The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data

Irina Pugach,*[,1] Ana T. Duggan,[2] D. Andrew Merriwether,[3] Françoise R. Friedlaender,[4] Jonathan S. Friedlaender,[4] and Mark Stoneking*[,1]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Department of Anthropology, McMaster University, Hamilton, Canada

[3]Department of Anthropology, Binghamton University, Binghamton, NY

[4]Department of Anthropology, Temple University, Philadelphia, PA

***Corresponding authors:** E-mails: irina_pugach@eva.mpg.de; mark_stoneking@eva.mpg.de.

Associate editor: Connie Mulligan

Abstract

A widely accepted two-wave scenario of human settlement of Oceania involves the first out-of-Africa migration circa 50,000 years ago (ya), and the more recent Austronesian expansion, which reached the Bismarck Archipelago by 3,450 ya. Whereas earlier genetic studies provided evidence for extensive sex-biased admixture between the incoming and the indigenous populations, some archaeological, linguistic, and genetic evidence indicates a more complicated picture of settlement. To study regional variation in Oceania in more detail, we have compiled a genome-wide data set of 823 individuals from 72 populations (including 50 populations from Oceania) and over 620,000 autosomal single nucleotide polymorphisms (SNPs). We show that the initial dispersal of people from the Bismarck Archipelago into Remote Oceania occurred in a "leapfrog" fashion, completely by-passing the main chain of the Solomon Islands, and that the colonization of the Solomon Islands proceeded in a bidirectional manner. Our results also support a divergence between western and eastern Solomons, in agreement with the sharp linguistic divide known as the Tryon–Hackman line. We also report substantial post-Austronesian gene flow across the Solomons. In particular, Santa Cruz (in Remote Oceania) exhibits extraordinarily high levels of Papuan ancestry that cannot be explained by a simple bottleneck/founder event scenario. Finally, we use simulations to show that discrepancies between different methods for dating admixture likely reflect different sensitivities of the methods to multiple admixture events from the same (or similar) sources. Overall, this study points to the importance of fine-scale sampling to understand the complexities of human population history.

Key words: **human migration, genetic admixture, demographic inference, evolutionary biology, genetic variation.**

Introduction

The Pacific is a vast region, encompassing an entire hemisphere of our planet, and the human settlement of the far-flung Pacific Islands has long been of intense interest. A convenient division of the Pacific consists of Near and Remote Oceania, with the border located between Makira, the most easterly island of the main Solomon Island Archipelago, and the islands of Santa Cruz (politically part of the Solomon Islands), Vanuatu, and New Caledonia (supplementary fig. S1, Supplementary Material online). Fossil, archaeological, and genetic evidence indicate that Near Oceania was initially colonized by 45,000–50,000 years ago (ya) (Groube et al. 1986; Roberts et al. 1990; O'Connell and Allen 2015; Malaspinas et al. 2016) or possibly even earlier (Clarkson et al. 2017); at this time, sea levels were considerably lower and most of Near Oceania (Australia, New Guinea, and many of the nearby islands) were connected as a single landmass called Sahul. However, Sahul was never connected to the continental Asia landmass (Sunda), which at the time encompassed the present islands of Sumatra, Java, and Borneo, and so humans had to cross water in order to colonize Sahul. These crossings would have been mostly intervisible, meaning that there would have been some indication of land ahead before losing sight of the land behind, and so may not have required particularly sophisticated boating technology or navigational skills. In contrast, the initial colonization of Remote Oceania would have involved crossing at least 400 km of open ocean, and was only accomplished ~3,500 ya (Spriggs 2003); colonization of the other islands of Remote Oceania involved crossing thousands of kilometers of open ocean.

The languages spoken today in Near Oceania are classified as either Papuan or Austronesian. The indigenous Papuan languages are so deeply rooted and varied that the relationships between them are still poorly understood (Hunley et al. 2007; Pawley 2007); they do not constitute a language family in the usual sense that the languages can be demonstrably related and they are also quite distinct from the Austronesian languages. In contrast, Austronesian languages can be demonstrably related (Blust 1999; Gray et al. 2009). Their introduction to Oceania was part of an expansion that probably started from Taiwan ~5,000 ya (Bellwood 2004; Ko et al. 2014) and arrived in the Bismarck Archipelago ~3,400 ya, in association with the appearance of Lapita pottery (Spriggs 2003; Bellwood and Dizon 2005). This Austronesian expansion then continued to, and was the first to cross, the border to Remote Oceania, arriving there by 3,200 ya (Kirch 2000; Bellwood 2004).

While there is growing consensus for the above "dual-wave" model for the colonization of Near and Remote Oceania, many questions remain. For example, there is an ~40,000–45,000 year hiatus and regional isolation between the initial colonization of Near Oceania and the Austronesian expansion; are there any detectable signals in the genomes of modern Oceanian populations of any other migrations during this time period? Whereas some studies of mtDNA variation have claimed to find evidence for pre-Austronesian contact between Southeast Asia and Near Oceania (Soares et al. 2011), the most detailed study to date of complete mtDNA genome sequences from Oceania estimated that mtDNA sequences attributable to any such pre-Austronesian contact constitute at most 2% of the current Oceanian mtDNA gene pool (Duggan et al. 2014). Y chromosome studies conducted to date in Near Oceania lack sufficient resolution to address the issue of pre-Austronesian migrations (Kayser et al. 2000, 2006; Delfin et al. 2012), while a recent analysis of genome-wide single nucleotide polymorphism (SNP) data did not find any evidence for pre-Austronesian contact (Bergström et al. 2017).

Another question concerns the timing of the contact that occurred between the Austronesian-speaking groups and the resident populations of Near Oceania (referred to here collectively as Papuans, while not implying any cultural, linguistic, or biological unity of these populations), prior to the colonization of Remote Oceania. The contemporary populations of Remote Oceania all carry both Asian-related and Papuan-related ancestry (Kayser et al. 2000, 2006, 2008; Friedlaender et al. 2008; Wollstein et al. 2010; Skoglund et al. 2016), and various methods for dating this signal of admixture provide estimates of ~3,500 ya (Wollstein et al. 2010; Pugach et al. 2011). This suggests that admixture occurred shortly after the arrival of Austronesian speakers in Near Oceania (most likely in the Bismarck Archipelago) and prior to the colonization of Remote Oceania. However, a recent study of ancient DNA obtained from skeletons dating to the initial occupation of Vanuatu and Tonga did not find any evidence of Papuan ancestry, and moreover dated the admixture (using the ALDER method) in contemporary Tongans to only ~1,500–2,300 ya (Skoglund et al. 2016). These results suggest that the initial colonization of Remote Oceania was by people whose ancestors had not yet mixed with Papuans, and that Papuan ancestry was introduced to Remote Oceania by more recent migrations. Indeed, it had previously been observed that there is higher and more variable amounts of Papuan ancestry in Fiji than elsewhere in Remote Oceania, which can be explained by additional migration(s) of Papuan people that reached as far as Fiji (Wollstein et al. 2010). Nonetheless, there is still a conflict between the dates of the Papuan-Asian admixture signal in current Polynesians of 1,500–2,300 ya estimated by the ALDER method (Skoglund et al. 2016) versus dates of ~3,500 ya estimated by other methods (Wollstein et al. 2010; Pugach et al. 2011).

A further question is whether the movement of people from Near to Remote Oceania followed a simple wave-of-advance model (i.e., proceeding from New Guinea to Santa Cruz/Vanuatu/New Caledonia via the main Solomon Islands archipelago, and then from this region to the rest of Remote Oceania), or whether movement was more complicated. There are already some indications that migration was more complicated than a simple wave-of-advance model; for example, obsidian from the Bismarck Islands is found in archaeological sites on Santa Cruz but not elsewhere in the Solomon Islands (Sheppard and Walter 2006), suggesting contact between Santa Cruz and the Bismarcks that bypassed the main Solomon Island archipelago. The Santa Cruz islands lie beyond the easternmost part of the Solomon Islands chain, and are separated from the other islands by 400 km of open ocean. Santa Cruz are thus the westernmost islands of Remote Oceania, colonized only when the Lapita people introduced navigational skills and technology which made long-distance ocean voyaging possible, with the earliest archaeological sites suggesting human occupation at ~3.2 kya (Gross 2016). The languages of Santa Cruz belong to a primary branch of Oceanic languages (Næss 2006; Ross and Næss 2007; Næss and Boerger 2008) and are so distantly related to other Austronesian languages in the Solomons, that they were originally classified as Papuan with perhaps some Austronesian substrate (Wurm et al. 1978). Santa Cruz is also unusual in having a much higher frequency of Papuan mtDNA and Y chromosome haplogroups than anywhere else in Remote Oceania (Friedlaender et al. 2002; Delfin et al. 2012; Duggan et al. 2014), and the diversity of Papuan mtDNA sequences is too high to be explained by a simple bottleneck or founder event (Duggan et al. 2014). Instead, it appears as if there was substantial contact between Santa Cruz and Near Oceania that did not penetrate much further into Remote Oceania.

To address these and other issues concerning the relationships among and between populations of Near and Remote Oceania, we have assembled a data set of 823 individuals from 72 populations, including 50 populations from Near and Remote Oceania (supplementary table S1, Supplementary Material online), all genotyped for ~620,000 SNPs on the Affymetrix Human Origins Array (Patterson et al. 2012). Our analyses indicate substantial complexity and waves of contact during the movement of people through Near Oceania to Remote Oceania. We also use computer simulations to address the discrepancy in dates for the Papuan/Asian admixture signal obtained by ALDER versus other methods, and show that this is likely to reflect the different effect of multiple episodes of admixture from the same (or similar) source populations on these methods. Overall, our study highlights how combining sampling on a fine scale with dense genome-wide data can illuminate new aspects of the peopling of the Pacific.

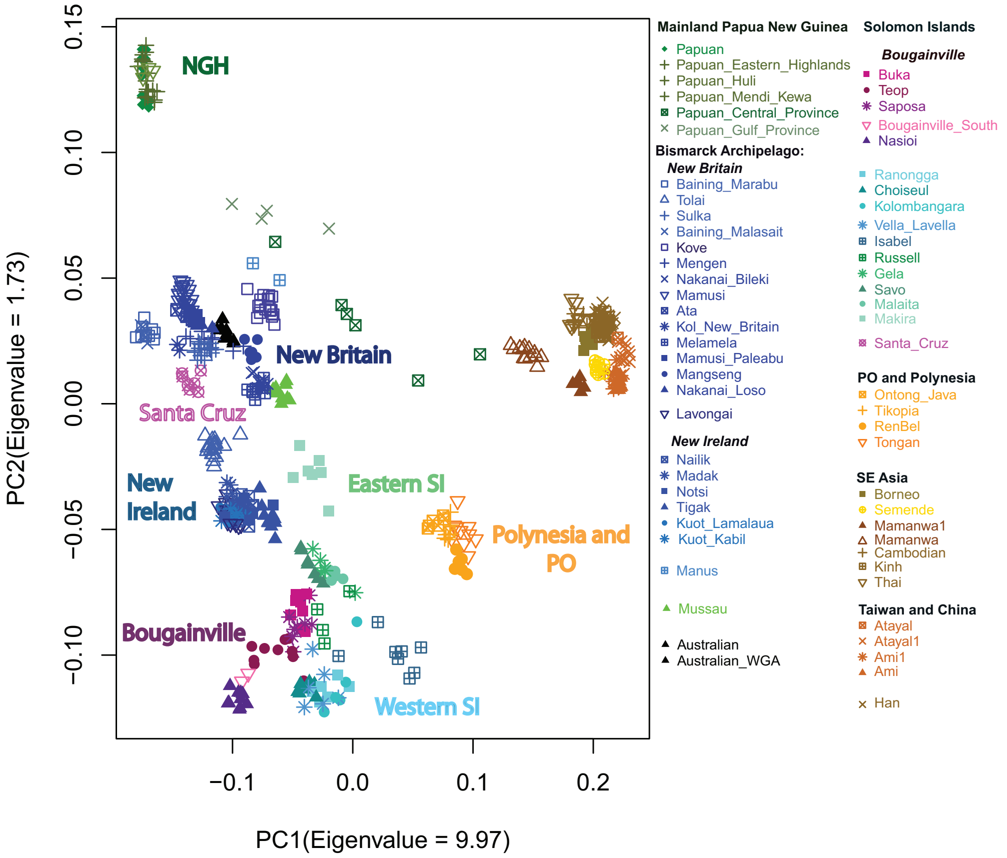

Fig. 1. Results of the PC analysis showing the genetic structure captured by the first two principal components. Each colored label represents an individual, color palettes are only used to identify a general geographic location (dark green = mainland Papua New Guinea [PNG], blue = Bismarck Archipelago, purple = Bougainville, aqua = Solomon Islands, except Santa Cruz, magenta = Santa Cruz, orange = PO and Polynesia, brown = Asia).

Results and Discussion

Large-Scale Population Structure

To obtain an overview of the data structure and potential associations between the samples in the data set, we applied two different approaches: principal components analysis (PCA), based on individual SNPs as well as on segments shared identical-by-descent (IBD), and the clustering algorithm ADMIXTURE (Alexander et al. 2009). PCA based on individual SNPs and performed on the entire data set (supplementary fig. S2, Supplementary Material online) highlights variation between main continental groups, while when we limit the analysis to populations from Oceania and Southeast Asia (fig. 1), we observe that the second main axis of variation is driven by the differences between the PNG highlanders (NGH) and Bougainville, with populations from the Bismarcks and Australia falling in the middle. Also, although it has been previously shown that when applied to human genetic data from some geographic regions, such as Europe, the first two principal axes correlate closely with geography (Lao et al. 2008; Novembre et al. 2008), we do not observe such correlation for populations of the Solomon Islands (SI). Notably, the samples from Santa Cruz (geographically in Remote Oceania) and the easternmost islands of the SI chain fall directly adjacent to populations from island New Britain (the largest island in the BA, which is geographically located northwest of Bougainville; fig. 1 and supplementary fig. S1, Supplementary Material online). Genetic proximity of eastern SI and the BA continues to be seen in higher PCs (supplementary fig. S3, Supplementary Material online) and in the analysis limited to the Oceanian populations only (supplementary fig. S4, Supplementary Material online).

PCA based on segments shared IBD (see Materials and Methods for details) indicates substantial haplotype sharing (fig. 2). Since the detection of IBD segments based on genotype data, rather than sequence data, is biased towards detection of longer fragments (Ralph and Coop 2013), we expect this analysis to reveal the history of recent gene flow (occurring mostly in the past 4,000 years; Ralph and Coop 2013). Notably, however, this signal of close recent relatedness is not observed for samples from Santa Cruz, which instead continue to exhibit closer distances to populations from New Britain (BA) and NGH (fig. 2).

Next, we used ADMIXTURE (Alexander et al. 2009), which we first applied to the complete data set (supplementary figs. S5A, S6A, and S7, Supplementary Material online), and the 13 separate ascertainment panels of the Human Origins Array (Patterson et al. 2012; supplementary fig. S8, Supplementary Material online). Since both analyses revealed little impact of Eurasian populations on genetic structure in Oceania, we repeated the analysis for the Oceanian data only

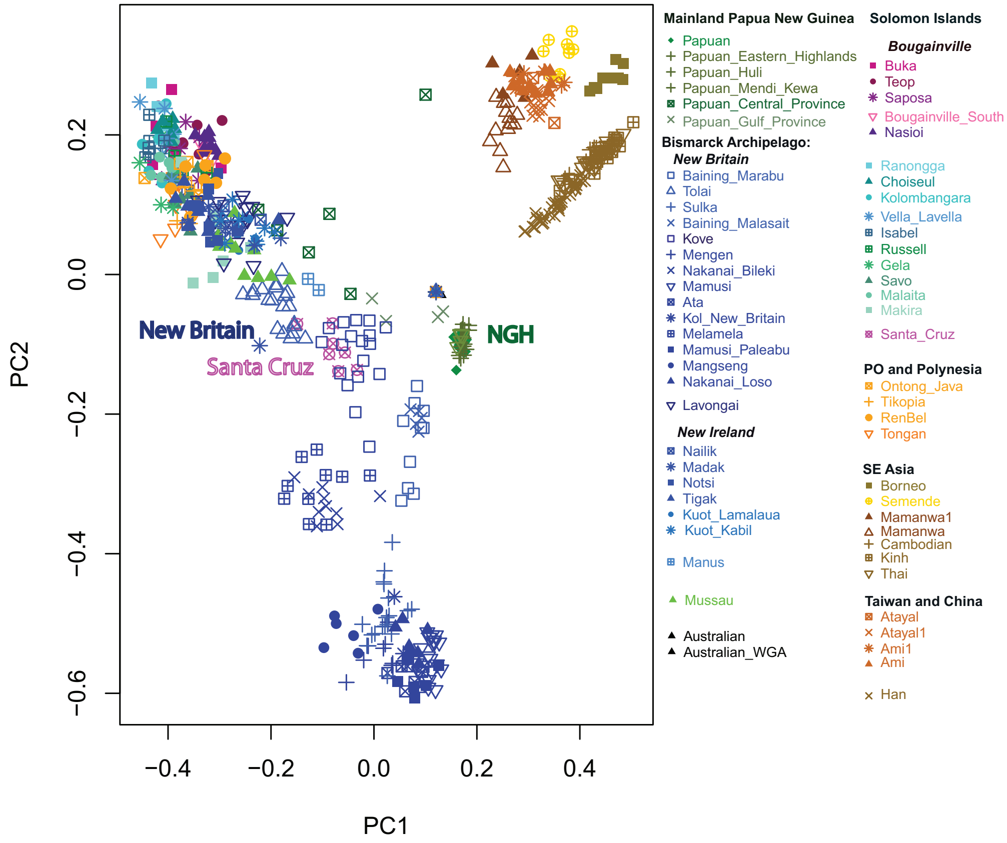

Fig. 2. Results of the PC analysis showing recent relatedness based on the number of IBD blocks shared between individuals. Each colored label represents an individual, colors are only used to identify a general geographic location (dark green = mainland Papua New Guinea, blue = Bismarck Archipelago, purple = Bougainville, aqua = Solomon Islands, except Santa Cruz, magenta = Santa Cruz, orange = PO and Polynesia, brown = Asia).

(supplementary fig. S5B, Supplementary Material online). For this subset the lowest cross-validation error is seen for $K = 5$ (supplementary fig. S6B, Supplementary Material online). Similar to the results obtained for the full data set (supplementary fig. S7, Supplementary Material online), the five Oceanian components could broadly be assigned to: New Guinea (dark green; seen at highest frequency in NGH): Polynesia (orange; seen at highest frequency in Polynesian outliers and Tonga), Bougainville (aqua; seen at highest frequency in Nasioi), and two components found at highest frequency in the BA, namely in the Baining (blue) and Mamusi (pink) populations of New Britain (supplementary fig. S9, Supplementary Material online).

When the results of the ADMIXTURE analysis for each population are superimposed onto a geographic map (fig. 3), several patterns become apparent. Firstly, the Papuan ancestry component (dark green) is found at high or appreciable frequency only in the west—on the main island of New Guinea, in Australia, and in some coastal populations of New Britain; it is almost completely absent from New Ireland and does not appear at all across the SI, with the exception of the easternmost islands of the chain—Makira and Santa Cruz—where it appears at substantial frequency. The two components which appear at high frequency on New Britain (blue and pink) are always observed together; they are prevalent throughout the BA and are almost absent from the western SI. However, they reach a substantial frequency in the eastern SI, accounting for >50% of the inferred ancestry in Santa Cruz, ~30% in Makira, and ~20% in other populations of the eastern SI. The aqua component, which is present at high frequency on Bougainville, is ubiquitous east of mainland New Guinea and seems to exhibit a clinal pattern, decreasing in frequency with increasing distance from Bougainville. The Polynesian component (orange) also seems to exhibit a geographical gradient, with its frequency diminishing in an east to west direction. Some small islands (e.g., Manus, Mussau) and coastal populations (e.g., Kove, Melamela) from western New Britain do not follow this cline, having a third of their ancestry assigned to this Polynesian component. An exception at the opposite extreme are individuals from Santa Cruz, who have very little (on average 5%) of this component. In addition, the aboriginal Australian samples reveal a baffling signal of admixture, showing both Papuan and Bougainville ancestry components in roughly equal proportion; the interpretation of this signal is considered below. Australian Aboriginals are assigned their own ancestry component when the value of K is increased to seven (supplementary fig. S5B, Supplementary Material online).

Taken together these results tentatively suggest different historical trajectories for western versus eastern SI (figs. 1–3); confirm the population of Santa Cruz as a genetic outlier with

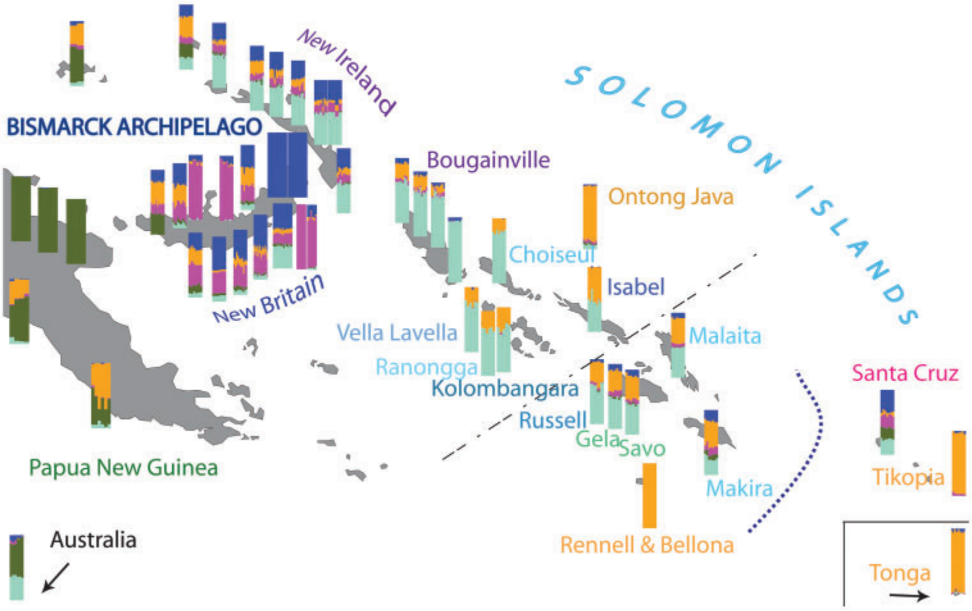

Fig. 3. ADMIXTURE results for $K = 5$ showing the approximate location of the Oceanian populations included in this study. For reasons of space the location of aboriginal Australians and Tongans does not correspond to their true location (which can be seen in supplementary fig. S1, Supplementary Material online). The curved dotted line marks the biogeographic boundary between Near and Remote Oceania, whereas the straight broken line denotes the Tryon–Hackman linguistic divide.

respect to its neighbors (figs. 1 and 3), as found in previous studies of uniparental markers (Friedlaender et al. 2002; Delfin et al. 2012; Duggan et al. 2014); and suggest both long-term isolation for some populations (e.g., Baining, Mamusi, Nasioi, and NGH) and ubiquitous recent post-Austronesian gene flow for others (SI, except Santa Cruz and the PO, fig. 2). However, because both PCA and ADMIXTURE are descriptive analyses that do not directly inform about the underlying historical processes, in the subsequent sections we apply additional methods to first validate and then to expand these tentative findings.

Evaluating Diverse Signals of Admixture in Oceania

Demographic processes not involving admixture, such as drift (Hudjashov et al. 2017), recent bottlenecks (Falush D, van Dorp L, Lawson D, unpublished data) as well as uneven sampling (Puechmaille 2016) could produce ADMIXTURE results indistinguishable from those which reflect actual admixture. To evaluate signals of admixture in Oceanian populations we first formally tested for admixture using the 3-population test (Patterson et al. 2012) followed by a TreeMix (Pickrell and Pritchard 2012) analysis. For those populations exhibiting significant signals of admixture, we inferred the order and dates of admixture events and explored further these signals in the patterns of sharing of segments inherited IBD.

3-Population Test and TreeMix

We first applied the 3-population test (Patterson et al. 2012) in the form of f3(C; A, B), which tests for "treeness," and where a significantly negative value of the f3 statistic would imply that population C has descended from an admixture event between A and B. We find signals of admixture for most, but not all, Oceanian populations (supplementary table S2, Supplementary Material online). The most significantly negative f3 values are observed when the admixture event involves as parental sources NGH, Baining, or Nasioi populations (all of which are assigned their own ancestry component in the ADMIXTURE analysis) and Taiwanese Aboriginals, probably reflecting the autochthonous ancestry and the Asian ancestry source associated with the Austronesian expansion, respectively (Spriggs 2003). If we search for other signals of admixture by excluding Asian sources, the most significant results are obtained with either Tonga or Isabel (western SI) as one of the sources (except for Papuans and Saposa, where the most significant result is obtained with Sulka and Ontong Java, respectively). This could reflect ancestry associated with gene flow postdating the Austronesian expansion, but since this result is confounded by the fact that both Tongans and Isabel show evidence of admixture from Taiwan, it remains to be further evaluated whether or not this signal reflects additional gene flow. In supplementary table S2, Supplementary Material online, we report for each population the top three results with the highest significance score, followed by the top three results when the source populations do not include Asia. Importantly, admixture was neither inferred via the f3 statistic for any of the populations which have been ascribed their own component in the ADMIXTURE analysis at $K = 5$ (namely NGH, Baining and Mamusi groups, Nasioi, Tikopia and Rennell and Bellona), nor was it inferred for the aboriginal Australians.

Next, we used TreeMix (Pickrell and Pritchard 2012) to build a population tree and identify admixture events (which in this approach are added sequentially to the constructed tree until the residuals between the observed data and the fitted data are reduced). First, we performed the TreeMix analysis on the full data set of 75 populations, adding the Yorubans as an outgroup. The inferred maximum-likelihood tree largely groups the populations as expected (supplementary fig. S10*A*, Supplementary Material online), but with Santa Cruz grouping with populations from New Britain, and not the Solomon Islands. Positive residuals of the standard error capture pairs of populations that are more closely related to each other than is suggested by the inferred tree, and are thus candidates for gene flow (Pickrell and Pritchard 2012). The residuals (supplementary fig. S10*B*, Supplementary Material online) show the greatest error for: 1) the Australian Aboriginals versus Indian and European populations; 2) the Australian Aboriginals versus the Papuan Highlanders; 3) Papuan Highlanders versus coastal populations of New Guinea, Kove from the coast of northwest New Britain and Manus—the largest of the Admiralty Islands to the north of PNG; 4) western SI versus Bougainville; 5) Taiwanese Aboriginals, Mamanwa (Philippines), Semende (Sumatra) and to a lesser extent island southeast Asia in general versus Polynesians and Isabel; and 6) Papuan Gulf Province versus Europeans. However, when we start adding migration edges to the tree the results appear to be meaningless (supplementary fig. S11*A*, Supplementary Material online). This is most likely explained by the large number of populations that TreeMix is trying to fit simultaneously, which is prohibitive for disentanglement of complex admixture histories (Lipson et al. 2013). Indeed, when we reduce the data set by omitting the populations from the BA, the migration edges inferred by TreeMix recover all the purported signals of admixture, such as gene flow from the Taiwanese Aboriginals to Polynesians, from NGH to the coastal groups and to the eastern SI, and from Bougainville to the western SI (supplementary fig. S11*B*, Supplementary Material online); importantly however, the migration edge involving NGH, Bougainville and Australian aboriginals is not inferred.

In summary, both the f3 statistics and TreeMix clearly identify the genetic impact of the Austronesian expansion on Oceanic populations, and TreeMix confirms many of the more local-scale admixture signals (while f3 results in this respect are harder to interpret, as they are confounded by the fact that most populations in Near Oceania are related through Austronesian ancestry). However, neither test infers NGH and Bougainville ancestry in aboriginal Australia, nor have any other studies of genome-wide data in aboriginal Australians (e.g., McEvoy et al. 2010; Pugach et al. 2013; Malaspinas et al. 2016). The likely explanation for this is as follows: in analyses such as ADMIXTURE, recent genetic drift will result in populations that are genetically distinct and hence are assigned pure ancestry components. If two such populations have independently undergone such drift, whereas a third related population has not, then the third population might be inferred to be admixed between the two drifted populations, simply because it has a larger population size and greater interindividual genetic variation (Falush D, van Dorp L, Lawson D, unpublished data). Thus, the most likely explanation for the apparent admixture in Australia is that Australia, NGH, and Bougainville all share ancestry but NGH and Bougainville experienced more genetic drift after population divergence. This enhanced drift is more easily detected as a separate ancestry component in the ADMIXTURE analysis, and since aboriginal Australians share ancestry with both ancestry components, they are (mistakenly) assigned as admixed.

Timing of Asian Admixture

To infer the date when Asia-related (presumably Austronesian-related) ancestry entered Near Oceania, we first used PCAdmix (Brisbin et al. 2012) to infer local ancestry along individual chromosomes for all populations in the BA, SI and for Tongans. For this inference PNG highlanders and either Taiwanese Aboriginals (as suggested by f3 and TreeMix analyses) or Polynesian Outliers (as suggested by ADMIXTURE) were used as the two source populations for the admixture. Once the ancestry blocks contributed by the two source populations were identified, we estimated the width of the ancestry blocks via wavelet transform (WT) analysis (Pugach et al. 2011), and the date of admixture was inferred based on the simulations described previously (Pugach et al. 2011). When Taiwanese Aboriginals are used as the source of Asian ancestry, the date of admixture inferred for all of the admixed populations falls between 77 and 104 (95% CI: 66–152) generations ago, which using 30 years for the generation time (Fenner 2005) corresponds to ∼2,300–3,100 years since admixture. If there was additional post-Austronesian gene flow from Polynesians, these dates are likely more recent than the dates of initial admixture, but are still within the chronological bounds of the Lapita period (Bellwood 2004) or just below them (fig. 4). Depending on the population, using Polynesians as the source of admixture results in admixture dates that are 150–1,000 years younger, which is not surprising given that the Polynesians are themselves of dual Asian-Papuan ancestry (e.g., Kayser et al. 2008; Friedlaender et al. 2008; Wollstein et al. 2010). The admixture date is substantially more recent for the Tongans (∼1,500 ya) because Tongans are themselves Polynesian, and there is very little differentiation between them and the Polynesian Outliers used as the parental group. With Taiwanese Aboriginals used as the parental group, the inferred date of admixture for Tongans is 83 (95% CI: 66–112) generations ago or circa 2,500 ya. This date of admixture is consistent with the dates of 90 generations (95% CI: 77–131; Pugach et al. 2011) and 99 generations (95% CI: 19–267; Wollstein et al. 2010) estimated previously by two independent methods for a diverse sample of Polynesian islands which included Tonga (Wollstein et al. 2010; Pugach et al. 2011). However, this date is older than an estimate of 50–80 generations ago obtained with the ALDER method (Loh et al. 2013) on the same Tongan samples (Skoglund et al. 2016). Similarly, all the admixture dates inferred using ALDER for the populations of Near Oceania appear to be 10–20 generations younger (Skoglund et al. 2016) than those inferred here using the WT method.

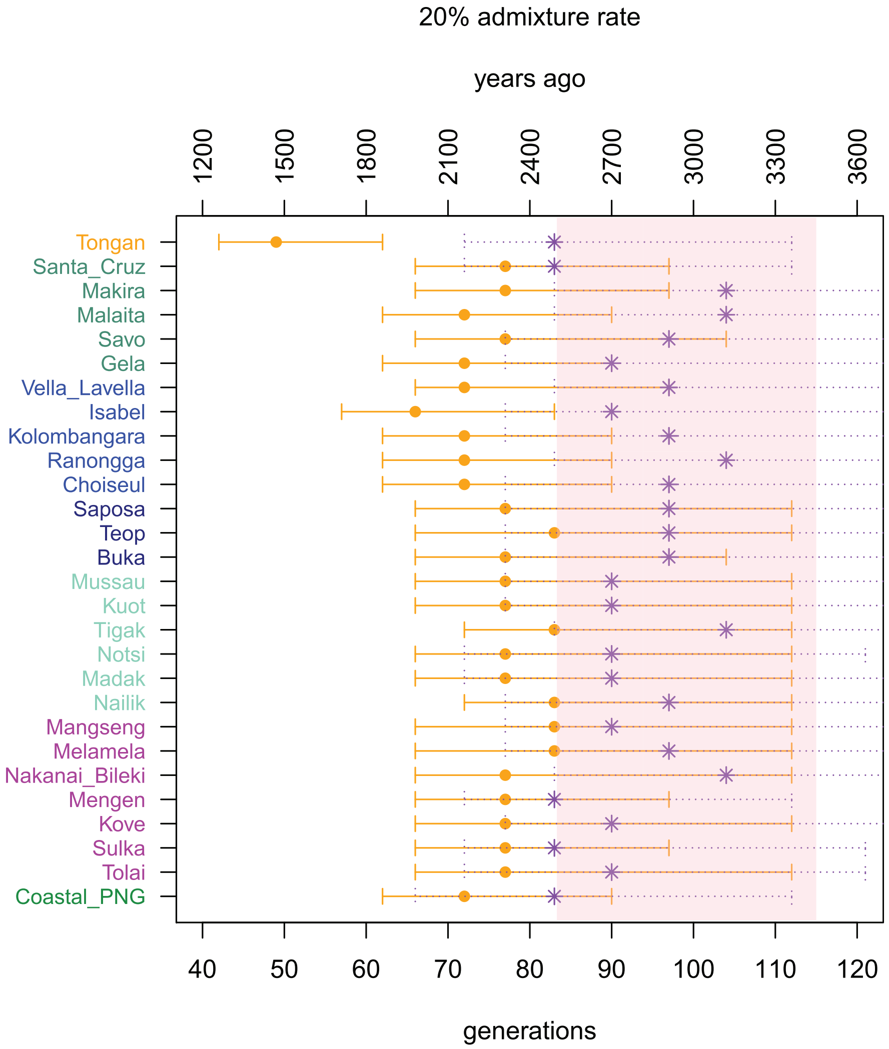

Fig. 4. The dates of admixture inferred via the Wavelet Transform Analysis. In assigning local ancestry along individual chromosomes, we used the PNG Highlanders as a proxy for the Near Oceanian ancestry, whereas for the Austronesian ancestry we either used the Taiwan Aboriginals (the inferred admixture dates are shown in purple) or the Polynesian Outliers (the inferred dates are shown in orange). Error bars represent the 95% confidence interval. The times of admixture are expressed in either generations or years, assuming a generation time of 30 years. The chronological bounds of the Lapita period are shown as a pink band.

To understand the source of this discrepancy we compared the WT method to ALDER using data simulated under a variety of different admixture scenarios. For this comparison, we used coalescent simulations implemented in the software MaCS (Chen et al. 2009). The basic simulation setup was as described previously (Pickrell et al. 2014 and supplementary fig. S12, Supplementary Material online). Briefly, for each admixture scenario we simulated data from nine populations, each consisting of 30 individuals with 10 independent chromosomes of 200 Mb. One of the simulated populations was the product of an admixture between two other populations in the simulated data set, with the admixture times set to 40, 70, 90, 100, 110, 120, 130, 140, 150, or 165 generations ago. Each simulated history was generated ten times. We then applied ALDER and the WT method to compare how each performs in recovering the simulated dates. Our results indicate that given a single episode of admixture, both methods perform reasonably well in recovering the simulated date (fig. 5). We then proceeded with the same simulation setup, but simulating two independent pulses of admixture from the same source population occurring at different time points. The first pulse of gene flow was set to have occurred 100, 120, or 140 generations ago, followed by the second pulse at 40, 60, or 90 generations ago, thereby simulating nine different admixture histories. As before, each of the admixture histories was simulated ten times. We then proceeded with ALDER and WT analysis to infer admixture dates. Note that while both methods have extensions which enable their use in situations involving multiple events of admixture from different sources (Pickrell et al. 2014; Pugach et al. 2016), here we are simulating multiple pulses of gene flow from the same source population. Our expectation is that a single estimate of the time of admixture would be inferred, and it would be a composite date reflecting both the older date of admixture as well as the more recent additional gene flow. The simulation results (fig. 5) indicate that in such situations ALDER tends to recover

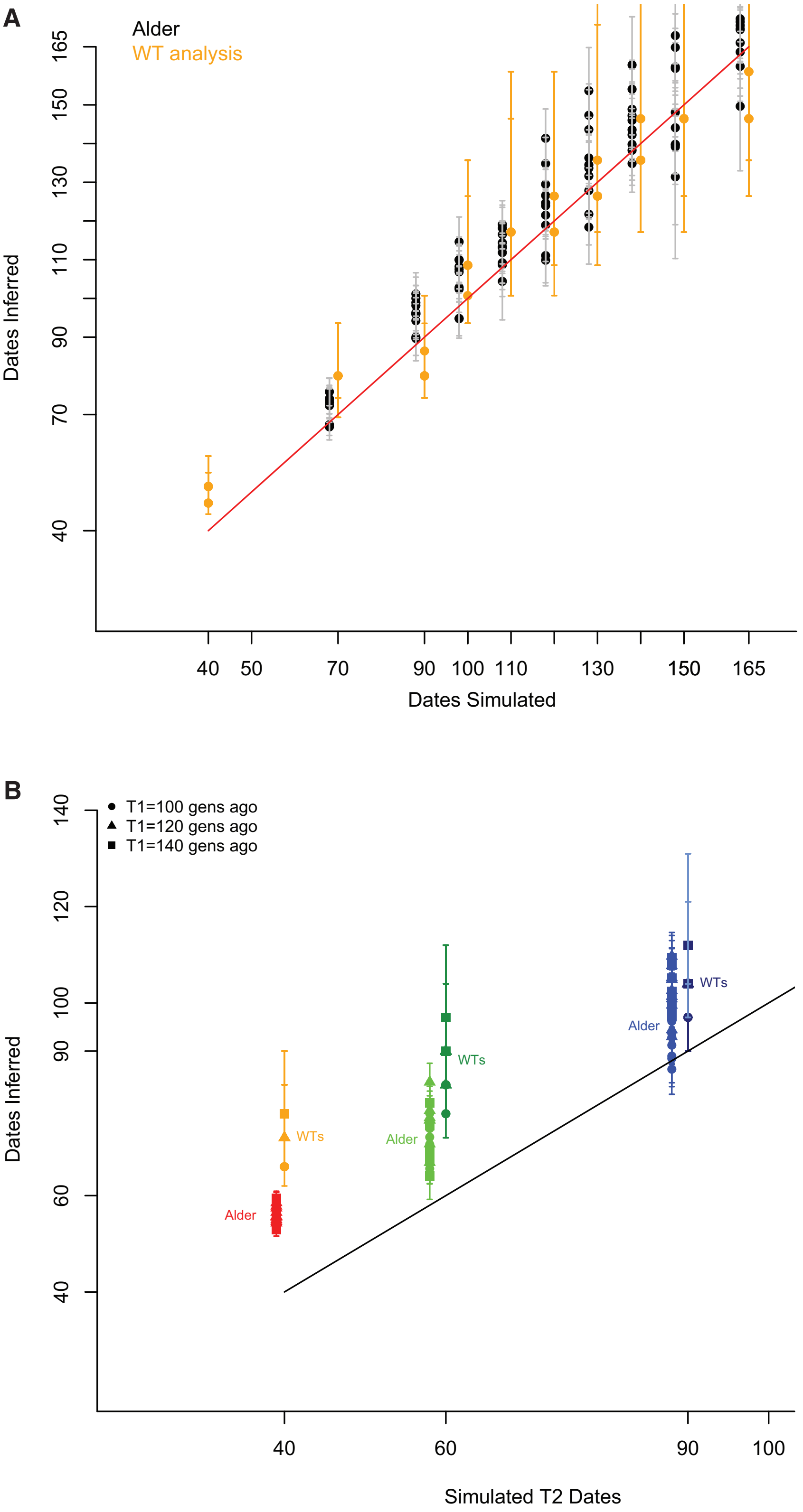

Fig. 5. Performance of the ALDER and the Wavelet Transform method in recovering simulated dates of admixture. Both methods were applied to (*A*) ten simulated data sets with different admixture histories, but involving only a single instantaneous event of gene flow occurring 40–165 generations ago, (*B*) nine simulated data sets, generated using two pulses of gene flow from the same source, the more recent episode occurring 40, 60, or 90 generations ago, with the earlier episode of admixture 100, 120, or 140 generations ago. For both (*A*) and (*B*) each admixture time point was simulated ten times. The error bars denote the 95% confidence interval. While running both methods a single event of admixture was assumed. The estimates from ALDER and the WT method are slightly off-set from each other for better visibility of the results.

a date which is close to the most recent date of admixture, whereas the date estimated with the WT method is closer to the mid-point between the two events. It therefore seems likely that the more recent dates of Austronesian gene flow inferred by ALDER versus the WT method for Near Oceania and Tonga reflect the fact that after the admixture resulting from the initial Austronesian expansion, most of these populations experienced additional gene flow from a population(s) related to Austronesians, that is, from Polynesia.

Recent Gene Flow Accessed through Sharing of IBD Segments

To more clearly elucidate the patterns of recent relatedness and local migrations postdating the Austronesian expansion, we analyzed IBD blocks, that is, segments of DNA that were inherited without recombination by each pair of individuals from their most recent common ancestor (Powell et al. 2010; Ralph and Coop 2013). Detection of these blocks based on genotype data, rather than sequence data, is biased towards detection of longer fragments, hence they are mainly informative about the events occurring in the past 4,000 years (Ralph and Coop 2013). We expect that recent gene flow between two populations would result in them sharing a higher number of longer IBD segments in comparison to others.

In terms of between population IBD sharing, we observe that most populations of New Britain (the largest island in the BA) show substantial sharing of IBD segments with their neighbors, namely populations from New Ireland (BA), Bougainville, and the eastern part of the SI chain, and also with the PO and Tongans. The exceptions are: 1) the Baining groups, who share only with the neighboring Sulka and Tolai, and with Santa Cruz (one of the Baining groups also shows sharing with Polynesians); and 2) the Nakanai Loso, Ata, and the Mamusi groups, who share only with their neighbors, and in agreement with the ADMIXTURE result (fig. 3) reveal no connection to New Ireland, the SI, or Polynesia (supplementary fig. S13*A*, Supplementary Material online). In general, all populations from across New Ireland share IBD segments extensively with their geographic neighbors (namely the Kuot groups, Nailik, Notsi, and Madak, which are located more centrally than the Tigak, who come from the northwestern tip of the island), as well as with Sulka, Tolai, and Kove from New Britain. The length of the IBD segments reflects that the contact must have been very recent. In contrast the length of the IBD segments shared with the SI and with Tonga and the PO suggests older ties; except for the Tigak who reveal a more recent sharing with Isabel (SI) and with Tongans. Like the coastal Tigak population, the Lavongai people from New Hanover, (located a few kilometers northwest of New Ireland) extensively share IBD segments with populations from New Ireland and New Britain; they also reveal a more recent contact with Bougainville and the western SI, and with the Polynesians (supplementary fig. S13*B*, Supplementary Material online). In contrast to populations from New Ireland, but similarly to Lavongai, the two populations residing on Manus and Mussau (the islands located furthest to the north) not only share IBD segments with the other populations from the BA, the Solomon Islands, PO and the Tongans, but also with the Papuans (consistent with results of the ADMIXTURE analysis, fig. 3). However, the length of shared IBDs is much lower in Manus, who share long IBDs only with the Nakanai Bileki from New Britain, with Isabel (SI) and with Tongans (supplementary fig. S13*C*, Supplementary Material online). Bougainville populations (Nasioi, Buka, Saposa, and Teop) show substantial sharing with each other and across the western SI, with little/moderate sharing with populations from eastern SI. Except for the Teop, they all share recent connections with the Tolai (New Britain) and Lavongai. Interestingly, Nasioi exhibits an equal amount of sharing with all populations of Polynesian ancestry, whereas the others share much longer tracts with the Tongans and Ontong Java. Similarly, the islands in the western SI (Ranongga, Choiseul, Kolombangara, and Vella Lavella; Isabel is discussed separately below) show substantial sharing with populations from Bougainville and with each other, but much less sharing with populations from the eastern SI. In addition, Choiseul and Kolombangara show recent links with the Tolai (New Britain). The western SI populations share more with Tonga than with the PO. In contrast, populations from the eastern SI (Russell, Gela, Savo, Malaita, and Makira) show more Papuan-related ancestry than the western SI and also exhibit more sharing with the BA than with Bougainville or the western SI (except Isabel, which is discussed below). They also do not show any particular recent links to Santa Cruz. In addition, like many other islands across the SI chain, Makira and Savo show a recent genetic relationship with the Tolai from New Britain (supplementary fig. S13*D*, Supplementary Material online).

Isabel and Santa Cruz depart from these general patterns of ancestry sharing for the SI. Isabel shows extensive recent genetic sharing with most populations from the BA and the SI (Santa Cruz being the notable exception), and with all sampled populations of Polynesian ancestry. Santa Cruz shares more segments IBD with New Britain than with New Ireland, Bougainville, or eastern or western SI (including Isabel). Peculiarly, Santa Cruz exhibits recent genetic ties with the Tolai, as do some other islands in the Solomons. Santa Cruz also shows equally low levels of sharing with all populations of Polynesian ancestry; this nonspecific sharing may indicate that Santa Cruz did not experience any additional post-Austronesian gene flow from Polynesia; by contrast, other populations exhibit ubiquitous sharing of long IBD segments with either Tongans (e.g., Sulka, Tigak, Choiseul, Savo), or Tongans and the PO (e.g., Buka, Saposa, Isabel, Makira), suggestive of contact postdating the Austronesian expansion (supplementary fig. S13*E*, Supplementary Material online).

In summary, the analyses based on the number and length of segments shared IBD highlight 1) a considerable recent post-Austronesian expansion gene flow, which is diverse both in terms of its sources and direction, across many of the islands in Near Oceania; 2) different patterns of sharing in western versus eastern Solomons (fig. 6), characterized by substantial local sharing of long IBD segments in the western Solomons versus fairly restricted local sharing in the eastern Solomons; 3) special roles of Isabel islanders and the Tolai people of the Gazelle Peninsula, New Britain, who show recent genetic ties with many populations across Oceania; 4)

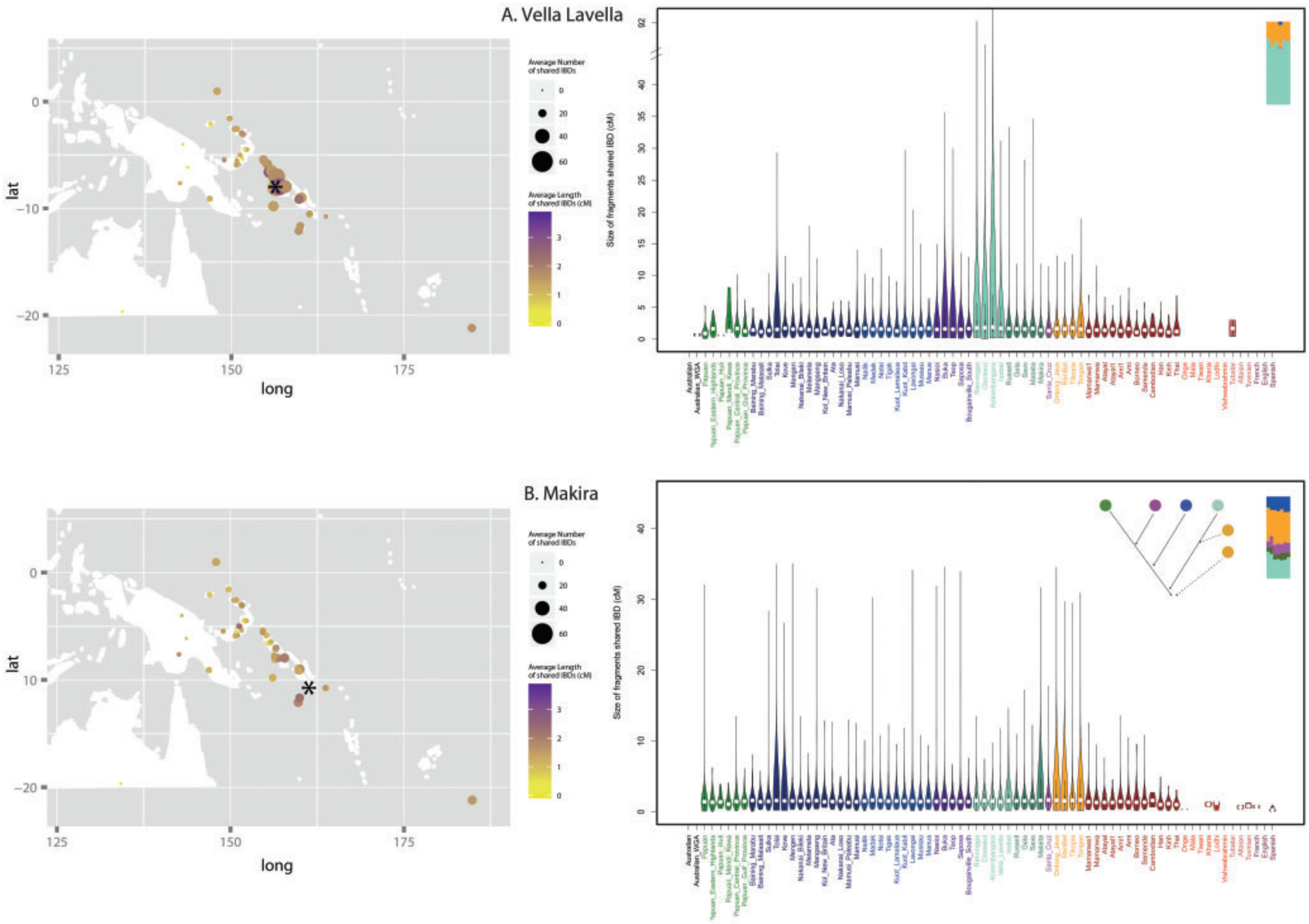

Fig. 6. Dissimilarity in the pattern of sharing of IBD blocks between populations from western (*A*) and eastern (*B*) Solomons. Left panel: each data point represents the results for the comparison of the population marked with an asterisk (*A*: Vella Lavella; *B*: Makira) to each of the other populations in the data set. Data points are placed on the map according to the sampling location of each population. The size of each circle is proportional to the mean number of IBD blocks shared, and the color intensity indicates the mean length of such shared blocks. Right panel: a violin plot, which displays the cumulative distribution of all ancestry blocks, inferred by the IBD analysis for the presented population (*A*: Vella Lavella; *B*: Makira); the plot captures the total abundance of blocks of each ancestry (x axis) of different genetic lengths in cM (y axis). The insets show an excerpt from the plot summarizing results of the ADMIXTURE analysis (fig. 3) for Vella Lavella (*A*) and Makira (*B*), and the Admixture History Graph inferring the order of admixture events for Makira (*B*); such inference was not possible for Vella Lavella (*A*).

notably low levels of sharing between Santa Cruz and other populations.

Recent Gene Flow via Inferred Chronology of Admixture Events

Next, to better understand the chronology of gene flow across Near Oceania, we applied the Admixture History Graph (AHG) approach (Pugach et al. 2016), which elucidates the sequence of admixture events in populations with more than two ancestry components inferred by the ADMIXTURE analysis.

As already discussed, most of the components inferred by the ADMIXTURE analysis in Oceania do not reflect distinct ancestries, yet an Austronesian-related signal is clearly present, as seen in the f3 results that indicate admixture involving Taiwan Aboriginals (supplementary table S2, Supplementary Material online). This signal is associated with the orange ancestry, and is seen in the highest frequency in PO (fig. 3). It therefore seems reasonable to interpret any AHG configuration in which the orange ancestry is not the most recent as suggesting additional post-Austronesian gene flow. In New Britain, distinguished by the higher frequency of the blue and pink ancestries, the AHG analysis infers Polynesian gene flow as appearing last only in the Mengen, Melamela, and Mangseng groups (supplementary fig. S13*A*, Supplementary Material online). For the other populations, the AHGs are not completely resolved for the four ancestries, but suggest either the pink (present at high frequency in the Mamusi groups) or the aqua (present at high frequency in Bougainville) ancestry as the most recent. The results of the AHG analysis for populations from New Ireland and Lavongai are inconclusive, which corroborates evidence from the IBD-sharing analysis of continuous admixture and/or a complex history of gene flow for these populations. Notably, neither the AHG nor the IBD-sharing analysis suggests that the Polynesian ancestry is the most recent in these groups (supplementary fig. S13*B*, Supplementary Material online). All Bougainville populations with more than two ancestries inferred (Buka, Saposa, and Teop) have identical AHG configurations, which infers the aqua ancestry as most recent (supplementary fig. S13*D*,

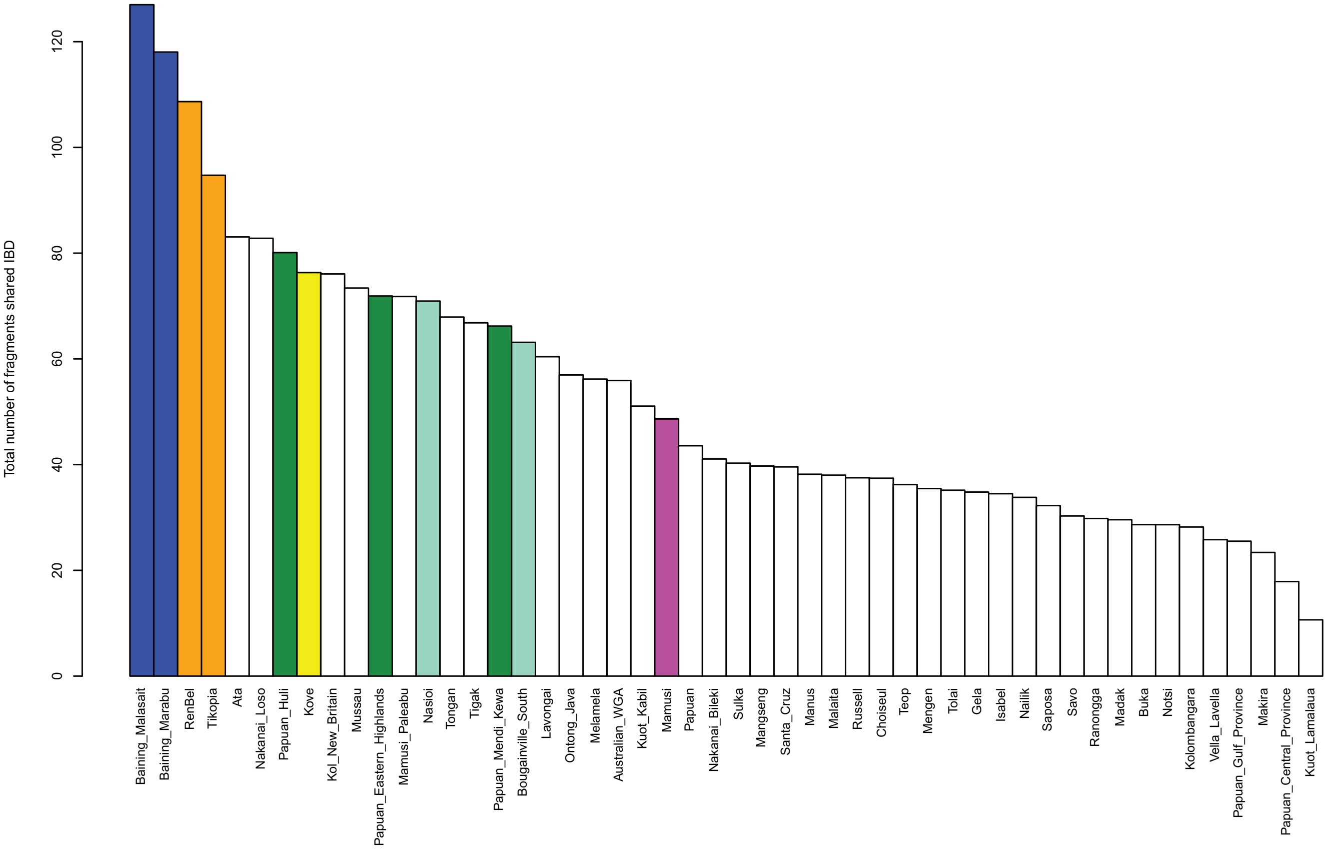

FIG. 7. Number of IBD blocks shared within populations. The blue, orange, dark green, aqua and pink colors identify the five main components inferred in the ADMIXTURE analysis at $K = 5$. The colored bars indicate populations which have their own ancestry component assigned to them in at $K = 5$ (yellow corresponds to the sixth main component, which at $K = 6$ is assigned to Kove, see supplementary fig. S5B, Supplementary Material online).

Supplementary Material online), indicating substantial recent local gene flow. AHG analysis for the western SI populations was not possible, because they exhibit only two ancestry components. The eastern SI populations (Russell, Gela, Savo, Malaita, and Makira) show more Papuan-related ancestry, and the AHGs were different than those inferred for the Bougainville populations, as the orange ancestry was coupled with the aqua (Bougainville) ancestry (supplementary fig. S13*D*, Supplementary Material online). For Makira the AHG is similar, but there is additional recent gene flow from an orange ancestry source, which is also corroborated by their sharing long IBD segments not only with Tonga (as is the case for the other islands in the eastern SI), but with the Polynesian Outliers as well (supplementary fig. S13*D*, Supplementary Material online). Finally, the AHG for Santa Cruz looks similar to those inferred for New Britain, but with additional gene flow from some Papuan-related source (supplementary fig. S13*E*, Supplementary Material online).

The AHG approach is based on differences in the covariance between the various ancestry components inferred by ADMIXTURE (Pugach et al. 2016). Since the results of the ADMIXTURE analysis are sometimes ambiguous and do not necessarily reflect admixture (as discussed above), it is important to note that in the absence of gene flow the results of the AHG test are simply inconclusive, as no difference in covariance between the ancestry components is observable. Therefore, the AHG approach will not result in a false interpretation of the ADMIXTURE analysis. Overall, the AHG tests strongly suggest that, with a few exceptions, most populations in Near Oceania exhibit a signal of post-Austronesian gene flow of diverse sources, both eastward from the BA and westward from Polynesia.

Recent Genetic Drift

We assessed recent genetic drift in the Oceanian populations by two approaches. Firstly, we analyzed patterns of sharing of IBD blocks within populations. Because the genomes of individuals in a drifted population would exhibit a congruence of genealogies, we expect to see more IBD segments shared by individuals within such populations (Gusev et al. 2012) as opposed to individuals from a population which did not experience recent drift (Ralph and Coop 2013; Harris and Nielsen 2013). When we consider the sharing of IBD segments within each population, we observe that the majority of populations assigned their own ancestry component in the ADMIXTURE analysis also exhibit extensive IBD-segment sharing within the population (fig. 7).

This pattern is further corroborated by the genome-wide analysis of runs of homozygosity (ROH), which are uninterrupted chromosome regions in which all loci in an individual are homozygous, due to identical chromosomal segments being inherited from both parents. Thus, as with IBD segments, ROHs provide a useful measure of parental relatedness and degree of endogamy, historical population size and geographical isolation (e.g., McQuillan et al. 2008). Again, all populations in the data set that were assigned their own ancestry component in the ADMIXTURE analysis (NGH, Baining, Mamusi, Nasioi, and Polynesian Outliers), also exhibit an excess of very long ROHs in comparison with other populations (supplementary fig. S14, Supplementary Material online), which is typical of populations with limited mate choice, whether due to geographical isolation (e.g., impassable mountainous terrain of interior New Guinea and New Britain, Bergström et al. 2017, where the NGH, Mamusi and Baining people reside), practice of consanguineous marriages (cross-cousin marriages, known to have been widely practiced by the Nasioi, Ogan 2005) or a strong historical bottleneck (Kirin et al. 2010), (e.g., the back migration from Polynesia of Polynesian Outliers, Kirch 1984).

Unusual genetic profile of the Santa Cruz islands and evidence for the "leapfrog" colonization from the Bismarcks

As previously mentioned, the placement of Santa Cruz in the PC analysis is rather unexpected; given the geographic location, we would expect Santa Cruz to fall closest to the other populations of Remote Oceania. Instead they are positioned closest to populations from New Britain in the BA, and remain so positioned at higher PCs (supplementary fig. S3, Supplementary Material online). The results of the ADMIXTURE analysis are equally perplexing—Santa Cruz is the furthest of the Solomon Islands from PNG, yet reveals the highest frequency of Papuan-like ancestry components, with only 5% coming from Polynesia (fig. 3).

The sharing of IBD segments within Santa Cruz is low, especially in comparison to the Polynesian Outliers (supplementary fig. S15, Supplementary Material online), which is further evidence against recent population bottlenecks (Gusev et al. 2012). The absence of a population bottleneck in Santa Cruz is also evident from the pattern of genome-wide LD and ROHs; namely there is no evidence of elevated LD or an excess of ROHs as observed in Baining and some of the Polynesian Outliers (supplementary figs. S16 and S14, Supplementary Material online), and as is typically associated with temporary reductions in population size (Pritchard and Przeworski 2001). Furthermore, it has been shown previously that the slow rate of decay in abundance of small IBD segments (2–10 cM) in comparison to longer segments is indicative of founder events (Palamara et al. 2012); such slow decay is observed in Polynesian Outliers, but not in Santa Cruz, similarly confirming that in addition to the absence of a population bottleneck, there is no evidence to support a founder event in Santa Cruz either.

Next, we examined the decay in correlation of genome-wide LD between Santa Cruz and other populations. The basic logic behind this analysis is that following divergence, the correlation in genome-wide LD between the two daughter populations will decay exponentially over time, with the rate of decay dependent only on the recombination distance between the markers (McEvoy et al. 2011). For this analysis, we have computed the correlation between the LD values for each pair of populations and for each recombination distance category, and then determined at which recombination distance the correlation decays to zero, surmising that for populations with recent genetic ties this correlation should persist for longer genomic distances (i.e., there would be less time for recombination to break down the correlation). Indeed, the correlation in LD between the French and any population from Oceania decays to zero at about the same recombination distance of ∼0.4 cM, suggesting that the French are approximately equally related to all populations in the analysis (supplementary fig. S17*A*, Supplementary Material online). In contrast, the correlation in LD between two Polynesian outliers, Rennell and Bellona (RB) and Ontong Java, persists for much longer recombination distances, >0.8 cM, whereas the correlation between RB and populations from the BA and the SI decays much faster, becoming zero at a distance between 0.45 and 0.65 cM (supplementary fig. S17*B*, Supplementary Material online). Interestingly, this correlation persists for longer distances between Santa Cruz and the BA (0.57 cM with Sulka and Notsi), than for Santa Cruz and the SI (0.55 cM with Makira; 0.52 cM with Choiseul; supplementary fig. S17*C*, Supplementary Material online). To compute the significance of this difference, that is, if it is consistent with zero, we used a weighted block jackknife procedure (Busing et al. 1999) by sequentially dropping each of the 22 chromosomes (see Methods) and recomputing the correlation in LD between Santa Cruz and BA, and Santa Cruz and SI, for each of the runs. A jackknife estimate of the difference in the rate of the correlation decay is highly significant with $Z = 3.96$ and $Z = 6.12$ standard errors from zero for the difference between SC and the BA versus SC and Makira, and between SC and the BA versus SC and Choiseul, respectively.

To summarize, in all analyses Santa Cruz falls consistently closer to populations from the Bismarcks than to those from the Solomon Islands, and appears to be quite different from the other populations of Polynesian origin. The high frequency of Papuan-like ancestry in Santa Cruz, and the gradient of this ancestry diminishing across the eastern SI in an east–west direction, is in accordance with previous suggestions (Sheppard 2011) that the first people to expand into Remote Oceania did so in a "leapfrog" manner (proceeding by leaps rather than by gradual steps) from the Bismarcks, completely bypassing the SI. The SI were then subsequently settled in a bidirectional manner, underlying the clear distinction in the genetic background between western and eastern parts of the chain (fig. 6). Notably, this distinction that we observe based on genetic evidence is mirrored by a high-order linguistic break, known as the Tryon–Hackman line (Ross 1988; Sheppard and Walter 2006), which demarcates the boundary between the Western Oceanic grouping of Austronesian languages (northern New Britain and New Ireland, and the northern and western Solomons to Isabel), and the Central/Eastern Oceanic grouping, which includes all Austronesian languages spoken to the east of this line (Lynch et al. 2002). There are also cultural differences between SI populations on either side of the Tryon–Hackman line (Sheppard and Walter 2006); the genetic evidence provides strong support for bidirectional settlement of the SI, and thus a potential explanation for the Tryon–Hackman line.

In conclusion, our analyses support the view that the initial dispersal of people from Bismarck Archipelago into Remote Oceania occurred in a "leapfrog" manner, completely bypassing the main chain of the Solomon Islands, and that the colonization of SI proceeded then in a bidirectional manner, underlying the divergence between western and eastern Solomons, in agreement with the Tryon–Hackman line. Santa Cruz remains a puzzle, exhibiting strong ties to the Bismarcks and high levels of Papuan ancestry not found elsewhere in the SI, yet also not exhibiting any signal associated with a founder event. It appears there was substantial Papuan-related gene flow to Santa Cruz that bypassed the rest of the SI. We also report ubiquitous recent post-Austronesian gene flow across the Solomons, both from the west and from the east, which could have been precipitated by natural calamities, such as volcanic eruptions (known to have occasioned the movement of the Tolai people) and droughts, wars, trade, or changes in cultural practices, like the introduction of sweet potato over the last four hundred years (Moore 2003), as well as the back migration from Polynesia of Polynesian Outliers (Kirch 1984). Finally, as more and more studies of human genetic variation emerge based on dense geographical sampling, it becomes apparent that geographically close populations can have quite different histories (e.g., in Siberia, Pugach et al. 2016, Indonesia, Hudjashov et al. 2017, and Papua New Guinea, Bergström et al. 2017), so dense genome-wide data and fine scale sampling are needed to unravel all the complexities of human population history.

Materials and Methods

Samples and Genotypes

The data set includes samples from 50 populations from Near and Remote Oceania (supplementary fig. S1, Supplementary Material online): aboriginal Australians, 3 populations from the highlands and 3 populations from the coastal parts of mainland Papua New Guinea, 23 populations from the Bismarck Archipelago (islands of New Britain, New Ireland, Lavongai, Mussau, and Manus), 5 populations from Bougainville, 11 populations from the Solomon Islands, 3 Polynesian "outlier" populations (Rennell and Bellona, Ontong Java, and Tikopia) sampled in the Solomon Islands, and Tonga. For comparative purposes we also included samples from Mainland and Island South East Asia, India and Western Eurasia (supplementary table S1 and fig. S1, Supplementary Material online). All samples were genotyped previously (Lazaridis et al. 2014; Qin and Stoneking 2015; Skoglund et al. 2016) on the Affymetrix Human Origins SNP Array, which includes 13 panels of SNPs ascertained in different populations, including Papuans (Patterson et al. 2012). Data curation was performed as described previously (Pugach et al. 2013). After quality filtering and data integration, the full data set comprised 823 individuals and over 620,000 autosomal SNPs. Not all data were used for all of the analyses. Although all individuals were sampled as unrelated, from the proportion of the genome in shared IBD blocks we have identified four pairs of samples in the Oceanian data set (including one pair with individuals belonging to two different populations from Bougainville Island) with the relationship coefficient indicative of them being first degree relatives (supplementary fig. S18, Supplementary Material online). Excluding one individual in any pair with a relationship coefficient higher than 0.2 (all first and second degree relatives, as well as some third and fourth degree relatives) had no impact on the results (supplementary fig. S19, Supplementary Material online), so we therefore kept them in the data set.

PCA

PCA was performed using the StepPCO software (Pugach et al. 2011) on the entire data set, on a subset excluding populations of Western Eurasia and India, and on a subset comprising only the Oceanian populations.

Admixture

To infer individual ancestry components and admixture proportions we used the ADMIXTURE software (Alexander et al. 2009). The analysis was performed for the entire data set and for the data set comprising only the Oceanian populations. The LD pruning for the analyses was done using the PLINK tool (Purcell et al. 2007) with the following settings, which define window size, step and the r^2 threshold: –indep-pairwise 200 25 0.4 (Rasmussen et al. 2011), which reduced the data set to 215,869 markers. We ran ADMIXTURE for $K = 3$ through $K = 12$ for the full data set, and from $K = 3$ through $K = 8$ for the Oceanian subset. One hundred and twenty independent runs for each value of K were performed for each data set. We confirmed consistency between runs and used the cross-validation procedure implemented in ADMIXTURE to find the best value of K (supplementary fig. S6, Supplementary Material online). In addition, to evaluate how the choice of SNPs affects the results, for the full data set we performed ten independent runs for each value of K from $K = 7$ through $K = 11$ using the SNPs from each of the thirteen ascertainment panels (Patterson et al. 2012) separately. Results for each of the panels for $K = 10$ are shown in supplementary figure S8, Supplementary Material online.

3-Population Test

Formal tests of admixture were carried out using the f3 statistics. These f3 statistics were calculated with the qp3Pop script from the AdmixTools package (Patterson et al. 2012).

Inferring Segments IBD

The data were phased using BEAGLE v3.3.2. Segments that are identical by descent (IBD) were inferred using the fastIBD method implemented in BEAGLE (Browning and Browning 2007, 2011). We ran the algorithm 10 times with different random seeds. The results were then combined and postprocessed as described previously (Ralph and Coop 2013; Pugach et al. 2016) to extract only those IBD segments seen at least twice in the ten BEAGLE runs and which had a significance score lower than 10^{-9}. We also merged any two segments separated by a gap shorter than at least one of the segments and no >5 cM long, thus removing artificial gaps potentially introduced because of low marker density or possible switch error during phasing (Ralph and Coop 2013; Pugach et al. 2016). All results were adjusted for differences in sample size, and blocks shorter than 2 cM were excluded from some of the analyses (Ralph and Coop 2013). To run PC analysis based on the results of the IBD calculation, we used the ade4 R package (Dray and Dufour 2007); the inverse of the matrix with the number of IBD blocks shared by each pair of individuals was used as a distance matrix (Pugach et al. 2016).

ROH

ROH (consecutive stretches of homozygous SNPs) provide information about relatedness among individuals and population isolation. ROH were identified using PLINK (Purcell et al. 2007) with default settings.

Linkage Disequilibrium (LD) and Correlation in LD

Genome-wide LD was estimated by binning the genome-wide data into 50 evenly spaced recombination distance categories (0.005–1.000 cM). For each population and for every pair of SNPs within each distance category, we calculated the squared correlation (rLD^2) in allele frequencies by randomly selecting seven individuals from each population (to ensure equal number of individuals per population, since the smallest sample size in this data set was 7) and further adjusting the measurement for each pair of SNPs to account for missing data (McEvoy et al. 2011; Pugach et al. 2013). The correlation between the LD values was computed for each pair of populations and for each recombination distance category (McEvoy et al. 2011; Pugach et al. 2013). Standard errors for the difference in the distance over which the correlation between the LD values decays to zero were computed using the weighted jackknife approach (Busing et al. 1999) with chromosome-sized blocks and weighting each replicate by the total number of SNPs in each of the runs.

AHGs

For populations exhibiting evidence of multiple admixture events, the order of the admixture events was inferred using AHGs as described previously (Pugach et al. 2016). All tests for covariance between the ancestry components were based on the results for $K = 5$ (supplementary fig. S13, Supplementary Material online) of the ADMIXTURE analysis based on the Oceanian populations only.

Dates of Admixture Inference

To infer local ancestry along individual chromosomes and obtain the genome-wide block-like signal of admixture we used a PCA-based approach implemented in the PCAdmix software (Brisbin et al. 2012). In choosing the proxies which would best represent the parental populations we followed the results of the 3-population test and ADMIXTURE analyses. Specifically, we used the PNG highlanders to represent the Near Oceanian ancestry component, and either the Polynesian Outliers or the Taiwan Aboriginals as the proxy for the Austronesian ancestry component. After retrieving the block-like admixture signal using PCAdmix, which assigns ancestry along individual chromosomes to either of the two source populations, we applied wavelet-transform analysis and used the WT coefficients to infer time since admixture by comparing the results to simulations, as described previously (Pugach et al. 2011).

Comparison between the WT and ALDER Methods

We used simulated data to assess the performance of these two methods for dating admixture events. For each admixture scenario we simulated data from nine populations with the MaCS software (Chen et al. 2009), each consisting of 30 individuals with 10 independent chromosomes of 200 Mb. One of the simulated populations was the product of an admixture between two other populations in the simulated data set (Pickrell et al. 2014 and supplementary fig. S12, Supplementary Material online), with the admixture rate set to 0.2. For the one pulse admixture scenarios, the admixture times were set to 40, 70, 90, 100, 110, 120, 130, 140, 150, or 165 generations ago, thus generating ten distinct admixture histories. For the scenarios involving two pulses of gene flow from the same source, the earlier event was set to 100, 120, or 140 generations ago, and on each of these backgrounds an additional admixture event was set to occur 40, 60, or 90 generations ago, thus producing nine distinct admixture histories. Each simulated admixture history was generated ten times. As an example, below is the MaCS command line used to simulate two pulses of admixture from the same source population, occurring 90 and 140 generations ago:

```
macs 540 200000000 -t 0.00004 -r 0.0004 -I 9 60 60 60 60 60 60 60 60 60 0 -em 0.0035 1 7 8000 -em 0.003525 1 7 0 -em 0.00225 1 7 8000 -em 0.002275 1 7 0 -ej 0.0125 7 8 -ej 0.0125125 1 2 -ej 0.0125125125 4 5 -en 0.0249 8 0.02 -ej 0.025 8 9 -en 0.0249249 2 0.02 -ej 0.02525 2 3 -en 0.0249249249 5 0.02 -ej 0.0252525 5 6 -en 0.0748 9 0.01 -ej 0.075 6 9 -en 0.1498 9 0.01 -ej 0.15 3 9
```

When inferring the times of admixture based on the WT analysis, PCAdmix (Brisbin et al. 2012) was first applied to the simulated data set to determine local ancestry along the chromosomes, with the source populations defined by the simulation set-up (supplementary fig. S12, Supplementary Material online). The WT analysis was then performed as described previously (Pugach et al. 2011), with wt coefficients compared with those obtained with simulations (Pugach et al. 2011) based on either the admixture rate of 0.2 for the one-pulse or 0.4 for the two-pulse scenarios. ALDER was run with default parameters, with all "unadmixed" simulated populations supplied as potential sources.

Data Availability

To comply with the informed consent under which the samples were obtained, we make the data available upon request by asking the person requesting the data to agree in writing to the following restrictions: 1) The data will only be used for studies of population history; 2) the data will not be used for medical or disease-related studies, or for studies of natural selection; 3) the data will not be distributed to anyone else; 4) the data will not be used for any commercial purposes; and

5) no attempt will be made to identify any of the sample donors.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all individuals who contributed samples for this study. We thank David Reich for assistance with genotyping, and Murray Cox and Cosimo Posth for valuable discussion. This research was funded by the Max Planck Society.

Author Contributions

I.P. and M.S. conceived and designed the experiments. D.A.M., F.R.F., and J.S.F. provided reagents. I.P. performed the experiments, analyzed the data with assistance from A.T.D., and contributed analytical tools. I.P. and M.S. wrote the paper with input from all authors.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.

Bellwood P. 2004. First farmers: the origins of agricultural societies. 1st ed. Malden (MA): Wiley-Blackwell.

Bellwood P, Dizon EZ. 2005. The Batanes archaeological project and the Out of Taiwan hypothesis for Austronesian dispersal. *J Austronesian Stud* 1:1–31.

Bergström A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, Alpers MP, Koki G, Pomat W, Siba P. 2017. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* 357(6356):1160–1163.

Blust R. 1999. Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. *Symp Ser Inst Linguist Acad Sin* 1:31–94.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG, Bustamante CD. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* 84(4):343–364.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 88(2):173–182.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5):1084–1097.

Busing FMTA, Meijer E, Leeden RVD. 1999. Delete-m jackknife for unequal m. *Stat Comput* 9(1):3–8.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res* 19(1):136–142.

Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L, Smith M, Roberts RG, Hayes E, Lowe K, Carah X. 2017. Human occupation of northern Australia by 65,000 years ago. *Nature* 547(7663):306–310.

Delfin F, Myles S, Choi Y, Hughes D, Illek R, van Oven M, Pakendorf B, Kayser M, Stoneking M. 2012. Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands. *Mol Biol Evol* 29(2):545–564.

Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 22(4):1–20.

Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, Kayser M, Stoneking M. 2014. Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *Am J Hum Genet* 94(5):721–733.

Falush D, van Dorp L, Lawson D. 2016. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots, https://www.biorxiv.org/content/early/2017/11/27/066431, last accessed January 5, 2018, unpublished data.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128(2):415–423.

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo J-H, Koki G, Hodgson JA. 2008. The genetic structure of Pacific Islanders. *PLOS Genet* 4(1):e19.

Friedlaender JS, Gentz F, Green K, Merriwether DA. 2002. A cautionary tale on ancient migration detection: mitochondrial DNA variation in Santa Cruz Islands, Solomon Islands. *Hum Biol* 74(3):453–471.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913):479–483.

Gross JL. 2016. Waipio valley: a Polynesian journey from Eden to Eden. Xlibris Corporation.

Groube L, Chappell J, Muke J, Price D. 1986. A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature* 324(6096):453–455.

Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol* 29(2):473–486.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet* 9(6):e1003521.

Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, Lansing JS, Hammer MF, Cox MP. 2017. Complex patterns of admixture across the Indonesian archipelago. *Mol Biol Evol* 34(10):2439–2452.

Hunley K, Dunn M, Lindström E, Reesink G, Terrill A, Norton H, Scheinfeldt L, Friedlaender FR, Merriwether DA, Koki G, Friedlaender JS. 2007. Inferring prehistory from genetic, linguistic, and geographic variation. In: Friedlaender JS, editor. Genes Lang cult Hist southwest Pac. New York: Oxford University Press. p. 141–154.

Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA, Moyse-Faurie C, Rutledge RB, Schiefenhoevel W, Gil D, et al. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol* 23(11):2234–2244.

Kayser M, Brauer S, Weiss G, Underhill P, Roewer L, Schiefenhövel W, Stoneking M. 2000. Melanesian origin of Polynesian Y chromosomes. *Curr Biol* 10(20):1237–1246.

Kayser M, Lao O, Saar K, Brauer S, Wang X, Nürnberg P, Trent RJ, Stoneking M. 2008. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet* 82(1):194–198.

Kirch PV. 1984. The Polynesian Outliers: continuity, change, and replacement. *J Pac Hist* 19(4):224–238.

Kirch PV. 2000. On the road of the winds: an archaeological history of the Pacific Islands before European contact. London: University of California Press.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* 5(11):e13996.

Ko AM-S, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, Stoneking M, Ko Y-C. 2014. Early Austronesians: into and out of Taiwan. *Am J Hum Genet* 94(3):426–436.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, et al. 2008. Correlation between genetic and geographic structure in Europe. *Curr Biol* 18(16):1241–1248.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513(7518):409–413.

Lipson M, Loh P-R, Levin A, Reich D, Patterson N, Berger B. 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol Biol Evol* 30(8):1788–1802.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193(4):1233–1254.

Lynch J, Crowley T, Ross M. (Malcolm D). 2002. In: Lynch J, Ross M, Crowley T, editors. The Oceanic languages. Richmond, Surrey: Curzon.

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, et al. 2016. A genomic history of Aboriginal Australia. *Nature* 538(7624):207–214.

McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, Wilton AN. 2010. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *Am J Hum Genet* 87(2):297–305.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21(6):821–829.

McQuillan R, Leutenegger A-L, Abdel-Rahman R, Franklin CS, 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(3):359–372.

Moore C. 2003. New Guinea: crossing boundaries and history. Honolulu: University of Hawaii Press.

Næss Å. 2006. Bound nominal elements in Äiwoo (reefs): a reappraisal of the "multiple noun class systems." *Ocean Linguist* 45(2):269–296.

Næss Å, Boerger BH. 2008. Reefs-Santa Cruz as Oceanic: evidence from the verb complex. *Ocean Linguist* 47:185–212.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. *Nature* 456(7218):98–101.

O'Connell JF, Allen J. 2015. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J Archaeol Sci* 56:73–84.

Ogan E. 2005. An introduction to Bougainville cultures. In: Regan AJ, Griffin HM, editors. Bougainville before the conflict. Canberra, Australia: Pandanus Books. p. 47–56.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91(5):809–822.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192(3):1065–1093.

Pawley A. 2007. Recent research on the historical relationships of the Papuan languages, or, what does linguistics say about the prehistory of Melanesia. In: Friedlaender, editor. Population genetics, linguistics, and culture history in the Southwest Pacific. New York: Oxford University Press. p. 36–58.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA* 111(7):2632–2637.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 11(11):800–805.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69(1):1–14.

Puechmaille SJ. 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol Resour* 16(3):608–627.

Pugach I, Delfin F, Gunnarsdóttir E, Kayser M, Stoneking M. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc Natl Acad Sci USA* 110(5):1803–1808.

Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. 2016. The complex admixture history and recent southern origins of Siberian populations. *Mol Biol Evol* 33(7):1777–1795.

Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol* 12(2):R19.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.

Qin P, Stoneking M. 2015. Denisovan ancestry in East Eurasian and Native American populations. *Mol Biol Evol* 32(10):2665–2674.

Ralph P, Coop G. 2013. The geography of recent genetic ancestry across Europe. *PLoS Biol* 11(5):e1001555.

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334(6052):94–98.

Roberts RG, Jones R, Smith MA. 1990. Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. *Nature* 345(6271):153–156.

Ross M. 1988. Proto Oceanic and the Austronesian languages of Western Melanesia. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Ross M, Næss Å. 2007. An Oceanic origin for Äiwoo, the language of the Reef Islands? Ocean. *Linguist* 46(2):456–498.

Sheppard PJ. 2011. Lapita colonization across the Near/Remote Oceania boundary. *Curr Anthropol* 52(6):799–840.

Sheppard PJ, Walter R. 2006. A revised model of Solomon Islands culture history. *J Polyn Soc* 115:47–76.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, et al. 2016. Genomic insights into the peopling of the Southwest Pacific. *Nature* 538(7626):510–513.

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo J-H, Thomson N, et al. 2011. Ancient voyaging and Polynesian origins. *Am J Hum Genet* 88(2):239–247.

Spriggs MK. 2003. Chronology of the Neolithic transition in Island Southeast Asia and the Western Pacific: a view from 2003. *Rev Archaeol* 24:57–80.

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Stoneking M, Kayser M. 2010. Demographic history of Oceania inferred from genome-wide data. *Curr Biol* 20(22):1983–1992.

Wurm S, Carrington L, Wurm S. 1978. Reefs-Santa Cruz: Austronesian, but . . .! Second Int Conf Austronesian Linguist Proc. Canberra (Australia) Pacific Linguistics: 969–1010.