Current Biology

Article

The genomic landscape of contemporary western Remote Oceanians

Highlights

- New genome-wide data for 1,433 contemporary ni-Vanuatu reveal fine-scale structure
- Admixture between East Asian- and Papuan-related peoples was spatially uneven
- Both Polynesian and non-Polynesian speakers in Vanuatu carry Polynesian ancestry
- Similarity in ancestry proportions among spouses supports social assortative mating

Authors

Lara R. Arauna, Jacob Bergstedt, Jeremy Choin, ..., Antoine Gessain, Lluis Quintana-Murci, Etienne Patin

Correspondence

lrubioar@pasteur.fr (L.R.A.),
quintana@pasteur.fr (L.Q.-M.),
epatin@pasteur.fr (E.P.)

In brief

Although the settlement of Remote Oceania is now well understood, the population processes occurring in Vanuatu for the last 3,000 years remain unclear. Arauna et al. show that the fine-scale genetic structure of present-day ni-Vanuatu has been shaped by spatially uneven admixture, Polynesian migrations, volcano eruptions, and marriage patterns.

Arauna et al., 2022, Current Biology *32*, 4565–4575
November 7, 2022 © 2022 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2022.08.055

CellPress

CellPress
OPEN ACCESS

Article

The genomic landscape of contemporary western Remote Oceanians

Lara R. Arauna,[1,15,*] Jacob Bergstedt,[1,2,3,13] Jeremy Choin,[1,4,13] Javier Mendoza-Revilla,[1,5,13] Christine Harmant,[1] Maguelonne Roux,[1,6] Alex Mas-Sandoval,[7] Laure Lémée,[8] Heidi Colleran,[9] Alexandre François,[10] Frédérique Valentin,[11] Olivier Cassar,[12] Antoine Gessain,[12] Lluis Quintana-Murci,[1,4,14,*] and Etienne Patin[1,14,16,*]

[1]Institut Pasteur, Université Paris Cité, CNRS UMR2000, Human Evolutionary Genetics Unit, Paris 75015, France
[2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden
[3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden
[4]Chair Human Genomics and Evolution, Collège de France, Paris 75005, France
[5]Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Perú
[6]Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, Paris 75015, France
[7]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK
[8]Institut Pasteur, Biomics Platform, Paris 75015, France
[9]BirthRites Independent Max Planck Research Group, Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany
[10]Langues, Textes, Traitements Informatiques, Cognition (LaTTiCe), UMR 8094, CNRS, Paris 75015, France
[11]MSH Mondes, UMR 8068, CNRS, Nanterre 92023, France
[12]Institut Pasteur, Université Paris Cité, CNRS UMR 3569, Oncogenic Virus Epidemiology and Pathophysiology Unit, Paris 75015, France
[13]These authors contributed equally
[14]Senior author
[15]Twitter: @lararubioarauna
[16]Lead contact
*Correspondence: lrubioar@pasteur.fr (L.R. A.), quintana@pasteur.fr (L. Q.-M.), epatin@pasteur.fr (E. P.)

https://doi.org/10.1016/j.cub.2022.08.055

SUMMARY

The Vanuatu archipelago served as a gateway to Remote Oceania during one of the most extensive human migrations to uninhabited lands ~3,000 years ago. Ancient DNA studies suggest an initial settlement by East Asian-related peoples that was quickly followed by the arrival of Papuan-related populations, leading to a major population turnover. Yet there is uncertainty over the population processes and the sociocultural factors that have shaped the genomic diversity of ni-Vanuatu, who present nowadays among the world's highest linguistic and cultural diversity. Here, we report new genome-wide data for 1,433 contemporary ni-Vanuatu from 29 different islands, including 287 couples. We find that ni-Vanuatu derive their East Asian- and Papuan-related ancestry from the same source populations and descend from relatively synchronous, sex-biased admixture events that occurred ~1,700–2,300 years ago, indicating a peopling history common to the whole archipelago. However, East Asian-related ancestry proportions differ markedly across islands, suggesting that the Papuan-related population turnover was geographically uneven. Furthermore, we detect Polynesian ancestry arriving ~600–1,000 years ago to Central and South Vanuatu in both Polynesian-speaking and non-Polynesian-speaking populations. Last, we provide evidence for a tendency of spouses to carry similar genetic ancestry, when accounting for relatedness avoidance. The signal is not driven by strong genetic effects of specific loci or trait-associated variants, suggesting that it results instead from social assortative mating. Altogether, our findings provide an insight into both the genetic history of ni-Vanuatu populations and how sociocultural processes have shaped the diversity of their genomes.

INTRODUCTION

Vanuatu, an archipelago located in western Remote Oceania, is a key region for understanding the peopling history of the Pacific. The cultural, anthropological, and genetic diversity of Vanuatu reflects three distinct phases of population movements. The first, which started in present-day Taiwan ~5,000 years ago (ya), was associated with the spread of Austronesian languages to Near and Remote Oceania.[1–3] The so-called Austronesian expansion led to the emergence of the well-characterized Lapita cultural complex, which developed in the Bismarck Archipelago and reached Vanuatu ~3,000 ya.[4–6] Morphometric and ancient DNA (aDNA) studies indicate that the Lapita people in Vanuatu carried East Asian-related ancestry, supporting a connection with the Austronesian expansion.[7,8] The second migration occurred after the Lapita period, ~2,500 ya, and involved the arrival of Papuan-related peoples who shared ancestry with contemporary Bismarck Archipelago islanders. aDNA studies have shown that these migrations triggered a dramatic shift in genetic ancestry, from the East Asian-related ancestry observed in first Remote Oceanians to the Papuan-related ancestry that has remained predominant since then.[9,10] Finally, it has been postulated that "Polynesian outliers" from Vanuatu (i.e., Polynesian speakers living outside the Polynesia triangle) are the descendants of migrations from Polynesia into western Remote Oceania,[11,12] a model that has received recent genetic support.[13]

Although aDNA studies have revealed that ni-Vanuatu are descended from at least three ancestral populations,[9,10,13] whether the settlement process was uniform across the multiple islands of the archipelago remains an open question. aDNA data indicate that individuals dated to ~2,500 ya from Central and South Vanuatu carried largely different proportions of East Asian-related ancestry,[9,10,13] in line with either a unique population turnover that was geographically heterogeneous or separate admixture events between different populations across islands. Furthermore, Vanuatu is the country with the world's largest number of languages per capita,[14] which supports the view that languages have rapidly diversified since the initial settlement and/or that the arrival of new groups to the archipelago further increased linguistic diversity. Nevertheless, these questions have been difficult to resolve because, to date, available ancient and modern DNA data from the region have remained sparse.[9,10]

Here, we generated genome-wide genotype data for 1,433 contemporary ni-Vanuatu and assessed their fine-scale genetic structure in order to address the following central questions about the settlement of western Remote Oceania: do all contemporary ni-Vanuatu derive their ancestry from three populations only? Was the contribution of these three ancestral populations different across the archipelago? Was the post-Lapita shift to Papuan-related ancestry heterogeneous across islands? In addition, the recent settlement of Vanuatu allows for the study of how sociocultural practices have shaped genetic diversity in humans over the past ~3,000 years in the form of sex-biased admixture, relatedness avoidance, and non-random mating.[15–17] We thus used the comprehensive set of population genomics data presented here, which includes 287 male-female couples, to answer these important questions relating to the genetic history of a population that descends from diverse ancestral populations: was admixture sex biased? Was admixture accompanied by language shifts? Has socio-cultural structure influenced mating? Can residence rules and urbanization affect human genetic structure?

RESULTS

The genetic variation of ni-Vanuatu is spatially structured

To shed light on the genetic make-up of ni-Vanuatu, we collected >4,000 blood samples from contemporary individuals between 2003 and 2005 and genotyped a selection of 1,433 of these sampled individuals at 2,358,955 SNPs. After merging with 179 high-coverage whole-genome sequencing data[18] and excluding 173 low-quality or duplicated samples, a total of 1,439 ni-Vanuatu samples was included in subsequent analyses, including those from 522 males and 917 females living on 29 islands and in 179 different villages (Figure 1A; Data S1A).

Principal-component analysis (PCA) and ADMIXTURE analyses indicate that contemporary ni-Vanuatu fall on a genetic gradient between East Asian-related and Papuan-related populations (Figures 1B and S1), supporting the view that their ancestry derives from these two population groups. When projecting ancient samples from Vanuatu, we found that Lapita individuals show higher affinity with present-day East Asian-related populations, whereas post-Lapita individuals are closer to Papuan-related populations, in line with a Papuan-related population turnover occurring after the Lapita period.[8–10,13] Furthermore, contemporary ni-Vanuatu show high genetic similarities with individuals from the Bismarck Archipelago, in line with the hypothesis that their Papuan-related ancestors originated from these islands.[8–10,13,18] However, the extensive geographic coverage of the dataset presented here enabled us to reveal a substantial genetic substructure among ni-Vanuatu islanders. PCA showed that genetic variation in ancient and contemporary individuals from Vanuatu is explained by two contiguous but distinct groups that broadly reflect geography (Figures 1C and S1). When considering the number of ancestral components that is most supported by ADMIXTURE (K_{ADM} = 9; Methods S1A), populations from different islands also show different ancestral components (Figure S1). These results reveal that genetic variation in contemporary ni-Vanuatu is structured, which could result either from different admixture histories during the settlement period and/or from the existence of barriers to gene flow that formed after the settlement of the archipelago.

To gain further insights into the genetic history of ni-Vanuatu, we next assessed the fine-scale genetic structure of the 1,439 sampled individuals, using ChromoPainter and fineSTRUCTURE.[19–21] Haplotype-based clustering revealed a first separation between ni-Vanuatu living north and south of the strait separating Epi and Tongoa islands (K_{FS} = 2; Figure S2A). At K_{FS} = 4, populations separate into clusters, here referred to as the Banks and Torres Islands cluster and the North East, North West, and Central-South Vanuatu clusters (Figures 2A and S2A; Data S1B), that are in general agreement with the classification of Oceanic languages spoken in the archipelago.[22–24] Of note, individuals from the north and south of Pentecost Island cluster with individuals from different neighboring islands, indicating that the sea does not necessarily act as a barrier to gene flow in the region, in line with linguistic and ethnographic data that reveal cultural networks between islands.[22,23,25] Likewise, in the Ambae Island, northwest and south inhabitants cluster separately (Figure 2A), suggesting that the periodic volcanic activity of the Ambae caldera has affected gene flow in the region. At K_{FS} = 20, i.e., the highest K_{FS} value for which statistical robustness remains maximal (STAR Methods),[20] we found that genetic clusters are often island-specific (Figures 2A and S2A; Data S1A; Methods S1B and S1C). These observations suggest that clusters inferred by fineSTRUCTURE are reliable, as they reflect expected geographic and linguistic barriers to gene flow.

The post-Lapita shift to Papuan-related ancestry was geographically uneven

We detected important differences among fineSTRUCTURE clusters in the proportions of East Asian-related ancestry,

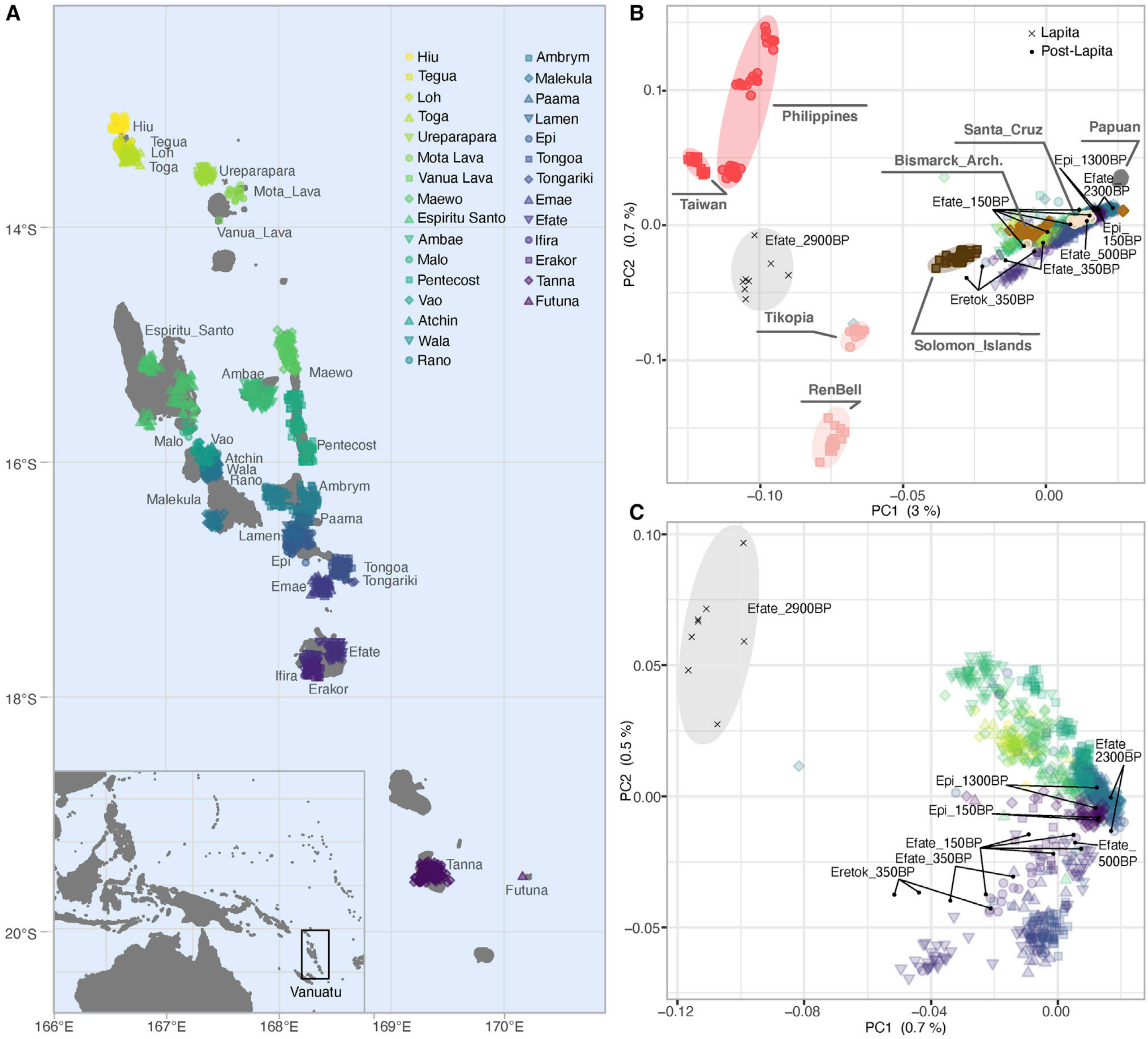

Figure 1. Sampling locations and genetic structure

(A) Map showing the sampling location of the 1,439 contemporary ni-Vanuatu individuals. The inset at the bottom left shows the location of Vanuatu in the Asia-Pacific region. Noise was added to sampling locations to facilitate visualization. See also Data S1A.

(B and C) Principal-component analysis (PCA) of genotypes of ni-Vanuatu at 301,774 SNPs, in the context of (B) the broad Asia-Pacific region and (C) the Vanuatu archipelago only. Black crosses and points indicate projected ancient samples from Vanuatu[9,13] dated to the Lapita and post-Lapita periods, respectively. Numbers in brackets indicate the proportion of variance explained by the corresponding PC. See also Figure S1.

Each point indicates an individual, colored according to the latitude of their island of residence.

i.e., ancestry related to contemporary populations from either Taiwan and the Philippines or Polynesia (range of SOURCEFIND estimates: 0.15–0.55; Figures 2C and S2B; Data S1C). East Asian-related ancestry was found to be the lowest in the North Vanuatu clusters (e.g., Malekula, Ambrym, and Epi; median = 0.230, SD = 0.056) and highest in the Central-South Vanuatu cluster (i.e., Mele and Imere in Efate Island; Makatea, Tongamea, and Vaitini in Emae Island; and Tongoa and Ifira islands; median = 0.323, SD = 0.077), where “Polynesian outlier” communities live today.[11,12] Nevertheless, East Asian-related ancestry is also high in islands where Polynesian ancestry is low, such as Ambae (median = 0.405, SD = 0.042; Figure 2D; Data S1C), suggesting that differences in East Asian-related ancestry are not solely due to differences in Polynesian ancestry.

To assess whether differences in East Asian-related ancestry originate from a geographically uneven ancestry turnover or separate admixture events between distinct populations, we dated admixture in each genetic cluster separately. Besides more recent events relating to Polynesian migrations (see next section), all estimates overlap the same time period that ranges from 1,700 to 2,300 ya (Figure 2B; Data S1D and S1E), suggesting that all ni-Vanuatu share the same admixture history. To test this hypothesis

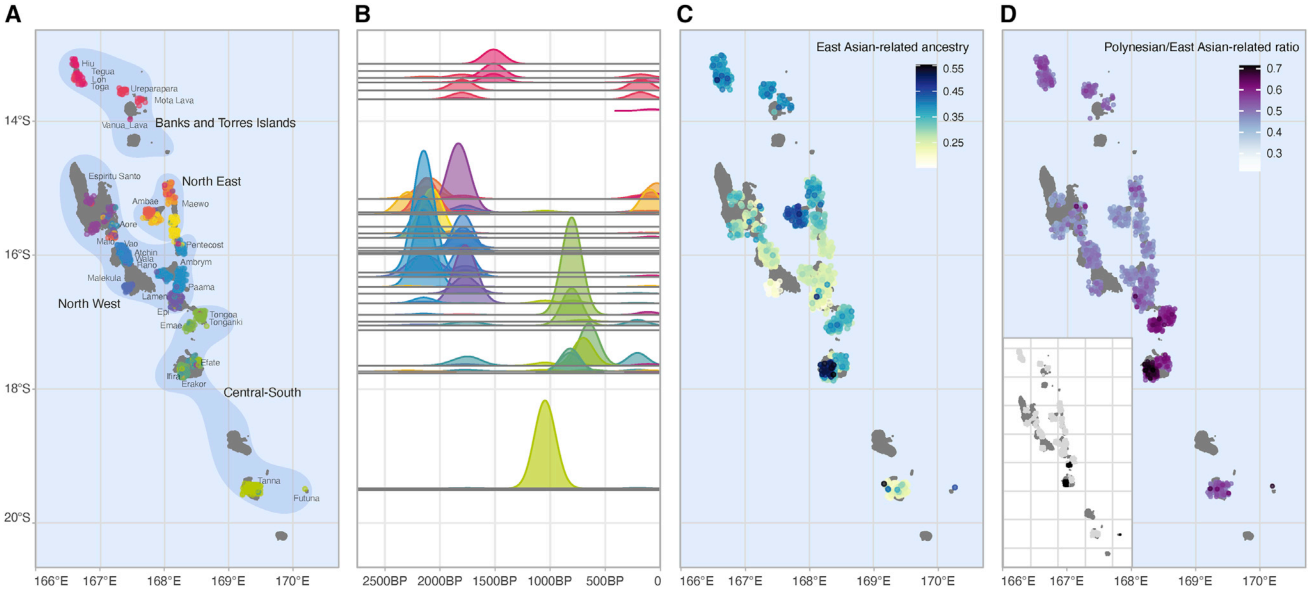

Figure 2. Fine-scale genetic structure and admixture

(A) Map showing the clustering of 979 ni-Vanuatu into 20 genetic clusters, according to fineSTRUCTURE ($K_{FS} = 20$). Each point indicates an individual, located according to their village of residence. Colors indicate genetic clusters, so the closer the colors, the closer the clusters. Noise was added to sampling locations to facilitate visualization. See also Figure S2A and Data S1B.

(B) Admixture date estimates for each genetic cluster, based on 100 bootstrap replicates. Colors are the same as those used in (A). The x axis shows the admixture date estimated by GLOBETROTTER, in years before present, assuming a generation time of 28 years. The y axis indicates the latitude of the islands assigned to each genetic cluster. The heights of the density curves are proportional to the sample size of each cluster. See also Data S1D.

(C) Proportions of East Asian-related ancestry estimated by SOURCEFIND.

(D) Ratio of the Polynesian and East Asian-related ancestry proportions, estimated by SOURCEFIND. The inset at the bottom left shows the location of Polynesian-speaking individuals, indicated by black points.

(C and D) Each point indicates an individual, colored according to their ancestry proportions.

See also Figure S2B and Data S1C.

more formally, we evaluated if the Papuan-related ancestry carried by present-day ni-Vanuatu derives from a single source. PCA and f_4-statistics of the form f_4(X, New Guinean highlanders; Solomon or Bismarck Archipelago islanders, East Asians) indicate that all ni-Vanuatu show similar genetic relatedness with populations from the Bismarck Archipelago, New Guinea, or the Solomon Islands (Methods S1D and S1E), in agreement with aDNA data.[13] Furthermore, the haplotype-based SOURCEFIND method detected that the same cluster of Bismarck Archipelago islanders is the source that contributed the most to all ni-Vanuatu (Methods S1F; Data S1E). Of note, SOURCEFIND also detected a small contribution from New Guinean highlanders and Santa Cruz islanders in all ni-Vanuatu clusters, which negatively correlates with Polynesian ancestry ($r = -0.378$, $p < 2.22 \times 10^{-16}$), probably because Polynesian source populations capture some Papuan-related ancestry. Collectively, these findings indicate that ni-Vanuatu are descended from relatively synchronous admixture events between the same sources of Papuan- and East Asian-related ancestry, yet their proportions differ markedly across islands, suggesting that the dramatic Papuan-related ancestry shift that started ~2,500 ya was geographically uneven.

Polynesian migrations did not necessarily trigger language shifts

Admixture proportions and date estimates indicate that ni-Vanuatu ancestry partly derives from a third and more recent migration originating from Polynesia (Figures 2B–2D and 3; Data S1C and S1D). We dated admixture events between 600 and 1,000 ya for genetic clusters predominant in Mele and Imere (Efate Island) and Makatea, Tongamea, and Vaitini (Emae Island), as well as in Ifira and Futuna islands (Figure 2B; Data S1D), where Polynesian languages are spoken today.[12,26] These clusters show higher Polynesian ancestry and, therefore, higher East Asian-related ancestry proportions, relative to Banks and Torres Islands and North Vanuatu clusters (SOURCEFIND estimates; Figures 2D and 3; Data S1C). Ni-Vanuatu Polynesian speakers show a higher ratio of Polynesian-to-East Asian ancestry, when compared with non-Polynesian speakers (ratio = 0.647 versus 0.479; Wilcoxon test $p < 2.22 \times 10^{-16}$; Figure 2D). Furthermore, f_4-statistics of the form f_4(X, East Asians; Tongans, New Guinean highlanders) suggest that ni-Vanuatu from Efate, Ifira, and Emae share more alleles with Polynesians than other ni-Vanuatu (Figure S3).

Interestingly, our analyses also revealed that Polynesian ancestry is not restricted to Vanuatu islands where Polynesian languages are spoken. Non-Polynesian-speaking populations assigned to the Central-South Vanuatu (e.g., Tongoa, Tongariki, and Tanna) and the Banks and Torres Islands clusters also show a higher Polynesian-to-East Asian ancestry ratio, relative to North East and North West clusters (ratio = 0.557 and 0.510, versus 0.461; Wilcoxon test $p < 2.22 \times 10^{-16}$; Figure 2D; Data S1C). Furthermore, estimated admixture dates are similar among groups from the Central-South Vanuatu cluster

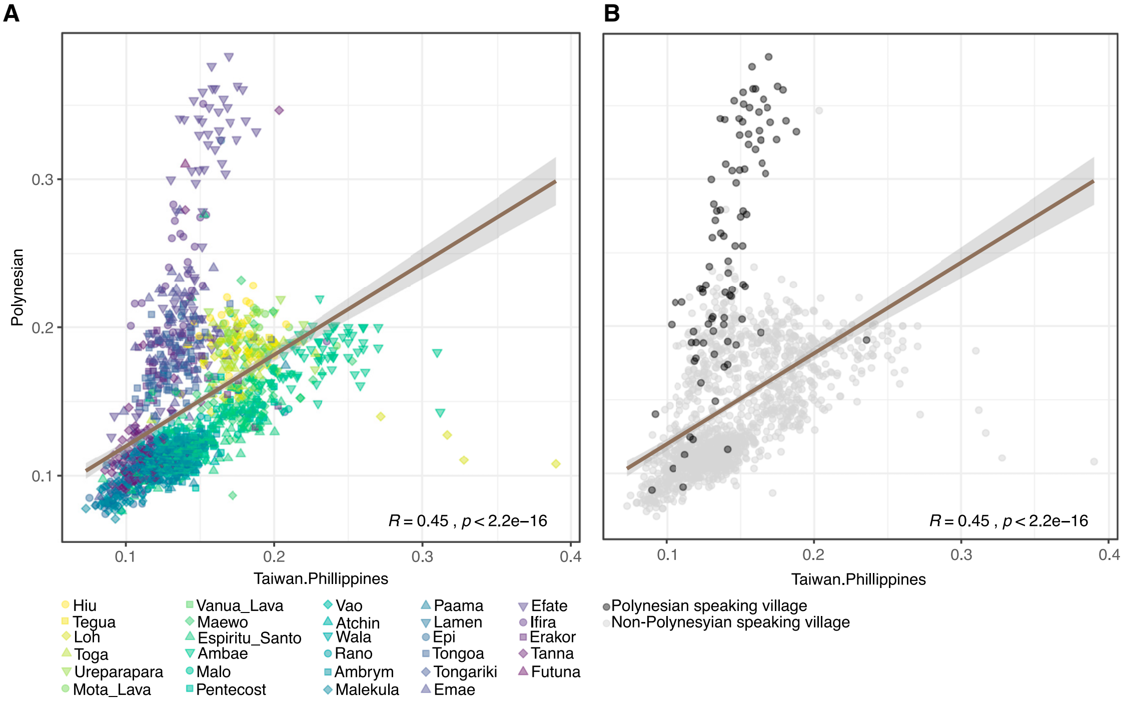

Figure 3. Genetic affinities to Taiwan/Philippines and Polynesian populations

Taiwan/Philippines and Polynesian ancestry proportions measure the proportion of ni-Vanuatu genomes that is closer to present-day populations from Taiwan and the Philippines or to Polynesians, respectively. Ancestry proportions were estimated by SOURCEFIND. See also Figure S3. The brown line indicates the regression line considering all the individuals from Vanuatu. The gray area indicates the 95% confidence level. *R* and p indicate Pearson's correlation coefficient and the corresponding p value (t test).

(A) Each individual is colored according to their island of residence, as in Figure 1A.

(B) The black and gray points indicate the individuals living in villages where Polynesian languages are spoken or not, respectively.

who speak or do not speak Polynesian languages (Figure 2B; Data S1D). These results support the view that Polynesian migrations, as well as subsequent contacts between Vanuatu islands[12,27] or with "Polynesian outliers" from the Solomon Islands,[28] also introduced Polynesian ancestry among non-Polynesian-speaking groups.

Conversely, we found no evidence of admixture with Polynesians in individuals assigned to the North Vanuatu cluster, including Epi islanders (Figures 2D and 3; Data S1C), despite the geographic proximity between Epi and Tongoa. Of note, ADMIXTURE, PCA, and fineSTRUCTURE analyses separate ni-Vanuatu into northern and southern populations, the frontier between the two being located between Epi and Tongoa (Figures 1C, 2A, S1, and S2A). The strait that separates the two islands today is the location of the Kuwae caldera,[29,30] whose volcanic activity may have been a barrier to gene flow and/or may have triggered large-scale population movements that disrupted the isolation by distance patterns. Together, these results reveal that, since 1,000 ya onward, Polynesians migrated to Vanuatu where they admixed with local populations and that such interactions did not necessarily result in a shift to Polynesian languages.

A limited genetic impact of European colonization

Our genetic data indicate that admixture between ni-Vanuatu and Europeans has been rare or has left few descendants in Vanuatu. In total, only 28 individuals out of 1,439 (1.95%) show a genetic contribution from Europeans higher than 1%, according to SOURCEFIND analyses (range: 0.03%–35.3%; Figure S2B; Data S1C). The two fineSTRUCTURE clusters with the highest European ancestry (median = 0.093 and 0.107, versus 0.002 for other clusters) show an admixture event in the last 120 years (mean = 78 ya, 95%CI: [74–81] and mean = 111 ya, 95%CI: [104–118]) (Figure 2B; Data S1D). Similarly, the three other clusters that include individuals carrying more than 1% of European ancestry show a pulse of admixture occurring in the last 200 years. These results are consistent with historical records; early, fleeting European contacts with ni-Vanuatu began in 1606, with subsequent documented exploratory voyages in 1768, 1774, and 1809, yet it was only from around 1829 onward that contacts became more common, when Christian missionaries and European colonists settled in the archipelago and the first intermarriages were reported.[31,32]

Genetic admixture in Vanuatu was sex biased

Genetic studies have suggested that Papuan-related migrations from the Bismarck Archipelago into Remote Oceania were male-biased because contemporary Polynesians and ancient individuals from present-day Vanuatu show lower Papuan-related ancestry on the X chromosome, relative to autosomes.[8,13] To confirm that admixture between the ancestors of present-day ni-Vanuatu was sex biased, we estimated Papuan- and East Asian-related ancestry in ni-Vanuatu on each chromosome

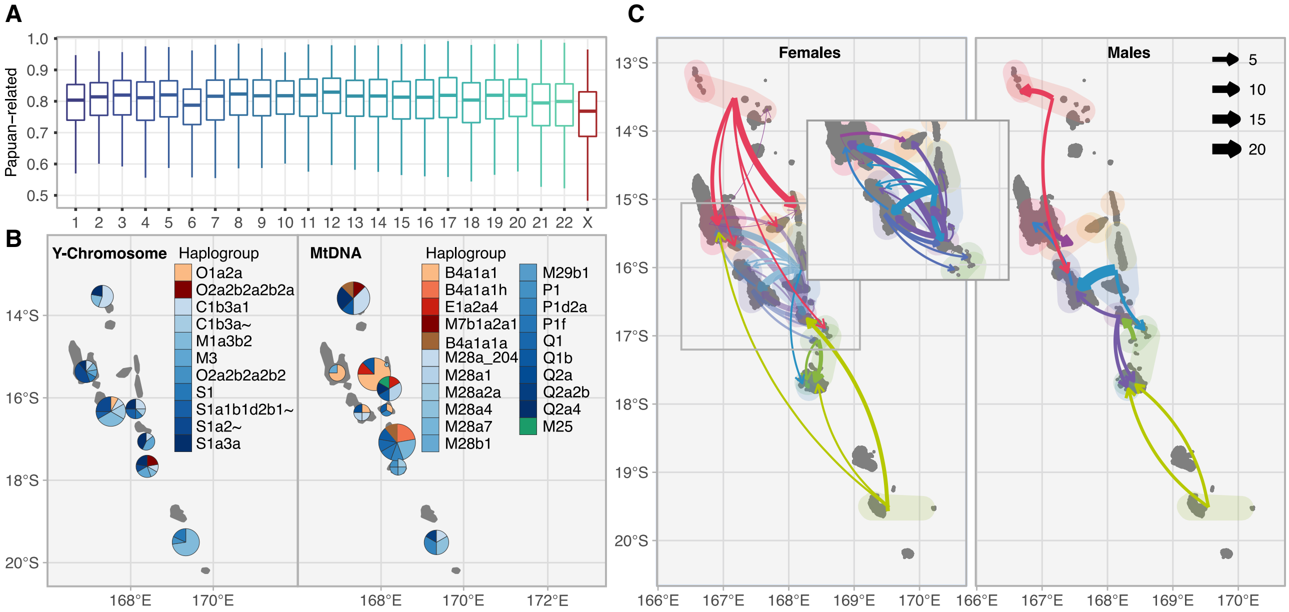

Figure 4. Sex-biased admixture and migration patterns

(A) Papuan-related ancestry proportions in ni-Vanuatu, estimated by RFMix for the 22 autosomes and the X chromosome separately. The line, box, and whiskers indicate the median, interquartile range, and 1.5× the interquartile range, respectively.

(B) Frequencies of Y chromosome and mtDNA haplogroups, colored according to their assumed origins (i.e., shades of blue or red indicate Papuan- or East Asian-related origins, respectively). Haplogroups were inferred from high-coverage genome sequencing data obtained for a subset of 179 ni-Vanuatu.[18] See also Data S1A.

(C) Recent migrations among Vanuatu islands, inferred based on fineSTRUCTURE clusters. The arrows connect the location of the genetic cluster to which individuals were assigned to their actual place of residence. The colors indicate the predominant genetic cluster in the island of origin. The width is proportional to the number of inferred migrant individuals, relative to the number of females or males in the genetic cluster.

See also Figure S4 and Data S1F.

separately, using local ancestry inference.[33] We found that Papuan-related ancestry is indeed significantly lower on the X chromosome, relative to autosomes (α_X = 75.2% versus α_{auto} = 79.8%, Wilcoxon test $p < 1.36 \times 10^{-5}$; Figure 4A), in agreement with aDNA results.[13] These values were similar between Polynesian and non-Polynesian speakers (Methods S1G; Wilcoxon test p = 0.13), which indicates that it is not explained by recent migrations from Polynesia. We replicated the results using high-coverage genome sequences from a subset of 179 ni-Vanuatu,[18] implying that our results are not biased due to SNP ascertainment (Methods S1G). Assuming that admixture proportions have reached equilibrium values,[34] we estimated that the genetic contribution of Papuan-related males to ni-Vanuatu was 27% higher than that of Papuan-related females (α_m = 93.5% versus α_f = 66.1%). Accordingly, Y chromosomes of ni-Vanuatu are dominated by haplogroups found at high frequency in Near Oceanians (e.g., M1a3b2 and S1), whereas mitochondrial DNAs (mtDNAs) show a high proportion of haplogroups typically found in East Asia (e.g., B4a1a1 and E1a2a4) (Figure 4B; Data S1A). Collectively, these results support the notion that ni-Vanuatu ancestry predominantly results from admixture between Papuan-related males and East Asian-related females.

Recent migrations are influenced by residence rules and urbanization

The genetic structure of contemporary ni-Vanuatu is also expected to reflect sociocultural practices (e.g., social networks, exchange, and marriage rules) that have culturally evolved since the settlement of the archipelago. We leveraged the high-resolution genetic data to infer recent migrations between Vanuatu islands—indicated by individuals who inhabit an island but belong to a genetic cluster that is prevalent in another island—and determined whether these migrations have involved mainly females or males, in line with virilocal or uxorilocal post-marital residence rules, respectively. We found that 5.70% of the sampled individuals (54 individuals) migrated at a large geographical scale (K_{FS} = 4), while 11.81% (112 individuals) migrated at a local scale (K_{FS} = 20; Data S1F), suggesting more genetic connections between closer islands. We estimated that local mobility among females is higher than among males (K_{FS} = 20; odds ratio = 2.20, Fisher's exact test $p = 9.16 \times 10^{-4}$; Figure 4C), as expected under virilocal residence and/or female exogamy.[35] The same trend was observed at a larger geographical scale, when considering migrations between four broader regions (K_{FS} = 4; odds ratio = 1.86, p = 0.075) and when restricting the analyses to the reported birth place of male-female couples included in the dataset (odds ratio = 1.96, $p = 6.89 \times 10^{-3}$). Notably, comparisons of the places of birth and residence of sampled individuals did not support female-biased migrations (odds ratio = 0.807, p = 0.18; Figure S4; Data S1F), possibly reflecting a bias in the self-reported birthplaces. This might occur if women are inclined to report their husband's or children's birthplace rather than their father's or mother's birthplace[36] or under other, more complex patterns.

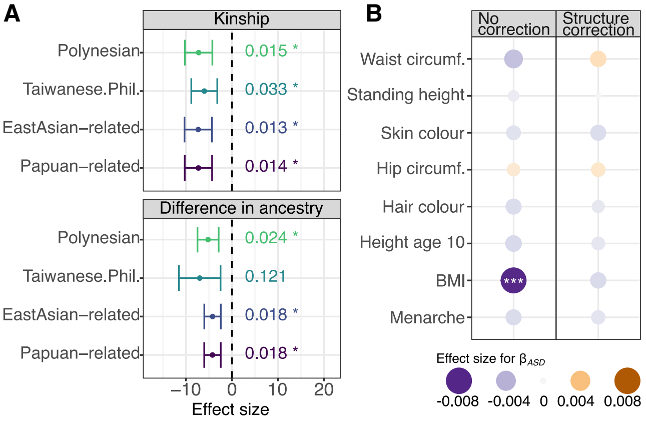

Figure 5. Ancestry- and trait-based assortative mating

(A) Effects of kinship and ancestry differences on partner choice. The point and error bars indicate the effect size and the 95% confidence interval. Effect sizes were estimated with a logistic regression model, while accounting for population structure (island of birth and village of residence). The effect size and p value were estimated using different ancestries as predictors, independently. See also Figure S5 and Data S1G.

(B) Increased (purple) or decreased (orange) genetic similarity among spouses (as measured by β_s^{ASD}) at trait-associated SNPs, relative to non-associated SNPs. Results on the left are based on a logistic model that includes only the genotype dissimilarity among spouses at each SNP, whereas results on the right are based on the logistic model that also controls for possible confounders (i.e., geography, kinship, and ancestry differences).

We also explored the direction of the inferred migrations and found that both female and male migrations mainly occurred from the northernmost and southernmost islands toward the center of the archipelago (Figure 4C), where Port Vila—the largest city—has developed.[37] Migrations are also common among North Vanuatu islands, consistent with a long-term network of cultural and material exchanges in the region.[25,28] Thus, our genomic data reflect the migration patterns that characterize the recent history of ni-Vanuatu, including residence rules and urbanization.

No evidence for endogamous practices among ni-Vanuatu spouses

Human kinship systems vary tremendously, regulating marriage, exchange, endogamy, or exogamy according to relational concepts.[38] Given the small census size of ni-Vanuatu populations, it is interesting to consider whether and how sociocultural features like marriage and exchange rules influence local levels of genetic relatedness. We found that genetic relatedness was higher within islands than between islands (Wilcoxon test $p < 2.22 \times 10^{-16}$; Figures S5A–S5C; Data S1G), consistent with an isolation by distance model (Mantel test p = 0.001; 1,000 permutations). Accordingly, out of the 287 male-female couples included in the dataset, 78.4% were born on the same island (Figure S5D). Genetic relatedness is also higher among inhabitants of the same village, relative to individuals living on the same island (Wilcoxon test $p < 2.22 \times 10^{-16}$) (Figure S5A; Data S1G), suggesting that the local community is often the source of marriage partners. However, despite the generally high levels of genetic relatedness observed between ni-Vanuatu, we did not observe an excess of genetic relatedness among couples. Specifically, spouses tend to show slightly lower kinship coefficients than random pairs of individuals from the same village, when excluding first-degree related individuals from all possible pairs (logistic regression model p = 0.057; resampling p = 0.079) (Methods S1H and S1I) and when adjusting for differences in ancestry ($p < 0.05$; Figure 5A). Assuming that our total sample is not biased toward related individuals, these results suggest that endogamy is not a common practice among contemporary ni-Vanuatu and more generally illustrate that populations can show high levels of genetic relatedness in the absence of endogamous marriage practices.

Ni-Vanuatu spouses tend to share similar genetic ancestry

Research on admixed populations from other parts of the world has shown that, in addition to low genetic relatedness, partners tend to show similar genetic ancestry,[16] because mating often occurs within sociocultural groups, which in turn can correlate with genetic ancestry. To verify whether this phenomenon is observed in Vanuatu, we implemented a logistic regression model that jointly estimates the effects of geography, genetic relatedness, and genetic ancestry on the probability to be partners (STAR Methods). These analyses confirmed that spouses tend to originate from the same island, while showing lower kinship coefficients than non-spouses (Figure 5A). Importantly, we found that spouses show lower differences in genetic ancestry than non-spouses, when considering Papuan-related ($\beta = -4.233$, p = 0.018), East Asian-related ($\beta = -4.221$, p = 0.018), or Polynesian ($\beta = -5.197$, p = 0.024) ancestry. These results suggest that ni-Vanuatu tend to mate with a partner who carries similar genetic ancestry.

Two hypotheses have been proposed to explain these observations. First, social assortative mating may underlie the observed signal, as genetic ancestry can correlate with sociocultural structure. Second, spouses may choose their partner because they share biological traits, such as physical appearance.[39,40] To test these hypotheses, we searched for genomic loci that could play a role in partner choice, by including in the logistic regression model a term that measures dissimilarity between individuals at each SNP (STAR Methods). No SNP showed statistical evidence for a significantly lower or higher genotype similarity between spouses, when accounting for multiple testing (Methods S1J). To test for assortative mating according to polygenic traits, we then evaluated if genotype similarity between spouses is significantly higher or lower at SNPs associated with candidate traits relating to physical appearance, when compared with non-associated SNPs. When we did not account for genetic structure and ancestry-associated assortative mating (i.e., effect sizes are estimated from a model where these confounders are not included; STAR Methods), we found evidence for assortative mating according to body mass index (BMI; Figure 5B). However, when accounting for such possible confounders, we found no statistical evidence that genotypes at trait-associated variants are more similar or dissimilar between spouses than expected. Collectively, our results do not support a marked tendency for partner choice according to genetic or phenotypic features, and they suggest instead the occurrence of assortative mating driven by social structure as the cause for ancestry-based assortments among ni-Vanuatu.

DISCUSSION

By leveraging an extensive genomic dataset of 1,439 contemporary individuals, we show here that ni-Vanuatu initially descend from admixture between the same ancestral populations: an East Asian-related population, which shares genetic affinities with groups living today in Taiwan and the Philippines, and a Papuan-related population, which shares genetic affinities with groups living today in the Bismarck Archipelago. We also estimate that admixture occurred after the Lapita period, ∼1,700–2,300 ya, and was relatively synchronous across islands, in agreement with a peopling history common to the whole archipelago. Thus, our results suggest that the high cultural diversity of ni-Vanuatu results from a rapid cultural diversification that developed *in situ*, as suggested by linguistic, archaeological, and archaeogenetic studies.[13,14,28] Nevertheless, our analyses cannot definitely rule out that, after the Lapita period, Vanuatu islands were settled by different, already admixed groups carrying varying levels of East Asian- and Papuan-related ancestry, as recently suggested.[18] Furthermore, we caution that admixture date estimates from modern DNA data are uncertain and can be biased downward when admixture was gradual,[41] which was likely the case in Vanuatu.[10] Additional aDNA time transects from multiple islands will be required to provide a definitive picture of the admixture history of ni-Vanuatu.

Our analyses reveal substantial differences in East Asian-related ancestry proportions between islands. We show that these differences do not result solely from Polynesian migrations, as our haplotype-based analyses could differentiate ancestry attributed to the Austronesian expansion from that introduced by Polynesians. A compelling example is Ambae, where East Asian-related ancestry is 1.8 times higher than in surrounding islands but where Polynesian ancestry is low. Assuming a simple admixture model, these findings imply that the major population turnover following the arrival of Papuan-related peoples was geographically uneven, possibly because, at the time of admixture, the two ancestral populations of ni-Vanuatu were of different sizes across islands, some of which being preferentially settled by East Asian-related groups and others by Papuan-related groups. Last, we confirm that admixture was sex biased;[8,13] either Papuan-related migrants were predominantly males or both males and females migrated, but admixture was more common between Papuan-related males and East Asian-related females.

A recent archaeogenetic study has reported that ancient ni-Vanuatu from the Chief Roi Mata's Domain in Efate show genetic similarities with Polynesians,[13] supporting the occurrence of migrations from Polynesia, which were previously postulated by linguistic studies.[11,12] We confirm that "Polynesian outlier" communities in Vanuatu are descended from admixture events between Polynesians and local populations. We dated these admixture events to 600–1,000 ya, in line with archaeological records.[11] Furthermore, we extend previous findings by mapping the genetic impact of Polynesian migrations to some Vanuatu islands where Polynesian languages are not spoken today (e.g., Makura, Tongoa, Tongariki, and Tanna).[12] These results indicate that genetic interactions between ni-Vanuatu and Polynesian incomers did not systematically trigger shifts to Polynesian languages. Intriguingly, Polynesian ancestry is not detected north of the Kuwae caldera, a large submarine volcano that separates Tongoa and Epi islands. Geological data have shown that the Kuwae volcano erupted in ca. 1452, producing among the largest volumes of magma and aerosol ever recorded,[29,30] and oral traditions and linguistic evidence suggest that after this eruption, Tongoa and Epi were repopulated by distant populations.[42,43] Our genetic results support this view; present-day Tongoa and Emae islanders, as well as Epi and southwest Malekula islanders, show close genetic affinities (Figure S2A), at odds with the gradual genetic differences expected under isolation by distance. Nowadays, ni-Vanuatu living north and south of Kuwae form two genetic groups with distinct sociocultural practices (e.g., grade-taking and chiefly title political systems),[44] indicating that the area has remained a genetic and cultural frontier.

By building upon the well-defined genetic history of Vanuatu, we also explored how genetic diversity has been shaped by cultural practices. While levels of genetic relatedness are high among ni-Vanuatu, we do not find evidence for generalized endogamy, which challenges the frequent association geneticists make between the two processes.[17,45–47] Nonetheless, even if ni-Vanuatu spouses are generally less related than non-spouses, we show that their genetic ancestries are more similar than expected, indicating that mating in Vanuatu is not random. Importantly, ancestry similarity between partners is not stronger at trait-associated SNPs, suggesting that ancestry-associated assortments are due to social structure, which may in turn be correlated with levels of East Asian-related and/or Polynesian ancestry. Other studies have suggested non-random mating according to ancestry in regions of the world where sociocultural structure is highly correlated with ancestry.[16,48] Our findings extend the occurrence of such sociocultural assortments to Oceanians, raising questions of how common this phenomenon is in human societies and whether non-random mating should systematically be accounted for in human genetic studies. Collectively, our study emphasizes the need to include diverse populations in genetic studies, not only to address key anthropological and evolutionary questions that are important for specific geographic regions but also to identify factors shaping the genetic diversity of human populations as a whole.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ SNP genotyping and quality filters
  - ○ Sample quality filters
  - ○ Merging with reference datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Genetic structure: PCA and ADMIXTURE

- ○ Haplotype phasing
- ○ Genetic structure: ChromoPainter and fineSTRUCTURE
- ○ Ancestry estimation
- ○ Admixture date estimation
- ○ Sex-biased admixture
- ○ Inferring migrations among islands
- ○ Tests for exogamy
- ○ Tests for ancestry-based assortative mating
- ○ Tests for SNP-based assortative mating
- ○ Tests for trait-based assortative mating

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.08.055.

ACKNOWLEDGMENTS

We thank all volunteers in Vanuatu who kindly provided samples and participated in this research. We thank the Ministry of Health of the Republic of Vanuatu (especially Myriam Abel, Rose Bahor, Yvannah Taga, Maturine Tary, and the former Minister Jerôme Ludvaune), the Vanuatu Family Health Association (especially Blandine Boulekone and Marie Nicklum Wala), the WHO Office in Port Vila (especially Corinne Capuano), and the Institut Pasteur of New Caledonia at Nouméa (especially Paul M.V. Martin and Eliane Chungue) for their continuous support and interest in this work. We are particularly thankful to Helene Walter, Woreka Mera, Françoise Charavay, and Sonia Treptow for their assistance in collecting samples. We thank S. Créno and the HPC Core Facility of Institut Pasteur (Paris) for the management of computational resources. L.R.A. was funded by a Pasteur-Roux-Cantarini fellowship from the Institut Pasteur. The laboratory of Human Evolutionary Genetics is supported by the Institut Pasteur, the Collège de France, the CNRS, the Fondation Allianz-Institut de France, 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), the Fondation de France (no. 00106080), the Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339214), and the French National Research Agency (ANR-19-CE35-0005).

AUTHOR CONTRIBUTIONS

Conceptualization, L.R.A., E.P., and L.Q.-M.; methodology, L.R.A., J.B., A.M.-S., and E.P.; software, L.R.A. and M.R.; formal analysis, L.R.A.; data generation, C.H. and L.L.; resources, O.C. and A.G.; writing – original draft, L.R.A., L.Q.-M., and E.P.; writing – review & editing, L.R.A., J.B., J.C., J.M.-R., M.R., A.M.-S., H.C., A.F., F.V., L.Q.-M., and E.P.; visualization, L.R.A.; supervision, L.Q.-M. and E.P.; funding acquisition, L.Q.-M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. We worked to ensure that the study questionnaires were prepared in an inclusive way.

Received: May 2, 2022
Revised: July 1, 2022
Accepted: August 18, 2022
Published: September 14, 2022

REFERENCES

1. Bellwood, P. (2004). First Farmers: The Origins of Agricultural Societies (Wiley-Blackwell).
2. Gray, R.D., Drummond, A.J., and Greenhill, S.J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science *323*, 479–483.
3. Hung, H.-C., and Carson, M.T. (2014). Foragers, fishers and farmers: origins of the Taiwanese Neolithic. Antiquity *88*, 1115–1131.
4. Bedford, S., and Spriggs, M. (2014). The Archaeology of Vanuatu: 3000 Years of History across Islands of Ash and Coral (Oxford University Press).
5. Petchey, F., Spriggs, M., Bedford, S., and Valentin, F. (2015). The chronology of occupation at Teouma, Vanuatu: use of a modified chronometric hygiene protocol and Bayesian modeling to evaluate midden remains. J. Archaeol. Sci. Rep. *4*, 95–105.
6. Petchey, F., Spriggs, M., Bedford, S., Valentin, F., and Buckley, H. (2014). Radiocarbon dating of burials from the Teouma Lapita cemetery, Efate, Vanuatu. J. Archaeol. Sci. *50*, 227–242.
7. Valentin, F., Détroit, F., Spriggs, M.J., and Bedford, S. (2016). Early Lapita skeletons from Vanuatu show Polynesian craniofacial shape: implications for Remote Oceanic settlement and Lapita origins. Proc. Natl. Acad. Sci. USA *113*, 292–297.
8. Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G.R., Reepmeyer, C., Petchey, F., Fernandes, D., et al. (2016). Genomic insights into the peopling of the southwest Pacific. Nature *538*, 510–513.
9. Lipson, M., Skoglund, P., Spriggs, M., Valentin, F., Bedford, S., Shing, R., Buckley, H., Phillip, I., Ward, G.K., Mallick, S., et al. (2018). Population turnover in remote Oceania shortly after initial settlement. Curr. Biol. *28*, 1157–1165.e7.
10. Posth, C., Nägele, K., Colleran, H., Valentin, F., Bedford, S., Kami, K.W., Shing, R., Buckley, H., Kinaston, R., Walworth, M., et al. (2018). Language continuity despite population replacement in Remote Oceania. Nat. Ecol. Evol. *2*, 731–740.
11. Flexner, J.L., Bedford, S., and Valentin, F. (2019). Who was Polynesian? Who was Melanesian? Hybridity and ethnogenesis in the South Vanuatu Outliers. J. Soc. Archaeol. *19*, 403–426.
12. Hermann, A., and Walworth, M. (2020). Approche interdisciplinaire des échanges interculturels et de l'intégration des communautés polynésiennes dans le centre du Vanuatu. J. Soc. Océanistes *151*, 239–262.
13. Lipson, M., Spriggs, M., Valentin, F., Bedford, S., Shing, R., Zinger, W., Buckley, H., Petchey, F., Matanik, R., Cheronet, O., et al. (2020). Three phases of ancient migration shaped the ancestry of human populations in Vanuatu. Curr. Biol. *30*, 4846–4856.e6.
14. François, A., Lacrampe, S., Franjieh, M., and Schnell, S. (2015). The Languages of Vanuatu: Unity and Diversity, *Volume 5* (Asia-Pacific Linguistics).
15. Heyer, E., Chaix, R., Pavard, S., and Austerlitz, F. (2012). Sex-specific demographic behaviours that shape human genomic variation. Mol. Ecol. *21*, 597–612.
16. Zou, J.Y., Park, D.S., Burchard, E.G., Torgerson, D.G., Pino-Yanes, M., Song, Y.S., Sankararaman, S., Halperin, E., and Zaitlen, N. (2015). Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. Proc. Natl. Acad. Sci. USA *112*, 13621–13626.
17. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. *19*, 220–234.
18. Choin, J., Mendoza-Revilla, J., Arauna, L.R., Cuadros-Espinoza, S., Cassar, O., Larena, M., Ko, A.M., Harmant, C., Laurent, R., Verdu, P., et al. (2021). Genomic insights into population history and biological adaptation in Oceania. Nature *592*, 583–589.
19. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.

20. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, et al. (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.

21. van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., Tarekegn, A., Thomas, M.G., Bradman, N., and Hellenthal, G. (2015). Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. PLoS Genet. *11*, e1005397.

22. Tryon, D.T. (1976). New Hebrides Languages: An Internal Classification (Australian National University).

23. François, A. (2011). Where *R they all? The geography and history of *R-loss in southern oceanic languages. Oceanic Linguist. *50*, 140–197.

24. Lynch, J.D., Ross, M.T., and Crowley, T. (2011). The Oceanic Languages (Routledge).

25. Huffman, K. (1996). Trading, cultural exchange and copyright: important aspects of Vanuatu art. In Arts of Vanuatu, J. Bonnemaison, K. Huffman, C. Kaufmann, and D. Tryon, eds. (Crawford House Press), pp. 182–194.

26. Walworth, M., Dewar, A., Ennever, T., Takau, L., and Rodriguez, I. (2021). Multilingualism in Vanuatu: four case studies. Int. J. Bilingualism *25*, 1120–1141.

27. Lynch, J.D., and Fakamuria, K. (1994). Borrowed moieties, borrowed names: sociolinguistic contact between Tanna and Futuna-Aniwa. Vanuatu. Pac. Stud. *17*, 79–91.

28. Bedford, S., and Spriggs, M. (2008). Northern Vanuatu as a pacific crossroads: the archaeology of discovery, interaction, and the emergence of the ‘‘ethnographic present’’. Asian Perspect. *47*, 95–120.

29. Monzier, M., Robin, C., and Eissen, J.-P. (1994). Kuwae (≈ 1425 A.D.): the forgotten caldera. J. Volcanol. Geotherm. Res. *59*, 207–218.

30. Gao, C., Robock, A., Self, S., Witter, J.B., Steffenson, J.P., Clausen, H.B., Siggaard-Andersen, M., Johnsen, S., Mayewski, P.A., and Ammann, C. (2006). The 1452 or 1453 A.D. Kuwae eruption signal derived from multiple ice core records: greatest volcanic sulfate event of the past 700 years. J. Geophys. Res. *111*, 1–11.

31. Flexner, J., Spriggs, M., Bedford, S., and Abong, M. (2016). Beginning historical archaeology in Vanuatu: recent projects on the archaeology of Spanish, French, and Anglophone colonialism. In Archaeologies of Early Modern Spanish Colonialism (Springer International Publishing), pp. 205–227.

32. Jolly, M. (2009). The sediment of voyages: re-membering Quirós, Bougainville and cook in Vanuatu. In Oceanic Encounters: Exchange, Desire, Violence (ANU Press), pp. 57–112.

33. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.

34. Goldberg, A., and Rosenberg, N.A. (2015). Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X chromosome. Genetics *201*, 263–279.

35. Vienne, B. (1984). Gens de Motlav. Idéologie et pratique sociale en Mélanésie. Publ. Soc. Océanistes *42*, 232–240.

36. Bolton, L. (1999). Women, place and practice in Vanuatu: a view from ambae. Oceania *70*, 43–55.

37. Petrou, K., and Connell, J. (2017). Rural-urban migrants, translocal communities and the myth of return migration in Vanuatu: the case of Paama. J. Soc. Océanistes *144–145*, 51–62.

38. Godelier, M. (2004). Métamorphoses de la parenté (Fayard).

39. Parra, F.C., Amado, R.C., Lambertucci, J.R., Rocha, J., Antunes, C.M., and Pena, S.D. (2003). Color and genomic ancestry in Brazilians. Proc. Natl. Acad. Sci. USA *100*, 177–182.

40. Alvarez, L., and Jaffe, K. (2004). Narcissism guides mate selection: humans mate assortatively, as revealed by facial resemblance, following an algorithm of ‘‘self seeking like’’. Evol. Psychol. *2*, 177–194.

41. Pugach, I., Duggan, A.T., Merriwether, D.A., Friedlaender, F.R., Friedlaender, J.S., and Stoneking, M. (2018). The gateway from Near into Remote Oceania: new insights from genome-wide data. Mol. Biol. Evol. *35*, 871–886.

42. Clark, R. (1996). Linguistic consequences of the Kuwae eruption. In Oceanic Culture History: Essays in Honour of Roger Green, J. Davidson, G. Irwin, F. Leach, A. Pawley, and D. Brown, eds. (New Zealand Journal of Archaeology Special Publication), pp. 275–285.

43. Ballard, C. (2021). Transmission’s end? Cataclysm and chronology in indigenous oral tradition. In The Routledge Companion to Global Indigenous History, A. McGrath, and L. Russell, eds. (Taylor & Francis Group), pp. 571–602.

44. Bonnemaison, J. (1996). Le tissu de nexus. In Arts of Vanuatu, J. Bonnemaison, K. Huffman, C. Kaufmann, and D. Tryon, eds. (Crawford House Press), pp. 176–177.

45. Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M., and Wilson, J.F. (2010). Genomic runs of homozygosity record population history and consanguinity. PLoS One *5*, e13996.

46. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., et al. (2008). Runs of homozygosity in European populations. Am. J. Hum. Genet. *83*, 359–372.

47. Bianco, E., Laval, G., Font-Porterias, N., García-Fernández, C., Dobon, B., Sabido-Vera, R., Sukarova Stefanovska, E., Kučinskas, V., Makukh, H., Pamjav, H., et al. (2020). Recent common origin, reduced population size, and marked admixture have shaped European Roma genomes. Mol. Biol. Evol. *37*, 3175–3187.

48. Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., et al. (2014). Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS Genet. *10*, e1004572.

49. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

50. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

51. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

52. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

53. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

54. Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics *32*, 2817–2823.

55. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

56. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

57. Chacón-Duque, J.C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., et al. (2018). Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nat. Commun. *9*, 5388.

58. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.

59. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

60. Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media).

61. Calabrese, C., Simone, D., Diroma, M.A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., and Attimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics *30*, 3115–3117.

62. Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. Mol. Biol. Evol. *35*, 1291–1294.

63. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

64. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

65. Arauna, L.R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G., and Comas, D. (2017). Recent historical migrations have shaped the gene pool of Arabs and berbers in North Africa. Mol. Biol. Evol. *34*, 318–329.

66. Mas-Sandoval, A., Arauna, L.R., Gouveia, M.H., Barreto, M.L., Horta, B.L., Lima-Costa, M.F., Pereira, A.C., Salzano, F.M., Hünemeier, T., Tarazona-Santos, E., et al. (2019). Reconstructed lost Native American populations from Eastern Brazil are shaped by differential Je/Tupi ancestry. Genome Biol. Evol. *11*, 2593–2604.

67. Lucotte, E.A., Skov, L., Jensen, J.M., Macià, M.C., Munch, K., and Schierup, M.H. (2018). Dynamic copy number evolution of X- and Y-linked ampliconic genes in human populations. Genetics *209*, 907–920.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Blood DNA samples for 1,433 contemporary ni-Vanuatu adults	This paper	N/A
Critical commercial assays		
Infinium Omni 2.5-8 kit	Illumina	Cat. No: 20037365
QIAamp DNA Blood Mini Kit	QIAGEN	Cat. No: 51106
Qubit dsDNA Broad-Range Assay Kit	Invitrogen	Cat. No: Q32853
Deposited data		
Newly generated Omni 2.5 array data for 1,433 contemporary ni -Vanuatu	This study	EGA: EGAS00001005910; https://ega-archive.org/studies/EGAS00001005910
Whole genome sequences of 179 contemporary ni-Vanuatu	Choin et al.[18]	EGA: EGAS00001004540; https://ega-archive.org/studies/EGAS00001004540
Human Origins array data for 823 contemporary individuals from Oceania and Eurasia	Pugach et al.[41]	Available upon request
Whole genome sequences of 98 contemporary individuals from Eurasia (SGDP)	Mallick et al.[49]	https://www.internationalgenome.org/data-portal/data-collection/sgdp
Genome-wide data for 14 ancient individuals from Vanuatu	Lipson et al.[9]	https://reich.hms.harvard.edu/datasets
Genome-wide data for 11 ancient individuals from Vanuatu	Lipson et al.[13]	https://reich.hms.harvard.edu/datasets
GWAS summary statistics for 8 candidate traits from the UK Biobank database	N/A	http://www.nealelab.is/uk-biobank
Software and algorithms		
GenomeStudio	Illumina	https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html
PLINK v.1.9b	Chang et al.[50]	https://www.cog-genomics.org/plink/
KING v.2.1	Manichaikul et al.[51]	https://www.kingrelatedness.com/Download.shtml
EIGENSOFT v.6.1.4	Patterson et al.[52]	https://github.com/DReichLab/EIG
ADMIXTURE v.1.3.0	Alexander et al.[53]	https://dalexander.github.io/admixture/index.html
pong v.1.4.7	Behr et al.[54]	https://github.com/ramachandran-lab/pong
ADMIXTOOLS v.6.0	Patterson et al.[55]	https://github.com/DReichLab/AdmixTools/releases
SHAPEIT v.2	Delaneau et al.[56]	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
ChromoPainter / ChromoCombine	Lawson et al.[19]	https://people.maths.bris.ac.uk/∼madjl/finestructure-old/chromopainter_info.html
fineSTRUCTURE v.2.0.7	Lawson et al.[19]	https://people.maths.bris.ac.uk/∼madjl/finestructure/
SOURCEFIND	Chacón-Duque et al.[57]	https://github.com/hellenthal-group-UCL/sourcefindV2
fastGLOBETROTTER	Hellenthal et al.[58]	https://github.com/sahwa/fastGLOBETROTTER
RFMix v.1.5.4	Maples et al.[33]	https://github.com/indraniel/rfmix
BWA v.0.7.13	Li, and Durbin[59]	https://github.com/lh3/bwa/releases
GATK v.3.8	Van der Auwera, and O'Connor[60]	https://github.com/broadinstitute/gatk/releases
MToolBox	Calabrese et al.[61]	https://github.com/mitoNGS/MToolBox
Yleaf	Ralf et al.[62]	https://github.com/genid/Yleaf

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Etienne Patin (epatin@pasteur.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The SNP array data generated in this study has been deposited to the European Genome-Phenome Archive (EGA) under accession number EGA: EGAS00001005910. This paper analyzes existing, publicly available data. The accession numbers for these datasets are listed in the key resources table. This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The sampling survey was conducted in the Republic of Vanuatu between April 2003 and August 2005. The purpose of the study was the estimation of the seroprevalence of HTLV-1 viral infection and the assessment of human genetic diversity in ni-Vanuatu. The recruitment of participants was carried out after the agreement of the Ministry of Health of Vanuatu, the head of each Directorate from the sampled province, and the chief of the sampled village. The data collectors, Olivier Cassar, Helene Walter, Woreka Mera and Antoine Gessain, were accompanied by the village chief and/or the head of the local dispensary. The nature and the scope of the study were explained in detail by Olivier Cassar in English, and by Helene Walter in Bislama (i.e., an English-based pidgin-creole[14] that is the main *lingua franca* of Vanuatu), during information meetings organized in each village. Participants could ask any questions after the information meeting. After several hours of reflection, each volunteer participant of at least 18 years of age was asked to sign a written informed consent form, including consent for research on human genetic diversity. Couples were identified through interviews in English or Bislama, and were preferentially sampled. Blood samples were collected either at the local dispensary, a gymnasium or a hut provided by the village chiefs.

Sex, age, birth place, village of residence and date of blood collection were collected through a structured questionnaire. Birth place was missing for 73 samples and the residence for 49 samples. From this information, we retrieved, for each sample, the geographical coordinates of the birth place (around the center of the island) and of the place of residence (around the village of residence), using MapAction reference maps (https://mapaction.org/). We considered as the locations of "Polynesian outliers" all the villages within Futuna, the villages of Mele and Imere in Efate, Ifira island and the villages of Makatea, Tongamea and Vaitini in Emae.[12] Among the 287 couples, 207 reported the same island of birth, while 57 reported a different island. This information was missing for 23 couples.

The study received approval from the Institutional Review Board of Institut Pasteur (n°2016-02/IRB/5) and the Ministry of Health of the Republic of Vanuatu. It was conducted in full respect of the legal and ethical requirements and guidelines for good clinical practice, in accordance with national and international rules. Namely, research was conducted in accordance with: (i) ethical principles set forth in the Declaration of Helsinki (Version: Fortaleza October 2013), (ii) European directives 2001/20/CE and 2005/28/CE, (iii) principles promulgated in the UNESCO International Declaration on Human Genetic Data, (iv) principles promulgated in the Universal Declaration on the Human Genome and Human Rights, (v) the principle of respect for human dignity and the principles of non-exploitation, non-discrimination and non-instrumentalization, (vi) the principle of individual autonomy, (vii) the principle of justice, namely with regard to the improvement and protection of health and (viii) the principle of proportionality. The rights and welfare of the participants have been respected, and the hazards did not outweigh the benefits of the study. Feedback to local communities and partners in Vanuatu is planned for 2023, under the guidance of H.C. The results of this study will be presented to key stakeholders, including the Ministry of Health and the Vanuatu Cultural Center (Port Vila). Written and recorded resources will be provided in Bislama and distributed to interested individuals and communities.

METHOD DETAILS

SNP genotyping and quality filters

Five milliliters of blood were obtained by venepuncture from each volunteer participant and transferred to the Institut Pasteur of New Caledonia, where plasma and buffy coats were isolated, frozen, and stored at –80°C. Samples were then transferred to the Institut Pasteur in Paris (France). Out of the 4,428 collected samples, 1,433 samples were selected for genetic analyses. These samples were selected in order to (i) cover the largest number of islands and villages, (ii) cover locations where "Polynesian outliers" exist nowadays and (iii) include as many couples as possible. After sample selection, a total of 179 different villages were covered, located on 29 islands (Data S1A). DNA was purified from frozen buffy coats at the Institut Pasteur of Paris (France), using QIAamp DNA Blood Mini Kit protocol (QIAGEN, Germany), and eluted in AE buffer. DNA concentration was quantified with the Invitrogen Qubit 3 Fluorometer using the Qubit dsDNA Broad-Range Assay (Invitrogen, United States). Prior to SNP array genotyping, DNA integrity was checked on agarose gels.

The 1,433 selected samples were genotyped on the Illumina Infinium Omni 2.5-8 array (San Diego, California). Genotype calling was performed using the Illumina GenomeStudio software. We excluded 2,491 SNPs with missing annotations, 9,661 duplicated SNPs, 1,772 SNPs with a GenTrain score < 0.4, SNPs with a missingness > 0.05 and SNPs that deviate from Hardy-Weinberg equilibrium (i.e., p value < 0.01 in more than one Vanuatu island). Only autosomal SNPs were kept for the analyses, unless otherwise stated. After filters, a total of 2,269,868 SNPs were kept. When the analyses required minor allele frequency (MAF) filters, a MAF > 0.01 threshold was applied. When a linkage disequilibrium pruning was required, we pruned the data using a window size of 50 Kb, a step size of 5 SNPs and a r^2 threshold of 0.5. After all these filtering steps were applied, the remaining number of SNPs was 301,774. All filters were applied using PLINK v.1.9b (Chang et al.[50]).

Sample quality filters

The highest genotype missingness per sample was 0.018. We removed 21 samples with outlier values for heterozygosity (mean ± 3 SD), suggestive of DNA contamination, leaving 1,412 samples. Cryptic relatedness between samples was detected using KING v.2.1 (Manichaikul et al.[51]). We excluded 456 samples with a kinship coefficient > 0.08 with another sample, whenever analyses required unrelated samples. For 12 samples, the reported sex did not match the genetic sex inferred by the Y- and X-chromosome call rates. These include one couple, for which the reported sex is the opposite to the genetic sex. We thus exchanged the sex of the two samples, as it is most likely an error on the sample annotation. We did not include the 10 remaining samples, when performing analyses relating to mating practices and sex-specific migrations. After quality filters, a total of 287 couples was present in the dataset.

Merging with reference datasets

We merged the new SNP array data with whole genome sequences of different populations across the Pacific[18] and worldwide populations from the SGDP project.[49] We also merged the dataset with SNP array data from Oceanian populations,[41] to perform some population genetics analyses. Datasets were merged using PLINK v.1.9b. Transversions were excluded, to avoid allele strand inconsistencies.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genetic structure: PCA and ADMIXTURE

We performed Principal Component Analyses (PCA) with the 'SmartPCA' algorithm implemented in EIGENSOFT v.6.1.4.[52] aDNA samples were projected on PCs by using the "lsqproject" and "shrinkmode" options. We inferred population structure with ADMIXTURE v.1.3.0,[53] using 10 different seeds and assuming that K_{ADM} varies from 2 to 17, and visualized ADMIXTURE results using pong v.1.4.7.[54] We computed f_4-statistics with ADMIXTOOLS v.6.0[55] and estimated standard errors by block jackknife.

Haplotype phasing

We phased the dataset using SHAPEIT v.2[56,63] using 500 conditioning states, 10 burn-in steps, 50 Markov chain Monte Carlo (MCMC) main steps, a window length of 1 cM and an effective population size of 15,000. We used 1000 Genomes Project data[64] as a reference panel and therefore removed all array SNPs that did not align with this data, prior to haplotype phasing.

Genetic structure: ChromoPainter and fineSTRUCTURE

We ran ChromoPainter[19] to infer haplotype sharing among individuals. This algorithm is based on a Hidden Markov Model in which each sample is treated as a "recipient", i.e., a mosaic of haplotypes from a set of "donor" samples. We performed different analyses with ChromoPainter. In Analysis 1, we inferred the genetic structure of ni-Vanuatu running ChromoPainter only for the unrelated ni-Vanuatu samples. We set each sample both as a recipient and as a donor for all the samples (-a mode). We first estimated the global mutation probability and the switch rate with 10 expectation-maximization (EM) iterations, by using the '-in -iM' options and by averaging estimates obtained for chromosomes 1, 7, 13 and 19. The estimated values were 1.542×10^{-4} and 81.415, respectively. Then, we ran ChromoPainter in the same mode for all chromosomes but fixing the parameters to estimated values. In Analysis 2, we repeated the same process independently for the reference populations. In this case, we ran ChromoPainter only for the samples outside Vanuatu[18,49] and repeated the two-step process described above. In this case, the global mutation probability was estimated to 4.53×10^{-4} and the switch rate was 255.42.

We then produced a coancestry matrix by summing the coancestry matrices for all chromosomes and estimated the *C* parameter with ChromoCombine. The *C* factor was estimated at 0.452 for the Vanuatu dataset (Analysis 1), and 0.498 for the reference population dataset (Analysis 2). We applied the model-based Bayesian clustering method fineSTRUCTURE v.2.0.7[19] on ChromoPainter coancestry matrices. We ran 1 million burn-in iterations (-y option) and 1 million MCMC iterations (-x option) and sampled values every 10,000 iterations. We then inferred the fineSTRUCTURE tree (-m T option). We repeated this process 5 times, with different random seeds.

Following a previous study,[20] we estimated for each individual the robustness of the clustering assignation, by comparing the final state (i.e., the state with the highest posterior probability) with the 100 MCMC samples. For each individual *i*, we estimated $x_i^{(m)}$, the number of samples that cluster with *i* for each of the *m* = 1,…,100 MCMC samples. We also estimated $y_{ik}^{(m)}$, the number of samples that cluster with i in the final state and in each of the MCMC samples, for each inferred cluster k. Finally, we computed $P_i = \sum_m [y_{ik}^{(m)} / x_i^{(m)}]$. Therefore, we ended up with a matrix of this sum P_i, for all individuals and for each genetic cluster, that shows the robustness in the assignation of each individual to its corresponding cluster in the final state. We repeated this process for $k = 2,…,K$, K being the maximum number of clusters identified by fineSTRUCTURE. We also repeated the process for the 5 random seeds and compared the performance of each seed. Methods S1B shows the number of ni-Vanuatu individuals with $P_i < 0.9$ for each seed s and for each $k = 2,…K$. These estimations allowed to test the robustness in the cluster assignation of each individual, to compare the performance across seeds and to determine the robustness at each k value.

To compare robustness among seeds, we followed a similar approach. We estimated $x_i^{(s)}$, where $s = 1,…,5$, which corresponds to the number of individuals clustering with individual i for each seed. We then estimated $y_{ik}^{(s)}$ and $\sum_s [y_{ik}^{(s)} / x_i^{(s)}]$, which shows the concordance in the assignation of each individual to a cluster across the different seeds. To summarize this information, we analyzed all the individual sums, for each seed and each cluster k, as well as the number of individuals who were not assigned to the same cluster, between one seed and all the other seeds (Methods S1C).

Based on these analyses, we chose the seed that showed the highest consensus among seeds (Methods S1C) and in which the largest number of samples are robustly assigned (Methods S1B) (seed number 289 for Analysis 1 and 161 for Analysis 2). Then, we defined the maximum k value based on the number of individuals who are robustly assigned to a cluster for each k value and we focused our analyses on this value. Finally, once we chose a seed and a k value, we removed samples that showed some uncertainty in the cluster assignation ($P_i \leq 0.9$). Following this process, we observed that the robustness decreases from $K_{FS} = 20$ and we removed 3 samples from Analysis 1 (Vanuatu). For the reference dataset (Analysis 2), all samples showed a $P_i > 0.9$ across all k values, therefore we used genetic clusters from the k value that best suited the hypotheses to test ($K_{FS} = 25$) (Data S1E).

For the subsequent analyses, we redefined populations based on fineSTRUCTURE results and therefore on the genetic data itself. This approach is rather trivial when applied to continental populations, like the reference populations presented here, because fineSTRUCTURE clusters coincide broadly with the geographical or cultural categories assigned to the populations. Yet, this strategy has proven useful when complex admixture shapes the genetic structure,[65,66] or when the studied populations are sampled from a small geographic region,[20] in which case, defining populations based on geography or culture may lead to a biased view of the genetic structure of the populations.

Ancestry estimation

We estimated admixture proportions for each ni-Vanuatu individual using SOURCEFIND.[57] We first ran ChromoPainter using the ‘‘donors’’ mode: we considered each ni-Vanuatu individual as a recipient only and defined donor populations by using the genetic clusters defined above (Analysis 2). We ran SOURCEFIND by considering all surrogates (‘‘all surrogates’’ analysis; Data S1E; Methods S1F) or by limiting as much as possible the number of surrogates (‘‘limited surrogates’’ analysis). In the latter case, we considered as surrogates West_Eurasia1, West_Eurasia2, West_Eurasia3, EastAsia1, EastAsia2, EastAsia3, Atayal, Paiwan, Southeast_Asia, Cebuano, RenBell, Tikopia, PNG1, PNG2 and PNG_SGDP. Then, we summed the estimated ancestry proportions for these surrogates and grouped them as follows: Taiwanese.Philippines (Atayal, Paiwan, Cebuano), EastAsian_mainland (EastAsia1, EastAsia2, EastAsia3, Southeast_Asia), Polynesian (RenBell, Tikopia), Papuan-related (PNG1, PNG2 and PNG_SGDP) and European (West_Eurasia1, West_Eurasia2, West_Eurasia3). The East Asian-related ancestry was estimated as the sum of Taiwanese.Philippines group and Polynesian group (the EastAsian_mainland was negligible, with a maximum proportion = 0.007). To facilitate visualization, we excluded an individual born in Futuna and living in Malekula, who shows a Polynesian ancestry of 0.745 in figures showing ancestry proportions (Figures 2, 3, and S2B). The next individual carrying the highest Polynesian ancestry shows 0.383. For both the ‘‘all surrogates’’ and ‘‘limited surrogates’’ analyses, we ran 200,000 MCMC iterations with a burn-in of 50,000 iterations, sampling every 5,000 iterations and not allowing for self-copy. We estimated the final ancestry of each ni-Vanuatu sample as the mean of the 30 sampled MCMC runs.

Admixture date estimation

We estimated admixture dates using fastGLOBETROTTER.[58] For this analysis, we ran ChromoPainter using the ‘‘donors’’ mode, by considering the ni-Vanuatu genetic clusters defined in (Analysis 1) as recipients and all the reference populations (Analysis 2) as both recipients and donors. Surrogates were the same as those considered for the ‘‘limited surrogates’’ SOURCEFIND analysis. For each of the 20 recipient ni-Vanuatu clusters, we performed 100 bootstraps, which are implemented as resamplings of chromosomes among the available samples. We set the ‘‘Null.ind’’ option to 1. We assumed a generation time of 28 years, to estimate dates from the number of generations.

Sex-biased admixture

We studied sex-biased admixture by comparing ancestry proportions estimated for the autosomes and the X chromosome. We estimated ancestry proportions using RFMix v.1.5.4 (Maples et al.[33]), considering as ancestral populations New Guinean highlanders (approximating Papuan-related ancestry) and Taiwanese Indigenous peoples and the Cebuano from the Philippines (approximating East Asian-related ancestry). We ran the ‘‘TrioPhased’’ algorithm, allowing for phase correction and 3 EM iterations. The window size was set at 0.03 cM (see Choin et al.[18]) and the admixture date was set at 50 generations. The same parameters were used for the X chromosome and the autosomes. We combined the X chromosome haploid data of males from the same island to obtain diploid individuals. We filtered out SNPs with an RFMix posterior probability < 0.9, those within centromeres and within 2 Mb from the telomeres. We then estimated East Asian-related and Papuan-related ancestry in the autosomes and the X chromosome, separately. We estimated α_f and α_m, that is, the proportion of female and male ancestors of ni-Vanuatu who carried Papuan-related ancestry, respectively, by assuming that admixture proportions have reached equilibrium values.[34] In such a case, $\alpha_f = 3\alpha_X - 2\alpha_{auto}$ and $\alpha_m = 4\alpha_{auto} - 3\alpha_X$, where α_{auto} and α_X are the average Papuan-related ancestry proportions estimated on the autosomes and the X chromosome, respectively.

To avoid potential biases due to SNP ascertainment, we also studied sex-biased admixture by comparing the autosomes and the sex chromosomes of 179 ni-Vanuatu sequenced at ~30× coverage.[18] Because genotypes on the X chromosome were not called in the previous study, we mapped *fastq* files on the X chromosome of the human reference genome (version hs37d5) and performed genotype calling for X-linked variants, as previously described,[18] setting the ploidy parameter to 1 for males and 2 for females. We kept only biallelic SNPs and filtered *vcf* files following GATK best practices[60]: QualByDepth < 2; FisherStrand value > 60; StrandOddsRatio > 3; RMSMappingQuality < 40; MappingQualityRankSumTest < -12.5; and ReadPosRankSum < -8. We also removed genotypes with a DP < 10 for females and < 3 for males, and those with a GQ < 30. We removed PAR1, PAR2 and XTR regions according to the annotations provided by the UCSC browser, and amplicon regions as reported in Lucotte et al.[67] We filtered out SNPs that were not in Hardy-Weinberg equilibrium in females (p value < 0.0001), those with a missingness > 0.05 and those with a MAF < 0.01. We estimated mtDNA haplogroups with MToolBox[61] from the *bam* and *fastq* files, and Y chromosome haplogroups were estimated with Yleaf.[62]

Inferring migrations among islands

We used the genetic structure inferred by fineSTRUCTURE to study migration patterns among Vanuatu islands. First, we assigned each genetic cluster to one or more islands, each time more than 25% of the sampled individuals inhabiting the island were assigned to that cluster. We then identified "outlier" individuals who inhabit an island but are assigned to a genetic cluster that is prevalent in another island. These individuals, either themselves or their close ancestors, likely migrated from the island/s where their genetic cluster is predominant to their current place of residence. We removed from these analyses Vanuatu_8 and Vanuatu_9 clusters (K_{FS} = 20) because those clusters are driven by recent European gene flow, as well as Vanuatu_15 because no island could be assigned to this cluster when using the 25% rule. To study sex-specific migrations, we calculated the proportion of female (male) migrants by dividing the number of "outlier" females (males) by the number of females (males) in each cluster. We did not include clusters where the number of males or females was < 2.

Tests for exogamy

We tested if spouses show lower genetic relatedness than non-spouses by testing if the average kinship coefficient between 287 observed couples is higher or lower than expected by chance, using either permutations or a logistic regression model. Kinship coefficients were computed between all possible pairs of individuals using KING v.2.1 (Data S1G). When using permutations, we accounted for isolation by distance by sampling random pairs of males and females among individuals born in the same island or living in the same village (Methods S1H). Because marriages between first-degree related individuals are very unlikely (they are indeed not observed in the data), we excluded from this analysis individuals who are first-degree relatives, but also tested how the inclusion of these individuals affect the results (Methods S1H). We performed 10,000 permutations for each island. In each permutation, we sampled equally many random pairs as there were observed pairs in the island. We calculated p values by comparing the average kinship coefficient among the observed couples to the null distribution. The null distribution was made of 10,000 average kinship coefficients between randomly sampled pairs of individuals from the same island or village.

Based on the sample scheme described above, we designed a logistic regression model that could be generalized to multiple explanatory variables, such as ancestry. Let the variable *i* index all possible pairs of males and females in the dataset, except pairs of individuals who are first-degree relatives, and define the dependent binary variable Y_i with $Y_i = 1$ if *i* indexes an observed pair of spouses and $Y_i = 0$ otherwise. Define the probability that a pair is an observed couple $p_i(x) = P(Y_i = 1)$, and introduce φ_i as the kinship coefficient between individuals in the *i*:th pair. We estimated the effect of kinship on mate choice by the parameter β^{φ} in the logistic regression model

$$\log\left[\frac{p_i(x)}{1-p_i(x)}\right] = \mu + \varphi_i\beta^{\varphi} + I(village_i)\beta^{v} + I(island_i)\beta^{I}, \qquad \text{(Equation 1)}$$

where $I(village_i) = 1$ if individuals of pair *i* are from the same village and $I(village_i) = 0$ otherwise, and $I(island_i) = 1$ if individuals of pair *i* are from the same island and $I(island_i) = 0$ otherwise. As a sensitivity analysis, we also considered a model in which pairs of first-degree related individuals were included (Methods S1I).

Tests for ancestry-based assortative mating

We extended the logistic regression model shown in Equation 1 to test for ancestry-based assortative mating, by testing if the genetic ancestry of spouses is more similar than that of non-spouses, accounting for population structure and relatedness avoidance. Let q_{i1}^{EA} and q_{i2}^{EA} be the proportion of East Asian-related ancestry for the individuals 1 and 2 of pair *i* and introduce the variable

$q_i^{EA} = \left|q_{i1}^{EA} - q_{i2}^{EA}\right|$.We estimated the effect of having similar proportions of East Asian-related ancestry between two individuals on the probability that a pair is an observed couple $p_i(x)$ by the parameter β^q in the logistic regression model,

$$\log\left[\frac{p_i(x)}{1 - p_i(x)}\right] = \mu + q_i^{EA}\beta^q + \varphi_i\beta^\varphi + I(village_i)\beta^v + I(island_i)\beta^I, \qquad \text{(Equation 2)}$$

where we use the same notations as in Equation 1. A negative (positive) effect size β^q is interpreted as evidence for assortative (disassortative) mating according to ancestry. Effect sizes for other ancestries were calculated similarly. The same model was also tested in ni-Vanuatu originating or not from islands where Polynesian languages are spoken. Note that in standard regression analyses, such as those used in GWAS, population stratification is usually corrected by principal components of the genetic relatedness matrix; here we have taken a more general approach by decomposing the genetic structure in two variables: the kinship and the ancestry, and have studied how both variables affect mate choice independently.

Tests for SNP-based assortative mating

We extended the logistic regression model shown in Equation 1 to test for SNP-based assortative mating, by testing if the genotypes at a given SNP are more similar between spouses than non-spouses, accounting for population structure, ancestry-based associated mating and relatedness avoidance. Define, for each SNP *s* and each pair *i* of individuals, the allele sharing distance (ASD) $d_{i,s}$ as $d_{i,s} = 0$ if both alleles are identical between individuals, $d_{i,s} = 1$ if only one allele is identical and $d_{i,s} = 2$ if none of the alleles are identical. We estimated the association between SNP *s* and mate choice by the parameter β_s^{ASD} in the model

$$\log\left[\frac{p_i(x)}{1 - p_i(x)}\right] = \mu + d_{i,s}\beta_s^{ASD} + q_i^{EA}\beta^q + \varphi_i\beta^\varphi + I(village_i)\beta^v + I(island_i)\beta^I, \qquad \text{(Equation 3)}$$

where we use the same notations as in Equation 2. A negative (positive) effect size β_s^{ASD} indicates higher (lower) similarity at SNP *s* between the members of the observed couples than between non-couples, in line with SNP-based assortative (disassortative) mating.

Tests for trait-based assortative mating

We tested for trait-based assortative mating by testing if genotypes of spouses are more similar or dissimilar at trait-associated SNPs, relative to non-associated SNPs. We obtained GWAS summary statistics for 8 candidate traits from the UK Biobank database (http://www.nealelab.is/uk-biobank). Candidate traits include traits relating to morphology and physical appearance. Let $y_s = \beta_s^{ASD}$, where β_s^{ASD} is the effect size of SNP *s* on mate choice estimated by Equation 1. We estimated if trait-associated SNPs are more similar or dissimilar between spouses by the parameter β^{trait} in the model

$$y_s = \mu + I(associated_s)\beta^{trait} + MAF_s\beta^{MAF} + GERP_s\beta^{GERP} + rec_s\beta^{rec}, \qquad \text{(Equation 4)}$$

where $I(associated_s) = 1$ if SNP *s* is significantly associated with the candidate trait in GWAS (with GWAS *P*-value $< 5\times10^{-8}$) and $I(associated_s) = 0$ otherwise, MAF_s is the minor allele frequency of SNP *s* in ni-Vanuatu, $GERP_s$ is the Genomic Evolutionary Rate Profiling (GERP) score of SNP *s* and rec_s is the interpolated recombination rate between SNP *s* and SNP $s - 1$, estimated in cM/Mb from the 1000 Genomes Phase 3 combined recombination map. We considered that there is trait-based assortative (disassortative) mating if β^{trait} is significantly negative (positive). We adjusted p values with the Bonferroni correction for multiple testing, to account for the number of traits tested.