

Population Genetic Structure and Origins of Native Hawaiians in the Multiethnic Cohort Study

Sung K. Kim^{1,2}, Christopher R. Gignoux¹, Jeffrey D. Wall¹, Annette Lum-Jones³, Hansong Wang³, Christopher A. Haiman⁴, Gary K. Chen⁴, Brian E. Henderson⁴, Laurence N. Kolonel³, Loic Le Marchand³, Daniel O. Stram⁴, Richa Saxena^{5,6}, Iona Cheng^{3*}

1 Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, **2** Sequenom, Incorporated, San Diego, California, United States of America, **3** Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, Hawaii, United States of America, **4** Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America, **5** Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America, **6** Center for Human Genetics Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America

Abstract

The population genetic structure of Native Hawaiians has yet to be comprehensively studied, and the ancestral origins of Polynesians remain in question. In this study, we utilized high-resolution genome-wide SNP data and mitochondrial genomes of 148 and 160 Native Hawaiians, respectively, to characterize their population structure of the nuclear and mitochondrial genomes, ancestral origins, and population expansion. Native Hawaiians, who self-reported full Native Hawaiian heritage, demonstrated 78% Native Hawaiian, 11.5% European, and 7.8% Asian ancestry with 99% belonging to the B4 mitochondrial haplogroup. The estimated proportions of Native Hawaiian ancestry for those who reported mixed ancestry (i.e. 75% and 50% Native Hawaiian heritage) were found to be consistent with their self-reported heritage. A significant proportion of Melanesian ancestry (mean = 32%) was estimated in 100% self-reported Native Hawaiians in an ADMIXTURE analysis of Asian, Melanesian, and Native Hawaiian populations of $K=2$, where K denotes the number of ancestral populations. This notable proportion of Melanesian admixture supports the "Slow-Boat" model of migration of ancestral Polynesian populations from East Asia to the Pacific Islands. In addition, approximately 1,300 years ago a single, strong expansion of the Native Hawaiian population was estimated. By providing important insight into the underlying population structure of Native Hawaiians, this study lays the foundation for future genetic association studies of this U.S. minority population.

Citation: Kim SK, Gignoux CR, Wall JD, Lum-Jones A, Wang H, et al. (2012) Population Genetic Structure and Origins of Native Hawaiians in the Multiethnic Cohort Study. PLoS ONE 7(11): e47881. doi:10.1371/journal.pone.0047881

Editor: Dana C. Crawford, Vanderbilt University, United States of America

Received: June 8, 2012; **Accepted:** September 18, 2012; **Published:** November 7, 2012

Copyright: © 2012 Kim et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the National Institute of Health grants (CA140636, CA063464, CA54281, and HG004049-01A2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that the author, Sung Kim, is employed by Sequenom, Inc. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: icheng@cc.hawaii.edu

Introduction

Population structure and genetic ancestry of Native Hawaiians, a racial/ethnic minority group within the U.S., have been understudied. Previous reports of the genetic admixture of Native Hawaiians have been largely based on ancestry informative markers (AIMs), which can be limited in their resolution in providing accurate estimates of the genetic contribution of ancestral populations [1,2]. Studying these patterns of population structure is imperative for genetic association testing in which the confounding effects of population stratification can lead to false positive associations [3,4]. In addition, individuals with mixed ancestry such as Native Hawaiians can be used to map susceptibility loci for complex traits via admixture mapping. This is a powerful approach for localizing risk loci of traits that have a higher prevalence in one ancestral population than another [5–11]. Such an approach has yet to be applied to Native Hawaiians due to the lack of a comprehensive admixture map for this group. Future admixture mapping studies in Native Hawaiians may be

informative for chronic diseases such as cardiovascular disease, diabetes, and breast cancer for which Native Hawaiians demonstrate a higher incidence than other Asian and European populations [12–16].

Native Hawaiians have a long and complicated demography, dating back to the earliest settlement of Hawai'i. Archeological and paleoecological evidence suggest that the Hawaiian archipelago was initially settled by Polynesian settlers between 300 AD and 800 AD [17]. Ancestral Polynesians are widely believed to originate from East Asia [18,19]. However, as they migrated across the Pacific (remote Oceania), their degree of admixture with the indigenous Melanesian population is unclear. Two leading yet polar demographic models have been proposed. The first model, the "Express Train" hypothesis, suggests that the migration of ancestral Polynesians across the Pacific was rapid with limited or no contact with the local Melanesian population. Evidence in support of this theory includes linguistics [18–21] and mitochondrial data [22–25]. The second model, the "Slow Boat" hypothesis, suggests that ancestral Polynesians intermixed significantly with Melanesians

prior to populating the Pacific Islands. Evidence in support of this hypothesis includes archaeological evidence [26] and Y chromosome data [27–30]. Previous genome-wide surveys [31–33] that addressed the ancestral origins of Polynesians have yielded conflicting results over these two hypotheses with the majority of studies focusing on microsatellite (short-tandem repeat) markers [31,32]. These discrepancies may be due to the use of microsatellite markers, which have inherently high mutation rates and a mutation model that is difficult to model [34–36]. More recently, Wollstein et al. [33] used genome-wide SNP data as a more precise and fine-resolution approach to understand the demographic history of Polynesians, estimating ~13% Melanesian ancestry with admixture occurring 3,000 years ago.

The population size of Native Hawaiians during the 17th–18th century ranged from 250,000–800,000 individuals [17]. Tragically, due to an epidemic of new diseases brought by visiting sailors and immigrant settlers, the Native Hawaiian population underwent a drastic population bottleneck, such that the 1900 census reported only ~40,000 Native Hawaiians. Most recently, the Native Hawaiian population has experienced nearly a five-fold increase in size with the U.S. 2000 census estimating over 200,000 individuals [37] (recognizing self-reporting preferences may influence this estimate, possibly leading to an overestimation in reporting of Native Hawaiian ancestry due to cultural preferences). Several diverse ethnic groups migrated to Hawai'i, following Captain James Cook's discovery of the Hawai'i islands in 1778, primarily due to the development of a sugar production industry and the resultant need for a large cadre of plantation workers. Historical records indicate that the first reported Chinese immigrants migrated to Hawai'i on merchant ships in 1788 [17,38] with as many as 56,000 additional Chinese immigrants migrating as contract workers between 1852 and 1899 [39]. In addition, over 200,000 Japanese also migrated to Hawai'i from 1885 through 1924 [40]. Overall, hundreds of thousands of individuals from all around the world, including the Philippines, Portugal, and Puerto Rico, migrated to Hawai'i via labor recruitment efforts [17]. Native Hawaiians have often intermarried with persons of other racial/ethnic groups, and these marriages have led to significant interbreeding among the diverse populations that migrated to Hawai'i [17]. The drastic reduction in population size of Native Hawaiians after the arrival of Europeans and their increase in numbers after annexation by the U.S. suggest a rich and complicated demographic history similar to other admixed populations within the U.S., such as Latinos and African Americans [17] [41–43].

In this study, we present a comprehensive fine-scale analysis of the genome-wide estimates of population structure and admixture in Native Hawaiians, using high-resolution genotype data from both the nuclear and mitochondrial genomes. Our genome-wide SNP data of the nuclear genome gives us an opportunity to examine the degree of population structure within Native Hawaiians. In addition, we test two competing hypotheses on the ancestral origins of Polynesians as represented by Native Hawaiians. Furthermore, our sequencing data of the mitochondrial genome allows for characterizing the patterns of mitochondrial DNA (mtDNA) diversity in Native Hawaiians as well as estimating the effective population size of Native Hawaiians as seen through time.

Results

Genome-wide data

We applied ADMIXTURE on genome-wide SNP data to finely estimate the degree of admixture in Native Hawaiians. We

analyzed 148 Native Hawaiians by using a merged dataset of 114,112 SNPs and 466 additional individuals from HGDP (see Methods). Genome-wide SNP analysis at $K > 7$ identified continental substructure (such as the differentiation between our representative African populations) whereas $K < 5$ identified major continental groups (data not shown as they provide little information in regards to the ancestral composition of Native Hawaiians). Figures 1A and 1B show ADMIXTURE's estimate of each individual's genome-wide proportion of ancestry when $K = 5$ and $K = 6$. $K = 5$ and $K = 6$ were selected as they provided the first clear differentiation of Native Hawaiians relative to the major ethnic populations within our study. Each color was assigned to represent one of the K putative ancestral populations by determining which reference population group contained the largest ancestral component. We find, as Figure 2A illustrates, that for $K = 5$, on average, 37.7% (the sum of all non-green components of ancestry) of the Native Hawaiian genomes of those who reported themselves and both their parents as being solely Native Hawaiian originated from a population outside Oceania. ADMIXTURE analysis at $K = 6$ suggests that these 100% self-reported Native Hawaiians have an average of 78% of their genomes classified as Native Hawaiian ancestry (min = 0.019, max = 0.99, median = 0.87, std. dev. = 0.25) with 11.5% and 7.8% classified as European and Asian ancestry, respectively (Figure 2B).

As expected, individuals with founders originating from non-Hawaiian populations typically have higher levels of admixture when compared to individuals with full Native Hawaiian lineage. For example, at $K = 5$, individuals with either parent belonging to Asian or European descent, ADMIXTURE estimated 58.0% and 47.5% of their genomes to contain Asian and European ancestry, respectively (Figure 2A). At $K = 6$, we find Native Hawaiians, whose grandparent (as assumed based on reporting of parental background) was Asian (European) to contain 22.3% (20.3%) of their genome to be of Asian (European) ancestry (Figure 2B).

We performed a MDS analysis of the genome-wide SNP data by using the MDS routine in PLINK to generate a graphical view of the genomic distance between each individual. Figure 3 illustrates the clustering of individuals belonging to select populations that represent three major groups: Europeans, Asians and Oceania. Consistent with ADMIXTURE estimates of ancestry, we find that Native Hawaiians typically cluster between all three divergent populations. Moreover, admixed Native Hawaiians with European (Asian) ancestry tended to cluster more towards the representative European (Asian) populations. These findings suggest a great level of variation among Native Hawaiians today, quite possibly due to historical and ongoing admixture events, and warrant careful estimation of admixture with more individuals.

Origins of Native Hawaiians

The “Express Train” and the “Slow Boat” models of Polynesian migration are expected to have uniquely distinct genetic signatures on present day genomes of Native Hawaiians. Under the “Express Train” model, the proportion of admixture in Native Hawaiians of Melanesian and Asian ancestry is expected to be near zero, whereas under the “Slow Boat” model, the proportion of admixture is expected to be substantially greater than zero. To test these two models, we conducted a supervised ADMIXTURE analysis using Papuan and Melanesians as one source population of Polynesians and Han Chinese, She, Cambodian, Japanese, Yakut, and Yi as surrogates for the second source population of Taiwanese aborigines [18,19]. Importantly, we did not fix ancestry for the Melanesians or Asians and therefore allowed for admixture within either ancestral groups—thus,

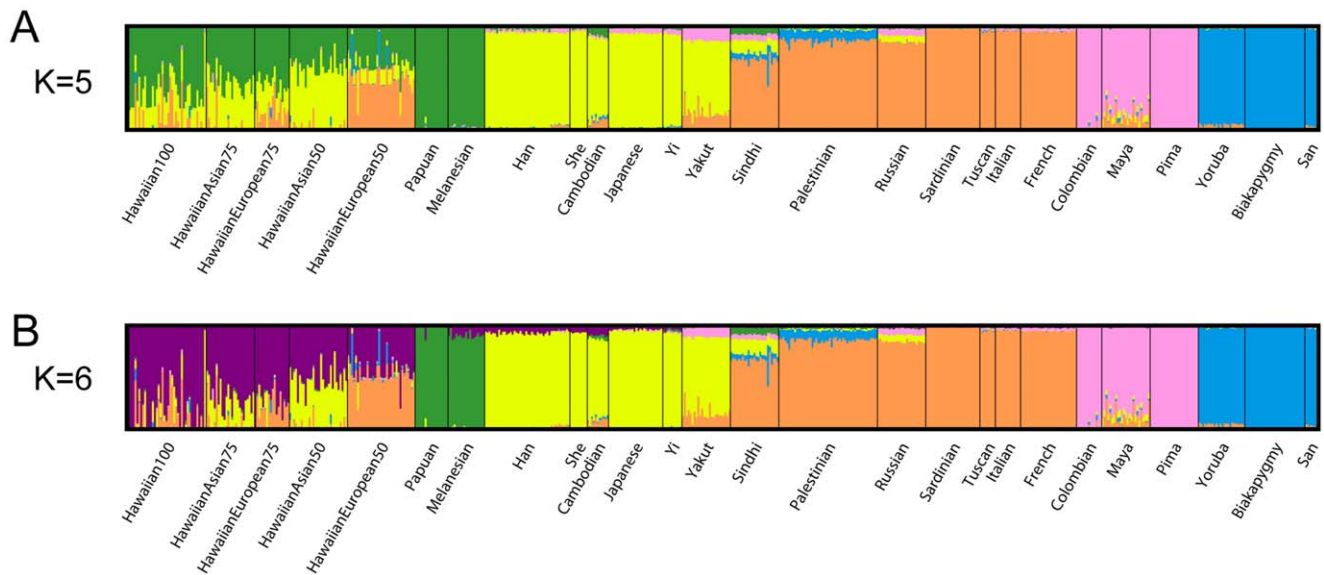


Figure 1. ADMIXTURE clustering of Native Hawaiians for K = 5 (A) and K = 6 (B). Figures 1A and 1B illustrate the clustering of Native Hawaiians and HGDP samples based on GWAS data. Each vertical bar represents an individual's proportion of K ancestral clusters (i.e. color) as estimated by ADMIXTURE.

doi:10.1371/journal.pone.0047881.g001

mitigating bias by earlier admixture processes and allowing for accurate clusters of ancestry membership. We set $K=2$ and estimated in 40 100% Native Hawaiians an average of 32% and 68% of their genomes to be derived from Melanesian and Asian origins, respectively (Figure 4). This notable proportion of Melanesian admixture (32%) among Native Hawaiians, substantially greater than zero, lends support of the “Slow Boat” model of ancestral origins.

Mitochondrial DNA data

The Polynesian motif is comprised of a 9-bp COII/tRNA^{Lys} intergenic deletion relative to the revised Cambridge Reference Sequence (rCRS) for the mitochondrial genome and accounts for >90% of Polynesian mtDNA [24]. Lum et al. [44] identified the Polynesian motif in ~95% of their Native Hawaiian samples, consistent with archeological and linguistic evidence of Polynesian origins [45]. In the 160 Native Hawaiian mitochondrial genomes for this study (100% self-reported Native Hawaiian ancestry), we find 80, 19, and 60 carry the B4a1a1, B4a1a1a1 and B4a1a1a3 motif, respectively. Furthermore, we identified one individual, who carried 7 of the 9 diagnostic markers for the Q1 motif (Table S5). Across all 160 Native Hawaiians, we identified 14 mtSNPs to have a minor allele frequency >1% (Table 1), of which 6 polymorphisms have not been previously identified by PhyloTree [46] to be a haplogroup identifying marker.

Figure S1 depicts the MDS analysis with select samples of mitochondrial haplogroups. We selected haplogroups P and B as they are commonly observed in Polynesian populations [47–49]. To provide a broad spectrum of ancestral haplogroups for comparison to our Native Hawaiian mitochondrial haplogroups, we included lineages from haplogroups H, J, K, L, M, D and G. Overall, we find consistent patterns with the known genealogy of our reference panel of mitochondrial sequences. The majority of our Native Hawaiians appears to cluster amongst themselves with some overlap with haplogroup B, a haplogroup found across East Asia, the Pacific, and the Americas. In light of the drastic isolation, genetic drift, population bottleneck and subsequent growth that

occurred in the Hawaiian Islands after settlement, we see a clear signature of Polynesian lineages as an outlying cluster.

The Bayesian Skyline analysis in Figure 5 shows a single, strong expansion beginning approximately 1,300 years before present (median estimate). From that point on, the Native Hawaiian effective population size expands two orders of magnitude within 1,000 years. The beginning of the expansion correlates strongly with archaeological records for the initial settlement of the Hawai'i islands. It is worth noting that if we used the phylogenetic mutation rate, our estimates would have been far too early (roughly 4,000 Years Before Present).

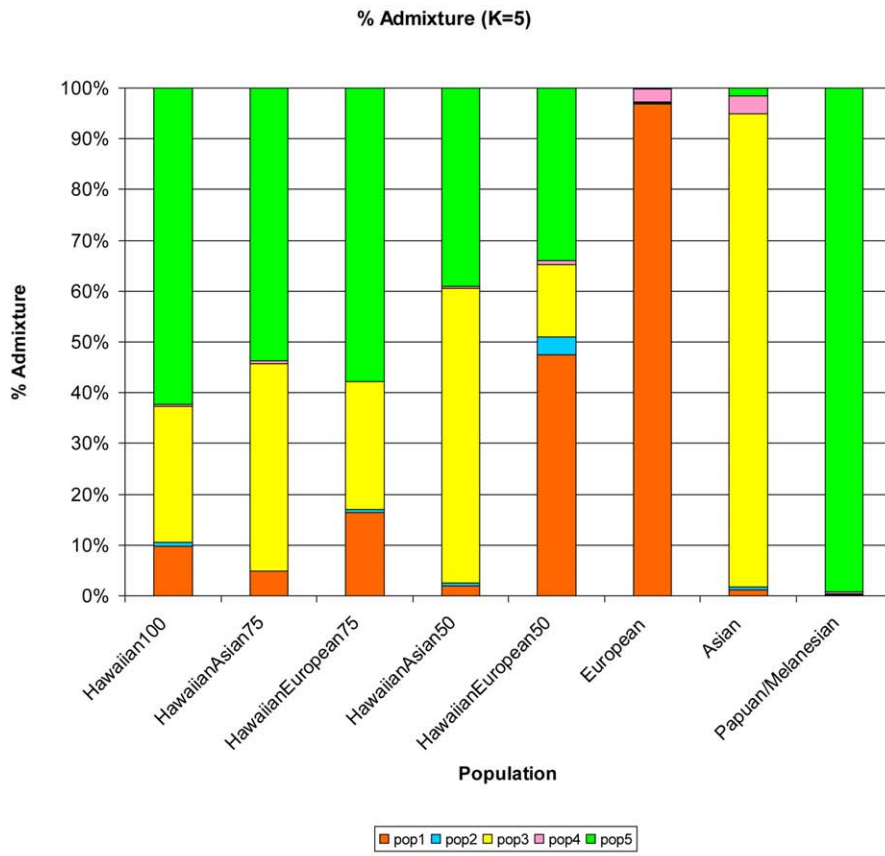
Discussion

In this study, we integrated nuclear and mitochondrial genomic data from Native Hawaiians with HGDP data to provide the first comprehensive survey of the three components of their demography—patterns of population structure, genetic origins, and population expansion over time.

Our genome-wide SNP data provides a unique view into the genetic architecture and the degree of admixture in Native Hawaiians. We find patterns of admixture when examining genome-wide SNP data in 100% self-reported Native Hawaiians such that on average 11.5% and 7.8% of their genomes contained European and Asian ancestry, respectively. In addition, we detected admixture to be consistent with the expected ancestral proportions based on varying degrees of self-reported Native Hawaiian ancestry. Given the high number of markers available in genome-wide array data, we are well-powered to determine fine-scale differences in cluster membership [50–52]; thus, we do not expect this measured admixture to be biased by ancient shared ancestry such as seen in Asian and Oceanic populations.

Overall, our results are roughly comparable to the report by Wang et al. [53] that estimated 74% Hawaiian ancestry based on 300 AIMs that were tested in Multiethnic Cohort subjects, who reported Native Hawaiian heritage. We recognize the possibility of errors in self-reported ancestry and our data set is hampered by the lack of male individuals that would allow for investigation of

A



B

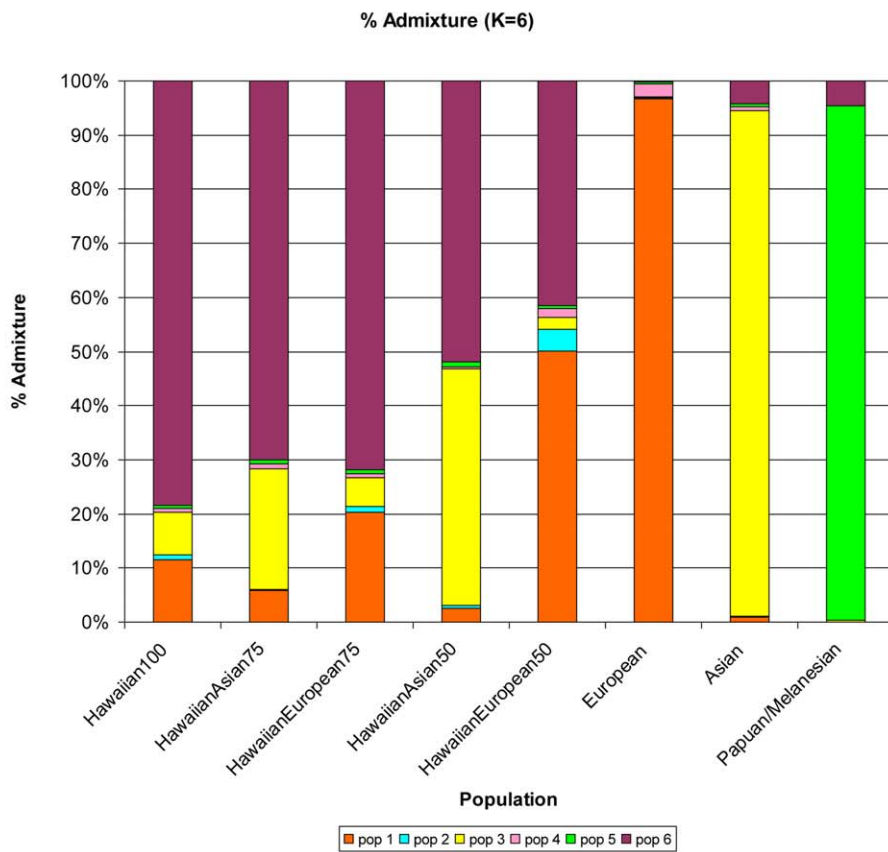


Figure 2. Average of each ancestral clustering estimated by ADMIXTURE for K = 5 and K = 6. Figures 2A and 2B illustrate the mean values for each proportion of ancestry based on GWAS data. Pop1-6 are the ancestral populations representing the European, African, East Asian, American, Oceanian, and Hawaiian populations, respectively. doi:10.1371/journal.pone.0047881.g002

sex-biased demography, a genomic signal detected in modern human populations [54], particularly in admixed populations, such as Puerto Ricans [55] and Polynesian populations [48]. To address these limitations and further refine the ancestral contribution of European and Asian ancestry in Native Hawaiians, future studies should develop specific AIMs for Native Hawaiians as well as include larger study populations of Native Hawaiian men and women with high-density genome-wide SNP data.

Recently, Kayser et al. [32] surveyed the nuclear genome with 377 microsatellite markers in 47 Pacific Islanders and identified 79% Asian and 21% Melanesian proportions of ancestry for Polynesians. These findings, along with evidence of excess Melanesian origin of the Y chromosome [27–30] and excess Asian origin of the mtDNA [22–25], suggest a sex-biased demography of admixture between ancestral Polynesians and indigenous Melanesians consistent with the “Slow Boat” hypothesis [48]. Contrary to Kayser et al. [32], Friedlaender et al. [31] also conducted a genome-wide survey with 687 microsatellites in 952 individuals from 41 Pacific populations and concluded that their data were supportive for the “Express-Train” hypothesis. Both studies are limited in genome-wide interpretations of the origins of Polynesians due to their use of microsatellites. Microsatellites are known to have higher mutation rates and a mutation model that is difficult to model, which are less reliable for studies geared toward demographic inference [34–36]. In addition,

previous studies on the Y chromosome and mtDNA are single-locus analyses and can provide only a limited view of past demographic events.

Our study of 488,961 autosomal SNPs of 40 100% Native Hawaiians provides the largest study to date that tests two divergent hypotheses regarding the ancestral migration patterns of Polynesians, specifically Native Hawaiians. Based on our supervised ADMIXTURE analyses, we estimated on average, 32% of the Native Hawaiian genome originates from Melanesian ancestry. This notable proportion of Melanesian admixture, substantially greater than zero, which would be expected under the “Express Train” model, conversely supports the “Slow Boat” hypothesis of ancestral origins. Although our conclusions are consistent with a recent genome-wide SNP study of 25 Polynesians by Wollstein et al [33], who examined highland Papuan groups, our estimation of Melanesian ancestry may have overestimated the true admixture proportions given that we based our analysis on ancestral populations (namely the coastal/island Melanesian groups found within the HGDP panel) that have historical Asian admixture via the Austronesian expansion. Furthermore, given the complex demographic history of Native Hawaiians, our current analyses cannot differentiate between long-range versus short range admixture events. While an analysis of variance of ancestry may provide insights into migration rates and estimate the time of

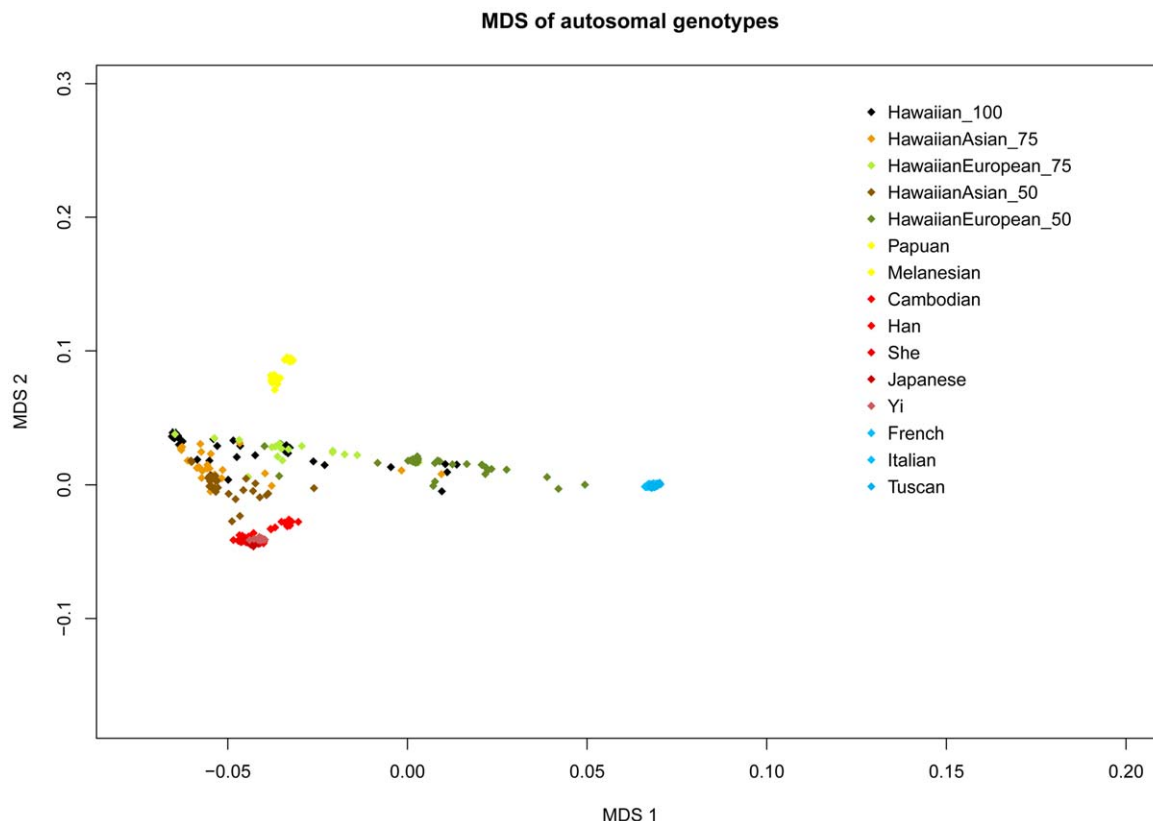


Figure 3. Multi-Dimensional Scaling analysis of GWAS data. HGDP representative samples with European, East Asian and Oceanian ancestry and are plotted against Native Hawaiians with various degrees of self-reported ancestry in MDS dimensions. doi:10.1371/journal.pone.0047881.g003

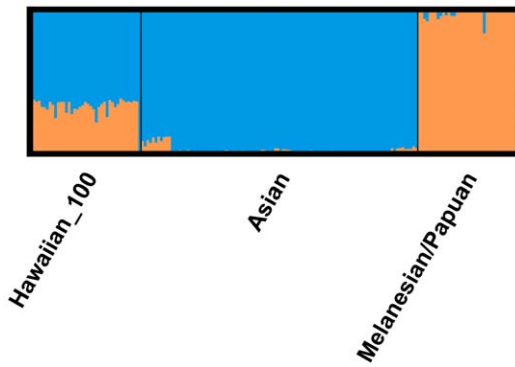


Figure 4. Supervised ADMIXTURE results for K=2 using Native Hawaiians, East Asian and Oceanian populations.

doi:10.1371/journal.pone.0047881.g004

the admixture events [56] such analysis may be confounded by overfit models.

Lastly, we merged 160 100% Native Hawaiian mitochondrial sequences with 544 a priori selected mitochondrial sequences available in the public domain to query the patterns of polymorphisms. The importance of understanding the genetic patterns of the mitochondria as demonstrated by Biffi et al. [57] is that the mtDNA can serve as a powerful tool to complement genome-wide data in assessing the confounding factors of population stratification in genome-wide association studies. Consistent with previous reports [44], we find that 99% of Native Hawaiians in our sample contain the Polynesian motif. We further show a clustering of our Native Hawaiian mitochondrial genomes towards the general B haplogroup, which previously has been identified among Melanesian, Micronesian, and Polynesian populations [58]. Given our observation of moderate levels of historical admixture with Europeans and Asians, the observed near fixation of the Polynesian motif in full Native Hawaiians suggests the occurrence of sex biased demography with an excess of European/Asian males and Native Hawaiian females in the ancestral population of Native Hawaiians.

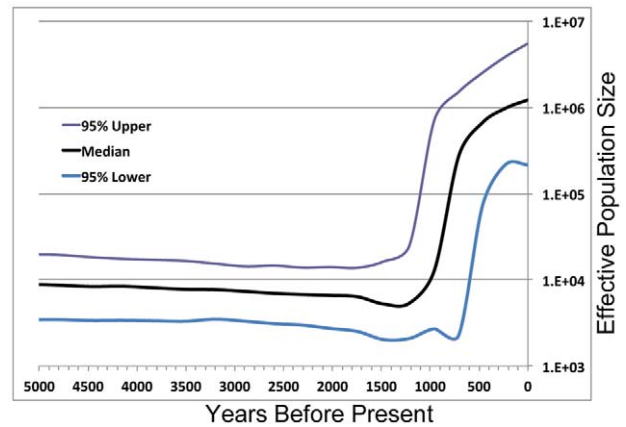


Figure 5. BEAST analysis of 160 Native Hawaiian mitochondrial genomes. All default values were used to generate the effective population size as a function of time with 95% confidence intervals.

doi:10.1371/journal.pone.0047881.g005

The Skyline analysis demonstrates a single, strong expansion after the initial settlement of the islands. In addition, the single mode of expansion is concordant with a single wave of settlers, rather than a two-wave process that has been put forward in the literature [17,59,60]. Although Polynesians contact with Westerners in the past few hundred years has resulted in a marked decrease in census population size, the effective population size was large enough so findings concerning the earlier period of expansion should not have been largely affected by such an event. The median population today appears to be overestimated, however, the credible interval is consistent with the current census population. At any rate Skyline analyses (consistent with other coalescent approaches) will have more uncertainty with extremely recent estimates due to the paucity of very recent substitution events [61,62].

Recent next-generation sequencing approaches have demonstrated the issues of demographic inference with mitochondrial data due to biased sampling approaches [63]. However, in our analysis this is not a concern as we selected 160 Native Hawaiians

Table 1. Native Hawaiian specific mitochondrial SNP with minimum allele frequency >1%.

mtDNA position	Reference allele	Native Hawaiian allele	Minor Allele Frequency	Haplogroup Marker
841	A	G	0.093	
1185	C	T	0.377	B4a1a1a3
1692	A	G	0.325	M27c, G2b2a, U6a7a
3398	T	C	0.038	M65b, I2a1
3639	A	G	0.013	
4314	T	C	0.013	P10
6905	A	G	0.129	B4a1a1a1, M5c2
8538	T	C	0.013	
9145	G	A	0.019	
10451	T	C	0.036	
12245	T	C	0.019	R3
14022	A	G	0.994	B4a1a1
15746	A	G	0.994	B4a1a, G3a, N8
15776	A	G	0.013	

doi:10.1371/journal.pone.0047881.t001

for sequencing from a population-based cohort and all samples were used in the Skyline analysis. The fact that they nearly all exhibit the Polynesian motif is consistent with Native Hawaiians being at the terminus of a serial expansion and limits our resolution to more ancient events. Therefore, the recent expansion signal observed in this analysis would not be due to bias in sampling but represents the colonization of the Hawaiian Islands.

It is feasible that this pattern of explosive growth may preserve an excess of rare variants across the entire genome of Native Hawaiians in comparison to populations that have not experienced such a recent rise in effective population size. Similar patterns have been detected in large candidate gene resequencing studies in diverse populations [64]. Future large-scale sequencing studies of Native Hawaiians should take into account that the allele frequency spectrum of the Polynesian component of this population would be notably shifted towards rare variants, even though they are at a terminus of a long serial founder effect model across the Pacific Ocean [65].

Given the dynamic and complex history of Native Hawaiians, along with the large influx of various ethnic groups into the islands over the past 200 years, it is possible that the observed admixture signals in individuals, who reported full Native Hawaiian heritage, are residuals of historical admixture events. Although our dataset is modest in comparison to studies of predominantly European populations that incorporate thousands of individuals genotyped on genome-wide SNP arrays [66], our study does, nevertheless, notably contribute to our basic understanding of the population genetics of Native Hawaiians, an understudied minority group. In addition, our estimation of the explosive growth of the Native Hawaiian population, likely leading to an increase in rare variants, has important implications for future sequencing studies of complex diseases in this population. The information gained here lays the foundation for future genetic association studies of Native Hawaiians by providing insights into the necessary corrections for errors incurred due to confounding factors of population stratification. Moreover, our findings suggest that mapping by admixed linkage disequilibrium may be a powerful tool in elucidating the genetic etiology of complex traits in Native Hawaiians.

Materials and Methods

Study Subjects

The Multiethnic Cohort Study is a large population-based prospective study of more than 215,000 men and women from Hawaii and California (mainly Los Angeles County). The cohort is composed predominantly of individuals from the following five racial/ethnic groups: African Americans, Native Hawaiians, Japanese, Latinos, and European Americans. Participants between the ages of 45 and 75 years were recruited from March 1993 through May 1996. Participants were asked to self-report their race/ethnicity and that of each of their parents, which included the option to report a mixed racial/ethnic background. Further details about this cohort are provided elsewhere [67].

We utilized high-resolution genome-wide SNP data and mitochondrial genomes of 148 and 160 Native Hawaiians, respectively (discussed below). In total, we had 192 individuals, who self-identified as solely Native Hawaiian ancestry and reported only Native Hawaiian ancestry for each parent (herein we refer as 100% self-reported Native Hawaiian). Thirty and 35 individuals identified one parent belonging to either an Asian or European-ancestry descent group, respectively, which we classified as 50% Native Hawaiian. The remaining 25 and 18 individuals comprised of those who reported one of their parents heritage as

Asian and/or European, which we classified as 75% Native Hawaiian. We classified Asian heritage as those who reported themselves as Japanese, Korean, Chinese, and/or Filipino (see Table S1 and Table S2) for descriptions of each individual as well as the available data). This study was approved by the institutional review boards at the University of Hawaii and the University of Southern California. Written informed consent was obtained from all participants.

Genome-wide data and analysis

To gain a high-resolution estimate of genome-wide admixture in Native Hawaiians, we utilized genome-wide data for 148 Native Hawaiians (of whom 40 self-reported as 100% Native Hawaiians). Genotyping was conducted using the Illumina Infinium 660W bead array at the University of Southern California as part of an on-going genome-wide association study of breast cancer (Native Hawaiian cases/controls = 79/69). Of the 561,490 SNP probes, 72,529 were excluded due to a low minor allele frequency ($MAF < 0.01$), poor completion rate (< 0.95), or poor concordance (< 0.99) across 75 intended replicate samples. We selected a subset of individuals from the Human Genetic Diversity Panel (HGDP; $n = 466$) [68-70] (Table S1) that included samples from two Oceanic populations (Papuan and Melanesians), European ancestry (Utah residents from the CEPH population, Sardinians, Tuscans, French, Italians, and Russians), Africa (Yoruba, Biaka Pygmies, and San), South-Central Asia/Middle East (Sindhi and Palestinians), Central/South America (Mayans, Colombians, and Pima), and Asia (Han Chinese, She, Cambodians, Japanese, Yakut, and Yi). Individuals from Asian and Oceania populations were selected to represent potential source populations for Native Hawaiian while the remaining groups were chosen as contrast for population structure analyses. After initial quality control filtering, data was available for 670,372 and 488,961 SNPs for the HGDP and Native Hawaiian samples, respectively. Using PLINK [71], we further filtered the joint set of autosomal SNPs for genotyping call rates $> 99.5\%$, minor allele frequencies $> 5\%$, HWE ($P > 0.001$), and SNPs in linkage disequilibrium ($r^2 < 0.4$). In total, 114,112 SNPs and 614 individuals were used by ADMIXTURE [72] to estimate each individual's proportion of ancestry for varying values of K ancestral clusters in which K ranged from 2 to 8. We utilized PLINK [71] to conduct a multi-dimensional scaling (MDS) analysis of the genome-wide data to generate a graphical view of the genomic distance between individuals.

Mitochondrial genome data and analysis

To identify the mitochondrial haplogroup diversity of Native Hawaiians and estimate their population expansion, we sequenced the 16.5 kb mitochondrial genome in 160 100% self-reported Native Hawaiians, using the Affymetrix GeneChip Human Mitochondrial Resequencing Array 2.0. This chip provides an economical and highly accurate method for assessing variation in the mitochondrial coding regions [73]. Mitochondrial DNA was amplified in two PCR reactions using 100 ng of genomic DNA. Fragmentation, labeling, and chip hybridization was conducted according to manufacturer's instruction. Sequence analysis was conducted using the Affymetrix GeneChip Sequence Analysis Software (GSEQ) v4.1 software. GSEQ uses an objective statistical framework, based on the ABACUS algorithm, to assign base calls to each position according to quality criteria (default settings were used). The Revised Cambridge Reference Sequence (rCRS) was used as the reference sequence. The average base call rate of the mitochondrial genome was 90.6%. We validated the use of the Affymetrix Mitochondrial Resequencing Array by independently sequencing 2% replicate samples using the Pyrosequencing

technology. We observed a 100% concordant rate in sequence calls.

6,686 full mitochondrial sequences were downloaded from NCBI (as of February 2010) using the following command: Homo [Organism] AND gene_in_mitochondrion [PROP] AND 14000:19000 [SLEN] NOT pseudogene [All Fields]. Using build 8 of PhyloTree [46], we extracted 2,621 full mitochondrial sequences with previously annotated mitochondrial haplogroup assignments. Each mitochondrial sequence was then independently aligned to rCRS via KALIGN version 2.0.3 [74], and all variant positions were annotated relative to rCRS base positions. For each of the 2,781 total mitochondrial sequences (including the 160 Native Hawaiians we sequenced), we extracted all aligned base information, excluding the hypervariable region, and identified the diagnostic mitochondrial SNP (mtSNP) from MITOMAP [75] for a given mitochondrial haplogroup as well as any mtSNP with a minor allele frequency greater than 2%. MDS was then performed on 252 mtSNPs and 704 mitochondrial sequences (160 of which were Native Hawaiian) by using the subroutine *cmdscale* within R software (Tables S3 and S4). We chose MDS, rather than Principal Component Analysis, as it can handle potential missing data without imputation. Detailed mitochondrial haplogroup categorization was determined for each Native Hawaiian mtDNA sequence by the presence or absence of the B4 sub-lineage haplogroup mutations found in build 8 of PhyloTree [46].

We further used the 160 Native Hawaiian mitochondrial genomes to estimate demographic parameters of the Hawaiian population, using the Bayesian Skyline process in the BEAST software package [76]. This Bayesian algorithm estimates effective population size (N_e) through time via coalescent theory by using a nonparametric Markov Chain Monte Carlo. The Bayesian Skyline has been shown to be a highly accurate method of detecting fluctuations in population sizes in humans [77–79]. In particular, rapid expansions, such as island colonization, can be captured by this method.

In order to run Bayesian Skylines, we first isolated the coding region of each mitochondrial genome and aligned it to rCRS by using the coding region coordinates described in MITOMAP. These regions were analyzed using default settings of the BEAST software. As the observed mutation rate in mtDNA is known to decrease with a higher number of coalescent events [80], we kept the mutational clock in generation-based units in BEAST and applied a post-run correction for the mutation rate at each

generation. This method has been shown to be more accurate and consistent with archaeological dating methods than assuming a constant mutation rate through time [79]. There is still a debate in the field as to why this mutation rate slowdown occurs [81,82], however, when estimating the timing of demographic events in the recent past, the empirical estimator calibrated from archaeological evidence [76] and a model incorporating purifying selection [77] give qualitatively similar results.

Supporting Information

Figure S1 MDS analysis of mitochondrial haplogroup. (TIF)

Table S1 Sample information for GWAS and mitochondrial sequence data. (XLS)

Table S2 Summary counts of subjects used for AIMS, GWAS, and mtDNA analysis. (XLS)

Table S3 Individuals used for MDS mtDNA sequence analysis and corresponding haplogroup. (XLS)

Table S4 Corresponding mtDNA position numbers used for mtDNA sequence analysis and rCRS allele. (XLS)

Table S5 Nine diagnostic markers for mtDNA haplogroup Q1 and allelic states for individual 52 and rCRS. (XLS)

Acknowledgments

We are indebted to the Native Hawaiian participants of the Multiethnic Cohort Study. We thank Lauren Hu, Mira Sharma, Awapuhi Lee, Ann Seifried, and Maarit Tiirikainen for their contributions to this project.

Author Contributions

Conceived and designed the experiments: SKK CRG JDW ALJ CAH RS IC. Performed the experiments: SKK ALJ. Analyzed the data: SKK CRG ALJ GKC. Contributed reagents/materials/analysis tools: SKK CRG GKC. Wrote the paper: SKK CRG JDW ALJ HW CAH GKC BEH LNK LLM DOS RS IC.

References

- Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, et al. (2012) Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet* 8: e1002554.
- Wang H, Haiman CA, Kolonel LN, Henderson BE, Wilkens LR, et al. (2010) Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum Genet* 128: 165–177.
- Thomas DC, Witte JS (2002) Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 11: 505–512.
- Wacholder S, Rothman N, Caporaso N (2002) Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 11: 513–520.
- Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A* 85: 9119–9123.
- Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet* 55: 809–824.
- Reich D, Patterson N (2005) Will admixture mapping work to find disease genes? *Philos Trans R Soc Lond B Biol Sci* 360: 1605–1607.
- Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, et al. (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet* 37: 177–181.
- Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, et al. (2004) Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 74: 979–1000.
- Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. *Am J Hum Genet* 74: 965–978.
- Montana G, Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. *Am J Hum Genet* 75: 771–789.
- Henderson SO, Haiman CA, Wilkens LR, Kolonel LN, Wan P, et al. (2007) Established risk factors account for most of the racial differences in cardiovascular disease mortality. *PLoS One* 2: e377.
- Pike MC, Kolonel LN, Henderson BE, Wilkens LR, Hankin JH, et al. (2002) Breast cancer in a multiethnic cohort in Hawaii and Los Angeles: risk factor-adjusted incidence in Japanese equals and in Hawaiians exceeds that in whites. *Cancer Epidemiol Biomarkers Prev* 11: 795–800.
- Moy KL, Sallis JF, David KJ (2010) Health indicators of Native Hawaiian and Pacific Islanders in the United States. *J Community Health* 35: 81–92.
- Mau MK, Sinclair K, Saito EP, Baumhofer KN, Kaholokula JK (2009) Cardiometabolic health disparities in native Hawaiians and other Pacific Islanders. *Epidemiol Rev* 31: 113–129.

16. Maskarinec G, Grandinetti A, Matsuura G, Sharma S, Mau M, et al. (2009) Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort. *Ethn Dis* 19: 49–55.
17. Nordyke CE (1989) *The Peopling of Hawai'i*. Honolulu, HI: The University of Hawai'i Press.
18. Diamond JM (2000) Taiwan's gift to the world. *Nature* 403: 709–710.
19. Blust R (1999) Subgrouping, circularity, and extinction: some issues in Austronesian comparative linguistics. 31–94.
20. Gray RD, Jordan FM (2000) Language trees support the express-train sequence of Austronesian expansion. *Nature* 405: 1052–1055.
21. Diamond JM (1988) Express train to Polynesia. *Nature* 336: 307–308.
22. Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro AS, et al. (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol* 12: 604–615.
23. Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, et al. (2005) Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol* 3: e247.
24. Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. *Am J Hum Genet* 57: 1463–1475.
25. Melton T, Peterson R, Redd AJ, Saha N, Sofro AS, et al. (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *Am J Hum Genet* 57: 403–414.
26. Kirch PV (2000) *On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact*. London: University of California Press.
27. Kayser M, Brauer S, Weiss G, Underhill PA, Roewer L, et al. (2000) Melanesian origin of Polynesian Y chromosomes. *Curr Biol* 10: 1237–1246.
28. Capelli C, Wilson JF, Richards M, Stumpf MP, Gratrix F, et al. (2001) A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am J Hum Genet* 68: 432–443.
29. Underhill PA, Passarino G, Lin AA, Marzuki S, Oefner PJ, et al. (2001) Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific. *Hum Mutat* 17: 271–280.
30. Hurles ME, Nicholson J, Bosch E, Renfrew C, Sykes BC, et al. (2002) Y chromosomal evidence for the origins of oceanic-speaking peoples. *Genetics* 160: 289–303.
31. Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, et al. (2008) The genetic structure of Pacific Islanders. *PLoS Genet* 4: e19.
32. Kayser M, Lao O, Saar K, Brauer S, Wang X, et al. (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet* 82: 194–198.
33. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, et al. (2010) Demographic history of Oceania inferred from genome-wide data. *Curr Biol* 20: 1983–1992.
34. Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* 16: 551–558.
35. Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, et al. (2002) Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet* 70: 625–634.
36. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5: 435–445.
37. US Census 2000. Hawaii Tumor Registry, Cancer Research Center of Hawaii, University of Hawaii.
38. Lum A, editor (1988) *Sailing for the Sun: The Chinese in Hawaii, 1789–1989*. Honolulu: Three Heroes.
39. Nordyke EC, Lee RKC (1989) The Chinese in Hawai'i: A Historical and Demographic Perspective. *The Hawaiian Journal of History* 23: 196–216.
40. Mitchell BM, Salisbury RE (1999) Japanese Americans. *Encyclopedia of multicultural education*. Westport: Greenwood Publishing Group, Inc.
41. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, et al. (2006) Population stratification confounds genetic association studies among Latinos. *Hum Genet* 118: 652–664.
42. Murray T, Beaty TH, Mathias RA, Rafaels N, Grant AV, et al. (2010) African and non-African admixture components in African Americans and an African Caribbean population. *Genet Epidemiol* 34: 561–568.
43. Yaeger R, Avila-Bront A, Abdul K, Nolan PC, Grann VR, et al. (2008) Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. *Cancer Epidemiol Biomarkers Prev* 17: 1329–1338.
44. Lum JK, Rickards O, Ching C, Cann RL (1994) Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. *Hum Biol* 66: 567–590.
45. Bellwood P (1987) *The Polynesians. Prehistory of an island people*. New York: Thames and Hudson, Inc.
46. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30: E386–394.
47. Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, et al. (2007) Melanesian mtDNA complexity. *PLoS One* 2: e248.
48. Kayser M, Brauer S, Cordaux R, Casto A, Lao O, et al. (2006) Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol* 23: 2234–2244.
49. Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, et al. (2005) Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proc Natl Acad Sci U S A* 102: 13034–13039.
50. Li M, Reilly MP, Rader DJ, Wang LS (2010) Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics* 26: 798–806.
51. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998–1003.
52. Alexander DH, Lange K (2011) Stability selection for genome-wide association. *Genet Epidemiol* 35: 722–728.
53. Wang H, Haiman CA, Kolonel LN, Henderson BE, Wilkens LR, et al. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum Genet* 128: 165–177.
54. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* 4: e1000202.
55. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al. (2010) Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci U S A* 107 Suppl 2: 8954–8961.
56. Gravel S (2012) Population genetics models of local ancestry. *Genetics* 191: 607–619.
57. Biffi A, Anderson CD, Nalls MA, Rahman R, Sonni A, et al. (2010) Principal-component analysis for assessment of population stratification in mitochondrial medical genetics. *Am J Hum Genet* 86: 904–917.
58. Schurr TG, Wallace DC (2002) Mitochondrial DNA diversity in Southeast Asian populations. *Hum Biol* 74: 431–452.
59. Alexander WD (1891) *A brief history of the Hawaiian people*. New York: American Book Company.
60. Bryan WA (1915) *Natural history of Hawaii*. Honolulu: The Hawaiian Gazette Co., Ltd.
61. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22: 1185–1192.
62. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475: 493–496.
63. Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M (2011) High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res* 21: 1–11.
64. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, et al. (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 1: 131.
65. Kirch PV, Green RC (1987) History, Phylogeny, and Evolution in Polynesia. *Current Anthropology* 28: 431–456.
66. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
67. Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, et al. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* 151: 346–357.
68. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. *Science* 296: 261–262.
69. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. *Science* 298: 2381–2385.
70. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, et al. (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* 1: e70.
71. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
72. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19: 1655–1664.
73. Hartmann A, Thieme M, Nanduri LK, Stempf T, Mochle C, et al. (2009) Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum Mutat* 30: 115–122.
74. Lassmann T, Sonnhammer EL (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6: 298.
75. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, et al. (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* 35: D823–828.
76. Ho SY, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol* 22: 1561–1568.
77. Atkinson QD, Gray RD, Drummond AJ (2008) mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol* 25: 468–474.
78. Kitchen A, Miyamoto MM, Mulligan CJ (2008) A three-stage colonization model for the peopling of the Americas. *PLoS One* 3: e1596.
79. Gignoux CR, Henn BM, Mountain JL (2011) Rapid, global demographic expansions after the origins of agriculture. *Proc Natl Acad Sci U S A* 108: 6044–6049.
80. Henn BM, Gignoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol Biol Evol* 26: 217–230.

81. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 84: 740–759.
82. Loogvali EL, Kivisild T, Margus T, Villems R (2009) Explaining the imperfection of the molecular clock of hominid mitochondria. *PLoS One* 4: e8260.