

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, et al.

▶ **To cite this version:**

Jeremy Choin, Javier Mendoza-Revilla, Lara R Arauna, Sebastian Cuadros-Espinoza, Olivier Cassar, et al.. Genomic insights into population history and biological adaptation in Oceania. Nature, 2021, 592 (7855), pp.583-589. 10.1038/s41586-021-03236-5 . pasteur-03205291

HAL Id: pasteur-03205291

https://pasteur.hal.science/pasteur-03205291

Submitted on 11 May 2021

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

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



Distributed under a Creative Commons Attribution 4.0 International License

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

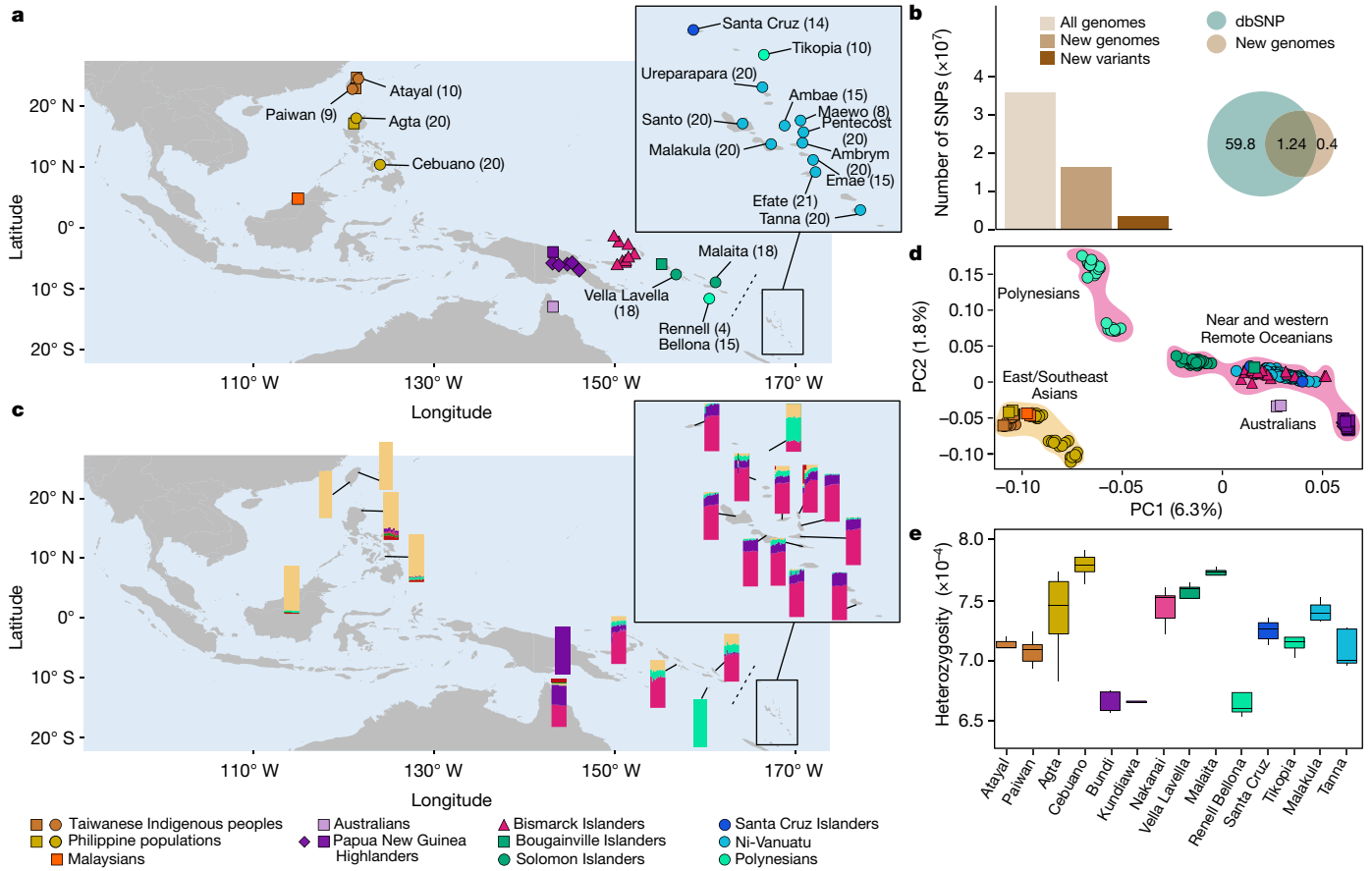


Fig. 1 | Whole-genome variation in Pacific Islanders. a, Location of studied populations. The indented map is a magnification of western Remote Oceania. Circles indicate newly generated genomes. Sample sizes are indicated in parentheses. Squares, triangles and diamonds indicate genomes from Mallick et 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].

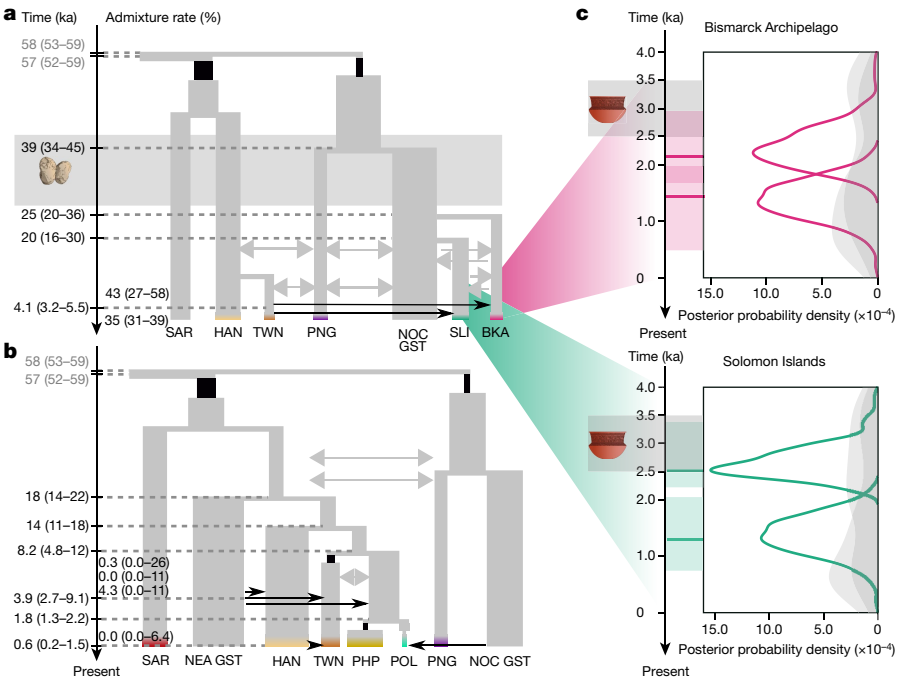


Fig. 2 | Demographic models of the human settlement of the Pacific. **a**, Maximum-likelihood model for Near Oceanian populations. Point estimates of parameters and 95% confidence intervals are reported in Supplementary Table 4. The grey area indicates the archaeological period for the settlement of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

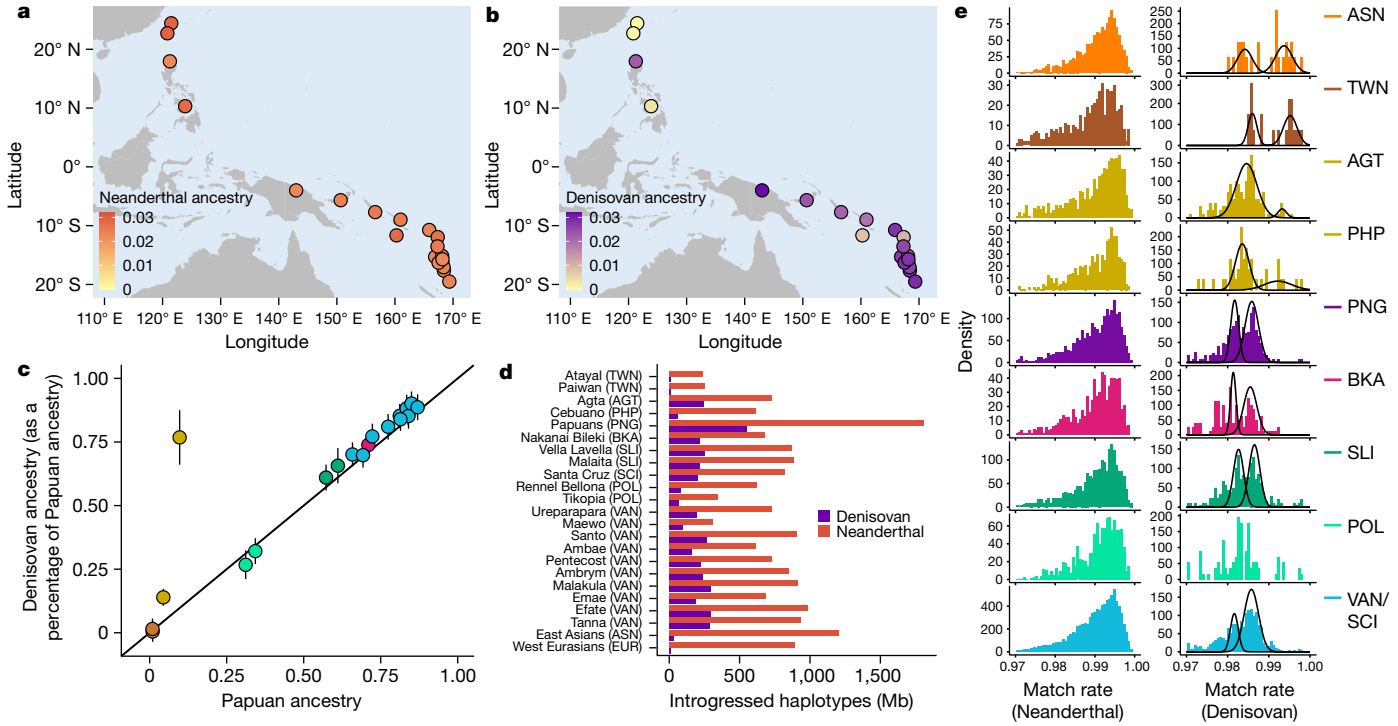


Fig. 3 | Neanderthal and Denisovan introgression across the Pacific. **a**, **b**, Estimates of Neanderthal (**a**) and Denisovan (**b**) ancestry on the basis of f_4-ratio statistics. Maps were generated using the maps R package[51]. **c**, Correlation between Papuan ancestry and Denisovan ancestry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

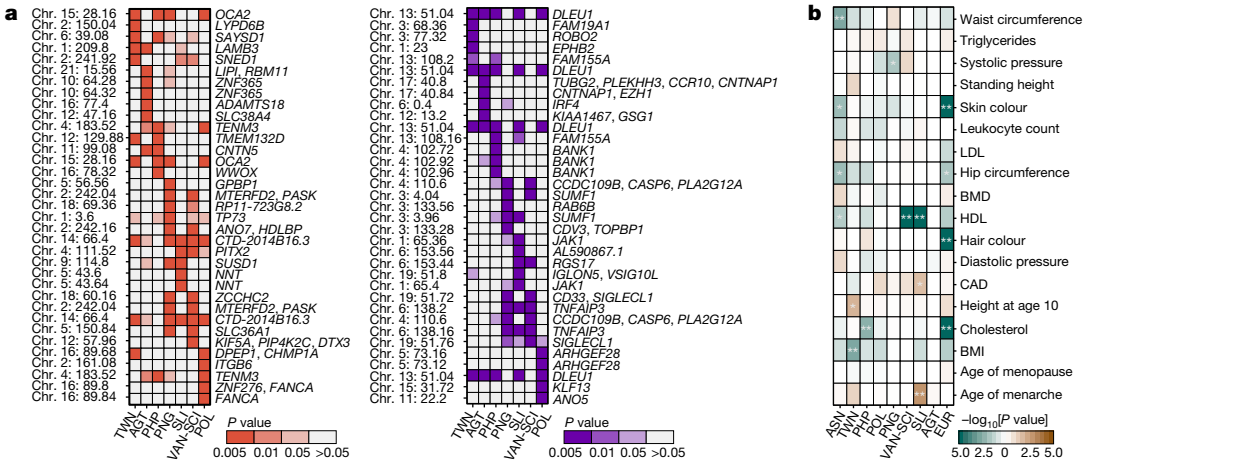


Fig. 4 | Mechanisms of genetic adaptation to Pacific environments. **a**, Genomic regions showing the strongest evidence of adaptive introgression from Neanderthals (red) and Denisovans (purple). Each row is a 40-kb window, each column is a Pacific population group, and each cell is coloured according 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).

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

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.

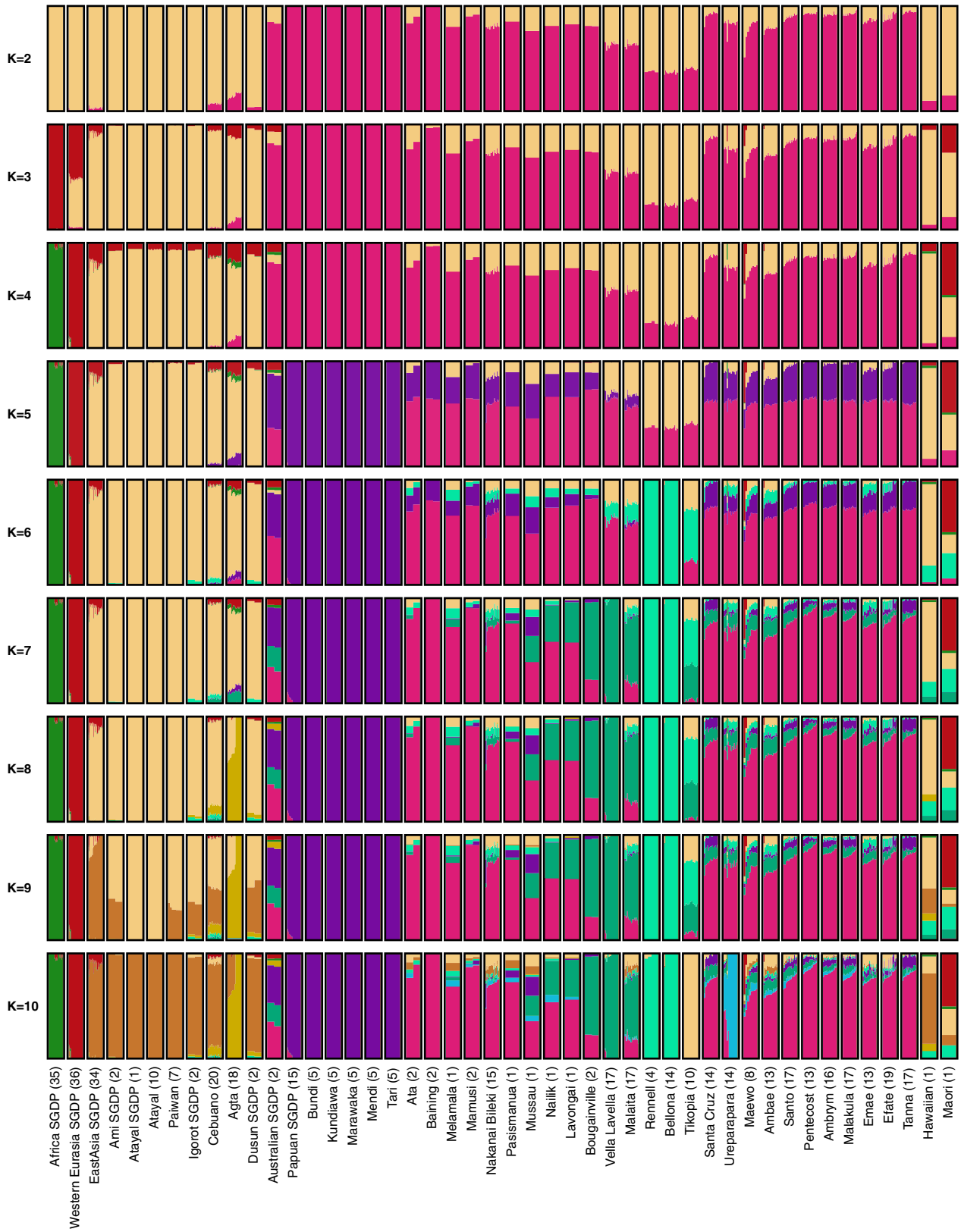
Additional information

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

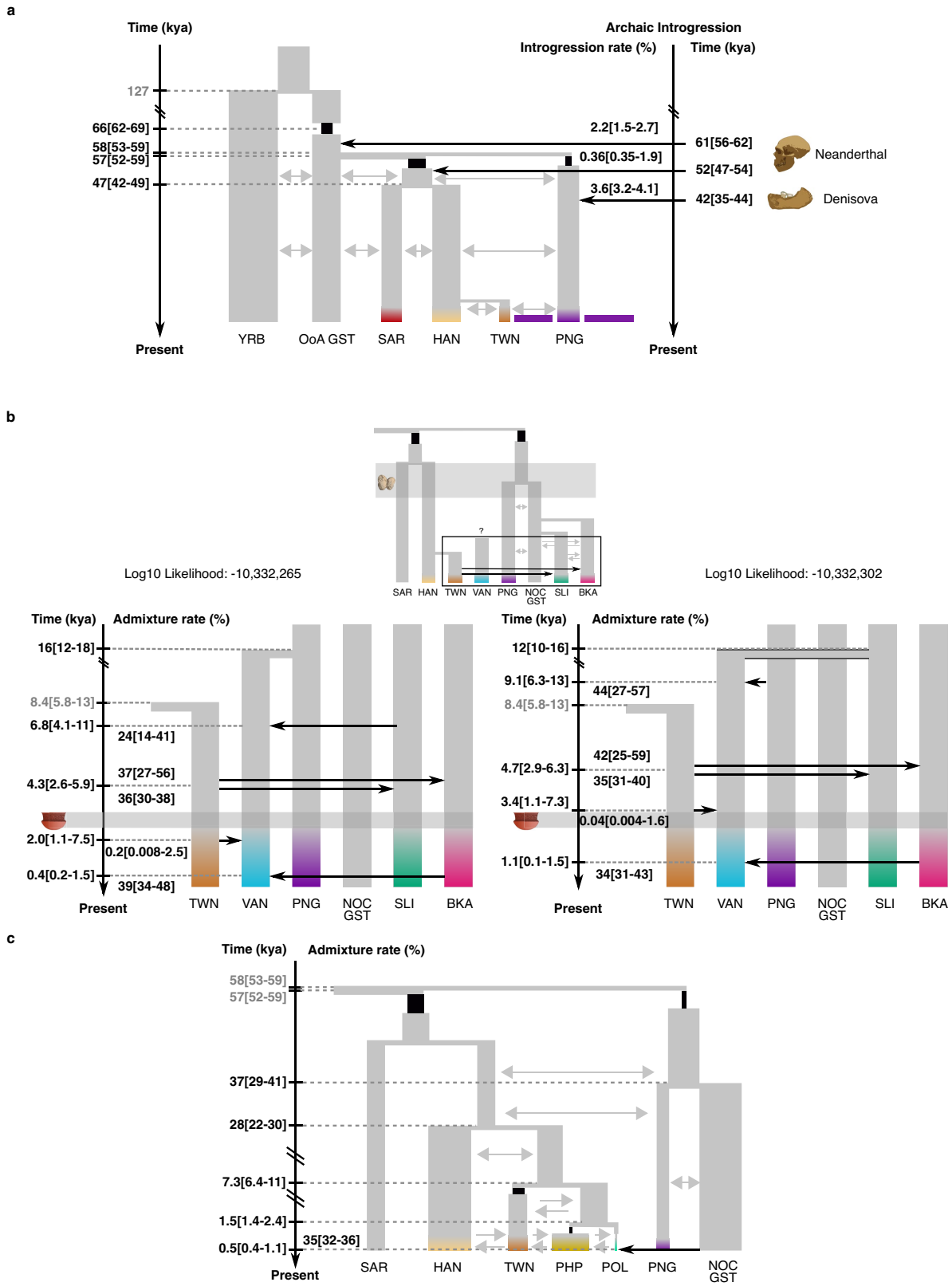
Correspondence and requests for materials should be addressed to E.P. or L.Q.-M.

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.

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

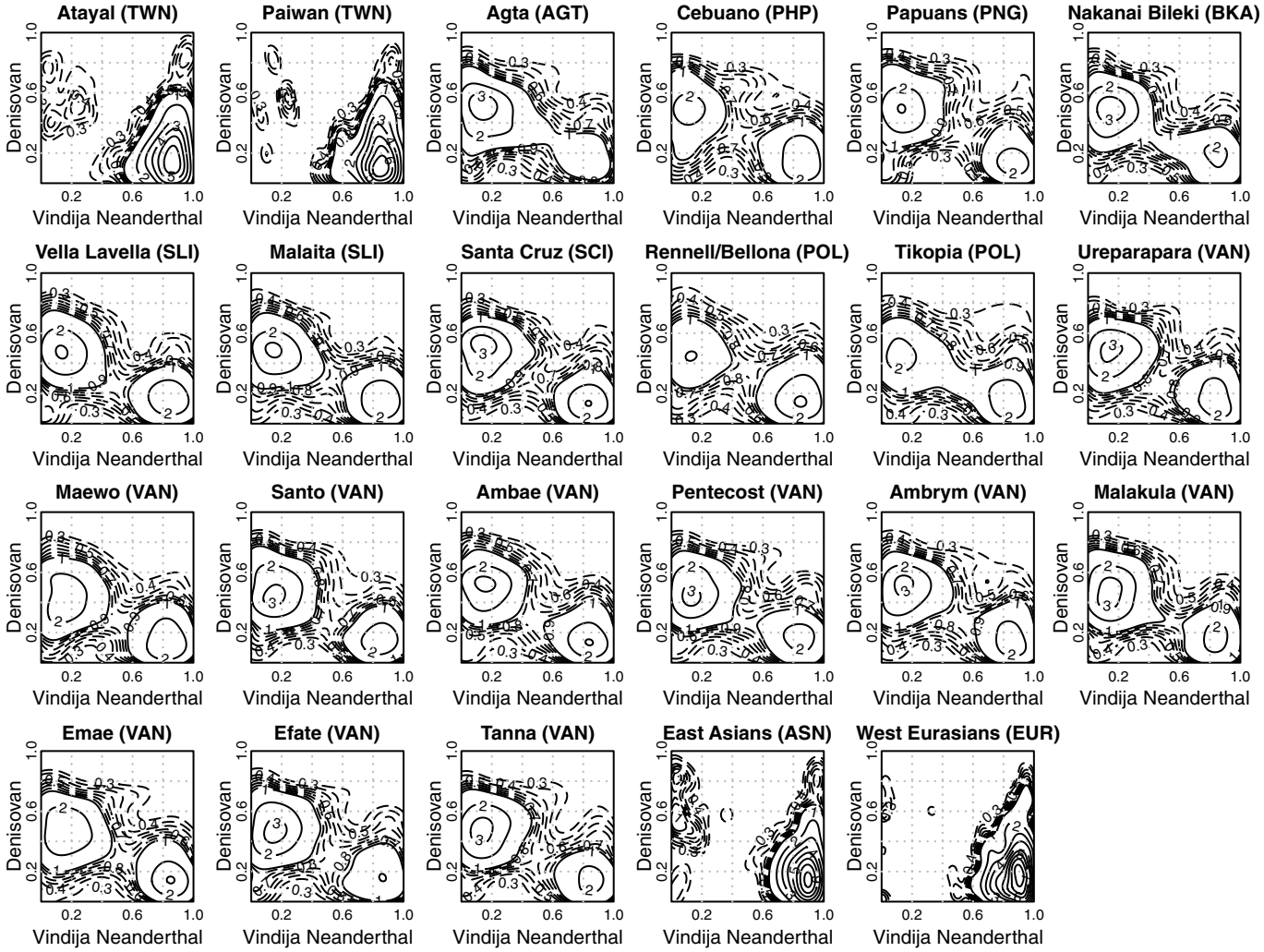


Extended Data Fig. 1 | Genetic structure of Pacific populations. ADMIXTURE ancestry components are shown from $K = 2$ (top) to $K = 10$ (bottom) for the 462 unrelated individuals. The lowest cross-validation error was obtained at $K = 6$ (Supplementary Fig. 5). Populations are delimited by black borders.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

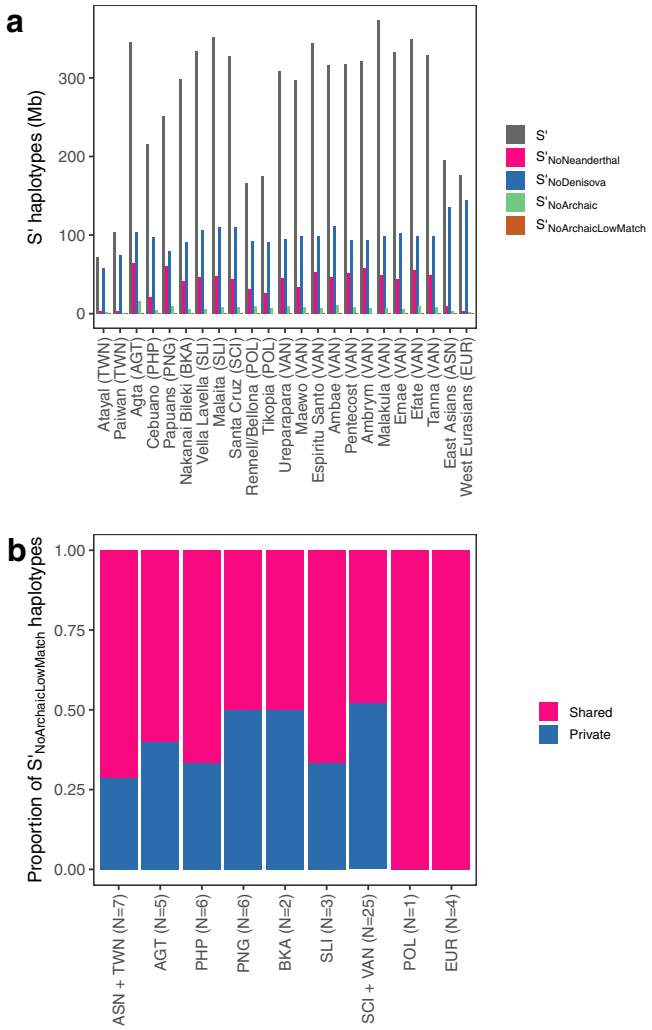


Extended Data Fig. 2 | See next page for caption.

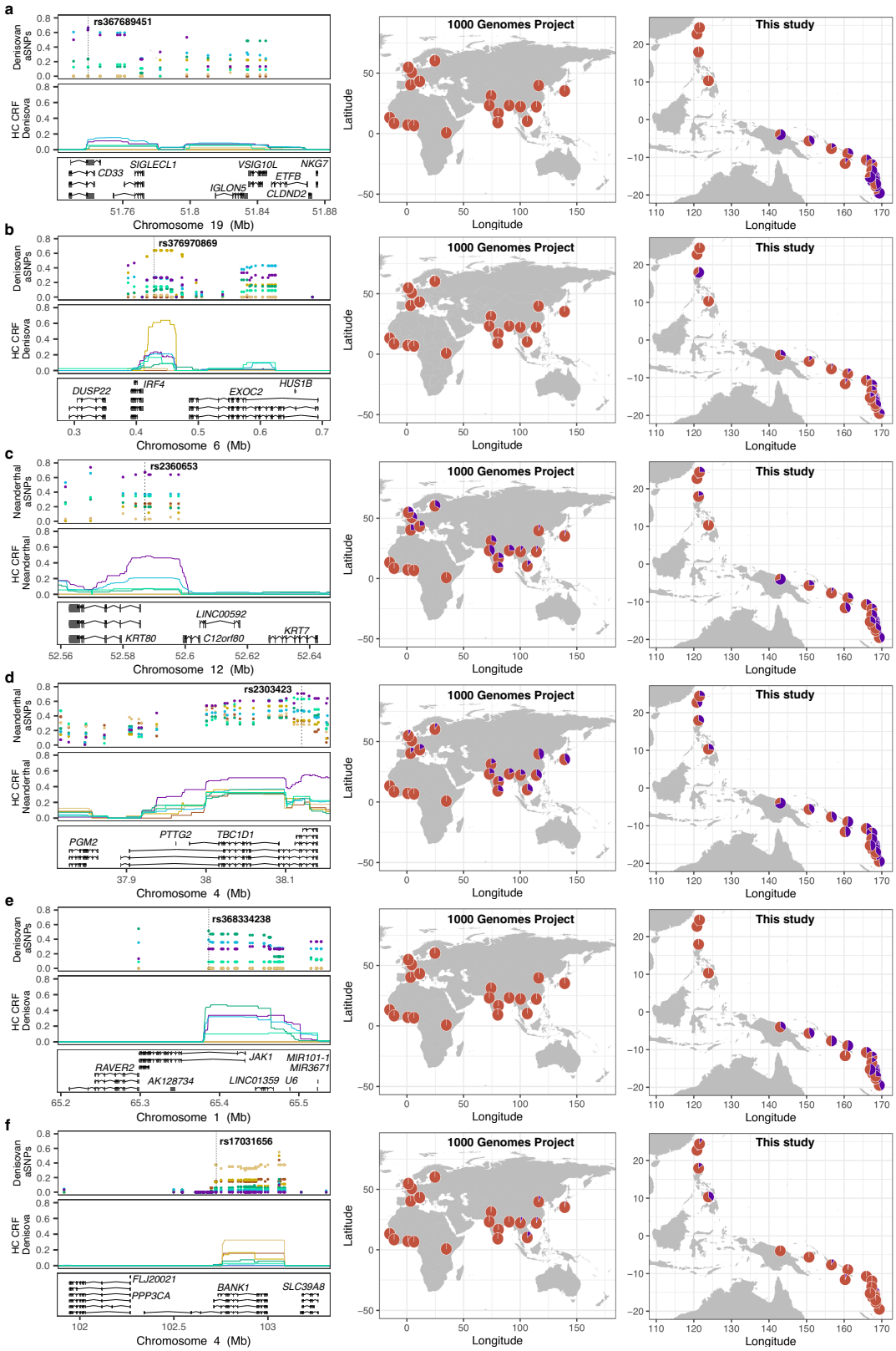
Extended Data Fig. 2 | Demographic models for Pacific populations. **a**, Maximum-likelihood demographic model for baseline populations. Point estimates of parameters and 95% confidence intervals are shown in Supplementary Table 2. **b**, Maximum-likelihood demographic models for western Remote Oceanian individuals (VAN). The likelihoods of the two models are not considered to be different. Point estimates of parameters and 95% confidence intervals are shown in Supplementary Table 5. The (VAN, PNG) model (left) assumes that the ni-Vanuatu diverged from Papua New Guinean Highlanders and then received gene flow from Solomon Islanders, Bismarck Islanders and Austronesian-speaking Taiwanese Indigenous peoples. The (VAN, SLI) (right) model assumes that the ni-Vanuatu diverged from the Solomon Islanders and then received gene flow from the other three groups. For the sake of clarity, only Taiwanese Indigenous, Near Oceanian and western Remote Oceanian populations are shown. **c**, Maximum-likelihood model for Austronesian-speaking populations, represented by Taiwanese Indigenous, Philippine Kankanaey and Tikopia Polynesian individuals. BKA, Bismarck Islanders; HAN, Han Chinese individuals (China); NOC GST, a meta-population of Near Oceanian individuals; OoA GST, an unsampled population to represent the Out-of-Africa exodus; PHP, Philippine individuals; PNG, Papua New Guinean Highlanders; POL, Polynesian individuals from the Solomon Islands; SAR, Sardinian individuals (Italy); SLI, Solomon Islanders; TWN, Taiwanese Indigenous peoples; VAN, ni-Vanuatu; YRB, Yoruba individuals (Nigeria). We assumed a mutation rate of 1.25×10^{-8} mutations per generation per site and a generation time of 29 years. Single-pulse introgression rates are reported as a percentage. The 95% confidence intervals are shown in square brackets. The larger the rectangle width, the larger the estimated effective population size (N_e), except for **b**. Bottlenecks are indicated by black rectangles. Grey and black arrows represent continuous and single pulse gene flow, respectively. One- and two-directional arrows indicate asymmetric and symmetric gene flow, respectively. We limited the number of parameter estimations by making simplifying assumptions regarding the recent demography of East-Asian-related and Near Oceanian populations in **a** and **c**, respectively (Supplementary Note 4). Sample sizes are described in Supplementary Note 4.



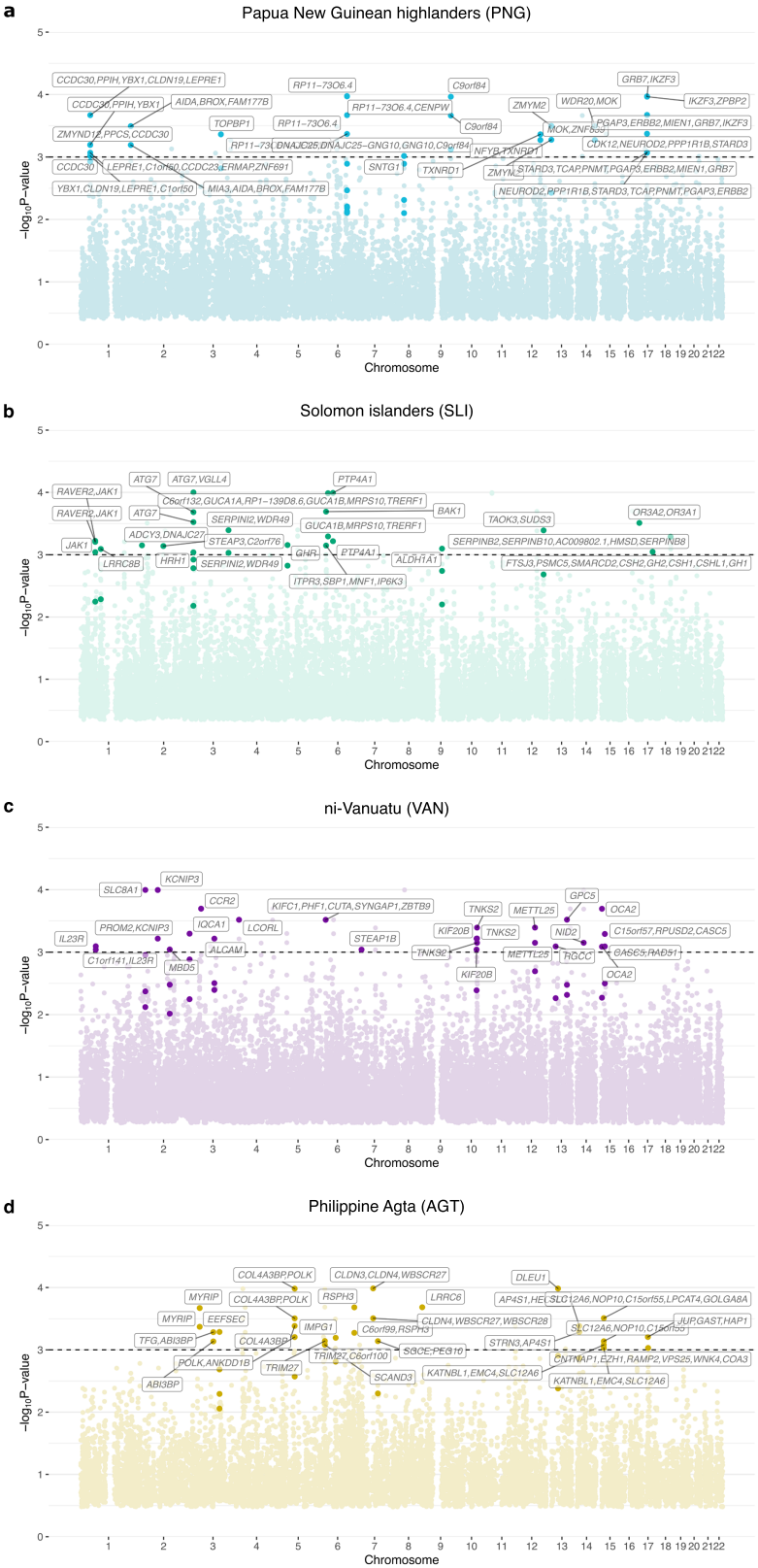
Extended Data Fig. 3 | Match rate of introgressed S′ haplotypes in Pacific populations to the Vindija Neanderthal and Altai Denisovan genomes. The match rate is the proportion of putative archaic alleles matching a given archaic genome, excluding sites at masked positions. Only S′ haplotypes with more than 40 sites outside archaic genome masks were included in the analysis. The numbers indicate the height of the density corresponding to each contour line. Contour lines are shown for multiples of 1 (solid lines) and multiples of 0.1 between 0.3 and 0.9 (dashed lines).



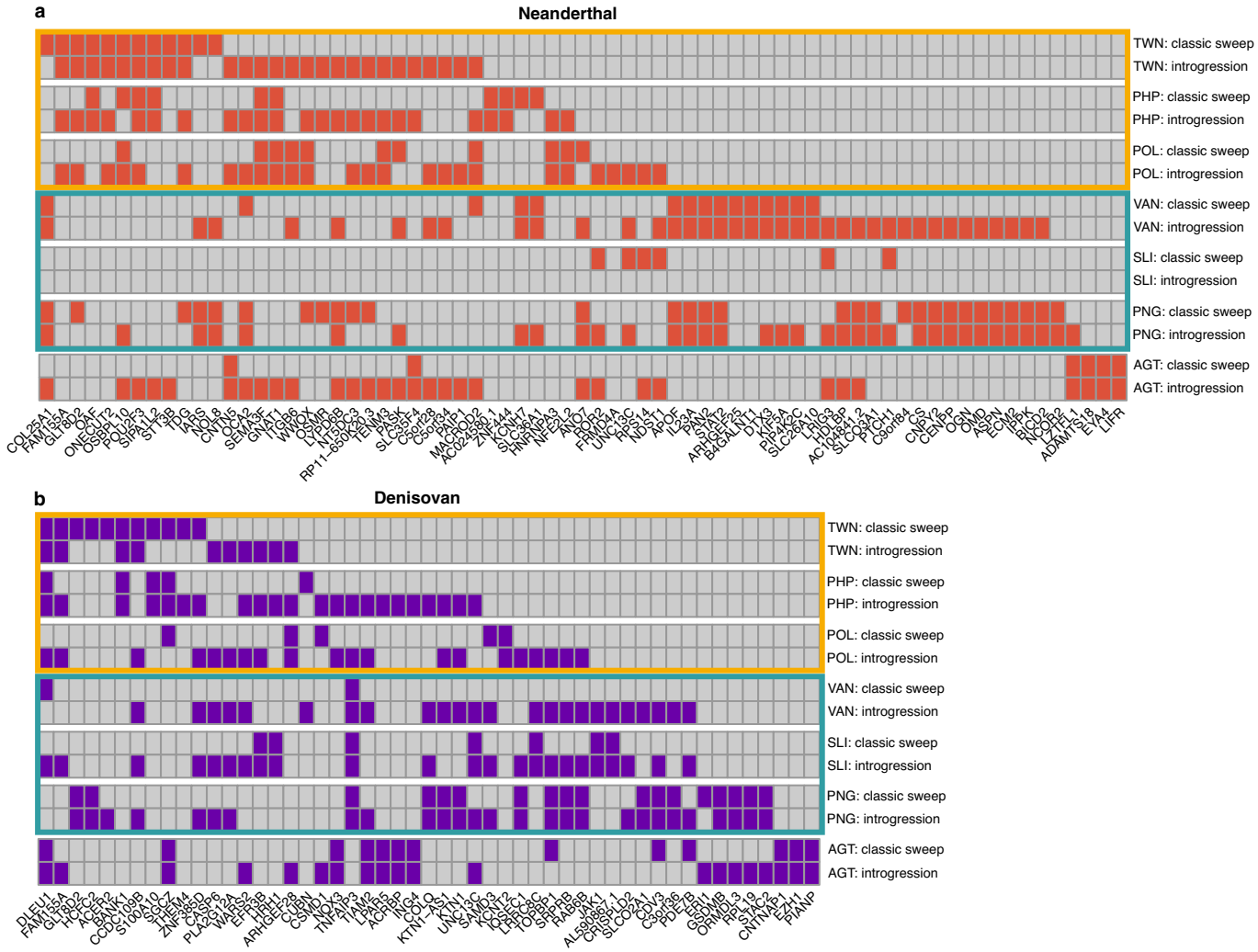
Extended Data Fig. 4 | Detection of introgressed haplotypes from an unknown archaic hominin. a, Cumulative length of S′ haplotypes retrieved among modern human populations (S′), after removing Neanderthal CRF haplotypes ($S'_{NoNeanderthal}$) or Denisovan CRF haplotypes ($S'_{NoDenisova}$) or both ($S'_{NoArchaic}$), 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.



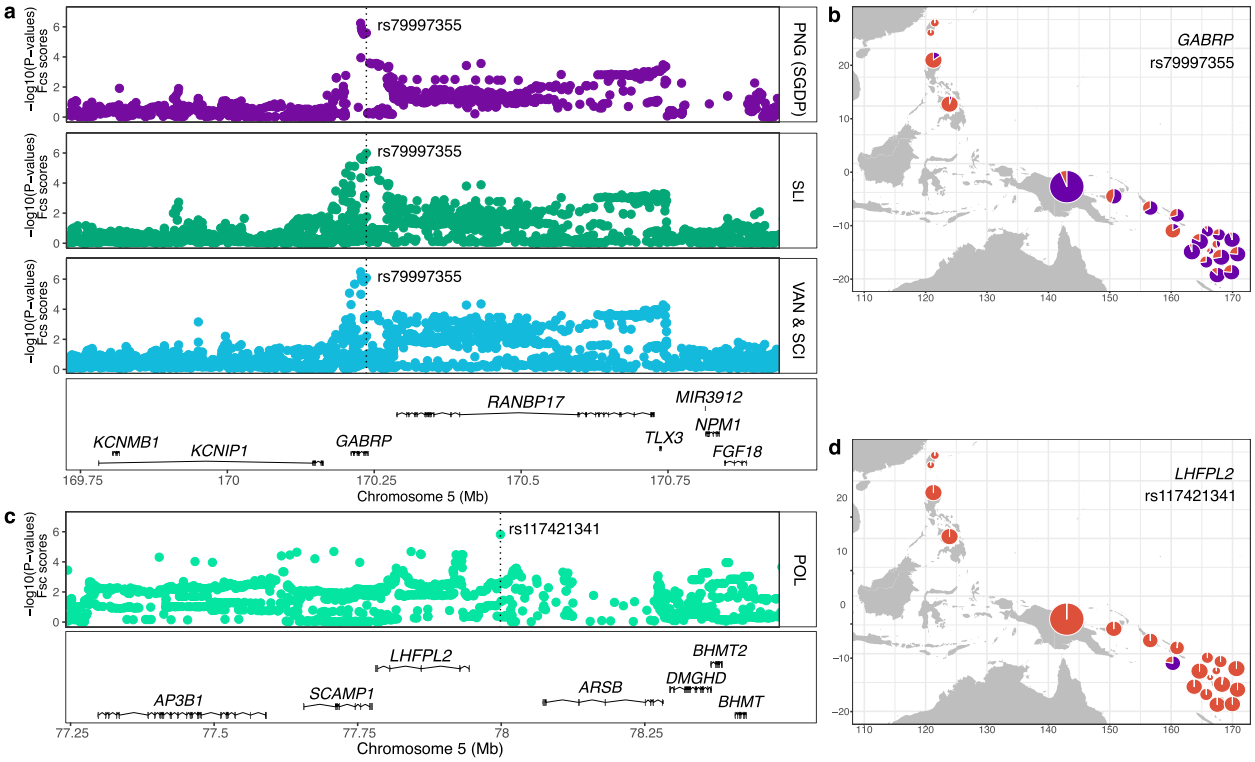
Extended Data Fig. 5 | Examples of candidate loci for adaptive introgression in Pacific populations. **a**, Adaptive introgression of Denisovan origin at the *CD33* locus. **b**, Adaptive introgression of Denisovan origin at the *IRF4* locus. **c**, Adaptive introgression of Neanderthal origin at the *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].



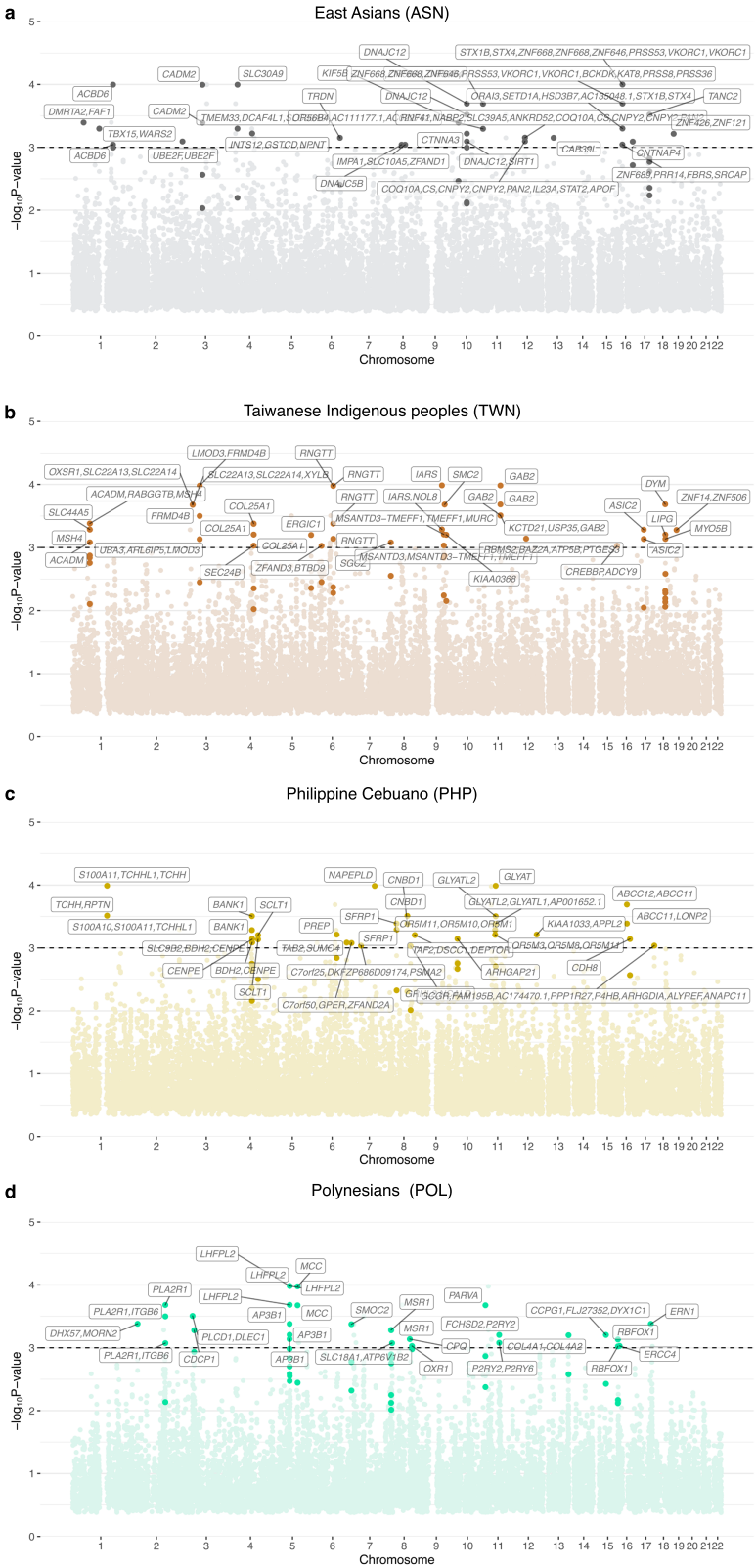
Extended Data Fig. 6 | Classic sweep signals detected in Papuan-related populations. a–**d**, Manhattan plots of classic sweep signals in Papua New Guinean Highlanders (**a**), Solomon Islanders (**b**), ni-Vanuatu (**c**) and Philippine Agta (**d**). **a**–**d**, The *y* axis shows the $-\log_{10}(P\text{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



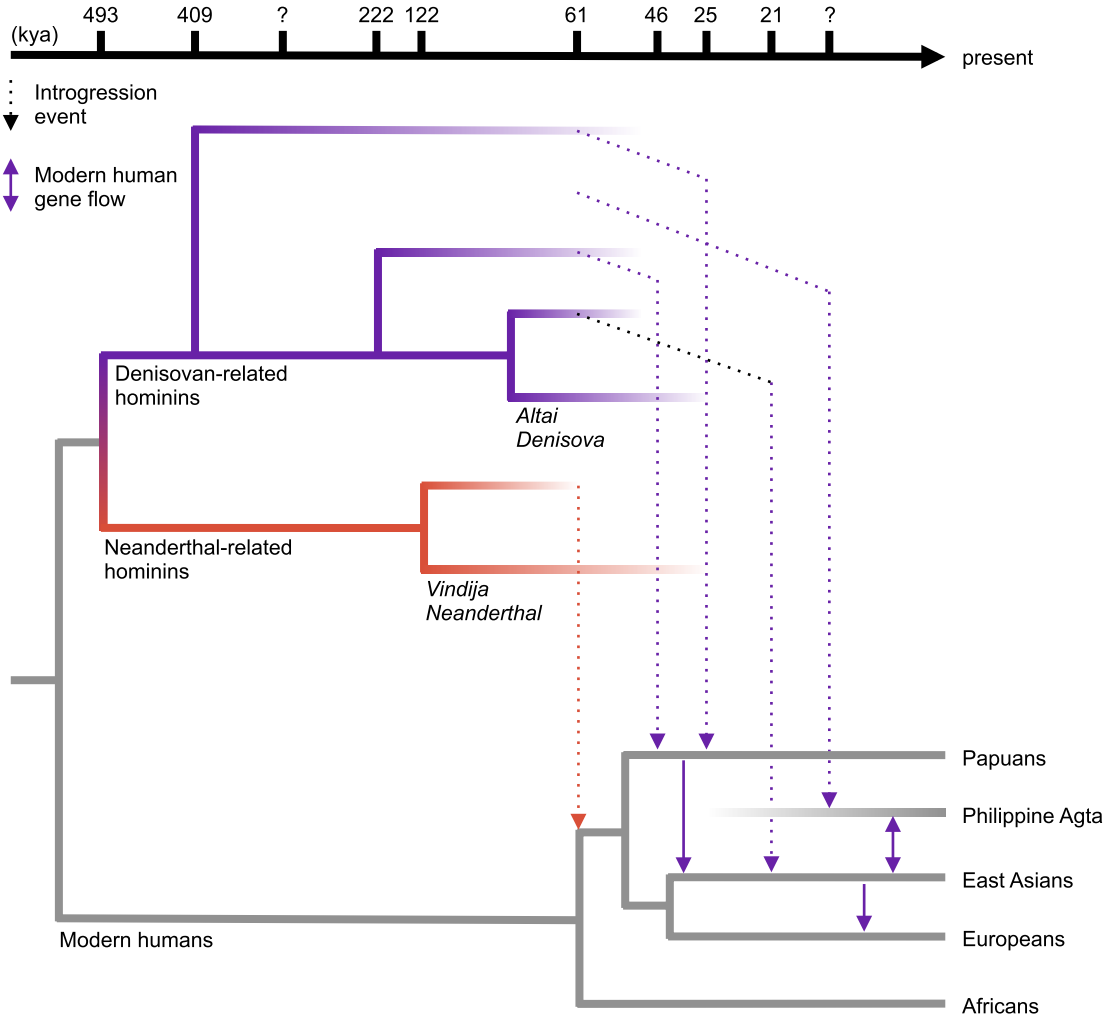
Extended Data Fig. 7 | Classic sweeps and adaptive archaic introgression. **a**, **b**, Coloured squares indicate genomic regions displaying signals of both a selective sweep and adaptive introgression from Neanderthals (**a**) or Denisovans (**b**). Yellow and blue frames indicate genomic regions identified 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* 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
[x]	[ ] Antibodies
[x]	[ ] Eukaryotic cell lines
[x]	[ ] Palaeontology and archaeology
[x]	[ ] Animals and other organisms
[ ]	[x] Human research participants
[x]	[ ] Clinical data
[x]	[ ] Dual use research of concern

Methods

n/a	Involved in the study
[x]	[ ] ChIP-seq
[x]	[ ] Flow cytometry
[x]	[ ] 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.