. For parameter inference, we relied on the full set of summary statistics, as these are likely to be more informative for the parameters of interest. For example, the length of the S′ introgressed haplotypes are expected to be highly informative for estimating the time of introgression. We estimated the performance of our ABC approach using only summary statistics with a correlation coefficient (*r*) higher than 0.1 to the parameter of interest (Supplementary Fig. 64). Based on prediction error (PE) as a performance measure, we found a high accuracy for the estimation of the divergence time between the introgressing Denisovan lineages and the Altai Denisovan, but only relatively moderate and low accuracy for the estimation of the time of introgression of these Denisovan components into Papuans and the introgression rate, respectively (Supplementary Fig. 65). We therefore caution in interpreting these parameter estimates. Assuming a 29-year generation time, the divergence times of the two distinct Denisovan lineages to the Altai Denisovan were dated to 409 ka (95% CI: 335–497 ka), and 222 ka (95% CI: 174–263 ka), respectively (Supplementary Fig. 66). The introgression pulses from these two lineages were dated to 25 ka (95% CI: 15–35 ka), and 46 ka (95% CI: 39–56 ka), respectively. Lastly, we estimated similar levels of Denisovan introgression rate for these two events, with a point estimate of 2.7% (95% CI: 1.1–4.6%) for the more recent pulse, and a point estimate of 3.2% (95% CI: 1.2–5.1%) for the more ancient pulse.

Our results are in agreement with those previously obtained by Jacobs et al. (2019)[92], supporting the presence of two deeply divergent Denisovan lineages in Papuans. However, we note that in the model proposed in the aforementioned study, the oldest introgression is from a Denisovan lineage more distantly related to the Altai Denisovan genome, than the one introgressing more recently. Conversely, in our model, the oldest introgression event is from a Denisovan lineage more *closely* related to the Altai Denisovan genome, than the Denisovan lineage introgressing more recently. In order to formally choose between these two models, we simulated an extra set of 50,000 simulations under each model. We used the same set of summary statistics as previously described, but used as limits for the uniform parameter priors the 95% CIs reported in this section or those in Jacobs et al. (2019), to simulate each model accordingly. A goodness-of-fit test showed that only our model was able to reproduce observed S′-based summary statistics (goodness-of-fit for our model: *P*-value>0.05, versus goodness-of-fit for Jacobs' model: *P*-value<0.01). By performing ABC model selection at a tolerance level of 0.005, 0.01, and 0.05, we found that our model was strongly favoured, resulting in a posterior probability of 99% for the three tolerance rates.

Interpretation

Our results support the presence of two deeply divergent Denisovan-related lineages in modern Papuans. In contrast to a previously proposed model[92], we favour a scenario in which the more distantly Denisovan-related lineage introgressed with Papuans more recently. Given the geographic location of the Altai Denisovan, we consider this model as more parsimonious: it suggests that the ancestors of Papuan-related populations, as they migrated from mainland Eurasia to Oceania, introgressed with Denisovan-related groups that were increasingly different from the Altai Denisovan. The older Denisovan introgression event, dated at 46 ka (95% CI: 39–56 ka) could have occurred before, or very close to, the colonization of Sahul. In turn, the more recent Denisovan introgression event occurred likely after the initial colonization of Sahul, which fits with our proposed date at 25 ka (95% CI: 15–35 ka). These results collectively suggest that Papuans received two independent Denisovan introgression pulses from two divergent Denisovan lineages, one being probably old and potentially representing an introgression event common to populations with Denisovan ancestry, followed by another independent pulse specific to Papuan-related populations.

In Extended Data Figure 10, we show a schematic model recapitulating the history of archaic introgression that is consistent with our data. Our results are in agreement with previous studies showing that the ancestors of non-Africans experienced a unique introgression event from a single Neanderthal population[52,56,59,71,92,95]. We estimate that interbreeding occurred ~61 ka (95%CI: 56–62 ka; Extended Data Fig. 2a, Supplementary Table S2), from a Neanderthal lineage that was closely related to the Vindija Neanderthal (divergence time at 122 ka [95%CI: 107–128 ka]).

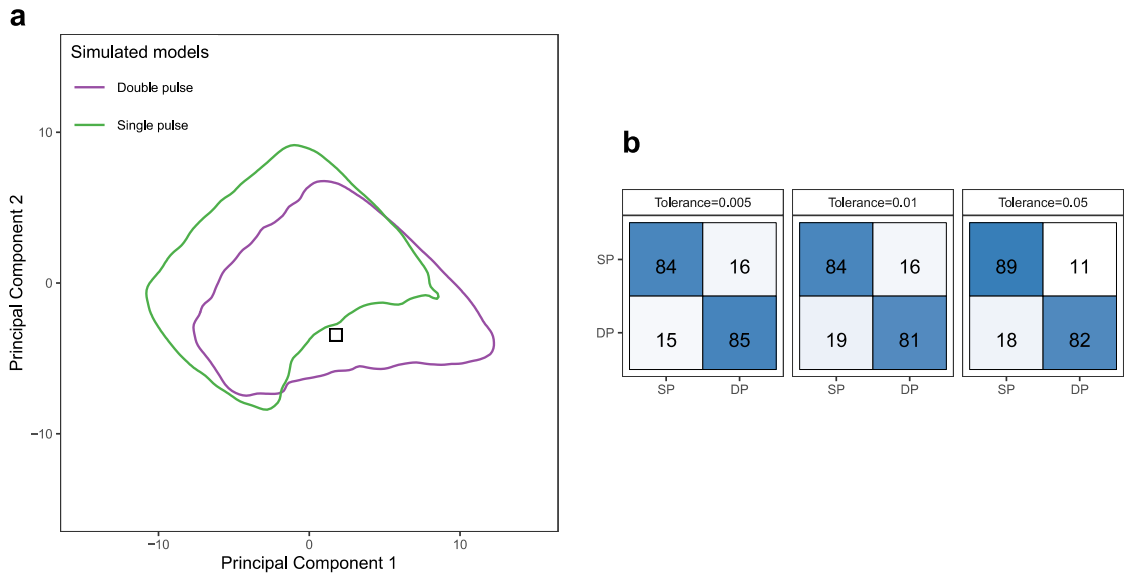
Conversely, we infer at least three independent introgression events from Denisovans into modern humans, and report suggestive evidence for a fourth event. The oldest inferred event of interbreeding occurred in the ancestors of Papuan-related groups, who were probably also the ancestors of Australian and Philippine Agta[5,7,86,91], ~46 ka (95% CI: 39–56 ka) from a lineage that diverged 222 ka (95%CI: 174–263 ka) from the Altai Denisovan. A putative location would therefore appear to be either in mainland Asia or in the Sunda Shelf, before the divergence of these populations. As East Asians carry only trace amounts (<1%) of this introgression event, we suggest the Denisovan ancestry in these populations was likely acquired through gene flow from Near Oceanians or Philippine 'Negritos'.

The second Denisovan introgression event was estimated to occur ~25 ka (95% CI: 15–35 ka), from a very divergent Denisovan lineage to the Altai Denisovan (divergence estimated at 409 ka [95%CI: 335–497 ka]). Evidence for interbreeding is restricted to PNG and Papuan-related populations from nearby islands. The estimated date suggests that the introgression event may have occurred in Sundaland or even further east of the Wallacea line.

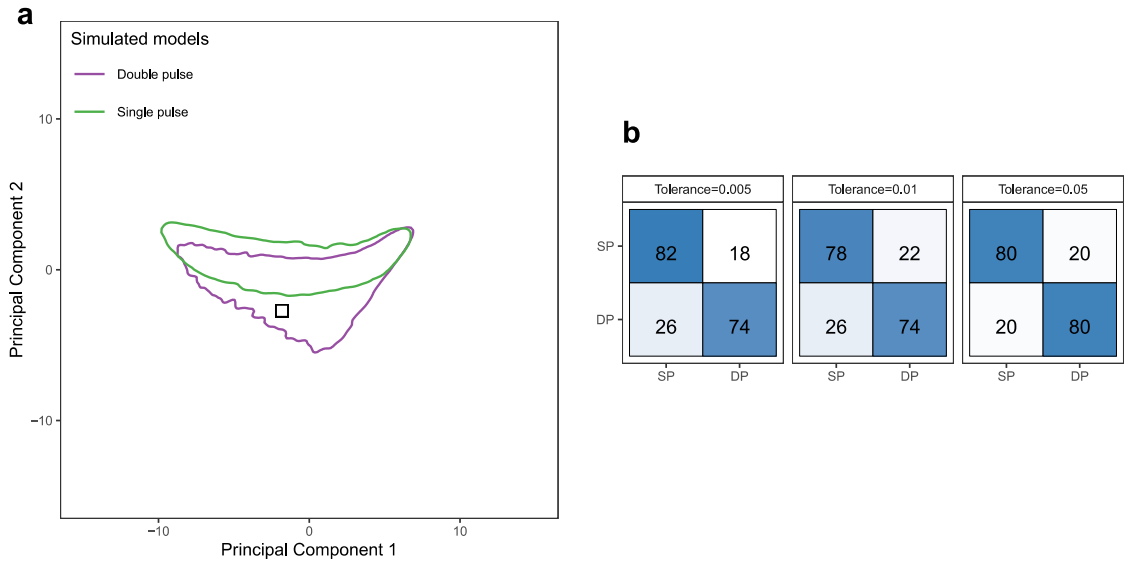
The third Denisovan interbreeding event is inferred in East Asians, from a Denisovan lineage closely related to the Altai Denisovan, as recently reported[71,92]. We date this

introgression to ~21 ka (95% CI: 15–26 ka; Supplementary Table 7). Given the strong genetic similarity of this Denisovan lineage with the Altai Denisovan, it is possible that the introgression event occurred in mainland Asia. The presence of this Denisovan component in the Philippine Agta and western Eurasians could have been acquired through gene flow with East Asian groups.

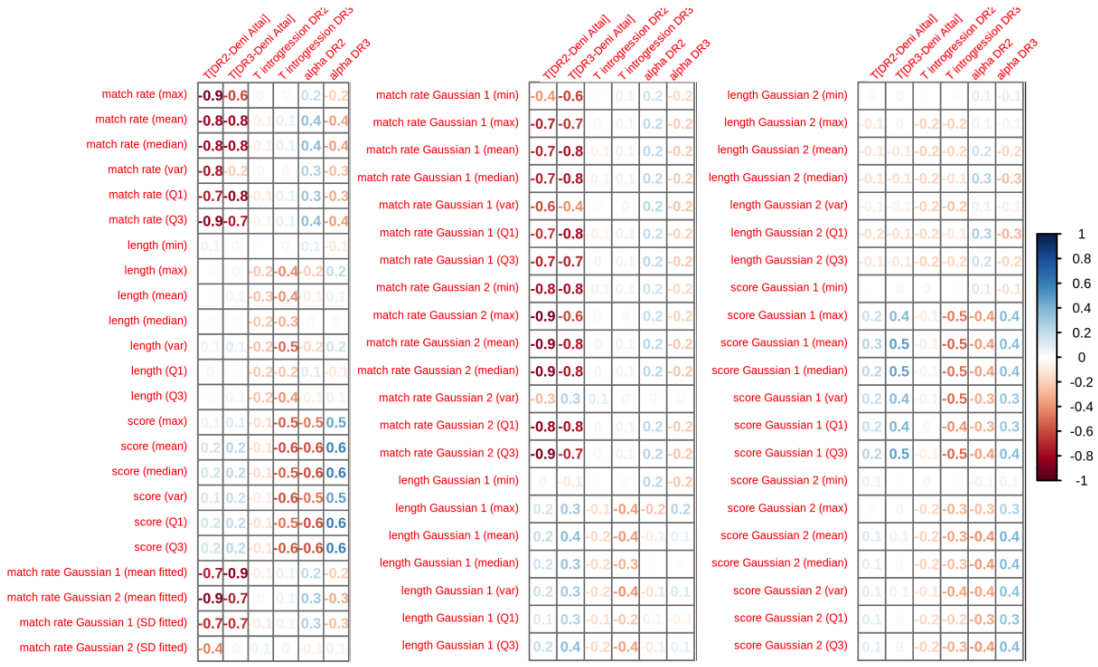
Lastly, although we infer that part of the Denisovan ancestry detected in the Philippine Agta was inherited through a common introgression event with PNG (Supplementary Note 11), the fact that (i) the Denisovan ancestry in Agta is disproportionately high, given their Papuan-related ancestry, and (ii) they show a total proportion of Denisovan ancestry comparable to that of PNG, despite their high East Asian–related ancestry (Extended Data Fig. 1), suggest that an additional, independent introgression event occurred in the ancestors of this Philippine group[7].



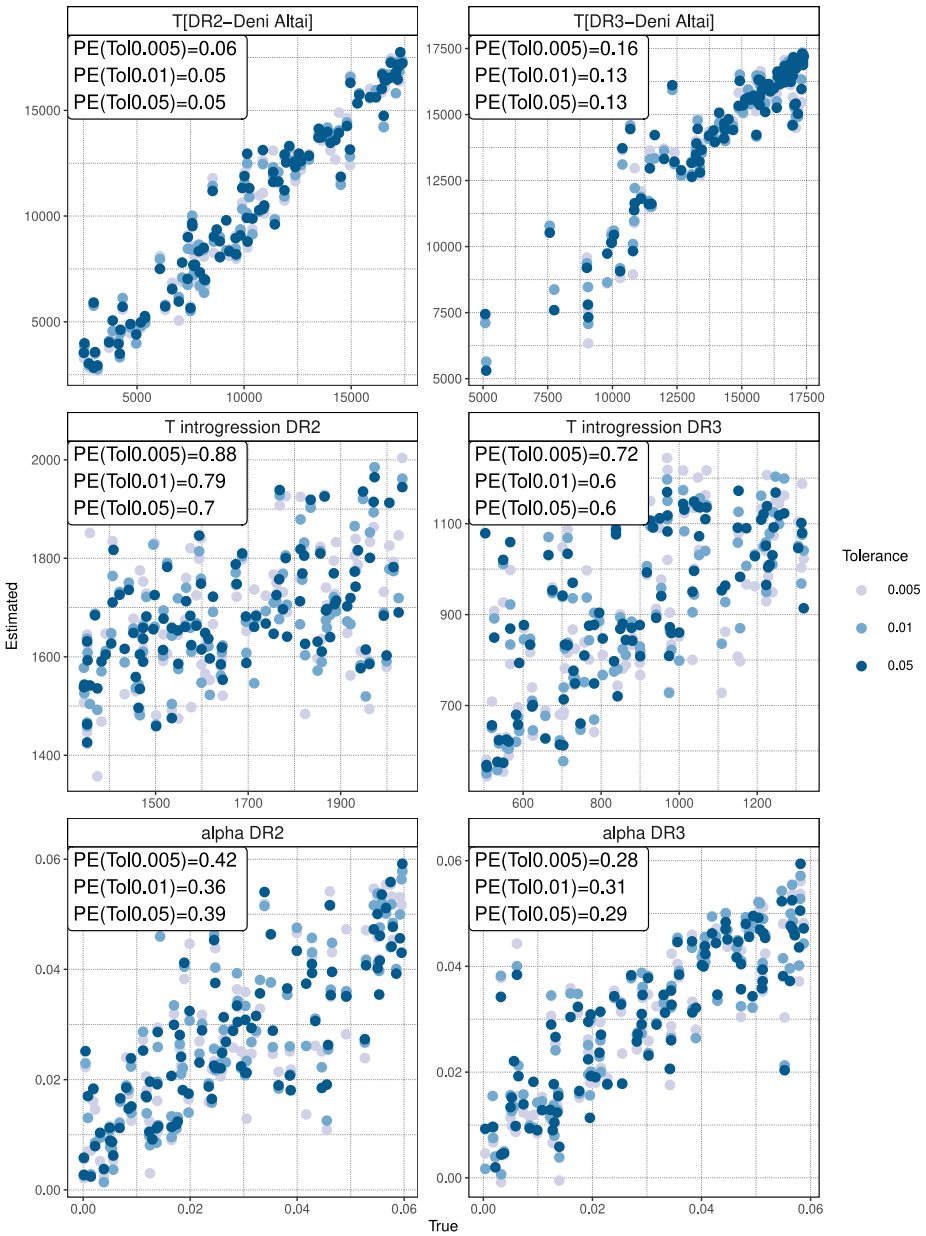
Supplementary Figure 62. A priori check and performance evaluation of the ABC approach used to differentiate between the single-pulse and double-pulse models of Denisovan introgression in PNG. **a,** PCA of the ABC summary statistics obtained for the two simulated introgression models (90% coloured contours) and the observed data (black square). **b,** Confusion matrix showing cross-validation prediction accuracy at different tolerance rates. SP and DP stand for single-pulse and double-pulse models of Denisovan introgression, respectively.



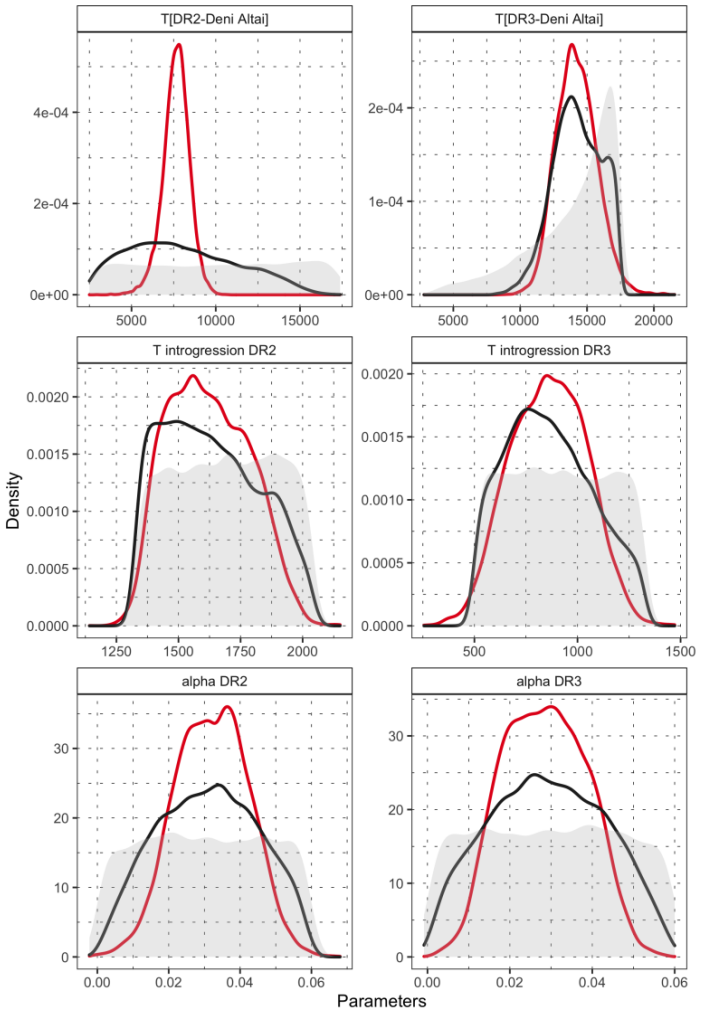
Supplementary Figure 63. A priori check and performance evaluation of the ABC approach used to differentiate between the single-pulse and double-pulse models of Denisovan introgression in PNG, based only on Denisovan match rate summary statistics. **a,** PCA of the ABC summary statistics obtained for the two simulated introgression models (90% coloured contours) and the observed data (black square). **b,** Confusion matrix showing cross-validation prediction accuracy at different tolerance rates. SP and DP stand for single-pulse and double-pulse models of Denisovan introgression, respectively.



Supplementary Figure 64. Pearson correlation coefficients between parameters and summary statistics employed in the ABC approach.



Supplementary Figure 65. Performance of the ABC estimation of parameters in the double pulse model of Denisova introgression. Prediction errors (PE), as a measure of the ABC performance for three different tolerance rates, are shown in the upper-left corner of each panel.



Supplementary Figure 66. ABC estimation of the parameters of the double pulse model of Denisovan introgression in Papuans. Prior (grey area) and posterior (red and black lines) distributions are shown for the split time between Denisovan lineages and the Altai Denisovan ($T_{\text{DR2-Deni Altai}}$ and $T_{\text{DR3-Deni Altai}}$), the time of introgression ($T_{\text{introgression-DR2}}$ and $T_{\text{introgression-DR3}}$) and the introgression rate (α_{DR2} and α_{DR3}) of the two distinct Denisovan lineages into PNG. Black and red curves indicate posterior distributions obtained with the rejection algorithm and neural networks, respectively.

Supplementary Note 13: Exploring Unknown Archaic Introgression

Rationale

The S′ approach is aimed at detecting introgressed haplotypes without the need of an archaic reference genome[71]. We therefore sought to characterize S′ introgressed haplotypes from archaic hominins other than Neanderthals or Denisovans, as they may potentially reveal introgression from unknown archaic humans. Notably, given the presence of archaic hominins in the Philippines (*Homo luzonensis*)[96] and Indonesia (*Homo floresiensis*)[97], it is of interest to study potential unknown archaic ancestry in the Agta and Cebuano from the Philippines.

Methods

To retain only S′ haplotypes introgressed from a potentially unknown hominin, we removed S′ haplotypes that are likely to be of either Neanderthal or Denisovan origin, based on their overlap with Neanderthal or Denisovan haplotypes detected by the CRF approach (Supplementary Note 8). To further characterize these S′ haplotypes, we estimated their match rate to the Vindija Neanderthal and Altai Denisovan genomes.

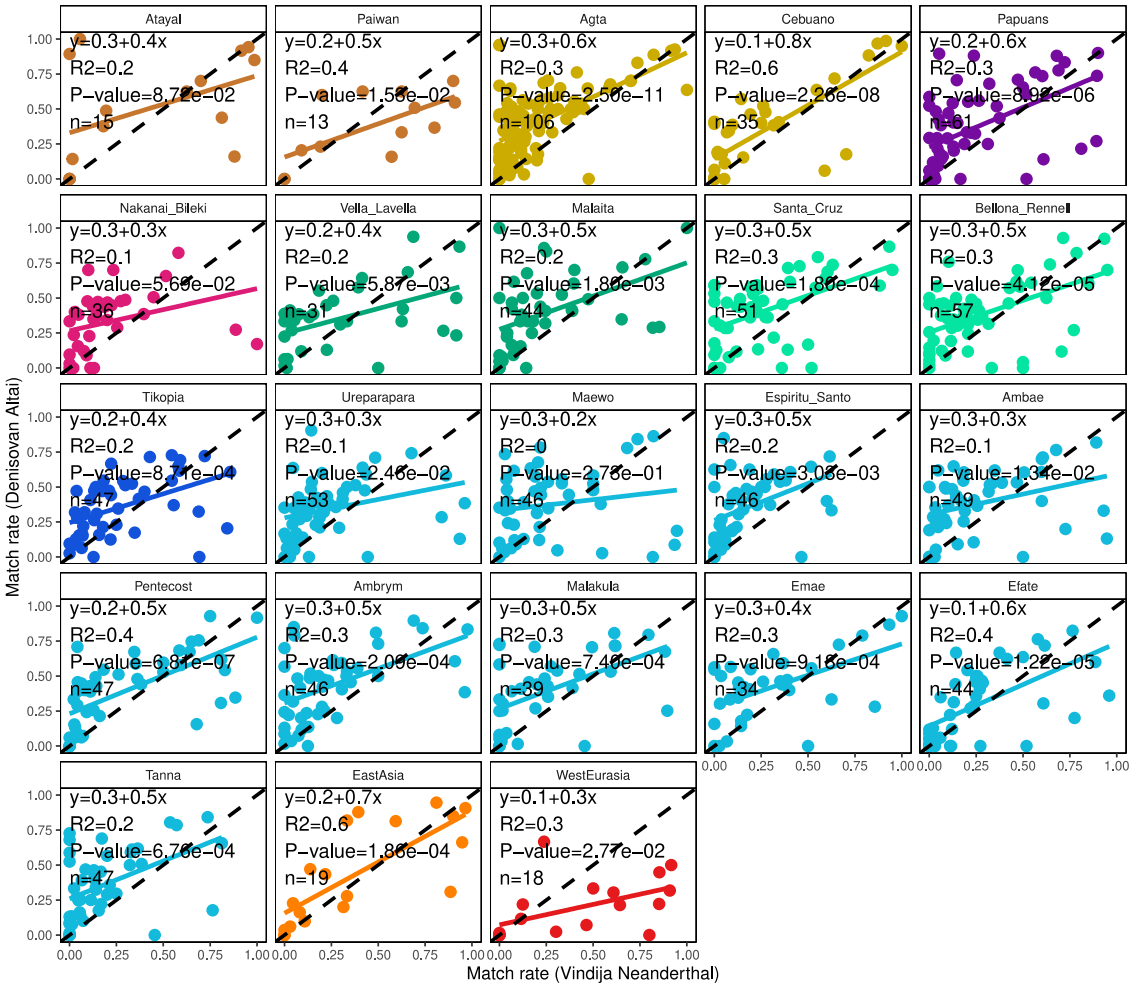
Results

The total amount of S′ haplotypes retrieved among populations showed that, as expected, populations harbouring both Neanderthal and Denisovan ancestry have the highest number of S′ haplotypes (Extended Data Fig. 4a). Note that S′ is estimated at the scale of the population sample, and we are simply analysing the total number of retrieved S′ haplotypes, so the highest numbers of introgressed S′ sequences are found among population samples with the largest sizes (Supplementary Table 1). We next removed S′ haplotypes of either Neanderthal or Denisovan origin, as determined by the CRF method. As expected, we observed a strong reduction in S′ haplotypes in populations with high levels of Neanderthal and Denisovan ancestry (Extended Data Fig. 4a), particularly in PNG, Papuan-related and Agta populations.

Characterizing these patterns in west Eurasians can be particularly informative as these populations carry minimal levels of Denisovan ancestry, as previously observed[92], and all S′ haplotypes should mostly be of Neanderthal origin. Accordingly, we observed a strong reduction of S′ sequences after removing Neanderthal haplotypes in this population group. Nevertheless, we observed a moderate reduction in the amount of S′ haplotypes in west Eurasians after removing Denisovan haplotypes (Extended Data Fig. 4a). Two patterns might explain this observation. First, in contrast to the CRF approach, the S′ approach detects introgressed haplotypes at the population level, and thus detects significantly longer haplotypes, as it basically concatenates several distinct introgressed haplotypes across chromosomes[71]. A single S′ haplotype can therefore overlap with several Neanderthal and/or Denisovan CRF haplotypes from different individuals. Second, it is also possible that these S′ haplotypes are false positives, or that the pattern is due to the low sensitivity of the CRF method to detect introgressed haplotypes, together with the stringent posterior probability threshold (set as >0.90) used to call CRF introgressed haplotypes (Supplementary Note 8).

To further characterize the remaining S′ haplotypes, we estimated their match rate to the Vindija Neanderthal and Altai Denisovan genomes. Noticeably, we found that several S′ haplotypes have high match rates to Vindija Neanderthal or Altai Denisovan (Supplementary Fig. 67). These S′ haplotypes are likely archaic segments of Neanderthal or Denisovan origin that were not detected by the CRF approach. This is most apparent in the west Eurasian population, where the majority of S′ haplotypes are located below the diagonal, and with high Neanderthal match rate values, similar to those found in our original S′ analysis (Supplementary Note 9). Likewise, in PNG, we observed several S′ haplotypes that are located above the diagonal and with a high match rate to the Denisovan genome. We also observed several S′ haplotypes that are of clearly ambiguous origin. This is most evident in the Agta, Cebuano, and East Asian populations, where many S′ haplotypes show a similar

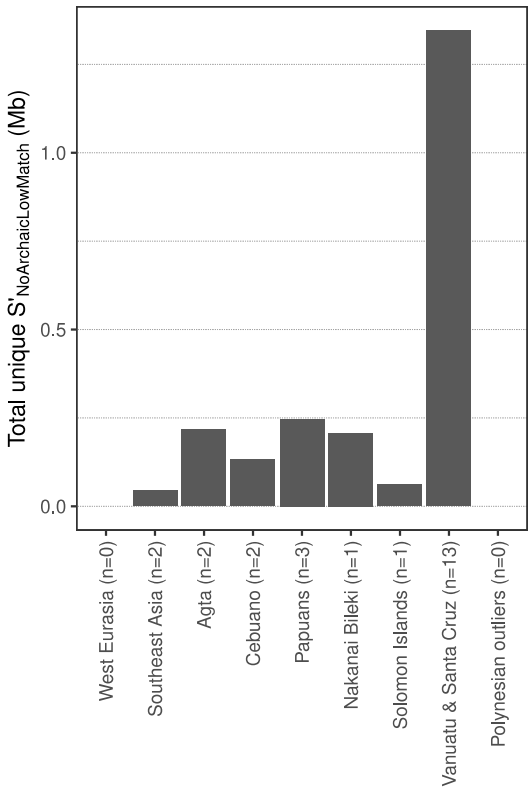
match to the Vindija Neanderthal and Altai Denisovan (i.e. located at the diagonal), which resulted in a strong correlation (see R^2 values in Supplementary Fig. 67). We therefore removed all S′ haplotypes with a match rate higher than 1% to either the Vindija Neanderthal or Altai Denisovan genome. The remaining S′ haplotypes, which we termed $S'_{NoArchaicLowMatch}$, do not overlap with CRF Neanderthal or Denisovan haplotypes.



Supplementary Figure 67. Match rate of S′ haplotypes to the Vindija Neanderthal and Altai Denisovan, after removing those that overlap with CRF haplotypes. The coloured line indicates the best fit regression line, and the black, dashed line, the identity line. The number of haplotypes used to compute the correlation, as well as the linear equation, the correlation coefficient R^2 and the corresponding *P*-value, are shown inside each panel.

After removing all haplotypes of potential Neanderthal or Denisovan origin based on the CRF method and estimated match rates, we retained a total of 59 S′ haplotypes among all populations (Extended Data Fig. 4b). If introgression occurred from an unknown archaic hominin and is present in some groups and not others (e.g. a local *Homo erectus* population in Southeast Asia, or *Homo luzonensis)*, we would expect to find that the remaining S′ haplotypes are not shared among populations. In contrast with this expectation, we found that most of these remaining S′ haplotypes are shared (Extended Data Fig. 4b). For example, we observed that all haplotypes in west Eurasians and Polynesian outliers are shared with other populations. Across the remaining populations, we also observed that ~50% of these haplotypes can be found in other populations. Further characterizing the S′ haplotypes that are unique to specific populations, we only retained <2 Mb of introgressed material per population group (Supplementary Fig. 68). For example, less than 1Mb of S′ haplotypes were

detected in the Agta and Cebuano from the Philippines, our two populations of interest. Overall, these results suggest limited evidence of introgression from hominins other than Neanderthal and Denisovan in Philippine populations, or, alternatively, that these hominins were closely related to Neanderthals or Denisovans.



Supplementary Figure 68. Total amount of population-specific $S'_{NoArchaicLowMatch}$ haplotypes. For each population, the number of haplotypes used to compute the total amount of population-specific $S'_{NoArchaicLowMatch}$ haplotypes is shown in brackets, next to the population label.

Supplementary Note 14: Adaptively-Introgressed Haplotypes

Methods

Two recently developed statistics have been used to detect candidate regions for adaptive introgression (AI), based on the number and derived allele frequency of sites that are uniquely shared between archaic hominins and modern humans[98]. Briefly, under AI, one would expect to find archaic introgressed alleles at high frequency in a population known to carry archaic ancestry, but absent (or at very low frequency) in a population without archaic ancestry. The $Q95(w,y,z)$[98] is defined as the 95th percentile of derived allele frequencies within a genomic window in a target population, where the derived allele frequency of these sites in an outgroup population (i.e. a population without the archaic ancestry of interest) is lower than $w$is higher than y in an archaic hominin, but lower than z in a different archaic hominin. Throughout this section, we refer to w, y, and z as the derived allele frequency in Africans, the Vindija Neanderthal genome, and the Altai Denisovan genome. To find Neanderthal-specific AI genomic windows, we therefore defined the $Q95_{Neanderthal}$ statistic as:

$$Q95_{Neanderthal}(w = 0.01, y = 1, z = 0).$$

For the sake of clarity, this statistic estimates the 95th percentile of the derived allele frequencies in a target population that are lower than 1% in Africans, fixed in the Vindija Neanderthal genome, but absent in the Altai Denisovan genome. Likewise, we defined the $Q95_{Denisova}$ statistic as:

$$Q95_{Denisova}(w = 0.01, y = 0, z = 1).$$

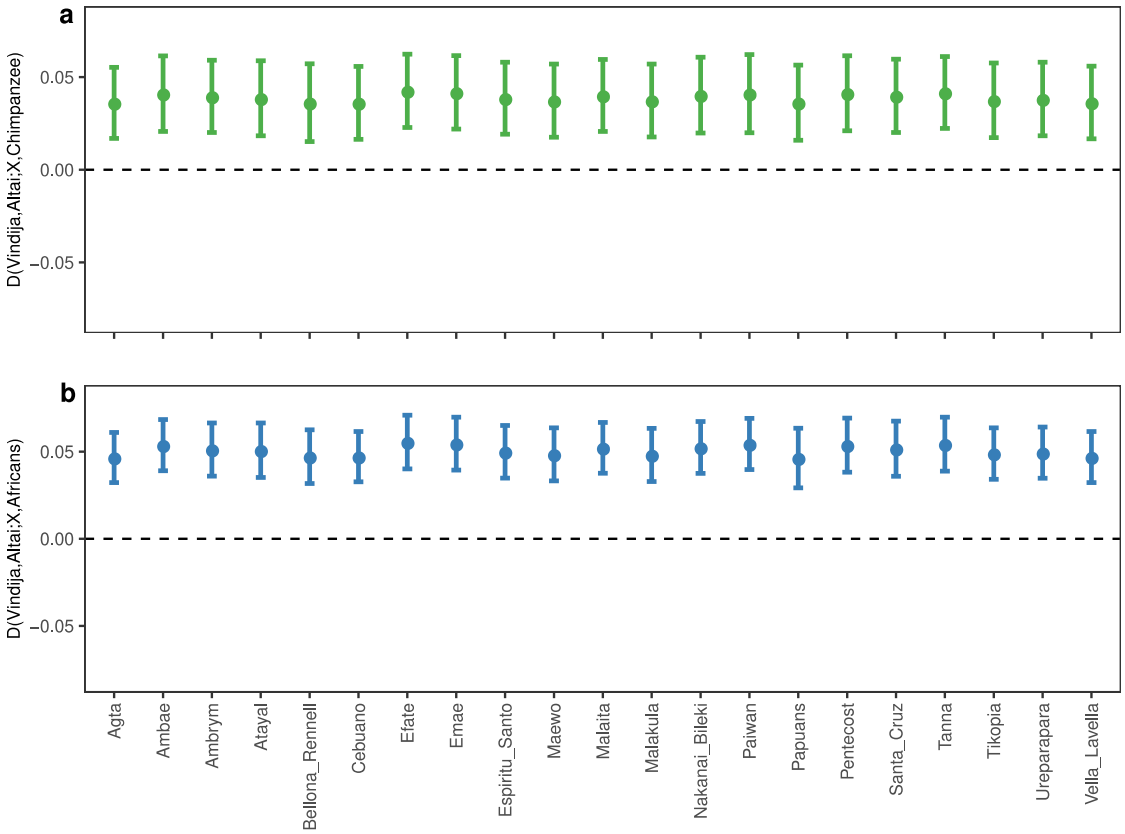
We also computed a complementary statistic – the $U(w,x,y,z)$ statistic[98]. Unlike the $Q95$ statistic, this statistic counts the *number of sites* within a genomic window where a target population has a derived allele frequency higher than x, where the derived allele frequency in an outgroup population is lower than w, and where the derived allele frequency in an archaic hominin is higher than y, but lower than z in a different archaic hominin. In the original study[98], the authors set the derived allele frequency x to 20%, 30% or 50%. This is a sensible approach as the archaic ancestry between different populations may vary greatly, and setting a unique threshold of derived allele frequency x for all populations of interest can result in a very lenient or conservative statistic. Instead of using several different thresholds, we decided to set a single derived allele frequency threshold x for each target population separately by computing the $Q95(w,y,z)$ statistic using *all* sites across the genome. We term this statistic $Q95_{Genome\ Archaic}$. The $Q95_{Genome\ Neanderthal}$ and $Q95_{Genome\ Denisova}$ values used to compute the *U*-statistic in each population are reported in Supplementary Table 13. Note that for the Atayal and Paiwan populations, the $Q95_{Genome\ Denisova}$ threshold is zero, which would mean that any site that is fixed in the Altai Denisovan (but absent in the Vindija Neanderthal genome) and at lower frequency than 1% in Africans would count towards the *U*-statistic computed in these populations. We defined the $U_{Neanderthal}$ statistic as follows:

$$U_{Neanderthal}(w = 0.01, x = Q95_{Genome\ Neanderthal}, y = 1, z = 0).$$

For the sake of clarity, this statistic estimates the number of sites that have a derived allele frequency higher than the population specific $Q95_{Genome\ Neanderthal}$ value, lower than 1% in Africans, fixed in the Vindija Neanderthal genome, but absent in the Altai Denisovan genome. Analogously, we defined the $U_{Denisova}$ statistic as:

$$U_{Denisova}(w = 0.01, x = Q95_{Genome\ Denisova}, y = 0, z = 1).$$

Given that there is only one archaic genome available for the analysis, a fixed or absent allele simply refers to a homozygous state for one allele or the other. We did not include the Altai Neanderthal in this analysis, as this individual is thought to be more distantly related to the Neanderthal population that introgressed with modern humans[56]. Our analyses, using *D*-statistics, confirmed this observation by showing that our target populations share significantly more derived alleles with the Vindija Neanderthal than with the Altai Neanderthal (Z-score > 2 for all populations comparisons) (Supplementary Fig. 69).



Supplementary Figure 69. Derived allele sharing between Vindija Neanderthal or Altai Neanderthal and modern human populations. **a**, Derived allele sharing using the Chimpanzee or **b**, Africans as outgroup populations. Points show derived allele sharing (*D*-statistic) and bars show two standard errors from the point estimate computed via a weighted-block jackknife procedure. Population sample sizes are reported in Supplementary Table 1.

We computed U and $Q95$ statistics in 40-kb non-overlapping windows along the genome of all target populations, with the exception of PNG[16] and Nakanai Bileki[15] individuals, because of ethical restrictions in this regard. We decided to use this window size because the mean length of introgressed haplotypes in ref.[59] was ~44kb. For both statistics, we used all 35 Africans from the SGDP dataset[17] as the outgroup population. We defined the ancestral/derived states of alleles using the chimpanzee reference genome and removed sites with any missing genotype, and discarded genomic windows with less than 5 sites, leaving a total of ~65,000 non-overlapping genomic windows in each population. Lastly, our candidate genomic windows of AI were considered as those with both U and $Q95$ statistics values in the top 0.5% of their respective genome-wide distribution (Supplementary Tables 14 and 15). Custom-generated codes to compute U and $Q95$ statistics are available on GitHub (www.github.com/h-e-g/evoceania).

Results

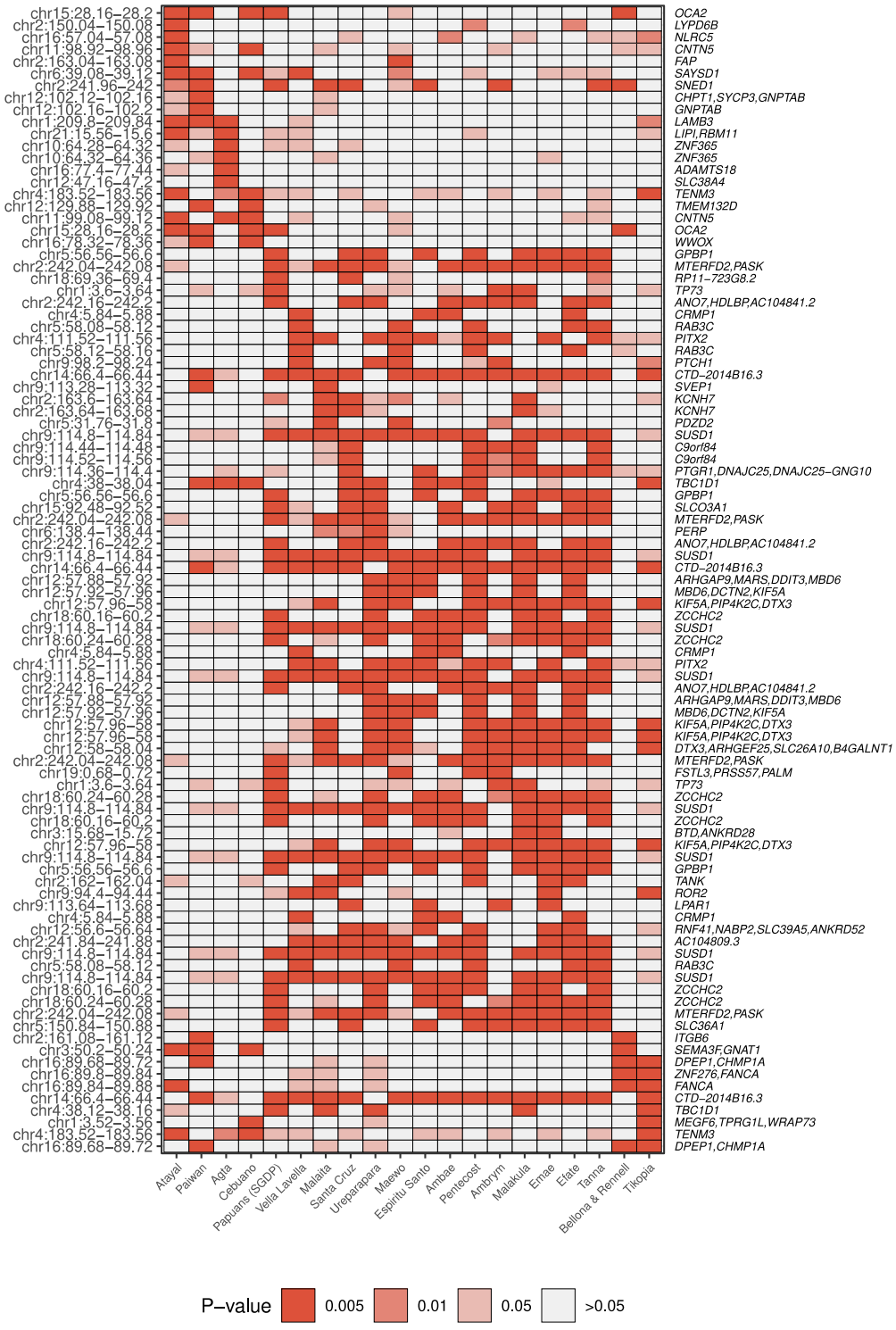
We identified a number of novel hits for Neanderthal adaptive introgression in the Pacific (Fig. 4a), many of which were shared among populations of the same ancestry (Supplementary Fig. 70 and Supplementary Table 14). For example, we detected a Neanderthal-introgressed ~18kb-long haplotype at high frequency in Near and Remote Oceanians (ranging from ~20% to >60%), which encompasses the 5'-UTR and intronic region of the *KRT80* gene (Extended Data Fig. 5c; left panel). *KRT80* is a protein-coding gene that encodes a type II epithelial keratin. Keratins are intermediate filament proteins responsible for the structural integrity of epithelial (skin) cells[99]. In accordance with a Neanderthal origin, the derived allele of the top archaic-like SNP (aSNP) (rs2360653-C) is found at moderate frequencies in Europeans and South East Asians, at low frequencies in East Asians, and is absent in sub-Saharan Africans from the 1000 Genomes Project[43] (Extended Data Fig. 5c; middle panel). Among Oceanian populations, the highest frequency is observed in PNG, Remote Oceanians and the Agta (Extended Data Fig. 5c; right panel). Notably, this aSNP acts as an expression quantitative trait locus (eQTL) of *KRT80* in sun exposed skin tissue (*P*-value = 6.1×10^{-5}; GTEx data[100]). The putative introgressed C allele is associated with a lower expression. This observation is in line with a recent study showing that Neanderthal introgressed alleles influence disease risk, including skin lesions resulting from sun exposure (i.e., keratosis)[101]. Our findings provide further support to the notion that Neanderthal alleles have been adaptive in different human populations due to their effects on skin[87,88,102], but it remains unclear how Oceanian populations benefited from Neanderthal alleles that reduce the expression of keratinocyte-related genes.

Another pertinent example of Neanderthal adaptive introgression in Oceanians was detected at the metabolism-related *TBC1D1* gene (Extended Data Fig. 5d), in line with other studies highlighting genes affecting lipid metabolism, type 2 diabetes risk, adipose tissue differentiation and body fat distribution[103,104]. *TBC1D1* transcripts have been reported to be highly expressed in skeletal muscle and adipose tissue[105] and are regulated through muscle contraction and energy depletion[106,107]. Furthermore, mutations at the human and murine *TBC1D1* have been associated with obesity[108-110]. In accordance with a Neanderthal origin, the derived allele of the top *TBC1D1* aSNP (rs2303423-C) is found at low frequencies in Europeans, moderate frequencies in South and East Asians, and absent in sub-Saharan Africans from the 1000 Genomes Project[43] (Extended Data Fig. 5d; middle panel). Among Oceanian populations, the top aSNP was found at the highest frequency in PNG (derived allele frequency [DAF] = 71%), Papuan-related Remote Oceanians (DAF ranging from ~30 to 60%), and, notably, in Polynesian outliers from Tikopia, where it is almost fixed (DAF = 90%) (Extended Data Fig. 5d; right panel). None of the aSNPs present in the Neanderthal introgressed haplotype have been reported to be associated with any trait by GWAS (as of March 27, 2020). Future functional studies should help to clarify the effect of this novel candidate introgressed variant on metabolic or obesity-related traits, given the growing health concern that obesity represents in this region of the world[111].

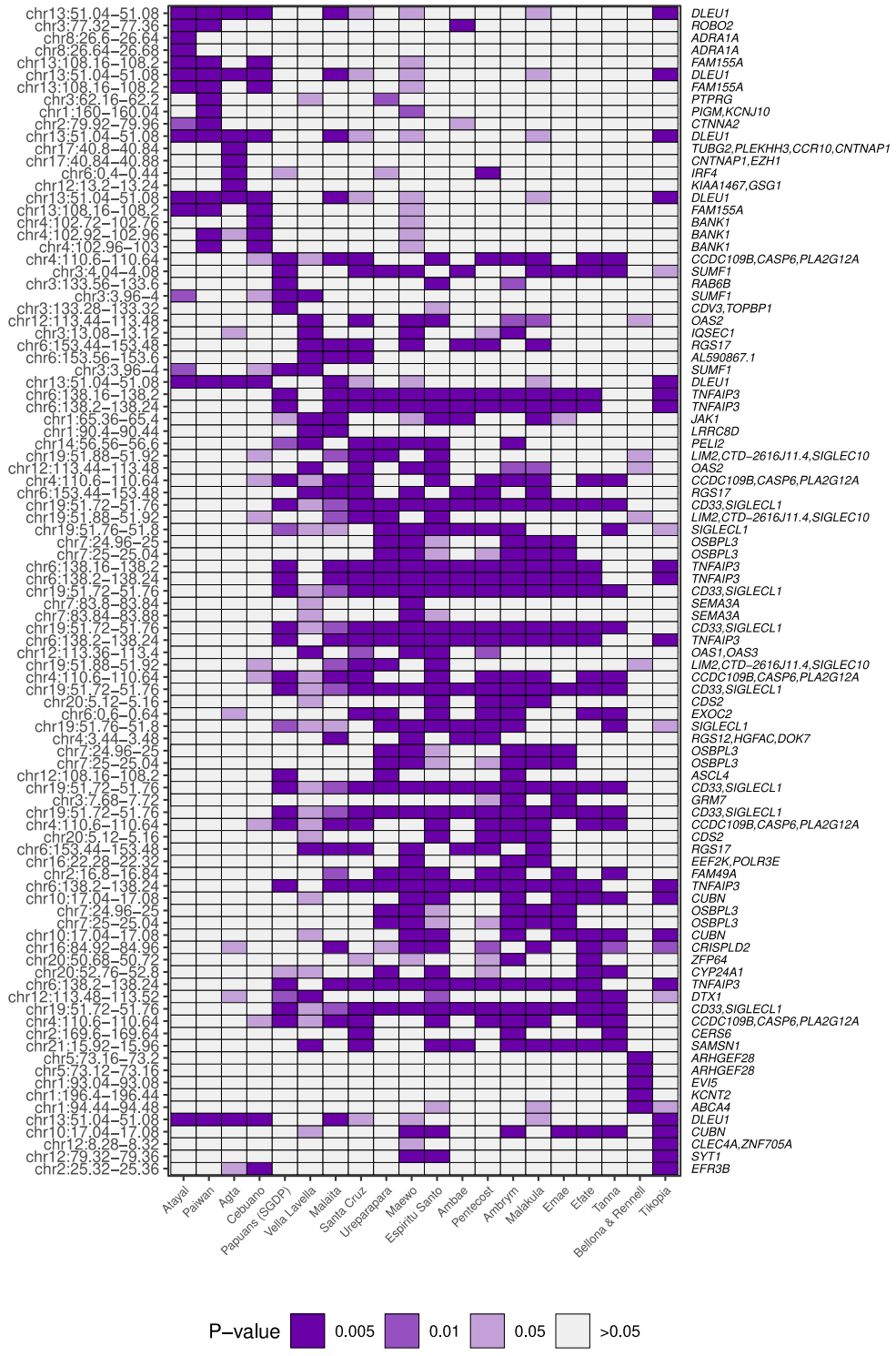
We also identified a number of novel hits of adaptive introgression from Denisovan (Fig. 4a), which were often shared among populations of the same ancestry (Supplementary Fig. 71 and Supplementary Table 15). For example, we found a ~30kb-long introgressed haplotype at *CD33*, which plays a key role in mediating cell-cell interactions and maintaining immune cells in a resting state[112]. This haplotype contains 7 high-frequency variants, including a non-synonymous variant predicted to be deleterious (rs367689451-A; SIFT score = 0) that is >66% frequency and is restricted to Oceanians (Extended Data Fig. 5a). We found that the frequency of Denisovan introgressed alleles at *CD33* were significantly higher than that of other genome-wide Denisovan-introgressed SNPs, in all groups independently of their levels of Papuan-related ancestry (Mann-Whitney U test; *P*-value$<1.5\times10^{-6}$), indicating that high Papuan ancestry is unlikely to explain the Denisovan adaptive signal at this locus. We also detected a strong signal at *IRF4*, which presents a ~29kb-long haplotype with 13 high-frequency variants at 64% frequency in the Agta (Extended Data Fig. 5b). IRF4 regulates interferon responses to viral infections and Toll-like-receptor signalling[113]. Furthermore, we identified a ~78kb-long introgressed haplotype at high frequency among Near and Remote Oceanian populations (ranging from ~30% to >50%), which encompasses

the 5'-UTR and intronic region of the *JAK1* gene (Extended Data Fig. 5e), a key mediator of cytokine signalling during important developmental, immune, and inflammatory responses[114,115]. The archaic allele with the highest derived frequency (rs368334238-A), which is located in the first intron of *JAK1*, is absent from all Africans and Eurasians, consistent with an introgression event from Denisovans into the common ancestors of Oceanian populations.

Another Denisovan adaptively introgressed signal includes a ~37kb-long haplotype, encompasses the *BANK1* gene, at high frequencies among populations with East Asian-related ancestry (ranging from ~15% to 37%) (Extended Data Fig. 5f; left panel). *BANK1* encodes a B-cell-specific scaffold protein that functions in B-cell receptor-induced calcium mobilization from intracellular stores[116]. Several variants in the *BANK1* gene have been associated with systemic lupus erythematosus (SLE), a prototypical autoimmune disease characterized by loss of immune tolerance to nuclear and cell surface antigens[117]. Although GWAS of SLE have been conducted in East Asians[118-120], none of the associated variants reported in the GWAS catalog included any of the high-frequency introgressed variants (as of April 1, 2020). In accordance with a Denisovan introgression event, the derived allele of the top archaic SNP (aSNP) (rs17031656-T) is absent among Africans and Europeans, and is only present in Southeast Asians from the 1000 Genomes Project[43] (Extended Data Fig. 5f; middle panel). Among Pacific populations, its highest frequency is found in the Cebuano (DAF = 37%) and Agta (DAF = 14%) from the Philippines, and the Atayal and Paiwan (DAF~15%) from Taiwan. In the remaining populations, its frequency is < 5%, being completely absent in PNG (Extended Data Fig. 5f; right panel). This frequency distribution suggests that this variant has been acquired through a Denisovan population that introgressed exclusively in the ancestors of East Asians. Notably, several other studies have shown that archaic introgressed alleles can influence present-day risk of autoimmune diseases in humans. Recent examples include signals of Neanderthal introgression in the chemokine receptor (*CCR*) gene family constituting the risk alleles for celiac disease[121], in the *ZNF365D* gene that is associated with a higher risk of Crohn's disease[87], and in the *TLR6-1-10* gene cluster that has been associated with greater susceptibility to allergies[122,123]. In light of this, the introgressed signal of Denisovan origin at *BANK1* may represent another case of evolutionary mismatch in modern humans, i.e., alleles that were beneficial in the past have become detrimental after important environmental changes[121-123].



Supplementary Figure 70. Genomic regions showing the strongest evidence of adaptive introgression from Neanderthal. Each row is a 40-kb window, each column is a Pacific population, and each cell is coloured according to whether the window is in the top 0.5%, 1%, 5%, or >5% of the U and $Q95$ statistics empirical distributions. The 5 most extreme genomic windows detected in population groups (Fig. 4a) are shown, and the genes within each window are shown on the right of each row.



Supplementary Figure 71. Genomic regions showing the strongest evidence of adaptive introgression from Denisovan. Each row is a 40-kb window, each column is a Pacific population, and each cell is coloured according to whether the window is in the top 0.5%, 1%, 5%, or >5% of the U and $Q95$ statistics empirical distributions. The 5 most extreme genomic windows detected in population groups (Fig. 4a) are shown, and the genes within each window are shown on the right of each row.

Supplementary Note 15: Gene Enrichment in Archaic Introgression

Introgressed haplotypes of archaic origin

To conduct enrichment analyses of Neanderthal and Denisovan introgressed haplotypes in gene set categories, we merged our Pacific populations into three population groups, based on their shared ancestry according to PCA and ADMIXTURE results (Supplementary Note 3). This included (i) a 'Papuan group' that consists of populations with high Papuan-related ancestry: PNG (SGDP samples only[17]), Solomon Islands, Santa Cruz, and the Vanuatu archipelago; (ii) an 'East Asian group' that consists of populations with high East Asian-related ancestry: East Asians (SGDP samples only[17]), Taiwanese indigenous peoples, Philippines (Cebuano), and Polynesian outliers; and (iii) the Philippine Agta. We then defined introgressed haplotypes in each population group, as the union of high-confidence introgressed haplotypes of Neanderthal or Denisovan origin, identified in each population that forms the particular population group (Supplementary Note 8). The union of introgressed haplotypes can therefore be thought of as a tiling path of inferred Neanderthal or Denisovan haplotypes among each population group.

Controlling for confounding factors

To establish that archaic introgression, rather than other factors, are driving the enrichment at a given gene set category, it is important to define the genomic features that can affect the occurrence of introgressed haplotypes across the genome. Based on a recent study[124], we considered the following genomic features: (i) recombination rate[43], (ii) density of conserved elements across mammals identified by PhastCons[125], (iii) density of regulatory elements based on the DNase I segments cumulated across all ENCODE cell types[126], (iv) deleteriousness based on CADD scores[127], and (v) number of SNPs.

For each autosomal protein-coding gene (Ensembl genes)[128], these genomic features were measured within 50-kb windows at the genomic centre of each gene, with the exception of the recombination rate, which was measured within 200-kb windows centred on genes. The reason to use 200-kb windows for the recombination rate estimates is that the sparsity of sites within the 1000 Genomes Phase 3 genetic map[43] would have resulted in recombination rate estimates based on few sites. We only considered genes with a recombination rate higher than 0.0005 cM/Mb to distinguish between genes where the recombination rate is 0 and genes within gaps in the genetic map. Deleteriousness was measured using the mean value of CADD scores in each gene. Genes that contained less than 5 sites with a genetic map position, less than 5 sites with associated CADD scores, or less than 5 SNPs were discarded.

Resampling-based enrichment analysis

We devised a resampling-based enrichment test for a given gene set (i.e. a set of genes composing a particular biological pathway) using a set of 'control' genes that were matched for all genomic features described above, to obtain empirical null distributions. Specifically, we matched each gene for all aforementioned genomic features based on quartiles (i.e. each gene was placed into one of four bins for each genomic feature). In doing so, each gene had a list of control genes with similar genomic features. As some genes can only have a small number of matching control genes (note that, by partitioning five genomic features into quartiles, we have $5^4 = 625$ possible bin combinations), we selected for further analysis only those genes with at least three matching control genes. The values that defined the quartiles for each genomic feature can be found in Supplementary Table 16. For a given gene set, we then estimated the number of genes that overlapped introgressed haplotypes. For each gene that composed this gene set, we then randomly sampled a control gene to obtain a control gene set of the same length. We repeated this resampling 100,000 times to obtain resampling *P*-values. *P*-values were calculated by counting the proportion of resamples where the number of control genes that overlapped with introgressed haplotypes were higher than, or equal to, the value observed for the tested gene set. All *P*-values were then adjusted

for multiple testing by the Benjamini-Hochberg method, to account for the number of gene sets tested. Gene sets with an adjusted *P*-value < 0.05 were considered as significantly enriched.

Enrichment analysis of adaptively introgressed genes

To test whether gene set categories are enriched for archaic adaptive introgression, we intersected the introgressed haplotypes with significant U and $Q95$ genomic windows (Supplementary Note 14). We considered adaptively introgressed haplotypes, in each population group, as introgressed haplotypes that significantly overlapped the Neanderthal or Denisovan U and $Q95$ genomic windows identified in each population that forms that population group. For this analysis, we considered adaptively introgressed U and $Q95$ genomic windows as those in the top 5% of their respective distribution. Note that this subset of introgressed haplotypes is composed of introgressed haplotypes at high frequency, a hallmark of positive selection. We then carried out the resampling-based enrichment test using only adaptively introgressed haplotypes as described in the section above.

Gene set categories

We considered the following gene set categories for the resampling-based enrichment analysis: (i) KEGG[129], (ii) Wikipathways[130], (iii) the GWAS catalog[131], and (iv) Gene Ontology (GO) (including biological process, molecular function, and cellular component)[132]. For the GO enrichment analysis, we restricted the list of GO terms to those between levels 3 and 7, to avoid redundancy. Furthermore, we analysed additional gene set categories, including 1,553 manually-curated genes involved in innate immunity[123], and 1,257 genes whose products are known to have physical interactions with viruses (VIPs)[124]. To limit the effect of genomic clusters of genes on the enrichment analysis, we only retained genes that were less than 200kb apart from the centre of other genes present in a given gene set category. In practice, for each gene set category, we calculated the distance between the gene centre for all pairs of genes, and removed the gene that had the highest number of genes within 200 kb from its genomic centre. Gene set categories with less than 10 genes after this procedure were discarded.

Results

In Papuan-related populations, we detected a significant enrichment in introgressed haplotypes at genes associated with BMI and obesity-related traits (FDR *P*-value < 0.05; Supplementary Table 17), suggesting preferential retention of archaic alleles in pathways related to lipid metabolism. In line with a role of archaic introgression in immune responses[123,124], we found that VIPs[124] were enriched in Denisovan-introgressed genetic material, and genes affecting 'immune response to measles' were enriched in signals of adaptive Neanderthal introgression (Supplementary Tables 17 and 18). In East Asian-related populations, genes affecting 'apoptotic cellular response to stress' and 'cancer' were enriched in Neanderthal ancestry and signals of adaptive introgression. Furthermore, we found an enrichment of adaptive Denisovan-introgressed genetic ancestry among genes related to 'sleep duration', presumably because of adaptation to daytime variation with latitude (Supplementary Table 17 and 18). Lastly, in the Philippine Agta, we found an enrichment of Neanderthal adaptive introgression at several pathways related to general cellular functions, and notably, an enrichment of Denisovan adaptive introgression at genes associated to obesity-related phenotypes (Supplementary Table 18).

Supplementary Note 16: Genome Scans for Classic Sweeps

Rationale
We searched for signatures of positive selection under the classic sweep model, by considering five different analyses (Supplementary Fig. 72 and Supplementary Table 19), which broadly correspond to different branches of the population tree where positive selection may have occurred.

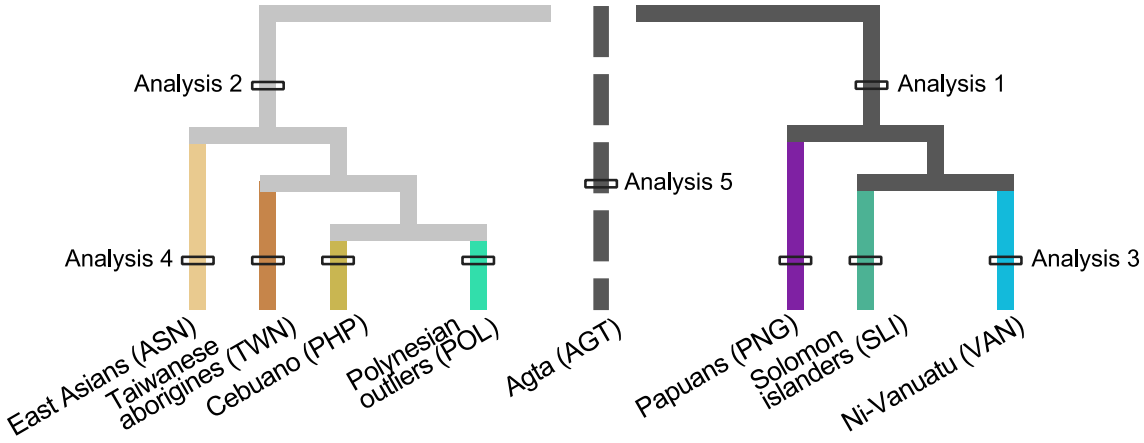
- *Analysis 1:* Detection of positive selection occurring in the **ancestral population of Oceanians** (populations with predominantly Papuan-related ancestry). To identify these signals, we searched for classic sweeps that are common to populations from PNG (SGDP samples[17]), the Solomon Islands and Vanuatu (note that populations from the Bismarck archipelago alone were not included in the analysis because it was not permitted by the informed consent they signed). We computed the inter-population statistics described below for each population separately: PNG, Solomon islanders and ni-Vanuatu. We used a pool of all populations with East Asian ancestry as reference population (East Asians, Taiwanese indigenous peoples and Philippine Cebuano, with the exception of Polynesian outliers), and sub-Saharan African or European samples as outgroups.

- *Analysis 2:* Detection of positive selection occurring in the **ancestral population of East Asians** (populations with predominantly East Asian-related ancestry). We followed a similar strategy as above. We also computed the inter-population neutrality statistics for each population separately: East Asians, Taiwanese indigenous peoples, Philippine Cebuano and Polynesian outliers. We used a pool of all populations with high Papuan-related ancestry as reference population (i.e., PNG, Bismarck, Solomon and Vanuatu islanders), and sub-Saharan African or European samples as outgroups.

- *Analysis 3:* Detection of positive selection occurring in **each specific population with high Papuan-related ancestry**. We compared each test population (PNG, Solomon Islands and Vanuatu Archipelago) to a reference population composed of a pool of all populations with high Papuan-related ancestry (PNG, Bismarck, Solomon and Vanuatu islanders), excluding the test population. Sub-Saharan African or European populations were used as outgroups.

- *Analysis 4:* Detection of positive selection occurring in **each specific population with high East Asian ancestry**. We compared each test population (East Asians, Taiwanese, Cebuano, and Polynesian outliers) to a reference population composed of a pool of all populations with high East Asian-related ancestry (East Asians, Taiwanese indigenous peoples, Cebuano, and Polynesian outliers) excluding the test population. Sub-Saharan African or European populations were used as outgroups.

- *Analysis 5:* Detection of positive selection occurring in **the Philippine Agta population**. To identify selection signals in this population, we used inter-population neutrality statistics with, as reference populations, either a pool of the Papuan-related or the East Asian-related populations. Sub-Saharan African or European populations were used as outgroups.



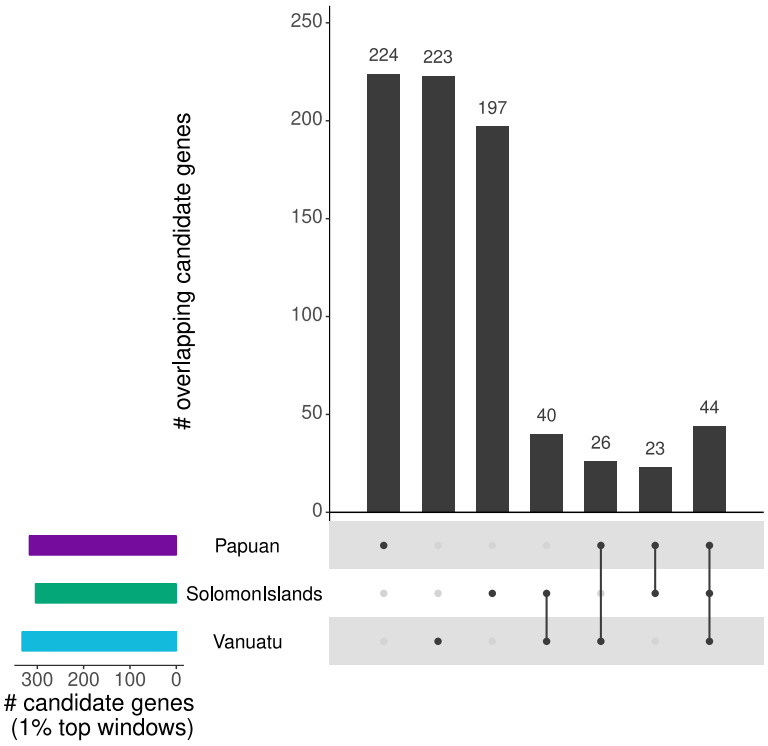
Supplementary Figure 72. Rationale used for the analyses of classic sweeps.

Methods

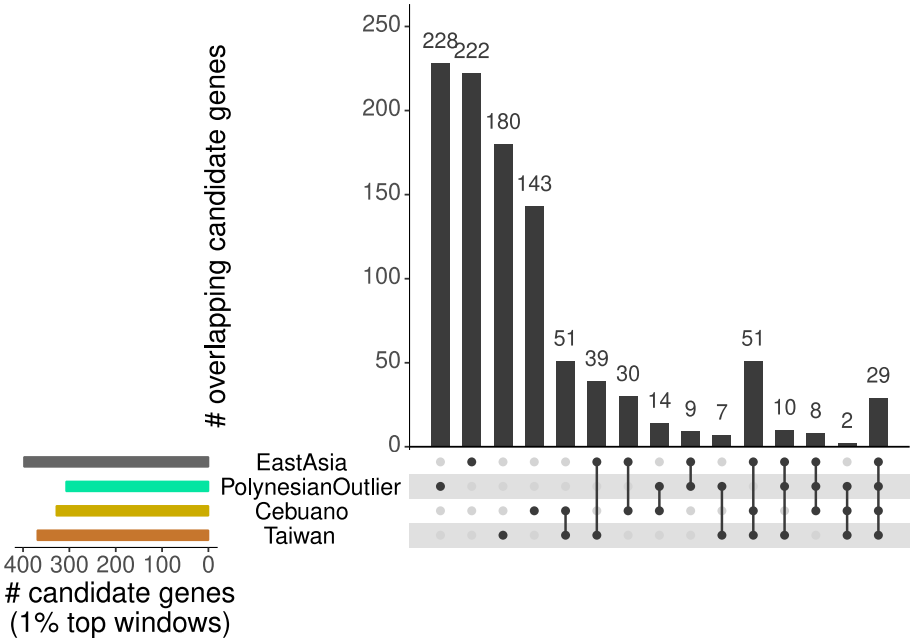
We scanned the genome for candidate loci under positive selection using the inter-population LSBL (Locus-specific branch lengths)[133] and XP-EHH (cross-population extended haplotype homozygosity)[134] statistics, combined into a Fisher's score (F_{CS}). We estimated the F_{CS} as the sum of the $-\log_{10}$(percentile rank of the statistic for a given SNP), for all the inter-population statistics. We defined outlier SNPs as those with an F_{CS} among the 1% highest of genome. Putatively selected regions were defined as genomic windows that show a proportion of outlier SNPs (i.e., number of outliers SNPs/total number of SNPs in the window) among the 1% highest of the genome, after partitioning all windows into five bins based on the number of SNPs. The test and reference populations (both for XP-EHH and for LSBL) and the outgroup populations (for LSBL) were defined for each analysis as described in Supplementary Table 19 and Supplementary Fig. 72. We estimated AMOVA-based F_{ST} to compute LSBL, and XP-EHH was computed in 100-kb sliding windows with a 50-kb step. The derived alleles were determined using the 4-way EPO ancestral sequence from the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/retired_reference/ancestral_alignments/). We normalized the XP-EHH scores in 40 separate bins of derived allele frequency. We kept only windows with >50 SNPs, and removed 500kb around gaps. Neutrality statistics were computed with the optimized, window-based algorithms implemented in *selink* (www.github.com/h-e-g/selink).

Results

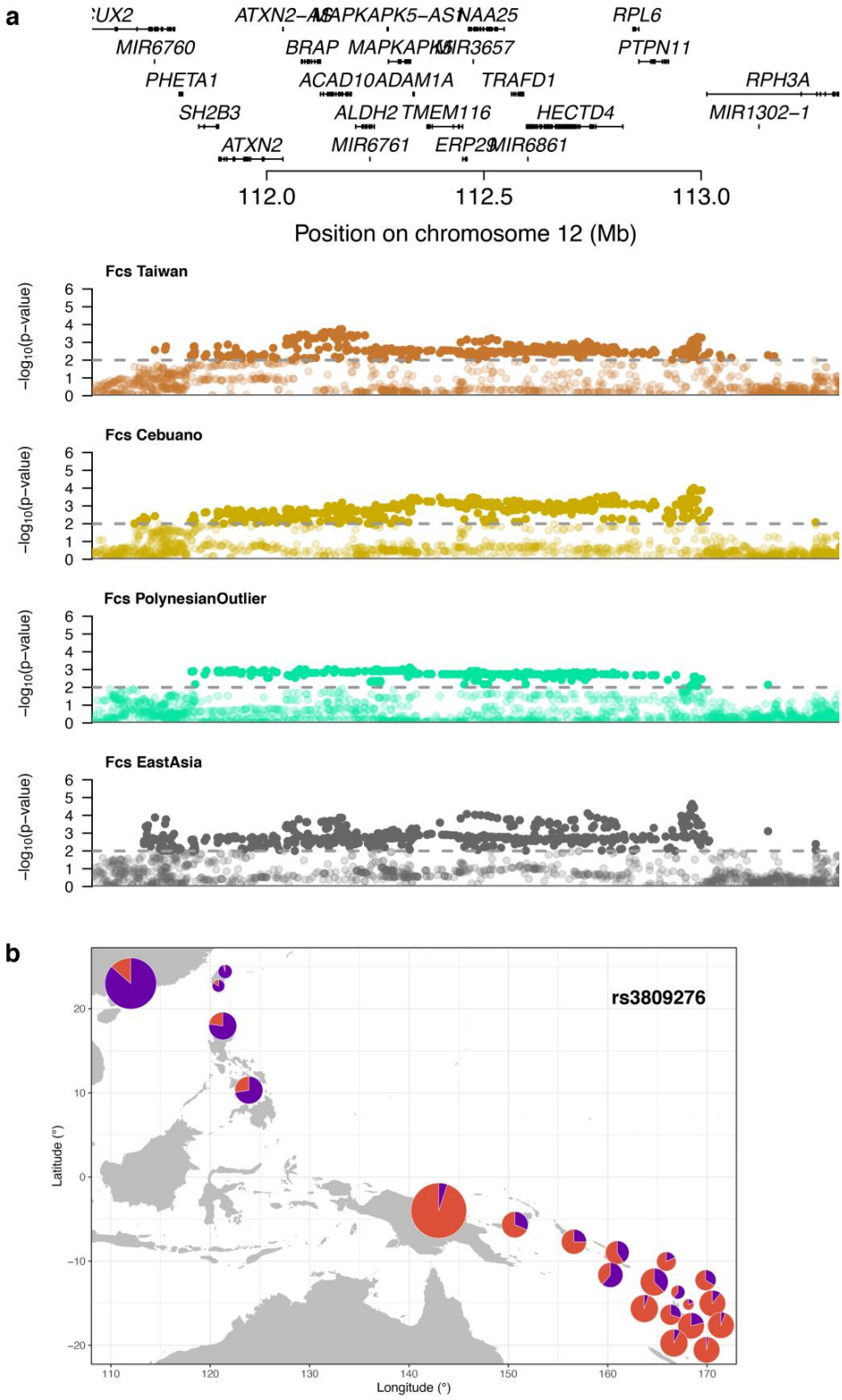
We found 44 candidate genes shared among the three different Papuan-related populations from Near and Remote Oceania (i.e., *Analysis 1*; Extended Data Fig. 6, Supplementary Figs. 72 and 73 and Supplementary Table 20), among which the strongest hit (*P*-value < 0.001) overlaps the *GABRP* and *RANBP17* genes (Extended Data Fig. 8a,b). We detected 29 candidate genes that were shared between the four East Asian-related populations (i.e., *Analysis 2*; Extended Data Fig. 9, Supplementary Figs. 72 and 74 and Supplementary Table 23). The shared region with the highest selection scores (*P*-value < 0.001) overlaps a ~1Mb-haplotype encompassing multiple genes (Supplementary Fig. 75), including *ALDH2*. Among *ALDH2* variants, the derived allele at rs3809276 is observed at >60% in East Asians, Taiwanese indigenous peoples, Philippine Cebuano and Polynesian outliers, while being at < 15% in Papuan-related groups. ALDH2 deficiency results in adverse reactions to alcohol consumption and is associated with increased survival in Japanese[135].



Supplementary Figure 73. Number of candidate genes for positive selection shared among populations of Papuan-related ancestry.



Supplementary Figure 74. Number of candidate genes for positive selection shared among populations of East Asian-related ancestry.



Supplementary Figure 75. Top candidate region for a classic sweep shared between four East Asian-related populations. **a**, The genomic region shows significant windows (*P*-value < 0.001) in the four populations tested. Each point in the Manhattan plot represents a SNP. The y axis shows the $-\log_{10}$(*P*-value) of the Fisher score for each SNP. **b**, Population frequencies of the rs3809276 derived allele (in purple) in Pacific populations. The map was generated using the *maps* R package.

Among population-specific signals, one of the strongest signals was observed in Solomon islanders at *ATG7* (Supplementary Table 21), which regulates cellular responses to nutrient deprivation[136], and has been associated with blood pressure[137]. Putatively selected variants at *ATG7* reach ~70% frequency in Solomon islanders, 10% in Papuans and < 5% worldwide. Another strong population-specific hit was detected at *LHFPL2* in Polynesian outliers (Extended Data Fig. 8c,d and Supplementary Table 24); variation in *LHFPL2* is associated with eye macula thickness — a highly variable trait across populations that is responsible for sharp vision[138]. *LHFPL2* variants reach ~80% frequency in Polynesian outliers only, in particular those from Rennell and Bellona, and are absent from current databases. In the Philippine Agta, the second strongest hit was detected at *DLEU1* (Extended Data Figs. 6d and Supplementary Table 22), which also showed a signal of adaptive Denisovan introgression[92] (Extended Data Fig. 7). Putatively-selected *DLEU1* variants (*P*-value < 0.002) are >83% frequency in the Agta and <50% in other Pacific populations, and include 5 high-frequency aSNPs likely introgressed from Denisova. Genetic variation at this locus is strongly associated with height[139] and waist-hip ratio[140], suggesting positive selection for introgressed archaic variants affecting height in the Agta from the Philippines.

Supplementary Note 17: Signals of Adaptive Admixture

Rationale

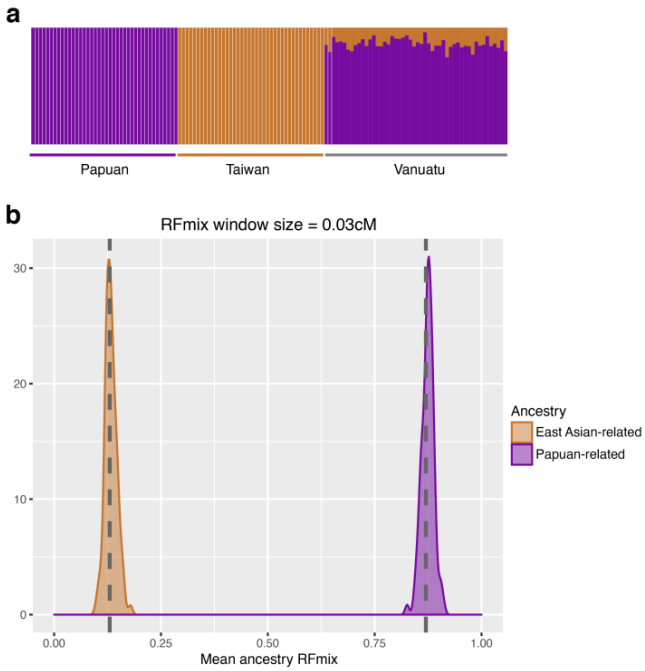
Several studies have provided empirical evidence that adaptive introgression – the acquisition of adaptive traits through hybridization with closely-related species – is a possible source of adaptive variation[141-145]. We and others have recently showed that, besides introgression, gene flow can also promote adaptation by spreading beneficial alleles between populations of the same species, a process called 'adaptive admixture' or 'adaptive gene flow'[25,146-148]. Because Oceanian populations globally result from pervasive admixture, to different extents, between populations of Papuan-related and East Asian-related ancestry, we sought to test if admixed Near and Remote Oceanians (i.e., here populations from the Solomon Islands and Vanuatu) have acquired advantageous alleles via gene flow from East Asian-related populations (here Taiwanese indigenous peoples and Philippine Cebuano). We also tested if Polynesian outliers have acquired advantageous alleles via gene flow from Papuan-related groups. We used deviations in local ancestry, combined with signatures of positive selection in parental populations, to identify examples of variants under putative adaptive admixture, following the procedure described in ref.[25].

Local ancestry simulations

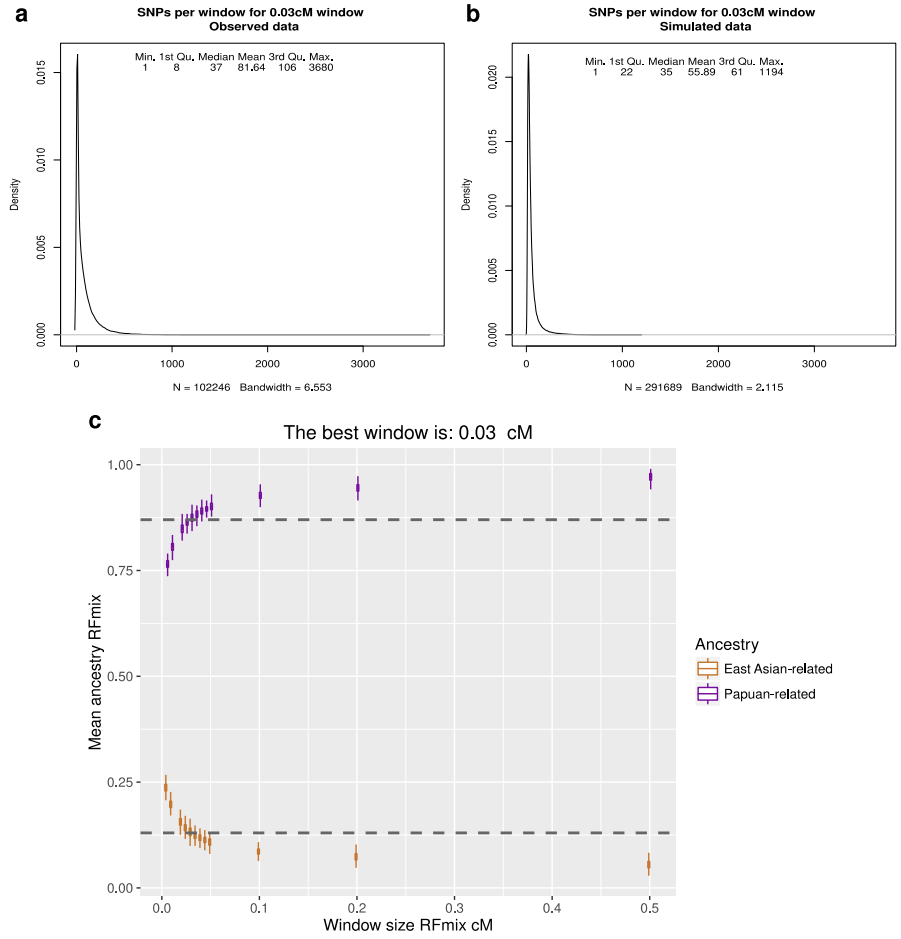
We first used simulations with *fastsimcoal2*[46] to estimate the number of standard deviations (SD) in local ancestry that is expected under neutrality, and to estimate the window size for local ancestry inference with RFMix v.1.5.4 (ref.[83]). We used the demographic model for western Remote Oceanians described in Supplementary Note 4 (Extended Data Fig. 2b and Supplementary Table 5) to simulate 20 diploid individuals representing the Paiwan, 20 for the Atayal, 40 for PNG and 50 admixed individuals from Vanuatu. We simulated 500 windows of 100-kb each, concatenated in one chromosome, giving a total of 50Mb per chromosome. We set a constant mutation rate of 1.25×10^{-8} mutation/generation/site[17,51]. The recombination rate for each simulated window was estimated by averaging the recombination rate from random 100-Kb windows sampled in the 1000 Genomes Phase 3 genetic map[43]. We kept only biallelic sites and alleles with a MAF > 0.01 and ran 100 simulations (Supplementary Fig. 76). We estimated local ancestry with RFMix, for 50 admixed, simulated samples from the Vanuatu using, as parental sources, a population composed of 40 East Asian (20 Atayal and 20 Paiwan) and 40 PNG samples. We used the TrioPhased algorithm implemented in RFMix assuming T_{adm} = 50 generations, 3 EM iterations, a minimum number of 5 reference haplotypes per tree node and different runs for window lengths of 0.005, 0.01, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.1, 0.2 and 0.5 cM, to test for the optimal value (Supplementary Fig. 77).

We selected the window length for which the mean ancestry estimated from RFMix was the closest to the admixture proportions estimated by ADMIXTURE[33]. We reasoned that East Asian-related ancestry could be underestimated when using large windows in RFMix, because East Asian admixture proportions in western remote Oceanians are low (13% in the Malakula population, i.e., the ni-Vanuatu population used to represent western Remote Oceanians in the demographic model; Supplementary Note 4). Indeed, a large window in the genome of the ni-Vanuatu will typically include few, small East Asian ancestry segments and many more Papuan ancestry segments, so RFMix will preferentially assign this window to the Papuan major ancestry. To determine the significance threshold of deviations in local ancestry, we estimated the number of false positives from the simulated data, according to the number of SDs in local ancestry considered. Namely, as the simulations are neutral and therefore no selection signals are expected, the proportion of loci with local ancestry higher or lower than the genome-wide average $\pm\, x$ SD is considered as an estimate of our FPR. We calculated the FPR for different x values for each ancestry separately. We estimated a FPR = 1% at $\pm\, x$ = 2.97 or 3.58 SD for Papuan and East Asian ancestries, respectively (Supplementary Fig. 78), therefore we set a threshold of $x \pm 3$ SD for both ancestries. We also checked whether our approach was impacted by phasing errors. We estimated our FPR

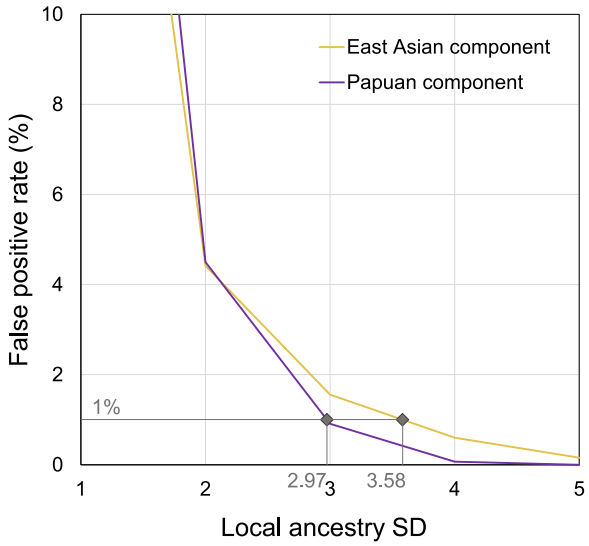
using the same simulations, except that we created unphased diploid individuals from the simulated haploid data, and phased simulated samples using SHAPEIT2 (refs.[84,85]) with the same parameters as for the observed data. RFMix was run on the phased simulated data using the same parameters as before, except that we used the PopPhased algorithm. Under these conditions, we estimated a FPR = 1% at $\pm$ *x* = 2.83 or 2.84 SD for Papuan and East Asian ancestries, indicating that our approach has low FPR, even in the presence of phasing errors.



Supplementary Figure 76. Genetic ancestry analyses of a representative simulation of the parental and admixed populations. **a**, ADMIXTURE clustering analysis, **b**, Distribution and mean (dashed line) of the genome-wide ancestry of 50 simulated Vanuatu samples, based on local ancestry inference by RFMix performed for 100 simulations, with a window size of 0.03 cM.



Supplementary Figure 77. Simulation-based estimation of RFMix parameters. Number of SNPs per window considering a window size of 0.03 cM in **a**, the observed data and **b**, the simulated data. **c**, Mean ancestry estimated with RFMix for different window sizes (in cM) in the simulated data. The window size for which RFMix estimates ancestry proportions closer to the simulated value (dashed line) is considered the best (0.03 cM).



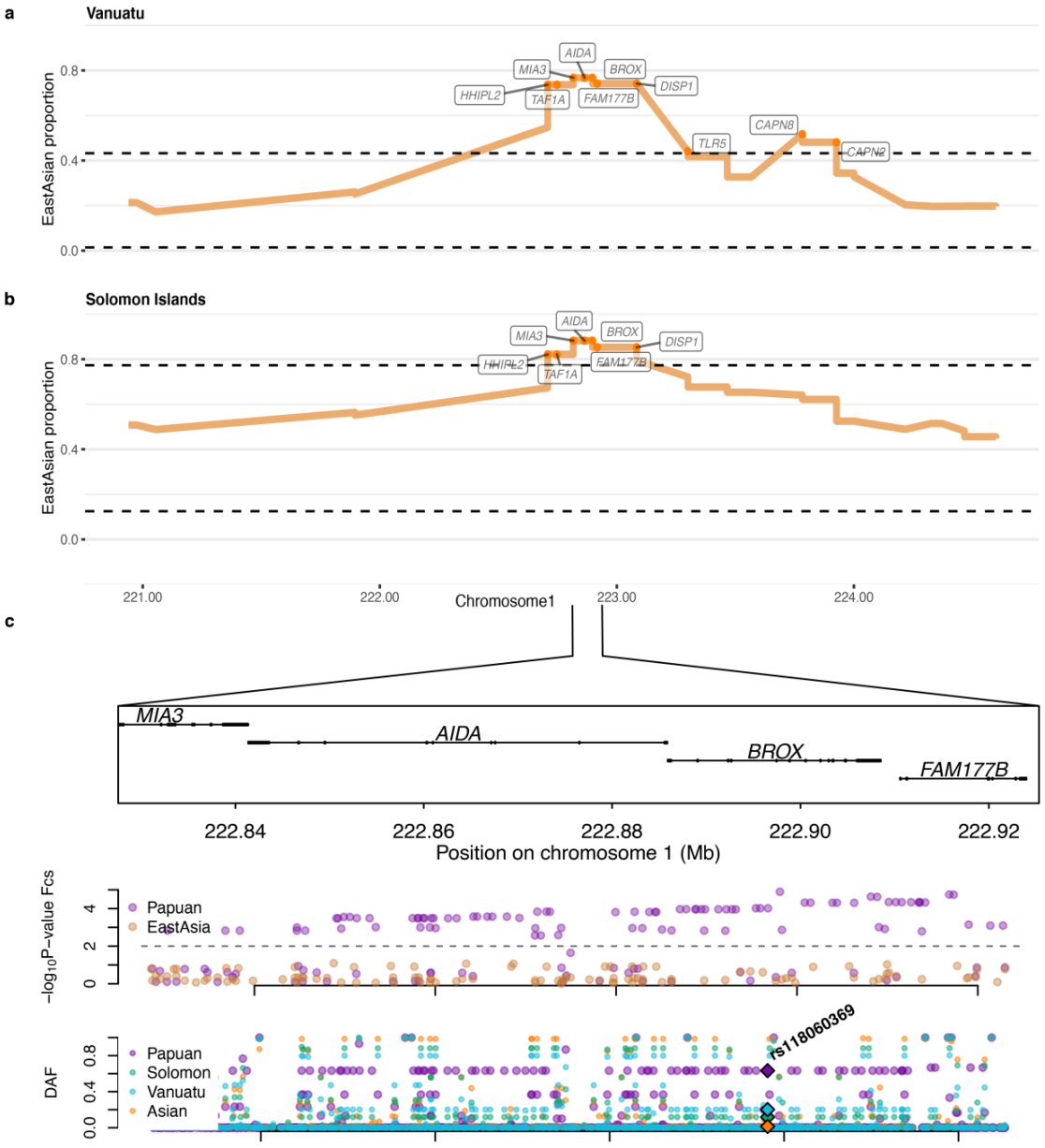
Supplementary Figure 78. False positive rates (FPR) for a genome scan of adaptive admixture. FPR was estimated based on neutral simulations. Results are shown for different significance thresholds, based on standard deviations (SD) from the genome-wide average of local ancestry.

Local ancestry inference

To estimate local ancestry in the genomes of the studied populations, we used RFMix v.1.5.4 (ref.[83]), with the same parameters as for the simulations, but allowing for phase correction (using the PopPhased algorithm implemented in RFMix) and fixing the window length to the optimal value of 0.03 cM (Supplementary Fig. 77). We set two groups of parental populations, one for the Papuan-related ancestry and another for the East Asian-related ancestry. In the first group, we included all the populations in the dataset from PNG (i.e., Bundi, Kundiawa, Marawaka, Mendi, Tari and Papuan_SGDP from refs.[16,17]). In the second group, we included Taiwanese indigenous peoples (Atayal and Paiwan Ami_SGDP, and Atayal_SGDP from ref.[17]) and the Cebuano from Philippines. The admixed populations for which we computed local ancestry were those from the Solomon Islands (Vella Lavella and Malaita), Vanuatu (Ureparapara, Santo, Malakula, Ambae, Maewo, Pentecost, Ambrym, Emae, Efate and Tanna), Polynesian outliers (Rennell, Bellona and Tikopia), and the Philippine Agta. Populations in each of the four population groups were analysed together. We kept only SNPs with a MAF > 0.01, leaving a total of 7,875,602 SNPs. We removed sites with a posterior probability lower than 0.9 from the local ancestry results. We also excluded centromeres, based on UCSC annotations[149], and 2Mb from the telomeres of each chromosome. To identify deviations in local ancestry, we estimated the proportion of ancestry in 100-Kb windows.

Results

No signals of post-admixture selection were detected at genes with classic sweep signals, such as *RANPB17*, *GABRP* and *ALDH2*, supporting the ancient nature of these selection events (Supplementary Table 25). We observed a unique, significant increase in East Asian ancestry among Vanuatu and Solomon islanders at the *BROX* gene (Supplementary Fig. 79). This gene includes a Bro1 domain that participates in the virus budding machinery, by interacting with the virus nucleocapsid and stimulating the production of virus-like particles[150]. Intriguingly, *BROX* showed a strong, classic sweep signal only in PNG (Supplementary Fig. 79). This suggests strong, local adaptation of PNG after their divergence from other Oceanians, resulting in PNG being a poor proxy, at the locus, of the Papuan-related source population of admixed Oceanians, when performing local ancestry inference. Alternatively, this may suggest post-admixture selection for the East Asian haplotype in populations from the Solomon and Vanuatu Archipelagos.



Supplementary Figure 79. Putative local signal of adaptive admixture in admixed Oceanians. Local proportions of East-Asian ancestry at the candidate locus in admixed populations from **a**, the Vanuatu and **b**, the Solomon Islands. **c**, Genes of the genomic region that shows an excess of East-Asian ancestry (top panel). Local signal of positive selection for Papuans (Analysis 1) and East Asians (Analysis 2) at the candidate locus (middle panel). The y axis shows the $-\log_{10}$(*P*-value) of the combined Fisher score (F_{CS}). Each point is a SNP. Derived allele frequency of SNPs at the locus, in the admixed Oceanian populations and the parental populations (bottom panel). The SNP with the highest F_{CS} is highlighted (rs118050369).

Supplementary Note 18: Signals of Polygenic Adaptation

Rationale

Because the genetic architecture of most adaptive traits is expected to be polygenic[151,152], we searched for evidence, in Pacific populations, of directional selection on candidate traits whose genetic architecture has been well described by genome-wide association studies (GWAS). Building upon previous work[153,154], we used an approach that tests if the integrated haplotype scores (iHS) of trait-increasing alleles are significantly different from those of random SNPs with similar allele frequency. This approach does not rely on effect size estimates, which can be biased due to partial correction for population stratification, resulting in spurious signals of polygenic selection[155,156]. Instead, it relies on the assumption that alleles affecting traits are the same in Oceanians and Europeans, and, moreover, that these alleles affect traits in the same direction. In light of these assumptions, which are relatively strong, we used in parallel an independent approach that tests for the co-localization of selection signals and trait-associated genes; this window-based approach makes the assumption that the same genomic regions affect the traits of interest in all human populations.

SNP-based approach

Methods. We obtained GWAS summary statistics for 25 candidate traits from the UK Biobank database[157] (http://www.nealelab.is/uk-biobank), which are less biased by population stratification than previous GWAS[155,156]. Traits were considered of interest if they are related to morphology, metabolism and immunity, as these phenotypes are strong candidates for responses, through natural selection, to changes in climatic, nutritional and pathogenic environments. We first classified SNPs as increasing or decreasing the candidate trait, based on the sign of UK Biobank effect sizes (β), considering a significance threshold of *P*-value ≤ 5×10^{-8}. A negative β indicates that the alternate allele is trait-decreasing, while a positive value indicates that it is trait-increasing. We thus changed the sign of β values when the alternative allele was ancestral (and the reference was derived), so that the sign of β values indicates the effect of the derived allele on the trait of interest. Next, we computed iHS (iHS = $\ln(iHH_a/iHH_d)$) using *selink* (www.github.com/h-e-g/selink), for each SNP and population, and standardized scores in 100 bins of DAF. We then polarized iHS, following previous studies[153,154], so that positive iHS indicates directional selection of the trait-decreasing allele, while negative iHS indicates directional selection of the trait-increasing allele. To do so, we simply changed the sign of iHS for the derived alleles with a negative β. We called the resulting statistic the polarized trait-iHS (tiHS).

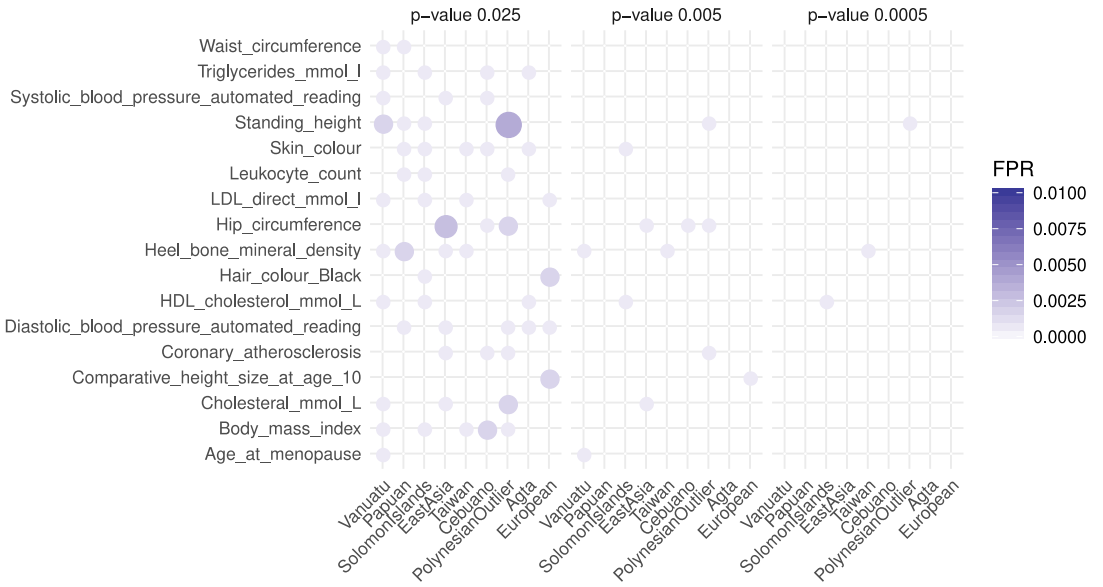
To test if a trait is under directional selection, we first kept, for each trait, trait-associated variants that are unlinked. Specifically, we partitioned the genome into 100-Kb non-overlapping contiguous windows, and kept for each window the variant with the lowest association *P*-value. We then compared the mean tiHS of the *x* independent, trait-associated alleles to the mean tiHS of 100,000 random samples of *x* SNPs with similar DAF, Genomic Evolutionary Rate Profiling (GERP) score[158], and surrounding recombination rate (based on 1000 Genomes phase 3 genetic map[43]), to account for the effects of background selection. GERP, recombination rate and DAF were grouped into 8 bins. We considered that directional selection has increased (or decreased) the trait if less than 2.5% (or 0.05% or 0.005%) of the resampled sets have a mean tiHS that is lower (or higher) than the observed tiHS, which we considered as empirical *P*-values. We adjusted *P*-values for multiple testing with the Benjamini-Hochberg method, to account for the number of traits and populations tested.

To estimate the FPR of our approach, we sampled 1,000 times *x* random genome-wide SNPs, *x* being the number of independent trait-associated alleles, and used the sampled SNPs as pseudo-data. We compared each of the 1,000 tiHS average values to a null distribution obtained by random sampling of *x* SNPs matched to pseudo-data. The FPR was estimated as the proportion, out of 1,000 pseudo-data, of tiHS average values that were within the 2.5%, 0.05% or 0.005% of the null distributions. We adjusted *P*-values for multiple

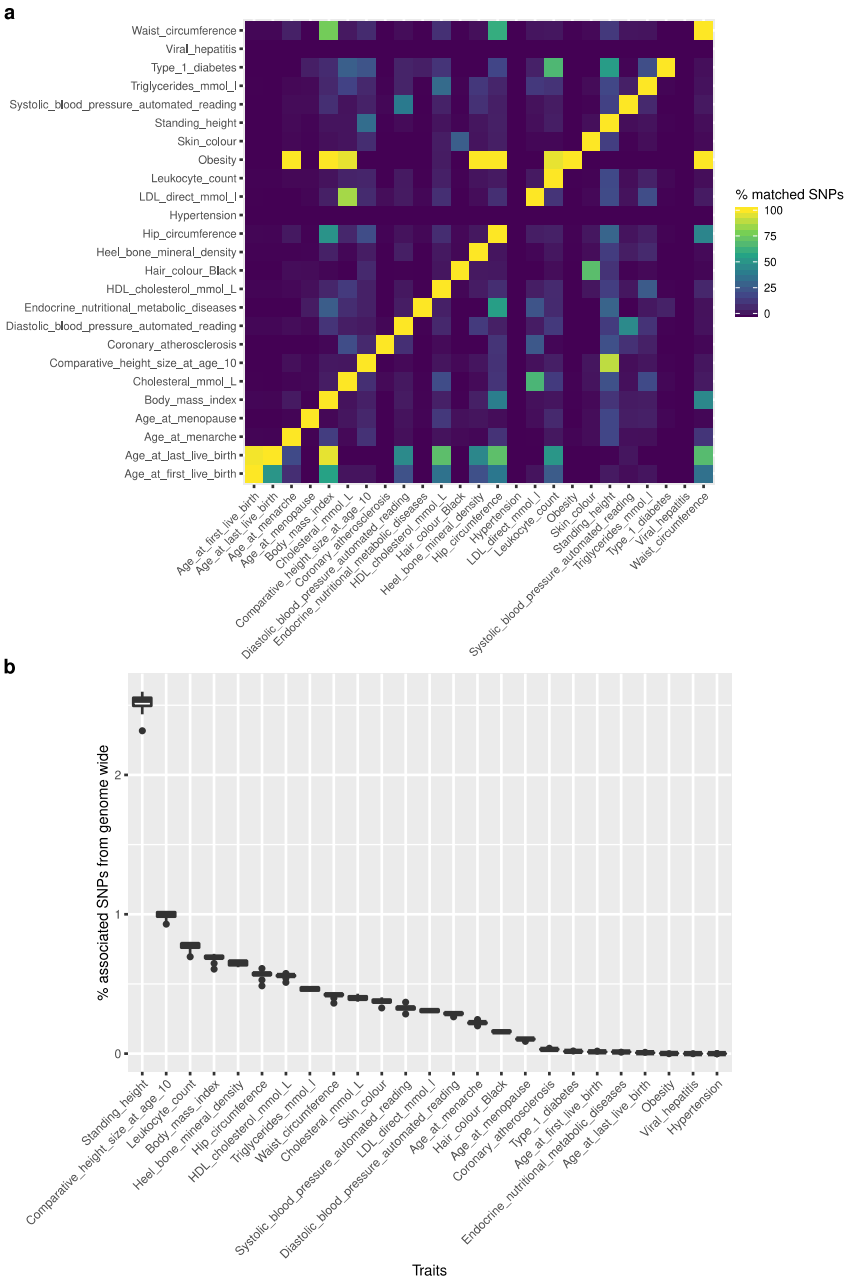
testing with the Benjamini-Hochberg method, to account for the number of traits and populations tested.

Results. We first estimated the FPR of our approach, based on resampling. The maximum FPR was 0.1% at *P*-value = 5×10^{-3} and 0.4% at *P*-value = 2.5×10^{-2} (Supplementary Fig. 80), which were thus used as the significance thresholds in subsequent analyses. As a positive control, we searched for signals of polygenic adaptation in European populations, where such signals have been extensively studied[153,155,156]. As expected, we found a signal of polygenic adaptation for lighter skin and hair pigmentation[153] and no signals for increased height[155,156] (Fig. 4b). We also identified a new signal for decreased cholesterol, which has not been previously reported. With respect to Pacific populations, we detected a signal for decreased BMI in Taiwanese indigenous peoples, and a unique, strong signal for decreased high-density lipoprotein (HDL) cholesterol in the Solomon Islands and the Vanuatu archipelago (Fig. 4b).

Because some of the traits tested for polygenic adaptation are pleiotropic, it is difficult to identify the specific trait that is adaptive. For Europeans, 70% of SNPs associated with hair colour were also associated with skin colour, suggesting that the two traits are highly pleiotropic (Supplementary Fig. 81). For East Asians, 47% of the SNPs associated with hip circumference were also associated with waist circumference. However, for Oceanians, <7% of variants associated with HDL levels were associated with other candidate traits, suggesting that pleiotropy plays a minor role in explaining these signals. Together, these findings support the occurrence of polygenic adaptation related to lipid metabolism in Oceanians, possibly in response to long-term fish consumption[159].



Supplementary Figure 80. Specificity of the SNP-based approach to detect polygenic selection. False positive rate (FPR) estimated based on 1,000 random samples of genome-wide SNPs used as pseudo-data. The *P*-value is obtained from the rank of the mean tiHS for resampled SNPs in a null distribution obtained by resampling. The FPR was estimated by counting the number of significant resamples at three different *P*-value thresholds: 0.025, 0.005 and 0.0005.



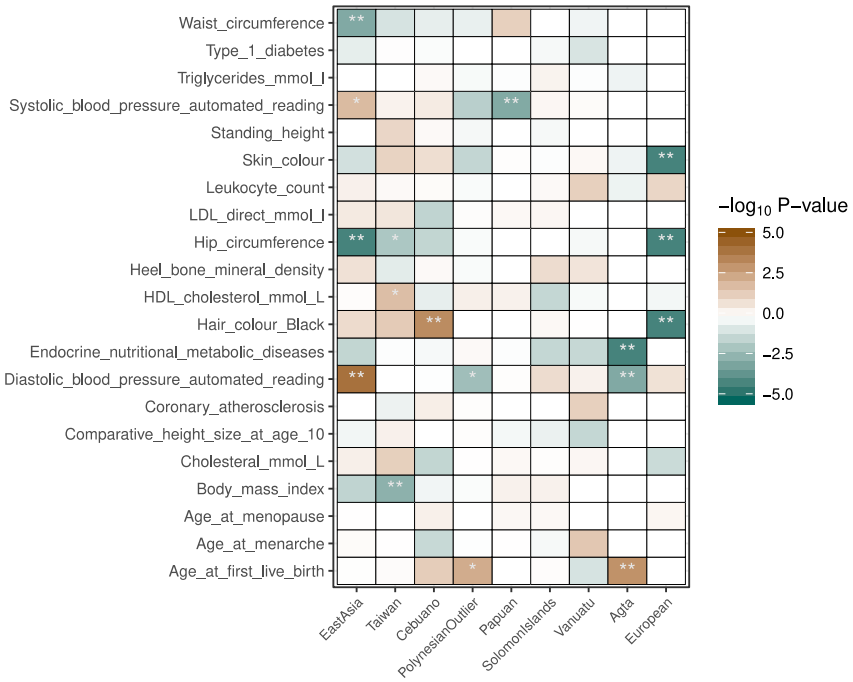
Supplementary Figure 81. Genetic architecture of candidate complex traits. **a**, Number of SNPs shared among candidate traits. Each column shows the percentage of SNPs associated with the trait that are also associated with other candidate traits (rows). **b**, Percentage of associated SNPs per candidate trait. The total number of SNPs varies across populations (n=9 populations), as it is the number of SNPs kept for selection analyses for each population. The boxplots indicate the median value, the first and fourth quartiles and the dots the outliers of the distribution.

Window-based approach

Methods. To statistically test if genes associated with a trait are preferential targets of positive selection, we first kept, for each trait, trait-associated variants that are unlinked. To do so, we partitioned the genome into 100-Kb non-overlapping contiguous windows and kept, for each window, only the variant with the lowest association *P*-value. We considered a window to be associated with a trait if at least one SNP within the window shows a *P*-value < 5×10^{-8}. For each window, we estimated the mean tiHS for each population (see previous

section) and calculated the mean GERP score[158], the mean recombination rate, the mean DAF and the number of SNPs per window. We then tested if the mean tiHS of trait-associated windows is higher than a null distribution, obtained from 100,000 sets of randomly-sampled windows, each set being matched to trait-associated windows in terms of GERP scores, recombination rate, DAF and number of SNPs. GERP, recombination rate and DAF were grouped into 8 bins. We calculated a *P*-value for each trait as the number of resamples, out of 100,000 resamples, where the mean tiHS was lower (or higher) than that observed for the trait-associated windows. We adjusted *P*-values for multiple testing with the Benjamini-Hochberg method, to account for the number of traits and populations tested.

Results. To relax the assumption that alleles affecting traits are the same in Oceanians and Europeans, we used another approach that tests for the co-localization of selection signals and trait-associated genes; this window-based approach assumes that the same genomic regions affect the traits of interest in all human populations. At a significance threshold of *P*-value < 0.005, we replicated a signal for decreased skin and hair colour in Europeans (Supplementary Fig. 82). With respect to the SNP-based approach, the window-based approach detected several additional signals, which may suggest either higher power, because trait-associated SNPs are actually not portable in Pacific populations, or stronger effects of pleiotropy, because genomic windows supposed to be trait-associated have not been associated *per se* with the trait of interest. Conversely, some of the strongest signals detected using the SNP-based approach (e.g., HDL in Vanuatu, Santa Cruz and Solomon Islands) were not significant when using the window-based approach, suggesting reduced power. Importantly, the polygenic adaptation signal for HDL cholesterol in Oceanians was replicated when decreasing the size of genomic windows (*P*-value < 0.05), suggesting that local signatures of positive selection are too weak to be detected when using 100-kb genomic windows. Among signals that were not detected with the SNP approach, we found signals related to blood pressure; specifically, lower systolic blood pressure in PNG, higher diastolic blood pressure in East Asians, and lower diastolic blood pressure in the Philippine Agta. We also detected a signal for decreased hip and waist circumference, increased hair pigmentation, and increased age at last reproduction in East Asian-related groups. GWAS of morphological and life-history traits in Pacific populations, which are largely underrepresented in genomics research, are required to confirm these results.



Supplementary Figure 82. Window-based detection of polygenic adaptation in Pacific populations. Colours indicate the -$\log_{10}$(*P*-value) for a significant decrease (in blue; tiHS > 0) or increase (in brown; tiHS < 0) of the candidate trait. *P*-values were computed for each trait as the number of resamples, out of 100,000 resamples, where the mean tiHS was lower (or higher) than that observed for the trait-associated windows (two-sided test). P-values were adjusted for multiple testing with the Benjamini-Hochberg method, to account for the number of traits and populations tested. Significance is indicated by stars, with *P-value < 0.025 and ***P*-value < 0.005.

Supplementary Information References

1 Cassar, O. *et al.* Human T lymphotropic virus type 1 subtype C melanesian genetic variants of the Vanuatu Archipelago and Solomon Islands share a common ancestor. *J Infect Dis* **196**, 510-521, doi:10.1086/519167 (2007).

2 Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479-483, doi:10.1126/science.1166858 (2009).

3 Ko, A. M. *et al.* Early Austronesians: into and out of Taiwan. *Am J Hum Genet* **94**, 426-436, doi:10.1016/j.ajhg.2014.02.003 (2014).

4 Shikama, T., Ling, C. C., Shimoda, N. & Baba, H. Discovery of Fossil *Homo sapiens* from Cho-chen in Taiwan. *J Anthrop Soc Nippon* **84**, 131-138 (1976).

5 Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* **89**, 516-528, doi:10.1016/j.ajhg.2011.09.005 (2011).

6 Détroit, F. *et al.* Upper Pleistocene *Homo sapiens* from the Tabon cave (Palawan, The Philippines): description and dating of new discoveries. *Cr Palevol* **3**, 705-712, doi:10.1016/j.crpv.2004.06.004 (2004).

7 Jinam, T. A. *et al.* Discerning the Origins of the Negritos, First Sundaland People: Deep Divergence and Archaic Admixture. *Genome Biol Evol* **9**, 2013-2022, doi:10.1093/gbe/evx118 (2017).

8 Reid, L. A. Who are the Philippine negritos? Evidence from language. *Hum Biol* **85**, 329-358, doi:10.3378/027.085.0316 (2013).

9 Spriggs, M. in *Bougainville: Before the Conflict* (eds A. J. Regan & H. Griffin) (Pandanus Press, 2005).

10 Kirch, P. V. *On the road of the winds: An archeological history of the Pacific islands before European contact*. (University of California Press, 2017).

11 Delfin, F. *et al.* Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands. *Mol Biol Evol* **29**, 545-564, doi:10.1093/molbev/msr186 (2012).

12 Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215, doi:10.1093/nar/16.3.1215 (1988).

13 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

14 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

15 Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235-239, doi:10.1126/science.aad9416 (2016).

16 Malaspinas, A. S. *et al.* A genomic history of Aboriginal Australia. *Nature* **538**, 207-214, doi:10.1038/nature18299 (2016).

17 Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206, doi:10.1038/nature18964 (2016).

18 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

19 The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:10.1038/nature04072 (2005).

20 Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41**, D64-69, doi:10.1093/nar/gks1048 (2013).

21 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-33, doi:10.1002/0471250953.bi1110s43 (2013).

22 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).

23 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

24 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

25 Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543-546, doi:10.1126/science.aal1988 (2017).
26 Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318-323, doi:10.1093/bioinformatics/btu668 (2015).
27 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).
28 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).
29 Walter, R. & Sheppard, P. J. in *Lapita: Ancestors and descendants* (eds P.J. Sheppard, T. Thomas, & G.R. Summerhayes) 32-72 (New Zealand Archaeological Association Monograph Series, 2009).
30 Pugach, I. *et al.* The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data. *Mol Biol Evol* **35**, 871-886, doi:10.1093/molbev/msx333 (2018).
31 Posth, C. *et al.* Language continuity despite population replacement in Remote Oceania. *Nat Ecol Evol* **2**, 731-740, doi:10.1038/s41559-018-0498-2 (2018).
32 Lipson, M. *et al.* Population Turnover in Remote Oceania Shortly after Initial Settlement. *Curr Biol* **28**, 1157-1165 e1157, doi:10.1016/j.cub.2018.02.051 (2018).
33 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).
34 Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817-2823, doi:10.1093/bioinformatics/btw327 (2016).
35 Wollstein, A. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr Biol* **20**, 1983-1992, doi:10.1016/j.cub.2010.10.040 (2010).
36 Lipson, M. *et al.* Reconstructing Austronesian population history in Island Southeast Asia. *Nat Commun* **5**, 4689, doi:10.1038/ncomms5689 (2014).
37 Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* **9**, 3258 (2018).
38 Skoglund, P. *et al.* Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510-513, doi:10.1038/nature19844 (2016).
39 Kirch, P. V. The Polynesian outliers: Continuity, change, and replacement. *J Pac Hist* **19**, 224–238 (1984).
40 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).
41 Lipson, M. *et al.* Three Phases of Ancient Migration Shaped the Ancestry of Human Populations in Vanuatu. *Curr Biol*, doi:https://doi.org/10.1016/j.cub.2020.09.035 (2020).
42 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265, doi:10.1093/bioinformatics/bth457 (2005).
43 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
44 Tenesa, A. *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**, 520-526, doi:10.1101/gr.6023607 (2007).
45 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).
46 Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**, e1003905, doi:10.1371/journal.pgen.1003905 (2013).
47 Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931-942 (2000).
48 de Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477-481, doi:10.1126/science.aag2602 (2016).
49 Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182-188, doi:10.1038/s41586-019-1279-z (2019).

50 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, doi:10.1002/ajpa.20188 (2005).
51 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).
52 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:doi:10.1038/nature13810 (2014).
53 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, doi:10.1111/j.1755-0998.2010.02847.x (2010).
54 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).
55 Pouyet, F., Aeschbacher, S., Thiery, A. & Excoffier, L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife* **7**, doi:10.7554/eLife.36317 (2018).
56 Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658, doi:10.1126/science.aao1887 (2017).
57 Yang, M. A. *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288, doi:https://doi.org/10.1126/science.aba0909 (2020).
58 Liu, L. & Chen, X. *The Archaeology of China: From the Late Paleolithic to the Early Bronze Age*. (Cambridge Univ. Press, 2012).
59 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).
60 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
61 Groube, L., Chappell, J., Muke, J. & Price, D. A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature* **324**, 453-455, doi:10.1038/324453a0 (1986).
62 Kirch, P. V. *The Lapita peoples* (Blackwell, 1999).
63 Summerhayes, G. R. & Allen, J. The transport of Mopir obsidian to Late Pleistocene New Ireland. *Archaeology in Oceania* **28**, 144-148 (1993).
64 Summerhayes, G. R. Obsidian network patterns in Melanesia—Sources, characterization and distribution. *IPPA Bulletin* **29**, 109-123 (2009).
65 Bergstrom, A. *et al.* A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160-1163, doi:10.1126/science.aan3842 (2017).
66 Sheppard, P. J., Chiu, S. & Walter, R. Re-dating Lapita movement into Remote Oceania. *J. Pacific Archaeol.* **6**, 26-36 (2015).
67 Rieth, T. M. & Athens, T. 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 (2017).
68 Harris, D. N. *et al.* Evolutionary history of modern Samoans. *Proc Natl Acad Sci U S A* **117**, 9458-9465 (2020).
69 Hung, H.-C. & Carson, M. T. Foragers, fishers and farmers: Origins of the Taiwanese Neolithic. *Antiquity* **88**, 1115-1131, doi:doi:10.1017/S0003598X00115352 (2014).
70 Bellwood, P. *First Farmers: the Origins of Agricultural Societies*. (Blackwell, 2005).
71 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 e59, doi:10.1016/j.cell.2018.02.031 (2018).
72 Chen, C. H. *et al.* Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum Mol Genet* **25**, 5321-5331, doi:10.1093/hmg/ddw346 (2016).
73 Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035 (2002).
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,* doi: 10.22541/au.160157625.155585861 (2020).

75 Verdu, P. & Rosenberg, N. A. A general mechanistic model for admixture histories of hybrid populations. *Genetics* **189**, 1413-1426, doi:10.1534/genetics.111.132787 (2011).
76 Weir, B. S. & Cockerham, C. C. Estimating *F*-statistics for the analysis of population structure. . *Evolution* **38**, 1358-1370 (1984).
77 Bowcock, A. M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455-457, doi:10.1038/368455a0 (1994).
78 Pudlo, P. *et al.* Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859-866, doi:10.1093/bioinformatics/btv684 (2016).
79 Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol* **3**, 475-479 (2012).
80 Jay, F., Boitard, S. & Austerlitz, F. An ABC Method for Whole-Genome Sequence Data: Inferring Paleolithic and Neolithic Human Expansions. *Mol Biol Evol* **36**, 1565-1579, doi:10.1093/molbev/msz038 (2019).
81 Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607-619, doi:10.1534/genetics.112.139808 (2012).
82 Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953-967, doi:10.1534/genetics.114.162362 (2014).
83 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, doi:10.1016/j.ajhg.2013.06.020 (2013).
84 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).
85 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2011).
86 Qin, P. & Stoneking, M. Denisovan Ancestry in East Eurasian and Native American Populations. *Mol Biol Evol* **32**, 2665-2674, doi:10.1093/molbev/msv141 (2015).
87 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357, doi:10.1038/nature12961 (2014).
88 Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-1021, doi:10.1126/science.1245938 (2014).
89 Vernot, B. & Akey, J. M. Complex history of admixture between modern humans and Neandertals. *Am J Hum Genet* **96**, 448-453, doi:10.1016/j.ajhg.2015.01.006 (2015).
90 Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199-209, doi:10.1534/genetics.112.148213 (2013).
91 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, doi:10.1016/j.cub.2016.03.037 (2016).
92 Jacobs, G. S. *et al.* Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **177**, 1010-1021 e1032, doi:10.1016/j.cell.2019.02.035 (2019).
93 Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012, doi:https://doi.org/10.1126/science.aay5012 (2020).
94 Douka, K. *et al.* Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature* **565**, 640–644, doi:https://doi.org/10.1038/s41586-018-0870-z (2019).
95 Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113-1118, doi:10.1126/science.aaa0114 (2014).
96 Detroit, F. *et al.* A new species of *Homo* from the Late Pleistocene of the Philippines. *Nature* **568**, 181-186, doi:10.1038/s41586-019-1067-9 (2019).
97 Brown, P. *et al.* A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055-1061, doi:10.1038/nature02999 (2004).
98 Racimo, F., Marnetto, D. & Huerta-Sanchez, E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Mol Biol Evol* **34**, 296-317, doi:10.1093/molbev/msw216 (2017).
99 Rogers, M. A. *et al.* Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13. *J Invest Dermatol* **124**, 536-544, doi:10.1111/j.0022-202X.2004.23530.x (2005).

100 The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).
101 Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737-741, doi:10.1126/science.aad2149 (2016).
102 Dannemann, M. & Kelso, J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *Am J Hum Genet* **101**, 578-589, doi:10.1016/j.ajhg.2017.09.010 (2017).
103 Williams, A. L. *et al.* Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97-101, doi:10.1038/nature12828 (2014).
104 Racimo, F. *et al.* Archaic Adaptive Introgression in TBX15/WARS2. *Mol Biol Evol* **34**, 509-524, doi:10.1093/molbev/msw283 (2017).
105 Taylor, E. B. *et al.* Discovery of TBC1D1 as an insulin-, AICAR-, and contraction-stimulated signaling nexus in mouse skeletal muscle. *J Biol Chem* **283**, 9787-9796, doi:10.1074/jbc.M708839200 (2008).
106 Bruss, M. D., Arias, E. B., Lienhard, G. E. & Cartee, G. D. Increased phosphorylation of Akt substrate of 160 kDa (AS160) in rat skeletal muscle in response to insulin or contractile activity. *Diabetes* **54**, 41-50, doi:10.2337/diabetes.54.1.41 (2005).
107 Deshmukh, A. *et al.* Exercise-induced phosphorylation of the novel Akt substrates AS160 and filamin A in human skeletal muscle. *Diabetes* **55**, 1776-1782, doi:10.2337/db05-1419 (2006).
108 Chadt, A. *et al.* Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. *Nat Genet* **40**, 1354-1359, doi:10.1038/ng.244 (2008).
109 Meyre, D. *et al.* R125W coding variant in TBC1D1 confers risk for familial obesity and contributes to linkage on chromosome 4p14 in the French population. *Hum Mol Genet* **17**, 1798-1802, doi:10.1093/hmg/ddn070 (2008).
110 Stone, S. *et al.* TBC1D1 is a candidate for a severe obesity gene and evidence for a gene/gene interaction in obesity predisposition. *Hum Mol Genet* **15**, 2709-2720, doi:10.1093/hmg/ddl204 (2006).
111 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, doi:10.1111/ahg.12132 (2015).
112 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 U S A* **98**, 5764-5769, doi:10.1073/pnas.091097198 (2001).
113 Negishi, H. *et al.* Negative regulation of Toll-like-receptor signaling by IRF-4. *Proc Natl Acad Sci U S A* **102**, 15989-15994, doi:10.1073/pnas.0508327102 (2005).
114 Eletto, D. *et al.* Biallelic JAK1 mutations in immunodeficient patient with mycobacterial infection. *Nat Commun* **7**, 13992, doi:10.1038/ncomms13992 (2016).
115 Rodig, S. J. *et al.* Disruption of the Jak1 gene demonstrates obligatory and nonredundant roles of the Jaks in cytokine-induced biologic responses. *Cell* **93**, 373-383, doi:10.1016/s0092-8674(00)81166-6 (1998).
116 Kurosaki, T. Regulation of B-cell signal transduction by adaptor proteins. *Nat Rev Immunol* **2**, 354-363, doi:10.1038/nri801 (2002).
117 Murphy, G., Lisnevskaia, L. & Isenberg, D. Systemic lupus erythematosus and other autoimmune rheumatic diseases: challenges to treatment. *Lancet* **382**, 809-818, doi:10.1016/S0140-6736(13)60889-2 (2013).
118 Han, J. W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* **41**, 1234-1237, doi:10.1038/ng.472 (2009).
119 Yang, W. *et al.* Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet* **6**, e1000841, doi:10.1371/journal.pgen.1000841 (2010).
120 Yang, W. *et al.* Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. *Am J Hum Genet* **92**, 41-51, doi:10.1016/j.ajhg.2012.11.018 (2013).
121 Taskent, R. O. *et al.* Variation and Functional Impact of Neanderthal Ancestry in Western Asia. *Genome Biol Evol* **9**, 3516-3524, doi:10.1093/gbe/evx216 (2017).

122 Dannemann, M., Andres, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am J Hum Genet* **98**, 22-33, doi:10.1016/j.ajhg.2015.11.015 (2016).
123 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, doi:10.1016/j.ajhg.2015.11.014 (2016).
124 Enard, D. & Petrov, D. A. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **175**, 360-371 e313, doi:10.1016/j.cell.2018.08.034 (2018).
125 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).
126 The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
127 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, doi:10.1093/nar/gky1016 (2019).
128 Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res* **47**, D745-D751, doi:10.1093/nar/gky1113 (2019).
129 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).
130 Kutmon, M. *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res* **44**, D488-494, doi:10.1093/nar/gkv1024 (2016).
131 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, doi:10.1093/nar/gky1120 (2019).
132 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
133 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).
134 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918, doi:10.1038/nature06250 (2007).
135 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*, doi:10.1038/s41431-019-0518-y (2019).
136 Lee, I. H. *et al.* Atg7 modulates p53 activity to regulate cell cycle and survival during metabolic stress. *Science* **336**, 225-228, doi:10.1126/science.1218395 (2012).
137 Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* **51**, 51-62, doi:10.1038/s41588-018-0303-9 (2019).
138 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, doi:10.1093/hmg/ddy422 (2019).
139 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186, doi:10.1038/ng.3097 (2014).
140 Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166-174, doi:10.1093/hmg/ddy327 (2019).
141 Pardo-Diaz, C. *et al.* Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genet* **8**, e1002752, doi:10.1371/journal.pgen.1002752 (2012).
142 Huerta-Sanchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197, doi:10.1038/nature13408 (2014).
143 Arnold, B. J. *et al.* Borrowed alleles and convergence in serpentine adaptation. *Proc Natl Acad Sci U S A* **113**, 8320-8325, doi:10.1073/pnas.1600405113 (2016).
144 Clarkson, C. S. *et al.* Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nat Commun* **5**, 4248, doi:10.1038/ncomms5248 (2014).

145 Song, Y. *et al.* Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr Biol* **21**, 1296-1301, doi:10.1016/j.cub.2011.06.043 (2011).
146 Hodgson, J. A. *et al.* Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proc Biol Sci* **281**, 20140930, doi:10.1098/rspb.2014.0930 (2014).
147 Laso-Jadart, R. *et al.* The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan. *Am J Hum Genet* **101**, 977-984, doi:10.1016/j.ajhg.2017.09.025 (2017).
148 Jeong, C. *et al.* Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* **5**, 3281, doi:10.1038/ncomms4281 (2014).
149 Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858, doi:10.1093/nar/gky1095 (2019).
150 Popov, S., Popova, E., Inoue, M. & Gottlinger, H. G. Divergent Bro1 domains share the capacity to bind human immunodeficiency virus type 1 nucleocapsid and to enhance virus-like particle production. *J Virol* **83**, 7185-7193, doi:10.1128/JVI.00198-09 (2009).
151 Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* **20**, R208-215, doi:10.1016/j.cub.2009.11.055 (2010).
152 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, doi:10.1146/annurev-genom-083115-022316 (2019).
153 Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764, doi:10.1126/science.aag0776 (2016).
154 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, doi:10.1038/s41588-019-0484-x (2019).
155 Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, doi:10.7554/eLife.39725 (2019).
156 Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, doi:10.7554/eLife.39702 (2019).
157 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).
158 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).
159 Summerhayes, G. R. & Ford, A. in *Southern Asia, Australia and the Search for Human Origins* (eds R. Dennell & M. Porr) (Cambridge University Press, 2014).

5.3 Summary of results

To obtain insight into the peopling and demographic past of Pacific islanders we jointly inferred the parameters characterizing their demographic history using multidimensional site frequency spectra and the maximum likelihood framework. We first explored different branching topologies and estimated the demographic parameters of Near Oceanians (i.e. Papuans, Bismark and Solomon islanders). We found that the settlement of the region was accompanied by a strong founder event - around five time stronger than that of Eurasians - and that the different groups diverged in the late Pleistocene, between 40,000 and 25,000 years ago. These results point to a rapid genetic isolation of the different groups of Near Oceania, after the initial settlement dated to around 45,000 years ago (archaeological data, (O'Connell et al. 2018a; O'Connell and Allen 2015)).

Similarly, we tested different topologies and estimated demographic parameters for western Remote Oceanians. We found that Ni-Vanuatu received post-Lapita gene flow from Bismarck islanders in agreement with ancient DNA (Posth et al. 2018; Lipson et al. 2018). Furthermore, the best-fitted models indicate that the Bismarck archipelago ancestry alone is not enough to represent the Papuan-related genetic diversity found today in Vanuatu. Because of a lack of continuity between first and present-day Ni-Vanuatu as shown by ancient DNA and craniometric studies (Posth et al. 2018; Lipson et al. 2018; Valentin et al. 2016), interpretation of demographic models using modern DNA is very limited.

We also reconstructed the demographic history of the East/Southeast Asian ancestors of Near and Remote Oceanians. Assuming an isolation with migration model, we estimated that Taiwanese Indigenous peoples and Malayo-Polynesian speakers diverged around 7,300 years ago at odds with the "Out-of-Taiwan" model - hypothesis that predicts a dispersal event from Taiwan around 4,800 years ago and that brought both the agriculture and Austronesian languages to Oceania (Bellwood 1997). We obtained consistently older divergence times, even when considering gene flow into Austronesian-speaking groups, but with broader confidence intervals. These results suggest a population structure of Austronesian speakers that predate the appearance of agriculture in Taiwan. However, because of the large uncertainty in the estimates further analyses using ancient genomes are needed.

After having investigated divergence times, we wanted to obtain insight into the mode and tempo of gene flow between East/Southeast Asians and Near Oceanians. We used an Approximate Bayesian Computation (ABC) approach to test for a single-pulse model versus a two-pulse model or continuous model of gene flow. We found that a two-pulse

model best matched the summary statistics. We dated a first admixture event in Near Oceanians (Bismarck and western Solomon islanders) at around 3,000 years ago and a second at around 1,500 years ago, indicating mutliple contacts between East/Southeast Asians and Near Oceanians.

We also shed light on the archaic genetic legacy found in the Pacific region. Using allele frequency-based methods we found that, while the level of Neanderthal ancestry is fairly homogeneous across the region, the level of Denisovan ancestry is heterogeneous. For example, the Agta foragers from the Philippines present around 3% of Denisovan ancestry while neighbour populations have around 0% (e.g. Cebuano Filipinos). The identification and analysis of Denisovan haplotypes in the genome of present-day Pacific islanders suggest multiple episodes of interbreeding between Denisovans and Pacific groups. Using an ABC approach we found that two highly divergent groups of Denisovans introgressed with Papuan-related groups, around 45,000 and 25,000 years ago, respectively.

Finally, we searched for the occurrence of classic sweeps and other modes of genetic adaptation such as adaptive admixture/introgression and polygenic adaptation. We found that unlike Neanderthal introgression which facilitated the adaptation of modern humans related to a large range of phenotypes (e.g. metabolism, pigmentation and neuronal development), Denisovan introgression mainly targeted immune-related functions (e.g. *CD33* and *IRF4* genes). We identified 44 shared genetic regions targeted by classical positive natural selection (i.e. classic sweeps) in Papuan-related groups. The strongest hit includes the *RANBP17* gene, which is involved in Body Mass Index and HDL cholesterol. We identified 29 genetic regions shared between East-Asian related groups (including Polynesian groups). One of our strongest signal fall within the *ALDH2* locus which is involved in alcohol metabolism.

Collectively, our analyses provide novel insights into the genetic history of Pacific populations, including various interactions with archaic hominins, early splits during the late Pleistocene, recent range expansions in the Holocene period and a complex history of interactions between peoples from East/Southeast Asia and Oceania. Our result also increased our understanding of the mechanisms of biological adaptation experienced by Pacific islanders.

Chapter 6

Result 2
Selection efficacy in insular populations: the case of Pacific islanders

I present here the first results obtain for the second part of my thesis, which aims to evaluate the efficacy of natural selection in Pacific populations, and ultimately better understand their present-day relation to diseases.

6.1 Context

Alleles associated with diseases are part of the human genetic diversity and mutation, genetic drift and natural selection, thus govern their occurrence, frequency, and population distribution. At mutation-selection equilibrium and stationary demography, the rate at which deleterious mutations are removed from the populations, i.e., the efficacy of natural selection depends on the product between the effective population size (N_e) and the selection coefficient (s) (Charlesworth 2009). Hence, in theory, the efficacy of natural selection to remove deleterious mutation depends on the demographic fluctuations experienced by populations, i.e., the demographic history. The burden of deleterious mutations has often been quantified through the measure of the mutational load, which corresponds to the reduction in fitness owing to the accumulation of deleterious mutations in genomes, compared with the optimal fitness (by convention set to 1) (Knudson 1979; Lopez et al. 2018a; Paul 1987; Simons and Sella 2016). Theoretically, in small or bottlenecked populations - because of a strong genetic drift - the mutational load is transiently high (Simons et al. 2014; Balick et al. 2015) due to a drop in the efficacy of natural selection and the prevalence of recessive diseases may thus increase. These predictions are strengthened by epidemiological studies, which reported cases of unusually frequent recessive disorders in isolated or small-island populations (O'Brien et al. 1988; Carr, Morton, and Siegel 1971; Eickhoff and Beighton 1985).

In an attempt to validate empirically these predictions in humans, a large number of genomic studies has compared the pattern of deleterious mutations between Sub-Saharan Africans and non-African groups (Lopez et al. 2018a; Simons and Sella 2016; Simons et al. 2014; Do et al. 2015; Henn et al. 2016b; Henn et al. 2015b; Lohmueller et al. 2008; Lohmueller 2014; Fu et al. 2013; Pedersen et al. 2017a; Font-Porterias et al. 2021). Because the individual's fitness cannot be easily calculated in humans, these studies used different metrics and definitions of the burden of deleterious mutations and efficacy of natural selection, which led to conflicting interpretations. For example, Henn et al. (Henn et al. 2016b), using simulations and selection coefficients approximated from sequence conservation-based score (GERP (Cooper et al. 2005)) categories, predicted significant differences in the additive mutational load between human groups. Conversely, Do et al. (Do et al. 2015) counted the differences in the number of derived deleterious mutations between African and European individuals and concluded that the Out-of-Africa bottleneck did not affected the efficacy of natural selection. Although there is increasing evidence to suggest that bottleneck and recent population growth had a negligible impact on the additive genetic load and the efficacy of natural selection (Lopez et al. 2018a; Simons and

Sella 2016; Simons et al. 2014; Do et al. 2015), long-standing and strong bottlenecks, as experienced by Greenlandic Inuit, appeared to have impacted the number and frequency of recessive deleterious mutations (Pedersen et al. 2017a). Likewise, recent studies also highlighted the role of recent admixture in balancing the effect of strong genetic drift on the burden of recessive deleterious mutations (Lopez et al. 2018a; Font-Porterias et al. 2021).

The region of Oceania, spanning from Papua New Guinea up to the Polynesian Triangle includes thousands of scarcely populated islands. Archaeological records suggest that Near Oceania, which includes New Guinea, the Bismarck Archipelago and the Solomon Islands, was first inhabited around 45,000 years ago (ya) (O'Connell et al. 2018b; Gosling and Matisoo-Smith 2018a). Remote Oceania, which includes Micronesia, the Reef/Santa Cruz, Vanuatu, New Caledonia, Fiji, and Polynesia, remained unoccupied until the recent arrival of Austronesian-speaking people originating from Taiwan and Islands Southeast Asia around 3,200 ya (Gosling and Matisoo-Smith 2018a; Kirch 2017). Genomic studies shed light on a demographic past characterized by a strong founder event associated with the peopling of the ancient Sahul continent and northeastern islands lying off (effective population size (Ne)= 153-1,788 diploids) (Choin et al. 2021; Malaspinas et al. 2016b), low effective population sizes notably for Polynesian groups (Choin et al. 2021; Harris et al. 2020) and recent admixture between Papuan-related and East/Southeast Asian-related groups (Choin et al. 2021; Pugach et al. 2018a; Posth et al. 2018; Lipson et al. 2018; Lipson et al. 2020). Moreover, the World Health Organization (WHO) also reports a high prevalence of metabolic disorders such as Type 2 diabetes, obesity and gout in this region. Yet, little is known about the burden of deleterious mutations and whether the strong genetic drift experienced by Pacific islanders (especially Polynesians) resulted in a reduction in the efficacy of natural selection. More generally, the region of Oceania, by its almost unique geographic context and sharp demographic events, provides with an excellent model to evaluate the extent to which recent demographic events have impacted the occurrence and distribution of deleterious mutations in the human genome.

6.2 Results

6.2.1 Dataset

We combined a previously generated WGS dataset composed of Pacific islanders (Choin et al. 2021) with sequences from a number of worldwide groups (Malaspinas et al. 2016b; Bergstrom et al. 2020). This dataset includes a total of 150 individuals distributed in 15

Sub-Saharan Africans (Yoruba), 15 Europeans (French), 15 East Asians (Han Chinese), 15 Southeast Asians (Cebuano), 15 Polynesian outliers (Rennell and Bellona), 15 New Guineans (Highlanders) 15 western Solomon Islanders (Vella Lavella), 15 Eastern Solomon Islanders (Malaita), 15 Southern Ni-Vanuatu (Tanna) and 15 central Ni-Vanuatu (Malakula). Focusing on Near and Remote Oceanians (New Guineans, Solomon islanders, Ni-Vanuatu and Polynesian outliers), we identified 73,219 quality-filtered segregating missense and 636 stop gained loss-of-functions variants (hereafter referred as LoF) within exons of 18,300 genes. Considering only 169 LoF variants absent or at low frequency in gnomAD (Karczewski et al. 2020)(Supplementary Table 1) we did not find significant enrichments in LoF genes for any gene ontology (GO) categories, after correcting for multiple testing (Benjamini & Hochberg method, Top 20 GO enrichment results are given in Supplementary Table 2).

6.2.2 Evaluate the efficacy of natural selection

We investigated whether the demographic history of Oceanian groups, mainly characterized by strong founder effects, low effective population sizes, as well as recent admixture events, impacted their burden of deleterious mutations. We first assessed the allele frequency spectra of deleterious variants, using a sequence conservation-based score (i.e., GERP RS score (Cooper et al. 2005)) that is free from genome reference biases. We found that the derived frequency spectrum of all populations is enriched in rare variants, the proportion of which increases with deleteriousness (Figure 1a). We also observed that rare variants (singletons) are enriched in deleterious variants, mainly for the "Moderate" and "Strong" deleteriousness categories (Figure 1b). Altogether, these results are consistent with the effect of purifying selection acting on worldwide human populations. Interestingly, Polynesian outliers (RenBell) harbour an excess of neutral mutations and a default of deleterious mutations in singletons, compared to other groups (Figure 1b), but the lowest proportion of rare non-deleterious variants was also observed in this population (Figure 1c), suggesting that differences in the allele distribution of deleterious variants could be, at least partially, explained by stronger genetic drift among Polynesian groups.

We next tested whether the observed population differences in the shape of deleterious SFS could also result from a difference among populations in the efficacy of purifying selection (which depends on $N_e s$). We thus calculated the ratio of the fixation probability (u) for a new deleterious mutation versus a neutral mutation, to quantify the efficacy of purifying selection to remove mutations, relative to genetic drift ($u_{\text{del}}/u_{\text{neu}}$, the smaller ($u$) the greater the efficacy of natural selection). To calculate the fixation probability of new

deleterious mutations (u_{del}), we used the parameters of the distribution of fitness effects (DFE) inferred with the algorithm implemented in $\partial a \partial i / Fit \partial a \partial i$. We first estimated a 3-epoch demographic model using one-dimensional synonymous SFS for each of the 10 groups. Then, conditional on the demographic parameters, we inferred the parameters for the DFE of non-synonymous mutations. We found subtle differences in the $u_{\text{del}}/u_{\text{neu}}$ ratio between Oceanians and other continental reference populations (Yoruba, Han Chinese and French). Notably, Oceanians tend to have a higher ratio (Table 1, Supplementary Figure 1), especially Papuans and Polynesians, suggesting a slightly reduced efficacy of natural selection in these groups. However, we caution that the DFE and thus the $u_{\text{del}}/u_{\text{neu}}$ ratio of some Oceanian groups, particularly Polynesians, should be interpreted with caution because of the very poor fit between observed and expected non-synonymous SFS (Supplementary Figure 1).

6.2.3 Evaluate the mutational load

We compared the empirical mutation load of present-day Pacific and continental reference groups (Yoruba, Han Chinese and French). Previous studies (Simons et al. 2014; Do et al. 2015; Lohmueller 2014) have reported that differences in genetic load between groups depend on the functional category of coding variants (e.g. GERP RS score (Cooper et al. 2005)) and the dominance model. We thus approximated the load under an additive and a recessive model for different GERP categories (Cooper et al. 2005), using between-population ratios of the mean number of derived alleles per individual (N_{alleles}) or between-population ratios of the mean number of homozygous derived genotypes (N_{hom}) respectively. We found that all Oceanian populations present the same level of genetic load as continental reference groups (for all GERP categories) under an additive model (corrected p values > 0.05, Figure 2a, Supplementary Table 3). However, Pacific groups, to the exclusion of Cebuano, harbour a significant higher recessive load than Africans for the strongly deleterious mutation category (adjusted p-value = 0.03 for all ratios, Figure 2b, Supplementary Table 3).

Simons and colleagues (Simons et al. 2014) have suggested that bottlenecks and population growth have only a minor impact on the additive genetic load, owing to (i) the compensation between the number of segregating variants, including deleterious mutations, and their frequency and (ii) because these demographic events are too recent or did not last long enough. In line with this, we found that Polynesians and Papuan highlanders show the lowest number of deleterious variants for all GERP categories and mutations that segregate on average at higher frequency than in any other Oceanian and non-Oceanian groups (Figure 3). We obtained similar patterns using stop-gained

Loss-of-Function variants (Supplementary Figure 2).

6.2.4 Effect of the Papuan-related ancestry and runs of homozygosity

Runs of homozygosity (ROH) correspond to long genomic regions in which all loci for an individual are homozygous. These long genomic segments are considered identical by descent, and are found homozygous in the individual either because the parents of the individual are related due to cultural endogamy, or because he is part of a population where relatedness is high. Long ROH can thus be used to measure population size and parental relatedness. We found that Polynesian outliers (RenBell) and to a lesser extent Papuan Highlanders (PNG) and Vella Lavella western Solomon islanders presented the highest levels of cumulative long ROH (cROH, Figure 4, Supplementary Figure 3) suggesting strong recent bottleneck, isolation or parental relatedness. We then tested whether mutational load is correlated with the individual cumulative length of long ROH, controlling for varying levels of Papuan-related genetic ancestry. We found that both additive and recessive mutational loads are not associated with Papuan-related ancestry proportions carried by Pacific islanders (corrected p value > 0.05). However, the number of derived homozygous genotypes correlated significantly with the cumulative length of long ROH, for different categories of GERP RS score tested (adjusted p values = 0.02, 0.007, 0.003 and slope = 1.50×10^{-7}, 1.42×10^{-7}, 1.63×10^{-7} for "Neutral", "Moderate" and "Strong" categories of GERP RS respectively, Supplementary Table 4 and Supplementary Table 5).

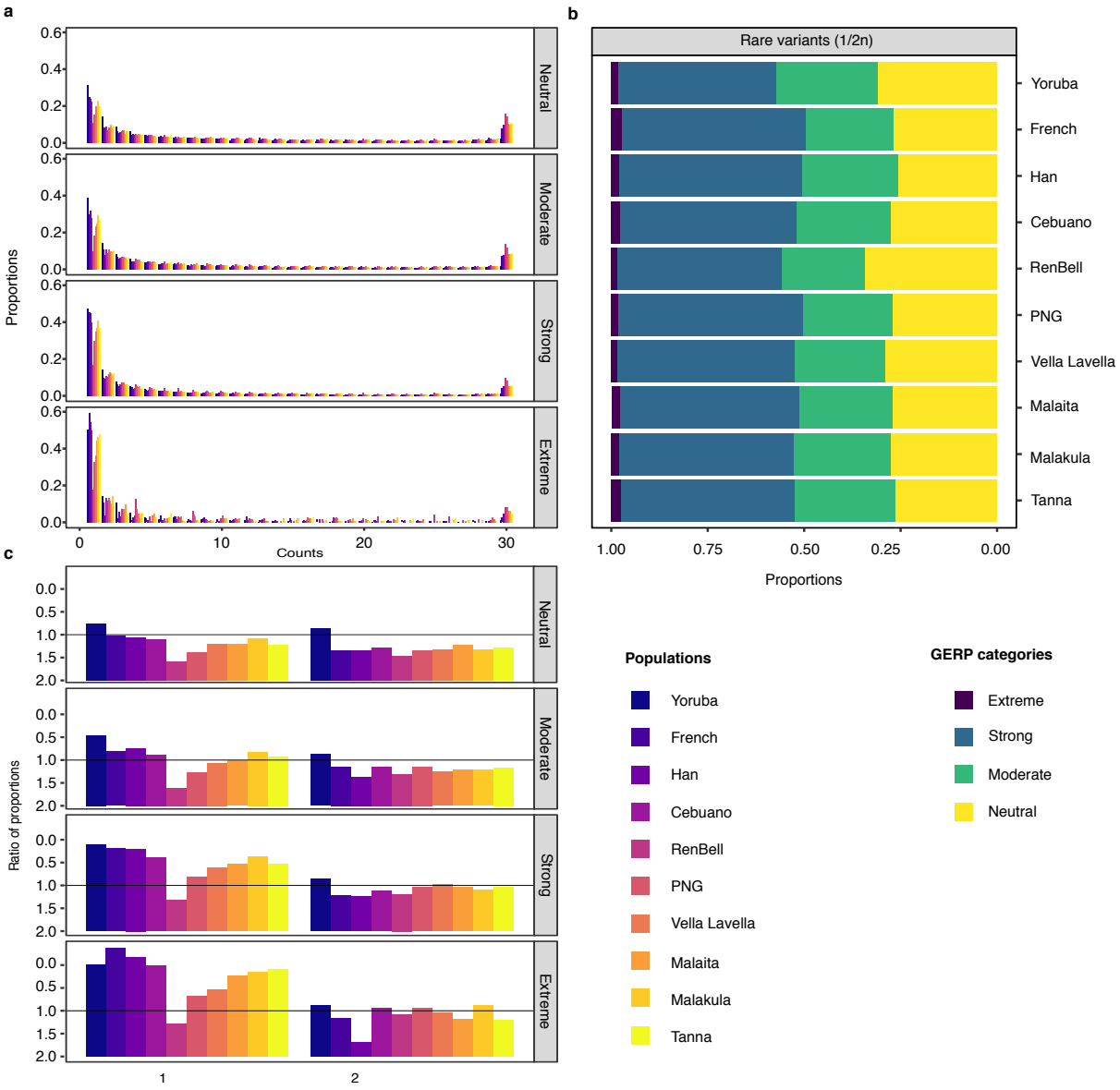


Figure 1. Allele frequency spectra of deleterious mutations. (a) Derived allele frequency spectra of non-synonymous mutations for different bin of GERP score (category of deleteriousness) for different world-wide populations including Africans (Yoruba), Europeans (French), East-Asians (Han), Filipinos (Cebuano), Polynesians (RenBell), Papuan highlanders (PNG), Solomon islanders (Vella-Lavella and Malaita) and Ni-Vanuatu (Malakula and Tanna). (b) Proportion of derived non-synonymous singletons assigned to different GERP score categories (deleteriousness categories). (c) First and second bin of derived allele frequency spectra of non-synonymous mutations for different bin of GERP score (category of deleteriousness) and normalized by the derived allele frequency spectra expected under constant effective population size and no natural selection. The sample size is equal to 15 for each group.

Population	Sample size (2n)	Nw	Beta	E[s]	Nw.E[s]	*u* del	*u* neu	*u* del/*u* neu
Yoruba	15	11343.64496	0.171 [0.160-0.184]	0.01 [0.007-0.013]	112.6 [82.5-153.2]	1.42E-05 [1.39e-05-1.45e-05]	4.41E-05	0.322 [0.316-0.328]
French	15	8409.283233	0.172 [0.161-0.187]	0.01 [0.007-0.014]	86.7 [61.4-120.7]	2.00E-05 [1.95e-05-2.03e-05]	5.95E-05	0.336 [0.328-0.342]
Han	15	7756.813185	0.175 [0.160-0.190]	0.01 [0.007-0.014]	74.3 [54.8-110.1]	2.19E-05 [2.14e-05-2.24e-05]	6.45E-05	0.340 [0.332-0.347]
Cebuano	15	7959.579458	0.122 [0.108-0.134]	0.06 [0.03-0.1]	457.8 [264.5-1019.6]	2.26E-05 [2.21e-05-2.32e-05]	6.28E-05	0.360 [0.353-0.370]
PNG	15	6930.722878	0.119 [0.107-0.132]	0.06 [0.03-0.1]	424.6 [237.9-860.4]	2.68E-05 [2.62e-05-2.73e-05]	7.21E-05	0.371 [0.364-0.379]
RenBell	15	7149.817662	0.087 [0.099-0.102]	0.98 [0.266-0.287]	7034.7 [1900.6-2055.4]	2.58E-05 [2.51e-05-2.60e-05]	6.99E-05	0.369 [0.359-0.372]
Malaita	15	8334.841315	0.148 [0.136-0.163]	0.02 [0.01-0.03]	149.3 [97.1-237.6]	2.12E-05 [2.06e-05-2.16e-05]	6.00E-05	0.353 [0.344-0.360]
Vella Lavella	15	8284.849957	0.130 [0.113-0.145]	0.04 [0.02-0.09]	316.4 [165.3-714.7]	2.16E-05 [2.11e-05-2.21e-05]	6.04E-05	0.358 [0.349-0.365]
Tanna	15	7973.785943	0.151 [0.137-0.165]	0.02 [0.01-0.03]	160.0 [110.1-265.6]	2.14E-05 [2.09e-05-2.20e-05]	6.27E-05	0.341 [0.334-0.351]
Malakula	15	7948.626794	0.147 [0.135-0.159]	0.02 [0.01-0.03]	142.8 [96.5-224.8]	2.25E-05 [2.20e-05-2.30e-05]	6.29E-05	0.358 [0.350-0.366]

Table 1. **DFE parameters (Beta and E[S]) and fixation probability of a new mutation for each group**. Nw corresponds to the weighted N_e across the 3-epoch model inferred with ∂a∂i and calculated as in (Lopez et al. 2018). *u* del corresponds to the fixation probability of a new deleterious mutation, u_{neu} to the fixation probability of a new neutral mutation and u_{del}/u_{neu} to the ratio. 95%CI are given in brackets and were calculated by bootstrapping by site 100 times.

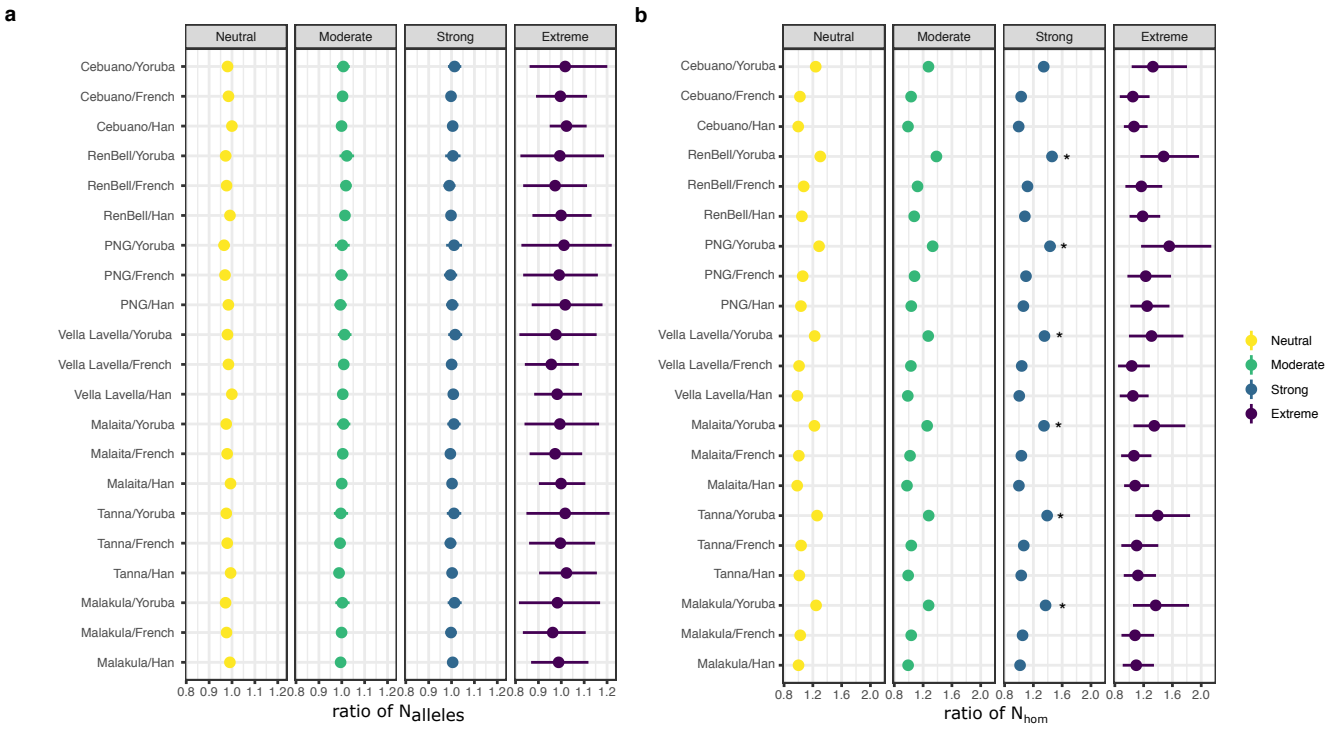


Figure 2. Ratios of the mean per-individual number of (a) derived alleles ($N_{alleles}$) and (b) homozygous derived genotypes (N_{hom}) between Pacific islanders and non-Oceanian groups represented by Yoruba Africans, Han Chinese and French Europeans. Dots indicate point estimates and lines, the 95% confidence intervals obtained by block bootstrap. * indicates a significant adjusted p-value (lower or

equal than 0.05) computed by comparing the bootstrap distribution for the considered GERP RS category to that of the “neutral” category.

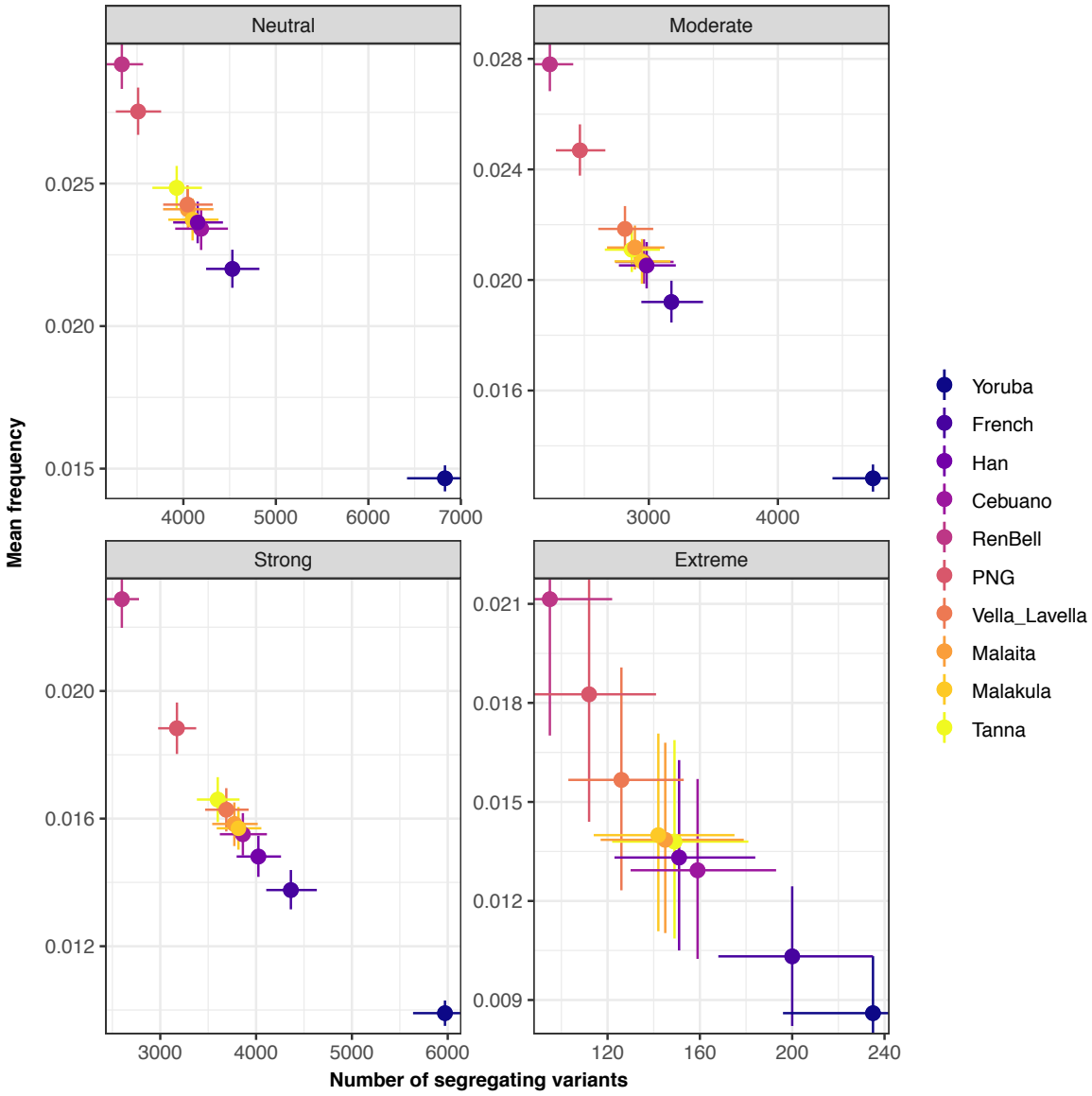


Figure 3. **Number of segregating deleterious mutations per category of GERP RS and their mean frequency per group**. Dots indicate point estimates and lines give the 95% confidence intervals obtained by block bootstrap.

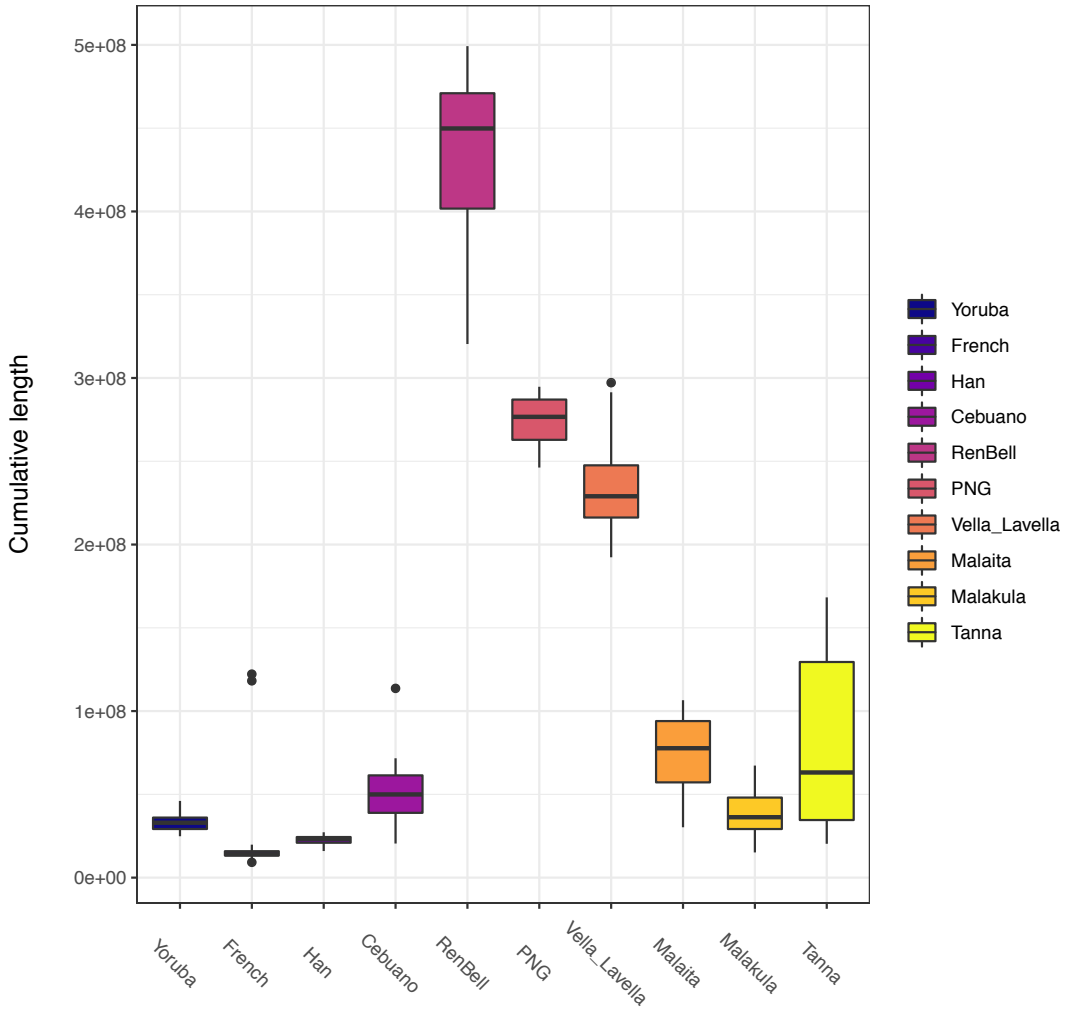


Figure 4. Cumulative length of long ROH (class C) per group.

6.3 Conclusion

6.3.1 Summary of results and short-term perspectives

The impact of bottlenecks, recent expansions and gene flow on the burden of deleterious mutations in human has been deeply investigated in the last decade (Do et al. 2015; Font-Porterias et al. 2021; Fu et al. 2013; Henn et al. 2015b; Henn et al. 2016b; Lohmueller 2014; Lohmueller et al. 2008; Lopez et al. 2018a; Pedersen et al. 2017a; Simons and Sella 2016; Simons et al. 2014). Most of these studies focused on the differences in the mutational load and efficacy of natural selection between continental populations such as African and European groups. Here, we investigated the burden of deleterious mutations and efficacy of natural selection of Oceanian islanders, who experienced strong founder effects, population collapses and recent admixture (Choin et al. 2021; Harris et al. 2020; Malaspinas et al. 2016b). We find that Oceanians show only subtle differences in the efficacy of natural selection (Table 1 and Supplementary Figure 1) and the current genetic load (Figure 2 and Supplementary Table 3), relative to continental reference groups (Yoruba, Han and French). However, we find that deleterious variants, including Loss-of-Function (LoF) variants tend to segregate at higher frequency in Polynesian and Papuan highlander groups, likely due to a stronger genetic drift (Figure 3 and Supplementary Figure 2). Yet, we need to evaluate whether these observations are also true for variants associated with metabolic disorders (e.g. BMI, Type-2-Diabetes) and whether a stronger drift (as for Polynesians), increases the genetic variance at metabolic associated genomic regions (Barton and Turelli 2004).

Additional analyses are required to investigate and dissect in greater detail the impact of (i) the recent Southeast Asian admixture and (ii) the apparent higher cumulative length of run of homozygosity in some Oceanians, on the current genetic load. Furthermore, we need to monitor the trajectory of the load through time using forward-in-time simulations. Similarly, we want also to investigate the role of each demographic event experienced by Oceanian islanders in shaping the occurrence and the distribution of deleterious mutations using forward-in-time simulations.

6.3.2 Limitations

As most of our analyses rely on allele frequency-based methods that are sensitive to sample size (e.g. SFS comparison, DFE), we randomly sampled 15 individuals per population. However, the number of deleterious mutations and the probability to observe rare variants, e.g., strongly or extremely deleterious mutations maintained at very low frequency by

natural selection, depends on the number of samples used. Consequently, our low sample size can reduce the power to detect differences in mutation load between Oceanian and continental reference groups. Likewise, as we do not have any phenotypic data, we assessed the deleteriousness of variants using a conservation-based prediction score (here GERP score (Cooper et al. 2005)) as in (Font-Porterias et al. 2021; Henn et al. 2015b; Henn et al. 2016b; Lopez et al. 2018a; Pedersen et al. 2017a). Nevertheless, this score is not always proportional to the deleteriousness of a given variant as recently shown by (Huber, Kim, and Lohmueller 2020).

Furthermore, because Simons and Sella found that the number of derived alleles (N_{alleles}) is the only statistic directly correlated with the mutational load and not biased by demographic events (Simons and Sella 2016), we thus approximated the additive mutational load using this approach. However, Pedersen et al., based on simulations suggest that the number of derived alleles is likely underpowered to detect narrow differences in load across human groups.

Our estimates of the fixation probability (u) for a new deleterious mutation versus a neutral mutation suggest a reduction in the efficacy of natural selection especially for Polynesians and Papuan highlanders (PNG) (Table 1). Nevertheless, we also found a very poor fit between observed and expected non-synonymous SFS for Polynesians and to a lesser extinct, for Papuan highlanders (Supplementary Figure 1). This poor fit can be due to the effect of a strong genetic drift or/and a demographic history that is not well considered when fitting the DFE: the exacerbated drift experienced by Polynesians could have strongly distorted the synonymous and non-synonymous SFS but led also to fewer segregating variants (fewer SNP to fit both the 3-epoch demographic model and the DFE with $\partial a \partial i / Fit \partial a \partial i$ (Kim, Huber, and Lohmueller 2017)).

6.4 Material and Methods

Whole-genome sequencing data. HGDP (Bergstrom et al. 2020) 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.0.7.13 (Li and Durbin 2009) and duplicates were marked with Picard Tools. Base quality scores were recalibrated with the Genomic Analysis ToolKit (GATK) software v.3.8 (DePristo et al. 2011). Variant calling was performed following the GATK best-practice recommendations (McKenna et al. 2010). All samples were genotyped individually with 'HaplotypeCaller' in gvcf mode. For Malaspinas et al. (Malaspinas et al.

2016b), and Choin et al. (Choin et al. 2021) sequences we started from gvcf files generated in Choin et al. (Choin et al. 2021). The raw multisample VCF containing all individuals was then generated with the 'GenotypeGVCFs' tool. Using BCFtools v.1.9, we applied different hard quality filters on invariant and variant sites, based on coverage depth, genotype quality, Hardy–Weinberg equilibrium and genotype missingness. Heterozygosity was assessed with PLINK v.1.90 (Purcell et al. 2007; Chang et al. 2015) and cryptically related samples were detected with KING v.2.1 (Manichaikul et al. 2010). For all analyses we allowed 0% of missingness except for the description of Loss-of-Function (LoF) and missense variants where we allowed up to 10%.

Variant annotation. To investigate the burden of deleterious mutations, we restricted our analyses on bi-allelic synonymous and non-synonymous single nucleotide polymorphisms (SNPs). To do so, we first kept SNP within CDS based on a downloaded bed files containing genomic positions (hg19) of coding sequence regions (CDS) for each canonical transcript form the UCSC 'Table browser' database (https://genome.ucsc.edu/). We then classified variants into 'missense' or 'synonymous' using ensembl-vep tool (VEP) version 100.2 (McLaren et al. 2016). Stop gained loss-of-function (LoF) variants were annotated with LOFTEE (available at https://github.com/konradjk/loftee) implemented in VEP (Lipson et al. 2020).

LoF and missense variants were intersected with gnomAD (Karczewski et al. 2020) for frequency annotations. Supplementary Table 1 was generated using only Oceanian samples (without Africans, Han Chinese, French and Cebuano individuals) and including related samples.

We assessed the deleteriousness of missense variants using a reference-free method based on the sequence conservation score "GERP RS" (Cooper et al. 2005). We then classified variants according to bin of GERP score (Lopez et al. 2018a; Henn et al. 2016b; Font-Porterias et al. 2021): Neutral: $-2 \leq GERP < 2$; Moderately deleterious: $2 \leq GERP < 4$; Strongly deleterious: $4 \leq GERP < 6$; Extremely deleterious: $GERP \geq 6$. For each class of deleteriousness, we generated unfolded and folded site frequency spectra using a custom python script using the $Fit\partial a\partial i$ library (Kim, Huber, and Lohmueller 2017). We also calculated the number of segregating variants and their mean frequency per population and per category of GERP RS. 95%CI were obtained by bootstrapping by blocks of 2Mb.

Gene Ontology enrichment. To test whether LoF variants detected in Oceanians and absent or at low frequency in gnomAD (Karczewski et al. 2020) database (max frequency in gnomAD of 0.01%) targeted specific biological functions, we tested for Gene ontology

enrichment using the R package GOseq (Young et al. 2010) which corrects for gene length. We corrected p-values (multiple testing) using an FDR approach (Benjamini & Hochberg method).

DFE of new non-synonymous mutations. We used $\partial a \partial i / Fit \partial a \partial i$ (Kim, Huber, and Lohmueller 2017; Gutenkunst et al. 2009) to infer the DFE of new non-synonymous mutations. We used the synonymous and missense mutations as neutral and deleterious classes, respectively. We fitted a three-epoch demographic model to synonymous SFS per population. $Fit \partial a \partial i$ infers the mean (E(s)) of a gamma distributed DFE model, fitted on the non-synonymous SFS, accounting for demography. Parameters are scaled by $2N_{Anc}$, with N_{Anc} estimated using the following equation $\theta_s = 4N_{Anc}\ \mu L_s$ with μ the mutation rate and L_s the length of the sequence where synonymous mutations can arise. We used here a mutation rate equal to 1.5×10^{-8} (Segurel, Wyman, and Przeworski 2014) and a ratio $L_{NS}/L_S = 2.31$ (Huber et al. 2017) to estimate L_S and $L_{NS}$from $L_S + L_{NS}$. We calculated a weighted Ne over the inferred demographic changes through time as in (Lopez et al. 2018a; Font-Porterias et al. 2021). We computed the average fixation probability of a new mutation (u) by integrating over the DFE inferred for each population separately. We computed the fixation probability of a new deleterious mutation (u_{del}) and calculated the ratio of u_{del} over the fixation probability of a neutral mutation (u_{neu}) as a way to quantify the relative strength of selection versus drift at removing deleterious mutations. We calculated confidence intervals for estimated parameters by bootstrapping by site 100 times.

Approximation of the mutational load. We used the $N_{alleles}$ and N_{hom} statistics (Simons et al. 2014; Henn et al. 2016b) to approximate the additive and recessive mutational load of present-day worldwide human groups: $N_{alleles} = N_{het} + 2N_{hom}$ with N_{het} and N_{hom} corresponding to the numbers of heterozygous and derived homozygous genotypes, respectively. We stratified these summary statistics for different categories of deleteriousness based on the GERP RS score (Cooper et al. 2005). We computed the average number of $N_{alleles}$ and N_{hom} per group using a custom python script and calculated between-population ratio for each GERP score categories. We used a 2Mb-block paired bootstrapping approach to obtain the 95% confidence intervals of between-population ratios (1,000 resamples with replacement). P-values were obtained by comparing the bootstrap distributions of deleterious categories with that of the neutral category. P-values were corrected for multiple testing using an FDR (Benjamini & Hochberg) method.

Runs of homozygosity. We call Runs Of Homozygosity (ROH) with GARLIC (Szpiech, Blant, and Pemberton 2017) v1.1.6 (https://github.com/szpiech/garlic) using the weighted LOD calculation (–weighted flag) to account for linkage disequilibrium between loci and recombination events (Blant et al. 2017). We used the –auto-winsize flag to automatically

guess the best window size based on the SNP density, –gl-type GQ to account for the quality of th genotypes, –auto-overlap-frac flag and a mutation rate equal to 1.25×10^{-8}. ROH were called per population and classified in 3 clusters based on their length using a Gaussian mixture model implemented in GARLIC. We focused our analyses on the longest class of ROH, the third class (class C) because it likely represents ROH due to recent parental relatedness, isolation or recent bottleneck. Linear regressions between mutational load ($N_{alleles}$ and N_{hom}) and length of long ROH were performed using the *lm()* function of R and adjusted by the different levels of PNG ancestry (proportion taken from Choin et al. (Choin et al. 2021)) and population effect:

$lm(Load \sim PNG_{ancestry} + cumulativeROH + pop_{PNG} + pop_{Malakula} + pop_{Tanna} + pop_{Malaita} + pop_{VL} + pop_{PolOut})$

pop_x corresponds to a binary vector that takes the value "1" if the sample belongs to the population x and "0" otherwise. We performed linear regressions only for the class "Neutral", "Moderate" and "Strong" of GERP RS, because the range of the values taken by $N_{alleles}$ and N_{hom} for the last category "Extreme" was tight (discrete values). P-values were corrected for multiple testing by using an FDR (Benjamini & Hochberg) method

6.5 Bibliography

Balick, D. J., R. Do, C. A. Cassa, D. Reich, and S. R. Sunyaev. 2015. 'Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck', *Plos Genetics*, 11.

Barton, N. H., and M. Turelli. 2004. 'Effects of genetic drift on variance components under a general model of epistasis', *Evolution*, 58: 2111-32.

Bergstrom, A., S. A. McCarthy, R. Y. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanche, J. F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. L. Xue, R. Durbin, and C. Tyler-Smith. 2020. 'Insights into human genetic variation and population history from 929 diverse genomes', *Science*, 367: 1339-+.

Blant, A., M. Kwong, Z. A. Szpiech, and T. J. Pemberton. 2017. 'Weighted likelihood inference of genomic autozygosity patterns in dense genotype data', *Bmc Genomics*, 18.

Carr, R. E., N. E. Morton, and I. M. Siegel. 1971. 'Achromatopsia in Pingelap Islanders. Study of a genetic isolate', *Am J Ophthalmol*, 72: 746-56.

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *Gigascience*, 4.

Charlesworth, B. 2009. 'Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation', *Nat Rev Genet*, 10: 195-205.

Choin, J., J. Mendoza-Revilla, L. R. Arauna, S. Cuadros-Espinoza, O. Cassar, M. Larena, A. M. S. Ko, C. Harmant, R. Laurent, P. Verdu, G. Laval, A. Boland, R. Olaso, J. F. Deleuze, F. Valentin, Y. C. Ko, M. Jakobsson, A. Gessain, L. Excoffier, M. Stoneking, E. Patin, and L. Quintana-Murci. 2021. 'Genomic insights into population history and biological adaptation in Oceania', *Nature*, 592: 583-+.

Cooper, G. M., E. A. Stone, G. Asimenos, Nisc Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow. 2005. 'Distribution and intensity of constraint in mammalian genomic sequence', *Genome Res*, 15: 901-13.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43: 491-+.

Do, R., D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. 2015. 'No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans', *Nature Genetics*, 47: 126-31.

Eickhoff, S., and P. Beighton. 1985. 'Genetic disorders on the island of St Helena', *S Afr Med J*, 68: 475-8.

Font-Porterias, N., R. Caro-Consuegra, M. Lucas-Sanchez, M. Lopez, A. Gimenez, A. Carballo-Mesa, E. Bosch, F. Calafell, L. Quintana-Murci, and D. Comas. 2021. 'The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm', *Mol Biol Evol*, 38: 2804-17.

Fu, W. Q., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, J. M. Akey, and NHLBI Exome Sequencing Project. 2013. 'Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants (vol 493, pg 216, 2013)', *Nature*, 495: 270-70.

Gosling, A. L., and E. A. Matisoo-Smith. 2018. 'The evolutionary history and human settlement of Australia and the Pacific', *Current Opinion in Genetics & Development*, 53: 53-59.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. 'Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data', *Plos Genetics*, 5: e1000695.

Harris, D. N., M. D. Kessler, A. C. Shetty, D. E. Weeks, R. L. Minster, S. Browning, E. E. Cochrane, R. Deka, N. L. Hawley, M. S. Reupena, T. Naseri, S. T. McGarvey, T. D. O'Connor, Trans-Omics Precision Med, and TOPMed Population Genetics. 2020. 'Evolutionary history of modern Samoans', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 9458-65.

Henn, B. M., L. R. Botigue, C. D. Bustamante, A. G. Clark, and S. Gravel. 2015. 'Estimating the mutation load in human genomes', *Nat Rev Genet*, 16: 333-43.

Henn, B. M., L. R. Botigue, S. Peischl, I. Dupanloup, M. Lipatov, B. K. Maples, A. R. Martin, S. Musharoff, H. Cann, M. P. Snyder, L. Excoffier, J. M. Kidd, and C. D. Bustamante. 2016. 'Distance from sub-Saharan Africa predicts mutational load in diverse human genomes', *Proc Natl Acad Sci U S A*, 113: E440-9.

Huber, C. D., B. Y. Kim, and K. E. Lohmueller. 2020. 'Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution', *Plos Genetics*, 16: e1008827.

Huber, C. D., B. Y. Kim, C. D. Marsden, and K. E. Lohmueller. 2017. 'Determining the factors driving selective effects of new nonsynonymous mutations', *Proc Natl Acad Sci U S A*, 114: 4465-70.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Consortium Genome Aggregation Database, B. M. Neale, M. J. Daly, and D. G. MacArthur. 2020. 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581: 434-43.

Kim, B. Y., C. D. Huber, and K. E. Lohmueller. 2017. 'Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples', *Genetics*, 206: 345-61.

Kirch, P. V. 2017. *On the road of the winds: An archeological history of the Pacific islands before European contact* (University of California Press).

Knudson, A. G. 1979. 'Our Load of Mutations and Its Burden of Disease', *American Journal of Human Genetics*, 31: 401-13.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

Lipson, M., P. Skoglund, M. Spriggs, F. Valentin, S. Bedford, R. Shing, H. Buckley, I. Phillip, G. K. Ward, S. Mallick, N. Rohland, N. Broomandkhoshbacht, O. Cheronet, M. Ferry, T. K. Harper, M. Michel, J. Oppenheimer, K. Sirak, K. Stewardson, K. Auckland, A. V. S. Hill, K. Maitland, S. J. Oppenheimer, T. Parks, K. Robson, T. N. Williams, D. J. Kennett, A. J. Mentzer, R. Pinhasi, and D. Reich. 2018. 'Population Turnover in Remote Oceania Shortly after Initial Settlement', *Current Biology*, 28: 1157-+.

Lipson, M., M. Spriggs, F. Valentin, S. Bedford, R. Shing, W. Zinger, H. Buckley, F. Petchey, R. Matanik, O. Cheronet, N. Rohland, R. Pinhasi, and D. Reich. 2020. 'Three Phases of

Ancient Migration Shaped the Ancestry of Human Populations in Vanuatu', *Current Biology*, 30: 4846-+.

Lohmueller, K. E. 2014. 'The distribution of deleterious genetic variation in human populations', *Current Opinion in Genetics & Development*, 29: 139-46.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. 2008. 'Proportionally more deleterious genetic variation in European than in African populations', *Nature*, 451: 994-U5.

Lopez, M., A. Kousathanas, H. Quach, C. Harmant, P. Mouguiama-Daouda, J. M. Hombert, A. Froment, G. H. Perry, L. B. Barreiro, P. Verdu, E. Patin, and L. Quintana-Murci. 2018. 'The demographic history and mutational load of African hunter-gatherers and farmers', *Nature Ecology & Evolution*, 2: 721-30.

Malaspinas, A. S., M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergstrom, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. L. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvag, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F. X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev. 2016. 'A genomic history of Aboriginal Australia', *Nature*, 538: 207-+.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen. 2010. 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26: 2867-73.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. 'The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20: 1297-303.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17.

O'Brien, E., L. B. Jorde, B. Ronnlof, J. O. Fellman, and A. W. Eriksson. 1988. 'Founder effect and genetic disease in Sottunga, Finland', *Am J Phys Anthropol*, 77: 335-46.

O'Connell, J. F., J. Allen, M. A. J. Williams, A. N. Williams, C. S. M. Turney, N. A. Spooner, J. Kamminga, G. Brown, and A. Cooper. 2018. 'When did Homo sapiens first reach Southeast Asia and Sahul?', *Proceedings of the National Academy of Sciences of the United States of America*, 115: 8482-90.

Paul, D. B. 1987. 'Our Load of Mutations Revisited', *Journal of the History of Biology*, 20: 321-35.

Pedersen, C. E. T., K. E. Lohmueller, N. Grarup, P. Bjerregaard, T. Hansen, H. R. Siegismund, I. Moltke, and A. Albrechtsen. 2017. 'The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit', *Genetics*, 205: 787-801.

Posth, C., K. Nagele, H. Colleran, F. Valentin, S. Bedford, K. W. Kami, R. Shing, H. Buckley, R. Kinaston, M. Walworth, G. R. Clark, C. Reepmeyer, J. Flexner, T. Maric, J. Moser, J. Gresky, L. Kiko, K. J. Robson, K. Auckland, S. J. Oppenheimer, A. V. S. Hill, A. J. Mentzer, J. Zech, F. Petchey, P. Roberts, C. Jeong, R. D. Gray, J. Krause, and A. Powell.

2018. 'Language continuity despite population replacement in Remote Oceania', *Nature Ecology & Evolution*, 2: 731-40.

Pugach, I., A. T. Duggan, D. A. Merriwether, F. R. Friedlaender, J. S. Friedlaender, and M. Stoneking. 2018. 'The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data', *Molecular Biology and Evolution*, 35: 871-86.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*, 81: 559-75.

Segurel, L., M. J. Wyman, and M. Przeworski. 2014. 'Determinants of mutation rate variation in the human germline', *Annu Rev Genomics Hum Genet*, 15: 47-70.

Simons, Y. B., and G. Sella. 2016. 'The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives', *Current Opinion in Genetics & Development*, 41: 150-58.

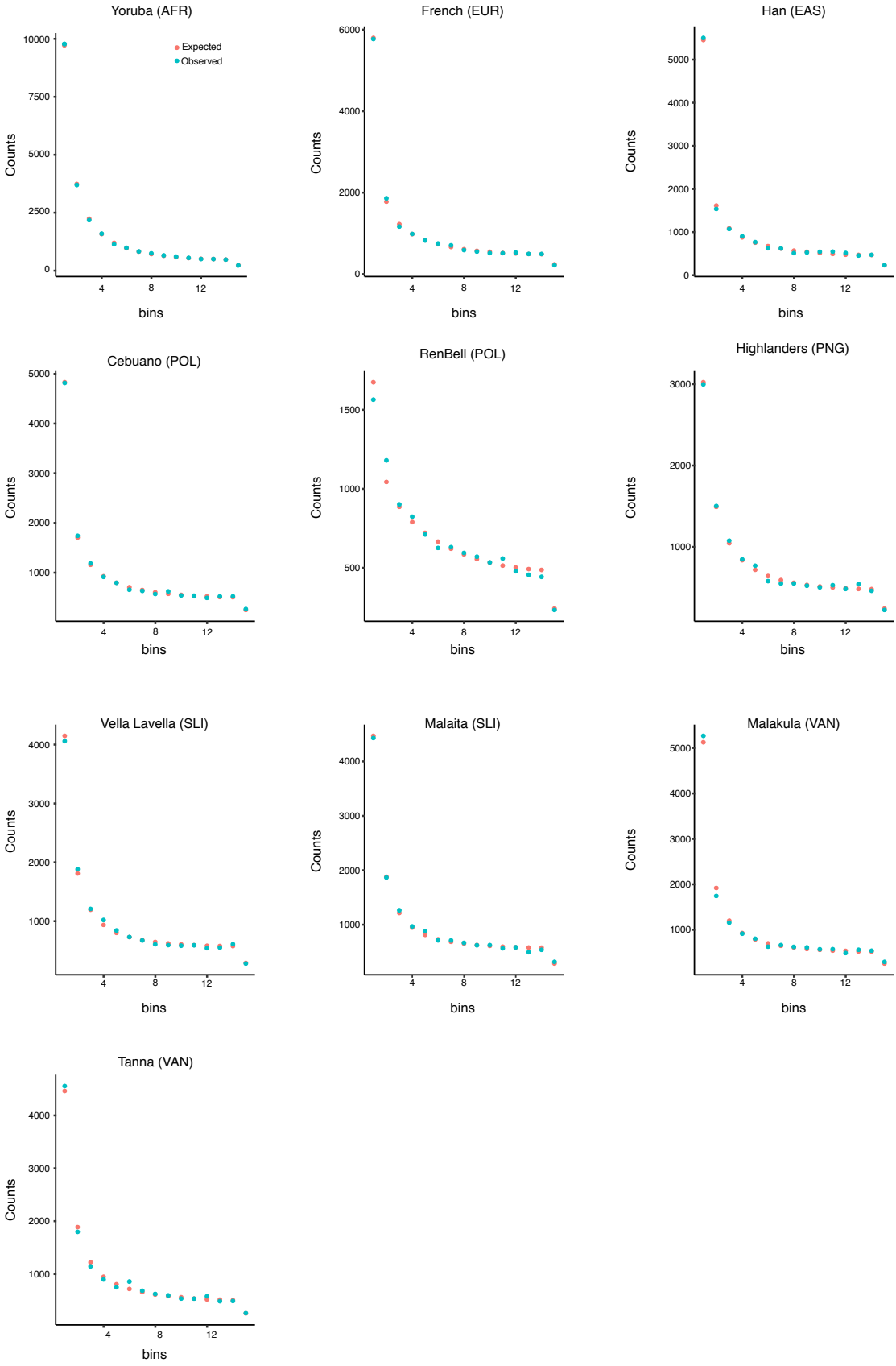
Simons, Y. B., M. C. Turchin, J. K. Pritchard, and G. Sella. 2014. 'The deleterious mutation load is insensitive to recent population history', *Nature Genetics*, 46: 220-+.

Sirugo, G., S. M. Williams, and S. A. Tishkoff. 2019. 'The Missing Diversity in Human Genetic Studies', *Cell*, 177: 26-31.

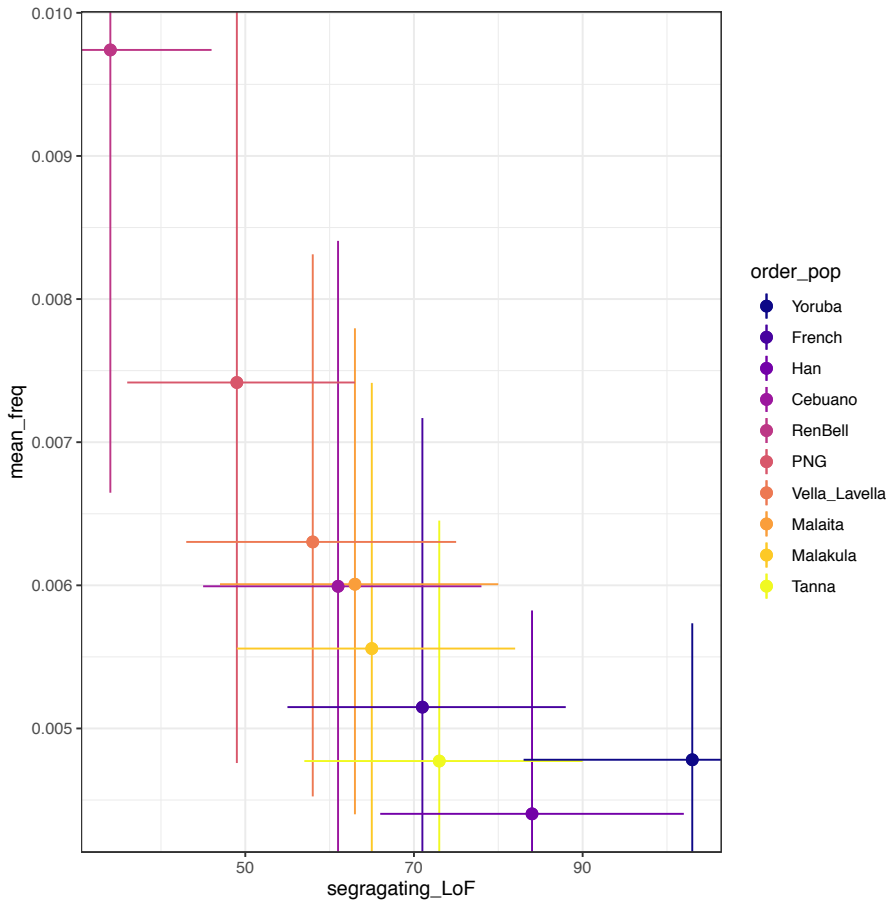
Szpiech, Z. A., A. Blant, and T. J. Pemberton. 2017. 'GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification', *Bioinformatics*, 33: 2059-62.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack. 2010. 'Gene ontology analysis for RNA-seq: accounting for selection bias', *Genome Biology*, 11: R14.

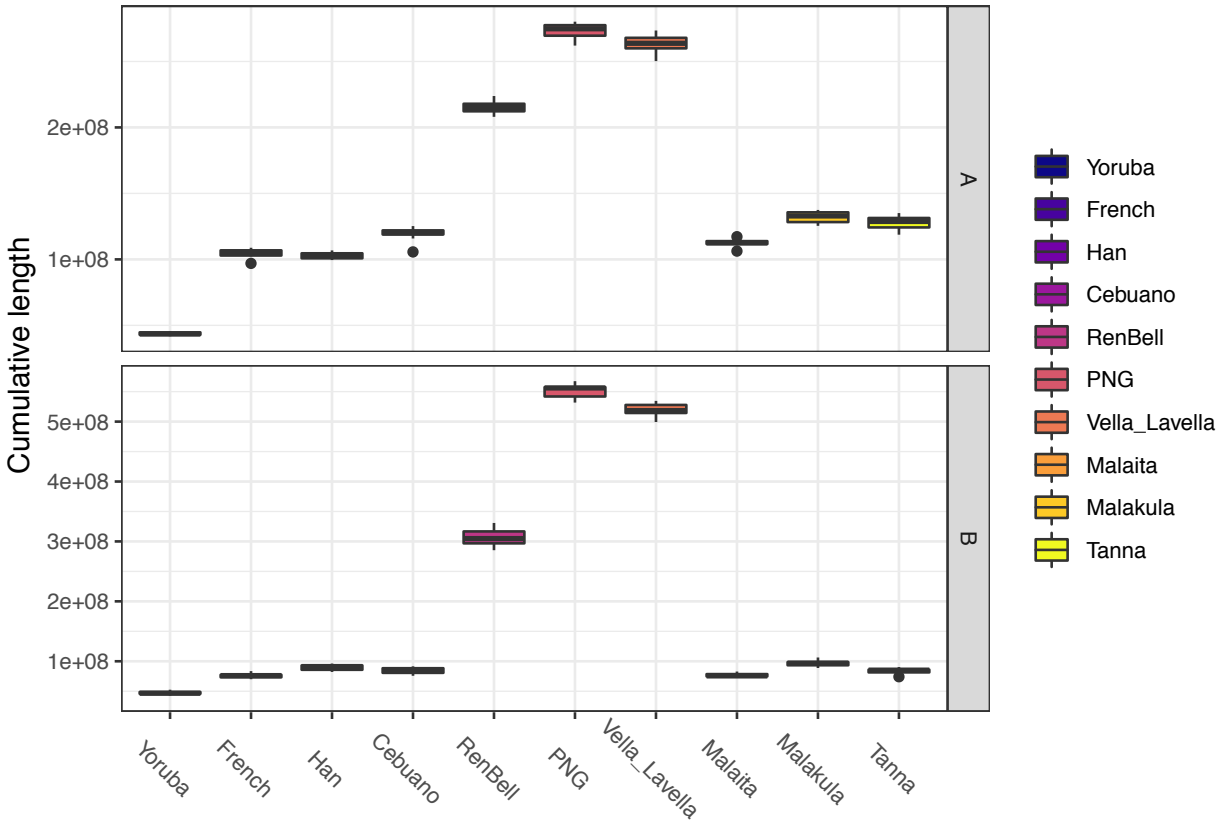
6.6 Supplementary information



Supplementary Figure 1. Fitted non-synonymous SFS with Fit∂a∂i. Observed (blue) and expected (salmon) 1 dimensional folded SFS (n=15 for each group).



Supplementary Figure 2. Number of segregating Loss-of-Function mutations and their mean frequency per group. Dots indicate point estimates and lines give the 95% confidence intervals obtained by block bootstrap.



Supplementary Figure 3. Cumulative length of ROH for class A and B per group.

Supplementary Table 1. **LoF variants present in 308 Oceanians (including related samples, Papuan highlanders, Solomon Islanders, Reef/Santa Cruz islanders, Ni-Vanutu, RenBell and Tikopia Polynesians outliers) absent or at low frequency in gnomAD (< 1/1000).** Count gives the number of LoF alleles in the 308 Oceanians individuals; n(HET), the number of heterozygous genotypes; n(HOM), the number of LoF homozygote genotypes; max Freq gnomAD give the maximal frequency observed in gnomAD and max Pop gnomAD, the population where the maximal frequency is observed.

[provided as a excel file]

Supplementary Table 2. **Top 20 Gene ontology (GO)**. P.values provided here are uncorrected for multiple testing (adjusted p.value = 1 for all GO categories).

category	p.value	term
GO:0062023	0.00021261	collagen-containing extracellular matrix
GO:0004867	0.00068207	serine-type endopeptidase inhibitor activity
GO:0032982	0.00177242	myosin filament
GO:0051015	0.0021489	actin filament binding
GO:0035381	0.00231927	ATP-gated ion channel activity
GO:0032838	0.00261907	plasma membrane bounded cell projection cytoplasm
GO:0045742	0.00285371	positive regulation of epidermal growth factor receptor signaling pathway
GO:0036289	0.00299293	peptidyl-serine autophosphorylation
GO:2001225	0.00308808	regulation of chloride transport
GO:1901186	0.0034293	positive regulation of ERBB signaling pathway
GO:0042491	0.0036445	inner ear auditory receptor cell differentiation
GO:0060401	0.00393939	cytosolic calcium ion transport
GO:0019896	0.00398365	axonal transport of mitochondrion
GO:0019428	0.00413488	allantoin biosynthetic process
GO:0019628	0.00413488	urate catabolic process
GO:0035253	0.00434579	ciliary rootlet
GO:0031012	0.00467057	extracellular matrix
GO:0097014	0.00521012	ciliary plasm
GO:0005770	0.0053969	late endosome
GO:0008092	0.00557218	cytoskeletal protein binding

Supplementary Table 3. Adjusted p-values (bootstraping) of ratios of the mean per-individual number of derived alleles ($N_{alleles}$) and derived genotypes (N_{hom}) between Pacific islanders and non-Oceanian groups represented by Yoruba Africans, Han Chinese and French Europeans.

Pop	Moderate_Nalleles	Strong_Nalleles	Extreme_Nalleles	Moderate_Nhom	Strong_Nhom	Extreme_Nhom
Cebuano/Yoruba	0.541745825	0.427097290	0.833013728	0.687717435	0.071992801	0.833013728
Cebuano/French	0.554344566	0.699623788	0.935401197	0.833013728	0.833013728	0.884071593
Cebuano/Han	0.972502750	0.817742226	0.833013728	0.833013728	0.972502750	0.699623788
RenBell/Yoruba	0.173582642	0.474252575	0.928257174	0.341135117	0.025197480	0.687717435
RenBell/French	0.116538346	0.733246675	0.972502750	0.341135117	0.435706429	0.741225877
RenBell/Han	0.427097290	0.833013728	0.960468659	0.525787421	0.452621405	0.554344566
PNG/Yoruba	0.427097290	0.341135117	0.833013728	0.683131687	0.029397060	0.607259274
PNG/French	0.442467753	0.554344566	0.928257174	0.733246675	0.569326851	0.622562134
PNG/Han	0.833013728	0.687717435	0.833013728	0.972502750	0.687717435	0.435706429
Vella_Lavella/Yoruba	0.451616377	0.427097290	0.972502750	0.588162236	0.025197480	0.833013728
Vella_Lavella/French	0.435706429	0.687717435	0.833013728	0.687717435	0.591171652	0.928257174
Vella_Lavella/Han	0.846338896	0.833013728	0.907822425	0.960468659	0.687717435	0.822565112
Malaita/Yoruba	0.435706429	0.427097290	0.928257174	0.687717435	0.025197480	0.736405233
Malaita/French	0.427097290	0.687717435	0.960468659	0.833013728	0.627646326	0.833013728
Malaita/Han	0.833013728	0.789060820	0.960468659	0.867132518	0.699623788	0.627646326
Tanna/Yoruba	0.687717435	0.427097290	0.833013728	0.833013728	0.029397060	0.733246675
Tanna/French	0.733246675	0.687717435	0.935401197	0.972502750	0.627646326	0.833013728
Tanna/Han	0.849375257	0.833013728	0.833013728	0.699623788	0.733246675	0.687717435
Malakula/Yoruba	0.530772729	0.341135117	0.960468659	0.733246675	0.029397060	0.808702913
Malakula/French	0.525787421	0.569326851	0.928257174	0.833013728	0.687717435	0.833013728
Malakula/Han	0.928257174	0.688446410	0.972502750	0.833013728	0.833013728	0.697130287

Supplementary Table 4. Adjusted p-values of the linear regressions ($N_{alleles}$ / N_{hom} ~ $PNG\ ancestry + cumulative\ ROH + pop\ PNG + pop\ Malakula + pop\ Tanna + pop\ Malaita + pop\ VL + pop\ PolOut$). PNG means Papua New Guinea(n) and pop, population

	Nalleles Neutral	Nalleles Moderate	Nalleles Strong	Nhom Neutral	Nhom Moderate	Nhom Strong
PNG ancestry	8.17E-01	9.30E-01	9.30E-01	3.15E-01	8.27E-01	8.61E-01
cumulative long ROH	1.42E-01	8.61E-01	9.84E-01	2.13E-02	6.66E-03	2.70E-03
pop PNG	5.48E-01	9.30E-01	9.30E-01	3.23E-01	8.17E-01	8.61E-01
pop Malakula	7.44E-01	9.69E-01	9.30E-01	3.24E-01	8.46E-01	8.17E-01
pop Tanna	7.44E-01	9.84E-01	9.30E-01	3.38E-01	8.17E-01	8.17E-01
pop Malaita	7.40E-01	9.30E-01	9.30E-01	1.65E-01	7.44E-01	8.61E-01
pop Vella Lavella	6.22E-01	9.20E-01	9.30E-01	7.64E-02	5.48E-01	9.84E-01
pop Polynesian Outliers	5.80E-02	5.50E-01	9.84E-01	3.86E-01	8.61E-01	9.71E-01

Supplementary Table 5. Slopes of the linear regressions ($N_{alleles}$ / N_{hom} ~ $PNG\ ancestry + cumulative\ ROH + pop\ PNG + pop\ Malakula + pop\ Tanna + pop\ Malaita + pop\ VL + pop\ PolOut$). PNG means Papua New Guinea(n) and pop, population

	Nalleles Neutral	Nalleles Moderate	Nalleles Strong	Nhom Neutral	Nhom Moderate	Nhom Strong
PNG ancestry	1.48E+02	-3.29E+01	-5.17E+01	1.81E+02	5.94E+01	-5.08E+01
cumulative long ROH	2.00E-07	-4.15E-08	-1.60E-09	1.50E-07	1.42E-07	1.63E-07
pop PNG	-2.38E+02	3.26E+01	4.74E+01	-1.74E+02	-6.15E+01	5.05E+01
pop Malakula	-1.52E+02	1.98E+01	4.46E+01	-1.50E+02	-4.82E+01	5.69E+01
pop Tanna	-1.54E+02	9.36E+00	4.11E+01	-1.49E+02	-5.51E+01	5.99E+01
pop Malaita	-1.02E+02	2.02E+01	2.24E+01	-1.15E+02	-4.37E+01	2.49E+01
pop Vella Lavella	-1.18E+02	3.54E+01	3.31E+01	-1.38E+02	-5.97E+01	2.31E+00
pop Polynesian Outliers	-1.33E+02	5.09E+01	-1.87E+00	-4.30E+01	-1.23E+01	-3.11E+00

Chapter 7

Discussion

7.1 A complex demographic history

7.1.1 Near Oceania: A highly structured region

Previous demographic inferences estimated a deep divergence time between northern and southern Sahul (New Guineans and Aboriginal Australians) occurring at least 37,000 years ago (Malaspinas et al. 2016a). Our work on the demographic history of Near Oceanian populations confirms but also extends these findings to the two other archipelagos that compose the region of Near Oceania. This population structure is among the oldest estimated at a scale of a continent, excluding Africa. Indeed, the genetic isolation of the different Near Oceanian groups is almost as old as between Europeans and East Asians (Wollstein et al. 2010; Malaspinas et al. 2016a). Similarly, at a finer scale, a study (Bergstrom et al. 2017) based on SNP-array genotyping data of 381 Papua New Guineans shed light to a strong intra New Guinea population structure, between lowlanders and highlanders dated back to around 20,000 years ago. However, the genetic structure within highlanders is more recent, dated back to around 10,000 years ago (Bergstrom et al. 2017). Archaeological studies indicate an *in situ* emergence of agriculture in highland New Guinea around 10,000 years ago (Golson et al. 2017) and more recently, the associated "Neolithic" behaviour changes (social and economic changes) between 5,050 and 4,200 years ago (Shaw et al. 2020). This congruence between archaeology and genetic data suggests that the spread of agriculture in the highland of New Guinea played a key role in re-shaping after the initial settlement, the genetic makeup of New Guinean highlanders and could explain in part, the strong but recent genetic structure observed today in the region (Bergstrom et al. 2017). Likewise, a recent study also indicates that environmental factors (e.g. climate, topography) are not enough to explain the current geographic distribution of New Guinean languages and that other factors such as population movement can also explain the language diversity observed today in the region (Antunes et al. 2020).

Very little is known about the peopling history, population structure and time of divergence between Solomon islanders. Studies based on SNP array genotyping datasets (Pugach et al. 2018b; Isshiki et al. 2020) or mtDNA and Y-chromosome datasets (Delfin et al. 2012) point towards a different demographic history between western and eastern islands of the archipelago. Western Solomon islands were peopled at least 30,000 years ago as attested by the only Pleistocene archaeological site of the Buka island (Wickler and Spriggs 1988). Some Solomon islanders carry specific NRY lineages dated back to around 9,500 years ago (Delfin et al. 2012). Interestingly, we inferred younger divergence time between Solomon Islanders and other Near Oceanians when replacing Vella Lavella

western Solomon Islanders by Malaita eastern Solomon Islanders (around 20,000 years ago versus around 9,500 years ago) supporting a different peopling history of these two parts of the archipelago. However, this result needs to be confirmed and extended by for example, reconstructing the joint demographic history of different Solomon Island groups.

Our demographic inference of Near Oceania islanders is based on a limited number of groups, only one per archipelago, mainly because of the complexity of the demographic models and the limits of our approach (see 7.2 Inferring demographic models with SFS-based methods). How well do our models represent the peopling history of Near Oceania? As previously mentioned, the intra-archipelago genetic structure could be very high because of an early isolation of the different groups or because of different admixture histories within islands/archipelagos and/or with East/Southeast Asians. Consequently, many divergent genetic lineages might not be represented in our models and it is also possible that unsampled groups have a different population history. In addition, the different volcanic eruptions that occurred during the Pleistocene and Holocene periods forced different groups to migrate, to colonize new territories or to replace other groups (Torrence et al. 2004; Torrence, Neall, and Boyd 2009). It is thus likely that some Pleistocene groups disappeared and did not contributed to the current gene pool. Ancient DNA would help to better portrait the genetic makeup of the first settlers and better understand the current genetic diversity of Near Oceania.

7.1.2 Dissociating language, culture and genes?

The study of the Austronesian language phylogeny supports a "pulse-pause" model of the Pacific settlement from Taiwan (Gray, Drummond, and Greenhill 2009). The analyses indicate a first pause between Taiwan and the Philippines around 4,000 years ago and a second pause in western Polynesia around 2,800 years ago, as predicted by the "Out-of-Taiwan" model (Bellwood 1997). These results are also in agreement with ancient DNA studies, which linked the first Lapita settlers of Remote Oceania to Taiwan and the Philippines (Posth et al. 2018; Lipson et al. 2020; Lipson et al. 2018). However, a recent genetic study (Larena et al. 2021) based on 1,028 individuals from the Philippines and on two ancient individuals dated to around 8,000 year ago from the Liang islands (between Mainland East Asia and Taiwan) questioned the language-culture package proposed by the "Out-of-Taiwan" model. The two Liangdao samples form the oldest link between Mainland East Asians and present-day Austronesian speakers. The analyses suggest that the Cordilleran Austronesian speakers migrated into the Philippines before the start of agriculture in the region. Similarly, our demographic models of Formosan (using Taiwanese aborigines) and Malayo-Polynesian speakers (using Kankanaey Filipinos,

Solomon islanders and Polynesians) suggest a population structure of Austronesian speakers that predates the appearance of agriculture in Taiwan and Southeast Asia. Furthermore, the analysis of Oryza japonica (Gutaker et al. 2020), the main cultivated subspecies of rice, revealed a very recent diffusion in Island Southeast Asia starting around 2,500 years ago. Altogether, these different studies suggest that the spread of agriculture in Islands Southeast Asia and later in Oceania is not the consequence of a demic diffusion but rather a diffusion of ideas involving limited gene flow.

7.1.3 Multiple origins and/or migrations for the Lapita?

The archaeologist Noury, based on the analysis of the motifs found on Lapita potteries, hypothesized that multiple founder groups were at the origin of the Bismarck Lapita societies (Noury and Galipaud 2011; Noury 2005). These multiple migrations would originate from Island Southeast Asia (Borneo/Sulawesi) and the Philippines through the Marianna islands in Micronesia. The Marianna archipelago is of great interest since archaeological and paleoenvironmental evidence suggest the presence of the Lapita culture as old as in the Bismarck archipelago around 3,500 years ago or even older, up to 4,500 years ago (Carson 2020; Athens and F. 2004). Recently the study of two skeletons from Guam island (Marianna islands) dated to around 2,200 years ago revealed that the first settlers of the archipelago likely came from the Philippines (Pugach et al. 2021). This study also highlights the close genetic relationship between the two Guam individuals and the Lapita individuals from the Vanuatu and Tonga. This suggests an alternative route through the Marianna islands for the peopling of western Remote Oceania and ultimately Polynesia. Hence, although not formally tested, this study proposes an alternative hypothesis that gives to Micronesia a key role in the peopling history of Polynesia (Figure 7.1). This archipelago and more generally Micronesia should thus receive careful attention and be further investigated.

Our analyses indicate multiple interactions between East/Southeast Asia and Near Oceania, at least two, with the most recent gene flow dated at around 1,500 years ago. In 1992, from the study of western Pacific rock arts, C. Ballard proposed that the Austronesian Painting Tradition (APT) did not spread from Taiwan or the Philippines with the initial Austronesian settlers but rather more recently from Island Southeast Asia around 2,000 years ago (Ballard 1992). Does the second gene flow that we detected correspond to most recent Island Southeast Asian influences in Near Oceania? Does it reflect a settlement of western Remote Oceania and Polynesia via multiple routes (i.e. through New Guinea and the Bismarck or through the Marianna Islands)? The mode and tempo of East/Southeast Asian gene flow should be further evaluated and extended to other Near and Remote

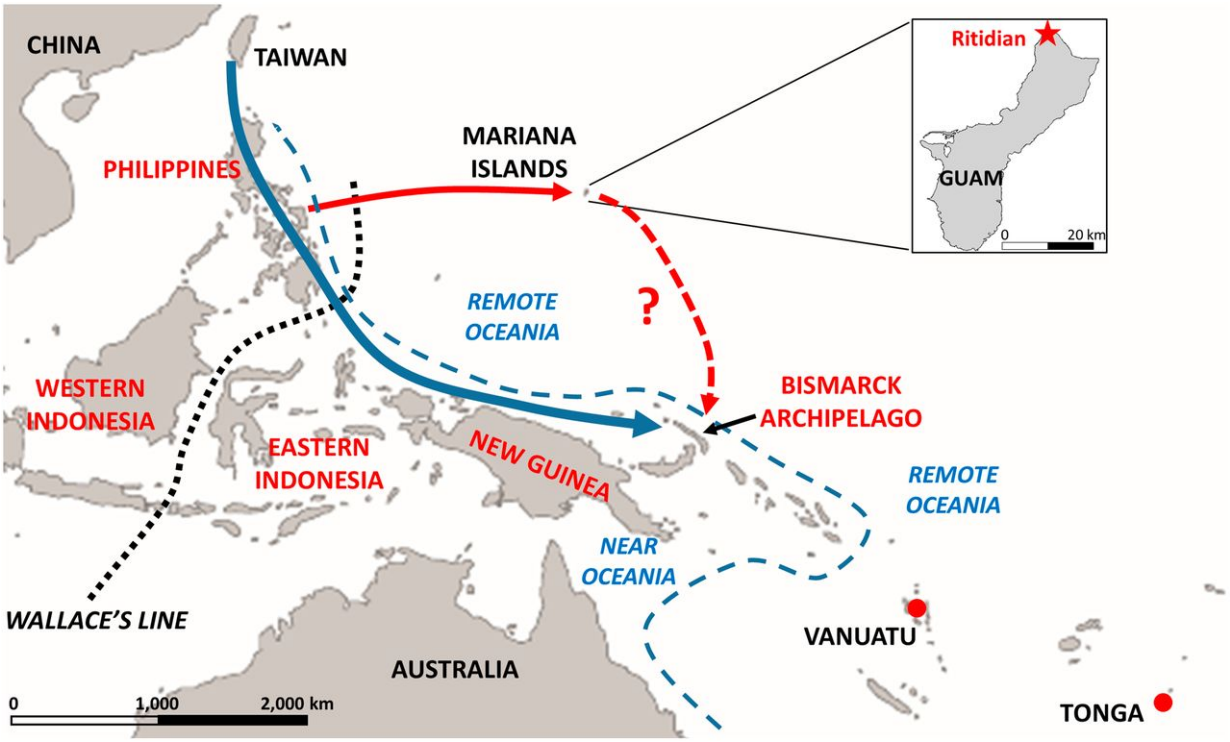


Figure 7.1: **Routes taken by the settlers of Remote Oceania (Pugach et al. 2021)**. Red dots indicate the locations of the Lapita samples from Vanuatu and Tonga. The blue and red arrows indicate the standard route taken by Austronesian speakers and the route for the peopling of the Mariana Islands respectively. The dashed red arrow indicates the likely alternative route for the peopling of Remote Oceania.

Oceanian groups. Ancient DNA from Near Oceania before 3,500 years ago will also give insight about the hypothesis of early Holocene (around 6,000 years ago or even before) connexions between Island Southeast Asia and Near Oceania.

7.2 Inferring demographic models with SFS-based methods

7.2.1 Obtaining unbiased estimates

One of the main objectives of my thesis was to reconstruct the demographic history of Near and Remote Oceanians using modern DNA. The aim of the analyses was to estimate unbiased demographic parameters and particularly, times of divergence between the different Near and Remote Oceanians groups. The admixed nature of Oceanian islanders makes inaccurate the use of standard methods such as Relate (Speidel et al. 2019) or MSMC (Schiffels and Durbin 2014) to estimate divergence times between groups or effective population sizes. The joint estimations of the parameters characterizing the demographic past of Near and Remote Oceanians as well as of the East/Southeast Asian ancestry of these groups were performed using a multidimensional SFS-based inference and the maximum likelihood framework implemented in *Fastsimcoal2* (Excoffier et al. 2013; Excoffier et al. 2021). The SFS is a powerful summary statistic to infer part of the demographic parameters including effective populations size or divergence time (Gutenkunst et al. 2009; Excoffier et al. 2013; Excoffier et al. 2021; Marchi, Schlichta, and Excoffier 2021). However, in some cases, the same SFS can be explained by different demographic scenarios (Terhorst and Song 2015; Myers, Fefferman, and Patterson 2008) and other, more informative summary statistics can be used to infer the number, the nature, the time and rate of genetic interactions between populations (Cooke and Nakagome 2018; Gravel 2012; Liang and Nielsen 2014). For these reasons, asymmetric and symmetric migrations between geographically close groups (migrations following a stepping stone model) as well as single pulse admixture events should be considered in these analyses more as nuisance parameters. Questions related to gene flow between Near Oceanians and East/Southeast Asians as well as archaic introgression with both Neanderthal and Denisovan archaic hominins, were instead investigated in detail using an ABC approach with informative summary statistics.

To obtain unbiased estimates (less biased as possible) of the divergence times, we considered in all our models: (i) Neanderthal and Denisovan introgression events,

(ii) continuous migrations between neighbour groups (asymmetric or symmetric), (iii) East/Southeast Asian gene flow and (iv) effective population size changes over time including bottlenecks, population contractions and expansions. In some of our models, we included ghost populations to capture gene flow from unsampled groups and also population structure. The number of parameters increases very rapidly which limits the number of populations that can be included in the models. For some of them, the parameter space was very large increasing the risk of overfitting and model misspecification (Marchi, Schlichta, and Excoffi 2021; Terhorst and Song 2015). Although, we assessed the accuracy and uncertainty of the estimated parameters, it is almost impossible to ensure that the likelihood converged towards the global maximum.

7.2.2 Obtaining uncertainty of the estimates

Unlike ABC approaches enabling the calculation of the 95% confidence intervals, the algorithm implemented in *Fastsimcoal2* does not. To do so, we calculated the 95% confidence intervals using a non-parametric block-bootstrap approach as recommended in the literature (de Manuel et al. 2016; Malaspinas et al. 2016a; Sikora et al. 2019) and *Fastsimcoal2* best practices (https://groups.google.com/g/fastsimcoal). The individuals used to represent the different sampled populations were selected based on their mean sequencing coverage but also based on other analyses such as PCA and ADMIXTURE. We considered through our bootstrapping strategy that all the variability of the demographic parameter estimation came from the selected independent regions of the genome (each haplotype has its own history) and not from the individuals. However, the individuals chosen to represent a population can also be a source of variability, for example in case of recent admixture there is a variance in the ancestry proportions. This variability could thus be considered in the resampling strategy when calculating the uncertainty of the parameters.

For most of the model that we inferred, we replicated the same models replacing groups by others from the same archipelago, for example Vella Lavella by Malaita Solomon Islanders or Malakula by Emae Ni-Vanuatu. For some of the parameters the confidence intervals were overlapping, strengthening the accuracy of these estimates. Nevertheless, what does it mean for those that did not replicate? It is challenging to know whether it is because these parameters are not correctly inferred or because it reflects true differences in the demographic past of these groups.

7.2.3 Model comparison

I would like also to discuss about the model comparison with *Fastsimcoal2*. The SFS that we used for model inference were build using all SNP found outside genes and CpG islands. Because of the presence of linked-SNP, the likelihood computed by *Fastsimcoal2* is a composite likelihood. Composite likelihood provides unbiased parameter estimates but the likelihood itself is inflated and cannot be used for model comparison using AIC or BIC. An alternative approach has been proposed which consists in re estimating (100 times) with more simulations the likelihood of each model included in the comparison (de Manuel et al. 2016; Malaspinas et al. 2016a; Sikora et al. 2019). We considered that a model was the most likely if the initial expected log10(likelihood) under this model is higher than that of the alternative models, and the difference between the mean of the 100 re-estimated log10(likelihoods) of this model and that of other models is higher than 50 as in (Sikora et al. 2019). We estimated that using these criteria, the true model was selected in 81% of the cases, but it is likely that this true positive rate depends on the complexity of the demographic model to estimate. Therefore, the threshold used to consider a model as the most likely should be a priori evaluated and adapted to each demographic model tested.

7.2.4 "All models are wrong but some are useful"

Finally, it is important to mention that all demographic inferences rely on mathematical simulations and thus on assumptions that, if violated, can lead to biased estimates. For example, coalescent simulations and demographic inferences assume neutrality. However, part on neutral variants found outside genes can also be affected by linked selection, especially background selection which corresponds to the elimination of neutral variants owing to negative selection acting on linked deleterious mutations. Genomic regions affected by background selection have a lower genetic diversity mimicking a signal of low effective population size and recent expansion (Ewing and Jensen 2016; Marchi, Schlichta, and Excoffi 2021; Schrider, Shanku, and Kern 2016). Moreover, the dates of events in absolute time ("years ago") also rely on two parameters: the mutation rate and the generation time. Despite the mutation rate varies greatly along the genome (as mentioned in chapter 2) we assumed in our simulations a constant mutation rate, 1.25×10^{-8} per site and per generation as in (Malaspinas et al. 2016a; Schiffels and Durbin 2014; Sikora et al. 2019).

The space of possible demographic models is very large and it is not possible to explore all of them. Models presented in chapter 5 correspond to the most likely demographic

scenarios among a subset of models that we compared. Despite I referred in this thesis to the "inference of complex demographic models", all models presented here (and in other studies) are too simple to represent the real demographic history of Oceanian populations. However, simple models are sometimes useful to test and rise hypotheses as well as to pave the way for a better understanding of the complex peopling and demographic history of human populations.

7.3 Future directions

7.3.1 Toward fine-scale and transdisciplinary studies

Our work, combined with other recent genomic studies (Posth et al. 2018; Lipson et al. 2018; Pugach et al. 2018b; Malaspinas et al. 2016a), provide a more detailed picture of the population history and the genetic diversity observed today in Near and Remote Oceania. However, the answers provided still remain too general for this region of the world with the richest cultural and linguistic diversity combined with a very deep continental genetic structure.

We have seen so far in this thesis (i) the old divergence time and deep population structure of Near Oceanians, (ii) the role of migrations (Lapita and post-Lapita) in shaping the biological and cultural makeup of Remote Oceanians and (ii) the heterogeneity in the admixture history of the different groups. How and when were Central and Eastern Solomon Islands peopled? Is there a population continuity in this archipelago since initial settlement? Who is/are the Lapita people(s)? Did the Lapita societies originate from one or multiple sources? Who were the first western Remote Oceanians and Polynesians? What were the consequences of Europeans in the population structure and health of Oceanians? To my mind, these questions should be addressed by considering the inputs of different disciplines and by shifting toward fine-scale studies: archipelago, island or even burial-based studies (for ancient DNA) as suggested by K.R Veeramah (Veeramah 2018). Generally, population geneticists try to validate/invalidate hypotheses or models proposed by archaeologists, anthropologists and linguists. I think, it is time now that new hypotheses and new models emerge from a common and constructive discussion between disciplines.

In her review, K.R Veeramah (Veeramah 2018) discussed the issue with the concept of "migration" that tends to be simplified in genomic and paleogenetic studies. First, because geneticists typically use only a limited number of samples to represent a likely

socially heterogenous group and second, because paleogenetic studies do not address the question of the nature of the migrations (e.g. back migrations, leapfrogging, continuous). While Valentin et al. (Valentin et al. 2016; Valentin et al. 2014) referred to "Secondary movement of people", geneticists referred to "population replacement". In this context, transdisciplinary approaches will allow to harmonize the different terms that are used by both archaeologists and geneticists but that do not have the same definition or precision (e.g. "Population replacement" vs "Secondary movement of people"). This approach will also bring to the same level stories, sometimes different, tell by genes, language and culture.

The perception of blood, DNA or other part of the body changes from culture to another. Similarly, the destruction of bones to extract ancient DNA or the post-mortem manipulation of the body in sacred lands can sometimes be perceived as unethical by autochthonous groups (e.g. the study of the Kennewick man and other native American groups (Wagner et al. 2020; Rasmussen et al. 2015; Bhattacharya et al. 2018)). For decades, it was very complicated to obtain DNA samples from Pacific groups, and although some communities recently consented to be part of genetics studies, others still refuse. For example, from 2017 the customary senate of New Caledonia refuses the involvement of indigenous Kanak people in population genetic studies. Archaeologists, anthropologists and linguists, spend a tremendous amount of time in the fieldwork, where they create stable relationships, communicate with local authorities and engage, when desired, autochthonous groups. In this regard, I am personally convinced that geneticists will benefit from this collaborative and transdisciplinary work.

7.3.2 Lack of diversity in databases

As recently as January 2019, around 78% of individuals found in Genome Wide Association Studies (GWAS) were of European descent while only less than 0.20% were Oceanians (Figure 7.2) (Sirugo, Williams, and Tishkoff 2019). More generally, genomic resources such as ascertained SNP arrays (Lachance and Tishkoff 2013) or databases such as HGDP (Bergstrom et al. 2020) and gnomAD (Karczewski et al. 2020), epidemiological and clinical databases, e.g., ClinVar (Landrum et al. 2018) or OMIM (Amberger et al. 2015) are European-centered. However, many studies point to the low portability, from an ancestry to another, of SNP effect sizes and corresponding polygenic scores, i.e., individual's genetic predisposition score for a given tested trait (Peterson et al. 2019; Majara et al. 2021; Amariuta et al. 2020; Sirugo, Williams, and Tishkoff 2019). This is a burning issue since variants associated with diseases in Oceanians or other underrepresented populations (e.g., Native Americans) but rare in Europeans, are still lacking (Kessler et al. 2019; Kessler et al. 2016; Landry et al. 2018).

Future works need to include phenotypic data of Pacific groups especially metabolic-related traits in order to identify population specific genetic variants associated with such disorders. Similarly, future work would also include new whole genome sequences or whole exome sequences of more Pacific groups such as western and eastern Polynesians who are almost completely absent from genetic databases. These new data with a reduced ascertainment bias, combined with phenotypic, time transect ancient DNA, archaeological, anthropological and linguistic data will allow (i) to better portrait the biological diversity of the region, (ii) to fill in some gaps in the peopling history of Near and Remote Oceania, (iii) evaluate the impact of Europeans on population structure and health of Oceanians, (iv) shed light on biological functions that contributed to adaption to their insular environment and (v) to better understand their present-day relation to diseases.

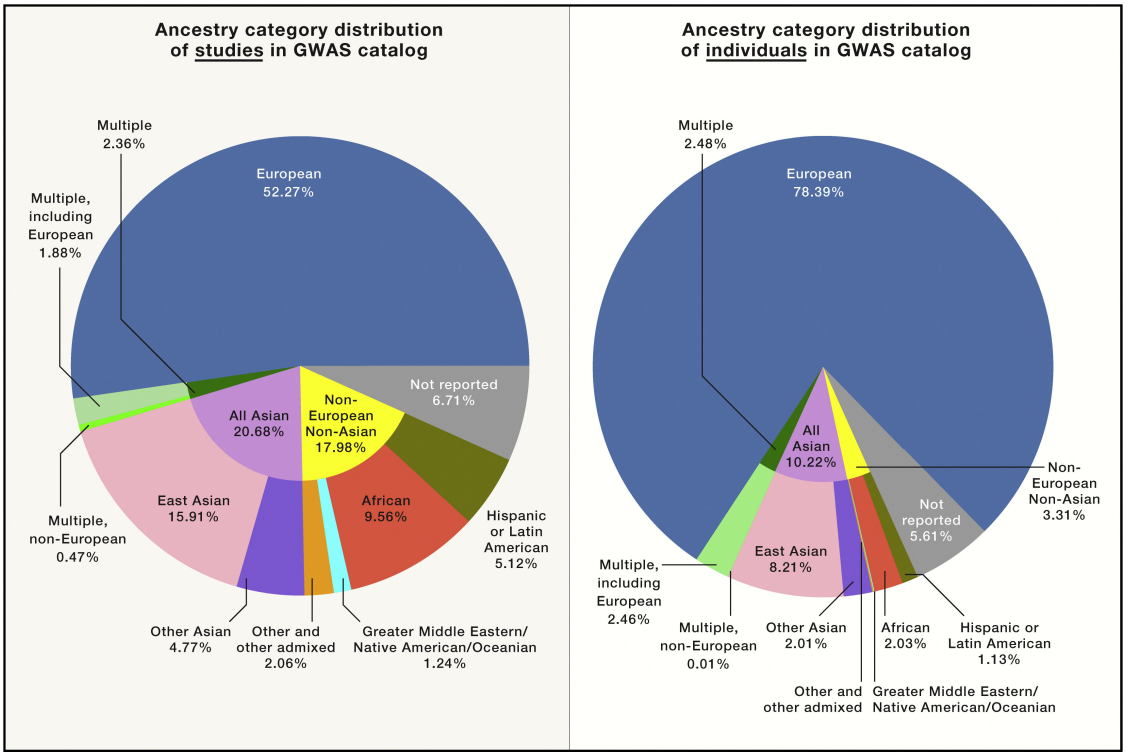


Figure 7.2: **Ancestry distribution in GWAS Catalog studies (January 2019) (Sirugo, Williams, and Tishkoff 2019)**. Percentage of each ancestry based either on studies (left) or on the total number of individuals in GWAS studies (right).

Bibliography

Amariuta, T., K. Ishigaki, H. Sugishita, T. Ohta, M. Koido, K. K. Dey, K. Matsuda, Y. Murakami, A. L. Price, E. Kawakami, C. Terao, and S. Raychaudhuri. 2020. 'Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements', *Nature Genetics*, 52: 1346-54.

Amberger, J. S., C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh. 2015. 'OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders', *Nucleic Acids Res*, 43: D789-98.

Antunes, N., W. Schiefenhovel, F. D'Errico, W. E. Banks, and M. Vanhaeren. 2020. 'Quantitative methods demonstrate that environment alone is an insufficient predictor of present-day language distributions in New Guinea', *PLoS One*, 15.

Athens, J. S., and Dega M. F. 2004. 'Austronesian colonization of the Mariana Islands: The palaeoenvironmental evidence', *Bull. Indo-Pacific Prehistory Assoc.*, 24: 21-30.

Ayub, Q., L. Moutsianas, Y. Chen, K. Panoutsopoulou, V. Colonna, L. Pagani, I. Prokopenko, G. R. Ritchie, C. Tyler-Smith, M. I. McCarthy, E. Zeggini, and Y. Xue. 2014. 'Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes', *Am J Hum Genet*, 94: 176-85.

Balick, D. J., R. Do, C. A. Cassa, D. Reich, and S. R. Sunyaev. 2015. 'Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck', *Plos Genetics*, 11.

Ballard, C. 1992. *Painted rock art sites in western Melanesia: locational evidence for an 'Austronesian' tradition* (Australian Rock Art Research Association).

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. 2008. 'Natural selection has driven population differentiation in modern humans', *Nat Genet*, 40: 340-5.

Barreiro, L. B., and L. Quintana-Murci. 2010. 'From evolutionary genetics to human immunology: how selection shapes host defence genes', *Nature Reviews Genetics*, 11: 17-30.

Barton, N. H., and M. Turelli. 2004. 'Effects of genetic drift on variance components under a general model of epistasis', *Evolution*, 58: 2111-32.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. 'Approximate Bayesian computation in population genetics', *Genetics*, 162: 2025-35.

Bedford, S., and M. Spriggs. 2008. 'Northern Vanuatu as a Pacific Crossroads: The Archaeology of Discovery, Interaction, and the Emergence of the "Ethnographic Present"', *Asian Perspectives*, 47: 95-120.

———. 2019. *Debating Lapita: Distribution, Chronology, Society and Subsistence* (ANU Press).

Bellwood, P. 1997. *Prehistory of the Indo-Malaysian Archipelago: Revised Edition* (ANU Press).

Bergstrom, A., S. A. McCarthy, R. Y. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanche, J. F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. L. Xue, R. Durbin, and C. Tyler-Smith. 2020. 'Insights into human genetic variation and population history from 929 diverse genomes', *Science*, 367: 1339-+.

Bergstrom, A., N. Nagle, Y. Chen, S. McCarthy, M. O. Pollard, Q. Ayub, S. Wilcox, L. Wilcox, R. A. H. van Oorschot, P. McAllister, L. Williams, Y. L. Xue, R. J. Mitchell, and C. Tyler-Smith. 2016. 'Deep Roots for Aboriginal Australian Y Chromosomes', *Current Biology*, 26: 809-13.

Bergstrom, A., S. J. Oppenheimer, A. J. Mentzer, K. Auckland, K. Robson, R. Attenborough, M. P. Alpers, G. Koki, W. Pomat, P. Siba, Y. L. Xue, M. S. Sandhu, and C. Tyler-Smith. 2017. 'A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea', *Science*, 357: 1160-+.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. 'Genetic signatures of strong recent positive selection at the lactase gene', *Am J Hum Genet*, 74: 1111-20.

Betty, D. J., A. N. Chin-Atkins, L. Croft, M. Sraml, and S. Easteal. 1996. 'Multiple independent origins of the COII/tRNA(Lys) intergenic 9-bp mtDNA deletion in aboriginal Australians', *Am J Hum Genet*, 58: 428-33.

Bhattacharya, S., J. Li, A. Sockell, M. J. Kan, F. A. Bava, S. C. Chen, M. C. Avila-Arcos, X. Ji, E. Smith, N. B. Asadi, R. S. Lachman, H. Y. K. Lam, C. D. Bustamante, A. J. Butte, and G. P. Nolan. 2018. 'Whole-genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia', *Genome Res*, 28: 423-31.

Bindon, J. R., and P. T. Baker. 1997. 'Bergmann's rule and the thrifty genotype', *American Journal of Physical Anthropology*, 104: 201-10.

Bird, M. I., R. J. Beaman, S. A. Condie, A. Cooper, S. Ulm, and P. Veth. 2018. 'Palaeogeography and voyage modeling indicates early human colonization of Australia was likely from Timor-Roti', *Quaternary Science Reviews*, 191: 431-39.

Bird, M. I., S. A. Condie, S. O'Connor, D. O'Grady, C. Reepmeyer, S. Ulm, M. Zega, F. Saltre, and C. J. A. Bradshaw. 2019. 'Early human settlement of Sahul was not an accident', *Sci Rep*, 9: 8220.

Bittles, A. H., and H. A. Hamamy. 2010. 'Endogamy and Consanguineous Marriage in Arab Populations', *Genetic Disorders among Arab Populations, Second Edition*: 85-108.

Blant, A., M. Kwong, Z. A. Szpiech, and T. J. Pemberton. 2017. 'Weighted likelihood inference of genomic autozygosity patterns in dense genotype data', *Bmc Genomics*, 18.

Blust, R. 1999. *Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics* (Academica Sinica).

———. 2009. *The Austronesian Languages* (Research School of Pacific and Asian Studies).

———. 2019. 'The Austronesian Homeland and Dispersal', *Annual Review of Linguistics, Vol 5*, 5: 417-34.

Bradshaw, C. J. A., K. Norman, S. Ulm, A. N. Williams, C. Clarkson, J. Chadoeuf, S. C. Lin, Z. Jacobs, R. G. Roberts, M. I. Bird, L. S. Weyrich, S. G. Haberle, S. O'Connor, B. Llamas, T. J. Cohen, T. Friedrich, P. Veth, M. Leavesley, and F. Saltre. 2021. 'Stochastic models support rapid peopling of Late Pleistocene Sahul', *Nat Commun*, 12: 2440.

Bradshaw, C. J. A., S. Ulm, A. N. Williams, M. I. Bird, R. G. Roberts, Z. Jacobs, F. Laviano, L. S. Weyrich, T. Friedrich, K. Norman, and F. Saltre. 2019. 'Minimum founding populations for the first peopling of Sahul', *Nature Ecology & Evolution*, 3: 1057-63.

Buckley, H. R. 2007. 'Possible gouty arthritis in Lapita-associated skeletons from Teouma, Efate Island, Central Vanuatu', *Current Anthropology*, 48: 741-49.

Campbell, C. D., J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, L. Vives, B. J. O'Roak, P. H. Sudmant, J. Shendure, M. Abney, C. Ober, and E. E. Eichler. 2012. 'Estimating the human mutation rate using autozygosity in a founder population', *Nat Genet*, 44: 1277-81.

Carlson, J. C., S. L. Rosenthal, E. M. Russell, N. L. Hawley, G. Sun, H. Cheng, T. Naseri, M. S. Reupena, J. Tuitele, R. Deka, S. T. McGarvey, D. E. Weeks, and R. L. Minster. 2020. 'A missense variant in CREBRF is associated with taller stature in Samoans', *Am J Hum Biol*, 32: e23414.

Carr, R. E., N. E. Morton, and I. M. Siegel. 1971. 'Achromatopsia in Pingelap Islanders. Study of a genetic isolate', *Am J Ophthalmol*, 72: 746-56.

Carson, M. T. 2020. 'Peopling of Oceania: Clarifying an Initial Settlement Horizon in the Mariana Islands at 1500 Bc', *Radiocarbon*, 62: 1733-54.

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. 'The application of molecular genetic approaches to the study of human evolution', *Nat Genet*, 33 Suppl: 266-75.

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *Gigascience*, 4.

Charlesworth, B. 2009. 'Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation', *Nat Rev Genet*, 10: 195-205.

Chikhi, L., V. C. Sousa, P. Luisi, B. Goossens, and M. A. Beaumont. 2010. 'The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes', *Genetics*, 186: 983-95.

Choin, J., J. Mendoza-Revilla, L. R. Arauna, S. Cuadros-Espinoza, O. Cassar, M. Larena, A. M. S. Ko, C. Harmant, R. Laurent, P. Verdu, G. Laval, A. Boland, R. Olaso, J. F. Deleuze, F. Valentin, Y. C. Ko, M. Jakobsson, A. Gessain, L. Excoffier, M. Stoneking, E. Patin, and L. Quintana-Murci. 2021. 'Genomic insights into population history and biological adaptation in Oceania', *Nature*, 592: 583-+.

Clarkson, C., Z. Jacobs, B. Marwick, R. Fullagar, L. Wallis, M. Smith, R. G. Roberts, E. Hayes, K. Lowe, X. Carah, S. A. Florin, J. McNeil, D. Cox, L. J. Arnold, Q. Hua, J. Huntley, H. E. A. Brand, T. Manne, A. Fairbairn, J. Shulmeister, L. Lyle, M. Salinas, M. Page, K. Connell, G. Park, K. Norman, T. Murphy, and C. Pardoe. 2017. 'Human occupation of northern Australia by 65,000 years ago', *Nature*, 547: 306-+.

Colley, R. , and P. Ash. 1971. *The geology of Erromango* (New Hebrides Geol. Surv. Reg. Rept).

Cooke, N. P., and S. Nakagome. 2018. 'Fine-tuning of Approximate Bayesian Computation for human population genomics', *Curr Opin Genet Dev*, 53: 60-69.

Cooper, G. M., E. A. Stone, G. Asimenos, Nisc Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow. 2005. 'Distribution and intensity of constraint in mammalian genomic sequence', *Genome Res*, 15: 901-13.

Copeland, J. 1866. *Lecture on the New Hebrides Islands, the New Hebrides natives, and the New Hebrides mission* (Mills, Dick & Co).

Dannemann, M., and F. Racimo. 2018. 'Something old, something borrowed: admixture and adaptation in human evolution', *Current Opinion in Genetics & Development*, 53: 1-8.

de Manuel, M., M. Kuhlwilm, P. Frandsen, V. C. Sousa, T. Desai, J. Prado-Martinez, J. Hernandez-Rodriguez, I. Dupanloup, O. Lao, P. Hallast, J. M. Schmidt, J. M. Heredia-Genestar, A. Benazzo, G. Barbujani, B. M. Peter, L. F. K. Kuderna, F. Casals, S. Angedakin, M. Arandjelovic, C. Boesch, H. Kuhl, L. Vigilant, K. Langergraber, J. Novembre, M. Gut, I. Gut, A. Navarro, F. Carlsen, A. M. Andres, H. R. Siegismund, A. Scally, L. Excoffier, C. Tyler-Smith, S. Castellano, Y. L. Xue, C. Hvilsom, and T. Marques-Bonet. 2016. 'Chimpanzee genomic diversity reveals ancient admixture with bonobos', *Science*, 354: 477-81.

Delfin, F., S. Myles, Y. Choi, D. Hughes, R. Illek, M. van Oven, B. Pakendorf, M. Kayser, and M. Stoneking. 2012. 'Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands', *Mol Biol Evol*, 29: 545-64.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43: 491-+.

Diamond, J. 2003. 'The double puzzle of diabetes', *Nature*, 423: 599-602.

Diamond, J., and P. Bellwood. 2003. 'Farmers and their languages: The first expansions', *Science*, 300: 597-603.

Diamond, J. M. 1988. 'Archaeology - Express Train to Polynesia', *Nature*, 336: 307-08.

Do, R., D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. 2015. 'No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans', *Nature Genetics*, 47: 126-31.

Docker, E. W. 1970. *The blackbirders: The recruiting of South Seas labour for Queensland, 1863–1907* (Angus & Robertson).

Duggan, A. T., B. Evans, F. R. Friedlaender, J. S. Friedlaender, G. Koki, D. A. Merriwether, M. Kayser, and M. Stoneking. 2014. 'Maternal History of Oceania from Complete mtDNA Genomes: Contrasting Ancient Diversity with Recent Homogenization Due to the Austronesian Expansion', *American Journal of Human Genetics*, 94: 721-33.

Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. J. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. H. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. L. S. Xi, K. Y. Yip, J. L. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. P. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. L. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. H. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roeder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. L. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. A. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. E. Wang, J. Wrobel, Y. B. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. J. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. J. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Y. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee,

A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. C. Zhang, T. Y. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. L. Li, K. S. Sandhu, M. Z. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. A. Ruan, Y. J. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. C. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. Mccue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. S. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. J. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. K. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. D. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Q. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. M. J. Mu, H. O'Geen, Z. Q. Ouyang, D. Patacsil, B. K. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Y. Shi, C. Sisu, T. Slifer, H. Witt, L. F. Wu, X. Q. Xu, K. K. Yan, X. Q. Yang, K. Y. Yip, Z. D. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. J. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Q. Yang, M. Snyder, K. P. White, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. P. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Z. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Q. Yan, M. H. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S.

Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Y. Huang, Q. H. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. P. Weng, S. Iyer, X. J. Dong, M. Greven, X. Y. Lin, J. Wang, H. L. S. Xi, J. L. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. M. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, E. Birney, and ENCODE Project Consortium. 2012. 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489: 57-74.

Eickhoff, S., and P. Beighton. 1985. 'Genetic disorders on the island of St Helena', *S Afr Med J*, 68: 475-8.

Ewing, G. B., and J. D. Jensen. 2016. 'The consequences of not accounting for background selection in demographic inference', *Molecular Ecology*, 25: 135-41.

Excoffier, L., I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa, and M. Foll. 2013. 'Robust demographic inference from genomic and SNP data', *PLoS Genet*, 9: e1003905.

Excoffier, L., N. Marchi, D. A. Marques, R. Matthey-Doret, A. Gouy, and V. C. Sousa. 2021. 'fastsimcoal2: demographic inference under complex evolutionary scenarios', *Bioinformatics*.

Fagundes, N. J., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier. 2007. 'Statistical evaluation of alternative models of human evolution', *Proc Natl Acad Sci U S A*, 104: 17614-9.

Fan, S., M. E. Hansen, Y. Lo, and S. A. Tishkoff. 2016. 'Going global by adapting local: A review of recent human adaptation', *Science*, 354: 54-59.

Fisk, E. K. 1966. *New Guinea on the threshold : aspects of social, political, and economic development* (Australian National University Press).

Flexner, J. 2013. 'Mission archaeology in Vanuatu: Preliminary findings, problems, and prospects', *Australasian Historical Archaeology*, 31: 14-24.

———. 2016. *An Archaeology of Early Christianity in Vanuatu: Kastom and Religious Change on Tanna and Erromango, 1839–1920* (ANU Press).

Flexner, J. L., and M. Spriggs. 2015. 'Mission sites as indigenous heritage in southern Vanuatu', *Journal of Social Archaeology*, 15: 184-209.

Font-Porterias, N., R. Caro-Consuegra, M. Lucas-Sanchez, M. Lopez, A. Gimenez, A. Carballo-Mesa, E. Bosch, F. Calafell, L. Quintana-Murci, and D. Comas. 2021. 'The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm', *Mol Biol Evol*, 38: 2804-17.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd, G. K. Chambers, R. A. Lea, J. H. Loo, G. Koki, J. A. Hodgson, D. A. Merriwether, and J. L. Weber. 2008. 'The genetic structure of Pacific islanders', *Plos Genetics*, 4.

Fu, W. Q., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, J. M. Akey, and NHLBI Exome Sequencing Project. 2013. 'Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants (vol 493, pg 216, 2013)', *Nature*, 495: 270-70.

Garcia-Dorado, A. 2008. 'A Simple Method to Account for Natural Selection When Predicting Inbreeding Depression', *Genetics*, 180: 1559-66.

Genomes Project, Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. 'A global reference for human genetic variation', *Nature*, 526: 68-74.

Golson, Jack, Tim Denham, Philip Hughes, John D. Muke, Pamela Swadling, and Australian National University Press. 2017. *Ten thousand years of cultivation at Kuk Swamp in the highlands of Papua New Guinea* (ANU Press: Acton, A.C.T.).

Gordon, J. (editor). 1863. *The Last Martyrs of Eromanga, being a Memoir of the Rev. George N. Gordon, and Ellen Catherine Powell, his Wife* (MacNab and Shafer).

Gosden, C. 1995. 'Arboriculture and agriculture in coastal Papua New Guinea', *Antiquity*, 69: 807-17.

Gosling, A. L., H. R. Buckley, E. Matisoo-Smith, and T. R. Merriman. 2015. 'Pacific Populations, Metabolic Disease and 'Just-So Stories': A Critique of the 'Thrifty Genotype' Hypothesis in Oceania', *Annals of Human Genetics*, 79: 470-80.

Gosling, A. L., and E. A. Matisoo-Smith. 2018a. 'The evolutionary history and human settlement of Australia and the Pacific', *Current Opinion in Genetics & Development*, 53: 53-59.

———. 2018b. 'The evolutionary history and human settlement of Australia and the Pacific', *Current Opinion in Genetics & Development*, 53: 53-59.

Gosling, A. L., E. Matisoo-Smith, and T. R. Merriman. 2014. 'Hyperuricaemia in the Pacific: why the elevated serum urate levels?', *Rheumatol Int*, 34: 743-57.

Gower, G., P. I. Picazo, M. Fumagalli, and F. Racimo. 2021. 'Detecting adaptive introgression in human evolution using convolutional neural networks', *Elife*, 10.

Gravel, S. 2012. 'Population genetics models of local ancestry', *Genetics*, 191: 607-19.

Gray, R. D., A. J. Drummond, and S. J. Greenhill. 2009. 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323: 479-83.

Green, R. C. 2003. *in Pacific Archaeology: Assessments and Anniversary of the First Lapita Excavation (July 1952)* (Le Cahiers de l'Archeologie en Nouvelle-Caledonie, Noumea, New Caledonia).

Gutaker, R. M., S. C. Groen, E. S. Bellis, J. Y. Choi, I. S. Pires, R. K. Bocinsky, E. R. Slayton, O. Wilkins, C. C. Castillo, S. Negrao, M. M. Oliveira, D. Q. Fuller, J. A. D. Guedes, J. R. Lasky, and M. D. Purugganan. 2020. 'Genomic history and ecology of the geographic spread of rice', *Nature Plants*, 6: 492-502.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. 'Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data', *PLoS Genet*, 5: e1000695.

Hage, P., and J. Marck. 2003. 'Matrilineality and the melanesian origin of Polynesian Y chromosomes', *Current Anthropology*, 44: S121-S27.

Hamblin, M. T., and A. Di Rienzo. 2000. 'Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus', *American Journal of Human Genetics*, 66: 1669-79.

Hamid, I., K. L. Korunes, S. Beleza, and A. Goldberg. 2021. 'Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde', *Elife*, 10.

Harris, D. N., M. D. Kessler, A. C. Shetty, D. E. Weeks, R. L. Minster, S. Browning, E. E. Cochrane, R. Deka, N. L. Hawley, M. S. Reupena, T. Naseri, S. T. McGarvey, T. D. O'Connor, Trans-Omics Precision Med, and TOPMed Population Genetics. 2020. 'Evolutionary history of modern Samoans', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 9458-65.

Hartl, Daniel L., and Andrew G. Clark. 2007. *Principles of population genetics* (Sinauer Associates: Sunderland, Mass.).

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. 2014. 'A genetic atlas of human admixture history', *Science*, 343: 747-51.

Henn, B. M., L. R. Botigue, C. D. Bustamante, A. G. Clark, and S. Gravel. 2015a. 'Estimating the mutation load in human genomes', *Nat Rev Genet*, 16: 333-43.

———. 2015b. 'Estimating the mutation load in human genomes', *Nature Reviews Genetics*, 16: 333-43.

Henn, B. M., L. R. Botigue, S. Peischl, I. Dupanloup, M. Lipatov, B. K. Maples, A. R. Martin, S. Musharoff, H. Cann, M. P. Snyder, L. Excoffier, J. M. Kidd, and C. D. Bustamante. 2016a. 'Distance from sub-Saharan Africa predicts mutational load in diverse human genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 113: E440-E49.

———. 2016b. 'Distance from sub-Saharan Africa predicts mutational load in diverse human genomes', *Proc Natl Acad Sci U S A*, 113: E440-9.

Huber, C. D., B. Y. Kim, and K. E. Lohmueller. 2020. 'Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution', *Plos Genetics*, 16: e1008827.

Huber, C. D., B. Y. Kim, C. D. Marsden, and K. E. Lohmueller. 2017. 'Determining the factors driving selective effects of new nonsynonymous mutations', *Proc Natl Acad Sci U S A*, 114: 4465-70.

Hudjashov, G., T. Kivisild, P. A. Underhill, P. Endicott, J. J. Sanchez, A. A. Lin, P. Shen, P. Oefner, C. Renfrew, R. Villems, and P. Forster. 2007. 'Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis', *Proc Natl Acad Sci U S A*, 104: 8726-30.

Hughes, R. G., and M. A. Lawrence. 2005. 'Globalisation, food and health in Pacific Island countries', *Asia Pacific Journal of Clinical Nutrition*, 14: 298-306.

Huoponen, K., T. G. Schurr, Y. Chen, and D. C. Wallace. 2001. 'Mitochondrial DNA variation in an aboriginal Australian population: evidence for genetic isolation and regional differentiation', *Hum Immunol*, 62: 954-69.

International HapMap, Consortium. 2003. 'The International HapMap Project', *Nature*, 426: 789-96.

Isshiki, M., I. Naka, Y. Watanabe, N. Nishida, R. Kimura, T. Furusawa, K. Natsuhara, T. Yamauchi, M. Nakazawa, T. Ishida, R. Eddie, R. Ohtsuka, and J. Ohashi. 2020. 'Admixture and natural selection shaped genomes of an Austronesian-speaking population in the Solomon Islands', *Sci Rep*, 10: 6872.

Jeong, C., G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. B. Witonsky, J. K. Pritchard, C. M. Beall, and A. Di Rienzo. 2014. 'Admixture facilitates genetic adaptations to high altitude in Tibet', *Nat Commun*, 5: 3281.

Jordan, F. M., R. D. Gray, S. J. Greenhill, and R. Mace. 2009. 'Matrilocal residence is ancestral in Austronesian societies', *Proceedings of the Royal Society B-Biological Sciences*, 276: 1957-64.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E.

V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Consortium Genome Aggregation Database, B. M. Neale, M. J. Daly, and D. G. MacArthur. 2020. 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581: 434-43.

Karlsson, E. K., D. P. Kwiatkowski, and P. C. Sabeti. 2014. 'Natural selection and infectious disease in human populations', *Nature Reviews Genetics*, 15: 379-93.

Kayser, M. 2010. 'The human genetic history of Oceania: near and remote views of dispersal', *Curr Biol*, 20: R194-201.

Kayser, M., S. Brauer, G. Weiss, P. A. Underhill, L. Roewer, W. Schiefenhovel, and M. Stoneking. 2000. 'Melanesian origin of Polynesian Y chromosomes', *Curr Biol*, 10: 1237-46.

Kealy, S., J. Louys, and S. O'Connor. 2017. 'Reconstructing Palaeogeography and Inter-island Visibility in the Wallacean Archipelago During the Likely Period of Sahul Colonization, 65-45000Years Ago', *Archaeological Prospection*, 24: 259-72.

———. 2018. 'Least-cost pathway models indicate northern human dispersal from Sunda to Sahul', *Journal of Human Evolution*, 125: 59-70.

Kessler, M. D., N. W. Bateman, T. P. Conrads, G. L. Maxwell, J. C. Dunning Hotopp, and T. D. O'Connor. 2019. 'Ancestral characterization of 1018 cancer cell lines highlights disparities and reveals gene expression and mutational differences', *Cancer*, 125: 2076-88.

Kessler, M. D., L. Yerges-Armstrong, M. A. Taub, A. C. Shetty, K. Maloney, L. J. B. Jeng, I. Ruczinski, A. M. Levin, L. K. Williams, T. H. Beaty, R. A. Mathias, K. C. Barnes, Americas Consortium on Asthma among African-ancestry Populations in the, and T. D. O'Connor. 2016. 'Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry', *Nat Commun*, 7: 12521.

Kim, B. Y., C. D. Huber, and K. E. Lohmueller. 2017. 'Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples', *Genetics*, 206: 345-61.

Kimura, M. 1991. 'The neutral theory of molecular evolution: a review of recent evidence', *Jpn J Genet*, 66: 367-86.

Kinaston, R., S. Bedford, M. Richards, S. Hawkins, A. Gray, K. Jaouen, F. Valentin, and H. Buckley. 2014. 'Diet and Human Mobility from the Lapita to the Early Historic Period on Uripiv Island, Northeast Malakula, Vanuatu', *PLoS One*, 9.

Kinaston, R., H. Buckley, F. Valentin, S. Bedford, M. Spriggs, S. Hawkins, and E. Herrscher. 2014. 'Lapita Diet in Remote Oceania: New Stable Isotope Evidence from the 3000-Year-Old Teouma Site, Efate Island, Vanuatu', *PLoS One*, 9.

Kingman, J. F. 2000. 'Origins of the coalescent. 1974-1982', *Genetics*, 156: 1461-3.

Kirch, P. V. 2017. *On the road of the winds: An archeological history of the Pacific islands before European contact* (University of California Press).

Klamer, M. 2019. 'The dispersal of Austronesian languages in Island South East Asia: Current findings and debates', *Language and Linguistics Compass*, 13.

Knudson, A. G. 1979. 'Our Load of Mutations and Its Burden of Disease', *American Journal of Human Genetics*, 31: 401-13.

Krishnan, M., T. J. Major, R. K. Topless, O. Dewes, L. Yu, J. M. D. Thompson, L. McCowan, J. de Zoysa, L. K. Stamp, N. Dalbeth, J. Harre Hindmarsh, N. Rapana, R. Deka, W. W. H.

Eng, D. E. Weeks, R. L. Minster, S. T. McGarvey, S. Viali, T. Naseri, M. Sefuiva Reupena, P. Wilcox, D. Grattan, P. R. Shepherd, A. N. Shelling, R. Murphy, and T. R. Merriman. 2018. 'Discordant association of the CREBRF rs373863828 A allele with increased BMI and protection from type 2 diabetes in Maori and Pacific (Polynesian) people living in Aotearoa/New Zealand', *Diabetologia*, 61: 1603-13.

Lachance, J., and S. A. Tishkoff. 2013. 'SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it', *Bioessays*, 35: 780-86.

Landrum, M. J., J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott. 2018. 'ClinVar: improving access to variant interpretations and supporting evidence', *Nucleic Acids Res*, 46: D1062-D67.

Landry, L. G., N. Ali, D. R. Williams, H. L. Rehm, and V. L. Bonham. 2018. 'Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice', *Health Affairs*, 37: 780-85.

Larena, M., F. Sanchez-Quinto, P. Sjodin, J. McKenna, C. Ebeo, R. Reyes, O. Casel, J. Y. Huang, K. P. Hagada, D. Guilay, J. Reyes, F. P. Allian, V. Mori, L. S. Azarcon, A. Manera, C. Terando, L. Jamero, G. Sireg, R. Manginsay-Tremedal, M. S. Labos, R. D. Vilar, A. Latiph, R. L. Saway, E. Marte, P. Magbanua, A. Morales, I. Java, R. Reveche, B. Barrios, E. Burton, J. C. Salon, M. J. T. Kels, A. Albano, R. B. Cruz-Angeles, E. Molanida, L. Granehall, M. Vicente, H. Edlund, J. H. Loo, J. Trejaut, S. Y. W. Ho, L. Reid, H. Malmstrom, C. Schlebusch, K. Lambeck, P. Endicott, and M. Jakobsson. 2021. 'Multiple migrations to the Philippines during the last 50,000 years', *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Larson, G., T. Cucchi, M. Fujita, E. Matisoo-Smith, J. Robins, A. Anderson, B. Rolett, M. Spriggs, G. Dolman, T. H. Kim, N. T. D. Thuy, E. Randi, M. Doherty, R. A. Due, R. Bollt, T. Djubiantono, B. Griffin, M. Intoh, E. Keane, P. Kirch, K. T. Li, M. Morwood, L. M. Pedrina, P. J. Piper, R. J. Rabett, P. Shooter, G. Van den Bergh, E. West, S. Wickler, J. Yuan, A. Cooper, and K. Dobney. 2007. 'Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in island southeast Asia and Oceania', *Proceedings of the National Academy of Sciences of the United States of America*, 104: 4834-39.

Leavesley, M. G. 2005. 'Prehistoric Hunting Strategies in New Ireland, Papua New Guinea: The Evidence of the Cuscus (Phalanger orientalis) Remains from Buang Merabak Cave', *Asian Perspectives*, 44: 207-18.

Leavesley, M. G., and J. Chappell. 2004. 'Buang Merabak: additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea', *Antiquity*, 78.

Leavesley, Matthew. 2006. 'Late Pleistocene complexities in the Bismarck Archipelago.' in Ian Lilley (ed.), *Archaeology of Oceania: Australia and the Pacific Islands* (Blackwell Publishing).

Lee, P. H., C. O'Dushlaine, B. Thomas, and S. M. Purcell. 2012. 'INRICH: interval-based enrichment analysis for genome-wide association studies', *Bioinformatics*, 28: 1797-99.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

———. 2011. 'Inference of human population history from individual whole-genome sequences', *Nature*, 475: 493-6.

Liang, M., and R. Nielsen. 2014. 'The lengths of admixture tracts', *Genetics*, 197: 953-67.

Lipson, M., P. R. Loh, S. Sankararaman, N. Patterson, B. Berger, and D. Reich. 2015. 'Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes', *PLoS Genet*, 11: e1005550.

Lipson, M., P. Skoglund, M. Spriggs, F. Valentin, S. Bedford, R. Shing, H. Buckley, I. Phillip, G. K. Ward, S. Mallick, N. Rohland, N. Broomandkhoshbacht, O. Cheronet, M. Ferry, T. K. Harper, M. Michel, J. Oppenheimer, K. Sirak, K. Stewardson, K. Auckland, A. V. S. Hill, K. Maitland, S. J. Oppenheimer, T. Parks, K. Robson, T. N. Williams, D. J. Kennett, A. J. Mentzer, R. Pinhasi, and D. Reich. 2018. 'Population Turnover in Remote Oceania Shortly after Initial Settlement', *Current Biology*, 28: 1157-+.

Lipson, M., M. Spriggs, F. Valentin, S. Bedford, R. Shing, W. Zinger, H. Buckley, F. Petchey, R. Matanik, O. Cheronet, N. Rohland, R. Pinhasi, and D. Reich. 2020. 'Three Phases of Ancient Migration Shaped the Ancestry of Human Populations in Vanuatu', *Current Biology*, 30: 4846-+.

Lohmueller, K. E. 2014. 'The distribution of deleterious genetic variation in human populations', *Current Opinion in Genetics & Development*, 29: 139-46.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. 2008. 'Proportionally more deleterious genetic variation in European than in African populations', *Nature*, 451: 994-U5.

Loos, R. J. 2016. 'CREBRF variant increases obesity risk and protects against diabetes in Samoans', *Nat Genet*, 48: 976-8.

Lopez, M., A. Kousathanas, H. Quach, C. Harmant, P. Mouguiama-Daouda, J. M. Hombert, A. Froment, G. H. Perry, L. B. Barreiro, P. Verdu, E. Patin, and L. Quintana-Murci. 2018a. 'The demographic history and mutational load of African hunter-gatherers and farmers', *Nature Ecology & Evolution*, 2: 721-30.

———. 2018b. 'The demographic history and mutational load of African hunter-gatherers and farmers', *Nature Ecology & Evolution*, 2: 721-30.

Luque, M., and C. Mondragon. 2005. 'Faith, fidelity and fantasy - Don Pedro Fernandez de Quiros and the 'foundation, government and sustenance' of La Nueba Hierusalem in 1606', *Journal of Pacific History*, 40: 133-48.

Majara, L., A. Kalungi, N. Koen, H. Zar, D. J. Stein, E. Kinyanda, E. G. Atkinson, and A. R. Martin. 2021. 'Low generalizability of polygenic scores in African populations due to genetic and environmental diversity', *bioRxiv*.

Malaspinas, A. S., M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergstrom, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvag, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F. X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev. 2016a. 'A genomic history of Aboriginal Australia', *Nature*, 538: 207-14.

Malaspinas, A. S., M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergstrom, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. L. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvag, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F. X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev. 2016b. 'A genomic history of Aboriginal Australia', *Nature*, 538: 207-+.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen. 2010. 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26: 2867-73.

Marchi, N., F. Schlichta, and L. Excoffi. 2021. 'Demographic inference', *Current Biology*, 31: R276-R79.

Matisoo-Smith, E. 1994. 'The Human Colonization of Polynesia - a Novel-Approach - Genetic Analyses of the Polynesian Rat (Rattus-Exulans)', *Journal of the Polynesian Society*, 103: 75-87.

———. 2015. 'Ancient DNA and the human settlement of the Pacific: A review', *Journal of Human Evolution*, 79: 93-104.

Matisoo-Smith, E., J. S. Allen, R. M. Roberts, G. J. Irwin, and D. M. Lambert. 1999. 'Rodents of the sunrise: Mitochondrial DNA phylogenies of Polynesian Rattus exulans and the settlement of Polynesia', *Pacific from 5000 to 2000 Bp*: 259-+.

Matisoo-Smith, E., and J. H. Robins. 2004. 'Origins and dispersals of Pacific peoples: evidence from mtDNA phylogenies of the Pacific rat', *Proc Natl Acad Sci U S A*, 101: 9167-72.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. 'The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20: 1297-303.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17.

Melton, T., R. Peterson, A. J. Redd, N. Saha, A. S. Sofro, J. Martinson, and M. Stoneking. 1995. 'Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis', *Am J Hum Genet*, 57: 403-14.

Minster, R. L., N. L. Hawley, C. T. Su, G. Sun, E. E. Kershaw, H. Cheng, O. D. Buhule, J. Lin, M. S. Reupena, S. Viali, J. Tuitele, T. Naseri, Z. Urban, R. Deka, D. E. Weeks, and S. T. McGarvey. 2016. 'A thrifty variant in CREBRF strongly influences body mass index in Samoans', *Nat Genet*, 48: 1049-54.

Mona, S., M. Tommaseo-Ponzetta, S. Brauer, H. Sudoyo, S. Marzuki, and M. Kayser. 2007. 'Patterns of Y-chromosome diversity intersect with the Trans-New Guinea hypothesis', *Mol Biol Evol*, 24: 2546-55.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. 'A fine-scale map of recombination rates and hotspots across the human genome', *Science*, 310: 321-4.

Myers, S., C. Fefferman, and N. Patterson. 2008. 'Can one learn history from the allelic spectrum?', *Theoretical Population Biology*, 73: 342-48.

Narasimhan, V. M., K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. E. R. Barnes, A. H. Barnett, C. Bates, S. Bellary, N. A. Bockett, K. Giorda, C. J. Griffiths, H. Hemingway, Z. L. Jia, M. A. Kelly, H. A. Khawaja, M. Lek, S. McCarthy, R. McEachan, A. O'Donnell-Luria, K. Paigen, C. A. Parisinos, E. Sheridan, L. Southgate, L. Tee, M. Thomas, Y. L. Xue, M. Schnall-Levin, P. M. Petkov, C. T. -Smith, E. R. Maher, R. C. Trembath, D. G. MacArthur, J. Wright, R. Durbin, and D. A. Heel. 2016. 'Health and population effects of rare gene knockouts in adult humans with related parents', *Science*, 352: 474-77.

Neel, J. V. 1962. 'Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?', *Am J Hum Genet*, 14: 353-62.

Nielsen, R., and M. A. Beaumont. 2009. 'Statistical inferences in phylogeography', *Mol Ecol*, 18: 1034-47.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. 'Genomic scans for selective sweeps using SNP data', *Genome Res*, 15: 1566-75.

Noury, A. 2005. *Le reflet de l'ame lapita: Essai d'Interprétation des Décors des Poteries Lapita en Melanesie et Polynesie Occidentale (3200-2700 BP)* (Arnaud Noury).

Noury, A., and J-C. Galipaud. 2011. *Les Lapita, nomades du Pacifique*.

O'Brien, E., L. B. Jorde, B. Ronnlof, J. O. Fellman, and A. W. Eriksson. 1988. 'Founder effect and genetic disease in Sottunga, Finland', *Am J Phys Anthropol*, 77: 335-46.

O'Connell, J. F., and J. Allen. 2015. 'The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago', *Journal of Archaeological Science*, 56: 73-84.

O'Connell, J. F., J. Allen, M. A. J. Williams, A. N. Williams, C. S. M. Turney, N. A. Spooner, J. Kamminga, G. Brown, and A. Cooper. 2018a. 'When did Homo sapiens first reach Southeast Asia and Sahul?', *Proc Natl Acad Sci U S A*, 115: 8482-90.

———. 2018b. 'When did Homo sapiens first reach Southeast Asia and Sahul?', *Proceedings of the National Academy of Sciences of the United States of America*, 115: 8482-90.

Olivares, G., B. Pena-Ahumada, J. Penailillo, C. Payacan, X. Moncada, M. Saldarriaga-Cordoba, E. Matisoo-Smith, K. F. Chung, D. Seelenfreund, and A. Seelenfreund. 2019. 'Human mediated translocation of Pacific paper mulberry [Broussonetia papyrifera (L.) L'Her. ex Vent. (Moraceae)]: Genetic evidence of dispersal routes in Remote Oceania', *PLoS One*, 14.

Oppenheimer, S. J., and M. Richards. 2001. 'Polynesian origins. Slow boat to Melanesia?', *Nature*, 410: 166-7.

Patin, E., M. Lopez, R. Grollemund, P. Verdu, C. Harmant, H. Quach, G. Laval, G. H. Perry, L. B. Barreiro, A. Froment, E. Heyer, A. Massougbodji, C. Fortes-Lima, F. Migot-Nabias, G. Bellis, J. M. Dugoujon, J. B. Pereira, V. Fernandes, L. Pereira, L. Van der Veen, P. Mouguiama-Daouda, C. D. Bustamante, J. M. Hombert, and L. Quintana-Murci. 2017. 'Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America', *Science*, 356: 543-46.

Patin, E., and L. Quintana-Murci. 2018. 'The demographic and adaptive history of central African hunter-gatherers and farmers', *Current Opinion in Genetics & Development*, 53: 90-97.

Paul, D. B. 1987. 'Our Load of Mutations Revisited', *Journal of the History of Biology*, 20: 321-35.

Pedersen, C. E. T., K. E. Lohmueller, N. Grarup, P. Bjerregaard, T. Hansen, H. R. Siegismund, I. Moltke, and A. Albrechtsen. 2017a. 'The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit', *Genetics*, 205: 787-801.

Pedersen, C. T., K. E. Lohmueller, N. Grarup, P. Bjerregaard, T. Hansen, H. R. Siegismund, I. Moltke, and A. Albrechtsen. 2017b. 'The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit', *Genetics*, 205: 787-801.

Pedro, N., N. Brucato, V. Fernandes, M. Andre, L. Saag, W. Pomat, C. Besse, A. Boland, J. F. Deleuze, C. Clarkson, H. Sudoyo, M. Metspalu, M. Stoneking, M. P. Cox, M. Leavesley, L. Pereira, and F. X. Ricaut. 2020. 'Papuan mitochondrial genomes and the settlement of Sahul', *J Hum Genet*, 65: 875-87.

Peter, B. M., E. Huerta-Sanchez, and R. Nielsen. 2012. 'Distinguishing between selective sweeps from standing variation and from a de novo mutation', *PLoS Genet*, 8: e1003011.

Peterson, R. E., K. Kuchenbaecker, R. K. Walters, C. Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, C. E. Carey, A. R. Martin, J. L. Meyers, J. Su, J. Chen, A. C. Edwards, A. Kalungi, N. Koen, L. Majara, E. Schwarz, J. W. Smoller, E. A. Stahl, P. F. Sullivan, E. Vassos, B. Mowry, M. L. Prieto, A. Cuellar-Barboza, T. B. Bigdeli, H. J. Edenberg, H. Huang, and L. E. Duncan. 2019. 'Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations', *Cell*, 179: 589-603.

Phadnis, N., and J. D. Fry. 2005. 'Widespread correlations between dominance and homozygous effects of mutations: Implications for theories of dominance', *Genetics*, 171: 385-92.

Posth, C., K. Nagele, H. Colleran, F. Valentin, S. Bedford, K. W. Kami, R. Shing, H. Buckley, R. Kinaston, M. Walworth, G. R. Clark, C. Reepmeyer, J. Flexner, T. Maric, J. Moser, J. Gresky, L. Kiko, K. J. Robson, K. Auckland, S. J. Oppenheimer, A. V. S. Hill, A. J. Mentzer, J. Zech, F. Petchey, P. Roberts, C. Jeong, R. D. Gray, J. Krause, and A. Powell. 2018. 'Language continuity despite population replacement in Remote Oceania', *Nature Ecology & Evolution*, 2: 731-40.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. 'The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation', *Curr Biol*, 20: R208-15.

Przeworski, M., G. Coop, and J. D. Wall. 2005. 'The signature of positive selection on standing genetic variation', *Evolution*, 59: 2312-23.

Pugach, I., A. T. Duggan, D. A. Merriwether, F. R. Friedlaender, J. S. Friedlaender, and M. Stoneking. 2018a. 'The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data', *Mol Biol Evol*, 35: 871-86.

———. 2018b. 'The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data', *Molecular Biology and Evolution*, 35: 871-86.

Pugach, I., A. Hubner, H. C. Hung, M. Meyer, M. T. Carson, and M. Stoneking. 2021. 'Ancient DNA from Guam and the peopling of the Pacific', *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Pugach, I., R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking. 2011. 'Dating the age of admixture via wavelet transform analysis of genome-wide data', *Genome Biology*, 12.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*, 81: 559-75.

Qin, P., and M. Stoneking. 2015. 'Denisovan Ancestry in East Eurasian and Native American Populations', *Mol Biol Evol*, 32: 2665-74.

Quintana-Murci, L. 2016. 'Understanding rare and common diseases in the context of human evolution', *Genome Biology*, 17.

———. 2019. 'Human Immunology through the Lens of Evolutionary Genetics', *Cell*, 177: 184-99.

Quintana-Murci, L., and A. G. Clark. 2013. 'Population genetic tools for dissecting innate immunity in humans', *Nat Rev Immunol*, 13: 280-93.

Racimo, F., S. Sankararaman, R. Nielsen, and E. Huerta-Sanchez. 2015. 'Evidence for archaic adaptive introgression in humans', *Nat Rev Genet*, 16: 359-71.

Rasmussen, M., M. Sikora, A. Albrechtsen, T. S. Korneliussen, J. V. Moreno-Mayar, G. D. Poznik, C. P. E. Zollikofer, M. S. P. de Leon, M. E. Allentoft, I. Moltke, K. Jonsson, C. Valdiosera, R. S. Malhi, L. Orlando, C. D. Bustamante, T. W. Stafford, D. J. Meltzer, R. Nielsen, and E. Willerslev. 2015. 'The ancestry and affiliations of Kennewick Man', *Nature*, 523: 455-U159.

Redd, A. J., and M. Stoneking. 1999. 'Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations', *Am J Hum Genet*, 65: 808-28.

Redd, A. J., N. Takezaki, S. T. Sherry, S. T. Mcgarvey, A. S. M. Sofro, and M. Stoneking. 1995. 'Evolutionary History of the Coii/Trna(Lys) Intergenic 9-Base-Pair Deletion in Human Mitochondrial Dnas from the Pacific', *Molecular Biology and Evolution*, 12: 604-15.

Reich, D., N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M. Ko, Y. C. Ko, T. A. Jinam, M. E. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Paabo, and M. Stoneking. 2011. 'Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania', *Am J Hum Genet*, 89: 516-28.

Richards, M., S. Oppenheimer, and B. Sykes. 1998. 'mtDNA suggests Polynesian origins in Eastern Indonesia', *Am J Hum Genet*, 63: 1234-6.

Rieth, T. M. , and Athens J. S. 2017. 'Late Holocene Human Expansion into Near and Remote Oceania: A Bayesian Model of the Chronologies of the Mariana Islands and Bismarck Archipelago', *The Journal of Island and Coastal Archaeology*, 14: 5-16.

Rosenberg, N. A., and M. Nordborg. 2002. 'Genealogical trees, coalescent theory and the analysis of genetic polymorphisms', *Nat Rev Genet*, 3: 380-90.

Ross, M. 1995. *Reconstructing Proto Austronesian verbal morphology: evidence from Taiwan.* (Taipei: Institute of History and Philology, Academia Sinica).

Rotival, M., P. Cossart, and L. Quintana-Murci. 2021. 'Reconstructing 50,000 years of human history from our DNA: lessons from modern genomics', *C R Biol*, 344: 177-87.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, Consortium International HapMap, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S.

Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. 2007. 'Genome-wide detection and characterization of positive selection in human populations', *Nature*, 449: 913-8.

Sankararaman, S., S. Mallick, N. Patterson, and D. Reich. 2016. 'The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans', *Curr Biol*, 26: 1241-7.

Scheinfeldt, L., F. Friedlaender, J. Friedlaender, K. Latham, G. Koki, T. Karafet, M. Hammer, and J. Lorenz. 2006. 'Unexpected NRY chromosome variation in Northern Island Melanesia', *Mol Biol Evol*, 23: 1628-41.

Schiffels, S., and R. Durbin. 2014. 'Inferring human population size and separation history from multiple genome sequences', *Nat Genet*, 46: 919-25.

Schorr, D. B. 2018. 'Savagery, Civilization, and Property: Theories of Societal Evolution and Commons Theory', *Theoretical Inquiries in Law*, 19: 507-31.

Schrider, D. R., and A. D. Kern. 2017. 'Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome', *Mol Biol Evol*, 34: 1863-77.

Schrider, D. R., A. G. Shanku, and A. D. Kern. 2016. 'Effects of Linked Selective Sweeps on Demographic Inference and Model Selection', *Genetics*, 204: 1207-23.

Segurel, L., M. J. Wyman, and M. Przeworski. 2014. 'Determinants of mutation rate variation in the human germline', *Annu Rev Genomics Hum Genet*, 15: 47-70.

Seplyarskiy, V. B., and S. Sunyaev. 2021. 'The origin of human mutation in light of genomic data', *Nat Rev Genet*.

Shaw, B., J. H. Field, G. R. Summerhayes, S. Coxe, A. C. F. Costers, A. Ford, J. Haro, H. Arifeae, E. Hull, G. Jacobsen, R. Fullagar, E. Hayes, and L. Kealhofer. 2020. 'Emergence of a Neolithic in highland New Guinea by 5000 to 4000 years ago', *Science Advances*, 6.

Sheppard, P. J. 2011. 'Lapita Colonization across the Near/Remote Oceania Boundary', *Current Anthropology*, 52: 799-840.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. 'dbSNP: the NCBI database of genetic variation', *Nucleic Acids Res*, 29: 308-11.

Shriver, M. D., G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar, J. Huang, J. M. Akey, and K. W. Jones. 2004. 'The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs', *Hum Genomics*, 1: 274-86.

Sikora, M., V. V. Pitulko, V. C. Sousa, M. E. Allentoft, L. Vinner, S. Rasmussen, A. Margaryan, P. D. Damgaard, C. de la Fuente, G. Renaud, M. A. Yang, Q. M. Fu, I. Dupanloup, K. Giampoudakis, D. Nogues-Bravo, C. Rahbek, G. Kroonen, M. Peyrot, H. McColl, S. V. Vasilyev, E. Veselovskaya, M. Gerasimova, E. Y. Pavlova, V. G. Chasnyk, P. A. Nikolskiy, A. V. Gromov, V. I. Khartanovich, V. Moiseyev, P. S. Grebenyuk, A. Y. Fedorchenko, A. I. Lebedintsev, S. B. Slobodin, B. A. Malyarchuk, R. Martiniano, M. Meldgaard, L. Arppe, J. U. Palo, T. Sundell, K. Mannermaa, M. Putkonen, V. Alexandersen, C. Primeau, N. Baimukhanov, R. S. Malhi, K. G. Sjogren, K. Kristiansen, A. Wessman, A. Sajantila, M. M. Lahr, R. Durbin, R. Nielsen, D. J. Meltzer, L. Excoffier, and E. Willerslev. 2019. 'The population history of northeastern Siberia since the Pleistocene', *Nature*, 570: 182-+.

Simmons, M. J., and J. F. Crow. 1977. 'Mutations Affecting Fitness in Drosophila Populations', *Annual Review of Genetics*, 11: 49-78.

Simons, Y. B., and G. Sella. 2016. 'The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives', *Current Opinion in Genetics & Development*, 41: 150-58.

Simons, Y. B., M. C. Turchin, J. K. Pritchard, and G. Sella. 2014. 'The deleterious mutation load is insensitive to recent population history', *Nature Genetics*, 46: 220-+.

Sirugo, G., S. M. Williams, and S. A. Tishkoff. 2019. 'The Missing Diversity in Human Genetic Studies', *Cell*, 177: 26-31.

Skoglund, P., C. Posth, K. Sirak, M. Spriggs, F. Valentin, S. Bedford, G. R. Clark, C. Reepmeyer, F. Petchey, D. Fernandes, Q. M. Fu, E. Harney, M. Lipson, S. Mallick, M. Novak, N. Rohland, K. Stewardson, S. Abdullah, M. P. Cox, F. R. Friedlaender, J. S. Friedlaender, T. Kivisild, G. Koki, P. Kusuma, D. A. Merriwether, F. X. Ricaut, J. T. S. Wee, N. Patterson, J. Krause, R. Pinhasi, and D. Reich. 2016. 'Genomic insights into the peopling of the Southwest Pacific', *Nature*, 538: 510-+.

Soares, P., T. Rito, J. Trejaut, M. Mormina, C. Hill, E. Tinkler-Hundal, M. Braid, D. J. Clarke, J. H. Loo, N. Thomson, T. Denham, M. Donohue, V. Macaulay, M. Lin, S. Oppenheimer, and M. B. Richards. 2011. 'Ancient voyaging and Polynesian origins', *Am J Hum Genet*, 88: 239-47.

Specht, J., T. Denham, J. Goff, and J. E. Terrell. 2014. 'Deconstructing the Lapita Cultural Complex in the Bismarck Archipelago', *Journal of Archaeological Research*, 22: 89-140.

Speidel, L., M. Forest, S. N. Shi, and S. R. Myers. 2019. 'A method for genome-wide genealogy estimation for thousands of samples', *Nature Genetics*, 51: 1321-+.

Spriggs, M. 1997. *The Island Melanesians* (Oxford:Blackwells).

Stoneking, M., L. B. Jorde, K. Bhatia, and A. C. Wilson. 1990. 'Geographic variation in human mitochondrial DNA from Papua New Guinea', *Genetics*, 124: 717-33.

Su, B., L. Jin, P. Underhill, J. Martinson, N. Saha, S. T. McGarvey, M. D. Shriver, J. Chu, P. Oefner, R. Chakraborty, and R. Deka. 2000. 'Polynesian origins: insights from the Y chromosome', *Proc Natl Acad Sci U S A*, 97: 8225-8.

Summerhayes, G. R. 2009. 'Obisidian network patterns in Melanesia - sources, characterisation and distribution', *Bulletin of the Info-Pacific prehistory association*, 29: 109-23.

———. 2010. *Lapita Obsidian Sources and Distribution/ Les Sources et la Répartition de l'obsidienne Lapita, Lapita: Ancêtres Océaniens* (Somogy Edition/Musée du Quai Branly).

Summerhayes, G. R., M. Leavesley, A. Fairbairn, H. Mandui, J. Field, A. Ford, and R. Fullagar. 2010. 'Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago', *Science*, 330: 78-81.

Sykes, B., A. Leiboff, J. Low-Beer, S. Tetzner, and M. Richards. 1995. 'The origins of the Polynesians: an interpretation from mitochondrial lineage analysis', *Am J Hum Genet*, 57: 1463-75.

Szpiech, Z. A., A. Blant, and T. J. Pemberton. 2017. 'GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification', *Bioinformatics*, 33: 2059-62.

Tadmouri, G. O., P. Nair, T. Obeid, M. T. Al Ali, N. Al Khaja, and H. A. Hamamy. 2009. 'Consanguinity and reproductive health among Arabs', *Reprod Health*, 6: 17.

Terhorst, J., and Y. S. Song. 2015. 'Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 7677-82.

Terrell, J. E. 2004. 'The 'sleeping giant' hypothesis and New Guinea's place in the prehistory of Greater Near Oceania', *World Archaeology*, 36: 601-09.

Thomson, V. A., O. Lebrasseur, J. J. Austin, T. L. Hunt, D. A. Burney, T. Denham, N. J. Rawlence, J. R. Wood, J. Gongora, L. G. Flink, A. Linderholm, K. Dobney, G. Larson, and A. Cooper. 2014. 'Using ancient DNA to study the origins and dispersal of ancestral Polynesian chickens across the Pacific', *Proceedings of the National Academy of Sciences of the United States of America*, 111: 4826-31.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. 'Convergent adaptation of human lactase persistence in Africa and Europe', *Nat Genet*, 39: 31-40.

Torrence, R., V. Neall, and W. E. Boyd. 2009. 'Volcanism and Historical Ecology on the Willaumez Peninsula, Papua New Guinea', *Pacific Science*, 63: 507-35.

Torrence, R., V. Neall, T. Doelman, E. Rhodes, C. Mckee, H. Davies, R. Bonetti, A. Gugliemetti, A. Manzoni, M. Oddone, J. Parr, and C. Wallace. 2004. 'Pleistocene colonisation of the Bismarck Archipelago: new evidence from West New Britain', *Archaeology in Oceania*, 39: 101-30.

Torrence, R., and P. Swadling. 2008. 'Social networks and the spread of Lapita', *Antiquity*, 82: 600-16.

Valentin, F., F. Detroit, M. J. T. Spriggs, and S. Bedford. 2016. 'Early Lapita skeletons from Vanuatu show Polynesian craniofacial shape: Implications for Remote Oceanic settlement and Lapita origins', *Proceedings of the National Academy of Sciences of the United States of America*, 113: 292-97.

Valentin, F., E. Herrscher, S. Bedford, M. Spriggs, and H. Buckley. 2014. 'Evidence for Social and Cultural Change in Central Vanuatu Between 3000 and 2000 BP: Comparing Funerary and Dietary Patterns of the First and Later Generations at Teouma, Efate', *Journal of Island & Coastal Archaeology*, 9: 381-99.

Veeramah, K. R. 2018. 'The importance of fine-scale studies for integrating paleogenomics and archaeology', *Current Opinion in Genetics & Development*, 53: 83-89.

Veeramah, K. R., D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins, G. Destro-Bisol, H. Soodyall, L. Louie, and M. F. Hammer. 2012. 'An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data', *Mol Biol Evol*, 29: 617-30.

Vernot, B., S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf, R. M. Gittelman, M. Dannemann, S. Grote, R. C. McCoy, H. Norton, L. B. Scheinfeldt, D. A. Merriwether, G. Koki, J. S. Friedlaender, J. Wakefield, S. Paabo, and J. M. Akey. 2016. 'Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals', *Science*, 352: 235-9.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. 'A map of recent positive selection in the human genome', *PLoS Biol*, 4: e72.

Wagner, J. K., C. Colwell, K. G. Claw, A. C. Stone, D. A. Bolnick, J. Hawks, K. B. Brothers, and N. A. Garrison. 2020. 'Fostering Responsible Research on Ancient DNA', *American Journal of Human Genetics*, 107: 183-95.

Walser, J. C., L. Ponger, and A. V. Furano. 2008. 'CpG dinucleotides and the mutation rate of non-CpG DNA', *Genome Res*, 18: 1403-14.

Walter, R., and P. J. Sheppard. 2009. 'A review of Solomon Island archaeology.' in (New Zealand Archaeological Association).

Watson, J. D., and F. H. Crick. 1953. 'Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid', *Nature*, 171: 737-8.

Wickler, S., and M. Spriggs. 1988. 'Pleistocene human occupation of the Solomon Islands, Melanesia', *Antiquity*, 62: 703-06.

Wollstein, A., O. Lao, C. Becker, S. Brauer, R. J. Trent, P. Nurnberg, M. Stoneking, and M. Kayser. 2010. 'Demographic history of Oceania inferred from genome-wide data', *Curr Biol*, 20: 1983-92.

Wright, S. 1929. 'Fisher's theory of dominance', *Am. Nat.*, 63: 274-79.

———. 1931. 'Evolution in Mendelian Populations', *Genetics*, 16: 97-159.

———. 1943. 'Isolation by Distance', *Genetics*, 28: 114-38.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, and J. Wang. 2010. 'Sequencing of 50 human exomes reveals adaptation to high altitude', *Science*, 329: 75-8.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack. 2010. 'Gene ontology analysis for RNA-seq: accounting for selection bias', *Genome Biology*, 11: R14.

Zhang, M., G. P. Sun, L. L. Ren, H. B. Yuan, G. H. Dong, L. Z. Zhang, F. Liu, P. Cao, A. M. S. Ko, M. A. Yang, S. M. Hu, G. D. Wang, and Q. M. Fu. 2020. 'Ancient DNA Evidence from China Reveals the Expansion of Pacific Dogs', *Molecular Biology and Evolution*, 37: 1462-69.

Annexes

Report

Current Biology

Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest

Highlights

- A strong selective sweep at *TRPS1* occurred in African rainforest hunter-gatherers
- Pleiotropic height genes lead to polygenic selection signals for reproductive age
- Pathogen-driven selection, mostly viral, has been pervasive among hunter-gatherers
- Post-admixture selection has maintained adaptive variation in hunter-gatherers

Authors

Marie Lopez, Jeremy Choin, Martin Sikora, ..., Paul Verdu, Etienne Patin, Lluís Quintana-Murci

Correspondence

etienne.patin@pasteur.fr (E.P.), quintana@pasteur.fr (L.Q.-M.)

In Brief

Lopez et al. search for genomic evidence of local adaptation of hunter-gatherers to the African rainforest. They find signals of classic sweeps, polygenic adaptation, and post-admixture selection at height, development, and immune response genes. They show that pleiotropy of height genes leads to polygenic selection signals for life-history traits.

Lopez et al., 2019, Current Biology *29*, 2926–2935
September 9, 2019
https://doi.org/10.1016/j.cub.2019.07.013

CellPress

CellPress

Current Biology
Report

Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest

Marie Lopez,[1,2] Jeremy Choin,[1] Martin Sikora,[3] Katherine Siddle,[1,11] Christine Harmant,[1] Helio A. Costa,[4] Martin Silvert,[1,2] Patrick Mouguiama-Daouda,[5] Jean-Marie Hombert,[6] Alain Froment,[7] Sylvie Le Bomin,[8] George H. Perry,[9] Luis B. Barreiro,[10] Carlos D. Bustamante,[4] Paul Verdu,[8] Etienne Patin,[1,12,*] and Lluís Quintana-Murci[1,12,13,*]

[1]Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris 75015, France
[2]Sorbonne Universités, Ecole Doctorale Complexité du Vivant, 75005 Paris, France
[3]Centre for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark
[4]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA
[5]Laboratoire Langue, Culture et Cognition (LCC), Université Omar Bongo, 13131 Libreville, Gabon
[6]CNRS UMR 5596, Université Lumière-Lyon 2, 69007 Lyon, France
[7]Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, 75005 Paris, France
[8]UMR7206, Muséum National d'Histoire Naturelle, CNRS, Université Paris Diderot, Paris 75016, France
[9]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA
[10]Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
[11]Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[12]Senior author
[13]Lead Contact
*Correspondence: etienne.patin@pasteur.fr (E.P.), quintana@pasteur.fr (L.Q.-M.)
https://doi.org/10.1016/j.cub.2019.07.013

SUMMARY

African rainforests support exceptionally high biodiversity and host the world's largest number of active hunter-gatherers [1–3]. The genetic history of African rainforest hunter-gatherers and neighboring farmers is characterized by an ancient divergence more than 100,000 years ago, together with recent population collapses and expansions, respectively [4–12]. While the demographic past of rainforest hunter-gatherers has been deeply characterized, important aspects of their history of genetic adaptation remain unclear. Here, we investigated how these groups have adapted—through classic selective sweeps, polygenic adaptation, and selection since admixture—to the challenging rainforest environments. To do so, we analyzed a combined dataset of 566 high-coverage exomes, including 266 newly generated exomes, from 14 populations of rainforest hunter-gatherers and farmers, together with 40 newly generated, low-coverage genomes. We find evidence for a strong, shared selective sweep among all hunter-gatherer groups in the regulatory region of *TRPS1*—primarily involved in morphological traits. We detect strong signals of polygenic adaptation for height and life history traits such as reproductive age; however, the latter appear to result from pervasive pleiotropy of height-associated genes. Furthermore, polygenic adaptation signals for functions related to responses of mast cells to allergens and microbes, the IL-2 signaling pathway, and host interactions with viruses support a history of pathogen-driven selection in the rainforest. Finally, we find that genes involved in heart and bone development and immune responses are enriched in both selection signals and local hunter-gatherer ancestry in admixed populations, suggesting that selection has maintained adaptive variation in the face of recent gene flow from farmers.

RESULTS

Exome Sequencing Dataset and Population Structure

African rainforest hunter-gatherers (RHGs)—historically grouped under the term "Pygmies"—live along the dense tropical rainforests of central Africa, in the western and eastern part of the Congo Basin [1–3]. Genetic studies have deeply investigated the demographic history of these groups, characterized by long-term isolation since the Upper Paleolithic and substantial admixture with neighboring Bantu-speaking farmers in the last 1,000 years [4–12]. However, their adaptive history has received less attention. Natural selection studies in RHGs have primarily focused on small adult body size as the only trait characterizing the "pygmy" phenotype [13–20], and used SNP genotyping data [14, 15, 19–21] or whole-genome/exome sequencing of a few individuals or populations [4, 6, 18, 22, 23].

To understand human genetic adaptation to the rainforest, we generated and analyzed whole-exome sequencing data (~40× coverage) for seven RHG groups from Cameroon, Gabon, and Uganda, as well as, for comparison purposes, seven sedentary groups of Bantu-speaking agriculturalists (AGRs) (Figure 1A; Table S1). After quality filters, we obtained a final dataset of 566 individuals (298 RHGs and 268 AGRs), consisting of 266 newly generated exomes that were analyzed with 300 previously reported exomes [4] (Figure S1).

Genetic differentiation among RHG groups was higher than that between RHGs and AGRs (among-RHG, F_{ST} = 0.025; among-western RHG, F_{ST} = 0.021; RHG-AGR, F_{ST} = 0.017;

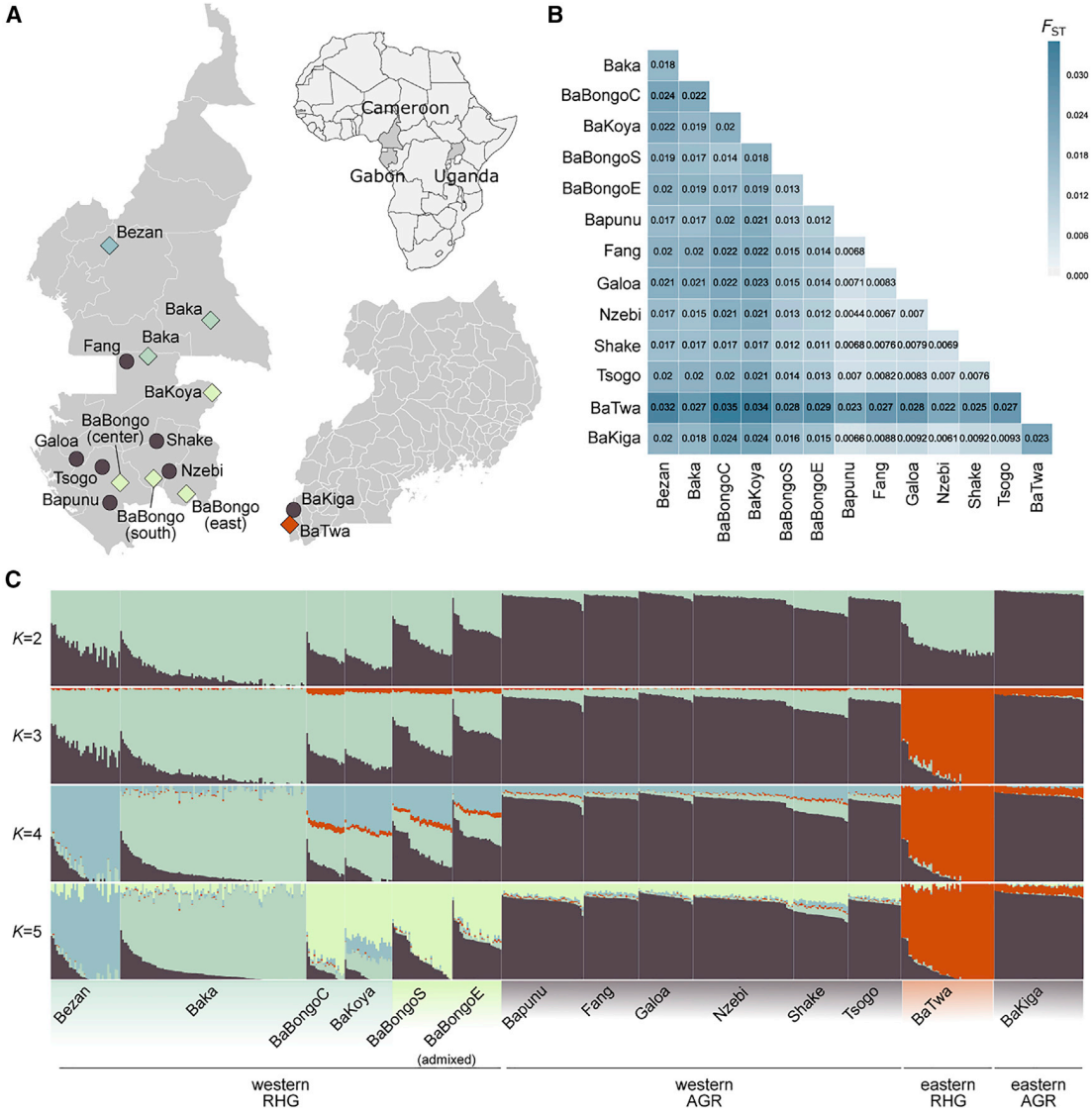


Figure 1. Location, Genetic Differentiation, and Structure of Central African Populations

(A) Geographic location of the populations analyzed. Populations of rainforest hunter-gatherers (diamonds) and neighboring farmers (circles) originating from the three countries are shown in the map of Africa. Colors indicate the dominant membership in each population, based on ADMIXTURE results (C).

(B) Levels of genetic differentiation between populations measured by pairwise F_{ST} calculated on the exome data.

(C) Cluster membership proportions estimated by ADMIXTURE on the merged exome and SNP array data. Cross-validation values were lowest at $K = 5$ clusters.

(B and C) BaBongoC, BaBongoS, and BaBongoE stand for BaBongo populations from the center, south, and east of Gabon, respectively.

See also Figure S1 and Table S1.

among-AGR, $F_{ST} = 0.007$; Figure 1B). To increase SNP density, particularly in the non-coding genome, we combined the exome data with SNP array data for the same individuals [12, 24, 25], yielding a total of 1,253,548 SNPs. When using ADMIXTURE [26] on the dataset pruned for allele frequency (MAF > 5%) and linkage disequilibrium ($r^2 < 0.5$), RHGs separated into four clusters at $K = 5$ (Figure 1C), corresponding to Bezan, Baka, BaBongo and BaKoya, and BaTwa groups. As previously observed [5, 12, 14, 24], membership proportions to the cluster assigned to AGRs were non-negligible and similar among RHG groups (~4%–9%; Table S1), with the exception of the BaBongo of east and south Gabon, who presented high AGR proportions (~43% [SD = 11%] and ~24% [SD = 17%], respectively). Membership proportions to the cluster assigned to RHGs were also non-negligible among AGRs (~10%–30%). Our results show that RHG populations are highly structured, emphasizing the importance of considering these groups separately in subsequent analyses.

Searching for Signals of Local Genetic Adaptation in Central Africans

For all natural selection analyses, we increased SNP density to 9,129,103 high-quality variants (MAF > 1%), through genotype imputation using (1) newly generated whole genomes from

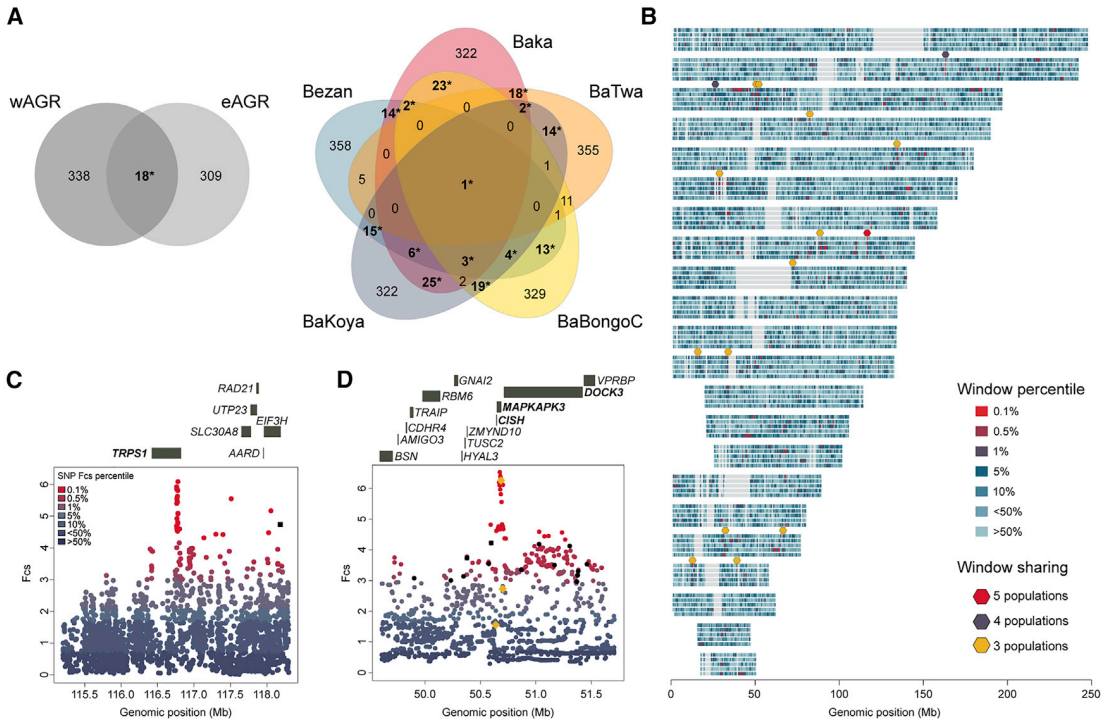


Figure 2. Shared Signals of Classic Sweeps among Rainforest Hunter-Gatherers

(A) Number of candidate windows for classic sweeps (i.e., windows with proportions of outlier SNPs among the 1% highest of the genome) common to western and eastern AGR populations (wAGR and eAGR), as well as common to RHG populations. p values obtained based on 10,000 resamples are shown: *$p < 10^{-4}$.

(B) Genome-wide map of classic sweep signals in RHG groups. The autosomes of each of the five RHG populations (from top to bottom: Bezan, Baka, lowly admixed BaBongo, BaKoya, and BaTwa) are shown. Colored dots indicate genomic regions that are common to at least three RHG populations.

(C) Selective sweep signal at the locus containing the *TRPS1* gene (chr8:116702422-116802422) in the Baka RHGs.

(D) Selective sweep signal at the locus containing *CISH*, *MAPKAPK3*, and *DOCK3* genes (chr3:50610197-50710197 and chr3:50660197-50760197) in the BaTwa RHGs.

(C and D) Dot colors indicate SNP F_{CS} percentiles, black squares indicate non-synonymous mutations, and black dots indicate eQTLs (q value < 0.005) [33]. eQTLs of *MAPKAPK3* (rs107457 and rs9879397) and *DOCK3* (rs12629788) are shown as yellow diamonds. Not all genes of the genomic region are shown for convenience.

See also Figures S2 and S3 and Data S1.

20 RHG Baka and 20 AGR Nzébi from Gabon (5–6× coverage) and (2) the 1000 Genomes Phase 3 panel [27] (STAR Methods; Figure S1). We focused on the five RHG populations presenting the lowest average levels of AGR ancestry and analyzed the highly admixed RHG groups differently (see Recent Genetic Adaptation of Admixed Rainforest Hunter-Gatherers). To identify signals of strong sweeps, we searched for variants with both high allele frequency and extended haplotype homozygosity in RHGs, relative to AGRs (STAR Methods). Genome-wide ranks of PBS [28] and XP-EHH [29] were combined into a Fisher's score (F_{CS}), and to reduce false positives, candidate regions were defined as 100-kb windows with the 1% highest proportion of outlier SNPs of the genome.

We first scanned the genomes of AGR populations (Figure S2), the evolutionary history of whom is well characterized [24, 29–32]. We found 18 candidate regions for positive selection in both western and eastern AGRs, while only ~3.5 were expected to be shared if candidate loci were false positives (10,000 random samples; resampling $p < 10^{-4}$) (Figure 2A; Data S1). Among candidates, we replicated, for example, the signal encompassing the *LARGE* gene, involved in Lassa virus infectivity [34]. These results provide evidence that the genomic regions detected by our approach are enriched in true signals.

A Strong, Shared Selective Sweep at *TRPS1* across All Hunter-Gatherer Groups

Our search for sweeps in RHGs identified candidates that were shared by RHG groups more than expected by chance (resampling $p < 10^{-4}$) (Figure 2A; Data S1). Remarkably, we identified a single genomic region that exhibits sweep signals in all RHG populations, but not in AGRs (Figures 2A–2C and S3). This region lies upstream of the 5′ UTR of *TRPS1*, which encodes a transcription factor (TF) with multiple pleiotropic effects, including skeletal development and inflammatory T_H17 cell differentiation [35–37]. The six variants presenting the highest frequency differences between RHGs and AGRs (Data S1) define a 5,777 bp region that contains a primate-specific THE1B endogenous retrovirus sequence, known to control the expression of nearby genes [38]. Given the high expression of *TRPS1* in monocytes [39], we analyzed published RNA sequencing (RNA-seq) data from monocytes of individuals of central African ancestry to test if candidate variants affect *TRPS1* expression

[40]. A highly differentiated variant that falls within the THE1B fragment was associated with increased expression of a short, non-canonical *TRPS1* transcript upon immune stimulation (rs111351287; regression $p = 5 \times 10^{-6}$). These findings suggest that the most robust signal of adaptation to the African rainforest can be ascribed to *TRPS1*, possibly in relation with variation in morphological and/or immunological traits.

Detection of Other Classic Sweep Signals in Rainforest Hunter-Gatherers

Other selective sweep signals were specific to a smaller number of RHG groups (Figure S2; Data S1). These include the known 150-kb region encompassing *CISH*, *MAPKAPK3*, and *DOCK3* [6, 14], which we show here to be shared among western and eastern RHGs (Baka, BaKoya, and BaTwa). We searched the GTEx database [33] for regulatory variation at these genes (eQTLs) and found two *cis*-eQTLs for *MAPKAPK3* (rs107457 and rs9879397), one for *DOCK3* (rs12629788), and none for *CISH* (Data S1). Selection scores at these eQTLs were among the highest of the region, particularly for *MAPKAPK3* (Figure 2D), which affects hepatitis C virus (HCV) infectivity [41].

We also detected two contiguous regions at the *IFIH1* locus [18], which present strong enrichments in selection scores that are shared by all western RHG groups. Candidate variation at this locus (rs12479043) controls the expression of the nearby *FAP* gene [33], which regulates fibroblast and myofibroblast growth and wound healing during chronic inflammation [42]. We also identified two windows—shared by Bezan, Baka, and BaKoya—encompassing *RASGEF1B*, whose expression is induced in macrophages by lipopolysaccharide, a membrane component of Gram-negative bacteria [43]. Finally, we found a window in the Bezan, BaBongo, and BaKoya that overlaps *PITX1*, recently identified as a selection candidate in RHGs [22]. PITX1 modulates the core development of limb [44], is associated with height variation [45], acts as an early TF in the developing pituitary gland [46], and regulates interferon-α virus induction [47]. These results support the hypothesis that development and immunity are key traits in local adaptation to the rainforest.

Evidence for Polygenic Selection Favoring the "Pygmy" Phenotype

Given the polygenic nature of most adaptive traits [48, 49], we searched for evidence of polygenic adaptation focusing on 12 candidate quantitative traits. These include height, body mass index, skin pigmentation, life history traits, and immune cell counts, the genetic architectures of which have been extensively studied [50]. We compared the distribution of mean F_{CS} scores in non-overlapping, 100-kb genomic windows containing trait-associated SNPs to that of randomly sampled windows, accounting for SNP density, LD levels, and background selection (STAR Methods). Stature-related traits showed the most significant polygenic selection signals, in all RHG groups (adjusted $p < 0.05$) while being non-significant in AGRs (Figure 3A). Life-history traits related to reproduction also exhibited selection signals in various RHG groups, consistent with the proposed adaptive nature of early reproduction in RHGs [51, 52]. Furthermore, we replicated selection signals for cardiovascular traits in the BaTwa (adjusted $p < 0.001$) [23]. Notably, we found significant signals in "Leukocyte count" in the Baka and the BaBongo (adjusted $p < 0.05$), suggesting polygenic adaptation related to immunity.

We next examined whether signals of polygenic selection could result from pleiotropy; e.g., advantageous height-associated variants affect other correlated traits [49]. Using the UK Biobank dataset [50], we computed the genetic correlations from LD-score regressions between "Standing height" and the remaining traits, and found significant correlations for eight of them (STAR Methods; Data S1). For these, we repeated the analysis after excluding windows associated with "Standing height" or "Comparative height at age 10," and the significance of selection signals was lost or dramatically reduced (Figures 3B and S4). Conversely, when excluding windows associated with non-height traits (e.g., reproduction-related traits), we found that "Standing height" was still significant in four RHG populations (adjusted $p < 0.05$) (Figure 3C). These results show that height has been an adaptive trait in RHGs, resulting in spurious polygenic selection signals for other correlated traits because of pleiotropy.

Evidence of Pervasive Pathogen-Driven Selection in the Equatorial Rainforest

We further investigated genomic signatures of polygenic adaptation, by searching for excesses in mean F_{CS} among windows related to 5,354 gene ontology (GO) terms [53] (STAR Methods). We detected 38 terms that were significant in at least three RHG groups, but not in AGRs (Figure 3D; Data S1). Among these, we found positive regulation of "mast cell degranulation" and "the phosphatidylinositol 3-kinase (PI3K) pathway" (false discovery rate [FDR] $p < 5\%$). Recognition by mast cells of allergens and antigens induces degranulation, a process mediated by the PI3K pathway that results in inflammation and allergy [54]. Enrichments were also found in the IL-2 signaling pathway, which activates the PI3K pathway and regulates immune tolerance [55]. All enrichments remained significant after removing windows associated with height (FDR $p < 5\%$), excluding potential pleiotropic effects. To gain further insights into pathogen-driven selection, we next focused on 1,553 innate immunity genes (IIGs) [56] and 1,257 genes encoding virus-interacting proteins (VIPs) [57]. We found significant enrichments in selection signals for both gene sets in RHGs, but not in AGRs, in particular for VIPs interacting with double-stranded DNA (dsDNA) and single-stranded RNA (ssRNA) viruses (FDR $p < 5\%$; Table S2; Data S1). These results collectively support the notion that pathogens have been a major driver of local adaptation in the African rainforest.

Recent Genetic Adaptation of Admixed Rainforest Hunter-Gatherers

To search for evidence of recent selection in RHG since their admixture with AGRs, we focused on the highly admixed BaBongo (Figure 1C) and performed local ancestry inference with RFMix [58], using as putative parental populations western RHG and AGR individuals with the lowest AGR and RHG membership proportions, respectively (STAR Methods). Six contiguous windows on chromosome 1 showed both evidence of selection (i.e., top 1% of the proportion of outlier SNPs) and an excess of RHG local ancestry (i.e., higher than the genome-wide average + 2 SD) in admixed RHG (Figures 4A and S2; Data S1). Among the

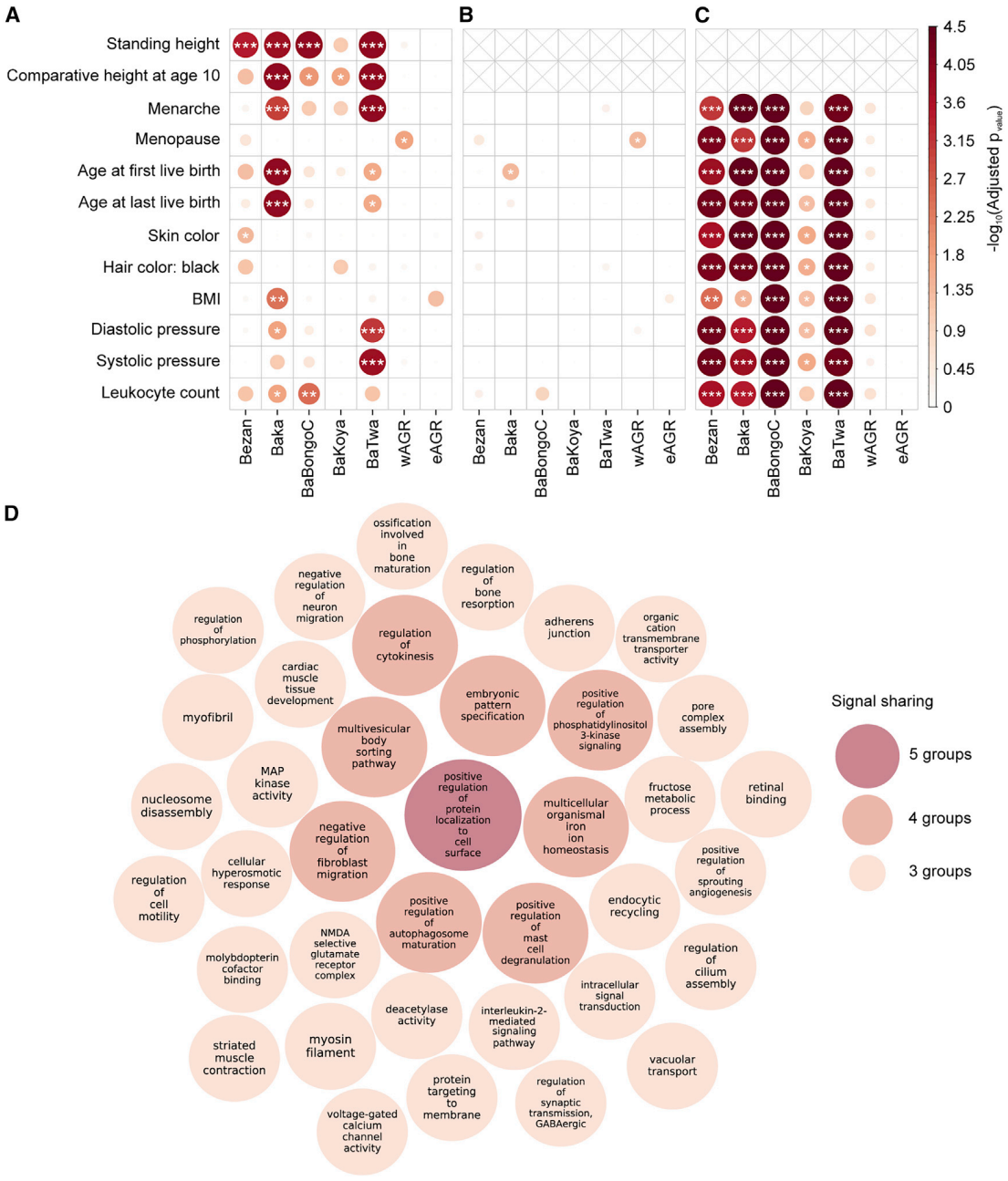


Figure 3. Signals of Polygenic Selection in African Rainforest Hunter-Gatherers

(A) Signals of polygenic selection for 12 candidate quantitative traits, based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations.

(B) Signals of polygenic selection for the candidate quantitative traits, based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations, after removing windows associated with "Standing height" and "Comparative height at age 10." Loss of significance was not explained by the reduced number of windows tested (Figure S4).

(C) Signals of polygenic selection for "Standing height," based on higher mean F_{CS} of trait-associated windows relative to genome-wide expectations, after removing windows associated with each of the remaining quantitative traits.

(A–C) Color gradient and circle sizes are proportional to $-\log_{10}$(adjusted p) with adjusted *p < 0.05, **p < 0.01, and ***p < 0.001. Multiple testing corrections were performed using the Benjamini-Hochberg method. wAGR and eAGR stand for western and eastern AGR groups. Signals were generally stronger in Baka and BaTwa RHGs, probably because of their larger sample size.

(D) Gene Ontology (GO) terms enriched in selection scores (FDR p < 5%) in RHG, but not in AGR, populations, considering the window mean F_{CS} as selection score. Circle color and size indicate the number of RHG populations that show significant evidence of polygenic selection for a given GO term.

See also Figure S4, Table S2, and Data S1.

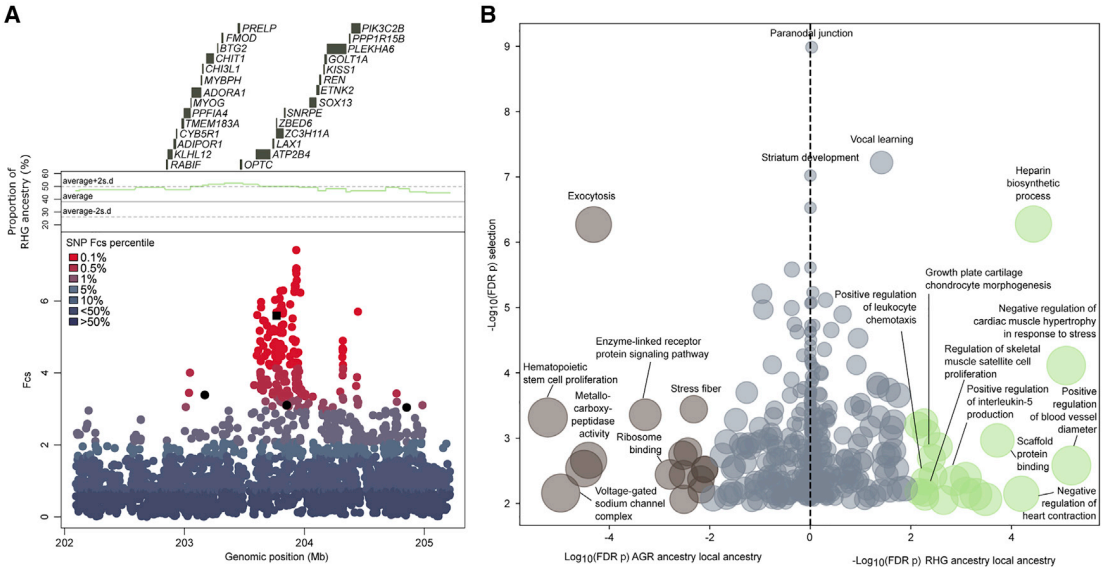


Figure 4. Selection Signals in Highly Admixed Rainforest Hunter-Gatherers

(A) Selective sweep signal and average local RHG ancestry at the chr1:203564464-203764464 locus in the highly admixed RHG BaBongo. Dot colors indicate SNP F_{CS} percentiles, the black square indicates the non-synonymous variant (rs6697388) at *ZBED6*, and black dots indicate eQTLs (q value < 0.005) [33].

(B) GO terms enriched in both local ancestry in the highly admixed RHG BaBongo, and selection scores in each of the two putative parental populations, with respect to the rest of the genome (FDR p < 5%). Green (brown) dots indicate GO terms enriched in both western RHG (western AGR) local ancestry and selection scores in parental western RHG (western AGR) populations (FDR p < 5%). Enrichments were assessed using the Mann-Whitney-Wilcoxon rank-sum test. See also Data S1.

strongest candidate variants, we found a non-synonymous mutation (rs6697388) in *ZBED6*, which encodes a TF that controls muscle growth through *IGF2* repression [59]. *ZBED6* is located within the intron of the *ZC3H11A* gene, whose product is required for the efficient growth of several nuclear-replicating viruses [60]. The rs6697388 G allele (p.Leu391Arg) is present at the highest frequency in admixed BaBongo (51%), with lower frequencies in parental RHG (42%) and AGR (15%) groups. With respect to the strong, shared selective sweep detected at *TRPS1* (Figure 2C), the locus also presented selection signals in the BaBongo but no excess of RHG or AGR ancestry (Figures S2 and S3), suggesting weaker or no positive selection at *TRPS1* since admixture.

Finally, we searched for evidence of polygenic selection since admixture, by testing for excesses in AGR or RHG local ancestry in genomic windows related to GO terms in the admixed BaBongo (STAR Methods). We found 21 GO terms that were enriched in both RHG local ancestry and selection signals in the parental RHGs (Figure 4B; Data S1), an overlap that was significantly larger than expected (7.3% versus 4.7%, χ^2 test, p = 0.042). These terms were mostly related to cardiac and skeletal development and immune functions, and included "heparin biosynthetic process," which participates in mast cell-mediated immune and inflammatory responses [61], echoing the signals detected for "mast cell degranulation" in weakly admixed RHGs (Figure 3D). We also found 16 GO terms that were enriched in both AGR local ancestry and selection signals in the parental AGRs (Figure 4B; Data S1), including stem cell proliferation, exocytosis, and muscle composition. Together, these results support further the notion that heart and bone development as well as immune responses have been an important substrate of selection in RHGs, before and after their admixture with neighboring farmers.

DISCUSSION

Here we present the first exome-based survey of multiple geographically dispersed groups of African rainforest hunter-gatherers, with the aim of investigating how populations have adapted to the challenging habitats of the equatorial rainforest. Because positive selection often targets regulatory regions [62], we combined the exome dataset with SNP array data, to cover both genic and intergenic regions. In doing so, we found evidence of a unique, strong sweep that is shared by all RHG groups, targeting the regulatory region of *TRPS1*, mutations in which can cause growth retardation, distinctive craniofacial features [63], and hypertrichosis [64]. Furthermore, the transcription factor TRPS1 regulates STAT3, a mediator of inflammation and immunity [65], and RUNX2, controlling facial features and viral clearance [66, 67]. Interestingly, *TRPS1* has been recently shown to carry signals of archaic introgression in western Africans [68]. Functional studies should help determine the adaptive nature—developmental and/or immune-related—of variation at this locus, which possibly introgressed from extinct African hominins [18, 68, 69].

This study also extends previous findings of a sweep targeting the *CISH-MAPKAPK3-DOCK3* region [6, 14], by delineating *MAPKAPK3* as the most likely target. *MAPKAPK3* expression is regulated by two eQTLs that are among the strongest candidates for positive selection at the locus in RHG populations. MAPKAPK3 directly interacts with HCV and regulates cell infectivity [41]. A lower prevalence of HCV infection has been reported in RHG, with respect to AGR [70, 71]. Our results strengthen the evolutionary importance of the *CISH-MAPKAPK3-DOCK3* region in both western and eastern RHGs, and pinpoint

MAPKAPK3 variation as a putative, additional risk factor for HCV infection in Africans.

Our analyses provide robust evidence for polygenic selection of height, which we replicate in various RHG groups. Importantly, our results are not affected by biased genome-wide association study (GWAS) summary statistics due to partial control for population stratification, which can result in spurious polygenic selection signals [72, 73]. Our approach tests for the co-localization of selection signals and trait-associated genes; thus, it does not depend on effect size estimates and does not assume that associated variants are the same across populations. More generally, polygenic selection of height is unlikely to result from sexual selection [74] but from genetic adaptation to equatorial forest environments [75]. Our study sheds new light onto the debated adaptive nature of height, and supports that the early reproductive age of RHGs is not the cause of their small body size, as previously suggested [51, 52]. Instead, our results suggest that directional selection of height has resulted in changes in life-history traits because of pervasive pleiotropy of height-associated genes.

We also found signals of polygenic selection in RHGs at functions related to the IL-2 pathway, the sensing of allergens and microbes, and interactions with dsDNA and ssRNA viruses. Interestingly, higher seropositivity for more than 30 viruses has been reported in the BaTwa from Uganda, with respect to AGRs, particularly for dsDNA viruses [76]. That we also found an excess of RHG ancestry related to heparin biosynthesis, interleukin production, and leukocyte chemotaxis in highly admixed RHGs suggests preferential retention of RHG variation at immune-related functions. This finding supports a long-standing history of adaptation of RHGs to high pathogen pressures. This contrasts with a study in southern Africa, which reported a low exposure and adaptation to pathogens of hunter-gatherers of the Kalahari Desert, except for those who recently came in contact with other populations [77].

Collectively, our analyses uncover height, development, and immune response as iconic adaptive traits of African RHGs. It is interesting to note that the PI3K signaling pathway—under polygenic selection in four RHG populations—modulates inflammatory responses [78], body energy homeostasis [79, 80], and insulin secretion [81]. Several studies have highlighted the reciprocal relationship between proinflammatory cytokines and the regulation of the growth hormone through the IGF-1 axis [82]. It is thus tempting to speculate that pleiotropic effects between developmental growth and immunity could have further participated in the "pygmy" phenotype. Epidemiological work on the infectious disease burden in hunter-gatherers should increase our understanding of how historical high pathogen-driven selection has contributed to the reduced stature characterizing populations of the rainforest.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.07.013.

ACKNOWLEDGMENTS

We thank all the participants for providing the DNA samples used in this study. We also thank Guillaume Laval and Maxime Rotival for useful discussions, Muh-Ching Yee for laboratory assistance, and the "Paléogénomique et Génétique Moléculaire" Platform from the MnHn at the *Musée de l'Homme* for assistance in sample preparation. This work was supported by the Institut Pasteur, the Centre National de la Recherche Scientifique, the "Histoire du Génome des Populations Humaines Gabonaises" project (Institut Pasteur/Republic of Gabon), and an Agence Nationale de la Recherche grant "AGRHUM" (ANR-14-CE02-0003-01) to L.Q.-M. M.L. was supported by the Fondation pour la Recherche Médicale (FDT20170436932), and J.C. by the INCEPTION program and the "Ecole Doctorale FIRE - Programme Bettencourt."

AUTHOR CONTRIBUTIONS

E.P. and L.Q.-M. conceived and supervised the study. M.L. conducted all the analyses and analyzed the data, with contributions from J.C., M. Silvert, and E.P. C.H. performed laboratory work. M. Sikora, K.S., H.C., and C.D.B. generated and/or analyzed whole-genome data. P.M.-D., J.-M.H., A.F., S.L.B., G.H.P., L.B.B., and P.V. assembled the samples. M.L., E.P., and L.Q.-M. wrote the manuscript, with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 14, 2019
Revised: June 26, 2019
Accepted: July 4, 2019
Published: August 8, 2019

REFERENCES

1. Hewlett, B.S. (2014). Hunter-Gatherers of the Congo Basin: Cultures, Histories and Biology of African Pygmies (Transaction Publishers).
2. Perry, G.H., and Verdu, P. (2017). Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. Quat. Int. *448*, 150–157.
3. Bahuchet, S. (2012). Changing language, remaining pygmy. Hum. Biol. *84*, 11–43.
4. Lopez, M., Kousathanas, A., Quach, H., Harmant, C., Mouguiama-Daouda, P., Hombert, J.M., Froment, A., Perry, G.H., Barreiro, L.B., Verdu, P., et al. (2018). The demographic history and mutational load of African hunter-gatherers and farmers. Nat Ecol Evol *2*, 721–730.
5. Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.M., et al. (2009).

Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr. Biol. *19*, 312–318.

6. Hsieh, P., Veeramah, K.R., Lachance, J., Tishkoff, S.A., Wall, J.D., Hammer, M.F., and Gutenkunst, R.N. (2016). Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. Genome Res. *26*, 279–290.
7. Patin, E., Laval, G., Barreiro, L.B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K.K., Kidd, J.R., Van der Veen, L., Hombert, J.M., et al. (2009). Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet. *5*, e1000448.
8. Batini, C., Lopes, J., Behar, D.M., Calafell, F., Jorde, L.B., van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D. (2011). Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol. Biol. Evol. *28*, 1099–1110.
9. Veeramah, K.R., Wegmann, D., Woerner, A., Mendez, F.L., Watkins, J.C., Destro-Bisol, G., Soodyall, H., Louie, L., and Hammer, M.F. (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol. Biol. Evol. *29*, 617–630.
10. Aimé, C., Laval, G., Patin, E., Verdu, P., Ségurel, L., Chaix, R., Hegay, T., Quintana-Murci, L., Heyer, E., and Austerlitz, F. (2013). Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. Mol. Biol. Evol. *30*, 2629–2644.
11. Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S., et al. (2008). Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc. Natl. Acad. Sci. USA *105*, 1596–1601.
12. Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., et al. (2014). The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat. Commun. *5*, 3163.
13. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
14. Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G., et al. (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet. *8*, e1002641.
15. Migliano, A.B., Romero, I.G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., Huang, D.W., Sherman, B.T., Siddle, K., Scholes, C., et al. (2013). Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. Hum. Biol. *85*, 251–284.
16. Becker, N.S., Verdu, P., Georges, M., Duquesnoy, P., Froment, A., Amselem, S., Le Bouc, Y., and Heyer, E. (2013). The role of GHR and IGF1 genes in the genetic determination of African pygmies' short stature. Eur. J. Hum. Genet. *21*, 653–658.
17. Pemberton, T.J., Verdu, P., Becker, N.S., Willer, C.J., Hewlett, B.S., Le Bomin, S., Froment, A., Rosenberg, N.A., and Heyer, E. (2018). A genome scan for genes underlying adult body size differences between Central African hunter-gatherers and farmers. Hum. Genet. *137*, 487–509.
18. Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell *150*, 457–469.
19. Mendizabal, I., Marigorta, U.M., Lao, O., and Comas, D. (2012). Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum. Genet. *131*, 1305–1317.
20. Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nédélec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S.L., et al. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proc. Natl. Acad. Sci. USA *111*, E3596–E3603.
21. Amorim, C.E., Daub, J.T., Salzano, F.M., Foll, M., and Excoffier, L. (2015). Detection of convergent genome-wide signals of adaptation to tropical forests in humans. PLoS ONE *10*, e0121557.
22. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., et al. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. *20*, 82.
23. Bergey, C.M., Lopez, M., Harrison, G.F., Patin, E., Cohen, J.A., Quintana-Murci, L., Barreiro, L.B., and Perry, G.H. (2018). Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. Proc. Natl. Acad. Sci. USA *115*, E11256–E11263.
24. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543–546.
25. Fagny, M., Patin, E., MacIsaac, J.L., Rotival, M., Flutre, T., Jones, M.J., Siddle, K.J., Quach, H., Harmant, C., McEwen, L.M., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat. Commun. *6*, 10047.
26. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
28. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.
29. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
30. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., et al.; 1000 Genomes Project (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.
31. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
32. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
33. Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

34. Andersen, K.G., Shylakhter, I., Tabrizi, S., Grossman, S.R., Happi, C.T., and Sabeti, P.C. (2012). Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 868–877.

35. Fantauzzo, K.A., and Christiano, A.M. (2012). Trps1 activates a network of secreted Wnt inhibitors and transcription factors crucial to vibrissa follicle morphogenesis. Development *139*, 203–214.

36. Wuelling, M., Kaiser, F.J., Buelens, L.A., Braunholz, D., Shivdasani, R.A., Depping, R., and Vortkamp, A. (2009). Trps1, a regulator of chondrocyte proliferation and differentiation, interacts with the activator form of Gli3. Dev. Biol. *328*, 40–53.

37. Yosef, N., Shalek, A.K., Gaublomme, J.T., Jin, H., Lee, Y., Awasthi, A., Wu, C., Karwacz, K., Xiao, S., Jorgolli, M., et al. (2013). Dynamic regulatory network controlling TH17 cell differentiation. Nature *496*, 461–468.

38. Dunn-Fletcher, C.E., Muglia, L.M., Pavlicev, M., Wolf, G., Sun, M.A., Hu, Y.C., Huffman, E., Tumukuntala, S., Thiele, K., Mukherjee, A., et al. (2018). Anthropoid primate-specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. PLoS Biol. *16*, e2006337.

39. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell *175*, 1701–1715.e16.

40. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell *167*, 643–656.e17.

41. Ngo, H.T., Pham, L.V., Kim, J.W., Lim, Y.S., and Hwang, S.B. (2013). Modulation of mitogen-activated protein kinase-activated protein kinase 3 by hepatitis C virus core protein. J. Virol. *87*, 5718–5731.

42. Tillmanns, J., Hoffmann, D., Habbaba, Y., Schmitto, J.D., Sedding, D., Fraccarollo, D., Galuppo, P., and Bauersachs, J. (2015). Fibroblast activation protein alpha expression identifies activated fibroblasts after myocardial infarction. J. Mol. Cell. Cardiol. *87*, 194–203.

43. Andrade, W.A., Silva, A.M., Alves, V.S., Salgado, A.P., Melo, M.B., Andrade, H.M., Dall'Orto, F.V., Garcia, S.A., Silveira, T.N., and Gazzinelli, R.T. (2010). Early endosome localization and activity of RasGEF1b, a toll-like receptor-inducible Ras guanine-nucleotide exchange factor. Genes Immun. *11*, 447–457.

44. Nemec, S., Luxey, M., Jain, D., Huang Sung, A., Pastinen, T., and Drouin, J. (2017). *Pitx1* directly modulates the core limb development program to implement hindlimb identity. Development *144*, 3325–3335.

45. Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. *14*, e1007371.

46. Szeto, D.P., Rodriguez-Esteban, C., Ryan, A.K., O'Connell, S.M., Liu, F., Kioussi, C., Gleiberman, A.S., Izpisúa-Belmonte, J.C., and Rosenfeld, M.G. (1999). Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. Genes Dev. *13*, 484–494.

47. Island, M.L., Mesplede, T., Darracq, N., Bandu, M.T., Christeff, N., Djian, P., Drouin, J., and Navarro, S. (2002). Repression by homeoprotein pitx1 of virus-induced interferon a promoters is mediated by physical interaction and trans repression of IRF3 and IRF7. Mol. Cell. Biol. *22*, 7120–7133.

48. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.

49. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.

50. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

51. Migliano, A.B., Vinicius, L., and Lahr, M.M. (2007). Life history trade-offs explain the evolution of human pygmies. Proc. Natl. Acad. Sci. USA *104*, 20216–20219.

52. Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., De Souza, R., Djurovic, G., Hames, R., Hurtado, A.M., Kaplan, H., et al. (2006). Growth rates and life histories in twenty-two small-scale societies. Am. J. Hum. Biol. *18*, 295–311.

53. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

54. Kim, M.S., Rådinger, M., and Gilfillan, A.M. (2008). The multiple roles of phosphoinositide 3-kinase in mast cell biology. Trends Immunol. *29*, 493–501.

55. Malek, T.R., and Castro, I. (2010). Interleukin-2 receptor signaling: at the interface between tolerance and immunity. Immunity *33*, 153–165.

56. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. *98*, 5–21.

57. Enard, D., and Petrov, D.A. (2018). Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. Cell *175*, 360–371.e13.

58. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.

59. Younis, S., Schönke, M., Massart, J., Hjortebjerg, R., Sundström, E., Gustafson, U., Björnholm, M., Krook, A., Frystyk, J., Zierath, J.R., and Andersson, L. (2018). The ZBED6-IGF2 axis has a major effect on growth of skeletal muscle and internal organs in placental mammals. Proc. Natl. Acad. Sci. USA *115*, E2048–E2057.

60. Younis, S., Kamel, W., Falkeborn, T., Wang, H., Yu, D., Daniels, R., Essand, M., Hinkula, J., Akusjärvi, G., and Andersson, L. (2018). Multiple nuclear-replicating viruses require the stress-induced protein ZC3H11A for efficient growth. Proc. Natl. Acad. Sci. USA *115*, E3808–E3816.

61. Humphries, D.E., Wong, G.W., Friend, D.S., Gurish, M.F., Qiu, W.T., Huang, C., Sharpe, A.H., and Stevens, R.L. (1999). Heparin is essential for the storage of specific granule proteases in mast cells. Nature *400*, 769–772.

62. Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009). Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. *26*, 649–658.

63. Momeni, P., Glöckner, G., Schmidt, O., von Holtum, D., Albrecht, B., Gillessen-Kaesbach, G., Hennekam, R., Meinecke, P., Zabel, B., Rosenthal, A., et al. (2000). Mutations in a new gene, encoding a zinc-finger protein, cause tricho-rhino-phalangeal syndrome type I. Nat. Genet. *24*, 71–74.

64. Fantauzzo, K.A., Tadin-Strapps, M., You, Y., Mentzer, S.E., Baumeister, F.A., Cianfarani, S., Van Maldergem, L., Warburton, D., Sundberg, J.P., and Christiano, A.M. (2008). A position effect on TRPS1 is associated with Ambras syndrome in humans and the Koala phenotype in mice. Hum. Mol. Genet. *17*, 3539–3551.

65. Hillmer, E.J., Zhang, H., Li, H.S., and Watowich, S.S. (2016). STAT3 signaling in immunity. Cytokine Growth Factor Rev. *31*, 1–15.

66. Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R.B., Pérez, G.M., et al. (2016). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat. Commun. *7*, 11616.

67. Chopin, M., Preston, S.P., Lun, A.T.L., Tellier, J., Smyth, G.K., Pellegrini, M., Belz, G.T., Corcoran, L.M., Visvader, J.E., Wu, L., and Nutt, S.L. (2016). RUNX2 mediates plasmacytoid dendritic cell egress from the bone marrow and controls viral immunity. Cell Rep. *15*, 866–878.

68. Durvasula, A., and Sankararaman, S. (2019). Recovering signals of ghost archaic introgression in African populations. bioRxiv. https://doi.org/10.1101/285734.

69. Hsieh, P., Woerner, A.E., Wall, J.D., Lachance, J., Tishkoff, S.A., Gutenkunst, R.N., and Hammer, M.F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res. *26*, 291–300.

70. Foupouapouognigni, Y., Mba, S.A., Betsem à Betsem, E., Rousset, D., Froment, A., Gessain, A., and Njouom, R. (2011). Hepatitis B and C virus infections in the three Pygmy groups in Cameroon. J. Clin. Microbiol. *49*, 737–740.

71. Kowo, M.P., Goubau, P., Ndam, E.C., Njoya, O., Sasaki, S., Seghers, V., and Kesteloot, H. (1995). Prevalence of hepatitis C virus and other blood-borne viruses in Pygmies and neighbouring Bantus in southern Cameroon. Trans. R. Soc. Trop. Med. Hyg. *89*, 484–486.

72. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.

73. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, e39702.

74. Becker, N., Touraille, P., Froment, A., Heyer, E., and Courtiol, A. (2012). Short stature in African pygmies is not explained by sexual selection. Evol. Hum. Behav. *33*, 615–622.

75. Perry, G.H., and Dominy, N.J. (2009). Evolution of the human pygmy phenotype. Trends Ecol. Evol. *24*, 218–225.

76. Harrison, G.F., Sanz, J., Boulais, J., Mina, M.J., Grenier, J.-C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C.M., Elledge, S.J., et al. (2019). Natural selection contributed to immunological differences between human hunter-gatherers and agriculturalists. bioRxiv. https://doi.org/10.1101/487207.

77. Owers, K.A., Sjödin, P., Schlebusch, C.M., Skoglund, P., Soodyall, H., and Jakobsson, M. (2017). Adaptation to infectious disease exposure in indigenous Southern African populations. Proc. Biol. Sci. *284*, 20170226.

78. Hawkins, P.T., and Stephens, L.R. (2015). PI3K signalling in inflammation. Biochim. Biophys. Acta *1851*, 882–897.

79. Hopkins, B.D., Pauli, C., Du, X., Wang, D.G., Li, X., Wu, D., Amadiume, S.C., Goncalves, M.D., Hodakoski, C., Lundquist, M.R., et al. (2018). Suppression of insulin feedback enhances the efficacy of PI3K inhibitors. Nature *560*, 499–503.

80. Fruman, D.A., Chiu, H., Hopkins, B.D., Bagrodia, S., Cantley, L.C., and Abraham, R.T. (2017). The PI3K pathway in human disease. Cell *170*, 605–635.

81. Odegaard, J.I., and Chawla, A. (2013). Pleiotropic actions of insulin resistance and inflammation in metabolic homeostasis. Science *339*, 172–177.

82. Smith, T.J. (2010). Insulin-like growth factor-I regulation of immune function: a potential therapeutic target in autoimmune diseases? Pharmacol. Rev. *62*, 199–236.

83. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

84. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

85. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

86. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

87. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

88. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

89. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

90. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., et al. (2018). GOATOOLS: a Python library for Gene Ontology analyses. Sci. Rep. *8*, 10872.

91. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

92. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

93. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 1–33.

94. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

95. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

96. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

97. Cavalli-Sforza, L.L. (1986). African Pygmies (Academic Press).

98. Diamond, J.M. (1991). Anthropology. Why are pygmies small? Nature *354*, 111–112.

99. Shea, B.T., and Bailey, R.C. (1996). Allometry and adaptation of body proportions and stature in African pygmies. Am. J. Phys. Anthropol. *100*, 311–340.

100. Froment, A. (2001). Hunter-gatherers: an interdisciplinary perspective. In Hunter-Gatherers: An Interdisciplinary Perspective, L.R.-C. Panter-Brick, ed. (Cambridge University Press), pp. 239–266.

101. Becker, N.S., Verdu, P., Hewlett, B., and Pavard, S. (2010). Can life history trade-offs explain the evolution of short stature in human pygmies? A response to Migliano et al. (2007). Hum. Biol. *82*, 17–27.

102. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
Nextera Rapid Capture Expanded Exome kit	Illumina	Cat#FC-140-1006
HumanOmniExpress-24 v1.1 DNA Analysis Kit	Illumina	N/A
Deposited Data		
Exome and whole-genome sequencing	This paper	EGAS00001003722
Software and Algorithms		
PLINK v1.9	[83]	http://www.cog-genomics.org/plink/1.9/
KING v1.4	[84]	http://people.virginia.edu/~wc9c/KING/history.htm
ADMIXTURE	[26]	http://software.genetics.ucla.edu/admixture/download.html
BWA v.0.7.7	[85]	http://bio-bwa.sourceforge.net/
Picard Tools v.1.94	N/A	http://broadinstitute.github.io/picard
GATK v3.5	[86]	https://software.broadinstitute.org/gatk/download/
SHAPEIT2	[87]	http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
IMPUTE v.2	[88]	http://mathgen.stats.ox.ac.uk/impute/impute_v2.1.0.html
LDSC	[89]	https://data.broadinstitute.org/alkesgroup/LDSCORE/
GOATOOLS	[90]	https://github.com/tanghaibao/goatools
RFMIX v1.5.4	[58]	https://sites.google.com/site/rfmixlocalancestryinference/
BEAGLE	[91]	https://faculty.washington.edu/browning/beagle/beagle.html
GERP++	[92]	http://mendel.stanford.edu/SidowLab/downloads/gerp/

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents. Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Lluís Quintana-Murci (quintana@pasteur.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection

Sampling consisted in human saliva or blood from 157 rainforest hunter-gatherers and 120 farmers from western and eastern central Africa (Figure S1), including 208 males and 69 females. Informed consent was obtained from all participants in this study, which was overseen by the institutional review board of Institut Pasteur (2011-54/IRB/8), the *Comité National d'Ethique du Gabon* (0016/2016/SG/CNE), the University of Chicago (IRB 16986A) and Makerere University, Kampala, Uganda (IRB 2009-137). The 277 new samples collected for exome sequencing were analyzed together with 317 exomes of central Africans from Lopez et al. 2018 [4] and 101 Europeans from Quach et al. 2016 [40] (Table S1).

METHOD DETAILS

Exome Sequencing

Sample libraries were prepared with the Nextera Rapid Capture Expanded Exome Kit, which delivers 62Mb of genomic content per individual, including exons, untranslated regions and microRNAs, and were sequenced on Illumina HiSeq2500 machines. Using the GATK Best Practices recommendations [93], pairs of 101-bp reads were mapped onto the human reference genome (GRCh37) with Burrows-Wheeler Aligner (BWA) version 0.7.7 [85], using 'bwa mem -M -t 4 -R', and reads duplicating the start position of another read were marked as duplicates with Picard Tools version 1.94 (http://broadinstitute.github.io/picard/), using 'MarkDuplicates'. We used GATK version 3.5 [86] for base quality score recalibration ('Base Recalibrator'), insertion/deletion (indel) realignment ('IndelRealigner'), and SNP and indel discovery ('Haplotype Caller') for each sample. Individual variant files were combined with 'GenotypeGVCFs' and filtered with 'VariantQualityScoreRecalibration'. We used high confidence variants from the 1000G Phase 1 and HapMap 3 projects [94, 95] as VQSR training callsets, and applied a tranche sensitivity threshold of 99.5%. From the 947,523 sites detected, we removed indels as well as SNPs that (i) were located on the sex chromosomes,

(ii) were not biallelic, (iii) were monomorphic in our total sample, (iv) had a depth of coverage < 5 ×, (v) had a genotype quality score (GQ) < 20, (vi) presented missingness > 15%, and (vii) presented a Hardy-Weinberg test $p < 10^{-6}$ in at least one of population. As criteria to remove low-quality samples, we required a total genotype missingness < 15% (21 excluded samples). In addition, we checked for unexpectedly high or low heterozygosity values, suggesting high levels of inbreeding or DNA contamination, and excluded 3 individuals presenting heterozygosity levels 4 SD higher than their population average. We thus retained exome data for 671 individuals, with an average depth of coverage after duplicate removal of 38 × (SD: 9 ×), ranging from 25 × to 95 ×. The application of these quality-control filters resulted in a final dataset of 682,468 SNPs (Figure S1), of which 107,621 SNPs were polymorphic only in the 268 newly-sequenced individuals.

SNP Array Data

In addition to exome sequencing, we retrieved the genotyping data of the same 671 individuals from Quach et al. 2016 [40], Patin et al. 2014 [12], Patin et al. 2017 [24] and Fagny et al. 2015 [25] (Figure S1; Table S1). We removed SNPs located on the X and Y chromosomes, problematic genotype clustering profiles (i.e., Illumina GenTrain score < 0.35) or with call rate < 95%. We kept 599,559 SNPs common to different genotyping SNP arrays. We removed a total of 53 C/G or A/T SNPs to prevent misaligned SNPs, and excluded a total 5 additional SNPs that were under Hardy-Weinberg disequilibrium in at least one of the populations ($p < 10^{-6}$) using PLINK [96], leading to a final dataset of 559,501 SNPs.

We applied additional filters on the genotyping dataset of the 671 individuals retained for exome sequencing. We removed two individuals with heterozygosity levels higher or lower than the population mean ± 4 SD Although related individuals were avoided during the sampling and for exome sequencing (based on published SNP array data) [5, 12, 17, 24, 25], we sought to exclude possibly remaining pairs of cryptically related individuals. Indeed, RHG populations are small isolated communities, where individuals can be related to many others. We considered that two individuals were strongly (cryptically) related if they presented a first-degree relationship (kinship coefficient > 0.177), as inferred by KING [84]. Following this criterion, only one individual was removed. Additionally, we removed another individual who did not present any first-degree relatedness but was related in second-degree to many others. After removing these two individuals, the dataset included 77 and 232 pairs of second-degree (kinship coefficient > 0.0884) and third-degree (kinship coefficient > 0.0442) related RHG individuals, respectively. The application of these quality-control filters resulted in a final genotyping dataset of 667 individuals and 599,501 SNPs (Figure S1).

Merging Exome and SNP Array Data

Before merging the genotyping array and the exome data from the 667 high-quality individuals in common, we flipped alleles for 8,393 SNPs with incompatible allelic states, and removed 9 SNPs with alleles that remained incompatible after allele flipping from the genotyping dataset. The total concordance rate was evaluated on 28,403 SNPs common to both datasets. The concordance rates for each of the 667 individuals exceeded 98%, confirming an absence of errors during DNA sample processing. The entire genotyping and exome datasets (599,492 and 682,468 SNPs, respectively) were then merged, yielding a final dataset of 1,253,548 SNPs for 667 individuals, 566 of whom were African farmers or hunter-gatherers (Figure S1).

Whole-Genome Sequencing

We generated whole genomes of 20 RHG Baka and 20 AGR Nzébi of Gabon, which were also part of the exome and SNP array datasets. All the samples were processed using the paired-end library preparation protocol from Illumina. Libraries were sequenced on Illumina HiSeq 2000 machines at the Stanford Center for Genomics and Personalized Medicine. 101-pb reads were aligned to the human reference genome (GRCh37) using BWA [85], followed by base quality recalibration and realignment around known indels with GATK [86]. Genotyping was carried out across all 40 individuals jointly using GATK 'UnifiedGenotyper', and called variants were stratified into variant quality tranches using 'VariantQualityScoreRecalibration' tool (VQSR) from GATK. SNPs with a VQSR tranche > 99.0 were considered as confidently called. Genotype calls were refined and improved based on LD using BEAGLE [91], yielding a final dataset of 17,687,206 variants (Figure S1). All individuals presented very low rates of missing values ranging from 0.5% to 4%, and a mean depth of coverage of 6.5 × (ranging from 4 × to 13 ×).

Imputation of SNP Array and Exome Data

Before imputation, we phased the data with SHAPEIT2 using 100 states, 20 MCMC main steps, 7 burnin and 8 pruning steps [87]. SNPs and allelic states were then aligned with the 1000 Genomes Project imputation reference panel (Phase 3 [27]), referred to as 'reference panel 1', as well as the 40 whole genomes of Baka RHG and Nzébi AGR of Gabon, referred to as 'reference panel 2' (Figure S1). We removed from the reference panels SNPs with MAF < 1%, SNPs with C/G or A/T alleles and 414,679 multiallelic SNPs in the reference panel 1. We evaluated the allelic concordance between the two reference panels and excluded 9,649 additional sites from the reference panel 2, yielding to final datasets of 11,501,018 SNPs in the reference panel 1 and 14,252,666 SNPs in the reference panel 2.

Genotype imputation was performed with IMPUTE v.2 [88] considering 1-Mb windows and both reference panels simultaneously, with the '-merge_ref_panels' option. We used genotype calls instead of genotype probabilities, which are not handled by downstream programs, and considered as confident genotype calls genotypes with posterior probability > 0.8. Of the 13,092,258 SNPs obtained after imputation, we removed SNPs that: (i) presented an information metric < 0.8, (ii) had a duplicate, (iii) presented a call rate < 95%, and (iv) were monomorphic. The final imputed dataset included 10,262,236 SNPs, and 9,129,103 after filtering

SNPs with MAF < 1%. To evaluate imputation accuracy, we estimated correlation coefficients r^2 between true genotypes (i.e., obtained by Illumina genotyping array or exome sequencing) and imputed genotypes for the same SNPs (i.e., obtained by artificially removing genotyped SNPs from the data before imputation and then imputing them). The average correlation coefficient across all genotyped SNPs with information metric > 0.8 were 0.86 and 0.85 for reference panels 1 and 2, respectively, showing that our quality filters ensure to keep accurately imputed SNPs for further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome Scans for Selective Sweeps

Genomic regions candidate for positive selection were detected in seven populations of RHG (Bezan, Baka, BaBongo of central Gabon, BaKoya, BaBongo of south and east Gabon and BaTwa) and two populations of AGR (western and eastern AGR), with an outlier approach that considers two interpopulation statistics: PBS (Population Branch Score [28]), and XP-EHH [29]. We combined these scores into a Fisher's score (F_{CS}) equal to the sum, over the two statistics, of $-\log_{10}$(rank of the statistic for a given SNP/number of SNPs). Interpopulation statistics require a reference population, and PBS statistics an outgroup population. We performed separate scans of classic sweeps for each population, using Europeans as outgroup, and different reference populations: western AGR for each western RHG population, eastern AGR for eastern RHG, pooled western RHG for western AGR, and eastern RHG for eastern AGR. PBS was calculated for each SNP using AMOVA-based F_{ST} values computed with home-made scripts (available upon request). The derived allele of each SNP was defined based on the 6-EPO alignment. XP-EHH was computed in 100-kb sliding windows with a 50-kb pace, with home-made scripts (available upon request). Only SNPs with a derived allele frequency (DAF) between 10% and 90% were analyzed further. XP-EHH scores were normalized in 40 separate bins of DAF. An outlier SNP was defined as a SNP with an F_{CS} among the 1% highest of the genome. A putatively selected genomic region was defined as a 100-kb window presenting a proportion of outlier SNPs among the 1% highest of all windows, in five bins of SNP numbers. Windows containing less than 50 SNPs were discarded as well as 500-kb regions around gaps, to avoid biases in the outlier enrichment scores.

Polygenic Selection of Complex Traits

We retrieved the results of the Genome Wide Association studies from UK BIOBANK (round 2, http://www.nealelab.is/uk-biobank/) of 12 complex traits that we selected as candidates for adaptation of RHG, based on previous hypotheses from biological anthropology studies [51, 73, 97–101]. Our genomic dataset was split into non-overlapping 100-kb windows. We considered a window as associated with a trait if it included a SNP with a genome-wide significant association with this trait ($P_{assoc}<5 \times 10^{-8}$). We computed for each genomic window, associated or not with the trait, the average F_{CS}, the proportion of conserved SNP positions based on GERP scores > 2 [92], and the recombination rate using the combined HapMap genetic map [102], to account for the confounding effects of background selection.

In order to test for polygenic selection, we generated a null distribution by randomly sampling *x* windows (*x* being the number of windows associated with a tested trait) among windows with a similar number of SNPs, proportion of GERP > 2 sites and recombination rate observed in the trait-associated windows. We then calculated the average of the mean of the F_{CS} across the *x* resampled windows. We resampled 100,000 sets of *x* windows for each trait. To test for significance, we computed a resampling *P*-value by calculating the proportion of resampled windows which mean F_{CS} was higher than that observed for the tested trait. All *P*-values for polygenic adaptation were then adjusted for multiple testing by the Benjamini-Hochberg method, to account for the number of traits tested, and traits with an adjusted $p < 0.05$ were considered as candidates for polygenic selection.

To test if polygenic selection signals are due to pleiotropy of height-associated genes, we first estimated genetic correlations between candidate traits from LD-score regression using the ldsc tool [89]. We used precomputed European LD-scores (https://data.broadinstitute.org/alkesgroup/LDSCORE/). *P*-values were corrected for multiple testing using the Bonferroni correction, and adjusted *P*-values < 0.05 were considered as significant.

To correct for pleiotropy for each trait genetically correlated with height, we removed windows significantly associated with 'Standing Height' and 'Comparative height at age 10' in both windows associated with the candidate trait and resampled windows. Similarly, we re-tested for polygenic adaptation on ''Standing height'' and ''Comparative height at age 10'' associated regions using the same approach, but by removing all trait-associated windows, except height-associated windows. To test if loss of significance was due to a decrease in power, we down-sampled the number of tested trait-associated windows to the same number as after removing height-associated windows. We down-sampled a 100 times trait-associated windows, and estimated a hundred *P*-values as described above. We finally compared the distribution of the 100 obtained *P*-values with the estimated *P*-value (non-adjusted for multiple testing) both before and after removing height-associated windows.

Polygenic Selection of Gene Ontologies

To detect enrichment of F_{CS} scores in sets of genes corresponding to a given biological pathway, we compared the distributions of F_{CS} between genes that were part of the gene ontology (GO) term tested, relative to the rest of the genes of the genome, using a Mann-Whitney-Wilcoxon rank-sum test. To limit the effect of clusters of genes on the enrichment calculation, we assigned to each 100-kb non-overlapping genomic window both a GO term, based on the presence of at least one gene from the corresponding term, and a mean F_{CS} score. We tested if mean F_{CS} of windows assigned to a given GO term were different from genome-wide expectations, accounting for multiple testing. We restricted the enrichment analysis to 5,354 GO terms with levels comprised between

levels 3 and 7 [53], using the python library goatools [90], and that include at least 5 genes. We examined a total of 15,503 windows and determined *P*-values corresponding to 5% and 1% of false discoveries, FDR p = 9.24×10^{-3} and FDR p = 4.03×10^{-4}, respectively, by randomly resampling *y* genes (*y* being sampled from the distribution of the number of genes assigned to each GO term). We also studied additional gene sets, including 1,553 manually-curated genes involved in innate immunity [56] and 1,257 genes encoding proteins known to have physical interactions with multiple families of viruses [57].

Local Ancestry Inference

To perform local ancestry inference in the genomes of the highly-admixed BaBongo RHG from south and east Gabon, we first constituted putative parental populations that were representative of RHG and AGR ancestry. We considered as the parental AGR population, 163 individuals with less than 20% of their ancestry assigned to the RHG component, based on the ADMIXTURE analysis at $K = 5$. Likewise, we considered as the parental RHG population, 101 individuals with less than 5% AGR ancestry. The genomes of the highly-admixed BaBongo were decomposed into segments of RHG or AGR ancestry with RFMix v.1.5.4 [58], including two EM steps. We excluded 2-Mb regions from the telomeres of each chromosome. Based on RFMix ancestry estimations, the mean AGR ancestry was 94% [SD = 1.6%] in the parental AGR population, 62% [SD = 5.9%] in the highly-admixed BaBongo, and 27% [SD = 3.7%] in the parental RHG population. These ancestry proportions were highly correlated with ADMIXTURE membership proportions at $K = 2$ (Pearson's correlation coefficient $R^2 = 0.99$). We then searched for excesses in RHG or AGR ancestry in pathways by assigning ancestry proportions to 100-kb windows across the genome, with the same approach used for GO enrichments.

DATA AND CODE AVAILABILITY

The newly generated exomes (n = 277) and genomes (n = 40) of central African rainforest hunter-gatherers and agriculturalists have been deposited in the European Genome-phenome Archive (EGA). The accession number for the newly generated data reported in this paper is EGA: EGAS00001003722. Data accessibility is restricted to academic research on human genetic history and adaptation. Exome sequencing data for the remaining, previously published samples are available under accession codes EGA: EGAS00001002457 and EGA: EGAS00001001895.

Current Biology, Volume *29*

Supplemental Information

Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest

Marie Lopez, Jeremy Choin, Martin Sikora, Katherine Siddle, Christine Harmant, Helio A. Costa, Martin Silvert, Patrick Mouguiama-Daouda, Jean-Marie Hombert, Alain Froment, Sylvie Le Bomin, George H. Perry, Luis B. Barreiro, Carlos D. Bustamante, Paul Verdu, Etienne Patin, and Lluís Quintana-Murci

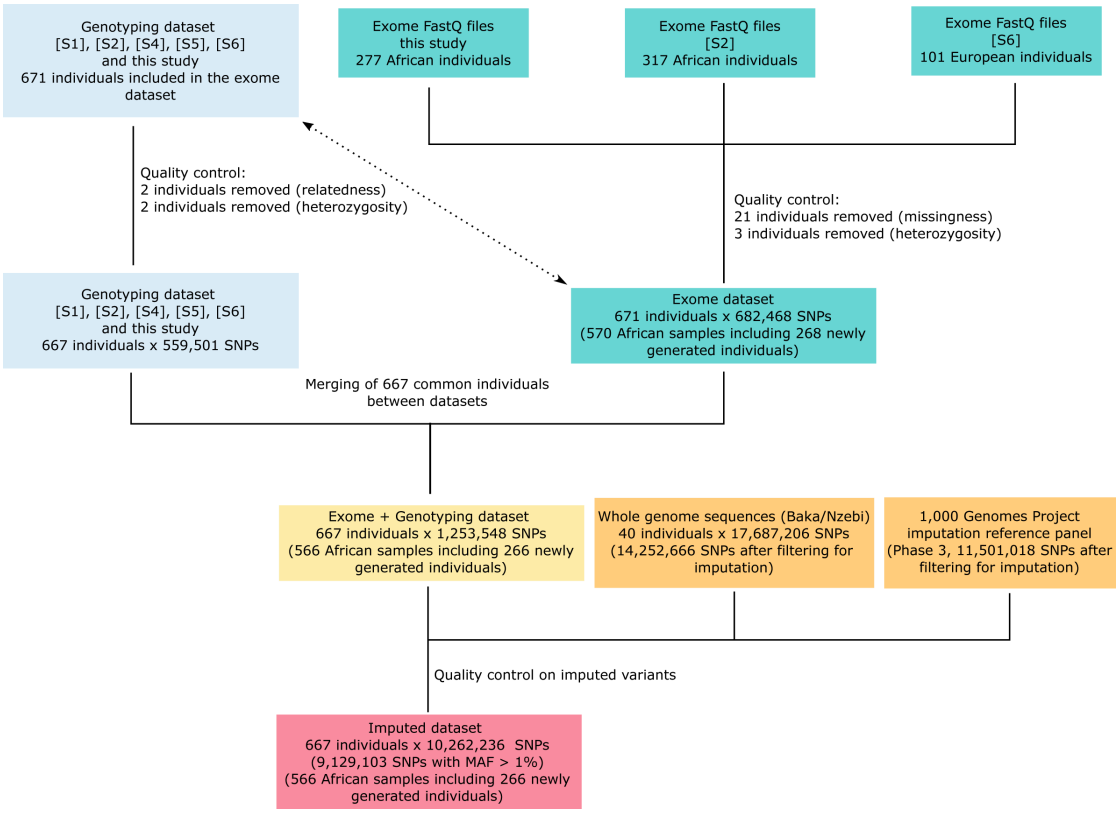


Figure S1. Summary of the Data Processing Performed in this Study. Related Figure 1. The arrow indicates the correspondence between individuals analyzed in both datasets.

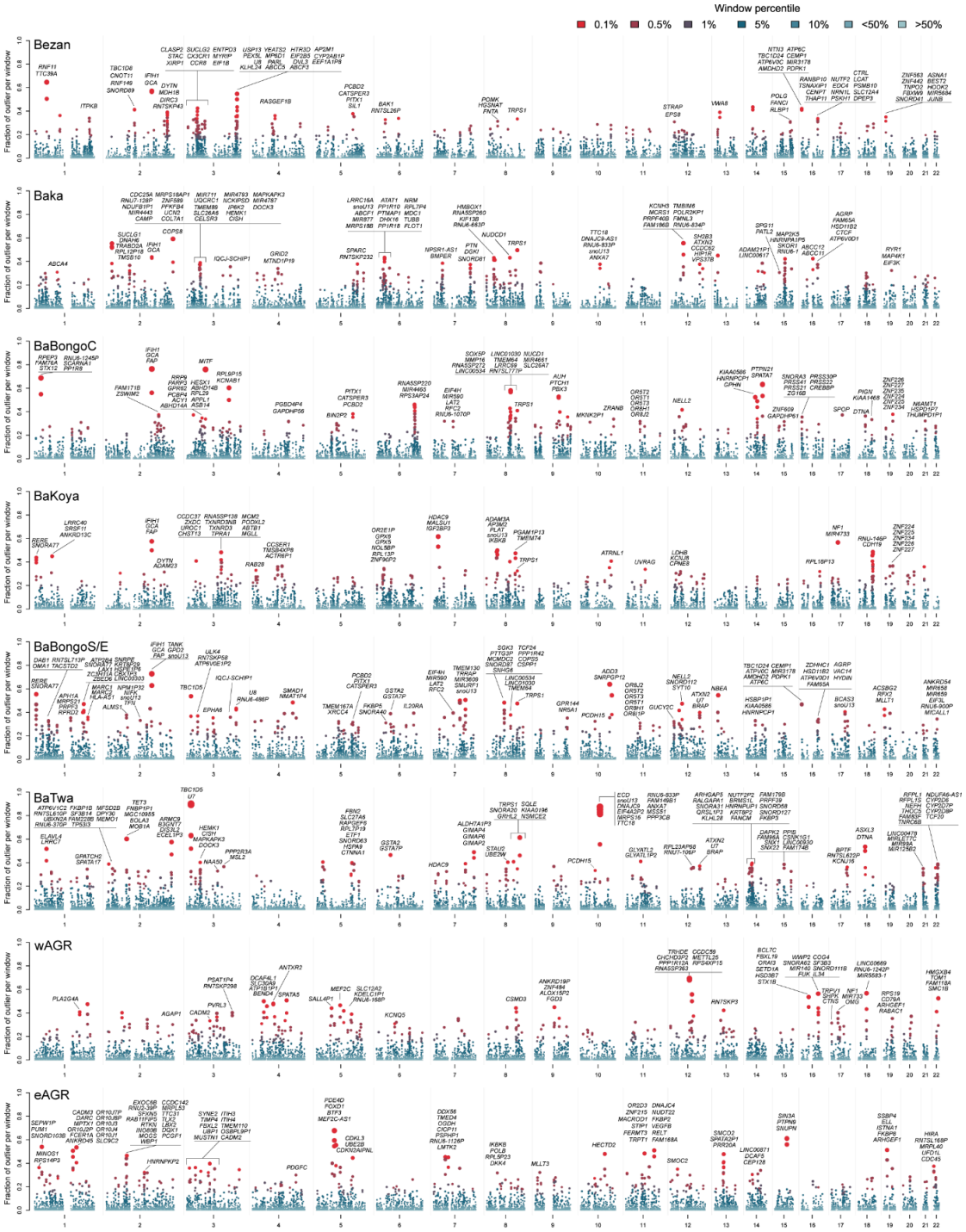


Figure S2. Genome-Wide Signals of Classic Sweeps in Central African Populations. Related to Figure 2.

Proportions of outlier SNPs (i.e., F_{CS} in the top 1% of the empirical distribution) in 100-kb windows along the genome of RHG and AGR populations. Gene names are shown for candidate windows with a proportion of F_{CS} outlier SNPs > 30%.

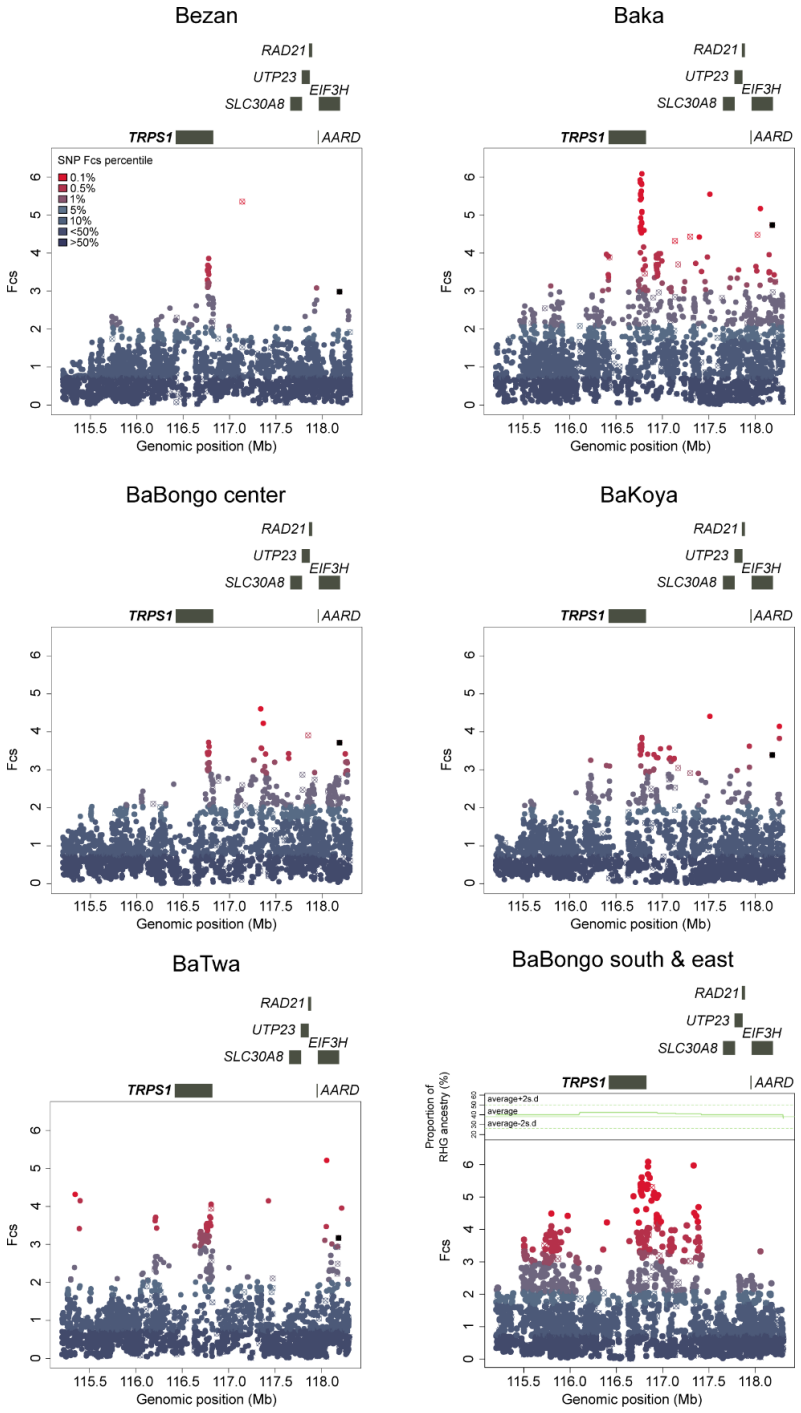


Figure S3. Selective Sweep Signal at the *TRPS1* Locus in African Rainforest Hunter-Gatherers. Related to Figure 2C.

Local genomic signals of classic sweeps at the candidate windows containing the *TRPS1* gene (chr8:116702422-116802422) in all RHG populations. Dot colors indicate SNP F_{CS} percentiles, black squares indicate non-synonymous mutations and circled crosses indicate non-imputed SNPs. Average local RHG ancestry is shown for the admixed BaBongo of south and east Gabon.

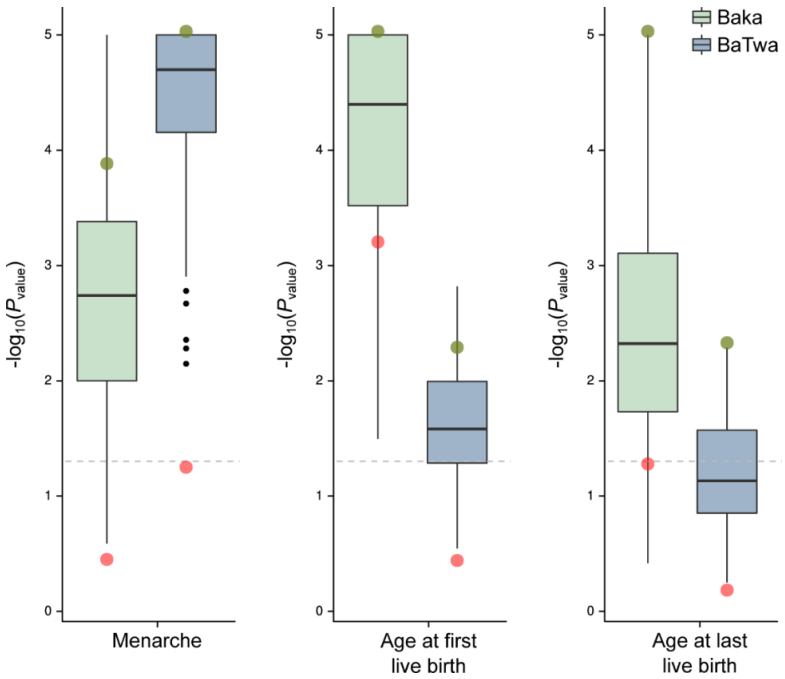


Figure S4. Significance of Tests for Polygenic Selection when Accounting for Pleiotropy or Reduced Number of Genomic Windows. Related to Figure 3A-C.
Red dots indicate $-\log_{10}$(non-adjusted *P*) *when accounting* for pleiotropy (i.e., after excluding height-associated windows). Green dots indicate $-\log_{10}$(non-adjusted *P*) *when not accounting* for pleiotropy. Boxplots correspond to $-\log_{10}$(non-adjusted *P*) of 100 random samples of *x* trait-associated windows, where *x* is the number of windows associated to the trait tested, *when accounting* for pleiotropy. The grey dashed line indicates the significance threshold $-\log_{10}(0.05)$. When red points are below both the dashed line and box plots, this indicates that the significant signals of polygenic selection are no longer significant, because of our correction for pleiotropy, and not because of the reduced number of windows.

Group	Population	Country	N	Reference for SNP array data (accession number)	Reference for exome sequencing (accession number)	Mean AGR ancestry	SD of AGR ancestry	Minimum AGR ancestry	Maximum AGR ancestry
wAGR	Tsogo	Gabon	29	[S1] (EGAS00001002078)	This study	80.7%	1.8%	77.4%	84.3%
wAGR	Galoa	Gabon	30	[S1] (EGAS00001002078)	This study	86.80%	3.2%	78.6%	92.5%
wAGR	Shake	Gabon	30	[S1] (EGAS00001002078)	This study	67%	2.90%	59.6%	71.3%
wAGR	Fang	Gabon	31	[S1] (EGAS00001002078)	This study	85.3%	1.1%	82.3%	87.8%
wAGR	Bapunu	Gabon	44	[S1] (EGAS00001002078)	[S2] (EGAS00001002457)	82.5%	3.6%	66.4%	88.2%
wAGR	Nzébi	Gabon	55	[S1] (EGAS00001002078)	[S2] (EGAS00001002457)	82.7%	3.1%	73%	87.7%
eAGR	BaKiga	Uganda	49	[S3] (EGAS00001000908)	[S2] (EGAS00001002457)	88.6%	1.9%	84.5%	92.8%
wRHG	Bezan	Cameroon	38	[S3] (EGAS00001000605)	This study	9.5%	13.3%	0%	45.5%
wRHG	BaBongo (center)	Gabon	21	This study	This study	9.4%	9.1%	0%	39.8%
wRHG	BaBongo (east)	Gabon	27	[S4] (EGAS00001000605)	This study	43.3%	11.2%	31.3%	82.8%
wRHG	BaBongo (south)	Gabon	33	[S4] (EGAS00001000605)	This study	24.3%	17.4%	0%	59.1%
wRHG	BaKoya	Gabon	26	[S1] (EGAS00001002078)	This study	4.1%	5.5%	0%	20.8%
wRHG	Baka	Cameroon/ Gabon	72/ 30	[S5] (EGAS00001001066) [S4] (EGAS00001000605)	[S2] (EGAS00001002457) This study	8.1%	10.7%	0%	51.4%
eRHG	BaTwa	Uganda	51	[S3] (EGAS00001000908)	[S2] (EGAS00001002457)	8.7%	12.2%	0%	43.5%
EUR	Belgian	Belgium	101	[S6] (EGAS00001001895)	[S6] (EGAS00001001895)	NA	NA	NA	NA

Table S1. Population description, sample size, and AGR ancestry proportions of the final dataset of 667 individuals. Related to Figure 1.

Ancestry proportions were estimated in AGR and RHG populations with ADMIXTURE at *K*=5 clusters.

Immune traits	Bezan	Baka	BaKoya	BaBongoC	BaBongoS/E	BaTwa	wAGR	eAGR
All II genes	2.13×10^{-2}	6.44×10^{-2}	0.324	$\mathbf{2.95\times10^{-3}}$	3.51×10^{-2}	0.674	3.33×10^{-2}	0.210
Adaptors	0.224	$\mathbf{2.54\times10^{-3}}$	0.034	9.25×10^{-3}	0.359	0.574	0.752	0.310
Regulators	0.140	0.497	0.166	1.84×10^{-2}	$\mathbf{3.36\times10^{-3}}$	$\mathbf{4.81\times10^{-3}}$	0.675	0.731
Secondary receptors	$\mathbf{2.26\times10^{-3}}$	$\mathbf{1.42\times10^{-4}}$	0.682	5.05×10^{-2}	0.144	0.697	0.985	0.487
Signal transducers	0.382	0.074	0.623	$\mathbf{1.55\times10^{-3}}$	2.53×10^{-2}	0.925	0.399	0.387
Sensors	$\mathbf{3.89\times10^{-4}}$	0.424	0.338	$\mathbf{5.99\times10^{-3}}$	0.507	0.157	$\mathbf{8.57\times10^{-3}}$	0.350
Transcription factors	0.536	0.931	0.930	0.450	0.496	0.362	1.67×10^{-2}	0.758
Accessory molecules	0.904	0.777	0.946	0.894	0.548	0.991	0.063	0.287
Effectors	0.250	0.302	0.140	0.410	0.699	0.418	0.387	0.395
Uncharacterized	0.413	0.154	0.047	0.939	0.678	0.906	0.134	3.09×10^{-2}
All VIP genes	0.179	0.199	0.083	$\mathbf{6.85\times10^{-4}}$	1.17×10^{-2}	0.236	0.193	0.323
dsDNA	0.268	$\mathbf{8.88\times10^{-3}}$	$\mathbf{3.93\times10^{-3}}$	2.46×10^{-2}	4.22×10^{-2}	0.073	0.648	0.787
ssRNA	0.182	0.339	0.084	$\mathbf{4.21\times10^{-4}}$	$\mathbf{7.98\times10^{-3}}$	0.562	0.336	0.316
ssDNA	0.919	0.093	$\mathbf{2.31\times10^{-3}}$	1.29×10^{-2}	0.429	0.619	0.592	0.736
dsDNART	0.365	0.985	0.710	0.382	0.152	0.459	0.111	0.569
ssRNART	0.213	0.496	0.864	0.162	0.258	0.069	0.118	0.135

Table S2. Polygenic Selection Signals for Immune-Related Traits in Central Africans. Related to Figure 3D.

Evidence for polygenic selection across 1,553 innate immunity (II) and 1,257 viral interacting protein (VIP) genes, based on their enrichment in high F_{CS} selection scores (FDR P<5%; in bold), relative to genome-wide expectations. Families of viral interacting proteins include host genes interacting with: double-stranded DNA virus (dsDNA), double-stranded DNA retrovirus (dsDNART), single-stranded DNA virus (ssDNA), single-stranded RNA virus (ssRNA) and single-stranded RNA retrovirus (ssRNART).

Supplemental References

S1. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543-546.

S2. Lopez, M., Kousathanas, A., Quach, H., Harmant, C., Mouguiama-Daouda, P., Hombert, J.M., Froment, A., Perry, G.H., Barreiro, L.B., Verdu, P., et al. (2018). The demographic history and mutational load of African hunter-gatherers and farmers. Nat Ecol Evol *2*, 721-730.

S3. Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nedelec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S.L., et al. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proc Natl Acad Sci U S A *111*, E3596-3603.

S4. Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Regnault, B., Lemee, L., Gravel, S., et al. (2014). The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun *5*, 3163.

S5. Fagny, M., Patin, E., MacIsaac, J.L., Rotival, M., Flutre, T., Jones, M.J., Siddle, K.J., Quach, H., Harmant, C., McEwen, L.M., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat Commun *6*, 10047.

S6. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell *167*, 643-656 e617.