The genomic landscape of contemporary western Remote Oceanians
Lara R. Arauna, Jacob Bergstedt, Jeremy Choin, Javier Mendoza-Revilla, Christine Harmant, Maguelonne Roux, Alex Mas-Sandoval, Laure Lémée, Heidi Colleran, Alexandre François, et al.

To cite this version:

HAL Id: hal-03885644
https://hal.science/hal-03885644v2
Submitted on 5 Dec 2022

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

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

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
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.
The genomic landscape of contemporary western Remote Oceanians

Lara R. Arauna,1,15,* Jacob Bergstedt,1,2,3,15 Jeremy Choin,1,4,13 Javier Mendoza-Revilla,1,5,13 Christine Harmant,1 Maguelonne Roux,1,8 Alex Mas-Sandoval,1 Laure Lémée,1 Heidi Colleran,7 Alexandre François,10 Frédérique Valentin,11 Olivier Cassar,12 Antoine Gessain,12 Lluis Quintana-Murci,1,4,14,* and Etienne Patin1,14,16,*

1Institut Pasteur, Université Paris Cité, CNRS UMR2000, Human Evolutionary Genetics Unit, Paris 75015, France
2Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden
3Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden
4Chair Human Genomics and Evolution, Collège de France, Paris 75005, France
5Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Perú
6Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, Paris 75015, France
7Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK
8Institut Pasteur, Biomics Platform, Paris 75015, France
9BirthRites Independent Max Planck Research Group, Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany
10Langues, Textes, Traitements Informatiques, Cognition (LaTTiCe), UMR 8094, CNRS, Paris 75015, France
11MSH Mondes, UMR 8068, CNRS, Nanterre 92023, France
12Institut Pasteur, Université Paris Cité, CNRS UMR 3569, Oncogenic Virus Epidemiology and Pathophysiology Unit, Paris 75015, France
13These authors contributed equally
14Senior author
15Twitter: @larrubioarau
16Lead 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.13 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 firstRemote Oceanians to the Papuan-related ancestry that has remained predominant since then.\textsuperscript{8,10} 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,\textsuperscript{1,12} a model that has received recent genetic support.\textsuperscript{13}

Although aDNA studies have revealed that ni-Vanuatu are descended from at least three ancestral populations,\textsuperscript{5,10,13} whether the settlement process was uniform across the multiple islands of the archipelago remains an open question. aDNA data indicate that individuals dated to \textasciitilde{}2,500 ya from Central and South Vanuatu carried largely different proportions of East Asian-related ancestry,\textsuperscript{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,\textsuperscript{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.\textsuperscript{5,10}

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 \textasciitilde{}3,000 years in the form of sex-biased admixture, relatedness avoidance, and non-random mating.\textsuperscript{9,10,11} 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\textsuperscript{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.\textsuperscript{5,8,10,13} 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.\textsuperscript{5,10,13,18} 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.\textsuperscript{18–21} Haploype-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.\textsuperscript{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.\textsuperscript{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),\textsuperscript{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,
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. 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 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 Vanuatu9,10 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.
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 \( f_4(X, \text{New Guinean highlanders}; \text{Solomon or Bismarck Archipelago islanders, East Asians}) \) indicate that all ni-Vanuatu show similar genetic relatedness with populations from the Bismarck Archipelago, New Guinea, or the Solomon Islands (Methods S1 D 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 S1 E). 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 \( \sim 2,500 \) 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 \((\text{ratio} = 0.647 \text{ versus } 0.479; \text{Wilcoxon test } p < 2.22 \times 10^{-16}; \text{Figure 2D})\). Furthermore, \( f_4 \)-statistics of the form \( f_4(X, \text{East Asians}; \text{Tongans, New Guinean highlanders}) \) 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 \((\text{ratio} = 0.557 \text{ and } 0.510, \text{versus } 0.461; \text{Wilcoxon test } p < 2.22 \times 10^{-16}; \text{Figure 2D; Data S1C})\). Furthermore, estimated admixture dates are similar among groups from the Central-South Vanuatu cluster.

---

**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.
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 or with “Polynesian outliers” from the Solomon Islands, 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 2Da 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, S1A, and S2A). The strait that separates the two islands today is the location of the Kuwae caldera, 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. 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. 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
separately, using local ancestry inference. We found that Papuan-related ancestry is indeed significantly lower on the X chromosome, relative to autosomes (\(a_X = 75.2\%\) versus \(a_{\text{auto}} = 79.8\%\), Wilcoxon test \(p < 1.36 \times 10^{-3}\); Figure 4A), in agreement with aDNA results. 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, implying that our results are not biased due to SNP ascertainment (Methods S1G). Assuming that admixture proportions have reached equilibrium values, we estimated that the genetic contribution of Papuan-related males to ni-Vanuatu was 27% higher than that of Papuan-related females (\(a_m = 93.5\%\) versus \(a_f = 66.1\%\)). 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. 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 or under other, more complex patterns.
founders (i.e., geography, kinship, and ancestry differences).

Right are based on the logistic model that also controls for possible genotype dissimilarity among spouses at each SNP, whereas results on the SNPs. Results on the left are based on a logistic model that includes only the

islands than between islands (Wilcoxon test $p < 2.22 	imes 10^{-3}$ relatedness. We found that genetic relatedness was higher within

like marriage and exchange rules influence local levels of genetic
tercepts.38 Given the small census size of ni-Vanuatu populations,

Human kinship systems vary tremendously, regulating marriage,
exchange, endogamy, or exogamy according to relational con-
cepts.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–SSC; 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 SSD). Genetic relatedness is also higher among inhabitants of the same village, relative to individuals living on the same island (Wilcoxon test $p < 2.22 \times 10^{-16}$[38,41] (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,38 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.

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 $b$38,41) 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).
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.

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. Furthermore, we caution that admixture date estimates from modern DNA data are uncertain and can be biased downward when admixture was gradual, which was likely the case in Vanuatu. 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; 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, supporting the occurrence of migrations from Polynesia, which were previously postulated by linguistic studies. 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. 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). 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, and oral traditions and linguistic evidence suggest that after this eruption, Tongoa and Epi were repopulated by distant populations. Our genetic results support this view: present-day Tongoa and Emae islanders, as well as Epi and southwest Malekula islanders, show close genetic affinities with east Asian-related ancestry, 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), 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. 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. 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
o Haplotype phasing
o Genetic structure: ChromoPainter and fineSTRUCTURE
o Ancestry estimation
o Admixture date estimation
o Sex-biased admixture
o Inferring migrations among islands
o Tests for exogamy
o Tests for ancestry-based assortative mating
o Tests for SNP-based assortative mating
o 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 Jérôme Ludvaune), the Vanuatu Family Health Association (especially Blandine Boulekone and Marie Nickllum 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 Chunge) for their continuous support and interest in this work. We are particularly thankful to Helene Walter, Woreka Mera, Françoise Charavay, and Sonia Trepтов 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 College 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 DEC2020180339214), and the French National Research Agency (ANR-19-CE35-0005).

AUTHOR CONTRIBUTIONS


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 that 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


### STAR METHODS

#### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood DNA samples for 1,433 contemporary ni-Vanuatu adults</td>
<td>This paper</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Critical commercial assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infinium Omni 2.5-8 kit</td>
<td>Illumina</td>
<td>Cat. No: 20037365</td>
</tr>
<tr>
<td>QIAamp DNA Blood Mini Kit</td>
<td>QIAGEN</td>
<td>Cat. No: 51106</td>
</tr>
<tr>
<td>Qubit dsDNA Broad-Range Assay Kit</td>
<td>Invitrogen</td>
<td>Cat. No: Q32853</td>
</tr>
<tr>
<td><strong>Deposited data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newly generated Omni 2.5 array data for 1,433 contemporary ni-Vanuatu</td>
<td>This study</td>
<td>EGA: EGAS00001005910; <a href="https://ega-archive.org/studies/EGAS00001005910">https://ega-archive.org/studies/EGAS00001005910</a></td>
</tr>
<tr>
<td>Whole genome sequences of 179 contemporary ni-Vanuatu</td>
<td>Choin et al.(^{18})</td>
<td>EGA: EGAS00001004540; <a href="https://ega-archive.org/studies/EGAS00001004540">https://ega-archive.org/studies/EGAS00001004540</a></td>
</tr>
<tr>
<td>Human Origins array data for 823 contemporary individuals from Oceania and Eurasia</td>
<td>Pugach et al.(^{41})</td>
<td>Available upon request</td>
</tr>
<tr>
<td>Genome-wide data for 14 ancient individuals from Vanuatu</td>
<td>Lipson et al.(^{9})</td>
<td><a href="https://reich.hms.harvard.edu/datasets">https://reich.hms.harvard.edu/datasets</a></td>
</tr>
<tr>
<td>Genome-wide data for 11 ancient individuals from Vanuatu</td>
<td>Lipson et al.(^{13})</td>
<td><a href="https://reich.hms.harvard.edu/datasets">https://reich.hms.harvard.edu/datasets</a></td>
</tr>
<tr>
<td>GWAS summary statistics for 8 candidate traits from the UK Biobank database</td>
<td>N/A</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td><strong>Software and algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenomeStudio</td>
<td>Illumina</td>
<td></td>
</tr>
<tr>
<td>PLINK v.1.9b</td>
<td>Chang et al.(^{50})</td>
<td><a href="https://www.cog-genomics.org/plink/">https://www.cog-genomics.org/plink/</a></td>
</tr>
<tr>
<td>KING v.2.1</td>
<td>Manichaikul et al.(^{51})</td>
<td><a href="https://www.kingrelatedness.com/Download.shtml">https://www.kingrelatedness.com/Download.shtml</a></td>
</tr>
<tr>
<td>EIGENSOFT v.6.1.4</td>
<td>Patterson et al.(^{52})</td>
<td><a href="https://github.com/DReichLab/EIG">https://github.com/DReichLab/EIG</a></td>
</tr>
<tr>
<td>ADMIXTURE v.1.3.0</td>
<td>Alexander et al.(^{53})</td>
<td><a href="https://dalexander.github.io/admixtiple/index.html">https://dalexander.github.io/admixtiple/index.html</a></td>
</tr>
<tr>
<td>pong v.1.4.7</td>
<td>Behr et al.(^{54})</td>
<td><a href="https://github.com/ramachandran-lab/pong">https://github.com/ramachandran-lab/pong</a></td>
</tr>
<tr>
<td>ADMIXTOOLS v.6.0</td>
<td>Patterson et al.(^{55})</td>
<td><a href="https://github.com/DReichLab/AdmixTools/releases">https://github.com/DReichLab/AdmixTools/releases</a></td>
</tr>
<tr>
<td>SHAPEIT v.2</td>
<td>Delaneau et al.(^{56})</td>
<td><a href="https://mathgen.stats.ox.ac.uk/">https://mathgen.stats.ox.ac.uk/</a> genetics_software/shapeit/shapeit.html</td>
</tr>
<tr>
<td>ChromoPainter / ChromoCombine</td>
<td>Lawson et al.(^{19})</td>
<td><a href="https://people.maths.bris.ac.uk/~mdjll/chromopainter_info.html">https://people.maths.bris.ac.uk/~mdjll/chromopainter_info.html</a></td>
</tr>
<tr>
<td>fineSTRUCTURE v.2.0.7</td>
<td>Lawson et al.(^{19})</td>
<td><a href="https://people.maths.bris.ac.uk/~mdjll/finestructure-old/">https://people.maths.bris.ac.uk/~mdjll/finestructure-old/</a></td>
</tr>
<tr>
<td>SOURCEFIND</td>
<td>Chacón-Duque et al.(^{57})</td>
<td><a href="https://github.com/hellenthal-group-UCL/sourcefindV2">https://github.com/hellenthal-group-UCL/sourcefindV2</a></td>
</tr>
<tr>
<td>fastGLOBETROTTER</td>
<td>Hellenthal et al.(^{38})</td>
<td><a href="https://github.com/sahwa/fastGLOBETROTTER">https://github.com/sahwa/fastGLOBETROTTER</a></td>
</tr>
<tr>
<td>RFMix v.1.5.4</td>
<td>Maples et al.(^{23})</td>
<td><a href="https://github.com/indraniel/rfmix">https://github.com/indraniel/rfmix</a></td>
</tr>
<tr>
<td>BWA v.0.7.13</td>
<td>Li, and Durbin(^{59})</td>
<td><a href="https://github.com/broadinstitute/bwa/releases">https://github.com/broadinstitute/bwa/releases</a></td>
</tr>
<tr>
<td>GATK v.3.8</td>
<td>Van der Auwer, and O’Connor(^{60})</td>
<td><a href="https://github.com/broadinstitute/gatk/releases">https://github.com/broadinstitute/gatk/releases</a></td>
</tr>
<tr>
<td>MToolBox</td>
<td>Calabrese et al.(^{61})</td>
<td><a href="https://github.com/mitoNGS/MToolBox">https://github.com/mitoNGS/MToolBox</a></td>
</tr>
<tr>
<td>Yleaf</td>
<td>Ralf et al.(^{62})</td>
<td><a href="https://github.com/genid/Yleaf">https://github.com/genid/Yleaf</a></td>
</tr>
</tbody>
</table>
RESOURCES 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 and Antoine Gessain in creole14 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.

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 QIAGEN 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 2,269,888 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.).

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.). 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 and worldwide populations from the SGDP project. We also merged the dataset with SNP array data from Oceanian populations 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 aDNA samples were projected on PCs by using the “lsqproject” and “shrinkmode” options. We inferred population structure with ADMIXTURE v.1.3.0 using 10 different seeds and assuming that K varies from 2 to 17, and visualized ADMIXTURE results using pong v.1.4.7. We computed f2-statistics with ADMIXTOOLS v.6.0 and estimated standard errors by block jackknife.

Haplotype phasing
We phased the dataset using SHAPEIT v.256,63 using 500 conditioning states, 10 burn-in steps, 50 Markov chain Monte Carlo (MCMC) main steps, a window size of 1 Cm and an effective population size of 15,000. We used 1000 Genomes Project data 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 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 analyses with ChromoPainter. In Analysis 1, we inferred the genetic structure of ni-Vanuatu running ChromoPainter only for the un-
samples that cluster with $i$ in the final state and in each of the MCMC samples, for each inferred cluster $k$. Finally, we computed $P_i = \sum_m [y^{(m)}_i / y^{(m)}]$. 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 $k = 2, \ldots, K$, $K$ 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 and for each $k = 2, \ldots, K$. 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 $x^{(s)}_i$, where $s = 1, \ldots, 5$, which corresponds to the number of individuals clustering with individual $i$ for each seed. We then estimated $y^{(s)}_i$ and $\frac{\sum_s [y^{(s)}_i / x^{(s)}_i]}$, 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, or when the studied populations are sampled from a small geographic region, 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 SOURCEFINET. 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 SOURCEFINET 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. 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” SOURCEFINET 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.), 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.) 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 telo-
meres. We then estimated East Asian-related and Papuan-related ancestry in the autosomes and the X chromosome, separately. We estimated \( a_i \) and \( d_{im} \), 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.\(^3\) In such a case, \( a_i = 3a_{XX} - 2a_{auto} \) and \( d_{im} = 4d_{auto} - 3a_{XX} \), where \( a_{auto} \) and \( a_X \) 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 ~30x coverage.\(^16\) 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,\(^15\) 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\(^16\): 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 ampicon 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 respectively, by assuming that admixture proportions have reached equilibrium values.\(^3\) In such a case, \( a_i = 3a_{XX} - 2a_{auto} \) and \( d_{im} = 4d_{auto} - 3a_{XX} \), where \( a_{auto} \) and \( a_X \) are the average Papuan-related ancestry proportions estimated on the autosomes and the X chromosome, respectively.

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 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 S1 G). 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 S1 H). 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, for each island. In each permutation, we sampled equally many random pairs as there were observed pairs in the island. We but also tested how the inclusion of these individuals affect the results (Methods S1 H). 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 \( P(Y_i = 1) \), and introduce \( \phi_i \) as the kinship coefficient between individuals in the \( i \)th pair. We estimated the effect of kinship on mate choice by the parameter \( \beta^r \) in the logistic regression model

\[
\log \left( \frac{P_i(x)}{1 - P_i(x)} \right) = \mu + \phi \beta^r + I(\text{village}_i)\beta^v + I(\text{island}_i)\beta^l. \tag{Equation 1}
\]

where \( I(\text{village}_i) = 1 \) if individuals of pair \( i \) are from the same village and \( I(\text{village}_i) = 0 \) otherwise, and \( I(\text{island}_i) = 1 \) if individuals of pair \( i \) are from the same island and \( I(\text{island}_i) = 0 \) otherwise. As a sensitivity analysis, we also considered a model in which pairs of first-degree related individuals were included (Methods S1 I).
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 \( \beta^q \) in the logistic regression model,

\[
\log \left[ \frac{p_i(x)}{1 - p_i(x)} \right] = \mu + q_{EA} \beta^q + q_{VA} \beta^v + I(village_i) \beta^v + I(island_i) \beta^i,
\]

(Equation 2)

where we use the same notations as in Equation 1. A negative (positive) effect size \( \beta^q \) 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) \( d_{is} \) as \( d_{is} = 0 \) if both alleles are identical between individuals, \( d_{is} = 1 \) if only one allele is identical and \( d_{is} = 2 \) if none of the alleles are identical. We estimated the association between SNP \( s \) and mate choice by the parameter \( \beta_{ASD}^s \) in the model

\[
\log \left[ \frac{p_i(x)}{1 - p_i(x)} \right] = \mu + d_{is} \beta_{ASD}^s + q_{EA} \beta^q + q_{VA} \beta^v + I(village_i) \beta^v + I(island_i) \beta^i,
\]

(Equation 3)

where we use the same notations as in Equation 2. A negative (positive) effect size \( \beta_{ASD}^s \) 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 \( y_s = \beta_{ASD}^s \), where \( \beta_{ASD}^s \) 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 \( \beta_{trait}^s \) in the model

\[
y_s = \mu + I(associated_s) \beta_{trait}^s + MAF_s \beta_{MAF} + GERP_s \beta_{GERP} + rec_s \beta_{rec},
\]

(Equation 4)

where \( I(associated_s) = 1 \) if SNP \( s \) is significantly associated with the candidate trait in GWAS (with GWAS P-value < 5 × 10^{-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 \( \beta_{trait}^s \) is significantly negative (positive). We adjusted p values with the Bonferroni correction for multiple testing, to account for the number of traits tested.
Supplemental Information

The genomic landscape of contemporary western Remote Oceanians

Lara R. Arauna, Jacob Bergstedt, Jeremy Choin, Javier Mendoza-Revilla, Christine Harmant, Maguelonne Roux, Alex Mas-Sandoval, Laure Lémée, Heidi Colleran, Alexandre François, Frédérique Valentin, Olivier Cassar, Antoine Gessain, Lluis Quintana-Murci, and Etienne Patin
Figure S1. Genetic structure of ni-Vanuatu and a selection of worldwide populations, Related to Figure 1.

(A) ADMIXTURE analyses of the SNP array data for 1,439 ni-Vanuatu, together with whole genome sequences of other populations from the Pacific region\textsuperscript{18}.

(B) Principal Component Analyses of the SNP array dataset for ni-Vanuatu. Each point is an individual, colored according to the geographical latitude coordinates of their island of residence, as in Figure 1A. Percentages indicate the proportion of variance explained. PC1 and PC2 are shown in Figure 1C. Black crosses and points indicate projected ancient samples from Vanuatu\textsuperscript{9,13} dated to the Lapita and Post-Lapita periods, respectively.
Figure S2. Fine-scale genetic clusters in Vanuatu and ancestry proportions, Related to Figure 2.

(A) Fine-scale genetic clusters according to fineSTRUCTURE, from $K_{FS} = 2$ to 18. Each point indicates an individual, located according to their village of residence. Colors indicate clusters, so that the closer the colors, the closer the clusters. Noise was added to sampling locations to facilitate visualization.

(B) Ancestry proportions in ni-Vanuatu, estimated by SOURCEFIND. Each point indicates an individual, colored according to their ancestry proportions. The East Asian-related ancestry is the sum of the Taiwan.Philippines and Polynesian ancestry. Noise was added to sampling locations to facilitate visualization.
Figure S3. Allele sharing of ni-Vanuatu with Polynesians, Related to Figure 3.
The x-axis indicates $f_4(X, \text{New Guinean Highlanders}; \text{EastAsia}, \text{Australian})$, which measures allele sharing of ni-Vanuatu and Near Oceanians (X) with East Asians. The y-axis indicates $f_4(X, \text{EastAsian}; \text{Polynesian, Tolai})$, which measures allele sharing of ni-Vanuatu and Near Oceanians (X) with Polynesians (Rennell and Bellona [RenBel], Tikopia, Tonga), relative to their East Asian ancestry. The regression line is estimated considering only modern individuals from Vanuatu. The bars show two standard errors. Deviations upper from the regression line indicate a higher affinity with the Polynesian population tested. Black crosses and points indicate projected ancient samples from Vanuatu 9,13 dated to the Lapita and Post-Lapita periods, respectively.
Figure S4. Proportions of female and male ni-Vanuatu migrants, Related to Figure 4. Migrations were defined based on genetic clusters (left column) or self-reported information (right column) (STAR★Methods).
Figure S5. Genetic relatedness among contemporary ni-Vanuatu, Related to Figure 5.

(A) Comparison of kinship levels among ni-Vanuatu living in the same village, the same island or the entire archipelago. Significance of the differences was tested by a Wilcoxon test.

(B) Degrees of genetic relatedness of ni-Vanuatu from the different Vanuatu islands. Bar plots indicate the proportion of pairs of related individuals according to the inferred degree of genetic relatedness, in each island. The black points and indicate the mean kinship per island. Significance was tested by a Fisher’s exact test. *P < 0.05, **P < 0.01, ***P < 0.001.

(C) Genetic relatedness and islands of residence of related ni-Vanuatu. Each line connects two individuals who are related up to third degree.

(D) Genetic relatedness and islands of residence of ni-Vanuatu spouses. Each line connects the two spouses, and the color indicates the degree of genetic relatedness.

(C,D) Islands are order from north to south (clockwise).